

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25

**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<sup>1\*</sup>, Philip L. Tzou<sup>2\*</sup>, Soo-Yon Rhee<sup>2\*</sup>, Malaya K. Sahoo<sup>3</sup>, Benjamin A. Pinsky<sup>4</sup>, Miri Krupkin<sup>5</sup>,  
Joseph D. Puglisi<sup>5</sup>, Elisabetta V. Puglisi<sup>5</sup>, Robert W. Shafer<sup>2</sup>

\*Each of these authors contributed equally to the work

**AFFILIATIONS**

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

Corresponding authors: PLT ([philiptz@stanford.edu](mailto:philiptz@stanford.edu)), RWS ([rshafer@stanford.edu](mailto:rshafer@stanford.edu))

Abstract word count: 300

Text word count: 2,897

26 **ABSTRACT**

27 **Background:** HIV-1 RT initiation depends on interaction between viral 5'-leader RNA, RT, and host  
28 tRNA<sub>3<sup>Lys</sup></sub>. We therefore sought to identify co-evolutionary changes between the 5'-leader and RT in  
29 viruses developing RT-inhibitor resistance mutations. **Methods:** We sequenced 5'-leader positions 37-  
30 356 of paired plasma virus samples from 29 individuals developing the NRTI-resistance mutation M184V,  
31 19 developing an NNRTI-resistance mutation, and 32 untreated controls. 5'-leader variants were defined  
32 as positions where  $\geq 20\%$  of NGS reads differed from the HXB2 sequence. Emergent mutations were  
33 defined as nucleotides undergoing  $\geq 4$ -fold change in proportion between baseline and follow-up.  
34 Mixtures were defined as positions containing  $\geq 2$  nucleotides each present in  $\geq 20\%$  of NGS reads.  
35 **Results:** Among 80 baseline sequences, 87 positions (27.2%) contained a variant; 52 contained a  
36 mixture. Position 201 was the only position more likely to develop a mutation in the M184V (9/29 vs.  
37 0/32;  $p=0.0006$ ) or NNRTI-resistance (4/19 vs. 0/32;  $p=0.02$ ; Fisher's Exact Test) groups than the control  
38 group. Mixtures at positions 200 and 201 occurred in 45.0% and 28.8%, respectively, of baseline  
39 samples. Because of the high proportion of mixtures at these positions, we analyzed 5'-leader mixture  
40 frequencies in two additional datasets: five publications reporting 294 dideoxyterminator clonal  
41 GenBank sequences from 42 individuals and six NCBI BioProjects reporting NGS datasets from 295  
42 individuals. These analyses demonstrated position 200 and 201 mixtures at proportions similar to those  
43 in our samples and at frequencies several times higher than at all other 5'-leader positions. **Conclusions:**  
44 Although we did not convincingly document co-evolutionary changes between RT and 5'-leader  
45 sequences, we identified a novel phenomenon, wherein positions 200 and 201, immediately  
46 downstream of the HIV-1 primer binding site exhibited an extraordinarily high likelihood of containing a  
47 nucleotide mixture. Possible explanations for the high mixture rates are that these positions are  
48 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 structure and functions have been studied using phylogenetic methods  
53 [2], biochemical experiments with site-directed mutants [3–5], RNA structural probing methods [6,7],  
54 NMR [8], and cryo-EM [9,10]. These and many other studies have shown that the 5'-leader has multiple  
55 functional units including TAR, which is required for initiating transcription; several RNA elements  
56 required for RT initiation including the nearly invariant primer binding site (PBS); and several  
57 downstream elements responsible for viral splicing, dimerization, and packaging [1,4,8,9,11,12].

58 The 5'-leader region adapts at least two conformations including a monomeric form that  
59 interacts with the cellular translation machinery to favor the synthesis of HIV-1 proteins and a dimeric  
60 form that results in genomic packaging and virus assembly (reviewed in [13,14]). The 5'-leader also  
61 initiates reverse transcription through a three-way interaction among the viral RNA genome, the reverse  
62 transcriptase (RT) enzyme, and tRNA<sub>3<sub>Lys</sub></sub> (reviewed in [4,15]). During RT initiation, RT binds to the viral  
63 RNA-tRNA<sub>3<sub>Lys</sub></sub> duplex forming the reverse transcription initiation complex [9].

64 RT initiation represents a bottleneck to HIV-1 replication as evidenced by an approximately 100-  
65 3000-fold reduced rate of nucleotide incorporation during initiation compared with elongation and by  
66 the much slower synthesis of the first 200 HIV-1 nucleotides compared to the remaining nucleotides of  
67 proviral DNA [15]. We hypothesized that during RT initiation there may be an interplay between changes  
68 in RT and the 5'-leader viral RNA. In particular, we questioned whether the 5'-leader would evolve to  
69 adapt to the development of drug-resistance mutations (DRMs) in the RT enzyme. Therefore, we  
70 sequenced the 5'-leader region of plasma virus samples obtained from individuals before and after the  
71 development of the most common nucleoside RT inhibitor (NRTI) mutation M184V, which confers

72 resistance to the cytidine analogs lamivudine (3TC) and emtricitabine (FTC), or a non-nucleoside RT  
73 inhibitor (NNRTI)-resistance mutation.

74

## 75 **METHODS**

### 76 **Individuals and samples**

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

87 The remnant plasma samples used in this study were obtained between the years 2000 and  
88 2017. The Stanford University Institutional Review Board approved this study under protocol 6633,  
89 entitled “Human Immunodeficiency Virus Quasispecies During Antiviral Therapy”. The approval included  
90 a waiver of consent because the study used anonymized de-identified data and remnant plasma  
91 samples. The samples were sequenced and the data were collected for research purposes between July  
92 2019 and January 2021.

93

### 94 **Sequencing protocol**

95 Total nucleic acids were extracted from 200 uL of plasma using the EZ1 Virus Mini Kit V2.0 on  
96 the automated EZ1 advanced XL (both from Qiagen), according to manufacturer's instructions. Nucleic  
97 acids were eluted in 60uL of AVE buffer and the 5'-leader region was amplified using the following  
98 amplification strategy. First, one-step RT-PCR (SuperScript III) was performed using HIV-1 specific  
99 primers with 5' tags to enable sample indexing during a second PCR. The PCR primers were  
100 complementary to HXB2 positions 469-490 (5'-GACCAGATCTGAGCCTGGGAGC) and 863-844 (5'-  
101 CCCCTGGCCTTAACCGAAT). The resulting amplicon encompassed HXB2 positions 491-843 which  
102 included the 5'-leader positions 37-356. A second PCR was then performed to multiplex samples with  
103 dual indexes using NEBNext Multiplex Oligos for Illumina (New England Biolabs). Library quality and  
104 concentration were measured using the Agilent DNA 1000 kit (Agilent Technologies). The PCR products  
105 showing non-specific peaks were purified again using E-gel SizeSelect II Agarose Gel, 2% (Invitrogen)  
106 according to manufacturer's instructions. Amplicons were pooled at equimolar concentration and  
107 purified using AMPure XP beads (Beckman Coulter). The library was spiked with 12.5% PhiX and loaded  
108 onto an Illumina MiSeq v2 Cartridge at 8 pM and was sequenced using 2 x 250 bp runs. Both ends of  
109 each DNA fragment were sequenced (i.e., paired-end sequencing) to obtain bi-directional sequence  
110 information.

