BPscore: an effective metric for meaningful comparisons of structural chromosome segmentations

Rafał Zaborowski[1*], Bartek Wilczyński[1*]

1 Faculty of Mathematics, Informatics and mechanics, University of Warsaw, Warsaw, Poland

* bartek@mimuw.edu.pl, r.zaborowski@mimuw.edu.pl

Abstract

Studying the 3D structure of chromosomes is an emerging field flourishing in recent years because of rapid development of experimental approaches for studying chromosomal contacts. This has led to numerous studies providing results of segmentation of chromosome sequences of different species into so called Topologically Associating Domains (TADs). As the number of such studies grows steadily and many of them make claims about the perceived differences between TAD structures observed in different conditions, there is a growing need for good measures of similarity (or dissimilarity) between such segmentations. We provide here a BP score, which is a relatively simple distance metric based on the bipartite matching between two segmentations. In this paper, we provide the rationale behind choosing specifically this function and show its results on several different datasets, both simulated and experimental. We show that not only the BP score is a proper metric satisfying the triangle inequality, but that it is providing good granularity of scores for typical situations occuring between different TAD segmentations. We also introduce local variant of the BP metric and show that in actual comparisons between experimental datasets, the local BP score is correlating with the observed changes in gene expression and genome methylation. In summary, we consider the BP score a good foundation for analysing the dynamics of chromosome structures. The methodology we present in this work could be used by many researchers in their ongoing analyses making it a popular and useful tool.

Author summary

Many researchers are interested in the chromosomal structure, its function and dynamics. Over the recent years, chromosome conformation capture (3C) methods have become the main source of experimental data on the subject and the Topologically Associating Domains (TADs) have become the *de-facto* standard unit of chromosomal structure. Many methods have been developed for TAD calling and the 3C experiments have been done in multiple conditions giving us a multitude of chromosomal segmentations describing the most atomic differences in chromosomal structure between conditions. Until now, such segmentations were compared mostly by very rough measures, such as the Jaccard coefficient or TAD overlaps or very general metrics like the variation of information coefficient. This has limited the researchers in the analysis of differential TAD segmentations, and practically prevented any proper analysis of TAD dynamics between conditions. Our approach has the potential to facilitate such analyses by providing researchers with mathematically sound metric that is designed specifically for the purpose and tested on both simulated and experimental data. Additionally, we provide a local variant of our measure that is a natural derivative of the BP score that can indicate which parts of the chromosomes are undergoing the most significant structural reorganizations.

Introduction

Scientific studies of the cell nucleus and its structure started more than a hundered years ago when embryologists were performing their experiments on the fertilized eggs and subsequent rapid cell divisions. In the last 50 years, after discovery of the DNA structrue and its role, the topic of chromatin structure was considered to be of slightly lower importance, as the researchers have been mostly preoccupied with studying the DNA sequence of the genomes to uncover its function. In recent years, genome conformation studies gained significant attention again, for at least two reasons. On the technical side, this is mainly due to development of Chromosome Conformation Capture, or 3C technique [1] and its derivative methods like Hi-C [2], 4C [3, 4] or 5C [5]. On the other hand, a perhaps more important reason why there is so much interest in the newly acquired data on chromosomal conformation is that current efforts in explaining the function of non-coding regulatory elements without the information on the contacts between enhancers and promoters are proving it to be an extremely difficult problem [19]. The problem of gene regulation is difficult even given that the positions of such elements can be identified by DNAse Seq or related techniques [20]. This leads to the situation where many researchers are probing chromosome structure using 3C related techniques and the research in the field is progressing rapidly. This, as always is the case in blooming scientific fields, leads to new technical challenges in describing the chromosomal structures and comparing them between cell-types and experimental conditions.

Going back to the technical side of the chromosome capture techniques, in the Hi-C assay DNA is cross-linked, shared into pairs of fragments and ligated. Ligated fragments are then pulled down and merged with adapters. Finally, fragments are paired-end sequenced to produce library of paired reads indicating pairs of genomic regions that are in close proximity in the three-dimensional space of the nucleus. Even though the method is not exact in the sense of mapping the 3D-distance to the probability of ligation followed by succesful identification of a pair, it has been proven to be very reproducible provided that we are interested in the average contact frequency over a large pool of cells statistically sampled by deep sequencing [21]. Single cell hi-c techniques are much more recently developed, but there is significant progress there as

well and we should expect soon to be able to compare chromosomal contact structures also between individual cells [22] [23]. Also, other methods, that are not ligation-dependent have been developed for probing chromosomal contacts, most notably the genome architecture mapping (GAM) results corroborate the hi-c derived chromosomal structures [24].

Once the sequencing is completed and contacting pairs of segments are identified, usually the analysis of interactions is based on construction of contact maps: symmetric matrices that summarize contacts between bins of a pre-defined size. If we are interested in cis-chromosomal contacts, which is usually the case, such large genome-wide matrix is usually broken down into separate matrices for each chromosome. In these matrices, each cell contains the number of interactions between 2 particular regions of the same chromosome. Such matrices usually need to be normalized, to account for several technical issues, such as GC content bias and diverse mappability across genomic regions that lead to non-uniform read coverage along chromosomes. There are methods that are specialized to a given hi-c protocol [25] as well as more generic normalization methods based on the assumption of uniform coverage [16].

In recent years, many studies have suggested that in addition to the well established chromosomal compartments representing the division of chromatin into active euchromatin and inactive heterochromatin, there exists a finer structure of Topologically Associating domains or TADs for short [6] [26]. The presence of TAD-like structures was confirmed in almost all metazoan Hi-C data, with notable exclusion of some chromosomes in C. elegans [27], while there appear to be no sign of TADs in yeast and plants [28]. The majority of animals, including all studied mammals, show significant concentrations of contacts into TADs and these appear to be largely conserved between species and conditions [6]. Even though, mostly due to variablity in hi-c protocols and TAD-calling methods, it still remains a challenge to assess physical properties and biological function of TADs in individual cells, it was shown beyond doubt that regions of genome demarcated by TADs are indeed correlated with gene regulation [7]. In particular it was shown that a deletion of TAD boundary regions may lead to different developmental disorders [8] [9].

Numerous studies exploiting Hi-C to compare genome conformation between different cell types or conditions were conducted [10–12, 15]. The purpose of such analyses is to capture differences in spatial organization of chromatin and its relationships with gene expression or epigenetic modifications. One approach is to spot and quantify regions of chromosome where TAD arrangement is (dis)similar between two experiments. A common method for this purpose is to count number or overlapping TAD boundaries that was used among others by [6] [13] [14]. However this approach suffers one serious disadvantage that large domains with relatively small shift in their boundaries contribute to lowering the overall dissimilarity as much as small domains with relatively large shifts. Another approach, that is not as boundary-centric is by measuring variation of information between different TAD-segmentations as used in [15]. This approach, however suffers from the fact that it does not take into account the linear structure of the genome and usually overestimates the deviation from the randomized control and therefore overestimates the TAD structure conservation. Given the rapid advancement of the experimental research in the field and constantly growing flow of new data with TAD structures, it is important to have a proper metric to compare different experimental results as well as different computational methods for TAD detection. In perfect scenario, such dissimilarity measure should be a proper metric satisfying triangle inequality.

