UNIVERSITAT DE BARCELONA

FACULTAT DE BIOLOGIA

DEPARTAMENT DE GENÈTICA, MICROBIOLOGIA I ESTADÍSTICA

Programa de Doctorat en Biomedicina

Determining the Three-dimensional Structure of Genomes and Genomic Domains Integrating Chromosome Conformation Capture Data and Microscopy Images

Memòria presentada per

David Castillo Andreo

per optar al grau de doctor per la Universitat de Barcelona

Treball realitzat al Centre de Nacional d'Anàlisi Genòmica (CNAG)

Doctorand	Director	Tutor
David Castillo Andreo	**Marc A. Martí-Renom**	**Modesto Orozco López**

Acknowledgments

Thanks to my director Marc for all the support and friendship and all past and present *Marcians*: Julen, François, Silvia, Aleks, Fra, María, ... Special thanks to Marco and Irene for the beautiful *colazia's* which are surely the source of my scarce scientific knowledge. Thanks also to my fellows Huy and Shyamtanu on the other side of the ocean.

Y gracias a mi familia por aguantarme; a mi pareja Carmen, a mi hijo cantarín Lucas y a la hija mas bella del mundo, Laia.

En memoria de mi abuelo Juan Antonio al que siempre apasionó el conocimiento.

Abstract

Microscopy and Chromosome Conformation Capture (3C) are the two main techniques for studying the three-dimensional (3D) organization of the genome. Microscopy, allowing the visualization of genomic *loci* in individual nuclei, pioneered the field of structural genomics and became the gold-standard for the validation of new discoveries. 3C and 3C-based techniques, identifying the number of contacts between pairs of genomic *loci*, have already been key to unveil the importance of the 3D genome organization in many cellular processes. Both techniques are continuously evolving pushing forward the technologies and giving rise to innovative assays that require the support of new computational methods for data collection, analysis and modeling.

In this thesis, I have contributed to provide these essential computational methods to the Structural Genomics community. In Microscopy, I participated in the design and implementation of OligoFISSEQ, a novel multiplexing imaging technology to visualize multiple genomic regions in hundreds and thousands of individual cells. In 3C-based techniques, I contributed to the development of a tool for the reconstruction of the 3D organization of chromatin from highly-sparse 3C-based datasets (*e.g.* Promoter Capture Hi-C). Finally, I have introduced pTADbit, a novel approach for the reconstruction of the 3D Genome organization integrating both Microscopy and 3C data via the application of Machine Learning methods.

Table of Contents

General Introduction

Nuclear Organization

Each human cell contains around 2 meters of DNA packed inside its nucleus which has a diameter of approximately 10-15 micrometers. Despite this level of compaction, DNA folds and unfolds efficiently allowing the dynamic interactions that are essential for the transcriptional and regulatory processes occurring inside the nucleus. Such processes are only possible by virtue of a highly organized nuclear structure. Furthermore, although containing identical genomic sequence, nuclei of different cell types exhibit specialized types of processes which are only feasible by the acquirement of specific organizations (Winick-Ng et al. 2021)

The first level of organization of the DNA is the wrapping of the double helix molecule around the histone proteins forming nucleosomes and preventing the stretching and possible breakages of the chain (van Emmerik et al. 2019). Free linker DNA connects adjacent nucleosomes constituting a "beads-on-a-string" 10-nm structure that we refer as the chromatin fiber. The specific positions of the nucleosomes and linker DNA in the chromatin denotes a further degree of organization and compaction. The spacing between consecutive nucleosomes differ between cells and regions and it influences gene expression by controlling DNA accessibility of many binding proteins and regulatory elements (Bai et al. 2010). The 10-nm chromatin fiber is further compacted and organized in higher-order structures up to 30 nm width in conformations that are still focus of study.

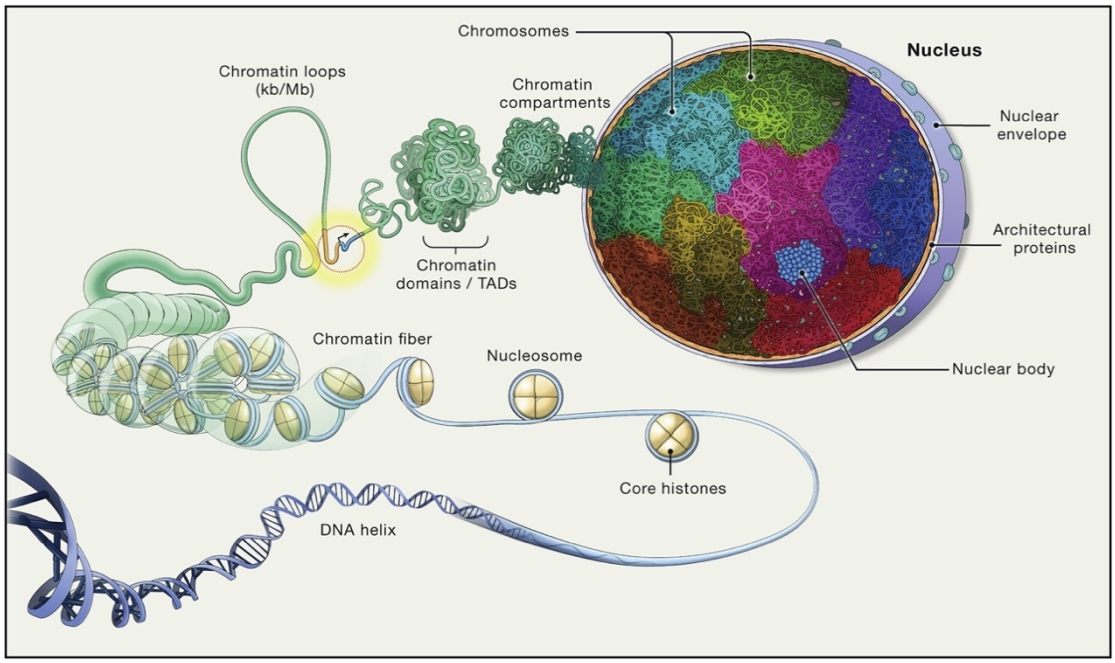


Figure 1. The Organization of the Eukaryotic Genome. DNA is hierarchically organized, first wrapped around histone proteins assembled in nucleosomes, forming all together the chromatin fiber. Chromatin folds into loops, often bringing gene regulatory elements (yellow), like enhancers, into proximity to promoters of genes (gold/blue) to control their transcription (black arrow). Then it is further organized in chromatin domains, referred as TADs, and in larger domains referred as compartments. Each chromosome occupies distinct volumes, or chromosome territories inside the cell nucleus. The nucleus also contains RNA and protein aggregates which form nuclear bodies (blue). Figure from (Misteli 2020).

In mammalian cells, evidences exist that cohesin, a ring-shaped structural maintenance of chromosome (SMC) complex, extrude the chromatin fiber until it finds a CCCTC-binding factor (CTCF) site in what is known as the loop-extrusion model (Sanborn et al. 2015). The mechanism would produce the folding of the chromatin in globular 3D conformations that have been observed in single cells imaging experiments (Bintu et al. 2018). Although the high variability between cells (Nagano et al. 2013; Szabo et al. 2020), a preferential position of the borders of the blobs of chromatin in CTCF motifs explains the chromatin organization of Topologically Associating Domains (TADs) and loops observed in population-based Chromosome Conformation Capture (3C) experiments (Lieberman-Aiden et al. 2009; Dixon et al. 2012; S.S. Rao et al. 2014).

TADs were identified in the context of Hi-C and 5-C assays, 3C-based experiments (Dekker et al. 2002). TADs have been defined as domains of contiguous sequential regions of dense self-interactions and have been highlighted as architectural chromatin units in many cellular contexts (Nora et al. 2012; S.S. Rao et al. 2014). TADs are stable across different cell types and highly conserved across mammalian species and often

isolate and constraint the essential interactions for gene regulation, for example between enhancers and gene promoters (Bonev et al. 2017; Zhan et al. 2017). During differentiation, the likelihood of co-regulation of genes located within the same TAD is maximized (Zhan et al. 2017) and during replication, TADs are stable units of replication timing (Pope et al. 2014).

At larger scales, Hi-C has also revealed the existence of another layer of organization constituted by the preferential long-range interaction of domains that segregates chromosomes in two different compartments, referred as "A" and "B" (Lieberman-Aiden et al. 2009). The A compartment contains active and open chromatin (euchromatin) while the B compartment inactive and close chromatin (heterochromatin). Later studies have further subdivided those compartments by the association of different epigenetic marks (S.S. Rao et al. 2014; Xiong et al. 2019; Liu et al. 2021; Vilarrasa-Blasi et al. 2021).

Cohesin depletion experiments has shown the abolishment of the preferential position of TAD boundaries at CTCF sites but also the prevalence of TAD-like structures and compartments in bulk Hi-C and single cells (S.S.P. Rao et al. 2017; Schwarzer et al. 2017; Bintu et al. 2018), which indicates the existence of multiple mechanisms complementary to loop-extrusion in chromatin folding. One of such mechanisms that has been proposed to shape the structure of the genome is phase-separation. In this thermodynamic process, a high-concentrated macromolecule in a mixture is partitioned into two or more distinct phases with different physical and chemical properties. During the process, the macromolecules condense into a dense phase characterized by the formation of liquid-like compartments which coexists with a dilute phase (Banani et al. 2017; Alberti et al. 2019). Indeed, membrane-less assemblies in the form of liquid-phase condensates have been shown to be involved in transcriptional control (Boija et al. 2018; Sabari et al. 2018) and gene regulation (Larson et al. 2017; Guo et al. 2019).

Polymer models strengthen the idea that both mechanisms, loop-extrusion and phase separation, co-exist having complementary tasks in the shaping of chromatin architecture. For instance, loop-extrusion establishing TAD borders and loop interactions and phase separation segregating different regions and creating less variable regulatory structures (Nuebler et al. 2018; Conte et al. 2021).

Chromosomes occupy defined regions of the genome called Chromosome Territories (CT) (Cremer et al. 2001) with gene-rich chromosomes preferentially positioned towards the nuclear interior, whereas gene-poor chromosomes preferentially positioned towards the periphery (Tanabe et al. 2002). During interphase, long-range chromatin repositioning occurs only during a relatively short time window, after which chromatin movements are constrained within small nuclear subdomains (Walter et al. 2003). CTs intermingle significantly creating regions that have been shown to potentially contain transcription factories and shape the chromatin structure (Branco et al. 2006). One of the most visible and notorious example of nonrandom inter-chromosomal assembly in human nuclei is the formation of the nucleolus (Figure 1) in which five different acrocentric chromosomes come into physical proximity in the human genome (Maass et al. 2018).

Surrounding the nucleus, the lamina constitutes the outer functional organization of the genome with the association of chromatin in lamina-associated domains (LADs). LADs are blocks of chromatin ranging from 50kb to 10Mb in size that are in close proximity to the nuclear lamina. They represent a strongly repressive chromatin type exhibiting heterochromatic features, including low gene density, low transcriptional activity and late replication timing (Buchwalter et al. 2019). If we aggregate the contacting regions with the lamina over multiple individual cells, LADs account for approximately 35% to 40% of the mammalian genome (Guelen et al. 2008). In reality, given the stochastic nature of the contacts between LADs and the lamina, any given *locus* may be contacting the lamina only in a subpopulation of cells.

All these highlighted genomic features reveal a complex and hierarchical organization of the genome emerging from the stochastic nature of the chromatin movements. A coordinated system of segregation and congregation in which compartmentalization is key to guarantee the correct regulation and function of the nuclear processes. An organization flexible enough to ensure the creation of the needed micro-environments in which the most essential and conserved gene transcription occur but also able to facilitate the creation of more transient domains.

Methods to study chromatin structure

Microscopy

Microscopy has been one of the main techniques traditionally used to study the nuclear structure. However, it has been long time undermined by the low statistical power of low-throughput techniques both in terms of the number of *loci* that can be visualized in each individual cell and in the number of cells that can be analyzed in a single sample. It is only now that the latest developments in the field (see next sections), to which I contributed (Nguyen et al. 2020), have allowed us to reach the required resolution and statistical power to study in detail the organization of chromatin.

New imaging techniques have been designed to capture the high variability observed in the cell population by increasing the number of *loci* detected in each individual nucleus and optimizing the processes to provide large numbers of imaged cells. A key breakthrough in the field have been the development of array-based oligonucleotide synthetic probes, being Oligopaints (Beliveau et al. 2012) the variant used by the majority of imaging methods in the field due to its improved computational design and probe synthesis. Oligopaints are computationally designed DNA sequence-specific probes with additional non-genomic sequences (streets) that enable additional functionalities, including amplification, indirect labeling, barcode-based multiplexing, and sequential and combinatorial labeling.

The following technologies deserve special mention in the field of multiplexed genomic imaging:

- Multiplexed diffraction-limited FISH (Wang et al. 2016; Bintu et al. 2018): it relies on the sequential labeling and imaging of multiple DNA *loci* with Oligopaints. The oligonucleotides targeting each genomic locus contain unique streets with specific sequences or barcodes that could be independently read by complementary, fluorescently labeled oligonucleotide readout probes. The amplification of the signal is achieved by the aggregation of hundreds of probes per target that share the same barcode. Strand displacement and photobleaching is used to extinguish the signal between rounds.

The technique has been used to study the conformation of regions of approximatively 2Mbp regions in around 40 sequencing rounds using two imaging channels.

- Optical reconstruction of chromatin architecture (ORCA) (Mateo et al. 2019): regions of interest (100-700 kb) are tiled in short sections (2–10 kb) targeted by primary Oligopaint probes with unique barcodes that are similar to the ones used in multiplexed error-robust fluorescence in situ hybridization (MERFISH) that allows the measurement of hundreds to thousands RNA molecules within a single cell (K.H. Chen et al. 2015).
The primary probes are labelled with fluorophores and imaged sequentially. One of the streets is labelled with a fiducial fluorophore (fiducial oligo) that improves the registration of the images through the sequencing rounds and therefore enhance the genomic resolution. The readout probe (readout oligo) binds to the barcode sequence in the other street, is imaged together with the fiducial oligo and subsequently removed by strand displacement. The process is repeated for each barcode. Each barcoded region contains at least 20 primary probes allowing the resolution of the targets in diffraction-limited images.
ORCA improves the resolution attained by other methods by focusing on specific regions of interest and improving the registration of round-to-round images. The use of strand displacement for the removal of the imaged readout oligo allows for repeated measurements of the same barcode which is especially useful for error quantification and correction.

- Hi-M (Cardozo Gizzi et al. 2019): it also relies on the sequential labeling and imaging of multiple DNA loci with Oligopaints. The probes in this technique contain a cleavable bond allowing the elimination of the fluorescence signal of a particular barcode from one cycle to the next. The amplification of intensity in the images is achieved by the aggregation of hundreds of oligos per detected target. By multiple sequential cycles of hybridization, washing, and imaging of each barcode, Hi-M has proven to simultaneously label around 20 different loci.
The incorporation of combinatorial labeling schemes should make it possible to considerably increase the number of detected loci without increasing the number of hybridization cycles.

- OligoFISSEQ (Nguyen et al. 2020): is the combination of fluorescent in situ sequencing (FISSEQ) technologies developed originally for in situ transcript localization and quantification (Je Hyuk Lee et al. 2015) and Oligopaints. In OligoFISSEQ, barcodes

are embedded in the Oligopaint streets and sequenced *in situ*. By bringing together hundreds to thousands of identically barcoded Oligopaints to a genomic target, OligoFISSEQ does not require the type of amplification FISSEQ needs; the signal is strong enough to acquire diffraction-limited images using conventional microscopes achieving throughputs on the order of hundreds to thousands of cells and thus providing the statistical power necessary for addressing cell-to-cell variability.
OligoFISSEQ allows also an exponential increase of the number of targeted regions with the number of hybridization rounds, potentially reaching thousands of different loci in 5 rounds and 4 channels.

- DNA-MERFISH (Su et al. 2020): it is also a natural extension of MERFISH that allows the measurement of hundreds to thousands RNA molecules within a single cell (K.H. Chen et al. 2015). In DNA-MERFISH, synthetic single-stranded DNA probes are used to image many chromatin loci simultaneously in each round of sequencing. Each probe incorporates two distinct readout sequences corresponding to the sequencing round in which the locus is expected to be detected. The distinct identities of the loci are determined based on the combinations of rounds in which they appear and that match the designed barcodes. Not all combinations of rounds correspond to a possible barcode increasing the error-robustness of the scheme. Bringing together hundreds of consecutive oligos allows for the use of diffraction-limited images.
The technology has been used to image thousands of genomic loci in 50 rounds of hybridization and 2 color channels per round.

- seqFISH+ (Takei et al. 2021): it is based on seqFISH, a sequential barcoding scheme to multiplex different mRNAs. SeqFISH+ is based on sequential hybridization and the aggregation of hundreds of synthetic single-stranded DNA probes per target to amplify the signal intensity. Primary probes are ligated to the DNA binding sites and padlocked (Nilsson et al. 1994) at the binding sites after the initial hybridization to stabilize them during the sequential rounds. Each primary probe is flanked by the several unique readout probe binding sequences which are hybridized by fluorescent oligos, imaged and stripped over sequential rounds.
The schema allows the identification of 2,460 different loci in 80 rounds of sequential hybridization and 2 channels.

- *In situ* genome sequencing (IGS) (Payne et al. 2021): it uses a different approach to identify multiple targets per cell combining *in situ* and *ex situ* sequencing. The main idea behind the method is the imaging of the positions of genomic *loci* without specifically target DNA motifs and the determination of their genomic sequence *a posteriori*.

To accomplish that, IGS randomly incorporates DNA sequencing adaptors into fixed genomic DNA with Tn5 transposase preserving genomic fragments in their spatial positions (X. Chen et al. 2016). Tn5 transposase selectively inserts the adaptors into accessible chromatin loci in the living cells. Those transposed fragments are circularized *in situ* by the ligation of two DNA hairpins that contain a unique molecular identifier (UMI) and primer sites used for *in situ* and *ex situ* DNA sequencing, followed by the amplification of the circular templates using rolling circle amplification. The spatial positions of the amplicons are determined using sequential rounds of *in situ* sequencing by ligation (SBL) and fluorescence imaging (J. H. Lee et al. 2014). Then, the amplicons are dissociated and amplified using PCR to produce an *in vitro* sequencing library. Finally, *in situ* amplicon positions and *ex situ* paired-end sequencing reads are computationally matched.

IGS allows the simultaneous sequencing and imaging of genomes within intact biological samples, spatially localizing thousands of genomic loci in individual nuclei. Due to its genome-wide sampling frequency (at most ~1 Mb), IGS is currently limited in its ability to examine specific genetic loci at higher-resolutions.

Chromosome Conformation Capture (3C) and derived technologies

Chromosome Conformation Capture (3C) techniques appeared already twenty years ago (Dekker et al. 2002) and have been key to unveil the importance of the 3D structure of the genome in many cellular processes by resolving finer details of the genome structure. 3C-based assays provide rich, high-throughput and genome-wide data describing genome topology and enabling systematic studies at high resolutions.

The principal strategy of 3C techniques to study chromatin topology is the quantification of the contact frequencies between distal *loci* in cell populations. The main steps of 3C-based protocols are similar and consist in chromatin crosslinking using most often formaldehyde, digestion by sonication and/or using restriction enzymes and re-ligation of the resulting sticky ends to form chimeric molecules (Figure 2).

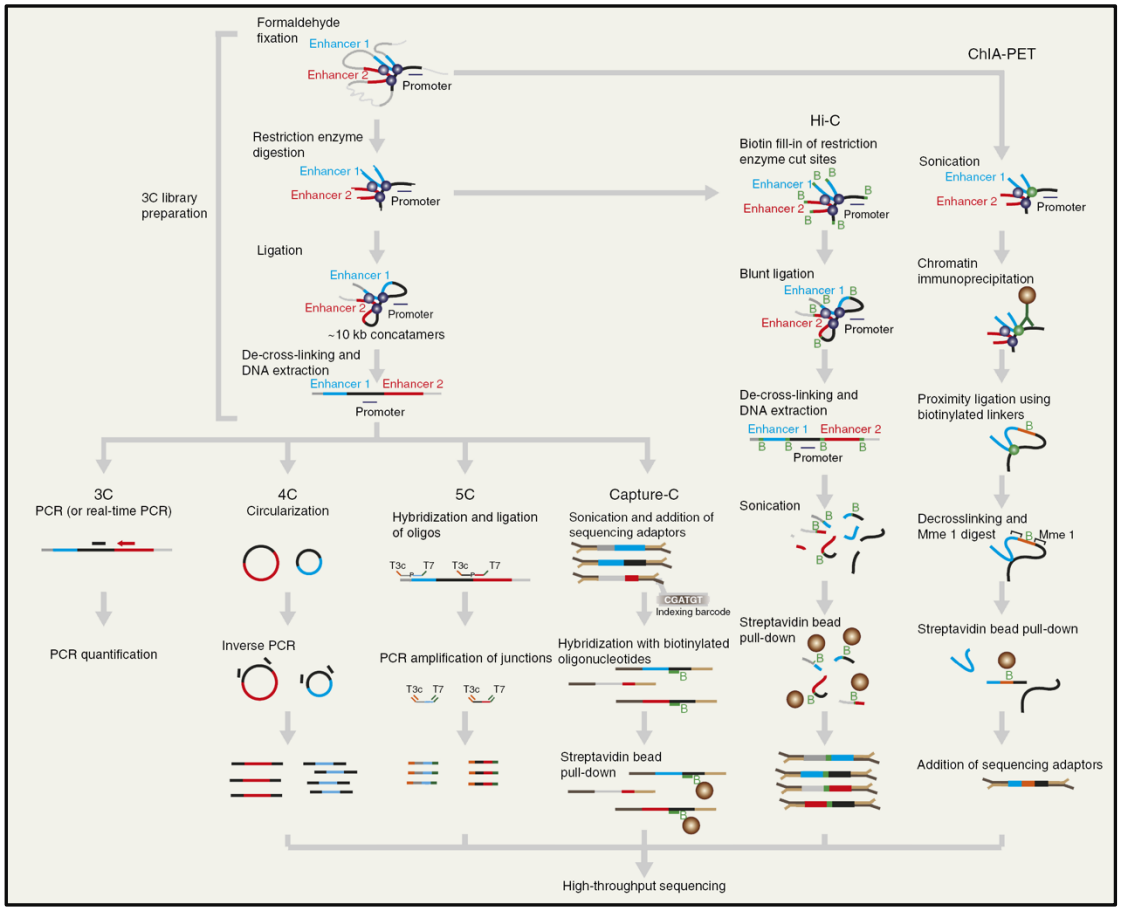


Figure 2. Comparison of different 3C-based methodologies. Figure adapted from (Davies et al. 2017)

After reversing the crosslink, the molecules (3C templates) are amplified by PCR, sequenced and mapped to the appropriate reference genome. The crosslinked DNA fragments may be distant in the genomic sequence, but they are close in 3D space allowing the inference of the chromosomal conformations by counting the number of occasions that those fragments co-occur in the chimeric molecules.

The following techniques are part of the 3C-based family:

- Chromosome Conformation Capture (3C) (Dekker et al. 2002): after the re-ligation of the fragments and using PCR primers designed to amplify specific ligation junctions, 3C can retrieve interactions between two targeted loci in the cell population. The kind of information obtained with this technique is a *one-versus-one* interaction profile. The requirement for PCR primers designed to amplify regions of interest limits the method to the detection of spatial relationships between known DNA sequences.

- Circular Chromosome Conformation Capture (4C) (Simonis et al. 2006; Zhao et al. 2006): the most important innovation of the 4C techniques is that it allows the detection of unknown DNA regions with a targeted region of interest (viewpoint). The 4C protocol differs after reversing the crosslink; a second digestion with a different restriction enzyme creates fragments that can re-ligate and circularize. Primers binding the known DNA fragment are used to amplify the DNA circles containing the viewpoint and its interacting DNA and the result can be analyzed by microarrays or next generation sequencing (NGS).

- Chromosome Conformation Capture Carbon Copy (5C) (Dostie et al. 2006): it enhances the main 3C technique by incorporating special primers designed with oligonucleotides containing universal sequences. Thanks to those sequences, all 3C templates can be simultaneously amplified in a multiplex PCR reaction. The junctions can be analyzed by microarrays or by NGS. The information obtained by the technique is the interacting profiles of a set of continuous regions of interest in a "many versus many" form. The main disadvantage of the 5C protocol is the primer design needed to interrogate the region of interest, making it unfeasible for genome-wide studies.

- Hi-C (Lieberman-Aiden et al. 2009): after the digestion with restriction enzymes the sticky ends are filled with biotin-labeled nucleotides. The 3C templates are re-ligated, sheared and purified by biotin pull-down using streptavidin beads. The purification ensures that only junctions with biotin are selected for high-throughput sequencing. The chimeric reads are mapped to the reference genome allowing the construction of matrices of interactions between all fragments in the genome providing "all versus all" information.
A variation of Hi-C that increases considerably the resolution of the obtained contact matrices is Micro-C (Hsieh et al. 2015) in which micrococcal nuclease is used instead of restriction enzymes to fragment chromatin and obtain single nucleosome resolutions.
Further adaptations of the Hi-C technique have allowed the application of the protocol to individual cells (Nagano et al. 2013; Ramani et al. 2017).

- Chromatin Interaction Analysis by Paired-end Tag sequencing (ChIA-PET) (Fullwood et al. 2009): combines 3C with chromatin immunoprecipitation (ChIP) to study chromatin interactions bound by one specific protein. The 3C templates are

enriched by ChIP using a specific antibody and DNA sequences tethered together and to the protein of interest are re-ligated with oligonucleotide DNA linkers, the sequence of which contains restriction sites for a posterior digestion. The resulting Paired-End Tags (PETs) are sequenced and mapped to the reference genome. This technique provides information of interactions between regions brought together by proteins. An improved version of ChiA-PET is HiChIP (Mumbach et al. 2016) which lower the requirements in terms of number of input cells while achieving better signal-to-background ratios than in situ Hi-C.

- Capture-C (Hughes et al. 2014): combines 3C, NGS and oligonucleotide capture technology (OCT). After the standard 3C experiment, the 3C templates are sonicated and paired-end sequencing adaptors are added. Then, capture probes with biotin hybridize in a set of fragments of interest and are pull-down by streptavidin beads. The captured DNA fragments are amplified and sequenced allowing the generation of genome-wide contact profiles from hundreds of selected loci at a time with a reduction of the costs compared to standard 3C experiments.

The application of the Capture-C strategy can be used to enrich Hi-C libraries in a technique known as Capture Hi-C (CHi-C) (Mifsud et al. 2015) enabling deep sequencing of target fragments and excluding uninformative background.

Three-dimensional modelling of the genome from 3C data

The inherent nature of proximity-based crosslinking of the 3C techniques do not allow a direct measurement of the physical distances between regions in the genome. Instead, they provide a quantification of the frequencies of contact between distal *loci* that is *a proxy* for its spatial distance. The inference of those distances from interaction frequencies is a transformation to which we refer as modelling. The reconstruction of 3D chromatin structures from its interaction data allows the analysis of the genome in metaphase and interphase in its spatial context providing richer information to the scientist.

The modelling strategies used to obtain 3D conformations from interaction data can be divided in two categories: data-driven modeling and thermodynamics-based modelling.

Data-driven modelling

Data-driven modelling methods provide solutions, generally faster than thermodynamics-based approaches, that are compatible with the given input data subject to the constraints of the considered environment. In data-driven models all parameters can be derived from the input data. Generally, data-driven methods adopt simplified representations of chromatin using spheres or points as chromosome regions or *loci* adopting a coarse-grained “beads-on-a-string” configuration.

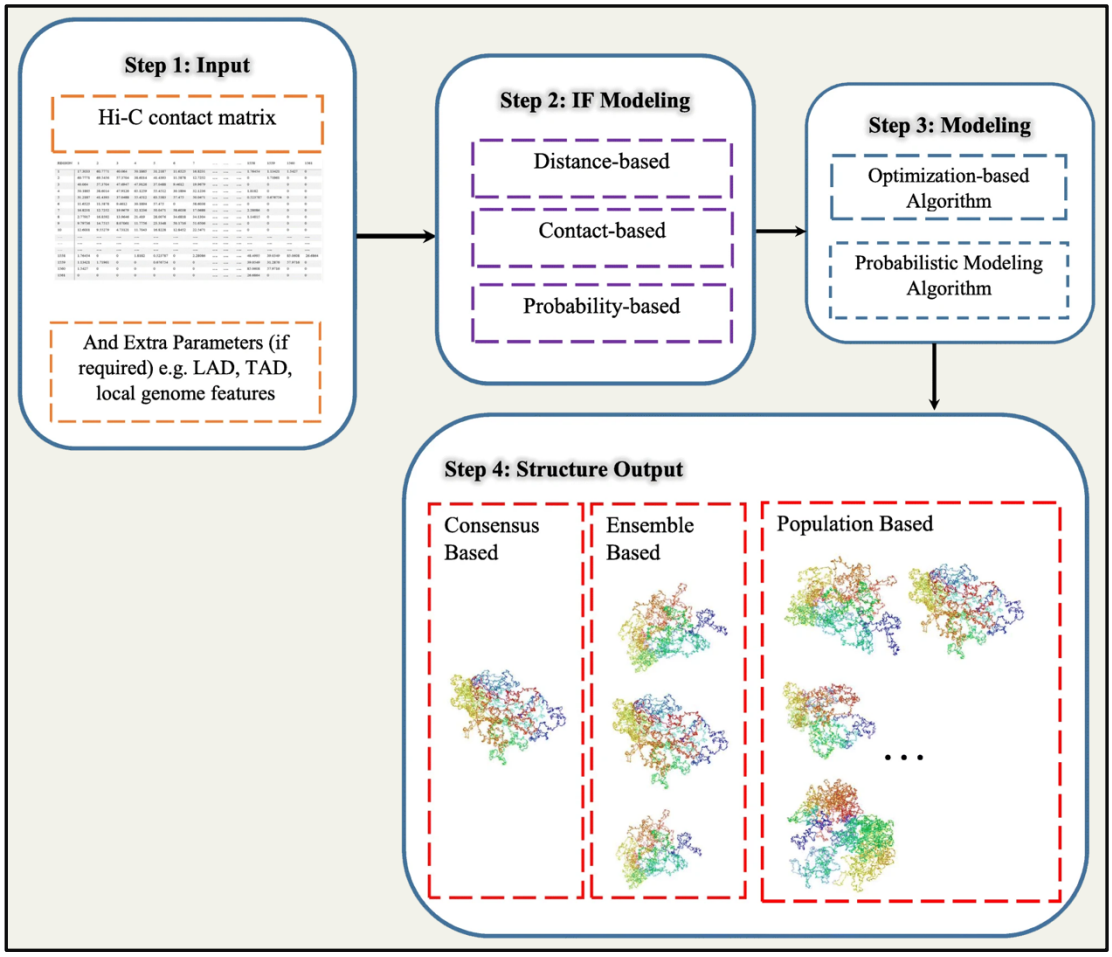


Figure 3. Genomic regions 3D structure reconstruction workflow of data-driven modelling methods. Step 1: The input preparation, usually, 3C-based contact data and sometimes empirical extra parameters. Step 2: The three data-driven modelling approaches depending on the strategy used to model the interaction frequency (IF). Step 3: Structural modelling with each tool defined sampling strategy, and Step 4: generation of a consensus average structure or a group of structures. Figure adapted from (Oluwadare et al. 2019).

According to the type of structures generated data-driven methods can be largely classified in two main groups: consensus-based (Hu et al. 2013; J. Paulsen et al. 2015; Rieber et al. 2017; J. Li et al. 2018; Abbas et al. 2019; F.Z. Li et al. 2020) or ensemble-based (Rousseau et al. 2011; Tjong et al. 2016; Tuan Trieu et al. 2016b; T. Trieu et al. 2016a; Jonas Paulsen et al. 2017; Serra et al. 2017; Zhu et al. 2018; T. Trieu et al. 2019). Consensus-based methods transform the interaction frequencies into a single 3D conformation which, in the case of 3C data produced from a population of cells, represent an average solution of the ill-defined modelling problem. Ensemble-based methods, instead, take into consideration that the information has been produced from an ensemble of cells that could eventually adopt different conformations. Therefore, they explore solution spaces in which individual structures satisfy not all but some of the imposed restraints. Different conformations satisfy different sets of input restraints forming a final

ensemble that satisfy most of the defined restraints. Ensemble-based methods aim to reproduce the heterogeneity of the cell to cell variability of the population.
According to the method used in the modelling, data-driven methods can be divided in three categories: distance-based, contact-based and probability-based methods.

- Distance-based methods (Tuan Trieu et al. 2016b; Jonas Paulsen et al. 2017; Rieber et al. 2017; Serra et al. 2017; J. Li et al. 2018; T. Trieu et al. 2019): are characterized by the initial conversion of the interaction frequencies to physical distances and the reconstruction of the spatial coordinates that satisfy those distances. It is the most followed approach in the determination of 3D structures probably inspired by classical multidimensional scaling (Torgerson 1958). The main differences between the methods in this category are how the interaction frequencies are converted to distances and the method used to infer the coordinates from them. In a distance-based method a 3D structure is initialized and an objective function is used to quantify the difference between the inferred 3D structure and the distances expected from the obtained transformation. The 3D structure is iteratively updated to minimize the objective function using multidimensional scaling or other optimization techniques (J. Paulsen et al. 2015).
Although it is commonly assumed that the interaction frequency between two *loci* is inversely related to its distance, the scaling factor of that relation might be different from organism to organism and even among different cell types. The scaling factor is one of the main parameters that the different methods try to optimize.
Some methods introduce additional restraints obtained empirically like the minimum and maximum distances between adjacent loci, the positions of telomeres and centromeres or the confinement of the nuclear lamina to produce more accurate reconstructions of the 3D models.
One of the main drawbacks of distance-based approaches is that weak interaction frequencies, strongly affected by noise, are normally unreliable for the prediction of long-range distances.

- Contact-based methods (T. Trieu et al. 2016a; Jonas Paulsen et al. 2017; Zhu et al. 2018): this group of methods use the interaction frequencies directly to model 3D structures. As such, they do not require the pairs of regions to satisfy a specific distance. Some methods falling into this category require distances to be below a certain threshold as to simulate the 3C crosslinking (T. Trieu et al. 2016a). Others, model the frequencies

as neighboring affinities to construct interaction networks and derive structures with optimization processes inspired by manifold learning (Zhu et al. 2018).

- Probability-based methods (Rousseau et al. 2011; Hu et al. 2013; Tjong et al. 2016): methods in this category model the interaction frequencies using probabilistic frameworks. Considering that most of 3C-based assays are conducted in cell populations, probabilistic methods are appropriate to consider the outcome data as an average of an undetermined ensemble. The main advantage of these methods is that the uncertainties in the experimental data can be overcome through a probabilistic representation. Systematic biases such as GC content and the uneven distribution of the restriction enzyme cutting sites need to be considered in the probabilistic models.
Probabilistic-based methods infer ensemble of structures, through Bayesian inference (Hu et al. 2013) or maximum likelihood optimization (Rousseau et al. 2011; Tjong et al. 2016), that are statistically consistent with the input data as the best approximation of the underlying true population of structures given the available data.

Thermodynamics-based modeling

The modelling strategies based on thermodynamics apply polymer physics principles to simulate the dynamics of the chromatin fiber. These modelling approaches treat each chromosome as a biopolymer frequently represented as a coarse-grained "beads-on-a-string" model seeking to reproduce and understand the underlying principles of chromatin organization by applying a set of parameters that characterize its global properties and motion. The defined functions and properties governing the chromatin dynamics in the simulations can be known from statistical physics or hypothesized from empirical observations. With them, polymer physics models have been successfully used to simulate chromosome folding at large-scales.

Different types of polymer models have been proposed to explain the observed behavior of chromatin in the nucleus. Before the birth of Hi-C, a densely knotted and compact conformation in equilibrium referred as "globule" was commonly proposed to simulate chromatin (Münkel et al. 1998). But the measure of the contact probability of two intra-chromosomal loci depending on their genomic distance, brought the “fractal globule” model to the scientists’ attention. The fractal globule is a long-lived and non-equilibrium

state proposed in the nineties (Grosberg et al. 1993) in which a compact unentangled polymer crumples into a series of small globules under certain topological constraints (Mirny 2011). Such polymer state is unknotted facilitating the unfolding and refolding in the cell cycle and during gene activation and repression. Furthermore, a polymer in such state tend to form spatial domains at the mega-base scale of the size observed in 3C data. In contrast, the equilibrium globule is highly knotted and do not present similar spatial domains.

However, the fractal globule model fails to explain certain experimental observations. One of them is the plateau at large genomic distances observed in FISH experiments when measuring the mean-square spatial distance between two genomic regions as a function of their genomic distance. Such plateau is produced by the organization of chromosomes into territories. Another observation that is not explained by the fractal globule model is the variability of the exponential decay of the contact probability among different regions and cell types. Moreover, one needs highly specific simulation constraints for a polymer to reach the specific conditions of the fractal globule state in which it stays briefly before converging to a different equilibrium state.

More recent polymer models are able to simulate more accurately the experimental observations. The "String and Binders Switch model" (SBS) (Barbieri et al. 2012) explains the genome folding as the effect of the binding of macromolecules (binders) on chromosomes (string). Each chromosome has different binding sites to which certain binders have specific affinities. The introduction in the simulations of the binders in certain concentrations would explain the formation of domains and other observed phenomena. Furthermore, the SBS model has been shown to be compatible with thermodynamics mechanisms of phase separation in single cells (Conte et al. 2020) in which chromatin adopt two main states: one in which is randomly folded and another where it is organized in segregated globules. The concentration and affinity of the binders switches the system from one state to the other in what is referred as phase separation. The globular state allows the establishment of stable environments where specific contacts are highly favored over stochastic encounters. The coexistence of the distinct states in the cell population give rise to many different single-molecule conformations which are compatible with the highly variable structural and temporal patterns of contacts observed within TADs.

The "loop-extrusion" model (Sanborn et al. 2015; G. Fudenberg et al. 2016; Gassler et al. 2017) is based on the binding of loop-extrusion factors which extrude chromatin and form the observed organizational domains by their continuous loading and unloading and the presence of boundary elements.

Another suggested model hypothesizes on the formation of loops in the loop-extrusion model as supercoiling processes induced by transcription (Racko et al. 2018). Supercoiled chromatin would better explain the increase of inter-contacts within TADs.

Objectives

The main objective of this thesis is to develop computational tools for the analysis of the three-dimensional structure of the genome contributing to the main two approaches used to study it: microscopy and Chromosome Conformation Capture (3C) technologies. Additionally, to combine Hi-C analysis techniques with innovative microscopy technologies to relate the genome structure of individual cells with the average population. To achieve the main objective, the following projects were conducted:

1. The provision of the indispensable computational tools to decode and analyze OligoFISSEQ images. First, the design and implementation of an automated decoding pipeline prepared to handle the high-throughput nature of the technology. Second, the analysis of the results and the interpretation of the structural information provided by OligoFISSEQ.

2. Contribute to the development of a new method for the modelling of genomic regions from sparse 3C-based information.

3. The development of probabilistic TADbit (pTADbit) that produces ensembles of three-dimensional structures of genomic regions combining Hi-C data and information from imaging experiments using Machine Learning (ML).

Impact and authorship report of the publications

This thesis dissertation is composed of three scientific publications to which David Castillo has significantly contributed. The first two manuscripts have been published in *Nature Methods* and *NAR Genomics and Bioinformatics*. The third article will be submitted to peer-review in the following months and a pre-print version is already available in *BioRxiv*.

The first and third articles in which David is co-first and first author respectively constitute the main projects of his PhD. The impact factors of the journals and the specific contributions of David to each manuscript are indicated in the following section.

Marc A. Marti-Renom

3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing

Huy Q. Nguyen[*1], Shyamtanu Chattoraj[*1], **David Castillo**[*2], Son C. Nguyen, Guy Nir, Antonios Lioutas, Elliot A. Hershberg, Nuno M. C. Martins, Paul L. Reginato, Mohammed Hannan, Brian J. Beliveau, George M. Church, Evan R. Daugharthy, Marc A. Marti-Renom & C.-ting Wu

[*] These authors contributed equally

[1] Department of Genetics, Harvard Medical School, Boston, MA, USA

[2] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

- Published in Nature Methods, July 2020
- 5-year impact factor: 34.975
- URL: https://doi.org/10.1038/s41592-020-0890-0
- Author's contribution: This is the main published article of David's thesis. He showed all his outstanding skills in designing and implementing the decoding pipeline to de-multiplex the information in the OligoFISSEQ high-throughput images. David's contribution was essential to the analysis of the raw images and the interpretation of the structural information provided by OligoFISSEQ. Huy Nguyen and Shyamtanu Chattoraj designed the OligoFISSEQ protocol and perform the described experiments.

3D reconstruction of genomic regions from sparse interaction data

Julen Mendieta-Esteban[*], Marco Di Stefano, **David Castillo**, Irene Farabella, Marc A Marti-Renom

- Published in NAR Genomics and Bioinformatics, March 2021
- 5-year impact factor: Not determined (<2 years old journal)
- URL: https://doi.org/10.1093/nargab/lqab017
- Author's contribution: The article introduces a new method to reconstruct the chromatin structural (3D) organization from sparse 3C-based datasets such as pcHi-C. David contributed to this work in the development of the computational

framework TADdyn, a tool that integrates restraint-based modelling and molecular dynamics.

Probabilistic 3D-modelling of genomes and genomic domains by integrating high-throughput imaging and Hi-C using machine learning

David Castillo*, Julen Mendieta-Esteban, Marc A Marti-Renom

- Pre-print version available in *BioRxiv*, September 2022
- URL: https://www.biorxiv.org/content/10.1101/2022.09.19.508575v1
- Author's contribution: This manuscript corresponds to the other big share of David's thesis after the publication of the first article. David was involved in all the sections of the article, from the design and implementation of the Machine Learning neural networks to the development of the software for the modelling of chromatin from 3C data.

Publications

Chapter I

3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing

There is a need for methods that can image chromosomes with genome-wide coverage, as well as greater genomic and optical resolution. We introduced OligoFISSEQ, a suite of three methods that leverage fluorescence *in situ* sequencing (FISSEQ) of barcoded Oligopaint probes to enable the rapid visualization of many targeted genomic regions. Applying OligoFISSEQ to human diploid fibroblast cells, we show how four rounds of sequencing are sufficient to produce 3D maps of 36 genomic targets across six chromosomes in hundreds to thousands of cells, implying a potential to image thousands of targets in only five to eight rounds of sequencing. We also used OligoFISSEQ to trace chromosomes at finer resolution, following the path of the X chromosome through 46 regions, with separate studies showing compatibility of OligoFISSEQ with immunocytochemistry. Finally, we combined OligoFISSEQ with OligoSTORM, laying the foundation for accelerated single-molecule super-resolution imaging of large swaths of, if not entire, human genomes.

ARTICLES
https://doi.org/10.1038/s41592-020-0890-0

nature methods

Check for updates

3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing

Huy Q. Nguyen[1,14], Shyamtanu Chattoraj[1,14], David Castillo[2,14], Son C. Nguyen[1,12], Guy Nir[1,3], Antonios Lioutas[1], Elliot A. Hershberg[4], Nuno M. C. Martins[1], Paul L. Reginato[1,3,5], Mohammed Hannan[1], Brian J. Beliveau[4,6], George M. Church[1,3], Evan R. Daugharthy[1,3,7,8,13], Marc A. Marti-Renom[2,9,10,11] ✉ and C.-ting Wu[1,3] ✉

There is a need for methods that can image chromosomes with genome-wide coverage, as well as greater genomic and optical resolution. We introduce OligoFISSEQ, a suite of three methods that leverage fluorescence in situ sequencing (FISSEQ) of barcoded Oligopaint probes to enable the rapid visualization of many targeted genomic regions. Applying OligoFISSEQ to human diploid fibroblast cells, we show how four rounds of sequencing are sufficient to produce 3D maps of 36 genomic targets across six chromosomes in hundreds to thousands of cells, implying a potential to image thousands of targets in only five to eight rounds of sequencing. We also use OligoFISSEQ to trace chromosomes at finer resolution, following the path of the X chromosome through 46 regions, with separate studies showing compatibility of OligoFISSEQ with immunocytochemistry. Finally, we combined OligoFISSEQ with OligoSTORM, laying the foundation for accelerated single-molecule super-resolution imaging of large swaths of, if not entire, human genomes.

A capacity to view genomes in situ, in their entirety and at high genomic resolution is becoming increasingly important, with one potentially enabling class of methods being fluorescence in situ hybridization (FISH)[1]. Indeed, it was FISH that enabled the pioneering work demonstrating chromosome territories in interphase cells[2,3]. Of the several methods for FISH, a number are oligomer (oligo) based[1]; one such method is Oligopaints[4] (see Supplementary Note 1 for additional examples), which appends nongenomic sequences (Mainstreet and Backstreet) to enable multiple functionalities, including amplification, indirect visualization via fluorophore-conjugated (secondary) oligonucleotides, barcode-based multiplexing and sequential and combinatorial labeling of DNA or RNA[4–21]. In the context of megabase-level coverage, some studies have used these functionalities to walk along contiguous megabases of the genome[13,14], with others labeling up to 40 regions on single chromosomes to reveal chromosomal paths[9,21], and still other studies visualizing entire, or nearly entire, genomes, one chromosome or one chromosome arm at a time[15,19]. Here we demonstrate how streets enable a new technology, OligoFISSEQ, which vastly increases the number of targets that can be visualized, putting us within reach of genome-wide imaging via the visualization of a multitude of subchromosomal regions. As OligoFISSEQ is compatible with the single-molecule localization method OligoSTORM[5,10], it also accelerates the speed with which genomic regions can be visualized at super-resolution.

OligoFISSEQ is based on FISSEQ technologies that have been honed for in situ detection of transcripts[22,23] (see Supplementary Note 2 for recent iterations and earlier studies). Here we present three strategies that direct the sequencing to barcodes embedded in Oligopaint streets, wherein one strategy uses sequencing by ligation (SBL), another uses sequencing by synthesis (SBS) and a third strategy uses sequencing by hybridization (SBH). Focusing on OligoFISSEQ with SBL, we map 66 genomic regions in human diploid PGP1 skin fibroblast cells (XY; PGP1f) using only four rounds of sequencing. We next introduce a method to improve barcode detection and, in conjunction with OligoFISSEQ, trace the human X chromosome by mapping 46 regions along its length. We demonstrate that OligoFISSEQ is compatible with immunofluorescence (IF) and then conclude by combining OligoFISSEQ with OligoSTORM to achieve a much accelerated rate at which multiple genomic regions (ranging in size from tens of kilobases to megabases) can be visualized simultaneously at super-resolution.

Results

Principle and validation of OligoFISSEQ. FISSEQ technologies[22,23] leverage next-generation sequencing methods[24,25] to provide in situ 3D spatial maps of transcripts that have been reverse transcribed and then amplified. As FISSEQ can also be used for in situ decoding of barcodes introduced during the generation of cDNA, we reasoned that it might be possible for FISSEQ to read barcoded

[1]Department of Genetics, Harvard Medical School, Boston, MA, USA. [2]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [3]Wyss Institute, Harvard Medical School, Boston, MA, USA. [4]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [5]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA. [7]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [8]ReadCoor, Cambridge, MA, USA. [9]CRG, BIST, Barcelona, Spain. [10]Pompeu Fabra University, Barcelona, Spain. [11]ICREA, Barcelona, Spain. [12]Present address: Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA. [13]Present address: ReadCoor, Cambridge, MA, USA. [14]These authors contributed equally: Huy Q. Nguyen, Shyamtanu Chattoraj, David Castillo. ✉e-mail: martirenom@cnag.crg.eu; twu@genetics.med.harvard.edu

Oligopaints. Furthermore, by targeting hundreds to thousands of identically barcoded Oligopaints to a genomic region, the combination of Oligopaints with FISSEQ, which we call OligoFISSEQ, could both obviate the need for target amplification, typically required by FISSEQ, and render the targeted chromosomal structure amenable to imaging. Finally, as FISSEQ is carried out using diffraction-limited microscopy, we anticipated a capacity of OligoFISSEQ to image the same genomic regions in hundreds to thousands of cells and thus provide the computational and statistical power necessary for addressing cell-to-cell variability.

We began by designing an Oligopaint library that targeted 18,536 oligonucleotides to a 4.8-Mb single-copy region on human chromosome 19 (Chr19-20K; Extended Data Fig. 1a) and then tested whether it could be sequenced in situ, focusing first on SBL to effect ligation-based interrogation of targets (LIT) and then on SBS to effect synthesis-based interrogation of targets (SIT), implementing hybridization-based interrogation of targets (HIT) only later (Fig. 1a–e). Importantly, as Oligopaint streets can accommodate multiple barcodes, we were able to design a single library to accommodate the sequencing chemistries of both LIT and SIT, with the primer binding site and barcode for LIT embedded on Mainstreet (5′ end of the Oligopaint oligonucleotide) and the primer binding site and barcode for SIT embedded on Backstreet (3′ end of the Oligopaint oligonucleotide; Fig. 1a). We use LIT and SIT to refer to the steps of sequencing per se, and OligoFISSEQ-LIT (O-LIT) and OligoFISSEQ-SIT (O-SIT) to refer to the use of LIT and SIT, respectively, in the context of OligoFISSEQ.

With O-LIT (Fig. 1c and Extended Data Fig. 1b), the barcode was read with SOLiD chemistry[24], wherein each barcode digit (defined as the smallest unit of a barcode; five nucleotides per digit) was read by cleavable 8-mers carrying one of four fluorophores. In brief, a sequencing primer was hybridized to the street, and a subsequent barcode readout began by binding of the first barcode digit by a labeled 8-mer, which was then ligated and imaged. The 8-mer was then cleaved between nucleotides five and six, leaving the first five nucleotides and removing the label, allowing the next digit to be read. Excluding the primer binding site, barcodes were 23 nucleotides in length and sufficient to accommodate four rounds of sequencing ((four rounds of sequencing × five nucleotides per digit) + three nucleotides uncleaved after the fourth round of sequencing); when fully utilized, four- or eight-digit barcodes have the potential to distinguish 256 (4^4) or 65,536 (4^8) targets, respectively. Using O-LIT on Chr19-20K, we recovered four-digit barcodes from 92.1% ± 5.7% of PGP1f cells ($n = 85$ cells from four replicates; Fig. 1f).

In the case of O-SIT (Fig. 1d and Extended Data Fig. 1b), barcodes were sequenced using Illumina NextSeq chemistry[24] via the extension of primers one base at a time and using only two fluorophores; one fluorophore was assigned to deoxycytidine (C), the other was assigned to deoxythymidine (T), both fluorophores were assigned simultaneously to deoxyadenosine (A), and deoxyguanosine (G) was left unlabeled (Fig. 1d,f). With each digit of the barcode being only a single nucleotide, SIT barcodes are compact, with an eight-nucleotide-long barcode theoretically able to identify 65,536 targets (4^8). Following the application of O-SIT to Chr19-20K, we recovered four-digit barcodes from 90.8% ± 5.6% of PGP1f cells ($n = 66$ cells from four replicates; Fig. 1f).

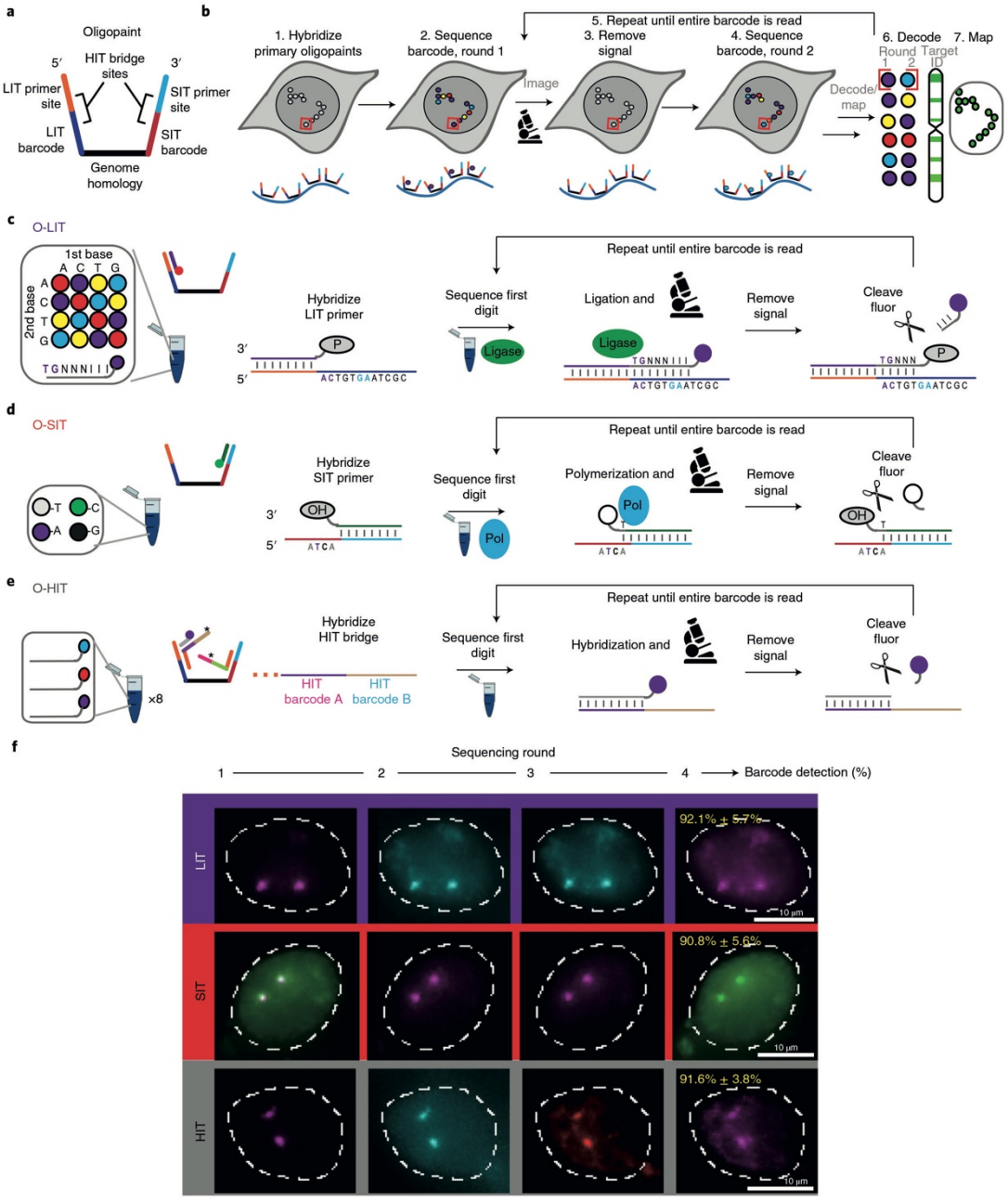
Chr19-20K can also be co-opted for HIT through SBH (Fig. 1a), reminiscent of strategies that have enabled Oligopaints to facilitate transcriptome profiling[6,8,12,18]. Here, we introduce SBH for 3D spatial mapping of chromosomal DNA. In particular, we implemented OligoFISSEQ-HIT (O-HIT) by appending SBH barcodes via two bridge oligonucleotides[14,19,20,26]—one hybridizing to the junction of the LIT barcode and its primer sequence on Mainstreet and the other hybridizing to the junction of the SIT barcode and its primer sequence on Backstreet; SBH barcodes can also be embedded directly into the streets. As each bridge carries two 20-nucleotide barcode positions, each position encoding one of six possible barcodes, the resulting 24 (4 × 6) barcodes had the potential to identify 1,296 (6^4) targets (Fig. 1e). Each barcode was identified via complementary labeled secondary oligonucleotides, and thus, using three fluorophore species, eight rounds of hybridization (8 × 3) were sufficient to identify all 24 barcodes in this iteration of O-HIT, with the option to increase target capacity through additional barcode positions, barcode sequences and/or fluorophore species. By using O-HIT on Chr19-20K, we successfully recovered four-digit barcodes from 91.6% ± 3.8% of PGP1f cells ($n = 79$ cells from four replicates; Fig. 1f and Extended Data Fig. 1b).

Mapping 66 genomic regions with O-LIT. We next assessed the potential of OligoFISSEQ to address multiple regions on multiple chromosomes. We chose to work with O-LIT because it is expected to scale without the increased costs predicted to accompany the scaling up of purely hybridization-based technologies, such as O-HIT, for which the number of species of labeled oligonucleotides, and thus their cost, would increase as the number of targets increases. In contrast, O-LIT reagents would remain the same regardless of whether they target one region or hundreds or thousands of regions. Furthermore, because the five-nucleotide O-LIT barcode digits are relatively compact, they decrease the requisite length of Oligopaint oligonucleotides, further reducing costs. In addition, because O-LIT delivers a positive signal at each round of sequencing, its barcoding is more robust, in contrast to O-SIT and O-HIT, which contain 'blank' readouts.

To assess the scalability of O-LIT, we designed an Oligopaint library (36plex-5K; Fig. 2a) targeting six regions along each of six chromosomes: chromosome 2 (Chr2; 242 Mb), Chr3 (198 Mb), Chr5 (181 Mb), Chr16 (90 Mb), Chr19 (58 Mb) and ChrX (156 Mb), with a unique barcode for each of the 36 targets. Thus, 36plex-5K targeted a total of 66 regions in PGP1f cells (six targets for each of two homologs of the five autosomes and six targets on the single X chromosome), each represented by 5,000 Oligopaint oligonucleotides and, together, encompassing 31.6 Mb, with targeted regions ranging in size between 642 kb and 1.22 Mb (876 kb average). We chose gene-poor chromosomes (5.4–6.1 genes per Mb; Chr2, Chr3,

Fig. 1 | Using OligoFISSEQ to sequence barcoded Oligopaints in situ. a, Oligopaint oligonucleotides used for OligoFISSEQ. Portions of the LIT and SIT primer sites and barcodes can function as binding sites for HIT bridges (**e**), as well as priming sites to amplify the Oligopaint library. **b**, OligoFISSEQ workflow. **c**, O-LIT workflow. After the phosphorylated LIT primer (P) is hybridized, it is ligated to an 8-mer (TGNNNIII), the first two nucleotides of which correspond to a specific fluorophore; as Oligopaint barcodes are predefined, each fluorophore corresponds to only a single barcode digit. N denotes a mixture of A, C, T or G; I denotes deoxyinosine[32], a universal base. **d**, O-SIT workflow. SIT primers contain 3′ hydroxyls (OH). A (purple) and C (green) are conjugated to distinct fluorophores and T (gray) is conjugated to two fluorophores, with G (black) remaining unlabeled. **e**, O-HIT workflow. In this iteration, two bridge oligonucleotides (asterisks) bring in four barcode positions, for each of which there are six possible barcode sequences. As each round of hybridization brings in three fluorophore-conjugated secondary oligonucleotides, each corresponding to one barcode sequence, eight rounds of hybridization (24 labeled oligonucleotides) are sufficient in this case to determine the sequence at each barcode position. **f**, Representative images after four rounds of O-LIT, O-SIT and O-HIT using Chr19-20K on PGP1f cells. Images are representative of maximum-intensity *z*-projections. The first round of SIT identified deoxyadenosine (labeled by a combination of purple and green and thus appearing white). Mean barcode detection efficiencies with s.d. values are shown from four replicates for LIT, SIT and HIT representing 85, 66 and 79 total cells, respectively.

Chr5 and ChrX) and gene-rich chromosomes (10.8 and 23 genes per Mb; Chr16 and Chr19, respectively), as well as large chromosomes (242 Mb; Chr2) and small chromosomes (58 Mb; Chr19). We positioned three targets along each chromosome arm—one target as close as possible to the telomere, one in the center of the arm and one as close as possible to the centromere, with intertarget distances ranging from 7 Mb to 74.9 Mb (average of 28.8 Mb). The number of Oligopaint oligonucleotides per target (5,000) was kept constant to assess the robustness of LIT with respect to target size and different densities of oligonucleotide binding sites (4–7.7 binding sites per



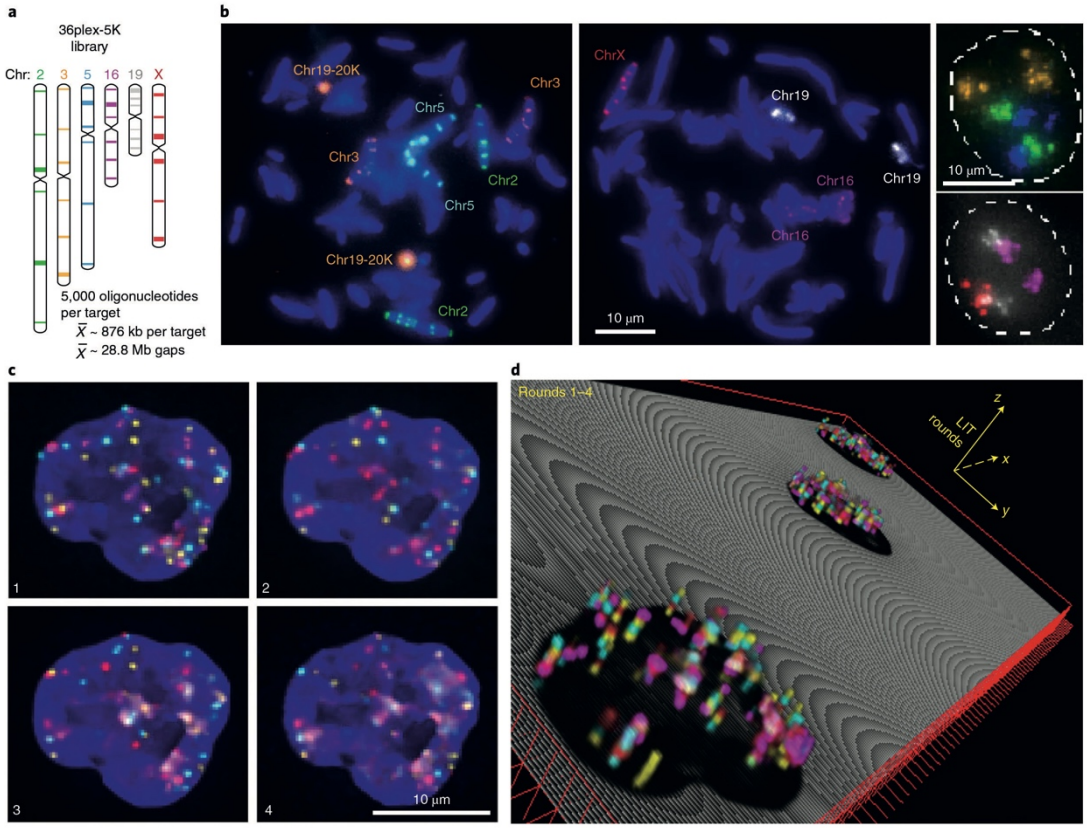


Fig. 2 | OligoFISSEQ-LIT on 36plex-5K. a, Chromosome numbers are color coded to correspond with images in **b**. Each target corresponds to a unique barcode. **b**, Metaphase chromosome spreads of male lymphoblast cells (left; cells from Applied Genetics; Methods) and interphase nuclei from PGP1f cells (right) are representative of four replicates. All six targets on any single chromosome were labeled with secondary oligonucleotides carrying the same species (color) of fluorophore. Chr19-20K was used as a positive control in metaphase chromosome spreads. Images are representative of maximum-intensity *z*-projections. **c**, Four rounds of O-LIT off both streets of 36plex-5K. Images were deconvolved and represent five-color merged maximum-intensity *z*-projections; $n=1$. **d**, 3D representation of the field of view (FOV) containing three cells sequenced with four rounds of O-LIT. Sequencing rounds are represented on the *z* axis, with the first round being closest to the DAPI-determined nuclear outline (black). The maximum-intensity *z*-projection of the sequencing signal from each round was taken, duplicated (a total of two images for better visualization) and then stacked on top of each other. The lower left cell corresponds to the cell in **c**.

kb, average of 5.8). In addition, because all 36plex-5K Oligopaint oligonucleotides targeting the same chromosome shared the same reverse primer sequence, it was possible to use indirect labeling to produce a six-banded pattern along all targeted chromosomes in metaphase and distinctly colored territories in interphase cells (Fig. 2b). This outcome confirmed the accuracy of the library.

An every-pixel automated analysis pipeline. To improve target detection, we sequenced simultaneously off Mainstreet and Backstreet (Fig. 2c,d and Extended Data Fig. 1c–f), which, in the case of 36plex-5K, carried the same barcode. Indeed, this strategy identified 100% of the 66 targeted regions in PGP1f cells via manual decoding ($n=2$ from two replicates; Extended Data Fig. 1f). However, as manual decoding does not scale well, we developed an automated pipeline to address a range of signal intensities and sizes by interrogating every pixel individually (Fig. 3a); a centroid-based pipeline did not perform as well as the every-pixel pipeline (29.93% ± 4.9% versus 62.8% ± 4.8%, $n=111$ cells from three replicates; Extended Data Fig. 1g).

The every-pixel pipeline detected 95% ± 5.15% of 36plex-5K targeted regions but with many false positives (FPs; 574.86 ± 325.38 FPs per nucleus; $n=611$ cells from 15 replicates; Extended Data Fig. 2a,b). Thus, we developed a two-tier system (Fig. 3a) in which tier 1 filtered out pixels below a minimum signal intensity and/or patch size, reducing FPs 165-fold (3.49 ± 1.36 FPs per nucleus; 5.29% ± 2.06%) while detecting 62.2% ± 6.68% of the targeted regions (~41/66) in each nucleus ($n=611$ cells from 15 replicates; Extended Data Fig. 2c,d). In tier 2, the requirements for pixel intensity and patch size were lowered, after which barcode subsampling was applied, and all newly detected signals from the same chromosome were required to be within 4.5 μm of tier 1 detected regions. This proximity-based filtering reflects the propensity of chromosomes to occupy distinct territories[2], as well as measurements of distances between consecutive tier 1 regions along a chromosome (Methods; Supplementary

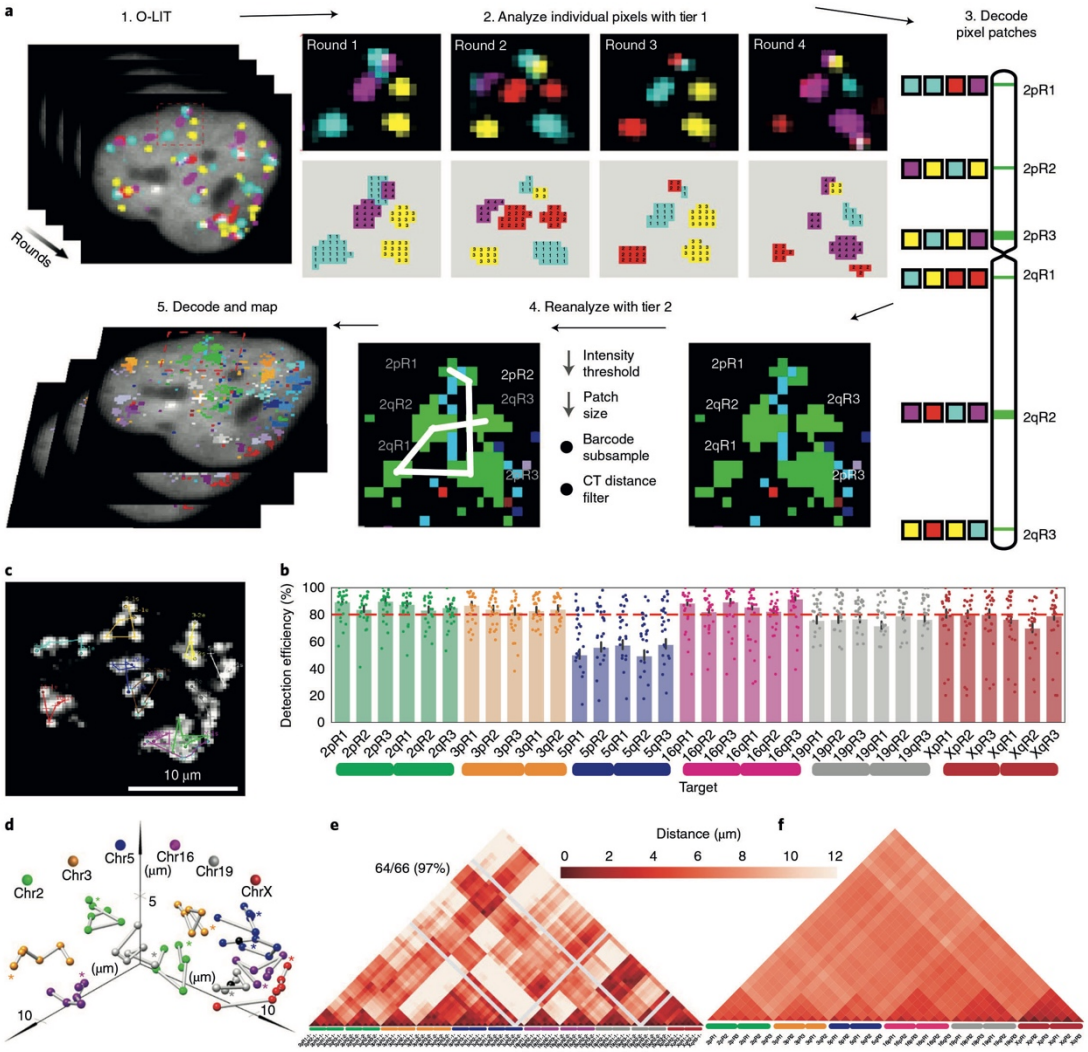


Fig. 3 | Every-pixel analysis pipeline on 36plex-5K. a, Sequencing rounds (step 1) were analyzed at the level of individual pixels using tier 1 parameters with thresholds for signal intensity and pixel patch size (step 2), and pixel patches were then decoded (step 3). Missing targets and FPs were filtered by reanalyzing images with tier 2 parameters (step 4) to produce traces (step 5). Tier 2 decreased the thresholds for signal intensity and pixel patch size, subsampled barcodes and applied filters for chromosome territories. Barcodes and color codes were designated as follows: 1, FITC; 2, Cy3; 3, TxRed; 4, Cy5. **b**, Tier 2 detection efficiency of 36plex-5K after sequencing off both streets; 80.2% ± 7.3% of targeted regions were detected in 611 cells from 15 replicates. **c**, Detection efficiencies from individual replicates are shown, with chromosomal targets on the *x* axis. The dashed red line marks the mean of all chromosomal targets. 3qR3 and 5pR3 shared a barcode and were not included. Error bars represent the 95% bootstrap confidence interval (CI) of the mean. **c**, Chromosome traces of Fig. 2c nucleus after tier 2. In total, 64 of 66 (97%) targeted regions were detected; $n = 1$. **d**, Ball-and-stick traces of the nucleus referred to in **c**. Colored spheres represent targets; black spheres represent undetected targets and were positioned by calculating the median proportionate distance between flanking detected spheres. Gray lines between signals denote extrapolations. The asterisks mark the beginning of chromosomes. **e**, Single-cell pairwise spatial distance matrix after tier 2 detection of the nucleus referred to in **b**. Homologs are displayed separately. Centroids of targets were used for this and all subsequent spatial distance matrices. Gray lines denote undetected targets. **f**, 36plex-5K population pairwise spatial distance measurements after tier 1 detection ($n = 611$ cells from 15 replicates). Homologous target measurements were combined.

Fig. 1), although in the context of chromosome rearrangements it would need to be modified. Tier 2 eliminated all FPs while detecting 80.2% ± 7.3% (~52/66) of targeted regions in each nucleus with at least 70% (~46/66) of targeted regions recovered in ~70% of cells (Fig. 3b and Extended Data Fig. 2e–g). The centroids of all detected targets were then conceptually connected to produce ball-and-stick

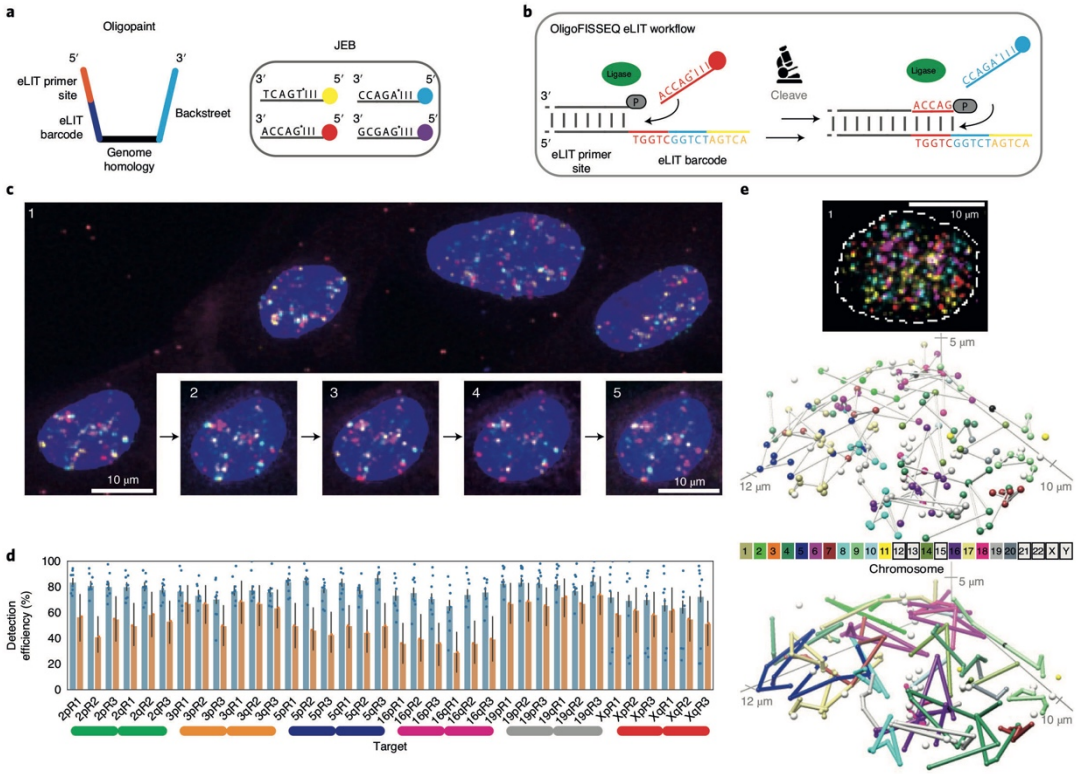


Fig. 4 | Improving O-LIT by using JEB. a, Design of Oligopaint oligonucleotides that use eLIT (left) and JEB-labeled 8-mers complementary to the five-nucleotide eLIT barcode digit (right). eLIT is compatible with a variety of barcode configurations; our strategy used barcodes consisting of five digits each, in which each digit was one of only four distinct five-nucleotide sequences. To further reduce the complexity of the pool of eight-nucleotide oligonucleotides, we also used deoxyinosine[32] in positions 6, 7 and 8. In short, JEB technology reduced the pool of labeled oligonucleotides from 1,024 to four (Extended Data Fig. 5a,b). **b**, eLIT workflow with JEB. **c**, Five rounds of sequencing with O-eLIT. PGP1f cells after the first round of sequencing (cropped field of view) and images from five rounds (1–5) of sequencing of one nucleus (inset). Extranuclear puncta are fiducial TetraSpeck beads (Thermo Fisher). Images are deconvolved maximum-intensity *z*-projections; $n=1$. **d**, Tier 2 target detection efficiency of 36plex-1K after five rounds of O-LIT with SOLiD reagents (orange; average of 54.6%; $n=41$) or O-eLIT with JEB (blue; average of 74% ± 11.2%; $n=440$ from nine replicates). Detection efficiencies from individual replicates are plotted. Error bars represent the 95% bootstrap CI of the mean. **e**, First O-eLIT round of 129-plex (top; deconvolved maximum-intensity *z* projection; $n=1$). Tier 2 tracings (middle; white spheres are tier 1 duplicated barcodes that did not move to tier 2, with untraced chromosomes boxed in color key). Sticks color-coded to facilitate visualization (bottom). Oligonucleotide target density was 5.8 to 11.9 per kb.

renditions of chromosomes, with undetected targets positioned by calculating the median distance between flanking centroids (Fig. 3c,d); ball-and-stick strategies have been used in other studies to trace chromosome paths and are useful when assessing chromosome structure and positioning[9,13,17,20,21]. Note that targets 3qR3 and 5pR3, which were designed to share barcodes, were both detected at 69% efficiency, boding well for the consistency and robustness of barcode recovery. Similarly, 15 replicates using PGP1f cells produced similar ranges of barcode recovery, with no remarkable batch effects as shown in the principal-component analysis (Extended Data Fig. 2h).

Development of eLIT to interrogate fine-scale genome organization. O-LIT mapping of 36plex-5K revealed the paths of all six chromosomes (Fig. 3c,d and Extended Data Fig. 3a,b), producing single-cell spatial genomics data (Fig. 3e,f and Extended Data Fig. 3c–e) that align with previous studies and thus argue the potential of OligoFISSEQ to be informative. First, the chromosomes fell into different territories[3], with the smaller chromosomes (Chr16 and Chr19) and larger chromosomes (Chr2, Chr3, Chr5 and ChrX) positioned toward the center and periphery of the nucleus, respectively (Extended Data Fig. 3f), in line with observations of a radial positioning of chromosomes that places smaller chromosomes more centrally[2,27]. Consistent with this, median inter-homolog distances for the smaller chromosomes were less than those for the larger chromosomes across hundreds of cells (Extended Data Fig. 4a; $P=4.3\times10^{-37}$). These robust sample sizes also enabled consideration of suggestions that diploid genomes can, under some circumstances, separate into two spatially distinct haploid sets[28–30]. Here, cluster analyses of 36plex-5K maps revealed that the five targeted PGP1f autosomes spatially separated into two haploid sets in 6.9% (18/258) of cells (Extended Data Fig. 4b–e), which, however, was statistically similar to proportions expected from randomized controls (5% and 5.4% for directed random and completely random). While definitive descriptions await the analysis of complete genomes, this observation, compounded with studies of homolog pairing and anti-pairing[31], highlights the possibility that it is in

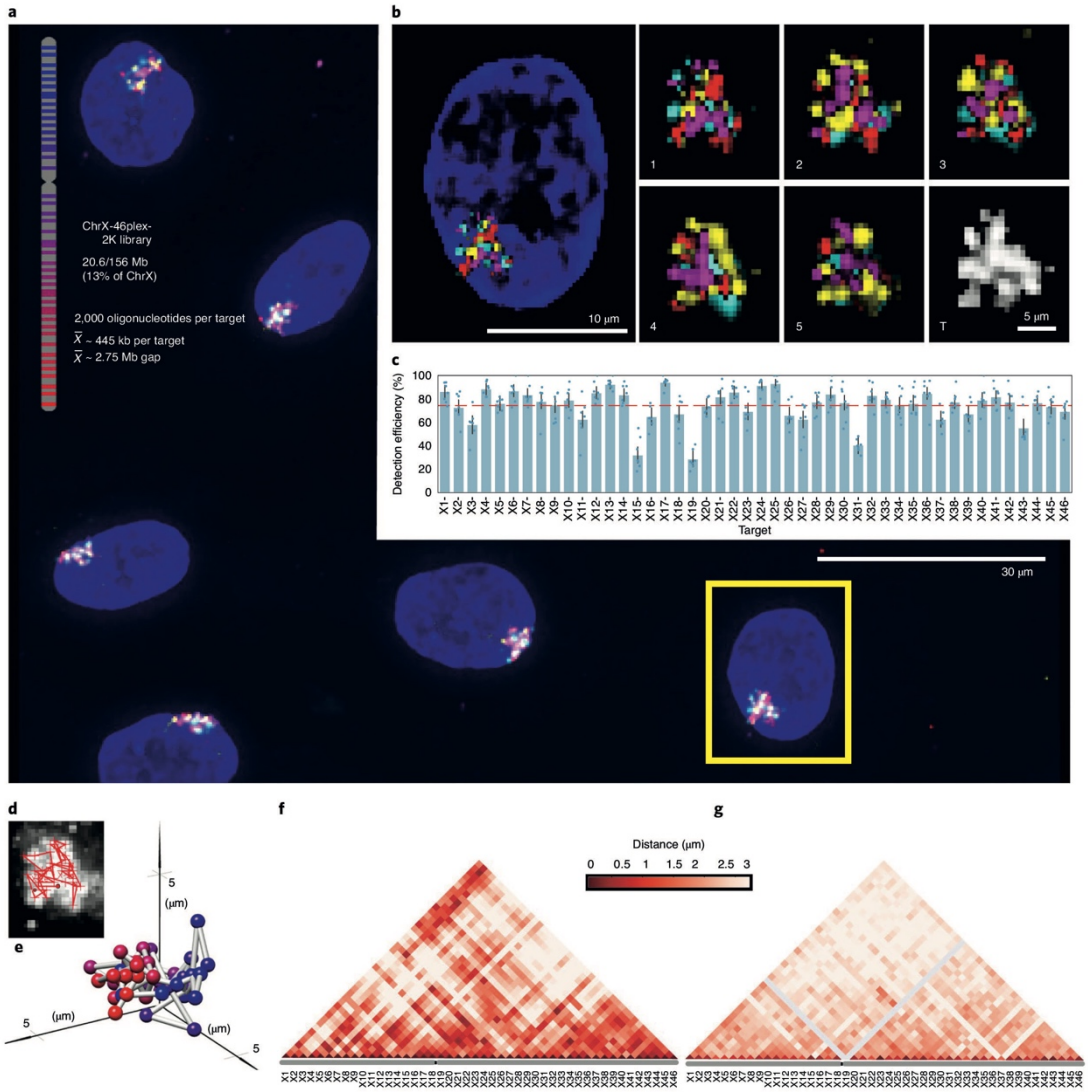


Fig. 5 | Tracing 46 regions along the X chromosome. a, Targets of ChrX-46plex-2K and nuclei after the first round of O-eLIT sequencing off both streets in PGP1f cells. Images are from deconvolved maximum-intensity *z*-projections. $n=1$. **b**, Five rounds of sequencing with O-eLIT off both streets; nucleus from **a** (yellow square). DAPI-stained nucleus after the first round of sequencing (left). T, totality of targets labeled simultaneously with a secondary oligonucleotide complementary to a barcode present on all oligonucleotides. Images are from deconvolved maximum-intensity *z*-projections; $n=1$. **c**, Tier 2 target detection efficiency after five rounds of O-eLIT off both streets in PGP1f cells. The mean detection efficiency (red dashed line) was 74.29% ± 2.5% ($n=177$ from seven replicates), and the average detection efficiency off one street was 73.7% ± 2.97% ($n=122$ from five replicates) and off both streets was 75.3% ± 1.97% ($n=55$ from two replicates). Detection efficiencies from individual replicates are plotted. Error bars represent the 95% bootstrap CI of the mean. **d**,**e**, Chromosome traces (**d**) and 3D visualization (**e**) of the nucleus from **b** after tier 2 analysis and interpolation of missing targets. Sphere color corresponds to chromosome cartoon in **a**; $n=1$. **f**, Single-cell pairwise spatial distances after interpolation of missing targets from the nucleus in **b**. **g**, Population pairwise spatial distances ($n=177$ from seven replicates) after tier 1 detection (combining reads off Mainstreet with reads off both streets).

cell types that do not segregate the genome into haploid sets that inter-homolog interactions will prevail.

We also aggregated single-cell 36plex-5K data from 611 cells to produce an average distance matrix, but this time combining data for homologous chromosomes (Fig. 3f). The comparison of this matrix to a Hi-C map of PGP1f cells[14] revealed a strong correlation ($r=0.705$, $P=1.77\times10^{-174}$; Extended Data Fig. 3e), once more indicating the robustness of O-LIT. Nevertheless, the matrices also

differed, with O-LIT producing subchromosomal stripes of greater or lesser distance, and the Hi-C matrix being more mottled. While stripes may reflect discontinuities along a chromosome, they may also suggest chromosome-specific[9,21] and interchromosome-specific signatures. For example, chromosomal regions that are overall further from other regions may be relatively more buried within a chromosome territory or nearer the nuclear membrane, while chromosomal regions that are closer to other regions may be nearer to the surface of chromosomal territories or less constrained to the nuclear membrane. As for the mottled appearance of the Hi-C matrix, it suggests that, at the scale of whole chromosomes, distances on the order of microns may not always correlate with interaction frequencies and distances amenable to Hi-C; indeed, an absence of correlation may indicate that proximity and interaction are distinct features. Thus, O-LIT matrices of distance and Hi-C matrices of interaction frequency may, together, provide layers of information that neither matrix alone can provide.

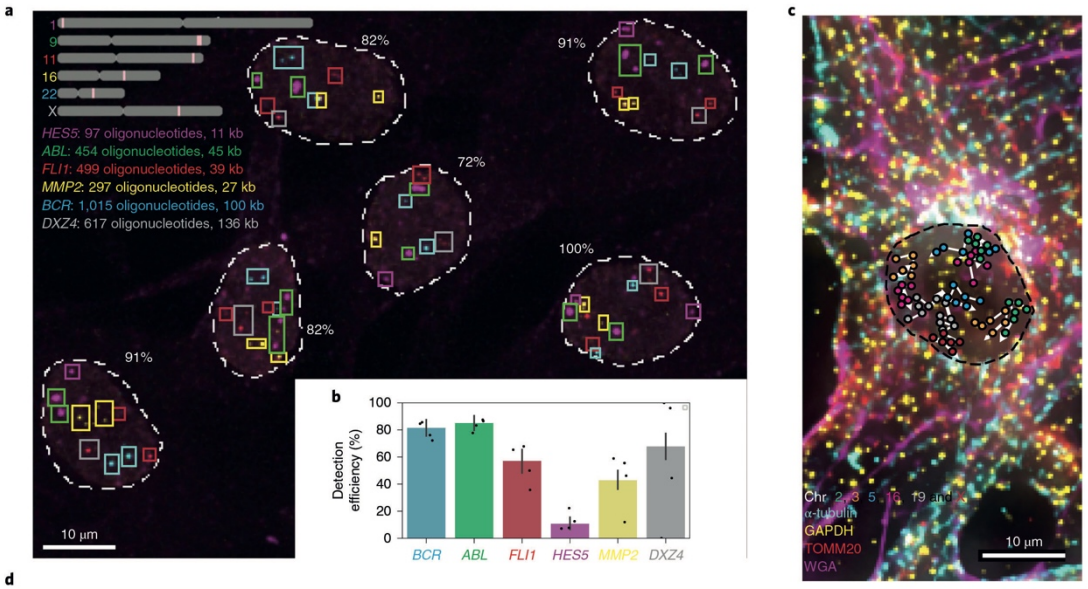
We next refined O-LIT so that it could target smaller genomic regions, as well as trace chromosomes at higher genomic resolution. However, because the commercial production of SOLiD reagents was discontinued at this juncture in our studies, we focused first on developing an alternative to the SOLiD reagents, the outcome of which was a method that ultimately improved signal detection. SOLiD chemistry reads sequences as dinucleotides using labeled eight-nucleotide oligonucleotides (TGNNNIII, where the first two positions represent all 16 dinucleotide combinations, positions 3–5 are degenerate and positions 6–8 are universal), thus entailing 1,024 (16×4^3) oligonucleotide species[24]. Because this level of complexity is excessive for O-LIT, where barcodes are defined by the user, we aimed to reduce the complexity of the oligonucleotide pool to the minimum necessary for decoding O-LIT barcodes, reasoning further that a minimally complex oligonucleotide pool might increase signal over background measurements. Thus, taking advantage of the universal base deoxyinosine[32], we reduced the complexity of the oligonucleotide pool from 1,024 to 4, referring to this strategy as 'just enough barcodes' (JEB) and the LIT chemistry using this strategy as eLIT (Fig. 4a,b). Application of OligoFISSEQ using eLIT (O-eLIT) to a library targeting 9,267 Oligopaint oligonucleotides to Chr19 (Chr19-9K) proved successful, yielding a 3.3-fold brighter signal-to-nuclear-background ratio as compared to the application of LIT to the same library using SOLiD oligonucleotides ($n=55$ cells for SOLiD and 57 cells for JEB from two replicates; Extended Data Fig. 5a,b).

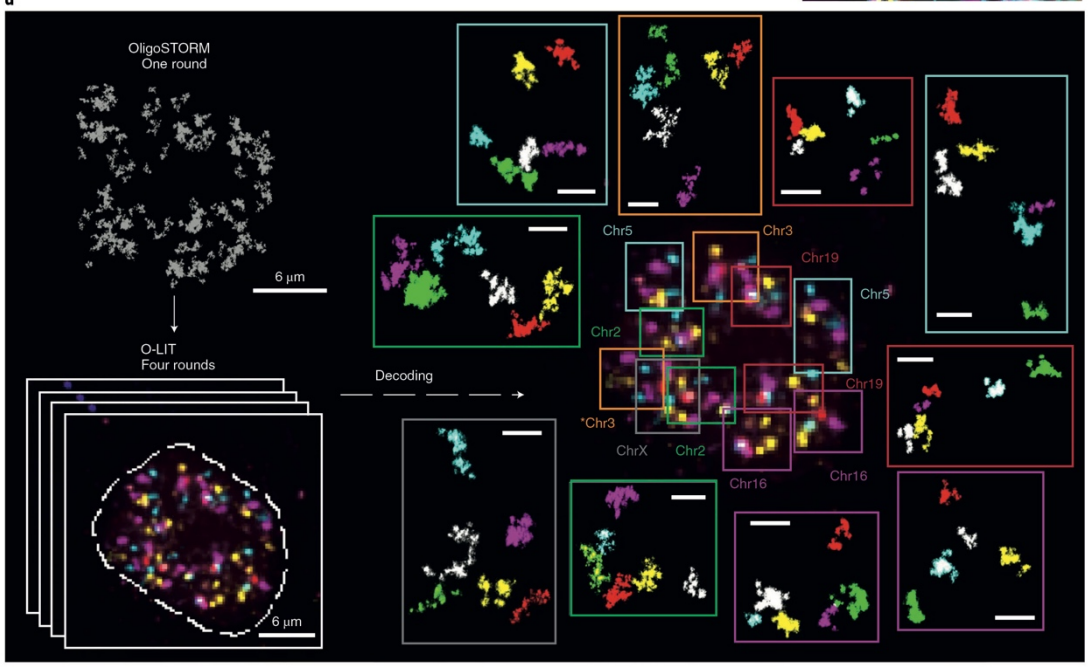
Anticipating that the improved signal-to-nuclear background ratio would improve genomic resolution, we generated a library identifying smaller genomic regions (average of 173 kb) by directing Oligopaint oligonucleotides to only the first 1,000 of the 5,000 oligonucleotide targets defined by 36plex-5K for each designated genomic region (Extended Data Fig. 5c). Then, to benchmark this library, called 36plex-1K, against 36plex-5K, we adopted the same barcodes for 35 of 36 targets, with the exception being 5pR3, which was given a new barcode; 5pR3 had previously shared a barcode with 3qR3 to enable assessment of barcode detection across different regions. Five rounds of O-eLIT using only Mainstreet of 36plex-1K yielded a tier 2 barcode recovery efficiency of 74% ± 11.2% (48 of 66; $n=440$ cells from nine replicates), which was higher than that obtained with five rounds of O-LIT (54.6%, $n=41$; Fig. 4c,d and Extended Data Fig. 5c,d). O-eLIT of 36plex-1K gave homolog-resolved data (Extended Data Fig. 5e–i). These findings argued that O-eLIT would be useful genome wide. We recently imaged 249 regions with a genome-wide library (129plex) corresponding to 129 100-kb targets spanning all the autosomes (120 targets), ChrX (6 targets) and ChrY (3 targets). Five rounds of sequencing confirmed genome-wide capacity (Methods); although inadvertent barcode duplications complicated analyses, tier 2 can nevertheless detect 95% (165 of 174) of unique barcodes, while tier 1 can detect 44% (33 of 75) of duplicated barcodes (Fig. 4e and Supplementary Table 12).

Fine ChrX tracing and suggestions of chromosome signatures. To test the potential of O-eLIT to achieve finer genomic resolution, we applied an Oligopaint library, ChrX-46plex-2K, targeting 2,000 oligonucleotides to each of the 46 regions along the human X chromosome, the number of targets aligning with a previous study that used a hybridization-based Oligopaint strategy to image 40 regions of this chromosome[9]. The targets ranged in size from 253 kb to 1.22 Mb (average of 445 kb), with an average distance between targets of 2.75 Mb and total coverage of 20.6 Mb or 13.3% of the chromosome (Fig. 5a). As such, ChrX-46plex-2K served as an informative proxy for assessing the capacity of OligoFISSEQ to accommodate all other chromosomes. Here we applied O-eLIT to both streets and achieved a tier 2 barcode recovery efficiency of 74.3% ± 2.5% in PGP1f cells (~34/46 targeted regions, $n=177$ from seven replicates; Supplementary Fig. 2, Fig. 5b,c and Extended Data Fig. 6a,b), interpolating the positions of any target that had escaped detection (Methods). Although three targets were difficult to recover (X15, X19 and X31), the quality of the data nevertheless permitted 176 traces spanning the entirety of the X chromosome, single-cell spatial distance matrices and a population-based spatial distance matrix that was strongly correlated with a corresponding Hi-C map ($r=0.641$, $P=7.074\times10^{-245}$) and inversely correlated with Hi-C interaction frequencies ($r=-0.84$, $P=5.08\times10^{-275}$), the latter producing an exponential factor of 0.18 (Fig. 5d–g and Extended Data Fig. 6c–j), similarly to that observed previously[9]. Furthermore, the chromosome traces revealed two major clusters (Extended Data Fig. 7a–c; Calinski–Harabasz index of 213.71) that differed in their radii of gyration ($t=-10.1$; $P=3.9\times10^{-19}$; Extended Data Fig. 7d), one cluster consisting of 20 chromosomes (11%) and the other comprising 156 (89%) chromosomes. While the basis for this heterogeneity will require additional study, whether it is the cell cycle, chromatin accessibility and/or overall chromosome activity, these findings emphasize the potential of O-eLIT to advance understanding of the manner in which chromosomal material can be packaged and whether that packaging correlates with function.

Fig. 6 | OligoFISSEQ extensions and applications. a, O-eLIT detection of single-gene targets after sequencing off both streets. Colored squares mark gene targets identified after five rounds of sequencing. Values reflect the percentage of targets detected out of 11 (5 autosomal genes × 2, in addition to *DXZ4* on ChrX). Images are from deconvolved maximum-intensity *z*-projections and are representative of two replicates. **b**, Tier 1 target detection efficiency from the experiment in **a** ($n=61$ cells from two replicates). Tier 2 is inapplicable due to a lack of targets from the same chromosome. Detection efficiencies from individual replicates are plotted. Error bars represent the 95% bootstrap CI of the mean. **c**, Combining O-LIT and IF. 36plex-5K was sequenced for four rounds with O-LIT off both streets, followed by IF and staining with wheat germ agglutinin (WGA). Images are from deconvolved maximum-intensity *z*-projections with chromosome traces overlaid. $n=1$. **d**, 36plex-5K was hybridized to PGP1f cells and imaged with one round of OligoSTORM (2 h) to visualize all 66 regions simultaneously, followed by four rounds of O-LIT (2–3 h per round) to decode targets. OligoSTORM image showing the entire FOV with all unidentified targets (top left). Micrograph from the first round of O-LIT; image from deconvolved maximum-intensity *z*-projection (bottom left). All six chromosomes were identified and arrayed, in super-resolution, around the central nucleus (right; central image decorated with colored squares, color coded by chromosome). All 66 regions except for one region on Chr16 were detected and identified by O-LIT, with one homolog of Chr3 (asterisk) not captured by OligoSTORM because it fell outside the FOV. All scale bars for OligoSTORM images represent 1 µm.

Oligopaint libraries 36plex-5K and 36plex-1K have also enabled analyses of chromosome folding. Combining the two datasets (for 36plex-5K, $n=611$ cells from 15 replicates; for 36plex-1K, $n=440$ cells from 9 replicates), we evaluated the angles formed by the chromosomal segments flanking the centromeres (Extended Data Fig. 8a) and observed that only a minority, if any, of the chromosomes extend their p and q arms in polar opposite directions or are folded into a hairpin; median values for the angles ranged from 74° to 94° (Extended Data Fig. 8b,c). Furthermore, assessment of the angles formed by the two contiguous chromosomal segments lying within each arm (Extended Data Fig. 8a) showed that the p and q arm angles were significantly different for Chr2, Chr3, Chr16 and Chr19 ($n=686$, 668, 586 and 760, respectively; $P=4.15\times10^{-16}$, 0.004, 1.36×10^{-14} and 3.33×10^{-11}, respectively; Extended Data Fig. 8c). As





the larger angle was associated with the p (shorter) arm of Chr2 and Chr19 and with the q (longer) arm of Chr3 and Chr16, these findings cannot be explained solely by relative arm lengths. Consistent with this, arm angle and arm length were not significantly correlated ($r=0.26$, $P=0.42$; Extended Data Fig. 8d), leaving open the possibility that arm angles reflect the impact of centromere structure on flanking genomic regions and/or interdependence of the p and q arms, the constraints of chromosomal territories or other intrinsic organizational principles, Rabl configurations resulting from the last cell division and/or the state of gene activity, such as accessibilities underlying allelic skewing. Regardless of the reasons, these observations of X-chromosome conformations (Extended Data Fig. 7a–d) and arm angles (Extended Data Fig. 8a–d) demonstrate the potential of chromosome-wide imaging to address whether there are chromosome-level structural signatures, such as may be indicative of cell type, cell state and/or cellular health or age, with evidence from a recent study of two chromosomes in *Caenorhabditis elegans* aligning with these possibilities[21]. Chromosome organization may also reflect the evolutionary history of a chromosome[33,34]. The capacity of OligoFISSEQ to generate large datasets will facilitate the study of these potential paradigms of genome organization.

Single-gene identification, IF and acceleration of super-resolution imaging. OligoFISSEQ has proven versatile, capable of imaging single regions in the size range of tens of kilobases and accommodating IF, as well as accelerating super-resolution imaging (Fig. 6a,d, Extended Data Figs. 9a,b and 10a,b and Supplementary Fig. 4a). With respect to single regions, we applied O-eLIT to six genes ranging in size from 11 kb to 136 kb (Fig. 6a,b): *HES5* (11 kb, Chr1), *MMP2* (27 kb, Chr16), *FL11* (39 kb, Chr11), *ABL* (45 kb, Chr9), *BCR* (100 kb, Chr22) and *DXZ4* (136 kb, ChrX). Detection of the larger targets hovered between 43% and 80%, reaching as high as 83.7% ± 4.38% for *ABL* ($n=61$ cells from two replicates; Fig. 6b), and although detection of the smallest target *HES5* was low (9.82% ± 3.79%), with the incorporation of amplification strategies[35,36] we expect that detection of targets as small as, or even smaller than, *HES5* should become robust. Regarding IF, we conducted four rounds of O-LIT using 36plex-5K and sequencing off both streets, followed by immunocytochemical detection of antibodies directed against α-tubulin, GAPDH and TOMM20, and we were able to trace all six chromosomes, as well as obtain strong signals for all three proteins (Fig. 6c and Extended Data Fig. 10a). We have also applied ChrX-46plex-2K to IMR-90 human fibroblast cells (XX) and then distinguished the active X (Xa) from the inactive X (Xi) chromosome through IF to macroH2A.1, which preferentially binds the latter (Extended Data Fig. 9a–l). Xi displayed a lower radius of gyration ($P=9.07\times10^{-5}$; Extended Data Fig. 9h) and megadomain structures (Extended Data Fig. 9k,l), consistent with Hi-C and FISH studies[9,37–43] and further validating the use of O-eLIT for high-resolution chromosome tracing. Taken together, these findings confirm the potential of OligoFISSEQ to enable discoveries regarding the genome-wide spatial relationship between genes and their epigenetic partners.

Lastly, we demonstrated the capacity of OligoFISSEQ to improve the speed with which genomic regions can be imaged using single-molecule localization microscopy. Here we focused on OligoSTORM[5,10], which combines Oligopaints[4] with stochastic optical reconstruction microscopy[44], to provide super-resolution images of genomic regions in a space-filling fashion and thus reveal detailed volumetric structures[5,7,13,14,16]. The throughput of OligoSTORM, however, hovers at ten to a few hundred cells per experiment, with imaging times of up to 2 h. In contrast, because OligoFISSEQ can be carried out with diffraction-limited microscopy, it has the capacity to image hundreds to thousands of cells per experiment, with relatively negligible imaging times. Thus, we explored the possibility of accelerating super-resolution genome imaging by combining O-LIT with OligoSTORM (Fig. 6d, Extended Data Fig. 10b and Supplementary Fig. 4a).

First, using 36plex-5K and bridge oligonucleotides containing binding sites for secondary oligonucleotides conjugated with a fluorophore suitable for OligoSTORM (Alexa Fluor 647), we captured all 66 targets simultaneously in a single 2-h round of OligoSTORM (Fig. 6d and Extended Data Fig. 10b; see also Chr2-6plex in Supplementary Fig. 4a). Then, with only four rounds of O-LIT, we identified all 66 regions. Thus, by combining OligoSTORM with OligoFISSEQ, we enabled a 36-fold reduction in imaging time and data storage demands (from ~2.73 TB to ~76 GB; Fig. 6d), while achieving 17 nm ± 5 nm of lateral precision and 50 nm ± 10 nm of axial precision, and 40 nm ± 5 nm of lateral resolution and 60 nm ± 5 nm of axial resolution. Extrapolating to all 46 chromosomes of a diploid human nucleus and anticipating many more than six targets per chromosome, this study demonstrates the feasibility of simultaneously 'OligoSTORMing' hundreds of regions of the genome. O-LIT should also permit OligoSTORM walking along the genome, with many walks per nucleus. Previously, we accomplished multi-walk imaging through temporal barcoding[14]. Here, multiple rounds of OligoSTORM could produce super-resolved walks in multiple regions of the genome, simultaneously, after which all regions could be identified with O-LIT. In summary, given the potential of O-LIT to identify hundreds to perhaps thousands of regions, OligoSTORM should scale similarly.

Discussion

There is a growing need for methods that will enable the imaging of entire genomes at high genomic and optical resolution while also supporting the levels of throughput and reproducibility that are becoming increasingly essential for understanding biological entities as dynamic as the genome. To this end, we have described OligoFISSEQ, a set of three methods for in situ genome mapping, demonstrating the potential of these methods to scale toward whole-genome imaging. OligoFISSEQ also has the capacity to meld with other technologies and thus extend its usefulness further. For example, when combined with homolog-specific Oligopaints (HOPs)[5], it should enable genome-wide studies in the context of parent-of-origin and, with adjustments to the barcodes, OligoFISSEQ could also enable multiplexed and/or multicolor visualization of chromosome folding in combination with other technologies, such as OligoDNA-PAINT[5], Hi-M[17] and optical reconstruction of chromatin architecture (ORCA; ref. [20]). In terms of scaling, our capacity to map 46 regions on ChrX at ~1 genomic target per 2.75 Mb predicts that OligoFISSEQ could accommodate a thousand or more targets in human nuclei, with the potential to increase that number through a reduction in target size, temporal barcoding to better resolve targets, additional rounds of sequencing and incorporation of expansion microscopy[45]; preliminary studies show that Chr19-9K can support eight rounds of O-LIT (Extended Data Fig. 10c) and that OligoFISSEQ is feasible in the context of hydrogels (Extended Data Fig. 10d,e). Scaling could also be enhanced via microfluidics, which would significantly reduce the time required for each round of sequencing by 15–20%. Indeed, with the advent of improved enzymatics, methods for amplifying signal (for example, SABER[35] and ClampFISH[36]) and superior imaging, OligoFISSEQ should become applicable to the study of smaller targets, such as enhancers and promoters. As important will be improvements in image analysis. For example, implementation of point spread function-fitting algorithms should improve spatial resolution and thus scalability[46], while a reduction in the dependence on the proximity of signals to affirm true signal would permit better detection of chromosome rearrangements, where targets that are expected to be near each other are instead widely separated. Finally, OligoFISSEQ should interface beautifully with other FISSEQ-based technologies to achieve multi-omic views of the genome, with each round of sequencing visualizing DNA, RNA[22,23] and protein[47] simultaneously.

We note that, as OligoFISSEQ has the capacity for significant genome coverage and the potential to consistently identify the same targets across thousands of cells, it is well suited for studying variability at a handful of regions as well as addressing this challenging topic at the level of the entire genome. Structural variability of specific genomic features has now been widely observed[7,9,11,13,14,16,17,20,21,48,49], and, while often thought of locally, the impact of this structural variability may reach globally[14]. Even a minor, seemingly inconsequential change in one part of the nucleus may have a profound 'butterfly effect' (ref. [50]) on the global scale, with its impact potentially contributing to and/or propagating gene regulatory states and phase separations, perhaps even constituting essential, potentially heritable signatures of the genome. Thus, although variability may appear random at the local level, a genome-wide perspective may reveal that apparent randomness actually reflects global responsiveness and an exquisitely controlled regulatory program that directs structural conformations across the entire nucleus, as much the outcome of evolution as any other honed genetic function.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0890-0.

Received: 15 December 2019; Accepted: 8 June 2020;
Published online: 27 July 2020

References

1. Hu, Q., Maurais, E. G. & Ly, P. Cellular and genomic approaches for exploring structural chromosomal rearrangements. *Chromosome Res.* **28**, 19–30 (2020).
2. Bolzer, A. et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3**, e157 (2005).
3. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* **2**, a003889 (2010).
4. Beliveau, B. J. et al. Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc. Natl Acad. Sci. USA* **109**, 21301–21306 (2012).
5. Beliveau, B. J. et al. Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nat. Commun.* **6**, 7147 (2015).
6. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
7. Boettiger, A. N. et al. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* **529**, 418–422 (2016).
8. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016).
9. Wang, S. et al. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598–602 (2016).
10. Beliveau, B. J. et al. In situ super-resolution imaging of genomic DNA with OligoSTORM and OligoDNA-PAINT. *Methods Mol. Biol.* **1663**, 231–252 (2017).
11. Cattoni, D. I. et al. Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. *Nat. Commun.* **8**, 1753 (2017).
12. Eng, C.-H. L., Shah, S., Thomassie, J. & Cai, L. Profiling the transcriptome with RNA SPOTs. *Nat. Methods* **14**, 1153–1155 (2017).
13. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, eaau1783 (2018).
14. Nir, G. et al. Walking along chromosomes with super-resolution imaging, contact maps and integrative modeling. *PLoS Genet.* **14**, e1007872 (2018).
15. Rosin, L. F., Nguyen, S. C. & Joyce, E. F. Condensin II drives large-scale folding and spatial partitioning of interphase chromosomes in *Drosophila* nuclei. *PLoS Genet.* **14**, e1007393 (2018).
16. Szabo, Q. et al. TADs are 3D structural units of higher-order chromosome organization in *Drosophila*. *Sci. Adv.* **4**, eaar8082 (2018).
17. Cardozo Gizzi, A. M. et al. Microscopy-based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms. *Mol. Cell* **74**, 212–222 (2019).
18. Eng, C.-H. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).
19. Fields, B. D., Nguyen, S. C., Nir, G. & Kennedy, S. A multiplexed DNA FISH strategy for assessing genome architecture in *Caenorhabditis elegans*. *eLife* **8**, e42823 (2019).
20. Mateo, L. J. et al. Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* **568**, 49–54 (2019).
21. Sawh, A. N. et al. Lamina-dependent stretching and unconventional chromosome compartments in early *C. elegans* embryos. *Mol. Cell* **78**, 96–111 (2020).
22. Ke, R. et al. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
23. Lee, J. H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
24. Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
25. Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).
26. Player, A. N., Shen, L.-P., Kenny, D., Antao, V. P. & Kolberg, J. A. Single-copy gene detection using branched DNA (bDNA) in situ hybridization. *J. Histochem. Cytochem.* **49**, 603–612 (2001).
27. Heride, C. et al. Distance between homologous chromosomes results from chromosome positioning constraints. *J. Cell Sci.* **123**, 4063–4075 (2010).
28. Mayer, W., Smith, A., Fundele, R. & Haaf, T. Spatial separation of parental genomes in preimplantation mouse embryos. *J. Cell Biol.* **148**, 629–634 (2000).
29. Hua, L. L. & Mikawa, T. Mitotic antipairing of homologous and sex chromosomes via spatial restriction of two haploid sets. *Proc. Natl Acad. Sci. USA* **115**, E12235–E12244 (2018).
30. Reichmann, J. et al. Dual-spindle formation in zygotes keeps parental genomes apart in early mammalian embryos. *Science* **361**, 189–193 (2018).
31. Joyce, E. F., Erceg, J. & Wu, C.-t. Pairing and anti-pairing: a balancing act in the diploid genome. *Curr. Opin. Genet. Dev.* **37**, 119–128 (2016).
32. Watkins, N. E. & SantaLucia, J. Jr. Nearest-neighbor thermodynamics of deoxyinosine pairs in DNA duplexes. *Nucleic Acids Res.* **33**, 6258–6267 (2005).
33. Yunis, J. J. & Prakash, O. The origin of man: a chromosomal pictorial legacy. *Science* **215**, 1525–1530 (1982).
34. Tjong, H. et al. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc. Natl Acad. Sci. USA* **113**, E1663–E1672 (2016).
35. Kishi, J. Y. et al. SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. *Nat. Methods* **16**, 533–544 (2019).
36. Rouhanifard, S. H. et al. ClampFISH detects individual nucleic acid molecules using click chemistry–based amplification. *Nat. Biotechnol.* **37**, 84–89 (2019).
37. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
38. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
39. Deng, X. et al. Bipartite structure of the inactive mouse X chromosome. *Genome Biol.* **16**, 152 (2015).
40. Minajigi, A. et al. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* **349**, aab2276 (2015).
41. Darrow, E. M. et al. Deletion of *DXZ4* on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl Acad. Sci. USA* **113**, E4504–E4512 (2016).
42. Giorgetti, L. et al. Structural organization of the inactive X chromosome in the mouse. *Nature* **535**, 575–579 (2016).
43. Wang, C.-Y., Jégu, T., Chu, H.-P., Oh, H. J. & Lee, J. T. SMCHD1 merges chromosome compartments and assists formation of super-structures on the inactive X. *Cell* **174**, 406–421 (2018).
44. Rust, M. J., Bates, M. & Zhuang, X. W. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3**, 793–795 (2006).
45. Chen, F., Tillberg, P. W. & Boyden, E. S. Expansion microscopy. *Science* **347**, 543–548 (2015).
46. Sage, D. et al. Super-resolution fight club: assessment of 2D and 3D single-molecule localization microscopy software. *Nat. Methods* **16**, 387–395 (2019).
47. Kohman, R. E. & Church, G. M. Fluorescent in situ sequencing of DNA barcoded antibodies. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.27.060624 (2020).
48. Finn, E. H. et al. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* **176**, 1502–1515 (2019).
49. Finn, E. H. & Misteli, T. Molecular basis and biological function of variability in spatial genome organization. *Science* **365**, eaaw9498 (2019).
50. Lorenz, E. N. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141 (1963).

Methods

Materials. Lists of reagents and catalog numbers (Supplementary Table 1), oligonucleotide sequences (Supplementary Tables 2–8 and 16) and library information such as coordinates, barcodes and density (Supplementary Table 12) are presented as supplementary information.

Oligopaint library design. All Oligopaint oligonucleotide sequences and coordinates for libraries used in this study can be found in Supplementary Tables 2–6. Oligopaints[4] leverages the ability to computationally design and synthesize sequence-specific oligonucleotide probes for FISH[4] (see Supplementary Note 1 for additional examples). Oligopaint FISH probes were computationally designed for optimal hybridization and high specificity. Oligopaint genome-binding sequences were obtained from the Oligopaints website (https://oligopaints.hms.harvard.edu/; ref. [51]), using the hg19 genome with 'Balanced' settings. 129plex sequences were obtained using OligoMiner on soft-masked hg38 sequence using a Tm window of 42–47 °C and a length range of 30–37 nucleotides[51]. Genome homology sequences of other libraries ranged from 35–41 nucleotides. Universal forward- and reverse-priming sequences were appended to each Oligopaint oligonucleotide using OligoLEGO (https://github.com/gnir/OligoLego/), allowing the libraries to be PCR amplified and renewable. The universal priming sequences also served as various OligoFISSEQ primer and bridge sites. Each library used in this study was designed with specific features and is described in detail in the supplementary file specific for each set.

LIT. For the Chr19-20K library, a portion of the universal forward-priming sequence was used as the LIT primer binding site, followed by the LIT barcode. Barcode and color-code designation was as follows: 4, Cy5/Alexa Fluor 647; 3, TxRd; 2, Cy3; 1, FITC/Alexa Fluor 488.

The 36plex-5K library shared the same universal forward-priming sequence among all oligonucleotides and contained chromosome-specific universal reverse-priming sequences. Individual chromosome targets could be amplified, hybridized and detected by using the universal reverse-priming sequence. Universal forward-priming sequences were used as LIT primer binding sites for 18-nucleotide primers. In cases where O-LIT was performed off both Mainstreet and Backstreet, a LIT primer binding site was hybridized to the Backstreet. Barcodes were specified using sequences from OligoLego (https://github.com/gnir/OligoLego/). Candidate barcode sequences were decoded to reveal color codes using a MATLAB script (https://www.mathworks.com/). To maintain color-code diversity between neighboring targets, barcodes were manually assigned to targets (for example, barcodes were specified so that neighboring targets would have different colors in the first round). Each LIT barcode digit required a five-nucleotide sequence, while the last barcode digit required eight nucleotides to allow adequate space for 8-mer binding. Thus, a four-digit barcode required 23 nucleotides in total. For 36plex-5K, the targets 3qR3 and 5pR2 contained the same barcode sequences to assess barcode recovery from separate genomic targets.

JEB/O-eLIT barcodes. The 36plex-1K library selected a subregion of 36plex-5K targets, with 1,000 Oligopaint oligonucleotides per target instead of 5,000 oligonucleotides. Additionally, 36plex-1K targets contained JEB-compatible barcode digits. The 36plex-1K targets contained the same barcode digit color coding as for 36plex-5K, with the exception of 5pR3. 36plex-1K could only be sequenced using Mainstreet and not both streets.

The ChrX-46plex library was designed to span the entire human X chromosome with 2,000 Oligopaint oligonucleotides per target. The library was divided into two sublibraries, ChrX-23plex-odd and ChrX-23plex-even, with each sublibrary targeting either odd (X1, X3, X5, ...) or even (X2, X4, X6, ...) targets. Each sublibrary contained the same universal forward-priming sequences and different universal reverse-priming sequences. ChrX-46plex barcodes contained JEB digits and were also manually assigned to maintain color-code diversity between neighboring targets. ChrX-46plex is compatible with sequencing off both streets.

The six-gene library shared the same universal forward-priming sequence and different universal reverse-priming sequences. Barcodes were manually specified using JEB digits. The six-gene library is compatible with sequencing off both streets.

The 129plex genome-wide library aims at imaging each chromosome arm of the human genome using OligoFISSEQ. We selected the regions based on the density of Oligopaint oligonucleotides that could be targeted (average, 8.6 oligonucleotide targets per kb) and position on the chromosome arm. First, using a custom-curated R script, we used a sliding window of 100 kb along all chromosomes to calculate oligonucleotide target densities. Then, wherever possible, we selected three regions for each chromosome arm: one near the telomere, another near the centromere, and a third more centrally located, selecting regions where the density of oligonucleotide targets would be above 6 per kb. For some chromosome arms, we selected fewer than three regions owing to the constraints of oligonucleotide target density. Each region corresponded to a 5-digit barcode. The 129plex was sequenced off both streets. Due to 21 inadvertently duplicated barcodes, 42 of the targets could not be assigned (Supplementary Table 12).

SIT barcode. For the Chr19-20K library, the universal reverse-priming sequence was used as the SIT primer binding site, followed by the SIT barcode sequence. Barcode and color-code designation was as follows: 4, Cy5; 3, Cy5 + Cy3; 2, Cy3; 1, blank. For 36plex-1K, the universal reverse-priming sequence was used as the SIT primer binding site, followed by SIT barcodes. Target color coding was designed to be the same as for 36plex-5K but with SIT reagents.

HIT barcode. For the Chr19-20K library, bridging oligonucleotides (HIT bridges) were designed to hybridize to Mainstreet and Backstreet. HIT bridges contained binding sites for HIT readout oligonucleotides. HIT readout oligonucleotide sequences were derived from OligoLego. Barcode and color-code designation was as follows: 0, blank; 1, Alexa Fluor 647/Cy5; 2, Cy3B/Cy3; 3, FAM/Alexa Fluor 488.

For the 36plex-5K library, HIT bridges were designed to hybridize to street-specific sequences for each target. This was done by designing bridges flanking universal priming sites (forward and reverse), as well as the 5′ or 3′ ends of LIT barcodes, due to similar LIT barcodes being present on both streets. HIT bridges contained binding sites for HIT readout oligonucleotides derived from OligoLego.

Oligopaint probe synthesis. Oligopaint oligonucleotides were purchased as single-stranded oligonucleotide pools from CustomArray (http://www.customarrayinc.com/oligos_main.htm/) or Twist Bioscience (https://www.twistbioscience.com/) in 12,000 and 92,000 chip formats. Oligonucleotide pools were amplified as previously described[10,14] with minor modifications (a step-by-step protocol can be found in Supplementary Protocol 1). Briefly, PCR conditions for each library and sublibrary were optimized using real-time PCR to obtain optimal template concentration, primer concentration and annealing temperature. Next, libraries were linearly amplified with low-cycle PCR using Kapa Taq reagents. dsDNA PCR products were purified using Zymo columns and eluted with ultra-pure water (UPW). T7 RNA promoter sequence was then appended to Oligopaints using REV primers containing the T7RNAP on the 5′ end. Note that some users may opt to add the T7RNAP straight from the raw library. dsDNA PCR products were purified using Zymo columns and eluted with UPW. PCR products were then in vitro transcribed using HiScribe (NEB, E2040S) overnight at 37 °C to make RNA.

RNA products were reverse transcribed with Thermo Maxima H Minus Reverse Transcriptase (Thermo Fisher, EP0753) to make cDNA. RNA was then digested to leave single-standed DNA. This product was purified using Zymo columns. Final single-standed DNA Oligopaint oligonucleotides were resuspended at 100 µM in UPW and stored at −20 °C until use. Linear PCR, touched-up PCR and single-standed DNA Oligopaint oligonucleotides were quality checked by running on 2% agarose DNA gels to confirm single bands were migrating at the expected sizes during synthesis.

Other oligonucleotides. Sequences for all other oligonucleotides can be found in Supplementary Tables 7 and 8. Primers, secondary fluorophore-labeled oligonucleotides, LIT sequencing primers, SIT sequencing primers, JEB oligonucleotides and molecular inversion probes were purchased from IDT (https://www.idtdna.com/). HIT secondary oligonucleotides were purchased from Bio-Synthesis (https://www.biosyn.com/). Alexa Fluor 405 activator fluorophore was purchased from Thermo Fisher (https://www.thermofisher.com/).

Cell culture. Our study used two human cell lines: PGP1f and IMR-90. PGP1f cells are primary human fibroblasts taken from the PGP1 male donor from the personal Genome Project (Coriell, GM23248; ref. [52]). They were previously found to be of normal karyotype[14,53]. PGP1f cells were cultured in DMEM (Gibco) supplemented with 10% FBS (Thermo Fisher; A3160401), 1× penicillin–streptomycin (Thermo Fisher, 15140122) and 1× nonessential amino acids (Thermo Fisher, 11140050). PGP1f cells were cultured for no more than five passages before thawing new cultures. IMR-90 cells were cultured in DMEM supplemented with 10% FBS and 1× penicillin–streptomycin. Cells were cultured at 37 °C in a 5% CO_2 incubator.

Sample preparation for OligoFISSEQ. Ibidi Sticky Slide VI (https://ibidi.com/, 80608) was used for all experiments except for metaphase spreads (Fig. 2b) and hydrogels (Extended Data Fig. 10d,e). Ibidi slides were assembled and allowed to cure overnight at 37 °C before use. Each well required 100–200 µl of reagent, and we generally designated one hole as the inlet and the other hole as the outlet. PGP1f cells from ~70% confluent 10-cm dishes were detached from the dishes using 1 ml of trypsin (Thermo Fisher, 25-200-056), neutralized with 2–3 ml of fresh medium. Next, 100 µl of cells in suspension was added to each Ibidi well and allowed to adhere and recover overnight at 37 °C in an incubator. The following day, the medium was aspirated and cells were washed with 1× PBS and fixed for 10 min with 4% formaldehyde (Electron Microscopy Sciences, 15710) in a final concentration of 1× PBS (Thermo Fisher, 10010-023). Fixative was removed and cells were rinsed with 1× PBS. Cells were then permeabilized with 0.5% Triton X-100 (Sigma-Aldrich, T8787-250ML) in a final concentration of 1× PBS for 15 min on a rotator. Permeabilization reagent was aspirated and cells were rinsed in 0.1% Triton/1× PBS and stored in either this or PBS at 4 °C until use. Samples were used within 2–3 weeks of fixation.

Cell samples for the molecular inversion probe and hydrogel experiments were grown on rectangular glass microscope slides. Cells were plated similarly to the Ibidi slides, except 150 µl of cells in suspension was plated onto discrete areas on rectangular slides (previously etched with a glass etching pen to note the region) and incubated overnight at 37 °C in a 10-cm petri dish. The following day, the same steps were performed as with Ibidi slides but in 50-ml Coplin jars. Cells were stored in 1× PBT in Coplin jars until use. Metaphase spreads were purchased from Applied Genetics (product: HMM).

DNA FISH. Step-by-step protocols can be found in Supplementary Protocols 2 and 3, which were adapted from Beliveau et al. [4] and based on previous studies[54,55]. All OligoFISSEQ methods begin with hybridization of primary Oligopaint libraries overnight and then deviate. The following steps are common to LIT, SIT and HIT with Ibidi slides (all steps were completed on a rotator unless specified otherwise). Ibidi wells were washed with 0.1% PBT at room temperature for 5 min and incubated with 0.1 N HCl for 8 min. Two SSCT washes were performed. Cellular RNA was digested with 50 µl of 2 µg ml^{-1} RNase A (Thermo Fisher, EN0531) in 2× SSCT for each well. Slides were incubated in 37 °C in a humidified chamber for 1 h. RNase A was washed out by adding 2× SSCT. Prehybridization began by adding 50% formamide/2× SSCT for 10 min at room temperature. Prehybridization continued with prewarmed (60 °C) 50% formamide/2× SSCT added, and the slide was placed on top of the heat block set in a 60 °C water bath for 20 min. Next, the primary Oligopaint library was added, the samples were aspirated and 50 µl total of primary Oligopaint oligonucleotide library (2 µM final concentration) was added in hybridization mix (50% formamide, 2× SSCT and 10% dextran sulfate). Samples with primary Oligopaint oligonucleotide libraries were then denatured, wells were sealed with parafilm to prevent evaporation and the slide was placed on a preheated hot block in an 80 °C water bath for 3 min under the weight of a rubber plug. Oligopaint oligonucleotide library hybridization to samples was performed by placing samples in a humidified chamber at 42 °C to incubate for >16 h. The next day, probes that did not hybridize were washed out by adding prewarmed (60 °C) 2× SSCT directly to each well containing primary hybe mix and were then aspirated. New prewarmed 2× SSCT was added and samples were incubated on a hot block for 15 min. This was repeated once and then again at room temperature. After this wash, the protocol deviates for the techniques (see below). Note that cellular DNA was stained after every two rounds of sequencing to maintain adequate DAPI signal.

For detection of Oligopaints via secondary hybridization, samples were then prepared for secondary oligonucleotide hybridization to primary oligonucleotide streets for detection. Samples were washed with 30% formamide/2× SSCT for 8 min and 50 µl in total of secondary oligonucleotides and/or bridge oligonucleotides was added at 1.2 µM in 30% formamide/2× SSCT to each well. Samples were incubated in a humidified chamber for 45 min at room temperature in darkness. Nonhybridized secondary oligonucleotides were washed out with 30% formamide/2× SSCT added directly to the samples, which were then aspirated and incubated twice for 15 min on a rotator. Samples were washed twice with 2× SSCT for 5 min. In some experiments, DNA was counterstained with DAPI (Thermo Fisher, D1306) in PBS for 10 min. Samples were then washed with 1× PBS twice for 5 min and imaged in 1× PBS or imaging buffer containing PBS, PCD, PCA and Trolox (Supplementary Protocol 3).

For cells on rectangular slides, the same overall protocol as above was performed but in Coplin jars, and wash volumes were scaled accordingly (25-µl volumes for primary and secondary hybridizations). The protocol was modified as follows: RNase was added directly to cells on a rectangular slide, which was covered with a 22 × 22 mm^2 coverslip. Post-RNase washes were performed by transferring the slide and coverslip to a Coplin jar and 'sliding' the coverslip off. The same approach was used for secondary hybridization. Primary Oligopaint hybridization was performed by adding primary Oligopaint mix directly to cells on a rectangular slide, covering with a 22 × 22 mm^2 coverslip and sealing the edges with rubber cement (Elmer's). Rubber cement was allowed to dry for 3 min and the sample was denatured on a heat block, similar to the process for Ibidi slides.

LIT. LIT is built upon Oligopaint[4], SBL[56] and FISSEQ technologies[22,23,57] (see Supplementary Note 2 and Supplementary Protocol 3 for recent iterations and the step-by-step protocol). After hybridization of the primary Oligopaint library, samples for O-LIT required treatment with phosphatase to deplete endogenous phosphates that could prime ligation, contributing to background and poor signal. The samples were washed with 50 µl of 1× NEB CutSmart buffer for 8 min. Next, 50 µl of shrimp alkaline phosphatase (rSAP; NEB, M0371L; 7.5 µl rSAP in 1× CutSmart) was added to each well followed by incubation at 37 °C with humidity for 1 h. To inactivate phosphatase, the sample was then transferred to a preheated heat block in a 65 °C water bath for 5 min and washed twice with preheated (65 °C) 2× SSCT on the heat block for 5 min each. The slides were washed for 5 min in 2× SSCT at room temperature. Samples were then prepared for LIT primer binding by washing with 30% formamide/2× SSCT for 8 min, and 50 µl of LIT sequencing primer was added at 1.2 µM in 30% formamide/2× SSCT to each well. Samples were incubated in humidified chambers for 45 min. Nonhybridized LIT primers were washed out with 30% formamide/2× SSCT being washed directly in, aspirated and incubated twice for 15 min on a rotator. Samples were washed with 2× SSCT twice for 5 min. Next, samples were prepared for the first round of LIT by adding 100 µl of 1× Quick Ligation buffer (NEB, B6058S) for 8 min and aspirated. LIT reaction mix (see Supplementary Protocol 3 for the recipe) was prepared on ice. Before adding ligases, vigorous vortexing was performed on the LIT reaction mix. After vortexing, ligases were added and mixed thoroughly by pipetting. O-eLIT reagent was performed similarly but, instead of SOLiD purple reagent mix, 40 pmol of each JEB oligonucleotide was added to each sample and UPW was adjusted accordingly. Next, 100 µl of this mix was added to each well and samples were incubated in a humidified chamber at 25 °C for 55 min. LIT reaction mix was then aspirated and samples were rinsed with 1 M guanadine hydrochloride (GHCL; Sigma-Aldrich, G3273) and washed twice for 15 min on a rotator at room temperature. Samples were washed in 1× PBS for 5 min. Cellular background fluorescence was reduced by treating the samples with 100 µl True Black (Biotum, 23007) in 70% ethanol for 2 min. Three 1× PBS quick rinses and a 10-min wash were performed. Samples were then imaged in 1× PBS or imaging buffer (see Supplementary Protocol 3 for recipe). Before proceeding to the next LIT round, nonligated phosphates were treated with phosphatase (Quick CIP; NEB, M0508L) for 30 min at 37 °C. Quick CIP was then washed out with three GHCL washes for 5 min. The previous LIT round was cleaved to release the fluorophore and regenerate the 5′ phosphate by rinsing and incubation for 15 min at room temperature on a rotator with cleave 1, followed by the same for cleave 2. Samples were then rinsed three times with GHCL and washed twice for 5 min. The next round of LIT could proceed with the pre-ligation step. After the last barcode digit was read, the fluorophore was cleaved and all targets were detected by hybridizing specific bridges and fluorophores as described above.

SIT. SIT is based on Oligopaint[4] and SBS[58] technologies using the Illumina NextSeq 500/550 TG Kit (Illumina, TG-160-2002). After hybridization of the primary Oligopaint library, samples were prepared for SIT primer binding by washing with 30% formamide/2× SSCT for 8 min, and 50 µl of LIT sequencing primer was added at 1.2 µM in 30% formamide/2× SSCT to each well. Samples were incubated in humidified chambers for 45 min. Nonhybridized SIT primers were washed out with 30% formamide/2× SSCT, which was added directly to the samples, aspirated and incubated twice for 15 min on a rotator. Samples were washed with 2× SSCT twice for 5 min. The first round of SIT proceeded by rinsing with 100 µl of prewarmed (60 °C) NextSeq polymerase solution (from reservoir 31) and then incubation on a 60 °C heat block in a water bath for 5 min. The samples were aspirated and washed with 2× SSCT three times for 10 min. The samples were washed in 1× PBS and then imaged in 1× PBS or imaging buffer. Before proceeding onto the next SIT round, samples were treated with NextSeq cleave solution (from reservoir 29) with a rinse and then incubated for 5 min on a 60 °C heat block in a water bath. Samples were then washed three times for 10 min in 2× SSCT. The next round of SIT could then proceed. For all target identification, SIT primers containing Alexa Fluor 488 were used, or secondary oligonucleotides with bridges were added.

HIT. HIT is based on Oligopaint[4] and SBH technologies[6,12,59]. After hybridization of the primary Oligopaint library, samples were prepared for HIT bridge oligonucleotide hybridization to primary oligonucleotide streets for detection. HIT bridges for 36plex-5K were designed to span the universal priming region and part of either the Mainstreet barcode or Backstreet barcode. Samples were washed with 30% formamide/2× SSCT for 8 min, and 50 µl of bridge oligonucleotides was added at 1.2 µM in 30% formamide/2× SSCT to each well. Samples were incubated in humidified chambers for 45 min at room temperature in darkness. Nonhybridized bridge oligonucleotides were washed out with 30% formamide/2× SSCT, which was added directly to the samples, aspirated and incubated twice for 15 min on a rotator. The first round of HIT commenced with the addition of 50 µl to each well with HIT secondary oligonucleotides specific to each round added at 1.2 µM in 30% formamide/2× SSCT for 45 min at room temperature in a dark humidified chamber. Nonhybridized HIT secondary oligonucleotides were washed out with 30% formamide/2× SSCT, which was added directly to the samples, aspirated and incubated twice for 15 min on a rotator. Samples were washed with 2× SSCT twice for 5 min and then with 1× PBS for 5 min. Samples were imaged in 1× PBS or imaging buffer. Before proceeding to the next round, the secondary oligonucleotide fluorophores from the previous HIT round were cleaved via rinsing and incubation for 15 min with 1 mM TCEP (Sigma-Aldrich, 646547-10×1ml). Samples were rinsed three times with PBS and the next HIT round commenced.

Immunofluorescence. To visualize proteins, samples were subjected to IF. After OligoFISSEQ, Oligopaint oligonucleotides were removed by washing with 80% formamide/2× SSCT twice for 7 min. Next, samples were washed with 2× SSCT for 3 min, rinsed with 1× PBS and fixed in 4% formaldehyde/PBS for 10 min. After PBS rinses and permeabilization in 0.5% Triton/PBS for 10 min, samples were blocked in 3% BSA/PBT for 1 h. Primary antibodies diluted in 1% BSA/PBT were then added to each well, and wells were sealed with parafilm and incubated overnight at 4 °C for >12 h. The next day, primary antibodies were removed and three PBT washes were performed. Secondary antibodies (Supplementary Table 1) diluted in 1% BSA/PBT were then added at a 1:500 dilution for each, for 1 h at room temperature on a shaker. WGA (Thermo Fisher, W11261; 1:20) was also added during the second incubation step. Three PBT washes for 5 min each were

performed, and samples were restained with DAPI (1:1,000) for 10 min and imaged in imaging buffer.

Hydrogel. Hydrogel embedding was based on work by Moffitt et al.[60] (see Supplementary Protocol 4 for the step-by-step protocol). Cells for hydrogel embedding were grown on rectangular glass slides. FISH was performed on these slides as described in 'DNA FISH'. After primary Oligopaint library hybridization, samples were washed at 60 °C in 2× SSCT for 20 min, then for 10 min at room temperature and then with 1× PBS for 5 min. In preparation for hydrogel embedding, slides were air-dried for 5 min and the area around cells was wiped dry with a Kimwipe. Hydrogel reagents were combined in Eppendorf tubes on ice and degassed on ice in a vacuum chamber (Thermo Fisher, 08-642-7) during incubations. Cells were then washed for 10 min at 4 °C with hydrogel mix without APS and TEMED. Hydrogel mix was then removed from samples, and ~20 µl of hydrogel solution (recipe in Supplementary Protocol 4) was spotted onto parafilm on a gelation chamber slide (rectangular slide wrapped in parafilm, using two 22 × 22 mm² coverslips as spacers on each end of the slide), then the slide sample was flipped onto hydrogel solution/gelation chamber, being careful to spread the hydrogel solution without forming bubbles. The sample was then incubated at 37 °C for 1 h in a vacuum chamber. After incubation, the gelation chamber was carefully removed. The edges of the hydrogel disc were trimmed, and a diamond etching pen was used to break the rectangular slide, preserving the gel/glass slide portion. The gel/glass slide portion was then transferred to a 35-mm petri dish and digested in 2 ml of digestion buffer (recipe in Supplementary Protocol 4; ref. [60]) overnight at 37 °C. After overnight digestion, the cell/hydrogel dissociates from the glass slide, so extra care was taken to avoid hydrogel damage. The digestion buffer and glass slide were removed, and the hydrogel was washed in 2× SSCT three times for 20 min each. The hydrogel was divided into smaller pieces for downstream applications. To note orientation, hydrogel pieces were cut into distinct shapes, to facilitate imaging and alignment downstream. After cutting, the hydrogel sample was transferred to 1.5-ml Eppendorf tubes for easier handling.

Metaphase FISH. Unless otherwise stated, all steps were performed using Coplin jars. Treatment commenced by adding 25 µl RNase A to the slides, sandwiching under a 22 × 22 mm² coverslip and then incubating in a humidified chamber. Primary Oligopaint hybridization was performed in the same way.

Diffraction-limited microscopy. OligoFISSEQ and diffraction-limited microscopy were carried out using a widefield epifluorescence setup. A Nikon Eclipse Ti body was equipped with a 60× 1.4-NA Plan Apo lambda objective (Nikon MRD01605), Andor iXon Ultra EMCCD camera (DU-897U: 512 × 512 pixel FOV, 16-µm pixel size), X-Cite 120 LED Boost light source, motorized stage and off-the-shelf filter sets from Chroma (~488 nm 49308 C191880; ~532 nm 49309 C191881; ~594 nm 49310 C191882; and ~647 nm 49009 C177216). Images were obtained with ND4 and ND8 filters in place. Microscope operation was handled by Nikon NIS Elements software. In general, *z*-stacks were obtained with 0.3-µm slices with an exposure time of 200–300 ms and LED intensity of 20–60%, depending on the library being imaged. *zxy* stage position was maintained within .nd2 metadata and was essential for returning to the same FOV. Orientation of the sample into the stage and sample holder was carefully maintained so as to enable returning to the same FOV. This was important as the sample was removed after imaging and between sequencing rounds.

OligoSTORM imaging. To combine OligoFISSEQ with OligoSTORM, we first performed one round of OligoSTORM imaging on all targets (Chr2-6plex or 36plex-5K) inside PGP1f male fibroblast cells by hybridizing Alexa Fluor 647-labeled secondary oligonucleotides that bind to the bridges (present in the Backstreet of individual Oligopaint oligonucleotides, with each chromosome containing a specific barcode), which contain a binding site for secondary oligonucleotides. OligoSTORM samples were imaged on a Vutara 352 biplane system with a ×60 1.3-NA silicone objective (UPLSAPO60XS2, Olympus). For single-molecule blinking, we used a switching buffer containing 2-mercaptoethanol and GLOXY[14]. The excitation laser power was set at 60% on the software (6.3 kW cm^{-2} at the objective) for the 640-nm laser and 0.5% on the software (0.08 kW cm^{-2} at the objective) for the photoactivation laser of 405 nm. We used 30–40 *z*-slices of 0.1-µm thickness for each *z*-slice. Approximately 10–12 photoswitching cycles of 250 frames per cycle were used for each *z*-slice.

The OligoSTORM images were analyzed using Vutara SRX software[14]. The DBSCAN clustering algorithm was used to identify the clusters from the raw image. Fifty particles within a 0.1-µm distance were used for clustering. The mean axial precision was 50 ± 10 nm in *z*, and the mean radial precision was 17 ± 5 nm in *xy*. The resolution of the super-resolved structures was calculated by Fourier ring correlation analysis (a built-up feature in SRX software). Resolution was 40 ± 5 nm in *xy* and 60 ± 5 nm in *z*.

Data visualization. Images were processed using either Nikon Elements or ImageJ/FIJI[61]. Image files (.nd2) were imported using the Bio-formats plugin[62]. Figure 2d was generated using ImageJ (under Plugins > 3D viewer)[63]. Chromosome schematics were generated using ChromoMap[64]. Figures were assembled in Adobe Illustrator. Micrograph images for publication figures were post-processed using brightness and contrast enhancement (ImageJ > Image > Adjust > Brightness/contrast). GraphPad Prism was also used for graphs. Molecular graphics and analyses were performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization and Informatics at the University of California and supported by the National Institutes of Health (NIH; grant no. P41-GM103311; ref. [65]).

Tier 1 detection. *Preprocessing.* Each round of OligoFISSEQ was imaged using five channels: Alexa Fluor 647, Texas Red, Cy3, Alexa Fluor 488 and DAPI and a series of *z*-slices. The *z*-stacks were deconvolved and background corrected using 20 iterations of the Richardson–Lucy algorithm using a theoretically calculated point spread function with Nikon software[66].

Rounds were compiled into hyperstacks composed of the five channels, a series of *z*-slices and one frame per round. If an image had all the puncta labeled, as with in toto images, it was included as a new additional frame. The hyperstacks were aligned using the Fiji plugin 'Correct 3D Drift' (ref. [67]). Images of DAPI-stained nuclei were used to perform threshold segmentation and extract each individual cell from the initial image as a separate region of interest. The segmentation provided information about the location and the envelope of the individual nuclei that made up each hyperstack. Nuclei with areas below 25 µm² were discarded.

Detection of barcodes. To compare intensities from different channels, images were normalized by dividing their intensities by the maximum intensity among the values of all the *z*-slices in the same round and channel.

For the detection of barcodes and for each round, the intensities of every pixel position were compared across different channels. A centroid-based pipeline using TrackMate[68] did not perform as well in our study; thus, we moved forward with this every-pixel approach. The channel with the highest value was kept as the prevalent channel. At every pixel position, the transition between channels along the different rounds was compared with the list of expected barcodes. A barcode was assigned to a pixel position if the set of transitions coincided with the one associated with the barcode. A maximum-intensity projection image was built by averaging the intensities of the prevalent channels from every round. Connected pixels that had the same barcode were grouped to form 3D patches. The following information was collected and saved for each patch:

- Barcode
- Center position
- Number of pixels forming part of the patch (size)
- Maximum intensity of the pixels of the patch
- Pixel position having the maximum intensity of the pixels of the patch

For an image with all puncta labeled, information on the intensity of each pixel position was stored in an additional file.

Tier 2 detection. *Chromosome tracing.* Patches composed of a single pixel location were discarded and the remainder were used in the tracing, disregarding the intensity or size of the patch.

Patches with high intensity values were selected as the most confident and were used to find the chromosome centers. We used an implementation of the constrained *k*-means algorithm[69] to find the center of the set of barcodes belonging to the same chromosome. To separate the homologs, we used a cannot-link constraint in the two copies of the same region to avoid having them in the same cluster. We used a sphere of radius 4.5 µm with origin in the centers to delimit the chromosome territory and filter out patches located outside.

The Domino sampler of the Integrative Modeling Platform[70] was the core element of the chromosome tracing. In Domino, each locus is represented by a particle with a finite set of different possible locations in the image. The locations are extracted from the list of patches having the same barcode as the one assigned to the locus. The remaining factors of the proposed problem are encoded in the system as restraints to the list of possible solutions. The following restraints are imposed by the system to filter compatible solutions:

- Two particles cannot share the same location or patch.
- Two consecutive particles of the same chromosome should be closer than a distance of 4 µm for the 36plex dataset and 1 µm for ChrX-46plex.
- Chromosomes must be confined in territories modeled as spheres of radius 4.5 µm.

Chromosome territory and the distance between consecutive regions were inferred as explained in 'Inferring chromosome territory and maximum distance between consecutive regions'. By applying these additional constraints to the barcodes, we were able to use patches that had intensities below, but not far from, the detection thresholds (Supplementary Table 14) and were likely to be true positives. Patches with higher intensities and sizes are most likely to be true-positive regions. Therefore, a score based on intensity and size was assigned to each patch as a measure of the likelihood of the patch being a true-positive detection. The list of patches was sorted by score and used as input data as an iterative process to find the most probable path of each chromosome (Supplementary Fig. 3).

The iterative process of tracing the chromosomes started by assigning patches with high scores to the corresponding regions. The process was executed once per

chromosome, considering all homologs at the same time because barcodes were not designed to distinguish them. Domino was used to list all possible solutions that were compatible with the imposed restraints. Each solution had a total score, obtained by summing the scores of the individual patches that were selected in that particular solution. We selected the conformation that had the highest total score. In the case where two or more solutions yielded an identical total score, we selected the solution that conformed to the shortest chromosome spatial length. An iterative process was performed for assignment of regions, whereby the threshold was lowered to allow more patches as input, and the previous approach was used to select the remaining unassigned regions. This iterative process was finished when all regions had been identified or there were no more input data to feed Domino.

Detection efficiency and false-positive ratios. To calculate the detection efficiency per barcode, the datasets were filtered using intensity thresholds (Supplementary Table 14) that were optimized for every experimental condition. Patches formed by a single pixel were also discarded regardless of the intensity of the patch.

For 36plex datasets, we calculated the mean of the barcodes detected per nuclei, excluding barcodes assigned to the X chromosome. In the ideal case, and due to the ploidy, we expected two barcodes per nucleus. In reality, the datasets may eventually include false positives or duplicates of patches that probably belong to the same oligonucleotide, which will increase the ratio. Nuclei with a mean of more than 2.5 barcodes were discarded because they were most likely in a mitotic process. For ChrX-46plex, we followed a similar procedure and discarded nuclei with a mean of detected barcodes that was greater than 1.5.

For each of the remaining nuclei, we computed the ratio of detected barcodes versus expected barcodes. We expected two barcodes per cell, except for the barcodes belonging to the X chromosome. The ratios per barcode and per cell were capped to 1 and averaged over all cells to produce the detection efficiency. For the false-positive ratio of the barcode, we instead calculated the excess of detections as the detected value minus the expected value in cases where the detected value was over the expected value, and we then computed the ratio of excess detections versus the expected values.

Distance heat maps and Hi-C maps. For every traced nucleus, we calculated all pairwise distances between the detected regions and averaged the results among all cells. For the average heat map of 36plex-5K, LIT dataset regions 3qR3 and 5pR3 were not taken into consideration because they shared the same barcode and were therefore indistinguishable. Hi-C maps of PGP1f cells were obtained from previous in situ Hi-C experiments[14]. The values of the interaction frequencies in the included Hi-C maps were extracted from the observed values of interaction matrices produced at a resolution of 5 Kb. The submatrices formed by the genomic regions of each pair of probes were aggregated to obtain the interregional observed interaction. Single-cell heat maps were built with the identification of homologous chromosomes. The list of barcodes was traced according to the procedure described in above in the 'Chromosome tracing' section of 'Tier 2 detection'. All pairwise distances of the traced regions were calculated. Non-identified regions appear as gray columns and rows.

Inferring chromosome territory and the maximum distance between consecutive regions. To infer the maximum distance between consecutive regions used in the chromosome tracing, the list of detected barcodes for all 36plex datasets was filtered to discard mitotic cells as explained in 'Detection efficiency and false-positive ratios'. Patches formed by a single pixel were also filtered out. After the filtering process, the 36plex dataset comprised 1,171 nuclei and 48,352 barcodes. Then, we calculated the distances between consecutive regions for each chromosome in each nucleus (Supplementary Fig. 1). The histograms of those distances show the expected bimodal distributions for the chromosomes, except for chromosome X as foreseen from male cells. Bimodality is more evident in bigger chromosomes because those tend to be in the periphery of the nucleus, while smaller chromosomes prefer the interior. After inspection of the histograms, we selected 4 µm as the general maximum distance between consecutive regions and a slightly higher value of 4.5 µm for the chromosome territory.

For the ChrX-46plex dataset, we followed a similar approach. After the filtering process, ChrX-46plex contained 189 nuclei and 7,752 barcodes. Based on the histograms of distances between consecutive regions, we selected 2.5 µm as the general maximum distance (Supplementary Fig. 2).

Clustering of 3D structures for ChrX-46plex. After tier 2 detection, we had 177 cells for the ChrX-46plex library, with an average of 34 detected regions. We discarded one of the cells that had fewer than 23 identified barcodes so as to meet the required 50% detection efficiency per cell in all the 3D structures. Next, we calculated the pairwise distances for each chromosome between all of their detected targets and used these as a measure of similarity to the built distance matrix. We used the coincident distances between structures to cluster them hierarchically using the Ward method. The Calinski–Harabasz criterion for clustering evaluation was used to evaluate the optimal number of clusters.

ChrX-46plex-2K tracing in IMR-90 cells. O-eLIT with the ChrX-46plex-2K library in IMR-90 cells was performed as in PGP1f cells. Five rounds of sequencing were performed off both streets, followed by immunostaining for MacroH2A.1 (Abcam, ab183041) at 1:250 dilution to mark the inactive X chromosome. For the every-pixel analysis, chromosome traces with fewer than 13 identified regions were filtered out.

MacroH2A.1 IF images were aligned and segmented with the DAPI channel of their OligoFISSEQ correspondence. For each nucleus, the position of maximum intensity of the IF image was compared with the geometric center of the traced X chromosomes. To filter out images without a clear IF signal, we only considered nuclei where their maximum IF intensity was greater than two times the average intensity inside. If the center was closer than 3.5 µm, the X chromosome was considered IF positive and annotated as inactive (Xi). The other X chromosome in the nucleus was annotated as active (Xa). In cases where both homologs were closer than 3.5 µm to the IF signal, the closest homolog was annotated as Xi and the farthest was annotated as Xa. The nuclei were manually checked to discard errors, mainly due to overlapping cells that resulted in 40 Xi chromosomes, for which we traced 31 homologs that were identified as Xa.

Generation of random nuclei for haploid separation. For the directed random nuclei, we first calculated the mean and s.d. of the distance to the nuclear envelope for every chromosome (Supplementary Table 15). We used this information to generate a set of random nuclei where the chromosomes were randomly placed following a normal distribution in which the mean and s.d. were equal to the values calculated in the observed data. The positions of the large chromosomes in the synthetic nuclei were biased toward the periphery, while the positions of the small chromosomes were biased toward the nuclear interior. No spatial bias was used for the completely random nuclei.

Histogram of split homologs by *k*-means. For the analysis, we selected 258 nuclei for which all centers of the 11 chromosomes were known. We used the conventional *k*-means algorithm to cluster the positions of the chromosomes into two groups and reported how many autosomes were split by the clustering, that is, how many autosomes had one copy in one of the groups and the homolog in the other group.

Method for the alignment of nuclei. For the analysis, we selected 258 nuclei for which all centers of the 11 chromosomes were known. We used an implementation of the constrained *k*-means algorithm[69] to cluster the chromosomes into two groups: one group contained one copy of each autosome and the other group contained the homolog. The X chromosome was assigned to the closest group. The geometric centers of the clusters were joined and the resulting segment, together with all the positions of the chromosomes, was rotated to be parallel to the x axis and moved to leave the middle point toward the origin; $x=0$, $y=0$. In the rotation of the nuclei, we kept the group containing the X chromosome at the left of the y axis.

Density plots. The density plots were built using the kernel density estimation of the projection to the xy plane of the position of the chromosomes.

Number of split homologs. We checked each aligned nucleus and reported how many autosomes could be split by a virtual line along the y axis, that is, the number of autosomes with one of the copies on the left of the y axis and the other on the right of the y axis.

Number of split homologs left to right. We checked each aligned nucleus and reported how many autosomes could be split by a virtual line parallel to the y axis at different distances from the origin, that is, the number of autosomes with one of their homologs on the left of the line and the other on the right of the line.

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

Data availability

All data are available in the main text or the supplementary materials, and materials are available upon request. Information regarding all datasets (for example, cells, replicates and filters) can be found in Supplementary Table 9. Source data are provided with this paper.

Code availability

All code is available at https://github.com/3DGenomes/OligoFISSEQ/.

References

51. Beliveau, B. J. et al. OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. *Proc. Natl Acad. Sci. USA* **115**, E2183–E2192 (2018).
52. Ball, M. P. et al. A public resource facilitating clinical use of genomes. *Proc. Natl Acad. Sci. USA* **109**, 11920–11927 (2012).
53. Zhang, K. et al. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods* **6**, 613–618 (2009).

54. Pardue, M. L. et al. Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proc. Natl Acad. Sci. USA* **64**, 600–604 (1969).
55. Bauman, J. G., Wiegant, J., Borst, P. & van Duijn, P. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. *Exp. Cell Res.* **128**, 485–490 (1980).
56. Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
57. Valouev, A. et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).
58. Guo, J. et al. Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc. Natl Acad. Sci. USA* **105**, 9145–9150 (2008).
59. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
60. Moffitt, J. R. et al. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl Acad. Sci. USA* **113**, 14456–14461 (2016).
61. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
62. Linkert, M. et al. Metadata matters: access to image data in the real world. *J. Cell Biol.* **189**, 777–782 (2010).
63. Schmid, B., Schindelin, J., Cardona, A., Longair, M. & Heisenberg, M. A high-level 3D visualization API for Java and ImageJ. *BMC Bioinformatics* **11**, 274 (2010).
64. Anand, L. ChromoMap: an R package for interactive visualization and annotation of chromosomes. Preprint at *bioRxiv* https://doi.org/10.1101/605600 (2019).
65. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
66. Richardson, W. H. Bayesian-based iterative method of image restoration. *J. Opt. Soc. Am.* **62**, 55–59 (1972).
67. Parslow, A., Cardona, A. & Bryson-Richardson, R. J. Sample drift correction following 4D confocal time-lapse imaging. *J. Vis. Exp.* 51086 (2014).
68. Tinevez, J.-Y. et al. TrackMate: an open and extensible platform for single-particle tracking. *Methods* **115**, 80–90 (2017).
69. Wagstaff, K., Cardie, C., Rogers, S. & Schroedl, S. Constrained *k*-means clustering with background knowledge. In *Proc. 18th International Conference on Machine Learning*, 577–584 (2001).
70. Russel, D. et al. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* **10**, e1001244 (2012).

Acknowledgements

We acknowledge members of the Marti-Renom and Wu laboratories for technical and conceptual support, especially T. Ryu, A. Lioutas and S. Aufmkolk as well as J. AlHaj Abed, S. D. Lee, J. Erceg and T. Hatkevich; B. Beliveau, H. Sasaki, J. Horrell, L. Cai, J. Kishi and P. Soler-Vila for discussion; D. Barclay, R. Kohman, E. Iyer, K. Rodgers, A. Skrynnyk, J. Tam and R. Terry for discussion about FISSEQ and sequencing reagents; S. Alon, F. Chen, Z. Chiang, D. Goodwin, A. Payne, A. Sinha and O. Wassie for discussion about FISSEQ; C. Ebeling, J. Rosenberg and J. Stuckey for discussion and technical assistance; F. Pan and A. Hutchinson for assistance in procuring SOLiD reagents; P. Montero-Llopis and the MicRoN imaging core at Harvard Medical School; the ImageJ discussion forum; and StackOverflow. This work was supported by a Damon Runyon Dale F. Frey Breakthrough Award (to B.J.B.) to support B.J.B. and E.A.H., awards from the NSERC of Canada (PGS D) to P.L.R., the NIH (HG005550 and HG008525) and NSF (DGE1144152) to E.R.D., the European Research Council under the Seventh Framework Program (FP7/2007–2013 609989), the European Union's Horizon 2020 Research and Innovation Program (676556) and the Spanish Ministerio de Ciencia, Innovación y Universidades (BFU2017-85926-P) to M.A.M.-R., the Centro de Excelencia Severo Ochoa 2013–2017 (SEV-2012-0208) and the CERCA Programme/Generalitat de Catalunya to the CRG, from the NIH to GMC (RM1HG008525-03) and the NIH (DP1GM106412, R01HD091797 and R01GM123289) to C.-t.W.

Author contributions

H.Q.N., S.C., D.C., S.C.N., G.M.C., E.R.D., M.A.M.R. and C.-t.W. conceived the study with the original conceptualization of OligoFISSEQ contributed by S.C.N and E.R.D.; G.N., A.L. and N.M.C.M. provided guidance for barcode design and angle analysis; A.L., E.A.H. and B.J.B. provided guidance for Oligopaint sequences and barcode design. P.L.R. supported early protocol development; M.H. provided technical support; H.Q.N. and S.C. designed and performed the experiments. H.Q.N., S.C., D.C., G.M.C., M.A.M.R. and C.-t.W. analyzed the data; H.Q.N. wrote the manuscript with S.C., D.C., M.A.M.R. and C.-t.W. with input from all authors; C.-t.W. oversaw the project.

Competing interests

Harvard University has filed patent applications on behalf of C.-t.W., H.Q.N. and S.C., pertaining to Oligopaints and related oligonucleotide-based methods for genome imaging. E.R.D. is currently an employee of ReadCoor and has an equity interest in ReadCoor. Potential conflicts of interest for G.M.C. are listed on http://arep.med.harvard.edu/gmc/tech.html/. C.-t.W. has an equity interest in ReadCoor and an active research collaboration with Bruker Nano in her laboratory at Harvard Medical School.

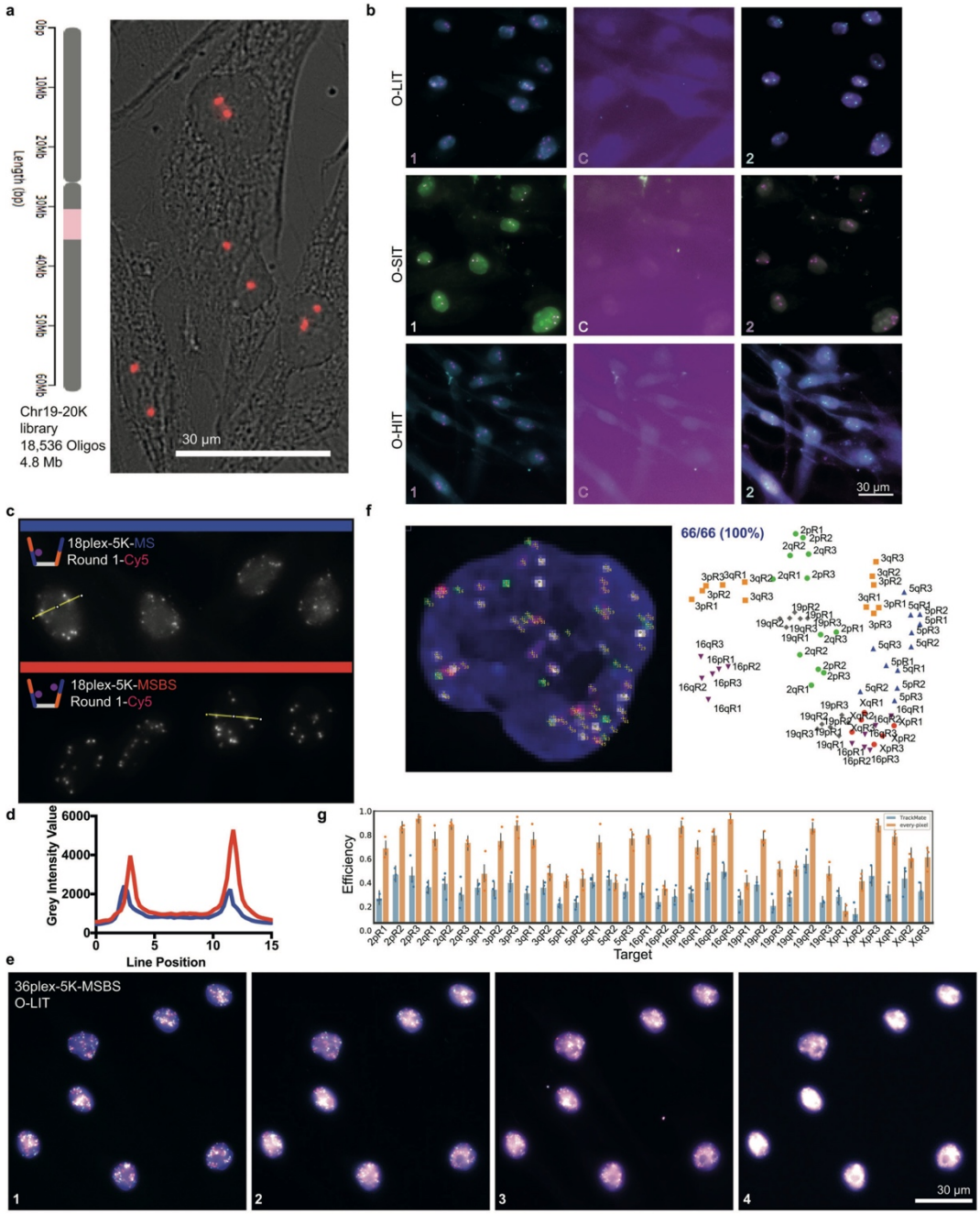
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-020-0890-0.

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0890-0.

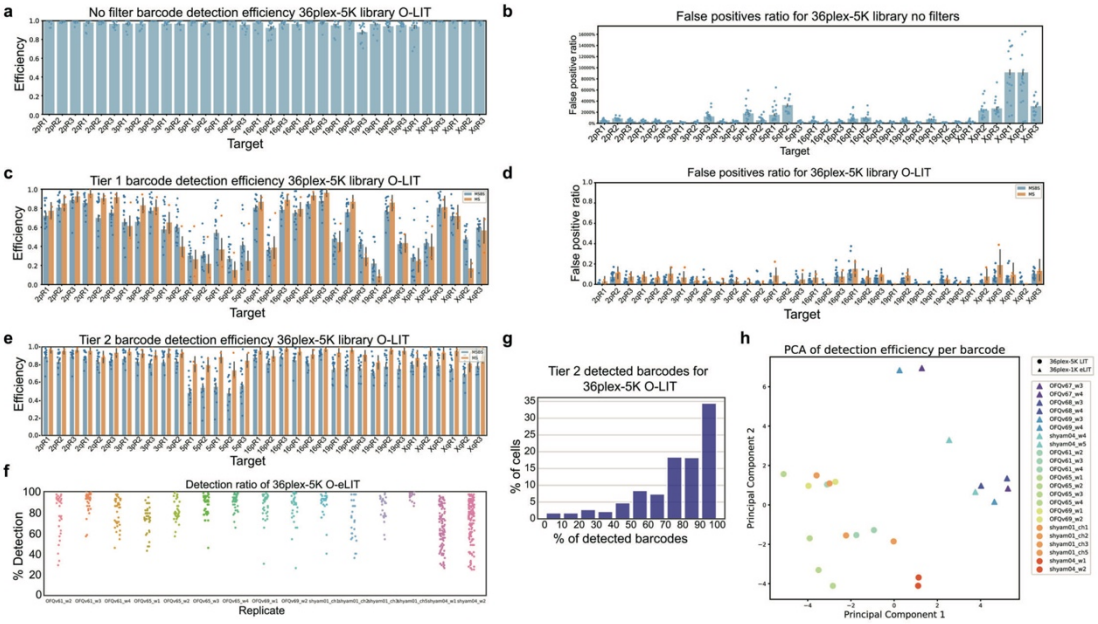
Correspondence and requests for materials should be addressed to M.A.M.-R. or C.-t.W.

Peer review information Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

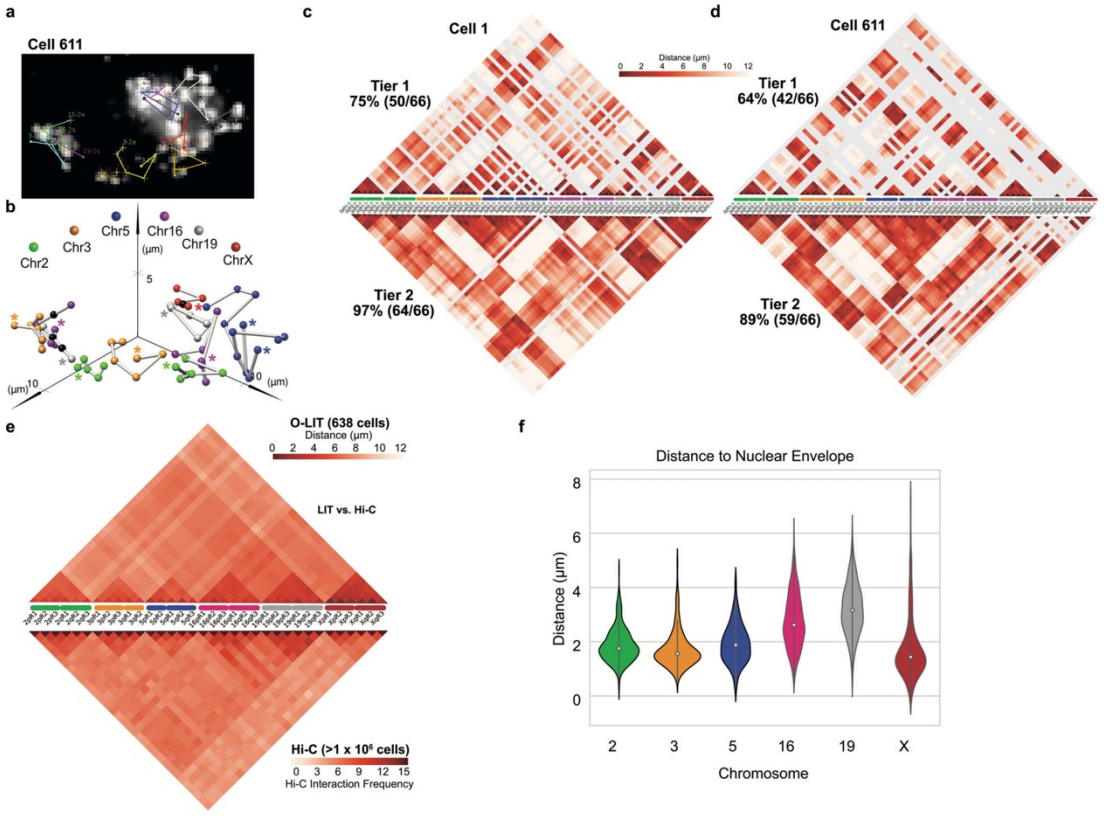


Extended Data Fig. 1 | See next page for caption.

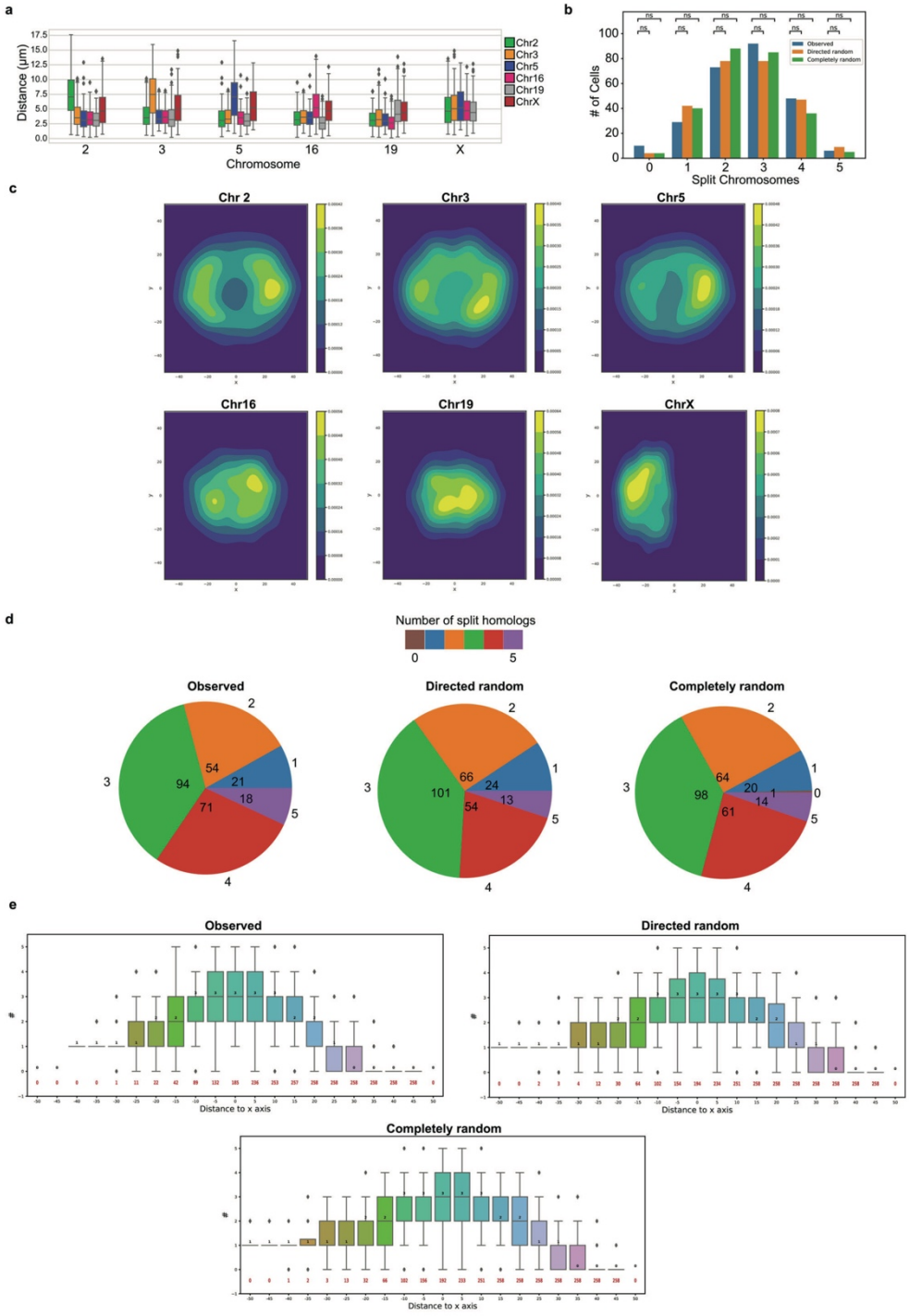
Extended Data Fig. 1 | Chr19-20K and 36plex-5K-O-LIT optimization. a, Chr19-20K targets 18,536 Oligopaint oligos to human chromosome 19. Right, Chr19-20K detection with secondary oligo (red) in PGP1f cells representative of 5 replicates. **b**, Signal is completely removed in each OligoFISSEQ method after cleavage. Images showing two rounds of sequencing with a cleavage step (C) and representative of 4 replicates. **c**, 36plex-5K O-LIT off of both Mainstreet and Backstreet (MSBS; bottom, red) produces stronger signal than off of Mainstreet (MS; top, blue). Cy5 channel from first round of O-LIT. n = 1. **d**, O-LIT off of both streets produces stronger signal than off of MS. Grey intensity value measurements from yellow lines in panel c. n = 1. **e**, Raw, non-deconvolved field of view of cell from Figs. 2c, d and 3a–c. Maximum z-projection. n = 1. **f**, Manual decoding of cell from panel c and Figs. 2c, d and 3a–c yields 100% target recovery. n = 1. **g**, Tier1 detection efficiency after 36plex-5K O-LIT off of both streets and detected with TrackMate (blue, 29.93 ± 4.9%) or Every-pixel (orange, 62.8% ± 4.8%). n = 111 cells from 3 replicates. Detection efficiency from individual replicates are plotted. Error bars represent 95% bootstrap confidence interval of the mean.



Extended Data Fig. 2 | Detection efficiency after 36plex-5K O-LIT. a, Detection efficiency without filtering after 36plex-5K O-LIT off of both streets. 95 ± 5.15% of targeted regions are detected (n = 611 from 15 replicates). Detection efficiency from individual replicates are plotted. Error bars represent 95% bootstrap confidence interval of the mean. **b**, False positive (FP) discovery rate from panel a. FP discovery rate from individual replicates are plotted. Error bars represent 95% bootstrap confidence interval of the mean. **c**, Tier 1 detection efficiency after 36plex-5K O-LIT off of Mainstreet (orange, 61.93 ± 12%, n = 53 from 2 replicates) versus off of both streets (blue, 62.17% ± 6.68%, n = 611 cells from 15 replicates). Detection efficiency from individual replicates are plotted. Error bars represent 95% bootstrap confidence interval of the mean. **d**, FP discovery rate from panel c. Using Mainstreet = 8.64% and using both streets = 5.29%. FP discovery rate from individual replicates are plotted. Error bars represent 95% bootstrap confidence interval of the mean. **e**, Tier 2 detection efficiency after 36plex-5K off of Mainstreet (orange, 92.3% ± 3.42% from 53 cells from 2 replicates) versus off of both streets (blue, 80.19 ± 7.29%, n = 611 cells from 15 replicates). Detection efficiency from individual replicates are plotted. Error bars represent 95% bootstrap confidence interval of the mean. **f**, Detection efficiency after 36plex-5K O-LIT off of both streets for individual cells from 15 replicates in panel **e**. **g**, Percentage of cells displaying a range of efficiencies of barcode detection after 36plex-5K O-LIT off of both streets. Data taken from panel **e**. **h**, Principal component analysis showing lack of batch effect in 36plex datasets (n = 1171 cells from 15 36plex-5K O-LIT replicates and 8 36plex-1K O-eLIT replicates).

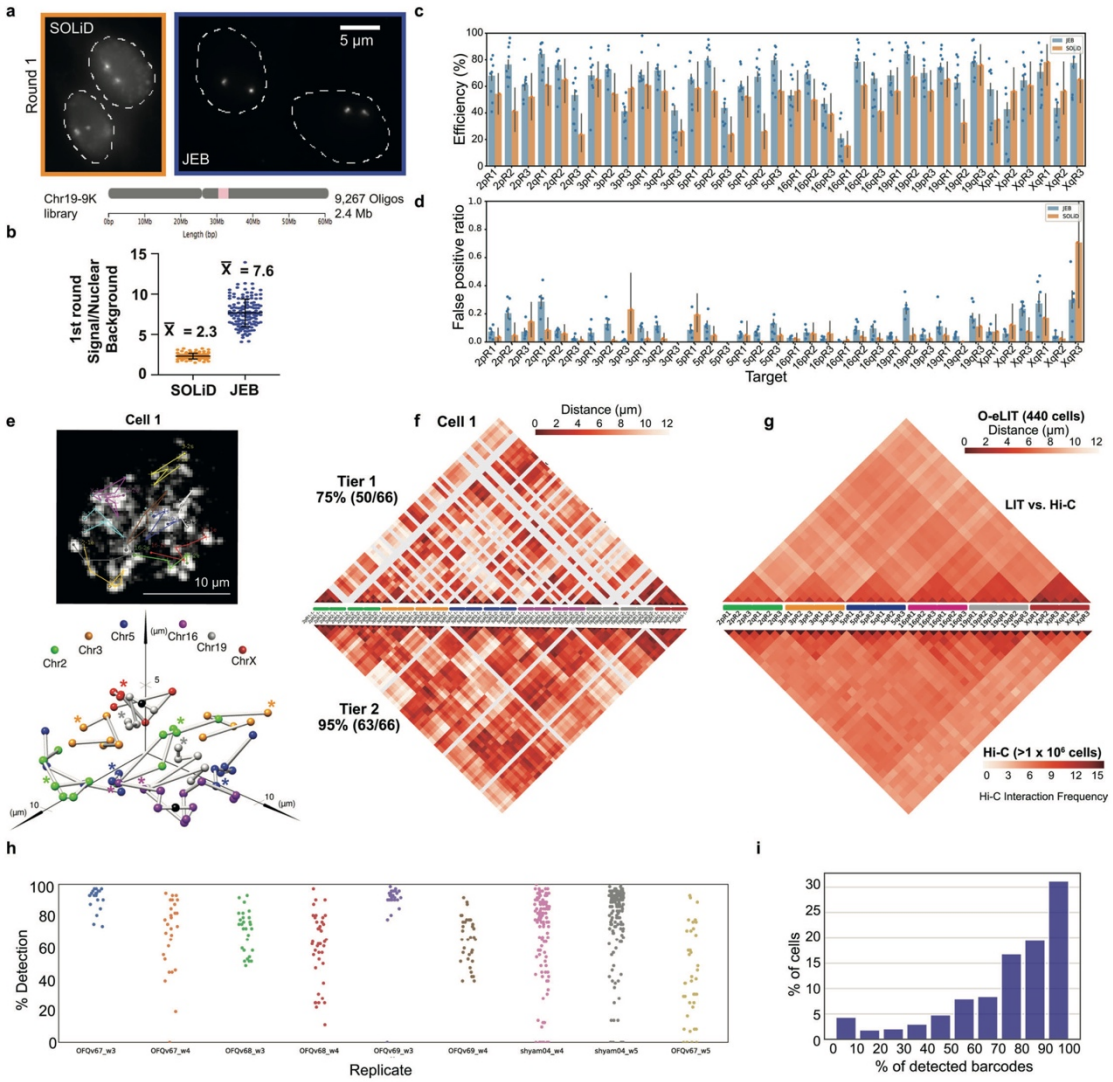


Extended Data Fig. 3 | O-LIT with 36plex-5K to interrogate genome organization. a, Chromosome traces of Cell 611 after Tier 2 detection of cell 611 after four rounds of O-LIT 36plex-5K off of both streets. 59/66 (89%) of 36plex-5K targets were detected. Image is from the first round of O-LIT with target identities. n = 1. **b**, Ball and stick of Cell 611. Colored spheres represent chromosomal targets, while black spheres represent targets that were not detected and, thus, were placed by calculating the median proportionate distance between flanking detected targets. Beginning of chromosome (for example 2pR1) marked by an asterisk. **c**, Single-cell pairwise spatial distance matrix after Tier 1 (top) and Tier 2 (bottom) detection of the nucleus in Fig. 3. Targets are represented on the x-axis with homologs separately displayed. Undetected targets are represented by grey lines. **d**, Single-cell pairwise spatial distance matrix after Tier 1 (top) and Tier 2 (bottom) detection of Cell 611. Targets are represented on the x-axis with homologs separately displayed. Undetected targets are represented by grey lines. **e**, 36plex-5K population pairwise spatial distances (top, from Fig. 3f). Average pairwise spatial distances from cell population after Tier 1 detection (n = 611 from 15 replicates). (Spearman's rank correlation 0.705, two-sided p-value for a hypothesis test whose null hypothesis is that two sets of data are uncorrelated = 1.77e-174). Measurements from homologous targets were combined. Bottom, Hi-C data of 36plex-5K targets obtained from (Nir et al. 2018). **f**, Average distances between the nuclear membrane and the closest of the six targets imaged for each chromosome. (n = 686, 668, 364, 586, 760, and 494 for Chr2, 3, 5, 16, 19, and X, respectively.) The thick line in each violin plot represents the Interquartile range (IQR), the white dot marks the median and the thin lines extend 1.5 times the IQR.

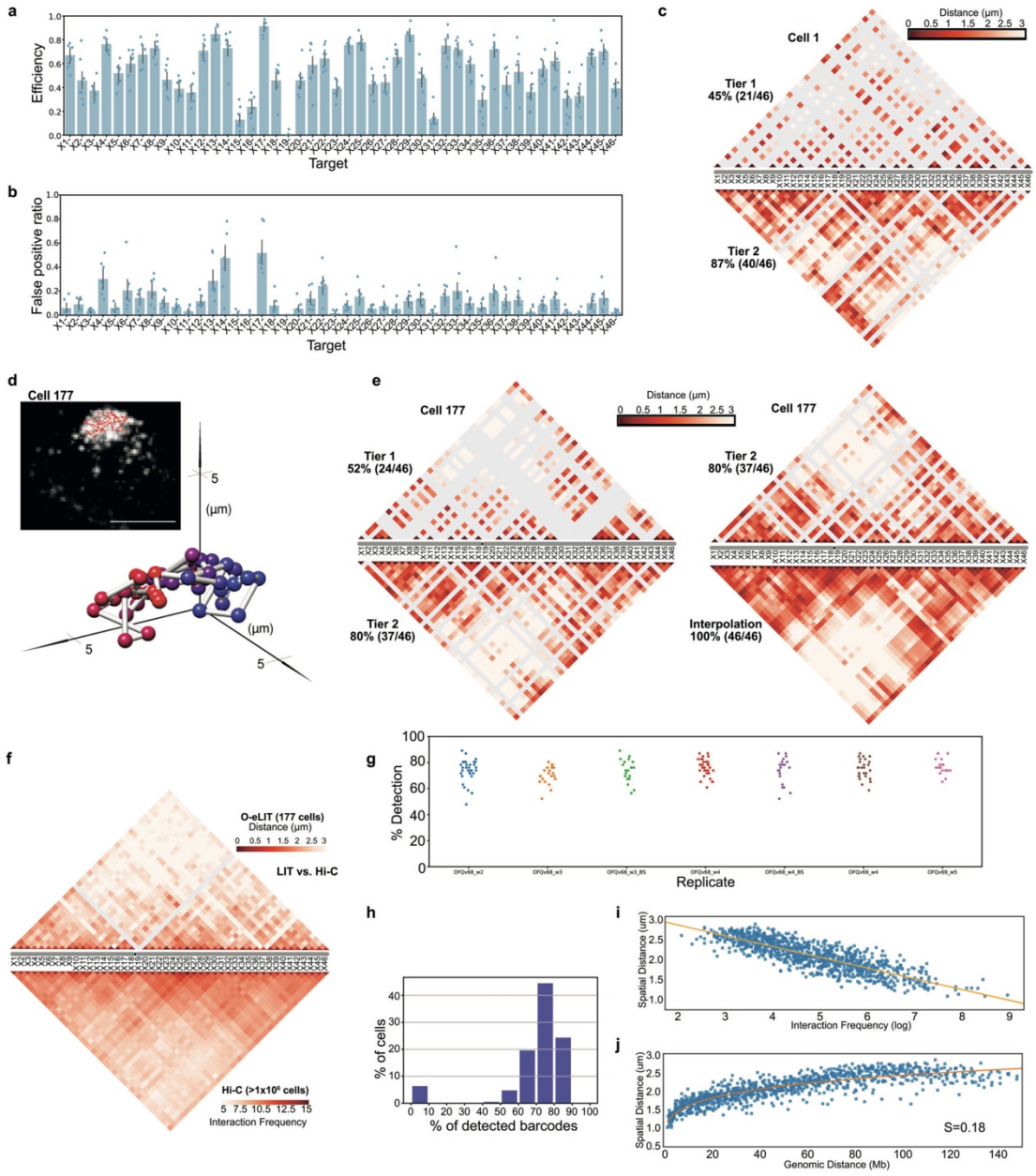


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | O-LIT with 36plex-5K to interrogate homolog organization. a, Minimum distances between heterologous and homologous chromosomes. All measurements represent distances between the geometric centers of chromosomes for which all six targets were imaged. Distances between a chromosome and a heterologous chromosome is the shorter of the two distances between that chromosome and the two homologous copies of the heterologous chromosome (n = 686, 668, 364, 586, 760, and 494 for Chr2, 3, 5, 16, 19, and X, respectively). Inter-homolog distances for Chr16 and 19 are less than those for Chr2, 3, and 5 (independent-samples t-test p = 4.28 ×10-37). Boxes represent the IQR (25th, 50th and 75th percentiles) and whiskers extend 1.5 times the IQR. **b**, Number of cells with varying numbers of homologs split by K-means clustering. The K-means algorithm was applied to 258 nuclei, individually, to cluster chromosomes into two groups based on proximity and then report the number of homolog pairs that were split by the clustering. A value of "5" indicates that the homologs from each five pairs of imaged autosomes in a single nucleus clustered into two spatially separate groups. Observed, PGP1f cells. Directed random, raw positions in Observed but with the chromosome identities of all positions randomized, with the larger chromosomes (2, 3, 5) biased towards the nuclear periphery and smaller chromosomes (16 and 19) biased towards the nuclear interior. Completely random category, randomization of the chromosome identities carried out with no spatial bias. The significance of each pair was evaluated from a two proportion z-test with n = 258 for each category with a null hypothesis of equal proportion and a significance level of 0.05. **c**, Density plots of homolog positions. Built by using Kernel density estimation (KDE) of nuclei projected and aligned along the x-y plane of the position of the chromosomes. **d**, Pie charts of total number of cells with homologs split by a virtual line along the y-axis. **e**, Number of aligned cells with homologs split by a virtual line parallel to the y-axis at different distances from the origin, that is, number of autosomes with one of their homologs on the left of the line and the other on the right (n = 258 for each category). Boxes represent the IQR (25th, 50th and 75th percentiles) and whiskers extend 1.5 times the IQR.

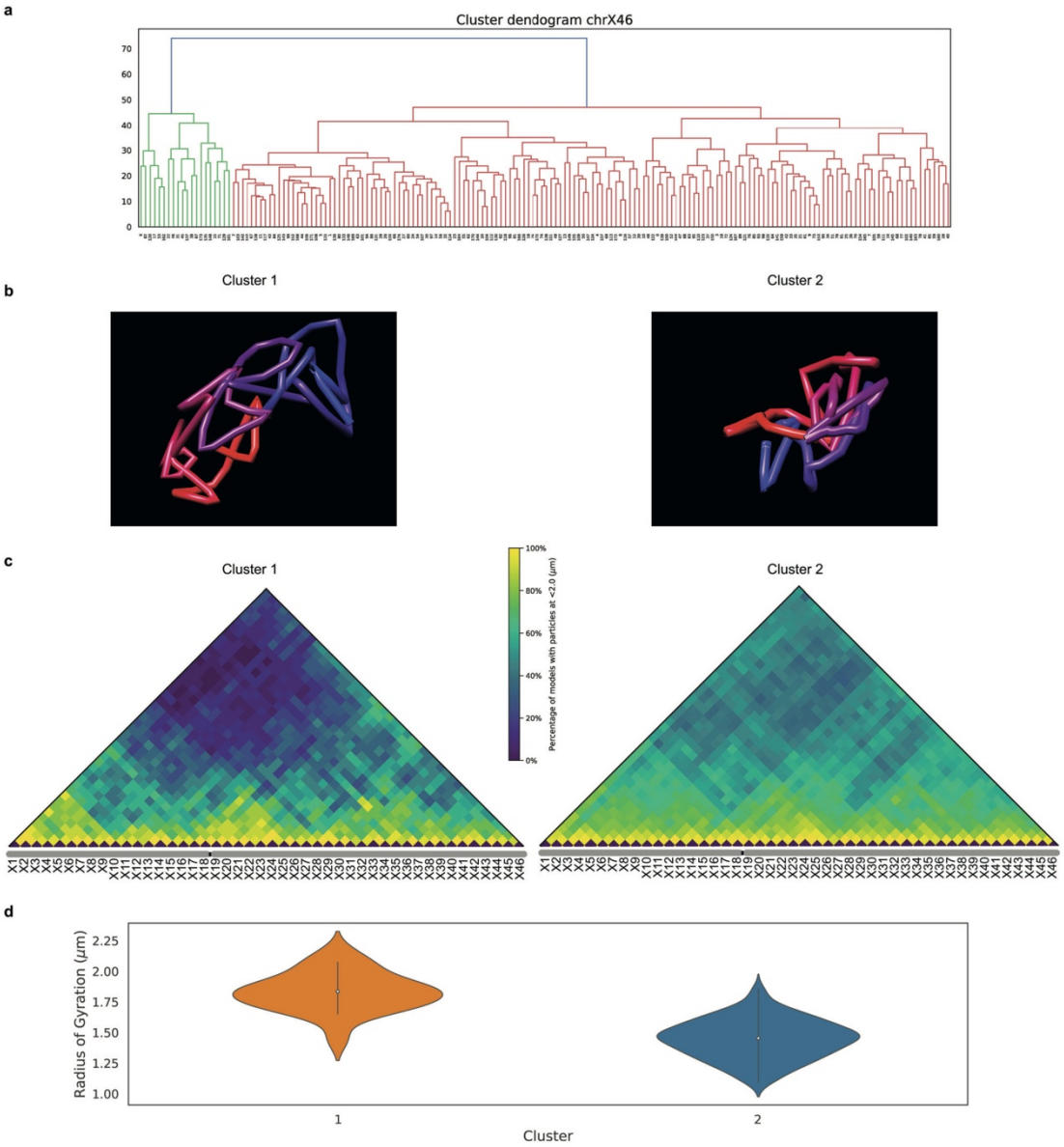


Extended Data Fig. 5 | O-eLIT with JEB. a, Chr19-9K. One round of O-LIT (SOLiD) or O-eLIT (JEB) off of Mainstreet. Maximum z-projections representative of 2 replicates. **b**, Chr19-9K signal over nuclear background measurements after one round of O-LIT (orange; n = 113 puncta from 55 cells from 2 replicates) or O-eLIT (blue; n = 136 puncta from 57 cells from 2 replicates). Bar is the mean and SD. **c**, Tier 1 detection of 36plex-1K after five rounds of O-LIT with SOLiD reagents (orange; average of 51.75%, n = 41) or O-eLIT with JEB (blue; average of 61.2 ± 10.2%, n = 440 from 9 replicates). Detection efficiency from individual replicates are plotted. Error bars represent 95% bootstrap confidence interval of the mean. 36plex-1K library shares first 1,000 Oligopaint oligos of each target in 36plex-5K. For example, for target 2pR1, 36plex-5K spans the chromosomal region from nt position 1,002,895 to 1,660,898 (~658 kb), whereas 36plex-1K spans the region from nt 1,002,895 to 1,147,495 (~144 kb). **d**, FP discovery rate from panel c. SOLiD = 7.49% and JEB = 8.95%. FP discovery rate from individual replicates are plotted. Error bars represent 95% bootstrap confidence interval of the mean. **e**, Chromosome traces and ball and stick of Fig. 4c cell after Tier 2 detection and five rounds of O-eLIT 36plex-1K. 63/66 (95%) targets were detected. Asterisks, beginning of chromosomes. n = 1. **f**, Single-cell pairwise spatial distance matrices of panel C cell. **g**, 36plex-1K population pairwise spatial distance measurements (top, from Fig. 3f). Average pairwise spatial distance from cell population after Tier 1 detection (n = 440 from 9 replicates). Measurements from homologous targets were combined. Bottom, Hi-C data of 36plex-5K targets obtained from (Nir et al. 2018). **h**, 36plex-1K detection rate for individual cells from 9 replicates. **i**, Percentage of cells displaying a range of efficiencies of barcode detection after 36plex-1K O-eLIT off of Mainstreet.

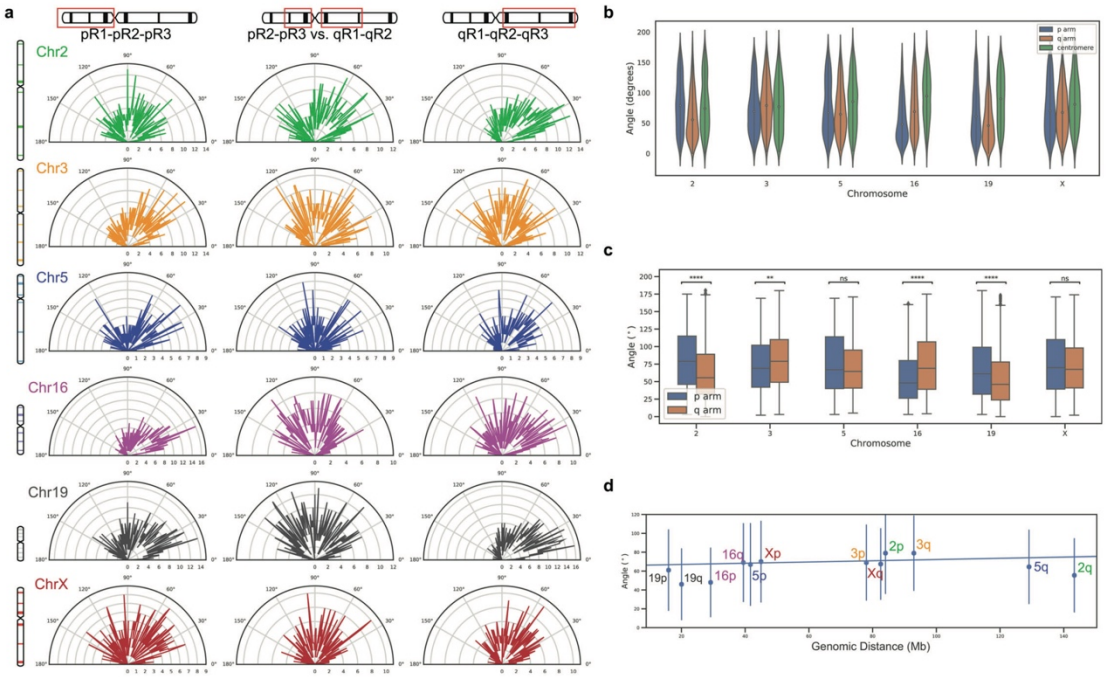


Extended Data Fig. 6 | See next page for caption.

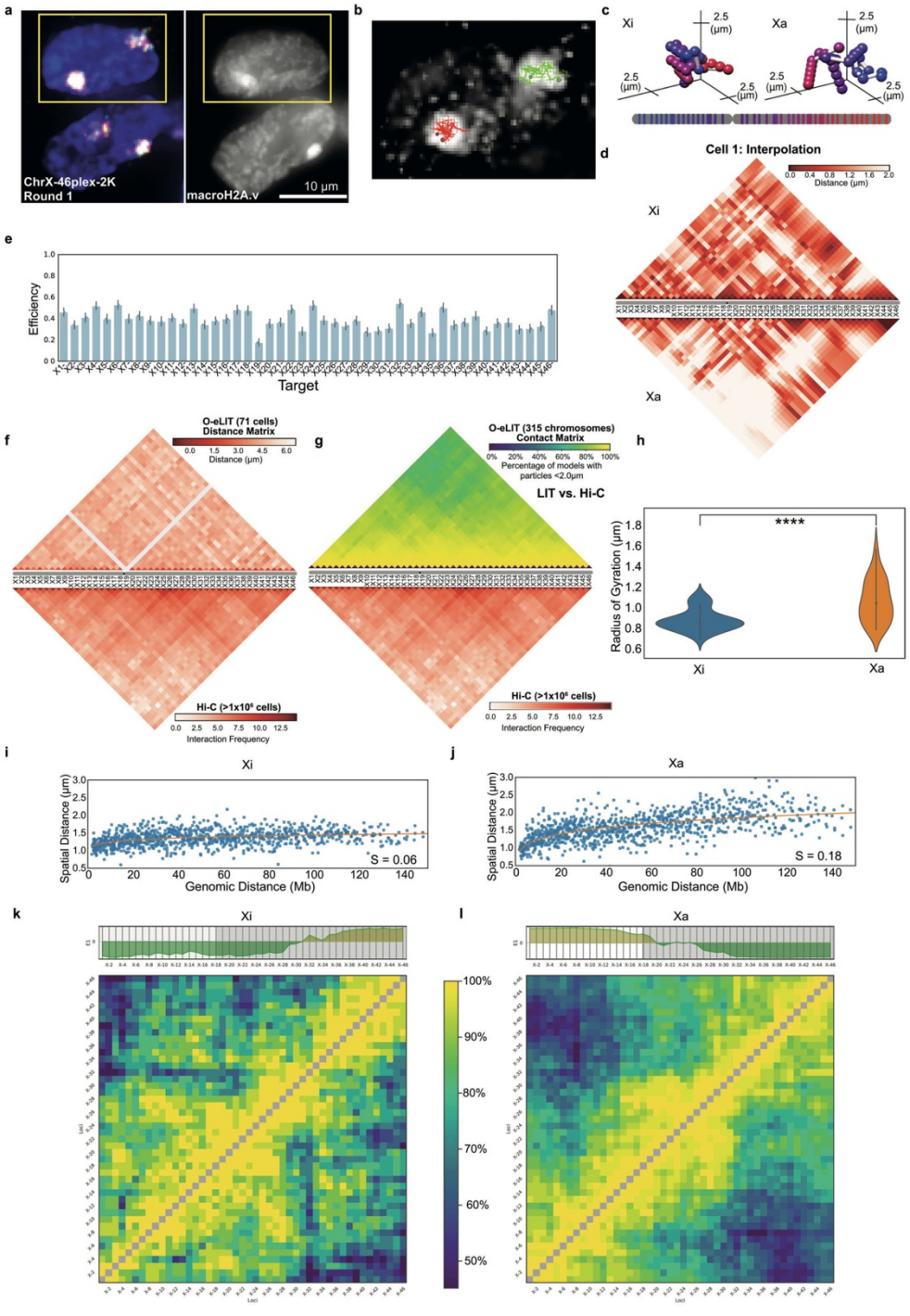
Extended Data Fig. 6 | O-eLIT with ChrX-46plex-2K. a, ChrX-46plex-2K O-eLIT Tier 1 detection off of one street and off of both streets combined (52.86 ± 5.78% from 177 cells from 7 replicates). Detection efficiency from individual replicates are plotted. Error bars represent 95% bootstrap confidence interval of the mean. **b**, FP discovery rate from panel a. Error bars represent 95% bootstrap confidence interval of the mean. **c**, Single-cell pairwise spatial distance matrix after Tier 1 (top) and Tier 2 (bottom) detection of Cell 1 from Fig. 5b. Undetected targets are represented by grey lines. **d**, Chromosome traces (top) and ball and stick representation (bottom) of Cell 177 after Tier 2 detection and interpolation and five rounds of O-eLIT on ChrX-46plex-2K off of both streets. Image is from the first round of O-eLIT with target identities. n = 1. **e**, Single-cell pairwise spatial distance matrix after Tier 1 (top), Tier 2 (bottom) of Cell 177 (left), and Tier 2 (top) and interpolation (bottom) of same cell (right). Undetected targets are represented by grey lines. **f**, ChrX-46plex-2K population pairwise spatial distances (top). Average pairwise spatial distances from cell population after Tier 1 detection (n = 177 from 7 replicates). Bottom, Hi-C (Nir et al. 2018) data of ChrX-46plex-2K targets. (Spearman's rank correlation 0.641, two-sided p-value for a hypothesis test whose null hypothesis is that two sets of data are uncorrelated = 7.074e-245). **g**, ChrX-46plex-2K detection rate for individual cells from 7 replicates. **h**, Percentage of cells displaying a range of efficiencies of barcode detection after ChrX-46plex-2K O-eLIT. **i**, Mean spatial distance versus Interaction frequency of Hi-C (Nir et al. 2018) of ChrX-46plex-2K targets. Pearson correlation coefficient (r = −0.84) and p-value = 5.08E-275 (two-sided, using slope = 0 for null hypothesis and Wald Test with t-distribution as test statistic) of the linear least-squares regression. **j**, Mean spatial distance versus genomic distance for all pairwise ChrX-46plex-2K targets (n = 177 from 7 replicates).



Extended Data Fig. 7 | O-eLIT identifies clusters after ChrX-46plex O-eLIT. a, Hierarchical clustering based on structure of ChrX traces from ChrX-46plex after 5 rounds of O-eLIT and Tier 2 detection yielded two clusters (Cluster 1 = 20; Cluster 2 = 156). See Methods for more details. **b**, ChrX representative models (existing traces that are closer to the virtual centroid) of the two clusters obtained after Hierarchical clustering in panel **a**. **c**, ChrX-46plex-2K population contact matrix of two clusters derived after Hierarchical clustering in panel a where pairwise spatial distances are considered to be in contact if less than 2 μm apart. **d**, Radius of gyration for the two clusters (Cluster 1 = 20; Cluster 2 = 156) derived after the hierarchical clustering shown in panel a. The thick line in each violin plot represents the Interquartile range (IQR), the white dot marks the median and the thin lines extend 1.5 times the IQR.

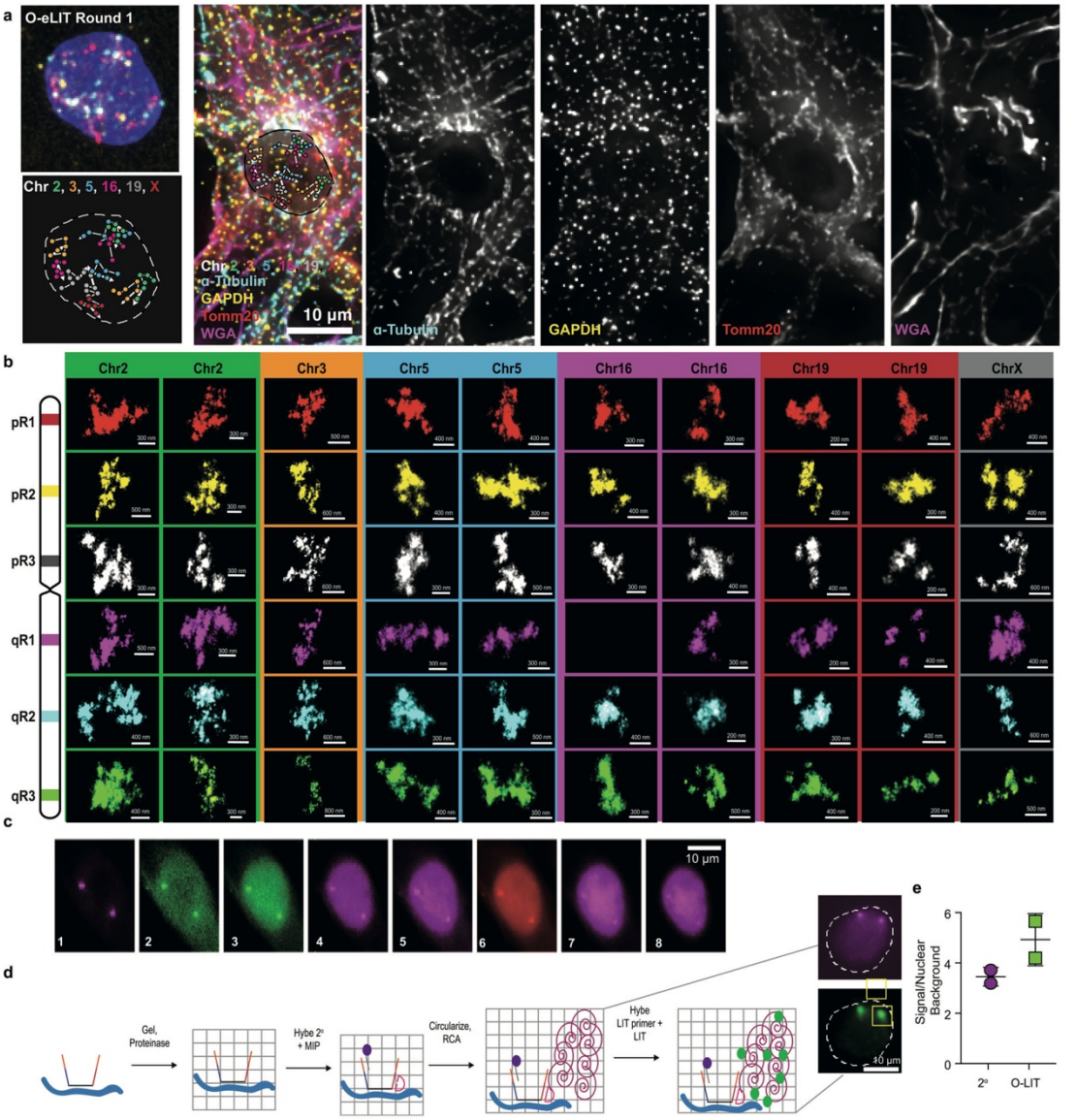


Extended Data Fig. 8 | Angles from 36plex. a, Measurements of angles formed by three points along the p arm (left), q arm (right), and intersection of vectors formed by pR2-pR3 and qR1-qR2 (middle) for each chromosome. Measurements were obtained by combining data from 36plex-5K and 36plex-1K analyses and selecting chromosomes that had all six targets identified. Chr2: n = 686, Chr3: n = 668, Chr5: n = 363, Chr16: n = 586, Chr19: n = 760, ChrX: n = 493 (n = 1,051 cells from 24 replicates; for 36plex-5K, n = 611 from 15 replicates; for 36plex-1K, n = 440 from 9 replicates). **b**, Distribution of angles formed by segments in panel a. The thick line in each violin plot represents the Interquartile range (IQR), the white dot marks the median and the thin lines extend 1.5 times the IQR. **c**, Box plots comparing p and q arm angles. Two-sided student's t-test with null hypothesis of equal mean was performed to compare arms, ns $p > 0.05$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$. Boxes represent the IQR (25th, 50th and 75th percentiles) and whiskers extend 1.5 times the IQR. Sample size information in a). Exact p-values for each chromosome: Chr.2 = 4.149e-16, Chr.3 = 0.004, Chr.5 = 0.093, Chr.16 = 1.357e-14, Chr.19 = 3.325e-11, Chr.X = 0.101. **d**, Linear least-squares regression between arm angle and arm length with Pearson correlation coefficient r = 0.26 and p-value = 0.42 (two-sided, using slope = 0 for null hypothesis and Wald Test with t-distribution as test statistic).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | O-eLIT comparison of X chromosomes in female IMR-90 cells after ChrX-46plex-2K O-eLIT off of both streets. a, First round of O-eLIT sequencing. MacroH2A.1 immunostaining after five rounds of O-eLIT marks the Xi. n = 1. **b**, **c**, Xi and Xa traces (b) and ball and stick (c) of panel a nucleus after Tier 2 analysis and interpolation of missing targets. Sphere color corresponds to chromosome cartoon. n = 1. **d**, Single-cell pairwise spatial distances after interpolation of missing targets in panel **a**. **e**, Tier 2 target detection efficiency after five rounds of O-eLIT. 38.57% of targeted regions are detected in 71 cells. Detection efficiency from individual replicates are plotted. Error bars represent 95% bootstrap confidence interval of the mean. **f**, Population pairwise spatial distances after Tier 1 detection (n = 71 cells) and Hi-C data of IMR-90 cells (Rao et al. 2014). **g**, Population contact maps (top) where two targets are considered to be in contact if less than 2 µm apart (n = 315 chromosomes). Bottom, Hi-C data as in panel f. (Spearman's rank correlation with the Hi-C matrix is r = 0.733, two-sided p-value for a hypothesis test whose null hypothesis is that two sets of data are uncorrelated = 2.564 ×10e-175). **h**, Radius of gyration for the Xi (n = 40 chromosomes) and Xa (n = 31 chromosomes). The thick line in each violin plot represents the Interquartile range (IQR), the white dot marks the median and the thin lines extend 1.5 times the IQR. P-value = 7.08 ×10e-6 (two-sided t-test whose null hypothesis is equal means). **i**, **j**, Linear plot of the mean spatial distance versus the genomic distance for all pairwise targets for Xi (n = 40 chromosomes) and Xa (n = 31 chromosomes). **k–l**, Population contact maps for Xi (n = 40 chromosomes) and Xa (n = 31 chromosomes) with eigenvector analysis used to identify different domains. X1-X18 (white) and X19-X46 (grey) targets p and q arms, respectively.



Extended Data Fig. 10 | OligoFISSEQ applications. a, O-eLIT and immunofluorescence (IF). 36plex-1K was sequenced 5 rounds with O-eLIT off Mainstreet. Then, the same sample was prepared for IF and stained with antibodies. Samples were counterstained with wheat germ agglutinin (WGA) to stain membranes. Images are from deconvolved, maximum z-projections representative of 2 replicates. **b**, Chromosomal regions imaged with OligoSTORM from Fig. 6d enlarged and displayed separately. Orientation may differ from Fig. 6d. n = 1. **c**, 8 rounds of O-LIT sequencing of Chr19-9K off of Mainstreet. Images are maximum z-projections. Signal is detectable in all rounds even though the imaging was conducted without the advantage of eLIT, suggesting that 8 rounds of O-eLIT will produce even stronger signals. Images are representative of 2 replicates. **d**, O-LIT is compatible with gel embedding and target amplification via rolling circle amplification (RCA). Chr19-9K was hybridized to PGP1f cells, after which the sample was embedded in a hydrogel and then cleared of cellular background with proteinase. Next, a molecular inversion probe (MIP) was hybridized to a Chr19-9K specific barcode on Backstreet as well as a fluorophore labeled (purple) secondary oligo to Mainstreet to visualize Chr19-9K Oligopaint oligos. MIPs were circularized via ligation and RCA, after which the first digit of the barcode was sequenced using O-LIT (green). Images are representative of two replicates. **e**, Comparison of secondary fluorophore signal (2°) versus first round sequencing signal (LIT) from puncta in panel b images. Center values are mean values (3.4 for 2o and 4.9 for O-LIT) with SD.

nature research

Corresponding author(s): Dr. Chao-ting Wu (twu@genetics.med.harvard.edu) and Dr. Marc A. Marti-Renom (martirenom@cnag.crg.eu)

Last updated by author(s): 05/23/2020

Reporting Summary

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

Statistics

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

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

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

Software and code

Policy information about availability of computer code

Data collection	Nikon Elements (NIS ElementsAR ver. 5.02.01.) and Vutara SRX software were used to acquire images.
Data analysis	ImageJ/Fiji (version 2.0.0-rc-69/1.52p) were used to align, normalize, contrast, overlay, and measure images as described in the Methods section. Python (version 2.7) with custom scripts was used for image analysis. Constrained K-means algorithm (version 1.5) https://zenodo.org/record/831850 was used for clustering. Integrative Modelling Platform () was used for Tier 2 (https://www.ncbi.nlm.nih.gov/pubmed/22272186). Domino (version) was used for Tier 2 tracing. Seaborn package for Python was used to generate plots. R (version 3.6.1) and R-Studio (version 1.2.1335) was used for initial image analysis. GraphPad Prizm (version 8.2) was used to generate plots. Microsoft Excel (version 16.16.7) was used to generate tables. Adobe Illustrator (version 22.0.1) was used to assemble figures. ChromoMap package for R by Lakshay Anand was used to generate chromosome cartoons (https://doi.org/10.1101/605600). Nikon Elements (NIS ElementsAR ver. 5.02.01.) was used to process images and deconvolution. Chimera was used for ball and stick visualizations (https://doi.org/10.1002/jcc.20084). Python scripts will be available on GitHub (https://github.com/3DGenomes/OligoFISSEQ).

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

Data

Policy information about availability of data

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

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

The data that support the findings of this study are available from the corresponding author upon reasonable request. All raw and processed data will be made available upon request.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No sample size calculation was performed as we aimed to obtain images of as many cells as possible given experimental constraints during technology optimization. For all experiments with analysis, a minimum of 3 technological replicates were performed to confirm reproducibility. We deemed this to be sufficient due to low observed variability between samples. Datasets were imaged to assess and compare the efficiency of the different versions of OligoFISSEQ. The samples were aggregated to study the structural variability of the cell population. We verified that the sample sizes were sufficient to capture such variability by comparing OligoFISSEQ distance matrices with interaction frequency matrices obtained with Hi-C experiments as an orthogonal method (see "Distance heat-maps and Hi-C maps" of the Material and Methods).
Data exclusions	Cells that did not pass initial quality control filtering were not included in downstream processing and analysis. Cells in mitotic process were discarded from the analysis following the procedure described in the section "Detection efficiency and False Positives ratios" of the Material and Methods. A second exclusion is applied for cells which total detection efficiency in Tier 1 is below 25%. Those cells are mainly presenting imaging distortions in one or more channels or are falling in the border of the images. Exclusion criteria was pre-established, as we focused on interphase cells that were entirely imaged.
Replication	All replication attempts were successful and detailed in Fig. S3, S6, S7, and Table S13. Preferential chromosome positioning as identified in our study (Fig. S4, S5), was in line with reported observations in the literature. Additionally, we found that chromosomes segregated into distinct regions (territories) in the nucleus, also in line with observations from the literature (Fluorescent In Situ Hybridization and Hi-C studies).
Randomization	Cells used for imaging were selected randomly and all imaged cells that passed quality filters were used for analysis, therefore, there was no requirement for randomization.
Blinding	Blinding was not performed as experimental conditions were evident from the image data. Analysis and quantifications were performed using computational pipeline applied equally to all conditions and replicates for a given Oligopaint oligo library. Thresholds for detecting puncta were chosen for each Oligopaint oligo library (see Table S14) on graphs with objective properties that appeared indistinguishable across conditions.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

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

Antibodies

Antibodies used	Anti-Alpha Tubulin (Sigma-Aldrich: T9026) used at 1:500. Anti-GAPDH (Abcam: ab9483) used at 1:200 Anti-TOMM20 (Abcam: ab78547) used at 1:500 Anti-macroH2A.1 (Abcam: ab183041) used at 1:250 Donkey Anti Mouse Cy5 (Jackson ImmunoResearch Laboratories: 715-175-150) used at 1:500 from 1.25mg/mL stock Donkey Anti-Rabbit Cy3 (Jackson ImmunoResearch Laboratories: 711-165-152) used at 1:500 from 1.25mg/mL stock Bovine Anti-Goat Alexa Fluor 594 (Jackson ImmunoResearch Laboratories: 805-585-180) used at 1:500 from 1.25mg/mL stock
Validation	Anti-Alpha Tubulin has been validated by Sigma-Aldrich to be specific in human cell lines (osteosarcoma and breast cancer) using western blotting and in HeLa cells by immunofluorescence microscopy (https://www.sigmaaldrich.com/catalog/product/sigma/t9026?lang=en&region=US). Anti-GAPDH has been validated by Abcam to produce positive signal in whole cell lysates from HeLa as well as human brain tissue lysate as well as positive immunofluorescence signal in HeLa cells (https://www.abcam.com/gapdh-antibody-loading-control-ab9483.html). Anti-TOMM20 has been validated by Abcam to produce positive signal in HEPG2 whole cell lysate and positive immunofluorescence signal in HEPG2 cells (https://www.abcam.com/tomm20-antibody-mitochondrial-marker-ab78547.html). Anti-macroH2A.1 has been validated by Abcam to produce positive signal in HAP1 lysates and reduced signal in HAP1 m2A1 knockouts (https://www.abcam.com/mh2a1-antibody-epr93592-ab183041.html).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	PGP1f (Human male fibroblasts, Coriell: GM23248), IMR-90 (Human female fibroblasts, ATCC: CCL-186)
Authentication	None of the cell lines have been authenticated.
Mycoplasma contamination	Cell lines were not tested for mycoplasma contamination but no indication of contamination was observed.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Chapter II

3D reconstruction of genomic regions from sparse interaction data

Chromosome conformation capture (3C) technologies measure the interaction frequency between pairs of chromatin regions within the nucleus in a cell or a population of cells. Some of these 3C technologies retrieve interactions involving non-contiguous sets of loci, resulting in sparse interaction matrices. One of such 3C technologies is Promoter Capture Hi-C (pcHi-C) that is tailored to probe only interactions involving gene promoters. As such, pcHi-C provides sparse interaction matrices that are suitable to characterize short- and long-range enhancer–promoter interactions. We introduced a new method to reconstruct the chromatin structural (3D) organization from sparse 3C-based datasets such as pcHi-C. Our method allows for data normalization, detection of significant interactions and reconstruction of the full 3D organization of the genomic region despite of the data sparseness. Specifically, it builds, with as low as the 2–3% of the data from the matrix, reliable 3D models of similar accuracy of those based on dense interaction matrices. Furthermore, the method is sensitive enough to detect cell-type-specific 3D organizational features such as the formation of different networks of active gene communities.

Published online 22 March 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 1 **1**
doi: 10.1093/nargab/lqab017

3D reconstruction of genomic regions from sparse interaction data

Julen Mendieta-Esteban[1], Marco Di Stefano[1], David Castillo[1], Irene Farabella[1,*] and Marc A. Marti-Renom[1,2,3,4,*]

[1]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain, [2]Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08003 Barcelona, Spain, [3]Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain and [4]ICREA, 08010 Barcelona, Spain

Received December 05, 2020; Revised February 08, 2021; Editorial Decision February 24, 2021; Accepted March 02, 2021

ABSTRACT

Chromosome conformation capture (3C) technologies measure the interaction frequency between pairs of chromatin regions within the nucleus in a cell or a population of cells. Some of these 3C technologies retrieve interactions involving non-contiguous sets of loci, resulting in sparse interaction matrices. One of such 3C technologies is Promoter Capture Hi-C (pcHi-C) that is tailored to probe only interactions involving gene promoters. As such, pcHi-C provides sparse interaction matrices that are suitable to characterize short- and long-range enhancer–promoter interactions. Here, we introduce a new method to reconstruct the chromatin structural (3D) organization from sparse 3C-based datasets such as pcHi-C. Our method allows for data normalization, detection of significant interactions and reconstruction of the full 3D organization of the genomic region despite of the data sparseness. Specifically, it builds, with as low as the 2–3% of the data from the matrix, reliable 3D models of similar accuracy of those based on dense interaction matrices. Furthermore, the method is sensitive enough to detect cell-type-specific 3D organizational features such as the formation of different networks of active gene communities.

INTRODUCTION

Chromatin within the nucleus is organized into higher order structures that emerge at different genomic scales, from chromosome territories (at tens of megabases scale), active and inactive chromatin domains (at few megabases scale) (1), self-interacting domains or TADs (at hundreds of kilobases scale) (2,3,4) and long-range chromatin loops between regulatory elements (at tens of kilobases scale). This multi-scale organization has a direct impact on many biological processes, such as gene regulation, DNA replication and cell differentiation (5,6,7). Indeed, genome structure typically reflects cell-type-specific differences in the transcription pattern, and it is frequently rewired upon cell state changes and disease onset (8). Thus, investigating the principles shaping chromosome three-dimensional (3D) structure is pivotal to shed light into the relationship between genome structure and function.

Several experimental techniques are available to examine chromatin organization (9). Amongst them, molecular biology methods, such as chromosome conformation capture (3C) and its derivatives are widely used (10). These experiments retrieve information about the frequency of interaction between loci in single (11,12,13) or in populations of thousands to millions of cells and have been designed to analyse the chromatin landscape at different genomic scales (1,14,15,16). For example, some cell population-based experiments allow the retrieval of unspecified interactions in the whole genome (e.g. Hi-C (1), Micro-C (14), GAM (15) and SPRITE (16)). Complementarily, other 3C-based experiments are tailored to capture interactions centred on a specific locus with the rest of the genome (e.g. 4C (17) and multi-contact 4C (MC-4C) (18)) or on sets of dispersed loci in the genome, such as loci enriched for a specific protein (HiChIP) (19) or loci harbouring gene promoters (pcHi-C) (20). Each class of 3C-based experiments provide different but complementary insights on particular aspects of the genome organization, and their analysis is dependent on the experimental genomic resolution and on the inherent technical biases of each experimental procedures.

A variety of physics- and data-driven approaches for genome 3D reconstruction have been developed to expose the principles shaping chromosome 3D structure (21,22,23,24). For instance, data-driven (restraint-based) modelling approaches as PGS (25,26), TADbit (27), 4Cin (28) and TADdyn (29) have been implemented to re-

*To whom correspondence should be addressed. Tel: +34 934 020 542; Fax: +34 934 037 279; Email: martirenom@cnag.crg.eu
Correspondence may also be addressed to Irene Farabella. Tel: +34 934 031 945; Email: irene.farabella@cnag.crg.eu

construct ensembles of chromatin 3D models from cell population-based datasets. Others are focused on the 3D modelling of chromatin based on single-cell Hi-C data, like manifold based optimization (30) and NucDynamics (31). However, the majority of the data-driven methods are based on interaction experiments that have been designed to retrieve dense contact information from a continuous set of loci or the whole genome, whilst other interaction experiments are characterized by data sparseness (e.g. HiChIP or pcHi-C). As such, data-driven methods for sparse data modelling are needed.

Generally, the interaction profiles of sparse 3C-based datasets have specific properties that set them apart from other 3C-like techniques characterized by a dense interaction profile. Indeed, protein or promoter capture-based interaction profiles are heavily biased on interactions between captured fragments and devoid of interactions between non-captured fragments. This fact poses the question of whether this lack of information prevents the 3D reconstruction of the whole loci of interest and its analysis, or whether it is sufficient to allow for accurate 3D modelling. To answer this question, we have implemented a new method, which is tailored to integrative modelling and analysis of sparse 3C-based datasets. We have also validated the procedure comparing the resulting reconstructed models with available dense experimental datasets, unveiling that the 3D chromatin organization can be well recovered by interrogating only a small percentage of loci. Additionally, we have designed new tools to facilitate a robust differential analysis of the resulting models and showcased their usability in comparative analyses using the β-globin locus as a test case. Interestingly, comparing different cell-types, we unveiled that the β-globin locus in cord-blood Erythroblasts (cb-Ery), where its foetal and adult β-globin genes are highly expressed, is hierarchically organized in a 3D network of active gene communities that follows an expression gradient.

MATERIALS AND METHODS

Experimental datasets

Structural data were obtained from publicly available 3C-based chromatin interaction experiments of GM12878 cells (Hi-C GEO: GSE63525 and pcHi-C ArrayExpress: E-MTAB-2323) (6,32), and cord-blood derived Erythroblasts (cb-Ery), naive CD4+ T-cells (nCD4), and Monocytes (Mon) (pcHi-C EGA: EGAS00001001911) (33).

Hi-C datasets processing. The reads for each replicate were mapped onto the GRCh38 reference genome, filtered and merged using TADbit with default parameters (27). Then, starting from the merged filtered fragments, the genome-wide raw interaction maps were binned at 5 kilo-base (kb) and normalized using OneD (34) as implemented in TADbit (27).

pcHi-C datasets processing. For each experiment, the reads were mapped onto the GRCh38 reference genome using TADbit (27) and were filtered applying the following filters: (i) self-circles, (ii) dangling-ends, (iii) errors, (iv) extra dangling-ends, (v) duplicated reads and (vi) random breaks. Next, we computed the reproducibility score to measure the similarity between replicates from each pcHi-C dataset (35). Then, for each cell-type, the different replicates from the same experiment were merged into one dataset for further analysis, making an exception with replicate ERR436029 from the GM12878 pcHi-C dataset (E-MTAB-2323), which was discarded due to a clearly low reproducibility score when compared with the rest of the replicates (average of 0.24 with the other replicates as compared to the average of 0.84 obtained between the other replicates). Using the merged filtered fragments, the genome-wide raw interaction maps of each cell-type were binned at 5 kb and normalized using the PRoportion of INTeraction approach (PRINT, next section).

Sparse data normalization PRoportion of INTeraction approach (PRINT). PRINT, a multi-stage normalization procedure, weighs each pair of interacting bins with the same philosophy as the visibility approach for Hi-C (36). Starting from a raw interaction matrix as input, PRINT first transforms the raw interaction between two bins (i and j) into a percentage of interaction with respect to the rest of the genome as:

$$\text{value}_{ij} = \frac{\text{bin}_{\text{ij}}}{\sum \text{row}_i + \sum \text{row}_j - \text{bin}_{ij}}$$

where (bin_{ij}) represent the number of times in which bin i and j interact, and $\sum \text{row}_i$ and $\sum \text{row}_j$ are the sum of all the interactions of bins i and j, respectively, with all the genome (self-interactions included). Then, the non-baited interactions (that is, those bins containing only pcHi-C off-target reads) are filtered out.

PRINT assessment. Using the *benchmarking datasets* described above, each stage of PRINT normalization (raw pcHi-C (pcHi-C-raw), pre-normalized pcHi-C (pcHi-C-pre) and normalized pcHi-C (pcHi-C-norm)) was assessed in comparison with the dense Hi-C interaction matrix by calculating the Spearman's rank correlation coefficient between interactions (bin_{ij}) present in both interaction matrices.

Reconstructed 3D genomic regions

Benchmarking datasets. We selected 12 genomic regions of interest (Supplementary Table S1) as defined by Rao *et al.* (6). This set of genomic regions were predicted to result in reliable 3D models based on their >0.7 MMP scores (37) (Supplementary Table S2). Briefly, MMP score takes into account the interaction matrix size, the contribution of significant eigenvectors in the matrix and the skewness and kurtosis of the z-scores distribution of the matrix to assess their potential for being modelled (37).

Comparative analysis datasets. We selected a genomic region around a locus of interest (here the β-globin) defining it in a semi-automatic manner in each cell-type. Briefly, a viewpoint, which may be constituted by a bin or a set of bins of interest, is selected. Here, as viewpoint we used bins

enclosing the active haemoglobin genes in cb-Ery (HBB, HBD, HBG1 and HBG2). Then, all the other bins that interacted with the viewpoint bins in the normalized genome-wide interaction matrix were selected. Each of these bins were then scored by their cumulative normalized interaction frequency values with the viewpoint bins. From this set only the top intra-chromosomal 200 bins were selected since, by visual inspection, they were the bins spanning the genomic region that best enclosed the viewpoint. Then an unweighted interaction network was generated with the nodes corresponding to the top 200 bins and the viewpoint bins. Edges between nodes were added if their pairwise cumulative normalized interaction frequency value was in the top 200 interacting bins. Then, a series of transformations were applied to the unweighted interaction network: (i) nodes that are highly proximal in 1D genomic resolution (closer than 25 kb) were merged into one node; and (ii) poorly connected nodes in the network that had <5 edges were filtered out (average number of edges per node in Mon, nCD4 and cb-Ery were 200, 214 and 214, respectively). The extreme nodes in terms of genomic coordinates were selected from the final unweighted interaction network to represent the optimal genomic region around the viewpoint. Here, to perform comparative analysis, we defined the optimal genomic region around the viewpoint as the broader genomic region that enclosed all of the genomic coordinates identified in each cell-type.

3D chromosome ensemble reconstruction from sparse datasets

Model representation. Each genomic region was described with a beads-on-string model based-on the previously implemented protocols (29,38) without bending rigidity potential. Thus, a chromosome was represented with N spherical beads with diameter $\sigma = 50$ nm that contain 5 kb of chromatin which determined the genomic unit length of each model.

System set up for molecular dynamics. All simulations were done using TADdyn (29). A generic random self-avoiding walk algorithm was used to define the initial conformation of each model. The potential energy of each system comprised the terms of the Kremer–Grest polymer model (39) including chain-connectivity (Finitely Extensible Nonlinear Elastic, FENE) (40) and excluded volume (purely repulsive Lennard-Jones) interactions. The initial conformation was placed randomly inside a cubic simulation box of size 1000 σ centred at the origin of the Cartesian axis O = (0.0, 0.0, 0.0), tethered at the centre of the box using a harmonic ($K_t = 50.0\ k_BT/\sigma^2$ and $d_{eq} = 0.0\ \sigma$) to avoid any border effect and energy minimized using a short run of the Polak–Ribiere version of the conjugate gradient algorithm (41) to favour smooth adaptations of the implementations of the excluded volume and chain connectivity interaction.

Encoding sparse data into TADdyn restraints. TADdyn (29) empirically identifies the three optimal parameters to be used for modelling based on a grid search approach. These are: (i) maximal distance between two non-interacting particles (*maxdist*); (ii) a lower-bound cut-off to define particles that do not frequently interact (*lowfreq*); and (3) an upper-bound cut-off to define particles that frequently interact (*upfreq*). All possible combinations of the parameters were explored in the intervals *lowfreq* = (−1.0,−0.5, 0, 0.5), *upfreq* = (−1, −0.5, 0, 0.5), *maxdist* = (200, 300, 400, 500) nm and assessing each combination using distance thresholds to determine if two particles are in contact (*dcutoff*) at 100,150, 200, 250, 300, 350, 450, 500 nm. For each of the combinations an ensemble of 100 3D models was generated and the Spearman correlation coefficient between the contact map derived from each ensemble and the experimental input interaction matrix was calculated. The top set of parameters for each region in each cell-type were set for those resulting in the highest Spearman correlation coefficient between the models contact map and the input interaction matrix. To allow for a robust comparative analysis ('Materials and Methods' section) the optimal *maxdist* and the *dcutoff* parameters were selected based on the consensus within the top optimal values for each region in each cell-type. Optimal *maxdist* and the *dcutoff* were set at 300 and 200 nm, respectively for the ensembles of models reconstructed from the GM12878, cb-Ery, nCD4 and Mon pcHi-C datasets. Once the three optimal parameters were defined, the type of restraints between each pair of particles was set considering an inverse relationship between the frequencies of interactions of the contact map and the corresponding spatial distances. Non-consecutive particles with contact frequencies above the upper-bound cut-off were restrained by a harmonic oscillator at an equilibrium distance, whilst those below the lower-bound cut-off were maintained further apart than an equilibrium distance by a lower-bound harmonic oscillator. To identify 3D models that best satisfy all the imposed restraints, the optimization procedure was then performed using a steered molecular dynamic protocol. A total of 1000 replicate trajectories were generated for each genomic region and dataset. Each of the 1000 replicate trajectories, the conformation at the end of the steering protocol (when the target spring constant and equilibrium distance are reached) was retained to form the final ensemble of 1000 3D models. For the cb-Ery, nCD4 and Mon datasets, to account for possible mirrored 3D models within the final ensemble of 3D models, each ensemble was then clustered based on structural similarity score as implemented in TADbit (27) and only the models from the most populated cluster were retained for further analysis.

Steered molecular dynamics protocol. A steered molecular dynamics protocol was used to progressively favour the imposition of the defined set of restraints between non-consecutive particles. For each restraint, the equilibrium distance was set to 1 particle diameter (σ). The spring constant $k(L,t)$ was weighted with the sequence-separation L between the constrained beads as in TADdyn (29) to ensure that the steering process was not dominated by the target pairs at the largest sequence separation. However, here the $k(L,t)$ was smoothly ramped during the steering phase from zero to its maximum value.

3D chromosome ensemble reconstruction from dense datasets

The reconstruction of 3D models of genomic regions from dense data followed the modelling protocol described above. That is, a grid search approach was used to select for the optimal parameters to be used for modelling. The optimal *maxdist* and the *dcutoff* parameters were selected based on the consensus within the top optimal values for each region in the GM12878 pcHi-C dataset and set at 300 and 200 nm, respectively. Using these parameters, the final ensemble of 1000 3D models was obtained starting from the computed 1000 steered molecular dynamics trajectories.

3D chromosome ensemble reconstruction from Virtual pcHi-C derived from dense datasets

A dataset of Virtual pcHi-C interaction matrices was produced starting from the normalized Hi-C interaction matrices at 5 kb resolution (GM12878 cells GEO: GSE63525; 'Materials and Methods' section) and from the liftover (https://genome.ucsc.edu/cgi-bin/hgLiftOver) list of captured fragments in pcHi-C GM12878 experiment (32). The obtained Virtual pcHi-C interaction matrices comprised only interactions (bin_{ij}) in which either i or j enclose the coordinates of a captured fragment. These interaction matrices were used as input for the reconstruction of 3D models of genomic regions following the modelling protocol described above. The optimal *maxdist* and the *dcutoff* parameters were set at 300 and 200 based on their consensus with the parameters used in the GM12878 pcHi-C dataset. A total of 1,000 steered molecular dynamics trajectories were computed, and for each trajectory the conformations satisfying the majority of the imposed constraints within a radius of 2 σ were retained.

3D chromosome ensemble reconstruction from 'synthetic' sparse dataset

We used a previously published 'toy genome' (37) (that is, the ensemble of models accounting for the formation of TAD-like architecture with low structural variability and high noise levels that comprises a total of 626 particles at the highest genomic resolution) to randomly select 10 sets of 22 loci from the toy genome contact map (or synthetic interaction maps). These loci mimic pcHi-C to generate reliable sparse interaction matrices comprising only interactions (bin_{ij}) in which either i or j have been selected as random captured loci. Each of these sets was then randomly subsampled to generate 'synthetic' capture matrices with 2, 4, 6, 10, 14 and 18 selected captured loci. The obtained 'synthetic' capture matrices (70 in total) were next used as input for the reconstruction of 3D models of genomic regions following the modelling protocol described above. The optimal *maxdist* and the *dcutoff* parameters were set at 500 and 200 nm. Using these parameters, a final ensemble of 100 3D models was reconstructed for each 'synthetic' capture matrices comprising the conformations that best satisfied the imposed restraints in each of the computed 100 steered molecular dynamics trajectories.

Analysis of the ensemble of 3D models

Contact map generation. For each ensemble of 3D models, a contact map was calculated at 5 kb resolution to visualize the frequencies of contacts in the ensemble. Two beads were considered to constitute a contact when their euclidean distance was below 200 nm cut-off.

Matrix comparison. The degree of similarity between two matrices was computed by comparing each cell from the matrices, or a subset of them, using the Spearman's rank correlation coefficient (r_s) as implemented in the Python library SciPy (42,43):

$$r_s = 1 - \frac{6\sum_{i=1}^{n}\left(r_{\mathrm{bin}_{x_i}} - r_{\mathrm{bin}_{y_i}}\right)^2}{n\left(n^2-1\right)}$$

where $r_{\mathrm{bin}_{x_i}}$ is the rank of the *ith* observation in one matrix, $r_{\mathrm{bin}_{y_i}}$ is the rank of the *ith* observation in the other matrix and n states for the number of pairs of observations.

Particle-to-particle median distance correlation (ppMdC). For each ensemble of 3D models, we differentiated three sets comprising particles enclosing the coordinates of: (i) captured loci (capture), (ii) non-captured loci (other) and (iii) all the loci (all). For each of the pairs of particles in a given set we calculated the particle-to-particle median distance. Then, the degree of similarity between two given sets was computed using the Spearman's rank correlation coefficient between their particle-to-particle median distances. The ppMdC measure varies between −1.0 and 1.0 for comparisons where the particle-to-particle median distances perfectly anti-correlate or correlate, respectively.

Hierarchical clustering of ensembles of 3D models. Multiple ensembles of 3D models were merged in a unique set and the models were structurally superpose using pairwise rigid-body superposition. Next, the all-vs-all distance root-mean-square deviation (dRMSD) was calculated and the resulting dRMSD matrix was hierarchically clustered using Ward's sum of squares method (44) as implemented in the Python library SciPy (42).

Cell-specific expression profile. Publicly available (33) expression matrix containing the expression values (log(FPKM)) of each gene in cb-Ery, nCD4 and Mon cell-types was downloaded (GeneExpressionMatrix.txt.gz at https://osf.io/u8tzp/). The three datasets had two or more replicates each (two cb-Ery, five Mac and eight nCD4, respectively), thus the average expression value of each gene from all replicates was used. Then, a cell-specific per-bin cumulative expression profile of the chr11:3 795 000–8 505 000 genomic region at 5 kb resolution was obtained assigning the mean expression value of each gene (with log(FPKM) > 0) to bins enclosing for the coordinates of its transcription start site (coordinates retrieved from bioMart (45)).

3D enrichment analysis. To study the spatial co-localization of different regulatory elements and the local levels of transcription (based on genome-wide ChIP-

and RNA-seq data) around a selected locus (central viewpoint) we implement a *3D enrichment analysis tool* (named 'radial-plot') that allows the comparison of heterogeneous sets of data from multiple data sources. Per each cell-type a per-particle binarized chromatin marks profile in the genomic region was generated starting from the ChIP-seq signal of H3K27ac, H3K36me3, H3K4me1, H3K4me3, H3K9me3 and H3K27me3 in cb-Ery, nCD4 and Mon cell-types (33). A particle was considered enclosing for a chromatin mark if a peak was present. Similarly, we also constructed, for each cell-type, a per-particle binarized transcription profile starting from the cell-specific expression profile ('Materials and Methods' section). Then the 3D spatial distribution of the 3D enrichment based on the per-particle binarized profile around the chosen central viewpoint was calculated as follow: (i) starting from the central viewpoint an initial sphere with a radius of 200 nm was constructed; (ii) a series of spherical shells, that occupied a volume equal the initial sphere, were added; (iii) each model in the ensemble of 3D models a particle of the binarized profile was assigned to a spherical shell based on its relative distance to the central viewpoint; (iv) each spherical shell we performed Fisher's exact tests for 2×2 contingency tables comparing the amount of particles with or without signal in the spherical shell with the outside ones, and the log of the odd ratios was assigned to the shell if the P-value < 0.01. The obtained 3D enrichment was then visualized as a 2D radial plot.

Defining gene communities: co-occurrence of expressed genes. For each ensemble of 3D models, based on their cell-specific expression profile ('Materials and Methods' section), we defined the set of expressed particles ($\log(\text{FPKM}) > 0$). Then, considering this set of particles, an all-versus-all pairwise distances matrix was calculated in each model and hierarchically clustered using Ward's sum of squares method (44) as implemented in the Python library SciPy (42). Then the Calinski–Harabasz index (46), as implemented in the Python library Scikit-learn (47), was used to determinate the optimal number of clusters in each dendrogram. Then, for each ensemble, a co-occurrence matrix was generated considering the percentage of models in which a pair of particles belonged to the same cluster. The co-occurrence measure varies between 0 and 100, where 0 indicates absence of co-occurrence and 100 indicates a stable co-occurrence within the ensemble of 3D models. The co-occurrence matrix was next hierarchically clustered using Ward's sum of squares method (44) and communities of co-occurrent active genes were identified using the Calinski –Harabasz index analysis in the dendrogram.

Communities stability within the ensemble of models. To assess the stability of each community within the ensemble we introduced the inter-community co-occurrence score that defines the degree of unstable compositions of a community. It is computed as the mean co-occurrence values between each gene in a community and the rest of the communities.

Distance between communities and within community. To describe the spatial arrangement of each community for a given ensemble of 3D models, we treated each community as a rigid body and calculated its centre of mass (COM) in each 3D model of the ensemble. Per each model the all-versus-all pairwise distances between the COMs of each communities were computed and the mean distance values assigned as the typical distance between communities. Similarly, per each model, we also calculated the distance of each particle in a given community and the COM of its community. The within community distance of a given particle was defined by its mean value in the ensemble of 3D models.

RESULTS

Overall modelling strategy for sparse 3C data

Sparse 3C datasets provide information of interactions that involve a limited number of specific loci in the genome. pcHi-C, for example, provides a promoter-centred view of chromatin interactions, helping to assign distal regulatory regions to their target genes, thus providing insights on how gene expression might be controlled (32,33,48) and how disease-associated genomic variation could affect gene regulation (49). The main limitation of these sparse technologies, however, is the scarcity of specialized tools for their analysis. Here, we have developed an integrative 3D modelling method capable of dealing with data sparsity, enabling the analysis and interpretation of pcHi-C data, and tested it on 12 distinct loci (Benchmarking datasets; 'Materials and Methods' section and Supplementary Table S1). Our method follows an integrative modelling procedure comprising five steps (50): (i) gather experimental data and process them to obtain the input interaction matrix for the modelling approach, (ii) represent the selected chromatin regions using a bead-spring polymer model with a particle size proportional to the genomic resolution of the experimental data, (iii) transform the frequency of interactions into spatial retrains, (iv) sample the conformational space by steered molecular dynamics and (v) analyse and validate the obtained ensemble of 3D models ('Materials and Methods' section and Figure 1A).

In this work, we gathered pcHi-C interaction data ('Materials and Methods' section), whose processing step is pivotal to minimize the experimental biases from the capture protocol. To this end, we designed a multi-stage normalization procedure named PRINT ('Materials and Methods' section). PRINT weighs each interaction by dividing it by the cumulative whole-genome interaction frequencies of both of the interacting bins, regularizing the interaction patterns for the fact that captured loci are highly enriched in contacts. It also removes the pcHi-C unspecific interactions between non-probed bins. To test quantitatively the performance of our normalization procedure, we compared each of the normalization stages of the pcHi-C matrices with the respective Hi-C matrices normalized with OneD in each of the selected loci (34). The median correlation between bins with interaction data in both matrices was 0.27 ($\pm$0.025 Median Absolute Deviation (MAD)) for raw pcHi-C matrices (pcHi-C-raw), increasing to 0.44 ($\pm$0.032 MAD) with the pcHi-C pre-normalization step (pcHi-C-pre) and reaching 0.60 ($\pm$ 0.056 MAD) for fully normalized pcHi-C matrices (pcHi-C-norm) (Supplementary Figure S1A), suggesting that PRINT reduced successfully the target biases. Then,

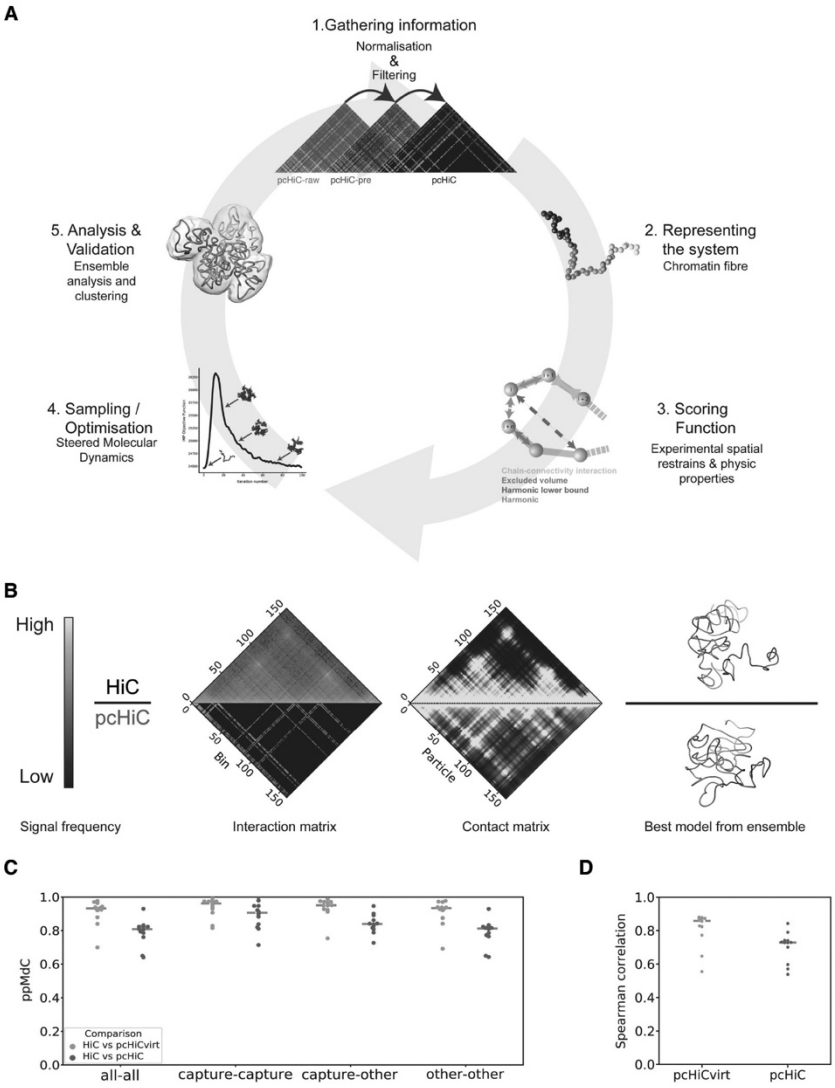


Figure 1. Integrative modelling for sparse datasets efficiently reconstructs the 3D organization of genomic loci. (**A**) Workflow of the integrative modelling approach followed to build ensembles of chromatin 3D models from pcHi-C: (i) gathering the input interaction matrices with subsequent normalization and filtering; (ii) representation of the chromatin fibre as a polymer with the particle size proportional to the resolution of the experiment; (iii) definition of the scoring function used in the modelling procedure. Here, the scoring function comprises spatial restraints derived directly from the input interaction data and from properties of the chromatin fibre ('Materials and Methods' section); (iv) sampling the conformational space by steered molecular dynamics ('Materials and Methods' section); and (v) validation of the obtained ensemble of models and further analysis. Model images in all panels were created with Chimera (74). (**B**) Representation of the input and output data from region 2 (Supplementary Table S1). The upper half of the panel refer to the dense dataset (Hi-C), whereas the lower half refer to the sparse-datasets (pcHi-C). From left to right, the matrices of normalized interaction frequency ('Materials and Methods' section) between each pair of bins, the contact matrix obtained from the ensemble of models of region 2 displays the percentage of models in which two bins are found bellow the defined distance cut-off for the contact ('Materials and Methods' section), and the best model from the ensemble as assesses by the scoring function. The colour bar shows the colour coding from low (blue) to high (yellow) interaction or contact frequencies signal. (**C**) Comparison between model ensembles derived from sparse (pcHi-Cvirt and pcHi-C in grey and blue, respectively) and dense (Hi-C) datasets assessed by the particle-to-particle median distance correlation (ppMdC; 'Materials and Methods' section). Three subsets of particles have been compared given the enclosed loci: (i) captured loci (capture), (ii) non-captured loci (other) and (iii) all the loci (all). The grey dashed line indicates the median ppMdC in the 12 analysed regions. (**D**) Element-wise Spearman correlation coefficients between the experimental Hi-C interaction matrices and the contact maps derived from the model ensembles reconstructed from sparse data (pcHi-Cvirt and pcHi-C in grey and blue, respectively). The grey dashed line indicates the median element-wise Spearman correlation coefficients in the 12 analysed regions.

we represented the selected loci as a bead-spring polymer model with a particle size set to 5 kb, taking into account the restriction fragment lengths distribution in the benchmarking datasets (Supplementary Figure S1B). Similarly to TADbit (27) and TADdyn (29), to simulate the structural conformation of genomic loci, we then transformed the interaction frequencies associated with each bin pair into spatial restraints ('Materials and Methods' section). The latter were then imposed on the model using steered molecular dynamics as sampling method in which the spring constant associated to each restraint was ramped up as a function of simulation time from zero to the value computed from the interaction data. Lastly, we implemented new means for a robust quantitative spatial differential analysis of genomic loci.

Comparison between sparse and dense 3C-derived models

Dense 3C data have been extensively used to reconstruct the 3D organization of genomic loci (25,27,29,30). Here, to test the reliability of our modelling approach, we used sparse and dense datasets to build ensembles of models of the same loci. Specifically, we applied our integrative method for sparse data modelling to previously published pcHi-C datasets of GM12878 cells (32) to reconstruct 3D model ensembles of 12 distinct loci (Figure 1B and Supplementary Table S1) at a 5 kb resolution and compared them with the corresponding ones reconstructed using Hi-C (6) at the same genomic resolution. Additionally, to quantify the effect of sparsity in the comparison independently of the experimental protocol biases, we generated virtual pcHi-C (pcHi-Cvirt) interaction matrices from the normalized Hi-C datasets extracting the rows and columns probed in the pcHi-C experiment ('Materials and Methods' section). These virtual sparse matrices were then used to reconstruct 3D model ensembles of the selected loci.

The comparison between the sparse and dense derived 3D model ensembles revealed that it is possible to recover most of the 3D organization of the dense dataset in spite of the data sparsity (Figure 1C). Indeed, the all-versus-all particle-to-particle median distance correlation (ppMdC) between the sparse and dense derived 3D model ensembles was 0.81 (±0.019 MAD) and 0.93 (±0.024 MAD) for both pcHi-C and pcHi-Cvirt. Additionally, when comparing distances between particles that have both been captured in the pcHi-C experiment (capture–capture), the ppMdC was higher, reaching 0.91 (±0.054 MAD) for pcHi-C and 0.96 (±0.019 MAD) for pcHi-Cvirt. Consistently, when comparing distances between non-captured particles with captured particles (capture-other) or between non-captured particles (other–other), the ppMdC indicated good agreement with values of 0.84 (±0.03 MAD) and 0.95 (±0.02 MAD), and 0.81 (±0.02 MAD) and 0.93 (±0.02 MAD), respectively, for pcHi-C and pcHi-Cvirt in both comparisons (Figure 1C). The results indicate that the sparse derived ensembles of 3D models are a good representation of the dense experiment and that the intrinsic experimental biases of the capture experiment only minorly affect the 3D reconstruction. Indeed, comparing the whole contact map computed from the 3D model ensembles derived from sparse data directly with the whole experimental Hi-C interaction matrices revealed that the reconstructed ensembles of models are in good agreement with the dense experimental data having an element-wise Spearman's rank correlation coefficient of 0.73 (±0.02 MAD) and 0.86 (±0.02 MAD), for pcHi-C and pcHi-Cvirt derived ensembles of models, respectively (Figure 1D). Overall, this suggest that the ensembles of models reconstructed by our approach represent well the 3D organization of the selected genomic regions and, more importantly, recover the spatial arrangements of loci that are not interrogated by the sparse experiment.

Reconstruction efficiency and data sparsity

To investigate the relationship between the reconstruction efficiency and data sparsity, we simulated 'synthetic' capture data. Briefly, we generated 10 different sets of 'synthetic' capture matrices that represent generic capture-like experiments. We started from the contact matrix derived from a 3D toy-genome models ensemble that simulates roughly a one Mb length genome (comprising more than 600 particles) with a TAD-like architecture, a high level of interaction noise and low variability between models (37) ('Materials and Methods' section and Figure 2A). To build each of the 10 'synthetic' sets, we randomly selected 22 captured loci and constructed 6 additional datasets of different sparsity downsampling each set considering 2, 4, 6, 10, 14 and 18 loci at a time, which mimics the distribution of captured probes per Mb present in a typical genome-wide pcHi-C experiment (Figure 2B). The constructed 70 capture-like matrices thus aim to represent typical pcHi-C experimental design. Using our integrative modelling method for sparse datasets, we reconstructed, from each of the 'synthetic' capture matrices in the dataset and their downsampled counterparts, ensembles of 100 models and compared them with the reference toy-genome ensemble (Figure 2A). Independently of the sets, the ppMdC between the sparse and dense model ensembles increased with the number of captured particles used in the modelling procedure reaching a median correlation between sets of 0.82 (+/- 0.02 MAD with just 10 captures per Mb (Figure 2C). Notably, also with 4 and 6 captures per Mb the ppMdC reached 0.69 (±0.04 MAD) and 0.79 (±0.05 MAD) for four and six captures, respectively, although with greater variation within sets. This suggests that with 10 captured loci per Mb the uncertainty in the input information is smaller, leading to more precisely reconstructed models. Nevertheless, it is possible to reconstruct good models also with fewer as four captured loci per Mb although with a higher degree of variability. To quantify the effect of data sparseness on model reconstruction, we next measured the amount of input information used during the modelling as the percentage of all possible interaction pairs in the contact matrix (dense data input) and then assessed it with the ppMdC. The results indicate that it was possible for the majority of the sets (8/10) to reliably reconstruct the reference toy genome (ppMdC > 0.8) with just 2–3% of all the interaction pairs in the contact matrix used as restraints (Figure 2D and Supplementary Figure S2). Taken together, this analysis shows that it is possible to consistently recover most of the 3D organization of a region of interest with 10 captured loci per Mb and with just 2–3% of all possible interactions within a region captured.

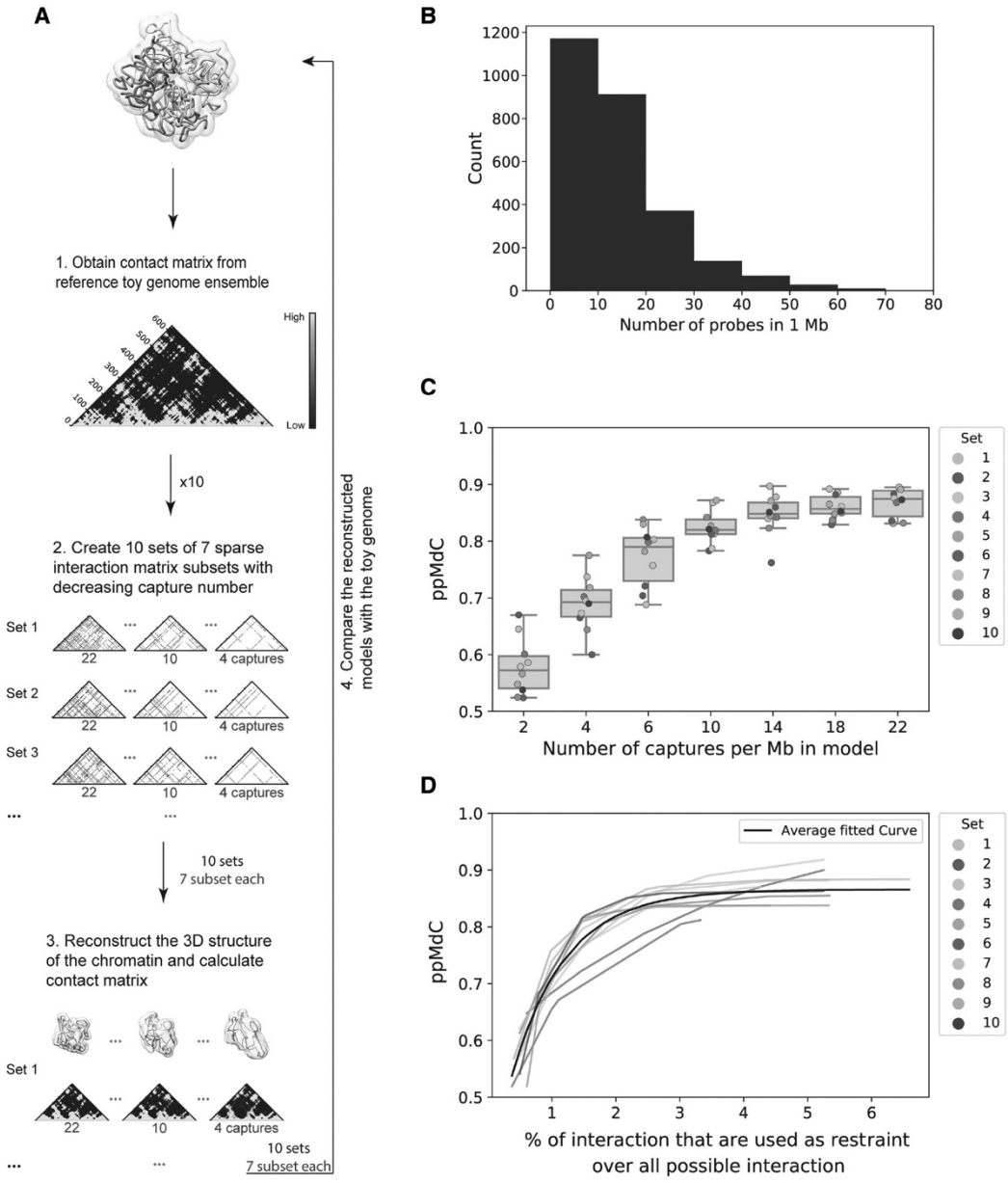


Figure 2. A low percentage of the interaction data is needed to produce reliable 3D reconstructions. (**A**) Workflow for the generation of 3D model ensembles from 'synthetic' sparse datasets and comparison with the toy genome. A total of 70 'synthetic' captured map were generated representing 10 different capture experiments with different level of data sparsity ('Materials and Methods' section). Model images were created with Chimera (74). (**B**) Distribution of pcHi-C probes per megabase windows in the genome (32). (**C**) Distribution of the ppMdC between the 'synthetic' models and the toy genome grouped by subsets of captures per megabase. Box boundaries represent first and third quartiles, middle line represents median and whiskers extend to 1.5 times the interquartile range. The 10 sets of captured positions are displayed with the colour code shown in the insert. (**D**) Relationship between the ppMdC and the percentage of cells in the matrix used as restraints in each set represented with an exponential fit. The used colour code is the same as in (C), the grey line represents the mean fit of all the datasets in analysis.

Cell-type-specific organization of the β-globin locus

To illustrate the utility of our integrative approach in unveiling the differential organization of loci, we applied it to the genomic region surrounding the β-globin locus in three different cell-types (cb-Ery, nCD4 and Mon; 'Materials and Methods' section) for which pcHi-C data are available (33). The selected genomic region contains five coding genes (HBB, HBD, HBG1, HBG2 and HBE1) with developmental-stage-dependent expression (51), which is finely regulated by a set of upstream enhancers known as the locus control region (LCR) (52). This locus is known to be in an active conformation in cb-Ery, where the LCR is interacting mainly with expressed genes as HBB and HBD, but not in nCD4 and Mon cells (33).

First, we defined the optimal region to be modelled based on the interaction networks (in all cell-types) of the embryonic (HBG1 and HBG2) and adult (HBB and HBD) globin genes with the rest of the genome at 5 kb resolution ('Materials and Methods' section). The defined region spanned 4.7 Mb of chr11 (chr11:3 795 000–8 505 000 base-pairs (bp)) comprising several neighbouring genes and multiple long-range regulatory elements. By applying our integrative approach, we generated an ensemble of 1000 3D models for each cell-type. The packing of the genomic region was significantly different in each cell-types with median radius of gyration of 248 ± 3, 242 ± 2 and 237 ± 2 nm for cb-Ery, nCD4 and Mon, respectively (P-values < $9.1e^{-163}$ in each of the pairwise comparisons using two-samples Kolmogorov–Smirnov statistics) (Supplementary Figure S3A), with the topology of the region in cb-Ery being less tightly packed than in nCD4 and Mon. Each ensemble was then clustered by structural similarity (27) and the models from the most populated cluster were selected for the comparative analysis between cell-types. Clustering by dRMSD, confirmed that the topology of the region was markedly different in the three cell-types, with nCD4 and Mon folds being more similar between each other than with cb-Ery (Figure 3B). Particularly interesting is how the topology of the β-globin locus (chr11:5 201 270–5 302 470) varied in the three cell-types. Indeed, in Erythroblasts the β-globin locus appeared to be located further from the main core of the region as compared with naïve CD4+ T cells and Monocytes, with median distances between the centre of mass of the β-globin locus of 286, 243 and 207 nm in cb-Ery, nCD4 and Mon, respectively (P-values < $3.46e^{-101}$ in all the pairwise cell-type comparisons; two-samples Kolmogorov–Smirnov statistic) (Supplementary Figure S3B).

To characterize this further, we focused specifically on the β-globin locus and quantified its spatial organization with respect to hypersensitive site 3 (HS3) in the LCR, which is forming an intricate network of interaction with the β-globin genes (53) and is required for their activation (54). In line with this evidence, in the 3D ensemble of models representing cb-Ery cells, HS3 was significantly closer to HBB, HBD, HBG1, HBG2 and HBE1 genes than in the 3D ensemble of models representing nCD4 and Mon (P-values < 0.007, two-samples Kolmogorov–Smirnov test). In the latter two cell-types HS3 had a similar distance distribution with HBB, HBD, HBG1 and HBG2 genes (P-values > 0.01, two samples Kolmogorov–Smirnov test) (Figure 3C).

Performing 3D enrichment analysis of varied epigenetic features and expression levels around HS3 ('Materials and Methods' section), we unveiled a stark enrichment of active chromatin marks (H3K27ac, H3K36me, H3K4me1 and H3K4me3) and expression levels, and a clear depletion of inactive marks (H3K9me3 and H3K27me3) in cb-Ery. This 3D functional signature could not be inferred from the 2D genomic track (Supplementary Figure S4A) and was absent in nCD4 and Mon, where active chromatin marks and transcript levels were depleted (Figure 3D and E; Supplementary Figure S5). Overall, our models recapitulated the different 3D organization of the β-globin locus and highlight the existence of a specific 3D functional signature enriched in active chromatin features that characterized the active β-globin locus in cb-Ery.

Active gene communities in cb-Ery: a cell-type-specific 3D signature

To examine whether the specific 3D functional signature of the active β-globin locus influence its genomic neighbourhood, we investigated its long-range interaction patterns. Comparative analysis of the distance profile between HBG2 (the most expressed gene in cb-Ery) and each of the selected loci (chr11: 3 795 000–8 505 000 bp), revealed the existence of an intricate cell-type-specific network of spatially proximal expressed genes (Figure 4A), in line with previous observations of transcribed genes co-localizing in space (24,55,56,57,58). This network comprised distal transcribed sites (even located at 1.4 Mb away as STIM1) that showed cell-type-specific spatial proximity. Indeed, HBG2 in cb-Ery was in closer proximity with all other expressed loci of the genomic neighbourhood than in nCD4 and Mon (Figure 4B).

To further characterize the cell-type-specific spatial distribution of these transcribed loci, we clustered their relative distances within the ensembles of 3D models and identified communities of expressed genomic loci (Figure 4C–E and 'Materials and Methods' section). Then, we quantified the amount of times a given community of expressed genomic loci occurred within the ensembles of 3D models (i.e. the co-occurrence score, 'Materials and Methods' section) and used this quantification as a proxy to define the 'community stability'. This analysis revealed the existence of highly variable communities of expressed genomic loci that followed a cell-type-specific segregation in the 3D space. Interestingly, the organization of these communities was overall more stable in cb-Ery than in nCD4 and Mon, where less defined communities were identified. Indeed, as assessed by the mean inter-community co-occurrence scores ('Materials and Methods' section), the cb-Ery network was characterized by the presence of four stable communities ('Materials and Methods' section and Table 1). Whilst the nCD4 network was formed by three communities with overall low co-occurrence (although community 2 in this network showed a stability in line with the communities in the cb-Ery network), and the Mon network formed by only two unstable communities ('Materials and Methods' section and Table 1). Overall, the results highlight the presence of more defined 3D communities of expressed genes in cb-Ery as compared to nCD4 and Mon, suggesting that

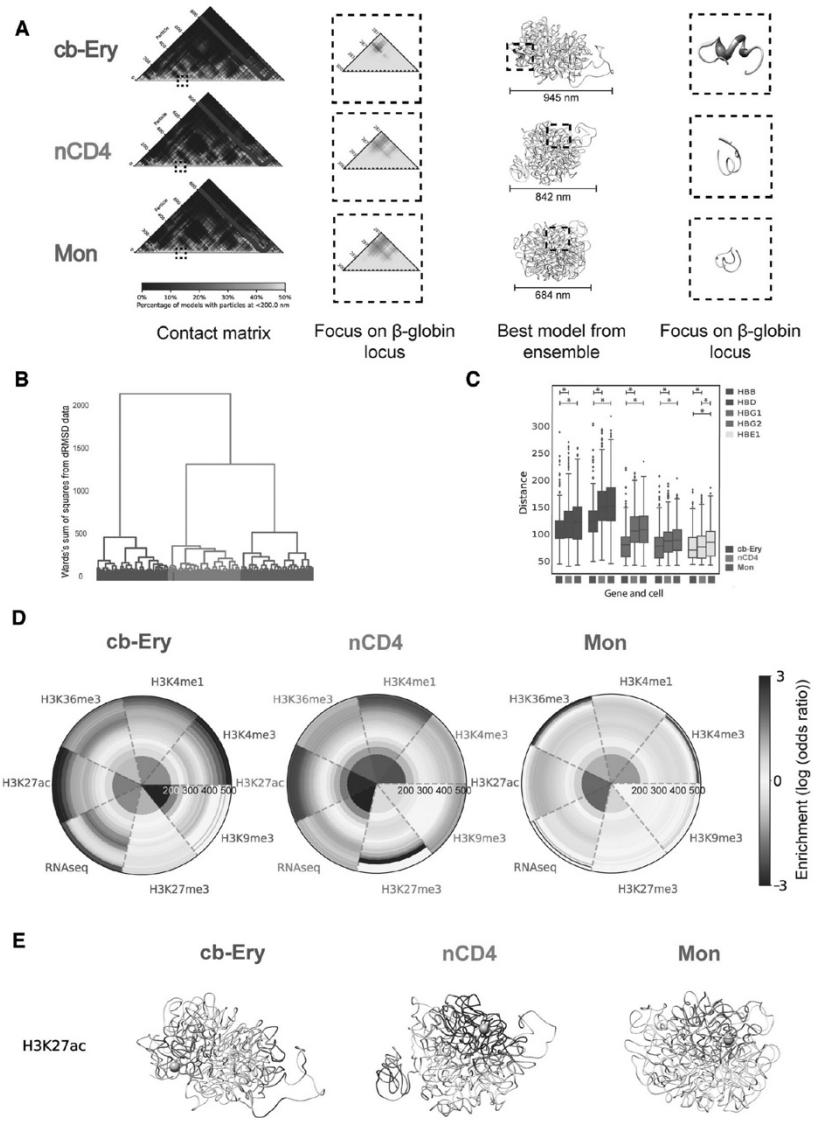


Figure 3. Cell-type-specific organization patterns of the β-globin locus. (**A**) β-globin locus in cb-Ery, nCD4 and Mon cell-types. From left to right: representation of the contact matrix derived from each of the model ensembles colour-coded from low (blue) to high (yellow) contact frequency (columns filtered due to low interaction data are coloured grey); zoom in of the β-globin locus in the matrix; best model from ensemble as assessed by the scoring function; zoom up of the β-globin locus in the model. Models are represented as a tube with thickness proportional to the cell-type expression profile ('Materials and Methods' section), the regulatory elements and genes in the β-globin locus are coloured as follow: HBB and HBD in red, HBG1 and HBG2 in green, HBE1 in yellow, LCR in blue and 3'HS1 and HS5 in orange. Model images were rendered with the Chimera visualization software (74). (**B**) Clustering tree (see 'Hierarchical clustering of ensembles of 3D models' section in Chromatin ensemble 3D analysis) of cb-Ery (purple), nCD4 (orange) and Mon (pink) model ensembles. (**C**) Cell-type-specific distance distributions between the particle containing HS3 site of the LCR and the β-globin genes (HBB, HBD, HBG1, HBG2, and HBE1, colour coded as in (A)) as observed in the ensemble of models. Box boundaries represent first and third quartiles, middle line represents median, and whiskers extend to 1.5 times the interquartile range (two-samples Kolmogorov–Smirnov test, asterisk indicate $P < 0.007$). (**D**) Radial plot showing the 3D enrichment around HS3 ('Materials and Methods' section). Each circumference shows the enrichment or depletion of features around HS3 on layers (up to 560 nm away from HS3) of non-overlapping volumes equal to the one of the initial sphere with radius of 200 nm. The colour bar shows the colour coding from highly depleted (blue) to highly enriched (red) features. (**E**) The representative 3D model of each of the ensembles (cb-Ery, Mon and nCD4) is represented as a tube and colour-coded by the 3D enrichment analysis of H3K27ac (from highly depleted in blue to highly enriched in red) around HS3 (represented as a light blue sphere).

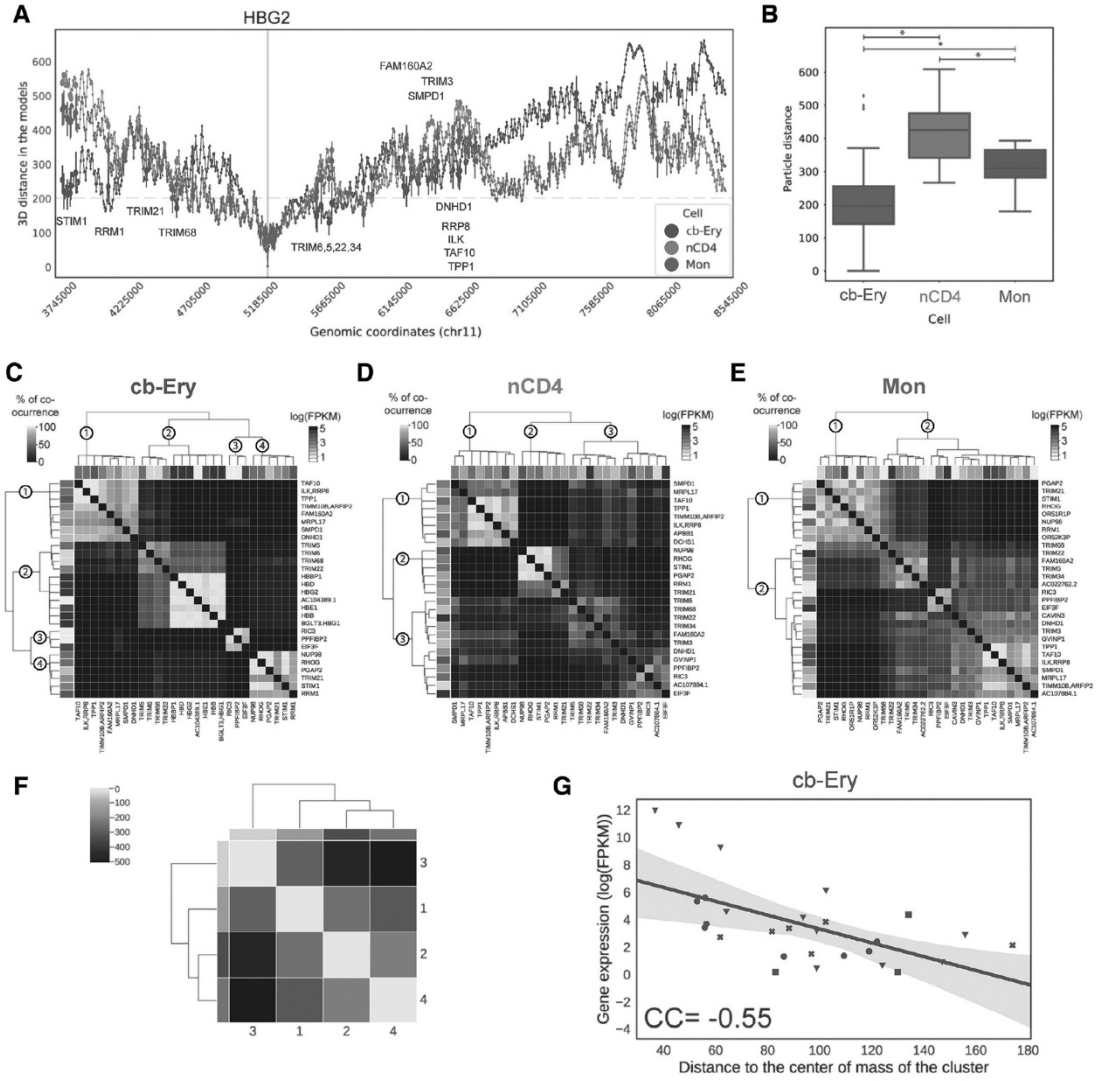


Figure 4. Communities of active genes as a cell-type-specific 3D signature in cb-Ery. (**A**) Line plot of the mean distances between the TSS of HBG2 (focus point, blue vertical line) and all other particles in the genomic region (chr11:3 795 000–8 504 999 bp) for cb-Ery (purple), nCD4 (orange) and Mon (pink) as calculated in each model ensembles. Error bar, indicating one standard deviation, is displayed for particles enclosing a transcribed gene (in at least one cell). The grey dashed line indicates 200 nm cut-off used in the analysis ('Materials and Methods' section). (**B**) Cell-type-specific distance distribution between particles enclosing the HBG2 gene and all transcribed genes in the genomic region (chr11:3 795 000–8 504 999 bp) for cb-Ery (purple), nCD4 (orange), and Mon (pink) as calculated in each model ensembles. Box boundaries represent first and third quartiles, middle line represents median, and whiskers extend to 1.5 times the interquartile range (two-samples Kolmogorov–Smirnov test, asterisk indicate *P*-values $< 7.5e^{-6}$). (**C**–**E**) Hierarchical clustering of each genes based on the co-occurrence analysis ('Materials and Methods' section) in cb-Ery (C), nCD4 (D), and Mon (E). Co-occurrence value range from 0 (low, dark blue) to 100 (high, bright yellow). In each hierarchical tree the communities are labelled at their root branch. Per each gene the relative expression (log(FPKM) is shown in a scale of reds from 0 to 5. (**F**) Hierarchical clustering of the distances between the communities defined in cb-Ery ('Materials and Methods' section). Distance values are coloured in the matrix from dark blue to bright yellow and the average expression in log(FPKM) per community is coloured by ranking from lowest (lightest) to highest (darkest) in three different shades of red. (**G**) Relationship between gene expression in log(FPKM) and the median distance of the gene particles to the centre of mass of its own community in cb-Ery ensemble of models ('Materials and Methods' section). Purple line denotes the linear regression fit, the shading around the regression line represents the confidence interval, each community is represented with different symbols (circle community 1; inverse triangle community 2; square community 3; and ex community 4).

Table 1. Communities stability assessment

Cell	Community	Mean inter-community co-occurrence	Average inter-community co-occurrence per cell
cb-Ery	1	2.96	3.06
	2	4.90	
	3	0.54	
	4	3.85	
nCD4	1	11.49	9.16
	2	3.83	
	3	12.17	
Mon	1	10.33	10.33
	2	10.33	

Description — Cell: the cell-type data used to reconstruct the chromatin; Community: the defined communities by Ward's clustering; Mean inter-community co-occurrence: communities stability score as defined in 'Materials and Methods' section; Average inter-community co-occurrence per cell: average mean inter-community co-occurrence value of all the communities in each of the cells.

the co-occurrence of these segregated communities within an ensemble of possible folds is part of the cell-type-specific 3D signature.

Next, we investigated whether the stability of the 3D communities of expressed genes in cb-Ery could be related to the high levels of expression of the β-globin genes (highest as HBG2 with 10.86 FPKM, whilst the mean expression of all the other expressed genes in nCD4 and Mon was 2.45 and 2.10 FPKM, respectively). Clustering the distance distribution between the centres of mass of each community in cb-Ery (Figure 4F) revealed a clear hierarchical organization with the most expressed community, which included the highly expressed β-globin locus (Supplementary Table S3), located in the centre, and the least expressed community in the periphery. This pattern was not present in nCD4, and impossible to address in Mon with just two communities (Supplementary Figure S6A and B). This suggests a hierarchical organization in cb-Ery, in which the location in space of each of the communities and their levels of expression are related. Surprisingly, this hierarchy was also overall present at the community level in cb-Ery, where the distance between each gene to the centre of mass of the community and its expression were negatively correlated (CC: −0.55, *P*-value = 0.002; Figure 4G). This suggests the formation in cb-Ery of a gradient of expression within the community were the most expressed genes are located in the centre of their communities and the less expressed ones are preferentially located in the periphery in line with the organization previously observed for the alpha-globin locus (24). This overall community organization was not evident in nCD4 and Mon (Supplementary Figure S6C and D), thus suggesting that the high expression of the β-globin loci in cb-Ery could be associated with the establishment of a hierarchical organization in the loci.

DISCUSSION

Here, we have introduced an integrative modelling method for the 3D reconstruction, analysis and interpretation of sparse 3C-based datasets such as pcHi-C. We also demonstrate its usability in the comparative 3D analysis of genomic regions using the β-globin locus as an example, showing that our method can detect cell-type-specific 3D organizational features within genomic regions that can lead to several important implications on the relationship between genomic function and spatial genome organization, such as the expression dependent organization of active loci.

Generally, the analysis and interpretation of sparse 3C-datasets is not trivial and specialized analytical tools are required. In the case of pcHi-C, the available tools (ChiCMaxima, Chicago, Chicdiff, Gothic, HiCapTools (59,60,61,62,63)) are mainly focused on the implementation of normalization strategies to reduce the impact of non-biological biases and on strategies to detect interaction between captured loci. Conversely, the integrative modelling method presented in this study has been designed for the analysis and interpretation of sparse 3C-datasets in their third dimension, allowing for data normalization, detection of significative interaction, and most importantly, the recovery of the full structural organization of a genomic region despite of the data sparseness.

Indeed, here we extensively tested our procedure by comparing models reconstructed directly from sparse and dense datasets, showing that 3D models reconstructed by the integrative modelling method for sparse data modelling are a good representation of the dense experiment. In fact, model reconstruction is only minorly affected by the intrinsic experimental biases of the capture experiment. Additionally, and most importantly, our model procedure reproduces remarkably well the whole 3D organization of the selected genomic regions even recovering the organization of loci that are not included as input restraints and are not readily observable in the sparse experiment.

Next, to assess whether the 3D reconstructed models were not only a *bona fide* representation of models based on Hi-C datasets, we used a 'synthetic' toy genome with known 3D organization (37) and proved that we can efficiently model sparse pcHi-C-like datasets using as few as 2–3% of all possible interaction data. Importantly, this quantification highlights how the degree of sparseness of the data is related to the efficiency of the 3D reconstruction process and provide a general guideline for sparse data modelling. In light of this, we speculate that our integrative approach could easily be applied to different type of 3C datasets with similar sparseness. For example, protein-centric chromatin conformation method such as HiChIP (19) could be used as input experiment to reconstruct the chromatin folding, assuming that the protein-capture biases of this type of experiments are similar to the promoter-capture biases observed in the pcHiC experiments.

Finally, to illustrate the utility of our integrative approach, we applied it to the β-globin locus, whose 3D organization has been extensively studied (51,53,64,65,66). We investigated this locus in three different cell-types (cb-Ery, nCD4 and Mon) and performed a comparative analysis between them. In agreement with previous studies (33), our models show that the topology of the β-globin locus varies in the three cell-types owing to their differential expression. Interestingly, our models also unveil that the globin HBG2 gene is embedded in an epigenetically ac-

tive and highly transcribed neighbourhood in cb-Ery giving rise to a locus-specific 3D functional signature. This functional signature is absent in the models of other cell-types (nCD4 and Mon), where the locus is not expressed. We also show that this cell-specific organization, not only occurs proximally to the β-globin genes but also involves loci located at longer genomic distances (more than 1 Mb away). Indeed, our 3D comparative analysis unveiled the existence of an intricate cell-type-specific network of spatially proximal expressed genes that forms gene communities that are segregated in the 3D space in a cell-type-specific fashion. The identified communities are compatible with the formation of chromatin foci in which transcribed genes co-localize as a general mechanism to organize gene transcription (24,55,56,57,58,67). Interestingly, we observed that the co-occurrence within the ensemble of models of the identified cell-type-specific communities is cell-type dependent, with the cb-Ery communities network formed by more persistent communities than the nCD4 and Mon community networks. This suggests that also the degree of co-occurrence of the communities within the ensemble is an important feature for the identification of a cell-type-specific 3D signature. Additionally, we observed that in cb-Ery, where the β-globin genes are highly expressed, the communities present an overall hierarchical spatial organization, both between and within communities. This topology is dependent on the level of transcription with highly expressed entities (entire community or specific gene within a community) located in the core of the hierarchical 3D organization and low expressed entities found at the periphery. We hypothesize that the observed communities could represent cell-type-specific transcription factories (24,67,68,69) or phase-separated foci (70,71,72) organized following a gradient of transcription with high concentration of nascent transcripts and macromolecular protein complexes in the core of the assemblies that create a 'sticky' environment for the less expressed peripheral loci. This hierarchical organization is only marginally present in nCD4 and Mon, suggesting that it contributes to the cell-type-specific 3D signature characterizing the β-globin region in cb-Ery. However, the long-range interactions between the active β-globin locus and other active gene loci have been seen to be not dependent on the process of ongoing transcription or on the binding of RNAPII to regulatory elements (73), suggesting that the observed communities' organization is more likely dependent on high concentrations of other macromolecular protein complexes in the 'sticky' core of the hierarchical 3D organization.

In summary, we have shown that sparse datasets like pcHi-C can be effectively used to model in 3D the spatial conformation of genomic domains. The resulting models retain most of the genomic region organization and recover also the organization of loci that are not readily observable in the sparse experiment. Importantly, this is achievable with a very small percentage (∼2–3%) of all possible interaction data in the genomic region. Additionally, our study not only provides a novel approach for sparse-data 3D modelling but also introduces new tools for the comparative analysis of genomic regions. Thus, it will aid the discovery of cell-type-specific 3D signatures and help deciphering complex mechanism underlying the cell-type-specific 3D genome organization.

DATA AVAILABILITY

Hi-C data for GM12878 cell line were obtained from Gene Expression Omnibus (GEO) at the accession number GSE63525. pcHi-C data from GM12878 cell line were obtained from ArrayExpress at the accession number E-MTAB-2323. pcHi-C data for cb-Ery, nCD4 and Mon cells were obtained from the European Genome-phenome Archive at the accession number EGAS00001001911. Expression matrix for cb-Ery, nCD4 and Mon cells was downloaded from https://osf.io/u8tzp/ (GeneExpressionMatrix.txt.gz).

The code used to reconstruct 3D models from the sparse-data modelling approach, to analyze data and to generate figures is available in the GitHub repository (https://github.com/3DGenomes/SparseDataModelling).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank all the current and past members of the Marti-Renom lab for their continuous discussions and support. Dr Irene Miguel-Escalada and Dr Wouter de Laat for helpful discussions. The 4D genome unit at CRG for data availability. The Javierre lab for providing access to the ChIP-seq peaks for the β-globin locus in different cell-types. We acknowledge the ENCODE consortium and the ENCODE production laboratories that generated the datasets used in the manuscript. This study makes use of data generated by the PCHI-C Consortium available in the EGA European Genome-Phenome Archive (National Institute for Health Research of England, UK Medical Research Council (MR/L007150/1) and UK Biotechnology and Biological Research Council (BB/J004480/1)).

FUNDING

European Research Council under the 7th Framework Program FP7/2007–2013 [609989, in part]; European Union's Horizon 2020 Research and Innovation Programme [676556]; Spanish Ministerio de Ciencia, Innovación y Universidades [BFU2013–47736-P, BFU2017–85926-P to M.A.M-R; IJCI-2015–23352 to I.F]; Fundació la Marató de TV3 [201611 to M.A.M-R.]; CRG acknowledges support from Centro de Excelencia Severo Ochoa 2013–2017'; SEV-2012–0208; CERCA Programme/Generalitat de Catalunya; Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III and the EMBL partnership; Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement; European Regional Development Fund (ERDF) by the Spanish Ministry of Science and Innovation corresponding to the Programa Opertaivo FEDER Plurirregional de España (POPE) 2014–2020; Secretaria d'Universitats i Recerca, Departament d'Empresa i Coneixement of the Generalitat de Catalunya corresponding to the programa Operatiu FEDER Catalunya 2014–2020. Funding for open access charge: Spanish Ministerio de Ciencia, Innovación y Universidades [BFU2017–85926-P].

Conflict of interest statement. None declared.

REFERENCES

1. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
2. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
3. Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., van Berkum,N.L., Meisig,J., Sedat,J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.
4. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Leblanc,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, **148**, 458–472.
5. Hsieh,T.S., Cattoglio,C., Slobodyanyuk,E., Hansen,A.S., Rando,O.J., Tjian,R. and Darzacq,X. (2020) Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Mol. Cell*, **78**, 539–553.
6. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
7. Bonev,B., Mendelson Cohen,N., Szabo,Q., Fritsch,L., Papadopoulos,G.L., Lubling,Y., Xu,X., Lv,X., Hugnot,J.P., Tanay,A. *et al.* (2017) Multiscale 3D genome rewiring during mouse neural development. *Cell*, **171**, 557–572.
8. Zheng,H. and Xie,W. (2019) The role of 3D genome organization in development and cell differentiation. *Nat. Rev. Mol. Cell Biol.*, **20**, 535–550.
9. Kempfer,R. and Pombo,A. (2020) Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.*, **21**, 207–226.
10. Dekker,J., Marti-Renom,M.A. and Mirny,L.A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, **14**, 390–403.
11. Nagano,T., Lubling,Y., Stevens,T.J., Schoenfelder,S., Yaffe,E., Dean,W., Laue,E.D., Tanay,A. and Fraser,P. (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**, 59–64.
12. Ramani,V., Deng,X., Qiu,R., Lee,C., Disteche,C.M., Noble,W.S., Shendure,J. and Duan,Z. (2020) Sci-Hi-C: a single-cell Hi-C method for mapping 3D genome organization in large number of single cells. *Methods*, **170**, 61–68.
13. Flyamer,I.M., Gassler,J., Imakaev,M., Brandao,H.B., Ulianov,S.V., Abdennur,N., Razin,S.V., Mirny,L.A. and Tachibana-Konwalski,K. (2017) Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, **544**, 110–114.
14. Hsieh,T.H., Weiner,A., Lajoie,B., Dekker,J., Friedman,N. and Rando,O.J. (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, **162**, 108–119.
15. Beagrie,R.A., Scialdone,A., Schueler,M., Kraemer,D.C., Chotalia,M., Xie,S.Q., Barbieri,M., de Santiago,I., Lavitas,L.M., Branco,M.R. *et al.* (2017) Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, **543**, 519–524.
16. Quinodoz,S.A., Ollikainen,N., Tabak,B., Palla,A., Schmidt,J.M., Detmar,E., Lai,M.M., Shishkin,A.A., Bhat,P., Takei,Y. *et al.* (2018) Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*, **174**, 744–757.
17. van de Werken,H.J., de Vree,P.J., Splinter,E., Holwerda,S.J., Klous,P., de Wit,E. and de Laat,W. (2012) 4C technology: protocols and data analysis. *Methods Enzymol.*, **513**, 89–112.
18. Allahyar,A., Vermeulen,C., Bouwman,B.A.M., Krijger,P.H.L., Verstegen,M., Geeven,G., van Kranenburg,M., Pieterse,M., Straver,R., Haarhuis,J.H.I. *et al.* (2018) Enhancer hubs and loop collisions identified from single-allele topologies. *Nat. Genet.*, **50**, 1151–1160.
19. Mumbach,M.R., Rubin,A.J., Flynn,R.A., Dai,C., Khavari,P.A., Greenleaf,W.J. and Chang,H.Y. (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, **13**, 919–922.
20. Schoenfelder,S., Furlan-Magaril,M., Mifsud,B., Tavares-Cadete,F., Sugar,R., Javierre,B.M., Nagano,T., Katsman,Y., Sakthidevi,M., Wingett,S.W. *et al.* (2015) The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.*, **25**, 582–597.
21. Bendandi,A., Dante,S., Zia,S.R., Diaspro,A. and Rocchia,W. (2020) Chromatin compaction multiscale modeling: a complex synergy between theory, simulation, and experiment. *Front. Mol. Biosci.*, **7**, 15–21.
22. Oluwadare,O., Highsmith,M. and Cheng,J. (2019) An overview of methods for reconstructing 3-D chromosome and genome structures from Hi-C data. *Biol. Proc. Online*, **21**, 7.
23. Serra,F., Di Stefano,M., Spill,Y.G., Cuartero,Y., Goodstadt,M., Bau,D. and Marti-Renom,M.A. (2015) Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett.*, **589**, 2987–2995.
24. Baù,D., Sanyal,A., Lajoie,B.R., Capriotti,E., Byron,M., Lawrence,J.B., Dekker,J. and Marti-Renom,M.A. (2011) The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.*, **18**, 107–114.
25. Tjong,H., Li,W., Kalhor,R., Dai,C., Hao,S., Gong,K., Zhou,Y., Li,H., Zhou,X.J., Le Gros,M.A. *et al.* (2016) Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc. Natl Acad. Sci. U.S.A.*, **113**, E1663–E1672.
26. Hua,N., Tjong,H., Shin,H., Gong,K., Zhou,X.J. and Alber,F. (2018) Producing genome structure populations with the dynamic and automated PGS software. *Nat. Protoc.*, **13**, 915–926.
27. Serra,F., Bau,D., Goodstadt,M., Castillo,D., Filion,G.J. and Marti-Renom,M.A. (2017) Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.*, **13**, e1005665.
28. Irastorza-Azcarate,I., Acemel,R.D., Tena,J.J., Maeso,I., Gomez-Skarmeta,J.L. and Devos,D.P. (2018) 4Cin: a computational pipeline for 3D genome modeling and virtual Hi-C analyses from 4C data. *PLoS Comput. Biol.*, **14**, e1006030.
29. Di Stefano,M., Stadhouders,R., Farabella,I., Castillo,D., Serra,F., Graf,T. and Marti-Renom,M.A. (2020) Transcriptional activation during cell reprogramming correlates with the formation of 3D open chromatin hubs. *Nat. Commun.*, **11**, 2564.
30. Paulsen,J., Gramstad,O. and Collas,P. (2015) Manifold based optimization for single-cell 3D genome reconstruction. *PLoS Comput. Biol.*, **11**, e1004396.
31. Stevens,T.J., Lando,D., Basu,S., Atkinson,L.P., Cao,Y., Lee,S.F., Leeb,M., Wohlfahrt,K.J., Boucher,W., O'Shaughnessy-Kirwan,A. *et al.* (2017) 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, **544**, 59–64.
32. Mifsud,B., Tavares-Cadete,F., Young,A.N., Sugar,R., Schoenfelder,S., Ferreira,L., Wingett,S.W., Andrews,S., Grey,W., Ewels,P.A. *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.
33. Javierre,B.M., Burren,O.S., Wilder,S.P., Kreuzhuber,R., Hill,S.M., Sewitz,S., Cairns,J., Wingett,S.W., Varnai,C., Thiecke,M.J. *et al.* (2016) Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, **167**, 1369–1384.
34. Vidal,E., le Dily,F., Quilez,J., Stadhouders,R., Cuartero,Y., Graf,T., Marti-Renom,M.A., Beato,M. and Filion,G.J. (2018) OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. *Nucleic Acids Res.*, **46**, e49.
35. Yang,T., Zhang,F., Yardimci,G.G., Song,F., Hardison,R.C., Noble,W.S., Yue,F. and Li,Q. (2017) HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.*, **27**, 1939–1949.
36. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.
37. Trussart,M., Serra,F., Bau,D., Junier,I., Serrano,L. and Marti-Renom,M.A. (2015) Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. *Nucleic Acids Res.*, **43**, 3465–3477.

38. Di Stefano,M., Rosa,A., Belcastro,V., di Bernardo,D. and Micheletti,C. (2013) Colocalization of coregulated genes: a steered molecular dynamics study of human chromosome 19. *PLoS Comput. Biol.*, **9**, e1003019.
39. Kremer,K. and Grest,G.S. (1990) Dynamics of entangled linear polymer melts: a molecular-dynamics simulation. *J. Chem. Phys.*, **92**, 5057–5086.
40. Rosa,A. and Everaers,R. (2008) Structure and dynamics of interphase chromosomes. *PLoS Comput. Biol.*, **4**, e1000153.
41. Polak,E. and Ribiere,G. (1969) Note sur la convergence de méthodes de directions conjuguées. *Rev. Fran Inf. Rech. Op.*, **16**, 35–43.
42. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.
43. Zwillinger,D. and Kokoska,S. (2000) In: *RC Standard Probability and Statistics Tables and Formulae*. Chapman & Hall/CRC, Boca Raton, FL.
44. Ward,J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Statist. Assoc.*, **58**, 236–244.
45. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–W598.
46. Caliński,T. and Harabasz,J. (1974) A dendrite method for cluster analysis. *Commun. Stat.*, **3**, 1–27.
47. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
48. Jung,I., Schmitt,A., Diao,Y., Lee,A.J., Liu,T., Yang,D., Tan,C., Eom,J., Chan,M., Chee,S. *et al.* (2019) A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.*, **51**, 1442–1449.
49. Miguel-Escalada,I., Bonas-Guarch,S., Cebola,I., Ponsa-Cobas,J., Mendieta-Esteban,J., Atla,G., Javierre,B.M., Rolando,D.M.Y., Farabella,I., Morgan,C.C. *et al.* (2019) Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.*, **51**, 1137–1148.
50. Russel,D., Lasker,K., Webb,B., Velazquez-Muriel,J., Tjioe,E., Schneidman-Duhovny,D., Peterson,B. and Sali,A. (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.*, **10**, e1001244.
51. Palstra,R.J., Tolhuis,B., Splinter,E., Nijmeijer,R., Grosveld,F. and de Laat,W. (2003) The beta-globin nuclear compartment in development and erythroid differentiation. *Nat. Genet.*, **35**, 190–194.
52. Levings,P.P. and Bungert,J. (2002) The human beta-globin locus control region. *Eur. J. Biochem.*, **269**, 1589–1599.
53. Liu,X., Zhang,Y., Chen,Y., Li,M., Zhou,F., Li,K., Cao,H., Ni,M., Liu,Y., Gu,Z. *et al.* (2017) In situ capture of chromatin interactions by biotinylated dCas9. *Cell*, **170**, 1028–1043.
54. Fraser,P., Pruzina,S., Antoniou,M. and Grosveld,F. (1993) Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes. *Genes Dev.*, **7**, 106–113.
55. Fraser,P. and Bickmore,W. (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature*, **447**, 413–417.
56. Jackson,D.A., Hassan,A.B., Errington,R.J. and Cook,P.R. (1993) Visualization of focal sites of transcription within human nuclei. *EMBO J.*, **12**, 1059–1065.
57. Osborne,C.S., Chakalova,L., Brown,K.E., Carter,D., Horton,A., Debrand,E., Goyenechea,B., Mitchell,J.A., Lopes,S., Reik,W. *et al.* (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.*, **36**, 1065–1071.
58. Osborne,C.S., Chakalova,L., Mitchell,J.A., Horton,A., Wood,A.L., Bolland,D.J., Corcoran,A.E. and Fraser,P. (2007) Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. *PLoS Biol.*, **5**, e192.
59. Ben Zouari,Y., Molitor,A.M., Sikorska,N., Pancaldi,V. and Sexton,T. (2019) ChiCMaxima: a robust and simple pipeline for detection and visualization of chromatin looping in Capture Hi-C. *Genome Biol.*, **20**, 102.
60. Cairns,J., Freire-Pritchett,P., Wingett,S.W., Varnai,C., Dimond,A., Plagnol,V., Zerbino,D., Schoenfelder,S., Javierre,B.M., Osborne,C. *et al.* (2016) CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.*, **17**, 127.
61. Cairns,J., Orchard,W.R., Malysheva,V. and Spivakov,M. (2019) Chicdiff: a computational pipeline for detecting differential chromosomal interactions in Capture Hi-C data. *Bioinformatics*, **35**, 4764–4766.
62. Mifsud,B., Martincorena,I., Darbo,E., Sugar,R., Schoenfelder,S., Fraser,P. and Luscombe,N.M. (2017) GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PLoS One*, **12**, e0174744.
63. Anil,A., Spalinskas,R., Akerborg,O. and Sahlen,P. (2018) HiCapTools: a software suite for probe design and proximity detection for targeted chromosome conformation capture applications. *Bioinformatics*, **34**, 675–677.
64. Brown,J.M., Leach,J., Reittie,J.E., Atzberger,A., Lee-Prudhoe,J., Wood,W.G., Higgs,D.R., Iborra,F.J. and Buckle,V.J. (2006) Coregulated human globin genes are frequently in spatial proximity when active. *J. Cell Biol.*, **172**, 177–187.
65. Schubeler,D., Francastel,C., Cimbora,D.M., Reik,A., Martin,D.I. and Groudine,M. (2000) Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus. *Genes Dev.*, **14**, 940–950.
66. Huang,P., Keller,C.A., Giardine,B., Grevet,J.D., Davies,J.O.J., Hughes,J.R., Kurita,R., Nakamura,Y., Hardison,R.C. and Blobel,G.A. (2017) Comparative analysis of three-dimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element. *Genes Dev.*, **31**, 1704–1713.
67. Sanyal,A., Bau,D., Marti-Renom,M.A. and Dekker,J. (2011) Chromatin globules: a common motif of higher order chromosome structure? *Curr. Opin. Cell Biol.*, **23**, 325–331.
68. Sutherland,H. and Bickmore,W.A. (2009) Transcription factories: gene expression in unions? *Nat. Rev. Genet.*, **10**, 457–466.
69. Iborra,F.J., Pombo,A., Jackson,D.A. and Cook,P.R. (1996) Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. *J. Cell Sci.*, **109**, 1427–1436.
70. Gurumurthy,A., Shen,Y., Gunn,E.M. and Bungert,J. (2019) Phase separation and transcription regulation: are super-enhancers and locus control regions primary sites of transcription complex assembly? *Bioessays*, **41**, e1800164.
71. Boija,A., Klein,I.A., Sabari,B.R., Dall'Agnese,A., Coffey,E.L., Zamudio,A.V., Li,C.H., Shrinivas,K., Manteiga,J.C., Hannett,N.M. *et al.* (2018) Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell*, **175**, 1842–1855.
72. Cho,W.K., Spille,J.H., Hecht,M., Lee,C., Li,C., Grube,V. and Cisse,II. (2018) Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, **361**, 412–415.
73. Palstra,R.J., Simonis,M., Klous,P., Brasset,E., Eijkelkamp,B. and de Laat,W. (2008) Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PLoS One*, **3**, e1661.
74. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

Chapter III

Probabilistic 3D-modelling of genomes and genomic domains by integrating high-throughput imaging and Hi-C using machine learning

Among the existing techniques for interrogating the genome structure, Hi-C assays have become the most performed experiments and constitute the majority of the publicly available datasets. As a result, there is a continuous demand to create and improve algorithms and methods to assist the scientific community in the interpretation of Hi-C experimental data. Here we introduce probabilistic TADbit (pTADbit), a new approach that combines Deep Learning and restraint-based modelling to infer the three-dimensional (3D) structure of genome and genomic domains interrogated by Hi-C experiments. pTADbit uses thousands of microscopy-based distances between genomic loci to train a neural network model that aims at predicting the population distribution of the spatial distance between two genomic loci based solely on their Hi-C interaction frequency. pTADbit produces more accurate chromatin models compared to the original TADbit as well as other available 3D modeling methods, while drastically reducing the required computation time. The resulting ensemble of models not only agree consistently with independent measures obtained by imaging experiments but also better capture the heterogeneity of the cell population. The development of pTADbit lays the basis for the integration of data produced from high-throughput imaging assays into the 3D modelling genomes and genomic domains.

Probabilistic 3D-modelling of genomes and genomic domains by integrating high-throughput imaging and Hi-C using machine learning.

David Castillo[1], Julen Mendieta-Esteban[1,^], and Marc A. Marti-Renom[1,2,3,4,*]

1. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain.

2. Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, 08003 Barcelona, Spain.

3. Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain.

4. ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain.

*To whom correspondence should be addressed: M.A.M-R. martirenom@cnag.crg.eu

^Current address: University of Navarra, Pamplona, Spain.

Abstract

Among the existing techniques for interrogating the genome structure, Hi-C assays have become the most performed experiments and constitute the majority of the publicly available datasets. As a result, there is a continuous demand to create and improve algorithms and methods to assist the scientific community in the interpretation of Hi-C experimental data. Here we introduce probabilistic TADbit (pTADbit), a new approach that combines Deep Learning and restraint-based modelling to infer the three-dimensional (3D) structure of genome and genomic domains interrogated by Hi-C experiments. pTADbit uses thousands of microscopy-based distances between genomic loci to train a neural network model that aims at predicting the population distribution of the spatial distance between two genomic loci based solely on their Hi-C interaction frequency. pTADbit produces more accurate chromatin models compared to the original TADbit as well as other available 3D modeling methods, while drastically reducing the required computation time. The resulting ensemble of models not only agree consistently with independent measures obtained by imaging experiments but also better capture the heterogeneity of the cell population. The development of pTADbit lays the basis for the integration of data produced from high-throughput imaging assays into the 3D modelling genomes and genomic domains.

Introduction

There are now clear evidences of the gene regulatory roles of the three-dimensional (3D) folding of the DNA inside the nucleus [1-3], many of them unveiled by Chromosome Conformation Capture (3C)-based technologies developed already twenty years ago [4]. Together with the proliferation of the 3C techniques, there has been a parallel development of methods for the inference of the 3D structure of the genome. Examples of recent developments include the reconstruction of structural models from sparse interaction data [5], from single-cell Hi-C information [6], or genome-wide low-resolution models of diploid genomes [7]. Thanks to the developed methods and their resulting models, we are gaining key insights on specific biological processes in the field of structural genomics [8]. The long list of implemented algorithms [9, 10] is a prove of the effort of the scientific community to provide the needed tools to analyze and interpret 3C-based experiments.

Among the 3C techniques, Hi-C experiments are the most applied 3C assays resulting in the majority of the publicly available datasets [11]. The main output of a Hi-C experiment is an interaction matrix representing the frequency at which two regions (or loci) of the genome are found crosslinked together within the nucleus in thousands to millions of cells. This population-based interaction matrix provides fundamental information of the 3D structure of the genome but it is, at the same time, difficult to interpret and integrate with other biological evidences as it is not a direct measure of the spatial distances at which interactions occur [12]. Therefore, an accurate measure or estimation of the physical distances at which genomic regions interact is essential for accurately characterizing how nuclear processes occur.

The inference of the 3D structure of genomes based on experiments is a process that is referred as 3D modelling. One of the strategies for the determination of those 3D conformations is restraint-based modelling in which the interaction frequency between fragments of DNA is transformed into a set of spatial restraints that are then satisfied in the resulting structures [13]. In general, finding the optimal equivalence between the frequencies and the physical distances is a key step, and requires either computationally intensive algorithms or the introduction of empirical parameters. In the first case, it is often found that the imposed restraints are similarly satisfied in conformations at different scales. In the second, the inclusion of empirical parameters might introduce some degree of arbitrary in the results. One of the available restraint-based solutions for the modelling of Hi-C information is TADbit [14], a complete Python library covering all steps in the analysis of 3C-based data. TADbit models have already provided significant biological insights (see for example [15-17]). The modeling step of the TADbit pipeline consists in the building of 3D ensembles from Hi-C interaction matrices. The conversion of interaction frequencies to physical distances involves a comprehensive and computationally expensive search of the optimal parameters that will produce models having the proper scale. Moreover, and importantly, the resulting ensemble of models, which is based on the population average contact frequency, has been shown not to fully reflect the variability observed in

the cell population [18]. This inability to completely reproduce the cell-to-cell heterogeneity, far from being exclusive to TADbit, is a common drawback in many 3D modelling approaches.

Here we present *probabilistic* TADbit (pTADbit), developed to overcome the mentioned limitations by using Machine Learning (ML) in the modelling of three-dimensional genomic regions. The main idea of the new method is to use the abundant information of the recent high-throughput imaging datasets [19, 20] to produce more accurate chromatin models. Classically, measures from imaging assays like Fluorescent In Situ Hybridization (FISH) have been exclusively used to validate the accuracy of the resulting models. Nowadays, and thanks to the recent advances in the imaging of the genome [19-21], the amount of available large-scale datasets, both in terms of interrogated loci and number of cells imaged, has exponentially increased allowing for the development of predictive models based on Artificial Intelligence (AI). In pTADbit, the distances between genomic loci obtained from imaging experiments are used to train Neural Networks (NN). Such tens of thousands of image-based distances between two particular loci are sufficient to generate a smooth histogram, which can be then fitted to reconstruct a given mathematical function. Next, the NNs are trained to predict the necessary parameters to reconstruct a probability density function of distances solely from the Hi-C matrix and the genomic distance of the interacting loci (**Fig. 1** and **Methods**). Convolutional Neural Networks (CNN) are widely used in image classification and recognition tasks due to its efficient use of two-dimensional convolutional layers. pTADbit benefits from that efficiency for the extraction of features and the recognition of patterns in the Hi-C matrix, but instead of using the CNN for the classification of the images, it combines the feature extraction with a regression layer to predict a set of parameters. pTADbit results in more accurate 3D models of genomes and genomic domains when compared with the original TADbit and other available methods. It also results in ensemble of conformations with variability closer to that observed by imaging. Finally, pTADbit reduces the computation time attained by the original TADbit and other restraint-based methods.

Results

Neural Network validation

The distances from the public imaging datasets were grouped into histograms, which were next approximated by exponentially modified Gaussian functions. Each function was depicted using three parameters: *K, loc and scale*. The goal of the trained Neural Networks (NN) was thus to predict those three parameters using only as input Hi-C data and the genomic distance between the pairs of loci which distance needs to be predicted (**Methods**). Distances between regions of 30Kbp and 250Kbp were used to train the short-range NN and the long-range NN, respectively. It is important to note that the histograms could be better approximated by other existing mathematical functions or higher-order fitting expressions like the *Beta* or a mixture of gaussian curves, but those approximations required the estimation of more variables to reconstruct the histograms. For example, the unnormalized *Beta* function is defined by four parameters compared to the three of the exponentially modified Gaussian function. Nevertheless, the error incurred in the approximation of the histograms did not significantly differ from the one obtained using other higher-order functions (**Fig. 2a**). Thus, we focused on developing NN to predict the three parameters for exponentially modified Gaussian functions.

The error in the estimation of the histograms to an exponentially modified Gaussian increased when pairs of loci were closer in genomic distance as those histograms adopt shapes that better approximate to a decreasing exponential than to a normal function (**Fig. 2a**). Additionally, the fitting of the histograms to exponential Gaussian functions resulted on a large *K* variability for pairs of consecutive loci (**Fig. 2b**). Finally, the obtained *K* values as a function of the Hi-C normalized interaction frequency (**Fig. 2c**) indicated that pairs of loci interacting with similar frequencies could be approximated by functions which *K* value considerably differed (**Fig. 2c**). Similar to the histograms of distances between 30Kbp region, we approximated the histograms of distances between 250Kbp regions with exponential Gaussian functions (**Fig. 2d**). The SSE error incurred in the approximation was considerably smaller and more constant than the obtained in the short-range histograms. We also observed a higher SSE in short genomic distances (**Fig. 2d**), which translated in a lower accuracy in the prediction of the histograms in that regime. A large variability of the *K* values in short genomic distances was also observed in the fitting of the histograms of distances between 250kbp regions (**Fig. 2e**).

Next, after determining the best function type to approximate the observed image distances, the short-range NN was trained with 11,723 different histograms from 33,755 imaged cells using 70% of the input data as a training set and the remaining 30% as a validation set. The predicted parameters *K, loc and scale* of the histograms were compared to the expected values, which resulted in high correlation coefficients (that is, *K* with r=0.96, *loc* with r=0.99, and *scale* with r=0.98, **Fig. 3a**).

Importantly, the trained NN for short-range distances resulted in very good agreement for the *loc* and *scale* parameter in all genomic distances tested. However, the agreement was not as good for the *K* value, particularly in short genomic distances characterized by high interaction frequencies, which probably reflects the above discussed inaccuracy of the histogram approximation for pairs of consecutive loci.

Next, to assess the minimum number of imaged distances (or cells) required for a good accuracy prediction by the NN, the CNNs were retrained and tested with increasing sample sizes of randomly selected cells from the *K562_chr21-28-30Mbp* dataset, which is the one set with the largest number of imaged cells (13,997 cells in total). The NN resulted in a plateau Mean Square Error around 1,500-2,000 imaged cells for the three measures of *K*, *loc*, and *scale* (**Fig. 3b**). Interestingly, this is a similar number of cells required to obtain dense Hi-C interaction maps using the so-called "low-input" protocols [22], which may indicate that this is the minimum number of cells required to properly capture the variability in genome structure in a population.

Finally, the long-range NN was trained with 61,789 histograms from 4,848 imaged cells using 70% of the input data as training set and the remaining 30% as validation set. The predictions resulted in high correlation coefficients with the expected values for each parameter *K* (r=0.87), *loc* (r=0.94) and *scale* (r=0.93) (**Fig. 3c**).

pTADbit benchmarking

Numerous algorithms for the 3D modelling of chromatin exist [9, 13]. It is, however, difficult to find implemented methods publicly available that can be directly compared with pTADbit. Many of the existing packages follow a different approach by providing unique consensus solutions instead of ensembles of structures. Others are tailored to model the genome at lower resolutions and are simply not easy to adapt to building structures at 30Kbp. Despite these limitations, together with the original TADbit method [14], we executed the Lorentzian 3D Genome (LorDG) [23] and the Chrom3D [24] packages and compared their resulting models to those obtained by pTADbit. The three methods, as for pTADbit, are able to provide ensembles of structures at the high-resolution (30Kbp) using as input solely the Hi-C interaction matrix. Next, for the three methods, we generated an ensemble of 1,000 models of the region 40Mbp-42.5Mbp in chromosome 21 in IMR90 using as input a Hi-C interaction matrix (**Fig. 4a-d**). As this genomic region has also been imaged but never used for the NN training, we were able to directly compare the results of the models against observed image distances.

An indirect way of benchmarking the generated models is to assess the agreement of a contact map calculated from the generated ensemble of 3D confirmation with that of the input Hi-C interaction matrix [14]. This is accomplished by producing contact maps simulating the crosslinking in the

models at different cutoff distances. For example, the comparison of the 30Kbp resolution contact map of the genomic region chr21:28-30Mbp in IMR90 obtained with pTADbit with 400nm cutoff distance (**Fig. 4b**) results in a Stratum-adjusted Correlation Coefficient (SCC), a metric designed specifically to compare Hi-C matrices [25], of 0.81. All the ensemble of models built by the compared methods result in high SCC values for cutoff distances below 400nm except Chrom3D with values that are slightly lower than the other ensembles. pTADbit and Chrom3D results are more consistent across all distance cutoffs until 500nm (**Fig. 4c**). Interestingly, the original TADbit has a level of accuracy similar to pTADbit for the short range contacts as TADbit indeed optimizes the most likely distance of an interaction given the input matrix [14]. However, the original TADbit clearly suffered in identifying longer-range interactions present in the input Hi-C. This is not the case of the results from pTADbit.

To have a more direct benchmarking of the models, we turned to independent imaging experiments never used for modeling. We compared the pairwise median distances between all loci in the models with the pairwise median distances of high-throughput images from the literature (**Fig. 4d**) [20]. As LorDG and Chrom3D do not explicitly have a scaling factor to a priory assess the real size of the resulting ensemble of models, we adjusted their size multiplying the model coordinates by the median distances of the pTADbit ensemble. That is, LorDG model distances were multiplied by 92.19 and Chrom3D by 49.49. The scaling of the LorDG and Chrom3D ensembles simplified the comparison with the other ensembles. All methods resulted in good correlations for all modeled cases (r=0.92 for TADbit, r=0.96 for pTADbit, 0.90 for Chrom3D, and 0.95 for LorDG, and **Fig. 4d**) but the match of the distances with the ones obtained in the images differed considerably. The original TADbit ensemble exhibit a transformation that undervalued short distances and did not grow linearly. The median distances in the pTADbit ensemble of models grew linearly at almost the same rate as in the images but with a slightly scale offset. This scale offset could be caused by differences in the protocols or conditions used in the acquisition of the images in this dataset compared to the ones used in the datasets of the training of the NNs. Chrom3D did not consistently reproduce the distances as the algorithm emphasizes a subset of bead pairs that significantly interact instead of optimizing distances between large number of pairs. Finally, LorDG transformation was linear but it also undervalued short distances and overvalued the long ones.

Next, to assess if the models reproduce the distance variability observed in the images, we plotted the standard deviation of the pairwise distances with the genomic distance obtained in the ensembles (**Fig. 4e**). We observed that TADbit and LorDG models were too deterministic for all ranges of genomic distances. That is, did not result in an ensemble that captured the variability observed in the images. In turn, Chrom3D models resulted also in lower variability in very short genomic distances. This was likely a consequence of constraining only pairs which interaction values are statistically significant. pTADbit ensemble resulted in a constant distance variability with the

exception of a decrease in variability for very short genomic distances. Although pTADbit and Chrom3D exhibit an increase of the variability of the ensemble of solutions observed in imaging experiments, such variability is still lower for the generated models. The reduced variability of distances in the solutions is further addressed in the **Discussion** section.

Finally, the computational burden of generating an ensemble of 1,000 models using the four methods was assessed on a computational workstation with a 24 core Intel(R) Core (TM) i9-7960X @2.80GHz with 128 Gb of RAM. The ensemble of models was generated for models of increasing size between 1Mbp and 30Mbp (**Fig. 4f**). All methods followed an exponential increase trend as the size of models increase with the exception of LorDG, which appeared to follow a more linear trend. However, LorDG compared worse against all other methods in all tested genomic sizes. pTADbit, in contrast, favorably compared against all other methods for models larger than 7-8Mb, with a reduction of computational time larger as the size of the models increased. In average, pTADbit required about 3.5h of computational time to generate 1,000 models of 30Mbp of size at 30Kbp resolution (that is, 1,000 particles), which is about two thirds the time required for TADbit or Chrom3D.

Modeling additional regions

After validating the ability of pTADbit to recover regions of the genome used in the training phase, we next generated with pTADbit an ensemble of 1,000 structural models for each of the regions 54Mbp-58Mbp and 151Mbp-155Mbp of chromosome 4 in human foreskin fibroblasts (HFFs). The contact maps obtained from the ensembles (**Fig. 5a,b**) resulted in high correlation with their equivalent Hi-C matrices (SCC at 400nm of 0.60 and 0.81, respectively). The median distances between various particles in the ensembles of models with the distances of publicly available imaging data [26] labelling a total of 18 regions with bacterial artificial chromosome (BAC) probes also resulted in high correlations of r=0.93 (**Fig. 5c**) and r=0.89 (**Fig. 5d**) for the 54Mbp-58Mbp and 151Mbp-155Mbp in chromosome 4 in HFF, respectively. Finally, we verified that the contacts maps of the ensembles agreed with the published measures at different distance cutoffs (150, 200 and 350 nm, **Fig. 5e**). The contact maps resulted in high correlation with the published percentages for each of the cutoffs (r=0.83, 0.84 and 0.86 respectively), confirming the accuracy of the obtained models.

Modelling full chromosome 19

With the reduction of the computation times in the generation of the ensembles, pTADbit can now be applied to model large regions of the genome at high resolutions, including entire chromosomes.

We next modeled the entire human chromosome 19 at 30Kbp (**Fig. 6a**), which results in a total of 1,971 particles. The ensemble of 1,000 models required a total of about 30h of computational time on a single workstation with a 24 core Intel(R) Core (TM) i9-7960X @2.80GHz with 128 Gb of RAM. As in the previous validations, the contact map of the ensemble (**Fig. 6b**) resulted in high correlation with the Hi-C matrix of the chromosome (SCC=0.79). Finally, the standard deviation of the pairwise distances across different genomic distances indicated that pTADbit may not capture full variability for very short distances while resulting in larger standard deviations for larger ones (**Fig. 6c**). This was likely the result of combining the modelling of low- and high-resolution structures where the Monte Carlo simulations of the low-resolution models are twice as the ones used in the high-resolution ones (**Methods**).

Discussion

pTADbit provides an update of the original TADbit [14], which makes use for the first time of large-scale image data to train the required transformation of frequency of interactions observed using Hi-C and the physical distance between loci measured by imaging technologies. Indeed, this transformation is key to any algorithm aiming at modeling genomes and genomic domains. Briefly, for the methods benchmarked in this study, the original TADbit included a configurable scale factor that sets the amount of DNA in base-pairs contained in one nanometer. The scale factor drives the structures to the proper range of distances during the optimization step in which a grid search is conducted to find the optimal transformation. Chrom3D relies on two configurable parameters to produce the structures with the appropriate scale: the nuclear radius and the volume of the modelled chromosome as a percentage of it. It does not include any scaling factor for the modelling of genomic regions shorter than full chromosomes. LorDG optimizes $\propto$ in the transformation:

$$d_{ij} = \frac{1}{IF_{ij}^{\propto}}$$

where d_{ij} is the distance between particles *i* and *j* and IF_{ij} its interaction frequency but it does not include any scaling factor. The new pTADbit is thus the first method to implicitly incorporate this transformation in the NNs by integrating the training imaging datasets. The explicit transformation from Hi-C data to distances is one the strengths of pTADbit compared to existing methods because it allows the reduction of the computation times in the generation of the ensembles.

The approach to incorporate the imaging information in the modelling method consists in the grouping of single cell distances into population-based histograms in which the information of the exact conformation adopted by chromatin in each individual cell is lost. Those histograms are then predicted and used in the production of ensembles of individual structures that follow the probabilistic distributions. The impossibility of the prediction of the precise histogram solely from the Hi-C information is overcome by their approximation to exponential Gaussian curves, which can be depicted using three parameters *K*, *loc* and *scale* that can be accurately predicted by the NNs. The approximation of the histograms using other higher-order curves decreased the error of the fitting particularly for distances between consecutive regions but made their prediction more computationally intensive as it required the optimization of additional parameters. Importantly, the accuracy of the prediction for distances between consecutive loci did not significantly change the shape of the chromatin structures, which is mainly driven by distances between far apart regions.

The limited number of high-resolution high-throughput imaging datasets is a restricting factor for the size of the modelled regions. The distances used in the short-range NN were obtained from traces of chromatin regions of 2Mbp length at 30Kbp. Therefore, the histograms predicted by the short-range NN are, conceptually, not valid for pairs of fragments that are farther than the 2Mbp, which

restricts the total length of the modelled region. To overcome this limitation pTADbit implements a multi-resolution approach where distances from traces at lower resolution (250Kbp) are used to train the long-range NN. Then, in the modelling process, low-resolution models are built first to determine the shape of the structures at genomic distances that are above the limitation imposed by the short-range NN. The distances of those low-resolution models are then used to produce the final high-resolution conformations.

Recent studies highlight the importance of providing ensembles of structures as opposed to single consensus solutions better capturing the heterogeneity of the cell population [7]. This is indeed one of the main objectives of pTADbit. However, pTADbit benchmarking still results in variabilities smaller than those observed by microscopy (although significantly larger than other compared approaches). The main reason relies in the assignment of the restraints during the modelling approach (**Methods**). Each reconstructed histogram is used as a probability distribution function from where to sample the distance restraints. The restraints imposed are mutually independent and randomly obtained from the distributions. If, during the sampling process, a distance from the tail of the distribution (far from the median) is assigned between fragment *i* and *j*, the probability of having a similar distance assigned to *i* and the neighboring fragments of *j* is low. As a result, the obtained structures, although having a high degree of variability, tend to penalize structures which pairs of distances have very low probability to occur in the population.

We demonstrate that the ensemble of structures obtained with pTADbit not only are in high agreement with both Hi-C interaction matrices and independent imaging data but they are also closer to represent the large heterogeneity of the cell population. In that respect, we could have further increased the degree of variability observed in the imaging data by increasing the number of Monte Carlo simulations and keeping only the structures that have best satisfied the imposed restraints. However, that would have increased the computation time, which is another of the advantages of pTADbit. Moreover, although having imaging data from only a few regions of the genome, the method is applicable to the rest of the human genome. However, we expect small biases in the distance predictions caused by the scarcity of datasets used for the training of the NNs. Indeed, only imaging data of specific regions of chromosome 21 are available at high-resolution. We anticipate that the release of new high-throughput imaging datasets of different regions will allow us to increase the accuracy of the predictions. Moreover, pTADbit is not restricted to the existing trained NNs and it is prepared to use other future TensorFlow networks trained with more extensive datasets.

In summary, pTADbit is a novel approach for the modelling of chromatin fiber that makes use of imaging information to accelerate the generation of ensembles of 3D structures and to reproduce more accurate models that better capture heterogeneity of the cell population.

Methods

pTADbit architecture

TADbit, a broader pipeline for the analysis of Hi-C data, from the mapping of the sequenced reads to the production of 3D ensemble, now includes pTADbit as part of its python package. Although pTADbit represents a completely different approach in the generation of ensemble of genomic regions, TADbit is the perfect container providing the required tools to produce the Hi-C matrices and analyze the resulting structures. The neural networks used in pTADbit were built using TensorFlow [27] and can be replaced by other TensorFlow models as far as input and output parameters are conserved.

Prediction of distance distributions

A short-range Convolutional Neural Network (CNN) and a long-range Neural Network (NN) were trained to predict the distribution of distances between two chromatin loci from its Hi-C interaction frequency and neighborhood. The short-range CNN was trained to predict distributions in Hi-C matrices at 30Kbp resolution and can be used to determine the structure of the models at the finer scale of TADs and sub-TADs (shorter than 1.5Mb). The long-range NN predicts distributions using Hi-C matrices of 250Kbp and is used to shape the models in genomic ranges that are usually larger than the average size of mammalian TADs (larger than 1.5Mb).

Histogram fitting

The short-range CNN was trained to predict histograms of distances obtained from extensive imaging between genomic loci *i and j* using as input the interaction frequency, the genomic distance and the corresponding Hi-C sub-matrices centered at *ij*. The long-range NN, instead, only used the interaction frequency and the genomic distance as input. Distances were obtained from publicly available high-throughput datasets of labelled regions in individual cells [19, 20].

For each cell, coordinates of the different imaged loci were converted to pairwise distances between loci *i and j*. Next, the distribution of the distances *ij* imaged in the same cell type and genomic region were stacked for 200 bins as “Observed data” (**Fig. 1** second panel). Next, an exponentially modified Gaussian distribution was fit to each histogram (**Fig. 1** third panel). The probability density function of the exponentially modified Gaussian distribution was:

$$f(x;\mu,\sigma,\lambda) = \frac{\lambda}{2} e^{\frac{\lambda}{2}(2\mu+\lambda\sigma^2-2x)} erfc\left(\frac{\mu+\lambda\sigma^2-x}{\sqrt{2}\sigma}\right)$$

where $erfc$ is the complementary error function defined as:

$$erfc(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt$$

Such fitting was used as implemented in the *SciPy* python package (https://scipy.org). The three parameters *K, loc and scale* in the parameterization thus corresponded to having *loc* and *scale* equal to σ and μ, respectively, and *K*=1/($\sigma\lambda$). The objective of this procedure was to represent and be able to reconstruct the histograms using the fewer number of parameters as possible; in our case the triad *K*, *loc* and *scale*, which in turn will be predicted by our trained NNs.

Short-range Convolutional Neural Network

The short-range CNN (**Fig. 7a**) was composed by two inputs: one encoder for the submatrix that was formed by three convolutional and two pooling layers and another encoder for the genomic distance with a fully-connected layer. They both converged to a fully-connected layer with three outputs *K*, *loc* and *scale*. Weights were trained using the Adam optimizer [28] to minimize the mean-squared error (MSE) between the input and the output. Rectified linear unit (ReLU) activation functions were used for hidden layers and softplus functions were used for the outputs to prevent the prediction of negative values. The CNN was trained using as input datasets from Bintu et al. 2018 [19] (**Table 1)**. The list of coordinates of the centers of the imaged 30Kbp segments were obtained from https://github.com/BogdanBintu/ChromatinImaging and used to calculated the *K*, *loc* and *scale* parameters of each pairwise distance as described in the previous section. Additionally, Hi-C matrices of chromosome 21 were obtained from GEO database (GSE63525 [2] and GSE104334[29]), normalized using Vanilla coverage [30] and scaled to the range 0 to 1. Then, for each pair of genomic loci *i and j,* its *K*, *loc* and *scale* parameters were matched to the Hi-C 19x19 submatrix centered in the intersection of *ij.* Together with the genomic distance between loci *i and j,* the Hi-C submatrix was used as input data to train the CNN. In addition, each *K*, *loc* and *scale* parameters was assigned to Hi-C matrices with different sequencing depths to assure that the predictions were not tied to a specific quality of the matrix. To evaluate the performance of the CNN, a K-Fold Cross Validation [31] with 5 folds and 3 repeats was applied and an average mean absolute error of 337.66 with a standard deviation of 18.39 was obtained. The reduced standard deviation with respect to the average indicates that the performance of the CNN did not depend on any specific partitioning of the training and validation sets and was independent of their random selection.

Long-range Neural Network

The long-range NN (**Fig. 7b**) was composed by two fully-connected layers with the interaction frequency as input and three outputs *K*, *loc* and *scale*. Weights were trained using the Adam optimizer to minimize the mean-squared error (MSE) between the input and output. Rectified linear unit (ReLU) activation functions were used for hidden layers and softplus functions were used for the outputs to prevent the prediction of negative values. The NN was trained with the imaging dataset of the p-arm of chromosome 2 from Su et al. 2020 [20] (**Table 1)**. Similarly, the Hi-C matrix was obtained from the GEO database (GSE104334[29]). We assigned the list of coordinates of the centers of the imaged segments to their equivalent 250Kbp bins of the Hi-C matrix and proceed similarly to the short-range CNN but in this case using just the Hi-C interaction frequency and the genomic distance of each pair of loci. Therefore, the long-range neurons are trained to perform a regression without the image recognition explained in the short-range CNN.

Bead-on-a-string models

Each 30Kbp bin (column or row) of the input Hi-C interaction matrix was represented by a spherical particle which size was proportional to the number of nucleotides contained in the DNA fragment. Those particles form a connected chain that mimic the polymeric nature of DNA in what is commonly referred as a bead-on-a-string model.

Assignment of spatial restraints and scoring

To combine short and long-range restraints a multi-scale approach was adopted in the modelling process by building low-resolution models and using their distances to restraint long-range interactions. To produce the ensemble of models at low resolution, the input Hi-C matrix was first downscaled to 250Kbp and used as input for the long-range NN to predict their corresponding *K*, *loc* and *scale* parameters from the input frequency and genomic distance between the interrogated loci. Using the predicted parameters of the exponentially modified Gaussian function, the distribution of distances between two particles *i and j* in the population of models was recreated. Distances were randomly sampled from the distributions and assigned as spatial restraints between *i* and *j* with the following criteria:

- Consecutive particles were always restrained to guarantee the continuity of the polymer chain. The absence of restraints between two consecutive particles might result in the disconnection of the chain, which is incompatible with the polymeric nature of chromatin.
- For non-consecutive particles, 60% of the possible pairs were randomly selected in each individual model and restraint with a distance sampled from the distributions. Selecting different restraint pairs in each structure increased the heterogeneity of the resulting

ensemble. By using a small number of restraints, the computation time of the model was reduced at the same time that the variability of the resulting conformations increased. Indeed, the assignment of all possible restraints in each individual model would favor the introduction of contradictory restraints if, for example, regions were brought together to close distance while neighboring regions were taken apart. It is important to note that restraints that could potentially produce triangle inequalities were discarded during the assignment of restraints in each individual model. That is, for every triad of particles *i*, *j*, *k* the assignment of more than two distance restraints between them was prevented. We found the percentage balance 60/40 between restraint and non-restraint pairs to be a good compromise between computation time and variability of the resulting ensembles.

- The remaining 40% of the non-consecutive pairs were not restrained.

The spatial restraints used between pair of particles were implemented as harmonic oscillators that penalize quadratically deviations from the given equilibrium distance. The mathematical function of the restraint was:

$$UH_{ij} = k_{ij}(d_{ij} - d_0)^2,$$

where d_{ab} is the distance between particle *i* and particle *j* in the model, d_0 is the predicted equilibrium distance sampled from the NN distribution and k_{ab} is the harmonic constant that depends on the *loc* parameter of the predicted distribution as follows:

$$k_{ij} = k(loc_{max} - loc_{ij} + 1)^2,$$

being loc_{max} the maximum *loc* parameter predicted for the model.

The sum of all the imposed harmonic restraints forms an objective function to minimize. To reach that goal, a Monte Carlo simulated annealing sampling approach was used where the conformation of the 3D model was randomly modified and the new configuration accepted or rejected according to the Metropolis criteria [32]. The simulation was repeated with the spheres at different starting random positions, generating each one a single model. The number of structures produced were two times the number initially requested, conserving at the end of the process only the half that best satisfied the imposed restraints. By increasing the number of simulations resulted in structures that were more compatible with the imposed restraints. Considering the low-resolution of these models, increasing the number of requested structures did not compromise the computation time. Finally, the distances between pairs of particles in the low-resolution models were used to build the models at higher resolution.

Next, low-resolution models at 250Kbp were used to produce high-resolution models at 30Kbp by imposing as restraints all distances at 250Kbp for pairs of particles which genomic distance was above the applicability of our short-range NN (that is, 1.5 Mb). Briefly, each 30Kbp particle in the

high-resolution model was assigned to the 250Kbp containing it as distances of the low-resolution models were considered as a good approximation to the distances of the high-resolution structures. Each low-resolution structure was then used in the production of one high-resolution structure, as to maintain the composition of the already optimized low-resolution ensemble. During the building of the high-resolution models, the Hi-C submatrices at 30Kbp resolution were used as input to the short-range CNN to predict the *K*, *loc* and *scale* parameters of the exponentially modified Gaussian functions of each pair of loci *i* and *j*, which genomic distance was below 1.5Mbp. For pairs which genomic distance was larger than 1.5Mbp, instead of sampling from the distributions, the optimized distances obtained in each of the low-resolution models were used as restraints.

Analogously to the low-resolution models, the distribution of distances between pairs of particles which genomic distance was below 1.5Mbp were recreated and randomly sampled to assign the distances as spatial restraints. The assignments were as follows:

- Consecutive particles were always restrained to guarantee the continuity of the polymer chain.
- For non-consecutive particles which genomic distance was below 1.5Mbp, 30% of the possible pairs were randomly selected in each individual model and restraint with a distance sampled from the distributions. We found that the assignment of 30% of the restraints in the short genomic regime was enough to recreate the structural features without compromising the computation times.
- For non-consecutive particles which genomic distance was above 1.5Mbp, 60% of the possible pairs were randomly selected in each individual model and restrained with the distances obtained from its starting low-resolution model.
- The remaining non-consecutive were not restrained.

Finally, the ensemble of conformations was built by minimizing the objective function with a Monte Carlo sampler.

Three-dimensional (3D) modeling

The 3D models in pTADbit were generated by assigning the predicted distances between loci as spatial restraints that were then satisfied using the Integrative Modeling Platform (IMP) [33]. The procedure is in many ways similar to the previously published TADbit [14].

Stratum-adjusted correlation coefficients (SCC)

To calculate the SCC coefficient, first the matrix is smoothed with a 2D mean filter to minimize the effect of noise and biases and second, the Hi-C data is stratified according to their genomic distance.

The smoothing filter is characterized by the span size *h* and the correlation is limited to a maximum number of diagonals starting from the main one. The SCC measures in this manuscript were calculated using a span size *h* of 3. In the case of the full chromosome 19, 4Mbp was used as the maximum distance from the diagonal at which SCC was computed. All the diagonals were taken into account for the rest of the matrices.

Production of the ensemble of models with TADbit, LorDG and Chrom3D

TADbit. The ensemble of 1,000 models generated with the original TADbit algorithm [14] was obtained by performing an optimization step to find optimal values of -0.4 for *lowfreq*, 0 for *upfreq* and 430 nm for the distance *cutoff* using a *scale* of 0.007. With those values, the final ensemble of models was obtained as previously described [14].

Chrom3D. The ensemble of 1,000 models generated with the original Chrom3D algorithm [24] was obtained following the publicly available protocol in https://github.com/Chrom3D using a *cooling-rate* of 0.001. The generation of individual models was parallelized with an in-house python script that assigned a different seed number to each run.

LorDG. The ensemble of 1,000 models generated with the original LorDG algorithm [23] was obtained following the publicly available protocol in https://github.com/BDM-Lab/LorDG. LorDG estimated the $\propto$ parameter in the relation between the interaction frequency and the physical distances to be 0.6.

Table 1. Image datasets used in the different trained CNN.

CNN	Label	Cell line	Region (hg38)	# Cells	# Pairwise distances
Short-range	IMR90_chr21-18-20Mb	IMR90	18,627,714-20,577,518	1,277	2,080
Short-range	IMR90_chr21-28-30Mb	IMR90	28,000,071-29,949,939	4,871	2,080
Short-range	K562_chr21-28-30Mb	K562	28,000,071-29,949,939	13,997	2,080
Short-range	HCT116_chr21-28-30Mb_untreated	HCT116	28,000,071-29,949,939	1,979	2,080
Short-range	HCT116_chr21-34-37Mb_untreated	HCT116	34,628,096-37,117,534	11,631	3,403
Long-range	Chr2-p-arm_replicate	IMR90	1-94,750,001	4,848	61,789

Figures

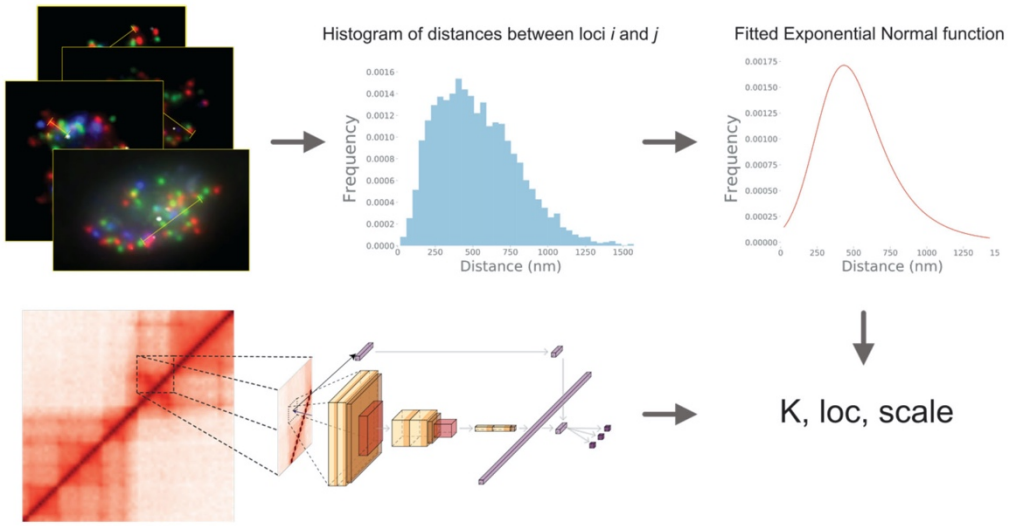


Figure 1. Schema of pTADbit prediction of histograms. Imaging distances are compiled into histograms which are approximated by exponential Gaussian functions. *K*, *loc* and *scale* parameters depicting the functions are predicted by Neural Networks using Hi-C matrices.

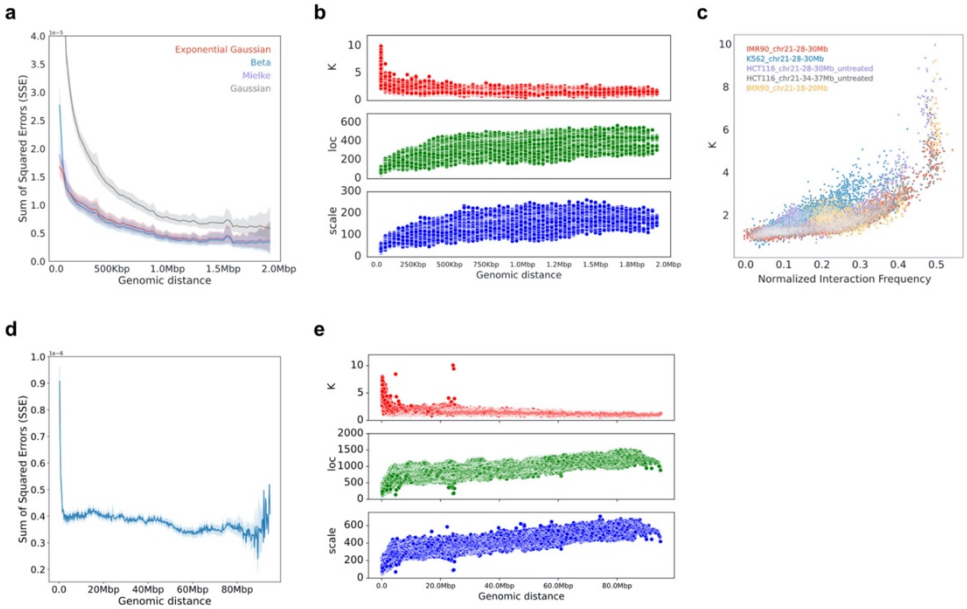


Figure 2. Histogram fitting. a Average Sum of Square Errors (SSE) of the fitting of the histograms of distances between pairs of 30Kbp regions to four mathematical curves as a function of the genomic distance. **b** *K*, *loc* and *scale* values of the fitted histograms of distances between pairs of 30Kbp regions to exponential Gaussian curves as a function of the genomic distance. **c** *K* values of the fitted histograms of distances between pairs of 30Kbp regions to exponential Gaussian curves as a function of the interaction frequency. **d** Average SSE of the fitting of the histograms of distances between pairs of 30Kbp regions to exponential Gaussian curves as a function of the genomic distance. **e** *K*, *loc* and *scale* values of the fitted histograms of distances between pairs of 250Kbp regions to exponential Gaussian curves as a function of the genomic distance.

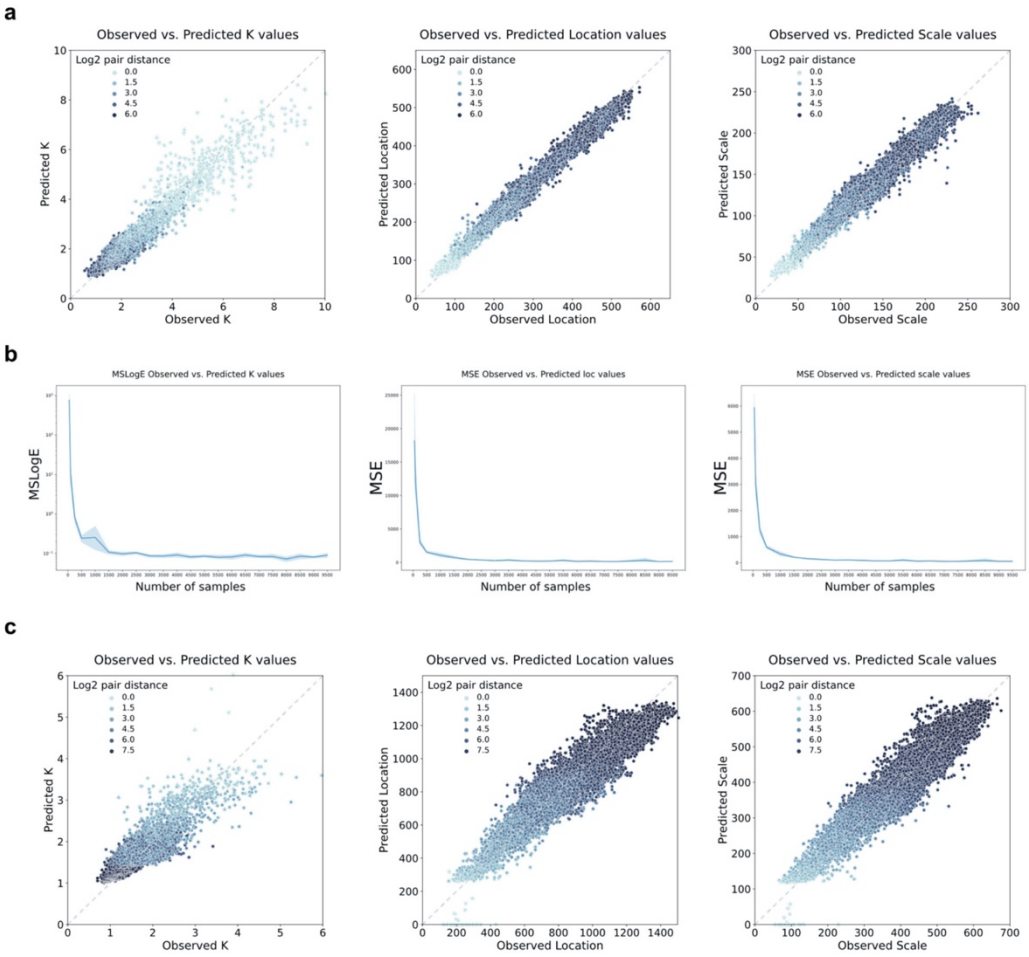


Figure 3. Observed vs. predicted values. a *K*, *loc* and *scale* values for the validation set of the short-range CNN composed by 19,298 predictions. The Pearson correlation coefficients are 0.96, 0.99 and 0.98 for *K, loc* and *scale*, respectively. **b** Mean Square Log Error of the observed vs. predicted *K* values and Mean Square Error of the observed vs. predicted *loc* and *scale* values for the validation set of the short-range CNN using an increasing number of imaged cells. The training of the CNN was repeated 5 times for each number of cells using random training and validation sets. **c** Observed vs. predicted *K, loc* and *scale* values for the validation set of the long-range NN composed by 111,669 predictions. The Pearson correlation coefficients (r) are 0.65, 0.93 and 0.93 for *K, loc* and *scale*, respectively.

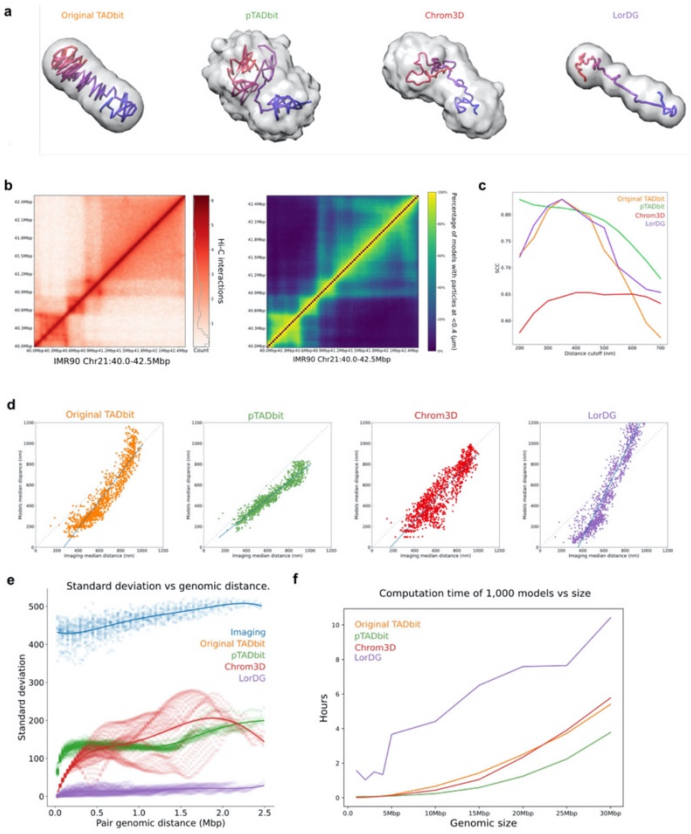


Figure 4. Model benchmarking. a Ensemble of models at 30Kbp of the genomic region 28Mbp-30Mbp in chromosome 21 in IMR90 generated using TADbit (orange), pTADbit (green), Chrom3D (red) and LorDG (purple). The centroid model (the one closer to the average) is depicted by a tubular shape colored from blue (40Mbp) to red (42.5Mbp) and the occupancy of the ensemble is represented by a semi-transparent grey shadow. **b** Normalized Hi-C matrix of the genomic region used to produce the ensembles in **a** (left panel) and contact map of the pTADbit ensemble using a distance cutoff of 400 nm (right panel). **c** Stratum-adjusted Correlation Coefficient (SCC) between the contact map of the ensembles in **a** and the normalized Hi-C matrix in **b** using different distance cutoffs and for all compared methods. **d** Comparison of the pairwise median distances between each 30Kbp probe in the imaged genomic region 40Mbp-42.5Mbp in chromosome 21 in IMR90 (n=7,591 cells) and all loci in the ensemble of 1,000 models of the same region produced by the original TADbit (orange), pTADbit (green), Chrom3D (red), and LorDG (purple). The Pearson correlation coefficients (r) were 0.92 for TADbit, 0.96 for pTADbit, 0.9 for Chrom3D and 0.95 for LorDG. **e** Standard deviation of the distance between pairs of loci depending on its genomic distance in the ensemble of 1,000 models. **f** Computation time to produce an ensemble of 1,000 models of different genomic sizes for the modelled region. Results were obtained using 24 cores in a workstation with an Intel(R) Core (TM) i9-7960X @ 2.80GHz with 128 Gb of RAM.

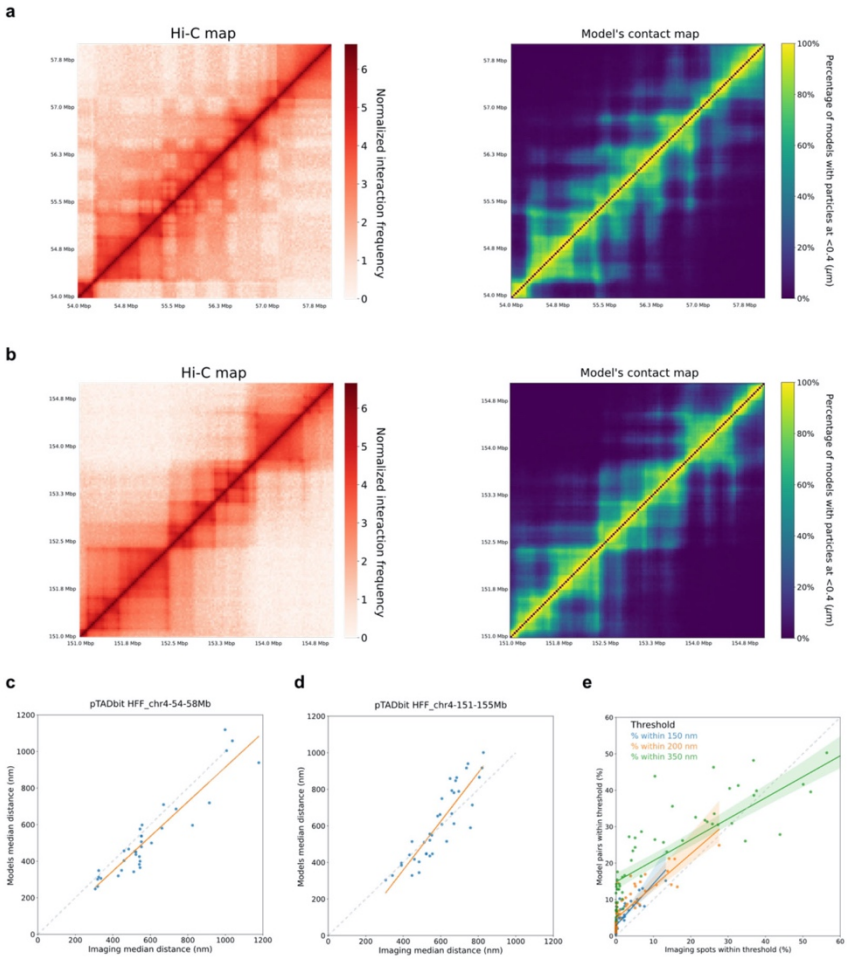


Figure 5. Model accuracy. **a** Hi-C matrix of the genomic region 54Mbp-58Mbp in chromosome 4 in HFF (left) and contact map (right) using a distance cutoff of 400 nm of the ensemble of 1,000 models generated from the matrix in a using pTADbit. **b** Hi-C matrix of the genomic region 151Mbp-155Mbp in chromosome 4 in HFF (left) and contact map (right) using a distance cutoff of 400 nm of the ensemble of 1,000 models generated from the matrix in a using pTADbit. **c** Comparison of the pairwise median distances between each of the 9 probes in the imaged genomic region 54Mbp-58Mbp in chromosome 4 in HFF (n=50,197 distances) and the corresponding loci in the ensemble of 1,000 models of the same region produced by pTADbit. The Pearson correlation coefficient (r) is 0.93. **d** Same as **b** for the genomic region 151Mbp-155Mbp (n=116,047 distances) resulting in a Pearson correlation coefficient (r) of 0.89. **e** Comparison of the percentages of pairs of the 9 imaged probes which distance is within 150, 200 and 350 nm each in the imaged genomic regions 54Mbp-58Mbp and 151Mbp-155Mbp in chromosome 4 in HFF and the corresponding loci in the ensemble of 1,000 models of the same region produced by pTADbit. The Pearson correlation coefficients (r) are 0.83, 0.84 and 0.86 for 150, 200 and 350 nm, respectively.

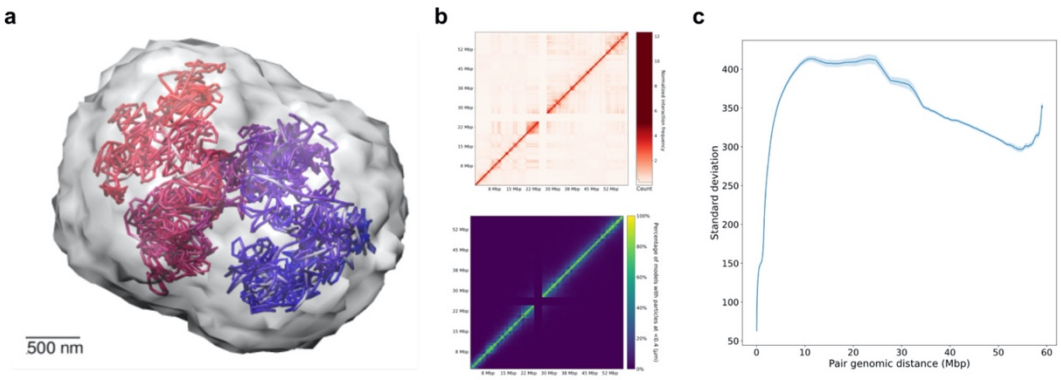


Figure 6. Model accuracy for an entire chromosome. **a** Ensemble of models at 30Kbp of the chromosome 19 in IMR90 generated using pTADbit. The centroid model (the one closer to the average) is depicted by a tubular shape colored from blue to red and the occupancy of the ensemble is represented by a semi-transparent grey shadow. **b** Normalized Hi-C matrix of the chromosome 19 used to produce the ensemble of models (top panel) and contact map of the pTADbit ensemble using a distance cutoff of 400 nm (lower panel). **c** Standard deviation of the distance between pairs of loci depending on its genomic distance in the ensemble of 1,000 models.

