Physical Principles of HIV Integration

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

Understanding the mechanisms underlying the non-random pattern of HIV integration within a host genome remains an unsolved problem in biophysics. To address this issue, we formulate a generic multi-scale physical framework which models viral incorporation into DNA and chromosomes as stochastic "reconnections" between 3D-proximal polymer segments. Our model naturally predicts that successful integration events occur more often within more flexible or highly bent DNA or genomic regions, in agreement with long-standing observations. We further find that large-scale chromosomal organisation biases integration events towards transcriptionally active genomic regions. We finally propose and solve a reaction-diffusion model which captures the observed distribution of HIV integration sites within the nuclear environment of T-cells. Our general framework can be used to predict HIV integration dynamics within different cell-lines or displaying different size, fraction of transcriptionally inactive genome and nuclear genomic arrangements.

Retroviruses are pathogens which infect organisms by integrating their DNA within the genome of the host. Once integrated, they exploit the transcription machinery of the host cell to proliferate and propagate itself into other host cells or organisms. This remarkable strategy provides a powerful way to ensure that the viral information remains engrained in the host cell, and gets transmitted to its daughters when it divides. It is also the reason why between 5% and 10% of human DNA 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 [1–4]. These ancient viral repeats are a potential danger for the cell, and the gene regulatory network has evolved so as to keep most of these sequences transcriptioanlly silent, so that they are inactive at all times (as activation usually leads to disease such as cancer [4]). At the same time, retroviruses are potentially appealing for clinical medicine, as they can be used as vectors for gene therapies [5–7].

Experiments have provided a wealth of important observations on the mechanism through which retroviruses work. First, classical experiments have shown that the retro-viral integration complex (or "intasome") displays a marked tendency to target bent DNA regions, so that integration occurs more easily within chromatin, where DNA bends to wrap tightly around histone octamers, rather than within naked DNA [8–14]. This is clearly advantageous for retroviruses which target eukaryotes, since their DNA is packaged into chromatin [1, 15]. Second, recent experiments employing genomic sequencing have shown that, for most retroviruses, a key feature that determines integration sites within chromosomes is the local chromatin state [16]. Thus, gammaretroviruses, deltaretroviruses and lentiviruses including HIV-1, display a strong preference for transcriptionally active chromatin [16, 17].

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 biophysical models that investigate the generic biophysical mechanisms of integration; because many aspects of this problem appear to be shared across several genera of retroviruses (e.g., the preference for bent DNA and open euchromatin), such an approach may provide a useful complement to experiments. Therefore, here we propose a multi-scale biophysical model for retroviral integration in host cells, focussing on the case of HIV. First, we introduce and study a polymer physics model in which retroviral DNA and host genomes are coarse grained into semi-flexible polymers, and integration is modelled as a "reconnection" move, which consists in swapping (i.e., cutting and rejoining) two DNA/chromatin segments as they become proximal in 3D. Then, at a larger scale, we propose and exactly solve a reaction-diffusion problem to study HIV integration within a human nucleus.

At all scales considered, ranging from that of nucleosomes (< 100 nm) to that of the cell nucleus ($> 10\,\mu$m), our model compares favourably with experiments, both qualitatively and quantitatively. Our model can thus provide new generic mechanistic insight into the physical principles driving the preferential integration of HIV into DNA and chromatin. It suggests that simple reconnection moves, based on simple physical features such as local DNA elasticity and chromatin accessibility, are sufficient to explain most of the existing experimental data on HIV integration. The model also yields several predictions, which can be tested in future experiments.

