Science Translational Medicine
AAAS

stm.sciencemag.org/cgi/content/full/12/543/eaaz0802/DC1

Supplementary Materials for

Single-cell transcriptional landscapes reveal HIV-1–driven aberrant host gene transcription as a potential therapeutic target

Runxia Liu, Yang-Hui Jimmy Yeh, Ales Varabyou, Jack A. Collora, Scott Sherrill-Mix, C. Conover Talbot Jr., Sameet Mehta, Kristen Albrecht, Haiping Hao, Hao Zhang, Ross A. Pollack, Subul A. Beg, Rachela M. Calvi, Jianfei Hu, Christine M. Durand, Richard F. Ambinder, Rebecca Hoh, Steven G. Deeks, Jennifer Chiarella, Serena Spudich, Daniel C. Douek, Frederic D. Bushman, Mihaela Pertea, Ya-Chi Ho*

*Corresponding author. Email: ya-chi.ho@yale.edu

Published 13 May 2020, *Sci. Transl. Med.* **12**, eaaz0802 (2020)
DOI: 10.1126/scitranslmed.aaz0802

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at stm.sciencemag.org/cgi/content/full/12/543/eaaz0802/DC1)

Data file S1 (Microsoft Excel format). Primary data.
Data file S2 (Microsoft Excel format). HIV-1 SortSeq probe sequences.
Data file S3 (.docx format). Location of HIV-1 SortSeq probes.

Data file S4 (Microsoft Excel format). List of HIV-1 SortSeq samples.
Data file S5 (Microsoft Excel format). Differentially expressed genes between HIV-1 Sortseq^{+} and Sortseq^{-} cells.

Materials and methods

Induction of HIV-1 RNA expression from resting CD4$^+$ T cells

Resting CD4$^+$ T cells from ART-treated, virally suppressed, HIV-1-infected individuals were isolated using magnetic negative depletion with EasySep human CD4$^+$ T cell enrichment kit (STEMCELL), CD69 MicroBead Kit, CD25 MicroBead Kit, and HLA-DR Microbead Kit (Miltenyi Biotec). Cells were cultured in the presence of enfuvirtide (10 µM) to block new rounds of infection. Enfurvirtide was chosen to block viral entry and to avoid false-positive measurement of the incoming virion as HIV-1 RNA expression following latency reversal (*38*). Cells were cultured at 1 million/ml culture media, 1 µg/ml PMA, 1 µM ionomycin, and 30 units/ml interleukin-2 for 16–18 hours to induce HIV-1 RNA expression without inducing cellular proliferation and in vitro infection. PMA and ionomycin were chosen because they have been validated as one of the most potent latency reversing agents in clinical samples (*76*). Cells were examined within 24 hours to characterize the early cellular responses upon latency reversal. Aliquots of 5–25 million CD4$^+$ T cells from HIV-1-infected individuals were aliquoted per tube. Twenty five million primary CD4$^+$ T cells from HIV-1-uninfected individuals were used as negative controls. Five million activated CD4$^+$ T cells from HIV-1-uninfected individuals were infected with NL4-3 virus in vitro as positive controls.

HIV-1 SortSeq

Using HIV-1 RNA expression as a surrogate, HIV-1 SortSeq captures HIV-1-infected cells upon early latency reversal and overcomes the rarity of CD4$^+$ T cells in HIV-1-infected, ART-treated, virally suppressed individuals. Readily detectable levels of both 5' and 3' HIV-1 RNA indicate the presence of inducible and putatively intact HIV-1 (*38*). The fluorescent in situ hybridization process (FISH) is modified from a previously reported method (*37*) with modifications described below. All reagents are certified RNAase free from the vendor. Centrifugation was performed at 400 *g* for 10 min before permeabilization and 800 *g* for 10 min after permeabilization (to reduce cell loss) in a swing-bucket centrifuge.

Probes

We designed 96 probes targeting 5' HIV-1 RNA and another 96 probes targeting 3' HIV-1 RNA over conserved regions (data file S2 and S3). The use of one set of 96 fluorescently labeled probes targeting 5' HIV-1 RNA and another set of 96 fluorescently labeled probes targeting 3' HIV-1 RNA overcomes the HIV-1 sequence diversity in clinical samples and maximizes HIV-1 RNA signal without RNA degradation caused by branch-DNA amplification (*22*). HIV-1 probes were obtained from Stellaris LGC Biosearch Technologies. Each pool of 48 probes were diluted with 200 µl TE into 25 µM, immediately aliquoted into 5 µl single-use aliquots and stored at -20°C.

Viability staining

Aliquots of 5–25 million cells were pelleted, stained with LIVE/DEAD Fixable Near-IR Dead Cell Stain Kit (Thermo Fisher L10119) in RNase-free 1X phosphate buffered saline (PBS) (Thermo Fisher 10010023) at 4°C for 15 min, and washed twice with 1X PBS supplemented with 2% heat-inactivated fetal bovine serum. Cells were pelleted into a protein-low bind tube (Eppendorf, 022431064) during the final wash.

Fixation

Cells were fixed with freshly made, methanol free 1% formaldehyde (made fresh from methanol-free 16% formaldehyde (VWR 100503-916), 10X PBS (Thermo Scientific AM9624) and RNase free H2O same day immediately before use). The use of methanol free formaldehyde, instead of the methanol containing paraformaldehyde, preserves RNA integrity. Cells were fixed for exactly 5 minutes at room temperature as longer incubation may cause over-crosslinking. Cells were pelleted at 400 *g* for 10 min at room temperature.

Permeabilization

Cells were permeabilized by 1 ml freshly made 70% ethanol (made fresh immediately before use) at 4°C for 30 min to overnight. Instead of commercially available permeabilization reagents, this ensures that RNA is incubated in an RNA-free solution. An overnight incubation is typically applied, although 30 min incubation still allows sufficient permeabilization. Cells can be transferred to biosafety level-1 containment after this step.

Hybridization

Cells in each microfuge tube were pelleted and washed with 1 ml freshly made FISH Wash buffer (2X SSC, 10 µg/ml bovine serum albumin (BSA) from UltraPure BSA (Thermo Scientific AM2616) and 40% formamide made from 20X SSC (Thermo Scientific AM9770), formamide (Thermo Scientific AM9342)) at room temperature. Cells were then hybridized in 40 nM of each probe set in the RNA preserving hybridization buffer (RPHB) made from 20X SSC (Thermo Scientific AM9770), ammonium sulfate (Sigma Alderich A4418-100G), EDTA (Thermo Scientific AM9260G), formamide (Themo Scientific AM9342), tRNA from *Escherichia coli* (Roche 10109541001) and bovine serum albumin (Thermo Scientific AM2616) pH 5.2 in a final volume of 200 µl per tube. Hybridization was performed on a metallic block in a 30°C incubator. Cells were incubated from 2 hours to 18 hours. Longer incubation may lead to brighter signals but lower RNA integrity. Cells were then washed once with 1 ml freshly made FISH Wash buffer, pelleted, follow by two times of resuspension in 1 ml FISH Wash buffer and incubation at 30°C for 30 min. Cells were then washed once with 1 ml freshly made RNase-free 2X SSC, pelleted, and then resuspended in 300 µl 2X SSC. The samples were kept on ice until sorting.

Controls

Activated CD4$^+$ T cells infected with NL4-3 virus 2–3 days prior to HIV-1 SortSeq were used as positive controls. Samples were fixed, permeabilized, and hybridized in three different aliquots: 5’ HIV-1 RNA probes alone, 3’ HIV-1 RNA probes alone, and 5’ and 3’ HIV-1 RNA probes. Uninfected cells were treated with PMA/ionomycin side-by-side with the clinical samples and used as negative control. The uninfected cells, typically 25 million per experiment, were hybridized with both 5’ and 3’ HIV-1 RNA probes as negative controls.

Sorting

Cells were sorted by Sony SH800 cell sorter under purity mode. Cells were sorted into protein low-bind microfuge tubes containing 46.25 µl of reverse-cross linking buffer (100 mM NaCl (Thermo Scientific AM9759), 10 mM Tris 8.0 (Thermo Scientific AM9856), EDTA 1 mM (Thermo Scientific AM9260G), RNasin 40 U (Promega N2111). Tubes were immediately spun and snap frozen on dry ice. While cell sorters are capable of sorting into 96-well plates, sorting into individual microfuge tubes allow immediate processing to preserve RNA integrity. Sorted cells are stored at -80°C.

Reverse cross-linking

Cells in the reverse cross-linking buffer were incubated at 50°C for 1 hour on an Eppendorf Thermomixer after adding 2.5 µl of 10% SDS (Promega V6551) and 1.25 µl of 20 mg/ml proteinase K (Thermo Scientific AM2548) for each tube. Cells were placed on ice immediately after the incubation.

RNA extraction and library preparation

RNA was extracted using Zymo Quick-RNA MicroPrep (R1050) with on-column DNase treatment and eluted into 10 µl of H2O. The cDNA library was prepared using SMARTer Stranded Total RNA-Seq Kit v1 (samples from participant 21, 108, 159 and 361) or v2 (samples from participant 154, 218, 256, 348, 1001, 1002, 1022, 1023, 1025, 1027, 1029, UCSF2006 and UCSF2147) - Pico Input Mammalian Strand-Specific Illumina Sequencing Libraries (Takara) according to the manufacturer's suggestions. This method used random hexamers to construct the cDNA library of the total RNA followed by ribosomal RNA depletion to avoid 3' bias of mRNA capture using oligo-dT primers. cDNA concentration was measured by Qbit. The cDNA library quality was examined by Agilent Bioanalyzer. Samples with quality sufficient for RNAseq were normalized to equal molecular ratio. Samples with different barcodes were pooled into one high-output NextSeq 400M 2x150 or HiSeq 300M platforms, targeting the estimated reads of 20-40 million per sample.

HIV-1 landscape mapping in HIV-1 SortSeq

To generate cell-specific HIV-1 landscapes, mates of each set of preprocessed paired-end sequencing reads were aligned independently against the HXB2 HIV-1 reference genome with HISAT2 spliced end-to-end alignment option Alignments of two mates were merged together using SAMtools. All unaligned sequences were realigned against the HXB2 HIV-1 reference genome with Bowtie2 local alignment option and merged with the HISAT2 alignment to produce a full cell-specific HIV-1 RNA landscape. Sequences from individual single cells which were mapped to HIV-1 were inspected using the Integrative Genomic Browser (IGV, Broad Institute). These HIV-1 sequences were examined by HIV-1 BLAST (Los Alamos National Library HIV Sequence Database) to exclude plasmid contaminations and HIVAlign (Los Alamos National Library HIV Sequence Database) to examine the location in the HIV-1 genome. As there is no efficient way to check the purity of the rare SortSeq$^+$ positive cells from HIV-1-infected individuals, the HIV-1 RNA sequences are used as the quality control to ensure HIV-1 SortSeq$^+$ cells are authentic HIV-1-infected cells.

Determine the sensitivity of HIV-1 SortSeq

To test the sensitivity of detection, we sorted known number of NL4-3-BFP infected, BFP positive primary CD4$^+$ T cells into uninfected primary CD4$^+$ T cells (5 in 5 million, 50 in 5 million, 500 in 5 million and 5,000 in 5 million) and performed HIV-1 SortSeq to calculate the sensitivity of HIV-1-infected cells captured in HIV-1 SortSeq.

