Bioinformatics, 34, 2018, i475–i483
doi: 10.1093/bioinformatics/bty265
ISMB 2018

Quantifying the similarity of topological domains across normal and cancer human cell types

Natalie Sauerwald and Carl Kingsford*

Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Three-dimensional chromosome structure has been increasingly shown to influence various levels of cellular and genomic functions. Through Hi-C data, which maps contact frequency on chromosomes, it has been found that structural elements termed topologically associating domains (TADs) are involved in many regulatory mechanisms. However, we have little understanding of the level of similarity or variability of chromosome structure across cell types and disease states. In this study, we present a method to quantify resemblance and identify structurally similar regions between any two sets of TADs.
Results: We present an analysis of 23 human Hi-C samples representing various tissue types in normal and cancer cell lines. We quantify global and chromosome-level structural similarity, and compare the relative similarity between cancer and non-cancer cells. We find that cancer cells show higher structural variability around commonly mutated pan-cancer genes than normal cells at these same locations.
Availability and implementation: Software for the methods and analysis can be found at https://github.com/Kingsford-Group/localtadsim
Contact: carlk@cs.cmu.edu

1 Introduction

Three-dimensional chromosome structure has been shown to be an influential factor in diverse aspects of cellular functioning. Since the introduction of chromosome conformation capture (Dekker *et al.*, 2002) and its many variants including a high-throughput experiment permitting genome-wide structural measurements termed Hi-C (Lieberman-Aiden *et al.*, 2009), there have been many studies associating chromosome structure with numerous cellular processes. Among these include several studies linking chromosome structure to gene expression and regulation (Cavalli and Misteli, 2013; Cremer and Cremer, 2001; Duggal *et al.*, 2014; Le Dily *et al.*, 2014; Sauerwald *et al.*, 2017), and more specifically changes in structure have been associated with various human diseases and disabilities, including several cancers (Fudenberg *et al.*, 2011; Hnisz *et al.*, 2016; Meaburn *et al.*, 2009; Misteli, 2010), as well as deformation or malformation of limbs during development (Lupiáñez *et al.*, 2016). On the mechanistic side, structural components have been implicated in replication timing (Ay *et al.*, 2014; Moindrot *et al.*, 2012; Pope *et al.*, 2014; Ryba *et al.*, 2010) and associated with DNA accessibility and nuclear organization (Ramani *et al.*, 2016).

Although studies of chromosome structure have provided meaningful biological insights such as those mentioned above, many questions remain about the precise role and variability of the chromosomal architecture. In particular, one key question is the extent to which chromosome structure is conserved between cell types, or how much it differs between normal and diseased tissue, e.g. cancer tissue. A deeper understanding of the level of structural similarity across cell types would reveal mechanistic insights into the role of three-dimensional folding of the chromosomes and demonstrate the relative cell-type specificity of the arrangement, yet very limited work has been devoted to this question. We address this question through quantifying structural similarity in pairwise comparisons, and apply this method to compare chromosome structure across many cell types, as well as between cancer and normal cells.

Chromosome structure is described in terms of several different scales of components, from multi-megabase compartments to sub-megabase topologically associating domains (TADs) and subTADs (Bonev and Cavalli, 2016). Compartments divide chromosomes into two broad categories: loosely packed, gene-rich areas termed A compartments and densely packed inactive areas termed B compartments. They can be identified in a straightforward way from the correlation matrix of the Hi-C map (Lieberman-Aiden *et al.*, 2009). TADs, visually identifiable as squares along the diagonal of the Hi-C contact map with enriched contact density, represent smaller regions that interact significantly more with other loci within the same TAD than with those outside of it (Dixon *et al.*, 2012). Although TADs are somewhat visible in Hi-C maps, it has proven challenging to definitively classify them computationally.

TADs have been shown to correlate with several epigenetic features, including histone markers and CCCTC-binding factor (CTCF) (Ong and Corces, 2014). Histone modifications have proved very tightly linked to Hi-C data, leading to several methods for identifying TADs or predicting Hi-C maps based on ChIP-seq data from a range of histone marks (Bednarz and Wilczyński, 2014; Di Pierro *et al.*, 2017; Huang *et al.*, 2015; Sefer and Kingsford, 2015). Beyond epigenetics, TADs seem to be involved in several other cellular functions. TAD boundaries correlate well with replication timing domains and thus are involved in cell reproduction (Dileep *et al.*, 2015). Lamina-associated domains (LADs), regions near the nuclear lamina associated with gene repression, also frequently coincide with TAD domains (van Steensel and Belmont, 2017). Interruption of TADs has also been shown to alter enhancer/promoter interactions (Lupiáñez *et al.*, 2015), further implicating TAD structure in gene regulatory mechanisms.

Many methods have been developed to identify TADs, first through an HMM-based method (Dixon *et al.*, 2012), and later through optimization of various scoring functions such as InsulationScore (Crane *et al.*, 2015) and Armatus (Filippova *et al.*, 2014). It is not yet clear how to evaluate TAD finder accuracy with no settled ground truth, but two recent benchmarking studies evaluated the performance of 7 different TAD callers, 6 of which overlapped between the two studies, and found no clear consensus on optimal performance (Dali and Blanchette, 2017; Forcato *et al.*, 2017).

Though there is some preliminary evidence that TAD structure is conserved across cell types (Rao *et al.*, 2014) and possibly even species (Dixon *et al.*, 2012), these previous studies have not attempted to identify the locations of structural similarity, nor which genomic features or disease states may correlate with conserved structures. Hi-C data itself is highly variable and likely full of false contacts and missing true contacts, and it is impacted significantly by the choice of data processing and normalization techniques, making it difficult to compare Hi-C maps directly (Yang *et al.*, 2017). Spurious differences like coverage variance can have a strong impact on the apparent similarity of two Hi-C maps, even if the underlying structures are similar. The variability within and between chromosomes is also large, which could mask intrinsic similarity in a global metric. For these reasons, we choose to compare TAD structure rather than Hi-C measurements directly, and we seek regions of locally similar structures rather than one global measure of similarity.

