Supplementary Material for "Characterization of mutation spectra with ultra-deep sequencing: application to HIV-1 drug resistance"

C. Wang et. al.

Table 1: GS20 Sequencing and Mapping Statistics for Four Plasmid Clones and Eight Clinical Samples

Sample	Sequencing				Mapping	
	Reads	Bases	Length	Quality	Percentage[c]	Coverage[d]
NL43[a]	5894	612551	104(±15)	26(±7)	97.7	357
NL43[ab]	7755	803906	104(±15)	27(±7)	97.4	463
7295[ab]	9377	996894	106(±13)	27(±7)	97.1	576
7303[ab]	6557	672347	103(±14)	26(±6)	96.3	383
92_2664[b]	4624	471027	102(±19)	27(±7)	83.8	364
F56272	9956	1092790	110(±16)	27(±7)	91.0	921
V10606	5286	545201	103(±18)	27(±8)	86.3	436
V11909	5177	530512	102(±18)	28(±9)	88.9	437
V47683	9419	1002358	106(±14)	27(±7)	93.2	855
V63557	5546	584861	105(±14)	27(±7)	89.3	483
V65472	6486	687950	106(±16)	27(±7)	91.5	580
V9878	5857	605960	103(±18)	27(±8)	91.1	509
Average	6827	717196	105(±16)	27(±7)	92.0	530

[a] Clonal sequences between 2147-3779 were amplified from recombinant HIV plasmids, the others are RT-PCR products from clinical samples between 2211-3279.

[b] Sequencing was done at the Stanford Genome Technology Center; the others were done at the 454 Life Science Company, Branford, CT.

[c] The percentage of total nucleotides mapped onto the entire HIV-1 genome sequence (accession number M19921).

[d] For clonal sequences, the coverage averaged over HIV genome region between 2147-3779. For the clinical samples, the coverage averaged over HIV genome regions between 2211-3279.

Table 2: Performance of Three Alignment Algorithms when GS20 Plasmid Clones are Aligned to a Subtype C Consensus Sequence

Sample	Percentage of nucleotides			Mismatch error rate[a]		
	ASW	SW	BLAST	ASW	SW	BLAST
NL43	91.4	89.5	86.3	0.0024	0.0027	0.0024
NL43	91.7	90.3	88.2	0.0017	0.0023	0.0019
7295	90.6	89.4	87.9	0.0027	0.0030	0.0029
7303	87.7	86.0	84.3	0.0013	0.0017	0.0014
Average	90.4	88.8	86.7	0.0020	0.0024	0.0022

[a] The mismatch error rate is calculated by comparing each mapped read with the consensus sequencing from all mapped reads.

