nature biotechnology

Genome architectures revealed by tethered chromosome conformation capture and population-based modeling

Reza Kalhor[1,2], Harianto Tjong[1], Nimanthi Jayathilaka[1,2], Frank Alber[1] & Lin Chen[1,3,4]

We describe tethered conformation capture (TCC), a method for genome-wide mapping of chromatin interactions. By performing ligations on solid substrates rather than in solution, TCC substantially enhances the signal-to-noise ratio, thereby facilitating a detailed analysis of interactions within and between chromosomes. We identified a group of regions in each chromosome in human cells that account for the majority of interchromosomal interactions. These regions are marked by high transcriptional activity, suggesting that their interactions are mediated by transcriptional machinery. Each of these regions interacts with numerous other such regions throughout the genome in an indiscriminate fashion, partly driven by the accessibility of the partners. As a different combination of interactions is likely present in different cells, we developed a computational method to translate the TCC data into physical chromatin contacts in a population of three-dimensional genome structures. Statistical analysis of the resulting population demonstrates that the indiscriminate properties of interchromosomal interactions are consistent with the well-known architectural features of the human genome.

The three-dimensional (3D) organization of the eukaryotic genome plays important roles in nuclear functions[1,2]. However, few structural details of chromatin organization have been delineated at the genomic scale. For instance, individual chromosomes are localized in spatially distinct volumes known as the chromosome territories[3], which tend to occupy preferential positions with respect to the nuclear periphery[4,5]. Moreover, the territories of different chromosomes form extensive interactions[6], and high-density gene clusters can extend outside of the bulk of their chromosome's territory[7]. Nevertheless, the internal organization of chromosome territories and the mechanisms that govern the interactions between them are not well understood.

Chromosome conformation capture–based techniques have emerged as powerful tools for mapping chromatin interactions[8–16]. The genome-wide application of these techniques has revealed that functional activity can determine the association preferences of loci within each chromosome[10]. Further understanding of the spatial organization of chromosomes, however, is limited by several factors. Low signal-to-noise ratios in conformation capture experiments compromise their ability to map low-frequency interactions, especially those between chromosome territories. Additionally, the data represent an ensemble average of genome structures in the cell population, wherein individual structures may substantially differ from each other[17–19]. Coupled with the enormous size of the genome, this heterogeneity of genome architecture makes translating conformation capture data into 3D structural models challenging. As a result, even as genome-wide conformation capture data have been used to propose theoretical folding models[10], they have not yet been employed for determining the corresponding 3D structures of the entire genome in mammalian cells.

For the genome-wide mapping of chromatin contacts, we have developed the TCC technology, a modified conformation capture method in which key reactions are carried out on solid phase instead of in solution. This tethering strategy leads to higher signal-to-noise ratios, enabling an in-depth analysis of interchromosomal interactions. We show that a specific group of functionally active loci are more likely to form interchromosomal contacts and that most of these contacts are a result of indiscriminate encounters between loci that are accessible to each other. We also introduce a structural modeling procedure that calculates a population of 3D genome structures from the TCC data. We show that the calculated population reproduces the hallmarks of chromosome territory positioning in agreement with independent fluorescence *in situ* hybridization (FISH) studies. This population-based approach allows for a probabilistic analysis of the spatial features of the genome, a capability that can accommodate the wide range of cell-to-cell structural variations that are observed in mammalian genomes[17,20].

RESULTS

Detecting genome-wide chromatin contacts using TCC

To identify chromatin interactions using TCC (**Fig. 1**), we preserved native chromatin contacts by chemically crosslinking DNA and proteins. The DNA was then digested with a restriction enzyme, and, after cysteine biotinylation of proteins, the protein-bound fragments were immobilized at a low surface density on streptavidin-coated beads. The immobilized DNA fragments were then ligated while tethered to the surface of the beads. Finally, ligation junctions were purified, and ligation events were detected by massively parallel sequencing, a process that revealed the genomic locations of the pairs of loci that had formed the initial contacts (**Fig. 1**).

[1]Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, California, USA. [2]Program in Genetic, Molecular and Cellular Biology, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [3]Department of Chemistry, University of Southern California, Los Angeles, California, USA. [4]USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. Correspondence should be addressed to F.A. (alber@usc.edu) or L.C. (linchen@usc.edu).

Received 22 February; accepted 2 November; published online 25 December 2011; doi:10.1038/nbt.2057

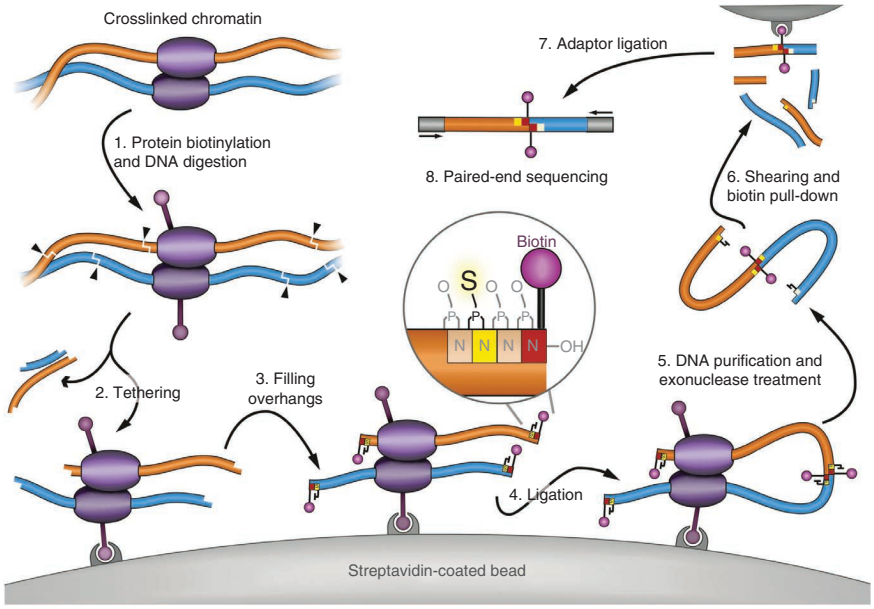

Figure 1 Overview of TCC. Cells are treated with formaldehyde, which covalently crosslinks proteins (purple ellipses) to each other and to DNA (orange and blue strings). (1) The chromatin is solubilized and its proteins are biotinylated (purple ball and stick). DNA is digested with a restriction enzyme that generates 5′ overhangs. (2) Crosslinked complexes are immobilized at a very low density on the surface of streptavidin-coated magnetic beads (gray arc) through the biotinylated proteins; noncrosslinked DNA fragments are removed. (3) The 5′ overhangs are filled in with an α-thio-triphosphate–containing nucleotide analog (the yellow nucleotide in the inset), which is resistant to exonuclease digestion, and a biotinylated nucleotide analog (the red nucleotide with the purple ball and stick in the inset) to generate blunt ends. (4) Blunt DNA ends are ligated. (5) Crosslinking is reversed and DNA is purified. The biotinylated nucleotide is removed from nonligated DNA ends using *Escherichia coli* exonuclease III whereas the phosphorothioate bond protects DNA fragments from complete degradation. (6) The DNA is sheared and fragments that include a ligation junction are isolated on streptavidin-coated magnetic beads, but this time through the biotinylated nucleotides. (7) Sequencing adaptors are added to all DNA molecules to generate a library. (8) Ligation events are identified using paired-end sequencing.

We applied TCC, using HindIII as the restriction enzyme, to map the chromatin contacts in GM12878 human lymphoblastoid cells (**Supplementary Table 1**). As an example of nontethered conformation capture, we also applied Hi-C[10] to the same cell line using identical cell counts and crosslinking conditions. The resulting contact frequency maps (**Fig. 2a**,**b** and **Supplementary Fig. 1a**) showed that TCC accurately reproduces the patterns observed in Hi-C results (Pearson's r for genome-wide comparison = 0.96, $P < 10^{-16}$). Additionally, the general features of genome-wide conformation capture data that were described previously[10] were also observed in our data (**Fig. 2a**,**b** and **Supplementary Fig. 1a**–**c**).

Improved signal-to-noise ratio in tethered libraries

One of the main sources of noise in conformation capture experiments is random intermolecular ligations between DNA fragments that are not crosslinked to each other[9,21]. Because randomly selected DNA fragments are more likely to originate from different chromosomes, these ligations tend to be overwhelmingly interchromosomal. Therefore, we measured the fraction of interchromosomal ligations in our tethered (TCC) and nontethered (Hi-C) HindIII libraries to compare their relative noise levels (**Fig. 2c**). In the tethered library, this fraction is almost half that of the nontethered library. We also compared the average difference between the observed interchromosomal contact frequencies in each library and those expected from completely random intermolecular ligations. This difference is twice as large in the tethered library compared to the nontethered library (**Supplementary Methods**). Together, these observations indicate that the noise from random intermolecular ligations is considerably lower in the tethered library.

