Physical Principles of HIV Integration in the Human Genome

D. Michieletto[1], M. Lusic[2], D. Marenduzzo[1], E. Orlandini[3]
[1] *SUPA, School of Physics and Astronomy, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, UK.*
[2] *Department of Infectious Diseases, Integrative Virology, Heidelberg University Hospital and German Center for Infection Research, Im Neuenheimer Feld 344, 69120 Heidelberg, Germany.*
[3] *Dipartimento di Fisica e Astronomia and Sezione INFN, Universitá di Padova, Via Marzolo 8, Padova 35131, Italy.*

The Human Immunodeficiency Virus (HIV) integrates its genetic material in a small and non-random fraction of all possible integration sites along the host genome. The mechanisms driving this non-uniform selection are not known. To address this issue, we formulate a generic physical model where we treat the viral incorporation into DNA and chromosomes as a stochastic quasi-equilibrium process. Our model rationalises a number of long-standing unexplained observations: we show that HIV integration is favoured in nucleosomal rather than naked DNA, and in flexible over rigid DNA. We find that these biases arise due to the difference in the bending energy barrier associated with DNA insertions. By coupling our model to a well-established framework for large-scale 3D genome organisation we discover that the non-random integration of HIV in human chromatin may be explained as due to large-scale accessibility of interphase chromosomes. Finally we propose and solve a reaction-diffusion model that recapitulates the distribution of HIV hot-spots within the nucleus of human T-cells. With few generic assumptions, our model can explain much of the current evidence on HIV integration and may be used to predict integration patterns in human cells.

Retroviruses are pathogens which infect organisms by inserting their DNA within the genome of the host. Once integrated, they exploit the transcription machinery already in place to proliferate and propagate themselves into other cells or organisms [1–4]. This unique strategy ingrains the viral DNA in the host cell ensuring its transmission to the daughter cells. This is the reason why about $5-10\%$ of the human genome is made up by retroviral nucleotide repeats: these are the remnants of viral DNA which integrated long ago and mutated in such a way that it is no longer able to replicate itself [2, 3, 5, 6]. Whilst many retroviruses clearly pose a danger to health, they are also potentially appealing for clinical medicine, as they can be used as vectors for gene therapies [1, 7, 8].

Experiments have provided a wealth of important observations on the mechanisms through which retroviruses work. First, classical experiments have shown that the retroviral integration complex (or "intasome") displays a marked tendency to target bent DNA regions and in particular those where DNA wraps around histone octamers, i.e. chromatin, rather than naked DNA [9–17]. This is clearly advantageous for retroviruses which target eukaryotes, since their DNA is extensively packaged into chromatin [2, 18]. Second, more recent experiments strongly suggest that the integration sites displayed by most classes of retroviruses are correlated with the underlying chromatin state [19]. For instance, gammaretroviruses, deltaretroviruses and lentiviruses – including HIV – display a strong preference to insert their DNA into transcriptionally active chromatin [12, 19, 20]. Importantly, the preference for transcriptionally active regions remains significantly non-random even after knock-out of known tethering factors such as LEDGF/p75 [12, 19, 21–24].

In stark contrast with the abundance of experimental studies on retroviral and HIV integration in DNA and chromatin, there remains to date a distinct and surprising lack of models that investigate the generic biophysical mechanisms of integration into host genomes. Because many aspects of this problem appear to be shared across several classes of retroviruses, such an approach may provide a useful complement to experiments and can shed light into their universal behaviour.

Here, we propose a generic biophysical model for retroviral integration in host cells, focussing on the case of HIV. We first introduce and study a framework in which retroviral DNA and host genomes are modelled as semi-flexible polymers, and integration events are accounted for by performing local stochastic recombination moves between 3D-proximal polymer segments. Then, at larger scales, we formulate and solve a reaction-diffusion problem to study HIV integration within the nuclear environment of human cells.

At all scales considered, ranging from that of nucleosomes (< 100 nm) to that of the cell nucleus ($> 10\,\mu$m), our model compares remarkably well with experiments, both qualitatively and quantitatively. In light of this, we argue that our framework can provide new mechanistic insight into the non-random selection of HIV integration sites into DNA and chromatin. Indeed it suggests that simple physical features, such as DNA elasticity and large-scale chromosome folding, may be sufficient to explain most of the existing experimental data on HIV integration patterns. In particular, our results rationalise the currently poorly understood correlation between integration probability and underlying epigenetic state [19]. Finally, our model yields several predictions which can be tested in future experiments.

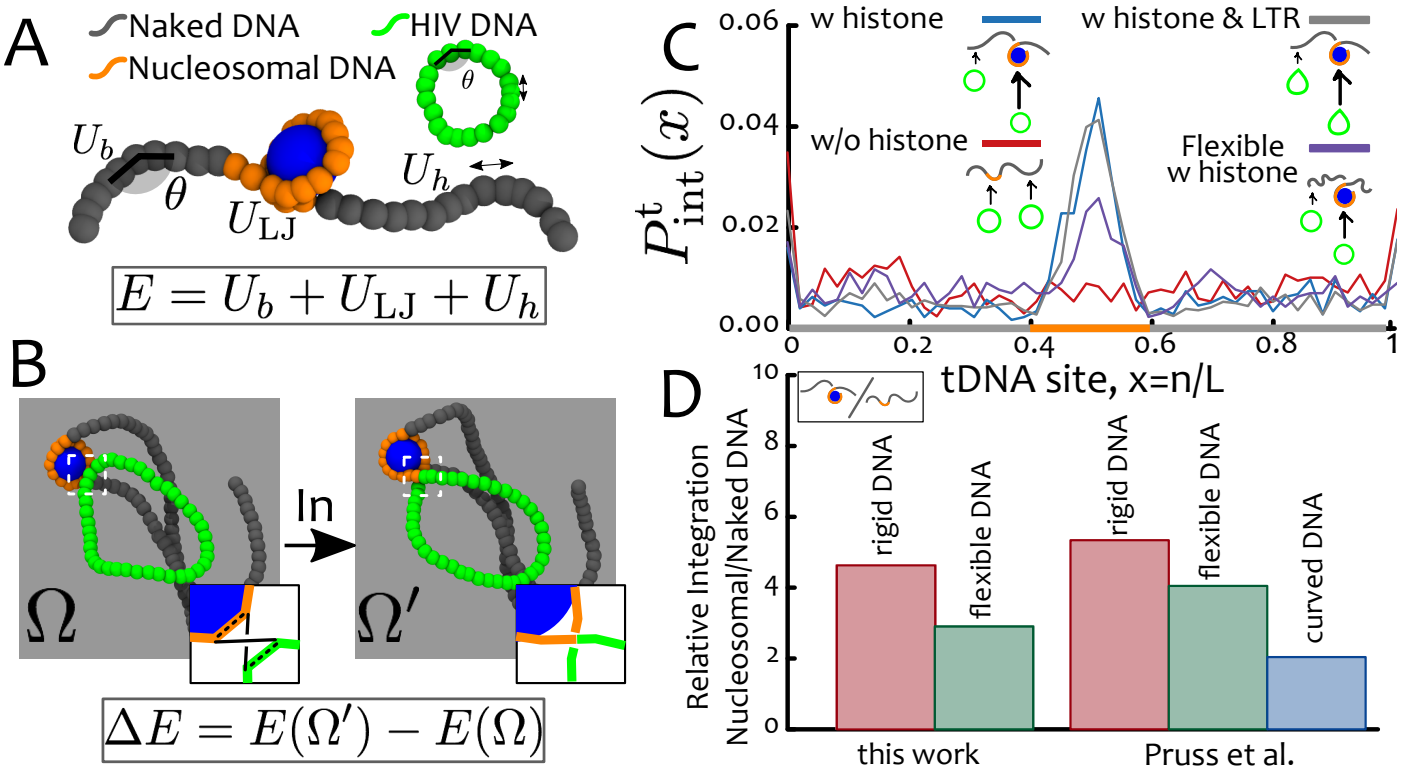

FIG. 1. **DNA elasticity biases HIV integration within mono-nucleosomes.** **A** Model for tDNA and vDNA as diffusing bead-spring polymers with bending rigidity. The potentials associated with bending ($U_b(\theta)$), steric/attractive interactions (U_{LJ}) and stretching of the bonds (U_h) contribute to the energy E of a given configuration. **B** Our quasi-equilibrium stochastic integration takes into account the energy of the old configuration Ω (before integration) and of the new one Ω' (after integration) to determine an integration probability $p = \min\{w, 1\}$, with $w = \exp(-\Delta E/k_BT)$, with which the integration move is accepted. **C** The integration probability $P_{\rm int}(x)$ as a function of the relative tDNA site, $x = n/L$, displays a $\sim$ 4-fold enhancement in the region wrapped around the histone-like protein. The same behaviour is observed when a kinked site (corresponding to the intasome flanked by LTR) is included in the vDNA. Considering flexible tDNA ($l_p = 30$ nm) weakens this preference. **D** Direct quantitative comparison with data from Ref. [9]. Our simulations can predict that there is a preferential integration within nucleosomes and that this bias is strongest for rigid tDNA and weakened in more flexible or curved substrates. The integration profiles are generated by averaging over 1000 independent simulations and the dynamics of the simulated process can be seen in Suppl. Movies 1,2.

A Model for HIV Integration as a Quasi-Equilibrium Stochastic Process

When HIV enters the nucleus of a cell it does so in the form of a pre-integration complex (PIC) [25]. This complex is made of the viral DNA (vDNA), the integrase (IN) enzyme which joins the long-terminal-repeats (LTRs) into the intasome structure and a number of host enzymes that facilitate nuclear import and trafficking [26]. For simplicity, we here focus on a simpler model for HIV integration that relies only on the presence of the vDNA and IN. Indeed, these two elements are the only ones necessary and sufficient to perform successful integrations in vitro [10].

To formulate our model, we start from a broadly-employed generic polymer model for DNA and chromatin [27–30]. Both viral DNA and target DNA (tDNA) are treated as semi-flexible bead-spring chains made of beads of size σ and with persistence length l_p typically set to 50 nm for DNA [18] and 30 nm for chromatin [31] if not otherwise stated (Fig. 1A). The dynamics of the chains are evolved by performing Molecular Dynamics simulations in Brownian mode, which implicitly accounts for the presence of a solvent which also works as a thermostat. In practice, this means that vDNA and tDNA explore space diffusively, and that the vDNA searches for its integration target via 3D diffusion, as seen in vitro [32].

Although HIV integration is a complex process that requires many intermediate steps [14, 33], here we are interested in studying a simple model that can capture the essential physics of the process. We thus choose to condense HIV integration into one stochastic step which consists in swapping the bonds of two polymer segments which are transiently close in 3D space (see Fig. 1B). This "integration move" is performed every time any vDNA site is within a distance $R_c = 2\sigma = 5$ nm from a tDNA site. If successful, the vDNA is inserted into the tDNA and it is irreversibly trapped in place, thus becoming a provirus; if rejected, the vDNA is not inserted into the host DNA and resumes its diffusive search. Accounting for the precise position of the intasome along the vDNA does not change our results and we discuss this refinement in the SI.

Because HIV integration does not require ATP to be carried out *in vitro* [8, 9] we argue that the integrase enzyme must work in thermal equilibrium. Thus, we choose to assign an equilibrium acceptance probability to the integration move by computing the total internal energy of the polymer configurations before (Ω) and after (Ω') the move (Fig. 1B). This energy is made of contributions from the bending of the chains, stretching of the bonds and steric interactions. The energy difference $\Delta E = E(\Omega') - E(\Omega)$ is then used to assign the (Metropolis) probability $p = \min\{1, e^{-\Delta E/k_B T}\}$ for accepting or rejecting the integration attempt. Notice that because a successful integration event is irreversible, in reality this process is only in quasi-equilibrium as it violates detailed balance.

It is important to note that, whilst our stochastic quasi-equilibrium model clearly does not reproduce the correct sequence of molecular events leading to integration, it still correctly captures the integration kinetics at longer timescales. This is because such kinetics depend on steric interactions and the energy barrier associated with integration, both included in our model. As the host DNA needs to be severely bent upon integration [16, 33], and as this deformation expends energy that is not provided by ATP [8], we expect that also in reality the IN enzyme will effectively probe the substrate for regions with lower energy barriers against local bending deformations, as targeting these regions leads to faster integration.

Another physical model for HIV integration in DNA was considered in Refs. [15, 16]. An important difference with respect to our work is that in [15, 16] the authors considered the probability of integration to be equal to the Boltzmann weight of the elastic energy of DNA, equilibrated after insertion. Here, instead, we consider a quasi-equilibrium stochastic process in 3D where the energy *barrier* against local deformations and diffusive search are the main determinants of integration profiles. Additionally, Refs. [15, 16] considered 1D models for sequence-dependent naked DNA elasticity, whereas here we consider fully 3D models incorporating nucleosomes and interchromatin interactions.

