nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-021-00998-1

Single-cell measurement of higher-order 3D genome organization with scSPRITE

Mary V. Arrastia[1,3], Joanna W. Jachowicz[2,3], Noah Ollikainen[2], Matthew S. Curtis[1], Charlotte Lai[2], Sofia A. Quinodoz[2], David A. Selck[1], Rustem F. Ismagilov[1,2] ✉ and Mitchell Guttman[2] ✉

Although three-dimensional (3D) genome organization is central to many aspects of nuclear function, it has been difficult to measure at the single-cell level. To address this, we developed 'single-cell split-pool recognition of interactions by tag extension' (scSPRITE). scSPRITE uses split-and-pool barcoding to tag DNA fragments in the same nucleus and their 3D spatial arrangement. Because scSPRITE measures multiway DNA contacts, it generates higher-resolution maps within an individual cell than can be achieved by proximity ligation. We applied scSPRITE to thousands of mouse embryonic stem cells and detected known genome structures, including chromosome territories, active and inactive compartments, and topologically associating domains (TADs) as well as long-range inter-chromosomal structures organized around various nuclear bodies. We observe that these structures exhibit different levels of heterogeneity across the population, with TADs representing dynamic units of genome organization across cells. We expect that scSPRITE will be a critical tool for studying genome structure within heterogeneous populations.

In eukaryotes, linear DNA is packaged in a 3D arrangement in the nucleus. This includes organization of DNA regions from the same chromosome (chromosome territories)[1], which are further subdivided into megabase-sized, self-associating TADs[2,3] based on gene activity (active/inactive or A/B compartments)[1], and local interactions between regulatory elements (enhancer–promoter loops)[4–6]. Additionally, DNA regions from multiple chromosomes are organized around nuclear bodies that form higher-order structural units[7,8].

Genome organization in a single nucleus affects various nuclear functions, including DNA replication[9], transcription[5,10] and RNA processing[11,12]. Indeed, genome structure is known to dynamically change between cell types and in individual cells across time to reflect differences in biological state[5,13,14]. For example, during the cell cycle, DNA structure undergoes dramatic rearrangement from open chromatin during interphase to highly condensed metaphase chromosomes[15–17]. Similarly, gene expression levels are heterogeneous among populations of cells[18,19], suggesting that there might be differences in enhancer–promoter contacts present in individual cells in the population.

Currently, most methods used to study nuclear organization measure ensemble structures across millions of cells and can obscure critical information about the genome organization of any given cell. For example, measuring cells across the cell cycle and averaging their DNA contacts would mask cell cycle-dependent dynamics. Additionally, several studies showed that observation of genome structures such as TADs[13–17,20] in single cells do not always match structures predicted from ensemble measurements[1,3,21]. Accordingly, genome organization observed in bulk assays might not accurately reflect specific structures that exist within biological populations.

The two main techniques for measuring genome architecture of single cells are microscopy and single-cell Hi-C (scHi-C). Microscopy provides the capability to study a broad range of genomic interactions in single cells but is generally limited to measurements of a small number of loci simultaneously[13,14,20] and does not provide a genome-wide view. In contrast, scHi-C provides a genome-wide view of nuclear structure in single cells, but it requires specialized equipment (for example, robotics), generates data for low cell numbers and is limited to low-resolution structures (~10-Mb resolution per cell)[15,16,22]. Additionally, because scHi-C relies on proximity ligation to measure interactions, it has limited ability to capture long-range and higher-order interactions, such as those organized around nuclear bodies[7,23].

To address these technological gaps, we developed scSPRITE to provide comprehensive, high-resolution, genome-wide maps of DNA structure from thousands of single cells. scSPRITE measures both inter- and intra-chromosomal interactions and dramatically increases the number of detected DNA contacts per cell relative to existing methods. To demonstrate its utility, we measured 3D genome structures in 1,000 individual mouse embryonic stem cell (mESC) nuclei and observed chromosome territories, A/B compartments and TADs in hundreds of single nuclei. We identified higher-order structures in hundreds of single cells, including inter-chromosomal contacts around centromeres, nucleoli and nuclear speckles. Notably, we identified cell-to-cell heterogeneity in mESC genome structure at different levels of resolution, including at promoter–enhancer contacts of the key pluripotency gene, *Nanog*. Together, these observations demonstrate that scSPRITE accurately measures genome structure and provides insights into genome organization. We expect that this approach will enable future studies examining the relationship between genome organization and nuclear function in individual cells.

Results

scSPRITE maps 3D structure in thousands of individual cells. To understand 3D genome organization in individual cells, we extended our previously described SPRITE protocol[7] to enable

[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA. [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [3]These authors contributed equally: Mary V. Arrastia, Joanna W. Jachowicz. ✉e-mail: rustem.admin@caltech.edu; mguttman@caltech.edu

single-cell measurements. scSPRITE works as follows: We dissociate cells into a single-cell suspension, crosslink DNA and protein complexes in situ, isolate and permeabilize nuclei, digest DNA using a restriction enzyme and perform two sets of split-and-pool barcoding to (1) tag DNA fragments contained in the same nucleus and (2) tag the 3D spatial arrangement of these fragments (Fig. 1a).

To map all DNA fragments originating from one nucleus, we performed split-and-pool barcoding to generate a unique cell-specific barcode (cell barcode) for all DNA molecules contained in a single nucleus. Briefly, we distributed permeabilized nuclei across a 96-well plate (~200,000 nuclei), where each well contained a unique DNA barcode tag, and performed ligation such that all DNA molecules in the same nucleus were labelled with the same tag. We then pooled nuclei and repeated the split-and-pool process twice more to ensure that the number of barcode combinations ($96^3 = 884{,}736$) exceeded the cell number (Methods). Because single nuclei can form aggregates in suspension, we filtered nuclei to remove potential clumps before proceeding to the next step (Extended Data Fig. 1a).

To verify that this approach accurately tags DNA contained in a single nucleus, we tested this first set of split-and-pool barcoding in permeabilized nuclei in a mixed population of human (HEK293T) and mouse (mESC) cells. After split-pool barcoding and sequencing, we clustered reads into groups based on their cell barcodes and computed the percentage of reads that aligned exclusively to the mouse or human genome (Methods). We found that only 3.4% of cells contained reads from both species (Fig. 1b and Extended Data Fig. 1b), indicating that most cell barcodes represent single cells. Because we cannot identify collisions that lead to mixing in the same species, we extrapolate a total collision rate (~10%) from the detected collisions.

Having developed an approach to accurately tag DNA in a single nucleus, we next sought to map these DNA fragments relative to each other in 3D space. To do this, we withdrew a small fraction of the single cell-tagged nuclei (~1,500 nuclei) and sonicated them to generate spatial clusters of chromatin. We then performed three additional rounds of split-and-pool barcoding, such that all DNA fragments contained in a spatial cluster obtained the same barcode combinations, whereas molecules in distinct spatial clusters obtained different combinations. After sequencing, we identified DNA molecules within the same spatial complex by matching all six barcode sequences and all complexes arising from the same nucleus by matching the first three barcodes (Fig. 1a, Extended Data Fig. 1c and Methods).

To validate the method, we applied scSPRITE to mESCs because their genome structure has been extensively studied[3,15], and they display known functional heterogeneity[17,24–26]. We sequenced ~1,500 single cells and analytically excluded cell barcodes that were likely to represent cell aggregates using the detected collision rates measured from the previously described mixing experiment (Fig. 1c and Extended Data Fig. 1b). To focus on the most informative single cells, we restricted our analysis to the 1,000 cells containing the highest number of spatial clusters per cell (Methods).

To confirm that spatial barcoding in scSPRITE accurately measures known genome structures, we merged individual cell barcodes from scSPRITE (referred to as ensemble scSPRITE) and compared heat maps to those previously generated by bulk SPRITE in mESCs[7] (Fig. 1d). We found that these maps are highly similar across all levels of resolution (Pearson correlation $r = 0.92$, 1 Mb genome-wide; $r = 0.97$, 200 kb on chr2; $r = 0.95$, 40 kb across chr6: 48–54 Mb).

Together, our results demonstrate that scSPRITE tags single cells with minimal collisions and accurately measures 3D organization at different levels of resolution. Although we analyzed 1,500 single cells in this experiment, the number of cells analyzed by scSPRITE can be adjusted by modifying the number of rounds of split-and-pool barcoding such that the number of barcode combinations exceeds the number of single cells (>100-fold excess; Methods).

scSPRITE measures multiway interactions in single cells. Because each individual cell contains a single genome, and contacts detected in multiple cells cannot be pooled together (as they are in bulk measurements), single-cell genome structure methods need to maximize the number of contacts detected in each cell. This is the main challenge and limitation for all single-cell genomic methods.

Currently, existing single-cell genome structure methods (for example, scHi-C) use proximity ligation and are, therefore, limited to measuring pairwise DNA contacts[16,22,27–30]. Although these measurements are averaged across multiple cells, this is not possible in a single cell because a specific DNA region can be measured only once per allele. Accordingly, even with perfect efficiency, pairwise methods would be unable to measure all possible contacts present in a given structure (Fig. 1e and Extended Data Fig. 1e). In contrast, SPRITE captures multiway contacts among DNA molecules, which dramatically increases the structural resolution that can be obtained for an individual cell. This is because the maximal number of interactions that can be captured increases quadratically with the size of a complex[23] (Extended Data Fig. 1d). For example, if a crosslinked complex contains four DNA fragments, the maximum number of contacts that can be observed by pairwise methods is two, whereas the maximal number of pairwise contacts that can be identified with multiway contacts is six (Extended Data Fig. 1e).

Indeed, we observe an increase in the number of pairwise contacts detected for each cell using scSPRITE (average of 34,992,080 per cell) compared to scHi-C[16] (average of 375,470 per cell), even though the number of sequencing reads per cell is ~10-fold lower for scSPRITE (average of 83,318 per cell) than for scHi-C (average of 751,172 per cell) (Fig. 1f and Supplementary Note 1). We observe uniform coverage across all 1-Mb bins in virtually all cells and across all 100-kb bins in more than 80% of cells (Fig. 1g) with almost no bias toward any chromosome (with the exception of chromosome 8 due to a trisomy in our cell line; Methods) (Fig. 1h). Notably, we observe low variability in genomic coverage across the analyzed cells (median absolute deviation (MAD) = 14, MAD range: 0–49, median = 35; Extended Data Fig. 1f,g).

scSPRITE detects chromosome territories and compartments. To determine which DNA structures can be observed in single cells, we generated DNA contact maps from each of the 1,000 individual cells. For every structure identified in the ensemble data, we computed a normalized detection score that reflects how well each single cell contact map resembles this structure compared to a randomized contact map. Briefly, for each structure, we calculated an observed detection score, which defines whether each pair of genomic bins in a structure were in contact. A cell that contains all possible pairwise contacts in a given structure would have a detection score of 1, whereas a cell containing none of the expected pairwise contacts would have a detection score of −1. We normalized these observed scores to a distribution of scores generated by randomly permuting the locations of each structure (Methods and Supplementary Note 4).

We focused on genomic structures that were previously reported to occur in single cells—chromosome territories and A/B compartments[1]. Chromosome territories are structures containing high frequencies of intra-chromosomal interactions with minimal inter-chromosomal interactions (Fig. 2a). First, we looked at the contacts between chromosome 1 (chr1) and chr2 and detected clear separation of contacts into chromosome territories in both the ensemble data (Fig. 2a) and in more than 75% of single cells (score > 0; Fig. 2b and Extended Data Fig. 2a). Next, we quantified detection scores for every pair of chromosomes in every cell (Extended Data Fig. 2b and Supplementary Table 1). Although some chromosomes show stronger self-interactions than others, all chromosomes organize into territories (average score = 0.08, s.d. = 0.06; Fig. 2c

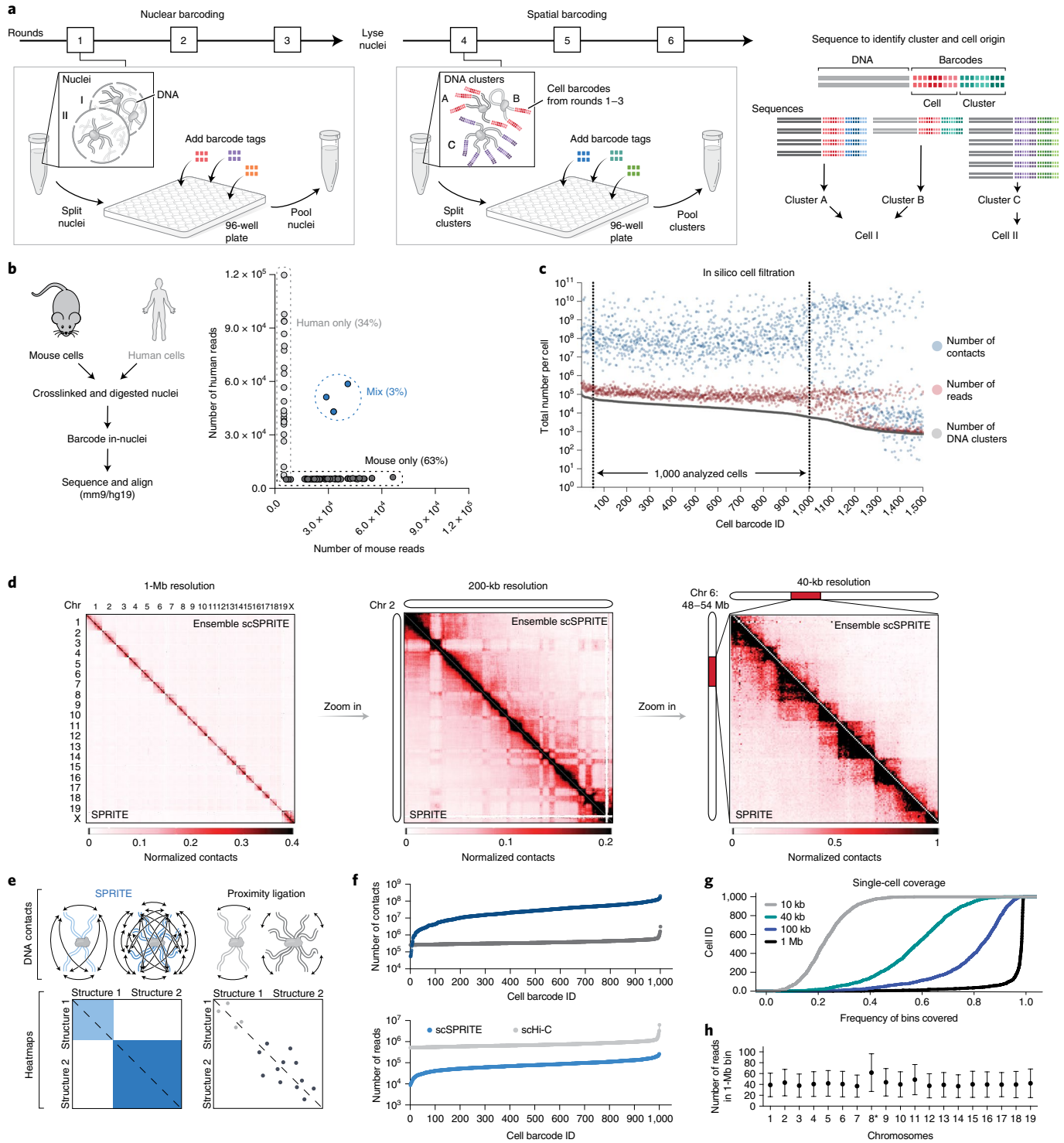


Fig. 1 | scSPRITE—a single-cell method to map DNA structure genome-wide. a, Schematic of scSPRITE protocol. **b**, Validation of in-nuclei barcoding step on mixed cell population (human–mouse cells); the number of reads for each identified cell barcode ID is plotted. Threshold of >95% single-species reads was applied to identify mouse- or human-only cells; cell barcodes >1,000 reads are plotted. **c**, Number of contacts (blue), reads (red) and DNA clusters (gray) plotted for the 1,500 cells. Dashed lines represent filtration steps: left of the dashed lines—cell aggregates estimated based on detected collision rate from Fig. 1b; right of the dashed lines—cells with low number of reads/contacts **d**, Comparison of merged scSPRITE (upper diagonal, 'ensemble scSPRITE') and bulk SPRITE[7] (lower diagonal). Chromosome territories across all chromosomes at 1-Mb resolution (left); A/B compartments on chromosome 2 at 200-kb resolution (middle); TADs within an 18-Mb region of chromosome 6 at 40-kb resolution (right). **e**, Schematic illustration of multiway interactions (SPRITE-derived methods) and pairwise interactions (proximity ligation methods) and examples of heat maps. **f**, Number of contacts (top) and number of reads (bottom) obtained from scSPRITE (blue) and scHi-C[16] (gray). **g**, Genomic coverage per 1-Mb, 100-kb, 40-kb and 10-kb bins in individual 1,000 cells. **h**, Average number of reads per single cell in 1-Mb bins of each chromosome ($n = 1{,}000$ cells). Average (dots) and s.d. (bars) are shown; asterisk marks chromosome with detected trisomy.

and Extended Data Fig. 2c; see Methods for chr8). We observe that 95% of cells contain well-defined territories (Fig. 2d,e), and only a small fraction of cells (<50 cells) do not contain observable chromosome territories (Fig. 2d and Extended Data Fig. 2d) and might reflect cell states containing distinct organization, such as mitotic chromosomes.

