Published online 25 August 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 20* ***10345–10355***
doi:10.1093/nar/gks753

Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing

Karen E. Ocwieja[1], Scott Sherrill-Mix[1], Rithun Mukherjee[2], Rebecca Custers-Allen[1], Patricia David[3], Michael Brown[4], Susana Wang[4], Darren R. Link[3], Jeff Olson[3], Kevin Travers[4], Eric Schadt[4,5] and Frederic D. Bushman[1,*]

[1]Perelman School of Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA, [2]Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, [3]RainDance Technologies, Lexington, MA, [4]Pacific Biosciences, Menlo Park, CA and [5]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York City, NY, USA

Received May 17, 2012; Revised July 12, 2012; Accepted July 17, 2012

ABSTRACT

Alternative RNA splicing greatly expands the repertoire of proteins encoded by genomes. Next-generation sequencing (NGS) is attractive for studying alternative splicing because of the efficiency and low cost per base, but short reads typical of NGS only report mRNA fragments containing one or few splice junctions. Here, we used single-molecule amplification and long-read sequencing to study the HIV-1 provirus, which is only 9700 bp in length, but encodes nine major proteins via alternative splicing. Our data showed that the clinical isolate HIV-$1_{89.6}$ produces at least 109 different spliced RNAs, including a previously unappreciated ~1 kb class of messages, two of which encode new proteins. HIV-1 message populations differed between cell types, longitudinally during infection, and among T cells from different human donors. These findings open a new window on a little studied aspect of HIV-1 replication, suggest therapeutic opportunities and provide advanced tools for the study of alternative splicing.

INTRODUCTION

Alternative splicing greatly expands the information content of genomes by producing multiple mRNAs from individual transcription units. Approximately 95% of human genes with multiple exons encode RNA transcripts that are alternatively spliced, and mutations that affect alternative splicing are associated with diseases ranging from cystic fibrosis to chronic lymphoproliferative leukemia (1–5). Work to decipher an RNA 'splicing code' has revealed that multiple interactions between *trans*-acting factors and RNA elements determine splicing patterns, though regulation is little understood for most genes (6).

The integrated HIV-1 provirus is ~9700 bp in length and has a single transcription start site, but according to the published literature yields at least 47 different mRNAs encoding 9 proteins or polyproteins, making HIV an attractive model for studies of alternative splicing (7). HIV mRNAs fall into three classes: the unspliced RNA genome, which encodes Gag/Gag-Pol; partially spliced transcripts, ~4 kb in length, encoding Vif, Vpr, a one-exon version of Tat, and Env/Vpu; and completely spliced mRNAs of roughly 2 kb encoding Tat, Rev and Nef (Figure 1A). Additional rare 'cryptic' splice donors (5′ splice sites) and acceptors (3′ splice sites) contribute even more mRNAs (8–13). A complex array of positive and negative *cis*-acting elements surrounding each splice site regulates the relative abundance of the HIV-1 mRNAs, and disrupting the balance of message ratios impairs viral replication in several models (14–21). Studies have suggested strain-specific splicing patterns may exist (7,22,23). However, detailed studies of complete message populations have not been reported for clinical isolates of HIV-1.

Several groups have demonstrated tissue- and differentiation-specific splicing of cellular genes (2,24,25). Importantly for HIV, these include changes during T-cell activation (26,27), raising the question of how cell-specific splicing affects HIV replication. While most studies of HIV-1 splicing have been conducted in cell lines using lab-adapted viral strains, limited works in PBMCs from infected patients, monocytes and macrophages have

*To whom correspondence should be addressed. Tel: +1 215 573 8732; Fax: +1 215 573 4856; Email: bushman@mail.med.upenn.edu

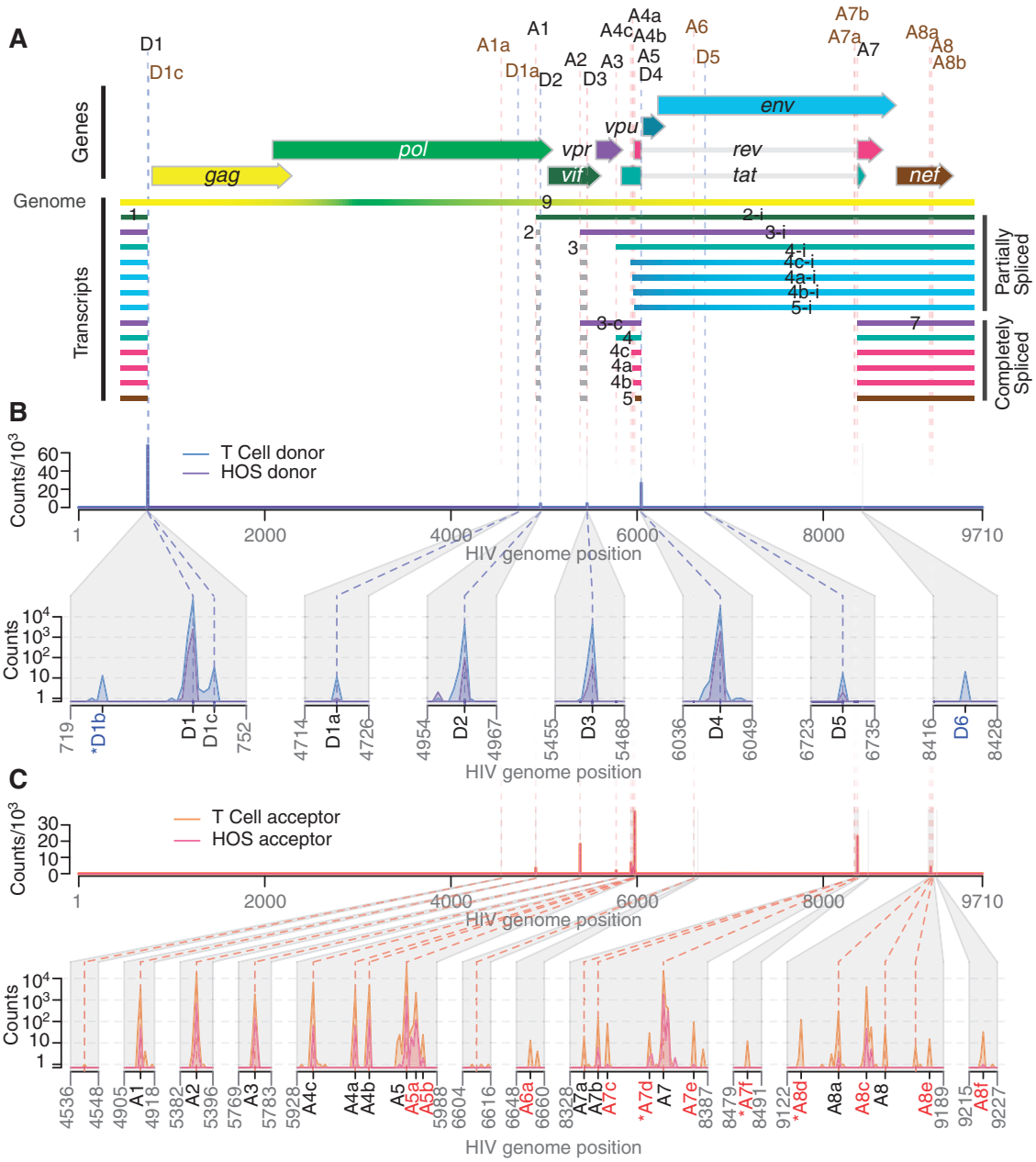

Figure 1. Mapping the splice donors and acceptors of HIV-$1_{89.6}$. PacBio® sequence reads of HIV-$1_{89.6}$ cDNA from infected HOS-CD4-CCR5 (HOS) and $CD4^+$ T cells were aligned to the HIV-$1_{89.6}$ genome shown in (**A**). Exons of the conserved HIV-1 transcripts are colored according to the encoded gene. Non-coding exons 2 and 3 are variably included in each transcript where indicated. Conserved (black) and published cryptic (brown) splice donors ('D') and acceptors ('A') are shown. Numbering is according to previous convention (7). Gaps in HIV-1 sequence alignments with at least one end located at a published or verified splice donor or acceptor were defined as introns. For each base of the HIV-$1_{89.6}$ genome, the number of sequence reads in which that base occurred at the 5′-end (**B**) or 3′-end (**C**) of an intron is plotted for each cell type. Putative splice donors and acceptors, numbered according to nearest published site, were defined as loci that were found in at least 10 reads to be at the 5′- and 3′-ends of introns, respectively, in sequence alignments from T-cell infections. Regions containing splice sites are enlarged for clarity. Coordinates of the splice donors and splice acceptors are provided in Supplementary Table S2. The novel acceptor A8c was further verified. Asterisks indicate putative splice donors and acceptors that are adjacent to dinucleotides other than the consensus GT and AG, respectively.

