Heliyon 10 (2024) e39474

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Characterizing transcripts of HIV-1 different substrains using direct RNA sequencing

Weizhen Li [a,b,1], Yong Huang [b,1], Haowen Yuan [c,1], Jingwan Han [b], Zhengyang Li [a,b], Aiping Tong [d], Yating Li [b], Hanping Li [b], Yongjian Liu [b], Lei Jia [b], Xiaolin Wang [b], Jingyun Li [b], Bohan Zhang [b,**], Lin Li [a,b,*]

[a] School of Public Health and Health Management, Gannan Medical University, Ganzhou, Jiangxi, 341000, China
[b] State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences, Beijing, 100071, China
[c] Department of Microbiological Laboratory Technology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, 250012, China
[d] State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University, Chengdu, 610041, China

ARTICLE INFO

Keywords:
HIV-1 transcripts
Alternative splicing
Polyadenylation
RNA modification
Nanopore direct RNA sequencing

ABSTRACT

Post-transcriptional processing and modification of viral RNA, including alternative splicing, polyadenylation, and methylation, play crucial roles in regulating viral gene expression, enhancing genomic stability, and increasing replication efficiency. These processes have significant implications for viral biology and antiviral therapies. In this study, using Oxford Nanopore Technology (ONT) direct RNA sequencing (DRS), we provided a comprehensive analysis of the transcriptome and epitranscriptome features of the HIV-1 B (NL4-3) subtype strain and, for the first time, characterized these features in the CRF01_AE (GX2005002) subtype strain. We identified 11 novel splicing sites among the 61 RNA isoforms in NL4-3 and defined the splicing sites for GX2005002 based on its 63 RNA isoforms. Furthermore, we identified 74 and 79 chemically modified sites in the transcripts of NL4-3 and GX2005002, respectively. Although differences in poly(A) tail length were observed between the two HIV-1 strains, no specific correlation was detected between poly(A) tail length and the number of modification sites. Additionally, three distinct N6-methyladenosine (m^6A) modification sites were identified in both NL4-3 and GX2005002 transcripts. This study provides a detailed analysis of post-transcriptional processing modifications in HIV-1 and suggests promising avenues for future research that could potentially be applied as new therapeutic targets in HIV treatment.

** Corresponding author.

* Corresponding author. School of Public Health and Health Management, Gannan Medical University, Ganzhou, Jiangxi, 341000, China.

E-mail addresses: y900205@163.com (W. Li), yonghuang1987@126.com (Y. Huang), yuanhaowen0218@163.com (H. Yuan), hanjingwan@outlook.com (J. Han), lizhengyang0806@163.com (Z. Li), aipingtong@scu.edu.cn (A. Tong), lyt59790709@163.com (Y. Li), hanpingline@163.com (H. Li), yongjian325@sina.com (Y. Liu), jialeihaowawa@gmail.com (L. Jia), woodsxl@163.com (X. Wang), lijyjk@163.com (J. Li), zbhforjob@163.com (B. Zhang), dearwood@sina.com (L. Li).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e39474
Received 26 May 2024; Received in revised form 7 October 2024; Accepted 15 October 2024
Available online 17 October 2024

1. Introduction

Viral gene expression within host cells is regulated at both transcriptional and post-transcriptional levels, playing crucial roles in the viral life cycle. Through mechanisms such as alternative splicing, polyadenylation, and methylation modifications, viruses can regulate gene expression, increase genomic stability, and enhance replication efficiency [1–3]. Extensive efforts have been made to decipher the molecular mechanisms of viral RNA transcription and post-transcriptional modification, which are important for understanding viral biology and identifying antiviral therapeutic targets [4]. However, the transcriptional regulation of viral genes is highly complex, and many questions remain unanswered [5]. Human immunodeficiency virus (HIV), the causative agent of acquired immune deficiency syndrome (AIDS), exploits multiple strategies to replicate in host cells, serving as an important model for understanding the diversity of post-transcriptional processing and modifications in viruses [6,7].

The HIV-1 genome is approximately 9.2–9.8 kb in length and encodes nine open-reading frames (ORFs) [8]. The transcription of the HIV-1 provirus results in a single type of transcript—full-length genomic RNA—which functions as mRNA for the major structural proteins, such as the Gag and Gag/Pro/Pol precursor polyproteins. Other proteins, including the envelope protein (Env), regulatory proteins (Tat, Rev), and accessory proteins (Vif, Nef, Vpr, Vpu), are encoded by various spliced transcripts. Based on the extent of splicing, HIV-1 mRNAs can be classified into three categories: unspliced, partially spliced, and completely spliced [9]. In the laboratory HIV strain NL4-3, alternative splicing depends on four major splice donors (SDs) and nine splice acceptors (SAs), leading to the production of over 50 different mRNA variants [10]. Along with alternative splicing, RNA 5'-cap methylation and 3'-end polyadenylation occur simultaneously. The dynamic regulation of poly(A) plays a crucial role in eukaryotic gene expression by maintaining mRNA stability and regulating translation [3]. Previous studies have shown that the poly(A) tail participates in mRNA export from the nucleus to the cytoplasm, and its length varies greatly among different species [11–13]. In viruses, polyadenylation is often closely related to viral fate in host cells, as the poly(A) tail typically enhances translation efficiency and promotes mRNA stability [3,7,14]. However, the poly(A) length of HIV-1 transcripts has not yet been demonstrated. Therefore, the dynamic regulation of HIV-1 RNA polyadenylation is crucial for maintaining mRNA integrity [7]. Additionally, RNA modifications can dynamically reshape gene expression and regulate RNA stability, metabolism, splicing, translation, localization, transport, and interactions with other RNAs or RNA-binding proteins [15–17]. This is exemplified by existing studies showing that several RNA methylation modifications directly increase the stability and translation efficiency of HIV-1 RNA or indirectly facilitate viral replication by enabling the virus to evade innate immune recognition of viral transcripts [2,17,18]. Exploring potential sites of modification on HIV-1 RNA can provide targets for antiviral therapeutic drugs and host antiviral immune responses [19].

Although HIV-1 transcripts have been reported previously, all studies were based on mRNA reverse transcription and PCR amplification [10,20,21]. These processes caused numerous chimeras and hindered accurate information retrieval, leading to deviations in evaluating splice variants, base modifications, and single RNA molecule analysis [10,22]. In recent years, the emerging nanopore direct RNA sequencing (DRS) technology has enabled the direct sequencing of long-read RNA without PCR amplification, also allowing for the simultaneous identification of base modifications within the nucleotide sequence [23]. Considering the importance of RNA modifications in host cell and viral RNA regulation, information acquired through DRS is essential for a comprehensive understanding of the HIV-1 lifecycle and pathogenicity. Currently, DRS technology has been successfully applied multiple times to study alternative splicing events in human transcripts and to explore the transcriptional landscape of viruses [22,24].

In this study, to investigate the diversity of HIV-1 transcripts, we employed DRS technology to sequence polyadenylated viral RNA from laboratory variants of the subtype B strain (NL4-3) and the CRF01_AE strain (GX2005002), which is widely prevalent in China and derived from an original HIV isolate, in human lymphoma cells (MT2). We conducted a transcriptomic analysis based on in vitro transcription data to comprehensively compare the detailed characteristics of post-transcriptional processing and modifications in the RNA of these two strains. The findings of this research will contribute to a better understanding of the mechanisms governing post-transcriptional processing and modifications of HIV-1 RNA, deepen our knowledge of the HIV-1 lifecycle, and enrich the transcriptional landscape information concerning prevalent HIV-1 strains in China. These insights will provide new perspectives for blocking HIV replication and developing novel antiviral strategies.

