Position effects influence HIV latency reversal

Heng-Chang Chen[1,2], Javier P. Martinez[3], Eduard Zorita[1,2], Andreas Meyerhans[3,4] & Guillaume J. Filion[1,2]

[1]Genome Architecture, Gene Regulation, Stem Cells and Cancer Programme, Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
[2]University Pompeu Fabra, Doctor Aiguader, 08003 Barcelona, Spain
[3]Infection Biology Group, Department of Experimental and Health Sciences, University Pompeu Fabra, Doctor Aiguader, 08003 Barcelona, Spain
[4]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

SUMMARY

The main obstacle to curing HIV is the presence of latent proviruses in the body of infected patients. In the recent years, its has been proposed to use drugs to activate the expression of HIV while maintaining the antiretroviral therapy in hope of purging the latent reservoir. However, such therapies do not reactivate all the latent proviruses and their efficiency is thus limited. Here, we address the potential role of chromatin on the prevention of latency reversal by developing a method called Barcoded HIV Ensembles (B-HIVE) to map the chromosomal location of thousands of individual proviruses and to track their transcriptional activity in an infected cell population. Using B-HIVE in Jurkat cells, we discovered that the induction of HIV latency depends on the insertion site, and that it is strongest away from the active enhancers of the host cell. The insertion site also has an effect on the effectiveness of drugs used to reactivate latent viruses. We found that two latency reversing agents, phytohaemagglutinin and vorinostat, reactivate proviruses at distinct genomic locations. Based on these results, we propose that combinations of drugs that reactivate complementary subsets of latent proviruses will be most effective. Overall, our data suggests that the insertion context of HIV is a critical determinant of both the fate of the infection and the response of the virus to reactivation therapies.

INTRODUCTION

The development of a treatment for the human immunodeficiency virus (HIV) was a turning point in the fight against AIDS. Current antiretroviral therapy (ART) suppresses HIV replication to undetectable levels in the plasma. However, some infected cells do not respond to ART and HIV rapidly rebounds when the treatment is interrupted[1–4]. The existence of a viral reservoir escaping ART is the major hurdle towards the development of a cure for HIV.

Latent infections of resting CD4+ T lymphocytes are an important component of the treatment-resistant HIV reservoir. As HIV is transcriptionally silent in these cells, the viruses are invisible to antiviral drugs and to the immune system. Recent research efforts were geared towards purging the reservoir by reactivation therapies[5]. The principle is to maintain the patients under ART while using additional drugs, so-called latency reversing agents, to force the expression of the provirus. This should then render the infected cells susceptible to clearance by the immune system or by cytopathic effects.

One main question is how to estimate the efficiency of this strategy *in vivo*[6]. Viral outgrowth assays give an estimate of the reservoir size at 0.1-10 cells with a latent infection per million resting CD4+ T lymphocytes[6,7]. However, they only provide a minimal approximation of the frequency of latently-infected cells. In contrast, PCR-based methods may offer an overestimate of the purgeable reservoir as they cannot distinguish between replication-competent proviruses and defective or hypermutated proviruses. Moreover, replication-competent proviruses have variable propensities to reactivation. This defines an “inducible” reservoir composed of viruses that can be reactivated, and a “non-inducible” reservoir composed of viruses that cannot[8]. Approximately 10% of the non-inducible viruses are replication-competent[8]. It is still unknown how these viruses are maintained silent, indicating that a critical element is missing to understand HIV latency.

It is known since the 1930s that the position of a gene has a critical influence on its expression. When X-ray mutagenesis made it possible to induce chromosomal rearrangements, it was observed that some genes become silent when translocated near the centromere[9]. This phenomenon, collectively known as “position effects” was later explained by the role of histones in transcription. The post-translational modifications of the histone tails, typically found at the centromere, can recruit repressors and shut down the expression of a gene, even if its sequence is unchanged[10]. Position effects do not only take place in pericentromeric

heterochromatin, but also in the rest of the genome. For instance, it has long been observed that transgenes are silenced in ways that depend on their insertion site[11]. More recently, genome-wide assays such as TRIP (Thousands of Reporters Integrated in Parallel) have revealed that transgenes have similar expressions in similar chromatin contexts[12]. However, how the context modifies gene expression has not yet been elucidated. It is still an open question, for instance, whether the histone marks found on chromosome arms are implicated in the same way as in pericentromeric heterochromatin. In summary, the genomic context of a gene plays a major role in its expression, but the forces at work are largely unknown.

The above suggests that the insertion context may influence the expression of HIV. This idea is supported by experimental data on cell lines showing that the LTR promoter is sensitive to the local chromatin environment[13] and that latent proviruses are often found near heterochromatin[14]. Interestingly, the insertion of HIV in the human genome is nonrandom. The provirus preferentially integrates in active genes and gene-rich chromosomes[15,16]. It was recently shown that this pattern corresponds to loci in physical proximity to the nuclear pores[17], but it is unclear whether the distribution of insertion sites in the human genome represents an evolutionary optimum or an accident. At any rate, there is some evidence that HIV latency may be induced by the chromatin context of the insertion site.

In contrast, the potential role of the insertion site in reactivation therapies has never been addressed. For instance, the first drug used in clinical trials for reactivation therapy is a histone deacetylase (HDAC) inhibitor[18]. This strategy relies on the assumption that chromatin is a key component of HIV latency, but it disregards the fact that HDACs target only a subset of the genome[19]. It is thus doubtful that HDAC inhibitors have the same effect on all latent proviruses, independently of their insertion site. Besides, many genes are not activated upon blocking HDACs[19], so HDAC inhibitors are expected to be ineffective when the provirus is silenced by alternative mechanisms. Overall, the possibility that the failures of reactivation therapy are due to position effects was given little consideration. The reason for this gap is the lack of technologies to study the role of the genomic context in HIV reactivation. More generally, it is important to know whether a latency reversing agent reactivates only a subset of the latent proviruses, but this is presently impossible because they produce indistinguishable virions.

Here we address this question by developing the Barcoded HIV Ensembles (B-HIVE) technology to track the expression of individual HIV proviruses in a heterogeneous population. The strategy is to insert DNA barcodes into the genome of recombinant HIV and use these barcodes to identify all the transcripts produced by a provirus. Using B-HIVE, we show that the expression of HIV depends on the insertion site, confirming that the provirus is sensitive to position effects. B-HIVE further revealed that expression hotspots do not coincide with integration hotspots, but with promoters and enhancers of the host cell. Conversely, latent proviruses are typically found far away from promoters and enhancers. Finally, B-HIVE revealed that different latency reversing agents activate different subsets of latent proviruses. These results have important implications for reactivation therapy as they suggest to develop cocktails of drugs with complementary spectra instead of drugs with synergistic interactions. In summary, our work provides a novel high throughput technology to study HIV latency and to assay latency reversing agents.

RESULTS

Principle of B-HIVE

In order to determine how the local sequence context influences both HIV integration, latency, and response to drugs, we developed B-HIVE, a genome-wide method to map insert-specific expression across thousands of integrated viruses. B-HIVE is inspired from TRIP[12], a genome-wide method to study position effects. Briefly, the principle of B-HIVE is to tag individual HIV genomes with a unique barcode of 20 nucleotides to track the viral transcripts produced by each provirus of the infected cell population (Figure 1, Supplementary Figure S2). The barcodes are randomly generated during the library preparation (see Methods and Supplementary Figure S1) and their sequence is unknown until the analysis stage. They can still be used as universal identifiers because the complexity of the library is several orders of magnitude larger than the total number of HIV infections in the cell population, so the probability of two proviruses sharing the same exact barcode is negligible (see derivation in Methods).

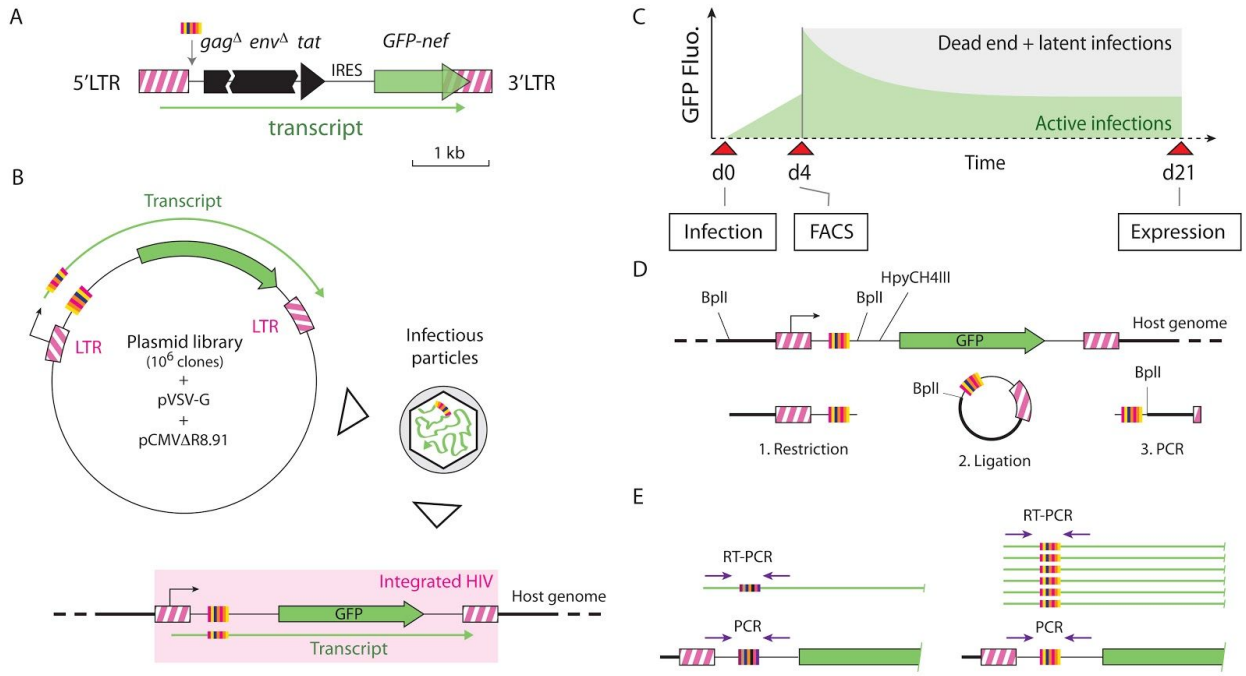


Figure 1. Principles of the B-HIVE technology

(A) Structure of the parental minimal lentiviral vector used for B-HIVE. This construct expresses Tat and GFP under the control of the HIV-1 LTR. The arrow and the multicolor tag indicates the position where the barcode is inserted. (B) Preparation of the barcoded library and infection. Barcodes are represented as multicolor tags. Infectious viral particles are prepared by co-transfection of 293T cells with two additional plasmids expressing the HIV-1 gag-pol and the VSV-G proteins, respectively (See Methods). Barcoded pseudotypes (HIV_{BCD}) are used to infect Jurkat cells. The region highlighted in pink shows the structure of an integrated provirus. Each provirus is distinguished by its own barcode. (C) Experimental outline of HIV_{BCD} infections and latency. Jurkat cells are infected with barcoded virions at MOI close to 0.5 and GFP(+) cells are FACS-sorted 4 days post infection (Supplementary Figure S4). Immediately after sorting, a founder pool of 20.000 GFP(+) cells is expanded for 17 days. During the expansion phase, about half of the cells lose GFP expression. (D) Provirus mapping is carried out by inverse PCR. The genome of the cell population is digested by the restriction enzymes BpII and HpyCH4III, which creates hybrid human-HIV fragments including the barcodes. The digestion products are self-ligated, PCR-amplified and sequenced paired-end. (E) Quantification of the proviral expression. The provirus on the left has a lower expression than the provirus on the right. Barcode abundance is measured by taking the ratio of read counts after RT-PCR on the RNA pool and PCR on the DNA pool.

We used a 4.6 kb minimal HIV construct containing the Tat transcriptional activator and a GFP-Nef fusion gene allowing us to visually monitor the transcriptional activity (Figure 1A). This construct was shown to recapitulate the initial phase of HIV infection and to spontaneously enter latency in cultured cells[14]. We generated a barcoded HIV library with over one million distinct barcodes and obtained infectious pseudotype particles, named here HIV_{BCD}, by co-transfection

with a VSV-G expression plasmid (Figure 1B). Viral stocks were titrated using the TZM-bl indicator cell line[20]. Jurkat cells were initially infected with a viral inoculum corresponding to a MOI of around 0.5. The efficiency of two independent HIV_{BCD} infections in Jurkat cells was comparable to that of the non-barcoded pseudotypes (Supplementary Figure S3), indicating that barcodes do not seem to disable the viruses.

HIV_{BCD}-infected GFP(+) Jurkat cells are sorted by flow cytometry and used to establish a pool of 20,000 infected founder cells (Figure 1C). This cell pool is then expanded in culture so that each founder cell produces a clone with identical proviruses. The purpose of this approach is threefold. First, it ensures that every cell in the founder population is infected with at least one transcriptionally active barcoded virus. Second, the initial, relatively low cell number in the founder pool decreases the chances of two viruses having the same barcode. Third, the LTR-dependent GFP expression allows us to identify latency events by the loss of fluorescence (see below). The integrated proviruses are mapped by inverse PCR[21,22] followed by high throughput sequencing (Figure 1D). This step also reveals the sequences of the barcodes, which is critical to associate each transcript to a single genomic location within the host cell (Figure 1B, Supplementary Figure S5). The expression of individual proviruses is assayed by measuring the abundance of the associated barcode in the RNA pool compared to the DNA pool of the population (Figure 1E). This normalization by copy number is important because the founder cells may not divide at the same rate, thus creating an imbalance in the representation of the barcodes. Given this information, it is possible to obtain the expression landscape of individual proviruses inserted at distinct genomic locations.

Barcoded viruses have wild type insertion patterns

HIV has a characteristic nonrandom pattern of insertion into the host genome[15,16]. In T cells, HIV preferentially targets active genes and favours gene-rich chromosomes over gene-poor chromosomes. In order to analyse whether the barcoded viruses follow the same trend, we carried out two independent infections of barcoded viruses in Jurkat cells as explained above. Figure 2A shows the chromosomal distribution of integrated barcoded viruses for the two replicates. We detected HIV_{BCD} insertions throughout the genome, though not uniformly, and with a clear enrichment in chromosomes 16, 17 and 19 in both replicates. These results are in agreement with a previous study that analysed the global distribution of HIV insertions in the human chromosome[17]. It is also apparent from the map that chromosomes X and Y are targeted

less frequently than autosomes. This is expected since Jurkat cells have two copies of each autosome but only one copy of the sex chromosomes. The small fraction of insertions that cannot be mapped unambiguously was discarded (*e.g*. when the provirus is inserted in a repeated sequence). Overall, we recovered 1,245 and 1,227 HIV insertion sites, in each replicate respectively, which together recapitulate the known features of HIV integration.

