**Deciphering human genome codes shaping HIV proviral transcription and fate**

**An integrated genomics approach deciphering human genome codes shaping HIV proviral transcription and fate**

Holly Ruess1, Jeon Lee1, Carlos Guzman?2,3, Chi Pak?2,4, Venkat Malladi1, and Iván D’Orso2,5

*1Bioinformatics Core Facility, The University of Texas Southwestern Medical Center, UT Southwestern, Dallas, TX 75390, USA*

*2Department of Microbiology, The University of Texas Southwestern Medical Center, UT Southwestern, Dallas, TX 75390, USA*

*3Present address: Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA 92093*

*4Present address?*

*5Lead contact*: [Ivan.Dorso@utsouthwestern.edu](mailto:Ivan.Dorso@utsouthwestern.edu)

**SUMMARY**

HIV integrates semi-randomly into the genome of immune cells and thus proviruses that persist in patients under long-term, highly active suppressive therapy can be detected in various positions (inside and outside) and orientations (same, convergent and divergent) respective to genes, promoters, and enhancers. Thus, this integration landscape heterogeneity can influence HIV transcription activity thereby dictating proviral fate (active vs latent). However, the effect of the integration site (here referred to as “integration code paradigm”) to proviral transcription activity has so far remained elusive. Here we integrate open-source, large-scale datasets including epigenetics, transcriptome, and 3D genome architecture to interrogate the chromatin states, transcription activity landscape, nuclear sub-compartments, and topological associated domains around HIV integration sites in CD4+ T cell-based models to decipher ‘codes’ in the human genome shaping proviral transcription. Our results provide evidence of the importance of HIV proviral placement on the human genome to its transcriptional activity. Specifically, we found a positive correlation of expression for proviruses for the intragenic (same and convergent orientation) and intergenic (same orientation) groups based on proximity to the nearest human gene promoter, and both intergenic and intragenic (convergent orientation) based on HIV-associated human gene transcriptional activity, and no apparent contribution of proximal enhancers based on linear and not 3D scale. Furthermore, amachine-learning logistic regression modelof a 2-kb region around HIV integration sites reveals upstream chromatin accessibility and transcription activity and categorical nuclear sub-compartments as optimal features predicting HIV transcriptional outcomes (*resulting in a predicted expression level sensitivity of 68.42% and a specificity of 59.10% for test data*). Importantly, our studies provide clinical relevance as intact proviruses in patients (which can resume infection after therapy interruption) are enriched for intergenic positions more frequently positioned in an opposite orientation relative to host genes or located in either relative proximity to or increased distance to gene promoters and chromatin accessible regions. Our studies highlight the importance of specific chromatin states and genome architecture in the control of HIV transcription activity and proviral fate. Future interrogation of larger interrogation libraries in patient samples combined with deep learning will inform rational approaches to target the various genome-intact HIV groups.

**INTRODUCTION**

Add section of Patients (**Figure 1A**) and ART (**Figure 1B**)

The Human Immunodeficiency Virus (HIV) preferentially integrates into chromatin-accessible sites and transcriptionally active regions of the human genome {Schroder, 2002 #301;Marini, 2015 #343;Lucic, 2019 #796}, but proviruses are detected on every human chromosome, in various chromatin landscapes (euchromatic and heterochromatic), and at different locations (inside and outside) and directions (same or opposite) respective to human genes (**Figure 1C**). Thus, the integration landscape is highly diverse in sequence and chromatin structure, and could influence HIV transcription levels {Bukrinsky, 1991 #797;Jordan, 2001 #300;Sherrill-Mix, 2013 #340}, ultimately impacting proviral fate (active or latent) (**Figure 1B**). The provirus could be active or be converted into the latent state (probably if the integration locus was transiently induced during infection and then decayed over time –like a human inducible gene that responds to stimulation). Despite this simple binary classification, HIV is a collection of species having different degrees of transcription activation states. As such, the active state is not only one state, but a collection of states with different degrees of activation here referred to as “activity continuum” from low-to-high (**Figure 1?**). In addition, the latent state can, in certain circumstances such as exposure to stimuli (cytokines of the environmental milieu), become reactivated.

Although proviral fate is seemingly governed by its integration site, the genomic context that leads to provirus activation or latency still remains poorly understood. Given that HIV contains its own promoter, one would assume that the virus could function in a cell-autonomous manner (irrespective of the integration site). However, work by the Verdin lab in 2001 has suggested that the integration site could control HIV transcription {Jordan, 2001 #300}. As any other gene promoter, HIV can be regulated at the basal level and in response to environmental stimulation. Basal transcription alone is not enough for the production of the large amount of viral products required to perpetuate the infection. When infected immune cells are exposed to stimulation, the viral genome is transcribed by the sequential action of cellular (e.g., NF-B) and viral (Tat) activators, which promote a positive feedback loop guaranteeing the production of high-levels of viral products. Despite our current knowledge on how the HIV transcriptional program operates, we have no clear understanding on the role of the integration landscape on proviral fate. Given the variable fate of the infection upon integration into the host, it is logical to speculate that HIV is a cell-non-autonomous unit strictly relying on the integration site for productive infection. However, it remains unclear how the integration site in the human genome shapes proviral fate. The integration landscape might provide instructions –here referred to as “*integration code*”– where combinations of genetic/epigenetic features (1D space) and/or its location respective to nuclear territories (3D/4D space) influences HIV proviral activity and fate. Position effect variegation (PEV) was previously described in *Drosophila{Reuter, 1992 #798;Akhtar, 2013 #799}.* The long-term objective of these studies was to define if there is any pattern of host genetic/epigenetic features that predictably influences HIV expression, latency and reversal. To study the influence of the integration landscape, host genome expression, and architecture, to HIV proviral transcription and fate we first collected/integrated a series of next-generation sequencing (NGS) datasets including position information from the B-HIVE dataset{Chen, 2017 #423}, epigenetic marks and chromHMM information{Ernst, 2017 #805}, lamin subcompartments, identification of TADs and loops, human CD4 T cell expression, genome accessibility, genome annotations (promoters and enhancers) and implementation of machine learning (**Figure 1E**).

For over several million of years, a broad range of retroviruses has invaded our genomes. Since retroviruses integrate in a “semi-random” manner into chromatin accessible sites of the human genome, it seems logical to speculate that the integration site could contain information (here referred to as “***integration code***”) that determines the magnitude of HIV transcription and proviral fate. Given that retroviruses like HIV contain their own promoter elements, they strictly rely on genetic/epigenetic elements in the human genome to be transcribed and replicate. One of the key questions that remain to be answered in the field is “How does site integration affect HIV transcription, latency and reversal?” The dogma in the field poses that “local genetic/epigenetic features influence HIV transcription, latency and reversal”. However, a comprehensive description of features in the human genome controlling proviral fate remains poorly understood. Elucidating the molecular bases of this mechanism will allow us not only to gain a better knowledge of this poorly understood process but also provide the rationale basis for clinical interventions to cure patients from HIV infection. Here, we use a collection of large-scale datasets datasets to interrogate the “*integration code paradigm*” and we also provide a detailed “compass” to guide future investigations to solve this long-standing biomedical research problem. Systematic analyses of transcripts and regulatory information are essential for the identification of genes and regulatory regions, and are an important resource for the study of human health and disease. Such analyses would provide comprehensive views on the potential roles of integration landscapes in the human genome across cellular contexts and individuals.

