VIRULENCE
2022, VOL. 13, NO. 1, 386–413
https://doi.org/10.1080/21505594.2022.2031583

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS Check for updates

# HIV-1 provirus transcription and translation in macrophages differs from pre-integrated cDNA complexes and requires E2F transcriptional programs

Albebson L. Lim[a,b], Philip Moos[a], Christopher D. Pond[a], Erica C. Larson[a,c], Laura J. Martins[d], Matthew A. Szaniawski[d], Vicente Planelles[d], and Louis R. Barrows[a]

[a]Department of Pharmacology and Toxicology, University of Utah, Salt Lake City, Utah, USA; [b]Marine Science Institute, University of the Philippines Diliman, Quezon City, Philippines; [c]Department of Microbiology & Molecular Genetics, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; [d]Department of Pathology, University of Utah, Salt Lake City, Utah, USA

**ABSTRACT**

HIV-1 cDNA pre-integration complexes persist for weeks in macrophages and remain transcriptionally active. While previous work has focused on the transcription of HIV-1 genes; our understanding of the cellular milieu that accompanies viral production is incomplete. We have used an *in vitro* system to model HIV-1 infection of macrophages, and single-cell RNA sequencing (scRNA-seq) to compare the transcriptomes of uninfected cells, cells harboring pre-integration complexes (PIC), and those containing integrated provirus and making late HIV proteins. scRNA-seq can distinguish between provirus and PIC cells because their background transcriptomes vary dramatically. PIC cell transcriptomes are characterized by NFkB and AP-1 promoted transcription, while transcriptomes of cells transcribing from provirus are characterized by E2F family transcription products. We also find that the transcriptomes of PIC cells and Bystander cells (defined as cells not producing any HIV transcript and thus presumably not infected) are indistinguishable except for the presence of HIV-1 transcripts. Furthermore, the presence of pathogen alters the transcriptome of the uninfected Bystander cells, so that they are distinguishable from true control cells (cells not exposed to any pathogen). Therefore, a single cell comparison of transcriptomes from provirus and PIC cells provides a new understanding of the transcriptional changes that accompany HIV-1 integration.

**ARTICLE HISTORY**
Received 18 October 2021
Revised 2 December 2021
Accepted 11 January 2022

**KEYWORDS**
HIV-1; macrophages; single cell RNA sequencing; infection; pre-integrated cDNA complex transcription; provirus transcription

## Introduction

The major barrier to curing HIV-1 in patients is a small reservoir of cells that are latently infected and impervious to immune recognition and clearance [1–3]. The study of HIV-1 latency is complicated by the fact that latently infected cells *in vivo* are extremely rare. It is a drawback that many studies of latency have relied on bulk sequencing endpoints. Under these conditions, the specific parameters defining the latently infected cell are diluted in the context of a vast heterogeneous population. In addition, multiple mechanisms can result in latency reversal and therefore one latently infected cell may differ from the next [4]. Thus, averaging data from a heterogeneous cell population, such as data obtained by bulk sequencing studies, leads to confusion rather than clarification.

Following infection and reverse transcription, the pre-integration cDNA complex (PIC), in which the HIV-1 genome may take linear or circular forms, serves as a template for transcription [5–7]. In dividing T cells, the PIC is short-lived, and the transient transcription of genes from the PIC is considered irrelevant [6]. However, in quiescent T cells, the PIC is longer- lived and even results in sufficient transcription for virion production in response to latency reversal agents [8]. HIV-1 latency in macrophages is multi-factorial and the dynamics of PIC integration are different [9–12]. Macrophages are known for prolonged viral production while exhibiting resistance to host cell lethality, they reside in multiple tissue reservoirs where viral production appears to be influenced by the micro-environments. They also exhibit T cell-like post-integration latency, showing sensitivity to the chromatin environment, the absence of transcriptional activation, the presence of transcriptional repressors, and host antiviral processes [9–12]. Pre-integration latency is the process that we address here. Macrophages have been shown to harbor transcriptionally active pre-integrated HIV-1 cDNA for months; however, this PIC mRNA is not thought to result in the production of virus [10–12].

**CONTACT** Louis R. Barrows lbarrows@pharm.utah.edu

The study of macrophage latency *in vivo* has proven difficult due to the very scarce number of cells. We have used a reporter HIV-1 infection of PMA activated THP-1 cells to model infection of macrophages. This model was selected to provide consistent induction responses in repeated experiments to provide statistical certainty for our conclusions (see Results, below).

We used single-cell RNA sequencing (scRNA-seq) techniques [13,14] to show that in activated THP-1 cells, many of the cells in a culture infected with our HIV-1 construct are producing HIV transcripts, while only a minority are producing Gag/p24 and other HIV-1 proteins. Levels of HIV-1 transcripts do not correlate with virus production, since many of the cells harboring PIC complexes have transcript loads comparable to cells transcribing from integrated provirus. This means that a high load of viral transcript is not a sufficient switch to reverse latency. Overall, within the limitations of 10X Genomics technology [15], transcripts for gag-pol, tat, env, and nef are found in higher numbers of “PIC cells”, compared to “Provirus cells” (those transcribing from integrated HIV-1 cDNA and producing mCherry, gag and other proteins). However, the loads of these transcripts detected in PIC and Provirus cells are similar. Transcripts for gag, vif, vpr, rev, vpu, and the marker gene mCherry are detectable in relatively equivalent percentages of cells transcribing from PIC or provirus, and at similar levels per cell. In no case did Provirus cells appear to be producing any of the detected HIV-1 transcripts at a higher prevalence than PIC cells.

Quite notably, the background transcriptomes of cells harboring HIV-1 PIC are not detectably altered by PIC transcription. Unsupervised clustering shows cells containing PIC transcripts to be distributed equally throughout the PIC/Bystander cell cluster. “Bystander” cells are defined as those cells not containing any detectable HIV-1 transcript. Thus, cells harboring HIV-1 PIC appear “oblivious” to the presence of HIV-1 gene transcription at the transcriptome level, even though some have been reported in the literature to be producing detectable levels of Nef, other HIV-1 proteins, and chemokines [9,10,15–19].

Transcription Factor Targeting analysis [20–22] shows that NFkB and AP-1 transcription products are predominant in the transcriptomes of PIC and Bystander cells. In contrast, Provirus cluster cells, on average, contain higher total amounts of viral transcripts, and their transcriptomes predominantly exhibit E2F promoter family transcription products. These cells make detectable amounts of p24, mCherry, and Vpu proteins. This seems counter-intuitive because NFkB and AP-1 are transcription sites in the HIV-1 LTR that promote transcription from the provirus [23,24]. In addition, E2F has been reported to suppress HIV-1 transcription [25–28]. Nevertheless, in our model it is clear that when cells are making late HIV-1 gene proteins, the transcriptomes exhibit activation of E2F regulated transcripts, while NFkB and AP-1 regulated genes are relatively suppressed. Western blotting data agrees with the transcription factor analysis in that Rb phosphorylation is increased in the Provirus cluster cells, which correlates with increased E2F activation.

## Results

### *DHIV3 infection of activated THP-1 macrophages yields cells with transcriptomes containing DHIV3 mRNA but otherwise identical to uninfected cells*

We used a VSV-G-pseudotyped DHIV3 virus that expressed mCherry to promote consistent levels of viral entry in PMA-activated THP-1 cells and to allow for easier interpretation of post-cell entry events (Fig. S-1) [29,30]. DHIV3 is replication-deficient due to the replacement of part of the nef gene with the reporter gene. Following infection, flow cytometry data clearly revealed two clusters of THP-1 cells, mCherry positive cells (usually from 2 to 20% of the total population, depending on DHIV3 titer) and mCherry negative cells (Figure 1(a,**b**)).

Quantifying DHIV3 infection by scRNA-seq closely agreed with the flow cytometry data in that two clusters of cells were clearly identified, based on an individual cell’s background transcriptome, in percentages that agreed with flow cytometry of replicate cultures (Figure 1(c)). Data analysis is described in detail in the Methods. Briefly, raw fastq files were generated, aligned to a custom reference genome (GRCh38 augmented with mCherry and HIV genes) and per cell gene counts generated with the 10X Cell Ranger pipeline. Following basic QC and filtering as suggested by Seurat, we generated both UMAP and t-SNE clustering projections [15]. Library construction targeted 5,000 cells and routinely yielded greater than 4,000 cells following Seurat analysis and quality control. Library construction and sequencing experiments were performed with technical repeats. The technical repeats were found to be statistically identical and were therefore combined (Fig. S-2).

Following Seurat analysis, the number of mean reads per cell was approximately 35,000, with a median of more than 2,500 genes detected per transcriptome, and greater than 15,000 different gene transcripts detected in the overall library. Control cultures yielded slightly larger numbers of cells captured, with greater than

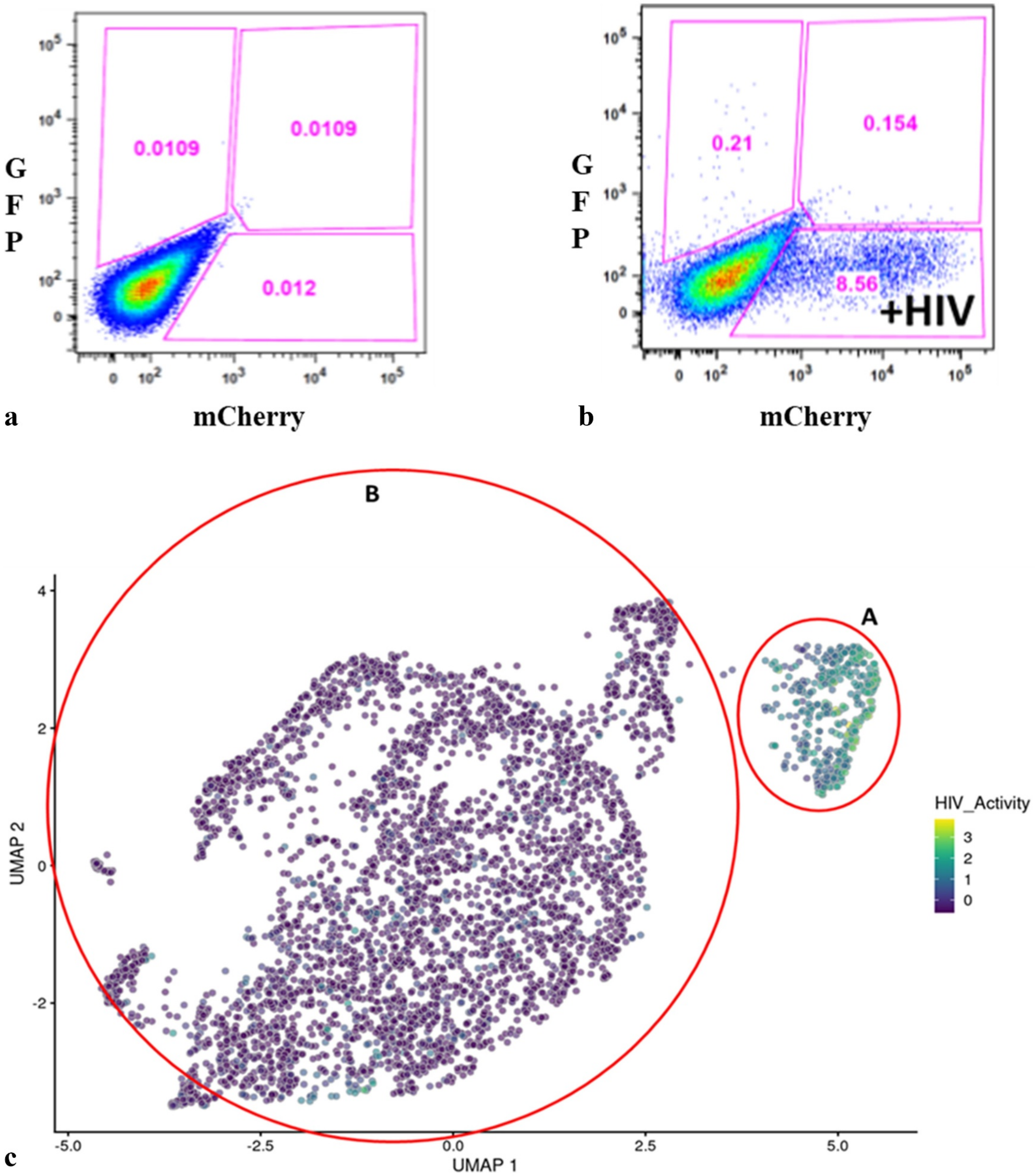

**Figure 1.** Flow cytometry analysis of DHIV3-mCherry infected THP-1 cells and UMAP projection of scRNA seq data from replicate experiment. Panel **A**) mock infection of PMA activated THP-1 cells. **B**) PMA activated THP-1 cells infected with DHIV3-mCherry. Absicca, mCherry (Texas Red) emission. Ordinate, GFP (FITC) emission. mCherry positive cells equal approximately 8.5% of the total viable cell population. **C)** UMAP projection of a duplicate culture (HIVreplicate1) is shown in panel C. Greater than 14,000 different cellular genes were detected in this analysis, including the 9 viral genes and mCherry message originating from DHIV3-mCherry. A semi-supervised two cluster model was adopted, the smaller "Provirus cluster" (cluster A) was 8.1% of the total cell population, approximately equivalent to the percentage of mCherry positive cells from panel B. The two semi-supervised clusters are circled in red. PIC/Bystander cluster is indicated as cluster B. The HIV activity scale presents the Seurat module score that is described in methods. Input data in this analysis included 33,819 PCA entries. Bar codes of the same cells tracked to the Provirus clusters, regardless of whether the clusters were generated using UMAP or Seurat-tSne tools (Fig. S-3).

6,000 cells each and with more than 3,000 genes detected per cell. All experiments were conducted with parallel cultures that were analyzed by flow cytometry for mCherry production. In the HIVreplicate1 experiment, flow cytometry indicated 8.6% mCherry positive cells (Figure 1(a)).

Figure 1(b) shows the scRNA-seq UMAP analyses of HIVreplicate1 (t-Sne projection shown in Fig. S-3). When we quantified HIV-1 transcripts in individual cells, we found that cells in the smaller, "Provirus" cluster (defined as what we now understand to be cells transcribing provirus template) were characterized by transcriptomes generally containing higher minimum levels of HIV-1 transcripts. This “Provirus” cluster accounted for approximately 8.1% of the total cell population, and closely matched the percentage of cells identified as mCherry-positive by flow cytometry in the duplicate culture, shown above. In the second “PIC/ Bystander” semi-supervised cluster of mCherry negative cells, we found that many of the macrophages (more than 50%) were producing HIV-1 transcript (Fig. S-3 plots cells with any detected HIV-1 transcript). We now understand that these cells are “PIC” cells, defined as cells containing HIV-1 mRNA transcribed from pre-integrated cDNA complexes. Some PIC cell transcriptomes appeared to contain HIV-1 transcript loads as high as many cells detected in the Provirus cluster, but most expressed lower amounts (Figure 1(b); S-3). Remarkably, these PIC cells had no other statistically significant changes to their transcriptomes that would differentiate them from the truly uninfected “Bystander” cells (again, defined as cells in the PIC/ Bystander cluster not containing detectable HIV-1 mRNA). Thus, PIC and Bystander cells made up the “PIC/Bystander” cluster.

We used a Feature map to plot the influence of the cell cycle, number of genes detected or mitochondrial transcript number on the distribution of cells containing HIV-1 transcripts (Fig S-4 B-D) [31,32]. None of these factors had any significant influence on the distribution of PIC cells throughout the PIC/Bystander cluster. We then used unsupervised clustering and violin plots of HIV transcript load to examine the distribution of cells containing HIV-1 transcripts throughout the Provirus and PIC/Bystander clusters (Figure (2)). Using K-nearest neighbors clustering, at K10, 2 clusters of cells (clusters 6 and 8) were identified that accounted for most of the Provirus cluster cells (372 of the 380 provirus cells determined by semi-supervised clustering). Therefore, the combined transcriptomes of cells in clusters 6 and 8 were compared to the combined transcriptomes of cells in clusters 1–5, 7, and 9–10, and used to generate the DGE analyses shown below. Unsupervised clustering using a range of designated K values from 10 to 130 is shown in Figure S-5.

Clusters 6 and 8 (Figure 2(b)) were characterized as containing fewer cells with a low HIV transcript load. Thus, cells in what we define as the Provirus cluster differ significantly from PIC cells in terms of average minimal DHIV3 transcript load. The key to their clustering, as defined in our semi-supervised model, was the fact that they vary significantly in background transcriptome. The multiple clusters generated in the unsupervised clustering are likely stochastic. We conclude this because, they were not influenced by the presence or absence of PIC cells (Figure 2(b)). PIC cells are distributed throughout all the PIC/Bystander clusters, regardless of the specified K value. So, for our purposes the two cluster, semi-supervised, model shown in Figure 1(b) was considered to accurately represent the interpretation gained from flow cytometry (Figure 1(a)), namely: either the cells were making mCherry protein, or they were not.

Inferring from our flow cytometry percentages, we hypothesized that only cells in the Provirus cluster were making fluorescing mCherry protein. To confirm that cells making mCherry were also making late viral proteins, the production of p24 capsid protein was

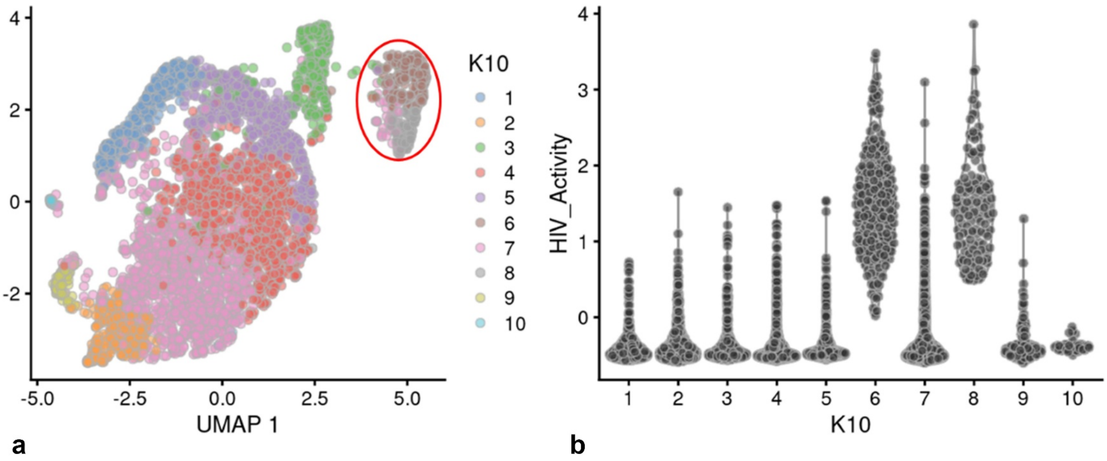

**Figure 2.** Unsupervised clustering of UMAP shown in Figure 1. Panel **A**) shows unsupervised clustering obtained at K equals 10. **B**) Violin plot of HIV-1 transctipts/cell in the 10 clusters identified at K10 (Scran’s buildSSNGraph using the PCA as input). PIC cells with detectable HIV-1 transcripts, were distributed throughout clusters 1–5, 7 and 9–10. Clusters 6 and 8 contained 372 of the 381 cells included in the semi-supervised Provirus cluster (circled in red). Stipulation of lower K values means that during analysis any one given cell is clustered with a smaller number of cells with similar transcriptomes.

quantified using flow cytometry and monoclonal mouse IgG-AG3.0 anti-Gag p24 antibody. Cells were fixed in formaldehyde, and secondary anti-mouse Alexa Fluor antibody allowed detection during flow cytometry. When cells were analyzed using FACS Canto, only cells making mCherry were found to stain for p24 (Fig. S-6). This suggests that the presumed PIC cells are not making late viral proteins in detectable amounts, while only Provirus cells appear to be making late viral proteins.