2. Materials and methods

2.1. Infection of MT2 cell

The HIV-1 subtype B strain NL4-3 and subtype CRF01_AE strain GX2005002 used in this study were both stored in our laboratory. NL4-3 was produced by transfecting 293T cells with the pNL4-3 infectious molecular clone of the acquired immunodeficiency syndrome-associated retrovirus [25]. GX2005002 is a primary strain of HIV-1 CRF01_AE subtype isolated in Guangxi Province, China [26]. Infectious virions were detected by tissue culture infectious dose 50 (TCID50), which was calculated by infecting TZM-b1 cells. The viral supernatant was diluted in a 1:3 ratio into seven gradients, with four replicates per gradient. TZM-b1 cells (110,000 cells/mL) were treated with DMEM medium. Each well received 90 μL of cell suspension and 10 μL of the diluted virus, while the negative control group received 10 μL of medium. The cells were incubated at 37 °C and 5 % CO_2 for 48 h. Then, 20 μL of Bright-Glo fluorescence mixture was added to each well. After a 10-min incubation in the dark, relative luminescence units (RLU) were measured. Wells with RLU values higher than the mean RLU of the negative control plus the standard deviation were considered positive. TCID50 was calculated based on these results.

MT2 cells [American Type Culture Collection (ATCC)], a human T-lymphotropic virus type I infected cell line [27], were used for HIV-1 infection. The cells were cultured in RPMI medium 1640 (gibco, cat #22400-089) plus 10%fetal bovine serum FBS (gibco, cat

#10099-141) with added penicillin and streptomycin, and then incubated at 5 % CO_2 and 37 °C for 3 days. MT2 cells (3×10^6) were infected with NL4-3 and GX2005002 virus at a multiplicity of infection (MOI, MOI = TCID50/cell number) of 1 respectively. Four hours after infection, the medium was supplemented to a total volume of 5 ml, and the infected cells and supernatants were collected after culturing for 4, 12, 24, 36, 48, and 60 h, respectively. The cytopathic effect in infected MT2 cells were observed by microscope.

2.2. Total RNA and mRNA isolation

Total RNA was isolated from MT2 cell pellets ($<1 \times 10^7$ cells) using an RNA extraction kit (TAKARA, cat #9767). Cells were lysed using buffer RL + DTT provided in the kit, following the manufacturer's instructions for subsequent processing. Finally, the total RNA was eluted in 55 μl of nuclease-free water. Total concentration of total RNA samples was measured with Qubit fluorometer using the Qubit RNA HS Assay Kit (Invitrogen, cat #Q32852).

mRNA from NL4-3 and GX2005002 strain-infected MT2 cells were separately isolated from cell pellets using an mRNA Isolation Kit (NEB, cat #S1550), and the concentration was measured with the Qubit fluorometer RNA HS Assay (Invitrogen, cat #Q32852).

2.3. HIV-1 viral load and p24 concentration detection

The total RNA was diluted 100-fold and 20 μl of the diluted RNA was used to detect the viral load of HIV-1 with an HIV-1 Nucleic Acid Assay Kit (Daan, cat#0331) followed by qPCR. Additionally, the supernatant of MT2 cells was collected at 4, 12, 24,36, 48 and 60 h after infection. A P24 antigen detection kit (Abcam, cat#ab218268) was used to measure P24 antigen content in the supernatant by ELISA. Both HIV-1 viral load and P24 concentration were measured according to the manufacturer's instructions.

2.4. In vitro transcription

Reverse transcription and PCR were performed with 300 ng of total RNA from NL4-3-infected and GX2005002-infected MT2 cells using the Takara PrimeScript™ One-Step RT-PCR Kit (cat #RR055A) with virus-specific primers (Fig. S1). Templates for in vitro transcription were prepared by RT-PCR, followed by agarose gel purification using the Promega Wizard SV Gel and PCR Clean-Up System (cat #A9282), in vitro transcription with the Vazyme T7 High Yield Transcription Kit (cat #TR101), and RNA purification using the Invitrogen MEGAclear™ transcription purification kit (cat #AM1908). The oligonucleotides used in this study are listed in Table S1.

2.5. Nanopore direct RNA sequencing and data analysis

For nanopore sequencing of HIV-1-infected MT2 cells, 500 ng of mRNA was used for library preparation following the manufacturer's instructions (the Oxford Nanopore DRS protocol, SQK-RNA002). The libraries for all samples were quantified using the Qubit fluorometer DNA HS assay and loaded onto a FLO-MIN106D flow cell (R9.4), followed by a 42-h sequencing run on the MinION device (Oxford Nanopore Technologies).The nanopore direct sequencing data used Guppy (v6.4.6) [28] to generate fastq files. The quality of the data was assessed using Nanoplot (v1.40.0) based on the raw fastq files.

2.6. The splicing events analysis of HIV-1 transcriptome

Reads were mapped directly to the NL4-3 and GX2005002 references using *minimap2* (v2.17) [29,30] with the splice preset, followed by filtering, viewing, and sorting with Samtools (v1.15.1). The resulting sorted BAM file was used as input for splicing site analysis using megadepth software (v1.2.0) (http://github.com/ChristopherWilks/megadepth). Reads were screened for potential SD and SA sites by identifying exon start and end positions from the BAM file analyses. If the locations of the potential sites were unknown, only the sites associated with the canonical GT-AG splicing site pair were annotated. The Integrative Genomics Viewer (IGV, v2.13.1) [10,31] was used for genomics data visualization.

2.7. Poly(A) tail length detection

Raw sequencing data of NL4-3 and GX2005002 were indexed using Nanopolish (v0.14.0). The software was used with default parameters to assess poly(A) tail lengths in all sequenced libraries of NL4-3 and GX2005002. Processed reads were aligned against the NL4-3 and GX2005002 reference genomes using minimap2 with the map-not preset. The Samtools sort module was used to convert the mapped SAM file into a sorted BAM file. The resulting BAM files of the alignments were exported and processed with the *nanopolish polya* module to extract poly(A) length information. The initial step for statistical analysis of the same transcript type in the two strains involved conducting a standard normal distribution test, followed by Mann-Whitney *U* tests on the data that did not pass the normal distribution test. A P-value<0.05 was considered to indicate statistical significance.

2.8. RNA modification data analysis

Raw fast5 and fastq files were indexed using Nanopolish (v0.10.1) for [28] IVT reads and viral RNA reads, respectively. The sequence alignments were further refined by re-aligning the identified signals of viral reads to the viral genome using Nanopolish event align, and redundancy was removed using Nanocompore eventalign_collapse (v1.0.4) [23]. IVT product reads and modified base detection were analyzed by sample-level comparison using the "Nanocompore Sampcomp" mode. Nanocompore Sampcomp produced a table of statistical test results for each kmer. A logistic regression (logit) test was performed on the sample label assignments, with significant P-values ($P < 0.005$) and LOR values ($P > 0.5$) used in conjunction with the GMM p-value to filter for higher probability modifications. The P-value from the Kolmogorov-Smirnov (KS) test was set to <0.005.

2.9. m^{6}A methylation sites detection

The "index" and "eventalign" steps performed in Nanopolish were used for the resquiggling process. Segmented raw signals generated in the previous step and contained in the event align text file were pre-processed with 'm6anet-dataprep', and the predictions of m^{6}A modifications in DRACH motifs were obtained via 'm6anet-run_inference', algorithms implemented in the m6anet program (v2.0.2) [32]. The output file was set to probability modified >0.5 and contrasted with nanocompore software.

