Topological data analysis reveals principles of chromosome structure throughout cellular differentiation

Natalie Sauerwald[1], Yihang Shen[1] and Carl Kingsford[1,*]

[1]Computational Biology Department, Carnegie Mellon University, Pittsburgh, 15213, USA.
[*] Corresponding author

February 4, 2019

Abstract

Three-dimensional chromosome structure has a significant influence in many diverse genomic processes and has recently been shown to relate to cellular differentiation. Many methods for describing the chromosomal architecture focus on specific substructures such as topologically-associating domains (TADs) or compartments, but we are still missing a global view of all geometric features of chromosomes. Topological data analysis (TDA) is a mathematically well-founded set of methods to derive robust information about the structure and topology of data sets, making it well-suited to better understand the key features of chromosome structure. By applying TDA to the study of chromosome structure through differentiation across three cell lines, we provide insight into principles of chromosome folding generally, and observe structural changes across lineages. We identify both global and local differences in chromosome topology through differentiation, identifying trends consistent across human cell lines.

Availability: Scripts to reproduce the results from this study can be found at https://github.com/Kingsford-Group/hictda

Contact: carlk@cs.cmu.edu

1 Introduction

The three-dimensional shape of chromosomes has significant influence in many critical cellular processes, including gene expression and regulation (Cremer and Cremer, 2001; Cavalli and Misteli, 2013; Le Dily *et al.*, 2014; Duggal *et al.*, 2014; Rennie *et al.*, 2018), replication timing (Ryba *et al.*, 2010; Moindrot *et al.*, 2012; Pope *et al.*, 2014; Ay *et al.*, 2014)), and overall nuclear organization (Yaffe and Tanay, 2011; Ramani *et al.*, 2016; Chen *et al.*, 2018). Alterations in the 3D structure of the genome have been tied to many cancers (Meaburn *et al.*, 2009; Misteli, 2010; Fudenberg *et al.*, 2011; Hnisz *et al.*, 2016), developmental conditions including deformation or malformation of limbs (Lupiáñez *et al.*, 2016), and severe brain anomalies (Spielmann *et al.*, 2018). The wide range of processes related to chromosome structure suggests that understanding this component is crucial to a broader understanding of many genomic mechanisms, yet it remains a challenge to study this architecture and identify meaningful structures within the complex system.

In particular, the process of differentiation, by which a cell changes to a new cell type, is critical to all multi-cellular life, but the mechanisms behind this process remain an active field of research with recent work suggesting a role for chromosome structure in this process. Structural changes have been observed in chromosomes through lineage specification, both across several stages of human cardiogenesis (Fields *et al.*, 2017) as well as across human embryonic stem cells (ESCs) and four human ES-cell-derived lineages (Dixon *et al.*, 2015). Fields *et al.* (2017) identified both global and local structural dynamics, observing transitions from repressive to active compartments around cardiac-specific genes as they are upregulated through differentiation. Dixon *et al.* (2015) also noted structural dynamics across hierarchical scales during development, with some corresponding gene expression changes. In addition, it has been shown that different chromatin configurations can determine different paths during development by physically separating or connecting enhancers with particular developmental genes (Kragesteen *et al.*, 2018). All of this work suggests that chromosome structure plays an important role in the process of cellular differentiation, with structural alterations related to regulatory mechanisms underlying this key cellular process.

Chromosome structure can be measured by a number of variants of the chromosome conformation capture protocol (Dekker *et al.*, 2002), including Hi-C (Lieberman-Aiden *et al.*, 2009) which permits genome-wide measurements of the chromosomal architectures of a population of cells. Hi-C quantifies physical proximity by counting cross-linkage frequencies between genomic segments. Because of the dependence on cross-linking, which is likely to induce both false positive and false negative contacts, and the heterogeneity within cell populations, Hi-C can be very challenging to analyze. Many methods have focused on identifying local structures called topologically-associating domains (TADs) (Dixon *et al.*, 2012; Rao *et al.*, 2014; Crane *et al.*, 2015; Weinreb and Raphael, 2015; Fillipova *et al.*, 2014; Norton *et al.*, 2018), others on detecting differential interactions between two Hi-C matrices (Djekidel *et al.*, 2018; Lun and Smyth, 2015; Bunnik *et al.*, 2018), and still others on translating the Hi-C contact values into a 3D model of chromosome structure (Lesne *et al.*, 2014; Paulsen *et al.*, 2017; Serra *et al.*, 2017; Trieu and Cheng, 2015). However, it has proven challenging to study large-scale structures across the entire genome.

A class of techniques called "Topological data analysis (TDA)" has gained prominence recently as a generalized, mathematically grounded set of methods for identifying and analyzing topological and geometric structures underlying data. Emerging from work in applied topology and computational geometry, TDA aims to infer information about the robust structures of data (Chazal and Michel, 2017). These methods have already been applied to various biological contexts (Cámara, 2017), including in studies of gene expression at the single cell level (Rizvi *et al.*, 2017), viral reassortment (Chan *et al.*, 2013), horizontal evolution (Camara *et al.*, 2016), cancer genomics (Nicolau *et al.*, 2011; Arsuaga *et al.*, 2015), and other complex diseases (Li *et al.*, 2015; Hinks *et al.*, 2016). Similar methods have also been used in tools to enable large-scale biological database searching (Yu *et al.*, 2015). The two main methods of TDA are Mapper, a dimensionality reduction framework and visualization method, and persistent homology, an algorithm for extracting geometric and topological structures which describe the underlying data.

