AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

RESEARCH ARTICLE

Transcription Start Site Heterogeneity and Preferential Packaging of Specific Full-Length RNA Species Are Conserved Features of Primate Lentiviruses

Jonathan M. O. Rawson,[a] Olga A. Nikolaitchik,[a] Saurabh Shakya,[a] Brandon F. Keele,[b] Vinay K. Pathak,[c] Wei-Shau Hu[a]

[a]Viral Recombination Section, HIV Dynamics and Replication Program, National Cancer Institute at Frederick, Frederick, Maryland, USA
[b]AIDS and Cancer Virus Program, Frederick National Laboratory, Frederick, Maryland, USA
[c]Viral Mutation Section, HIV Dynamics and Replication Program, National Cancer Institute at Frederick, Frederick, Maryland, USA

Jonathan M. O. Rawson and Olga A. Nikolaitchik contributed equally to this study. The order of the two first authors' names was determined by a coin toss.

ABSTRACT HIV-1 must package its RNA genome to generate infectious viruses. Recent studies have revealed that during genome packaging, HIV-1 not only excludes cellular mRNAs, but also distinguishes among full-length viral RNAs. Using NL4-3 and MAL molecular clones, multiple transcription start sites (TSS) were identified, which generate full-length RNAs that differ by only a few nucleotides at the 5′ end. However, HIV-1 selectively packages RNAs containing one guanosine (1G RNA) over RNAs with three guanosines (3G RNA) at the 5′ end. Thus, the 5′ context of HIV-1 full-length RNA can affect its function. To determine whether the regulation of genome packaging by TSS usage is unique to NL4-3 and MAL, we examined 15 primate lentiviruses including transmitted founder viruses of HIV-1, HIV-2, and several simian immunodeficiency viruses (SIVs). We found that all 15 viruses used multiple TSS to some extent. However, the level of TSS heterogeneity in infected cells varied greatly, even among closely related viruses belonging to the same subtype. Most viruses also exhibited selective packaging of specific full-length viral RNA species into particles. These findings demonstrate that TSS heterogeneity and selective packaging of certain full-length viral RNA species are conserved features of primate lentiviruses. In addition, an SIV strain closely related to the progenitor virus that gave rise to HIV-1 group M, the pandemic pathogen, exhibited TSS usage similar to some HIV-1 strains and preferentially packaged 1G RNA. These findings indicate that multiple TSS usage and selective packaging of a particular unspliced RNA species predate the emergence of HIV-1.

IMPORTANCE Unspliced HIV-1 RNA serves two important roles during viral replication: as the virion genome and as the template for translation of Gag/Gag-Pol. Previous studies of two HIV-1 molecular clones have concluded that the TSS usage affects unspliced HIV-1 RNA structures and functions. To investigate the evolutionary origin of this replication strategy, we determined TSS of HIV-1 RNA in infected cells and virions for 15 primate lentiviruses. All HIV-1 isolates examined, including several transmitted founder viruses, utilized multiple TSS and selected a particular RNA species for packaging. Furthermore, these features were observed in SIVs related to the progenitors of HIV-1, suggesting that these characteristics originated from the ancestral viruses. HIV-2, SIVs related to HIV-2, and other SIVs also exhibited multiple TSS and preferential packaging of specific unspliced RNA species, demonstrating that this replication strategy is broadly conserved across primate lentiviruses.

KEYWORDS RNA genome packaging, human immunodeficiency virus, lentiviruses, transcription start site, virus evolution

Editor Cyprian C. Rossetto, University of Nevada, Reno

Address correspondence to Wei-Shau Hu, Wei-Shau.Hu@nih.gov.

The authors declare no conflict of interest.

Received 5 April 2022
Accepted 5 June 2022
Published 23 June 2022

Upon infection, human immunodeficiency virus type-1 (HIV-1) reverse transcribes its RNA genome into DNA and integrates the DNA copy into host chromosomes, forming a provirus (1). Host machinery transcribes the provirus to generate viral RNAs, some of which are spliced and translated to express various viral proteins. Other viral RNAs remain unspliced and serve two important roles in viral replication: template for translation to produce the Gag/Gag-Pol polyproteins and packaging into nascent virions as the viral genome (1). Recent studies of HIV-1 molecular clones NL4-3 and MAL have demonstrated that multiple neighboring transcription start sites (TSS) are used during the transcription of HIV-1 RNA, resulting in RNAs with three, two, or one guanosines at the 5′ end (referred to as 3G, 2G, and 1G RNA, respectively) (2–7). Interestingly, while the most abundant unspliced HIV-1 RNA in cells is 3G RNA, 1G RNA is significantly enriched in virions (2–7). Therefore, HIV-1 distinguishes between ~9-kb RNAs that differ by only a couple of nucleotides (nts) and selects a minor RNA species during genome packaging.

HIV-1 has been classified into four groups (M, N, O, and P) that arose from independent cross-species transmission events (8). HIV-1 group M is primarily responsible for the HIV pandemic and has further been categorized into numerous subtypes and circulating recombinant forms (CRFs). At this time, HIV-1 TSS have only been characterized in NL4-3 and MAL, which belong to group M subtypes B and A, respectively. However, HIV-1 is quite diverse in the human population, and these two viruses might not fully represent the breadth of TSS usage among HIV-1 groups and subtypes. In addition, the evolutionary origins of HIV-1 TSS heterogeneity and selective packaging of 1G RNA have not been investigated. HIV-1 groups M and N resulted from cross-species transmission of simian immunodeficiency virus from chimpanzees (SIVcpz) into humans, whereas HIV-1 groups O and P resulted from cross-species transmission of SIV from gorillas (SIVgor) into humans (8–12). Furthermore, SIVgor also originated from SIVcpz (11). SIVcpz, in turn, is thought to have arisen from recombination of at least two other SIVs: SIV from red-capped mangabeys (SIVrcm) and SIV from mustached, mona, or greater spot-nosed monkeys (SIVmus/mon/gsn) (13, 14). TSS usage in these SIVs has not yet been examined.

In addition to the HIV-1/SIVcpz/SIVgor lineage, there are many other distinct lineages of SIVs in primates. Of these, HIV-2 is a known human pathogen and causes AIDS. However, HIV-2 is significantly less pathogenic, has a different genome structure, and replicates more slowly in cell culture than HIV-1 (15, 16). HIV-2 resulted from cross-species transmission of SIV from sooty mangabey monkeys (SIVsmm) into humans. Likewise, cross-species transmission of SIVsmm into macaques generated SIVmac (17, 18). Thus, HIV-2, SIVsmm, and SIVmac represent a group of closely related viruses that are distinct from HIV-1. Furthermore, there are several other SIV lineages including SIVagm from African green monkeys that are clearly distinct from HIV-1- or HIV-2-related viruses. In total, there are at least 45 known SIVs, with evidence of frequent cross-species transmission and recombination events (14, 19). It is currently unclear whether TSS heterogeneity and selective packaging of a particular RNA species are conserved features of these primate lentiviruses.

In this study, we have investigated the TSS usage of cellular and virion RNA in 16 primate lentiviruses. In addition to NL4-3, we examined six HIV-1 viruses including multiple clinical isolates and four SIVs including several that are closely related to the progenitor viruses that were zoonotically transmitted into humans. All of the HIV-1 and HIV-1-related viruses we examined have three guanosines at the TSS, and they used multiple neighboring nts to initiate transcription. Furthermore, all of them preferentially packaged 1G RNA. These findings indicate that multiple TSS and selective 1G RNA packaging are features that preexisted in SIV prior to cross-species transmission and the generation of HIV-1. Intriguingly, the proportions of the transcribed RNA species varied greatly across viruses, even among those belonging to the same subtype. We also examined TSS usage in two HIV-2 molecular clones and two SIVs closely related to HIV-2 (i.e., SIVmac and SIVsmm). We found that these viruses also use multiple TSS and, except for SIVsmm, selectively package specific RNA species. Taken together, our study shows that primate lentiviruses use heterogenous TSS and most of these viruses also distinguish among unspliced RNAs with minor differences at the 5′ end, indicating that these features are part of a conserved replication strategy.