Quantifying the size of the latent reservoir using the viral outgrowth assay

The quantitative viral outgrowth assay was performed as previously described (*18*).

Confocal microscopy

Confocal microscopy was performed using Zeiss 710NLO Meta confocal microscope.

Viral RNA sequencing from viral outgrowth positive wells

Resting CD4$^+$ T cells were plated at a limiting dilution for viral outgrowth. Each end dilution well contains replication competent HIV-1 replicated from one HIV-1-infected cell. Cells were activated with Dynabeads human T-activator CD3/CD28 (ThermoFisher). One million MOLT-4/CCR5$^+$ cells were added to each well to allow exponential amplification of HIV-1 released into the culture. After 3 weeks, 180 µl of supernatant was subjected to p24 ELISA (PerkinElmer). Supernatant from p24 positive wells were collected for viral RNA isolation using Quick-RNA Viral Kit (Zymo Research). After cDNA synthesis using 3' *env* specific primer (ES8, CACTTCTCCAATTGTCCCTCA) and Superscript IV reverse transcriptase (ThermoFisher), V3-V4 of *env* was amplified as previously described using outer PCR primers ES7 (CTGTTAAATGGCAGTCTAGC) and ES8 (CACTTCTCCAATTGTCCCTCA) and inner PCR primers DLoop (GTCTAGCAGAAGAAGAGG) and Nesty8 (CATACATTGCTTTTCCTACT) as previously described (*39*) using Platinum Taq High Fidelity polymerase. PCR products examined on 1% agarose gels and bulk PCR products were extracted using a QIAquick Gel Extraction Kit (Qiagen) for Sanger sequencing. Sanger sequencing was used instead of TOPO cloning or deep sequencing because these end-dilution viral outgrowth culture wells contain replication competent viruses exponentially replicated from one HIV-1 provirus as previously described (*18*).

Phylogenetic analysis

We used the highly diverse V3-V4 region of HIV-1 *env* to determine whether different viruses originated from the same clone. Identical sequences in this region indicate clonal expansion of the infected cells as previously described (*26*, *39*). The neighbor-joining trees were constructed using MEGA7 (*77*). All positions containing gaps and missing data were eliminated.

Single-cell transcriptome expression analysis

To avoid batch effect in transcriptome analysis (due to different operator, sorting conditions and sequencing platform at Yale University versus Johns Hopkins University), only HIV-1 SortSeq performed at Yale University were included in transcriptome analysis. Reads were trimmed for quality and aligned to hg38 (gencode) using HiSAT2 (*78*). Alignments were processed using ballgown and per-gene counts were obtained (*77*). The Log_2(TPM+1) values were used for principal component analysis using Single Cell Toolkit (*79*). The resulting gene count matrix was then processed using DESingle (*74*). Briefly, DESingle normalizes using a modified median method, calculates a maximum likelihood estimation and detects differentially expressed genes using a Zero-Inflated Negative Bionomial (ZINB) model. DESingle's model overcomes technical dropout to classify differentially expressed genes into DE abundance, DE status, and DE general (indicated in data file S5). DEsingle was used to compare SortSeq$^+$ and SortSeq$^-$ cells using default parameters. P values less than 0.05 after Benjamini-Hochberg corrections were considered significant and used for further analysis (*75*). Gene ontology of the resulting gene list containing up-regulated and down-regulated genes was separately analyzed by Enrichr (*80*).

Construction of Jurkat T cell clones containing HIV-1 reporter proviruses at known integration sites

Jurkat T cells (NIH AIDS Reagents Program) were transduced with a single-round HIV-1 reporter virus NL4-3-d6-drEGFP (*55*) at a low multiplicity of infection (<1% GFP expression). This lentivirus has a full-length NL4-3 genome except that Env is replaced by a destabilized GFP. All viral genes were mutated with inactivating point mutations except for *tat* and *rev*. This HIV-1 reporter virus contains all HIV-1 splice sites and splice elements which allows splicing between HIV-1 and the host RNA. Three days after infection, GFP positive cells were sorted into single wells in 96 well plates. Three weeks after infection, cell line clones which grew into visible pellets were collected for expansion culture and integration site analysis using inverse PCR (*52*).

CRISPR-dCas9-based HIV-1 LTR-specific activation and inhibition system

Jurkat T cell clones (8B10, 1G2, 1D7 HIV-1-Jurkat cell line clones and uninfected Jurkat T cells) were transduced with EF1α-dCas9-VP64 activation domain tagged with a mCherry reporter (CRISPRa) or with EF1α-dCas9-Krab repression domain tagged with a mCherry reporter (CRISPRi)(*57*). Transduced cells were sorted by mCherry expression using flow cytometry. These are CRISPRa or CRISPRi-ready cells. Cells were transduced with lentiviral vectors (pU6-sgRNA-EF1α-puro-T2A-BFP)(*57*) carrying HIV-1-specific guide RNA (gRNA)(CTACAAGGGACTTTCCGCTG)(*58*) or non-targeting (NT) gRNA (GTATTACTGATATTGGTGGG)(*81*) with blue fluorescent protein (BFP) co-expression. U6-gRNA-BFP positive cells were sorted by flow cytometry and cultured for 4 weeks.

RNA landscape mapping in CRISPRa and CRISPRi Jurkat cell line clones

From CRISPRa and CRISPRi-ready, HIV-1-specific and non-targeting gRNA transduced Jurkat cell line clones (8B10, 1G2, 1D7 HIV-1-Jurkat cell line clones and uninfected Jurkat cells), 5,000 HIV-1-GFP positive cells from the CRISPRa/HIV-1-gRNA and CRISPRa/non-targeting-gRNA transduced cells were sorted into 50 µl of TRIzol. Similarly, 5,000 HIV-1-GFP positive cells from the CRISPRi/HIV-1-gRNA and CRISPRi/non-targeting-gRNA transduced cells were sorted into 50 µl TRIzol. After RNA extraction (Direct-zol RNA Microprep kit, Zymo Research R2060), library preparation (NEB Next single cell/low input RNA library prep kit for Illumina, NEB E6420L) and HiSeq 4000 2x150 sequencing. Adapters and raw reads with nucleotide read score lower than 30 were trimmed using Trimmomatic (*82*). The trimmed reads were then aligned to GRCh38 and NL4-3-d6-drEGFP using STAR (*83*). Generated bam files were normalized using transcript per million (TPM) in deepTools (*84*). Output BigWigs with 1 nt bin size were visualized using integrative genomic viewer (IGV).

Western blot

Four million Jurkat T cell clones were lyzed with RIPA lysis and extraction buffer (Thermo Fisher 89900). Denatured proteins were examined on PVDF membranes transferred from NuPAGE 4-12% Bis-Tris protein gels and stained with primary antibodies (VAV1 (1:250, Thermo Fisher MA5-17198), Rap1B (36E1) Rabbit mAb (1:200, Cell Signaling 2326S) and GAPDH (1:1000, Thermo Fisher MA5-15738)).

qPCR

Gene expression level of *IMPDH1*, *JAK1*, *UPF2* and *IKBKB* were analyzed by qPCR with of the cDNA libraries (if the remaining aliquot from single-cell RNAseq is sufficient for qPCR reactions). qPCR was performed using 1 µl of cDNA library, PerfeCTa qPCR ToughMix (Quantabio) and commercially validated qPCR primer/probe sets (Taqman Gene Expression Assay, Thermo Fisher) spanning two exons of the genes of interest: *IMPDH1* (Hs04190080_gH), *JAK1* (Hs01026983_m1), *UPF2* (Hs00922385_m1), *IKBKB* (Hs01559460_m1) and *EF1A1* (Hs00951278_m1). Normalization to EF1A1 expression as 2^(-ΔCt) values. Samples with undetected levels were arbitrarily assigned as Ct value of 60 for normalization. The violin plots were made using PRISM 8.

Identification of HIV-1-host chimeric RNA using bioinformatics analysis

Integration site identification

A custom Python script was implemented to search for sites of DNA fusion from two distinct genomes. This script is publicly available at https://github.com/alevar/chimFinder. In our study the tool was applied to identify possible HIV-1 integration events in the human genome based on the generated alignments. Reads were classified as belonging to an integration site if the following requirements were met: 1) there existed two distinct alignments, one to HIV-1 and one to the human genome; 2) non-overlapping substrings of the alignments were at least 20 nucleotides long each and 3) the overlap or gap between the two alignments was no greater than 5 bases (*23*). Reads with alignments to homopolymeric sequences were removed by computing the topological entropy score of each aligned substring and evaluating the score against the expected value of topological entropy. Additionally, a sigmoid scoring function was implemented to apply a cumulative soft threshold for each suggested integration site and provide a confidence metric for the manual verification of the reads using UCSC Genome Browser and LANL HIV BLAST search.

Additionally, by considering HIV-1 and human genome alignments of both R1 and R2 reads together, we successfully applied our method to detect sites of integration which occur within the sequence fragment spanned by the two mates of the paired-end reads. We reported a DNA fusion without single base resolution if the following requirements were met: 1) both reads came from the same paired-end fragment; 2) each mate had at most 30 unmapped and at least 20 mapped nucleotides; 3) all other filtering and scoring requirements from the single-end analysis were met.

HIV-1 Database cleanup

Total of 2557 HIV-1 assembled genomes were initially obtained for the analysis from the NCBI viral genomes database. In order to detect possible contamination of HIV-1 assemblies with human genome, all sequences in the reference dataset were aligned with BLAST against GRCh38.p8. Several sequences were noted to contain large contiguous fragments with >95% identity to the human genome. Since large fragments of human genome contained in HIV-1 references would present significant challenges for the task of identifying retroviral integration sites in human cells, we discarded each of the sequences.

Accession Number	**Alignment Length**	**Sequence Length**	**Query Start**	**Query End**	**Chromosome**	**Human Reference Start**	**Human Reference End**
AF133821.1	368	10035	0	373	chr7	11920632	11921017
AF133821.1	600	10035	9418	10035	chr20	13289301	13289915
AY352275.1	565	10280	9714	10280	chr7	111059856	111060424
KT427793.1	424	9167	26	458	chr7	47371091	47371522
KX232608.1	161	9668	317	484	chr10	3801952	3802119
KX232608.1	192	9668	130	323	chr10	3801969	3802162
KX232618.1	142	9439	0	288	chr3	156439817	156439960
KX232624.1	222	9171	0	222	chr19	31627386	31627608
M19921.2	1178	14825	13647	14825	chr16	315977	317155
M19921.2	1585	14825	9694	11284	chr17	82206777	82208367

Read alignment

Datasets were preprocessed with TrimGalore by trimming adapter sequences, removing end bases with quality less than 5 as well as 3’ and 5’ Ns to eliminate bad queries from the alignments. Additionally, sequences shorter than 45 base pairs were discarded as containing insufficient information for the downstream analysis (*23*).

The initial set of reads for the inference of HIV-1 integrations sites was obtained by aligning sequencing data first against the HIV-1 reference based on the assemblies downloaded from

NCBI and then realigning all reads against the GRCh38 reference sequence. Both alignments were performed using Bowtie2 with the very-sensitive-local alignment option, which allows long soft-clipping on either end of the sequence. No secondary alignments were reported by Bowtie2.

