SUPPLEMENTARY DATA TO THE MANUSCRIPT: ERROR CORRECTION OF NGS DATA

OSVALDO ZAGORDI, ROLF KLEIN, MARTIN DÄUMER AND NIKO BEERENWINKEL

FIGURES

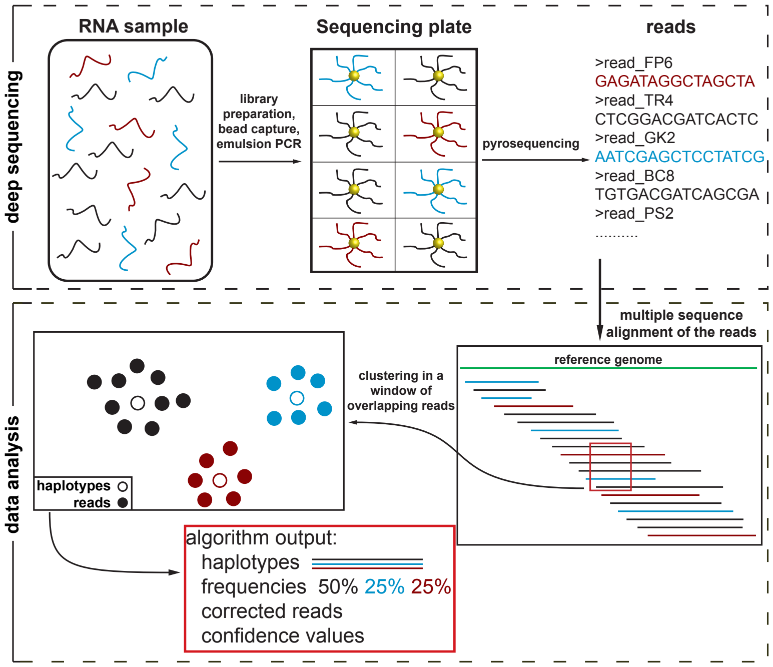

Figure S1. Schematic view of the analysis. The RNA sample is first reverse transcribed into DNA, then broken into small fragments (300-800 bp), converted into single stranded DNA and ligated with specific adapters. The ssDNA is immobilised onto special capture beads (shown in yellow). The emulsion PCR reaction enriches the beads such that each of them holds multiple copies of identical fragments. These beads are deposited onto a special sequencing plate in order to have on average one bead per micro well. The sequencing reaction takes place simultaneously and independently in each well, and one read is obtained from each bead. Reads are then aligned to the reference sequence and passed to the clustering algorithm. The proportion of initial sequences (shown in different colours) is maintained if the number of reads is high. The algorithm reports the haplotypes with a confidence value for their existence and their frequency in the sample.

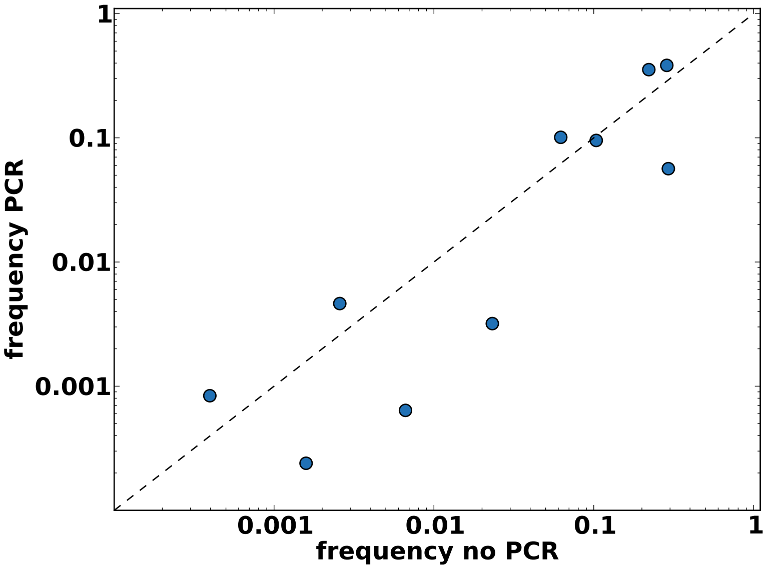

Figure S2. Frequency bias. Frequencies of the original haplotypes were estimated by mapping all reads obtained in the control experiments to all original sequences and assigning them to the best match (see Materials and Methods for details). In many cases, the proportions are not maintained after the PCR amplification.

