nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-12046-3 OPEN

# Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration

Bojana Lucic[1,11], Heng-Chang Chen[2,3,11], Maja Kuzman[4,11], Eduard Zorita[2,3,11], Julia Wegner[1,10], Vera Minneker[5], Wei Wang[6], Raffaele Fronza[6], Stefanie Laufs[6], Manfred Schmidt[6], Ralph Stadhouders[7,8], Vassilis Roukos[5], Kristian Vlahovicek[4], Guillaume J. Filion[2,3,9] & Marina Lusic[1]

HIV-1 recurrently targets active genes and integrates in the proximity of the nuclear pore compartment in CD4$^+$ T cells. However, the genomic features of these genes and the relevance of their transcriptional activity for HIV-1 integration have so far remained unclear. Here we show that recurrently targeted genes are proximal to super-enhancer genomic elements and that they cluster in specific spatial compartments of the T cell nucleus. We further show that these gene clusters acquire their location during the activation of T cells. The clustering of these genes along with their transcriptional activity are the major determinants of HIV-1 integration in T cells. Our results provide evidence of the relevance of the spatial compartmentalization of the genome for HIV-1 integration, thus further strengthening the role of nuclear architecture in viral infection.

[1] Department of Infectious Diseases, Integrative Virology, Heidelberg University Hospital and German Center for Infection Research, Heidelberg, Germany. [2] Genome Architecture, Gene Regulation, Stem Cells and Cancer Programme, Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain. [3] University Pompeu Fabra, Barcelona, Spain. [4] Bioinformatics Group, Division of Molecular Biology, Department of Biology, Faculty of Science, University of Zagreb, Zagreb, Croatia. [5] Institute of Molecular Biology (IMB), Mainz, Germany. [6] German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany. [7] Department of Pulmonary Medicine, Erasmus MC, Rotterdam, The Netherlands. [8] Department of Cell Biology, Erasmus MC, Rotterdam, The Netherlands. [9] Department of Biological Sciences, University of Toronto Scarborough, Toronto, ON, Canada. [10] Present address: Institute for Clinical Chemistry and Clinical Pharmacology, Universitätsklinikum Bonn, Bonn, Germany. [11] These authors contributed equally: Bojana Lucic, Heng-Chang Chen, Maja Kuzman, Eduard Zorita. [12] These authors jointly supervised this work: Kristian Vlahovicek, Guillaume J. Filion, Marina Lusic. Correspondence and requests for materials should be addressed to K.V. (email: kristian@bioinfo.hr) or to G.J.F. (email: guillaume.filion@gmail.com) or to M.L. (email: Marina.lusic@med.uni-heidelberg.de)

Integration of the proviral genome into the host chromosomal DNA is one of the defining features of retroviral replication[1–3]. Following integration, the viral genome can either be expressed or enter a transcriptionally dormant stage, establishing a reservoir of latently infected cells. Latently infected cells are indistinguishable from the non-infected ones and are therefore not eliminated by immune clearance mechanisms or recognized by current antiretroviral treatments[4,5]. Resting CD4$^+$ T cells of the memory phenotype represent the main reservoir of latent human immunodeficiency virus type 1 (HIV-1)[6]. However, it is still unclear how these reservoirs are established, as HIV-1 does not efficiently infect resting T cells due to different blocks at both pre-integration and integration levels[4,7–10]. One possible explanation is that some of the activated CD4$^+$ T cells revert back to the resting state upon infection with HIV-1, generating the reservoirs of silenced but replication-competent viruses[4]. What remains still to be defined is how this transition from activated to resting state occurs, and what changes in the cellular genome and chromatin are involved[11,12].

In activated CD4$^+$ T cells, the viral DNA enters the nucleus to access chromatin[13] passing through the nuclear pore complex (NPC)[14–16]. Nuclear pore proteins are important factors for the viral nuclear entry[17], as well as for the positioning and consequent integration of the viral DNA into the cellular genome[3,13–16,18,19]. Integration is not a random process, as HIV-1 predominantly integrates into active genes in gene-dense regions[20], mediated by the action of viral proteins integrase (IN) and capsid (CA). Through its interaction with LEDGF/p75[21–23], IN guides the integration into gene bodies. This pattern is shifted toward 5′ end regions of genes[22,24,25] or toward gene-poor regions[25] when LEDGF/p75 is depleted. Through its interaction with cleavage and polyadenylation specificity factor 6 (CPSF6), HIV-1 CA also contributes to the location of the viral genome[24,26,27]. Lack of CPSF6 arrests the incoming viral particles at the level of the NPC[27] or retargets the integrating viral DNA to the lamina-associated heterochromatin domains[26].

It is well established that HIV-1 targets open chromatin regions of active transcription and regions bearing enhancer marks[20,28,29]. Unlike typical enhancers, genomic elements known as super-enhancers (SEs) are defined by high levels of acetylated lysine 27 of histone 3 (H3K27ac) and binding of transcriptional co-activators, such as bromodomain-containing protein 4 (BRD4), the mediator complex[30], and the p300 histone acetyltransferase[30–32]. SEs control the expression of genes that define cell identity[30,32–34], and in case of CD4$^+$ T cells, relevant for HIV-1 infection, they control cytokines, cytokine receptors, and transcription factors regulating T cell-specific transcriptional programs[35]. Strikingly, one of the strongest immune-activation SEs[36,37] encoding for transcription factor BACH2 is among the most frequently targeted HIV-1 integration genes[38,39]. SE elements of cell identity genes were shown to be bound by nuclear pore proteins, which regulate their expression[40,41] and anchor them to the nuclear periphery[41]. Moreover, SEs seem to play a general role in organizing the genome through higher-order chromatin structures and architectural chromatin loops[42–44].

Evidence accumulated in the past decade has revealed that the chromosomal contacts, achieved by genome folding and looping, define separate compartments in the nucleus[45]. Hi-C data have shown that transcribed genes make preferential contacts with other transcribed genes, forming a spatial cluster known as the A compartment[46,47]. Reciprocally, silent genes and intergenic regions form a spatial cluster known as the B compartment. The loci of the B compartment are usually in contact with the nuclear lamina[48], i.e., at the periphery of the nucleus, where low levels of gene expression and heterochromatin histone signatures are found. In fact, these regions are almost completely avoided by HIV-1[18,25,26], whereas HIV-1 targets regions of open chromatin, which in some studies map in proximity to the NPC[18,19,49].

This suggests that a complex and dynamic interplay between the incoming virus, the host cell chromatin, and the dynamic nuclear organization contribute to the selection of genomic sequences into which HIV-1 integrates.

Here we find that HIV-1 integrates in proximity of SEs in patients and in T cell cultures in vitro. The observed phenomenon does not depend on the activity of SEs but on their position in spatial neighborhoods where HIV-1 insertion is facilitated. Consistently, HIV-1 integration hotspots cluster in the nuclear space and tend to contact SEs. Finally, we find that SE activity is critical to reorganize the genome of activated T cells, showing that they indirectly contribute to HIV-1 insertion biases.

## Results

**HIV-1 integrates in genes proximal to SEs**. We assembled a list of 4031 HIV-1 integration sites from activated primary CD4$^+$ T cells infected in vitro (ref. [50] and this study) and 9519 insertion sites from 6 studies from HIV-1 patients[38,39,51–54] (Supplementary Table 1). Ten thousand seven hundred and thirty-five integrations were in gene bodies (77% averaged over patient studies and 84% over in vitro infection studies), targeting a total of 5601 different genes (Supplementary Fig. 1a). This insertion dataset is not saturating (Supplementary Fig. 1b), yet we found that a subset of genes are recurrent HIV-1 targets, consistent with our previous findings[18]. We thus defined recurrent integration genes (RIGs) as genes with ≥1 HIV-1 integrations in at least 2 out of 8 datasets (see "Methods"), yielding a total of 1648 RIGs (Supplementary Fig. 1c).

To characterize RIGs, we extracted protein-coding genes without HIV-1 insertions in any dataset (called non-RIGs in the analysis, consisting of 13,140 genes) and compared their chromatin immunoprecipitation sequencing (ChIP-Seq) features in primary CD4$^+$ T cells. We first analyzed the levels of epigenomic features on protein-coding genes (Fig. 1a). As previously reported[18,50], we observed higher levels of H3K27ac, H3K4me1, and H3K4me3, as well as BRD4 and mediator of RNA polymerase II transcription subunit 1 (MED1) at transcription start sites of RIGs vs non-RIGs. Histone profiles of H3K36me3 and H4K20me1 were higher throughout RIG gene bodies, while the repressive transcription mark H3K27me3 was lower on RIGs vs non-RIGs. Of note, the mark of facultative heterochromatin H3K9me2 was depleted at transcription start sites of RIGs but remained unchanged throughout the gene body of RIGs vs non-RIGs.

In order to test the specificity of chromatin signatures of HIV-1 integration sites, we adapted the receiver operating characteristic (ROC) analysis[55,56]. We used control sites matched according to the distance to the nearest gene (see "Methods") and confirmed significant enrichment of the following genomic features: H3K27ac, H3K4me1, BRD4, MED1, H3K36me3, and H4K20me1 (Fig. 1b). The marks H3K27ac, H3K4me1, and H3K36me3, characteristic of active enhancers[57], cell type-specific enhancers[58], and bodies of transcribed genes[59], respectively, were the most enriched in the proximity of insertion sites. Consistent with the presence of H3K27ac and H3K4me1, we also found significant enrichment of BRD4, a constituent of SE genomic elements[30,32] (Fig. 1b). On average, 60% of insertion sites were significantly enriched in these chromatin marks (not shown) while we observed depletion of H3K27me3 and H3K9me2 in the proximity of insertion sites. Interestingly, we did not observe a statistically significant enrichment of H3K4me3 in the proximity of insertion sites.

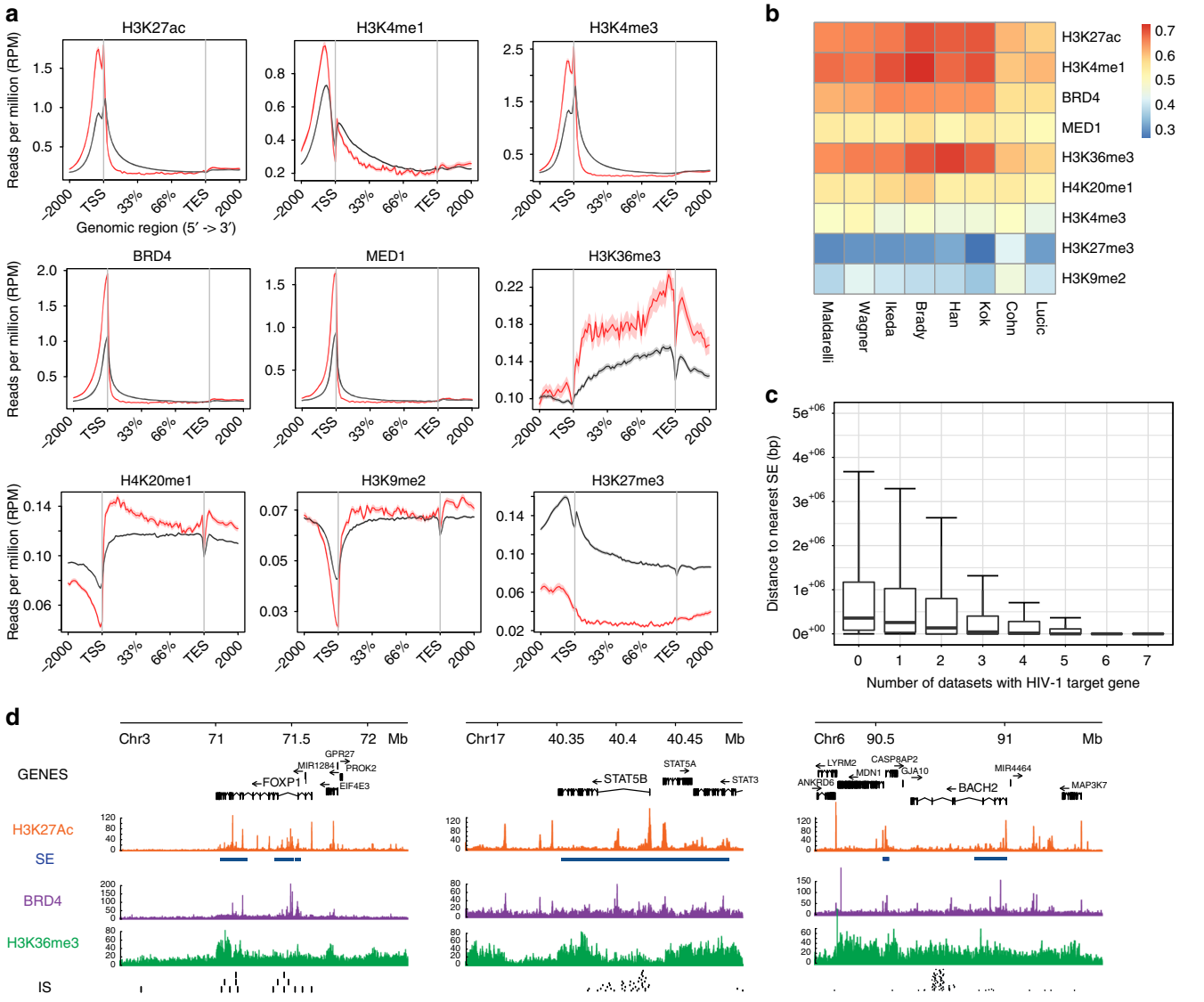



**Fig. 1** HIV-1 integration hotspots are within genes proximal to super-enhancers (SEs). **a** Metagene plots of H3K27ac, H3K4me1, H3K4me3, BRD4, MED1, H3K36me3, H4K20me1, H3K9me2, and H3K27me3 ChIP-Seq signals in recurrent integration genes (RIGs), which are protein coding in red and the rest of the protein-coding genes that are not targeted by HIV-1 (no RIGs) in black. **b** ROC analysis represented in heatmap summarizing the co-occurrence of integration sites and epigenetic modification obtained by ChIP-Seq for H3K27ac, H3K4me1, BRD4, MED1, H3K36me3, H4K20me1, H3K4me3, H3K27me3, and H3K9me2. HIV-1 integration datasets are shown in the columns, and epigenetic modifications are shown in rows. Associations are quantified using the ROC area method; values of ROC areas are shown in the color key at the right. **c** Distance to the nearest SE in activated CD4$^+$ T cells. Box plots represent distances from the gene to the nearest SE grouped by number of times the gene is found in different datasets. **d** *FOXP1*, *STAT5B*, and *BACH2* IS (black) superimposition on H3K27ac (orange), SE (blue), H3K36me3 (green), and BRD4 (violet) ChIP-Seq tracks

To confirm these trends, we identified SEs in activated CD4$^+$ T cells using H3K27ac ChIP-Seq and merged them with the SEs in activated CD4$^+$ T cells from dbSuper[60,61]. We obtained 2584 SEs, intersecting 564 RIGs (34.22%, Supplementary Fig. 1d). In addition, the more a RIG is targeted by HIV-1 (i.e., the higher the number of datasets where HIV-1 insertions are found in the gene), the closer it lies to SEs on average (Fig. 1c). In contrast, the insertion sites of the retrovirus HTLV-1[62] (human T lymphotropic virus type 1) were not enriched in SE marks (Supplementary Fig. 1e), while murine leukemia virus (MLV) showed a strong enrichment in all SE marks as expected[63]. Figure 1d shows the integration biases at gene scale on *FOXP1*, *STAT5B*, and *BACH2*, three highly targeted RIGs involved in T cell differentiation and activity. The ChIP-Seq profiles of H3K27ac, H3K36me3, and BRD4 indicate prominent clustering of HIV-1 insertion sites near the SEs defined by those marks. Thus HIV-1 displays specific preference to integrate into genes proximal to SEs, herein defined as genomic elements of retroviral integrations.

**RIGs are proximal to SEs regardless of their expression**. HIV-1 is known to integrate into highly expressed genes[20,29]. It is thus possible that genes with an SE are targeted more often because they are expressed at a higher level. To test whether this is the case, we measured the transcript abundance of protein-coding genes in CD3/CD28-activated CD4$^+$ T cells by RNA sequencing (RNA-Seq). The mean expression of genes with HIV-1 insertions is higher than those not targeted by HIV-1 (Fig. 2a). More specifically, 21.4% of protein-coding genes targeted by HIV-1 are in the top 10% most expressed genes, compared to 6.07% of non-targeted genes. Moreover, the genes more often targeted by HIV-1 (RIGs) are expressed at higher levels (Fig. 2b), thus confirming that HIV-1 is biased toward highly expressed genes.

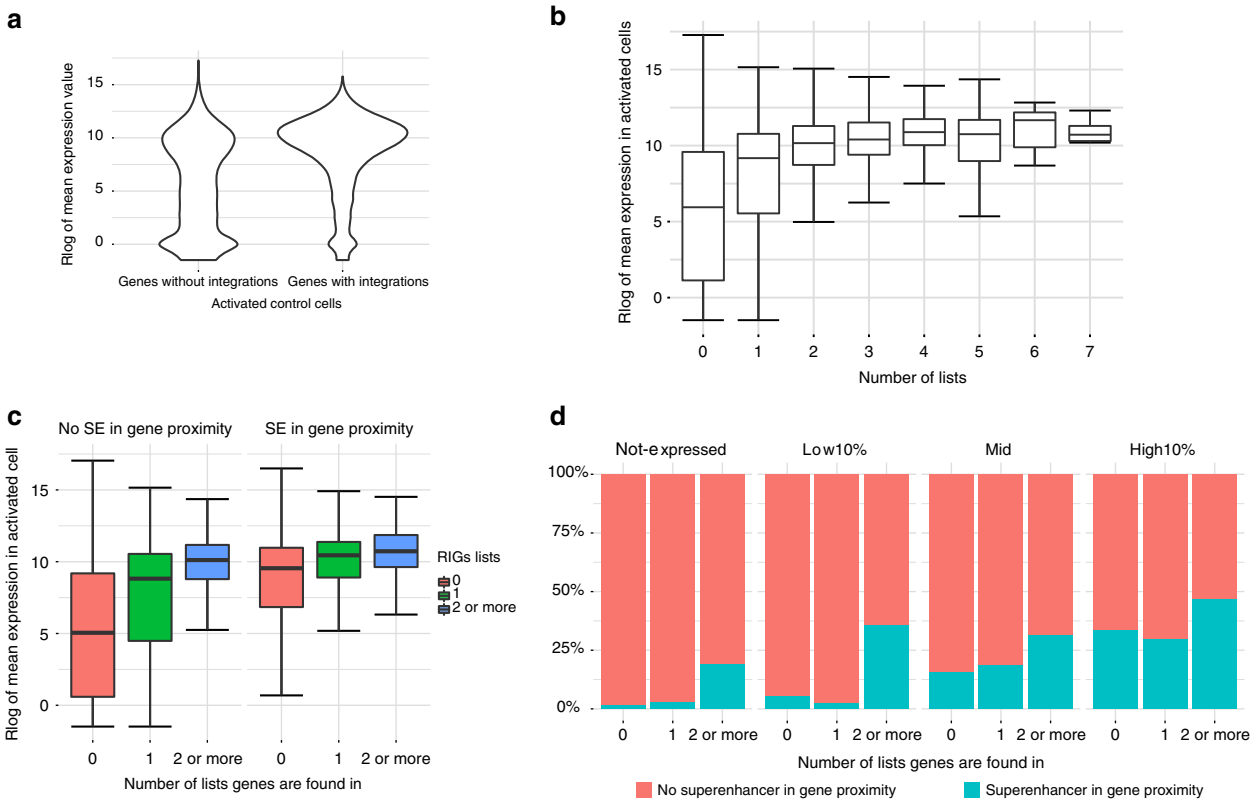



**Fig. 2** RIGs are proximal to super-enhancers regardless of their expression. **a** Regularized log-transformed read counts on protein-coding genes averaged over three replicates in activated CD4$^+$ T cells shown as violin plot for genes without HIV-1 integrations and genes with HIV-1 integrations. **b** Box plot for protein-coding genes grouped by number of HIV-1 lists they appear in. **c** Box plot for protein-coding genes grouped by the number of HIV-1 lists they appear in, with RIGs grouped together in ≥2 lists' group. Box plots are shown separately for genes that have super-enhancer 5 kb upstream of TSS or super-enhancer overlaps them (SE in proximity) and genes that do not have super-enhancer in proximity. Differences in median abundances of mRNA are statistically significant for all groups ($p$ value <2.2 x $10^{-16}$ for genes without HIV integrations and genes found on only one list and $p$ value 3.7 × $10^{-12}$ for RIGs, calculated by Wilcoxon rank-sum test). **d** Bar plots show the percentage of protein-coding genes that have super-enhancer in proximity, arranged by number of lists the gene is found in and by expression group

On average, genes with a SE are expressed at higher levels than those without (Fig. 2c). This trend is more subtle for RIGs, as they are expressed at a high level, with or without SEs (Fig. 2c, compare the blue boxes). However, RIGs are more often in the proximity of SEs than non-RIGs, irrespective of their expression (Fig. 2d). In particular, 19.05% of RIGs that are silent also have a proximal SE, while this is true for only 1.5% of the silent genes that were never found to be HIV-1 targets (Fig. 2d, leftmost panel). The trend remains the same for expressed genes (Fig. 2d) after dividing them into "low," "medium," and "high" expression groups (see "Methods"). In summary, our gene expression analysis suggests that genes recurrently targeted by HIV-1 have adjacent SE elements, irrespective of their transcriptional levels.