The human genome sequence provides the underlying code for the correct transcriptional regulation of most biological processes. Transcriptional regulatory programs are coordinated by the precise spatio-temporal interaction between *cis*-elements present at promoters and enhancers, and sequence-specific factors that recognize them, as well as by the position of genes and regulatory elements in the three-dimensional space and in relation to nuclear territories.

Over the past two decades, great efforts have been made in elucidating how HIV integrates into the human genome and how HIV transcription normally operates. Upon integration, the virus becomes transcriptionally active or silent, leading to active or latent infections, respectively (here referred to as proviral fate) (**Figure 1A**). One of the most exciting breakthroughs in biomedical research was the discovery of anti-retroviral therapy (ART), which curbs active infections to nearly undetectable levels. However, one major problem is that ART does not target latent infections, which can be reactivated in response to immune stimulation leading to viral rebound. Thus, HIV latency has become the center of attention for potential clinical interventions. As such, a large body of research has identified chromatin signatures (epigenetics) on the proviral genome correlating with latency establishment/maintenance, and discovered enzymes as potential therapeutic targets to induce latency reversal for the elimination of latently infected cells. However, these previous studies have used low-resolution approaches to partially interrogate the epigenetic landscape at the proviral promoter on a narrow number of model cell lines (which is far from representing the heterogeneous integration landscape in patients). Therefore, we currently lack coherent models that can broadly recapitulate proviral fate choice on a large-scale and can provide a mechanistic understanding of how the integration landscape shapes proviral fate (**Figure 1A**). This major biomedical challenge demands a comprehensive definition of the molecular rules modulating proviral fate before we can even leverage this knowledge in the clinical setting.

Given that HIV contains its own promoter, one would assume that the virus functions in a cell-autonomous manner (irrespective of the integration site). Nonetheless, work by the Verdin lab in 2001 proposed that the integration site controls basal and immune stimulation-dependent transcription, indicating that HIV possibly operates in a cell state- and integration site-dependent manner. Although HIV integrates into chromatin-accessible sites located near the nuclear pore, integrated proviruses exist at different locations (intergenic or intragenic) and in distinct orientations (sense, divergent or convergent) respective to human genes and regulatory elements (**Figure 1B**). Thus, the integration landscape is highly heterogeneous in sequence and chromatin state and potentially has variable effects on proviral transcription. Despite these great discoveries, it remains unknown what genomic features control proviral transcription and fate. The central hypothesis of this proposal is that the integration site contains “instructions” –here referred to as integration code (one or more genomic features located proximally and/or distally from the integration site)– that shape the organization and activity of proviruses in the context of the human genome. It has been known for years that the human genome provides the underlying code for the correct transcriptional regulation of most biological processes through precise interactions between *trans*-acting factors/non coding-RNAs and *cis*-elements (promoters, enhancers). These regulatory elements could provide “*local*” functions (related to the processes of transcription and chromatin accessibility) or “*distal*” functions (chromatin communication between human and viral genomes) that could influence HIV expression thereby modulating proviral fate (**Figure 1C**). Depending on the integration site, proviruses might function as constitutively active or silent transcriptional unit, and the silent proviruses could be reactivated at different levels by immune stimulation. One interesting scenario could be that the integration event creates new chromatin states or disrupts pre-existing ones thereby impacting cellular and/or viral behaviors. Given the large diversity of the integration landscape as well as the complexity in the regulation of transcription, it is conceivable that a combination of regulatory elements (including the location and orientation of proviruses respective to nearby genes and regulatory elements such as enhancers and mobile elements, chromatin states of the provirus and neighboring genomic domains, and long-range chromatin interactions) could define the integration code.

**RESULTS**

**Expression of HIV Integration Groups Defined Based on Their Position and Orientation in the Human Genome**

To start interrogating the relationship between HIV integration position and transcription activity, we first reanalyzed the B-HIVE dataset by Chen *et al*. in Jurkat CD4 T cells{Chen, 2017 #423;Ciuffi, 2017 #801} to obtain HIV position information (integration site) respective to human coordinates. Chen *et al*. have previously used a binary classification of HIV proviruses (active vs latent){Chen, 2017 #423}. However, we found significant problems with this classification because HIV proviruses show a continuum of expression from low (more silent) to high (more active), which better reflect the level of proviral activities and fates present in patient samples.

Because HIV integrates inside (intragenic) or outside (intergenic) genes, and in the same or opposite orientation, respective to the nearest human gene transcription start site (TSS) (**Figure 2A**), we first defined integration clusters based on their positions and orientations to examine any potential relationships with human genomic features modulating HIV proviral transcription and fate. This analysis gave rise to six different groups: Intergenic same (Group 1, n=127, 8.15%), Intergenic convergent (Group 2, n=71, 4.56%), Intergenic divergent (Group 3, n=77, 4.94%), Intragenic same (Group 4, n=581, 37.29%), Intragenic convergent (Group 5, n=642, 41.21%), and Intragenic overlapping (Group 6, n=60, 3.85%) (**Figures 2A and 2B**). Group 6 is a composite of 3 subgroups depending on the three possible combinations of HIV direction respective to the two overlapping genes [Group 6a: both genes in same direction with HIV in same direction (n=26); Group 6b: gene 1 and gene 2 in opposite directions (n=24); and Group 6c: both genes in same direction with HIV in opposite direction (n=10)] (**Figure 2A**). However, given the limited number of genes in each of the subgroups 6 we treat them as a single group to increase statistical power for all downstream analyses and for reasons explained in each of the separate analyses below. The large number of proviruses detected in the intragenic groups compared to the intergenic groups is consistent with several previous studies highlighting HIV integrations in intronic regions of highly transcribed genes{Schroder, 2002 #301;Ikeda, 2007 #795;Lucic, 2019 #796}.

Having established the HIV integration groups, we then examined the relationship between HIV positions and their transcription activity. For this, we first identified barcodes in the B-HIVE integrated HIV proviruses (DNA barcodes), mapped barcodes to integration sites, and quantified their expression (RNA barcodes/DNA barcodes) (see Methods). After visualizing the expression of each HIV integration site on a per integration group basis using Circos plots, we found that HIV proviruses from each group are detected in every single chromosome with various gene expression levels irrespective of their group (**Figure 2C–H**), indicating that the arrangement of HIV insertions in the position/orientation defined groups above is not a major determinant of HIV expression levels. As such, we reasoned that HIV activity is regulated by local and distal codes that are unique to the integration site and not shared within each integration group. Below we evaluate HIV expression in relation to various regulatory features including its distance to the nearest human gene TSS, expression level, enhancer position/activity as well as nuclear topology.