Genomes are further divided into A/B compartments, which are intra-chromosomal structures defined by open (A) or closed (B) chromatin states[1] (Fig. 2f). To measure A/B compartment patterns in single cells, we first focused on a region on chr2 that has a well-defined B-A-B compartment switch observed in the ensemble scSPRITE data (Fig. 2f). We calculated the detection score for that region in individual cells and observed segregation of DNA into A/B compartments in more than 65% of single cells (score > 0; Fig. 2g and Extended Data Fig. 2e). Next, using our ensemble data, we defined all regions that correspond to a compartment switch (B-A-B or A-B-A) genome-wide (224 regions; Supplementary Table 2) and quantified their detection scores for each cell (Extended Data Fig. 2f). We observed that individual regions are more variable in single cells than chromosome territories (average score = 0.03, s.d. = 0.06; Fig. 2h and Extended Data Fig. 2f) but are still present in ~95% of cells (Fig. 2i). We looked more closely into three regions with different average detection scores (Region 1, score = 0.12, s.d. = 0.16; Region 2, score = −0.01, s.d. = 0.14; Region 3, score = −0.10, s.d. = 0.11) and observed that the variability in the A/B compartment structure in single cells is indeed higher for the regions with lower detection scores (Fig. 2j and Extended Data Fig. 2g) (i.e., Region 3>Region 2>Region 1). This suggests that the detection score metric that we developed is useful to identify cells and regions of variable structures. We observe a small detection bias toward active regions (A compartments, 45% of observed reads versus 39% of expected reads) (Extended Data Fig. 2h).

Together, our results demonstrate that scSPRITE can detect known genomic interactions, such as chromosome territories and A/B compartments in single cells, and can be used to measure structural variability between individual cells.

Inter-chromosomal hubs are organized around nuclear bodies. The nucleus is further organized around various nuclear bodies that form higher-order inter-chromosomal contacts[7,8]. These contacts have not been previously explored in single cells at the genome-wide scale because existing single-cell proximity ligation methods are limited in their ability to detect inter-chromosomal contacts[16,22,23,29]. scSPRITE measures, on average, an almost tenfold increase in the proportion of inter-chromosomal contacts per cell than scHi-C (54% and 6%, respectively) (Fig. 3a), which makes it a suitable method to study higher-order organization in the nucleus. We focused on three types of known inter-chromosomal structures: inactive regions associated with nucleoli, active chromatin around nuclear speckles, and centromeric and peri-centric regions organized into chromocenters.

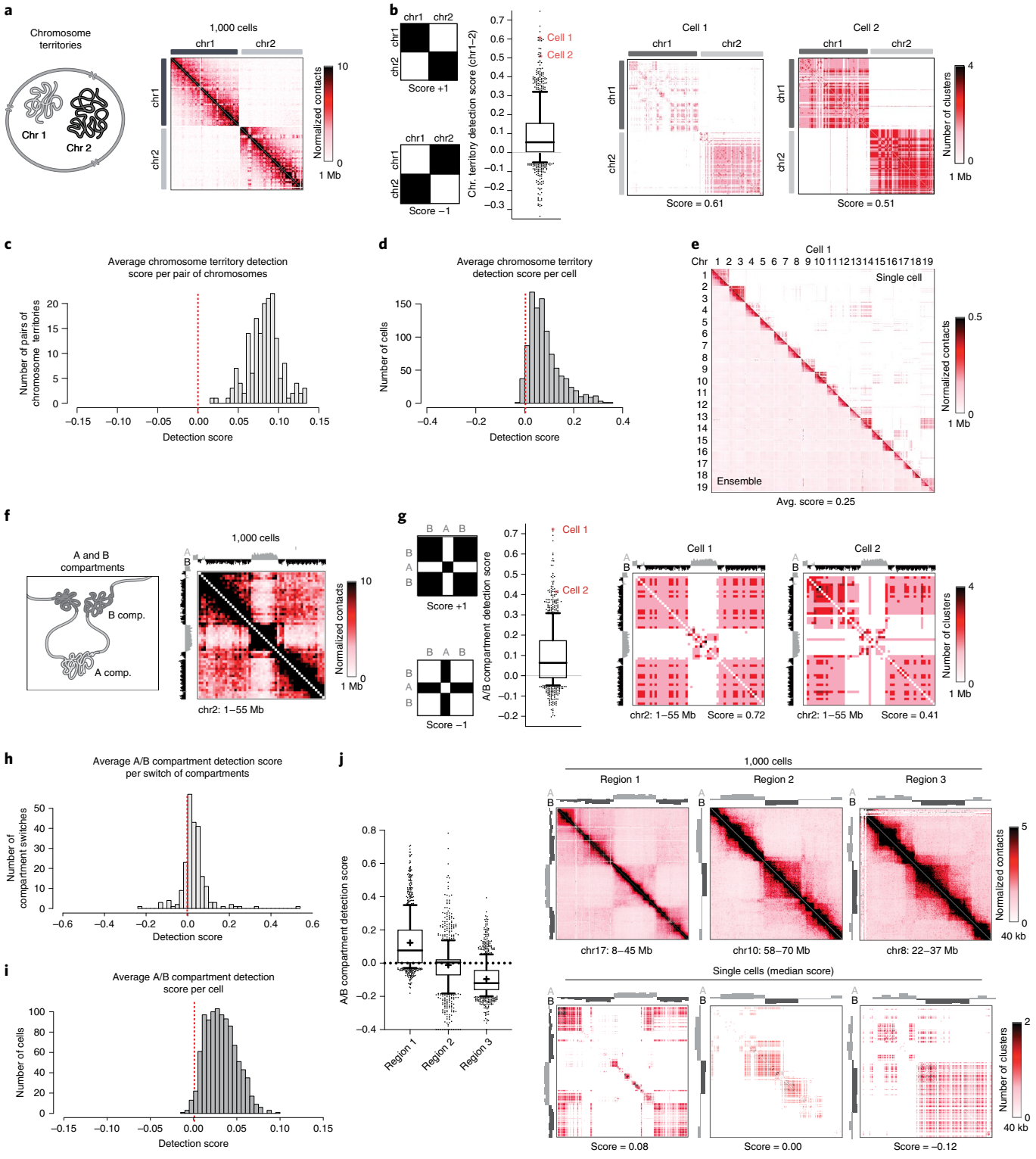
Inactive DNA hubs are known to organize around the nucleolus[7], a nuclear body that is formed around transcription of ribosomal DNA (rDNA) regions[12]. In mESCs, regions on chr12, chr15, chr16, chr18 and chr19 contain rDNA clusters that form nucleolar organizing regions (NORs). We first explored contacts between two NOR-containing regions on two pairs of chromosomes (chr18/chr19 and chr12/chr19) that were previously reported to form strong interactions in mESCs[7]. We observed similar interaction patterns between these regions in the ensemble SPRITE data and in individual cells (score > 0 in 54% and 61% of cells, respectively) (Fig. 3b,c and Extended Data Fig. 3a,b). We compared the frequencies of contacts detected by scSPRITE (specifically how often these two regions are in the same cluster) to the frequencies of their co-occurrence at the same nucleolus measured by microscopy (where the nucleolus is visualized by nucleolin immunostaining and DNA regions are visualized by DNA fluorescence in situ hybridization (DNA FISH))[7]. We focused our analysis specifically on 1-Mb regions targeted by DNA FISH probes (three NOR-containing chromosome pairs and two control chromosome pairs) and observed a strong correlation between these datasets (R^2 = 0.88; Extended Data Fig. 3c), indicating that single-cell measurements generated by scSPRITE are similar to those observed by microscopy. Similarly, we observed a strong correlation between scSPRITE and SPRITE data for these NOR regions (R^2 = 0.88; Extended Data Fig. 3d). To look at genome-wide interactions of NORs, we quantified the percentage of single cells that contain each nucleolar contact (Extended Data Fig. 3e) and observed that, on average, 38% of cells contained each nucleolar pair (Extended Data Fig. 3e). The most frequent contacts are formed between NORs on chr18 (3–10 Mb) and chr19 (25–28 Mb or 29–37 Mb), which are both observed in more than 50% of cells, and the least frequent contacts are observed between chr15 (67–71 Mb) and chr18 (57–60 Mb), which are observed in less than 20% of cells (Extended Data Fig. 3e). In all cases, we observed that NORs interact more frequently than random non-NOR-containing regions (Fig. 3d).

Next, we looked at active nuclear hubs organized around nuclear speckles—structures enriched in pre-mRNA splicing factors[11,31]. First, we focused on the previously reported inter-chromosomal interactions formed by precise regions of mouse chr2/chr4 and chr2/chr5 (ref. [7]) and observed these contacts in 53% and 38% of cells, respectively (score > 0; Fig. 3e,f and Extended Data Fig. 3f,g). Next, we quantified the percentage of single cells that contain each pair of interacting speckle regions (Extended Data Fig. 3h).

Fig. 2 | scSPRITE accurately measures single-cell DNA interactions at different resolutions by capturing multiway interactions. a, Illustration of chromosome territories for chr1 and chr 2 (left) and ensemble scSPRITE heat map (right) of the same structures; downweighted contact map at 1-Mb resolution. **b**, Chromosome territory normalized detection scores for 1,000 individual cells between chr1 and chr2. Left: representation of structures with max. score (+1) and min. score (−1). Center: box plot where whiskers represent the 10th and 90th percentiles; box limits represent the 25th and 75th percentiles; the black line represents the median; red dots represent single-cell examples (n = 1,000 cells). Right: single-cell examples of chr1 and chr2 territories, plotted as number of DNA clusters at 1-Mb resolution. **c**, Normalized detection scores across all 1,000 cells per each pair of chromosome territories detected in ensemble scSPRITE data; score = 0 (red line). **d**, Normalized detection scores across all pairs of chromosome territories detected in ensemble scSPRITE data per single cell; score = 0 (red line). **e**, Chromosome territories (chr1–19) in ensemble scSPRITE (left) and in a single cell (right, detection score = 0.25). **f**, Illustration of A/B compartment in chr2:0–55 Mb (left) and ensemble scSPRITE heat map (right); downweighted contact map at 1-Mb resolution. **g**, A/B compartments detection scores for 1,000 individual cells. Left: representation of structures with max score (+1) and min. score (−1). Center: box plot where whiskers represent the 10th and 90th percentiles; box limits represent the 25th and 75th percentiles; the black line represents the median; red dots represent single-cell examples (n = 1,000 cells). Right: single-cell examples of A/B compartments in chr2:0–55 Mb, plotted as number of DNA clusters at 1-Mb resolution. **h**, Normalized detection scores across all 1,000 cells per each compartment switch; score = 0 (red line). **i**, Compartment detection scores across all compartments per single cell; score = 0 (red line). **j**, Examples of three different regions containing a high (Region 1), medium (Region 2) and low (Region 3) median compartment switch score. For each region's box plot: whiskers represent the 10th and 90th percentiles; box limits represent the 25th and 75th percentiles; the black line represents the median; red dots represent single-cell examples (n = 1,000 cells). Heat maps for each region are shown in both ensemble scSPRITE (above) and single cell (below).

We detected speckle interactions in an average of 34% of cells (Extended Data Fig. 3h), with interactions between regions on chr4 (128–142 Mb) and chr5 (112–126 Mb) observed in more than 50% of cells and between chromosome 2 (117–181 Mb) and chromosome 13 (55–58 Mb) observed in less than 10% of cells (Extended Data Fig. 3h). When we calculated the frequency of contacts per 1-Mb bin of every interacting speckle region, we observed that most speckle regions interact more frequently than random regions but less frequently that NORs (Fig. 3d).

Finally, we explored centromeric and peri-centromeric heterochromatin (PCH) regions. Centromeres and peri-centromeres are long stretches of repetitive DNA essential for chromosome stability and segregation[32] and have been shown to come into close proximity to form inter-chromosomal structures called chromocenters[32] (Fig. 3g). Because PCH regions are not mapped in the genome, we focused our analysis on the first 10 Mb of each chromosome. First, we made single-cell contact maps and calculated detection scores for two pairs of PCH regions (chr1/chr11 and chr4/chr11)



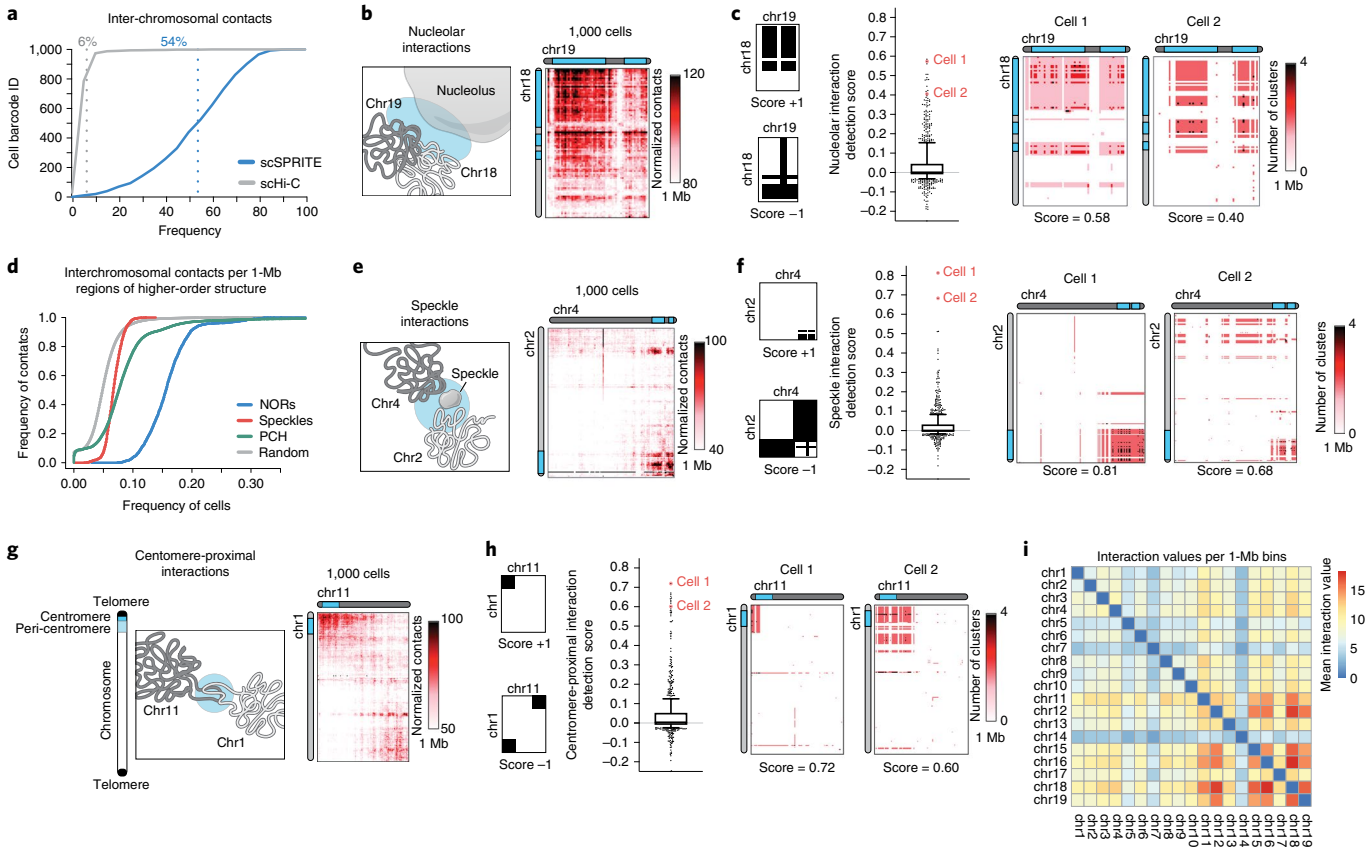


Fig. 3 | scSPRITE identifies inter-chromosomal structures genome-wide in hundreds of single mESCs. a, Quantification of inter-chromosomal contacts from the top 1,000 cells by scHi-C[16] (gray) and scSPRITE (blue). The dashed lines represent the mean percentage of inter-chromosomal contacts. **b**, Nucleolar interaction between chr18 and chr19: illustration (left) and ensemble scSPRITE heat map (right); contact map at 1-Mb resolution. **c**, Nucleolar interaction detection scores for 1,000 cells (middle). Box plot where whiskers represent the 10th and 90th percentiles; box limits represent the 25th and 75th percentiles; the black line represents the median; red dots represent single-cell examples ($n = 1{,}000$ cells). Representation of structures with max score (+1) and min. score (−1) (left). Single-cell examples (right), plotted as number of DNA clusters at 1-Mb resolution. **d**, Frequency of NOR (blue), speckle (red) and PCH (green) higher-order interactions in comparison to randomly shuffled regions of the same size (gray) in 1,000 individual cells. **e**, Speckle interaction between chr2 and chr4: illustration (left) and ensemble scSPRITE heat map (right); contact map at 1-Mb resolution. **f**, Speckle interaction detection scores for 1,000 individual cells (middle). Box plot where whiskers represent the 10th and 90th percentiles; box limits represent the 25th and 75th percentiles; the black line represents the median; red dots represent single-cell examples ($n = 1{,}000$ cells). Representation of structures with max score (+1) and min. score (−1) (left). Single-cell examples (right), plotted as number of DNA clusters at 1-Mb resolution. **g**, PCH interactions between chr1 and chr11: illustrations (left) and ensemble scSPRITE heat map (right); contact map at 1-Mb resolution. **h**, PCH region detection scores for 1,000 individual cells (middle). Box plot where whiskers represent the 10th and 90th percentiles; box limits represent the 25th and 75th percentiles; the black line represents the median; red dots represent single-cell examples ($n = 1{,}000$ cells). Representation of structures with max score (+1) and min. score (−1) (left). Single-cell examples (right), plotted as number of DNA clusters at 1-Mb resolution. **i**, Mean interaction value of inter-chromosomal PCH contacts (normalized to number of reads per region) for each pair of chromosomes. NOR-containing chromosomes are shown in bold.