We next assessed the relationship between HIV-1 integration and transcription of genes controlled by SEs by using JQ1, a bromodomain and extraterminal domain protein inhibitor that prevents BRD4 binding to acetylated chromatin[64] and causes a subsequent dysregulation of RNA Pol II binding[31].

*MYC* is known to be regulated by SEs[31], so we used the *MYC* RNA and protein levels as a control for the JQ1 treatment in CD4$^+$ T cells (Supplementary Fig. 2a). We compared the HIV-1 insertion profiles with or without JQ1 by inverse PCR (see "Methods"). We mapped a total of 38,964 HIV-1 insertion sites and did not observe, at the chromosome scale, that JQ1 affects the insertion biases (Supplementary Fig. 2b, left panel). Similarly, spatial localization of the provirus and two representative RIGs remained unchanged upon treatment (Supplementary Fig. 2c, d).

Transcriptional profiling of activated CD4$^+$ T cells confirmed that protein-coding genes proximal to SEs are significantly more upregulated or downregulated upon JQ1 treatment than coding genes without SEs (Supplementary Fig. 2e, f). This effect is more pronounced among RIGs than among non-targeted genes (Supplementary Fig. 2f). Of note, HIV-1 maintains its preferences for highly transcribed genes in both control and JQ1-treated cells (compare Fig. 2a and Supplementary Fig. 2g).

In summary, our gene expression analysis suggests that genes recurrently targeted by HIV-1 are adjacent to SE elements, irrespective of their transcriptional levels, but disruption of SEs does not impact HIV-1 integration patterns.

**HIV-1 insertion hotspots are clustered in the nuclear space.** Our previously published results showed that the majority of tested RIGs are distributed in the outer zones of the T cell nucleus[18], so we hypothesized that the enrichment of HIV-1 insertion sites near SEs may be due to their particular organization in the nuclear space. We thus performed Hi-C to get some insight into the conformation of the T cell genome.

In order to minimize issues caused by the heterogeneity of the biological material, we used the widely available Jurkat lymphoid T cellular model. To ensure that the behavior of HIV-1 is similar in both models, we compared a published collection of 58,240 insertion sites in Jurkat cells[28] to the 28,419 insertion sites in primary CD4$^+$ T cells from the current study (obtained by linear amplification-mediated and inverse PCR) and previous

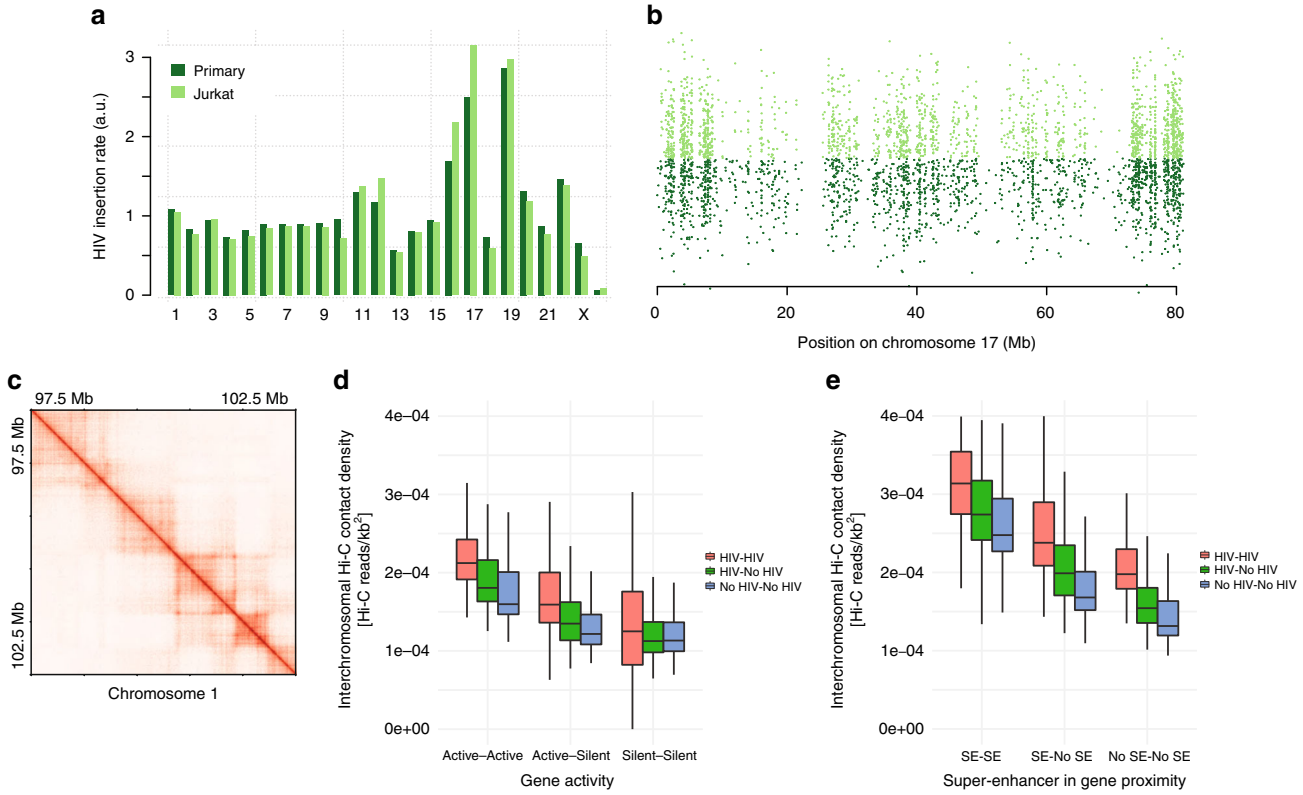



**Fig. 3** HIV-1 integration hotspots are clustered in the nuclear space. **a** Bar plot of HIV-1 insertion rate per chromosome (the genome-wide average is set to 1) in primary T and in Jurkat cells. **b** HIV-1 insertion cloud on chromosome 17 in primary T and Jurkat cells. Each dot represents an HIV-1 insertion site. The *x*-coordinate indicates the location of the insertion site on chromosome 17; the *y*-coordinate is random so that insertion hotspots appear as vertical lines. **c** Detail of the unnormalized Hi-C contact map in Jurkat in 5 kb bins. TADs and loop domains are clearly visible. **d** Box plot of inter-chromosomal Hi-C contact density (see "Methods"). Contact densities were computed between chromosomal aggregates of all gene fragments (5 kb) corresponding to Active and Silent genes, with (HIV) or without HIV insertions (No HIV). The distribution of densities are composed of the scores for all inter-chromosomal combinations. **e** Same as in **d**, but genes are classified between genes in proximity of super-enhancers (SE), i.e., within gene body or 5 kb upstream of TSS, or far from super-enhancers (No SE)

studies[38,39,51–54]. The insertion rates per chromosome are similar between cells (Fig. 3a); both show the characteristic approximately threefold increase on chromosomes 17 and 19. The apparent difference on chromosome 17 is possibly due to the use of different mapping technologies. For comparison, our previous measure of the insertion rates on chromosome 17 of Jurkat cells[29] (using the same inverse PCR technology) is very close to the current measure in primary CD4$^+$ T cells. The insertion cloud representation shows that the profiles are similar on chromosome 17, with the exception of a hotspot visible only in primary CD4$^+$ T cells at position ~57 Mb (Fig. 3b). We also found that the HIV-1 target genes are similar in Jurkat cells and in other CD4$^+$ datasets (Supplementary Fig. 3). In summary, apart from minor differences, HIV-1 insertion biases are comparable in primary CD4$^+$ T and in Jurkat cells.

Hi-C on uninfected Jurkat cells yielded ~1.5 billion informative contacts. Topologically associating domains (TADs) and loop domains are clearly visible on the raw Hi-C map in 5 kb bins (Fig. 3c), showing that the experiment captures the basic structural features of the Jurkat genome. We also verified that the A and B compartments are well defined and that they correspond to the regions of high and low gene expression, respectively (data not shown). To our knowledge, this dataset constitutes the highest-resolution Hi-C experiment presently available in Jurkat cells.

If the insertion pattern of HIV-1 reflects a particular organization of the genome, one predicts that the insertion hotspots occupy the same nuclear space and thus cluster together in three dimension (3D). We tested this hypothesis by measuring the amount of inter-chromosomal Hi-C contact densities among different classes of HIV-1 insertion sites (Fig. 3d). The loci most targeted by HIV-1 engage in stronger contact with each other than non-targeted loci. Also, the differences in contact strength are more pronounced when loci correspond to active genes. In addition, SEs tend to cluster together and with HIV-1 insertion hotspots in 3D (Fig. 3e, Supplementary Fig. 4), indicating that SEs locate in the physical proximity of HIV-1 insertion sites. Thus HIV-1 insertion sites form spatial clusters interacting with SEs in the nucleus, consistently with the view that the insertion process depends on the underlying 3D organization of the T cell genome.

**SEs and HIV-1 occupy the same 3D sub-compartment**. To better define the properties of HIV-1 insertion sites, we segmented the Jurkat genome into spatial clusters. For each chromosome, we generated 15 clusters of loci enriched in self interactions, which we coalesced down to 5 genome-wide clusters based on their inter-chromosomal contacts (Fig. 4a and see "Methods"). This approach yielded two A-type sub-compartments called A1 and A2, two B-type sub-compartments called B1 and B2, and one intermediate/mixed compartment called AB (Fig. 4b, c).

The AB- and B-type sub-compartments correspond to known types of silent chromatin: AB is richest in the Polycomb mark H3K27me3, B1 is richest in H3K9me3, and B2 is richest in lamin (Fig. 4d and Supplementary Fig. 5a). The two A-type sub-compartments are enriched in euchromatin marks, with higher coverage in A1 than in A2 (Fig. 4d and Supplementary Fig. 5a).

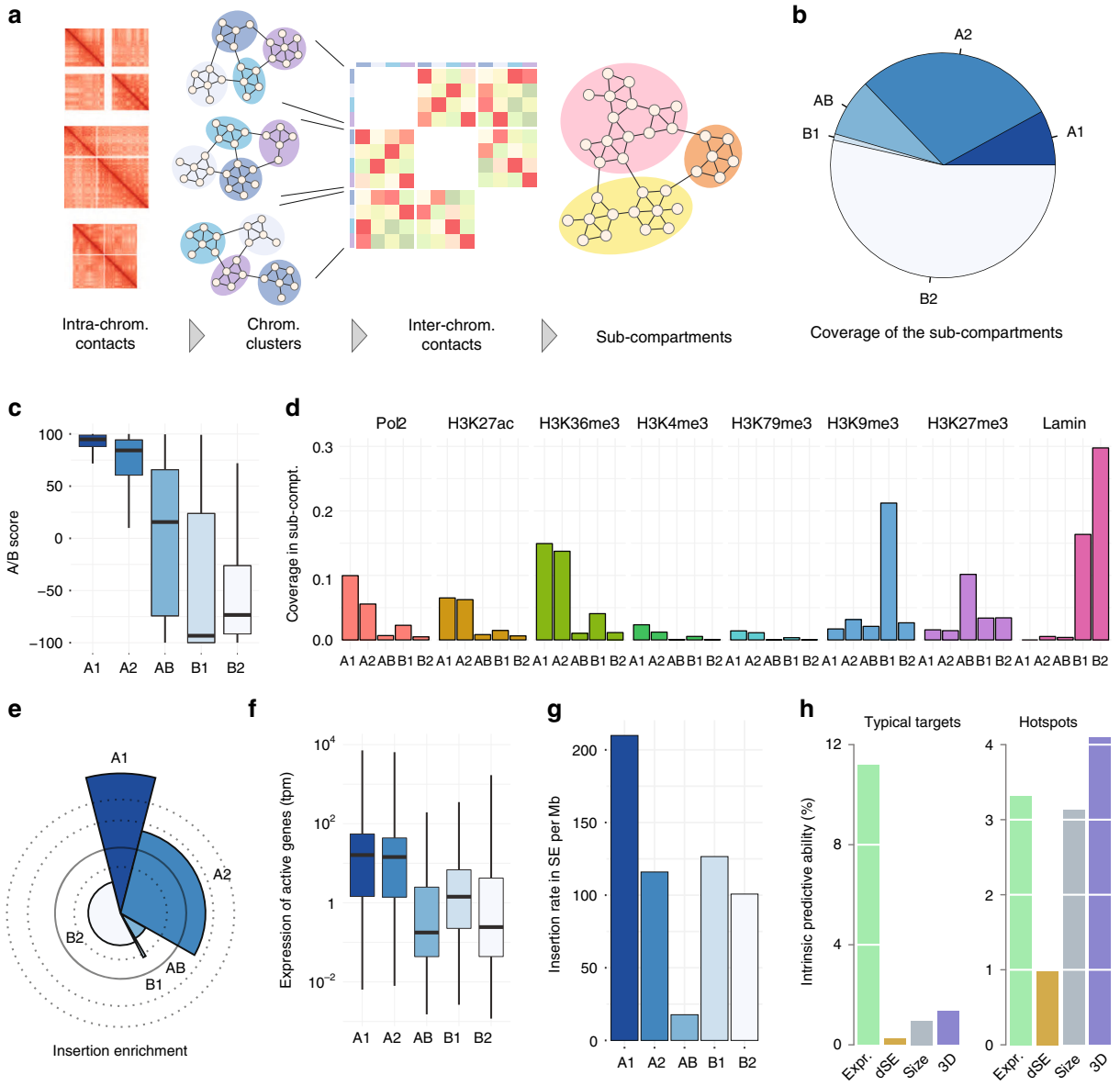



**Fig. 4** Super-enhancers and HIV-1 occupy the same 3D sub-compartment. **a** Definition and identification of 3D compartments. For each chromosome, 15 spatial communities were identified by clustering. The inter-chromosomal contacts between the communities were used as a basis for another round of clustering in five genome-wide spatial communities. **b** Pie chart showing the coverage of the sub-compartments in the Jurkat genome. **c** Distribution of AB scores in the 3D sub-compartments. The AB score measures the likelihood that a locus belongs to the A or B compartment. Extreme values +100 and −100 stands for "fully in A" or "fully in B", respectively. A score of 0 means "both or neither." **d** Proportion of 3D sub-compartments covered by major chromatin features. Coverage was computed as the span of enriched ChIP-Seq signal divided by the sub-compartment size. **e** Spie chart showing the observed vs expected HIV-1 insertions in the sub-compartments. The expected amount of insertions is the area of the wedge delimited by the circle in bold line, and the observed amount is the area of the colored wedge. Dotted lines represent the limit of the wedge for 2× and 3× enrichment outside the circle, and 0.5× depletion inside the circle. The observed/expected ratio is approximately 2.5 times higher in A1 than in A2. **f** Box plot showing the expression of protein-coding genes in the sub-compartments. The plot was rendered using defaults from the ggplot2 library. The *y*-axis has a logarithmic scale. **g** Bar plot showing the integration density of HIV in super-enhancers located in different 3D sub-compartments. **h** Bar plot showing the contribution of different predictors to the HIV-1 insertion sites in typical genes (left) or in hotspots (right). The *y*-axis represents the loss of accuracy when the corresponding variable is removed from the model. Expr. gene expression, dSE distance to nearest super-enhancer, Size gene size, 3D sub-compartment. See "Methods" for detail

Strikingly, the rate of HIV-1 insertion is 2.7 times higher in A1 than in A2 (Fig. 4e). In contrast, the coverage of euchromatin marks and the transcriptional activity are only slightly higher in A1 than in A2 (Fig. 4d, f), e.g., 1.04 times higher in H3K27ac coverage, 1.09 times in H3K36me3 coverage, and 1.12 times in median gene expression. More importantly, the ~2.5-fold enrichment of HIV-1 insertion is still present when controlling for gene expression (Supplementary Fig. 5b), indicating an intrinsic preference for the A1 sub-compartment. Of note, we obtained similar results when defining 10 sub-compartments instead of 5, where HIV-1 insertion rates are enriched in one sub-compartment covering ~10% of the genome (data not shown). Hence, our observation is robust with respect to the definition of sub-compartments. The 3D organization of the Jurkat T cell genome thus explains large differences of HIV-1 insertion rates between genes expressed at similar levels.

If HIV-1 targets SEs because of their location in the nuclear space, one predicts that the insertion rate of HIV-1 in the SEs of

A1 should be higher than in the SEs of A2. Figure 4g shows that, indeed, HIV-1 is ~1.5 times more likely to integrate in the SEs of A1 than in those of A2. Since the insertion rate in SEs depends primarily on their location, we conclude that the enrichment in SEs at genome-wide scale is due to their position in the 3D space of the nucleus, rather than to their activity or their chromatin features.

To quantify this statement and to clarify how different determinants contribute to HIV-1 insertion, we used a modeling approach based on logistic regression. We predicted either typical HIV-1 target genes (top 33% gene-wide insertion rate) or HIV-1 hotspots (top 2.5% bin-wise insertion rate, see "Methods"). Typical HIV-1 targets are almost entirely determined by gene expression (Fig. 4h), consistently with previous reports that HIV-1 integrates primarily in active genes[2,3,20]. On the other hand, HIV-1 hotspots are multifactorial and sub-compartments appear as the major determinants (Fig. 4h). These results establish that typical HIV-1 targets and hotspots, such as RIGs, are driven by different classes of mechanisms. Finally, they show that the 3D organization is a major contributor of HIV-1 hotspots.

**Genes proximal to SEs reposition upon T cell activation**. Our results so far suggest that HIV-1 insertion hotspots cluster near SEs because of their location in the structured genome of T cells, but they do not address the contribution of SEs to this structure.

We thus investigated the role of SEs in the spatial distribution of genes in T cells. RIGs belong to a subset of T cell genes that show the strongest response to T cell activation (Supplementary Fig. 6a), so we reasoned that their spatial positioning might change with the activation status of the cell. We therefore employed 3D immuno-DNA fluorescence in situ hybridization (FISH) to visualize gene positioning in resting and activated CD4$^+$ T cells. The cumulative frequency plots revealed that nine RIGs, seven of which have SEs *FOXP1*, *STAT5B*, *NFATC3* (Fig. 5a), *KDM2A*, *PACS1* (Fig. 5b), and *GRB2*, *RNF157* (Fig. 5c), change spatial positioning and relocalize further toward the outer shells of the T cell nucleus upon activation. Three RIGs, *NPLOC4*, *RPTOR*, and *BACH2*, were already peripheral before activation and remained so afterwards (Supplementary Fig. 6c). We recapitulated the overall distribution of 9 RIGs that displayed repositioning in activated ($n = 1690$ alleles) vs resting CD4$^+$ T cells ($n = 1700$ alleles, Fig. 5d). As expected, the frequency distribution of alleles in three zones of equal surface areas[18] showed a prominent shift toward the outer shells of the nucleus in activated T cells, corresponding to the area located <1 micron under the nuclear envelope. Of note, a pan nuclear distribution of *KDM2A* and *PACS1* in activated CD4$^+$ T cells was also observed[26].