We also generated tethered and nontethered libraries using the 4-cutter MboI instead of HindIII. MboI results in a higher concentration of, and shorter, DNA fragments, thereby increasing the probability of random intermolecular ligations. Consequently, the fraction of interchromosomal ligations increased substantially in the nontethered MboI library (**Fig. 2c**). By contrast, it showed only a modest increase in the tethered MboI library. This result demonstrates that tethered libraries are minimally affected by the concentration of DNA fragments, confirming that most ligations in these libraries are between DNA fragments that are crosslinked to each other.

An improved signal-to-noise ratio allows a more accurate analysis of contacts with relatively low frequencies such as interactions between chromosomes (**Supplementary Fig. 1d**). For instance, several interactions between the small arm of chromosome 2 and chromosomes 20, 21 and 22 are clearly enriched in the tethered HindIII library (**Fig. 2d**) but not the nontethered HindIII library (**Fig. 2e**).

Intrachromosomal contacts define two classes of regions

We first analyzed the contact pattern within each chromosome. We defined the contact profile of a region as the ordered list of frequency values for its contacts with all the other regions in the genome (Online Methods). The Pearson's correlation between two intrachromosomal contact profiles is a similarity measure for the corresponding regions' contact behaviors. Using this measure and confirming a previous study[10], we observed that each chromosome can be divided into two classes of regions with anti-correlated intrachromosomal contact profiles (**Fig. 3a** and **Supplementary Fig. 2a**). At any given genomic distance, regions in the same class contact each other more frequently than regions in different classes (**Supplementary Fig. 2b**). One of these classes, here referred to as the 'active class', is substantially enriched for the presence and expression of genes, DNase hypersensitivity and activating histone modifications[10] (**Supplementary Fig. 2c**). The other class, here referred to as 'inactive', displays the opposite behavior (**Supplementary Fig. 2c**).

We asked how the similarity between contact profiles changes with increasing genomic distance between the regions on a chromosome. Notably, the contact profiles of the active regions remain similar even when relatively long genomic distances separate them (**Fig. 3b**). For the inactive regions, in contrast, the contact profile similarity decreases more quickly and dissipates at longer distances (**Fig. 3b**). Therefore, inactive regions are more likely to associate with

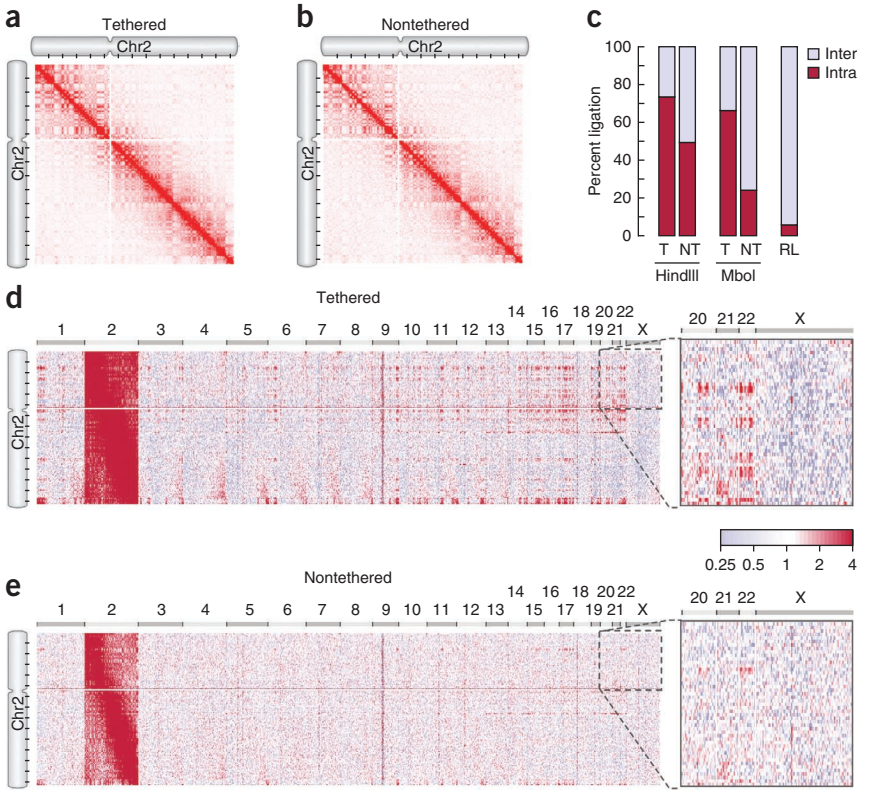

Figure 2 Tethering improves the signal-to-noise ratio of conformation capture. (**a**,**b**) TCC can reproduce the results obtained by Hi-C[10]. A genome-wide contact frequency map is compiled from the ligation frequency data generated by tethered (TCC) (**a**) and nontethered (Hi-C) (**b**) conformation capture. The portion of each map that corresponds to the intrachromosomal contacts of chromosome 2 is shown. The intensity of the red color in each position of the map represents the observed frequency of contact between corresponding segments of the chromosome, which are shown on the top and to the left of the map. In these maps, chromosome 2 is divided into segments that span 277 HindIII sites each, resulting in 258 segments of ~1 Mb. A pair of tick marks on the ideogram encompasses 4,986 HindIII sites. In this and other figures, the white lines in the heat maps mark the unalignable region of the centromeres. (**c**) The observed fractions of intra- and interchromosomal ligations in tethered (T) and nontethered (NT) libraries produced using HindIII or MboI. The random ligation (RL) bar represents the expected fractions if all ligations occurred between noncrosslinked DNA fragments. For the nontethered MboI library only, these fractions were determined by sequencing 160 individual DNA molecules from three replicates of the experiment. (**d**,**e**) The genome-wide enrichment map for chromosome 2, compiled from the tethered (**d**) and nontethered (**e**) HindIII libraries. Enrichment is calculated as the ratio of the observed frequency in each position to its expected value; expected values were obtained assuming completely random ligations (Online Methods). Red and light blue, respectively, indicate enrichment and depletion of a contact. Chromosome 2 (left) extends along the *y* axis whereas all 23 chromosomes (top) extend along the *x* axis. The zoomed panel to the right of each map magnifies the section that corresponds to contacts between the small arm of chromosome 2 and chromosomes 20, 21, 22 and X. For maps in **d** and **e**, each chromosome is divided into segments that span 558 HindIII sites, leading to respectively 116 and 1,384 segments of ~1.5 Mb for chromosome 2 and all other chromosomes. A pair of tick marks on chromosome 2 spans 5,022 HindIII sites.

their neighboring regions, whereas active regions can associate with a more diverse panel of long-range contact partners.

A special case of this behavior was observed in the interactions between inactive regions of large chromosomes (that is, chromosomes 1–6, 8 and 10). The similarity of contact profiles decreases abruptly for inactive regions separated by the centromere. Consequently, only inactive regions in the same chromosome arm have similar contact profiles (**Supplementary Fig. 3a**). The frequency of contacts between inactive regions in different chromosome arms is also substantially lower than would be expected from their sequence separation alone (**Supplementary Fig. 3b**). These characteristics give rise to a distinctive four-block pattern in the 'inactive-only' correlation matrices of the larger chromosomes (**Fig. 3c** and **Supplementary Fig. 3c**). In contrast, the contact profile similarity of active regions is largely unaffected by the centromere (**Fig. 3c** and **Supplementary Fig. 3a**,**c**). These results suggest that, in larger chromosomes, inactive regions from opposing chromosome arms are largely inaccessible to each other whereas active regions can still interact.

Interchromosomal contacts are mostly mediated by active class

We next analyzed the contacts between different chromosomes. We began by defining the interchromosomal contact probability index (*ICP*) as the sum of a region's interchromosomal contact frequencies divided by the sum of its inter- and intrachromosomal contact frequencies. *ICP*, therefore, describes the propensity of a region to forming interchromosomal contacts.