The Nucleosome is a Geometric Catalyst for HIV Integration

HIV integration on artificially designed DNA sequences in vitro revealed that the IN enzyme displays a pronounced preference for flexible or intrinsically curved DNA sequences [10]. At the same time, chromatinised substrates have been shown to be more efficiently targeted than naked DNA [9]. The affinity to histone-bound DNA is counter-intuitive as the nucleosomal structure may be thought to hinder intasome accessibility to the underlying DNA [14].

To address these unexplained findings, we use our model to simulate the integration of a short viral DNA (40 beads or 320 bp) within a DNA sequence made of 100 beads (or 800 bp) in which the central 20 beads (160 bp) are wrapped in a nucleosomal structure. [The precise lengths of vDNA and tDNA do not change our results as the integration moves are performed locally, see SI]. The nucleosome is modelled by setting a short-ranged attraction between the central segment (orange in Fig. 1A) and a histone-like protein of size $\sigma_h = 3\sigma = 7.5$ nm [27] (see SI for details).

In our simulations tDNA and the histone-like protein diffuse within a confined region of space and spontaneously assemble in a nucleosome as seen in Figure 1A. After the assembly of the nucleosome, we allow the diffusing vDNA to integrate anywhere along the substrate.

Strikingly, we observe that the probability of observing an integration event, $P_{\rm int}(x)$, as a function of the genomic position x displays a $\sim$4-fold increase within the nucleosome (Fig.1C). Integration is instead random, hence near-uniform, within naked DNA (Fig. 1C). In all cases, $P_{\rm int}(x)$ increases at the ends of the host polymer, as integration there entails a smaller bending energy barrier.

These results can be explained by noting that the region of tDNA bound to histones is highly bent. For this reason, the (bending) energy barrier associated with integration is smaller, and insertion is thus more likely.

We also point out that if all nucleosomal segments were fully wrapped we would expect a flat-top integration probability rather than the one observed, that is peaked at the centre of the nucleosomal segment (or at the dyad [18]). Because in our model nucleosomes are dynamic and may partially unravel, the most likely segment to be histone-bound at any time is, by symmetry, the inner central segment thus explaining the integration preference for that specific location (see Fig. 1C).

We conclude this section by directly comparing our findings with those from Refs. [9, 10] (Fig. 1D). First, we notice that our simple model predicts a $\sim 5-$fold enhancement of nucleosomal integration in remarkable agreement with the values reported in Ref. [9] for rigid DNA substrates. Second, the same authors show that this bias is weakened by considering flexible or intrinsically curved DNA substrates. Motivated by this finding, we repeat our simulations using a more flexible substrate with $l_p = 30$ nm and we remarkably observe the same weakening (see Fig. 1C,D). This behaviour can be rationalised within our simple argument: since more flexible (or curved) DNA segments display a much smaller conformational energy when wrapped around histones, the difference in energy barriers within and outside nucleosomal regions is largely reduced in these substrates.

It is finally important to stress that in Refs. [9, 10] the experiments were performed in absence of other enzymatic cofactors. Hence, the observed bias (and behaviour on different substrates) must be solely due to the viral integrase enzyme. This is fully consistent with our results, which show that the nucleosome acts as a "geometrical catalyst" for HIV integration.

HIV Integration in Supercoiled DNA

The double-helical structure endows DNA with torsional rigidity. In turn, this implies that local twist deformations can contribute to the conformational energy and potentially affect the integration process. Our simple model neglects twist rigidity and yet can predict the bias for nucleosomal DNA. We therefore argue that bending deformations may play a more important role than twist ones during HIV integration within mono-nucleosomes.

At the same time, it is well known that in supercoiled DNA, local twist deformations can be transformed into writhe [18], leading to conformations which display a large degree of bending focused at the tips of plectonemes [34]. Although supercoiling is a phenomenon which is formally possible only with circular DNA molecules, the presence of proteins and topological insulators can create effectively supercoiled loops in some regions of the eukaryotic genome [35].

A direct prediction of our model is that supercoiled regions should display an enhanced integration probability due to their larger bending, so that we expect the tips of the plectonemes to be especially targeted.

Experiments have shown that supercoiled DNA displays a $\sim 2-5$-fold enhancement in integration frequency [32] – this is in line with our model as a supercoiled DNA molecule will store more elastic bending energy with respect to a torsionally relaxed one. Yet, there is no direct evidence showing which sub-structure is most targeted within a supercoiled molecule. We propose that our prediction that plectoneme tips should be integration hotspots may be detected and tested in the future using set-ups with magnetic tweezers as done in Ref. [32].

Integration within Nucleosomes is Affected by Local Chromatin Structure

The results from the previous section point to an intriguing role of DNA elasticity in determining the observed preferred integration within mono-nucleosomes. Yet, it is also important to characterise integration site-selection within a poly-nucleosome (chromatin) fibre. To address this level of detail we now model a 290 bead (~ 2.1 kbp) long chromatin fibre, forming an array of 10 nucleosomes.

The self-assembly of the fibre is guided by the same principles of the previous section. Attractive interactions ($\epsilon = 4k_BT$) are assigned between nucleosomal DNA (20 beads or ~ 147 bp) and histone-like proteins (size $\sigma_h = 3\sigma = 7.5$ nm). Linker DNA (10 beads or ~ 74 bp) separates 10 blocks of nucleosomal DNA and the stiffness of the DNA is fixed at $l_p = 20\sigma = 50$ nm. The ground state of this model is an open chromatin fibre, similar to the 10-nm fibre (Fig. 2A). [While the size of our linker DNA is slightly above the average one in eukaryotes, this is chosen to accelerate the self-assembly kinetics of the fibre as the energy paid to bend the linker DNA is lower [36]].

To generate increasing levels of compaction, thus mimicking different local chromatin environments, we now introduce an affinity (or attraction) between selected histone-like proteins. We consider either the case where each nucleosome, labelled i, interacts with its nearest neighbours (nn) along the chain, $i \pm 1$, or with its next-to-nearest (nnn) neighbours, $i \pm 2$. The former case leads to bent/looped linker DNA [36, 37] while the latter a local zig-zag folding displaying straight linker DNA [38] (Fig. 2A). Importantly, recent evidence from both *in vitro* [39] and *in vivo* [40, 41] suggest that both these types of conformations may occur in different regions of the genome – so that the associated chromatin fibre is "heteromorphic" [36, 42]. For the nearest-neighbour case, we also distinguish a partially folded state (nnp) – obtained when $\epsilon_h = 40k_BT$ (Fig. 2B) – and a fully condensed structure (nnf) – when $\epsilon_h = 80k_BT$ (Fig. 2C).

By simulating quasi-equilibrium stochastic HIV integration within these chromatin fibres of different structure, we observe that chromatin compaction yields a notable effect. Whilst open fibres still display an integration probability within nucleosomes that is significantly enhanced with respect to random distributions, this bias is weakened for more compact fibres, especially for nearest-neighbour folding (Figs. 2E,G).

What underlies this change in trend? First, an analysis of the local bending energy landscape along the polymer contour reveals that for nearest-neighbour (nnp and nnf) folding establishing nucleosome contacts requires looping the short linker DNA (Fig. 2F). This increases the local bending stress in the linker DNA, potentially rendering it comparable to the one stored within histone-bound regions (Fig. 2E). In turn, this decreases the energy barrier towards integration within linker DNA. However, this argument does not explain why nucleosomal DNA becomes even less preferred than linker DNA in highly condensed structures with neareast-neighbour attraction. Indeed, they should at most be equally targeted. It also fails to explain why compaction also decreases the preference for nucleosomal DNA in zig-zagging fibers, where the linker DNA is straight (Fig. 2E. We therefore reason that a second important factor is dynamic accessibility: When nucleosomes are tightly packed against each other, there is less available 3D space to reach them diffusively (and it takes longer to do so), and this hinders integration efficiency. This is true especially for the highly condensed structure with nearest-neighbour (nnf) attraction, which indeed leads to the most striking reduction in nucleosomal integration (Fig. 2G).

Another notable result of our simulations is that the overall integration efficiency, measured by number of integrations $n_{\rm int}$ over the total simulation time, is reduced by chromatin compaction, and integration in a zig-zagging fibre yields the globally slowest process. This suggests that integration may be more efficient in open structures, such as euchromatin, with respect to compact ones, normally associated with heterochromatin.

Although a generic tendency of HIV integration to be suppressed in compacted chromatin has been shown in the

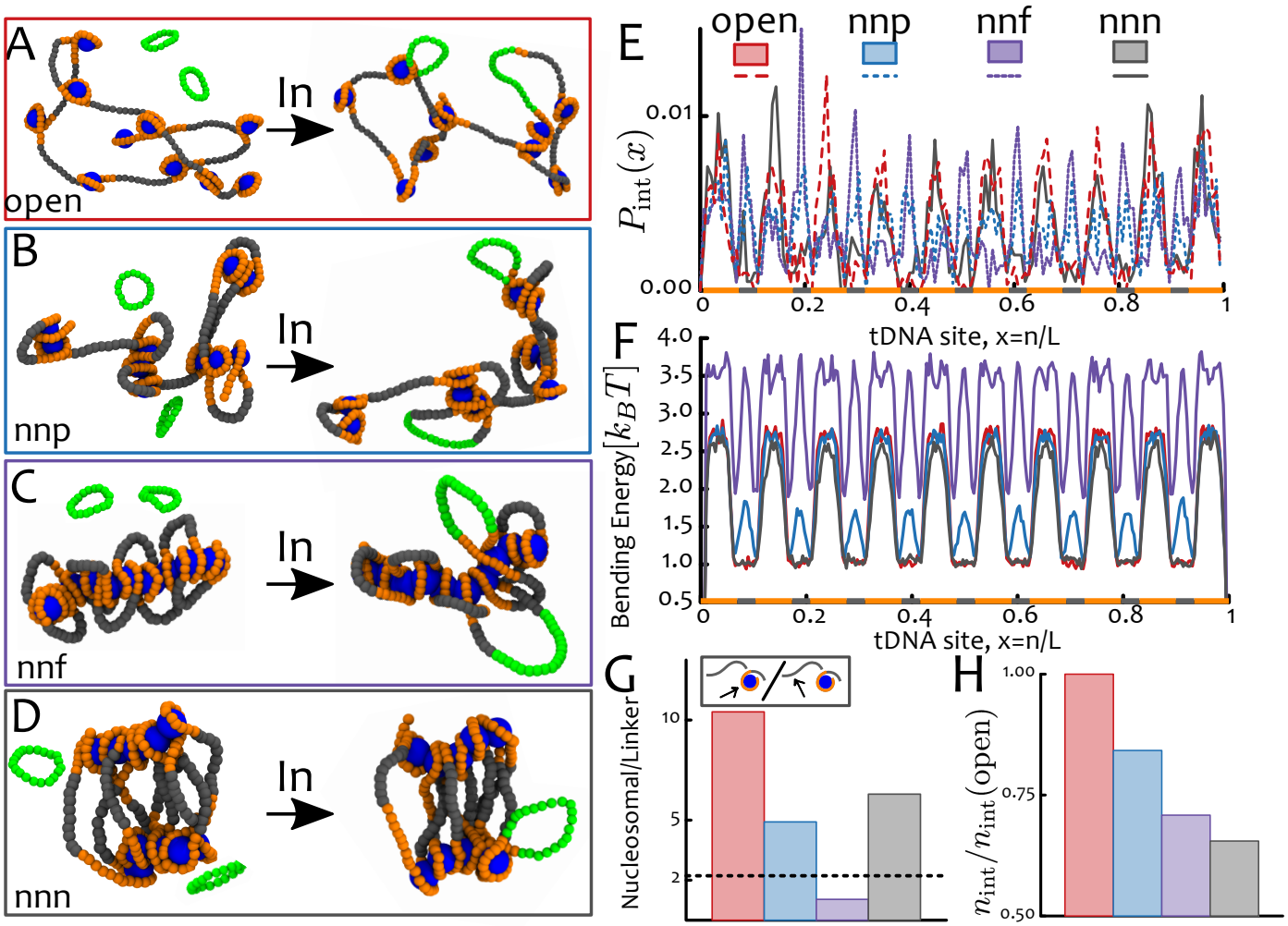

FIG. 2. **Local Chromatin Structure Affects Nucleosomal Integration.** **A** Snapshot of open chromatin fibre composed of 10 nucleosomes. Nearest-neighbour (nn) attraction between the histone-like particles induce partially folded (nnp, **B**) and fully condensed (nnf, **C**) structures. Next-nearest-neighbour attraction instead leads to zig-zagging fibres (nnn, **D**). **E** The integration probability $P_{\rm int}(x)$ as a function of the relative tDNA site $x = n/L$ displays peaks whose location depend on the compaction level. Open fibres are integrated mostly within nucleosomes while folded arrays also display peaks within linker DNA. **E** The bending energy profile shows that fibres with nearest-neighbour attractions (nnp and nnf) but not those with next-nearest-neighbour attraction, display stress within linker DNA. This only partially explains why these regions are targeted within these chromatin structures. **G** The ratio of nucleosomal versus linker DNA integrations suggest that not only energy barrier but also dynamic accessibility plays a role in determining the integration profiles (the expected value for random integration $200/90 = 2.2$ is shown as a dotted line). **H** In line with this, the number of successful integration events over the total simulation time, $n_{\rm int}$, decreases with chromatin condensation. In all cases, the fibre is reconstituted independently before performing the quasi-equilibrium stochastic integration. Data is generated by averaging over 2000 independent integration events. See Suppl. Movies 3,4,5,6 for the full dynamics.

past (see [43] and below), no existing experiment has accurately measured HIV integration profiles within chromatin fibres at different compaction levels, and with different kind of local secondary structures. Thus, we hope that in the future our predictions may be tested using for instance *in vitro* systems with reconstituted chromatin at different salt concentrations.