As the data came from an RNA sequencing experiment, a certain number of spliced reads is presumed to exist in the data. Spliced reads, however, cannot be accurately aligned with Bowtie2. Misalignments of spliced reads can introduce false positives into results, if a high degree of sequence similarity exists between regions in one of the HIV-1 genomes to the region of the human genome near the splice site. A list of reads for which a spliced alignment exists, such that no bases are trimmed or clipped at either the 3' or the 5' end of the read (end-to-end alignment) was generated by re-aligning HIV-1 mapped reads to the human reference genome with the HISAT2 spliced aligner. All end-to-end spliced reads mapped by HISAT2 to GRCh38 were removed from the Bowtie2 alignments via a custom Python script as being confidently matched against a single reference genome.

Merging human and HIV-1 alignments

Integration events based on alignments against HIV-1 and human reference genomes were detected by a custom Python script. The script first merges two alignments together to produce a record for each read that mapped to both genomes. Each resulting record contains read alignment information, mapping qualities of both alignments and raw read sequences corresponding to each alignment of the read with respect to the start and end position on the template. To reduce effects of the PCR amplification bias and include counts of reads derived from only unique molecules in the original samples, we removed any reads that were marked as PCR-duplicates (bit 0x400) by the alignment software. To further eliminate low-quality and low-confidence alignments from our analysis, we removed any reads that did not pass the quality control (bit0x200) as well as any secondary or supplemental alignments reported by the alignment software (bit 0x100 and bit 0x800).

Filtering of the records based on intermediate scores

A custom Python program was implemented to search for sites of DNA fusion from two distinct genomes. In our study the tool was applied to identify possible HIV-1 integration events in the human genome based on the generated alignments.

First, all reads with an overlap or gap of more than 5 base pairs between two alignments within a read were discarded (*23*). Next, reads with alignment lengths less than 20 base pairs of either human or HIV-1 fragments outside the overlap were discarded (*23*). The alignment length score was computed for each aligned fragment of the read independently (Definition 1). The arithmetic mean of the two scores was taken as representation of the alignment length score of the entire sequence.

A large number of ambiguous records due to long homopolymeric sequences in alignments was observed upon inspecting the resulting data. These homopolymers are >10 continuous Cs or Gs which are likely sequencing artifacts not HIV-1 sequences. These records are detected and filtered out in part by the function of normalized topological entropy[10] (Definition 2) implemented in our method. The threshold value of entropy was computed as the expected value of topological entropy (Definition 3) of the shortest alignment length of 20 base pairs permitted in our analysis, thus yielding a threshold value of ~0.84. Entropy score was computed and evaluated against the threshold independently for HIV-1 and human aligned substrings of the read.

It follows from the definition of the expected value of topological entropy (Definition 3) that the output of the function is discrete and depends only on the value of n. Since the value of n does not change on the sequence length interval $[17,65]$ based on the definition of n (Definition 2), it follows that the expected value of the entropy function does not change either within that interval. This observation implies that the expected value will always be a slight underestimate of the actual entropy of sequences in the alignment. We chose not to apply more stringent criteria and to clear any additional ambiguity during the manual verification, while still maintaining higher specificity and sensitivity of the final results by being able to detect some low-complexity sequences. A full description of the entropy-based method can be found in the original publication on topological normalized entropy of DNA sequences (*85*).

After filtering reads based on predefined thresholds, an intermediate quality score (Definition 4) was computed for each read based on the topological entropy and alignment length scores (Definition 4). Since each score is normalized to $[0,1]$ scale, it is possible to compare sequences based on additive effects of the measurements.

Definition 1

Alignment length score of a sequence w with length $|w|$ is defined as

$$S(w) = \frac{\frac{|w| - m}{\sqrt{d + (|w| - m)^2}} + 1}{2}$$

Where m is the threshold value, defined as the value of $w \vee$ for which the tangent to the sigmoid function has the greatest positive value. And d is defined as the slope steepness factor which regulates how great the tangent value is at the threshold point m thus regulating the strength of separation of values by the threshold.

Definition 2

For each sequence w of length $w \vee$ and a 4-character alphabet, n is defined as a unique integer such that

$$4^n + n - 1 \leq |w| < 4^{n+1} + (n + 1) - 1$$

The topological entropy H_{top} of a sequence w is thus defined as

$$H_{top}(w) = \frac{log_4 p_{w_1^{4^n+n-1}}(n)}{n}$$

where $w_1^{4^n+n-1}$ is defined as the first $4^n + n - 1$ nucleotides of the sequence and $p_{w_1^{4^n+n-1}}(n)$ is defined as the total number of unique substrings of length n in the given sequence.

Definition 3

The expected value of normalized topological entropy $E_{[H_{top}]}$ is defined as

$$E_{[H_{top}]} = \frac{log_4 4^n - 4^n\left(1 - \frac{1}{4^n}\right)^{4^n}}{n}$$

where the value of n was computed with regard to the shortest alignment length. In our case, where the minimum permitted alignment length was set to 20 base pairs, the value of $n = 2$, $4^n + n - 1 = 17$ and yielding the expected topological entropy score $E_{[H_{top}]} \approx 0.84$

Definition 4

The intermediate integration site score I_w of a sequence w is defined as

$$I_w = \frac{H_{p_{HIV}} + H_{p_{Hum}}}{2} \times \frac{(S_{HIV} + S_{Hum})(1 - p_{length}) + 2p_{length}}{2}$$

where $H_{p_{HIV}}$ and $H_{p_{Hum}}$ are the outputs of the normalized topological entropy function on HIV-1 and human substrings of the read. H_{HIV} and H_{Hum} are the outputs of the alignment length scoring function on HIV-1 and human sequence fragments of the read. p_{length} is the minimum alignment length score value permitted.

Clustering of the reads

All reads identified at this point are clustered into groups based on the suggested integration site location on the human and HIV-1 genomes with a 5 base pair window, to account for errors in breakpoint inference from alignments. For each cluster the number of supporting reads is computed and the sequence with the highest intermediate score is reported as representative of the integration event along with the intermediate score. Final scores of the integration sites (Definition 6) are recomputed by adding a sigmoid score term for the number of supporting reads (Definition 5). The final scores are then used to further filter inferred integration sites based on a cumulative score threshold value of 0.75. The final scores are also used to sort the integration sites by likelihood.

All integration sites from the final output were manually verified by realigning representative sequences using UCSC BLAT and LANL HIV BLAST against the human reference genome and LANL HIV database respectively.

Definition 5

The score of the number of reads which support a given integration site x was defined as

$$C(x) = \frac{\frac{x - m}{\sqrt{d + (x - m)^2}} + 1}{2}$$

Where m is the threshold value, defined as the value of x for which the tangent to the sigmoid function has the greatest positive value. And d is the defined as the slope steepness factor which regulates how great the tangent value is at the threshold point m thus regulating the strength of separation of values by the threshold.

Definition 6

$$score = I_W \times (C(1 - p_{count}) + p_{count})$$

Where $I_W = max(\{I_w(w_1), \ldots, I_w(w_n)\})$ of a set of reads $\{w_1, \ldots, w_n\}$ which support a given integration site, C and p_{count} are the output of the read count scoring function and the minimum read count score value respectively.

Pseudo paired-end analysis

In order to investigate potential integration events which occurred between the two sequenced ends of a paired-end library fragment, we performed a pseudo paired-end analysis based on single-end alignments. A pair of reads was considered indicative of an integration event if one read was successfully mapped to the human genome and the other read was mapped to one of the HIV-1 genomes from the database. Parameters to the scoring and filtering functions were kept identical to the single-end analysis, however an additional threshold of at most 30 unmapped nucleotides for each mate was applied in order to exclude potential misalignments.

Annotation of integration sites

PyBedtools package is used to annotate integration sites based on the intersection with features from GRCh38 annotation. Integration sites were also appropriately marked if no intersection with the annotation was available.

Intron retention and cryptic exon analysis

We produced intron coordinate files for genes with observed integration events based on the GRCh38 annotation. In order to avoid transcript-specificity in our analysis, we only considered intron fragments that do not overlap any known exons from any transcript in the annotation. Spliced alignments to the human genome from the previous step in our pipeline were used to identify all reads overlapping a given gene of interest using the Bedtools software package (*86*). We computed breadth of coverage for intronic and exonic regions of the gene separately, and only kept exonic regions if full transcription was observed through 100% breadth of coverage. Additionally, we computed the depth of coverage for each base in the intronic and exonic regions.

In order to provide a comparative metric for intron retention over the genes of interest, we report the final depth of coverage for each intron as a fraction between the mean intron depth of coverage and the mean exon depth of coverage.

All scripts, tools, dependency lists and documentation designed as part of this study are publicly available on GitHub (https://github.com/alevar/chimFinder) and Zenodo (https://doi.org/10.5281/zenodo.3740882).

Manual sequence validation of HIV-1-host chimeric RNA

Filtered sequence reads were examined by UCSC BLAT to examine the HIV-1-human RNA junction on the GRCh38/hg38 assembly. Sequences containing homopolymers (such as >10 consecutive Cs or >10 consecutive Gs), if not identified in the previous pipeline, were examined and excluded. Alignments were examined from reads with the highest score. The human cDNA junction was identified from valid alignments with the highest score. The alignment was examined at the UCSC Genome Browser (using the human Dec. 2013 GRCh38/hg38 assembly) to determine the genome orientation and the location of the junction (intron, exon, splice junction, noncoding region, or intergenic region).

The sequences at the non-matching side of the junction were then examined using HIV-1 BLAST (Los Alamos National Library HIV Sequence Database) to determine whether the sequences at the non-matching side of the junction are valid HIV-1 genome. The orientation and location of the HIV-1 junction was then examined to determine whether the junction is a LTR end or a known splice site.

Sequences with clear junctions mapped to an HIV-1 LTR end and a human genome were considered a read-through transcription. The orientation (same or convergent) of the HIV-1 genome in relation to the human genome was determined by manual inspection at the UCSC BLAT.

Sequences with junctions mapped to the HIV-1 genome which are neither LTR ends nor splice sites were examined with caution, since these sequences may be artifacts of the deep sequencing library preparation process (*47*). These sequences were in conjunction with human genomic sequences which were not at known splice sites. While it is possible that both HIV-1 and human genome activates a novel splice site and produced such spliced transcripts, we cannot otherwise confirm the authenticity of these HIV-1-host chimeric RNA transcripts. Thus, these sequences were not reported as authentic HIV-1-host chimeric RNA.

