Genome Biology

METHOD Open Access

SPIN reveals genome-wide landscape of nuclear compartmentalization

Yuchuan Wang[1], Yang Zhang[1], Ruochi Zhang[1], Tom van Schaik[2], Liguo Zhang[3,4], Takayo Sasaki[5], Daniel Peric-Hupkes[2], Yu Chen[3,6], David M. Gilbert[5], Bas van Steensel[2], Andrew S. Belmont[3] and Jian Ma[1*]

*Correspondence:
jianma@cs.cmu.edu
[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh 15213, PA, USA
Full list of author information is available at the end of the article

Abstract
We report SPIN, an integrative computational method to reveal genome-wide intranuclear chromosome positioning and nuclear compartmentalization relative to multiple nuclear structures, which are pivotal for modulating genome function. As a proof-of-principle, we use SPIN to integrate nuclear compartment mapping (TSA-seq and DamID) and chromatin interaction data (Hi-C) from K562 cells to identify 10 spatial compartmentalization states genome-wide relative to nuclear speckles, lamina, and putative associations with nucleoli. These SPIN states show novel patterns of genome spatial organization and their relation to other 3D genome features and genome function (transcription and replication timing). SPIN provides critical insights into nuclear spatial and functional compartmentalization.

Keywords: Nuclear compartmentalization, 3D genome organization, Nuclear bodies, Probabilistic graphical model

Background

In human and other higher eukaryotic cells, interphase chromosomes are organized spatially within the cell nucleus [1, 2], such that their packaging and folding lead to dynamic interactions between genomic loci [3]. A key determinant of this intranuclear chromosome packaging is the interaction between chromosomes and heterogeneous constituents in the nucleus—in particular, nuclear compartments or nuclear bodies—including nuclear pore complexes, lamina, nucleoli, and nuclear speckles [4, 5]. Earlier microscopy work demonstrated the important connections between spatial localization of chromosome regions and gene expression regulation [1, 6–8]. Therefore, characterizing nuclear compartmentalization is crucial toward a comprehensive delineation of the roles of nuclear organization in different cellular conditions [9]. Unfortunately, our understanding of the genome-wide chromatin interaction with different nuclear compartments remains surprisingly limited.

The advent of whole-genome mapping methods for chromatin interactions such as Hi-C has shown that, at megabase resolution, chromosomes are spatially segregated into A/B compartments genome-wide [10]. A/B compartments exhibit distinct correlations to active euchromatic and inactive heterochromatic regions of the genome, respectively, although such strict, binary compartment separation is a coarse-grained approximation [11]. Indeed, higher coverage Hi-C data generated from the human lymphoblastoid (GM12878) cells revealed that A/B compartments can be divided into five primary subcompartments, A1, A2, B1, B2, and B3, which harbor more refined associations with various functional features such as gene expression and histone modification [12]. However, the observations of chromosome spatial association with nuclear compartments derived from Hi-C are limited and intrinsically indirect. The large gap between what we can infer indirectly from Hi-C versus what we can see, albeit at lower throughput, in the microscope for nuclear compartmentalization has been a major bottleneck preventing a comprehensive view of nuclear organization.

Several genome-wide mapping methods have enabled more direct examination between chromosome regions and specific nuclear compartments. In [13], DamID was utilized to measure contact frequencies between chromatin with nuclear lamina, revealing that $\sim$ 35% of the human genome form lamina-associated domains (LADs). Recently, TSA-seq was developed to estimate cytological distance of chromatin toward nuclear speckles and nuclear lamina [14]. Even though TSA-seq data show a strong correlation with DamID, there are also clear differences. Most notably, the transitions of DamID scores are typically much more abrupt than TSA-seq maps that show gradual changes of signals over a chromatin trajectory [14], reflecting the differences in the methods, i.e., TSA-seq for cytological distance vs. DamID for molecular contact frequency. In addition, [14] (and later [15]) showed that TSA-seq scores relative to speckles and lamina are correlated with Hi-C subcompartments although TSA-seq scores reflect distance to subnuclear structures. These results manifested the potential of an integrative framework that simultaneously analyzes different but complementary mapping data to offer a more complete view of nuclear compartmentalization.

Here, we develop a new computational method called SPIN (Spatial Position Inference of the Nuclear genome) to identify genome-wide chromosome localization patterns relative to multiple nuclear compartments. SPIN integrates TSA-seq, DamID, and Hi-C in a unified framework based on hidden Markov random field (HMRF). As a proof-of-principle, we apply SPIN to TSA-seq (for nuclear speckles and nuclear lamina), DamID (for nuclear lamina and a marker of nucleoli), and Hi-C data to identify genome-wide spatial localization states in K562 cells. The “SPIN states” reveal new and detailed correlations with other features of genome structure and function, such as Hi-C subcompartments, topologically associating domains (TADs) [16, 17], histone modifications, levels of transcriptional activity, and DNA replication timing. Comparisons with Hi-C subcompartments and LADs from multiple cell types suggest constitutive patterns of compartmentalization. We also identify possible molecular determinants and sequence-level features that modulate different compartmentalization. Taken together, SPIN is an effective integrative method that combines different genome-wide mapping approaches of nuclear genome organization to infer global patterns of spatial localization of the chromosomes, which provide critical insights into nuclear spatial and functional compartmentalization.

Results

Overview of the SPIN method

The overview of the SPIN method is illustrated in Fig. 1a. Our goal is to identify genome-wide spatial compartmentalization patterns of the chromosomes by integrating TSA-seq [14] and DamID [13, 18] data together with Hi-C. TSA-seq and DamID provide complementary information to measure distance and contact frequency between chromosome regions and subnuclear structures. The rationale of including Hi-C is that the pairwise genomic regions spatially interacting with each other more often than expected (from Hi-C) are more likely to share similar spatial compartmentalization patterns. We formulate this objective by using the hidden Markov random field (HMRF) [19, 20], in which nodes represent non-overlapping genomic bins (with a size of 25kb) and edges represent either significant Hi-C interactions or adjacent genomic bins (see the "Methods" section). We assume that each node is associated with an unobserved spatial localization state that SPIN aims to reveal ("SPIN state"). The observations on each node include signals from TSA-seq and DamID for defined nuclear compartments. Given the observed data on each genomic bin across the entire genome, the goal of SPIN is to solve the estimation problem by maximizing the likelihood of assigning spatial compartmentalization states. Thus, the output of SPIN contains spatial localization state assignment, which is originally hidden, for each genomic bin throughout the genome.

SPIN is different from previous methods for chromatin domain segmentation based on a hidden Markov model (HMM) [18, 21, 22] where chromatin interaction is not utilized. Although SPIN shares similarity in its goal with Segway-GBR [23], the regularization in Segway-GBR uses significant Hi-C interactions as prior such that pairs of interacting genomic loci are encouraged to have the same label in genome annotation, which is not necessarily appropriate for refined spatial localization states of the chromosomes (see Additional file 1: Supplementary Results for more detailed comparisons). The transition probabilities between different states learned in SPIN generalize such constraints so that chromatin regions in spatial proximity would be assigned with states that correspond to similar but not necessarily the same localization. In addition to showing advantage of SPIN over other approaches using both simulation and real data (see Additional file 1: Supplementary Results), we validate our method by demonstrating that the SPIN states stratify functional genomic data and provide spatial interpretation for other 3D genome features, advancing our understanding of nuclear compartmentalization patterns (see later sections).

Note that in principle the input of SPIN on each genomic bin can also include functional genomic signals such as histone modifications, replication timing, and transcription levels. However, in this work, we explicitly limit the input signals to those that directly measure the spatial position of chromatin (TSA-seq and DamID) and use functional genomic data to evaluate the functional correlations of different SPIN states genome-wide.

SPIN identifies genome-wide patterns of nuclear compartmentalization

In this implementation of SPIN to infer genome-wide nuclear compartmentalization patterns, we used TSA-seq and DamID mapping data in K562 for nuclear speckles (SON TSA-seq) and the nuclear lamina (Lamin-B1 DamID and TSA-seq) [14, 24]. In addition, we generated new DamID maps using a Dam methylase fusion with a nucleolar

