nature biotechnology

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41587-020-0519-y

In the format provided by the authors and unedited.

GPSeq reveals the radial organization of chromatin in the cell nucleus

Gabriele Girelli[1,2,4], Joaquin Custodio[1,2,4], Tomasz Kallas[1,2,4], Federico Agostini[1,2], Erik Wernersson[1,2], Bastiaan Spanjaard[3], Ana Mota[1,2], Solrun Kolbeinsdottir[1,2], Eleni Gelali[1,2], Nicola Crosetto[1,2,5] ✉ and Magda Bienko[1,2,5] ✉

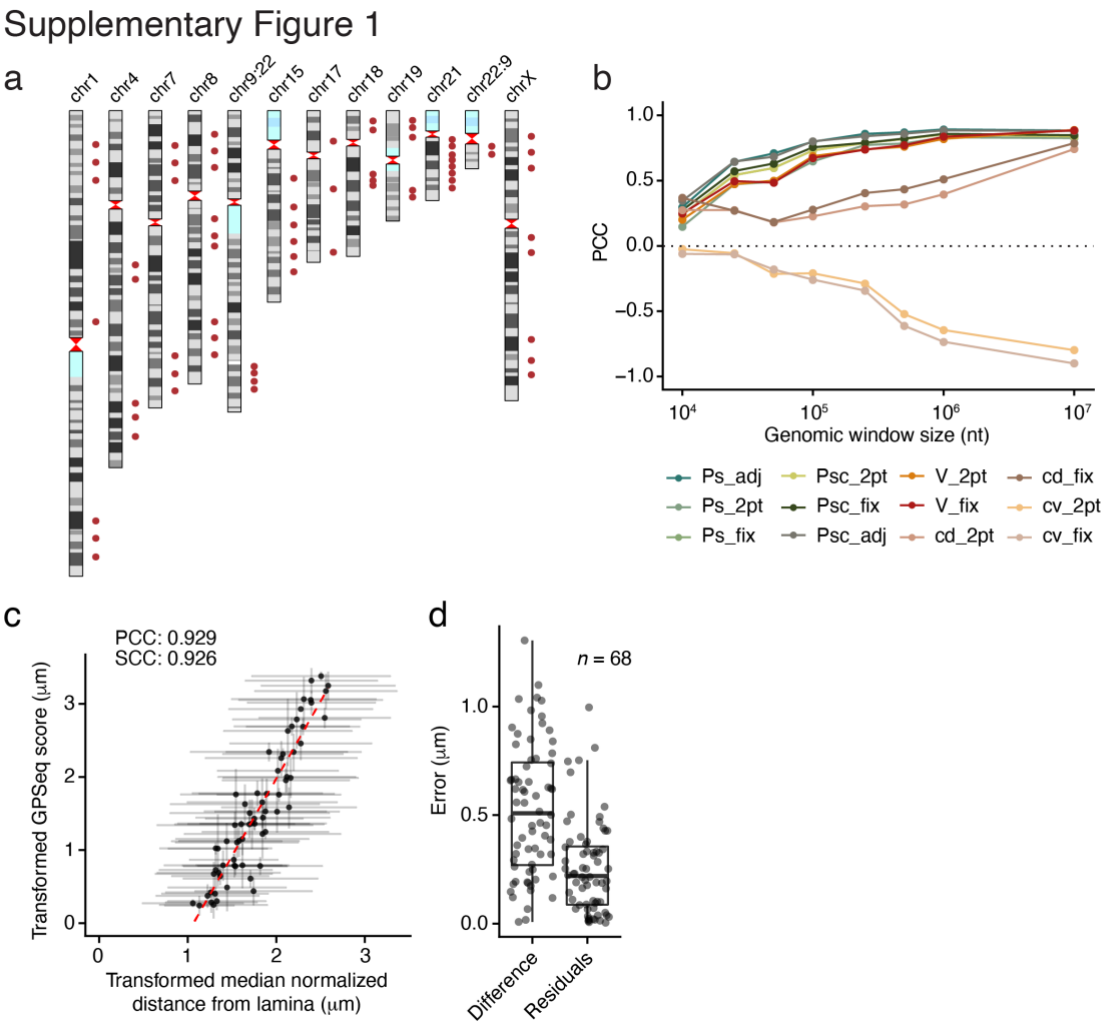
[1]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. [2]Science for Life Laboratory, Stockholm, Sweden. [3]Berlin Institute of Medical Systems Biology Max Delbrück Center, Berlin, Germany. [4]These authors contributed equally: Gabriele Girelli, Joaquin Custodio, Tomasz Kallas. [5]These authors jointly supervised this work: Nicola Crosetto, Magda Bienko. ✉e-mail: nicola.crosetto@ki.se; magda.bienko@ki.se

SUPPLEMENTARY INFORMATION

GPSeq reveals the radial organization of chromatin in the cell nucleus

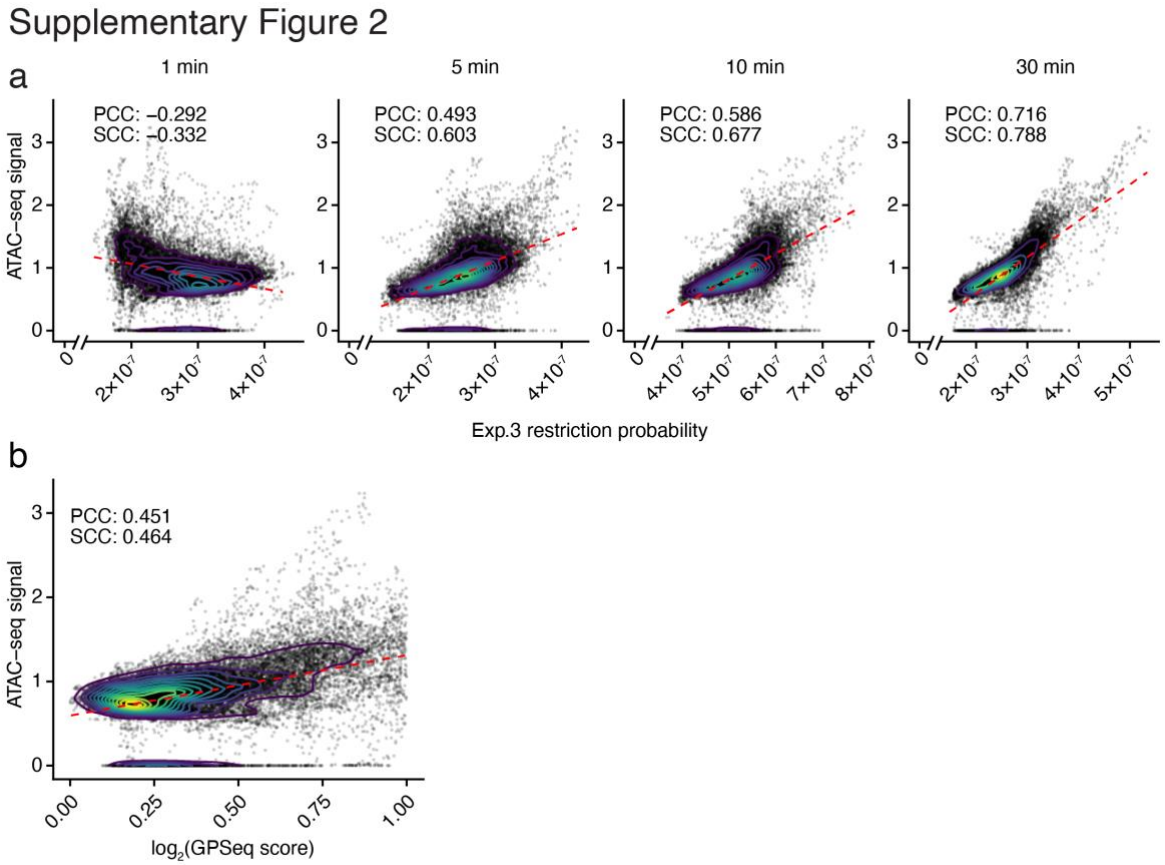
Gabriele Girelli, Joaquin Custodio, Tomasz Kallas, Federico Agostini, Erik Wernersson, Bastiaan Spanjaard, Ana Mota, Solrun Kolbeinsdottir, Eleni Gelali, Nicola Crosetto & Magda Bienko

1. Supplementary Figures



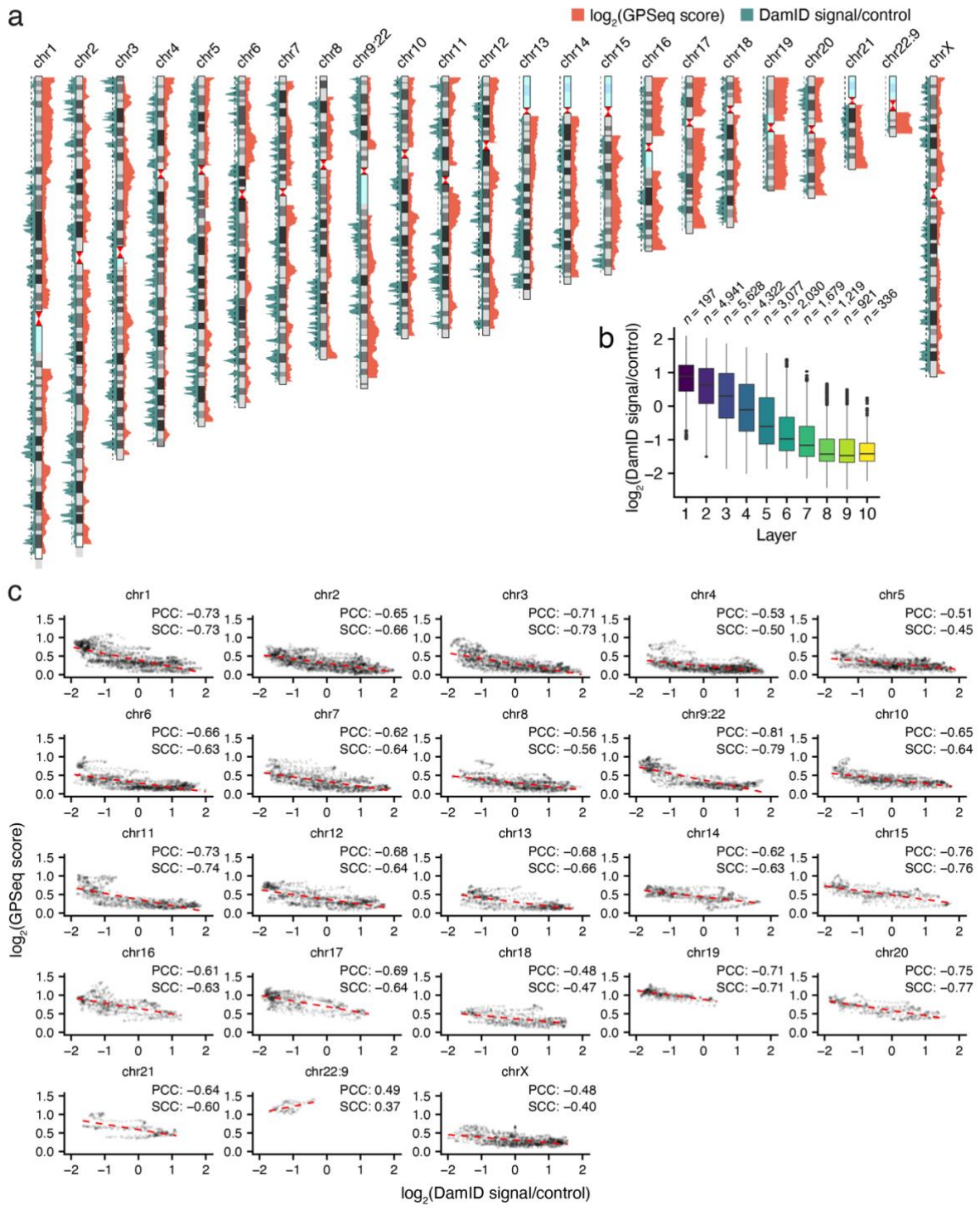
Supplementary Figure 1. Validation of GPSeq by DNA FISH. (**a**) Genomic location of the 68 DNA FISH probes (red dots) used to validate GPSeq results. In the chromosome ideograms, the color of the cytobands is based on the intensity of the Giemsa staining, peri-centromeric regions are colored in red, and acrocentric regions and variable heterochromatic regions are colored in cyan. chr9:22 and chr22:9 are the derivative chromosomes of the t(9;22)(q34;q11.2) translocation. (**b**) Pearson's correlation coefficient (PCC) between different radiality estimates, at different resolution (genomic window size), and the median 3D distance from the nuclear lamina as measured by DNA FISH, for each of the 68 probes shown in (a) at various resolutions. The estimate abbreviations indicate the estimate name (corresponding to the subscripts in the column 'Name' in **Supplementary Note 1 Table 1**) followed by the name of the approach used to combine different time points (adj: adjacent approach; 2pt: two-points approach; fix: fixed approach) (see **Supplementary Note 1**). (**c**) Correlation between the GPSeq score transformed

to absolute distance and the normalized median distance to the nuclear lamina measured by DNA FISH and converted to absolute distance (see **Supplementary Methods**). Horizontal error bars: ± s.d. of transformed distance to nuclear lamina across cells. Vertical error bars: ± s.d. of the transformed GPSeq score averaged across the four GPSeq experiments described in **Supplementary Table 2**. Each dot represents one of the 68 probes shown in (a). The number of cells analyzed for each probe and used to calculate the horizontal error bars is available in **Supplementary Table 11**. Dashed red line: linear regression. (**d**) Distribution of the transformed GPSeq score error quantified either as the absolute difference between the two axes in (c) (left boxplot) or as the residuals from the regression line in (c) (right boxplot). *n*, number of FISH probes used for the analysis. In the boxplots, each box extends from the 25th to the 75th percentile, the midline represents the median, and the whiskers extend from –1.5×IQR to +1.5×IQR from the closest quartile, where IQR is the inter-quartile range. Dots: outliers (data falling outside whiskers). All the source data for this figure are from HAP1 cells.



Supplementary Figure 2. Comparison of GPSeq with ATAC-seq. (**a**) Correlation between the restriction probability (see Eq. (7) in **Supplementary Note 1**) calculated for four restriction times (conditions) in one MboI experiment (Exp.3), and the ATAC-seq signal at 1 Mb resolution (overlapping genomic windows, 100 kb step size). (**b**) Correlation between the log2 of the GPSeq score (averaged across Exp.1–4, see **Supplementary Table 2**), and the ATAC-seq signal at 1 Mb resolution (overlapping genomic windows, 100 kb step size). In all the plots, PCC and SCC are the Pearson's and Spearman's correlation coefficient, respectively. *n* = 26,350 genomic windows (points) were analyzed in each plot. Density contours are shown as concentric curves. All the source data for this figure are from HAP1 cells.

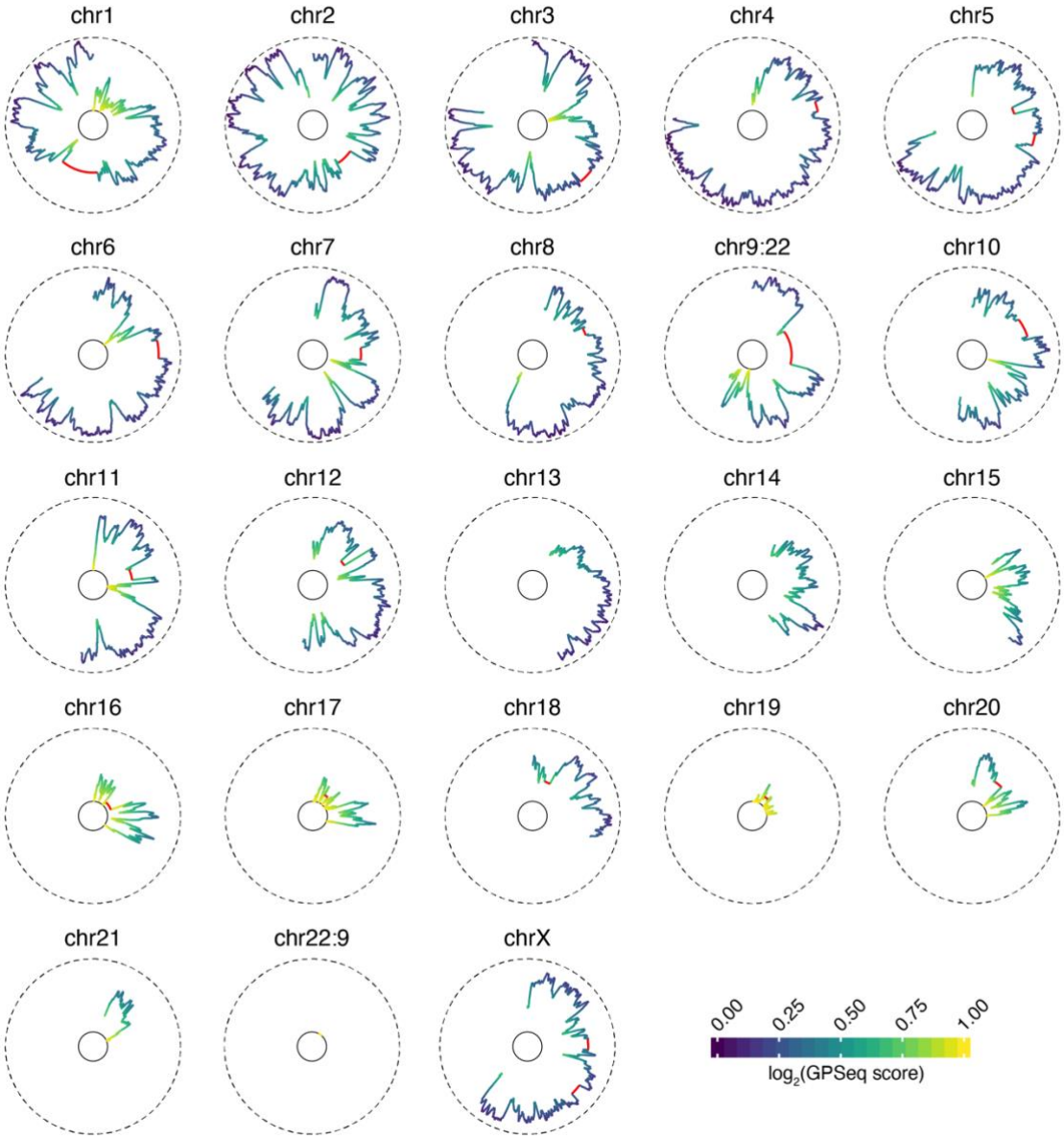
Supplementary Figure 3



Supplementary Figure 3. Comparison of GPSeq with Lamin B DamID. (**a**) Chromosome-wide profiles of the log2 of the GPSeq score (orange) and of the Lamin B DamID signal over control (green) (1 Mb overlapping genomic windows, 100 kb step size). Dashed line: DamID signal over control equal to 1. In the chromosome ideograms, the color of the cytobands is based on the intensity of the Giemsa staining, peri-centromeric regions are colored in red, and

