TITLE

HIV-1 5'-Leader Mutations in Plasma Viruses Before and After the Development of Reverse Transcriptase Inhibitor-Resistance Mutations

SHORT TITLE

HIV-1 5'-Leader Plasma Virus Sequences

AUTHORS

Janin Nouhin[1]*, Philip Tzou[2]*, Soo-Yon Rhee[2]*, Malaya K. Sahoo[3], Benjamin A. Pinsky[4], Miri A. Krupkin[5], Joseph D. Puglisi[5], Elisabetta V. Puglisi[5], Robert W. Shafer[2]

*Each of these authors contributed equally to the work

AFFILIATIONS

[1]Division of Infectious Diseases, Stanford University, Stanford, CA, USA and Virology Unit, Institut Pasteur du Cambodge, Pasteur Network, Phnom Penh, Cambodia; [2]Division of Infectious Diseases, Stanford University, Stanford, CA, USA; [3]Department of Pathology, Stanford University, Stanford, CA, USA; [4]Division of Infectious Diseases, Department of Medicine, Stanford University, Stanford, CA, USA; Department of Pathology, Stanford University, Stanford, CA, USA; [5]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, USA.

Abstract word count: 300

Text word count: 2,897

ABSTRACT

Background: HIV-1 RT initiation depends on interaction between viral 5'-leader RNA, RT, and host $tRNA3_{Lys}$. We therefore sought to identify co-evolutionary changes between the 5'-leader and RT in viruses developing RT-inhibitor resistance mutations. **Methods:** We sequenced 5'-leader positions 37-356 of paired plasma virus samples from 29 individuals developing the NRTI-resistance mutation M184V, 19 developing an NNRTI-resistance mutation, and 32 untreated controls. 5'-leader variants were defined as positions where ≥20% of NGS reads differed from the HXB2 sequence. Emergent mutations were defined as nucleotides undergoing ≥4-fold change in proportion between baseline and follow-up. Mixtures were defined as positions containing ≥2 nucleotides each present in ≥20% of NGS reads. **Results:** Among 80 baseline sequences, 87 positions (27.2%) contained a variant; 52 contained a mixture. Position 201 was the only position more likely to develop a mutation in the M184V (9/29 vs. 0/32; p=0.0006) or NNRTI-resistance (4/19 vs. 0/32; p=0.02; Fisher's Exact Test) groups than the control group. Mixtures at positions 200 and 201 occurred in 45.0% and 28.8%, respectively, of baseline samples. Because of the high proportion of mixtures at these positions, we analyzed 5'-leader mixture frequencies in two additional datasets: five publications reporting 294 dideoxyterminator clonal GenBank sequences from 42 individuals and six NCBI BioProjects reporting NGS datasets from 295 individuals. These analyses demonstrated position 200 and 201 mixtures at proportions similar to those in our samples and at frequencies several times higher than at all other 5'-leader positions. **Conclusions:** Although we did not convincingly document co-evolutionary changes between RT and 5'-leader sequences, we identified a novel phenomenon, wherein positions 200 and 201, immediately downstream of the HIV-1 primer binding site exhibited an extraordinarily high likelihood of containing a nucleotide mixture. Possible explanations for the high mixture rates are that these positions are particularly error-prone or provide a viral fitness advantage.

49 **INTRODUCTION**

50 The first 356 nucleotides of the RNA genome, the 5'-leader extends from the start of the
51 transactivating response (TAR) element at HXB2 position 455 in the proviral DNA sequence to HXB2
52 position 810 (1). The 5'-leader exists as a mixture of monomeric and dimeric species and its secondary
53 structure has been studied using biochemical, NMR(2) , and cryo-EM (3,4) studies. The 5'-leader has
54 multiple functional units including TAR, which is required for initiating transcription; several RNA
55 elements required for RT initiation including the invariant primer binding site (PBS); and several
56 downstream elements responsible for viral splicing, dimerization, and packaging.

57 One rate-limiting step in retroviral replication is the initiation of reverse transcription. This
58 process requires a three-way interaction between the viral RNA genome, the reverse transcriptase (RT)
59 enzyme, and $tRNA3_{Lys}$. During this process, RT binds to the viral RNA-$tRNA3_{Lys}$ duplex forming the reverse
60 transcription initiation complex (3). RT initiation represents a bottleneck to HIV-1 replication as
61 evidenced by an approximately 100-3000-fold reduced rate of nucleotide incorporation during initiation
62 compared with elongation and by the much slower synthesis of the first 200 HIV-1 nucleotides
63 compared to the remaining nucleotides of proviral DNA (5).

64 We hypothesized that during RT initiation there may be an interplay between changes in RT and
65 the 5-leader viral RNA. In particular, we questioned whether the 5'-leader would evolve to adapt to the
66 development of drug-resistance mutations (DRMs) in the RT enzyme. Therefore, we sequenced the 5'-
67 leader region of plasma virus samples obtained from individuals before and after the development of
68 the most common nucleoside RT inhibitor (NRTI) mutation M184V or a non-nucleoside RT inhibitor
69 (NNRTI)-resistance mutation.

70

71 **METHODS**

72 **Individuals and samples**

73 We identified individuals undergoing two or more genotypic resistance tests at different times
74 for clinical purposes who had cryopreserved remnant plasma samples meeting one of the following
75 criteria: (1) Had a baseline sample lacking the 3TC/FTC-resistance mutation M184V/I and a follow-up
76 sample containing M184V ("M184V group"); (2) Had a baseline sample lacking an NNRTI-resistance
77 mutation and a follow-up sample containing an NNRTI-resistance mutation ("NNRTI group"); and (3) Had
78 two or more samples at different times lacking M184V/I or an NNRTI resistance mutation ("Control
79 group"). Individuals meeting the first criteria were required to not have a history of a sample with
80 M184V/I and individuals meeting the second criteria were required to not have a history of a sample
81 with an NNRTI-resistance mutation. To be eligible for next-generation sequencing (NGS), samples were
82 required to have a plasma HIV-1 RNA level ≥10,000 copies/ml.

83 The remnant plasma samples used in this study were obtained between the years 2000 and
84 2017. The Stanford University Institutional Review Board approved this study under protocol 6633,
85 entitled "Human Immunodeficiency Virus Quasispecies During Antiviral Therapy". The approval included
86 a waiver of consent because the study used anonymized de-identified data and remnant plasma
87 samples. The authors did not access data enabling them to identify individual participants. The samples
88 were sequenced and the data were collected for research purposes between July 2019 and January
89 2020 and between September 2020 and January 2021.