**Relationship Between HIV Expression and Distance to Nearest Human Gene Promoter**

Given HIV can be located at various distances respective to nearby gene promoters (TSS) (**Figure 3A**), we interrogated if there is any relationship between the expression level of each provirus in the six above defined integration groups and their distance to nearby TSS based on a linear scale (one-dimensional analysis) (**Figures 3B–G**). For this analysis we did not include Group 6 because the overlapping nature of the two genes associated to the HIV insertion compromised the analysis. Interestingly, we found that the expression of proviruses in intergenic regions is poorly correlated with the distance to the nearest TSS irrespective of their orientation to the most proximal TSS [Group 1: Intergenic same (r=-0.13, p=0.14) (**Figure 3B**), Group 2: Intergenic convergent (r=-0.19, p=0.11) (**Figure 3C**) and Group 3: Intergenic divergent (r=-0.16, p=0.15) (**Figure 3D**), potentially indicating that the orientation of intergenic proviruses is not a major feature controlling proviral expression. In addition, we observed a potential pattern for proviruses in the “Intragenic” groups, which seem to be, in general, more correlated than the “Intergenic” clusters, including Group 4: Intragenic same (r=-0.11, p=0.0012) (**Figure 3E**) and Group 5: Intragenic convergent (r=-0.23, p=0.0000000064) (**Figure 3F**). Surprisingly, the expression of proviruses in Group 5: Intragenic convergent was more correlated with the expression of its associated gene than the expression of proviruses in Group 6: Intragenic same, suggesting that the orientation of the HIV proviruses respective to the nearest TSS has a major impact on proviral expression, in potential agreement with theories of transcription interference raised on a few number of CD4 T cell clones{Lenasi, 2008 #345}.

Collectively, while the transcription activity of both “Intragenic” groups is more correlated with the distance to the nearest TSS, these proviruses are typically defective in patients{Maldarelli, 2014 #303;Einkauf, 2019 #660}, potentially due to their location inside genes and the high level of co-transcriptional processing. Surprisingly, the orientation of the “Intergenic” groups seems to be an important feature of HIV expression levels, with the activity of the “Intergenic same” group being more correlated to the distance to the nearest TSS than in proviruses located in opposite orientations (either convergent or divergent). Specifically, we found a positive correlation of expression of proviruses for the intragenic groups, and more significant for the intragenic convergent class (**Figure 3F**) based on proximity to the nearest human gene promoter.The fact that intact HIV proviruses in patients under long-term suppressive therapy are more prevalent in non-genic than in genic positions{Einkauf, 2019 #660}, consistent with our large-scale data analysis from *in vitro* integrations, raises some interesting observations for future studies on the molecular bases of HIV persistence.

**Relationship Between HIV Activity and the Expression of its Associated Human Gene**

Besides the distance of HIV integration to the nearest human gene promoter in the lineal scale, another regulatory feature shaping HIV proviral transcription can be the level of expression of the HIV-associated gene (e.g., the gene in which HIV is integrated into if intragenic or the nearest gene if intergenic). Thus, it is possible that the activity of virus 1 is greater than virus 2 if the gene associated with virus 1 (gene X) is expressed at higher levels than the gene linked to virus 2 (gene Y) (**Figure 4A**). To test this hypothesis, we calculated the expression of proviruses in five out of the six HIV integration groups (Figure 2A) and the expression of the HIV-associated human gene (**Figures 4B–F**). Once again, we excluded Group 6 (“Overlapping”) because of HIV association with two genes. For this analysis, we compared the expression of HIV proviruses in each integration group derived from the number of barcodes obtained from RT-qPCR normalized to the copy of DNA barcodes (log10(RNAmean/DNAmean)) to the expression of HIV-associated human genes extracted from total RNA-seq from Jurkat cells (see Methods). Surprisingly, from the three intergenic groups, we found no apparent correlations for the “Intergenic same” (r=0.09, p=0.29) (**Figure 4B**) and “Intergenic divergent” (r=0.11, p=0.32) (**Figure 4D**). However, the “Intergenic convergent” group showed a slight, but statistical significant correlation (r=0.26, p=0.03) (**Figure 4C**), potentially indicating that this HIV arrangement is somehow benefited from host transcription activity. Again, this observation is rather interesting given that intact HIV proviruses in patients under long-term suppressive therapy were preferably located in non-genic positions{Einkauf, 2019 #660}.

Given that HIV preferably integrates inside genes with no apparent difference in orientation respective to host gene transcription activity, we were particularly interested in testing whether there is any correlation between the expression of HIV in the intragenic sense and convergent groups. Strikingly, we found that the correlation level of the “Intragenic same” group (r=-0.002, p=0.94) (**Figure 4E**) is much lower than the correlation of the “Intragenic convergent” group (r=0.07, p=0.06), which is borderline to statistical significance (**Figure 4F**). These results suggest that the convergent orientation offers HIV a molecular benefit potentially linked to the lack of interference of Pol II molecules transcribing in the same orientation.

Notably, studying a similar cell-based model of post-integration latency, Lenasi *et al*. described that HIV integration in the sense orientation led to transcriptional interference from neighboring human gene promoters caused by the elongating Pol II transcribing through the HIV promoter triggering the physical exclusion of the transcription pre-initiation complex (PIC) on the viral promoter {Lenasi, 2008 #345}. Taken together, the analysis of a large number of clones containing intragenic integration events largely expands these initial observations {Lenasi, 2008 #345}, indicating that transcription interference accounts for at least part of the observed proviral latency in the conditions tested in this cell-based model. However, since the correlation levels are low it is likely than a combination between orientation respective to nearest gene promoters and other regulatory features. Below we also interrogate HIV expression as a function of its position in the nuclear space and in relation to nuclear sub-compartments and chromatin states.

Is it worth adding your analysis of repetitive elements at least for Supplementary?

