Published in final edited form as:

Proc IEEE Int Symp Biomed Imaging. 2021 April ; 2021: 788–791. doi:10.1109/isbi48211.2021.9433812.

STATISTICAL COMPARISONS OF CHROMOSOMAL SHAPE POPULATIONS

Carlos J. Soto[1,2], **Peiyao A. Zhao**[3], **Kyle N. Klein**[3], **David M. Gilbert**[3], **Anuj Srivastava**[2]

[1]Department of Statistics, Pennsylvania State University, State College, PA, USA

[2]Department of Statistics, Florida State University, Tallahassee, FL, USA

[3]Department of Biological Science, Florida State University, Tallahassee, FL, USA

Abstract

This paper develops statistical tools for testing differences in shapes of chromosomes resulting from certain gene knockouts (KO), specifically RIF1 gene KO (RKO) and the cohesin subunit RAD21 gene KO (CKO). It utilizes a *two-sample test* for comparing shapes of KO chromosomes with wild type (WT) at two levels: (1) *Coarse shape analysis*, where one compares shapes of full or large parts of chromosomes, and (2) *Fine shape analysis*, where chromosomes are first segmented into (TAD-based) pieces and then the corresponding pieces are compared across populations. The shape comparisons – coarse and fine – are based on an elastic shape metric for comparing shapes of 3D curves. The experiments show that the KO populations, RKO and CKO, have statistically significant differences from WT at both coarse and fine levels. Furthermore, this framework highlights local regions where these differences are most prominent.

IndexTerms—

shape comparisons; chromosomal shapes; knock-out; TAD; elastic shape analysis

1. INTRODUCTION

Understanding genetic makeup and its constituents from genome conformation capture data is an important problem in biological data science. One part of genome analysis is estimation and analysis of chromosome structures using different sensing techniques. Hi-C [1] has emerged as an important sensing modality for learning chromosome structures. and a number of methods have been developed to estimate chromosomal structures from the Hi-C contact data (see e.g. BACH [2], SIMBA3D [3]). Using these methods one can estimate chromosomal shapes for Hi-C data collected from different populations. These tools enable investigations of an important scientific question: *Does a genetic mutation natural or artificial – induce a significant change in chromosome morphology?* For instance, if one knocks out a gene in a laboratory setting, what is the corresponding effect on the

8. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of The Florida State University.

shape of the chromosome? Is this effect predominantly local spatially, *i.e*. only some parts of of the chromosome are affected, or is it global, *i.e*. the whole chromosome is affected? The development of tools for addressing such questions is the main focus of this paper.

Comparisons of populations using sampled data are naturally performed using two-sample tests. Hagwood et al. [4] derived several nonparametric two-sample tests that rely only on pairwise shape metric – quantification of differences in shapes of curves – without any need for statistical modeling. We utilize a similar two-sample test that compares chromosomal shapes between and across classes, to decide if any two classes are similar or different. The actual quantification of differences in shapes is performed using *elastic shape analysis* [5]. For comparing shapes of chromosomal populations, there are at least two possibilities. Firstly, one can compare them using full structures. We will call this *coarse or global shape analysis*. Secondly, one can split full chromosomes into smaller structural units and then compare shapes of corresponding units across populations. We will call this *fine or local shape analysis*. An important related challenge is to split chromosomes into small units consistently across populations and maintain correspondences, so that comparisons are meaningful. To ensure this correspondence, we will utilize the concept of topologically associated domains or TADs for splitting chromosomes into smaller units.

For the experiments presented in this paper, we use the *Structural Inference via Multiscale Bayesian Approach* [3], termed SIMBA3D, for estimating chromosome shapes from Hi-C contact data. We treat several structures resulting from the same Hi-C matrix as independent random samples from the population representing that conformation. In this setting, comparisons of structures from Hi-C matrices for WT and KOs implies a two-sample test between two underlying populations.

