

Decoding Human Genome Regulatory Features That Influence HIV-1 Proviral Expression and Fate Through an Integrated Genomics Approach

Bioinformatics and Biology Insights
Volume 16: 1–17
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322211072333



Holly Ruess¹ , Jeon Lee¹ , Carlos Guzman^{2,3},
Venkat S Malladi¹ and Iván D'Orso²

¹Lyda Hill Department of Bioinformatics, The University of Texas Southwestern Medical Center, Dallas, TX, USA. ²Department of Microbiology, The University of Texas Southwestern Medical Center, Dallas, TX, USA. ³Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA, USA.

ABSTRACT: Fundamental principles of HIV-1 integration into the human genome have been revealed in the past 2 decades. However, the impact of the integration site on proviral transcription and expression remains poorly understood. Solving this problem requires the analysis of multiple genomic datasets for thousands of proviral integration sites. Here, we generated and combined large-scale datasets, including epigenetics, transcriptome, and 3-dimensional genome architecture to interrogate the chromatin states, transcription activity, and nuclear sub-compartments around HIV-1 integrations in Jurkat CD4⁺ T cells to decipher human genome regulatory features shaping the transcription of proviral classes based on their position and orientation in the genome. Through a Hidden Markov Model and ranked informative values prior to a machine learning logistic regression model, we defined nuclear sub-compartments and chromatin states contributing to genomic architecture, transcriptional activity, and nucleosome density of regions neighboring the integration site, as additive features influencing HIV-1 expression. Our integrated genomics approach also allows for a robust experimental design, in which HIV-1 can be genetically introduced into precise genomic locations with known regulatory features to assess the relationship of integration positions to viral transcription and fate.

KEYWORDS: Data integration, human genome, genomics, enhancers, sub-compartments, ChromHMM, HIV-1, integration, transcription

RECEIVED: August 10, 2021. **ACCEPTED:** December 9, 2021.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: I.D. was in part funded by the US National Institutes of Health under award number R01AI114362. H.R., J.L., and V.S.M. were supported by the Cancer Prevention Research Institute (CPRIT) under award number RP150596. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Iván D'Orso, Department of Microbiology, The University of Texas Southwestern Medical Center, 6000 Harry Hines Blvd., NL3.110B, Dallas, TX 75390-9048, USA. Email: Ivan.Dorso@utsouthwestern.edu

Introduction

Great efforts have been made, over the last 2 decades, to understand how HIV-1 integrates into the human genome and how proviral transcription functions at individual integration positions.^{1–6} HIV-1 preferentially integrates into chromatin-accessible sites within or near transcriptionally active regions in CD4⁺ T cells of patient samples and in ex vivo infection studies,^{7–10} and select integration sites can promote T cell proliferation ex vivo.¹¹ However, integrated proviruses are detected on every human chromosome, in various chromatin landscapes (euchromatic and heterochromatic), and at different locations (intergenic or intragenic) and orientations (sense, divergent, or convergent) relative to human genes and regulatory elements.¹² Because the integration neighborhood is highly variable in terms of sequence, predicted chromatin structure, and transcriptional activity,^{13–16} it is possible that integration sites contain information regulating the amplitude of proviral transcription and hence shaping its fate (active vs latent infection). Indeed, previous works on tens of Jurkat CD4⁺ T cell clones containing HIV-1 placed in distinct positions suggest that the integration site controls basal and immune stimulation-dependent transcription,^{13,15,17,18} implying that HIV-1 operates in an integration

site-dependent manner influenced by the human genome context and/or architecture. Despite previous research, it is still unknown which regulatory features have the most influence on HIV-1 proviral transcription and whether numerous factors contribute in an additive or synergistic manner. Given the huge number of possible regulatory features, including nuclear sub-compartments, enhancers, expression of genic and non-genic domains, genome accessibility, and functional chromatin states (Figure 1A), HIV-1 transcription and expression could be regulated at multiple levels.

Here, we devise an integrative genomics strategy for determining the contribution of individual or combinations of regulatory features to HIV-1 proviral expression (Figure 1A and B). First, we defined the nuclear sub-compartments, transcription activity landscape, chromatin accessibility, and chromatin states around HIV-1 integration sites by combining HIV-1 integration and expression datasets with new and open-source, large-scale datasets including 3-dimensional (3D) genome architecture, transcriptome, genome accessibility, and epigenetics (chromatin marks) (Table 1). Second, we predicted upstream chromatin accessibility, transcription activity, and categorical nuclear sub-compartments as optimal features shaping HIV-1 expression outcomes through a machine learning (ML) logistic



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

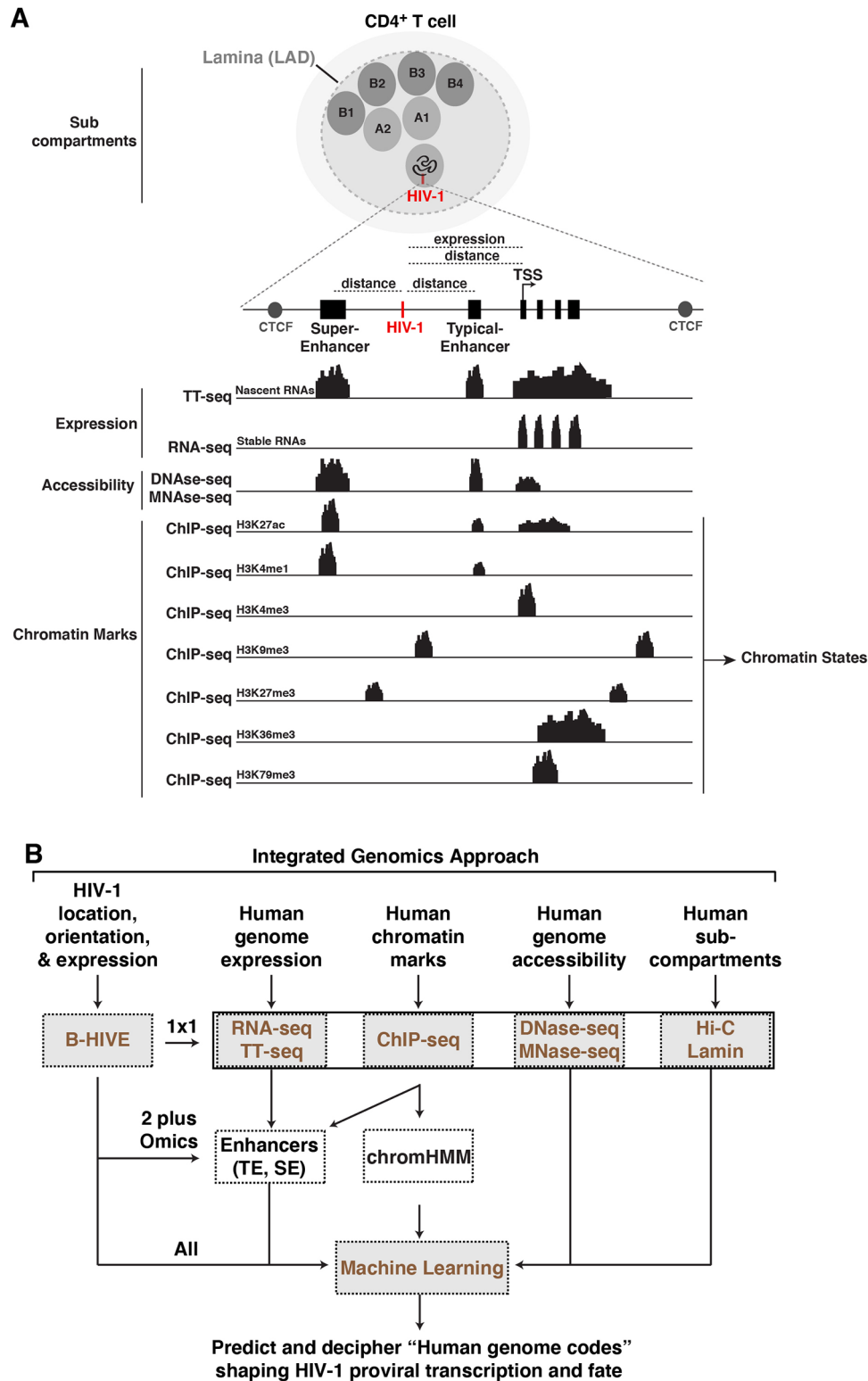


Figure 1. Flowchart of the integrated genomics approach to elucidate human genome codes contributing to HIV-1 proviral transcription and fate. (A) Scheme depicting the positions of HIV-1 proviruses relative to the multiple regulatory features evaluated in this study including nuclear architecture (with its A and B sub-compartments and lamina-associated domains), typical enhancers (TEs) and super enhancers (SEs), host expression, nucleosome density and accessibility, and chromatin marks collectively used to demarcate functional chromatin states. (B) Barcoded *HIV-1* Ensembles (B-HIVE) expression data are compared with each individual dataset 1-by-1 (1×1). Then, B-HIVE expression data are compared with datasets that combine multiple datasets ("2 plus Omics") (eg, TEs and SEs combine TT-seq and ChIP-seq). B-HIVE is then compared with "all" datasets with a machine learning model. From there, future HIV-1 patients' datasets can be integrated with the long-term objective to predict human genome codes leading to clinical decision-making.

Table 1. Genomics datasets used in this study.

SEQUENCING TYPE	TARGET	SRA/ENCODE	INPUT SRA/ ENCODE	REPLICATE	REFERENCE
DNase-seq	DNase I hypersensitive	ENCFF001DPG	N/A	1	Encode Consortium
	DNase I hypersensitive	ENCFF001DPF	N/A	2	Encode Consortium
MNase-seq	Micrococcal nuclease	GSM4295147	N/A	1	This study
ChIP-seq	H3K27me3	GSM569085	GSM569086	1	R. Young Lab, unpublished
	H3K4me3	ENCLB868HNG	GSM945268	1	Encode Consortium
	H3K4me3	ENCLB676HDJ	GSM945268	2	Encode Consortium
	H3K4me3	GSM1603213	GSM1603229	3	Reeder et al ¹⁹
	H3K27ac	GSM1697882	GSM1697880	1	Hnisz et al ²⁰
	H3K27ac	GSM1519638	GSM1519637	2	Mansour et al ²¹
	H3K27ac	GSM1519642	GSM1519640	3	Mansour et al ²¹
	H3K27ac	GSM1603211	GSM1603229	4	Reeder et al ¹⁹
	H3K4me1	GSM1603225	GSM1603229	1	Reeder et al ¹⁹
	H3K36me3	GSM1603209	GSM1603229	1	Reeder et al ¹⁹
	H3K79me3	GSM1603215	GSM1603229	1	Reeder et al ¹⁹
	H3K9me3	GSM1603227	GSM1603229	1	Reeder et al ¹⁹
RNA-seq	Total RNA	GSM4290736	N/A	1	I. D'Orso Lab, unpublished
	Total RNA	GSM4290737	N/A	2	I. D'Orso Lab, unpublished
	Total RNA	GSM4290738	N/A	3	I. D'Orso Lab, unpublished
TT-seq	Transient transcriptome	GSM2260187	N/A	1	Michel et al ²²
	Transient transcriptome	GSM2260188	N/A	2	Michel et al ²²
Hi-C	3D genome architecture	GSM3489136	N/A	1	Lucic et al ⁶
	3D genome architecture	GSM3489137	N/A	2	Lucic et al ⁶
DamID-seq	Lamina-associated protein lamin B1	GSM2492607	GSM2492606	1	Robson et al ²³
	Lamina-associated protein lamin B1	GSM2492609	GSM2492608	2	Robson et al ²³
B-HIVE (Barcoded HIV-1 Ensembles)	HIV-1 Expression DNA	GSM2182756, GSM2182757	N/A	1	Chen et al ²⁴
	HIV-1 Expression DNA	GSM2182758, GSM2182759	N/A	2	Chen et al ²⁴
	HIV-1 Expression RNA	GSM2182760, GSM2182761	N/A	1	Chen et al ²⁴
	HIV-1 Expression RNA	GSM2182762, GSM2182763	N/A	2	Chen et al ²⁴

regression model of a 2 kb region around HIV-1 integration sites to interrogate neighboring effects.

