Comparison of the Hi-C, GAM and SPRITE methods using polymer models of chromatin

Luca Fiorillo1,6, Francesco Musella1,6, Mattia Conte1,6, Rieke Kempter2,3, Andrea M. Chiariello1, Simona Bianco1,2, Alexander Kukalev2, Ibai Irastorza-Azcarate2, Andrea Esposito1, Alex Abraham1, Antonella Prisco4, Ana Pombo2,3 and Mario Nicodemi1,2,5✉

Hi-C, split-pool recognition of interactions by tag extension (SPRITE) and genome architecture mapping (GAM) are powerful technologies utilized to probe chromatin interactions genome wide, but how faithfully they capture three-dimensional (3D) contacts and how they perform relative to each other is unclear, as no benchmark exists. Here, we compare these methods in silico in a simplified, yet controlled, framework against known 3D structures of polymer models of murine and human loci, which can recapitulate Hi-C, GAM and SPRITE experiments and multiplexed florescence in situ hybridization (FISH) single-molecule conformations. We find that in silico Hi-C, GAM and SPRITE bulk data are faithful to the reference 3D structures whereas single-cell data reflect strong variability among single molecules. The minimal number of cells required in replicate experiments to return statistically similar contacts is different across the technologies, being lowest in SPRITE and highest in GAM under the same conditions. Noise-to-signal levels follow an inverse power law with detection efficiency and grow with genomic distance differently among the three methods, being lowest in GAM for genomic separations >1Mb.
The SBS model of each of the loci considered was inferred from available Hi-C data and used to derive an ensemble of 3D structures. Those 3D structures were in turn employed to benchmark the performance of in silico Hi-C, SPRITE and GAM experiments. For the Sox9 locus, we also analyzed a polymer model inferred from GAM data and found similar results.

To validate our approach, we demonstrated that in silico average Hi-C, GAM and SPRITE data all successfully compare against corresponding independent experiments, and that our model returns a bona fide representation of chromatin conformations by comparison against independent, single-cell, multiplexed FISH imaging data available for the human HCT116 cell locus. That provides evidence that the architecture of the loci considered is well described by our polymer models and that they can be used to compare the performance of the three technologies with respect to key experimental parameters, including detection efficiency, genomic separation and cell numbers.

We found that in silico Hi-C, GAM and SPRITE bulk data are overall faithful to the reference 3D structures of the polymer models of the loci considered. The intrinsic variability of single-molecule conformations renders single-cell contact data much less faithful to the underlying 3D structure and strongly different across replicates. We identified the minimal number of cells required for replicate experiments to return statistically consistent data, which is shown to be different across the technologies—lowest in SPRITE and highest in GAM under the same conditions. The noise-to-signal level in contact matrices grows as a power law by decreasing efficiency, which implies that experiments using large cell numbers may be required to reduce noise effects, and it varies with genomic distance differently in the three methods, with GAM being the least affected by noise at larger genomic separations.

**Results**

**Derivation of in silico contact maps from known single-molecule 3D structures.** For comparison of in silico Hi-C, GAM and SPRITE data, we focused first on the case study of a 6-Mb region around the Sox9 gene (chr11:109–115 Mb, mm9) in mESCs and its SBS polymer model. The SBS is a model of chromatin where molecules, such as transcription factors, form DNA loops by bridging distal cognate binding sites. It has been shown to accurately describe Hi-C, GAM and FISH data across loci and cell types.

The genomic locations of the binding sites of the Sox9 locus model were inferred from its Hi-C data by the PRISMR algorithm, a machine learning method that determines the minimal set of binding sites (and cognate binders) best describing the input data from only polymer physics. Here, we considered the published model of the locus at 40-kb resolution and explored an ensemble of single-molecule 3D polymer structures derived by molecular dynamics (MD) simulations in the thermodynamic steady state of the system (Fig. 1a and Methods).

To derive Hi-C, GAM and SPRITE in silico contact data, we computationally implemented the steps of the three methods on those 3D structures (Fig. 1a). In brief, with in silico Hi-C we fragmented in equal segments the two polymer chains representing the two Sox9 alleles in each cell, ligated crosslinked fragments and counted ligation products to derive an in silico analog of Hi-C contact frequencies (Methods). The overall efficiency of the process is the product of the in silico crosslinking, digestion, biotinylation, ligation and sequencing efficiencies. In silico SPRITE was similarly implemented by counting chain fragments tagged with the same barcode. Finally, in silico GAM was performed by cutting randomly oriented slices from a sphere (representing the nucleus) where two single-molecule 3D structures (the two ‘alleles’) had been randomly positioned, and by listing the polymer sites falling within each slice, to derive the cosegregation matrix (Methods). The overall efficiency comprises the detection and sequencing efficiency of such sites. The nuclear radius and slice thickness are set to match typical experimental values (Methods).

By such a procedure we derived in silico contact maps of the known polymer 3D structures and investigated how the output of the different technologies is affected by the overall detection efficiency, by the number of pairs, N, of 3D single-molecule structures included in the analysis (below, for simplicity, we refer to N as the number of in silico cells) and how that varies with genomic separation.

**In silico Hi-C, SPRITE and GAM reproduce experimental 3D structure data.** Because our polymer model is inferred from Hi-C data, to check that the derived in silico bulk Hi-C map—that is, contact data averaged over the ensemble of 3D structures—reproduces real bulk Hi-C data in the Sox9 locus (Fig. 1b), we measured their correlation and found that the coefficients Spearman (rs), Pearson (r) and HiRep (stratum adjusted correlation coefficient (scc)) have high values: rs = 0.83, r = 0.83 and scc = 0.80, respectively (Supplementary Table 1a), as previously reported. Similar results were obtained for the HoxD locus in mESC and the Eph4 locus in CHXL-12 cells (Supplementary Figs. 1a,b and 2a and Methods).

To validate our approach, we next compared the in silico SPRITE and GAM contact matrices derived from the same ensemble of model 3D structures with the corresponding, independent SPRITE and GAM experimental matrices, and we found high correlations between model and experiment—respectively, rs = 0.92 and r = 0.79, rs = 0.75 and r = 0.80 (Fig. 1b and Supplementary Table 1a). The HiRep score, albeit designed for comparison of Hi-C data, is also statistically significantly high—respectively, scc = 0.57 and scc = 0.40, (Methods and Supplementary Fig. 3). In the comparison we used published SPRITE bulk mESC data and a GAM dataset produced for the 4D Nucleome Consortium (Methods) composed of 1,122 nuclear slices from F123 mES cells, compared with the output from 1,128 in silico slices. The lower correlation between experimental and in silico GAM contact matrices (derived from Hi-C-based polymers) raises the possibility that Hi-C and GAM may capture some distinct specific contacts, although those differences could derive from noise. Again, similar results were found for the mESC HoxD locus (Supplementary Fig. 1a,b; SPRITE and GAM data are not available for the Eph4 locus in CHXL-12 cells).

To demonstrate that the SBS model 3D structures are a bona fide representation of chromatin conformations in single cells, we took advantage of published multiplex FISH super-resolution microscopy data for a 2.5-Mb region in human HCT116 cells (chr21:34.6–37.1 Mb), because we can compare experimental and model single-molecule 3D structures and Hi-C data (GAM and SPRITE data are not available for that cell type). We repeated the procedure described for the Sox9 locus and compared all-against-all the model-predicted 3D structures with those from imaging data (Fig. 2a). To find the best match between model and experimental structures, each SBS model single-molecule conformation was univocally associated with a corresponding imaged 3D structure, by searching the minimum root mean squared deviation (RMSD) of the coordinates of pairs of rotated and centered 3D structures. To test the significance of the association, as a control we considered self-avoiding random-walk (SAW) polymer chains having the same number of beads and the same average gyration radius—that is, same size—as the real images of the locus (Methods), and we univocally associated each SAW structure with an experimental structure by the least RMSD criterion. Next, we compared the RMSD distribution between SBS structures and their best-matching experimental structure to that between SAW structures and their best-matching experiment (Fig. 2b). The two distributions were found to be statistically different (two-sided Mann–Whitney U-test P = 0), with 93% of the former falling below the first tertile of the
latter (Fig. 2b). We also compared the RMSD distribution between the experimental structures and their best-matching SBS model conformation with the control distribution of RMSD between experimental structures and their best-matching SAW conformation (Fig. 2c). Again, the two distributions are statistically different (P=0.15), with 70% of the former below the first tertile of the control. Finally, we verified that the distribution of RMSD between experimental structures and their best-matching SBS model conformation is statistically indistinguishable from that between the best-matching pairs of experimental structures (P=0.15).