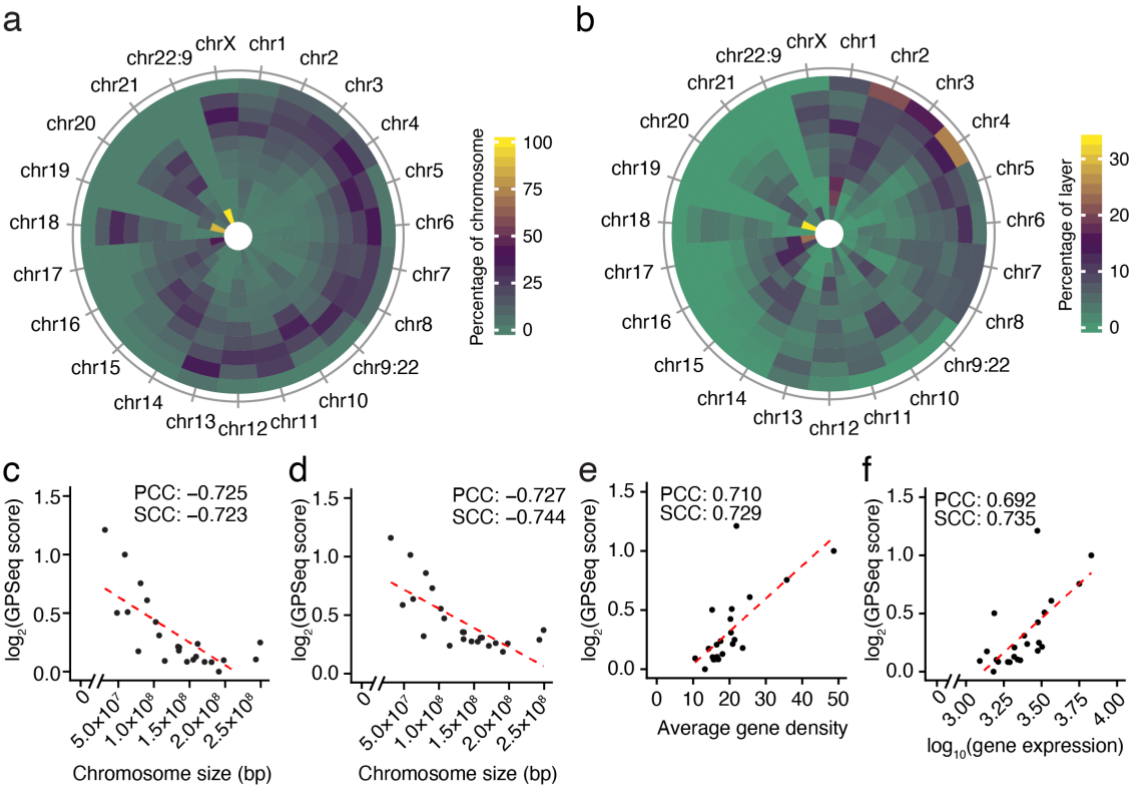
acrocentric regions and variable heterochromatic regions are colored in cyan. (**b**) Distribution of the Lamin B DamID signal over control (1 Mb resolution) in ten nuclear layers defined based on the GPSeq score as described in the **Supplementary Methods**. *n*, number of genomic windows analyzed. In all the boxplots, each box extends from the 25th to the 75th percentile, the midline represents the median, and the whiskers extend from –1.5×IQR to +1.5×IQR from the closest quartile, where IQR is the inter-quartile range. Dots: outliers (data falling outside whiskers). (**c**) Correlation between the log2 of the GPSeq score and Lamin B DamID signal over control (1 Mb overlapping genomic windows, 100 kb step size), separately for each chromosome. In all the scatterplots, PCC and SCC are the Pearson's and Spearman's correlation coefficient, respectively. chr9:22 and chr22:9 are the derivative chromosomes of the t(9;22)(q34;q11.2) translocation. Dashed red lines: linear regression. Sample size information for (c) is available in **Supplementary Table 11**. All the source data for this figure are from HAP1 cells.

Supplementary Figure 4



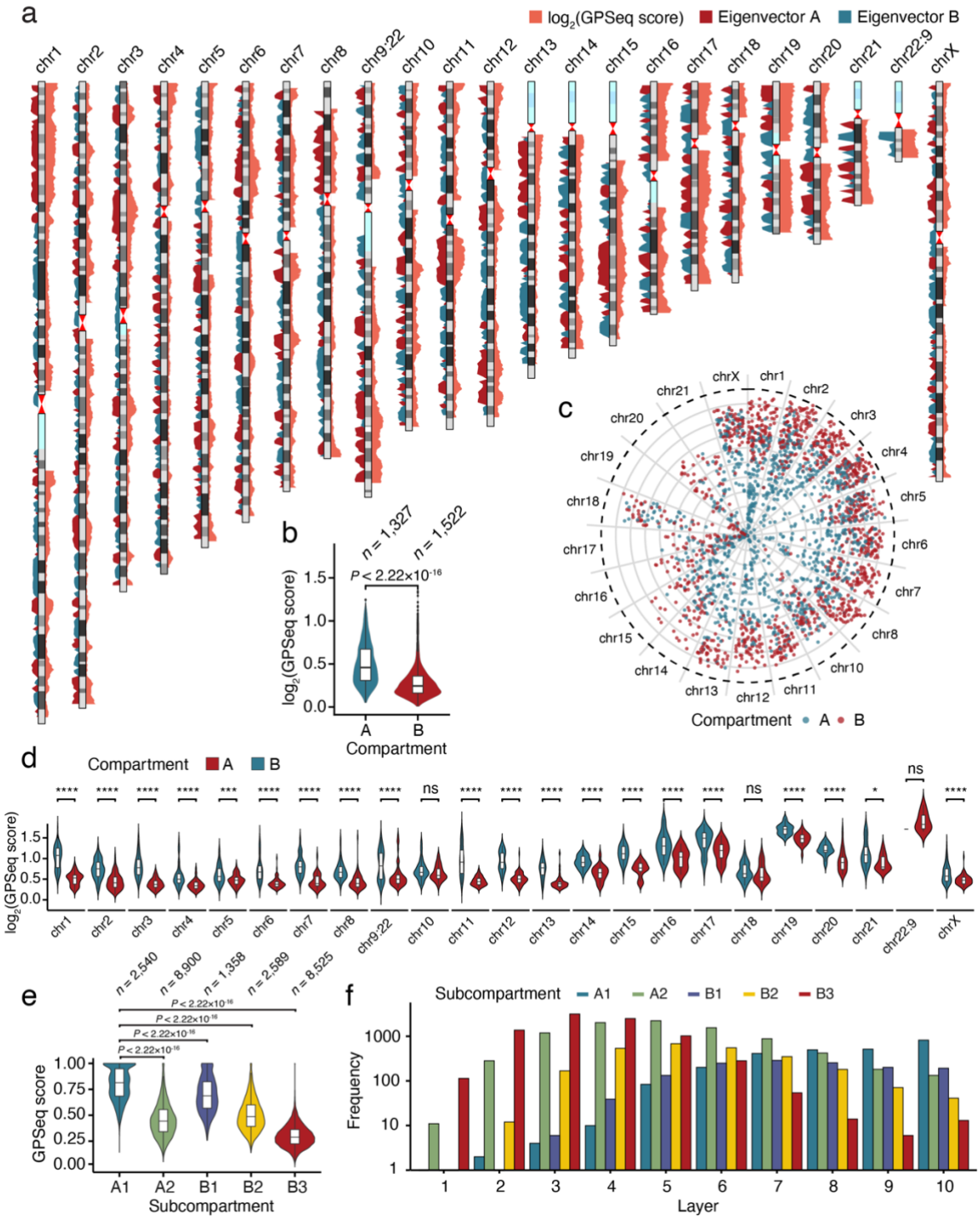
Supplementary Figure 4. Radial arrangement of chromosomes. Circular plots showing the radial position of 1 Mb consecutive genomic windows along all the human autosomes and chrX. chr9:22 and chr22:9 are the derivative chromosomes of the t(9;22)(q34;q11.2) translocation. Dashed circles: nuclear lamina. Solid circles: nuclear center. Red segments represent repetitive sequences that were masked out before calculating the GPSeq score (see **Supplementary Table 10**). Sample size information for each plot is available in **Supplementary Table 11**. All the source data for this figure are from HAP1 cells.

Supplementary Figure 5



Supplementary Figure 5. Predictors of radial organization. (**a**) Tiled 'pizza-plot' showing the distribution of each chromosome in ten concentric nuclear layers. All the tiles belonging to the same slice (chromosome) sum up to 100% and the tiles are colored based on the fraction (%) of each chromosome in a given layer. (**b**) Tiled 'pizza-plot' showing the frequency of each chromosome in ten concentric nuclear layers. All the tiles in the same layer sum up to 100%, and the tiles are colored based on the fraction (%) of each layer occupied by a different chromosome. In both (a) and (b), the center of the pizza-plot corresponds to the nuclear interior. chr9:22 and chr22:9 are the derivative chromosomes of the t(9;22)(q34;q11.2) translocation. (**c**) Correlation between the log2 of the GPSeq score (calculated per chromosome-wide windows) and chromosome size in base-pairs (bp). (**d**) Correlation between the log2 of the GPSeq score (calculated per 1 Mb overlapping genomic windows with 100 kb steps and averaged per chromosome) and chromosome size in base-pairs (bp). (**e**) Correlation between the log2 of the GPSeq score (chromosome resolution) and the mean number of transcription start sites (TSS, one TSS per gene) per 1 Mb per chromosome (gene density). (**f**) Correlation between the log2 of the GPSeq score (chromosome resolution) and the average RNA-seq read count per 1 Mb per chromosome. In (c-f), each dot represents one chromosome. All the source data for this figure are from HAP1 cells.

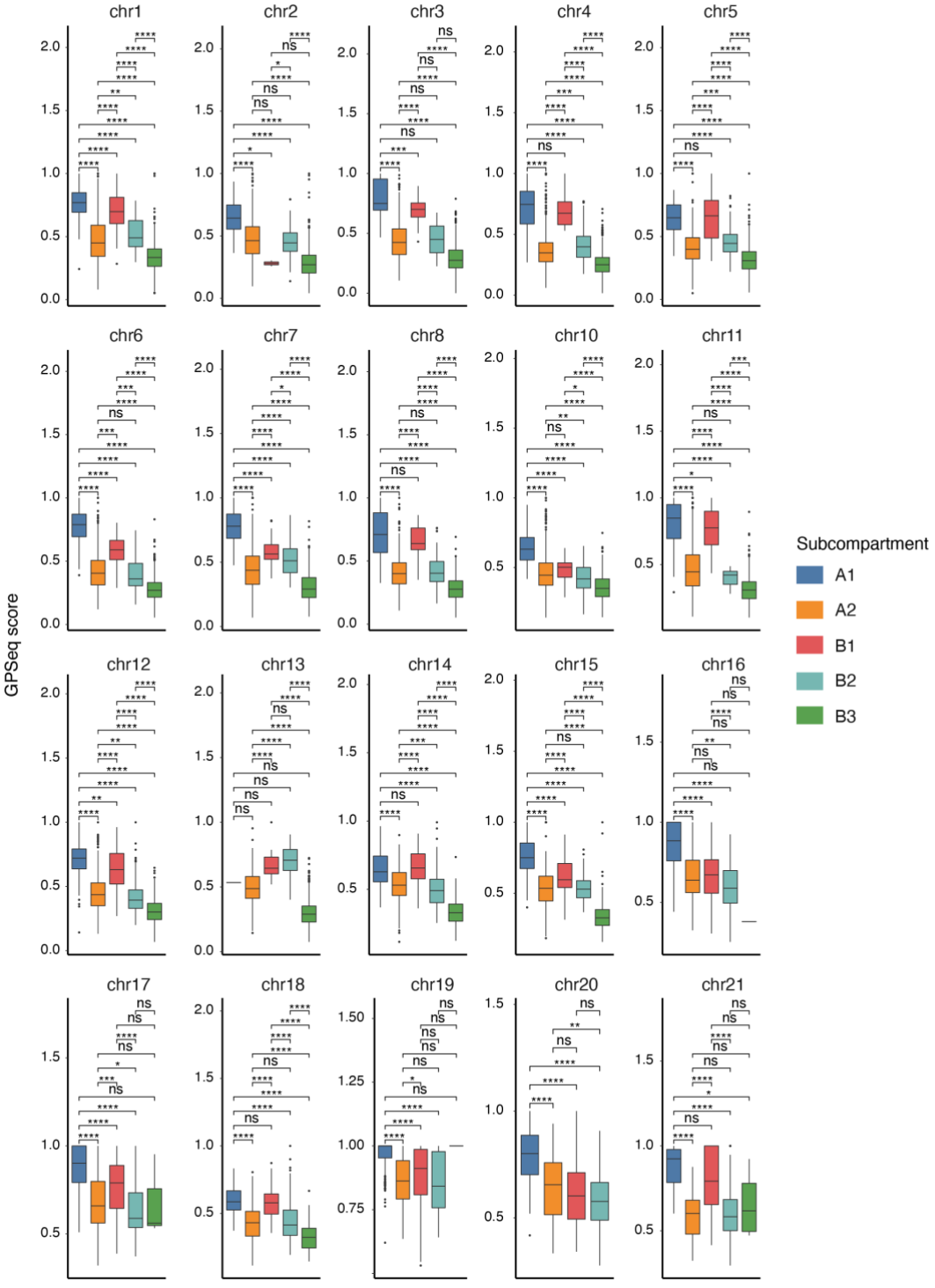
Supplementary Figure 6



Supplementary Figure 6. Radial organization of A/B compartments and subcompartments. (**a**) Chromosome-wide profiles of the log2 of the GPSeq score (orange) and of the absolute Hi-C Eigenvector (1 Mb overlapping genomic windows, 100 kb step size). A/B compartments defined based on the Eigenvector values are shown in two different colors. Eigenvector A and Eigenvector B show the absolute values of the Eigenvector values for A and B-compartment

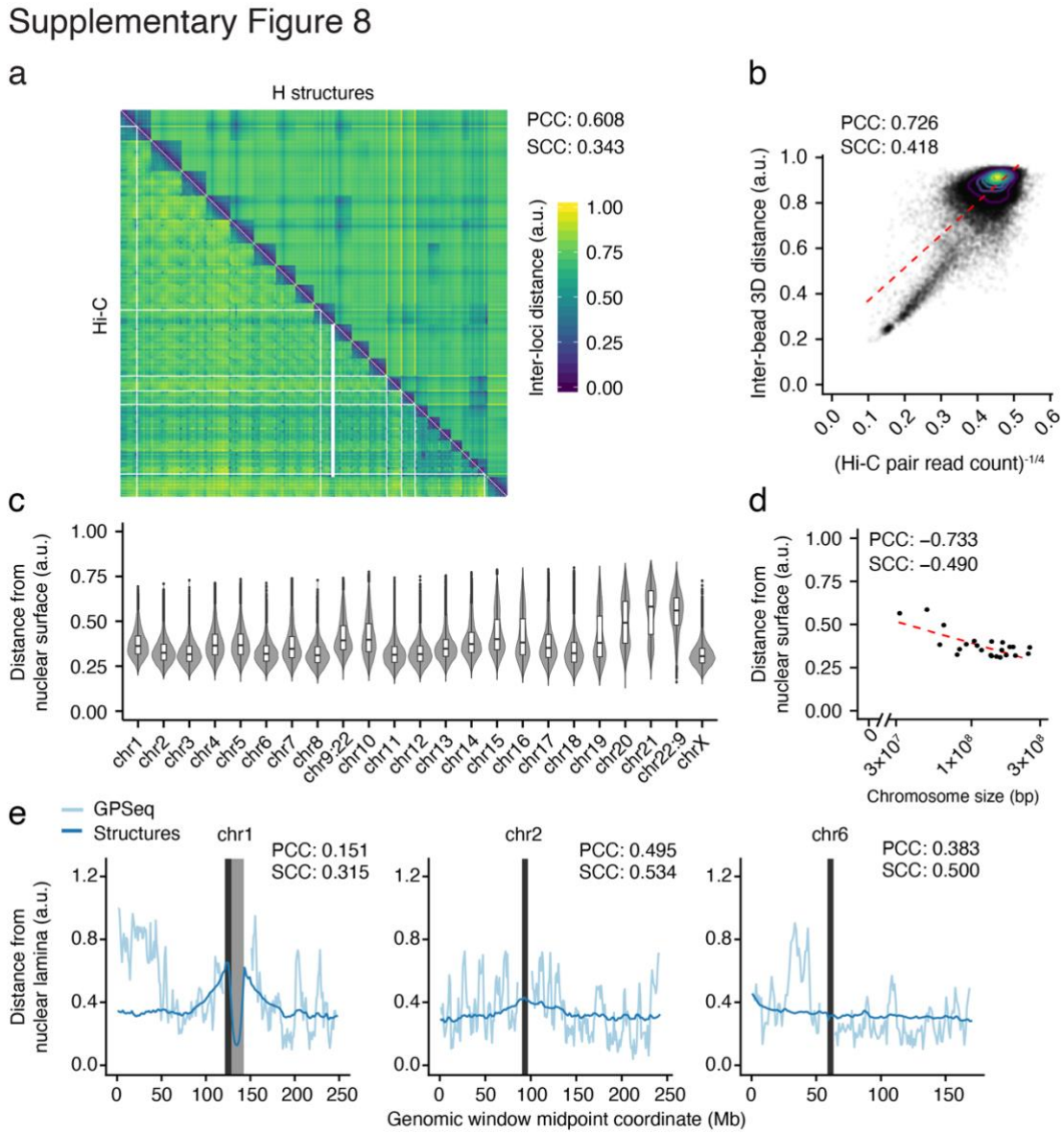
regions, respectively. In the chromosome ideograms, the color of the cytobands is based on the intensity of the Giemsa staining, peri-centromeric regions are colored in red, and acrocentric regions and variable heterochromatic regions are colored in cyan. chr9:22 and chr22:9 are the derivative chromosomes of the t(9;22)(q34;q11.2) translocation. (**b**) Distributions of the log2 of the GPSeq scores of non-overlapping 1 Mb windows belonging to either A or B compartment. *P*-value: Wilcoxon test, two-sided. *n*, number of genomic windows analyzed. (**c**) 1 Mb non-overlapping genomic windows (dots) radially arranged based on their GPSeq score, and colored based on the compartment type. Dashed line: nuclear lamina. Continuous grey lines: concentric nuclear layers. (**d**) Distribution of the log2 of the GPSeq score calculated in non-overlapping 1 Mb windows, separately for the A and B compartments of individual chromosomes. Asterisks indicate the *P*-value (Wilcoxon test, two-sided) of the comparison for each chromosome. Asterisks: * $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$; **** $P \leq 0.0001$. For exact *P* and *n* values see **Supplementary Table 11**. (**e**) Distributions of the GPSeq scores of non-overlapping 100 kb windows belonging to different A/B subcompartments. *P*-values: Wilcoxon test, two-sided. *n*, number of genomic windows analyzed. (**f**) Number of 100 kb genomic windows belonging to different A/B subcompartments across ten concentric nuclear layers. In all the violin plots in the figure, each box spans from the 25th to the 75th percentile and whiskers extend from –1.5×IQR to +1.5×IQR from the closest quartile, where IQR is the inter-quartile range. Dots: outliers (data falling outside whiskers). All the source data for this figure are from HAP1 cells.

