Published online 23 June 2022

Nucleic Acids Research, 2022, Vol. 50, No. 17 ***e98***
https://doi.org/10.1093/nar/gkac516

Selective ablation of 3′ RNA ends and processive RTs facilitate direct cDNA sequencing of full-length host cell and viral transcripts

Christian M. Gallardo [1,2], Anh-Viet T. Nguyen[1], Andrew L. Routh [3,4] and Bruce E. Torbett [1,2,5,6,7,*]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA, [2]Center for Immunity and Immunotherapies, Seattle Children's Research Institute, Seattle, WA 98101, USA, [3]Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX 77555, USA, [4]Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, TX 77555, USA, [5]Institute for Stem Cell & Regenerative Medicine, University of Washington, Seattle, WA 98109, USA, [6]Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, Seattle, WA 98101, USA and [7]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA 98101, USA

Received January 31, 2022; Revised April 25, 2022; Editorial Decision May 27, 2022; Accepted June 01, 2022

ABSTRACT

Alternative splicing (AS) is necessary for viral proliferation in host cells and a critical regulatory component of viral gene expression. Conventional RNA-seq approaches provide incomplete coverage of AS due to their short read lengths and are susceptible to biases and artifacts introduced in prevailing library preparation methodologies. Moreover, viral splicing studies are often conducted separately from host cell transcriptome analysis, precluding an assessment of the viral manipulation of host splicing machinery. To address current limitations, we developed a quantitative full-length direct cDNA sequencing strategy to simultaneously profile viral and host cell transcripts. This nanopore-based approach couples processive reverse transcriptases with a novel one-step chemical ablation of 3′ RNA ends (termed CASPR), which decreases ribosomal RNA reads and enriches polyadenylated coding sequences. We extensively validate our approach using synthetic reference transcripts and show that CASPR doubles the breadth of coverage per transcript and increases detection of long transcripts (>4 kb), while being functionally equivalent to PolyA+ selection for transcript quantification. We used our approach to interrogate host cell and HIV-1 transcript dynamics during viral reactivation and identified novel putative HIV-1 host factors containing exon skipping or novel intron retentions and delineated the HIV-1 transcriptional state associated with these differentially regulated host factors.

INTRODUCTION

Alternative splicing (AS) greatly increases protein diversity encoded by the human genome, and has been estimated to occur in up to 95% of genes with multiexonic transcripts (1). This process is tightly regulated by *cis*- and *trans*-acting elements, chromatin accessibility and other signaling pathways (2). AS has been shown to be a driver of human proteome diversity (3,4) and a critical regulatory component in the tissue-specific expression of human transcriptomes (5). Recently, increasing use of massively parallel RNA-seq pipelines has allowed population-scale transcriptome studies that have revealed naturally occurring variants that modulate AS and influence disease susceptibility (6).

Viral infections commonly alter the host cell splicing landscape, as shown by genes that appear differentially spliced upon viral infection in transcriptomic studies or splicing-related genes that appear differentially enriched or phosphorylated in proteomic studies (7). In cells infected with HIV-1 (HIV), alternatively spliced host cell transcripts have been shown to promote a permissive environment for viral activation and proliferation via induction of alternative transcription start/end sites (8) and via functional enrichment of HIV replication-related pathways (9). Similarly, proteomic studies have shown induction of signaling pathways involved in mRNA splicing in T lymphocytes upon HIV entry (10), with phosphorylation of canonical splice factors being the apparent regulatory mechanism. Additionally, splicing-related host factors have been reported that bind HIV accessory proteins and act as *trans*-regulatory elements, including the binding of U2AF65 and SPF45 by Rev (11) and SR proteins by Vpr (12), as well as the interactions between POLR2A and Tat (13).

*To whom correspondence should be addressed. Tel: +1 206 884 1140; Email: betorbet@uw.edu

AS is also a critical regulatory mechanism of viral gene expression (14). In HIV, a single unspliced 9.2-kb RNA serves as both the genome and mRNA for both Gag and Gag–Pol polyproteins, while alternatively spliced mRNA variants code for the seven remaining gene products by dynamically and specifically interacting with regulatory elements, thereby generating over 50 physiologically relevant transcripts that can be grouped in 'partially spliced' (4 kb) and 'completely spliced' (1.8 kb) groups (15–18). The underlying mechanism in AS regulation of HIV transcripts is the placement of the open-reading frames (ORFs) of each gene in close proximity to the 5′ cap of HIV RNA, thus optimizing the coding potential of HIV genes by translating different proteins from a common mRNA. The completely spliced 1.8 kb class is particularly important during the early infection phase, and it includes Tat and Rev transcripts that, respectively, aid in transcription and export of partially spliced transcripts from the nucleus. An eventual shift in splicing dynamics, partially attributed to Rev, results in increased production of partially spliced and unspliced mRNAs (11). Thus, carefully orchestrated splicing dynamics are critical for regulating the dynamics of HIV gene expression and any resulting interactions with host factors.

Conventional RNA-seq approaches, while robust and reproducible, are limited by their read length in providing full coverage of AS events (such as alternate donor/acceptor sites, exon skipping, alternate exon usage and intron retention). Moreover, library preparation techniques can introduce biases/artifacts due to PCR amplification bias, artifactual recombination, fragmentation or targeted enrichment methods for coding sequences (CDS) (19). The read-length limitation in short-read RNA-seq, coupled with the biases and artifacts introduced in prevailing library preparation methodologies, can prevent an assessment of full-exon connectivity in a quantitative manner, resulting in loss of information on transcript isoform diversity, including splice variants (20). The limitations of current RNA-seq approaches are particularly exacerbated when assessing transcript expression in polycistronic HIV RNA where all transcripts are flanked by identical 5′ and 3′ end exons (only varying in their internal splicing sites) and vary greatly in overall transcript length. Previous attempts to address these constraints have used primer sets for each transcript class or gene product, relied on molecular barcoding or used emulsion PCR to ameliorate PCR skewing or sampling biases (18,21). However, use of different primer sets prevents the quantitative comparison between transcripts and does not provide full exon coverage, while molecular barcoding approaches were used with short-read next-generation sequencing (NGS) approaches. Previous HIV splicing studies were not implemented within the context of a host cell transcriptome analysis, precluding a direct assessment of the viral manipulation of host splicing machinery or further insights into virus–host interaction dynamics (22). Since the regulation of HIV gene expression depends on the ability of the virus to co-opt host cell splicing machinery, understanding host cell transcriptional state and its resulting HIV mRNA splicing signature would identify novel molecular signatures of HIV infection and provide opportunities for drug/probe development based on novel viral–host factor interactions.

To address current RNA-seq limitations, we developed and validated a quantitative full-length RNA-seq strategy for the simultaneous profiling of polyadenylated host and viral transcripts from unamplified cDNA. This nanopore sequencing-based approach is supported by use of processive reverse transcriptases (RTs) and oligo-d(T) priming, coupled with a one-step chemical ablation of 3′ RNA ends (named CASPR, Chemical Ablation of Spuriously Priming RNAs), which decreases ribosomal RNA (rRNA) reads and enriches polyadenylated transcripts. We validate both RT conditions and CDS enrichment strategies using synthetic reference transcripts and show that while CASPR is functionally equivalent to PolyA+ selection for transcript quantification purposes, it provides critical advantages in doubling the breadth of coverage per transcript and significantly increasing the efficiency of capture of long transcripts >4 kb in size. This improved practical throughput and likelihood of capturing full-exon connectivity. We then demonstrate the utility of our pipeline by interrogating host cell and HIV transcript dynamics in reactivated J-Lat 10.6 cells, a widely used cell line model of HIV reactivation (23,24). We identify putative host factor correlates of HIV transcriptional reactivation that contain exon skipping events (PSAT1) or novel intron retentions (PSD4) and delineate the HIV transcriptional state associated with these differentially regulated host factors. We anticipate that this pipeline will allow greater insights into host cell–pathogen transcript dynamics involved in viral infection and activation.

MATERIALS AND METHODS

Cell culture

J-Lat 10.6 cells, a Jurkat-derived cell line that is latently infected with HIV (23), were obtained from the NIH AIDS Reagent Program (clone #10.6, from Dr. Eric Verdin). The J-Lat 10.6 clone contains a single R7/ΔEnv strain integrated into the SEC16A locus, and EGFP inserted into the nef ORF. For control experiments, the Jurkat E6-1 clone was obtained from the NIH AIDS Reagent Program [cat #177, from Dr. Arthur Weiss (25)]. J-Lat 10.6 cells were activated with 10 ng/ml TNF-α (PeproTech, 300-01A) for 24 h, which induces latency reversal of integrated provirus, resulting in positive GFP expression and p24 production, which are, respectively, detected via flow cytometry and p24 ELISA. Cell lines were maintained in RPMI 1640 (Life Tech) supplemented with 10% fetal bovine serum (Hyclone) and 1% penicillin/streptomycin at 37°C and 5% CO_2.

Total RNA isolation

Total RNA was isolated from cell pellets ($<1 \times 10^7$ cells) using the RNeasy Mini Kit (QIAGEN, cat #74134). Cells were lysed with RLT buffer (with no β-ME) and processed according to manufacturer's instructions, and eluted in 25–50 μl nuclease-free water. Total RNA sample quality was assessed via Agilent Bioanalyzer using the RNA 6000 Nano Kit, resulting in RNA integrity (RIN) scores of 10, for all samples.

Synthetic RNA reference standards

Spike-in RNA variants (SIRVs) that include Iso Mix E0, ERCC and long SIRV modules were purchased from Lexogen (SIRV-Set 4, 141.03). SIRVs were resuspended in 10 μl nuclease-free water to a concentration of 5.35 ng/μl. Resuspended SIRVs were then admixed into total RNA preparations prior to PolyA+ selection, CASPR treatment or reverse transcription, to a concentration of 0.13 ng of SIRVs per μg of total RNA sample.

Chemical Ablation of 3' RNA ends (CASPR)

Sodium periodate ($NaIO_4$) was purchased from Millipore Sigma (311448-5G). A 2× buffered periodate solution (BP) was prepared fresh each time by measuring $NaIO_4$ powder and resuspending to a concentration of 4 mg/ml in aqueous solution of 200 mM sodium acetate (pH 5.5) (Invitrogen, AM9740). Input RNA (up to 5 μg) was mixed with an equal volume of 2× BP and incubated at room temperature in the dark for 30 min. Following treatment, RNA was cleaned with RNA Clean & Concentrator (Zymo Research, R1013) according to the manufacturer's instructions, and eluted in nuclease-free water.

PolyA selection

Polyadenylated transcripts were enriched from total RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490S), according to the manufacturer's instructions.

Reverse transcription and second-strand synthesis

Reverse transcription was carried out with SuperScript IV Reverse Transcriptase (SSIV RT; Thermo Fisher, 18090010) or MarathonRT (MRT; Kerafast, EYU007). Reactions were carried out in a 20 μl volume with the following components and final concentrations: 1× reaction buffer, dNTPs (0.5 mM), RNAseOUT (2 U/μl), oligo-d(T) primer (1 μM) or 4609-bp gene-specific primer (0.1 μM), 5 mM DTT [for SSIV only], RNA input (<5 μg), and MRT (0.5 μM or 20 U) or SSIV RT (200 U). Primers were initially annealed to template RNA in the presence of dNTPs, by heating to 65°C for 5 min, followed by snap cooling to 4°C for 2 min. After snap cooling, the rest of the components were added, followed by reverse transcription for 1.5 h at 42°C for MRT and 50°C for SSIV. Reactions were stopped by heat inactivation at 85°C for 5 min. Second-strand synthesis was carried out using a modified Gubler and Hoffman procedure (26) adapted from Invitrogen's A48570 kit, in a single-pot format involving direct addition of second-strand buffer, dNTPs, *Escherichia coli* DNA Pol I, RNAse H and *E. coli* DNA ligase to the heat-inactivated first-strand reaction. Second-strand synthesis was carried out at 16°C for 2 h, followed by DNA Clean with the Monarch Kit for downstream processing. Verification of yield and quality of cDNA was done via NanoDrop spectrometry, and by running on an 0.8% E-Gel NGS and imaged using Azure c600 (Azure Biosystems).

Nanopore sequencing

All samples were barcoded with Native Barcoding Kit (EXP-NBD104) prior to nanopore library preparation using the Ligation Sequencing Kit (SQK-LSK109). All samples were sequenced with MinION R9.4.1 flow cells, basecalled with Guppy basecaller 3.4.5 and demultiplexed with Guppy barcoder.

Reference sequences

A custom rRNA reference file was created by concatenating the fasta sequences for 28S (gene ID: 100008589), 5.8S (gene ID: 100008587), 5S (gene ID: 100169751) and 18S (gene ID: 100008588) rRNA sequences. LncRNA transcripts in fasta format were downloaded from Gencode release 31 (GRCh38.p12). For human reference alignment, the UCSC analysis set of December 2013 human genome (GCA_000001405.15) without the alt-scaffolds was used along with its associated gtf annotation file when appropriate. A custom reference sequence for R7 viral strain present in J-Lat cells was generated by extracting mapped reads from previous HIV alignments, size filtering, assembling with Unicycler (https://github.com/rrwick/Unicycler), polished with Medaka and manually inspected with SnapGene against HXB2 originating background sequence to rule out structural variants.