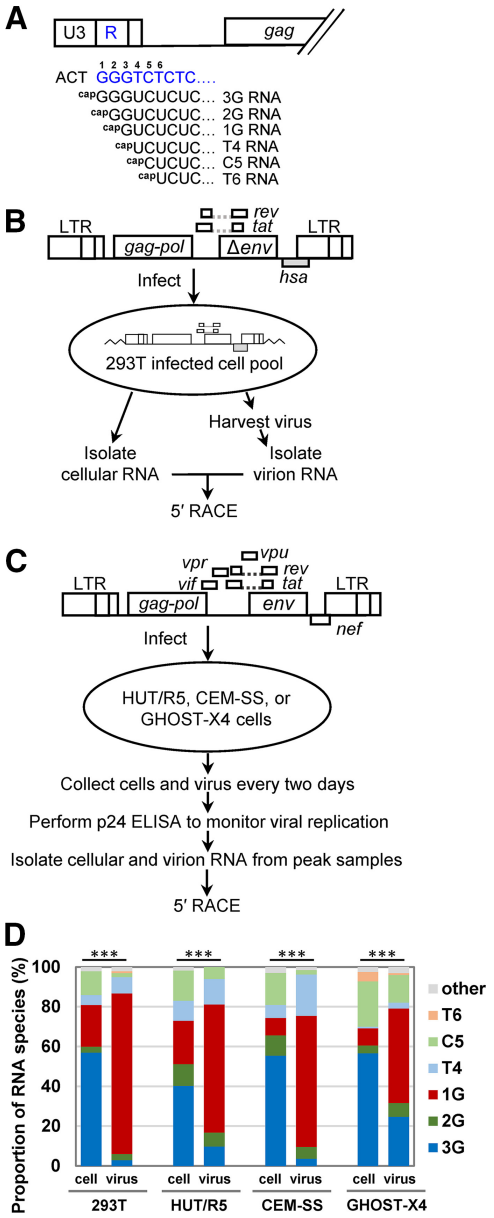

FIG 1 System used to study the TSS of HIV-1 RNA. (A) The sequences near the TSS of NL4-3 and RNA species generated from using several TSS. (B and C) Experimental workflow used to determine TSS in cells infected with and viruses produced by a modified construct H0 (B) and unmodified replication-competent HIV-1 (C). H0 expresses Gag/Gag-Pol, Tat, and Rev and can carry out a single round of virus replication when supplemented with Env. (D) Proportions of HIV-1 full-length RNA species in different cells infected with and viruses produced by NL4-3 as determined by 5′ RACE. The results are summarized from at least two independent experiments with sequencing the following total number of clones: 100 (293T cell), 99 (293T virus), 119 (HUT/R5 cell), 101 (HUT/R5 viruses), 137 (CEM-SS cell), 135 (CEM-SS viruses), 127 (GHOST-X4 cell), and 101 (GHOST-X4 virus). ***, $P < 0.001$, chi-square test.

RESULTS

HIV-1 NL4-3 exhibits TSS heterogeneity and preferentially packages 1G RNA in multiple cell types. Using modified HIV-1 constructs, we and others have previously shown that transcription of NL4-3 initiates at multiple neighboring sites, generating several RNA species (Fig. 1A) (2–4, 7). First, we examined whether NL4-3 TSS usage is affected by experimental conditions such as cell types from which viruses are produced or by modifications to the viral genome intended to limit the replication cycle (Fig. 1B and C). To recapitulate our previous studies, we used the same single-cycle vector called H0 (20), an NL4-3-derived construct that contains intact long terminal repeats (LTRs), 5′ untranscribed region (UTR), *gag-pol*, *tat*, and

rev, but has inactivating deletions in *vif*, *vpr*, *vpu*, and *env*. Additionally, this construct contains a heat stable antigen (*hsa*) marker gene in *nef* (Fig. 1B). To ensure capture of the authentic TSS distribution, we infected 293T cells with HIV-1 virions pseudotyped with vesicular stomatitis virus G protein (VSV G) (Fig. 1B) and used flow cytometry to detect HSA to determine the multiplicity of infection. H0 expresses functional Gag/GagPol; thus, H0-infected cells produce viral particles that are noninfectious due to the absence of Env. We isolated total RNA from infected cells and viral particles produced from these cells and analyzed the TSS using 5′ rapid amplification of cDNA ends (5′ RACE). Primers that anneal to the *gag* gene were used for cDNA synthesis so that only unspliced RNAs were analyzed. For both cellular and virion RNA, ~100 clones were sequenced per virus, combined across at least two independent experiments. We obtained results similar to those previously published (7); briefly, 3G RNA was the most abundant species in cells, whereas 1G RNA was enriched in virions (Fig. 1D).

To examine TSS usage in intact, replication-competent NL4-3 during spreading infection, we used two T cell lines, HUT/R5 and CEM-SS, as well as an engineered osteosarcoma cell line GHOST-X4 that expresses CD4 and coreceptor CXCR4 (Fig. 1C). After infection, cells and virus supernatants were collected every 2 days, and viral replication was monitored by p24 ELISA. Cell and virion RNAs were extracted from samples collected at the peak of virus production and analyzed using 5′ RACE. In all cases, we found that in the cytoplasm, 3G RNA was the most abundant species, although a significant proportion of 1G RNA was also detected. Additionally, we identified several other RNA species starting at TSS slightly downstream of the three guanosines at T4, C5, and T6 (Fig. 1A). Most transcripts with these downstream TSS had 5′ guanosines consistent with 5′ capping, indicating that they are not artifacts of 5′ RACE. These results agree with previous observations (7) and confirm the validity of our 5′ RACE method.

We next determined the TSS in NL4-3 virions produced from infected HUT/R5, CEM-SS, and GHOST-X4 cells. In all cases, NL4-3 packaged mostly 1G RNA into virions (Fig. 1D). Additionally, TSS distributions were significantly different between cells and virions in all four cell types ($P < 0.001$, chi-square test; Fig. 1D). Specifically, 1G RNA was enriched in virions, while 3G RNA was depleted ($P < 0.001$ in all cases, Fisher's exact test). The proportions of most other RNA species did not significantly differ between cells and virions, except for C5 RNA, which was significantly depleted in virions from 293T, HUT/R5, and CEM-SS cells ($P = 0.01$, 0.03, and < 0.001, respectively, Fisher's exact test), but not GHOST/X4 cells ($P = 0.09$, Fisher's exact test). These results show that the distribution of NL4-3 RNA species in cells and virions is largely consistent across cell types and between single-cycle vectors and intact replication-competent viruses. Because results from the H0 vector and replication-competent NL4-3 were similar, in our analyses below for other viruses we used either H0-like constructs to infect 293T cells or replication-competent viruses to infect T cells or GHOST cells. In all cases, infection was used instead of transfection to better recapitulate authentic TSS patterns. Furthermore, viral genomes were packaged by the appropriate authentic Gag/Gag-Pol proteins, and the LTRs, 5′ UTR, and *gag-pol* sequences were unaltered to reflect authentic transcription regulation and genome packaging.

TSS usage in HIV-1 group M subtype B viruses. Although widely used in research settings, NL4-3 is a lab-adapted chimera that may not reflect the properties of subtype B viruses from patients. To address this, we determined the TSS usage of three subtype B transmitted founder viruses: TRJO, CH058, and CH106 (21–24). Transmitted founder molecular clones represent the sequences of transmitted viruses, as determined by sequencing samples from acutely infected patients and mathematical modeling (22). Similar to NL4-3, TRJO, CH058, and CH106 also have three guanosines near the putative TSS and identical 6-nucleotide (nt) sequences immediately surrounding the guanosines (Fig. 2A). In HUT/R5 cells, NL4-3 and TRJO had similar TSS usage, producing mostly 3G RNA but also a significant proportion of 1G RNA (Fig. 2B). However, CH058 and CH106 exhibited a strikingly distinct pattern of TSS usage. These viruses generated mostly C5 (rather than 3G) RNA in cells, but also a significant proportion of 1G RNA. Compared to NL4-3 and TRJO, both CH058 and CH106 contain a 1-nt deletion between the CATA/TATA box (25) and the TSS (Fig. 2A). This difference might explain the altered TSS usage for these viruses, as the TSS is partly determined

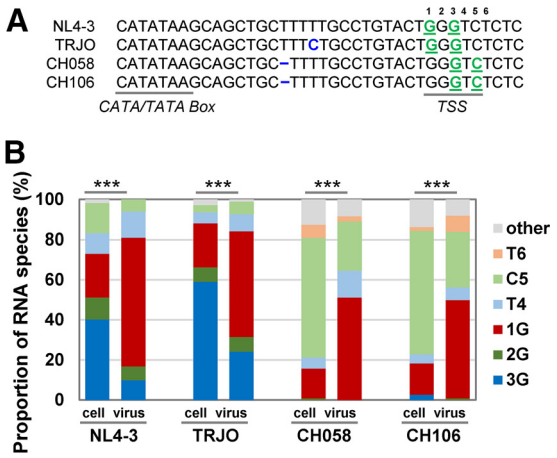

FIG 2 TSS usage of HIV-1 group M subtype B transmitted founder viruses. (A) Alignment of sequences between the CATA/TATA box (25) and near the TSS compared to NL4-3. Green letters denote major start sites, which were defined as >20% of the sequences from cells or virions. Sequence differences relative to NL4-3 are shown in blue. (B) Proportions of RNA species in cells infected with and viruses produced by the indicated HIV-1 strains. TRJO, CH058, and CH106 are unmodified full-length viruses and were propagated in HUT/R5 cells. NL4-3 results from Fig. 1D (HUT/R5) are shown here for comparison. Results are summarized from two independent experiments with sequencing the following total number of clones: 110 (TRJO cell), 95 (TRJO virus), 127 (CH058 cell), 119 (CH058 virus), 109 (CH106 cell), and 112 (CH106 virus). ***, $P < 0.001$, chi-square test.

by distance from the TATA box (see Discussion). Nonetheless, all four viruses had significantly different TSS distributions between cells and virions ($P < 0.001$ in all cases, chi-square test). Specifically, 1G RNA was enriched in virions, while 3G (NL4-3, TRJO) or 5C (CH058, CH106) RNAs were depleted ($P < 0.001$ in all cases, Fisher's exact test). These results demonstrate that 1G RNA is selectively packaged over 3G or 5C RNAs in all the examined subtype B viruses.