We then asked whether the observed gene redistribution is an exclusive feature of genes proximal to SEs or a general feature of all expressed genes, independent of HIV-1 targeting. We therefore evaluated the spatial distribution of two groups of control genes: expressed genes with SEs and expressed genes without SEs. The *MYC* gene, a gene harboring five well-described SEs, changed its radial position toward the outer shells of the nucleus upon T cell activation (Supplementary Fig. 6d). The same trend was observed for two other regions proximal to SEs that are not targeted by the virus: one on chromosome 1 covering the gene *LMNA* and the other on chromosome 11 encompassing *SLC43A1*, *UBE2L6*, and *TIMM10*. Both regions showed statistically significant repositioning toward the more exterior shells of the nucleus with T cell activation.

In contrast, when we assessed the spatial distribution of three highly expressed genes without SEs, *TAP1*, *CCNC*, and *MCM4*, we did not observe any statistically significant allele redistribution upon T cell activation (Supplementary Fig. 6e).

Next, we wanted to understand whether disruption of SEs impacts the nuclear position of genes proximal to these elements during T cell activation. To do so, we pretreated resting T cells with JQ1 before activating them with CD3/CD28 beads and observed that the two tested RIGs, *STAT5B* and *GRB2*, retained their position in the center of the nucleus (Supplementary Fig. 6g, h), supporting the notion that SEs contribute to the positioning of genes prior and during T cell activation.

As HIV-1 target genes group together on linear chromosomes[18] and integration hotspots cluster in the nuclear space (Fig. 3e and Supplementary Fig. 4), we assessed their spatial relationships during T cell activation. Two highly targeted regions (top 10% of RIGs density) on chromosomes 11 and 17 were visualized by dual-color FISH coupled to high-throughput imaging (HTI)[65,66]. We observed that *KDM2A* and *PACS1*, two genes proximal to SEs lying at 1.1 Mb from each other on chromosome 11 (Fig. 5e), clustered together in the nuclear space, with a minimized median distance of 0.42 µm in both resting and activated state (data summarized in Supplementary Fig. 6f). Similarly, we found that, in the hotspot region mapping to q25.1-3 on chromosome 17 containing 35 RIGs (Fig. 5f), three RIGs, *GRB2*, *TNRC6C*, and *RNF157*, cluster together (Fig. 5f, data summarized in Supplementary Fig. 6f). This clustering is not a mere consequence of the linear distances between these genes, as two other genes from the same locus, *NPLOC4* and *RPTOR*, despite being at a similar linear distance, are not spatially associated (measured distances given in Supplementary Fig. 6f).

In summary, our results show that seven out of nine RIGs proximal to SEs change their radial positioning upon T cell activation, moving to the outer shell of the nucleus. This is a feature pertinent also to genes proximal to SEs that are not HIV-1 targets, suggesting that SEs contribute to the spatial organization of the genome and that in dependence of the activation state could be more exposed to HIV-1 insertions.

## Discussion

The integration of the viral DNA into the host cell genome is responsible for the long-term persistence of HIV-1 in cellular reservoirs[67]. The persistence of HIV-1 is influenced by the chromosomal context at the sites of integration, with a strong impact on the outcome of viral infection[29]. Here we characterized the genomic features of integration sites identified from patients[38,39,51–54] and from in vitro infections of activated CD4$^+$ T cells (ref. [50] and this study). By analyzing these large datasets, we confirmed that HIV-1 recurrently integrates into a subset of transcriptionally active cellular genes. We show that HIV-1 recurrently integrates into a group of genes proximal to SE genomic elements in activated CD4$^+$ T cells and in patients (Fig. 1c). Yet, neither the activity of SEs nor their effect on gene expression alone explain the integration biases (Supplementary Fig. 2b). Instead, we found that the correlation can be attributed to the enrichment of SEs in the A2 and especially in the A1 sub-compartments, where HIV-1 integrates at higher frequency than in the rest of the genome (Fig. 4e).

The contribution of gene expression levels to the insertion rate of HIV-1 is intricate. On one hand, HIV-1 shows a clear bias toward expressed genes, even upon JQ1 treatment where it still integrates preferentially into genes that are most active after JQ1 treatment. However, HIV-1 recurrently integrates into genes proximal to SEs (Fig. 2d), among which there are both upregulated and downregulated genes (Supplementary Fig. 2f). One potential explanation is that the HIV-1 IN has a strong affinity for some protein present in transcribed regions (e.g., LEDGF). The

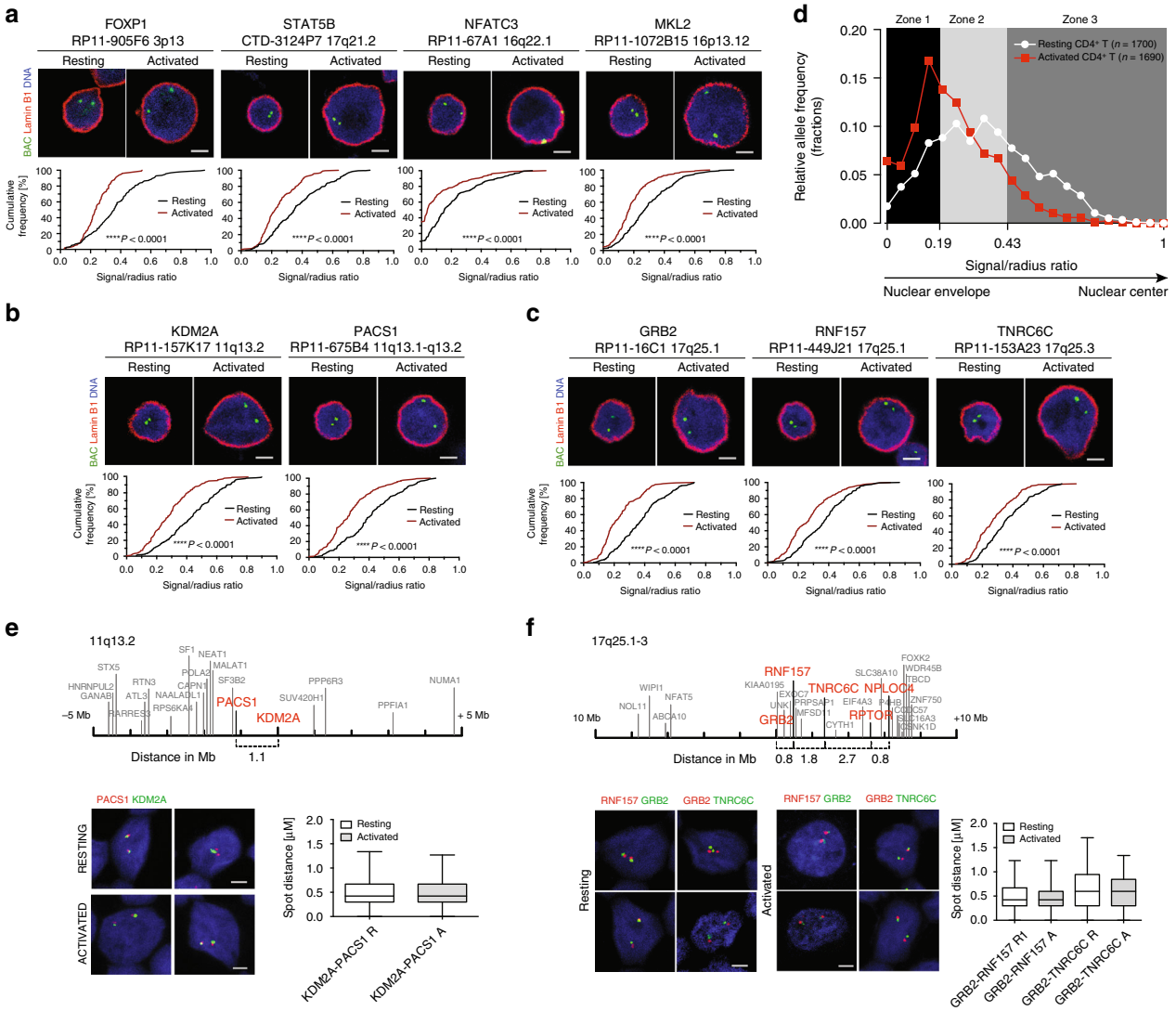



**Fig. 5** Genes proximal to super-enhancers change their nuclear positioning upon T cell activation. Three-dimensional immuno-DNA FISH of nine RIGs in resting and activated (anti-CD3/anti-CD28 beads, IL-2 for 48 h) CD4$^+$ T cells (green: BAC/gene probe, red: lamin B1, blue: DNA staining with Hoechst 33342, scale bar represents 2 µm). Cumulative frequency plots show combined data from both experiments ($n = 100$, black: resting cells, red: activated cells). The *p* values of the Kolmogorov–Smirnov tests are indicated. Box plots represent minimized distances (5th−95th percentile) for the analyzed gene combinations in resting (white) and activated (gray) CD4$^+$ T cells, obtained by high-throughput imaging and subsequent computational measurements. In the box plots, the center line represents the median, the bounds of the box span from 25% to 75% percentile, and the whiskers visualize 5% and 95% of the data points. Representative images for **a** *FOXP1*, *STAT5B*, *NFATC3*, and *MKL2*; **b** *KDM2A* and *PACS1*; and **c** *GRB2*, *RNF157*, and *TNRC6C*. **d** Allele fraction density plot for all resting and activated alleles that displayed peripheral repositioning. The *y*-axis shows the allele fraction density for genes *FOXP1*, *STAT5B*, *NFATC3*, *MKL2*, *KDM2A*, *PACS*, *GRB2*, *RNF157*, and *TNRC6C*. The *x*-axis represents ratios of distance from nuclear envelope (lamin B1 staining) and radius (signal to radius ratio) for alleles in resting cells ($n = 1700$) and activated cells ($n = 1690$ alleles). Binning into three equal concentric zones of the nucleus is performed as in ref. [18]. **e** Schematic representation of chromosomal region 11q13.2 within 10 Mb: RIGs (bold red) and single HIV-1 integration sites (plain gray) and HTI of *KDM2A* and *PACS1*. **f** Schematic representation of chromosomal region 17q25.1-3 within 20 Mb: RIGs (bold red) and single HIV-1 integration sites (plain gray) and HTI of *GRB2*, *RNF157*, and *TNRC6C*

complete absence of such proteins in non-transcribed regions would have more influence on the signal than its quantitative variations in transcribed regions. In any event, gene expression and chromatin are not the sole contributors to HIV-1 insertion patterns. The A1 sub-compartment is targeted more frequently than the rest of the genome (Fig. 4e), even when controlling for chromatin and gene expression (Supplementary Fig. 5b). This indicates that the 3D genome organization of activated T lymphocytes is an important determinant of the HIV-1 insertion process. SEs most likely contribute to this organization[42] because their dismantling prior to T cell activation prevents repositioning of genes with SEs toward the outer shells of the nucleus (Supplementary Fig. 6g, h).

Predictive modeling helps clarify this conclusion. An important insight is that HIV-1 insertion hotspots do not obey the same rules as typical target genes (Fig. 4h). There are thus two processes at work: one that attracts viruses to active genes, and another one, more complex, that provokes recurrent integrations within the same genes (i.e., the RIGs). 3D compartmentalization plays a role only in the second process, explaining why studies with different definitions of HIV-1 targets may come to different conclusions.

Although we show that SEs do not affect HIV-1 integration patterns in activated T cells, we find that, during T cell activation, genes with SEs move toward the outer shells of the nucleus (Fig. 5a–c). In line with the previously shown association of nuclear pore proteins with HIV-1[18,19], and their proximity to SEs and enrichment in the A1 sub-compartment defined here (ref. [41] and data not shown), it is tempting to speculate that the A1 sub-compartment corresponds to genomic loci associated with the nuclear pore. None of the chromatin features mapped in Jurkat cells is known to discriminate active genes at the nuclear pore from other active genes, and the chromatin of A1 is otherwise similar to that of A2 (Supplementary Fig. 5a). Interestingly, the density of SEs is similar between A1 and A2 (Supplementary Fig. 5a), so it is unlikely that the A1 sub-compartment simply emerges from the clustering of SEs. More plausibly, SEs are one of many contributors to the segregation of the genome in spatial clusters. More generally, the existence of two separate clusters of active genes in the 3D space of the nucleus is itself an intriguing observation that will require more work to be fully understood.

While the spatial positioning of the A1 and A2 sub-compartments in T cells still needs to be mapped, a recent study proposed an alternative concept to the one where nuclear periphery represents solely transcriptionally repressive environment[68,69]. Instead, and consistently with our findings, they predict distinctive localization of active A1 and A2 Hi-C sub-compartments. Transcriptionally active regions are divided into two groups: a transcriptional "hot zone" close to nuclear speckles corresponds to the A1 sub-compartment and another one far from speckles corresponds to A2[68]. Interestingly, transcriptional hot zones confined within the A1 sub-compartment are enriched in SEs and highly expressed genes, traits we observed to be strongly associated with HIV-1 insertional hotspots.

It is well established that the main components that mediate HIV-1 integration into actively transcribing units are the viral proteins IN and CA[3,17]. Their cellular partners LEDGF/p75 and CPSF6 could chaperone the virus into clusters of SE domains in the A1 compartment. LEDGF/p75 interacts with a large number of splicing factors and directs HIV-1 integration to highly spliced transcription units[22], making this a plausible link to the A1 compartment.

Likewise, CPSF6 as part of the mRNA polyadenylation machinery, could guide HIV-1 integrations toward the nuclear compartment with high transcription and mRNA processing rates (such as A1[68]). Alternatively, the CA–CPSF6 axis could regulate HIV-1 targeting independently of the polyadenylation role of CPSF6[24,26,70].

Among the factors that are binding putative SEs and could play a role in integration site selection, p300 and BRD4 seem to be the most promising candidates. p300, a histone acetyltransferase used to identify typical[71,72] and SEs[32,73,74], is an interaction partner of the HIV-1 IN. p300 promotes the DNA-binding activity of IN[75] and could serve to direct viral integration toward genes with SEs in the A1 compartment, though a role for p300 in HIV-1 integration targeting has yet to be established.

BRD4, on the other hand, does not bind HIV-1 IN[76,77] but has a well-established role in HIV-1 latency[78,79]. The mechanism of action has recently been ascribed to the short isoform of BRD4, which recruits a repressive SWI/SNF complex to the viral long terminal repeat (LTR)[80]. Loss of the short isoform, occurring rapidly upon JQ1 treatment, leaves the long isoform engaged in the transcriptional activation of the viral genome[80]. The same mechanism could account for the activation of cellular genes upon JQ1 treatment[31]. In fact, our RNA-Seq data show that genes proximal to SEs are both upregulated and downregulated upon JQ1 (Supplementary Fig. 2e). Furthermore, genes targeted by HIV-1 are more responsive to JQ1 than non-HIV-1 targets (Supplementary Fig. 2g). This implies that HIV-1 preferentially targets genes that have a rapid and tightly regulated transcriptional response. Given the opposing role of BRD4 on viral LTR and cellular genes, insertion into genes proximal to SEs might represent a source of transcriptional fluctuations[81,82] and play an important role in either establishment or reversal of latency.

Based on our findings that the majority of tested RIGs and genes with SEs reposition from the nuclear interior to the periphery during T cell activation, it could be envisaged that RIGs differ between resting and activated $CD4^+$ T cells. Meta-analysis of the only available integration sites dataset[50] from these two cell activation states showed, however, no significant difference. Additional work will thus be required to assess comprehensively the RIGs that are used by HIV-1 in resting T cells.

Overall, we show that HIV-1 insertion sites form spatial clusters interacting with SEs of A1 compartment, highlighting the importance of the underlying 3D genome organization for HIV-1 integration. While additional studies will be needed to decipher the mechanism of such site selection, our results identify hotspots of integration that could improve characterization and enable targeting of latent HIV-1 reservoirs.

## Methods

**Primary cell isolation, culture, treatments, and infection**. For $CD4^+$ T cells isolation, whole blood was mixed with RosetteSep Human $CD4^+$ T cell enrichment cocktail beads according to the manufacturer's instructions and $CD4^+$ T cells were separated using Histopaque Ficoll gradient by centrifugation. Cells were cultured in complete T cell medium (RPMI-1640+10% fetal bovine serum (FBS) + primocin), left in resting state or activated with Dynabeads Human T-Activator CD3/CD28 and plated in complete medium supplemented with 5 ng/ml IL-2 for 20–72 h at 37 °C.

Cells were treated when indicated with 500 nM JQ1(+) or dimethyl sulfoxide (DMSO) for 6 h at 37 °C.

In all, $1 \times 10^6$ activated $CD4^+$ T cells were infected with 0.5–1 µg of p24 of virus by spinoculation for 90 min at 2300 rpm at room temperature (RT) in the presence of polybrene at 37 °C. Virus stocks were produced from the viral clone HIV-$1NL_{4\_3}$ and a mutant that harbors a frameshift (FS) mutation in the *env* gene ($pNL_{4\_3}$-envFS) and was pseudotyped with vesicular stomatitis virus glycoprotein, resulting in a FS virus that performs a single-round infection (HIV-$1NL_{4\_3}$ FS). Cells were then incubated for 72 h at 37 °C. When indicated, 14 h after infection with HIV-$1NL_{4\_3}$, cells were treated with the fusion inhibitor T20 to prevent multiple infection and integration. All viral stocks were generated by transfecting viral DNA in HEK 293T cells and collecting supernatants after 48–72 h following sucrose gradient purification of virus articles. Viral production was quantified in the supernatants for HIV-1 p24 antigen content using the Innotest HIV Antigen mAB Kit (INNOGENETICS N.V. Gent, Belgium). The human Jurkat T cell line (obtained from the cell collection of the Center for Genomic Regulation, Barcelona) was grown at 37 °C under a 95% air and 5% $CO_2$ atmosphere, in RPMI 1640 medium (Gibco) supplemented with 10% FBS (Gibco), 1% penicillin–streptomycin (Gibco), and 1% GlutaMAX (100×) (Gibco). Jurkat cells were passaged every 2 days with a 1:5 dilution. Cells were tested for mycoplasma regularly.

**Fluorescence in situ hybridization**. Approximately $3 \times 10^5$ $CD4^+$ T cells were plated on the PEI-coated coverslips placed into a 24-well plate for 1 h at 37 °C. Cells were treated with 0.3× phosphate-buffered saline (PBS) to induce a hypotonic shock and fixed in 4% paraformaldehyde (PFA)/PBS for 10 min Coverslips were extensively washed with PBS and cells were permeabilized in 0.5% triton X-100/PBS for 10 min. After three additional washings with PBS-T (0.1% tween-20), coverslips were blocked with 4% bovine serum albumin (BSA)/PBS for 45 min at RT and primary antibody anti-lamin B1 ab16048, from Abcam (1:500 in 1% BSA/PBS), was incubated overnight at 4 °C. Following three washings with PBS-T, fluorophore-coupled secondary antibody (anti-rabbit, coupled to Alexa 488 #11034, Alexa 568 #A11011, or Alexa 647 #A27040 from Invitrogen, diluted 1:1000 in 1% BSA/PBS) were incubated for 1 h at RT, extensively washed, and post fixed with ethylene glycol bis(succinimidyl succinate) (EGS) in PBS. Coverslips were washed three times with PBS-T and incubated in 0.5% triton X-100/0.5% saponin/PBS for 10 min. After three washings with PBS-T, coverslips were treated with 0.1 M HCl for 10 min, washed three times with PBS-T, and additionally permeabilized step in 0.5% triton X-100/0.5% saponin/PBS for 10 min. After extensive PBS-T washings, coverslips were equilibrated for 5 min in 2× saline sodium citrate (SSC) and then put in hybridization solution overnight at 4 °C. For the HIV-1 FISH, RNA digestion was additionally performed beforehand using RNAse A (100 µg/ml).