111

### 112 **Sequence analysis pipeline**

113 The Fastp program was applied to each FASTQ file to trim adapters, remove regions with low  
114 phred scores, and stich paired reads [16]. Trimmed sequence reads were aligned to the HXB2 5'-leader  
115 sequence using Minimap2 and the alignments were saved in Sequence Alignment Map (SAM) text files  
116 [17]. Samtools were used to convert the resulting SAM text files into binary BAM files and BAI index files  
117 [18]. PySam was used to read each BAM file to construct nucleotide frequency files containing the  
118 proportions of each nucleotide and indel at each 5'-leader position. Paired samples for which both

119 baseline and follow-up contained a read depth of  $\geq 200$  at all 5'-leader positions between 37 and 356  
120 were retained for analysis.

121 We then created consensus sequences using IUPAC codes, whenever one or more nucleotide  
122 was present in  $\geq 20\%$  of sequence reads at the same position. These sequences were submitted to  
123 GenBank (accession numbers: OQ814268-OQ814427) and used to construct a neighbor-joining  
124 phylogenetic tree. The phylogenetic tree employed a weighted distance matrix, which utilized the  
125 Jaccard distance to account for the overlap among nucleotides when more than one nucleotide was  
126 present at the same position. The 160 unedited fastq files were submitted to the NCBI Sequence Read  
127 Archive (BioProject PRJNA954829).

128 For each baseline and follow-up sequence, we reported all nucleotides, insertions, and deletions  
129 present in  $\geq 20\%$  of sequence reads. Mutational changes between baseline and follow-up were defined  
130 as nucleotides or indels for which the proportion of reads containing them changed by  $\geq 4$ -fold and that  
131 were present at follow-up in  $\geq 20\%$  of reads. This requirement therefore occasionally necessitated noting  
132 which nucleotides and indels in the baseline sequence were present between 5% and 20% of reads.

133

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

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

142 Each sequence was first aligned to the HXB2 reference 5'-leader sequence using the Smith-  
143 Waterman algorithm included in the European Molecular Biology Open Software Suite (EMBOSS)  
144 package [19]. Next a multiple sequence alignment of the 5'-leader regions was performed using MAFFT  
145 with a set of 157 pre-aligned subtype reference sequences from LANL as a seed alignment using the  
146 option for adding unaligned sequences [20]. The alignment was then manually adjusted to increase the  
147 consistency of indel placement. For each 5'-leader position we determined the frequency of each  
148 nucleotide and each indel.

149 Clonal sequences set: Among the 85 publications reporting HIV-1 group M 5'-leader sequences  
150 from  $\geq 5$  persons, we identified five publications that reported an average of at least three clones per  
151 person with active virus replications [21–25]. Four publications reporting at least three clones per  
152 person were excluded because the clones were from proviral DNA samples from persons with virological  
153 suppression and they contained large numbers of deletions.

154 NGS set: We reviewed the NCBI Sequence Read Archive (SRA) to identify published studies in  
155 which NGS encompassed the 5'-leader in at least five persons. We analyzed the sequences from each of  
156 these studies using the NGS pipeline described above (“Sequence analysis section”). Overall, there were  
157 246 SRA BioProjects including 63 containing 5'-leader sequences as determined by our sequence  
158 analysis pipeline. Six publications contained sequences from 295 persons (between 8 and 105 persons  
159 per publication) [7,26–30].

160

## 161 RESULTS

### 162 Individuals and samples

163 Samples before and after ART were available for 29 individuals who developed M184V and for  
164 19 individuals who developed an NNRTI-resistance mutations including K103N (n=13), Y181C (n=5), and  
165 G190A (n=1). Samples from two time points were available from 32 control individuals who were ART-

166 naïve and who did not develop any RTI-resistance mutations. Table 1 summarizes the demographics,  
167 ART histories, and baseline laboratory values for each group of individuals. Among the 29 individuals in  
168 the M184V group, 19 were NRTI-naïve at the time of the first sample. Nine had previously received 3TC  
169 or FTC but never developed M184V; one had received NRTIs other than 3TC or FTC. Among the 19  
170 individuals in the NNRTI-resistance group, 16 were NNRTI-naïve and three had previously received an  
171 NNRTI but never developed NNRTI resistance.

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

180

### 181 **Baseline sequences**

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



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

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

202           Among the 56 individuals who were ART-naïve at the time baseline sequencing was performed,  
203 22 were found by Sanger sequencing of the *pol* gene to have ambiguous nucleotides (i.e., mixtures) at  
204  $< 0.5\%$  of positions, an established marker suggestive of recent infection [31,32] while 34 were found to  
205 have ambiguous nucleotides at  $\geq 0.5\%$  of positions. The probabilities of having mixtures at positions 200  
206 and 201 were significantly correlated with the proportion of ambiguous nucleotides in *pol* (position 200:  
207 spearman rho=0.28, p=0.03; position 201: spearman rho=0.33; p=0.01).

208           Compared to HXB2, two individuals had large deletions. In one individual, positions 119 to 161  
209 were deleted and in the other individual, positions 281 to 288 were deleted. Both of these deletions  
210 were also present in the follow-up sequence from the same individual suggesting that these deletions  
211 were not sequencing artifacts. An additional 42 baseline samples had deletions of 1 to 3 nucleotides.  
212 The most frequent positions with deletions were between positions 301 to 303 (20 samples) and

213 between positions 150 to 153 (13 samples). Compared to HXB2, 54 samples had one or more insertions,  
214 with the most frequent occurring between positions 300 to 305 (26 samples), 254 to 255 (6 samples),  
215 165 to 167 (5 samples), and 198 to 202 (4 samples). Eight samples had insertions containing more than  
216 three nucleotides.

217

## 218 **Follow-up sequences**

219 Figure 4 summarizes the number of individuals by the number of positions at which a  $\geq 4$ -fold  
220 change in the proportion of a nucleotide or indel occurred between baseline and follow-up for each of  
221 the three sample groups. The median number of changes per individual was higher in the M184V (3;  
222 IQR:1-4;  $p=0.014$  Wilcoxon Rank Sum Test) and NNRTI groups (2; IQR: 1-4;  $p=0.12$ ; Wilcoxon Rank Sum  
223 Test) than in the control group (0.5; IQR:0-4). The three control group individuals with large changes  
224 included one individual who may have experienced a re-infection as evidenced by a nucleotide distance  
225 of 5.4% between their baseline and follow-up HIV-1 *pol* sequence.