2. HI-C DATA AND CHROMOSOMAL STRUCTURE

In this section we provide some background material on the problem of estimating conformations from Hi-C data. The experimental data used in this paper comes in the form of contact matrices from Hi-C [6] data. Hi-C data is a form of chromosome conformation capture technique that measures interactions between parts of chromosomes. It generates contact maps that describe the probability of observing interactions between any two regions of the genome that, in turn, relate to the pairwise distance matrices between pairs of genomic loci. In Fig. 1 we show a Hi-C contact matrix as an image in which we see most of the interactions take place around the main diagonal, meaning physically closer regions have larger probabilities of interacting with each other.

The problem of inferring 3D structures from contact matrices thus requires solving an inverse problem – infer pairwise distances between regions from the given probability of interactions. To do this we use the SIMBA3D methodology [3]. Since this methodology uses a gradient-based approach, one obtains a set of conformations representing different outcomes corresponding to different local solutions, even for the same contact matrix. Different conformations resulting from random initializations of SIMBA3D are viewed as random samples from an underlying population associated with a single contact matrix.

Fig. 1 shows three examples of these conformations and in Fig. 4 we display additional examples.

3. TWO-SAMPLE SHAPE TEST

In this section we describe a statistical procedure for comparing two sets of curves in terms of their shapes. This is a nonparametric approach and is based entirely on the choice of a shape metric.

Shape Metric:

To develop statistical analysis of shapes, we need a metric (or a distance) that quantifies shape distances between any two curves. We will use *elastic shape metric* [5] for this purpose. It is summarized here briefly. Let $\mathscr{F}$ be the set of all parametrized curves, viewed as smooth maps from [0, 1] to $\mathbb{R}^3$. Since the shape of a curve is invariant to rigid motions, global scales, and re-parameterization, we seek a metric that is invariant to those transformations. Let $\mathbb{O}(3)$ be the set of all 3 × 3 orthogonal matrices and let $\Gamma = \{\gamma : [0, 1] \to [0, 1] | \gamma(0) = 0, \gamma(1) = 1, \gamma$ is a diffeomorphism$\}$ be the set of all re-parameterizations. For any curve $f \in \mathscr{F}$, its rotation by $O \in \mathbb{O}(3)$ and re-parametrization by $\gamma \in \Gamma$ are given by $f \mapsto O(f \circ \gamma)$. In order to compare shapes of curves, this method uses an elastic metric that is computed using the notion of *square-root velocity function*, or SRVF, of curves. The SRVF q of a parameterized curve f is defined as, $q(t) = \dot{f}(t)/\sqrt{|\dot{f}(t)|}$ if $|\dot{f}(t)| > 0$ and 0 if $|\dot{f}(t)| = 0$. Since the SRVF is defined in terms of derivatives, we get invariance to translation automatically. In order to be invariant to scales, we rescale the SRVF of a curve according to $q \to q/\|q\|$, where $\| \cdot \|$ denotes the $\mathbb{L}^2$ norm of a curve. The space of all unit norm SRVFs is an infinite-dimensional sphere $\mathbb{S}_\infty$. Thus, the elastic shape distance between any two curves f_1 and f_2, represented by their scaled SRVFs q_1 and q_2, is defined as

$$d_s(f_1, f_2) = \min_{O \in \mathbb{O}(3), \gamma \in \Gamma} \left(\cos^{-1}\langle q_1, O(q_2, \gamma)\rangle\right)$$

We use Procrustes methodology to find the optimal rotation O^* and Dynamic Programming to find optimal parameterization γ^* as described in [5].

Statistical Comparison of Shape Populations:

Now we have a way of comparing the shapes of curves in a pairwise fashion using d_s. However our goal is to compare populations of shapes, using shapes sampled from each population. The specific question we pose is: *Do the two sets of shapes, representing two samples, come from the same underlying population or not?* The binary hypotheses of interest are $H_0 : F_{(1)} = F_{(2)}$ versus $H_1 : F_{(1)} \neq F_{(2)}$ where $F_{(1)}$ and $F_{(2)}$ are any two population distributions (such as WT, RKO, or CKO). There are several two sample-tests for testing equality of distributions, but many of these assume that data points lie in a Euclidean space.