3. Results

3.1. DRS sequencing data were sufficient for transcriptome analysis of HIV-1

We infected human lymphoma cells (MT2) with the laboratory HIV-1 subtype B strain (NL4-3) and an original HIV-1 CRF01_AE isolate (GX2005002) (see Methods). To enhance the comprehensiveness of the HIV-1 RNA transcriptome and obtain optimal sequencing data, we measured viral load and p24 concentration at various time points (Fig. S2). Cytopathic effects were observable after 12 h. At 36 h post-infection, both NL4-3 and GX2005002 showed higher viral loads and p24 concentrations compared to other time points (Figs. S2C and D). Therefore, we extracted total polyadenylated mRNAs from the infected MT2 cells 36 h post-infection and

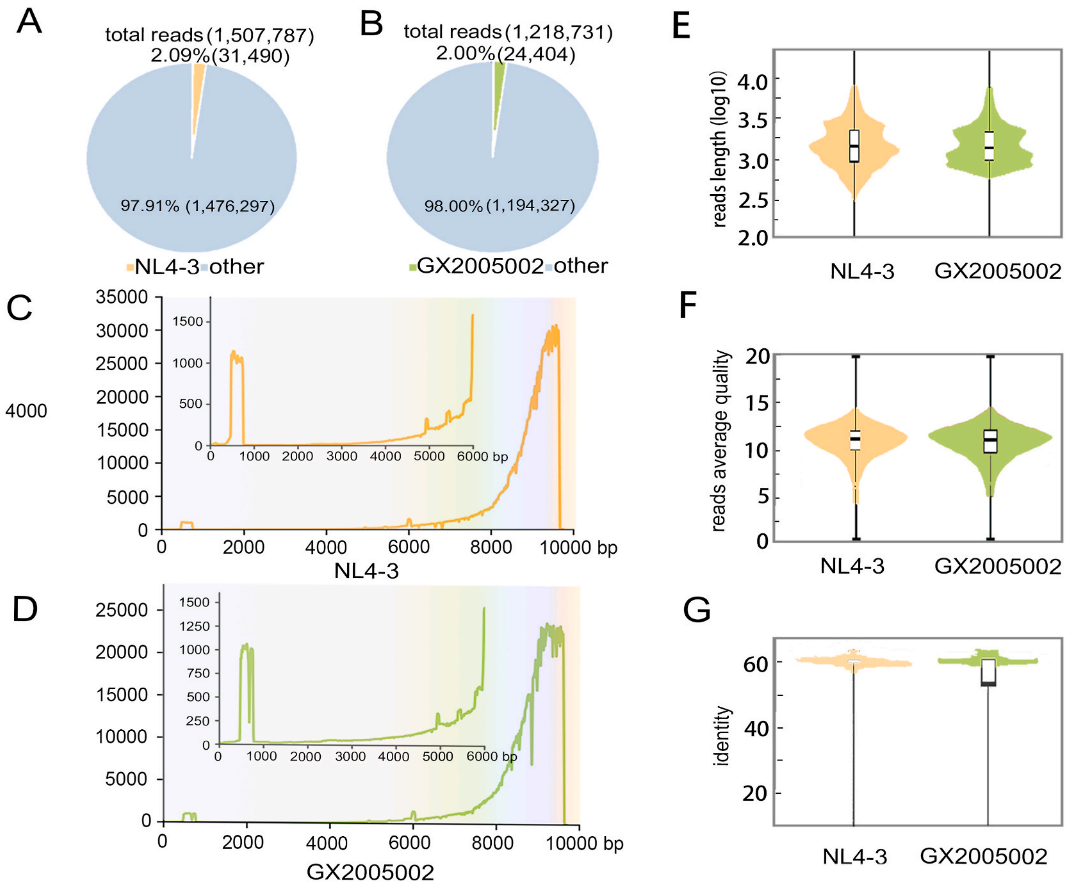

Fig. 1. The statistics and features of ONT direct RNA sequencing data. (A) Read count of total reads from MT2 cells infected with HIV-1 subtype B strain NL4-3. (B) Read counts of total reads from MT2 cells infected with HIV-1 subtype CRF01_AE GX2005002. (C) Genome coverage of NL4-3 sequencing data is shown in Fig. 1A. (D) Genome coverage of sequencing data is shown in Fig. 1B. (E) The distribution of read length of NL4-3 and GX2005002 sequencing data. (F) Average quality values of NL4-3 and GX2005002 sequencing data. (G) Alignment identity of mapped NL4-3 and GX2005002 direct RNA sequencing data.

performed nanopore DRS. A total of 1,507,787 and 1,218,731 raw reads were generated, of which 2.09 % (31,490) and 2.00 % (24,404) reads were mappable to NL4-3 and GX2005002, respectively (Fig. 1A and B). These reads covered the whole genome of the HIV-1 virus, with particularly higher coverage at the 3’ end of both strains compared to other regions (Fig. 1C and D). This reflects the 3’ end bias in nanopore DRS sequencing, as it involves directional sequencing from the 3’ end of RNA. Despite the bias introduced by DRS, the length, average quality, and alignment identity of the mapped reads indicate that the quality and reliability of the raw data are sufficient for subsequent analysis (Fig. 1E, F, G).

3.2. Alternative splicing event analysis of NL4-3 and GX2005002

The reads were mapped to the NL4-3 and GX2005002 reference genomes, and splicing site analysis was performed using megadepth. Reads were screened for potential SD and SA sites by identifying exon start and end positions from BAM files. We observed 61 RNA isoforms from NL4-3 and 63 RNA isoforms from GX2005002 (Fig. 2A, Fig. S3). The splicing occurred at the broadly conserved