Supplementary Figure 7



Supplementary Figure 7. Radial organization of A/B subcompartments along individual chromosomes. For each chromosome, the distributions of the GPSeq scores calculated in non-overlapping 100 kb windows are shown separately for each subcompartment. Asterisks indicate the *P*-value (Wilcoxon test, two-sided) of the corresponding comparison between two different subcompartments (ns: $P > 0.05$; * $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$; **** $P \leq 0.0001$). For

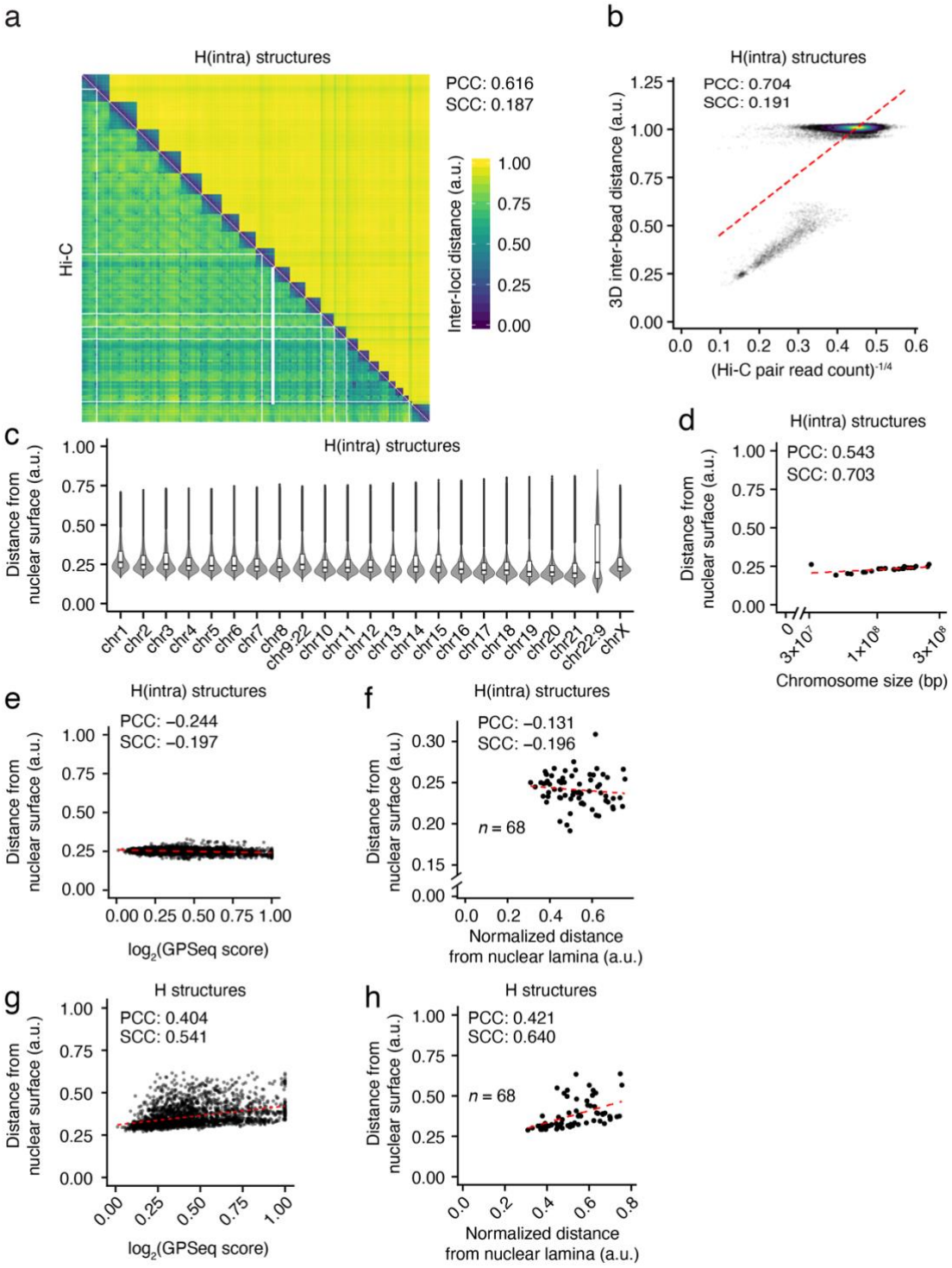
exact *P* and *n* values see **Supplementary Table 11**. In all the boxplots, boxes span from the 25th to the 75th percentile and whiskers extend from –1.5×IQR to +1.5×IQR from the closest quartile, where IQR is the inter-quartile range. Dots: outliers (data falling outside whiskers). All the source data for this figure are from HAP1 cells.



Supplementary Figure 8. Analysis of *chromflock* structures generated using only Hi-C data (all contacts) (H structures). (**a**) Distance matrix heatmap. The upper triangle shows the inter-bead 3D distances calculated from 10,000 H structures. The bottom triangle shows the KR-normalized Hi-C contact frequency matrix, with each element raised to the power of –0.25. The reported correlation coefficients are for 1 Mb resolution, while for simplicity the plot shows averaged values over 10 Mb genomic windows (points). (**b**) Correlation between the average inter-bead 3D distance in H structures and the KR-normalized Hi-C contact frequency, with each element raised to the power of –0.25. Each dot represents a pair of 10 Mb non-overlapping genomic windows, each obtained by averaging 1 Mb non-overlapping bins. n = 47,531 genomic window pairs (points) were analyzed. Density contours are shown as concentric curves. (**c**)

Distribution of the average distances from the modeled nuclear surface of 1 Mb beads in H structures, separately for each chromosome. In all the boxplots inside the violin plots, each box spans from the 25th to the 75th percentile, whiskers extend from –1.5×IQR to +1.5×IQR from the closest quartile, where IQR is the inter-quartile range. Dots: outliers (data falling outside whiskers). chr9:22 and chr22:9 are the derivative chromosomes of the t(9;22)(q34;q11.2) translocation. (**d**) Correlation between the average chromosome distance from the modeled nuclear surface in H structures and chromosome size in base-pairs (bp). Each dot corresponds to one chromosome. (**e**) Examples of radiality profiles along chr1, 2 and 6, based on GPSeq (light blue) or H structures (dark blue). Dark gray: centromeric regions. Light gray: variable heterochromatic region on chr1. The size of the gray bands is based on the hg19 Giemsa track retrieved from the UCSC Table Browser. In all the figure, PCC and SCC are the Pearson's and Spearman's correlation coefficient, respectively. Dashed red lines: linear regression fits.

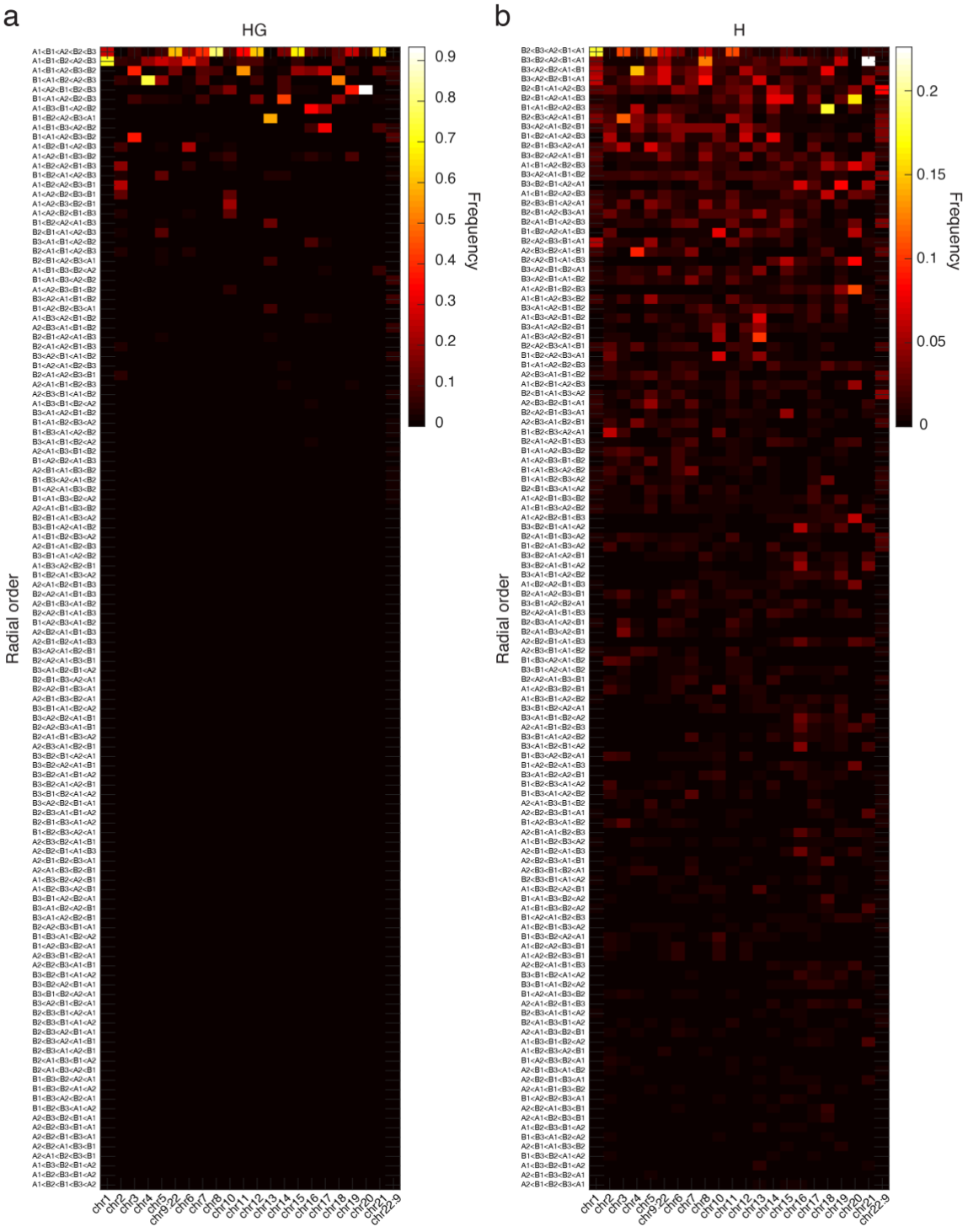
Supplementary Figure 9



Supplementary Figure 9. Analysis of *chromflock* structures generated using all (H structures) or only intra-chromosomal Hi-C contacts (H(intra) structures). (**a**) Distance matrix heatmap. The upper triangle shows the inter-bead 3D distances calculated from 10,000 H(intra) structures. The bottom triangle shows the KR-normalized Hi-C contact frequency matrix, with

each element raised to the power of –0.25. The reported correlation coefficients are for 1 Mb resolution, while for simplicity the plot shows averaged values over 10 Mb genomic windows (points). (**b**) Correlation between the average inter-bead 3D distance in H(intra) structures and the KR-normalized Hi-C contact frequency, with each element raised to the power of –0.25. Each dot represents a pair of 10 Mb non-overlapping genomic windows, each obtained by averaging 1 Mb non-overlapping bins. n = 47,531 genomic window pairs (points) were analyzed. (**c**) Distribution of the average distance from the modeled nuclear surface of 1 Mb beads in H(intra) structures, separately for each chromosome. In all the violin plots, each box spans from the 25th to the 75th percentile, whiskers extend from –1.5×IQR to +1.5×IQR from the closest quartile, where IQR is the inter-quartile range. Dots: outliers (data falling outside whiskers). chr9:22 and chr22:9 are the derivative chromosomes of the t(9;22)(q34;q11.2) translocation. (**d**) Correlation between the average chromosome distance from the modeled nuclear surface H(intra) structures and chromosome size in base-pairs (bp). Each dot corresponds to one chromosome. (**e**) Correlation between the radial position of 1 Mb genomic windows in H(intra) structures and the log2 of the GPSeq score of the same windows. Each dot represents a single 1 Mb genomic window. n = 2,627 genomic windows (points) were analyzed. (**f**) Correlation between the radial position in H(intra) structures and the median 3D distance to the nuclear lamina measured by DNA FISH. Each dot represents one of the 68 FISH probes shown in **Supplementary Fig. 1a**. (**g, h**) Same as in (e) and (f), respectively, but for H structures generated using all Hi-C contacts. In all the figure, PCC and SCC are the Pearson's and Spearman's correlation coefficient, respectively. Dashed red lines: linear regression fits.

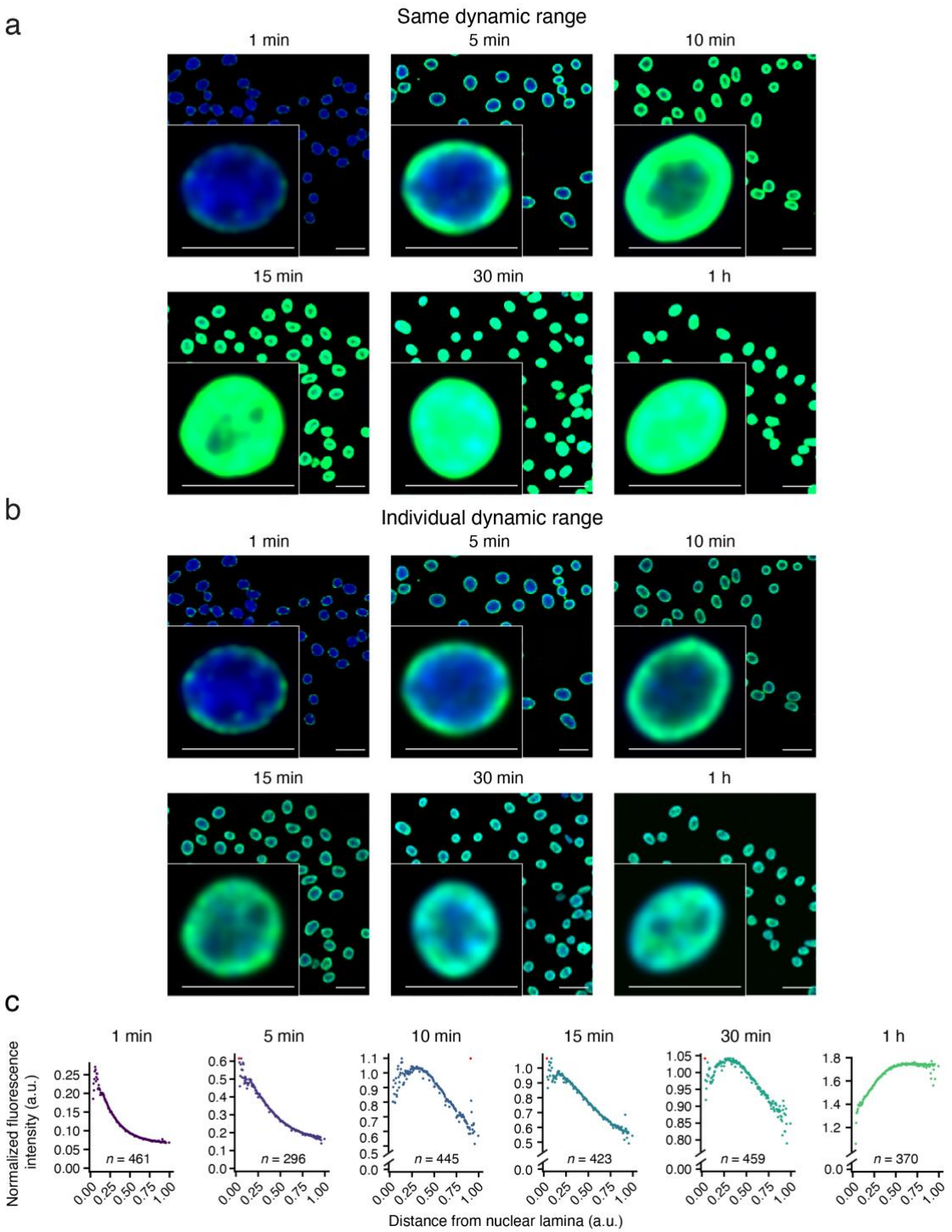
Supplementary Figure 10



Supplementary Figure 10. Radial order of A/B subcompartments in 1,000 *chromflock* structures at 100 kb resolution. (**a**) Mean frequency of all possible radial arrangements of A/B subcompartments in each chromosome, in structures built using GPSeq and Hi-C (HG). The arrangements are shown in order of decreasing frequency from top to bottom. (**b**) Same as in

(a), but for structures built only using Hi-C information (H). chr9:22 and chr22:9 are the derivative chromosomes of the t(9;22)(q34;q11.2) translocation.