suggested that differences may indeed exist in relevant cell types (9,22,28,29). Moreover, human splicing patterns differ between individuals, but such polymorphisms have not been investigated in the context of HIV infection (30,31).

Here, we use deep sequencing to comprehensively characterize the transcriptome of an early passage clinical isolate, HIV-$1_{89.6}$ (32), in primary $CD4^+$ T cells from seven human donors and in the human osteosarcoma (HOS) cell line. Many deep sequencing techniques provide short reads, which rarely query more than a single exon–exon junction. To distinguish the full structure of HIV-1 mRNAs, which can contain several splice junctions, we used Pacific Biosciences (PacBio®)

sequencing technology, which yields read lengths up to 10 kb (33). We used RainDance Technologies single-molecule PCR enrichment to preserve ratios of RNAs during preparation of sequencing templates. We identified previously published and novel HIV-1 transcripts and determined that HIV-$1_{89.6}$ encodes a minimum of 109 different splice forms. These included a new size class of transcripts, some of which contain novel open reading frames (ORFs) that encode new proteins. We also found significant variation between cell types, over time during infection of HOS cells and among individuals. These data reveal unanticipated complexity and dynamics in HIV-1 message populations, begin to clarify a little studied dimension of HIV-1 replication and suggest possible targets for therapeutic interventions.

MATERIALS AND METHODS

Cell culture and viral infections

HIV-$1_{89.6}$ was generated by transfection and subsequent expansion in SupT1 cells. Primary T cells were isolated by the University of Pennsylvania Center for AIDS Research Immunology core and confirmed to be homozygous for the wild-type CCR5 allele as shown in Supplementary Table S1 and described in Supplementary Methods. HOS-CD4-CCR5 cells (34,35) were obtained through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH from Dr Nathaniel Landau. Single round infections in T cells and HOS-CD4-CCR5 cells were performed using standard methods (see Supplementary Methods).

RNA and reverse transcription

Total cellular RNA was purified using the Illustra RNA kit (GE Life Sciences, Fairfield, CT, USA) from 5×10^6 cells per infection. Viral cDNA was made using a reverse transcription primer complementary to a sequence in U3 (RTprime, Supplementary Table S2). We used Superscript III reverse transcriptase (Invitrogen) in the presence of RNaseOUT (Invitrogen) to conduct first-strand cDNA synthesis from equal amounts of total cellular RNA from each HOS-CD4-CCR5 time point (15.2 μg) and from each T-cell infection (3 μg) according to the manufacturer's instructions for gene-specific priming of long cDNAs, and then treated with RNaseH (Invitrogen). We checked for full reverse transcription of the longest (unspliced) viral cDNAs by PCR using primers that bind in the first major intron of HIV-$1_{89.6}$ (keo003, keo004, Supplementary Table S2, data not shown)

Bulk RT–PCR and cloning

Transcripts were amplified from cellular RNA using the Onestep RT-PCR kit (Qiagen) with primer pairs keo056/keo057 and keo058/keo059 (Supplementary Table S2) with the following amplification: 5 cycles of 30 s at 94°C, 12 s at 56°C, 40 s at 72°C; then 30 cycles of 30 s at 94°C, 14 s at 56°C, 40 s at 72°C; and finally 10 min at 72°C. For verification of dynamic changes, primers F1.2 and R1.2 were used with 35 cycles of 30 s at 94°C, 30 s at 56°C and 45 s at 72°C followed by 10 min at 72°C. Products were resolved on agarose gels (Nusieve 3:1, Lonza for verification of dynamic changes, Invitrogen for cloning) stained with ethidium-bromide (Sigma) for visualization, or SYBR Safe DNA gel stain (Invitrogen) for cloning (keo056/keo057 amplified material). DNA was purified using Qiaquick gel extraction kit (Qiagen) and cloned using the TOPO TA cloning kit (Invitrogen). Plasmid DNA was prepared using Qiaprep Spin Miniprep kit (Qiagen). Inserts were identified and verified using Sanger sequencing. The cDNAs for *tat^8c, tat* (1 and 2 exon), *ref*, *rev* and *nef*, and the transcript with exon structure 1-5-8c were cloned into the expression vector pIRES2-AcGFP1 (Clonetech) as described in Supplementary Methods.

Assays of protein activity and HIV replication

Activity and HIV replication assays were performed as described in Supplementary Methods. Tat activity expressed from each cDNA was measured in TZM-bl cells (36) (gift of Dr Robert W. Doms). Rev activity was assayed in HEK-293T cells co-transfected with pCMVGagPol-RRE-R, a reporter plasmid from which Gag and Pol are expressed in a Rev-dependent manner (gift of David Rekosh) (37). Intracellular and released supernatant p24 was measured from cells transfected with expression constructs and infected with HIV-$1_{89.6}$.

Western blotting

HEK-293T cells were transfected with expression constructs and treated with MG132 (EMD Chemicals) to inhibit the proteosome or DMSO (Supplementary Methods). Proteins were detected by immunoblotting using a mouse antibody that recognizes the carboxy terminus of HIV-1 Nef diluted 1:1000 in 5% milk (gift of Dr James Hoxie) (38). Horseradish peroxidase (HRP)-conjugated secondary rabbit-anti-mouse antibody (p0260, DAKO) was used for detection with SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific). Beta-tubulin was used as a loading control, detected by the HRP-conjugated antibody (ab21058, Abcam).

Single-molecule amplification

Amplification was performed by RainDance Technologies using a protocol similar to that previously reported (detailed description in Supplementary Methods) (39). Amplification was carried out in droplets to suppress competition between amplicons. PCR droplets were generated on the RDT 1000 (RainDance Technologies) using the manufacturer's recommended protocol. The custom primer libraries for this study contained 18 (HOS-CD4-CCR5 cells) or 20 (primary T cells) PCR primer pairs designed to amplify different HIV RNA isoforms (Supplementary Table S2).

Single-molecule sequencing

DNA amplification products from the RainDance PCR droplets were converted to SMRTbell™ templates using

the PacBio® *RS* DNA Template Preparation Kit. Sequencing was performed by Pacific Biosciences using the PacBio SMRT™ sequencing technology as described (33). Sequence information was acquired during real time as the immobilized DNA polymerase translocated along the template molecule. Prior to sequence acquisition, hairpin adapters were ligated to each DNA template end so that DNA polymerase could traverse DNA molecules multiple times during rolling circle replication [SMRTbell™ template sequencing (40)], allowing error control by calculating the consensus ('circular consensus sequence' or CCS). For raw reads, the average length was 2860 nt, and 10% were >5000 nt. After condensing into consensus reads, the mean read length was 249.5 nt, due to the use of a shorter Pacific Biosciences sequencing protocol to accommodate the small size of many amplicons. Consensus reads of 1% were >1100 nt. Sequencing data were collected in 45-min movies.