(Fig. 3g,h and Extended Data Fig. 3i,j); we detected formation of these inter-chromosomal interactions in 54% and 80% of cells, respectively (score > 0; Fig. 3h and Extended Data Fig. 3i,j). Next, we looked at genome-wide interactions of PCH regions and quantified the percentage of single cells that contain each PCH contact (Extended Data Fig. 3k). We observed that, on average, 49% of cells contained two different PCH-containing regions in close proximity. Notably, the PCH region of chr11 forms pairs with other PCH regions most frequently (80% of cells), and PCH region of chr14 interacts least frequently with other PCH regions (30% of cells) (Extended Data Fig. 3k). More generally, when we calculated frequency of PCH interactions per each 1-Mb region of PCH, we observed that these regions form pairs more frequently than random regions of the same size (Fig. 3d). We note that, after size normalizations (Methods), chromosomes that contained NORs displayed a higher contact frequency between their centromeric regions (Fig. 3i), consistent with previous observations by microscopy[33,34].

The results of these analyses demonstrate that scSPRITE can capture various higher-order contacts reflecting inter-chromosomal interactions across multiple cells and involving structures of different sizes and transcriptional output (active versus inactive hubs). We note that centromere-proximal and nucleolar contacts were not detectable even in the ensemble scHi-C data[16] (Extended Data Fig. 3l). Although the ensemble scHi-C[16] was able to identify speckle interactions, the single-cell interaction maps lacked information on these structures (Extended Data Fig. 3l).

TADs are heterogeneous across individual cells. TADs are intra-chromosomal structures in which contiguous regions of the genome have been shown to interact more with themselves than with surrounding regions[2,3,35]. However, these observations are based mainly on bulk measurements, and whether TADs exist in single cells has been debated[13,15,16,20]. Specifically, it is unclear whether the inability to observe TADs in single cells reflects technical limitation

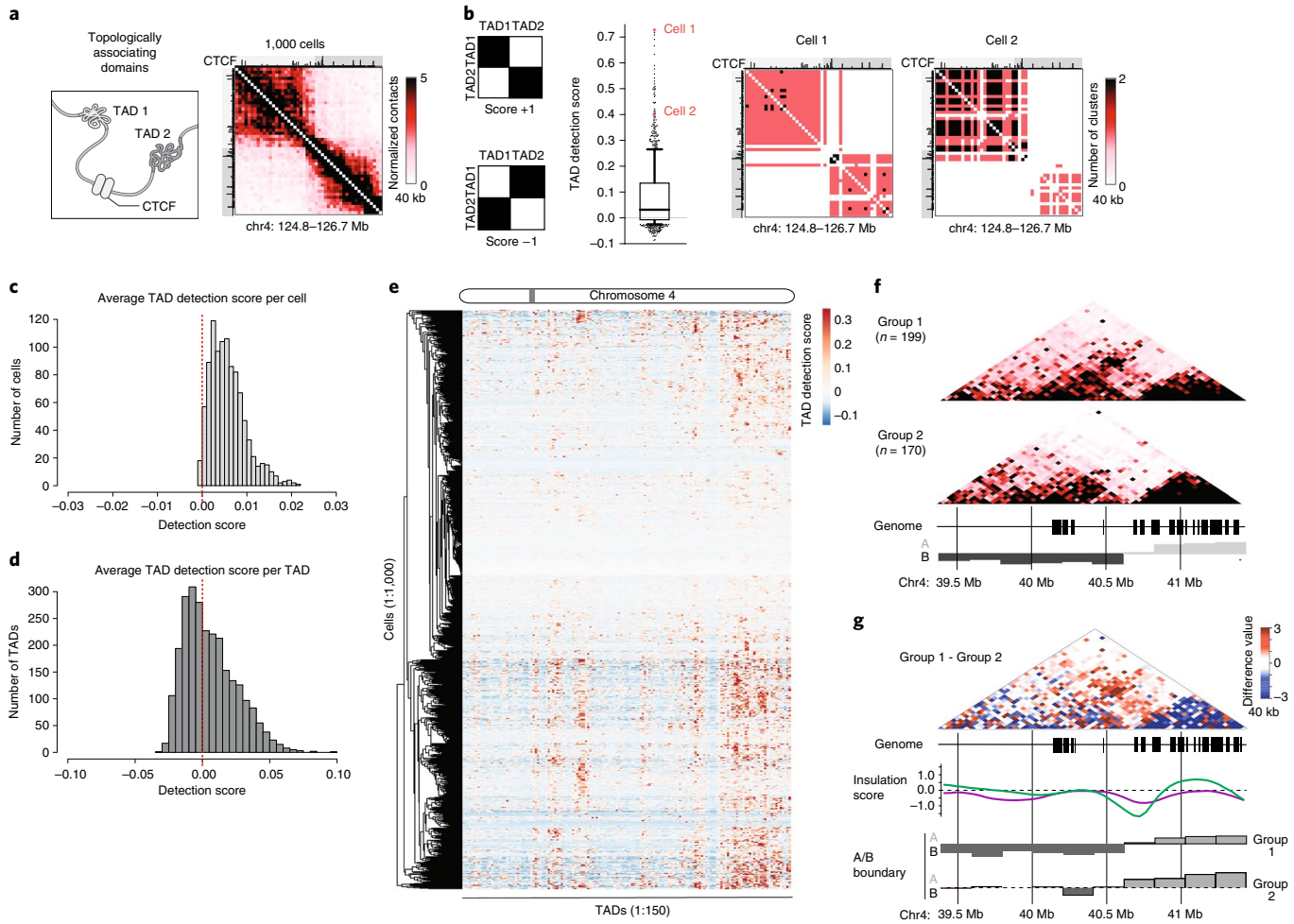


Fig. 4 | TADs are heterogeneous units present in the genomes of individual mESCs. a, TAD structure between 124.8 Mb and 126.7 Mb of chr4: illustration (left) and scSPRITE heat map (right); pairwise contact map at 40-kb resolution. **b**, TAD detection scores for 1,000 cells (middle). Box plot where whiskers represent the 10th and 90th percentiles; box limits represent the 25th and 75th percentiles; the black line represents the median; red dots represent single-cell examples ($n = 1{,}000$ cells). Representation of structures with max score (+1) and min. score (−1). Single-cell examples (right), plotted as number of DNA clusters at 40-kb resolution. **c**, Normalized detection scores across all 1,000 cells per each TAD detected in ensemble scSPRITE data; red line marks score = 0. **d**, TAD detection scores across all TADs detected in ensemble scSPRITE data per single cell; red line marks score = 0. **e**, TAD detection scores across 1,000 cells (clustered based on score similarity pattern): columns represent the strength of TAD detection scores for all TADs detected across chr4 in ensemble scSPRITE; gray bar indicates the variable region described in Extended Data Fig. 4c. **f**, Ensemble heat maps across the 39.4–41.4-Mb region of chr4 representing cells containing (Group 1, top) or lacking (Group 2, bottom) the contact emerging over the boundary of A/B compartment. **g**, Difference contact map across 39.4–41.4 Mb of chr4 made by subtracting the normalized contacts in Group 2 from Group 1 (Fig. 4f). Insulation scores for cells in Group 1 (purple) and Group 2 (green) are plotted.

of current single-cell methods (for example, low-resolution structures) or if these DNA structures are not present in individual genomes. Because scSPRITE generates higher-resolution structures in individual cells, we asked whether it can detect TADs in single cells.

We first defined all TADs present in mESCs using the ensemble scSPRITE data (Supplementary Table 3), which are similar to TADs defined from Hi-C data[3] (Pearson correlation $r = 0.70$; Extended Data Fig. 4a,b). We used these genomic coordinates to score each of these TADs in every single cell. First, we focused our analysis on a region of chromosome 4 (124.8–126.7 Mb) where we observed strong evidence for TADs in the ensemble scSPRITE dataset (Fig. 4a). Using the genomic locations defined from the ensemble data, we detected TAD-like structures in more than 75% of single cells (score > 0; Fig. 4b and Extended Data Fig. 4c), suggesting that most individual cells contain this specific TAD structure with the same boundaries.

To explore the heterogeneity of TAD structures in single cells, we performed two analyses. First, we looked at the average representation of all TADs in each cell by averaging the TAD detection score for each region (identified in the ensemble dataset) in each individual cell; we found that most cells contain TADs (95% of cells with score > 0; Fig. 4c). Second, we explored whether individual TADs are more or less variable across individual cells by averaging the TAD detection score for individual TADs across cells. We found that most TADs are highly variable between cells (65% of cells with score < 0; Fig. 4d) and noticed that highly variable TADs are not randomly distributed but cluster in shared genomic regions (variable TAD regions; Fig. 4e and Extended Data Fig. 4d).

To explore these variable TAD regions, we focused on a specific example that showed a low detection score, suggesting its structural variability (chr4: 38.5–43.6 Mb, average score across the three TADs identified in this region = 0.00, s.d. = 0.06; Extended Data Fig. 4f). We identified two groups of cells containing differences in genome

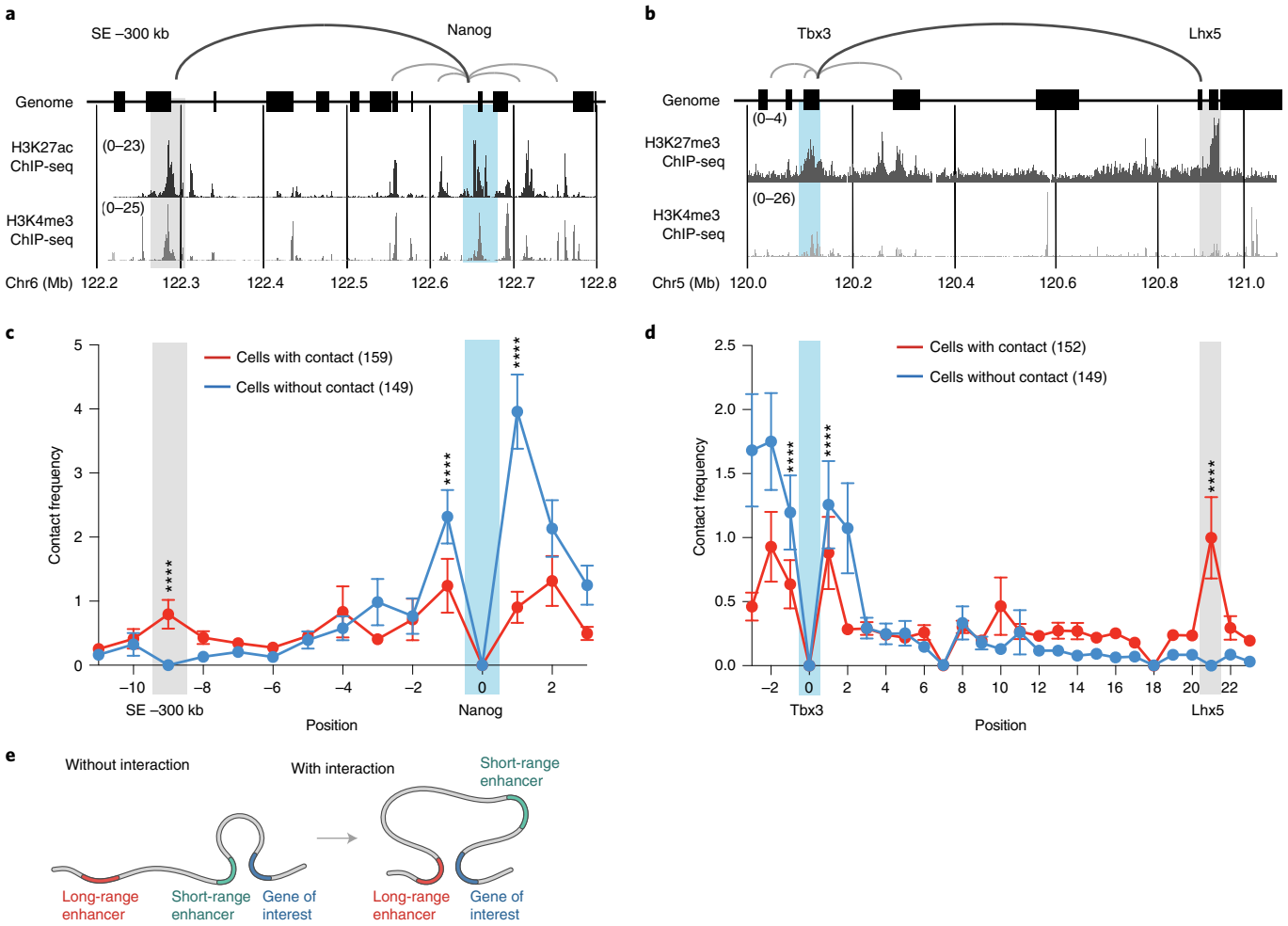


Fig. 5 | Heterogeneous structural states formed by *Nanog* and *Tbx3* loci in individual mESCs. a, Representation of the *Nanog* locus and its DNA interactions with SEs: 122.2–122.8-Mb region in chr6 with corresponding ChIP-seq tracks for H3K27ac and H3K4me3; *Nanog*–SE interaction (black lines). **b**, Representation of *Tbx3* locus and its DNA interactions with *Lhx5*: 120.0–121.0-Mb region in chr5 with the corresponding ChIP-seq tracks for H3K27me3 and H3K4me3; *Tbx3*–*Lhx5* interaction (black line). **c**, Normalized contact frequency plot between *Nanog* locus and 122.2–122.8-Mb surrounding region in chr6. Shown are cells containing (red) or lacking (blue) the contact between the *Nanog* locus and SE −300 kb. Each position refers to a 40-kb bin. Asterisks denote statistical significance ($P < 0.0001$, unpaired two-sided *t*-test with Welch's correction) between the two groups at the specified positions ($n = 1{,}000$ random bootstrap groups for each of the two groups). Error bars represent 1 s.d. **d**, Normalized contact frequency plot between *Tbx3* locus and 120.0–121.0-Mb surrounding region in chr5. Shown are cells containing (red) or lacking (blue) the contact between the *Tbx3* locus and *Lhx5*. Each position refers to a 40-kb bin. Asterisks denote statistical significance ($P < 0.0001$, unpaired two-sided *t*-test with Welch's correction) between the two groups at the specified positions ($n = 1{,}000$ random bootstrap groups for each of the two groups). Error bars represent 1 s.d. **e**, Schematic illustrating differences in structure when a gene of interest lacks (left) or contains (right) the long-range enhancer interaction. ChIP-seq, chromatin immunoprecipitation with sequencing.

organization at that region (Fig. 4f). Specifically, we detected a population of cells that contain an alternative TAD that spans the boundary of the ensemble-defined A/B compartment (Fig. 4f and Extended Data Fig. 4f). When focusing exclusively on cells that contain this alternative TAD, we found that the A/B compartments defined in those cells are distinct from the ensemble population (Fig. 4f,g). We confirmed that these distinct structural states are not explained by differences in cell cycle (Extended Data Fig. 4g) or by other major structural changes between these two groups of cells (Extended Data Fig. 4h). This suggests that this region is present in at least two distinct—and mutually exclusive—structural states in different cells in the population.

Together, our results demonstrate that the scSPRITE method can detect TAD-like genome organization in individual cells and identifies structural differences at the level of TADs in single cells. More studies are required to define if these cell-to-cell variabilities and region-to-region differences are functionally relevant and if they are characterized by other features such as transcription, specific chromatin marks or weak insulation boundaries (Supplementary Note 2).

scSPRITE detects heterogeneity across long-range contacts. We next asked if scSPRITE could detect structural changes that reflect biologically significant long-range DNA contacts, such as the interactions between promoters and super enhancers (SEs) or between regions enriched in polycomb group proteins (PcGs)[36]. SEs are large domains enriched in H3K27 acetylation that are thought to modulate gene expression by forming loops with promoters[6]. Bulk genome-wide studies have shown that SEs can form long- and short-range interactions with the same promoter[37–39], but it remains unclear whether these interactions occur simultaneously in the same cell. Similarly, DNA regions bound by PcGs have been shown

to interact across long distances to regulate gene expression;[36] however, it remains unclear how heterogenous these long-range interactions are in a population of cells.

We focused on two examples of long-range interactions in mESCs: (1) the *Nanog* locus, a key pluripotency factor in embryonic stem cells whose promoter interacts with multiple enhancers over a broad range of distances (up to 300 kb)[37,40] (Fig. 5a); and (2) the *Tbx3* locus, a transcription factor involved in the maintenance of pluripotency[41] whose locus interacts with another PcG-enriched gene, *Lhx5* (760 kb downstream) (Fig. 5b).