Integration within Euchromatin is Enhanced by Large-Scale Chromatin Folding

Polymer modelling of large-scale chromatin organisation in 3D has led to some breakthroughs in our current understanding of genome architecture *in vivo* [28, 29, 45–52]. Some of these models strongly suggest that epigenetic patterns made of histone post-translational modifications – such as H3K4me3 or H3K9me3 – play a crucial role in folding the genome in 3D [46, 49, 50]. Based on this evidence we thus ask whether a polymer model of viral integration

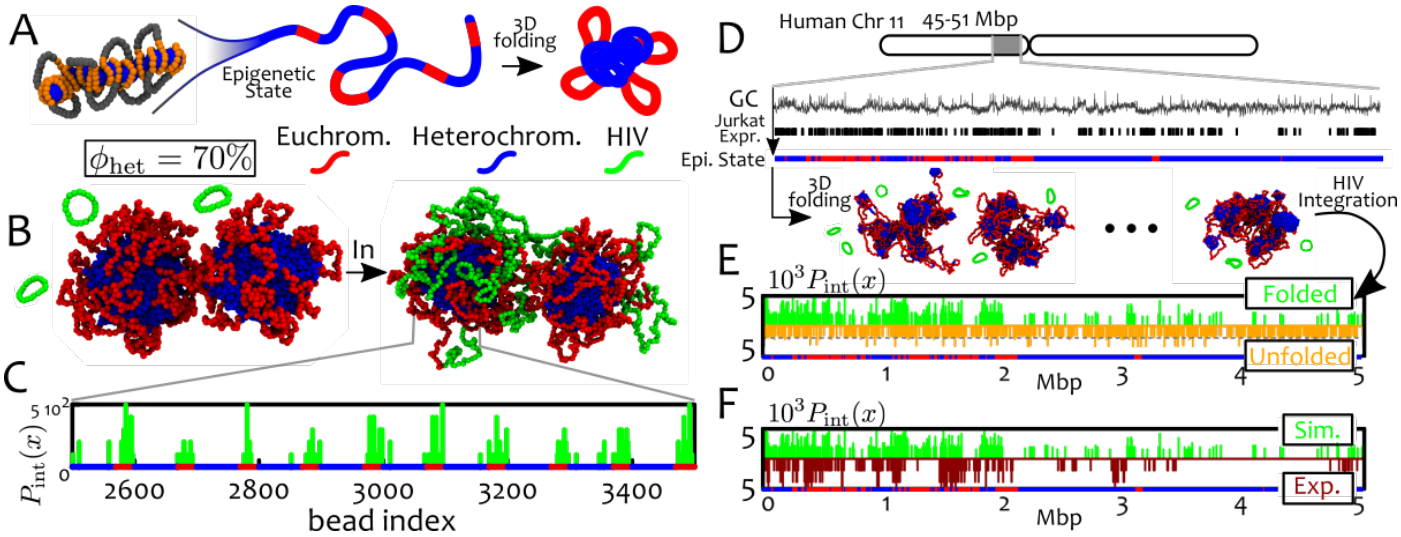

FIG. 3. **Large-Scale 3D Chromatin Folding Enhances Euchromatic Integration.** **A** Pictorial representation of our coarse-grained model which describes chromatin as a fibre with epigenetic marks. These marks dictate 3D folding by self-association through proteins and transcription factors [27, 44]. **B** Snapshots of our polymer model where the fraction of heterochromatin is set at $\phi_{\rm het} = 70\%$. We show two typical configurations, before and after integration events. **C** The integration probability displays a strong enrichment in euchromatic regions. **D-F** Simulations of a 5 Mbp region of human chromosome 11 (45-51 Mbp) modelled at 1 kb resolution with a polymer $N = 5000$ beads long. **D** In this model, expression level in Jurkat T-cells and GC content are used to label beads as euchromatic (red) or heterochromatic (blue) respectively. We assign attractive interactions ($\epsilon = 3\ k_BT$) between heterochromatin beads so that the fixed epigenetic pattern guides the folding of the chromatin fibre (see snapshots). Steady state conformations are then used as hosts for $n = 500$ integration events of a 10 kbp viral DNA. **E** Comparison between the distribution of integration sites in folded and unfolded chromatin conformations. The latter is obtained by assigning no self-attraction between heterochromatin beads. Viral integration within unfolded chromatin is uniform ($P_{\rm int} = 1/n$, dashed line) while it is not uniform (i.e. non-random) for folded chromatin. **F** Comparison between simulated and experimentally-measured distribution of integration sites in Jurkat T-cells [20]. The agreement between simulations and experiments is highly significant, with a p-value $p < 0.001$ when a Spearman Rank is used to test the null-hypothesis that the distributions are independent. This result can be compared with the p-value $p = 0.6$ obtained when the same test is performed to test independence of the integration profiles in experiments and unfolded chromatin. The dynamics corresponding to one of our simulations is shown in full in Suppl. Movie 7.

in a chromatin fibre that is folded in 3D based upon its epigenetic patterns can give us some insight of how large-scale 3D chromatin architecture determines the distribution of integration sites.

To do so, we coarse-grain our poly-nucleosomal fibre found in the previous section and model it as a polymer of thickness $\sigma = 10$ nm (about the size of a nucleosome). We further assume that the histones carry epigenetic marks which then drive the large-scale folding and perform our quasi-equilibrium sotchastic integration process within these folded substrates. To study this model we start from an idealised block co-polymer model in which 50 blocks of 100 beads are portioned into 30 euchromatic and 70 heterochromatic beads (fraction of heterochromatin $\phi_{\rm het} = 70\%$, see Fig. 3). Heterochromatic compaction is driven by implicit multivalent bridges [44, 53], which are effectively accounted for by endowing heterochromatic beads with a weak self-attraction ($\epsilon = 3\ k_BT$, see SI and Refs. [46, 49]). In contrast, we assume that euchromatic beads interact only by steric repulsion, for simplicity. This model naturally drives the phase-separation of the system into compartments of compact heterochromatin (or "B") decorated by swollen loops of euchromatin (or "A") [54].

To dissect the effect of 3D organisation from that of flexibility discussed previously, we now set the same persistence length everywhere along the fibre ($l_p = 3\sigma$), and measure the steady-state distribution of integration sites.

As shown in Figure 3C, we observe that the probability of integration is highly enriched in euchromatic regions while viral incorporation into heterochromatic regions is strongly suppressed. Local chromatin folding therefore provides a second driver, besides flexibility, favouring HIV integration in active region, in excellent qualitative agreement with experiments [1, 19, 20].

Our simulations give a mechanistic insight into the biophysical mechanism which may underlie this phenomenon. Inspection of the simulations trajectories suggests that a recurrent structure in our model is a daisy-like configuration (see Fig. 3A), with one or more large heterochromatic cores, "screened" by many euchromatic loops (petals). The latter are therefore the regions which first encounter the diffusing viral DNA; a similar organisation is expected near nuclear pores of inter-phase nuclei, where "channels" of low density chromatin separate inactive and lamin-associated

regions of the genome [55].

Our Generic Polymer Model Accurately Predicts HIV Integration Profiles in Human T-Cells

To quantitatively test our generic co-polymer model for inter-phase chromosomes, we consider a region of the chromosome 11 in Jurkat T-cell (45-51 Mbp). We coarse-grain the chromatin fibre into beads of size $\sigma = 1$ kbp $\simeq 10$ nm, and label them as euchromatin (red) if the corresponding genomic location simultaneously display high GC content and high expression in the Jurkat cell line (see Fig. 3D, data available from ENCODE [56] and Ref. [20]). The remaining beads are marked as heterochromatin (blue). The threshold in GC content and expression level is set in such a way that the overall heterochromatin content is $\sim 70\%$, as in the previous case. We then compare the statistics of integration events that occur within a folded chromatin fibre (by imposing a weak heterochromatin self-attraction, as before $\epsilon = 3\ k_BT$, see SI and Refs. [46]) and within a non-folded substrate (by imposing repulsive interaction between any two beads irrespectively if hetero- or euchromatic).

Our results confirm that the reason behind the non-random distribution of integration sites within our framework is indeed the 3D folding of the chromatin fibre, as we instead find a uniform probability of integration events in the unfolded case (see Fig. 3E and see also Suppl. Movie 3). We finally compare the distribution of predicted integration sites with those detected by genome-wide sequencing in Ref. [20] (Fig. 3F). We do this by testing the independence of the integration profiles in real T-cells and in silico using a Spearman Rank test. This reports a highly significant agreement ($p < 0.001$) between experimental and simulated HIV integration profiles in folded substrates and it confirms that there is no correlation ($p \simeq 0.6$) between experimental profiles and the ones found along unfolded chromatin substrates.

Our results thus suggest that the large-scale 3D chromatin organisation is an important physical driver that can bias the distribution of HIV integration sites even when the substrate displays uniform elasticity. Because of the daisy-like conformation assumed by folded chromosomes – with a heterochromatic core screened by euchromatic "petals" – the integration events are more likely to occur on euchromatin regions as these are the most easily accessible.

We thus conclude this section by suggesting that large-scale chromosome folding is a generic physical driver that underlies integration site-selection for all families of retrovirus that target inter-phase nuclei. Specifically, we find that the bias for open chromatin is a direct consequence of diffusive target search along a pre-folded substrate and we argue that this mechanism is at work even in absence of known tethering factors such as LEDGF/p75. Although this nuclear protein enhances the preference of HIV for euchromatin [19], it is also well-established that this preference remains significantly above random in cells where LEDGF/p75 is knocked-out [12, 19, 22].

Varying Heterochromatin Content Affects the Statistics and the Rate of Integration Events

Distinct cell types may display dramatically different amounts of active and inactive chromatin, and this aspect has been shown to affect HIV integration efficiency, at least in some cases. Most notably, a "resting" T-cell, which contains a larger abundance of the H3K9me3 mark [57, 58] and of cytologically-defined heterochromatin, has been shown to be less likely to be infected by HIV with respect to an "activated" T-cell. It is also known that the few resting cells which get infected do so after a sizeable delay [59].

Our previous results showed that, within a single chromatin fibre, integration within euchromatin is more likely than in heterochromatin. However, these do not lead to a direct prediction of difference in infection (or integration) time on chromatin fibres with different global epigenetic composition. To address this issue, we now consider a block co-polymer chromatin model with varying fraction of heterochromatin content ($\phi_{\rm het}$ =30%, 50% and 80%).

In Figure 4A-B we show typical 3D structures of chromatin fibres with different heterochromatic content. When the latter is small ($\phi_{\rm het}$ =30%), heterochromatin self-organises into globular compartments of self-limiting size, surrounded by long euchromatin loops which entropically hinder the coalescence of heterochromatic globules [31, 60]. For large heterochromatin content ($\phi_{\rm het}$ =80%), inactive domains merge to form a large central core, "decorated" by short euchromatic loops, resembling the above-mentioned daisy-like structure.

Our simulations confirm that viral loops integrate preferentially in open, euchromatin regions in all these cases. Additionally, we observe that the total time taken for the viral loops to integrate within the genome increases, at least exponentially, with the abundance of heterochromatin (Fig. 4C). In biological terms this implies that "resting", heterochromatin rich, T-cells are much more difficult to infect with respect to activated T-cells. The same results additionally suggest that stem cells, which are euchromatin-rich, should be infected more quickly with respect to differentiated cells, which contain more heterochromatin [61–63]. This is in qualitative agreement with existing experiments on lentivirus infection [59, 64].

A further surprising result is that the efficiency of viral integration in the euchromatic parts of the genome increases with the total fraction of heterochromatin. This can be quantified by measuring the integration probability within a given epigenetic state "s" as

$$P^s_{\rm int} = \sum_{i=1}^{N} P_{\rm int}(i)\delta(s(i) - s) \qquad (1)$$

where $s(i)$ is the epigenetic state of the i-th bead. For random integration events, i.e., constant $P_{\rm int} = 1/N$, one obtains $P^s_{\rm random} = \phi_s$. Hence the change in integration efficiency due to the 3D organisation can be quantified as $\chi_s = P^s_{\rm int}/\phi_s$ (see Fig. 4D).