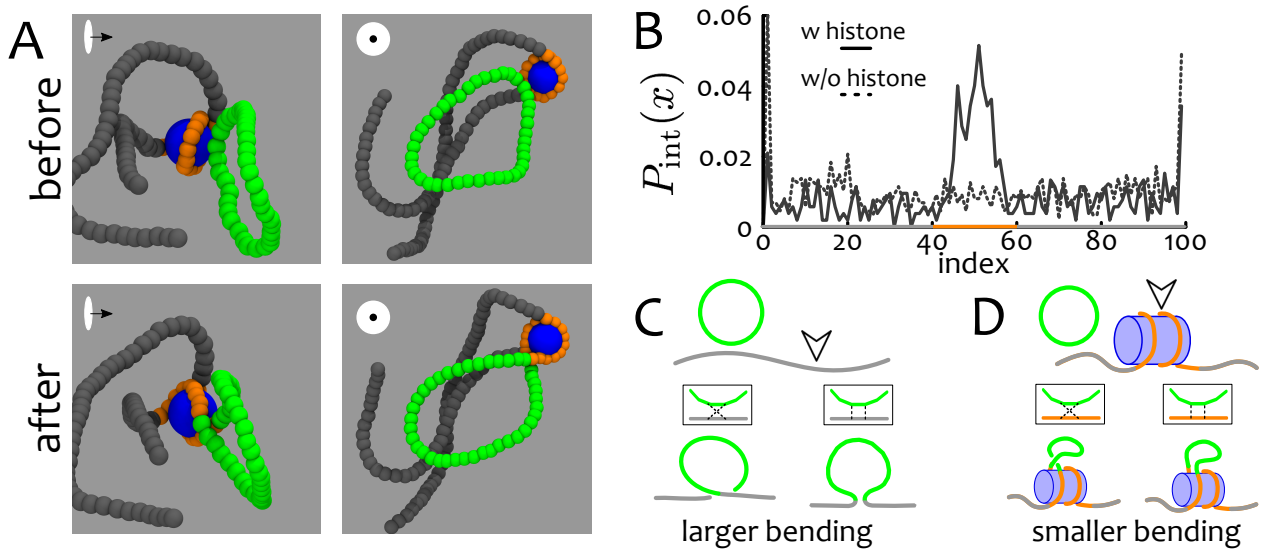

FIG. 1. **A** Snapshots from Brownian Dynamics simulations capturing before and after an integration event (left:side view, right:top view). The viral DNA (green) integrates on a DNA segment (grey and orange) wrapped around a histone-like particle (blue). The dynamics of this simulation is shown in Suppl. Movie 1. **B** The integration probability $P_{\rm int}(x)$ as a function of the bead index, x, is ∼5-fold enhanced in the region wrapped around the histone-like protein (average over 1000 replicas). The integration profile is also observed to peak at the two ends of the host DNA. **C-D** Cartoons describing the change in local curvature during integration events. These are likely to reduce the local curvature in regions that are wrapped around histones, whereas they are likely to increase it within naked DNA.

Nucleosomes and Highly Bent Regions of DNA are Preferred Integration Sites

Long-standing evidence strongly suggest that the intasome complex preferentially targets highly bent regions of DNA, in particular those wrapped around histones, rather than naked DNA [8, 9]. To shed light on the mechanism underlying this bias, we perform simulations in which both viral and host DNAs are coarse-grained into semi-flexible bead-spring polymers with beads of size $\sigma = 2.5$ nm (or about 8 base-pairs) connected by harmonic springs ($\kappa = 10$ k_BT and $r_0 = 1.1\sigma$, see SI) and with persistence length $l_p = 20\sigma = 50$ nm, as appropriate for naked DNA. We include a histone-like particle of size $\sigma_h = 3\sigma = 7.5$ nm which interacts with a section of the host DNA (20 beads, or 160 bp, in the middle of a $N = 100$ bead, or 800 bp, polymer) via an attractive short-ranged attraction ($\epsilon = 4$ k_BT and $r_c = 1.8(\sigma_h + \sigma)/2$, see SI) [18, 19]. This drives the self-assembly of a nucleosome-like structure where the host DNA wraps around the histone-like particle. [Notice that we do not consider torsional stiffness of DNA and that the model nucleosome can be wrapped in a right- or left-handed configuration, in contrast with real chromatin where the wrapping is always left-handed. This detail is irrelevant for our conclusions.]

As the intasome complex does not consume ATP [6], we choose to model reconnection moves – or "integration events" – as a close-to-equilibrium process, where acceptance is determined through a Metropolis test based on detailed balance and which penalises the creation of configurations with large internal energies, e.g., if a reconnection introduces extensive bending. Once the viral DNA is integrated, it is trapped in place and the topology of the system is quenched.

By performing Brownian Dynamics (BD) simulations of this model we find that most of the recombination events occur in the region of DNA that associates to the histone-like particle (orange segment in Fig. 1A). Further quantification reveals that the probability of successful integration event $P_{\rm int}(x)$ as a function of the genomic position x displays a ∼5-fold increase at the nucleosome with respect to the background probability $1/N$ (Fig. 1B). If no histone-like particle is included in the simulations, $P_{\rm int}(x)$ is approximately constant along the contour. In both cases, $P_{\rm int}(x)$ shows enhancement at the extremities of the host polymer, as they have less bending penalty.

Within our framework, the preference for nucleosomal DNA can be understood by noticing that the local bending of a naked DNA segment wrapped around a histone protein is large. Thus, any integration event is likely to reduce the local curvature and hence to be accepted as a valid recombination move in the Metropolis step (Fig. 1D). On the other hand, insertion of a circular DNA in a semi-flexible naked DNA segment is likely to create larger bending (Fig. 1C), and hence to be rejected. We conjecture that a similar mechanism is at play in the experiments, both *in vitro* and *in vivo*, as the local deformation of DNA in nucleosomes may lower the energy barrier required for the intasome to successfully integrate the viral DNA.