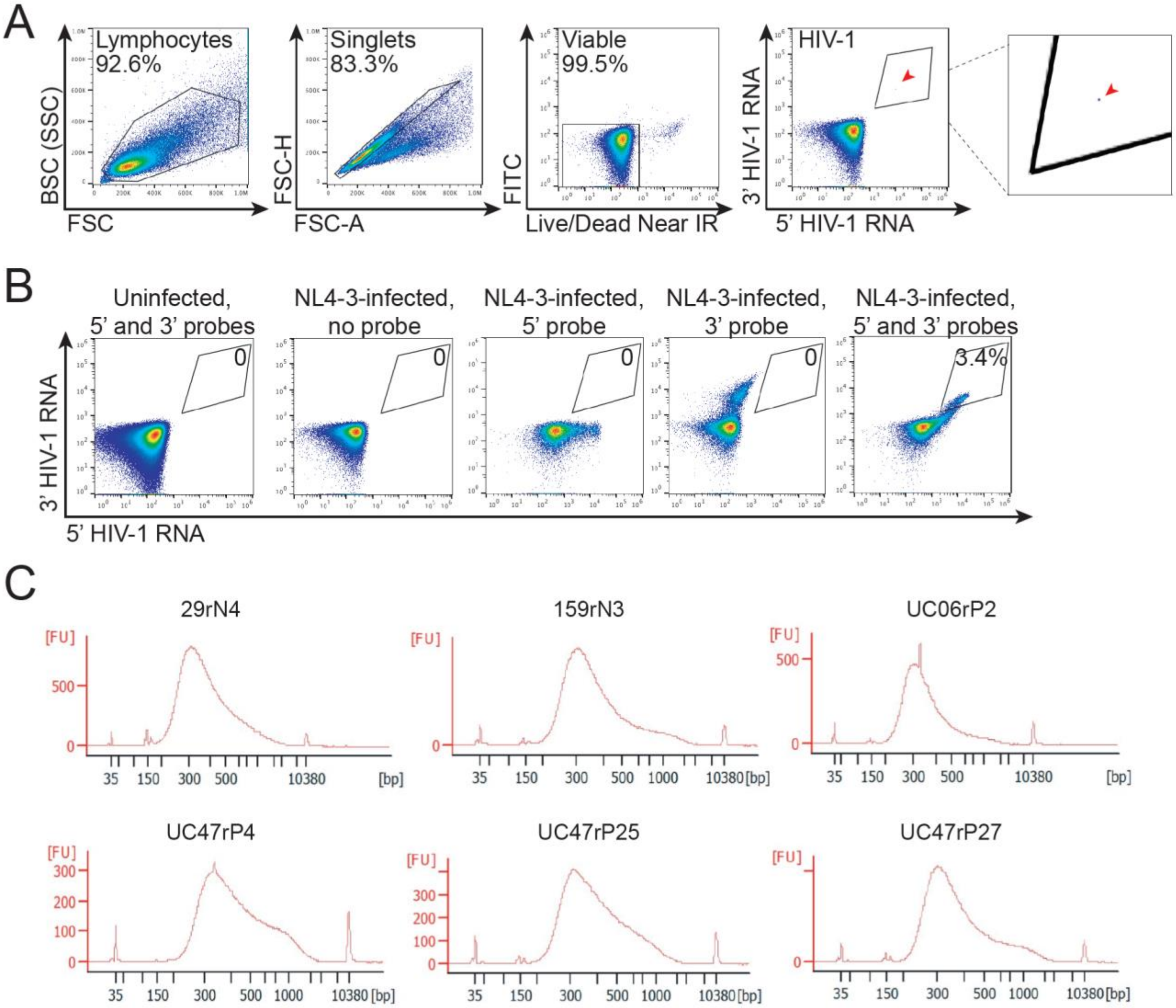


Fig. S1. Flow cytometry gating strategies and quality control of HIV-1 SortSeq. (**A**) Gating strategy of HIV-1 SortSeq. Cells are gated based on the size and granularity, doublet discrimination, viability, and HIV-1 5’ RNA and HIV-1 3’ RNA staining. (**B**) Activated primary CD4$^+$ T cells infected with NL4-3 reference strain demonstrate colocalized HIV-1 5’ and 3’ RNA expression. (**C**) After cDNA synthesis and library preparation, the concentration of the HIV-1 SortSeq$^+$ and SortSeq$^-$ cells were determined by Qbit fluoremetric quantitation. The size and quality of the library is analyzed by electropherogram (Agilent Bioanalyzer).

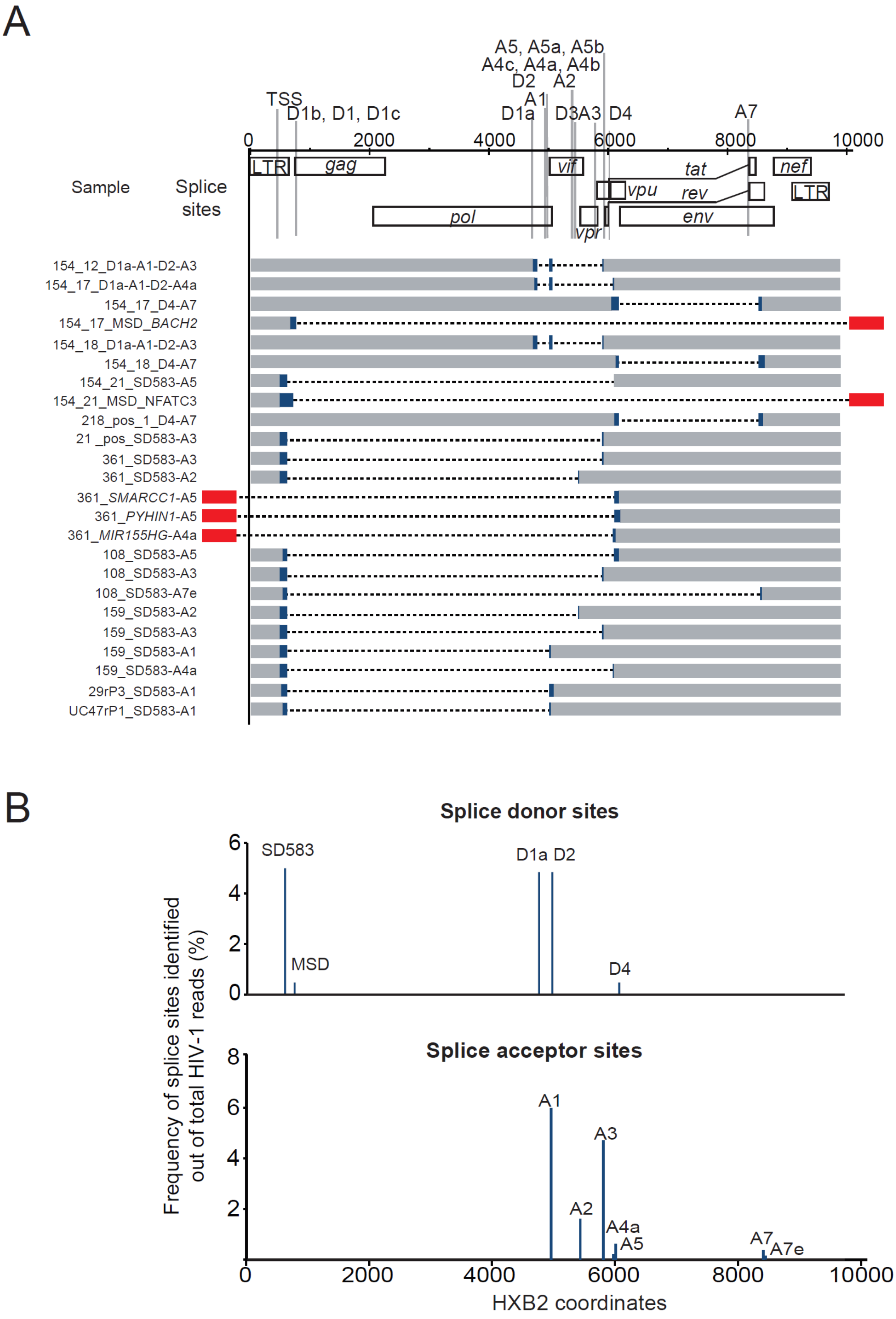

A
A5, A5a, A5b
A4c, A4a, A4b
D2 A2
A1
TSS
D1b, D1, D1c
D1a
D3A3 D4
A7
0
2000
4000
6000
8000
10000
LTR
gag
vif
tat
nef
vpu
rev
LTR
pol
env
vpr
Sample
Splice sites
154_12_D1a-A1-D2-A3
154_17_D1a-A1-D2-A4a
154_17_D4-A7
154_17_MSD_BACH2
154_18_D1a-A1-D2-A3
154_18_D4-A7
154_21_SD583-A5
154_21_MSD_NFATC3
218_pos_1_D4-A7
21 _pos_SD583-A3
361_SD583-A3
361_SD583-A2
361_SMARCC1-A5
361_PYHIN1-A5
361_MIR155HG-A4a
108_SD583-A5
108_SD583-A3
108_SD583-A7e
159_SD583-A2
159_SD583-A3
159_SD583-A1
159_SD583-A4a
29rP3_SD583-A1
UC47rP1_SD583-A1
B
Splice donor sites
6
4
2
0
SD583
MSD
D1a D2
D4
Splice acceptor sites
8
6
4
2
A1
A3
A2
A4a
A5
A7
A7e
0
2000
4000
6000
8000
10000
HXB2 coordinates
Frequency of splice sites identified out of total HIV-1 reads (%)

Fig. S2. HIV-1 SortSeq$^+$ cells from HIV-1–infected individuals harbor spliced HIV-1 RNA. (**A**) HIV-1 uses canonical splice sites to splice into the host gene (red bar) and HIV-1. The known splice sites were defined as previously described (*5*, *87*). The exact splice junction is captured on individual RNAseq reads. The blue bars represent the reads captured in RNAseq. The dashed lines indicate HIV-1 introns which are spliced out. The grey bars represent the sequences not captured in RNAseq. D: splice donors; A: splice acceptors. (**B**) The frequency of HIV-1 splice donor sites (D) and splice acceptor sites (A) were calculated based on the total reads mapped to HIV-1 in the single cells.