90

91 **Sequencing protocol**

92 Total nucleic acids were extracted from 200 uL of plasma using the EZ1 Virus Mini Kit V2.0 on
93 the automated EZ1 advanced XL (both from Qiagen), according to manufacturer's instructions. Nucleic
94 acids were eluted in 60uL of AVE buffer and the 5'-leader region was amplified using the following
95 amplification strategy. First, one-step RT-PCR (SuperScript III) was performed using HIV-1 specific
96 primers with 5' tags to enable sample indexing during a second PCR. The PCR primers were

97 complementary to HXB2 positions 469-490 (5'-GACCAGATCTGAGCCTGGGAGC) and 863-844 (5'-
98 CCCCCTGGCCTTAACCGAAT). The resulting amplicon encompassed HXB2 positions 491-843 which
99 included the 5'-leader positions 37-356. A second PCR was then performed to multiplex samples with
100 dual indexes using NEBNext Multiplex Oligos for Illumina (New England Biolabs). Library quality and
101 concentration were measured using the Agilent DNA 1000 kit (Agilent Technologies). The PCR products
102 showing non-specific peaks were purified again using E-gel SizeSelect II Agarose Gel, 2% (Invitrogen)
103 according to manufacturer's instructions. Amplicons were pooled at equimolar concentration and
104 purified using AMPure XP beads (Beckman Coulter). The library was spiked with 12.5% PhiX and loaded
105 onto an Illumina MiSeq v2 Cartridge at 8 pM and was sequenced using 2 x 250 bp runs.
106

107 **Sequence analysis pipeline**

108 The Fastp program was applied to each FASTQ file to trim adapters, remove regions with low
109 phred scores, and stich paired reads (6). Trimmed sequence reads were aligned to the HXB2 5'-leader
110 sequence using Minimap2 and the alignments were saved in Sequence Alignment Map (SAM) text files
111 (7). Samtools were used to convert the resulting SAM text files into binary BAM files and BAI index files
112 (8). PySam was used to read each BAM file to construct nucleotide frequency files containing the
113 proportions of each nucleotide and indel at each 5'-leader position. Paired samples for which both
114 baseline and follow-up contained a read depth of ≥200 at all 5'-leader positions between 37 and 356
115 were retained for analysis.

116 We then created consensus sequences using IUPAC codes, whenever one or more nucleotide
117 was present in ≥20% of sequence reads at the same position. These sequences were submitted to
118 GenBank (accession numbers: OQ814268-OQ814427) and used to construct a neighbor-joining
119 phylogenetic tree. The phylogenetic tree employed a weighted distance matrix, which utilized the
120 Jaccard distance to account for the overlap among nucleotides when more than one nucleotide was

121 present at the same position. The 160 unedited fastq files were submitted to the NCBI Sequence Read
122 Archive (BioProject PRJNA954829).

123 For each baseline and follow-up sequence, we reported all nucleotides, insertions, and deletions
124 present in ≥20% of sequence reads. Mutational changes between baseline and follow-up were defined
125 as nucleotides or indels for which the proportion of reads containing them changed by ≥4-fold and that
126 were present at follow-up in ≥20% of reads. This requirement therefore occasionally necessitated noting
127 which nucleotides and indels in the baseline sequence were present between 5% and 20% of reads.

128

129 **Analysis of previously published HIV-1 group M 5'-leader sequences**

130 One-per-person sequence set: We searched the June 2021 version of the Los Alamos National
131 Laboratory HIV Sequence Database (LANL) for non-problematic HIV-1 5'-leader group M sequences with
132 a minimal length of 250 nucleotides. Sequences were grouped into submission sets according to the
133 GenBank "Title" and "Author" fields. Submission sets that were not linked to a PubMed reference or
134 that contained sequences from fewer than five persons were excluded from analysis. When multiple
135 sequences were available for the same person, we randomly selected one sequence for analysis. The
136 search yielded 85 studies containing 1,417 one-per-person 5'-leader sequences.

137 Each sequence was first aligned to the HXB2 reference 5'-leader sequence using the Smith-
138 Waterman algorithm included in the European Molecular Biology Open Software Suite (EMBOSS)
139 package (9). Next a multiple sequence alignment of the 5'-leader regions was performed using MAFFT
140 with a set of 157 pre-aligned subtype reference sequences from LANL as a seed alignment using the
141 option for adding unaligned sequences (10). The alignment was then manually adjusted to increase the
142 consistency of indel placement. For each 5'-leader position we determined the frequency of each
143 nucleotide and each indel.

144 Clonal sequences set: Among the 85 publications reporting HIV-1 group M 5'-leader sequences
145 from ≥5 persons, we identified five publications that reported an average of at least three clones per
146 person with active virus replications (11–15). Four publications reporting at least three clones per
147 person were excluded because the clones were from proviral DNA samples from persons with virological
148 suppression and they contained large numbers of deletions.

149 NGS set: We reviewed the NCBI Sequence Read Archive (SRA) to identify published studies in
150 which NGS encompassed the 5'-leader in at least five persons. We analyzed the sequences from each of
151 these studies using the NGS pipeline described above ("Sequence analysis section"). Overall, there were
152 246 SRA BioProjects including 63 containing 5'-leader sequences as determined by our sequence
153 analysis pipeline. Six publications contained sequences from 295 persons (between 8 and 105 persons
154 per publication) (16–21).

155

156 **RESULTS**

157 **Individuals and samples**

158 Samples before and after ART were available for 29 individuals who developed M184V and for
159 19 individuals who developed an NNRTI-resistance mutations including K103N (n=13), Y181C (n=5), and
160 G190A (n=1). Samples from two time points were available from 32 control individuals who were ART-
161 naïve and who did not develop any RTI-resistance mutations. Table 1 summarizes the demographics,
162 ART histories, and baseline laboratory values for each group of individuals. Among the 29 individuals in
163 the M184V group, 19 were NRTI-naïve at the time of the first sample. Nine had previously received 3TC
164 or FTC but never developed M184V; one had received NRTIs other than 3TC or FTC. Among the 19
165 individuals in the NNRTI-resistance group, 16 were NNRTI-naïve and three had previously received an
166 NNRTI but never developed NNRTI resistance.