### The DHIV3 mRNA in "PIC" cells is due to PIC transcription

Integrase inhibitors have proven to be useful tools in the study of PIC transcription [8,16]. We used 25 nM MK-2048 [33,34] for our experiments to prove the presumed "Provirus" versus "PIC" status of our cell clusters. MK-2048 is a second-generation integrase inhibitor active against several integrase mutations [33]. When integrase inhibitor-treated cultures were analyzed by flow cytometry, mCherry producing cells were routinely reduced from an average of about 12% to less than 2% of the total (Figure 3). The use of the viability stain in this experiment assured that this was not due to the death of the infected and drug-treated cells. We then analyzed protein preparations from parallel cultures of these cells by Western blotting. MK-2048 treatment achieved >80% inhibition of mCherry production in these experiments. For the Western analysis, we initially purified populations of viable mCherry positive and mCherry negative cells using FACS analysis. However, because of the laborious difficulty in obtaining sufficient quantities of protein from the sorted cells, we resorted to comparing proteins of whole cell cultures obtained either in the presence or in the absence of integrase inhibitor. This latter approach yielded results similar to those obtained with proteins from sorted cells (Figure 4(a,b)). Thus, even for the DHIV3-infected whole culture preparations, in the absence of integrase inhibitor, where the majority of protein (>80%) came from the mCherry negative cells, m-Cherry protein was still readily detectable. All protein preparations were compared to an equivalent amount of protein from Control cells. Control cell protein preparations were from PMA activated THP-1 cells not exposed to DHIV3.

As anticipated, we found protein preparations from DHIV3-mCherry cultures, in the absence of integrase inhibitor, to contain high levels of mCherry protein (Figure 4(b)). In addition, these same protein preparations contained p24 and p24 precursor proteins (Figure 4 (c)). The p24 antibody we used is known to bind both p24 and Gag precursor proteins [10,3536]. As expected, p24 protein was also detectable in protein from the MK-2048 treated culture. In the literature, this observation has been attributed to residual p24 protein lingering from the initial infection [10]. We obtained support for this interpretation by taking a 48 hour time point (Figure 4(d)). The extended time point showed p24

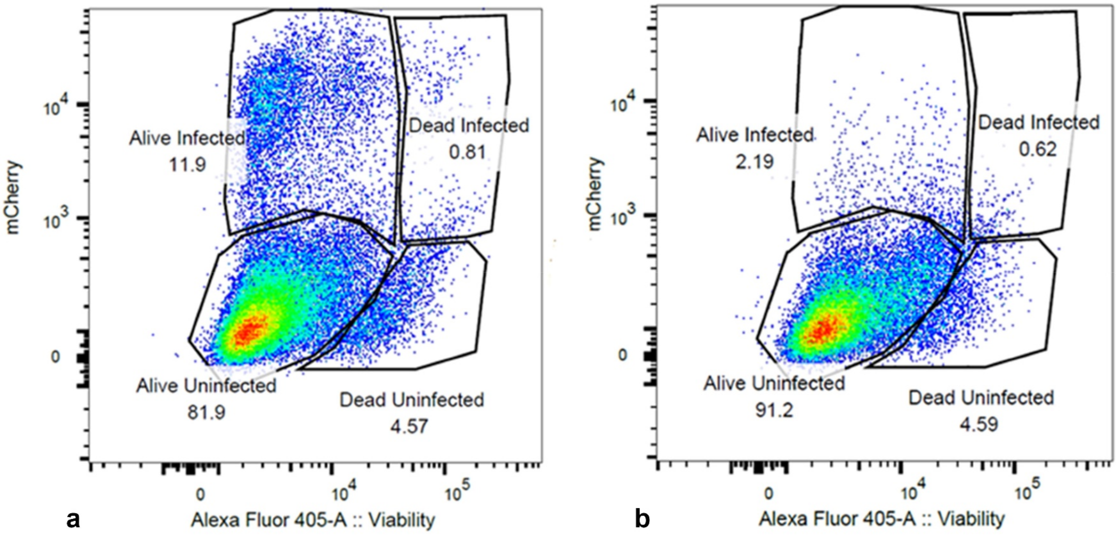

Figure 3. Integrase-inhibitor treatment selectively reduces mCherry positive cells. Panel **A**) flow cytometry analysis of DHIV3-mCherry infected THP-1 cells, versus viability stain. Abscissa shows viability stain intensity, ordinate shows mCherry intensity. Infected (Provirus), mCherry-producing cells account for approximately 12% of the cell population. Panel **B**) Same as A except with the addition of 25 nM MK-2048 integrase inhibitor at time of infection. Integrase inhibitor effectively reduces number of mCherry producing cells, without decreasing cell viability.

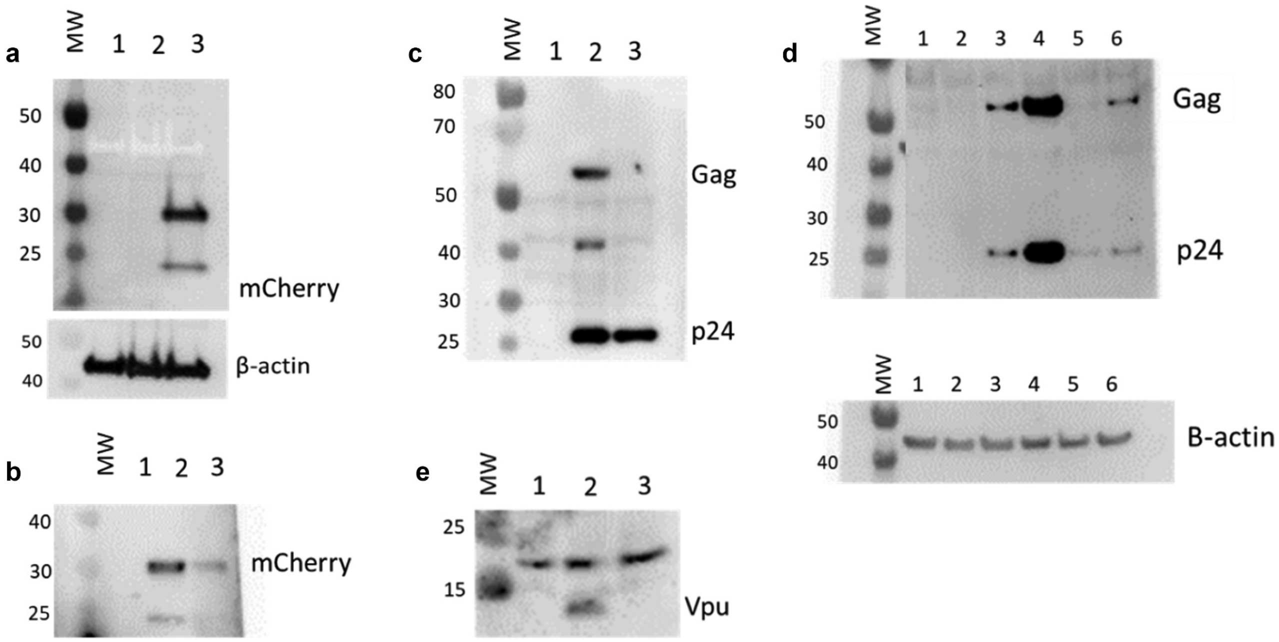

Figure 4. Effect of integrase inhibitor on mCherry, p24, Gag and Vpu protein production in cultures containing DHIV3-mCherry infected cells. MW, molecular weight markers. **A)** Cells infected with DHIV3-mCherry were purified by FACS sorting based on their expression of mCherry fluorescence. Lane 1, Protein from Control cells; Lane 2, Protein from PIC/Bystander cells; Lane 3, Protein from Provirus cells. Antibody used was goat anti-mCherry, developed with HRP linked anti-goat secondary. mCherry protein was only detectable in sorted Provirus cells. **B-E)** Lane 1, Control cell protein; Lane 2, protein from DHIV3 infected culture; Lane 3, protein from DHIV3 infected cultures treated with integrase-inhibitor (25 nM MK-2048) as shown above in Figure 3. **B)** Lane 2, mCherry protein was readily detectable in protein from cultures containing Provirus cells, with anti-mCherry antibody used in A. Lane 3, a small amount of mCherry signal was detected in MK-2048 treated cultures. **C)** Lane 2, p24 and Gag precursor proteins visualized with p24 antibody used above in Fig. S-6, and HRP linked anti-mouse secondary antibody. The p24 band in lane 3 is residual from infection as reported in the literature [10]. The presence of precursor proteins in lane 2 shows p24 synthesis in cultures containing Provirus cells. **D)** Lanes 1 and 2, Control cell protein at 24 and 48 hrs respectively; lanes 3 and 4, protein from DHIV3 infected culture at 24 and 48 hrs respectively; lanes 5 and 6, protein from DHIV3 infected cultures treated with integrase inhibitor (as above) at 24 and 48 hrs respectively. At 24 hrs post infection, we only found both p24 and precursor Gag proteins in the protein samples from DHIV3 infected cells in the absence of integrase inhibitor. At 48 hrs post-infection, in the absence of integrase inhibitor, the amounts of detectable p24 and Gag proteins were dramatically increased from levels at 24 hrs post infection. As seen initially (Panel C), some p24 protein was detectable in integrase inhibitor-treated cultures at 24 hrs post infection, however, Gag is not detectable at this time. At 48 hrs post infection in the integrase inhibitor treated cultures, some Gag protein does become detectable, reflecting production in cells that escaped complete integrase inhibition. This is in agreement with our flow cytometry analysis that showed suppressed, but still detectable numbers of mCherry positive cells in the integrase inhibitor treated cultures (Figure 3). The Gag precursor proteins only appear in the integrase inhibitor treated culture proteins 48 hrs after treatment. All antibodies, sources and dilutions are provided in Methods. **E)** Lane 2, Vpu detected with rabbit antibody, visualized using HRP linked anti-rabbit secondary. The resolution of the image is slightly compromised due to the small size of Vpu protein.

precursor proteins only appearing in the integrase inhibitor cultures 48 hours after treatment. In contrast, p24 and Gag precursor proteins were increasingly detectable in cell proteins from 24- and 48-hour integrase competent infections. Thus, p24 detected in DHIV3 infected cells was likely due to residual p24 from the original infection. We found both p24 and precursor p24 proteins only in the protein samples from cells without integrase inhibitor.

We tried antibodies for all the other major HIV-1 proteins, although not exhaustively, to test the correlation of protein expression with the detection of the transcript. One antibody that yielded a clear result was the polyclonal antibody against Vpu. In this case, we see a result very similar to that obtained with mCherry and p24, in that protein was clearly detected only in samples from infected cell cultures not treated with integrase inhibitor (Figure 4(e)). Vpu was not detectable in protein from the integrase inhibitor-treated cells. It is an interesting side note that Vpu has been associated with inhibition of NFkB promoted transcription [36,37].

Final confirmation of the PIC versus Provirus status of our semi-supervised cluster cell populations was obtained using real-time polymerase chain reaction (real-time PCR) analysis [38,39]. DNA from the respective cell treatment groups described above (Control, DHIV3 infected, DHIV3 infected with integrase inhibitor) was isolated and analyzed by PCR using multiple sets of primers. For detection of integrated proviral DNA, a set of primers [38] was used to amplify from random nascent human genomic Alu sequences to an internal HIV LTR sequence. This initial amplification step was followed by a second PCR amplification step using nested primers, which would only amplify discrete DNA products that contain integrated provirus [38]. Integrated provirus was detected in much higher abundance from DNA of DHIV3infected cells in the absence of integrase inhibitor (Fig. S-7 A, B, C). Very small amounts of integrated provirus DNA were detectable in MK2048 treated DHIV3-mCherry infected cell DNA, when using higher amounts (200 ng) of starting DNA. This is in close agreement with Flow cytometry results (Figure 3) and Western analysis (Figure 4) indicating small amounts of proviral DHIV3-mCherry DNA in our integrase inhibitor-treated cultures.

To demonstrate unequivocally that our PIC cluster cells do indeed contain PIC HIV cDNA, we used the primers of Brussels and Sonigo [39], which are internal to the HIV-1 LTR sequence. These primers were oriented in a way so as to detect only circular 2-LTR PIC HIV-1 DNA by bridging the 2-LTR junction [38,39]. We found that we required 2 rounds of PCR amplification to obtain the predicted 2-LTR amplicon, suggesting that this particular PIC species is in low abundance in our model cells (Fig. S-7D, E, F). We then tested this conclusion using bracketing primers (see Methods) to generate a product to contain the predicted amplicon of Brussels and Sonigo [37], and then followed with a round of amplification using the previous primers to generate a nested 2-LTR junction product. This approach also generated equal amounts of the predicted amplicon product from DHIV3-mCherry infected culture DNA, whether in the presence of integrase inhibitor or not. Thus, the circular 2-LTR PIC DNA was detectable in PIC and Provirus cells. The 2-LTR PIC cDNA PCR product was not detected in Control cell DNA.

To confirm that total PIC cDNA is in relatively high abundance in our DHIV3-mCherry infected cultures, we adapted the previous primers to amplify total DHIV3 LTR DNA. With these primers, similarly high levels of total PIC cDNA was detectable in DHIV3 infected cells, whether in the presence of integrase inhibitor or not (Fig S-7 G, H, I). This confirmed that both PIC and Provirus cells contain PIC cDNA.

In summary, the Provirus and PIC status assigned to our cell clusters were confirmed by real-time PCR. The Western blot analysis confirmed our flow cytometry observations that mCherry producing cells were also producing p24. Both approaches show that the mCherry cells are selectively suppressed by integrase inhibitor treatment and lead to the conclusion that only Provirus cluster cells are making mCherry, p24 or Vpu proteins from the DHIV3 transcripts. Conversely, transcripts detected in PIC cells, due to PIC transcription, do not lead to detectable mCherry, p24, or Vpu synthesis.

## scRNA-seq analysis of integrase inhibitor-treated DHIV3 infected cultures shows suppression of the Provirus cluster

DHIV3 infected THP-1 cells were treated with 25 nM MK-2048 at the time of infection, using our established protocol, and analyzed by scRNA-seq. In this experiment, the integrase inhibitor blocked ~87.5% of mCherry production by flow cytometry analysis of a parallel culture (Figure 3). The UMAP image of DHIV3 transcript features is shown in Figure 5(a). Transcriptome clustering using varying nearest neighbor K-values (K10, K50, K90, and K130) yielded 3 to 7 clusters, depending on the K-value applied (Fig. S-8). Regardless of the specified K-value, we did not detect a Provirus cluster in MK-2048 cultures, as was observed with integrase competent infections (Figure 5(c)). Again, the distribution of HIV-1 transcript containing PIC cells throughout the semi-supervised integrase inhibitor cell cluster was not affected by cell cycle, percent mitochondrial gene expression or numbers of genes detected per cell (Fig. S-9). Thus, the scRNA seq result confirms that integrase inhibitor-treatment selectively suppresses cells in the Provirus cluster, agreeing with results obtained by Western blot and PCR analysis, and confirms that the DHIV3 mRNA detected in the PIC cluster cells is due to transcription of PIC complexes.

