Supporting Information

Jabara et al. 10.1073/pnas.1110064108

SI Materials and Methods

vRNA Extraction and cDNA Synthesis. Viral RNA was extracted from three plasma samples taken longitudinally from an individual infected with subtype B HIV-1 who was participating in a protease inhibitor efficacy trial (M94-247). Two samples were collected at ~6 mo before and immediately before the addition of the protease inhibitor ritonavir to a failed therapy regimen (plasma viral loads of 285,360 copies of viral RNA/mL and 321,100 copies of viral RNA/mL, respectively), and one sample was collected during ritonavir therapy (at approximately 2 mo on therapy, 349,920 copies of viral RNA/mL) but during a time of apparent intermittent compliance. For each plasma sample, vRNA was extracted from pelleted (25,000 × *g* for 2 h) viral particles using the QiaAMP Viral RNA kit (Qiagen). Approximately 10,000 copies of viral RNA from each sample were present in the cDNA synthesis reaction as described (1–3). The tagging primer used was, 5′-GCCTTGCCAGCACGCTCAGGCCTTGCA(BARCODE)CGNNNNNNNNTCCTGGCTTTAATTTTACTGGTACAGT-3′. The barcode represented TCA, GTA, and TAT for study days 58, 248, and 303, respectively. The 3′ end of the tagging primer targeted downstream of the protease coding domain (HXB2 2568–2594). The oligonucleotides were purchased from IDT and were purified by standard desalting.

Amplification of Tagged Sequences. The single-stranded cDNA was column purified using the PureLink PCR Purification Kit (Invitrogen), using Binding Buffer HC (high cutoff) and three washes to remove the cDNA primer. Primer removal was verified by electropherogram analysis using an Experion HighSense RNA microfluidic chip (Bio-Rad Laboratories). Samples were amplified by nested PCR, using upstream primers 5′-GAGAGACAGGCTAATTTTTTAGG-3′ (HXB2 2071–2093) and 5′-ATAGACAAGGAACTGTATCC-3′ (HXB2 2224–2243); the downstream primers targeted the 5′ portion of the cDNA tagging primer 5-GCCTTGCCAGCACGCTCAGGC-3′ then 5′-CCAGCACGCTCAGGCCTTGCA-3′. The PCR was done using Platinum Taq DNA Polymerase High Fidelity (Invitrogen). Each reaction contained 1× High Fidelity PCR Buffer, 0.2 mM of each dNTP, 2 mM $MgCl_2$, 0.2 μM of each primer, 1.5 units of Platinum Taq DNA Polymerase. The purified cDNA template was split to 2 × 50 μl for the first round PCR, and 1 μl of the purified first round product was used for nested PCR. Samples were denatured at 94 °C for 2 min, followed by 30 cycles of 94 °C for 15 s, 55 °C for 30 s, 68 °C for 1 min, and a final extension at 68 °C for 5 min.

Samples were column purified after the first round of PCR using the MinElute PCR Purification Kit (Qiagen), and eluted into 30 μl of buffer EB. Second round PCR product was gel purified using a 2% agarose gel and QIAquick gel extraction kit (Qiagen), with incubation of the solubilization buffer at room temperature. DNA was quantified by Qubit fluorometer using dsDNA High Sense assay (Invitrogen). Product generation, quality, and primer removal for both PCR rounds was verified using an Experion DNA microfluidic chip (Bio-Rad).

454 Pyrosequencing. Tagged samples from the three time points were combined and sequenced on the 454 GS FLX platform with XLR70 Titanium sequencing chemistry as per the manufacturer's instructions (Roche) but with under-loaded beads to minimize signal crosstalk. Sequences were processed from two independent 454 GS FLX Titanium runs (1/8th of a plate each).

Bioinformatic Pipeline for Raw Sequence Processing. A suite of programs was written to filter and parse raw 454 sequencing reads. In short, first, each sequence was placed in the correct orientation compared with a reference *pro* gene sequence. This alignment was then used to identify insertions or deletions caused by the 454 sequencing of homopolymers. When there was an insertion, the extraneous base was excised from the sequence. Deletions retained were largely resolved in the construction of the consensus sequence (see below). Second, they were evaluated for the presence of the cDNA primer 5′ tail, with the encoded information (barcode and Primer ID) exactly spaced. Third, individual sequences were binned by their barcodes, and then by their Primer ID. Fourth, sequences were trimmed to the protease coding domain (*pro* gene). Within a barcode bin, when three sequences contained an identical Primer ID, a consensus sequence was called by majority rule. Ambiguous nucleotide designations were used when there was a tie (Fig. S3*B*). Sequences are available under GenBank accession numbers JN820319-JN824997.

Population Analyses. A χ^2 test was used to test for significance changes in allele frequency between the two untreated time points. To control for multiple testing, collective assessment of significance was based on False Discovery Rate analysis (FDR = 0.05). Tests for linkage disequilibrium were computed by DnaSP v.5.10.01 (4). These tests were done on filtered populations devoid of sequences containing ambiguities or gaps. Tests for neutrality were computed by DnaSP and R (5) on filtered populations devoid of sequences containing ambiguities. Gaps and alleles represented by a single sequence were reverted to the consensus. Beta *P* values were calculated against the null hypothesis that $D = 0$, assuming that D follows a beta distribution after rescaling on [0, 1] (6).