We present a method to identify statistically significantly structurally similar regions of TAD structures, in two main steps. First, we use the information theoretic variation of information (VI) metric (Meilă, 2003) to measure the similarity of all subsets of the two TAD structures, using a dynamic programming algorithm that we designed to efficiently compute this metric. We then select the statistically significant chromosomal regions among those with a locally optimal VI measure through a rigorous null model, and eliminate redundancies from this set. We apply this method to evaluate the similarity of chromosome structure across all pairwise combinations of 23 human samples, across both cancer and non-cancer conditions. The following large-scale comparison of structural concordance and variability across cell types, both globally and on the chromosomal level, identifies biologically meaningful cell type pairs with high structural similarity, and a trend of low structural similarity among cancer cells can be seen at the locations of commonly mutated pan-cancer genes.

This study is the first large-scale study of human chromosomal structural similarity, providing a framework method for future work in this domain. Our comparison of cancer and normal cells reveals insight into the three-dimensional disruptions that occur in cancer genomics, corresponding to the known changes in genome sequence from mutations and structural variants.

2 Materials and methods

We introduce a method which, given two lists of TADs from different samples on one chromosome, identifies the sub-intervals in which the two TAD lists are significantly similar. This is done by optimizing a distance metric, selecting the statistically significant optima and removing redundant intervals with a heuristic.

2.1 Data

Hi-C data were taken from four different studies (Dixon *et al.*, 2012; Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014; The ENCODE Project Consortium, 2012) that were published over 7 years, representing 21 unique human cell types across healthy and diseased states, with 23 Hi-C samples in total, as summarized in Table 1. The samples were chosen to be publicly available and represent a wide array of cell types and conditions. All data were downloaded as raw read (.fastq) files, and processed through the same Hi-C Pro (version 2.8.0) (Servant *et al.*, 2015) pipeline into Hi-C maps, using iterative correction and eigenvector decomposition normalization (Imakaev *et al.*, 2012). All Hi-C maps were generated at 100 kb resolution, the highest shared by all four studies, meaning that each point in the Hi-C matrix corresponds to the number of contacts between two chromosomal intervals of 100 kb each. We call each of these 100 kb segments a genomic bin. This resolution is relatively low because only the more recent studies were sequenced deeply enough for significantly higher resolution. This may impact our results in that we can only capture relatively large-scale regions of structural similarity, but these larger regions are likely to be the most robust. The TAD sets were calculated using version 2.1 of the Armatus software (Filippova *et al.*, 2014), a principled method that is extremely efficient and has performed favorably in recent benchmarking studies (Dali and Blanchette, 2017; Forcato *et al.*, 2017). Armatus requires one parameter, γ, which varies the resolution of TADs that are predicted, biasing the algorithm towards choosing larger or smaller domains. There is no direct relationship between the γ value and the domain sizes, so in order to ensure that all TAD sets have the same approximate median TAD size, the γ value was chosen individually for each Hi-C map and chromosome. The γ value which returned TADs at the expected median size of 880 kb reported in Bonev and Cavalli (2016) was used in each case.

2.2 Overview of the approach

To compare two samples we quantify the similarity between their TAD boundary locations. A TAD is a genomic interval, consisting of a range of bins. A TAD set is then a collection of these intervals identified by a TAD caller. A TAD set can be thought of as a one-dimensional clustering for all of the genomic bins along a chromosome, where the bins within each TAD form a cluster. A natural way to compare clusterings is using a distance metric on these clusterings, and two highly similar clusterings, (i.e. TAD sets, in our case) will be identifiable by a low distance. To identify structurally similar regions, we compute the distances for all possible regions (e.g. all sub-intervals of the chromosome) and then select the regions

Table 1. Hi-C samples used for pairwise comparisons

Cell type	Description	Study	Resolution
GM06990	Blood lymphocyte	Lieberman-Aiden *et al.* (2009)	100kb
K562	Chronic myeloid leukemia	Lieberman-Aiden *et al.* (2009)	100kb
IMR90	Lung fibroblast	Dixon *et al.* (2012)	40kb
hESC	Embryonic stem cell	Dixon *et al.* (2012)	40kb
IMR90	Lung fibroblast	Rao *et al.* (2014)	5kb
GM12878	Blood lymphocyte	Rao *et al.* (2014)	1kb
HMEC	Mammary epithelial	Rao *et al.* (2014)	5kb
HUVEC	Umbilical vein endothelial	Rao *et al.* (2014)	5kb
K562	Chronic myeloid leukemia	Rao *et al.* (2014)	5kb
KBM7	Chronic myeloid leukemia	Rao *et al.* (2014)	5kb
NHEK	Epidermal keratinocyte	Rao *et al.* (2014)	5kb
A549	Adenocarcinomic alveolar basal epithelial	ENCODE (2016)	20kb
Caki2	Clear cell renal carcinoma (epithelial)	ENCODE (2016)	20kb
G401	Rhabdoid tumor kidney epithelial	ENCODE (2016)	20kb
LNCaP-FGC	Prostate carcinoma epithelial-like	ENCODE (2016)	20kb
NCI-H460	Large cell lung cancer	ENCODE (2016)	20kb
Panc1	Pancreas ductal adenocarcinoma	ENCODE (2016)	20kb
RPMI-7951	Malignant melanoma	ENCODE (2016)	20kb
SJCRH30	Rhabdomyosarcoma fibroblast	ENCODE (2016)	20kb
SKMEL5	Malignant melanoma	ENCODE (2016)	20kb
SKNDZ	Neuroblastoma	ENCODE (2016)	20kb
SKNMC	Neuroepithelioma	ENCODE (2016)	20kb
T47D	Ductal carcinoma	ENCODE (2016)	20kb

Note: Cell types listed in italics are non-cancer cell lines.

with statistically significantly low distance. More specifically, we compute VI between the TADs in $[i, j]$ in one sample to the TADs in $[i, j]$ in another sample, for all relevant $[i, j]$.

The distances between all sub-intervals on the chromosome can be represented in an $n \times n$ matrix, where n is the length of the chromosome, and every entry (i, j) represents the distance of the TAD structures in the region between genomic bins i and j. In this full matrix, the elements that are candidates for representing the most similar regions will appear as local minima in the sense that they are smaller than all eight surrounding values. These are intervals that are more similar than any neighboring interval. To determine which are significant we compute p-values for each of these local minima with a strict null model (Section 2.4). Once the statistically significant intervals have been identified, we further select only those which are dominating in the sense that every sub-interval within them has a higher distance measure. These intervals are then called significant structurally similar regions. An overview of the method is seen in Figure 1.