Determination of uniquely mapped reads

Reads were mapped to rRNA reference using *minimap2* with *map-ont* preset. Unmapped reads were extracted from the sam output using *samtools view* followed by conversion to fastq using *samtools bam2fq* (27). Fastq file containing unmapped rRNA reads was mapped to lncRNA reference with *minimap2* using *splice* preset, followed by extraction of unmapped reads and conversion to fastq as before. Unmapped lncRNA reads were remapped to human reference with *minimap2* using *splice* preset. Uniquely mapped reads were counted for each resulting sam file using *samtools view* with *-F260* flag to only count primary alignments and the *-c* option to output number of reads.

Gene body coverage, splice junction number and read distribution

For gene body coverage calculation (28), reads were mapped directly to hg38 analysis set reference using *minimap2* with *splice* preset and *–secondary=no* flag, with mapped reads converted to bam format, sorted and indexed using *samtools*. Gene body coverage was calculated with the geneBody_coverage.py script that is part of the RSeQC package (v3.0.1) using sorted and indexed bam files and the UCSC RefSeq (refGene) annotations in bed format. Splice junction quantification and saturation were calculated using the *junction_saturation.py* script, also within RSeQC package, and with identical inputs as before. For intragenic and intergenic read distributions, reads were mapped and processed as before using the gencode v31 human reference (GRCh38.p12). The comprehensive genome annotation gtf file was collapsed using GTEx collapse annotation script. Read distributions were computed from mapped

reads and collapsed annotations using RNA-SeQC (v2.3.4) with the following options: *–unpaired –coverage –base-mismatch=180 –mapping-quality 0 –detection-threshold=0 –legacy*.

Statistical analysis

Where indicated, *t*-tests were run between CASPR and PolyA selected samples within RT enzyme group (either MRT or SSIV). Analyses were performed within GraphPad Prism 8, assuming all rows are sampled from populations with same scatter (SD). Statistical significance was determined using the Holm–Sidak method, with $\alpha = 0.05$. Statistical significance was denoted as follows: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$.

HIV isoform analysis

Reads were mapped to R7 reference sequence with *minimap2* using *splice* preset, followed by filtering using *-F260* flag in *samtools view* and sorting. The resulting sorted bam file was used as input for the Pinfish pipeline (https://github.com/nanoporetech/pinfish). Briefly, bam files were used as input for *spliced_bam2gff* command using the *-M* option. The resulting gff file was clustered into isoform bins using *cluster_gff* command using the following options: *-c 3 -p 0*. Isoforms clusters were then polished using polish_clusters command with *-c 3* option. Polished clusters in fasta format were remapped to reference using minimap2 and processed using same settings as before. Polished clusters were visualized at this stage using IGV 2.7.2, and coverage maps for clustered isoforms were obtained with the *samtools depth* command with the *-a -d 0* options. The *spliced_bam2gff* command was then run with identical options as before and resulting polished clusters that were then collapsed with the *collapse_partials* command with the *-M -U* options. Resulting GFF files were then manually parsed to retain those isoforms that contain at least one full CDS.

Host cell transcript isoform analysis

Analysis of host cell isoforms was performed using the FLAIR pipeline (29) v1.4. Reads were mapped to UCSC hg38 reference using the *flair align* module using option *-p*, followed by splice junction correction with the *flair correct* module. Isoforms are collapsed using the *flair collapse* module with *–stringent –trust_ends* options to ensure 80% coverage per isoform cluster. Transcript lengths can be calculated with flair collapse outputs, by indexing the transcripts.fa file for each sample with *samtools faidx* and extracting the second column containing length of each sequence. The isoforms were then quantified with the *flair quantify* module using *–tpm –trust_ends* options. Outputs of this module were used to compute gene expression TPM correlation between samples and replicates. The *flair diffexp* module was finally used to generate differential gene/isoform expression analysis with default settings. Finally, the *flair diffsplice* module was used to determinate high-confidence AS events from the isoforms processed with previous modules. Differential gene, isoform or splicing outputs were filtered for maximum *P*-value of 0.1; those hits that remain were subject to additional false discovery rate (FDR) analysis with those with $P_{adj} < 0.1$ being highly significant. Transcript discovery sensitivity and specificity were calculated using *gffcompare* v0.11.5 (30) using gtf file outputs from the *flair collapse* module and the UCSC hg38 genome annotation in gtf format with the following command options: *-T -M -r*.

Cross-validation with published Illumina dataset that used J-Lat 10.6 clone and TNF-α induction

Raw Illumina PE150 data from Ma *et al.*'s manuscript (31) were downloaded from the SRA/ENA repository, with two replicates obtained for each TNF-plus and TNF-minus treatment (accessions: SRR13649912, SRR13649913, SRR13649916 and SRR13649917). For analysis of host transcripts, reads were mapped to the hg38 reference using HISAT2 with the following settings: *–rna-strandness RF*, and converted to a sorted bam file with samtools. Human reference mapping counts per replicate were used to normalize raw transcript count matrix obtained for this dataset from the Gene Expression Omnibus (accession: GSE166337), yielding counts per million (CPM) metric. Genes shown to be significant in our nanopore differential gene expression (DGE) data ($P_{adj} < 0.1$) were queried against normalized counts from the PE150 dataset to determine concordance in fold-change direction before and after TNF-α induction. For splicing site usage analysis, sorted bam files were processed with Portcullis (32) with the following settings: *–exon-gff –orientation RF –strandedness firststrand –separate*, to obtain spliced/unspliced bam and gff files. Splice site regions of interest are extracted from the gff file, and normalized to obtain CPM values for a given splice junction. For analysis of HIV transcripts from these short-read data, reads were mapped to the R7 HIV reference sequence with HISAT2 with the following settings: *–rna-strandness RF –max-intronlen 5000*, followed by conversion to sorted bam file using samtools. Sorted bam files were processed with Portcullis with same settings as before to obtain spliced and unspliced bam and gff files. Coverage of spliced HIV reads was computed using *samtools depth* with the following settings: *-a -d 0*, with resulting depth per position normalized to average number of reads for a given dataset (to allow comparison with other datasets).

DGE subset validation via relative quantification qPCR

Total RNA was isolated from J-Lat 10.6 cells treated with TNF-α ($n = 3$) or vehicle control ($n = 4$), with RNA integrity verified with Agilent Bioanalyzer RNA Nano Chip. Total RNA was treated with CASPR and reverse transcription was carried out with MRT and oligo-d(T) priming, followed by second-strand synthesis with identical conditions to those used in sequencing samples. Double-stranded cDNA purity and concentration were assessed with NanoDrop 1000, followed by sample dilution to 0.1–0.5 ng/μl. PrimeTime qPCR primers were ordered from Integrated DNA Technologies for the following gene targets:

- ACTB (Hs.PT.39a.22214847), RefSeq NM_001101, exons 1-2: ACAGAGCCTCGCCTTTG, CCTTGCACATGCCGGAG

- LIMD2 (Hs.PT.58.25965347.g), RefSeq NM_030576, exons 4-5: CACTGTCACACCAAGCTCAG, GTCGTAGTTGCCTTTGCTCTT
- MYO7B (Hs.PT.58.1328863), RefSeq NM_001080527, exons 38-40: CCGAATCCAGAAGGTCCTGA, CAACCACCTCCAGCATCTC
- PSAT1 (Hs.PT.58.20540177), RefSeq NM_058179, exons 5-6: GTCCTCAAACTTCCTGTCCAA, TCATCACGGACAATCACCAC
- TNFAIP3 (Hs.PT.58.1824217, RefSeq NM_006290, exons 6-7: TGATAGAAATCCCCGTCCAAG, TCCTGCCATTTCTTGTACTCAT

qPCR was carried out using 5 μl of cDNA input, 500 nM primers and Luna Universal qPCR Master Mix at 1× concentration (NEB, M3003L) in total reaction volume of 20 μl. Plates were run in a Roche Lightcycler II instrument with the following cycling parameters with ramp rates of 2.2°C/s: 95°C for 1 min (initial denaturation), 45 cycles of 95°C for 15 s and 60°C for 30 s (amplification), 95°C for 5 s and 60°C ramp to 95°C (melting curve) and 40°C for 10 s (cooling). qPCR was validated by running a serial dilution of Jurkat total RNA across 6 logs to determine PCR efficiency, standard curve slope, *Y*-intercept and R^2 error for each primer set: ACTB (1.799, −3.921, 16.91, 0.0273), LIMD2 (1.927, −3.512, 21.00, 0.00429), MYO7B (1.916, −3.541. 18.02, 0.0133), PSAT1 (1.937, −3.482, 19.02, 0.0158) and TNFAIP3 (1.980, −3.371, 22.91, 0.0579). Nontemplate controls were used for all reactions, with Ct values >40 cycles for all samples. Gene expression was calculated via relative quantification of LIMD2, MYO7B, PSAT1 and TNFAIP3 targets over ACTB reference, with two technical replicates used per sample. Gene expression for +/− TNF-α samples was determined using the Advanced Relative Quantification module in Lightcycler 480 software (version 1.5.0 SP3) in high-confidence mode.

RESULTS

Improvement of the specificity and yield of high-performing RTs for producing full-length transcripts for direct cDNA sequencing

Obtaining a readout of AS of host and viral transcripts involves end-to-end sequencing of reads, which provides for full-exon connectivity. To achieve this, processive RTs are required, along with an enrichment scheme to select for protein-coding sequences from total RNA isolates. For direct cDNA sequencing, an additional requirement is to maximize the yield of cDNA so as to dispense with the need for PCR amplification of transcripts. Taking into account these requirements, we first evaluated the high-performing RTs MRT, a eubacterial group II intron that has been shown to efficiently copy structured long RNAs (33,34), and SSIV, which has been considered a 'commercial gold standard' (35,36), for their yield of protein-coding transcripts from Nalm6 total RNA, a human leukemic B cell line.

Gel electrophoresis of double-stranded cDNA obtained via SSIV and MRT showed prominent bands of similar size to rRNA when Nalm6 total RNA is directly reverse transcribed with oligo-d(T) priming without any CDS enrichment strategy (i.e. control) (Figure 1A). The presence of putative rRNA bands when using total RNA was unsurprising given these structural RNAs are a major RNA cellular component (enriched up to 90% in total RNA) and source of interference in RNA-seq workflows (37,38). However, rRNAs are not polyadenylated, which raises the question on the source of this spurious priming. We hypothesized that these primer-independent products were the result of the RNAs themselves priming the RT initiation complexes, and that blocking 3′-OH ends of RNA inputs prior to reverse transcription could be beneficial in increasing the specificity of RT priming. For this purpose, we developed an approach dubbed CASPR that oxidizes vicinal 2′ and 3′ diols to selectively ablate 3′-hydroxyl ends in RNAs only, preventing their activity as interfering primers during RT and favoring RT initiation from the intended exogenous oligo-d(T) DNA primer. Pretreatment of input RNA with CASPR visibly improved RT specificity in both SSIV and MRT, resulting in a smear reminiscent of PolyA+ selection (PolyA+) (Figure 1A), albeit with greater mass yield compared to this established methodology (Figure 1B). The increases in specificity of oligo-d(T) priming elicited by CASPR treatment were particularly evident in MRT samples, where CASPR-treated lanes do not show any discernable rRNA bands, compared to residual rRNA bands present with SSIV. CASPR treatment also resulted in 5- and 10-fold improvements in cDNA yield compared to PolyA+ for SSIV and MRT, respectively ($P < 0.01$ and $P < 0.001$), with the CASPR MRT combination resulting in 50% greater cDNA yield compared to CASPR SSIV ($P < 0.05$). This increase in RT specificity was consistent when using total RNA from other human cell lines, and when using gene-specific priming modalities with *in vitro* transcribed HIV RNA (Supplementary Figure S1), suggesting that spurious priming from RNA inputs is prevalent. Importantly, CASPR treatment did not compromise RNA quality as measured on total RNA via Agilent Bioanalyzer, with RIN scores of 10 obtained from both CASPR and vehicle controls (Supplementary Figure S2).