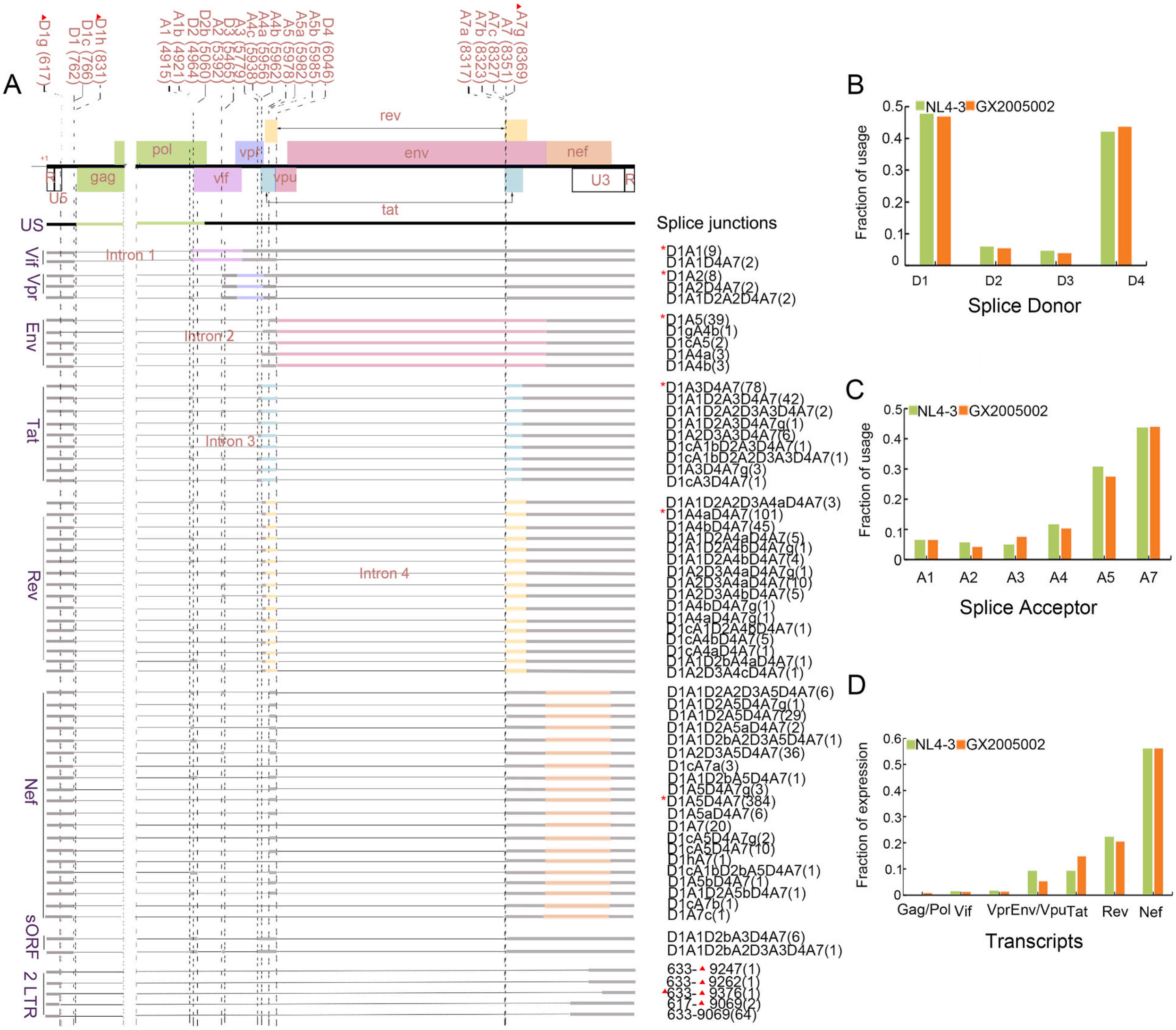

Fig. 2. The schematic representation of GX2005002 transcriptome and gene expression level, splice acceptor/donor usage for NL4-3 and GX2005002. (A) Schematic representation of GX2005002 RNA population detected by ONT sequencing in infected MT2 cells. The ORFs of HIV-1 were color-coded (pink for Env/Vpu, yellow for Nef, light yellow for Rev, blue for Tat, purple for Vif, and light purple for Vpr). Transcripts known or susceptible to encode the same viral proteins were grouped. The label of the red triangle in the splice site means the newly discovered sites in this study. The splice junctions that symbolized with “*” represent the most abundant splicing forms of each protein (with reads number in the parentheses). The splicing donor and acceptor sites as well as their positions were presented in red font. (B) The fraction of splicing donor sites of NL4-3 and GX2005002 calculated based on counts obtained isoforms cluster. (C) The fraction of splicing acceptor sites of NL4-3 and GX2005002 calculated based on counts obtained isoforms cluster. (D) The gene expression level of each protein of NL4-3 and GX2005002.

major splicing donors and acceptors, as well as at several previously reported cryptic loci (Fig. 2A, Fig. S3). Consistent with previous studies, D1 and D4 were the predominant SD sites, with 3970 and 2394 overlapping reads, respectively. Conversely, D2 and D3 were less frequently utilized, with 524 and 698 overlapping reads, respectively (Fig. 2B). Among the SA sites, A7 and A5 were the most abundant (Fig. 2C). Furthermore, eight new splicing sites were identified: SD sites at 617 (D1g), 633, and 831 (D1h), and SA sites at 8369 (A7g), 9,069, 9,247, 9,262, and 9376 bp. Among these, the SD site at 633 and the SA site at 9069 had high coverage, while the others did not. Most of the newly discovered sites involve splicing from the SD site at 633 to downstream 3' end SA sites.

Subsequently, we analyzed the expression levels of different transcripts and found that the expression levels of *nef* transcripts were the highest, followed by *rev*. The expression of *rev* transcripts in NL4-3-infected cells was higher than in GX2005002-infected cells, suggesting inherent differences between the two strains (Fig. 2D). We identified cryptic and unconventional splicing sites within our dataset, including 6 and 5 2-LTR RNAs in the NL4-3 and GX2005002 transcripts, respectively. The discovery of these 2-LTRs supports previous findings that they are naturally transcribed during HIV-1 infection and participate in virus replication [33]. Moreover, we observed several small open reading frames (sORF) in our dataset (Fig. 2A), indicating that low-expression transcripts can be detected by direct RNA sequencing technology [10]. These findings shed new light on the complexity of HIV-1 transcription.

In addition, eleven new potential splicing sites were detected in NL4-3-infected cells (Fig. S3), most of which were located in the UTR region. We detected two new putative splice acceptor (SA) sites in the vicinity of A1 (A1c) and A4 (A4d), respectively. Notably, although these potential SA and SD sites are infrequently used, they were detected in infected MT2 cells, corresponding to some false splicing sites mentioned in previous studies [9]. These findings will need to be validated in future research.

3.3. Detection of epigenetic modification sites of HIV-1 transcripts

Nanopore DRS has been widely used to detect RNA epi-transcriptomic modifications. To unambiguously investigate these modifications, we generated negative control RNAs by in vitro transcription (IVT) of the NL4-3 sequences and performed DRS on these unmodified controls (Fig. 3A). We obtained 2,941,513 reads for the IVT NL4-3 sample and 984,149 reads for the IVT GX2005002 sample, with 95.1 % and 98.3 % of the reads mapping to the NL4-3 and GX2005002 reference sequences, respectively (Fig. 3B and C). These reads covered the entire genome of the HIV-1 virus with high coverage (Fig. 3D and E). Next, the sequenced data from the IVT samples and previously sequenced data from mRNA samples of NL4-3- and GX2005002-infected MT2 cells were used to perform a comparative analysis using nanocompore, based on their electrical signal differences. Areas with low coverage (<30X) were excluded when calculating the modification sites. Seventy-four potential chemical modification sites were detected in NL4-3 transcripts, and 79 potential modification sites were identified in GX2005002 transcripts (Fig. 3F and G). Most modification sites in GX2005002 were ultimately found to be concentrated in the env region (Fig. 3G).

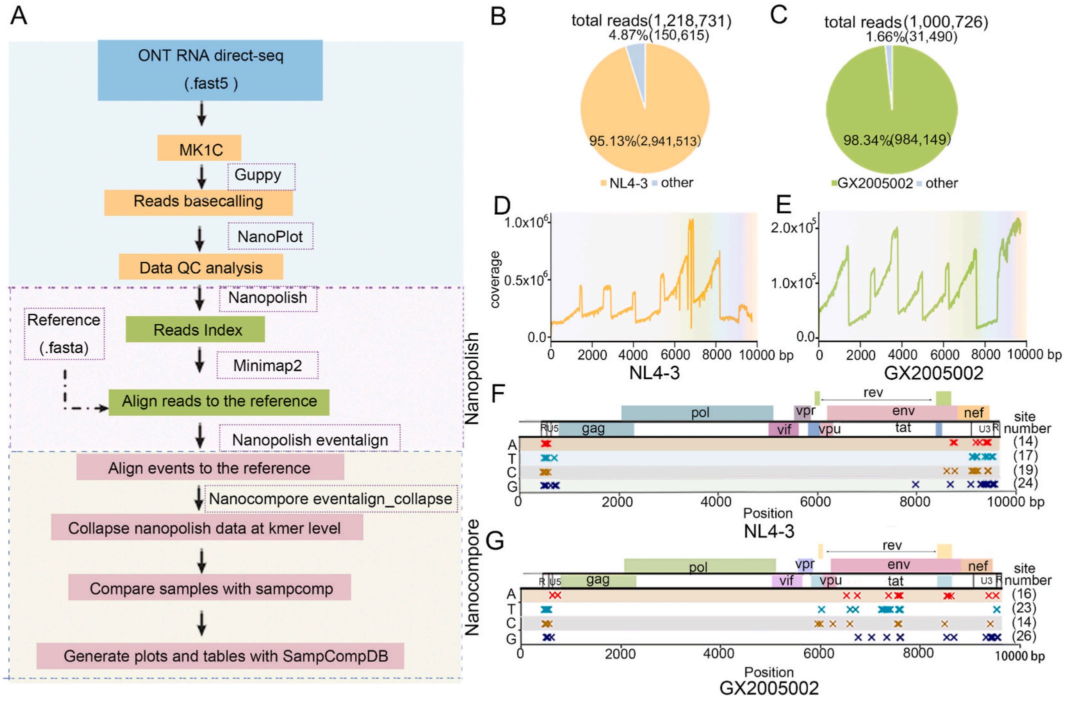

Fig. 3. Epigenetic modification sites of NL4-3 and GX2005002 transcripts. (A) Pipeline and strategies of RNA modification analysis of HIV-1 viruses. (B and C) Read counts of total RNA from in vitro transcription in NL4-3 and GX2005002. (D and E) Genome coverage of direct RNA sequencing data from in vitro transcription samples of NL4-3 and GX2005002. (F and G) Genomic locations of epigenetic modification sites of NL4-3 and GX2005002 transcript. The mark " × " represents the location of each modification site on genome.