For testing this hypothesis, we will utilize an *energy test* used in Hagwood et al. [4]. It is as follows. Suppose we have a pairwise distance matrix D between two sets of sizes $n_{(1)}$ and $n_{(2)}$ with composition as in Fig. 2. The energy of D, or the test statistic, is defined to be

$$S = \frac{n_{(1)} \cdot n_{(2)}}{n_{(1)} + n_{(2)}}\left[2 \cdot \bar{D}_{1,2} - \left(\bar{D}_{1,1} + \bar{D}_{2,2}\right)\right].$$

The $\bar{D}$ denotes the average distances for different combination: $\bar{D}_{I,J} = \frac{1}{n_{(I)} \cdot n_{(J)}} \sum_i \sum_j D_{I,J}(i,j)$. If the two classes are indeed different, then the distances between classes are larger, on average, than the distances within classes and the test statistic S will be large. On the other hand, if the classes are identical, the test statistic S will be closer to zero. In order to compute a p value for this test statistic we use a permutation test. This is based on a bootstrap method detailed in Algorithm 1.

Algorithm 1

Energy permutation test

```
Compute the full distance matrix $D$.
repeat
    Randomly generate an $N \times N$ permutation matrix $P$.
    Let $\tilde{D} := P^T D P$ be the permuted distance matrix.
    Compute the test statistic $\tilde{S}$ under $\tilde{D}$.
until Done $x$ number of times
The approximate p-value = $\frac{\#\{\tilde{S} < S\}}{x}$.
```

4. COARSE AND FINE METHODOLOGY

Next we focus on how to use this framework in comparing chromosome populations at coarse and fine levels. The coarse scale analysis is applied first, to test if there are any structural differences between any two populations at the full chromosome level. If that test detects any significant differences, then we apply fine scale analysis to find where do these differences lie. Together these two procedures help us determine the similarities and differences in parts of chromosomes across populations. For each chromosome dataset we consider all three classes. Our goal is to compare each of the knockouts with the wild type for each chromosome.

Coarse Scale Analysis Methodology:

For a pair of contact matrices, representing two classes, let $f_i^{(1)}$ represent the ith curve estimated from one class, and $f_j^{(2)}$ be the jth curve estimated from the other class. Let $n_{(1)}$ and $n_{(2)}$ denote the respective number of curves and let $N = n_{(1)} + n_{(2)}$. We compute a pairwise elastic shape distance matrix $D \in \mathbb{R}^{N \times N}$ with blocks $D_{I,J}$, $I, J \in \{1, 2\}$ as shown in Fig. 2. The entries of the submatrix $D_{I,J} = \left(D_{i,j}^{(I),(J)}\right)$ are $D_{i,j}^{(I),(J)} = d_s\left(f_i^{(I)}(t), f_j^{(J)}(t)\right)$. We then apply the energy test. With this test we can determine if the two sets of chromosomal shapes belong to the same underlying population or not. However, this test does not help us determine the subregions where the differences lie. For that we utilize the fine scale comparisons as described next.

Fine Scale Analysis Methodology

Here we take contact matrices and partition them diagonally into submatrices using the notion of Topological associated domains (TADs). There are several methods to estimate the TAD structure and we use an insulation score method [7] in this paper. In Fig. 3 we display a contact matrix (leftmost) with a red overlay of TAD segmentations and a portion boxed off in magenta, a blowup of the magenta portion (middle), and the four segments that correspond to the first four TADs from the blowup (right).

Let $f_{i,k}^{(l)}$ be the kth segment from of the ith conformation from the l^{th} class, e.g. $f_{8,24}^{(1)}$ is the 24th segment of the 8th curve from the first class of curves. For each segment location k, we create a large pairwise distance matrix $D(k) \in \mathbb{R}^{N \times N}$ and perform the energy test as earlier. The partition of this matrix into blocks is the same as in Fig. 2. The only difference is that we have a separate matrix and separate test for each location k. For each location one can decide if the corresponding populations differ in shape or not. Furthermore, the corresponding test statistic value indicates how different are the two populations being compared.