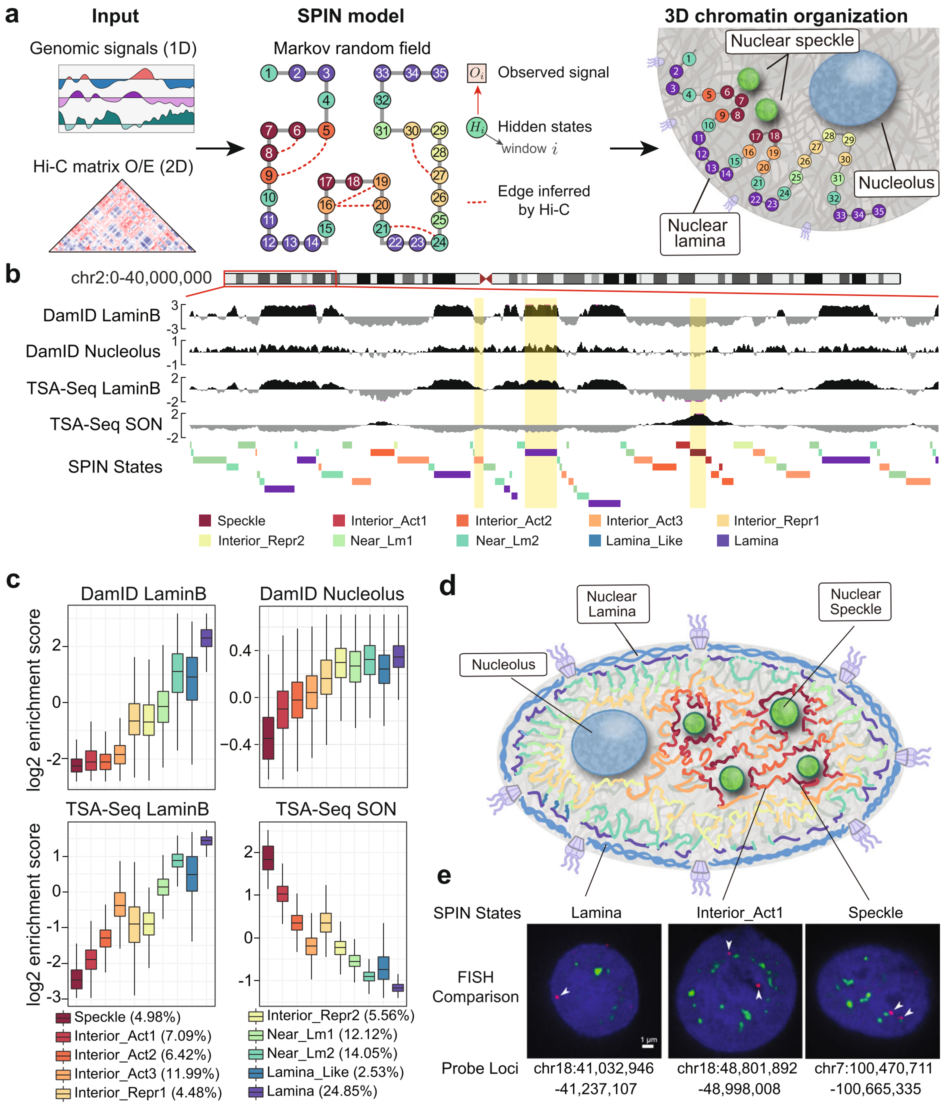

Fig. 1 Overview of the SPIN method and the SPIN states in K562. **a** Left panel: Input data for SPIN, including 1D genomic signals (TSA-seq and DamID), and Hi-C data. Middle panel: Graph representation of the hidden Markov random field. Genomic bins are represented as nodes and the hidden states of nodes are labeled by different colors. Right panel: Cartoon illustration of the SPIN states output for the nodes. **b** Genome browser view of the input data (TSA-seq and DamID) and output (SPIN states). Three regions with different SPIN states (Interior_Act3, Lamina, and Speckle) are highlighted. These regions have distinct distributions of TSA-seq and DamID signals. **c** Boxplots showing the distributions of the input TSA-seq and DamID data on different SPIN states. The genome-wide percentage (size) for each SPIN state is also shown. **d** Cartoon of the spatial positions of SPIN states relative to different nuclear compartments (nuclear lamina, nuclear speckles, nucleoli). **e** Comparison between the identified SPIN states and DNA FISH. Three genomic regions with different SPIN states (Lamina chr18:41,032,946-41,237,107, Interior_Act1 chr18:48,801,892-48,998,008, and Speckle chr7:100,470,711-100,665,335) are shown. DNA probes are labeled in red (with white arrowhead) and nuclear speckles (SON protein) is labeled in green. Nuclear DNA staining (DAPI) is in blue

targeting peptide repeat (4xAP3 [25]). We interpret these AP3-DamID data as putative nucleolus interactions; this will be analyzed in more detail elsewhere (van Schaik et al. *manuscript in prep.*). Hi-C data for K562 are from [12]. Details of data processing for TSA-seq, DamID, and Hi-C are in Additional file 1: Supplementary Methods. We partition the genome (chromosome 1–22 and X) into consecutive non-overlapping 25kb bins, which constitute the graph structure for the HMRF model in SPIN. Edges are derived from Hi-C and the adjacent genomic bins (including those caused by large-scale structural variants in K562; Methods and Additional file 1: Supplementary Methods). Figure 1b shows an example of the input signals from different measurements and the SPIN state annotations.

We identified 10 SPIN states that represent major nuclear compartmentalization patterns in K562. These 10 SPIN states are as follows: Speckle, Interior Active 1, 2, 3 (Interior_Act1, Interior_Act2, Interior_Act3), Interior Repressive 1, 2 (Interior_Repr1, Interior_Repr2), Near Lamina 1, 2 (Near_Lm1, Near_Lm2), Lamina_Like, and Lamina. The genome-wide percentage of each state is shown in Fig. 1c. The names of these states are partially informed by comparison to various functional genomic data, especially for the Interior states (see later), even though the input for SPIN does not use any functional genomic data. Note that we assessed the reliability of the identified SPIN states by using different TSA-seq, DamID, and Hi-C replicates as input and found that the states are highly consistent (Additional file 2: Figure S1a). In addition, we assessed the robustness of SPIN states based on its sensitivity to random initialization and found that SPIN can achieve consistent genome partitioning with randomly initialized states (Additional file 2: Figure S1b).

In Fig. 1c and Additional file 2: Figure S2, we show that each SPIN state has distinct distributions of TSA-seq and DamID signals for the input data, reflecting the spatial position for compartmentalization. For example, the Speckle state has the highest SON TSA-seq signals and the lowest lamina/nucleolus signals as compared to other states. Notably, although we group multiple states into larger classes such as Interior Active, Interior Repressive, and Near Lamina, the refined states do show their distinct patterns. For example, the Interior_Repr2 state has similar Lamin-B1 DamID and TSA-seq signals as compared to Interior_Repr1, but its nucleolus DamID score is significantly higher while its SON TSA-seq score is significantly lower (p value $<$ 2.2E−16). A recent report identified nucleolus associated domains (NADs) in mouse embryonic fibroblasts and found that there are two types of NADs [26]: type I NADs localize more frequently with both nucleoli and nuclear lamina and type II NADs localize with nucleoli but do not overlap with lamina. We also observed such distinctions related to nucleoli from our SPIN states for spatial localization. The Interior_Repr2 state has similar enrichment of nucleolous DamID scores as compared to the Near_Lm1-2, Lamina_Like, and Lamina states, but the Interior_Repr2 state has significantly lower enrichment with Lamin-B1 DamID and TSA-seq (Fig. 1c and Additional file 2: Figure S2) (p value $<$ 2.2E-16).

The identified SPIN states provide a comprehensive view of the spatial localization of the chromosomes in the nucleus relative to multiple subnuclear compartments, (see the cartoon in Fig. 1d). We compared the SPIN states to DNA FISH (fluorescence *in situ* hybridization) imaging data. In Fig. 1e, we show three genomic regions (with detailed genomic coordinates) that correspond to different SPIN states with comparisons to DNA FISH data from [14]. The probe in the FISH image of the Speckle state has an average distance of 0.16μm from nuclear speckle (SON protein). The probe in the FISH image of the Lamina state has an average distance of 0.98μm from nuclear speckles and is located $<$ 0.5μm from the nuclear periphery. The probe in the FISH image of the Interior_Act1 state has an average distance of 0.47μm from nuclear speckles. The comparison further suggests the reliability and advantage of having genome-wide SPIN states relative to multiple nuclear compartments.

SPIN states provide spatial interpretation for Hi-C subcompartments

The five primary Hi-C subcompartments (A1, A2, B1, B2, and B3) defined from [12], which exhibit strong associations with various genomic and epigenomic features, provide more detailed compartmentalization patterns from Hi-C data than the binary A/B compartment separation. However, the spatial localization context of Hi-C subcompartments has not been clearly revealed except that [14] used the Hi-C subcompartments to identify the two transcriptional hot-zones based on TSA-seq scores by comparing to A1/A2 subcompartments, suggesting that the A1 subcompartment was significantly closer than the A2 subcompartment to nuclear speckles (Hi-C subcompartments defined in GM12878 which has extremely high coverage Hi-C data). The recently developed algorithm SNIPER [15] facilitates the identification of subcompartments in Hi-C data with low to moderate coverage and provides Hi-C subcompartment annotations specifically in K562. Here, we directly compare the 10 SPIN states with Hi-C subcompartments in K562.

Figure 2a shows the overall comparison of Hi-C subcompartments in different SPIN states. Specifically, we found that the Speckle and Interior_Act1 states are strongly associated with A1 subcompartment (fold change enrichment 8.5 and 4.7, p value < 2.2E−16; Additional file 1: Supplementary Methods). Interior_Act2 is strongly associated with A1, A2, and B1 subcompartments (fold change enrichment 3.8, 3.2, and 3.6, respectively, p value < 2.2E−16). Interior_Act3 is enriched with A2 subcompartment (fold change enrichment 3.1, p value < 2.2E−16). The Interior_Repr1 and Interior_Repr2 states overlap more with B1 subcompartment (fold change enrichment 2.8, and 5.3, p value < 2.2E−16). We found that the Lamina_Like state is strongly enriched with B2 subcompartment (fold change enrichment 4.95, p value < 2.2E−16), while Lamina state is associated with both B2 and B3 subcompartments (fold change enrichment 3.16 and 5.58, p value < 2.2E−16). Together, different SPIN states have a strong correlation with different Hi-C subcompartments, supporting that Hi-C subcompartments reflect spatial positions relative to nuclear structures. However, the SPIN states offer a much more direct and refined interpretation of Hi-C subcompartments in terms of spatial compartmentalization. For example, although the Speckle, Interior_Act1, and Interior_Act2 states are all enriched with A1 subcompartment, they show distinguishable distributions regarding SON TSA-seq signals (Fig. 1c). This suggests that SPIN is able to further subdivide Hi-C subcompartment annotations into additional distinguishable spatial states of nuclear compartmentalization.

SPIN states stratify patterns of transcription activity and histone modification

Earlier studies have shown the correlation between the genome compartmentalization patterns and transcriptional activities [10, 12, 15]. We sought to assess whether the SPIN states, which offer more detailed compartmentalization patterns, further stratify the transcriptional activity based on spatial locations of the chromatin. We first compared the SPIN states with 11 histone modification ChIP-seq datasets in K562 from the ENCODE project, including H2A.Z, H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K36me3, H3K79me2, H3K9ac, H3K9me1, and H3K9me3. Overall, we found that the SPIN states have a strong correlation with histone modifications (Fig. 2a).