226 Fifty-three of the 63 deletions found in either baseline or follow-up samples were present at  
227 both time points. Fifty-three of the 70 insertions present in either baseline or follow-up samples were  
228 present at both time points. A neighbor-joining phylogenetic tree using the Jaccard distance to weight  
229 the distance between overlapping IUPAC nucleotides is shown in Figure 5.

230 Figure 6A shows the positions at which at least one nucleotide changed its proportion by  $\geq 4$ -fold  
231 and which occurred in  $\geq 20\%$  of sequence reads. The positions at which nucleotide changes occurred  
232 most frequently are shown in Figure 6B. Positions 200 and 201, which were the positions most likely to  
233 have mixtures at baseline, were also the positions most likely to exhibit a  $\geq 4$ -fold change in the  
234 proportion of a nucleotide. Position 201 was significantly more likely to exhibit a  $\geq 4$ -fold change in its  
235 proportions in the M184V (9/29 vs. 0/32;  $p=0.0006$ ; Fisher's Exact Test) and the NNRTI-resistance groups

236 (4/19 vs. 0/32;  $p=0.02$ ; Fisher's Exact Test) compared with the control group. There were no significant  
237 differences between each of the three groups at any other position including position 200.

238 Although position 201 frequently displayed a change between baseline and follow-up in the  
239 M184V and NNRTI groups, there was no consistent pattern of nucleotide change. For example, among  
240 the nine individuals in the M184V group that experienced a 4-fold change in the proportion of a  
241 nucleotide, the baseline nucleotides were C in five individuals, T in three individuals, and G in one  
242 individual. Among the five individuals with C at baseline, four changed to T and one to G. Among the  
243 three individuals with T at baseline, one changed to C, another to G, and the third to G and A. The one  
244 individual with G changed to T.

245

#### 246 **Distribution of mixtures in previously published clonal sequences and in NGS**

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

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

259

260 **DISCUSSION**

261 To our knowledge, this is one of the largest studies of 5'-leader sequences in persons living with  
262 HIV-1 and the first study of paired HIV-1 5'-leader sequences from individuals who did or did not  
263 develop RT-associated DRMs. The fact that all of the samples were from plasma suggests that the  
264 observed variants were most likely replication competent, which is much less often the case for proviral  
265 HIV-1 DNA sequences obtained from peripheral blood mononuclear cells (PBMCs) [33]. The availability  
266 of sequences from two time points also provides confirmation that the few observed uncommon  
267 variants, such as large deletions, were consistent with virus replication.

268 Contrary to our initial hypothesis, we did not observe co-evolutionary changes between RT and  
269 5'-leader sequences. The only 5'-leader position (position 201) that demonstrated a higher likelihood of  
270 evolution in samples from individuals who developed RT-associated DRMs is one of the most variable 5'-  
271 leader positions. However, our study uncovered a previously unreported phenomenon, wherein specific  
272 5'-leader positions, particularly positions 200 and 201, exhibited an extraordinarily high likelihood of  
273 containing two or more nucleotides each in more than 20% of NGS reads. We corroborated this finding  
274 by analyzing two previously published datasets: five published studies of clonal dideoxy-terminator  
275 Sanger sequences and six NCBI SRA NGS datasets encompassing the HIV-1 5'-leader.

276 Although HIV-1 is a quasispecies characterized by the presence of multiple circulating variants,  
277 the markedly high prevalence of nucleotide mixtures at two particular positions is striking. The  
278 prevalence of mixtures in our 80 baseline sequences was 45% for position 200 and 29% for position 201.  
279 In our analysis of previously published clonal sequences, these prevalences were 38% and 26%,  
280 respectively, and in our analysis of previously published NGS data, these prevalences were 35% and  
281 21%, respectively. Moreover, in these analyses, positions 200 and 201 contained mixtures at a  
282 frequency several times higher than that observed at any other 5'-leader position or, in the NCBI SRA  
283 NGS dataset, at any RT position. The fact that the previously published clonal and NGS datasets

284 contained large numbers of non-subtype B isolates suggests that the propensity of positions 200 and  
285 201 to contain nucleotide mixtures is not limited to one subtype.

286         Positions 200 and 201 are immediately downstream of the PBS, situated opposite to the  
287 consecutive A nucleotides at tRNA positions 58 and 57, respectively [9]. The significance of  
288 complementarity between the 5'-leader and tRNA at these two positions, however, has not been  
289 studied. There are several plausible non-exclusive explanations for the high proportion of mixtures at  
290 these two positions. First, these positions may be particularly error prone as they represent the first two  
291 nucleotides added following the second strand transfer step during HIV-1 reverse transcription [15].  
292 Second, the presence of multiple circulating variants at these positions may provide an as yet unknown  
293 fitness advantage within an individual patient. Third, HIV-1 replication has been shown to be  
294 occasionally primed by tRNA<sup>5</sup><sub>Lys</sub>, as well as tRNA<sup>3</sup><sub>Lys</sub> [34]; but it is not known whether or how this might  
295 influence the development of mutations just following the PBS. Finally, if the tRNA molecules co-  
296 packaged with genomic RNA re-annealed to the PBS following our nucleic acid extraction procedure but  
297 before reverse transcription, this might also influence mutations just following the PBS.

298         Our study has several limitations. First, our sequences did not encompass the first 36  
299 nucleotides of TAR. The plasma virus samples that we sequenced begin at the start of TAR and could not  
300 be amplified without designing primers that bind to this region making it impossible to sequence the  
301 nucleotides bound to or upstream of our 5'-prime PCR primer. Several studies have reported that the 5'-  
302 leader structure is exquisitely sensitive to mutations in TAR [35–37]. Indeed, a single nucleotide  
303 difference in the transcription start of TAR has been found to influence the overall structure of the 5'-  
304 leader [38].

305         Second, our study is also limited because we studied convenience samples. As a result, we were  
306 unable to match the three groups of patients for several important characteristics such as year of  
307 infection, time between samples, and CD4 count. However, because we found few differences in the

308 evolution of 5'-leader sequence between the three groups of patients and because our main finding, the  
309 extraordinary high proportion of mixtures at positions 200 and 201, was apparent in the baseline  
310 sequences, this limitation is unlikely to have influenced our study's conclusions.

311 Third, M184V has been shown to increase the fidelity of the HIV-1 RT enzyme [39,40]. Thus, we  
312 cannot be certain that the number of mutations in the M184V group was influenced by the  
313 development of this mutation. Nonetheless, M184V does not appear to significantly limit the  
314 evolutionary potential of viruses containing this mutation *in vitro* and likely *in vivo* [41–43]. Finally, we  
315 performed NGS without using unique molecular identifiers (UMIs) [44]. The use of UMIs would have  
316 allowed us to precisely quantify each variant within the HIV-1 quasispecies and to thus reliably identify  
317 linkages between positions in the same variant [44].