In this work we present new distance measure for comparing TAD sets and prove that our measure fulfill metric properties. We use simulated and real data to show that our metric is able to capture interesting properties of different domain sets. We also

perform comparison with simple boundaries overlap count based approach concluding that our metric exhibit smoother behaviour allowing for more precise analysis. Finally we present how our metric can be used to estimate (dis)similarity of subchromosomal regions enabling for more in depth analysis.

Results

BP-distance

In this chapter we give the definition of BP distance and explain notation. Precise definitions can be found in chapter Materials and methods.

Let A and B define 2 intervals of equal length. Each of them is partitioned on subintervals $A[i]$ and $B[j]$ (see fig. 1). A pair of subintervals $A[i]$, $B[j]$ may induce segment $o[k]$, which is non empty intersection between those subintervals. BP distance between domain sets A and B is defined as:

$$d(A,B) = 1 - \frac{1}{N}\sum_{i}^{|o|}\frac{l(o[i])^2}{\max\left(l\left(f_{A,B}(i)\right), l\left(f_{B,A}(i)\right)\right)} \tag{1}$$

where $f_{A,B}(i)$ is a function mapping $o[i]$ to subinterval in A, which induced $o[i]$ and $l(o[i])$ is function mapping $o[i]$ to its length.

Fact 1. *The above defined function is a metric.*

The proof can be found in chapter Materials and methods. Below we present some properties of function $d(A,B)$:

1. every change that is introduced has proportional impact to the length of the interval affected,
2. inclusions, deletions and shifts are treated equally,
3. we can compare domain sets of different sizes.

The above properties help avoid problems, which arise in alternative TADs similarity measures, where one assess similarity based on the number of TAD boundary overlaps across chromosome (see next section).

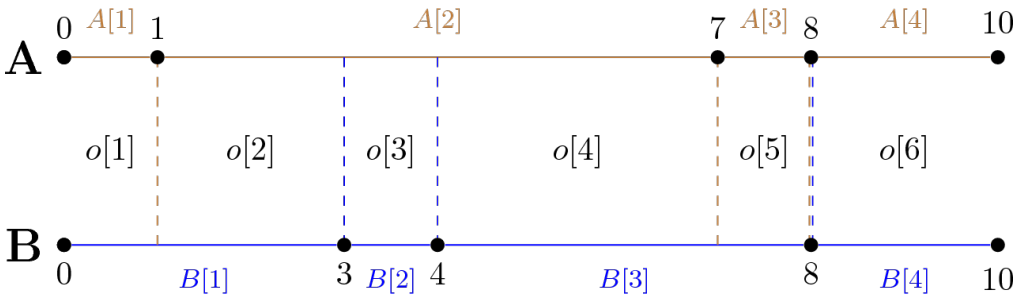

Fig 1. Chromosome partitioning and the induced segmentation.

Comparison with Existing Methods

To demonstrate advantage of our metric over boundaries overlap count approach we simulated artificial sets of TADs using following procedure. First we simulate 2 sets, each consisting of 600 TADs, the approximate number of TADs on chromosome 1 determined using DP approach [15] on human ESC Hi-C data [6]:

1. The TADs lengths were drawn to reflect real TAD length distribution (Negative Binomial distribution with parameters adjusted manually),
2. TAD length was fixed to 40 bins for every TAD.

Then for each of the above 2 sets we generated match sets according to 3 scenarios:

- each domain boundary is moved at most ε bins left or right of its initial position (fig. 2a),
- new boundary is introduced at most ε bins to the left of each TAD boundary, except of first one (fig. 2b),
- $2^n - 1$ boundaries are introduced into every TAD according to binary interval partitioning scheme (fig. 2c).