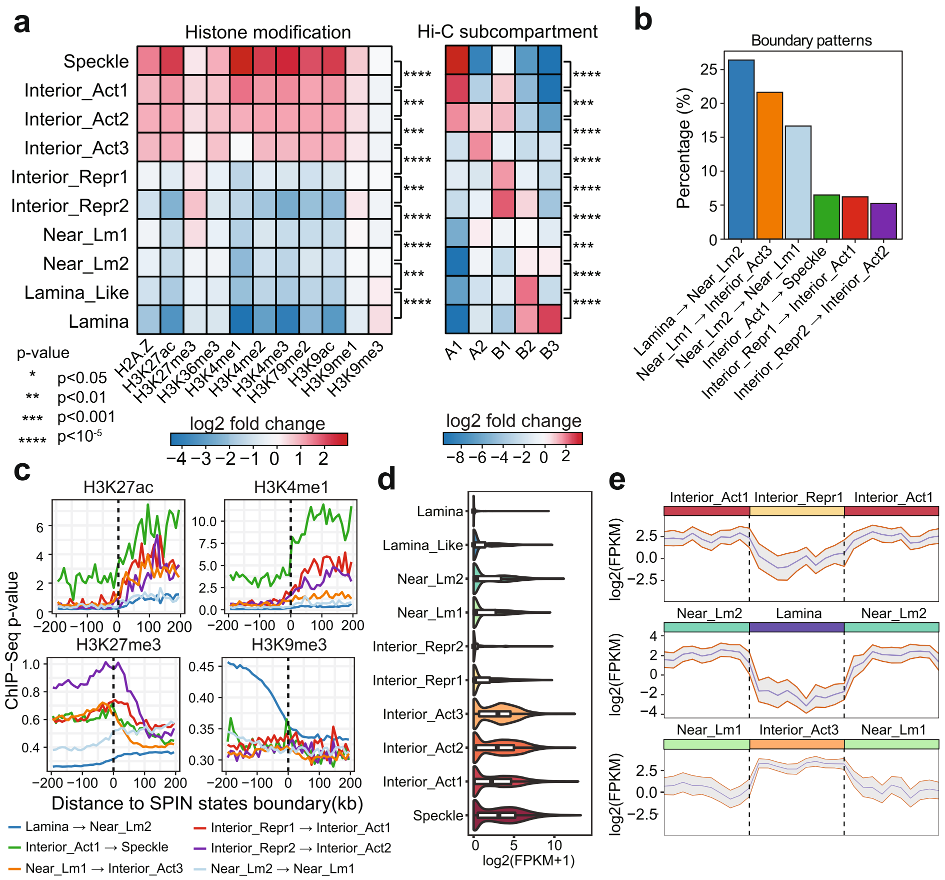

Fig. 2 SPIN states correlate with various genomic and epigenomic features. **a** Stratification of histone marks and Hi-C subcompartments as compared to SPIN states. Heatmap shows signal fold enrichment calculated as the ratio of observed signals over expected. K562 subcompartment annotations are produced by SNIPER [15]. **b** Most frequently observed SPIN state boundary transition patterns. We count the percentage of different SPIN boundary patterns where symmetrical patterns are merged. **c** Histone marks (H3K27ac, H3K4me1, H3K27me3, and H3K9me3) on SPIN state transition boundaries. **d** Violin plot of transcript levels based on GRO-seq in different SPIN states. **e** Transcript levels based on GRO-seq over three consecutive SPIN states. Three consecutive SPIN states are normalized by length (mean: blue line; standard deviation: gray)

In addition, we also show that by adding Hi-C data into the SPIN model we can achieve state calling to better stratify histone modifications as compared to the baseline HMM-based model (Additional file 2: Figure S3). From the SPIN states, as chromatin localization changes from nuclear periphery to the interior (i.e., the lamina to speckle axis), we observed a dramatic increase of ChIP-seq signal *p* value of active histone marks (e.g., H3K27ac, H3K4me1, H3K4me3, H3K9ac) and, in general, a decrease in repressive marks (e.g., H3K9me3) (Additional file 2: Figure S4). Specifically, histone marks that are associated with transcriptional activation, including H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, and H3K79me2, show a dramatic increase of signal in the Speckle state (> 5-fold increase on average, *p* value < 2.2E−16) (Fig. 2a). This result is consistent with previous studies that transcriptionally active chromatin regions are spatially localized preferentially near nuclear speckles and toward the interior [8, 14]. In contrast, heterochromatin mark H3K9me3 shows stronger presence in the Lamina state (*p* value < 2.2E−16), agreeing with earlier reports that LADs are often heterochromatic with inactive genes [8, 18, 21]. In addition, the H3K27me3 mark, known to be associated with repressed transcription [27], is more abundant in Interior_Repr2 (p value $< 2.2\text{E}{-}16$) compared with the Near_Lm1 and Interior_Repr1 states (Fig. 2a). Importantly, we found that there is an increase of nucleolus DamID signals in Interior_Repr2 compared with the Interior_Repr1 and also the Interior_Act states (Fig. 1c), suggesting a possible localization preference of Interior_Repr2 toward the nucleolus even though Interior_Repr1 and Interior_Repr2 have overall similar distance to nuclear lamina (Fig. 1c). This is in concordance with the recent report that H3K27me3 marks are enriched on type II NADs [26], which are found associated with nucleoli but not with LADs. Our results suggest that SPIN states reflect different associations with histone marks, and chromatin enriched for H3K27me3 and H3K9me3 have distinct spatial localization preferences.

To further demonstrate that the SPIN states clearly stratify functional genomic signals, we analyzed the patterns of histone modification signals across the transition boundaries between neighboring SPIN states. We selected the top six boundary types that are most frequently observed (Fig. 2b). Since SPIN states do not distinguish DNA strands, we therefore merged transition patterns in both directions on the genome. For example, Lamina to Near_Lm2 transition and Near_Lm2 to Lamina transition are considered as the same type of transition boundary for this analysis. For each transition type, we calculated the average histone modification signals at +/- 200 kb surrounding the transition boundaries. We found that many histone modifications show a clear, dramatic change across the SPIN state transition boundaries (Fig. 2c, Additional file 2: Figure S5). In particular, the active histone marks, such as H3K4me1 and H3K27ac, show a pronounced, > 2-fold signal increase, when the chromatin trajectory is going from Interior_Act1 to Speckle. H3K9me1 signals exhibit a gradual rather than a sharp increase across the transition boundaries (Additional file 2: Figure S5). Additionally, we observed the opposite trend of signal enrichment across the boundaries for the repressive marks such as H3K27me3 and H3K9me3 (Fig. 2c). We further compared histone mark changes at the transition boundaries of SPIN states when the boundaries were defined by Hi-C subcompartments in K562 [15]. This reveals a sharper transition of histone marks at SPIN state boundaries as compared to Hi-C subcompartments (Additional file 2: Figure S5; especially H3K9me1, H3K9me3, H3K4me1, H3K4me2, H3K4me3), further suggesting that SPIN states offer a more accurate and refined definition of nuclear compartmentalization as compared to Hi-C subcompartments.

Next, we explored how transcription activity varies in different SPIN states. We compared SPIN states with GRO-seq data [28] that measures through run-on transcription the density of engaged RNA Pol2 polymerases across protein coding genes in K562 (Fig. 2d). We found that genes in the Speckle and Interior_Act states have high transcription levels, as expected, with average FPKM > 40 for these nascent transcripts. The majority (over 90%) of the top 10% actively transcribed genes are from the Speckle or Interior_Act states. In contrast, genes in the Lamina and Interior Repr states are highly repressed. In addition, as shown in Fig. 2e for the consecutive SPIN states, nascent transcription in Interior_Act vs. Interior_Repr exhibit significant difference (p value $< 2.2\text{E}{-}16$), despite the fact that both states are likely to localize at relatively similar radial positions in the nucleus (based on TSA-seq). Also, genes in the Near_Lm and Lamina_Like states have higher transcription compared with the Lamina state (p value $= 5.689\text{E}{-}11$). These analyses suggest that the SPIN states demarcate spatial patterns of chromosome regions in fine-scale separated into transcriptionally active and repressed regions.

SPIN states are predictive of DNA replication timing

DNA replication timing (RT) is a vital genome function that is highly aligned with large-scale compartmentalization [29]. To further evaluate the functional connections of the SPIN states, we generated 7-fraction Repli-seq data for K562, where each fraction corresponds to the DNA replicated during 6 stages of S-phase (i.e., S1 to S6) as well as the stage entering G2-phase representing the latest DNA to replicate (Additional file 1: Supplementary Methods). For each genomic bin (5kb), we calculated the percentage of DNA replicated in each fraction (among all 7 fractions) which was used to compute the fold change score of SPIN states on different replication fractions genome-wide. We found that RT can be clearly stratified by the SPIN states (Fig. 3a). The Speckle and Interior_Act states are found in early replicated regions (S1, S2, fold change score > 1.5). The Interior_Repr1, Interior_Repr2, Near_Lm1, and Near_Lm2 states are replicated in the middle of S phase (S3, S4, fold change score > 1.3). The Lamina_Like and Lamina states are replicated late (S5, S6, and G2, fold change score > 1.3). Overall, the SPIN states show a striking separation of the multi-fraction Repli-seq. In addition, using the definition of constitutive and developmentally regulated replication domains (RDs) [29], we found the SPIN states have distinct correlation with different patterns of constitutive and developmental RDs (Fig. 3b). Constitutive RDs can be further separated as constitutive early (CE) and constitutive late (CL) domains. We found that 85% of the genomic regions in the Speckle state are CE and 55% of the genomic regions in the Lamina state are CL. In contrast, other SPIN states contain a higher proportion of developmentally regulated RDs. In Fig. 3c, we show that the SPIN states also correlate with evolutionary patterns of RT between human and non-human primates based on the annotations from [30]. Here, the RT patterns are separated in five groups based on their conservation across primates: early (all primate species have early RT), late (all primate species have late RT), weakly early (4 out of 5 species have early RT), weakly late (4 out of 5 species have late RT), and unconserved (the rest). We found that 96% of the genomic regions in Speckle states have conserved early RT pattern. In addition, the genomic regions in three interior_Act states mostly have conserved early RT. In contrast, 56% of Lamina state regions have conserved late RT. Similar to our observation of constitutive and developmentally regulated RDs, other SPIN states overlap more with unconserved RT. Collectively, these analyses reveal a strong correlation between the detailed nuclear spatial compartmentalization identified by SPIN and the DNA RT program as well as its constitutive patterns in different cell types and across different species.

