

The evolution of regulatory elements in the emerging promoter variant strains of HIV-1

Disha Bhange^a, Nityanand Prasad^a, Swati Singh^a, Harshit Kumar Prajapati^a, Shesh Prakash Maurya^b, Bindu Parachalil Gopalan^c, Sowmya Nadig^c, Devidas Chaturbhuj^d, Jayaseelan Boobalan^e, Thongadi Ramesh Dinesha^e, Syed Fazil Ahamed^c, Navneet Singh^a, Anangi Brahmaiah^a, Kavita Mehta^a, Yuvrajsinh Gohil^a, Pachamuthu Balakrishnan^f, Bimal Kumar Das^b, Mary Dias^c, Raman Gangakhedkar^g, Sanjay Mehendale^h, Ramesh Paranjape^g, Shanmugam Saravanan^e, Anita Shet^c, Sunil Suhas Solomon^{i,j}, Madhuri Thakar^d, and Udaykumar Ranga^{a*}

^aMolecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research (JNCASR), India 560064;

^bDepartment of Microbiology, HIV Immunology Laboratory, All India Institute of Medical Sciences (AIIMS), India 110029;

^cDivision of Microbiology/ Infectious Diseases Unit, St. John's National Academy of Health Sciences, India 560034;

^dDepartment of Serology and Immunology, National AIDS Research Institute (NARI), India 411026;

^eDepartment of Molecular Biology and Genotyping, Y. R. Gaitonde Centre for AIDS Research and Education (YRG CARE), India-600113;

^fInfectious Diseases Laboratory, Y. R. Gaitonde Centre for AIDS Research and Education (YRG CARE), Chennai, India 600113;

^gDepartment of Clinical Sciences, National AIDS Research Institute (NARI), India 411026;

^hP. G. Hinduja National Hospital and Medical Research Centre, Mumbai, India 400016;

ⁱY. R. Gaitonde Center for AIDS Research and Education (YRG CARE), Chennai, Tamil Nadu, India 600113;

^jDepartment of Medicine, Johns Hopkins University, School of Medicine, Baltimore, Maryland, United States of America.

***Corresponding author:** Udaykumar Ranga, HIV AIDS Laboratory, Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research Centre, Jakkur, Bengaluru, India 560064; Phone number: +91-80-2208 2830.

Email: udaykumar@jncasr.ac.in

Author Contributions: D.B. performed research, analyzed data, and wrote the paper, N.P., S.P.M., B.P.G., S.N., D.C., J.B., T.R.D, S.F.A., N.S., A.B. and K.M. performed research, S.S.^a, and Y.G. analyzed data, H.K.P. analyzed NGS data, P.B., B.K.D., M.D., R.G., S.M., R.P., S.S.^f, A.S., S.S.S., and M.T. designed research, U.R. designed research and wrote the paper.

Competing Interest Statement: The authors declare no competing interests.

Classification: Biological sciences, Microbiology

Keywords: HIV-1, subtype C, evolution, sequence duplication, latency

This PDF file includes:

Main Text, figures (1-5), tables (1,2)

42 **Abstract**

43

44 In a multicentric, observational, investigator-blinded, and longitudinal clinical study of 764 ART-naïve
45 subjects, we identified nine different promoter-variant strains of HIV-1 subtype C (HIV-1C) emerging in
46 the Indian population, with some of these variants being reported for the first time. Unlike several
47 previous studies, our work here focuses on the evolving viral regulatory elements, not coding
48 sequences. The emerging viral strains contain additional copies of the existing transcription factor
49 binding sites (TFBS), including TCF-1 α /LEF-1, RBEIII, AP-1, and NF- κ B, created by sequence
50 duplication. The additional TFBS are genetically diverse and may blur the distinction between the
51 modulatory region of the promoter and the viral enhancer. In a follow-up analysis, we found trends, but
52 not significant associations between any specific variant promoter and prognostic markers, probably
53 because the emerging viral strains might not have established mono infections yet. Illumina sequencing
54 of four clinical samples containing a co-infection indicated the domination of one strain over the other
55 and establishing a stable ratio with the second strain at the follow-up time-points. Since a single
56 promoter regulates viral gene expression and constitutes the master regulatory circuit with Tat, the
57 acquisition of additional and variant copies of the TFBS may significantly impact viral latency and latent
58 reservoir characteristics. Further studies are urgently warranted to understand how the diverse TFBS
59 profiles of the viral promoter may modulate the characteristics of the latent reservoir, especially following
60 the initiation of antiretroviral therapy.

61

62

63 **Significance Statement**

64

65 A unique conglomeration of TFBS enables the HIV-1 promoter to accomplish two diametrically opposite
66 functions – transcriptional activation and transcriptional silencing. The various phases of viral latency -
67 establishment, maintenance, and reversal - collectively determine the replication fitness of individual
68 viral strains. A profound variation in the TFBS composition of the viral promoter may significantly alter
69 the viral latency properties and the latent reservoir characteristics. Although the duplication of certain
70 TFBS remains a quality unique to HIV-1C, the high-level genetic recombination of HIV-1 may promote
71 the transfer of such molecular properties to the other HIV-1 subtypes. The emergence of several
72 promoter-variant viral strains may make the task of a ‘functional cure’ more challenging in HIV-1C.

73

74

75 **Introduction**

76

77 Based on phylogenetic association, the viral strains of HIV-1 are classified into four groups (M, N, O,
78 and P), and within group M, into ten different genetic subtypes, A, B, C, D, F, G, H, J, K, L (1), and
79 numerous recombinant forms. Of the various genetic subtypes of HIV-1 unevenly distributed globally,
80 HIV-1C and its recombinant forms are responsible for nearly half of the global infections (2–4). Despite
81 the high prevalence of HIV-1C, only a limited number of studies are available examining the causes
82 underlying the expansion of these viral strains and their impact on disease manifestation.

83

84 Although the basic architecture of HIV-1 LTR is broadly conserved among the diverse HIV-1 genetic
85 families, subtype-associated differences are manifested (5). The configuration of transcription factor
86 binding sites (TFBS), including those of NF- κ B, NF-AT, AP-1, and other regulatory elements such as
87 the TATA box, and the TAR region, in HIV-1C LTR (C-LTR) differs from that of the other viral subtypes
88 (5). Of the TFBS variations, differences in the copy number and sequence of the NF- κ B motif are unique
89 to HIV-1C. HIV-1C LTR typically contains three or four NF- κ B motifs in the enhancer region compared
90 to only one motif present in HIV-1A/E or two motifs in all the other HIV-1 subtypes (6). Further, the
91 additional copies of the NF- κ B motif in C-LTR are genetically variable, alluding to the possibility of the
92 viral promoter being receptive to a diverse and broader range of cellular signals. For instance, the four
93 copies of the NF- κ B motif in the enhancer of 4-kB viral strains represent three genetically distinct NF-
94 κ B binding sites (6). Apart from the NF- κ B motif, other regulatory elements, including AP-1, RBEIII, and
95 TCF-1 α /LEF-1, also show subtype-associated variations, although the impact of such variations has
96 not been examined.

97

98 Several publications reported the insertion or deletion of TFBS in HIV LTR. One example is the
99 sequence duplication of the TCF-1 α /LEF-1, RBF-2, AP-1, and c-EBP α binding motif in the modulatory
100 region of the LTR, technically called the most frequent naturally occurring length polymorphism
101 (MFNLP) (7–10). Approximately 38% of the HIV-1B viral isolates contain MFNLP, a phenomenon

102 believed to be a compensatory mechanism to ensure the presence of at least one functional RBEIII site
103 in the LTR (11). Although RBEIII duplication has been found in several subtypes, the significance of
104 this phenomenon has been examined predominantly in HIV-1B. It, however, remains inconclusive
105 whether the presence of RBEIII duplication is directly associated with reduced viral replication or slower
106 disease progression (12). A small number of reports examined RBEIII duplication in HIV-1C infection,
107 however, without evaluating its effect on the replication fitness of the viral strains and disease
108 progression (13).