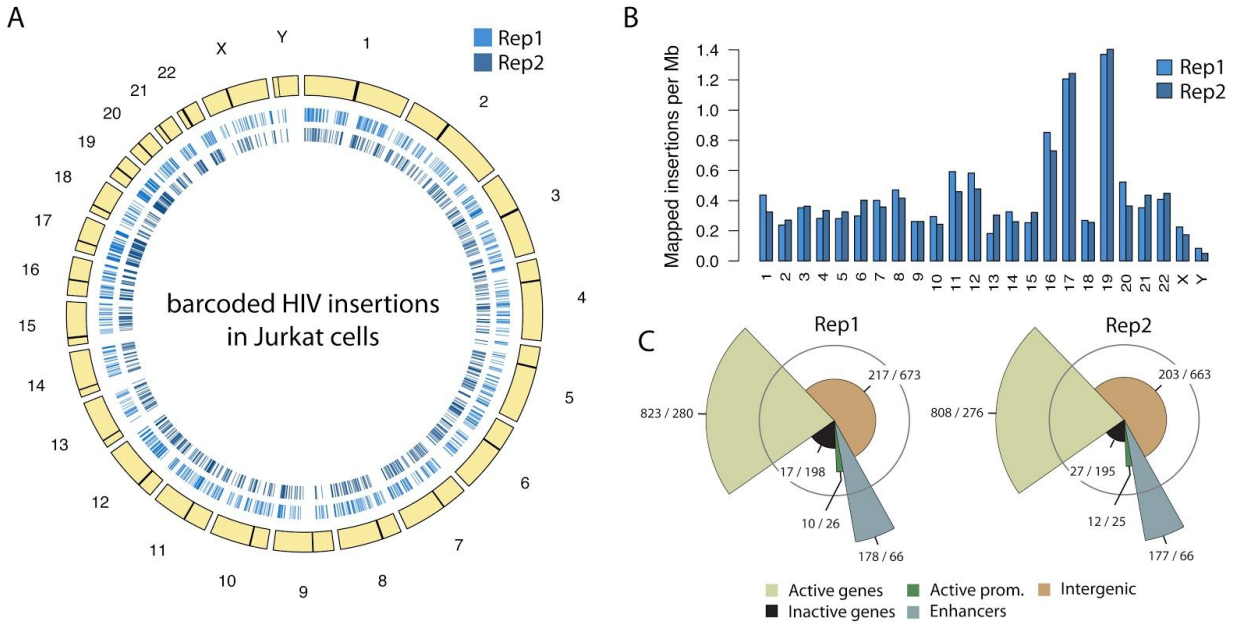


Figure 2. Insertion landscape of barcoded viruses

(A) Genome-wide map of HIV insertions. Two infections were carried out independently. Each blue tick indicates the position of a barcoded HIV insertion mapped by B-HIVE. Dense clusters are clearly visible on chromosomes 17 and 19. (B) Quantification of the insertion rate. The number of mapped insertion per Mb is represented as a bar graph for each infection separately. The small number of insertions on the Y chromosome is partly due to the low mappability in this repeat-rich chromosome. (C) HIV insertions in active (protein-coding) genes, inactive (protein-coding) genes, active promoters, enhancers and intergenic regions. The spie chart represents observed versus *expected fractions as follows: the angles of the wedges are proportional to the expected fraction of insertions and their area is proportional to the observed fraction of insertions. The grey circle shows the expected area of the wedges, so that wedges inside the circle are underrepresented and wedges that extend beyond the circle are overrepresented. The observed and expected numbers of insertions are shown in this order next to each category. HIV shows strong insertion biases towards the bodies of active genes (olive) and enhancers (blue). The results between independent infections are very similar.*

The chromosomal insertion biases are highly reproducible between independent infections (Figure 2B). The insertion rate is about 3 times higher than average in chromosomes 17 and 19

and about 2 times higher than average on chromosome 16. Figure 2C shows the insertion biases of barcoded proviruses with respect to the genomic features of Jurkat cells. The spie chart represents the number of HIV insertions in relation with the genomic coverage of five types of regions (active genes, inactive genes, active promoters, enhancers and intergenic regions, see Methods). The insertion rates in active genes and enhancers are 2.9-fold and 2.6-fold higher than expected, respectively. In contrast, insertion rates in inactive genes, intergenic regions and active promoters are 8.9-fold, 3.2-fold and 2.3-fold lower than expected, respectively. Preference towards active genes and regions bearing enhancer marks is a well known feature of HIV[15,23]. In summary, the barcoded viruses show the same global integration pattern as wild type HIV.

Proximity to regulatory elements affects HIV expression

During the growth period following FACS sorting, Jurkat cells lose fluorescence and the levels stabilize around 40% (Figure 1C). At that point, the expression levels in the population are at a steady-state. We measured the individual expression of mapped proviruses by taking the ratio of barcode counts in the RNA pool *versus* the DNA pool (Figure 1E). We could quantify expression for 889 and 966 insertions in each replicate, respectively. In the other cases, the barcodes were missing from either the RNA or the DNA pool, so the data was considered unreliable. The measurements of HIV_{BCD} expression are accurate, as shown by the reproducibility between technical replicates (Figure 3A). The precision of the measurement was also confirmed by template switch RT-qPCR on selected barcodes (Supplementary Figure S6). Interestingly, the levels of expression span more than four orders of magnitude. In B-HIVE, the sequencing power is dedicated to the detection of a single transcript, which gives higher resolution than for a typical transcriptome analysis. As a result, low expression levels can be measured with high accuracy.

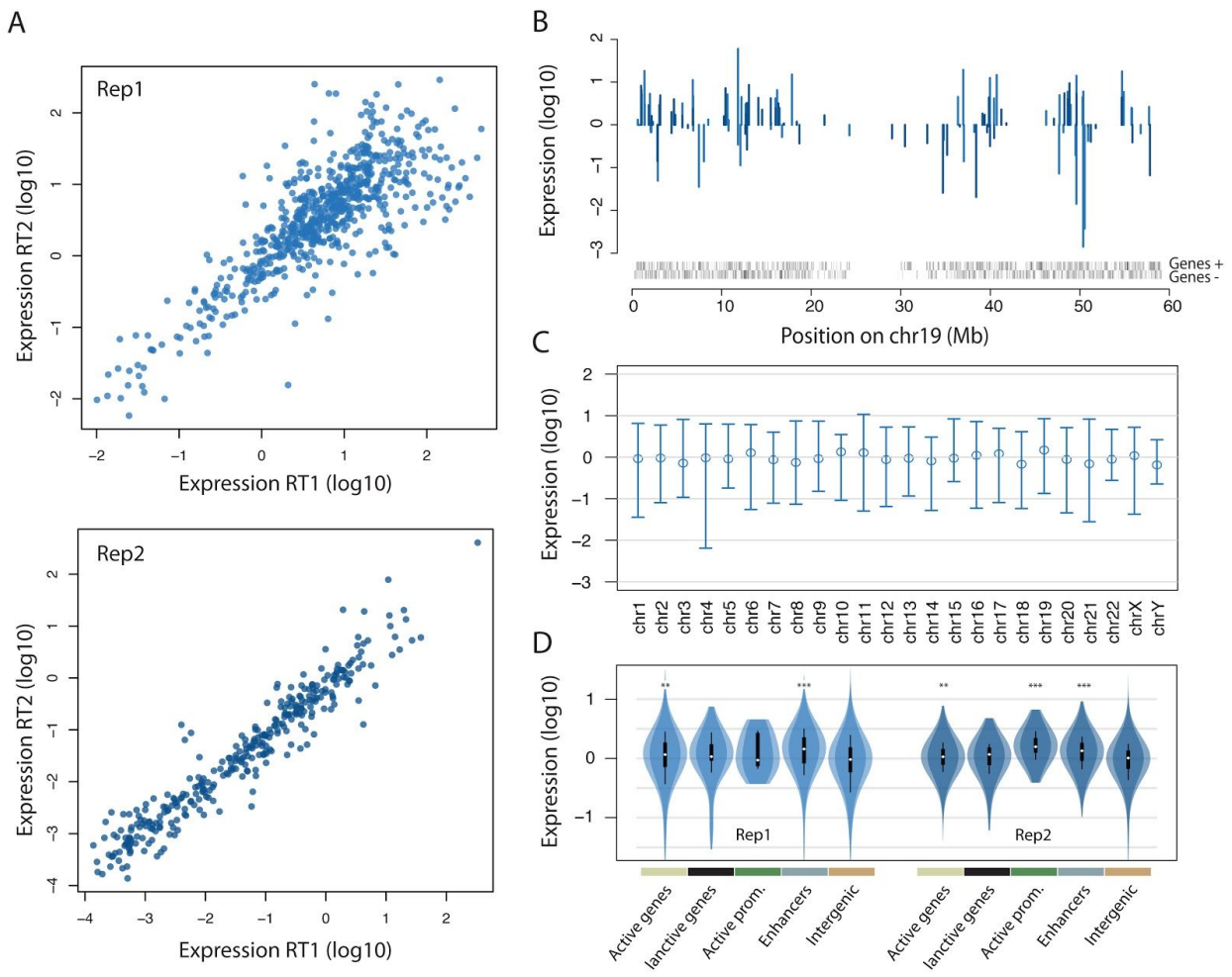


Figure 3. Quantifying the expression of individual barcoded viruses by B-HIVE.

(A) Replicability of the measurement by B-HIVE. The scatter plots are drawn separately for each infection. Each dot represents the same barcoded provirus, whose expression is measured by two distinct RT-PCR (R^2 values 0.69 and 0.94, respectively, note the logarithm on both axes). (B) Example expression profiles of HIV. The position of each vertical bar represents the location of a provirus and the height represents its mean-centered expression level (average of technical replicates, note the logarithm on the y axis). The few proviruses inserted close to the centromere are expressed less than average. Genes+ and Genes- indicate the locations of genes on the + and - strand, respectively. (C) Inter chromosomal variations of HIV expression. The expression of HIV on each chromosome is represented by the average (circle) and an interval containing 90% of the observed values (vertical bar going from the 5-th to the 95-th percentiles, note the logarithm on the y axis). The expression of HIV shows an over 100-fold variation, but it does not depend on the host chromosome, as the average is nearly constant. (D) Expression of HIV in different genomic regions. The violin plots show the mean-centered expression of HIV (note the logarithm on the y axis). The stars indicate the significance of the Wilcoxon test using the intergenic insertions as reference population. Replicate 1: active genes (1.4-fold) P = 0.002, enhancers (2.1-fold) P ~ 10^{-6}. Replicate 2: active genes (1.3-fold) P = 0.005, active promoters (3.4-fold) P ~ 10^{-8} , enhancers P ~ 10^{-12} (2.1-fold).

An example of expression profile of HIV is shown in Figure 3B. We chose to represent chromosome 19 because it contains the highest density of insertions. As previously shown for HIV integrations *in vivo*, HIV_{BCD} is mostly integrated in gene-rich regions in Jurkat cells, *i.e.* away from the centromere. The few proviruses inserted close to the centromere have lower expression than average, even though proviruses with low expression are also found in gene-rich regions.

One may expect that HIV inserts preferentially in chromosomes favoring high expression of the provirus. However, this is not the case: the expression of HIV is nearly constant across chromosomes and the expression on chromosomes 16, 17 and 19 is not significantly higher than in the rest of the genome (ANOVA P = 0.48, Figure 3C). So the actual insertion site on the chromosome seems to matter more than the chromosome targeted for integration. This conclusion is confirmed by inspecting the features surrounding the insertion site. Proviruses inserted within 5 kb of enhancers have a higher expression level (Figure 3D). Proviruses inserted within 5 kb of promoters were also expressed at higher level in the second replicate but not in the first, most likely due to the low number of insertions in this category (Figure 2C). The regulatory elements of the host genome thus seem to influence the expression of HIV. This is consistent with the long standing observation that enhancers do not discriminate self *versus* transgenes and typically activate most genes in their neighborhood regardless of their origin[24].

Surprisingly, viruses inserted in active genes, silent genes and intergenic regions had similar expression levels (expression in active and silent genes is respectively 1.35 times higher and not significantly different from expression in intergenic regions). More generally, when HIV is inserted in a gene, the expression of the provirus is practically uncorrelated with the expression of the host gene (Pearson correlation r = 0.07, P = 0.03, Supplementary Figure S7). Thus, it seems that the transcription of HIV is stimulated by endogenous regulatory elements but not by ongoing transcription. Taken together, our results show that the insertion context has a predictable influence on the expression of HIV.

The insertion site impacts entry into latency

Since the insertion site influences HIV expression, it is interesting to investigate whether it also affects the entry into latency. To tackle this question, we FACS-sorted the population of Jurkat

cells 17 days post infection and separated GFP(+) from GFP(-) subpopulations (Figure 4A). Since the ancestors of all those cells were GFP(+) on day 4 post infection, the loss of fluorescence may be explained by the loss of the virus or the loss of its expression. The first case corresponds to dead end infections, *i.e.* short-lived viruses with a circularized genome that can be expressed but not inserted in the host genome[25], and the second case corresponds to latency, *i.e.* a complete shutdown of HIV expression. It was shown in other studies that HIV becomes latent early after the infection, typically one day[26]. Note that in the experiments presented here, such early latency events are discarded by the FACS-sorting on day 4 post infection, so in what follows, latency will refer only to events occurring at later stages.