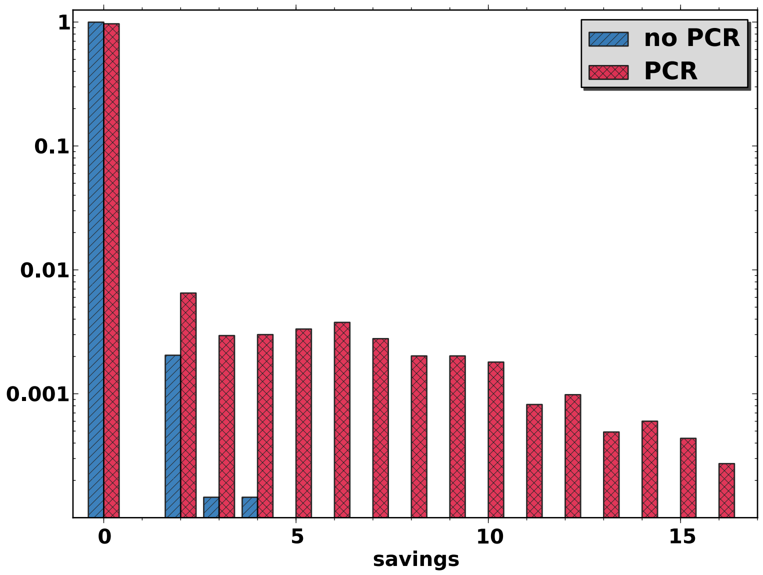

Figure S3. **Distribution of saved alignment costs saved when admitting a recombination event between two of the original haplotypes.** When reads are aligned to the haplotypes to find their closest match, there are cases where the alignment cost is lower if one considers the possibility that the read comes from a template obtained by the recombination of two of the original haplotypes. For the non-PCR dataset we observe values of savings as computed by the software *Recco* up to four, while in the PCR dataset 1.9% of the reads present higher values (up to 16).

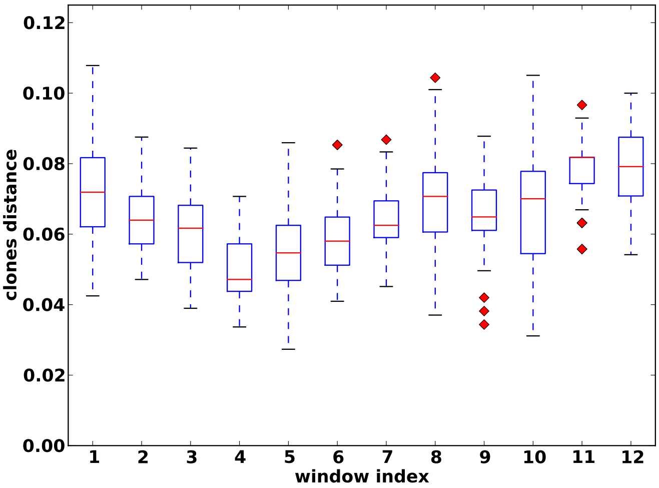

FIGURE S4. **Local diversity of the original haplotypes: windows drawn on the non-PCR-amplified sample.** The original haplotypes have an average distance of 6.8% among the 45 pairs when computed on their full sequence length. At a local level, the haplotypes display a wider range of diversity, as shown in several boxplots. The horizontal line and the box are median and lower-upper quartile, respectively. Whiskers and outliers show the range of values.