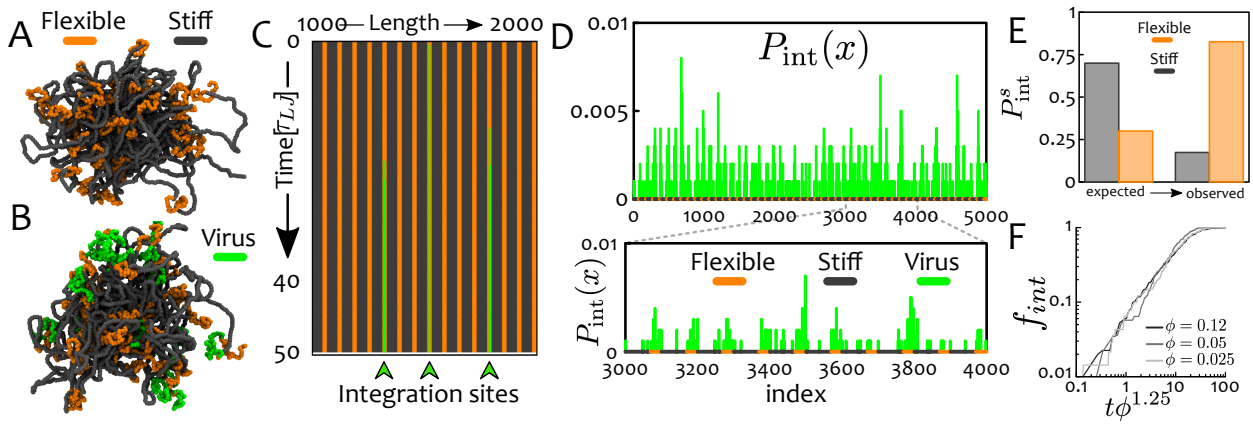

FIG. 2. **Di-block model for DNA and chromatin. A-B** Snapshots from BD simulations showing equilibrated configurations of the host genome before and after the integration events. **C** Kymograph showing multiple integration events occurring near one another in flexible (orange) segments. **D** The probability distribution of integration sites as a function of bead index, x, is strongly enriched within flexible regions genome-wide. **E** We observe that the probability of recording an integration event within a region marked as "s" (s=stiff or flexible), $P^s_{\rm int}$, displays a behaviour in marked contrast to that expected for random events (i.e., occurring with uniform probability throughout the host DNA). **F** The fraction of integrated loops scales with the overall density of the system as $f_{\rm int} \sim t^1 \phi^{-1.25}$, close to the scaling expected for the number of collisions of a gas of particles at density ϕ. The dynamics corresponding to one of our simulations is shown in full in Suppl. Movie 2.

A Di-Block Polymer Model for DNA and Chromatin Predicts Preferential Integration in Flexible Regions of the Genome

In order to probe the effect of DNA and chromatin flexibility on viral integration sites at a larger scale, we now consider a generic model for a host genome that displays a mixture of stiff and flexible regions [20, 21]. Our model genome is made of $N = 5000$ beads regularly partitioned into 50 blocks; each block features a rigid segment of $N_r = 70$ beads and a flexible one made of $N_f = 30$ beads. We model the latter as having a persistence length $l_p = 1\sigma$ ($\sim$ 3 or 10 nm for DNA or chromatin, respectively), and the former as having $l_p = 10\sigma$ ($\sim$ 30 or 100 nm for DNA or chromatin, respectively). [The precise values of persistence lengths do not matter for the steady state statistics; only the ratio of the two does.]

In the context of chromatin, the stiffer segments of the host polymer are meant to model regions of the genome that are transcriptionally silent and enriched in linker histone H1, such as gene deserts [22] (within these regions, it is reasonable to assume that H1 leads to local chromatin stiffening [20]). Flexible portions represent transcriptionally active regions, such as euchromatin [23, 24]. In Fig. 2A we show a typical equilibrium configuration of the host polymer, where the overall DNA/chromatin density is $\sim 10\%$ (kept using a periodic box and relevant for chromatin within mammalian nuclei [20, 25]). Current evidence suggests that viral DNA is chromatinised after entering the nucleus, i.e., it wraps around histones to form nucleosomes [26]. For this reason, we use the same bead size ($\sigma = 1kbp \simeq 10$ nm) and persistence length corresponding to open chromatin ($l_p = 1\sigma$) to represent the viral polymer. Within this framework, we perform BD simulations (see SI) of the integration dynamics of 20 viral loops each made of $N_v = 50$ beads (about 50 kbp, for chromatin). [Changing loop size does not change our results qualitatively.]