## Hallmark and REACTOME analyses indicate mitosis-associated pathways are upregulated in Provirus cluster cells whereas viral restriction and interferon-associated pathways are upregulated in PIC cluster cells

Differential gene expression (DGE) in Provirus versus PIC/Bystander cluster transcriptomes was analyzed using GSEA with Hallmark or REACTOME

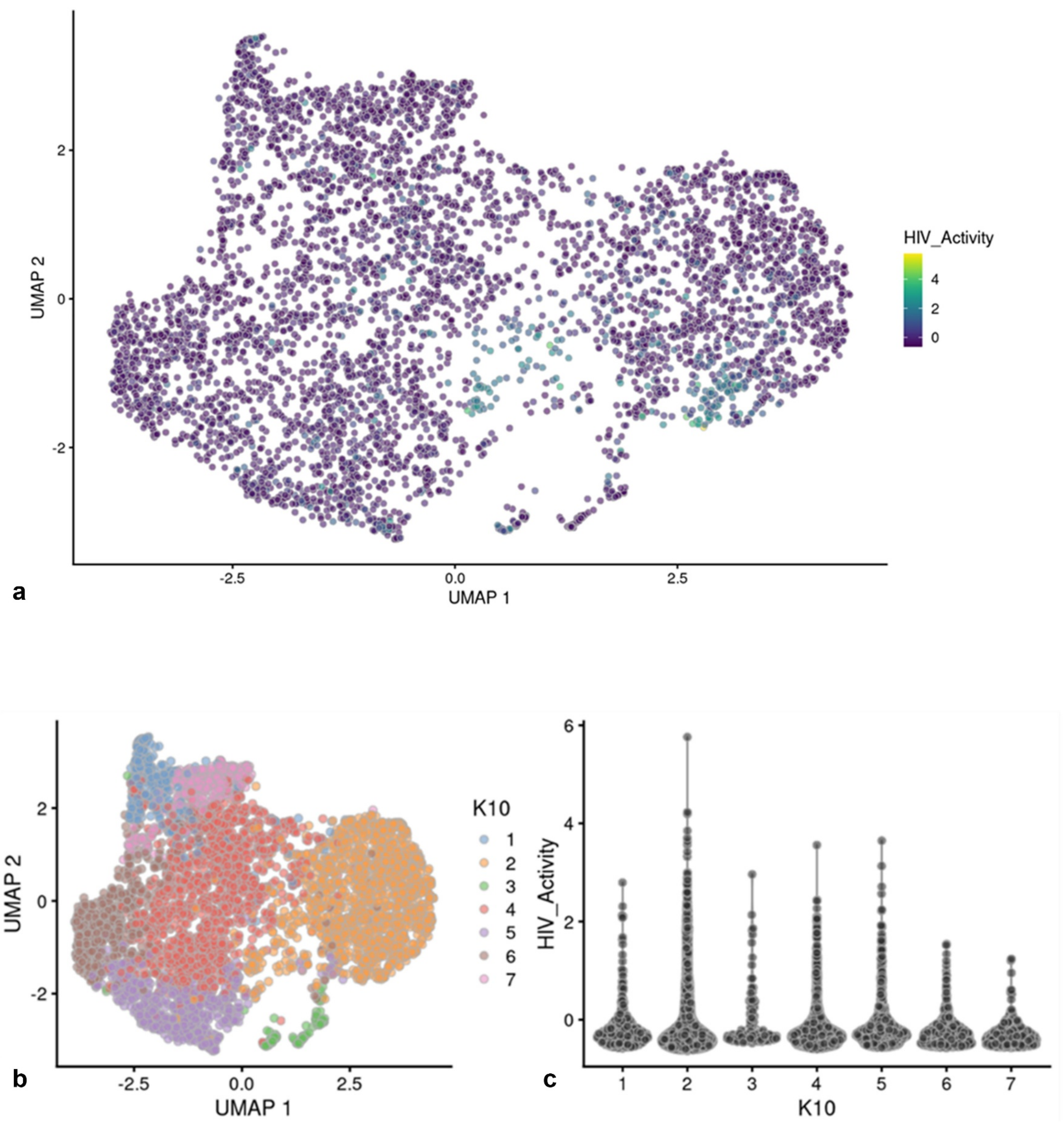

**Figure 5.** UMAP analysis of integrase-inhibitor treated DHIV3-mCherry infected THP-1 cells. Experiment performed as shown in Figure 3, with 25 nM MK-2048 added at the time of DHIV3 addition. Data were analyzed identically to data shown in Figure 2. Panel **A)** Feature plot showing the distribution of cells containing HIV-1 transcript, generated as described. Panel **B)** K10 unsupervised clustering generated 7 clusters (Scran's buildSSNGraph using the PCA as input). HIV-1 transcripts were distributed equally throughout all of them. No cluster corresponding to the "Provirus" cluster detected in Figure 2 was detected, regardless of K value used (see Fig. S-8). These data agree with the concept that integrase inhibitors selectively target and reduce the number of Provirus cluster cells.

gene lists (Appendix I), the pairwise T-Tests function from Scran was used to determine the statistically significant differentially expressed genes between groups. Table 1 presents the statistically significant Hallmark results. By comparing the DGE between the Provirus and PIC/Bystander cell clusters using Hallmark and REACTOME tools, it was clear that Provirus cluster cells were differentiated from PIC/Bystander cluster cells in several key ways. In general, the transcriptomes of cells in the Provirus cluster were characterized by gene transcripts associated with cell replication, whereas the transcriptomes of cells in the PIC/Bystander cluster were characterized by pathways associated with immune-response and interferon signaling. In the Provirus cells, the analyses clearly showed upregulation of cell replication, oxidative phosphorylation, protein synthesis and E2F family targeted pathways. On the other hand, NFkB, AP1, interferon responsive, and immune response pathways are relatively upregulated in the PIC/

Table 1. **Hallmark analysis of gene pathways up or down regulated detected in Provirus versus PIC/Bystander cluster cells (respectively)**. GSEA Hallmark analysis (fgsea R package) of metabolic pathways negatively or positively regulated ($p < 0.1$) in the Provirus cluster transcriptome versus the PIC/Bystander cluster transcriptome. Pathways up-regulated in Provirus cells include E2F, Myc targets, G2-M checkpoint, spermatogenesis and oxidative phosphorylation. Down-regulated pathways identified are more numerous, but notably included TNFα signaling via NFkB, inflammatory response genes, apoptosis and interferon γ response. The pairwise T-Tests function from Scran was used to determine the significant of genes between groups. The significant DGE subsets were used for all comparisons.

| Pathway | pval | padj | ES | NES | nMore Extreme | Size | Leading Edge (representative) | Enriched |
|---|---|---|---|---|---|---|---|---|
| Tnfa Signaling Via Nfkb | 0.00033 | 0.001399 | −0.74028 | −2.31254 | 0 | 187 | NINJ1, SAT1, IER3, IL1B, NFKBIA | negative |
| Complement | 0.000316 | 0.001399 | −0.67706 | −2.07107 | 0 | 160 | CTSL, CTSD, CTSB, LIPA, TIMP1 | negative |
| Inflammatory Response | 0.000318 | 0.001399 | −0.66804 | −2.05591 | 0 | 166 | C5AR1, IL1B, NFKBIA, TIMP1, CDKN1A, | negative |
| Coagulation | 0.000286 | 0.001399 | −0.70045 | −2.00402 | 0 | 96 | MMP9, CTSB, TIMP1, DUSP6, C3 | negative |
| Cholesterol Homeostasis | 0.000272 | 0.001399 | −0.71023 | −1.94144 | 0 | 70 | FABP5, SQLE, LPL, ATF5, S100A11, | negative |
| Epithelial Mesenchymal Transition | 0.000307 | 0.001399 | −0.63938 | −1.93759 | 0 | 147 | SAT1, DAB2, VIM, TIMP1, CXCL8 | negative |
| Hypoxia | 0.000318 | 0.001399 | −0.6227 | −1.91128 | 0 | 164 | IER3, PLIN2, S100A4, CDKN1A, PPP1R15A, | negative |
| Mtorc1 Signaling | 0.000336 | 0.001399 | −0.54085 | −1.69467 | 0 | 195 | SQLE, INSIG1, CALR, CDKN1A, PPP1R15A, | negative |
| UV Response Up | 0.0012 | 0.004614 | −0.52872 | −1.59832 | 3 | 141 | EIF5, NFKBIA, PPIF, ATP6V1F, SOD2, | negative |
| Apoptosis | 0.001493 | 0.00533 | −0.52273 | −1.57706 | 4 | 139 | SAT1, IER3, IL1B, LMNA, TIMP1 | negative |
| Kras Signaling Up | 0.002443 | 0.008142 | −0.50409 | −1.52704 | 7 | 146 | MMP9, IL1B, MAFB, G0S2, PPP1R15A, | negative |
| Il6 Jak Stat3 Signaling | 0.00326 | 0.010187 | −0.59782 | −1.63624 | 11 | 71 | IL1B, CD36, TNFRSF1B, IFNGR2, HMOX1 | negative |
| P53 Pathway | 0.003922 | 0.01032 | −0.47746 | −1.48489 | 11 | 180 | CTSD, NINJ1, SAT1, IER3, S100A4 | negative |
| Il2 Stat5 Signaling | 0.003552 | 0.01032 | −0.47887 | −1.48033 | 10 | 173 | PLIN2, COL6A1, TNFRSF1B, SNX9, KLF6 | negative |
| Xenobiotic Metabolism | 0.008534 | 0.021335 | −0.4758 | −1.43988 | 27 | 145 | NINJ1, TDO2, APOE, CD36, PGD | negative |
| Pi3k Akt Mtor Signaling | 0.009983 | 0.022688 | −0.52419 | −1.49267 | 34 | 93 | CALR, CDKN1A, SQSTM1, RPS6KA1, VAV3 | negative |
| Apical Junction | 0.009855 | 0.022688 | −0.47146 | −1.43125 | 31 | 150 | MMP9, INSIG1, RAC2, ZYX, CD276 | negative |
| Angiogenesis | 0.011132 | 0.023192 | −0.72619 | −1.6345 | 46 | 24 | LPL, TIMP1, S100A4, SPP1, THBD | negative |
| Reactive Oxigen Species Pathway | 0.010978 | 0.023192 | −0.61972 | −1.58502 | 41 | 47 | FTL, SOD2, NQO1, JUNB, MBP | negative |
| Interferon Gamma Response | 0.015349 | 0.030699 | −0.43862 | −1.36348 | 46 | 179 | NFKBIA, SOD2, CDKN1A, LY6E,, SPPL2A | negative |
| Myogenesis | 0.017438 | 0.033535 | −0.45597 | −1.37958 | 57 | 142 | CDKN1A, CD36, GSN, SPHK1, COL6A2 | negative |
| Protein Secretion | 0.022533 | 0.041727 | −0.49591 | −1.41213 | 78 | 93 | CD63, ATP6V1H, ABCA1, ARF1, BNIP3 | negative |
| Androgen Response | 0.023975 | 0.042813 | −0.50467 | −1.42398 | 85 | 85 | SAT1, INSIG1, SGK1, CCND1, B2M | negative |
| TGF Beta Signaling | 0.026772 | 0.046158 | −0.57256 | −1.47921 | 101 | 50 | PPP1R15A, IFNGR2, JUNB, RAB31, FKBP1A | negative |
| UV Response Down | 0.047167 | 0.078612 | −0.44009 | −1.30127 | 158 | 122 | DAB2, INSIG1, MGLL, RND3, SDC2 | negative |
| Allograft Rejection | 0.061941 | 0.099906 | −0.41295 | −1.25319 | 200 | 149 | MMP9, IL1B, TIMP1, IFNGR2, ITGB2 | negative |
| E2f Targets | 0.000142 | 0.001399 | 0.738487 | 2.126021 | 0 | 199 | STMN1, CDKN2C, SMC4, H2AFZ, CKS1B | positive |
| Myc Targets V1 | 0.000141 | 0.001399 | 0.650748 | 1.875111 | 0 | 200 | H2AFZ, TYMS, DUT, RPLP0, EEF1B2 | positive |
| G2-M Checkpoint | 0.000142 | 0.001399 | 0.633325 | 1.818855 | 0 | 195 | STMN1, CDKN2C, SMC4, H2AFZ, HMGN2 | positive |
| Spermatogenesis | 0.000156 | 0.001399 | 0.692643 | 1.799937 | 0 | 83 | CDKN3, RPL39L, PEBP1, GFI1, CCNB2 | positive |
| Oxidative Phosphorylation | 0.003883 | 0.01032 | 0.511797 | 1.459765 | 26 | 183 | LDHB, MPC1, UQCRH, COX8A, SLC25A5, | positive |

Bystander cluster cell transcriptomes. Intuitively, this makes sense, but it runs contrary to established literature that associates E2F transcription factors with decreased viral production [24,25]. This finding would not be obvious without the use of single-cell analysis. In the absence of single-cell analysis, the Provirus cluster's differentially expressed gene transcripts would have been swamped out by the 90% of mRNA obtained from PIC and Bystander cells. This led us to hypothesize that the cell's transcriptome background reflects the regulation of virus protein production to some extent, and that E2F transcription contributes to the favorable environment.

Using the UMAP Feature plots shown in Figure S-10, we visualized the distribution of cells containing some of the most statistically significant differentially expressed transcripts in the Provirus cell transcriptomes compared to the PIC/Bystander cell transcriptomes. The GSEA lists of the differentially expressed genes are provided in Supplementary Appendix I. Unlike the random distribution of PIC cells throughout the PIC/Bystander cluster, the distribution of these differentially expressed genes throughout the PIC/Bystander cluster is often clustered. We interpret this indicate the lack of impact PIC transcription has on the overall background PIC cell overall transcriptome (and vice versa).

## Biological repeat experiments confirm observations

To confirm the conclusions obtained from the analyses presented above, an independent biological repeat experiment was conducted. The repeat experiments captured over 4,500 cells, with an average greater than 3,000 genes per transcriptome. The control cell cultures again yielded a slightly larger number of cells captured with more than 3,000 genes transcripts per cell. The biological repeat experiments, HIVreplicate2 and WT2 (Control experiment number 2) were also conducted with parallel cultures that were analyzed by flow cytometry. Flow cytometry indicated 2.6% mCherry-positive cells in the HIVreplicate2 experiment. Figure 6 and S-11 show UMAP analysis of the biological repeat experiments HIVreplicate1 and HIVreplicate2. Figure S-12 shows unsupervised

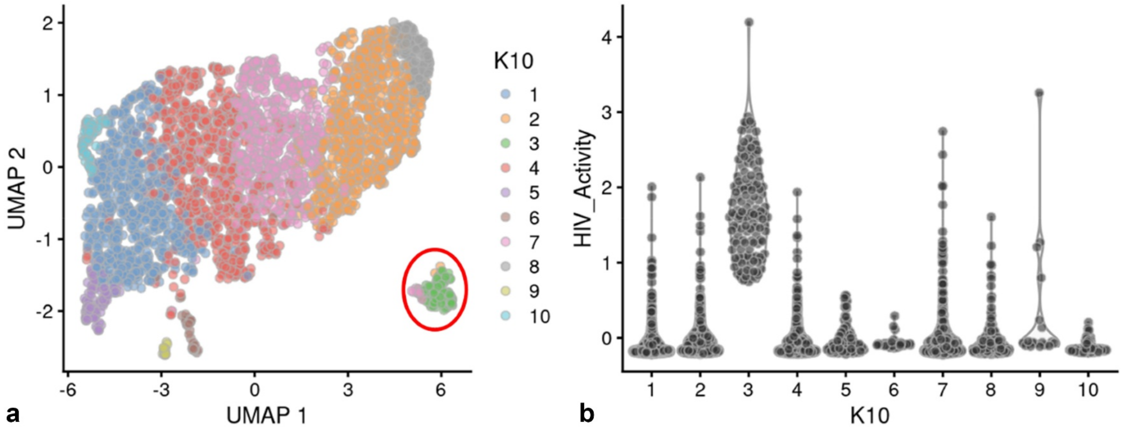

**Figure 6.** Unsupervised clustering of HIVrepeat2. Panel **A**) shows unsupervised clustering obtained at K equals 10. Panel **B**) Violin plot of HIV-1 transctipts/cell in the 10 clusters identified at K10 (Scran's buildSSNGraph using the PCA as input). PIC cells with detectable HIV-1 transcripts were distributed throughout clusters 1, 2 and 4–10. Cluster 3 contained 135 of the 227 cells in the semi-supervised Provirus cluster (circled in red).

clustering of HIVreplicate2 using K nearest neighbor values from 10 to 130.

Several comparisons were made to confirm that transcriptomes from Provirus and PIC/Bystander clusters in the repeat experiments were identical. In the first comparison, differentially expressed genes (positive or negative) were identified between the Provirus and PIC/Bystander clusters in the respective biological repeats. The log2 fold changes from these gene sets were then compared to test if the differences between the transcriptomes of Provirus and PIC/Bystander cells in the repeat experiments was consistent (Fig. S-13). Because there were almost 4 times as many Provirus cluster cells in the HIVreplicate1 experiment, the detected differentially expressed genes were significantly increased in the HIVreplicate1 case, compared to the HIVreplicate2 case. Nevertheless, the two gene sets were positively correlated. Correspondingly, GSEA with Hallmark and REACTOME gene sets from HIVreplicate2 identified many of the same pathways as differentially regulated as did HIVreplicate1 (Appendix I).

To obtain additional statistical certainty for our conclusion that the semi-supervised cluster transcriptomes obtained for HIVreplicate1 were repeated in HIVreplicate2, we compared log2 fold change values from the Provirus or PIC/Bystander clusters in HIV infected cells to the log2 fold change values from the Control (WT) PMA-treated THP-1 cultures. This 8-way comparison (shown in Fig. S-14) provided statistical certainty that DGE sets from the Provirus cluster and PIC/Bystander cluster gene sets from the biological replicates were not different. The replicate Provirus and PIC/Bystander gene sets have a generally strong concordance amongst themselves and there is a modest to strong non-zero mean trend in logFC among genes that changed in at least one of the contrasts between replicates (FDR 5%). In every comparison, a significant positive correlation was obtained from the commonly detected, significantly differentially expressed genes of Provirus or PIC/Bystander clusters in the two biological repeats when compared to the Control samples. The weakest correlations were observed in comparing PIC/Bystander to Control cell DEGs, especially Control experiment 2 (probably because there is more commonality between genes detected in PIC/Bystander cells and Control cells than there is between Provirus cluster cells and Control cells); nevertheless the data between HIVreplicate1 and HIVreplicate2 were statistically concordant. Therefore, the transcriptome data obtained from the two biological repeat experiments were not different. In other words, statistically identical representative transcriptomes for Provirus and PIC/Bystander clusters were obtained in independent biological repeat experiments.

As was the case for HIVreplicate1, a clear Provirus cluster was detectable in the HIVreplicate2 (Figure 6, S-11). However, because the level of Provirus infection in HIVreplicate2 was lower than in HIVreplicate1, the frequency of DHIV3 transcript detection in the Provirus cluster cells was proportionately lower, while the absolute number of detectable PIC and Bystander cells was relatively higher. Again, consistent with the observation for HIVreplicate1, PIC cells were randomly distributed throughout the Bystander cluster.

## M1 and M2 marker gene transcript expression is detectable in PIC/Bystander and Provirus cell clusters

We have used HIV-1 infection of PMA activated THP-1 cells to model infection of macrophages [40,41]. This model was selected to provide consistent induction responses in repeated experiments in order to provide statistical certainty for our conclusions. PMA activated THP-1 cells are not dividing and are, essentially, M0 macrophage cells [40–42], the type of macrophage most susceptible to HIV-1 infection [41]. In our protocol, THP-1 cells were incubated with 32 nM PMA for 24 hours, at which time they are morphologically differentiated, adherent, and flattened with rare mitoses. While THP-1 cells are known to de-differentiate over 72 hours when PMA is removed [42], in our protocol, following DHIV-1 addition for infection, the cells remain in 16 nM PMA, well within literature ranges of THP-1 cell activation by PMA [40–46].

While the PMA-activated THP-1 cells are not dividing [42], they can be further differentiated into M1 or M2 macrophage-like cells, by LPS and IFN-γ or IL-4 and IL-13, respectively [40]. This differentiation response resembles the M1 or M2 differentiation in primary macrophages when analyzed by transcriptome changes [40,46–48]. However, THP-1 induction responsiveness in terms of gene expression has been reported to be lower and not completely representative of induction responses obtained with naive primary macrophage cultures [40,49,50].

To determine if our Provirus or PIC cell transcriptomes were consistent with differentiation to M1 or M2 [40,46–48], and to determine if the M1 or M2 linked transcriptome changes could correlate with the Provirus versus PIC/Bystander transcriptome differences identified above, gene transcripts associated with M1 or M2 differentiation were compared between Provirus cells, PIC cells, and Control cells. Figure 7 shows the dotplot of these results. The overall expression pattern of M0, M1, or M2 marker genes in Provirus cells was not found to differ from Control cells, indicating no change in THP-1 differentiation state with viral integration. Thus, the distinction between Provirus and Control cell cluster transcriptomes was not due to the differentiation of Provirus cells toward an M1 or M2 phenotype. PIC and Bystander cells, however, were found to have increases in the level and fraction of cells expressing MAFB (M0) and IL1B, HLA-DRB1, and CD68 (M1) differentiation marker genes when compared to Control or Provirus cells. While the degree of a THP-1 cell's differentiation toward M1 or M2 phenotypes, relative to the expression of these marker transcripts, is a vague relationship, the differences described here may be associated with some of the transcriptomic differences identified between the PIC/Bystander cluster and Control or Provirus clusters.

