Title: Analysis of HIV-1 expression level and sense of transcription by high-throughput sequencing of the infected cell.

Authors: Lefebvre G, Desfarges S, Uyttebroeck F, Muñoz M, Beerenwinkel N, Rougemont J, Telenti A, Ciuffi A

Journal: Journal of virology

Year: 2011 Jul

Volume: 85

Issue: 13

Pages: 6205-11

DOI: 10.1128/JVI.00252-11

Analysis of HIV-1 expression level and sense of transcription by high-throughput sequencing of the infected cell

Running title: HIV-1 expression by SAGE-Seq

Gregory Lefebvre[1], Sébastien Desfarges[2], Frédéric Uyttebroeck[2], Miguel Muñoz[2], Niko Beerenwinkel[3,4], Jacques Rougemont[1], Amalio Telenti[2*§] and Angela Ciuffi[2*§].

Institute of Microbiology, University Hospital Center and University of Lausanne, Lausanne, Switzerland.

[1]Bioinformatics and Biostatistics Core Facility, School of Life Sciences, EPFL, Lausanne, Switzerland.

[2]Institute of Microbiology, University Hospital Center and University of Lausanne, Lausanne, Switzerland.

[3]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland.

[4]SIB – Swiss Institute of Bioinformatics, Basel, Switzerland

[§] Last co-authors

[*]Correspondent footnote:

Angela Ciuffi

Institute of Microbiology

University Hospital Center and University of Lausanne

Bugnon 48

CH-1011 Lausanne

Switzerland

Phone: +4121314.40.99

Fax:+4121314.40.60

e-mail: Angela.Ciuffi@chuv.ch

Amalio Telenti

Institute of Microbiology

University Hospital Center and University of Lausanne

Bugnon 48

CH-1011 Lausanne

Switzerland

Phone: +4121314.40.97

Fax:+4121314.40.60

e-mail: Amalio.Telenti@chuv.ch

Abstract

Next-generation sequencing offers an unprecedented opportunity to jointly analyze cellular and viral transcriptional activity, without prerequisite knowledge on the nature of the transcripts. SupT1 cells were infected with a VSV-G pseudotyped HIV vector. At 24 hours post-infection, both cellular and viral transcriptomes were analyzed by Serial Analysis of Gene Expression followed by high-throughput sequencing (SAGE-Seq). Read mapping resulted in 33 to 44 million tags aligning to the human transcriptome and 0.23 to 0.25 million tags aligning to the genome of the HIV-1 vector. Thus, at peak infection, one transcript in 143 is of viral origin (0.7%), including a small component of antisense viral transcription. Out of the detected cellular transcripts, 826 (2.3%) were differentially expressed between mock and HIV-infected samples. The approach also assessed whether HIV-1 infection modulates expression of repetitive elements or endogenous retroviruses. We observed very active transcription of these elements, with one transcript in 237 being of such origin, corresponding in average to 123,123 reads in mock (0.40%) and 129,149 reads in HIV-1 (0.45%) mapping to the genomic Repbase repository. This analysis highlights key details in the generation and interpretation of high-throughput data in the setting of HIV-1 cellular infection.

Introduction

There is significant interest in determining the level of HIV-1 transcription in the context of the infected cell. The impact of HIV-1 infection on cellular gene expression has been investigated in the past by gene expression arrays (3, 7, 9, 10, 15-18, 20, 21, 29, 31, 36, 37, 42, 47, 48, 51). However, use of these arrays is limited by the set of probes that are included in the chip. Recently, it has even been proposed that unbiased analysis of transcription activity by deep sequencing will soon replace gene expression profiling by microarrays (6, 45, 49). The first application of the novel technologies in the field of HIV-1 has been in the assessment of viral sequence variation, in particular mutations present at low frequency in complex (quasispecies) populations (2, 8, 12, 19, 22, 23, 43, 46, 50, 52, 55). Pyrosequencing approaches have also been instrumental in the assessment of small non-coding RNAs in HIV-1 infected cells (54). The next goal is the joint analysis of the viral and host transcriptome of the infected cell.

Two general methods of deep sequencing are defined on the basis of the mode of sample preparation (28). One approach uses fractionation of polyadenylated RNAs (valid also for non poly(A)+ RNAs) in short fragments, followed by reverse transcription using hexamers. Adapters are ligated to both ends, and used for sequencing. This method allows the detailed characterization of the transcript structure. A second approach uses Serial Analysis of Gene Expression (SAGE) to allow precise quantization of poly(A)+ RNA and to facilitate information on strandedness, a key for the understanding of antisense transcription of the HIV-1 genome. An important consideration regarding these technologies is the growing consensus that they will represent the new gold-standard in the analysis of transcription – deep sequencing-based

expression analysis has shown to represent a major advance in robustness, resolution and inter-lab portability over multiple microarray platforms (6, 45, 49).

In the present study, we used Super-SAGE, followed by high-throughput sequencing (SOLiD, Life Technologies), also known as Digital Gene Expression (DGE) tag profiling, 3' tag DGE, tag sequencing (Tag-Seq), or SAGE-Seq (32, 33, 38, 53), to analyze the transcriptome of a T cell line 24 hours post-infection.

Materials and Methods

Cells. HEK 293T cells were cultured in Dulbecco's Modified Eagle Medium (DMEM, Invitrogen), supplemented with 10% heat-inactivated fetal calf serum (FCS) and 50 μg/ml gentamycin (D-10 culture medium). SupT1 cells (a T-cell line) were cultured in RPMI-1640 (Invitrogen), supplemented with 10% heat-inactivated FCS and 50 μg/ml gentamycin (R-10 culture medium).

HIV-vector production. To produce HIV-based vector particles, 293T cells were co-transfected with two plasmids (20 μg total) using the calcium phosphate method (Invitrogen) and according to the manufacturer's instructions. One plasmid, pNL4-3Δenv-eGFP, codes for all viral proteins except the envelope, which was disrupted and replaced by GFP (56). The second plasmid, pMD.G, encodes the vesicular stomatitis virus G envelope protein (VSV-G) (39). Forty-eight hours after transfection, culture supernatant containing viral particles was collected, centrifuged to pellet cell debris, filtered through 0.22 μm filters, concentrated using Centricon-Plus 70-100K (Millipore), treated with 100 U/ml DNaseI (Roche) and stored frozen at -80°C. Virion

concentration was assessed by measuring the CA (p24) antigen by ELISA (Murex HIV Ag MAB; Abbott), according to instructions.

HIV infection. SupT1 cells ($5x10^6$) were infected with 15 μg p24 equivalent of HIV NL4-3Δenv-eGFP/VSV-G in presence of 5 μg/ml polybrene (Sigma) by centrifugation for 30 min at 1500 g. Mock cells were treated similarly but without HIV-based particles. Cells were washed with culture medium, resuspended at 10^6 cells/ml in R-10 and further incubated.