167 **Table 1. Description of the 80 Individuals Undergoing Baseline and Follow-Up HIV-1 5'-Leader Sequencing**
168

		Control Group (n=33)	M184V Group (n=29)	NNRTI-Resistance Group (n=19)	P value[1]		
					Control vs. M184V	Control vs. NNRTI	M184V vs. NNRTI
Sex		30 male / 2 female	24 male / 5 female	16 male / 3 female	NS	NS	NS
Age (IQR)		39 (26.8-51)	42 (33-47)	41 (33-48)	NS	NS	NS
Year (IQR)		2011 (2006-2014)	2001 (2001-2004)	2004 (2001-2010)	<0.001	0.003	NS
Baseline ART history[2]		Naïve: 32	NRTI-naïve: 19 NRTI-experienced: 10	NNRTI-naïve: 16 NNRTI-experienced: 3	-	-	-
CD4 (IQR)	*Baseline*	402.5 (296-503.5)	83 (35-252)	160 (78-259)	<0.001	<0.001	NS
	Follow-up	315 (221-472.8)	185 (81-312)	192 (98-255)	0.007	0.01	NS
VL (IQR)	*Baseline*	4.6 (4.1-4.9)	5.2 (4.6-5.6)	4.6 (4.4-5.6)	0.002	NS	NS
	Follow-up	4.8 (4.3-5)	4.4 (4.1-4.8)	4.8 (4.5-5)	NS	NS	NS
Months (IQR)[3]		12.5 (2-21.5)	20 (7-36)	39 (16.5-54.5)	0.04	0.01	0.04
***Pol* genetic distance (IQR)[4]**		0.75% (0.44%-1.23%)	2.17% (1.6%-2.84%)	1.92% (1.47%-2.98%)	<0.001	<0.001	NS

169
170 Footnotes: [1]Wilcoxon Rank Sum Tests were used to compare all median values. Fisher's exact tests were used to compare proportions (e.g.,
171 proportion that were male). [2]In the M184V Group, 9 of the 10 individuals who were NRTI-experienced individuals at baseline had received a 3TC-
172 or FTC-containing regimen but none had developed M184V (or M184I) prior to initial sampling. In the NNRTI-resistance group, 6 of the 16
173 individuals who were NNRTI-naïve at baseline were completely ART-naïve; none of the 3 NNRTI-experienced individuals had previous NNRTI-
174 resistance mutations. [3]Months between the baseline and follow-up samples. [4]The percentage of nucleotides that differed between baseline and
175 follow-up in the HIV-1 *pol* sequence performed for clinical purposes at the time the samples were initially obtained. The *pol* sequence usually
176 encompassed the complete protease and the first 300 RT codons.

177 At the time the initial plasma samples were obtained, the individuals in the control group had a
178 higher median CD4 count (402.5 cells/mm^3) compared to those in the M184V (83 cells/mm^3) and NNRTI-
179 resistance (160 cells/mm^3) groups. At the time of follow-up sampling, the median CD4 count in the
180 control group decreased while the median CD4 count in the M184V and NNRTI groups increased (Table
181 1). The median time between the two samples differed between the three groups: 12.5 months for the
182 control group, 20 months for the M184V group, and 39 months for the NNRTI-resistance group. The
183 median uncorrected genetic distance in the protease and RT genes was 0.75% for the individuals in the
184 control group compared with 2.17% and 1.92% in the M184V and NNRTI-resistance groups, respectively.
185

186 **Baseline sequences**

187 The median number of reads per sample was 2883 (IQR: 1264 – 4981). Within each sample, the
188 coverage was highly uniform across positions 37-356. Across all samples, 98.4% of paired reads
189 encompassed ≥300 nucleotides. All of the sequences belonged to subtype B based on phylogenetic
190 analysis of the *pol* gene. Among the 80 baseline sequences, 87 (27.2%) positions had a nucleotide
191 difference from HXB2 present in ≥20% of sequence reads (Figure 1). At positions 47, 200, 201, 213, 227,
192 265, and 305 more than one-half of the nucleotides differed from HXB2 but were the same as the
193 subtype B consensus sequence.
194

195 Figure 1

196 HIV-1 5'-leader nucleotide differences from HXB2 observed in the 80 baseline sequences superimposed
197 on a diagram of its RNA secondary structure. Nucleotides observed in ≥20% of reads in more than 10
198 sequences are shown in orange; those present in fewer than 10 sequences are shown in purple.
199 Insertions and deletions are not shown. Positions 1 to 36 were not sequenced. TAR: trans-activation
200 response element (positions 1-57); PolyA: polyadenylation signal (positions 58-104); H1: helix 1 tertiary

201 structure; H2: helix 2 tertiary structure; PBS: primer binding site (positions 182-199); DIS: dimer
202 initiation site (positions 242-278); psi: packaging signal (positions 312-315). The Matrix initiation codon
203 is surrounded by a green box; it is base-paired to the U5 triplet. The primer activation signal nucleotides
204 are colored green, the C-rich region nucleotides are colored yellow, and the PBS nucleotides are colored
205 maroon.
206

207 The poly A motif (AATAAA; positions 73-78), primer activating signal (PAS; positions 125-130)
208 and primer binding site (PBS; positions 182-199) were completely conserved in the 80 baseline
209 sequences. The U5 region was represented by CGT (91% of sequences) or CAT (9% of sequences) each of
210 which maintained complementarity to the Matrix initiation codon (positions 336-338). The dimerization
211 signal (DIS; positions 257-262) was represented by GCGCGC in all but five sequences. The RNA packaging
212 signal (PSI; positions 312-325) contained the same nucleotides in all but three sequences.

213 Figure 2 shows that the distribution of nucleotides and indels was highly similar between the 80
214 baseline samples and the 1,417 samples downloaded from the LANL database (Supplementary Figure 1).
215 There were no apparent differences in the proportions of nucleotides or indels between the sequences
216 from the 56 individuals who were ART-naïve and the 24 individuals who were ART-experienced at
217 baseline (Supplementary Figure 2). Figure 3 shows that 52 positions had more than one nucleotide
218 detected in ≥20% of sequence reads. At positions 200, 201, 304, 224, and 354, more than 10% of
219 samples had a mixture of two nucleotides.

220

221 Figure 2

222 Distribution of HIV-1 5'-leader nucleotide differences from HXB2 and indels in the 80 baseline sequences
223 and in 1,417 one-per-person Los Alamos National Laboratory HIV Sequence Database (LANL) dataset.

224 Figure 3

225 HIV-1 5'-leader positions ranked according to the frequency with which they contained a mixture of two
226 or more nucleotides. Each nucleotide in a mixture was required to present in ≥20% of NGS reads.

227 Supplementary Figure 1

228 HIV-1 5-leader nucleotide differences from the HXB2 sequence observed in 1,417 sequences in
229 the Los Alamos National Laboratories HIV Sequence Database dataset. Superscripts indicate the
230 percentage of times that a nucleotide was reported. Highlighted regions include (1) TAR: trans-

231 activation response element; (2) PolyA: polyadenylation signal loop with a box surrounding the
232 poly A motif; (3) U5; (4) PAS: primer activation signal; (5) PBS: primer binding site; (6) DIS: dimer
233 initiation signal loop with a box surrounding the palindromic DIS; (7) major 5' splicing site; (8)
234 PSI packaging signal; (9) Matrix protein with a box surrounding the start codon.

235 Supplementary Figure 2

236 Distribution of HIV-1 5'-leader nucleotide differences from HXB2 and indels in the 56 baseline sequences
237 from ART-naïve individuals compared with the 24 baseline sequences from ART-experienced individuals.