## PIC and Provirus cells express all DHIV3 genes at equivalent levels, although higher numbers of PIC cells detectably express some transcripts

We were interested to know if early HIV-1 gene transcripts (tat, nef, or rev) predominated in PIC cell transcriptomes, versus later transcripts in the Provirus cells. We used Feature plots to determine the distribution of selected HIV gene transcripts expressed in individual cells. It was found that cells expressing the respective early or later gene transcripts were distributed throughout the image (Figure 8(a)).

We then used violin plots to compare the load of individual DHIV3 gene transcripts in Provirus cluster cells to PIC cells (Figure 8(b)). The Provirus cluster (from Figure 2) contained transcriptomes of 372 cells, the PIC/Bystander

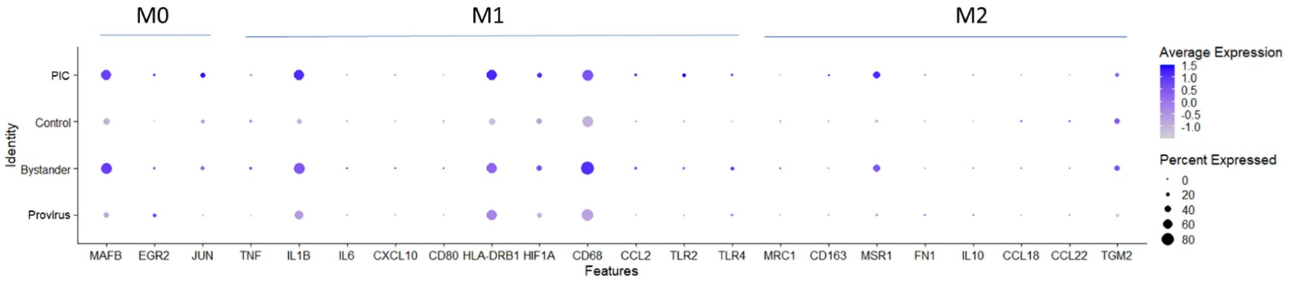

Figure 7. Dotplot of genes associated with M0, M1 and M2 differentiation states. The expression of M0, M1 or M2 marker genes in Provirus cells was not found to differ from those found in Control cells. The overall gene expression pattern did not change appreciably with HIV integration indicating that there was not a change in the differentiated THP-1 state with the viral infection. Minor changes are observed in the PIC and Bystander cells. These conditions were found to have higher overall levels of expression of the MAFB (M0) and IL1B, HLA-DRB1, and CD68 (M1) differentiation marker genes. This analysis shows relative upregulation M0 and M1 markers transcripts in PIC and Bystander cells when compared to Control or Provirus cells.

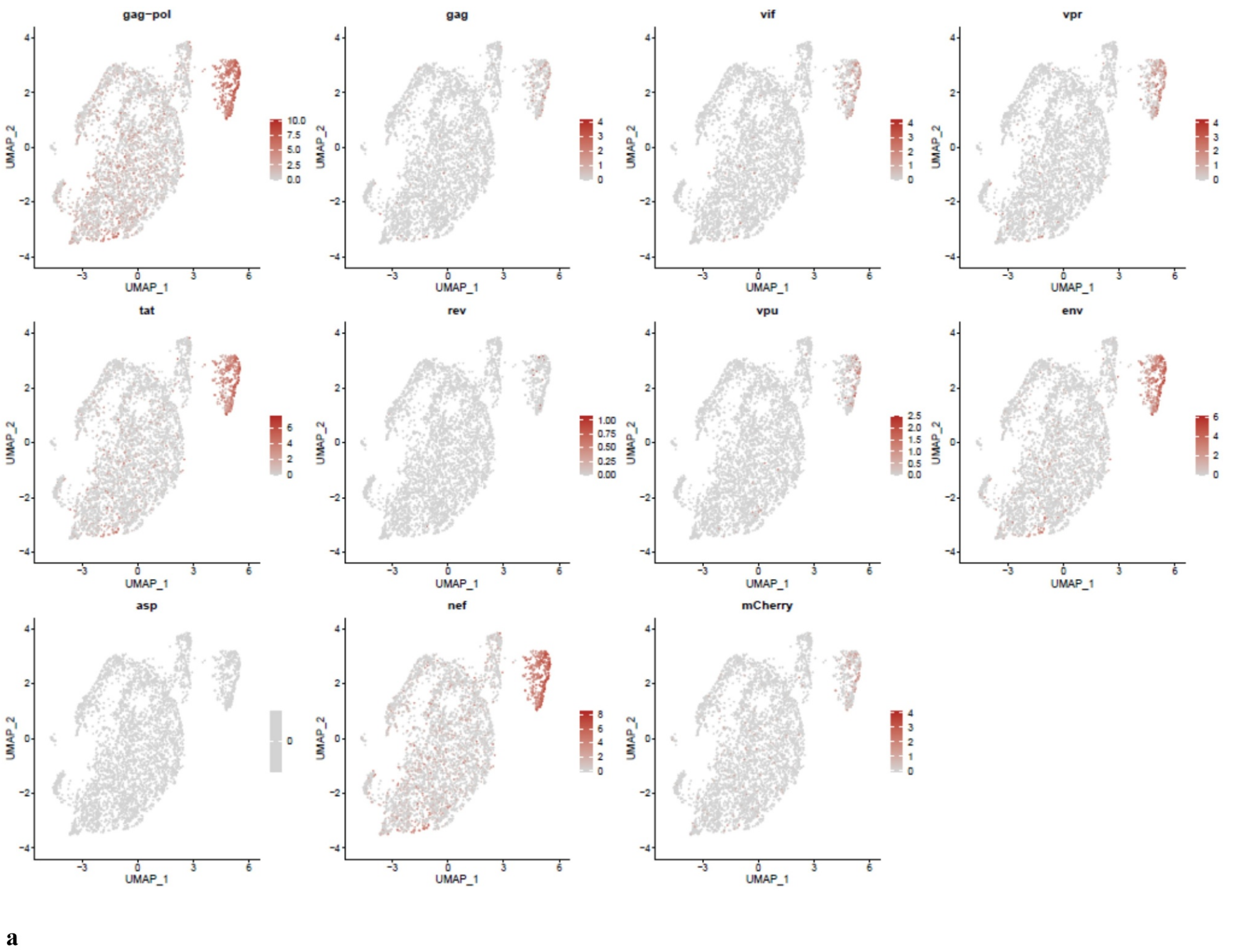

**Figure 8. The Distribution of HIV-1 transcripts throughout Provirus and PIC/Bystander clusters**. Panel **A**) Feature plot showing the distribution of cells from UMAP in Figure 2 that contain detectable DHIV3-mCherry transcripts. As described above, these UMAP projections were made with Seurat's FeaturePlot function. They are colored by the expression of individual genes (UMAP projection colored by walktrap, normalized log2 values). ASP is a negative control, bacterial gene transcript sequence. **B**) Violin plots of DHIV3-mCherry transcript/cell in cells from the Provirus and PIC clusters showing transcript level and cell number. The provirus cluster contained transcriptomes of 371 cells, the number of PIC cells in the PIC/Bystander cluster was 569 cells, thus the Provirus/PIC cell number ratio was 0.65. The plots were made with Seurat's VlnPot function. They show normalized log2 transcript levels. Two patterns of transcript distribution are evident. The first pattern is seen with gag-pol, tat, env, and nef, in which relatively high numbers of cells in the PIC/Bystander cluster express the transcripts, with the transcripts being detected in fewer numbers of Provirus cluster cells. The second pattern is seen with gag, vif, vpr, rev, vpu, and mCherry, in which relatively equal absolute numbers of cells in the Provirus and PIC/Bystander clusters are detected with the transcript sequences, remembering that there are more PIC cells than Provirus cells. The relative transcript loads per PIC cell versus the Provirus cells overlap. Negative control sequence (asp) shows no distribution.

cluster contained 569 cells, thus the ratio of Provirus to PIC cells was ~0.65. This visualization was much more informative and yielded a more nuanced understanding. Some transcripts, such as gag-pol, tat, env and nef were detected in far more cells in the PIC cluster. Experiment HIVreplicate1 suggested that the level of expression of this set of genes was significantly elevated in Provirus versus PIC cells, but this observation did not repeat in experiment HIVreplicate2 (Figure 9). In contrast, gag, vif, vpr, rev, vpu, and mCherry, were clearly detectable, but in fewer numbers of Provirus and PIC cells. Cells containing these transcripts were comparably prevalent in the two groups. The level of expression of this set of genes was comparable between Provirus and PIC cell groups in both experiments. In conclusion, all DHIV-1 transcripts were easily detectable in both Provirus cells and PIC cells. Furthermore, there is a clear overlap in the amount of HIV-1 gene transcripts detected in Provirus and PIC cells.

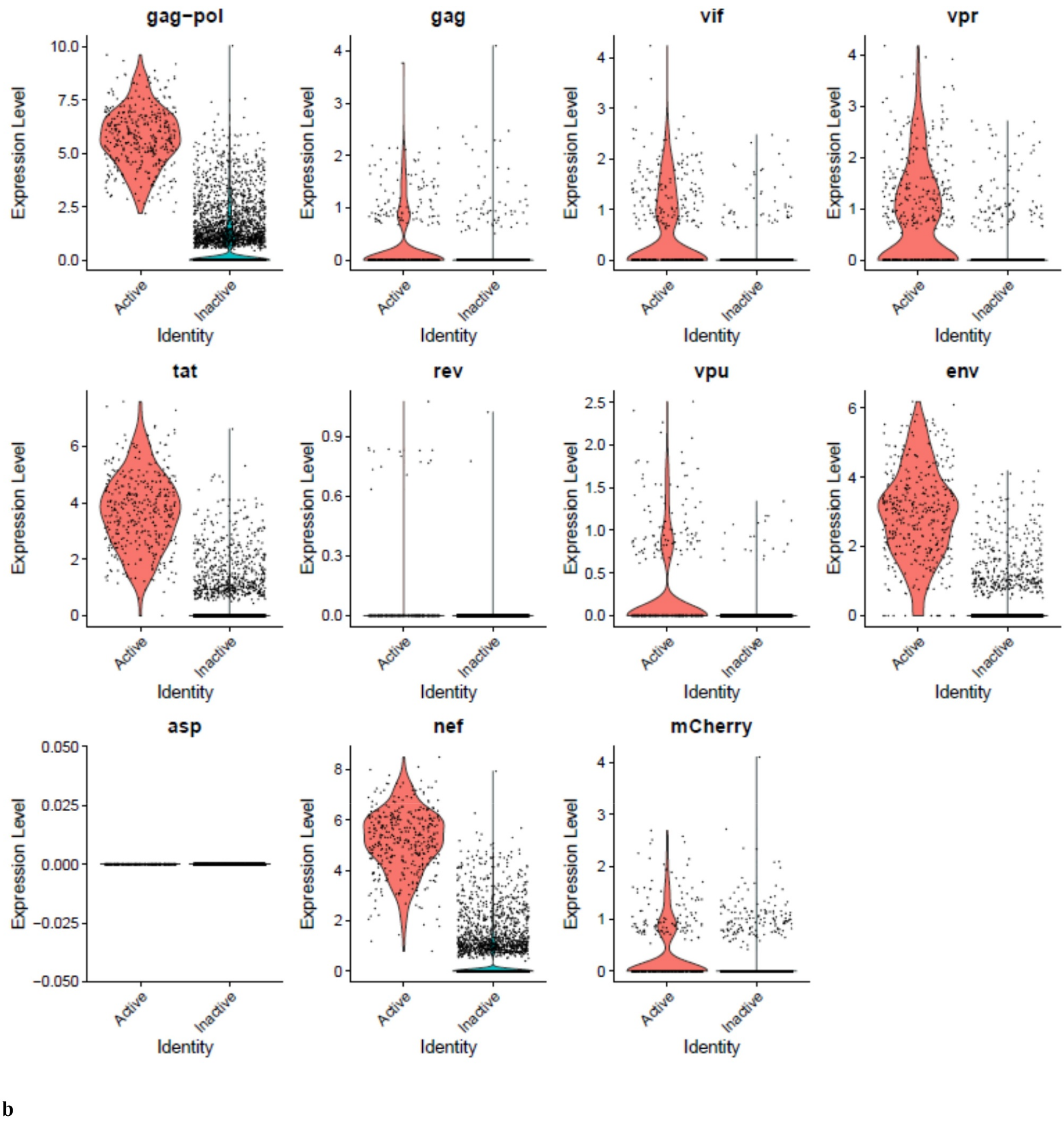

**Figure 8.** (Continued).

Unsupervised clustering of HIVreplicate2 cells (Figure 6, S-12) generated 10 clusters at a K nearest neighbor values of 10. Clusters 1, 2 and 4–10 contained most of the cells in the semi-supervised PIC/Bystander cluster. Cluster 3 contained 135 of the 227 cells in the semi-supervised Provirus cluster (circled in red, Figure 6). The transcriptome representing these Provirus cells was compared to the combined transcriptomes of the remaining clusters to generate the violin plots in Figure 9. The patterns of transcription observed in HIVreplicate1 were confirmed in this experiment (Figure 9, S-15). The transcripts of gag, vif, vpr, vpu, and mCherry were detectable in PIC cells at frequencies and levels of expression similar to those observed in the Provirus cells. In contrast, even correcting for the Provirus/PIC ratio of 0.17, the transcripts of gag-pol, tat, env, and nef were again detectable in proportionally higher number of PIC cells (Figure 9). Again, there was overlap in the numbers of HIV-1 gene transcripts detected in Provirus and PIC cells. No cells expressing rev were detected in the Provirus cluster in this

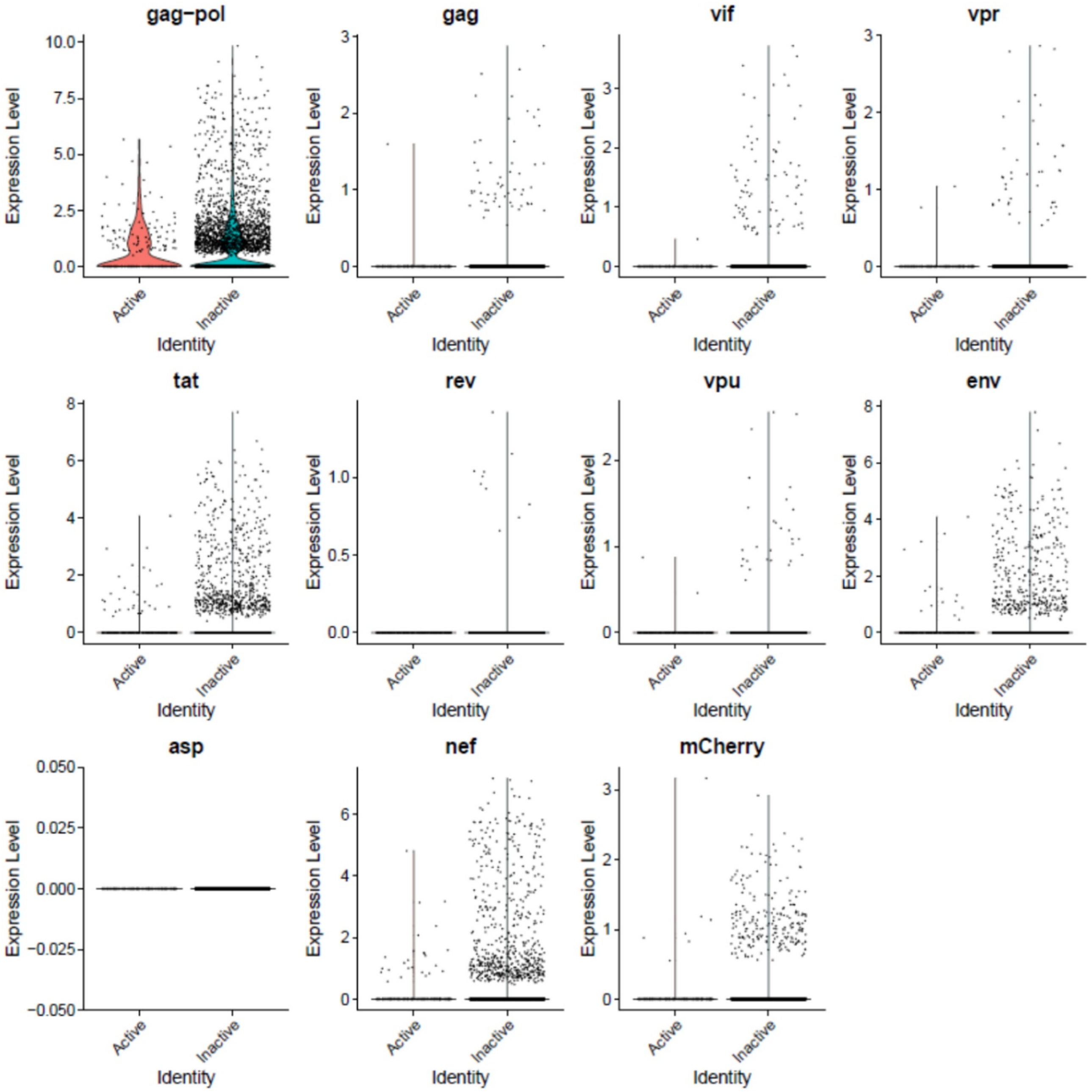

**Figure 9. The Distribution of HIV-1 transcripts throughout Provirus and PIC/Bystander clusters in HIVrepeat2**. Violin plots of DHIV3-mCherry transcript/cell in cells from the Provirus and PIC clusters showing transcript level and cell number. As described above, these were made with Seurat's VlnPot function. They show normalized log2 transcript levels. The two patterns of transcript distribution observed in HIVrepeat1 are evident. The first pattern is seen with gag-pol, tat, env and nef, in which higher numbers of cells in the PIC/Bystander cluster detectably express the transcripts. The second pattern is seen with gag, vif, vpr, vpu, and mCherry, in which fewer Provirus or PIC cluster cells are detected expressing the transcripts, but those cells expressing the transcripts are doing so at slightly higher average levels of transcripts per cell. It is difficult to compare transcript loads in the Provirus cluster cells to the results in HIVrepeat1 (Figure 8) due to the lower number of Provirus cells detected in this HIVrepeat2 experiment. In this experiment, the ratio of Provirus cells to PIC cells was 0.17. Nevertheless, the relative patterns observed in HIVrepeat1 are observed here. Following Seurat QC, no Provirus cells expressing rev were detected. Negative control sequence (asp) shows no distribution.

repeat, due either to the low efficiency of detecting this transcript in the 10X system, or because rev is expressed only at low levels in relatively few cells, or both.