TSS usage in HIV-1 group M subtype C viruses. To investigate whether other HIV-1 group M viruses have similar TSS usage and selective packaging of 1G RNA, we examined the 5′ ends of unspliced viral RNAs for MJ4 and CH185, two subtype C viruses (26, 27), which is the most prevalent subtype globally (28). In addition, CH185 is a transmitted founder virus identified from an acutely infected individual. Compared to NL4-3, MJ4 and CH185 viruses have the same three guanosines and immediate surrounding sequences at the putative TSS (Fig. 3A). Surprisingly, unlike NL4-3 and other subtype B viruses, the subtype C viruses MJ4 and CH185 produced mostly 1G RNA in cells (Fig. 3B), with small amounts of 3G RNA also detected. Similar results were observed using an MJ4-based H0-like vector in 293T cells and full-length replication-competent MJ4 in T cells (Fig. 3B). Thus, these two subtype C viruses had less transcriptional heterogeneity than the subtype B viruses we examined. Even so, MJ4 and CH185 had statistically significant differences in TSS distributions between cells and virions ($P < 0.001$ [MJ4 293T, CH185], $P = 0.008$ [MJ4 HUT/R5], chi-square test). In all cases, 3G RNA was depleted in virions ($P < 0.001$ [MJ4 293T], $P = 0.005$ [MJ4 HUT/R5], $P = 0.02$ [CH185], Fisher's exact test). In addition, 1G RNA was enriched in virions in all cases ($P < 0.001$ [MJ4 293T], $P = 0.049$ [MJ4 HUT/R5], $P = 0.045$ [CH185], Fisher's exact test). In sum, the subtype C viruses we examined predominantly generated 1G RNA and appeared to have less transcriptional heterogeneity than subtype B viruses; however, they still selectively packaged 1G RNA into virions.

TSS usage in other HIV-1 and HIV-1-related viruses. TSS heterogeneity and selective packaging of 1G RNA may have arisen during the spread of HIV-1 in humans or prior to zoonotic transmission of SIV in chimpanzees or gorillas. To distinguish between these possibilities, we next determined TSS in unspliced viral RNA from cells infected with and virions generated from SIVcpzMB897 and SIVcpzEK505. SIVcpzMB897 is closely related to the ancestral virus that jumped species and was transmitted into humans becoming HIV-1 group M, whereas SIVcpzEK505 is closely related to the ancestral virus that became HIV-1 group N (29). Both SIVcpzMB897 and SIVcpzEK505 have three guanosines at the putative TSS and flanking sequences identical to NL4-3 (Fig. 4A). In 293T cells, both SIVcpzMB897

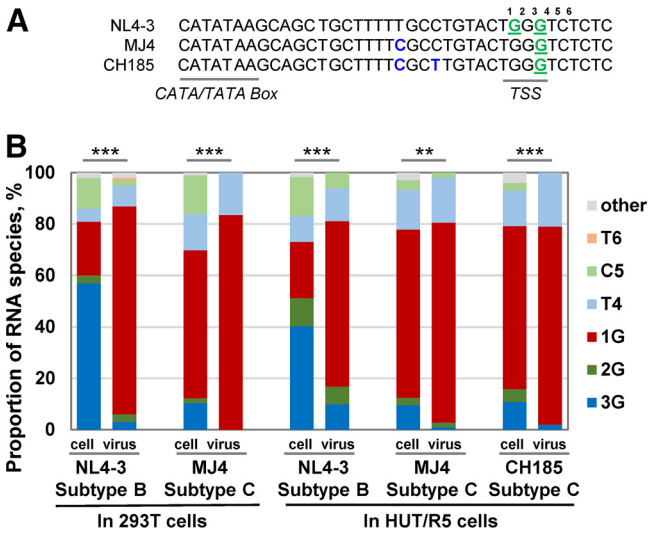

FIG 3 TSS usage of HIV-1 subtype C viruses. (A) Alignment of sequences between the CATA/TATA box and near the TSS compared to NL4-3. Green letters denote major TSS, whereas blue letters denote sequence differences relative to NL4-3. (B) Proportions of RNA species in cells infected with and viruses produced by the indicated HIV-1 strains. MJ4 results in 293T cells were generated using the previously described constructs CH0 and CT6 (37). Full-length MJ4 and CH185 were propagated in HUT/R5 cells. NL4-3 results from Fig. 1D are shown for comparison. Results are summarized from at least two independent experiments with sequencing data from the following total number of clones: 106 (MJ4 cell, 293T), 97 (MJ4 virus, 293T), 104 (MJ4 cell, HUT/R5), 108 (MJ4 virus, HUT/R5), 101 (CH185 cell), and 100 (CH185 virus). ***, $P < 0.001$; **, $P < 0.01$, chi-square test.

and SIVcpzEK505 produced multiple RNA species with 3G RNA being the most abundant species (Fig. 4B). Additionally, in both viruses, 1G RNA was enriched in virions, whereas 3G RNA was depleted ($P < 0.001$ in all cases, Fisher's exact test). These results indicate that TSS heterogeneity and selective packaging of 1G RNA likely predate the emergence of HIV-1 groups M and N.

HIV-1 groups O and P were introduced into humans by cross-species transmission of SIVgor (8, 12, 30). However, like HIV-1 groups M and N, SIVgor is also thought to have arisen from SIVcpz in a cross-species transmission to gorilla (11, 12). Thus, we also investigated TSS usage in HIV-1 group O CMO2.5 and SIVgor2139 viruses (Fig. 4A) (11, 31). We found that both viruses produced multiple unspliced RNA species in 293T cells, including 3G and 1G RNA. However, 1G RNA was the most abundant species in both viruses. Like HIV-1 group M and SIVcpz, 3G RNA was depleted and 1G RNA was selectively packaged into virions ($P < 0.001$ in all cases, Fisher's exact test). These results indicate that transcriptional heterogeneity and selective packaging of 1G RNA are highly conserved features of the HIV-1/SIVcpz/SIVgor lineage of primate lentiviruses.

Next, we analyzed SIVrcmGAB1, which also has three guanosines at the putative TSS (Fig. 4A). SIVrcm is thought to be one of the precursors of SIVcpz, although the origin of the SIVcpz LTR (which includes the TSS) is still uncertain (8, 13, 14). In GHOST-CCR2b cells, SIVrcmGAB1 produced almost exclusively 3G RNA (Fig. 4B), unlike any of the other viruses we analyzed. Furthermore, there was no significant difference in TSS distribution between cells and virions ($P = 0.10$, chi-square test). Thus, SIVrcmGAB1 did not exhibit selective packaging of 1G RNA or depletion of 3G RNA. These results clearly differ from those observed for SIVcpz and suggest that SIVcpz TSS heterogeneity and selective packaging of 1G RNA did not arise from SIVrcm.

TSS usage of HIV-2, SIVmac, SIVsmm, and SIVagm. To determine whether other lineages of primate lentiviruses exhibit TSS heterogeneity and selective packaging of a particular RNA species, we next analyzed HIV-2 and SIVs related to HIV-2: SIVmac, SIVsmm, and SIVagm. We examined the TSS usage of unspliced RNA in infected cells and virions for HIV-2 ROD, HIV-2 ST, SIVmac239, and SIVsmmE543. HIV-1 sequences nearly always have GGG at the TSS. However, most HIV-2 sequences have CAG at the putative TSS, although

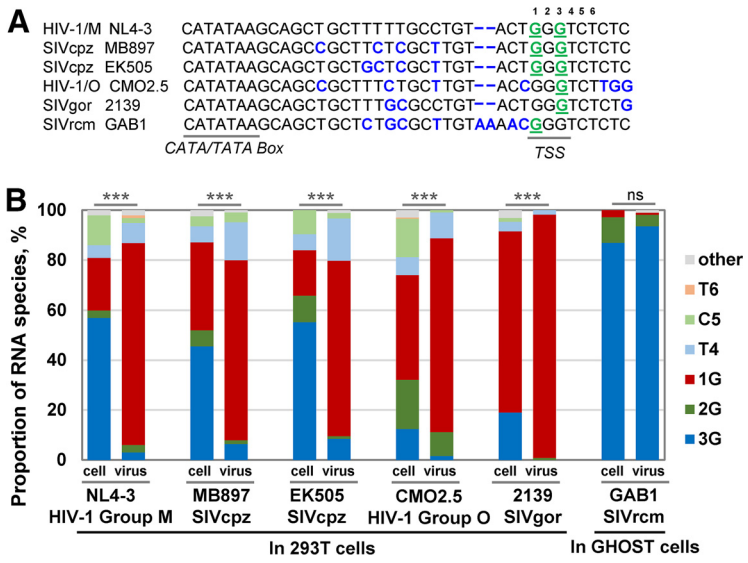

FIG 4 TSS usage of SIVcpz, HIV-1 group O, SIVgor, and SIVrcm. (A) Alignment of sequences between the CATA/TATA box and near the TSS compared to NL4-3. Green letters denote major TSS, whereas blue letters denote sequence differences relative to NL4-3. (B) Proportions of RNA species in cells infected with and viruses produced by a given virus. SIVcpz MB897 molecular clone was engineered to be encoded by two plasmids containing an overlapping region and inactivated *env* gene. To generate virus, these two plasmids were cotransfected with a plasmid encoding VSV G protein into 293T cells, and the resulting viruses were used to infect 293T cells, from which cellular and virion RNAs were isolated and analyzed. The same approach was used to generate samples from SIVcpzEK505 and SIVgor2139. HIV-1 group O results were generated using a previously described vector (38) similar to H0 (Fig. 1B). Full-length unmodified SIVrcmGAB1 was propagated in GHOST-CCR2b cells. NL4-3 results from Fig. 1D are shown for comparison. Results are summarized from at least two independent experiments with sequencing data from the following total number of clones: 125 (MB897 cell), 125 (MB897 virus), 94 (EK505 cell), 94 (EK505 virus), 208 (HIV-1 group O cell), 125 (HIV-1 group O virus), 131 (SIVgor cell), 116 (SIVgor virus), 108 (SIVrcm cell), and 109 (SIVrcm virus). ***, $P < 0.001$; ns, not significant, chi-square test.