Materials and Methods

Cell lines

Jurkat, Clone E6-1 (ATCC TIB-152) was obtained from the American Type Culture Collection (ATCC, Manassas, VA) and Jurkat J-Lat 10.6 clone was obtained from the lab of Dr Eric Verdin. Cells were cultured in Roswell Park Memorial Institute

(RPMI) 1640 Medium (HyClone, Logan, UT, SH30027.FS) supplemented with 8% fetal bovine serum (FBS) (Millipore Sigma, Burlington, MA, H9268) and 1% Penicillin/Streptomycin (MP Biomedicals, Irvine, CA, 091670049).

MNase-seq library preparation

Jurkat CD4⁺ T cells were cultured in RPMI 1640 media supplemented with 10% FBS and 1× Penicillin/Streptomycin at 37°C with 5% CO₂ at optimal density of 0.5 to 1 × 10⁶ cells per

mL. Cells were passaged every 2 days at a 1/3 dilution. Cell suspensions were transferred to 50 mL conical tubes and pelleted at $420 \times g$ for 10 min. Cells were then resuspended in phosphate-buffered saline (PBS) at a density of 1×10^6 cells/mL for crosslinking with 0.5% methanol-free formaldehyde (ThermoFisher, Waltham, MA, 28908) at room temperature with rotation for 10 minutes. The reaction was then quenched with 150 mM glycine (PBS buffer pH 7.5) for 10 minutes at room temperature. Cells were then pelleted by centrifuging at $420g$ for 10 minutes at 4°C and then washed twice with 20 mL cold PBS each time. Nuclei were collected by lysing the cells in Farnham's lysis buffer (5 mM PIPES pH 8.0, 85 mM KCl, and 0.5% NP-40 freshly supplemented with 1 mM PMSF [phenylmethylsulfonyl fluoride] and EDTA-free Protease inhibitor cocktail), washed once with cold MNase buffer (20 mM Tris-HCl pH 7.5, 15 mM NaCl, 60 mM KCl, and 2 mM CaCl_2) and then resuspended in MNase buffer at a concentration of 10×10^6 cells/mL before digestion. Micrococcal nuclease (New England BioLabs, Ipswich, MA, M0247S) digestion was performed using 1100 U enzyme/ 10^6 cells at 37°C for 10 minutes to achieve roughly 80% mono-nucleosome–20% di-nucleosome populations. The reaction was interrupted with Stop buffer (20 mM EDTA pH 8.0, 20 mM Ethylene glycol tetraacetic acid (EGTA) pH 8.0, and 0.4% Sodium lauryl sulfate (SDS)) and then centrifuged at $21\,000g$ for 5 minutes at 4°C . The supernatants were saved, and the small white pellet discarded. Samples (100 μL) were mixed with 1 volume of $2 \times$ Proteinase K buffer (4 mM EDTA pH 8.0, 40 mM Tris-HCl pH 6.8, 1 M NaCl, and 1 mg/mL Proteinase K) for reverse crosslinking at 65°C for 16 hours. After reverse crosslinking, the samples were first extracted with 1 volume of phenol-chloroform-isoamyl alcohol (25:24:1 ratio) with centrifugation at $21\,000g$ for 5 minutes at 4°C and later with 1 volume of chloroform-isoamyl alcohol (24:1 ratio). The aqueous phase was transferred to a 1.5 mL epitube and precipitated with 2.5 volumes of 100% cold ethanol and 1/10 volume of 3M NaOAc with centrifugation at $21\,000g$ for 15 minutes at 4°C . Samples were finally washed with 75% ethanol, air dried for ~5 minutes, and resuspended in 20 μL of water. The DNA concentration was measured by Qubit/Nanodrop. About 3 μg of DNA were loaded onto a 1.5% DNA agarose gel to verify the expected nucleosomal size distribution (80% mono- and 20% di-nucleosome). The mono-nucleosome band (~150bp size) was excised, and gel cleaned up using DNA Clean & Concentrator kit (Zymo Research, Irvine, CA, D4013) following the manufacturer's instructions. The DNA was eluted with 20 μL water (25 ng/ μL final concentration). Replicate DNA samples were analyzed on high-sensitivity DNA tape on Agilent 2200 TapeStation and used for library preparation. Library was prepared with ~375 ng of mono-nucleosomal DNA using the KAPA Hyper Prep Kit (KAPA Biosystems, Wilmington, MA, KK8502) according to the manufacturer's instructions. For the PCR amplification step, we inputted ~7.5 ng and performed 9

cycles of amplification, obtaining 700 ng (~35 ng/ μL). The quality control of the MNase-seq library was done on Agilent 2200 TapeStation. A single peak with average size of 306 bp (including ligated adapters) was observed. The MNase-seq DNA library was diluted to 4.2 nM for sequencing on an Illumina NextSeq 500 instrument as a 2×75 bp library. Illumina bcl2fastq (v 2.19.0) software was used for basecalling.

Barcodes and HIV-1 integration site mapping on the human genome and HIV-1 barcode clustering and quantification

We re-analyzed the Barcoded *HIV-1* Ensembles (B-HIVE) dataset²⁴ in the human genome (GRCh38) to prevent problems that can arise with lifting over data from previous genome versions. The B-HIVE data were processed with the B-HIVE for single provirus transcriptomics docker container (https://github.com/gui11aume/BHIVE_for_single_provirus_transcriptomics) with a change to the `expr.nf` file (see GitHub scripts for the updated script). For HIV-1 integration and expression analysis from B-HIVE dataset, we first identified barcodes in the HIV-1 proviruses (DNA barcodes), then mapped barcodes to integration sites, and quantified their expression. Quantitative Reverse Transcription PCR (RT-qPCR) normalized to the copy of DNA barcodes ($\log_{10}[\text{RNA}_{\text{mean}}/\text{DNA}_{\text{mean}}]$). The B-HIVE expression dataset was subdivided into 6 different groups: (1) Intergenic—Same, (2) Intergenic—Convergent, (3) Intergenic—Divergent, (4) Intragenic—Same, (5) Intragenic—Convergent, and (6) Intragenic—Overlapping (which consists of 3 subgroups; Figure 2A) depending on the relationship to the nearest gene from GENCODE version 25. For each group, a Circos plot version 0.696²⁵ was created to show the relationship of HIV-1 expression to genome location (Figure 2C to H). HIV-1 expression versus distance to nearest transcription start site (TSS) was plotted in R version R/3.3.2-gccmkl²⁶ using ggplot2.²⁷

ChIP-seq data analysis

The Nextflow²⁸ BICF ChIP-seq Analysis Workflow version 1.0.0²⁹ processed all ChIP-seq files, merging separate experiments as technical replicates. Briefly, reads were trimmed with trimgalore version 0.4.1³⁰ (parameters: `-q 25 --illumina --gzip --length 35`), aligned with bwa aln (`-q 5 -l 32 -k 2`) and then bwa samse (standard parameters) version 0.7.12,³¹ sorted and indexed with SAMtools version 1.3³² (`-F 1804 -q 30`), and duplicates removed with Sambamba version 0.6.6³³ (standard options). Bam files were converted to tagAlign with bedtools version 2.26.0 bamtoBed,³⁴ after which, samples were checked for quality control using deeptools version 2.5.0.1 multiBamSummary, plotCorrelation, plotCoverage, and plotFingerprint (all standard protocols),³⁵ and cross-correlation analysis with phantompeakqualtools version 1.2.^{36,37} Peaks were called with

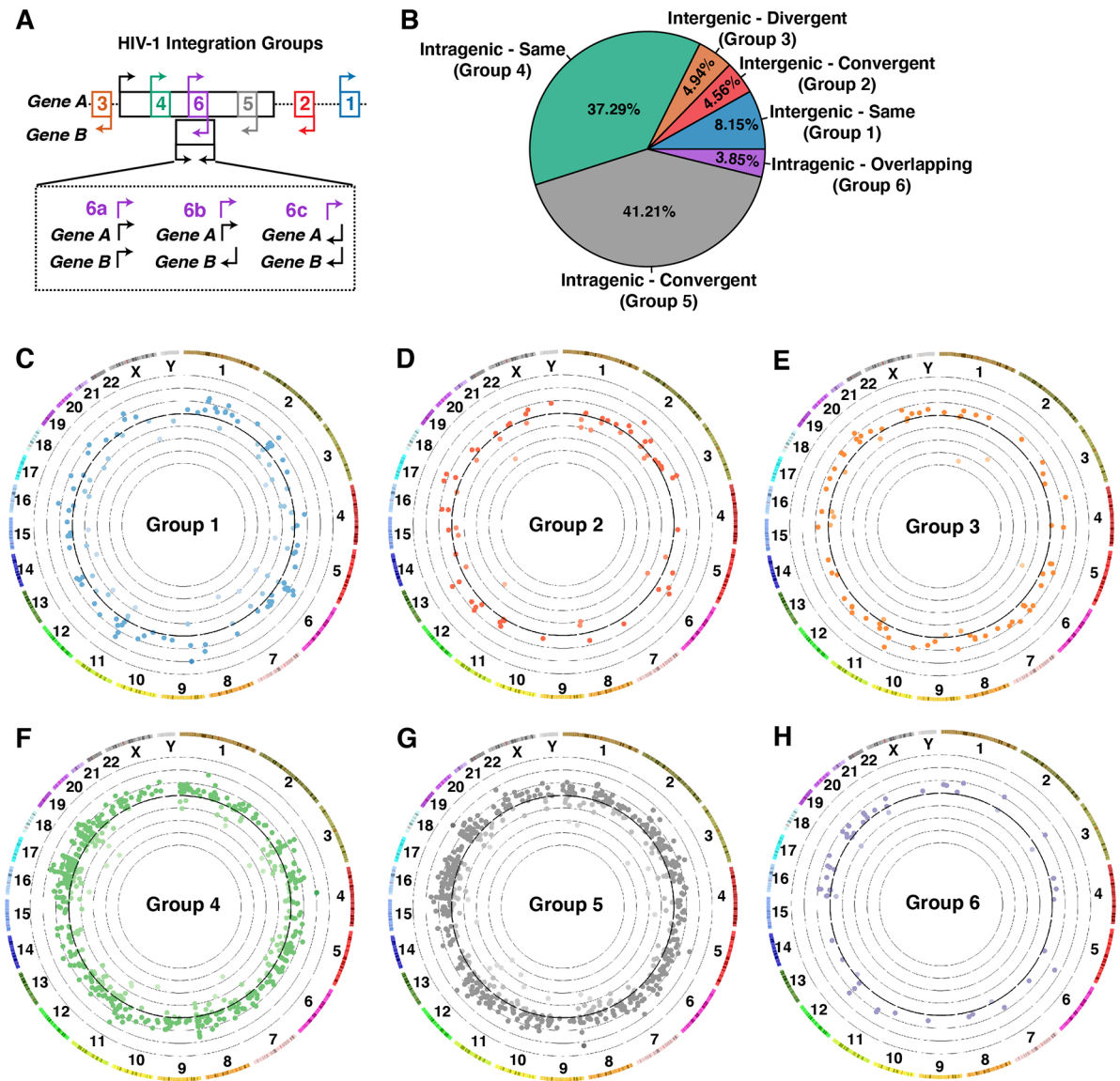


Figure 2. Defining the expression of HIV-1 proviral integration groups based on their position and orientation relative to human genes: (A) Diagram of 6 HIV-1 integration groups relative to nearest gene(s). Group 1: Intergenic—Same (blue), Group 2: Intergenic—Convergent (red), Group 3: Intergenic—Divergent (orange), Group 4: Intragenic—Same (green), Group 5: Intragenic—Convergent (gray), and Group 6: Intergenic—Overlapping (purple). Group 6 (overlapping genes at the same position) comprises 3 subgroups (6a, 6b, 6c). (B) Pie chart of the percentage of Barcoded HIV-1 Ensembles (B-HIVE) insertions (n = 1558) into each of the 6 HIV-1 integration groups. (C to H) Circos plot of each of the 6 HIV-1 integration groups as described in panel (A). Each circle represents B-HIVE chromosomal distribution and expression, with the inner most line, a \log_{10} of HIV-1 expression = -4, and the outer most line, a \log_{10} of HIV-1 expression = 3.