Next, we sought to investigate the functional significance of SPIN states in terms of how important the SPIN states are, among other epigenomic features, in predicting RT. We built a predictive framework based on a random forest regression model to predict the multi-fraction Repli-seq signals along the genome by using the SPIN states together with various histone mark data (Fig. 3d) (see the "Methods" section). We specifically calculated the importance of each input feature based on how much each feature decreases the weighted impurity in a decision tree in the random forest. We found that the SPIN state is the most important feature, followed by H3K9me1, H3K9ac, H3K4me1, and H3K36me3, which are the top 5 most informative features (Additional file 2: Figure S6).

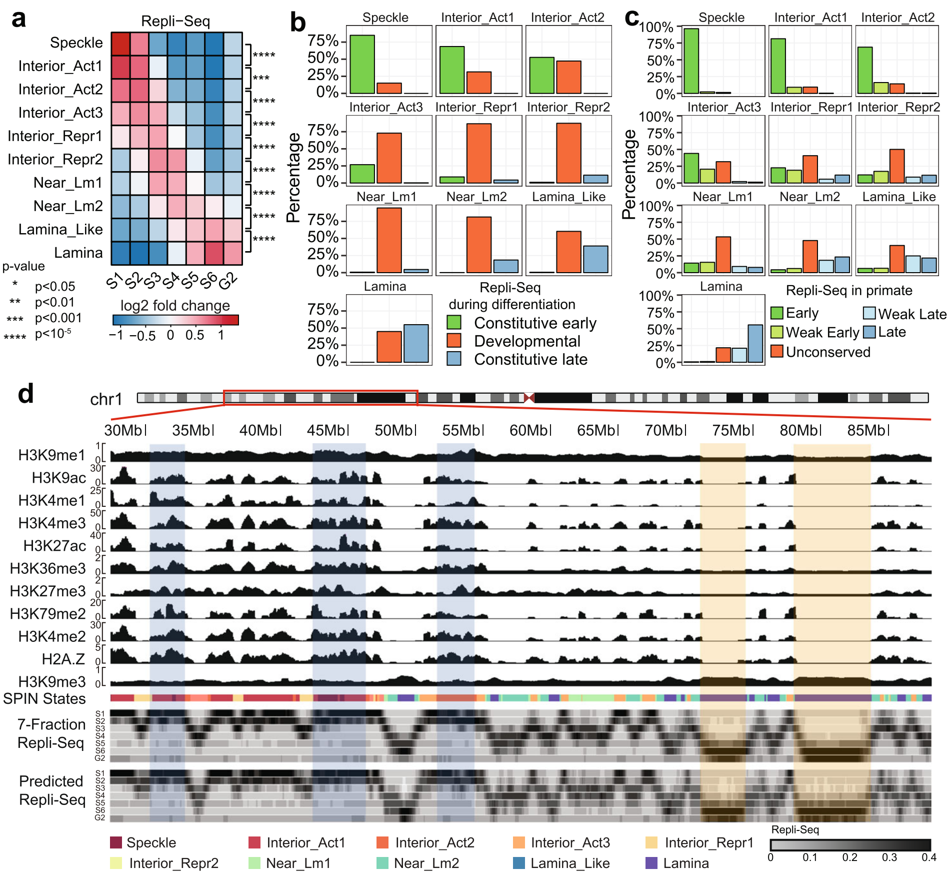

Fig. 3 SPIN states are predictive of DNA replication timing. **a** Stratification of multi-fraction Repli-Seq based on the SPIN states. **b** Comparison with constitutive and developmentally regulated replication timing (RT) domains. The barplot shows what percentage of each SPIN state belongs to constitutive early, constitutive late, and developmentally-regulated RT domains, respectively. **c** Comparison with RT evolutionary patterns defined by Phylo-HMGP [30] between human and other non-human primate species. **d** A Genome Browser view of the prediction of Repli-seq signals based on the SPIN states together with histone marks. The tracks are (from top to bottom): histone modifications, SPIN states, multi-fraction Repli-seq, and the predicted Repli-seq signals. For multi-fraction Repli-seq, the signal in each genomic bin is normalized across all fractions and the color scale is shown at the bottom. The three highlighted regions in light blue are examples of early replicating domains. The two highlighted regions in light orange are examples of late replicating domains

Together, the comparison with multi-fraction Repli-seq demonstrates that, by integrating different nuclear genome mapping data (TSA-seq, DamID, and Hi-C), the SPIN states delineate the detailed connections between nuclear compartmentalization and replication timing.

SPIN states offer new perspectives for other nuclear organization units

We evaluated the significance of the SPIN states with respect to providing new insights for other nuclear genome features. We assessed the interplay between the SPIN states and the known 3D genome structural features, including LADs, TADs, and chromatin loops. We also asked whether the SPIN states are indicative of the constitutive patterns of nuclear compartmentalization across different cell types.

By combining TSA-seq and DamID, we identified several types of nuclear periphery states with different localization relative to the nuclear lamina (Lamina, Lamina_Like, and

Near_Lm1-2, Fig. 1c). To further assess how each SPIN state corresponds to LADs across multiple cell types, we used Lamin-B1 DamID data in 6 human cell lines from the 4DN portal, including HCT116, K562, RPE-hTERT, HAP-1, HFFc6, and H1-hESC (see Additional file 3: Table S1). Based on the assignment of LADs in 6 cell lines, we separated LADs into two categories: constitutive LADs (cLADs) and facultative LADs (fLADs). cLADs are defined as genomic regions characterized as LADs in at least 5 out of 6 cell lines. fLADs are defined as genomic regions characterized as LADs in at least 2 but fewer than 5 cell lines. In Fig. 4a, we show that there is a large difference between cLADs and fLADs in terms of the overlap with different SPIN states. Seventy-one percent of cLADs are in the Lamina state in K562 as well as 22% in Near_Lm2, 3% in Lamina_Like, and 3% in Near_Lm1 in K562. In contrast, for fLADs, only 23% are in K562 Lamina state, but 39% are in Near_Lm2 and 23% are in Near_Lm1 in K562. This suggests that the SPIN states in one cell type (i.e., K562 in this study) can already separate fLADs and cLADs, as well as extending the concept of LADs into two separate categories. These results are consistent with the recently reported HiLands chromatin domains relative to the nuclear lamina based on both Lamin-B1 DamID and histone marks (in mES cells) [21, 31], where

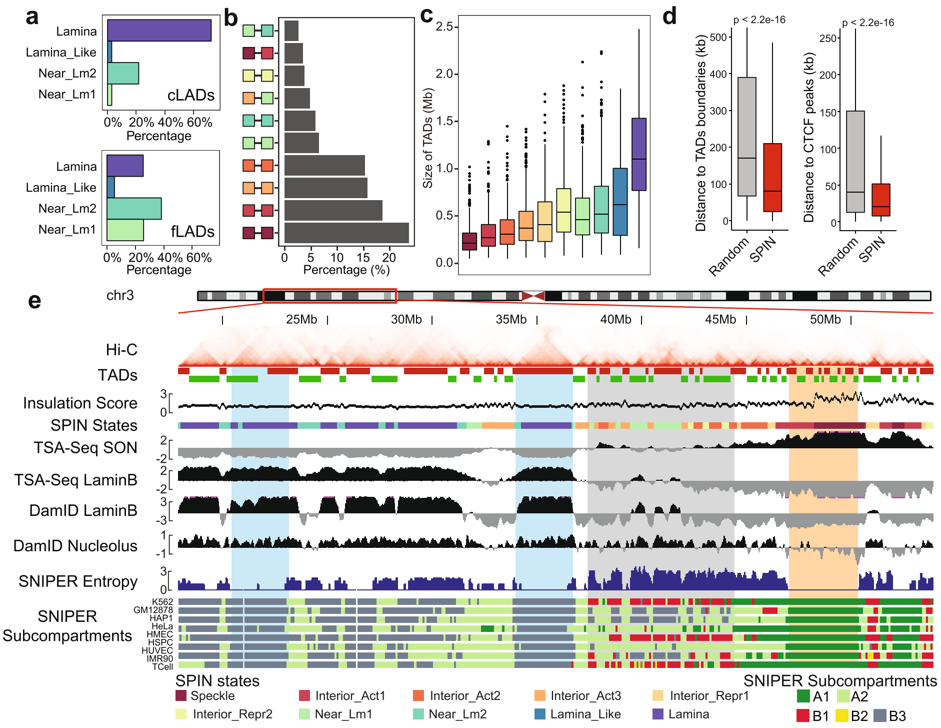

Fig. 4 Comparison between the SPIN states and other nuclear organization units. **a** Percentage of constitutive LADs (cLADs) and facultative LADs (fLADs) in SPIN states Lamina, Lamina_Like, and Near_Lm1-2. cLADs and fLADs are called from DamID data from 6 different human cell lines. **b** Distributions of SPIN states for significant interaction pairs (CTCF-mediated chromatin loops) based on CTCF ChIA-PET data. **c** Size of the TADs within each SPIN states. TADs are called by the directionality index (DI) method. **d** SPIN states boundaries are close to TADs boundaries and CTCF peaks. **e** A Genome Browser snapshot of the SPIN states together with TADs and cross cell-type Hi-C subcompartments. K562 Hi-C contact map is shown at the top. TADs are called by the DI method. Hi-C subcompartments in multiple cell lines [15] are shown at the bottom. Two highlighted regions in light blue are constitutive B3 subcompartments that correspond to the Lamina SPIN state. The light gray region has more dynamic subcompartment annotations in different cell types. The light orange region highlights a constitutive A1 subcompartment that corresponds to the Speckle SPIN state

HiLands-B and HiLands-P are two distinct chromatin states that correspond to the facultative and constitutive LADs. For the two types of LADs, both the Lamina SPIN state and HiLands-P have higher Lamin-B1 DamID signals and higher H3K9me3 modification, while the Near_Lm SPIN state and HiLands-B have lower Lamin-B1 DamID signals and higher H3K27me3 enrichment.