We selected cells with coverage over the *Nanog* and *Tbx3* regions of interest and split them into two groups based on whether we observed a contact between the target locus and the long-range enhancer (300 kb upstream for *Nanog* and 760 kb downstream for *Tbx3*) (Extended Data Fig. 5a). We computed the frequency of contacts between the target locus and all 40-kb bins for each group of cells (Fig. 5c,d). We noticed that, in the group with long-range interactions detected, short-range interactions were significantly weaker ($P < 0.001$) and, on average, three times less frequent (Fig. 5c,d). Additionally, we observed that detected long-range interactions span across a TAD border identified in the ensemble dataset for both *Nanog* and *Tbx3* examples (Extended Data Fig. 5a,d). We confirmed that the observed structural differences were not caused by technical differences (for example, number of reads in each group of cells) (Extended Data Fig. 5b,e) or different cell cycle phases (Extended Data Fig. 5c,f).

Our results demonstrate that, in cells where either the *Nanog* or *Tbx3* locus contacts the long-range region, the locus is less likely to form a contact with the short-range region (and vice versa) (Fig. 5e). Surprisingly, we detected long-range and short-range interactions in a similar number of cells, suggesting that both of these states are present at similar frequencies in mESCs. Whether such heterogeneity is a more global occurrence or restricted to specific loci (for example, transcription factors regulating pluripotency), and what (if any) functional role these distinct structures might play, remain to be determined (Supplementary Note 3).

Discussion

We have described scSPRITE, a method to generate high-resolution, genome-wide maps of 3D DNA organization in thousands of single cells. scSPRITE expands the toolkit of genome-wide, single-cell sequencing-based methods with an approach that enables high-resolution structural views across a broad spectrum of DNA interactions into high-throughput contact maps of the entire genome. In contrast to existing methods, scSPRITE does not require specialized equipment, techniques or training, and provides increased resolution from a lower number of sequencing reads across a larger number of cells. Because of this, we expect that it will expand the availability of single-cell genome structure measurements to any molecular biology laboratory. Additionally, we expect that scSPRITE can be scaled to work with as few as hundreds or as many as several thousands of cells simultaneously.

Our results reveal several novel insights about the heterogeneity of genome organization in mESCs. Specifically, we detected long-range higher-order interactions of both active (nuclear speckle) and inactive (centromeres and nucleolar contacts) chromatin regions as well as heterogenous organization of TADs and enhancer–promoter contacts between individual single cells. We note that our experiments were performed in a population of mESCs cultured using a '2 inhibitor' (2i) cocktail that is thought to promote ground state pluripotency and display more homogenous expression profiles across single cells[24,26,42] (Supplementary Note 3). Nonetheless, our results suggest that, even in these conditions, nuclear organization can be heterogeneous. Whether these cell-to-cell differences in 3D structure affect gene expression or have other functional significance remains to be determined.

Although our initial study focused on mESCs, scSPRITE can be applied to different cell types or homogenized tissues that are composed of mixed cell populations. One of the current challenges with studying complex tissues (for example, brain) or disease states (for example, tumors) is the heterogeneity of their cellular composition. The application of scSPRITE to such cell populations will enable studies of intrinsically heterogeneous systems and provide an accurate global view of their 3D genome organization. Accordingly, we expect that scSPRITE will provide the field with a path toward understanding the relationship between 3D genome organization and genome function in single cells.

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/s41587-021-00998-1.

Received: 20 July 2020; Accepted: 25 June 2021;
Published online: 23 August 2021

References

1. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289 (2009).
2. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
3. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
4. Dekker, J. & Mirny, L. The 3D genome as moderator of chromosomal communication. *Cell* **164**, 1110–1121 (2016).
5. Freire-Pritchett, P. et al. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *eLife* **6**, e21926 (2017).
6. Whyte, WarrenA. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
7. Quinodoz, S. A. et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174**, 744–757.e724 (2018).
8. Mao, Y. S., Zhang, B. & Spector, D. L. Biogenesis and function of nuclear bodies. *Trends Genet.* **27**, 295–306 (2011).
9. Miura, H. et al. Single-cell DNA replication profiling identifies spatiotemporal developmental dynamics of chromosome organization. *Nat. Genet.* **51**, 1356–1368 (2019).
10. Kagey, M. H. et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010).
11. Chen, Y. et al. Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J. Cell Biol.* **217**, 4025–4048 (2018).
12. Pederson, T. The nucleolus. *Cold Spring Harb. Perspect. Biol.* **3**, a000638 (2011).
13. Finn, E. H. et al. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* **176**, 1502–1515 (2019).
14. Wang, S. et al. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598 (2016).
15. Stevens, T. J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).
16. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).
17. Ma, X., Ezer, D., Adryan, B. & Stevens, T. J. Canonical and single-cell Hi-C reveal distinct chromatin interaction sub-networks of mammalian transcription factors. *Genome Biol.* **19**, 174 (2018).
18. Mohammed, H. et al. Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Rep.* **20**, 1215–1228 (2017).
19. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
20. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, eaau1783 (2018).
21. Giorgetti, L. et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**, 950–963 (2014).
22. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).

23. O'Sullivan, J. M., Hendy, M. D., Pichugina, T., Wake, G. C. & Langowski, J. The statistical-mechanics of chromosome conformation capture. *Nucleus* **4**, 390–398 (2013).
24. Kolodziejczyk, A. A. et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**, 471–485 (2015).
25. Guo, F. et al. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* **27**, 967–988 (2017).
26. Ghimire, S. et al. Comparative analysis of naive, primed and ground state pluripotency in mouse embryonic stem cells originating from the same genetic background. *Sci. Rep.* **8**, 5884 (2018).
27. Lee, D.-S. et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat. Methods* **16**, 999–1006 (2019).
28. Zhou, S., Jiang, W., Zhao, Y. & Zhou, D.-X. Single-cell three-dimensional genome structures of rice gametes and unicellular zygotes. *Nat. Plants* **5**, 795–800 (2019).
29. Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).
30. Ramani, V. et al. Sci-Hi-C: a single-cell Hi-C method for mapping 3D genome organization in large number of single cells. *Methods* **170**, 61–68 (2020).
31. Spector, D. L. & Lamond, A. I. Nuclear speckles. *Cold Spring Harb. Perspect. Biol.* **3**, a000646 (2011).
32. Guenatri, M., Bailly, D., Maison, C. L. & Almouzni, G. V. Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *J. Cell Biol.* **166**, 493–505 (2004).
33. Almouzni, G. & Probst, A. V. Heterochromatin maintenance and establishment: lessons from the mouse pericentromere. *Nucleus* **2**, 332–338 (2011).
34. Strongin, D. E., Groudine, M. & Politz, J. C. R. Nucleolar tethering mediates pairing between the *IgH* and *Myc* loci. *Nucleus* **5**, 474–481 (2014).
35. Dowen, J. M. et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374–387 (2014).
36. Pachano, T., Crispatzu, G. & Rada-Iglesias, A. Polycomb proteins as organizers of 3D genome architecture in embryonic stem cells. *Brief. Funct. Genomics* **18**, 358–366 (2019).
37. Blinka, S., Reimer, Michael, H. Jr., Pulakanti, K. & Rao, S. Super-enhancers at the *Nanog* locus differentially regulate neighboring pluripotency-associated genes. *Cell Rep.* **17**, 19–28 (2016).
38. Novo, C. L. et al. Long-range enhancer interactions are prevalent in mouse embryonic stem cells and are reorganized upon pluripotent state transition. *Cell Rep.* **22**, 2615–2627 (2018).
39. Schoenfelder, S. et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
40. Apostolou, E. et al. Genome-wide chromatin interactions of the nanog locus in pluripotency, differentiation, and reprogramming. *Cell Stem Cell* **12**, 699–712 (2013).
41. Russell, R. et al. A dynamic role of TBX3 in the pluripotency circuitry. *Stem Cell Rep.* **5**, 1155–1170 (2015).
42. Kalmar, T. et al. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol.* **7**, e1000149 (2009).

Methods

Cell types and culture conditions. We developed scSPRITE using mouse and human cells, focusing primarily on mESCs because their genome structure has been extensively studied[3,15].

We used a male embryonic stem cell line (bsps derived from V6.5 embryonic stem cell line, provided by K. Plath) and cultured them in serum-free 2i/LIF medium as previously described[43]. We suspect that this mESC line displays trisomy in chromosome 8 because the average number of reads aligning to chr8 is about 33% greater than the average number of reads across the other chromosomes (Fig. 1h).

HEK293T, a female human embryonic kidney cell line transformed with the SV40 large T antigen, was obtained from the American Type Culture Collection (no. CRL-1573) and cultured in complete media consisting of DMEM (no. 11965092, Gibco, Life Technologies) supplemented with 10% FBS (Seradigm Premium Grade HI FBS, VWR), 1× penicillin–streptomycin (Gibco, Life Technologies), 1× MEM non-essential amino acids (Gibco, Life Technologies), 1 mM sodium pyruvate (Gibco, Life Technologies) and maintained at 37 °C under 5% CO_2. For maintenance, 800,000 cells were seeded into 10 ml of complete media every 3–4 d in 10-cm plates.

scSPRITE protocol. *Cell crosslinking.* Media from mESCs was removed and washed once with 1× PBS. Cells on the 10-cm plates were then trypsinized using 2 ml of 0.025% trypsin-EDTA (pre-warmed to 37 °C). Plates were incubated at 37 °C for 5 min, and the trypsinized cells were mixed by pipetting to break up any clumps. We added 8 ml of pre-heated wash solution (DMEM/F12 + BSA, pre-warmed to 37 °C) to the plate to inactivate trypsin before transferring the cells to a conical tube. Cells were centrifuged at 330*g* for 3 min, and the supernatant was discarded. Cells were washed once with 1× PBS at a ratio of 4 ml of PBS per 1×10^7 cells and centrifuged again at 330*g* for 3 min. After the wash, 4 ml of 2 mM disuccinimidyl glutarate (DSG, Life Technologies, no. 20593) prepared in 1× PBS was added per 1×10^7 cells to the conical tube, and the solution was mixed thoroughly by pipetting to remove clumps. The cells in DSG solution were gently shaken for 45 min at room temperature. After incubation with DSG, 200 µl of 2.5 M glycine was added per 1 ml of DSG solution previously added to quench the reaction, and the tube was gently shaken for 5 min at room temperature. Cells were then centrifuged at 1,000*g* for 4 min, and the supernatant was discarded. Cells were washed with 1× PBS at a ratio of 4 ml of PBS per 1×10^7 cells and centrifuged again at 1,000*g* for 4 min. After the wash, 4 ml of 1% formaldehyde (16% wt/vol formaldehyde ampules, Life Technologies, no. 28908, prepared in pre-warmed (37 °C) 1× PBS) was added per 1×10^7 cells to the conical tube, and the solution was mixed thoroughly by pipetting to remove clumps. The cells in formaldehyde solution were then gently shaken for 10 min at room temperature. After incubation with formaldehyde, 200 µl of 2.5 M glycine was added per 1 ml of formaldehyde solution previously added to quench the reaction, and the tube was gently shaken for 5 min at room temperature. Cells were then centrifuged at 1,000*g* for 4 min, and the supernatant was removed. Cells were twice washed with cold 1× PBS + 0.5% BSA (wt/vol) solution, and centrifugation was done at 4 °C at 1,000*g* for 4 min. After the washes, enough cold 1× PBS + 0.5% BSA solution was added to get a cell concentration of 5×10^6 cells per ml. Crosslinked cells were then aliquoted in new 1.5-ml LoBind Eppendorf tubes, centrifuged (2,000*g* for 5 min) to remove the supernatant and flash-frozen in liquid nitrogen. Cells were kept at −80 °C until used for analyses.

Cell lysis and nuclei preparation. Crosslinked cells were thawed from −80 °C and were kept on ice during the cell lysis procedures. Initially, 1.4 ml of lysis buffer #1 (50 mM HEPES pH 7.4, 1 mM EDTA pH 8.0, 1 mM EgTA pH 8.0, 140 mM NaCl, 0.25% Triton X-100, 0.5% IGEPAL CA-630, 10% glycerol, 1× proteinase inhibitor cocktail (PIC)) was added per 1×10^7 cells. The cell solution was mixed thoroughly before incubating on ice for 10 min. Cells were pelleted afterwards at 900*g* for 8 min at 4 °C, and the supernatant was removed. Next, 1.4 ml of lysis buffer #2 (10 mM Tris-HCl pH 8, 1.5 mM EDTA, 1.5 mM EgTA, 200mM NaCl, 1× PIC) was added per 1×10^7 cells. Again, the cell solution was mixed thoroughly before incubating on ice for 10 min. Cells were pelleted afterwards at 900*g* for 9 min at 4 °C, and the supernatant was removed. Afterwards, the cells were washed in 800 µl of 1.2× CutSmart solution (from 10× CutSmart stock (NEB, no. B7204S)) and pelleted at 900*g* for 2 min. The supernatant was removed, and a fresh 400 µl of 1.2× CutSmart solution was added carefully to not resuspend the pellet. Then, 6 µl of 20% SDS was added to the tube, and the cells were thoroughly resuspended. The cell solution was mixed on an Eppendorf ThermoMixer C at 1,200 r.p.m. for 60 min at 37 °C to isolate nuclei. Next, 40 µl of 20% Triton X-100 was added to the same tube to quench the reaction, and the solution was left mixing on the same instrument at 1,200 r.p.m. for 60 min at 37 °C. Lastly, 30 µl of 5,000 U ml^{-1} HpyCH4V (NEB, no. R0620L) was added to the same tube to allow for DNA to be digested in-nuclei. In-nuclei digestion was performed for 4 h at 37 °C while shaking at 1,200 r.p.m. HpyCH4V is a 4-base pair (bp) restriction enzyme that performs blunt-end cutting at TGCA sequences. This particular enzyme was chosen because it was able to perform in-nuclei enzymatic restriction digestion and eliminated the need to perform any additional DNA strand repair steps after restriction digest. After 4 h of restriction digest, the average DNA fragment size was 823 bp.

After digestion, nuclei were pelleted at 900*g* for 2 min, the supernatant was removed and nuclei were washed three times with 1× PBS, 1 mM EDTA, 1 mM EgTA and 0.1% Triton X-100 solution at 900*g* for 2 min. After the washes, the nuclei concentration was assessed by loading 6 µl of the solution into a disposable hemocytometer (4-Chip Disposable Hemocytometer, Bulldog Bio, no. DHC-N420). After determining nuclei concentration, 5×10^5 nuclei were transferred by pipetting into a new 1.5-ml LoBind Eppendorf tube. In this new tube, 25 µl of dA-tail reaction buffer and 10 µl of Klenow fragment were added to the nuclei (both reagents were part of NEBNext dA-Tailing Module (NEB, no. E6053L)). The tube was filled to 250 µl using nuclease-free H_2O, and dA-tailing was performed in-nuclei at 37 °C for 90 min while shaking at 1,200 r.p.m. The reaction was then stopped with the addition of 200 µl of 1× PBS, 50 mM EDTA, 50 mM EgTA and 0.1% Triton X-100. The nuclei pellet was spun down at 900*g* for 2 min and washed twice using 400 µl of 1× PBS, 1 mM EDTA, 1 mM EgTA and 0.1% Triton X-100 solution. After the washes, the nuclei were resuspended in fresh 1× PBS, 1 mM EDTA, 1 mM EgTA and 0.1% Triton X-100 solution, and nuclei concentration was determined again using the hemocytometer, as described previously.

In-nuclei combinatorial barcoding. To uniquely identify DNA sequences originating from the same cell, combinatorial barcoding was performed in-nuclei (Fig. 1a). In our specific experiments, we used three rounds of combinatorial barcoding in the following order: 'DNA phosphate modified' (DPM), 'odd' tagging and 'even' tagging (these tags are described in the original SPRITE paper[7]). The resulting tags were pre-loaded onto a 96-well plate, with each well containing 2.4 µl of a uniquely barcoded tag at a concentration of 45 µM. Nuclei previously dA-tailed were washed twice in a solution of 1× PBS, 0.1% Triton X-100 and 0.3% BSA (wt/vol), and nuclei concentration was reassessed using a hemocytometer, as described previously.

To perform in-nuclei barcoding, 2×10^5 nuclei were withdrawn and transferred into a new 1.5-ml LoBind Eppendorf tube and filled to 1,125 µl using a solution of 1× PBS, 0.1% Triton X-100 and 0.3% BSA (wt/vol). The nuclei solution was well-mixed before loading 11.2 µl of nuclei solution into each well of a 96-well plate. Each well was then supplemented with 6.4 µl of ligation mix (220 µl of 2× Instant Sticky-end Ligase Master Mix (NEB, no. M0370), 352 µl of 5× Quick Ligase Buffer (NEB, no. B6058S) and 132 µ of 1,2-propanediol (Sigma-Aldrich, no. 398039)). The 96-well plate was sealed after loading a ligation mix and was mixed on an Eppendorf ThermoMixer C at 20 °C. The reaction was performed for 3 h while mixing at 1,600 r.p.m. for 30 s every 5 min.

After performing in-nuclei DNA ligation, 20 µl of 1× PBS, 50 mM EDTA, 50 mM EgTA and 0.1% Triton X-100 solution was added to each well and incubated for 10 min at 20 °C to stop the ligation reaction. Next, a solution of 80 µl of 1× PBS, 50 mM EDTA, 50 mM EgTA and 0.1% Triton X-100 (wt/vol) was added to each well, and all the contents of the well plate were pooled together into a new 15-ml conical tube. The 96-well plate was washed once with a solution of 100 µl of 1× PBS, 50 mM EDTA, 50 mM EgTA and 0.1% Triton X-100 and pooled together into the same conical tube. Nuclei were pelleted at 800*g* for 10 min, and all but 1 ml of supernatant was removed from the tube. The nuclei were resuspended before transferring to a new 1.5-ml non-LoBind Eppendorf tube. In the new Eppendorf tube, nuclei were washed twice with a solution of 500 µl of 1× PBS, 0.1% Triton X-100 and 0.3% BSA (wt/vol) at 900*g* for 2 min. This in-nuclei ligation process was repeated two more times, resulting in a total of three tags (the 'DPM', 'odd' and 'even' tags) being ligated to DNA fragments.