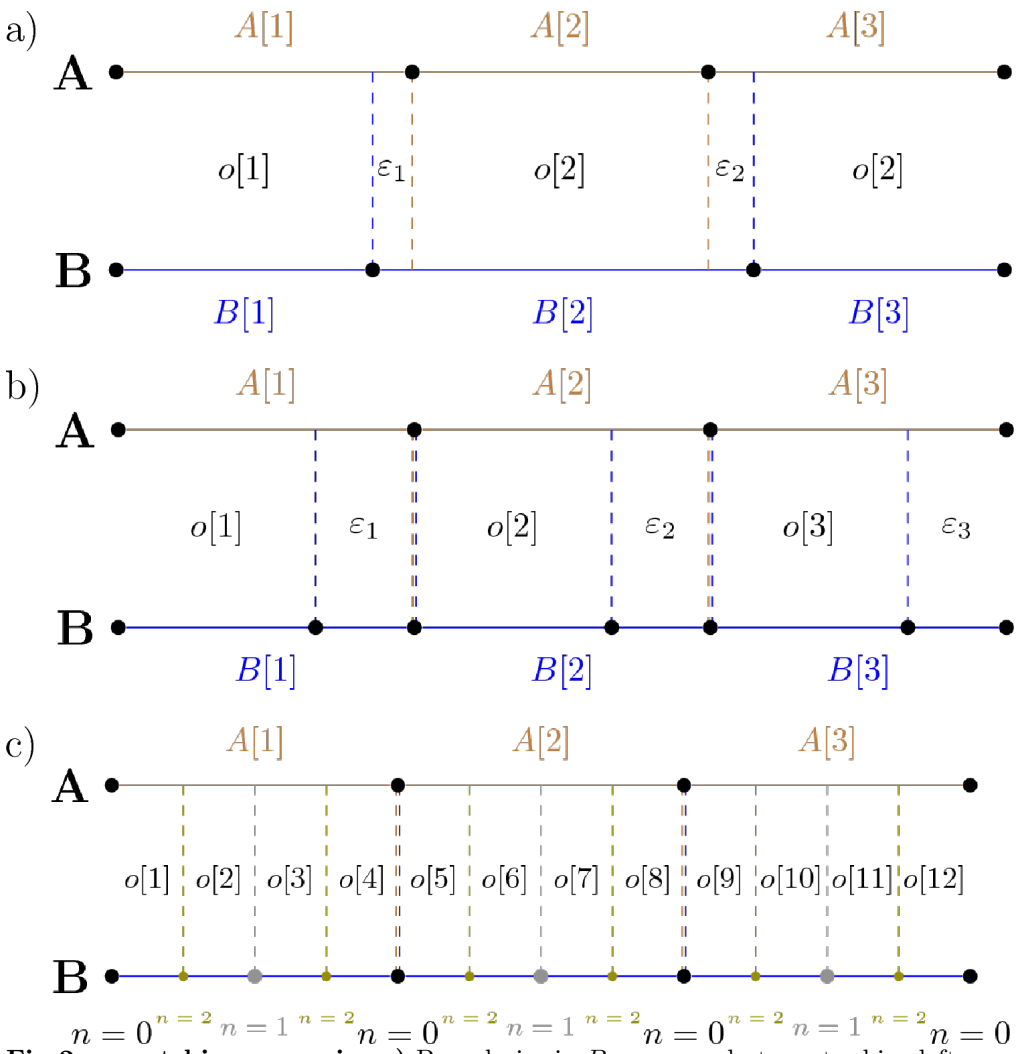

Fig 2. ε-matching scenarios a) Boundaries in B are moved at most ε bins left or right. **b)** New boundary is inserted into B at most ε bins to the left of existing boundary. **c)** Domains in B are partitioned according to binary interval partitioning scheme.

Moreover each of the above match sets is replicated 3 times with increasing parameter ε or n. Scenario 1 can represent situation where structural organization of 2 chromosomes is almost identical however due to noise factors coordinates of detected TADs do not overlap. Scenario 2 can be considered either a noisy match (just like 1) or large mismatch. The former case may occur when a little interval like non-TAD region (often produced in Directionality Index method [6] or DP approach [15]) is located next to large domain and both of them overlap with some large domain in corresponding TAD set. The latter one is possible when 2 consecutive TADs of similar length overlap with 1 large TAD, a situation indicating poor match. Scenario 3 can be considered mismatch except for $n = 0$. In particular we can imagine complete mismatch if a TAD in experiment A have length N and in experiment B there are N TADs, each of length 1.

Additionally we compare our metric with Variation of Information metric adopted to assess similarity of segmentation as described in [15].

For each scenario and possible parameter value we report 3 numbers (fig. 3): our distance (refered to as BP), Jaccard Index distance (refered to as JI) and Variation of Information distance (refered to as VI). The JI distance is calculated by dividing the number of TAD boundary overlaps between A and B by the number of all boundaries (overlaps are counted once) and subtracting resulting fraction from one. Calculation of VI distance is described in Materials and methods.

As can be seen in situations where perfect TADs overlap is disturbed by subtle noise the BPscore distance, similarly to the VI distance, outperforms boundary overlap count (Jaccard Index) approach in recovering true similarity.

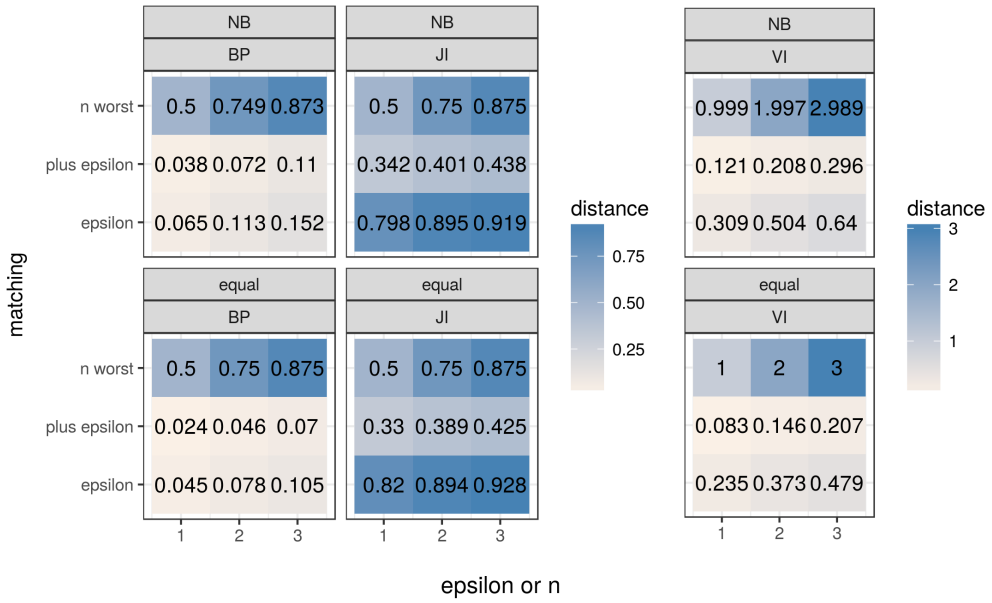

Fig 3. Comparison with Jaccard Index and Variation of Information distance.

Differentiating Technical From Biological Replicates Data

To demonstrate usefulness of our metric we present a case where one might be interested in assessing structural similarities genomewide for multiple samples. We selected publicly available sets of Hi-C data consisting of 6 cell lines: ESC, MSC, MDC, NPC, TDC and IMR90 [6] [11]. ESC was available in 4 technical replicates and

remaining lines in 2 technical replicates. We processed each dataset using Armatus software [15] to produce sets of TADs. Finally we made pairwise comparisons for every pair of TAD sets. Figure 4 illustrates comparisons divided on two categories: within replicates and between cells for BP metric (left), JI metric (center) and VI metric (right). As can be seen all metrics discriminate between two groups under consideration,

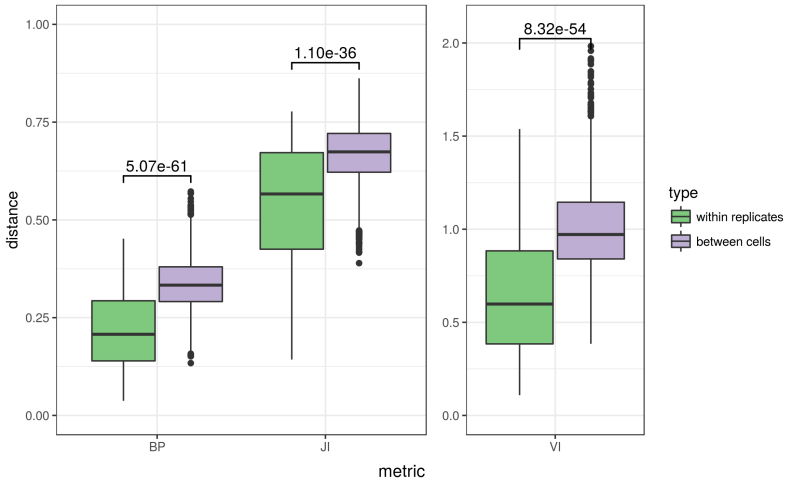

Fig 4. Pairwise comparisons of real Hi-C datasets with different metrics.

i.e. the scores of between-replicates comparisons are consistently lower than between-conditions comparisons. While all metrics give significant discrimation, we note that the significance of the difference, as well as the separation between the two groups is better for the VI and BP scores than the Jaccard index. Of the two metrics with better separation, BP score is giving slightly lower p-value, however this is unlikely to be a difference of practical importance. Similar behaviour is observed when comparing individual chromosomes (S1 Fig and S2 Fig).