During transcription of pro-virus, HIV-1 does not produce all transcripts in equal numbers or at the same time [51,52]. The specific processed gene transcripts produced initially in viral replication differ from transcripts produced later on. Furthermore, 10X Genomics scRNA-seq library production is known for significant numbers of dropouts, and cDNA copying of various gene transcripts during library construction varies in efficiency [53–54]. In addition, in using poly-T primers in the cDNA library construction, the 10X process introduces a 3' bias toward the detection of given sequences in a transcript [15]. Thus, it is not possible to make quantitative comparisons between the different transcripts using this approach. Nevertheless, the overarching take-away from this single cell analysis is that cells making fully spliced transcripts such as nef, tat, and rev are equally prevalent with cells making gag-pol and env transcripts (Figure 8,9). Furthermore, there appears to be two patterns of transcription. One pattern, observed with gag-pol, tat, env, and nef, is characterized by gene transcripts being more frequently detectable in PIC cells than Provirus cells. The other pattern, observed with gag, vif, vpr, rev, vpu, and mCherry, suggests relative equal frequency of transcription in Provirus and PIC cells.

### *Psupertime analysis indicates progression of cluster transcriptomes from Control to PIC/Bystander to Provirus*

To understand the transcriptome transitions needed to move from unexposed and uninfected "Control" cells to PIC/Bystander cells, and on to Provirus cluster cells, we performed a psupertime analysis [55–58] of the respective cell cluster transcriptomes. Psupertime is a supervised pseudotime [55] technique. It explicitly uses sequential condition labels as input. Psupertime is based on penalized ordinal logistic regression that places the cells in the ordering specified by the sequence of labels. This allows for targeted characterization of processes in single-cell RNA-seq data.

One thousand cells were randomly selected from each transcriptome cluster (combined Control, PIC/Bystander, and Provirus data sets, respectively) and their transcriptomes were combined for psupertime analysis. Imposition of Cluster identity yielded the image shown in Figure 10(a). The psupertime-type analysis showed closer similarity between Control and PIC/Bystander transcriptomes than between Control and Provirus transcriptomes, and closer similarity between PIC/Bystander and Provirus transcriptomes than between Control and Provirus. The GSEA list of the DGEs that contributed to this faux progression from Control to PIC/Bystander to Provirus is presented in Appendix II.

When we examined the expression of DHIV3 transcripts through the psupertime progression, the analysis showed no obvious preference for early gene transcription in cells belonging to the PIC/Bystander versus the Provirus clusters (Figure 10(b)).

When questioning which transcription factors were regulating the transcriptome transitions, we searched the contributory psupertime DGE transcripts for transcription factors. Many transcription factors that differed in expression in the contrasted transcriptomes were identified (Appendix II). However, this yielded a complicated picture and did not clarify which transcription factors

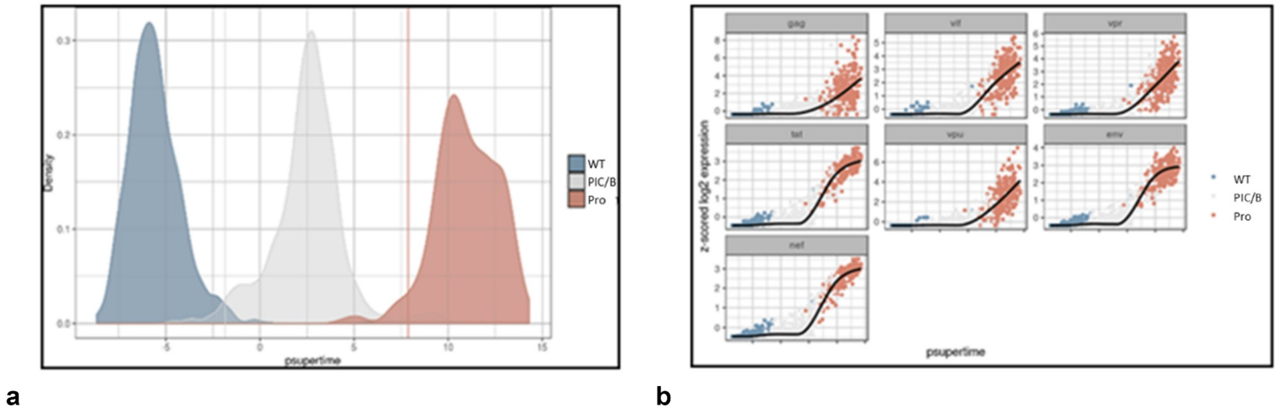

**Figure 10. Psupertime analysis of Control, PIC/Bystander, and Provirus cell transcriptomes**. Psupertime analysis is a supervised pseudotime approach that explicitly uses sequential labels as input. It uses a regression-based model that acknowledges the cell labels to identify genes relevant to the process. Panel **A**) one thousand randomly Control (WT), PIC/Bystander (PIC/B), and Provirus (Pro) cell transcriptomes were randomly selected and analyzed. Imposition of identity revealed a pseudo-evolution of Control to PIC/Bystander to Provirus cell transcriptomes. Panel **B**) distribution of HIV-1 transcripts through these clusters agrees with results shown in Figure 8, showing no bias toward early or later gene transcripts.

might be most important for controlling the transcriptome transitions from Control to PIC/Bystander to Provirus clusters. However, because the activity of most transcription factors is regulated by activation of proteins already present within the cell, and not at the transcription level, we speculated that Transcription Factor Targeting analysis might be more informative as to which transcription factors were key to cluster transitions.

### E2F, NF-kB and AP1 control phenotype transitions between PIC/Bystander cells and Provirus cells

We used Transcription Factor Targeting analysis to identify transcription factors that controlled DGEs in our Control, PIC/Bystander and provirus clusters. This analysis was consistent with the aforementioned Hallmark and REACTOME analyses. The E2F family of transcription factors predominate in regulating the Provirus cluster DGEs (Table 2). Twenty out of the 29 possible promoter-associated transcription factor interactions positively associated with the transition from the PIC/Bystander to the Provirus transcriptome identified with the E2F transcription factor family. Thus, E2F is associated with pathways that determine the phenotype of cells in the Provirus cluster. Conversely, 19 possible promoter-associated transcription factor interactions were negatively associated with DGEs reflecting the transition from PIC/Bystander to Provirus cells. Of these 19 possible promoter complexes, 8 different promoter interactions were identified to be associated with NFkB and AP1 transcription (Table 2) suggesting that downregulation of NFkB and AP1 also plays a key role in shaping the Provirus cluster cell transcriptome.

AP-1 and NFkB also appear to play roles in the maintenance of the PIC/Bystander cell transcriptome as well. Correspondingly, Transcription Factor Targeting identified differences between Control and PIC/Bystander cell transcriptomes. Simple exposure of activated THP-1 cells to DHIV3 was sufficient to decrease E2F signaling in PIC/Bystander cells, compared to Control cells, and to increase AP-1 and NFkB signaling (Appendix III). We presume this effect on the PIC/Bystander cells is through PAMP and/or interferon signaling, but we have not investigated this further. Again, these results are consistent with the results obtained with Hallmark and REACTOME analysis of the DGEs.

### Western blot analysis of Retinoblastoma protein phosphorylation shows an increase in Provirus cells

To confirm a role for E2F and NFkB in regulating the transcriptomes of Provirus and PIC/Bystander clusters (respectively), we sorted Provirus (mCherry positive cells) from PIC /Bystander (mCherry negative) cells using the FACS Canto. Rb phosphorylation is associated with activation of E2F promoter family transcription. We hypothesized Provirus cluster cells would exhibit retinoblastoma (Rb) phosphorylation consistent with E2F activation [27,59]. We used anti-T821 Phospho-Rb antibody. Phosphorylation of Rb at threonine-821 (T821) blocks pocket protein binding, including E2F family proteins, and activates E2F family promoter gene transcription [59]. Isolated Provirus cells had the greatest phosphorylation of Rb (pRb), compared to Control and PIC/Bystander cells (Figure 11). Interestingly, phosphorylation of Rb in PIC/Bystander cells was lower than that detected in Control cells (Figure 11).

These findings agreed closely with the Transcription Factor Targeting analysis, which showed that E2F promoter family transcription was predominant in Provirus cluster cells. It also agreed with the Transcription Factor Targeting analysis in that PIC/Bystander cells exhibited reduced Rb phosphorylation, and presumably E2F driven transcription, when compared to either Provirus cluster of Control cells. We did not find a difference in NFkB deactivation in the Provirus cells as would be implied by phospo-IkB S32, (Figure 11), and hypothesize that other mechanisms must account for the relative decrease in NFkB driven transcription in Provirus cluster cells.

### Cells transcribing from provirus are more likely to produce viral proteins upon second infection than PIC or Bystander cells

If Provirus cluster cells have already committed to the production of virus, represented by the switch of their background transcriptome to favor E2F transcription factor interactions, we hypothesized that they should be more efficient at producing virus upon second infection. We sequentially infected activated THP-1 cells with DHIV3-mCherry followed by an infection with DHIV3-GFP 24 hours later (DHIV3-mCherry infection at 0 hour and DHIV3-GFP infection at 24 hours). We found a higher percentage of cells positive for mCherry and GFP after 48 hours compared to GFP alone (Figure 12). At his time point, which was 24 hours after DHIV3-GFP infection, about half the mCherry positive cells were also GFP positive. Whereas less than one-quarter of the mCherry negative cells were expressing GFP protein. This trend continued out to 72 hours post DHIV3-mCherry infection, 48 hours after DHIV3-GFP addition, where about 60% of the mCherry cells were also GFP

**Table 2. Transcription Factor Targeting analysis of DGE contrasting PIC/Bystander and Provirus cells.** TFT analysis (GSEA with the fgsea R package and the C3 collection from msig) suggests that at least 3 transcription factor families control the transition from PIC/Bystander transcriptomes to Provirus cluster transcriptomes. These are E2F, NFkB and AP1 family promoter proteins. In particular, increased E2F regulated transcription appears to correspond with the transition to production of viral proteins. The pseudo-transition from Control to PIC/Bystander is characterized by a down regulation of E2F family regulated transcripts and up regulation of NFkB and AP1 regulated transcripts Appendix III. In comparing Provirus to PIC/Bystander transcriptomes, E2F family promoted transcripts are up regulated, while NFkB and AP1 transcription products are down regulated. Comparing Provirus to Control transcriptomes shows that overall Provirus cells have increased E2F regulated transcripts and decreased NFkB transcripts (with no significant change detected in AP1 regulation).

| Pathway | pval | padj | ES | NES | nMore Extreme | Size | Leading Edge (representative) | Enriched |
|---|---|---|---|---|---|---|---|---|
| TGTYNNNNNRGCARM. UNKNOWN | 0.000271 | 0.008414 | −0.67086 | −1.80812 | 0 | 66 | FRMD4A, ZEB2, CAMK1, BTG2, P2RX4 | negative |
| NFKB.Q6_01 | 0.000338 | 0.008809 | −0.553 | −1.71204 | 0 | 179 | MMP9, IL4I1, NFKBIA, MRPS6, DUSP6 | negative |
| AP1.Q6 | 0.00035 | 0.008809 | −0.51872 | −1.63047 | 0 | 202 | MMP9, LMNA, VIM, CDKN1A, LAPTM5 | negative |
| ELF1.Q6 | 0.000352 | 0.008809 | −0.512 | −1.61358 | 0 | 206 | LIMS1, TYROBP, SAT1, VIM, ARRB2 | negative |
| AP1.Q4_01 | 0.000349 | 0.008809 | −0.5116 | −1.61056 | 0 | 203 | MMP9, CD68, CDKN1A, PPP1R15A, FABP4 | negative |
| TCANNTGAY.SREBP1_01 | 0.000452 | 0.010535 | −0.47289 | −1.57952 | 0 | 386 | CTSD, TM4SF19, ATP6V1F, CALR, PSAP | negative |
| RGAGGAARY.PU1_Q6 | 0.000457 | 0.010535 | −0.42754 | −1.43088 | 0 | 393 | MMP9, TYROBP, IL4I1, VIM, PLD3 | negative |
| NFKAPPAB65.01 | 0.001017 | 0.021695 | −0.48751 | −1.51962 | 2 | 187 | MMP9, IER3, NFKBIA, SLAMF8, TNFRSF1B | negative |
| LXR.Q3 | 0.00106 | 0.021815 | −0.65715 | −1.73161 | 3 | 57 | MAFB, NFKBIA, SGK1, FKBP2, APOC1 | negative |
| CREL.01 | 0.001398 | 0.026582 | −0.4735 | −1.49185 | 3 | 205 | MMP9, IER3, NFKBIA, DUSP6, SLAMF8 | negative |
| AP1.Q6_01 | 0.001395 | 0.026582 | −0.46803 | −1.46954 | 3 | 200 | LMNA, SGK1, PPP1R15A, FABP4, SDCBP | negative |
| TGANNYRGCA.TCF11 MAFG_01 | 0.001431 | 0.026582 | −0.46304 | −1.46874 | 3 | 216 | MMP9, EIF5, SQSTM1, TPM3, RHOG | negative |
| CEBP.C | 0.001889 | 0.033994 | −0.51299 | −1.5463 | 5 | 141 | SAT1, NFKBIA, PTPN12, H3F3B, ALDOA | negative |
| BACH1.01 | 0.002099 | 0.036631 | −0.46449 | −1.46004 | 5 | 202 | HMGA1, LMNA, SGK1, CDKN1A, PPP1R15A | negative |
| AP1.C | 0.002449 | 0.041493 | −0.45856 | −1.4439 | 6 | 204 | MMP9, LMNA, CD68, PPP1R15A, FABP4 | negative |
| CCCNNGGGAR.OLF1_01 | 0.002967 | 0.046195 | −0.43587 | −1.39853 | 7 | 246 | IL4I1, ATF5, MTSS1, NFKBIA, LAS | negative |
| AP1.01 | 0.004175 | 0.06329 | −0.45533 | −1.43094 | 11 | 201 | LMNA, CDKN1A, VAT1, SQSTM1, EMP3 | negative |
| NFKB.Q6 | 0.004423 | 0.063825 | −0.45705 | −1.42695 | 12 | 189 | IL4I1, ATF5, NFKBIA, LASP1, SLAMF8 | negative |
| NRF2.Q4 | 0.004432 | 0.063825 | −0.45156 | −1.41242 | 12 | 191 | FRMD4A, SQSTM1, H3F3B, ALDOA, IDS | negative |
| E2F.Q3_01 | 0.000139 | 0.005046 | 0.659464 | 1.909187 | 0 | 208 | PCLAF, STMN1, H2AFZ, HMGN2, RPS19 | positive |
| E2F.03 | 0.000138 | 0.005046 | 0.650807 | 1.893047 | 0 | 219 | PCLAF, STMN1, H2AFZ, HMGN2, RPS20 | positive |
| E2F1.Q4_01 | 0.00014 | 0.005046 | 0.646185 | 1.865393 | 0 | 203 | PCLAF, STMN1, H2AFZ, HMGN2, RPS19 | positive |
| E2F.Q6_01 | 0.000139 | 0.005046 | 0.642509 | 1.863053 | 0 | 212 | PCLAF, STMN1, H2AFZ, HMGN2, RPS19 | positive |
| E2F.Q4_01 | 0.000139 | 0.005046 | 0.627736 | 1.818187 | 0 | 211 | PCLAF, STMN1, H2AFZ, HMGN2, RPS19 | positive |
| E2F.Q3 | 0.00014 | 0.005046 | 0.620013 | 1.787787 | 0 | 200 | STMN1, H2AFZ, HMGN2, RANBP1, PRKDC | positive |
| E2F.Q6 | 0.000139 | 0.005046 | 0.613171 | 1.774788 | 0 | 207 | PCLAF, STMN1, H2AFZ, HMGN2, RANBP1 | positive |
| E2F.Q4 | 0.000139 | 0.005046 | 0.610694 | 1.768827 | 0 | 211 | PCLAF, STMN1, H2AFZ, HMGN2, RANBP1 | positive |
| E2F1.Q6_01 | 0.000139 | 0.005046 | 0.581008 | 1.686153 | 0 | 215 | STMN1, HMGN2, RPS19, RANBP1 | positive |
| E2F1.Q6 | 0.000139 | 0.005046 | 0.579747 | 1.678268 | 0 | 209 | PCLAF, STMN1, H2AFZ, H2AFV, RPS20 | positive |
| E2F1DP1.01 | 0.000139 | 0.005046 | 0.568415 | 1.645244 | 0 | 207 | PCLAF, STMN1, H2AFZ, H2AFV, RPS20 | positive |
| E2F1DP2.01 | 0.000139 | 0.005046 | 0.568415 | 1.645244 | 0 | 207 | PCLAF, STMN1, H2AFZ, H2AFV, RPS20 | positive |
| E2F4DP2.01 | 0.000139 | 0.005046 | 0.568415 | 1.645244 | 0 | 207 | PCLAF, STMN1, H2AFZ, H2AFV, RPS20 | positive |
| E2F.02 | 0.000139 | 0.005046 | 0.568048 | 1.644182 | 0 | 207 | PCLAF, STMN1, H2AFZ, H2AFV, RPS20 | positive |
| E2F4DP1.01 | 0.000139 | 0.005046 | 0.562316 | 1.628381 | 0 | 210 | PCLAF, STMN1, H2AFZ, H2AFV, RPS20 | positive |
| E2F1DP1RB.01 | 0.00014 | 0.005046 | 0.561019 | 1.620205 | 0 | 204 | PCLAF, STMN1, H2AFZ, HMGN2, CBX5 | positive |
| E2F1.Q4 | 0.000277 | 0.008414 | 0.554186 | 1.611179 | 1 | 218 | STMN1, HMGN2, HMGB1, ZFP36L2, RANBP1 | positive |
| E2F1.Q3 | 0.000278 | 0.008414 | 0.551892 | 1.601657 | 1 | 215 | PCLAF, STMN1, H2AFZ, HMGN2, ATAD2 | positive |
| USF2.Q6 | 0.00056 | 0.012404 | 0.539697 | 1.558627 | 3 | 204 | STMN1, COMMD3, CDKN2C, HMGN2, REEP3 | positive |
| SMAD.Q6 | 0.002545 | 0.041876 | 0.519518 | 1.491178 | 17 | 190 | STMN1, CKS1B, BMP4, RPS14, CBX5 | positive |
| SGCGSSAAA. E2F1DP2_01 | 0.002916 | 0.046195 | 0.545308 | 1.522242 | 19 | 149 | PCLAF, H2AFZ, H2AFV, RPS20, RANBP1 | positive |
| CDP.02 | 0.004721 | 0.06633 | 0.614792 | 1.569087 | 29 | 73 | MEF2C, RPA3, PHACTR3, BHLHE22, PTMA | positive |
| PAX2.01 | 0.005566 | 0.074992 | 0.678893 | 1.597401 | 33 | 43 | HIST1H4C, HOXA10, ACTN4, MBNL1, JMJD1C | positive |
| E2F.01 | 0.005648 | 0.074992 | 0.647435 | 1.58611 | 34 | 56 | SMC4, RANBP1, PRKDC, DNMT1, RMI2 | positive |
| OCT1.02 | 0.005729 | 0.074992 | 0.542596 | 1.496063 | 38 | 132 | CDKN2C, HMGB2, RPS19, HOXA10, CPNE1 | positive |
| COMP1.01 | 0.006881 | 0.086168 | 0.594609 | 1.536162 | 43 | 80 | PCLAF, HMGB1, SKA2, HOXA10, CDK6 | positive |
| CRX.Q4 | 0.006754 | 0.086168 | 0.508209 | 1.442548 | 46 | 172 | CDKN2C, HMGN2, ZFP36L2, SATB1, RPA3 | positive |
| E2F.Q2 | 0.007399 | 0.090675 | 0.516606 | 1.446676 | 50 | 152 | STMN1, COMMD3, HMGN2, BMI1, UQCRH | positive |
| MEIS1AHOXA9.01 | 0.007979 | 0.095745 | 0.586922 | 1.521992 | 50 | 82 | CDKN2C, SKA2, SATB1, PDLIM1, HLX | positive |

positive, compared to about 40% GFP positive in mCherry negative cells. In repeat experiments and in experiments using primary macrophage and T-cell cultures (Fig. S-16) the results were the same. Provirus cluster, mCherry positive cells, were 2X to 5X more likely to make DHIV3-GFP protein upon second infection than PIC/Bystander cluster cells. Thus, cells already committed to making virus were more likely to make virus from a second infection than PIC/Bystander cells on the first infection.