To further validate the model, we checked that the SBS-predicted and microscopy-imaged mean distance matrices, as well as the model and Hi-C bulk contact matrices, have a high correlation (respectively, r_s = 0.96 and r_s = 0.94) (Fig. 2d and Supplementary Fig. 4a,b,d). Importantly, as in the other loci considered (see below), the in silico SPRITE and GAM average matrices also faithfully represent the mean distance data (correlation, r_s = −0.98 and r_s = −0.99, respectively; Supplementary Fig. 4c). As an additional check, to compare the distance matrices of the imaged data, of the SBS and of the SAW models, we computed their genomic-distance-corrected Pearson correlation coefficient, r’ (Methods). We found (Supplementary Fig. 5a–c and Methods) that the mean distance matrix of the SAW model is featureless, with no TADs or patterns, and it has a much lower correlation with the experimental one (r’ = 0.32) than the SBS model mean distance matrix (r’ = 0.84). Finally, we extended the comparison to single-molecule distance data. We computed the distribution of r’ values between the pairs experiment–experiment, experiment–SBS and experiment–SAW single-molecule distance matrices, and found that while the first and second distributions are not statistically distinguishable (two-sided Mann–Whitney U-test P=0.19), the experiment–SAW distribution is clearly different (P=0; Supplementary Fig. 5d).

Taken together, the agreement between model and experiments provides a validation of our polymer model because its 3D structures inferred from Hi-C data accurately recapitulate independent SPRITE, GAM and microscopy data, even at the single-molecule level, consistently across different experiments, loci and cell types. The consistent agreement also shows that our in silico approach has no particular biases favoring Hi-C, SPRITE or GAM.

**In silico bulk Hi-C, SPRITE and GAM data describe the benchmark average distance matrix.** Next, we investigated how well in silico Hi-C, SPRITE and GAM data reflect the spatial architecture of the underlying ensemble of model conformations. In the case study of the Sox9 locus, we computed the average distance matrix of the known 3D structures and compared it with in silico Hi-C, SPRITE and GAM bulk data—that is, averages over in silico cells (Fig. 3). The absolute values of Spearman correlation coefficients (as well as of Pearson and HiCRep correlations; Supplementary Table 1b) of the three methods with the average distance matrix are all high (r_s < −0.89; values are negative because large physical distances correspond to small contact frequencies).

Interestingly, the patterns visible in the in silico Hi-C, SPRITE and GAM bulk data are similar to each other, albeit GAM better highlights longer-range contacts between TADs (Fig. 3a). In particular, all three in silico methods identify the known TADs of the locus9,10,19 (Fig. 3b, different colors in the bottom bar). Additionally, Hi-C, SPRITE and GAM data match the domain-like patterns of the average 3D distance matrix, which represent the typical folding of the reference ensemble of model conformations (Fig. 3b). Similar results are found for the murine HoxD and Eph4 loci and, as already stated, the human locus (Supplementary Figs. 1c,d, 2b,c and 4b,c).

Taken together, our results support the view that bulk data from Hi-C, SPRITE and GAM are faithful to the overall spatial structure of the underlying 3D conformations in murine and human loci and, albeit that differences exist, they provide comparable information on average distances.
Stochasticity of single-cell data reflects the intrinsic variability of single-molecule 3D conformations. Whereas bulk Hi-C data are comparatively similar across replicates, single-cell data exhibit strong variability. Here we explore two sources of such variability: limited detection efficiency and, importantly, inherent differences across single-molecule conformations of chromatin.

To investigate the impact on contact maps of the structural variability of single molecules, we discuss first the ideal case of in silico experiments where the efficiency is set to 100%. Consistent with single-cell imaging data, single-molecule conformations vary widely in the ensemble of model 3D structures (Fig. 4a, top) and their single-cell distance matrices (Fig. 4a, bottom) have broadly varying Spearman correlations with the average distance matrix (Fig. 4b; mean $r_s = 0.88$; Supplementary Table 1c). Additionally, the correlation of an in silico single-cell Hi-C, SPRITE or GAM contact map (Fig. 4c) with its corresponding single-cell distance matrix is much lower than in the case of the bulk data previously discussed, with, on average, $r_s = -0.37$ and $r_s = -0.46$ for, respectively, in silico Hi-C and SPRITE (Fig. 4d and Supplementary Table 1d). As expected, for GAM the correlation between single-cell maps is even lower (average $r_s = -0.15$) and its distribution much broader, in the range $-0.4 < r_s < 0$. That is also a consequence of the different experimental procedures: while a single-cell in silico Hi-C or SPRITE experiment returns the contacts measured over an entire in silico nucleus—that is, two independent polymer structures representing the alleles—a single-cell in silico GAM experiment probes the polymer content of only a single slice of an in silico nucleus—that is, a tiny fraction of the two polymers.

Contact data from single-cell experiments become further deteriorated at lower values of detection efficiency and have worse correlations with the corresponding single-cell distance maps (Supplementary Table 1e). Consequently, the variability of replicates from in silico single-cell experiments increases and the correlation between their contact maps correspondingly decreases. For example,
for an efficiency of 0.5, we found that the average correlation between in silico single-cell replicates is around $r_s = 0.2$, 0.4 and 0.1 for Hi-C, SPRITE and GAM maps, respectively. While the impact of limited efficiency on contact maps is systematically investigated in the following section, here we stress that the values of correlation measured between single-cell replicates are consistent with those reported in real experimental studies. For example, in real single-cell experiments in CD4 T\textsubscript{h}1 cells with an efficiency of approximately 0.025 (ref. 55), the average Spearman correlation between different Hi-C maps of the Sox9 locus was found to be $r_s = 0.01$, which is numerically equal to the value found for the same efficiency in our model of mESC (Methods and Supplementary Table 1b).

To summarize, the variability of in silico single-cell Hi-C, SPRITE and GAM data reflects the intrinsic structural differences across chromatin single molecules, and they are less faithful than bulk data to the corresponding single-cell distances even in the ideal case of

---

**Fig. 3 | Bulk Hi-C, SPRITE and GAM data are faithful overall to average 3D distances.**

**a.** In silico bulk Hi-C, SPRITE and GAM maps of the Sox9 locus are shown and compared to the average 3D distance matrix in **b** of the known single-molecule 3D conformations of the locus model. The color scale represents the percentiles in each dataset. Spearman correlation coefficients are reported, on the bottom, between each contact map and the average 3D distance matrix (Pearson and HiCRep are reported in Supplementary Table 1b).

**b.** Average 3D distance matrix derived from the ensemble of in silico model single-molecule 3D conformations.

**Fig. 4 | Stochasticity of single-cell contact maps reflects the intrinsic variability of single-molecule 3D conformations.**

**a.** In silico single-cell distance maps (top) and average distance matrix (top right) of the Sox9 locus are shown. The variability of the corresponding single-molecule conformations of the model is represented in the bottom of the panel while the color scheme reflects the colors of the TADs of the locus\textsuperscript{19}, shown in the color bar. The distribution of Spearman correlations between in silico single-cell distance maps and the average distance map is shown (mean value $r_s = 0.88$). Mean Pearson and HiCRep correlations are reported in Supplementary Table 1c.

**b.** The distribution of Spearman correlation coefficients between in silico single-cell distance matrices at efficiency $= 1$ and their corresponding in silico single-cell distance matrices.

---

_**NATuRE METHodS** | VOL 18 | MAy 2021 | 482–490 | www.nature.com/naturemethods_
100% detection efficiency. Lower efficiencies further increase fluctuations, to the point where single-cell replicates can have correlations well below 0.1 at realistic efficiencies, hence explaining the variability of single-cell experiments.4,54–57.