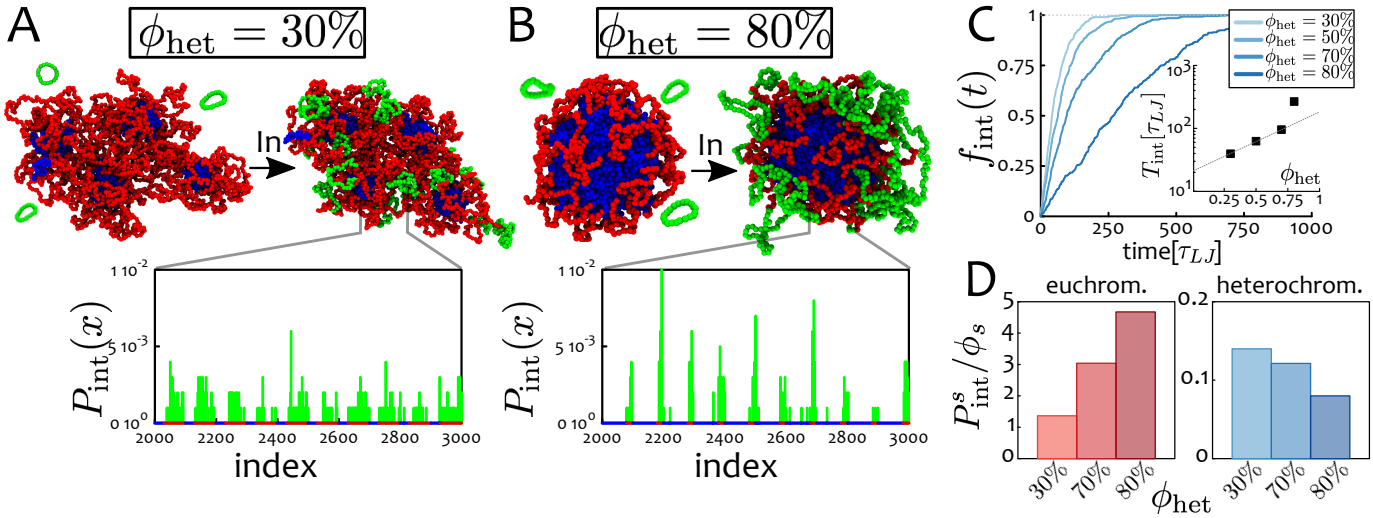

FIG. 4. **Integration is Slowed Down in Cells with Large Heterochromatin Content. A-B** Snapshots and probability distribution for a system with $\phi_{\rm het} = 30\%$ and 80%, respectively. **C** Fraction of integrated loops $f_{\rm int}$ as a function of time, for different levels of heterochromatin. In the inset the integration time $T_{\rm int}$ defined as $f_{\rm int}(T_{\rm int}) = 0.5$ is shown to (super-)exponentially increase as a function of $\phi_{\rm het}$. **D** Integration probability in state s – with s being either euchromatin (red) or heterochromatin (blue) – normalised by the total fraction of the host polymer in state s, ϕ_s. The plots show that, counter-intuitively, the larger the fraction of heterochromatin, the more likely it is for a loop to be integrated in euchromatin.

We find that $\chi_{\rm eu}$ increases as a function of $\phi_{\rm het}$ while $\chi_{\rm het}$ decreases. This counterintuitive observation can be understood as a direct consequence of 3D chromatin architecture. The more heterochromatin is present in the nucleus, the stronger the inactive ("B") compartments and the more they are screened by euchromatin loops. As far as we know, this finding has never been directly observed. It would be interesting to see whether future experiments may detect this change across different cell lines and confirm our prediction.

A Reaction-Diffusion Model for Retroviral Integrations in Human Nuclei

Having observed that large-scale chromosome folding can affect the distribution of HIV integration sites trough chromatin accessibility, we now aim to put this finding into the context of a realistic inter-phase nuclear environment. Because performing polymer simulations of a full genome is not currently feasible, we consider the observations made in the previous sections to formulate a continuum model of whole cell nuclei. We do so by coarse graining the behaviour of retroviral DNA in the nucleus as a random walk inside a sphere of radius R, which can integrate into the host genome at a rate κ. In general, the diffusion constant D and the integration rate κ will depend on the position of the viral loop in the nuclear environment. Indeed we have seen before that local epigenetic state and chromatin architecture play important roles in determining HIV integration rate and patterns.

Within this model, the probability $\rho(\boldsymbol{x}, t)$ of finding a viral loop in the nucleus at position $\boldsymbol{x}$ and time t obeys the following reaction-diffusion equation:

$$\partial_t \rho(\boldsymbol{x}, t) = \nabla \left(D(\boldsymbol{x}) \nabla \rho(\boldsymbol{x}, t) \right) - \kappa(\boldsymbol{x}) \rho(\boldsymbol{x}, t) \,. \qquad (2)$$

For simplicity, we assume spherical symmetry, i.e. $\rho(r, \theta, \phi, t) = \rho(r, t)$, and piecewise constant functions for D and κ (see below). With these assumptions, Eq. (2) becomes $\partial_t \rho = D/r^2 \partial_r \left(r^2 \partial_r \rho \right) - \kappa \rho$, where we have dropped, for notational simplicity, all dependences on r and t. In order to obtain the steady-state probability of integration sites, we thus need to find the time-independent distribution $\rho(r, t) = \rho(r)$ by solving the equation

$$\frac{D}{r^2} \partial_r \left(r^2 \partial_r \rho \right) - \kappa \rho = 0 \,. \qquad (3)$$

In the simplest case in which D and κ are uniform throughout the nucleus, the solution of Eq. (3) is

$$\rho(r) = \mathcal{N} \frac{\sinh(r/l)}{r} \,, \qquad (4)$$

where $l = \sqrt{D/\kappa}$ is a "penetration length", measuring the typical lengthscale that vDNA diffuses before integrating into the host, while $\mathcal{N}^{-1} = \int_0^R dt \sinh(t)/t$ is a normalisation constant (see SI for details).

To solve Eq. (2) in more general cases, we need to make some assumptions on how D and κ may vary within the nuclear environment. In line with our previous results at smaller scale, we assume that these parameters depend on local chromatin state – as we shall discuss, this is often dependent on nuclear location.

First, we need to model viral diffusivity in euchromatin and heterochromatin. Assuming faster or slower diffusion

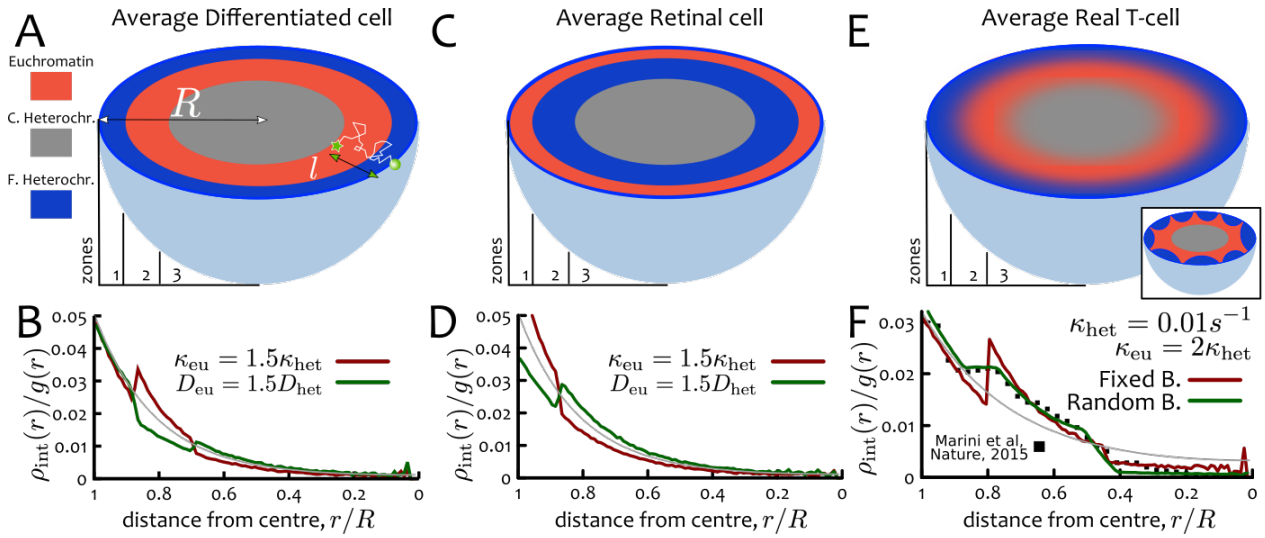

FIG. 5. **HIV Penetration Length Depends on the Nuclear Environment. A, C, E** Different cell lines display different chromatin organisations at the nuclear scale. **A** Shows a typical differentiated cells, modelled as a sphere with 3 concentric shells of equal volume. Zones 1 and 3 are populated by facultative (or lamin-associated) and constitutive heterochromatin, respectively; both are broadly transcriptionally silent. Zone 2, the middle layer, is populated by euchromatin. This configuration may be viewed as an angularly averaged model and it is appropriate to study HIV integration in population averages. **C** Shows the model for a "retinal" cell, where the two outer layers are inverted [65]. **E** Shows the model for a realistic population of T-cells (typical configuration of a single cell is shown in the inset). Here the location of the boundaries between zones 1 and 2, and between zones 2 and 3, is varied to account for local density variations and cell-to-cell fluctuations (see text). **B, D, F** Nuclear distribution of HIV integration sites in (**B**) differentiated cells, (**D**) retinal cells and (**F**) T-cells. The result with uniform D and κ (yielding $l \simeq 5$ μm) is shown in grey in each panel. The number of integrations at distance r, $\rho_{\rm int}(r)$, is divided by the area of the shell, $g(r) = 4\pi r^2 dr$, and normalised to unity. Filled squares in **F** denote data from Ref. [66].

in euchromatin are both potentially reasonable choices: the former assumption describes situations where euchromatin is more open and less compact [17, 67]. Assuming slower diffusivity in euchromatin may instead model local gel formation by mesh-forming architectural proteins such as SAF-A [68]. Second, the recombination rate κ may be thought of as effectively depending on local DNA/chromatin flexibility and 3D conformation, and given our previous results we expect it to be larger in euchromatin-rich nuclear regions.

For simplicity, we additionally posit that chromatin is organised in the nucleus into 3 main concentric zones. Each of these zones displays an enrichment of a particular chromatin state. This is the situation of typical differentiated cells, where it is well established that the most inner and outer zones are generally populated by transcriptionally inactive chromatin (heterochromatin and lamin-associated-domains, respectively [2, 69]), whereas the middle layer is commonly enriched in transcriptionally active euchromatin [65]. To mimic this organisation, and for simplicity, in our model D and κ vary spatially, but we assume a constant value within each of the three layers. Whilst this may be a crude approximation for individual cells, which are known to display heterogeneity in the local chromosome organisation [61, 63], our model may be more suitable to capture trends corresponding to ones obtained from a population of cells displaying the same average nuclear arrangement (see Fig. 5).

To highlight the effect of nuclear organisation on the spatial distribution of HIV integration sites, we compare the case just discussed of a differentiated cells, displaying a conventional layering, with cells displaying an "inverted" organisation, such as the retinal cells of nocturnal animals [65] (see Fig. 5C).

By measuring the integration profile $\rho_{\rm int}(r)$ (normalised by the area of the shell $g(r) = 4\pi r^2 dr$), we how the integration profile changes because of non-uniform D and κ (see Fig. 5B,D). As expected, we find that setting the recombination rate in euchromatin, $\kappa_{\rm eu}$, larger than the one in heterochromatin, $\kappa_{\rm het}$, enhances the probability of integration in the middle euchromatic layer in differentiated cells. On the other hand, faster diffusion in euchromatin-rich regions has the opposite effect (as fast diffusion depletes virus concentration). We also predict that the distribution of integration events in retinal cells should be very different. Here, larger $\kappa_{\rm eu}$ enhances the probability of integration near the periphery and, as before, increasing $D_{\rm eu}$ has the opposite effect.

Observed Distribution of HIV Hot-Spots in T-Cells is Recapitulated by a Refined Model

We now quantitatively compare our reaction-diffusion model with the experimentally measured distribution of

HIV recurrent integration genes (RIGs) in T-cells [66]. Remarkably, we find that our simple theory with uniform $D = 0.05\ \mu\text{m}^2/s$ and $\kappa = 0.002\, s^{-1}$ (leading to a penetration length $l = \sqrt{D/\kappa} = 5\mu\text{m}$) is already in fair overall agreement with the experimental curve (Fig. 5F). The agreement can be improved by progressively refining our model and adding more stringent assumptions. As we show below, these refinements also lead us to obtain more physical insight into the nuclear organisation of chromatin in real T-cells.