MACS2 version 2.1.0-20151222,³⁸ using the predominant fragment length from the cross-correlation analysis as --ext-size (other parameters: -p 1e-2 --nomodel --shift 0 --keep-dup all -B --SPMR). Consensus peaks were called (bedtools version 2.26.0)³⁴, and annotated (library ChIPseeker in R^{26,39}; if at least 2 replicates or pseudo-replicates contained a peak).

RNA-seq data analysis

FASTQ files were processed with the BICF RNA-seq Analysis workflow version 0.5.5. Briefly, reads with Phred quality scores less than 20 and less than 35 bp after trimming

were removed from further analysis using trimgalore version 0.4.1.³⁰ Quality-filtered reads were then aligned to the human reference genome (GRCh38) using the HISAT version 2.0.1⁴⁰ aligner using default settings and marked duplicates using Sambamba version 0.6.6.³³ Aligned reads were quantified to coding sequences of known transcripts using “feature-count” version 1.4.6⁴¹ per gene ID against GENCODE version 25. HIV-1 expression versus \log_{10} Fragments Per Kilobase Million (FPKM) of the nearest gene (eg, the gene in which HIV-1 is integrated into if intragenic or the nearest gene if intergenic) was plotted in R version R/3.3.2-gccmkl²⁶ using ggplot2.²⁷

TT-seq data analysis

FASTQ files were processed by a modified version of the BICF RNA-seq Analysis workflow version 0.5.5. Briefly, reads with Phred quality scores less than 20 and less than 35bp after trimming were removed from further analysis using trimgalore version 0.4.1.³⁰ Quality-filtered reads were then aligned to the human reference genome (GRCh38) using the HISAT version 2.0.1⁴⁰ aligner using default settings and marked duplicates using Sambamba version 0.6.6.³³ Aligned reads were quantified to the entire annotated transcript region using “featurecount” version 1.4.6⁴¹ per gene ID against GENCODE version 25.

MNase-seq data analysis

FASTQ files were processed with a modified Nextflow,²⁸ BICF ChIP-seq Analysis Workflow version 1.0.0.²⁹ Briefly, we used trimgalore version 0.4.1³⁰ on the raw reads to remove reads shorter than 35bp and with Phred quality scores less than 20bp and then aligned trimmed reads to the human reference genome (GRCh38) using default parameters in BWA samse version 0.7.12.³¹ The aligned reads were subsequently filtered for quality and uniquely mappable reads were retained for further analysis using SAMtools version 1.3³¹ and Sambamba version 0.6.6,³³ and bedtools version 2.26.0³⁴ bamtobed converted the bed file to tagAlign. Peaks were called with iNPS version 1.2.2⁴² and filtered for a $-\log_{10}$ (P value_of_peak) of less than .05.

DNase-seq data analysis

FASTQ files were processed with a modified Nextflow,²⁸ BICF ChIP-seq Analysis Workflow version 1.0.0.²⁹ Briefly, we used trimgalore version 0.4.1³⁰ on the raw reads to remove reads shorter than 35bp and with Phred quality scores less than 20bp and then aligned trimmed reads to the human reference genome (GRCh38) using default parameters in BWA samse version 0.7.12.³¹ The aligned reads were subsequently filtered for quality and uniquely mappable reads were retained for further analysis using SAMtools version 1.3³² and Sambamba version 0.6.6.³³ Relaxed peaks were called using MACS2 version 2.1.0-20151222³⁸ with the following parameters: `-p 1e-2 --nomodel --shift -100 --extsize 200 --keep-dup all -B --SPMR`. Peaks that overlap at least 50% between replicates were retained.

Hi-C data analysis

Reads were pooled by library and ran through the standard Hi-C pipeline using HOMER version 4.10.4.⁴³ Briefly, reads were trimmed with homerTools trim `-3 GATC -mis 0 -match-Start 20 -min 20`, mapped to human reference genome (GRCh38) with bowtie2 version 2.2.8,⁴⁴ and converted to tag

directory (`makeTagDirectory -genome hg38 -checkGC -restrictonSite GATC`). Matrices are normalized with analyze-HiC (standard protocol).

Typical and super enhancer databases

Strand specific, TT-seq was used to identify possible typical enhancers (TEs) and super enhancers (SEs). For this purpose, enhancers are defined as regions of the genome that are bidirectionally transcribed, and not in a gene or its promoter, or an annotated linc-RNA (Supplementary Figure S1). A 4-state Hidden Markov Model (HMM) on TT-seq both strands, TT-seq forward strand and TT-seq reverse strand, identified regions of the genome that are actively transcribed. Four states were chosen over a 2-state model because there were various amounts of transcription found in the genome; genic regions were easily identified, but low expression intergenic regions could not be identified with 2 states. Thus, 3 of the 4 states were coded for transcribed, and 1 state was labeled as non-transcribed. A database of possible enhancers was created by identifying regions of the genome that were identified as transcribed for all data and overlapped with regions that were both forward and reverse transcribed. Also added to the possible enhancer list were regions where there were overlapping forward and reverse transcription, but not identified as transcribed in all data. From this list, protein-coding genes with 2kb upstream and downstream were removed. Next, annotated regions from RNA-seq with an FPKM greater than 1 for all replicates and not protein-coding genes were removed. This final list of 20943 regions is purported enhancers.

Super enhancers were identified using Rose v0.1,^{45,46} stitching together a 12.5 kb distance, excluding 2.5 kb from TSS. Purported enhancer regions from TT-seq were used as previously identified enhancer regions. Merged, filtered, and read mapped duplicates removed bam files of histone marks (H3K27ac, H3K4me3, and H3K4me1) were used to rank the possible enhancers. As not all enhancers are identifiable with the 3-histone marks, TT-seq bam files were also used to identify SEs, and the output was filtered again for transcribed, annotated regions of the genome (eg, genes and linc-RNAs). The final SE database contained 767 merged regions (Supplementary Figure S1 and Supplementary Table S1), of which 360 were identified with H3K27ac alone, 436 identified with H3K4me3 alone, 301 identified with H3K4me1 alone, and 115 identified with TT-seq alone. The 767 SE regions were removed from the 20943 purported enhancer regions leaving 18357 possible enhancer regions. Of these regions, 701 overlap with H3K27ac peaks, 262 overlap with H3K4me3 peaks, 702 overlap with H3K4me1 peaks, and 1301 overlap with forward and reverse TT-seq transcribed regions with bidirectional transcription. The final merged regions contain 2180 enhancers (Supplementary Figure S1 and Supplementary Table S1). The closest TE and SEs to each provirus within the

6 HIV-1 integration groups (Figure 2A) were identified with bedtools closest (version 2.26.0).³⁴ HIV-1 expression versus distance to nearest TE or SE was plotted in R version R/3.3.2-gccmkl²⁶ using ggplot2.²⁷

De Novo identification of chromatin states with ChromHMM

We implemented ChromHMM,⁴⁷ which uses a multivariate HMM to calculate the probabilistic nature of a multi-state model and the biological nature of the state of chromatin at that location, to discover chromatin states in Jurkat T cells using epigenomics information derived from 7 individual ChIP-seq marks (H3K27me3, H3K4me3, H3K27ac, H3K4me1, H3K36me3, H3K79me3, and H3K9me3) known as state emissions. Filtered bam files, with mapping read duplicates removed, for each of the 7 histones, were individually converted to binary bin files with ChromHMM BinarizeBam v1.19.⁴⁷ We first obtained the state emissions of 15 different chromatin states defined as described on the Roadmap epigenomics project (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html) based on the observed data for the above 7 histone modifications. We used a core 15-state model for our analyses as it captured all the key interactions between the chromatin marks, and because larger numbers of states (eg, “expanded 18-state model”) did not apparently capture sufficiently distinct interactions. To *de novo* generate the core 15-state model in Jurkat T cells, we compared the relative abundance of the state emissions in Jurkat with known chromatin states for the 3 ENCODE cell lines most genetically and phenotypically linked with Jurkat (E115: Dnd41T cell leukemia, E116: GM12878 B cell lymphoblastoid, and E123: K562T cell leukemia) (Supplementary Figure S2A). To assign biologically meaningful *mnemonics* to the 15 chromatin states, we used the ChromHMM package to compute the overlap and neighborhood enrichments of each chromatin state relative to various types of functional annotations including the ChromHMM built in RefSeq annotations of (1) CpG islands; (2) genes; (3) exons; (4) introns; (5) TSS, 2 kb windows around TSS (TSS flanking), transcription termination sites (TTSs), and 2 kb windows around TTS (TTS flanking) based on the GENCODE v27 annotation; (6) Zinc finger (ZNF) genes obtained from ChromHMM; and (7) TEs and SEs obtained as described above (Figure 4A).