238

239 Compared to HXB2, two individuals had large deletions. In one individual, positions 119 to 161
240 were deleted and in the other individual, positions 281 to 288 were deleted. Both of these deletions
241 were also present in the follow-up sequence from the same individual suggesting that these deletions
242 were not sequencing artifacts. An additional 42 baseline samples had deletions of 1 to 3 nucleotides.
243 The most frequent positions with deletions were between positions 301 to 303 (20 samples) and
244 between positions 150 to 153 (13 samples). Compared to HXB2, 54 samples had one or more insertions,
245 with the most frequent occurring between positions 300 to 305 (26 samples), 254 to 255 (6 samples),
246 165 to 167 (5 samples), and 198 to 202 (4 samples). Eight samples had insertions containing more than
247 three nucleotides.
248
249 **Follow-up sequences**
250 Figure 4 summarizes the number of individuals by the number of positions at which a ≥4-fold
251 change in the proportion of a nucleotide or indel occurred between baseline and follow-up for each of
252 the three sample groups. The median number of changes per individual was higher in the M184V (3;
253 IQR:1-4; p=0.014 Wilcoxon Rank Sum Test) and NNRTI groups (2; IQR: 1-4; p=0.12; Wilcoxon Rank Sum
254 Test) than in the control group (0.5; IQR:0-4). The three control group individuals with large changes
255 included one individual who may have experienced a re-infection as evidenced by a nucleotide distance
256 of 5.4% between their baseline and follow-up HIV-1 *pol* sequence.
257
258 Figure 4
259 Number of HIV-1 5'-leader positions at which a nucleotide or indel exhibited a ≥4-fold change in its
260 proportion between baseline and follow-up and which was present at a level of ≥20% at follow-up in
261 each of the three patient groups.
262

263 Fifty-three of the 63 deletions found in either baseline or follow-up samples were present at
264 both time points. Fifty-three of the 70 insertions present in either baseline or follow-up samples were
265 present at both time points. A neighbor-joining phylogenetic tree using the Jaccard distance to weight
266 the distance between overlapping IUPAC nucleotides is shown in Figure 5.
267
268 Figure 5
269 Neighbor-joining tree containing the 80 baseline and follow-up sequences from this study.
270 Sample IDs for the M184V group are colored blue; those for the NNRTI group are colored
271 yellow; and those for the Control group are colored red. Bootstrap values are indicated at each
272 node of the tree.
273

274 Figure 6A shows the positions at which at least one nucleotide changed its proportion by ≥4-fold
275 and which occurred in ≥20% of sequence reads. The positions at which nucleotide changes occurred
276 most frequently are shown in Figure 6B. Positions 200 and 201, which were the positions most likely to
277 have mixtures at baseline, were also the positions most likely to exhibit a ≥4-fold change in the
278 proportion of a nucleotide. Position 201 was significantly more likely to exhibit a ≥4-fold change in its
279 proportions in the M184V (9/29 vs. 0/32; p=0.0006; Fisher's Exact Test) and the NNRTI-resistance groups
280 (4/19 vs. 0/32; p=0.02; Fisher's Exact Test) compared with the control group. There were no significant
281 differences between each of the three groups at any other position including position 200. At position
282 201, all four nucleotides were present at baseline and there was no predilection for any particular
283 change between baseline and follow-up.
284
285 Figure 6
286 Distribution of positions at which a nucleotide or indel changed its prevalence by ≥4-fold
287 according to patient group (A) and a table indicating the ten positions that changed most often
288 (B).
289

290 **Distribution of mixtures in previously published clonal sequences and in NGS**

291 The five publications describing a minimum of three clones per individual from at least five
292 individuals reported a total of 294 sequences from 42 individuals. At five positions, ≥10% of individuals
293 had a mixture of ≥2 nucleotides in different clones including positions 200 (n=16 individuals; 38.1%), 201
294 (n=11 individuals; 26.2%), 305 (n=7 individuals; 16.7%), 152 (n=5 individuals; 11.9%), and 281 (n=5
295 individuals; 11.9%).

296 The six publications in the NCBI SRA describing 5'-leader sequences from 295 individuals
297 reported that the most common positions containing a mixture at least two nucleotides (i.e., each
298 present in ≥20% of NGS reads) were also positions 200 (n=103 individuals; 34.9%) and 201 (n=62
299 individuals; 21.1%). Table 2 illustrates the 15 positions that most frequently contained mixtures in this
300 study and the frequency with which these positions contained mixtures in the GenBank and NCBI SRA
301 datasets. In contrast to the 5'-leader, among 384 RT sequences from 379 individuals in the same six
302 NCBI SRA dataset publications, no position had a mixture in ≥6% of positions.

303 **Table 2. Proportion of HIV-1 5'-Leader Positions Containing a Mixture of Nucleotides in Sampled Viruses in this Study, in Previously Published**
304 **Studies in GenBank, and in NGS Studies in the NCBI Sequence Read Archive (SRA)**
305

Position	Baseline Samples from this Study (n=80 individuals)[1]	Published Studies of Individuals with Multiple Clones (GenBank) (n=42 individuals)[2]	Published Studies of Individuals Undergoing NGS (NCBI SRA) (n=295 individuals)[3]
200	45.0%	38.1%	34.9%
201	28.8%	26.2%	21.1%
304	15.0%	4.8%	4.1%
224	10.0%	9.5%	6.1%
354	10.0%	2.4%	1.4%
305	8.8%	16.7%	6.0%
152	7.5%	11.9%	6.8%
286	6.3%	9.5%	5.8%
96	6.3%	6.9%	8.2%
263	6.3%	2.4%	2.7%
281	5.0%	11.9%	1.5%
47	5.0%	4.8%	8.2%
215	5.0%	7.1%	2.5%
227	5.0%	4.8%	2.7%
353	5.0%	2.4%	1.0%

306
307 Footnote: [1]A position was considered to have a mixture if at least one non-consensus nucleotide was present in ≥20% of reads. [2]Data were
308 obtained from five publications reporting ≥3 clones per individual from ≥5 individuals. A total of 294 sequences were performed (i.e., a mean of
309 7 clones per individual). A position was considered to have a mixture if at least one clone had a non-consensus nucleotide. Thirty-one individuals
310 (74%) had subtype B viruses and 11 patients had subtype C viruses. [3]A position was considered to have a mixture if at least one non-consensus
311 nucleotide was present in ≥20% of reads. Sixty-four percent of sequences belonged to subtype C; 21% to subtype B; and 15% to other subtypes
312 or circulating recombinant forms.