Transcriptional profiling by SAGE-Seq and bioinformatic analysis. At 24h post-exposure, cells were washed once with phosphate-buffered saline (PBS) and resuspended in 0.5 ml RNALater (Ambion). Total RNA was purified using miRvana isolation kit (Ambion), or TRIzol (Invitrogen) for the second experiment, according to recommendations. RNA quality was assessed by electrophoresis on the Bioanalyzer 2100 (Total RNA Nano; Agilent). SAGE libraries were prepared with SOLiD SAGE kit (Applied Biosystems) at the Functional Genomics Center Zurich. High-throughput sequencing was performed using SOLiD3 technology (Applied Biosystems). Sequence mapping was done with the bowtie software (27), using defined analysis parameters (read length (rl), number of mismatches (n) and alignments authorized (m)). Tag hits (*i.e.* successfully aligned reads) were normalized according to the number of locations they mapped to. Alignment used the human genome GRCh37, the human transcriptome RefSeq database, and HIV-1 genome *HIV NL4-3Δenv-eGFP*. Differential expression analysis was performed using R programming language (14) and based on negative-binomial modeling of count data as described in the Bioconductor package DESeq (1). Genes with significantly

modulated expression in HIV-1 samples compared to mock samples were selected with a false discovery rate (FDR) of 0.05 and annotation was retrieved using the biomaRt package (11).

Northern hybridization. Total RNA (10 μg) from mock-treated or HIV vector-infected cells was supplemented with RNA loading buffer (Bioline), heated 5 min at 65°C and separated by electrophoresis on 1.2% agarose-MOPS-formaldehyde gel (44). The gel was washed 3x10 min with 10xSSC. RNA transfer from gel to nitrocellulose membrane (Hybond N+; GE Healthcare) was performed over night by capillarity using 20x SSC. The membrane was rinsed with 6xSSC, crosslinked with UV (GeneLinker; BioRad), rinsed briefly with water and added to 10 ml preheated QuickHyb Hybridization solution (Agilent) for 30 min at 65°C. 200 pmol oligonucleotide probes were labeled with 20 μCi [γ–^{32}P]dATP with 10 U T4 polynucleotidyl kinase (T4 PNK; New England Biolabs) for 1h at 37°C. After heat inactivation of the enzyme for 10 min at 95°C, probes were purified on Qiaquick Nucleotide removal kit (Qiagen), mixed with 100 μl sonicated salmon sperm DNA [10 mg/ml] (Eppendorf), and added to the membrane in QuickHyb Hybridization solution for 1h at 65°C. Following hybridization, membrane was washed 2x15 min with 2xSSC/0.1% SDS at room temperature, 1x30 min in 2xSSC/0.1% SDS at 52°C, exposed to an intensifying screen over night at -80°C and revealed using ImageQuantTL v2005 software on Typhoon scanner (GE Healthcare). For re-use, membrane was stripped with boiling 0.1xSSC/0.1%SDS twice for 15 min. Oligonucleotide probes were 5'-TCTCTCTCAGGGTCATCCATTCCA-3' for peak 1 sense transcripts, 5'-GCTCGTCCTTGTACAGCTCGTCCA-3' for peak 3 sense transcripts, 5'-ATGGTGTTTTACTAATCTTTTCCATGTGTT-3' for peak 4 sense transcripts, 5'-CTTGTTACACCCTGTGAGCCTGCA-3' for peak 1 antisense transcripts, 5'-

CCGCCGCCGGGATCACTCTCGGCA-3' for peak 3 antisense transcripts and 5'-GTAGACAGGATGAGGATTAACACATGGAAA-3' for peak 4 antisense transcripts. As positive control for hybridization, 2 μg pNL4-3Δenv-eGFP digested with BglII and AflII was used.

Results and Discussion

SupT1 cells ($5x10^6$ cells) were mock-treated or infected with a VSV-G pseudotyped HIV-based vector expressing GFP (56) in duplicate (**Figure 1A**). At 36h post-infection, FACS analysis revealed that 93% of cells were successfully transduced, expressing GFP; other cell types were tested and found less efficient in transduction (**Supplemental File 1A**). The use of 24h time point in the experimental model takes into consideration cell loss and viability, initiation of translation, as measured by GFP expression (as soon as 16 hours), and the completion of the viral cycle, as measured by the production and release of p24 (as soon as 20 hours), reflecting viral particle release in SupT1 cells (**Supplemental File 1B**). Cells were collected for total RNA extraction and poly(A)+ RNAs were reverse transcribed and processed for SAGE library preparation. cDNAs were digested with NlaIII, a frequent 4-bp cutter, and ligated to an adaptor sequence, containing the recognition site of EcoP15I; an enzyme that cuts DNA asymmetrically, 25/27bp away. A second adapter was ligated leading to a 27bp transcript-specific tag, surrounded by two distinct adapter sequences. High-throughput sequencing was performed using the SOLiD 3 system with a universal primer annealing to the adaptor sequence. Reads were mapped either to the human genome (GRCh37), the human RefSeq database (40), the human repetitive-element database (24), or the HIV-1 vector genome (*HIV NL4-3Δenv-eGFP/VSV-G*) (**Figure 1B**).

Assessment of sequence alignment parameters. Variation of mapping parameters for HIV-1 tags was tested to maximize the number of tags mapping to the viral genome while minimizing the number of tags spuriously mapping to the host genome (**Supplemental File 2**). After shortening the reads by right-trimming, we used the read lengths of 21bp, 24bp, and 27bp. In addition, we tested different number of mismatches (n=0, 1, or 2) and multiple hits (m=1, 2, or 10). We retained condition rl=24, n=1 and m=1 as optimal for read mapping to the human genome, rl=24, n=1 and m=10 for mapping to the human transcriptome, and rl=24, n=2 and m=2 for mapping to the HIV-1 genome.

HIV-1 RNA tags identify sense and antisense transcription. Read mapping resulted in 33 to 44 million tags aligning to the human transcriptome and 0.23 to 0.25 million tags aligning to the genome of the HIV-1 vector (**Table 1**). These data suggest that, at peak infection, one transcript in 143 is of viral origin (0.7%). Overall, 0.33% were HIV-1-specific (1 in 309) and 0.37% were vector GFP-specific (1 in 267).

The distribution of tags on the HIV-1 vector genome emphasized several relevant aspects of HIV-1 transcription (**Figure 2A**). Five viral genomic regions carried 87.8% of total tags detected. A high proportion of tags (24.2%) mapped to the 3'end of the HIV-1 vector genome (**Figure 2A, Peak 1**), corresponding to sense transcription with a known functional polyadenylation site of HIV-1 (genome positions 9455-9460), and the first upstream NlaIII restriction site (pos. 9150). The second signal (**Figure 2A, Peak 2**), carrying 5.1% of total tags, corresponds to sense transcription and the same known functional polyadenylation site of HIV-1 or an alternative putative signal (at pos. 9108-9113), up to the second upstream NlaIII restriction site (pos. 8757). The third signal (**Figure 2A, Peak 3**), with 51.6% total tags, mapped at the end of the *gfp* orf,

with sense transcription, ending at a putative polyadenylation signal at pos. 7133-7138, reaching an upstream NlaIII restriction site at pos. 7072. The massive presence of these gfp tags suggests that viral transcription is skewed in favor of stable gfp transcripts, rendering this vector particularly suited as a reporter of HIV infection. Nevertheless, it seems unlikely that this putative poly(A) signal (located in the *env* orf) is used during transcription of wild type HIV-1. Further investigation of viral transcripts using wild-type HIV-1 should confirm this. The fourth signal (**Figure 2A, Peak 4**), representing 4.8% of total tags, is consistent with antisense viral transcription using a polyadenylation signal at pos. 4908-4903, reported previously by Landry *et al.* (26) , up to a first downstream NlaIII restriction site at pos. 5099. The fifth signal (**Figure 2A, Peak 5**), with 2.1% of total tags, corresponds to antisense transcription ending at an unidentified polyadenylation site, and up to the first downstream NlaIII restriction site at pos. 1226. The remaining 12.2% of tags mapping elsewhere to the genome correspond mostly to the use of further NlaIII restriction sites or alternative, putative poly(A) sites.