318 In conclusion, this is the first study in which paired circulating plasma HIV-1 5'-leader sequences  
319 were obtained from a large number of individuals with well characterized ARV treatment histories at  
320 two time points. The study uncovered a previously unreported phenomenon in which several 5'-leader  
321 positions, particularly positions 200 and 201, which are immediately downstream of the PBS, often  
322 contain high proportions of different nucleotides in the same sample. Data from this study also suggests  
323 that nucleotide mixtures at these positions increase with the duration of infection. Future studies  
324 employing deep sequencing will provide more insight into this phenomenon by determining whether  
325 different nucleotides at positions 200 and 201 are associated with changes at other 5'-leader positions.  
326 Additionally, mutational studies will be required to determine the potential biological significance of  
327 variations at these positions.

328

329

330

331 Acknowledgements

332 JN and RWS have been funded in part by an NIH grant R03 AI147632. EVP, JDP and MK have been

333 funded in part by an NIH grant U54 AI170856.

334

335 Conflicts of interest

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

337

338 **FIGURE LEGENDS**

339 **Figure 1**

340 HIV-1 5'-leader nucleotide differences from HXB2 observed in the 80 baseline sequences superimposed  
341 on a diagram of its RNA secondary structure. Nucleotides observed in  $\geq 20\%$  of reads in more than 10%  
342 of sequences are shown in orange; those present in fewer than 10% of sequences are shown in purple.  
343 Insertions and deletions are not shown. Positions 1 to 36 were not sequenced. TAR: trans-activation  
344 response element (positions 1-57); PolyA: polyadenylation signal (positions 58-104); H1: helix 1 tertiary  
345 structure; H2: helix 2 tertiary structure; PBS: primer binding site (positions 182-199); DIS: dimer  
346 initiation site (positions 242-278); psi: packaging signal (positions 312-315). The Matrix initiation codon  
347 is surrounded by a green box; it is base-paired to the U5 triplet. The primer activation signal nucleotides  
348 are colored green, the C-rich region nucleotides are colored yellow, and the PBS nucleotides are colored  
349 maroon.

350

351 **Figure 2**

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

354

355 **Figure 3**

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

358

359 **Figure 4**



360 Number of HIV-1 5'-leader positions at which a nucleotide or indel exhibited a  $\geq 4$ -fold change in its  
361 proportion between baseline and follow-up and which was present at a level of  $\geq 20\%$  at follow-up in  
362 each of the three patient groups.

363

364 Figure 5

365 Neighbor-joining tree containing the 80 baseline and follow-up sequences from this study.  
366 Sample IDs for the M184V group are colored blue; those for the NNRTI group are colored  
367 yellow; and those for the Control group are colored red. Bootstrap values are indicated at each  
368 node of the tree.

369

370 Figure 6

371 Distribution of positions at which a nucleotide or indel changed its prevalence by  $\geq 4$ -fold  
372 according to patient group (A) and a table indicating the ten positions that changed most often  
373 (B).

374

375 Supplementary Figure 1

376 HIV-1 5'-leader nucleotide differences from the HXB2 sequence observed in 1,417 sequences in  
377 the Los Alamos National Laboratories HIV Sequence Database dataset. Superscripts indicate the  
378 percentage of times that a nucleotide was reported. Highlighted regions include (1) TAR: trans-  
379 activation response element; (2) PolyA: polyadenylation signal loop with a box surrounding the  
380 poly A motif; (3) U5; (4) PAS: primer activation signal; (5) PBS: primer binding site; (6) DIS: dimer  
381 initiation signal loop with a box surrounding the palindromic DIS; (7) major 5' splicing site; (8)  
382 PSI packaging signal; (9) Matrix protein with a box surrounding the start codon.

383

384 Supplementary Figure 2

385 Distribution of HIV-1 5'-leader nucleotide differences from HXB2 and indels in the 56 baseline sequences

386 from ART-naïve individuals compared with the 24 baseline sequences from ART-experienced individuals.

387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433

## REFERENCES

1. Berkhout, B. Structure and Function of the Human Immunodeficiency Virus Leader RNA. In *Progress in Nucleic Acid Research and Molecular Biology*; Cohn, W.E., Moldave, K., Eds.; Academic Press, 1996; Vol. 54, pp. 1–34.
2. Abbink, T.E.M.; Berkhout, B. A Novel Long Distance Base-Pairing Interaction in Human Immunodeficiency Virus Type 1 RNA Occludes the Gag Start Codon. *J. Biol. Chem.* **2003**, *278*, 11601–11611, doi:10.1074/jbc.M210291200.
3. Song, R.; Kafaie, J.; Laughrea, M. Role of the 5' TAR Stem-Loop and the U5-AUG Duplex in Dimerization of HIV-1 Genomic RNA. *Biochemistry* **2008**, *47*, 3283–3293, doi:10.1021/bi7023173.
4. Isel, C.; Ehresmann, C.; Marquet, R. Initiation of HIV Reverse Transcription. *Viruses* **2010**, *2*, 213–243, doi:10.3390/v2010213.
5. van Bel, N.; Ghabri, A.; Das, A.T.; Berkhout, B. The HIV-1 Leader RNA Is Exquisitely Sensitive to Structural Changes. *Virology* **2015**, *483*, 236–252, doi:10.1016/j.virol.2015.03.050.
6. Tomezsko, P.J.; Corbin, V.D.A.; Gupta, P.; Swaminathan, H.; Glasgow, M.; Persad, S.; Edwards, M.D.; Mcintosh, L.; Papenfuss, A.T.; Emery, A.; et al. Determination of RNA Structural Diversity and Its Role in HIV-1 RNA Splicing. *Nature* **2020**, *582*, 438–442, doi:10.1038/s41586-020-2253-5.
7. Ye, L.; Gribling-Burrer, A.-S.; Bohn, P.; Kibe, A.; Börtlein, C.; Ambi, U.B.; Ahmad, S.; Olguin-Nava, M.; Smith, M.; Caliskan, N.; et al. Short- and Long-Range Interactions in the HIV-1 5' UTR Regulate Genome Dimerization and Packaging. *Nat Struct Mol Biol* **2022**, *29*, 306–319, doi:10.1038/s41594-022-00746-2.
8. Keane, S.C.; Summers, M.F. NMR Studies of the Structure and Function of the HIV-1 5'-Leader. *Viruses* **2016**, *8*, doi:10.3390/v8120338.
9. Larsen, K.P.; Mathiharan, Y.K.; Kappel, K.; Coey, A.T.; Chen, D.-H.; Barrero, D.; Madigan, L.; Puglisi, J.D.; Skiniotis, G.; Puglisi, E.V. Architecture of an HIV-1 Reverse Transcriptase Initiation Complex. *Nature* **2018**, *557*, 118–122, doi:10.1038/s41586-018-0055-9.
10. Ha, B.; Larsen, K.P.; Zhang, J.; Fu, Z.; Montabana, E.; Jackson, L.N.; Chen, D.-H.; Puglisi, E.V. High-Resolution View of HIV-1 Reverse Transcriptase Initiation Complexes and Inhibition by NNRTI Drugs. *Nat Commun* **2021**, *12*, 2500, doi:10.1038/s41467-021-22628-9.
11. Dutilleul, A.; Rodari, A.; Van Lint, C. Depicting HIV-1 Transcriptional Mechanisms: A Summary of What We Know. *Viruses* **2020**, *12*, 1385, doi:10.3390/v12121385.
12. Emery, A.; Swanstrom, R. HIV-1: To Splice or Not to Splice, That Is the Question. *Viruses* **2021**, *13*, 181, doi:10.3390/v13020181.
13. Lu, K.; Heng, X.; Summers, M.F. Structural Determinants and Mechanism of HIV-1 Genome Packaging. *Journal of Molecular Biology* **2011**, *410*, 609–633, doi:10.1016/j.jmb.2011.04.029.
14. Mailler, E.; Bernacchi, S.; Marquet, R.; Paillart, J.-C.; Vivet-Boudou, V.; Smyth, R.P. The Life-Cycle of the HIV-1 Gag-RNA Complex. *Viruses* **2016**, *8*, 248, doi:10.3390/v8090248.
15. Krupkin, M.; Jackson, L.N.; Ha, B.; Puglisi, E.V. Advances in Understanding the Initiation of HIV-1 Reverse Transcription. *Current Opinion in Structural Biology* **2020**, *65*, 175–183, doi:10.1016/j.sbi.2020.07.005.
16. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor. *Bioinformatics* **2018**, *34*, i884–i890, doi:10.1093/bioinformatics/bty560.
17. Li, H. Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* **2018**, *34*, 3094–3100, doi:10.1093/bioinformatics/bty191.
18. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079, doi:10.1093/bioinformatics/btp352.