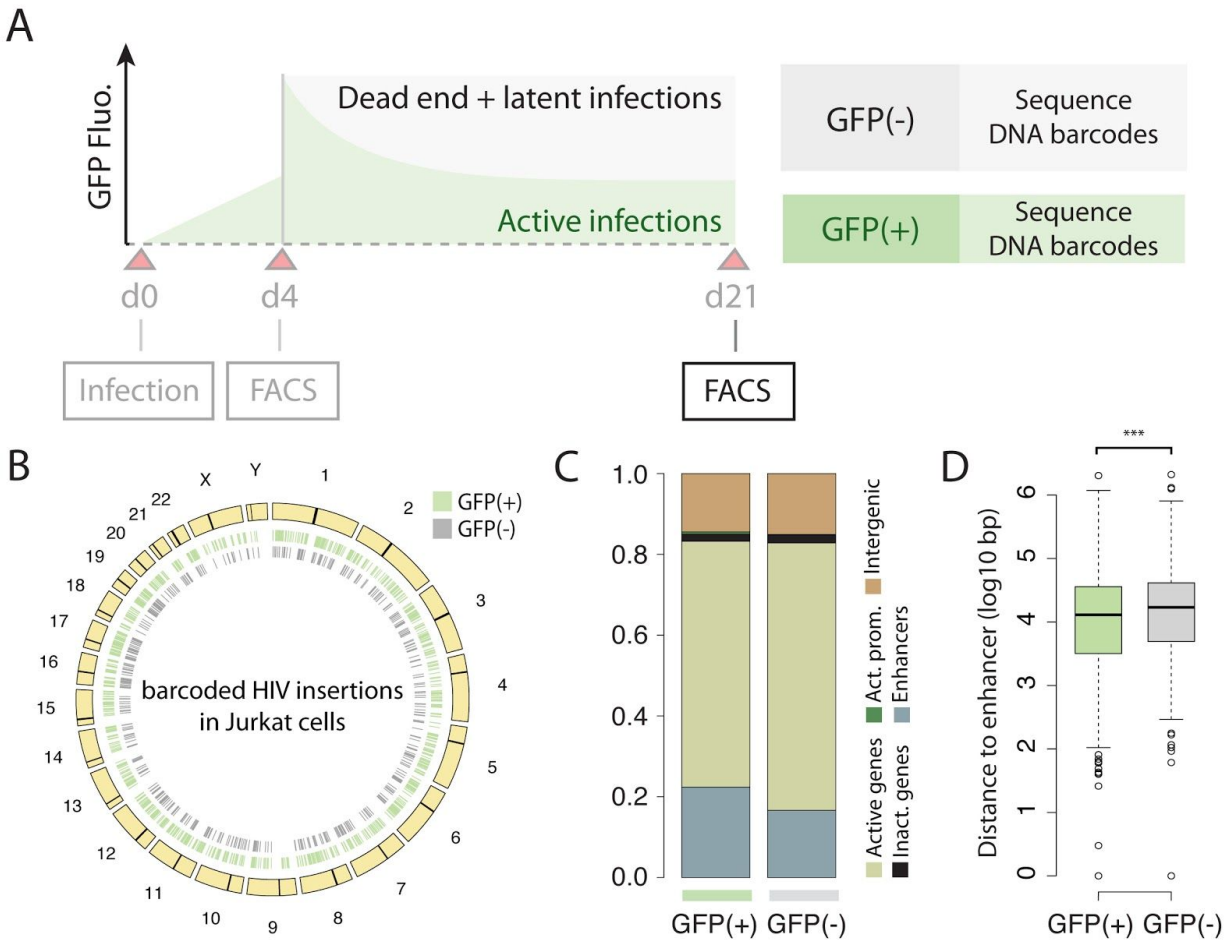


Figure 4. Location of latent proviruses.

(A) Schematic representation of the separation between latent and active populations. 21 days post infection, the GFP fluorescence stabilizes near 40%. GFP(+) and GFP(-) were isolated (1 million cells each) and expanded for 7 days. Genomic DNA was harvested and HIV insertions were identified by sequencing the barcodes. (B) Genome-wide distribution of latent versus *non-latent insertions. Each tick represents a mapped insertion in the*

GFP(+) population (green, 1,468 insertions) or the GFP(-) population (grey, 565 insertions). Latent proviruses are distributed over the whole genome. (C) Distribution of latent versus *non-latent insertions in genomic features. The stacked bar plots represent the proportion of HIV insertions in the five types of regions defined in Figure 2. The distributions are not significantly different. (D) Distance to enhancers. The box plot represents the distance between the HIV insertions and the closest enhancers. Latent insertions are approximately 2 times further away from enhancers than non-latent insertions (Wilcoxon test, P = 0.001). The box plots were generated with the default R parameters.*

We examined the distribution of proviruses in GFP(+) *versus* GFP(-) cells. Latent proviruses are found over the whole genome and no particular pattern emerges at a large scale (Figure 4B). The proportions of insertions in the five categories of genomic features were not significantly different (Figure 4C). However, latent insertions tend to be located further away from enhancers (Figure 4D), consistently with the previous finding that proviruses inserted in proximity of regulatory elements have higher expression levels. Overall, the insertion site influences the entry of HIV into latency, but genomic features are poor predictors of the outcome. This suggests that latency results from complex relationships between HIV and the host genome.

Latency reversing agents target distinct proviruses

Our results so far demonstrate that the insertion context plays a major role on the fate of HIV infection. It is thus possible that it also plays a role in the outcome of potential treatments. So far it has not been possible to test whether a given treatment reactivates all or a fraction of the latent proviruses. B-HIVE provides the unique opportunity to pinpoint which proviruses are reactivated by a drug, and to quantify the magnitude of the individual response.

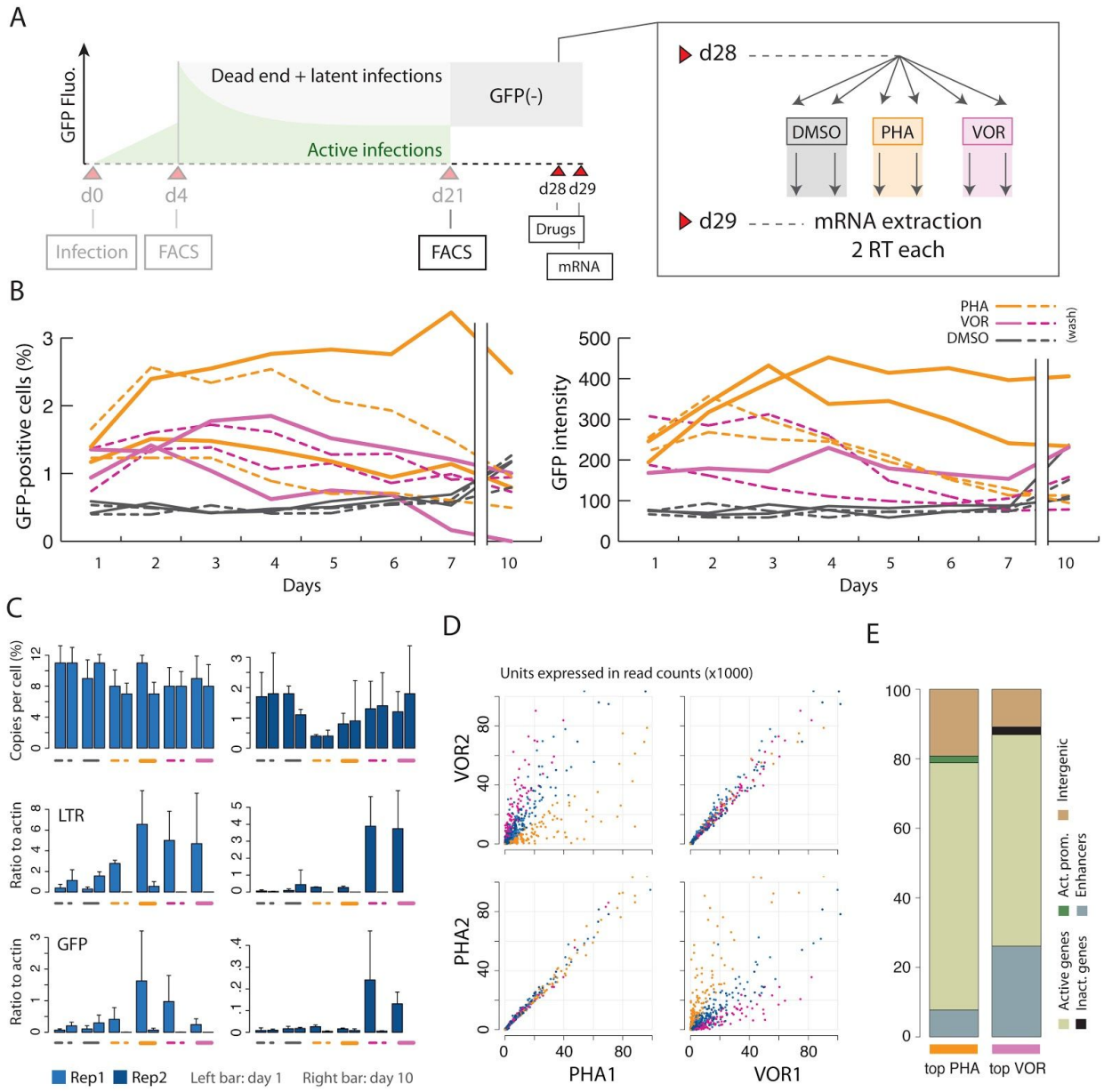


Figure 5. Effect of latency reversing agents on individual proviruses.

(A) Schematic representation of the drug treatment on the GFP(-) population. (B) Kinetics of HIV_{BCD} reactivation by phytohaemagglutinin (PHA), and Vorinostat (VOR). Left: evolution of the proportion of GFP(+) cells after treatment with either PHA (orange), VOR (pink) or DMSO (grey), average of two replicates. Dashed lines represent experiments where the drugs were removed after 24 hours of treatment. At day 1, all the drug treatments significantly increased the proportion of GFP(+) cells compared to DMSO (Student's t test, all $P < 0.006$). Right: evolution of the mean fluorescence intensity of GFP(+) cells, average of two replicates. The second VOR treatment is not represented because it produced aberrant values as the population declined (see left). At day 1, all the drug treatments significantly increased the fluorescent intensity compared to DMSO (Student's t test, all $P < 0.008$). For comparison, the same data acquired on the non-latent GFP(+) population is shown in Supplementary FIgure S9. (C)

Top: qPCR quantification of viral copies using HBB as a reference. Middle and bottom: RT-qPCR of viral expression using actin as a reference. Bars are paired to show the value at day 1 next to the value at day 10. Each blue bar shows the average of 4 values. The error bars show the estimated standard deviation of the measurements. Treatment is indicated below the bars with the same legend as in panel (B). Note that expression is not corrected for the fact that some populations contain fewer viruses. (D) B-HIVE measurement of individual provirus expression after drug treatment. The scatter plots represent the mRNA tag count expression of the same provirus in different conditions. Replicate treatments induce similar expression patterns; different treatments induce distinct expression patterns. This shows that PHA and VOR induce different proviruses. The top 15% PHA and VOR responders are highlighted in orange and pink, respectively. (E) Localizations of top 15% PHA and VOR responders in the genomic categories. Proviruses that better respond to VOR are enriched at enhancers (Chi-square test P = 0.037).

As a proof of principle, we tested two widely characterized latency reversing agents[27]. The first, phytohemagglutinin (PHA), elicits a strong activation of T cells and is routinely used for benchmarking; the second, vorinostat (VOR), is an HDAC inhibitor used in clinical trials[28]. We divided the GFP(-) population into distinct pools and treated them with PHA, with VOR, or with the control drug DMSO (Figure 5A). We also performed a pulse treatment of 24 hours followed by a relaxation phase where the drugs were washed. Figure 5B shows the kinetics of the proportion of GFP(+) cells and of the mean fluorescence intensity of the positive cells. Upon removal of the drugs, HIV expression comes back to control values in about 10 days. Control quantifications by qPCR and RT-qPCR confirm that the viruses are still present after 10 days and that their expression is lower (Figure 5C). An important implication of this result is that there exists a mechanism that actively silences latent proviruses. If latency were under the control of a bistable on/off switch, the active state would perpetuate itself after the removal of the drugs, and the proviruses would remain expressed. Instead, silencing takes place again when the drugs are washed. In the experimental setup depicted in Figure 4A, HIV latency is not a historical accident of the infection but a process instructed by the host genome.

Do PHA and VOR reactivate different subsets of proviruses or do they reactivate the same subset with a different strength? To answer this question, we measured the expression of individual barcodes in the RNA pool after treatment by either drug after 24 treatment (Figure 5D). The expression of each provirus was very similar between replicate treatments with the same drug, showing that the response of the provirus is reproducible. In contrast, we observed substantial variation when comparing the expression of the same provirus treated by different drugs (Figure 5D and Supplementary Figure S8). In other words, the same provirus may

respond more strongly to PHA than to VOR, or *vice versa*. Thus, the activities of the drugs are due in part to their different spectra, *i.e.* to the subset of latent proviruses they stimulate.

These results suggest that one may identify the proviruses that respond more to one drug or to the other. To this end we ranked the proviruses as a function of their response to PHA or VOR and formed two groups representing the top 15% of each class (see Methods). The proviruses that better respond to VOR and those that better respond to PHA have different distributions in the genomic categories (Figure 5E). In particular, the top 15% VOR responders are more frequently in the proximity of enhancers than the top 15% PHA responders. The insertion context thus carries some information about the potential response of a latent provirus to a given drug.

To further characterize the difference between these two classes of proviruses, we measured their genomic distances to chromatin features defined by histone modifications[29–31] (Figure 6A). The top 15% VOR responders are on average closer to peaks of H3K27ac, H3K4me3 and H3K4me1 which are characteristic marks of active regulatory elements. In contrast, the distances to H3K36me3, H3K79me3 and H3K9me3 domains were not different between the top 15% VOR responders and the top 15% PHA responders. The first two marks are characteristic of the bodies of active genes, and the last is typically found in heterochromatin. Thus, proviruses that have a stronger response to VOR than to PHA tend to lie closer to active regulatory elements. Taken together, our results show that the insertion context of HIV is critical to determine not only the fate of the infection, but also the response of the virus to reactivation therapies.

DISCUSSION

Here we developed and used the B-HIVE technology to study position effects on individual HIV integrations in the human genome (Figure 1). As expected, we observed a strong enrichment of insertions in chromosomes 17 and 19 (Figure 2B). We also showed that the insertion site impacts the expression of the provirus. The chromosome itself has little influence (Figure 3C), but proximity to the regulatory elements of the host induces higher expression levels (Figure 3D). We next showed that latent proviruses tend to locate further away from the enhancers of the host genome than non-latent proviruses (Figure 4), and finally that drugs used to reactivate

latent HIV target proviruses inserted at different locations (Figure 5C). Overall, these results illustrate the power of barcoding strategies and highlight the importance of the chromosomal context in the fate of an HIV infection.