The technical procedure, from SAGE library preparation to high-throughput sequencing, keeps track of strand specificity. Sequences are read from a universal primer in the first adapter (on the poly(A) side) towards the NlaIII site. Sense tags will map 3' from the NlaIII site, while antisense tags will map 5' of the NlaIII site. Globally, 20.5% of tags mapped to the HIV-1 vector genome in the antisense orientation (**Figure 2A**). Peaks 4 and 5 are comprised mostly of antisense tags (88% and 92%, respectively); however, antisense tags were also present at the 3' end of the viral genome (40.5% and 49.0%, for peaks 1 and 2, respectively). In contrast, a negligible amount of tags in the antisense orientation mapped to the third peak, corresponding to GPF (only 1 out of 125,970 tags). With the exception of peak four, we did not identified canonical poly(A) signals (AATAAA, ATTAAA, AGTAAA, AAGGAA) that would associate with the observed

179 accumulation of antisense tags. Confirming and understanding HIV-1 antisense transcription is
180 important because of the a possible roles for such transcripts in the regulation of viral expression
181 (54), the generation of antisense proteins (26, 30, 34, 35), or the production of cryptic epitopes
182 (4). However, the assessment of antisense transcription and cognate poly(A) motifs is not
183 straightforward. Specifically, Landry *et al.* indicated that RT-PCR cannot reliably separate sense
184 vs anti-sense transcripts due to endogenous RT priming, and would therefore require the use of 5'
185 LTR-deleted pNL4-3 constructs , and cloning and sequencing of the 3'RACE amplified products
186 (26). First, we repeated a deep sequencing experiment, using barcoded adapters (initially
187 designed for sample multiplexing) during the library preparation of SAGE-Seq, as it was
188 suggested that barcoding samples might generate less artifacts (53). Although the overall
189 distribution profile of the tags was comparable, only about 0.9% antisense tags were detected. To
190 further estimate the proportion of viral antisense transcription, we performed Northern blots
191 using specific sense and antisense probes (**Figure 2B**). 2-kb, 4-kb and 9-kb classes of viral sense
192 transcripts were detected expectedly according to the specific probe used. However, none of the
193 probes designed to anneal to antisense transcripts revealed any detectable signal. These data
194 suggest that viral antisense transcription is more consistent with 0.9% abundance than with
195 20.5% abundance, calling for caution when analyzing antisense transcription by novel high-
196 throughput technologies, and highlighting the necessity to validate data by alternative methods.
197
198 Three additional technical aspects deserve comment. The use of the HIV-1 vector containing a
199 GFP ORF in the place of *env* could disturb transcription and splicing. The Northern blot in **Figure 2B**
200 excludes significant defects in viral transcription and splicing. The integrity of transcription was
201 further assessed by RT-qPCR using primers flanking the major introns of HIV to confirm the detection of
202 multiply, singly and unspliced transcripts (data not shown). We examined the degree of conservation

of the tag sequence to assess whether the tags displayed a perfectly matched sequence corresponding to the HIV-1 vector DNA. We tested 0, 1 and 2 mismatches to ensure that we would capture the putative error rate of the viral reverse transcriptase, as well as base miscalling of the SOLiD sequencing software. Upon HIV vector-specific tag analysis, 96.97% of the tags displayed a perfect match with the HIV vector genome, 2.92% presented one mismatch and only 0.11% presented two mismatches. We also identified a limited number of reads (<0.01%) that mapped to the HIV-1 vector genome in the mock infected cells (**Table 1**). Upon inspection, all tags were HIV-specific, indicating a small level of contamination occurring through the experimental procedure (from culture to library preparation).

The cellular transcriptome during infection. The high-throughput sequencing analysis of SAGE-RNA allowed the detection of 36,271 human expressed genes (79.8%), according to the RefSeq database (**Table 1**). On average, each transcript was represented by 815 tags (ranging from 0 to 3.8 millions, and with a median of 14.7). While the overall distribution in expression levels was not different for mock and HIV-1 infected cells (**Figure 3**), the identity of transcripts contained in mock or HIV-infected samples was significantly modified. Out of the detected transcripts, 826 (2.3%) were significantly differentially modulated between mock and HIV-infected samples at a stringent adjusted *P* value of 10^{-4} at 24h post-infection. Ingenuity pathway analysis (www.ingenuity.com) identified "cellular growth and proliferation" and "RNA post-transcriptional modification" as prominent modulated functions, including as relevant networks those related to cell cycle, DNA replication and repair, gene expression, and cell death. There was down regulation in HIV-infected cells of many of the genes associated to those networks, consistent with marked cellular compromise and stress. The complete set of differentially expressed genes is included in **Supplemental File 3**. Of notice, some of the differentially

expressed genes include non-protein coding transcripts such as spliceosomal and small nucleolar RNAs that are not generally thought of as polyadenylated, and should not have been captured by SAGE. However, there is increasing evidence that poly(A) tails can be added to such RNAs, possibly marking them for degradation (5, 25).

We assessed the overlap between the set of genes that were identified as differentially expressed in the current study with those from various microarray studies in the literature that investigated differential expression in CD4+ T cells (42), CD8+ T cells (41), monocytes (17), and lymph nodes (29) during HIV-1 infection *in vivo*, as well as with a set of validated genes compiled by Giri *et al.* (16) from microarray studies published between 2000 and 2006. The overlap varied from 4% in the study of circulating monocytes in HIV-infected individuals (17) to 8% in the analysis of lymphatic tissue in the setting of various stages of HIV-1 infection *in vivo* (29). There was also a 6% overlap with a set of validated genes curated by Giri *et al.* (16). The final shortlist of 52 genes common between the present and one or more of the previous studies (**Supplemental File 4**) was also enriched for genes involved in cell cycle, DNA replication and repair, and cell proliferation. However, it should be underscored that differences in the nature of the techniques (deep-sequencing vs. microarray), a lack of recent studies using new generation microarrays for the transcriptome analysis of *in vitro* infection of T cell lines - the comparable experiment as the one completed herein, and general differences in study design makes comparison between studies of unclear significance.