Given its rigorous mathematical foundation and ability to identify important topological structures, TDA is very well-suited to the analysis of Hi-C data. Emmett *et al.* (2015) first applied these methods to human Hi-C data a few years ago, though computation limitations at the time restricted this analysis to only one chromosome at 1Mb resolution. More recently, TDA was used to analyze the similarities between single-cell Hi-C maps (Carriere and Rabadan, 2018). Carriere and Rabadan (2018) first computed pairwise distance between all single-cell Hi-C contact matrices, and applied TDA to the distance matrix between single cells rather than applying TDA directly to the Hi-C data, analyzing the results with Mapper. In this paper, we use persistent homology to identify the geometric structures in human chromosomes, and study how they change throughout lineage specification and differentiation.

This work presents the first use of TDA to study the chromosome structures of all 22 human autosomal chromosomes, providing insight into the structural changes involved in cellular differentiation. We describe the patterns underlying geometric structures of Hi-C data, noting that many of these patterns can be explained by the linearity of the chromosome. Additionally, we compare the topologies of 14 cell types representing various stages of differentiation and various cell lines, and note that the topological similarity is largely dictated by cell line rather than differentiation stage. Looking more closely at differentiation, changes along each lineage on each chromosome are quantified, demonstrating that several chromosomes display local changes consistent across cell lines, and others appear stable throughout differentiation.

2 Methods

2.1 Overview of TDA

The premise of TDA is that data points are sampled from an unknown continuous geometric structure. This structure can be described by topological properties preserved under continuous deformations of the space, such as the number and size of connected components, loops or holes it contains. TDA approximates a continuous geometry by building a *simplicial complex*, or a network of edges and triangles, from the nodes of the given data points. More complex simplices of n dimensions can be generated for high-dimensional data, but for the purposes of Hi-C , which is only three dimensional, our simplicial complexes are made up only of simplices of dimension at most 2, i.e., nodes, edges, and triangles. We use a Vietoris-Rips (VR) complex, which is a set of simplices produced by adding edges between all nodes with distance less than a given α, and a triangle between all sets of three nodes for which each pair is no more than α apart. Together these components describe a structure built from the data, from which the topological features of the underlying space can be described and quantified through a process called *persistent homology.*

Definition: Vietoris-Rips complex Given a set of points X in a metric space (M, d) and a real number $\alpha \geq 0$, The Vietoris-Rips complex is the set of simplices $\{[x_0, ..., x_k]\}$ such that $d(x_i, x_j) \leq \alpha$ for all

(i, j), with k less than or equal to a given maximum dimension (Chazal and Michel, 2017).

The analysis of simplicial complexes and their topological properties is based in homology theory, which defines the topological properties of any given dimension of a space. These properties can be represented by homology groups $H_0(X), H_1(X), H_2(X), ..., H_n(X)$. A homology group H_k represents k-dimensional "holes". For example, H_0 represents the connected components of the VR complex, H_1 represents one-dimensional loops, and H_2 represents two-dimensional voids (Spanier, 1966).

Given a set of data points X, we build VR complexes for different values of parameter α. The basis of persistent homology is the idea that features that persist in the VR complexes across values of α are the key topological features of the space generated by the data. A feature from persistent homology is therefore described by an interval $[b, d]$, where b represents the birth time of the feature, or the smallest value of α at which the feature is found, and d, called death time, the smallest value of α at which the feature no longer exists. These features are visualized in two ways: *persistence diagrams* and *barcode plots*. Persistence diagrams are sets of (b, d) points in the Euclidean half-plane above the diagonal. Barcode plots display the same information, but with each homology group shown as an interval $[b, d]$, and the y value is the index within the set of homology groups. For more technical details on TDA and persistent homology, see Carlsson (2009), Carlsson (2014), Edelsbrunner and Harer (2010), or Wasserman (2018).

2.2 Applying TDA to Hi-C

The methods of TDA use a distance matrix that describes the distances between all data points. Although Hi-C data is interpreted as describing the 3D distances between chromosomal segments, the values of a Hi-C matrix are contact counts rather than distance values, where a high contact count implies a low distance. We use the following transformation to convert a normalized Hi-C matrix M to a distance matrix K:

$$K_{i,j} = 1 - \begin{cases} 1 & i = j \\ \frac{1}{m}\log(M_{i,j} + 1) & i \neq j \end{cases}$$

where $m = 1.01 \max_{i,j \leq D}(\log(M_{i,j}) + 1)$, and D is the number of rows in the contact matrix. A pseudo-count of 1 is added to all off-diagonal values in the Hi-C matrix to avoid taking a logarithm of zero, and the factor of 1.01 is included to ensure that all distances where $i \neq j$ are nonzero.

We use GUDHI (Maria *et al.*, 2014), a Python library for TDA, to compute persistent homology from these distance matrices, with the maximum dimension of simplices created is 2.

2.3 Null models

The H_1 structures identified by TDA represent "loops" in the input data, or one-dimensional holes in the simplicial complex. We will use the term loop interchangeably with H_1 structure but it is important to note that these are not loops in the traditional sense of chromatin loops. The loops identified by TDA may be surrounded by non-consecutive genomic segments, unlike the continuous loops generally studied in chromatin. In order to understand the TDA loop structures, three separate null models of distance matrices representing various properties of the Hi-C data were also analyzed and compared to the H_1 structures of the original Hi-C data.

The three null models are defined as follows.

- Random permutation: all Hi-C distance values are permuted randomly, preserving only the symmetry of the distance matrix and the values themselves.

- Edge permutation: the distance values along each row of the distance matrix were permuted randomly, preserving both the degree of and set of distances for each node but randomly changing their assignments. The corresponding columns were permuted in the same way to preserve symmetry.

- Linear dependence: a new distance matrix is created, in which each diagonal beyond the main diagonal preserves the same mean and standard deviation of the original data, with Gaussian noise added. The dominant pattern of the Hi-C distance matrices is a decrease in distance as the difference between the row index and column index increases. This model represents this same pattern, but does not include any additional structures such as TADs or compartments.