109
110 Over the past several years, our laboratory has documented the emergence of LTR-variant strains of
111 HIV-1C in India and elsewhere (14–16). While the appearance of genetic diversity and such diversity
112 impacting viral evolution are common to the various genetic subtypes of HIV-1, the genetic variation we
113 describe in HIV-1C is non-sporadic and radically different in an important aspect. Viral evolution in HIV-
114 1C appears to be directional towards modulating transcriptional strength of the promoter by creating
115 additional copies of the existing TFBS, such as NF- κ B, AP1, RBEIII, and TCF-1 α /LEF-1 motifs, by
116 sequence duplication and co-duplication. A single viral promoter in HIV-1 regulates two diametrically
117 opposite functions critical for viral survival - transcriptional activation and silencing. Hence, any variation
118 in the constitution of the TFBS (copy number difference and/or genetic variation), may have a profound
119 impact on viral replication fitness.

120
121 Here, in a multicentric, observational, non-interventional, investigator-blinded, and longitudinal clinical
122 study, we examined the promoter sequences of 455 primary viral isolates derived from ART-naïve
123 subjects. We show that the magnitude of TFBS variation is much larger than we reported previously. At
124 least nine different TFBS variant viral strains have emerged in recent years. Using the Illumina MiSeq
125 platform, we attempted to characterize the proviral DNA of a selected subset of viral variants containing
126 the RBEIII motif duplication. The data allude to the possibility that some of the emerging strains could
127 achieve greater replication fitness levels and may establish expanding epidemics in the future, which
128 requires monitoring. This work provides important insights into the HIV-1 evolution taking place at the
129 level of population in India.

130
131

132 Results

133

134 The magnitude of TFBS variation in HIV-1C LTR

135

136 We collected 764 primary clinical samples from four different clinical sites in India, All India Institute of
137 Medical Sciences (AIIMS), New Delhi; National AIDS Research Institute (NARI), Pune; St. John's
138 Medical Hospital, Bangalore; and Y. R. Gaitonde Centre for AIDS Research and Education (YRG
139 CARE), Chennai, between 2017-2019. Using the genomic DNA, we determined the sequence of the
140 U3 region in the LTR of 520 of 764 viral samples, whereas the amplification of the rest of the samples
141 failed. The genetic typing of 455 of 520 sequences could be accomplished successfully. The proportion
142 of viral variants across the four clinical sites was comparable without geographic skewing (*SI Appendix*,
143 Table S3). The large majority of the variant LTRs contain additional copies of the existing TFBS,
144 including that of NF- κ B, RBF-2, AP-1, and TCF-1 α /LEF-1 binding sites (Fig. 1, *SI Appendix*, Fig. S1).
145 Based on the TFBS profile, the sequences of the copied TFBS, and their temporal location, the various
146 viral promoters may be classified into three categories.

147

148 Category-1 is represented by the viral promoters containing three NF- κ B motifs without any other TFBS
149 duplication. This group consists of the canonical LR-HHC-LTR and a new variant LR-FHC-LTR (see
150 below). The canonical LR-HHC-LTR represents the largest group among all HIV-1C promoters,
151 comprising 303 of 455 sequences (66.6%). This LTR contains three tandemly arranged NF- κ B sites in
152 the enhancer representing two distinct κ B-motifs, two H- κ B motifs (5'-GGGACTTTCC-3'), and one C-
153 κ B motif (5'-GGGGCGTTTTCC-3', differences underlined). Immediately upstream of the viral enhancer,
154 an RBF-2 binding site (R, RBEIII motif, 5'-ACTGCTGA-3') and further upstream a TCF-1 α /LEF-1 site
155 (L, 5'-TACAAA/GG/A-3') is located. The canonical HIV-1C promoter is identified here as LR-HHC-LTR
156 to denote the specific arrangement of the three categories of TFBS from 5' to 3'. The second member
157 of the group contains a variant LR-FHC-LTR comprising 40 of 455 (8.8%) sequences. The characteristic
158 feature of these viral strains is the presence of three genetically distinct NF- κ B motifs in the viral
159 enhancer. Thus, the two viral strains of category-1 formed the major proportion of all the LTRs, 343/455
160 (75.4%) (Fig. 1). Multisequence alignments of several representative viral strains of the canonical LR-

161 HHC (Fig. 2A) and the variant LR-FHC (Fig. 2B) are presented. The LR-FHC-LTR, being reported here
162 for the first time, may have originated from the 4-kB viral strain FHHC of category 2 described below.
163

164 Category-2 viral LTRs contain an additional (fourth) NF- κ B binding site (F- κ B site, 5'-GGGACTTTCT-
165 3') located downstream of the RBEIII site. The 4-kB LTRs thus, contain one F-, two H-, and one C- κ B
166 motifs, in that order, hence are labeled LR-FHHC. An alignment of several representative 4-kB viral
167 sequences shows a high degree of sequence conservation at all the TFBS (Fig. 2C).
168

169 **Double RBEIII LTRs show profound variation in number, genetic sequence, and position of TFBS**

170

171 Category-3 viral strains, representing 18.7% (85/455) sequences analyzed here, are characterized by
172 the duplication of the RBEIII site in the modulatory region upstream of the viral enhancer, which is
173 analogous of MFNLP described previously in HIV-1B (Fig. 1). Several unique molecular properties
174 qualify the duplication of the RBEIII site as described below.
175

176 Firstly, the duplicated sequence consists of three different elements - two of the elements invariably co-
177 duplicated, with the third element being variable. The two elements co-duplicated are the eight base
178 RBEIII core motif (5'-ACTGCTGA-3') which binds the RBF-2 factor, and a down-stream seven base
179 motif (5'-TGACACA-3') that forms a binding site for a heterodimer of c-Jun and ATF (*SI Appendix*, Fig.
180 S1). Notably, the 3'-TGA residues of the RBEIII core sequence overlap with the binding site of the c-
181 Jun: ATF heterodimer. Thus, a part of the duplicated sequence (5'-**ACTGCTGAC**ACA-3', the bolded
182 sequence binds RBF-2, the underlined sequence is expected to bind c-Jun, and the italicized sequence
183 ATF) appears to mediate the binding of a complex of transcription factors consisting of RBF-2, c-Jun,
184 and ATF. Notably, the entire sequence consisting of the eight base core RBF-2 binding motif,
185 overlapping c-Jun, and ATF binding motif is strictly conserved in the RBEIII motif duplication. Secondly,
186 in addition to the duplication of the 5'-ACTGCTGACACA-3' sequence, additional sequences are also
187 co-duplicated, forming the basis for further classification of the viral strains. In a canonical HIV-1C LTR,
188 the RBEIII motif is flanked by an upstream TCF-1 α /LEF-1 site (5'-TACAAA/GG/A-3') and a downstream
189 NF- κ B element (5'-GGGACTTTCC-3'). When the RBEIII motif is duplicated, one of these two TFBS is
190 also co-duplicated, forming at least two sub-groups - viral strains containing a TCF-1 α /LEF-1 or NF- κ B
191 site co-duplication. Thirdly, in the variant LTRs, the two RBEIII sites are not contiguous but are invariably
192 separated by an intervening sequence that usually forms a binding site for NF- κ B or TCF-1 α /LEF-1, as
193 described above.
194

195 Based on the nature of the intervening sequence, the viral strains may be classified into two major
196 subgroups, one containing the presence of a canonical NF- κ B (5'-GGGACTTTCC-3', LRHR-HC), or
197 κ B-like motif (5'-GGGACTTTCA-3', which is denoted as 'h' κ B- motif in the present work, strains LRhR-
198 HC and LRhR-HHC) and the other a TCF-1 α /LEF-1 (5'-TACAAA/GG/A-3', LRLR-HC and LRLR-HHC)
199 or an incomplete TCF-1 α /LEF-1 binding sequence. A third sub-group may also be identified where the
200 intervening sequence does not seem to form a binding site for a defined host factor (LRXR-HC and
201 LRXR-HHC). Multisequence alignments of the variant viral promoters are presented (Fig. 2D-J). Of
202 note, the three NF- κ B motifs in LRHR-HC, LRhR-HC, and LRhR-HHC strains are not arranged in
203 tandem, thus, blurring the distinction between the viral enhancer consisting of the only NF- κ B motifs
204 and the upstream modulatory region.
205

206 In a phylogenetic analysis of 461 LTR sequences, compared with 33 reference sequences, all the viral
207 strains of the cohort grouped with HIV-1C ascertaining their genetic identity (Fig. 3). A single sequence
208 A105-9811 is the only exception that clustered with HIV-1K reference sequences. Notably, the various
209 viral sequences combined homogeneously without forming separate clusters based on the TFBS
210 variation.
211

212 **Longitudinal analysis of prognostic markers**

213

214 Increased transcriptional strength of the LTR may augment plasma viral load modulating various
215 prognostic markers and immune activation markers. To this end, we monitored a few prognostic
216 markers, such as the plasma viral load, CD4 cell count, and soluble CD14 levels, in the blood samples
217 at the baseline and at two or three follow-up time-points spaced six months apart. Unfortunately, this
218 objective could be fulfilled only partially given practical constraints. Following the primary screening and
219 typing of the viral promoters, we could recruit only 208 of the 455 study participants for the follow-up
220 analysis. Subsequently, with the implementation of the 'Test and Treat' policy in 2017 in India, several

221 participants enrolled for the follow-up were excluded from the study. Consequently, the longitudinal
222 analysis could be accomplished only with a small number of study participants, who did not prefer to
223 switch to ART.