To validate that CASPR treatment was reducing rRNA, cDNA samples were sequenced with Oxford Nanopore Technologies (ONT) MinION to determine the effect of CASPR treatment at the read mapping level (Figure 1C). As expected, the most prominent effect of CASPR treatment is the reduction of reads mapping to rRNA reference from 84% to 24% in SSIV and from 75% to 12% in MRT, respectively ($P < 0.0001$ for both). This reduction in rRNA-mapped reads in CASPR-treated samples was associated with a proportional increase in percent of reads mapping to the human genome (hg38) reference, from 10% to 55% in SSIV and from 18% to 66% in MRT ($P < 0.0001$ for both), which compares favorably with hg38 enrichment levels in PolyA+ samples (75–80%). Compared to PolyA+, reads mapped to lncRNA mapping fractions were mostly nominal after CASPR treatment in both SSIV and MRT samples. Despite substantial CASPR-elicited increases in oligo-d(T) priming specificity, the reductions in rRNA were not fully penetrant compared to PolyA+, which routinely reduced rRNA reads to ∼1% irrespective of RT used. However, the read mapping fractions also show that each RT is not equally susceptible to rRNA interference, with MRT showing 2-fold lower rRNA fractions and 20%

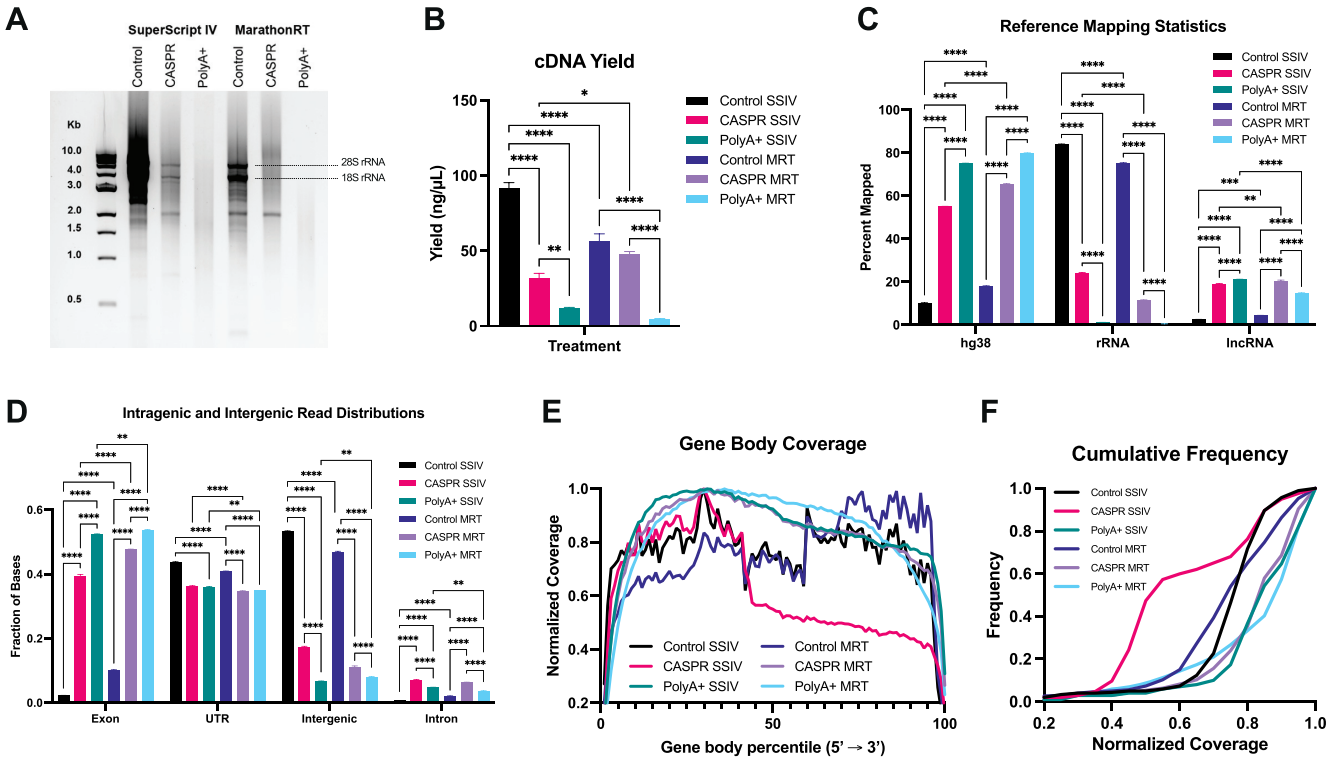

Figure 1. CASPR improves the specificity of oligo-d(T) primed RT when using total RNA inputs by reducing rRNA and increasing coverage evenness of protein-coding transcripts. (**A**) One percent agarose gel electrophoresis of double-stranded cDNA products that were reverse transcribed with oligo-d(T) priming with SSIV or MRT with no CDS enrichment (control), CASPR or PolyA+ selection. (**B**) cDNA yield of different RT and CDS enrichment combinations as measured spectrophotometrically. (**C**) Fraction of reads uniquely mapped to the listed references using nanopore sequencing. (**D**) Intragenic and intergenic read distributions. (**E**) Gene body coverage of protein-coding transcripts and (**F**) cumulative frequency distribution of gene body coverage. All values are means ± SEM. Statistical significance was calculated with two-way ANOVA with Tukey multiple comparison test: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$.

higher hg38 fractions after CASPR treatment compared to SSIV, suggesting MRT is more amenable to the priming specificity improvements elicited by ablation of 3′ RNA ends when using total RNA inputs. Given improvements observed in mapped read distributions elicited by CASPR, we next evaluated its effect on the distribution of intergenic and intragenic reads (Figure 1D). As expected, the most notable effect of CASPR and PolyA+ was a dramatic reduction in intergenic reads, with an associated increase in proportion of reads mapping to exonic and intronic regions ($P < 0.0001$ for all comparisons). Interestingly, both CASPR and PolyA+ slightly reduced read mappings to UTR regions in both RTs despite the associated increases in exonic reads that were observed for either treatment. All of this points to largely equivalent effects of CASPR and PolyA+ in increasing proportion of reads mapping to the intragenic features that delineate exon connectivity.

In addition to mapping statistics, the coverage along the length of protein-coding transcripts is critical to reveal full-exon connectivity. For this purpose, hg38 mapped reads were cross-referenced with the RefSeq genome annotation file to delineate the coverage along the 5′ to 3′ axis of each expressed transcript, an approach known as gene body coverage (28). The gene body coverage when using total RNA without CDS enrichment shows inconsistent coverage, with the control SSIV samples having clear 5′ and 3′ end biases (and associated low coverage in middle region of gene body) and control MRT showing consistent 3′ end bias (Figure 1E). Conversely, PolyA+ samples show even coverage with normalized coverage values consistently not dipping below 0.8 across 70% of gene body for both SSIV and MRT (Figure 1F) and no appreciable 5′ or 3′ end biases. These data suggest that the reduction in rRNA reads not only increases the number of reads mapping to protein-coding transcripts, but also increases their evenness of coverage across the length of the transcript. Interestingly, MRT samples that are CASPR treated show a gene body coverage distribution similar to that observed in PolyA+ samples, also consistently above 0.8 normalized coverage across majority of transcript body (Figure 1F). However, this same effect is not observed with CASPR-treated SSIV samples, with 0.5 median coverage compared to the >0.7 coverage values observed for all other treatment and RT combinations (Figure 1F). Overall, these data underscore the importance of CDS enrichment strategies and processive RTs in obtaining full-exon connectivity, while highlighting potential benefits of CASPR as an alternative to PolyA+ selection to substantially increase RT yield and priming specificity when using total RNA inputs.

Analytical performance validation of CDS enrichment strategies and RT conditions using synthetic RNA reference standards

Initial optimization of RT conditions using processive enzymes and a novel CDS enrichment strategy suggests that the combination of MRT with CASPR is well suited for direct cDNA sequencing using ONT. However, despite compelling data showing CASPR as a higher yield analogue of PolyA+ selection, and the coverage improvements elicited with MRT, neither of these interventions has been formally validated with reference standards. Synthetic RNA reference standards, which include ERCCs, SIRVs and sequins, have recently emerged for validating full RNA-seq workflows (39), and contain synthetic polyadenylated mono- and/or multiexonic transcripts of varied characteristics and in known concentration ranges. Given the synthetic nature of these transcripts, resulting reads obtained via sequencing can be cross-referenced with ground truth annotations to evaluate quantitative features of our workflow, the sensitivity and breadth of transcript capture, length biases due to RT processivity constraints and other performance variables. We used a SIRV (SIRV-Set 4) mix that was spiked into Nalm6 total RNA isolations prior to any enrichment interventions or RT with the goal of validating analytical performance of MRT and CASPR against established gold standards in the field. Three different CDS enrichment strategies (control, CASPR, PolyA+) and two different RTs (SSIV, MRT) were tested in triplicate using the same SIRV-spiked total RNA sample and run in parallel (Figure 2A). All the resulting CDS enrichment and RT combinations (six combinations, with three replicates each) are thus technical replicates of each other, and sequenced and analyzed (Figure 2B) in parallel to allow for robust comparisons between each condition.

Consistent with previous findings, direct cDNA sequencing of SIRV-spiked Nalm6 showed that CDS enrichment strategies are critical for enrichment of polyadenylated synthetic transcripts (Figure 3A). Specifically, CASPR treatment of total RNA prior to RT increased SIRV mapping by 5-fold in SSIV and 2.5-fold in MRT ($P < 0.001$ and $P < 0.01$, respectively). Moreover, the enrichment of SIRV reads with CASPR was comparable with that of PolyA+, with differences between enrichment strategies for each RT not statistically significant for SSIV and modest for MRT ($P < 0.05$). Given the lack of meaningful SIRV mapping fractions without CDS enrichment, CASPR and PolyA+ samples were sequenced deeper to allow for more sensitive analysis. One such analysis involves the quantification of ERCC controls within the SIRV mix, which are present in known concentrations spanning 6 logs. Cross-referencing of measured expression of ERCC transcripts with known input amounts showed that cDNA measurements are quantitative, with R^2 values averaging 0.9 for all CDS enrichment strategies and RT combinations (Figure 3B). This robustness in cDNA quantitation translates to actual measurements of human transcript abundance with all TPM correlations strongly trending in a linear manner irrespective of RT or CDS enrichment strategy tested (Figure 3C). ERCC data and hg38 gene expression correlations are strongly suggestive of CASPR treatment being functionally equivalent to PolyA+ selection with regard to ability to accurately quantify cDNA levels despite residual rRNA and marginally lower hg38 mapping fractions (Figure 1C). However, this does not provide clarity on the extent of coverage of these transcripts, a critical variable for full-length sequencing.

Isoform-level analysis can add an additional layer on the breadth of transcript coverage elicited by different RTs and CDS enrichment strategies. Isoform collapse and quantification of SIRV transcripts using FLAIR (29), followed by cross-referencing to known SIRVome reference annotation files, show that transcript capture sensitivities are largely equivalent between CASPR and PolyA+; however, CASPR provides distinct improvements in the transcript discovery sensitivity at the base level (Figure 3D). Base level transcript discovery corresponds to the number of exon bases within a query transcript that are reported at the same coordinate as the reference annotation. Therefore, a higher base level number would be reflective of a longer section of sequenced transcripts overlapping the reference annotation, and indicative of a greater breadth of coverage (30). In this regard, CASPR treatment shows 2-fold higher transcript discovery sensitivity at the base level compared to PolyA with both SSIV and MRT ($P < 0.001$ and $P < 0.0001$, respectively) and 40–60% higher at the locus level ($P < 0.05$ for SSIV, $P < 0.0001$ for MRT). This suggests that even though CASPR and PolyA result in equivalent number of read counts per transcript, CASPR provides significantly higher coverage per captured transcript, resulting in increased practical throughput and higher likelihood of capturing full-exon connectivity. Finally, long SIRVs ranging from 4 to 12 kb were quantified after sequencing for all RTs and CDS enrichment conditions to evaluate the propensity of each treatment combination to result in size biases related to the inherent processivity constraints of RTs for RNA inputs >5 kb in length, which was previously reported by us (33). Compared to PolyA+, CASPR trended toward increased sensitivity for capture of long synthetic transcripts >5 kb in size for all size classes (Figure 3E). Of particular note is the statistically significant increase in sensitivity of capture of 8 kb transcripts elicited by CASPR treatment, resulting in 6-fold increases in capture for both SSIV and MRT ($P < 0.01$ and $P < 0.0001$, respectively), and with MRT resulting in 2-fold higher sensitivity for this transcript size class as compared to SSIV ($P < 0.0001$). This increase in sensitivity of capture in CASPR-treated samples also translated to increased breadth of coverage for all transcript classes, with MRT in combination with CASPR showing more even coverage across all long SIRV transcript size classes, compared to limited coverage obtained with PolyA+ for both RTs (Figure 3F). Overall, these data validate that CASPR is functionally equivalent to PolyA+ selection while providing distinct advantages such as greater transcript coverage sensitivity and greater capacity to capture long transcripts. In addition, these data confirm that MRT, in combination with CASPR, has superior sensitivity and breadth of coverage than SSIV for capturing long polyadenylated transcripts from complex mixtures of host cell mRNAs.