We observed large differences in the distribution of *ICP* between the active and inactive classes. In the inactive class, the vast majority of regions have relatively low *ICP*s with the exception of a few cases (**Fig. 4a**, **Supplementary Fig. 4a**,**b** and **Supplementary Table 2**). Most of these exceptions flank the unalignable regions of the centromeres, and their high *ICP* is due to interaction with the centromeric regions of other chromosomes (**Supplementary Fig. 5a**). Additionally, the centromeric regions of the acrocentric chromosomes are more likely to contact each other than the centromeric regions of the metacentric chromosomes (**Supplementary Fig. 5b**). Furthermore, we found the highest centromere contact frequencies between chromosomes 13 and 21 and between chromosomes 14 and 22 (**Supplementary Fig. 5c**). All of these observations are in excellent agreement with previous imaging studies in lymphocytic cells[22–24].

In the active class, on the other hand, many regions have high *ICP*s. In fact, the vast majority of regions with a large *ICP* belong to the active class (**Fig. 4a**, **Supplementary Fig. 4a**,**b** and **Supplementary Table 2**). For example, in chromosome 2, 90% of the regions with a top 25% *ICP* are members of the active class (**Fig. 4a**). Nevertheless, not all the active regions have a large *ICP*. For instance, ~40% of the active regions in chromosome 2 form relatively few interchromosomal contacts, and their *ICP*s are similar to those of the inactive regions (**Fig. 4a**). This nonuniform contact behavior may reflect functional variations within this class. Indeed, we observed that those active regions with larger *ICP*s also show higher RNA polymerase II binding (**Fig. 4b**) as well as higher total

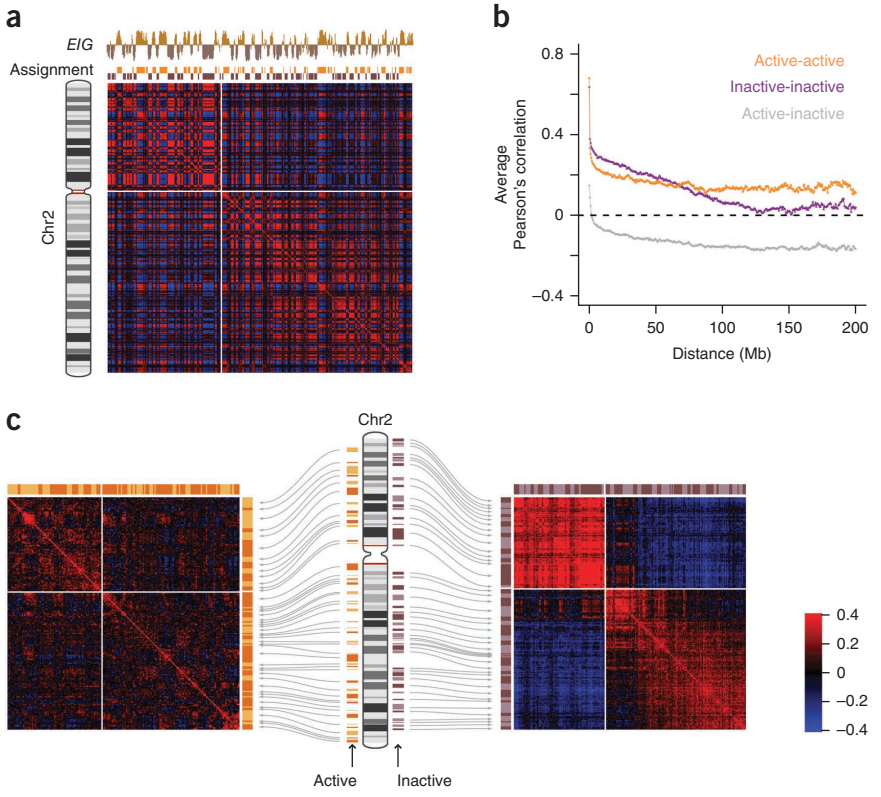

Figure 3 Intrachromosomal interactions. (**a**) Correlation map and class assignment for chromosome 2. The color of each position in the map represents the Pearson's correlation between the intrachromosomal contact profiles of the corresponding two segments of the chromosome to the left and on top (the ideogram of the chromosome has only been shown to the left, but the *x* axis of the map also represents the chromosome). To assign each segment to the active (orange blocks on top of the map) or the inactive (purple blocks on top of the map) class, principal component analysis was used to calculate *EIG* (value of the first principal component, plotted on top of the assignment blocks) for each segment (Online Methods). Segments with a positive *EIG* are assigned to the active class, whereas those with a negative *EIG* are assigned to the inactive class. Segments with *EIG* values close to zero have not been assigned to either class. The size of each chromosome band is based on the number of HindIII sites it contains. (**b**) The genome-wide average Pearson's correlation between intrachromosomal contact profiles of two active segments (orange), two inactive segments (dark purple), and an active and an inactive segment (gray) plotted against their genomic distance. (**c**) Active-active (left) and inactive-inactive (right) correlation maps for chromosome 2. Each correlation map is calculated following the procedure in **a**, except contacts between only the active segments (left) or only the inactive segments (right) are considered. The ideogram of chromosome 2 in the middle shows the active (orange bars on the left) and inactive (purple bars on the right) segments. The arrows mark the positions of these segments in the corresponding maps. In this figure, the tethered HindIII library is used and each chromosome is divided into segments of 138 HindIII sites, resulting in, respectively, 517 and 6,000 segments of ~0.5 Mb for chromosome 2 and the entire genome.

gene expression (Pearson's $r = 0.54$, $P < 10^{-15}$), indicating that higher transcriptional activity is associated with an increased probability of forming interchromosomal contacts.

We asked whether the regions' differences in *ICP* are reflected in their localization within their chromosomes' territories. Previous fluorescence imaging studies have shown that highly transcribed regions can frequently extend outside of the bulk territory of their chromosome[25,26]. One of these studies analyzed several loci on chromosome 11 in lymphoblastoid cells[27]. Remarkably, we found that the reported average distances of these loci from the edge of their chromosome territory is strongly correlated with their *ICP*s (Pearson's $r = 0.98$, $P < 10^{-3}$) (**Fig. 4c** and **Supplementary Fig. 4c**). Moreover, the loci that showed preferential localization in the bulk of the chromosome territory in the imaging study are inactive in the TCC data, whereas those that showed more frequent localization beyond the bulk of the territory are active and have large *ICP*s (**Fig. 4c**). Although more fluorescence imaging experiments are required to extend this observation to the entire genome, these examples suggest that *ICP* can also reflect the preferred positions of a locus within the territory of its chromosome.

Indiscriminate interactions between chromosome territories

To further examine the interactions between chromosomes, we analyzed those interchromosomal contacts with frequencies clearly above noise level. We refer to these contacts as 'significant interactions' (**Fig. 4d**). Most of these significant interactions are formed by active regions, in particular by those with high *ICP*s (**Fig. 4d**). Notably, most of these regions interact with numerous other high-*ICP* active regions throughout the genome (**Fig. 4d** and **Supplementary Fig. 6a**). For instance, each of the high-*ICP* active regions on chromosome 19 forms significant interactions with >40% of all the high-*ICP* active regions on chromosome 11 (**Fig. 4d**) and many more on other chromosomes (**Supplementary Fig. 6a**). Moreover, none of these interactions appears to be dominant, and they all have relatively low frequencies (**Fig. 4d** and **Supplementary Fig. 1d**). In the case of chromosomes 11 and 19, the significant interchromosomal interactions between high-*ICP* active regions are on average more than 70 times less frequent than intrachromosomal contacts between neighboring ~1-Mb regions. The numerosity of these interactions and their low frequencies suggest that each can be present in only a fraction of the cells.

Strikingly, the larger the *ICP* of the interchromosomal contact partners, the higher the observed frequency of their interaction (**Supplementary Fig. 6a**). Indeed, the contact frequency between a pair of high-*ICP* active regions shows a positive correlation with the product of their *ICP*s (**Fig. 4e** and **Supplementary Fig. 6b**,**c**). Based on these observations, it appears that for many high-*ICP* active regions, the probability of forming interchromosomal interactions is independent of the identity of their interaction partners. We already established that *ICP* can be an indicator for the relative position of a region from the edge of the chromosome territory. This correlation, therefore, suggests that the propensity for forming interchromosomal contacts between high-*ICP* active regions is largely governed by the spatial accessibility of the contact partners.

To confirm the existence of interchromosomal interactions between high-*ICP* active regions, we measured the colocalization frequency of one probe on chromosome 19 with each of four different probes on

chromosome 11 using 3D DNA FISH (**Fig. 4f–h** and **Supplementary Table 3**). The chromosome 19 probe was located in a high-*ICP* active region, whereas the four chromosome 11 probes were equally split between inactive and high-*ICP* active regions. These measurements showed that, in a small but significant fraction of the cells ($P < 0.05$), the high-*ICP* active region on chromosome 19 colocalizes with each of its active counterparts on chromosome 11 (**Fig. 4h**). In contrast, the same region on chromosome 19 is unlikely to localize in proximity to either inactive region on chromosome 11. These results support the conclusion that high-*ICP* active regions on different chromosomes can interact and that each interaction occurs in only a small fraction of the cells.