Diversity across and within populations was computed through customized bioinformatics suites. Unfiltered sequences were used in the analysis, and ambiguities, gaps, and alleles represented by a single sequence were removed from the final tabulation (Fig. 2 and Table S1).

SNPs were graphically displayed through the *Highlighter* tool (www.hiv.lanl.gov).

Phylogenetic Resolution of Sequences. The phylogeny for the population of consensus sequences from all three time points was resolved using two alternative methods and on populations devoid of sequences containing gaps or ambiguities. When only one example of a SNP was present across all sequences, it was converted to the consensus on the assumption that it was likely generated by residual method error. First, the Neighbor-Joining tree using the Kimura translation for pairwise distance and a bootstrap of 100 iterations was constructed with QuickTree v.1.1 (7).

Second, Maximum likelihood phylogeny was inferred using the PHYLIP package, version 3.69 (8), and the calculated phylogeny is available upon request. The PHYLIP program *seqboot* was used to create 100 bootstraps. Resulting bootstraps were submitted to the PHYLIP program *dnamlk* for maximum likelihood inference subject to a strict molecular clock. The consensus tree of all boostrap results was constructed using the PHYLIP program *consense*.

Both phylogenetic trees were visualized by a customized modification of Figtree v.1.3.1. (9).

Additional Considerations. ***Degenerate base synthesis in the cDNA primer.*** The degenerate bases (Primer ID) in the cDNA synthesis primer were randomized using machine mixing during oli-

gonucleotide synthesis. All four DNA phosphoramidite monomer bases are introduced to the column at the same time, but due to slight differences in binding or delivery, a strict equimolar ratio of dA, dT, dC, and dG may not be realized, resulting in a Primer ID bias (Fig. S7). When there is a Primer ID bias, there is an increased probability that a particular Primer ID will tag multiple templates because sequence tags with over-represented nucleotides will be more abundant than sequence tags with under-represented nucleotides. Because the bias is amplified over the length of the Primer ID the skewing can be significant. We observed a bias of ~40% dC in one of our Primer ID syntheses, and at the extreme dC_8 would be present at a 40-fold excess over the sequence frequency expected if all nucleotides were present at equal concentration. Similarly, we observed 15% dA in one synthesis which would result in a 60-fold decrease in the expected frequency of dA_8. This appears to be the result in variation in primer synthesis because the bias varied in the different barcode bins and therefore was not a constant feature of the cDNA synthesis step. However, this phenomenon is somewhat mitigated when a consensus sequence is formed, as whatever template was resampled to the greater extent within a mixed Primer ID population would be recorded.

Frameshift mutations. Pyrosequencing commonly miscalls homopolymers, resulting in a frameshift mutation by either calling too few or too many nucleotides in the homopolymer run. The HIV-1 *pro* gene contains several homopolymeric stretches. We took advantage of a known length (conserved in a coding region) to align individual reads against a reference sequence. Given this bias we removed the insertions to retain the correct length of the homopolymer run. Deletions were retained. Through consensus sequence generation, the deleted base was often recovered when the other resampled reads contained the missing base. Although consensus sequence generation reduced the spread and frequency of deletions in the final resolved, consensus reads, it did not eliminate deletions altogether (Fig. S8).

1. Abrahams MR, et al. (2009) Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *J Virol* 83:3556–3567.
2. Keele BF, et al. (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci USA* 105:7552–7557.
3. Salazar-Gonzalez JF, et al. (2008) Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol* 82:3952–3970.
4. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.
5. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290.
6. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
7. Howe K, Bateman A, Durbin R (2002) QuickTree: Building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* 18:1546–1547.
8. Felsenstein J (2005) *PHYLIP (Phylogeny Inference Package) version 3.69* (Department of Genome Sciences, University of Washington, Seattle, WA).
9. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.

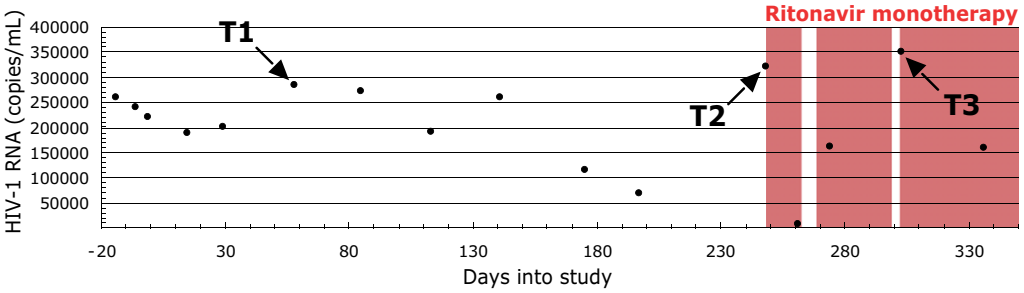

Fig. S1. Longitudinal sampling of blood plasma from a single individual infected with HIV-1 subtype B pre- and post- a failed ritonavir monotherapy regime. Two time-points ~6 mo apart were sampled before ritonavir therapy (T1 and T2). One time point was sampled after failed, intermittent ritonavir monotherapy (T3). The shaded areas represent times of therapy compliance based on self-report.