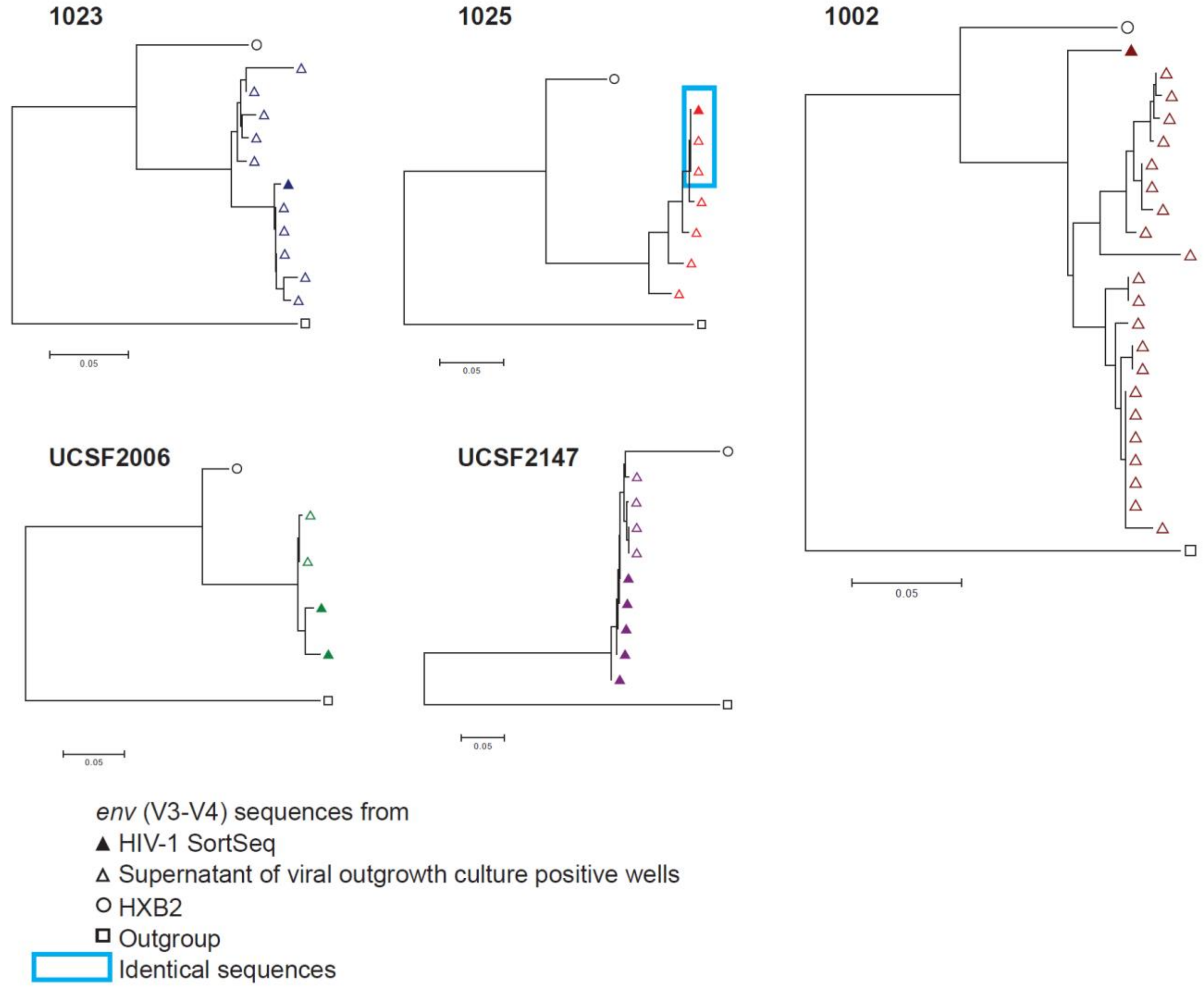


Fig. S3. HIV-1 SortSeq can identify clonally expanded HIV-1–infected cells harboring replication competent HIV-1. Phylogenetic analysis by the neighbor-joining method of *env* sequences (V3-V4) identifies clonally expanded CD4^{+} T cells containing replication-competent HIV-1 proviruses as previously described (*39*). Box indicates the same *env* sequences identified from HIV-1 SortSeq and replication-competent viruses from the viral outgrowth positive culture supernatant from participant 1025.

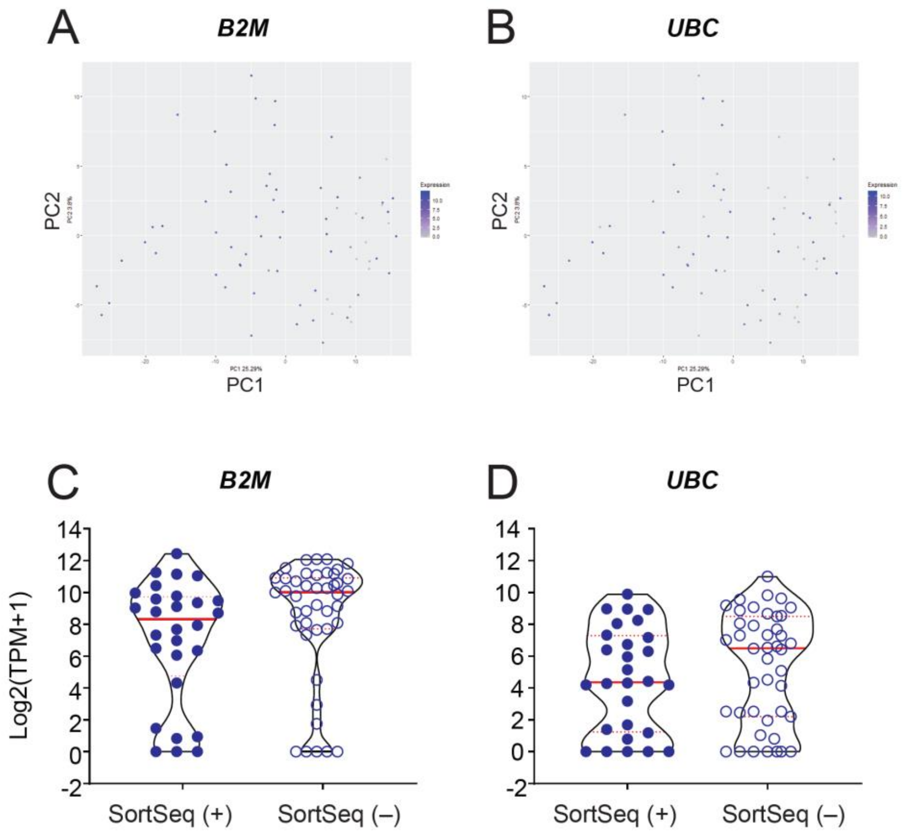


Fig. S4. Transcriptional profile of housekeeping genes *B2M* and *UBC* in HIV-1 SortSeq$^+$ and SortSeq$^-$ single cells. (**A**–**B**) The principal component analysis (PCA) demonstrates transcription levels of housekeeping genes *B2M* (**A**) and *UBC* (**B**) from ART-treated, virally suppressed, HIV-1-infected individuals. The PCA analysis was performed using Single Cell Toolkit (*79*). (**C**–**D**) Expression levels of selected significantly differentially expressed genes in HIV-1 SortSeq$^+$ versus HIV-1 SortSeq$^-$ single cells. Red lines denote median levels of expression. Dashed red lines denote 75$_{th}$ percentile of expression. P values calculated using DESingle (*74*) after Benjamini-Hochberg correction (*75*) do not reach statistical significance. The violin plots were made using PRISM.

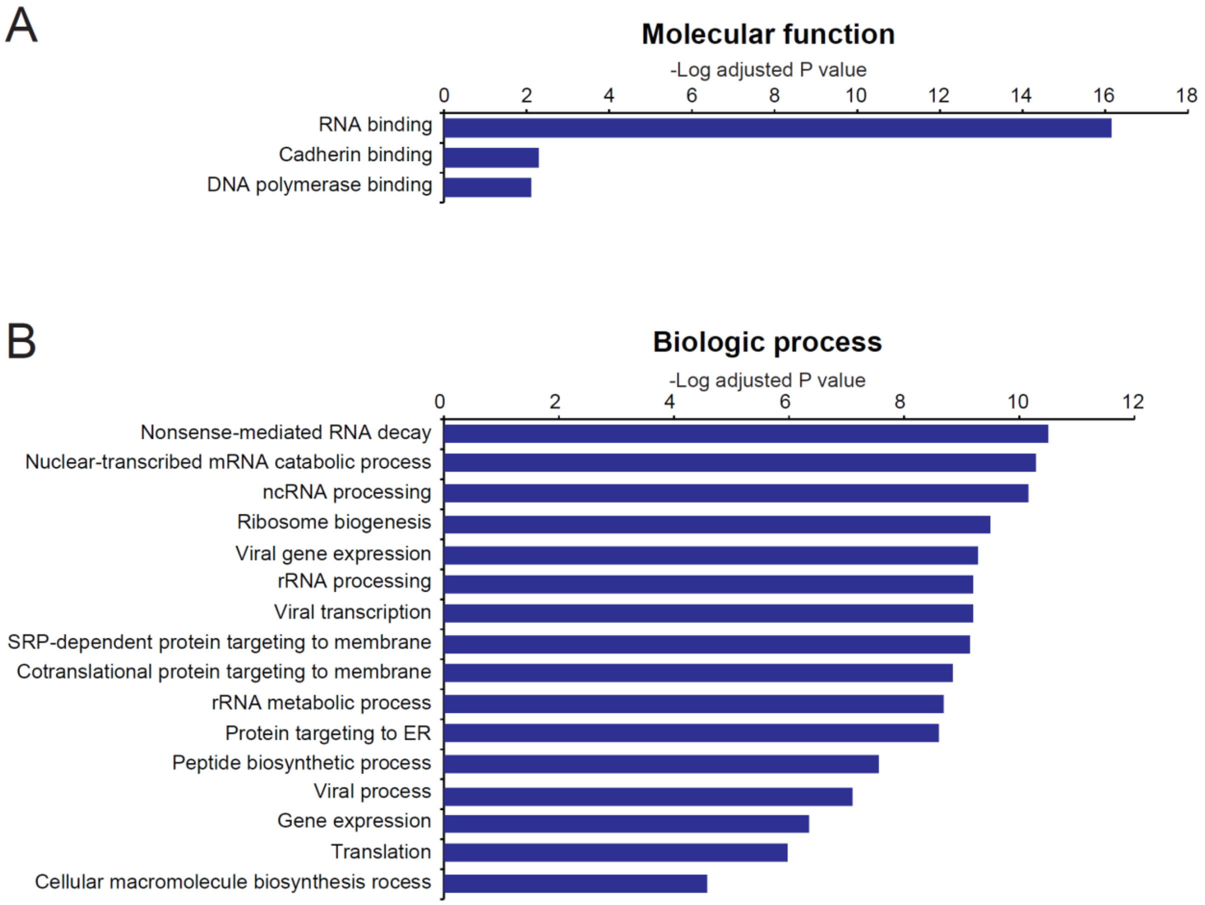


Fig. S5. Gene ontology analysis of differentially expressed genes in HIV-1 SortSeq$^+$ versus SortSeq$^-$ cells from ART-treated, virally suppressed, HIV-1–infected individuals. To understand the different transcriptional profiles in HIV-1 SortSeq$^+$ versus SortSeq$^-$ cells, we performed gene ontology (molecular function (**A**) and biologic function (**B**)) analysis of 395 up-regulated genes identified by DEsingle (*74*) was analyzed by Enrichr (*80*). P values less than 0.05 after Benjamini-Hochberg corrections were considered significant and used for differential gene expression analysis and gene ontology analysis (*75*). See data file S5 for the list of genes in each category.