**Contribution of Human Genome Enhancers to HIV Proviral Transcription and Fate**

HIV displays specific preference to integrate into genes proximal to high density of enhancers, which were recently defined as genomic elements of retroviral integration{Lucic, 2019 #796}. Enhancers 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 and often become deregulated in diseased states{Smith, 2014 #427;Li, 2016 #422;Kim, 2015 #111;Whyte, 2013 #255}. Thus, since enhancers play key roles in shaping cell-type specific responses, and because the relationship between human enhancers and HIV expression has not been previously explored, we interrogated whether the position of the HIV integration site respective to human enhancers is a key regulatory element for determining proviral activity level and consequently its fate. To this end, we generated a rigorous and comprehensive database of active enhancers based on a combination of conventional epigenetic and transcription signatures including: 1) a unique chromatin state demarcated by high H3K27ac, high H3K4me1, and low H3K4me3 derived from ChIP-seq, and 2) symmetrical bi-directional enhancer RNA transcription (eRNA) derived from TT-seq (**Figure 5A**). At least two types of active enhancer classes have been described: typical enhancers (TE) and super enhancers (SE). Whereas TE contain the classic composition of features indicated above, SE are locally grouped clusters of enhancers (defined as H3K27ac, H3K4me3, H3K4me1, and active transcription (TT-seq) domains within 12.5-kb of each other) (**Figure 5A**) driving high levels of transcription of nearby cell-identity genes{Whyte, 2013 #255}.

A Hidden-Markov model of eRNA Watson and Crick strands identified possible TE’s that when overlapped with enhancer histone marks (H3K27ac, H3K4me1, and H3K4me3), identified 2,180 active, intergenic TE (**Figure 5B**,see Materials and Methods for a detailed description). We defrayed from identifying intragenic enhancers given the high content of genic transcription and histone modifications 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 TT-seq (**Figures 5A and 5B**). Of note, we did not use the dbSuper database (Khan and Zhang, 2016) like Lusic et al. did recently{Lucic, 2019 #796} because it has been deprecated.

Chromatin state profiles were evaluated by constructing heat maps for enhancers ranked by decreasing levels of DNase hypersensitive sites (DHS), revealing accessible chromatin at the center of all predicted traditional enhancers as shown by DNase-seq, and chromatin signatures surrounding the nucleosome free region in a ‘peak-valley-peak’ pattern that is consistent with enhancer signatures{Pundhir, 2016 #802} (**Figures 5B and 5C**).

To test the hypothesis that HIV proviral expression is correlated to its proximity to human enhancers (**Figure 5D**), which predicts that proviruses closer to enhancers will be transcribed at higher level than those located at farther distances, we computed the distance of each provirus in the dataset to the closest TE and SE. Then, with the rigorously assembled enhancer and HIV position databases for all HIV integration groups, we measured the distance of each of the six different HIV integration groups to the center of their nearest TE and observed poor correlation and no statistical significance (p<0.05) for all groups, suggesting that proximity to a TE alone is not a good predictor of HIV proviral activity (**Figure 5E**).

Given that genes located near SE are transcribed to much higher levels compared to genes located near TE, if the proximity to enhancers is a true regulator of proviral transcription we would then expect that, compared to less active or latent proviruses, the most active proviruses should be positioned nearer to SE. Therefore, we also evaluated the distance of all six HIV integration groups to the nearest SE and found low correlations and no statistical significance for all four intergenic groups (same, convergent, divergent, and overlapping) (**Figure 5F**). However, interestingly, the two intragenic groups showed a statistical significant correlation: intragenic same (p=0.02) and intragenic convergent (p=0.002), consistent with the better correlations between the expression of “Intragenic same” and “Intragenic convergent” groups and the distance to the nearest TSS, proposing that intragenically positioned HIV may benefit from both its positions nearby to human regulatory elements such as promoters and SEs. Taken together, the preference for HIV to integrate into genes proximal to super enhancers{Lucic, 2019 #796}, may also have the dual benefit of increasing proviral transcriptional levels.

**Contribution of Nuclear Spatial Sub-compartments to HIV Proviral Transcription and Fate**

Within the nucleus of a mammalian cell, the genome within is partitioned into contact domains (A and B) that segregate into six sub-compartments (A1, A2, B1, B2, B3, and B4) and are associated with distinct patterns of histone marks {Rao, 2014 #804} (**Figure 6A**). AB and B-type sub-compartments correspond to known types of silent chromatin marks. AB, B1, B2 and B3 have very different properties. Sub-compartment B1 correlates positively with H3K27me3 and negatively with H3K36me3, suggestive of facultative heterochromatin. Sub-compartments B2 and B3 tend to lack all of the above-noted marks, includes 62% of pericentromeric heterochromatin and is enriched at the nuclear lamina and at nucleolar-associated domains. Sub-compartment B3 is enriched at the nuclear lamina (1.6-fold), but strongly depleted at nucleolar-associated domains. Sub-compartment B4 is mostly on chr19 and comprises a handful of regions, each of which contains many KRAB-ZNF superfamily genes (130 of the 278 KRAB-ZNF genes in the genome, a 65-fold enrichment). The two A-type sub-compartments are enriched in euchromatin marks, with higher coverage in A1 than in A2.

To evaluate the distribution of HIV proviruses respective to these nuclear sub-compartments we first re-analyzed a Hi-C dataset in Jurkat T cells{Lucic, 2019 #796} and found a roughly equal distribution of HIV proviruses into the A1, A2, B1, and B2 compartments and preferential enrichment in the B3 sub-compartment, and as expected a low percentage of proviruses (0.35%) in the B4 compartment given its lower representation in the genome (**Figure 6B**). Interestingly, the data revealed the known intrinsic HIV integration preference for the A1 >> A2 (relative to B).

I made these notes in our discussions: define nomenclature of TADs, define nomenclature of CTCF loops (domain boundaries conserved across cell-types{Weintraub et al Cell 2017}), score TADs (AB or ABC), discuss what is inside TADs and CTCF loops (e.g., chromatin states, gene content, geometric mean / logistic regression).

Conclusion: Proviruses in A1 subcompartment are expressed at higher levels than proviruses in other compartments (**Figure 6C**)

* Provide coverage for the different lamin compartments (**Figure 6B**)
* How do we quality control the definition of the different compartments?

Provide conclusion for HIV expression by lamin subcompartment (**Figure 6C**).

Provide conclusion for HIV expression by position respective to TADs and CTCF loops (**Figure 6E**).

* What is the raw map of contacts (size) 10kb?

**Contribution of Chromatin States to HIV Proviral Transcription and Fate**

Since chromatin marks do not denote a singular function, and we want to annotate regions of the genome, we used a 15-state Hidden Markov Model (chromHMM; {Ernst, 2017 #805}) of 7 histone marks (H3K27me3, H3K4me3, H3K27ac, H3K4me1, H3K36me3, H3K79me3, and H3K9me3) to annotate chromatin states similar to the 15 state annotations from Roadmap Epigenomics Project (Ernst et al., 2012). This approach supersedes recent classifications of “chromatin states” calling in CD4+ T cells using Encode datasets (Ref Verdin eLife). After overlapping the states with 3 epigenomes (E115, E116, E123, see supplementary materials) and comparing each histone mark function, we rearranged the state number and relabel each state to correspond to identical labels from Roadmap Epigenomics Project.