We observe that our model displays a strong bias favouring integration within flexible regions (Fig. 2B). In the context of naked DNA, this is in agreement with *in vitro* experiments studying intasome binding to DNA molecules with different sequences (hence different flexibility) [8, 9]. Within chromatin, this bias is reminiscent of the one favouring integration within transcriptionally active chromatin.

In Fig. 2C we also a kymograph of viral integration events, which highlights the occurrence of some intriguing "hierarchical" integration events, where viral DNA integrates either into, or close-by, already integrated viral DNA. At steady state (Fig. 2D), about 70% of the integrations occur in the flexible segments, even though they cover only the 30% of the host polymer (Fig. 2E).

We further analyse how the integration rate varies as a function of the system density (by varying the linear size of the periodic box, $L = 10$, 20 or 30σ). We find (Fig. 2F) that the number of integrated HIV molecules scales linearly with time while with a power $\alpha = 1.25$ as a function of density ϕ. Equivalently, this can be viewed as the total time needed for all loops to be integrated as scaling as $T_{\rm int} \sim \phi^{-1.25}$. This is close to the scaling expected for random collisions in a gas of particles at density ϕ, for which $T_m \sim \phi^{-1}$. The small deviation from this simple scaling may be due to correlation in the organisation of flexible and rigid segments under confinement.

Non-Equilibrium Integration Events Yield a Uniform Distribution of Integration Sites

To support our interpretation that it is the Metropolis test which leads to the biases we have reported thus far in our simulations, we compare our observations with the case in which any reconnection move is accepted, disregarding the energetic cost it entails. In this case, we observe that the integration dynamics is about 10-time faster than the one we described previously. Additionally, we find that the integration probability now follows a random (uniform) statistics, as expected. It is tempting to speculate that detailed balance violation (through the use of ATP), would not be a good infection strategy for the virus, as this may lead the intasome complex to frequently target heterochromatic regions. Since these are not actively transcribed, the viral DNA would not be able to proliferate.

A Polymer Model with Epigenetic Patterns Reveals that Large-Scale 3D Genome Organisation Enhances Euchromatic Integration

Polymer modelling of large-scale chromatin organisation in 3D has led to some breakthroughs in our current understanding of genome architecture *in vivo* [27–35]. Some of these models strongly suggest that epigenetic patterns of histone post-translational modifications – such as H3K4me3 or H3K9me3 – play a crucial role in folding the genome in 3D [28, 33, 34, 36]. Based on this evidence we thus ask whether a polymer model of viral integration in a chromatin fibre that is folded in 3D based upon its epigenetic pattern can give us some insight of how large-scale chromatin architecture determines the distribution of integration sites.

To do so, we start from an idealised block co-polymer model for chromatin in which 50 blocks of 100 beads are portioned into 30 "euchromatic" and 70 "heterochromatic" beads (fraction of heterochromatin $\phi_{\rm het} = 70\%$). In order to mimic heterochromatic compaction driven by multivalent bridges [22, 36], we endow heterochromatic beads with a weak self-attraction ($\epsilon = 3\ k_BT$). In contrast, we assume euchromatic beads to interact only by steric repulsion, for simplicity. This model naturally drives the phase-separation of the system into heterochromatic inactive "B" compartments decorated by swollen loops of euchromatin which appear as "A" compartments in the contact map (see Fig. 3A,B).

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 [7, 16, 17].

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 the euchromatic loops. The latter are therefore the regions which first encounter the diffusing viral DNA; a similar organisation is expected near nuclear pores, where "channels" of low density chromatin separate inactive and lamin-associated regions of the genome [37].

To quantitatively test our generic block co-polymer model, 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 they display high GC content [1] and high expression in the Jurkat cell line (see Fig. 3D, data available from ENCODE [38] and from Ref. [17]). The rest of the beads are then marked as heterochromatin (blue). The threshold for 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 $\epsilon = 3\ k_BT$) and ones which occur on a non-folded substrate (by imposing only repulsive interaction between any two beads).

Our results (Fig. 3E) 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 also Suppl. Movie 3). We finally compare (Fig. 3F) the disitribution of predicted integration sites with those detected by genome sequencing in Ref. [17] and find a strikingly good agreement (the p-value returned by the Spearman Rank test is 10^{-11} for folded and experimental datasets while 0.6 for unfolded and experimental datasets). Our results thus suggest that the large-scale organisation of chromosomes plays an important role in biasing the distribution of retroviral integration events and, in particular, it enhances euchromatic integrations.

The Abundance of Heterochromatin Affects the Statistics and the Rate of Integration Events

Different cell types are characterised by different global 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 [39, 40] and of heterochromatin (defined cytologically), is less likely to be infected by HIV with respect to an "activated" one.