The number of cells required for replicate reproducibility differs among Hi-C, SPRITE and GAM. The quality of in silico Hi-C, SPRITE and GAM contact maps improves when the number of in silico cells (N) in the experiment increases (Fig. 5a). Figure 5b shows, for example, the effect of N in the case of efficiencies comparable to typical experimental values: we set the Hi-C efficiency to 0.05, taken as an upper limit of values reported in recent studies;54–56 the same value is used as an estimate of the efficiency for SPRITE. Because the experimental efficiency in GAM is roughly one order of magnitude larger than in Hi-C and SPRITE, in the example shown we used an in silico GAM efficiency of 0.5 (Methods).

Importantly we checked that, in the large N limit, (1) the contact matrices overall are not dependent on the efficiency value considered and (2) the average over a large number of cells compensates for reduced efficiencies (see below and Methods). Indeed, as shown in Fig. 5a,b, the patterns of contact matrices become sharper and stabilize when N becomes sufficiently large, as also observed in experimental investigations.13,52–54. However, our data show that the threshold value of N to reach saturation is strongly dependent on the efficiency level and varies with different technologies.

We aimed to identify the minimal number of cells that, at a given efficiency, is required for replicate in silico experiments to return similar outputs—that is, to approach the bulk limit. To measure the similarity of pairs of identical experiments in the Sox9 locus (all with the same N and efficiency; Fig. 5c), we computed the average Pearson correlation between their contact maps (Fig. 5d,e; Spearman and HiCRep correlations returned analogous results; Supplementary Fig. 7). The correlation grows when N is increased and plateaus to 1 at the large N limit (Fig. 5d,e), independently of the efficiency of the in silico experiment. For each given efficiency, we heuristically defined the minimal number of cells, M, required for statistically reproducible results across replicates as the value of N where the correlation grows larger than a given threshold, r = 0.9 (Methods). Interestingly, we found that M is different in the different technologies: for example, if the efficiency is 0.1 we found that M is 200, 100 and 2,000 for Hi-C, SPRITE and GAM, respectively (Fig. 5d). Figure 5e shows the correlation between replicates at varying N obtained for efficiencies close to those reported in real Hi-C, SPRITE and GAM experiments—that is, as specified above, 0.05 for Hi-C and SPRITE and 0.5 for GAM: in those cases, M is approximately 650, 250 and 800, respectively. Additionally, we checked that
our estimates of $M$ compare well against available experimental estimations (Methods and Supplementary Fig. 6b).

Next, we systematically investigated how the quality of in silico data is affected by the efficiency of the experiment (Supplementary Fig. 11). We found that the number of cells required for replicate similarity ($M$) is strongly dependent on efficiency (Fig. 5f): $M$ diverges approximately as an inverse squared power law as the efficiency becomes small. In other words, halving the efficiency requires the quadrupling of cell number to achieve the same quality levels. In general, we find that $M$ for SPRITE is two times smaller than the corresponding value for Hi-C and one order of magnitude smaller than for GAM. Additionally our investigation shows that, even in the ideal case of efficiency $= 1$, single-cell replicates have below-threshold correlations because $M > 10$ even for SPRITE, due to the intrinsic variability of single-molecule 3D structures, as reported above.

Similar results were found for the murine HoxD, Epha4 and human HCT116 loci (Supplementary Figs. 8a–d, 9a–d and 10a–d). The sets of in silico single-molecule 3D structures employed in all our analyses were produced using polymer models inferred from Hi-C data\textsuperscript{25–28}. However, for the Sox9 locus we tested that our results remained unchanged overall also when the polymer model of the locus is inferred from GAM contact maps rather than from Hi-C\textsuperscript{28} (Supplementary Figs. 13 and 14a–d and Methods). Additionally, to assess the general validity of our analyses, we applied the in silico approach to 3D conformations of a toy block-copolymer, unrelated to real chromatin loci, finding similar results (Supplementary Fig. 15 and Methods). Finally, we checked that the above definition and features of $M$ can be fully grounded on the central limit theorem (CLT; Methods and Supplementary Fig. 12).

GAM cosegregation data, as mentioned before, include both random and nonrandom cosegregation events (that is, specific interactions) that can be dissected by the use of statistical methods such as SLICE\textsuperscript{9} (Methods). Hence, we investigated the performance of SLICE on in silico GAM data when both $N$ and efficiency varied (Supplementary Fig. 16). SLICE returns, in particular, the single-cell interaction probability of pairs, and multiplets, of DNA sites\textsuperscript{9}. We found that SLICE bulk interaction probabilities are faithful to the interaction probability of pairs, and multiplets, of DNA sites\textsuperscript{9}. We found that SLICE bulk interaction probabilities are faithful to the known average distance matrix ($r = -0.95$, $r_s = -1.00$, $sc\epsilon = -0.99$; Supplementary Fig. 16a,b) and the SLICE matrices behave with both $N$ and efficiency as found for GAM contact maps. Because by definition SLICE specifically detects significant interactions, however, we found that the average number of in silico cells ($M$) needed to return statistically reproducible results across replicates is approximately half that required for GAM alone under the same conditions (Supplementary Fig. 16c–e). For a realistic efficiency of 0.5, for example, we found that $M = 400$ for SLICE whereas $M = 800$ for GAM. In that respect, SLICE can be employed as a useful tool to enhance the performance of GAM, especially in applications where the number of available cells is small, such as in the analysis of sample tissues or biopsies.

Our findings illustrate how the level of variability of in silico contact matrices is affected by both the $N$ value and experimental efficiency, and how different technologies perform under different situations. Consistent with the CLT, the number of cells required for replicate similarity ($M$) grows as an inverse squared power law as efficiency decreases.

Noise-to-signal levels vary differently with genomic distance in Hi-C, SPRITE and GAM. Finally, we investigated the noise-to-signal level of the entries of contact matrices and how it varies with genomic separation, with $N$ and with the efficiency of in silico experiments. For each entry of a contact map, the noise-to-signal ratio is defined as the ratio of the standard deviation, $\sigma$, to the mean value, $\mu$, across replicate experiments under the same conditions. For a given $N$ and efficiency, we observed that the average noise-to-signal ratio, $\sigma/\mu$, is strongly dependent on genomic distance (Fig. 6a and Methods). In the Sox9 locus, we found for both Hi-C and SPRITE that $\sigma/\mu$ increases by more than one order of magnitude as genomic separation increases from 50 kb to 5 Mb. In particular, there is a steep increase above the 1-Mb level. SPRITE has the lowest $\sigma/\mu$ ratio at genomic scales below 1 Mb but, interestingly, GAM has an overall lower varying noise-to-signal level, especially at large genomic separations (>1 Mb) where it is almost one order of magnitude lower than Hi-C and SPRITE.

At a given genomic distance and efficiency, as expected, the noise-to-signal ratio decreases as $N$ increases (Fig. 6b). Consistent with the CLT, it follows an inverse squared power law in $N$ (that is, $N^{-1/2}$). One consequence of such a scaling behavior is that single-cell ($N=1$) contact maps become highly noisy at large genomic separations. For example, at the 1-Mb scale and for a detection efficiency of 0.5, the noise-to-signal ratio for $N=1$ is $>100\%$ for all three methods, with Hi-C having the largest fluctuations ($\sigma/\mu>1,000\%$). As expected, the noise-to-signal ratio is also strongly affected by experimental efficiency (Fig. 6c): in brief we find that, in our in silico study, for a given genomic distance and a given $N$, $\sigma/\mu$ decreases roughly as an inverse power law of efficiency.

Similar results were obtained for the murine HoxD, Epha4 and human HCT116 loci (Supplementary Figs. 8e–g, 9e–g and 10e–g). Finally, we extended all the analyses done on the model of the Sox9 locus derived from Hi-C data to the polymer model derived from GAM data\textsuperscript{28} and acquired results fully consistent with our previous findings, supporting the broader validity of our approach (Supplementary Fig. 14e–g).

Discussion

Hi-C, SPRITE and GAM are powerful technologies utilized to probe DNA contacts genome wide. We discuss a quantitative benchmark to assess how well these different methods represent the 3D structure
of the genome and how they perform relative to each other under different experimental conditions. Our approach is based on computer simulations of their performance in capturing the architecture of a known set of polymer 3D structures. We focused on the models of regions around genes Sox9 and HoxD in murine ES cells, around the Epha4 gene in CHLX-12 cells and in human HCT116 cells, as well as on a toy block-copolymer model. We analyzed in silico Hi-C, SPRITE and GAM with different experimental parameters including the cell numbers considered, detection efficiency and genomic separation scales (Table 1). There is a consistent agreement between independent Hi-C, SPRITE and GAM data and our in silico models across the studied loci and cell types (Fig. 1). Additionally, in the human HCT116 locus for which single-cell microscopy data are available, we verified that the model polymer conformations provide a bona fide representation of the experimental single-molecule 3D structures (Fig. 2).