Supplementary Figure 11



Supplementary Figure 11. Quantification of gradual antibody diffusion in the nucleus. (**a**) Gradual penetration of an antibody against histone H2A revealed by immunofluorescence and wide-field epifluorescence microscopy. Green: H2A. Blue: DNA stained with Hoechst 33342. Scale bars: 20 µm (field-of-view) and 10 µm (insets). Times indicate the duration of incubation

with HindIII. Optical midsections are shown. The same dynamic range was used for all digestion times. (**b**) Same as in (a) but using a different dynamic range for each digestion time, in order to highlight the pattern of digestion in individual samples. (**c**) Normalized IF fluorescence intensity at various distances from the nuclear lamina, for each of the times shown in (a) and (b). The IF signal was normalized over the fluorescence intensity of DNA stained with Hoechst 33342. Each dot represents the median intensity of each of 200 radiality layers. All the source data for this figure are from HAP1 cells.

2. Supplementary Methods

Information about antibodies, cell lines, data and code availability, and statistics is provided in the **Life Sciences Reporting Summary** available online.

Experimental methods

Preparation of cells for YFISH and GPSeq. We obtained HAP1 cells from Horizon Discovery (cat. no. C859) and cultured them in Iscove's Modified Dulbecco's Medium (IMDM, Merck, cat. no. 51471C) supplemented with 10% fetal bovine serum (FBS, Thermo Fisher Scientific, cat. no. F2442). We obtained GM06990 cells from the Coriell Cell Repository (cat. no. GM06990) and cultured them in Roswell Park Memorial Institute Medium 1640 supplemented with 2 mM L-glutamine (RPMI, Sigma, cat. no. R8758) and 15% fetal bovine serum (Thermo Fisher Scientific, cat. no. F2442). All cells tested negative for Mycoplasma. None of the cell lines used is included in the list of misidentified cell lines; therefore we did not authenticate them. In the case of HAP1, we seeded the cells directly onto 22x22 mm coverslips placed in 6-well plates and grew them until they reached ~70% confluency in each well. In the case of GM06990, which grow in suspension, we first centrifuged them for 5 min at 300 g and re-suspended them in 1X PBS. We then dispensed the cell suspension onto 22x22 mm coverslips pre-coated with Poly-L-Lysine (Sigma, cat. no. P8920-100 ml) and placed inside a 6-well plate and incubated for 10 min at room temperature (RT). We fixed the cells in 0.4X PBS (Thermo Fisher Scientific, cat. no. AM9625)/4% paraformaldehyde (EMS, cat. no. 15710) for 10 min at RT, followed by quenching of unreacted paraformaldehyde in 1X PBS/125 mM glycine for 5 min at RT. Subsequently, we washed the cells three times, 5 min each, with 1X PBS/0.05% Triton X-100 at RT and permeabilized them in 1X PBS/0.5% Triton X-100 for 20 min at RT. Following overnight incubation in 1X PBS/20% glycerol at RT, we subjected the cells to four cycles of freeze-and-thaw in liquid nitrogen, and then washed them three times, 5 min each in 1X PBS/0.05% Triton X-100 at RT. Afterwards, we incubated the cells in 0.1 N HCl for 5 min at RT and quickly rinsed them twice in 1X PBS/0.05% Triton X-100 at RT. Lastly, we rinsed the cells in 2X SSC buffer (Thermo Fisher Scientific, cat. no. AM9763) and stored them in 2X SCC/0.05% NaN_3 up to one month at 4 °C.

Preparation of YFISH and GPSeq adapters. The list of oligonucleotides (oligos) used to prepare YFISH and GPSeq adapters is available in **Supplementary Table 1**. We purchased individual oligos from Integrated DNA Technologies as desalted oligos at 100 mM

concentration in nuclease-free water. UMIs were generated by random incorporation of the four standard dNTPs using the 'Machine mixing' option. We diluted 10 µl of each forward oligo in 90 µl final volume of a mix containing 10 µl of 10X Polynucleotide Kinase (PNK) buffer and 2 µl of T4 PNK (NEB, cat. no. M0201). We incubated the samples for 1 h at 37 °C, after which we added 10 µl of the corresponding antisense oligonucleotides, and incubated the samples for 5 min at 95 °C, followed by gradual cooling down to 25 °C over a period of 45 min (–1.55 °C/min) in a PCR thermocycler. Ready-to-use adapters can be stored at –20 °C up to one year.

3D DNA FISH. We designed and produced all the DNA FISH probes used to validate GPSeq with the iFISH pipeline, which we recently established[1]. The genomic coordinates of the probes are available in **Supplementary Table 3**. We performed DNA FISH using a modification of the original 3D DNA FISH protocol[2], which we recently described[3]. Briefly, we first incubated the coverslips with the pre-hybridization buffer (PHB) containing 2X SSC/5Denhardt's solution/50 mM sodium phosphate buffer/1 mM EDTA/100 ng/ml ssDNA/50% formamide, pH 7.5–8.0, for 1 h at 37 °C. During this time, we prepared the first hybridization mix (HM-1) by mixing the single-locus probes (up to 6 single-locus probes together, typically using a final concentration of 3.2 nM per probe) at 1:9 vol./vol. ratio with 1.1X first hybridization buffer (HB-1) containing 2.2X SSC/5.5 Denhardt's solution/55 mM sodium phosphate buffer/1.1 mM EDTA/111 ng/ml ssDNA/55% formamide/11% dextran sulfate, pH 7.5–8.0. We removed the coverslips from PHB and placed each of them on 10 µl of HM-1 on a microscope slide. We then sealed the coverslips with fixogum (MP Biomedical, cat. no. 11FIXO0125) and waited until the fixogum solidified. We performed DNA denaturation for 3 min at 75 °C on a heating block. Afterwards, we incubated the samples for 18 h at 37 °C. The next day, we washed the coverslips three times, 10 min each at 37 °C in 2X SSC/0.2% Tween, while shaking, followed by two washes, 7 min each at 60 °C in 0.2X SSC/0.2% Tween pre-warmed at 60 °C, inside a water bath, a brief wash in 4X SSC/0.2% Tween at RT, two brief washes in 2X SSC and one final short wash in 2X SSC/25% formamide. Next, we placed the coverslips on 100 µl of the second hybridization mix (HM-2) containing the secondary fluorescently labeled oligos (one color per locus, up to 6 colors together), each at a final 20 nM concentration in 2X SSC/25% formamide/10% Dextran sulfate/1mg/ml *E. coli* tRNA/0.02% BSA, and incubated the samples for 3 h at 30 °C in a humidity chamber. Afterwards, we washed the coverslips for 1 h at 30 °C in 2X SSC/25% formamide, followed by 30 min at 30 °C in 0.1 ng/µl Hoechst 33342 in 2X

SSC/25% formamide. We imaged all the samples using wide-field microscopy as described below.

Visualization of DSBs by immunofluorescence (IF). We visualized DSBs in HAP1 cells using an anti-phospho-Histone H2A.X (Ser139) antibody (Millipore, cat. no. 05-636). We performed IF by adapting a previously published protocol for nuclear proteins[4]. Briefly, we incubated the samples with the primary antibody diluted 1:200 vol./vol. in a blocking buffer consisting of 1X PBS/0.1% Tween 20/1% BSA, for 1 hour at room temperature (RT). For primary antibody detection, we incubated the samples with the secondary antibody diluted 1:500 vol./vol. in the blocking buffer, for 1 hour at RT. We imaged all the samples using wide-field microscopy as described below.

Gradual antibody diffusion. We used the following antibodies: rabbit Anti-Histone H2A (Cell Signaling Technology, cat. no. 12349S), Anti-rabbit IgG ATTO 488 conjugate (Abcam, cat. no. ab150077) and Anti-rabbit IgG Alexa Fluor 647 conjugate (Abcam, cat. no. ab150075). We fixed HAP1 cells following the same protocol used for YFISH and GPSeq. We incubated different samples for 1, 5, 10, 15, 30 min or 1 h in the presence of the primary antibody diluted 1:500 vol./vol. in blocking buffer (1X PBS/0.1% Tween 20/1% BSA). Immediately after incubation, we post-fixed the samples in 1X PBS/4% PFA for 10 min at RT, followed by incubation with the secondary antibody diluted 1:500 vol./vol. in blocking buffer for 1 hour at RT. We imaged all the samples using wide-field microscopy as described below.

Wide-field epifluorescence microscopy. We used wide-field microscopy to image all the YFISH, 3D DNA FISH as well as IF samples. Briefly, we rinsed the samples twice in 2X SSC before mounting them with ProLong Diamond Antifade Mountant (Thermo Fisher Scientific, cat. no. P36965). We imaged all the samples using a 100X 1.45 NA objective mounted on a custom-built Eclipse Ti-E inverted microscope system (Nikon) controlled by the NIS Elements software (Nikon) and equipped with an iXON Ultra 888 ECCD camera (Andor Technology). For YFISH and 3D DNA FISH, we acquired multiple image stacks per sample, each consisting of 40–110 focal planes spaced 0.2 or 0.3 µm apart. We imaged TetraSpeck Microspheres (0.1 µm, fluorescent blue/green/orange and dark red, Thermo Fisher Scientific, cat. no. T7279) before or after each imaging session and used the images to correct for chromatic aberrations and shifts between channels using our in-house suite DOTTER.

STED microscopy. We performed super-resolution 3D-STED imaging of YFISH samples on a Leica SP8 3X STED system equipped with lasers for the depletion of fluorophores emitting in the blue/green (592 nm, MPB Communications Inc.), orange (660 nm, Laser Quantum) and red/far-red (775 nm, OneFive GmbH). For super-resolution imaging of YFISH samples, we excited the ATTO647N fluorophore with a tunable pulsed white-light fiber laser (Leica Microsystems) and an excitation wavelength of 640 nm together with 775 nm stimulated emission depletion. To image DNA stained with Hoechst 33342, we used a 405 nm excitation laser delivered by a diode-laser. We imaged all the samples with a chromatically optimized oil-immersion objective (HC PL APO 100X/1.40 OIL STED WHITE, Leica Microsystems). We passed the fluorescence signals through a 0.9–1.0 Airy unit pinhole and detected them using sensitive photodetectors (Leica Hybrid Detectors). We filtered the emitted light by appropriate dichroic mirrors and selecting an appropriate wavelength window (Hoechst 33342: 420–480 nm; ATTO647N: 650–730 nm) in the Acousto-Optical Beam Splitter (AOBS, Leica Microsystems). To block STED laser light in excess, we used a sharp notch filter in front of the ATTO647N detector. We acquired dual-color axial stacks sequentially, frame-by-frame at a scan speed of 400 lines per sec. We tuned the super-resolution STED pixel size based on the depletion power applied laterally and axially (80% laterally and 20% axially). Before the analysis, we deconvolved all the stacks with the Huygens Software (Scientific Volume Imaging).

Computational methods

YFISH image analysis. We converted raw images from ND2 format (Nikon) to uncompressed TIFF format using the *nd2_to_tiff* script from the *pygpseq* Python3 package available on GitHub (https://github.com/ggirelli/pygpseq/). We deconvolved all the channels using the Huygens Professional v17.04 Software (Scientific Volume Imaging) with the following parameters: CMLE algorithm, null background, and signal-to-noise ratio (SNR) of 7, in 50 iterations. To perform 3D deconvolution, we estimated a theoretical point spread function using the same software, considering both our microscope setup and the optical configuration used to acquire the images. We then estimated the background level of both DNA and YFISH channels as the median of the background intensities, and subtracted these values from the original image (we set negative intensity voxels to zero). After deconvolution, we performed 3D segmentation of the nuclei stained with Hoechst 33342, in each field of view, using the *tiff_auto3dseg* script of the *pygpseq* package. Briefly, the script combines two binary masks

generated with the *threshold_otsu* and *threshold_local* methods from the *scikit-image.filters* package[5] with a logical AND operation. Then, the script discards objects touching the XY contour of the image, fills any holes in the masks, and performs a dilate-fill-erode operation. To identify out-of-focus images, we used the *tiff_findoof* (v0.3.1) tool from the *pygpseq* package. The script discards the stacks where the peak of the gradient magnitude of the stack intensity over Z does not fall in a range of 50% of the stack around the mid slice. For each segmented nucleus, we estimated the volume, shape, surface, flattened size (in Z-projection), sum of intensity, average intensity, a shape descriptor, and the center of mass. Specifically, we ran the script using the following parameters: *--compressed -rn -a 200 130 130 --an-type 3d -uy --nuclear-sel -t 10*. In particular, the *-u* parameter allows extracting all the nuclear bounding boxes for further analysis. For each voxel, the pipeline produced all channel images, nuclear masks, and EDT-based matrices with absolute distance from the nuclear lamina and center (defined as the 1% most distant voxels from the lamina). Then, we used the *pygpseq-scripts* (v0.0.1, DOI: 10.5281/zenodo.3365634) suite of Python3 and R scripts, available at https://github.com/ggirelli/pygpseq-scripts to analyze the extracted nuclear bounding boxes. First, we ran the *extract_nuclear_features.py* script to obtain the characteristics of all extracted nuclei, which we fed in turn to the *select_nuclei.R* script, with parameter *–k 2*, to select only nuclei from cells in the G1 phase of the cell cycle. To achieve this, we fitted a sum of Gaussians model to both the distribution of nuclear volume (in voxels) and the integral over the full nuclear volume of the DNA staining intensity. If the fitting failed, we instead fit a single Gaussian. We selected nuclei that fell in a range of $\pm k$ around the peak of the major Gaussian. If also the single Gaussian fitting failed, we selected nuclei falling in a range of $\pm FWHM/2$ around the major distribution peak, where *FWHM* is the full-width-half-maximum. We retained for further analysis only nuclei selected in such a manner. Then, we used the *extract_nuclear_vx.py* script to obtain the following measurements, for every single voxel of the selected nuclei: intensity from each channel, absolute distance from nuclear lamina, absolute distance from nuclear center, normalized distance from nuclear lamina (defined as the absolute distance from the nuclear lamina divided by the sum of the absolute distance from lamina and center). Next, we fed the voxel data to the *extract_condition_profiles.py*, *extract_nuclear_profiles.py*, and *extract_nuclear_radii.py* scripts. These scripts build radiality profiles from the nuclear periphery inwards for: (1) each condition (*e.g.*, time of digestion) by pooling all the nuclear voxels (*extract_condition_profiles.py*); (2) each nucleus (*extract_nuclear_profiles.py*); and (3) single straight-line trajectories, going from the nuclear center of mass (CoM) to the nuclear surface (*extract_nuclear_radii.py*). Specifically, we ran

the *extract_nuclear_radii.py* script with default parameters, to draw 200 straight trajectories and sample 100 points homogeneously in each of them, for 500 randomly picked nuclei for each condition. The 200 trajectories all depart from the nuclear CoM, are homogeneously spread in space using a spherical Fibonacci lattice, and terminate at their point of intersection with the nuclear surface, determined by a triangular meshing algorithm over the nuclear 3D mask. In all the three cases, we built radiality profiles by first dividing the nuclear radius into 200 bins, and then by assigning the intensity values of the voxels (or sampled points) of interest to the corresponding bins, based on the normalized distance from the nuclear lamina. Then, we calculated the median of each bin and reported it alongside the midpoint of each bin. Finally, we ran the *extract_profile_descriptors.R* script to retrieve information on the peak, inflection point, and contrast of the profile. Specifically, the script fits a 5th-degree polynomial curve to the profile data and uses the *uniroot.all* function from the *rootSolve* R package (v1.7) to identify the peak and inflection point of the profile, and the relative intensities. The profile contrast is then calculated as the ratio of the intensity at the peak over the intensity at the inflection point.

3D DNA FISH image analysis. We deconvolved the DNA staining channel and performed 3D segmentation as described above for YFISH. To identify fluorescence dots corresponding to the detected DNA loci in 3D, and to correct for chromatic aberrations, we used our in-house suite DOTTER. In order to calculate the radial position of the selected FISH dots, we passed the output of DOTTER to the *gpseq_fromfish* (v7.0.2) script from the *pygpseq* package. For each FISH dot, we computed its distance from the nuclear edge (contour of segmented DNA stain) as well as from the nuclear center. Then, we normalized the distances dot-wise, by dividing them by the sum of the corresponding distances from the nuclear edge and center. We defined the nuclear center as the set of voxels in the top percentile of the distribution of distances from the nuclear edge. For each voxel, we calculated its 3D distance from the nuclear edge using an anisotropic 3D Euclidean transform on the nuclear mask using the *ndimage.morphology.distance_transform_edt* script from *scipy* (v1.1.0). We then normalized the distance of each voxel from the nuclear edge, using the same procedure as for the DNA FISH dots. This analysis was implemented as a snakemake flow[6], also available on GitHub (https://github.com/ggirelli/iFISH-singleLocus-analysis).

Pre-processing of GPSeq sequencing data. We demultiplexed raw sequencing data based on the RA5 indexes, either using the BaseSpace Sequence Hub cloud service of Illumina, or manually with bcl2fastq (v2.18). We quality-checked the generated fastq files with *fastqc*

(v0.11.4). We selected the reads containing the full prefix (UMI_barcode_restriction site) using *scan_for_matches*. Afterwards, we trimmed the reads to remove the prefix (including the restriction site) and aligned the remaining part against the human reference genome (Grch37/hg19 GCA_000001405.1) using *bwa-mem*. We retained primary alignments with a mapping quality equal to or higher than 30, while we discarded unmapped reads, chimeric reads, and reads mapped to chrY that is not present in HAP1 cells. We retained only the reads whose 5′ end is less than 20 bp away from the position of a HindIII or MboI recognition site (RS) in the reference human genome, depending on which enzyme was used. The reason why we do not strictly require all the reads to align exactly to the position of RS is that the T7 polymerase and reverse transcriptase used during the GPSeq library preparation are prone to occasionally skip some bases, leading to the resulting reads to align slightly downstream of the RS. Afterwards, we recovered the UMI sequence of each aligned read and we filtered the reads based on the quality of the UMI sequence using an approach similar to the one used by the *fastq_quality_filter* tool (http://hannonlab.cshl.edu/fastx_toolkit/index.html). We de-duplicated UMI sequences mapped to the same restriction site and generated a BED file containing the genomic coordinate and number of de-duplicated UMIs associated with each restriction site. We performed all the above steps using *gpseq-seq-gg* (v2.0.2), a bash/Python/R custom-designed pipeline that is available on GitHub (https://github.com/ggirelli/gpseq-seq-gg). A newer version of the pipeline (v2.0.3) with improvements to the efficiency of the UMI de-duplication script is available at DOI: 10.5281/zenodo.3264757. This version is optimized to deal with datasets generated with 4-base cutters such as MboI used in this study. Finally, we corrected the BED files generated by *gpseq-seq-gg* for the presence of the t(9;22) translocation using an *ad hoc* Python script available at http://github.com/ggirelli/bed-fix-chrom-rearrangement (v0.0.1, DOI: 10.5281/zenodo.3365906), using the following parameters: *-1 chr9:133681295 -2 chr22:23632359*.