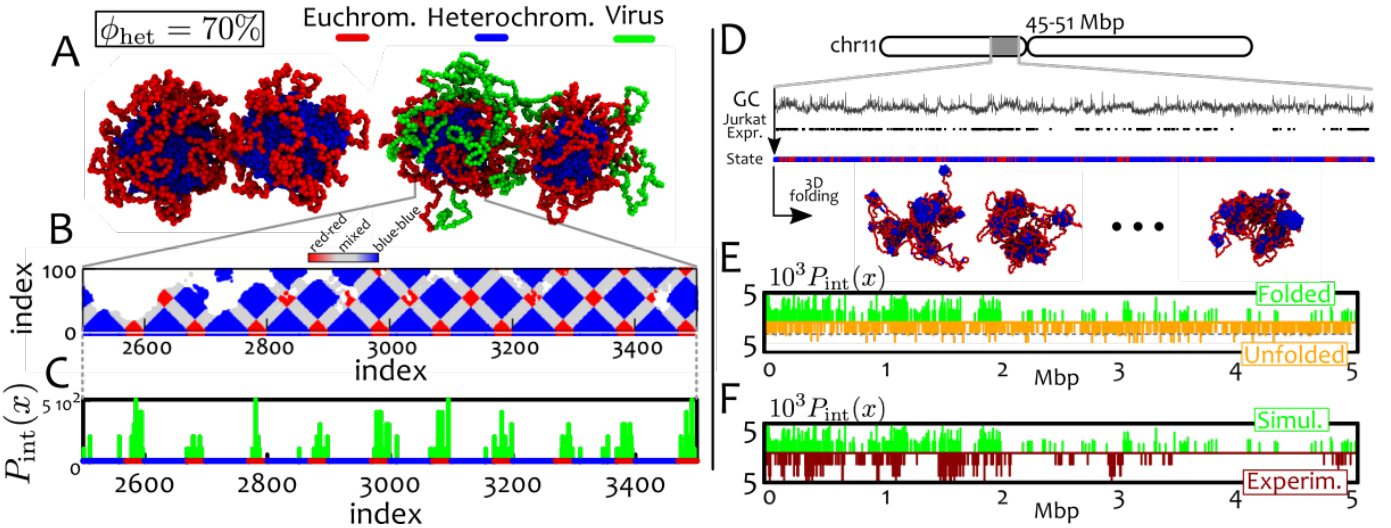

FIG. 3. **The 3D Folding of Chromatin Enhances Euchromatic Integration.** **A** Snapshots of our polymer model where the fraction of heterochromatin is $\phi_{\rm het} = 70\%$. We show two typical configurations, before and after integration events. **B** A section of the average contact map showing that heterochromatin and euchromatin regions create compartments. **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 non-random for folded chromatin. **F** Comparison between simulated and experimentally-measured distribution of integration sites in Jurkat T-cells [17]. The agreement between simulations and experiments is highly significant (p-value from Spearman Rank test is $p = 10^{-11}$, to be compared with $p = 0.6$ obtained from unfolded against experimental datasets). The dynamics corresponding to one of our simulations is shown in full in Suppl. Movie 3.

It is also known that the few resting cells which get infected do so after a sizeable delay [41].

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 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 hinder coalescence between inactive domains [20, 21]. For large heterochromatic content ($\phi_{\rm het}$ =80%), inactive domains merge to form a large central core, "decorated" by short euchromatic loops.

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-cell are much more difficult to infect dynamically 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 [42–44]. This is in qualitative agreement with experiments on lentivirus infection [41, 45], and more broadly suggests that possessing a high heterochromatin content provides a physical defence mechanism for the cell to slow down HIV infection.

A further surprising result is that the efficiency of viral integration in euchromatic parts of the genome increases with the total fraction of heterochromatin. In Fig. 4D we show the probability a viral loop is integrated into a region with epigenetic state "s", normalised by the fraction of the genome that is in that state, $\chi_s = P_{\rm int}(s)/\phi_s$. In other words, we plot the observed integration probabilities normalised by the ones expected for random (uniform) integration statistics. The figure shows that $\chi_{\rm eu}$ increases as a function of $\phi_{\rm het}$ while $\chi_{\rm het}$ decreases. This counter-intuitive observation can be understood as a direct consequence of chromatin architecture. The more heterochromatin is present in the nucleus, the stronger the inactive