Data analysis

Raw reads were processed to produce CCSs. Raw reads were also retained to help in primer identification and to avoid biasing against long reads. Reads were aligned against the human genome using Blat (41). Misprimed reads matching the RT primer, reads with a CCS length shorter than 40 nt or raw length shorter than 100 nt and reads matching the human genome were discarded. Filtered reads were aligned against the HIV-$1_{89.6}$ reference genome. Potential novel donors and acceptors were found by filtering putative splice junctions in the Blat hits for a perfect sequence match 20 bases up- and downstream of the junction, ignoring homopolymer errors, and requiring that one end of the junction be a known splice site. Local maximums within a 5-nt span with >9 such junctions were called as novel splice sites.

Filter-passed reads were aligned against all expected fragments based on primers and known and novel junctions. Primers were identified in CCS reads by an edit distance ≤1 from the primer in the start or end of the read, in raw reads by an edit distance ≤5 from a concatenation of the primer, hairpin adapter and the reverse complement of the primer, and in both types of reads by a Blat hit spanning an entire expected fragment.

Gaps in Blat hits were ignored if ≤10 bases long or in regions of likely poor read quality ≤20 bases long where an inferred insertion of unmatched bases in the read occurred at the same location as skipped bases in the reference. Any Blat hits with a gap >10 nt remaining in the query read were discarded. If HIV sequence was repeated in a given read (likely due to PacBio® circular sequencing), the alignments were collapsed into the union of the coverage. Gaps in the HIV sequence found in uninterrupted query sequence were called as tentative introns. Splice junctions were assigned to conserved or previously identified (published or in this work) splice sites and reads appearing to contain donors or acceptors further than 5 nt away from these sites were discarded. Reads with Blat hits outside the expected primer range were discarded from that primer grouping. The assigned primer pair, observed junctions and exonic sequence were used to assign each read to a given spliceform (specific transcript structure) or set of possible spliceforms. Partial sequences that did not extend through both primers were assigned to specific transcripts if the read contained enough information to rule out all other spliceforms or if all other possible spliceforms contained rare (<1% usage) donors or acceptors (Supplementary Table S3). Otherwise, the read was called indeterminate.

To calculate the ratios of transcripts within the partially spliced class, we counted the number of reads for each assigned spliceform amplified by primer pair 1.3 and divided by the total number of assigned partially spliced reads amplified with these primers (Supplementary Figure S1 and Supplementary Table S2). Assigned sequences amplified with primer pairs 1.4 and 4.1 (full-length cDNAs, T cells only) were used to calculate ratios of transcripts within each of the two completely splice classes (~2 and ~1 kb). To compare ratios of ~2 kb transcripts calculated within reads from primer pairs 1.4 and 4.1, we normalized ratios from pair 4.1 to the *nef 2* transcript (containing exons 1, 5 and 7). Due to size biases inherent in the approach, we did not compare across size classes, and unspliced transcripts were not included in ratio analysis. For all ratio analysis, transcripts including cryptic or novel junctions were counted only if they appeared in at least five reads, otherwise they were excluded from the analysis and from the count of total assigned reads.

To estimate the minimum total number of transcripts present, partial sequence reads were included. Each exon–exon junction occurring in at least five reads and not previously assigned to a particular transcript (Figure 2) was counted as evidence of an additional transcript (47 additional junctions were detected, see Supplementary Table S4). If two such junctions could conceivably occur in a single mRNA, we counted only one unless we could verify from sequence reads that they were amplified from separate cDNAs, resulting in 31 additional transcripts. The minimum transcript number calculated by a greedy algorithm treating introns as events in a scheduling problem agreed with the above calculation.

For studies of transcript dynamics, reads from primer pairs 1.2, 1.3 and 1.4 containing junctions between D1 or any donor and each of five mutually exclusive acceptors, A3, A4c, A4a, A4b, A5 and A5a, were collected and their ratios calculated.

Statistical analysis

Statistical modeling was performed using generalized linear modeling as described in Supplementary Report S2. All analyses were performed in R 2.14.0 (R Development Core) (42).

Data access

Sequence data is available in the SRA database (http://www.ncbi.nlm.nih.gov/sra/, 6 August 2012, date last accessed) with the following accession numbers: SRP014319. Samples were submitted and numbers are pending.