In summary, our observations indicate that most active regions do not interact exclusively with only a few specific regions on other chromosomes, rather they can form interactions indiscriminantly with many high-*ICP* active regions at different times. These contacts may be present only in the fraction of cells where both interaction partners are mutually accessible.

3D genome structures from conformation capture data

We then asked whether the indiscriminate and numerous low-frequency chromosome interactions can be reconciled with the non-random positioning of chromosome territories with the preferred radial positions seen in other studies[3–5]. Chromatin contacts are

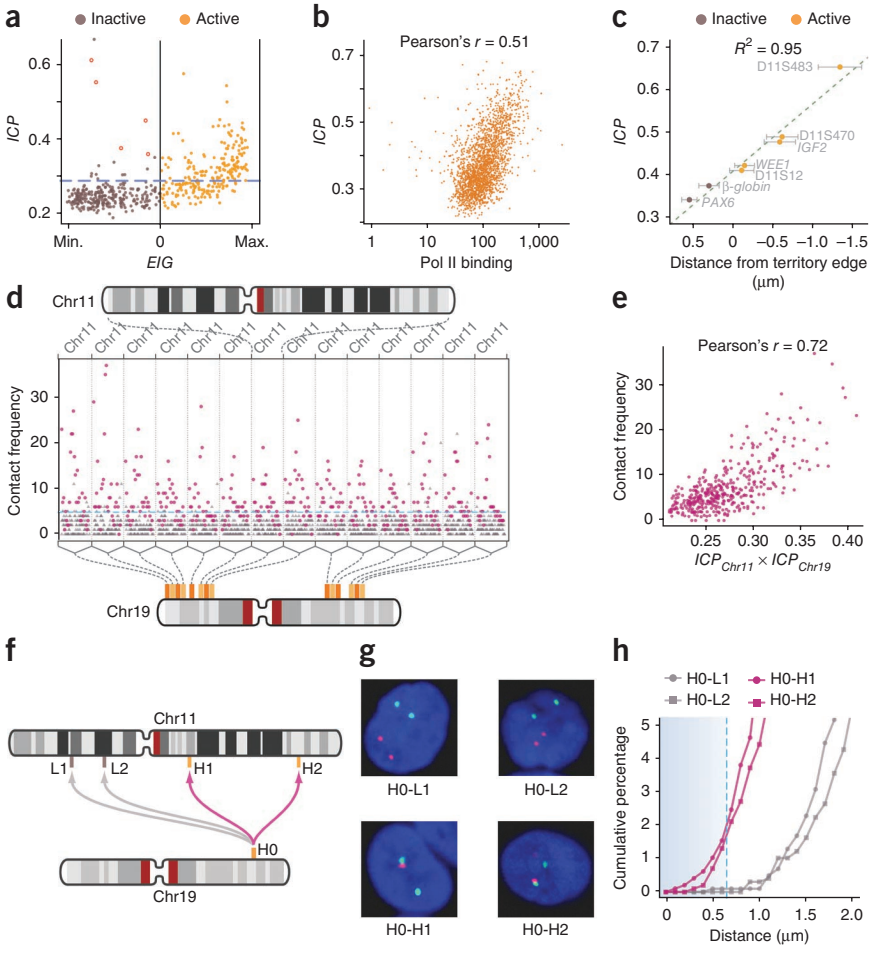

Figure 4 Interchromosomal interactions. (**a**) For all segments of chromosome 2, *ICP* is plotted against *EIG*. The blue dashed line separates high-*ICP* segments: values above the line are significantly larger than the average *ICP* for inactive segments ($P < 0.05$, nonparametric). The open red dots mark those inactive segments with a large *ICP* that also flank the centromere. Chromosome 2 is divided into 517 segments of ~0.5 Mb, each spanning 138 HindIII sites. Data from the tethered HindIII library are used in all panels. (**b**) For all active segments in the genome, *ICP* is plotted against the binding of RNA polymerase II (pol II). Pol II binding values are reproduced from a ChIP-Seq study[39] on the GM12878 cells and are in arbitrary units based on alignment frequency. $P < 10^{-16}$. Each point represents a segment of the genome that spans 138 HindIII sites. The *x* axis is plotted in a logarithmic scale. (**c**) For seven loci on the small arm of chromosome 11, the *ICP* value is plotted against their average distance from the edge of chromosome 11 territory as measured by FISH[27]. Positive distance values denote localization within the bulk territory, whereas negative values denote localization away from the bulk territory. Error bars, ± 95% confidence interval[27]. (**d**) Plotted are the frequencies of all contacts between high-*ICP* active segments on chromosome 19 and all the segments on chromosome 11. Purple dots represent contacts involving high-*ICP* active segments on chromosome 11, and gray triangles represent contacts involving all the other segments of chromosome 11. Contacts plotted between vertical dotted lines involve the same high-*ICP* active segment on chromosome 19 and all the segments of chromosome 11. The locations of the high-*ICP* active segments in chromosome 19 are marked by an orange bar on the ideogram of the chromosome on the bottom of the panel. The different shades of orange are used only to differentiate the adjacent segments. Frequencies above the dashed blue line are significantly higher than the average frequency of contacts between high-*ICP* active segments on chromosome 19 and inactive segments on chromosome 11 ($P < 0.04$, nonparametric). These frequencies can be considered significantly larger than the noise level, defined as the false-positive contact frequencies due to random intermolecular ligations. Each chromosome was divided into ~1 Mb segments that span 277 HindIII sites, resulting in a total of 143 segments for chromosomes 11 and 43 segments for chromosome 19. Among those, 14 segments on chromosome 19 and 28 segments on chromosome 11 were classified as high-*ICP* active. (**e**) For all possible pairs of high-*ICP* active segments from chromosomes 11 and 19, their contact frequency has been plotted against the product of their *ICP*s. Same interactions are marked with purple color in **d**. The *P*-value of the correlation is nominal. Other parameters are the same as in **d**. (**f**) The layout of 3D-FISH experiments where the localization of a high-*ICP* active locus on chromosome 19 (H0) relative to four loci on chromosome 11 (H1, H2, L1 and L2) was analyzed in about 1,000 cells per pair of loci. H1 and H2 are high-*ICP* active, whereas the L1 and L2 are inactive. The blocks on the chromosomes' ideograms mark the position of each locus (orange for high-*ICP* active and brown for inactive), and the arrows mark the pair combinations that are analyzed (purple for active-active and gray for active-inactive). (**g**) An example nucleus from each pair of loci analyzed in 3D-FISH. Nuclei are counterstained with DAPI (blue). In all four nuclei, the hybridization signal of H0 is shown in red and that of the other locus is shown in green. (**h**) Cumulative percentage of nuclei that show a pair of hybridization signals closer than a given distance is plotted. Only the closest pair of signals for each nucleus is considered. Distances smaller than 0.65 μm (dashed blue line, arbitrarily selected for visualization purposes) represent colocalizations in a close vicinity where a direct interaction between loci is possible.