Local BP score

Another potentially important application of comparing chromosomal segmentations is the search for locally re-organized fragments induced by 2 partitions for selected fragment of a chromosome. For example to check if differential gene expression or methylation pattern correlates with segmentation pattern. For this kind of analysis we suggest to use local BP score - a measure associated with our metric. Local BP score of segment i is defined as:

$$d^{\mathbf{BP}}_{A,B}(i) = 1 - \frac{l(o_{A,B}[i])}{\max\left(l\left(f_{A,B}(i)\right), l\left(f_{B,A}(i)\right)\right)} \tag{2}$$

Local BP score satisfies: $0 \leq d_{A,B}(i) < 1$. The closer it is to 0, the larger the overlap between $f_{A,B}(i)$ and $f_{B,A}(i)$. Conversely, the closer it is to 1, the higher the mismatch. Value of 0.5 may represent a boundary situation between match and mismatch case. Fig. 5a illustrates segmentation of human chromosome 1 (100 to 321 bins fragment in 40 kb resolution) between ESC and MSC cells. Bottom axis ticks represent bins, top triangles are TAD boundaries from MSC and bottom triangles are TAD boundaries belonging to ESC. Black, dashed vertical lines mark partitions, which increase matching (TAD boundaries in both sets overlap) while grey, dashed vertical lines denote partitions increasing mismatch.

Although in [15] the authors only consider the VI measure for global segmentation comparisons and do not discuss any local version of this measure, we thought that a

natural measure of local segmentation related to Variation of Information would be local MI (Mutual Information) expressed with following formula:

$$d_{A,B}^{\mathbf{MI}}(i) = -p_{A,B}(i) \cdot \log_2 \left(\frac{p_{A,B}(i)}{p_A(i) \cdot p_B(i)} \right) \tag{3}$$

where: $p_A(i) = \frac{l(f_{A,B}(i))}{N}$, $p_B(i) = \frac{l(f_{B,A}(i))}{N}$, $p_{A,B}(i) = \frac{l(o_{A,B}[i])}{N}$. In local MI score segments tend to be first ordered by their size (descending) and second by similarity of domains, which induce them (from match to mismatch), see fig. 5b.

Based on hypothesis that TADs define regulatory landscapes and may limit activity of certain regulatory elements [8] [9] we decided to check if there is a correlation between local score and differential gene expression or methylation pattern genome wide. To validate our prediction we derived differential gene expression and methylation data from publicly available resources. We then created all by all pairing of hi-c datasets for available cells and for each pair we assigned local BP and MI scores and fold change of gene expression or methylation for every gene (S3 Fig, S8 Fig). Due to large amount of genes we aggregated genes into 30 quantiles based on the local score value to reduce noise. Then for each group we plotted its quantile versus mean fold change (with standard deviation). This revealed corelations between gene expression or methylation fold change and the local domain rearrangement score (S4 Fig, S9 Fig).

Finally fig. 5c and fig. 5d illustrates the relationship between fold change of gene expression and methylation respectively as a function of local score using IMR90 and NPC cells as the example. Here genes were grouped based on local score so in summary there are 30 groups (quantiles). For each group median local score vs median fold change is shown. As can be seen the relationship is not strong, albeit visible and significant. Images for remaining pairs of cells can be found in supplement (S5 Fig, S10 Fig). The visible tendency of the local BP score to better correlate with methylation fold change and the MI score to better correlate with gene expression is generally representative of most datasets we analyzed. Corelations measured using PCC and Spearmans rho as well as significances are illustrated in S6 Fig and S7 Fig for gene expression, S11 Fig and S12 Fig for methylations and summarized in S1 Table and S2 Table.

Discussion

In this paper, we have considered the problem of comparisons between different segmentations of chromosomes into TADs. Currently, such comparisons are mainly based on comparing the locations of TAD boundaries and quantification of discrepancies between two sets of such locations. This is usually done using the Jaccard coefficient or some analogous function. We show in this work, that under a number of realistic scenarios, where the differences between TAD segmentations are driven by experimental noise, this is not an effective strategy. By utilising different ϵ perturbations of TAD segmentations, we can show that Jaccard Index is a poor technique for discerning significantly changed segmentations, from those with minor perturbations.

We provide an alternative scoring scheme, so called BP score, deriving its name from the bi-partite segment overlap graph. This score has several advantages over the Jaccard Index. Firstly, it is naturally scaling with the size of chromosomes and the resolution at which the segmentation is done, that is any change in the TAD structure will contribute to the metric proportionally to its size. Additionally, the method is treating equally all kinds of changes (insertions, deletions and shifts of domain boundaries) and can be meaningfully used to compare domain sets with different granularity, i.e. containing significantly different numbers of domains. On the technical level, the BP score is a proper distance metric, which has advantages for applications that need symmetry or