224
225 The demographic features of the 208 study participants at the baseline are summarised (Table 1). Of
226 the 208 study participants, the percentage of female, male, and transgender are 55.3% (115/208),
227 43.8% (91/208), and 1.0% (2/208), respectively. The average (mean) age of the study participants was
228 34.4 ± 8.45 years (median = 33 years). For the subsequent analyses, all the viral strains were classified
229 into four categories based on the nature of the TFBS variations identified in the viral promoter – HHC
230 (the conventional LTRs containing the HHC κ B-binding sites), FHC (all the three κ B-binding sites are
231 genetically distinct in these LTRs), FHHC (the LTRs contain four κ B-binding sites), and RR (double-
232 RBEIII strains; the seven viral variant strains are pooled into a single category, given the limited number
233 of samples in individual groups).

234
235 We compared the levels of plasma viral load, CD4 cell count, and sCD14 among these four categories
236 at the baseline in a cross-sectional analysis. This analysis did not show a significant difference in any
237 of the three parameters among the four groups (Fig. 4, left panels). The median values of all the
238 parameters were comparable among the four groups. At M0, the median PVL values for the HHC, FHC,
239 FHHC and, RR groups were 12,609.0, 13,553.0, 10,440.0, and 6,321.0 copies/ml, respectively (Fig. 4,
240 left panel). The median values of CD4 cell count and sCD14 were also similarly comparable among the
241 four groups. We also compared the three clinical parameters between the baseline and 12 M time-point
242 for PVL and sCD14 and baseline, 6 M and 12 M for CD4 cell count (Fig. 4, right panels). All the three
243 parameters appeared to remain stable without a significant change between the baseline and 12 M.
244 The promoter configuration did not appear to make a significant difference for any of the three
245 parameters examined. For instance, the median PVL values at 12 M were 18,450.0, 52,539.0, 13,266.0,
246 and 6,090.0 copies/ml, for HHC, FHC, FHHC, and RR groups, respectively. The CD4 cell count
247 remained stable over the 12-month observation period among all four groups. The sCD14 levels
248 appeared to show an increasing trend among all four groups at the follow-up; these differences,
249 however, were not statistically significant. Of note, in complete-case analysis, a trend of low-level
250 plasma viral load was manifested for the RR variants compared to the three other groups. However,
251 the viral load increased between the time points among all the groups (*SI Appendix*, Fig. S3A).

252 253 **The coexistence of viral variants in natural infection**

254
255 Of the various promoter variant viral strains described in the present work, the emergence of LTRs
256 containing RBEIII duplication is relevant to HIV-1 latent reservoirs. In the context of HIV-1B, the RBEIII
257 site, as well as the AP-1 motif, are known to play a predominantly suppressive role, especially in the
258 absence of cellular activation (17–19). The relative proportion of reads representing single Vs. double
259 RBEIII motif-containing viral sequences in a co-infection may offer leads as per the biological
260 significance of RBEIII duplication in natural infection.

261
262 To this end, we identified a subset of four of 85 subjects of our cohort who showed the presence of a
263 co-infection of single- and double-RBEIII viral strains in Sanger sequencing. The clinical profile of the
264 four subjects (2079, 3767, 4084, and VFSJ020) is summarised (Table 2). We performed an NGS
265 analysis, using the MiSeq Illumina platform (*SI Appendix*, Fig. S4), of the whole blood genomic DNA
266 and plasma viral RNA of the four subjects at the baseline and two or three follow up time points.

267
268 The NGS data confirmed the coexistence of single- and double-RBEIII viral strains in all four subjects
269 in both the DNA and RNA compartments. Importantly, in two subjects (2079 and 4084), the single-
270 RBEIII strains represented a significantly larger proportion of reads in the plasma RNA compared to the
271 double-RBEIII strains (Fig. 5). Only in subject VFSJ020, the double-RBEIII reads dominated the single-
272 RBEIII reads at all the time points, whereas in subject 3767 a mixed profile was observed. The data
273 between the replicate samples are consistent with each other ascertaining the reproducibility of the
274 analysis. A broad level concurrence between the plasma RNA and genomic DNA was also noted.

275
276
277

278 **Discussion**

279

280 The key finding of the present work is the continuing evolution of the HIV-1C viral promoter. In 2004,
281 we reported the emergence of HIV-1C strains containing four copies of the NF- κ B binding motif in the
282 viral enhancer for the first time in India (14, 16). The 4- κ B viral strains dominated the canonical viral
283 strains containing three copies of NF- κ B motifs in natural infection and under all experimental
284 conditions, alluding to the additional copy of NF- κ B motif conferring replication advantage (6).
285 Subsequently, the 4- κ B viral strains were also detected in Brazil and several African countries,
286 suggesting that the phenomenon is not specific to a single country (15). Although our initial focus was
287 limited to the NF- κ B motif duplication and its impact on viral gene expression, we also observed the
288 duplication of other TF binding motifs, including RBEIII and AP-1, though at a lower frequency (6, 16).

289

290 **Sequence motif duplication in HIV-1C differs from that of other HIV-1 families**

291

292 Gene duplication accompanied by sequence variation played a crucial role in the acquisition of novel
293 properties leading to the evolutionary success of organisms (20). In viruses, the duplication of
294 biologically important sequence motifs may have provided the same survival advantage as gene
295 duplication has done in higher organisms (18). The significance of sequence motif duplication in the
296 coding sequences of HIV-1 has attracted more attention conventionally compared to that of regulatory
297 sequences (21–23). Further, numerous publications have reported the deletion or duplication of
298 different regions in the enhancer, core promoter, and modulatory regions that removed or added copies
299 of TFBS (19, 24–26). However, such sequence modifications have been sporadic, typically limited to
300 the individual or a small number of viral strains and cannot be generalized.

301

302 One notable exception to this observation is the occurrence of the MFNLP, which broadly represents
303 the duplication of the RBEIII motif in the viral modulatory region with the concomitant co-duplication of
304 the flanking sequences for other host factors. RBEIII motif duplication in HIV-1B was found in
305 approximately 38% of primary viral isolates (8). Importantly, RBEIII motif duplication in HIV-1B is
306 believed to ensure the presence of a binding site for RBF-2 when the original copy becomes non-
307 functional due to mutations (27, 28).

308

309 RBEIII motif duplication in HIV-1C differs from that of HIV-1B in two crucial qualities. First, the creation
310 of an additional RBEIII motif is not associated with the inactivation of the original motif. In other words,
311 nearly all the double-RBEIII viral strains in our cohort of HIV-1C contained two copies of the intact motif
312 without any mutations in the core sequence. Preliminary leads from our laboratory confirm the functional
313 activity of both the motifs in such LTRs. Of note, the participants of the present study are all reportedly
314 ART-naïve by self-declaration. We cannot rule out the possibility of ART exposure in HIV-1C, leading
315 to the inactivation of the original RBEIII motif necessitating the need to create a second and functional
316 RBEIII motif in the promoter. Second, the co-duplication of the RBEIII and NF- κ B motifs is unique to
317 HIV-1C, a property not seen in any other HIV-1 genetic subtype. Thus, HIV-1C appears to exploit the
318 strategy of sequence motif duplication differently compared to other viral subtypes.