We find that in silico Hi-C, SPRITE and GAM bulk contact data, as well as SLICE interaction probabilities, faithfully represent the known spatial conformations of model polymers (Fig. 3). Single-cell contact data reflect the intrinsic broad structural variability of chromatin single molecules and are much less faithful to their corresponding single-cell distance matrices (Fig. 4). Because single-cell GAM captures a single slice of a nucleus rather than the entire nucleus, its fluctuations are even stronger than in single-cell Hi-C and SPRITE.

The minimal number of in silico cells (M) to be considered in an experiment for replicates to return sufficiently similar contact patterns (Fig. 5) increases approximately as an inverse squared power of the cell number employed in in silico experiments, and grows approximately as a power law by decreasing efficiency (Fig. 6). Genomic distance also impacts the noise-to-signal ratio (Fig. 6): for identical conditions, GAM is less noise sensitive at large genomic separation (say, >1 Mb).

Overall, our computational analyses are consistent across the investigated cases (models of both real loci and toy models), supporting the view that quantitative comparison of the performance of in silico Hi-C, SPRITE and GAM has a more general validity.

**Online content**
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-021-01135-1.

Received: 29 May 2020; Accepted: 31 March 2021; Published online: 7 May 2021

### References
1. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* **21**, 207–226 (2019).
2. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
3. Fullwood, M. J. et al. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
4. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
5. Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
6. Cuddaio, A. M. et al. Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. *Nat. Genet.* **50**, 1744–1751 (2018).
7. Hsieh, T. H. S. et al. Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell* **162**, 108–119 (2015).
8. Kristenstein, N. et al. Ultrastructural details of mammalian chromosome architecture. *Mol. Cell* **78**, 554–565 (2020).
9. Beagrie, R. A. et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).
10. Quinodoz, S. A. et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174**, 744–757 (2018).
11. Bickmore, W. A. The spatial organization of the human genome. *Annu. Rev. Genomics Hum. Genet.* **14**, 67–84 (2013).
12. Dekker, J. & Misteli, T. Long-range chromatin interactions. *Cold Spring Harb. Perspect. Biol.* **7**, a019356 (2015).
13. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16**, 245–257 (2015).
14. Dekker, J. & Mirny, L. The 3D genome as moderator of chromosomal communication. *Cell* **164**, 1110–1121 (2016).
15. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Nat. Rev. Genet.* **19**, 453–467 (2018).
16. Finn, E. H. & Misteli, T. Molecular basis and biological function of variability in spatial genome organization. *Science* **365**, eaaw9498 (2019).
17. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
18. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
19. Fraser, I. et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.* **11**, 852 (2015).
20. Cattoni, D. L. et al. Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. *Nat. Commun.* **8**, 1753 (2017).
21. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, eaau1783 (2018).
22. Cardozo Guizi, A. M. et al. Microscopy-based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms. *Mol. Cell* **74**, 212–222 (2019).
23. Finn, E. H. et al. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* **176**, 1502–1515 (2019).
24. Chiarriello, A. M., Annunziatella, C., Bianco, S., Esposito, A. & Nicodemi, M. Polymer physics of chromosome large-scale 3D organisation. *Sci. Rep.* **6**, 29775 (2016).
25. Bianco, S. et al. Modeling single-molecule conformations of the HoxD region in mouse embryonic stem and cortical neuronal cells. *Cell Rep.* **28**, 1574–1583 (2019).
26. Bianco, S. et al. Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat. Genet.* **50**, 662–667 (2018).

### Table 1 | Summary of the performance of in silico Hi-C, SPRITE and GAM under identical conditions

| Method | Faithfulness to 3D structure | Replicate similarity versus cell number | Noise versus detection efficiency |
|--------|-----------------------------|----------------------------------------|-------------------------------|
|        | Bulk data                   | Single-cell                            | <1Mb                          | >1Mb |
| Hi-C   | +++                         | +                                       | +++                           | +    |
| SPRITE | +++                         | +++                                    | +                             | +    |
| GAM    | +++                         | +                                       | +                             | +++  |
28. Conte, M. et al. Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation. Nat. Commun. 11, 3289 (2020).
29. Lupiáñez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene–enhancer interactions. Cell 161, 1012–1025 (2015).
30. Andrey, G. et al. A switch between topological domains underlies HoxD genes collinearity in mouse limbs. Science 340, 1234167 (2013).
31. Noordermeer, D. et al. The dynamic architecture of Hox gene clusters. Science 334, 222–225 (2011).
32. Li, Q. et al. The three-dimensional genome organization of Drosophila melanogaster through data integration. Genome Biol. 18, 145 (2017).
33. Serra, F. et al. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput. Biol. 13, e1005665 (2017).
34. Nir, G. et al. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. PLoS Genet. 14, e1007872 (2018).
35. Lin, D., Bonora, G., Yarhamci, G. G. & Noble, W. S. Computational methods for analyzing and modeling genome structure and organization. Wiley Interdiscip. Rev. Syst. Biol. Med. 11, e1355 (2018).
36. Tjong, H. et al. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proc. Natl Acad. Sci. USA 113, E1663–E1672 (2016).
37. Bohn, M. & Heereman, D. W. Diffusion-driven looping provides a consistent framework for chromatin organization. PLoS ONE 5, e12218 (2010).
38. Barbieri, M. et al. Complexity of chromatin folding is captured by the strings and binders switch model. Proc. Natl Acad. Sci. USA 109, 16173–16178 (2012).
39. Brackley, C. A., Taylor, S., Papantonis, A., Cook, P. R. & Marenduzzo, D. Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. Proc. Natl Acad. Sci. USA 110, E3605–E3611 (2013).
40. Jost, D., Carrivain, P., Cavalli, G. & Vaillant, C. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Res. 42, 9553–9561 (2014).
41. Sanborn, A. I. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl Acad. Sci. USA 112, E6456–E6465 (2015).
42. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. Cell Rep. 15, 2038–2049 (2016).
43. Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G. & Onuchic, J. N. Transferable model for chromosome architecture. Proc. Natl Acad. Sci. USA 113, 12168–12173 (2016).
44. Buckle, A., Brackley, C. A., Boyle, S., Marenduzzo, D. & Gilbert, N. Polymer simulations of heteromorphic chromatin predict the 3D folding of complex genomic loci. Mol. Cell 72, 786–797 (2018).
45. Fiorillo, L. et al. A modern challenge of polymer physics: novel ways to study, interpret, and reconstruct chromatin structure. Wiley Interdiscip. Rev. Comput. Mol. Sci. 10, e1454 (2019).
46. Shi, G., Liu, L., Hyeon, C. & Thrulallal, D. Interphase human chromosome exhibits out of equilibrium glassy dynamics. Nat. Commun. 9, 3161 (2018).
47. Nicodemi, M. & Frisco, A. Thermodynamic pathways to genome spatial organization in the cell nucleus. Biophys. J. 96, 2168–2177 (2009).
48. Fiorillo, L. et al. Inference of chromosome 3D structures from G4M data by a physics computational approach. Methods 181–182, 70–79 (2020).
49. Barbieri, M. et al. Active and poised promoter states drive folding of the extended HoxB locus in mouse embryonic stem cells. Nat. Struct. Mol. Biol. 24, 515–524 (2017).
50. Kragstein, B. K. et al. Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. Nat. Genet. 50, 1463–1473 (2018).
51. Yang, T. et al. HiCReP: assessing the reproducibility of Hi-C data using a stratified-adjusted correlation coefficient. Genome Res. 27, 1939–1949 (2017).
52. Dekker, J. et al. The 4D nucleome project. Nature 549, 219–226 (2017).
53. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. Cell 171, 305–320 (2017).
54. Stevens, T. J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature 544, 59–64 (2017).
55. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59–64 (2013).
56. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature 547, 61–67 (2017).
57. Flyunger, I. M. et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature 544, 110–114 (2017).
58. Ramani, V. et al. Massively multiplex single-cell Hi-C. Nat. Methods 14, 263–266 (2017).
59. Lando, D., Stevens, T. J., Basu, S. & Laue, E. D. Calculation of 3D genome structures for comparison of chromosome conformation capture experiments with microscopy: an evaluation of single-cell Hi-C protocols. Nucleus 9, 190–201 (2018).
60. Diaz, N. et al. Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. Nat. Commun. 9, 4938 (2018).
Methods

The SBS model. The SBS model describes a chromatin segment as a self-avoiding polymer chain comprising beads interacting with diffusing molecules, called binders. Along the polymer chain, binders can interactively attach with beads and act as building blocks. The interaction is specific, such that different types of binding sites interact only with their associated binders. The model can be visually represented as a chain with beads of different colors, where only same-color bead–binder pairs interact. Because the interaction is multivalent, a binder can attach to multiple beads. The number of colors (that is, types of binding sites) and their location along the polymer chain and the concentration of binders regulate the folding dynamics and equilibrium conformations of an SBS polymer.