We then evaluated the likelihood that HIV would insert into each of the 15 epigenetic states, and then by using the BHIVE data we used a hypergeometric test to predict the probability that the insertions are by chance or whether HIV is more or less likely to insert into that particular state. HIV was more likely to insert into states 2-8, 11-12, and 14 and less likely to insert into state 13 and 15. This suggest that HIV is more likely to insert in and around active genes. (Add a sentence here to relate it to a past paper). We also wanted to compare the expression of HIV in each state to determine if the median between states is different. HIV expression was not found to be normally distributed or have similar variances between states, so a Kruskal-Wallis rank sum test computed significant difference of the median between different states. Surprisingly HIV expression seems to be higher in State 7, Active enhancer, than States 4 or 5 (Strong transcription and Weak transcription, respectively)(Figures 7B and 7C). This may be due to or explained by ???? (4 and 5 have many more outliers towards the low expression). 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) (Figures 7B and 7C). Since we found almost no correlation between the distance to the nearest TE or SE (Figures 5E and 5F), we could conclude that our list of TE and SE are incomplete and that by being near a \*possible\* enhancer has more of an effect on HIV expression. It’s also possible that HIV is able to recruit pioneer transcription factors in order to open a genomic region and begin transcription. This may help explain why state 15 (Quiescent/Low), although has statistically less insertions , still has insertions with high expression. These outliers may have more to do with their position in the genome’s 3D structure and availability of pioneer transcription factor marks to be recruited.

**Machine Learning Approach to Train a Model Predicting HIV Expression Levels**

Since a single feature (e.g. histone marks, gene transcription, and distance to enhancers) did not statistically correlate with HIV expression, a machine learning model approach was used to train a linear regression model for HIV expression prediction. Each feature’s information value (IV) related to its importance in HIV fate prediction was calculated using *smbinning* R package (**Supplementary Table 1**). When a threshold of IV ≥ 2 was applied, 26 features were determined as optimal features for the prediction task, which include all 20 Lamin bins, 5 up-stream bins in H3K27ac (H3K27ac\_-600\_-400; H3K27ac\_-800\_-600; H3K27ac\_-1600\_-1400; H3K27ac\_-1800\_-1600; H3K27ac\_-2000\_-1800), and 1 up-stream bin in MNase (MNase\_-1200\_-1000). Through a logistic regression model training with the optimal features, their estimated weights and corresponding odds ratios, and the standard errors of the estimated weights were obtained (**Supplementary Table 2**). In addition, the model shows that, with the HIV expression level, 3 H3K27ac bins (H3K27ac\_-1600\_-1400; H3K27ac\_-1800\_-1600; H3K27ac\_-2000\_-1800) and 1 Lamin-B2 bin (lamin\_0\_200B2) have significant relationships (Pr(>|z|) < 0.01) and a mild relationship (Pr(>|z|) < 0.1) respectively while the others have no significant relationships. For the test dataset, HIV expression levels were predicted through the logistic regression model. First, we focused on HIV insertions of ‘High’ or ‘Low’ expression only so to evaluate the model as a binary classifier. The evaluation metrics were calculated as 68.42% of sensitivity, 59.10% of specificity and 64.71% of AUROC, confirming that the model has decent prediction power (**Figure 8B**). Next, we compared predicted HIV expression values among the three expression categories with t-test. The predicted HIV expression values for the ‘High’ expression category were found significantly higher than those for the ‘Low’ (p-val=0.00025) category. Meanwhile, the predicted values for the ‘Intermediate’ and ‘Low’ categories showed not significant difference (p-val=0.27) (**Figure 8C**). Furthermore, the actual and predicted HIV expression values were found to have a positive, moderate correlation (*R*=0.19, *p*=0.00018) (**Figure 8D**).

**Combination of In Vitro HIV Integration-Expression Data with Primary and Patient Datasets**

Our analysis of HIV integration-expression in the immortalized CD4 T cell system revealed: i) that the larger HIV integration groups are intragenic (same and convergent) but location respective to genes and transcription directionality does not seem to influence transcriptional output (**Figure 2**), ii) that there is correlation between HIV expression and distance to proximal TSS for intragenic (same and convergent) (**Figure 3**), and iii) that there is “some” influence of proximity to human gene TSS, no apparent evidence of transcription interference (TI), and intergenic same group shows increased correlation compared with intergenic (convergent and divergent) so orientation of intergenic HIV respective to proximal gene TSS may matter to dictate influence on HIV (if on the same direction). Furthermore, when analyzing HIV genome expression to host gene expression for the different clusters we found that “Intergenic convergent” is the best correlated (0.26, p=0.03) to proximal gene expression (**Figure 4**).

There is strong selection of intact proviruses with features of deeper viral latency during prolonged ART. Thus, our previous results were interesting given the enrichment of intact HIV proviruses compared to defective ones, for non-genic chromosomal positions and more frequently showed opposite orientation (divergent and convergent) relative to host genes, and intact HIV proviruses were preferentially integrated in either relative proximity to or increased distance from active transcriptional start sites and to accessible chromatin regions {Einkauf, 2019 #660}.

In lieu of the previous analysis in the immortalized model, the next most challenging goal was to translate the *in vitro* discoveries to patient outcomes thereby broadening our discoveries especially because the disease cannot be fully modeled in tissue culture given problems with immortalized models of latency and genetic heterogeneity of proviral integration landscapes and sequences *in vivo*. For this, we first assembled HIV integration site datasets from both primary CD4 and patient samples (**Figure 9A**). We particularly focused on defining recurrent integration sites (RIGs) and intact proviral genomes to better provide biologically relevance, then define overlay between the *in vitro* dataset and the primary/patient datasets.

Some questions that I had

1. What is the relationship between Recurrent Integration Genes (RIGs) and intact proviruses?
2. Are the BHIVE integrants enriched in RIGs? Yes, Lucic did this analysis in Supp Figure 3.
3. Does HIV expression correlate with the activity of its neighborhood?
4. Does HIV activity correlate with its directionality respective to most proximal genome regulatory elements (e.g., promoters, enhancers)?
5. Is there any correlation between HIV activity and human chromatin states?
6. 3D/topological analysis
7. Does HIV activity correlate with its position respective to spatial nuclear compartments (TADs and LADs)?

**DISCUSSION**

Our analysis suggests that the proximity of HIV to human enhancers is not a good predictor of its expression. This assertion contradicts a major but erroneous assertion from the Chen *et al*. study that “the expression of HIV is strongest close to endogenous enhancers9”. Whereas enhancers could theoretically still contribute to position effects on HIV proviral fate, we provide compelling evidence that distance to the nearest enhancer alone is not predictive of transcriptional status. Thus, it is possible that other regulatory features (i.e., chromatin landscape at both the provirus and neighboring domains, proximal and distal effects, as well as HIV provirus positioning in the 3D/4D nuclear space) could contribute to proviral transcriptional rates in basal and stimulated conditions.