First, we find that a better agreement is achieved if D remains uniform, but κ varies and $\kappa_{\text{eu}}/\kappa_{\text{het}} \simeq 1.5$ (equivalently, a model with $D_{\text{het}}/D_{\text{eu}} \simeq 1.5$ and uniform κ would yield the same result). A second improvement is found by re-sizing the three concentric nuclear shells as follows. We maintain the total mass of heterochromatin fixed at twice that of euchromatin, as realistic *in vivo* [2]. The volume of each layer has to adapt according to the fact that active chromatin is less dense than heterochromatin [17, 55]. It is possible to derive an equation relating the ratio between the density of heterochromatin and euchromatin, $\rho_{\text{het}}/\rho_{\text{eu}}$, to the positions of the boundaries between layers, R_{1-2} and R_{2-3}, as (see SI)

$$\frac{\rho_{\text{het}}}{\rho_{\text{eu}}} = \frac{2\left(R_{2-3}^3 - R_{1-2}^3\right)}{R^3 + R_{1-2}^3 - R_{2-3}^3}\,. \tag{5}$$

For a nucleus of radius R, we find that setting $R_{2-3}/R = 0.8$ and $R_{1-2}/R = 0.445$ (to match the data from Ref. [66]) we obtain $\rho_{\text{het}}/\rho_{\text{eu}} = 1.6$. Incidentally, this value is in pleasing agreement with recent microscopy measurements, reporting a value of 1.53 [70].

Two final refinements that we consider here are allowing for small fluctuations ($\pm 0.5\,\mu$m) in the position of the boundaries in each simulated nucleus and imposing that the innermost boundary, between euchromatin and constitutive heterochromatin, cannot always be crossed by the viral loop (this is done by assigning a successful crossing probability of $p = 0.5$). The former assumption accounts for the heterogeneity in a population of cells while the latter is related to the fact that the central zone of nuclei is often occupied by the nucleolus, a structure largely devoid of DNA. By including these realistic refinements, our theory matches extremely well the experimental measurements (Fig. 5F).

Discussion

In this work we have proposed a generic physical model for HIV integration in DNA and human chromatin, which is based on 3D diffusive target search and quasi-equilibrium stochastic integration (dependent on local energy barriers). Importantly, our model purposely neglects the interaction of the pre-integration complex (PIC) with other co-factors, such as LEDGF/p75 and nuclear pore proteins [12, 19, 71]. We make this choice both for simplicity and to focus on the key physical ingredients that are necessary and sufficient to recapitulate a bias in the site-selection process.

It is worth noting that the interaction of HIV with nucleoporins and cellular proteins is mainly relevant to ensure successful nuclear entry of the PIC [4, 25]. Similarly, the cleavage and polyadenylation specificity factor 6 (CPSF6) is known to bind to the viral capsid (CA), and is required for nuclear entry [4]. Recent evidence suggest that CPSF6 may also contribute to the site selection process [71]; yet, while the viral CA is present in the nucleus of primary macrophages [72], there is no evidence suggesting that CA enters the nucleus of primary T cells, thus questioning the relevance of CPSF6 in this cell lineage (on which we focus here when comparing to human chromosome HIV integration patterns).

Finally, while it is well-established that the presence of functional LEDGF/p75 enhances euchromatin integration [4, 19], this preference is found to persist significantly above random when LEDGF/p75 is knocked-out [12, 19, 22]. All this calls for a model that can explain non-random HIV site-selection independently of other co-factors, such as the one we have proposed here.

In the future, it would be possible to consider a refinement of our model in which a euchromatin tethering factor such as LEDGF/p75 is accounted for by setting specific attractive interactions between the vDNA polymer and euchromatic regions. This refined model is expected to naturally result in an enhancement of euchromatic integrations since the vDNA would spend more time in their vicinity. While we here find that this element is not necessary to recapitulate the preference for euchromatin within our simple model of chromosomes, we realise that it may be interesting to study its role within the context of a more realistic interphase nuclear environment, where the PIC has to traffic through a complex and crowded space. We leave this investigation for subsequent studies.

We finally argue that because our model is based on few generic assumptions, i.e. that of diffusive search and energy barrier sensing, our results are expected to hold for a number of retroviral families as long as their members undergo diffusion within the nucleus and require bending of the tDNA substrate to perform integration. Important exceptions are the families of alpha- and beta-retroviruses. First, they are different as they possess a unique intasome structure with an octomeric integrase enzyme that can accommodate unbent tDNA [73, 74]. We thus expect that local substrate deformations play a minor role for these retroviral families and that they may not display a bias for flexible or nucleosomal DNA. Second, these viruses are unlike HIV as they cannot traverse the nuclear envelope and must rely on its dissolution during mitosis to infect the host genome. As a consequence, their typical chromatin substrate is that of mitotic chromosomes, well-known to possess a unique large-scale chromatin structure with respect to the one we consider in this work for interphase chromosomes. Thus, we do not expect our simulations to capture the integration patterns of these viruses. Yet, future studies may address these dissimilarities in more

detail.

Conclusions

In this work we propose a generic biophysical model to rationalise the problem of how HIV can display non-random distributions of integration sites along the genome of the host. Our model identifies two key physical features underlying this non-trivial selection: local genome elasticity and large-scale chromatin accessibility. These two biophysical drivers are active at multiple length scales, and create trends which are in qualitative and quantitative agreement with experimental observations. Importantly, we stress that these two mechanisms are at play even in absence of known co-factors, for instance in vitro or in knock-out experiments [9, 19, 22], and should thus be considered as forming the physical basis of HIV integration on top of which several important biological and biochemical processes need to be accounted for in order to achieve a realistic picture of HIV integration.

By modelling a typical integration event as a stochastic quasi-equilibrium process we find a bias towards highly bent or flexible regions of the genome, in quantitative agreement with long-standing experimental observations (see Fig. 1). This bias can be explained as resulting from the difference in energy barrier against local deformation of the underlying tDNA substrate. For instance the nucleosome structure can be readily associated to a location with low energy barrier and thus acts as a geometric catalyst for the integration process.

At larger scales, our model predicts HIV integration patterns closely matching experimental ones and showing a marked preference towards transcriptionally active euchromatin (see Fig. 3). This finding can be explained by considering that 3D chromosome folding determines local chromatin accessibility. Specifically, we find that the integration patterns are dictated by the underlying epigenetic marks it displays typical "daisy-like" conformations in which euchromatic loops screen dense heterochromatin cores from external HIV integrations (see Fig. 3-4). In general, we argue that any viral DNA that probes the substrate diffusively is bound be affected by the large-scale 3D folding of the substrate. In line with this, we further predict that cell lines displaying a larger abundance of hetechromatin – such as resting T-cells and differentiated cells – are infected more slowly than ones richer in active chromatin – such as activated T-cells and stem cells (Fig. 4). Finally, we propose and solve a simple reaction-diffusion model that can capture the distribution of integration hotspots within the nuclear environment in human T-cells [66] (Fig. 5).

Besides rationalising much existing evidence on HIV integration by using a minimal number of assumptions, our model also leads to a number of predictions. For instance, it suggests that integration events that are not in "quasi-equilibrium", i.e. that consume ATP to deform the substrate, cannot probe the free energy landscape of the substrate and should thus not display any bias towards nucleosomal DNA. This scenario may be relevant if intasome complexes expending ATP are found, or artificially built. At the chromosome level, our model can be used to predict the distribution of integration sites within a chromatin fibre for which the epigenetic patterns are known. Thus, it can potentially be used to predict HIV integration profiles in a number of different cell lineages and organisms. These results could finally be compared, or combined, with "chromosome conformation capture" (CCC or Hi-C) analysis to provide new insight into the relationship between HIV integration and large-scale chromatin organisation in living cells [66, 75–77]. At the whole cell level, our reaction-diffusion model can predict how the distribution of HIV hot-spots may change in cells with non-standard genomic arrangements, such as retinal cells in nocturnal animals [65], but also senescent [78] and diseased cells in humans and mammals.

Acknowledgements We thank D. W. Sumners and J. Allan for inspiring discussions. This work was supported by ERC (CoG 648050, THREEDCELLPHYSICS).

Methods

We model DNA and chromatin as semi-flexible polymers made of beads of size σ connected by springs. Integration moves are performed by attempting swaps between bonds connecting consecutive beads in viral and host DNA. These moves are accepted or rejected according to a Metropolis algorithm, where the acceptance probability is based on the energy difference caused by the reconnection event. Molecular dynamics simulations are performed within the LAMMPS engine [79] in Brownian dynamics mode with a velocity-Verlet integration step of $\Delta t = 0.001\tau_{Br}$. Further details on the simulations and on the analytical solutions of the reaction-diffusion equations are presented in the SI.

[1] Craigie R, Bushman FD (2012) HIV DNA integration. *Cold Spring Harb. Persp. Med.* 2(7):1–18.
[2] Alberts B, Johnson A, Lewis J, Morgan D, Raff M (2014) *Molecular Biology of the Cell.* (Taylor & Francis), p. 1464.
[3] Cook P (2001) *Principles of Nuclear Structure and Function.* (Wiley).
[4] Lusic M, Siliciano RF (2017) Nuclear landscape of HIV-1 infection and integration. *Nat. Rev. Microbiol.* 15(2):69–82.
[5] Dewannieux M, et al. (2006) Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res.* 16(12):1548–1556.
[6] Griffiths DJ (2001) Endogenous retroviruses in the human genome sequence. *Genome Biol.* 2(6):reviews1017.1–1017.5.
[7] Bushman F, et al. (2005) Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* 3(11):848–58.

[8] Panganiban AT (1985) Retroviral DNA integration. *Cell* 42(1):5–6.

[9] Pruss D, Reeves R, Bushman F, Wolffe A (1994) The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J. Biol. Chem.* 269(40):25031–25041.

[10] Pruss D, Bushman F, Wolffe A (1994) Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc. Natl. Acad. Sci. USA* 91(June):5913–5917.

[11] Müller HP, Varmus HE (1994) DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* 13(19):4704–14.

[12] Schrijvers R, et al. (2012) HRP-2 determines HIV-1 integration site selection in LEDGF/p75 depleted cells. *Retrovirology* 9:1–7.

[13] Matysiak J, et al. (2017) Modulation of chromatin structure by the FACT histone chaperone complex regulates HIV-1 integration. *Retrovirology* 14(1):39.

[14] Maskell DP, et al. (2015) Structural basis for retroviral integration into nucleosomes. *Nature* 523(7560):366–369.

[15] Benleulmi M, et al. (2015) Intasome architecture and chromatin density modulate retroviral integration into nucleosome. *Retrovirology* 12(1):13.

[16] Naughtin M, et al. (2015) DNA physical properties and nucleosome positions are major determinants of HIV-1 integrase selectivity. *PLoS One* 10(6):1–28.

[17] Gilbert N, et al. (2004) Chromatin architecture of the human genome: Gene-rich domains are enriched in open chromatin fibers. *Cell* 118(5):555–566.

[18] Calladine CR, Drew H, Luisi FB, Travers AA (1997) *Understanding DNA: the molecule and how it works.* (Elsevier Academic Press).

[19] Kvaratskhelia M, Sharma A, Larue RC, Serrao E, Engelman A (2014) Molecular mechanisms of retroviral integration site selection. *Nucleic Acids Res.* 42(16):10209–10225.

[20] Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD (2007) HIV integration site selection : Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* 17:1186–1194.

[21] Ciuffi A, et al. (2005) A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* 11(12):1287–1289.

[22] Marshall HM, et al. (2007) Role of PSIP 1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS One* 2(12).

[23] Shun MC, et al. (2007) LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev.* 21(14):1767–1778.

[24] Koh Y, et al. (2013) Differential Effects of Human Immunodeficiency Virus Type 1 Capsid and Cellular Factors Nucleoporin 153 and LEDGF/p75 on the Efficiency and Specificity of Viral DNA Integration. *J. Virol.* 87(1):648–658.

[25] Engelman A, Cherepanov P (2012) The structural biology of HIV-1: Mechanistic and therapeutic insights. *Nat. Rev. Microbiol.* 10(4):279–290.

[26] Matreyek KA, Engelman A (2013) Viral and cellular requirements for the nuclear entry of retroviral preintegration nucleoprotein complexes. *Viruses* 5(10):2483–2511.

[27] Brackley CA, Allan J, Keszenman-Pereyra D, Marenduzzo D (2015) Topological constraints strongly affect chromatin reconstitution in silico. *Nucleic Acids Res.* 43(1):63–73.

[28] Barbieri M, et al. (2012) Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci. USA* 109(40):16173–8.

[29] Mirny LA (2011) The fractal globule as a model of chromatin architecture in the cell. *Chromosom. Res.* 19(1):37–51.

[30] Rosa A, Everaers R (2008) Structure and dynamics of interphase chromosomes. *PLoS Comp. Biol.* 4(8):1.

[31] Cook PR, Marenduzzo D (2009) Entropic organization of interphase chromosomes. *J. Cell. Biol.* 186(6):825–34.