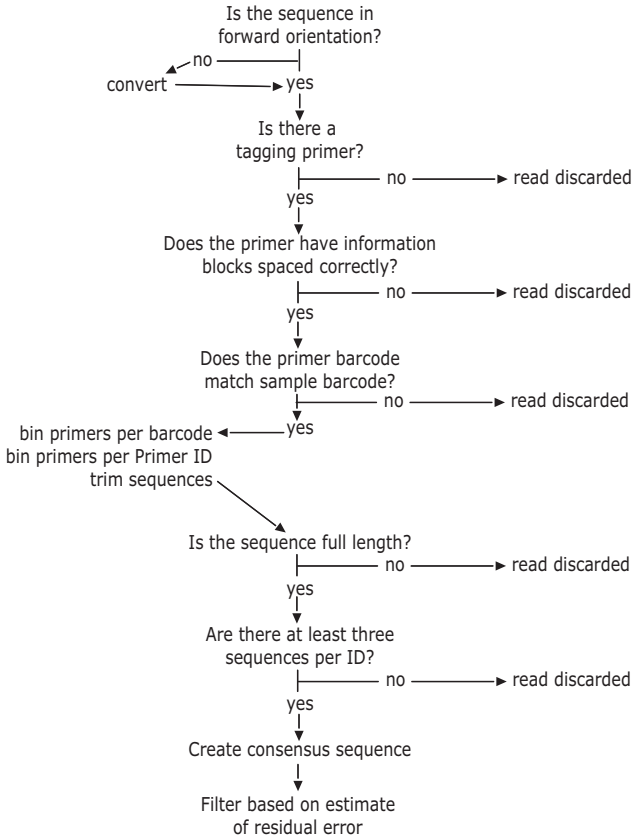

Fig. S2. Logic flow of the bioinformatic pipeline that processed raw sequence reads into consensus sequences. First, when applicable, reads were converted to forward orientation. Next, reads were assessed for the cDNA synthesis tagging primer containing correctly spaced sample and primer identifying information (barcode and Primer ID, respectively). Sequences were then binned based on the barcode, and within each barcode, binned by Primer ID, then trimmed to just the protease coding domain. For full-length protease sequences, when at least 3 sequences within a barcode bin contained an identical Primer ID, a consensus sequence was made based on majority-rule and the use of ambiguous nucleotide designations for ties. Sequences were then further filtered based on background estimates of error for the in vitro RT cDNA synthesis and the first round of Taq DNA polymerase synthesis.