5. EXPERIMENTAL RESULTS

We study a total of 10 datasets, each observed under three conditions – WT, RKO, and CKO. For each of these 30 contact matrices, we generate 60 conformations using SIMBA3D. In Fig. 1 we display some conformations examples. The color is simply a gradient representation of the start (blue) and end (yellow) of the curve. As in [3], we initialize each conformation with a random multivariate Gaussian, so there is some variability in the curves even within the same class. Even though we initialize each curve independently we notice there are structural similarities which are most evident in the locations of the colors across classes.

Coarse Analysis:

Fig. 4 shows the process for the first dataset of interest. Leftmost we show two contact matrices for this dataset, the top being RKO and the bottom being WT. In the middle we show some conformations estimated from these matrices, and lastly, rightmost we display the pairwise distance matrix comparing these two classes. We perform the energy test with 10000 replications and yield a test statistic $S = 1.5107$ with p-value= 0. Thus, we can conclude that the RKO and WT populations are structurally significantly different for this dataset.

We repeat this process for each dataset comparing WT to RKO and WT to CKO and present the results in Table 1. In every case, when comparing WT to KO, we have a large test statistic and a small p-value. Thus, we can conclude WT conformations have significantly different shapes than both RKO and CKO. Next we seek subregions where the differences are most significant.

Fine shape analysis:

Earlier we concluded that the WT curves are statistically significantly different than both the RKO and CKO curves, next we aim to locate the differences. For the first dataset, we have 46 TAD segments per curve, so we calculate 46 – 120 × 120 pairwise distance matrices, and perform the energy test on each matrix (i.e. for each segment). In Fig. 5 we display these 46 test statistic values as a heatmap, along with their corresponding color bars. For some segments, the shape differences across populations are very large, but for most of the segments the differences are not significant. We also study these differing segments visually and confirm structural differences.

The results are similar for other datasets. It is interesting to note that the structural differences in chromosomes, resulting from knocking out of genes, are very localized and most of the chromosomes remain unchanged.

6. CONCLUSIONS

This paper develops a framework for comparing shapes of chromosomal populations that are created by gene knockouts. These comparisons are performed at both global (large portion of chromosomes) and local (piece by piece) level. The testing is based on computing a certain test statistic involving elastic shape metric between curves. The global tests show that the chromosome shapes are altered by gene knockouts. The local tests show, however, that these differences are restricted to only a small number of segments. These results pave way for understanding biological significances of localized changes induced in chromosome structures by mutations and gene knockouts.

ACKNOWLEDGEMENTS

This research was supported in part by the grant NIH R01 GM126558 and NSF CDS&E DMS 1953087. There were no conflicts of interest in this research.

9. REFERENCES

[1]. Oluwadare O, Highsmith M, and Cheng J, “An overview of methods for reconstructing 3-D chromosome and genome structures from HiC data,” Biological procedures online, vol. 21, no. 1, pp. 7, 2019. [PubMed: 31049033]

[2]. Hu M et al. , “Bayesian inference of spatial organizations of chromosomes,” PLoS computational biology, vol. 9, no. 1, pp. e1002893, 2013. [PubMed: 23382666]

[3]. Rosenthal M et al. , “Bayesian estimation of three-dimensional chromosomal structure from single-cell hi-c data,” Journal of Computational Biology, 2019.

[4]. Hagwood C et al. , “Testing equality of cell populations based on shape and geodesic distances,” IEEE Transactions on Medical Imaging, vol. 32, no. 12, 2013.

[5]. Srivastava Anuj and Klassen Eric P, Functional and shape data analysis, Springer, 2016.

[6]. Van Berkum NL et al. , “Hi-C: a method to study the three-dimensional architecture of genomes.,” JoVE (Journal of Visualized Experiments), , no. 39, pp. e1869, 2010.