3.4. Polyadenylation of mRNA plays an important role in the maturation of HIV-1 transcripts

As nanopore DRS is based on the single-molecule detection of mRNA, it offers a unique opportunity to detect multiple epitranscriptomic features of individual RNA molecules. We measured the poly(A) length distribution (Tables 1 and 2) of NL4-3 and GX2005002 using Nanopolish (v 0.14.0), with the nanopolish *polya* module (Fig. 4A and B). *Nef* mRNA had a relatively shorter poly(A)

Table 1
Average poly(A) length of NL4-3 strain splicing forms of different transcripts.

Proteins	Splicing site	Average poly(A) length
Env/vpu	D1A1D2A5	118
	D1A2D3A5	112
	D1A4a	108
	D1A4b	107
	D1A4c	104
	D1A5*	120
	D1A5a	97
	D1cA5	143
	D1cA5c	55
	D1A2D3A4b	107
Nef	D1A1D2A2D3A5D4A7	110
	D1A1D2A5D4A7	109
	D1A1D2A5aD4A7	21
	D1A1bD2A5D4A7	141
	D1A2D3A5D4A7	125
	D1A2D3A5aD4A7	289
	D1A5D4A7*	107
	D1A5D4aA7	166
	D1A5D4bA7	156
	D1A5aD4A7	119
	D1A7	109
	D1cA1D2A5D4A7	186
	D1cA1bD2A5D4A7	220
	D1cA1cD2A2D3A5D4A7	144
	D1cA5D4A7	89
	D1A5bD4A7	57
Rev	D1A1D2A2D3A4aD4A7	56
	D1A1D2A4aD4A7	131
	D1A1D2A4aD4aA7	76
	D1A1D2A4bD4A7	141
	D1A2D3A4aD4A7	145
	D1A2D3A4bD4A7	112
	D1A2D3A4cD4A7	183
	D1A4aD4A7*	101
	D1A4aD4aA7	102
	D1A4bD4A7	129
	D1A4bD4aA7	220
	D1A4cD4A7	103
	D1cA1D2A2D3A4bD4A7	142
	D1cA2D3A4aD4A7	110
	D1cA4bD4A7	50
	D1cA4dD4A7	131
Tat	D1A1D2A2D3A3D4A7	161
	D1A1D2A3D4A7	116
	D1A1D2bA3	92
	D1A2D3A3D4A7	140
	D1A3D4A7*	125
	D1A3	49
	D1cA1bD2A3D4A7	60
Vif	D1A1D4A7	231
	D1A1*	188
Vpr	D1A2D4A7	282
	D1A2*	135
sORF	D1A1D2bA3D4A7	NA
2 LTR	615–9293	NA
	634–9252	112
	634–9247	110
	555–9162	NA
	617–9069	107
	634–9069[a]	160

[a] The transcript with the most detected splicing forms in different transcripts.

Table 2
Average poly(A) length of GX2005002 strain splicing forms of different transcripts.

Proteins	Splicing site	Average poly(A) length
Env/vpu	D1A4a	108
	D1A4b	192
	D1A5*	146
	D1cA5	140
	D1gA4b	143
Nef	D1A1D2A2D3A5D4A7	193
	D1A1D2A5aD4A7	188
	D1A1D2A5bD4A7	258
	D1A1D2A5D4A7	103
	D1A1D2A5D4A7g	98
	D1A1D2bA2D3A5D4A7	144
	D1A1D2bA5D4A7	249
	D1A2D3A5D4A7	98
	D1A5aD4A7	120
	D1A5bD4A7	78
	D1A5D4A7g	219
	D1A5D4A7*	121
	D1A7	82
	D1cA1bD2bA5D4A7	14
	D1cA5D4A7g	65
	D1cA5D4A7	102
	D1hA7	NA
	D1cA7a	NA
	D1cA7b	31
	D1A7c	227
Rev	D1A1D2A2D3A4aD4A7	213
	D1A1D2A4aD4A7	150
	D1A1D2A4bD4A7	129
	D1A1D2bA4aD4A7	90
	D1A2D3A4aD4A7	129
	D1A2D3A4bD4A7	86
	D1A4aD4A7g	121
	D1A4aD4A7*	126
	D1A4bD4A7	114
	D1cA4bD4A7	125
	D1cA4aD4A7	NA
	D1A1D2A4bD4A7g	NA
	D1A2D3A4aD4A7g	NA
	D1cA1D2A4bD4A7	NA
	D1A2D3A4cD4A7	NA
	D1A4bD4A7g	NA
Tat	D1A1D2A2D3A3D4A7	118
	D1A1D2A3D4A7	118
	D1A1D2A3D4A7g	146
	D1A2D3A3D4A7	121
	D1A3D4A7g	55
	D1A3D4A7*	130
	D1cA1bD2A2D3A3D4A7	159
	D1cA1bD2A3D4A7	52
	D1cA3D4A7	NA
Vif	D1A1*	105
	D1A1D4A7	287
Vpr	D1A1D2A2D4A7	263
	D1A2[a]	115
	D1A2D4A7	220
sORF	D1A1D2bA3D4A7	140
	D1A1D2bA2D3A3D4A7	81
2 LTR	633–9247	NA
	633–9262	NA
	633–9376	NA
	617–9069	156
	633–9069[a]	155

[a] The transcript with the most detected splicing forms in different transcripts.

tail than other transcripts in GX2005002, while *tat* mRNA in NL4-3 had a longer poly(A) tail. Next, the poly(A) tail length of the same transcript in the two strains was statistically analyzed. The results showed that the poly(A) lengths of *env/vpu*, *rev*, and *nef* transcripts in NL4-3 strains were significantly different from those in GX2005002 strains (Fig. 4C). To investigate the correlation between a transcript's poly(A) length and modification accessibility, we sub-grouped the transcripts based on their poly(A) tail length and detected