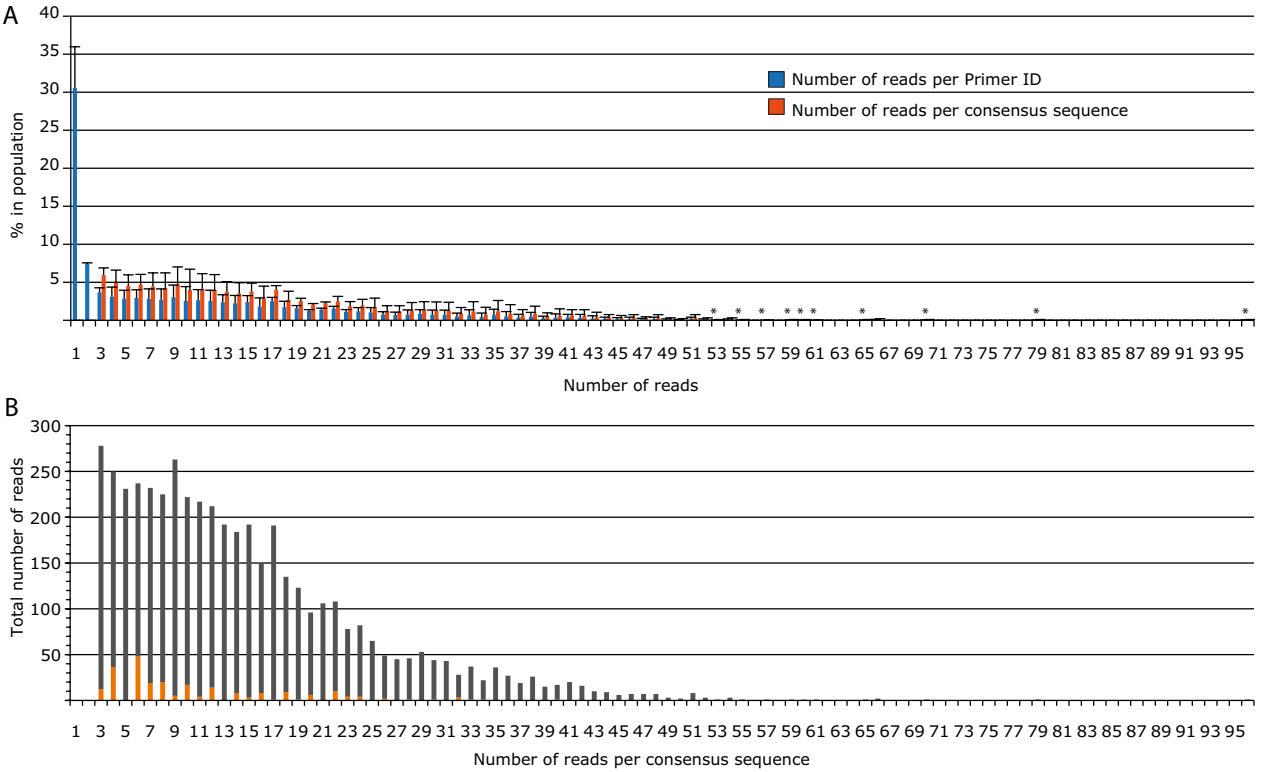

Fig. S3. (*A*) Distribution of the number of reads per Primer ID or consensus sequence. Blue bars represent the distribution of resampling of the filtered sequence population immediately before consensus sequence generation. Within a single Primer ID, when three or more sequences were present, a consensus sequence was formed. The orange bars represent the distribution of the number of reads that went into each consensus sequence. The values shown represent the mean for the data from the three time points with the error bars representing the SD between the three samples. Starred bars are included to mark positions where a single sequence had high resampling occurrence. (*B*) Number of consensus sequences containing an ambiguity as a function of extent of resampling. All three time points were combined. Gray bars represent consensus sequences without an ambiguity, and orange bars represent consensus sequences with an ambiguity. There is a discernible pattern of an increased number of ambiguities going out to 22 reads/consensus sequence for those consensus sequences created from an even number of reads, the result of having a tie between two different sequences at one position. However, this represents only a small fraction of the total reads (5.4%). The amino acid position with the highest ambiguity total was used per Primer ID subpopulation.

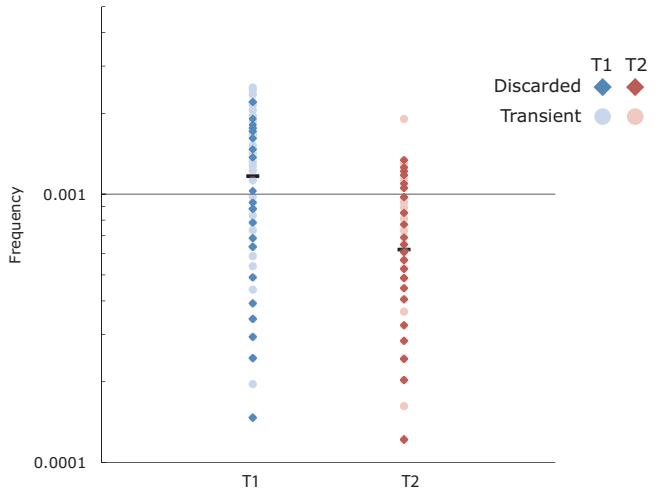

Fig. S4. Analysis of low abundance variants for the distribution of allelic skewing. We used discarded sequences (i.e., unique sequences represented by a single Primer ID) and transient genomes defined as having a low abundance SNP in the preconsensus population per untreated time point. Transient sequences were defined as having at least two sequences at only one of the untreated time points, or one copy at one of the untreated time points and then again at the third time point. These sequences were used to define a set of sequences that could be compared for low frequency abundance in the total data set versus the consensus sequences. The horizontal bars represent the measured frequency of a single copy sequences in the consensus population at T1 and T2. Dark points represent discarded genomes, and light points represent transient genomes with their position indicating their abundance in the total sequence population before construction of the consensus sequences. Blue points represent sequences present at T1, red points represent sequences present at T2. These data show that allelic skewing of 2-fold upward and 10-fold downward is common before the formation of the consensus sequence.