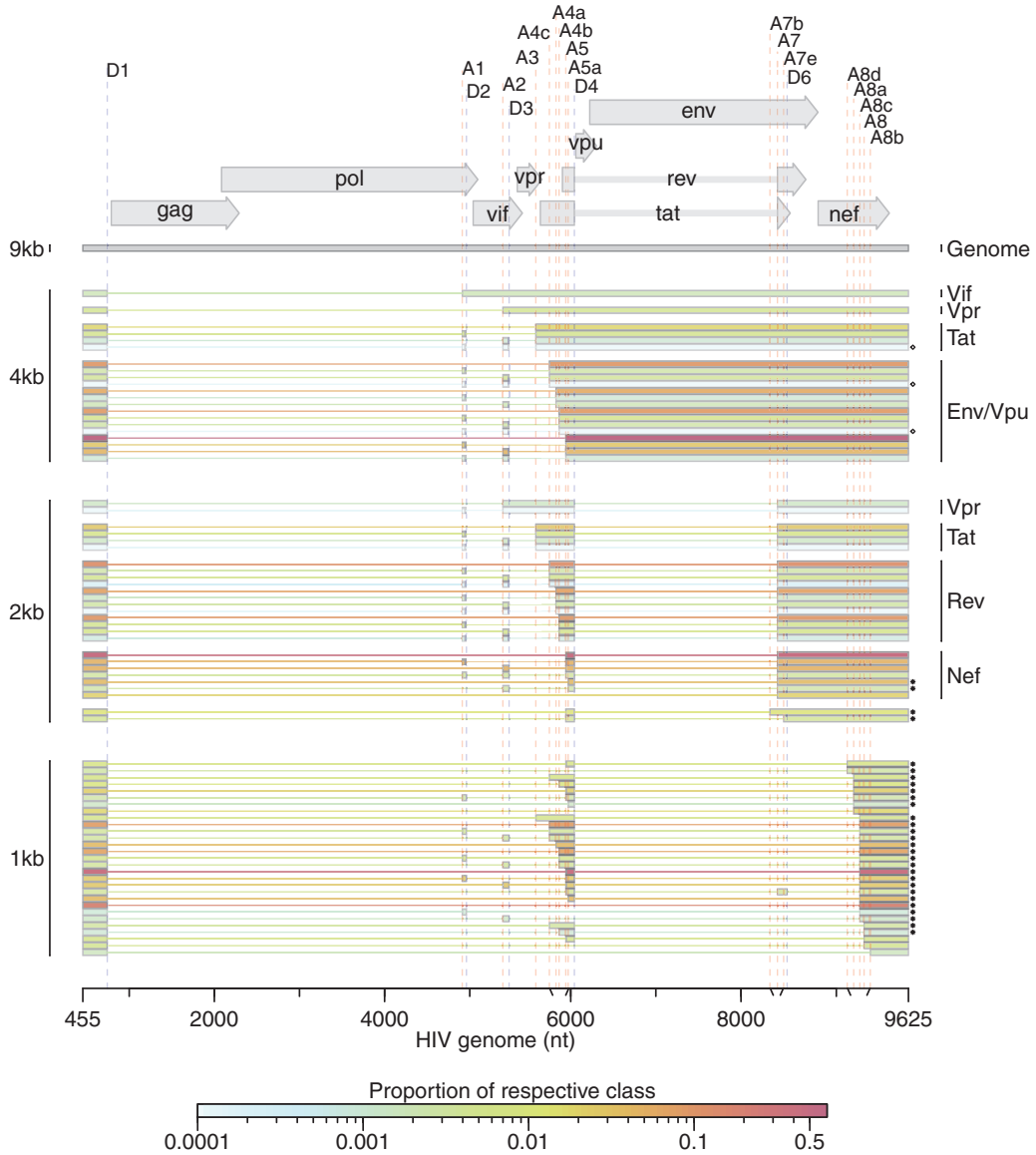

Figure 2. Spliced transcripts produced from HIV-$1_{89.6}$. HIV-$1_{89.6}$ transcripts in T cells for which the full message structure was determined are shown arranged by size class (unspliced genome, partially spliced or 4 kb, completely spliced or 2 kb, and a new completely spliced 1 kb class). Thick bars correspond to exons and thin lines to excised introns. For the well-conserved transcripts, encoded proteins are indicated. The relative abundance of each transcript within its size class is indicated by color according to the scale displayed. Asterisks denote transcripts that have not been reported previously to our knowledge. Of the 47 conserved HIV-1 transcripts, three were detected in fewer than five reads (one *tat* and two *env*/*vpu* messages, indicated, ◇), and two messages were not detected and are not shown (one encoding Vpr and one encoding Env/Vpu). Depicted non-conserved transcripts (using novel or cryptic splice sites) were each detected in at least five independent sequence reads across samples from at least two different human T-cell donors.

RESULTS

Sequencing HIV-1 transcripts produced in primary T cells and HOS cells

In order to characterize HIV-1 transcript populations, we prepared viral cDNA from primary $CD4^+$ T cells of seven different healthy human donors infected *in vitro* with HIV-$1_{89.6}$, an early passage dual-tropic clade-B clinical isolate (Supplementary Figure S1, human donor data in Supplementary Table S1) (32). We also studied HIV messages produced in infected HOS cells engineered to express CD4 and CCR5 (HOS-CD4-CCR5) because these cells support efficient HIV replication and engineered variants are widely used in HIV research. HOS cells were harvested at 18, 24 and 48 hours post infection (hpi) to investigate longitudinal changes during infection, and for comparison to 48 h infected T cells.

To preserve the relative proportions of template molecules while amplifying the cDNA, we used RainDance Technologies' single-molecule micro-droplet based PCR (39). Droplet libraries containing multiple overlapping primer pairs were designed to query all message forms and allow later calculation of relative abundance (Supplementary Table S2 and Supplementary Figure S1).

Each primer was unique so that sequences could be assigned to a specific primer pair, which helped reconstruct the origin of sequence reads and deduce message structures. Amplified DNA products were sequenced using Single Molecule Real-Time (SMRT®) technology from Pacific Biosciences (33,40). We obtained 847 492 filtered reads of amplified HIV-1 transcripts in primary $CD4^+$ T cells and 89 350 in HOS cells. The longest sequenced continuous stretch of HIV-1 cDNA was 2629 bp.

Splice donors and acceptors

We aligned PacBio® reads containing HIV sequences to the HIV-$1_{89.6}$ genome and identified candidate introns as recurring gaps in our sequences. Using this approach, we observed splicing at each of the widely conserved major splice donors and acceptors and several published cryptic sites (Figure 1A, hereafter referred to by their identifications shown in this figure, 'D' for donors, 'A' for acceptors).

In addition, we identified 13 putative novel splice sites: 2 donors and 11 acceptors (Figure 1 and Supplementary Table S3). In order to be selected as a bona fide splice site and remove artifacts possibly created by recombination during sample preparation, we required that the new acceptor or donor was observed spliced to previously reported splice donors or acceptors in >10 sequence reads in $CD4^+$ T cells. The most frequently used novel splice site was an acceptor that we have termed A8c because it lies near A8, A8a and A8b (discussed in detail below). Additional novel sites are further discussed in Supplementary Report S1.

Most of the new splice sites adhered to consensus sequences for the standard spliceosome (Supplementary Table S3). However, there appeared to be one splice donor upstream of D1 with a cytidine in place of the usual uracil 2 nt downstream of the splice site. Similar 'GC donors' appear in 1% of known splice junctions in humans (43). Of the novel splice acceptors, three were preceded by dinucleotides other than the consensus AG. Alternative dinucleotides are used infrequently as splice acceptors (44–47); however, it is possible that our deep sequencing method allowed us to observe rare events.

Structures of spliced HIV-$1_{89.6}$ RNAs

To quantify the populations of HIV-1 transcripts, we aligned all reads to the collection of 47 well-established spliced HIV-1 transcripts and detected 45 of them (Figure 2). We additionally aligned reads to the HIV-$1_{89.6}$ genome allowing all possible combinations of splice junctions—canonical, cryptic or novel—determined from the sequencing data (Figure 1), yielding an additional 32 complete transcripts, 19 of which were novel. The data also provide evidence for more novel splice junctions but in incomplete sequences, implying the existence of additional new transcripts (Supplementary Table S4 and Supplementary Report S1). The full data set taken together provides evidence for least 109 different HIV-$1_{89.6}$ transcripts in primary T cells.

Amplification primers that isolated the two main classes of spliced messages allowed us to determine the ratios of mRNAs in each (Figure 2 and Supplementary Table S5). Within the partially spliced class of transcripts, *env*/*vpu*, *tat* (1-exon), *vpr* and *vif* messages existed in an average ratio of 96:4:<1:<1 in $CD4^+$ T cells. The ratio of *nef*:*rev*:*tat*:*vpr* within the ~2 kb transcript class was 64:33:3:<1. Consistent with previous reports, the most abundant transcript in each class contained the splice junction from D1 to A5 (D1^A5)—an *env*/*vpu* transcript contributing 64% of the partially spliced class, and a completely spliced *nef* transcript contributing 47% of ~2 kb messages (Figure 2) (7,48). The relatively low abundance of transcripts encoding Tat suggests that Tat sufficiently stimulates HIV transcription elongation at low concentrations, or that the *tat* transcripts must be efficiently translated. Due to biases inherent in the reverse transcription step, we could only compare transcripts within each size class, and we note that our methods have not been validated for empirical quantification. However, the ratios were roughly confirmed using overlapping sequence reads obtained with alternate primer pairs and by end point RT–PCR analysis of HIV-1 RNAs (data not shown).

Exons 2 and 3 are non-coding exons whose inclusion in transcripts other than *vif* and *vpr* has no known function. We found that they were included in other messages infrequently, each in ~7–8% of transcripts in the ~2 kb completely spliced class of transcripts and 5% of partially spliced transcripts accumulating in T cells. This is consistent with previous measurements in the partially spliced class but much lower than has been estimated for completely spliced transcripts in HeLa cells, suggesting cell-type-specific splicing patterns may influence inclusion of these exons (7).

A novel ~1 kb class of completely spliced transcripts

Primers placed near the 5′- and 3′-ends of the HIV-$1_{89.6}$ genome amplified a second class of completely spliced transcripts ~1 kb in length. In place of A7, these transcripts use a set of little studied splice acceptors located ~800 bp downstream within the 3′-TR. Two groups have previously observed splicing from D1 to acceptors A8, A8a and A8b in this region, yielding messages of this size class in patient samples; however, none of these could be translated to a protein of significant length (9,13). We determined the complete structure of 29 members of the 1-kb class (Figure 2 and Supplementary Table S5). The most abundant messages observed in this class use the novel acceptor A8c to define their terminal exon. For $HIV_{89.6}$, acceptor A8c was used nearly as frequently as A7, which gives us the 2-kb class of transcripts (Supplementary Table S3), and this was supported by end point RT–PCR analysis (data not shown).