319

320 Importantly, the addition of more copies of NF- κ B to the viral promoter may be beneficial by enhancing
321 the transcriptional strength of the LTR. However, a stronger LTR can be detrimental to maintaining
322 stable latency. HIV-1C appears to have found two different solutions to the paradox of gene expression
323 regulation – limiting the copy number of the NF- κ B motifs to three and duplicating the RBEIII motif.

324

325 **Limiting the number of NF- κ B motif copy number in the viral enhancer**

326

327 Three viral strains, LR-HHC, LR-FHHC, and LR-FHC, lack RBEIII duplication. The prevalence of the
328 LR-FHHC viral strains was only 2% (13 of 607 primary viral isolates) in a southern Indian cohort when
329 discovered during 2000-2003 for the first time (6). The prevalence of these strains increased to
330 approximately 25% (39/159) during 2010-2011, evaluated at four different clinical sites of India,
331 suggesting replication success of 4- κ B viral strains at the population level (6). However, in the present
332 study, the prevalence of the LR-FHHC viral strains dropped to 5.9% (27 of 455) during 2017-2019.
333 Notably, a new variant viral strain LR-FHC representing the second-largest proportion among the
334 emerging variants with 8.8% (40 of 455) was identified here for the first time. Given the reduction in the
335 prevalence of LR-FHHC strains and the concomitant appearance of the LR-FHC strains, it is possible
336 that the former is being replaced by the latter.

337

338 This observation leads to three logical conclusions. First, LR-FHHC strains, given the stronger
339 transactivation properties of the promoter, may lack replication competence over a sustained period
340 explaining the transient nature of their prevalence in the population. Second, the 4-κB viral strains must
341 relinquish one κB-motif to regain the 3-κB formulation of the enhancer to down modulate the
342 transcriptional strength of the viral promoter. The LR-FHHC strains relinquished one of the two H-κB
343 sites to this end to become LR-FHC. Three, both the canonical LR-HHC and the variant LR-FHC strains
344 contain the same number of NF-κB motifs in the enhancer. However, the three κB-motifs of the FHC-
345 LTR are genetically variable. We propose that the LR-FHC-LTR is likely to be responsive to a broader
346 range of cellular activation signals compared to the LR-HHC-LTR, given the NF-κB motif variation. Thus,
347 by deleting one H-κB site from the LR-FHHC-LTR, HIV-1C appears to have down-modulated
348 transcriptional strength of the viral promoter on the one hand but retained the broader reception
349 potential to cellular signals on the other hand. If the LR-FHC viral strains enjoy a replication advantage
350 at the population level, they are expected to replace the canonical HHC strains in the coming years.

351

352 **Is the RBEIII motif duplication to impose avid latency of a stronger viral promoter?**

353

354 Seven different variant LTRs identified in this work contain a second copy of the RBEIII motif added by
355 sequence-motif duplication (Fig.1 and 2D-J). Unlike in HIV-1B where a new RBEIII site is created as a
356 compensatory mechanism when the original copy is mutated (8), in HIV-1C, both the RBEIII motifs are,
357 in contrast, intact without a mutation in the core motif (5'-ACTGCTGA-3'). Thus, RBEIII duplication in
358 HIV-1C appears to confer a novel function or an enhanced phenotype of the existing function but not
359 compensating for a loss of function.

360

361 A second quality of the RBEIII motif duplication in HIV-1C is also relevant, especially for HIV latency.
362 While RBEIII motif duplication is common to all the HIV-1 genetic subtypes (*SI Appendix*, Fig. S2 and
363 see 7, 10, 11, 29), one significant distinction unique to HIV-1C is the co-duplication of the NF-κB motif,
364 not seen in other subtypes. One variant LTR, LRhR-HHC, contains a total of four NF-κB motifs like the
365 old FHHC-LTR. However, the variant LTR contains two RBEIII sites, unlike the FHHC-LTR that has
366 only one (Fig. 1). Additionally, the duplicated κB-motif of LRhR-HHC is genetically distinct (5'-
367 GGGACTTTCA-3') from the other three types (C-, H-, and F-κB) described above. The Single
368 Nucleotide Mutation Model predicted the 5'-GGGACTTTCA-3' motif to bind the p50 homodimer with
369 reduced affinity compared to the consensus NF-κB motif 5'-GGGACTTTCC-3'. This binding prediction
370 was supported by the bimolecular dsDNA microarray analysis (30). Lastly, the duplicated κB-motif of
371 LRhR-HHC-LTR is separated from the viral enhancer by one RBEIII motif thus, obliterating the
372 distinction between the viral modulatory and enhancer elements. The biological significance of creating
373 one of each RBEIII and NF-κB motifs in the LRhR-HHC-LTR is of interest.

374

375 In the absence of cell activation, the RBEIII motif functions predominantly as a repressive element by
376 recruiting RBF-2 comprising three different cell factors, including TFII-I (27). While TFII-I can activate
377 several cellular genes, it can also suppress gene expression from several other cellular promoters,
378 including *c-fos* (31). Thus, the presence of two copies of the RBEIII motif in the LTR may have a
379 profound impact on viral latency, probably by stabilizing the latency phase. The NF-κB binding motifs,
380 in contrast, play a predominantly positive role in enhancing transcription from the LTR, under the
381 conditions of cell activation. Thus, a higher copy number of NF-κB (4 copies Vs. 3) in the promoter may
382 offset the negative impact of the RBEIII motifs, especially when the provirus is induced out of latency.
383 Of note, unlike the variant LRhR-HHC, a different variant strain LRhR-HC contains one less NF-κB motif
384 (two RBEIII but only three NF-κB motifs). Preliminary results from our laboratory show that the LRhR-
385 HC-LTR requires a profoundly stronger activation signal, compared to LRhR-HHC-LTR or the canonical
386 LR-HHC-LTR, for latency reversal in Jurkat cells or primary CD4 cells (Bhange D et al, unpublished
387 observations).

388

389 A different variant promoter pair (LRLR-HHC and LRLR-HC) is also of interest in this respect. This pair
390 also contains an RBEIII motif duplication where the motifs are accompanied by the co-duplication of the
391 TCF-1α/LEF-1 motif, not the NF-κB site. One member of the pair contains three NF-κB motifs (LRLR-
392 HHC) while the other only two (LRLR-HC). This variant promoter pair may have similar gene expression
393 properties as that of the LRhR-HHC and LRhR-HC LTRs if the additional copy of TCF-1α/LEF-1 is a
394 functional equivalent of the NF-κB motif.

395

396

397

398 **The implication of promoter variation on HIV-1 pathogenesis and evolution**

399

400 Since a single promoter regulates the expression of all the HIV-1 proteins and controls latency, a
401 profound variation in the TFBS composition is expected to have a significant impact on the various
402 properties of the virus, including latency, viral load, disease progression, and viral evolution. The
403 evolution of regulatory elements may play a role as essential or even more important than that of coding
404 sequences (20). However, little attention was focused on the evolution of the regulatory elements in
405 HIV-1, unlike the protein-coding regions (32).

406

407 In our study, the cross-sectional and longitudinal analyses did not find a statistically significant
408 difference in the levels of any of the prognostic markers across the promoter variants categorized into
409 four groups, HHC, FHC, FHHC, and RR (Fig. 4). A significant difference in PVL and CD4 cell count was
410 found in a previous study from our laboratory when a cohort of eighty patients was divided into HHC
411 and FHHC groups (6). The present study did not find such differences, except for the RR group
412 manifesting a trend in the complete-case analysis, which was not statistically significant. The analytical
413 power of the present study was profoundly compromised given the loss of available samples due to the
414 implementation of the test-and-treat policy.

415

416 Commensurate with our findings, earlier studies of the RBEIII motif duplication in HIV-1B also failed to
417 see an association between the LTR profile and clinical or transcriptional phenotypes in a cross-
418 sectional cohort (8). A different study could not see a correlation between the RBEIII motif duplication
419 and syncytium-inducing property of envelope and, thus, disease progression (33). Likewise, Koken S.,
420 et al. demonstrated the domination of a viral strain containing a single copy of the RBEIII motif over a
421 counterpart containing two copies of the TFBS in 28 days in cell culture. However, no such differences
422 were observed in patients alluding to an association between the RBEIII copy-number and disease
423 progression (33).