2.4 Metrics to compare persistence diagrams

In order to derive stability results for TDA, Carlsson (2014) proposed a metric called the *bottleneck distance* that quantifies the difference between two persistence diagrams. The bottleneck distance is based on a perfect bipartite matching g between two persistence diagrams dgm_1 and dgm_2, where points in either persistence diagram can also be matched to any point along the diagonal. The formula for computing the bottleneck distance d_B is:

$$d_B(\mathrm{dgm}_1, \mathrm{dgm}_2) = \inf_{\text{matching } g} \max_{(x,y) \in g} ||x - y||_\infty.$$

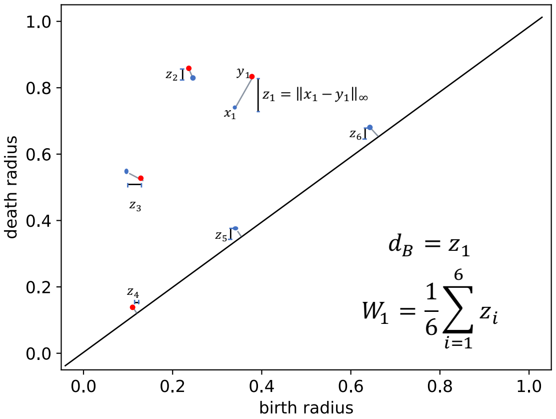

Figure 1: Illustration of the bottleneck distance (d_B) and the Wasserstein distance with $p = 1$ (W_1). Blue and red dots represent points of two different persistence diagrams, and grey lines denote the matching between them. The bottleneck distance represents the largest distance between two matched points, while the Wasserstein distance is the average of all distances between matched points.

The bottleneck distance quantifies the similarity between two persistence diagrams by the maximum distance between two points in a matching. It is therefore a measure of the greatest outlier, rather than the closeness of all pairs of points.

In order to avoid this concern, the related Wasserstein distance metric can also be used to quantify the difference between two persistence diagrams. The Wasserstein distance, W_p, is defined, for some $p \geq 1$, as:

$$W_p(\mathrm{dgm}_1, \mathrm{dgm}_2) = \inf_{\text{matching } g} \sum_{(x,y)\in m} ||x - y||_\infty^p.$$

This measures the total distance between matched loop structures, and therefore gives an overall quantification of the global similarity (Chazal and Michel, 2017). For this work, $p = 1$ to make these values comparable to the bottleneck distances, and we additionally normalized by the cardinality of the matching, resulting in a value which represents the average distance between all points in the two persistence diagrams. See Figure 1 for an illustration of the two distance metrics.

3 Results

3.1 Data

We analyzed Hi-C samples from 14 conditions, representing several differentiation lineages from two different studies (Fields *et al.*, 2017; Dixon *et al.*, 2015). Two of these lineages represent paths through human cardiogenesis, starting with stem cells and continuing through the mesoderm (MES), cardiac progenitor (CP), and cardiac myocyte (CM) stages. One line, from RUES2 cells, begins with embryonic stem cells (ESC), and also includes a fetal heart tissue sample. The study authors also collected data from WTC11 cells, beginning with a human induced pluripotent stem cell (PSC), then collecting data at the same stages as the RUES2 cells: MES, CP and CM (Fields *et al.*, 2017). The third Hi-C data set represents still another differentiation starting point, using H1 ESC cells to generate four human ES-cell-derived lineages: mesendoderm (ME), mesenchymal stem (MS) cells, neural progenitor (NP) cells, and trophoblast-like (TB) cells (Dixon *et al.*, 2015). All data is described in Table 1, including accession codes. Samples from all 14 conditions included two replicates each. All of the Hi-C data was processed from raw reads to normalized contact matrices at 100kb using the HiC-Pro pipeline (Servant *et al.*, 2015) and iterative correction and eigenvector decomposition (ICE) normalization (Imakaev *et al.*, 2012). In order to maximize coverage, we combined all of the reads from replicates to produce one Hi-C matrix per sample.

We generated ten of each of the null models for every RUES2 cell type and each chromosome from 13 to 22. The null models on longer chromosomes proved not to be computationally feasible, but the patterns across the chromosomes we were able to model were remarkably consistent (see Figures S15, S16, and S17), suggesting that the additional data from all three null models on chromosomes 1 through 12 would follow similar patterns.

3.2 Persistent homology in Hi-C data

We visualize the persient homology groups in the two ways described previously, persistence diagrams and barcode plots. We focus here on H_0 and H_1 structures and observe a very distinctive pattern in both structure classes across chromosomes and cell types in all of our Hi-C data. The majority of H_0 structures die off at a radius of somewhere between $\alpha = 0.1$ and $\alpha = 0.2$, suggesting that many new edges are formed near these values, and relatively few H_0 structures persist after this. TDA therefore quickly recovers the linear structure of the chromosome. The H_1 structures tend to have short lifespans (they are close to the diagonal in the persistence diagrams), and most are born at $\alpha \sim 0.6 - 0.8$, though there are consistently a few loops born earlier. A representative barcode plot and persistence diagram can be seen in Figure 2, and barcode plots for all samples on all autosomal chromosomes can be seen in Figures S1–S14.

Table 1: All Hi-C samples used for this study.