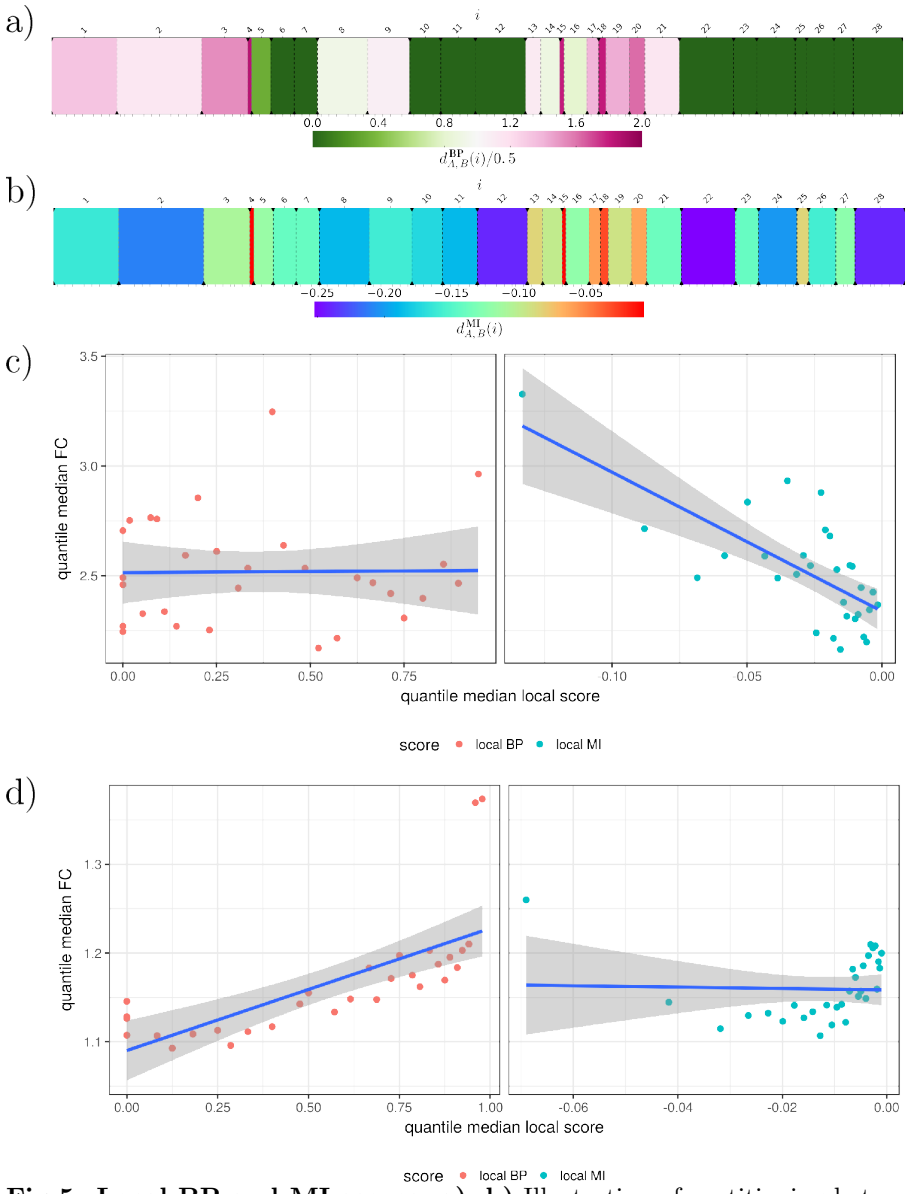

Fig 5. Local BP and MI scores. a), b) Illustration of partitioning between human ESC and MSC TADs of 4000 kb - 12840 kb (40 kb resolution) region of chromosome 1 with local BP score and local MI score respectively. **c)** Gene expression fold change vs local BP score or local MI score of 30 quantiles between human IMR90 and NPC cells . **d)** Same as in c), but with methylation fold change instead of gene expression.

triangle inequality, including clustering methods. The global BP metric is similar in the applicability to the Variation of Information metric suggested for the task earlier [15]. Not only are these methods giving more sensible results than JI in the three simulated scenarios of ϵ-matchings, but are also giving us a better discernibility between segmentations derived from experimental hi-c replicates and actual segmentations originating from different experimental conditions. This shows, that the researchers trying to obtain a meaningful score for TAD segmentation divergence should have more granular results with BP and VI scores than with JI or similar approaches.

In addition to the global VI and BP scores, we propose the analysis of local versions

of the two: local BP and local MI scores. The usefulness these local measures of domain rearrangement is supported by our analysis of experimental data. We show that in actual comparisons between different TAD segmentations performed on different cell lines, the regions with high local BP and local MI scores, i.e. those identified by the BP and MI score as the most rearranged, show the highest fold-changes in gene expression and methylation changes. This might be very useful in uncovering the mechanisms connecting chromosome structure rearrangement to functional changes like gene expression.

We provide an implementation of the BP, VI and JI scores that allows all researchers to test them on their own data. The implementation of the BP score is relatively simple as it can be computed as a sum of local scores for each of the atomic segment induced from the two segmentations being compared. This allows also to identify local contributions to the BP score, therefore leading to the identification of fragments of chromosomes that yield the most divergent parts of the chromosomes between the two segmentations under consideration. This might be specifically useful to the researches studying chromosomal domain dynamics, allowing them to focus their attention on parts of chromosomes that undergo the most dynamic TAD rearrangements.

Conclusion

All these results together show that the BP score is a valid metric for comparisons of TAD segmentations. It is superior to the widely used Jaccard Index, and can be useful in many comparative studies of chromatin domains. Not only is the BP score a better metric than the JI for the global comparison of segmentations, but its local version can give us an interesting measure of domain rearrangement that we show to be correlated with functional measurements.

We have focused our analyses of the BP score solely on its application to the TAD segmentations originating from hi-c experiments. This is however not an exclusive application of this measure. Mathematically, there is nothing in the BP score that would limit its adoption in the field of other chromosome segmentation situations, such as haplotype inference, epigenetic domains and many others.

Materials and methods

Hi-C data and its preprocessing

Raw Hi-C data was downloaded from Gene Expression Omnibus, repositories: GSE35156 and GSE52457. Reads were processed using Iterative Correction [16] pipeline: mapping, bining, filtering, contact maps generation and iterative correction steps with default parameters.

TADs generation

TAD boundaries were determined using Armatus software [15] (with default parameters) and processed to fill inter TAD gaps with artificial domains. As a result a set of TADs on any chromosome comprised sequence of nonoverlapping, consecutive intervals having total chromosome length the same for each cell type. Unmappable regions (contact map row/column sums to zero) were marked and excluded in all later analyses.

Calculation of Variation of Information

The VI distance between 2 partitions of the same chromosome is calculated as described in [15]. Briefly given 2 partitions A and B of length N:

$$d^{\mathbf{VI}}_{A,B} = H_A + H_B - 2 \cdot MI_{A,B} \tag{4}$$

where: H_A, H_B are entropies of partitions A and B respectively and $MI_{A,B}$ is their Mutual Information. Entropy of partition X is given by:

$$H_X = -\sum_{i}^{n} p_X(i) \cdot \log_2 \left(p_X(i)\right) \tag{5}$$

and Mutual Information of partitions A and B is expressed with:

$$MI_{A,B} = \sum_{i}^{|o|} p_{A,B}(i) \cdot \log_2 \left(\frac{p_{A,B}(i)}{p_A(i) \cdot p_B(i)} \right) \tag{6}$$

where: $p_X(i) = \frac{l(X[i])}{N}$, $p_A(i) = \frac{l(f_{A,B}(i))}{N}$, $p_B(i) = \frac{l(f_{B,A}(i))}{N}$, $p_{A,B}(i) = \frac{l(o_{A,B}[i])}{N}$ and n is a number of domains in X.

Differential gene expression

Gene expression data (bam files) was taken from http://epigenome.ucsd.edu/differentiation/download.html as pointed in [17]. Bam files were processed using edgeR [18] as described in manual to produce differential gene expression.

Methylation data

Methylations data was taken from: http://epigenome.ucsd.edu/differentiation/download.html [17] and processed using in house scripts to obtain fold changes for each segment in every cell type. Briefly each methylated position was asigned to segment and then for each segment a ratio of methylated by unmethylated counts was calculated.

Source code

The source coude is publicly available on github: https://github.com/rz6/bp-metric.