424

425 The coexistence of viral strains could be a likely explanation for the absence of association between
426 LTR variant forms and prognostic markers in our cohort. Deep sequencing of the samples identified the
427 presence of a co-infection in all four subjects in our study. At the current time, it appears that the double-
428 RBEIII viral strains appear only as a co-infection along with the single-RBEIII strains. In three of the four
429 subjects, single-RBEIII viral strains seem to dominate the double-RBEIII variants in both the genomic
430 DNA and RNA compartments and at most of the follow-up time-points. If RBEIII duplication indeed
431 manifests a suppressive effect on viral gene expression, a distinct association between the duplication
432 and the prognostic markers may become evident in a mono-infection, but not in a coinfection. The
433 dominant influence of the RBEIII motif duplication on viral gene expression has been confirmed using
434 panels of engineered viral clones (Bhange D et al, unpublished observations).

435

436 In summary, our work records the emergence of several promoter variant viral strains in HIV-1C of India
437 over recent years. Sequences representing the variant viral forms are also found in the sequence
438 databases derived from different global regions where HIV-1C is predominant. Sequence motif
439 duplication creates additional copies of TFBS that play a crucial role in regulating HIV latency and even
440 blurs the distinction between the viral enhancer and modulatory regions. Given that the RBEIII and AP-
441 1 sites play a significant role in regulating latency (19 and 20), the influence of RBEIII site duplication,
442 especially when accompanied by the co-duplication of NF- κ B motifs, needs experimental evaluation.
443 Consistent monitoring will be necessary to understand which variant viral strains will survive to establish
444 spreading epidemics in coming years. Detailed investigations are warranted to evaluate the impact of
445 the TFBS profile differences on HIV-1 latency and latent reservoir properties. ART administration may
446 have a profound impact on the promoter variations described here by exacerbating such sequence
447 duplications.

448

449

450 **Materials and methods**

451

452 **Study participants and samples**

453

454 Participants were recruited at four different sites in India for primary screening (PS) and longitudinal
455 study (LS)- All India Institute of Medical Sciences (AIIMS), New Delhi (PS=107, LS=73); National AIDS
456 Research Institute (NARI), Pune (PS=61, LS=38); St. John's Medical Hospital, Bangalore (PS=116,
457 LS=60); and Y. R. Gaitonde Centre for AIDS Research and Education (YRG CARE), Chennai (PS=171,

458 LS=37). Subjects above 18 years of age with a documented evidence of serological positive test for
459 HIV-1 were recruited to the study. All the study subjects were ART naïve at baseline as per self-
460 reporting.

461 462 **Ethics statement**

463
464 Written informed consent was obtained from all study participants, following specific institutional review
465 board-approved protocols. Ethical approval for the study was granted by the Institutional Review Board
466 of each clinical site. All the clinical sites screened the potential subjects, counseled, recruited the study
467 participants, and maintained the clinical cohorts for the present study. The Human Ethics and Biosafety
468 Committee of Jawaharlal Nehru Centre for Advanced Scientific Research (JNCASR), Bangalore,
469 reviewed the proposal and approved the study.

470 471 **Primary screening: LTR amplification and molecular typing of the viral promoter**

472
473 For the molecular typing of the viral promoter, 15 ml of peripheral blood were collected from every
474 participant at one time. Genomic DNA was extracted from 200 µl of the whole blood, and the U3 region
475 of LTR was amplified using a nested-PCR strategy (details in *SI Appendix*). The amplified LTR
476 sequences were analyzed using Sanger Dideoxy sequencing and were subjected to further quality
477 control by multiple sequence alignment and phylogenetic analysis with the in-house laboratory
478 sequence database using the ClustalW algorithm of BioEdit sequence alignment editor and MEGA6.0
479 software, respectively. Different viral strains were categorized by analyzing the LTR spanning
480 modulatory and enhancer region from the LEF motif up to the Sp1III motif.

481 482 **Phylogenetic analysis**

483
484 The phylogenetic analysis of the HIV-1C LTR variants derived from 461 patient samples was performed
485 using MEGA7.0 software. The analysis was performed with 1,000 bootstrap values. The percentage of
486 clustering is shown using the Maximum Composite Likelihood method. The values presented represent
487 the number of base substitutions per site. All the positions containing gaps and missing data were
488 eliminated. A total of 412 positions were included in the final dataset.

489 490 **The follow-up clinical procedures**

491
492 Following successful characterization of the viral promoter at JNCASR, the clinical sites were advised
493 to recruit specific study subjects without disclosing the nature of the viral LTR. The clinical sites were,
494 thus, blinded to the identity of the viral LTR. All the clinical procedures were performed at the clinical
495 sites using the same protocols and kits as described in *SI Appendix*. The analysis of CD4 cell count,
496 plasma viral load, and soluble CD14 was performed at longitudinal time-points (details in *SI Appendix*).

497 498 **RNA isolation and RT-PCR for the next-generation sequencing**

499
500 RNA was extracted from 1 ml of the stored plasma samples, and the complementary DNA (cDNA) was
501 synthesized using HIV-specific primers listed in the *SI Appendix* Table S1 (details in *SI Appendix*). The
502 cDNA was used for the amplification of LTR.

503 504 **The next-generation sequencing**

505
506 The PCR products containing the RBEIII motif duplication were subjected to the NGS analysis using
507 the Miseq Illumina platform. Each sample was amplified in duplicates using primers containing a unique
508 8 bp barcode sequence specific for each sample. The amplification of the U3 region (~300-350 bp)
509 using genomic DNA or cDNA prepared from plasma RNA was performed the same way as described
510 for the primary screening except that the primers contained a unique sequence barcode at the 5'-end
511 as listed in *SI Appendix*, Table S1 and 2. The concentration of the purified PCR product was determined
512 using the Qubit™ dsDNA BR assay kit, and all the samples were pooled at an equal concentration and
513 were processed further. The final sample included a pulsed LTR amplicon of Indie-C1, a reference HIV-
514 1C molecular clone, as internal quality control for sequencing. The pooled sample was further
515 processed for Illumina MiSeq sequencing (details in *SI Appendix*).

516

517 Data analysis was performed using a pipeline as depicted (*SI Appendix*, Fig. S4) and the details are
518 mentioned in *SI Appendix*. Following the analysis procedure, the percentage prevalence of single and
519 double RBEIII variants was obtained (Fig. 5).

520

521 **Statistical analysis**

522

523 The data were analyzed using GraphPad prism 9, except the sequences used for phylogeny
524 determination. P values of 0.05 or less were considered statistically significant. A nonparametric,
525 Kruskal–Wallis (for multiple comparisons) test was applied to evaluate statistical significance in the case
526 of a cross-sectional analysis of plasma viral load (Fig. 4A and *SI Appendix*, Fig. S3A). One-way ANOVA
527 was used to evaluate statistical significance for cross-sectional analysis of CD4 cell count and sCD14
528 levels (Fig. 4B, C and *SI Appendix*, Fig. S3B, C). Two-way ANOVA was used to evaluate statistical
529 significance in the case of longitudinal analysis of all the three parameters, plasma viral load, CD4 cell
530 count, and sCD14. (Fig. 4, and *SI Appendix*, Fig. S3).

531

532 **Data availability statement**

533

534 All the sequences reported in this paper are available from the GenBank database under accession
535 nos. MN840242.1- MN840356.1, MT847032 - MT847207, MT593868 - MT594037. The raw data files
536 for Illumina MiSeq are available under accession no. PRJNA720640.

537

538

539 **Acknowledgments and funding sources**

540

541 We thank Dr. Kushgra Bansal for the discussions during the NGS data analysis. We thank the study
542 participants for their participation and cooperation during the complete duration of the clinical study.
543 This work was supported by the Department of Biotechnology, Ministry of Science and Technology,
544 Government of India (Sanction order no. BT/PR7359/MED/29/651/2012).

545

546

547 **Abbreviations**

548

549 TFBS: Transcription factor binding site, HIV-1C: HIV-1 subtype C, LTR: Long terminal repeats, C-LTR:
550 HIV-1C LTR, NGS: Next-generation sequencing, RT-PCR: Reverse transcription polymerase chain
551 reaction, ART: Antiretroviral therapy.