## Discussion

Early research into the HIV-1 life cycle identified transcription of HIV-1 PIC cDNA in T-cells and macrophages [5–12]. The unintegrated viral DNA can take

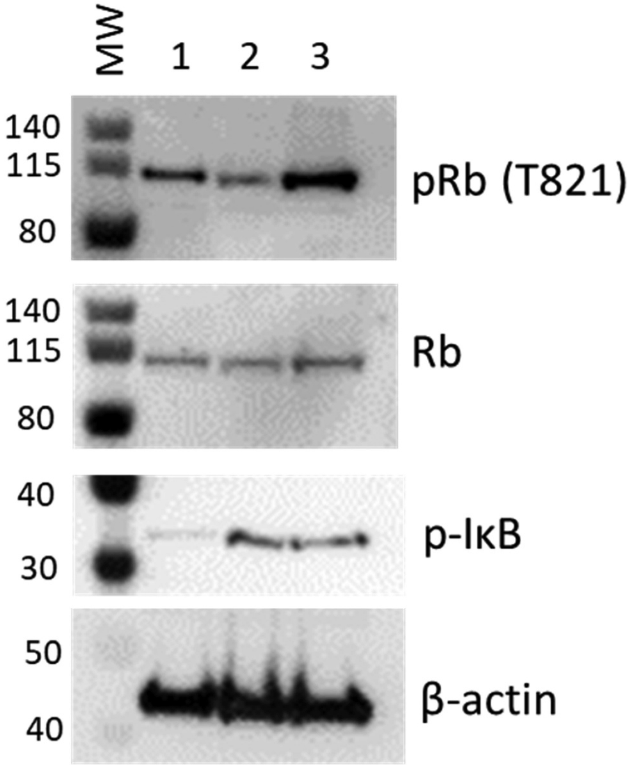

**Figure 11. Western blot analysis for phospho- Rb or IkB in protein from mCherry negative versus mCherry positive cells**. Cells infected with DHIV3-mCherry were purified by FACS sorting based on their expression of mCherry fluorescence. Lane 1, Protein from Control cells; Lane 2, Protein from PIC/Bystander cells; Lane 3, Protein from Provirus cells. Phospho-Rb (Phospho-T821 Rb antibody) was used to quantify Rb pocket phosphorylation, anti-Rb control antibody was used to quantify Rb protein levels relative to actin (visualized with beta-actin antibody). PIC/Bystander cells show the lowest level of Rb phosphorylation, Provirus show the highest, in close agreement with Transcription Factor Targeting results. Phospho-IkB S32 antibody was used to quantify activated IkB. Control cells show the lowest level of IkB phosphorylation, no difference was detectable between Provirus and PIC Cluster cells.

several forms, linear and the 1-LTR- and 2-LTR circles, with much of the transcription thought to emanate from 1-LTR circles [7,60]. In macrophages, it has been shown that HIV-1 PIC cDNA can persist and be actively transcribed for months. However, it is generally agreed that PIC HIV-1 transcription in macrophages does not routinely produce infectious virus [12]. The use of single-cell techniques has enabled us to quantify both the numbers of cells expressing given HIV transcripts in a mixed culture, and the relative transcript loads of each HIV-1 gene in each infected cell [13–15]. It also allows us to put the cells containing viral transcripts into the context of their background transcriptomes. This provides the opportunity to compare cells producing late viral proteins to those producing transcript but not late viral proteins to better understand the cellular metabolic background necessary for virus production.

We show that scRNA-seq can differentiate between macrophage cells that transcribe from PIC HIV-1 cDNA and macrophage cells that transcribe from integrated HIV-1 provirus. In proving this observation, we discovered that the synthesis of many HIV-1 proteins is only detectable in provirus cells, even though HIV-1 RNA transcripts can be detected equally in cells transcribing from provirus or PIC HIV-1 cDNA. Single-cell RNA sequencing can distinguish between the two cell types because the background transcriptomes vary dramatically. PIC cell transcriptomes are characterized by NFkB and AP-1 promoted transcription, while transcriptomes of cells transcribing from provirus are characterized by E2F family transcription products. We also find that the transcriptomes of PIC cells and Bystander cells are identical, suggesting that PIC cells are "oblivious", at the transcriptome level, to the fact that they are making HIV-1 transcripts. Furthermore, the presence of pathogen alters the transcriptome of the uninfected Bystander cells, so that they are readily distinguishable from true Control cells (cells not exposed to any pathogen). Thus, to understand the transcriptional changes caused by HIV-1 infection in these cell populations, it is not appropriate to compare bulk RNA sequencing data from HIV-1/host cell co-cultures to sequencing data from control cultures not exposed to pathogen. The correct comparisons to make are between cells transcribing from provirus and those transcribing PIC HIV-1 cDNA, or Bystander cells.

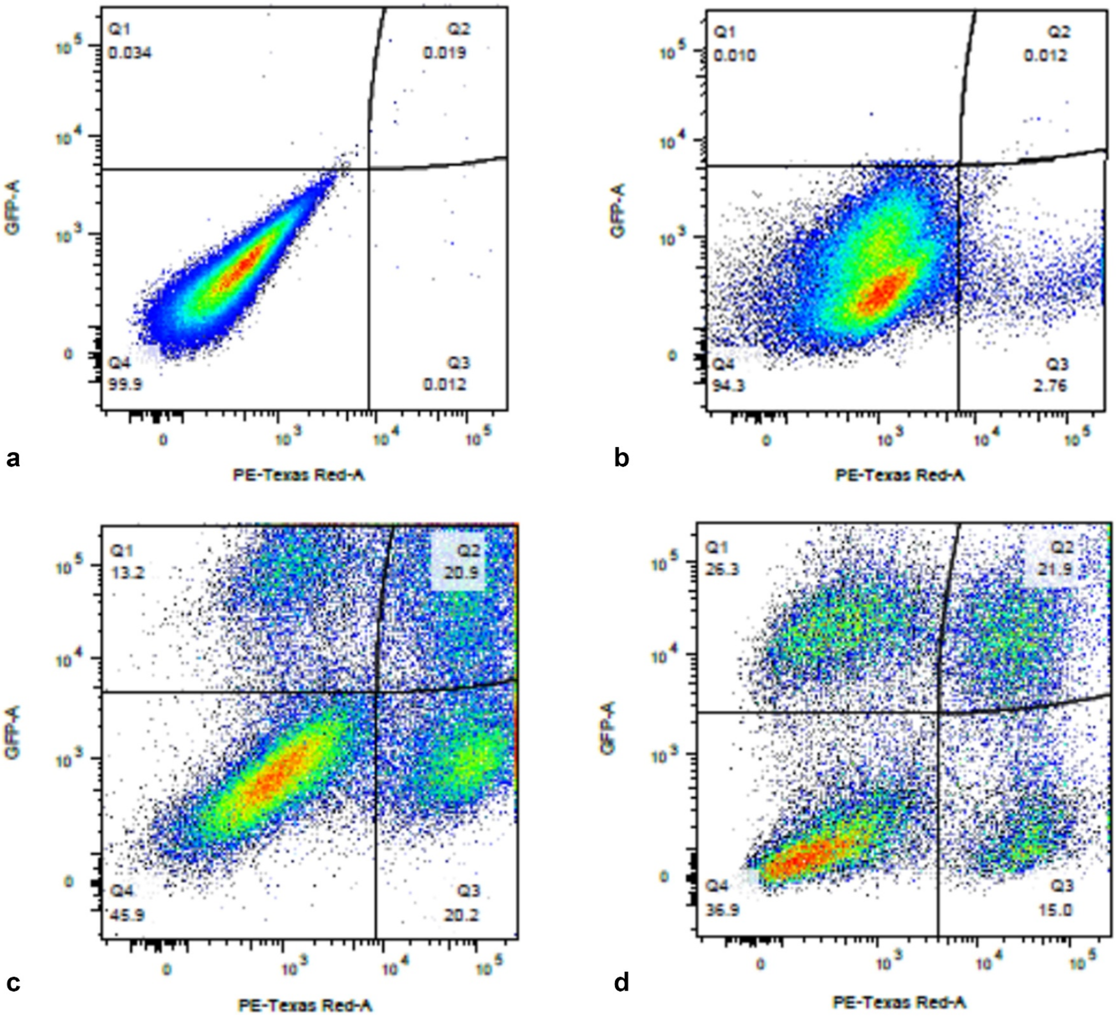

Figure 12. **Sequential infection of THP-1 cells with DHIV3-mCherry followed 24 hrs later with GFP DHIV3**. Abscissa, mCherry signal, Ordinate, GFP signal. Provirus cluster, mCherry positive, cells were 2 to 5 times more likely to make HIV-1 encoded GFP protein upon the second infection than PIC/Bystander cells upon the second infection. Panel **A**) time equal 0 hrs; addition of DHIV3-mCherry. Panel **B**) time equal 24 hrs; addition of DHIV3-GFP. Panel **C**) time equals 48 hrs after DHIV3-mCherry addition, 24 hrs after DHIV3-GFP addition. Panel **D**) time equals 72 hrs after DHIV3-mCherry addition, 48 hrs after DHIV3-GFP addition. The percentage of mCherry cells also producing GFP, compared to cells producing mCherry only, is always 2 to 5 times higher than the percentage of cells making only GFP, compared to those cells not producing either mCherry or GFP.

As reported by Marsh and Wu and colleagues [10], "transcription in the absence of integration is selective and skewed towards certain viral early genes such as nef and tat, with highly diminished rev and vif". In general, our single cell analysis agrees with this conclusion, but provides a more nuanced picture. In our data, compared to cells in the same co-culture transcribing from provirus, many more PIC cells are producing tat, nef, gag-pol, and env transcripts. However, the level of a given transcript load per cell is not detectably different from Provirus cluster cells. In contrast, the prevalence of PIC cells that make rev, gag, and accessory gene transcripts constitute a smaller fraction of the PIC cells. Nevertheless, those few PIC cells that are producing rev, gag and accessory gene transcripts are doing so at levels equivalent to Provirus cells. In Provirus cells, the numbers of cells detectably making fully spliced versus un-spliced transcripts are equal. The levels of these transcripts per cell overlaps with levels of transcripts produced by PIC cells. Thus, it is difficult to generalize that high overall transcript levels are a trigger for virus production.

Because DHIV3 is replication-deficient, we measure the production of HIV-1 proteins as a surrogate marker of virion production. Only the Provirus cells appear to be making p24, mCherry or Vpu protein. Using Western blot analysis, we only see the production of p24, mCherry and Vpu in purified Provirus cell protein, or protein samples containing Provirus cells. It is not clear whether our failure to detect the HIV proteins in preparations enriched for PIC/Bystander cells is because very few cells in this cluster were making the proteins (and thus below our level of detection), or if there was some restriction mechanism preventing protein production, or both. Single-cell protein analysis technology might be able to address this point. The robust detection of mCherry, p24, and Vpu in protein samples enriched for Provirus cells confirms that late protein synthesis is an attribute of the Provirus cells. Nef protein is reported in the literature to be detectable in protein from both Provirus and PIC cluster cells [10]. Nef protein is not made in cells infected with the DHIV3-mCherry, due to the mCherry sequence replacing the 5' portion of the nef gene, and we could not detect Nef protein from any DHIV3 infected cultures.

Integrated provirus transcription is required for virus production [12]. It is regulated by promoter elements in the HIV LTR. Thierry and coworkers have recently shown that PIC cDNA and provirus are differentially responsive to NFkB promotion [60]. As others have found, they report provirus transcription is enhanced by NFkB and AP1 binding. However, they find PIC cDNA transcription to be inhibited by NFkB activation. Our data adds a layer of complexity to this picture. We can show that early and late gene transcripts are detectable in Provirus cells, at levels overlapping with the respective transcript levels in PIC cells. We also show that, in cells that have transitioned from PIC/Bystander to Provirus, the background transcriptome reflects an overall down-regulation of NFkB and AP1 transcripts, and up-regulation by E2F family promoted transcripts. E2F is not a promoter in the HIV-1 LTR, and we do not suggest that E2F regulation of provirus transcription is the key to the PIC to Provirus transition. However, we do propose that an E2F family promoter-dominated transcriptome is required for virus production.

This proposition appears counter to literature, in which E2F is thought to suppress viral transcription [25,61]. Nevertheless, the pathways upregulated by E2F are those consistent with what one would intuitively anticipate a being required for viral production, and our transcription factor analysis suggests that E2F family promotors play an outsized role in Provirus transcription. E2F is an Rb-pocket binding protein, and phosphorylation of Rb at T821 is known to activate E2F transcription. Indeed, we show that levels of phospho-T821 Rb are higher in Provirus cells. A proposed role for E2F promoter family proteins in virus production is not new and is consistent with literature citing a role for Rb phosphorylation and E2F activation in HIV-1 linked tumorigenesis [62]. Consistent with this was our observation in our THP-1 system, and in primary cultures of human macrophage and T-cells, the Provirus cells are more likely than PIC cells to make virus upon second infection. If this interpretation is correct, then the study of genome-wide changes that accompany provirus integration and amplify E2F signaling might be key to understanding the switch in transcriptome necessary for viral protein production as PIC cells transition to Provirus cells.

Several recent scRNA-seq studies using donor macrophage or T-cell preparations have revealed complicated relationships between transcriptome and HIV-1 production. In general, these studies preselect for Provirus cells before amplification and report complicated and possibly stochastic, transcriptome heterogeneity in both latent and active HIV-1 infected cells [63–65]. In contrast, we focused on the distinction between GFP positive cells (transcribing from provirus) compared to GFP negative cells (either Bystander cells or those transcribing from PIC HIV-1 cDNA), in cell cultures containing both infected and uninfected cells. We found that a change in transcriptome background that includes E2F transcription accompanies Provirus integration, and this also accompanies the production of late viral proteins.

## Materials and methods

### Reagents

THP-1 cells, a monocytic cell line, were obtained from ATCC (Cat#TIB-202). HyClone™ RPMI 1640, kanamycin sulfate, Corning™ Accutase™ detachment solution and phorbol 12-myristate 13-acetate (PMA) were obtained from Fisher Scientific (cat. # SH30011.03, BP906-5, MT25058CI, and BP685-1, respectively). Fetal bovine serum was purchased from Atlanta Biologicals (cat. # S11150). BD Horizon™ Fixable Viability Stain 450 was obtained from BD Biosciences (cat. # 562,241).

### Cell culture

THP-1 cells were grown in RPMI 1640 supplemented with 10% FBS and KAN (50 μg/mL) at 37°C, 5% $CO_2$.

### Generation of DHIV3-mCherry

DHIV3-mCherry virus was generated by calcium phosphate transfection (25). In brief, HEK293FT cells were grown to 80% confluence. DHIV3-mCherry plasmid and VSV-g plasmid were mixed with calcium chloride (2.5 M) and HEPES buffered saline solution. The calcium phosphate-DNA suspension was added dropwise to the cells. Chloroquine (100 mM) was subsequently added and the HEK293FT cells were incubated overnight. The medium was replaced with fresh DMEM and incubated for an additional 48 hours. Supernatant was collected and filtered (0.45 μm). Optimal viral titers were determined by titrating the virus in THP-1 cells. A titer volume of 100 μL in 500 μL total (~5 x 106 TU/mL in THP-1s) was chosen due to its high DHIV infection and minimal effects on viability. Titer volume (up to 400 μL of 500 μL) was increased as infectivity fell off in stocks over time.

## *DHIV3-mCherry and THP-1 co-culture*

For the THP-1s, cells were preincubated overnight in PMA (20 ng/mL) at 500,000 cells/well to generate differentiated macrophages. Medium was replaced for THP-1 cells 1 hour before the addition of DHIV3-mCherry. Cells were incubated at 37°C for 24 hours followed by preparation for analysis by flow cytometry and cell imaging.

## *Flow cytometry*

Adherent THP-1 cells were incubated with Accutase™ for 15 minutes at 37°C. THP-1 cells were transferred to 5-mL tubes and washed with PBS. Cells were resuspended in BD Horizon™ Fixable Viability Stain 450 (0.25 μg/mL) and incubated at 4°C for 30 minutes. Cells were then fixed in 2% formaldehyde for 30 minutes at 4°C. Cells were analyzed using a FACS Canto. Percent infection by DHIV was quantified as a subset of the live population (FSC/V450/50-). Gates for infection were set according to the uninfected “mock” THP-1 cell controls. Three independent biological replicates were completed for all treatment conditions, each in triplicate wells per experiment. Population analysis was then done using FlowJoTM v10.7, to assess if infection levels and cell viability were consistent similar in all replicates. The Flow Cytometry figures are representative plots obtained from one of the replicates.

## *10X Genomics library construction and sequencing*

Two biological replicate cultures of (HIV+) THP-1 cells (HIVreplicate1 and HIVreplicate2) and (HIV-) THP-1 cells (hereafter referred to as Control, or WT1 and WT2 for wild type) were processed through the 10X Genomics Chromium Single Cell Controller with Single Cell Gene Expression 3’ Solution (v2 chemistry). Sequencing was done on an Illumina HiSeq 2500 instrument.