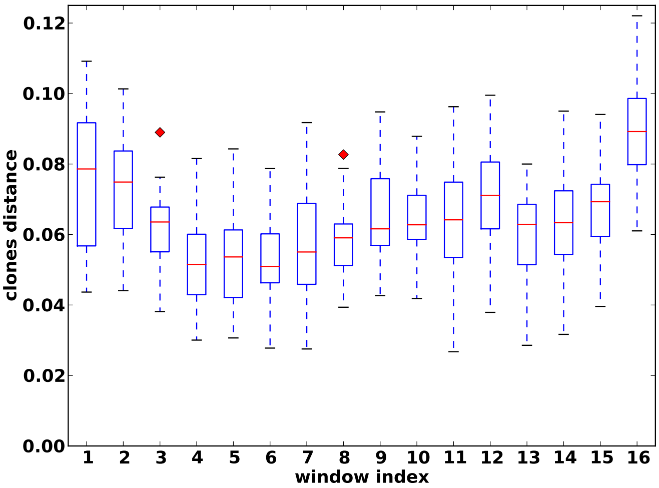

FIGURE S5. **Local diversity of the original haplotypes: windows drawn on the PCR-amplified sample.** The original haplotypes have an average distance of 6.8% among the 45 pairs when computed on their full sequence length. At a local level, the haplotypes display a wider range of diversity, as shown in several boxplots. The horizontal line and the box are median and lower-upper quartile, respectively. Whiskers and outliers show the range of values.

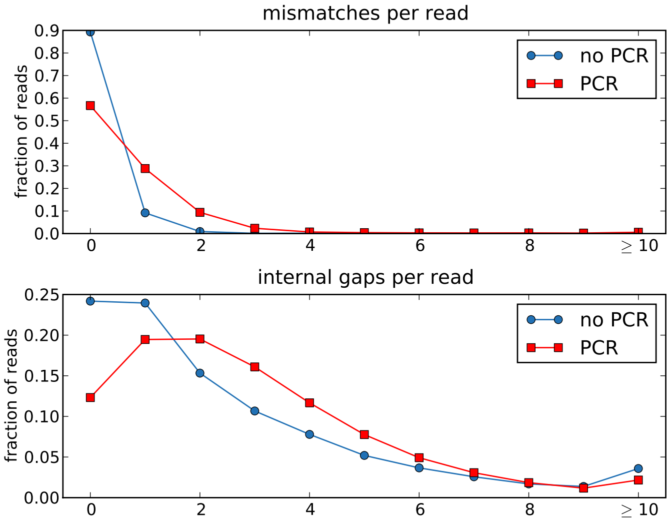

Figure S6. **Comparison of mismatches and gaps found on the reads.** Reads were aligned to all haplotypes in order to find their best match. The distribution of substitutions and gaps found on each read is reported for the non-PCR-amplified sample and for the amplified one.