The approach also assessed whether HIV-1 infection modulates expression of repetitive elements or endogenous retroviruses. We observed very active transcription of these elements, with one transcript in 237 being of such origin, corresponding in average to 123,123 reads in mock

(0.40%) and 129,149 reads in HIV-1 (0.45%) mapping to the genomic Repbase repository (**Figure 3**). Of these, 4021 and 4451 average reads mapped to human endogenous retroviruses (HERV) in mock and in HIV-infected cells, respectively (**Figure 3**). The modest increase in HERV in the HIV-infected cell results mainly from the contribution of HERVK which increases from 3,331 to 3,814 average tags. The biological role of the prominent transcriptional activity of endogenous retroelements, *i.e.* regulatory activity or generation of translated products, needs further analysis (13).

Conclusions. Analysis of deep sequence data using SAGE allowed the precise measurement of expression level of the proviral genome in HIV-1-infected cells and identified a small component of antisense viral transcription. Improvements in RNA-Seq will increasingly deliver information on both strand specificity and nature of the transcript, including splice forms, and sequence variation, that will facilitate the study of the dynamics of viral-host interactions (28). Extending this approach to replication-competent HIV-1 isolates, and to different cellular backgrounds, may reveal differences in viral-host interactions due to specific strain or cellular factors.

Authors' contributions

GL and JR performed the bioinformatics analyses, FU and MM carried out HIV-1 production, cell infections and RNA extraction, SD performed Northern blot analyses, AT and AC lead the project and wrote the manuscript, NB edited and proofread the manuscript.

Acknowledgements

We thank Dr Marzanna Künzli and Dr Sirisha Aluri from the Functional Genomics Center Zurich (FGCZ) facility for SAGE library preparation and SOLiD sequencing, with the great help and support from Dr Gerrit Kuhn (Life Technologies). The following reagent was obtained through the NIH AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH: pNL4-3-deltaE-EGFP (Cat# 11100) from Drs. Haili Zhang, Yan Zhou, and Robert Siliciano (56). This work was supported by the Swiss National Science Foundation, grant 310030-130699.

Competing interests

The authors declare that they have no competing interests.

References

1. **Anders, S., and W. Huber.** 2010. Differential expression analysis for sequence count data. Nature Precedings**:**1-5.

2. **Archer, J., M. S. Braverman, B. E. Taillon, B. Desany, I. James, P. R. Harrigan, M. Lewis, and D. L. Robertson.** 2009. Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4 (CXCR4)-using HIV-1 with ultra-deep pyrosequencing. AIDS (London, England) **23:**1209-1218.

3. **Arendt, C. W., and D. R. Littman.** 2001. HIV: master of the host cell. Genome Biol **2:**REVIEWS1030.

4. **Bansal, A., J. Carlson, J. Yan, O. T. Akinsiku, M. Schaefer, S. Sabbaj, A. Bet, D. N. Levy, S. Heath, J. Tang, R. A. Kaslow, B. D. Walker, T. Ndung'u, P. J. Goulder, D. Heckerman, E. Hunter, and P. A. Goepfert.** 2010. CD8 T cell response and evolutionary pressure to HIV-1 cryptic epitopes derived from antisense transcription. J Exp Med **207:**51-59.

5. **Bayne, E. H., S. A. White, and R. C. Allshire.** 2007. DegrAAAded into silence. Cell **129:**651-653.

6. **Blow, N.** 2009. Transcriptomics: The digital generation. Nature **458:**239-242.

7. **Borjabad, A., A. I. Brooks, and D. J. Volsky.** 2010. Gene expression profiles of HIV-1-infected glia and brain: toward better understanding of the role of astrocytes in HIV-1-associated neurocognitive disorders. J Neuroimmune Pharmacol **5:**44-62.

8. **Bushman, F. D., C. Hoffmann, K. Ronen, N. Malani, N. Minkah, H. M. Rose, P. Tebas, and G. P. Wang.** 2008. Massively parallel pyrosequencing in HIV research. AIDS (London, England) **22:**1411-1415.

9. **Carroll, E. E., R. Hammamieh, N. Chakraborty, A. T. Phillips, S. A. Miller, and M. Jett.** 2006. Altered gene expression in asymptomatic SHIV-infected rhesus macaques (Macacca mulatta). Virology journal **3:**74.

10. **Chung, H. K., C. A. Pise-Masison, M. F. Radonovich, J. Brady, J. K. Lee, S. Y. Cheon, P. Markham, A. Cristillo, and R. Pal.** 2008. Cellular gene expression profiles in rhesus macaques challenged mucosally with a pathogenic R5 tropic simian human immunodeficiency virus isolate. Viral Immunol **21:**411-423.

11. **Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber.** 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics **21:**3439-3440.

12. **Eriksson, N., L. Pachter, Y. Mitsuya, S. Y. Rhee, C. Wang, B. Gharizadeh, M. Ronaghi, R. W. Shafer, and N. Beerenwinkel.** 2008. Viral population estimation using pyrosequencing. PLoS computational biology **4:**e1000074.

13. **Garrison, K. E., R. B. Jones, D. A. Meiklejohn, N. Anwar, L. C. Ndhlovu, J. M. Chapman, A. L. Erickson, A. Agrawal, G. Spotts, F. M. Hecht, S. Rakoff-Nahoum, J. Lenz, M. A. Ostrowski, and D. F. Nixon.** 2007. T cell responses to human endogenous retroviruses in HIV-1 infection. PLoS pathogens **3:**e165.

14. **Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G.**

Smyth, L. Tierney, J. Y. Yang, and J. Zhang. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5:**R80.

15. **George, M. D., D. Verhoeven, Z. McBride, and S. Dandekar.** 2006. Gene expression profiling of gut mucosa and mesenteric lymph nodes in simian immunodeficiency virus-infected macaques with divergent disease course. J Med Primatol **35:**261-269.

16. **Giri, M. S., M. Nebozhyn, L. Showe, and L. J. Montaner.** 2006. Microarray data on gene modulation by HIV-1 in immune cells: 2000-2006. Journal of leukocyte biology **80:**1031-1043.

17. **Giri, M. S., M. Nebozyhn, A. Raymond, B. Gekonge, A. Hancock, S. Creer, C. Nicols, M. Yousef, A. S. Foulkes, K. Mounzer, J. Shull, G. Silvestri, J. Kostman, R. G. Collman, L. Showe, and L. J. Montaner.** 2009. Circulating monocytes in HIV-1-infected viremic subjects exhibit an antiapoptosis gene signature and virus- and host-mediated apoptosis resistance. J Immunol **182:**4459-4470.

18. **Harman, A. N., M. Kraus, C. R. Bye, K. Byth, S. G. Turville, O. Tang, S. K. Mercier, N. Nasr, J. L. Stern, B. Slobedman, C. Driessen, and A. L. Cunningham.** 2009. HIV-1-infected dendritic cells show 2 phases of gene expression changes, with lysosomal enzyme activity decreased during the second phase. Blood **114:**85-94.

19. **Hoffmann, C., N. Minkah, J. Leipzig, G. Wang, M. Q. Arens, P. Tebas, and F. D. Bushman.** 2007. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic acids research **35:**e91.