some have CGG instead. Among the viruses we used, HIV-2 ROD has CGG, whereas HIV-2 ST, SIVmac239, and SIVsmmE543 have CAG at the putative TSS (Fig. 5A). In 293T and HUT/R5 cells, we found that HIV-2 ROD RNAs most often initiated at the 5′ guanosine in the CGG sequence (referred to as G1; Fig. 5B). These results are consistent with a previous report (32). However, we also observed a significant proportion of RNAs initiating from the 3′ guanosine (referred to as G2; Fig. 5B). HIV-2 ST, SIVmac239, and SIVsmmE543 RNAs in cells most often initiated at the adenosine in the CAG sequence (referred to as A1; Fig. 5B). However, HIV-2 ST also generated significant proportions of RNA from the cytosine (C-1) and guanosine (G2). Likewise, SIVmac239 also generated a significant proportion of RNA from the guanosine (G2). In contrast, SIVsmmE543 did not utilize alternative TSS to a significant extent. Thus, like most HIV-1-related viruses, most HIV-2-related viruses exhibited substantial transcriptional heterogeneity, although we identified exceptions in both cases.

We next determined the TSS of unspliced viral RNA in virions produced by HIV-2 and HIV-2-related SIVs. In all cases, we found that the predominant TSS in virions was G1 (HIV-2 ROD) or A1 (HIV-2 ST, SIVmac239, SIVsmmE543) (Fig. 5B). Furthermore, we found that TSS distributions were often significantly different between cells and virions. Specifically, for HIV-2 ROD, HIV-2 ST, and SIVmac239, G1 or A1 RNAs were enriched in virions ($P = 0.01$ [HIV-2 ROD 293T], $P = 0.002$ [HIV-2 ROD HUT/R5], $P < 0.001$ [HIV-2 ST, SIVmac239], Fisher's exact test). However, TSS distributions did not significantly differ between cells and virions for SIVsmmE543 ($P = 0.24$, chi-square test), as A1 RNA comprised a higher proportion of cellular RNA for this virus. Thus, HIV-2 and some of the HIV-2-related viruses also selected certain RNA species for genome packaging based on TSS.

Lastly, we analyzed the TSS usage of SIVagm, which belongs to a lineage of SIV distinct from HIV-1- and HIV-2-related viruses. SIVagm sequences have CAG near the putative TSS (Fig. 5A). In 293T cells, SIVagmTAN1 generated mostly A1 RNA, similar to HIV-2-related viruses (Fig. 5B). However, small amounts of C-1 and G2 RNAs were also detected. In virions, A1 RNA

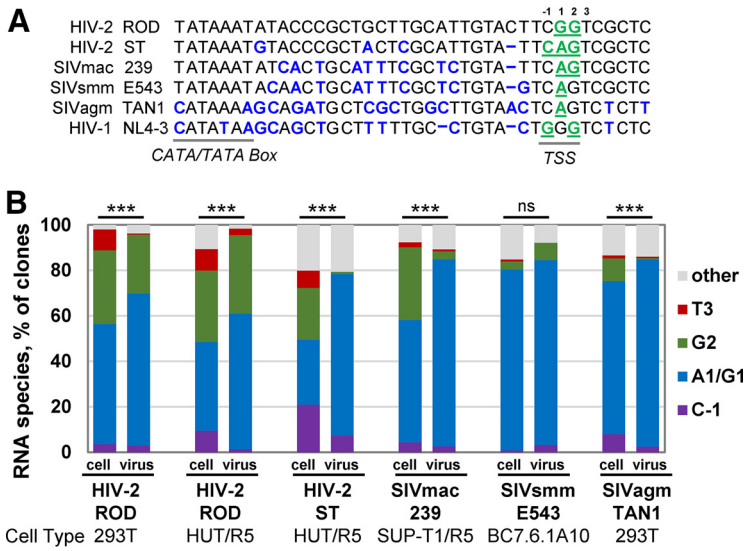

FIG 5 TSS usage of HIV-2, SIVmac, SIVsmm, and SIVagm. (A) Alignment of sequences between the CATA/TATA box and near the TSS compared to HIV-2 ROD isolate. Green letters denote major TSS, whereas blue letters denote sequences differences relative to ROD. (B) Proportions of RNA species in cells infected with and viruses produced by a given virus. HIV-2 ROD and SIVagm results in 293T cells were generated using previously described constructs (39). Full-length unmodified HIV-2 ROD and HIV-2 ST were propagated in HUT/R5 cells, whereas SIVmac239 and SIVsmm were propagated in SUP-T1/R5 cells and BC7.6.1A10 cells, respectively. Results are summarized from at least two independent experiments with sequencing data from the following total number of clones: 163 (ROD cell, 293T), 166 (ROD virus, 293T), 105 (ROD cell, HUT/R5), 136 (ROD virus, HUT/R5), 105 (ST cell), 107 (ST virus), 134 (SIVmac cell), 113 (SIVmac virus), 113 (SIVsmm cell), 91 (SIVsmm virus), 150 (SIVagm cell), and 159 (SIVagm virus). ***, $P < 0.001$; ns, not significant, chi-square test.

was the predominant species, and it was slightly enriched in virions relative to cells ($P = 0.003$, Fisher's exact test). Taken together, our results demonstrate that TSS heterogeneity and selective packaging of a particular RNA species are conserved features of primate lentiviruses. However, the extent of TSS heterogeneity and the degree of selection during packaging vary widely, even among closely related viruses.

DISCUSSION

Recent studies have demonstrated that HIV-1 generates multiple full-length viral RNAs via heterogenous TSS usage and selectively packages a specific full-length viral RNA species (1G RNA) (2–7), distinct from an earlier report (33). In this current study, we examined whether these features are conserved across primate lentiviruses by determining TSS of viral RNAs in cells and virions for 15 viruses. Based on phylogenetic analysis (Fig. 6), these viruses can be divided into 4 main groups: HIV-1-related viruses, HIV-2-related viruses, SIVagm, and SIVrcm. HIV-1-related viruses, including HIV-1, SIVcpz, and SIVgor, usually have GGG at the TSS (Fig. 6). We found that these viruses exhibited TSS heterogeneity and selectively packaged 1G RNA into virus particles. Nonetheless, we identified significant differences in TSS usage among HIV-1-related viruses, even within the same group or subtype. HIV-2-related viruses, including HIV-2, SIVmac, and SIVsmm, usually have CAG at the TSS, although HIV-2 ROD has CGG (Fig. 6). Like HIV-1, most of these viruses exhibited TSS heterogeneity and selective packaging of a particular RNA species. SIVagm has CAG at the TSS, like most HIV-2-related viruses (Fig. 6). SIVagm also displayed TSS heterogeneity and selective packaging of a specific RNA species. In contrast, SIVrcm has variable sequences near the TSS. Of the four SIVrcm sequences available (http://www.hiv.lanl.gov), three have CAG at the TSS (like HIV-2 and SIVagm), whereas one has GGG (like HIV-1) (Fig. 6). We analyzed SIVrcmGAB1, which has GGG at the TSS, because reagents were not available for the other strains. We found that SIVrcmGAB1 generated almost exclusively 3G RNA, with no significant difference in the distribution of RNA species between cells and virions. Collectively, our results demonstrate that heterogeneity in TSS usage and selective packaging of a particular RNA species are common features in primate lentiviruses and are