As a distance measure, we use the well-established VI metric, which evaluates the level of agreement between two clusterings based on information theoretic quantities (Meilă, 2003). The VI of two clusterings C and C' can be computed as the normalized sum of the two conditional entropies, where n is the number of elements (genomic bins, in our case) in C and C', as shown below.

$$VI(C, C') = \frac{H(C|C') + H(C'|C)}{\log(n)} \tag{1}$$

where the conditional entropy is defined as

$$H(C|C') = \sum_{i=1}^{k} \sum_{j=1}^{k'} P(i, j) \log \frac{P(j)}{P(i, j)} \tag{2}$$

and C and C' contain k and k' clusters, respectively, and $P(i) = \frac{|C_i|}{n}$, $P(i, j) = \frac{|C_i \cap C'_j|}{n}$. This metric was also used by Filippova *et al.* (2014) to compare their TAD calls with previous methods.

In practice, rather than calculating the entire matrix of VI values for every possible chromosomal sub-interval, we only compute sub-intervals that begin and end at TAD boundaries. Although it seems intuitive that the minimum VI distance would occur exclusively at cluster boundaries, this is not strictly true, as the VI formulation holds no such theoretical guarantees. However, in 10 randomized empirical tests, we observed over 97% of local minima occur at boundary points. Biologically, outside of TAD boundaries we have little understanding of fine-scale chromosome structure, and therefore, it is difficult to interpret the meaning of structural similarity away from these demarcations. We therefore calculate VI values only at TAD boundaries, and analyze this much smaller set of sub-intervals.

While some TAD callers return a partition of the chromosome with no gaps between TADs, Armatus does not explicitly require each bin to be within a TAD. This results in occasional gaps, or non-TAD domains, though they are very rare; on average across all cell types and parameter values, TADs cover 92.02% of the genome. Our method does not distinguish between these non-TADs and TADs; we consider all domains in the same way. The result of this is that we are practically measuring the partition of the chromosome induced by the TAD set, rather than the exact TADs themselves, but this remains a measurement of structural similarity.

2.3 Dynamic programming to compute multiple VI distances

In order to further improve efficiency, we use a dynamic programming algorithm to compute VI for every pair of boundaries. The algorithm is initialized by calculating the VI for every single-TAD interval in both TAD sets. We then proceed by adding the subsequent TAD to each of these intervals, computing the VI of the new interval as a function of the VI values of the smaller two intervals composing it. After computing VI values for every interval of two TADs in each TAD set, we continue increasing the intervals by one TAD until all sub-intervals have been covered.

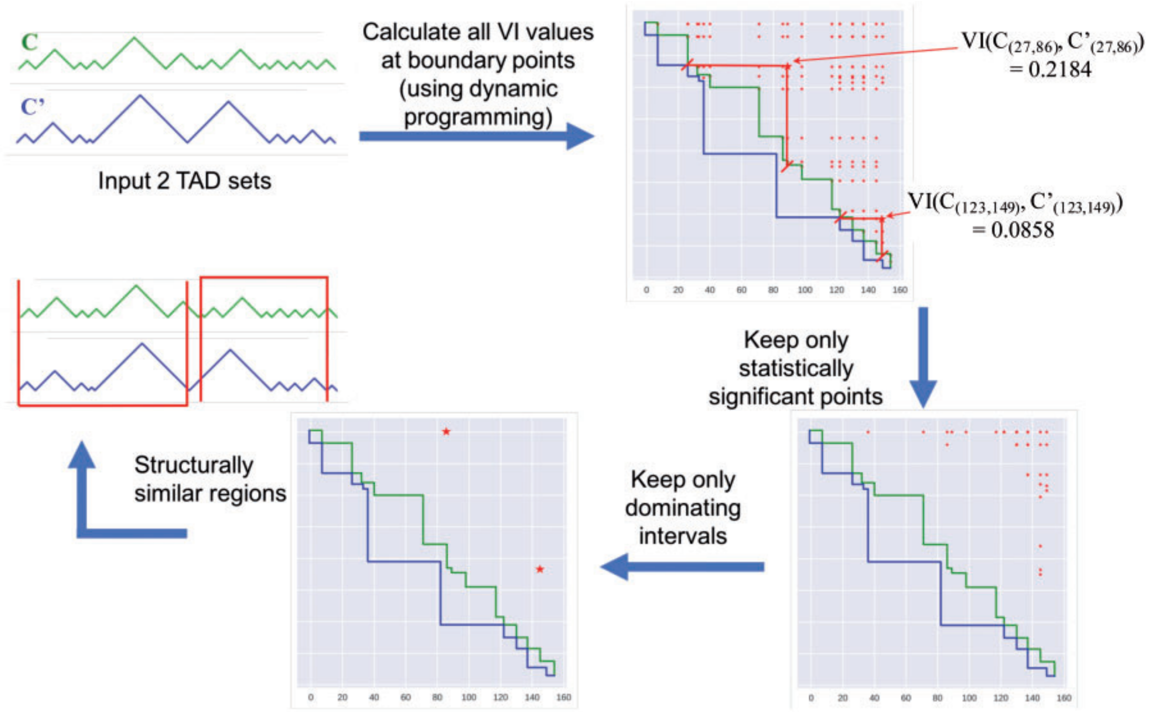
C
C'
Input 2 TAD sets
Calculate all VI values at boundary points (using dynamic programming)
VI(C(27,86), C'(27,86)) = 0.2184
VI(C(123,149), C'(123,149)) = 0.0858
0 20 40 60 80 100 120 140 160
Keep only statistically significant points
Keep only dominating intervals
Structurally similar regions

Fig. 1. Overview of major steps to identify structurally similar intervals

After initialization, at each step we have the VI of both sub-intervals to be combined into a larger interval. Let the sub-interval (i, j) be covered by TAD sets or clusterings C and C', and the sub-interval $(j+1, k)$ be covered by D and D'. We then define the sets CD and $C'D'$ as the concatenation of C and D, and C' and D', respectively, which cover (i, k). In order to compute $VI(CD, C'D')$, there are two cases to consider, illustrated in Figure 2. In the simpler case, there is no TAD in either TAD set that crosses the boundary at j, and the new VI is simply a rescaled sum of the previously calculated VIs:

$$VI(CD, C'D') = \frac{j-i+1}{k-i+1} VI(C, C') + \frac{k-j}{k-i+1} VI(D, D') \tag{3}$$