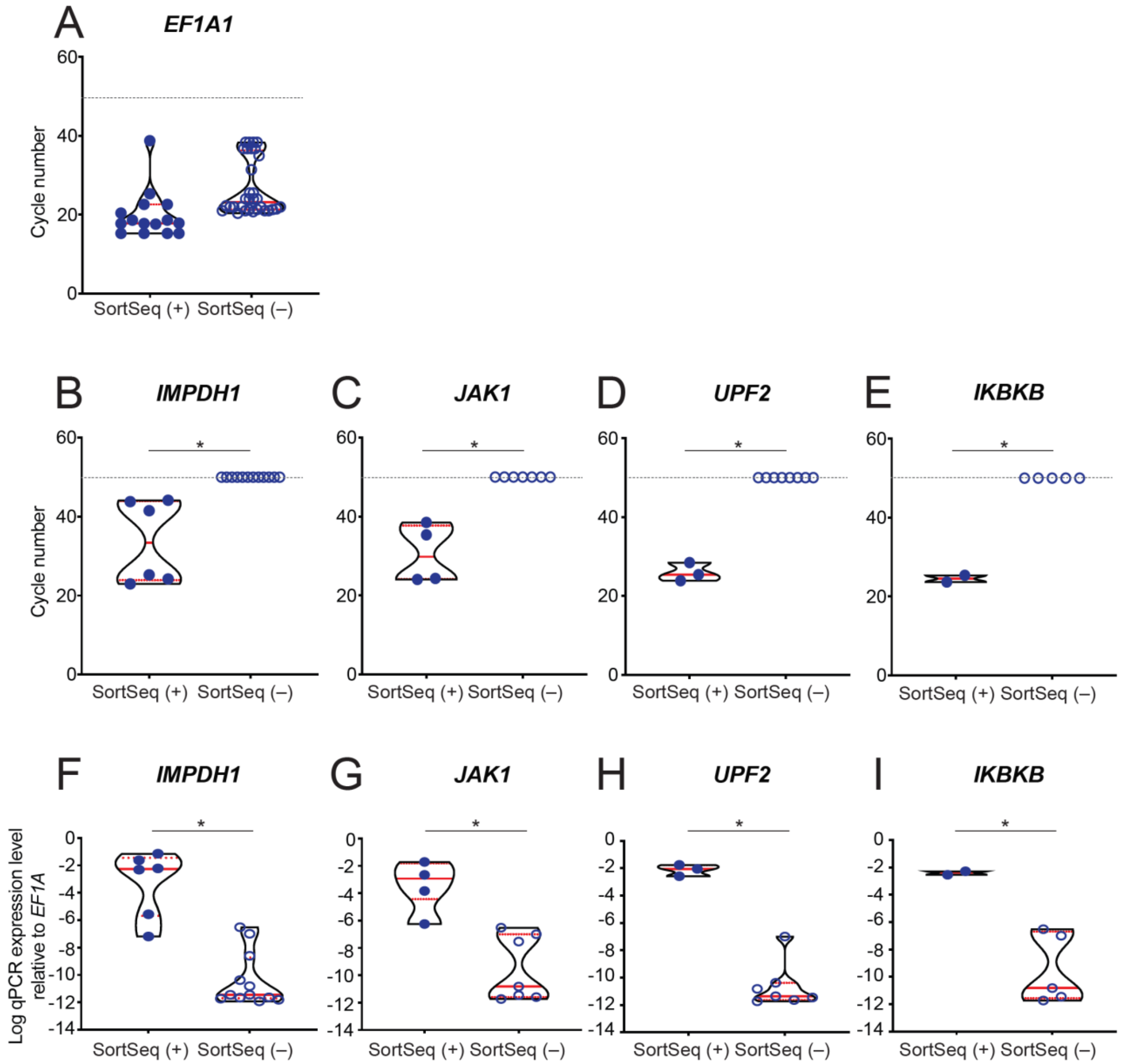
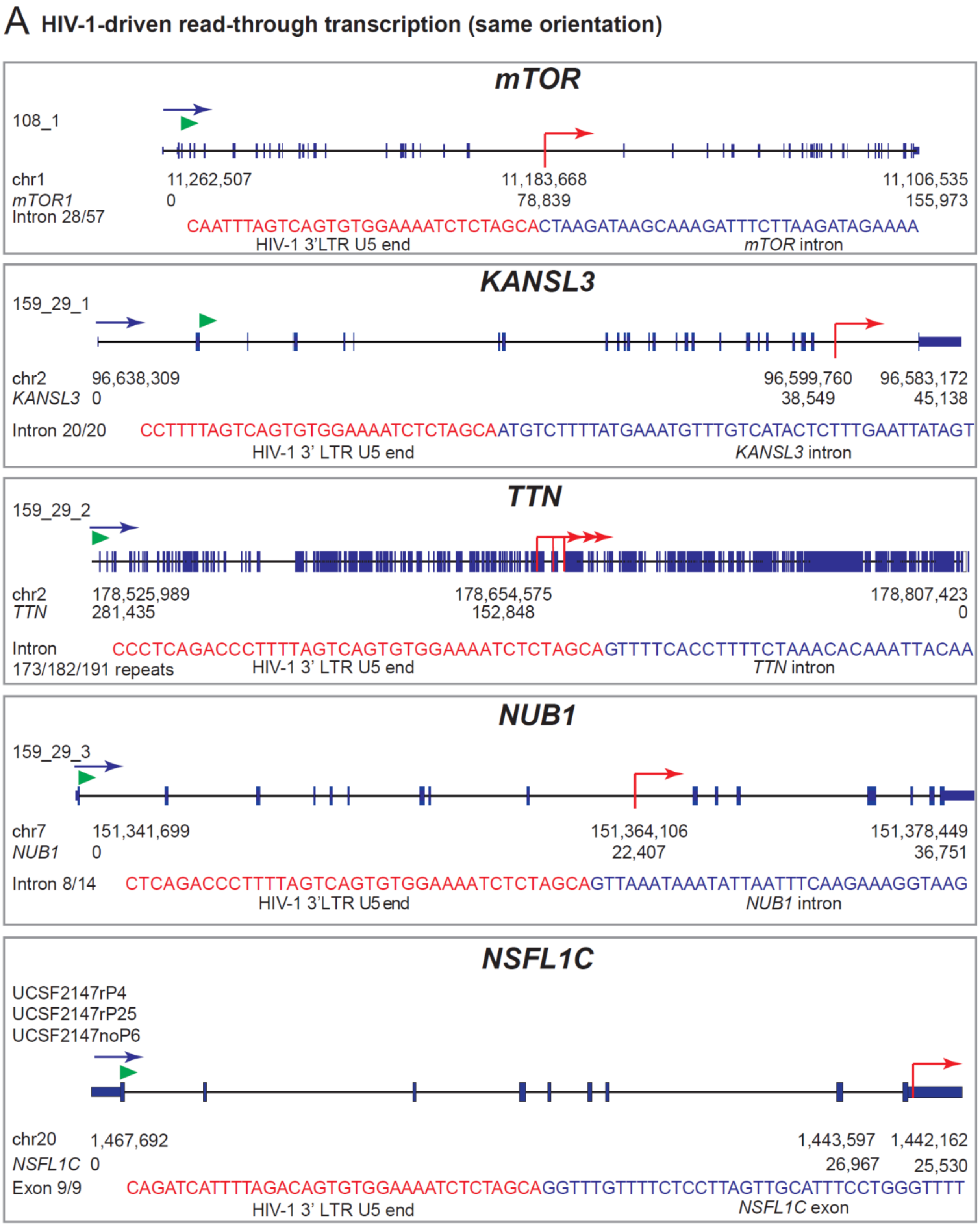
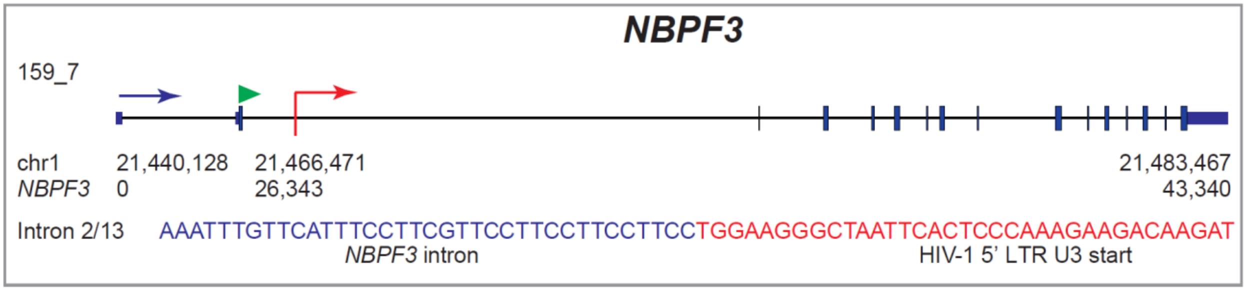


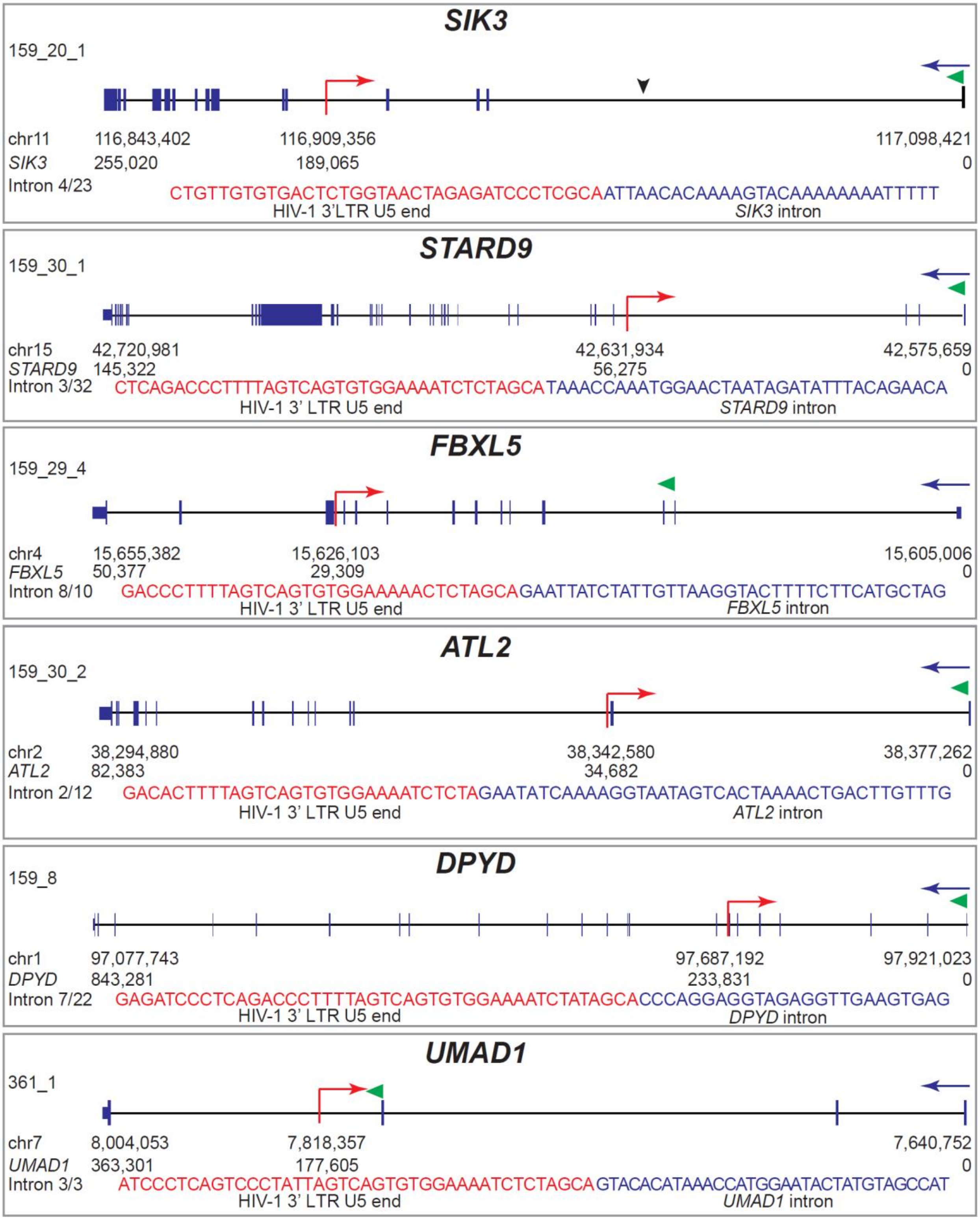
Fig. S6. Expression levels of *IMPDH1*, *JAK1*, *UPF2*, and *IKBKB* in HIV-1 SortSeq$^+$ and SortSeq$^-$ single cells measured by qPCR. The expression of selected up-regulated genes were measured by qPCR of the cDNA libraries of single cells from the HIV-1 SortSeq$^+$ and SortSeq$^-$ samples described in Fig. 3. **(A)** The expression level of *EF1A1* is readily measurable in all samples and used as a reference for normalization. (**B**–**E**) qPCR cycle numbers of *IMPDH1*, *JAK1*, *UPF2*, and *IKBKB* in HIV-1 SortSeq$^+$ and SortSeq$^-$ single cells. (**F**–**I**) *IMPDH1*, *JAK1*, *UPF2*, and *IKBKB* expression normalized to *EF1A1*. Gray dashed lines denote the detection limit of the qPCR cycle number 50. Asterisks denote P value <0.05 by Mann Whitney U Test.