For FISH without immunofluorescence (IF) for HTI, $1–2 \times 10^6$ $CD4^+$ T cells in 500 µl of medium were adhered to coverslips by centrifugation at $350 \times g$ for 10

min at RT. The coverslips were washed in PBS and the cells were fixed in 4% PFA/PBS for 10 min followed by extensive PBS washing. Permeabilization was performed by incubation in 0.5% triton X-100/0.5% saponin/PBS for 20 min. After three washings with PBS, cells were treated with 0.1 M HCl for 15 min. Coverslips were washed twice for 10 min with 2× SSC and put in hybridization solution overnight at 4 °C.

For DNA probe labeling, bacterial artificial chromosome (BAC) or P1 artificial chromosome (PAC) DNA was extracted using a Nucleobond Xtra Maxiprep or amplified by the Illustra GenomiPhi V2 DNA Amplification Kit according to the manufacturer's instructions. HIV-1 plasmid HXB2 was purified using the Qiagen Plasmid Extraction Kit. FISH probes were generated in a Nick translation reaction using three different protocols. All BACs/PACs are listed in Supplementary Table 5.

BACs were labeled with digoxigenin (DIG)-coupled dUTPs. Three micrograms of BAC DNA were diluted in $H_2O$ in a final volume of 16 µl. Four microliters of DIG-Nick translation mix (Roche) were added and the labeling reaction was carried out at 15 °C for up to 15 h. The labeling reaction was performed by using a fluorophore-coupled dUTPs in the same concentration as biotin-16-dUTP in ref. [2].

For HIV-1 labeling, a biotin-dUTP nucleotide mix containing 0.25 mM dATP, 0.25 mM dCTP, 0.25 mM dGTP, 0.17 mM dTTP, and 0.08 mM biotin-16-dUTP in $H_2O$ was prepared. Three micrograms of pHXB2 were diluted with $H_2O$ in a final volume of 12 µl, and 4 µl of each nucleotide mix and Nick translation mix (Roche) were added. Labeling was performed at 15 °C for 3–6 h.

For dual-color FISH or improvement of signal-to-noise ratio in single-color FISH, probes were labeled using the fluorophore-coupled nucleotides SpectrumGreen dUTP (Abbott), SpectrumOrange dUTP (Abbott), and Red 650 dUTP (Enzo).

In all, 1–3 µg of BAC DNA were diluted in a final volume of 22.5 µl $H_2O$. Also, 2.5 µl of 0.2 mM fluorophore-coupled dUTP, 5 µl of 0.1 mM dTTP, 10 µl of dNTP mix containing 0.1 mM of each dATP, dCTP, and dGTP, and 5 µl of 5× Nick translation buffer (Abbott) were added and reagents were mixed well by vortexing. The reaction was started by addition of 5 µl Nick translation enzymes (Abbott) and incubated at 15 °C for 13–14 h. The probes were checked for their size on a 1% agarose gel, and 200–500 bp probes were purified using Illustra Microspin G-25 columns according to the manufacturer's instructions. Probes were precipitated in ethanol, dissolved in formamide and 4× SSC/20% dextran sulfate (1:1), and stored at −20 °C prior to use.

For probe hybridization, 1–6 µl of probe was loaded on glass coverslips and heat denatured in metal chamber at 80 °C for 8 min in a water bath. Hybridization was carried out for 48 h at 37 °C. Four washings in 2× SSC (10 min each) at 37 °C were followed with 2 washings in 0.5× SSC at 56 °C.

FISH development for DIG-labeled BACs was performed by using fluorescein isothiocyanate (FITC)-labeled anti-DIG antibody (Roche), whereas biotin-labeled HIV-1 probes were detected by TSA Plus system from Perkin Elmer, that allows significant amplification of the signal, by using an anti-biotin antibody (SA-HRP) and a secondary antibody with a fluorescent dye (usually FITC for HIV).

For the directly labeled probes after initial washings, nuclei were stained with Hoechst 33342 (1:5000 in PBS), washed in PBS, and then mounted using mowiol.

**Microscopy and image analysis**. For the classical confocal microscopy and manual image analysis, 3D stacks were acquired with a Leica TCS SP8 confocal microscope using a ×63 oil immersion objective. Distance measurements were performed using Volocity (Perkin Elmer). The smallest distance between the FISH signal and the nuclear lamina, stained by IF for lamin B1, was determined, and measurements were normalized to the nuclear radius (defined as half of the maximum diameter of the lamin B1 ring). Signal-to-radius ratios were either binned into three classes of equal surface (zones 1–3)[18] or plotted on a cumulative frequency plot. Kolmogorov–Smirnov (KS) tests were performed to compare the distributions of positioning of a gene between two conditions (resting vs activated or DMSO vs JQ1).

For HTI and image analysis of dual-color FISH, images were acquired with a spinning disk Opera Phenix High Content Screening System (PerkinElmer), equipped with four laser lines (405 nm, 488 nm, 568 nm, 640 nm). Images of FISH experiments to calculate 3D distances were acquired in confocal mode using a ×40 water objective lens (NA 1.1) and two 16 bit CMOS cameras (2160 by 2160 pixels), with camera pixel binning of 2 (corresponding to 299 nm pixel size). For each sample, 11 *z*-planes separated by 0.5 µm were obtained for a total number of at least 36 randomly sampled fields, which acquired per condition a minimum of $16 \times 10^3$ cells. Image analysis was performed using the Harmony high-content imaging and analysis software (version 4.4, Perkin Elmer), using custom-made image analysis building blocks. Nuclei were segmented based on the Hoechst nuclei staining signal of maximum projected images using the algorithm B and cells in the periphery of the image were excluded from further analysis. FISH probe detection was performed by using the spot detection algorithm C and custom-made scripts were used to calculate the Euclidean distances between all the different colored probes per cell. Single cell-level data were then exported and custom-made R scripts were used to select the minimum distance between the different FISH probes per allele basis. To exclude spurious spot detection events from the analysis, only the distances of cells with two FISH probes detected per channel were calculated and plotted (Graph Pad, Prism).

**Quantitative real-time PCR (qPCR)**. Up to $5 \times 10^6$ $CD4^+$ T cells were used for RNA extraction with the InviTrap Spin Kit (Stratec Biomedical) according to the manufacturer's instructions and up to 500 ng of RNA was retro-transcribed using Moloney MLV reverse transcriptase from Invitrogen according to the manufacturer's instructions. Gene expression analysis were performed in duplicates using IQ supermix from Biorad in CFX96/C1000 Touch Real-Time PCR system, as described in Lusic et al., 2013. Statistical analysis of qPCR data was performed using Graphpad. Taqman assays used were: for MYC Taqman Hs00153408_m1 FAM/MGB and for GAPDH 4310884E VIC/TAMRA.

**Western blotting**. In all, $5 \times 10^6$ cells were harvested and homogenized in lysis buffer (20 mM Tris-HCl, pH 7.4, 1 mM EDTA, 150 mM NaCl, 0.5% Nonidet P-40, 0.1% sodium dodecyl sulfate (SDS), 0.5% sodium deoxycholate) supplemented with protease inhibitors (Roche) for 10 min at 4 °C and sonicated (Bioruptor) for 5 min. Equal amounts of total cellular proteins (20 µg), as measured with Bradford reagent (Biorad), were resolved by 10% SDS-polyacrylamide gel electrophoresis, transferred onto nitrocellulose membrane (GE Healthcare), and then probed with primary antibody, followed by secondary antibody conjugated with horseradish peroxidase. The immuno-complexes were visualized with enhanced chemiluminescence kits (GE Healthcare). Antibodies used were: for MYC 9E10, # sc-40 (1:500) from Santa Cruz and for actin Anti-β-Actin AC-74, # A5316 (1:5000) from Sigma Aldrich.

**Flow cytometric analysis**. T cell activation with CD3/CD28 activating beads was controlled with CD25 and CD69 activation markers. Approximately 150,000 were fixed in 3% PFA for 10 min at RT. Cells were washed in 1% FBS/PBS and stained with the corresponding antibody for 45 min on ice, (1:50 dilution was used for CD25 FITC, #555431 from BD and CD69 BV510 #310929 from Biolegend). Cells were extensively washed and profiled using BD FACSVerse™ instrument. Gates for activation marker-positive cells were set by utilizing unstained controls. FlowJo software was used for the data analysis. Gating strategy is described in Supplementary Fig. 7.

**Chromatin immunoprecipitation**. In all, $20 \times 10^6$ $CD4^+$ T cells were washed 1 time in PBS prior to crosslinking with 1% formaldehyde for 10 min at RT, followed by termination of the reaction with 125 mM glycine on ice. Cell pellet was washed 2 times with PBS at 4 °C and was lysed in 0.5% NP-40 buffer (10 mM Tris-Cl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 1 mM PMSF, and Protease Inhibitors). For histone ChIPs, obtained nuclei were washed once in the same buffer without NP-40. Nuclei were resuspended in 0.5% NP-40 buffer supplemented with 0.15% SDS and 1.5 mM $CaCl_2$. Nuclei were incubated at 37 °C for 10 min prior to addition of Micrococcal Nuclease (16 units of the enzyme), and the reaction was stopped after 7 min with 3 mM EGTA. DNA was additionally sheared by sonication (Covaris or Bioruptor, Diagenode) to an average size of DNA fragments <500 bps. Extracts were then diluted up to 0.01% SDS, 1% Triton-X, 20 mM Tris pH 8, 150 mM NaCl, and 2 mM EDTA. Extracts were precleared by 1-h incubation with protein A/G Magna ChIP beads at 4 °C and diluted with 5× IP buffer to a final concentration of 140 mM NaCl and 1% NP-40. Lysate corresponding to $3-4 \times 10^6$ million of cells was then incubated with 2–4 µg of the indicated antibody overnight at 4 °C, followed by a 2.5-h incubation with Magna ChIP Protein A/G Magnetic Beads (Millipore). Beads were then washed thoroughly with RIPA150, with LiCl-containing buffer and with TE buffer, RNAse treated for 1 h at 37 °C, and Proteinase K treated for 2 h at 56 °C. Decrosslinking of protein–DNA complexes was performed by an overnight incubation at 65 °C. Additional 1 h of Proteinase K digestion was performed at 56 °C and DNA was then extracted using Agencourt AMPure XP beads (Beckman Coulter) and quantified by real-time PCR. The following antibodies were used for ChIP: H3K27ac (ab4729), H3K4me3 (ab8580) H3K36me3 (ab9050), IgG Rabbit (ab46540).

**ChIP-Seq and RNA-Seq**. ChIP-Seq: Approximately 10 ng of the corresponding inputs and ChIP-ed DNA from primary $CD4^+$ T cells: H3K27ac, H3K4me3, H3K36me3, H4K20me1, and H3K9me2, IPs were prepared for sequencing using the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina®.

RNA-Seq: $5 \times 10^6$ DMSO and 500 nM JQ1-treated $CD4^+$ T cells from three independent donors were used for RNA extraction with the InviTrap Spin Kit (Stratec Biomedical) according to the manufacturer's instructions and libraries for sequencing were prepared by using the rRNA Depletion Kit NEBNext® and NEBNext® Ultra™ RNA Library Prep Kit for Illumina®. Sequencing was performed with $2 \times 75$ bp read length on the NextSeq platform.

**In situ Hi-C protocol**. Hi-C was performed based on the protocol published by Rao et al.[83] with modifications. Briefly, one million cells were crosslinked with 1% formaldehyde for 10 min at RT with gentle rotation. Nuclei were permeabilized by 0.25 ml freshly prepared ice-cold Hi-C lysis buffer [10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal CA630 (Sigma, I8896–50ML), and 1× Roche complete protease inhibitors (Roche, 11836153001)]. DNA was digested with 100 units of MboI (NEB, R0147M) at 37 °C overnight, and the ends of digested fragments were filled in by using 0.4 mM biotinylated deoxyadenosine triphosphate (biotin-14-dATP; Life Technologies, #65001) and ligated in 1 ml by incubating at 24 °C

overnight with gentle rotation. After reversal of the crosslinks, ligated DNA was purified and sheared to a length of 400 bp. Ligation junctions were pulled down with 75 µl of 10 mg/ml streptavidin C1 beads. Ten microliters of DNA-on-beads were amplified in 50 µl standard Herculase II Fusion DNA Polymerase reaction mix (Agilent Technologies, #600675) with 1 µM NEBNext Universal primer and index primer (NEB, E6040S). The cycling conditions were as follows: 98 °C for 2 min; 98 °C for 20 s, 65 °C for 30 s, and 72 °C for 45 s (8 cycles); and 72 °C for 3 min. PCR products were purified with 1.0× Agencourt AMPure XP beads (BECKMAN COULTER, A63880). Libraries ran as a smear on 1.5% agarose gel and estimate of the size of a smear was around 300 bp. The quality of the libraries was assessed by digesting with ClaI (NEB, R0197S) and checking that the smear shifts downwards.

**Reporting summary**. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

All relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding authors on reasonable request. The RNA- Seq of resting and activated CD4$^+$ T cells is available from Gene Expression Omnibus (GEO; Lucic et al. GSE122735), ChIP-Seq on primary CD4$^+$ T cells (Lucic et al. GSE GSE122826), in situ Hi-C data (Chen et al. GSE122958). Integration site raw data on in vitro infected CD4$^+$ T cells are available from GSE134382. A reporting summary for this Article is available as a Supplementary Information file.

## Code availability

Code for processing raw sequences to get integration sites is available here: https://github.com/gui11aume/genome_structure_and_HIV_integration/blob/master/maja/Brady_Integration_Sites.md. Integration sites used in this analysis are available as an R object containing a list of GRanges objects, one list element for each dataset used; https://github.com/gui11aume/genome_structure_and_HIV_integration/blob/master/maja/is.Robj. The matrix that contains the number of lists each gene is found in all 100 randomizations can be found here (in RDS format): https://github.com/gui11aume/genome_structure_and_HIV_integration/raw/master/maja/Replicates.RDS. Genuine Hi-C contacts were validated with the Hi.C pipeline (https://github.com/ezorita/hi.c).

Received: 11 May 2018 Accepted: 19 August 2019
Published online: 06 September 2019

## References

1. Coffin, J. M., Hughes, S. H. & Varmus, H. E. in *Retroviruses* (eds Coffin, J. M., Hughes, S. H. & Varmus, H. E.) (Cold Spring Harbor Laboratory Press, 2011).
2. Craigie, R. & Bushman, F. D. HIV DNA integration. *Cold Spring Harb. Perspect. Med.* **2**, a006890 (2012).
3. Lusic, M. & Siliciano, R. F. Nuclear landscape of HIV-1 infection and integration. *Nat. Rev. Microbiol.* **15**, 69–82 (2017).
4. Sengupta, S. & Siliciano, R. F. Targeting the latent reservoir for HIV-1. *Immunity* **48**, 872–895 (2018).
5. Churchill, M. J., Deeks, S. G., Margolis, D. M., Siliciano, R. F. & Swanstrom, R. HIV reservoirs: what, where and how to target them. *Nat. Rev. Microbiol.* **14**, 55–60 (2016).
6. Chomont, N. et al. HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat. Med.* **15**, 893–900 (2009).
7. Zack, J. A., Kim, S. G. & Vatakis, D. N. HIV restriction in quiescent CD4$^+$ T cells. *Retrovirology* **10**, 37 (2013).
8. Dai, J. et al. Human immunodeficiency virus integrates directly into naive resting CD4+ T cells but enters naive cells less efficiently than memory cells. *J. Virol.* **83**, 4528–4537 (2009).
9. Agosto, L. M. et al. The CXCR4-tropic human immunodeficiency virus envelope promotes more-efficient gene delivery to resting CD4+ T cells than the vesicular stomatitis virus glycoprotein G envelope. *J. Virol.* **83**, 8153–8162 (2009).
10. Pace, M. J. et al. Directly infected resting CD4+T cells can produce HIV Gag without spreading infection in a model of HIV latency. *PLoS Pathog.* **8**, e1002818 (2012).
11. Dahabieh, M. S., Battivelli, E. & Verdin, E. Understanding HIV latency: the road to an HIV cure. *Annu. Rev. Med.* **66**, 407–421 (2015).
12. Lusic, M. & Giacca, M. Regulation of HIV-1 latency by chromatin structure and nuclear architecture. *J. Mol. Biol.* **427**, 688–694 (2015).
13. Suzuki, Y. & Craigie, R. The road to chromatin - nuclear entry of retroviruses. *Nat. Rev. Microbiol.* **5**, 187–196 (2007).
14. Ocwieja, K. E. et al. HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog.* **7**, e1001313 (2011).
15. Di Nunzio, F. et al. Nup153 and Nup98 bind the HIV-1 core and contribute to the early steps of HIV-1 replication. *Virology* **440**, 8–18 (2013).
16. Koh, Y. et al. Differential effects of human immunodeficiency virus type 1 capsid and cellular factors nucleoporin 153 and LEDGF/p75 on the efficiency and specificity of viral DNA integration. *J. Virol.* **87**, 648–658 (2013).
17. Yamashita, M. & Engelman, A. N. Capsid-dependent host factors in HIV-1 infection. *Trends Microbiol.* **25**, 741–755 (2017).
18. Marini, B. et al. Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521**, 227–231 (2015).
19. Lelek, M. et al. Chromatin organization at the nuclear pore favours HIV replication. *Nat. Commun.* **6**, 6483 (2015).
20. Schröder, A. R. W. et al. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
21. Ciuffi, A. et al. A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med* **11**, 1287–1289 (2005).
22. Singh, P. K. et al. LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes Dev.* **29**, 2287–2297 (2015).
23. Cherepanov, P. et al. HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J. Biol. Chem.* **278**, 372–381 (2003).
24. Sowd, G. A. et al. A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc. Natl Acad. Sci. USA* **113**, E1054–E1063 (2016).
25. Vranckx, L. S. et al. LEDGIN-mediated inhibition of integrase-LEDGF/p75 interaction reduces reactivation of residual latent HIV. *EBioMedicine* **8**, 248–264 (2016).
26. Achuthan, V. et al. Capsid-CPSF6 interaction licenses nuclear HIV-1 trafficking to sites of viral DNA integration. *Cell Host Microbe* **24**, 392.e8–404.e8 (2018).
27. Bejarano, D. A. et al. HIV-1 nuclear import in macrophages is regulated by CPSF6-capsid interactions at the nuclear pore complex. *Elife* **8**, e41800 (2019).
28. 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).
29. Chen, H.-C., Martinez, J. P., Zorita, E., Meyerhans, A. & Filion, G. J. Position effects influence HIV latency reversal. *Nat. Struct. Mol. Biol.* **24**, 47–54 (2017).
30. Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
31. Lovén, J. et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).
32. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
33. Parker, S. C. J. et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl Acad. Sci. USA* **110**, 17921–17926 (2013).
34. Hnisz, D., Day, D. S. & Young, R. A. Insulated neighborhoods: structural and functional units of mammalian gene control. *Cell* **167**, 1188–1200 (2016).
35. Witte, S., O'Shea, J. J. & Vahedi, G. Super-enhancers: asset management in immune cell genomes. *Trends Immunol.* **36**, 519–526 (2015).
36. Roychoudhuri, R. et al. BACH2 represses effector programs to stabilize T (reg)-mediated immune homeostasis. *Nature* **498**, 506–510 (2013).
37. Tsukumo, S.-I. et al. Bach2 maintains T cells in a naive state by suppressing effector memory-related genes. *Proc. Natl Acad. Sci. USA* **110**, 10735–10740 (2013).
38. Wagner, T. A. et al. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570–573 (2014).
39. Maldarelli, F. et al. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179–183 (2014).
40. Toda, T. et al. Nup153 interacts with Sox2 to enable bimodal gene regulation and maintenance of neural progenitor cells. *Cell Stem Cell* **21**, 618–634.e7 (2017).
41. Ibarra, A., Benner, C., Tyagi, S., Cool, J. & Hetzer, M. W. Nucleoporin-mediated regulation of cell identity genes. *Genes Dev.* **30**, 2253–2258 (2016).
42. Rao, S. et al. Cohesin loss eliminates all loop domains, leading to links among superenhancers and downregulation of nearby genes. *Cell.* **171**, 305.e24–320.e24 (2017).
43. Olley, G. et al. BRD4 interacts with NIPBL and BRD4 is mutated in a Cornelia de Lange-like syndrome. *Nat. Genet.* **50**, 329–332 (2018).
44. Beagrie, R. A. et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).
45. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 772 (2016).
46. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
47. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