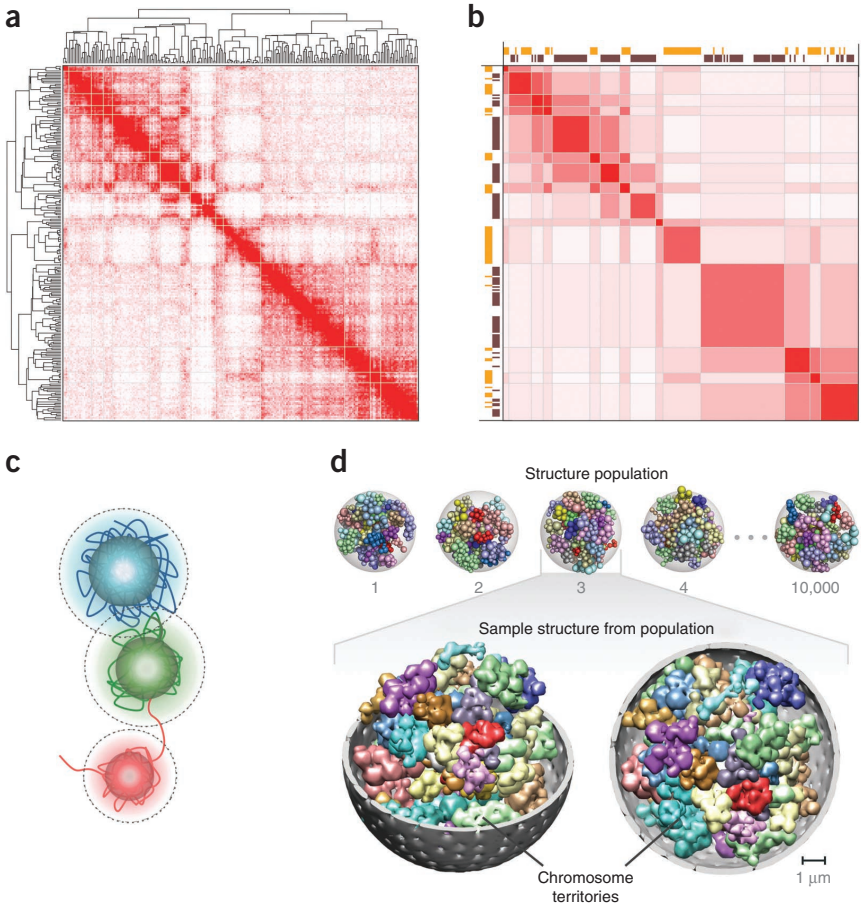

Figure 5 Coarse-graining of the contact frequency maps and structural representation of the genome. (**a**) The contact frequency map of chromosome 11 from the tethered HindIII library. The chromosome has been divided into 237 segments each of which covers 166 HindIII sites. Hierarchical constrained clustering was applied using the Pearson's correlation between the segments' contact profiles as the similarity measure (Methods). The dendrogram of constrained clustering is shown to the left and on top of the map. The intensity of the red color in the map represents the observed frequency of contact between corresponding chromosome segments. (**b**) Coarse-grained block matrix of chromosome 11. To identify the blocks, we determined a clustering cutoff following a previously described procedure[40]. In the block map, the value of an element is the average contact frequency of all the corresponding elements in the contact frequency map. The dimension of the initial contact frequency map is reduced to 15 elements for the block matrix of chromosome 11 and a total of 428 for the block matrices of all 23 chromosomes. Spearman's rank correlation coefficient between this block matrix and the contact frequency map in **a** is 0.78. Assignment of segments to the active (orange blocks) and inactive (dark brown blocks) classes are shown to the left and on top of the matrix. (**c**) Sphere representation for chromatin regions in a block. The sphere for each block is defined by two different radii. First, its hard radius (solid sphere), which is estimated from the block sequence length and nuclear occupancy of the genome; the sphere cannot be penetrated within this radius (Online Methods). Second, its soft radius (dotted line), which is twice that of the hard sphere radius. A contact between two spheres is defined as an overlap between the spheres' respective soft radii. Also shown is a schematic view of the 30 nm chromatin fiber. (**d**) Population of 10,000 genome structures. A schematic view of the calculated structure population is shown on top. A randomly selected sample from the population is magnified at the bottom. All 46 chromosome territories are shown. Homologous pairs share the same color. The nuclear envelope is displayed in gray. For visualization purposes, the spheres are blurred in the magnified structure because the use of 2 × 428 spheres to represent the genome makes the territories appear more discrete than they actually are.

observed with a wide range of frequencies, suggesting that many potential contacts are present in only a fraction of cells. In other words, the contacts in TCC data describe not necessarily one structure but represent the average contacts of numerous genome structures in different cells. Therefore, a population of genome structures must be generated in which the resulting variety of structures is statistically consistent with the data. We express this task as an optimization problem with three main components[28,29]: (i) a structural representation of chromosomes at an appropriate level of resolution; (ii) a scoring function quantifying the structure population's accordance with the data; and (iii) a method for optimizing the scoring function to yield a population of genome structures.

We first defined a coarse-grained structural representation of the chromosomes. The plaid appearance of the contact frequency maps suggests that each chromosome can be partitioned into 'blocks' of consecutive regions that share similar contact profiles. To identify these blocks, we applied constrained clustering using the Pearson's correlation between the regions' contact profiles as a similarity measure (**Fig. 5a** and Online Methods). Optimizing the clustering cutoff divided the haploid genome into 428 'chromatin-block' regions (**Supplementary Fig. 7a** and Online Methods). The resulting block-based contact frequency map (**Fig. 5b**) is highly correlated with the original frequency map (Spearman's correlation 0.81, $P < 10^{-16}$), confirming that the characteristic long-range contact patterns are preserved (**Fig. 5a**,**b**). Several observations indicate that large portions of chromatin regions in any given block are in spatial proximity and predominately occupy the same specific subterritory in the nucleus. First, the vast majority of contacts are between regions inside a block. Second, across the block borders, the contact probability between neighboring regions is abruptly reduced, and an abrupt change in contact profiles is observed. As a first approximation, we defined the subterritory that is largely occupied by each block region as a globular volume whose spherical radius is approximated by the block size (**Fig. 5c** and **Supplementary Table 4**). The structure of a genome is then given by a spatial arrangement of these spheres. Our goal is then to determine a population of genome structures, where in each structure all the 856 spheres of the diploid genome are packed into the nucleus in such a way that their contacts across the population are entirely consistent with the TCC data (**Fig. 5d**).

We converted the TCC contact frequencies into a set of contact restraints between spheres in all the structures of the population. A restraint can be thought of as generating a 'force' between the spheres so that they form a contact. Notably, any given contact can be enforced only in the fraction of models in the population corresponding to its

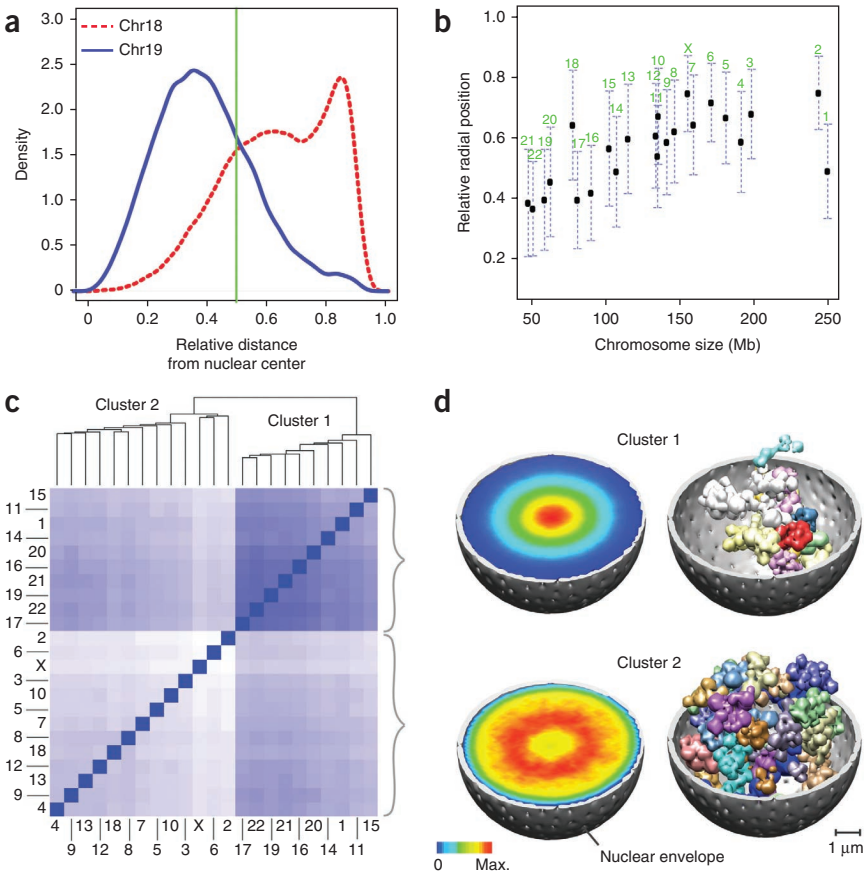

Figure 6 Population-based analysis of chromosome territory localizations in the nucleus. (**a**) The distribution of the radial positions for chromosomes 18 (red dashed line) and 19 (blue solid line), calculated from the genome structure population. Radial positions are calculated for the center of mass of each chromosome and are given as a fraction of the nuclear radius. (**b**) The average radial position of all chromosomes plotted against their size. Error bars, s.d. (**c**) Clustering of chromosomes with respect to the average distance between the center of mass of each chromosome pair in the genome structure population. The clustering dendrogram, which identifies two dominant clusters is shown on top. The matrix of average distances between pairs of chromosomes is shown at the bottom. The intensity of blue color increases with decreasing distance. (**d**) (Left panels) The density contour plot of the combined localization probability for all the chromosomes in cluster 1 (top panel) and cluster 2 (bottom panel) calculated from all the structures in the genome structure population. The rainbow color-coding on the central nuclear plane ranges from blue (minimum value) to red (maximum value). (Right panels) A representative genome structure from the genome structure population. Chromosome territories are shown for all the chromosomes in cluster 1 (top) and all the chromosomes in clusters 2 (bottom). The localization probabilities are calculated following a previously described procedure[28].