In the case of a TAD that overlaps the boundary between the two sub-intervals, one conditional entropy term can simply be rescaled as before, but we must adjust the entropy term conditioned on the TAD set including the overlapping TAD. If there is a TAD in $C'D'$ which begins at $s \le j$ and ends at $e > j$ which we refer to as $C'D'_{se}$ (made up of C'_k, the last TAD in C' and D'_1, the first TAD in D'), the new conditional entropies are given below:

$$H(C'D'|CD) = \frac{j-i+1}{k-i+1} H(C'|C) + \frac{k-j}{k-i+1} H(D'|D) \tag{4}$$

$$H(CD|C'D') = \frac{j-i+1}{k-i+1} H(C|C') + \frac{k-j}{k-i+1} H(D|D') \tag{5}$$

$$- \frac{1}{k-i+1} \sum_a |C_a \cap C'_k| \log \frac{|C'_k|}{|C_j \cap C'_k|}$$

$$- \frac{1}{k-i+1} \sum_a |D_a \cap D'_1| \log \frac{|D'_1|}{|D_j \cap D'_1|}$$

$$+ \frac{e-s+1}{k-i+1} H(CD|C'D'_{se})$$

We only compute VI at locations with a boundary in one of the two TAD sets, so we do not encounter the case in which there is an overlapping TAD in both TAD sets. The algorithm ensures that for each VI calculation, at least one TAD set will have a boundary at the point joining the two sub-intervals. In a timing test on 10 randomly chosen cell type pairs and chromosomes, the dynamic programming algorithm reduced the time to compute VI at all boundary points by 58.24% (from 3.384 to 1.413 s). When computing similar intervals and using a permutation test for significance (Section 2.4) between all cell types using all chromosomes, this savings is significant.

2.4 Identifying statistically significant sub-intervals

Once the VI values for all candidate sub-intervals have been calculated, we select the statistically significant regions through an adapted permutation test. For each sub-interval, we fix each TAD set and randomly shuffle the TADs from the other set 1000 times, calculating the VI at each reshuffling. The p-value is then the average of the two fractions of shuffles in which a lower VI was found than the original. The strictness of this null model comes from looking at each interval separately rather than shuffling the TADs across the entire chromosome at once, as well as keeping the TAD lengths fixed in the shuffling. For each interval, we are therefore calculating the likelihood of achieving a more closely matched TAD set while keeping the exact same number of TADs and their lengths. After computing this probability, we control the false discovery rate at a level of 0.05 through the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), keeping only the intervals for which we cannot reject the null hypothesis at this level.

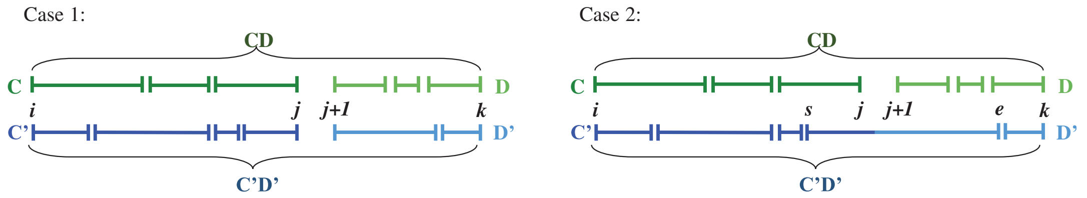

Fig. 2. The two possible cases for dynamic programming algorithm. Case 1 shows the combination of TAD sets where both have a boundary at j, while case 2 illustrates a TAD in one set which overlaps the boundary at j

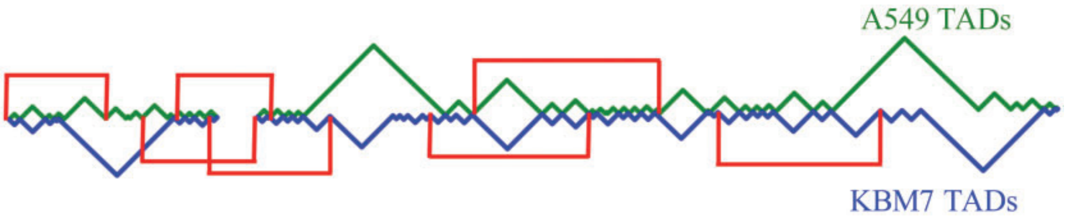

Fig. 3. A sample output of our method, from chromosome 18 of A549 (green) and KBM7 (blue), with the significant, dominating intervals marked by red brackets. The blank space with no TADs in either set corresponds to the centromere, where no reads can be mapped in the Hi-C data

2.5 Dominating intervals

The set of statistically significant intervals still includes many nested intervals, so to remove redundant results we introduce the notion of dominating intervals. An interval is defined as dominating through three tests. First, it must have a *P*-value that passes the statistical significance tests described above. Next, we keep only the intervals that do not contain any sub-intervals with a lower VI value. Finally, if there are still intervals among this set that begin or end at the same point, we keep only the longest. Our method therefore outputs statistically significant intervals that are optimal in the sense that there is no significant sub-interval that represents a higher similarity score. These significant, dominating intervals are the final result of the method, representing chromosomal intervals with significantly similar TAD structures.

2.6 Run time

The method was implemented in Go, and in a test of 10 randomly chosen cell type pairs and chromosomes, identifying the statistically significant dominating intervals took 1h 33min single-threaded with a peak memory usage of 15.9 GB. The timing is heavily dependent on the total number of TADs in both TAD sets, ranging in our test from 9 s to 49 min for a single chromosome.