Next, we compared the SPIN states with ChIA-PET/Hi-C chromatin loops and TADs derived from Hi-C. For CTCF-mediated ChIA-PET chromatin loops, we discarded loops within 25kb range to focus on longer-range interactions. We found that most loops are formed within the same SPIN states with more loops toward the interior states with higher transcriptional activity (Fig. 4b). We also observed similar results in Pol2-mediated ChIA-PET loops (Additional file 2: Figure S7a). In addition to ChIA-PET loops, we compared the SPIN states to the Hi-C loops in K562 identified by HiCCUPS [12] (Additional file 2: Figure S7c). Similar to the comparisons with ChIA-PET loops, the majority (81.4%) of the HiCCUPS-identified chromatin loops are formed by genomic loci from the same SPIN states (Additional file 2: Figure S7b-c).

For TADs defined by the directionality index (DI) method [16], we found that the TADs tend to stay within the same SPIN state. Specifically, 82.3% of TADs have only one SPIN state labeled. It is rare (0.4%) that one TAD spans more than two different SPIN states (p value $<$ 2.2E−16). Importantly, we observed the increase of TAD size when the SPIN state trajectory is changing from the nuclear interior to the periphery, with the average TAD size as 1.12Mb in the Lamina state and 0.19 Mb in the Speckle state, respectively (Fig. 4c, Fig. 4e, and Additional file 2: Figure S8). The boundaries of the SPIN states are close to TAD boundaries and CTCF peaks than expected at random (Fig. 4d) (p value $<$ 2.2E−16). We calculated the Hi-C insulation score [32] to represent TAD boundary strength. We found that the insulation score on the Speckle state is on average two times higher than the score in the Lamina state, indicating that there are stronger TAD/subTAD boundaries in the Speckle states (Additional file 1: Supplementary Methods). In addition, we performed analysis for TAD-TAD level interactions and showed that TADs from the same SPIN state tend to form spatially separated cliques (i.e., a group of TADs with long-range interactions; see Additional file 2: Figure S9 and Additional file 1: Supplementary Results). Taken together, these results show that SPIN states stratify TADs and chromatin loops by providing spatial context.

We sought to analyze how conserved the spatial compartmentalization patterns are across different cell types. Here, we use Hi-C subcompartments in multiple cell lines as an estimation of chromosome spatial localization in different cell types. As we have already shown, Hi-C subcompartments are highly correlated with SPIN states although the SPIN states provide more detailed and explicit compartmentalization views relative to subnuclear bodies. We compared the SPIN states in K562 with the SNIPER Hi-C subcompartments across 9 human cell types [15], including K562, GM12878, HAP1, HeLa, HMEC, HSPC, HUVEC, IMR90, and TCell. We calculated the SNIPER entropy score as the metric of conservation for Hi-C subcompartments across cell lines (Additional file 2: Figure S10). The Speckle state has the lowest SNIPER entropy score (0.1 on average), strongly suggesting that Hi-C subcompartments on Speckle (mostly A1) are largely conserved across cell lines (Fig. 4e). The Lamina_Like and Lamina states (mostly B2 and B3 subcompartments) also have relatively low SNIPER entropy scores (0.75 on average). The most dynamic SPIN states across cell types are Interior_Repr2 and Near_Lm1. This comparison with cross-cell type SNIPER Hi-C subcompartments suggests that different SPIN states have distinct patterns across cell types with Speckle being the most conserved state.

Discussion

We identified 10 SPIN states in K562 based on TSA-seq and DamID data together with Hi-C. We showed that different SPIN states represent different spatial localization preferences within the nucleus. Further analysis indicates that SPIN states have strong correlation with and also better stratify other functional genomic features, such as histone modification, transcription activity, and DNA replication timing, suggesting that the detailed SPIN states have important relationships to genome function. The SPIN states also facilitate the identification of potential molecular determinants and sequence features that may play roles in modulating nuclear genome compartmentalization (see Additional file 1: Supplementary Results, Additional file 2: Figure S11, S12, S13, and S14), paving the way for further experimental validation to pinpoint the mechanisms that give rise to specific compartmentalization. Our computational framework is flexible to incorporate more data for other nuclear structures (such as PML bodies, nuclear pores, and pericentromeric heterochromatin) when they become available (e.g., through the 4D Nucleome project [9]) to achieve even more complete characterization of nuclear compartmentalization. We therefore expect that SPIN has the potential to become an important method for revealing nuclear compartmentalization in different cellular conditions.

Nuclear compartmentalization analysis has primarily focused on A/B compartments based on Hi-C data. Although five major Hi-C subcompartments were revealed in [12] in GM12878 cells, such analysis has not been possible in other datasets with low to moderate coverage until recently [15]. Additionally, prior work on LADs and inter-LADs also made the binary distinction of chromatin domains associated with the nuclear lamina, where LADs and inter-LADs largely correspond to A/B compartment separations. Our SPIN states significantly advance our understanding of the detailed spatial localization patterns, much beyond binary separation of the chromatin domains in terms of nuclear compartmentalization. For example, our results further extend the concept of LADs into two separate types of SPIN states, i.e., Lamina and Near Lamina, that can distinguish constitutive LADs and facultative LADs. In addition, we have demonstrated that SPIN states offer new spatial interpretation of Hi-C subcompartments, clarifying the compartmentalization patterns of specific subcompartments relative to nuclear speckles, the nuclear lamina, and nucleolus. Besides, SPIN states reveal more refined compartmentalization patterns as compared to Hi-C subcompartments, supported by comparisons with other genomic and epigenomic features.

Our SPIN states can be further evaluated and compared using other analysis methods, e.g., polymer simulations [33], 3D genome structure population modeling [34], and integration between chromatin interactome and regulatory network [35]. In addition, recently published new genome-wide mapping methods [36–38] and approaches for assessing chromatin interaction dynamics [39, 40] could be incorporated into our framework. In this work, we made additional attempt to reveal the patterns of consecutive SPIN states on the chromosomes to reveal potential chromatin fiber trajectories with distinct functions (see Additional file 1: Supplementary Results, Additional file 2: Figure S15), which can be validated by Oligopaints/OligoSTORM imaging [41–43], further bridging the gap between what we can see from microscope and what we can produce from genomic mapping data.

The molecular determinants that modulate the maintenance and movement of compartmentalization remain largely elusive. Earlier microscopy studies identified genes associated with chromatin targeting to specific nuclear structure (e.g., Hsp70 transgene [7]). Very recent work from [44] postulated the roles of molecular determinants for the global changes of chromosome compartmentalization, although the exact players have yet to be identified. Our SPIN states facilitate the identification of potentially important sequence features for specific compartmentalization, which provides promising tool to help elucidate the mechanisms that maintain and modulate compartmentalization. We made initial effort to identify sequence features enriched in different SPIN states (Additional file 1: Supplementary Results, Additional file 2: Figure S11, S12, S13, and S14). This can be further facilitated by SPIN compartmentalization states in other cell types to prioritize important sequence features. Such analysis can be validated by genome engineering experiments.

Conclusions

In this work, we developed SPIN, a probabilistic graphical model that integrates TSA-seq, DamID, and Hi-C to provide a comprehensive view of genome-wide nuclear compartmentalization to nuclear speckles, nuclear lamina, and nucleolus. Overall, SPIN represents an important step forward in developing integrative computational tools to offer new perspectives of spatial organization of the chromosomes in the nucleus and their interplay with various subnuclear structures.

Methods

Data acquisition, processing, and availability

TSA-seq and DamID data

K562 cells were obtained from the ATCC. For TSA-seq, the cells were cultured following the ENCODE project recommendations. For DamID, the cells were cultured according to 4DN guidelines following the ATCC recommendations. TSA-seq data generation of SON and Lamin-B1 was described and reported in [14]. DamID of Lamin-B1 data generation was described and reported in [24]. For nucleolus DamID, a tandem repeat of 4 copies of the nucleolus targeting domain of AP3D1, linked with flexible GGSGG-linkers (4xAP3 in short [25]), was codon optimized for expression in human cells (IDT). NheI and SalI restriction sites were added on the flanks and used to replace the LMNB1 gene with the 4xAP3 repeat in the Dam expression vector. The 4xAP3 Dam vector was used to generate nucleolus contact data identical to Lamin-B1 DamID-seq ([24], van Schaik et al. *manuscript in prep.*). Sequencing reads from TSA-seq and DamID were first mapped to the human reference genome (hg38; chromosome 1–22 and X). For TSA-seq data, PCR duplicates were removed using Samtools [45] (rmdup command with default parameters). Next, for TSA-seq data, we calculated the number of reads mapped in sliding 20kb windows with 1kb step size on the genome. The normalized TSA-seq enrichment score was calculated as the log2 ratio of read counts between TSA pull-down sample and the input normalized by sequencing depth [14]. The TSA-seq score was further smoothed by using Hanning window of length 21 following the same smoothing approach used in [14]. For DamID data, scores were similarly calculated as the log2 ratio of mapped reads between

Dam-target and Dam-only samples [18]. The signal was then averaged on sliding 20 kb windows with 1 kb step size with additional smoothing by Hanning window of length 21. The smoothed TSA-seq and DamID signals were binned in 25 kb resolution. We chose 25 kb as the resolution for SPIN states to balance between the signal resolution limit in TSA-seq and DamID and the refined resolution of nuclear compartmentalization that SPIN can achieve (see Additional file 2: Figure S16 and Additional file 1: Supplementary Results).