[7]. Schwarzer W et al. , “Two independent modes of chromatin organization revealed by cohesin removal,” Nature, vol. 551, no. 7678, pp. 51, 2017. [PubMed: 29094699]

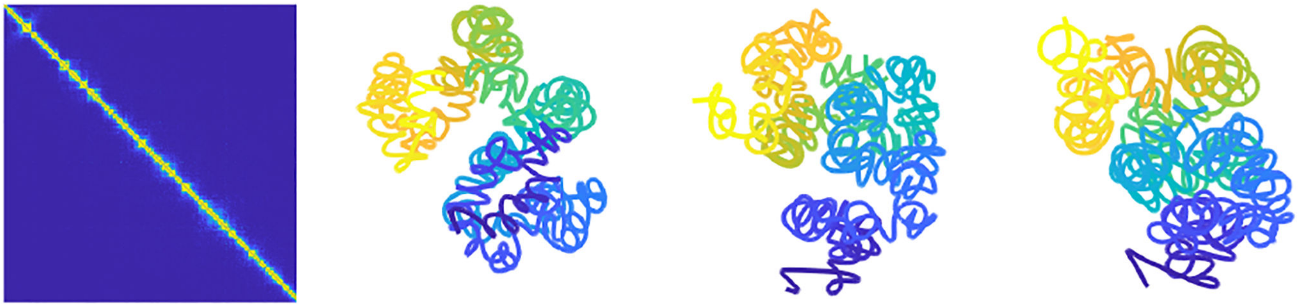

Fig. 1.

Left to right: A contact matrix viewed as an image and conformations generated using SIMBA3D for WT class, RKO, and CKO classes, respectively.

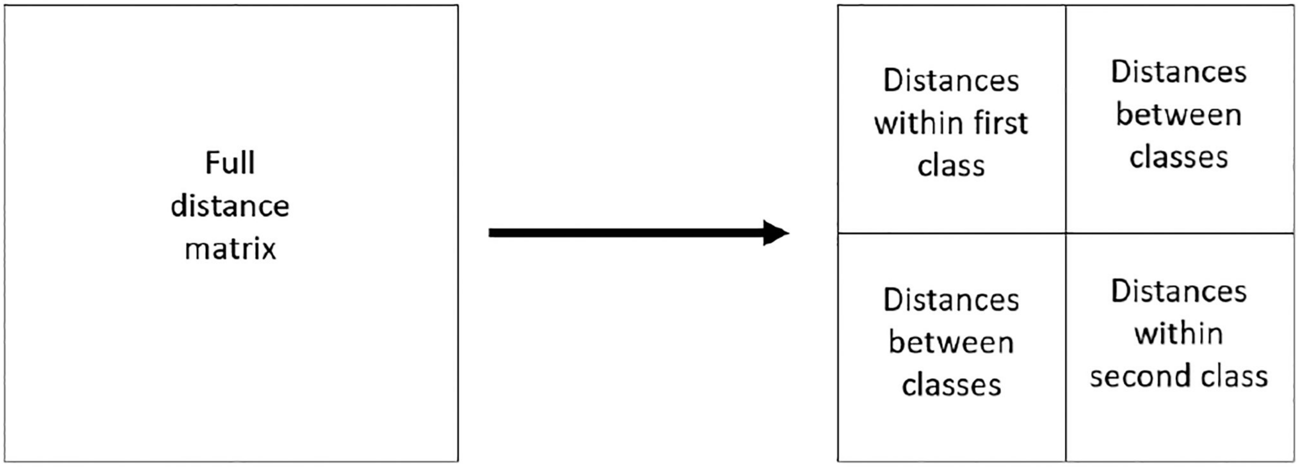
Full distance matrix
Distances within first class
Distances between classes
Distances between classes
Distances within second class

Fig. 2.

The pairwise distance matrix. Left: the full distance matrix; Right: block structure for two classes.