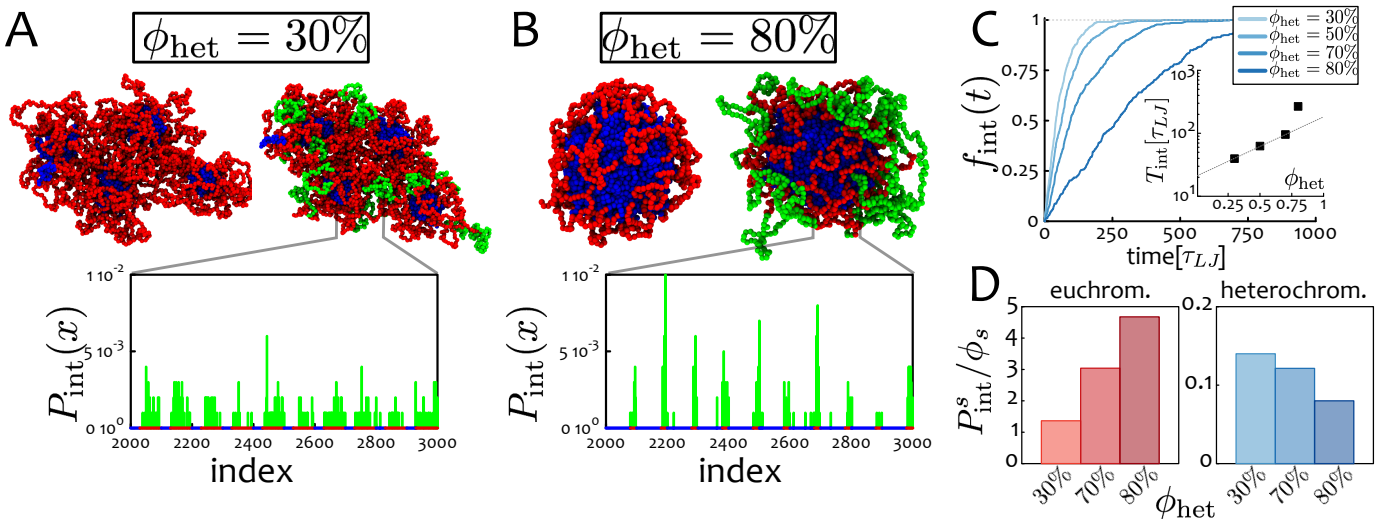

FIG. 4. **Integration Process is Affected by the Concentration of Heterochromatin. A-B** Snapshots and probability distribution for a system with $\phi_{\rm het} = 30\%$ and 80%, respectively. **C** Fraction of integrated loops $f_{\rm int}$ as a function of time, for different levels of heterochromatin. 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.

compartments are. As a consequence, the euchromatic regions are more strongly excluded by the "B" domains, which are thus more more effectively screened from integration.

A Reaction-Diffusion Model for Retroviral Integrations in Nuclei

In this section we formulate a yet larger-scale, continuum, model to predict the distribution of integration sites in whole 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.

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:

$$\partial_t \rho(\boldsymbol{x}, t) = \nabla \left(D(\boldsymbol{x}) \nabla \rho(\boldsymbol{x}, t) \right) - \kappa(\boldsymbol{x}) \rho(\boldsymbol{x}, t) \,. \qquad (1)$$

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.(1) 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

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

In the simplest case in which D and κ are uniform throughout the nucleus, the solution of eq. (2) is

$$\rho(r) = \mathcal{N} \frac{\sinh(r/l)}{r} \,, \qquad (3)$$

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

To solve eq. (1) in more general cases, we need to make some assumptions on how D and κ may vary in the nuclear environment. In line with our previous simulation 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 in euchromatin are both potentially reasonable choices: the former assumption describes situations where euchromatin is more open and less compact [23, 24]. Assuming slower diffusivity in euchromatin may instead model local gel formation by mesh-forming architectural proteins such as SAF-A [46]. 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