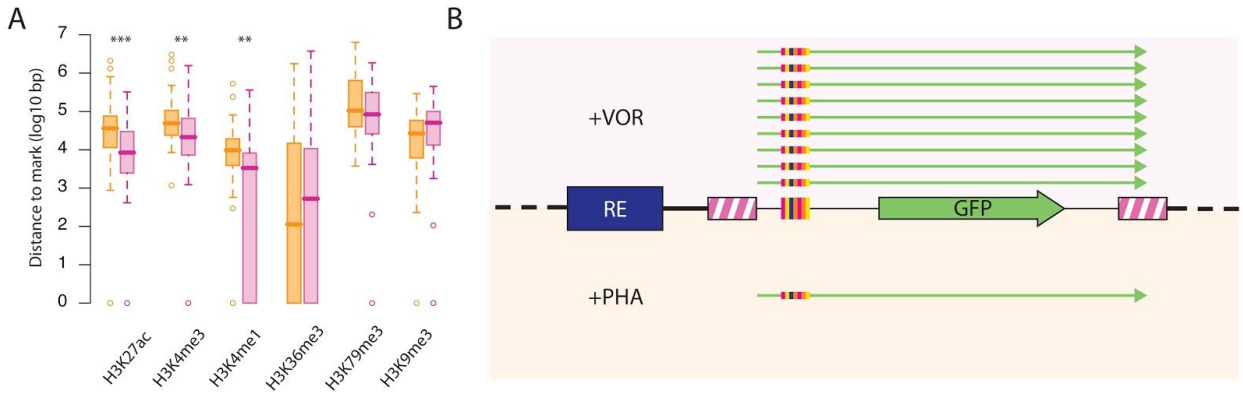


Figure 6. Model for the differential effect of latency reversing agents. *(A) The box plots show the distance to closest histone marks of proviruses among the top 15% PHA (orange) and top 15% VOR responders (pink). The stars indicate the significance of the Wilcoxon test (HeK37ac $P = 1.8\ 10^{-4}$, H3K4me3 $P = 2.6\ 10^{-3}$ and H3K4me1 $P = 3.2\ 10^{-3}$). (B) Graphical summary. Latent proviruses that respond more to VOR than to PHA are located closer to endogenous regulatory elements. VOR may affect the activity or the targeting of regulatory element, inducing proviruses inserted in the vicinity.*

The strong bias of HIV insertions suggests that the virus has evolved to select the most appropriate sites for the infection cycle. Surprisingly, the preferential targeting of chromosomes 16, 17 and 19 does not seem to bear a relation with the expression of the provirus (Figure 3D). It is likely that those chromosomes are simply more accessible to the provirus, either because of the state of their chromatin, or because of their proximity to the nuclear pore[17,32]. On the other hand, we observed that HIV also tends to target enhancers (Figure 2C), where it is expressed at higher levels (Figure 3D). It is possible that this enrichment has been somehow optimized by natural selection, but the potential evolutionary advantages remain speculative.

It is clear from previous work on chromatin, and in particular on position effect variegation[9] that the genomic context can have a strong influence on gene expression. The possibility that chromatin may play a role in latency was raised previously[14], but the theory was not significantly developed before this work. Here we observed fewer latent HIV proviruses near the enhancers of the host (Figure 4C). At least two mechanisms may account for this observation. In the first, local enhancers may suppress latency by continuously stimulating the expression of HIV. In the

second, silencing mechanisms similar to those occasionally observed in cancer[33] may take place away from enhancers. Our results support the view that HIV can be silenced by the genome of the host (Figure 5B), but they do not exclude the first mechanism. It is also important to highlight that HIV can be latent when inserted in the vicinity of enhancers and that the insertion patterns of latent and non-latent viruses are overall similar (Figure 4C). Thus the decision to enter latency cannot be reduced to a single factor such as the distance to the nearest enhancer.

Intriguingly, our results show that only 3% of the viruses can be reactivated under maximum stimulation (Figure 5B), while 12% of the cells are infected (Figure 5C). Part of the discrepancy may be due to sensitivity issues: the GFP in the HIV construct is translated via an IRES (Figure 1A) which typically produce moderate amounts of protein. This discrepancy also suggests that some latent proviruses in this model system are non-inducible, and the question whether they are replication-competent still needs to be addressed. Considering that latent proviruses tend to be located further away from enhancers (Figure 4D), it is possible that several of them are out of reach from drugs such as VOR, which seem to be more efficient when the insertions are closer to enhancers.

It is still debated whether HIV latency is an evolutionary accident or an intrinsic regulatory process[34–36]. There is substantial evidence that HIV has evolved autonomous mechanisms to shut down its own expression upon infection[8,37,38], however, this does not prevent the chromatin of the host from also having an effect. Autonomous triggers of latency happen early upon infection[26], whereas our experimental setup focuses on latency occurring from 4 days post infection (Figure 4A). It is likely that HIV latency may result from multiple causes. It will be important to identify in the near future the mechanisms inducing the type of latency that resists current reactivation therapies.

Our results suggest that it is possible to predict the response of proviruses to different drugs based on the features at or near the insertion site. The classification is still partial, but we envision that the accumulation of ChIP-seq and B-HIVE data set will make it possible to fully predict and understand how latent HIV proviruses respond to different treatments.

One of the present limitations of B-HIVE is that it is carried out in cellular models. This work was based on experiments performed *in vitro*, on a single immortalized cell line. It will be important to establish whether our conclusions also hold *in vivo* and in the clinic. In its present form, B-HIVE can be applied to track latent infections in animal models, but it cannot be used for clinical samples. For this purpose, *in situ* barcoding techniques will be required, *i.e.* methods to barcode latent viruses integrated in patient cells. The steady progress of genome editing via CRISPR/Cas9 may make such technologies available in the future.

Finally, we identified for the first time the differential targets of two drugs used to reactivate HIV. This result naturally raises the question of the mechanism by which the same transcriptional unit, namely HIV, can have different responses to the same treatment. The most likely explanation is that several pathways can silence HIV and that each drug targets a different one. Chromatin silencing mechanisms are poorly understood and many distinct mechanisms can take place. We suggest that PHA and VOR have distinct effects on human regulatory elements, such that latent HIV inserted in their vicinity would be more activated by VOR (Figure 6). More generally, our results suggest that cocktails will be more efficient than single drugs. Previous studies came to the same conclusion[39], but here we advocate the use of complementary drugs, which do not have to show synergistic effects. In this respect, B-HIVE paves the way to develop combinations of drugs with wide reactivation spectra. In summary, we foresee that the B-HIVE technology and its derivatives will help to further understand the processes governing HIV latency and contribute to develop more efficient eradication strategies.

ACKNOWLEDGEMENTS

We would like to thank Dr. Albert Jordan for providing research material, Dr. Petrus Stienen and Dr. Lucas Carey for their critical feedback on the manuscript, and the Sequencing core facility of the CRG for their technical support. Finally, we would like to thank the anonymous reviewers for their significant contribution to the manuscript. This research was supported by the Government of Catalonia and the Spanish Ministry of Economy and Competitiveness (Plan Nacional BFU2012-37168, Centro de Excelencia Severo Ochoa 2013–2017 SEV-2012-0208). JM and AM were supported by a grant from the Spanish Ministry of Economy and Competitiveness and FEDER (SAF2013-46077-R). EZ and GF are supported by the European Research Council (Synergy Grant 609989).

AUTHOR CONTRIBUTIONS

Conceptualization, G.F.; Methodology, H.C. and J.M.; Software, G.F. and E.Z., Formal Analysis, G.F., Investigation, G.F., H.C., E.Z., and J.M.; Resources, G.F.; Data curation, E.Z, Writing – Original Draft, G.F., H.C., J.M.; Writing – Review & Editing, E.Z. and A.M.; Visualization, G.F., H.C. and E.Z.; Supervision, G.F., Project Administration, G.F. and A.M.; Funding Acquisition, G.F. and A.M.

REFERENCES

1. Chun, T. W. *et al.* Relationship between pre-existing viral reservoirs and the re-emergence of plasma viremia after discontinuation of highly active anti-retroviral therapy. *Nat. Med.* **6,** 757–761 (2000).
2. Davey, R. T., Jr *et al.* HIV-1 and T cell dynamics after interruption of highly active antiretroviral therapy (HAART) in patients with a history of sustained viral suppression. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 15109–15114 (1999).
3. Durand, C. M., Blankson, J. N. & Siliciano, R. F. Developing strategies for HIV-1 eradication. *Trends Immunol.* **33,** 554–562 (2012).
4. Rosenberg, E. S. *et al.* Immune control of HIV-1 after early treatment of acute infection. *Nature* **407,** 523–526 (2000).
5. Deeks, S. G. HIV: Shock and kill. *Nature* **487,** 439–440 (2012).
6. Eriksson, S. *et al.* Comparative analysis of measures of viral reservoirs in HIV-1 eradication studies. *PLoS Pathog.* **9,** e1003174 (2013).
7. Chun, T. W. *et al.* Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387,** 183–188 (1997).
8. Ho, Y.-C. *et al.* Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155,** 540–551 (2013).
9. Muller, H. J. Types of visible variations induced by X-rays inDrosophila. *J. Genet.* **22,** 299–334 (1930).
10. Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406,** 593–599 (2000).
11. Kellum, R. & Schedl, P. A position-effect assay for boundaries of higher order chromosomal domains. *Cell* **64,** 941–950 (1991).

12. Akhtar, W. *et al.* Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel. *Cell* **154,** 914–927 (2013).
13. Jordan, A., Defechereux, P. & Verdin, E. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *EMBO J.* **20,** 1726–1738 (2001).
14. Jordan, A., Bisgrove, D. & Verdin, E. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *EMBO J.* **22,** 1868–1877 (2003).
15. Schröder, A. R. W. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110,** 521–529 (2002).
16. Lewinski, M. K. *et al.* Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J. Virol.* **79,** 6610–6619 (2005).
17. Marini, B. *et al.* Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521,** 227–231 (2015).
18. Archin, N. M. *et al.* Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature* **487,** 482–485 (2012).
19. Wang, Z. *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138,** 1019–1031 (2009).
20. Sarzotti-Kelsoe, M. *et al.* Optimization and validation of the TZM-bl assay for standardized assessments of neutralizing antibodies against HIV-1. *J. Immunol. Methods* **409,** 131–146 (2014).
21. Ochman, H., Gerber, A. S. & Hartl, D. L. Genetic applications of an inverse polymerase chain reaction. *Genetics* **120,** 621–623 (1988).
22. Triglia, T., Peterson, M. G. & Kemp, D. J. A procedure for in vitro amplification of DNA segments that lie outside the boundaries of known sequences. *Nucleic Acids Res.* **16,** 8186

(1988).

23. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* **17,** 1186–1194 (2007).
24. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13,** 613–626 (2012).
25. Meyerhans, A., Breinig, T., Vartanian, J.-P. & Wain-Hobson, S. HIV Sequence Compendium 2003. 14–21 (2003).
26. Matsuda, Y. *et al.* Epigenetic heterogeneity in HIV-1 latency establishment. *Sci. Rep.* **5,** 7701 (2015).
27. Spina, C. A. *et al.* An in-depth comparison of latent HIV-1 reactivation in multiple cell model systems and resting CD4+ T cells from aviremic patients. *PLoS Pathog.* **9,** e1003834 (2013).
28. Bullen, C. K., Laird, G. M., Durand, C. M., Siliciano, J. D. & Siliciano, R. F. New ex vivo approaches distinguish effective and ineffective single agents for reversing HIV-1 latency in vivo. *Nat. Med.* **20,** 425–429 (2014).
29. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351,** 1454–1458 (2016).
30. Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346,** 1373–1377 (2014).
31. Reeder, J. E., Kwak, Y.-T., McNamara, R. P., Forst, C. V. & D'Orso, I. HIV Tat controls RNA Polymerase II and the epigenetic landscape to transcriptionally reprogram target immune cells. *Elife* **4,** (2015).
32. Dieudonné, M. *et al.* Transcriptional competence of the integrated HIV-1 provirus at the

nuclear periphery. *EMBO J.* **28,** 2231–2243 (2009).

33. Esteller, M. Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum. Mol. Genet.* **16 Spec No 1,** R50–9 (2007).
34. Rouzine, I. M., Weinberger, A. D. & Weinberger, L. S. An evolutionary role for HIV latency in enhancing viral transmission. *Cell* **160,** 1002–1012 (2015).
35. Dahabieh, M. S., Emilie, B. & Eric, V. Understanding HIV Latency: The Road to an HIV Cure. *Annu. Rev. Med.* **66,** 407–421 (2015).
36. Ruelas, D. S. & Greene, W. C. An integrated overview of HIV-1 latency. *Cell* **155,** 519–529 (2013).
37. Razooky, B. S., Pai, A., Aull, K., Rouzine, I. M. & Weinberger, L. S. A hardwired HIV latency program. *Cell* **160,** 990–1001 (2015).
38. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P. & Schaffer, D. V. Stochastic Gene Expression in a Lentiviral Positive-Feedback Loop: HIV-1 Tat Fluctuations Drive Phenotypic Diversity. *Cell* **122,** 169–182 (2005).
39. Laird, G. M. *et al.* Ex vivo analysis identifies effective HIV-1 latency-reversing drug combinations. *J. Clin. Invest.* **125,** 1901–1912 (2015).

ONLINE METHODS

Cell culture

The human Jurkat T cell line (obtained from the cell collection of the Center for Genomic Regulation, Barcelona) was grown at 37 ºC under a 95% air/5% CO_2 atmosphere, in RPMI 1640 medium (Gibco) supplemented with 10% fetal bovine serum (Gibco), 1% Pen Strep (Gibco) and 1% GlutaMAX (100x) (Gibco). Jurkat cells were passaged every 2 days with 1 to 5 dilution. HEK 293T cells were grown under the same conditions in Dulbecco's modified Eagle's medium (Gibco). Cells were yearly checked for mycoplasma.

Cloning of HIV-based vector and construction of barcoded HIV library

All the infection experiments described in this study were carried out with a single-round of replication minimal HIV construct derived from plasmid pEV731[14] kindly provided by A. Jordan (IBMB, Barcelona). pEV731 is an HIV-based vector coding for the two-exon form of the HIV-1 Tat gene and a GFP marker gene under the control of the HIV-1 LTR. The sequence of HIV was excised from pEV731 and cloned into a 2.9 kb backbone. The resulting vector pHCC1 was used to prepare libraries of viral vectors by barcoding PCR as detailed below (Figure S1).