A
HIV-1-driven read-through transcription (same orientation)
mTOR
108_1
chr1
11,262,507
11,183,668
11,106,535
mTOR1
0
78,839
155,973
Intron 28/57
CAATTTAGTCAGTGTGGAAAATCTCTAGCACTAAGATAAGCAAAGATTTCTTAAGATAGAAAA
HIV-1 3'LTR U5 end
mTOR intron
KANSL3
159_29_1
chr2
96,638,309
96,599,760
96,583,172
KANSL3
0
38,549
45,138
Intron 20/20
CCTTTTAGTCAGTGTGGAAAATCTCTAGCAATGTCTTTTATGAAATGTTTGTCATACTCTTTGAATTATAGT
HIV-1 3' LTR U5 end
KANSL3 intron
TTN
159_29_2
chr2
178,525,989
178,654,575
178,807,423
TTN
281,435
152,848
0
Intron
173/182/191 repeats
CCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGTTTTCACCTTTTCTAAACACAAATTACAA
HIV-1 3' LTR U5 end
TTN intron
NUB1
159_29_3
chr7
151,341,699
151,364,106
151,378,449
NUB1
0
22,407
36,751
Intron 8/14
CTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGTTAAATAAATATTAATTTCAAGAAAGGTAAG
HIV-1 3'LTR U5 end
NUB1 intron
NSFL1C
UCSF2147rP4
UCSF2147rP25
UCSF2147noP6
chr20
1,467,692
1,443,597
1,442,162
NSFL1C
0
26,967
25,530
Exon 9/9
CAGATCATTTTAGACAGTGTGGAAAATCTCTAGCAGGTTTGTTTTCTCCTTAGTTGCATTTCCTGGGTTTT
HIV-1 3' LTR U5 end
NSFL1C exon

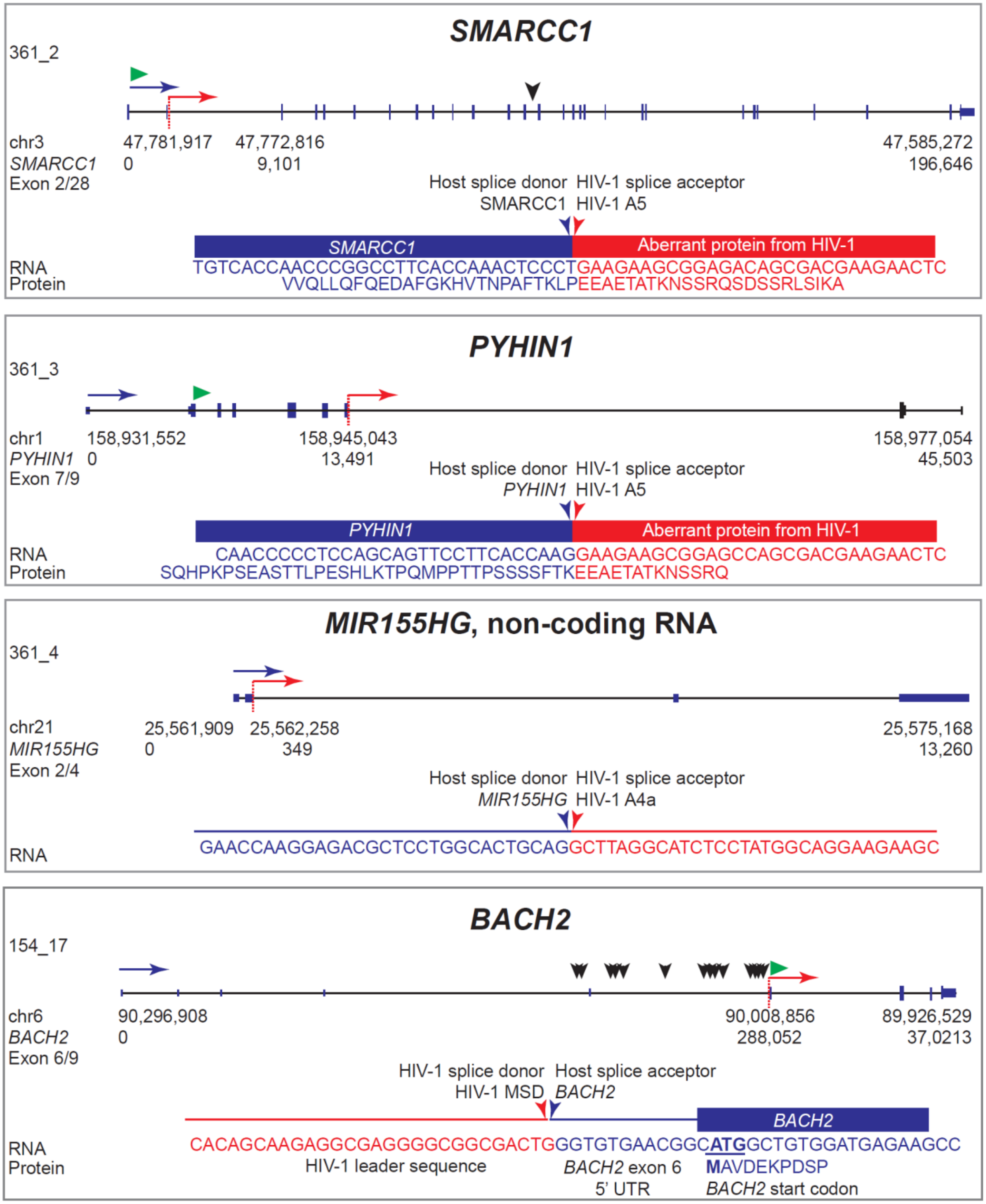
B **Host-driven read-through transcription (same orientation)**



C **HIV-1-driven read-through transcription (convergent orientation)**


SIK3
159_20_1
chr11 116,843,402 116,909,356 117,098,421
SIK3 255,020 189,065 0
Intron 4/23
CTGTTGTGTGACTCTGGTAACTAGAGATCCCTCGCAATTAACACAAAAGTACAAAAAAAATTTTT
HIV-1 3'LTR U5 end
SIK3 intron
STARD9
159_30_1
chr15 42,720,981 42,631,934 42,575,659
STARD9 145,322 56,275 0
Intron 3/32
CTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCATAAACCAAATGGAACTAATAGATATTTACAGAACA
HIV-1 3' LTR U5 end
STARD9 intron
FBXL5
159_29_4
chr4 15,655,382 15,626,103 15,605,006
FBXL5 50,377 29,309 0
Intron 8/10
GACCCTTTTAGTCAGTGTGGAAAAACTCTAGCAGAATTATCTATTGTTAAGGTACTTTTCTTCATGCTAG
HIV-1 3' LTR U5 end
FBXL5 intron
ATL2
159_30_2
chr2 38,294,880 38,342,580 38,377,262
ATL2 82,383 34,682 0
Intron 2/12
GACACTTTTAGTCAGTGTGGAAAATCTCTAGAATATCAAAAGGTAATAGTCACTAAAACTGACTTGTTTG
HIV-1 3' LTR U5 end
ATL2 intron
DPYD
159_8
chr1 97,077,743 97,687,192 97,921,023
DPYD 843,281 233,831 0
Intron 7/22
GAGATCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTATAGCACCCAGGAGGTAGAGGTTGAAGTGAG
HIV-1 3' LTR U5 end
DPYD intron
UMAD1
361_1
chr7 8,004,053 7,818,357 7,640,752
UMAD1 363,301 177,605 0
Intron 3/3
ATCCCTCAGTCCCTATTAGTCAGTGTGGAAAATCTCTAGCAGTACACATAAACCATGGAATACTATGTAGCCAT
HIV-1 3' LTR U5 end
UMAD1 intron

D Aberrant host-to-HIV-1 and HIV-1-to-host RNA splicing



↱ Host transcription orientation and transcription start site
▶ Host translation start site
↱ HIV-1-host chimeric RNA junction
↱ HIV-1 integration site
▼ HIV-1 integration site previously identified in HIV-1-infected individuals (Maldarelli *et al.* and Wagner *et al.*)

Fig. S7. HIV-1–host chimeric RNA landscape. (A) Read-through HIV-1-host chimeric RNA transcripts of HIV-1 to host with the same orientation were identified from HIV-1 SortSeq$^+$. The chimeric RNA sequences read from U5 end of HIV-1 3' LTR to host genes *mTOR*, *KANSL3*, *TTN*, *NUB1*, or *NSFL1C*. **(B)** Read-through RNA transcript of host *NBPF3* to HIV-1 with the same orientation were identified. The transcript sequence reads from *NBPF3* intron to U3 start of HIV-1 5'LTR. **(C)** Read-through RNA transcripts were identified when HIV-1 is integrated in the convergent orientation as the host transcription unit. Transcripts read from U5 end of HIV-1 3' LTR to the intron of host *SIK3*, *STARD9*, *FBXL5*, *ATL2*, *DPYD* or *UMAD1*. **(D)** Aberrant splicing transcripts from host canonical splice sites (blue arrowhead) into HIV-1 canonical splice acceptor sites (A4a and A5, red arrowhead) is captured in *SMARCC1*, *PYHIN1* and *MIR155HG*. HIV-1 splicing from the major splice donor (MSD or D1, red arrowhead) into the canonical acceptor site of the host gene *BACH2* was identified. HIV-1 genome is shown in red and human genome is shown in blue.