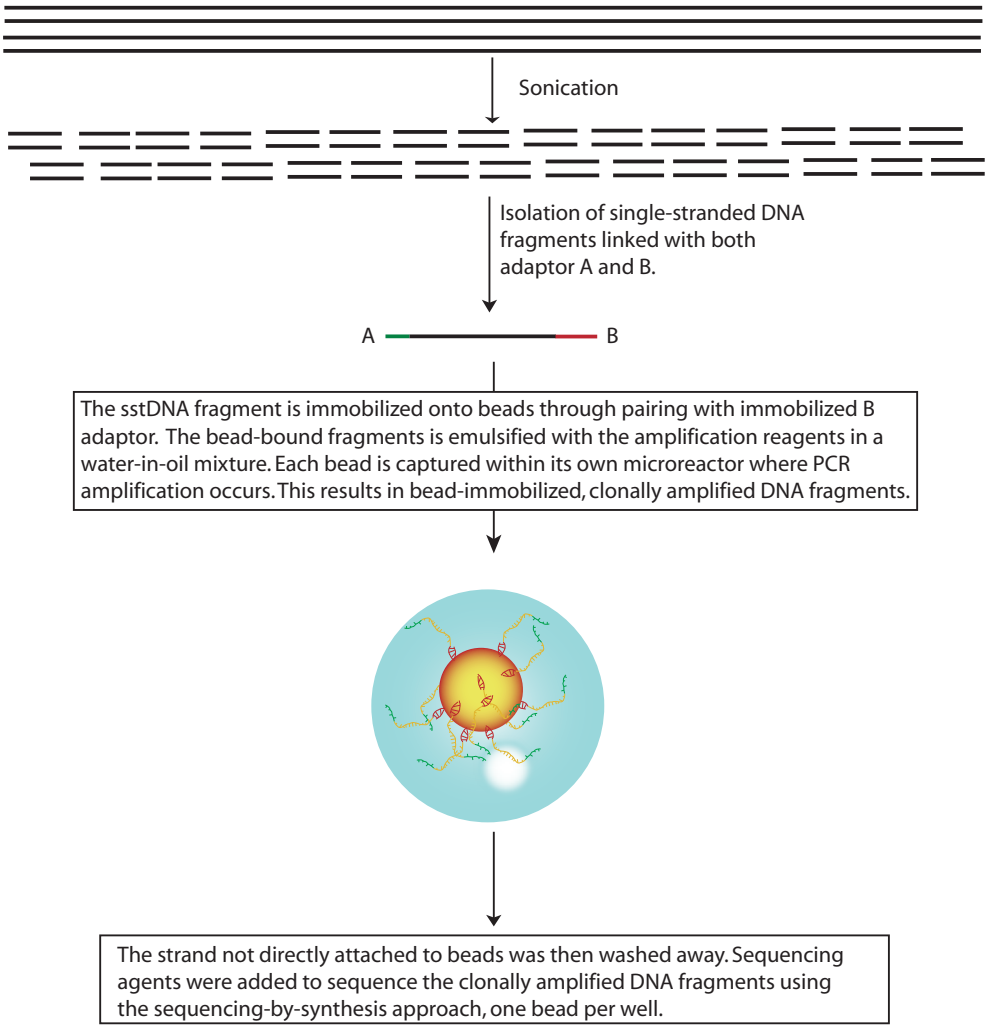

Figure 1: The diagram of the ultra-deep pyrosequencing procedure.