Acceptor A8c is not well conserved in HIV-1/SIVcpz (14%), although it is conserved in clade G viruses (>95%) and most HIV-2/SIVsmm genomes (86%) (49). This is due to the poor conservation of an adenosine at the wobble base position of the 123rd codon (proline) of the Nef reading frame, which creates the AG dinucleotide generally required at splice acceptors. Since any base at

this position would code for proline, there does not seem to be strong selection for a splice acceptor here. However, A8c is displaced from nearby well-conserved (>90%) cryptic acceptors A8a and A8b by multiples of 3 bp (12 and 21 bp, respectively), so splicing to any of these three acceptors would create similar ORFs. All HIVs and SIVs maintain at least one of these three acceptors, suggesting possible function (49). We confirmed that the 1 kb transcripts using A8a, A8b and A8c were present in infected HOS and T cells by end point RT–PCR using additional primer pairs and by Sanger sequencing of cloned transcripts (Figure 3A and B; data not shown).

The 1-kb transcript containing exons 1, 4 and 8c (1-4-8c, where exon 8c begins at A8c and extends to the poly-adenylation site) encodes the first exon of Tat followed by 25 novel amino acids (termed Tat^8c). Tat^ 8c showed activity when overexpressed in cells containing a Tat reporter construct (Figure 3C, nucleotide and amino acid sequences in Supplementary Table S6). Transcripts with exon structures 1-4a/b/c-8c encode a novel fusion of the amino-terminal 26 amino acids of Rev and the carboxy-terminal 80 amino acids of Nef, hereafter referred to as Ref. We did not detect Rev activity on overexpression of the *ref* transcript, and Ref did not appear to interfere with the normal function of Rev or with HIV replication (Supplementary Figure S2). Ref was detectable by western blot using antibodies targeting the C terminus of Nef after inhibition of the proteosome, suggesting that the fusion is expressed but not stable (Figure 3D). Thus, Ref has the potential to encode a new epitope potentially relevant in immune detection of HIV. The transcripts with exon structures 1-5-8c and 1-8c encode at most a short peptide, and so are candidates for acting as regulatory RNAs.

Temporal dynamics of transcript populations

To assess longitudinal variation, we investigated HIV-$1_{89.6}$ transcript populations during the course of a single round of infection in HOS-CD4-CCR5 cells. A sensitive method for comparison among conditions involves quantifying utilization of six mutually exclusive splice acceptors A3, A4c, A4a, A4b, A5 and a novel acceptor just downstream of A5 termed A5a. Splicing at these acceptors determines the relative levels of messages encoding Tat and Env/Vpu in the partially spliced class and messages encoding Tat, Rev and Nef in the completely spliced class.

We observed longitudinal changes in the levels of these messages in HOS cells over 12–48 h that were statistically significant ($P < 10^{-10}$; generalized linear model described in Supplementary Report S2). This pattern was especially evident in junctions involving donor 1 spliced to each of these acceptors (Figure 4A). Most dramatically, transcripts with splicing junctions between D1 and A3 (*tat* messages) increased with time ($P < 10^{-10}$), while D1^A4b junctions (used in *env*/*vpu* or *rev* messages) were used reciprocally less ($P < 10^{-10}$). Such kinetic changes affecting specific transcripts both with and without the Rev-response element cannot be explained by the accumulation of Rev, and they may reflect differential transcript stability or HIV-induced alterations to the host splicing machinery. Temporal changes in HOS cells were confirmed using end point RT–PCR and analysis after electrophoresis on ethidium-stained gels (Figure 4B).

Cell-type-specific splicing patterns

We also compared splicing between T cells and HOS cells and found significant cell type differences ($P < 10^{-10}$). For example, while transcripts with D1^A5 junctions were dominant in both cell types, messages using the D1^A4c splice junction (encoding Env/Vpu or Rev) made up the bulk of the remaining transcripts in T cells but were a minor species in HOS-CD4-CCR5 cells. Likewise, Tat messages (using A3), which were quite abundant in HOS cells at all time points, contributed relatively little to populations of transcripts in primary T cells harvested at 48 hpi (Figure 4A). We also used end point PCR and analysis on ethidium-bromide-stained gels to confirm that the relative ratios of transcripts containing junctions to A3, A4a, A4b and A4c were different in HOS and T cells (Figure 4B).

Human variation in HIV-1 splicing

Quantitative comparisons also revealed modest differences in splicing between primary $CD4^+$ T cells isolated from different human donors that were statistically significant ($P < 10^{-10}$) under a generalized linear model (Figure 4A). The magnitudes of predicted differences were small, all <33% and most <10%.

DISCUSSION

Use of single-molecule enrichment and long-read single-molecule sequencing has made possible the most complete study to date of the composition of HIV-1 message populations, revealing several new layers of regulation. Studies of the low-passage $HIV_{89.6}$ isolate in a relevant cell type showed numerous differences from studies of lab-adapted HIV strains in transformed cell lines, highlighting the importance of studying the most relevant models. These data also illustrate the limitations of gel-based assays for studying HIV-1 message population. Multiple different combinations of HIV-1 exons yield mRNAs of similar sizes that are easily confused in typical assays using gel electrophoresis. Thus, in many settings the more detailed information provided by single-molecule amplification and single-molecule DNA sequencing is more useful.

Using these methods, we have detected significant variations between HIV message populations generated in T cells from different human donors. The differences were modest compared to those observed between cell types or time points, perhaps not surprisingly since any human polymorphisms strongly affecting mRNA processing might interfere with normal gene expression. However, because tight calibration of message levels is important to HIV-1, the observed differences in message ratios might affect HIV-1 acquisition or disease progression. The variation in observed transcripts could also be affected by different kinetics of infection in T cells from the different donors. In either case, these data suggest that human polymorphisms may exist that affect HIV-1 message