Identification of nuclear sub-compartments

The 2 Jurkat Hi-C libraries (see above) were combined with HOMER v4.10.4 makeTagDirectory using a standard protocol.⁴³ The eigen values of the first principal component of each chromosome was calculated with HOMER v4.10.4 run-HiCpca using a standard protocol.⁴³ The sign of the eigen values divide each chromosome; however, it does not state if either positive or negative values are representative of the A or B

sub-compartments. So, for each chromosome, the positive and negative eigen values were overlaid with Jurkat’s 15-state chromatin marks (ChromHMM v1.19 OverlapEnrichment).⁴⁷ The A and B sub-compartments were clearly defined with the B sub-compartments preferentially segregating in states 9, 13, and 15 (heterochromatin, repressed Polycomb, and quiescent/Low, respectively). The sign of the eigen values were then corrected so that positive values represented the A sub-compartment and negative values, B sub-compartment. *K*-means of $k=2$ was calculated on the A sub-compartment, and *k*-means $k=2$ through $k=5$ on the B sub-compartment using R v3.5.1 *k*-means. The A sub-compartment was labeled A1 or A2 based on the results of overlaying the *k*-means on the chromatin states; with A1 having much higher values than A2. The B sub-compartment was overlaid with the chromatin states for all *k*-mean $k=2$ through $k=5$; however, the results did not split the B sub-compartment on expected chromatin marks (Supplementary Figure S3A). Thus, the B sub-compartment was not subdivided. The GM12878 sub-compartments were retrieved from the Gene Expression Omnibus (GEO) database (GSE63525). The coordinates were lifted over to GRCh38 with UCSC liftOver.⁴⁸

Machine learning

To study the immediate landscape surrounding HIV-1 insertions (1559 unique insertions in total) and its possible effects on expression, we looked at 2 kb regions, in 200 bp increments around the HIV-1 integration sites, for a total of 280 features per integration site (Figure 6A and Supplementary Table S2). Reads per kilobase per million (RPKM) values of each 200 bp bin (20 bins in total) were calculated for the 7 histone marks (H3K27ac, H3K4me3, H3K4me1, H3K36me3, H3K79me3, H3K9me3, and H3K27me3), RNA-seq, MNase-seq, DNase-seq, and TT-seq using RPKM.py (https://git.biohpc.swmed.edu/venkat.malladi/miscellaneous_scripts/blob/master/scripts/rpkm.py). Discrete values for ChromHMM states and lamin sub-compartment states were also noted for each 200 bp region. The ChromHMM states were then converted from a categorical into a numerical value based on our understanding on its openness: U1 (most open), U4, U3, U6, U2, U7, U5, U10, U8, U11, U12, U14, U13, U9, and U15 (most close) in order. The HIV-1 expression level was normalized by z-transform and was annotated as “Low” if the normalized expression is lower than -0.5 ($n=351$), as “High” if higher than 0.5 ($n=455$), and otherwise as “Intermediate” ($n=753$) (Figure 6A). To determine optimal features, which have predictive power in HIV-1 fate prediction, and train a prediction model with them, an ML approach was taken. As a first step, the genetic landscape dataset was randomly split into a training dataset (75% of HIV-1 insertions) and a test dataset (25% of HIV-1 insertions). To select optimal features for HIV-1 expression level prediction, an R package, *smbinning* (<https://rdr.io/cran/smbinning/>), was used for feature selection in the training

dataset consisting of “High” and “Low” expression instances only. It returned each feature’s Information Value (IV), which is a powerful classifier that is relevant to its importance in the prediction task but does not explain how the features contribute to the prediction. We used the conventional threshold of $IV \geq 2$ as optimal features to train our model (Supplementary Table S3). After training a logistic regression model with the training dataset of these optimal features, the model was evaluated with the unseen test dataset (Supplementary Table S4). Note that while the “Intermediate” expression instances were excluded for the model training to get a better model, we included them for model evaluation to understand how the optimal features are contributing to the prediction. A metagene profile plot of 2kb region surrounding HIV-1 insertion in H3K27ac and MNase-seq (Figure 6B and C) as well as heatmaps of HIV-1 insertions and mean expression in each sub-compartment (Figure 6D and E) were created.

Quantification and statistical analysis

P-values were determined as indicated in each analysis and were indicated in the following manner: **P* < .05, ***P* < .01, ****P* < .001, *ns*, denotes non-significant. We considered *P* < .05 to be statistically significant.

Results

Expression of HIV-1 integration groups defined based on their position and orientation relative to genes in the human genome

To start addressing the relationship between HIV-1 integration and expression, we used the B-HIVE dataset from Jurkat CD4⁺ T cells,²⁴ which exploited the thousands of reporters integrated in parallel (TRIP) assay,⁴⁹ to simultaneously obtain HIV-1 positions and expressions (semi-quantitatively measured by RT-qPCR). Cells were selected for Green fluorescent protein (GFP) expression prior to gene expression analysis which removes any proviruses integrated into sites incompatible with transcription at the time of sorting and thus bias the data to initially active proviruses. While the virus used in the B-HIVE study is a minimalistic, replication-defective 5'-end barcoded virus, its integration pattern matches events observed in people living with HIV-1 on ART,²⁴ thus allowing us to use it as a model to define the effect of viral integration sites to expression.

Using this dataset, and because HIV-1 integrates within (intragenic) or between (intergenic) genes, and in the same or opposite orientation relative to the nearest human gene TSS, we first defined 6 “HIV-1 integration groups” based on their positions and orientations, relative to the nearest protein-coding gene, to explore any potential relationships of each group with human genomic features (Figure 2A). These 6 HIV-1 integration groups were implemented to increase the likelihood of obtaining group-specific trends, given that the analysis

in the entire dataset did not retrieve any high correlations for the various analyses (data not shown), thus reinforcing the idea that HIV-1 integration grouping may help identify regulatory features.

The 6 HIV-1 integration groups include *Group 1*, intergenic location and same orientation relative to nearby protein-coding gene (*n* = 127); *Group 2*, intergenic location and opposite orientation (convergent) relative to nearby protein-coding gene (*n* = 71); *Group 3*, intergenic location and opposite orientation (divergent) relative to nearby protein-coding gene (*n* = 77); *Group 4*, intragenic location and same orientation relative to protein-coding gene in which HIV-1 is integrated into (*n* = 581); *Group 5*, intragenic location and opposite (convergent) orientation relative to protein-coding gene in which HIV-1 is integrated into (*n* = 642); and *Group 6*, which is a composite of 3 subgroups depending on the 3 possible combinations of HIV-1 directions relative to the 2 human overlapping genes (*n* = 60) (*Group 6a*: HIV-1 in same direction to both genes [*n* = 26]; *Group 6b*: HIV-1 in same direction to only 1 gene [*n* = 24]; and *Group 6c*: HIV-1 in opposite direction to both genes [*n* = 10]) (Figure 2A). Given the limited number of integration events in each of the subgroups of Group 6, we treated them as a single group to increase statistical power.

Expectedly, there was a larger number of proviruses detected in the intragenic groups (Groups 4-6) relative to the intergenic groups (Groups 1-3) in the Jurkat CD4⁺ T cell model (Figure 2B).^{8,10,50,51} Having established the HIV-1 integration groups, we then examined the relationship between HIV-1 chromosomal distributions in each group and their expression levels by creating Circos plots²⁵ (Figure 2C to H). In these plots, each circle represents an individual HIV-1 position relative to chromosomes with their matched expression, in which the inner most line represents a \log_{10} of HIV-1 expression = -4, and the outer most line represents a \log_{10} of HIV-1 expression = 3. After visualizing the expression of each HIV-1 integration site on the 6 integration groups, we found that HIV-1 proviruses from each group were detected in every single chromosome with various expression levels irrespective of their group (Figure 2C to H).

Together, HIV-1 proviral transcription activity might be regulated by local and/or distal codes that are unique to the integration site and not shared within each integration group. As such, below we extend the analysis of the B-HIVE dataset to explore the contribution of individual, and combination of features to assess their effect to HIV-1 expression including (1) distance to nearby gene TSS (Figure 3A), (2) expression of nearby gene (Figure 3B), and (3) distance to nearby active enhancer (Figure 3C).

Relationship between HIV-1 expression and distance to, or expression of, nearest human gene

HIV-1 proviruses could be found at various distances relative to nearby protein-coding genes. Thus, to test the hypothesis that proviruses located closer to gene TSS are more

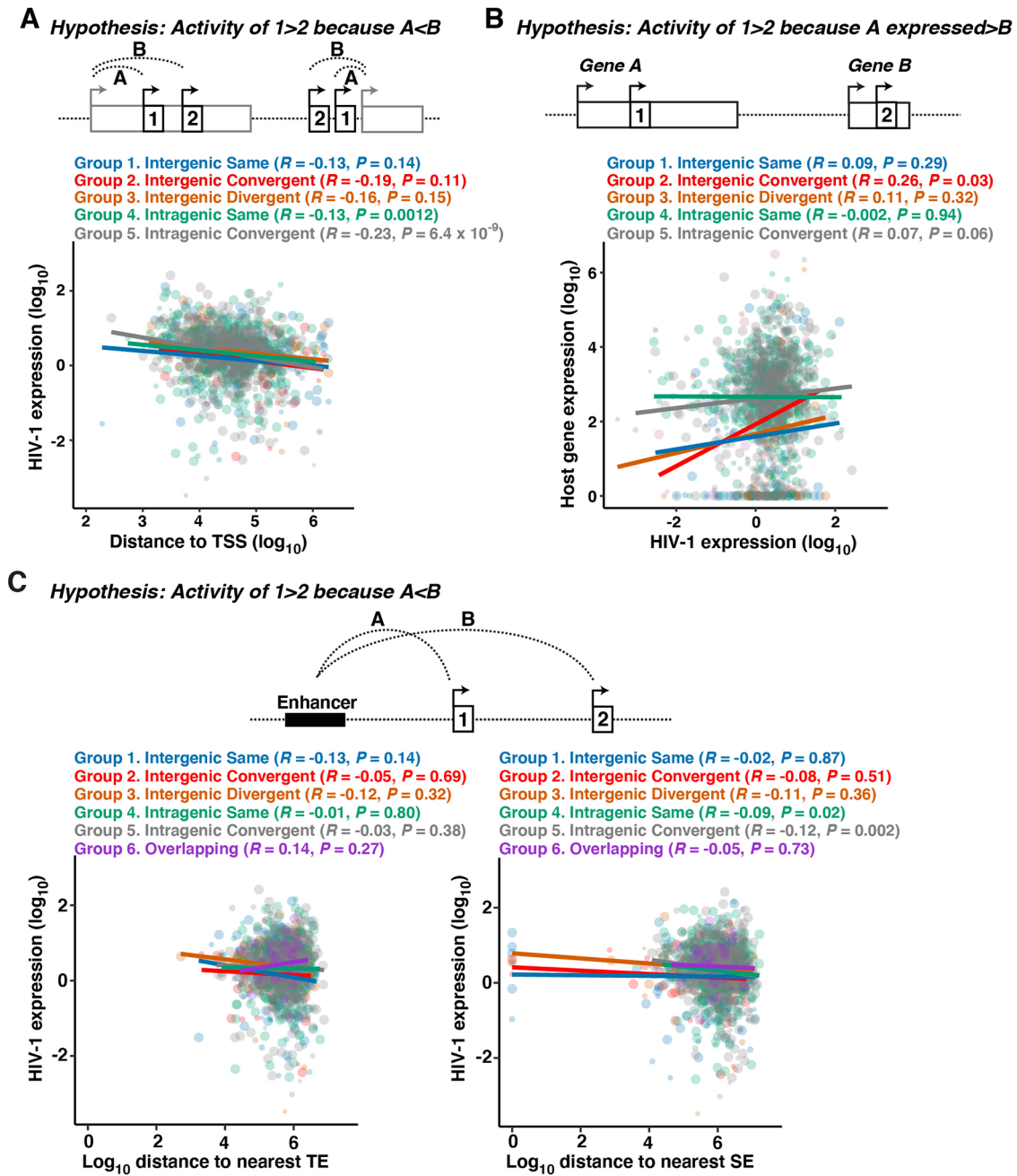


Figure 3. Scatter plots of the Pearson correlation coefficients and *P* values between expression of the proviral groups and various genomic features. (A) Hypothesis that HIV-1 expression is stronger the closer it is to the nearest gene TSS. Correlation of HIV-1 expression (\log_{10} RNA barcodes/DNA barcodes) to \log_{10} distance (bp) of the nearest gene TSS for each of the 5 HIV-1 integration groups. Group 6 (Overlapping) was excluded in A and B because of the complexity of HIV-1 association with 2 overlapping genes. (B) Hypothesis that HIV-1 expression ($\log_{10}[\text{RNA}_{\text{mean}}/\text{DNA}_{\text{mean}}]$) is correlated to the expression of its nearest gene, where if the nearest gene has higher expression, HIV-1 will also have higher expression. The opposite is hypothesized for lower expressed genes. Correlation of HIV-1 expression (\log_{10} RNA barcodes/DNA barcodes) as a function of host gene expression (\log_{10} FPKM) for each of the 5 HIV-1 integration groups. (C) Hypothesis that HIV-1 expression is higher the closer it is to an enhancer. Correlation of HIV-1 expression (\log_{10} RNA barcodes/DNA barcodes) to the \log_{10} distance (bp) of the nearest TE or SE. For all panels, the color-coding for each group is as follows: Group 1: Intergenic—Same (blue), Group 2: Intergenic—Convergent (red), Group 3: Intergenic—Divergent (orange), Group 4: Intragenic—Same (green), Group 5: Intragenic—Convergent (gray), and Group 6: Intergenic—Overlapping (purple). FPKM indicates Fragments Per Kilobase Million; TE, typical enhancer; TSS, transcription start site; SE, super enhancer.

active than those located at farther distances, we compared HIV-1 proviral expressions from B-HIVE dataset,²⁴ to their distances to nearby gene TSS derived from RefSeq annotations (Figure 3A). It is worth mentioning that this HIV-1 expression data cannot distinguish between readthrough

from host or intrinsic viral Long terminal repeat LTR-driven promoters. For this analysis, we omitted Group 6, which is composed of HIV-1 insertions within “2” overlapping genes, thus making it impossible to determine if 1 or both genes influence HIV-1 expression.