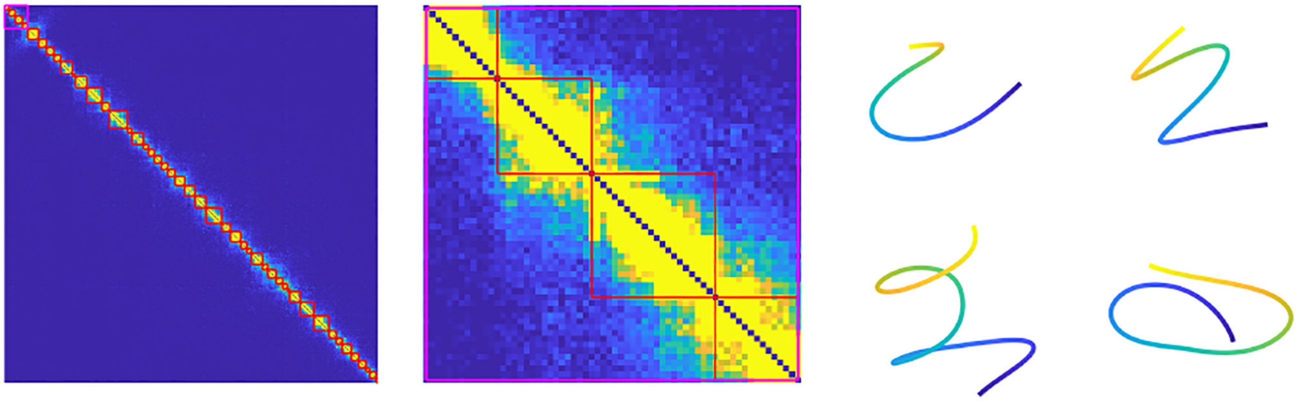

Fig. 3.

Left: A contact matrix with TAD overlay in red and a portion in magenta which is the middle image. Right: The first four segments of the curve which correspond with the TADs from the middle image.

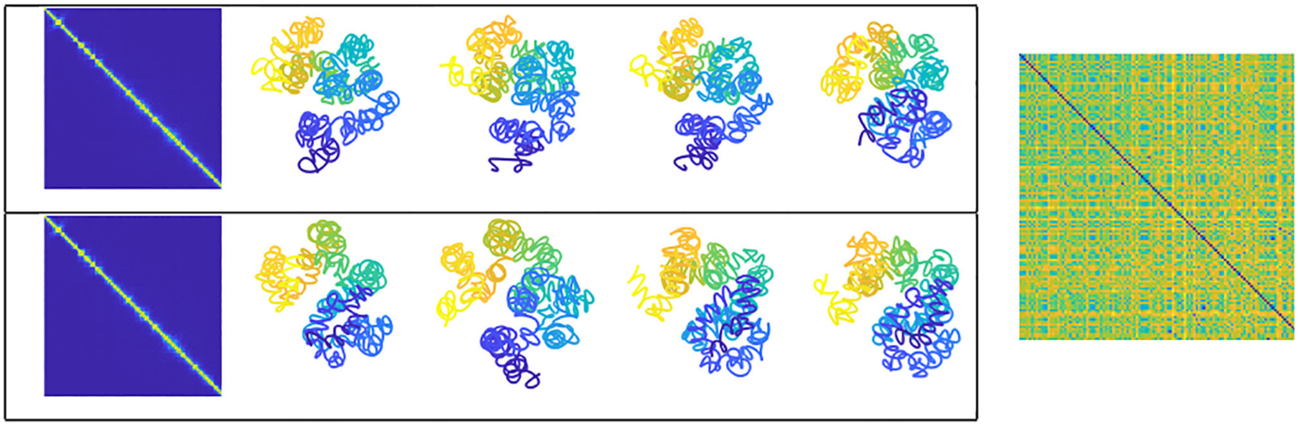

Fig. 4.

Left: RKO and WT contact matrices. Middle: Four random curves generated using SIMBA3D from each case. Right: The full pairwise distance matrix *D*.