TCC frequency (**Supplementary Methods**). If a contact is not enforced, no assumptions are made about the relative positions of the corresponding spheres. Therefore, in contrast to other approaches[12,30], our method does not correlate contact frequencies with average distances; it relies purely on the TCC data by incorporating only the presence or absence of chromatin contacts.

In a diploid cell, most loci are present in two copies. Because the TCC data do not distinguish between these copies, the optimal assignment of each sphere to a specific contact is determined as a part of our optimization process[31] using the integrative modeling platform[28,29].

Finally, starting from random positions, we simultaneously optimized the positions of all the spheres in a population of 10,000 genome structures to a score of zero, indicating that no restraint violations remained (**Supplementary Methods**).

To test how consistent this structure population is with the experiment, we calculated the block contact frequency map from the population of structures and compared it with the original data. The two were strongly correlated with an average Pearson's correlation of 0.94, confirming the excellent agreement between contact frequencies in the structure population and experiment (**Supplementary Fig. 7b–d**). Furthermore, three independently calculated populations showed that our structure population was highly reproducible (Pearson's $r > 0.999$), which also indicates that, at this resolution, the size of the model population was sufficiently large (**Supplementary Methods**).

Structural features of the genome population

Because chromatin contacts in the TCC data are observed over a wide range of frequencies, the resulting population shows a fairly large degree of structural variation (**Supplementary Fig. 8a**,**b**). For instance, on average only 21% of contacts are shared between any two structures in the population (**Supplementary Fig. 8c**). Despite this large heterogeneity, the structure population reveals a distinct and nonrandom chromosome organization. Specifically, the population clearly identifies the preferred radial positions of chromosomes (**Fig. 6a**,**b** and **Supplementary Fig. 9b**). These positions strongly agree with independent FISH studies in lymphoblasts[4,5]. The Pearson's correlation between the FISH- and population-based average positions was 0.71 ($P < 10^{-3}$) for the 22 chromosomes whose radial positions were previously determined[4]. Instead, radial positions in a control population generated without TCC data did not agree with the FISH data (Pearson's $r = -0.2$, **Supplementary Fig. 9a**), indicating that TCC data are sufficient for generating the correct radial distributions seen in the imaging experiments[4]. In general, the radial chromosome positions tend to increase with their size, with some noticeable exceptions (**Fig. 6b**). One of these cases is the radial positions of chromosomes 18 and 19 which, despite their similar size, we observed at different positions[5]. Chromosome 19 is located closer to the center of the nucleus, whereas chromosome 18 is preferentially located closer to the nuclear envelope (**Fig. 6a**). Furthermore, the homologous copies of chromosome 18 are often distant from each other whereas those of chromosome 19 are often closely associated (**Fig. 6a** and **Supplementary Fig. 9b**), in agreement with independent experimental evidence[5].

Structure-based analysis of territory colocalizations

When chromosome territories are clustered based on their average distances, two main groups can be identified (**Fig. 6c**). The first group (chromosomes 1, 11, 14–17 and 19–22) tends to occupy the central region of the nucleus as is evident from their population-based joint localization probabilities (**Fig. 6d**). These chromosomes also tend to have relatively high gene densities[32]. The second group

(chromosomes 2–10, 12, 13, 18 and X) preferentially occupies the periphery of the nucleus (**Fig. 6d**).

Finally, we observed differences in the local packing between the spheres composed of mainly active or inactive regions. The average distances between spheres of mainly active regions are statistically larger (**Supplementary Fig. 9c**), suggesting that inactive regions are more densely packed in the structure population in comparison to the active regions.

DISCUSSION

TCC offers improved sensitivity in identifying chromatin interactions. In particular, libraries generated with the tethering strategy have a lower level of random intermolecular ligation compared to those generated by a nontethered approach (**Fig. 2c**). The reduced noise level facilitates the analysis of low-frequency contacts such as interchromosomal interactions, which can otherwise be lost in the higher background noise (**Fig. 2d**,**e** and **Supplementary Fig. 10**). Because the intermolecular ligation noise remains low even at substantially increased DNA concentrations, this method also facilitates higher resolution analyses with enzymes that cut the chromatin more frequently.

Two main factors may contribute to this reduction of random intermolecular ligations in the tethered libraries. First, DNA fragments can be immobilized only when they are crosslinked to proteins and are otherwise washed out of the reaction (**Fig. 1**). Therefore, 'naked' DNA fragments, which would produce only false-positive contacts, are unlikely to participate in ligation. Second, immobilized protein-DNA complexes cannot diffuse freely, markedly reducing encounters between noncrosslinked molecules during ligation. When combined with a sufficiently low surface density of complexes that reduces their chance of immobilizing in close vicinity, these conditions can effectively reduce intermolecular ligations.

The TCC data provide new insights into the internal organization of the chromosome territories. The regions of the inactive class preferentially associate with neighboring inactive regions, whereas the regions of the active class have a diverse panel of long-range contact partners (**Fig. 3b** and **Supplementary Fig. 2b**). A pronounced instance of this behavior can be observed across the centromeres. In large chromosomes, inactive regions on opposing sides of the centromere show few interactions with each other (**Fig. 3c** and **Supplementary Fig. 3**). At the same time, active regions on different arms show extensive interactions (**Fig. 3c** and **Supplementary Fig. 3**). This behavior is consistent with previous reports in *Drosophila melanogaster* where interactions between some inactive polycomb-associated regions were constrained within a chromosome arm[33,34]. These observations are also consistent with the more dense packing of the inactive regions seen in our genome structure population (**Supplementary Fig. 9c**).

More clues into the spatial organization of loci is provided by their propensity to form interchromosomal contacts. With the interchromosomal *ICP*, we have introduced a quantitative measure of interchromosomal contact propensity for each region (**Fig. 4a**, **Supplementary Fig. 4a**,**b** and **Supplementary Table 2**). *ICP* appears to be an indicator of the relative position of a region within the chromosome territory (**Fig. 4c**). Based on the available localization data[26], we found that active regions with higher *ICP*s show more frequent localization beyond the bulk or at the border of the territory (**Fig. 4c**). Another important property of *ICP* is that it correlates with the functional characteristics of loci. For instance, active regions with larger *ICP* values show more binding by RNA polymerase II (**Fig. 4b**) and higher levels of gene expression.

Our results reveal new insights into interactions between chromosomes. Most of these interactions are mediated by active regions with relatively high *ICP*s. Each of these regions forms significant interactions with numerous high-*ICP* active regions on other chromosomes (**Fig. 4d**). Notably, the frequencies of these interactions increase with the *ICP* of the interaction partners (**Fig. 4e** and **Supplementary Fig. 6**). As these regions tend to localize at the territory borders more frequently with increasing *ICP*s (**Fig. 4c**), their interaction frequency may be largely governed by their accessibility rather than other factors. In other words, interchromosomal interactions can form indiscriminately between high-*ICP* active regions that are accessible to each other. Accessibility may be determined by factors such as radial position or regional transcriptional activity in each cell.

We also observed that the propensity to form interchromosomal contacts is correlated with a region's transcriptional activity (**Fig. 4b**). Because transcription is often focused at discrete sites (that is, transcription factories)[35], this correlation may be a consequence of the active regions being recruited to the same factory, thereby supporting previous suggestions that transcription factories play an important role in stabilizing interchromosomal interactions[2,36,37]. The indiscriminate nature of these interactions suggests that, based on accessibility in each cell, different combinations of loci associate in one factory. Nevertheless, the association of a specific transcription factor with only some of the transcription factories, as reported before[37], can make the recruitment of its targets to the same factories more likely. Moreover, as transcription is not the only nuclear function that is concentrated at discrete sites[1,38], it is possible that other factories, such as those of splicing and DNA repair, also mediate the indiscriminate interactions between chromosome territories.