We found the expression of proviruses in intergenic regions is not correlated with the distance to the nearest gene TSS irrespective of their orientation—Group 1: Intergenic—Same ($R=-0.13$, $P=.14$), Group 2: Intergenic—Convergent ($R=-0.19$, $P=.11$), and Group 3: Intergenic—Divergent ($R=-0.16$, $P=.15$) (Figure 3A), potentially indicating that the orientation of intergenic proviruses is not a major feature controlling their expression. However, the expression of proviruses in intragenic regions was statistically significant, despite the low correlation coefficients—Group 4: Intragenic—Same ($R=-0.13$, $P=.0012$) and Group 5: Intragenic—Convergent ($R=-0.23$, $P=6.4 \times 10^{-9}$). Collectively, while HIV-1 proviral expressions in the various integration groups is not strongly correlated with their distance to the nearby gene TSS, the intragenic groups showed a statistically significant trend, especially Group 5 (Convergent orientation), signifying that this genomic feature, perhaps in combination with other features, may be a factor influencing HIV-1 expression.

Another regulatory feature shaping HIV-1 expression can be the level of expression of the nearest protein-coding gene. Thus, it is possible that the activity of provirus 1 is greater than provirus 2 if the gene associated with provirus 1 (Gene A) is expressed at higher levels than the gene linked to provirus 2 (Gene B) (Figure 3B). To test this hypothesis, we calculated the expression of proviruses in Groups 1 to 5 (derived from the B-HIVE dataset),²⁴ and of the nearest HIV-associated human protein-coding gene (derived from RNA-seq in Jurkat T cells collected in this study; Table 1). Once again, we excluded Group 6 (Overlapping) because of the complexity of HIV-1 association with 2 overlapping genes.

We found no statistically significant correlations for Group 1: Intergenic—Same ($R=0.09$, $P=.29$) and Group 3: Intergenic—Divergent ($R=0.11$, $P=.32$), but Group 2: Intergenic—Convergent, showed a low, but statistically significant, correlation ($R=0.26$, $P=.03$), suggesting that the convergent arrangement may offer HIV-1 an advantage for its expression, potentially linked to the lack of transcription interference by RNA polymerase II molecules transcribing host genes positioned in the same orientation as HIV-1.⁵² Given HIV-1 preferably integrates inside genes, we tested whether there is any correlation between the expression of HIV-1 in Group 4 (Intragenic—Same) and Group 5 (Intragenic—Convergent) but found that neither intragenic group was statistically significant nor correlated—Group 4: Intragenic—Same ($R=-0.002$, $P=.94$); Group 5: Intragenic—Convergent ($R=0.07$, $P=.06$) (Figure 3B). This indicates that transcription interference accounts for at least part of the observed proviral expression effects.

Contribution of human enhancers to HIV-1 proviral expression

HIV-1 displays preference to integrate into genes proximal to high density of enhancers,⁸ which are short DNA sequences

that act as transcription factor binding hubs controlling key transcriptional programs by fine-tuning target gene promoter activity across vast linear distances.^{46,53,54} Enhancers were also proposed to facilitate proviral expression.²⁴ However, because in this study enhancers were predicted based on the level of intergenic H3K27ac (a marker associated with transcription activity), but without incorporating transcription activity data, we carefully revisited this idea to explore whether HIV-1 positions relative to enhancers is a key regulatory element for determining proviral expression.

To this end, we first generated a rigorous and comprehensive database of active enhancers based on a combination of accepted epigenetic and transcriptional signatures including (1) a unique chromatin state demarcated by high H3K27ac, high H3K4me1, and low H3K4me3 levels derived from our ChIP-seq datasets in Jurkat T cells,¹⁹ and (2) symmetrical bidirectional enhancer RNA transcription (eRNA) derived from transient transcriptome (TT-seq) datasets.²² At least 2 types of active enhancer classes have been described: TEs and SEs (Figure 1A and Supplementary Figure S1A). The TEs contain the classic composition of features indicated above, whereas SEs are locally grouped clusters of enhancers (defined as a higher signal of H3K27ac, H3K4me3, H3K4me1, and eRNA transcription) within 12.5 kb of each other (Supplementary Figure S1A) driving high levels of transcription of nearby cell-identity genes.⁴⁶

To generate a database of TE and SE, an HMM was used to classify transcribed eRNA (sense and antisense strands in relation to the reference genome) regions that when overlapped with known enhancer histone marks (H3K27ac, H3K4me3, and H3K4me1), identified 2180 active intergenic TE (Supplementary Figure S1B and S1C). We defrayed from identifying intragenic TE given the high content of genic transcription and histone marks (eg, H3K27ac) potentially obscuring accurate identification of this class of enhancers. Using this approach, we also identified 767 SEs containing the conventional high density of clustered H3K27ac, H3K4me3, H3K4me1, and eRNA transcription activity (Supplementary Figure S1B and S1C, Supplementary Table S1).

To test the hypothesis that HIV-1 expression is correlated to its proximity to enhancers (Figure 3C), we used our assembled enhancer (TE and SE) databases to measure both the expression and distance of each provirus of the 6 different HIV-1 integration groups to the nearest TE and observed poor correlation and no statistical significance ($P<.05$) for all groups, suggesting that HIV-1 proximity to a TE alone is not a good predictor of HIV-1 proviral activity (Figure 3C), which is contrary to the previous study using less rigorous enhancer annotations.²⁴

Given previous reports that genes located near SE are transcribed to much higher levels compared with genes located near TE,⁴⁶ we hypothesized that if proximity to enhancers is a true regulator of proviral transcription, we would then expect that, compared with less active or latent proviruses, the most active

proviruses should be positioned nearer to SE. As such, we evaluated the distance of each provirus of the 6 HIV-1 integration groups to the nearest SE and found low correlations and no statistical significance for all 3 intergenic groups (Groups 1-3: same, convergent, divergent, respectively) (Figure 3C). However, interestingly, the intragenic groups (Groups 4-5) showed low, but statistically significant correlation—Group 4: Intragenic—Same ($R=-0.09$, $P=.02$) and Group 5: Intragenic—Convergent ($R=-0.12$, $P=.002$) (Figure 3C), consistent with the better correlations between the expression of both HIV-1 intragenic groups and their distance to the nearest gene TSS (Figure 3A) and transcription activity (Figure 3B). This suggests HIV-1 integrated into genes proximal to SE⁸ may have the dual benefit of increasing proviral expression.

Although HIV-1 is preferentially integrated into intragenic regions and HIV-1 expressions from these sites are significantly correlated with their distance to TSS, convergent orientation, and distance to SE, our study shows that each individual correlation is low and not the main genomic features influencing HIV-1 expression. It is evident from our analysis that HIV-1 expression might be influenced by multiple genomic features without apparent single-key determinants. Below, we investigate the relationship of HIV-1 integration and expression by further dividing the human host genome into chromatin states and spatial nuclear sub-compartments.

Contribution of chromatin states to HIV-1 proviral expression

Previous studies have provided low-resolution information on chromatin marks (eg, H3K4me3, H3K9me3, H3K27ac) content within HIV-1 proviruses in immortalized (Jurkat) models of latency.^{55,56} In addition, a recent survey analyzed the content of chromatin marks derived from uninfected CD4⁺ T cells surrounding (500bp regions centered) HIV-1 integration sites in primary CD4⁺ T cells, and suggested that productive integration events were associated with active chromatin including transcribed genes (H3K36me3) or enhancers (H3K4me1), while non-productive integration events appeared biased toward heterochromatin (H3K27me3 and H3K9me3) and non-accessible regions.¹³⁻¹⁶ However, chromatin marks alone do not provide a complete picture because they do not denote a singular function, and thus do not accurately enable representation of functional states. Conversely, “chromatin states” better demarcate functional genomic domains,⁵⁷ but have not been previously implemented to assess HIV-1 expression.

To classify genomic domains to precisely define chromatin states and to explore their contribution to HIV-1 expression, we applied “chromatin state learning” in Jurkat T cells with ChromHMM⁴⁷ (Figure 4A), like the annotations from the Roadmap Epigenomics Project in other cell types. ChromHMM is based on a multivariate HMM and thus allows capturing significant combinatorial interactions between

multiple chromatin marks in their spatial context. For our analysis, we used 7 chromatin marks derived from ChIP-seq datasets collected by our lab and others (Table 1). We chose a core 15-state model, as it captured all key interactions between the epigenetic marks and their chromatin states, and because the expanded (18-state) model did not improve chromatin states definition (data not shown).

To *de novo* generate the core 15-state model, we compared the enrichment of the 7 chromatin marks in a particular state in Jurkat T cells with known chromatin states for the 3 ENCODE cell lines most genetically and phenotypically linked to Jurkat including E115 (Dnd41T cell leukemia), E116 (GM1282878 B cell lymphoblastoid), and E123 (K562T cell leukemia) (Supplementary Figure S2A). This comparison allowed for the proper rearrangement and relabeling of the chromatin state numbering to correspond to identical labels from the Roadmap Epigenomics Project. Importantly, the chromatin state relabeling was validated by the presence of the expected chromatin marks in the different states (eg, H3K4me3 and H3K27ac in state 1—active TSS, and H3K27me3 in state 13—repressed Polycomb) as well as by the distribution of the chromatin states relative to transcription start and termination sites (Supplementary Figure S2B).

To first evaluate the chromatin states nomenclature, we defined the HIV-1 integration landscape as a function of the 15 states. Given that HIV-1 integrates within and/or near open chromatin based on accessibility and epigenetic data,⁸ we expected HIV-1 to be inserted into open chromatin states. In fact, HIV-1 was more likely to insert into “accessible” chromatin—states 2 to 8, 11 to 12, and 14 ($P<.05$, 2-proportions z -test)—and less likely to insert into “inaccessible” chromatin—states 13 and 15 ($P<.001$, 2-proportions z -test) (Supplementary Figure S2C). Notably, these data functionally validated the chromatin states definition, thus allowing us to address, for the first time, the contribution of chromatin states to HIV-1 proviral expression.