48. Noordermeer, D. et al. The dynamic architecture of Hox gene clusters. *Science* **334**, 222–225 (2011).
49. Di Primio, C. et al. Single-cell imaging of HIV-1 provirus (SCIP). *Proc. Natl Acad. Sci. USA* **110**, 5636–5641 (2013).
50. Brady, T. et al. HIV integration site distributions in resting and activated CD4 + T cells infected in culture. *AIDS* **23**, 1461–1471 (2009).
51. Han, Y. et al. Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. *J. Virol.* **78**, 6122–6133 (2004).
52. Kok, Y. L. et al. Monocyte-derived macrophages exhibit distinct and more restricted HIV-1 integration site repertoire than CD4(+) T cells. *Sci. Rep.* **6**, 24157 (2016).
53. Cohn, L. B. et al. HIV-1 integration landscape during latent and active infection. *Cell* **160**, 420–432 (2015).
54. Ikeda, T., Shibata, J., Yoshimura, K., Koito, A. & Matsushita, S. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. *J. Infect. Dis.* **195**, 716–725 (2007).
55. Berry, C., Hannenhalli, S., Leipzig, J. & Bushman, F. D. Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput. Biol.* **2**, e157 (2006).
56. Brady, T. et al. Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev.* **23**, 633–642 (2009).
57. Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931–21936 (2010).
58. Heintzman, N. D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
59. Wagner, E. J. & Carpenter, P. B. Understanding the language of Lys36 methylation at histone H3. *Nat. Rev. Mol. Cell Biol.* **13**, 115–126 (2012).
60. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
61. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* **44**, D164–D171 (2016).
62. Shao, W. et al. Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes. *Retrovirology* **13**, 47 (2016).
63. LaFave, M. C. et al. MLV integration site selection is driven by strong enhancers and active promoters. *Nucleic Acids Res.* **42**, 4257–4269 (2014).
64. Filippakopoulos, P. et al. Selective inhibition of BET bromodomains. *Nature* **468**, 1067–1073 (2010).
65. Roukos, V. et al. Spatial dynamics of chromosome translocations in living cells. *Science* **341**, 660–664 (2013).
66. Roukos, V. & Misteli, T. Deep imaging: the next frontier in microscopy. *Histochem. Cell Biol.* **142**, 125–131 (2014).
67. Hughes, S. H. & Coffin, J. M. What integration sites tell us about HIV persistence. *Cell Host Microbe* **19**, 588–598 (2016).
68. Chen, Y. et al. Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J. Cell Biol.* **217**, 4025–4048 (2018).
69. Bickmore, W. A. The spatial organization of the human genome. *Annu. Rev. Genom. Hum. Genet.* **14**, 67–84 (2013).
70. Rasheedi, S. et al. The cleavage and polyadenylation specificity factor 6 (CPSF6) subunit of the capsid-recruited pre-messenger RNA cleavage factor I (CFIm) complex mediates HIV-1 integration into genes. *J. Biol. Chem.* **291**, 11809–11819 (2016).
71. Blow, M. J. et al. ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42**, 806–810 (2010).
72. Visel, A. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
73. Vahedi, G. et al. Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature* **520**, 558–562 (2015).
74. Witte, S., Bradley, A., Enright, A. J. & Muljo, S. A. High-density P300 enhancers control cell state transitions. *BMC Genomics* **16**, 903 (2015).
75. Cereseto, A. et al. Acetylation of HIV-1 integrase by p300 regulates viral integration. *EMBO J.* **24**, 3070–3081 (2005).
76. Sharma, A. et al. BET proteins promote efficient murine leukemia virus integration at transcription start sites. *Proc. Natl Acad. Sci. USA* **110**, 12036–12041 (2013).
77. De Rijck, J. et al. The BET family of proteins targets moloney murine leukemia virus integration near transcription start sites. *Cell Rep.* **5**, 886–894 (2013).
78. Boehm, D. et al. BET bromodomain-targeting compounds reactivate HIV from latency via a Tat-independent mechanism. *Cell Cycle* **12**, 452–462 (2013).
79. Li, Z., Guo, J., Wu, Y. & Zhou, Q. The BET bromodomain inhibitor JQ1 activates HIV latency through antagonizing Brd4 inhibition of Tat-transactivation. *Nucleic Acids Res.* **41**, 277–287 (2013).
80. Conrad, R. J. et al. The short isoform of BRD4 promotes HIV-1 latency by engaging repressive SWI/SNF chromatin-remodeling complexes. *Mol. Cell* **67**, 1001–1012.e6 (2017).
81. Rouzine, I. M., Weinberger, A. D. & Weinberger, L. S. An evolutionary role for HIV latency in enhancing viral transmission. *Cell* **160**, 1002–1012 (2015).
82. Razooky, B. S., Pai, A., Aull, K., Rouzine, I. M. & Weinberger, L. S. A hardwired HIV latency program. *Cell* **160**, 990–1001 (2015).
83. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping (Erratum). *Cell* **162**, 687–688 (2015).


## Acknowledgements

We thank the Infectious Diseases Imaging Platform (IDIP) and the platform coordinator Dr. Vibor Laketa (DZIF), as well as the genomics core facility of the CRG for their technical support. We also thank Monsef Benkirane and Thomas Gayraud from the IGH Montpellier for providing samples of cell sorted infected primary CD4+ T cell. This work was supported by German Center for Infection Research (DZIF) Thematic Translational Unit HIV-1 04.704 Infrastructural Measure to M.L. and by the Hector Grant M70 "HiPNose: HiV Positioning in the Nuclear Space" to M.L. and M.S. We acknowledge the financial support of the Spanish Ministry of Economy and Competitiveness ("Centro de Excelencia Severo Ochoa 2013–2017," Plan Nacional BFU2012–37168), of the CERCA (Centres de Recerca de Catalunya) Programme/Generalitat de Catalunya, and of the European Research Council (Synergy Grant 609989). K.V. and M.K. are supported by the European Structural and Investment Funds grant for the Croatian National Centre of Research Excellence in Personalized Healthcare (contract #KK.01.1.1.01.0010), Croatian National Centre of Research Excellence for Data Science and Advanced Cooperative Systems (contract KK.01.1.1.01.0009), and Croatian Science Foundation (grant IP-2014–09–6400).

## Author contributions

M.L, B.L. and G.F. designed the research; B.L., H.C., J.W., V.M. and W.W. performed the experiments; M.K., R.S., E.Z., R.F., K.V., and G.F. performed bioinformatics analysis; B.L., M.K., E.Z., V.R., K.V., G.F. and M.L analyzed the data; B.L., M.K., G.F. and M.L. wrote the manuscript; M.L., G.F., K.V., M.S. and S.L. provided funding.

## Additional information

**Supplementary Information** accompanies this paper at https://doi.org/10.1038/s41467-019-12046-3.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

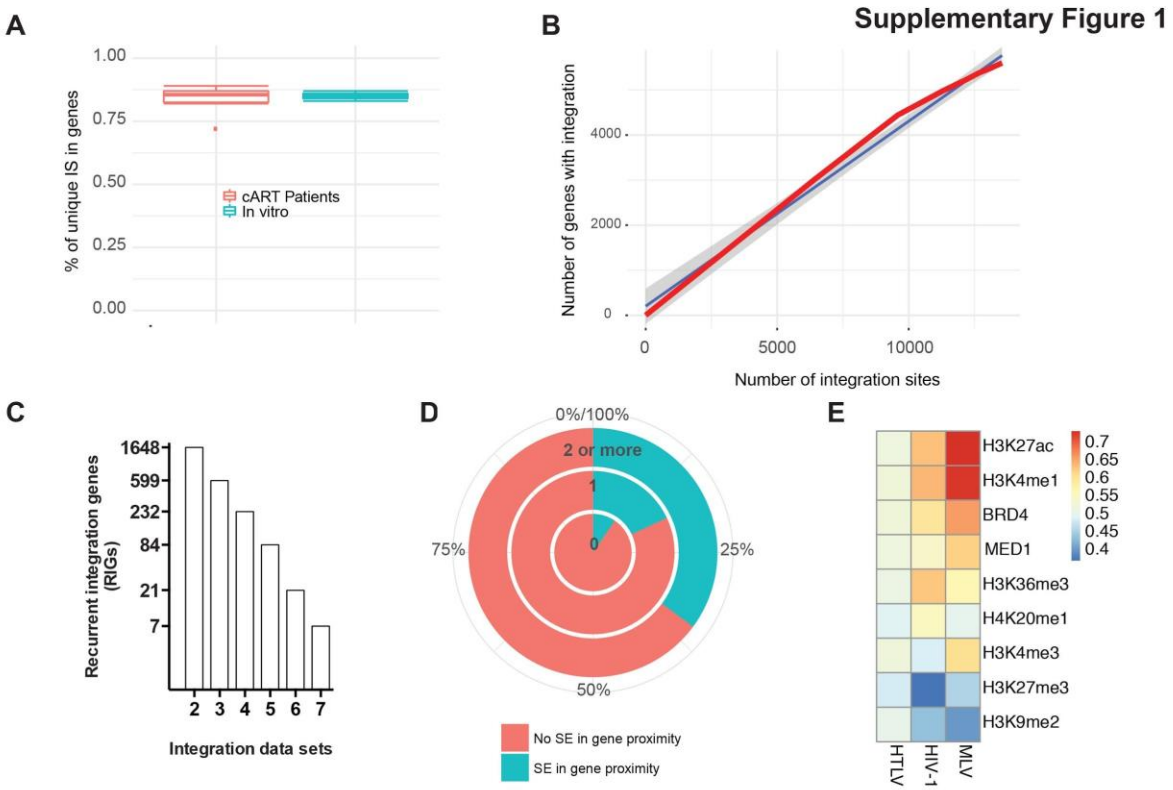
**Peer review information**: *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**NCOMMS-18-14402D**
**Supplementary information**

# Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration

Lucic et al.




**Supplementary Figure 1.**

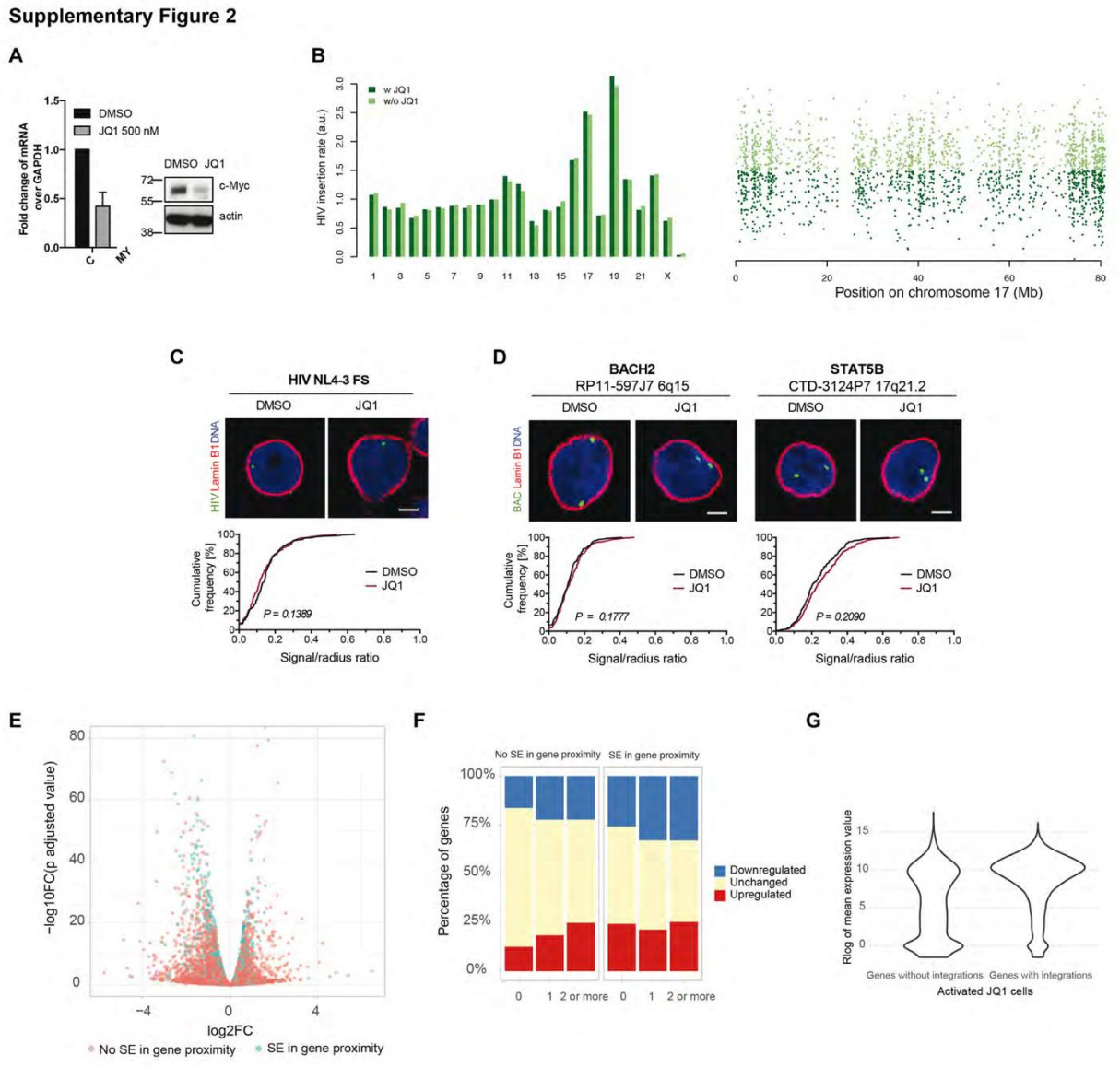
**A)** HIV-1 integration sites inside genes. The box plot represents the percentage of integration sites inside genes for cART-treated patients in rose (6 lists in total) and in vitro infections in blue (2 lists in total). The whiskers stretch from 5$^{th}$ to 95$^{th}$ percentile.

**B)** Analysis of the number of unique genes containing integration sites versus number of observed integrations. All integration data sets are sorted by decreasing size and the cumulative number of integration sites is plotted on the X-axis while the Y-axis shows the number of unique genes that have integrations. The number of unique genes found when analyzing different numbers of data sets linearly depends on the number of integrations in the observed data sets (adjusted $R^2 = 0.9896$, $p = 2.075e^{-8}$).

**C)** HIV-1 recurrently integrates into a subset of genes. The bar plot represents the number of genes (RIGs) shared among at least x different data sets. The number of RIGs shared among different data sets decreases exponentially as more data sets are taken into consideration.

**D)** Bar plot showing the percentage of genes that have a super-enhancer in proximity, in groups of genes on 0 list (without HIV-1 integrations), genes on 1 list and genes in 2 or more lists (RIGs).

**E)** ROC analysis represented as heatmap summarizing the co-occurrence density of integration sites and epigenetic modification obtained by ChIP-Seq for H3K27ac, H3K4me1, BRD4, MED1, H3K36me3, H4K20me1, H3K4me3, H3K27me3 and H3K9me2. HTLV, HIV-1 and MLV integration data sets are shown in the columns, and epigenetic modifications are shown in rows. Associations are quantified using the ROC area method; values of ROC areas are shown in the color key at the right.




**Supplementary Figure 2.**

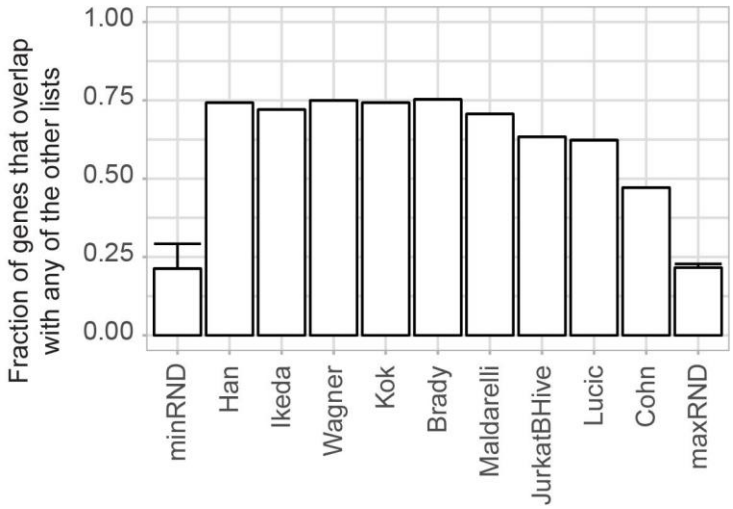
**A)** mRNA expression profiles and protein levels of MYC upon JQ1 treatment (500 nM JQ1 for 6 h). The mRNA levels are normalized over GAPDH, and mean and standard deviation are derived from three independent experiments. Representative protein levels of c-Myc and actin are shown on the western blot.

**B)** Bar plot of HIV-1 insertion rate per chromosome: control (w/o JQ1) or JQ1 treated (w JQ1) samples (left panel) and 'insertion cloud' representation on chromosome 17 with characteristic 3-fold enrichment (right panel). Each dot represents an HIV-1 insertion site. The x-coordinate indicates to the location of the insertion site on chromosome 17; the y-coordinate is random so that insertion hotspots appear as vertical lines. The insertion profile upon JQ1 treatment was flipped vertically

**C)** 3D immuno-DNA FISH images of HIV-1 in activated $CD4^+$ T cells pretreated with 500 nM JQ1 for 6 h and infected for 72 h (green: HIV-1 probe, red: lamin B1, blue: DNA staining with Hoechst 33342, scale bar represents 2 µm). Cumulative frequency plots show combined data

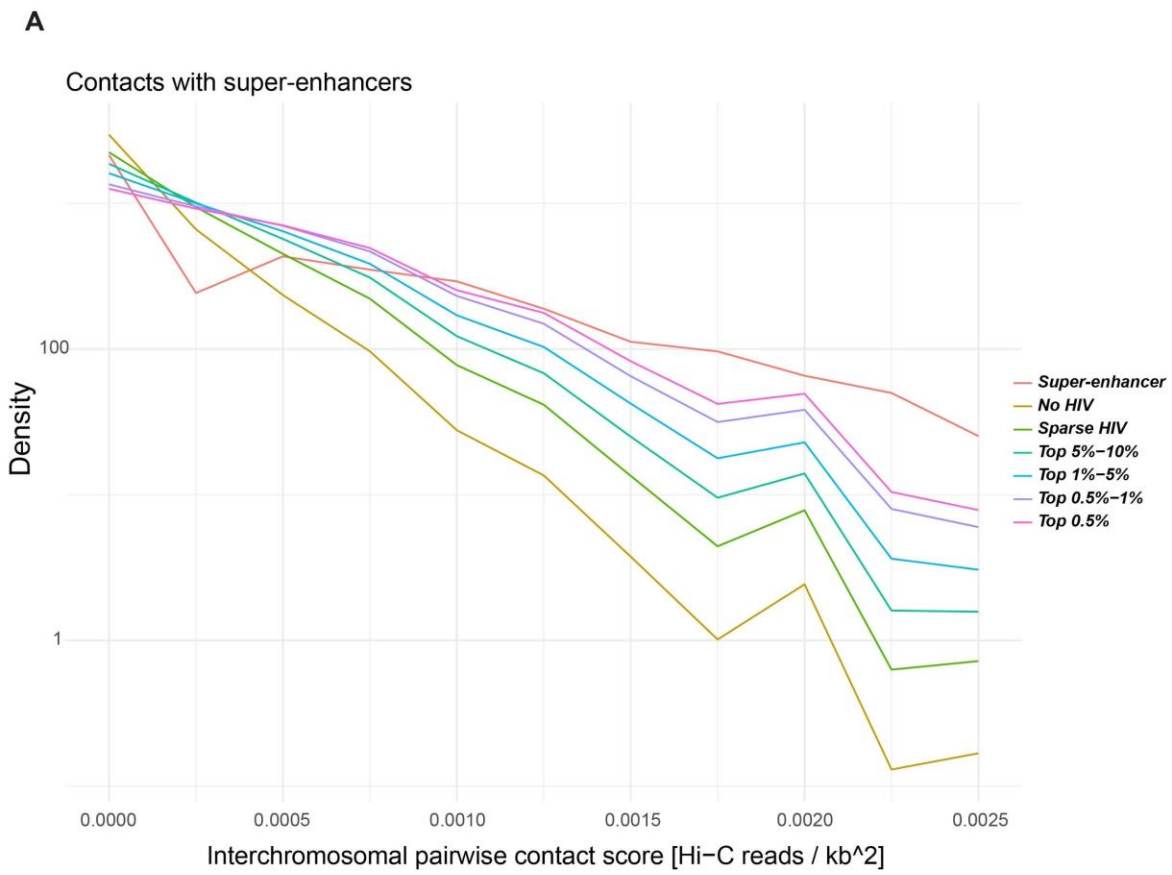
from both experiments (n = 100, black: DMSO, red: JQ1). The p-values of the Kolmogorov-Smirnov tests are indicated.
**D)** 3D immuno-DNA FISH images of *BACH2* and *STAT5B* upon 500 nM JQ1 treatment for 6 h in activated $CD4^+$ T cells (green: BAC/gene probe, red: lamin B1, blue: DNA staining with Hoechst 33342, scale bar represents 2 µm).
Regularized log transformed read counts on protein coding genes averaged over three replicates in activated JQ1 treated cells shown as violin plot for genes grouped by presence of HIV-1 integration in activated JQ1 treated cells.
**E)** Volcano plot showing the changes in mRNA levels of protein coding genes upon JQ1 treatment with respect to the vicinity to super-enhancers.
**F)** Bar plot showing the percentage of protein coding genes that are downregulated, unchanged, and upregulated upon JQ1 treatment. Genes are grouped by number of lists they occur in and by the presence or absence of super-enhancer in either gene body or 5 kb upstream.
**G)** Regularized log transformed read counts on protein coding genes averaged over three replicates in activated JQ1 treated cells shown as violin plot for genes grouped by presence of HIV-1 integration in activated JQ1 treated cells.