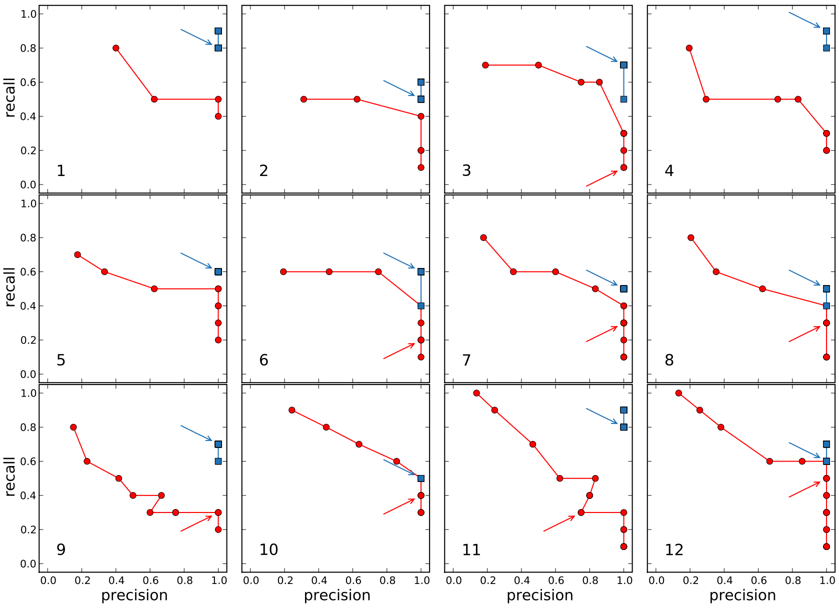

FIGURE S7. **Precision-recall analysis in individual windows: results for the non-PCR-amplified sample.** Red circles represent precision-recall values for the cut-off method, blue squares represent result for the clustering method. Arrows mark performance at values of threshold of 50 and 0.9 for the cut-off and clustering method, respectively.

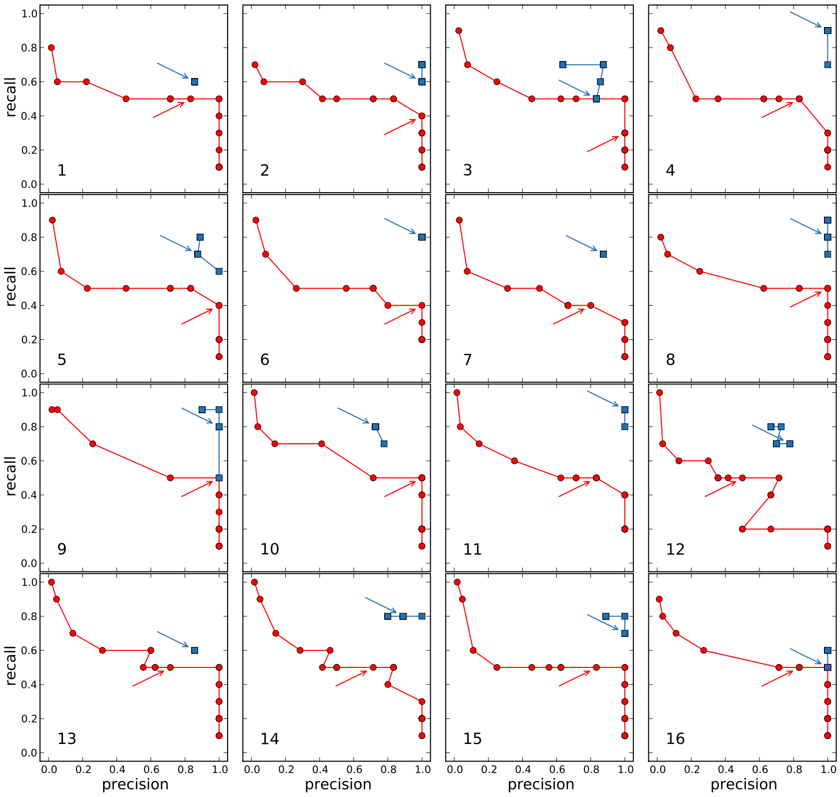

Figure S8. **Precision-recall analysis in individual windows: results for the PCR-amplified sample.** Blue circles represent precision-recall values for the cut-off method, red squares represent result for the clustering method. Arrows mark performance at values of threshold of 50 and 0.9 for the cut-off and clustering method, respectively.