313

314

315 **DISCUSSION**

316 To our knowledge, this is the first study of paired HIV-1 5'-leader sequences from individuals
317 who did or did not develop RT-associated DRMs. Contrary to our initial hypothesis, we did not observe
318 co-evolutionary changes between RT and 5'-leader sequences. The only 5'-leader position (position 201)
319 that demonstrated a higher likelihood of evolution in samples from individuals who developed RT-
320 associated DRMs is one of the most variable 5'-leader positions. However, our study uncovered a
321 previously unreported phenomenon, wherein specific 5'-leader positions, particularly positions 200 and
322 201, exhibited an extraordinarily high likelihood of containing two or more nucleotides each in more
323 than 20% of NGS reads. We corroborated this finding by analyzing two previously published datasets:
324 five published studies of clonal dideoxy-terminator Sanger sequences and six NCBI SRA NGS datasets
325 encompassing the HIV-1 5'-leader.

326 Although HIV-1 is a quasispecies characterized by the presence of multiple circulating variants,
327 the markedly high prevalence of nucleotide mixtures at two particular positions is striking. The
328 prevalence of mixtures in our 80 baseline sequences was 45% for position 200 and 29% for position 201.
329 In our analysis of previously published clonal sequences, these prevalences were 38% and 26%,
330 respectively, and in our analysis of previously published NGS data, these prevalences were 35% and
331 21%, respectively. Moreover, in these analyses, positions 200 and 201 contained mixtures at a
332 frequency several times higher than that observed at any other 5'-leader position or, in the NCBI SRA
333 NGS dataset, at any RT position. The fact that the previously published clonal and NGS datasets
334 contained large numbers of non-subtype B isolates suggests that the propensity of positions 200 and
335 201 to contain nucleotide mixtures is not limited to one subtype.

336 Positions 200 and 201 are immediately downstream of the PBS, situated opposite to the
337 consecutive A nucleotides at tRNA positions 58 and 57, respectively (3). The significance of
338 complementarity between the 5'-leader and tRNA at these two positions, however, has not been
339 studied. We believe that the most plausible explanation for the high proportion of mixtures at these two
340 positions may be that they are particularly error prone as they represent the first two nucleotides added
341 following the second strand transfer step during the process of HIV-1 reverse transcription (5).

342 Our study is limited because we performed NGS without using unique molecular identifiers
343 (UMIs) (22). The use of UMIs would have allowed us to precisely quantify each variant within the HIV-1
344 quasispecies and to thus reliably identify linkages between positions in the same variant (22). Our study
345 is also limited because we studied convenience samples and did not have information on the duration of
346 infection of the individuals in our study. If a high proportion of mixtures at positions 200 and 201 can be
347 found in recently infected individuals this would support the hypothesis that these two positions are
348 highly error-prone. In contrast, if the mixtures at these positions develops during the course of HIV-1
349 infection, the possibility that mixtures at these positions contribute to viral replication fitness may be
350 more likely.

351

352 Acknowledgements

353 JN and RWS have been funded in part by an NIH grant R03 AI147632. EVP, JDP and MK have been
354 funded in part by an NIH grant U54 AI170856.

355

356 Conflicts of interest

357 The author(s) declare that there are no conflicts of interest.

358
359 **REFERENCES**

360 1. Berkhout B. Structure and Function of the Human Immunodeficiency Virus Leader RNA. In: Cohn
361 WE, Moldave K, editors. Progress in Nucleic Acid Research and Molecular Biology [Internet].
362 Academic Press; 1996 [cited 2021 Oct 5]. p. 1–34. Available from:
363 https://www.sciencedirect.com/science/article/pii/S0079660308603591

364 2. Keane SC, Summers MF. NMR Studies of the Structure and Function of the HIV-1 5'-Leader. Viruses.
365 2016 21;8(12).

366 3. Larsen KP, Mathiharan YK, Kappel K, Coey AT, Chen DH, Barrero D, et al. Architecture of an HIV-1
367 reverse transcriptase initiation complex. Nature. 2018;557(7703):118–22.

368 4. Ha B, Larsen KP, Zhang J, Fu Z, Montabana E, Jackson LN, et al. High-resolution view of HIV-1 reverse
369 transcriptase initiation complexes and inhibition by NNRTI drugs. Nat Commun. 2021 May
370 4;12(1):2500.

371 5. Krupkin M, Jackson LN, Ha B, Puglisi EV. Advances in understanding the initiation of HIV-1 reverse
372 transcription. Curr Opin Struct Biol. 2020 Dec 1;65:175–83.

373 6. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics.
374 2018 Sep 1;34(17):i884–90.

375 7. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep
376 15;34(18):3094–100.

377 8. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
378 format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

379 9. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends
380 Genet. 2000 Jun 1;16(6):276–7.

381 10. Katoh K, Frith MC. Adding unaligned sequences into an existing alignment using MAFFT and LAST.
382 Bioinformatics. 2012 Dec 1;28(23):3144–6.

383 11. Novitsky VA, Montano MA, McLane MF, Renjifo B, Vannberg F, Foley BT, et al. Molecular Cloning
384 and Phylogenetic Analysis of Human Immunodeficiency Virus Type 1 Subtype C: a Set of 23 Full-
385 Length Clones from Botswana. J Virol. 1999 May;73(5):4427–32.

386 12. Rolland M, Tovanabutra S, deCamp AC, Frahm N, Gilbert PB, Sanders-Buell E, et al. Genetic impact
387 of vaccination on breakthrough HIV-1 sequences from the STEP trial. Nat Med. 2011 Mar;17(3):366–
388 71.

389 13. Ochsenbauer C, Edmonds TG, Ding H, Keele BF, Decker J, Salazar MG, et al. Generation of
390 Transmitted/Founder HIV-1 Infectious Molecular Clones and Characterization of Their Replication
391 Capacity in CD4 T Lymphocytes and Monocyte-Derived Macrophages. J Virol. 2012 Mar
392 1;86(5):2715–28.

393 14. Parrish NF, Gao F, Li H, Giorgi EE, Barbian HJ, Parrish EH, et al. Phenotypic properties of transmitted
394 founder HIV-1. Proc Natl Acad Sci U S A. 2013 Apr 23;110(17):6626–33.

395 15. Gondim MVP, Sherrill-Mix S, Bibollet-Ruche F, Russell RM, Trimboli S, Smith AG, et al. Heightened
396 resistance to host type 1 interferons characterizes HIV-1 at transmission and after antiretroviral
397 therapy interruption. Sci Transl Med. 2021 Jan 13;13(576):eabd8179.

398 16. Gall A, Ferns B, Morris C, Watson S, Cotten M, Robinson M, et al. Universal Amplification, Next-
399 Generation Sequencing, and Assembly of HIV-1 Genomes. J Clin Microbiol. 2012 Dec;50(12):3838–
400 44.

401 17. Illingworth CJR, Roy S, Beale MA, Tutill H, Williams R, Breuer J. On the effective depth of viral
402 sequence data. Virus Evol. 2017 Nov 14;3(2):vex030.