[32] Jones ND, et al. (2016) Retroviral intasomes search for a target DNA by 1D diffusion which rarely results in integration. *Nat. Commun.* 7(11409):1–9.

[33] Lesbats P, et al. (2016) Retroviral DNA Integration. *Chem. Rev.* 116:12730–12757.

[34] Bates A, Maxwell A (2005) *DNA topology.* (Oxford University Press).

[35] Naughton C, et al. (2013) Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat. Struct. Mol. Biol.* 20(3):387–395.

[36] Grigoryev SA, Arya G, Correll S, Woodcock CL, Schlick T (2009) Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. *Proceedings of the National Academy of Sciences* 106(32):13317–13322.

[37] Bascom GD, Kim T, Schlick T (2017) Kilobase pair chromatin fiber contacts promoted by living-system-like dna linker length distributions and nucleosome depletion. *The Journal of Physical Chemistry B* 121(15):3882–3894.

[38] Szerlong HJ, Hansen JC (2010) Nucleosome distribution and linker dna: connecting nuclear function to dynamic chromatin structure. *Biochemistry and Cell Biology* 89(1):24–34.

[39] Grigoryev SA, Woodcock CL (2012) Chromatin organization - The 30nm fiber. *Exp. Cell Res.* 318(12):1448–1455.

[40] Risca VI, Denny SK, Straight AF, Greenleaf WJ (2017) Variable chromatin structure revealed by in situ spatially correlated dna cleavage mapping. *Nature* 541(7636):237.

[41] Ou HD, et al. (2017) ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science (80-.).* 357(6349).

[42] Garces R, Podgornik R, Lorman V (2015) Antipolar and Anticlinic Mesophase Order in Chromatin Induced by Nucleosome Polarity and Chirality Correlations. *Phys. Rev. Lett.* 238102(June):1–5.

[43] Taganov KD, et al. (2004) Integrase-Specific Enhancement and Suppression of Retroviral DNA Integration by Compacted Chromatin Structure In Vitro. *J. Virol.* 78(11):5848–5855.

[44] Brackley CA, Taylor S, Papantonis A, Cook PR, Marenduzzo D (2013) Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. *Proc. Natl. Acad. Sci. USA* 110(38):E3605–11.

[45] Brackley CA, et al. (2017) Ephemeral Protein Binding to DNA Shapes Stable Nuclear Bodies and Chromatin Domains. *Biophys J.* 112(6):1085–1093.

[46] Brackley CA, Johnson J, Kelly S, Cook PR, Marenduzzo D (2016) Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains. *Nucleic Acids Res.* 44(8):3503–3512.

[47] Di Stefano M, Rosa A, Belcastro V, di Bernardo D, Micheletti C (2013) Colocalization of Coregulated Genes: A Steered Molecular Dynamics Study of Human Chromosome 19. *PLoS Comput. Biol.* 9(3):1–13.

[48] Ulianov SV, et al. (2016) Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res.* 26(1):70–84.

[49] Michieletto D, Orlandini E, Marenduzzo D (2016) Polymer Model with Epigenetic Recolouring Reveals a Pathway for the de novo Establishment and 3D Organisation of Chromatin Domains. *Phys. Rev. X* 6:041047.

[50] Jost D, Carrivain P, Cavalli G, Vaillant C (2014) Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Research* 42(15):1–9.

[51] Michieletto D, et al. (2018) Shaping Epigenetic Memory via Genomic Bookmarking. *Nucleic Acids Res.* 46(1):83–93.

[52] Michieletto D, Orlandini E, Marenduzzo D (2017) Epigenetic Transitions and Knotted Solitons in Stretched Chromatin. *Sci. Rep.* 7:14642.

[53] Sexton T, et al. (2012) Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell* 148(3):458–472.

[54] Rao SSP, et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665–1680.

[55] Cremer T, et al. (2015) The 4D nucleome: Evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments. *FEBS Lett.* 589(20):2931–2943.

[56] Dunham I, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.

[57] Phan AT, Goldrath AW, Glass CK (2017) Metabolic and Epigenetic Coordination of T Cell and Macrophage Immunity. *Immunity* 46(5):714–729.

[58] Korfali N, et al. (2010) The Leukocyte Nuclear Envelope Proteome Varies with Cell Activation and Contains Novel Transmembrane Proteins That Affect Genome Architecture. *Mol. Cell. Proteom.* 9(12):2571–2585.

[59] Pace MJ, et al. (2012) Directly infected resting CD4+T cells can produce HIV Gag without spreading infection in a model of HIV latency. *PLoS Pathog.* 8(7):15.

[60] Marenduzzo D, Micheletti C, Cook PR (2006) Entropy-driven genome organization. *Biophys. J.* 90(10):3712–21.

[61] Nagano T, et al. (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502(7469):59–64.

[62] Beagrie RA, et al. (2017) Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* 543(7646):519–524.

[63] Stevens TJ, et al. (2017) 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 544(7648):59–64.

[64] Bartholomae CC, et al. (2011) Lentiviral vector integration profiles differ in rodent postmitotic tissues. *Mol. Ther.* 19(4):703–710.

[65] Solovei I, et al. (2009) Nuclear Architecture of Rod Photoreceptor Cells Adapts to Vision in Mammalian Evolution. *Cell* 137(2):356–368.

[66] Marini B, et al. (2015) Nuclear architecture dictates HIV-1 integration site selection. *Nature* 521(7551):227–231.

[67] Boettiger AN, et al. (2016) Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* 529(7586):418–422.

[68] Nozawa RS, et al. (2017) SAF-A regulates interphase chromosome structure through oligomerisation with chromatin-associated RNAs. *Cell.*

[69] Kind J, et al. (2013) Single-cell dynamics of genome-nuclear lamina interactions. *Cell* 153(1):178–192.

[70] Imai R, Nozaki T, Tani T, Kaizu K, Hibino K (2017) Density imaging of heterochromatin in live cells using orientation-independent-DIC microscopy. *Mol. Cell. Biol.* pp. 1–57.

[71] Sowd GA, et al. (2016) A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc. Nat. Acad. Sci. USA* 113(8):E1054–E1063.

[72] Peng K, et al. (2014) Quantitative microscopy of functional hiv post-entry complexes reveals association of replication with the viral capsid. *Elife* 3.

[73] Ballandras-Colas A, et al. (2016) Cryo-EM reveals a novel octameric integrase structure for betaretroviral intasome function. *Nature* 530(7590):358–361.

[74] Grawenhoff J, Engelman AN (2017) Retroviral integrase protein and intasome nucleoprotein complex structures. *World J. Biol. Chem.* 8(1):32.

[75] Cavalli G, Misteli T (2013) Functional implications of genome topology. *Nat. Struct. Mol. Biol.* 20(3):290–9.

[76] Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* 295(February):1306–1312.

[77] Lieberman-Aiden E, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(october):289–93.

[78] Zirkel A, et al. (2017) Topological Demarcation By HMGB2 Is Disrupted Early Upon Senescence Entry Across Cell Types And Induces CTCF Clustering. *bioRxiv* 144.

[79] Plimpton S (1995) Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comp. Phys.* 117(1):1–19.

[80] Sides SW, Grest GS, Stevens MJ, Plimpton SJ (2004) Effect of end-tethered polymers on surface adhesion of glassy polymers. *Journal of Polymer Science Part B: Polymer Physics* 42(2):199–208.

[81] Bosse JB, et al. (2015) *Remodeling nuclear architecture allows efficient transport of herpesvirus capsids by diffusion.* Vol. 112, pp. E5725–E5733.

Supporting Information

Computational Details

To model DNA and chromatin we here consider a broadly employed coarse-grained model for biopolymers that has been successfully used in the past to describe the behaviour of DNA and chromatin in vivo and in vitro [29, 30, 44, 49]. In this model, DNA and chromatin are considered as semi-flexible bead-spring chains made of M beads. Each bead has a diameter of σ, which is taken to be $\sigma = 2.5$ nm (or 7.35 bp) for DNA and $\sigma = 10$ nm for chromatin. We simulate the dynamics of the fibre by performing molecular dynamics (MD) simulations in Brownian scheme, i.e. we include a stochastic force on each monomer to implicitly account for the solvent and noisy environment.

As commonly done in MD simulations, we express properties of the system in multiples of fundamental quantities. Energies are expressed in units of k_BT, where k_B is the Boltzmann constant and T is the temperature of the solvent. Distances are expressed in units of σ, which, as defined above, is the diameter of the bead. Further, time is expressed in units of the Brownian time τ_{Br}, which is the typical time for a bead to diffuse a distance of its size – more precisely, $\tau_{Br} = \sigma^2/D$, where D is the diffusion constant for a bead.

The interactions between the beads are governed by several potentials that are standard in polymer physics. First, purely repulsive interactions are modelled by the standard Weeks-Chandler-Anderson potential

$$U^{ab}_{WCA}(r) = k_BT\left[4\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] + 1\right] \tag{6}$$

if $r < r_c = 2^{1/6}\sigma$ and 0 otherwise. Here, where r is the separation between the two beads and r_c is a typical cut-off to ensure that the interaction is repulsive. Second, bonds between consecutive beads are treated as finitely extensible (FENE) springs:

$$U^{ab}_{FENE}(r) = -\frac{K_f R_0^2}{2}\ln\left[1 - \left(\frac{r}{R_0}\right)^2\right](\delta_{b,a+1} + \delta_{b,a-1}), \tag{7}$$

where R_0 (set to 1.6σ) is the maximum separation between beads and K_f (set to $30k_BT/\sigma^2$) is the strength of the spring. The combination of the WCA and FENE potentials with the chosen parameters gives a bond length that is approximately equal to σ [44]. Third, we model the stiffness of the polymers via a Kartky-Porod term:

$$U^{ab}_{KP} = \frac{k_BTl_p}{\sigma}\left[1 - \frac{\boldsymbol{t}_a\cdot\boldsymbol{t}_b}{|\boldsymbol{t}_a|\,|\boldsymbol{t}_b|}\right](\delta_{b,a+1} + \delta_{b,a-1}), \tag{8}$$

where $\boldsymbol{t}_a$ and $\boldsymbol{t}_b$ are the tangent vectors connecting bead a to $a+1$ and b to $b+1$ respectively; l_p is the persistence length of the chain and is set to $l_p = 20\sigma = 50$ nm for DNA and to $l_p = 3\sigma \approx 30$ nm for chromatin [76] if not otherwise stated.

When needed, attractive interactions are modelled via a standard Lennard-Jones potential

$$U^{ab}_{LJ}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} - \left(\frac{\sigma}{r_c}\right)^{12} + \left(\frac{\sigma}{r_c}\right)^{6}\right] \tag{9}$$

if $r \le r_c = 1.8\sigma$ and 0 otherwise.

To summarise, the total potential energy related to bead a is the sum of all the pairwise and triplet potentials involving the bead:

$$U_a = \sum_{b\neq a}\left(U^{ab}_{WCA} + U^{ab}_{FENE} + U^{ab}_{KP} + U^{ab}_{LJ}\right). \tag{10}$$

The time evolution of each bead along the fibre is governed by a Brownian dynamics scheme with the following Langevin equation:

$$m_a\frac{d^2\boldsymbol{r}_a}{dt^2} = -\nabla U_a - \gamma_a\frac{d\boldsymbol{r}_a}{dt} + \sqrt{2k_BT\gamma_a}\boldsymbol{\eta}_a(t), \tag{11}$$

where m_a and γ_a are the mass and the friction coefficient of bead a, and $\boldsymbol{\eta}_a$ is its stochastic noise vector obeying the following statistical averages:

$$\langle\boldsymbol{\eta}(t)\rangle = 0; \quad \langle\eta_{a,\alpha}(t)\eta_{b,\beta}(t')\rangle = \delta_{ab}\delta_{\alpha\beta}\delta(t-t'), \tag{12}$$

where the Latin indices represent particle indices and the Greek indices represent Cartesian components. The last term of Eq. 11 represents the random collisions caused by the solvent particles. For simplicity, we assume all beads have the same mass and friction coefficient (i.e. $m_a = m$ and $\gamma_a = \gamma$). We also set $m = \gamma = k_B = T = 1$. The Langevin equation is integrated using the standard velocity-Verlet integration algorithm, which is performed using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) [79]. We set the integration time step to be $\Delta t = 0.001\,\tau_{Br}$, where τ_{Br} is the Brownian time as mentioned previously.

Finally, the recombination moves are performed using a in-house modified versions of the algorithm implemented in LAMMPS as `fix bond/swap` [80]. The modifications are tailored to our specific model, i.e. they allow us to perform recombination moves only between HIV and host polymers (inter-chain recombination), avoid intra-chain (or "self") reconnections and to perform integration moves bypassing the Metropolis test (see below). These codes are freely available and can be requested directly from one of the authors. The recombination moves are attempted at every timestep, with probability 0.5 and between beads that are at most $R_c = 2\sigma$ apart.