20. **Hyrcza, M. D., C. Kovacs, M. Loutfy, R. Halpenny, L. Heisler, S. Yang, O. Wilkins, M. Ostrowski, and S. D. Der.** 2007. Distinct transcriptional profiles in ex vivo CD4+ and CD8+ T cells are established early in human immunodeficiency virus type 1 infection

and are characterized by a chronic interferon response as well as extensive transcriptional changes in CD8+ T cells. Journal of virology **81:**3477-3486.

21. **Imbeault, M., M. Ouellet, and M. J. Tremblay.** 2009. Microarray study reveals that HIV-1 induces rapid type-I interferon-dependent p53 mRNA up-regulation in human primary CD4+ T cells. Retrovirology **6:**5.

22. **Ji, H., N. Masse, S. Tyler, B. Liang, Y. Li, H. Merks, M. Graham, P. Sandstrom, and J. Brooks.** 2010. HIV drug resistance surveillance using pooled pyrosequencing. PLoS ONE **5:**e9263.

23. **Jojic, V., T. Hertz, and N. Jojic.** 2008. Population sequencing using short reads: HIV as a case study. Pac Symp Biocomput**:**114-125.

24. **Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz.** 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res **110:**462-467.

25. **LaCava, J., J. Houseley, C. Saveanu, E. Petfalski, E. Thompson, A. Jacquier, and D. Tollervey.** 2005. RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. Cell **121:**713-724.

26. **Landry, S., M. Halin, S. Lefort, B. Audet, C. Vaquero, J. M. Mesnard, and B. Barbeau.** 2007. Detection, characterization and regulation of antisense transcripts in HIV-1. Retrovirology **4:**71.

27. **Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg.** 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10:**R25.

28. **Levin, J. Z., M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson, N. Friedman, A. Gnirke, and A. Regev.** 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature methods **7:**709-715.

29. **Li, Q., A. J. Smith, T. W. Schacker, J. V. Carlis, L. Duan, C. S. Reilly, and A. T. Haase.** 2009. Microarray analysis of lymphatic tissue reveals stage-specific, gene expression signatures in HIV-1 infection. J Immunol **183:**1975-1982.

30. **Ludwig, L. B., J. L. Ambrus, Jr., K. A. Krawczyk, S. Sharma, S. Brooks, C. B. Hsiao, and S. A. Schwartz.** 2006. Human Immunodeficiency Virus-Type 1 LTR DNA contains an intrinsic gene producing antisense RNA and protein products. Retrovirology **3:**80.

31. **Martinez-Marino, B., H. Foster, Y. Hao, and J. A. Levy.** 2007. Differential gene expression in CD8(+) cells from HIV-1-infected subjects showing suppression of HIV replication. Virology **362:**217-225.

32. **Matsumura, H., D. H. Kruger, G. Kahl, and R. Terauchi.** 2008. SuperSAGE: a modern platform for genome-wide quantitative transcript profiling. Curr Pharm Biotechnol **9:**368-374.

33. **Matsumura, H., K. Yoshida, S. Luo, E. Kimura, T. Fujibe, Z. Albertyn, R. A. Barrero, D. H. Kruger, G. Kahl, G. P. Schroth, and R. Terauchi.** 2010. High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. PLoS ONE **5**.

34. **Michael, N. L., M. T. Vahey, L. d'Arcy, P. K. Ehrenberg, J. D. Mosca, J. Rappaport, and R. R. Redfield.** 1994. Negative-strand RNA transcripts are produced in human immunodeficiency virus type 1-infected cells and patients by a novel promoter downregulated by Tat. Journal of virology **68:**979-987.

35. **Miller, R. H.** 1988. Human immunodeficiency virus may encode a novel protein on the genomic DNA plus strand. Science (New York, N.Y **239:**1420-1422.

36. **Mitchell, R., C. Y. Chiang, C. Berry, and F. Bushman.** 2003. Global analysis of cellular transcription following infection with an HIV-based vector. Mol Ther **8:**674-687.

37. **Montano, M., M. Rarick, P. Sebastiani, P. Brinkmann, M. Russell, A. Navis, C. Wester, I. Thior, and M. Essex.** 2006. Gene-expression profiling of HIV-1 infection and perinatal transmission in Botswana. Genes Immun **7:**298-309.

38. **Morrissy, A. S., R. D. Morin, A. Delaney, T. Zeng, H. McDonald, S. Jones, Y. Zhao, M. Hirst, and M. A. Marra.** 2009. Next-generation tag sequencing for cancer gene expression profiling. Genome Res **19:**1825-1835.

39. **Naldini, L., U. Blomer, P. Gallay, D. Ory, R. Mulligan, F. H. Gage, I. M. Verma, and D. Trono.** 1996. In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. Science (New York, N.Y **272:**263-267.

40. **Pruitt, K. D., T. Tatusova, and D. R. Maglott.** 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research **35:**D61-65.

41. **Rotger, M., J. Dalmau, A. Rauch, P. McLaren, S. Bosinger, R. Martinez, N. G. Sandler, A. Roque, J. Liebner, M. Battegay, E. Bernasconi, P. Descombes, I. Erkizia, J. Fellay, B. Hirschel, J. M. Miró, E. Palou, M. Hoffmann, M. Massanella, J. Blanco, M. Woods, H. F. Günthard, P. de Bakker, D. C. Douek, G. Silvestri, J. Martinez-Picado, and A. Telenti.** 2011. Comparative transcriptomics of extreme phenotypes of human HIV-1 infection and SIV infection in sooty mangabey and rhesus macaque. The Journal of clinical investigation **in press**.

42. **Rotger, M., K. K. Dang, J. Fellay, E. L. Heinzen, S. Feng, P. Descombes, K. V. Shianna, D. Ge, H. F. Gunthard, D. B. Goldstein, and A. Telenti.** 2010. Genome-wide

mRNA expression correlates of viral control in CD4+ T-cells from HIV-1-infected individuals. PLoS pathogens **6:**e1000781.

43. **Rozera, G., I. Abbate, A. Bruselles, C. Vlassi, G. D'Offizi, P. Narciso, G. Chillemi, M. Prosperi, G. Ippolito, and M. R. Capobianchi.** 2009. Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. Retrovirology **6:**15.

44. **Sambrook, J. F., E.F.; Maniatis, T.** 1989. Northern Hybridization, p. 7.39-37.52. *In* N. N. Ford, C.; Ferguson, M. (ed.), Molecular Cloning: A Laboratory Manual, 2nd ed, vol. 1. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

45. **Shendure, J.** 2008. The beginning of the end for microarrays? Nature methods **5:**585-587.

46. **Simen, B. B., J. F. Simons, K. H. Hullsiek, R. M. Novak, R. D. Macarthur, J. D. Baxter, C. Huang, C. Lubeski, G. S. Turenchalk, M. S. Braverman, B. Desany, J. M. Rothberg, M. Egholm, and M. J. Kozal.** 2009. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. The Journal of infectious diseases **199:**693-701.

47. **Sirois, M., L. Robitaille, R. Sasik, J. Estaquier, J. Fortin, and J. Corbeil.** 2008. R5 and X4 HIV viruses differentially modulate host gene expression in resting CD4+ T cells. AIDS research and human retroviruses **24:**485-493.