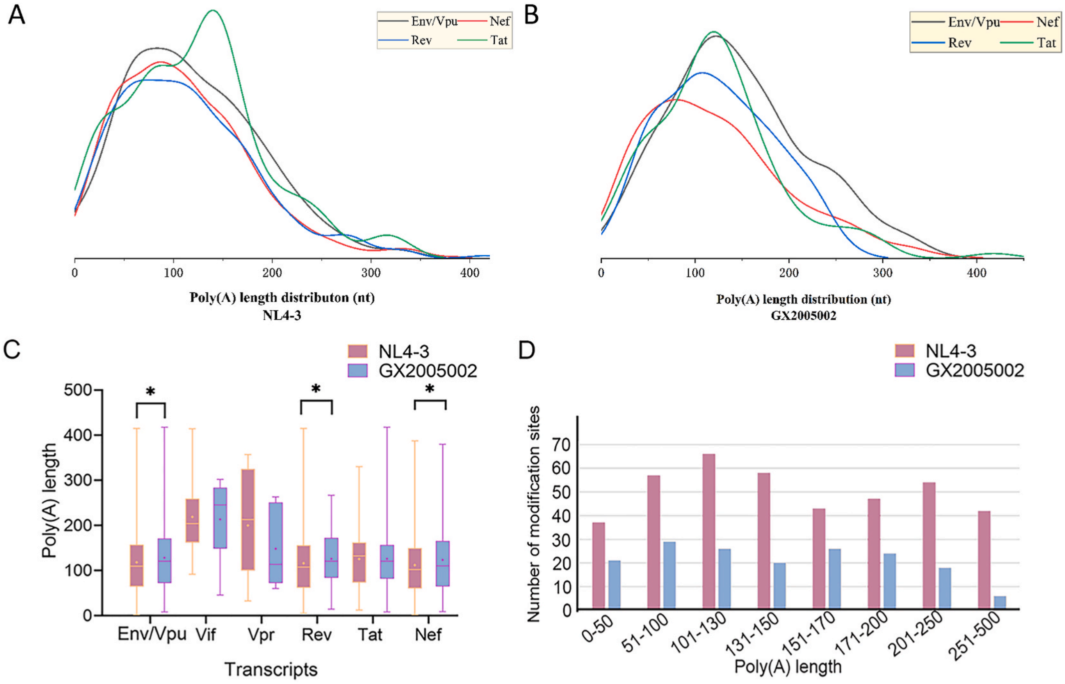

Fig. 4. Poly(A) length of NL4-3 and GX2005002 transcripts. (A and B) The poly(A) tail length distribution of the two viral sub-strains transcripts. (C) box patterns show statistical characteristics of poly(A) tail of different transcripts of NL4-3 and GX2005002. $*P < 0.05$. (D) The number of modifications was calculated by grouping transcripts based on different poly(A) lengths.

the modification numbers in each subgroup. Interestingly, we observed that the various lengths of poly(A) tails did not exhibit significant differences in modification numbers. However, notable distinctions were observed between the NL4-3 and GX2005002 strains (Fig. 4D). Furthermore, we examined whether there was a correlation between the expression abundance of different transcripts and poly(A) tail length, and found no significant correlation (Fig. S4).

3.5. m^{6}A modification analysis of HIV-1 transcripts

Nanocompore detected modifications of all types from the comparison of modified and unmodified samples, resulting in a large number of candidate sites. To specifically analyze m^{6}A modifications in HIV-1 transcripts, m6anet was used to predict m^{6}A modification sites. Sites that were identified in both Nanocompore and m6anet results were selected as m^{6}A modification sites to enhance the accuracy of the determination. There were three predicted m^{6}A modified sites in NL4-3 (A9430, A9458, A9472) and three in GX2005002 (A8517, A8548, A9377) (Fig. 5A and B). Interestingly, their distributions in NL4-3 differed from those in GX2005002. The dwell time and mean intensity differences between IVT and NL4-3/GX2005002-infected cell samples for one of these m^{6}A-modified sites are shown in Fig. 5C and D. Clearly, the dwell time of m^{6}A-modified sites differed from that of unmodified sites. However, the function of the discovered m^{6}A modification sites in both HIV-1 substrains requires further study, as it is of great significance for understanding HIV-1 infection and replication [34].

4. Discussion

In this study, we used nanopore DRS to characterize the transcriptome and epi-transcriptome of the HIV-1 subtype B strain NL4-3 and the subtype CRF_01AE strain GX2005002.Previous studies based on the NGS short platform, which rely on mRNA reverse transcription and PCR amplification, have not accurately provided information about the complete transcript. This limination makes it difficult to assess splice variants, base modifications, and the analysis of individual RNA molecules. Unlike traditional RT-sequencing methods, the DRS method avoids the biases introduced by reverse transcription and amplification. The advantage of long sequencing reads enabled us to comprehensively identify HIV-1 splicing events in both NL4-3 and GX2005002 infected cells. The splicing events identified in NL4-3 are consistent with previous results, with 11 additional novel and rare splice sites, revealing a highly complex landscape of HIV-1 RNA synthesis. Notably, this is the first investigation into the splicing events landscape of GX2005002. The results also highlighted differences in alternative splicing between these two HIV-1 strains. More splicing sites were found in NL4-3 than in GX2005002, indicating a more complex viral replication and protein expression for NL4-3. Overall, nanopore DRS proves to be a powerful tool for studying variable splicing in HIV-1. We observed a high degree of heterogeneity in the coverage of sequencing reads across different regions. Specifically, coverage was high at the 3' end, followed by the 5' end, and relatively low in the middle region.

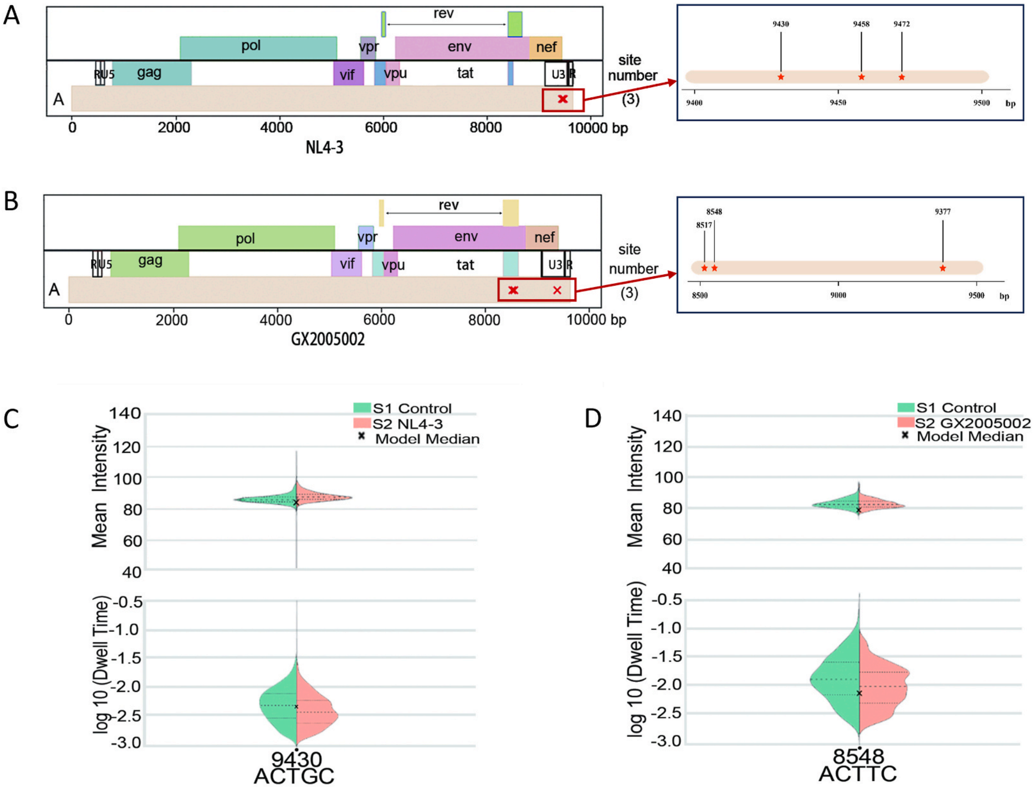

Fig. 5. The information of m6A modification in NL4-3 and GX2005002 transcriptome. (A and B) Prediction sites of m6A modification sites in NL4-3 and GX2005002. Transcripts were detected by nanocompore and m6Anet simultaneously. (C and D) Violin plots showing the dwell time and mean signal intensity of m6A modification sites of NL4-3 (A9340) and GX2005002 (A8548).

This phenomenon can be attributed to the bias in 3' end sequencing and variations in the transcription levels of different transcripts. Furthermore, among the newly discovered potential splicing sites, we noted that some sites were used less frequently, suggesting they may represent rare splicing events.