Integration Algorithm

In this section we discuss in more detail the integration algorithm used in this work (also see main text). The

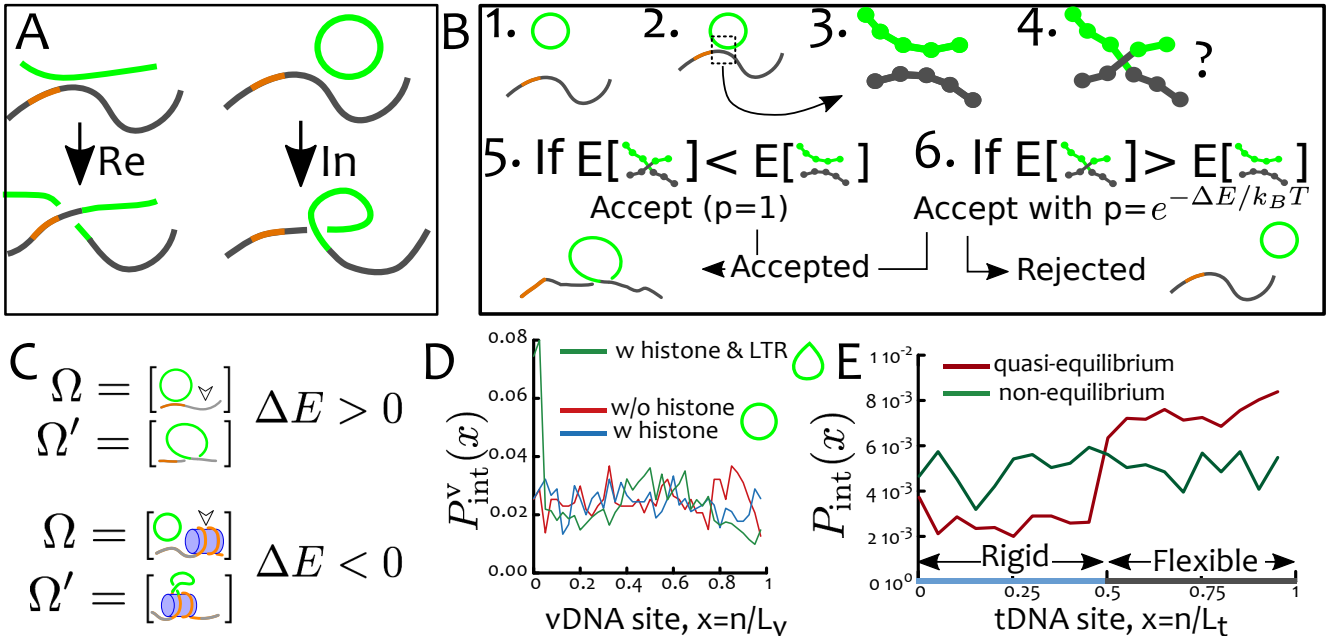

FIG. 6. **A** Schematics of recombination and integration events. **B** Model for integration as a quasi-equilibrium process. It can be divided in steps: (1-2) the viral DNA finds a candidate integration site by diffusive search; (3-4) an integration move is performed; (5) the difference in configurational energy is computed and the integration event is accepted if $\Delta E < 1$ or with probability $p = \exp[-\Delta E/k_BT]$. (6) If rejected the viral DNA can continue its diffusive search for the next candidate. **C** The preference for HIV integration in nucleosomal DNA can be explained by considering this DNA is tightly wrapped and hence heavily bent. For this reason an integration event is likely to reduce the bending of this segment. On the contrary integration in naked DNA must locally increase the bending. **D** Integration probability along the viral DNA. In our simplified model, any site along the viral DNA (of length $L_v = 40\sigma = 294$ bp) can be used as integration point. Considering a more refined model where the first bond along the viral DNA is kinked (thus mimicking the presence of intasome joining the long-terminal repeeats, or LTR, [33]) naturally favours the integration at that site (green line in the plot). At the same time, integration profile in the tDNA is unaltered (see main text). **E** Integration profile in a target DNA segment with heterogeneous flexibility (of length $L_t = 200\sigma = 1470$ bp). The flexible part is naturally favoured by our quasi-equilibrium model since the energy barrier for integration is lower. This is in agreement with experiments [10]. When a purely non-equilibrium model is used, i.e. no ΔE is computed and every event is accepted, the integration probability is uniform. Integrations profiles are obtained by averaging at least 1000 integration events.

integration of viral DNA into the host is a complex process that requires many steps [14]. In our model we summarise this complexity into two main processes: diffusive search of integration site and quasi-equilibrium integration event.

The former process is natural to account for as both target and viral DNAs are polymers diffusing in a noisy environment. Random and diffusive search has been seen in HIV integration in vitro [32] and it is very likely that it also occurs in vivo with even more pronounced complexity given the heterogeneous nuclear composition.

The latter, quasi-equilibrium integration, is our only assumption for the integration event: We start from the observation that HIV integration needs to locally deform the substrate in order to perform the double-strand cleavage and integration; then, we reason that in order to deform the substrate, the integrase enzyme needs to either expend some energy or to exploit thermal fluctuations that facilitate local deformations. In either case, we model this process by computing the total configurational energy immediately before (E) and immediately after (E') the integration move and compute the difference $\Delta E = E' - E$. If $\Delta E < 0$ then this means that the integration move is locally energetically favoured and the move is accepted. On the contrary, if $\Delta E > 0$ we accept it with probability $p = \exp -[\Delta E/k_BT]$. This effectively mimics thermal fluctuations (of order k_BT) that can induce spontaneous local deformations and thus facilitate integration. If the move is rejected the viral DNA resumes the diffusive search (see Fig. 6B for a schematics).

Integration within Single Nucleosomes

In the first part of our work, we study the integration of viral loops within nucleosomes. The size of viral loops is not a crucial parameter in our model as the reconnection moves occur only "locally", i.e. between two short polymer segments, and irrespectively of the rest of the chain. This is compatible with experiments in vivo and in vitro. Indeed, in the latter case HIV DNA is generally much smaller than in vivo and yet it is still successfully integrated within DNA substrates [9]. For this reason, we consider viral DNA

(vDNA) as made of 20 or 40 beads (about 150-294 bp) for computational efficiency.
We model the presence of a histone protein as represented by a sphere of size $\sigma_h = 3\sigma = 7.5$ nm. By assigning an attractive interaction between the histone and a selected segment of the polymer, we can model the wrapping of the DNA around the histone core. This is done using the LJ potential in eq. 9 and setting $r_c = 1.8\sigma$ and $\epsilon = 4k_BT$. With this simple model we cannot control the handedness of the wrapping. On the other hand, this is not crucial for our argument. Notice that this model has been used in the past to describe in vitro chromatin reconstitution [27].
Because the integration process is weighted by computing the change in energy before and after the recombination event, the kinetics of integration becomes very slow when we set the persistence length to be $l_p = 20\sigma$. This is because any bending introduced during the recombination heavily contributes to increasing the conformational energy. To speed up the kinetics while maintaining a realistic persistence length, we employ an extensible bond potential, i.e. instead of FENE bonds we use harmonic bonds

$$U^{ab}_{harm}(r) = \frac{\kappa}{2}\left(x - x_0\right)^2, \tag{13}$$

where $x_0 = 1.1\sigma$ and $\kappa = 20k_BT$. This effectively allows bonds to be more easily stretched and *uniformly* lowers the overall energy barrier along the DNA. Importantly, by using this potential for all bonds along the DNA we are not introducing explicit biases in the integration into any specific region. Any bias in the integration statistics is a result of the presence of histone-like particle and the consequent local larger bending introduced within the nucleosome core (see main text Fig. 1).

The observed bias for nucleosomal DNA can be readily understood by considering the fact that this DNA is tightly wrapped around the histone and hence heavily deformed and storing a large bending energy. This is because the nucleosomal DNA is comparable to DNA's persistence length. If now one imagines using a much larger contour length to wrap the same histone octamer, the resulting configuration is much looser, less bent and thus storing less bending energy. For this reason, one can imagine that an integration event on a segment of nucleosomal DNA lowers the total configurational energy and it is thus an energetically favoured process. on the contrary, integration in naked DNA requires a local deformation that locally increases the bending energy of the tDNA and it is thus energetically disfavoured but possible because of thermal fluctuations (see Fig. 6C for a schematics).

A Refined Model for Viral DNA with LTR

Up until now, we have considered a model where any segment along the viral DNA can be selected for the integration event. Indeed we observe a uniform integration probability along the viral DNA. Although this may seem biologically inaccurate, it is done for computational efficiency as it speeds up the search process. Restricting the integration event to occur in one segment on the vDNA would not change our results.

An improvement in this respect can be done as follows: we can consider a refined model where the viral DNA is now modelled as a loop with a "kink". More specifically, we set a fixed persistence length equal to $l_p = 50$ nm for all the segments of the viral DNA apart from one, where we set $l_p = 0$. By doing so we effectively allow the angle between beads $N-1, 0, 1$ to assume any sterically accessible conformation without paying any energy penalty. In practice, this translates into the viral DNA often displaying a "kink" at this location. This kink can be seen as the point where the intasome is located and where it keeps the long-terminal-repeats (LTR) together [33].

By performing integrations with this model we observe no change in the integration profile along the target DNA, as expected (indeed it is only the deformation of the substrate that determines the integration profile along the tDNA). Yet we observe a difference in the integration profile along the *viral* DNA. This is because the segment with zero persistence length is more easily deformable with respect to others. For this reason we expect, and observe, a marked increase in integration probability in this segment (see Fig. 6D). See Suppl. Mmovie 2 for an integration event with kinked HIV.

Integration in naked DNA with Heterogeneous Flexibility

In this section we discuss our results for integrations in substrates with heterogeneous flexibility. We consider a segment of naked DNA 200 beads long (about 1470 bp or 500 nm): the first half of the segment is rigid (persistence length $l_p = 50$ nm) while the second half is flexible (persistence length $l_p = 30$ nm). As one can see in Fig. 6E our simulations show that the integration of a 40 beads (294 bp) viral DNA is favoured in the flexible region, as seen experimentally [10]. This finding can be understood again in terms of energy barriers. The flexible regions allow more easily local deformations that are exploited by the integration machinery to integrate the viral DNA into the substrate.

Importantly, our model can account for this preference and for that shown for nucleosomal DNA with one simple assumption, i.e. that the HIV integration is a quasi-equilibrium process that needs to overcome an energy barrier to integrate the viral genetic material into the host (through local deformation of the substrate).

Non-Equilibrium Integration

In this section we discuss an alternative integration strategy that is fully non-equilibrium. Here, HIV integration maintains its diffusive search but it can perform integration within the host without being restricted to the Metropolis test. In other words, any integration move is always accepted. This strategy can be understood as mimicking the fact that an enzyme *may* consume ATP to actively deform the substrate and integrate the viral DNA without the need to wait for thermal fluctuations of the substrate.

We repeat the simulations performed in the previous section (in naked DNA substrates with heterogeneous flexibility) and report our findings in Figure 6E. As one can notice, this time the integration profile is uniform, i.e. the preference towards flexible regions is lost. This finding (flat integration profile in naked DNA with heterogeneous flexibility) is not seen in experiments considering HIV, which instead find a preference for flexible (or curved) DNA [10]. For this reason, we argue that our model for HIV integration as quasi-equilibrium process is a better model for real HIV integration as this gives integration profiles that are in quantitative agreement with experiments (see main text Fig. 1). Yet,our model predicts that retroviruses that do not require to bend the substrate or that employ ATP to deform the substrate should display integration profiles that are insensitive to the underlying DNA flexibility.

Integration within Nucleosomal Arrays

In the main text (Fig. 2) we consider a model in which several histones form a short reconstituted chromatin fibre. In formulating this model we are inspired by the idea that the chromatin fibre is not a static, crystalline structure but that it is a dynamic, "soft" assembly that minimises energetic costs and that can display heteromophic conformations [42]. Our model is motivated by in vitro [39] and in vivo [41] observations that chromatin assumes a range of possible structures rather than one typical structure.

In our model, the substrate DNA is made of 290 beads with 10 nucleosomal segments 20 beads (160 bp) long interspersed with linker DNA 10 beads long (80 bp). This choice considers a linker DNA slightly longer than the typical one in eukaryotes; yet this accelerates the self-assembly of the chromatin fibre as there is a smaller energy penalty to be paid to loop longer linker DNA.