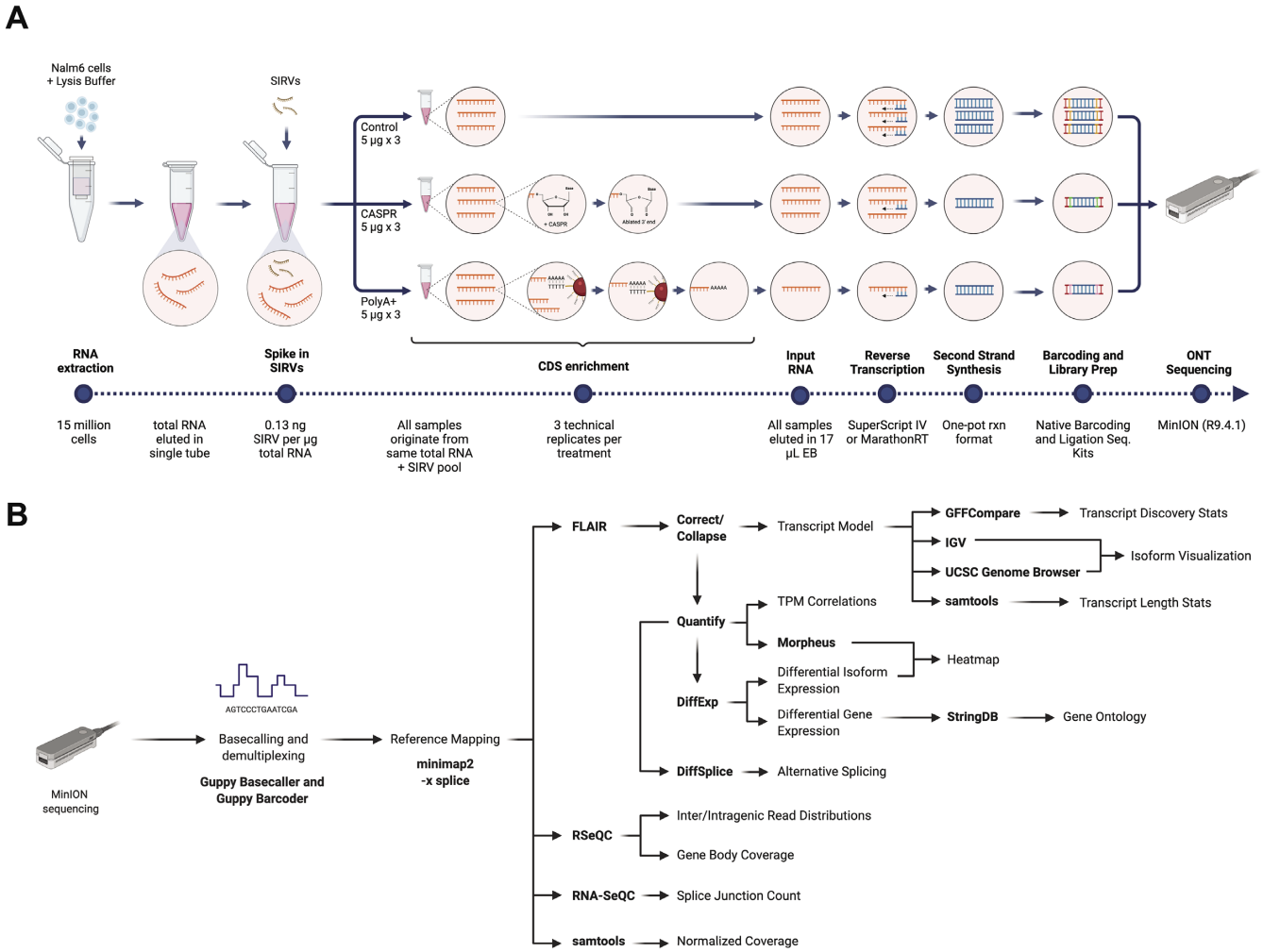

Figure 2. Assay and bioinformatic workflow for analytical performance validation. (**A**) Total RNA was isolated from Nalm6 cells and pooled into a single tube. Total RNA was spiked with SIRV-Set 4 at a concentration of 0.13 ng SIRVs per µg of total RNA. Three CDS enrichment conditions (control, CASPR and PolyA+ selection) were tested in parallel, all drawing identical RNA inputs from same total RNA sample in triplicate per condition. Following CDS enrichment, all samples are eluted in 17 µl of EB, followed by reverse transcription using 10 µl of input, and one-pot second-strand synthesis using a modified Gubler and Hoffman method. Following second-strand synthesis cleanup, identical volumes of double-stranded cDNA samples are then barcoded and prepared for sequencing using ONT Native Barcoding (EXP-NBD104, EXP-NBD-114) and Ligation Sequencing (SQK-LSK-109) Kits. Following library preparation, all samples are eluted in identical volumes of EB, and then equal volumes of each sample are pooled and sequenced via ONT MinION, using R9.4.1 chemistry. (**B**) Bioinformatic workflow used for data generation and analysis throughout manuscript. Typical outputs for each analysis are shown, with text in bold denoting specific tools used for a particular output.

Evaluation of RT and CDS enrichment strategies in the J-Lat 10.6 T cell line undergoing active HIV transcription

To determine whether our direct cDNA sequencing workflow can effectively capture HIV RNAs within a swarm of host cell transcripts, we evaluated both RTs and CDS enrichment conditions using the J-Lat 10.6 lymphocytic CD4 T cell line (23). This established and well-characterized Jurkat cell line has a single integrated provirus that contains all canonical splice sites and can be robustly induced to produce viral RNAs with TNF-α or other suitable HIV reactivation agents (24). Moreover, activation results in production of physiological levels of viral RNA, while also being representative of host transcriptional regulation dynamics of active infection (23). Thus, the J-Lat 10.6 cell line provides a stringent test case for evaluating efficiency of viral isoform capture within dynamically changing host cell transcripts without relying on PCR amplification to enrich for rare transcript variants, while allowing us to examine the effects of HIV reactivation on host cell transcript regulation.

J-Lat 10.6 cells were induced with 10 ng/ml TNF-α for 24 h, followed by assessment of p24 induction and EGFP expression, with all induction values normative to previous publications (Supplementary Figure S3). Both SSIV and MRT were tested for their performance with CASPR or PolyA selection, with all replicates and samples run in parallel. Consistent with previous data, host cell gene expression TPM values show concordance between CASPR treatment and PolyA+ selection when using either SSIV or MRT (Figure 4A) and it was reproducible across replicates (Supplementary Figure S4). Normalized gene body coverage values are consistent with those found in Nalm6 datasets, with CASPR MRT samples approaching the evenness observed

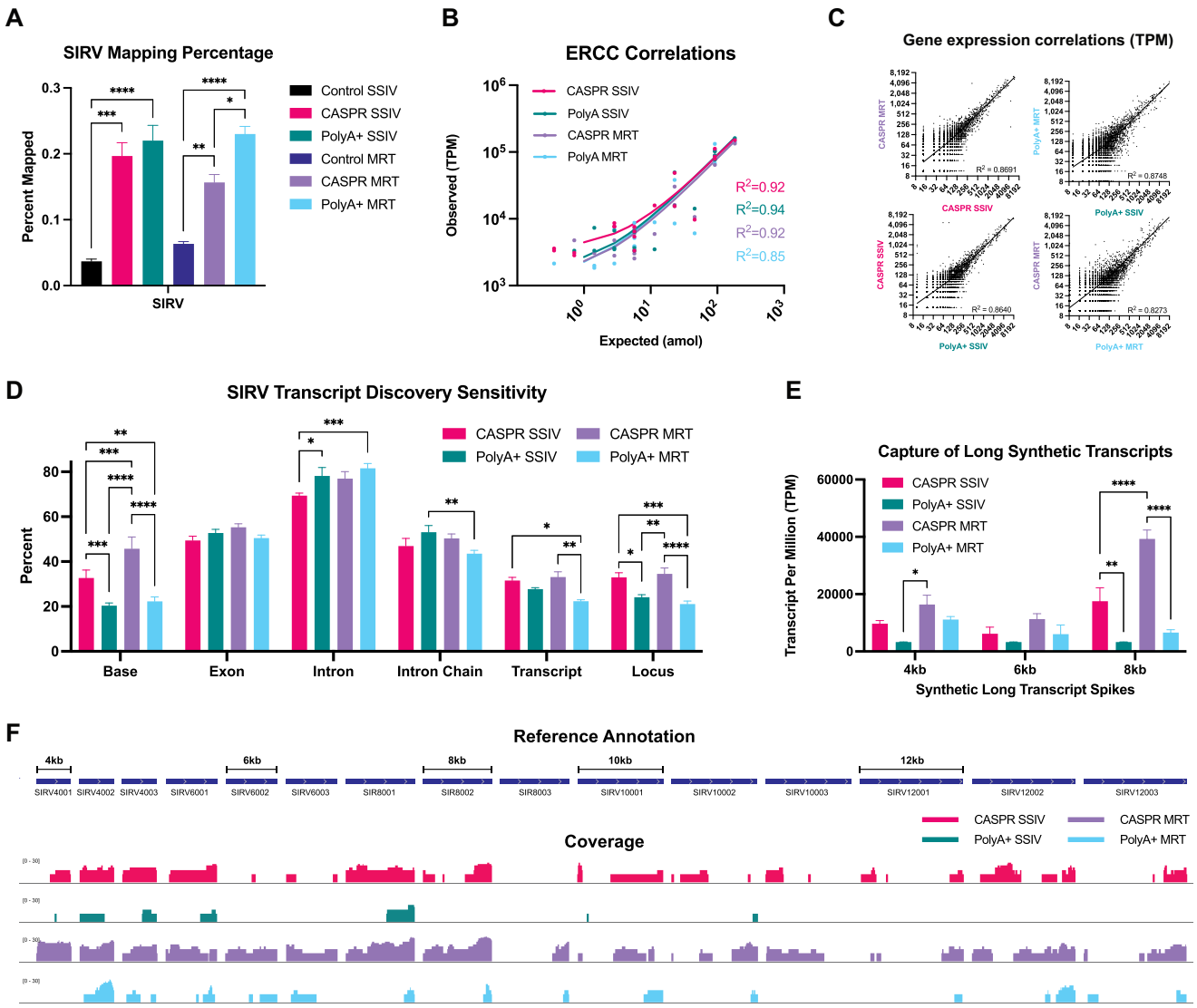

Figure 3. Validation with synthetic reference standards shows that CASPR is functionally equivalent to PolyA+ selection, but results in higher cDNA yield, coverage evenness and capture of long transcripts. (**A**) Percent of reads uniquely mapped to SIRV reference sequences. (**B**) Correlations of gene expression TPM values with absolute input amounts of each synthetic transcript (in attomoles) for ERCC subsets. (**C**) Hg38 gene expression correlations between different RTs and CDS enrichment strategies. (**D**) Transcript discovery sensitivity calculation using FLAIR-derived transcriptome and hg38 gtf annotation file. (**E**) Efficiency of capture of long SIRVs of 4, 6 and 8 kb size classes. (**F**) Raw coverage visualized via IGV of all long SIRVs for each RT and CDS enrichment strategy combination. All samples were run in triplicate ($n = 3$). All values are means ± SEM. Statistical significance was calculated with two-way ANOVA with Tukey multiple comparison test: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$.

in PolyA+ selected samples, and with SSIV showing measurable 5′ end bias consistent with previous data (Figure 4B). Compared to PolyA+, CASPR increases the fraction of long transcripts >4000 bp by 2.5- and 6-fold in SSIV and MRT, respectively ($P < 0.05$ for both), in a manner that is consistent with previously observed enrichments of long SIRVs (Figure 4C). The ability to capture longer transcripts positions CASPR well for the capture of HIV transcripts that are intrinsically difficult to reverse transcribe given high RNA structure (40) and their relatively long length (2–4 kb for spliced viral transcripts) compared to host cell coding transcripts (~1 kb average size).