Inference of the SBS model of a DNA locus from experimental data. The SBS model of a given genomic locus is a polymer with a convenient set of binding sites (colors) suitably arranged along the chain. To infer the genomic positions of different colors, a machine learning method, named PRISMR, is employed.

Briefly, PRISMR finds the SBS polymer that best reproduces a given input contact matrix, without any previous assumptions, based on only the principles of polymer physics. Input contact matrices—such as Hi-C or GAM—describing a genomic region of length L base pairs (bp) at resolution res are \( N_{\text{bin}} \times N_{\text{bin}} \)-sized, where \( N_{\text{bin}} = L / \text{res} \) is the number of DNA windows (bins) in the region considered. In general, the best SBS polymer describing such a DNA region will be composed by \( N_{\text{bin}} \times N_{\text{bin}} \times n \) beads, where \( n \) accounts for the presence of different binding sites in a bin and is given by PRISMR.

Once the polymer is found, for the locus of interest is found, its 3D conformation is derived by a massive parallel MD computer simulation at thermodynamic equilibrium. This is repeated for several independent runs to yield an entire ensemble of single-molecule 3D equilibrium structures. All MD simulations are carried out using the free software LAMMPS, with interaction potentials previously established in the literature. Full details of the models employed and simulation parameters can be found in the referenced papers. This entire approach has been successfully used to reproduce experimental data (FISH, Hi-C or GAM) and provided correct predictions on the impact of mutations over 3D DNA organization.

Model 3D structures of murine Sox9, HoxD and Epha4 and human HCT116 cell loci. The SBS polymer model of the murine Sox9 locus (chr1:1109–115 Mb, mm9) was inferred by PRISMR from mESC Hi-C data at 40-kb resolution, and consists of 2,250 beads and 15 types of binding site (colors). The ensemble of its 3D structures derived by MD is composed of 500 configurations. Full details on MD simulation parameters are previously described.

To check the robustness of the approach, we also employed a second ensemble of SBS 3D structures describing the same genomic region but derived from GAM mESC 40-kb cosegregation data. The polymer consists of \( N_{\text{bin}} = 2,250 \) and 15 colors and the ensemble comprises 450 equilibrium structures. Details of the MD simulation are previously published and the parameters used are the same as previously described.

The SBS polymer model for the murine HoxD locus (chr2:71–78 Mb, mm9) was derived by PRISMR from mESC Hi-C data at 40-kb resolution. The polymer comprises \( N_{\text{bin}} = 2,100 \), with 12 colors; the ensemble comprises 500 equilibrium configurations. Full details of the MD simulation are previously reported.

In regard to the murine Epha4 locus (chr17:73–9 Mb, mm9), the polymer model was obtained by PRISMR applied to CHLX-12-cell in situ Hi-C data at 10-kb resolution. Here, the polymer is \( N_{\text{bin}} = 12,600 \) beads long, with 21 colors. The ensemble comprises 500 equilibrium configurations. Full details of MD simulation parameters are previously reported.

The SBS model of the human HCT116 locus (chr2:34.6–37.1 Mb, hg38) was inferred from Hi-C data at 30-kb resolution. The polymer consists of 830 beads and four types of binding site, and MD simulations provided 1,000 3D equilibrium conformations. All further details of MD simulations can be found in previous work.

3D structures of the block-copolymer toy model. To test our algorithms and results, we also employed a toy SBS block-copolymer model unrelated to real DNA loci. We considered a polymer chain of 210 beads with only three different colors (Supplementary Fig. 15a); two of those colors (green and red) were arranged in separate blocks. Visually, we colored in green the first block and in red the second. In addition, to permit specific contacts across both blocks, we introduced a single bead of a third color, represented in blue, at approximately the middle position of each block. Potentials and parameters used for MD simulations were the same as previously described and, in particular, the interaction energy scale was set at \( 9k_{\text{B}}T \) for the red and green beads and at \( 12k_{\text{B}}T \) for the blue beads, with \( k_{\text{B}} \) being the Boltzmann constant and temperature \( T = 300 \) K. The interaction range was set at 1.5 \( r \) for the red and green beads and 1.5 \( r \) for the blue, where \( r \) is the bead diameter. Four binders of the blue type and 88 each of the red and green were employed to match the polymer fold. At equilibrium, the green and blue blocks segregate separately forming two globular domains while the isolated blue beads produce a long-range contact stretching out from their respective blocks (Supplementary Fig. 15a,b). The ensemble is composed of 150 structures. Here, a bin is assumed to be formed by a single bead, such that \( N_{\text{bin}} = N_{\text{bin}}^2 \).

GAM data. GAM data were obtained from mouse ESCs (clone F123), a male cell line derived from an F1 Mus musculus castaneus \( \times S.29/S.29 \) pair, kindly provided by B. Ren (University of California San Diego, San Diego, CA, USA). Normalized contact matrices from individual nuclear profiles (NPs) of the F123 cell line at 40 kb were obtained from the 4D Nucleome Consortium and are available on the 4D Nucleome data portal under accession no. 4DNFIPS01. Normalization of cosegregation frequencies was performed using pointwise mutual information (PMI). The PMI of genomic windows i and j describes the difference between the probability of both windows being found in the same NP (that is, their joint distribution) and their individual distributions across all NPs. PMI assumes that finding one window is independent of finding the second. Specifically, for windows i and j, the value of PMI is

\[
\text{PMI} = \log \left( \frac{p(i 
 j)}{p(i) \times p(j)} \right)
\]

where \( p(i) \) is the frequency whereby window i was found across the NPs (\( p(j) \) is analogously defined) and \( p(i 
 j) \) is the frequency that the two windows are found together across the NPs. PMI is then normalized (NPMI) by bounding between −1 and 1 as follows:

\[
\text{NPMI} = \frac{\text{PMI}}{\log p(i) \times p(j)}
\]

Comparison between experimental and in silico contact matrices. Quantitative comparisons between in silico and experimental matrices were performed computing the Spearman (\( r \)) and Pearson (\( r \)) correlation coefficients, and the stratum adjusted correlation coefficient (ssc) from the HiCRep method, developed specifically for Hi-C data. The experimental data used for comparisons (Fig. 1b, Supplementary Table 1a and Supplementary Fig. 13a) with the in silico data of the Sox9 region are Hi-C and SPRITE in mESC and GAM from the F123-cell 1,122 × INP dataset. All datasets are available on the 40-kb resolution. For the HoxD locus, the experimental data used (Supplementary Fig. 1a) are the same as for the Sox9 locus. For the Epha4 locus, the Hi-C data in mouse CHLX-12 cells at 10 kb were employed for comparison with our in silico Hi-C map (Supplementary Fig. 2a). All datasets are aligned to the mm9 assembly.

In silico distance matrices. To compute the distance matrix of a given polymer model, we calculate the Euclidean spatial distance between all pairs of beads, obtaining a two-dimensional (2D) matrix \( N_{\text{bin}} \times N_{\text{bin}} \). This entire approach has been successfully used to reproduce experimental data (FISH, Hi-C or GAM) and provided correct predictions on the impact of mutations over 3D DNA organization.