Cell suspensions were partitioned into an emulsion of nanoliter-sized droplets using a 10X Genomics Chromium Single Cell Controller and RNA sequencing libraries were constructed using the Chromium Single Cell 3’ Reagent Kit v2 (10X Genomics Cat#PN-120237). Briefly, droplets contained individual cells, reverse transcription reagents, and a gel bead loaded with poly(dT) primers that include a 16 base cell barcode and a 10 base unique molecular index. Lysis of the cells and gel bead enables priming and reverse transcription of poly-A RNA to generate barcoded cDNA molecules. Libraries were constructed by End Repair, A-Tailing, Adapter Ligation, and PCR amplification of the cDNA molecules. Purified cDNA libraries were qualified on an Agilent Technologies 2200 TapeStation using a D1000 ScreenTape assay (Agilent Cat#5067-5582 and Cat#5067-5583). The molarity of adapter-modified molecules was defined by quantitative PCR using the Kapa Biosystems Kapa Library Quant Kit (Kapa Biosystems Cat#KK4824).

HiSeq 125 Cycle Paired-End Sequencing v4: Sequencing libraries (25 pM) were chemically denatured and applied to an Illumina HiSeq v4 paired-end flow cell using an Illumina cBot. Hybridized molecules were clonally amplified and annealed to sequencing primers with reagents from an Illumina HiSeq PE Cluster Kit v4-cBot (PE-401-4001). Following the transfer of the flowcell to an Illumina HiSeq 2500 instrument (HCS v2.2.38 and RTA v1.18.61), either a 26 × 100 cycle or 125 cycle paired-end sequence run was performed using HiSeq SBS Kit v4 sequencing reagents (FC-401-4003). Basic html (notebook) files describing coding and QC data are provided in Appendix IV.

## Data analysis for UMAP *(Figure 1(b))* of HIVrepeat1

Raw FASTQ files from 10x Genomics were processed by 10x Genomics’ Cell Ranger software. Each library was processed with “cellranger count” pipeline with a common genomic reference made up of human and HIV genomes as well as mCherry. The human genomic reference was GRCh38 with gene annotation from Ensembl release 91, where only features with gene_biotype:protein_coding were kept. The HIV genome and annotation was acquired from NCBI genome (RefSeq

ID NC_001802.1). No warnings were issued by 10x Genomics regarding sequencing, alignment, or cell-based QC metrics; however, the samples could have been sequenced deeper as reflected in the sequencing saturation statistics.

In an attempt to recover those (perhaps lower quality) GEM partitions, the raw gene-barcode matrices from "cellranger count" (located in "outs/raw_gene_bc_matrices") was processed with the EmptyDrops algorithm (R package DropletUtils v1.2.2) to discriminate cells from background GEM partitions at a false discovery rate (FDR) of 0.001% [66]. GEM partitions with 2000 UMI counts or less were considered to be devoid of viable cells, while those with at least 10,000 UMI counts were automatically considered to be cells.

For each technical replicate, additional quality control measures were taken to filter out low-quality cells. Cell-based QC metrics were calculated with R package scater (v1.16.2) using the perCellQCMetrics function. Cells with extremely low UMI counts, extremely low gene counts, or an extremely high percentage of expression attributed to mitochondrial genes were also flagged as low quality. Extremeness in any of these three measures was determined by 3 median absolute deviations from the median with the scater is Outlier function. These cells suspected of being low quality were removed from downstream analysis with one exception in the HIV replicates; if cells exhibited above median HIV gene expression, they were not discarded as these were thought to hold potential value as examples of cells in which viral replication suppressed other gene expression (not observed). Further analysis of the HIV biological replicates showed there were remaining low quality cells as marked by unusual mitochondrial gene expression or low library size, which were removed to improve the signal-to-noise ratio. Specifically, HIV-infected cells with mitochondrial expression of 7.23% and above were removed as well as cells that have less than 3242 UMI counts. To ensure no cell type was discarded due to filtering, average gene expression was compared gene-wise between discarded and kept cells in scatter plots. There were no genes of interest that exhibited markedly different average gene expression between the discarded and kept cells, suggesting the filtering did not remove interesting subpopulations. After filtering low quality cells, technical replicates were combined into biological replicates HIVreplicate1, HIVreplicate2, wt1 and wt2.

For each biological replicate, cells were normalized [66] and scored for several important attributes. Each cell from HIV biological replicates was assigned a HIV activity score with Seurat's (v3.2.2) AddModuleScore function [58,67], where a high score indicates HIV gene expression was stronger in the cell relative to randomly selected genes of similar expression strength in the biological replicate. Cell cycle phases and scores were assigned cells with the cyclone method [56]. Cells were scored against a simulated doublet population of cells with scran's (v1.10.2) doubletCells function; those cells with extremely high doublet scores (5 median absolute deviations) were removed and remaining cells were re-normalized again with scran's quickCluster and scater's normalize methodology [57] for differences in sequencing depth between libraries.

## Data analysis for t-Sne insert *(Fig. S-3)*

10x Cell Ranger raw sequencing data was processed into UMI counts using the "mkfastq", "count", bioinformatics modules. Cell Ranger de-multiplexed cDNA libraries into FASTQ files with Illumina's bcl2fastq and aligned reads to a hybrid genomic reference composed of human (Ensemble GRCh38), HIV (NCBI ID: NC_001802.1), and mCherry genomic references with STAR aligner [68]. In addition to other QC metrics [69], CellRanger filtered cell barcodes and unique molecular identifiers (UMIs) in estimation of gene-cell UMI counts using only reads that mapped uniquely within the transcriptome. We specified an "expected cell number" of 3000 per library based on reported cell recovery rates.

The QC metrics reported by Cell Ranger indicated that our library construction was a success; the libraries averaged 97.9% valid cell barcodes, 60.6% of reads mapping to the transcriptome, and reported a median of 2402 genes detected per cell (mean of 15,262.2 genes per library). Only in HIV-infected samples did reads map to the HIV genome. Cell Ranger also evaluated dimension reduction, clustering, and differential gene expression analysis under default parameters. For further details of the Cell Ranger data processing and analysis pipeline, see https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/algorithms/overview.

Due to cost, this study was limited by low sequencing depth (mean of 41,433 reads per cell per library). This limitation was mitigated by removing genes with low sequencing coverage; specifically, a gene was filtered out if it did not have 1% of cells reporting at least 3 UMI; cells were filtered out if they did not have at least 200 genes with a UMI count. Cells with exceedingly high (top 2%) ribosomal and/or mitochondrial content were filtered out. To reduce mutliplets contaminating analysis, cells with the top 2.3% total UMI were removed (see 10x Genomics benchmarks).

### Dimension reduction, clustering, and differential expression

Highly variable genes were identified with scran's trendVar and decomposeVar. Specifically, loess smoothing was applied to the gene variance (dependent variable) and the mean gene expression (independent variable) after having corrected for the % mitochondrial expression and cell cycle effects on the cells. Genes with average expression below the first quartile were filtered out of consideration. Gene variance was decomposed into biological and technical components, where genes with variance above the mean trend (loess fit) were assumed to possess biological variation [70]. This process was repeated for the HIV biological replicates separately followed by scran's combineVar function applied to the combined HIV replicates to identify genes estimated to have positive biological variation and controlled with a false discovery rate of 0.05.

### Gene set enrichment analysis

A gene set enrichment analysis (GSEA) was conducted with R package fgsea (v1.8.0) [32]. The log2 fold change vector of strong-HIV vs. weak-HIV was evaluated for enrichment against 3 different collections of MSigDB gene sets: namely, Hallmark, REACTOME, and Transcription Factor Targets. The Benjamini-Hochberg false discovery rate (FDR) was controlled at 10%.

A DGE and GSEA analysis was conducted on HIV biological replicates in comparison to Control biological replicates. The Control (Wild Type) replicate expression data was subject to (nearly) identical quality control and identical pre-processing steps. As described above, quality control on HIV cells was subject to a greater degree of scrutiny.

To mitigate biases due to potential batch-specific variation, DGE and GSEA analyses contrasting Provirus -HIV and PIC/Bystander-HIV populations to Control cells leveraged consensus between various pairwise contrasts. For instance, in the HIV- Provirus vs. Control (WT) contrast, t-tests were evaluated in 4 distinct contrasts: (i) HIVreplicate1-Active vs. WT1; (ii) HIVreplicate1- Provirus vs. WT2; (iii) HIVreplicate2-Provirus vs. WT1; (iv) and HIVreplicate2- Provirus vs. WT2. The scran function combine Markers performed a meta-analysis across the 4 contrasts with the Simes method. The Simes meta-analysis tested whether any of the 4 contrasts manifest either a change a gene-wise expression for DGE analysis; that is to say the meta-analysis p-value encodes the evidence against the null hypothesis, which assumes the gene is not changed in any of the 4 comparisons. Similarly in the GSEA analysis, the log2 fold change statistics from the 4 comparisons were tested for enrichment of the 3 previously mentioned MSigDB gene set collections (Hallmark, REACTOME, Transcript Factor Targets). The results were also merged with the Simes meta-analysis. This same strategy for the HIV- Provirus vs. Control contrast was repeated in the HIV- PIC/Bystander vs. Control comparison.

There was interest in modeling the progression of infection from wild type to PIC/Bystander-HIV to Provirus -HIV clusters. Specifically, there was interest in identifying what genes exhibit a variable expression profile or non-constant trend when ordered from Control to PIC/Bystander-HIV to Provirus -HIV. Macnair and Claassen [71] developed a supervised psuedotime R package, called psupertime, that is tailored to this express purpose. In particular, a penalized logistic ordinal regression model was fit to the combined HIV and Control data. The input data was a subset of highly variable genes identified in the same manner as described above, but including Control data as well. The gene expression data had been normalized, log2 transformed, and followed by linear correction of effects due to percent mitochondrial expression and cell cycle phase. The model was able to clearly order cells that reflects the expected order of Control then PIC/Bystander-HIV then Provirus -HIV. The psupertime method also reports a small set of genes that strongly associate with the expected progression, which is based on the magnitude of the penalized coefficients in the logistic ordinal regression.

### Western protocol

To generate protein samples for the Western blot experiments, multiple (6–12) wells of differentiated THP-1 cells, in 12 well plates, were infected with DHIV-3. In the integrase inhibitor experiments, three independent wells were combined to generate samples from each of the three treatment conditions (control, DHIV infection, DHIV infection with integrase inhibitor). Proteins from these replicate wells were pooled for Western blotting. Two distinct experimental replicates were analyzed, the initial 24 hour infection experiment and then the second 24- and 48 hour repeat infections. These repeats used different passages of THP-1 cells and different batches of infectious DHIV-1. For the flow-sorted cells, 18 wells each of control and infected THP-1 cells were sorted to accumulate sufficient protein for the Western blots. Protein from the collected, sorted cell sample groups were pooled. While the cells used in the sorted cell experiment were from 1 passage of THP-1 cells, multiple batches of infectious DHIV-3 were used. The stored THP-1 cells were pelleted by centrifugation for 5 min at 10,000 x g. Precipitates were then washed twice in cold PBS. Afterward, cells were lysed using 50 mM Tris pH 7.4, 100 mM NaCl and 2 mM EDTA, complete® protease inhibitor, 2 mM NaF, 2 mM sodium orthovanadate and 10% SDS. Protein concentrations were determined using bovine serum albumin standard and Coomassie Plus Protein Reagent from Pierce Biotechnology (Rockford, IL). The amount of protein loaded for the sorted cell experiments was 7.5 μg/ lane, for the integrase inhibitor Western gels, 15 μg/ lane and separated using NuPAGE 4 – 12% Bis-Tris gradient gels (Invitrogen Life Sciences, Carlsbad, CA) and transferred to a PVDF membrane (Millipore, Billerica, MA). These were then blocked in 2% BSA in TBST for 20 min at room temperature, incubated overnight with primary antibodies at 4°C in blocking buffer solution and with the secondary antibody for 45 minutes at room temperature. Protein was detected using chemiluminescence and blots were visualized using a Protein Simple FluoroChem M system.

### Antibodies and fluors

P24 and Gag protein precursor production was detected with monoclonal mouse IgG-AG3.0, (NIH AIDS Res. Reagent Prog., Germantown, MD; Cat. # 4121), 1:500, and using an AlexaFlour 633 nm or 700 nm goat anti-mouse IgG (H + L) secondary antibody (Fisher Scientific, Pittsburgh, PA) for flow cytometry. Anti-HIV-1 NL4-3 VPU rabbit polyclonal, also from the NIH AIDS Res. Reagent Prog. (Cat. # 969), was used at 1:20 dilution for Western blots. Goat anti-mCherry, OriGene (TR150126), was used at a 1:5000 dilution for Western blots. ThermoFisher Scientific Anti-Rb (LF-MA0173, 32C8) was used at a 1:1000 dilution. Invitrogen antibodies: anti-pRB (T821, 710,314) was used at a 1:500 dilution, anti-pIκB (S32, 701,271) was used at a 1:500 dilution, and anti-B-actin (PA5-85,291) was used at a 1:5000 dilution.

Other antibodies obtained from the NIH AIDS Res. Reagent Prog., Germantown, MD, include anti-HIV-1 IIIB gp120 Polyclonal (Cat. # 57), anti-HIV-1 RF gp160 Polyclonal (HT7) (Cat. # 189), anti-HIV-1 Tat Polyclonal (Cat. # 705), anti-Nef Monoclonal (EH1) (Cat. # 3689), anti-HIV-1 Nef Polyclonal (Cat. # 2949), anti-HIV-1 Vpr 1–50 aa Polyclonal (Cat. # 11,836); anti-HIV-1 HXB2 Vif Polyclonal (Cat. #12,256), anti-HIV-1 HXB2 IN Polyclonal (Antigen 2) (Cat. # 12,877), anti-HIV-1 RT Monoclonal (MAb 21) (Cat. # 3483), anti-HIV-1 HXB2 RT Polyclonal (Antigen 2) (Cat. # 12,881), anti-HIV-1 Protease Polyclonal (Cat. # 13,564), anti-HIV-1 Rev Monoclonal (1G7) (Cat. # 7376), which were tested at various dilutions.

HRP-conjugated secondary antibodies were used to develop Western blots, anti-rabbit A0545 (1:5000), and anti-mouse A9044 (1:5000) from Sigman Chem. Co., and anti-goat 401,515 (1:10,000) from CalBiochem. Propidium Iodide was obtained from Molecular Probes, Eugene, OR. DAPI blue was obtained from Acros Organics, Geel, Belgium.

### Real-time PCR methods

For detection of integrated proviral DNA, DNA from control, DHIV3-mCherry infected, DHIV3-mCherry infected plus integrase inhibitor-treated THP1 cells was purified using Qiagen Blood and Tissue DNeasy kits. For the PCR experiments, we generated DNA from 2 replicate wells in a 12 well culture plate. Each RT PCR reaction was run with triplicate technical repeats. Every PCR experiment, using unique primer sets, was performed at least twice. The PCR evaluation of integrated HIV was performed using the primers and PCR conditions described by Chun et al. [38]. Briefly, the primers were:

Alu->LTR 5, 5'-TCCCAGCTACTCGGGAGGCTG AGG-3'

LTR->Alu 3', 5'-AGGCAAGCTTTATTGAGGCTTA AGC-3'

Nested secondary PCR primers (generating a 352 bp amplicon):

5', 5'-CACACACAAGGCTACTTCCCT-3'

3', 5'-GCCACTCCCCIGTCCCGCCC-3'

We used Ranger polymerase and buffer conditions (Meridian Bioscience, Thomas Scientific, Swedesboro, NJ) for the long-range PCR with Alu-LTR primer sets, and BioTaq polymerase and buffer conditions (Meridian Bioscience, Thomas Scientific, Swedesboro, NJ) for the nested PCR. The integrated HIV PCR was performed on an MJ PTC-200 with an MJR 2 × 48 and a Chromo-4 alpha unit for the long-range and nested PCR, respectively. The nested PCR was performed as a real-time assay using SYBR Green I to detect the amplicon progression curves and evaluate the melting curve.

For detection of total HIV DNA, to determine if comparable total HIV DNA was present in the samples the same samples described above, we utilized the 5' nested primer (5'-CACACACAAGGCTACTTCCCT-3') along with the LTR->Alu 3' primer (5'-AGGCAAG CTTTATTGAGGCTTAAGC-3') using PCR conditions similar to the nested PCR described above but with a 30 sec extension time for the 484 bp amplicon. This PCR was performed on a Roche LightCycler 480 instrument using BioTaq polymerase and buffer conditions with SYBR Green I detection of the 484 bp amplicon.

Detection of circular 2-LTR DHIV3-mCherry PIC DNA was performed as described in Brussel and Sonigo [37]. Briefly, the primers used were:

HIV F, 5' GTGCCCGTCTGTTGTGTGACT 3'

HIV R1, 5' ACTGGTACTAGCTTGTAGCACCATC CA 3'.

Initially we performed the PCR conditions used by Brussel and Sonigo [37], using BioTaq polymerase and buffer conditions with a 25 sec extension time and SYBR Green I detection, but we were unable to detect any amplicon 2LTR circle PIC product. Following the detection of the total PIC HIV data, we hypothesized that the 2-LTR content in these samples could be considerably lower at this 24 hour time point. Therefore, we reran the PCR again a second time and were able to detect an expected 231 bp amplicon. To verify this approach, we synthesized new PCR primers:

RU5 forward: 5'-GCTTAAGCCTCAATAAAG CTTGCCT-3' (this is the compliment of the LTR->Alu primer described above by Chun et al. [38]).

U3 reverse: 5'-ACAAGCTGGTGTTCTCTCCT-3'.

This primer set also did not generate the 2-LTR circular PIC amplicon within 50 PCR cycles but was designed to encompass the amplicon generated by Brussel and Sonigo primers above. Therefore, we used the HIV F and R1 primers as a nested set and ran a 1:20 dilution of this amplification for another 50 PCR cycles to obtain the expected 231 bp amplicon. This demonstrates the 2-LTR circular form of PIC cDNA was present at low levels in all the conditions where DHIV3-mCherry was used, but not in the control samples.

### *Sequential DHIV3-mCherry, DHIV3-GFP infection*

PBMCs from healthy human donors were isolated using lymphocyte separation medium (Biocoll separating solution; Biochrom) or lymphoprep (Stemcell). CD4 + T cells were negatively isolated using the RosetteSep™Human CD4 + T Cell Enrichment Cocktail (Stem Cell Technologies) or the EasySep™ Human Naïve CD4 + T Cell Isolation Kit (Stem Cell Technologies) according to the manufacturer's instructions. Primary rCD4s were cultured to a density of 5 x 106/mL in RPMI-1640 medium containing 10% FCS, glutamine (2 mM), streptomycin (100 mg/mL), penicillin (100 U/mL) and interleukin 2 (IL-2) (10 ng/mL).