Once the three rounds of in-nuclei barcoding process was completed, nuclei were filtered through a 10-µm mesh filter (pluriStrainer, no. 43-10010-50) into a new 1.5-mL non-LoBind Eppendorf tube to ensure that only single cells were isolated (Extended Data Fig. 1a). Filtered nuclei were then pelleted at 900*g* for 2 min, and the supernatant was removed. Nuclei were resuspended and washed twice in lysis buffer #3 (1.5 mM EDTA, 1.5 mM EgTA, 100 mM NaCl, 0.1% sodium deoxycholate, 0.5% sodium lauroyl sarcosinate) at 900*g* for 2 min. Next, the concentration of nuclei was determined, and 1,500 nuclei were withdrawn and used in the following steps of the protocol.

Scaling the number of cells to analyze and the number of barcoding rounds. In our experiments, we used a final concentration of 1,500 individual nuclei and performed three rounds of barcoding to generate 96^3 (884,736) barcode combinations. This results in 590-fold excess barcode combinations to the number of cells analyzed and results in <1 expected cell 'collisions' (where cells obtain the same complete barcode string). To provide some intuition on these numbers, we note that the probability that any two cells will have a 'collision' is defined by a Poisson distribution with a mean (λ) defined by the number of cells divided by the number of barcode combinations. The probability of observing two or more cells with the same barcode in this distribution is defined as the p(x>1). The expected number of collisions is the number of measured cells multiplied by this probability of collision. Accordingly, to analyze 10,000 cells with 100-fold barcode excess (100,000 barcode combinations) would yield <1 expected cell collisions. Thus, the number of cells analyzed can be adjusted to enable analysis of larger numbers (or smaller numbers) based on the needs of the application. Adjusting cell numbers might require adjusting the numbers of rounds of barcoding to enable accurate separation of individual cells. We recommend between 10- and 100-fold excess

barcode combinations to the number of cells analyzed. The exact excess used depends on how many potential collisions would be tolerated in the final output.

Sonication. Next, 1,500 nuclei were placed into a Covaris microtube-15 and filled to 15 µl using lysis buffer #3. The Covaris tube was placed in the Covaris M220 Focused-ultrasonicator, and sonication was performed for 2 min under specific settings (water temperature 6 °C, incident power 30 W, duty cycle 3.3) to release DNA complexes from nuclei. The tube was then removed from the instrument and set on ice.

At this step of the protocol, it is important to proceed with all the sonicated nuclei, as sampling them further will lead to a loss of nuclei fragments and will prevent the analysis of DNA structure in single cells. We also recommend adjusting the sonicated number of cells to sequencing abilities to achieve satisfactory coverage per cell.

N-hydroxysuccinimide beads coupling. After sonication, sample containing crosslinked DNA complexes was coupled to N-hydroxysuccinimide (NHS) beads as previously described[7]. Briefly, NHS-Activated Magnetic Beads (Life Technologies, no. 88826) were activated for coupling. First, 600 µl of NHS beads were withdrawn and placed in a 1.5-ml LoBind Eppendorf tube. The tube was placed on a DynaMag-2 magnet, and the the supernatant was removed. The beads were washed once with 600 µl of ice-cold 1M HCl, and the supernatant was removed again and replaced with 600 µl of ice-cold 1× PBS. After removing 1× PBS, the beads were resuspended in 500 µl of 1× PBS + 0.1% SDS. Additionally, 85 µl of 1× PBS + 0.1% SDS was added to the previously sonicated nuclei solution, mixed and added to the bead solution. The complexes were then coupled to NHS beads on an Eppendorf ThermoCycler C overnight at 4 °C while shaking at 1,200 r.p.m. After coupling, the flowthrough was removed, and 600 µl of 1M Tris-HCl pH 7.5, 0.5 mM EDTA, 0.5 mM EgTA and 0.1% Triton X-100 was added to the beads to quench the remaining NHS groups; this was done at 4 °C at 1,200 r.p.m. for 60 min. Once the beads were quenched, the flowthrough was removed, and the beads were washed twice in cold RLT2+ buffer (0.2% sodium lauryl sarcosinate, 1 mM EDTA, 1 mM EgTA, 10 mM Tris-HCl pH 7.5, 0.1% Triton X-100, 0.1% NP-40, filled to the final volume with RLT (Qiagen, no. 79216)). This was followed by three washes in M2 buffer (50 mM NaCl, 20 mM Tris-HCl pH 7.5, 0.2% Triton X-100, 0.2% NP-40, 0.2% sodium deoxycholate). The beads were then resuspended in a mix of M2 buffer and H_2O (58% M2, 42% H_2O) to attain a total volume of 1,125 µl of M2 buffer, H_2O and beads.

Spatial barcoding/complex-specific barcoding. Next, spatial barcoding of the DNA complexes on beads was performed as described previously[7]. First, the bead solution was well-mixed and loaded into each well of a 96-well plate (11.2 µl of bead solution per well). Each well of the plate contained 2.4 µl of uniquely barcoded tag at a concentration of 4.5 µM. Next, each well was supplemented with 6.4 µl of ligation mix (220 µl of 2× Instant Sticky-end Ligase Master Mix (NEB, no. M0370), 352 µl of 5× Quick Ligase Buffer (NEB, no. B6058S) and 132 µl of 1,2-propanediol (Sigma-Aldrich, no. 398039)). The 96-well plate was sealed after loading a ligation mix and was mixed on an Eppendorf ThermoMixer C at 20 °C. The reaction was performed for 60 min with mixing at 1,600 r.p.m. for 30 s every 5 min. Afterwards, the reaction was stopped by adding 60 µl of RLT2+ buffer to each well before pooling the solutions of each well into a 25-ml reservoir. Each well was then rinsed once with 100 µl of RLT2+ buffer to remove residual beads and pooled into the same 25-ml reservoir. The solution was then transferred to a 15-ml conical tube, which was then placed on a magnet to remove most of the RLT2+ buffer from the beads. With about 2 ml of RLT2+ buffer remaining, the beads were resuspended and transferred to a LoBind 1.5-ml Eppendorf tube, which was placed on a DynaMag-2 magnet to remove the remaining RLT2+ buffer. The beads were washed three times with 600 µl of M2 buffer. This process of split-pool barcoding on beads was repeated until the three additional tags were added. After the last round of split-pool barcoding was completed, the beads were resuspended in 600 µl of MyK buffer (20 mM Tris-HCl pH 8.0, 0.2% SDS, 100 mM NaCl, 10 mM EDTA, 10 mM EgTA, 0.5% Triton X-100) after the washes.

We performed three rounds of spatial tagging because it provided sufficient barcode combinations to uniquely label DNA complexes coming from each individual cell. Briefly, the mouse genome contains 2.5×10^9 nucleotides, which, when divided per the average fragment size of DNA after digestion (823 bp), results in 3.04×10^6 DNA fragments per cell. If we do three rounds of barcoding, we provide 884,736 number of combinations, which exceeds the number of DNA molecules 3.4 times. Notably, during scSPRITE barcoding, we distribute clusters of DNA molecules, not single molecules, so the actual number of barcode combinations will exceed the number of spatial clusters much more than our calculation.

Library preparation. To ensure that we capture all information coming from single cells, we need to sequence all DNA molecules that were bound to the beads. The bead solution was split equally into ten LoBind Eppendorf 1.5-ml tubes, with each tube containing 60 µl of beads in MyK buffer. Next, an additional 32 µl of MyK buffer and 8 µl of proteinase K (NEB, no. P8107S) were added to each tube. All ten tubes were placed on an Eppendorf ThermoCycler C, and reverse crosslinking proceeded overnight at 60 °C while shaking at 1,200 r.p.m. Next, the tubes were placed on a DynaMag-2 magnet, and the MyK and proteinase K solution was transferred to ten new LoBind Eppendorf tubes. The beads from each of the tubes were washed once with 20 µl of H_2O and then transferred to the same tube containing each respective MyK and proteinase K solution. DNA from each of the tubes were purified using the Clean & Concentrator-5 columns (Zymo, no. D4004) using 5× binding buffer to increase yield. Purified DNA from each column was eluted in ten new Eppendorf 1.5-ml tubes using 12 µl of H_2O. Each of the tubes were filled to 30 µl using 15 µl of Q5 Hot Start High-Fidelity 2× Master Mix (NEB, no. M0493S), 1.5 µl of 20× EvaGreen (Biotium, no. 31000-T), 1.2 µl of 25 µM indexed Illumina primers and 0.3 µl of H_2O. Real-time PCR amplification proceeded for 14 cycles, which was when the libraries entered exponential amplification but had not plateaued. After amplification, each of the libraries was diluted four-fold before running on a 1% agarose E-gel (Life Technologies, no. G402001) with an E-Gel 1-Kb Plus DNA Ladder (Life Technologies, no. 10488090) as a reference. After the run, the gel was cut between 300- and 1,000-bp marks to remove primer dimers, small non-specific amplicons and long DNA amplicons. Libraries from the gel were purified using a Gel Purification Kit (Zymo, no. D4002) as described by the manufacturer, and 20 µl of H_2O was used to elute libraries off the column.

To estimate the number of unique molecules in our libraries, the molarity of our libraries was determined using the concentration of our library from Qubit 3.0 Fluorometer (using the Qubit dsDNA HS Assay Kit) and the average library size (bp) using an Agilent TapeStation 2200 (using the Agilent High Sensitivity D1000 ScreenTape and reagents). This, in addition to estimated losses during library cleanup, allowed us to estimate the number of unique molecules in our libraries. The libraries were sequenced with a read depth of 2.4× to ensure that we are able to map the DNA contained in each cluster.

scSPRITE data generation. scSPRITE data were generated using Illumina paired-end sequencing on the NovoSeq through Novogene Corporation. Reads were sequenced with at least 120 bp in Read 1 for genomic DNA information and the DPM tag and 95 bp in Read 2 to read the other five remaining tags (odd - even - odd - even - Y-even) (Extended Data Fig. 1c). We generated 1,269,693,929 reads from the scSPRITE library made from ~1,500 cells. From the FastQC report, we observe a normal distribution of GC content per sequence (Read 1: normal distribution between ~15% and 71%; Read 2: normal distribution between ~27% and 59%).

Sequencing analysis pipeline. The full barcode sequence was identified by combining the DNA tag sequence from the beginning of Read 1 and the remaining five barcode tags from Read 2 (Extended Data Fig. 1c). The tags were identified from a table of known tag sequences, as previously described[7], with odd and even tags allowing up to two mismatches and DPM, and Y-even tags allowing zero mismatches. Out of 1,269,693,929 reads sequenced, we identified 26,546,674 (2.1%) reads with zero barcodes, 62,183,357 (4.9%) reads with one barcode, 116,086,266 (9.1%) reads with two barcodes, 291,410,130 (22.9%) reads with three barcodes, 33,689,683 (2.7%) reads with four barcodes, 107,535,755 (8.5%) reads with five barcodes and 632,242,064 reads (49.8%) that contained the full six-barcode sequence. Any reads that lacked the full six-barcode sequence (DPM - odd - even - odd - even - Y-even) in the expected order were discarded from further analysis and considered not usable for identifying cell of origin. The remaining 632,242,064 reads are, therefore, considered usable and were kept for downstream alignment and filtering. Before alignment, Read 1 was trimmed to a length of 100 bp (Extended Data Fig. 1c).

Alignment and filtering of reads. The trimmed reads containing the full six-barcode sequence were mapped to pre-indexed mm9 reference genome using STAR 2.6.1 using the following parameters: –outFilterMultimapNmax 50–outFilterScoreMin OverLread 0.30–outFilterMatchNminOverLread 0.30–outFilterIntronMotifs None–alignIntronMax 50000–alignMatesGapMax 1000–genomeLoad NoShared Memory–outReadsUnmapped Fastx–alignIntronMin 80–alignSJDBoverhangMin 5–sjdbOverhang 100–limitOutSJcollapsed 10000000–limitIObufferSize=300000000. SAMtools 1.9 was applied to filter-mapped reads, and only uniquely mapped reads (-q 255) were kept. Alignments that had overlapped a masked region as denoted by RepeatMasker (UCSC, milliDiv < 140) were removed using bedtools (version 2.25.0). Finally, reads that were aligned to a mm9 non-unique region of the genome were removed by excluding alignments that mapped to regions by the ComputeGenomeMask program (read length = 35 nucleotides). After these filtration steps, all BAM files that corresponded to the same sample but contained different Illumina primers at sequencing were pooled together before cluster identification (Extended Data Fig. 1c).

Cluster barcode and cell barcode identification. To identify SPRITE clusters, all reads that contained the same six-barcode sequences were grouped together into a single cluster. All reads containing the same six-barcode sequences that started at the same genomic position were excluded to remove possible PCR duplicates. This led to 161,989,473 remaining reads. Once identified, a SPRITE cluster file was generated where each line contained the cluster barcode name and corresponding

genomic alignments. Once the cluster barcodes were identified, the cell barcodes were identified by grouping clusters together that contained the same DPM, first odd and first even barcode sequences. This grouping can create on the order of hundreds of thousands of cell barcode files, but most of these files contain fewer than ten clusters. As a result, only the largest 4,000 cell barcode files based on file size were selected for downstream filtration, and the remaining cell barcode files were removed from the directory (Extended Data Fig. 1c).

Selecting single cells for analysis. Once the largest 4,000 cell barcode files were identified, these files underwent additional in silico filtration to select the most informative files for analysis (Extended Data Fig. 1c). The files were rank-ordered based on the number of clusters. The 1,500 cell barcode files with the largest number of clusters from the initial 4,000 files were selected, consistent with the initial number of cells used for the scSPRITE experiment. To ensure that we selected only single cells for downstream analysis and not cell doublets, we removed the top 3.4% of cells as determined from the detected collision rate calculated from the results of the human–mouse mixing experiment (Fig. 1b and Extended Data Fig. 1b). To ensure that we focus on the cells with the most information per cell (number of reads/cDNA cluster/DNA contacts), we selected the top 1,000 cell barcode files containing the most number of clusters per cell for downstream single-cell analysis. This led to 107,181,084 usable reads from the top 1,000 cell barcode files.

Next, in the 1,000 cells, we calculate the size distribution of DNA clusters per each cell and remove large clusters (>10,000 reads per cluster) from further analysis. We previously reported[7] that clusters larger than >10,000 reads per cluster contain less information about higher-resolution structures (that is, TADs) and most likely contain big chunks of nuclei that are composed of several chromosomes. We consider them less informative for the type of DNA interactions/structures (background) and, therefore, remove them from the further analysis. Excluding all reads in the >10,000-read clusters led to 83,318,292 remaining reads that were used for all downstream analyses.

Human–mouse mixing experiment. To determine the percentage of single cells that are mixed together during scSPRITE (from crosslinking until the end of in-nuclei barcoding), we performed an in-nuclei part of the scSPRITE experiment using cell types from different species—mouse and human. We perform only the in-nuclei barcoding step because we previously showed[7] that the spatial barcoding step used in bulk SPRITE leads to minimal collisions if the total number of NHS beads is in an excess to the total number of clusters in a sample (their mixing human–mouse experiment detects more than 99% of reads aligning to one species).

mESCs (bsps) and human cells (HEK293T) were harvested and resuspended into a single-cell solution, and then 30×10^6 cells per each cell type were mixed together in equal quantities and crosslinked, digested, dA-tailed and barcoded in-nuclei as described above. Additionally, for the experiments described in Extended Data Fig. 1b, we mixed equal numbers of mouse and human cells after crosslinking but (1) before digestion or (2) after digestion, or (3) we proceeded to the next step without mixing. Four rounds of in-nuclei barcoding were done (DPM - odd - even - Y-even) (8×10^7 barcode combinations and 2×10^5 cells). Next, nuclei were filtered through a 10-µm filter (pluriStrainer); 300 nuclei were removed as a new sample and reverse crosslinked; and we proceeded as described above. Next, 10% of the total purified libraries were sequenced using MiSeq; reads were then aligned to combined human/mouse genome using STAR alignment (hg19 and mm9 reference genomes). The best alignment was taken into consideration, and, if reads align equally well, they were considered as multi mappers and removed from further analysis. Reads were sorted into individual cells based on cell-specific barcodes, and we focused only on cell barcodes that had more than 1,000 reads per cluster.

Next, we calculated the percentage of reads that aligned to each genome for each identified cell barcode. We categorized cell barcodes as mouse- or human-derived when they contained more than 95% single-species reads and as mixed when they contained less than 95% single-species reads. We then calculated the fractions of human-only, mouse-only and mixed-cell barcodes (Fig. 1b) and reported the percentage of mixed-cell barcodes as detected collision rate. Detected collision rate was further used to estimate thresholds used for cell filtering (Extended Data Fig. 1c and Methods) and to calculate total collision rate. Total collision rate represents an estimation of all possible collisions and relies on the assumption that cells from the same species show similar collision rates as cells from mixed population, but we cannot detect them in our mixing experiment. It is calculated as follows: detected collision rate (mixed cells) + detected collision rate (human cells) + collision rate (mouse cells).