Sample name	Description	SRA Accessions	Study
RUES2 ESC	embryonic stem cell	SRX3375347, SRX3375348	Fields *et al.* (2017)
RUES2 MES	mesoderm	SRX3375349, SRX3375350	Fields *et al.* (2017)
RUES2 CP	cardiac progenitor	SRX3375351, SRX3375352	Fields *et al.* (2017)
RUES2 CM	cardiac myocyte	SRX3375353, SRX3375354	Fields *et al.* (2017)
RUES2 FetalHeart	fetal heart tissue	SRX3375355, SRX3375356	Fields *et al.* (2017)
WTC11 PSC	pluripotent stem cell	SRX4958481, SRX4958482	Fields *et al.* (2017)
WTC11 MES	mesoderm	SRX4958483, SRX4958484	Fields *et al.* (2017)
WTC11 CP	cardiac progenitor	SRX4958485, SRX4958486	Fields *et al.* (2017)
WTC11 CM	cardiac myocyte	SRX4958487, SRX4958488	Fields *et al.* (2017)
H1 ESC	embryonic stem cell	SRX378271, SRX378272	Dixon *et al.* (2015)
H1 ME	mesendoderm	SRX378273, SRX378274	Dixon *et al.* (2015)
H1 MS	mesenchymal stem cell	SRX378275, SRX378276	Dixon *et al.* (2015)
H1 NP	neural progenitor	SRX378277, SRX378278	Dixon *et al.* (2015)
H1 TB	trophoblast-like cells	SRX378279, SRX378280	Dixon *et al.* (2015)

3.3 Loop analysis

In order to understand the H_1 structures, the patterns observed in real Hi-C data were compared to our null models. Traditional chromatin loops have been shown to correlate with gene activation and bring together enhancers and promoters (Rao *et al.*, 2014). More work is needed to understand the biological significance of the loop structures from TDA, but Emmett *et al.* (2015) suggested that they may represent transcription factories or other regulatory interactions.

As a representative example, barcode plots of all loop structures of chromosome 14 in RUES2 MES cells can be seen in Figure 3a, along with the null models from this data.

One of the most striking patterns in comparing these null models to the true data is that the loops in Hi-C tend to be born at a much higher value of α, and survive only a short time. The model which most closely resembles this pattern is the linear dependence model, but the distribution of birth times shows that the linear dependence model has a long, significant tail towards the earlier birth times which is absent in real Hi-C data (Figure 3b). The short life span, which will generally correspond to smaller loops, is also much more consistent with the linear dependence model, although somewhat more pronounced in Hi-C (Figure 3c). The fact that the linear dependence null model shows these same patterns as Hi-C data suggests that the linear property (nodes that are close together in index, or linear distance, are also close in 3D space) is sufficient to explain the short loops and birth times concentrated at high values of α. Long loops appear to be created when nodes with a large difference in their indices (large linear distance) are close together, as demonstrated by the loops with very long life spans in the two permutation models. This appears to be very rare in Hi-C data; the majority of loop interactions we observe are relatively short, consistent with findings from more traditional chromosome loop structures which have been shown to be almost exclusively between loci less than 2Mb apart (Rao *et al.*, 2014).

The greatest difference between the linear dependence model and Hi-C data can be seen in the number of loops identified (Figure 3d). Although the linear dependence model is again the most similar to real data in this regard, it typically still contains over three times the number of loops as the corresponding Hi-C matrix. This observation could be explained by the existence of topologically associating domains (TADs) or compartments in real chromosomes, which are absent from the linear dependence model but a defining characteristic of Hi-C. These structures serve to isolate chromosome segments from each other, which likely prevents the formation of loops between them. If loops can largely be formed within a TAD or compartment, this would significantly restrict the total number of feasible loops which could explain the pattern we see in the persistence diagrams of Hi-C data.

3.4 Comparing topologies across differentiation

Although there are some evident changes in topological structures measured by TDA through differentiation, the larger differences exist between the three main cell lines (RUES2, WTC11, and H1, see Figure 4). Interestingly, there does not seem to be any more similarity, measured by bottleneck distance averaged over all chromosomes, between cells at the same stages of differentiation across the three cell lines than cells at different stages of differentiation. For example, the distance value between cardiac progenitor cells from RUES2 and WTC11 appears no lower than the value between RUES2 CP and WTC11 CM cells, and H1 ESC and RUES2 ESC seem no more similar to each other than H1 ESC and RUES2 MES or RUES2 ESC and H1 ME cells. Global topological features identified by TDA at the genome-wide scale therefore seem to be determined more by cell line than differentiation stage.

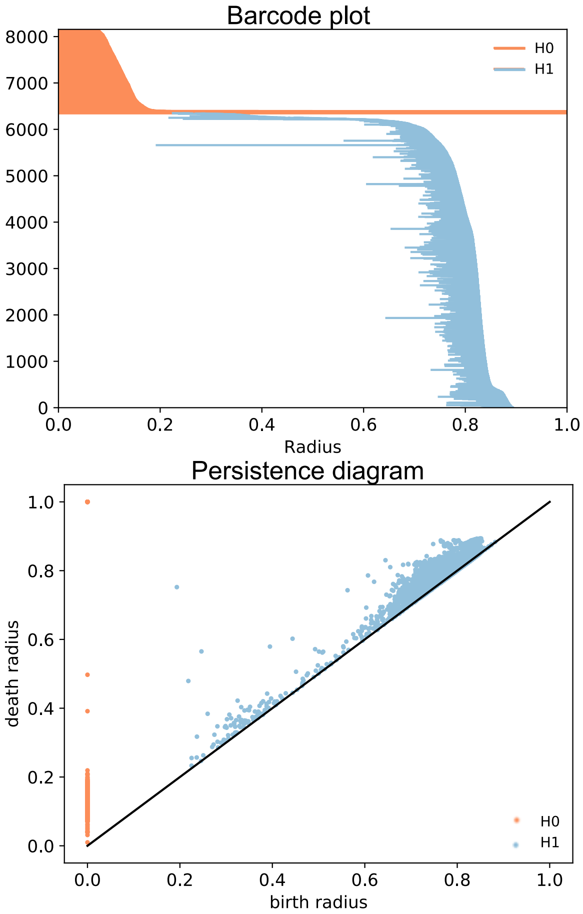

Figure 2: Representative example of a barcode plot and persistence diagram from Hi-C data. These figures were created from chromosome 5 of WTC11 cardiac progenitor cells, and summarize the output from the persistent homology computation. Each point or bar represents one structure, defined by the radius at which the structure can first be seen (birth radius), and the last radius before the structure no longer exists (death radius). These figures are two ways to represent the same persistent homology information.