To assess this, we compared the median proviral expression in each chromatin state (Figure 4B). Because HIV-1 expression was not found to be normally distributed between chromatin states or have similar variances between them (data not shown), a Kruskal-Wallis rank sum test was used to assess significant differences of the median between different states. Surprisingly, HIV-1 expression was higher in state 7 (active enhancer) than states 4 or 5 (strong and weak transcription, respectively, $P<.05$) (Figure 4B and C). Not surprisingly, state 15 (quiescent/low) has lower expression than states 4, 5, 7, and 12 (strong transcription, weak transcription, active enhancer, and bivalent enhancer, respectively, $P<.05$).

Collectively, at difference to previous studies using individual chromatin marks, we precisely identified functional chromatin states in Jurkat T cells with ChromHMM to assess the relationship between the states and HIV-1 expression. With this rigorous demarcation of human genomic domains, our

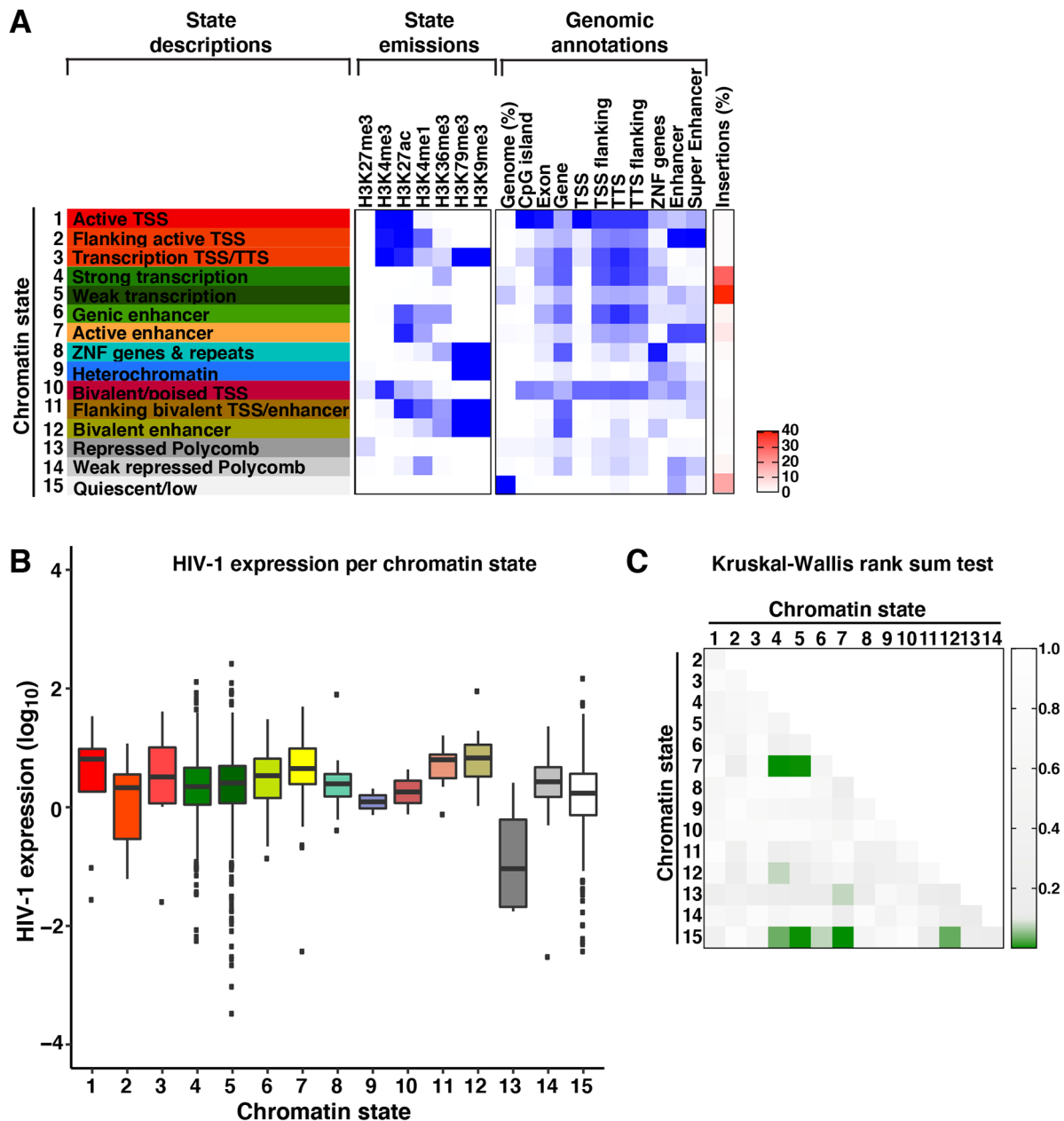


Figure 4. Expression of HIV-1 proviruses in relation to chromatin states defined using ChromHMM. (A) From left to right: ChromHMM plot containing the 15 states as defined by Roadmap Epigenomics Project. Histone marks used to create the states (state emissions). Overlap enrichment plots (ChromHMM) of RefSeq genomic annotations (CpG island, Exon, Gene, TSS, TSS flanking, TTS, TTS flanking, and ZNF genes), TE, and SE by 15 states. Heatmap of percentage of insertions in each state. (B) Box plot of B-HIVE expression by 15-state model. The color-coding is in reference to each of the chromatin states as shown in panel A. (C) Heatmap of P values comparing the median expression of B-HIVE in 2 different states by Kruskal–Wallis rank sum test. B-HIVE indicates *Barcoded HIV-1 Ensembles*; TE, typical enhancer; TSS, transcription start site; TTS, transcription termination site; SE, super enhancer; ZNF, Zinc finger.

results revealed that open regions associated with promoters and enhancers (both mono- and bidirectional) have higher HIV-1 mean expressions than those in inaccessible regions.

Contribution of nuclear spatial sub-compartments to HIV-1 proviral expression

The 3D organization of chromosomes enables long-range interactions between enhancers and promoters that are critical for building complex gene regulatory networks.^{58,59} As such,

these genomic contacts could be co-opted by HIV-1 to transcribe its genome to perpetuate the infection.

Interphase chromosomes occupy separate spaces known as nuclear territories,⁶⁰ and each chromosome is organized into dynamic, non-random structures containing stretches of transcriptionally active compartments interspersed with sections of transcriptionally inactive compartments.⁶¹ As such, the genome is partitioned into contact domains (A and B) segregating into 6 sub-compartments (A1, A2, B1, B2, B3, and B4) that (1) appear located in different nuclear territories, (2) are associated

with distinct patterns of histone marks, and (3) show different expression levels.⁶² To examine the integration–expression relationship of HIV-1 proviruses in the various sub-compartments, we used 2 complementary approaches. First, we re-analyzed a Hi-C dataset in Jurkat T cells,⁸ which is a method that interrogates the 3D architecture of whole genomes by coupling proximity-based ligation with massively parallel sequencing. Second, we predicted sub-compartments in Jurkat T cells using the lamin sub-compartments derived from GM12878 cells.⁶²

Using the Hi-C dataset from Jurkat T cells, we followed common practices in the field to identify the active and inactive (A and B, respectively) sub-compartments using the first eigenvector (first principal component), through genome partitioning into binary classifications based on genomic contacts and chromatin marks content (Figure 4A). We then further divided the A compartment into A1 and A2 using a k -means clustering of $k=2$. However, k -means clustering of the B compartment, with $k=2$ through $k=5$, did not yield a clearly defined separation of 3 regions based on expected content of chromatin marks (Supplementary Figure S3A), leaving us to keep all B sub-compartments together for the analysis of HIV-1 positions and relationship to expression relative to sub-compartments.

Expectedly, the distribution of proviruses revealed there is a preferential insertion into the A sub-compartments ($P < .001$, 2-proportions z -test), and reduced insertions within B sub-compartment ($P < .001$, 2-proportions z -test), than by random occurrence given the Jurkat T cell sub-compartments coverage (Supplementary Figure S3B to S3D), consistent with recent observations,⁸ indicating we were poised to evaluate the relationship between HIV-1 integrations and expression.

Importantly, the mean HIV-1 expressions in the A1 and A2 sub-compartments was significantly higher from the B compartment ($P < .001$, Kruskal–Wallis rank sum test) without any obvious, statistically significant difference between the A1 and A2 sub-compartments ($P = .43$) (Supplementary Figure S3D and S3E), suggesting HIV-1 integration into more accessible regions of the genome is, in general, beneficial for proviral expression.

Because we noticed the Jurkat Hi-C dataset was too sparse to divide into the 6 sub-compartments (A1, A2, B1, B2, B3, and B4), we then used the higher resolution sub-compartments defined in GM12878, which is a B cell line genotypically and phenotypically closely related to Jurkat, thus arguing that these results could be applicable to Jurkat, as domains are typically conserved (~80%) between cell types.⁶² Using these data, expectedly, we found a significant increase of HIV-1 proviruses in sub-compartments A1, A2, and B4, and significantly lower insertions in sub-compartments B1, B2, B3 ($P < .001$, 2-proportions z -test) (Figure 5A to C), thus signifying we were in the right track to test the relationship between HIV-1 integration and expression.

Notably, we found enrichment of mean HIV-1 expressions in sub-compartments A1 versus A2, B1, B2, or B3 ($P < .001$, Kruskal–Wallis rank sum test) (Figure 5D), suggesting the location of HIV-1 into categorical nuclear sub-compartments

overall contributes to increased expression. Importantly, these results are consistent with the Hi-C data analysis in Jurkat T cells (Supplementary Figure S3D and S3E), suggesting there is a conservation of sub-compartments between the 2 cell types, consistent with previous results.⁶²

Interestingly, sub-compartment B4 has an increase of HIV-1 proviruses insertions relative to its size in the genome ($P < .05$, 2-proportions z -test), but no significant increase or decrease in expression relative to other sub-compartments (Figure 5C and D). B4 consists of regions, mostly on chromosome 19, containing many of the KRAB-ZNF superfamily genes. In our chromatin state analysis, we also noticed an increase of HIV-1 insertions in state 8 (ZNF genes and repeats), but with no increase or decrease in relative expression to other chromatin states (Figure 4). Interestingly, this is consistent with the skewed prevalence of HIV-1 in this state in individuals who can immunologically control HIV-1.⁶³

Machine learning approach to train a model predicting HIV-1 proviral expression

To foresee regulatory features, a first-ever ML approach was employed to train a logistic regression model to predict HIV-1 expression from integration positions using genomic datasets. For this, we examined HIV-1 integration site-proximal (2kb) regions, in 200bp increments, around integration sites from the B-HIVE dataset (Figure 6A). HIV-1 integration sites were z -transformed and classified based on their normalized expression values as “Low” (< -0.5), “Intermediate” (-0.5 to 0.5), or “High” (> 0.5). Among the entire integration dataset, 75% of the Low and High expressed proviruses were used to train the model.