- 434 19. Rice, P.; Longden, I.; Bleasby, A. EMBOS: The European Molecular Biology Open Software Suite.  
435 *Trends in Genetics* **2000**, *16*, 276–277, doi:10.1016/S0168-9525(00)02024-2.
- 436 20. Katoh, K.; Frith, M.C. Adding Unaligned Sequences into an Existing Alignment Using MAFFT and  
437 LAST. *Bioinformatics* **2012**, *28*, 3144–3146, doi:10.1093/bioinformatics/bts578.
- 438 21. Novitsky, V.A.; Montano, M.A.; McLane, M.F.; Renjifo, B.; Vannberg, F.; Foley, B.T.; Ndung'u, T.P.;  
439 Rahman, M.; Makhema, M.J.; Marlink, R.; et al. Molecular Cloning and Phylogenetic Analysis of  
440 Human Immunodeficiency Virus Type 1 Subtype C: A Set of 23 Full-Length Clones from Botswana. *J*  
441 *Virology* **1999**, *73*, 4427–4432.
- 442 22. Rolland, M.; Tovanabutra, S.; deCamp, A.C.; Frahm, N.; Gilbert, P.B.; Sanders-Buell, E.; Heath, L.;  
443 Magaret, C.A.; Bose, M.; Bradfield, A.; et al. Genetic Impact of Vaccination on Breakthrough HIV-1  
444 Sequences from the STEP Trial. *Nat Med* **2011**, *17*, 366–371, doi:10.1038/nm.2316.
- 445 23. Ochsenbauer, C.; Edmonds, T.G.; Ding, H.; Keele, B.F.; Decker, J.; Salazar, M.G.; Salazar-Gonzalez,  
446 J.F.; Shattock, R.; Haynes, B.F.; Shaw, G.M.; et al. Generation of Transmitted/Founder HIV-1  
447 Infectious Molecular Clones and Characterization of Their Replication Capacity in CD4 T  
448 Lymphocytes and Monocyte-Derived Macrophages. *Journal of Virology* **2012**, *86*, 2715–2728,  
449 doi:10.1128/JVI.06157-11.
- 450 24. Parrish, N.F.; Gao, F.; Li, H.; Giorgi, E.E.; Barbian, H.J.; Parrish, E.H.; Zajic, L.; Iyer, S.S.; Decker, J.M.;  
451 Kumar, A.; et al. Phenotypic Properties of Transmitted Founder HIV-1. *Proc. Natl. Acad. Sci. U.S.A.*  
452 **2013**, *110*, 6626–6633, doi:10.1073/pnas.1304288110.
- 453 25. Gondim, M.V.P.; Sherrill-Mix, S.; Bibollet-Ruche, F.; Russell, R.M.; Trimboli, S.; Smith, A.G.; Li, Y.; Liu,  
454 W.; Avitto, A.N.; DeVoto, J.C.; et al. Heightened Resistance to Host Type 1 Interferons  
455 Characterizes HIV-1 at Transmission and after Antiretroviral Therapy Interruption. *Sci Transl Med*  
456 **2021**, *13*, eabd8179, doi:10.1126/scitranslmed.abd8179.
- 457 26. Gall, A.; Ferns, B.; Morris, C.; Watson, S.; Cotten, M.; Robinson, M.; Berry, N.; Pillay, D.; Kellam, P.  
458 Universal Amplification, Next-Generation Sequencing, and Assembly of HIV-1 Genomes. *J Clin*  
459 *Microbiol* **2012**, *50*, 3838–3844, doi:10.1128/JCM.01516-12.
- 460 27. Illingworth, C.J.R.; Roy, S.; Beale, M.A.; Tutill, H.; Williams, R.; Breuer, J. On the Effective Depth of  
461 Viral Sequence Data. *Virus Evol* **2017**, *3*, vex030, doi:10.1093/ve/vex030.
- 462 28. Fedonin, G.G.; Fantin, Y.S.; Favorov, A.V.; Shipulin, G.A.; Neverov, A.D. VirGenA: A Reference-Based  
463 Assembler for Variable Viral Genomes. *Briefings in Bioinformatics* **2019**, *20*, 15–25,  
464 doi:10.1093/bib/bbx079.
- 465 29. Vilsker, M.; Moosa, Y.; Nooij, S.; Fonseca, V.; Ghysens, Y.; Dumon, K.; Pauwels, R.; Alcantara, L.C.;  
466 Vanden Eynden, E.; Vandamme, A.-M.; et al. Genome Detective: An Automated System for Virus  
467 Identification from High-Throughput Sequencing Data. *Bioinformatics* **2019**, *35*, 871–873,  
468 doi:10.1093/bioinformatics/bty695.
- 469 30. Zhang, Y.; Wymant, C.; Laeyendecker, O.; Grabowski, M.K.; Hall, M.; Hudelson, S.; Piwowar-  
470 Manning, E.; McCauley, M.; Gamble, T.; Hosseinipour, M.C.; et al. Evaluation of Phylogenetic  
471 Methods for Inferring the Direction of Human Immunodeficiency Virus (HIV) Transmission: HIV  
472 Prevention Trials Network (HPTN) 052. *Clin Infect Dis* **2020**, *72*, 30–37, doi:10.1093/cid/ciz1247.
- 473 31. Kouyos, R.D.; von Wyl, V.; Yerly, S.; Böni, J.; Rieder, P.; Joos, B.; Taffé, P.; Shah, C.; Bürgisser, P.;  
474 Klimkait, T.; et al. Ambiguous Nucleotide Calls From Population-Based Sequencing of HIV-1 Are a  
475 Marker for Viral Diversity and the Age of Infection. *Clinical Infectious Diseases* **2011**, *52*, 532–539,  
476 doi:10.1093/cid/ciq164.
- 477 32. Andersson, E.; Shao, W.; Bontell, I.; Cham, F.; Cuong, D.D.; Wondwossen, A.; Morris, L.; Hunt, G.;  
478 Sönnernborg, A.; Bertagnolio, S.; et al. Evaluation of Sequence Ambiguities of the HIV-1 Pol Gene as  
479 a Method to Identify Recent HIV-1 Infection in Transmitted Drug Resistance Surveys. *Infection,*  
480 *Genetics and Evolution* **2013**, *18*, 125–131, doi:10.1016/j.meegid.2013.03.050.