Repeat masking

We manually annotated a list of masked regions (**Supplementary Table 10**), using the RepeatMasker track available from the UCSC Table browser for Grch37/hg19. The list includes peri-centromeric and peri-telomeric regions (up to 30 Mb) characterized by abnormally high GPSeq score across all experiments. The largest masked regions correspond to heterochromatic regions of variable length and to the p-arms of acrocentric chromosomes. Moreover, we also included two small regions on the q-arm of chrX and chr5, which yielded abnormally high read counts. When calculating the GPSeq score genome-wide, we masked out any genomic window

overlapping with the masked regions using the *-M* parameter of the *gpseqc_estimate* script, from the *gpseqc* Python3 package available on GitHub (https://github.com/ggirelli/gpseqc/).

Effect of resolution and sequencing depth. To investigate the effect of the resolution (*i.e.*, genomic window size) on the GPSeq score, we calculated the median number of reads, median number of enzyme recognition sites, median number of reads per recognition site, and standard deviation of the number of reads per recognition site, inside genomic windows of 5, 10, 15, 20, 25, 50, 100, 250, 500 kb, and 1, 5, 10 Mb for the longest digestion times (2 hours for HindIII, 30 min for MboI). To assess the effect of the sequencing depth, we iteratively removed 5% of the original de-duplicated UMIs, using an R script and assigning to each de-duplicated UMI the same probability of being removed at any cycle. First, we calculated the same genomic window statistics (window size of 1 Mb, with a step of 100 kb) as described above. Then, we calculated the GPSeq score from the sub-sampled BED files, both in a genome-wide fashion (1 Mb overlapping genomic windows with a step of 100 kb and 100 kb genomic windows) as well as using genomic windows (1 Mb and 100 kb) centered on the midpoint of the DNA FISH probes shown in **Supplementary Fig. 1a**. Following this, we calculated the Pearson's correlation coefficient between the GPSeq score at different simulated sequencing depths and the median distance from the nuclear lamina measured by DNA FISH. Lastly, we calculated the Spearman's correlation coefficient between the genome-wide GPSeq score of each experiment at different simulated sequencing depths, and all the other experiments at full sequencing depth.

Conversion of GPSeq scores to physical distances. To convert the GPSeq score into absolute physical distances, we first retrieved the average HAP1 nuclear volume (in voxels) from the *nuclear.summaries.csv* file automatically generated by the *pygpseq* pipeline and converted it into cubic nanometers (by multiplying it by our voxel size: 130×130×200 nm). Then, we multiplied the log2 GPSeq score by the absolute average nuclear radius to obtain a *transformed* GPSeq score (in nm).

GC-content. We obtained the fraction of guanines and cytosines (GC) in each genomic window (1 Mb and 100 kb) by applying the *letterFrequency* function from the Biostrings R package (https://www.bioconductor.org/packages/release/bioc/html/Biostrings.html) to the genome sequence provided by the BSgenome.Hsapiens.UCSC.hg19 R package (http://bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.UCSC.h

g19.html). We excluded genomic windows overlapping masked regions listed in **Supplementary Table 10** or for which it was not possible to obtain a GPSeq score.

ATAC-seq. We obtained ATAC-seq data for the HAP1 cell line from the GEO repository GSE111047 (**Supplementary Table 5**). We used the Parker Lab's ATAC-seq processing pipeline (https://github.com/ParkerLab/ATACseq-Snakemake) to obtain the genome-wide accessibility signal in bedGraph format. To directly compare the GPSeq scores and genome accessibility, we binned the ATAC-seq bedGraph files using the *bioTrackBinner* R script (v0.0.1) (http://github.com/ggirelli/bioTrackBinner, DOI: 10.5281/zenodo.3365977).

Lamin DamID. We downloaded Lamin B DamID data for the HAP1 cell line from the 4D Nucleome project web portal (https://www.4dnucleome.org/) (**Supplementary Table 5**). To directly compare the GPSeq scores at different resolutions and the Lamin B DamID signal over control, we binned the DamID bigWig files using the *bioTrackBinner* R script (v0.0.1) (http://github.com/ggirelli/bioTrackBinner, DOI: 10.5281/zenodo.3365977). To visualize the radial distribution of the DamID track, we used 'pizza-plots' as described below. To plot the distribution of DamID signal over control, we binned the log2 GPSeq score in 10 equally sized bins from 0 to 1. We obtained a pre-processed list of LAD and *inter*-LAD genomic locations with their classification into constitutive and facultative from ref. [7]. We assigned the GPSeq scores univocally to one of the four classes based on the location of the midpoint of the genomic windows.

GPSeq score profiles. We visualized GPSeq score profiles along chromosome ideograms by using the *ggkaryo2* R package (v0.0.3) available at http://github.com/ggirelli/ggkaryo2 (DOI: 10.5281/zenodo.3358887). We generated a circular chromosome-wise GPSeq score representation with the same package by reporting the log2 of the GPSeq score on the *y*-axis, the genomic chromosome coordinate (in bp) on the *x*-axis arranged circularly using the *coord_polar* function, and the connecting consecutive data points by segments colored based on the GPSeq score. We marked in red the peri-centromeric regions listed in **Supplementary Table 10**.

Pizza-plots. We generated scatter 'pizza-plots' using the *ggplot2* R package by setting the y-axis to follow the log2 GPSeq score and reporting the chromosomes on the x-axis arranged circularly using the *coord_polar* function. We then jittered the points at the same radial position

(*i.e.*, same GPSeq score) using the *geom_jitter* function and coloring the data points based on orthogonal data track scores. We generated tiled pizza-plots by dividing the log2 GPSeq score into 10 equally sized bins from 0 to 1 and assigning data points to each bin in a chromosome-wise fashion. Afterwards, we plotted each bin as in scatter pizza-plots but replacing the *geom_jitter* by the *geom_tile* function and normalizing the values to have the tiles of each layer (log2 GPSeq score bin) or slice (chromosome) summing up to 100%. To visualize the radial distribution of all chromosomes at once, we first filled a disk with circular sectors, one per chromosome. Then, for each of five layers of equal thickness (not shown) we updated the number of pixels per chromosome to be proportional to the number of genomic sites in that layer in our GPSeq data. The order of chromosomes was arbitrary.

DNA methylation and histone marks. We obtained DNA methylation and histone mark ChIP-seq and ChIPmentation data from ENCODE (**Supplementary Table 5**). For DNA methylation, we retrieved the corresponding BED file, while for ChIP-seq datasets we retrieved the corresponding bigWig file of the merged replicates signal ('fold change over control') mapped to the Grch37/hg19 genome assembly. ChIP-seq and ChIPmentation datasets for which signal tracks were not available were processed with a custom pipeline. Briefly, we removed the adapter sequences using *TrimGalore* (v0.4.4_dev) and discarded reads shorter than 20 nt. We then aligned the reads to the UCSC hg19/GRCh37 genome assembly using *bwa-mem* (v0.7.17-r1188) with default options. We filtered out the reads that failed to align and those with MAPQ ≤ 30, and discarded PCR duplicates using the *Picard MarkDuplicates* (v2.18.11) module. Finally, we generated genome coverage tracks in bigWig format using the *bamCoverage* module from *deeptools* (v3.2.1) with *--binSize* 50. We used these bigWig and BED tracks to calculate the signal at 1 Mb or 100 kb resolution, by applying *bioTrackBinner* to them. Afterwards, we combined ChIP-seq and DNA methylation binned tracks with the GPSeq score calculated at 1 Mb or 100 kb resolution and averaged the tracks signal across replicates. Finally, we built radiality profiles following the same procedure as described above for DamID.

Gene expression and RNA Pol II occupancy. We obtained HAP1 RNA-seq raw files from the GEO repository GSE95015 and RNA Pol II ChIP-seq data from GSE107599 (**Supplementary Table 5**). We removed the adapter sequences using *TrimGalore* (v0.4.4_dev) and discarded reads shorter than 16 nt. We mapped the reads using *Bowtie2* (v2.3.4.1; parameters: *--sensitive-local*) to a list of human rRNAs and tRNAs retrieved from NCBI

(https://www.ncbi.nlm.nih.gov/). We then mapped the reads that failed to align in the previous step to the UCSC hg19/GRCh37 genome assembly by using the *Gencode* (v27) gene annotation as a reference and the *STAR* software (v2.6.0c; parameters: *--twopassMode Basic --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --sjdbScore 1 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 100000 --outFilterMultimapNmax 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04*). We discarded PCR duplicates using the *Picard MarkDuplicates* module (v2.18.11). To quantify the RNA-seq signal at the gene level, we used the *QoRTs QC* module (v1.3.0; parameters: *--minMAPQ 255*) and the Gencode gene annotation (v27) as a reference. We normalized gene counts to TPM (Transcript Per Million tags). To quantify the RNA-seq signal at 1 Mb or 100 kb resolution, we used the *bamCount* function from the *bamsignals* package (v1.12.1) in R. For RNA Pol II ChIP-seq, we performed raw data processing and genomic tracks generation in the same way as described for histone marks. Finally, we built radiality profiles following the same procedure as described above for DamID.

Gene set radial enrichment. We obtained the list of homeobox genes from the Homeobox Database[8], the list of housekeeping genes from Eisenberg and Levanon[9], and the list of hallmark gene sets from MSigDB (http://software.broadinstitute.org/gsea/index.jsp). In one approach, we generated gene set enrichment plots by assigning to each gene the GPSeq score and layer of the 100 kb genomic window that comprised its TSS. Then, for each layer, we calculated the fraction of gene set entries over the fraction of all genes. In the other approach, we compared the GPSeq score distributions and calculated the statistical significance of the shift of each gene set from the background ("all") gene distribution using a two-sided Wilcoxon rank sum test.

TFBS radial distribution. We retrieved the list of transcription factor binding sites (TFBS) predicted in the UCSC hg19/GRCh37 genome assembly from JASPAR2020[10]. For each TFBS, we counted its occurrence within 1Mb and 100 kb genomic windows. We then calculated the GC-content of each TFBS as the average probability of each base in the motif being either a G or a C. Lastly, we used the *ggplot2* script to plot the Pearson's correlation coefficient between the number of TFBS per genomic window and the GPSeq score of the same window, against the GC-content of each TFBS.

Linear models of the GPSeq score. We retrieved chromosome sizes (C_s) based on the Grch37/hg19 genome assembly available from the UCSC Table Browser. We calculated gene

expression (E) as the poly(A) fraction read pile-up for each 1 Mb window or the 1 Mb pile-up averaged per chromosome (**Supplementary Methods**). We calculated gene density (D_g) as the number of transcription start sites (TSS) per 1 Mb window (one TSS per gene) or as the 1 Mb density averaged per chromosome (**Supplementary Methods**). We calculated the GC-content (P_{GC}) as the fraction of each 1 Mb window consisting of Gs or Cs. We then used the models to calculate the expected (predicted) GPSeq score (**Supplementary Methods**). We built uni- and multi-variable linear models using the *lm* function in R. Specifically, at chromosome-wide resolution, we built the following models (estimated parameters β, R-squared, and model P-values are available in **Supplementary Table 6**).

$$\log_2(\text{GPSeq score}) \sim \beta_0 + \beta_1 \cdot P_{GC} + \beta_2 \cdot C_s \quad (1)$$

$$\log_2(\text{GPSeq score}) \sim \beta_0 + \beta_1 \cdot P_{GC} + \beta_2 \cdot C_s + \beta_3 \cdot D_g \quad (2)$$

$$\log_2(\text{GPSeq score}) \sim \beta_0 + \beta_1 \cdot P_{GC} + \beta_2 \cdot C_s + \beta_3 \cdot \log_{10} E \quad (3)$$

At 1 Mb resolution, we built the following models instead.

$$\log_2(\text{GPSeq score}) \sim \beta_0 + \beta_1 \cdot P_{GC} + \beta_2 \cdot D_g + \beta_3 \cdot \log_{10} E + \beta_4 \cdot C_s \quad (4)$$

$$\log_2(\text{GPSeq score}) \sim \beta_0 + \beta_1 \cdot P_{GC} + \beta_2 \cdot D_g + \beta_3 \cdot \log_{10} E + \beta_4 \cdot C_s + \beta_5 \cdot X \quad (5)$$

We replaced X with different epigenetic mark and replication time data. A complete list of the estimated parameters, β, $R2$, and model P-values is available in **Supplementary 9**. We calculated the prediction error (PE) as the square root of the mean square prediction error.

Replication timing. We obtained Repli-seq data from ENCODE (**Supplementary Table 5**). We retrieved the bigWig files of the signal in each of the cell cycle phases (G1, S1, S2, S3, S4, G2) and of their wavelet transform. We binned the bigWig files with *bioTrackBinner* and combined them with the GPSeq score calculated at the same resolution. We used the cell cycle phase-specific signal to color data points in scatter pizza-plots.

Hi-C contacts, A/B compartments and subcompartments. We downloaded HAP1 Hi-C raw data from the 4D nucleome project web portal (https://www.4dnucleome.org/) (**Supplementary Table 5**). We processed the data using *Juicer*[11] by aligning the data against the human reference genome (Grch37/hg19 GCA_000001405.1) using *bwa-mem*. We classified the Hi-C contacts into four groups based on their position within the genome-wide contact frequency matrix. We defined as 'diagonal' and 'near-diagonal' the Hi-C contacts occurring between loci lying inside the same or to linearly adjacent genomic windows, respectively. We classified the remaining contacts as 'intra' or 'inter', depending on whether

the genomic loci were located on the same or on different chromosomes. For A/B compartment analyses, we computed Eigenvector tracks, intra-chromosome matrices, and inter-chromosome matrices using *Juicer Tools* (v1.9.8) with KR normalization at 1 Mb resolution. For each chromosome, we changed the sign of the Eigenvector, if needed, to have a negative correlation between the Eigenvector and the Lamin DamID-seq signal. For A/B subcompartment analyses, we retrieved the coordinates of subcompartment predictions for HAP1 cells at 100 kb resolution (see Source Data in ref. 12). We then assigned the subcompartment with the highest predicted probability to each non-overlapping 100 kb genomic window.

Radial position *vs.* chromosome size in *chromflock* structures. We converted the Hi-C KR-normalized matrix at 1 Mb resolution to a distance matrix, by elevating each element in the matrix to –0.25. For each pair of 1 Mb windows (beads), we computed their 3D spatial distances as the Euclidean distances between their centroids for each *chromflock* structure, averaged them across structures, and assembled them into a distance matrix. Finally, we calculated the Pearson's and Spearman's correlation coefficients between the full Hi-C distance matrix and the averaged 3D spatial bead distance matrix.

Radial distribution of A/B compartments in *chromflock* structures. To assess the polarity of A/B compartments in *chromflock* structures built using GPSeq and Hi-C data, we calculated the median radial position based on the distance of each bead from the nuclear surface, for all beads belonging to either A or B compartments, and then calculated the difference in the median radial position between A and B compartments for every structure.

Orientation of A/B subcompartments in *chromflock* structures. For simplicity, let us first assume that we are interested in studying the orientation of genomic compartments with two different properties, $S1$ and $S2$. For each chromosome in each *chromflock* structure, we first identify the beads with these properties: $T1 = \{b_i, b_i \in S1\}$ and $T2 = \{b_i, b_i \in S2\}$. We also use the set of beads with any of these two properties, *i.e.*, $T1 = T1 \cup T2$. We then define the normalized vector, d, from the center of mass of $T2$, $c(T2)$, to $c(T1)$, and name the normalized vector from $c(T)$ to 0 as n. We define the orientation, o as:

$$o = d \cdot n \in [-1,1] \quad (6)$$

i.e., $o = 1$ if beads with property $S1$ are comprised between $c(T2)$ and 0, and $o = -1$ if $c(T2)$ is between $c(T1)$ and 0. If a chromosome were completely polarized with respect to $S1$ and

$S2$, we would be able to draw a plane that completely separates $T1$ and $T2$. We defined polarity, p as the quotient of beads that can be correctly classified as $S1$ or $S2$ by a linear classifier. Thus, p can have any value in the range $[0,1]$, where 1 indicates perfect (linear) polarity, while any value close to 0.5 indicates that the two classes are mixed.

Radial distribution of SNPs and cancer SNVs. For SNPs, we retrieved the hg19 genomic coordinates of germline variants in the 1000 Genomes Project Phase 3[13] from which we selected only SNPs (>81 million events). For cancer SNVs, we obtained the hg18 genomic coordinates of substitutions found in four different tumor types from the Supplementary Tables in the corresponding publications described in ref. [14], and used the UCSC LiftOver utility[15] to convert hg18 coordinates to hg19. In both cases, we calculated the number of events per 100 kb genomic window and then averaged all the windows belonging to the same nuclear layer.

Radial distribution of gene fusions and mingling. We retrieved the cancer-related gene fusions from TCGA (https://www.tumorfusions.org/). We overlapped *chromflock* structure beads and gene fusions using the *foverlaps* function from the *data.table* R package, and labeled as "Fusions" the beads overlapping with at least one gene fusion junction in any cancer type and as "Controls" all the remaining beads.

Radial distribution of DSBs. To assess the radial distribution of endogenous DSBs, we used a published dataset (GSM3444988), which we previously obtained from K562 cells[16], by applying a modification of our BLISS method[17] (**Supplementary Table 5**). We computed the number of unique DSB ends per 100 kb genomic window. We defined as 'genic' the genomic windows with their center overlapping on either strand with any of the GENCODE v19 reference gene annotations and considered the remaining windows as 'intergenic'. We normalized the BLISS signal by dividing it by the length of the 'genic' or 'intergenic' portion of the genome in each of ten concentric nuclear layers. To analyze the radial distribution of DSBs in different A/B subcompartments, we assigned the genomic windows to different A/B subcompartments as described above for genic *vs.* 'intergenic regions, and then averaged the BLISS signal across all the windows in each layer.