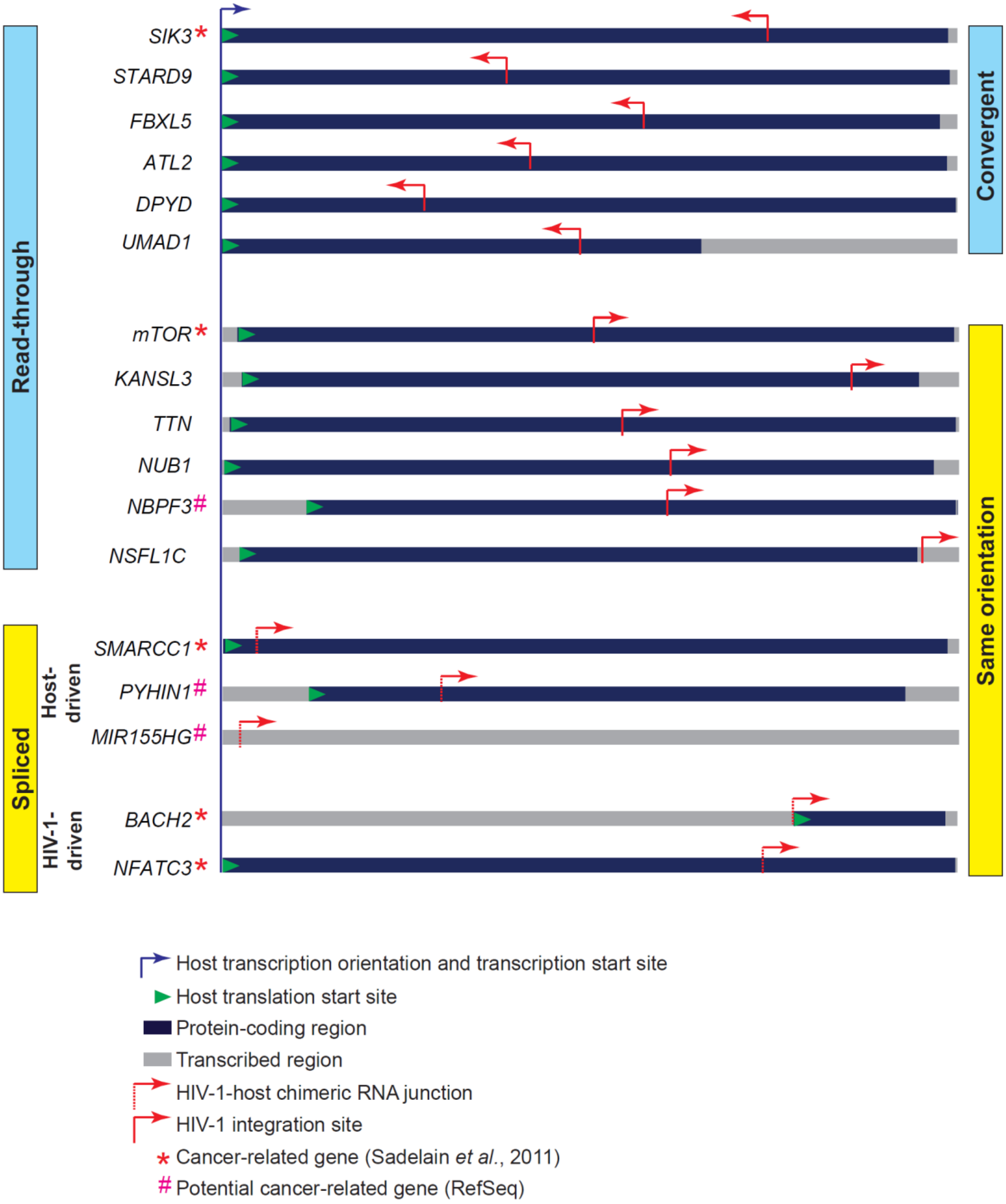


Fig. S8. Orientation and integration sites of induced HIV-1 proviruses in HIV-1 SortSeq$^+$ cells. The distance and orientation between the HIV-1-host chimeric RNA, the host gene transcriptional start site and the host gene translational start site were normalized to the length of the transcription unit. *, Cancer-related genes in the compiled cancer-related gene list (*49*). $^{\#}$, genes which are not categorized in the cancer-related gene list but related to cancer formation based on RefSeq descriptions.

Table S1. Characteristics of study participants.

ID	Age	Gender	Ethnicity	Current ART	Viral load (copies/ml)	Duration on ART (month)	Duration of undetectable viral load (month)	Analysis involved
21	53	M	AA	EFV/FTC/TDF	<50	119	26	B, V
29	54	M	W	ABC/DTG/3TC	<50	49	35	V
63	66	F	AA	FTC/RPV/TAF, DRV/r, DTG	<50	267	99	V
108	51	F	AA	ABC/DTG/3TC	<50	208	95	B, V*
111	59	F	AA	TAF/FTC	<50	123	87	V
154	69	M	AA	DRV/r, DTG	<50	168	12	B, S
159	57	M	AA	ABC/DTG/3TC	<50	77	77	B, S, V, T ,D
218	67	M	W	ABC, EFV, 3TC, RAL	<50	310	62	B, S
256	47	F	AA	FTC/TAF, RAL	<50	65	63	B, S
348	53	F	AA	FTC/RPV/TAF, DTG	<50	127	115	B, S
361	37	F	AA	EFV/FTC/TDF	<50	56	53	B
384	62	M	W	FTC/RPV/TAF	<50	184	180	V
404	70	M	AA	ABC/DTG/3TC	<50	248	65	V
1001	50	M	AA	FTC/ RPV/ TAF	<20	81	75	S, T
1002	47	M	AA	FTC/RPV/TDF	<20	288	57	S, T, P
1004	59	M	AA	EVG/COBI/FTC/TAF	<20	222	119	D
1009	57	M	AA	ABC/DTG/3TC	<20	230	89	D
1022	54	M	AA	FTC/TDF, DTG, DRV/r	<20	105	47	S, T
1023	51	M	W	ATV/r, FTC/TAF	<20	168	53	S, T, P
1025	47	M	W	FTC/RPV/TAF	<20	68	57	S, T, P
1027	53	M	Hispanic	FTC/RPV/TAF	<20	240	69	S, T
1029	58	F	AA	ATV, ddl, 3TC	<50	181	56	S, T
1035	67	F	AA	EFV/FTC/TDF	<40	326	83	D
UCSF2006	68	M	W	ABC/DTG/3TC	<40	253	238	S, T, P, D
UCSF2147	62	M	Asian	BIC/FTC/TAF	<40	246	172	S, T, P, D

AA, African American; W, White/Caucasian; 3TC, lamivudine; ABC, abacavir; ATV, atazanavir sulfate; BIC, Bictegravir; ddI, didanosine; DRV, darunavir; DTG, dolutegravir; EFV, efavirenz; FTC, emtricitabine; RAL, raltegravir; /r, ritonavir boost; RPV, rilpivirine; TAF, tenofovir alafenamide; TDF, tenofovir disoproxil.

Analysis involved: B, bulk HIV-1 SortSeq; D, drug treatment; P, phylogenetic analysis of viral outgrowth positive cultures; S, single HIV-1 SortSeq; T, single-cell transcriptome analysis; V, quantitative viral outgrowth; V*, quantitative viral outgrowth at two different time points.

Table S2. Genes in which HIV-1 is integrated.

		Sample name	Gene	Cancer-related gene	T cell acti-vation	House-keeping	Gene full name	Role
Read-through transcription	Same direction	108_bulk_1	*MTOR*	Yes	No	No	Mechanistic target of rapamycin (serine/ threonine kinase)	A central regulator of cellular metabolism, growth and survival in response to hormones, growth factors, nutrients, energy and stress signals. Mediate cellular responses to stresses such as DNA damage and nutrient deprivation.
		159_29_bulk_1	*KANSL3*	No	No	Yes	KAT8 regulatory NSL complex subunit 3	Acetylation of nucleosomal histone H4 on several lysine residues and therefore may be involved in the regulation of transcription.
		159_29_bulk_2	*TTN*	No	No	No	Titin	Assembly and functioning of vertebrate striated muscles.
		159_29_bulk_3	*NUB1*	No	No	Yes	Negative regulator of ubiquitin-like proteins 1	The protein encoded by this gene regulates the NEDD8 conjugation system post-transcriptionally by recruiting NEDD8 and its conjugates to the proteasome for degradation.
		159_7	*NBPF3*	Potential	No	No	Neuroblastoma breakpoint family member 3	Positive regulation of protein targeting to mitochondrion. Altered expression of some gene family members is associated with several types of cancer.
		UC47rP4 UC47rP25 UC47noP6	*NSFL1C*	No	No	Yes	N-ethylmaleimide-sensitive factor 1C	ATPases known to be involved in transport vesicle/target membrane fusion and fusions between membrane compartments
	Convergent direction	159_20_bulk_1	*SIK3*	Tumor suppressor	No	No	SIK family kinase 3 (SIK3)	Serine/threonine-protein kinase.
		159_30_bulk_1	*STARD9*	No	No	No	StAR-related lipid transfer protein 9	Microtubule-dependent motor protein required for spindle pole assembly during mitosis.
		159_29_bulk_4	*FBXL5*	No	No	Yes	F-box and leucine-rich repeat protein 5	Iron homeostasis by promoting the ubiquitination and subsequent degradation of IREB2/IRP2.
		159_30_bulk_2	*ATL2*	No	No	Yes	Atlastin GTPase 2	GTPase tethering membranes through formation of trans-homooligomers and mediating homotypic fusion of endoplasmic reticulum membranes.
		159_8	*DPYD*	No	No	No	Dihydropyrimidine dehydrogenase [NADP(+)]	Uracil and thymidine catabolism.
		361_bulk_1	*UMAD1*	No	No	No	UBAP1-MVB12-associated (UMA)-domain containing protein 1	A protein coding gene; function unclear.
Spliced chimeric RNA	Host to HIV-1	361_bulk_2	*SMARCC1*	Haplo-insufficiency	No	No	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1	Apart of the large ATP-dependent chromatin remodeling complex SNF/SWI and contains a predicted leucine zipper motif typical of many transcription factors.
		361_bulk_3	*PYHIN1*	Potential	No	No	Pyrin and HIN domain family member 1	Transcriptional regulation of genes important for cell cycle control, differentiation, and apoptosis. Down-regulation of this gene is associated with breast cancer. This protein acts as a tumor suppressor by promoting ubiquitination and subsequent degradation of MDM2, which leads to stabilization of p53/TP53.
		361_bulk_4	*MIR155HG*	Potential	No	No	MIR155 host gene	The long RNA transcribed from this gene is expressed at high levels in lymphoma and may function as an oncogene.
	HIV-1 to host	154_17	*BACH2*	Yes	No	No	BTB domain and CNC homolog 2	Plays an important role in coordinating transcription activation and repression by MAFK. Induces apoptosis in response to oxidative stress through repression of the antiapoptotic factor HMOX1.
		154_21	*NFATC3*	Yes	No	No	Nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 3 (NFATC3)	A member of the nuclear factors of activated T cells DNA-binding transcription complex. Regulation of gene expression in T cells and immature thymocytes.

Table S3. HIV-1–host RNA junctions.

		Sample name	Gene	chr	chr strand	Host gene junction	HIV-1 junction
Read-through transcription	Same direction	108_bulk_1	*MTOR*	chr1	-	11,183,668	3' LTR (U5 end)
		159_29_bulk_1	*KANSL3*	chr2	-	96,599,760	3' LTR (U5 end)
		159_29_bulk_2	*TTN*	chr2	-	178,654,575	3' LTR (U5 end)
		159_29_bulk_3	*NUB1*	chr7	+	151,364,106	3' LTR (U5 end)
		159_7	*NBPF3*	chr1	+	21,466,471	5' LTR (U3 begins)
		UC47rP4 UC47rP25 UC47noP6	*NSFL1C*	Chr20	-	1,443,598	3' LTR (U5 end)
	Convergent direction	159_20_bulk_1	*SIK3*	chr11	-	116,909,356	3' LTR (U5 end)
		159_30_bulk_1	*STARD9*	chr15	+	42,631,934	3' LTR (The beginning of R and the end of U5)
		159_29_bulk_4	*FBXL5*	chr4	-	15,626,103	3' LTR (The beginning of R and the end of U5)
		159_30_bulk_2	*ATL2*	chr2	-	38,342,580	3' LTR (U5 end)
		159_8	*DPYD*	chr1	-	97,687,192	3' LTR (U5 end)
		361_bulk_1	*UMAD1*	chr3	+	7,818,357	3' LTR (U5 end)
Spliced chimeric RNA	Host to HIV-1	361_bulk_2	*SMARCC1*	chr3	-	47,772,816	A5
		361_bulk_3	*PYHIN1*	chr1	+	158,945,043	A5
		361_bulk_4	*MIR155HG*	Chr21	+	25,562,258	A4a
	HIV-1 to host	154_17	*BACH2*	chr6	-	90,008,856	MSD
		154_21	*NFATC3*	chr16	+	68,183,240	The beginning of R - U5; MSD