- 481 33. Bruner, K.M.; Murray, A.J.; Pollack, R.A.; Soliman, M.G.; Laskey, S.B.; Capoferri, A.A.; Lai, J.; Strain,  
482 M.C.; Lada, S.M.; Hoh, R.; et al. Defective Proviruses Rapidly Accumulate during Acute HIV-1  
483 Infection. *Nat Med* **2016**, *22*, 1043–1049, doi:10.1038/nm.4156.
- 484 34. Das, A.T.; Vink, M.; Berkhout, B. Alternative TRNA Priming of Human Immunodeficiency Virus Type 1  
485 Reverse Transcription Explains Sequence Variation in the Primer-Binding Site That Has Been  
486 Attributed to APOBEC3G Activity. *Journal of Virology* **2005**, *79*, 3179–3181,  
487 doi:10.1128/jvi.79.5.3179-3181.2005.
- 488 35. Huthoff, H.; Berkhout, B. Mutations in the TAR Hairpin Affect the Equilibrium between Alternative  
489 Conformations of the HIV-1 Leader RNA. *Nucleic Acids Research* **2001**, *29*, 2594–2600,  
490 doi:10.1093/nar/29.12.2594.
- 491 36. Vrolijk, M.M.; Ooms, M.; Harwig, A.; Das, A.T.; Berkhout, B. Destabilization of the TAR Hairpin  
492 Affects the Structure and Function of the HIV-1 Leader RNA. *Nucleic Acids Research* **2008**, *36*,  
493 4352–4363, doi:10.1093/nar/gkn364.
- 494 37. Ding, P.; Kharytonchyk, S.; Kuo, N.; Cannistraci, E.; Flores, H.; Chaudhary, R.; Sarkar, M.; Dong, X.;  
495 Telesnitsky, A.; Summers, M.F. 5'-Cap Sequestration Is an Essential Determinant of HIV-1 Genome  
496 Packaging. *Proc Natl Acad Sci U S A* **2021**, *118*, e2112475118, doi:10.1073/pnas.2112475118.
- 497 38. Brown, J.D.; Kharytonchyk, S.; Chaudry, I.; Iyer, A.S.; Carter, H.; Becker, G.; Desai, Y.; Glang, L.; Choi,  
498 S.H.; Singh, K.; et al. Structural Basis for Transcriptional Start Site Control of HIV-1 RNA Fate.  
499 *Science* **2020**, *368*, 413–417, doi:10.1126/science.aaz7959.
- 500 39. Oude Essink, B.B.; Back, N.K.T.; Berkhout, B. Increased Polymerase Fidelity of the 3TC-Resistant  
501 Variants of HIV-1 Reverse Transcriptase. *Nucleic Acids Research* **1997**, *25*, 3212–3217,  
502 doi:10.1093/nar/25.16.3212.
- 503 40. Hsu, M.; Inouye, P.; Rezende, L.; Richard, N.; Li, Z.; Prasad, V.R.; Wainberg, M.A. Higher Fidelity of  
504 RNA-Dependent DNA Mismatch Extension by M184V Drug-Resistant than Wild-Type Reverse  
505 Transcriptase of Human Immunodeficiency Virus Type 1. *Nucleic Acids Research* **1997**, *25*, 4532–  
506 4536, doi:10.1093/nar/25.22.4532.
- 507 41. Jonckheere, H.; Witvrouw, M.; De Clercq, E.; Anné, J. Short Communication: Lamivudine Resistance  
508 of HIV Type 1 Does Not Delay Development of Resistance to Nonnucleoside HIV Type 1-Specific  
509 Reverse Transcriptase Inhibitors as Compared with Wild-Type HIV Type 1. *AIDS Research and*  
510 *Human Retroviruses* **1998**, *14*, 249–253, doi:10.1089/aid.1998.14.249.
- 511 42. Keulen, W.; van Wijk, A.; Schuurman, R.; Berkhout, B.; Boucher, C.A.B. Increased Polymerase Fidelity  
512 of Lamivudine-Resistant HIV-1 Variants Does Not Limit Their Evolutionary Potential. *AIDS* **1999**, *13*,  
513 1343.
- 514 43. Stockdale, A.J.; Saunders, M.J.; Boyd, M.A.; Bonnett, L.J.; Johnston, V.; Wandeler, G.; Schoffelen,  
515 A.F.; Ciaffi, L.; Stafford, K.; Collier, A.C.; et al. Effectiveness of Protease Inhibitor/Nucleos(t)ide  
516 Reverse Transcriptase Inhibitor-Based Second-Line Antiretroviral Therapy for the Treatment of  
517 Human Immunodeficiency Virus Type 1 Infection in Sub-Saharan Africa: A Systematic Review and  
518 Meta-Analysis. *Clinical Infectious Diseases* **2018**, *66*, 1846–1857, doi:10.1093/cid/cix1108.
- 519 44. Zhou, S.; Hill, C.S.; Spielvogel, E.; Clark, M.U.; Hudgens, M.G.; Swanstrom, R. Unique Molecular  
520 Identifiers and Multiplexing Amplicons Maximize the Utility of Deep Sequencing To Critically Assess  
521 Population Diversity in RNA Viruses. *ACS Infect. Dis.* **2022**, *8*, 2505–2514,  
522 doi:10.1021/acinfecdis.2c00319.
- 523

524 **Table 1. Description of the 80 Individuals Undergoing Baseline and Follow-Up HIV-1 5'-Leader Sequencing**  
 525

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

526

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

534

535 **Table 2. Proportion of HIV-1 5'-Leader Positions Containing a Mixture of Nucleotides in Sampled Viruses in this Study, in Previously Published**  
 536 **Studies in GenBank, and in NGS Studies in the NCBI Sequence Read Archive (SRA)**  
 537