48. **Solis, M., P. Wilkinson, R. Romieu, E. Hernandez, M. A. Wainberg, and J. Hiscott.** 2006. Gene expression profiling of the host response to HIV-1 B, C, or A/E infection in monocyte-derived dendritic cells. Virology **352:**86-99.

49. **t Hoen, P. A., Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. Vossen, R. X. de Menezes, J. M. Boer, G. J. van Ommen, and J. T. den Dunnen.** 2008. Deep

sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic acids research **36:**e141.

50. **Tsibris, A. M., B. Korber, R. Arnaout, C. Russ, C. C. Lo, T. Leitner, B. Gaschen, J. Theiler, R. Paredes, Z. Su, M. D. Hughes, R. M. Gulick, W. Greaves, E. Coakley, C. Flexner, C. Nusbaum, and D. R. Kuritzkes.** 2009. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. PLoS ONE **4:**e5683.

51. **Vahey, M. T., Z. Wang, Z. Su, M. E. Nau, A. Krambrink, D. J. Skiest, and D. M. Margolis.** 2008. CD4+ T-cell decline after the interruption of antiretroviral therapy in ACTG A5170 is predicted by differential expression of genes in the ras signaling pathway. AIDS research and human retroviruses **24:**1047-1066.

52. **Wang, G. P., A. Ciuffi, J. Leipzig, C. C. Berry, and F. D. Bushman.** 2007. HIV integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome Res **17:**1186-1194.

53. **Wu, Z. J., C. A. Meyer, S. Choudhury, M. Shipitsin, R. Maruyama, M. Bessarabova, T. Nikolskaya, S. Sukumar, A. Schwartzman, J. S. Liu, K. Polyak, and X. S. Liu.** 2010. Gene expression profiling of human breast tissue samples using SAGE-Seq. Genome Res **20:**1730-1739.

54. **Yeung, M. L., Y. Bennasser, K. Watashi, S. Y. Le, L. Houzet, and K. T. Jeang.** 2009. Pyrosequencing of small non-coding RNAs in HIV-1 infected cells: evidence for the processing of a viral-cellular double-stranded RNA hybrid. Nucleic acids research **37:**6575-6586.

55. **Zagordi, O., R. Klein, M. Daumer, and N. Beerenwinkel.** 2010. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic acids research.

56. **Zhang, H., Y. Zhou, C. Alcock, T. Kiefer, D. Monie, J. Siliciano, Q. Li, P. Pham, J. Cofrancesco, D. Persaud, and R. F. Siliciano.** 2004. Novel single-cell-level phenotypic assay for residual drug susceptibility and reduced replication capacity of drug-resistant human immunodeficiency virus type 1. Journal of virology **78:**1718-1729.

Figures

Figure 1 - Overview of experimental and analytical procedures.

Panel A, Experimental pipeline. Panel B, Bioinformatics pipeline.

Figure 2 - Distribution of HIV-specific tags along the viral genome.

(A) The HIV vector genome is depicted on the top panel with nucleotide positioning, NlaIII restriction sites (vertical black bars), the common sense transcription start site (TSS; red arrow) and the polyadenylation signal (poly(A); red vertical bar) are shown below. Multiple antisense transcription start sites (distributed between positions 9175 and 8714; only two are drawn here for figure simplicity; green arrows), and the poly(A) signal (vertical green bar), as described in Landry *et al.* (26), are indicated. HIV-1 open reading frames are also indicated, including the putative antisense protein (ASP). The bottom panel indicates the density of tags (vertical axis) distributed along the HIV-1 vector genome (x-axis), for both HIV-1 samples (HIV rep1 and HIV rep2) and mock samples (Mock rep1 and Mock rep2). The five major peaks are numbered at the bottom of the figure, displaying the number of tags at each peak, as well as the proportion of these tags representing sense or antisense transcription. (B) Total RNA from mock or HIV-based vector infected cells was extracted using miRvana or Trizol at 22h and 24h post-exposure and subjected to Northern blot hybridization. Left panel, total RNA gel electrophoresis. Right panels, Northern blots using strand-specific oligonucleotide probes aligning to the corresponding peak of tags detected by SAGE-Seq. Probes were designed to anneal with sense transcripts (red probes, top panels) or antisense transcripts (green probes, bottom panels), with nucleic acid sequence

complementary to the identified peak tag sequence. The plasmid encoding HIV vector genome digested with BglII and AflII was used as positive control for hybridization (+).

Figure 3 - Analysis of cellular transcripts in mock and HIV-infected cells.

Average tag density analysis in mock (black lines) and HIV-1 (red lines) samples for total cellular transcripts (lines), repetitive elements (dashed lines) and HERV elements (dotted lines). The figure indicates that most cellular transcripts were detected with 1 to 1000 tags (log 0 to 3, average 2.9 log), across the complete transcriptome. Only a few transcripts were identified in larger numbers (*e.g.* above 10000, log 4). Lesser number of tags was observed for the expressed repetitive elements (average 2.3 log) including HERVs (average 1.9 log). Statistical differences between mock and HIV-1 distributions were assessed by Wilcoxon test.

Tables

Table 1 - Mapping to the human genome, human transcriptome, and HIV-1 vector genome.

	Human		HIV-1 vector	
Sample	Genome tags[a]	Transcriptome tags[b]	HIV-specific tags[c]	GFP-specific tags[d]
Mock rep1	30,668,095	38,730,420	45	69
Mock rep2	31,611,573	43,691,287	2,334	2,862
HIV rep1	27,894,076	33,337,832	11,7391	13,5487
HIV rep2	29,484,387	36,666,667	10,9122	12,6344

[a]Alignment/mapping to GRCh37 allowed 1 mismatch (n=1), 1 hit (m=1), read length of 24 (rl=24).

[b]Alignment/mapping to RefSeq allowed 1 mismatch (n=1), 10 multiple hits (m=10), read length of 24 (rl=24).

[c]Aligmnent/mapping to *HIV NL4-3Δenv/eGFP* was performed allowing 2 mismatches (n=2) and 2 multiple hits (m=2), read length of 24 (rl=24).

Supplemental Material

Supplemental File 1 – Completion of viral cycle in SupT1 cells.

(A) SupT1 cells were highly susceptible to HIV-based vector infection (compared to CEM and Jurkat T cell lines). At 24h and 36h post-infection, 54% and 93% of cells were successfully transduced. (B) To establish the optimal time point to capture the completion of the viral replication cycle, as reflected by particle release (measured by p24 ELISA in the supernatant), we performed a time course collecting materials every 2 hours. The plots represent here non-cumulative estimates, *i.e. de novo* production, calculated by subtracting the measurement at a given time point by the prior measurement. Shown are the percentage of GFP+ cells and mean fluorescence intensity (MFI), representing the success of viral translation (green), which peaks at 22 hours, and followed by the peak of extracellular p24 (red) at 24 hours expected 2-4 after the peak of translation.

File format: pdf

This file can be viewed with: Adobe Acrobat Reader

Supplemental File 2 – Optimal mapping parameters for read analysis.

Read length (rl), number of mismatches (n) and multiple hits allowed (m) were tested to maximize the number of tags mapping to HIV-1 while minimizing the number of tags spuriously mapping to both viral and host genome. Y-axis: reads mapping to HIV-1 vector genome; x-axis:

number of reads mapping to HIV-1 but not to human genome. Parameters rl, n, and m are indicated only for the best conditions.