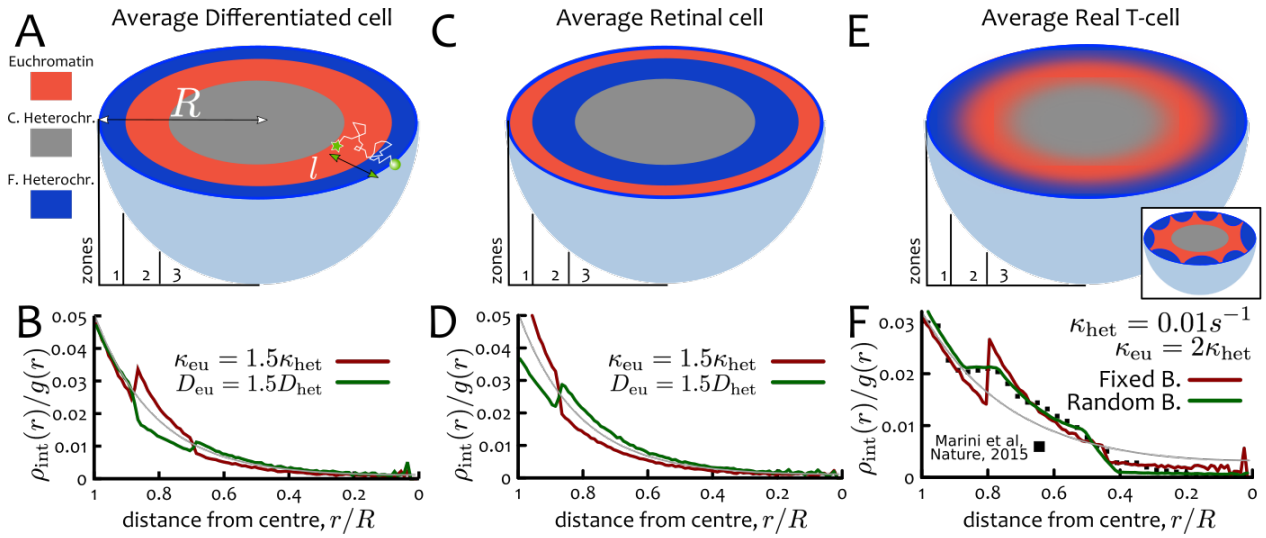

FIG. 5. **HIV Penetration length and Integration in the Nuclear Environment. A, C, E** Different cell lines display different chromatin organisations at the nuclear scale, and we consider this in our reaction-diffusion model for HIV integration in the nucleus. **A** We model a typical differentiated cells as a sphere with 3 concentric shells of equal volume. Zones 1 and 3 are populated by constitutive heterochromatin and lamin-associated-domains, 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** Model organisation in a "retinal" cell, with inversion of the two outer layers. **E** Model organisation within 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 for model differentiated cell (**B**), retinal (**D**) and T-cells (**F**). The result with uniform D and κ (yielding $l \simeq 5$ μm) is shown in gray 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.

outer zones are generally populated by transcriptionally inactive chromatin (heterochromatin and lamin-associated-domains, respectively [1, 47]), whereas the middle layer is commonly enriched in transcriptionally active euchromatin [48]. To mimic this organisation, and for simplicity, in our model D and κ vary spatially, but assume a constant value within each of the three layers. Whilst individual cells are known to display heterogeneity in the local chromosome organisation [42, 44], our model should capture trends corresponding to the distribution of integration sites across 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 cell", with a conventional heterochromatin-euchromatin-heterochromatin layering, with a "retinal cell", which displays an inverted euchromatin-heterochromatin-heterochromatin organisation (i.e., a switch between the middle and outer layer), and is observed in retinal cells of nocturnal animals [48] (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 show that the case with non-uniform D and κ predict deviations from the uniform case (see Fig. 5A-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 effects (as fast diffusion depletes virus concentration). We also predict that the distribution of integration events in "retinal" cells should be very different. Here, the larger $\kappa_{\rm eu}$ enhances the probability of integration near the periphery – as before, $D_{\rm eu}$ has the opposite effect.

More quantitatively, we now compare our model with the experimentally measured distribution of HIV recurrent integration genes (RIGs) in T-cells [49]. Remarkably, we find that our simple theory with uniform $D = 0.05$ μm$^2/s$ and $\kappa = 0.002\, s^{-1}$ (leading to a penetration length $l = \sqrt{D/\kappa} = 5\mu$m) is already in fair overall agreement with the experimental curve. The agreement can be improved by progressively refining our model. While adding more assumptions to match our theoretical results to the existing data, we also obtain insights 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 $\kappa_{\rm eu}/\kappa_{\rm het} \simeq 1.5$ (equivalently, a model with $D_{\rm het}/D_{\rm eu} \simeq 1.5$ and uniform κ would yield the same result). A second improvement is found by re-sizing the three concentric nuclear shells as follows. If the total amount of heterochromatin is maintained constant, at twice that of euchromatin (a realistic value *in vivo* [1]), the volume of the layers can change if we take

into account the fact that active chromatin is less dense than heterochromatin [23, 37]. It is possible to derive an equation relating the ratio between the density of heterochromatin and euchromatin, $\rho_{\rm het}/\rho_{\rm eu}$, to the positions of the boundaries between the first two and last two shells, R_{1-2} and R_{2-3}, as (see SI)

$$\frac{\rho_{\rm het}}{\rho_{\rm eu}} = \frac{2\left(R_{2-3}^3 - R_{1-2}^3\right)}{R^3 + R_{1-2}^3 - R_{2-3}^3}\,. \tag{4}$$

Thus, 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. [49]) we obtain $\rho_{\rm het}/\rho_{\rm eu} = 1.6$. Incidentally, we find that this value is in pleasing agreement with recent microscopy measurements (reporting a value of 1.53) [50].