3.5 Chromosome-specific topological changes through differentiation

Using both bottleneck and Wasserstein distances, we identify specific chromosomes with global and local changes through the various stages of differentiation (Figure 5). Recall that the bottleneck distance is an L_∞ norm, and therefore measures the maximum distance between paired H_1 structures (i.e. the greatest outlier), while the Wasserstein distance is an average distance between all H_1 structures. A low Wasserstein distance points to global similarity between conditions, because on average, structures from both conditions are close to each other. A high bottleneck distance suggests only that there is at least one structure which is significantly different between the two conditions, and therefore can point to more local differences. Therefore chromosomes with high bottleneck distance and low Wasserstein distance must contain few loop structures that differ considerably between the two homologies, suggesting some local changes rather than global changes between differentiation stages. Across all three lineages, chromosomes 1, 2, 10, 14, and 22 fit this profile, suggesting that while there are no major global changes to the topological structures of these chromosomes through differentiation as suggested by their low Wasserstein distance, there are some significant local differences between the Hi-C matrices suggested by the high bottleneck distance. Chromosome 21 stands out as having a consistently high Wasserstein distance across our three lineages, suggesting the topology of this chromosome may change more globally through differentiation.

By looking at the distances between H1 ESC cells and four of their possible progeny, we note that the biggest structural changes appear to occur early in differentiation of these lineages. Across all chromosomes, the distances between H1 ESC cells and neural progentitors are the smallest of the four, while there appear to be the most topological changes between ESC and mesendoderm, and ESC and mesenchymal cells. ME and MS represent earlier stages of differentiation, suggesting that the biggest changes in these lineages occur early on in the differentiation process. The one lineage we have beginning with PSC cells rather than ESC shows a different trend, where the PSC to MES transition has the lowest distance between topological structures rather than the highest (Figure 5d,f).