File format: pdf

This file can be viewed with: Adobe Acrobat Reader

Supplemental File 3 – Transcripts differentially expressed between mock and HIV-1 infected cells.

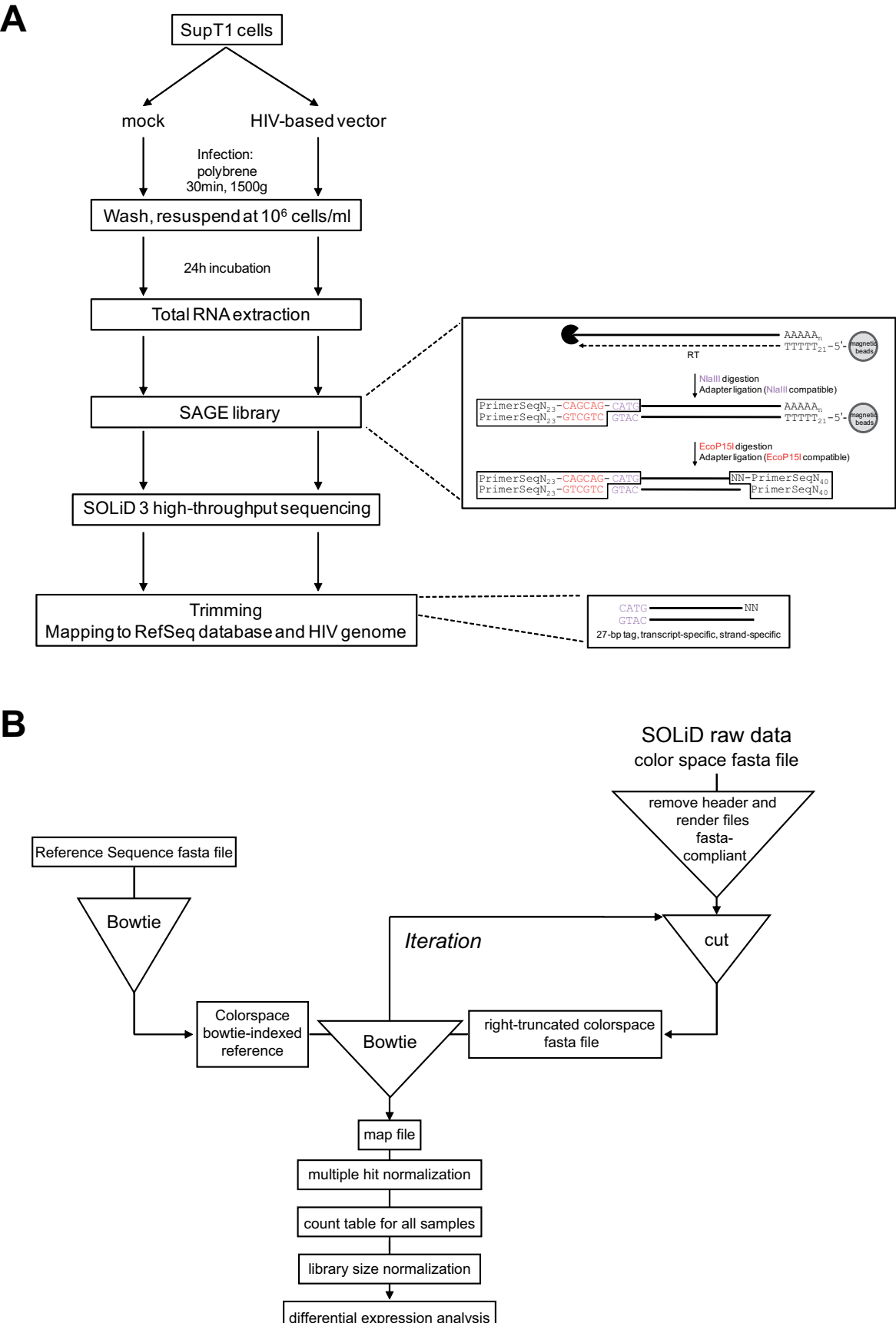
File format: xls

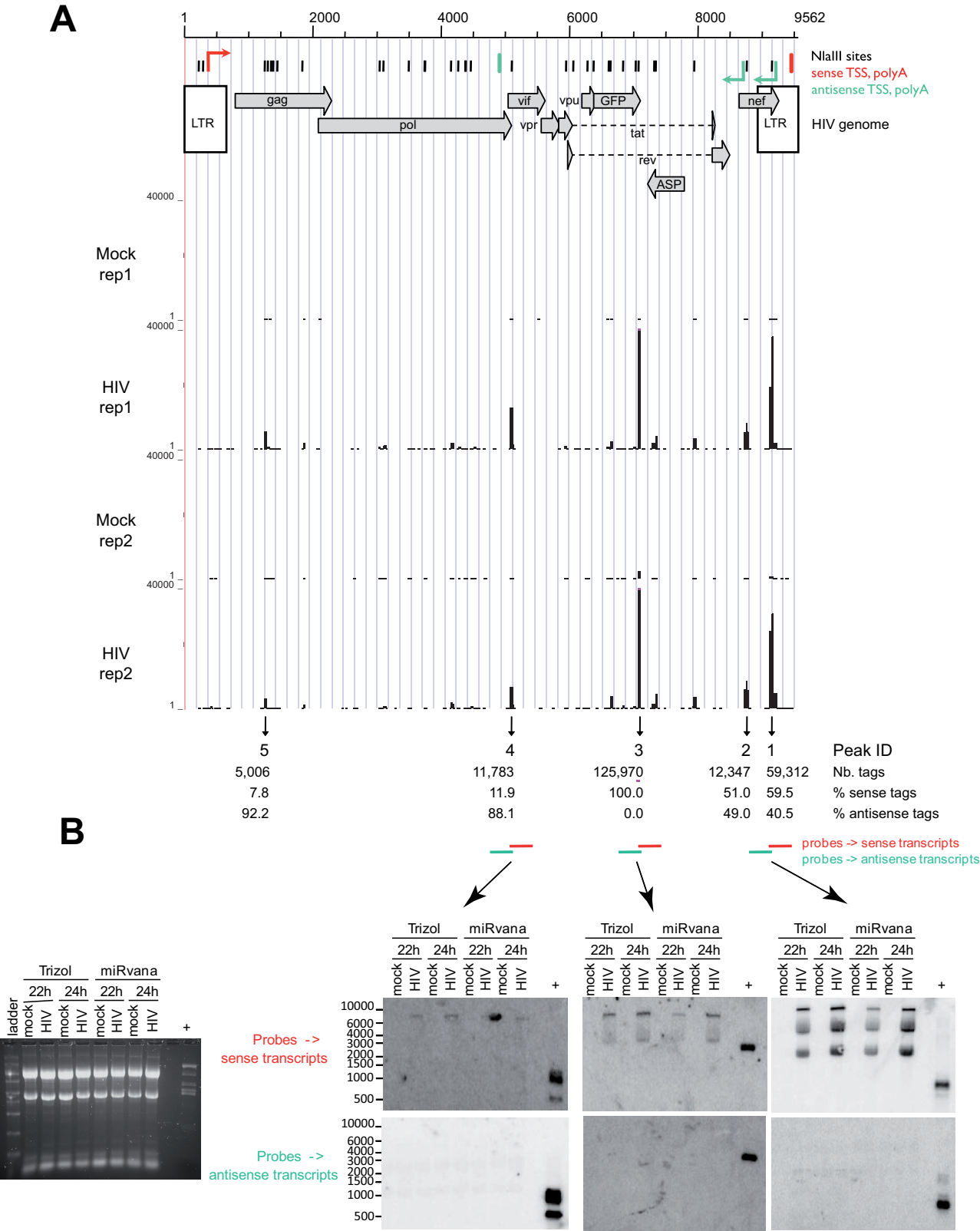
This file can be viewed with: Microsoft Excel Viewer