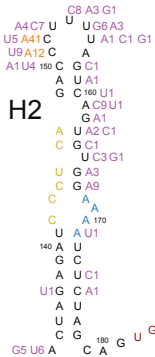
Position	Baseline Samples from this Study (n=80 individuals) <sup>1</sup>	Published Studies of Individuals with Multiple Clones (GenBank) (n=42 individuals) <sup>2</sup>	Published Studies of Individuals Undergoing NGS (NCBI SRA) (n=295 individuals) <sup>3</sup>
200	45.0%	38.1%	34.9%
201	28.8%	26.2%	21.1%
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%

538  
 539 Footnote: <sup>1</sup>A position was considered to have a mixture if at least one non-consensus nucleotide was present in  $\geq 20\%$  of reads. <sup>2</sup>Data were  
 540 obtained from five publications reporting  $\geq 3$  clones per individual from  $\geq 5$  individuals. A total of 294 sequences were performed (i.e., a mean of  
 541 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  
 542 (74%) had subtype B viruses and 11 patients had subtype C viruses. <sup>3</sup>A position was considered to have a mixture if at least one non-consensus  
 543 nucleotide was present in  $\geq 20\%$  of reads. Sixty-four percent of sequences belonged to subtype C; 21% to subtype B; and 15% to other subtypes  
 544 or circulating recombinant forms.