RNA modification can enhance viral replication and promote virus release [35]. In 2022, Wang et al. summarised the role of RNA modifications in HIV-1 mRNA, indicating their significant impact on the HIV-1 life cycle [20]. Previous studies have mainly focused on using NL4-3 plasmid to transfect 293T cells, investigating the transcriptional characteristics of packaged HIV-1 and host cells. However, the modifications of HIV-1 RNA are much more complex. As a cell model highly sensitive to HIV-1 infection, there is a significant lack of study on the post-transcriptional characteristics of HIV-1-infected MT2 cells. Using MT2 cells as a host has important practical significance. Using nanopore DRS and Nanocompore software, we identified 74 modification sites in NL4-3 transcripts and 79 modification sites in GX2005002 transcripts, respectively. The m^{6}A modification sites identified in the NL4-3 and GX2005002 strains in this study differ from those reported in previous research [36,37]. This discrepancy may be attributed to the expression of HTLV-1 tax in MT2 cells, which could act as a confounding factor affecting HIV-1 modifications. Therefore, the modification sites obtained in this study diverging from those observed in prior studies are acceptable. Unfortunately, Nanocompore could not distinguish modification types, only providing the probability of a nucleotide base being modified. Thus, the total number of modifications can change with adjustments to the threshold. Moreover, the accuracy of Nanocompore depends strongly on signal depth; therefore, low-coverage regions were excluded when predicting modification sites, leading to an uneven distribution of these sites. Furthermore, we measured the poly(A) length distribution of different transcripts to investigate the correlation between poly(A) length and modification accessibility. The number of modifications showed no significant differences along the poly(A) length. However, the average poly(A) length of NL4-3 transcripts was slightly shorter than that of GX2005002 transcripts.

Previous studies have shown that poly(A) tails contribute to transcription, translation, and stability in eukaryotic cells [11]. Similar to host cell transcripts, HIV-1 virus transcripts also possess a 5' cap structure and a 3' poly(A) tail [7], with 3' polyadenylation serving as a signal for transcription termination [38]. Variations in polyadenylation at the 3' end of viral transcripts may influence the stability and translation rate of HIV-1. Studies have found a relationship between changes in poly(A) tail length in bovine coronavirus-infected cells and infection duration and translation efficiency, suggesting that a longer poly(A) tail is conducive to coronavirus translation [39]. Therefore, this study measured the 3' polyadenylation levels of different transcripts in two HIV-1 strains to explore differences in poly(A) length between the strains. The results showed that the distribution of poly(A) length varied between the two strains. Additionally, the poly(A) lengths of the *env/vpu*, *rev*, and *nef* transcripts differed between the strains, which may affect the translation and abundance of related proteins. Detecting the poly(A) length of different HIV-1 strains provides insights for further analysis and understanding of HIV-1 replication. Furthermore, based on previous studies on the relationship between modified and unmodified poly(A) length in the COVID-19 virus [40], we explored whether there is a correlation between poly(A) length and the number of modification sites on modified HIV-1 transcripts. We aimed to determine if transcripts with more modification sites have shorter poly (A) tails. However, we found no significant difference in the number of modification sites distributed along the poly(A) length.

Despite the observed differences in splicing sites, it is noteworthy that the m^{6}A modification sites we identified are predominantly located near the 3' ends of the UTR region in NL4-3, consistent with previous studies [41]. In contrast, the GX2005002 strain had only one modification site detected in the UTR region, indicating a significant difference in RNA modification sites between the two strains. The discovery of these sites provides a new foundation for further research into m^{6}A modifications in HIV-1. To gain a deeper understanding of HIV-1, comparative and functional studies of the m^{6}A modification sites in the two strains should be conducted. In conclusion, our research offers valuable resources and directions for investigating the mechanisms of HIV-1 infection and replication. Transcripts with various splicing donors and acceptors may share certain regions. When modification sites are located in these regions, it becomes challenging to specify the exact transcript with the m^{6}A modification. Deep sequencing and experimental verification are necessary to facilitate the detection of epi-transcriptome modifications.

NGS-based RNA-seq requires reverse transcription and PCR amplification of the template RNA before sequencing, introducing various biases [42]. Pacific Biosciences single-molecule sequencing technologies enable sequencing of the entire transcript in a single read, allowing for clear identification of expressed genes and subtypes [43,44]. However, this approach still requires PCR and/or reverse transcription [45]. DRS provides a new opportunity for understanding HIV-1 RNAs [46]. This method has been successfully applied to detect the architecture of the SARS-CoV-2 transcriptome [40]. In this study, nanopore DRS was used to sequence native RNA molecules, exploring the transcriptomes and post-transcriptional processing modifications of HIV-1 subtype B (NL4-3) and CRF_01AE strains (GX2005002). The DRS method simplifies the library preparation process by directly performing RNA sequencing, thereby avoiding biases introduced by reverse transcription and PCR amplification of RNA samples. However, it is important to note that DRS technology is not suitable for degraded RNA samples. This study provides a new approach to understanding the HIV-1 life cycle and offers new opportunities for targeted HIV-1.

5. Conclusions

In summary, we employed the nanopore DRS method to directly investigate the post-transcriptional processing and modification of distinct HIV-1 strains. This approach effectively identified various RNA isoforms and novel splicing sites, quantified poly(A) length, and revealed RNA modification information. These findings offer new insights into the comprehensive understanding of the HIV-1 virus.

CRediT authorship contribution statement

Weizhen Li: Writing – original draft, Software, Formal analysis, Data curation. **Yong Huang:** Writing – original draft, Software. **Haowen Yuan:** Visualization. **Jingwan Han:** Resources. **Zhengyang Li:** Resources. **Aiping Tong:** Resources. **Yating Li:** Resources. **Hanping Li:** Supervision. **Yongjian Liu:** Methodology. **Lei Jia:** Formal analysis. **Xiaolin Wang:** Investigation. **Jingyun Li:** Supervision. **Bohan Zhang:** Writing – review & editing, Supervision, Methodology. **Lin Li:** Supervision, Resources, Project administration, Conceptualization.

Ethics approval

Not applicable.

Data availability statements

Sequencing data have been submitted to the National Genomics Data Center (NGDC) database and the accession number can be found at https://ngdc.cncb.ac.cn/gsa-human/, HRA005545.

Funding statement

This study was supported by the National Natural Science Foundation of China [82173583], the State Key Laboratory of Pathogen and Biosecurity (AMMS).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:Lin Li reports financial support was provided by the State Key Laboratory of Pathogen and Biosecurity (AMMS). Lin Li reports financial support was provided by National Natural Science Foundation of China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We thank Professor Shan Lu of the University of Massachusetts Medical School for the donation of MT2 cells.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e39474.