403 18. Fedonin GG, Fantin YS, Favorov AV, Shipulin GA, Neverov AD. VirGenA: a reference-based assembler
404 for variable viral genomes. Brief Bioinform. 2019 Jan 18;20(1):15–25.

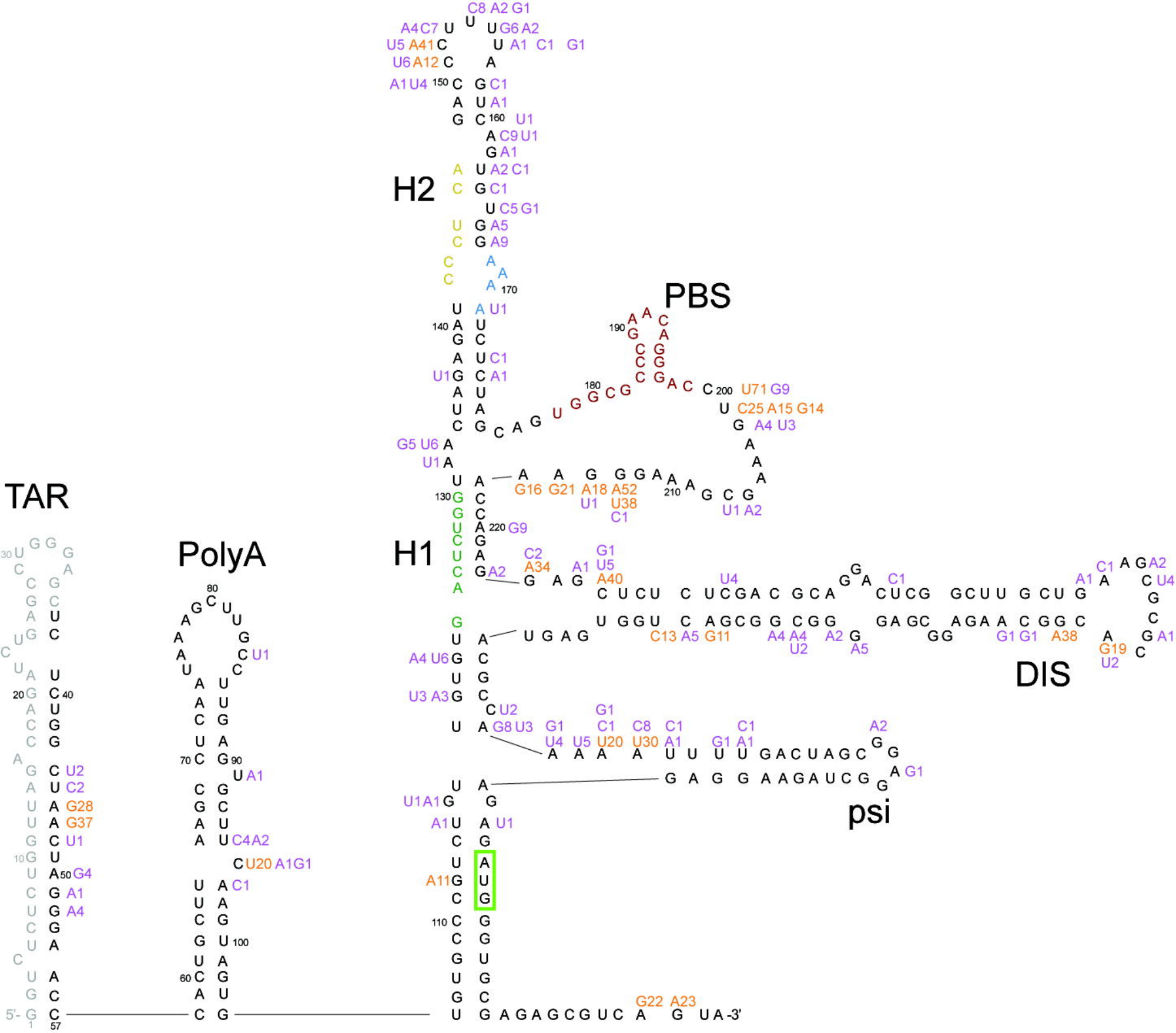
405 19. Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, et al. Genome Detective: an
406 automated system for virus identification from high-throughput sequencing data. Bioinformatics.
407 2019 Mar 1;35(5):871–3.

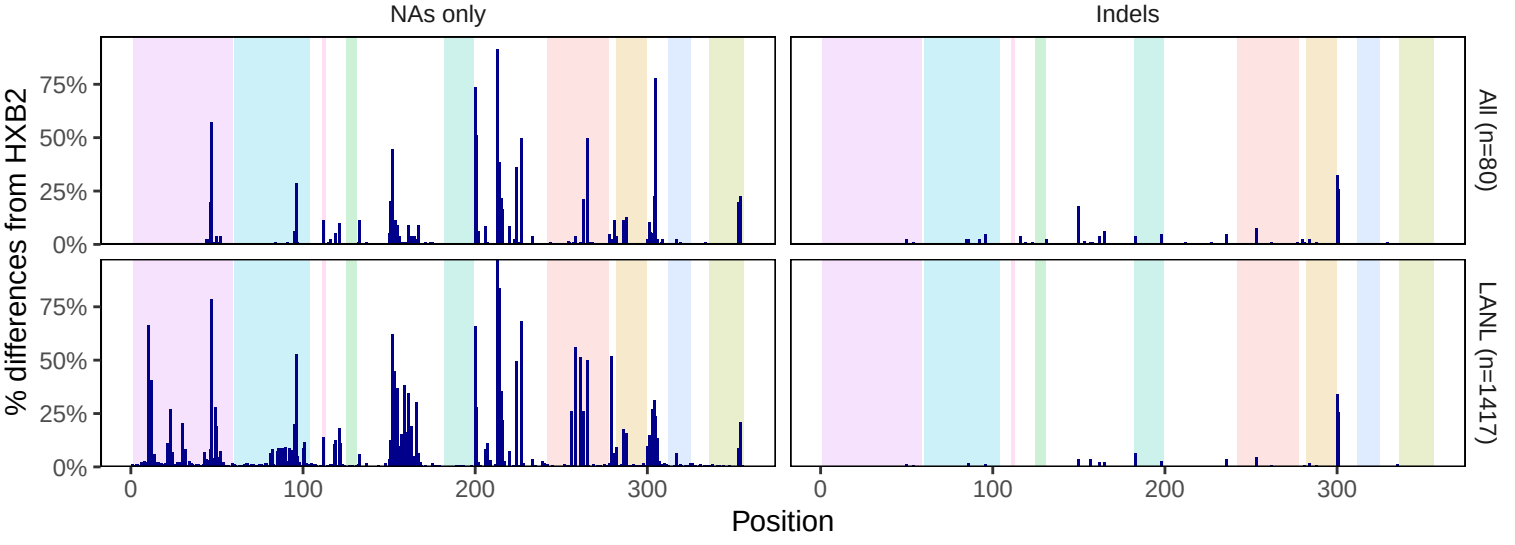
408 20. Zhang Y, Wymant C, Laeyendecker O, Grabowski MK, Hall M, Hudelson S, et al. Evaluation of
409 Phylogenetic Methods for Inferring the Direction of Human Immunodeficiency Virus (HIV)
410 Transmission: HIV Prevention Trials Network (HPTN) 052. Clin Infect Dis Off Publ Infect Dis Soc Am.
411 2020 Feb 7;72(1):30–7.