We note that, despite starting with equal numbers of human and mouse cells, we observe bias in the final libraries with a higher number of mouse cells than human cells, which results in better coverage per single human cell than mouse cell (Fig. 1b). This is likely caused by the fact that we observed that, during the full scSPRITE procedure (nuclei isolation, DNA digestion and in-cell barcoding), human HEK293 fibroblast cells are more susceptible to fragmentation and, as a consequence, lead to higher cell loss. We think that this results in an unequal read distribution observed in our experiment and is consistent with other mouse–human mixing experiments when genomic methods such as scHi-C are used[29].

Data analysis. *Contact maps.* Generation of ensemble heat maps from scSPRITE. The generation of pairwise contact frequency matrices for ensemble scSPRITE was done similarly as was done for SPRITE[7]. For each cluster in the ensemble scSPRITE dataset, we gathered all possible pairs of reads. The pairwise contact frequency for each genomic bin *i* and *j* was then determined by counting the pairs of reads from each cluster, where both reads in a pair overlap with both *i* and *j* bins. These are unweighted clusters. To minimize the effect that larger clusters contribute toward the number of pairwise contacts between any two bins, we also generated downweighted pairwise contact frequency matrices. The pairwise contact frequency was downweighted by a factor of $2/n$, where n represents the number of reads in each cluster. The unweighted and downweighted contact frequency matrices were then normalized using Hi-Corrector[44]. In addition, low-coverage bins and contacts in the same bin are masked in heat maps.

To assess how well ensemble scSPRITE mapped known genomic structures, we compared the mouse embryonic stem cluster file from ensemble scSPRITE with the original mouse embryonic stem cluster file from SPRITE[7]. We used unweighted pairwise contact frequency matrices for genome-wide (1-Mb resolution) and A/B compartment (200-kb resolution) for both ensemble scSPRITE and SPRITE but using clusters containing fewer than 1,000 reads per cluster. Downweighted pairwise contact frequency matrices were used for TAD comparison (40-kb resolution) for both ensemble scSPRITE and SPRITE but using all clusters.

Generation of single-cell heat maps from scSPRITE. Similarly to ensemble scSPRITE, single-cell contact frequency matrices were generated at 1-Mb and 40-kb resolutions for all 1,000 filtered cells. Contact frequency matrices were made similarly as described previously for ensemble scSPRITE, where each value in the matrix reflects the number of clusters containing a read pair at genomic bin *i* and *j*. Single-cell maps remained unweighted unless otherwise stated.

Comparison of ensemble and scSPRITE chromosome territory heat maps. Genome-wide 1-Mb resolution contact maps were generated for the ensemble data set by pooling clusters containing fewer than 10,000 reads per cluster from the filtered 1,000 single-cells dataset. The resulting contact matrix for the ensemble dataset represents the non-downweighted contact frequency for each pair of 1-Mb bins throughout the genome. The ensemble contact matrix was normalized by performing Hi-Corrector before plotting.

For the single-cell maps, genome-wide 1-Mb resolution contact maps were generated by using clusters fewer than 10,000 reads per cluster for each single cell. The resulting contact matrix for each single cell represents the number of clusters that contained each pair of 1-Mb bins throughout the genome. Each single-cell contact matrix was normalized by dividing every value in the contact matrix by the largest value in the matrix, resulting in a value between 0 and 1.

Insulation scores and A/B compartment annotation. Insulation scores and annotations for A and B compartments were calculated from the ensemble scSPRITE dataset using cworld (https://github.com/dekkerlab/cworld-dekker). Insulation scores were calculated using contact maps binned at 40-kb resolution, and A and B compartment annotations were calculated using contact maps binned at 200-kb and 1-Mb resolutions. Insulation scores were calculated using the script matrix2insulation.pl with the parameters '–ss 80000–im iqrMean–is 480000–ids 320000', and compartment annotations were calculated using the script matrix2compartment.pl with default parameters. We used the output file ending in 'insulation.boundaries.bed'. These TAD regions correspond to the interval between two insulation boundaries. To quantitatively compare TADs between ensemble scSPRITE and Hi-C, we computed the correlation coefficient between the insulation scores for each 40-kb genomic bin (using the '.insulation' file output by the matrix2insulation.pl script).

Detection scores for 3D genome structures. Detection scores were calculated to identify various 3D genome structures in single cells. These structures included chromosome territories, A/B compartments, TADs, centromere interactions, nuclear speckle interactions and nucleolar interactions. Each score reflects how clearly defined a given structure is in a single cell. The scores were calculated using a binary contact matrix for each cell, which defined whether each pair of genomic bins was in contact in that cell. For example, a clearly defined chromosome territory in a single cell consists of chromosomes interacting more with themselves than with each other (illustration, Fig. 2b).

To normalize detection scores, an expected detection score was calculated for each 3D genome structure in each cell. The expected detection score was calculated as the mean detection score for 1,000 randomized structures, which were generated by randomly shuffling the genomic coordinates of known structures. The normalized detection score for each structure in each cell was calculated as the observed detection score minus the expected detection score (Supplementary Note 4).

Detection scores were calculated for each structure in each cell as follows:

- Chromosome territories: (observed intra-chromosomal contacts) / (total possible intra-chromosomal contacts) − (observed inter-chromosomal contacts) / (total possible inter-chromosomal contacts). Genome-wide scores were calculated for every possible pair of chromosomes between chr1 and

chr19, excluding combinations between the same chromosomes (for example, chr1-chr1), amounting to 171 combinations (Supplementary Table 1) from binary matrices at 1-Mb resolution (171 combinations because chrA-chrB = chrB-chrA).

- Compartments: (observed intra-compartment contacts) / (total possible intra-compartment contacts) − (observed inter-compartment contacts) / (total possible inter-compartment contacts). Genome-wide scores were calculated for all 224 regions across all chromosomes in which we detected a compartment switch in our ensemble dataset (for example, chr1 has 21 regions, and chr3 has zero regions) (Supplementary Table 2). A compartment switch is defined as a transition between 'A to B to A' or 'B to A to B' compartments. Scores were calculated from binary matrices at 1-Mb resolution.
- TADs: (observed intra-TAD contacts) / (total possible intra-TAD contacts) − (observed inter-TAD contacts) / (total possible inter-TAD contacts). Genome-wide TAD scores were calculated ± 1 Mb from TAD boundary regions, and these were calculated for all 2,602 TAD boundary regions that we detected in our ensemble dataset (Supplementary Table 3). Scores were calculated from binary matrices at 40-kb resolution.
- Centromere interactions: (observed centromere–centromere contacts) / (total possible centromere–centromere contacts) − (observed centromere–non-centromere contacts) / (total possible centromere–non-centromere contacts). Centromere interactions were defined as interactions between positions 3 Mb and 13 Mb of each chromosome.
- Nuclear speckle interactions: (observed speckle–speckle contacts) / (total possible speckle–speckle contacts) − (observed speckle–non-centromere contacts) / (total possible speckle–non-centromere contacts). Nuclear speckle interactions were defined as interactions among the following nuclear speckles regions:[7] chr2 (164–174 Mb, 177–181 Mb), chr4 (128–142 Mb, 147–155 Mb), chr5 (112–126 Mb), chr8 (123–127 Mb), chr11 (95–103 Mb, 115–121 Mb), chr13 (55–58 Mb), chr15 (76–79 Mb) and chr17 (25–30 Mb).

Nucleolar interactions: (observed nucleolar–nucleolar contacts) / (total possible nucleolar–nucleolar contacts) − (observed nucleolar–non-centromere contacts + observed non-nucleolar–non-nucleolar contacts) / (total possible nucleolar–non-centromere contacts + total possible non-nucleolar–non-nucleolar contacts). Nucleolar interactions were defined as interactions among the following nucleolar regions:[7] chr12 (5–17 Mb, 25–32 Mb), chr15 (3–6 Mb, 67–71 Mb), chr16 (5–8 Mb), chr18 (3–10 Mb, 13–24 Mb, 25–33 Mb, 39–42 Mb, 57–60 Mb) and chr19 (11–24 Mb, 25–28 Mb, 29–37 Mb, 48–53 Mb, 58–61 Mb).

Calculation of MAD scores for scSPRITE. For each single cell in scSPRITE, we calculated the number of reads in each 1-Mb bin for every chromosome genome-wide (chr1–19). Once these reads were counted, we calculated the MAD value for each cell based on the number of reads in each 1-Mb bin genome-wide to determine the variability of coverage.

Analysis of higher-order structures. Comparison of intra-chromosomal versus inter-chromosomal contacts. The percentage of intra-chromosomal and inter-chromosomal contacts for each cell was calculated from the 1,000 cells in scHi-C[16] and from the filtered 1,000 cells from scSPRITE (cluster size threshold of <10,000 reads per cluster). For scHi-C, because every cluster is a pairwise contact, we counted the number of pairwise contacts that were intra-chromosomal contacts (two contacts in the same chromosome) and inter-chromosomal contacts (two contacts coming from different chromosomes). For scSPRITE, we counted all pairwise contacts per cluster, where the number of pairwise contacts can be expressed as a binomial coefficient of '*n* choose 2', where *n* is the number of reads per cluster. From this, we then counted the number of intra-chromosomal and inter-chromosomal contacts. This was repeated for all clusters in each cell. The percentage of inter-chromosomal contacts was determined by dividing the number of inter-chromosomal contacts by the sum of the number of intra- and inter-chromosomal contacts.

Frequencies of higher-order inter-chromosomal interactions. To determine the frequency of centromeric, speckle or nucleolar interactions in single cells, we used following metrics: (1) percentage of cells that contain a given interaction in each 1-Mb bin and (2) normalized mean interaction value.

The percentage of cells containing centromere-proximal, speckle or nucleolar interactions is determined by looking through the filtered 1,000 single cells, focusing on clusters below 10,000 reads per cluster and counting the number of cells containing at least one interaction between all 1-Mb bins (*i* and *j*) in the given centromere, speckle and nucleolar regions, respectively. The genomic regions of these higher-order structures were defined previously in the section titled 'Detection scores for 3D genome structures'. To determine the expected frequency of cells that would contain these interactions by chance, we generated random genomic regions that were size matched to each feature. We generated 1,000 random permutations of each feature. For each permutation, we computed the percentage of single cells showing a contact between these random bins.

We get the normalized mean interaction value by first calculating an interaction matrix between all 1-Mb genomic bins, where the values in the interaction matrix were the percentage of cells containing an interaction between each pair of 1-Mb bins, and then calculating the mean value for pairs of regions in this interaction matrix representing centromere-proximal, speckle or nucleolar regions. For example, to determine the mean interaction value for cells containing an interaction between the centromere-proximal regions on chromosome 1 and chromosome 2, we calculated the mean value in this interaction matrix for chromosome 1 positions 3,000,000–13,000,000 with chromosome 2 positions 3,000,000–13,000,000.

Higher-order structures in scHi-C data. Ensemble and single-cell contact maps from scHi-C[16] were plotted to visualize centromere, speckle and nucleolar interactions. The single-cell barcode from scHi-C that was referenced was 'hyb_2i-1CDES-1CDES_p10.H9-adj'.

DNA FISH comparison with ensemble scSPRITE analysis. For the FISH analysis, we focused on the same chromosomal loci pairs that were originally analyzed in SPRITE. These pairs include two control chromosomal pairs and four NOR chromosomal pairs. The chromosomal loci pairs are listed below:

- Control 1: chr3 (15–16 Mb) and chr15 (4–5 Mb)
- Control 2: chr3 (15–16 Mb) and chr19 (18–19 Mb)
- NOR 1: chr12 (6–7 Mb) and chr15 (4–5 Mb)
- NOR 2: chr15 (4–5 Mb) and chr18 (3–4 Mb)
- NOR 3: chr18 (3–4 Mb) and chr19 (18–19 Mb)

We first compared contact frequency values from the loci pairs listed above between ensemble scSPRITE and SPRITE to determine how well the two methods correlated with each other. For both ensemble scSPRITE and SPRITE, we generated 1-Mb resolution, genome-wide, pairwise contact frequency maps using clusters containing fewer than 10,000 reads per cluster. These contact maps were normalized using Hi-Corrector. Using both the ensemble scSPRITE and SPRITE normalized contact frequency maps, we then pulled out the contact frequency value from each of six loci pairs, plotted their values using a scatter plot and calculated the coefficient of determination (R^2).

We generated 1-Mb resolution, genome-wide contact frequency matrices for each single cell using clusters containing fewer than 10,000 reads per cluster. For each chromosomal loci pair, a cell contained that loci pair interaction if there was at least one read in the bin containing both loci. To calculate the percentage of cells for each loci pair, we divided the number of cells containing the loci pair interaction by the total number of cells used in the analysis.

Calculation of percentage of reads coming from A/B compartments. To get the expected percentage of reads that fell into either the A or B compartments in our ensemble dataset, we used the data from the ensemble, genome-wide compartment switch analysis. We counted the number of 1-Mb bins that were classified as being in either A or B compartments genome-wide (except for chr3 and chrX). To get the expected percentage of A or B reads in our ensemble dataset, we then divided the number 1-Mb bins in A or B compartments, respectively, by the total 1-Mb bins counted.

To get the percentage of reads in A or B compartments in single cells, we looked into the genome-wide reads (with the exception of chr3 and chrX) in each single-cell file. From there, we sorted reads into A or B compartments depending on the data from the ensemble, genome-wide compartment switch analysis. Once sorted, we then divided the number of reads that fell into A or B compartments by the total number of reads counted for that cell to determine the percentage of reads in A or B compartments, respectively.

Contact maps of regions with heterogeneous structures. To identify regions of heterogeneity, we manually looked through a genome-wide heat map using the ensemble scSPRITE dataset to look for emerging TAD-like structures in between designated A/B compartments and TAD regions based on the previously identified A/B regions and TAD boundary regions, respectively. Once a region was identified in a given chromosome, 40-kb weighted, single-cell contact maps were made for that specific chromosome, where the contact frequency values in each 40-kb bin are weighted by cluster size. In the 40-kb single-cell maps, the two 40-kb bins that made up the outermost interaction of the pseudo-TAD structure in the ensemble dataset were used to look for this same interaction in the single-cell dataset (further referred to as 'bin *A*' and 'bin *B*').

Long-range interactions. Detection of heterogeneity in long-range interactions. For the interactions studied in this paper (*Phc1* at the *Nanog* locus and *Lhx5* at the *Tbx3* locus), we used the 40-kb bins containing the locations of the *Phc1* enhancer and *Nanog* promoter and the locations of the *Lhx5* gene and the *Tbx3* gene as identified previously[38]. In every single cell, we first identified cells containing a contact anywhere along bin *A* and bin *B* in that chromosome to ensure that coverage was accounted for. For *Phc1* and *Nanog*, bins *A* and *B* are chr6 122,280,000–122,320,000 bp and chr6 122,640,000–122,680,000 bp, respectively. For *Tbx3* and *Lhx5*, bins *A* and *B* are chr5 120,120,000–120,160,000 bp and chr5 120,880,000–120,920,000 bp, respectively. On average, we detect a contact in 1/3 of the total cell number, which we think is technical and due to non-sufficient coverage of every region per cell. Once the cells with coverage were identified, we identified and grouped cells in this set that contained or lacked the interaction at

the intersection of bin *A* and bin *B*. For the SE–promoter interaction at the *Nanog* locus, we identified 308 cells with read coverage, of which 159 cells contained the *Nanog–Phc1* contact and 149 cells lacked the *Nanog–Phc1* contact. For the SE–promoter interaction at the *Tbx3* locus, we identified 301 cells with read coverage, of which 152 cells contained the *Tbx3–Lhx5* contact and 149 cells lacked the *Tbx3–Lhx5* contact.

For the cells with and without an interaction over the A/B compartment boundary (Fig. 4f), a similar approach was done as described above for grouping. We first identified cells based on coverage anywhere along bin *A* and bin *B*, which was chr4 40,120,000–40,160,000 bp and chr4 40,920,000–40,960,000 bp. This region was chosen because the ensemble scSPRITE map displayed a high contact frequency at this point, which happened to be over an A/B compartment transition. Once the cells with coverage were identified, we identified and grouped cells that contained or lacked an interaction within 120 kb of bin *A* and *B* (that is, chr4 40,080,000–40,200,000 and chr4 40,880,000–41,000,000 bp). Unlike the promoter–enhancer examples where a known bin contains the promoter and enhancer loci, this information does not exist for contacts over a compartment boundary. Therefore, we provided a wider bp range to sort the cells into those two groups. Of the 379 cells identified with read coverage, 199 cells contained an interaction over the A/B compartment, and 170 cells lacked this interaction.

Virtual 4C analysis. To identify contacts with a specific locus, such as in the *Nanog* and *Tbx3* examples, we first calculated a contact frequency matrix for all pairs of genomic bins at 40-kb resolution. Using the cells that were grouped into sets either containing or lacking interactions with a specific locus, we combined each cell's individual contact frequency matrix to create an ensemble contact frequency matrix for each set. Each ensemble contact frequency matrix was normalized by Hi-Corrector[44]. To convert this contact matrix to a one-dimensional profile of contacts, we simply used the values in the row of the contact matrix corresponding to the locus of interest.