Intermingling in *chromflock* structures. To assess mingling between heterologous chromosomes in the *chromflock* structures built using both GPSeq and Hi-C data (1 Mb resolution), we first extracted all the neighboring beads, N found in a sphere four times the bead

radius, for each bead, b_i belonging to a given chromosome, c_j. For every chromosome with beads in N, except c_j, we constructed the convex hull, H_j using the chromosome's beads in N. We defined as a mingling event the case in which the centroid of b_i falls inside H_j. We then defined the mingling frequency as the fraction of *chromflock* structures in a given set, in which a bead mingles.

3. Supplementary Tables

Supplementary Table 1. List of oligos used to make YFISH and GPSeq adapters. Because of its large size, the table is provided as a separate Excel file.

Supplementary Table 2. Summary of sequencing experiments. Because of its large size, the table is provided as a separate Excel file.

Supplementary Table 3. Genomic coordinates of the 68 DNA FISH probes used for GPSeq validation. Because of its large size, the table is provided as a separate Excel file.

Supplementary Table 4. Correlation between the log2 GPSeq score of genomic windows centered on the midpoint coordinates of the 68 FISH probes shown in Supplementary Fig. 1a and the median normalized distance from the nuclear lamina, as measured by DNA FISH. PCC and SCC are the Pearson's and Spearman's correlation coefficient, respectively. The number of genomic windows compared in each experiment is available in Supplementary Table 11.

Dataset	1 Mb		100 kb	
	PCC	SCC	PCC	SCC
Exp.1 (HindIII)	0.909	0.920	0.815	0.821
Exp.2 (HindIII)	0.899	0.897	0.767	0.768
Exp.3 (MboI)	0.921	0.910	0.889	0.906
Exp.4 (MboI)	0.904	0.888	0.895	0.882
Average	0.926	0.925	0.913	0.910

Supplementary Table 5. List of publically available datasets used in this study.

Cell line	Technique	Dataset	Accession
HAP1	ATAC-seq	Accessibility	SRR6766909, SRR6766910, SRR6766911
HAP1	DamID	Lamin B	4DNESUK5H9Y8
HAP1	Hi-C	Raw data	4DNFI1E6NJQJ
K562	Methyl-RRBS	DNA methylation	ENCFF001TOL, ENCFF001TOM
HAP1	ChIPmentation	H3K27ac	SRR6410066
HAP1	ChIPmentation	H3K4me1	SRR6410073, SRR6410074
HAP1	ChIPmentation	H3K4me3	SRR6410076, SRR6410077
HAP1	ChIPmentation	H3K36me3	SRR6410070, SRR6410071
HAP1	ChIP-seq	H3K56ac	GSM2871906, GSM2871907
HAP1	ChIPmentation	H3K9me3	SRR6410079, SRR6410080
HAP1	ChIPmentation	H3K27me3	SRR6410068
HAP1	ChIP-seq	RNA Pol II	GSM2871910, GSM2871911
HAP1	RNA-seq	Poly(A)+ RNA	GSM2493886, GSM2493887, GSM2493888, GSM2493898, GSM2493899, GSM2493900
HAP1	Hi-C	DNA loops	GSE74072
K562	Repli-seq	Wave and cell cycle phase-specific signal	GSM923448
K562	BLISS	DNA double-strand breaks	GSM3444988

Supplementary Table 6. Estimated parameter values, parameter *P*-values (calculated on t-statistic), adjusted R_2 (based on the Wherry formula), and model *P*-values (calculated from F-statistic) for multivariate models at chromosome-wide resolution (1-3) and 1 Mb resolution (4). All *P*-values were calculated based on the indicated number *n* of chromosomes (1-3) or 1 Mb genomic windows (4).

Model	Estimated parameters		Param. *P*-value	*n*	Adjusted R_2	*P*-value
(1)	β_0	-2.76 ± 0.2964	$1.04 \cdot 10^{-8}$	23	0.9386	$2.95 \cdot 10^{-13}$
	β_1	7.891 ± 0.645	$9.66 \cdot 10^{-11}$			
	β_2	$(-1.234 \pm 0.356) \cdot 10^{-9}$	0.0024			
(2)	β_0	-1.855 ± 5.229	0.0022	23	0.9469	$6.762 \cdot 10^{-13}$
	β_1	5.283 ± 1.414	0.0014			
	β_2	$(-1.251 \pm 0.331) \cdot 10^{-9}$	0.0013			
	β_3	$(8.759 \pm 4.3) \cdot 10^{-3}$	0.0558			
(3)	β_0	-2.709 ± 0.3124	$4.96 \cdot 10^{-8}$	23	0.9366	$3.662 \cdot 10^{-12}$
	β_1	7.036 ± 1.553	0.0002			
	β_2	$(-1.333 \pm 0.3964) \cdot 10^{-9}$	0.0032			
	β_3	0.09652 ± 0.158	0.5506			
(4)	β_0	$-1.142 \pm 9.89 \cdot 10^{-3}$	$< 2.2 \cdot 10^{-16}$	26,630	0.7407	$< 2.2 \cdot 10^{-16}$
	β_1	$(3.273 \pm 0.0729) \cdot 10^{-3}$	$< 2.2 \cdot 10^{-16}$			
	β_2	4.205 ± 0.0265	$< 2.2 \cdot 10^{-16}$			
	β_3	$(-3.422 \pm 0.0956) \cdot 10^{-2}$	$< 2.2 \cdot 10^{-16}$			
	β_4	$(-8.472 \pm 0.1486) \cdot 10^{-10}$	$< 2.2 \cdot 10^{-16}$			

Supplementary Table 7. MSigDB hallmark gene sets ranked by decreasing median GPSeq score. *P*-values: two-sided Wilcoxon rank sum test against the distribution of all the genes. *n*: number of genes in each set based on which the *P*-value was calculated.

Gene set	Genes (*n*)	GPSeq	*P*-value
HALLMARK_MYC_TARGETS_V2	55	0.74	0.00511670338629878
HALLMARK_DNA_REPAIR	140	0.733	0.000001154328146
HALLMARK_APICAL_SURFACE	42	0.711	0.229232829803953
HALLMARK_APICAL_JUNCTION	192	0.694	0.000579385889738793
HALLMARK_KRAS_SIGNALING_DN	190	0.678	0.110675080821622
HALLMARK_IL6_JAK_STAT3_SIGNALING	85	0.671	0.325410724331295
HALLMARK_ESTROGEN_RESPONSE_LATE	193	0.667	0.0237634796595256
HALLMARK_MYOGENESIS	191	0.662	0.000673442017787301
HALLMARK_MYC_TARGETS_V1	188	0.662	0.151093820082313
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	110	0.658	0.0253969030501521
HALLMARK_UV_RESPONSE_UP	150	0.648	0.00504504472329553
HALLMARK_P53_PATHWAY	191	0.646	0.0789674140809067
HALLMARK_PI3K_AKT_MTOR_SIGNALING	100	0.645	0.0908666914497851
HALLMARK_ADIPOGENESIS	190	0.619	0.346364712617381
HALLMARK_E2F_TARGETS	193	0.617	0.22339125118483
HALLMARK_PEROXISOME	99	0.605	0.460017232276344
HALLMARK_OXIDATIVE_PHOSPHORYLATION	195	0.603	0.159156650387645
HALLMARK_TGF_BETA_SIGNALING	52	0.602	0.579823597504093
HALLMARK_ESTROGEN_RESPONSE_EARLY	192	0.599	0.662777824849791
HALLMARK_ALLOGRAFT_REJECTION	188	0.599	0.972835327397186
HALLMARK_XENOBIOTIC_METABOLISM	188	0.596	0.620178055254321
HALLMARK_MTORC1_SIGNALING	195	0.594	0.793698067362589
HALLMARK_HEME_METABOLISM	189	0.594	0.917615818031126
HALLMARK_G2M_CHECKPOINT	190	0.589	0.88293752229152
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	46	0.587	0.359820422845325
HALLMARK_MITOTIC_SPINDLE	181	0.584	0.467568891619878
HALLMARK_WNT_BETA_CATENIN_SIGNALING	39	0.584	0.887730165691316
HALLMARK_HEDGEHOG_SIGNALING	35	0.575	0.148858769715641
HALLMARK_APOPTOSIS	153	0.57	0.663446872335927
HALLMARK_INTERFERON_GAMMA_RESPONSE	194	0.569	0.447037914566102
HALLMARK_GLYCOLYSIS	190	0.564	0.835490819326075
HALLMARK_COAGULATION	130	0.562	0.357891216770064
HALLMARK_CHOLESTEROL_HOMEOSTASIS	70	0.56	0.955504687507035
HALLMARK_HYPOXIA	196	0.559	0.945718996179087

Supplementary Table 8. Correlation between the TFBS density and the GPSeq score at (1 Mb overlapping genomic windows with 100 kb step). *n* = 26,630 genomic windows were compared. Because of its large size, the table is provided as a separate Excel file.

Supplementary Table 9. Estimated parameter values, parameter *P*-values (calculated on t-statistic), and R_2 adjusted (based on the Wherry formula) for a multivariate model where *X* corresponds to the indicated track (1 Mb resolution). Model *P*-values (based on F-statistic) are not reported because always $< 2.2 \cdot 10^{-16}$. For each track, *n* = 26,630 genomic windows were analyzed. Dataset accession numbers are listed in Supplementary Table 5.

Track	β_5		Adjusted R_2
	Estimated parameter	Param. *P*-value	
H3K27ac	$(3.439 \pm 0.3528) \cdot 10^{-2}$	$< 2.2 \cdot 10^{-16}$	0.7416
H3K27me3	$(-7.226 \pm 0.2255) \cdot 10^{-2}$	$< 2.2 \cdot 10^{-16}$	0.7504
H3K36me3	$(2.892 \pm 0.1006) \cdot 10^{-2}$	$< 2.2 \cdot 10^{-16}$	0.7482
H3K4me1	$(-5.009 \pm 1.658) \cdot 10^{-3}$	0.00251	0.7407
H3K4me3	$(2.996 \pm 0.2406) \cdot 10^{-2}$	$< 2.2 \cdot 10^{-16}$	0.7422
H3K56ac	$(0.4247 \pm 5.007) \cdot 10^{-3}$	0.085	0.7406
H3K9me3	$(6.878 \pm 0.2637) \cdot 10^{-2}$	$< 2.2 \cdot 10^{-16}$	0.7472
DNAme	$(3.222 \pm 0.4715) \cdot 10^{-1}$	$8.41 \cdot 10^{-12}$	0.7411
RNAPII	$(3.262 \pm 0.1656) \cdot 10^{-2}$	$< 2.2 \cdot 10^{-16}$	0.7444
Repli-seq wave	$(9.433 \pm 0.7425) \cdot 10^{-4}$	$< 2.2 \cdot 10^{-16}$	0.7422

Supplementary Table 10. List of masked manually curated telomeric and peri-centromeric regions. Because of its large size, the table is provided as a separate Excel file.

Supplementary Table 11. Specification of sample size (*n*) and *P*-values for Main, Extended Data Figures, and Supplementary Figures for which the values are too many to be included in the corresponding legend. Because of its large size, the table is provided as a separate Excel file.

Supplementary Table 12. Specification of sample size (*n*) and *P*-values for Supplementary Note 1 Figures for which the values are too many to be included in the corresponding legend. Because of its large size, the table is provided as a separate Excel file.

4. Supplementary Videos

Supplementary Video 1. Rendering of gradual gDNA digestion showing which parts of the genome are cut first (and, therefore, are more peripheral) and which are digested later. The GPSeq score appears along each chromosome ideogram as bars of increasing height. The height of the bars follows the time of enzyme diffusion, as shown in the cartoon on the left. The video is provided as a separate .mp4 file. In the chromosome ideograms, the color of the cytobands is based on the intensity of the Giemsa staining, peri-centromeric regions are colored in red, and acrocentric regions and variable heterochromatic regions are colored in cyan.

Supplementary Video 2–5. 3D rendering of selected examples of the 10,000 whole-genome structures generated by *chromflock* by integrating GPSeq and Hi-C information. In all the structures, each bead represents a 1 Mb genomic window (non-overlapping). Chromosomes are shown with distinct colors. Elements connecting the beads are shown in yellow. The modeled nuclear surface is shown in grey. All the videos are provided as separate .m4v files.

5. Supplementary Notes

Supplementary Note 1. Calculation of the GPSeq score

To convert the sequencing data generated by GPSeq into radiality maps, we have developed a computational framework that allows computing a centrality score, namely the GPSeq score. Below we describe in detail the centrality scores that we have tested and how the GPSeq score is calculated.

Introduction

A GPSeq experiment consists of n digestion times (herein named conditions), which yield $D_1, D_2, \dots, D_n$ sets of processed sequencing reads (one set for every digestion time). Each condition is characterized by a specific number of de-duplicated reads, $N_R(D_j)$, distributed along the genomic recognition sites, s, of the restriction enzyme used in the experiment. Let us denote $N_R(s, D_j)$ as the number of de-duplicated reads mapped to a given recognition site, s and $N_S(D_j)$ as the number of recognition sites for a given restriction enzyme considered in the condition D_j (for more details on how to consider recognition sites, please refer to the section "*Recognition site domain*" below). Thus, we have:

$$N_R(D_j) = \sum_{i=1}^{N_s(D_j)} N_R(s_i, D_j) \quad (1)$$

Let us then represent a genomic window, w by the genomic coordinates of the first, f and the last, l bases included in the window on a specific chromosome, c:

$$w = \{c_w, f_w, l | f_w < l_w\} \quad (2)$$

Each window is characterized by a specific number of de-duplicated reads, $N_R(w, D_j)$ and of recognition sites, $N_s(w, D_j)$ for a given restriction enzyme. Each recognition site can be considered as a small genomic window. Thus, the number of reads and recognition sites can be written as:

$$N_s(w, D_j) = |\{s | s \in w\}| \quad (3)$$

$$s \in w \Leftrightarrow c_w = c_S \wedge f_w \leq f_s < l_s \leq l_w \quad (4)$$

$$N_R(w, D_j) = \sum_{i=1}^{N_s(w, D_j)} N_R(s_i, D_j) \quad (5)$$

Considering only reads that are mapped to the recognition sites of the restriction enzyme in use, we can define the restriction (digestion) probability as:

$$P_w(w, D_j) = \frac{N_R(w, D_j)}{N_R(D_j)} \quad (6)$$

While the P_w of the same window can be compared across different conditions, it is difficult to use this measure to compare different windows in the same condition, since different genomic windows can contain a different number of recognition sites, which directly affects the restriction probability. In other words, genomic windows with a higher number of recognition sites are expected to show a higher P_w compared to windows with a lower number of recognition sites. To take this into account, we define the restriction probability as the average probability, $P_s(w, D_j)$ that a single recognition site in a given genomic window is cut:

$$P_s(w, D_j) = \frac{P_w(w, D_j)}{N_s(w, D_j)} = \frac{N_R(w, D_j)}{N_R(D_j) \cdot N_s(w, D_j)} \quad (7)$$

This probability takes $N_S(w, D_j)$ into consideration and can be compared both across different genomic windows in the same condition as well as across conditions. At the same time, each genomic window is characterized by a different distribution of sequencing reads along its recognition sites. Thus, we can define the mean of the number of reads per site, E and the variance, V as:

$$E(w, D_j) = \frac{N_R(w, D_j)}{N_s(w, D_j)} \quad (8)$$

$$V(w, D) = \frac{1}{N_s(w, D_j) - 1} \sum_{i=1}^{N_s(w, D_j)} \left(N_R(s_i, D_j) - E(w, D_j)\right)^2 \quad (9)$$

Definition of centrality scores

Initially, we computed the correlation between N_R calculated for genomic windows of 1 Mb and 100 kb centered on the 68 DNA FISH probes shown in **Supplementary Fig. 1a**, and the median distance from the nuclear lamina measured using the same probes, which showed a reproducible increase in the correlation with increasing digestion times (**Supplementary Note 1 Fig. 1a-e**). This observation prompted us to ponder on possible ways in which we could combine the different restriction conditions into a single centrality score. Specifically, we formulated two groups of candidate centrality estimates, based either on the read variance, V or on the restriction probability, P_S.