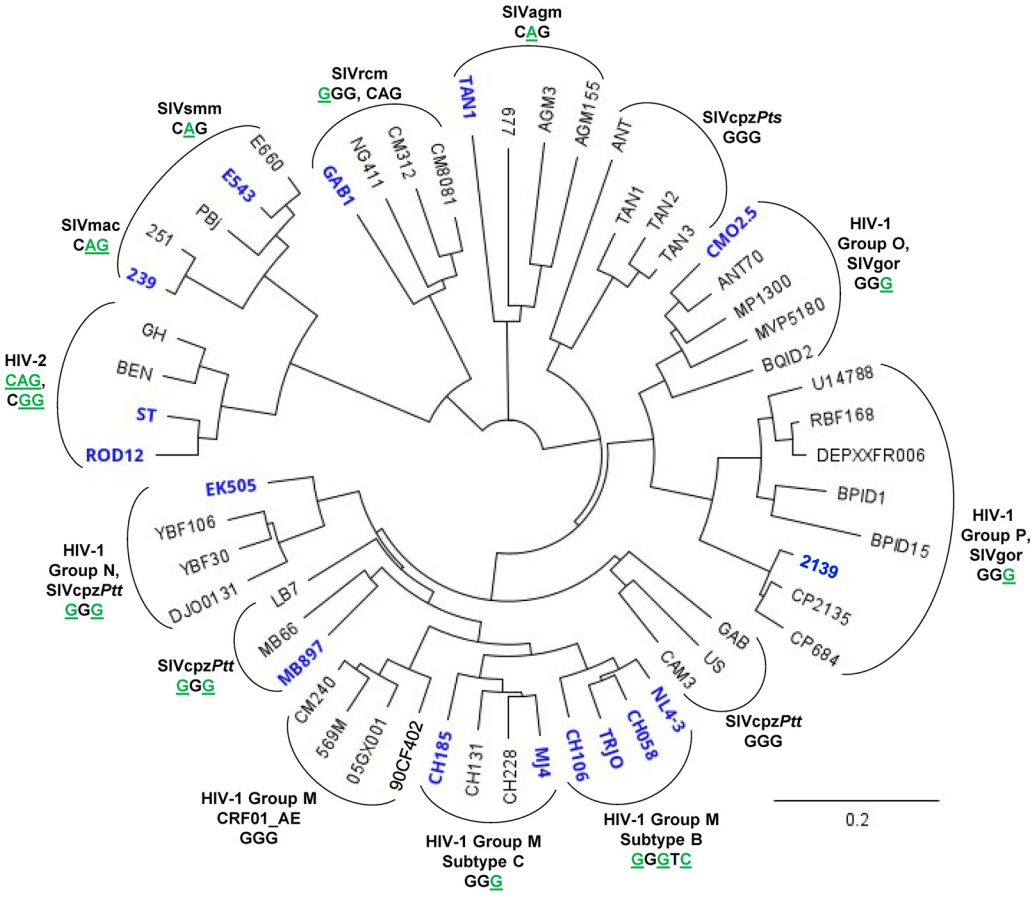

FIG 6 Evolutionary relationship of HIV/SIV LTR sequences. Phylogenetic trees based on LTR (U3-R-U5) sequences were constructed using the neighbor-joining method. Blue labels indicate viruses that were analyzed in this report. Sequences near the putative TSS (based on alignment with the NL4-3 TSS) are indicated below the label for each group. Green underlined nucleotides indicate major TSS that were identified in at least one virus from the indicated group. Major TSS were defined as sites used by ≥20% of the sequences from either the cellular RNA sample or the virion RNA sample. Scale bar: number of substitutions/site.

not unique to HIV-1-related viruses or viruses with GGG TSS sequences. However, we did identify two exceptions: SIVrcm and SIVsmm used multiple TSS but did not select a specific RNA species for genome packaging.

One surprising observation from our study is that closely related viruses can have striking differences in TSS usage. For example, among subtype B viruses, NL4-3 and TRJO produced mostly 3G and 1G RNAs, whereas CH058 and CH106 produced mostly 1G and 5C RNAs. Upon analyzing the promoter sequences of these viruses, we found that, compared to NL4-3, CH058 and CH106 have a 1-nt deletion between the TATA box and the TSS. In cellular genes that utilize TATA box-dependent transcription, TSS are determined by nt sequence as well as distance from the TATA box (34–36). Thus, it is possible that the 1-nt deletion altered the TSS usage of CH058 and CH106. However, not all of the differences in HIV-1 TSS usage can be explained by the distance between the TATA box and TSS. For example, the distances between the TATA box and TSS are identical among NL4-3, MJ4, and CH185, yet 1G RNA dominates the HIV-1 RNA species in cells infected with MJ4 and CH185, whereas 3G RNA dominates in cells infected with NL4-3. Further studies will be needed to determine the mechanism(s) that cause the distinct patterns of TSS usage in these viruses. Nonetheless, despite differences in TSS heterogeneity, all HIV-1-related viruses predominantly packaged 1G RNA into particles. In every case, 1G RNA was enriched in virions relative to cells, indicating that it was preferentially packaged over 3G and C5 RNAs. However, the degree to which 1G RNA was enriched in virions varied widely among viruses, mainly due to differences in the intracellular proportions of 1G RNA. Therefore, our findings suggest that selective packaging of 1G RNA is

highly conserved among HIV-1-related viruses. The distinct usage of TSS in closely related viruses can also be observed in HIV-2-related viruses. Among the viruses we examined, HIV-2 ST exhibited the greatest degree of TSS heterogeneity and selective packaging, while SIVsmmE543 exhibited the least. Among HIV-2 related viruses, there were sequence differences in the region between the CATA/TATA box and the TSS (Fig. 5A), which may have led to differences in TSS usage.

Previous studies have shown that 1G RNA and 3G RNA fold into distinct structures: 1G RNA, but not 3G RNA, folds into structure(s) that expose the RNA dimerization signal and Gag-binding sites, facilitating genome packaging (3, 7). It has been proposed that the 5′ context of the HIV-1 RNA also affects other RNA functions. This model, based on NL4-3 (subtype B) and MAL (subtype A), proposed that the 5′ cap is sequestered in 1G RNA, thereby preventing its translation, whereas the 5′ cap is exposed in 3G RNA, thereby promoting its translation (3, 5, 6). In our study, we found that different HIV-1 strains had very different ratios of 3G and 1G RNA; in some viruses such as CH058, CH106, MJ4, and CH185, very little 3G RNA was generated ($\leq$10%). Thus, it is unclear how these viruses produce sufficient RNA to serve as templates for Gag/Gag-Pol translation. One possibility is that in these viruses, RNA species other than 3G are preferentially translated. Alternatively, it is possible the 5′ context does not affect translation in these viruses. Further studies are needed to distinguish between these possibilities.

Lastly, our findings provide insights into the evolutionary origin of HIV-1 TSS heterogeneity and selective packaging of 1G RNA. In SIVcpz, the progenitor of HIV-1 groups M and N, we observed significant TSS heterogeneity, with mostly 3G and some 1G RNA in cells, and selective packaging of 1G RNA into virions. These results closely resemble those obtained for some HIV-1 strains (NL4-3 and TRJO), indicating that TSS heterogeneity and selective packaging of 1G RNA likely preceded the emergence of HIV-1. Thus, these features may have originated in SIVcpz or earlier SIV precursors. SIVcpz arose from recombination between at least two viruses: SIVrcm and SIVmus/mon/gsn (13). We found that SIVrcmGAB1 has GGG at the TSS like SIVcpz but generated and packaged almost exclusively 3G RNA. Thus, SIVrcm lacked TSS heterogeneity and selective packaging of 1G RNA, indicating that these features may have evolved directly in SIVcpz or may have arisen from a different SIV lineage.

MATERIALS AND METHODS

Molecular clones and viruses. Two types of viruses were used in this study: unmodified replication-competent viruses and modified viral constructs. All modified viral constructs had intact LTRs, 5′ UTR, and functional *gag-pol*; thus, transcription regulation and genome packaging were carried out by authentic elements from the same virus. The previously described NL4-3-based construct H0 (20) contains all the *cis*-acting elements required for HIV-1 replication and expresses functional Gag/Gag-Pol, Tat, and Rev. In addition, an *hsa* gene, internal ribosomal entry site (IRES) from encephalomyocarditis virus, and a mutated green fluorescent protein (*gfp*) gene were inserted in the *nef* gene (20); for simplicity, IRES-*gfp* is not described in the text or Fig. 1. Results of HIV-1 studies from 293T cells were generated using previously described constructs derived from other subtypes/groups but having H0-like structures. Specifically, MJ4 results were from a 293T cell line infected with CH0 and CT6 (37), and HIV-1 group O results were from cells infected with OHIG (38). CT6 has the same general structure as CH0 except CH0 expresses *hsa* whereas CT6 expresses a Thy1.2 gene (37). The previously described construct HIV2-H0G (39) is based on the ROD12 isolate of HIV-2 (40) and has a frameshift mutation in *vpr*, an inactivating deletion in *env*, and an insertion of a fragment containing *hsa*-IRES-inactivated *gfp* in the *nef* gene. The previously described SIVagm construct pTan-H0G is based on SIVagm molecular clone Tan-1 which has inactivating mutations in *vpr* and *env*, and an *hsa*-IRES-inactivated *gfp* fragment inserted in the *nef* gene (39).

The full-length infectious molecular clones pNL4-3, pTRJO, pCH058, pCH106, pMJ4 (21–24, 26), and pHIV-2/ST were used for spreading infections and were obtained from the National Institutes of Health AIDS Research and Reference Reagent Program (NIH ARP). The infectious molecular clone pROD12 is a kind gift from Keith Peden (FDA), whereas pCH185 is a kind gift from Christina Ochsenbauer and John Kappes (University of Alabama). SupT1/R5 cells stably infected with SIVmac239*, which contains a Q734stop mutation in the *env* gene and BC7.6.1A10 cells stably infected with SIVsmm-E543-3 were generous gifts from Julian Bess (AIDS and Cancer Virus Program, Frederick National Laboratory). Infectious virus stock for SIVrcm-GAB1 was obtained from NIH ARP.