Proof of fact 1

We start with defining concepts used throughout this proof.

Definition 1. *A segment is semi-closed discrete interval:* $(a, b] = \{x \in \mathbb{N}^+ \mid a < x \leq b\}$ *with following standard relations:*

(i) equality: $(a, b] = (c, d] \iff a = c \wedge b = d$

(ii) subset: $(a, b] \subseteq (c, d] \iff a \geq c \wedge b \leq d$

(iii) intersection:

$$(a, b] \cap (c, d] \iff \begin{cases} \varnothing, & \text{if } a \geq d \vee c \geq b \\ \left(max\left(a, c\right), min\left(b, d\right)\right], & \text{otherwise} \end{cases}$$

In this study we use segments to describe chromosomes, TADs and overlaps between them.

Remark 1. *Each chromosome can be partitioned on collection of non-overlapping, consecutive segments (domains). We restrict notation to one chromosome for simplicity and from now on assume that any partition refer to the same chromosome. This notation can be naturally extended to multiple chromosomes for example by assumming segmentations of concatenated chromosomes with fixed boundaries between them. We will use capital letters to distinguish partitions, and we will use indexes to refer to sorted segments (domains):* $X[i], Y[j]$.

Definition 2. *We define following functions on segments:*

(i) segment start: $s((a,b]) = a$,

(ii) segment end: $e((a,b]) = b$,

(iii) segment length: $l((a,b]) = b - a$,

As partitions being compared $X, Y, ...$ refer to the same chromosome we can write $l(X) = l(Y) = ... = N$.

Definition 3. *Intersecting two partitions* X *and* Y *induces a partition we call the segmentation* $o_{X,Y}$ *with following properties:*

(i) $\forall_i \forall_j X[i] \cap Y[j] \neq \varnothing \implies X[i] \cap Y[j] \in o_{X,Y}$

(ii) $\forall_i \forall_{j>i} s(o_{X,Y}[i]) < s(o_{X,Y}[j])$

Remark 2. *When considering triplet partitions* X,Y,Z *we distinguish between 2 types of segments, atomic and non-atomic (or divisible). We call a segment* $(a,b]$ *atomic and write* $o[i]$ *if there is no other segment* $o_{X,Y}[k]$, $o_{X,Z}[l]$ *or* $o_{Y,Z}[m]$ *that is smaller than* $(a,b]$ *and included in* $(a,b]$. *Otherwise we call segment non-atomic and write* $o_{X,Y}[i]$.

For example in figure 6 segment $o_{A,C}[3]$ is not atomic - it can be further partitioned into $o[3]$ and $o[4]$ both of which are atomic.

Definition 4. *We define the function* $f_{X,Y}(i)$, *which gives the original segment from partition* X, *that included the segment* $o[i]$. *More formally:*

$$f_{X,Y}(i) = (a,b] \;\; s.t. \;\; (a,b] \in X \wedge o_{X,Y}[i] \subseteq (a,b]$$

We will use the same function to find segments in the second partition by writing $f_{Y,X}$.

Definition 5 (BP distance)**.** *Given 2 partitions* X *and* Y *s.t.* $l(X) = l(Y)$ *their BP distance is defined as:*

$$d(X,Y) = 1 - \frac{1}{N} \sum_{i}^{|o_{X,Y}|} \frac{l(o[i])^2}{max(l(f_{X,Y}(i)), l(f_{X,Y}(i)))} \tag{7}$$

Remark 3. *From now on we will use* A,B,C *to distinguish between 3 partitions used to construct this proof and* X,Y *whenever we want to refer to any pair of partitions from set* A,B,C *s.t.* $X \neq Y$.

Claim 1. *Function* d *satisfies:* $d(A,A) = 0$ *(identity of indiscernibles).*

Proof.

$$\begin{aligned} d(A,A) &= 1 - \frac{1}{N}\sum_{i}^{|o_{A,A}|} \frac{l(o_{A,A}[i])^2}{\max\left(l\left(f_{A,A}(i)\right), l\left(f_{A,A}(i)\right)\right)} \\ &= 1 - \frac{1}{N}\sum_{i}^{|A|} \frac{l(A[i])^2}{l\left(f_{A,A}(i)\right)} \\ &= 1 - \frac{1}{N}\sum_{i}^{|A|} \frac{l(A[i])^2}{l\left(A[i]\right)} \\ &= 1 - \frac{1}{N}\sum_{i}^{|A|} l(A[i]) \\ &= 1 - \frac{1}{N}\cdot N = 0 \end{aligned} \tag{8}$$

□

Claim 2. *Function $d(X,Y)$ is symmetric for any A,B.*

Proof.

$$\begin{aligned} d(A,B) &= 1 - \frac{1}{N}\sum_{i}^{|o_{A,B}|} \frac{l(o_{A,B}[i])^2}{\max\left(l\left(f_{A,B}(i)\right), l\left(f_{B,A}(i)\right)\right)} \\ &= 1 - \frac{1}{N}\sum_{i}^{|o_{B,A}|} \frac{l(o_{B,A}[i])^2}{\max\left(l\left(f_{B,A}(i)\right), l\left(f_{A,B}(i)\right)\right)} = d(B,A) \end{aligned} \tag{9}$$

□

Claim 3. *Let us consider 3 domain sets A,B,C (refer to figure 6 as example). Function d satisfies:*

$$d(A,B) \leq d(A,C) + d(B,C) \tag{10}$$

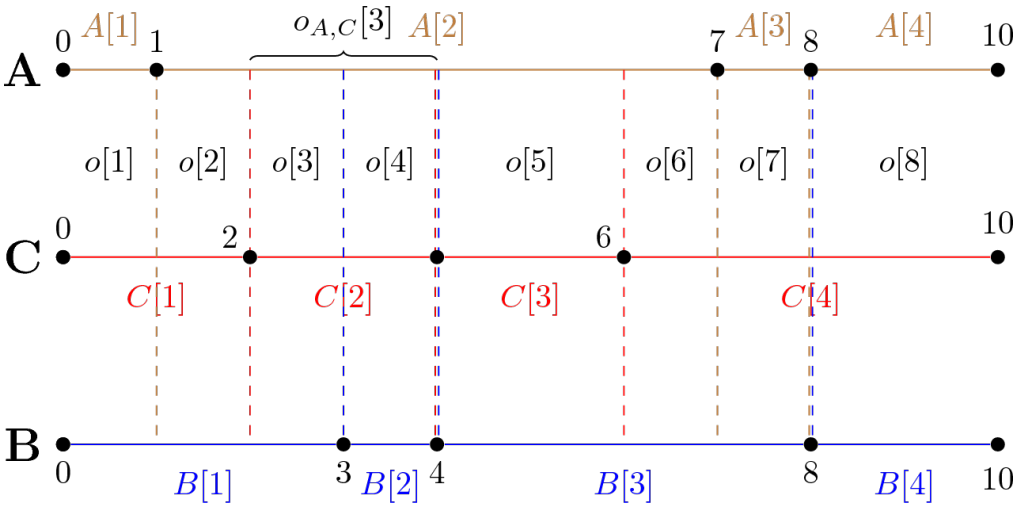

Fig 6. Ilustration of 3 partitions A, B, C and the induced segmentation.