The forward primer (GAT431) contains a 5' extension of 20 random nucleotides followed by the T7 promoter; the reverse primer (GAT432) contains a 5' extension that corresponds to the sequence of the Illumina PE1.0 primer (Figure S1A). The primers are 5'-phosporylated and anneal to the HIV sequence from pEV731 in divergent orientation. 100 pg of plasmid pHCC1 were used as template for PCR in a Phusion reaction mix (Thermo Fisher Scientific, F530S) in GC buffer complemented with 3% DMSO in the following cycling conditions: 98 ºC for 1 min // 98 ºC for 20 sec / 58 ºC for 1 min / 72 ºC for 10 min // - 24 cycles total / 72 ºC for 10 min. To digest the template, 1 µL 20,000 U/mL DpnI (NEB, R0176S) was added to the mix and the tubes were placed at 37 ºC for 30 min. The ends were made complementary and double-stranded by adding to the mix 1 µL 3,000 U/mL T4 DNA polymerase (NEB, M0203S) and placing the tubes at 12 ºC for 20 min. The reaction was stopped by adding 2 µL 0.5 M EDTA and placing the tubes at 65 ºC for 25 min. The products were purified by QIAquick Gel Extraction Kit (Qiagen) and self-ligated by T4 DNA ligase (Thermo Fisher Scientific, EL0013) with 5% PEG4000 at 22 ºC for 1 hour (Figure S1B). The ligated products (100-200 ng/µL) were precipitated by EtOH and resuspended in 11 µL distillated water. 20 µL ElectroMAX™ DH10B™

competent cells (Invitrogen, 18290015) were electroporated with 1 µL ligated products. Enough electroporations (typically 2-5) were carried out to obtain a final complexity over 1 million independent clones. A subsample of 1% of the electroporated bacteria was diluted 100 times and spread on ampicillin-containing agar plates to estimate the number of clones in the library; the rest was grown overnight in liquid medium and the plasmids were extracted the next day.

The complexity of the library estimated from the plating was ~1.3 million independent clones. To control the structure of the library (Figure S1C), 10 clones picked from the agar plate were analyzed by PCR (Figure S2A) and by Sanger sequencing (Figure S2B).

The primers used for barcoding PCR are listed in **Tables S1**.

Transfection and viral Infection

For viral particle preparation, two million HEK 293T cells in 10 cm dishes were transfected with 6.5 µg pCMVΔR8.91, 3.5 µg pVSV-G and either 10 µg pEV731 or 10 µg barcoded pHCC1. After 16 hours, the medium was replaced, and the supernatant containing viral particles was collected 48 and 72 hours post transfection. Transfection efficiency was validated by measuring the percentage of GFP(+) cells by FACS analysis (see section FACS sorting below). Viral stocks were titrated using the TZM-bl indicator cell line as described previously[20]. Briefly, 5-fold serial dilutions of virus stocks were titrated in quadruplicate in 100 µL culture medium in 96-well culture plates. 10,000 fresh TZM-bl cells were added to each well in 100 µL culture medium. After 48 hours of incubation, around half of the medium was removed from each well and replaced with an equal volume of the britelite Luc reporter gene assay system reagent (PerkinElmer, 6066769). After 2 min incubation, 150 µL of cell lysate were transferred to 96-well white solid plate (Corning-Costar) and luminescence was quantified using a Berthold Centro LB 960 luminometer (Berthold Technologies). Luciferase units were used to calculate $TCID_{50}$ (median tissue culture infective dose) with the Spearman & Kärber algorithm[40]. Jurkat cells were initially infected with a barcoded viral inoculum corresponding to an MOI around 0.5. Viral particles were used to infect 1 million Jurkat cells in a 6-well plate. The plate was centrifuged at 1,000 g at 32 ℃ for 90 min. The medium was replaced by 3 mL fresh RPMI 24 hours post infection. 48 hours post infection, the efficiency of infection was monitored by the FACS analysis (Figure S3).

FACS sorting

Cell infection levels were monitored by flow cytometry analysis of GFP expression with a FACSCalibur cell sorter (Becton Dickinson). GFP expression from the drug-treated GFP(-) cell fraction was measured at different days as indicated. Cells were washed and resuspended in phosphate-buffered saline (PBS). Viral infection efficiency was measured as the percentage of GFP(+) 48 hours post infection. Cells were sorted on day 4 position infection and on day 21 post infection (Figure S4) using a FACS Aria II-SORP (Becton Dickinson, San Jose, CA). Before sorting, cells were washed and resuspended in PBS containing 1 µL 4',6-diamidino-2-phenylindole (DAPI) at final concentration 1 µg/mL. Results shown throughout the manuscript correspond to representative data of experiments repeated at least three times.

Genomic DNA / RNA isolation and mRNA purification

Genomic DNA and total RNA from infected cell pools were extracted with the AllPrep DNA/RNA Mini Kit (Qiagen, 80284). The mRNA fraction was isolated from total RNA with the Oligotex mRNA Mini Kit (Qiagen, 70022). The mRNAs were eluted in 44 µL OEB buffer supplied with the kit.

Isolation of barcodes and HIV integration sites from genomic DNA

3 µg genomic DNA from infected Jurkat cells were digested by 2 µL 10 U/µL BpII (Thermo Fisher Scientific, ER1311) and 2 µL 5,000 U/mL HpyCH4III (NEB, R0618S) in buffer Tango complemented with 1X SAM in 50 µL final volume at 37 ℃ for 3 hours. The reaction was stopped by placing the tubes at 65 ℃ for 20 min. Since BpII and HpyCH4III ends are in general not compatible, fragments were blunt-ended by diluting the mix in 1X T4 DNA ligase buffer, adding 3.3 µL 10 mM dNTPs (Thermo Fisher Scientific, R0181), 4.2 µL 3,000 U/mL T4 DNA polymerase (NEB, M0203S), 0.8 µL 5,000 U/mL DNA polymerase I, klenow fragment (NEB, M0210S) and 4.2 µL 10,000 U/mL T4 polynucleotide kinase (NEB, M0201S) in 100 µL final volume at 12 ℃ for 20 min. The reaction was stopped by adding 2 µL 0.5 M EDTA and placing the tube at 65 ℃ for 20 min. The blunt-end products were diluted in 1 mL T4 DNA ligase buffer and self-ligated by adding 2 µL 30 U/µL T4 DNA ligase (Thermo Fisher Scientific, EL0013) and placing the tubes at 16 ℃ overnight. The ligation reaction was precipitated by EtOH the following day. The pellet was resuspended in 34 µL distilled water. To enhance the inverse PCR, the ligation products were linearized by cutting the HIV LTR with SacI. 2 µL 20,000 U/mL

SacI (NEB, R0156S) were added in 40 µL final volume and the tubes were placed at 37 °C for 3 hours. The reaction was stopped by placing the tubes at 65°C for 20 min.

Insertion sites were recovered from SacI-digested product by nested PCR. 8 µL SacI-digestion products were mixed in 50 µL standard Phusion polymerase reaction mix (Thermo Fisher Scientific, F530S) in GC buffer, with 0.1 µM primers GAT316 and GAT645 (annealing to the Illumina PE1.0 primer and to the U3 region of LTR, respectively). The cycling condition were as follows: 98 °C for 1 min // 98 °C for 20 sec / 65 °C for 1 min / 72 °C for 5 min // - 10 cycles / 72 °C for 5 min. For the second round of nested PCR, 10 µL were diluted in 50 µL of standard Phusion polymerase reaction mix in GC buffer with 0.1 µM primer GAT024 and an indexing primer GAT-int (annealing to the Illumina PE1.0 primer and to the 5' end of the LTR, respectively). The cycling condition were as follows: 98 °C for 1 min // 98 °C for 20 sec / 55 °C for 1 min / 72 °C for 5 min // - 2 cycles / 98 °C for 20 sec / 65 °C for 1 min / 72 °C for 5 min // - 15 cycles / 72 °C for 5 min. The reverse primers (an indexing primer GAT-int) of the second round of PCR adapt an index and the PE2.0 to the amplicons. The products contain the barcode and the human DNA flanking the HIV genome.

PCR products run as a smear on agarose gel (Figure S5B lane 12). The smears were specific, as they failed to appear when the cells were not infected, infected with barcode-less viruses or when no ligation was performed (Figure S5B). To confirm the structure of the amplicon, a smear obtained after the first round of nested PCR was cloned and individual clones were sequenced (Figure S5C).

4 different indexing primers GAT-int (GAT730, GAT731, GAT732 and GAT742) containing a 6 nucleotide index are used in this study. The primers and indices used are listed in **Tables S1** and **S2**, respectively.

Amplification of RNA and DNA barcodes for high throughput sequencing

Reverse transcription was performed on 10 µL purified mRNA, to which was added 1 µL 20 µM reverse primer GAT526 (annealing downstream of the T7 promoter) and 1 µL 10 mM dNTPs (Thermo Fisher Scientific, R0181). RNA was denatured at 95 °C for 1 minute and then placed on ice. 8 µL master mix containing 4 µL 5x cDNA synthesis buffer, 1 µL 0.1 M DTT, 1 µL 40 U/µL RNaseOUT™, 1 µL DEPC-treated water and 1 µL 15 U/µL ThermoScript™ (reagents

included in the ThermoScript™ RT-PCR System; Invitrogen, 11145-024) were added to the denatured RNA and the tube was placed at 65 °C for 1 hour. The reaction was stopped placing the tube at 85 °C for 5 minutes. 5 µL RT product were used as template in 50 µL standard Phusion polymerase reaction mix (Thermo Fisher Scientific, F530S) in GC buffer, with 1 µM primers GAT024 (annealing on the Illumina PE1.0 primer) and a barcode-specific GAT-bcd-amp primer (annealing on the T7 promoter). The cycling conditions were as follows: 98 °C for 1 min // 98 °C for 20 sec / 60 °C for 30 sec / 72 °C for 1 min // - 27 cycles / 72 °C for 5 min. The reverse primers adapt a 4 nucleotide index to the amplicons.

For DNA barcodes, 200 ng genomic DNA from infected Jurkat cells (representing the genome of approximately 20,000 cells) were added to a 50 µL PCR reaction mix identical to the one described above for RNA barcodes. The cycling condition were as follows: 98 °C for 1 min // 98 °C for 20 sec / 58 °C for 30 sec / 72 °C for 1 min // - 29 cycles / 72 °C for 5 min. For every condition, at least 5 tubes were pooled before sequencing (representing the genome of more than 100,000 cells).

24 different barcode-specific GAT-bcd-amp primers containing a 4 nucleotide index are used in this study. The primers and indices used are listed in **Tables S1** and **S2**, respectively.

Quantitative PCR (qPCR)

qPCR reaction mixes contained either 1 µL of Jurkat genomic DNA diluted 10 times or 2 µL of cDNA diluted 10 times, 5 µL 2X Power SYBR Green PCR Master Mix (Applied Biosystems, 4367659), 0.3 µL forward and reverse primers (300 nM final) in 10 µL final volume. The reactions were carried out in 384-well plates using the following cycling conditions: 95 °C for 10 min // 95 °C for 15 sec / 60 °C for 1.5 min // - 40 cycles. Primers GAT768 and GAT769 (annealing to GFP) were used to quantify the number of viral copies in infected cells. Primers GAT540 and GAT551 (annealing to the 5' LTR) were used to measure relative RNA expression after the drug treatment. Primers GAT750 and GAT751 (annealing to the human actin gene) and primers GAT1145 and GAT1146 (annealing to *hHBB*, the human hemoglobin subunit beta gene) were used as internal references. The primer sequence is listed in **Table S1**.

Drug treatment for provirus reactivation

GFP(-) Jurkat cells cultured in 10 mL RPMI 1640 medium were treated with 2% v/v phytohaemagglutinin (PHA, Gibco) or with 1 μM Vorinostat histone deacetylase inhibitor (VOR, suberoylanilide hydroxamic acid, also known as SAHA, Selleckchem) for 24 hours[41]. DMSO was used as a negative control at a final concentration of 0.1% v/v.

Sequencing library preparation

The quality of each sequencing library was visualized on a Bioanalyzer (Agilent Technologies) and quantified by qPCR using the KAPA library quantification kit (KAPABIOSYSTEMS, KK4835). The concentration of each library was calculated based on the formula provided in the KAPA library quantification kit. Libraries sequenced in the same lane were pooled together in the final concentration of 4 nM for high throughput sequencing. Mapping samples were sequenced as 76 bp paired end reads on a NextSeq sequencer (Illumina); expression and normalization samples were sequenced as 50 bp single read reads on a HiSeq2000 sequencer (Illumina).

Template switch T7 PCR and confirmation of provirus expression

2 μg genomic DNA from infected Jurkat cells were used for T7 promoter-driven *in vitro* transcription to obtain single-strand RNA (ssRNA). Genomic DNA was first digested in 20 μL final volume by 2 μL 5,000 U/mL HpyCH4III (NEB, R0618S) at 37 ℃ for 3 hours. The reaction was stopped by placing the tubes at 65 ℃ for 20 minutes. 2 μL were loaded on a gel for analysis, to the remaining 18 μL were added 2 μL T7 RNA Polymerase Mix (NEB, E2050S) and 10 μL NTP Buffer Mix (NEB, E2050S). The tubes were placed at 37 ℃ overnight. The reaction was stopped by adding 2 μL 10 mM $CaCl_2$ (Sigma-Aldrich, C4901-100G) and 1 μL DNaseI (NEB, M0303S) at 37 ℃ for 30 minutes, followed by adding 2 μL 0.5 M EDTA and placing the tubes at 70 ℃ for 10 minutes. ssRNA products were validated by RT-PCR with primers inside the minimal HIV 5'LTR region (Figure S6B left panel).