With regard to the capture efficiency of HIV transcripts, our pipeline was able to capture thousands of HIV reads despite constituting <1% of total dataset (Supplementary Figure S5). To compare the performance of RTs and CDS enrichment strategies in coverage evenness, reads were mapped to the HIV reference and normalized across length of the genome, with a normalized coverage of 1 indicating even sampling (Figure 4D). CASPR and SSIV show more consistent coverage across length of genome, with relative coverage being close to 1 for most of the genome tract length relevant to multiexonic transcripts (5000–10 000 bp). SSIV PolyA trails closely behind, but shows reduced coverage in regions associated with Vif and Vpr transcripts (5000–6000 bp), and overall lower coverage for regions coding for Gag and Gag–Pol. Compared to SSIV, MRT shows 3′ end bias, with coverage dropping between 7500 and 8300 bp. Of

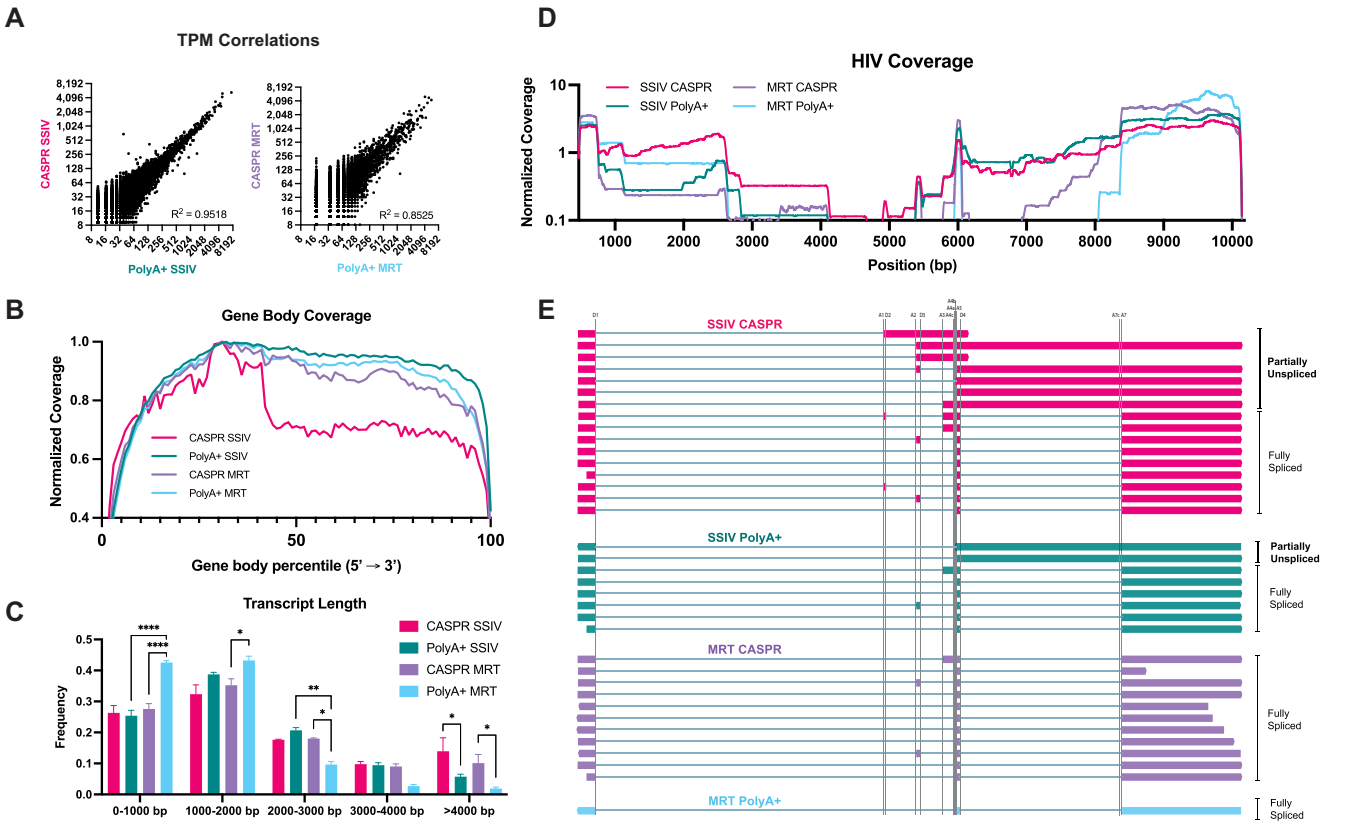

Figure 4. Evaluation of RT conditions and CDS enrichment strategies in capture of host and viral transcripts in cell line actively expressing HIV. (**A**) Host cell gene expression correlations for each CDS enrichment strategy when using SSIV and MRT. (**B**) Gene body coverage of protein-coding hg38 transcripts. (**C**) Frequency of host cell transcript lengths derived from the FLAIR isoform analysis pipeline, binned at 1000 bp intervals. (**D**) Coverage map of HIV reads. All samples were run in duplicate ($n = 2$). (**E**) Visualization of isoform structure of multiexonic HIV transcripts processed with the Pinfish pipeline. All values are means ± SEM. Statistical significance was calculated with two-way ANOVA with Tukey multiple comparison test: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$.

particular note, all samples show sharp increases in coverage at ∼2700 and ∼4200 bp, which are inconsistent with any splice junctions. However, the presence of long polyadenylated stretches in these two regions is suggestive of mispriming of oligo-d(T) being responsible for these artifactual increases in coverage.

To evaluate HIV isoform diversity in all treatments, HIV-mapped reads were grouped by exon boundaries into isoform clusters and collapsed into high-confidence multiexonic transcript models. This analysis pipeline worked robustly and identified splice sites that were consistent with those previously observed with long-read sequencing approaches (Supplementary Table S1). Multiexonic transcripts identified by Pinfish were then parsed to determine likely expressed gene based on which undisrupted ORF is closest to the 5′ end (Figure 4E). Consistent with normalized coverage data, MRT with PolyA+ selection did not capture overall HIV isoform diversity, with fully spliced species being favored. MRT with CASPR treatment performs nominally better than PolyA selection in increasing the isoform diversity of fully spliced transcripts; however, this treatment combination does not capture any partially unspliced transcripts coding for Env, Vpr and Vif. SSIV in combination with CASPR shows overall highest HIV isoform diversity, resulting in an assortment of fully spliced transcripts and 2–3-fold higher capture of partially spliced species compared to PolyA+. The detectable differences in viral isoform diversity captured with MRT and SSIV highlight the need to evaluate each RT enzyme independently of their performance in the capture of host cell transcripts and adopt strategies that take advantage of each RT's unique characteristics and strengths. For this purpose, an optimized approach to increase the likelihood of capturing both host and viral samples would rely on the interrogation of CASPR-treated total RNA using both SSIV and MRT, followed by the simultaneous sequencing of resulting cDNA.

Differential expression analysis using optimized RT and CDS enrichment conditions identifies alternatively spliced host factors of HIV assembly and defines their associated HIV splicing signature

Having evaluated the role of CASPR in increasing transcript capture efficiency and coverage metrics, and the identified strengths of SSIV and MRT for capture of respective viral and host transcripts, we set out to do a larger scale survey of viral reactivation dynamics within host cells in the J-Lat 10.6 cell line. The goal was the simultaneous identification of differentially regulated transcripts within host cells and their HIV isoform correlates. Taking into account

Table 1. Significant functional enrichments elicited by TNF-α

Term ID	Term description	Genes mapped	Enrichment score	FDR	Method	Sample group
GO:0033256	I-kappaB/NF-kappaB complex	4	5.53683	0.00055	afc	J-Lat (case)
GO:0005681	Spliceosomal complex	112	1.24205	0.00074	ks	J-Lat (case)
GO:0030667	Secretory granule membrane	59	0.566506	0.0032	ks	J-Lat (case)
GO:0022626	Cytosolic ribosome	68	0.22662	0.00054	ks	Jurkat (control)

Table 2. DGE in the J-Lat 10.6 case group elicited by TNF-α

Gene name	baseMean	$\log_2$FC	lfcSE	Stat	*P*-value	P_{adj}
TNFAIP3*	**28.3868218**	**2.02400656**	**0.29780197**	**6.79648477**	**1.07E−11**	**1.76E−08**
LIMD2*	**23.2264768**	**1.82090609**	**0.26816954**	**6.79013023**	**1.12E−11**	**1.76E−08**
NFKBIA*	**14.644418**	**1.98383297**	**0.33138767**	**5.98644168**	**2.14E−09**	**2.24E−06**
BIRC2*	**15.909459**	**1.83011777**	**0.31036165**	**5.89672651**	**3.71E−09**	**2.91E−06**
MYO7B*	**251.477162**	**−0.7096294**	**0.14296654**	**−4.9636048**	**6.92E−07**	**0.00043442**
FBLN2*	**37.7849837**	**0.91500472**	**0.21341341**	**4.28747525**	**1.81E−05**	**0.00945443**
NFKB2*	**14.0678839**	**1.456552**	**0.35152177**	**4.14356126**	**3.42E−05**	**0.01533418**
TRAF4*	**24.2017374**	**1.01974781**	**0.24850738**	**4.10349112**	**4.0696E−05**	**0.01596817**
RNASEK*	**64.4095627**	**0.68541206**	**0.18244994**	**3.75671294**	**1.72E−04**	**0.06004551**
SSBP2	45.5283287	0.67051415	0.18678805	3.58970585	3.31E−04	0.10391701
SNRPF	81.0650994	−0.4868181	0.1521773	−3.1990193	0.00137896	0.39350485
MKRN1	19.8006254	0.80917304	0.25865616	3.12837331	0.00175777	0.45980274
NFKB1	16.1716837	0.9648365	0.31784879	3.03552043	0.00240121	0.57979975
EFEMP1	12.5502742	0.92787868	0.31235625	2.97057829	0.0029724	0.61028546
ABCC1	29.8451097	0.65128555	0.21967884	2.96471677	0.00302962	0.61028546
PSAT1	47.7234616	−0.6819414	0.2320121	−2.9392494	3.29E−03	0.61028546
LSM7	33.4370785	−0.5771839	0.19914812	−2.8982645	0.00375234	0.61028546
EIF1	189.997178	−0.3583028	0.12546079	−2.8558944	0.00429158	0.61028546
TNIP1	20.3255463	0.88255174	0.30903108	2.85586724	0.00429195	0.61028546
HES4	40.1303992	−0.6558696	0.23071111	−2.8428174	0.00447167	0.61028546

The table lists 20 genes with lowest *P*-values in ascending order. Genes in bold and with asterisks have a P_{adj} value <0.1, and are thus highly significant. Genes provided in rows 1, 2, 6–9, 13 and 16 are also differentially expressed in the Jurkat control group with *P*-values <0.1. $\text{Log}_2\text{FC} = \log_2$ fold change (relative to TNF-α treatment).

our previous findings regarding the unique suitability for SSIV and MRT in the efficient capture of respective viral and host transcripts, total RNA was treated with CASPR and then split evenly to be reverse transcribed with SSIV and MRT, with resulting cDNA being used for sequencing. Since TNF-α induction is likely to cause global perturbations in host cell gene expression, the effect of TNF-α in the J-Lat 10.6 case group was compared with the differentially regulated transcripts elicited by TNF-α treatment in a control group of parental Jurkat cells lacking an integrated provirus. Those transcripts found to be differentially regulated by TNF-α in the Jurkat control group were 'subtracted' out from those differentially regulated in the J-Lat 10.6 case group, which will provide greater clarity on the host cell transcripts that will be uniquely up/downregulated by active HIV transcription, and not by the HIV reactivation agent itself.

An initial pilot run showed suitability of the approach in using both MRT and SSIV to maximize respective host cell and viral transcript capture efficiencies and coverage breadth during sequencing. Specifically, MRT showed 4-fold lower capture of artifactual rRNA-related hits in pilot differential isoform expression (DIE) analysis as compared with SSIV, with the latter showing that ∼40% of DIE hits can be traced to rRNA loci (Supplementary Figure S6). Given these initial results confirming suitability of our split MRT/SSIV approach, we proceeded to sequence additional biological replicates (up to a total of 5) in the presence or absence of TNF-α for both J-Lat (case) and Jurkat (control) groups (Supplementary Table S2). DGE analysis upon TNF-α induction in both case and control groups with *P*-values <0.1 revealed that 244 and 139 genes passed this filtering criterion in the J-Lat case and Jurkat control groups, respectively (Supplementary Figure S7). Of those genes passing *P*-value filtering criteria, 20 genes were found to be modulated by TNF-α induction in both J-Lat and Jurkat datasets, suggesting relatively low overlap between responses to TNF-α induction in case and control groups. To further determine the extent of TNF-α response overlap between case and control groups, DGE data were used to compute functional enrichment analysis with StringDB (41) version 11.0 with Gene Ontology (GO) framework at the Cellular Component and Biological Process levels. Highly significant (FDR < 0.01) GO Cellular Components enriched in the J-Lat case group include the 'NF-kappaB complex', the 'spliceosomal complex' and 'secretory granule membrane', which do not overlap with the single 'cytosolic ribosome' term found in the Jurkat control group (Table 1). Likewise, GO Biological Process terms do not overlap between case and control groups, except for NF-kappaB signaling, which is present in both case and controls groups, but 2-fold more enriched in the former group (Supplementary Figure S8). The activation of the NFKB complex observed in functional enrichment analysis is consistent with the highly significant (P_{adj} < 0.05) genes found to be differentially regulated in J-Lat cells treated with TNF-α, including TNFAIP3, NFKBIA, BIRC2 and NFKB2 (Table 2). Cross-comparison of our highly significant DGE hits