Molecular clones for SIVcpz-MB897, SIVcpz-EK505, and SIVgor-2139 were generous gifts from Beatrice Hahn. As these plasmids frequently acquired deletions or insertions during propagation, we split each molecular clone into two plasmids. SIVcpz-MB897-3′ was created by deleting sequences between the MluI restriction site upstream of the 5′ LTR and the NheI site in *env*, whereas SIVcpz-MB897-5′ was created by deleting sequences between the StuI restriction site in *env* and the SnaBI site downstream of the 3′ LTR. The SIVcpz-MB897-5′ and SIVcpz-MB897-3′ plasmids have a 458-bp overlap in viral sequences to facilitate homologous recombination during transfection. In addition, the *env* gene in SIVcpz-MB897-3′ was inactivated by introducing a frameshift

mutation using Klenow fill-in of an EcoRI site. To generate SIVcpz-EK505-3′, a region between the MluI restriction site upstream of the 5′ LTR and the BstZ17I site in *pol* was deleted; SIVcpz-EK505-5′ was generated by deleting sequences between the SbfI restriction site in *vif* and the NsiI site downstream of the 3′ LTR. The SIVcpz-EK505-5′ and SIVcpz-EK505-3′ plasmids contain a 2.5-kb overlap in viral sequences. In addition, a portion of the *env* gene in SIVcpz-EK505-3′ was deleted by digestion with ScaI and re-ligation. The SIVgor-2139 molecular clone we propagated contained a 1.3-kb bacterial sequence inserted in the *env* gene causing its inactivation; additionally, the TSS sequence was GTGG, which is atypical for SIVgor isolates. We generated SIVgor2139-3′ by deleting a region between the ZraI restriction site upstream of the 5′ LTR and the HpaI site in *pol*, whereas SIVgor2139-5′ was created by deleting sequences between SphI restriction sites in *env* and downstream of the 3′ LTR. The resulting SIVgor2139-5′ and SIVgor2139-3′ contain a 3.6-kb overlap in viral sequences. To generate SIVgor plasmids with the more representative sequence of TGGG at the TSS, 3′ and 5′ LTR fragments were synthesized (IDT) and used to replace the original sequences between the BbvCI and NotI sites in SIVgor2139-3′ or the MluI and StuI sites in SIVgor2139-5′, respectively. All cloning was performed using the NEBuilder HiFi DNA Assembly kit (NEB) or standard molecular cloning techniques. The viral genome portions in SIVcpz and SIVgor plasmids were sequenced in entirety to ensure sequence integrity.

Cell culture, infections, and flow cytometry. Human embryonic kidney 293T cells were obtained from American Type Culture Collection. CEM-SS (a derivative of human T cell line CEM), GHOST-X4, and GHOST-CCR2b cells were obtained from the NIH AIDS Reagent Program. GHOST are human osteosarcoma cells derived from HOS that stably express CD4 and either CXCR4 or CCR2b, respectively. HUT/R5 is a derivative of human T-cell line HUT78 that stably expresses CCR5 (41). Both SUP-T1/R5 and BC7.6.1A10 are human T cell lines derived from SUP-T1. GHOST and 293T cells were maintained in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum (FBS), 100 U/mL penicillin, and 100 μg/mL streptomycin. CEM-SS, HUT/R5, SUP-T1/R5, and BC7.6.1A10 cells were maintained in Roswell Park Memorial Institute (RPMI-1640) medium supplemented with 10% FBS, 100 U/mL penicillin, and 100 μg/mL streptomycin. All cells were maintained in humidified 37°C incubators with 5% CO_2.

To generate VSV-G-pseudotyped viruses, 293T cells were transfected using TransIT-LT1 (Mirus Bio) with viral constructs and the helper plasmids pSynGagPol and pHCMV-G that express codon-optimized HIV-1 Gag/Gag-Pol and VSV-G, respectively (42, 43). These viruses were used to infect fresh 293T cells to generate infected cell pools. SIVcpz and SIVgor plasmids were linearized prior to transfection to facilitate homologous DNA recombination. For this purpose, SIVcpzMB897-5′ and -3′ were digested with NheI, SIVcpz-EK505-5′ and -3′ were digested with ApaI and AccI, respectively, and SIVgor-TGGG-5′ and -3′ were digested with SphI and KpnI, respectively. Supernatants were harvested 24 to 48 h after transfection, clarified through a 0.45-μm filter (Millex), and in some cases concentrated 10- to15-fold by ultracentrifugation at 25,000 $\times$ *g* for 90 min through a 20% sucrose cushion. To generate pools of infected cells, more than 1 million 293T cells were infected with viruses at MOIs of 0.1–1.9 and analyzed 3 days postinfection by flow cytometry to determine the proportion of infected cells. Pools of infected cells consisted of at least 95,000 independent infection events.

To prepare virus stocks of NL4-3, TRJO, CH058, CH106, MJ4, CH185, ROD, and ST for spreading infections, 293T cells were transfected in 6-well plates using 3 μg plasmid/well and TransIT-LT1 (Mirus Bio). Virus supernatants were collected 48 h posttransfection and clarified through 0.45-μm filters. HIV-1 virus stocks were quantified by p24 ELISA (XpressBio), whereas HIV-2 virus stocks (ROD, ST) were quantified by p27 ELISA (XpressBio). HUT/R5 or CEM-SS cells were infected with 100 μL of undiluted virus, using 1 million cells in 3 mL volume. Infections were performed in duplicate using independent virus stocks. Starting 3 days postinfection, cells and virus supernatants were collected every other day. Briefly, 1.5 mL of culture was removed from each infection and centrifuged at 16,000 $\times$ *g* for 10 min. Viral supernatants and cell pellets (resuspended in 150 μL RNAlater) were frozen at −80°C. Lastly, 1.5 mL of fresh medium was added to the cultures. GHOST-X4 and GHOST-CCR2b cells were infected with NL4-3 or SIVrcmGAB1, respectively, as follows: 30,000 cells/well were seeded in 48-well plates using 200 μL medium/well and infected the next day with 50 μL of undiluted virus by spinoculation at 1200 $\times$ g for 1 h at room temperature in the presence of 20 μg/mL Polybrene. At 24 h postinfection, the cells were split into 6-well plates; cells and viral supernatants were collected 3 and 6 days postinfection. Virus replication was monitored by p24 (NL4-3, TRJO, CH058, CH106, MJ4, and CH185) or p27 (ROD, ST, and SIVrcm) ELISA.

To perform flow cytometry analysis of cells infected with HIV-1 or HIV-2 single-round viruses, cells were collected, washed twice with Dulbecco's phosphate-buffered saline (PBS) containing 2.5% fetal bovine serum (FBS), then stained with 0.4 μg/mL phycoerythrin- or allophycocyanin-conjugated anti-HSA antibodies (BioLegend) for 30 min, and washed two additional times with PBS containing 2.5% FBS. Infections by SIVcpz and SIVgor were quantified using internal p24 staining; briefly, infected 293T cells were fixed for 30 min with 2% formaldehyde in PBS containing 2.5% FBS, washed twice in PBS with 2.5% FBS, stained with RD1-conjugated anti-p24 antibodies (KC57-RD1, Beckman Coulter) in permeabilization buffer (PBS with 2.5% FBS and 0.5% Saponin), washed twice in PBS with 2.5% FBS, and fixed in 2% formaldehyde in PBS. Flow cytometry was performed using an LSR II system (BD Biosciences) and data were analyzed using FlowJo software (TreeStar, LLC).

RNA isolation and 5′ RACE. To isolate RNA from viral particles, supernatant from infected cells was harvested, clarified through a 0.45-μm filter (Millex), and concentrated by centrifugation at 25,000 $\times$ *g* for 90 min through a 20% sucrose cushion, and total RNA was isolated using a Mini Viral RNA kit (Qiagen) according to the manufacturer's protocol. Total cellular RNA from infected cells was isolated using the RNeasy Plus minikit (Qiagen) according to the manufacturer's protocol. Isolated RNA was converted into cDNA using the SMARTer RACE 5′/3′ kit (TaKaRa); RNA was denatured immediately prior to cDNA synthesis as described in the manufacturer's protocol. All primers for cDNA preparations targeted the *gag* gene to ensure that only full-length RNA is analyzed. Next, PCR was performed using a 5′ universal primer from the SMARTer RACE 5′/3′ kit and a 3′ virus sequence-specific primer. Primers used in this study are shown in Table 1. The products of 5′ RACE were gel-purified and cloned into the pRACE plasmid (TaKaRa). For the 5′ RACE reactions, $>1 \times 10^6$ copies of full-length viral RNA were typically used for cDNA synthesis; subsequently, $>1 \times 10^6$ copies of viral cDNA were used for PCR,