The template switch T7 PCR protocol was performed according to a previous report[42] with modifications (see Figure S6A for detail). 2 μL ssRNA were used for template switch T7 PCR. The reaction mix contained 1 μL 20 μM reverse primer GAT551 (annealing on the LTR), 1 μL 10 mM dNTPs (Thermo Fisher Scientific, R0181) and 8 μL 100 μM template switch oligonucleotide GAT997 in final volume 12 μL. ssRNAs were denatured at 65 ℃ for 5 minutes and then placed on ice. 7 μL master mix containing 4 μL 5X First-Strand Buffer (Invitrogen,

18064-022), 2 µL 0.1 M DTT (Invitrogen, 18064-022) and 1 µL 40 U/µL RNasin® Plus RNAse Inhibitor (Promega, N2611) were then added to the reaction and the tubes were placed at 42 °C for 2 minutes. 1 µL 200 U/µL SuperScript™ II reverse transcriptase (Invitrogen, 18064-022) was added and the tubes were further incubated at 42 °C for 90 minutes. The reaction was stopped by placing the tubes at 70 °C for 10 minutes. cDNAs amplified by template switch were validated by PCR showing a specific band at 105 bp (Figure S6B right panel).

The template switch RNA-DNA oligonucleotide used in this protocol contains three dGs at the 3' end. The dGs hybridize to dCs added to the 3' end of the cDNA strand newly synthesized by the reverse transcriptase. The reverse transcriptase then elongates the cDNA using the sequence of the oligonucleotide as a template. No PCR products were observed in the control samples where the template switch oligonucleotide did not contain the three dGs at the 3' end (data not shown).

To validate expression measurements obtained from B-HIVE, qPCR was performed with primers annealing to the partial sequence of selected barcodes (Figure S6C). The sequence of the chosen barcodes are shown in Figure S6C. Primers used for qPCR are listed in **Table S1**.

Estimating the frequency of barcode duplications

The complexity of the barcoded library was estimated at 1.3 million clones (see paragraph Cloning of HIV-based vector and construction of barcoded HIV library). Since 20,000 cells are FACS-sorted after infection, as many barcodes are drawn from a pool of 1.3 million. On average, a barcode is drawn 20,000 / 13,000,000 = 0.0015 times. This is a rare event, well described by the Poisson distribution. We can thus estimate the probability that a barcode is drawn more than one time as Prob(X > 1), where X has a Poisson distribution with mean 0.0015. This probability is equal to $1.12\ 10^{6}$, so duplications are negligible in the conditions of infection used here.

Data and bioinformatic analyses

HIV barcodes were extracted from paired-end reads through an inexact search of the T7 promoter sequence (TATAGTGAGTCGTA) allowing up to 3 errors (mismatches, insertions and deletions) using Seeq v1.1.2 (http://github.com/ezorita/seeq). The nucleotides upstream of the hit constitute the barcode. Each barcode was paired with the associated reverse read (where

the insertion site is sequenced), and reads without hit were discarded. Sequencing errors in the barcodes were reverted by sequence clustering using Starcode v1.0[43] allowing up to one mismatch and using the 'message passing' clustering algorithm. The reverse reads were mapped on GRCh37/hg19 with BWA-MEM 0.7.5a-r405[44] with default parameters and minimum mapping quality 20. The sequence of pHCC1 was added to the BWA index in order to identify sequencing and mapping artifacts (reads mapping to pHCC1 were discarded).

Candidate HIV insertion sites were considered identical if they were mapped within 100 bp of each other. To make unique barcode-locus assignments, we gave a confidence score to each pair of barcode and insertion site. The score of a pair was defined as the number of reads where the pair was found, divided by the highest of either (i) the number of reads where the barcode was found, or (ii) the number number of reads where the insertion site was found. A barcode was assigned to a location if and only if the score for this pair was higher than 0.90, *i.e.* if > 90% of the reads where the barcode was found were paired to the given location, and > 90% of the reads where the insertion site was found were paired with the given barcode. In practice, the score was recorded as -10 times the $\log_{10}$ of the complementary probability (like the mapping quality score of the SAM format), so that 10 was the minimum score to associate a barcode with a location.

RT-PCR on RNA barcodes and PCR on DNA barcodes contained a 4 nucleotide index in the read instead of the standard 6 nucleotide TruSeq Illumina index. We demultiplexed the reads through an inexact search of the constant region of the reads (TATAGTGAGTCGTATTAAAA) allowing up to 3 errors (mismatches, insertions and deletions) using Seeq. The barcodes correspond to the nucleotides upstream of the hit and the indices correspond to the four nucleotides downstream. Barcodes were clustered with Starcode allowing up to one mismatch and using the 'message passing' clustering algorithm, only exact hits to the indices were considered. Reads from technical replicates were added for each barcode, and the expression score of a given barcode was computed as the mean-centered $\log_{10}$ of the ratio between RNA counts and DNA counts.

For the analysis we used ChIP-seq profiles of H3K27ac, H3K9me3, H3K4me1, H3K4me3, H3K36me3 and H3K79me3[29–31] (GEO accessions GSM1697882, GSM1519634, GSM1519638, GSM1519642 and GSE65687), plus several mock ChIP-seq profiles (GEO accessions

GSM1697880 and GSE65687), all in Jurkat cells. The reads were mapped on GRCh37/hg19 with BWA-MEM 0.7.5a-r405 [44] with default parameters and minimum mapping quality 20. The targets were identified with Zerone v1.0 [45] using options -l and -c.99. Enhancers were defined as the enriched H3K27ac regions, yielding approximately 39,000 peaks.

The wild-type Jurkat mRNA-seq was prepared from 100 ng total RNA using the TruSeq Stranded mRNA Library Prep Kit (Illumina, RS-122-2101) according to manufacturer's instructions with 4.5 minutes fragmentation time. Libraries were sequenced on the HiSeq2000 sequencer (Illumina) with 50 bp single end using v3 sequencing chemistry. We mapped the reads to Ensembl Homo Sapiens cDNA, assembly GRCh37, release 75 using kallisto[46] with single-end mode (--single), sequence bias correction (--bias), fragment length of 300nt (-s300) and standard deviation of 100nt (-l100). The counts in transcripts per million were associated to the genes, and the count of the different isoforms were added-up, generating a total count per gene copy.

We divided the genome in 5 different regions: ‘active genes’, ‘silent genes’, ‘active promoters’, ‘enhancers’ and ‘intergenic’. Here, genes only refer to protein-coding genes. The division between active and silent genes was performed by cutting off at the 60% percentile of the transcript per million count (*i.e.* active genes were defined as the top 60% most expressed). The cut off value is equal to 0.68 transcripts per million. Active promoters were defined as the regions spanning 5,000 bp centered on the transcription start sites of active genes. Insertions were considered to be in the vicinity of an enhancer if their mapped location was within 2,500 bp of a H3K27ac mark. Insertions close to enhancers were classified in the enhancer category even if they were inserted inside a gene or a promoter. The rest of the genome was classified as intergenic.

mRNA barcodes obtained after drug treatment with DMSO, PHA and VOR were extracted as clustered as above. Barcodes that were present in all the experiments (across two independent infections) were considered PCR contaminants and were discarded. Only barcodes that had positive counts in at least 11 of the 12 replicate conditions were kept and the other were discarded.

To define the top 15% VOR and top 15% PHA responders, we computed a Student's t statistic (also called the effect size) for each barcode and ranked them based on this score. The score was computed as the t statistic from the 4 replicate values of mRNA barcode counts in PHA versus the 4 replicate values of mRNA barcode counts in VOR. If this value is high, all 4 replicate values in PHA are higher than those in VOR and *vice versa*. Barcodes with extreme scores represent insertions that most specifically respond to one drug or to the other.

The raw data of this study is available from GEO (accession number GSE82061). We documented all data processing steps in a Docker virtual machine where all analyses and figures can be reproduced. The Docker image also provides the source code used in this study. The B-HIVE Docker image is available for public download at http://hub.docker.com/r/gui11aume/bhive/.

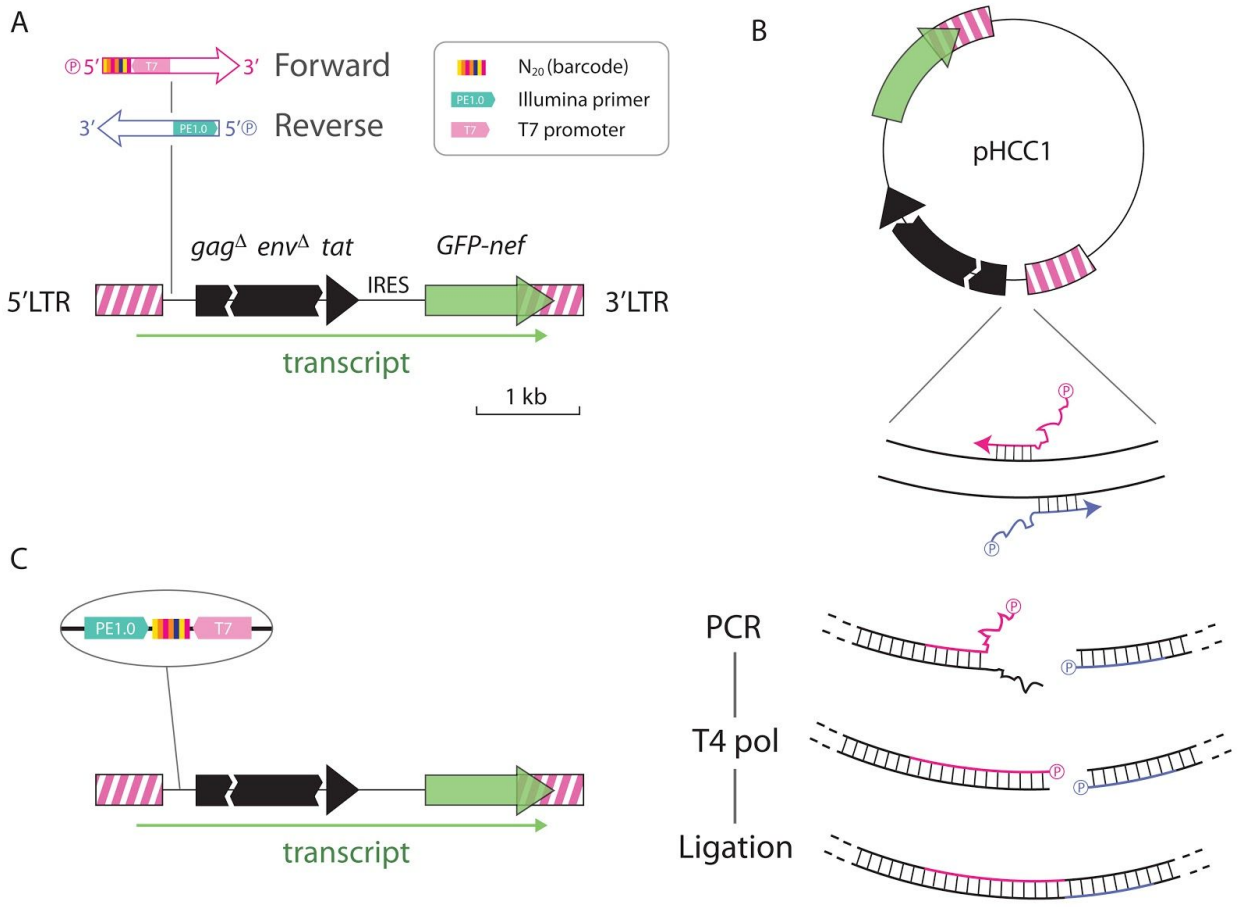
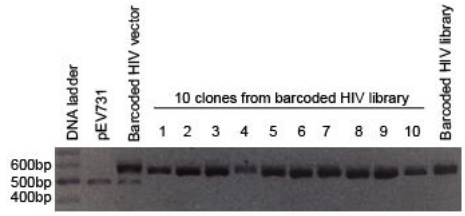


Figure S1. Barcoding PCR.

(A) Structure of the primers with their position on the construct. The forward primer contains a stretch of 20 random nucleotides at the 5' end that will constitute the barcode, followed by a T7 promoter. The reverse primer contains the sequence of the Illumina sequencing primer PE1.0 at the 5' end. Both primers are phosphorylated to allow the ligation of the PCR products. (B) Schematic of the reaction. The divergent primers amplify the whole backbone. Because the random nucleotides are different between primers, half of the products contain a non-double stranded end. The T4 polymerase removes the 3' overhangs and generates blunt products by using the barcode as a template. Finally, the blunt products are self ligated. (C) Structure of the HIV construct after the barcoding PCR. The net result is the addition of a 20 nucleotide barcode that is unique for every ligation product, flanked by the Illumina primer PE1.0 and the T7 promoter. This design has two purposes: it allows sequencing PCR and RT-PCR products directly without additional steps of library preparation, and it also allows quantifying barcode abundance by T7 PCR (see Figure S6).

Figure S2 (Related to Figure 1B)

A



B

	T7 promoter	barcode	Illumina PE1.0 primer sequence
Plasmid seq	TCTGGTTTCCCTTTCGCTTTTAATACGACTCACTATANNNNNNNNNNNNNNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCAAGTCCCTGTTCGGGCGCC		
clone #1	TCTGGTTTCCCTTTCGCTTTTAATACGACTCACTATAAAAACAAGGGACATTCACCG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCAAGTCCCTGTTCGGGCGCC	
clone #2	TCTGGTTTCCCTTTCGCTTTTAATACGACTCACTATATTCGACGAAAAAAGGGCCGG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCAAGTCCCTGTTCGGGCGCC	
clone #3	TCTGGTTTCCCTTTCGCTTTTAATACGACTCACTATAGACCATACTGTTAACAGAGC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCAAGTCCCTGTTCGGGCGCC	
clone #4	----------		
clone #5	TCTGGTTTCCCTTTCGCTTTTAATACGACTCACTATACCGCAGGGGCTACCCCAAGC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCAAGTCCCTGTTCGGGCGCC	
clone #6	TCTGGTTTCCCTTTCGCTTTTAATACGACTCACTATAGCTGTCAAATGGTAGTTAGC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCAAGTCCCTGTTCGGGCGCC	
clone #7	TCTGGTTTCCCTTTCGCTTTTAATACGACTCACTATAGTGGGACAGGCCTGGGATCC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCAAGTCCCTGTTCGGGCGCC	
clone #8	TCTGGTTTCCCTTTCGCTTTTAATACGACTCACTATACACACTCGACGCCACGCATG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCAAGTCCCTGTTCGGGCGCC	
clone #9	TCTGGTTTCCCTTTCGCTTTTAATACGACTCACTATAATTAATCGCCAAGCCCAAAC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCAAGTCCCTGTTCGGGCGCC	
clone #10	TCTGGTTTCCCTTTCGCTTTTAATACGACTCACTATAAGTGGACCATAACTAAGACC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCA AGTCCCTGTTCGGGCGCC	

Figure S2. Validation of barcoded HIV library.