Probability-based centrality scores

We first hypothesized that centrally located genomic windows should have higher P_s in samples digested for a longer time, compared to shortly digested samples, while peripheral genomic windows should show the opposite behavior. Therefore, we combined P_s into a probability-based centrality measure (C_{P_s}) by comparing each condition with the immediately preceding one:

$$C_{P_s}(w) = \sum_{j=2}^{n} \frac{P_s(w, D_j)}{P_s(w, D_{j-1})} \tag{10}$$

Additionally, we defined a centrality measure based on the cumulative restriction probability:

$$P_{s_c}(w, D_j) = \begin{cases} P_s(w, D_j), & j = 1 \\ P_s(w, D_j) + P_{s_c}(w, D_{j-1}), & 1 < j \leq n \end{cases} \tag{11}$$

$$C_{P_{s_c}}(w) = \sum_{j=2}^{n} \frac{P_{s_c}(w, D_j)}{P_{s_c}(w, D_{j-1})} \tag{12}$$

Variability-based centrality scores

We then reasoned that peripheral windows should be fully digested at any restriction condition, thus showing a stable read variance, V. On the other hand, internal windows would show increasing numbers of reads, N_R and at the same time a higher read variance, V. Hence, we defined a variability-based centrality score, (C_V) by comparing the last and first condition:

$$C_V(w) = \sum_{j=2}^{mn} \log\left(\frac{V(w, D_j)}{V(w, D_{j-1})}\right) = \log\left(\frac{V(w, D_n)}{V(w, D_1)}\right) \tag{13}$$

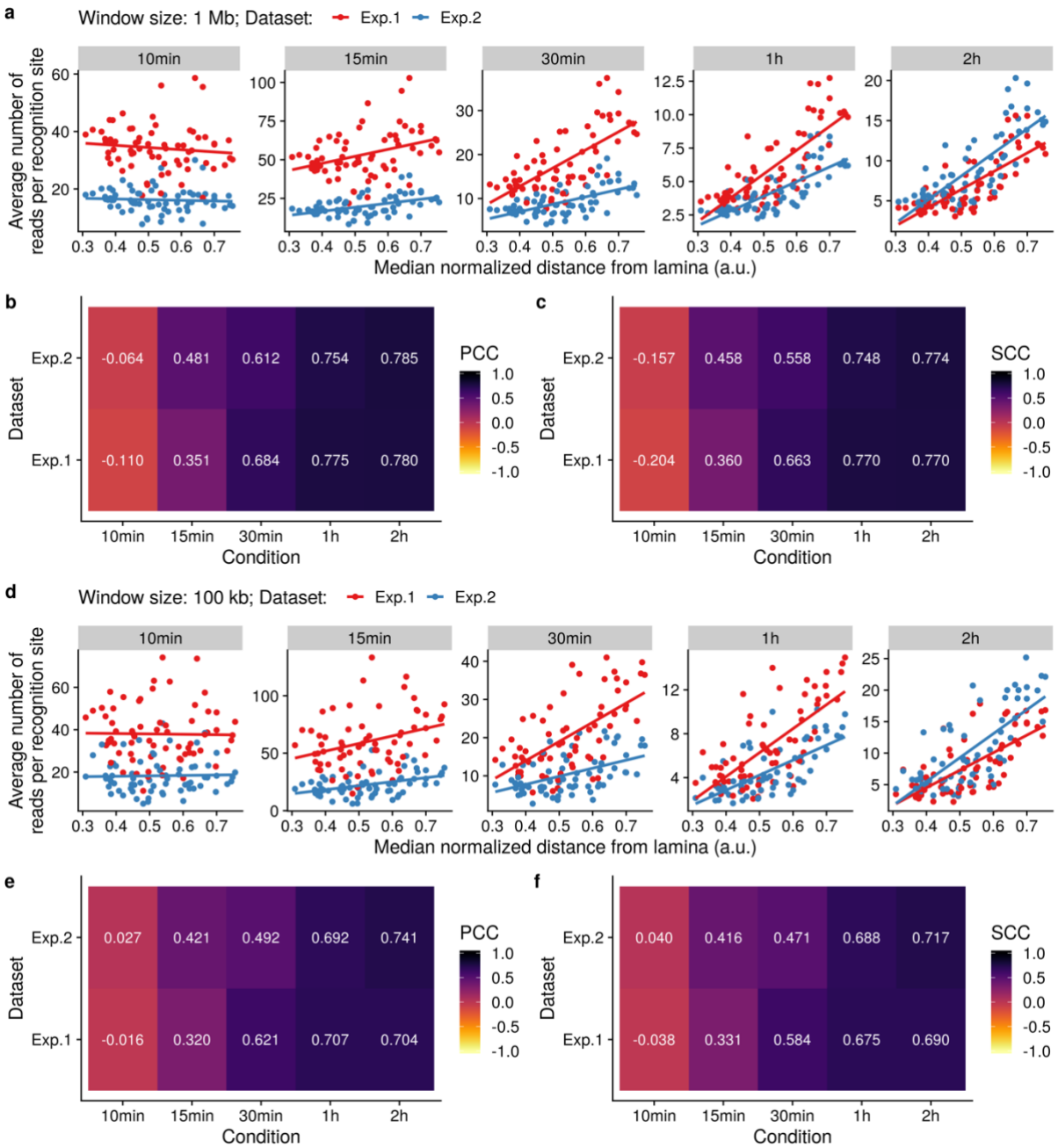
Furthermore, we defined similar centrality measures based on the coefficient of variation, (c_v) and of dispersion, (c_d):

$$c_v(w, D) = \frac{\sqrt{V(w, D)}}{E(w, D)} \tag{14}$$

$$C_{c_v}(w) = \sum_{j=2}^{n} [c_v(w, D_n) - c_v(w, D_1)] = c_v(w, D_n) - c_v(w, D_1) \tag{15}$$

$$c_d(w, D) = \frac{V(w, D)}{E(w, D)} \tag{16}$$

$$C_{c_d}(w, D) = \sum_{j=2}^{n} [c_d(w, D_n) - c_d(w, D_1)] = c_d(w, D_n) - c_d(w, D_1) \tag{17}$$



Supplementary Note 1 Fig. 1. (**a**) Correlation between the average number of reads per recognition site in 1 Mb windows centered on the DNA FISH probe shown in **Supplementary Fig. 1a**, and the median normalized distance from the nuclear lamina measured with these probes, for different GPSeq restriction conditions of Exp.1 and Exp.2 (see **Supplementary Table 2** for a detailed description of the experiments). (**b**) Pearson's correlation coefficient (PCC) for the data reported in (a). (**c**) Same as in (b) but showing Spearman's correlation coefficient (SCC) instead. (**d**) Same as in (a), but in 100 kb windows. (**e, f**) Same as in (b-c) but referring to (d). In (a) and (c), for each condition and experiment, n = 68 points (FISH probes) are shown.

Combination of restriction conditions

In the above formulations, we combined different restriction conditions by comparing each digestion time with the immediately preceding one: we refer to this as *adjacent approach*. At

the same time, we also tested two additional ways of combining different conditions: (*i*) the *fixed approach*, which compares each condition with the initial one, (D_j and D_1); and (*ii*) the *two-points approach*, which only combines the last and initial condition, (D_n and D_1). The different formulations, based on these three approaches, are summarized in **Supplementary Note 1 Table 1**. It is important to note how the *adjacent* and *two-points* approaches lead to the same formulation for variability-based centrality estimates, as they are based on differences rather than on ratios.

Type	Name	*Fixed approach*	*Two-points approach*	*Adjacent approach*
Probability-based	$C_{P_s}(w)$	$\sum_{j=2}^{n} \frac{P_s(w,D_j)}{P_s(w,D_1)}$	$\frac{P_s(w,D_n)}{P_s(w,D_1)}$	$\sum_{j=2}^{n} \frac{P_s(w,D_j)}{P_s(w,D_{j-1})}$
	$C_{P_{s_c}}(w)$	$\sum_{j=2}^{n} \frac{P_{s_c}(w,D_j)}{P_{s_c}(w,D_1)}$	$\frac{P_{s_c}(w,D_n)}{P_{s_c}(w,D_1)}$	$\sum_{j=2}^{n} \frac{P_{s_c}(w,D_j)}{P_{s_c}(w,D_{j-1})}$
Variability-based	$C_V(w)$	$\sum_{j=2}^{n} \log\left(\frac{V(w,D_j)}{V(w,D_1)}\right)$	$\log\left(\frac{V(w,D_n)}{V(w,D_1)}\right)$	
	$C_{c_v}(w)$	$\sum_{j=2}^{n} \left(c_v(w,D_j) - c_v(w,D_1)\right)$	$c_v(w,D_n) - c_v(w,D_1)$	
	$C_{c_d}(w,D)$	$\sum_{j=2}^{n} \log\left(\frac{V(w,D_j)}{V(w,D_1)}\right)$	$c_d(w,D_n) - c_d(w,D_1)$	

Supplementary Note 1 Table 1. Different centrality formulations, either probability- or variability-based, using different approaches to combine different restriction conditions of the same GPSeq experiment.

One of the most important effects of combining multiple conditions using one of the approaches described above is that this intrinsically normalizes scores across conditions. In other words, if a genomic window, w is characterized by a bias towards higher or lower read counts (*e.g.*, due to more or less chromatin accessibility or to biased restriction) such bias will be corrected by the normalization itself, provided that it has a multiplicative (*i.e.*, linear) effect on the number of reads, N_R or on the average number of reads, E or on the read variance, V.

Recognition site domain

Given that a GPSeq experiment comprises multiple conditions and that typically only a fraction of all the restriction sites that have been cut are sequenced in the same run, the sets of recognition sites identified in different conditions only partially overlap. In other words, even if cut, a recognition site might yield sequencing reads only in one condition, but not in another one. We can envision four different ways in which this can be taken into consideration, by defining a different domain of recognition sites to be used in the above centrality formulations.

Let us first define $S(D_j)$ as the set of recognition sites, s that are associated with at least one aligned read in a given condition:

$$S(D_j) = \{s | N_R(s, D_j) \geq 1\} \quad (18)$$

The *universal approach* considers all the genomic recognition sites, independently of whether any read was mapped to them. In this case, the number of sites in a given genomic window is:

$$N_{s_{UNIV}}(w, D_j) = |\{s | s \in w\}| \quad (19)$$

While this approach has the advantage of using the same domain for all the conditions, it also tends to include a large number of empty recognition sites (*i.e.*, sites with no associated mapped read), which skews the centrality calculation.

The *union approach* considers all the genomic recognition sites that have at least one mapped read associated to them in at least one condition:

$$S_{U_{UNION}} = S(D_1) \cup S(D_2) \cup \ldots \cup S(D_n) \quad (20)$$

$$N_{s_{UNION}}(w, D_j) = \left|\{s | s \in w \wedge \exists j \text{ s.t. } N_R(s, D_j) \neq 0\}\right| \quad (21)$$

As for the universal approach, also in this case the same domain is used for all conditions, while decreasing the number of empty recognition sites included in the calculation.

The *intersection approach* considers all the genomic recognition sites that have at least one mapped read associated to them in all the conditions:

$$S_{U_{INTER}} = S(D_1) \cap S(D_2) \cap \ldots \cap S(D_n) \quad (22)$$

$$N_{s_{INTER}}(w, D_j) = \left|\{s | s \in w \wedge N_R(s, D_j) \neq 0, j = 1,2,\ldots,n\}\right| \quad (23)$$

This domain can be readily applied to all conditions, but it has the tendency to discard a large fraction of recognition sites depending on the initial overlap between the different conditions.

Finally, the *separate approach* uses a separate recognition site domain for each condition, defined as the set of recognition sites with at least one mapped read (see (18)):

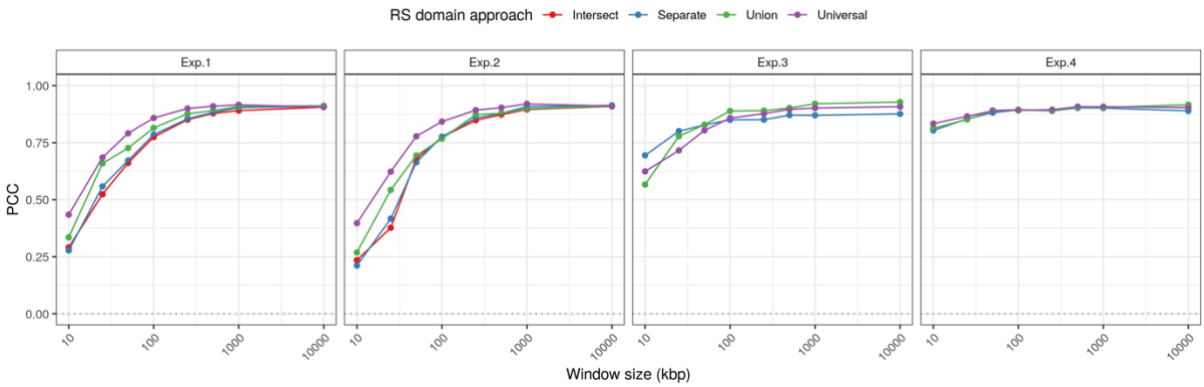
$$N_{s_{SEP}}(w, D_j) = \left|\{s | s \in w \wedge N_R(s, D_j) \neq 0\}\right| \quad (24)$$

While this approach defines a different recognition site domain for each condition, it also discards all empty sites, avoiding any skewing effect that might be caused by the sub-sampling steps.

Definition of GPSeq score

To identify the best centrality score, we first used the above formulations to compute the scores for 1 Mb and 100 kb windows centered on the 68 DNA FISH probes shown in **Supplementary Fig. 1a**, and then compared the scores with the median distance from the nuclear lamina

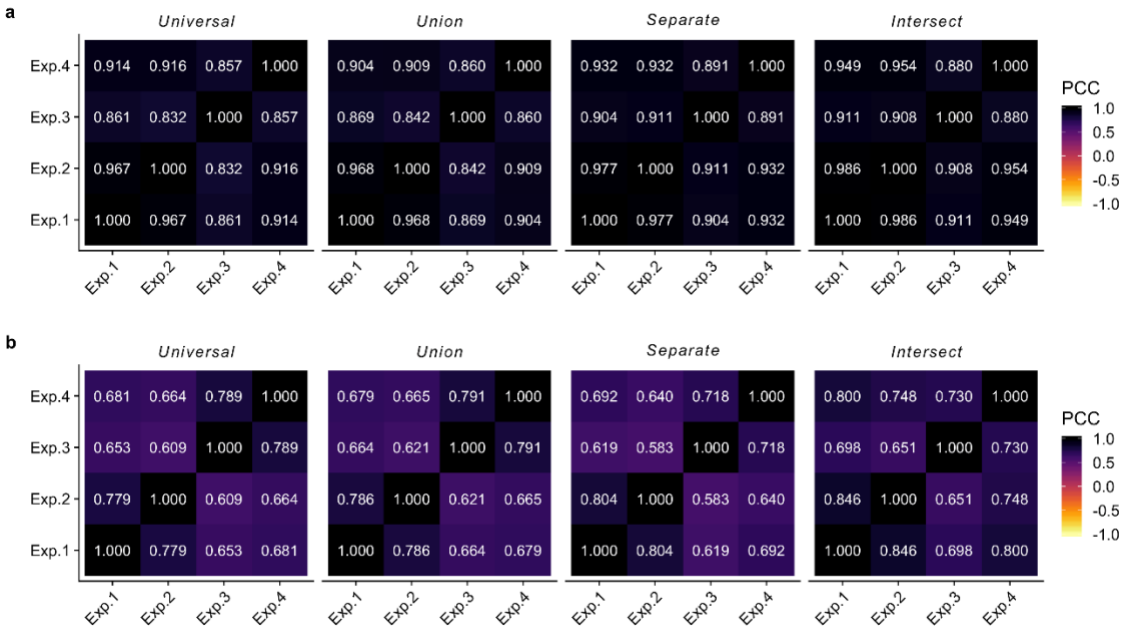
measured with the same probes. Importantly, we rescaled the scores to be able to compare them across different experiments (see "*Score rescaling*" section below). As shown in **Supplementary Fig. 1b**, the rescaled probability-based C_{P_S} formulation outperformed all the others, both at 1 Mb and at 100 kb resolution, yielding higher correlations at almost all resolutions tested. We then compared different recognition site domain approaches described above. The *separate* and *intersection approaches* yielded a higher correlation with DNA FISH in comparison to the *universal* and *union approaches* (**Supplementary Note 1 Fig. 2**). Furthermore, we calculated the genome-wide rescaled score, C_{P_S} for 1 Mb (overlapping with 100 kb steps) and 100 kb windows, using different approaches. Importantly, when calculating the score in a genome-wide fashion, we discarded the windows overlapping with a manually annotate BED file of peri-centromeric and peri-telomeric regions (**Supplementary Table 10**, see "Repeat masking" section in the **Supplementary Methods**). The *separate approach* showed a higher correlation between experiments than any other approach, at both resolutions (**Supplementary Note 1 Fig. 3**). Based on these results, we defined the rescaled centrality score calculated using the separate C_{P_S} approach as the GPSeq score and used it throughout the manuscript. We note that the GPSeq score is free from biases affecting individual genomic windows, as such biases are inherently corrected when the windows are compared across different conditions during the GPSeq score calculation.



Supplementary Note 1 Fig. 2. Pearson's correlation coefficient (PCC) between the centrality score, C_{P_S}, calculated with four different recognition site (RS) domain approaches, and the 3D distance to the nuclear lamina measured by DNA FISH, with window size of 1 Mb centered on the DNA FISH probes (n = 68) shown in **Supplementary Fig. 2a**.

GPSeq score rescaling

To rescale the GPSeq score in order to make it comparable across different experiments and resolutions, we first normalize the scores to have their values ranging between 0 to 1 (excluding outliers, defined as data points with a distance from the closest quartile (1st or 3rd) equal to or greater than 1.5 times the interquartile range). Then, to avoid negative values, we raise 2 to the power of the score, shifting non-outliers scores to the [1, 2] interval, using the default values *--score-outliers* and *--score-outlier-limit* of the *gpseqc_estimate* script described in the **Supplementary Methods**.



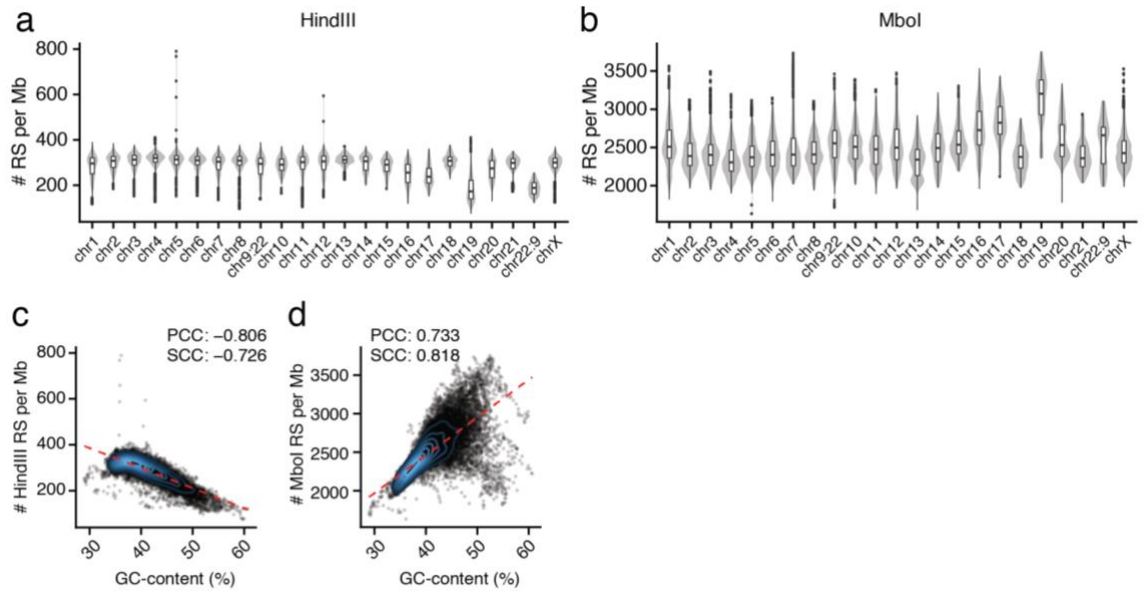
Supplementary Note 1 Fig. 3. (**a**) Pearson's correlation coefficient (PCC) between the GPSeq score calculated using different recognition site domain, for 1 Mb genomic windows overlapping with a 100 kb step. (**b**) Same as in (a), but for 100 kb non-overlapping windows. Sample size information for (a, b) is available in **Supplementary Table 12**.