As these interchromosomal interactions are both numerous and low frequency, each can be present only in a small fraction of the cells. In fact, in our FISH experiments, two pairs of high-*ICP* active regions were found to colocalize in only a few percent of the cells (**Fig. 4f**–**h**). These cell-to-cell differences are reflected in a fairly large variation between the genome structures in the population generated from the TCC data (**Fig. 6a**,**b** and **Supplementary Fig. 8**). Despite this variation, however, the structure population reproduces the previously described[4,5], preferred radial positions of chromosomes (**Fig. 6a**,**b** and **Supplementary Fig. 9a**,**b**). The structural analysis indicates that the genome-wide behavior of interchromosomal interactions, as observed in the TCC data, is in keeping with the previously described architectural features. Furthermore, this population demonstrates that the TCC data are sufficient to reproduce the distinct spatial distributions of chromosome territories (**Fig. 6a**,**b** and **Supplementary Fig. 9a**,**b**).

Our population-based modeling, therefore, provides a means of studying 3D genome architectures. By systematically translating the TCC data into a population of genome structures, this approach also allows a statistical interpretation of the genome organization (**Fig. 6** and **Supplementary Figs. 8** and **9b**,**c**). Although not every structure in the population may be a definitive structure of chromosomes, several lines of evidence indicate that, as a whole, this population is representative of the true configurations of the genome. The structure population is highly reproducible with independently generated populations reproducing the same statistical features with a high precision. More importantly, the population statistics agree with independent experimental data (such as FISH data) that were not used to generate the structures. Moreover, a structure population based only on part of the TCC data was able to correctly predict the missing data (**Supplementary Methods**).

In this work, we have focused on the locations of chromosome territories. But the resulting genome structure population provides a starting point for a higher resolution description of the spatial properties of the genome.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession code. All sequencing results and binary contact catalogs are publicly available in NCBI SRA under accession number SRA025848.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors would like to acknowledge P. Laird, J. Knowles and J. Aman and the USC Epigenome Center for assistance in high-throughput sequencing, M. Michael and A. Williams for assistance in confocal microscopy, N. Bottini and Q.-L. Ying and members of their laboratories for assistance in cell culture, A.D. Smith for access to cluster computing, I. Slaymaker for graphic design and the integrative modeling platform team for support. Structure calculations were done on USC HPCC. We also thank N. Arnheim, A.D. Smith, O. Aparicio, S. Forsburg, W. Li, M.S. Madhusudhan, K. Gong, S. Srivastava, S. Al-Bassam, M. Murphy, J. Peace and Z. Ostrow for useful discussions and comments on the manuscript. This work is supported by Human Frontier Science Program grant RGY0079/2009-C to F.A., Alfred P. Sloan Foundation grant to F.A.; US National Institutes of Health (NIH) grants GM064642 and GM077320 to L.C., NIH grant GM096089 to F.A. and NIH grant RR022220 to F.A. and L.C. F.A. is a Pew Scholar in Biomedical Sciences, supported by the Pew Charitable Trusts.

AUTHOR CONTRIBUTIONS

R.K. and L.C. conceived the TCC technique and R.K. performed the experiments and analyzed the contact data. R.K. and N.J. performed the FISH experiments and analyzed the results. H.T. and F.A. conceived the modeling strategy, and R.K. and L.C. provided input and discussions. H.T. performed the modeling experiments and analysis. R.K., F.A., H.T. and L.C. wrote the manuscript. All authors commented on and revised the manuscript. F.A. and L.C. supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/nbt/index.html.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787–800 (2007).
2. Branco, M.R. & Pombo, A. Chromosome organization: new facts, new models. *Trends Cell Biol.* **17**, 127–134 (2007).
3. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* **2**, 292–301 (2001).
4. Boyle, S. *et al.* The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* **10**, 211–219 (2001).
5. Cremer, M. *et al.* Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome Res.* **9**, 541–567 (2001).
6. Branco, M.R. & Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.* **4**, e138 (2006).
7. Sproul, D., Gilbert, N. & Bickmore, W.A. The role of chromatin structure in regulating the expression of clustered genes. *Nat. Rev. Genet.* **6**, 775–781 (2005).
8. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol. Cell* **10**, 1453–1465 (2002).
9. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).
10. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
11. Spilianakis, C.G. & Flavell, R.A. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nat. Immunol.* **5**, 1017–1027 (2004).
12. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
13. Wurtele, H. & Chartrand, P. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Res.* **14**, 477–495 (2006).
14. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra and interchromosomal interactions. *Nat. Genet.* **38**, 1341–1347 (2006).
15. van Steensel, B. & Dekker, J. Genomics tools for unraveling chromosome architecture. *Nat. Biotechnol.* **28**, 1089–1095 (2010).
16. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).
17. Cook, P.R. Predicting three-dimensional genome structure from transcriptional activity. *Nat. Genet.* **32**, 347–352 (2002).
18. Lanctot, C., Cheutin, T., Cremer, M., Cavalli, G. & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat. Rev. Genet.* **8**, 104–115 (2007).
19. Misteli, T. Self-organization in the genome. *Proc. Natl. Acad. Sci. USA* **106**, 6885–6886 (2009).
20. Misteli, T. Protein dynamics: implications for nuclear architecture and gene expression. *Science* **291**, 843–847 (2001).
21. Simonis, M., Kooren, J. & de Laat, W. An evaluation of 3C-based methods to capture DNA interactions. *Nat. Methods* **4**, 895–901 (2007).
22. Alcobia, I., Quina, A.S., Neves, H., Clode, N. & Parreira, L. The spatial organization of centromeric heterochromatin during normal human lymphopoiesis: evidence for ontogenically determined spatial patterns. *Exp. Cell Res.* **290**, 358–369 (2003).
23. Sullivan, G.J. *et al.* Human acrocentric chromosomes with transcriptionally silent nucleolar organizer regions associate with nucleoli. *EMBO J.* **20**, 2867–2877 (2001).
24. Alcobia, I., Dilao, R. & Parreira, L. Spatial associations of centromeres in the nuclei of hematopoietic cells: evidence for cell-type-specific organizational patterns. *Blood* **95**, 1608–1615 (2000).
25. Volpi, E.V. *et al.* Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. *J. Cell Sci.* **113**, 1565–1576 (2000).
26. Mahy, N.L., Perry, P.E., Gilchrist, S., Baldock, R.A. & Bickmore, W.A. Spatial organization of active and inactive genes and noncoding DNA within chromosome territories. *J. Cell Biol.* **157**, 579–589 (2002).
27. Mahy, N.L., Perry, P.E. & Bickmore, W.A. Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *J. Cell Biol.* **159**, 753–763 (2002).
28. Alber, F. *et al.* Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694 (2007).
29. Alber, F. *et al.* The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701 (2007).
30. Bau, D. *et al.* The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.* **18**, 107–114 (2011).
31. Alber, F., Kim, M.F. & Sali, A. Structural characterization of assemblies from overall shape and subcomplex compositions. *Structure* **13**, 435–445 (2005).
32. Kreth, G., Finsterle, J., von Hase, J., Cremer, M. & Cremer, C. Radial arrangement of chromosome territories in human cell nuclei: a computer model approach based on gene density indicates a probabilistic global positioning code. *Biophys. J.* **86**, 2803–2812 (2004).
33. Tolhuis, B. *et al.* Interactions among polycomb domains are guided by chromosome architecture. *PLoS Genet.* **7**, e1001343 (2011).
34. Chotalia, M. & Pombo, A. Polycomb targets seek closest neighbours. *PLoS Genet.* **7**, e1002031 (2011).
35. Cook, P.R. The organization of replication and transcription. *Science* **284**, 1790–1795 (1999).
36. Cook, P.R. A model for all genomes: the role of transcription factories. *J. Mol. Biol.* **395**, 1–10 (2010).
37. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* **42**, 53–61 (2010).
38. Lamond, A.I. & Spector, D.L. Nuclear speckles: a model for nuclear organelles. *Nat. Rev. Mol. Cell Biol.* **4**, 605–612 (2003).
39. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).
40. Kelley, L.A., Gardner, S.P. & Sutcliffe, M.J. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.* **9**, 1063–1065 (1996).

ONLINE METHODS

TCC. Twenty-five million GM12878 cells were crosslinked with 1% formaldehyde. Cells were lysed and treated with iodoacetyl-PEG2-biotin to biotinylate cysteine residues. Biotinylated chromatin was digested with either HindIII or MboI and immobilized on 400 µl MyOne Streptavidin T1 beads (Invitrogen), which has about 100 cm^2 surface area. The DNA ends were filled in using dGTPαS and biotin-14-dCTP nucleotide analogs and ligated. Crosslinking was reversed and DNA was purified and treated with *Escherichia coli* exonuclease III to remove the biotinylated residues from nonligated DNA ends. Fragments that contain ligation junctions were then purified by pull-down with streptavidin-coated magnetic beads and prepared for massively parallel sequencing.