(A) 10 clones randomly selected in a barcoded library were confirmed by PCR. Only the clones and barcoded HIV library containing an inserted product showed a 566 bp PCR product. (B) 10 clones were validated by Sanger sequencing. Each clone contained a T7 promoter (nucleotides with red background), a 17-22 nucleotide barcode and the sequence of the Illumina PE1.0 primer (nucleotides with blue background). Clone #4 could not be sequenced.

Figure S3 (Related to Figure 1B)

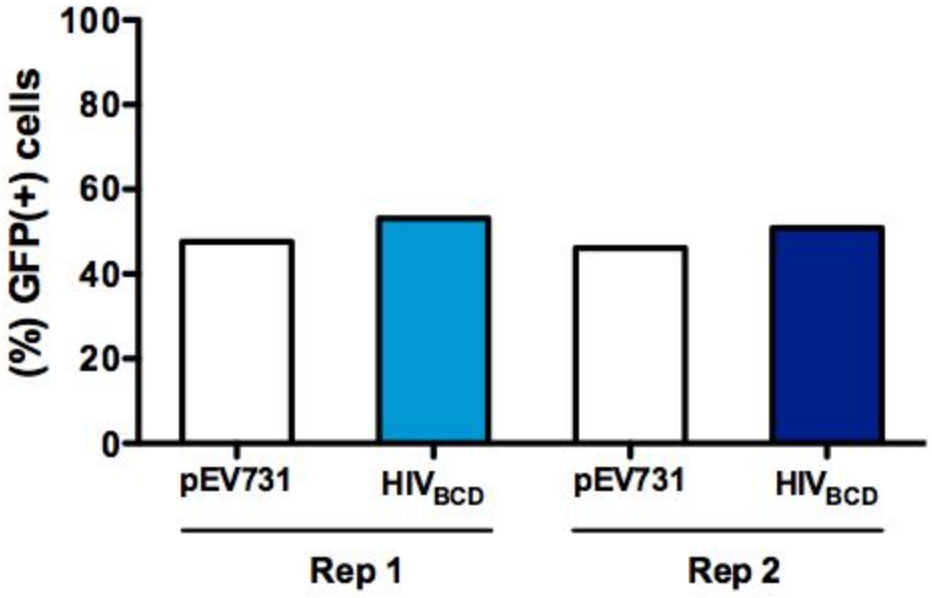


Figure S3. Efficiency of two independent HIV$_{BCD}$ infections in Jurkat cells.

Infection efficiency of non-barcoded minimal HIV construct derived from plasmid pEV731 (open bar) and barcoded HIV (HIV$_{BCD}$) in two independent infections (Rep1 and Rep2) in Jurkat cells were checked by measuring GFP expression 48 hours post infection by FACS analysis. The success of infection is similar with and without barcode.

Figure S4 (Related to Figure 1C and Figure 4A)

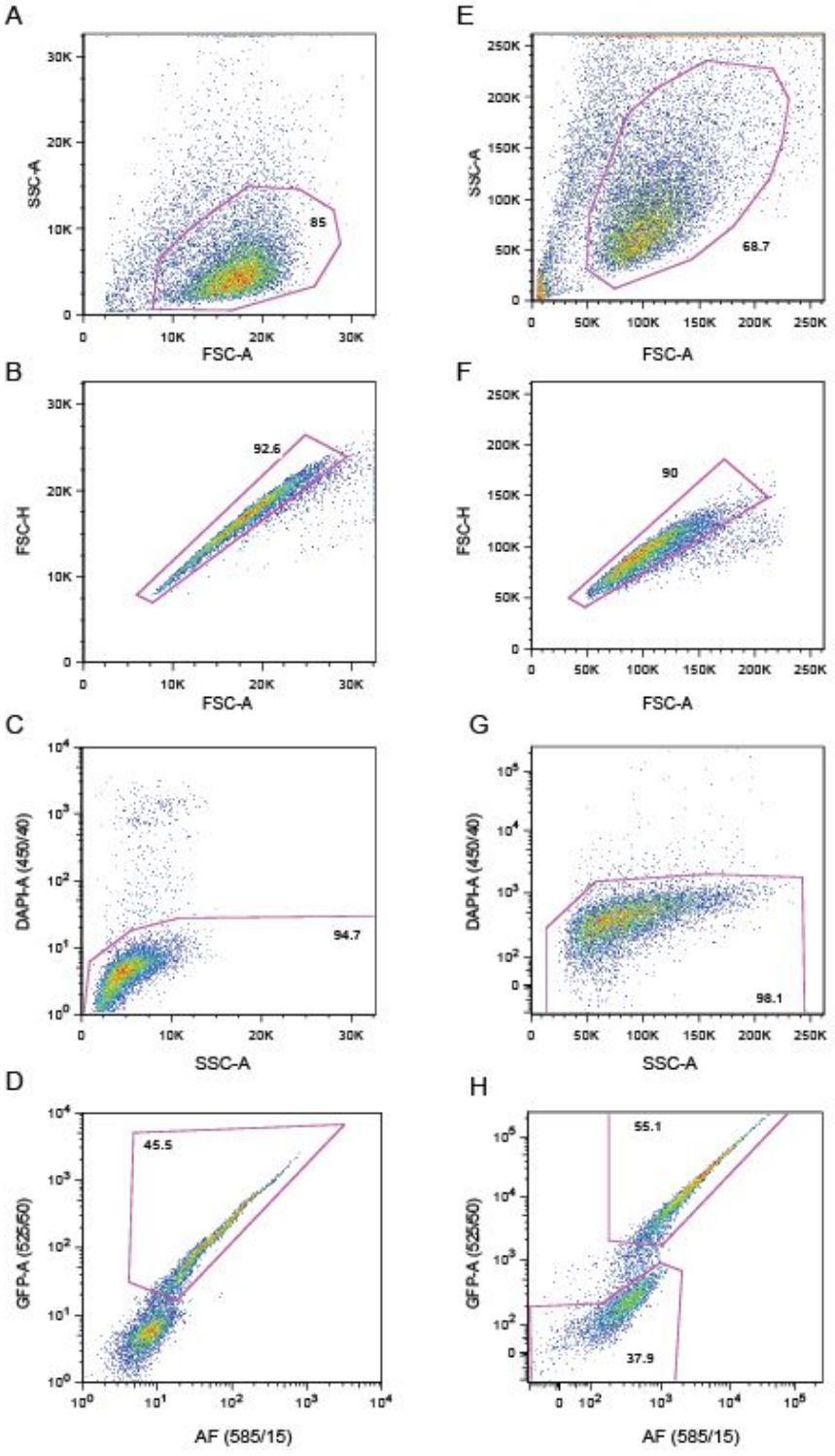


Figure S4. Isolation of GFP(+) and GFP(-) infected cells by FACS.

Representative FACS profiles for sorting GFP(+) cells 4 days post infection (A to D), and for separating GFP(+) from GFP(-) cells 21 days post infection (E to H). (A and E) Jurkat T cells were selected by FSC-A versus SSC-A (pink line-gated region). (B and F) Singlets were selected by FSC-A versus FSC-H (pink line-gated region). (C and G) Live cells (pink line-gated region) were selected by DAPI negative on a SSC-A versus DAPI-A (wavelength: 450/40). (D and H) GFP(+) cells (4 days post infection) as well as GFP(+) and GFP(-) cells (21 days post infection) were acquired by GFP fluorescence (wavelength: 525/50) versus autofluorescence (AF, wavelength: 585/15).

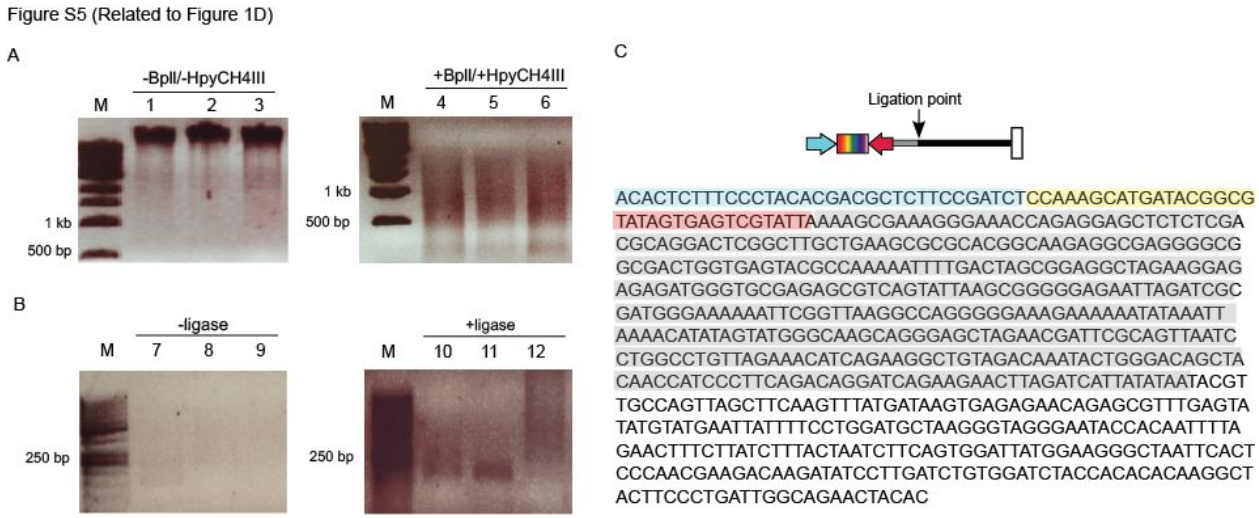


Figure S5. Validation of the mapping step of B-HIVE.

(A) Expected results from genomic DNA digested by BplI. The size of digested products is expected to lie between 0.5 and 1 kb. Lanes 1, 2 and 3: before BplI digestion; lanes 4, 5 and 6: after BplI digestion. Lanes 1 and 4: T cells only; lanes 2 and 5: genomic DNA from cells infected by wild type HIV; lanes 3 and 6: genomic DNA from cells infected by barcoded HIV. (B) Expected results from the preparation of samples from B-HIVE. 2.0% (wt/vol) agarose gel displaying a PCR smear corresponding to different integrations in cells infected by barcoded HIV (lane 12). Uninfected cells (lane 10), cells infected by non-barcoded HIV (lane 11) and no ligase controls (lanes 7, 8 and 9) did not yield any PCR product. (C) A PCR smear after the first inverse PCR was confirmed by Sanger sequencing. Nucleotides with blue background represent the Illumina sequencing primer. Nucleotides with red background represent the sequence of a T7 promoter. Nucleotides with yellow background represent a barcode. Nucleotides with grey background present sequence from a HIV vector. The sequence after the ligation point was confirmed to human DNA sequence from clone RP3-336H9 on chromosome 6p by NCBI BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi). White box: Partial of barcoded HIV 5'LTR; blue arrow: illumina universal TruSeq adaptor; multicolor box: barcode; red arrow: T7 promoter.

Figure S6 (Related to Figure 1E)

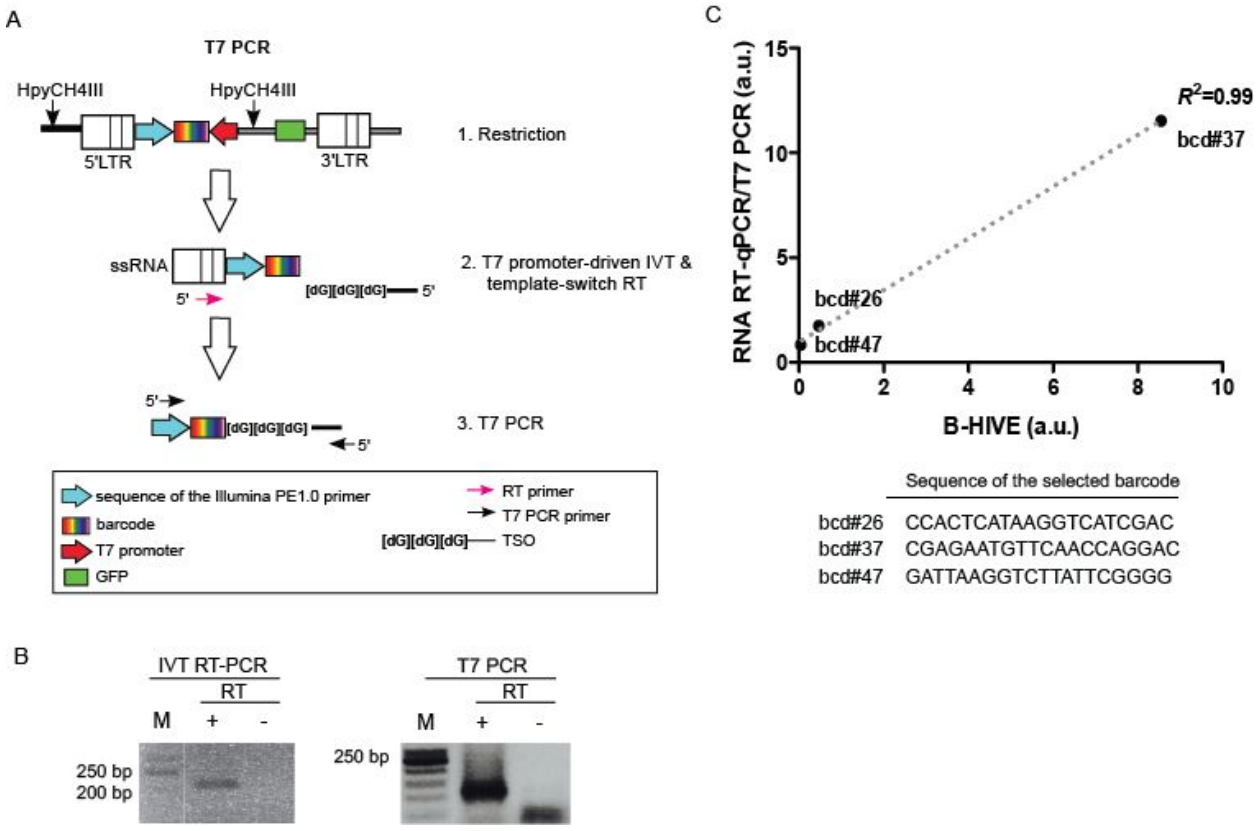


Figure S6. Template Switch T7 PCR

(A) Sketch of the key steps in template switch T7 PCR. 1. The 5-cutter restriction enzyme HpyCH4III was chosen to fragment genomic DNA between 500 bp to 1 kb. 2. *In vitro* ssRNA was transcribed from the T7 promoter inserted next to the barcode. Template switch reverse transcription was further performed with the RT primer annealed on the HIV 5'LTR together with the template switching oligonucleotide (TSO). 3. A primer annealing to Illumina PE1.0 and the TSO were used for DNA amplification. (B) *In vitro* transcribed ssRNA was validated by RT-PCR. Only the reaction containing reverse transcriptase yielded a correct 198 bp RT-PCR product; no PCR product was detected from the reaction without reverse transcriptase. Similarly, only the reaction containing reverse transcriptase yielded a correct 105 bp PCR product after T7 PCR was detected. (C) The scatter plot shows individual HIV expression measured by B-HIVE against expression measured by T7 PCR (a.u: arbitrary units).