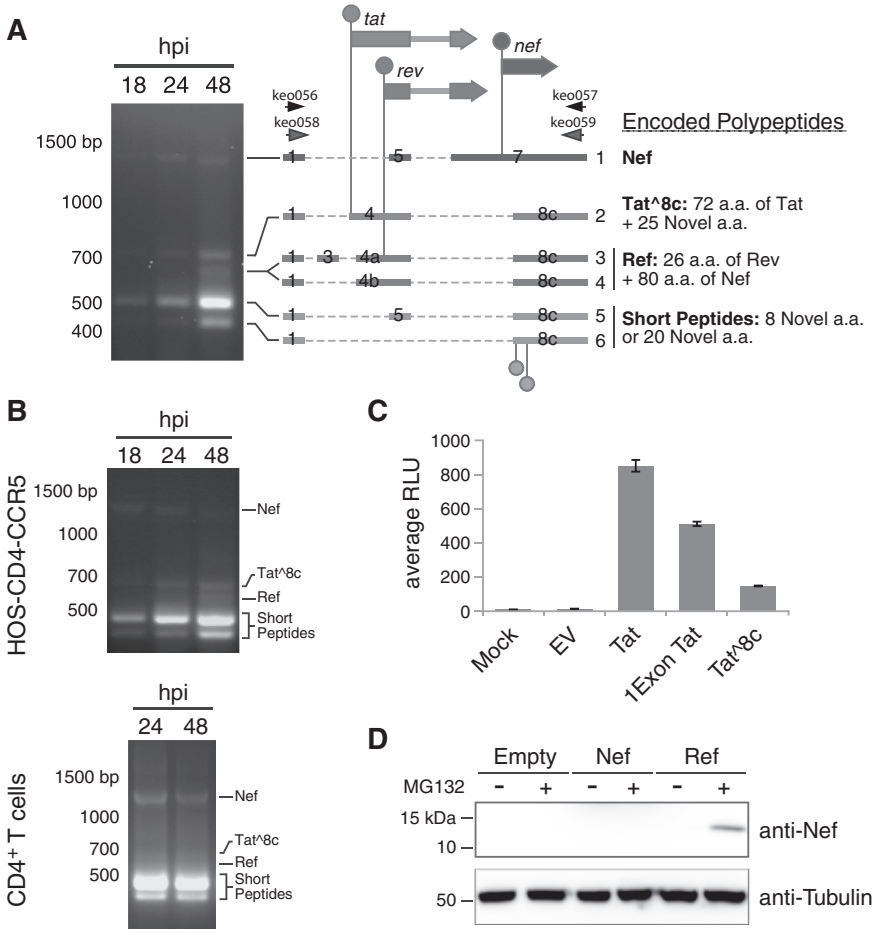

Figure 3. Novel transcripts utilizing acceptor A8c. (**A**) HIV-$1_{89.6}$ transcripts were amplified by RT–PCR using RNA from infected HOS-CD4-CCR5 cells with primers keo056 and keo057 (Supplementary Table S2). Major bands detected after gel electrophoresis were cloned from the 48 hpi sample and message structures determined by Sanger sequencing. Thick bars represent exons and dashed lines excised introns. Genes are shown above (not to scale) with start codons indicated by circles. Coding potentials of open reading frames are described. The first two start codons in messages 5 and 6, circles below, are not shared by known HIV-1 genes. Messages 1, 2, 4 and 5 were cloned into expression plasmids for activity assays. (**B**) Confirmation of presence of the ~1 kb message RNAs in HOS-CD4-CCR5 and primary $CD4^+$ T cells (human donor 1, harvested 24 and 48 hpi). An independent primer pair (keo058 and keo059) was used to amplify transcripts by RT–PCR. Expected amplicon sizes for transcripts in (A) are shown. (**C**) Tat activity was measured in Tzm-bl cells as Tat-dependent luciferase production after transient transfection with expression plasmids. (**D**) Western blot showing expression of protein of the predicted size for Ref (12.5 kb) in cells transfected with the Ref expression construct and treated with proteosome inhibitor MG132, detected by an antibody recognizing the carboxy-terminus of Nef. Expression plasmid encoding Nef was included to control for possible expression of partial Nef peptides or breakdown products from the Nef ORF.

populations in infected individuals, providing a new candidate mechanism connecting human genetic variation with measures of HIV disease.

Sequences from the 89.6 viral strain revealed a class of small (~1 kb) completely spliced transcripts, most contributed by splicing to a new poorly conserved acceptor A8c. These encoded two new proteins, one of which had Tat activity, and we showed that another, a Rev-Nef fusion termed Ref, could be detected in cells. HIV-$1_{89.6}$ is a particularly cytotoxic virus isolated from the CSF of a patient, and it forms unusually large syncitia in macrophages (32). The abundance of 1-kb transcripts produced by this virus provides a possible explanation for its unique properties. In addition to the novel acceptor A8c, we have also identified 3 putative novel splice donors and 11 putative novel acceptors, which require further studied to clarify possible functions.

The wealth of new messages found here in HIV-$1_{89.6}$ and in other HIV-1 isolates suggests there may be ongoing evolution of novel splice sites and new ORFs. Because splice acceptors in HIV-1 are weak (14), mutations creating sequences that even slightly resemble the 3′ splice site consensus may be occasionally recruited as novel acceptors, creating new mRNAs. In fact, new splice signals may evolve with relative ease—it has been estimated that reasonable matches to the consensus for splice donors, acceptors and branch-point sites occur within random sequence every 290, 490 and 24 bp, respectively (50), though sequence substitutions in HIV are usually also constrained by overlapping viral coding regions. We and others have observed appearance of novel exons within the major HIV-1 introns (8,10,11). Such long stretches of RNA relatively devoid of competing splice sites may be particularly poised to evolve new signals. On the other hand, most of the

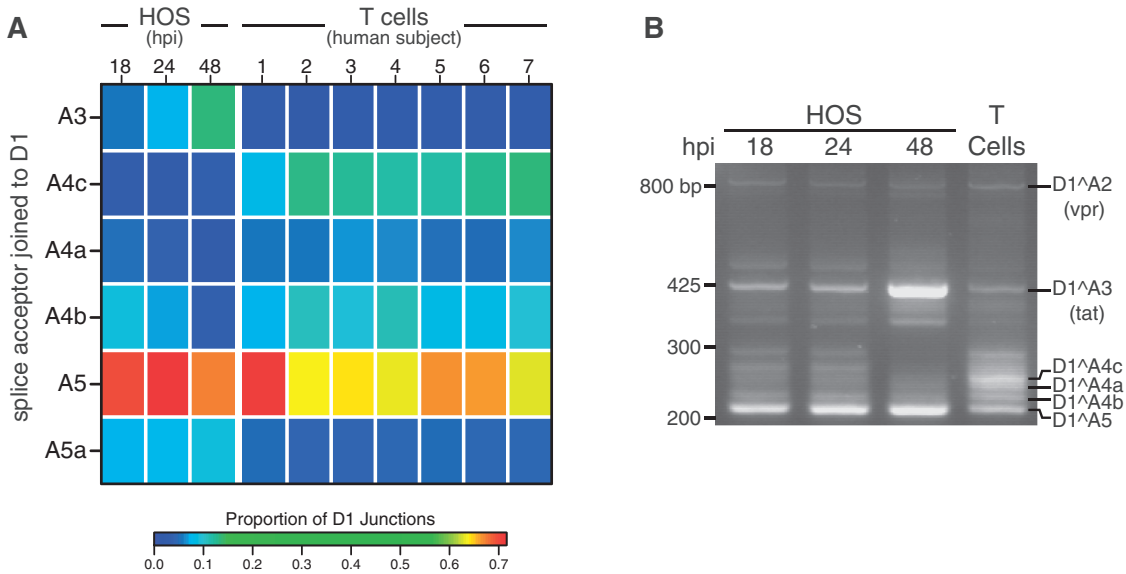

Figure 4. Temporal, cell type and donor variability in accumulation of HIV-1 messages. (**A**) In order to highlight changes in ratios of HIV-1 transcripts accumulating over time during infection and between HOS-CD4-CCR5 cells and primary T cells, we used PacBio® read counts to calculate proportions of transcripts with splicing from the first major splice donor, D1, to each of the mutually exclusive acceptors: A3 (required to make Tat), A4c, A4a, A4c (Env/Vpu and Rev), A5 (Env/Vpu and Nef) and the novel putative acceptor A5a. Sequences used in the analysis derived from templates amplified with primers F1.2 and R1.2 (Supplementary Table S2). The heat map shows average data for T cell and HOS cell samples in columns with the color tiles indicating the proportion of D1 splicing to each of the mutually exclusive acceptors (rows), according to the color scale shown. Statistics for this analysis based on a generalized linear model are provided in Supplementary Report S2. (**B**) Reverse transcription and bulk PCR amplification of HIV-$1_{89.6}$ transcripts from HOS cells and primary T cells from one human subject (subject 3) resolved by agarose gel electrophoresis and stained with ethidium bromide verified temporal and cell type changes shown in (A).

putative novel splice acceptors we observed clustered near previously identified acceptors in HIV-1, suggesting that conserved *cis*-acting splicing signals may recruit factors that act promiscuously on new nearby sequences. Clusters of splice sites might also provide redundancies that protect vital messages, as suggested previously (51,52). Frequent evolution of new splice sites may allow viruses to test out new combinations of exons, potentially yielding new RNAs and proteins, like those reported here. However, such novelty must compete with immune constraints—unstable novel polypeptides like Ref can be targeted to the proteosome and presented on MHC molecules as new epitopes for immune recognition.