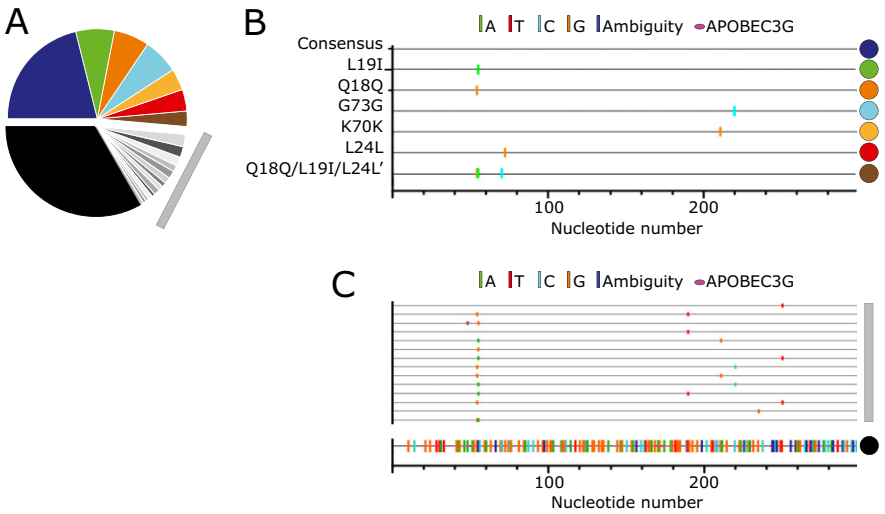

Fig. S5. Major and minor allelic variants in the untreated populations. (*A*) Frequency of major (colored) and minor (grayscale) unique *pro* gene sequences. Gray colors represent *pro* gene sequences present between 2.5 and 0.5% in frequency. Black represents the sum of all *pro* gene sequences individually present at <0.5%. (*B*) SNP distribution of the most abundant *pro* gene sequences (>2.5%), with the colored dots on the right indicating the corresponding sequences identified in the pie chart (*A*). (*C*) The gray bar corresponds to SNP distribution of variants present between 2.5 and 0.5%, the same sequences indicated in panel A with the gray bar. The line at the bottom indicated by the black circle represents the sum of all variants <0.5% in frequency for the sequences shown in black in the pie chart (*A*).

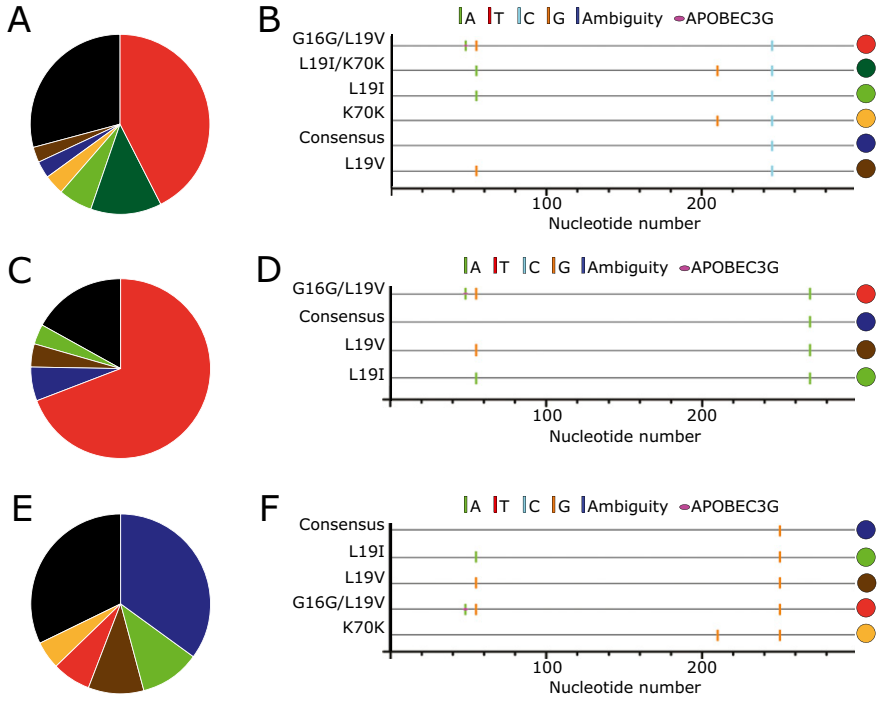

Fig. S6. Major and minor unique *pro* gene sequences in the major resistant populations V82A, L90M, and I84V. (*A*) Frequency of different unique *pro* gene sequences carrying the V82A mutation at high frequency (colored >2.5%) and low frequency (<2.5%, black and with the abundance pooled). (*B*) Highlighter plot showing the sequence changes from the consensus sequence for the major (>2.5%) *pro* gene variants carrying the V82A mutation. The V82A substitution is indicated by the nucleotide change at position 245 shown in light blue. (*C*) Frequency of different unique *pro* gene sequences carrying the L90M mutation at high frequency (colored >2.5%) and low frequency (<2.5%, black and with the abundance pooled). (*D*) Highlighter plot showing the sequence changes from the consensus sequence for the major (>2.5%) *pro* gene variants carrying the L90M mutation. The L90M substitution is indicated by the nucleotide change at position 268 shown in green. (*E*) Frequency of different unique *pro* gene sequences carrying the I84V mutation at high frequency (colored >2.5%) and low frequency (<2.5%, black and with the abundance pooled). (*F*) Highlighter plot showing the sequence changes from the consensus sequence for the major (>2.5%) *pro* gene variants carrying the I84V mutation. The I84V substitution is indicated by the nucleotide change at position 250 shown in orange.