A final refinement can be achieved by allowing for some small fluctuations ($\pm 0.5\,\mu$m) in the position of the boundaries in each simulated nucleus, and by 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 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, and thus denser, and largely devoid of DNA. With these refinements, our theory matches very closely the experimental measurements (Fig. 5F).

Discussion

In this work we propose a multiscale biophysical model for retroviral integration in DNA and chromatin, focussing on the case of HIV. Our simple model identifies two key mechanisms for the selection of integration sites within the host genome. First, the reconnection moves which we use to model integration are based on energy differences, so that integration events which lead to extensive DNA bending are less likely to occur with respect to moves which do not increase bending, or do so only moderately. Second, in our model the viral DNA samples space diffusively, hence steric accessibility is another key factor contributing to integration site selection. These two biophysical drivers are active at multiple length scales, and create trends which are in qualitative and quantitative agreement with experimental observations.

We observe a bias towards highly-bent or highly flexible regions of the genome, in quantitative agreement with long-standing experimental observations on single DNA molecules [9] (see Fig. 1). Further, we observe that integration of HIV occurs more favourably within transcriptionally active, open chromatin regions, as seen in the experiments [16]. Our model gives a mechanistic insight into this finding: the large-scale chromosomal organisation enhances the interaction of viral DNA with euchromatic regions, as they are most accessible. More detailed simulations also show that our model can reproduce the trend that cell lines with a larger abundance of hetechromatin (such as resting T-cells and differentiated cells) are infected more slowly with respect to cells richer in active chromatin (such as activated T-cells and stem cells). Finally, the distribution of integration events within the nuclear environment predicted by a reaction-diffusion model is in good agreement with the nuclear localisation of recurrent integration genes in populations of activated T-cells [49].

Besides rationalising the experimental evidence on HIV integration, our model also leads to a number of predictions. At both, DNA and local chromatin level, our model suggests that if the reconnection integration moves no longer satisfy detailed balance, then flexibility should no longer provide a bias to favour integration. This might be relevant if integration complexes expending ATP are found, or built artificially in the future. Second, at the chromosome level, our model can be used to predict the pattern of integration sites within a chromatin fibre for which the epigenetic pattern is known (Fig. 3). Therefore, it can be potentially used to predict HIV integration in a number of different cell lines and organisms. These results could be compared, or combined, with "chromosome conformation capture" (CCC or Hi-C) analysis to provide us with new insight into the relationship between HIV integration and the large-scale chromatin organisation in living cells [49, 51–53]. At the whole cell level, we finally predict that the profile of HIV integration sites within the nuclear environment strongly depends on the large-scale genome organisation, and that dramatically different profiles may be seen in cells with non-standard genomic arrangements, such as retinal cells in nocturnal animals [48], but also senescent and diseased cells in humans and mammals.

METHODS

We model DNA and chromatin as semi-flexible polymers made of beads of size σ connected by springs. For DNA we consider $\sigma = 2.5$ nm and harmonic bonds, while the persistence length is set to $l_p = 50$ nm; for chromatin, we consider $\sigma = 30$ nm, FENE bonds and persistence length $l_p = 100$ nm. The reconnection moves are performed by attempting swaps between bonds connecting consecutive beads in viral and host DNA or chromatin. The recombination 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 [54] 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] Alberts B, Johnson A, Lewis J, Morgan D, Raff M (2014) *Molecular Biology of the Cell.* (Taylor & Francis), p. 1464.

[2] Cook P (2001) *Principles of Nuclear Structure and Function.* (Wiley).

[3] 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.

[4] Griffiths DJ (2001) Endogenous retroviruses in the human genome sequence. *Genome Biol.* 2(6):reviews1017.1–1017.5.

[5] Bushman F, et al. (2005) Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* 3(11):848–58.

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

[7] Craigie R, Bushman FD (2012) HIV DNA integration. *Cold Spring Harb. Persp. Med.* 2(7):1–18.

[8] 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.

[9] 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.

[10] 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.

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

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

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

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

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

[23] 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.

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

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

[26] Ali I, Conrad RJ, Ott M (2016) Retrovirus Integration: Some Assembly Required? *Cell Host Microbe* 20(6):702–704.

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

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] Michieletto D, et al. (2017) Shaping epigenetic memory via genomic bookmarking. *Nucleic Acids Res.* 1(November):1–12.

[36] 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.

[37] 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.

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

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

[40] 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.

[41] 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.

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

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

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

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

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

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

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

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

[50] 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.

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

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

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

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