In silico simulations of Hi-C, SPRITE and GAM. The codes used to implement in silico Hi-C, SPRITE and GAM take as input the spatial coordinates of the beads of our polymer 3D structures. To take into account the two alleles present in each single cell, we explicitly consider pairs of independent structures in our simulations. By repeating the algorithms over different, randomly selected polymer pairs, we simulate experiments carried over a population of cells. All codes are written in the C programming language.

In silico Hi-C. For in silico Hi-C experiments we implemented a proxy of the key steps of a Hi-C protocol—the so-called ligation and contact matrix generation, as described in detail in the following sections. In particular we applied those steps separately to each in silico cell, as in a real single-cell Hi-C experiment.

Crosslinking. During real Hi-C crosslinking, DNA contacting sites are bound together with formaldehyde to fix the overall 3D structure. Formaldehyde binds to DNA–protein complexes and consequently fixes DNA sites that are linked by...
In our SBS polymers only same-colored beads interact with each other through a binder and only if they are closer than a threshold distance \( d \), fixed by the interaction energy cutoff. Thus, for in silico Hi-C we crosslink beads of the same color and that are closer than \( d \). This is done with efficiency \( p_{\text{b}} \), simulating the experimental one. To identify the sets of crosslinked beads, a customized version of the DBSCAN clustering algorithm is employed.

Digestion. After crosslinking, DNA is digested—that is, cut into fragments. In standard Hi-C experiments, digestion fragments have a median length in the range from few hundreds of base pairs to several kilobases, depending on the restriction enzyme used. Importantly, in the SBS models of the Sos9, HoxD, Epha4 and human loci, because the genomic content of each bead falls within that range a single polymer bead is a good representation of the average digestion fragment. Thus, we implement digestion by splitting the polymer chain into its own individual beads. As result, we acquire a set of independent clusters consisting of crosslinked beads.

Biotinylation. The next step is biotinylation, where DNA fragments in each crosslinked cluster are marked with biotin. Unmarked fragments cannot be detected in Hi-C. In our algorithm, biotinylation is implemented by removing beads from their clusters with probability \( 1 - p_{\text{b}} \) with \( p_{\text{b}} \) modeling biotinylation efficiency.

Ligation. In Hi-C, crosslinked and biotinylated pairs of fragments are randomly linked, forming contacting pairs. This is implemented by random selection of pairs of beads of the same color and that are closer than \( d \). To account for experimental ligation efficiency, each selected bead is ligated with a probability of only \( p_{\text{lig}} \) and is otherwise discarded.

Contact matrix generation. Next in Hi-C, ligated fragments are sequenced and a contact is counted between their corresponding bins, eventually producing a contact matrix \( N_{\text{c}} \times N_{\text{c}} \). Similarly, in our algorithm we produce a 2D \( N_{\text{c}} \times N_{\text{c}} \) matrix. For each polymer structure in input, ligated beads are counted as a contact with given detection probability \( p_{\text{d}} \)—modeling the sequencing efficiency of real experiments—and their corresponding matrix entry is incremented by 1. The procedure is iterated over the \( N \) simulated cells, and the final in silico matrix yields the total count of contacts between each possible pair of bins.

In silico SPRITE. For SPRITE, we implemented the main steps of its protocol—crosslinking, digestion, split-pool tagging and contact matrix generation.

Crosslinking. In SPRITE experiments, crosslinking is carried out as in Hi-C, so the same procedure described above for our in silico Hi-C is employed.

Digestion. After crosslinking, in SPRITE experiments DNA is fragmented first by sonication and then by DNase digestion, resulting in a collection of crosslinked fragments of approximately 150–1,000 bp. However, in our algorithm, we implement SPRITE chromatin digestion with an algorithm analogous to that used for in silico Hi-C.

Split-pool tagging. The split-pool tagging procedure allows identification of DNA fragments belonging to the same crosslinked cluster, because they are uniquely barcode linked and all DNA fragments belonging to the same cluster are associated with a specific tag sequence. In our in silico procedure, because the beads composing a given cluster are known, an explicit split-pool tagging implementation is not required. However, to take into account the fact that in real experiments some fragments may not be tagged successfully, we remove beads from their clusters with probability \( 1 - p_{\text{p}} \).

Contact matrix production. Experimentally, fragments with the same barcode are sequenced and assigned to their corresponding genomic windows—that is, the bins comprising the contact map. In this way it is possible to define those bins associated with a fixed cluster. A contact is then counted for every possible pair of bins associated with a cluster. The count is weighted by a corrective factor \( 2/n \), taking into account the size of the cluster, with \( n \) the number of fragments in the cluster. In our in silico procedure, because each fragment is represented by a polymer bead, we assign a count to a bin pair if at least one bead from each bin is found in the same cluster. As with in silico Hi-C, each bead is detected with only given probability \( p_{\text{d}} \), modeling sequencing efficiency. Each contact count over the population of \( N_{\text{c}} \) cells is then multiplied by its weight, \( 2/n \), and finally collected in a \( N_{\text{c}} \times N_{\text{c}} \) matrix.

In silico GAM. In GAM experiments, a nuclear slice is cut in random orientation from a cell nucleus, the DNA sites in the slice are sequenced and their co-occurrence in a collection of slices measured to construct a GAM cosegregation matrix. In our algorithm, nuclear slicing and cosegregation matrix generation were implemented as follows.

Slice cutting. We model a cell nucleus as a sphere containing two different, randomly placed, polymer structures of the locus of interest, to take into account both alleles present in the considered cell type. For each in silico cell we generate a randomly oriented slice within the sphere, and all polymer beads inside are counted as cosegregating. The in silico nuclear radius and slice thickness are set, respectively, to the scale of the experimental one. To match the scales of the experimental values in the considered cell types, the simulated slices may not contain the specific locus of interest, as in a real GAM experiment, where cellular slices contain only a fraction of the nucleus. To account for experimental detection and sequencing efficiencies, beads inside a simulated slice are counted only with certain probability \( p_{\text{s}} \).

Cosegregation matrix production. In GAM, bins found in the same slice are detected with a given efficiency and counted as cosegregating. Cosegregation frequencies are then arranged in a 2D \( N_{\text{c}} \times N_{\text{c}} \) matrix. Similarly, in our algorithm, we build a 2D \( N_{\text{c}} \times N_{\text{c}} \)-sized matrix. Reflecting the process of calling positive windows in experiments and GAM data, it is possible to detect a bin present in a simulated slice we consider the whole bin to be inside it. We therefore count all possible pairs of bins found in a slice and add that to the corresponding entry in the cosegregation matrix. We finally normalize the matrix by the number of slices employed, to generate cosegregation frequencies.

SLICE. We applied the statistical tool SLICE over in silico cosegregation data to determine its performance (Main Text and Supplementary Fig. 16). To do so, we considered the SLICE model statistics previously developed and implemented a version of the SLICE algorithm customized for the application on the in silico data of genomic loci.

The efficiency of in silico experiments. Each step of the in silico Hi-C algorithm has a specific efficiency: the probability \( p_{\text{b}} \) for the inclusion of a bead in a crosslinked cluster, the probability \( p_{\text{d}} \) for the survival of a bead in a cluster, the probability \( p_{\text{lig}} \) for ligation of a bead and the probability \( p_{\text{i}} \) for the detection of a ligated bead. Similarly, each step of in silico SPRITE and GAM has limited efficiency. By construction, in our algorithms the different steps are all independent and hence the overall in silico efficiency, \( \varepsilon \), that we discuss in Main Text is simply the product of single-step efficiencies. For example, for Hi-C, different values of \( p_{\text{b}}, p_{\text{d}}, p_{\text{lig}} \) and \( p_{\text{i}} \) corresponding to a given \( \varepsilon \) yield the same average output. The same holds for SPRITE and GAM.