Hi-C data

We obtained the Hi-C data of K562 cells from [12]. Both intra-chromosomal and inter-chromosomal interactions were processed at 25kb resolution and VC_SQRT normalization was applied to the Hi-C contact matrices. Hi-C data extraction and normalization were performed using Juicer [46]. For intra-chromosomal contact maps, we calculated the log2 ratio between observed (O) over expected (E) interactions (i.e., O/E) for each pair of interactions. The rationale is to consider genomic loci (not necessarily close on 1D distance) that share spatial localization with higher than expected genome-wide Hi-C interaction patterns to facilitate the identification of compartmentalization. Our simulation evaluation also supports this rationale (Additional file 1: Supplementary Results). For inter-chromosomal interactions, the expected number of interactions was set to be uniformly distributed between genomic loci on different chromosomes. For each chromosome, we fitted a Weibull distribution for Hi-C contacts and kept those interactions with p value $< 1\text{E}{-5}$ as significant interactions for subsequent steps as input to the SPIN method. For each inter-chromosomal interaction, we also used p value $< 1\text{E}{-5}$ as the cutoff for significant interactions, but for each pair (i, j), we required that all neighboring pairs between $i \pm 1$ and $j \pm 1$ should be also significant to increase the reliability of added edge.

Multi-fraction Repli-seq

Multi-fraction Repli-seq was performed using an extension protocol to the E/L Repli-seq [47]. Briefly, K562 interphase cells were labeled with BrdU and sorted into 7-fractions (S1, S2, S3, S4, S5, S6, and G2) (Additional file 1: Supplementary Methods). Note that cells in G1 fraction were collected at the very early side of G1 peak and are sequenced without BrdU Immunoprecipitation (IP); therefore, we used G1 fraction as a control to remove copy number and mappability bias. For each Repli-seq fraction, we mapped the sequenced reads to the reference genome hg38. Then, we calculated read counts on 10kb sliding windows with 1 kb step size. The total number of mapped reads were then normalized to 1 million read counts per fraction. The raw signals of each window were normalized by the signals from G1. Genomic bins with zero mapped reads in G1 were considered as unmappable regions. For each 1 kb bin, the replication timing signal was calculated as the percentage of the total signals over the seven fractions. Finally, the replication timing signals were also binned in 25 kb resolution.

Other epigenomic data and annotations

We compared SPIN states with other epigenomic datasets, such as Hi-C subcompartments, TADs, LADs, ChIP-seq, and GRO-seq. For Hi-C subcompartments, we used the K562 Hi-C subcompartment annotation produced by SNIPER [15]. Hi-C subcompartments in additional cell types were also from [15]. K562 histone mark and transcription factor ChIP-seq data were obtained from the ENCODE project [48]. Datasets with replicates were merged. For data sets with no processed *p* value available, we used MACS2 [49] to calculate ChIP-seq *p* values (for narrow peak call, command bdgpeakcall is used). We downloaded CTCF and POLR2A ChIA-PET in K562 from the ENCODE project. ChIA-PET reads were processed using ChIA-PET Tool [50] with default parameters. As for TADs, we used the DI method [16] to call TADs based on 10kb resolution Hi-C. In addition, we used CaTCH [51] to call hierarchical TADs. To identify LADs, we used a hidden Markov model to identify LADs and inter-LADs from K562 Lamin-B1 DamID data [13]. LADs annotations in additional cell types were collected from 4DN data portal. See Additional file 1: Supplementary Methods for additional data analysis details by comparing to SPIN states.

All datasets used in this work are listed in Additional file 3: Table S1.

Algorithm description of the SPIN method

Overall design of the model

SPIN (Spatial Position Inference of the Nuclear genome) is developed based on a type of probabilistic graphical model called hidden Markov random field (HMRF) [19, 20], with the goal to identify genome-wide patterns of nuclear compartmentalization by integrating TSA-seq, DamID, and Hi-C (see Fig. 1a for the model overview). HMRF can be represented as an undirected graph $G = (V, E)$, where each node $i \in V$ represents a non-overlapping 25kb genomic region and E represents the set of edges. For each node i, the observation $O_i \in \mathbb{R}^d$ is a vector of 1D TSA-seq and DamID signals on the genomic bins. Specifically, in this study, these observations are SON TSA-seq for nuclear speckles, Lamin-B1 TSA-seq for nuclear lamina, Lamin-B1 DamID for nuclear lamina, and 4xAP3 nucleoli DamID. The edges $(i,j) \in E$ in the graph G represent the following: (1) significant Hi-C interactions between two genomic loci (see Hi-C data processing), (2) adjacent nodes on the chromosomes, and (3) *de novo* adjacencies caused by large structural variations since K562 is a cancer cell line (see Additional file 1: Supplementary Methods).

Each node i has a hidden state H_i, which represents the spatial localization of genomic bin i relative to multiple nuclear compartments. H_i is only dependent upon O_i and the neighbors of i, i.e., $N(i) = \{j | j \in V, (i,j) \in E\}$. Given the number of states k (see below on how we estimate k), our goal is to estimate the hidden states H_i for all nodes that maximize the following joint probability:

$$P(\overrightarrow{H}, \overrightarrow{O}) \propto \frac{1}{Z} \prod_{i \in V} P_V(O_i | H_i) \prod_{(i,j) \in E} P_E(H_i, H_j) \tag{1}$$

where $\overrightarrow{H}$ represents the hidden states of all nodes, H_i is the hidden state of node i, $P_V(O_i|H_i)$ corresponds to the potential of node i that the observation is O_i given the hidden state H_i. $P_E(H_i, H_j)$ corresponds to the edge potential between two nodes i and j with hidden states H_i and H_j. Z is the constant used for normalization:

$$Z = \sum_{\overrightarrow{H}} \left(\prod_{i \in V} P_V(O_i | H_i) \prod_{(i,j) \in E} P_E(H_i, H_j) \right) \tag{2}$$

We assume that the observation of O_i given hidden state $H_i = h_a$ follows a multivariate Gaussian distribution, i.e.,

$$P_V(O_i|H_i = h_a) = \frac{1}{\sqrt{(2\pi)^d |\Sigma^{h_a}|}} \exp\left\{-\frac{1}{2}(O_i - \mu^{h_a})^T [\,\Sigma^{h_a}]^{-1}\,(O_i - \mu^{h_a})\right\} \tag{3}$$

where O_i follows multivariate Gaussian distribution $N(\mu^{h_a}, \Sigma^{h_a})$ given state $H_i = h_a$. The edge potential is defined by the transition probability between neighbor states h_a and h_b.

$$P_E(H_i = h_a, H_j = h_b) \propto t(h_a, h_b) \tag{4}$$

Initialization and model parameter estimation

To initialize $P_V(O_i|H_i)$ for each node i, we estimate it based on a Gaussian mixture model with given number of states k. Here, Gaussian mixture model assumes that the input data from TSA-seq and DamID for a given state are generated from a mixture of multivariate Gaussian distributions. We have:

$$P_\theta(O_i) = \sum_{h_a \in \mathbb{H}} P(O_i|H_i = h_a) \times \pi(H_i = h_a) \tag{5}$$

where $P_\theta(O_i)$ represents the mixture of k Gaussian distributions of different types of observed signals, $\pi(H_j = h_a)$ is the mixture proportion of the hidden states.

To initialize $P_E(H_i, H_j)$, we estimate it by the transition probability of initial states called from the Gaussian mixture model. For each bin i, we choose H_i to be the state h_a that maximizes $P(H_i = h_a|O_i)$ of Gaussian mixture model. The initial transition matrix between two state h_a and h_b can be calculated as:

$$\frac{\sum_{(i,j)\in E} [\,\mathbb{1}(H_i = h_a)\mathbb{1}(H_j = h_b)]}{\sum_{(i,j)\in E} [\,\mathbb{1}(H_i = h_a) + \mathbb{1}(H_j = h_b)]} \tag{6}$$

where $\mathbb{1}$ is the indicator function.

We use the expectation-maximization (EM) algorithm to estimate the parameters in the model, including parameters to define node potential and edge potential. At iteration t, we assume that our estimate of model parameters from previous iteration is θ^{t-1}. The goal of the EM algorithm is to maximize the expected value of the log likelihood. By using mean-field approximation [52, 53], we maximize the following Q function:

$$Q(\theta^t|\theta^{t-1}) = \mathbb{E}_{P(\vec{H}|\vec{O},\theta^{t-1})}\left[\log P(\vec{H}, \vec{O} \mid \theta^t)\right] \tag{7}$$

$$\begin{aligned} Q(\theta^t|\theta^{t-1}) \propto & \sum_{i\in V} \sum_{h_a \in \mathbb{H}} P(H_i = h_a|O_i, \theta^{t-1}) \log P(O_i|H_i = h_a, \theta^t) \\ & + \sum_{i\in V} \sum_{h_a \in \mathbb{H}} P(H_i = h_a|O_i, \theta^{t-1}) \log P(H_i = h_a|N(i), \theta^t) \end{aligned} \tag{8}$$

where $N(i)$ represents the neighboring nodes of node i, and we can use the estimated hidden states from last iteration to approximate:

$$P(H_i = h_a|N(i), \theta^t) = \prod_{j\in N(i)} P_E(H_i = h_a, H_j = h_j^{\theta^{t-1}}) \tag{9}$$

where $h_j^{\theta^{t-1}}$ is the loopy belief propagation estimated hidden state for node j at iteration $t-1$. We also have:

$$P(O_i|H_i = h_a, \theta^t) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_{\theta_t}^{h_a}|}} \exp\left\{-\frac{1}{2}(O_i - \mu_{\theta_t}^{h_a})^T [\,\Sigma_{\theta_t}^{h_a}]^{-1}\,(O_i - \mu_{\theta_t}^{h_a})\right\} \tag{10}$$

where O_i follows multivariate Gaussian distribution given state $H_i = h_a$.

The first part on the right-hand side of Eq. 8 corresponds to the node potential and the second part corresponds to the edge potential.