Deciphering human genome codes shaping HIV proviral fate will certainly require painstaking investigations of multiple genetic and epigenetic features. To avoid incorrect assumptions that one or more regulatory features could contribute to position effects, deep and precise interrogations will have to be conducted in the same system, which will require the generation of thousands of clones containing single proviruses integrated into a unique position in the human genome and the use of single-cell level approaches to simultaneously characterize regulatory features of both HIV and human genomes. Finally, our analysis demonstrates the importance of carefully and correctly characterizing enhancer elements when studying their potential role in genome regulation. Future studies should be cautious to employ the same rigorous standards when defining genomic domains to interrogate functional insights in human health and disease.

Einkauf et al. concluded that intact HIV sequences in non-genic positions (divergent and convergent) are more prevalent than in genic positions, and/or close to TSS and/or to chromatin accessible areas. Notably, both of these conclusions are in agreement with our predictive ML model.

However, the analysis of **Figure 4** has several caveats. First, it assumes that HIV integration does not apparently influences expression of the host gene where it is integrated (because host gene expression data derives from uninfected cells). Second, it assumes that the sequencing reads of the host gene where HIV is integrated only derive from the associated gene (and not from intragenic transcriptional units or enhancers). This is a problem because eukaryotic genes usually contain several coding units within such as micro RNAs and snRNAs, as well as non-coding intragenic enhancers. Thus, the conclusion of this analysis should be taken with caution. Third, given that read-through transcripts deriving from the activity of upstream human gene promoters have been previously documented{Bullen, 2014 #602}, the approach used to measure HIV RNAs cannot precisely decipher if the transcripts arise from the HIV promoter or upstream cellular promoters. This is important to note because read-through transcripts will lead to non-functional viral products and may not account for proviral fate effects. Fourth, another caveat of this previous analysis is that both HIV and host gene expression were measured using different approaches and not the same system (RNA-seq for host gene expression in uninfected Jurkat cells and the pipeline used for HIV expression analysis). Taken together, future studies where HIV and host gene expression are measured in the same system (either at the population or single-cell level) are required to more accurately answer this question.

**Discussion ML section:** Through a ML approach, optimal genetic features for HIV expression prediction, which also may have biological implications in HIV fates, were identified. A logistic regression model trained with them was shown to have decent prediction power. Since our model was built on a limited number of HIV insertion observations (n=1559), as the HIV genetic dataset grows, we expect that genetic landscape related to the HIV fate will be more prominent and, accordingly, our machine learning approach can bring in a better model. To the best of our knowledge, it is the first report on ML model’s success in HIV expression prediction based on genetic landscape.

Analysis 1. Comparison of BHIVE and Einkauf datasets

1. Determine the integration groups (1-6) for the “intact” and “defective” proviruses from Einkauf by applying the same pipeline as in Figure 2.
2. Einkauf also has ATAC-seq and RNA-seq so we can do further analysis to define specific features enriched in the intact vs defective proviruses (e.g., distance to TSS, distance to accessible genomic regions).
3. Determine the epigenomic landscape of intact vs defective proviruses using CD4 T cell datasets generated by ENCODE.
4. Define if the BHIVE dataset contains examples that match intact proviruses in the Einkauf dataset. For this, we can overlay integration sites in BHIVE and Einkauf to find the BHIVE proviruses matching the intact and defective proviruses in patient samples.
   * 1. For intragenic, we can call same integration site if in the same gene (irrespective of orientation?)
     2. For intergenic, we can call same integration site if in the vicinity (<10 or 20 kb from TSS or TTS from the HIV-associated gene).
5. Identify RIGs in the BHIVE and Einkauf datasets (Lusic et al did this in their Supplementary Figure 3) and then test if RIGs correlate with intact and/or defective proviruses in the Einkauf dataset. HIV recurrently integrates into similar genome regions/genes, Lucic defined the term “Recurrent Integrated Genes (RIGs)” as genes with ≥1 HIV-1 integrations in at least 2 out of 8 datasets analyzed in the Lusic et al work yielding a total of 1648 RIGs (Supplementary Fig. 1c).

Analysis 2. Comparison of BHIVE and 8 Lusic datasets

1. Determine the integration groups (1-6) for the proviruses from Lusic (Supplementary Figure 3).
2. Determine the overlay of integration sites and RIGs between BHIVE, Lusic and Einkauf to reinforce the claim that these are proviruses (potentially intact) found in patients.
3. Compare epigenomic features of “intact” and “defective” proviruses with those from the BHIVE analysis to provide fundamental principles of differential transcriptional regulation of these two classes of proviruses. There are a number of analyses and observations we can make. For example, defective proviruses (those that will not produce infection) are typically located intragenically compared with the intact proviruses, which appear to be preferentially located to intergenic regions. I need to think about this a bit more.
4. Note for discussion: relate this chromHMM conclusion to the idea that most intragenic proviruses are defective.

**MATERIALS AND METHODS**

**URLs**: The computer code used for this analysis is available on GitHub: *provide link*

**Barcodes and HIV Integration Site Mapping on the Human Genome and HIV barcode clustering and quantification**

BHIVE data was 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).

The BHIVE expression data set was subdivided into 6 different groups, 1) Intergenic – Same, 2) Intergenic – Convergent, 3) Intergenic – Divergen, 4) Intragenic – Same, 5) Intragenic – Convergent, 6) Intragenic – Overlapping (which consists of 3 groups, Figure 2A) depending on the relationship to the nearest gene from Gencode version 25. For each group, a Circos plot version 0.696 (Krzywinski et al., 2009) was created to show the relationship of expression to genome location (Figure2B – 2H). HIV expression versus distance to nearest Transcription Start Site (TSS) was plotted in R version R/3.3.2-gccmkl (R Core Team, 2016) using ggplot2 (Wickham, 2016).

**Traditional and Super enhancer Databases**

TTseq contains information on transient transcription which we used to identify possible enhancers and super enhancers. TT-seq also allows the transcription to be identified as the forward or reverse strand. For this purpose, enhancers are defined as regions of the genome that are bi-directionally transcribed, and not in a gene or its promoter or an annotated linc-RNA. A 4 state hidden markov model on TT-seq all data (not separated by strand), 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 untranscribed. A database of possible enhancers was created by identifing regions of the genome that were identified as transcribed for all data, and also overlapped with regions that were forward and reverse transcribed. Also added to the possible enhancer list were regions where there were overalpping forward and reverse transcription, but not identified as transcribed in all data. From this list, protein coding genes with 2 kb upstream and downstream were removed. Next, annotated regions from RNAseq with an fpkm greater than 1 for all replicates and not protein-coding were removed. This final list of 20,943 regions are purported enhancers.