Significance and variance estimation. To determine the variance and significance of the observed contacts between these two groups, we performed a bootstrap method. Specifically, we generated random groups of cells by sampling with replacement from the initially defined groups. This approach allows us to estimate how much of the observed signal is dependent on individual cells in the population and how stable these estimates are across cells in the group. We generated 1,000 random bootstrap groups for each of the two groups and computed the average and s.d. across these permutations. To define the significance of differences between these two groups, we computed a *P* value using the unpaired two-sided *t*-test with Welch's correction between the bootstrap values in group A versus group B.

Comparison of cells with and without SE–promoter contact. We compared the number of reads and contacts from the cells containing and lacking the SE–promoter contact from the *Nanog* and *Tbx3* examples to determine if there was any bias that contributed to differences in their respective virtual 4C plots. For the *Nanog–Phc1* example, we focused our analysis on the cells that contained or lacked the *Nanog–Phc1* contact, as described previously. For each cell, we went through every cluster and calculated the number of genome-wide reads and contacts from each cluster. We then summed the number of reads and contacts from all the clusters in each cell and then repeated this process for all the cells in the two groups. We then used the Kolmogorov–Smirnov test to calculate statistical significance between the two groups. This same analysis was repeated for the *Tbx3–Lhx5* example.

Chromatin immunoprecipitation with sequencing data. We downloaded the call sets from the ENCODE portal (https://www.encodeproject.org/) with the following identifiers: H3K27ac ENCSR000CDE, H3K4me3 ENCSR000CBG and H3K27me3 ENCSR000CFN.

Cell cycle analysis. We computationally sorted the cells into M, G1, G2 or S phases of cell cycle based on the parameters described previously[16]. After categorizing the cells by phase, we calculated the percentage of cells in each corresponding cell cycle phase in the sets that contained or lacked a particular interaction.

For the SE–promoter interaction at the *Nanog* locus, 152 of the 159 cells (95.6%) containing the *Nanog–Phc1* contact and 145 of the 149 cells (97.3%) lacking the *Nanog–Phc1* contact were sorted into cell cycle phases. For the SE–promoter interaction at the *Tbx3* locus, 146 of the 152 cells (96.1%) containing the *Tbx3–Lhx5* contact and 148 of the 149 cells (99.3%) lacking the *Tbx3–Lhx5* contact were sorted into cell cycle phases. For the chr4 A/B heterogeneity example, 195 of the 199 cells (98.0%) containing the interaction of the A/B compartment boundary and 166 of the 170 cells (97.6%) lacking this interaction were sorted into cell cycle phases. The other cells were identified as 'Unknown' and were not included in the cell cycle plot.

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

Data availability

The datasets (Figs. 1–5 and Extended Data Figs. 1–5) generated and analyzed in the current study are available in the Gene Expression Omnibus repository under accession number GSE154353 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154353).

Code availability

scSPRITE software is available at https://github.com/caltech-bioinformatics-resource-center/Guttman_Ismagilov_Labs.

References

43. Engreitz, J. M. et al. RNA–RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell* **159**, 188–199 (2014).
44. Li, W., Gong, K., Li, Q., Alber, F. & Zhou, X. J. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics* **31**, 960–962 (2015).

Acknowledgements

We would like to thank F. Gao from Caltech's Bioinformatics Resource Center and I. Antoshechkin from Caltech's Millard and Muriel Jacobs Genetics and Genomics Laboratory for assistance. We would also like to thank C. Chen, V. Trinh, E. Detmar, E. Soehalim, A. Narayanan and I. Goronzy for their contributions in helping develop scSPRITE and analysis. We would like to thank M. Thompson's laboratory for allowing us to use their MiSeq instrument and the ENCODE Consortium and the ENCODE production laboratory of B. Ren (University of California, San Diego) for making their data publicly available. We also thank N. Shelby and S. Hiley for contributions to the writing and editing this manuscript and I.-M. Strazhnik for helping with illustrations. Funding: This work was funded by the National Institutes of Health 4DN Program (U01 DA040612 and U01 HL130007), the National Human Genome Research Institute Genomics of Gene Regulation Program (U01 HG007910), the New York Stem Cell Foundation (NYSCF-R-I13), the Sontag Foundation and funds from Caltech. M.V.A. and S.A.Q. were funded by a National Science Foundation Graduate Research Fellowship Program fellowship. M.V.A. was additionally funded by the Earle C. Anthony Fellowship (Caltech). M. Guttman is an NYSCF-Robertson Investigator.

Author contributions

M.V.A. conducted the experiments to develop and validate the method, conceptualized and performed the analyses and wrote the manuscript. J.W.J. contributed to and supervised the experiments to develop and validate the method, conceptualized and performed the analyses and wrote the manuscript. N.O. conceptualized and performed analysis to validate the method, developed the pipeline for the workup of scSPRITE sequencing data and contributed to writing the manuscript. M.S.C. contributed to the experiments to develop the method. C.A.L. developed a pipeline to sort cells by cell-specific barcodes. S.A.Q. contributed to the experiments to develop and validate the method. D.A.S. contributed to conceptualize scSPRITE and to the experiments to develop the method. R.F.I. conceptualized scSPRITE and supervised the experiments and the analysis to develop the method. M.G. conceptualized scSPRITE, supervised the experiments and the analysis to validate the method and wrote the manuscript. For detailed author contributions, please see Supplementary Note 5.

Competing interests

This paper is the subject of a patent application filed by Caltech. R.F.I. has a financial interest in Talis Biomedical Corp. S.A.Q. and M.G. are inventors on a patent owned by Caltech on SPRITE. The remaining authors declare no competing financial interests.

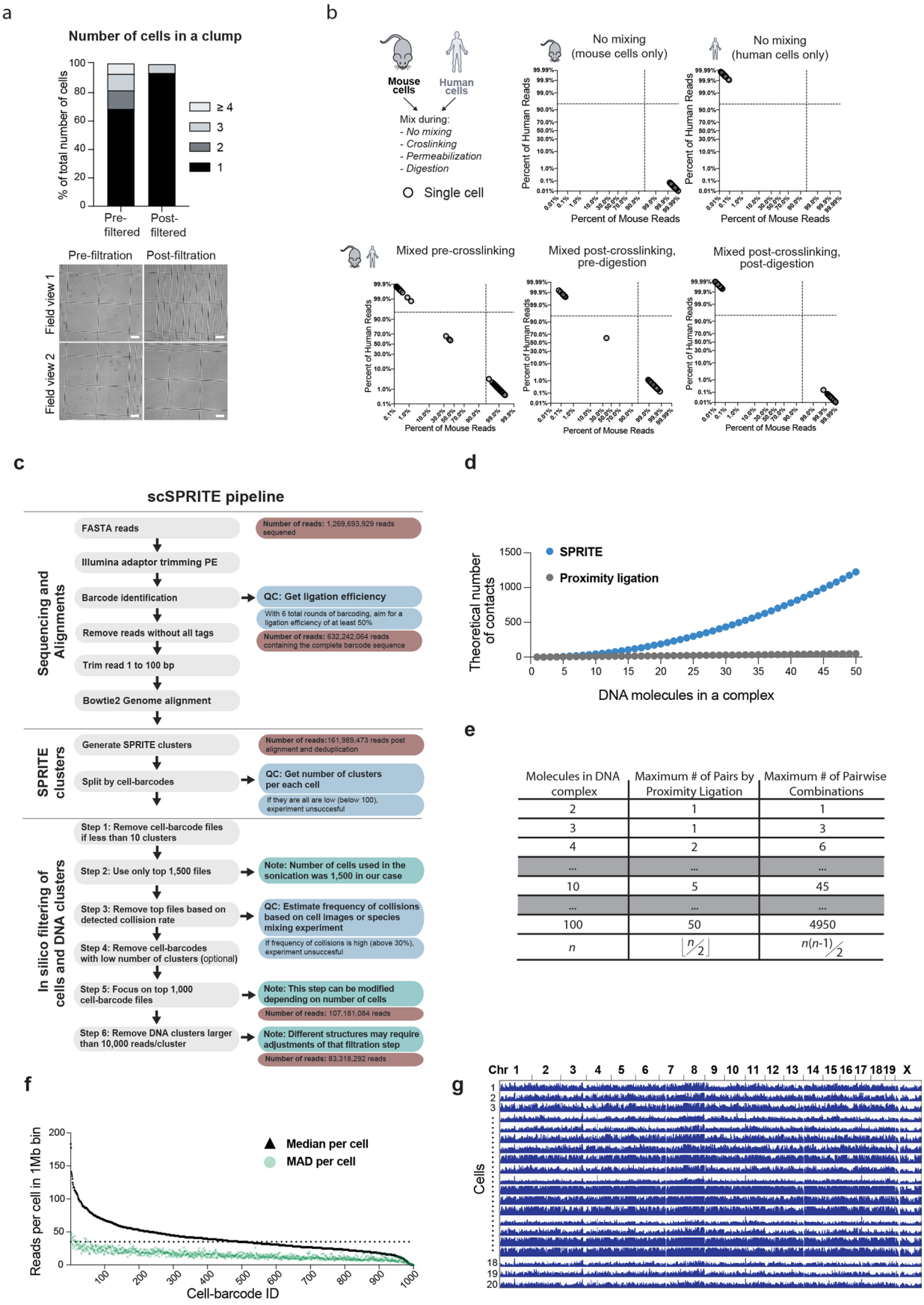
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-021-00998-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-00998-1.

Correspondence and requests for materials should be addressed to R.F.I. or M.G.

Peer review information *Nature Biotechnology* thanks Andrew Adey and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

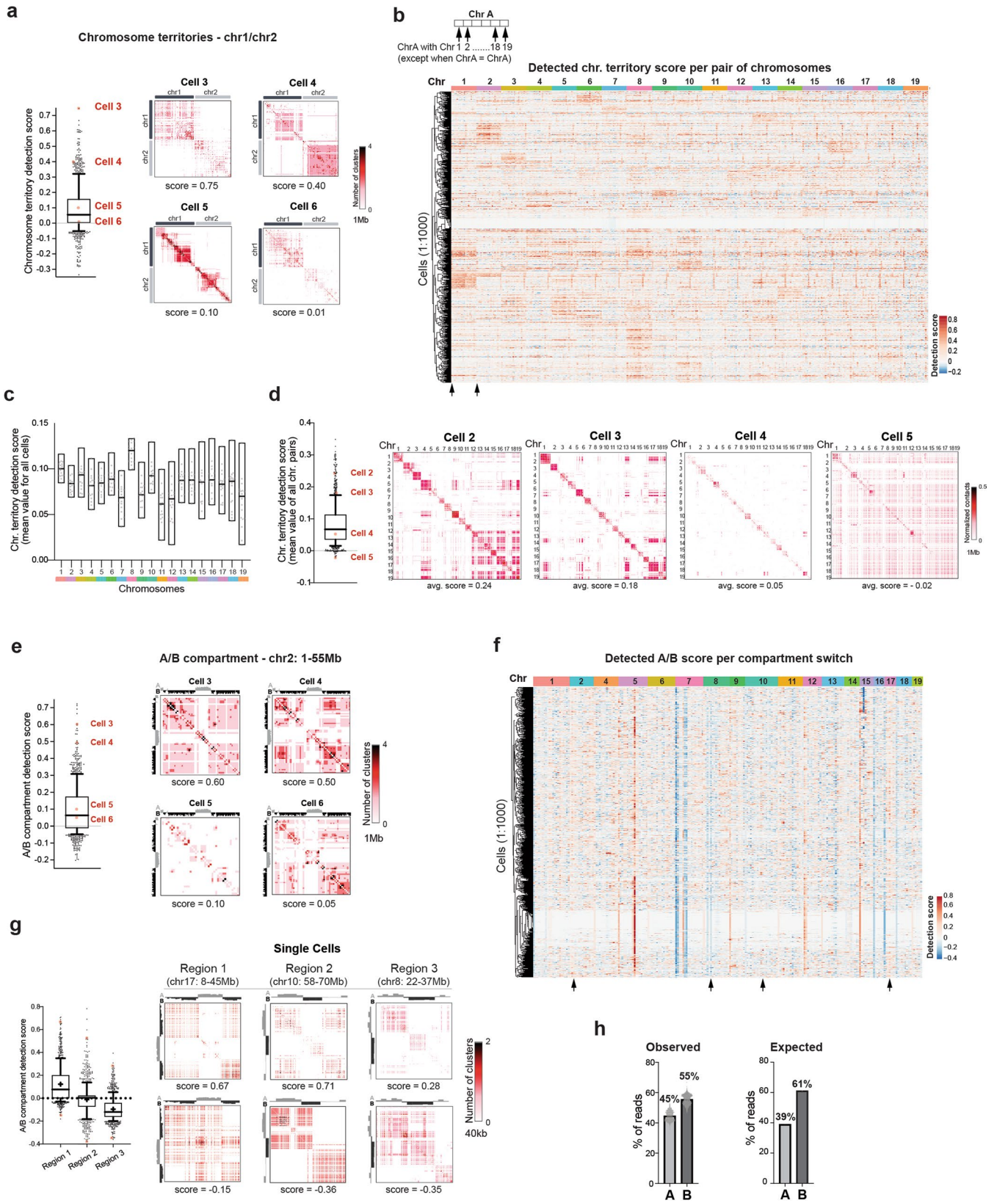


e

Molecules in DNA complex	Maximum # of Pairs by Proximity Ligation	Maximum # of Pairwise Combinations
2	1	1
3	1	3
4	2	6
...	...	...
10	5	45
...	...	...
100	50	4950
n	$\lfloor n/2 \rfloor$	$n(n-1)/2$

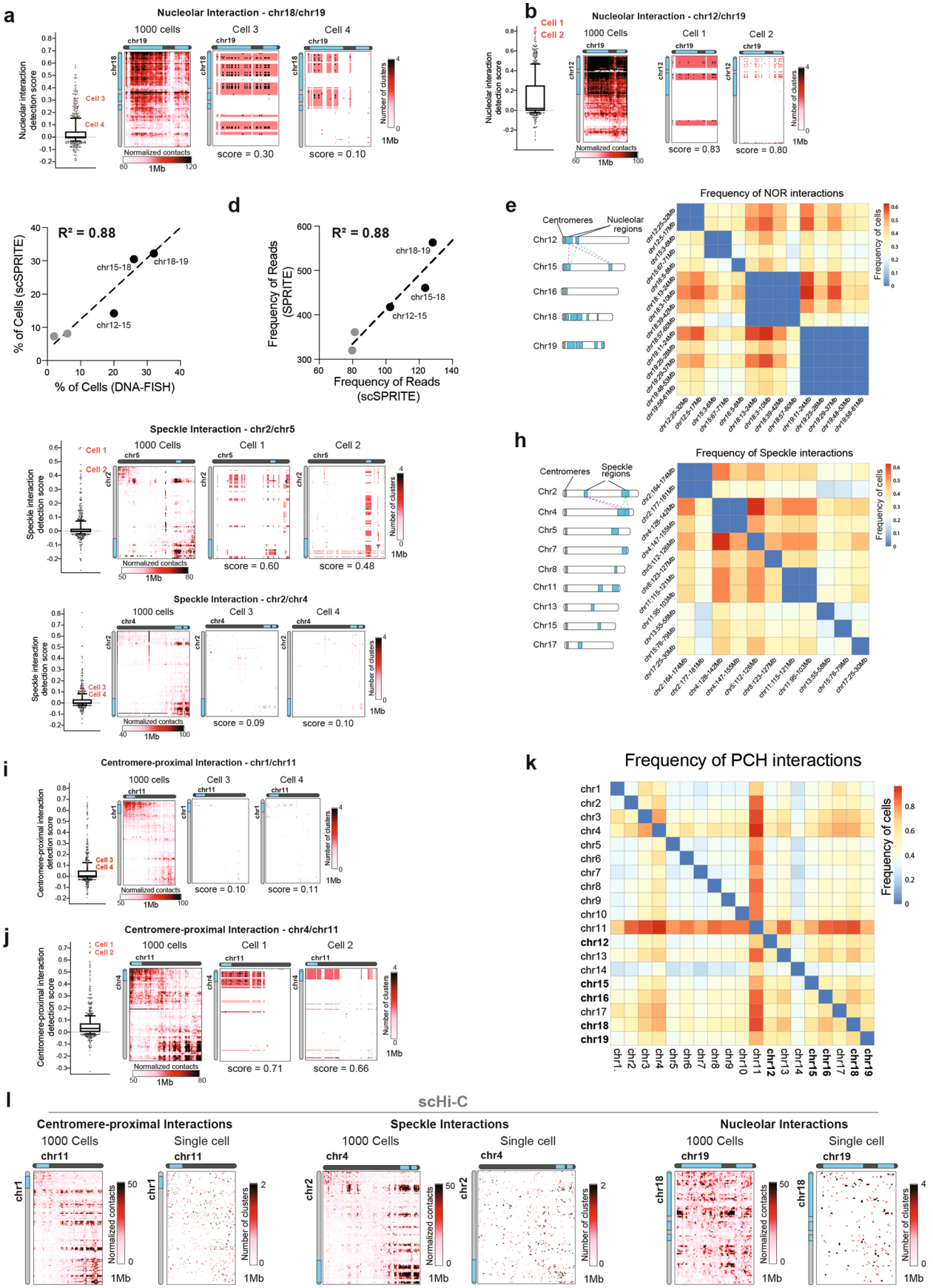
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | scSPRITE generate single cell maps with high genomic coverage. a. Quantification of cell aggregation. Top: number of cells in clumps pre- and post-filtration (singlets, doublets, triplets, etc). Bottom: microscope images (10x) of cells pre- and post-filtration step, scale bar 100 µm. **b**. Validation of In-nuclei barcoding step of the protocol on mixed cell population (human-mouse cells): no mixing (top middle and top right), mixing before crosslinking (bottom left), mixing after crosslinking (bottom middle), and mixing after in-nuclei restriction digest (bottom right). **c**. Schematic of the computational analysis pipeline for processing scSPRITE data. **d**. Theoretical number of contacts measured by SPRITE-derived methods and Hi-C-derived methods over increasing numbers of DNA molecules per complex. **e**. Maximum number of pairwise interactions that can be obtained from proximity ligation (Hi-C-derived methods) and complex barcoding (SPRITE-derived methods). **f**. Genome-wide coverage for the filtered 1,000 cells: the median (black triangular points) and median absolute deviation (MAD) (green circular points) values were calculated per cell using the number of reads per 1 Mb bin genome-wide (chr1-19). **g**. Genomic coverage of 20 random cell barcodes; 1 Mb bin per chromosome.