In the in silico overall efficiency refers to the probability of the detection of a polymer bead in our Hi-C, SPRITE and GAM algorithms, and can be mapped onto the corresponding experimental overall efficiency. For Hi-C, since the average length of a digestion fragment equals the size of a polymer bead in our models, in silico overall efficiency can be assumed to be a good proxy for single-cell efficiency in a real Hi-C experiment, and analogously for SPRITE. In the case of GAM, experimental efficiency is computed via the SLICE algorithm and provides the probability for detection of a DNA window (that is, a bin) present in a physical slice. The link between in silico and experimental bead efficiency can be derived as follows. In a simulated GAM experiment, if \( k \) beads of a bin fall within a slice, the probability of detection of such a bin is \( 1 - (1 - \varepsilon)^k \). Averaging over all the permitted values of \( k \), the following approximate relation links bead and GAM efficiency:

\[
\varepsilon_{\text{GAM}} = 1 - (1 - \varepsilon) - \frac{(1 - \varepsilon)^n}{2n}
\]

where \( \varepsilon_{\text{GAM}} \) is the experimental GAM efficiency, \( \varepsilon \) is the bead efficiency of in silico GAM and \( n \) is the number of beads composing a bin in the polymer structures. For example, we found that the value \( \varepsilon = 0.5 \) employed in Fig. 5 corresponds to \( \varepsilon_{\text{GAM}} = 0.97 \), which is very close to the experimental efficiency of 0.94 estimated for the published GAM dataset at 40-kb resolution. We also tested, using numerical simulations, that the above approximate relation holds.

In the in silico experiments, the efficiency value used can strongly affect the quality of the contact maps (Supplementary Fig. 11), which worsens when efficiency is reduced. Such impact is generally compensated by the use of a large number of cells. Indeed, bulk contact maps, even at very low efficiency values, well correlate with those obtained at efficiency \( \varepsilon = 1 \). For instance, the Spearman correlation between bulk contact maps at efficiency \( \varepsilon = 1 \) and efficiency \( \varepsilon = 0.025 \) is \( r = 0.96 \) for Hi-C, \( r = 0.97 \) for SPRITE and \( r = 0.99 \) for GAM. Hence, for large numbers of cells, contact matrices are overall independent of efficiency, albeit that certain effects remain, linked to the specificity of the protocols of the experimental technologies.

Analyses of in silico single-cell data. In silico Hi-C, SPRITE and GAM single-cell contact matrices were obtained from a single simulated cell—that is, a pair of independent polymer structures (Fig. 4c). Single-cell distance matrices were computed as described in the section In silico distance matrices. For a given simulated cell, we calculated Pearson, Spearman and HiCRep correlation coefficients between contact and distance maps and between the latter and the bulk distance matrix. This was repeated for 250 in silico single cells and the average correlation was extracted (Supplementary Table 1). The distributions of Spearman correlation are reported in Fig. 4d for contact maps at
efficiency $\eta = 1$. In the calculation of rank correlations (Spearman) we considered the dissimilarity of the data structure of SPRITE and of Hi-C and GAM in silico single-cell data, as the latter have only integer entries. Such a difference was taken into account by normalization and discretization of the entries in silico SPRITE matrices in integers.

### Analysis of noise-to-signal ratio

The noise-to-signal ratio of the entry $a_{ij}$ of a contact matrix, where $i$ and $j$ are bin indices, is defined as the ratio between standard deviation $\sigma_{ij}$ and the mean $\mu_{ij}$ of that entry, across replicates. To estimate two such quantities, we first fixed the number of cells $N$ and the efficiency $\epsilon$. We then performed an in silico version of Hi-C, SPRITE and GAM 50 times each (when $N = 1$, we ran 100,000 times to overcome low sampling), and obtained a set of replicate contact matrices whereby we could extract mean and standard deviations for each entry, $a_{ij}$. This procedure was applied for $N$ and $\epsilon$ ranging from, respectively, 1–50,000 and 0.05–0.90. In particular, in Fig. 6 and Supplementary Figs. 8e–g, 9e–g, 10e–g and 14e–g we report the noise-to-signal ratio under different conditions.

First, we fixed the number of cells ($N = 50,000$) and efficiency ($\epsilon = 0.5$) and studied variation in the noise-to-signal ratio with genomic distance (Fig. 6a and Supplementary Figs. 8, 9, 10 and 14f). With analogous $\rho$ all entries of a single-cell contact map $M$ is, are linearly proportional to each other—that is, noise-to-signal ratio. Such a difference was used to account for the global efficiency of the simulated experiment ($\eta$). Thus, the entry first term is the probability of $a_{ij}$ belonging to the interval $[a, a + da]$. By definition, mean and variance for the entry $(ij)$ are

$$\mu_{ij}(\epsilon) = \int da f_\eta(ax) a \quad \sigma_{ij}^2(\epsilon) = \int da f_\eta(ax) \left( a - \mu_{ij}(\epsilon) \right)^2.$$

If an in silico experiment with $N$ cells at efficiency $\epsilon$ is performed, the output is the sum of $N$ independent single-cell matrices:

$$M = \sum_{n=1}^{N} A_n.$$
considerations hold approximately true. For a simulated single-cell experiment with efficiency \( e \), we define the probability for the entry \( a_{ij} \) to be 1 as \( p_{ij} (e) \) and have
\[
p_{ij} (e) = p_{ij} (1)
\]
and
\[
\sigma_{ij}^2 (e) = p_{ij} (1) - p_{ij} (e)
\]
where \( p_{ij} (1) \) and \( p_{ij} (e) \) indicate mean and variance, respectively. Hence, following the notation of the previous section, we can write \( \rho(e) \) in this particular case as
\[
\rho(e) = \frac{2}{b(b-1)} \sum_{i,j} p_{ij}(e) \left( 1 - p_{ij}(e) \right)
\]
To find the relation between \( L(e) \) and \( e \) we need to write \( \rho(e) \) as function of \( e \), that is, to exhibit the dependence of \( p_{ij}(e) \) on \( e \). In complete generality, it holds:
\[
p_{ij} (e) = P \{ a_{ij} = 1 \} \approx P \{ j \text{ detected} \} P \{ i, j \text{ in contact}\}
\]
where \(\sigma_{ij}^2 (e) (P \{ j \text{ detected} \})\) is the probability of bin \( i \) being detected and \( P \{ i, j \text{ in contact}\} \) is the conditioned probability that bins \( i \) and \( j \) are in contact given that they have been detected. We now assume that
\[
P \{ i \text{ detected} \} = P \{ j \text{ detected} \} = e. \quad \text{This is a simplifying assumption, because a bin is actually composed of multiple beads. We have:}
\]
\[
p_{ij} (e) = \varepsilon^2 c_{ij},
\]
where, for sake of simplicity, we have indicated the conditioned probability with \( c_{ij} \). So we can write \( \rho(e) \) as function of \( e \):
\[
\rho(e) = \frac{2}{b(b-1)} \sum_{i,j} \varepsilon^2 c_{ij} \left( 1 - \varepsilon^2 c_{ij} \right)
\]
For small \( e \), we can approximate the denominator as 1 and have
\[
\rho(e)_{e<1} = \varepsilon^2 \frac{2}{b(b-1)} \sum_{i,j} c_{ij} = \varepsilon^2 c.
\]
Thus, the Bernoulli approximation leads to a predicted exponent of \(-2\) between \( L(e) \) and \( e \) for small values of \( e \):
\[
L(e) = \delta^{-2} \varepsilon^{-1} (e) \approx \delta^{-2} e^{-2}.
\]
These predictions are satisfactorily confirmed in Fig. 5f and Supplementary Figs. 8–10 and 14d. The fact that Hi-C and SPRITE entries are not binary would affect the functional form of \( \rho(e) \), adding correction terms to the above (Supplementary Fig. 6b). We calculated the average Spearman correlation between models 60 single-cell experiments. For all of these, we extracted 40-kb Hi-C maps for the same 40-kb bin pairs. We calculated the mean Spearman correlation between in silico replicates are comparable to those found in experiments. Additionally, to verify that our estimates of \( \rho(e) \) are consistent with available experimental data, we used it also for pairs of GAM and SPRITE data to return a comprehensive view of similarity measures, beyond Spearman and Pearson correlations. We verified that the Hi-CRep correlations (scc) between experimental and in silico Hi-C, SPRITE and GAM contact maps are statistically significantly high. In the case of the S9t locus (Fig. 1b and Supplementary Table 1a), we computed the correlation between best-matching experimental structure and the single-molecule distance maps of the imaged 3D structures22 and of the SBS model configurations23 and computed the correlations between all pairs (exp–SBS contact maps, for Hi-C, SPRITE and GAM. Randomization is performed by bootstrapping contact frequencies at each genomic distance. We found that the measured scc correlations between model and experimental contact maps are higher than the 90th percentile distribution for all three technologies (Supplementary Fig. 3). For all other loci, analogous results were obtained.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**
Published Hi-C, SPRITE, GAM and microscopy data used for analysis are available at the referenced papers. The new GAM data from the F123 cell line are available on the 4D Nucleome data portal under accession no. 4DNFIBFSQ1EO.