# TAR

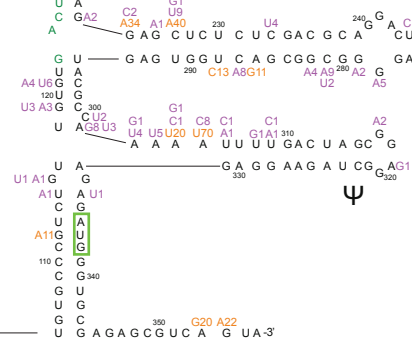


# Poly(A)

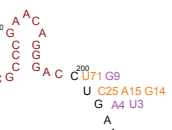


# H2

# H1



# PBS



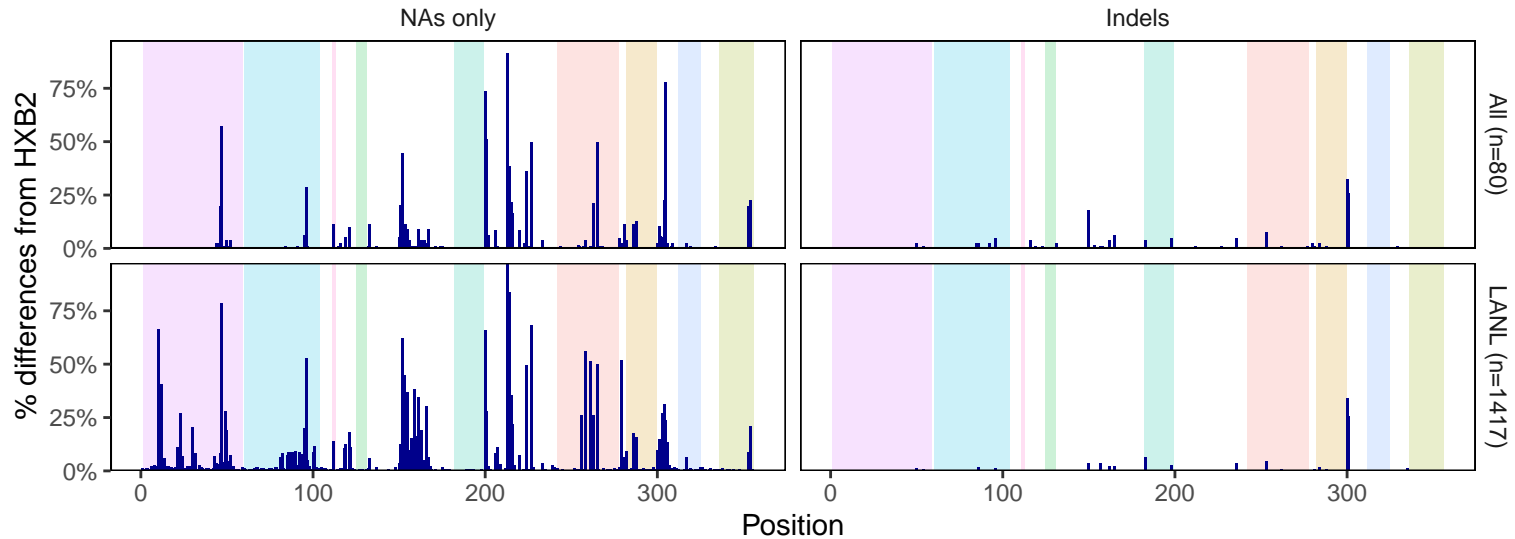
# DIS

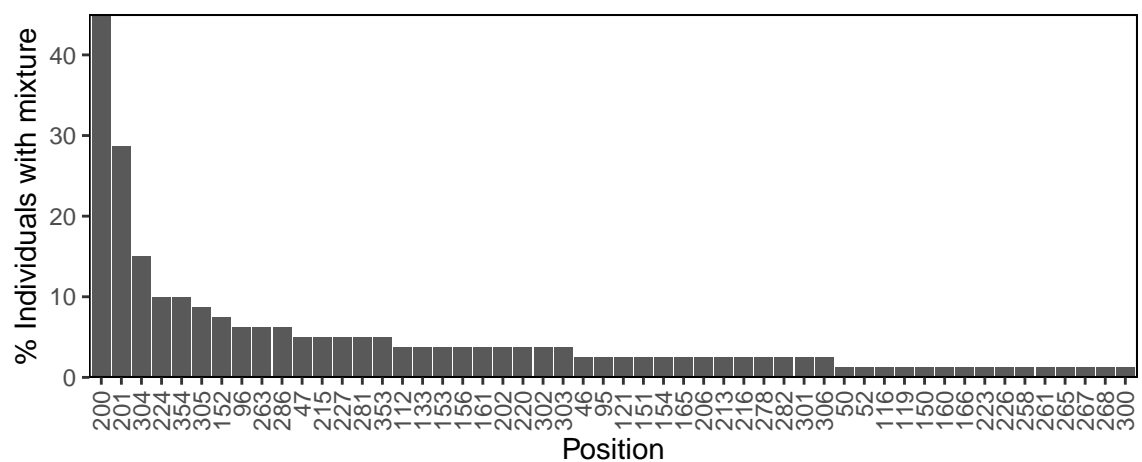


# ψ

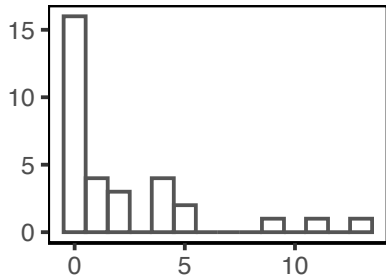




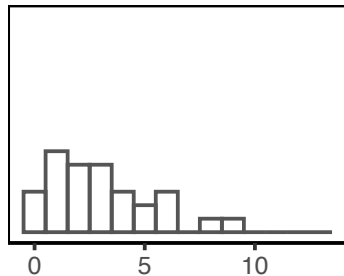




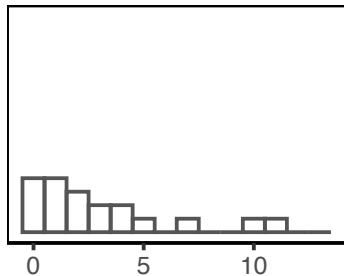
Control



M184V

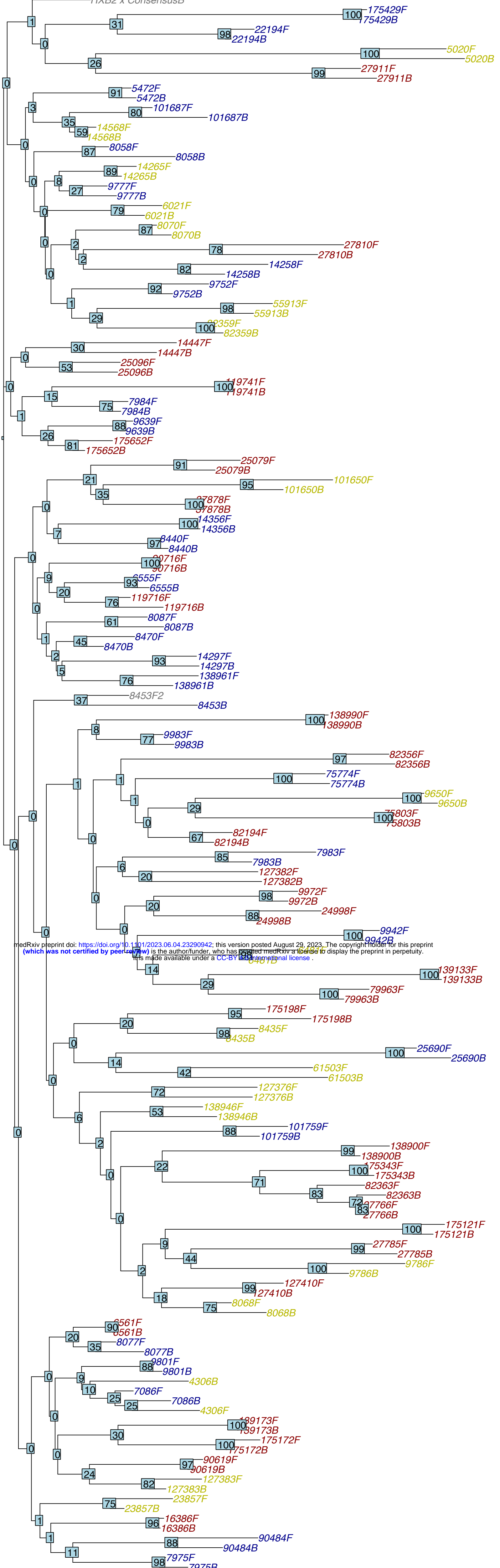


NNRTI



# Positions with  $\geq 4$ -fold change

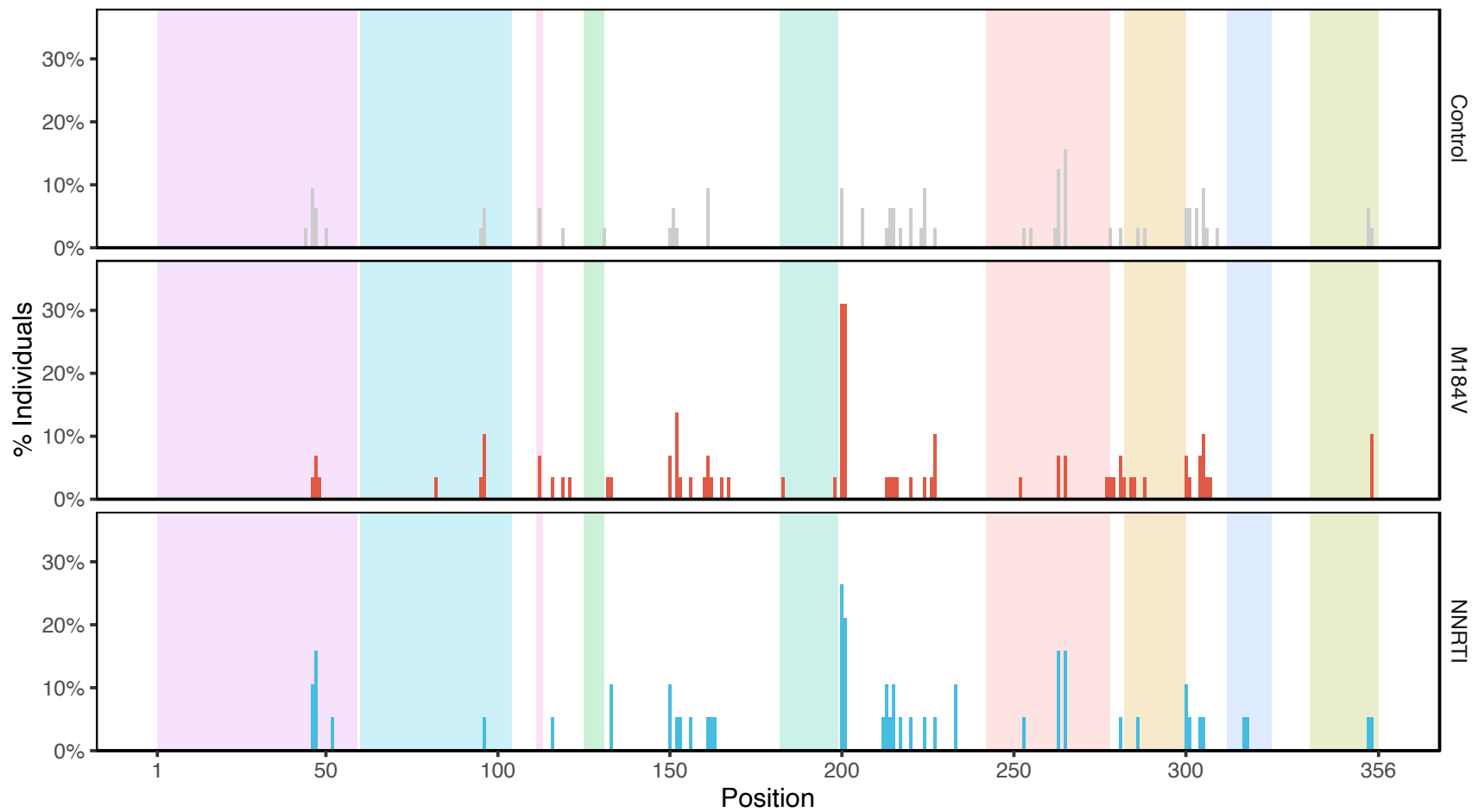
HXB2 x ConsensusB



medRxiv preprint doi: <https://doi.org/10.1101/2023.06.04.23290942>; this version posted August 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

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)