Super enhancers were identified using Rose v0.1 (Warren et al., 2013 and Loven et al., 2013), stitching together a 12,500 bp distatnce, exluding 2500 bp from TSS. Purported enhancer regions from TT-seq were used as previouly 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. Since not all enhancers are identifiable with the 3 histone marks, TT-seq bam files were also used to identify super enhancers, and the output was filtered again for transcribed, annotated regions of the genome (E.G. genes and linc-RNAs). The final super enhancer database contained 767 merged regions, 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 super enhancer regions were removed from the 20,943 purported enhancer regions leaving 18,357 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.

The closest enhancers and super enhancers to each of the 6 group HIV insertions were identified with bedtools closest (version 2.26.0; Quinlan and Hall, 2010). HIV expression versus distance to nearest enhancer or super enhancer was plotted in R version R/3.3.2-gccmkl (R Core Team, 2016) using ggplot2 (Wickham, 2016).

**Data availability**

Dataset listed in Table 1 were downloaded from NCBI Gene Expression Omnibus.

**NGS datasets uses in this study – ChiP-seq**

The Nextflow (Tommaso et al., 2017) BICF ChiP-seq Analysis Workflow version 1.0.0 (Barnes et al., 2019) processed all ChiP-seq files, merging separate experiments as technical replicates. Briefly, reads were trimmed with trimgalore version 0.4.1 (Martin, 2011) (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 (Li and Durbin, 2009), sorted and indexed with SAMtools version 1.3 (Li et al., 2009)(-F 1804 -q 30), and duplicates removed with Sambamba version 0.6.6 (Tarasov et al., 2015) (standard options). Bam files were converted to tagAlign with bedtools version 2.26.0 bamtobed (Quinlan and Hall, 2010), after which, samples were checked for quality control using deeptools version 2.5.0.1 multiBamSummary, plotCorrelation, plotCoverage, and plotFingerprint (all standard protocols) (Ramírez et al., 2016) and cross correlation analysis with phantompeakqualtools version 1.2 (Kharchenko et al., 2008 and Landt et al., 2012). Peaks were called with MACS2 version 2.1.0-20151222 (Zhang et al., 2008), using the predominant fragment length from the cross correlation analysis as --extsize (other parameters: -p 1e-2 --nomodel --shift 0 --keep-dup all -B --SPMR). Consensus peaks were called (bedtools version 2.26.0; Quinlan and Hall, 2010) and annotated (library ChIPseeker in R; Yu et al., 2015 and R Core Team, 2014) if at least 2 replicates or pseudoreplicates contained a peak.

**NGS datasets uses in this study – RNA-seq**

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 (Martin, 2011). Quality-filtered reads were then aligned to the human reference genome (GRCh38) using the HISAT version 2.0.1 (Pertea et al., 2016) aligner using default settings and marked duplicates using Sambamba version 0.6.6 (Tarasov et al., 2015). Aligned reads were quantified to coding sequences of known transcripts using ‘featurecount’ version 1.4.6 (Liao et al., 2014) per gene ID against Gencode version 25. HIV expression versus log 10 Fragments Per Kilobase Million (fpkm) of the nearest gene was plotted in R version R/3.3.2-gccmkl (R Core Team, 2016) using ggplot2 (Wickham, 2016).

**NGS datasets uses in this study – TT-seq**

Files were processed 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 35 bp after trimming were removed from further analysis using trimgalore version 0.4.1 (Martin, 2011). Quality-filtered reads were then aligned to the human reference genome (GRCh38) using the HISAT version 2.0.1 (Pertea et al., 2016) aligner using default settings and marked duplicates using Sambamba version 0.6.6 (Tarasov et al., 2015). Aligned reads were quantified to the entire annotated transcript region using ‘featurecount’ version 1.4.6 (Liao et al., 2014) per gene ID against Gencode version 25.

**NGS datasets uses in this study – MNase-seq**

Files were processed with a modified Nextflow (Tommaso et al., 2017), BICF ChiP-seq Analysis Workflow version 1.0.0 (Barnes et al., 2019). Briefly, we used trimgalore version 0.4.1 (Martin, 2011) on the raw reads to remove reads shorter than 35bp or with phred quality scores less than 20 and then aligned trimmed reads to the human reference genome (GRCh38) using default parameters in BWA sampe version 0.7.12 (Li and Durbin, 2009). The aligned reads were subsequently filtered for quality and uniquely mappable reads were retained for further analysis using SAMtools version 1.3 (Li et al., 2009) and Sambamba version 0.6.6 (Tarasov et al., 2015), and bedtools version 2.26.0 (Quinlan and Hall, 2010) bamtobed converted the bed file to tagalign. Peaks were called with iNPS version 1.2.2 (Chen et al., 2014) and filtered for a -log10(Pvalue\_of\_peak) of less than 0.05.

**NGS datasets uses in this study – Dnase-seq**

Files were processed with a modified Nextflow (Tommaso et al., 2017), BICF ChiP-seq Analysis Workflow version 1.0.0 (Barnes et al., 2019). Briefly, we used trimgalore version 0.4.1 (Martin, 2011) on the raw reads to remove reads shorter than 35bp or with phred quality scores less than 20 and then aligned trimmed reads to the human reference genome (GRCh38) using default parameters in BWA samse version 0.7.12 (Li and Durbin, 2009). The aligned reads were subsequently filtered for quality and uniquely, mappable reads were retained for further analysis using SAMtools version 1.3 (Li et al., 2009) and Sambamba version 0.6.6 (Tarasov et al., 2015). Relaxed peaks were called using MACS version 2.1.0 (Zhang et al., 2008) 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.

**NGS datasets uses in this study – Hi-C**

Reads were pooled per library an ran through the standard Hi-C pipeline using homer version 4.10.4 (Heinz et al., 2010). Briefly, reads were trimmed with homerTools trim -3 GATC -mis 0 -matchStart 20 -min 20, mapped to human reference genome (GRCh38, canonical) with bowtie2 version 2.2.8 (Langmead and Salzberg, 2012), and converted to tag directory (makeTagDirectory -genome hg38 -checkGC -restrictonSite GATC). Matrices are normalized with analyzeHiC (standard protocol), and tads/loops identified (findTADsAndLoops.pl -res 3000 -window 15000 -genome hg38 -p badRegions.bed). The USCS gap and genomicSuperDups were combined to make badRegions file. Tads and Loops from the 2 libraries were combined with merge2Dbed.pl

**Identification of chromatin states – chromHMM**

We implemented ChromHMM (Ernst et al. 2012), which uses a multivariate Hidden Markov Model (HMM) to calculate the probabilistic nature of a multi-state model and the biological nature of the state of chromatin at that location, in order to discover chromatin states in Jurkat 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 (Ernst et al., 2012). 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>) on the basis of the observed data for the above 7 histone modifications. We used a core 15-state model for our analyses since it captured all the key interactions between the chromatin marks, and because larger numbers of states (e.g., “expanded 18-state model”) did not apparently capture sufficiently distinct interactions. To *de novo* generate the core 15-state model in Jurkat cells we compared the relative abundance of the state emissions in Jurkat cells with known chromatin states for the 3 ENCODE cell lines most genetically and phenotypically linked with Jurkats (E115: Dnd41 T cell leukaemia, E116: GM1282878 lymphoblastoid, E123: K562 T cell leukaemia) (as shown in Figure X). In order 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 buildt in refseq annotations of: 1) CpG islands, 2) genes, 3)exons, 4)introns, 5) transcription start sites (TSS), 2-Kb windows around TSS (TSS flanking), transcription termination sites (TTS), and 2-Kb windows around TTS (TTS flanking) based on the GENCODE v27 annotation, 3) Zinc finger genes (ZNF genes) obtained from chromHMM, and 4) typical enhancers and super enhancers obtained as described in the “Enhancer” section above.