HIV has likely evolved to produce calibrated message populations in T cells which seem to be altered with relative ease, as in infection in HOS cells, suggesting that therapeutic disruption of correct splicing may be feasible. A few studies have begun to explore small molecule therapy to disrupt HIV-1 splicing (15,19). Several factors could be responsible for the differences we observed between HOS and T cells, including hnRNP A/B and H, SC35, SF2/ASF and SRp40 (53,54). Inhibition of SF2/ASF has already been shown to abrogate HIV-1 replication *in vitro* (15). Thus the lability seen here for function of these factors suggests they may be attractive antiretroviral targets.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–6, Supplementary Figures 1 and 2, Supplementary Methods, Supplementary Reports 1 and 2 and Supplementary References [55–58].

ACKNOWLEDGEMENTS

We would like to thank the University of Pennsylvania Center for AIDS Research (CFAR) for preparation of viral stocks and isolation of primary $CD4^+$ T cells; James A. Hoxie, Ronald G. Collman, Jianxin You, Robert W. Doms, Paul Bates, David Rekosh and members of the Bushman laboratory for reagents, helpful discussion and technical expertise. F.D.B., K.T., D.L., E.S., K.E.O. and R.M. conceived and designed the experiment. K.E.O. and R.C.A. carried out sample preparation and experimental validation. P.D. and J.O. performed single-molecule amplification. K.T. and S.W. performed sequencing. S.S.-M., K.E.O. and M.B. analyzed the data. K.E.O., F.D.B. and S.S.-M. wrote the manuscript.

FUNDING

Funding for open access charge: HIV Immune Networks Team (HINT) Consortium [P01 AI090935]; National Institutes of Health (NIH) institutional training grant [T32 AI 7324-19 to K.E.O.].

Conflict of interest statement. None declared.

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. 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.
3. Pagani,F., Raponi,M. and Baralle,F.E. (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl Acad Sci. USA*, **102**, 6368–6372.
4. Wang,G.S. and Cooper,T.A. (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.*, **8**, 749–761.
5. Wang,L., Lawrence,M.S., Wan,Y., Stojanov,P., Sougnez,C., Stevenson,K., Werner,L., Sivachenko,A., DeLuca,D.S., Zhang,L. *et al.* (2011) SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N. Engl. J. Med.*, **365**, 2497–2506.
6. Barash,Y., Calarco,J.A., Gao,W., Pan,Q., Wang,X., Shai,O., Blencowe,B.J. and Frey,B.J. (2010) Deciphering the splicing code. *Nature*, **465**, 53–59.
7. Purcell,D.F. and Martin,M.A. (1993) Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *J. Virol.*, **67**, 6365–6378.
8. Benko,D.M., Schwartz,S., Pavlakis,G.N. and Felber,B.K. (1990) A novel human immunodeficiency virus type 1 protein, tev, shares sequences with tat, env, and rev proteins. *J. Virol.*, **64**, 2505–2518.
9. Carrera,C., Pinilla,M., Perez-Alvarez,L. and Thomson,M.M. (2010) Identification of unusual and novel HIV type 1 spliced transcripts generated in vivo. *AIDS Res. Hum. Retroviruses*, **26**, 815–820.
10. Lutzelberger,M., Reinert,L.S., Das,A.T., Berkhout,B. and Kjems,J. (2006) A novel splice donor site in the gag-pol gene is required for HIV-1 RNA stability. *J. Biol. Chem.*, **281**, 18644–18651.
11. Salfeld,J., Gottlinger,H.G., Sia,R.A., Park,R.E., Sodroski,J.G. and Haseltine,W.A. (1990) A tripartite HIV-1 tat-env-rev fusion protein. *EMBO J.*, **9**, 965–970.
12. Schwartz,S., Felber,B.K., Benko,D.M., Fenyo,E.M. and Pavlakis,G.N. (1990) Cloning and functional analysis of multiply spliced mRNA species of human immunodeficiency virus type 1. *J. Virol.*, **64**, 2519–2529.
13. Smith,J., Azad,A. and Deacon,N. (1992) Identification of two novel human immunodeficiency virus type 1 splice acceptor sites in infected T cell lines. *J. Gen. Virol.*, **73(Pt 7)**, 1825–1828.
14. Stoltzfus,C.M. (2009) Chapter 1. Regulation of HIV-1 alternative RNA splicing and its role in virus replication. *Adv. Virus Res.*, **74**, 1–40.
15. Bakkour,N., Lin,Y.L., Maire,S., Ayadi,L., Mahuteau-Betzer,F., Nguyen,C.H., Mettling,C., Portales,P., Grierson,D., Chabot,B. *et al.* (2007) Small-molecule inhibition of HIV pre-mRNA splicing as a novel antiretroviral therapy to overcome drug resistance. *PLoS Pathog.*, **3**, 1530–1539.
16. Brass,A.L., Dykxhoorn,D.M., Benita,Y., Yan,N., Engelman,A., Xavier,R.J., Lieberman,J. and Elledge,S.J. (2008) Identification of host proteins required for HIV infection through a functional genomic screen. *Science*, **319**, 921–926.
17. Jablonski,J.A. and Caputi,M. (2009) Role of cellular RNA processing factors in human immunodeficiency virus type 1 mRNA metabolism, replication, and infectivity. *J. Virol.*, **83**, 981–992.
18. Konig,R., Zhou,Y., Elleder,D., Diamond,T.L., Bonamy,G.M., Irelan,J.T., Chiang,C.Y., Tu,B.P., De Jesus,P.D., Lilley,C.E. *et al.* (2008) Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, **135**, 49–60.
19. Tranell,A., Tingsborg,S., Fenyo,E.M. and Schwartz,S. (2011) Inhibition of splicing by serine-arginine rich protein 55 (SRp55) causes the appearance of partially spliced HIV-1 mRNAs in the cytoplasm. *Virus Res.*, **157**, 82–91.
20. Zhou,H., Xu,M., Huang,Q., Gates,A.T., Zhang,X.D., Castle,J.C., Stec,E., Ferrer,M., Strulovici,B., Hazuda,D.J. *et al.* (2008) Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe*, **4**, 495–504.
21. Zhu,Y., Chen,G., Lv,F., Wang,X., Ji,X., Xu,Y., Sun,J., Wu,L., Zheng,Y.T. and Gao,G. (2011) Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting multiply spliced viral mRNAs for degradation. *Proc. Natl Acad. Sci. USA*, **108**, 15834–15839.
22. Saltarelli,M.J., Hadziyannis,E., Hart,C.E., Harrison,J.V., Felber,B.K., Spira,T.J. and Pavlakis,G.N. (1996) Analysis of human immunodeficiency virus type 1 mRNA splicing patterns during disease progression in peripheral blood mononuclear cells from infected individuals. *AIDS Res. Hum. Retroviruses*, **12**, 1443–1456.
23. Delgado,E., Carrera,C., Nebreda,P., Fernandez-Garcia,A., Pinilla,M., Garcia,V., Perez-Alvarez,L. and Thomson,M.M. (2012) Identification of new splice sites used for generation of rev transcripts in human immunodeficiency virus type 1 subtype C primary isolates. *PloS One*, **7**, e30574.
24. Grabowski,P. (2011) Alternative splicing takes shape during neuronal development. *Curr. Opin. Genet. Dev.*, **21**, 388–394.
25. Llorian,M. and Smith,C.W. (2011) Decoding muscle alternative splicing. *Curr. Opin. Genet. Dev.*, **21**, 380–387.
26. Ip,J.Y., Tong,A., Pan,Q., Topp,J.D., Blencowe,B.J. and Lynch,K.W. (2007) Global analysis of alternative splicing during T-cell activation. *RNA*, **13**, 563–572.
27. Topp,J.D., Jackson,J., Melton,A.A. and Lynch,K.W. (2008) A cell-based screen for splicing regulators identifies hnRNP LL as a distinct signal-induced repressor of CD45 variable exon 4. *RNA*, **14**, 2038–2049.
28. Sonza,S., Mutimer,H.P., O'Brien,K., Ellery,P., Howard,J.L., Axelrod,J.H., Deacon,N.J., Crowe,S.M. and Purcell,D.F. (2002) Selectively reduced tat mRNA heralds the decline in productive human immunodeficiency virus type 1 infection in monocyte-derived macrophages. *J. Virol.*, **76**, 12611–12621.
29. Dowling,D., Nasr-Esfahani,S., Tan,C.H., O'Brien,K., Howard,J.L., Jans,D.A., Purcell,D.F., Stoltzfus,C.M. and Sonza,S. (2008) HIV-1 infection induces changes in expression of cellular splicing factors that regulate alternative viral splicing and virus production in macrophages. *Retrovirology*, **5**, 18.
30. Hull,J., Campino,S., Rowlands,K., Chan,M.S., Copley,R.R., Taylor,M.S., Rockett,K., Elvidge,G., Keating,B., Knight,J. *et al.* (2007) Identification of common genetic variation that modulates alternative splicing. *PLoS Genet.*, **3**, e99.
31. Kwan,T., Benovoy,D., Dias,C., Gurd,S., Serre,D., Zuzan,H., Clark,T.A., Schweitzer,A., Staples,M.K., Wang,H. *et al.* (2007) Heritability of alternative splicing in the human genome. *Genome Res.*, **17**, 1210–1218.
32. Collman,R., Balliet,J.W., Gregory,S.A., Friedman,H., Kolson,D.L., Nathanson,N. and Srinivasan,A. (1992) An infectious molecular clone of an unusual macrophage-tropic and highly cytopathic strain of human immunodeficiency virus type 1. *J. Virol.*, **66**, 7517–7521.
33. Eid,J., Fehr,A., Gray,J., Luong,K., Lyle,J., Otto,G., Peluso,P., Rank,D., Baybayan,P., Bettman,B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.
34. Deng,H., Liu,R., Ellmeier,W., Choe,S., Unutmaz,D., Burkhart,M., Di Marzio,P., Marmon,S., Sutton,R.E., Hill,C.M. *et al.* (1996) Identification of a major co-receptor for primary isolates of HIV-1. *Nature*, **381**, 661–666.
35. Landau,N.R. and Littman,D.R. (1992) Packaging system for rapid production of murine leukemia virus vectors with variable tropism. *J. Virol.*, **66**, 5110–5113.
36. Wei,X., Decker,J.M., Wang,S., Hui,H., Kappes,J.C., Wu,X., Salazar-Gonzalez,J.F., Salazar,M.G., Kilby,J.M., Saag,M.S. *et al.* (2003) Antibody neutralization and escape by HIV-1. *Nature*, **422**, 307–312.
37. Srinivasakumar,N., Chazal,N., Helga-Maria,C., Prasad,S., Hammarskjold,M.L. and Rekosh,D. (1997) The effect of viral regulatory protein expression on gene delivery by human immunodeficiency virus type 1 vectors produced in stable packaging cell lines. *J. Virol.*, **71**, 5841–5848.
38. Shugars,D.C., Smith,M.S., Glueck,D.H., Nantermet,P.V., Seillier-Moiseiwitsch,F. and Swanstrom,R. (1993) Analysis of human immunodeficiency virus type 1 nef gene sequences present in vivo. *J. Virol.*, **67**, 4639–4650.