**Supplementary Figure 3**



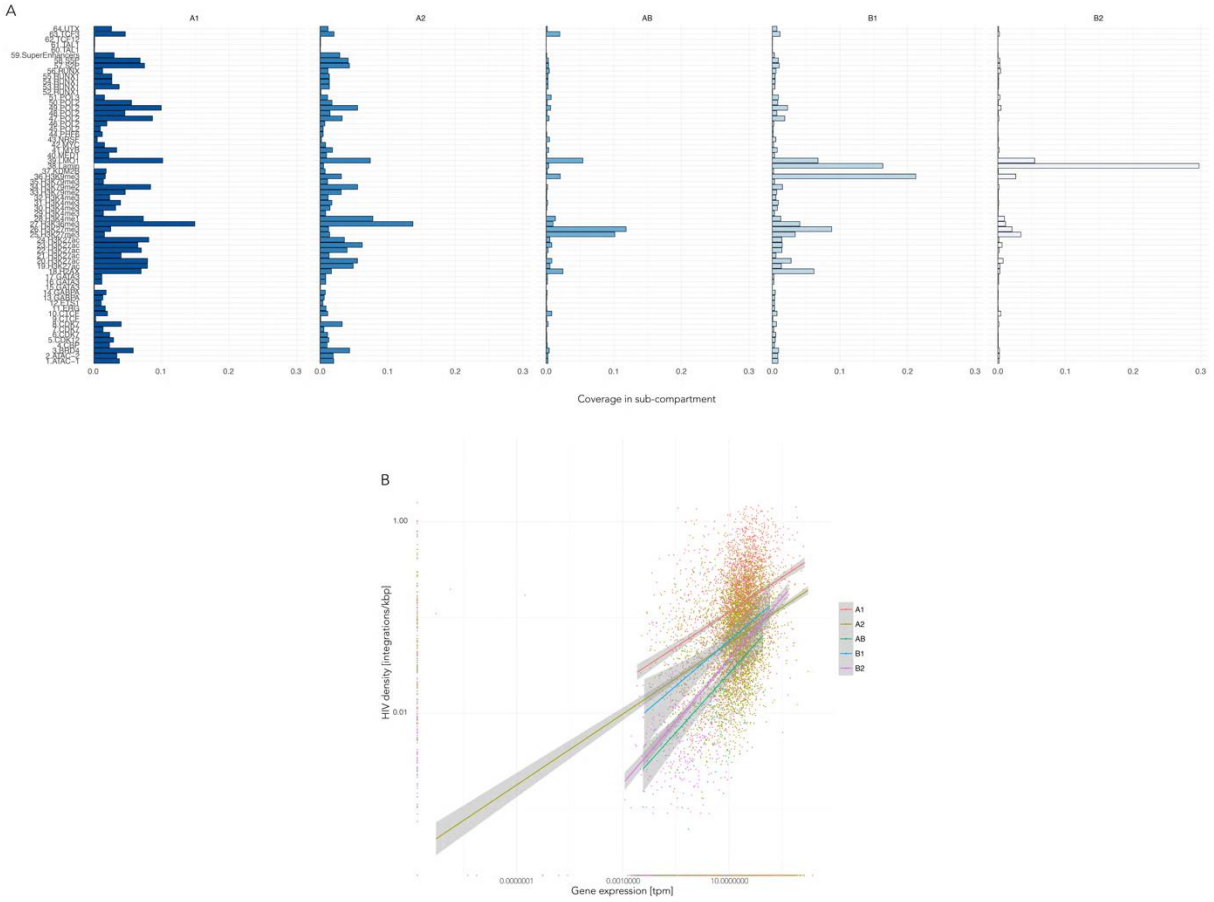

**Supplementary Figure 3.**
Comparison of insertion sites in $CD4^+$ T cells with B-HIVE insertion sites in Jurkat. The fraction of genes from a data set that is shared with at least one other data set is shown on the Y-axis. The bar plot shows that different data sets share most of targeted genes among each other, while randomly chosen subsets of genes (minRND and maxRND) are only partially shared with genes from other data sets.

**Supplementary Figure 4**



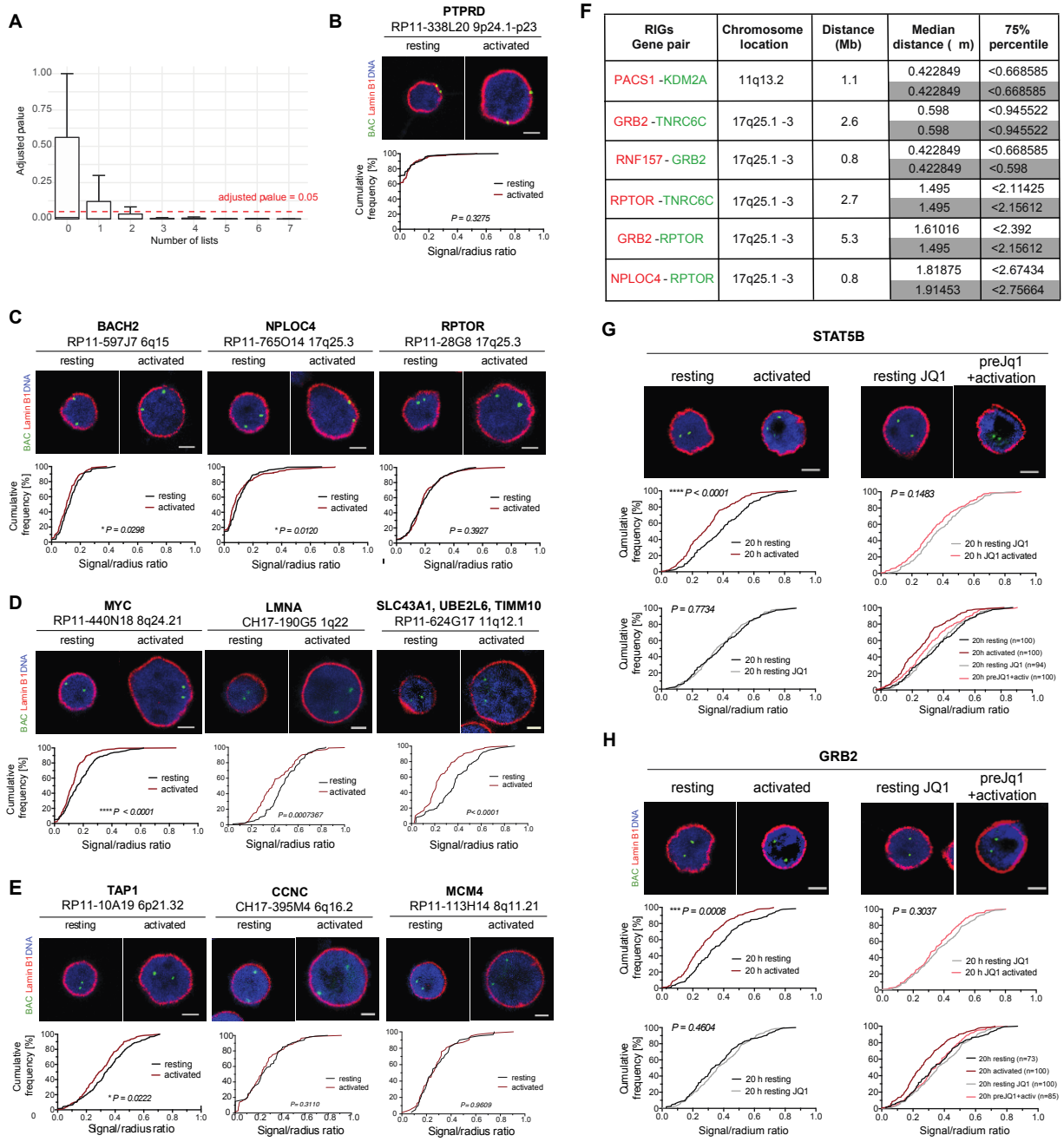

**Supplementary Figure 4.**
**A)** Density plot of inter-chromosomal pairwise contact scores (see Methods) between super-enhancers and HIV hotspots. HIV hotspots were identified within genomic bins of 100 kb, then sorted by count of HIV insertions and split in percentile groups, *e.g.*, top 0.5% in HIV-1 count, from 0.5% to 1%, *etc*. Super-enhancers show strongest inter-chromosomal contacts with other super-enhancers (observe the higher density at higher contact scores, red line), followed by contacts between super-enhancers and the most HIV-dense hotspots (pink line). The interaction scores show a monotonous decay as the HIV-1 hotspots are more sparse.

**Supplementary Figure 5**




**Supplementary FIgure 5.**
**A)** Proportion of 3D sub-compartments covered by Jurkat chromatin features available from the literature. Coverage was computed as the span of enriched ChIP-Seq signal divided by the sub-compartment size.
**B)** Scatter plot of HIV density in gene bodies versus endogenous expression in Jurkat cells. Each dot represents a protein coding gene. Dot colors identify the 5 different sub-compartments. Sub-compartments A1 and A2 show similar distributions of gene expression and almost identical effects of gene expression on HIV-1 density (see slopes of linear models). However, A1 shows higher HIV-1 density overall compared to A2 (see vertical shift of fitted lines), suggesting an intrinsic preference of HIV-1 for A1 independent of the gene expression level.

## Supplementary Figure 6




F

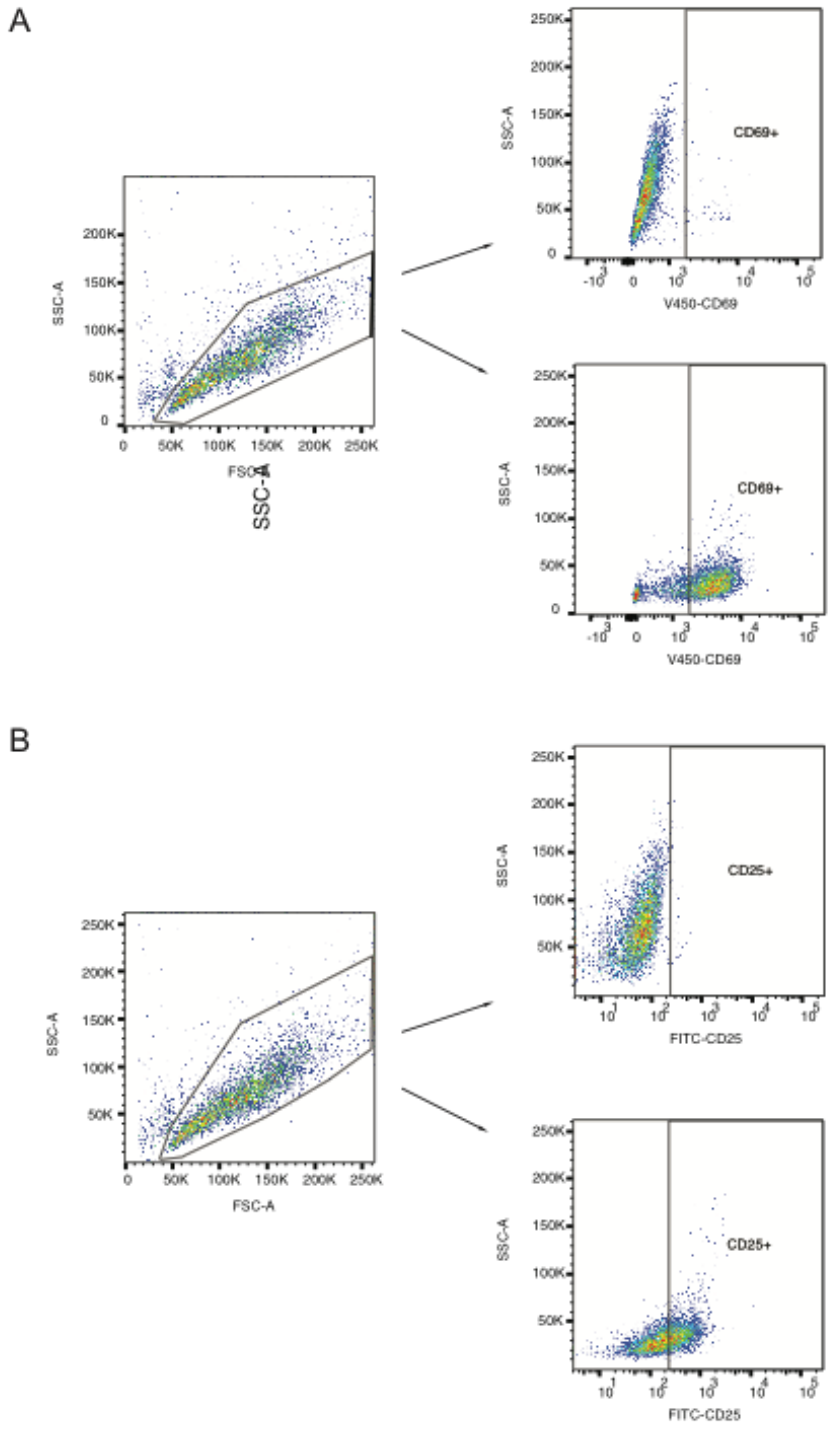
| RIGs Gene pair | Chromosome location | Distance (Mb) | Median distance ( m) | 75% percentile |
|---|---|---|---|---|
| PACS1 -KDM2A | 11q13.2 | 1.1 | 0.422849 | <0.668585 |
| | | | 0.422849 | <0.668585 |
| GRB2 -TNRC6C | 17q25.1 -3 | 2.6 | 0.598 | <0.945522 |
| | | | 0.598 | <0.945522 |
| RNF157 - GRB2 | 17q25.1 -3 | 0.8 | 0.422849 | <0.668585 |
| | | | 0.422849 | <0.598 |
| RPTOR -TNRC6C | 17q25.1 -3 | 2.7 | 1.495 | <2.11425 |
| | | | 1.495 | <2.15612 |
| GRB2 -RPTOR | 17q25.1 -3 | 5.3 | 1.61016 | <2.392 |
| | | | 1.495 | <2.15612 |
| NPLOC4 - RPTOR | 17q25.1 -3 | 0.8 | 1.81875 | <2.67434 |
| | | | 1.91453 | <2.75664 |

**Supplementary Figure 6.**

**A)** Adjusted p-value for change in expression of genes upon activation of CD4$^+$ T cells. Genes are grouped by number of HIV-1 lists they appear in. The dashed red line represents an adjusted p-value of 0.05.

3D immuno-DNA FISH in resting and activated CD4$^+$ T cells. Representative images of **B)** *PTPRD.* **C)** *BACH2*, *NPLOC4* and *RPTOR*. **D)** *MYC*, *LMNA* and *SLC43A1*, *UBE2L6*, *TIMM10*. **E)** *TAP1*, *CCNC* and *MCM4.* **F)** Table summarizing the spatial relationships on chromosome 11 and 17 in resting and activated CD4$^+$ T cells.

Effect of super-enhancer disruption by JQ1 on the T cell activation-induced movement of RIGs : 3D immuno-DNA FISH images of *STAT5B* **G)** and *GRB2* **H)** in $CD4^+$ T cells treated with 500 nM JQ1 or DMSO, and activated for 20 h: Green: gene, red: lamin B1, blue: DNA counterstaining with Hoechst 33342. Cumulative frequency plots in the lower panels show combined data from both experiments (n = 100, black: resting cells, red: activated cells). The p-values from the Kolmogorov-Smirnov tests are indicated.