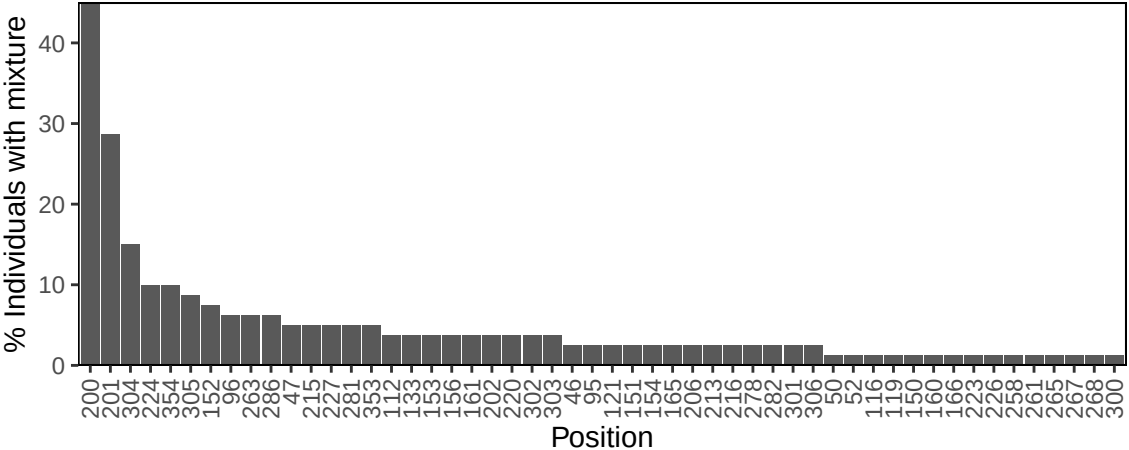
412 21. Ye L, Gribling-Burrer AS, Bohn P, Kibe A, Börtlein C, Ambi UB, et al. Short- and long-range
413 interactions in the HIV-1 5′ UTR regulate genome dimerization and packaging. Nat Struct Mol Biol.
414 2022 Apr;29(4):306–19.

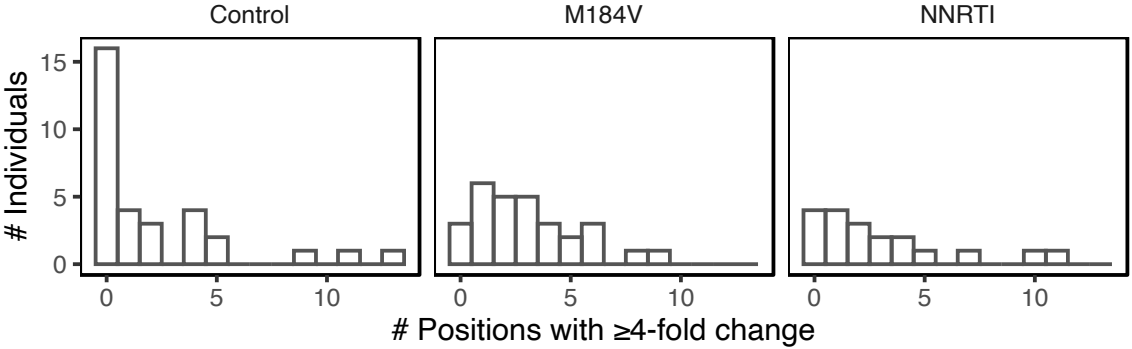
415 22. Zhou S, Hill CS, Spielvogel E, Clark MU, Hudgens MG, Swanstrom R. Unique Molecular Identifiers and
416 Multiplexing Amplicons Maximize the Utility of Deep Sequencing To Critically Assess Population
417 Diversity in RNA Viruses. ACS Infect Dis. 2022 Dec 9;8(12):2505–14.

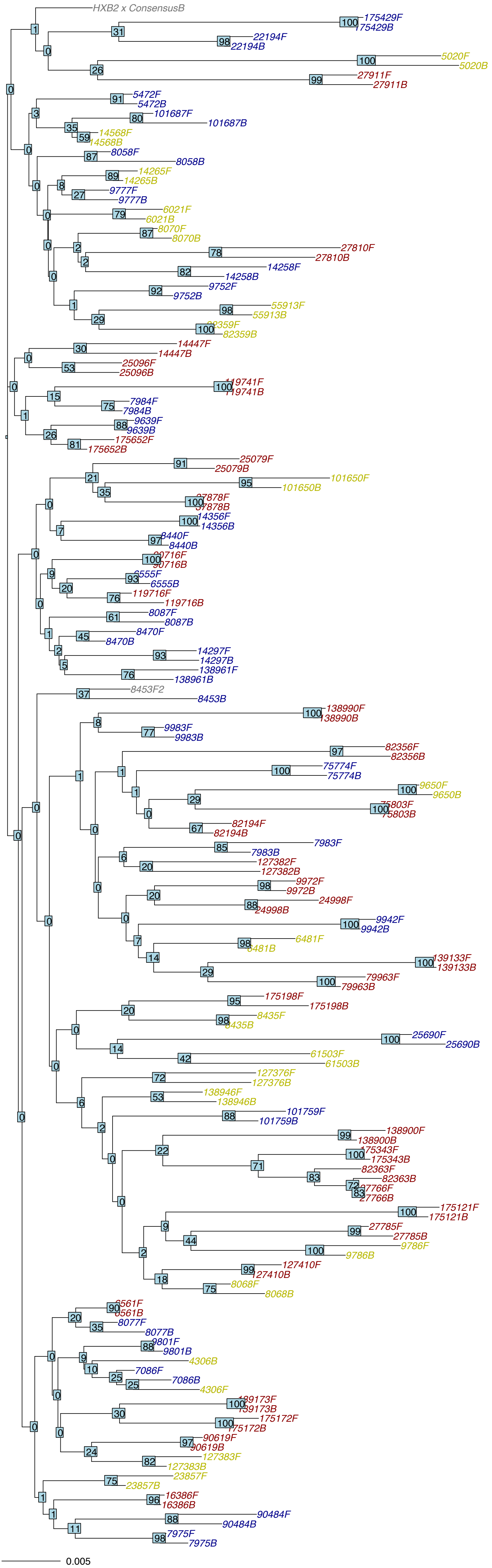
418


H2
PBS
TAR
PolyA
H1
DIS
psi
5'-
-3'