3 Results

3.1 Comparison of TAD similarity across all 253 pairs of cell types

The method described above was run on all pairwise combinations of the 23 Hi-C maps (253 pairs total), on all 22 autosomal chromosomes, resulting in an average of 5.908 significant intervals per pairwise comparison per chromosome. The average length of a region of structural similarity across all 253 pairwise comparisons is 15.25 Mb, with the longest spanning almost the entirety of chromosome 2 at 219.7 Mb, between NHEK and GM12878, and the shortest of length 1.4 Mb, on chromosome 9 between A549 and NCI-H460. An example of the output intervals can be seen in Figure 3. One artifact of the method is that when an interval intersects a TAD, the algorithm cannot distinguish whether the TAD truly has a boundary at the edge of the interval or not. This leads to the identified regions often containing a TAD at the edge of an interval which is much smaller than its corresponding TAD in the other set, but appears to be a perfect match to the algorithm and therefore is included in the optimal interval.

We can compare the relative conservation and variability of chromosomal regions by looking at the results at the chromosome level. We say that a genomic bin is structurally conserved in one pairwise comparison if it is contained within one of the significant, dominating intervals. On average, each 100 kb genomic bin is structurally conserved in 115.02 out of 253 possible pairwise comparisons, though this varies significantly by location. Figure 4 shows, at each genomic bin, the number of cell type pairs in which the bin was contained in a significant structurally similar interval, across two representative chromosomes. We expect the centromere to be conserved in all cell types, and it does appear as a highly conserved element though not in every pairwise comparison. The reason for this is our significance test, which ensures that no single-TAD interval will be considered significant. There must therefore be enough structural similarity in the regions flanking the centromere to deem any interval spanning the centromere significant. Outside of the centromere, overall variability of this bin-level similarity measure is fairly high. There appear to be chromosomal regions that are extremely similar across most cell types, while others share almost no similarity between any of the pairs we studied.

3.2 Quantifying genome-wide and chromosome-level similarity

The identified structurally similar regions can be further used to measure the genome-wide and chromosome-level similarity. The percent similarity between two genomes (or two chromosomes) was defined as the percentage of the genome (or chromosome) covered by a significant, dominating interval between each pair of cell types. The full set of pairwise percent similarity values is presented as a heat map in Figure 5. For further detail, the top 10 pairs in terms of percent similarity are shown in Table 2.

The two IMR90 samples rank somewhat highly (52.41%, ranked 35 out of 253) in terms of percent similarity, but the two K562 samples are very dissimilar (32.14%, ranked 246 out of 253). This could be explained by the markedly low average similarity of

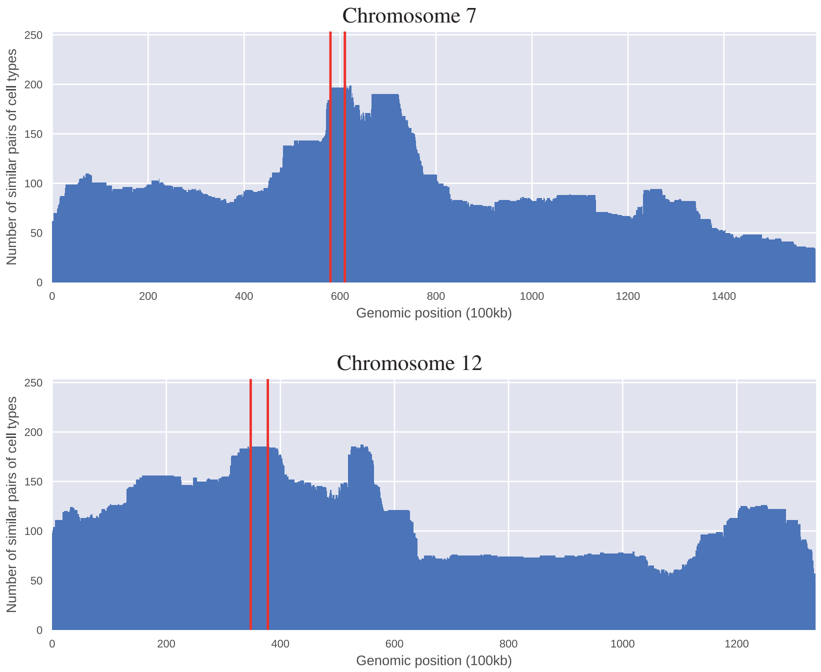

Fig. 4. Structural conservation by genomic location for several chromosomes. The height at each genomic bin represents the number of cell type pairs in which the bin was contained in a significant structurally similar interval. The red lines show the approximate location of the centromere, where reads cannot be mapped and therefore almost all Hi-C maps should be empty in this region, resulting in the appearance of a highly conserved structural element. The significance threshold enforces a minimum number of TADs that must be included in a significant interval, so there are some cell type pairs which differ enough in structure around the centromere that it does not appear as a conserved element in these comparisons

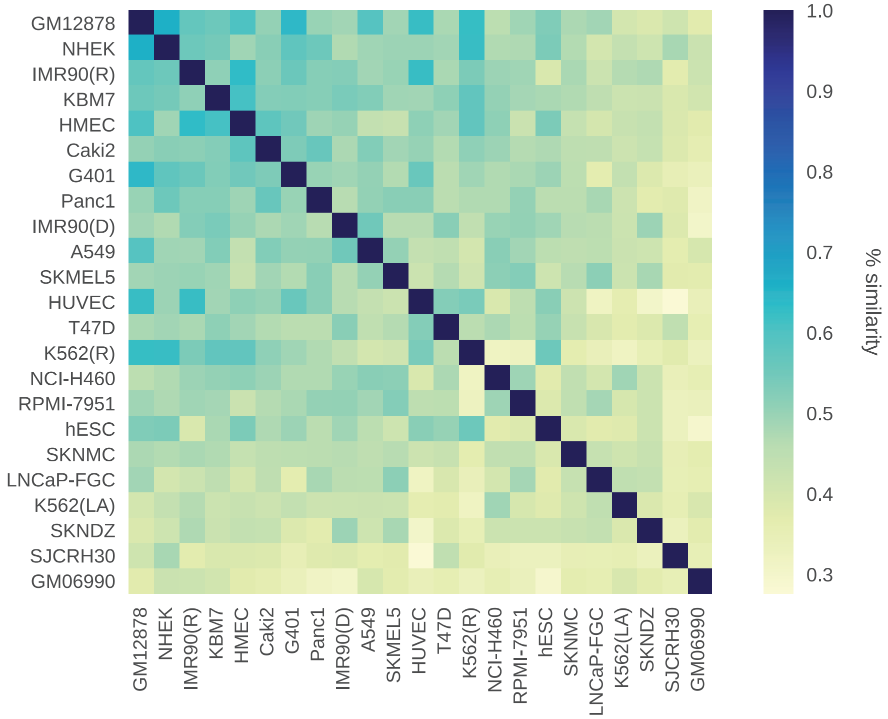

Fig. 5. Heat map of the genome-wide percent similarity between all pairs of cell types studied. Rows are ordered by highest to lowest average pairwise % similarity, calculated by summing the values across each row, dividing by the number of rows and sorting by this average