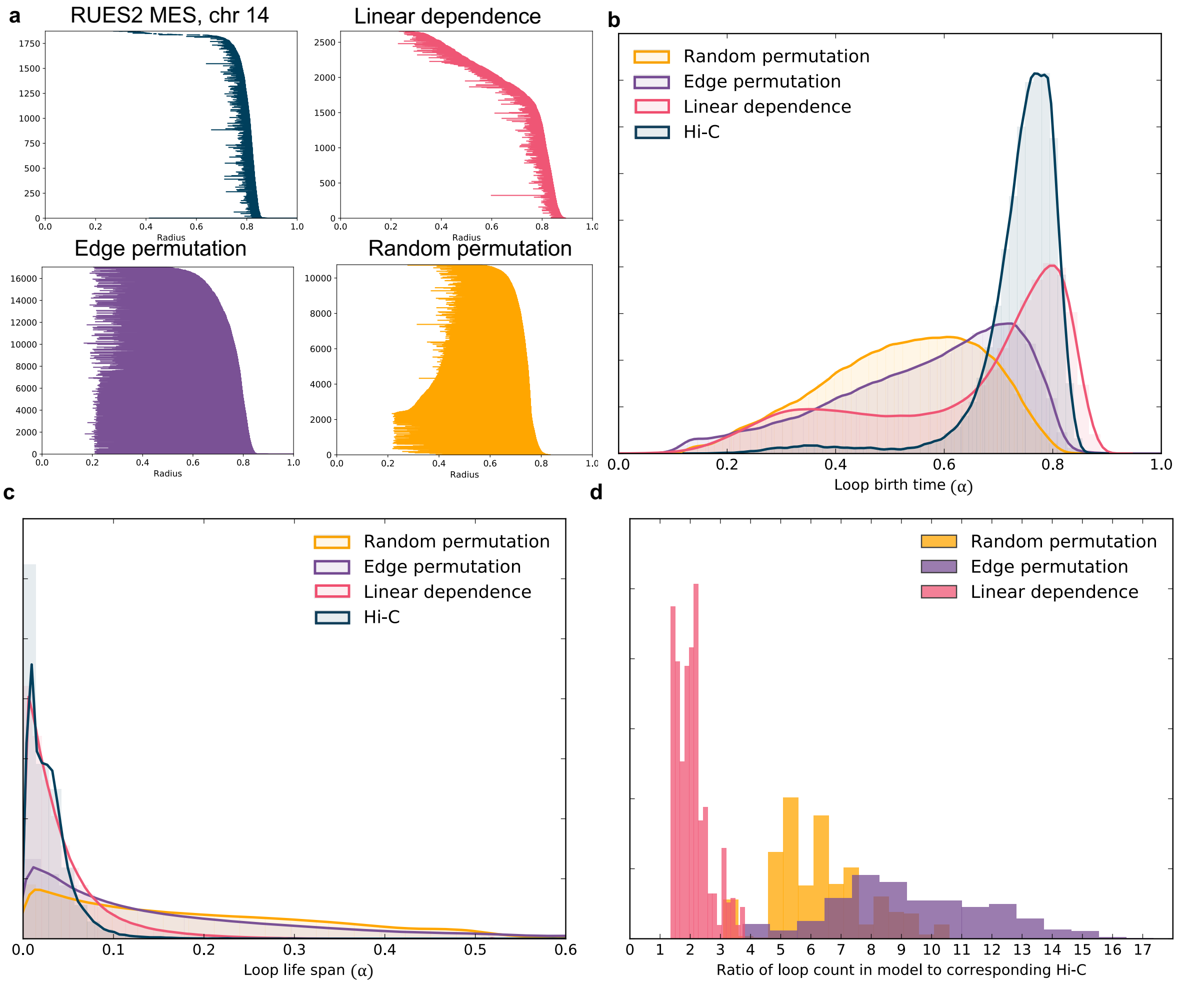

Figure 3: Loop analysis through multiple null models. **(a)** Barcode plots (showing only loop structures) from chromosome 14 of RUES2 MES cells, along with the corresponding barcode plots from the three null models. **(b)** Normalized histogram of the birth times of all loops in RUES2 data from true Hi-C and each null model. Hi-C clearly shows the tightest distribution, with almost no loop birth times before 0.6. **(c)** Normalized histogram of loop life spans (length of a barcode line) shows that the loops from real Hi-C data are very small, similar to the linear dependence model. **(d)** Normalized histogram of the ratios of numbers of loops in each null model to the number of loops in the corresponding Hi-C matrix, showing that all null models result in significantly more loop structures overall than those found in real Hi-C data.

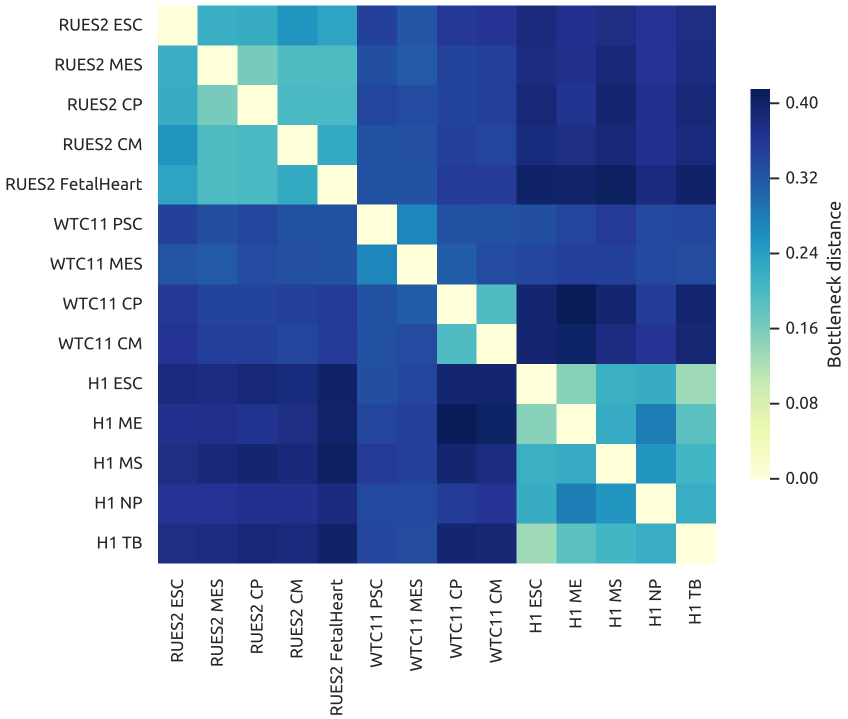

Figure 4: Comparison of all 14 samples studied. This heatmap represents the bottleneck distances between each pair of samples in our data, averaged over all 22 chromosomes. With the exceptions of the high distances between the two later stages and two earlier stages of WTC11 differentiation, the pattern of low distance within one cell line dominates these comparisons. This pattern can be seen by the lighter blocks along the diagonal with limits corresponding to the changes in cell line.