References

[1] Y. Lee, D.C. Rio, Mechanisms and regulation of alternative pre-mRNA splicing, Annu. Rev. Biochem. 84 (2015) 291–323.
[2] S. Chen, S. Kumar, C.E. Espada, et al., N6-methyladenosine modification of HIV-1 RNA suppresses type-I interferon induction in differentiated monocytic cells and primary macrophages, PLoS Pathog. 17 (2021) e1009421.
[3] C.J. Passmore LA, Roles of mRNA poly(A) tails in regulation of eukaryotic gene, Nat. Rev. Mol. Cell Biol. 23 (22) (2022) 93–106.
[4] R.A. Varier, T.K. Kundu, Chromatin modifications (acetylation/deacetylation/methylation) as new targets for HIV therapy, Curr. Pharmaceut. Des. 12 (2006) 1975–1993.
[5] E. Vicenzi, G. Poli, Regulation of HIV expression by viral genes and cytokines, J. Leukoc. Biol. 56 (1994) 328–334.
[6] J.M. Sung, D.M. Margolis, HIV persistence on antiretroviral therapy and barriers to a cure, Adv. Exp. Med. Biol. 1075 (2018) 165–185.
[7] J. Gao, Y.D. Tang, W. Hu, et al., When poly(A) binding proteins meet viral infections, including SARS-CoV-2, J. Virol. 96 (2022) e0013622.
[8] J. Nkeze, L. Li, Z. Benko, et al., Molecular characterization of HIV-1 genome in fission yeast Schizosaccharomyces pombe, Cell Biosci. 5 (2015) 47.
[9] A. Emery, R. Swanstrom, HIV-1: to splice or not to splice, that is the question, Viruses 13 (2021).
[10] N. Nguyen Quang, S. Goudey, E. Segeral, et al., Dynamic nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection, Retrovirology 17 (2020) 25.
[11] M. Edmonds, M.H. Vaughan Jr., H. Nakazato, Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship, Proc Natl Acad Sci U S A 68 (1971) 1336–1340.
[12] C.S. McLaughlin, J.R. Warner, M. Edmonds, et al., Polyadenylic acid sequences in yeast messenger ribonucleic acid, J. Biol. Chem. 248 (1973) 1466–1471.
[13] J. Jia, W. Lu, B. Liu, et al., An atlas of plant full-length RNA reveals tissue-specific and monocots-dicots conserved regulation of poly(A) tail length, Nat. Plants 8 (2022) 1118–1126.
[14] A.O. Subtelny, S.W. Eichhorn, G.R. Chen, et al., Poly(A)-tail profiling reveals an embryonic switch in translational control, Nature 508 (2014) 66–71.
[15] H. Shi, J. Wei, C. He, Where, when, and how: context-dependent functions of RNA methylation writers, readers, and erasers, Mol Cell 74 (2019) 640–650.
[16] B.S. Zhao, I.A. Roundtree, C. He, Post-transcriptional gene regulation by mRNA modifications, Nat. Rev. Mol. Cell Biol. 18 (2017) 31–42.
[17] S. N'Da Konan, E. Segeral, F. Bejjani, et al., YTHDC1 regulates distinct post-integration steps of HIV-1 replication and is important for viral infectivity, Retrovirology 19 (2022) 4.
[18] S. Riquelme-Barrios, C. Pereira-Montecinos, F. Valiente-Echeverría, et al., Emerging roles of N(6)-methyladenosine on HIV-1 RNA metabolism and viral replication, Front. Microbiol. 9 (2018) 576.
[19] S. Wang, H. Li, Z. Lian, et al., The role of RNA modification in HIV-1 infection, Int. J. Mol. Sci. 23 (2022).
[20] E. Delgado, C. Carrera, P. Nebreda, et al., Identification of new splice sites used for generation of rev transcripts in human immunodeficiency virus type 1 subtype C primary isolates, PLoS One 7 (2012) e30574.
[21] K.E. Ocwieja, S. Sherrill-Mix, R. Mukherjee, et al., Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing, Nucleic Acids Res. 40 (2012) 10345–10355.
[22] S. Mitsuhashi, S. Nakagawa, M. Sasaki-Honda, et al., Nanopore direct RNA sequencing detects DUX4-activated repeats and isoforms in human muscle cells, Hum. Mol. Genet. 30 (2021) 552–563.
[23] A. Leger, P.P. Amaral, L. Pandolfini, et al., RNA modifications detection by comparative Nanopore direct RNA sequencing, Nat. Commun. 12 (2021) 7198.
[24] M.T. Parker, K. Knop, A.V. Sherwood, et al., Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m(6)A modification, Elife 9 (2020).
[25] A. Adachi, H.E. Gendelman, S. Koenig, et al., Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone, J. Virol. 59 (1986) 284–291.
[26] J. Han, S. Liu, W. Guo, et al., Development of an HIV-1 subtype panel in China: isolation and characterization of 30 HIV-1 primary strains circulating in China, PLoS One 10 (2015) e0127696.
[27] W. Zhou, D. Zhang, X. Jia, [Apoptosis in MT2 cells induced by HepG2.2.15 cells], Zhonghua gan zang bing za zhi = Zhonghua ganzangbing zazhi = Chinese journal of hepatology 7 (1999) 34–35.
[28] R.R. Wick, L.M. Judd, K.E. Holt, Performance of neural network basecalling tools for Oxford Nanopore sequencing, Genome Biol. 20 (2019) 129.
[29] C.M. Gallardo, A.T. Nguyen, A.L. Routh, et al., Selective ablation of 3` RNA ends and processive RTs facilitate direct cDNA sequencing of full-length host cell and viral transcripts, Nucleic Acids Res. 50 (2022) e98.
[30] H. Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics 34 (2018) 3094–3100.
[31] J.T. Robinson, H. Thorvaldsdottir, W. Winckler, et al., Integrative genomics viewer, Nat. Biotechnol. 29 (2011) 24–26.
[32] C. Hendra, P.N. Pratanwanich, Y.K. Wan, et al., Detection of m6A from direct RNA sequencing using a multiple instance learning framework, Nat. Methods 19 (2022) 1590–1598.
[33] A. Brussel, P. Sonigo, Evidence for gene expression by unintegrated human immunodeficiency virus type 1 DNA species, J. Virol. 78 (2004) 11263–11271.
[34] X. Wang, Z. Lu, A. Gomez, et al., N6-methyladenosine-dependent regulation of messenger RNA stability, Nature 505 (2014) 117–120.
[35] R. Kumar, D. Mehta, N. Mishra, et al., Role of host-mediated post-translational modifications (PTMs) in RNA virus pathogenesis, Int. J. Mol. Sci. 22 (2020).
[36] E.M. Kennedy, et al., Post-transcriptional m6 A editing of HIV-1 mRNAs enhances viral gene expression, Cell Host Microbe 19 (2016) 675–685.
[37] N. Tirumuru, et al., N6 -methyladenosine of HIV-1 RNA regulates viral infection and HIV-1 Gag protein expression, Elife 5 (2016) e15528.
[38] Y. Huang, G.G. Carmichael, Role of polyadenylation in nucleocytoplasmic transport of mRNA, Mol. Cell Biol. 16 (1996) 1534–1542.
[39] H.Y. Wu, T.Y. Ke, W.Y. Liao, et al., Regulation of coronaviral poly(A) tail length during infection, PLoS One 8 (2013) e70548.
[40] D. Kim, J.Y. Lee, J.S. Yang, et al., The architecture of SARS-CoV-2 transcriptome, Cell 181 (2020) 914–921.e910.
[41] E.M. Kennedy, H.P. Bogerd, A.V. Kornepati, et al., Posttranscriptional m(6)A editing of HIV-1 mRNAs enhances viral gene expression, Cell Host Microbe 19 (2016) 675–685.
[42] D. Aird, M.G. Ross, W.S. Chen, et al., Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries, Genome Biol. 12 (2011) R18.
[43] D. Sharon, H. Tilgner, F. Grubert, et al., A single-molecule long-read survey of the human transcriptome, Nat. Biotechnol. 31 (2013) 1009–1014.

[44] J.L. Weirather, M. de Cesare, Y. Wang, et al., Comprehensive Comparison of Pacific Biosciences and Oxford Nanopore Technologies and Their Applications to Transcriptome Analysis, 2017, p. 6.

[45] J. Gleeson, A. Leger, Y.D.J. Prawer, et al., Accurate expression quantification from nanopore direct RNA sequencing with NanoCount, Nucleic Acids Res. 50 (2022) e19.

[46] D.R. Garalde, E.A. Snell, D. Jachimowicz, et al., Highly parallel direct RNA sequencing on an array of nanopores, Nat. Methods 15 (2018) 201–206.