552

553

554 **References**

555

- 556 1. J. Yamaguchi, *et al.*, Brief Report: Complete Genome Sequence of CG-0018a-01 Establishes HIV-1
557 Subtype L. *J. Acquir. Immune Defic. Syndr.* **83**, 319–322 (2020).
- 558 2. D. M. Tebit, E. J. Arts, Tracking a century of global expansion and evolution of HIV to drive understanding
559 and to combat disease. *Lancet Infect. Dis.* **11**, 45–56 (2011).
- 560 3. D. Locateli, *et al.*, Molecular Epidemiology of HIV-1 in Santa Catarina State Confirms Increases of Subtype
561 C in Southern Brazil. *J. Med. Virol.* **79**, 52–55 (2007).
- 562 4. J. Hemelaar, *et al.*, Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review,
563 global survey, and trend analysis. *Lancet Infect. Dis.* **19**, 143–155 (2019).
- 564 5. E. Ramírez de Arellano, V. Soriano, J. Alcamil, A. Holguín, New findings on transcription regulation across
565 different HIV-1 subtypes. *AIDS Rev.* **8**, 9–16.
- 566 6. M. Bachu, *et al.*, Multiple NF- κ B Sites in HIV-1 Subtype C Long Terminal Repeat Confer Superior
567 Magnitude of Transcription and Thereby the Enhanced Viral Predominance. *J. Biol. Chem.* **287**, 44714–
568 44735 (2012).
- 569 7. M. Ait-Khaled, J. E. McLaughlin, M. A. Johnson, V. C. Emery, Distinct HIV-1 long terminal repeat
570 quasispecies present in nervous tissues compared to that in lung, blood and lymphoid tissues of an AIDS
571 patient. *AIDS* **9**, 675–684 (1995).
- 572 8. M. C. Estable, *et al.*, Human immunodeficiency virus type 1 long terminal repeat variants from 42 patients
573 representing all stages of infection display a wide range of sequence polymorphism and transcription
574 activity. *J. Virol.* **70**, 4053–62 (1996).
- 575 9. S. E. C. Koken, J. L. B. Van Wamel, J. Goudsmit, B. Berkhout, J. L. M. C. Geelent, Natural variants of the
576 HIV-1 long terminal repeat: Analysis of promoters with duplicated DNA regulatory motifs. *Virology* **191**,
577 968–972 (1992).

- 578 10. L. Zhang, *et al.*, Identification of a replication-competent pathogenic human immunodeficiency virus type 1
579 with a duplication in the TCF-1alpha region but lacking NF-kappaB binding sites. *J. Virol.* **71**, 1651–1656
580 (1997).
- 581 11. M. C. Estable, B. Bell, M. Hirst, I. Sadowski, Naturally Occurring Human Immunodeficiency Virus Type 1
582 Long Terminal Repeats Have a Frequently Observed Duplication That Binds RBF-2 and Represses
583 Transcription. *72*, 6465–6474 (1998).
- 584 12. M. C. Estable, In search of a function for the most frequent naturally-occurring length polymorphism
585 (MFNLP) of the HIV-1 LTR: retaining functional coupling, of Nef and RBF-2, at RBEIII? *Int. J. Biol. Sci.* **3**,
586 318–27 (2007).
- 587 13. S. Singh, *et al.*, Intra-Clade C signature polymorphisms in HIV-1 LTR region: The Indian and African
588 lookout. *Virus Res.* **297**, 198370 (2021).
- 589 14. N. B. Siddappa, *et al.*, Identification of subtype C human immunodeficiency virus type 1 by subtype-specific
590 PCR and its use in the characterization of viruses circulating in the southern parts of India. *J. Clin. Microbiol.*
591 **42**, 2742–2751 (2004).
- 592 15. J. Boullosa, *et al.*, Genetic Diversity in HIV-1 Subtype C LTR from Brazil and Mozambique Generates New
593 Transcription Factor-Binding Sites. *Viruses* **6**, 2495–2504 (2014).
- 594 16. M. Bachu, *et al.*, Sequence Insertions in the HIV Type 1 Subtype C Viral Promoter Predominantly Generate
595 an Additional NF-κB Binding Site. *AIDS Res. Hum. Retroviruses* **28**, 1362–8 (2012).
- 596 17. W. Bernhard, K. Barreto, S. Raitatha, I. Sadowski, An upstream YY1 binding site on the HIV-1 LTR
597 contributes to latent infection. *PLoS One* **8**, e77052 (2013).
- 598 18. K. A. Kropp, A. Angulo, P. Ghazal, Viral enhancer mimicry of host innate-immune promoters. *PLoS Pathog.*
599 **10**, e1003804 (2014).
- 600 19. M. H. Naghavi, S. Schwartz, A. Sonnerborg, A. Vahlne, Long terminal repeat promoter/enhancer activity
601 of different subtypes of HIV type 1. *AIDS Res Hum Retroviruses* **15**, 1293–1303 (1999).
- 602 20. S. B. Carroll, Evolution at two levels: on genes and form. *PLoS Biol.* **3**, e245 (2005).
- 603 21. S. Guglietta, G. Pantaleo, C. Graziosi, Long sequence duplications, repeats, and palindromes in HIV-1
604 gp120: length variation in V4 as the product of misalignment mechanism. *Virology* **399**, 167–175 (2010).
- 605 22. S. Sharma, *et al.*, PTAP motif duplication in the p6 Gag protein confers a replication advantage on HIV-1
606 subtype C. *J. Biol. Chem.* **293**, 11687–11708 (2018).
- 607 23. N. Marlowe, *et al.*, Analysis of insertions and deletions in the gag p6 region of diverse HIV type 1 strains.
608 *AIDS Res. Hum. Retroviruses* **20**, 1119–25 (2004).
- 609 24. M. A. Montano, *et al.*, Divergent transcriptional regulation among expanding human immunodeficiency
610 virus type 1 subtypes. *J. Virol.* **71**, 8657–65 (1997).
- 611 25. J. Leonard, *et al.*, The NF-kappa B binding sites in the human immunodeficiency virus type 1 long terminal
612 repeat are not required for virus infectivity. *J. Virol.* **63**, 4919–24 (1989).
- 613 26. N. L. Michael, L. D'Arcy, P. K. Ehrenberg, R. R. Redfield, Naturally occurring genotypes of the human
614 immunodeficiency virus type 1 long terminal repeat display a wide range of basal and Tat-induced
615 transcriptional activities. *J. Virol.* **68**, 3163–74 (1994).
- 616 27. J. Chen, T. Malcolm, M. C. Estable, R. G. Roeder, I. Sadowski, TFII-I regulates induction of chromosomally
617 integrated human immunodeficiency virus type 1 long terminal repeat in cooperation with USF. *J. Virol.* **79**,
618 4396–406 (2005).
- 619 28. I. Sadowski, D. A. Mitchell, TFII-I and USF (RBF-2) regulate Ras/MAPK-responsive HIV-1 transcription in
620 T cells. *Eur. J. Cancer* **41**, 2528–36 (2005).
- 621 29. V. R. Gómez-Román, *et al.*, nef/long terminal repeat quasispecies from HIV type 1-infected mexican
622 patients with different progression patterns and their pathogenesis in hu- PBL-SCID mice. *AIDS Res. Hum.*
623 *Retroviruses* **16**, 441–452 (2000).
- 624 30. W. Du, J. Gao, T. Wang, J. Wang, Single-nucleotide mutation matrix: A new model for predicting the NF-
625 κB DNA binding sites. *PLoS One* **9** (2014).
- 626 31. A. L. Roy, Biochemistry and biology of the inducible multifunctional transcription factor TFII-I: 10 years
627 later. *Gene* **492**, 32–41 (2012).
- 628 32. I. Maljkovic Berry, *et al.*, Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1)
629 pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. *J. Virol.* **81**, 10625–
630 35 (2007).
- 631 33. S. E. Koken, J. L. van Wamel, J. L. Geelen, B. Berkhout, Functional Analysis of the ACTGCTGA Sequence
632 Motif in the Human Immunodeficiency Virus Type-1 Long Terminal Repeat Promoter. *J Biomed Sci* **1**, 83–
633 92 (1994).
- 634