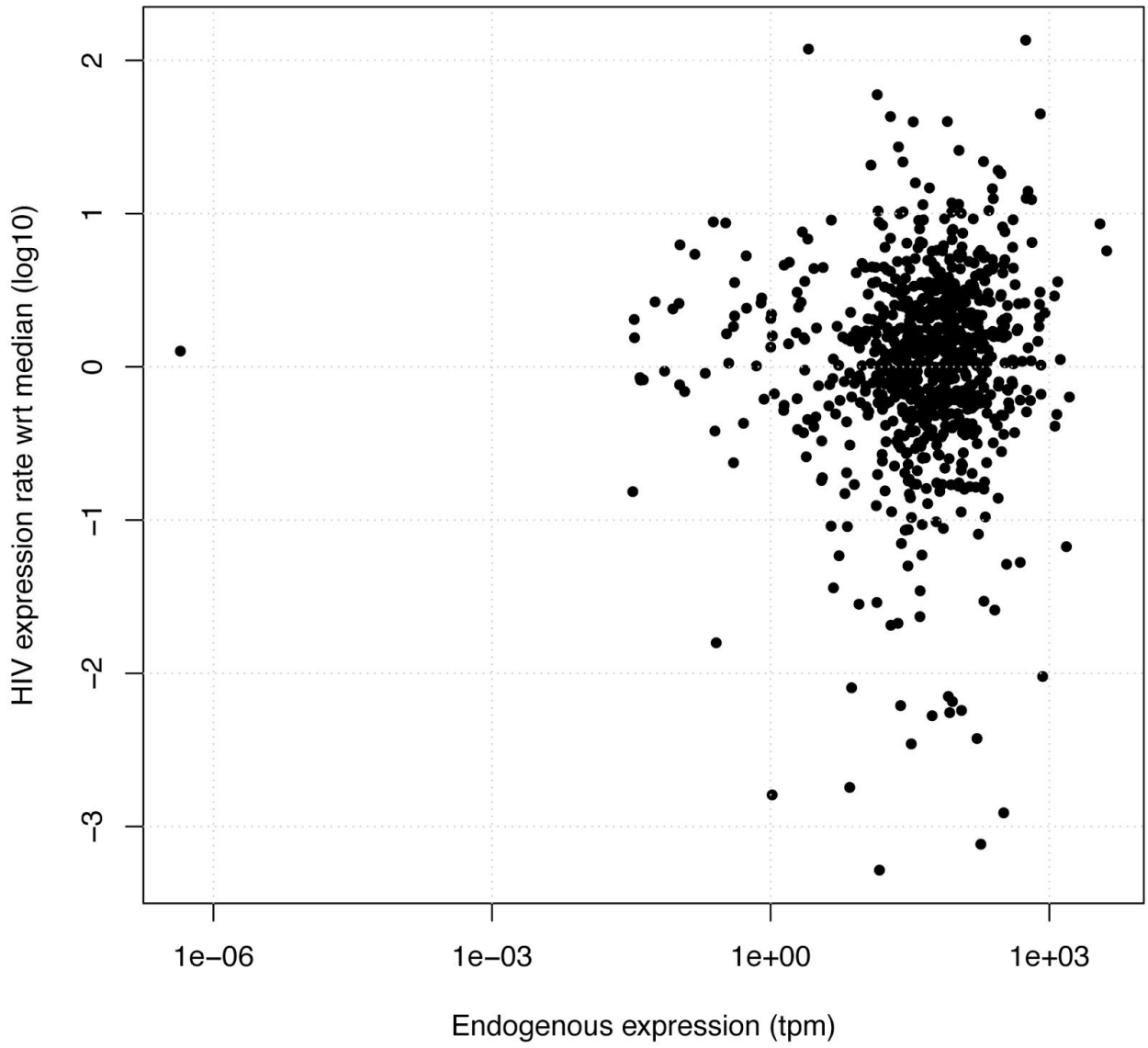
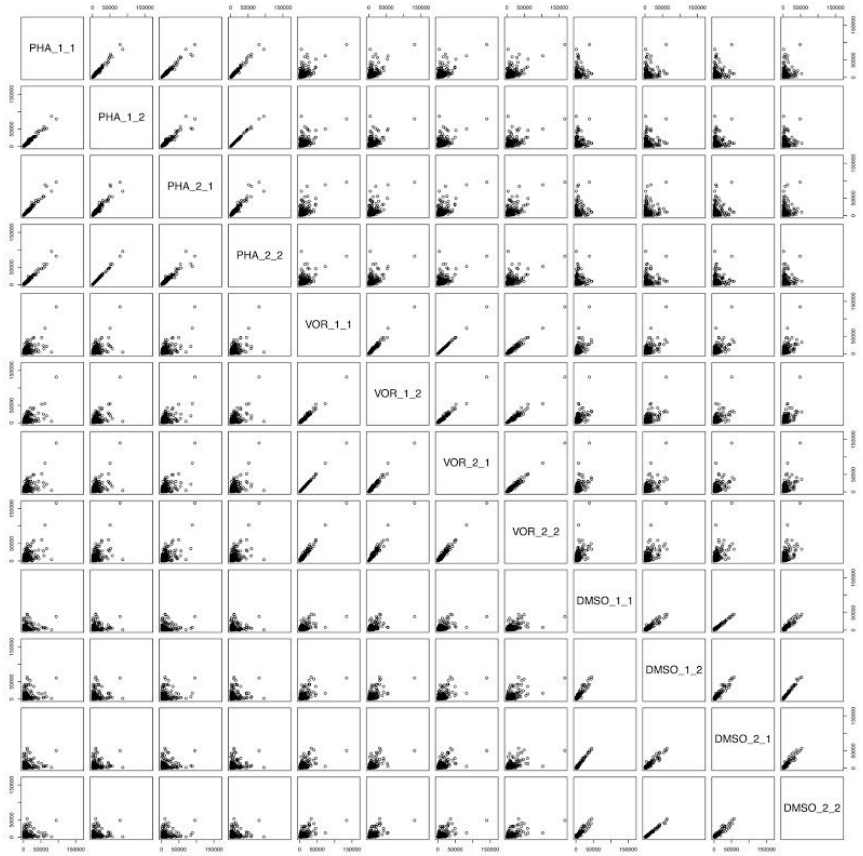


Figure S7. Expression of HIV versus expression of endogenous genes

Scatter plot where each dot represents a provirus inserted in a protein-coding gene. The X axis shows the expression of the gene measured in transcripts per million (tpm) and the Y axis shows the expression of barcoded HIV proviruses measured by B-HIVE. The expression of HIV does not show any dependence on the expression of the endogenous genes (Pearson correlation r = 0.07, P = 0.03).


PHA_1_1
PHA_1_2
PHA_2_1
PHA_2_2
VOR_1_1
VOR_1_2
VOR_2_1
VOR_2_2
DMSO_1_1
DMSO_1_2
DMSO_2_1
DMSO_2_2

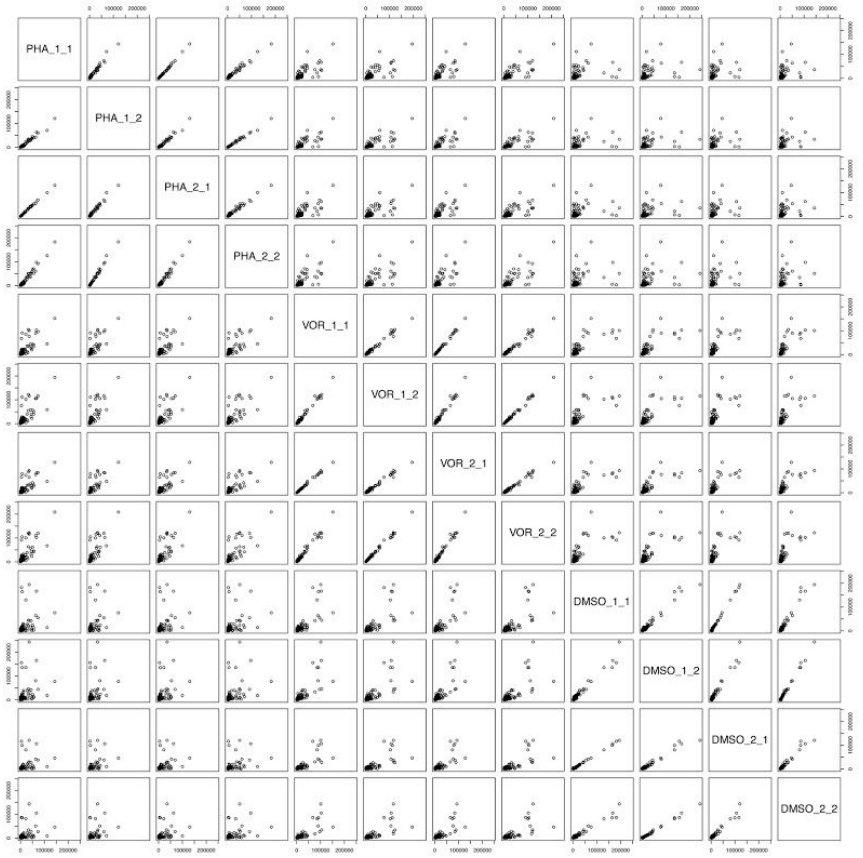

PHA_1_1
PHA_1_2
PHA_2_1
PHA_2_2
VOR_1_1
VOR_1_2
VOR_2_1
VOR_2_2
DMSO_1_1
DMSO_1_2
DMSO_2_1
DMSO_2_2

Figure S8. Comparison of drug activities on all HIV insertions.

Each panel is a scatter plot representing the mRNA tag counts of the same provirus (average of two replicates) in different conditions (PHA, VOR or DMSO). Top: first biological replicate; Bottom: second biological replicate. The expression of the proviruses is similar when they receive the same treatment, but it is variable when the treatment is different.

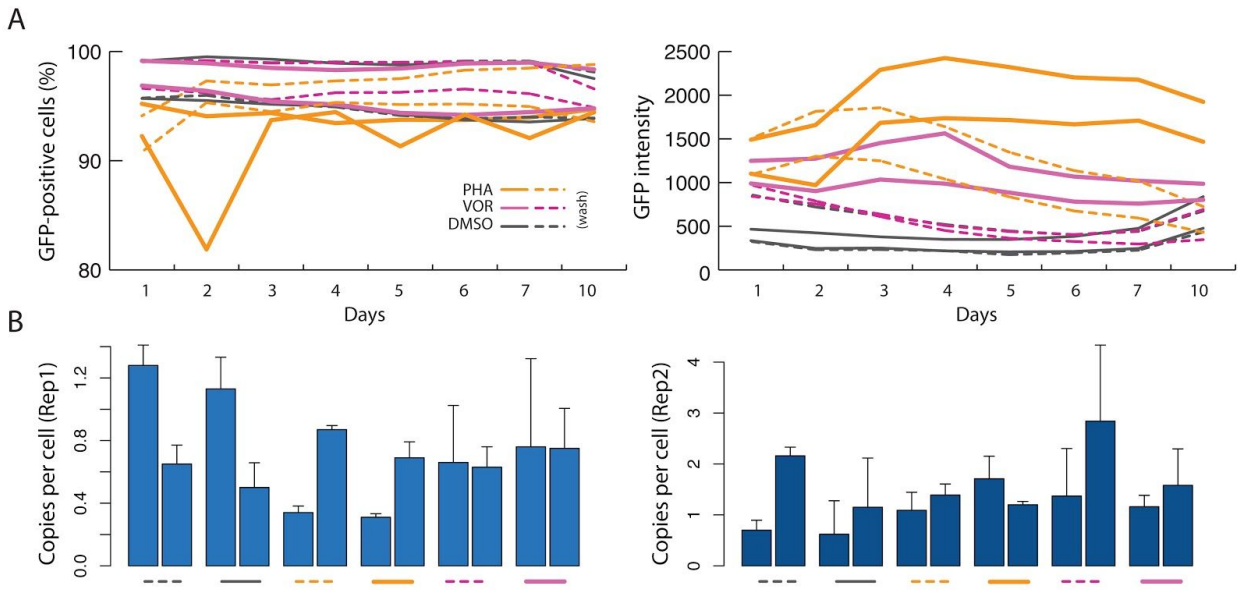


Figure S9. Drug treatment of the non-latent GFP(+) population.

(A) Left: Evolution of the percentage of GFP(+) cells upon PHA, VOR or control DMSO treatment. Dotted lines correspond to experiments where the drug is removed after 24 hours. The percentage remains high throughout the whole time period of 10 days. Right: fluorescence intensity of the GFP(+) cells. (B) Viral copy number estimated by qPCR, using HBB as a reference. The bars show the average of 4 measurements. The error bars show the estimated standard deviation of the measurements.

SUPPLEMENTARY REFERENCES

40. Hierholzer, J. C. & Killington, R. A. in *Virology Methods Manual* 25–46 (1996).
41. Dahabieh, M. S., Ooms, M., Simon, V. & Sadowski, I. A doubly fluorescent HIV-1 reporter shows that the majority of integrated HIV-1 is latent shortly after infection. *J. Virol.* **87,** 4716–4727 (2013).
42. Salimullah, M., Sakai, M., Mizuho, S., Plessy, C. & Carninci, P. NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb. Protoc.* **2011,** db.prot5559 (2011).
43. Zorita, E., Cuscó, P. & Filion, G. J. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31,** 1913–1919 (2015).
44. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
45. Cuscó, P. & Filion, G. J. Zerone: a ChIP-seq discretizer for multiple replicates with built-in quality control. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw336
46. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).