**Machine learning section**

In order to study the immediate landscape surround HIV insertion (1559 insertions in total) and its possible effects on expression, we looked at 2 kb region, in 200 bp increments around the sites. 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 subcompartment states were also noted for each 200 bp region. The chromHMM states were 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 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).To determine optimal features, which have predictive power in HIV fate prediction, and train a prediction model with them, a Machine Learning (ML) approach was taken. As a first step, the genetic landscape dataset was randomly split into a training dataset (75% of HIV insertions) and a test dataset (25% of HIV insertions). To select optimal features for HIV expression level prediction, an R package, *smbinning* (https://rdrr.io/cran/smbinning/), was applied to the training dataset consisting of ‘High’ and ‘Low’ expression instances only. It returned each feature’s information value (IV), which is relevant to its importance in the prediction task, and features of IV ≥ 2 were determined as optimal ones. After training a logistic regression model with the training dataset of the optimal features, the trained model was evaluated with the unseen test dataset. Note that ‘Intermediate’ expression instances were excluded for the model training to get a better model but included for the model evaluation.

**ACKNOWLEDGEMENTS**

We thank X and Y for critical reading of the manuscript. This research was supported in part by US National Institutes of Health grant R01AI114362 to ID. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

**AUTHOR CONTRIBUTIONS**

X developed the idea.

X designed the experiments.

X analyzed the data.

X wrote the manuscript.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

**FIGURE LEGENDS**

**Figure 1. Integrated genomics approach to elucidate position effects contributing to HIV proviral transcription and fate**

A) Introduction of HIV latency problem.

B) Introduction of proviral fate.

C) Introduce position effect hypothesis.

D) Introduce regulatory features as paradigm of regulators of position effects.

E) Introduce integrated genomics approach used to study the position effect hypothesis.

**Figure 2. Defining the expression of HIV integration groups based on their position and orientation respective to human genes**

A)Reintroduce that HIV can be found in different orientations and positions giving rise to at least 6 different integration groups.

B) Introduce the abundance of each group in the BHIVE model. Groups 4 and 5 are over-represented.

C) Distribution of each HIV integration cluster and their expression by chromosome.

D) The log10 scale for the Circos plots is -4 to +3.

**Figure 3. Relationship between the expressions of each HIV group to nearest TSS**

A) Hypothesis that activity of HIV group 1 is greater that group 2 because the distance from group 1 integration site to the nearest TSS is smaller than the distance from group 2 integration site to the nearest TSS.

B) Calculations of HIV expression (log10) as a function of distance to TSS (log10) for Group 1.

C) Calculations of HIV expression (log10) as a function of distance to TSS (log10) for Group 2.

D) Calculations of HIV expression (log10) as a function of distance to TSS (log10) for Group 3.

E) Calculations of HIV expression (log10) as a function of distance to TSS (log10) for Group 4.

F) Calculations of HIV expression (log10) as a function of distance to TSS (log10) for Group 5.

**Figure 4. Relationship between the expressions of each HIV group to human gene expression**

A) Hypothesis that activity of HIV group 1 is greater than the activity of group 2 because the expression of the gene in which HIV group 1 is integrated into or nearby to is greater than the expression of the gene in which HIV group 2 is integrated into or nearby to.

B) Calculations of HIV expression (log10) as a function of host gene expression (log10) for Group 1.

C) Calculations of HIV expression (log10) as a function of host gene expression (log10) for Group 2.

D) Calculations of HIV expression (log10) as a function of host gene expression (log10) for Group 3.

E) Calculations of HIV expression (log10) as a function of host gene expression (log10) for Group 4.

F) Calculations of HIV expression (log10) as a function of host gene expression (log10) for Group 5.

**Figure 5. Contributions of human genome enhancers to HIV proviral transcription and fate**

A) Features of promoters and enhancers (traditional and super). Promoters and enhancers are characterized by thepresence of DNase I hypersensitive sites (DHS) and H3K27ac-modified chromatin.However, compared to promoters, enhancers contain greater levels of H3K4me1-modifiedchromatin, and lower levels of H3K4me3-modified chromatin. Promoters, traditionalenhancers, and super enhancers support bidirectional transcriptional activity, but enhancersproduce a class of bidirectional and symmetric non-coding RNAs (eRNAs), whereaspromoters support asymmetric transcriptional activity (i.e., higher transcription levels areobserved in the coding strand).

B) Heatmaps of DHS, various epigenetic markers, andnascent transcriptional activity for traditional enhancers and super enhancers in Jurkat CD4T cells. Each row represents an enhancer and the genomic position respective to theenhancer center is shown. Color density reflects the magnitude of the respective signal ateach genomic position.

C) Genome browser track view of a representative traditionalenhancer and super enhancer.

D) Hypothesis…..

E) Correlation scatter plots showing log10 distance to nearest typical enhancer (TE) of active and latent proviruses. Each point represents a provirus(N = number). The red horizontal line indicates the median distance of thepopulation.

(F) Correlation scatter plots of HIV expression (log2 RNA/DNA ratio) of active proviruses and distance to nearest enhancer or super enhancer (log10 distance). Each point represents a unique provirus (N = number). The blue line indicates the best linear fit and the gray regions the 95% CI.

**Figure 6. Expression of HIV integration clusters in relation to three-dimenstional architecture**

**Figure 7. Expression of HIV integration clusters in relation to chromatin states**

* Need final overlapping plots (E115, E116, E123) for Supplemental.
* Need heatmap plot depicting “HIV expression” + statistical analysis for panel A.

**Figure 8. HIV expression level prediction based on surrounding genetic landscape.** (A) 7 histone marker and 4 other types of sequences and chromatin-status, secondary information calculated using them, were 200-bp binned and taken as an input matrix, ***X***, and HIV expression values as an output vector, ***Y***, for our ML approach. (B) The trained model’s AUROC (Area Under the Receiver Operating Characteristics) curve. (C) Predicted HIV expression comparison among three categories of ‘High’, ‘Intermediate’, and ‘Low’. (D) Linear regression and correlation between the actual and predicted HIV expression values.

**Figure 9. Patient data analysis**

**REFERENCES**