**Supplementary Figure 7.**
Gating strategy used to determine non activated (upper right panels) and activated cells (lower panels) with two markers, CD69(A) and CD25(B). Here shown is an example of cell activation of 20hrs, as explained in Supplementary Figure 6. Population of live unstained activated cells was used to set the gates: side scatter plus specific marker was used for activated T cell population identification.

| | Lucic | Maldarelli | Cohn | Han | Ikeda | Wagner | Kok | Brady | In Vitro | Patients | Total CD4 HIV |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Total number of unique IS | 3167 | 1723 | 6416 | 74 | 366 | 443 | 497 | 864 | 4031 | 9519 | 13546 |
| # of unique IS in genes | 2644 | 1506 | 4644 | 61 | 320 | 370 | 444 | 750 | 3394 | 7345 | 10735 |
| % of unique IS in genes | 83 | 87 | 72 | 82 | 87 | 84 | 89 | 87 | 84 | 77 | 79 |
| Total number of genes | 57763 | 57763 | 57763 | 57763 | 57763 | 57763 | 57763 | 57763 | 57763 | 57763 | 57763 |
| # of genes with IS | 2158 | 1199 | 3012 | 66 | 305 | 324 | 445 | 683 | 2579 | 4253 | 5601 |
| % of genes with IS | 4 | 2 | 5 | 0 | 1 | 1 | 1 | 1 | 4 | 7 | 1 |
| # of genes without IS | 55605 | 56564 | 54751 | 57697 | 57458 | 57439 | 57318 | 57080 | 55184 | 53510 | 52162 |

**Supplementary Table 1.**
HIV-1 integration sites from primary $CD4^+$ T cells.
Number of unique integration sites from each list used in this study (Total number of unique IS), number and percentage of unique IS in genes (# and % of unique IS in genes), the total number of genes used in this study (Total number of genes), number and percentage of genes with integration sites used in this study (# and % of genes with IS) and number of genes without integration isets (# of genes without IS).

| Primer ID | Sequence | FW or RV | Index | Description |
|---|---|---|---|---|
| GAT1786 | 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTTGGGAGTGAATTAGCCCTT-3' | RV | - | |
| GAT1787 | 5'-CAAGCAGAAGACGGCATACGAGATTAAATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | TAAATC | |
| GAT1788 | 5'-CAAGCAGAAGACGGCATACGAGATAATGCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | AATGCG | |
| GAT1789 | 5'-CAAGCAGAAGACGGCATACGAGATACTCGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | ACTCGG | |
| GAT1790 | 5'-CAAGCAGAAGACGGCATACGAGATTGTTTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | TGTTTG | |
| GAT1791 | 5'-CAAGCAGAAGACGGCATACGAGATGACTCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | GACTCC | Isolation HIV integration sites from 5'LTR |
| GAT1792 | 5'-CAAGCAGAAGACGGCATACGAGATAGCGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | AGCGGA | |
| GAT1793 | 5'-CAAGCAGAAGACGGCATACGAGATTTCGTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | TTCGTC | |
| GAT1794 | 5'-CAAGCAGAAGACGGCATACGAGATGAAACAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | GAAACA | |
| GAT1795 | 5'-CAAGCAGAAGACGGCATACGAGATTAACCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | TAACCG | |
| GAT1796 | 5'-CAAGCAGAAGACGGCATACGAGATTCTGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | TCTGGA | |
| GAT1797 | 5'-CAAGCAGAAGACGGCATACGAGATAGGGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | AGGGAC | |
| GAT1798 | 5'-CAAGCAGAAGACGGCATACGAGATTCATTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCCAGGGGTCAGATAT-3' | FW | TCATTA | |
| GAT1799 | 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGTCAGTGTGGAAAATCTCTAG-3' | FW | - | |
| GAT1800 | 5'-CAAGCAGAAGACGGCATACGAGATTAGTTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | TAGTTT | |
| GAT1801 | 5'-CAAGCAGAAGACGGCATACGAGATGCTAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | GCTAAG | |
| GAT1802 | 5'-CAAGCAGAAGACGGCATACGAGATGGACGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | GGACGA | |
| GAT1803 | 5'-CAAGCAGAAGACGGCATACGAGATCTCGAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | CTCGAA | |
| GAT1804 | 5'-CAAGCAGAAGACGGCATACGAGATCGTACCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | CGTACC | Isolation HIV integration sites from 3'LTR |
| GAT1805 | 5'-CAAGCAGAAGACGGCATACGAGATCCTTACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | CCTTAC | |
| GAT1806 | 5'-CAAGCAGAAGACGGCATACGAGATCCGTGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | CCGTGG | |
| GAT1807 | 5'-CAAGCAGAAGACGGCATACGAGATCGCCGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | CGCCGG | |
| GAT1808 | 5'-CAAGCAGAAGACGGCATACGAGATAGTTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | AGTTCT | |
| GAT1809 | 5'-CAAGCAGAAGACGGCATACGAGATAGACCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | AGACCC | |
| GAT1810 | 5'-CAAGCAGAAGACGGCATACGAGATGGATTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | GGATTT | |
| GAT1811 | 5'-CAAGCAGAAGACGGCATACGAGATGGTAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACTACTTTGAGCACTCAAG-3' | RV | GGTAGT | |

**Supplementary Table 2.**
List of primers used in inverse PCR to map insertion sites in primary $CD4^+$ T cells. Primer ID, Sequence, FW or RW, Indexes and description.

| ChIP | Source | Conducted on | Experiment | Accession number |
|---|---|---|---|---|
| BRD4 | Hertweck A et al. | | Homo sapiens Th1 RA Brd4 | SRR2971477 |
| Input | Uni W. Reference Epigenome Mapping Project | - | ChIP-Seq Input of CD4 Primary Cells | SRR787508 |
| Input | Uni W. Reference Epigenome Mapping Project | - | ChIP-Seq Input of CD4 Primary Cells | SRR787509 |
| Input | This study | 05.2015. | ChIP-Seq Input of CD4 Primary Cells | |
| Input | This study | 05.2015. | ChIP-Seq Input of CD4 Primary Cells | |
| Input | This study | 08.2015. | ChIP-Seq Input of CD4 Primary Cells | |
| Input | This study | 06.2016. | ChIP-Seq Input of CD4 Primary Cells | |
| Input | This study | 07.2016. | ChIP-Seq Input of CD4 Primary Cells | |
| H3K27ac | This study | 07.2016. | Histone H3K27ac ChIP-Seq of CD4 Primary Cells | |
| H3K27ac | This study | 07.2016. | Histone H3K27ac ChIP-Seq of CD4 Primary Cells | |
| H3K27ac | This study | 08.2015. | Histone H3K27ac ChIP-Seq of CD4 Primary Cells | |
| H3K27ac | This study | 08.2015. | Histone H3K27ac ChIP-Seq of CD4 Primary Cells | |
| H3K36me3 | This study | 07.2016. | Histone H3K27ac ChIP-Seq of CD4 Primary Cells | |
| H3K36me3 | This study | 07.2016. | Histone H3K27ac ChIP-Seq of CD4 Primary Cells | |
| H3K27me3 | Uni W. Reference Epigenome Mapping Project | - | Histone H3K27me3 ChIP-Seq of CD4 Primary Cells | SRR980445 |
| H3K27me3 | Uni W. Reference Epigenome Mapping Project | - | Histone H3K27me3 ChIP-Seq of CD4 Primary Cells | SRR980446 |
| H3K4me1 | Uni W. Reference Epigenome Mapping Project | | Histone H3K4me1 ChIP-Seq of CD4 Primary Cells | SRR980417 |
| H3K4me1 | Uni W. Reference Epigenome Mapping Project | | Histone H3K4me1 ChIP-Seq of CD4 Primary Cells | SRR980418 |
| H3K4me3 | This study | 05.2015. | Histone H3K4me3 ChIP-Seq of CD4 Primary Cells | |
| H3K4me3 | This study | 05.2015. | Histone H3K4me3 ChIP-Seq of CD4 Primary Cells | |
| H3K4me3 | This study | 08.2015. | Histone H3K4me3 ChIP-Seq of CD4 Primary Cells | |
| H3K9me2 | This study | 05.2015. | Histone H3K9me2 ChIP-Seq of CD4 Primary Cells | |
| H3K9me2 | This study | 08.2015. | Histone H3K9me2 ChIP-Seq of CD4 Primary Cells | |
| H3K9me2 | This study | 08.2015. | Histone H3K9me2 ChIP-Seq of CD4 Primary Cells | |
| H4K20me1 | This study | 08.2015. | Histone H4K20me1 ChIP-Seq of CD4 Primary Cells | |
| H4K20me1 | This study | 08.2015. | Histone H4K20me1 ChIP-Seq of CD4 Primary Cells | |
| MED1 | Hertweck A et al. | - | Homo sapiens Th1 RA Med1 | SRR2971479 |

**Supplementary Table 3.**
List of ChIP-Seq data.
All immuno-precipitated factors, source, the date experiment was performed, the name of the experiment and accession numbers.

| SAMPLE | Concentration(ng/ul) | Volume(ul) | Sample_Information |
|---|---|---|---|
| 1a | 56 | 14 | donor 41 CD4T cells CTRL resting |
| 1b | 56 | 14 | donor 41 CD4T cells CTRL resting |
| 2a | 70 | 14 | donor 41 CD4T cells JQ1 resting |
| 2b | 70 | 14 | donor 41 CD4T cells JQ1 resting |
| 3a | 95 | 12 | donor 42 CD4T cells CTRL resting |
| 3b | 95 | 12 | donor 42 CD4T cells CTRL resting |
| 4a | 82 | 12 | donor 42 CD4T cells JQ1 resting |
| 4b | 82 | 12 | donor 42 CD4T cells JQ1 resting |
| 5a | 111 | 12 | donor 44 CD4T cells CTRL resting |
| 5b | 111 | 12 | donor 44 CD4T cells CTRL resting |
| 6a | 81 | 12 | donor 44 CD4T cells JQ1 resting |
| 6b | 81 | 12 | donor 44 CD4T cells JQ1 resting |
| 7a | 842 | 10 | donor 41 CD4T cells CTRL activated |
| 7b | 842 | 10 | donor 41 CD4T cells CTRL activated |
| 8a | 576 | 10 | donor 41 CD4T cells JQ1 activated |
| 8b | 576 | 10 | donor 41 CD4T cells JQ1 activated |
| 9a | 522 | 10 | donor 42 CD4T cells CTRL activated |
| 9b | 522 | 10 | donor 42 CD4T cells CTRL activated |
| 10a | 470 | 10 | donor 42 CD4T cells JQ1 activated |
| 10b | 470 | 10 | donor 42 CD4T cells JQ1 activated |
| 11a | 882 | 10 | donor 44 CD4T cells CTRL activated |
| 11b | 882 | 10 | donor 44 CD4T cells CTRL activated |
| 12a | 684 | 10 | donor 44 CD4T cells JQ1 activated |
| 12b | 684 | 10 | donor 44 CD4T cells JQ1 activated |

**Supplementary Table 4**
List of RNA-Seq data.
Sample number, sample concentrations, volume and sample ID.

| Covered gene | BAC clone | Posi1on GRCh37 | Chr bands | Posi1on GRCh38 |
|---|---|---|---|---|
| BACH2 | RP11*597J7 | chr6:90,681,035*90,881,332 | q15 | chr6:89,971,316*90,171,613 |
| FOXP1 | RP11*905F6 | chr3:71,256,411*71,475,547 | p13 | chr3:71,207,260*71,426,396 |
| GRB2 | RP11*16C1 | chr17:73,269,580*73,422,789 | q25.1 | chr17:75,273,499*75,426,708 |
| KDM2A | RP11*157K17 | chr11:66,913,936*67,086,969 | q13.2 | chr11:67,146,465*67,319,498 |
| MKL2 | RP11*1072B15 | chr16:14,245,800*14,422,293 | p13.12 | chr16:14,151,943*14,328,436 |
| MYC | RP11*440N18 | chr8:128,596,756*128,777,986 | q24.21 | chr8:127,584,511*127,765,740 |
| NFATC3 | RP11*67A1 | chr16:68,111,243*68,156,174 | q22.1 | chr16:68,077,340*68,122,271 |
| NPLOC4 | RP11*765O14 | chr17:79,379,432*79,579,283 | q25.3 | chr17:81,405,632*81,612,257 |
| PACS1 | RP11*675B4 | chr11:65,815,062*65,953,271 | q13.1*q13.2 | chr11:66,047,591*66,185,800 |
| PTPRD | RP11*338L20 | chr9:8,981,678*9,142,717 | p24.1*p23 | chr9:8,981,678*9,142,717 |
| RNF157 | RP11*449J21 | chr17:73,999,159*74,183,053 | q25.1 | chr17:76,003,078*76,186,972 |
| RPTOR | RP11*28G8 | chr17:78,705,399*78,868,353 | q25.3 | chr17:80,731,599*80,894,553 |
| STAT5B | CTD*3124P7 | chr17:40,326,868*40,479,760 | q21.2 | chr17:42,174,850*42,327,742 |
| TAP1 | RP11*10A19 | chr6:32,735,717*32,915,875 | p21.32 | chr6:32,767,940*32,948,098 |
| TNRC6C | RP11*153A23 | chr17:76,004,952*76,182,689 | q25.3 | chr17:78,008,871*78,186608 |
| LMNA | CH17*190G5 | chr1:156,020,485*156,242,095 | 1q22 | chr1:156,050,694*156,272,304 |
| SLC43A1, UBE2L6, TIMM10 | RP11*624G17 | chr11:57,196,976*57,407,534 | 11q12.1 | chr11:57,429,503*57,640,061 |
| CCNC | CH17*395M4 | chr6:99,928,076*100,138,693 | 6q16.2 | chr6:99,480,200*99,690,817 |
| MCM4 | RP11*113H14 | chr8:48,815,695*48,977,857 | 8q11.21 | chr8:47,903,135*48,065,297 |
| MSN | CH17*413H2 | chrX:64,779,881*64,988,569 | Xq12 | chrX:65,560,001*65,768,727 |
| RECQL | RP11*501E24 | chr12:21,521,462*21,700,905 | 12p12.1 | chr12:21,368,528*21,547,971 |
| MARCH1 | CH17*454P4 | chr4:164,780,476*164,991,773 | 4q32.3 | chr4:163,859,324*164,070,621 |

**Supplementary Table 5** List of BACs.

Gene covered by bacterial artificial chromosomes used in 3D immuno-DNA FISH experiments (Covered gene), clone identification (BAC clone) and BACs genomic coordinates (Position GRCh37, Chr bands and Position GRCh38).

## Supplementary methods

### Linear amplification-mediated PCR (LAM-PCR) to map HIV-1 insertion sites in primary cells

The mapping was performed as described previously[1,2] using 1µg of genomic DNA from HIV-1 NL4-3 infected primary human $CD4^+$ T cells.

### Inverse PCR to map HIV-1 insertion sites in primary cells

The mapping of HIV was performed based on the protocol published by Chen *et al.* [3] with modifications. Briefly, 3 µg genomic DNA from HIV-1NL4_3-infected $CD4^+$ T cells treated with 500 nM JQ1 or DMSO before infection were digested by 2 µL 10,000 U/mL AluI (NEB, R0137S) and 2 µL 10,000 U/mL BglII (NEB, R0144S) in NEBuffer 2.1 in 50 µL final volume at 37 ℃ for 3 hours. The reaction was heat-inactivated at 80 ℃ for 20 min. BglII digestion aims to eliminate byproducts, which contain only the sequence of the HIV-1 backbone after AluI digestion. The double-digested products were diluted in 1 mL T4 DNA ligase buffer, then self-ligated by adding 2 µL 30 U/µL T4 DNA ligase (Thermo Fisher Scientific, EL0013) and incubating at 16 ℃ overnight. The ligation reaction was ethanol-precipitated the following day. The pellet was resuspended in 84 µL distilled water. To destroy non-circularized genomic DNA, 4 µL 25 mM ATP and 2 µL 10 U/µL Plasmid-Safe™ ATP-Dependent DNase (Epicentre, E3101K) were added with 10X Reaction Buffer in 100 µL final volume at 37 ℃ for 2 hours. The reaction was heat-inactivated at 70 ℃ for 30 min.

6 µL Plasmid-Safe-digested products were mixed in 50 µL standard Phusion polymerase reaction mix (Thermo Fisher Scientific, F530S) in GC buffer, with 0.1 µM primers GAT1786 (annealing to the Illumina PE1.0 primer) and one indexing primer GAT-int_5LTR (annealing to the 5' end of the LTR) or 0.1 µM primers GAT1799 (annealing to the Illumina PE1.0 primer) and one indexing primer GAT-int_3LTR (annealing to the 3' end of the LTR) for each condition of the sample. The cycling conditions were as follows: 98 ℃ for 1 min; 98 ℃ for 20 sec, 55 ℃ for 1 min, 72 ℃ for 5 min (2 cycles); 98 ℃ for 20 sec, 62 ℃ for 1 min, 72 ℃ for 5 min (27 cycles); 72 ℃ for 5 min. GAT-int_5LTR and GAT-int_3LTR primers add the Illumina PE2.0 primer and a 6-nucleotide index to the amplicons. PCR products ran as a smear on agarose gel . The primers used are described in **Supplementary Table 2**.

### Integration sites and genes

We analyzed eight lists of HIV-1 integration sites. [4–9] were downloaded from retroviral integration database (RID)[10]. We downloaded raw sequences from [11] (PRJNA531196) and one data set was provided by Lusic lab (previously unpublished).

We processed the raw sequences from [11] in the following way: To remove barcodes + LTR-CA sequence on the left side of the reads and linker sequence on the right, we used bbduk allowing for editdistance=2 and minlength=0. Next, reads were converted to fasta files and mapped to hg19 (GRCh37) by pBLAT with parameters -maxIntron=0 -minIdentity=98. We processed the resulting tables in R in the following way: we calculated the percentage of identity by dividing the number of matches by query length for each sequence. All alignments with percentage of identity calculated in this way smaller than 98.0 were not considered. All sequences which mapped to multiple positions in the genome were also not further considered. In this way we obtained in total 1475 uniquely mappable integration sites in activated cells. Some of those integration sites originated from clonally expanded cells, and those were collapsed for our analysis. We obtained 864 unique integration sites from activated cells. Code for processing raw sequences to get integration sites is available here: https://github.com/gui11aume/genome_structure_and_HIV_integration/blob/master/maja/Brady_Integration_Sites.md .

We used only unique integrations from each study. If the location of the integration was not precisely defined (spanning more than one nucleotide), we used the midpoint as the location for that integration. All sites were converted to hg19 (GRCh37) version of the genome using R rtracklayer package[12]. Gene coordinates were downloaded from Ensembl, GRCh37, February 2014. UCSC symbols were used for genes in UCSC (hg19), while others were named after ENSG identifier. Integration sites used in this analysis are available as an R object containing a list of GRanges objects, one list element for each data set used, https://github.com/gui11aume/genome_structure_and_HIV_integration/blob/master/maja/is.Robj .

In those cases where using all genes could have introduced bias to the results (for ChIP-Seq profiles on genes and expression data), only protein coding genes were used for the analysis. This was the case in the following figures: Figure 1A, Figure 2A, 2B, 2C, 2D, Figure 4F, Supplementary Figure 2 (E, G) and Supplementary Figure 5B. List of all genes from Ensembl, GRCh37 assembly from February 2014 was used for the Figures 1C, Supplementary Figure 1 (A, B, C, D) and Supplementary Figure 3.

We counted overlaps between integration sites and all GRCh37 genes disregarding strand and orientation.

Gene coordinates for genes and number of HIV integration lists can be found in **Supplementary Data 1 .**

To control for the JQ1 activity we then additionally sequenced 39k (14k of IS in non-treated ie control infections and 25k in JQ1 pretreated cells).

All data on primary cells (i.e. a total of ~28k) were then used for comparison with insertion sites in Jurkats.

**Redefinition of RIGs**

To each gene, we added a number representing number of lists that found HIV-1 integration inside this gene (**Supplementary Data 1**). We define recurrent integration genes (RIGs) as genes for which we found HIV integration in 2 or more data sets. To account for possible false positives, we assign an assessment of confidence to all RIGs we defined by doing the following: for each of the 8 mapping experiments, we produced 100 mock data sets with the same number of integration sites. Random mock HIV integration sites were chosen to match the distance to the nearest expressed gene, as explained for the ROC analysis. In this way we created 100 matched control lists of 8 “experiments”. For each gene, we counted the number of control “experiments” where given gene was targeted. This value ranges from 0 (if the gene was not targeted) to 8 (if the gene was targeted in all 8 random data sets). For each gene we collected those scores for the 100 randomizations and counted how often they were higher than the observed score. The matrix that contains number of lists each gene is found on in all 100 randomizations can be found here (in RDS format): https://github.com/gui11aume/genome_structure_and_HIV_integration/raw/master/maja/Replicates.RDS.

For each gene we assigned a number which represents the number of randomizations (out of 100) in which this gene scored worse than in real data sets (**Supplementary Data 1**, last column nRealBetter).

To assess the relationship between number of integration sites included in analysis and number of genes discovered to have integrations, we did the following: We made lists of genes found to have integrations in each study. Next, we sorted those lists in decreasing order, by number of integration sites found in a study. We plotted cumulative sum of number of integrations found in studies on x-axis, and number of genes targeted in that study and not in any other studies before that study on y axis. We used linear regression to model this relationship.