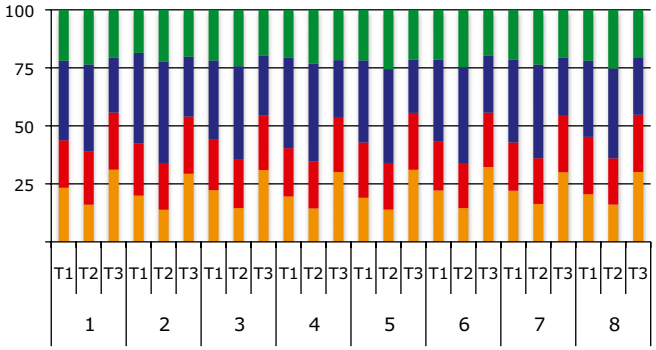

Fig. S7. Frequency of appearance of individual nucleotides at each Primer ID position (labeled 1–8) in resolved consensus sequences. Orange, red, blue, and green represent dA, dT, dC, and dG, respectively. On the horizontal axis, each Primer ID position is subdivided by time point (T1, T2, and T3).

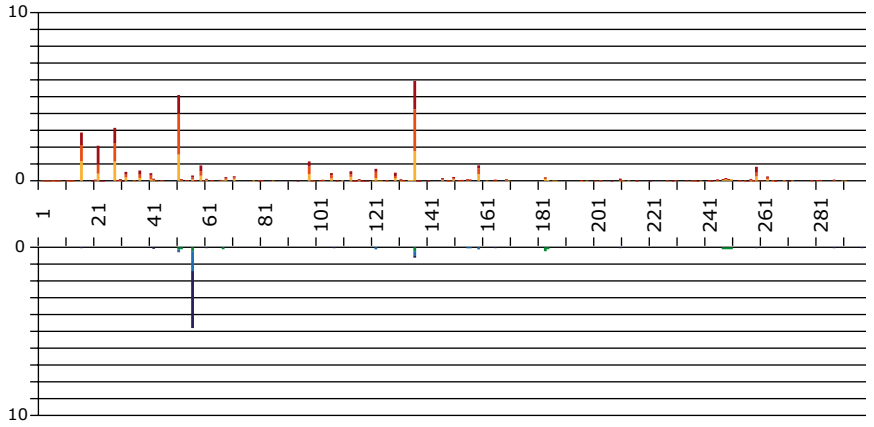

Fig. S8. Frequency of deletions in total versus consensus sequences. The percentage and nucleotide position of single nucleotide deletions are depicted in total (upward facing bars) or consensus (downward facing bars) sequences. Color corresponds to time point. T1 is yellow and green, T2 is orange and blue, and T3 is dark red and purple.

Table S1. Frequency of nonconsensus codons per position

Consensus			Nonsynonymous						Synonymous					
AA_{pos}[a]	AA_c[b]	C_c[c]	C_m[d]	AA_m[e]	T1[f]	T2[g]	T3[h]	$T3_s$[i]	$T3_r$[j]	C_m[k]	T1[l]	T2[m]	T3[n]	$T3_s$[o]	$T3_r$[p]
4	T	ACT	GCT	A		0.06	0.05		0.09						
5	L	CTT	CCT	P	0.12		0.05	0.14							
7	Q	CAA								CAG	0.35	0.12	0.09	0.14	0.09
8	R	CGA								CGG	0.12		0.05	0.14	
9	P	CCC													
10	L	CTC	TTC	F		0.19				CTT		0.19			
11	V	GTC	ATC	I	0.23	0.25				GTT		0.12			
14	K	AAG	AGG	R		0.12				AAA	1.17	0.19	0.59	0.29	0.72
15	I	ATA	GTA	V	1.17	0.12	0.14	0.14	0.18	ATC			0.09		0.18
16	G	GGG	AGG	R		0.06	0.05		0.09	GGA	2.22	3.54	38.86	17.70	45.97
17	G	GGG	AGG	R			0.09	0.29		GGA	0.35	0.19	0.18	0.43	0.09
18	Q	CAA	GAA	E	0.23	0.12				CAG	18.55	21.75	6.46	12.81	3.53
19	L	CTA	ACA	T	0.47										
			ATA	I	19.25	19.83	20.42	19.28	24.98	TTA	0.12	0.19	0.09	0.29	
			GTA	V	3.38	5.66	46.00	25.61	52.76						
20	K	AAG	AGG	R	0.12	0.12	0.05		0.09	AAA		0.31	0.86	0.29	1.27
21	E	GAA								GAG	0.12	0.06	0.05	0.14	
22	A	GCT								GCC	0.47	0.44	0.27	0.58	0.18
										GCG	0.23				
23	L	CTA								CTG		0.19			
24	L	TTA								CTA	0.35	5.72	1.31	2.16	0.63
										TTG	12.49	0.81	0.59	1.01	0.27
25	D	GAT	GGT	G	0.12	0.12				GAC	0.23	0.93	0.05	0.14	
26	T	ACA	GCA	A		0.12									
27	G	GGA								GGG	0.12	0.06			
28	A	GCA								GCG	0.12		0.09	0.14	
29	D	GAT	AAT	N	0.12		0.05		0.09	GAC	0.23	0.19			
30	D	GAT								GAC		0.06	0.09	0.14	0.09
31	T	ACA								ACG		0.12			
32	V	GTA								GTG		0.25			
33	L	TTA	GTA	V	0.47	0.06				CTA		0.25	0.14	0.29	0.09
										TTG	0.35	0.12	0.14	0.43	
34	E	GAA	GGA	G		0.12	0.05		0.09	GAG	0.12		0.05	0.14	
			CAA				0.09								
35	E	GAA	AAA	K	0.12	0.06	0.09	0.14							
36	M	ATG	ATA	I	0.82	0.81	0.27	0.43	0.27						
37	N	AAT	AGT	S		0.19	0.05			AAC		0.06	0.05	0.14	
			GAT	D	2.33	2.30	0.95	0.86	1.27						
38	L	TTG								TTA	0.23	0.62	0.05		0.09
39	P	CCA								CCT	0.23				
40	G	GGA								GGG	0.12	0.12			
41	K	AAA	AGA	R		0.06	0.18	0.14	0.27	AAG	4.08	1.43	0.50	1.15	0.27
42	W	TGG	CGG	R	0.12	0.06									
			TAG	_	0.12		0.05		0.09						
			TGA	_			0.14		0.27						
43	K	AAA	AGA	R		0.06	0.05		0.09	AAG	0.35		0.14	0.14	0.18
44	P	CCA								CCG		0.06	0.23	0.43	0.18
45	K	AAA	AGA	R	0.12	0.12	0.05		0.09	AAG	0.58	0.99	0.41	1.29	
46	M	ATG	ATA	I		0.12	0.09	0.14	0.09						
48	G	GGA	GAA	E			0.14	0.14	0.18	GGG	0.35	0.19			
49	G	GGA	GAA	E	0.12	0.06	0.05		0.09	GGG	0.23	0.12			
50	I	ATT								ATC	0.12	0.12			
51	G	GGA								GGG	0.12	0.06			
52	G	GGT	AGT	S	0.12	0.06	0.05	0.14		GGA		0.06	0.05	0.14	
										GGC	0.12	0.31	0.09	0.14	0.09
										GGG			0.14	0.43	
53	F	TTT								TTC	0.70		0.05	0.14	
54	I	ATC	ACC	T	0.12	0.06	0.05		0.09	ATT	0.35	0.06	0.14	0.14	
55	K	AAA	AGA	R	0.12		0.05		0.09	AAG	0.12	0.06			
56	V	GTA	ATA	I	0.12		0.05	0.14		GTG		0.75	0.14	0.14	0.18