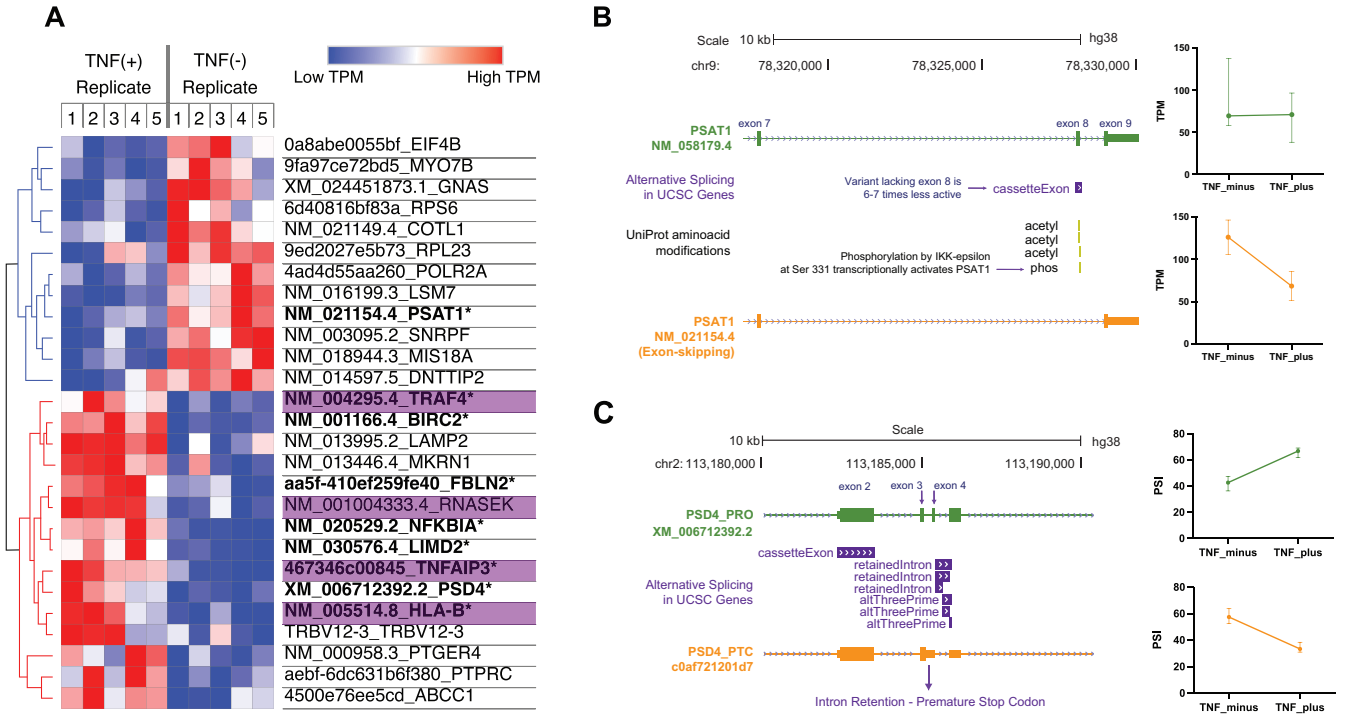

Figure 5. DIE analysis shows that putative HIV host factors PSAT1 and PSD4 are alternatively spliced in host cells upon HIV reactivation. (**A**) Heatmap showing hierarchically clustered TPM values for differentially expressed isoforms (*P*-value <0.1). Highly significant hits ($P_{adj} < 0.1$) are in bold, while isoforms shaded in purple are also present in the Jurkat control group. (**B**) PSAT1 isoform lacking functionally important exon 8 is differentially downregulated upon HIV reactivation with TNF-α. (**C**) Unproductive PSD4 isoform containing novel intron retention event is predominantly expressed in J-Lat cells prior to HIV induction. Upon HIV reactivation with TNF-α, intron retention event is downregulated and productive isoform is upregulated.

($P_{adj} < 0.1$) with a publicly available NGS dataset from a report using the J-Lat 10.6 clone and similar TNF-α induction conditions shows high concordance, with 78% of genes showing a consistent and highly significant fold-change direction in comparison with our data after TNF-α induction (Supplementary Figure S9) (31). Moreover, we performed qPCR validation of a subset of significant DGE hits and show high concordance with our sequencing findings in the direction of fold changes upon TNF-α induction and their statistical significance (Supplementary Figure S10). A significant fraction of the DGE hits in the J-Lat case group (including those related to the NFKB complex) were also found to be highly significant in the Jurkat group, underscoring the utility of our 'subtractive' approach to tease apart partially overlapping responses. The NFKB complex-related genes that were found to be differentially expressed exclusively in J-Lat cells include NFKBIA and BIRC2, which were previously found via RNA-seq to be upregulated upon latency reversal in SIV-infected ART-suppressed nonhuman primates (42). BIRC2 was also found to be a negative regulator of HIV transcription that could be antagonized with Smac mimetics for reversal of latency (43). The robust upregulation of BIRC2 we observed in our dataset despite active HIV transcription can be reconciled with the paradoxical role of this gene as both a positive modulator of the canonical NFKB pathway and a negative modulator of the noncanonical NFKB pathway (44), with our use of TNF-α engaging the canonical NFKB pathway.

To gain further insights into the specific transcript variants or isoforms eliciting gene expression changes, we plotted the TPM values of differentially expressed isoforms with *P*-value <0.01 in the J-Lat case group (Figure 5A). Hierarchical clustering shows two distinct populations, that are up- or downregulated upon TNF-α induction. Those isoforms also found to be differentially expressed in the Jurkat control group are highlighted in purple, and genes found to be highly significant ($P_{adj} < 0.1$) are in bold and denoted with an asterisk. Consistent with the DGE data, most of the highly significant DIE isoforms are upregulated upon TNF-α induction, with only a single isoform of PSAT1 being downregulated in this group. The DE isoform data confirm the involvement of the NFKB complex via significant 4-fold increases in relevant NFKBIA and BIRC2 isoform TPMs upon TNF-α treatment. Of particular note is the highly significant ($P_{adj} < 0.1$) paradoxical downregulation of a PSAT1 isoform given that previous studies have found this gene to be enriched during Tat-elicited cell proliferation in productive HIV infection (45) and during FOXO1 inhibition-elicited latency reversal in HIV-infected CD4 T cells (46). This paradoxical result can be reconciled by close inspection of its exon connectivity (Figure 5B), which reveals that the downregulated isoform (NM_021154.4) lacks exon 8 and results in a variant with 6–7-fold lower activity compared to the standard NM_058179.4 isoform that retains this exon (47). The downregulation of the PSAT1 isoform lacking exon 8 that we observed in our data is consistent with the statistically significant decrease in usage of

the exon 7–9 splice junction derived from our re-analysis of an independent and published Illumina dataset that used the J-Lat 10.6 clone and similar TNF-α induction conditions (Supplementary Figure S11) (31). Exon 8 contains a serine 331 residue that was shown to be phosphorylated by IKBKE, a known activator of the NFKB pathway, and this modification results in a downstream activation of the serine biosynthetic pathway (SBP) to support cell proliferation (48). Besides providing a putative link between the NFKB complex and the SBP in a latency reversal context, the coupling of exon connectivity along with DIE shows the utility of full-length approaches to clarify seemingly paradoxical mechanisms of transcriptional regulation.

To further investigate changes in splicing as a response to TNF-α-induced viral reactivation in host cells, we used the FLAIR DiffSplice module to call AS events from collapsed isoform clusters. An intron retention event between exon 3 and exon 4 in the PSD4 gene locus was found to be significantly ($P_{adj} < 0.05$) modulated upon TNF-α induction in J-Lat 10.6 cells (Figure 5C). This intron retention event, which is novel and not found in UCSC or SIB databases, was uniquely found in the J-Lat 10.6 case group and results in a premature termination codon that renders this transcript variant unproductive. DRIMSeq2 data were used to calculate the percent spliced in (PSI) of this intron retention event and showed that the nonproductive isoform was predominant in uninduced J-Lat (60% PSI value), but upon TNF induction, the intron retention was downregulated resulting in 65% PSI in the productive isoform. This dynamic is concomitant with the robust induction ($P_{adj} < 0.1$) of the productive XM006712392.2 PSD4 isoform upon TNF-α treatment as evidenced by its 2-fold increase in normalized isoform expression (Figure 5A). PSD4 belongs to a family of Pleckstrin and Sec7 domain containing proteins (PSD or EFA6), which are associated with the plasma membrane (PM) and interact with ARF6 proteins via their Sec7 guanine exchange factor domain to regulate PM and endosomal traffic (49). ARF6 has been previously found to be a molecular determinant of HIV-1 Gag association with the PM (50) via its activation of PIP5K lipid-modifying enzyme (51) that enhances PIP2 production, an acidic phospholipid that is specifically recognized by the highly basic region of HIV matrix for anchoring into PM (52). Despite the wealth of evidence of an ARF6 interaction with Sec7 domain containing proteins, PSD4 has not been directly associated with productive HIV infection or evaluated for its regulation via an intron retention mechanism.

In addition to host cell transcriptional correlates, our approach also captures the HIV transcriptional signature that is concomitant to TNF-α-induced viral reactivation in J-Lat 10.6 cells. Our nanopore approach to quantify HIV transcripts performs favorably with competing datasets with regard to coverage evenness, with cross-comparison of spliced reads with an Illumina-sequenced TNF-α-induced J-Lat 10.6 clone from an independent study showing uneven coverage across exons, and oversampling at splice junctions (Supplementary Figure S12). Isoform clustering and collapse analysis across four nanopore-sequenced replicates shows the capture of all canonical HIV splice sites and all multiexonic transcripts (Figure 6A). These transcripts are divided into 'completely spliced' (i.e. 2 kb) and 'incompletely spliced' (i.e. 4 kb) classes based on the presence or lack of a D4–A7 splice event. However, unlike previous approaches (18,21), direct comparison of enrichment between any transcript is possible in our approach irrespective of transcript class (Figure 6B). In addition to canonical HIV isoforms, our approach showed the presence of a novel Nef isoform that lacks the canonical A5–D4 exon while retaining a complete CDS. Additionally, a completely spliced variant of Vif was observed, which despite lacking the canonical intron retention between D4–A7 still contains a complete and undisrupted CDS upstream to this site. With regard to noncoding exons 2 and 3, these are present at much lower enrichment levels compared to previous studies (21), with noncoding exon 3 being more prevalent and associated with Rev/Nef/Tat/Env transcripts, and noncoding exon 2 being less prevalent and only associated with Tat and Nef transcripts. Gene assignment was based on two variables, with ORF proximity to the 5′ end of isoform being the initial variable, followed by the presence of a complete and undisrupted CDS. Using this system allows isoforms to be assigned to a gene unambiguously, particularly in cases of incompletely spliced transcripts containing A4 acceptors, where ORF proximity to 5′ cap alone would impute an unproductive Rev isoform, instead of the likely productive Env/Vpu transcript. By classifying isoforms into likely expressed genes (Figure 6C), relative gene expression can be determined, with highest enriched genes being Nef, Rev and Env accounting for 45%, 27% and 20% of transcripts, respectively. The high abundance of Nef and Rev compared to the relatively low level of Tat is consistent with previous studies (21,53) and concordant with splice acceptor usage in our data (Figure 6D). Moreover, the relatively high abundance of Rev is consistent with the requirement of this viral protein to oligomerize on RRE substrates to ensure the export of unspliced and partially unspliced transcripts out of the nucleus (54). As expected, the D1 splice donor shows highest usage followed closely by D4, the latter of which is consistent with the highest enrichment observed in transcripts containing the D4–A7 splice junction (i.e. fully spliced) (Figure 6E). HIV splicing dynamics can be further explored with a splice junction matrix (Figure 6F), showing all observed combinations of splice donor/acceptor junctions along with their enrichment, with D1–A5 and D4–A7 junctions being the most highly expressed junctions and correlating to Env and Rev/Nef transcripts, respectively.