both Lieberman-Aiden *et al.* (2009) cell types (K562 and GM06990) with all other cell types; both rank in the bottom four of average similarity. This is the oldest dataset we use, so the data may contain more errors or stronger batch effects than the more recently generated samples. If we instead compare the K562 data from the Rao *et al.* (2014) study to KBM7, which comes from the same cancer type (chronic myeloid leukemia), we see a similarity of 57.26%, which ranks them 12 out of 253 pairs. There is some biological

Table 2. Top 10 cell type pairs in percent similarity

Cell type pair	Cell type 1 Description	Cell type 2 Description	% similar
GM12878, NHEK	Blood lymphocyte	Epidermal keratinocyte	66.02
G401, GM12878	Rhabdoid tumor kidney epithelial	Blood lymphocyte	64.64
IMR90 (R), HMEC	Lung fibroblast	Mammary epithelial	63.24
GM12878, K562(R)	Blood lymphocyte	Chronic myeloid leukemia	62.56
K562(R), NHEK	Chronic myeloid leukemia	Epidermal keratinocyte	62.33
GM12878, HUVEC	Blood lymphocyte	Umbilical vein endothelial	62.30
IMR90 (R), HUVEC	Lung fibroblast	Umbilical vein endothelial	62.13
HMEC, KBM7	Mammary epithelial	Chronic myeloid leukemia	60.84
GM12878, HMEC	Blood lymphocyte	Mammary epithelial	60.15
A549, GM12878	Adenocarcinomic alveolar basal epithelial	Blood lymphocyte	59.22

Note: For the cell types which could come from two different samples, the initial of the first author of the data source is in parentheses.

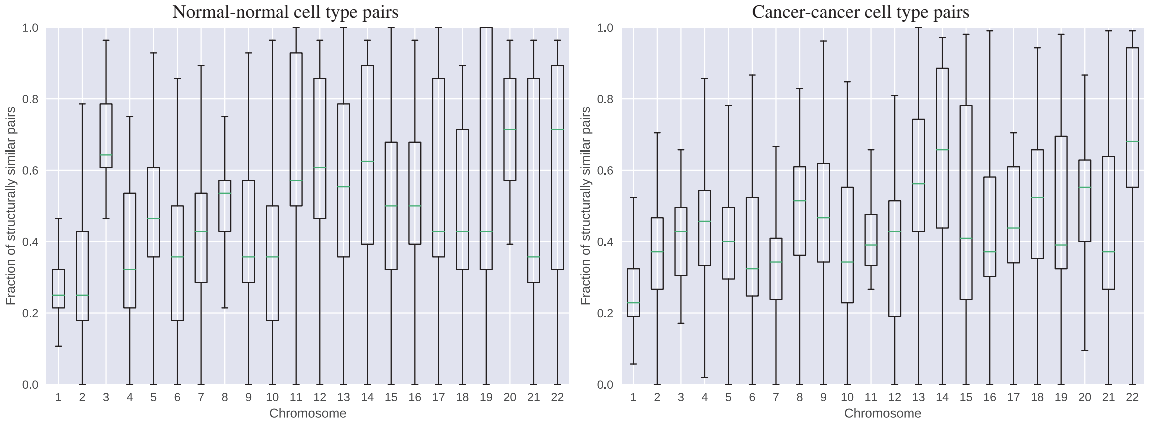

Fig. 6. Box plots showing the distributions of similarity measures across chromosomes, in pairs of cancer cell types and normal cell types. The distribution is over all genomic bins of the given chromosome, and the value at each genomic bin is the fraction of (cancer–cancer or normal–normal) pairs for which the bin is contained in a significant dominating interval

similarity and functional connection between the cell type pairs near the top of the structural similarity measure. The fourth most similar pair (GM12878 and K562) consists of a blood lymphocyte cell line and a chronic myeloid leukemia cell line of lymphoblast morphology, so these come from the same tissue and cell lineage. However, many of the most similar pairs have no apparent biological justification.

Though there are no previous methods quantifying structural similarity to which we can compare, two previous studies counted the number of TAD boundaries (computed using different methods) that they considered overlapping between certain pairs of cell types. Dixon *et al.* (2012), using their method referred to as DomainCaller, reported data indicating a Jaccard index (similarity score between 0 and 1) of 0.52165 between their IMR90 and hESC cell types. Our method reports a comparable similarity score of 0.4902 (fraction similarity across the genome, also between 0 and 1), despite using different methods for data normalization and TAD calling. Rao *et al.* (2014) similarly reported the number of shared TAD boundaries between pairs of cell types including GM12878, which was sequenced much more deeply than the others. Using their own TAD calling method, they identified significantly more TADs in GM12878 than any other cell type because of the higher resolution of the data, so overall their data gave Jaccard indices ranging from 0.2129 to 0.3033 for comparisons of GM12878 to each of IMR90, HMEC, HUVEC, K562, KBM7 and NHEK. However, because there are more GM12878 TADs than any other cell type, this comparison is somewhat skewed. Simply looking at the fraction of each cell type's shared TAD boundaries with GM12878 to its own overall number of TAD boundaries gives similar TAD boundary fractions in the 0.499–0.6688 range. In our analysis, these same cell type pairs ranged in percent genomic similarity levels from 0.5552 to 0.6603. Again, this is using yet another TAD caller and data normalization method, but the level of similarity measured seems to be fairly robust to all of these differences.