**ChIP-Seq data analysis**

We analysed ChIP-Seq data sets obtained from this study (H3K4me3, H3K36me3, H3K27Ac, H4K20me1 and H3K9me2) and publicly available data sets from University of Washington Human Reference Epigenome Mapping Project H3K4me1 (SRA accession number: SRX342315), H3K27me3 (SRX342313), input for CD4 primary T cells (SRX252742). Data sets for BRD4 (SRR2971477), MED1 (SRR2971478) and corresponding Input (GSM1527712) were downloaded from [13].ChIP-Seq reads were mapped to human genome (GRCh37) using Bbmap([14]) with parameters minid=0.98, qtrim=lr, minavgquality=20. Resulting bam files belonging to same experiments were merged and sorted using bamtools. Average binding profiles in reads per million across sets of genes were made using ngsplot [15]. Peaks were called using MACS2 [16], for every data set versus its matching input, with parameters --broad --broad-cutoff 0.1 -p 1e-9 -g 2.7e9 -B. All results were transformed to RPKM for downstream analysis and visualization. The list of ChIP-Seq data used in this study are in **Supplementary Table 3.**

For Jurkat cells, ChIP-Seq reads were mapped to hg19 using BWA-mem. BWA options were as follows: '-k17 -r1.3 -B2 -O4 -T22' for read lengths less or equal to 30 nt, '-k18 -B3 -O5 -T28' for read lengths less or equal to 40 nt and default options for longer reads. ChIP-Seq enriched regions were discretized using Zerone [17] with mapping quality cutoff 20 and enrichment confidence 0.99. We used publicly-available ChIP-Seq profiles for Jurkat cell line. All data sets were obtained from NCBI Gene Expression Omnibus with the following series accessions: ERG and GABPA (GSE49091) [18]; H3K27Ac, CDK7 and PolII (GSE50622, GSE60027) [19]; H3K36me3, H3K79me3, H3K4me1, H3K9me3, H3K27Ac, H3K4me3, PolII, S5P and S2P (GSE65687) [20]; NRSF (GSE53366) [21]; PolII and CDK12 (GSE72023) [22]; ETS1, CBP and RUNX (GSE17954) [23]; H3K27Ac (GSE51522) [24]; PolIII (GSE20309) [25]; H3K27Ac and H3K27me3 (GSE59257) [26]; H3K4me3, H3K27me3, H3K79me2 and PolII (GSE23080); PHF6 (GSE45864); KDM2B (GSE70624) [27]; RUNX1, GATA3, TAL1, LMO1, TCF3 and TCF12 (GSE29181) [28]**;** H3K27Ac, MED1 and MYB (GSE59657) [29]; H3K4me3 (GSE35583) [30]; Lamin (DamID, GSE94971) [31]; RUNX1 (GSE42575) [32]; TAL1 (GSE25000) [33]; H3K4me3 and H3K79me2 (GSE60104) [34]; PolII (GSE25494) [35]; RUNX1, GATA3, H3K27Ac and CTCF (GSE68976) [36]; MYC, BRD4 and CDK7 (GSE83777); RUNX1 and GATA3 (GSE76181) [37]; CTCF (GSE12889) [38]; H2AX (GSE25577) [39]; UTX (GSE72300) [33]; YY1 (GSE99521) [40].

**Super-enhancer calling**

We used super-enhancer data for all activated $CD4^+$ cell types CD4p_CD25-_Il17p_PMAstim_Th17 and CD4p_CD25-_Il17-_PMAstim_Th from dbSuper[41]. To define super-enhancers using our own ChIP-Seq data we followed the same procedures as in dbSuper. Thus, we defined super-enhancers using HOMER software [41,42] findPeaks with default parameters ('-style super') on our H3K27ac peaks (peak finding described above). Briefly, peaks found within a distance of 12.5 kb were stitched together into larger regions. Super-enhancer signal of each region was determined by the total normalized number of reads subtracted by normalized number of reads in the input peaks. Regions are sorted by score and super-enhancers are identified as regions with score higher than that defined by slope greater than 1. We defined a gene to be “proximal” to super-enhancer if it overlaps with one, or if we can find a super-enhancer element 5 kb upstream of transcription start site.

**Hi-C contacts**

Hi-C reads were mapped using BWA-MEM with the following options: ‘-P -k17 -U0 -L0,0 -T25’. Each read end was mapped independently. Genuine Hi-C contacts were validated with the Hi.C pipeline (https://github.com/ezorita/hi.c), using the following discard filters: (i) contact pairs with mapping quality below 10, (ii) self-circularized molecules and (iii) reads with inferred insert size

greater than 2000 bp after digestion and ligation. Hi-C contacts were then binned at 5kb resolution and stored in HDF5 format using Cooler (https://github.com/mirnylab/cooler).

**Hotspots and HIV-dense genes**

HIV genome-wide hotspots were identified dividing the genome in bins of 100 kb and sorting the bins by HIV insertion count. HIV density in genes followed a similar rationale but the bins were designed to match gene bodies, as described by ENSEMBL GTF GRCh37 release 75. HIV density was computed as the number of insertions per kb of gene body.

**AB score**

AB scores were derived from the first eigenvector of the Hi-C correlation matrix (as in Identification of 3D sub-compartments). We chose the reference A and B regions to be the most dissimilar 3D structures, i.e. the genomic bins with 10% top and bottom values of the first eigenvector, respectively. For each row of the correlation matrix we computed $A_{score}$ and $B_{score}$ as the sum of its values in the A and B reference regions, respectively. Finally, the AB score was computed as:

$$AB_{score} = \frac{A_{score} - B_{score}}{A_{score} + B_{score}} \cdot 100$$

Yielding values between 100 for A-like regions and -100 for B-like regions. Ambiguous regions that are equally in contact with the reference A and B regions, or that are not in contact with them at all, will have AB scores close to 0.

**Identification of 3D sub-compartments**

The following pre-processing steps were performed to prepare the matrix before clustering. Observed-over-expected normalization was applied to smooth the diagonal decay[43], followed by ICE row-sum balancing[44]. Outlier contacts, such as enhancer-promoter loops, were smoothed by thresholding the largest values of the matrix to the 90-th percentile. The correlation matrix was subsequently computed and the resulting diagonal was set to 0. The outliers of the correlation matrix were further smoothed by applying a linear scaling, *i.e.* values below the 5-th and above the 95-th percentiles were set to -1 and +1, respectively, and intermediate values were scaled proportionally. Spatial clusters were identified on the correlation matrix running k-means (10 restarts) with the first $k$=15 weighted eigenvectors, *i.e.* the 15 leading eigenvectors, each weighted by its respective eigenvalue. This process was repeated independently for each chromosome, delineating 15 spatial clusters per chromosome.

The choice of $k$ relied on previous Hi-C analyses, which reported six distinguishable clusters[45]. Comparatively, the present clustering is performed at higher resolution (5kb). For this reason, the value of $k$ was chosen large enough to allocate new potentially unresolved conformations. The enrichment of ChIP-Seq data on the intrachromosomal clusters showed, in most cases, five clear patterns (supplementary figure with ChIP heatmap, intrachromosomal). Therefore, $k$ was reduced to 5 in the subsequent interchromosomal clustering. The same analysis was repeated with increasing values of $k$, resulting in split subclusters which shared similar features.

To identify interchromosomal clusters, normalized interchromosomal scores were computed between each pair of chromosomal sub-compartments. Normalized scores were defined as the

total number of Hi-C reads between them, divided by the product of their sizes. The final sub-compartments were identified by k-means clustering on the normalized score matrix with the *k*=5 leading weighted eigenvectors. Compartment names A1, A2, AB, B1 and B2 were assigned based on their distribution on the AB score scale (Figure 5C). A1 and A2 had strong and moderate enrichment in active transcription marks, respectively. AB, B1 and B2 showed very low levels of active marks and moderate to strong enrichment in H3K27me3, H3K9me3 and Lamin, respectively (Supplementary figure with interchromosomal heatmap). The complete list of sub-compartments is available in the supplementary material.

The A1, A2 and B2 sub-compartments are robust to the implementation details of the definition: when we used different normalizations or different weights for the eigenvectors, they always appeared with similar coverage and chromatin features. On the other hand, AB and B1 varied in coverage and composition, suggesting that they are fuzzier than the other sub-compartments.

**Pairwise contact score**

Throughout the study, pairwise contact scores were used to quantify the amount of 3D interactions between multiple loci on different chromosomes. Intrachromosomal contacts were not considered in this computation in order to avoid the intrinsic bias produced by short-range 1D interactions (loci that are closer in 1D tend to interact more in 3D). Pairwise scores were computed as the sum of Hi-C contacts within the interchromosomal region covered by a pair of loci, divided by the product of their lengths (in kbp). The unit of this metric (number Hi-C reads per square kilobase pair) allows for fair comparison of different loci even if their spans are different.

**ROC analysis**

In order to assess if there is an enrichment of various chromatin features on sites of HIV-1 integration, we adapted the ROC curve areas method from [46] and [47]). In short, the strategy was to use “nested case controls” - a collection of integration sites sampled from the genome which would act as control sites and can be compared to true integration sites. For every chromatin feature and experiment we analysed, we compared density of values for this feature measured on integration sites, versus density of values of this feature measured on control sites. For every cut-point of value of measured feature, we measured the percentage of integration sites with value of this feature higher than the cutpoint (true positive rate) and percentage of control sites with value of this feature higher then the cut-point (false positive rate). Thus, we constructed the ROC curve by calculating the true and false positive rates for all possible cut-point values for analysed epigenomic feature. The area under the ROC curve was then calculated. For details, see supplementary Text S1 [47]). The control sites were generated to account for bias of integration towards genes - they were sampled to match true sites in distance to nearest gene. For each true integration site, we generated 10 matched control sites and compared various chromatin features of the matched sites with the chromatin features of the true site. For each true integration site and chromatin feature, we counted a fraction of control sites having a lower feature value (e.g. true site has higher H3K27ac value than n percent of control matched sites). We averaged the results over all experiments. This is explained in more detail in the following sections:

For random matched control sites**,**e generated 10 control sites for every integration site in the following way: First we generated random 100 million numbers from 1 to largest chromosome length with seed set to 23779. Next, we generated 100 million chromosome names, where names were chosen at random but with weights corresponding to the number of occurrences of each chromosome in our data set. This way we generated 100 million random possible positions for controls. Next, we excluded all the positions from this random set that were found in the blacklisted area of human GRCh37 genome [48]. Sites are available on request. We calculated the distance to the nearest gene for each possible integration site and divided them into subsets of 1000 base pair bins based on those distances. For each true integration site, we extracted a subset of all random possible positions that are located in the same bin of distance to their nearest gene as the integration site is to its nearest bin. Then, from those equidistant subset of random integration sites we randomly picked 10 to represent random matched controls for each real integration site.
To Add genomic feature value to integration sites and compare integration sites to its random matched controls we firstcut the genome into tiles of length 1000 base pairs and excluded blacklisted areas. We calculated $75^{th}$ quantile of RPKM values of each genomic feature (for each chromatin mark and transcription factor separately) over each tile. For super-enhancers we used 1 if super-enhancer exists in a tile, and a 0 if it does not instead of RPKM values. We assigned the value of the bin in which integration site is located to each integration site (and each random matched control). Next, we compared the value for each integration site only with its matched controls to determine the proportions of controls whose values equaled or exceeded that of the integration site. We scored each integration site in the following way: if this value was higher on true integration site than on matched control, we counted it as 1. If the value for the matched control was equal to true value on integration site, we counted it as ½, and if the value was lower, it was counted as 0. Final score for each integration site was calculated as average of those 10 values. Finally, we calculated empirical ROC area under the curve as average of all values for integrations in a data set. At last, we repeated this analysis on various bin sizes; 1Kb, 2Kb, 5Kb, 10Kb, 20Kb, 25Kb, 50Kb, 100Kb and on all genomic features and data sets and created a heatmap for every bin size (data not shown). We implemented p value calculation from [47] in R. Briefly, all comparisons utilize the Wald-test statistic and are referred to a Chi Square distribution to obtain p-values.

**RNA-Seq data analysis**
We mapped reads from RNA-Seq experiments to human genome (GRCh37 assembly, GENCODEV19) using BBMap with parameters maxindel=200000 xstag=unstranded ambiguous=random xmtag=t scoretag=t pairedonly=t minavgquality=20 maqb=51 minid=0.91. To calculate mean expression over replicates, we used rlog transformation from DESeq2 package [49] to normalize the counts over all replicates, and calculated mean over all replicates. We used a regularized logarithm transformation (rlog) to make our data more homoskedastic. The rlog transformation produces log2 scale transformed data which has been normalized with respect to library size. We used rlog instead of log2 transformation because it is more robust in the case when the size factors vary widely; this transformation is reducing the variance to avoid that the result becomes dominated by highly expressed, highly variable genes (original count scale data), or low expressed genes (if logarithm-transformed data are used). We used a widely utilised rlog approach of DESeq2, which enables transformation similar to a log2 for genes with high counts, and resolves the problem for genes with low counts by compressing together the values for different samples. Otherwise, a standard logarithm transformation would spread apart the data, ie random noise could overtake the real biological signal. Differential expression for JQ1 treated cells VS control cells was done with the same package, following the Bioconductor RNA-seq workflow [50]. Genes were divided to expression groups as follows: All genes with rlog of expression (averaged over replicates in activated non treated CD4+ T cells) lower or equal to

0 are considered to be not expressed. We divided the rest into 3 groups: low 10% being all genes with expression in the bottom 10% from genes that were considered as expressed. Analogously, we grouped all most highly expressed genes (top 10 quantile) into top 10% group, while the rest of the genes was grouped in mid group.

For Jurkat cells, gene expression levels were derived from mRNA-seq experiments in Jurkat . Sequencing reads were mapped to protein-coding Ensembl cDNA assembly GRCh37 release 75 using kallisto [51] with options '–single' (single-end mode), '–bias' (sequence bias correction, '–s300' (fragment length 300 nucleotides) and '–l100' (s.d. 100 nucleotides). The counts of the different isoforms were summed to make a total count per gene copy in transcripts per million reads (tpm).
The list of RNA-Seq data used in this study are in **Supplementary Table 4.**

**Logistic regression**
We used logistic regression to model the HIV-1 insertion landscape in Jurkat cells, based on the following four predictors: gene expression, distance of the gene to the closest super enhancer, sub-compartment of the gene and gene size. Gene size must be added to the model because it is a confounding factor. In model I, we predicted whether the gene was a typical HIV target, defined as belonging to the top 33% genes with highest HIV insertion rate (number of insertions divided by gene size). In a model II, we predicted whether the gene contained a hotspot, defined as a 10 kb bin with more than five HIV insertions (this corresponds to the top 2.5% genes with highest number of insertions in a single 10 kb bin).

The models were trained with the standard parameters of the glm function in R. We used 5-fold cross validation to test combinations of transforms (hyperbolic arcsine and logarithmic functions) and / or discretization in quantiles. The best cross-validation scores were obtained after discretizing the predictors in combinations of tertiles and quartiles, so models I and II were trained on discretized variables (the same discretization was used for both models). We then removed one of the four variables, retrained the model in the same conditions and measured the probability that a gene is classified as HIV-1 target (typical targets in model I, hotspots in model II) given that it is indeed an HIV-1 target. We chose this score instead of the classification accuracy because of the low amount of HIV-1 targets in model II – where the classification accuracy ranges from 97.5% to 100%. We used the loss of this score compared to the full model as an indicator of the intrinsic value of the predictor. The results did not change substantially when typical HIV targets were defined as top 50% or top 20%, neither when hotpots were defined as top 5% or top 1% (not shown).

**Supplementary References**


1. Bartholomae, C. C., Glimm, H., von Kalle, C. & Schmidt, M. Insertion Site Pattern: Global Approach by Linear Amplification-Mediated PCR and Mass Sequencing. in *Methods in Molecular Biology* 255–265 (2012).
2. Schmidt, M. *et al.* High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). *Nat. Methods* **4**, 1051–1057 (2007).
3. Chen, H. C., Zorita, E. & Filion, G. J. Using Barcoded HIV Ensembles (B-HIVE) for single provirus transcriptomics. *Curr. Protoc. Mol. Biol.* (2018).
4. Han, Y. *et al.* Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. *J. Virol.* **78**, 6122–6133 (2004).
5. Ikeda, T., Shibata, J., Yoshimura, K., Koito, A. & Matsushita, S. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. *J. Infect. Dis.* **195**, 716–725 (2007).
6. Maldarelli, F. *et al.* HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179–183 (2014).
7. Wagner, T. A. *et al.* HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570–573 (2014).
8. Cohn, L. B. *et al.* HIV-1 integration landscape during latent and active infection. *Cell* **160**, 420–432 (2015).
9. Kok, Y. L. *et al.* Monocyte-derived macrophages exhibit distinct and more restricted HIV-1 integration site repertoire than CD4(+) T cells. *Sci. Rep.* **6**, 24157 (2016).
10. Shao, W. *et al.* Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes. *Retrovirology* **13**, 47 (2016).
11. Brady, T. *et al.* HIV integration site distributions in resting and activated CD4+ T cells infected in culture. *AIDS* **23**, 1461–1471 (2009).
12. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
13. Hertweck, A. *et al.* T-bet Activates Th1 Genes through Mediator and the Super Elongation Complex. *Cell Rep.* **15**, 2756–2770 (2016).
14. *BBMap: A Fast, Accurate, Splice-Aware Aligner*. (2014).
15. Shen, L., Shao, N., Liu, X. & Nestler, E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15**, 284 (2014).
16. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
17. Cuscó, P. & Filion, G. J. Zerone: a ChIP-seq discretizer for multiple replicates with built-in quality control. *Bioinformatics* **32**, 2896–2902 (2016).
18. Sharma, N. L. *et al.* The ETS family member GABPα modulates androgen receptor signalling and mediates an aggressive phenotype in prostate cancer. *Nucleic Acids Res.* **42**, 6256–6269 (2014).
19. Kwiatkowski, N. *et al.* Targeting transcription regulation in cancer with a covalent CDK7

inhibitor. *Nature* **511**, 616–620 (2014).

20. 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).
21. Gasper, W. C. *et al.* Fully automated high-throughput chromatin immunoprecipitation for ChIP-seq: identifying ChIP-quality p300 monoclonal antibodies. *Sci. Rep.* **4**, 5152 (2014).
22. Zhang, T. *et al.* Covalent targeting of remote cysteine residues to develop CDK12 and CDK13 inhibitors. *Nat. Chem. Biol.* **12**, 876–884 (2016).
23. Hollenhorst, P. C. *et al.* DNA specificity determinants associate with distinct transcription factor functions. *PLoS Genet.* **5**, e1000778 (2009).
24. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
25. Oler, A. J. *et al.* Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat. Struct. Mol. Biol.* **17**, 620–628 (2010).
26. Navarro, J.-M. *et al.* Site- and allele-specific polycomb dysregulation in T-cell leukaemia. *Nat. Commun.* **6**, 6094 (2015).
27. Andricovich, J., Kai, Y., Peng, W., Foudi, A. & Tzatsos, A. Histone demethylase KDM2B regulates lineage commitment in normal and malignant hematopoiesis. *J. Clin. Invest.* **126**, 905–920 (2016).
28. Sanda, T. *et al.* Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell* **22**, 209–221 (2012).
29. 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).
30. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
31. Robson, M. I. *et al.* Constrained release of lamina-associated enhancers and genes from the nuclear envelope during T-cell activation facilitates their association in chromosome compartments. *Genome Res.* **27**, 1126–1138 (2017).
32. Kim, D. Y. *et al.* CBFβ stabilizes HIV Vif to counteract APOBEC3 at the expense of RUNX1 target gene expression. *Mol. Cell* **49**, 632–644 (2013).
33. Benyoucef, A. *et al.* UTX inhibition as selective epigenetic therapy against TAL1-driven T-cell acute lymphoblastic leukemia. *Genes Dev.* **30**, 508–521 (2016).
34. Orlando, D. A. *et al.* Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep.* **9**, 1163–1170 (2014).
35. Ip, J. Y. *et al.* Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Res.* **21**, 390–401 (2011).
36. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
37. Saint-André, V. *et al.* Models of human core transcriptional regulatory circuitries. *Genome Res.* **26**, 385–396 (2016).
38. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin

barrier regions reveals demarcation of active and repressive domains. *Genome Res.* **19**, 24–32 (2009).

39. Seo, J. *et al.* Genome-wide profiles of H2AX and γ-H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells. *Nucleic Acids Res.* **40**, 5965–5974 (2012).
40. Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573–1588.e28 (2017).
41. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* **44**, D164–71 (2016).
42. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
43. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
44. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
45. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
46. Brady, T. *et al.* Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev.* **23**, 633–642 (2009).
47. Berry, C., Hannenhalli, S., Leipzig, J. & Bushman, F. D. Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput. Biol.* **2**, e157 (2006).
48. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
50. Love, M. I., Anders, S., Kim, V. & Huber, W. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Res.* **4**, 1070 (2015).
51. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).