DISCUSSION

In this study, we introduce and validate a full-length direct cDNA sequencing pipeline for the simultaneous profiling of polyadenylated viral and host cell transcripts from unamplified cDNA. This approach is supported by the use of two high-performing RTs and oligo-d(T) priming, coupled to a novel one-step chemical ablation of 3′ RNA ends (termed CASPR), which reduces rRNA reads and enriches polyadenylated transcripts. We use this approach to simultaneously interrogate host and viral transcriptional dynamics within a full-length sequencing context in a relevant cell line model of HIV reactivation. This has allowed us to identify putative host factors of HIV transcriptional activation that contain exon skipping events (PSAT1) or novel

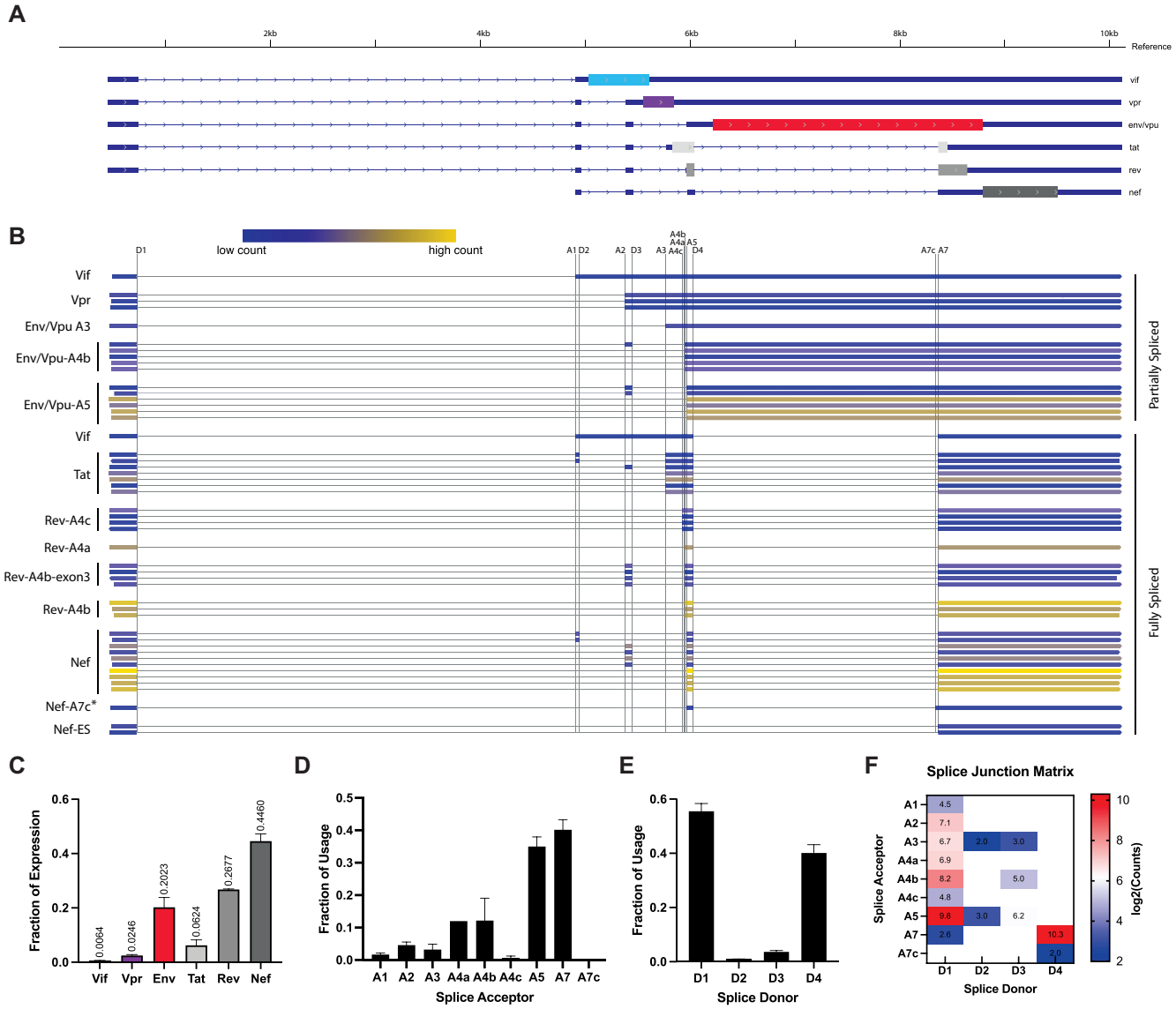

Figure 6. HIV transcriptional signature, gene expression and splice acceptor/donor usage for TNF-α-induced viral reactivation in J-Lat 10.6 cells. (**A**) Idealized splicing structures of HIV genes and their CDS regions. (**B**) HIV multiexonic isoform clusters observed across four replicates are color coded based on count numbers; isoform clusters are annotated with likely gene expressed and differentiating splice acceptor junction. Noncanonical/novel isoforms are labeled with an asterisk. (**C**) Gene expression fractions calculated based on counts obtained per isoform cluster, gene assignment based on proximity of ORF to 5′ end and presence of undisrupted CDS. Splice (**D**) acceptor and (**E**) donor usage. (**F**) Splice junction matrix with $\log_2$ normalized counts shows association and frequency of specific splice donor/acceptor junctions.

intron retentions (PSD4). In addition, our full-length RNA-seq pipeline is agnostic to sequencing methodology or library preparation approaches, and widely applicable for the study of viral transcription dynamics in host cells.

CASPR treatment in combination with MRT was a critical component in maximizing the quantitative capture of full-length host cell transcripts. The exact mechanism of CASPR-mediated improvements in obtaining full-length cDNA is beyond the scope of this manuscript; however, our data suggest that these improvements in priming specificity (via reduction of primer-independent products) are modulated by the 3′-OH ends of RNA inputs. The presence of nonspecific cDNAs generated in a primer-independent manner has been a largely overlooked artifact of reverse transcription. This has been cemented by the notion that exogenous DNA primers are an absolute requirement for reverse transcription, despite growing evidence of primer-independent cDNA generation in a variety of RTs, which has been variously reported in the field as ‘false priming’, ‘self-priming’ and ‘background priming’ (55–58). Moreover, the fact that the CASPR reagent resulted in improvements in the performance of both MRT and SSIV despite their different origins, and in a variety of RNA inputs and priming modalities, points to RT initiation in the absence of exogenous primer being a prevalent phenomenon. Primer-independent cDNA products are also a barrier in the study of replication dynamics of other RNA viruses where expression of negative-strand intermediate

transcripts is a hallmark of active viral replication, as is the case in dengue virus, West Nile virus, hepatitis C virus, SARS-CoV-2 and others (57,59–62). This suggests wide applicability of the CASPR reagent that, coupled with a suitable priming modality and a processive RT, could increase the breadth and sensitivity in the capture of full-length transcripts of interest in other relevant systems.

Given the polycistronic nature of HIV RNA, the full-exon connectivity provided by this pipeline is a critical component in the unambiguous assignment of detected isoforms to a likely expressed gene or in the identification of novel splice variants. This is not a minor problem for HIV, where a single intron retention event between two isoforms with seemingly identical splice junctions could result in expression of another viral gene. Full-length reads obtained in our pipeline allow straightforward isoform assignment and productivity analysis for the majority of HIV genes. However, the case of partially unspliced transcripts containing A4 or A5 splice sites constitutes an illustrative case where gene assignment can remain ambiguous even with full-length isoform information. Based on the premise that the closest ORF to the 5′ end of transcript constitutes the determinant factor in the gene that is expressed, partially unspliced isoforms containing A4 sites would translate to an unproductive Rev (since the CDS is disrupted by the D4/A7 intron retention), whereas those containing A5 would be translated as Env/Vpu. This ambiguity, however, is consistent with previous studies showing that HIV co-opts the host cell translation machinery in noncanonical ways to further regulate its gene expression via leaky ribosomal scanning or ribosome shunting (63). Thus, ORF proximity to 5′ end is a necessary but not sufficient factor in determining which gene is eventually expressed from a particular splice variant. In these cases, we used the presence of a complete and nondisrupted CDS as a second prioritization scheme for gene assignment, whereby a partially spliced variant containing an A4 junction is likely to code for productive Env/Vpu and not an unproductive Rev (i.e. prioritization of longest ORF). Given the dynamic nature of HIV RNA secondary structure proximal to splice junctions (64) and its inhibitory role in ribosome scanning, future studies coupling splice variant detection with DMS-MaP secondary structure probing (34) might provide additional clarity on Rev and Env/Vpu translational regulation, while allowing additional variables for consideration of gene assignment and productivity analyses.

Despite the moderate sequencing depth used in our studies, the yield and coverage increases elicited by CASPR allowed sufficient capture of host cell transcript variants for biologically meaningful DGE/DIE analyses while also detecting all canonical splice junctions in HIV isoforms. Sequencing throughput in our studies was a function of the MinION sequencer used, which allowed for rapid method development and validation studies at the expense of number of reads (compared to some large-scale transcriptomic studies of rare AS transcripts) (29). Any throughput constraints can be easily addressed in future studies by adopting higher throughput platforms available from ONT, including the GridION and PromethION each with 5- and 250-fold higher throughput. The higher sequencing depth provided by these platforms would enable detection of low-expression genes, resulting in higher sensitivity for rare genes, isoforms or splice variants. An additional consideration in our platform hinges on the number of cells required for dispensing with PCR amplification; currently, 50 000 cells are required to obtain sufficient total RNA. The required number of cells might not be unreasonable when using cultured cell lines, but when using primary cells or clinical samples, the requirement might be a limitation without further PCR amplification. For these types of samples, a cDNA amplification library preparation kit that attaches 5′ and 3′ adapters during RT can be used with CASPR-treated RNA inputs, followed by emulsion PCR with a single primer set and with a modest number of cycles to minimize PCR sampling bias (65), and allow for enrichment comparison between transcripts.

An interesting finding revealed by our study is the predominant intron retention event observed in the PSD4 locus of uninduced J-Lat cells, which results in expression of a truncated and inactive isoform due to a premature termination codon. The biological relevance of this AS event is not yet established; however, the role of other Sec7 domain containing proteins in targeting of viral components to the PM via its guanine exchange factor activity and interaction with ARF6 has been thoroughly documented (51). The reduction in expression of productive PSD4 could reduce the amount of active ARF6 and thus affect the balance of phosphatidylinositol that allows permissive assembly or entry of viral components proximal to the PM. However, intron retention events are widespread in cancer transcriptomes (66), and given the origin of J-Lat 10.6 cells from immortalized T cell leukemia peripheral blood mononuclear cells, the causal relationship between the modulation of PSD4 (and other AS isoforms) and HIV replicative capacity has to be thoroughly validated in primary cells.

In summary, we developed and systematically validated a full-length RNA-seq pipeline for assessing viral RNA transcript dynamics within a host cell transcriptome. This approach is supported by use of highly processive RTs, coupled with CASPR, a novel one-step CDS enrichment strategy that outperforms prevailing PolyA+ selection strategies in the breadth and sensitivity of capture of host cell and HIV transcripts. The simultaneous host and viral transcriptional signatures revealed in our approach can be used to interrogate transcriptional changes in response to a variety of viral induction methodologies, host gene manipulations (i.e. knockdown and knockouts) and viral sequence mutations, allowing greater granularity in the study of the interdependence of host and viral transcriptional regulation during infection with HIV and other RNA viruses.

DATA AVAILABILITY

Sequencing data have been submitted to the NCBI Sequence Read Archive under BioProject ID PRJNA801353.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Human Genome Research Institute [R01HG009622 to B.E.T.]; National Institute of Allergy and Infectious Diseases [U54AI150472 to B.E.T].

Funding for open access charge: National Institute of Allergy and Infectious Diseases [U54AI150472].
Conflict of interest statement. C.M.G. and B.E.T. are listed as inventors on a provisional patent filed by Seattle Children's Research Institute related to the CASPR methodology.