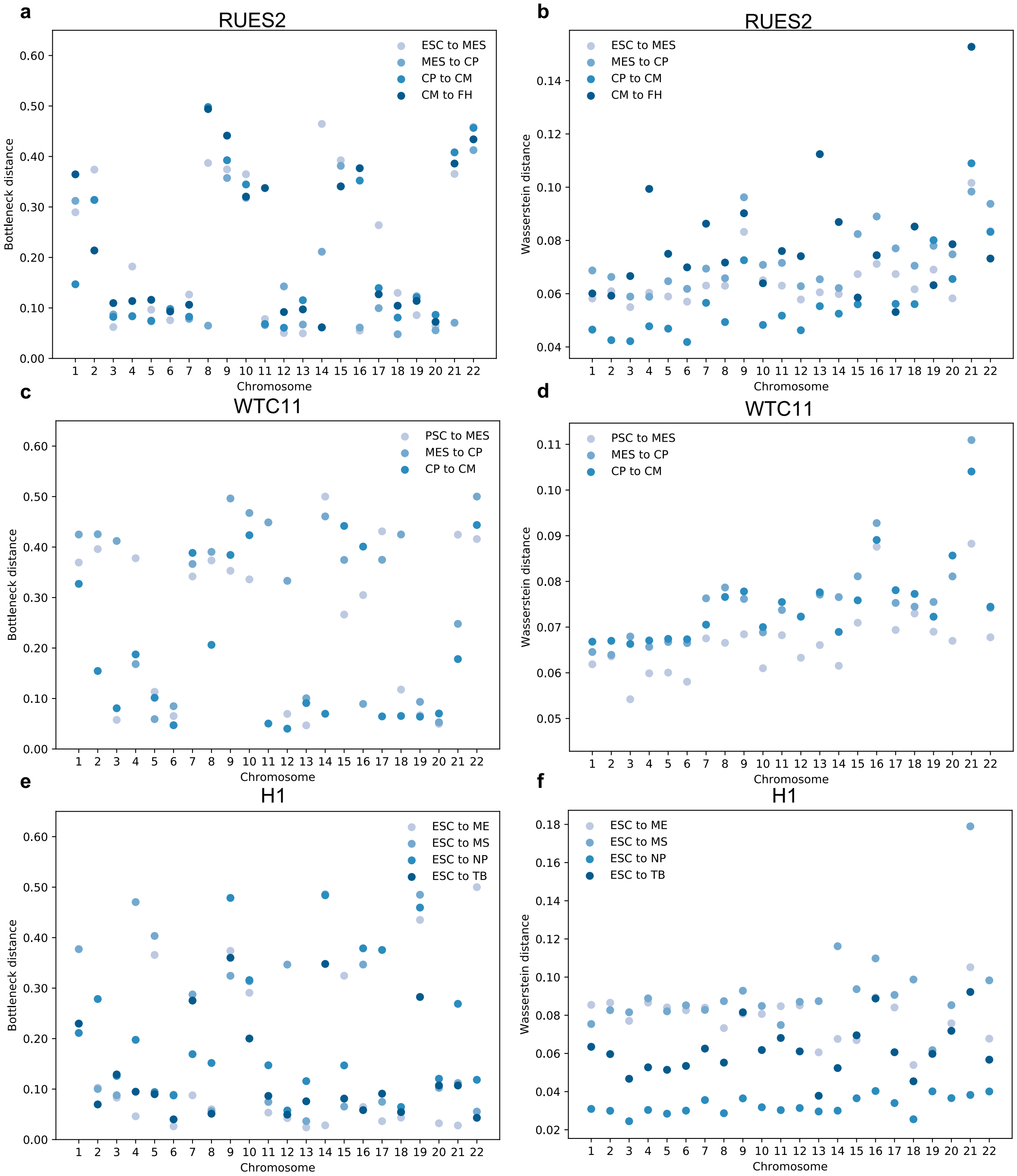

Figure 5: Distances between each consecutive stage of differentiation on all autosomal chromosomes. Each row represents one cell line, the left column shows bottleneck distances, and the right column gives Wasserstein distances. Each point represents the distance between the loop topologies of two consecutive stages of differentiation.

4 Discussion

One of the major challenges in the application of TDA to Hi-C data is the computational complexity of the methods combined with the scale of Hi-C. Due to computational limitations, the structures studied here are fairly large-scale; the Hi-C data analyzed is at a relatively low 100kb resolution, and our study only includes 14 samples. By studying smaller sections of the genome, perhaps near a gene of interest or within a particular structure of interest, higher resolution Hi-C or 5C could be used with TDA to identify small-scale topological structures. We were also only able to study intra-chromosomal matrices, but the topology of inter-chromosomal interactions would likely yield interesting insights as well. Another consequence of the computational complexity of TDA is our focus on only 0- and 1-dimensional features. Emmett *et al.* (2015) speculate that the 2-dimensional voids may represent interesting biological features such as transcription factories. Future improvements in the data structures and algorithms underlying TDA would significantly improve our ability to study the topology of the full genome.

It would additionally be very interesting to trace the loop structures back to the genomic locations that define these features, though topologically the existence of each of these features is all that matters, rather than defining exactly where in the data set they can be found. Unfortunately, this localization problem is challenging because each H_1 structure is an equivalence class rather than a specific loop. The definition of the “optimal” loop defining this equivalence class is therefore ambiguous. It remains nontrivial, even NP-hard (Chen and Freedman, 2011) under certain assumptions, to identify which member of this equivalence class is somehow representative or optimal, though this would facilitate the interpretation of these structures in the biological context.

The nature of Hi-C , as an experiment based on cross-linking over a full population, does not permit a transformation from the Hi-C counts to a true distance metric. Our distance matrices therefore do not satisfy the triangle inequality, which may affect the TDA results in unpredictable ways. One possibility to improve this concern is to use any of the methods that estimate a 3D structure from Hi-C, or select only the Hi-C values that satisfy a metric definition Duggal *et al.* (2013), and infer distances which would be geometrically consistent. However, this induces another source of error, and it is unclear whether the results would be more reliable.

5 Conclusion

We have presented the first application of TDA to study the topology of all 22 human autosomal chromosomes. By studying clusters and loops of 14 samples from various cell lines and stages of differentiation, we identify generative principles of chromosome structure. Our models suggest that the linearity of the chromosome is sufficient to explain the short lifespan of its loops, but additional structures in Hi-C likely lead to the small number of loops and their late birth times. We also show that topological structure is largely determined by cell line rather than stage of differentiation, and that there are few chromosome-wide changes through differentiation. We do, however, find evidence for local structural changes on several chromosomes consistent across all three cell lines studied. TDA shows promise for further analysis of Hi-C data, especially as computational limitations are overcome permitting analysis of higher dimensional features at higher resolution.

Acknowledgements

The authors would like to thank Alessandro Bertero and William S. Noble for useful information about their data, and Guillaume Marçais for comments on the manuscript.

Funding

This work has been supported in part by the Gordon and Betty Moore Foundation’s Data-Driven Discovery Initiative through Grant GBMF4554 to C.K., 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 and by the Richard K. Mellon Presidential Fellowship in Life Sciences to N.S. This work was partially funded by The Shurl and Kay Curci Foundation.

Financial disclosure

C.K. is a co-founder of Ocean Genomics, Inc.

References

Arsuaga, J. *et al.* (2015). Identification of copy number aberrations in breast cancer subtypes using persistence topology. *Microarrays*, **4**(3), 339–369.

Ay, F. *et al.* (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Research*, **24**(6), 999–1011.

Bunnik, E. M. *et al.* (2018). Changes in genome organization of parasite-specific gene families during the plasmodium transmission stages. *Nature Communications*, **9**(1), 1910.

Cámara, P. G. (2017). Topological methods for genomics: present and future directions. *Current Opinion in Systems Biology*, **1**, 95–101.

Camara, P. G. *et al.* (2016). Topological data analysis generates high-resolution, genome-wide maps of human recombination. *Cell Systems*, **3**(1), 83–94.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, **46**(2), 255–308.