When a threshold of IV, a ranking of variables used for feature selection in binary logistic regression models, ≥ 2 was applied, 26 regulatory features were determined as optimal for the prediction task, which included all 20 lamin sub-compartment bins (from Hi-C), 5 bins in H3K27ac (from ChIP-seq), and 1 bin in chromatin accessibility (from MNase-seq) (Supplementary Table S3). Our model revealed unique, optimal features for the prediction, including upstream histone acetylation (H3K27ac) and chromatin accessibility and categorical nuclear sub-compartments, especially sub-compartment A1 (Supplementary Table S4 and Figure 6B to E). It is not readily apparent when plotting the features in a metagene analysis, where bins are the most informative (Figure 6B and C). For instance, visual inspection of the MNase-seq bins (Figure 6C) looks noisy when comparing the signal of the High and Low integrants. However, the MNase-seq bin -1000 to -1200 bp upstream of HIV-1 integration was identified with the ML model as being informative (Figure 6C). Also, in the H3K27ac data analysis, nucleosome periodicity on the left was clear, and the signal density on the left versus right side near 2kb away from HIV-1 integration was similar and higher compared with signal proximal to the integration site (Figure 6B). However, the model only picked bins to the left and away

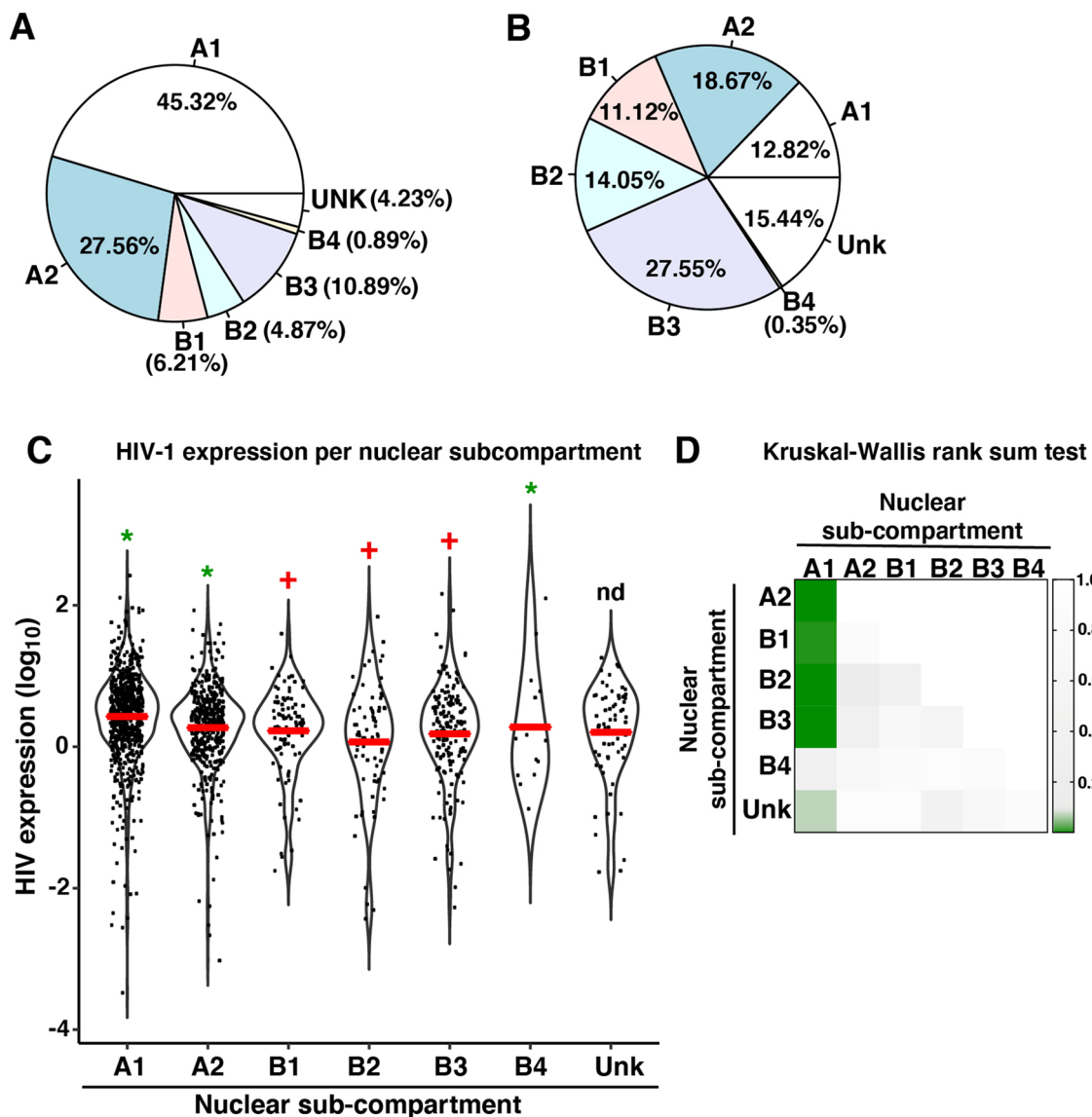


Figure 5. Expression of HIV-1 integration and expression in relation to 3D architecture. (A) Pie chart of percentage of HIV-1 integration per GM12878 nuclear sub-compartments. Unk denotes unknown sub-compartments. (B) Pie chart of GM12878 sub-compartments genomic coverage (as percentage). Unk denotes unknown sub-compartments. (C) Jitter and violin plot of HIV-1 expression in each GM12878 sub-compartments derived from Hi-C data. The red line represents mean expression. Green asterisks represent significantly more insertions relative to sub-compartment genomic coverage, and red crosses represent significantly less insertions; nd was not calculated due to unknown sub-compartments ($P < .05$, 2-proportions z-test). (D) Heatmap representing the P value (P) pairwise comparison of sub-compartments derived from GM12878 versus expression using a Kruskal–Wallis rank sum test.

from the integration site as informative ranked variables for the prediction task (Supplementary Table S4 and Figure 6B).

For the test dataset, we used the remaining 25% of Low and High expressing HIV-1 positions, and all Intermediate expressing HIV-1 positions. HIV-1 expression levels were predicted through the logistic regression model of the optimal features, their estimated weights and corresponding odds ratios, and the standard errors of the estimated weights were obtained (Supplementary Table S4). The evaluation metrics were calculated as 68.42% of sensitivity, 59.10% of specificity, and 64.71% of area under the receiver operating characteristics (Figure 6F), revealing the model has fair-to-low prediction power. Notably, the predicted HIV-1 expression values for the “High”

expression category were significantly higher than those for the “Low” category ($P = .00025$, Wilcoxon test) (Figure 6G), revealing the model can capture large expression differences. However, expectedly, the predicted values for the “Intermediate” and “Low” categories showed not significantly statistic differences ($P = .27$, Wilcoxon test) (Figure 6G), perhaps because the model was trained using high and low expressing groups. Furthermore, the actual and predicted HIV-1 expression values were found to have a positive, moderate correlation ($R = 0.19$, $P = .00018$) (Figure 6H), indicating the model can detect the expected differences in HIV-1 expression. Nonetheless, the modest prediction may be attributed to the complex biology of HIV-1 integration and expression, which the present model

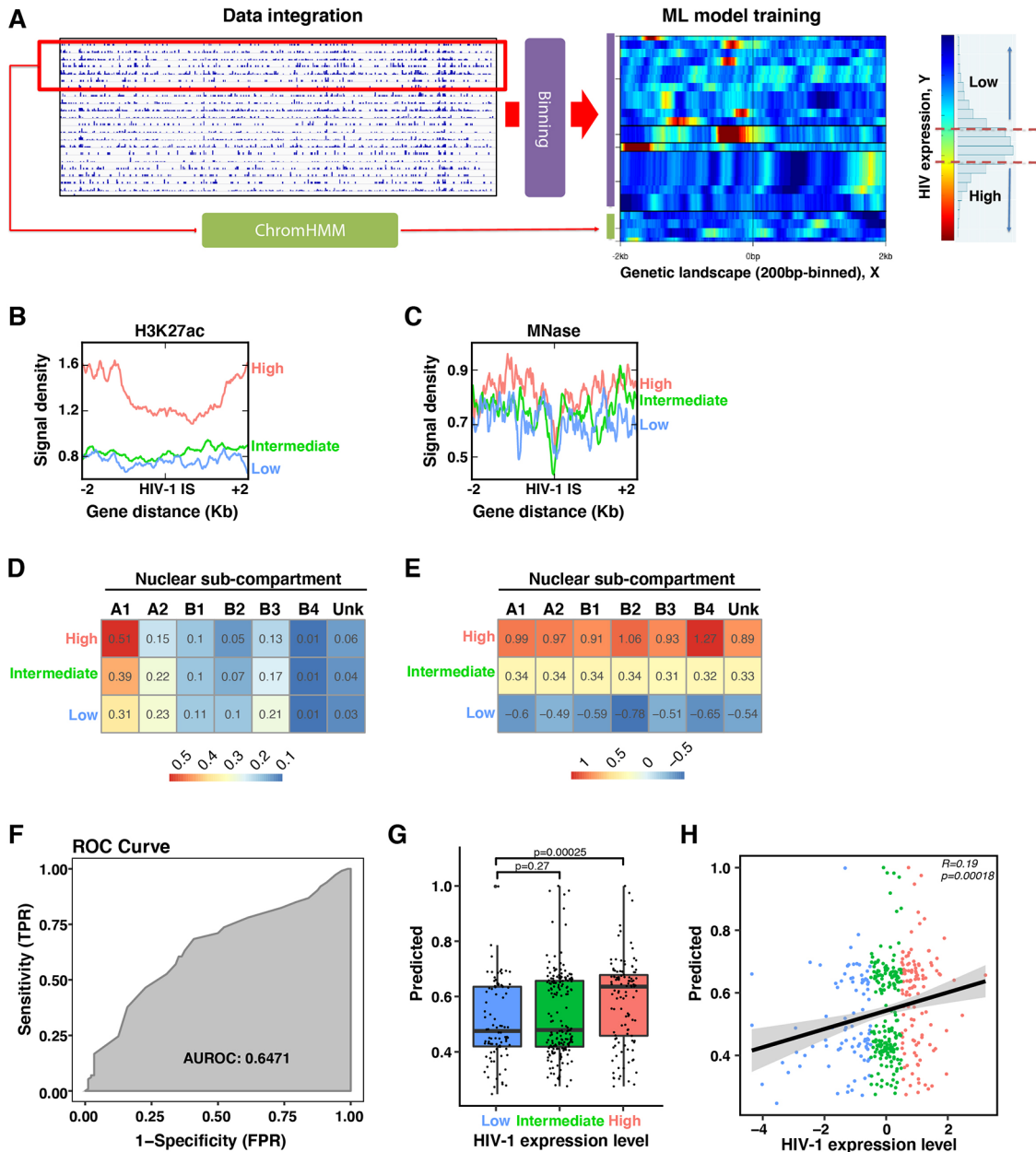


Figure 6. HIV-1 expression level prediction based on genetic and 3D landscape surrounding integration sites. (A) Schematic of data integration and training of the machine learning (ML) model. Genetic marks 2 kb surrounding HIV-1 insertions were binned every 200bp and measured. Barcoded HIV-1 Ensembles (B-HIVE) data were split into 3 groups (Low, Intermediate, and High), training the ML model for Low versus High. (B and C) Metagenes plots of optimal features (H3K27ac and MNase-seq) 2 kb surrounding B-HIVE insertions, split by the 3 categories based on expression (Low, Intermediate, and High). (D) Heatmap of the percentage of B-HIVE insertions, split by the 3 categories of Low, Intermediate, and High, versus GSM12878 sub-compartments. (E) Heatmap of the mean expression of B-HIVE insertions, split by the 3 categories (Low, Intermediate, and High) versus GSM12878 sub-compartments. (F) The trained model’s AUROC (area under the receiver operating characteristics) curve is shown. (G) Predicted HIV-1 expression comparison among 3 categories of Low, Intermediate, and High. *P* value (*P*) calculated from Wilcoxon test. (H) Linear regression and Pearson correlation test between the actual and predicted HIV-1 expression values.

may not be able to accurately predict using the combined datasets and low number of integration events ($n = 1558$).