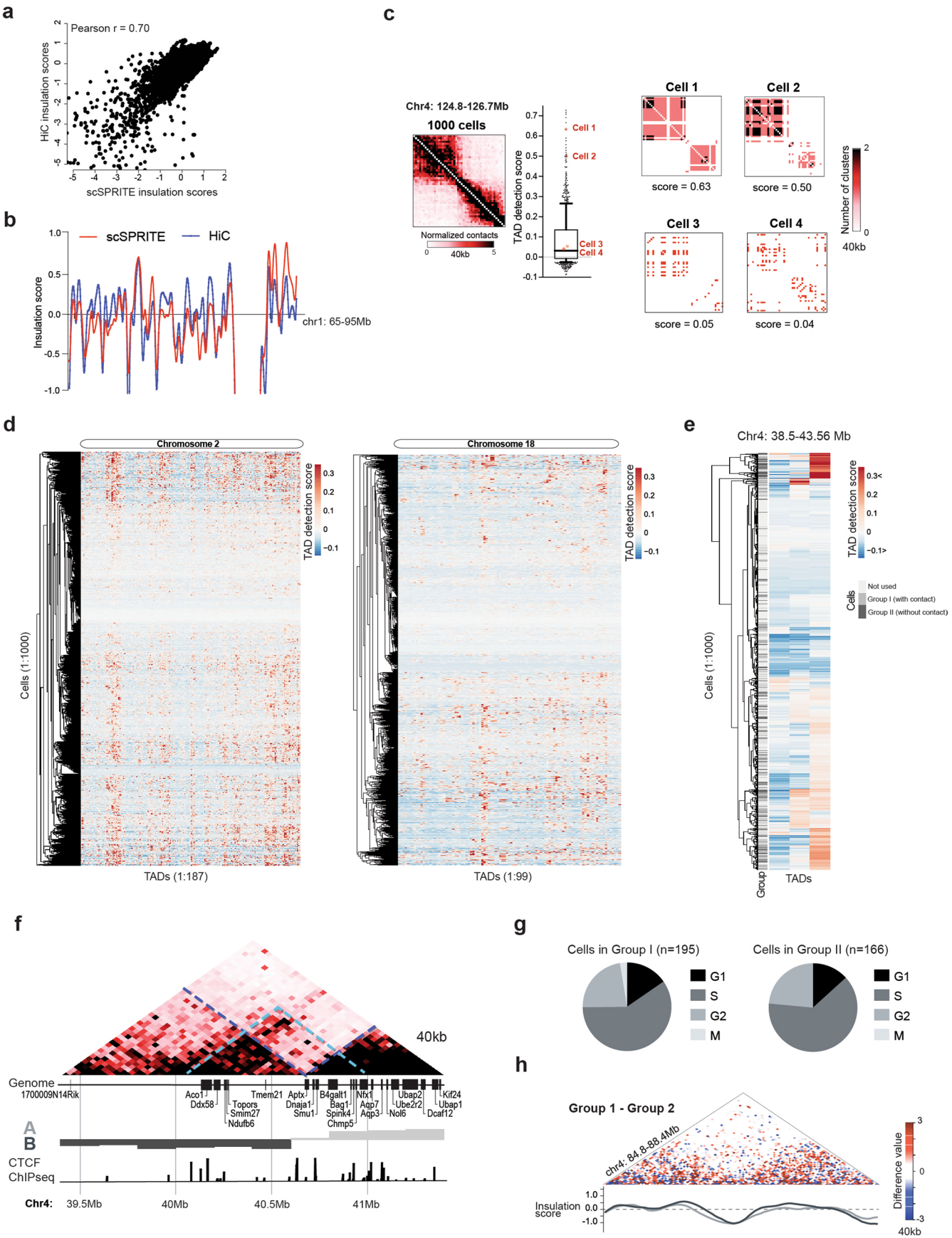
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Known chromosomal structures can be measured genome-wide in hundreds of single mESCs by scSPRITE. a. Additional single cell examples of chromosome territory structure between chr1 and chr2; plotted as number of DNA clusters at 1 Mb resolution. Box plot represents normalized detection scores between chr1 and chr2, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **b**. Chromosome territory scores across 1000 cells (clustered based on similarity pattern). Columns represent chromosome territory detection scores for all pairs of chromosomes with the reference chromosome. Arrows represent chromosome territory scores between chr1 and chr2, which were analyzed in this paper. **c**. Quantification of chromosome territory scores with respect to each chromosome. Boxplots show the range of chromosome territory scores, the average score (black line), and individual pairs of chromosome territory scores (grey dots). **d**. Box plot represents average chromosome territory detection scores from all genome-wide (chr1-19) chromosome pairs., where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells) (left).. Additional single cell examples of genome-wide (chr1-19) chromosome territories (right). **e**. Additional single cell examples of A/B compartments detected within 0-55Mb in chr2; plotted number of DNA clusters at 1 Mb resolution (right). Box plot represents normalized detection scores between 0-55Mb in chr2, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **f**. Representation of compartment switching scores across 1,000 cells (clustered based on score similarity pattern). Columns represent the strength of compartment switching detection scores for compartments that switched from "B-to-A-to-B" or "A-to-B-to-A" genome-wide (chr1-19). Arrows represent compartment switching scores for chr2 1-55 Mb, chr8 22-37 Mb, chr10 58-70 Mb, and chr17 8-45 Mb, all of which were analyzed in this paper. **g**. Additional single cell examples of compartment switching from Region 1, Region 2, and Region 3 (right). For each region's box plot: whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **h**. Expected (right) and observed (left) coverage of reads in the A and B compartment.



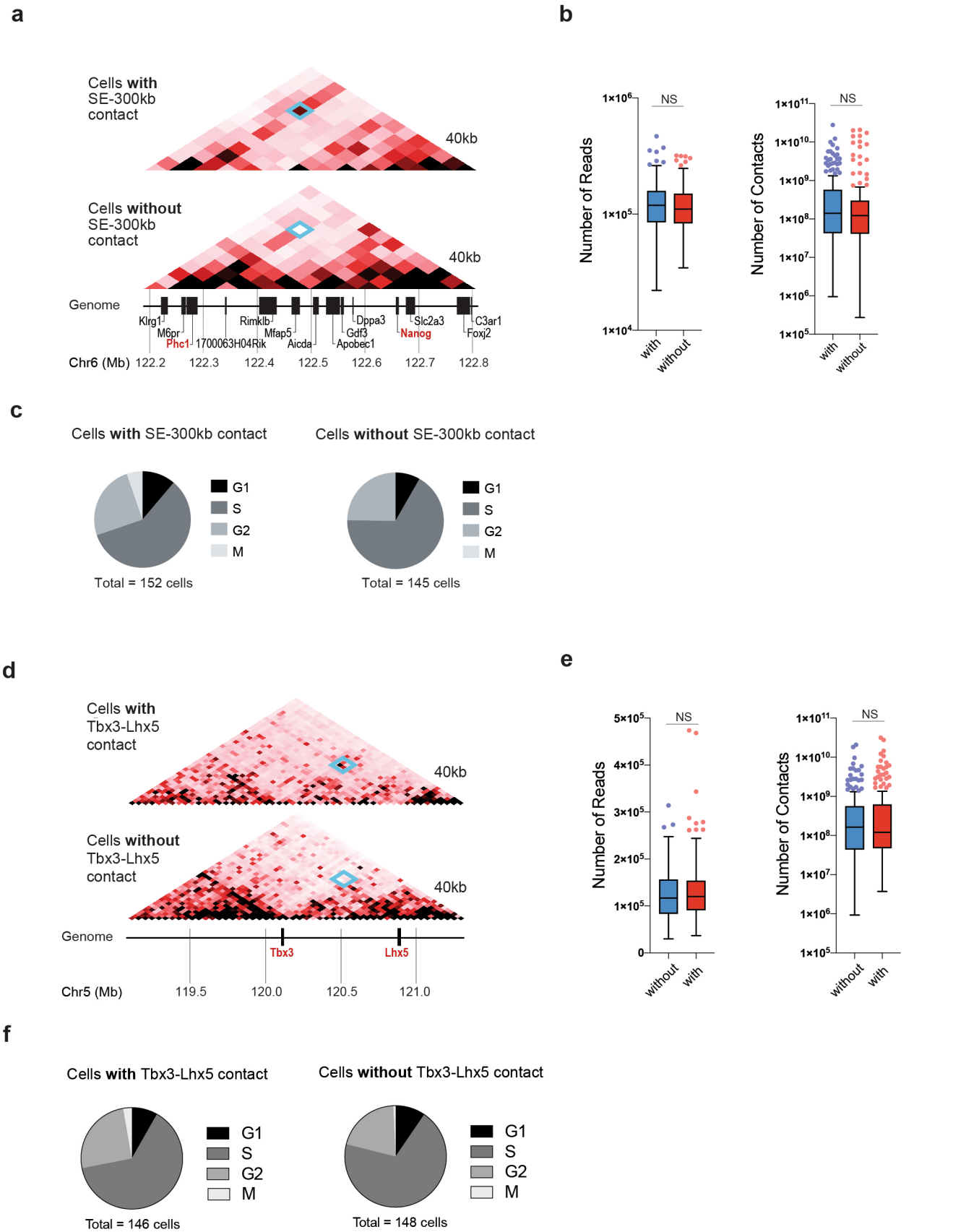
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Higher-order structures are identified genome-wide in hundreds of single mESC by scSPRITE method. a. Additional single cell examples of nucleolar interactions detected between chr18 and chr19; plotted number of DNA clusters at 1 Mb resolution; detection scores below contact map (right). Box plot represents normalized detection scores between chr18 and chr19, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **b**. Nucleolar interaction between chr12 and chr19: detection scores for 1000 cells (middle). Box plot where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). Representation of structures with max score (+1) and min. score (-1) (left) and ensemble scSPRITE heatmap (middle); contact map at 1 Mb resolution. Single cell examples (right); plotted number of DNA clusters at 1 Mb resolution. **c**. Relative correlation of the percent of cells from scSPRITE vs DNA-FISH containing inter-chromosomal interactions at specified 1 Mb regions targeted by DNA-FISH probes. Control chromosomes (grey points) and nucleolar associating chromosomes (black dots) are plotted. **d**. Relative correlation of the contact frequency from scSPRITE vs the contact frequency from SPRITE containing inter-chromosomal interactions targeted by DNA-FISH probes. Control chromosomes (grey points) and nucleolar associating chromosomes (black dots) are plotted. **e**. Frequency of cells containing inter-chromosomal nucleolar contacts (normalized to number of reads per region) for each pair of nucleolar associating chromosomes.. **f**. Single cell examples of speckle interaction detected between chr2 and chr5; plotted number of DNA clusters at 1 Mb resolution. Box plot represents normalized detection scores between chr2 and chr5, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **g**. Additional single cell examples of speckle interactions detected between chr2 and chr4; plotted number of DNA clusters at 1 Mb resolution. Box plot represents normalized detection scores between chr2 and chr4, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **h**. Frequency of cells containing inter-chromosomal speckle contacts (normalized to number of reads per region) for each pair of speckle associating chromosomes. **i**. Additional single cell examples of centromere-proximal interactions detected between chr1 and chr11; plotted number of DNA clusters at 1 Mb resolution. Box plot represents normalized detection scores between chr1 and chr11, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **j**. Single cell examples of chr4 and chr11 centromere-proximal regions interacting together; plotted number of DNA clusters at 1 Mb resolution. Box plot represents normalized detection scores between chr4 and chr11, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **k**. Frequency of cells containing inter-chromosomal centromeric contacts (normalized to number of reads per region) for each pair of chromosomes. **l**. Higher-order structures representation from scHi-C data[16] – centromere-proximal interactions, speckle interactions, and nucleolar interactions; Pairwise contact map from ensemble 1,000 cells (left), pairwise contact map from their best single cell (right).



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | TADs are heterogeneous units present in the genomes of individual mESCs. a. Genome-wide correlation of insulation scores between ensemble scSPRITE and Hi-C[3] from mouse ES cells at 40 kb resolution. **b**. Insulation score profile of ensemble scSPRITE (red) and Hi-C[3] (blue) at 40 kb resolution at chr1 65-95 Mb. **c**. Additional single cell examples of TAD-like structures between 124.8-126.7Mb of chr4; plotted number of DNA clusters at 40 kb resolution; detection scores below contact map. Box plot represents normalized detection scores between 124.8-126.7Mb of chr4, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **d**. TAD detection scores across 1,000 cells (clustered based on score similarity pattern) in chr2 (left) and chr18 (right). Columns represent the strength of TAD detection scores for all TADs detected across chr2 or chr18, respectively, in ensemble scSPRITE. **e**. TAD detection scores across 1,000 cells between 38.5-48.56 Mb of chr4. Each line represents the strength of TAD detection scores in this given region from a single cell. Cells are either in Group 1 or 2 in Fig. 4f or not used. **f**. Ensemble heatmap from all 1000 cells between 39.4-41.4Mb of chr4 representing strong TADs detected in bulk (blue lines), and weak emerging TADs (green line) over the A/B boundary. **g**. Fraction of cells in each cell cycle phase from the set of single cells containing (left) or lacking (right) the contact between the boundary region (Fig. 4f). **h**. Difference contact map across a control region 84.8-88.4 Mb of chr4 made by subtracting the normalized contacts from cells in Group II from Group I (Fig. 4f). Insulation scores for cells in Group I (dark grey) and Group II (light grey) are plotted.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Structural heterogeneity in long-range interactions is revealed by scSPRITE. a. Ensemble heatmaps across 122.2-122.8 Mb region in chr6 representing cells containing (top) or lacking (bottom) the contact between the *Nanog* locus and the -300 Kb SE. Blue square shows the contact. **b**. Number of genome-wide reads (left) and number of genome-wide contacts (right) for groups of cells with and without the *Nanog*-SE interaction. For each box plot, whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median (with = 159 cells, without = 149 cells). No statistical significance between the two groups were seen based on the Kolmogorov–Smirnov two-sided test. **c**. Fraction of cells in each cell cycle phase from the set of single cells containing (left) or lacking (right) the contact between the *Nanog* locus and the SE 300kb upstream of *Nanog*. **d**. Heatmaps between 119.24-121.28Mb in chr5 of pooled cells either containing (top) or lacking (bottom) the contact between the *Tbx3* locus and *Lhx5*. Blue square shows the contact. **e**. Number of genome-wide reads (left) and number of genome-wide contacts (right) for groups of cells with and without the *Tbx3*-*Lhx5* interaction. For each box plot, whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median (with = 152 cells, without = 149 cells). No statistical significance between the two groups were seen based on the Kolmogorov–Smirnov two-sided test. **f**. Fraction of cells in each cell cycle phase from the set of single cells containing (left) or lacking (right) the contact between the *Tbx3* locus and the *Lhx5*.

nature research

Corresponding author(s): Mitchell Guttman and Rustem F. Ismagilov

Last updated by author(s): Jun 11, 2021

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.

n/a	Confirmed	
☐	☒	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	cworld (https://github.com/dekkerlab/cworld-dekker); Snakemake pipeline for sequencing data (https://github.com/caltech-bioinformatics-resource-center/Guttman_Ismagilov_Labs)

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

GEO Accession # GSE154353 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154353). Secure token for viewing the GEO Accession is "grknyoyavvozdyh". Github link: https://github.com/caltech-bioinformatics-resource-center/Guttman_Ismagilov_Labs.

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	Around 1500 nuclei were originally loaded for in-nuclei cell-specific barcoding and subsequent sequencing. During in-silico analysis, to ensure we selected single cells for downstream analysis, the top 3.4% percent of cells, as determined from the results of the human-mouse mixing experiment, were removed. From the remaining files, to ensure that the analysis containing the highest coverage per cell were analyzed, we then selected the top 1000 cell barcode files containing the most number of clusters per cell for downstream single-cell analysis.
Data exclusions	All data, including any data removed during in-silico analyses, are being made available.
Replication	Because single cell measurements are done, each cell is a replicate.
Randomization	None - No treatment groups were present such that randomization is necessary.
Blinding	None - No treatment groups were present such that blinding is necessary.

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

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	The male bsps mouse embryonic stem cell line was provided by K. Plath. HEK293T, a female human embryonic kidney cell line transformed with the SV40 large T antigen was obtained from ATCC (#CRL-1573).
Authentication	Not applicable for cell line authentication
Mycoplasma contamination	Cells were routinely tested for mycoplasma using a kit from Millipore-Sigma (MP0025-1KT)
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines