bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

PLOS | SUBMISSION

ChromoTrace: Reconstruction of 3D Chromosome Configurations by Super-Resolution Microscopy

Sandro Morganella[1], Øyvind Ødegård[2], Stephanie Alexander[2], Jonas Ries[2], Tomas Fitzgerald[1], Jan Ellenberg[2] and Ewan Birney[1,*] ¶

1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

2 European Molecular Biology Laboratory (EMBL), Cell Biology and Biophysics Unit, Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany

* birney@ebi.ac.uk

Abstract

Motivation: The three-dimensional structure of chromatin plays a key role in genome function, including gene expression, DNA replication, chromosome segregation, and DNA repair. Furthermore the location of genomic loci within the nucleus, especially relative to each other and nuclear structures such as the nuclear envelope and nuclear bodies strongly correlates with aspects of function such as gene expression. Therefore, determining the 3D position of the 6 billion DNA base pairs in each of the 23 chromosomes inside the nucleus of a human cell is a central challenge of biology. Recent advances of super-resolution microscopy in principle enable the mapping of specific molecular features with nanometer precision inside cells. Combined with highly specific, sensitive and multiplexed fluorescence labeling of DNA sequences this opens up the possibility of mapping the 3D path of the genome sequence in situ.

Results: Here we develop computational methodologies to reconstruct the sequence configuration of all human chromosomes in the nucleus from a super-resolution image of a set of fluorescent in situ probes hybridized to the genome in a cell. To test our approach we develop a method for the simulation of chromatin packing in an idealized human nucleus. Our reconstruction method, ChromoTrace, uses suffix trees to assign a known linear ordering of in situ probes on the genome to an unknown set of 3D in situ probe positions in the nucleus from super-resolved images using the known genomic probe spacing as a set of physical distance constraints between probes. We find that ChromoTrace can assign the 3D positions of the majority of loci with high accuracy and reasonable sensitivity to specific genome sequences. By simulating spatial resolution, label multiplexing and noise scenarios we assess algorithm performance under realistic experimental constraints. Our study shows that it is feasible to achieve chromosome-wide reconstruction of the 3D DNA path in chromatin based on super-resolution microscopy images.

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

The primary nucleic acid sequence of the human genome is not sufficient to understand its functions and their regulation. Fitting the 6 billion basepairs or approximately 2 m of double-helical DNA into an approximately 10 μm nucleus requires tight packing of DNA into chromatin, where about 150 bp of DNA are wrapped around cylindrical nucleosome core particles, which in turn can be tightly packed due to interspersed flexible linker DNA [1]. In addition, each chromosomal DNA molecule occupies a discrete 3D volume inside the nucleus and the arrangement of these chromosome territories is non-random and changes with cell differentiation [2,3]. This remarkable management of 23 large linear polymer molecules controls crucial functions of the genome, such as gene expression, DNA replication, chromosome segregation, and DNA repair.

Structural biology techniques, such as electron microscopy, crystallography, and NMR have given atomic level insights into the physical structure of the DNA double helix and the nucleosome [4]. In vitro, also higher order structures such as nucleosomes stacked into 11 or 30 nm chromatin fibres can be observed and studied at high resolution. However, the existence of regular higher order nucleosome structures in vivo has not been demonstrated under physiological conditions. To date, little direct information is available about the functionally crucial 3D folding and structure of chromatin between the scale of single nucleosomes (approximately 5 nm) and the diffraction limit of light (200 nm), which can only resolve entire chromosome territories with a size of a few μm.

In situ, classically two general types of higher order chromatin organization have been distinguished at a coarser level, euchromatin which tends to be less compact and displays high gene density and activity, and heterochromatin, with a higher degree of compaction and lower gene density and activity [5]. Due to the arrangement of chromatin from individual chromosomes in territories, the majority of DNA-DNA interactions occur in *cis*, while *trans* interactions are more rare and mostly observed on the surface of or on loops outside of territories [6–8]. Within territories and across the whole nucleus euchromatin and heterochromatin are generally spatially separated [9], leading to heterochromatin rich and gene expression poor domains at the nuclear periphery and around nucleoli. Gene expression is intrinsically linked to the 3D structure of chromosomes, chromatin packing densities and the accessibility of DNA by the transcriptional machinery.

In the last 10 years, biochemical DNA crosslinking technologies based on chromosome conformation capture (3C), have been developed to address the issue of higher order chromatin structure in an indirect manner [10]. These methods have been widely used to measure the average linear proximity of genome sequences to each other in cell populations with good throughput and at kb resolution. The resulting contact frequency maps analyzed with computational models have indirectly inferred principles of genome organization [11]. A major result of these studies, is that chromosomes are organized into domains of 400-800 kb that are topologically associated. These TADs are the smallest structuring units of chromatin above the 150 bp nucleosome level that can be reliably detected biochemically so far. Although good correlations between contact frequency and function of regulatory elements has been shown for several genes [12], such crosslinking technologies cannot determine the 3D position and physical distances of genomic loci inside the nucleus directly.

Recent developments in light microscopy techniques, collectively called super-resolution microscopy, can determine the position of single fluorescent molecules with a precision of a few nanometers, much below the diffraction barrier. This allows the characterization of previously unobserved details of biological structures and processes [13–16]. First

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

studies have already explored the use of super-resolution microscopy to investigate chromatin structure [17–19], such as the organization of distinct epigenetic states in Drosophila cells [20] that suggested distinct folding mechanisms and packing densities that correlate with gene expression. Dissection of nucleosome organization inside the nucleus in single cells using super-resolution shows that higher nucleosome compaction corresponds to heterochromatin while lower compaction associates with active chromatin regions and RNA polymerase II, and that the spatial distribution, size and compaction of nucleosome correlate to cell pluripotency [18]. While these studies provide first new intriguing insights into chromatin organization they have so far largely focused on single loci without a complete 3D reconstruction of a chromosome or the genome.

However, the resolving power of super-resolution microscopy raises the tantalizing possibility to directly reconstruct the 3D path of large parts of the chromosomal DNA sequence. Super-resolution microscopy can resolve unprecedentedly small volume elements (approximately 20 x 20 x 20 nm [21]) inside the total nuclear volume (approximately 8 x10-6 μm^3), which will on average contain only up to 2 kb or a few nucleosomes. This fundamental increase in information of the relative positioning of defined loci in the genome can now be leveraged computationally.

This increase in resolution, which enables to distinguish around 60 million volume elements inside a single nucleus, can be combined with any sensitive and site specific fluorescence in situ hybridization (FISH) probe design that allows for spectral and/or temporal multiplexing. Several methods that fulfill these criteria have recently been developed, and fall within two general probe design categories. Either a primary imager strand with fluorophore-containing DNA is hybridized to the genome directly [22] or a primary genome-sequence specific DNA probe that facilitates transient binding of the fluorophore-containing secondary imager strand is used (DNA-PAINT) [23]. Our reconstruction algorithm should in principle allow the mapping of the genome sequence in 3D with a resolution of tens of nucleosomes, depending on their local packing density.

Carrying out such large-scale genome mapping studies by systematic super-resolution microscopy will critically depend upon choosing the best design of the necessary chromosome or genome wide fluorescent probe libraries and use sufficient resolution in the employed 3D super-resolution imaging technology. To prove that such studies are feasible and guide their probe design and microscope technology choices, we have developed an algorithm, called ChromoTrace that uses an efficient combinatorial search to test the theoretical possibility of complete three-dimensional reconstruction of chromosomal scale regions of DNA inside nuclei of single human cells (**Fig 1**). To thoroughly test our algorithm, we have developed a simulation to model chromatin packing of the human genome sequence within a realistic geometry of the nucleus. Our modeled 3D architecture of the genome reproduces several key characteristics of real chromosomes, and our ChromoTrace reconstruction algorithm, based on suffix trees, then maps the simulated 3D sequence label positions back to the reference genome. By simulating realistic resolution, label multiplexing and noise scenarios we assessed the algorithm performance for different experimental scenarios. Our results show that ChromoTrace can map the positions of the labeled probes to the genome with very high precision, while it has some limitations in terms of recall. Importantly, our study shows for the first time that it is feasible to achieve chromosome-wide reconstruction of the 3D DNA path in chromatin based on current super-resolution microscopy and DNA labeling technology and defines the required quality of experimental data to achieve a certain bp resolution and reconstruction completeness, which will be invaluable to guide experimental efforts to generate such data sets systematically.

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig1. Representation of a chromosome labelling scheme. (A) Linear DNA displayed as a ribbon with six genome regions labelled in three different colors. (B) 3D view of an expected super-resolution microscope experiment.

Materials and Methods

Simulation of chromatin structure of the nucleus

In order to generate simulations of the chromatin structure we use a well-known mathematical model known as Self-Avoiding Walk (SAW). A SAW is a sequence of distinct points in a d-dimensional (hyper) cubic lattice such that each point is a nearest neighbor of its predecessor. The generation of a SAW for dimension of 2 or greater is not trivial, and the complexity increases with the portion of the lattice that the SAW has to occupy (density). For example consider a generic d-lattice containing N points for the generation of a SAW composed of K points, it is clear that the complexity of this task increases with K and that it has the maximum complexity when $K = N$. A pivot algorithm was used to generate the SAWs. In particular, for each chromosome we generated a SAW of length N (N being the size of the corresponding chromosome), that is, we generated a sequence of points in 3D-space, $S = P_1, P_2, \ldots, P_N$ satisfying the SAW properties of adjacency and uniqueness. This process starts by randomly picking the initial position P_1 of each SAW among all the possible available positions: **i)** positions inside the nucleus that are not included in the nucleolus **ii)** positions that have not been already picked and included in any other walk. Let P_i be the current 3D-coordinate position associated with the ith point of the SAW ($i < N$), then pick the next position $P_{(i+1)}$ across all the available adjacent points, where all positions have the same probability to be picked. If the current point P_i does not have any available position (i.e., all adjacent positions have been already assigned to a SAW), then restart the computation of the current SAW from another point P_j of the SAW with $j < i$. Our choice was to select $j = i - 0.2i$. If the resulting $j \leq 0$ then restart the computation of the SAW from the first point, by picking a new random point P_1. **Fig 2** shows an example of a simulation obtained by using the described method. From this figure one can appreciate that the SAW-based model approach is able to produce a challenging scenario for the proposed reconstruction algorithm, generating visually different types of chromatin conformation (i.e., similar to open, fractal and compact).

Fig2. 3D view of the nucleus simulation. (A) One whole genome simulation, each chromosome (two copies of the 22 autosomes and the two sex chromosomes X and Y) is drawn with a different colour. Individual chromosomes show a random configuration with a high degree of compactness inside the nucleus. (B) One copy of chromosome 1 is highlighting the ability of the simulation to create a realistic chromatin structure having the expected conformations, open, fractal and compact.

ChromoTrace Algorithm

We propose a new algorithm, ChromoTrace, to reconstruct the position of the chromosomes from labelled points in the chromosome, defined below:

Main data structures: The two main inputs of our algorithm are a labeling and a segmentation file which are modeled through the following two data structures:
First, a list $\mathbf{L}$ modelling the labelling file, where each element $L^i \in \mathbf{L}$ is a concatenated list for the ith chromosome with $i = 1, 2, \ldots, 22, X, Y$.

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Each concatenated list contains $||L^i||$ elements, that is, the number of probes labelled on the ith chromosome and each pair of elements $L^i_j, L^i_{j+1} \in L^i$ correspond to two probes that are genomic adjacent in the labelling file.

Given a sequence of elements $L^i_j, \ldots, L^i_j \in L^i$ we define $col(L^i_j, \ldots, L^i_j)$ to be a function returning the sequence of associated colours. This function is defined on the discrete set of colours used to label the probes. For example, for a three colour labelling schema we have $col(L^i_j, \ldots, L^i_{j'}) :\rightarrow red, green, blue$

Second, a set $\mathbf{S}$ modelling the segmented file, where each element $S_k \in \mathbf{S}$ corresponds to a segmented point in the segmentation file.

It contains $||\mathbf{S}||$ elements (the number of segmented points), each element $S_k \in \mathbf{S}$ has three components specifying the coordinates in the 3D-lattice: $S^x_k, S^y_k, S^z_k \in N_0$. Given two points $S_k, S_{k'} \in \mathbf{S}$ we define their Euclidean distance:

$$d(S_k, S_{k'}) = \sqrt{(S^x_k - S^x_{k'})^2 + (S^y_k - S^y_{k'})^2 + (S^z_k - S^z_{k'})^2}, \text{with } d(S_k, S_{k'}) \in R_+$$

In addition, we define the function $col(S_k, \ldots, S_{k'})$ in agreement with the previous definition used for $\mathbf{L}$.

Step1: Computing the suffix tree and the unique color signatures:

A suffix tree $\mathbf{T}$ is used to model the labeling file. This tree is defined on the set of colours used to label the probes and it has the following properties:

Let $||T||$ be the maximum depth of the tree $\mathbf{T}$, and let $||\omega||$ be the number of colours used for the labelling then the tree has $||\omega||^{||T||}$leaves. Each node has an associated colour in the alphabet ω (e.g. $\omega \rightarrow red, green, blue$), and each node $T_t \in \mathbf{T}$ can be uniquely identified by using the sequence of colours (nodes) visited on the path from the root node (T_0) to this node: $col(T_0, \ldots, T_t)$.

In addition, each node $T_t \in \mathbf{T}$ has an associated number indicating the total times that the colour sequence $col(T_0, \ldots, T_t)$ is found in the labelling list, This number can be accessed by referring to $\mathbf{T}[col(T_0, \ldots, T_t)]$ (or more concisely $\#T_t$).

We define the set of unique *signature colours* $\mathbf{C}$ as the set of nodes in $T_t \in \mathbf{T}$ (and consequently the set of sequence colours) that has been found only once in the labelling list: $T_t \in \mathbf{C}$ if $f\#T_t = 1$.

Note that the suffix tree has to be reversible, so it needs to be computed by scrolling the elements in $\mathbf{L}$ from the first to the last probe and *vice versa*.

Step2: Computing the distance graph and the trivial path set:

The distance graph is computed from the list of segmented points $\mathbf{S}$, and it is modelled as an adjacency matrix $\mathbf{A}$ having the following properties:

$$A = ||S|| \times ||S||$$

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$A_{k,z} = \begin{cases} =1 \text{ if } d(S_k,S_z) \leq \text{ Threshold} \\ =0 \text{ Otherwise} \end{cases} \quad \text{with } k \neq z, \text{ and } k, x = 1, \ldots, ||S||$$

Here, the threshold used to define adjacent elements should be related to the degree of chromatin compactness. In particular, this value should decrease with the compactness observed in the segmentation data (super resolution image). In our performance assessment we used the minimum threshold value of 1 indicating that the produced segmentation is highly compact.

Given the adjacency matrix $\mathbf{A}$ we define a list of trivial paths $\mathbf{V}$ where each element $V_v \in \mathbf{V}$ is a sequence of segmented points $V_v = S_{k1}, S_{k2}, \ldots, S_N$ such that:

$$A_{k_i,k_{i+1}} = 1 \text{ for } i = 1, N-1$$

where points have to be adjacent, and

$$\sum_{z=1}^{||S||} A_{k_i,z} = 2 \text{ for } i = 2, \ldots, N-1$$

where the internal points of the sequence have to have exactly two adjacent points, and

$$\sum_{z=1}^{||S||} A_{k_i,z} > 0 \text{ and } \sum_{z=1}^{||S||} A_{k_i,z} \neq 2 \text{ for } i = 1, N$$

where the first and the last point of the sequence need to have either one or more than two adjacent points.

Step3: Locating trivial paths:

The aim of this step is to try to assign the genomic positions to the trivial paths. In particular, we look for a subset of trivial paths $\bar{\mathbf{V}} \subseteq \mathbf{V}$ that have a unique colour signature, that is, we look for $\bar{\mathbf{V}} \subseteq \mathbf{V}$, such that:

$$\forall\ \bar{V}_v = \{S_{k_1}, S_{k_2}, \ldots, S_N\} \in \bar{\mathbf{V}} : \mathbf{T}[col(S_{k_1}, S_{k_2}, \ldots, S_N)] \in \mathbf{C}$$

Step4: Expansion of trivial paths:

After **Step3** we know exactly the genomic position of the segmented probes in $\bar{\mathbf{V}}$, in fact we have a unique association between each segmented element in $\bar{\mathbf{V}}$ and one element in $\mathbf{L}$. That is, $\forall\ \bar{V}_v = \{S_{k_1}, S_{k_2}, \ldots, S_N\} \in \bar{\mathbf{V}}$ we have the association $[S_{k_p} \rightarrow L^i_j]$(with $p = 1, \ldots, N$).

Now we can use these associations to try to extend the trivial paths by using the colour information contained in the labelled file. The expansion is separately performed on the first and last element of each trivial path (i.e. S_{k_1} and S_N).

Let S_{k_p} be either the first or the last element of one trivial path in $\bar{\mathbf{V}}(p = 1, N)$, then we know that for this element we have an associated element in $\mathbf{L}$, $[S_{k_p} \rightarrow L^i_j]$. Let L^i_{j-1} be the adjacent node of L^i_j, the aim of the expansion is to find a segmented element $S_z \in \mathbf{S}$, such that:

$$A_{k_p,z} = 1 \text{ and } col(S_z) = col(L^i_{j-1})$$

If such an element is found the trivial path is updated, by concatenating the new association $[S_z \rightarrow L^i_{j-1}]$. For example, if $S_{k_p} = S_{k_1}$, then the updated trivial path is

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\bar{V}_v = \{S_z, S_{k_1}, S_{k_2}, \ldots, S_N\}.$$

This step is iteratively repeated until the above condition is satisfied exactly for one probe, and it stops if we do not find any adjacent probe with the expected colour.

Alternatively, when we have ambiguous situations, that is, more than one adjacent probe of S_{k_p} have the expected colour then we expand our search to the next adjacent probes, in other words, we look for the pair of segmented elements that satisfy the following condition:

$$col(S_z, S_{z'} = col(L^i_{j-1}, L^i_{j-2}))$$

where $A_{z,z'} = 1$. If again we have more than one pair of segmented probes satisfying this condition the algorithm stops.

End of the algorithm:

When all the described steps are performed, we will find in $\bar{\mathbf{V}}$ all the segmented probes for which we could assign a genomic position. Therefore, we can remove from $\mathbf{S}$ the segmented points in $\bar{\mathbf{V}}$, and from $\mathbf{L}$ the elements having an association with the elements in $\bar{\mathbf{V}}$. This update of $\mathbf{L}$ and $\mathbf{S}$ results in a new set of unique colour signature $\mathbf{C}$, a new adjacency matrix $\mathbf{D}$, and consequently a new set of trivial paths $\mathbf{V}$. All the described steps, including the update of $\mathbf{L}$ and $\mathbf{S}$, are iteratively run until no new association $[S_z \rightarrow L^i_{j-1}]$ is found and full iterative closure has been achieved.

Results

Simulation of reasonable chromatin structure in the nucleus

The nucleus is delimited by the nuclear envelope and contains the nucleolus and chromatin. To simulate the results of super-resolved detection of large-scale probe hybridizations, we need to build a model of reasonable packing densities of DNA in the human nucleus. The precise local density of DNA in the human nucleus is surprisingly unclear due to the uncertainty regarding the in vivo structure of chromatin. To tackle this problem with reasonable computational complexity at the scale relevant for super-resolution microscopy our simulation uses an intermediate grained self-avoiding-walk (SAW) model for chromatin on a grid of points. We assume random packing of chromatin and an average density corresponding to the highest values estimated in human cells, to conservatively estimate the sequence reconstruction challenge at the single cell level. Although real biological data will likely show stronger heterogeneity in local packing density, this would aid rather than hinder the reconstruction task (**Discussion**). We focus here purely on the single cell reconstruction problem, although cell-to-cell structure conservation is likely to improve the reconstruction ability by allowing averaging of conserved structures of the same genomic sequence between cells (**Discussion**).

The SAW model has been widely used in literature to simulate the structure of polymer chains [24], and as such should also be suitable for the generation of chromatin conformation producing a structure with similar features to those observed in real experiments. For this simulation we wanted to generate a model that was in agreement with a simple yet realistic structure of the human nucleus (**Fig 1B**). We used a 3D sphere of 500 μm^3 volume, with an internal sphere of 50 μm^3 volume devoid of chromatin that represents the nucleolus. In this space the simulation laid the 46 human chromosomes (two copies of the 22 autosomes and the two sex chromosomes X and Y for a total of 6,053,303,898 bp). Each chromosome was then modeled as a separate SAW with a length proportional

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to the real chromosome size and the simulated chromatin paths were forced to remain inside the nucleus but not allowed to enter the nucleolus.

Fig 2 shows the packing density and folding characteristics for a synthetic whole chromosome set generated by SAWs to simulate the whole genome within the nucleus. Overall the simulated chromosomes display a number of structural characteristics that are very similar to previous experimental data [7, 25, 26]. Interestingly, although we assumed random packing and no biologically driven heterogeneity in density, the simulation results in a variety of chromatin conformations: similar to the observed open, fractal, and compact [25]. Furthermore, each chromosome resides roughly within its own chromosome territory (**Fig 2**).

In order to assess the properties of our simulated genome architectures statistically, we next simulated 100 synthetic nuclear chromosome sets. To quantify the packing density of chromatin in our simulated genomes we used the average number of contacts between labeled probe positions within the SAWs across the 3D search space. For each chromosome **Fig 3A** shows the average number of contacts made within the same chromosomal SAWs (intra-chromosome contacts) and the average number of contacts made between different chromosomal SAWs (inter-chromosome contacts). Unsurprisingly the average number of contacts within chromosomes is highly correlated with chromosome length (**Fig 3A**) and after adjusting of chromosome length each chromosome shows proportionally similar numbers (**Fig 3B**).

Fig3. Packing density of the simulated chromatin configurations. The packing density was obtained as the number of contacts in the 3D-lattice space between non-adjacent points within the SAWs and presented as the average number of contacts across all the 100 generated simulations. (A) The number of intra-chromosome and the inter-chromosome overall. (B) The number of intra-chromosome and the inter-chromosome overall after adjusting for chromosome length. (C) Linear distance against Euclidean distance for all points at five different linear distances, as the linear distance between two points increase so does both the value and variance of Euclidean distances.

It is reasonable to expect that as the genomic distance between two loci on the same chromosome increases than on average so should their physical distance in 3D space. To test this, we walked along the simulated genomes using five different genomic distances and calculated the Euclidean distance that our simulation assigned to all pairs of such spaced loci for all chromosomes. As expected, the average physical distance between two loci in 3D space is highly correlated to their genomic distance and the variance in Euclidean distance increases with genomic distance (**Fig 3C**).Loci with a large genomic distance but short 3D distance are reminiscent of chromatin looping behavior. Overall we conclude that we have a reasonable simulation of genome architecture, without heterogeneity in packing density, that recapitulates many features of chromatin.

Testing the mapping of chromosomal DNA sequence to 3D positions of labeled loci

Confident in our ability to simulate realistic chromatin paths, we then explored under which experimental conditions, and with what computational methods, the 3D positions of fluorescently labeled genomic loci could be mapped back to the linear chromosomal DNA sequences. Computationally the inputs to the method is a description of the linear labeling of the genome with different colors, which we are able to experimentally design to optimize reconstruction, and the results of the super-resolution image determination,

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

providing a set of the 3D coordinates (x,y,z) and color classification but without the indication of the locus (**Fig 4A**). The goal of the reconstruction algorithm is to assign each of the in situ locus with a specific (x,y,z) position.

Fig4. Illustration of the ChromoTrace algorithm. (A) The 3D coordinates that would be obtained from super-resolution microscope imaging are converted into an adjacency graph. Given the pre-specified linear labelling sequence of green-red-blue-blue-green a trivial path is detected. (B) Extension is attempted is both forward and reverse directions mapping colors back to the probe design, position A has only one option in the adjacency graph and is unambiguously mapped to green, position F is mapped to green but on two different paths, finally position G two two options, red and blue but there is only the option to mapp position G to the color blue given the probe design.

This problem is computationally challenging, as we have vastly more loci along the chromosomal sequence than different colors to create distinguishable sequence specific probes. Further challenges will occur due to errors in the labeling and imaging experiment. We proposed to solve this problem using the fact that the linear sequence of the probe design on the sequence dramatically constrains the search space for solutions. Furthermore we can use efficient string based data structures, such as a suffix tree, to efficiently explore compatible places on the design space relative to the 3D space. We named this combined combinatorial exploration followed by expansion ChromoTrace method (**Methods**).

Our simulations puts us in a position to explore these experimental and technical constraints in a controlled manner, since we can vary the design of the probe library both in terms of number of colors and spacing along the linear genome sequence and use the 3D chromatin simulations in our nuclear sets to predict the outcome if these probes were imaged by super-resolution microscopy. Since we know the underlying ground truth of sequence identity and probe color, we can test the hypothesis that the high resolution of 3D position determination and high reliability of color classification provided by super-resolution microscopy should provide enough information to find unique solutions for mapping back multicolored probe positions to the linear DNA sequence.

We created probe designs using a regular fixed spacing between probes (in our hands, a 10.8 kb spacing is optimal), resulting in an effective spatial imaging resolution of 4.3 x10^{-5} μm^3 volume which is well within the limits of super-resolution. We then convert the 3D positions of the simulated imaging data to a graph of potential adjacencies, using a threshold distance of 3672 nm relating to the maximum distance between two sequential probe positions in space (10.8 kb). The resulting potential adjacencies graph should in theory contain most of the true path of the probes along the genome plus spurious links of physically close but non-adjacent probes. We then create a suffix tree of depth 20 (the maximum length from the root to any leaf) containing the expected probe colors along the genome and capturing the two possible directions of reading the labels (p to q and q to p direction) resulting in a reversible suffix tree with path information for both forward and reverse genome direction. At depth of 20 the vast majority of leaves are unique (this depends on the number of colors, but often this achieved around depth 7). The algorithm then iteratively explores the adjacency graph to find regions with a unique solution of matching potentially physically adjacent color combinations with the genome sequence (**Fig 4A**)Once such anchor regions are found, the algorithm has a vastly reduced search space and extends them into the adjacency graph until it hits regions with high combinatorial complexity (**Fig 4B**), such as for example repetitive or highly compact regions.

To test the performance of ChromoTrace for determining the DNA path through the

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

nucleus we first loaded the labeling file into the reversible suffix tree and jointly searched the suffix tree and adjacency graph (x, y, z) to find unique sequences of labels (colors) found in both. We performed this analysis for all 100 synthetic nuclear sets, for all 22 probe designs, for all chromosomes separately as well as for the whole genome. We choose to use precision (specificity) and recall (or sensitivity) to assess algorithm performance. Recall is the ratio of the number of relevant records retrieved to the total number of relevant records and precision is the ratio of the number of relevant records retrieved to the total number of relevant and irrelevant records retrieved. Since the ground truth (labeling design) is known *a priori* in our simulation there is no ambiguity in how to measure performance.

Analyzing these 55,000 reconstruction attempts shows that the algorithm is highly precise (mean of 0.97 across all simulations), however the recall rate is much more variable (**Fig 5A**). This variability can largely be explained by two factors **i)** the number of colors available in the probe design **ii)** the density of probe positions in 3D space. For individual chromosomes the mean recall rate is approximately 50% when using a 10 color probe design, however for the same probe design genome wide the mean recall rate drops to 5.5% (**Fig 5A**). This reflects the increased number of ambiguous sequence paths available when the spatial search space is more densely packed, due to labeled sequences from physically close chromosome territories.

Fig5. Reconstruction performance for the main simulations. This figures shows the reconstruction algorithms performance in terms of the relationship between precision and recall given the number of colors in the probe design. Figure A shows Recall against precision genome wide (circles) and for chromosome 20 only (triangles). Precision is good for both genome and chromosome scale regions for all the different probe designs whereas recall is much more dependent on the number of available colors and improves as the number of colors is increased. Figure B shows the total number of contacts in 100 kb windows against the area under the precision-recall curve given the number of colors in the probe design. Regions with a higher packing density of probes are more difficult to reconstruct and the PR-AUC values are strongly negatively correlated with the total number of contacts for all different probe designs. The critical density where the PR-AUC values from the mean probe design (14 colors) drops below 0.5 is 61 or more total contacts within a 100 kb window relating to approximately 32% of possible spaces being occupied.

As expected the mean number of contacts for synthetic chromatin paths per chromosome are highly correlated to chromosome length (**Fig 3A**). To assess the reconstruction performance in dependence of the spatial probe position density (i.e. chromatin compaction) we show the area under the precision-recall curve values (PR AUC) against the total number of intra-chromosomal contacts in 100 kb windows across all autosomes and for all probe designs (**Fig 5B**). The contacts are defined as the total number of occupied spaces around each labeled probe position given the adjacency graph distance threshold. There is a clear trend for increased PR AUC values for probe designs with a greater number of colors irrespective of chromatin density. Across all probe designs there is a marked drop in performance as the chromatin density increases, and this drop is much sharper for probe designs with fewer colors (**Fig 5B**). The chance for creating ambiguous sequence paths is reduced as the number of unique color options increases, and given the strict (single-point) distance threshold of 3672 nm there will be 19 spaces around each probe that could potentially be occupied by another color. Exactly which colors these spaces may contain is unknown (3D folding of chromatin) so even for probe designs with greater than 19 colors there will almost certainly be at least one ambiguous

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequence path extension possible for a probe that has a relatively large proportion of its adjacent spaces occupied. The point at which the mean PR AUC values drop below 0.5 is 61 or more total contacts within a 100 kb window (**Fig 5B**), relating to approximately 32% of possible spaces in 3D space being occupied and a critical compaction of 31.1 kb within 1.48 x10^{-6} μm^3 nuclear volume for the mean probe design (14 colors). The amount of chromatin compaction that can be tolerated at the 0.5 PR AUC level is dependent on the number of colors available to the labeling design, with designs containing larger numbers of colors allowing more compacted genomic regions to be resolved.

Overall for probe designs with 7 or more colors, and for chromosomal scale regions we provide highly accurate reconstructions for up to 97% of the probes (mean=72%, sd=11%). Unsurprisingly smaller chromosomes can be more completely reconstructed, and the larger the number of colors the better the reconstruction.

Robustness and error tolerance

Real experimental super-resolution data will contain noise, likely from two major sources, firstly missing probes due to hybridization failure and secondly mislabeled probes, either due to chemical mislabeling or crosstalk between different dyes in the super-resolution microscope. To assess the performance of the reconstruction algorithm in the presence of errors we simulated 99 datasets for each error mode, containing error rates ranging from 1% to 99%, across all 22 probe designs for the 100 simulated nuclear sets, for all chromosomes separately as well as for the whole genome (a total of over 5.4 Million simulations).

For all probe designs the proportion of mislabeled probes has a dramatic effect on the reconstruction precision and we observe a clear decrease in precision as the proportion of probes with the wrong color is increased (**Fig 6A**). At only 10% mislabeled probes for the 24 color probe designs the mean precision is 0.9 (sd=0.012), dropping to 0.85 (sd=0.017) for 11 colors and to 0.5 (sd=0.064) for 3 colors. Recall rates are even more strongly effected by the proportion of mislabeled probes, starting from a maximum recall rate of approximately 0.8 for the 24 color probe designs with no mislabeled probes, recall rates drop sharply for all probe designs as the proportion of mislabeled probes increases (**Fig 6B**). At 10% mislabeled probes for the 24-color probe designs the mean recall is 0.34 (sd=0.026), dropping to 0.23 (sd=0.022) for 11 colors and to 0.044 (sd=0.012) for 3 colors. Above 60% of mislabeled probes both precision and recall is reduced below useful levels. The rapid drop in performance for recall compared to precision is not unexpected considering the suffix tree matching used was exact, did not allow for gaps and that the process for simulating noise due to mislabeled probes put them in random positions along the genome, where they will consequently terminate search paths.

Fig6. Robustness to missing and mislabeled probes. The relationship between the amount of error for two different modes (missing and mislabeled probes) and the overall reconstruction performance given the number of colors in the probe design is displayed in panels A through D. The number of colors in the probe design is indicated using different shades of black-blue. Panels A and C show the proportion of error against precision for mislabeled and missing probe errors respectively and panels B and D show the proportion of error against recall. The relationship between percentage error and recall is very similar for the two different error modes with a strong decrease in recall for modest increases in error for all different probe designs. Whereas the tolerance to error for precision shows a clear difference between the two modes with missing probes being far less likely to introduce incorrect path extensions than mislabeled probes.

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For missing probes the relationship between recall and percentage of errors is very similar (**Fig 6B and 6D**). This is not surprising since either removing or replacing probes with a wrong color in a sequence of colors is likely to stop the extension of correct paths at a similar rate (distance threshold). Precision however, only starts to drop at a much higher percentage of missing, compared to mislabeled probes (**Fig 6A and 6C**). This suggests that the chance of creating an error in path extension when removing probes is lower than if mislabeled probes are present. If DNA paths were linear in 3D space this would be entirely expected as the distance threshold between sequential probes would ensure that most paths are not incorrectly extended across missing probe locations, while mislabeled probes will not only terminate extension but also cause mismatches to the genome sequence. These results suggest that removing relatively large numbers of probes is unlikely to cause incorrect path extensions across a majority of the simulated chromatin space. However, it is worth noting that the reconstruction performance is greatly influenced by chromatin compaction (**Fig 5B**), and since the number of extension choices for any given path will be higher in dense regions, the presence of missing probes in highly compacted regions can also lead to incorrect path extension.

Encouragingly even for the probe designs with the lowest numbers of colors (3) precision remains greater than 0.75 with a missing probe rate of 25%. Furthermore, precision is also relatively robust to mislabeled probe errors, remaining above 0.75 with more than 10% of mislabeled probes for probe designs with greater than 7 colors. As expected recall is far more sensitive to error and there is only a marginal difference observed between the two error modes. When maintaining a recall rate above 0.5, error tolerance improves by approximately 0.2% for both error modes with each additional color and reaches a maximum of 5% for the probe designs with the highest number of colors (24).

Differences in chromatin packing density

We are interested in creating the right computational complexity for the reconstruction problem. It is unclear how much of the available volume chromatin occupies locally within the nucleus under physiological conditions, but the literature suggests nucleosome concentrations of $142 \pm 28\ \mu M$ with nucleosomes every 185 bp in HeLa cells leading to a packing density of 10 % when assuming a nucleosome volume of 1296 nm^3 [27]. To be conservative, our simulation used a higher than average density of chromatin, with 34% of the available local volume occupied by chromatin (545 thousand points per genome from a 1.59 million point grid). An increased density of the SAWs will result in a harder reconstruction problem, because a higher number of occupied adjacent spaces within the simulation lead to an increase in the number of ambiguous choices for path extension.

To assess the effect of lowering the chromatin density we performed some additional simulations by omitting the nucleolus and doubling the number of grid points resulting in filling approximately 6.9% of the nuclear volume with chromatin (982 thousand points per genome from a 14.1 million point grid). Unsurprisingly these SAWs are less densely packed, an effect that can be visualized by looking at the proportion of adjacent spaces that are occupied, given the distance threshold, for all labeled genomic locations in the simulations (**Fig 7**). While in our original simulations the median proportion of occupied spaces around each probe position from the labeling design is 0.52 (**Fig 7A**), in the lower density simulation this is decreased to 0.37 (**Fig 7B**).

Fig7. Differences in simulation packing densities. Reconstruction performance in when decreasing the packing density of the simulations. (A-B) For all positions across the simulations, the proportion of directly adjacent spaces that are occupied for the

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

new (blue) and original (red) simulations respectively. The distribution is left shifted for the new simulations compared to the original and the median number of occupied spaces is reduced reflecting a decrease in density. (C) Genome wide performance of the reconstruction algorithm for the new (triangles) and original (circles) simulations in terms of precision and recall given the number of colours in the probe design.

To test how this affects the reconstruction performance, we generated 100 synthetic nuclear sets using the approach described above and produced 22 different probe designs containing 3 to 24 colors for the lower density simulations. We then reconstructed using the ChromoTrace algorithm for all synthetic data sets for each chromosome separately and for the whole genome. As expected performance, in terms of both precision and recall, is improved for the less densely packed simulation (**Fig 7C**). The genome wide mean precision remains high (greater than 0.9) for all probe designs, increasing gradually with more colors and reaching a maximum of 0.98 for the probe designs with 24 colors. The difference in recall is much more pronounced with mean recall rates of 0.62, 0.37 and 0.035, compared to 0.19, 0.06 and 0.003, for probe designs with 24, 11 and 3 colors for the lower density compared to the higher density simulations respectively. Importantly when comparing the lower to the higher density simulations the recall rate is improved by a mean factor of 5.5 across all different color probe designs (**Fig 7C**). This marked improvement in sensitivity reflects the decreased number of occupied adjacent 3D spaces around each individual probe position and consequently a reduced number of ambiguous sequence path extension choices when lowering the density of the simulated chromatin paths (**Fig 7A and 7B**).

Overall across all probes designs the lower density simulation has a genome wide mean recall rate of 0.39 compared to 0.09 for the higher density simulation, a more than 4 fold difference. Interestingly, the reconstruction performance increase is much less striking for individual chromosomes with only marginal differences observed in both precision and recall (**Supplementary Figure 1**) suggesting that even in highly compacted chromatin regions, additional challenges arise due to the close proximity of neighboring chromosome territories and reflect the increased combinatorial complexity when adding more probes into the reversible suffix tree.

Supplementary Fig 1 3D view of the nucleus simulation when decreasing the packing density. (A) One whole genome simulation with each chromosome is drawn in a different colour confined within the defined nuclear space. (B) One copy of chromosome 1 showing the different type of compaction across the chromsome.

Discussion

Although we simulated reasonable chromatin paths and deliberately used a challenging density of chromatin in the nucleus, our simulation of 3D chromosome folding is coarse grained and does at this time not take the known structural heterogeneity of chromatin packing of different genomic sequences into account, for example eu- and heterochromatic domains or TADs. It is therefore necessary to consider how such structural heterogeneity would affect the reconstruction problem. For a given packing density, such structures should lead to one of two outcomes, firstly that the entire chromosome (or probed region of interest) is overall more compact that simulated, leading to a significantly smaller volume of the chromosome territory. This would effectively reduce the amount of resolvable spatial information present for the reconstruction. Such a result would be disappointing in terms of the reconstruction algorithm, but fascinating in terms of how such chromosomal domains are created and maintained. However, the extended conformation

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of many chromosomes seen previously [28], along with the distribution of their contacts to the nuclear lamina [29], suggest that overall compaction is an unlikely configuration, except for specific cases such as mitotic chromosomes or the inactive X chromosome. The second outcome is that the more highly packed regions than we simulated are interspersed with more extended regions. The extended regions would be easier to reconstruct, as the better resolved 3D information will be more accurately able to place these regions to a unique position on the genome. At the extreme of this model one would have a series of resolvable linkers with interspersed globules of packed chromatin that would not be resolvable. In such a scenario integration with the HiC data or other contact maps, whose resolution is good in these more dense regions [30] would be very interesting.

On the other hand, when the density of chromatin in the nucleus is lower, the reconstruction improves dramatically in terms of recall. In experimental HiC data if unusual numbers of contacts are observed relative to chromosome size it may be indicative of biological processes effecting chromatin condensation [31]. It is feasible to resolve a large fraction of chromosomal scale regions with a resolution of 10.8 kb and reconstruction at this level would provide very high-resolution chromosomal scale chromatin maps (including the internal structure of TADs, TAD boundaries and inter-TAD regions). Even if the very fine details of high density chromatin structures remain challenging with the currently available imaging technology, the spatial information provided by even partial reconstruction of the chromatin path is certain to increase our understanding of how chromosome folding and partitioning is related to active processes such as gene expression [32] as chromatin density is thought to be lower in active and higher in inactive TADs [33]. Interestingly, in addition to resolving chromatin paths and map interactions within resolved paths, the reconstruction completeness of ChromoTrace will provide an indirect measure of density and thus allow chromatin state inference across chromosomes.

The other important consideration is the number of differentiatable fluorescence colors that the reconstruction requires. The number of flourophores compatible with 3D super-resolution microscopy and in-situ hybridization probes is currently limited to about three dyes that can be reliably spectrally separated if imaged at the same time. Since DNA in situ probes can be coupled to more than one flurophore, combinatorial labeling can create different color ratios. In our simulations, up to 10 colors for simultaneous detection could easily be generated in this manner, however will also introduce noise due to chemical labeling errors (the chance by which a probe will be labeled with a different color ratio than intended) which would lead to wrong probe assignments. However, since any given color will have only a finite set of possible neighboring mistakes with associated error rates, a substitution matrix of possible errors can easily be integrated into both the extension phase and exploration phase of the suffix tree, changing the formulation of the problem into a likelihood model of seeing the 3D position of probes (Data) given a certain path labeling. In addition, recent advances in labeling techniques such as the 'Exchange-PAINT' method now allow sequential hybridization and image capture, allowing to separate up to 10 pseudocolors based on a single dye in time [21]. This labeling technology requires long super-resolution image acquisition times, but could massively increase the number of probes available for the reconstruction algorithm. For example, a binary code with 2 colors and 10 labeling rounds could distinguish in the region of 2^{10} labels, which would make reconstruction almost trivial. It is therefore very likely that a well-designed combination of spectral and temporal multiplexing of fluorescent dyes, will make it possible to generate image data with sufficiently large numbers of differently 'colored' probes and reasonable data acquisition times to allow high resolution reconstruction of the chromatin paths for individual chromosomes within the nucleus. Our ChromoTrace algorithm should prove valuable to guide the optimal

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

design of such probes, since it allows to simulate the effect of different designs on the reconstruction performance rapidly in silico.

Conclusion

In this paper we proposed a novel algorithm, ChromoTrace, to in theory, leverage super-resolution microscopy of thousands to millions of in situ genome sequence probes to provide accurate physical reconstructions of 3D chromatin structure at the chromosomal scale in single human cells. To test this algorithm we have made simulations of chromatin paths in realistic nuclear geometries, and explored different labeling strategies of in situ probes. Our study shows, that near complete resolution of a chromosome with 10 kb resolution can be achieved with realistic microscope resolution and fluorescent probe multiplexing parameters. Extensions to this method such leveraging between nucleus consistency effects and using a likelihood-based scheme will allow even more sophisticated modeling of experimental error sources in the future.

There is currently no suitable experimental data in to substantiate this work; this is firmly a theoretical exploration of the possibility to achieve this and the constraints any experimental method would need to satisfy for successful reconstruction. For example, it is clear that minimizing mislabeling is more important than minimizing missing probes. However, our simulations are based on known and realistic experimental parameters where available, and we have tested our method under challenging chromatin density levels and aggressive error models of missing or misreported data. Our algorithm and assumptions are compatible with leading super-resolution techniques; in particular our method assumes isotropic resolution of the probes, which has been shown using methods such as direct stochastical optical reconstruction microscopy combined with interference [21,34]. Nevertheless real experimental data will likely have properties that we have not anticipated. Some of these properties, such as systematic error behavior, or changes in resolution across the nucleus might hinder our reconstruction. On the other hand, properties such as structured heterogeneity in packing density and cell-to-cell structure conservation are likely to improve our ability to reconstruct. Our reconstructions based on single cell image data are initially most likely to work in a patchwork manner across a chromosome, and will be very complementary to the contact based maps based on HiC or promoter-capture HiC [35]. Combining super resolution imaging and contact mapping should provide fundamentally new insights into chromatin organization and packing within the nucleus.

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Kornberg RD, Lorch Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell. 1999;98(3):285-94.

2. Misteli T. Concepts in nuclear architecture. Bioessays. 2005;27(5):477-87.

3. Parada LA, McQueen PG, Misteli T. Tissue-specific spatial organization of genomes. Genome Biol. 2004;5(7):R44.

4. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature. 1997;389(6648):251-60.

5. Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell. 2004;118(5):555-66.

6. Brown JM, Green J, das Neves RP, Wallace HA, Smith AJ, Hughes J, et al. Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. J Cell Biol. 2008;182(6):1083-97.

7. Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet. 2001;2(4):292-301.

8. Cremer T, Cremer M. Chromosome territories. Cold Spring Harb Perspect Biol. 2010;2(3):a003889.

9. Naumova N, Dekker J. Integrating one-dimensional and three-dimensional maps of genomes. J Cell Sci. 2010;123(Pt 12):1979-88.

10. Dekker J. The three 'C' s of chromosome conformation capture: controls, controls, controls. Nat Methods. 2006;3(1):17-21.

11. 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.

12. Pombo A, Dillon N. Three-dimensional genome architecture: players and mechanisms. Nat Rev Mol Cell Biol. 2015;16(4):245-57.

13. Castillo O, Rojas H, Dominguez Z, Angles-Cano E, Marchi R. Endothelial fibrinolytic response onto an evolving matrix of fibrin. BMC Hematol. 2016;16:9.

14. Meng XQ, Dai YY, Jing LD, Bai J, Liu SZ, Zheng KG, et al. Subcellular localization of Pyk2 during oocyte fertilization and early-embryo development in mice. J Reprod Dev. 2016.

15. Schermelleh L, Heintzmann R, Leonhardt H. A guide to super-resolution fluorescence microscopy. J Cell Biol. 2010;190(2):165-75.

16. Szymborska A, de Marco A, Daigle N, Cordes VC, Briggs JA, Ellenberg J. Nuclear pore scaffold structure analyzed by super-resolution microscopy and particle averaging. Science. 2013;341(6146):655-8.

17. Fabre PJ, Benke A, Joye E, Nguyen Huynh TH, Manley S, Duboule D. Nanoscale spatial organization of the HoxD gene cluster in distinct transcriptional states. Proc Natl Acad Sci U S A. 2015;112(45):13964-9.

bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

18. Ricci MA, Manzo C, Garcia-Parajo MF, Lakadamyali M, Cosma MP. Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. Cell. 2015;160(6):1145-58.

19. Wani AH, Boettiger AN, Schorderet P, Ergun A, Munger C, Sadreyev RI, et al. Chromatin topology is coupled to Polycomb group protein subnuclear organization. Nat Commun. 2016;7:10291.

20. Boettiger AN, Bintu B, Moffitt JR, Wang S, Beliveau BJ, Fudenberg G, et al. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature. 2016;529(7586):418-22.

21. Huang F, Sirinakis G, Allgeyer ES, Schroeder LK, Duim WC, Kromann EB, et al. Ultra-High Resolution 3D Imaging of Whole Cells. Cell 2016;166(4):1028-40.

22. Bienko M, Crosetto N, Teytelman L, Klemm S, Itzkovitz S, van Oudenaarden A. A versatile genome-scale PCR-based pipeline for high-definition DNA FISH. Nat Methods. 2013;10(2):122-4.

23. Jungmann R, Avendano MS, Woehrstein JB, Dai M, Shih WM, Yin P. Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. Nat Methods. 2014;11(3):313-8.

24. Flory PJ. Principles of Polymer Chemistry. Science. 1953;119.

25. Barbieri M, Chotalia M, Fraser J, Lavitas LM, Dostie J, Pombo A, et al. Complexity of chromatin folding is captured by the strings and binders switch model. Proc Natl Acad Sci U S A. 2012;109(40):16173-8.

26. Wang C, Liu C, Roqueiro D, Grimm D, Schwab R, Becker C, et al. Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. Genome Res. 2015;25(2):246-56.

27. Weidemann T, Wachsmuth M, Knoch TA, Muller G, Waldeck W, Langowski J. Counting nucleosomes in living cells with a combination of fluorescence correlation spectroscopy and confocal imaging. J Mol Biol. 2003;334(2):229-40.

28. Lassadi I, Kamgoue A, Goiffon I, Tanguy-le-Gac N, Bystricky K. Differential chromosome conformations as hallmarks of cellular identity revealed by mathematical polymer modeling. PLoS Comput Biol. 2015;11(6):e1004306.

29. Harr JC, Luperchio TR, Wong X, Cohen E, Wheelan SJ, Reddy KL. Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and A-type lamins. J Cell Biol. 2015;208(1):33-52.

30. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16:259.

31. Martin RM, Cardoso MC. Chromatin condensation modulates access and binding of nuclear proteins. FASEB J. 2010;24(4):1066-72.

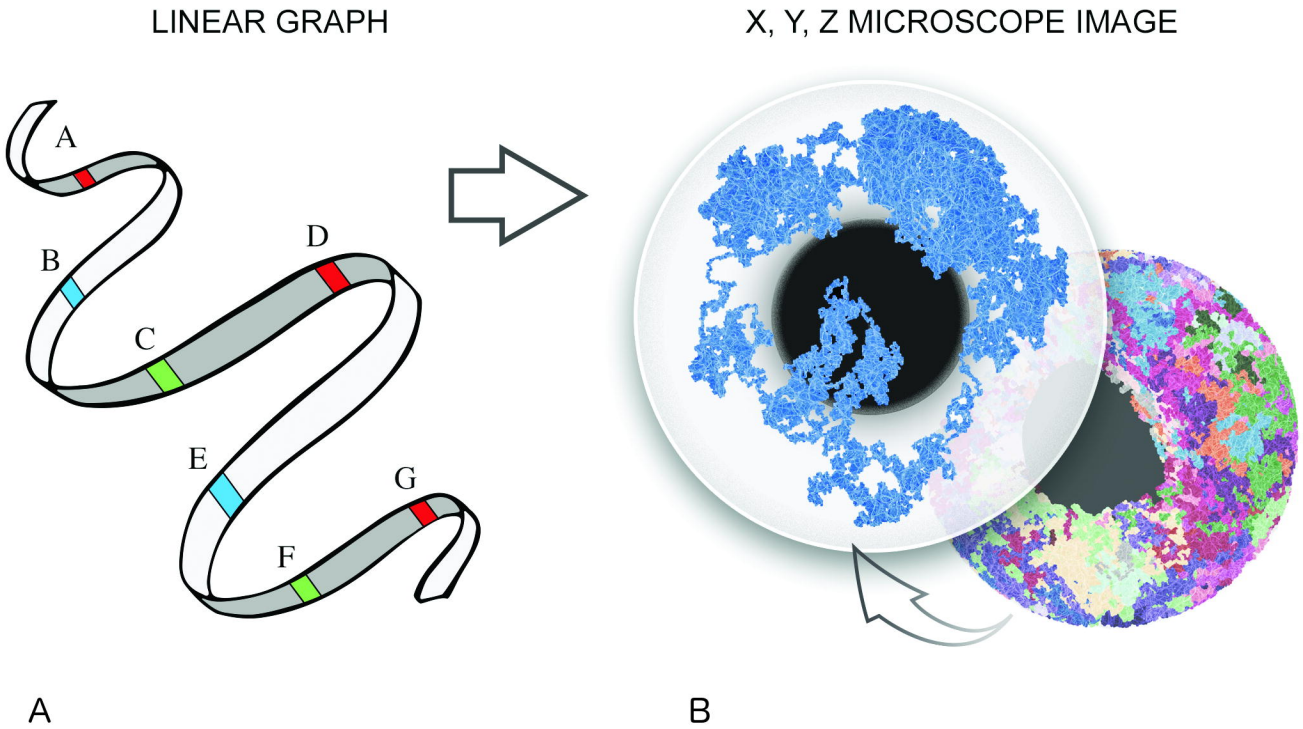
32. Criscione SW, De Cecco M, Siranosian B, Zhang Y, Kreiling JA, Sedivy JM, et al. Reorganization of chromosome architecture in replicative cellular senescence. Sci Adv. 2016;2(2):e1500882.

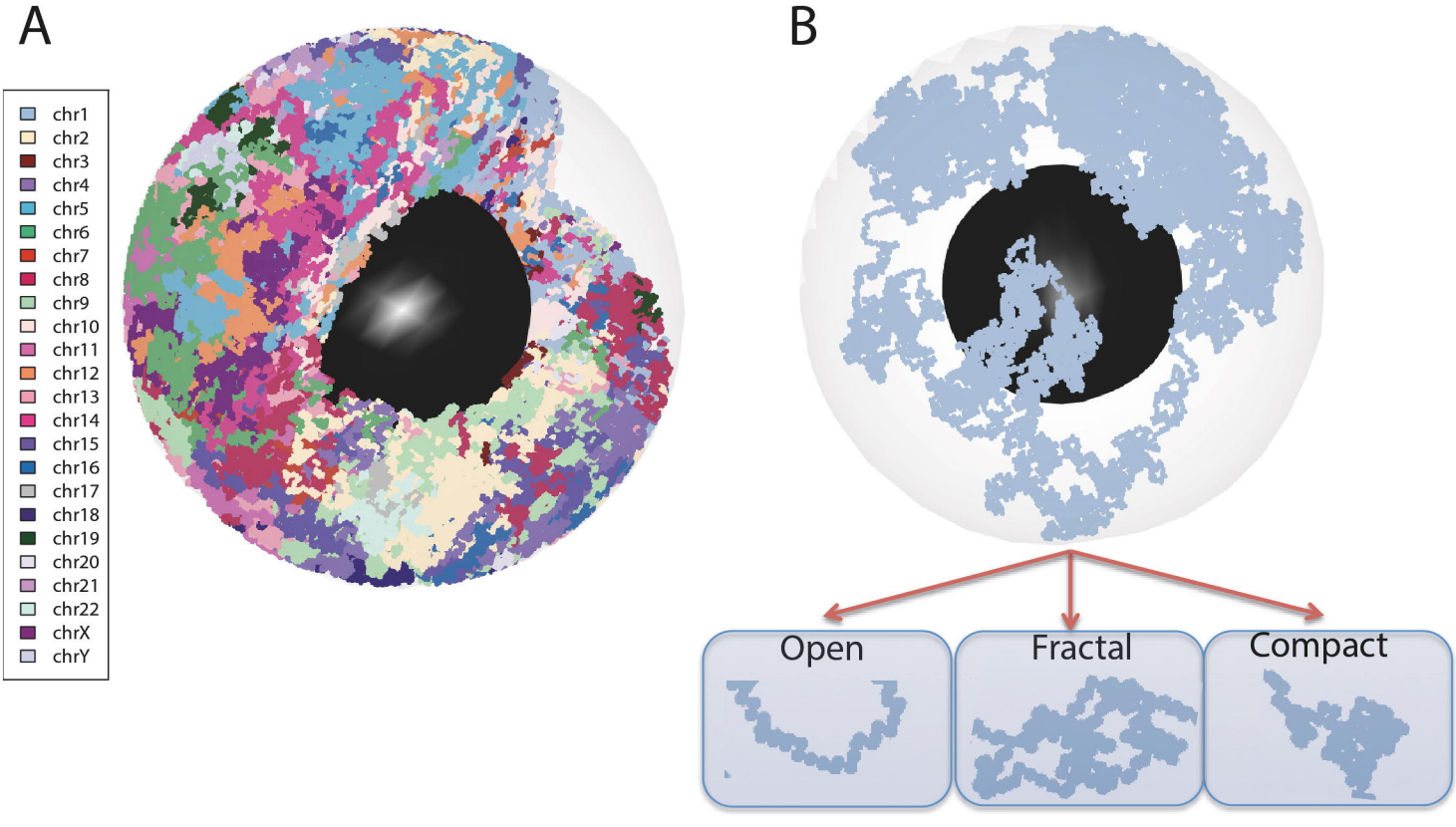
33. Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. Genome Res. 2016;26(1):70-84.

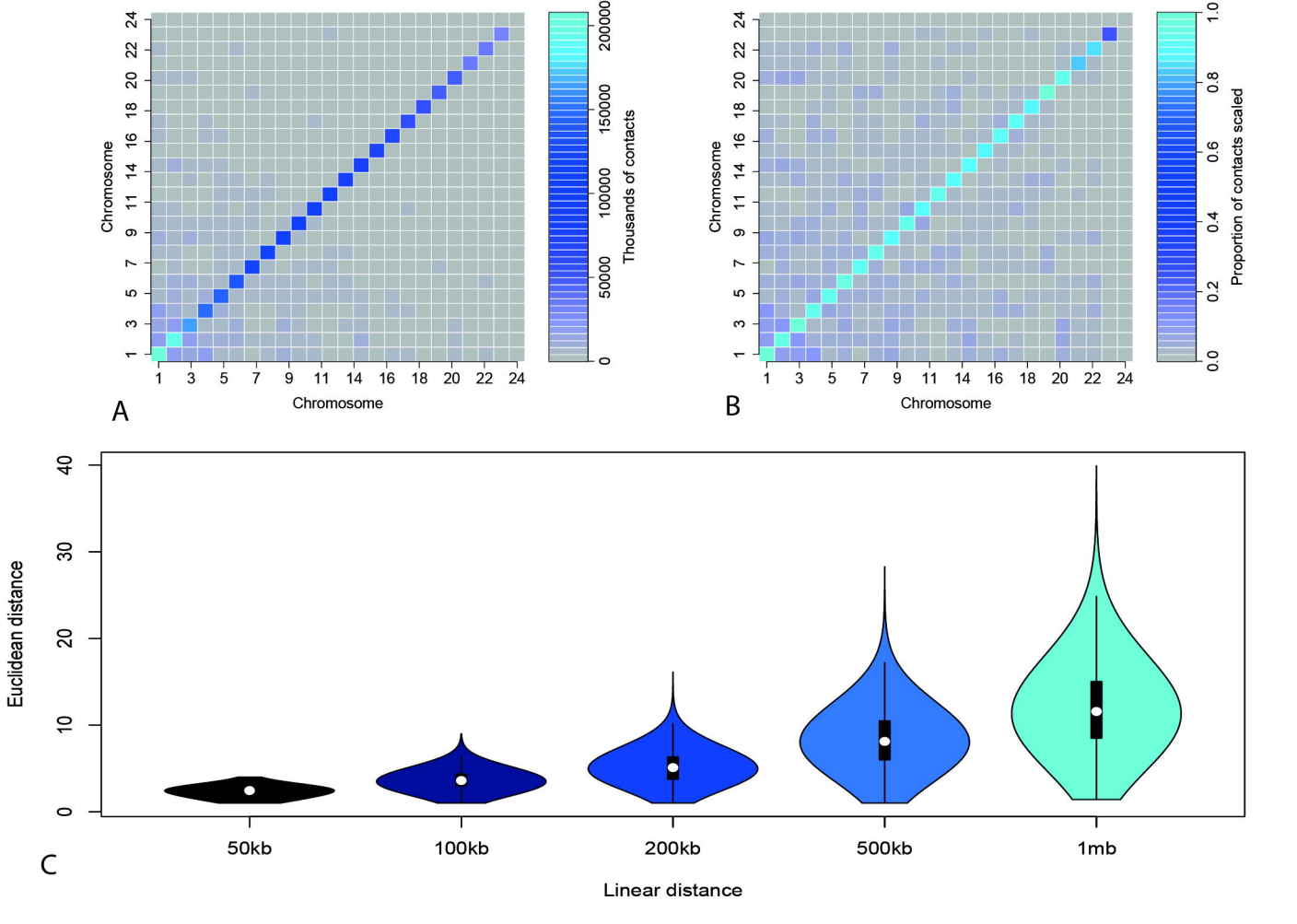
bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

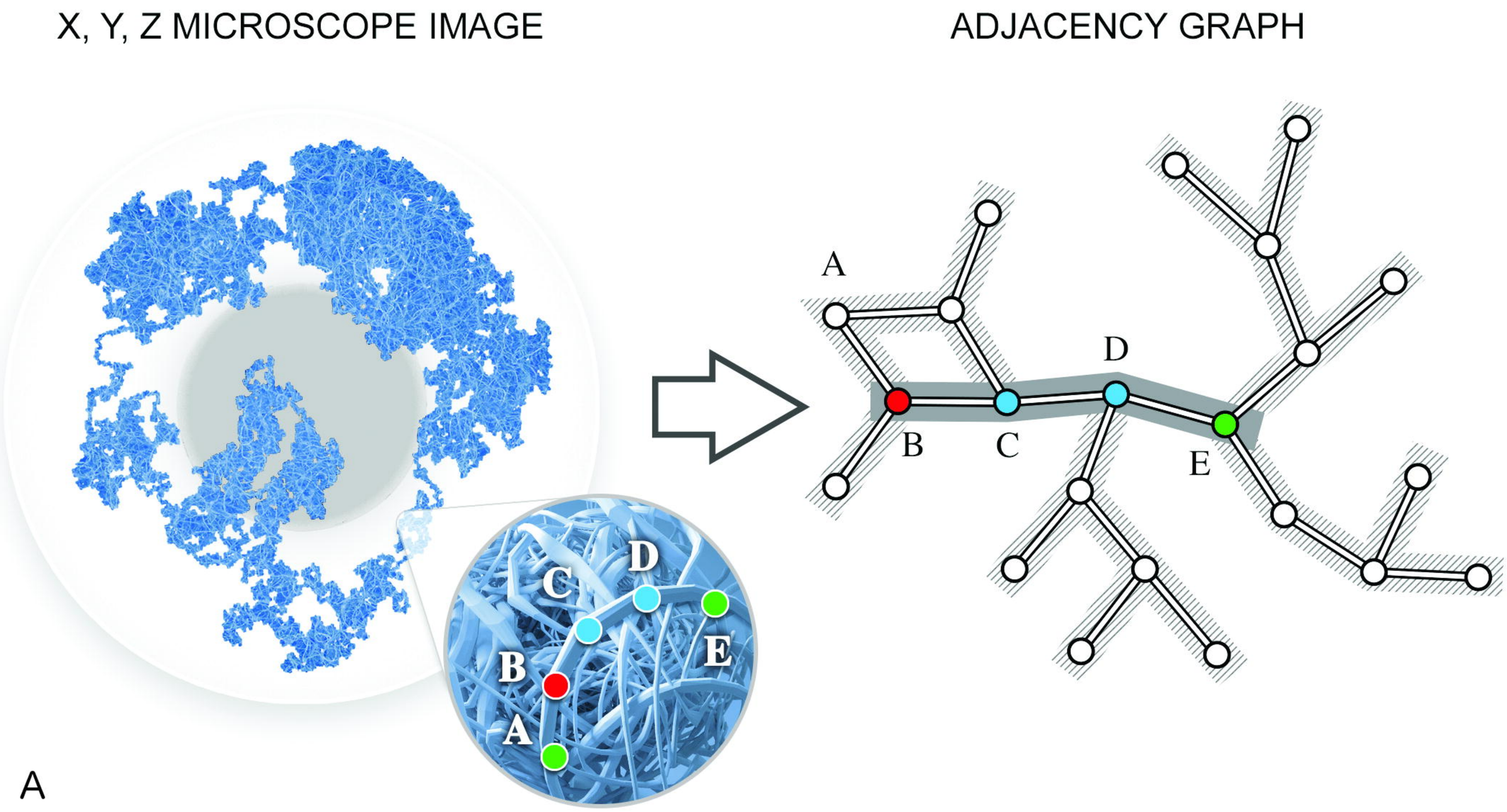
34. Hell SW. Toward fluorescence nanoscopy. Nat Biotechnol. 2003;21(11):1347-55.

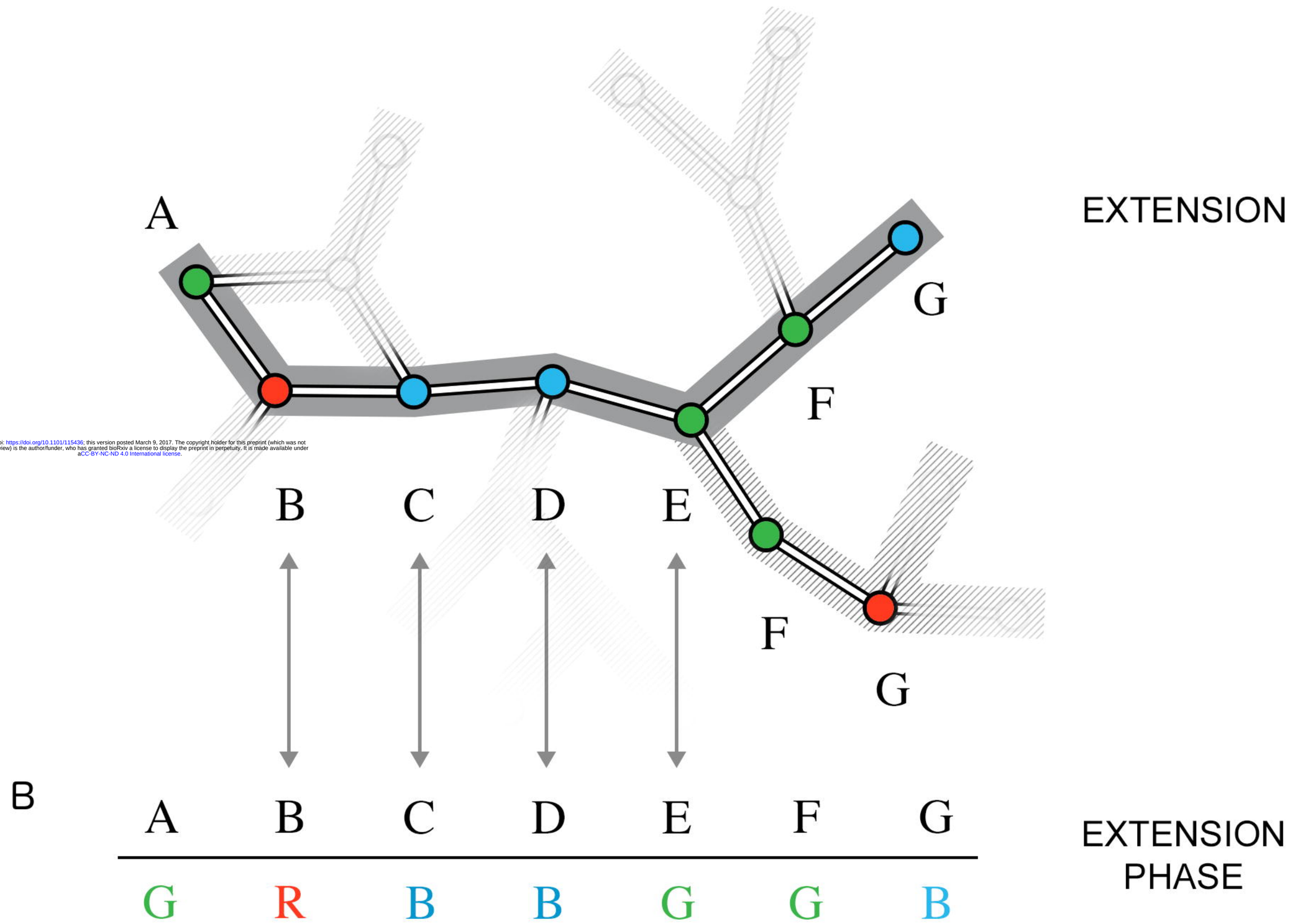
35. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011;43(11):1059-65.


LINEAR GRAPH
X, Y, Z MICROSCOPE IMAGE
A
B
C
D
E
F
G
A
B

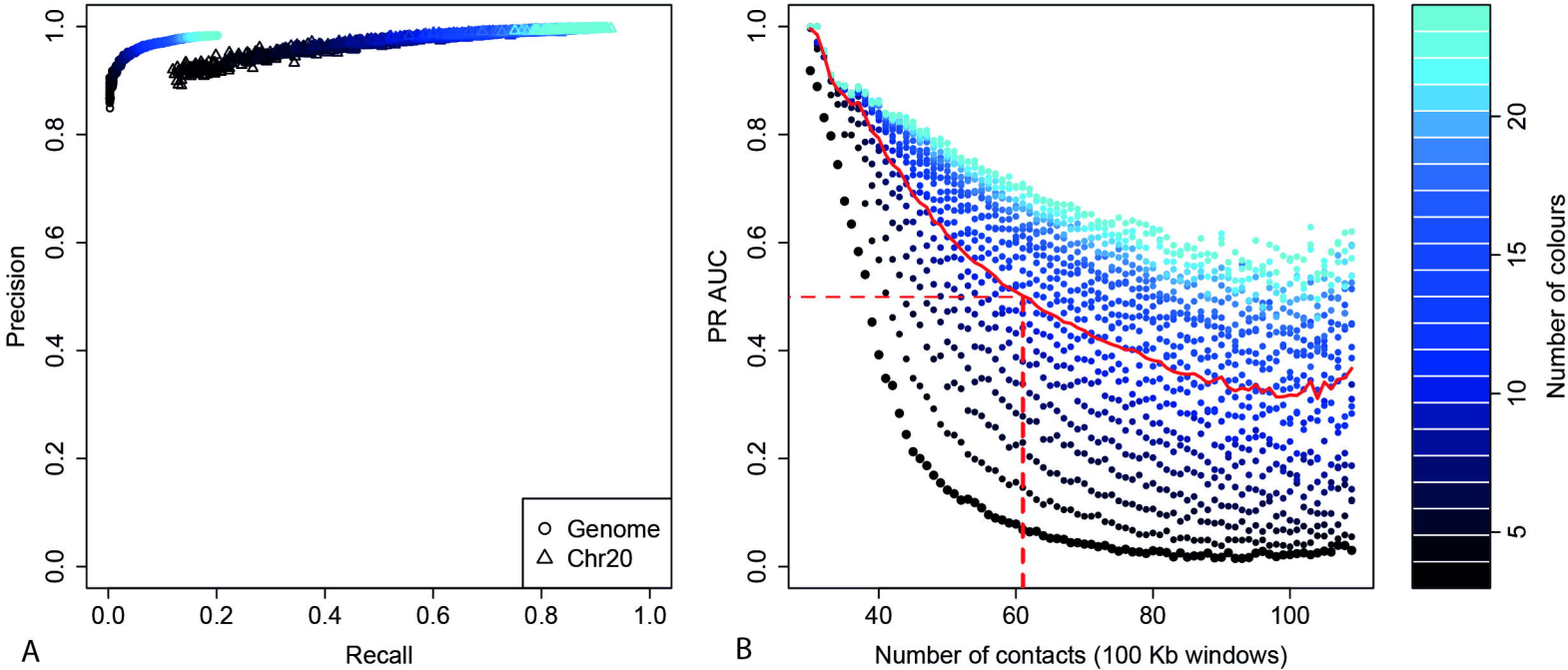

A
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
B
Open
Fractal
Compact

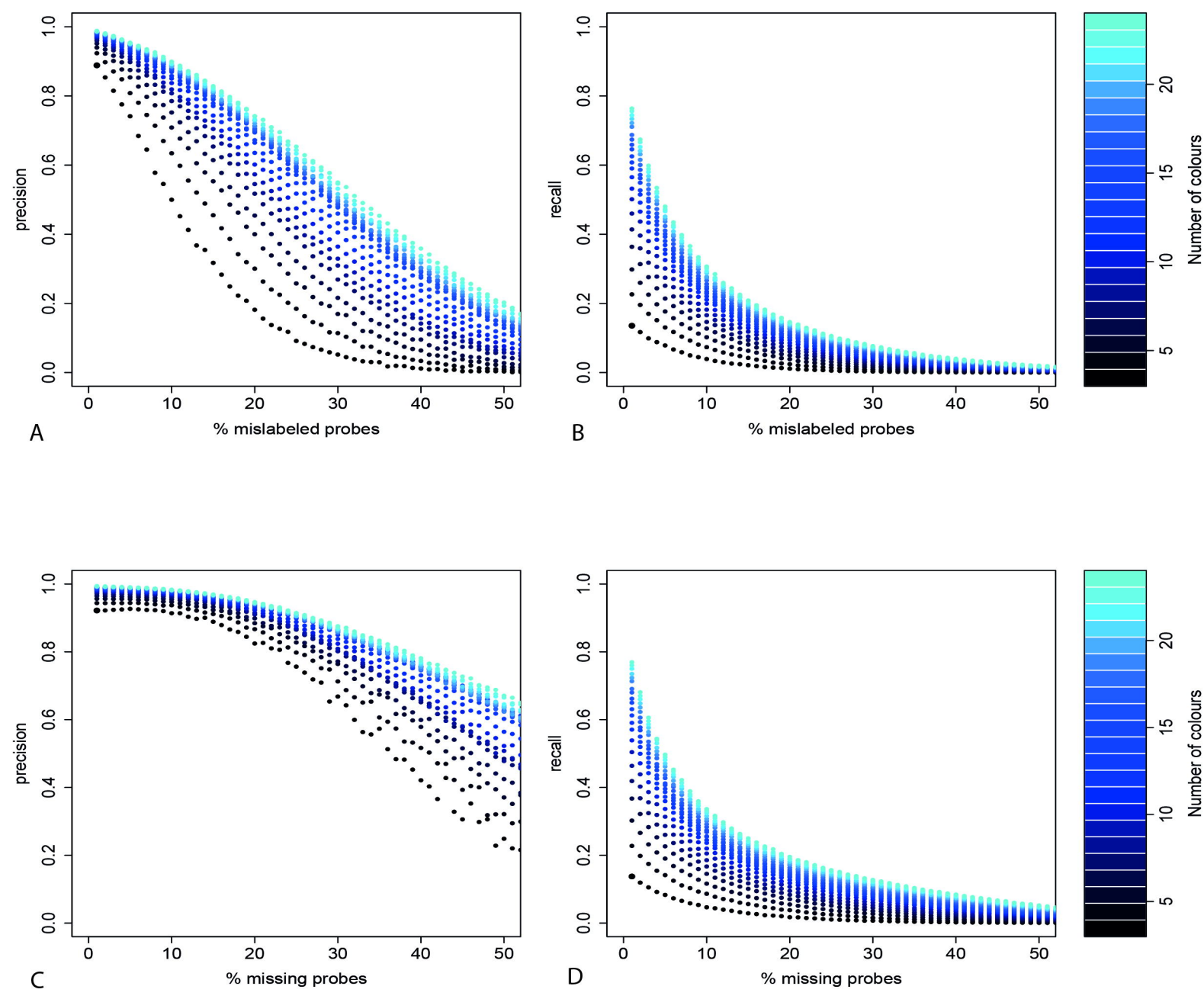

A
Chromosome
Chromosome
Thousands of contacts
B
Chromosome
Chromosome
Proportion of contacts scaled
C
Euclidean distance
50kb
100kb
200kb
500kb
1mb
Linear distance

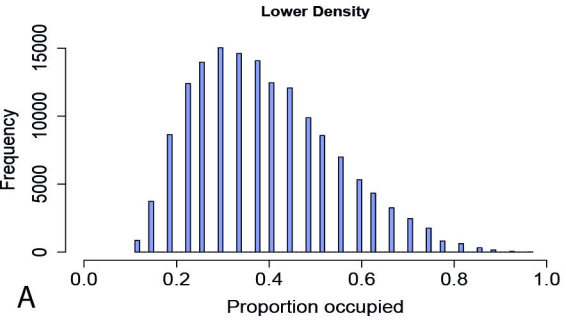


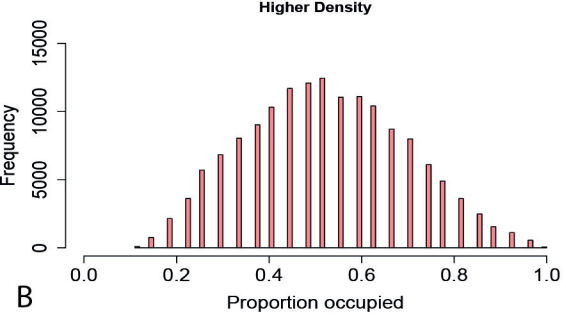


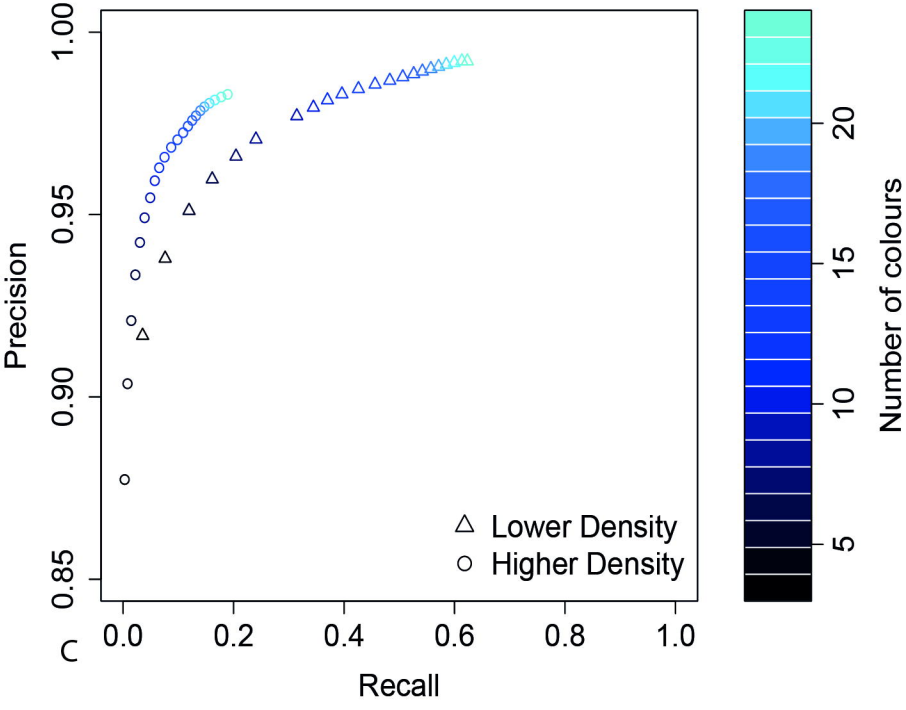
bioRxiv preprint doi: https://doi.org/10.1101/115436; this version posted March 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


Precision
Recall
Genome
Chr20
A
PR AUC
Number of contacts (100 Kb windows)
B
Number of colours


precision
% mislabeled probes
A
recall
% mislabeled probes
B
Number of colours
precision
% missing probes
C
recall
% missing probes
D
Number of colours


Lower Density
Frequency
0
5000
10000
15000
0.0
0.2
0.4
0.6
0.8
1.0
Proportion occupied
A


Higher Density
Frequency
0
5000
10000
15000
0.0
0.2
0.4
0.6
0.8
1.0
Proportion occupied
B


Precision
0.85
0.90
0.95
1.00
0.0
0.2
0.4
0.6
0.8
1.0
Recall
C
Lower Density
Higher Density
Number of colours
5
10
15
20