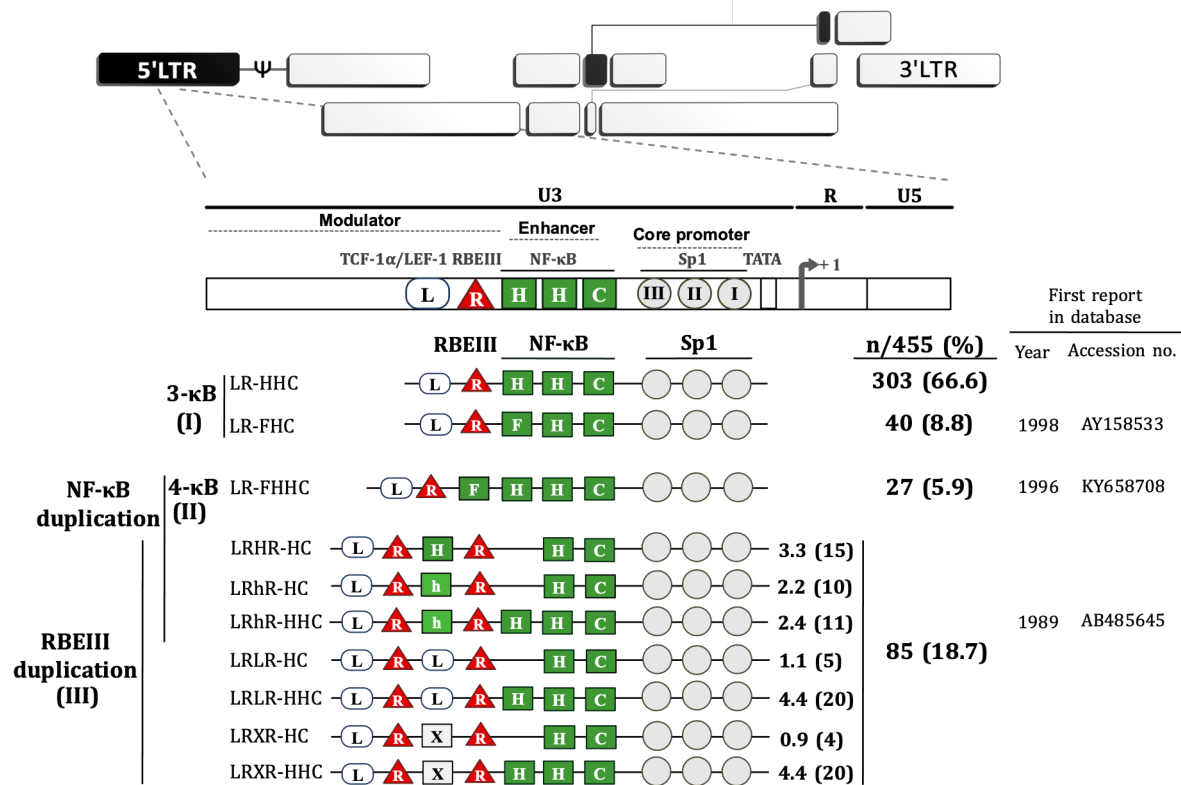


Fig. 1. The magnitude of TFBS variation in HIV-1C LTR. The upper panel represents genome organization of HIV-1 followed by the TFBS arrangement in the canonical HIV-1C LTR. Sp1 motifs are depicted as grey circles, RBEIII motifs (R) as red triangles, and TCF-1α/LEF-1 sites (L) as open rectangular boxes. The various types of NF-κB binding sites (H, C, F, and h) are depicted as green square boxes. The various HIV-1C viral strains are classified into three main categories based on the NF-κB and/or RBEIII motif duplication. (I) The 3-κB LTR viral strains. The canonical viral strains (LR-HHC) continue to remain the major category. The LR-FHC strains contain three different NF-κB motifs, and these strains represent a new variant identified for the first time. (II) The canonical 4-κB LTR viral strains. The LR-FHHC-LTR contains four tandem copies of NF-κB motif in the enhancer and its frequency appears to be dropping since the previous reports (Bachu M. et. Al., 2014). (III) The viral strains containing the RBEIII site duplication. The two RBEIII sites are separated by an interceding sequence that constitutes an additional copy of a κB-motif (H), κB-like motif (h), TCF-1α/LEF-1 motif (L) or sequence without a distinct pattern (X). The analysis is from 455 samples as we could not type 65 of 520 LTR sequences.

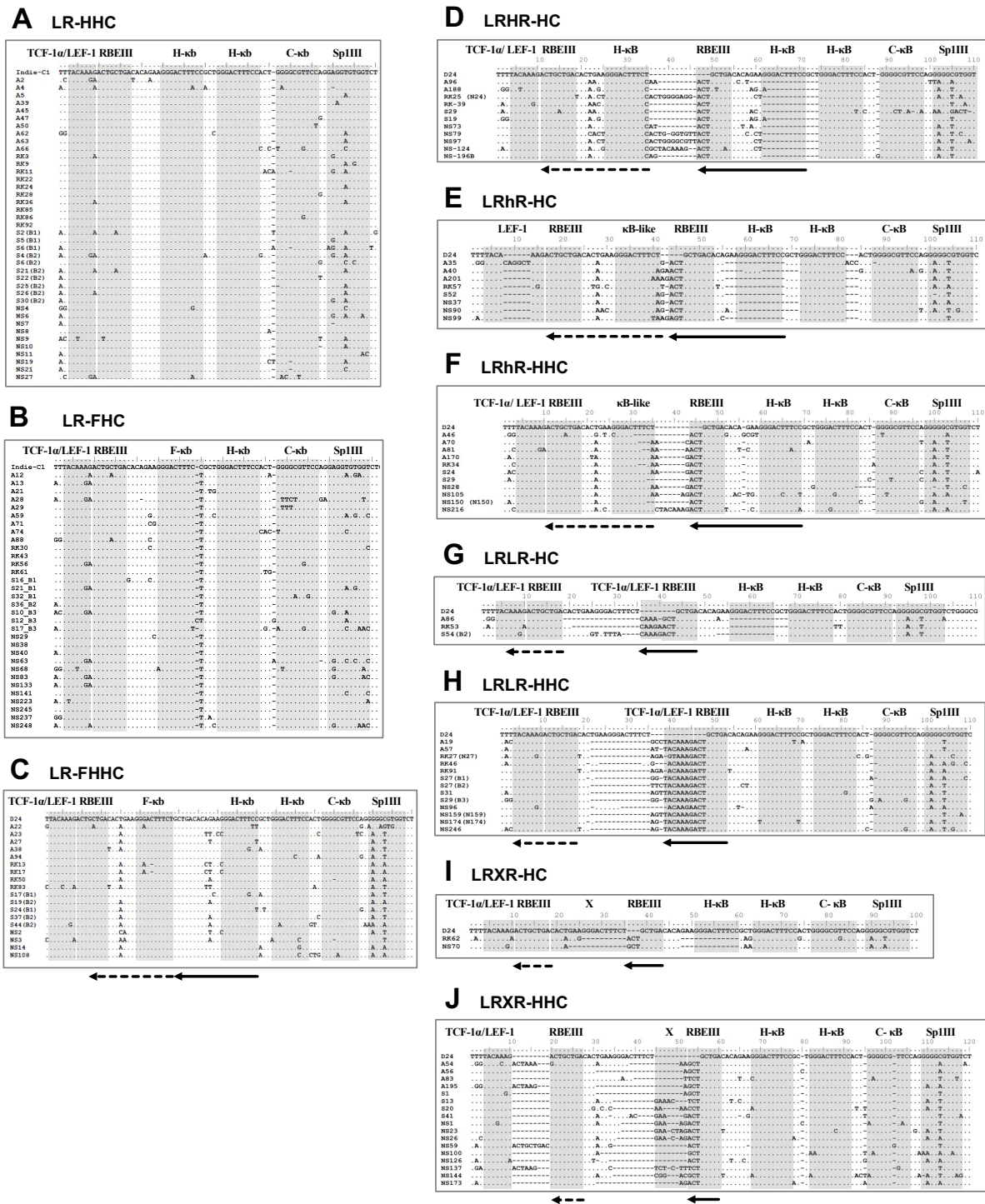


Fig. 2. Multiple sequence alignment of the LTR sequences of HIV-1C strains. The alignment contains a few representative sequences under each category. Patient identity is depicted on the left side of alignment. TFBS of significance are highlighted using grey shade boxes and labelled on the top. (A) The LR-HHC (the canonical 3-kB strains) sequences are aligned with the Indie C1 (AB023804.1) reference sequence. (B) The alignment of the LR-FHC variant viral sequences (C) The alignment of the LR-FHHC (4-kB variants) sequences with the D24 (EF469243.2) reference sequence. (D, E, F, G, H, I and J) Sequence alignment of seven types of double-RBEIII variant viral strains. The intervening sequences between the two RBEIII sites represent an H-kb (H), κB-like (h), hLEF (L), or un-typable (X) motif. The solid and dotted arrows represent original and duplicated sequence, respectively.