The reconstitution is simulated by using a model where each of the 10 histones in the system experiences a short-ranged attraction (as before mediated by the LJ potential with $r_c = 1.8\sigma$ and $\epsilon = 4k_BT$) with a specific nucleosomal segment. In other words, for 10 nucleosomal segments and 10 histone-like particles, we assume attractive interactions between each pair (1-1, 2-2, 3-3, ...) and purely steric interactions with all other possible pairs (e.g., 1-2, 2-4, 5-9, ...). Although this is far from realistic (as histones do not have one binding site in the entire genome) we are not here interested in the process of reconstitution *per se* but on the integration statistics within a reconstituted chromatin fibre acting as substrate. For this reason, we employ this simple strategy to reconstitute many (> 1000) chromatin fibres in open, partially folded and condensed states and study the statistics of HIV integration on these different substrates with varying local chromatin structure (see main text). These states are generated as follows: in the open chromatin state histone-like particles display purely repulsive interactions between one another (i.e. via WCA potential described above with $\sigma_{\text{histone}} = 3\sigma$). Starting from this open state, we then include nearest-neighbour attractions $(i, i \pm 1)$between the histone-like particles: depending on the strength of the attraction this leads to either partially folded fibres (nnp, $r_c = 9\sigma = 3\sigma_{\text{histone}} = 22.5$ nm and $\epsilon = 40k_BT$) or fully condensed fibres (nnf, $r_c = 9\sigma = 3\sigma_{\text{histone}} = 22.5$ nm and $\epsilon = 80k_BT$). Alternatively, we also consider next-nearest-neighbour attractions which lead to zig-zagging fibres (nnn, $r_c = 9\sigma = 3\sigma_{\text{histone}} = 22.5$ nm and $\epsilon = 60k_BT$). These large values of binding affinity are required to stabilise short DNA loops which are formed by the linker DNA (10 beads is half the persistence length of DNA in this model) and overcome steric hindrance of the DNA.

We stress that although this model may be seen as oversimplified, it allows us to reproduce a range of chromatin substrates with conformations that are not too far from the heteromorphic structures seen in vitro or in vivo [39, 41, 42]. More sophisticated and realistic models for chromatin reconstitution employing patchy-particles can in principle be used (see Ref. [27]) but the main physical drivers of HIV integration, i.e. bending and accessibility, are already fully captured by our simplified model.

Random Walk Model for Viral Integration in Nuclei

In this section we discuss the random walk model to describe the behaviour of viral DNA entering cell nuclei. We assume that the viral DNA enters the nucleus from the nuclear envelope and begins an unbiased Brownian walk with diffusion coefficient D, i.e. satisfying the Langevin equation

$$\boldsymbol{r}_{t+\Delta t} = \boldsymbol{r}_t + \sqrt{2Dk_BT\Delta t}\boldsymbol{\eta} \tag{14}$$

where $\boldsymbol{\eta}$ is a vector of Gaussian numbers with zero mean and unit variance. Furthermore, we describe the integration process as happening at rate κ. In other words, at each time-step we draw a random number and if it is smaller than $\kappa\Delta t$, we stop the random walk and record its position among the "integrated" viruses.

We average over typically 10^5 random walks starting from positions uniformly distributed on the surface of a

sphere of radius R and obtain a distribution of integration events. This distribution is then binned in the radial location r of the integrated event and the count for each bin is divided by the area of the shell at position r, i.e. $4\pi r^2 dr$. This distribution is finally normalised so that its integral over the range $[0, R]$ is unity.

In this model, we take $\sigma = 50$ nm and $R = 200\sigma = 10\,\mu m$. By taking the viscosity of the nucleoplasm at $\eta = 150cP$ we can then obtain $\tau_{br} = 50$ ms. We set $D = 0.05\,\mu m^2/s$, close to the diffusivity of viral capsids in nuclei [81] and choose $\kappa = 0.002\,s^{-1}$ (a parameter largely unknown). The HIV penetration length is thus $l = \sqrt{D/\kappa} = 5\,\mu m$. Importantly, Eq. (14) can be simulated with spatially varying diffusion coefficient and reaction rates.

Solution of the Reaction-Diffusion Equation for Viral Integration in the Nucleus: Uniform case

In this section we describe the solution of the reaction-diffusion equation

$$\frac{D}{r^2}\partial_r\left(r^2\partial_r\rho(r)\right) - \kappa\rho = 0 \tag{15}$$

with D and κ constants. The equation can be written as

$$\partial_{rr}\rho + \frac{2}{r}\partial_r\rho - \frac{\kappa}{D}\rho = 0\,. \tag{16}$$

By assuming that the solution takes the form of $\rho(r) = r^n f$ we can write

$$\begin{aligned} &r^n\partial_{rr}f + 2\left[n+1\right]r^{n-1}\partial_r f + \\ &+\left[2nr^{n-2} + n(n-1)r^{n-2} - \frac{\kappa}{D}r^n\right]f = 0 \\ &r^2\partial_{rr}f + 2\left[n+1\right]r\partial_r f - \left[+r^2\frac{\kappa}{D} - n(n+1)\right]f = 0\,. \end{aligned}$$

By setting $n = -1/2$ we find

$$r^2\partial_{rr}f + r\partial_r f - \left[r^2\frac{\kappa}{D} + \frac{1}{4}\right]f = 0\,, \tag{17}$$

which is the modified Bessel equation

$$x^2\partial_{xx}y + x\partial_x y - \left[x^2 + m^2\right]y = 0 \tag{18}$$

with $m = 1/2$, $x = r/l$ and $l = \sqrt{D/\kappa}$. The generic solution of this equation is

$$y = C_1 I_m(x) + C_2 K_m(x)\,, \tag{19}$$

with $I_m(x)$ and $K_m(x)$ the modified Bessel functions of order m of the first and second kind respectively. Using this generic form, the steady state distribution $\rho(r) = r^{-1/2}f$ is thus

$$\rho(r) = C_1\frac{I_{1/2}(r/l)}{\sqrt{r}} + C_2\frac{K_{1/2}(r/l)}{\sqrt{r}}\,. \tag{20}$$

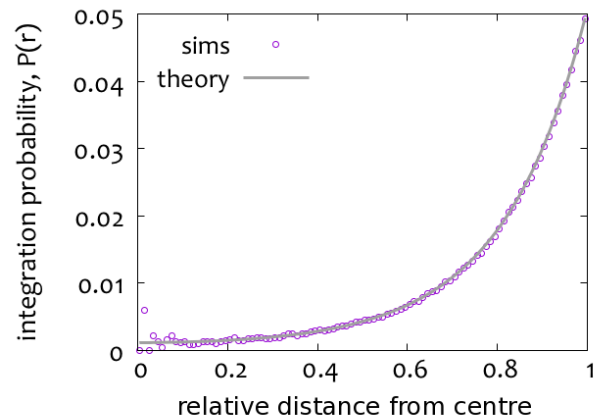

FIG. 7. Probability distribution $\rho(r)$. Data points are obtained averaging over $5\,10^5$ random walks starting from the surface of a sphere of radius R and allowed to diffuse and react within the sphere at constant D and κ. The line is the prediction obtained from the analytic solution to the reaction-diffusion problem (see eq. (21)).

Here, we note that $\rho(r)$ must not diverge at $r \to 0$; this entails that $C_2 = 0$ and we thus find

$$\rho(r) = \frac{\sinh\left(r/l\right)}{r\,\mathrm{shi(R/l)}} \tag{21}$$

with the normalisation hyperbolic sine integral $\mathrm{shi(t)} \equiv \int_0^{\mathrm{R}}\sinh(\mathrm{t})/\mathrm{t}\,\mathrm{dt}$ obtained imposing $\int_0^R \rho(r)dr = 1$.

Alternatively, one can define $\rho = f/r$ and find a simple equation for f leading to the following solution of Eq. (16)

$$f = Ae^{-\sqrt{\kappa/D}r} + Be^{\sqrt{\kappa/D}r} \tag{22}$$

with A and B coefficients to be determined. In order for this solution to be well behaved at $r \to 0$ we need $A = -B$, reobtaining Eq. (21).

Reaction-Diffusion Equation in Heterogeneous Nuclei

In this case we need to solve the system of equations:

$$\begin{aligned} &D_1\left(\partial_{rr}\rho_1(r) + \frac{2\partial_r\rho_1(r)}{r}\right) - k_1\rho_1(r) = 0 \\ &D_2\left(\partial_{rr}\rho_2(r) + \frac{2\partial_r\rho_2(r)}{r}\right) - k_2\rho_2(r) = 0 \\ &D_3\left(\partial_{rr}\rho_3(r) + \frac{2\partial_r\rho_3(r)}{r}\right) - k_3\rho_3(r) = 0 \end{aligned} \tag{23}$$

with the continuity conditions

$$\begin{aligned} &\rho_1(R_1) = \rho_2(R_1)\ ,\ \rho_2(R_2) = \rho_3(R_2) \\ &\partial_r\rho_1(R_1) = \partial_r\rho_2(R_1)\ ,\ \partial_r\rho_2(R_2) = \partial_r\rho_3(R_2) \\ &\partial_r\rho_1(\epsilon \to 0) = 0\,. \end{aligned}$$

In these equations, R_1 and R_2 are the position of the boundaries between zones 1-2 and 2-3 respectively. We solve this system of equations with Mathematica `DSolve` which gives

$$\rho_1 = \mathcal{N}\frac{\sinh\left(r/l_1\right)}{r}$$
$$\rho_2 = \frac{\mathcal{N}}{l_1 r}\left[l_1 \sinh\left(\frac{\mathrm{R1}}{l_1}\right)\cosh\left(\frac{r-\mathrm{R1}}{l_2}\right) + l_2\cosh\left(\frac{\mathrm{R1}}{l_1}\right)\sinh\left(\frac{r-\mathrm{R1}}{l_2}\right)\right]$$
$$\rho_3 = \frac{\mathcal{N}}{4\kappa_2 l_1 l_2 r} e^{-\frac{r+\mathrm{R2}}{l_3}-\frac{\mathrm{R1}+\mathrm{R2}}{l_2}-\frac{\mathrm{R1}}{l_1}}\left\{\kappa_2\left[l_2 l_3\left(e^{\frac{2\mathrm{R1}}{l_1}}+1\right)\left(e^{\frac{2r}{l_3}}-e^{\frac{2\mathrm{R2}}{l_3}}\right)\left(e^{\frac{2\mathrm{R1}}{l_2}}+e^{\frac{2\mathrm{R2}}{l_2}}\right)+8l_1\sinh\left(\frac{\mathrm{R1}}{l_1}\right)e^{\frac{r+\mathrm{R2}}{l_3}+\frac{\mathrm{R1}+\mathrm{R2}}{l_2}+\frac{\mathrm{R1}}{l_1}}\right.\right.$$
$$\left.\left.\left(l_2\cosh\left(\frac{r-\mathrm{R2}}{l_3}\right)\cosh\left(\frac{\mathrm{R1}-\mathrm{R2}}{l_2}\right)-l_3\sinh\left(\frac{r-\mathrm{R2}}{l_3}\right)\sinh\left(\frac{\mathrm{R1}-\mathrm{R2}}{l_2}\right)\right)\right]-D_2\left(e^{\frac{2\mathrm{R1}}{l_1}}+1\right)\left(e^{\frac{2r}{l_3}}+e^{\frac{2\mathrm{R2}}{l_3}}\right)\left(e^{\frac{2\mathrm{R1}}{l_2}}-e^{\frac{2\mathrm{R2}}{l_2}}\right)\right\}$$

where $l_i = \sqrt{D_i/\kappa_i}$ and $\mathcal{N}$ is a normalisation factor that can be set by imposing

$$\int_0^R \left[\Theta(x,0,R_1)\rho_1(r) + \Theta(x,R_1,R_2)\rho_2(r) + \Theta(x,R_2,R)\rho_3(r)\right] dr = 1 \tag{24}$$

where $\Theta(x,b_1,b_2)$ is here defined as unity if $b_1 \leq x < b_2$ and zero otherwise.

RESIZING NUCLEAR SHELLS

In order to obtain quantitative agreement between our model and the data from experiments [66] we observe that we need to reshape the concentric regions in the nucleus of a model T-cell. First we fix the mass (or amount of genetic material) in heterochromatin to be twice the one in euchromatin, i.e.

$$2\rho_{\mathrm{eu}}V_{\mathrm{eu}} = \rho_{\mathrm{het}}V_{\mathrm{het}}\,. \tag{25}$$

We further impose that the volume occupied by heterochromatin is the inner and outer layer, whereas the one of euchromatin is the middle one, i.e.

$$V_{\mathrm{het}} = \frac{4}{3}\pi R_1^3 + \left(\frac{4}{3}\pi R^3 - \frac{4}{3}\pi R_2^3\right)$$
$$V_{\mathrm{eu}} = \frac{4}{3}\pi R_2^3 - \frac{4}{3}\pi R_1^3\,. \tag{26}$$

In these equations, R_1 and R_2 are the boundaries between the first and second layers and between the second and the third, respectively. We now insert eq. (26) into eq. (25) and can solve for $\rho_{\mathrm{eu}}/\rho_{\mathrm{het}}$:

$$\frac{\rho_{\mathrm{het}}}{\rho_{\mathrm{eu}}} = \frac{2\left(R_2^3 - R_1^3\right)}{R^3 + R_1^3 - R_2^3}\,. \tag{27}$$

obtaining the equation reported in the main text.