Table S1. Cont.

Consensus			Nonsynonymous						Synonymous					
AA_{pos}[a]	AA_c[b]	C_c[c]	C_m[d]	AA_m[e]	T1[f]	T2[g]	T3[h]	$T3_s$[i]	$T3_r$[j]	C_m[k]	T1[l]	T2[m]	T3[n]	$T3_s$[o]	$T3_r$[p]
57	R	AGA	AAA	K	0.23					AGG	0.23	0.87	0.14	0.14	0.18
58	Q	CAG	TAG	_			0.05		0.09	CAA	0.93	0.50	0.23	0.29	0.27
60	D	GAT	AAT	N		0.12									
			GGT	G		0.12									
61	Q	CAA	CGA	R	0.12	0.06	0.05	0.14		CAG		0.19	0.23	0.58	
			TAA	_	0.12	0.06	0.05		0.09						
62	I	ATA	GTA	V	0.35	0.06									
63	L	CTC	CCC	P	0.12		0.41	0.58	0.36	CTT	11.32	5.41	1.27	2.88	0.45
64	I	ATA	GTA	V	1.05	0.06	0.09		0.18						
			ATG	M	0.23		0.05	0.14							
65	E	GAA	AAA	K			0.09	0.14	0.09	GAG	0.35	0.06	0.05		
66	I	ATC								ATA		0.25	0.18	0.58	
										ATT	1.98	0.19			
67	C	TGT								TGC	0.35	0.12	0.05	0.14	
68	G	GGA								GGG	0.23	0.12	0.05	0.14	
69	H	CAT	TAT	Y	0.23	0.06	0.09	0.14		CAC	0.82	0.31	0.14	0.29	0.09
70	K	AAA	CAA	Q	0.47	0.12	0.41	1.29		AAG	3.27	10.88	15.27	6.62	25.34
71	A	GCT	ACT	T		0.12	0.09								
72	I	ATA	GTA	V	0.12	0.12									
73	G	GGT								GGC	0.47	18.09	7.05	15.68	3.62
74	T	ACA								ACG	0.23	0.12			
75	V	GTA	ATA	I	0.23	0.06	0.05			GTG	1.87	0.99	0.27	0.43	0.27
			GCA	A			0.09		0.18						
76	L	TTA								CTA		0.12	0.09		0.18
										TTG	0.93	0.62	0.27	0.43	0.18
77	V	GTA	ATA	I	0.23	0.56	0.72	2.01	0.18	GTG	0.82	0.62	0.23	0.58	
			CTA	L			0.14								
78	G	GGA								GGG	1.17	1.24	0.09	0.14	
79	P	CCT								CCC	1.17	0.31	0.54	1.29	0.18
81	P	CCT								CCC	0.12	0.19			
										CCG	1.52	0.44			
82	V	GTC	ATC	I		0.06	1.27	3.60		GTA	0.35	0.31	0.05		
			CTC	L		0.06	1.08	3.45		GTT	1.05	0.75	0.41	1.01	
			GCC	A		0.12	49.89		99.91						
			TTC	F			0.14	0.43							
83	N	AAC	AGC	S	0.12		0.05		0.09	AAT	8.17	6.40	3.62	4.75	4.16
84	I	ATA	GTA	V			5.15								
85	I	ATT								ATA		0.12	0.05	0.14	
										ATC	0.12	0.12	0.05		
86	G	GGA								GGG		0.12			
										GGT	0.12	0.06			
87	R	AGA	AAA	K	0.12	0.06	0.05		0.09	AGG	0.58	0.37	0.05	0.14	
			GGA	G		0.06	0.09	0.14	0.09						
88	N	AAT								AAC	0.35	0.93			
89	L	CTA	ATA	I		0.12				CTG	1.17	0.68	1.36	1.87	1.54
										TTA	1.98	0.56	1.27	0.14	2.44
90	L	TTG	ATG	M	0.12		13.56		0.09	CTG	0.47		0.09	0.14	0.09
			TCG	S	0.12		0.05		0.09	TTA	0.47	0.19	0.14	0.43	
91	T	ACT	GCT	A		0.06	0.05		0.09	ACC	0.12	0.06	0.09	0.14	0.09
										ACG	0.12	0.12	0.77		1.54
92	Q	CAG								CAA	0.23	0.19	0.14		
93	I	ATT	CTT	L	0.12	0.06				ATC	0.23		0.09	0.14	0.09
94	G	GGT	GAT	D	0.12	0.06				GGA	0.23				
										GGC	1.28	0.25	0.50	1.29	0.18
										GGG	0.23	0.06	0.09	0.14	
95	C	TGC								TGT	0.70	0.12	0.14		0.27
96	T	ACT								ACA	0.12		0.09	0.14	0.09
										ACC	0.70	0.12	0.23	0.43	0.09
										ACG		0.06	0.05	0.14	
97	L	TTA								CTA	0.58		0.05	0.14	
										TTG	0.12	0.25	0.27	0.43	0.27