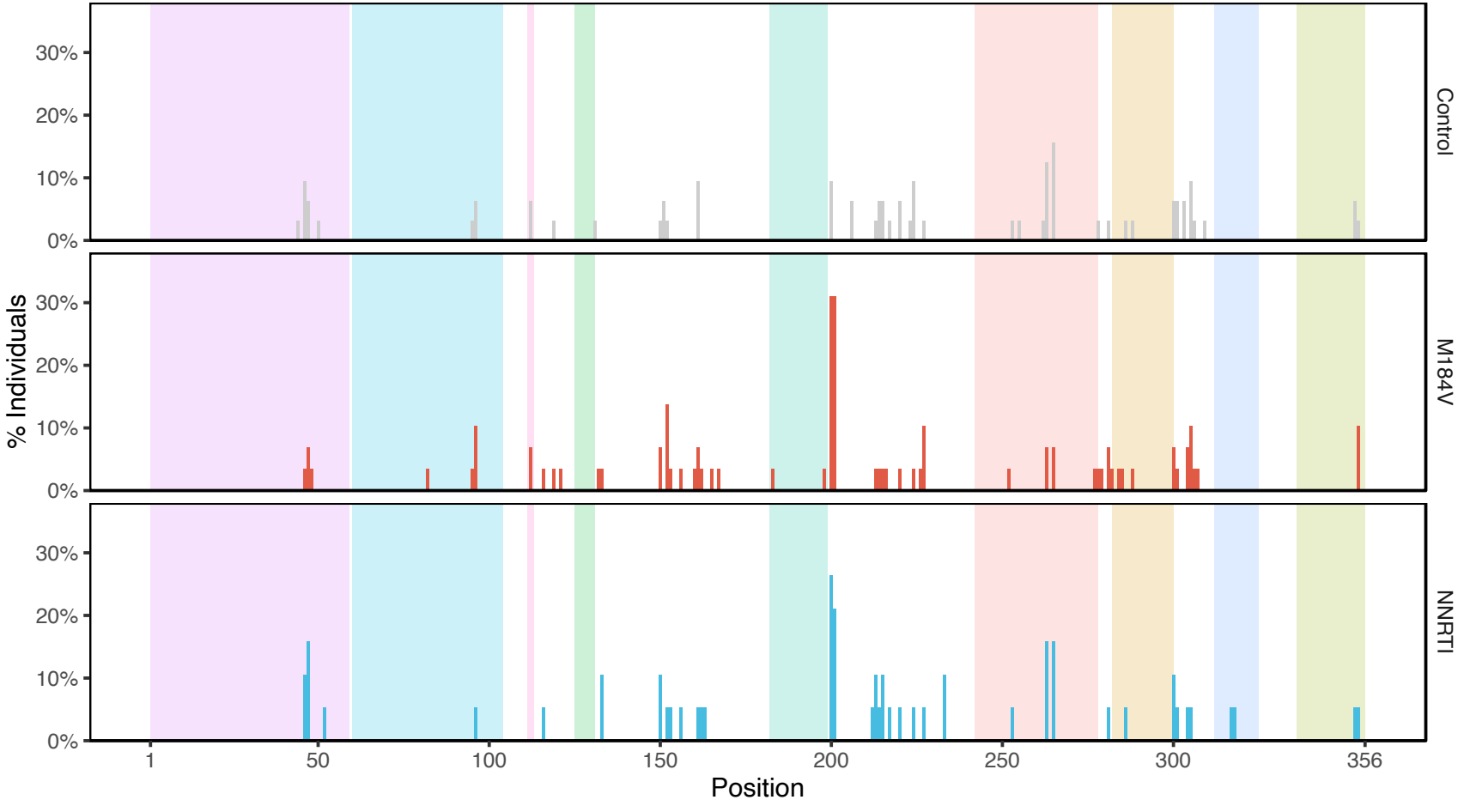

NAs only
Indels
All (n=80)
LANL (n=1417)
% differences from HXB2
75%
50%
25%
0%
0
100
200
300
Position


% Individuals with mixture
0
10
20
30
40
200
201
204
224
354
305
152
96
263
286
47
215
227
281
353
112
133
153
156
161
202
220
302
303
46
95
121
151
154
165
206
213
216
278
282
301
306
50
52
116
119
150
160
166
223
226
258
261
265
267
268
300
Position


Control
M184V
NNRTI
Individuals
15
10
5
0
0
5
10
Positions with ≥4-fold change


HXB2 x ConsensusB
175429F
175429B
22194F
22194B
5020F
5020B
27911F
27911B
5472F
5472B
101687F
101687B
14568F
14568B
8058F
8058B
14265F
14265B
9777F
9777B
6021F
6021B
8070F
8070B
27810F
27810B
14258F
14258B
9752F
9752B
55913F
55913B
82359F
82359B
14447F
14447B
25096F
25096B
119741F
119741B
7984F
7984B
9639F
9639B
175652F
175652B
25079F
25079B
101650F
101650B
97878F
97878B
14356F
14356B
8440F
8440B
90716F
90716B
6555F
6555B
119716F
119716B
8087F
8087B
8470F
8470B
14297F
14297B
138961F
138961B
8453F2
8453B
138990F
138990B
9983F
9983B
82356F
82356B
75774F
75774B
9650F
9650B
75803F
75803B
82194F
82194B
7983F
7983B
127382F
127382B
9972F
9972B
24998F
24998B
9942F
9942B
6481F
6481B
139133F
139133B
79963F
79963B
175198F
175198B
8435F
8435B
25690F
25690B
61503F
61503B
127376F
127376B
138946F
138946B
101759F
101759B
138900F
138900B
175343F
175343B
82363F
82363B
27766F
27766B
175121F
175121B
27785F
27785B
9786F
9786B
127410F
127410B
8068F
8068B
6561F
6561B
8077F
8077B
9801F
9801B
4306B
7086F
7086B
4306F
139173F
139173B
175172F
175172B
90619F
90619B
127383F
127383B
23857F
23857B
16386F
16386B
90484F
90484B
7975F
7975B
0.005

A



B

Top 10 positions

Position	Control (n=32)	M184V (n=29)	NNRTI (n=19)
200	9% (3)	31% (9)	26% (5)
201	0% (0)	31% (9)	21% (4)
152	3% (1)	14% (4)	5% (1)
305	9% (3)	10% (3)	5% (1)
96	6% (2)	10% (3)	5% (1)
227	3% (1)	10% (3)	5% (1)
354	3% (1)	10% (3)	5% (1)
265	16% (5)	7% (2)	16% (3)
263	12% (4)	7% (2)	16% (3)
47	6% (2)	7% (2)	16% (3)