Proof. We start with putting equation 7 into inequality 10:

$$\begin{aligned} &1-\frac{1}{N}\sum_{i}^{|o_{A,B}|}\frac{l(o_{A,B}[i])^2}{\max\left(l\left(f_{A,B}(i)\right),l\left(f_{B,A}(i)\right)\right)}\leq \\ &1-\frac{1}{N}\sum_{i}^{|o_{A,C}|}\frac{l(o_{A,C}[i])^2}{\max\left(l\left(f_{A,C}(i)\right),l\left(f_{C,A}(i)\right)\right)}+ \\ &1-\frac{1}{N}\sum_{i}^{|o_{B,C}|}\frac{l(o_{B,C}[i])^2}{\max\left(l\left(f_{B,C}(i)\right),l\left(f_{C,B}(i)\right)\right)} \end{aligned} \tag{11}$$

Using multinomial theorem we can substitute divisible segments with atomic ones in the following way:

$$l\left(o_{X,Y}[k]\right)^2=\left(\sum_{i}^{n(k)}l\left(o[i]\right)\right)^2=\sum_{i}^{n(k)}l\left(o[i]\right)^2+2\sum_{i}^{n(k)-1}\sum_{j}^{n(k)-i}l\left(o[i]\right)\cdot l\left(o[j]\right) \tag{12}$$

where: $n(k)$ is the number of atomic segments in divisible segment $o_{X,Y}[k]$. Obviously $n(k)$ also depends on X and Y, but for simplicity we decided to leave it out of the notation here assuming it follows from the formula. Using equation 12 we can rewrite inequality 11:

$$\begin{aligned} &N-\sum_{i}^{n}\frac{l\left(o[i]\right)^2}{\max\left(l\left(f_{A,B}(i)\right),l\left(f_{B,A}(i)\right)\right)} \\ &\qquad-2\sum_{k}^{n_{A,B}}\sum_{i}^{n(k)-1}\sum_{j}^{n(k)-i}\frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(f_{A,B}(k),f_{B,A}(k)\right)}\leq \\ &N-\sum_{i}^{n}\frac{l\left(o[i]\right)^2}{\max\left(l\left(f_{A,C}(i)\right),l\left(f_{C,A}(i)\right)\right)} \\ &\qquad-2\sum_{k}^{n_{A,C}}\sum_{i}^{n(k)-1}\sum_{j}^{n(k)-i}\frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(f_{A,C}(k),f_{C,A}(k)\right)}+ \\ &N-\sum_{i}^{n}\frac{l\left(o[i]\right)^2}{\max\left(l\left(f_{B,C}(i)\right),l\left(f_{C,B}(i)\right)\right)} \\ &\qquad-2\sum_{k}^{n_{B,C}}\sum_{i}^{n(k)-1}\sum_{j}^{n(k)-i}\frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(f_{B,C}(k),f_{C,B}(k)\right)} \end{aligned} \tag{13}$$

where: n is number of atomic segments and $n_{X,Y}$ is number of divisible segments induced by X,Y partitioning. As atomic segments are common for A, B and C we can omit the subscript in n.

Definition 6 (islands of segments)**.** *We define a family of segments $I_k=\{o[i]\mid o[i]\subseteq C[k]\}$ referred to as islands of segments, as satisfying the following condition:*

$$\underset{o[i]\in I_k}{\forall}\ \underset{\substack{o[j]\in I_k\\ i\neq j}}{\exists}\left[f_{A,C}(i)=f_{A,C}(j)\vee f_{B,C}(i)=f_{B,C}(j)\right]$$

Intuitively islands of segments give rise to product sums on the right side of inequality 13.

Claim 4. *Any atomic segment $o[i]$ satisfies:*

$$\forall_{1\leq i\leq n}\left[\frac{l\left(o[i]\right)^2}{max\left(l\left(f_{A,B}(i)\right),l\left(f_{B,A}(i)\right)\right)}\geq\frac{l\left(o[i]\right)^2}{max\left(l\left(f_{A,C}(i)\right),l\left(f_{C,A}(i)\right)\right)}\right.$$
$$\left.\vee\ \frac{l\left(o[i]\right)^2}{max\left(l\left(f_{A,B}(i)\right),l\left(f_{B,A}(i)\right)\right)}\geq\frac{l\left(o[i]\right)^2}{max\left(l\left(f_{B,C}(i)\right),l\left(f_{C,B}(i)\right)\right)}\right] \tag{14}$$

Proof.

1. first we can simplify notation by replacing $f_{X,Y}\left(i\right)$ with: $f_X\left(i\right)$ and $f_{Y,X}\left(i\right)$ with: $f_Y\left(i\right)$ as we are only concerned with atomic segments,
2. if $f_A(i) > f_B(i)$, then:
$$l\left(f_A(i)\right)\leq\max\left(l\left(f_A(i)\right),l\left(f_C(i)\right)\right)$$
3. otherwise $f_A(i)\leq f_B(i)$ and:
$$l\left(f_B(i)\right)\leq\max\left(l\left(f_B(i)\right),l\left(f_C(i)\right)\right)$$

This allows us to split squared terms from the right hand side of inequality 13 and merge them into 2 groups (S - smaller, R - remaining), both of size n:

- S is sum of terms from either A,C or B,C, such that each term satisfies condition 2 (if it is a term from A,C) or condition 3 (if it is a term from B,C),
- R are remaining terms, i.e. they may not satisfy the above conditions.