Taken together, the model succeeded at predicting informative variables in the human genome leading to large HIV-1 expression differences. Notably, the ranked variables obtained can assist with experimental design, selecting the most informative genomic assays, for future larger-scale experiments.

Discussion

In this work, we have applied an integrated genomics approach, by combining new and open-source datasets to interrogate how the position of HIV-1 proviruses into the human genome shapes their expression. We also implemented an ML approach to delineate features in the human genome predicting HIV-1 expression. It is the combination of these 2 approaches that

have enabled us to start defining relationships between HIV-1 integration and transcription and to fuel the most critical questions for future experimental design.

With the addition of larger integration datasets, possibly encompassing the entire integration landscape (chromatin states in each of the sub-compartments and including inter- and intragenic integrants) and their measured transcription profiles, and not steady-state expression (which cannot distinguish between transcription and downstream regulatory events), more refined models with better predictive powers could be achieved. As the datasets grow, we expect that the genetic landscape related to HIV-1 transcription will be more prominent and, accordingly, future ML approaches can be used to provide a compass compatible with clinical decision-making (Figure 1B).

To date, no other research or analysis has addressed, at this scale, how human genome codes effect HIV-1 proviral transcription. Of course, importantly, the results and predictions of the various data analyses must be followed up by experimental testing and must be validated in models other than Jurkat (eg, primary T cells), which certainly imposes other challenges. As such, we envision that future work will be needed (1) to study differences in epigenomic landscapes before and after HIV-1 insertion; (2) to implement a Clustered regularly interspaced short palindromic repeats (CRISPR) based integration library in which HIV-1 is uniquely positioned in defined integration landscapes to study and develop better, more complete ML models; and (3) to study HIV-1 integrants at the single-cell level to define their locations and intactness (MIP-seq),⁶⁴ expression (scRNA-seq), and chromatin landscapes (scATAC-seq). These single-cell studies are the only ones that will allow to study the relationship between HIV-1 position, intactness, and expression as well as viral-host chromatin landscapes. However, a significantly large number of insertions must occur in enough locations of all combinations of nuclear sub-compartments by chromatin state by inter/intragenic location, and measured along with HIV-1 expression to have the required power to obtain statistically meaningful data.

Finally, our analysis demonstrates the importance of carefully and correctly characterizing regulatory features (eg, functional chromatin states) when studying their potential role in genome regulation. Future studies should be cautious to use the same rigorous standards when defining genomic domains to interrogate functional insights in human health and disease.

Acknowledgements

We thank members of the D'Orso laboratory (Nora-Guadalupe Ramirez, Ashutosh Shukla, Jinli Wang, and Usman Hyder) for critically reading of the manuscript and Chi Pak for generating the MNase-seq dataset.

Author Contributions

HR and ID developed the idea. HR, JL, and ID designed the experiments. HR, JL, and ID analyzed the data. HR and ID

wrote the manuscript with feedback from all co-authors. ID obtained funding to support the studies.


Availability of Data and Materials

Requests for further information and reagents may be directed to the Lead Contact, Dr Iván D'Orso, at the University of Texas Southwestern Medical Center (ivan.dorso@utsouthwestern.edu). All software used in this study are individually listed above under each data analysis tool. The computer code used for this analysis is available on GitHub: https://github.com/utsw-bicf/HIVproviral_fate. NGS datasets used in this study (Table 1) were downloaded from NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/)⁶⁵ or ENCODE project (www.encodeproject.org). MNase-seq raw and processed sequencing results have been submitted to NCBI Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE144753. RNA-seq raw and processed sequencing results have been submitted to NCBI Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE144552.

ORCID iDs

Holly Ruess  <https://orcid.org/0000-0001-9148-6672>

Jeon Lee  <https://orcid.org/0000-0002-9071-7157>

Venkat S Malladi  <https://orcid.org/0000-0002-0144-0564>

Iván D'Orso  <https://orcid.org/0000-0002-1409-2351>

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

- Hughes SH, Coffin JM. What integration sites tell us about HIV persistence. *Cell Host Microbe*. 2016;19:588-598.
- Mbonye U, Karn J. Transcriptional control of HIV latency: cellular signaling pathways, epigenetics, happenstance and the hope for a cure. *Virology*. 2014;454-455:328-339.
- Mbonye U, Karn J. The molecular basis for human immunodeficiency virus latency. *Annu Rev Virol*. 2017;4:261-285.
- Michieletto D, Lusic M, Marenduzzo D, Orlandini E. Physical principles of retroviral integration in the human genome. *Nat Commun*. 2019;10:575.
- Morton EL, Forst CV, Zheng Y, et al. Transcriptional circuit fragility influences HIV proviral fate. *Cell Rep*. 2019;27:154-171.
- Ott M, Geyer M, Zhou Q. The control of HIV transcription: keeping RNA polymerase II on track. *Cell Host Microbe*. 2011;10:426-435.
- Cohn LB, Silva IT, Oliveira TY, et al. HIV-1 integration landscape during latent and active infection. *Cell*. 2015;160:420-432.
- Lucic B, Chen HC, Kuzman M, et al. Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration. *Nat Commun*. 2019;10:4059.
- Marini B, Kertesz-Farkas A, Ali H, et al. Nuclear architecture dictates HIV-1 integration site selection. *Nature*. 2015;521:227-231.
- Schroder AR, Shinn P, Chen H, et al. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*. 2002;110:521-529.
- Yoon JK, Holloway JR, Wells DW, et al. HIV proviral DNA integration can drive T cell growth ex vivo. *Proc Natl Acad Sci U S A*. 2020;117:32880-32882.
- Shukla A, Ramirez NP, D'Orso I. HIV-1 proviral transcription and latency in the new era. *Viruses*. 2020;12:555.
- Battivelli E, Dahabieh MS, Abdel-Mohsen M, et al. Distinct chromatin functional states correlate with HIV latency reactivation in infected primary CD4(+) T cells. *eLife*. 2018;7:e34655.
- Bukrinsky MI, Stanwick TL, Dempsey MP, Stevenson M. Quiescent T lymphocytes as an inducible virus reservoir in HIV-1 infection. *Science*. 1991;254:423-427.

15. Jordan A, Defechereux P, Verdin E. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *EMBO J.* 2001;20:1726-1738.
16. Sherrill-Mix S, Lewinski MK, Famiglietti M, et al. HIV latency and integration site placement in five cell-based models. *Retrovirology.* 2013;10:90.
17. Lucic B, Lucic M. Connecting HIV-1 integration and transcription: a step toward new treatments. *FEBS Lett.* 2016;590:1927-1939.
18. Maxfield LF, Fraize CD, Coffin JM. Relationship between retroviral DNA-integration-site selection and host cell transcription. *Proc Natl Acad Sci U S A.* 2005;102:1436-1441.
19. Reeder JE, Kwak YT, McNamara RP, Forst CV, D'Orso I. HIV Tat controls RNA Polymerase II and the epigenetic landscape to transcriptionally reprogram target immune cells. *eLife.* 2015;4:e08955.
20. Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. *Cell.* 2013;155:934-947.
21. Mansour MR, Abraham BJ, Anders L, et al. Oncogene regulation: an oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science.* 2014;346:1373-1377.
22. Michel M, Demel C, Zacher B, et al. TT-seq captures enhancer landscapes immediately after T-cell stimulation. *Mol Syst Biol.* 2017;13:920.
23. Robson MI, de Las Heras JJ, Czapiewski R, Sivakumar A, Kerr ARW, Schirmer EC. Constrained release of lamina-associated enhancers and genes from the nuclear envelope during T-cell activation facilitates their association in chromosome compartments. *Genome Res.* 2017;27:1126-1138.
24. Chen HC, Martinez JP, Zorita E, Meyerhans A, Filion GJ. Position effects influence HIV latency reversal. *Nat Struct Mol Biol.* 2017;24:47-54.
25. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* 2009;19:1639-1645.
26. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2014.
27. Wickham H. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag; 2016.
28. Di Tommaso P, Chatzou M, Floden EW, et al. Nextflow enables reproducible computational workflows. *Nat Biotechnol.* 2017;35:316-319.
29. Barnes SD, Ruess H, Mathews JA, Cheng B, Malladi VS. *BICF ChIP-seq Analysis Workflow* (publish_1.0.5). Geneva, Switzerland: Zenodo; 2019. doi:10.5281/zenodo.2648844.
30. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 2011;17:1-10.
31. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25:1754-1760.
32. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25:2078-2079.
33. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P, Sambamba: fast processing of NGS alignment formats. *Bioinformatics.* 2015;31:2032-2034.
34. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841-842.
35. Ramirez F, Ryan DP, Gruning B, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 2016;44:W160-W165.
36. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol.* 2008;26:1351-1359.
37. Landt SG, Marinov GK, Kundaje A, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 2012;22:1813-1831.
38. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9:R137.
39. Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics.* 2015;31:2382-2383.
40. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc.* 2016;11:1650-1667.
41. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 2014;30:923-930.
42. Chen W, Liu Y, Zhu S, et al. Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. *Nat Commun.* 2014;5:4909.
43. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* 2010;38:576-589.
44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9:357-359.
45. Loven J, Hoke HA, Lin CY, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell.* 2013;153:320-334.
46. Whyte WA, Orlando DA, Hnisz D, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell.* 2013;153:307-319.
47. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc.* 2017;12:2478-2492.
48. Hinrichs AS, Karolchik D, Baertsch R, et al. The UCSC genome browser database: update 2006. *Nucleic Acids Res.* 2006;34:D590-598.
49. Akhtar W, de Jong HA, Pindyurin AV, et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell.* 2013;154:914-927.
50. 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.* 2007;195:716-725.
51. Maldarelli F, Wu X, Su L, et al. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science.* 2014;345:179-183.
52. Lenasi T, Contreras X, Peterlin BM. Transcriptional interference antagonizes proviral gene expression to promote HIV latency. *Cell Host Microbe.* 2008;4:123-133.
53. Kim TK, Shiekhattar R. Architectural and functional commonalities between enhancers and promoters. *Cell.* 2015;162:948-959.
54. Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat Rev Genet.* 2016;17:207-223.
55. Lusic M, Marcello A, Cereseto A, Giacca M. Regulation of HIV-1 gene expression by histone acetylation and factor recruitment at the LTR promoter. *EMBO J.* 2003;22:6550-6561.
56. Pearson R, Kim YK, Hokello J, et al. Epigenetic silencing of human immunodeficiency virus (HIV) transcription by formation of restrictive chromatin structures at the viral long terminal repeat drives the progressive entry of HIV into latency. *J Virol.* 2008;82:12291-12303.
57. Roadmap Epigenomics C, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518:317-330.
58. Dekker J, Mirny L. The 3D genome as moderator of chromosomal communication. *Cell.* 2016;164:1110-1121.
59. Yu M, Ren B. The three-dimensional organization of mammalian genomes. *Annu Rev Cell Dev Biol.* 2017;33:265-289.
60. Meaburn KJ, Misteli T. Cell biology: chromosome territories. *Nature.* 2007;445:379-781.
61. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 2009;326:289-293.
62. Rao SS, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.* 2014;159:1665-1680.
63. Jiang C, Lian X, Gao C, et al. Distinct viral reservoirs in individuals with spontaneous control of HIV-1. *Nature.* 2020;585:261-267.
64. Einkauf KB, Lee GQ, Gao C, et al. Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J Clin Invest.* 2019;129:988-998.
65. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 2013;41:D991-D995.