REFERENCES

1. Pan,Q., Shai,O., Lee,L.J., Frey,B.J. and Blencowe,B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
2. Fu,X.D. and Ares,M. Jr (2014) Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 689–701.
3. Nilsen,T.W. and Graveley,B.R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**, 457–463.
4. Liu,Y., Gonzàlez-Porta,M., Santos,S., Brazma,A., Marioni,J.C., Aebersold,R., Venkitaraman,A.R. and Wickramasinghe,V.O. (2017) Impact of alternative splicing on the human proteome. *Cell Rep.*, **20**, 1229–1241.
5. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
6. Park,E., Pan,Z., Zhang,Z., Lin,L. and Xing,Y. (2018) The expanding landscape of alternative splicing variation in human populations. *Am. J. Hum. Genet.*, **102**, 11–26.
7. Ashraf,U., Benoit-Pilven,C., Lacroix,V., Navratil,V. and Naffakh,N. (2019) Advances in analyzing virus-induced alterations of host cell splicing. *Trends Microbiol.*, **27**, 268–281.
8. Imbeault,M., Giguère,K., Ouellet,M. and Tremblay,M.J. (2012) Exon level transcriptomic profiling of HIV-1-infected $CD4^+$ T cells reveals virus-induced genes and host environment favorable for viral replication. *PLoS Pathog.*, **8**, e1002861.
9. Byun,S., Han,S., Zheng,Y., Planelles,V. and Lee,Y. (2020) The landscape of alternative splicing in HIV-1 infected CD4 T-cells. *BMC Med. Genomics*, **13**, 38.
10. Wojcechowskyj,J.A., Didigu,C.A., Lee,J.Y., Parrish,N.F., Sinha,R., Hahn,B.H., Bushman,F.D., Jensen,S.T., Seeholzer,S.H. and Doms,R.W. (2013) Quantitative phosphoproteomics reveals extensive cellular reprogramming during HIV-1 entry. *Cell Host Microbe*, **13**, 613–623.
11. Pabis,M., Corsini,L., Vincendeau,M., Tripsianes,K., Gibson,T.J., Brack-Werner,R. and Sattler,M. (2019) Modulation of HIV-1 gene expression by binding of a ULM motif in the Rev protein to UHM-containing splicing factors. *Nucleic Acids Res.*, **47**, 4859–4871.
12. Lapek,J.D., Lewinski,M.K., Wozniak,J.M., Guatelli,J. and Gonzalez,D.J. (2017) Quantitative temporal viromics of an inducible HIV-1 model yields insight to global host targets and phospho-dynamics associated with protein Vpr. *Mol. Cell. Proteomics*, **16**, 1447–1461.
13. Mueller,N., Pasternak,A.O., Klaver,B., Cornelissen,M., Berkhout,B. and Das,A.T. (2018) The HIV-1 Tat protein enhances splicing at the major splice donor site. *J. Virol.*, **92**, e01855-17.
14. Zhou,C., Liu,S., Song,W., Luo,S., Meng,G., Yang,C., Yang,H., Ma,J., Wang,L., Gao,S. *et al.* (2018) Characterization of viral RNA splicing using whole-transcriptome datasets from host species. *Sci. Rep.*, **8**, 3273.
15. Karn,J. and Stoltzfus,C.M. (2012) Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harb. Perspect. Med.*, **2**, a006916.
16. Kutluay,S.B., Emery,A., Penumutchu,S.R., Townsend,D., Tenneti,K., Madison,M.K., Stukenbroeker,A.M., Powell,C., Jannain,D., Tolbert,B.S. *et al.* (2019) Genome-wide analysis of heterogeneous nuclear ribonucleoprotein (hnRNP) binding to HIV-1 RNA reveals a key role for hnRNP H1 in alternative viral mRNA splicing. *J. Virol.*, **93**, e01048-19.
17. Esquiaqui,J.M., Kharytonchyk,S., Drucker,D. and Telesnitsky,A. (2020) HIV-1 spliced RNAs display transcription start site bias. *RNA*, **26**, 708–714.
18. Emery,A., Zhou,S., Pollom,E. and Swanstrom,R. (2017) Characterizing HIV-1 splicing by using next-generation sequencing. *J. Virol.*, **91**, e02515-16.
19. Shi,H., Zhou,Y., Jia,E., Pan,M., Bai,Y. and Ge,Q. (2021) Bias in RNA-seq library preparation: current challenges and solutions. *Biomed. Res. Int.*, **2021**, 6647597.
20. Byrne,A., Beaudin,A.E., Olsen,H.E., Jain,M., Cole,C., Palmer,T., DuBois,R.M., Forsberg,E.C., Akeson,M. and Vollmers,C. (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.*, **8**, 16027–16027.
21. Ocwieja,K.E., Sherrill-Mix,S., Mukherjee,R., Custers-Allen,R., David,P., Brown,M., Wang,S., Link,D.R., Olson,J., Travers,K. *et al.* (2012) Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic Acids Res.*, **40**, 10345–10355.
22. Nguyen Quang,N., Goudey,S., Ségéral,E., Mohammad,A., Lemoine,S., Blugeon,C., Versapuech,M., Paillart,J.-C., Berlioz-Torrent,C., Emiliani,S. *et al.* (2020) Dynamic nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection. *Retrovirology*, **17**, 25.
23. Jordan,A., Bisgrove,D. and Verdin,E. (2003) HIV reproducibly establishes a latent infection after acute infection of T cells *in vitro*. *EMBO J.*, **22**, 1868–1877.
24. Spina,C.A., Anderson,J., Archin,N.M., Bosque,A., Chan,J., Famiglietti,M., Greene,W.C., Kashuba,A., Lewin,S.R., Margolis,D.M. *et al.* (2013) An in-depth comparison of latent HIV-1 reactivation in multiple cell model systems and resting $CD4^+$ T cells from aviremic patients. *PLoS Pathog.*, **9**, e1003834.
25. Weiss,A., Wiskocil,R. and Stobo,J. (1984) The role of T3 surface molecules in the activation of human T cells: a two-stimulus requirement for IL 2 production reflects events occurring at a pre-translational level. *J. Immunol.*, **133**, 123–128.
26. Gubler,U. and Hoffman,B.J. (1983) A simple and very efficient method for generating cDNA libraries. *Gene*, **25**, 263–269.
27. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Subgroup,G.P.D.P. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
28. Wang,L., Nie,J., Sicotte,H., Li,Y., Eckel-Passow,J.E., Dasari,S., Vedell,P.T., Barman,P., Wang,L., Weinshiboum,R. *et al.* (2016) Measure transcript integrity using RNA-seq data. *BMC Bioinformatics*, **17**, 58.
29. Tang,A.D., Soulette,C.M., van Baren,M.J., Hart,K., Hrabeta-Robinson,E., Wu,C.J. and Brooks,A.N. (2020) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.*, **11**, 1438.
30. Pertea,G. and Pertea,M. (2020) GFF utilities: GffRead and GffCompare. *F1000Research*, **9**, ISCB Comm J-304.
31. Ma,X., Chen,T., Peng,Z., Wang,Z., Liu,J., Yang,T., Wu,L., Liu,G., Zhou,M., Tong,M. *et al.* (2021) Histone chaperone CAF-1 promotes HIV-1 latency by leading the formation of phase-separated suppressive nuclear bodies. *EMBO J.*, **40**, e106632.
32. Mapleson,D., Venturini,L., Kaithakottil,G. and Swarbreck,D. (2018) Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *GigaScience*, **7**, giy131.
33. Zhao,C., Liu,F. and Pyle,A.M. (2018) An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA*, **24**, 183–195.
34. Guo,L.T., Adams,R.L., Wan,H., Huston,N.C., Potapova,O., Olson,S., Gallardo,C.M., Graveley,B.R., Torbett,B.E. and Pyle,A.M. (2020) Sequencing and structure probing of long RNAs using MarathonRT: a next-generation reverse transcriptase. *J. Mol. Biol.*, **432**, 3338–3352.
35. Stahlberg,A., Kubista,M. and Pfaffl,M. (2004) Comparison of reverse transcriptases in gene expression analysis. *Clin. Chem.*, **50**, 1678–1680.
36. Zucha,D., Androvic,P., Kubista,M. and Valihrach,L. (2019) Performance comparison of reverse transcriptases for single-cell studies. *Clin. Chem.*, **66**, 217–228.
37. O'Neil,D., Glowatz,H. and Schlumpberger,M. (2013) Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Curr. Protoc. Mol. Biol.*, **Chapter 4**, Unit 4.19.

38. Zhao,S., Zhang,Y., Gamini,R., Zhang,B. and Schack,D. (2018) Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci. Rep.*, **8**, 4781.
39. Hardwick,S.A., Deveson,I.W. and Mercer,T.R. (2017) Reference standards for next-generation sequencing. *Nat. Rev. Genet.*, **18**, 473–484.
40. Watts,J.M., Dang,K.K., Gorelick,R.J., Leonard,C.W., Bess,J.W., Swanstrom,R., Burch,C.L. and Weeks,K.M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711–716.
41. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H. and Bork,P. (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
42. Nixon,C.C., Mavigner,M., Sampey,G.C., Brooks,A.D., Spagnuolo,R.A., Irlbeck,D.M., Mattingly,C., Ho,P.T., Schoof,N., Cammon,C.G. *et al.* (2020) Systemic HIV and SIV latency reversal via non-canonical NF-κB signalling *in vivo*. *Nature*, **578**, 160–165.
43. Pache,L., Dutra,M.S., Spivak,A.M., Marlett,J.M., Murry,J.P., Hwang,Y., Maestre,A.M., Manganaro,L., Vamos,M., Teriete,P. *et al.* (2015) BIRC2/cIAP1 is a negative regulator of HIV-1 transcription and can be targeted by smac mimetics to promote reversal of viral latency. *Cell Host Microbe*, **18**, 345–353.
44. Hrdinka,M. and Yabal,M. (2019) Inhibitor of apoptosis proteins in human health and disease. *Genes Immun.*, **20**, 641–650.
45. Jarboui,M.A., Bidoia,C., Woods,E., Roe,B., Wynne,K., Elia,G., Hall,W.W. and Gautier,V.W. (2012) Nucleolar protein trafficking in response to HIV-1 Tat: rewiring the nucleolus. *PLoS One*, **7**, e48702.
46. Vallejo-Gracia,A., Chen,I.P., Perrone,R., Besnard,E., Boehm,D., Battivelli,E., Tezil,T., Krey,K., Raymond,K.A., Hull,P.A. *et al.* (2020) FOXO1 promotes HIV latency by suppressing ER stress in T cells. *Nat. Microbiol.*, **5**, 1144–1157.
47. Baek,J.Y., Jun,D.Y., Taub,D. and Kim,Y.H. (2003) Characterization of human phosphoserine aminotransferase involved in the phosphorylated pathway of L-serine biosynthesis. *Biochem. J.*, **373**, 191–200.
48. Xu,R., Jones,W., Wilcz-Villega,E., Costa,A.S., Rajeeve,V., Bentham,R.B., Bryson,K., Nagano,A., Yaman,B., Olendo Barasa,S. *et al.* (2020) The breast cancer oncogene IKKε coordinates mitochondrial function and serine metabolism. *EMBO Rep.*, **21**, e48260.
49. Sztul,E., Chen,P.W., Casanova,J.E., Cherfils,J., Dacks,J.B., Lambright,D.G., Lee,F.S., Randazzo,P.A., Santy,L.C., Schürmann,A. *et al.* (2019) ARF GTPases and their GEFs and GAPs: concepts and challenges. *Mol. Biol. Cell*, **30**, 1249–1271.
50. Chukkapalli,V. and Ono,A. (2011) Molecular determinants that regulate plasma membrane association of HIV-1 Gag. *J. Mol. Biol.*, **410**, 512–524.
51. Van Acker,T., Tavernier,J. and Peelman,F. (2019) The small GTPase Arf6: an overview of its mechanisms of action and of its role in host–pathogen interactions and innate immunity. *Int. J. Mol. Sci.*, **20**, 2209.
52. Freed,E.O. (2006) HIV-1 Gag: flipped out for PI(4,5)P_2. *Proc. Natl Acad. Sci. U.S.A.*, **103**, 11101–11102.
53. Erkelenz,S., Hillebrand,F., Widera,M., Theiss,S., Fayyaz,A., Degrandi,D., Pfeffer,K. and Schaal,H. (2015) Balanced splicing at the Tat-specific HIV-1 3′ss A3 is critical for HIV-1 replication. *Retrovirology*, **12**, 29.
54. Fernandes,J., Jayaraman,B. and Frankel,A. (2012) The HIV-1 Rev response element. *RNA Biol.*, **9**, 6–11.
55. Lanford,R.E., Chavez,D., Chisari,F.V. and Sureau,C. (1995) Lack of detection of negative-strand hepatitis C virus RNA in peripheral blood mononuclear cells and other extrahepatic tissues by the highly strand-specific rTth reverse transcriptase PCR. *J. Virol.*, **69**, 8079–8083.
56. Haddad,F., Qin,A.X., Giger,J.M., Guo,H. and Baldwin,K.M. (2007) Potential pitfalls in the accuracy of analysis of natural sense–antisense RNA pairs by reverse transcription-PCR. *BMC Biotechnol.*, **7**, 21.
57. Tuiskunen,A., Leparc-Goffart,I., Boubis,L., Monteil,V., Klingström,J., Tolou,H.J., Lundkvist,A. and Plumet,S. (2010) Self-priming of reverse transcriptase impairs strand-specific detection of dengue virus RNA. *J. Gen. Virol.*, **91**, 1019–1027.
58. Frech,B. and Peterhans,E. (1994) RT-PCR: 'background priming' during reverse transcription. *Nucleic Acids Res.*, **22**, 4342–4343.
59. Lim,S.M., Koraka,P., Osterhaus,A.D. and Martina,B.E. (2013) Development of a strand-specific real-time qRT-PCR for the accurate detection and quantitation of West Nile virus RNA. *J. Virol. Methods*, **194**, 146–153.
60. Lerat,H., Berby,F., Trabaud,M.A., Vidalin,O., Major,M., Trépo,C. and Inchauspé,G. (1996) Specific detection of hepatitis C virus minus strand RNA in hematopoietic cells. *J. Clin. Invest.*, **97**, 845–851.
61. Fehr,A.R. and Perlman,S. (2015) Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol. Biol.*, **1282**, 1–23.
62. Sawicki,S.G. (2008) Coronavirus genome replication. In: Raney,K., Gotte,M. and Cameron,C. (eds). *Viral Genome Replication*. Springer, Boston, MA, pp. 25–39.
63. Guerrero,S., Batisse,J., Libre,C., Bernacchi,S., Marquet,R. and Paillart,J.C. (2015) HIV-1 replication and the cellular eukaryotic translation apparatus. *Viruses*, **7**, 199–218.
64. Tomezsko,P.J., Corbin,V.D.A., Gupta,P., Swaminathan,H., Glasgow,M., Persad,S., Edwards,M.D., McIntosh,L., Papenfuss,A.T., Emery,A. *et al.* (2020) Determination of RNA structural diversity and its role in HIV-1 RNA splicing. *Nature*, **582**, 438–442.
65. Gallardo,C.M., Wang,S., Montiel-Garcia,D.J., Little,S.J., Smith,D.M., Routh,A.L. and Torbett,B.E. (2021) MrHAMER yields highly accurate single molecule viral sequences enabling analysis of intra-host evolution. *Nucleic Acids Res.*, **49**, e70.
66. Dvinge,H. and Bradley,R.K. (2015) Widespread intron retention diversifies most cancer transcriptomes. *Genome Med.*, **7**, 45.