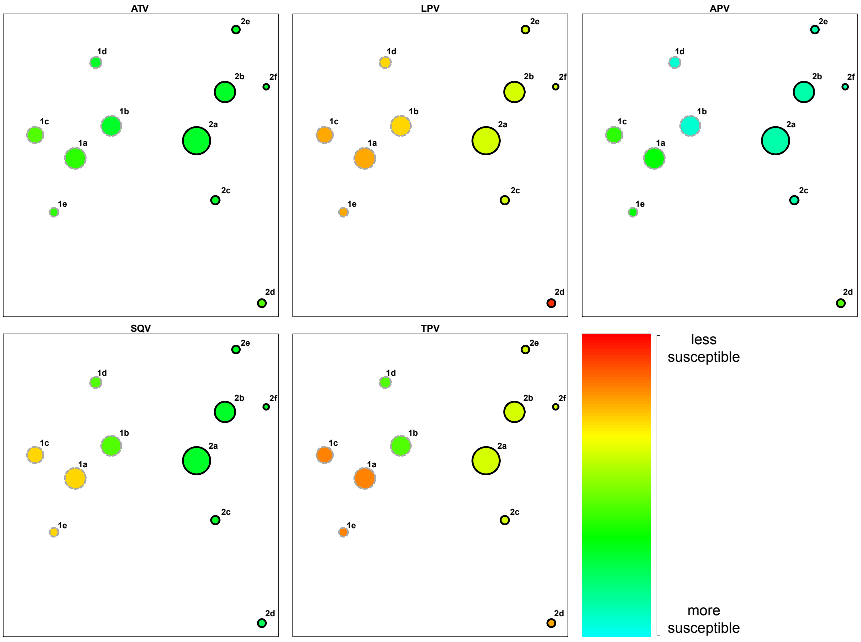

Figure S9. **Structure of the viral quasispecies and predicted resistance to five protease inhibitors.** Circles represent detected haplotypes translated into amino acid sequences. The size reflects the frequency of the amino acid sequences, while the fill colour indicates the predicted resistance to the protease inhibitors atazanavir (ATV), lopinavir (LPV), amprenavir (APV), saquinavir (SQV) and tipranavir (TPV), respectively shown in subfigures. Green indicates higher and red lower levels of predicted drug susceptibility. Haplotypes inferred in patient 1 are depicted with dashed grey border line, while haplotypes in patient 2 with a black border line. The circles are positioned in the plot such that their distance approximately preserves the Hamming distance of the amino acid sequences. The number and the letter next to each circle denote, respectively, the patient and the protease sequence reported in Table S1.

TABLES

Table S1. **Frequencies of the protease haplotype** The reconstructed haplotypes have been translated into amino acid sequences and are represented by a list of amino acid substitution with respect to the consensus sequence (see text). The table reports predicted fold change resistance to five antiviral drugs, the sum of the frequencies and the number of nucleotide haplotypes (nt hap) translated into the same protease. The higher coverage obtained in the sample from patient 2 allows to detect rarer variants, one of which (haplotype d) with predicted higher resistance to four out of five drugs. Letters in the second column (hap) mark haplotypes in Figure S9.

pat	hap	L10	V11	A22	D25	V32	M46	Q58	I62	C67	T74	L89	ATV	LPV	APV	SQV	TPV	%	# nt hap
1	a	I							V			M	2.0	3.5	1.7	2.7	2.1	46.7	5
1	b	I							V				1.7	3.2	1.1	1.9	1.6	36.0	3
1	c	I										M	2.0	3.4	1.7	2.7	2.0	12.8	3
1	d	I							V	S			1.7	3.2	1.1	1.9	1.6	3.0	1
1	e	I				A			V			M	2.0	3.5	1.7	2.7	2.1	1.5	1
2	a	V							V				1.8	2.7	1.3	1.5	1.8	86.7	9
2	b	V	A						V				1.8	2.7	1.3	1.5	1.8	12.1	2
2	c	V			G				V				1.8	2.7	1.3	1.5	1.8	0.4	1
2	d	V			G		I		V		A		2.2	4.7	1.9	1.4	2.0	0.3	1
2	e	V	A					H	V				1.8	2.7	1.3	1.5	1.8	0.3	1
2	f	V	A	T					V				1.8	2.7	1.3	1.5	1.8	0.2	1