Supplementary Note 2. Considerations on sequencing depth and genomic resolution

Here, we aim at recapitulating the interconnected effects that the GPSeq experiment design, sequencing depth, and genomic resolution have on the GPSeq score. We first discuss the crucial steps during the design and execution of a GPSeq experiment. We then consider the effects of sequencing depth and resolution on the distribution of GPSeq reads and their statistics and more generally on the GPSeq score.

Choosing the restriction enzyme

The first crucial step when designing a GPSeq experiment is choosing the right restriction enzyme. In this study, we have used two different restriction enzymes (the 4-base cutter MboI and the 6-base cutter HindIII), which have a different frequency and genomic distribution of recognition sites (RS) (**Supplementary Note 2 Fig. 1a, b**). Using a 4-base cutter implies that a larger number of genomic loci can potentially yield fragments to be sequenced. Thus, using a 4-base cutter requires a higher sequencing depth compared to using a 6-base cutter. This also implies that the density of RS will be higher in any genomic window, potentially allowing to achieve a higher resolution in radiality maps. In general, when aiming at a resolution higher than 100 kb (*e.g.*, 10 kb), we recommend using a 4-base cutter, which requires deeper and more costly sequencing. When a lower resolution is still acceptable (*e.g.*, 250 kb), we recommend using a 6-base cutter, which allows for a more shallow and cost-effective sequencing. The actual distribution of the RS also plays an important role. First and foremost, it is essential for the RS of the chosen restriction enzyme to be distributed as homogeneously as possible along the genome, with the least possible number of genomic windows devoid of RS ("RS deserts"). An RS desert would not be able to generate any fragment to be sequenced, rendering impossible the calculation of the GPSeq score for that region. The number of RS available in any genomic window is already accounted for by the GPSeq score formulation, as described in **Supplementary Note 1**. Moreover, the GPSeq score definition inherently corrects for any linear effect that the number of RSs might have on the number of reads. For example, HindIII RS (AAGCTT) correlate with GC-content (**Supplementary Note 2 Fig. 1c**), which would lead to higher number of reads in high GC-content regions. Anti-correlation is instead observed in the case of MboI RS (GATC) (**Supplementary Note 2 Fig. 1d**), leading to a lower number of reads in high GC-content regions. Nonetheless, we note that HindIII (Exp.1 and 2) and MboI experiments (Exp.3 and 4) were highly correlated at different resolutions (**Extended Data Fig. 2h-j**).



Supplementary Note 2 Fig. 1. (**a**) Number of HindIII restriction sites (RS) per 1 Mb genomic window, separately for each chromosome. (**b**) Same as in (a), but for MboI. In all the boxplots inside the violin plots, each box spans from the 25th to the 75th percentile and the whiskers extend from –1.5×IQR to +1.5×IQR from the closest quartile, where IQR is the inter-quartile range. Dots: outliers (data falling outside whiskers). (**c**) Correlation between the GC-content and the number of HindIII recognition sites (RS) per 1 Mb genomic windows overlapping with 100 kb step (masked based on the regions listed in **Supplementary Table 10**). *n* = 26,350 genomic windows (points) were analyzed. Dashed red line: linear regression. Density contours are shown as concentric curves. (**d**) Same as in (d), but for MboI.

Choosing the number of restriction times

In this study, we have used a specific set of digestion times, which differ in between the two enzymes that we used. Our choice of the digestion times was directed by the empirical observation of how the restriction enzyme diffuses throughout the nucleus, as revealed by YFISH. In the case of MboI, we observed a clear fluorescent band at the nuclear periphery already after only one minute of incubation in the presence of the enzyme, whereas in the case of HindIII this required a longer time (10 min). During the course of GPSeq development, we noticed that the YFISH signal pattern no longer changed beyond a certain duration of incubation with a given restriction enzyme (2 hours for HindIII; 30 min for MboI). In fact, when we compared the GPSeq score calculated using all the digestion times up to 2 hours with the GPSeq score obtained using all the times up to 6 hours, the two scores did not differ in any significant manner even when substituting the 2-hour sample with the 6-hour sample. We conclude that, as long as the longest digestion yields a YFISH signal homogenously spread throughout the whole nuclear volume, the precise timing of it is not critical. We strongly

recommend to always empirically determine the set of digestion times for a given enzyme, using YFISH prior to any new GPSeq experiment. This is particularly important when applying GPSeq to a new cell line that has not been assessed by YFISH before. The number of incubation times in between the two extremes depends on how many clearly distinguishable YFISH signal patterns one is able to observe for a given cell type (the patterns should vary from a very thin fluorescence layer confined at the nuclear periphery to a homogenous staining throughout the nucleus). However, we note that, adding more times in between the shortest and longest incubations, does not substantially improve the correlation between radiality estimates by GPSeq and DNA FISH measurements (data not shown). Therefore, as a rule of thumb we suggest using 4–5 incubation times in total. We note however that this might need to be adjusted when aiming to achieve higher resolutions than in this study.

Amplification steps during library preparation

The GPSeq protocol includes two amplification steps: *in vitro* transcription (IVT) and polymerase chain reaction (PCR) (see step-by-step protocol at **Protocol Exchange**, DOI: 10.21203/rs.3.pex-570/v1). IVT allows for linear amplification of the genomic fragments of interest, while PCR is used to attach Illumina sequencing adapters and indexes to the IVT-amplified fragments. At the same time, PCR exponentially amplifies DNA sequences in a potentially biased manner. In other words, PCR might lead to the over-representation of certain fragments based on their sequence. We minimize this issue by performing a low number of PCR cycles during the GPSeq library preparation (typically, 9 PCR cycles for 50–300 ng of IVT input, see **Supplementary Table 2**). To maximize the complexity of the final sequencing libraries, we also recommend splitting the PCR reactions into multiple tubes (e.g., 4–5 25 µl PCR reactions for 300 ng of IVT input in the case of HindIII or 50 ng IVT input in the case of MboI).

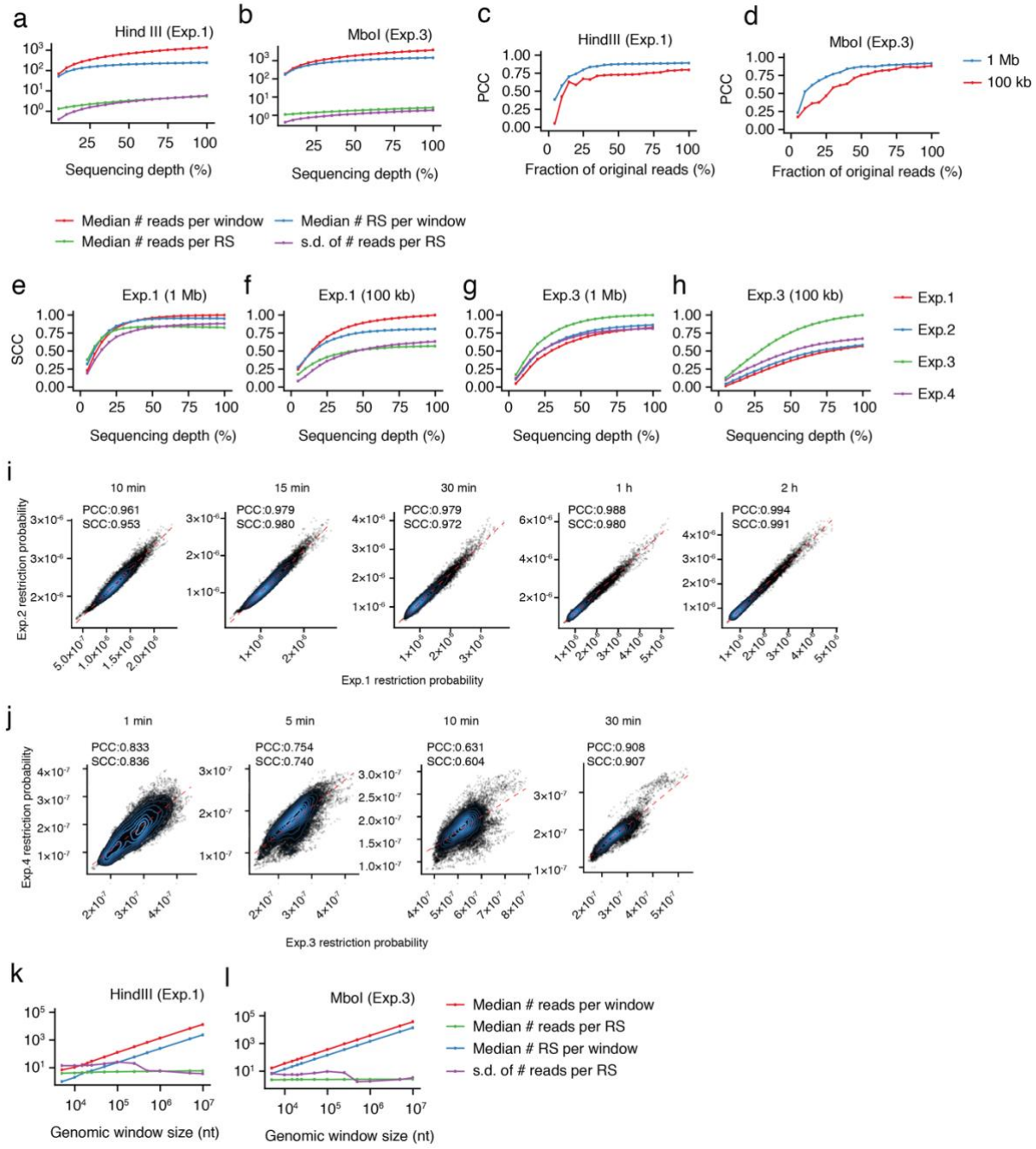
Effects of sequencing depth on the GPSeq score

Once the proper restriction enzyme has been chosen and GPSeq libraries have been prepared, they need to be sequenced. The choice of the sequencing platform is crucial as it defines the maximum achievable sequencing depth. In this study, we sequenced all our GPSeq libraries on the NextSeq 550 sequencing platform from Illumina, which in single-end mode typically yields 400–500 million reads (75 nt). To investigate the effect of different sequencing depths on GPSeq, we simulated different sequencing depths by iteratively removing 5% of the original reads (**Supplementary Methods**). The depth of sequencing will first and foremost affect the

number of reads and their overall statistics. We found that the number of reads and the number of cut RS per 1 Mb genomic window scaled linearly with the sequencing depth, both in the case of a 6-base cutter (Exp.1) and 4-base cutter (Exp.3) (**Supplementary Note 2 Fig. 2a, b**). At the same time, the read distribution across the RS (in terms of median and standard deviation) remained largely unaffected, until reaching a depth so shallow to hinder any type of analysis (*e.g.*, a depth lower than 20% of the original number of reads, see **Supplementary Note 2 Fig. 2a, b**).

The effect of low sequencing depths on the reads will, in turn, affect the GPSeq score calculation. We first investigated how a shallower depth affects the correlation between the distance from the nuclear lamina measured by DNA FISH and the GPSeq score calculated for 1 Mb windows centered around the DNA FISH probes. The correlation remained stable until the depth was lowered to 25% and 50% of the original for HindIII and MboI, respectively (**Supplementary Note 2 Fig. 2c, d**). At sequencing depths above the aforementioned values, the correlation remained stable, indicating that the achieved depth was well over saturation at a resolution of 1 Mb.

Next, we focused on the effect of lower sequencing depths on the GPSeq score, calculated in a genome-wide fashion using genomic windows of 1 Mb (overlapping with a step of 100 kb) or 100 kb. To this aim, we investigated how a reference GPSeq experiment at decreasing sequencing depth correlates to the four other experiments, individually and at full depth. The results showed a consistent decrease in the inter-experiment correlation when decreasing the sequencing depth. In the case of HindIII experiments, the correlation appeared to be stable at depths above 50% and 75% of the original at 1 Mb and 100 kb resolution, respectively (**Supplementary Note 2 Fig. 2e, f**). On the other hand, MboI experiments did not show a stable correlation at any depth at 100 kb resolution and were correlated only above 75% depth at 1 Mb resolution (**Supplementary Note 2 Fig. 2g, h**). This is due to the fact that MboI libraries have a higher molecular complexity requiring a higher sequencing depth. This was further confirmed by inspecting the read count profiles corresponding to individual digestion times, which showed that the inter-experiment reproducibility was lower for MboI compared to HindIII, especially in the case of shorter digestion times (**Supplementary Note 2 Fig. 2i, j**). These results underscore the need for a higher sequencing depth when using higher-frequency cutters or when aiming for higher genomic resolutions.



Supplementary Note 2 Fig. 2. Effect of sequencing depth and enzyme cutting frequency on the GPSeq score. (**a, b**) Median number of reads per genomic window, median number of restriction sites (RS) per genomic window, median number of reads per RS, and standard deviation (s.d.) of the number of reads per RS, in two experiments at different simulated sequencing depths (1 Mb overlapping genomic windows with 100 kb steps). (**c, d**) Pearson' s correlation coefficient (PCC) between the GPSeq score and the median distance from the nuclear lamina measured by DNA FISH, in two experiments at different simulated sequencing depths. n = 68 FISH probes were used (see **Supplementary Fig. 1a**). (**e, h**) Spearman's correlation coefficient (SCC) between the GPSeq score calculated based on one experiment and the GPSeq score calculated based on three other GPSeq experiments, at different simulated sequencing depths and resolutions. Note that for the reference experiment (indicated in the title of each plot) all the sequencing reads were used. The experiments shown in the figure are described in detail in

Supplementary Table 2. (**i**) Correlation between the restriction probability (see Eq. (7) above) calculated for five restriction times (conditions) in two HindIII experiments (Exp.1 and 2), using 1 Mb overlapping genomic windows with 100 kb steps. Each dot represents a single 1 Mb genomic window. PCC and SCC represent the Pearson's and Spearman's correlation coefficient, respectively. Dashed red lines: linear regression. Density contours are shown as concentric curves. n = 26,350 genomic windows (points) were analyzed. (**j**) Same as in (i), but for two MboI experiments (Exp.3 and 4). (**k, l**) Median number of reads per genomic window, median number of restriction sites (RS) per genomic window, median number of reads per RS, and standard deviation (s.d.) of the number of reads per RS, in two experiments at different resolutions. Sample size information for (a, b), (e-h) and (k, l) is available in **Supplementary Table 12**. All the source data for this figure are from HAP1 cells.

Effects of genomic resolution on the GPSeq score

As discussed above, depending on the desired genomic resolution, it is crucial to choose an appropriate enzyme and sequencing platform to achieve the necessary sequencing depth. In the case of HindIII and MboI, the distribution of read counts along the RS (in terms of average and standard deviation) is largely unaffected by the genomic resolution (**Supplementary Note 2 Fig. 2k, l**). We have investigated the effect of higher genomic resolutions on the correlation between the GPSeq score and the distance from the nuclear lamina measured by DNA FISH (**Fig. 2f**). In the case of HindIII (Exp.1-2), the correlation drastically decreased for resolutions higher than 100 kb. On the other hand, the correlation decreased less steeply in the case of MboI (Exp.3), with a Pearson's correlation coefficient above 0.75 even at 10 kb resolution in the case of one experiment (Exp.4) with higher sequencing depth.

Conclusions

The desired genomic resolution, choice of enzyme, and achievable depth (which depends on the available sequencing platform) are critical aspects that should be considered whenever designing a new GPSeq experiment. The higher the desired resolution, the higher the cutting frequency of the restriction enzyme should be, thus requiring a higher sequencing depth. The accuracy of the GPSeq score — determined by comparing it to direct measurements of radial distances by 3D DNA FISH — depends on the frequency along the genome of the RS of the restriction enzyme in use. On the other hand, the precision of the GPSeq score — assessed by evaluating its reproducibility across multiple experiments — depends on the coverage of the RS of the restriction enzyme in use. This means that, given a sufficient sequencing depth, a more frequent cutter will allow achieving a more accurate radiality estimate at the same genomic resolution. As a rule of thumb, a 6-base cutter such as HindIII is sufficient to obtain reproducible radiality maps at 100 kb resolution, using 5–6 different restriction times and

sequencing all the samples in a single sequencing run on Illumina's NextSeq 500 (400–500 Mreads). Using a 4-base cutter such as MboI allows to further increase the accuracy of the radiality estimates, although a higher sequencing depth is needed to achieve high reproducibility at the same resolution.

6. Supplementary References

1. Gelali, E. *et al.* iFISH is a publically available resource enabling versatile DNA FISH to study genome architecture. *Nat. Commun.* **10**, 1636 (2019).
2. Solovei, I. & Cremer, M. 3D-FISH on cultured cells combined with immunostaining. *Methods Mol. Biol. Clifton NJ* **659**, 117–126 (2010).
3. Gelali, E. *et al.* An Application-Directed, Versatile DNA FISH Platform for Research and Diagnostics. *Methods Mol. Biol. Clifton NJ* **1766**, 303–333 (2018).
4. Kodiha, M., Umar, R. & Stochaj, U. Optimized immunofluorescence staining protocol to detect the nucleoporin Nup98 in different subcellular compartments. (2009).
5. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
6. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinforma. Oxf. Engl.* **34**, 3600 (2018).
7. Kind, J. *et al.* Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134–147 (2015).
8. Zhong, Y.-F. & Holland, P. W. H. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evol. Dev.* **13**, 567–568 (2011).
9. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet. TIG* **29**, 569–574 (2013).
10. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).
11. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
12. Xiong, K. & Ma, J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat. Commun.* **10**, 5069 (2019).
13. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
14. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
15. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
16. Gothe, H. J. *et al.* Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations. *Mol. Cell* **75**, 267-283.e12 (2019).
17. Yan, W. X. *et al.* BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat. Commun.* **8**, 15058 (2017).