TABLE 1 RT and PCR Primers used for 5′ RACE

Virus	RT primer	Reverse PCR primer[a]
NL4-3	GGTGGCTCCTTCTGATAATG	*GATTACGCCAAGCTT*-TCGTTCTAGCTCCCTGCTTG
CH058	GGTGGCTCCTTCTGATAATG	*GATTACGCCAAGCTT*-TCGTTCTAACTCCCTGCTTG
CH106	GGTGGCTCCTTCTGATAATG	*GATTACGCCAAGCTT*-TCGTTCTAGCTCCCTGCTTG
TRJO	GGTGGCTCCTTCTGATAATG	*GATTACGCCAAGCTT*-TCGTTCTAGCTCCCTGCTTG
MJ4	GGTGGCTCCTTCTGATAATG	*GATTACGCCAAGCTT*-AGCTCCCTGCTTGCCCATAC
CH185	GGTGGCTCCTTCTGATAATG	*GATTACGCCAAGCTT*-AGCTCCCTGCTTGCCCATAC
CMO2.5	GGGAGATGGCACTATGTACC	*GATTACGCCAAGCTT*-CACACCAGAGCACTGCTATTGTGTTC
HIV-2 ROD	TACCCAGGCATTTAGGGTTC	*GATTACGCCAAGCTT*-TCTGTCCAATTCATTCGCTGCCCACAC
HIV-2 ST	TACCCAAGCATTTAGGGTTC	*GATTACGCCAAGCTT*-TCTGTCCAATTCATTCGCTGCCCACAC
SIVgor2139	GAGATCGCAAGCCATCTGAC	*GATTACGCCAAGCTT*-CGTTCCCATGCATCCAATCG
SIVcpzMB897	TCTTGCCCACATTTCCA	*GATTACGCCAAGCTT*-AGCTCCCTGCTTGCCCATAC
SIVcpzEK505	TCTTGCCCACATTTCCA	*GATTACGCCAAGCTT*-AGCTCCCTGCTTGCCCATAC
SIVmac239	TACCCAGGCATTTAATGTTC	*GATTACGCCAAGCTT*-CTATCTAATTCATTTGCTGCCC
SIVsmmE543	TACCCAAGCATTTAATGTTC	*GATTACGCCAAGCTT*-TCTGTCCAATTCATTTGCTGC
SIVrcmGAB1	GACCCAGGCATTTAGAGTCC	*GATTACGCCAAGCTT*-CCATACCAAATGCTTCAGCATG
SIVagmTAN1	GCGGGCTCAATACTTCTATG	*GATTACGCCAAGCTT*-CCAAATTCCTCCCTGACAGTGCCGAGTG

[a]5' extension with InFusion sequence *GATTACGCCAAGCTT* for cloning into pRACE plasmid is in *italic*. The forward PCR primer was provided by the 5′ RACE kit.

and $>5 \times 10^9$ copies of the PCR product were used for cloning. Most experiments yielded >1,000 transformed bacterial colonies. On average, >100 clones per sample from 2 to 5 independent experiments were analyzed. The 5′ UTR sequences were analyzed manually in CloneManager v2 (Sci Ed Software LLC) or with a custom R script.

Alignments, phylogenetic trees, and statistical analysis. Most LTR (U3-R-U5) sequences were obtained from the Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov). Several others were obtained from GenBank under the following accession numbers: MJ4 (HIV-1 group M, subtype C), AF321523.1; CMO2.5 (HIV-1 group O), AY623602; and RBF168 (HIV-1 group P), GU111555. LTR sequences were imported into Geneious Prime v2021.2.2 (Biomatters) and aligned with MUSCLE v3.8.425 (44). Neighbor-joining phylogenetic trees were constructed in Geneious Prime using the Tamura-Nei genetic distance model.

Chi-square tests were used to determine whether the distributions of TSS were different between cellular RNA and virion RNA. Fisher's exact tests were used to determine whether the proportions of a particular RNA species were different between cellular RNA and virion RNA. Two-sided tests were used in all cases. Both chi-square and Fisher's exact tests were performed in GraphPad Prism v9.2.0 (GraphPad Software, LLC).

Data availability. All data required to support conclusions are presented in the article.

ACKNOWLEDGMENTS

We thank Eric Freed and Frank Maldarelli for helpful discussions.

This work was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Cancer Institute (NCI), Center for Cancer Research, by NIH Intramural AIDS Targeted Antiviral Program grant funding (to W.-S.H. and to V.K.P.), and by the Innovation Award and Preliminary Budget Fund, Office of AIDS Research, NIH. This work was also supported in part with federal funds from the NCI, NIH, under contract 75N91019D00024. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. government.

REFERENCES

1. Freed EO, Martin MA. 2013. Human immunodeficiency virus: replication, p 1502–1560. *In* Knipe DM, Howley PM (eds), Fields virology, 6th ed, vol 2. Lippincott, Williams, & Wilkins, Philadelphia, PA.
2. Masuda T, Sato Y, Huang YL, Koi S, Takahata T, Hasegawa A, Kawai G, Kannagi M. 2015. Fate of HIV-1 cDNA intermediates during reverse transcription is dictated by transcription initiation site of virus genomic RNA. Sci Rep 5:17680. https://doi.org/10.1038/srep17680.
3. Kharytonchyk S, Monti S, Smaldino PJ, Van V, Bolden NC, Brown JD, Russo E, Swanson C, Shuey A, Telesnitsky A, Summers MF. 2016. Transcriptional start site heterogeneity modulates the structure and function of the HIV-1 genome. Proc Natl Acad Sci U S A 113:13378–13383. https://doi.org/10.1073/pnas.1616627113.
4. Pollpeter D, Parsons M, Sobala AE, Coxhead S, Lang RD, Bruns AM, Papaioannou S, McDonnell JM, Apolonia L, Chowdhury JA, Horvath CM, Malim MH. 2018. Deep sequencing of HIV-1 reverse transcripts reveals the multifaceted antiviral functions of APOBEC3G. Nat Microbiol 3:220–233. https://doi.org/10.1038/s41564-017-0063-9.
5. Brown JD, Kharytonchyk S, Chaudry I, Iyer AS, Carter H, Becker G, Desai Y, Glang L, Choi SH, Singh K, Lopresti MW, Orellana M, Rodriguez T, Oboh U, Hijji J, Ghinger FG, Stewart K, Francis D, Edwards B, Chen P, Case DA, Telesnitsky A, Summers MF. 2020. Structural basis for transcriptional start site control of HIV-1 RNA fate. Science 368:413–417. https://doi.org/10.1126/science.aaz7959.
6. Ding P, Kharytonchyk S, Kuo N, Cannistraci E, Flores H, Chaudhary R, Sarkar M, Dong X, Telesnitsky A, Summers MF. 2021. 5'-Cap sequestration is an essential determinant of HIV-1 genome packaging. Proc Natl Acad Sci U S A 118:e21124751. https://doi.org/10.1073/pnas.2112475118.
7. Nikolaitchik OA, Liu S, Kitzrow JP, Liu Y, Rawson JMO, Shakya S, Cheng Z, Pathak VK, Hu WS, Musier-Forsyth K. 2021. Selective packaging of HIV-1 RNA genome is guided by the stability of 5' untranslated region polyA stem. Proc Natl Acad Sci U S A 118:e2114494118. https://doi.org/10.1073/pnas.2114494118.
8. Sharp PM, Hahn BH. 2011. Origins of HIV and the AIDS pandemic. Cold Spring Harb Perspect Med 1:a006841. https://doi.org/10.1101/cshperspect.a006841.
9. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM, Hahn BH. 1999. Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. Nature 397:436–441. https://doi.org/10.1038/17130.

10. Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH. 2001. The origins of acquired immune deficiency syndrome viruses: where and when? *Philosophical transactions of the*. Philos Trans R Soc Lond B 356:867–876. https://doi.org/10.1098/rstb.2001.0863.
11. Takehisa J, Kraus MH, Ayouba A, Bailes E, Van Heuverswyn F, Decker JM, Li Y, Rudicell RS, Learn GH, Neel C, Ngole EM, Shaw GM, Peeters M, Sharp PM, Hahn BH. 2009. Origin and biology of simian immunodeficiency virus in wild-living western gorillas. J Virol 83:1635–1648. https://doi.org/10.1128/JVI.02311-08.
12. Van Heuverswyn F, Li Y, Neel C, Bailes E, Keele BF, Liu W, Loul S, Butel C, Liegeois F, Bienvenue Y, Ngolle EM, Sharp PM, Shaw GM, Delaporte E, Hahn BH, Peeters M. 2006. Human immunodeficiency viruses: SIV infection in wild gorillas. Nature 444:164. https://doi.org/10.1038/444164a.
13. Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, Marx PA, Hahn BH, Sharp PM. 2003. Hybrid origin of SIV in chimpanzees. Science 300:1713. https://doi.org/10.1126/science.1080657.
14. Bell SM, Bedford T. 2017. Modern-day SIV viral diversity generated by extensive recombination and cross-species transmission. PLoS Pathog 13:e1006466. https://doi.org/10.1371/journal.ppat.1006466.
15. Nyamweya S, Hegedus A, Jaye A, Rowland-Jones S, Flanagan KL, Macallan DC. 2013. Comparing HIV-1 and HIV-2 infection: lessons for viral immunopathogenesis. Rev Med Virol 23:221–240. https://doi.org/10.1002/rmv.1739.
16. Tchounga B, Ekouevi DK, Balestre E, Dabis F. 2016. Mortality and survival patterns of people living with HIV-2. Curr Opin HIV AIDS 11:537–544. https://doi.org/10.1097/COH.0000000000000299.
17. Santiago ML, Range F, Keele BF, Li Y, Bailes E, Bibollet-Ruche F, Fruteau C, Noë R, Peeters M, Brookfield JFY, Shaw GM, Sharp PM, Hahn BH. 2005. Simian immunodeficiency virus infection in free-ranging sooty mangabeys (*Cercocebus atys atys*) from the Tai Forest, Côte d'Ivoire. J Virol 79:12515–12527. https://doi.org/10.1128/JVI.79.19.12515-12527.2005.
18. Peeters M, Jung M, Ayouba A. 2013. The origin and molecular epidemiology of HIV. Expert Rev Anti Infect Ther 11:885–896. https://doi.org/10.1586/14787210.2013.825443.
19. Foley BT, Leitner T, Paraskevis D, Peeters M. 2016. Primate immunodeficiency virus classification and nomenclature: Review. Infect Genet Evol 46:150–158. https://doi.org/10.1016/j.meegid.2016.10.018.
20. Rhodes TD, Nikolaitchik OA, Chen J, Powell D, Hu WS. 2005. Genetic recombination of human immunodeficiency virus type 1 in one round of viral replication: effects of genetic distance, target cells, accessory genes, and lack of high negative interference in crossover events. J Virol 79:1666–1677. https://doi.org/10.1128/JVI.79.3.1666-1677.2005.
21. Ochsenbauer C, Edmonds TG, Ding H, Keele BF, Decker J, Salazar MG, Salazar-Gonzalez JF, Shattock R, Haynes BF, Shaw GM, Hahn BH, Kappes JC. 2012. Generation of transmitted/founder HIV-1 infectious molecular clones and characterization of their replication capacity in CD4 T lymphocytes and monocyte-derived macrophages. J Virol 86:2715–2728. https://doi.org/10.1128/JVI.06157-11.
22. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, Wei X, Jiang C, Kirchherr JL, Gao F, Anderson JA, Ping L-H, Swanstrom R, Tomaras GD, Blattner WA, Goepfert PA, Kilby JM, Saag MS, Delwart EL, Busch MP, Cohen MS, Montefiori DC, Haynes BF, Gaschen B, Athreya GS, Lee HY, Wood N, Seoighe C, Perelson AS, Bhattacharya T, Korber BT, Hahn BH, Shaw GM. 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A 105:7552–7557. https://doi.org/10.1073/pnas.0802203105.
23. Salazar-Gonzalez JF, Salazar MG, Keele BF, Learn GH, Giorgi EE, Li H, Decker JM, Wang S, Baalwa J, Kraus MH, Parrish NF, Shaw KS, Guffey MB, Bar KJ, Davis KL, Ochsenbauer-Jambor C, Kappes JC, Saag MS, Cohen MS, Mulenga J, Derdeyn CA, Allen S, Hunter E, Markowitz M, Hraber P, Perelson AS, Bhattacharya T, Haynes BF, Korber BT, Hahn BH, Shaw GM. 2009. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. J Exp Med 206:1273–1289. https://doi.org/10.1084/jem.20090378.
24. Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, Keele BF, Derdeyn CA, Farmer P, Hunter E, Allen S, Manigart O, Mulenga J, Anderson JA, Swanstrom R, Haynes BF, Athreya GS, Korber BTM, Sharp PM, Shaw GM, Hahn BH. 2008. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. J Virol 82:3952–3970. https://doi.org/10.1128/JVI.02660-07.
25. van Opijnen T, Kamoschinski J, Jeeninga RE, Berkhout B. 2004. The human immunodeficiency virus type 1 promoter contains a CATA box instead of a TATA box for optimal transcription and replication. J Virol 78:6883–6890. https://doi.org/10.1128/JVI.78.13.6883-6890.2004.
26. Ndung'u T, Renjifo B, Essex M. 2001. Construction and analysis of an infectious human Immunodeficiency virus type 1 subtype C molecular clone. J Virol 75:4964–4972. https://doi.org/10.1128/JVI.75.11.4964-4972.2001.
27. Parrish NF, Gao F, Li H, Giorgi EE, Barbian HJ, Parrish EH, Zajic L, Iyer SS, Decker JM, Kumar A, Hora B, Berg A, Cai F, Hopper J, Denny TN, Ding H, Ochsenbauer C, Kappes JC, Galimidi RP, West AP, Bjorkman PJ, Wilen CB, Doms RW, O'Brien M, Bhardwaj N, Borrow P, Haynes BF, Muldoon M, Theiler JP, Korber B, Shaw GM, Hahn BH. 2013. Phenotypic properties of transmitted founder HIV-1. Proc Natl Acad Sci U S A 110:6626–6633. https://doi.org/10.1073/pnas.1304288110.
28. Buonaguro L, Tornesello ML, Buonaguro FM. 2007. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. J Virol 81:10209–10219. https://doi.org/10.1128/JVI.00872-07.
29. Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, Loul S, Ngole EM, Bienvenue Y, Delaporte E, Brookfield JFY, Sharp PM, Shaw GM, Peeters M, Hahn BH. 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science 313:523–526. https://doi.org/10.1126/science.1126531.
30. Neel C, Etienne L, Li Y, Takehisa J, Rudicell RS, Bass IN, Moudindo J, Mebenga A, Esteban A, Van Heuverswyn F, Liegeois F, Kranzusch PJ, Walsh PD, Sanz CM, Morgan DB, Ndjango J-BN, Plantier J-C, Locatelli S, Gonder MK, Leendertz FH, Boesch C, Todd A, Delaporte E, Mpoudi-Ngole E, Hahn BH, Peeters M. 2010. Molecular epidemiology of simian immunodeficiency virus infection in wild-living gorillas. J Virol 84:1464–1476. https://doi.org/10.1128/JVI.02129-09.
31. Tebit DM, Zekeng L, Kaptue L, Gurtler L, Fackler OT, Keppler OT, Herchenroder O, Krausslich HG. 2004. Construction and characterization of an HIV-1 group O infectious molecular clone and analysis of *vpr*- and *nef*-negative derivatives. Virology 326:329–339. https://doi.org/10.1016/j.virol.2004.05.027.
32. Guyader M, Emerman M, Sonigo P, Clavel F, Montagnier L, Alizon M. 1987. Genome organization and transactivation of the human immunodeficiency virus type 2. Nature 326:662–669. https://doi.org/10.1038/326662a0.
33. Menees TM, Muller B, Krausslich HG. 2007. The major 5' end of HIV type 1 RNA corresponds to G456. AIDS Res Hum Retroviruses 23:1042–1048. https://doi.org/10.1089/aid.2006.0275.
34. Haberle V, Stark A. 2018. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol 19:621–637. https://doi.org/10.1038/s41580-018-0028-8.
35. Compe E, Egly JM. 2021. The long road to understanding RNAPII transcription initiation and related syndromes. Annu Rev Biochem 90:193–219. https://doi.org/10.1146/annurev-biochem-090220-112253.
36. Petrenko N, Struhl K. 2021. Comparison of transcriptional initiation by RNA polymerase II across eukaryotic species. Elife 10:e67964. https://doi.org/10.7554/eLife.67964.
37. Chin MP, Rhodes TD, Chen J, Fu W, Hu WS. 2005. Identification of a major restriction in HIV-1 intersubtype recombination. Proc Natl Acad Sci U S A 102:9002–9007. https://doi.org/10.1073/pnas.0502522102.
38. Nikolaitchik OA, Galli A, Moore MD, Pathak VK, Hu WS. 2011. Multiple barriers to recombination between divergent HIV-1 variants revealed by a dual-marker recombination assay. J Mol Biol 407:521–531. https://doi.org/10.1016/j.jmb.2011.01.052.
39. Chen J, Powell D, Hu WS. 2006. High frequency of genetic recombination is a common feature of primate lentivirus replication. J Virol 80:9651–9658. https://doi.org/10.1128/JVI.00936-06.
40. Ryan-Graham MA, Peden KW. 1995. Both virus and host components are important for the manifestation of a Nef- phenotype in HIV-1 and HIV-2. Virology 213:158–168. https://doi.org/10.1006/viro.1995.1556.
41. Wu L, Martin TD, Vazeux R, Unutmaz D, KewalRamani VN. 2002. Functional evaluation of DC-SIGN monoclonal antibodies reveals DC-SIGN interactions with ICAM-3 do not promote human immunodeficiency virus type 1 transmission. J Virol 76:5905–5914. https://doi.org/10.1128/JVI.76.12.5905-5914.2002.
42. Kotsopoulou E, Kim VN, Kingsman AJ, Kingsman SM, Mitrophanous KA. 2000. A Rev-independent human immunodeficiency virus type 1 (HIV-1)-based vector that exploits a codon-optimized HIV-1 *gag-pol* gene. J Virol 74:4839–4852. https://doi.org/10.1128/jvi.74.10.4839-4852.2000.
43. Yee JK, Miyanohara A, LaPorte P, Bouic K, Burns JC, Friedmann T. 1994. A general method for the generation of high-titer, pantropic retroviral vectors: highly efficient infection of primary hepatocytes. Proc Natl Acad Sci U S A 91:9564–9568. https://doi.org/10.1073/pnas.91.20.9564.
44. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.