39. Tewhey,R., Warner,J.B., Nakano,M., Libby,B., Medkova,M., David,P.H., Kotsopoulos,S.K., Samuels,M.L., Hutchison,J.B., Larson,J.W. *et al*. (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat. Biotechnol*., **27**, 1025–1031.
40. Travers,K.J., Chin,C.S., Rank,D.R., Eid,J.S. and Turner,S.W. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res*., **38**, e159.
41. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
42. R Development Core Team. (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/ (6 August 2012, date last accessed).
43. Thanaraj,T.A. and Clark,F. (2001) Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res*., **29**, 2581–2593.
44. Aebi,M., Hornig,H., Padgett,R.A., Reiser,J. and Weissmann,C. (1986) Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell*, **47**, 555–565.
45. Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res*., **28**, 4364–4375.
46. Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2001) SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res*., **29**, 255–259.
47. Sheth,N., Roca,X., Hastings,M.L., Roeder,T., Krainer,A.R. and Sachidanandam,R. (2006) Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res*., **34**, 3955–3967.
48. Guatelli,J.C., Gingeras,T.R. and Richman,D.D. (1990) Alternative splice acceptor utilization during human immunodeficiency virus type 1 infection of cultured cells. *J. Virol*., **64**, 4093–4098.
49. Kuiken,C.F.B., Leitner,T., Apetrei,C., Hahn,B., Mizrachi,I., Mullins,J., Rambaut,A., Wolinsky,S. and Korber,B. (eds), (2010) *HIV Sequence Compendium 2010*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM.
50. Burge,C.B., Tuschl,T.H. and Sharp,P.A. (1999) Splicing of precursors to mRNAs by the spliceosomes. In: Gesteland,R.F., Cech,T. and Atkins,J.F. (eds), *The RNA World*, 2nd edn. Cold Spring Harbor Laboratory Press, Plainview, NY, pp. 525–560.
51. Abbink,T.E. and Berkhout,B. (2008) RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor. *J. Virol*., **82**, 3090–3098.
52. Verhoef,K., Bilodeau,P.S., van Wamel,J.L., Kjems,J., Stoltzfus,C.M. and Berkhout,B. (2001) Repair of a Rev-minus human immunodeficiency virus type 1 mutant by activation of a cryptic splice site. *J. Virol*., **75**, 3495–3500.
53. Caputi,M., Freund,M., Kammler,S., Asang,C. and Schaal,H. (2004) A bidirectional SF2/ASF- and SRp40-dependent splicing enhancer regulates human immunodeficiency virus type 1 rev, env, vpu, and nef gene expression. *J. Virol*., **78**, 6517–6526.
54. Zahler,A.M., Damgaard,C.K., Kjems,J. and Caputi,M. (2004) SC35 and heterogeneous nuclear ribonucleoprotein A/B proteins bind to a juxtaposed exonic splicing enhancer/exonic splicing silencer element to regulate HIV-1 tat exon 2 splicing. *J. Biol. Chem*., **279**, 10077–10084.
55. Almarza,D., Bussadori,G., Navarro,M., Mavilio,F., Larcher,F. and Murillas,R. (2011) Risk assessment in skin gene therapy: viral-cellular fusion transcripts generated by proviral transcriptional read-through in keratinocytes transduced with self-inactivating lentiviral vectors. *Gene Ther*., **18**, 674–681.
56. Bohne,J., Wodrich,H. and Krausslich,H.G. (2005) Splicing of human immunodeficiency virus RNA is position-dependent suggesting sequential removal of introns from the 5′ end. *Nucleic Acids Res*., **33**, 825–837.
57. Exline,C.M., Feng,Z. and Stoltzfus,C.M. (2008) Negative and positive mRNA splicing elements act competitively to regulate human immunodeficiency virus type 1 vif gene expression. *J. Virol*., **82**, 3921–3931.
58. Asang,C., Hauber,I. and Schaal,H. (2008) Insights into the selective activation of alternatively used splice acceptors by the human immunodeficiency virus type-1 bidirectional splicing enhancer. *Nucleic Acids Res*., **36**, 1450–1463.