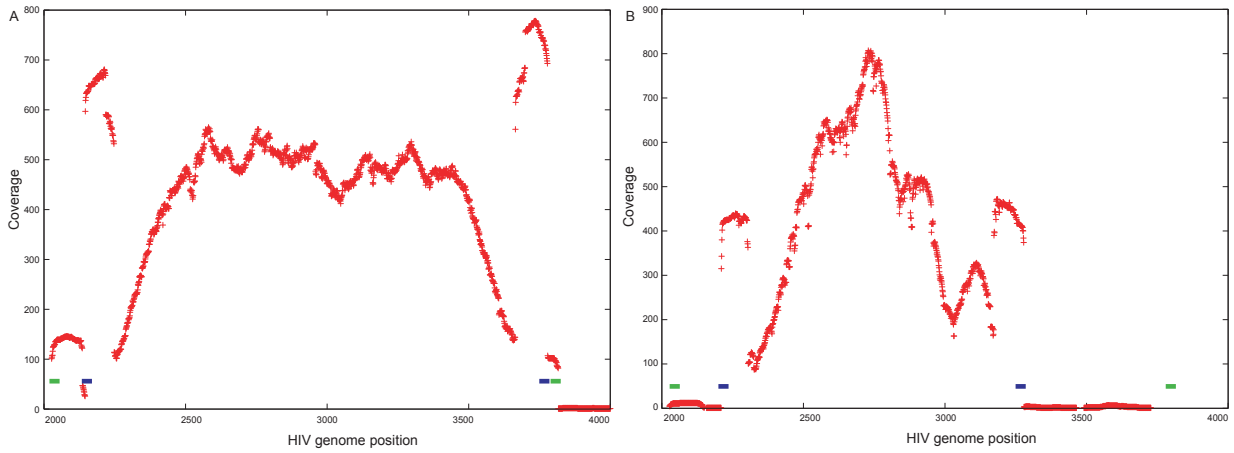

Figure 2: The distribution of coverage for NL43 plasmid clonal sequences (Panel A) and for the clinical sample V11909 (Panel B) mapped on HIV-1 NL43 (accession number M19921). The first round PCR primers for both samples (marked as green bars) are TTGGAAATGTGGAAAGGAAGGAC> and <CTCAAACAGTTATGGGGAGGGAATCA. The second round PCR primers for NL43 plasmid clone (blue bars on Panel A) are CAGAGCCAACAGCCCCACCA> and <CCTGTCTCATAACCGTTCGGTGGACC. The second round PCR primers for the clinical sample V11909 (blue bars on Panel B) are GAAGCAGGAGCCGATAGACAAGG> and <GGACTATTTACCTGTCATGTCCGAT.

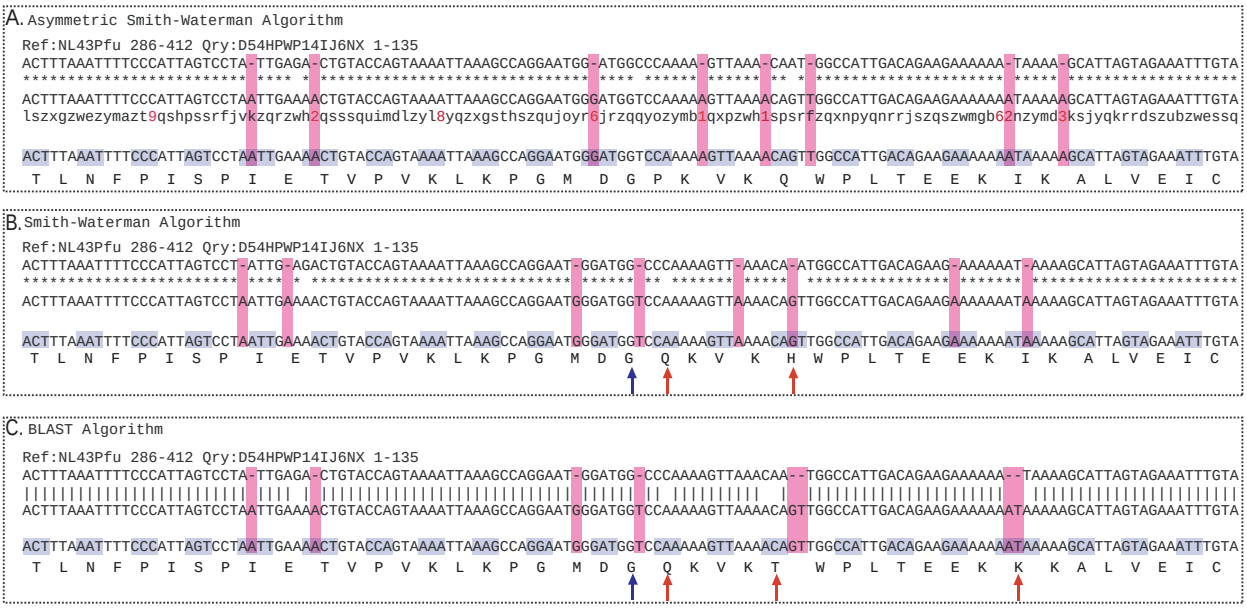

Figure 3: A GS20 read mapped to the Sanger sequence using different alignment methods. The upper nucleic acid sequence in each panel is the sequence of the NL43 clone between positions $286-412$ (encompassing the last four protease and first 38 RT codons). The middle nucleic acid sequence in each panel is the same GS20 read aligned slightly differently by the Asymmetric Smith-Waterman (A), Smith-Waterman (B), and BLAST algorithms (C). Panel A shows the Phred-equivalent quality scores beneath the GS20 read. Quality scores between 0 to 9 are indicated with the appropriate digit (red); scores between $10-35$ are indicated using the lower case alphabet in ascending order. The predicted insertion (all of which are presumed to be artifactual sequencing errors) are highlighted with red rectangles. Alternate grey shading is used to highlight codon grouping. The arrows demonstrate nucleotides that differ in the final sequence following insertion removal. Red arrows indicate those differences that result in amino acid changes and blue arrows indicate those that result in silent changes. Only the ASW algorithm produces the correct amino acids sequence.