At the chromosomal level, these percent similarities and even the ranking of pairwise similarity can vary significantly. Similarity levels averaged over all pairwise comparisons per chromosome vary from 33.70% on chromosome 1 to 69.05% on chromosome 22. For an individual pair, similarity can cover an entire chromosome as in the case of the Caki2 and HMEC which are 100% similar on chromosome 1. In contrast, some pairs have almost no similarity on a chromosome, such as SKMEL5 and the IMR90 sample from Dixon *et al.* (2012), which have 0.963% similarity on chromosome 1. Box plots of the distribution of overall similarity among all normal-normal and cancer-cancer cell type pairs are shown in Figure 6, where several chromosomes, such as 3 and 20, stand out as being particularly more structurally similar among normal cell type pairs than cancer cell type pairs.

3.3 Comparing structural conservation between cancer and non-cancer cell type pairs

Several studies have shown that chromosome structure can be disrupted in a broad range of cancer types (Fudenberg *et al.*, 2011;

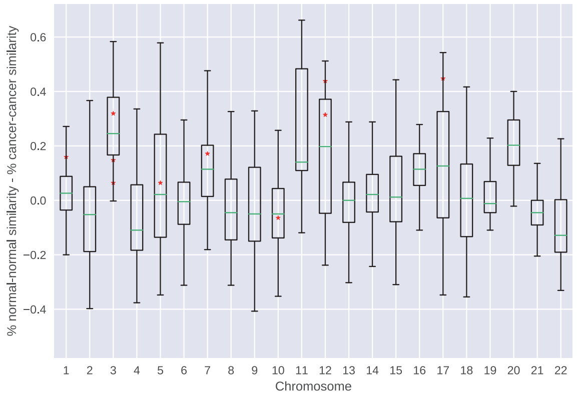

Fig. 7. Box plot showing the chromosome-level distributions of differences between level of structural similarity at all genes in normal cell type pairs and cancer cell type pairs. The red stars represent the differences observed at the 10 most commonly mutated pan-cancer genes from (Kandoth *et al.*, 2013)

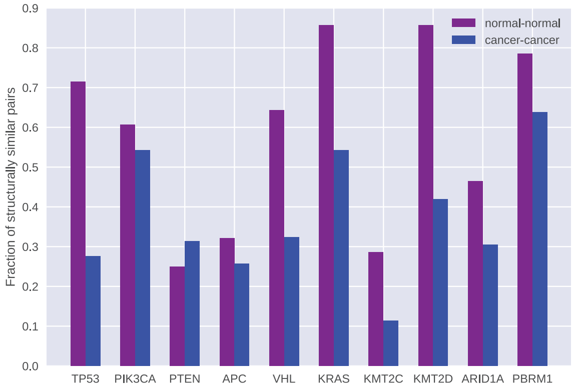

Fig. 8. Relative conservation of cancer–cancer and normal–normal cell type pairs at 10 prominent pan-cancer gene locations. For the cases in which the gene spans multiple bins, the bin for the gene location was chosen as the bin containing the gene's midpoint

Hnisz *et al.*, 2016; Meaburn *et al.*, 2009; Misteli, 2010), and the comparison method above can give a genome-wide view of structural similarity among cell type pairs of all combinations of normal and cancer cell types. Among the 21 unique cell types in our dataset, 14 come from cancer cell lines and the other 7 are non-cancerous (see Table 1). Including the two duplicate cell types, this gives 28 pairs of two normal cell types, and 105 pairs of two cancer cell types. Globally, the normal-normal pairs show slightly higher average structural conservation, but the difference is not significant: 44.17% average similarity among cancer-cancer pairs, and 49.02% similarity among normal-normal pairs.

However, we find that there is more structural conservation at the regions around established pan-cancer genes in normal-normal cell type pairs than in cancer-cancer pairs, which may point to the structural disruption that occurs in conjunction with cancer mutations. Looking at the top 10 most commonly mutated pan-cancer genes from a large-scale study of data from The Cancer Genome Atlas (Kandoth *et al.*, 2013), we can see that the structure around most of these genes is more conserved among normal cell type pairs than cancer pairs (Figs. 7 and 8). Figure 7 shows the distributions for each chromosome of the percent similarity among cancer-cancer pairs subtracted from normal-normal pairs. A value above zero indicates higher structural similarity among normal-normal cell type pairs. Despite 10 out of 22 chromosomes having lower than zero average difference, 9/10 cancer genes are located on chromosomes with a positive average value. In addition, we note that three of these genes are located on chromosome 3, which has the highest average difference between structural similarity in normal-normal pairs compared to cancer-cancer pairs. The prevalence of mutations in cancer cells on genes located on chromosome 3 and the disruption caused by the mutations may result in variable structural changes in cancer cells.

Looking more closely at these 10 gene locations, we note that normal-normal pairs are more structurally similar at nine of these 10 gene locations (Fig. 8). Over all human gene loci, 57.77% show a higher fraction of structurally similar normal-normal pairs than cancer-cancer pairs, which gives a probability of 0.03441 (using the hypergeometric test) of pulling at least 9/10 random genes with higher normal-normal structural conservation, suggesting that the pattern of Figure 8 is statistically significant. If we further restrict the null model to the probability of finding at least 9/10 genes from the same chromosomes as our pan-cancer genes, the p-value increases to 0.1425, which is expected based on the distributions shown in Figure 7. Though this value is above the traditional 0.05 p-value cutoff, the combination of results suggests a role for 3D structure disruption around mutated genes in cancer cell types. In order to validate and confirm this conclusion, we would need more Hi-C samples.

4 Discussion and conclusions

We have presented the first method to quantify local chromosomal structural similarity and have used it to perform a large-scale comparison of TAD structure across 23 human samples from both cancer and noncancerous conditions. We note the variability among structural components both globally and by chromosome, as well as between cancer and normal cell types. This led to the new observation that pan-cancer gene locations show more structural variability among cancer cells than among normal cells.

Though the analysis was performed using only TADs from the Armatus software at one individually optimized parameter setting, our results are in line with the levels of structural similarity reported by other studies using other TAD finders and pre-processing pipelines. Further study in this area will involve testing the robustness of these results to the choice of TAD caller, as well as the Hi-C data resolution and normalization. Another tunable aspect of this method is the choice of distance metric, for which we used VI. Though VI is a well-established and general metric for calculating clustering similarity, there are many other metrics which fit the same criteria.