Table S1. Cont.

Consensus			Nonsynonymous							Synonymous					
AA_{pos}[a]	AA_c[b]	C_c[c]	C_m[d]	AA_m[e]	T1[f]	T2[g]	T3[h]	$T3_s$[i]	$T3_r$[j]	C_m[k]	T1[l]	T2[m]	T3[n]	$T3_s$[o]	$T3_r$[p]
98	N	AAT								AAC	0.23	0.12	0.14		0.18
99	F	TTT	CTT	L		0.06	0.18	0.29	0.09	TTC	1.05	0.50	1.54	1.44	1.54
			GTT	V	0.23										

Only positions of diversity and SNPs that were represented by more than 1 sequence are shown.
[a]Amino acid position, protease.
[b]Consensus amino acid in untreated population.
[c]Consensus codon in untreated population.
[d]Coding nonconsensus amino acid.
[e]Coding nonconsensus codon.
[f]Frequency of SNP in first untreated time point.
[g]Frequency of SNP in second untreated time point.
[h]Frequency of SNP in third time point, treated.
[i]Frequency of SNP in third time point, treated, susceptible population (not V82A, I84V, L90M).
[j]Frequency of SNP in third time point, treated, population containing major ritonavir resistant variant V82A.
[k]Silent nonconsensus codon.
[l]Frequency of SNP in first untreated time point.
[m]Frequency of SNP in second untreated time point.
[n]Frequency of SNP in third time point, treated.
[o]Frequency of SNP in third time point, treated, susceptible population (not V82A, I84V, L90M).
[p]Frequency of SNP in third time point, treated, population containing major ritonavir resistant variant V82A.

Table S2. Summary of nucleotide variation in sampled time points

Variable	T1	T2	T3	$T3_s$	$T3_r$
No. of sequences	810	1449	1925	547	970
No. of polymorphic (segregating) sites	104	115	110	71	69
Total number of mutations	115	129	121	75	73
Average number nt differences, k	2.38809	2.33683	3.08838	2.43819	2.05962
Nucleotide diversity, π	0.00804	0.00787	0.01040	0.00821	0.00693
Theta (per sequence)	14.29822	14.63943	13.51412	10.31864	9.25678
Theta (per site)	0.04814	0.04929	0.04550	0.03474	0.03117
Tajima's *D*	−2.3541	−2.3164	−2.0937	−2.1606	−2.1209
Beta *P* value	0.0013	0.0014	0.0070	0.0065	0.0071

T1 and T2 are untreated populations, and T3 is a population intermittently exposed to ritonavir monotherapy. Within T3, $T3_s$ represents the sensitive (not V82A, I84V, or L90M) portion of the population. $T3_r$ represents the major drug resistance clade V82A.