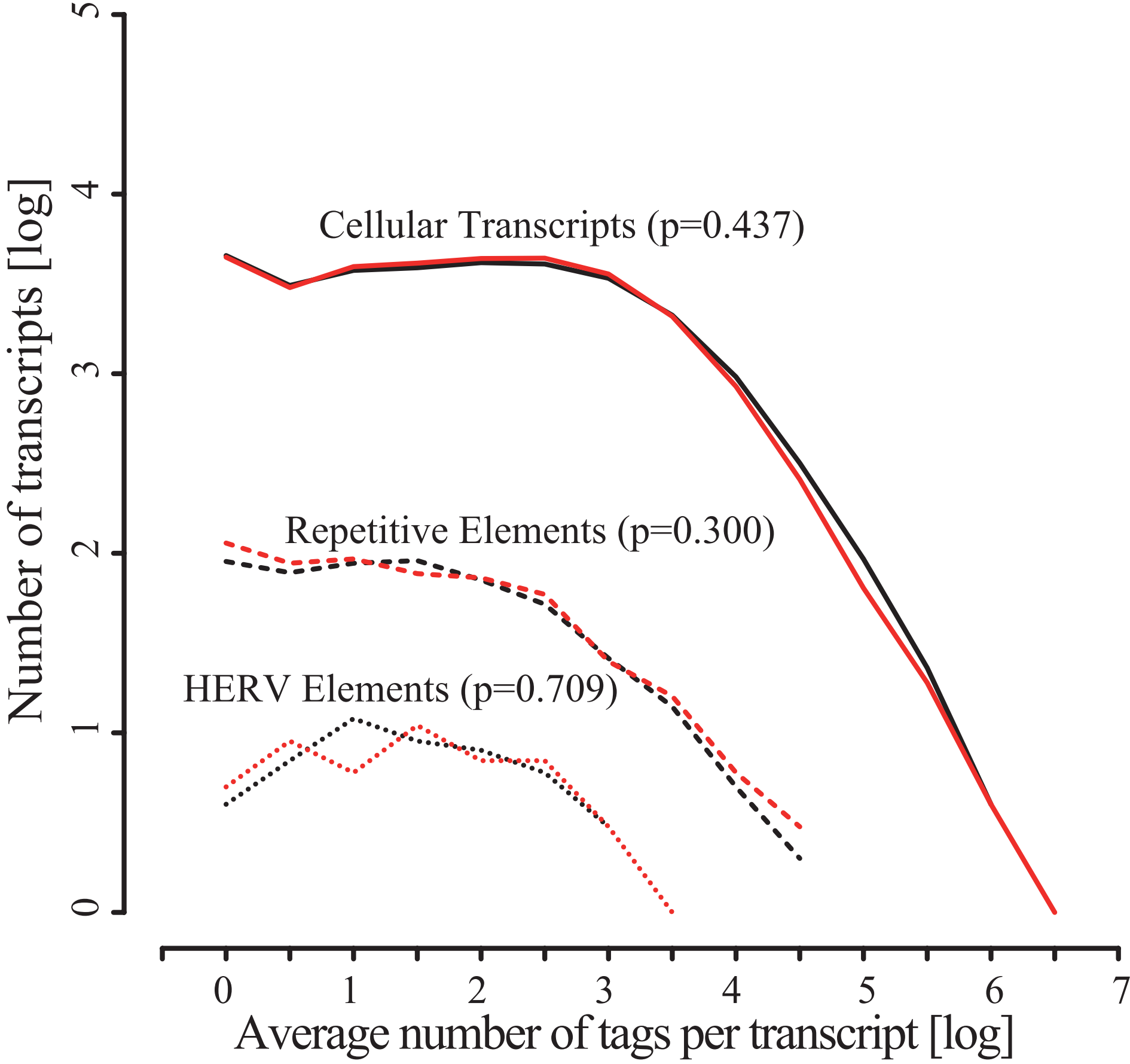
Supplemental File 4 – Transcripts identified by SAGE-Seq common to previous studies using microarray technology.

File format: xls

This file can be viewed with: Microsoft Excel Viewer


A
SupT1 cells
mock
HIV-based vector
Infection:
polybrene
30min, 1500g
Wash, resuspend at 10^6 cells/ml
24h incubation
Total RNA extraction
SAGE library
SOLiD 3 high-throughput sequencing
Trimming
Mapping to RefSeq database and HIV genome
AAAAAn
TTTTT21-5'-
magnetic beads
RT
NlaIII digestion
Adapter ligation (NlaIII compatible)
PrimerSeqN23-CAGCAG-CATG
PrimerSeqN23-GTCGTC GTAC
EcoP15I digestion
Adapter ligation (EcoP15I compatible)
NN-PrimerSeqN40
PrimerSeqN40
CATG
GTAC
NN
27-bp tag, transcript-specific, strand-specific
B
SOLiD raw data
color space fasta file
remove header and render files fasta-compliant
Reference Sequence fasta file
Bowtie
Iteration
cut
Colorspace bowtie-indexed reference
Bowtie
right-truncated colorspace fasta file
map file
multiple hit normalization
count table for all samples
library size normalization
differential expression analysis


A
1
2000
4000
6000
8000
9562
NlaIII sites
sense TSS, polyA
antisense TSS, polyA
HIV genome
LTR
gag
pol
vif
vpr
vpu
GFP
tat
rev
ASP
nef
LTR
Mock rep1
HIV rep1
Mock rep2
HIV rep2
40000
1
5
4
3
2
1
Peak ID
5,006
11,783
125,970
12,347
59,312
Nb. tags
7.8
11.9
100.0
51.0
59.5
% sense tags
92.2
88.1
0.0
49.0
40.5
% antisense tags
B
probes -> sense transcripts
probes -> antisense transcripts
Trizol
miRvana
22h
24h
mock
HIV
ladder
+
Probes -> sense transcripts
Probes -> antisense transcripts
10000
6000
4000
3000
2000
1500
1000
500


Cellular Transcripts (p=0.437)
Repetitive Elements (p=0.300)
HERV Elements (p=0.709)
Number of transcripts [log]
0
1
2
3
4
5
0
1
2
3
4
5
6
7
Average number of tags per transcript [log]