Beyond the methodological choices, our results are somewhat dictated by the available Hi-C samples. Hi-C is a fairly expensive and time-consuming protocol, so the amount of data available is much smaller than other genomic data types such as RNA-seq. We selected samples from prominent studies in the field, but without more data it is difficult to determine whether chromosome structure can be tissue-specific or cancer type-specific, or any number of other possibilities. As more data becomes available, the robustness of the results of such a structural comparison will significantly increase.

Given the set of samples we used, it is difficult to determine the level of batch effects or other protocol-specific differences influencing our results. The extremely low similarity values for both samples from the Lieberman-Aiden *et al.* (2009) study seem to suggest some batch effects or protocol-specific variations, but otherwise the similarity clustering did not simply group cell types from the same studies. This concern could be further studied or mitigated with more Hi-C samples.

Another concern with the data is specific to the cancer samples, which are likely to be highly mutated and contain genomic structural variants. Despite this, we still map them to the reference (non-cancer) genome. Some of the areas where we see structural differences across cancer cells may simply be due to an inability to map reads with high mutation levels, rather than a variation in three-dimensional structure. Through further advances in long-read technology and genome mapping and assembly, it may become easier to avoid these concerns and study three-dimensional structure more directly. Some work has begun in this area, combining structural variant detection with Hi-C data (Chakraborty and Ay, 2017).

Our method and analysis represents a first step towards understanding the conservation and changes in chromosome structure across human cell types and disease states. We provide the first genome-wide structural comparison of cancer and non-cancer genes, as well as a systematic pairwise analysis of similarity across 23 human cell types. As Hi-C data becomes more widely available and reliable, the ability to compare and identify structurally similar or variable regions may provide even more insight into the mechanisms and influence of chromosome architecture on gene regulation and cellular functioning.

Acknowledgements

The authors would like to thank the members of the Kingsford Group for valuable discussions and comments on the manuscript, especially Mingfu Shao, Guillaume Marçais, Brad Solomon and Hongyu Zheng.

Funding

This research was supported in part by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through Grant GBMF4554 to C.K., by the US National Science Foundation (CCF-1256087, CCF-1319998) and by the US National Institutes of Health (R01HG007104 and R01GM122935). Research reported in this publication was supported by the NIGMS of the NIH under award number P41GM103712. This study was partially funded by The Shurl and Kay Curci Foundation.

Conflict of Interest: none declared.

References

Ay,F. *et al.* (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.*, **24**, 999–1011.

Bednarz,P. and Wilczyński,B. (2014) Supervised learning method for predicting chromatin boundary associated insulator elements. *J. Bioinformatics Comput. Biol.*, **12**, 1442006.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B.*, **57**, 289–300.

Bonev,B. and Cavalli,G. (2016) Organization and function of the 3D genome. *Nat. Rev. Genet.*, **17**, 661.

Cavalli,G. and Misteli,T. (2013) Functional implications of genome topology. *Nat. Struct. Mol. Biol.*, **20**, 290.

Chakraborty,A. and Ay,F. (2017) Identification of copy number variations and translocations in cancer cells from Hi-C data. *Bioinformatics*, **34**, 338–345.

Crane,E. *et al.* (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, **523**, 240.

Cremer,T. and Cremer,C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, **2**, 292.

Dali,R. and Blanchette,M. (2017) A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res.*, **45**, 2994–3005.

Dekker,J. *et al.* (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.

Di Pierro,M. *et al.* (2017) De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Proc. Natl. Acad. Sci. USA*, **114**, 12126–12131.

Dileep,V. *et al.* (2015) Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program. *Genome Res.*, **25**, 1104–1113.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376.

Duggal,G. *et al.* (2014) Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Res.*, **42**, 87–96.

Filippova,D. *et al.* (2014) Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.*, **9**, 14.

Forcato,M. *et al.* (2017) Comparison of computational methods for Hi-C data analysis. *Nat. Methods*, **14**, 679.

Fudenberg,G. *et al.* (2011) High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.*, **29**, 1109–1113.

Hnisz,D. *et al.* (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, **351**, 1454–1458.

Huang,J. *et al.* (2015) Predicting chromatin organization using histone marks. *Genome Biol.*, **16**, 162.

Imakaev,M. *et al.* (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999.

Kandoth,C. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333.

Le Dily,F. *et al.* (2014) Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev.*, **28**, 2151–2162.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Lupiáñez,D.G. *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012–1025.

Lupiáñez,D.G. *et al.* (2016) Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet.*, **32**, 225–237.

Meaburn,K.J. *et al.* (2009) Disease-specific gene repositioning in breast cancer. *J. Cell Biol.*, **187**, 801–812.

Meilă,M. (2003) Comparing clusterings by the variation of information. In: Schölkopf,B. and Warmuth,M.K. (eds.) *Learning Theory and Kernel Machines*. Springer, Berlin, Heidelberg, pp. 173–187.

Misteli,T. (2010) Higher-order genome organization in human disease. *Cold Spring Harb. Perspect. Biol.*, **2**, a000794.

Moindrot,B. *et al.* (2012) 3D chromatin conformation correlates with replication timing and is conserved in resting cells. *Nucleic Acids Res.*, **40**, 9470–9481.

Ong,C.-T. and Corces,V.G. (2014) CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.*, **15**, 234.

Pope,B.D. *et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature*, **515**, 402.

Ramani,V. *et al.* (2016) Mapping 3D genome architecture through in situ DNase Hi-C. *Nat. Protoc.*, **11**, 2104.

Rao,S.S.P. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Ryba,T. *et al.* (2010) Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.*, **20**, 761–770.

Sauerwald,N. *et al.* (2017) Chromosomal dynamics predicted by an elastic network model explains genome-wide accessibility and long-range couplings. *Nucleic Acids Res.*, **45**, 3663–3673.

Sefer,E. and Kingsford,C. (2015) Semi-nonparametric modeling of topological domain formation from epigenetic data. In: Pop,M. and Touzet,H. (eds.) *Algorithms in Bioinformatics*, *WABI 2015*. Lecture Notes in Computer Science, Vol. 9289. Springer, Berlin, Heidelberg, pp. 148–161.

Servant,N. *et al.* (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*, **16**, 259.

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57.

van Steensel,B. and Belmont,A.S. (2017) Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell*, **169**, 780–791.

Yang,T. *et al.* (2017) HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.*, **27**, 1939–1949.