Hi-C. As an example of nontethered conformation capture, Hi-C was carried out as described previously[10] on 25 million GM12878 cells. Crosslinking conditions were identical to that of the TCC experiments. Digestion was carried out with either HindIII or MboI. The ligation step was carried out in a total volume of 40 ml.

Contact frequency maps. Unless otherwise stated, analyses described in this article have been carried out using the tethered HindIII library. Moreover, in all the analyses of this library, intrachromosomal contacts between regions closer than 30,000 bp have been removed from consideration (**Supplementary Methods**).

To generate the contact frequency maps, the genome was divided into contiguous 'segments' spanning an equal number of restriction sites. The contact matrix $\boldsymbol{F}$ was defined such that the matrix entry $f_{i,j}$ is based on the number of observed ligation products between segments i and j (**Supplementary Methods**)[9,10,41]. Depending on the resolution that was desired, the number of restriction sites in each segment may have varied. For example, in the contact frequency maps shown in **Figure 2a**,**b**, chromosome 2 was divided into segments spanning 277 HindIII sites, dividing it into 258 segments.

Contact profile. The contact profile of region i is the ith row-vector of the matrix ($\boldsymbol{F}$), which entails the ordered list of contact frequencies of segment i with all other segments in the genome.

Contact enrichment (expected value). The expected value for the frequency of a contact between segments i and j ($e_{i,j}$) was calculated as

$$e_{i,j} = \gamma \times s_i \times s_j$$

where s_i and s_j are the total of all observed contact frequencies involving segments i and j, respectively, and γ is a normalization constant. For example, in **Figure 2d**,**e**, (γ is chosen such that the average observed/expected frequency ($f_{i,j}/e_{i,j}$) of all interchromosomal contacts is equal to 1.

Correlation maps. For each chromosome all contact frequencies were first normalized by the average contact frequency of all pairs of segments with the same distance in the map. Then each element in the correlation map, $p_{i,j}$, was defined as Pearson's correlation between the intrachromosomal contact profiles of segments i and j.

Principal component analysis and assignment of the active and inactive classes. The first principal component of each intrachromosomal correlation map (defined as the eigenvector with the largest eigenvalue) was calculated. The projection of each segment's intrachromosomal correlation profile on this eigenvector was taken as the value of its first principal component (*EIG*). Of the two possible directions for the eigenvector, the one that would result in a positive correlation between *EIG* and RNA polymerase II (pol II) binding was chosen. Segments with a positive *EIG* were then assigned to the active and others to the inactive class. For the analyses that required a high-confidence assignment of the classes (that is, **Figs. 3c** and **4d** and **Supplementary Fig. 3**), only the segments with positive *EIG* values that were larger than a third of the maximum chromosome-wide *EIG* were assigned to the active class, and only those with negative *EIG* values that were smaller than a third of the minimum chromosome-wide *EIG* were assigned to the inactive class. The remaining segments were left unassigned. With these criteria, ~77% of all segments in autosomal chromosomes were assigned to one of the two classes.

RNA polymerase II binding. Raw pol II ChIP-seq data in GM12878 cells were obtained from another study[39]. The ChIP-seq data were aligned to the human genome (GRCh37/hg19). The binding of pol II to each segment was calculated as the number of reads that aligned to the segment in anti-pol II ChIP divided by number of aligned reads in anti-IgG negative control.

Gene expression. Raw RNA-seq (poly-A enriched) data for GM12878 cells were obtained from another study[39] and aligned to the human genome (GRCh37/hg19). The expression level of UCSC known canonical genes in hg19 was estimated using a two-parameter generalized Poisson model, as described[42]. Total gene expression for each segment was measured as the sum of the expressions (Theta values) of all genes that overlap with that segment.

Histone modifications. Raw histone modification ChIP-seq data in GM12878 cells were obtained from the ENCODE project[43] (generated at the Broad Institute and in the Bradley E. Bernstein laboratory at the Massachusetts General Hospital/Harvard Medical School). The ChIP-seq data were aligned to the human genome (GRCh37/hg19). Each histone modification level was calculated as the number of reads that aligned to the segment in the corresponding antibody pull-down experiment divided by the number of aligned reads in the input negative control.

DNase hypersensitivity. Raw DNaseI sensitivity sequencing data in GM12878 cells were obtained from the ENCODE project[43] (these data were generated using the Digital DNaseI methodology[44] by the UW ENCODE group). The Digital DNase sequencing reads were aligned to the human genome (GRCh37/hg19). The total number of alignments to each segment was taken as the total amount of DNase hypersensitivity in that segment.

3D-FISH. BACs were obtained from the BACPAC Resource Center (BPRC) at Children's Hospital Oakland Research Institute. 3D-FISH experiments were carried as described previously[45]. The only BAC that aligns to chromosome 19 (RP11-50I11) was labeled with digoxigenin, whereas the other BACs (RP11-651M4, RP11-220C23, RP11-169D4 and RP11-770J1), all of which align to chromosome 11, were labeled with biotin in nick-translation reactions. In each hybridization reaction, roughly 300 ng of each labeled probe and 5 µg of CotI DNA were used. Each label was detected with two layers; avidin-FITC and mouse anti-digoxigenin as the first layer, and goat anti-avidin-FITC and sheep anti-mouse-Cy3 as the second layer. Total DNA was counterstained by DAPI. Confocal microscopy was carried out using an Olympus FluoView FV1000 imaging system equipped with a 60X/1.42 PlanApo objective. Optical sections (z stacks) of 0.20 µm apart were obtained in the sequential mode in DAPI, FITC and Cy3 channels. Center-to-center distances between the probes were calculated using the Smart 3D-FISH plug-in for ImageJ as described[46]. Each pair of probes was processed in duplicates with 1,011 total nuclei for H0-L1, 987 for H0-L2, 976 for H0-H1 and 998 for H0-H2.

Modeling the 3D organization of the genome. To identify the clustering cutoff, we used a penalty function designed to simultaneously minimize the number of clusters and the variation within each cluster[40].

The genome of the diploid cell was represented by 856 spheres, whose relative radii depend on the genomic length of the chromatin regions in a block (see **Fig. 5b** and **Supplementary Methods** for the definition of the blocks). Each sphere is represented by two concentric spheres, a hard sphere and a soft sphere (**Fig. 5c**). The radius of the hard sphere of a block was defined as (**Supplementary Table 4**)

$$R_i = \left(\frac{O_{\text{nuc}} l_i}{2 \sum_{k=1}^{N} l_k} \right)^{1/3} R_{\text{nuc}},$$

with l_i as the genomic length of the block region i, R_{nuc} as the nuclear radius, and N as the number of blocks in the haploid genome. The chromatin occupancy volume O_{nuc} was set to 20%. The radius of the soft sphere is twice the radius of the hard sphere.

The scoring function captures all the information about the genome structure and is the sum of restraints of various types. These restraints ensure that all spheres are positioned within the nuclear volume. The overlap between hard spheres is prevented, allowing for a defined genome occupancy in the nucleus. A contact restraint enforces that the soft radii of two spheres are overlapping. Contacts are enforced based on the contact information from the HindIII-TCC library. Our procedure ensures that only a fraction of models in the population enforces a contact according to the observed contact frequency. The scoring function was implemented and optimized in the integrative modeling platform[28,47].

The optimization relies on conjugate gradients and molecular dynamics with simulated annealing. It starts with a random configuration of spheres and then iteratively moves these spheres so as to minimize violations of the restraints to a score of zero, resulting in a population of 10,000 genome structures that are consistent with the input data. Details of the computational methods are described in **Supplementary Methods**.

41. Tanizawa, H. *et al.* Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.* **38**, 8164–8177 (2010).
42. Srivastava, S. & Chen, L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.* **38**, e170 (2010).
43. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
44. Sabo, P.J. *et al.* Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods* **3**, 511–518 (2006).
45. Beatty, B., Mai, S. & Squire, J. *FISH: A Practical Approach.* (Oxford University Press, Oxford, 2002).
46. Gue, M., Messaoudi, C., Sun, J.S. & Boudier, T. Smart 3D-FISH: automation of distance analysis in nuclei of interphase cells by image processing. *Cytometry A* **67**, 18–26 (2005).
47. Alber, F., Forster, F., Korkin, D., Topf, M. & Sali, A. Integrating diverse data for structure determination of macromolecular assemblies. *Annu. Rev. Biochem.* **77**, 443–477 (2008).

doi:10.1038/nbt.2057