**Code availability**
The codes used in our work are based on standard, publicly available software packages, as detailed in Methods. Molecular dynamics simulations use LAMMPS, v30July2016. Analyses and plots were produced with the Anaconda package v4.7.12. HiCrep correlations were computed with R v3.5.1. 3D structure visualizations were produced with POV-Ray, v3.7. The algorithms for simulation of Hi-C, SPRITE, GAM and SILICE in silico use standard routines, such as DRSCAN, and are described in full detail in Methods. The Hi-C, SPRITE and GAM algorithms are available at https://github.com/fmsella/In-silico-Hi-C_GAM_SPRITE.
References

61. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. 
   *J. Comput. Phys.* **117**, 1–19 (1995).

62. Kremer, K. & Grest, G. S. Dynamics of entangled linear polymer melts: 
   a molecular-dynamics simulation. *J. Chem. Phys.* **92**, 5057–5086 (1990).

63. Gribnau, J., Hochedlinger, K., Hata, K., Li, E. & Jaenisch, R. Asynchronous 
   replication timing of imprinted loci is independent of DNA methylation, but 
   consistent with differential subnuclear localization. *Genes Dev.* **17**, 759–773 
   (2003).

64. Rosa, A. & Everaers, R. Structure and dynamics of interphase chromosomes. 
   *PLoS Comput. Biol.* **4**, e1000153 (2008).

65. Bystricky, K., Heun, P., Gehlen, L., Langowski, J. & Gasser, S. M. Long-range 
   compaction and flexibility of interphase chromatin in budding yeast analyzed 
   by high-resolution imaging techniques. *Proc. Natl Acad. Sci. USA* **101**, 
   16495–16500 (2004).

66. Gavrilov, A., Razin, S. V. & Cavalli, G. In vivo formaldehyde cross-linking: it 
   is time for black box analysis. *Brief. Funct. Genomics* **14**, 163–165 (2015).

67. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for 
   discovering clusters in large spatial databases with noise. In Proc. 2nd 
   International Conference on Knowledge Discovery and Data Mining (eds 
   Simoudis, E. et al.) 226–231 (AAAI Press, 1996).

68. Tahara, M. et al. Cell diameter measurements obtained with a handheld cell 
   counter could be used as a surrogate marker of G2/M arrest and apoptosis in 
   colon cancer cell lines exposed to SN-38. *Biochem. Biophys. Res. Commun.* 
   **434**, 753–759 (2013).

69. Yang, F. et al. Dielectrophoretic separation of colorectal cancer cells. 
   *Biomicrofluidics* **4**, 13204 (2010).

Acknowledgements

We received no specific funding for this work. M.N. acknowledges support from 
CINECA ISCRA (nos. HP10CYFPS5 and HP10CRTY8P), Einstein BIH Fellowship 
Award (nos. EVF-BIH-2016 and 2019), Regione Campania SATIN Project 2018–2020 
and computer resources from INFIN, CINECA, ENEA CRESCO/ENEAGRID and Scope/ 
ReCAS/Ibisco at the University of Naples. A. Pombo thanks the Helmholtz Association 
(Germany) for support. M.N. and A. Pombo acknowledge support from the National 
Institutes of Health Common Fund 4D Nucleome Program (grant nos. 1U54DK167977- 
01 and 1U1HG011585-01) and the EU H2020 Marie Curie ITN (no. 812282). I.I.-A. 
acknowledges support from the Federation of European Biochemical Societies (FEBS 
Long-Term Fellowship). A.M.C. and S.B. acknowledge support from CINECA ISCRA 
(grant no. HP10CCZ4KN).

Author contributions

M.N. designed the project with input from A. Pombo and A. Prisco. L.F. and 
M.C. developed the modeling. L.F. and F.M. ran computer simulations and performed 
data analyses with help from M.C., A.M.C., S.B., A.E. and A.A. R.K., A.K. and I.I.-A. 
produced and normalized the GAM dataset. M.N., L.F., F.M., A. Pombo and A. Prisco 
 wrote the manuscript, with input from all the authors.

Competing interests

A. Pombo and M.N. hold a patent on ‘Genome Architecture Mapping’: Pombo, A., 
Edwards, P. A. W., Nicodemi, M., Beagrie, R. A. & Scialdone, A. Patent no. PCT/ 
EP2015/079413 (2015).

Additional information

Supplementary information The online version contains supplementary material 
available at https://doi.org/10.1038/s41592-021-01135-1.

Correspondence and requests for materials should be addressed to M.N.

Peer review information Nature Methods thanks Zhijun Duan and the other, 
anonymous, reviewer(s) for their contribution to the peer review of this work. Lei 
Tang was the primary editor on this article and managed its editorial process and peer 
review in collaboration with the rest of the editorial team. Reprints and permissions 
information is available at www.nature.com/reprints.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  
  Only common tests should be described solely by name; describe more complex techniques in the Methods section.

- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  
  Give P values as exact values whenever suitable.

- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

No software was used

Data analysis

The codes used in our work are based on standard, publicly available software packages, as detailed in the Methods section. Molecular Dynamics simulations use LAMMPS, version 30July2016. Analyses and plots were produced with the Anaconda package version 4.7.12. HiCRep correlations were computed with R, version 3.5.1. 3D structure visualisations were produced with POV-Ray, version 3.7. The algorithms for simulating Hi-C, SPRITE, GAM and SLICE in-silico also use standard routines, such as DBSCAN, and are described in full detail in the Methods section and in the original papers. All the codes are available upon request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Published Hi-C, SPRITE, GAM and microscopy data used are available at the referenced papers in the manuscript. New GAM F123 data are available on the 4D Nucleome data portal under accession number 4DNFIFBSQUEO.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences  ☐ Behavioural & social sciences  ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | The computational analyses were performed over previously published data of a 6Mb-long genomic locus at 40kb, a 7Mb-long genomic locus at 40kb, a 6Mb-long genomic locus at 10kb and a 2.5Mb-long genomic locus at 30kb. Then we also analysed a 210 beads long block-copolymer. |
|-------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions | No data were excluded for our analysis. |
| Replication | To verify the reproducibility of our findings, we applied our computational approach to four different genomic loci and on a block-copolymer and found analogous results. |
| Randomization | To test the significance of the HiCRep correlations (scc) between our model matrices and the experimental ones, we benchmarked against a control distribution of scc derived from 100 randomized matrices. We verified that the scc between our in-silico matrices and the corresponding experimental matrices are above the 90th percentile of the random control. To test that our ensemble of model polymer structures is a bona-fide representation of real configurations of chromatin in nuclei, in the case of the human locus, we performed an RMSD analysis between available imaging data and our model structures and compared with the results from a random control, i.e. between the imaging data and self-avoiding-walk polymers. We found that the imaging-model RMSD distribution is statistically different than the control. |
| Blinding | This is not relevant for our study as we do not have clinical data from different patients. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☒  | Antibodies           |
| ☒  | Eukaryotic cell lines |
| ☒  | Palaeontology and archaeology |
| ☒  | Animals and other organisms |
| ☒  | Human research participants |
| ☒  | Clinical data        |
| ☒  | Dual use research of concern |

### Eukaryotic cell lines

**Policy information about cell lines.**

**Cell line source(s)**

Mouse F123 ESCs (a male, hybrid cell line, derived from S129/jae and Cast) (Gribnau et al., 2003) were kindly provided by Bing Ren, University of California San Diego, San Diego, CA, USA

**Authentication**

Whole genome sequencing confirmed positions of SNPs, immunofluorescence for cell stage markers confirmed pluripotent state of mESC line

**Mycoplasma contamination**

The used cell line was tested negative for mycoplasma

**Commonly misidentified lines**

(See iCLAC register)

No commonly misidentified cells lines were used in this study

### Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☒  | ChIP-seq             |
| ☐  | Flow cytometry       |
| ☒  | MRI-based neuroimaging |