In the E-step, we compute the expected states of all nodes. Given the parameter estimation θ^{t-1}, we calculate the posterior probability:

$$P(H_i = h_a|O_i, N(i), \theta^{t-1}) = \frac{P(O_i|H_i = h_a, \theta^{t-1})P(H_i = h_a|N(i), \theta^{t-1})}{\sum\limits_{h'_a \in \mathbb{H}} P(O_i|H_i = h'_a, \theta^{t-1})P(H_i = h'_a|N(i), \theta^{t-1})} \tag{11}$$

In the M-step, we use the maximum likelihood estimation (MLE) to maximize the Q function:

$$\theta^{t*} = \arg\max_{\theta^t} Q(\theta^t|\theta^{t-1}) \tag{12}$$

Loopy belief propagation for state estimation

Given the parameters and observations in the graph, the hidden state inference problem is solved by the loopy belief propagation (LBP) algorithm [54]. LBP works by passing messages among neighboring nodes in the Markov random field structure. Each node passes messages to neighboring nodes when it has received all incoming messages. The passed message from node i to node j about hidden state $H_j = h_b$ is calculated as:

$$m_{i\to j}(H_j = h_b) = \sum_{h_a \in \mathbb{H}} \left[P_V(O_i|H_i = h_a) \times P_E(H_i = h_a, H_j = h_b|O_i, O_j) \right. \\ \left. \times \prod_{k \in N(i)\backslash j} m_{k\to i}(H_i = h_a) \right] \tag{13}$$

where $\mathbb{H}$ is the set of all states, $m_{i\to j}(H_j = h_b)$ represents the message passing from node i to node j about hidden state $H_j = h_b$. $N(i) \setminus j$ refers to the neighbors of node i other than node j. The complete message passed between nodes should be normalized before sending. We normalize the sum of message $m_{i\to j}$ to be 1, i.e.,

$$m_{i\to j}(H_j = h_b) = \frac{m_{i\to j}(H_j = h_b)}{\sum\limits_{h'_b \in \mathbb{H}} m_{i\to j}(H_j = h'_b)} \tag{14}$$

After we send messages from all nodes to their neighbors, we calculate the belief of each nodes based on the node potential and the incoming messages. The belief of node i with hidden state H_i is calculated as:

$$b(H_i = h_a) = \frac{1}{Z'} \times P_V(O_i|H_i = h_a) \times \prod_{k \in N(i)} m_{k\to i}(H_i = h_a) \tag{15}$$

where Z' is the normalization constant:

$$Z' = \sum_{h_a \in \mathbb{H}} \left[P_V(O_i|H_i = h_a) \times \prod_{k \in N(i)} m_{k\to i}(H_i = h_a) \right] \tag{16}$$

Belief is the normalized product of all incoming messages and node potentials, which approximates the marginal probability of each node. Based on belief, we can update the estimated states of each node. To do that, we simply go through all possible hidden states and choose the one with highest belief. LBP runs by iteratively passing messages among neighboring nodes and updating all messages to be sent simultaneously based on previous incoming messages. At the first iteration, the initial messages are all set to 1 before they are normalized. As for the termination condition, we will stop iterations if there is no change of belief or maximum iteration number is reached. We set the maximum iteration to 500 but the method can typically terminate within 100 iterations. All computations are performed in log space to avoid numerical underflow.

Estimation of the number of states

To estimate the number of states, we employed the following strategies: (1) we use Elbow method based on K-means clustering. (2) We calculate Akaike information criterion (AIC) and Bayesian information criterion (BIC) scores to determine the appropriate number of states. (3) We further use the "NbClust" package [55] to identify the best number of states based on various types of clustering metrics.

We applied the Elbow method on the input TSA-seq and DamID data based on K-means clustering. Specifically, we assessed the total within-cluster sum of squares as a function of the number of clusters. The total within-cluster sum of squares is calculated as:

$$\sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{4} (O_{ij} - \overline{O}_{kj})^2 \tag{17}$$

where K is the number of states, C_k is the set of cluster k, and j refers to the 4 different input data types (TSA-seq and DamID). $\overline{O}_{kj}$ is the average score for data j in cluster k. We determined the appropriate cluster number K where additional cluster would not lead to much improvement in terms of the total within-cluster sum of squares. We found that the appropriate number of states may range from 10 to 15 (Additional file 2: Figure S17a).

We first applied the "NbClust" package [55] to assess the optimal number of states. The NbClust package determines the optimal number of states using 30 different metrics for the quality of clustering results. We set the distance measure to "euclidean", cluster analysis method to "kmeans," and the range of state numbers from 2 to 15. We used TSA-seq/DamID signals as input for NbClust to assess the quality of clustering results. We found that the most frequently identified state number is 10 (Additional file 2: Figure S17b).

We further calculated AIC/BIC scores for different number of states (k). The AIC/BIC scores are calculated as follows:

$$AIC = -2\ln(L) + 2K \quad BIC = -2\ln(L) + K\ln(n) \tag{18}$$

where L is maximum likelihood of the model, which is estimated as the product of beliefs for predicted SPIN states on each node, i.e., $L = \prod_{i \in V} b(H_i)$, K is the total number of parameters estimated in the model, and n is the total number of nodes. As shown in Additional file 2: Figure S17c-d, both AIC and BIC scores decrease as the number of states increases. However, we found that the slope of the curve drops close to zero as the number of states exceeds 10.

Taken together, we conclude that the choice of SPIN state number 10 is most appropriate for the datasets we used in this work.

Method for predicting DNA replication timing from SPIN states

We developed a predictive model to demonstrate that multi-fraction Repli-seq can be predicted based on SPIN states and histone marks signals. The model takes SPIN states and 11 histone modification ChIP-seq signals as input, and the 7-fraction Repli-seq score as predictive output. Here, we used 25kb as the window size. We then calculated the average ChIP-seq signals (p value given by MACS2) with each window for 11 histone marks, H2A.Z, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me1, and H3K9me3. Regions with missing values in any dataset were discarded. The discrete SPIN states were transformed into integer numbers ranging from 1 to 10, ranked by the distance from nuclear speckles (1 as Lamina state and 10 as Speckle state). The predicted Repli-seq score for each bin is a seven-dimensional vector, where each dimension corresponds to a specific fraction (S1-S6, and G2).

We then utilized the random forest regressor in scikit-learn [56]. For the prediction model, the input contains SPIN states and histone marks averaged over 25kb bins. The performance of our model was measured by the average R^2 score between real Repli-seq signal and the predicted one in all fractions. We performed a cross-validation on different chromosomes to avoid over-fitting, where we left out one chromosome as test set, and used the remaining chromosomes as training set. The process was repeated for every chromosome and then the results from each fold were averaged. To improve the predictive performance, we performed a parameter scanning for the random forest model and used the parameter set with the highest R^2 score on the training set. The parameters that were tuned include the number of trees (1000), the maximum number of features in each tree (square root of the total number of features), and the maximum depth of the tree (100). The feature importance reported by the random forest regressor was used to select informative features.

We were able to achieve 0.95 R^2 score on the training set and 0.923 R^2 score on the testing set with consistent performance across chromosomes (Additional file 2: Figure S6).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-020-02253-3.

Additional file 1: Supplementary Methods and Supplementary Results.

Additional file 2: Figure S1–S22.

Additional file 3: Table S1. Datasets used in this paper.

Additional file 4: Table S2. List of structural variations in K562 used in this work (combination of [61] and [62]).

Additional file 5: Table S3. GO analysis of genes near ChIP-seq peaks in different TF clusters based on the SPIN states. ChIP-seq peaks on each clusters are merged and used as input for GREAT to calculate GO term enrichment.

Additional file 6: Review history.

Acknowledgments

The authors would like to thank other members in the laboratories of Ma, Belmont, Gilbert, and van Steensel for helpful discussions.

Peer review information

Barbara Cheifet and Andrew Cosgrove were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 6.

Authors' contributions

Conceptualization, JM; methodology, YW and JM; software, YW; resources, LZ, TvS., TS, YC, DPH, DMG, BvS, ASB; investigation, YW, YZ, RZ, DMG, BvS, ASB, and JM; writing – original draft, YW, YZ, RZ, and JM; writing – review & editing,

YW, YZ, RZ, DMG, BvS, ASB, and JM; funding acquisition, ASB, BvS, DMG, and JM. The authors read and approved the final manuscript.

Funding
This work was supported by the National Institutes of Health Common Fund 4D Nucleome Program grants U54DK107965 (A.S.B., B.v.S., D.M.G., and J.M.) and UM1HG011593 (J.M., A.S.B., and D.M.G.) and National Institutes of Health grant R01HG007352 (J.M.). J.M. is additionally supported by a Guggenheim Fellowship from the John Simon Guggenheim Memorial Foundation.

Availability of data and materials
The source code of SPIN can be accessed at: https://github.com/ma-compbio/SPIN [57] under the MIT Open Source License. The DOI for the source code version used in this paper is available at Zenodo: https://doi.org/10.5281/zenodo.4245640 [58]. The datasets generated during the current study are available in the GEO repository: GSE148362 [59] and GSE148609 [60]. All other public datasets used in this work can be found in Additional file 3: Table S1.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh 15213, PA, USA. [2]Oncode Institute and Division of Gene Regulation, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [3]Department of Cell and Developmental Biology, University of Illinois, Urbana 61801, IL, USA. [4]Present Address: Whitehead Institute for Biomedical Research, Cambridge 02142, MA, USA. [5]Department of Biological Science, The Florida State University, Tallahassee 32304, FL, USA. [6]Present Address: Department of Molecular & Cell Biology, University of California, Berkeley 94720, CA, USA.