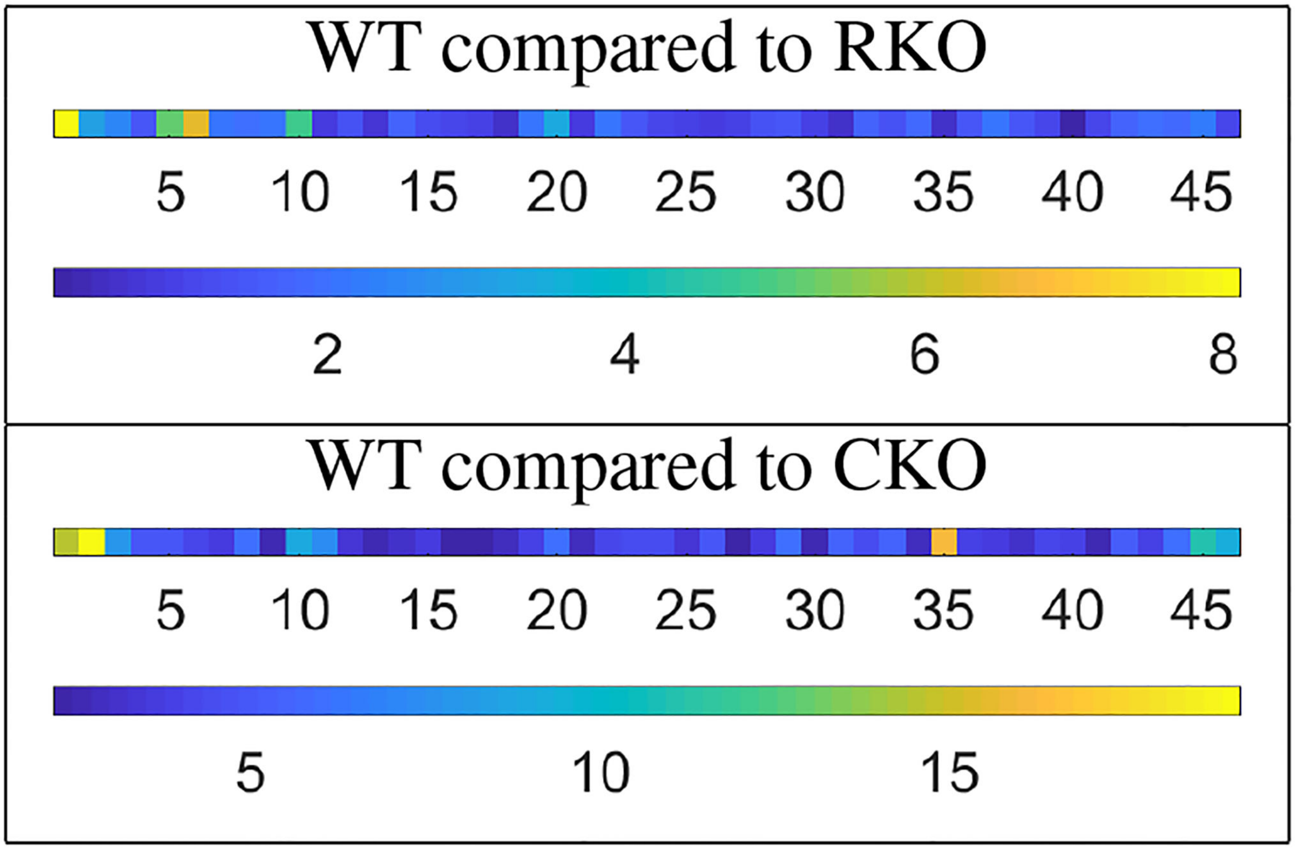

Fig. 5.

The local test statistics *S* for 46 TAD segments. The three largest *S* are 1, 6, and 5 for the top and 1, 35, and 2 for the bottom. The three smallest *S* are 40, 18, and 31 for the top and 17, 27, and 16 for the bottom.

Table 1.

Top: Description of each dataset. Middle & Bottom: Coarse analysis Energy test results with 10000 bootstrap replications when comparing two classes.

Dataset	1	2	3	4	5
Chromosome	1	1	1	2	2
Start point	916	1396	3479	2	1445
End point	1919	2399	4482	991	2433
WT compared to RKO					
Test Statistic	1.5107	1.3992	1.3695	1.9682	2.4141
p-value	0	0	0	0	0
WT compared to CKO					
Test Statistic	2.3633	2.4305	2.3992	3.3762	3.6824
p-value	0	0	0	0	0