Carlsson, G. (2014). Topological pattern recognition for point cloud data. *Acta Numerica*, **23**, 289–368.

Carriere, M. and Rabadan, R. (2018). Topological data analysis of single-cell Hi-C contact maps. *arXiv:1812.01360*.

Cavalli, G. and Misteli, T. (2013). Functional implications of genome topology. *Nature Structural and Molecular Biology*, **20**(3), 290–299.

Chan, J. M. *et al.* (2013). Topology of viral evolution. *Proceedings of the National Academy of Sciences*, pages 18566–18571.

Chazal, F. and Michel, B. (2017). An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv:1710.04019*.

Chen, C. and Freedman, D. (2011). Hardness results for homology localization. *Discrete & Computational Geometry*, **45**(3), 425–448.

Chen, Y. *et al.* (2018). Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *The Journal of Cell Biology*, **217**(11), 4025–4048.

Crane, E. *et al.* (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, **523**(7559), 240–244.

Cremer, T. and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, **2**(4), 292–301.

Dekker, J. *et al.* (2002). Capturing chromosome conformation. *Science*, **295**(5558), 1306–1311.

Dixon, J. R. *et al.* (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398), 376–380.

Dixon, J. R. *et al.* (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**(7539), 331–336.

Djekidel, M. N. *et al.* (2018). FIND: difFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Research*, **28**(1), 412–422.

Duggal, G. *et al.* (2013). Resolving spatial inconsistencies in chromosome conformation measurements. *Algorithms for Molecular Biology*, **8**, 8.

Duggal, G. *et al.* (2014). Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Research*, **42**(1), 87–96.

Edelsbrunner, H. and Harer, J. (2010). *Computational topology: an introduction*. American Mathematical Society.

Emmett, K. *et al.* (2015). Multiscale topology of chromatin folding. *arXiv:1511.01426*.

Fields, P. A. *et al.* (2017). Dynamic reorganization of nuclear architecture during human cardiogenesis. *bioRxiv*, page 222877.

Fillipova, D. *et al.* (2014). Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, **9**, 14.

Fudenberg, G. *et al.* (2011). High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nature Biotechnology*, **29**(12), 1109–1113.

Hinks, T. S. *et al.* (2016). Multidimensional endotyping in patients with severe asthma reveals inflammatory heterogeneity in matrix metalloproteinases and chitinase 3–like protein 1. *Journal of Allergy and Clinical Immunology*, **138**(1), 61–75.

Hnisz, D. *et al.* (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, **351**(6280), 1454–1458.

Imakaev, M. *et al.* (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, **9**(10), 999–1003.

Kragesteen, B. K. *et al.* (2018). Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nature Genetics*, **50**(10), 1463–1473.

Le Dily, F. *et al.* (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & Development*, **28**(19), 2151–2162.

Lesne, A. *et al.* (2014). 3D genome reconstruction from chromosomal contacts. *Nature Methods*, **11**(11), 1141.

Li, L. *et al.* (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine*, **7**(311), 311ra174.

Lieberman-Aiden, E. *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950), 289–293.

Lun, A. T. and Smyth, G. K. (2015). diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, **16**(1), 258.

Lupiáñez, D. G. *et al.* (2016). Breaking TADs: how alterations of chromatin domains result in disease. *Trends in Genetics*, **32**(4), 225–237.

Maria, C. *et al.* (2014). The Gudhi library: Simplicial complexes and persistent homology. In *International Congress on Mathematical Software*, pages 167–174. Springer.

Meaburn, K. J. *et al.* (2009). Disease-specific gene repositioning in breast cancer. *The Journal of Cell Biology*, **187**(6), 801–812.

Misteli, T. (2010). Higher-order genome organization in human disease. *Cold Spring Harbor Perspectives in Biology*, **2**(8), a000794.

Moindrot, B. *et al.* (2012). 3D chromatin conformation correlates with replication timing and is conserved in resting cells. *Nucleic Acids Research*, **40**(19), 9470–9481.

Nicolau, M. *et al.* (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, pages 7265–7270.

Norton, H. K. *et al.* (2018). Detecting hierarchical genome folding with network modularity. *Nature Methods*, **15**(2), 119–122.

Paulsen, J. *et al.* (2017). Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome Biology*, **18**(1), 21.

Pope, B. D. *et al.* (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature*, **515**(7527), 402–405.

Ramani, V. *et al.* (2016). Mapping 3D genome architecture through in situ DNase Hi-C. *Nature Protocols*, **11**(11), 2104–2121.

Rao, S. S. P. *et al.* (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**(7), 1665–1880.

Rennie, S. *et al.* (2018). Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. *Nature Communications*, **9**(1), 487.

Rizvi, A. H. *et al.* (2017). Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*, **35**(6), 551.

Ryba, T. *et al.* (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Research*, **20**(6), 761–770.

Serra, F. *et al.* (2017). Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Computational Biology*, **13**(7), e1005665.

Servant, N. *et al.* (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, **16**(1), 259.

Spanier, E. H. (1966). Algebraic topology. 1966. *MacGraw-Hill, New York*.

Spielmann, M. *et al.* (2018). Structural variation in the 3D genome. *Nature Reviews Genetics*, **19**, 453–467.

Trieu, T. and Cheng, J. (2015). MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. *Bioinformatics*, **32**(9), 1286–1292.

Wasserman, L. (2018). Topological data analysis. *Annual Review of Statistics and Its Application*, **5**, 501–532.

Weinreb, C. and Raphael, B. J. (2015). Identification of hierarchical chromatin domains. *Bioinformatics*, **32**(11), 1601–1609.

Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, **43**(11), 1059–1065.

Yu, Y. W. *et al.* (2015). Entropy-scaling search of massive biological data. *Cell systems*, **1**(2), 130–140.