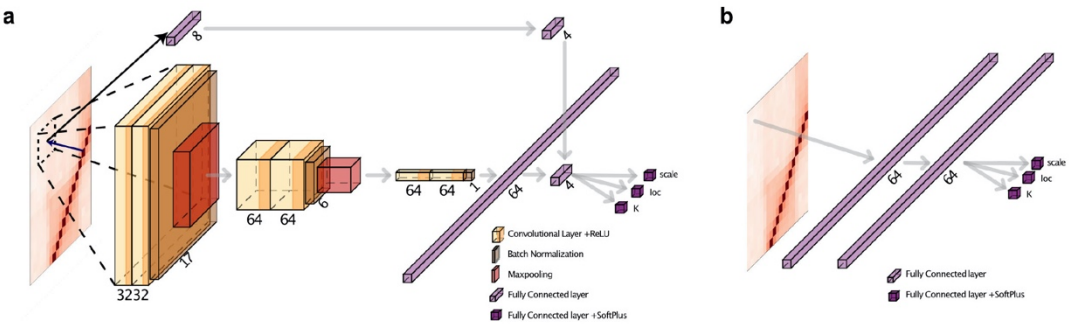


Figure 7. Convolutional Neural Network architectures. **a** Short-range Neural Network architecture. The 19x19 pixels input sub-matrices pass through the initial convolutional layers that performs a hierarchical decomposition of the information allowing the CNN to learn a wide range of features, from the very local to the more global ones. The subsequent fully connected layers learn non-linear combinations of the extracted features, combine them with the genomic distance of the center of the sub-matrix and does a linear regression to estimate *K*, *loc* and *scale*. **b** Long-range Neural Network architecture. The long-range NN is a simplified version of the short-range where the convolutional layers are removed and the input consists only in the interaction frequency and the genomic distance.

Acknowledgments

MAM-R acknowledges support by the Spanish Ministerio de Ciencia e Innovación (PID2020-115696RB-I00) and the National Human Genome Research Institute of the National Institutes of Health under Award Number RM1HG011016. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. CRG acknowledges support from 'Centro de Excelencia Severo Ochoa 2013-2017', SEV-2012-0208 and the CERCA Programme/ Generalitat de Catalunya as well as support of the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III and the EMBL partnership, the Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement, and the Co-financing with funds from the European Regional Development Fund (ERDF) by the Spanish Ministry of Science and Innovation corresponding to the Programa Operativo FEDER Plurirregional de España (POPE) 2014-2020 and by the Secretaria d'Universitats i Recerca, Departament d'Empresa i Coneixement of the Generalitat de Catalunya corresponding to the programa Operatiu FEDER Catalunya 2014-2020.

References

1. Hsieh TS, Cattoglio C, Slobodyanyuk E, Hansen AS, Rando OJ, Tjian R, et al. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Mol Cell. 2020. doi: 10.1016/j.molcel.2020.03.002.
2. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-80. doi: 10.1016/j.cell.2014.11.021.
3. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell. 2017;171(3):557-72 e24. doi: 10.1016/j.cell.2017.09.043. PubMed Central PMCID: PMCPMC5651218.
4. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295(5558):1306-11. Epub 2002/02/16. doi: 10.1126/science.1067799 295/5558/1306 [pii].
5. Mendieta-Esteban J, Di Stefano M, Castillo D, Farabella I, Marti-Renom MA. 3D reconstruction of genomic regions from sparse interaction data. NAR Genom Bioinform. 2021;3(1):lqab017. doi: 10.1093/nargab/lqab017. PubMed Central PMCID: PMCPMC7985034.
6. Ramani V, Deng X, Qiu R, Lee C, Disteche CM, Noble WS, et al. Sci-Hi-C: A single-cell Hi-C method for mapping 3D genome organization in large number of single cells. Methods. 2020;170:61-8. doi: 10.1016/j.ymeth.2019.09.012. PubMed Central PMCID: PMCPMC6949367.
7. Boninsegna L, Yildirim A, Polles G, Zhan Y, Quinodoz SA, Finn EH, et al. Integrative genome modeling platform reveals essentiality of rare contact events in 3D genome organizations. Nat Methods. 2022;19(8):938-49. doi: 10.1038/s41592-022-01527-x. PubMed Central PMCID: PMCPMC9349046.
8. Yildirim A, Boninsegna L, Zhan Y, Alber F. Uncovering the Principles of Genome Folding by 3D Chromatin Modeling. Cold Spring Harb Perspect Biol. 2022;14(6). doi: 10.1101/cshperspect.a039693. PubMed Central PMCID: PMCPMC9248826.
9. Oluwadare O, Highsmith M, Cheng J. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. Biol Proced Online. 2019;21:7. doi: 10.1186/s12575-019-0094-0. PubMed Central PMCID: PMCPMC6482566.
10. Boninsegna L, Yildirim A, Zhan Y, Alber F. Integrative approaches in genome structure analysis. Structure. 2022;30(1):24-36. doi: 10.1016/j.str.2021.12.003. PubMed Central PMCID: PMCPMC8959402.
11. Reiff SB, Schroeder AJ, Kirli K, Cosolo A, Bakker C, Lee S, et al. The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. Nat Commun.

2022;13(1):2365. doi: 10.1038/s41467-022-29697-4. PubMed Central PMCID: PMCPMC9061818.

12. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013;14(6):390-403. doi: 10.1038/nrg3454. PubMed Central PMCID: PMCPMC3874835.
13. Serra F, Di Stefano M, Spill YG, Cuartero Y, Goodstadt M, Bau D, et al. Restraint-based three-dimensional modeling of genomes and genomic domains. FEBS Lett. 2015;589(20 Pt A):2987-95. doi: 10.1016/j.febslet.2015.05.012.
14. Serra F, Bau D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput Biol. 2017;13(7):e1005665. doi: 10.1371/journal.pcbi.1005665. PubMed Central PMCID: PMCPMC5540598.
15. Farabella I, Di Stefano M, Soler-Vila P, Marti-Marimon M, Marti-Renom MA. Three-dimensional genome organization via triplex-forming RNAs. Nat Struct Mol Biol. 2021;28(11):945-54. doi: 10.1038/s41594-021-00678-3.
16. Stik G, Vidal E, Barrero M, Cuartero S, Vila-Casadesus M, Mendieta-Esteban J, et al. CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response. Nat Genet. 2020;52(7):655-61. doi: 10.1038/s41588-020-0643-0.
17. Miguel-Escalada I, Bonas-Guarch S, Cebola I, Ponsa-Cobas J, Mendieta-Esteban J, Atla G, et al. Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. Nat Genet. 2019;51(7):1137-48. doi: 10.1038/s41588-019-0457-0. PubMed Central PMCID: PMCPMC6640048.
18. Le Dily F, Bau D, Pohl A, Vicent GP, Serra F, Soronellas D, et al. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. 2014;28(19):2151-62. doi: 10.1101/gad.241422.114. PubMed Central PMCID: PMCPMC4180976.
19. Bintu B, Mateo LJ, Su JH, Sinnott-Armstrong NA, Parker M, Kinrot S, et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science. 2018;362(6413). doi: 10.1126/science.aau1783. PubMed Central PMCID: PMCPMC6535145.
20. Su JH, Zheng P, Kinrot SS, Bintu B, Zhuang X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. Cell. 2020;182(6):1641-59 e26. doi: 10.1016/j.cell.2020.07.032. PubMed Central PMCID: PMCPMC7851072.
21. Nguyen HQ, Chattoraj S, Castillo D, Nguyen SC, Nir G, Lioutas A, et al. 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. Nat Methods. 2020;17(8):822-32. doi: 10.1038/s41592-020-0890-0. PubMed Central PMCID: PMCPMC7537785.

22. Diaz N, Kruse K, Erdmann T, Staiger AM, Ott G, Lenz G, et al. Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. Nat Commun. 2018;9(1):4938. doi: 10.1038/s41467-018-06961-0. PubMed Central PMCID: PMCPMC6265268.
23. Trieu T, Cheng J. MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. Bioinformatics. 2016;32(9):1286-92. doi: 10.1093/bioinformatics/btv754.
24. Paulsen J, Sekelja M, Oldenburg AR, Barateau A, Briand N, Delbarre E, et al. Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. Genome Biol. 2017;18(1):21. doi: 10.1186/s13059-016-1146-2. PubMed Central PMCID: PMCPMC5278575.
25. Yang T, Zhang F, Yardimci GG, Song F, Hardison RC, Noble WS, et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res. 2017;27(11):1939-49. doi: 10.1101/gr.220640.117. PubMed Central PMCID: PMCPMC5668950.
26. Finn EH, Pegoraro G, Brandao HB, Valton AL, Oomen ME, Dekker J, et al. Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. Cell. 2019;176(6):1502-15 e10. doi: 10.1016/j.cell.2019.01.020. PubMed Central PMCID: PMCPMC6408223.
27. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation; Savannah, GA, USA: USENIX Association; 2016. p. 265–83.
28. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv. 2014;1412.6980v9.
29. Rao SSP, Huang SC, Glenn St Hilaire B, Engreitz JM, Perez EM, Kieffer-Kwon KR, et al. Cohesin Loss Eliminates All Loop Domains. Cell. 2017;171(2):305-20 e24. doi: 10.1016/j.cell.2017.09.026. PubMed Central PMCID: PMCPMC5846482.
30. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012;9(10):999-1003. Epub 2012/09/04. doi: 10.1038/nmeth.2148. PubMed Central PMCID: PMC3816492.
31. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2; Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.; 1995. p. 1137–43.
32. Metropolis N, Ulam S. The Monte Carlo method. J Am Stat Assoc. 1949;44(247):335-41. Epub 1949/09/01.
33. Russel D, Lasker K, Webb B, Velazquez-Muriel J, Tjioe E, Schneidman-Duhovny D, et al. Putting the pieces together: integrative modeling platform software for structure

determination of macromolecular assemblies. PLoS Biol. 2012;10(1):e1001244. Epub 2012/01/25. doi: 10.1371/journal.pbio.1001244. PBIOLOGY-D-11-02156 [pii]. PubMed Central PMCID: PMC3260315.

Discussion

Until the development of 3C techniques, imaging methods were the main approach used to study the structure of the genome. The use of imaging, predominantly FISH, has allowed the discovery of many important features of genome conformation, like the existence of chromosome territories (Cremer et al. 2001) or the striking differences in nuclear position and topology between some similarly sized chromosomes (Croft et al. 1999). Those used imaging techniques were characterized by their low-throughput nature and its inability to evaluate high numbers of simultaneous loci in each individual cell. Then, the emergence of 3C technology revolutionized the field of structural genomics by bringing unprecedent resolutions and providing the necessary sample size to apply the power of statistics to drive the scientific conclusions (S.S. Rao et al. 2014). Imaging was then somewhat relegated, thanks to the orthogonality of the information provided, to a validation method for 3C predictions. But the last innovations in microscopy, mostly in the massive development of FISH probes, have given a new impulse to imaging technologies. Indeed, the advances in the massive synthesis of custom and complex oligonucleotides, led by Oligopaints (Beliveau et al. 2012), has opened a new era of multi-targeted and high-throughput oligo-based microscopy. OligoFISSEQ, introduced in Chapter I of this thesis, is part of this new age of oligo-based technologies.

Despite the unquestionable value of the provided information, 3C technology introduce its inherent biases to the data and do not provide a direct measure of the physical distances in the experiment. The analysis of cross-linked data is further challenged by the inclusion of biases that are specific to the 3C method used. In Chapter II we provide a modelling strategy that corrects those biases in the case of Promoter Capture Hi-C (pcHi-C).

This thesis project contributes to the provision of analysis tools and pipelines for both imaging and 3C data and concludes with the development of pTADbit in Chapter III, a tool that combines high-throughput imaging data and Hi-C.

3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing

In Chapter I of this thesis, we designed and developed the decoding pipeline of the oligoFISSEQ technique. In OligoFISSEQ images, the signal is entangled between the different channels and sequencing rounds. The encoding of the information used by the technique allows the imaging of large numbers of loci in individual cells in a reduced

number of sequencing rounds. The multiplexing of the information, however, implies the use of specialized algorithms to decode the signal.

We demonstrated the potential of OligoFISSEQ with the design and use of two libraries: one targeting six regions along each of six human chromosomes: 2, 3, 5, 16, 19 and X (36plex) and the other labelling 46 regions along the human X chromosome (ChrX-46plex). The use of the 36plex library in male PGP1f cells allowed us to validate the technique by reproducing well-known structural features like chromosome territories or the tendency of smaller chromosomes to be positioned toward the center and larger chromosomes towards the periphery of the nucleus. It also allowed us to show the potential application of OligoFISSEQ to study the genome structure with enough number of samples to reach statistical significance. We next showed the power of OligoFISSEQ to trace chromosomes with finer genomic resolution by applying the ChrX-46plex library to male PGP1f cells. We were able to generate 176 traces spanning the entire X chromosome permitting both single-cell as well as population-based analysis. The clustering of the traces revealed two major groups that differed significantly in their radii of gyration, one cluster consisting of 20 chromosomes (11%) and the other comprising 156 (89%).

To exhibit the potential use of OligoFISSEQ with other labelling technologies, we applied the ChrX-46plex library in female IMR90 cells in conjunction with immunofluorescence (IF) to macroH2A.1 which preferentially binds the inactive X chromosome. Using OligoFISSEQ we traced and analyzed separately the active and inactive X chromosomes and were able to validate known features like their difference in occupied volume and the separation in two megadomains of the inactive X chromosome.

Finally, we demonstrated the capacity of OligoFISSEQ to accelerate super-resolution imaging. Combining OligoSTORM with four rounds of OligoFISSEQ we achieve a 36-fold reduction in imaging time.

OligoFISSEQ is one of the multiplexing imaging technologies in the undeclared race to image the full genome providing as high-resolution as possible. Bintu et al 2018 (Bintu et al. 2018) is probably the first manuscript in which acquired imaging data enabled enough resolution and throughput to match 3C-based data in the 2Mbp targeted regions.

Not only that, up to date no multiplexing imaging dataset provides thousands of different cells in a resolution as high as 30kbp. Their findings demonstrate that TADs are structures present in single cells and not an emergent property of population averaging. The main difference between OligoFISSEQ protocol and their multiplexed approach is the entanglement of the signal in different wavelengths used by the former. That allows OligoFISSEQ to image similar number of targets in fewer number of rounds. The sequel manuscript from the same research group Su et al 2020 (Su et al. 2020) further improves their multiplexing strategy using three different imaging wavelenghts and more than 200 sequencing rounds to image almost the entirety of chromosome 21 at 50kbp resolution. Such an increase of the number of targets using the original protocol would have boosted dramatically the cost of the experiments because the number of imaged loci grows only linearly with the number of sequencing rounds. The cost factor is one of the advantages of OligoFISSEQ against other technologies due to its reduced number of sequencing rounds. Scaling to longer regions allowed them to better characterize certain domain properties and demonstrate the tendency of loci in A and B compartments to spatially segregate, although often incompletely, in single cells. In the same manuscript, they accomplished the imaging of the full human genome in thousands of different cells using a variation of MERFISH (K.H. Chen et al. 2015) tailored to DNA which translated in a 10-fold reduction in the number of rounds compare to the sequential approach. Although the low resolution attained, the visualization of the full human genome is an astonishing breakthrough in the field. Full genome imaging provides an unprecedent vision of the genome organization that allowed them the study of the chromosome overlapping that suggests the existence of substantial trans-chromosomal interactions.

SeqFISH+ (Takei et al. 2021) reaches also the full genome scale imaging mouse embryonic stem cell nuclei at 1Mbp resolution using two microscope channels. Simultaneously, they imaged several 1.5Mbp regions at 25kbp resolution using a third one. In total they reach 3,660 different targets using 80 rounds of sequential hybridization for the 1Mbp localizations and another 80 rounds for the 25kbp detailed regions, together with 17 chromatin marks and subnuclear structures and the expression profile of 70 RNAs. With this strategy they observed that many DNA loci, especially active gene loci, reside at the surface of nuclear bodies and zone interface.

Another recent addition to new multiplexing imaging technologies is *in situ* genome sequencing (IGS) (Payne et al. 2021) which uses a different approach to reach full genome visualization by combining *in situ* and *ex situ* sequencing. The use of *ex situ*

sequencing allows the identification of the imaged loci with the precision required to distinguish different alleles and spot their structural differences. However, their un-targeted approach makes the method un-suitable for the examination of the structure of specific loci compared to methods like OligoFISSEQ in which the targets are bioinformatically designed. Indeed, the use of *in situ* transposition to incorporate the DNA-sequencing adapters randomly in the genome prevents the selection of a set of specific loci.

All the mentioned technologies, including OligoFISSEQ, rely on customized detection and analysis pipelines that combine image analysis and statistical approaches. Few initiatives exist that englobe the needed tools to process all these different imaging techniques. Maybe the closest to such platform would be Starfish (Axelrod et al. 2021), a Python library that contains the basic functions for the analysis of spatial genomics.

3D reconstruction of genomic regions from sparse interaction data

In Chapter II of this thesis we presented a novel tool for the 3D reconstruction and analysis of chromatin regions from the sparse interaction data obtained with 3C-based experiments. These types of assays are conceived to capture interactions between specific regions and the rest of the genome, for example, *loci* enriched in a specific protein (Mumbach et al. 2016) or regions known to contain gene promoters (Schoenfelder et al. 2015). Capturing specific interactions allows the production of high-resolution interacting profiles. Contrary to the output of experiments providing continuous matrices of pair-wise interactions, these profiles when represented as matrices are characterized by data sparsity because the large majority of the cells in the interaction matrix belong to non-captured fragments and, as such, are empty. Additionally, they are heavily biased on interactions between captured fragments. Most of the computational tools for the analysis of 3C experiments are designed to handle dense interaction matrices and are ill-suited for the sparsity of the capture experiment's profiles, for instance in the reconstruction of 3D structural models.

In the case of pcHi-C, the existing tools (ChiCMaxima (Ben Zouari et al. 2019), Chicago (Cairns et al. 2016), …) are mainly focused on the implementation of normalization strategies to reduce the impact of biases and on the assessment of the significance of interactions between captured loci. Contrarily, the work presented in this thesis allows

the analysis and interpretation of pcHi-C assays by producing ensembles of three-dimensional structures that are compatible with the experimental data. This methodology covers the normalization of the data, the detection of significative interactions and finally the recovery of the full structural organization of a genomic region in study.

In our approach, a genomic region is reconstructed using a restraint molecular dynamics approach with TADdyn (Di Stefano et al. 2020). The polymer is constituted by spherical beads of 50 nm of diameter each containing 5kb of chromatin fiber and is subjected to the potential energy composed by the chain stiffness, connectivity and excluded volume. The known interaction frequencies are converted to spatial restraints that are imposed progressively to the interacting beads using a steered molecular dynamics protocol.

Using the explained methodology, we modelled 12 genomic regions from Promoter Capture Hi-C (pcHi-C) data and compared the ensemble of structures with their equivalent ensembles reconstructed from Hi-C. Additionally, to quantify the effect of sparsity in the comparison, we reconstructed ensembles from virtual capture interaction matrices (pcHi-Cvirt) built by selecting from normalized Hi-C matrices the rows and columns of the regions captured in the experiments.

The comparison between the sparse and dense derived 3D model ensembles revealed that it is possible to recover most of the 3D organization of the dense dataset in spite of the data sparsity. Indeed, we obtained high median distance correlation between the sparse and dense derived 3D model ensembles for both pcHi-C and pcHi-Cvirt. In summary, these results indicate that the sparse derived ensembles of 3D models are a good representation of the dense experiments.

Probabilistic 3D-modelling of genomes and genomic domains by integrating high-throughput imaging and Hi-C using machine learning

In Chapter III of this thesis, we introduced pTADbit which combines Hi-C interaction data with high-throughput microscopy information to produce ensembles of 3D structures that reproduce more reliably the heterogeneity of the cell population. In this regard, pTADbit lays the foundations for the application of Artificial Intelligence in the modelling of chromatin from 3C data.

We demonstrated that pTADbit produces 3D ensembles that are in high agreement with both Hi-C interaction matrices and independent imaging data with inferred structures that are closer to represent the large heterogeneity observed in imaging experiments. Moreover, pTADbit compares favorably against other modelling frameworks while drastically reducing the required computation time.

Although having imaging data from only a few regions of the genome, the method is applicable, not surprisingly, to the rest of the human genome. However, we cannot discard the presence of small biases in the distance predictions caused by the scarcity of datasets used for the training of the NNs. We anticipate an increase in the accuracy of the predictions with the release of new high-throughput imaging datasets of different regions. Additionally, pTADbit is not restricted to the existing trained NNs and it is prepared to use other future Tensorflow networks trained with more extensive datasets.

The use of distances from imaging experiments in neural networks (NNs) allows pTADbit to overcome one of the major difficulties in the modelling of genomic regions: the inference of the equivalence between interaction frequencies and distances. The direct transformation from Hi-C data to distances is one the strengths of pTADbit compared to existing methods.

Among the innumerable tools for the three-dimensional modelling of chromatin, it is difficult to find tools that combine high-throughput imaging data and Hi-C. The main reason is probably that imaging datasets are scarce and quite recent. A tool integrating Hi-C and FISH data to obtain more accurate three-dimensional models is GEM-FISH (Abbas et al. 2019). GEM-FISH allows the reconstruction of 3D models of chromosomes integrating Hi-C and FISH data and prior biophysical knowledge of polymer physics. As pTADbit, GEM does not rely on any specific conversion between the Hi-C contact frequencies and the corresponding spatial distances, but directly encodes the proximity of cross-linked data and FISH distances as spatial restraints. It also uses a divide-and-conquer approach similar to pTADbit in which a TAD-level lower-resolution structure is first computed and finally integrated with higher-resolution conformations of each TAD to complete the final 3D model of the chromosome.

As for the use of Artificial Intelligence (AI) to 3C data, we find several examples where deep learning is used to predict contact frequencies using as input different chromatin features. For example, in DeepC (Schwessinger et al. 2020) or Akita (Geoff Fudenberg et al. 2020) contact frequencies are predicted using the DNA sequence and different types of neural networks. These applications of AI in the field of chromatin folding have the objective of predicting Hi-C data while pTADbit focuses on the reconstruction of the 3D structures from it.

To the best of our knowledge, no computational approach has been proposed previously to predict interacting chromatin distances through AI using Hi-C data.

Conclusion

The two main approaches to study the structure of the genome are 3C technologies and imaging. Although many tools exist for the analysis of 3C-based data, the innumerable variations and adaptations of the technology makes still essential the provision of new computational libraries. On the one hand, as part of this thesis, we developed a pipeline for the reconstruction and analysis of Promoter Capture Hi-C (pcHi-C) a 3C technique that generates sparse data. On the other hand, during the last years, imaging has experienced a major revolution with the appearance of new high-throughput multiplexing techniques capable of reaching unprecedent resolutions in single cells. In this thesis we developed OligoFISSEQ, one of those emerging technologies and provide tools for the analysis of its generated information. Finally, we developed a tool that combines 3C-based and high-throughput imaging data to bring together information of both worlds to enhance the three-dimensional reconstruction of genomic regions.

From Chapter I, we can specifically conclude:

1. We designed and implemented OligoFISSEQ a novel imaging technology to visualize multiple genomic regions in hundreds and thousands of individual cells.
2. We demonstrated the capacity of OligoFISSEQ to study chromatin organization in both individual cells and populations.
3. Thanks to its multiplexing strength, OligoFISSEQ has the potential to scale the labelled regions in individual cells to hundreds or even thousands of *loci*.
4. We showed the capacity of OligoFISSEQ to accelerate the rate at which multiple genomic regions can be visualized in super-resolution images.

From Chapter II, we can specifically conclude:

1. We developed a bioinformatics tool for the reconstruction of the 3D organization of chromatin from sparse pcHi-C datasets.
2. The structures reconstructed with our methodology are highly similar to those obtained with benchmarked tools using dense datasets.
3. The designed methodology can be easily adapted to other sparse 3C-based data sets.

From Chapter III, we can specifically conclude:

1. We developed pTADbit, a novel approach for the reconstruction of the 3D organization of chromatin from Hi-C data that makes use of imaging information and Machine Learning (ML).
2. pTADbit produce ensembles of 3D models that reproduce more accurately the heterogeneity of the cell population.
3. The inference of the equivalence between interaction frequencies and distances from high-throughput imaging experiments allows pTADbit to dramatically reduce the computation times in the generation of 3D structures.

Bibliography

Abbas, Ahmed, Xuan He, Jing Niu, Bin Zhou, Guangxiang Zhu, Tszshan Ma, . . . Jianyang Zeng. 2019. "Integrating Hi-C and FISH data for modeling of the 3D organization of chromosomes." *Nature Communications* 10 (1): 2049. https://doi.org/10.1038/s41467-019-10005-6. https://doi.org/10.1038/s41467-019-10005-6.

Alberti, S., A. Gladfelter, and T. Mittag. 2019. "Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates." *Cell* 176 (3): 419-434. https://doi.org/10.1016/j.cell.2018.12.035.

Axelrod, Shannon, Matthew Cai, Ambrose Carr, Jeremy Freeman, Deep Ganguli, Justin Kiggins, . . . Kevin Yamauchi. 2021. "starfish: scalable pipelines for image-based transcriptomics." *Journal of Open Source Software* 6: 2440. https://doi.org/10.21105/joss.02440.

Bai, L., and A. V. Morozov. 2010. "Gene regulation by nucleosome positioning." *Trends Genet* 26 (11): 476-83. https://doi.org/10.1016/j.tig.2010.08.003.

Banani, S. F., H. O. Lee, A. A. Hyman, and M. K. Rosen. 2017. "Biomolecular condensates: organizers of cellular biochemistry." *Nat Rev Mol Cell Biol* 18 (5): 285-298. https://doi.org/10.1038/nrm.2017.7.

Barbieri, Mariano, Mita Chotalia, James Fraser, Liron-Mark Lavitas, Josée Dostie, Ana Pombo, and Mario Nicodemi. 2012. "Complexity of chromatin folding is captured by the strings and binders switch model." *Proceedings of the National Academy of Sciences* 109 (40): 16173-16178. https://doi.org/doi:10.1073/pnas.1204799109. https://www.pnas.org/doi/abs/10.1073/pnas.1204799109.

Beliveau, Brian J., Eric F. Joyce, Nicholas Apostolopoulos, Feyza Yilmaz, Chamith Y. Fonseka, Ruth B. McCole, . . . Chao-ting Wu. 2012. "Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes." *Proceedings of the National Academy of Sciences* 109 (52): 21301-21306. https://doi.org/10.1073/pnas.1213818110. https://www.pnas.org/content/pnas/109/52/21301.full.pdf.

Ben Zouari, Yousra, Anne M. Molitor, Natalia Sikorska, Vera Pancaldi, and Tom Sexton. 2019. "ChiCMaxima: a robust and simple pipeline for detection and visualization of chromatin looping in Capture Hi-C." *Genome Biology* 20 (1): 102. https://doi.org/10.1186/s13059-019-1706-3. https://doi.org/10.1186/s13059-019-1706-3.

Bintu, Bogdan, Leslie J. Mateo, Jun-Han Su, Nicholas A. Sinnott-Armstrong, Mirae Parker, Seon Kinrot, . . . Xiaowei Zhuang. 2018. "Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells." *Science* 362 (6413): eaau1783. https://doi.org/doi:10.1126/science.aau1783. https://www.science.org/doi/abs/10.1126/science.aau1783

Boija, A., I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, . . . R. A. Young. 2018. "Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains." *Cell* 175 (7): 1842-1855.e16. https://doi.org/10.1016/j.cell.2018.10.042.

Bonev, B., N. Mendelson Cohen, Q. Szabo, L. Fritsch, G. L. Papadopoulos, Y. Lubling, . . . G. Cavalli. 2017. "Multiscale 3D Genome Rewiring during Mouse Neural Development." *Cell* 171 (3): 557-572.e24. https://doi.org/10.1016/j.cell.2017.09.043.

Branco, M. R., and A. Pombo. 2006. "Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations." *PLoS Biol* 4 (5): e138. https://doi.org/10.1371/journal.pbio.0040138.

Buchwalter, A., J. M. Kaneshiro, and M. W. Hetzer. 2019. "Coaching from the sidelines: the nuclear periphery in genome regulation." *Nat Rev Genet* 20 (1): 39-50. https://doi.org/10.1038/s41576-018-0063-5.

Cairns, Jonathan, Paula Freire-Pritchett, Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, . . . Mikhail Spivakov. 2016. "CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data." *Genome Biology* 17 (1): 127. https://doi.org/10.1186/s13059-016-0992-2. https://doi.org/10.1186/s13059-016-0992-2.

Cardozo Gizzi, Andrés M., Diego I. Cattoni, Jean-Bernard Fiche, Sergio M. Espinola, Julian Gurgo, Olivier Messina, . . . Marcelo Nollmann. 2019. "Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms." *Molecular Cell* 74 (1): 212-222.e5. https://doi.org/https://doi.org/10.1016/j.molcel.2019.01.011. https://www.sciencedirect.com/science/article/pii/S1097276519300115.

Chen, K. H., A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. 2015. "RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells." *Science* 348 (6233): aaa6090. https://doi.org/10.1126/science.aaa6090.

Chen, X., Y. Shen, W. Draper, J. D. Buenrostro, U. Litzenburger, S. W. Cho, . . . H. Y. Chang. 2016. "ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing." *Nat Methods* 13 (12): 1013-1020. https://doi.org/10.1038/nmeth.4031.

Conte, Mattia, Luca Fiorillo, Simona Bianco, Andrea M. Chiariello, Andrea Esposito, and Mario Nicodemi. 2020. "Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation." *Nature Communications* 11 (1): 3289. https://doi.org/10.1038/s41467-020-17141-4. https://doi.org/10.1038/s41467-020-17141-4.

Conte, Mattia, Ehsan Irani, Andrea M. Chiariello, Alex Abraham, Simona Bianco, Andrea Esposito, and Mario Nicodemi. 2021. "Loop-extrusion and polymer phase-separation can co-exist at the single-molecule level to shape chromatin folding." *bioRxiv*: 2021.11.02.466589. https://doi.org/10.1101/2021.11.02.466589. https://www.biorxiv.org/content/biorxiv/early/2021/11/02/2021.11.02.466589.full.pdf.

Cremer, T., and C. Cremer. 2001. "Chromosome territories, nuclear architecture and gene regulation in mammalian cells." *Nat Rev Genet* 2 (4): 292-301. https://doi.org/10.1038/35066075.

Croft, J. A., J. M. Bridger, S. Boyle, P. Perry, P. Teague, and W. A. Bickmore. 1999. "Differences in the localization and morphology of chromosomes in the human nucleus." *J Cell Biol* 145 (6): 1119-31. https://doi.org/10.1083/jcb.145.6.1119.

Davies, James O. J., A. Marieke Oudelaar, Douglas R. Higgs, and Jim R. Hughes. 2017. "How best to identify chromosomal interactions: a comparison of approaches." *Nature Methods* 14 (2): 125-134. https://doi.org/10.1038/nmeth.4146. https://doi.org/10.1038/nmeth.4146.

Dekker, J., K. Rippe, M. Dekker, and N. Kleckner. 2002. "Capturing chromosome conformation." *Science* 295 (5558): 1306-11. https://doi.org/10.1126/science.1067799.

Di Stefano, Marco, Ralph Stadhouders, Irene Farabella, David Castillo, François Serra, Thomas Graf, and Marc A. Marti-Renom. 2020. "Transcriptional activation during cell reprogramming correlates with the formation of 3D open chromatin hubs." *Nature Communications* 11 (1): 2564. https://doi.org/10.1038/s41467-020-16396-1. https://doi.org/10.1038/s41467-020-16396-1.

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, . . . B. Ren. 2012. "Topological domains in mammalian genomes identified by analysis of chromatin interactions." *Nature* 485 (7398): 376-80. https://doi.org/10.1038/nature11082.

Dostie, J., T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, . . . J. Dekker. 2006. "Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements." *Genome Res* 16 (10): 1299-309. https://doi.org/10.1101/gr.5571506.

Fudenberg, G., M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny. 2016. "Formation of Chromosomal Domains by Loop Extrusion." *Cell Rep* 15 (9): 2038-49. https://doi.org/10.1016/j.celrep.2016.04.085.

Fudenberg, Geoff, David R. Kelley, and Katherine S. Pollard. 2020. "Predicting 3D genome folding from DNA sequence with Akita." *Nature Methods* 17 (11): 1111-1117. https://doi.org/10.1038/s41592-020-0958-x. https://doi.org/10.1038/s41592-020-0958-x.

Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, . . . Y. Ruan. 2009. "An oestrogen-receptor-alpha-bound human chromatin interactome." *Nature* 462 (7269): 58-64. https://doi.org/10.1038/nature08497.

Gassler, J., H. B. Brandão, M. Imakaev, I. M. Flyamer, S. Ladstätter, W. A. Bickmore, . . . K. Tachibana. 2017. "A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture." *Embo j* 36 (24): 3600-3618. https://doi.org/10.15252/embj.201798083.

Grosberg, A. Y., Yitzhak Rabin, Shlomo Havlin, and A. Neer. 1993. "Crumpled Globule Model of the Three-Dimensional Structure of DNA." *Europhysics Letters (epl)* 23: 373-378. https://doi.org/10.1209/0295-5075/23/5/012.

Guelen, Lars, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B. Faza, Wendy Talhout, . . . Bas van Steensel. 2008. "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions." *Nature* 453 (7197): 948-951. https://doi.org/10.1038/nature06947. https://doi.org/10.1038/nature06947.

Guo, Y. E., J. C. Manteiga, J. E. Henninger, B. R. Sabari, A. Dall'Agnese, N. M. Hannett, . . . R. A. Young. 2019. "Pol II phosphorylation regulates a switch between transcriptional and splicing condensates." *Nature* 572 (7770): 543-548. https://doi.org/10.1038/s41586-019-1464-0.

Hsieh, T. H., A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando. 2015. "Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C." *Cell* 162 (1): 108-19. https://doi.org/10.1016/j.cell.2015.05.048.

Hu, M., K. Deng, Z. Qin, J. Dixon, S. Selvaraj, J. Fang, . . . J. S. Liu. 2013. "Bayesian inference of spatial organizations of chromosomes." *PLoS Comput Biol* 9 (1): e1002893. https://doi.org/10.1371/journal.pcbi.1002893.

Hughes, J. R., N. Roberts, S. McGowan, D. Hay, E. Giannoulatou, M. Lynch, . . . D. R. Higgs. 2014. "Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment." *Nat Genet* 46 (2): 205-12. https://doi.org/10.1038/ng.2871.

Larson, A. G., D. Elnatan, M. M. Keenen, M. J. Trnka, J. B. Johnston, A. L. Burlingame, . . . G. J. Narlikar. 2017. "Liquid droplet formation by HP1α suggests a role for

phase separation in heterochromatin." *Nature* 547 (7662): 236-240. https://doi.org/10.1038/nature22822.

Lee, J. H., E. R. Daugharthy, J. Scheiman, R. Kalhor, J. L. Yang, T. C. Ferrante, . . . G. M. Church. 2014. "Highly multiplexed subcellular RNA sequencing in situ." *Science* 343 (6177): 1360-3. https://doi.org/10.1126/science.1250212.

Lee, Je Hyuk, Evan R. Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C. Ferrante, Richard Terry, . . . George M. Church. 2015. "Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues." *Nature Protocols* 10 (3): 442-458. https://doi.org/10.1038/nprot.2014.191. https://doi.org/10.1038/nprot.2014.191.

Li, F. Z., Z. E. Liu, X. Y. Li, L. M. Bu, H. X. Bu, H. Liu, and C. M. Zhang. 2020. "Chromatin 3D structure reconstruction with consideration of adjacency relationship among genomic loci." *BMC Bioinformatics* 21 (1): 272. https://doi.org/10.1186/s12859-020-03612-4.

Li, J., W. Zhang, and X. Li. 2018. "3D Genome Reconstruction with ShRec3D+ and Hi-C Data." *IEEE/ACM Trans Comput Biol Bioinform* 15 (2): 460-468. https://doi.org/10.1109/tcbb.2016.2535372.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, . . . J. Dekker. 2009. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science* 326 (5950): 289-93. https://doi.org/10.1126/science.1181369.

Liu, Yuanlong, Luca Nanni, Stephanie Sungalee, Marie Zufferey, Daniele Tavernari, Marco Mina, . . . Giovanni Ciriello. 2021. "Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes." *Nature Communications* 12 (1): 2439. https://doi.org/10.1038/s41467-021-22666-3. https://doi.org/10.1038/s41467-021-22666-3.

Maass, Philipp G., A. Rasim Barutcu, and John L. Rinn. 2018. "Interchromosomal interactions: A genomic love story of kissing chromosomes." *Journal of Cell Biology* 218 (1): 27-38. https://doi.org/10.1083/jcb.201806052. https://doi.org/10.1083/jcb.201806052.

Mateo, L. J., S. E. Murphy, A. Hafner, I. S. Cinquini, C. A. Walker, and A. N. Boettiger. 2019. "Visualizing DNA folding and RNA in embryos at single-cell resolution." *Nature* 568 (7750): 49-54. https://doi.org/10.1038/s41586-019-1035-4.

Mifsud, B., F. Tavares-Cadete, A. N. Young, R. Sugar, S. Schoenfelder, L. Ferreira, . . . C. S. Osborne. 2015. "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C." *Nat Genet* 47 (6): 598-606. https://doi.org/10.1038/ng.3286.

Mirny, Leonid A. 2011. "The fractal globule as a model of chromatin architecture in the cell." *Chromosome Research* 19 (1): 37-51. https://doi.org/10.1007/s10577-010-9177-0. https://doi.org/10.1007/s10577-010-9177-0.

Misteli, Tom. 2020. "The Self-Organizing Genome: Principles of Genome Architecture and Function." *Cell* 183 (1): 28-45. https://doi.org/https://doi.org/10.1016/j.cell.2020.09.014. https://www.sciencedirect.com/science/article/pii/S0092867420311557.

Mumbach, Maxwell R., Adam J. Rubin, Ryan A. Flynn, Chao Dai, Paul A. Khavari, William J. Greenleaf, and Howard Y. Chang. 2016. "HiChIP: efficient and sensitive analysis of protein-directed genome architecture." *Nature Methods* 13 (11): 919-922. https://doi.org/10.1038/nmeth.3999. https://doi.org/10.1038/nmeth.3999.

Münkel, Christian, and Jörg Langowski. 1998. "Chromosome structure predicted by a polymer model." *Physical Review E* 57 (5): 5888.

Nagano, T., Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, . . . P. Fraser. 2013. "Single-cell Hi-C reveals cell-to-cell variability in chromosome structure." *Nature* 502 (7469): 59-64. https://doi.org/10.1038/nature12593.

Nguyen, Huy Q., Shyamtanu Chattoraj, David Castillo, Son C. Nguyen, Guy Nir, Antonios Lioutas, . . . C. ting Wu. 2020. "3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing." *Nature Methods* 17 (8): 822-832. https://doi.org/10.1038/s41592-020-0890-0. https://doi.org/10.1038/s41592-020-0890-0.

Nilsson, M., H. Malmgren, M. Samiotaki, M. Kwiatkowski, B. P. Chowdhary, and U. Landegren. 1994. "Padlock probes: circularizing oligonucleotides for localized DNA detection." *Science* 265 (5181): 2085-8. https://doi.org/10.1126/science.7522346.

Nora, E. P., B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, . . . E. Heard. 2012. "Spatial partitioning of the regulatory landscape of the X-inactivation centre." *Nature* 485 (7398): 381-5. https://doi.org/10.1038/nature11049.

Nuebler, Johannes, Geoffrey Fudenberg, Maxim Imakaev, Nezar Abdennur, and Leonid A. Mirny. 2018. "Chromatin organization by an interplay of loop extrusion and compartmental segregation." *Proceedings of the National Academy of Sciences* 115 (29): E6697-E6706. https://doi.org/doi:10.1073/pnas.1717730115. https://www.pnas.org/doi/abs/10.1073/pnas.1717730115.

Oluwadare, Oluwatosin, Max Highsmith, and Jianlin Cheng. 2019. "An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data." *Biological Procedures Online* 21 (1): 7. https://doi.org/10.1186/s12575-019-0094-0. https://doi.org/10.1186/s12575-019-0094-0.

Paulsen, J., O. Gramstad, and P. Collas. 2015. "Manifold Based Optimization for Single-Cell 3D Genome Reconstruction." *PLoS Comput Biol* 11 (8): e1004396. https://doi.org/10.1371/journal.pcbi.1004396.

Paulsen, Jonas, Monika Sekelja, Anja R. Oldenburg, Alice Barateau, Nolwenn Briand, Erwan Delbarre, . . . Philippe Collas. 2017. "Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts." *Genome Biology* 18 (1): 21. https://doi.org/10.1186/s13059-016-1146-2. https://doi.org/10.1186/s13059-016-1146-2.

Payne, Andrew C., Zachary D. Chiang, Paul L. Reginato, Sarah M. Mangiameli, Evan M. Murray, Chun-Chen Yao, . . . Fei Chen. 2021. "In situ genome sequencing resolves DNA sequence and structure in intact biological samples." *Science* 371 (6532): eaay3446. https://doi.org/doi:10.1126/science.aay3446. https://www.science.org/doi/abs/10.1126/science.aay3446.

Pope, B. D., T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, . . . D. M. Gilbert. 2014. "Topologically associating domains are stable units of replication-timing regulation." *Nature* 515 (7527): 402-5. https://doi.org/10.1038/nature13986.

Racko, D., F. Benedetti, J. Dorier, and A. Stasiak. 2018. "Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes." *Nucleic Acids Res* 46 (4): 1648-1660. https://doi.org/10.1093/nar/gkx1123.

Ramani, V., X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, . . . J. Shendure. 2017. "Massively multiplex single-cell Hi-C." *Nat Methods* 14 (3): 263-266. https://doi.org/10.1038/nmeth.4155.

Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, . . . E. L. Aiden. 2014. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." *Cell* 159 (7): 1665-80. https://doi.org/10.1016/j.cell.2014.11.021.

Rao, S. S. P., S. C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K. R. Kieffer-Kwon, . . . E. L. Aiden. 2017. "Cohesin Loss Eliminates All Loop Domains." *Cell* 171 (2): 305-320.e24. https://doi.org/10.1016/j.cell.2017.09.026.

Rieber, L., and S. Mahony. 2017. "miniMDS: 3D structural inference from high-resolution Hi-C data." *Bioinformatics* 33 (14): i261-i266. https://doi.org/10.1093/bioinformatics/btx271.

Rousseau, M., J. Fraser, M. A. Ferraiuolo, J. Dostie, and M. Blanchette. 2011. "Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling." *BMC Bioinformatics* 12: 414. https://doi.org/10.1186/1471-2105-12-414.

Sabari, B. R., A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, . . . R. A. Young. 2018. "Coactivator condensation at super-enhancers links phase separation and gene control." *Science* 361 (6400). https://doi.org/10.1126/science.aar3958.

Sanborn, A. L., S. S. Rao, S. C. Huang, N. C. Durand, M. H. Huntley, A. I. Jewett, . . . E. L. Aiden. 2015. "Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes." *Proc Natl Acad Sci U S A* 112 (47): E6456-65. https://doi.org/10.1073/pnas.1518552112.

Schoenfelder, S., M. Furlan-Magaril, B. Mifsud, F. Tavares-Cadete, R. Sugar, B. M. Javierre, . . . P. Fraser. 2015. "The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements." *Genome Res* 25 (4): 582-97. https://doi.org/10.1101/gr.185272.114.

Schwarzer, W., N. Abdennur, A. Goloborodko, A. Pekowska, G. Fudenberg, Y. Loe-Mie, . . . F. Spitz. 2017. "Two independent modes of chromatin organization revealed by cohesin removal." *Nature* 551 (7678): 51-56. https://doi.org/10.1038/nature24281.

Schwessinger, Ron, Matthew Gosden, Damien Downes, Richard C. Brown, A. Marieke Oudelaar, Jelena Telenius, . . . Jim R. Hughes. 2020. "DeepC: predicting 3D genome folding using megabase-scale transfer learning." *Nature Methods* 17 (11): 1118-1124. https://doi.org/10.1038/s41592-020-0960-3. https://doi.org/10.1038/s41592-020-0960-3.

Serra, F., D. Baù, M. Goodstadt, D. Castillo, G. J. Filion, and M. A. Marti-Renom. 2017. "Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors." *PLoS Comput Biol* 13 (7): e1005665. https://doi.org/10.1371/journal.pcbi.1005665.

Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, . . . W. de Laat. 2006. "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)." *Nat Genet* 38 (11): 1348-54. https://doi.org/10.1038/ng1896.

Su, Jun-Han, Pu Zheng, Seon S. Kinrot, Bogdan Bintu, and Xiaowei Zhuang. 2020. "Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin." *Cell* 182 (6): 1641-1659.e26. https://doi.org/https://doi.org/10.1016/j.cell.2020.07.032. https://www.sciencedirect.com/science/article/pii/S0092867420309405.

Szabo, Q., A. Donjon, I. Jerković, G. L. Papadopoulos, T. Cheutin, B. Bonev, . . . G. Cavalli. 2020. "Regulation of single-cell genome organization into TADs and

chromatin nanodomains." *Nat Genet* 52 (11): 1151-1157. https://doi.org/10.1038/s41588-020-00716-8.

Takei, Y., J. Yun, S. Zheng, N. Ollikainen, N. Pierson, J. White, . . . L. Cai. 2021. "Integrated spatial genomics reveals global architecture of single nuclei." *Nature* 590 (7845): 344-350. https://doi.org/10.1038/s41586-020-03126-2.

Tanabe, H., S. Müller, M. Neusser, J. von Hase, E. Calcagno, M. Cremer, . . . T. Cremer. 2002. "Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates." *Proc Natl Acad Sci U S A* 99 (7): 4424-9. https://doi.org/10.1073/pnas.072618599.

Tjong, H., W. Li, R. Kalhor, C. Dai, S. Hao, K. Gong, . . . F. Alber. 2016. "Population-based 3D genome structure analysis reveals driving forces in spatial genome organization." *Proc Natl Acad Sci U S A* 113 (12): E1663-72. https://doi.org/10.1073/pnas.1512577113.

Torgerson, Warren S. 1958. *Theory and methods of scaling.Theory and methods of scaling.* Oxford, England: Wiley.

Trieu, T., and J. Cheng. 2016a. "MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data." *Bioinformatics* 32 (9): 1286-92. https://doi.org/10.1093/bioinformatics/btv754.

Trieu, T., O. Oluwadare, and J. Cheng. 2019. "Hierarchical Reconstruction of High-Resolution 3D Models of Large Chromosomes." *Sci Rep* 9 (1): 4971. https://doi.org/10.1038/s41598-019-41369-w.

Trieu, Tuan, and Jianlin Cheng. 2016b. "3D genome structure modeling by Lorentzian objective function." *Nucleic Acids Research* 45 (3): 1049-1058. https://doi.org/10.1093/nar/gkw1155. https://doi.org/10.1093/nar/gkw1155.

van Emmerik, C. L., and H. van Ingen. 2019. "Unspinning chromatin: Revealing the dynamic nucleosome landscape by NMR." *Prog Nucl Magn Reson Spectrosc* 110: 1-19. https://doi.org/10.1016/j.pnmrs.2019.01.002.

Vilarrasa-Blasi, Roser, Paula Soler-Vila, Núria Verdaguer-Dot, Núria Russiñol, Marco Di Stefano, Vicente Chapaprieta, . . . José Ignacio Martin-Subero. 2021. "Dynamics of genome architecture and chromatin function during human B cell differentiation and neoplastic transformation." *Nature Communications* 12 (1): 651. https://doi.org/10.1038/s41467-020-20849-y. https://doi.org/10.1038/s41467-020-20849-y.

Walter, J., L. Schermelleh, M. Cremer, S. Tashiro, and T. Cremer. 2003. "Chromosome order in HeLa cells changes during mitosis and early G1, but is stably maintained during subsequent interphase stages." *J Cell Biol* 160 (5): 685-97. https://doi.org/10.1083/jcb.200211103.

Wang, S., J. H. Su, B. J. Beliveau, B. Bintu, J. R. Moffitt, C. T. Wu, and X. Zhuang. 2016. "Spatial organization of chromatin domains and compartments in single chromosomes." *Science* 353 (6299): 598-602. https://doi.org/10.1126/science.aaf8084.

Winick-Ng, Warren, Alexander Kukalev, Izabela Harabula, Luna Zea-Redondo, Dominik Szabó, Mandy Meijer, . . . Ana Pombo. 2021. "Cell-type specialization is encoded by specific chromatin topologies." *Nature* 599 (7886): 684-691. https://doi.org/10.1038/s41586-021-04081-2. https://doi.org/10.1038/s41586-021-04081-2.

Xiong, Kyle, and Jian Ma. 2019. "Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions." *Nature Communications* 10 (1): 5069. https://doi.org/10.1038/s41467-019-12954-4. https://doi.org/10.1038/s41467-019-12954-4.

Zhan, Y., L. Mariani, I. Barozzi, E. G. Schulz, N. Blüthgen, M. Stadler, . . . L. Giorgetti. 2017. "Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes." *Genome Res* 27 (3): 479-490. https://doi.org/10.1101/gr.212803.116.

Zhao, Z., G. Tavoosidana, M. Sjölinder, A. Göndör, P. Mariano, S. Wang, . . . R. Ohlsson. 2006. "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions." *Nat Genet* 38 (11): 1341-7. https://doi.org/10.1038/ng1891.

Zhu, G., W. Deng, H. Hu, R. Ma, S. Zhang, J. Yang, . . . J. Zeng. 2018. "Reconstructing spatial organizations of chromosomes through manifold learning." *Nucleic Acids Res* 46 (8): e50. https://doi.org/10.1093/nar/gky065.