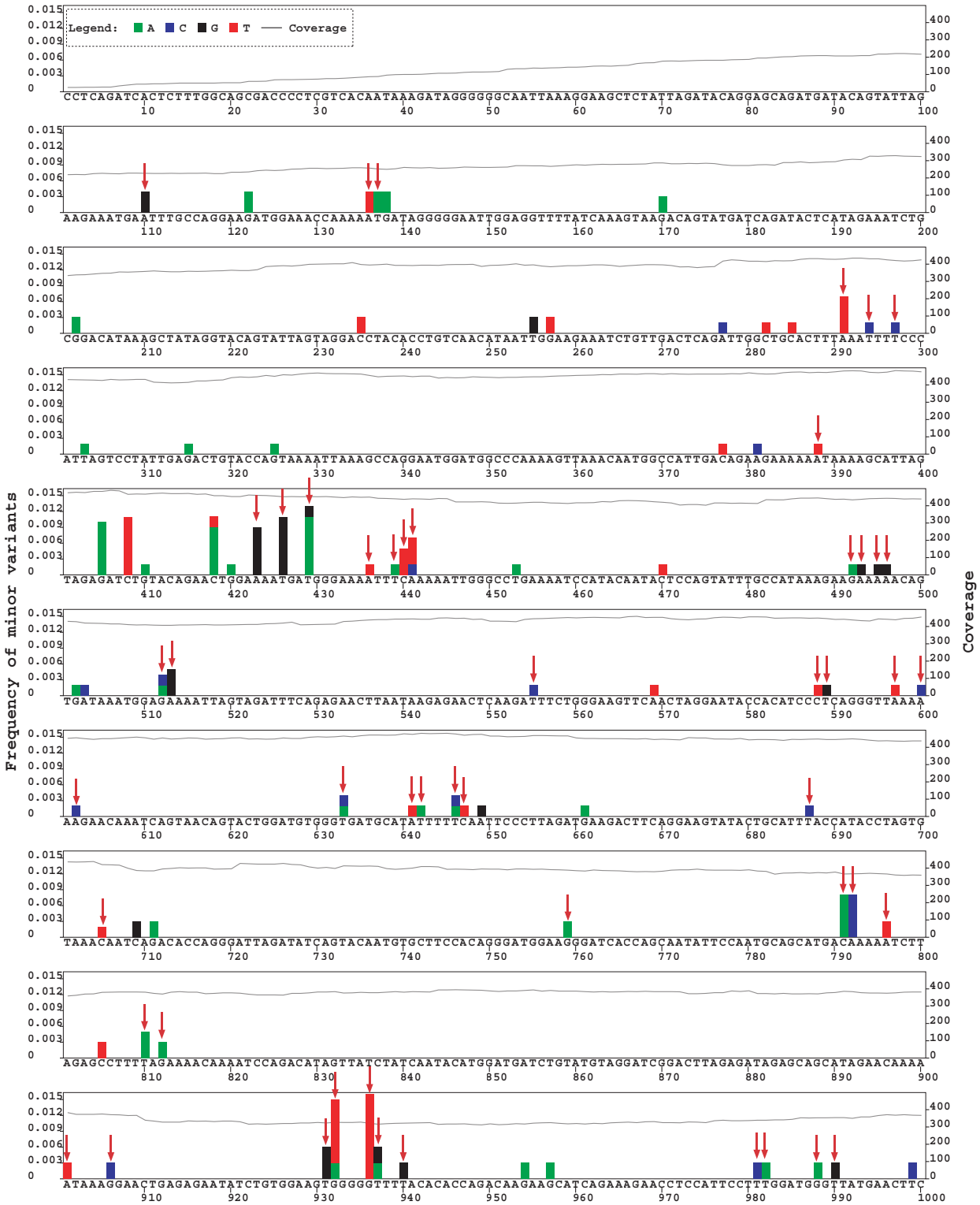

Figure 4: Sequence errors for the first 1,000 nucleotides of plasmid clone 7303 including complete protease and part of RT coding region. The sequence at the bottom of each lane represents the consensus sequence. Sequencing errors in homopolymeric regions are marked with red arrows.