Fig. 3. Phylogenetic analysis of HIV-1C LTR variant sequences. A total of 461 viral sequences isolated from study participants are included in the analysis. The analysis also includes four HIV-1C reference sequences and three sequences representing each major genetic subtype of HIV-1 as described in materials and methods. All the reference sequences are represented in bold. Different LTR variant types are represented using different symbols as depicted. The analysis was performed with 1000 bootstrap and the percentage of clustering are shown. The tree is drawn to scale with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree and the scale is shown at the beginning of the tree. There were a total of 412 positions in the final dataset. Evolutionary analyses were performed using MEGA7.0 software.

Table 1. Characteristics of study participants in the longitudinal study at the time of recruitment

Parameters	
Total number of participants	208
Gender	
Female	55.3% (n=115)
Male	43.8% (n=91)
Transgender	1.0% (n=2)
Age	
Mean and SD	34.4±8.45
Median	33
ART status	Naïve

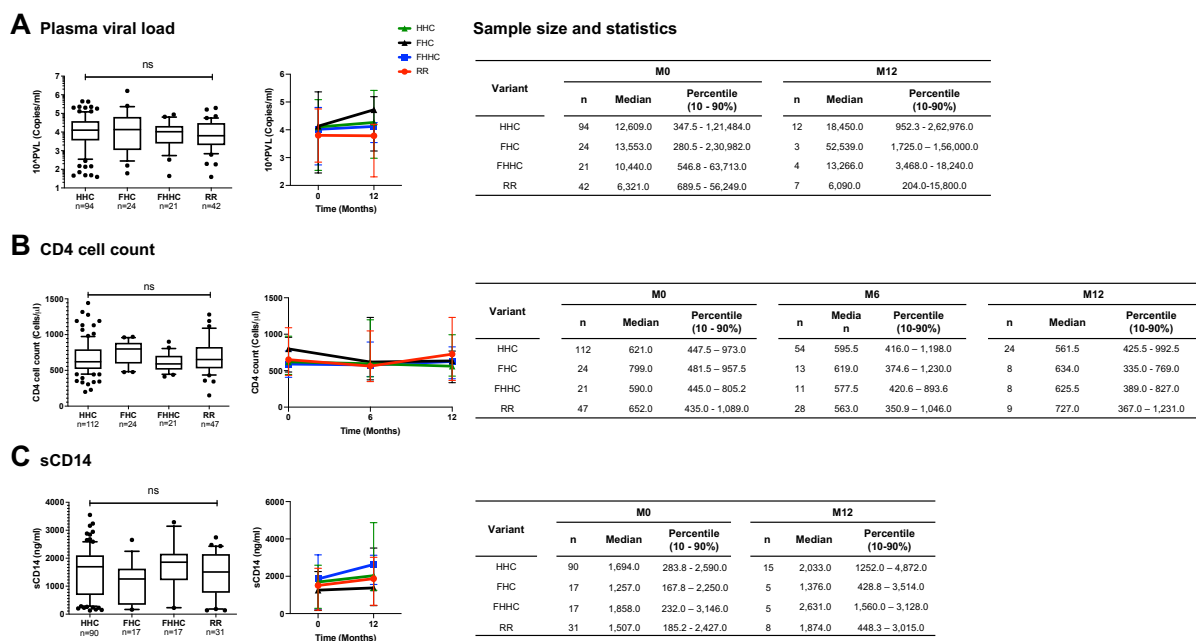


Fig. 4. Cross-sectional and longitudinal analysis of prognostic markers. Plasma viral load (A), CD4 cell count (B), and soluble CD14 (C) levels among the four study arms are presented at the baseline (left panels) and follow up points (right panels). The number of samples included under each evaluation are presented in the tables. Given the limited sample numbers, several groups were pooled under the double-RBEIII arm. A non-parametric test i.e. Kruskal-Wallis test was applied for the statistical analysis of the plasma viral load. One-way ANOVA analysis was applied to CD4 count and sCD14.

Table 2. The clinical profile of the four study subjects containing a duplication of the RBEIII motif

Subject ID	Age/ Gender	Enrollment date	Promoter variant	PVL (number of RNA copies/ml)		CD4 (cells/ul)			sCDs14		
				M0	M12	M0	M6	M12	M24	M0	M12
2079	32/F	01/06/2016	LR-HHC and LRLR-HHC	185	9,989	989	332	558	1,439	1,146.00	729.56
3767	40/F	29/06/2016	LR-HHC and LRhR-HHC	3,742	-	566	292	661	430	139.00	-
4084	38/F	07/05/2016	LR-HHC, LR- HC, and LRhR- HC	6,924	5,179	509	415	442	613	983.00	448.28
VFSJ020	38/F	22/09/2016	LR-HHC and LRhR-HHC	21,200	-	461	-	579	516	2,204.43	-

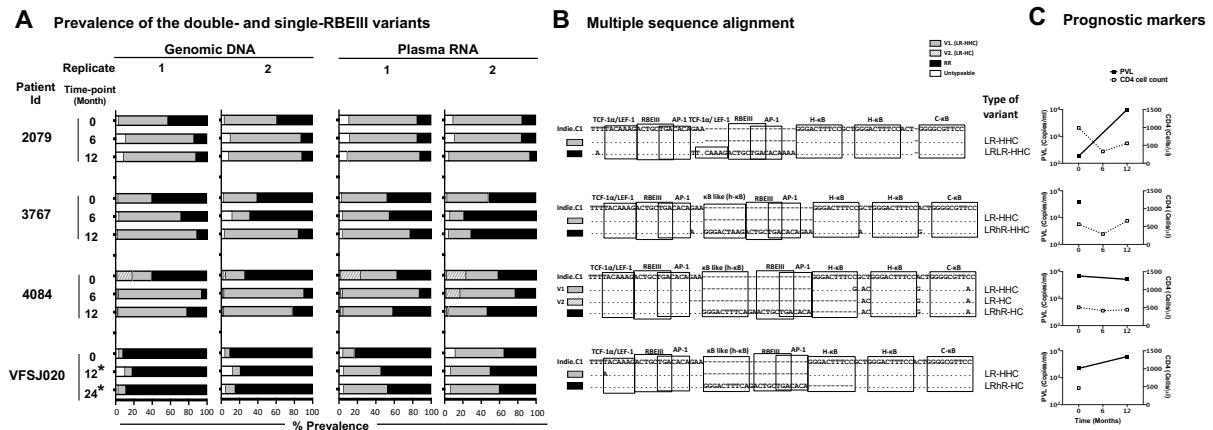


Fig. 5. The frequencies of single- and double-RBEIII variants in a subset of study participants. (A) Two independent rounds of analyses were performed (replicates 1 and 2) using both the whole blood genomic DNA and plasma viral RNA. The samples were collected at six-month intervals as shown. An asterisk (*) represents the samples collected post-ART. The dark, grey and hollow bars represent the percentage prevalence of double-RBEIII, single-RBEIII and minority/un-typable viral strains, respectively. (B) Multiple sequence alignment of single- and double-RBEIII viral variants in respective subjects. A few TFBS of relevance are marked using open square boxes. The viral variants are aligned with the Indie.C1 reference sequence, which was pulsed in the sequencing sample as an internal control. Dashes represent sequence deletion and dots sequence homology. (C) Prognostic markers, plasma viral load (PVL) and CD4 cell count are represented by filled box with a solid line and open box with a dotted line, respectively.