Monocyte-derived macrophages (MDMs) were obtained by stimulation of PBMC cultures with 15 ng/mL recombinant human M-CSF (R&D systems) and 10% human AB serum (Sigma Aldrich) in DMEM supplemented with glutamine (2 mM), streptomycin (100 mg/mL) and penicillin (100 U/mL) for 6 days.

### *Statistical analysis*

The pairwise T-Tests function from Scran was used to determine statistically significant differential expression of genes between groups. This was performed for all comparison sets. Only those genes which were significantly different were included in Hallmark, REACTOME, pseudotime, psupertime, and TFT analyses. Other default statistical standards were adopted from the various software recommendations during data analyses unless otherwise specified.

## Acknowledgments

This work was supported by the University of Utah HCI Flow Cytometry Facility and Metabolomics core facilities, and the HSC High Throughput Genomics core facility. The help of Christopher J. Conley and Brian Lohman in the bio-statistical analysis is gratefully acknowledged. Mr. A.L. Lim was partially supported during this work by the grant U01TW008163 (P.I., Margo Haygood, University of Utah). This work was also supported in part by R21 AI124823 (LRB & VP) and a University of Utah seed grant (LRB).

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the National Institutes of Health [R21 AI124823]; National Institutes of Health [U01TW008163]; University of Utah [Seed Grant].

## Author summary

A single cell analysis can distinguish between HIV-1 infected macrophage cells that are transcribing pre-integrated HIV-1 cDNA and those transcribing HIV-1 provirus. Only cells transcribing HIV-1 provirus are making detectable p24, marker mCherry, and Vpu proteins, which corresponds with a change in the host cell's background transcriptome from one expressing viral restriction and immunological response genes to one that is expressing genes associated with cell replication and oxidative phosphorylation.

## Data availability statement

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials. The data that support the findings of this study are openly available in "figshare" at the following dois:

Appendix I - https://doi.org/10.6084/m9.figshare.16834633.v1

Appendix II - https://doi.org/10.6084/m9.figshare.16834639.v1

Appendix III - https://doi.org/10.6084/m9.figshare.16834636.v1

Appendix IV - https://doi.org/10.6084/m9.figshare.16834663.v1

## References

[1] Finzi D, Hermankova M, Pierson T, et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. Science. 1997;278(5341):1295–1300.

[2] Chun TW, Carruth L, Finzi D, et al. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. Nature. 1997;387(6629):183–188.

[3] Wong JK, Hezareh M, Günthard HF, et al. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. Science. 1997 Nov 14;278(5341):1291–1295. PMID: 9360926.

[4] Szaniawski MA, Spivak AM, Bosque A, et al. Sex influences SAMHD1 activity and susceptibility to human immunodeficiency virus-1 in primary human macrophages. J Infect Dis. 2019 Feb 15;219(5):777–785. PMID: 30299483; PMCID: PMC6376916.

[5] Pang S, Koyanagi Y, Miles S, et al. High levels of unintegrated HIV-1 DNA in brain tissue of AIDS dementia patients. Nature. 1990 Jan 4;343(6253):85–89. PMID: 2296295.

[6] Zhou Y, Zhang H, Siliciano JD, et al. Kinetics of human immunodeficiency virus type 1 decay following entry into resting CD4+ T cells. J Virol. 2005Feb;79(4):2199–2210. PMID: 15681422; PMCID: PMC546571.

[7] Hamid FB, Kim J, Shin CG. Distribution and fate of HIV-1 unintegrated DNA species: a comprehensive update. AIDS Res Ther. 2017 Feb 16;14(1):9. PMID: 28209198; PMCID: PMC5314604.

[8] Chan CN, Trinité B, Lee CS, et al. HIV-1 latency and virus production from unintegrated genomes following direct infection of resting CD4 T cells. Retrovirology. 2016 Jan 5;13(1):1. PMID: 26728316; PMCID: PMC4700562.

[9] Wu Y, Marsh JW. Early transcription from nonintegrated DNA in human immunodeficiency virus infection. J Virol. 2003Oct;77(19):10376–10382. PMID: 12970422; PMCID: PMC228496.

[10] Kelly J, Beddall MH, Yu D, et al. Human macrophages support persistent transcription from unintegrated HIV-1 DNA. Virology. 2008 Mar 15;372(2):300–312. Epub 2007 Dec 4. PMID: 18054979; PMCID: PMC2276161.

[11] Kruize Z, Kootstra NA. The role of macrophages in HIV-1 persistence and pathogenesis. Front Microbiol. 2019 Dec 5;10:2828. PMID: 31866988; PMCID: PMC6906147.

[12] Englund G, Theodore TS, Freed EO, et al. Integration is required for productive infection of monocyte-derived macrophages by human immunodeficiency virus type 1. J Virol. 1995May;69(5):3216–3219. PMID: 7707554; PMCID: PMC189028.

[13] Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018 Aug 7;50(8):96. PMID: 30089861; PMCID: PMC6082860.

[14] Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. Front Genet. 2019 Apr 5;10:317. PMID: 31024627; PMCID: PMC6460256.

[15] 10X Genomics Home Page. Accessed 02/02/2022. https://www.10xgenomics.com/?utm_medium%5B0%5D=search&utm_medium%5B1%5D=search&utm_source%5B0%5D=google&utm_source%5B1%5D=google&utm_campaign=sem-goog-2021-website-page-ra_g-brand-amer&useroffertype=website-page&userresearcharea=ra_g&userregion=amer&userrecipient=customer&usercampaignid=7011P000001PCfa&utm_content=Pure%20Brand&utm_term=10x%20genomics&gclid=Cj0KCQiA3smABhCjARIsAKtrg6L5CIu8nsc24EDzVxlSXMQ1IjYrYk8KP9bBVXEva4QP15dnxLtfwjUaAteHEALw_wcB&gclsrc=aw.ds

[16] Sloan RD, Donahue DA, Kuhl BD, et al. Expression of Nef from unintegrated HIV-1 DNA downregulates cell surface CXCR4 and CCR5 on T-lymphocytes. Retrovirology. 2010 May 13;7(1):44. PMID: 20465832; PMCID: PMC2881062.

[17] Wu Y, Marsh JW. Selective transcription and modulation of resting T cell activity by preintegrated HIV DNA. Science. 2001 Aug 24;293(5534):1503–1506. PMID: 11520990.

[18] Poon B, Chang MA, Chen IS. Vpr is required for efficient Nef expression from unintegrated human immunodeficiency virus type 1 DNA. J Virol. 2007Oct;81(19):10515–10523. Epub 2007 Jul 25. PMID: 17652391; PMCID: PMC2045493.

[19] Gelderblom HC, Vatakis DN, Burke SA, et al. Viral complementation allows HIV-1 replication without integration. Retrovirology. 2008 Jul 9;5(1):60. PMID: 18613957; PMCID: PMC2474848.

[20] Essaghir A, Toffalini F, Knoops L, et al. Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. Nucleic Acids Res. 2010Jun;38(11):e120. Epub 2010 Mar 9. PMID: 20215436; PMCID: PMC2887972.

[21] Banks CJ, Joshi A, Michoel T. Functional transcription factor target discovery via compendia of binding and expression profiles. Sci Rep. 2016 Feb 9;6(1):20649. PMID: 26857150; PMCID: PMC4746627.

[22] ENCODE Transcription Factor Targets Dataset. Accessed 02/02/2022. https://maayanlab.cloud/Harmonizome/dataset/ENCODE+Transcription+Factor+Targets

[23] Gaynor R. Cellular transcription factors involved in the regulation of HIV-1 gene expression. AIDS. 1992 Apr;6(4):347–364.

[24] Schiralli Lester GM, Henderson AJ. Mechanisms of HIV transcriptional regulation and their

contribution to latency. Mol Biol Int. 2012;2012:614120. Epub 2012 Jun 3. PMID: 22701796; PMCID: PMC3371693.

[25] Kundu M, Guermah M, Roeder RG, et al. Interaction between cell cycle regulator, E2F-1, and NF-kappaB mediates repression of HIV-1 gene transcription. J Biol Chem. 1997 Nov 21;272(47):29468–29474. PMID: 9368006.

[26] Mbonye U, Karn J. Transcriptional control of HIV latency: cellular signaling pathways, epigenetics, happenstance and the hope for a cure. Virology. 2014Apr;454-455:328–339Epub 2014 Feb 22. PMID: 24565118; PMCID: PMC4010583.

[27] Frolov MV, Dyson NJ. Molecular mechanisms of E2F-dependent activation and pRB-mediated repression. J Cell Sci. 2004 May 1;117(Pt 11):2173–2181. PMID: 15126619.

[28] Bracken AP, Ciro M, Cocito A, et al. E2F target genes: unraveling the biology. Trends Biochem Sci. 2004Aug;29(8):409–417. PMID: 15362224.

[29] Bosque A, Planelles V. Induction of HIV-1 latency and reactivation in primary memory CD4+ T cells. Blood. 2009 Jan 1;113(1):58–65. Epub 2008 Oct 10. PMID: 18849485; PMCID: PMC2614643.

[30] Hotter D, Bosso M, Jønsson KL, et al. IFI16 targets the transcription factor Sp1 to suppress HIV-1 transcription and latency reactivation. Cell Host Microbe. 2019 Jun 12;25(6):858–872.e13. Epub 2019 Jun 4. PMID: 31175045; PMCID: PMC6681451.

[31] Mascolini M Merck Offers Unique Perspective on Second-Generation Integrase Inhibitor 10th International Workshop on Clinical Pharmacology of HIV Therapy, Amsterdam, 2009 Apr 15-17

[32] Ikebe E, Matsuoka S, Tezuka K, et al. Activation of PERK-ATF4-CHOP pathway as a novel therapeutic approach for efficient elimination of HTLV-1-infected cells. Blood Adv. 2020 May 12;4(9):1845–1858. PMID: 32369565; PMCID: PMC7218438.

[33] NIH AIDS Reagent Program, Reagent Datasheet Detail: monoclonal mouse IgG-AG3.0. Accessed 02/02/2022, (NIH AIDS Res Reagent Prog, Germantown, MD; Cat. # 4121) https://www.hivreagentprogram.org/Catalog/HRPMonoclonalAntibodies/ARP-4121.aspx

[34] Sauter D, Hotter D, Van Driessche B, et al. Differential regulation of NF-κB-mediated proviral and antiviral host gene expression by primate lentiviral Nef and Vpu proteins. Cell Rep. 2015 Feb 3;10(4):586–599. Epub 2015 Jan 22. PMID: 25620704; PMCID: PMC4682570.

[35] Langer S, Hammer C, Hopfensperger K, et al. HIV-1 Vpu is a potent transcriptional suppressor of NF-κB-elicited antiviral immune responses. Elife. 2019 Feb 5;8:e41930. PMID: 30717826; PMCID: PMC6372280.

[36] Chun TW, Stuyver L, Mizell SB, et al. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. Proc Natl Acad Sci U S A. 1997 Nov 25;94(24):13193–13197. PMID: 9371822; PMCID: PMC24285.

[37] Brussel A, Sonigo P. Evidence for gene expression by unintegrated human immunodeficiency virus type 1 DNA species. J Virol. 2004Oct;78(20):11263–11271. PMID: 15452245; PMCID: PMC521838.

[38] Tedesco S, De Majo F, Kim J, et al. Convenience versus biological significance: are PMA-differentiated THP-1 cells a reliable substitute for blood-derived macrophages when studying *in vitro* polarization? Front Pharmacol. 2018 Feb 22;9:71. PMID: 29520230; PMCID: PMC5826964.

[39] Chanput W, Mes JJ, Wichers HJ. THP-1 cell line: an in vitro cell model for immune modulation approach. Int Immunopharmacol. 2014Nov;23(1):37–45. Epub 2014 Aug 14. PMID: 25130606.

[40] Gažová I, Lefevre L, Bush SJ, et al. the transcriptional network that controls growth arrest and macrophage differentiation in the human myeloid leukemia cell line THP-1. Front Cell Dev Biol. 2020 Jul 3;8:498. PMID: 32719792; PMCID: PMC7347797.

[41] Cassol E, Cassetta L, Alfano M, et al. Macrophage polarization and HIV-1 infection. J Leukoc Biol. 2010Apr;87(4):599–608. Epub 2009 Dec 30. PMID: 20042468.

[42] Spano A, Barni S, Sciola L. PMA withdrawal in PMA-treated monocytic THP-1 cells and subsequent retinoic acid stimulation, modulate induction of apoptosis and appearance of dendritic cells. Cell Prolif. 2013Jun;46(3):328–347. PMID: 23692091; PMCID: PMC6496477.

[43] Aldo PB, Craveiro V, Guller S, et al. Effect of culture conditions on the phenotype of THP-1 monocyte cell line. Am J Reprod Immunol. 2013Jul;70(1):80–86. Epub 2013 Apr 29. PMID: 23621670; PMCID: PMC3703650.

[44] Daigneault M, Preston JA, Marriott HM, et al. The identification of markers of macrophage differentiation in PMA-stimulated THP-1 cells and monocyte-derived macrophages. PLoS One. 2010 Jan 13;5(1):e8668. PMID: 20084270; PMCID: PMC2800192.

[45] Italiani P, Boraschi D. From monocytes to M1/M2 macrophages: phenotypical vs. Functional differentiation. Front Immunol. 2014 Oct 17;5:514. PMID: 25368618; PMCID: PMC4201108.

[46] Genin M, Clement F, Fattaccioli A, et al. M1 and M2 macrophages derived from THP-1 cells differentially modulate the response of cancer cells to etoposide. BMC Cancer. 2015 Aug 8;15(1):577.PMID: 26253167; PMCID: PMC4545815.

[47] Bosshart H, Heinzelmann M. THP-1 cells as a model for human monocytes. Ann Transl Med. 2016Nov;4(21):438. PMID: 27942529; PMCID: PMC5124613.

[48] Maeß MB, Wittig B, Lorkowski S. Highly efficient transfection of human THP-1 macrophages by nucleofection. J Vis Exp. 2014Sep;2;(91(4):e51960. PMID: 25226503; PMCID: PMC4828023.

[49] Schwartz S, Felber BK, Benko DM, et al. Cloning and functional analysis of multiply spliced mRNA species of human immunodeficiency virus type 1. J Virol. 1990Jun;64(6):2519–2529. PMID: 2335812; PMCID: PMC249427.

[50] Sheldon M, Ratnasabapathy R, Hernandez N. Characterization of the inducer of short transcripts, a human immunodeficiency virus type 1 transcriptional element that activates the synthesis of short RNAs. Mol Cell Biol. 1993Feb;13(2):1251–1263. PMID: 8423790; PMCID: PMC359010.

[51] Svensson V, Natarajan KN, Ly LH, et al. Power analysis of single-cell RNA-sequencing experiments. Nat

Methods. 2017Apr;14(4):381–387. Epub 2017 Mar 6. PMID: 28263961; PMCID: PMC5376499.

[52] Qiu P. Embracing the dropouts in single-cell RNA-seq analysis. Nat Commun. 2020 Mar 3;11(1):1169. PMID: 32127540; PMCID: PMC7054558.

[53] Macnair W, Claassen M, Psupertime: supervised pseudotime inference for single cell 2 RNA-seq data with sequential labels. 2019. Accessed 02/02/2022, Mar 28. https://www.biorxiv.org/content/10.1101/622001v1

[54] Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017Oct;14(10):979–982. Epub 2017 Aug 21. PMID: 28825705; PMCID: PMC5764547.

[55] Street K, Risso D, Fletcher RB, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018 Jun 19;19 (1):477. PMID: 29914354; PMCID: PMC6007078.

[56] Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018Jun;36(5):411–420. Epub 2018 Apr 2. PMID: 29608179; PMCID: PMC6700744.

[57] Scialdone A, Natarajan KN, Saraiva LR, et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods. 2015 Sep 1;85:54–61. Epub 2015 Jul 2. PMID: 26142758.

[58] Sergushichev AA. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. 2016;bioRxiv:060012. 10.1101/060012

[59] Rubin SM. Deciphering the retinoblastoma protein phosphorylation code. Trends Biochem Sci. 2013Jan;38 (1):12–19. Epub 2012 Dec 3. PMID: 23218751; PMCID: PMC3529988.

[60] Thierry S, Thierry E, Subra F, et al. Opposite transcriptional regulation of integrated vs unintegrated HIV genomes by the NF-κB pathway. Sci Rep. 2016 May 11;6(1):25678. PMID: 27167871; PMCID: PMC4863372.

[61] Kundu M, Srinivasan A, Pomerantz RJ, et al. Evidence that a cell cycle regulator, E2F1, down-regulates transcriptional activity of the human immunodeficiency virus type 1 promoter. J Virol. 1995Nov;69(11):6940–6946. PMID: 7474112; PMCID: PMC189612.

[62] Ambrosino C, Palmieri C, Puca A, et al. Physical and functional interaction of HIV-1 Tat with E2F-4, a transcriptional regulator of mammalian cell cycle. J Biol Chem. 2002 Aug 30;277(35):31448–31458. Epub 2002 Jun 7. PMID: 12055184.

[63] Bradley T, Ferrari G, Haynes BF, et al. Single-cell analysis of quiescent HIV infection reveals host transcriptional profiles that regulate proviral latency. Cell Rep. 2018 Oct 2;25(1):107–117.e3. PMID: 30282021; PMCID: PMC6258175.

[64] Cohn LB, da Silva IT, Valieris R, et al. Clonal CD4$^{+}$ T cells in the HIV-1 latent reservoir display a distinct gene profile upon reactivation. Nat Med. 2018May;24 (5):604–609. Epub2018 Apr 23. PMID: 29686423; PMCID: PMC5972543.

[65] Golumbeanu M, Cristinelli S, Rato S, et al. Single-cell RNA-seq reveals transcriptional heterogeneity in latent and reactivated HIV-infected CELLS. Cell Rep. 2018 Apr 24;23(4):942–950. PMID: 29694901.

[66] Lun ATL, Riesenfeld S, Andrews T, Gomes T, et al. Participants in the 1st human cell atlas Jamboree, Marioni JC. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. 2019 Mar 22;20(1):63. PMID: 30902100; PMCID: PMC6431044.

[67] McCarthy DJ, Campbell KR, Lun AT, et al. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017 Apr 15;33(8):1179–1186. PMID: 28088763; PMCID: PMC5408845.

[68] Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.

[69] How does Cell Ranger process and filter UMIs? https:kb.10xgenomics.com/hc/en-us/articles/115003133812-How-does-cellranger-count-process-and-filter-UMIs-, 10xgenomics. Accessed 02/02/2022.

[70] loess: Local Polynomial Regression Fitting. Accessed 02/02/2022. https://rdrr.io/r/stats/loess.html

[71] SnapGene software (from Insightful Science; available at snapgene.com)