Received: 8 June 2020 Accepted: 18 December 2020
Published online: 14 January 2021

References

1. Kumaran RI, Thakar R, Spector DL. Chromatin dynamics and gene positioning. Cell. 2008;132(6):929–34.
2. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. 2016;17(11):661.
3. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013;14(6):390.
4. Spector DL. Snapshot: cellular bodies. Cell. 2006;127(5):1071–1.
5. Dundr M, Misteli T. Biogenesis of nuclear bodies. Cold Spring Harb Perspect Biol. 2010;2(12):000711.
6. Takizawa T, Meaburn KJ, Misteli T. The meaning of gene positioning. Cell. 2008;135(1):9–13.
7. Khanna N, Hu Y, Belmont AS. HSP70 transgene directed motion to nuclear speckles facilitates heat shock activation. Curr Biol. 2014;24(10):1138–44.
8. Van Steensel B, Belmont AS. Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. Cell. 2017;169(5):780–91.
9. Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, Mirny LA, O'shea CC, Park PJ, Ren B, et al. The 4D nucleome project. Nature. 2017;549(7671):219.
10. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–93.
11. Kempfer R, Pombo A. Methods for mapping 3d chromosome architecture. Nat Rev Genet. 2020;21(4):207–26.
12. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80.
13. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008;453(7197):948–51.
14. Chen Y, Zhang Y, Wang Y, Zhang L, Brinkman EK, Adam SA, Goldman R, Van Steensel B, Ma J, Belmont AS. Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. J Cell Biol. 2018;217(11):4025–48.
15. Xiong K, Ma J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. Nat Commun. 2019;10(1):5069.
16. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376.
17. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485(7398):381.

18. Meuleman W, Peric-Hupkes D, Kind J, Beaudry J-B, Pagie L, Kellis M, Reinders M, Wessels L, van Steensel B. Constitutive nuclear lamina–genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. 2013;23(2):270–80.
19. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging. 2001;20(1):45–57.
20. Koller D, Friedman N, Bach F. Probabilistic Graphical models: principles and techniques. Cambridge: MIT Press; 2009.
21. Zheng X, Kim Y, Zheng Y. Identification of lamin B–regulated chromatin regions based on chromatin landscapes. Mol Biol Cell. 2015;26(14):2685–97.
22. Marco E, Meuleman W, Huang J, Glass K, Pinello L, Wang J, Kellis M, Yuan G-C. Multi-scale chromatin state annotation using a hierarchical hidden Markov model. Nat Commun. 2017;8(1):1–9.
23. Libbrecht MW, Ay F, Hoffman MM, Gilbert DM, Bilmes JA, Noble WS. Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. Genome Res. 2015;25(4):544–57.
24. Leemans C, van der Zwalm MC, Brueckner L, Comoglio F, van Schaik T, Pagie L, van Arensbergen J, van Steensel B. Promoter-intrinsic and local chromatin features determine gene repression in LADs. Cell. 2019;177(4):852–64.
25. Scott MS, Boisvert F-M, McDowall MD, Lamond AI, Barton GJ. Characterization and prediction of protein nucleolar localization sequences. Nucleic Acids Res. 2010;38(21):7388–99.
26. Vertii A, Ou J, Yu J, Yan A, Pagès H, Liu H, Zhu LJ, Kaufman PD. Two contrasting classes of nucleolus-associated domains in mouse fibroblast heterochromatin. Genome Res. 2019;29(8):1235–49.
27. Ferrari KJ, Scelfo A, Jammula S, Cuomo A, Barozzi I, Stützer A, Fischle W, Bonaldi T, Pasini D. Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity. Mol Cell. 2014;53(1):49–62.
28. Niskanen EA, Malinen M, Sutinen P, Toropainen S, Paakinaho V, Vihervaara A, Joutsen J, Kaikkonen MU, Sistonen L, Palvimo JJ. Global SUMOylation on active chromatin is an acute heat stress response restricting transcription. Genome Biol. 2015;16(1):153.
29. Dileep V, Ay F, Sima J, Vera DL, Noble WS, Gilbert DM. Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program. Genome Res. 2015;25(8):1104–13.
30. Yang Y, Gu Q, Zhang Y, Sasaki T, Crivello J, O'Neill RJ, Gilbert DM, Ma J. Continuous-trait probabilistic model for comparing multi-species functional genomic data. Cell Syst. 2018;7(2):208–18.
31. Zheng X, Hu J, Yue S, Kristiani L, Kim M, Sauria M, Taylor J, Kim Y, Zheng Y. Lamins organize the global three-dimensional genome from the nuclear periphery. Mol Cell. 2018;71(5):802–15.
32. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods. 2015;72:65–75.
33. Nuebler J, Fudenberg G, Imakaev M, Abdennur N, Mirny LA. Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc Natl Acad Sci. 2018;115(29):6697–706.
34. Hua N, Tjong H, Shin H, Gong K, Zhou XJ, Alber F. Producing genome structure populations with the dynamic and automated PGS software. Nat Protoc. 2018;13(5):915.
35. Tian D, Zhang R, Zhang Y, Zhu X, Ma J. MOCHI enables discovery of heterogeneous interactome modules in 3D nucleome. Genome Res. 2020;30(2):227–38.
36. Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, Lai MM, Shishkin AA, Bhat P, Takei Y, et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. Cell. 2018;174(3):744–57.
37. Beagrie RA, Scialdone A, Schueler M, Kraemer DC, Chotalia M, Xie SQ, Barbieri M, de Santiago I, Lavitas L-M, Branco MR, et al. Complex multi-enhancer contacts captured by genome architecture mapping. Nature. 2017;543(7646):519.
38. Zheng M, Tian SZ, Capurso D, Kim M, Maurya R, Lee B, Piecuch E, Gong L, Zhu JJ, Li Z, et al. Multiplex chromatin interactions with single-molecule precision. Nature. 2019;566(7745):558–62.
39. Finn EH, Pegoraro G, Brandao HB, Valton A-L, Oomen ME, Dekker J, Mirny L, Misteli T. Extensive heterogeneity and intrinsic variation in spatial genome organization. Cell. 2019;176(6):1502–15.
40. Belaghzal H, Borrman T, Stephens AD, Lafontaine DL, Venev SV, Marko JF, Weng Z, Dekker J. Compartment-dependent chromatin interaction dynamics revealed by liquid chromatin Hi-C. bioRxiv. 2019704957.
41. Beliveau BJ, Joyce EF, Apostolopoulos N, Yilmaz F, Fonseka CY, McCole RB, Chang Y, Li JB, Senaratne TN, Williams BR, et al. Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc Natl Acad Sci. 2012;109(52):21301–6.
42. Nir G, Farabella I, Estrada CP, Ebeling CG, Beliveau BJ, Sasaki HM, Lee SH, Nguyen SC, McCole RB, Chattoraj S, et al. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. PLoS Genet. 2018;14(12):1007872.
43. Bintu B, Mateo LJ, Su J-H, Sinnott-Armstrong NA, Parker M, Kinrot S, Yamaya K, Boettiger AN, Zhuang X. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science. 2018;362(6413):1783.
44. Falk M, Feodorova Y, Naumova N, Imakaev M, Lajoie BR, Leonhardt H, Joffe B, Dekker J, Fudenberg G, Solovei I, et al. Heterochromatin drives compartmentalization of inverted and conventional nuclei. Nature. 2019;570(7761):395–9.
45. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and samtools. Bioinformatics. 2009;25(16):2078–9.
46. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3(1):95–8.
47. Marchal C, Sasaki T, Vera D, Wilson K, Sima J, Rivera-Mulia JC, Trevilla-García C, Nogues C, Nafie E, Gilbert DM. Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. Nat Protoc. 2018;13(5):819.
48. Consortium EP, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.
49. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):137.

50. Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, Ariyaratne PN, Mohamed YB, Ooi H-S, Tennakoon C, et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol. 2010;11(2):22.
51. Zhan Y, Mariani L, Barozzi I, Schulz EG, Blüthgen N, Stadler M, Tiana G, Giorgetti L. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. Genome Res. 2017;27(3):479–90.
52. Celeux G, Forbes F, Peyrard N. EM procedures using mean field-like approximations for markov model-based image segmentation. Pattern Recog. 2003;36(1):131–44.
53. Zhang J. The mean field theory in EM procedures for markov random fields. IEEE Trans Signal Process. 1992;40(10): 2570–83.
54. Murphy KP, Weiss Y, Jordan MI. Loopy belief propagation for approximate inference: an empirical study. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc.; 1999. p. 467–75.
55. Malika C, Ghazzali N, Boiteau V, Niknafs A. Nbclust: an R package for determining the relevant number of clusters in a data set. J Stat Softw. 2014;61:1–36.
56. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
57. Wang Y, Zhang Y, Zhang R, van Schaik T, Zhang L, Sasaki T, Peric-Hupkes D, Chen Y, Gilbert DM, van Steensel B, Belmont AS, Ma J. SPIN reveals genome-wide landscape of nuclear compartmentalization. GitHub. 2020. https://github.com/ma-compbio/SPIN. Accessed 18 Dec 2020.
58. Wang Y, Zhang Y, Zhang R, van Schaik T, Zhang L, Sasaki T, Peric-Hupkes D, Chen Y, Gilbert DM, van Steensel B, Belmont AS, Ma J. SPIN reveals genome-wide landscape of nuclear compartmentalization. Zenodo. 2020. https://doi.org/10.5281/zenodo.4245640. Accessed 18 Dec 2020.
59. Wang Y, Zhang Y, Zhang R, van Schaik T, Zhang L, Sasaki T, Peric-Hupkes D, Chen Y, Gilbert DM, van Steensel B, Belmont AS, Ma J. SPIN reveals genome-wide landscape of nuclear compartmentalization. Multi-fraction Repli-seq on K562 cell line. Gene Expr Omnibus. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148362. Accessed 18 Dec 2020.
60. Wang Y, Zhang Y, Zhang R, van Schaik T, Zhang L, Sasaki T, Peric-Hupkes D, Chen Y, Gilbert DM, van Steensel B, Belmont AS, Ma J. SPIN reveals genome-wide landscape of nuclear compartmentalization. Genome-wide maps of nucleolus interactions in human cells using 4xAP3 DamID. Gene Expr Omnibus. 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148609. Accessed 18 Dec 2020.
61. Li Y, Zhou S, Schwartz DC, Ma J. Allele-specific quantification of structural variations in cancer genomes. Cell Syst. 2016;3(1):21–34.
62. Dixon JR, Xu J, Dileep V, Zhan Y, Song F, Le VT, Yardımcı GG, Chakraborty A, Bann DV, Wang Y, et al. Integrative detection and analysis of structural variation in cancer genomes. Nat Genet. 2018;50(10):1388.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.