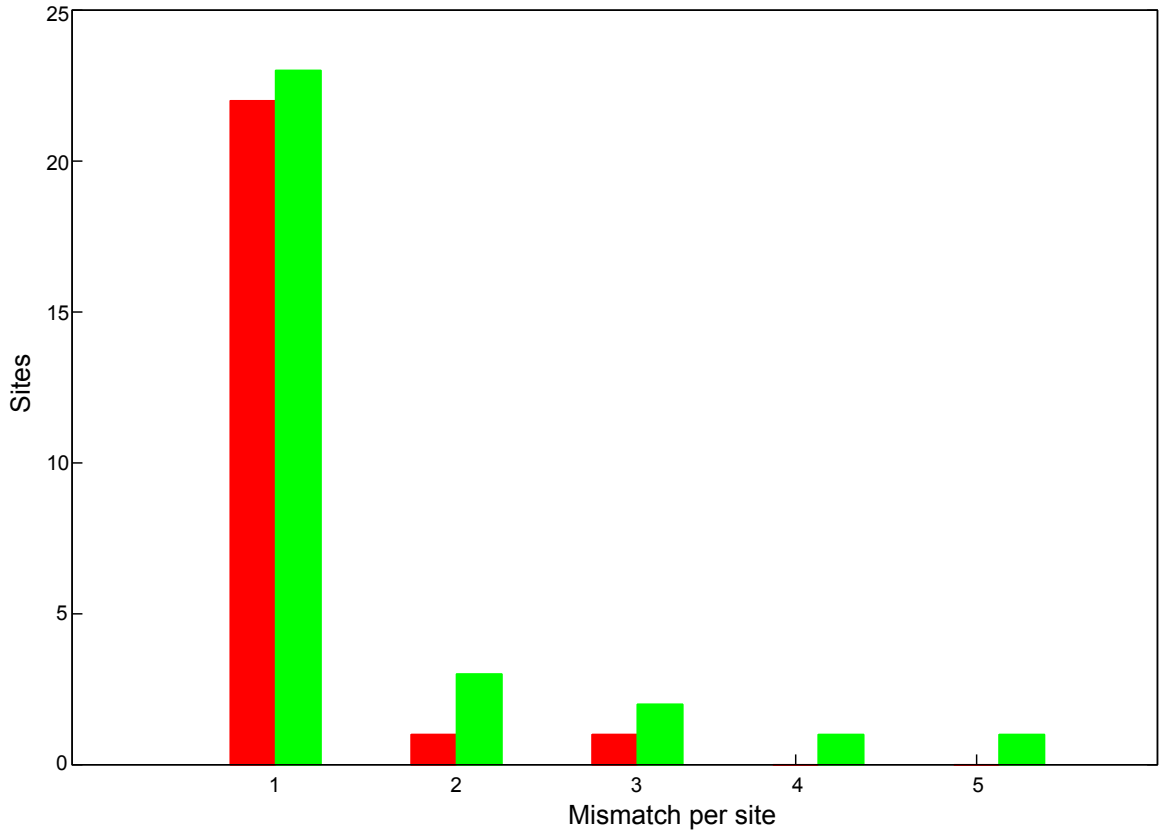

Figure 5: Distribution of the number of sites with one, two, three, four, and five errors in homopolymeric (green bars) and non-homopolymeric regions (red bars) for the plasmid clone 7303 in regions with coverage between 300 to 400.