□

We can rewrite inequality 13:

$$\begin{aligned}N-\sum_i^n\frac{l\left(o[i]\right)^2}{\max\left(l\left(f_{A,B}(i)\right),l\left(f_{B,A}(i)\right)\right)}\\-2\sum_k^{n_{A,B}}\sum_i^{n(k)-1}\sum_j^{n(k)-i}\frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(f_{A,B}(k),f_{B,A}(k)\right)}\leq\\N-S-2\sum_k^{n_{A,C}}\sum_i^{n(k)-1}\sum_j^{n(k)-i}\frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(f_{A,C}(k),f_{C,A}(k)\right)}+\\N-R-2\sum_k^{n_{B,C}}\sum_i^{n(k)-1}\sum_j^{n(k)-i}\frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(f_{B,C}(k),f_{C,B}(k)\right)}\end{aligned} \tag{15}$$

Now:

$$\sum_i^n\frac{l\left(o[i]\right)^2}{\max\left(l\left(f_A(i)\right),l\left(f_B(i)\right)\right)}\geq S$$

so we know that:

$$\begin{aligned}N-\sum_i^n\frac{l\left(o[i]\right)^2}{\max\left(l\left(f_A(i)\right),l\left(f_B(i)\right)\right)}\\-2\sum_k^{n_{A,B}}\sum_i^{n(k)}\sum_j^{n(k)-i}\frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(l\left(f_{A,B}(k)\right),l\left(f_{B,A}(k)\right)\right)}\leq N-S\end{aligned} \tag{16}$$

After using 16 to simplify 15, what is left to show is that:

$$R + 2\sum_{k}^{n_{A,C}}\sum_{i}^{n(k)-1}\sum_{j}^{n(k)-i}\frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(l\left(f_{A,C}(k)\right), l\left(f_{C,A}(k)\right)\right)} + 2\sum_{k}^{n_{B,C}}\sum_{i}^{n(k)-1}\sum_{j}^{n(k)-i}\frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(l\left(f_{B,C}(k)\right), l\left(f_{C,B}(k)\right)\right)} \leq N \tag{17}$$

We note that no two segments $o_{A,C}[k]$ and $o_{B,C}[l]$ can generate two different atomic segments $o[i]$, $o[j]$ that would be properly included in them.

Claim 5. *There are no 2 segments $o_{A,C}[k]$ and $o_{B,C}[l]$, such that for two different indexes i,j $(i < j)$: $o[i] \subset o_{A,C}[k] \wedge o[j] \subset o_{A,C}[k]$ and $o[i] \subset o_{B,C}[l] \wedge o[j] \subset o_{B,C}[l]$.*

Proof.

1. assume that there are such 2 segments, which means that:
   (a) $s(o_{A,C}[k]) \leq s(o[i]) < e(o[i]) \leq s(o[j]) < e(o[j]) \leq e(o_{A,C}[k])$
   (b) $s(o_{B,C}[l]) \leq s(o[i]) < e(o[i]) \leq s(o[j]) < e(o[j]) \leq e(o_{B,C}[l])$
2. if $o[i] \subset o_{A,C}[k] \wedge o[j] \subset o_{A,C}[k]$ then:
   (a) $\exists_{u,v,u\neq v} e(B[u]) = e(o[i]) \wedge s(B[v]) = s(o[j])$
3. but 2 contradicts 1b as by definition of $o_{B,C}[l]$ we have:
   (a) $\exists_{w,z} s(o_{B,C}[l]) = \max(s(B[w]), s(C[z]))$
   (b) $\exists_{w,z} e(o_{B,C}[l]) = \min(e(B[w]), e(C[z]))$

   so: $s(B[w]) < e(o[i])$ and $s(o[j]) < e(B[w])$

□

Now we merge both product sums from inequality 17 and split them into m groups P_k corresponding to islands I_k:

$$P_k = \{(i,j) \mid i \neq j,\ o[i] \in I_k,\ o[j] \in I_k\}$$

We also simplify notation:

1. as we consider islands of segments we can replace $f_{C,A}(k)$ and $f_{C,B}(k)$ with $f_C(k)$,
2. for any $(i,j) \in P_k$ we introduce the notation $f_{A\vee B}(i)$:

$$f_{A\vee B}(i) = \begin{cases} f_{A,C}(i), & \text{if } f_{A,C}(i) = f_{A,C}(j) \\ f_{B,C}(i), & \text{otherwise} \end{cases}$$

Rewriting inequality 17 gives:

$$R + 2\sum_{k}^{m}\sum_{(i,j)}^{|P_k|}\frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right)} \leq N \tag{18}$$

The number of elements in P_k can be upperbounded by:

$$|P_k| \leq \binom{n}{2}$$

This allow us to upperbound the left hand side of inequality 18:

$$\begin{aligned} R + 2\sum_{k}^{m}\sum_{\substack{i,j \\ j>i}}^{|P_k|} \frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right)} \leq \\ R + 2\sum_{k}^{m}\sum_{i}^{n(k)-1}\sum_{\substack{j \\ j>i}}^{n(k)-i} \frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right)} \end{aligned} \tag{19}$$

We can now split R on 2 groups:

1. segments $o[i]$ such that: $\exists_{(u,v)\in P_k} i = u \vee i = v$,
2. remaining segments,

So we can write:

$$\begin{aligned} R = R_1 + R_2 = \sum_{k}^{m}\sum_{i}^{n(k)} \frac{l(o[i])^2}{\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right)} \\ + \sum_{i}^{n_r} \frac{l(o[i])^2}{\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right)} \end{aligned} \tag{20}$$

Now we put equation 20 into the right hand side of inequality 19:

$$\begin{aligned} \sum_{i}^{n_r} \frac{l(o[i])^2}{\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right)} + \sum_{k}^{m}\left[\sum_{i}^{n(k)} \frac{l(o[i])^2}{\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right)}\right. \\ \left. + 2\sum_{i}^{n(k)-1}\sum_{\substack{j \\ j>i}}^{n(k)-i} \frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right)}\right] \end{aligned} \tag{21}$$

In order to further upperbound the left hand side of inequality 21 we need to select the minimum possible denominator. It can be easily shown that it is minimum when:

$$\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right) = l\left(f_C(k)\right)$$

This follows from the definition of islands of segments as each atomic segment $o[i] \in I_k$ also satisfies: $o[i] \subset C[k]$ meaning $l(f_{A\vee B}(i)) \leq l(f_C(k))$. The latter upperbound let us write again:

$$\begin{aligned} &\sum_{i}^{n_r} \frac{l(o[i])^2}{\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right)} + \sum_{k}^{m}\left[\sum_{i}^{n(k)} \frac{l(o[i])^2}{\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right)}\right. \\ &\qquad \left. + 2\sum_{i}^{n(k)-1}\sum_{\substack{j \\ j>i}}^{n(k)-i} \frac{l\left(o[i]\right)\cdot l\left(o[j]\right)}{\max\left(l\left(f_{A\vee B}(i)\right), l\left(f_C(k)\right)\right)}\right] \\ &\leq \sum_{i}^{n_r} \frac{l(o[i])^2}{l\left(f_C(k)\right)} + \sum_{k}^{m}\left[\frac{1}{l\left(f_C(k)\right)}\left(\sum_{i}^{n(k)} l(o[i])^2 + 2\sum_{i}^{n(k)-1}\sum_{\substack{j \\ j>i}}^{n(k)-i} l\left(o[i]\right)\cdot l\left(o[j]\right)\right)\right] \\ &= \sum_{i}^{n_r} \frac{l(o[i])^2}{l\left(f_C(k)\right)} + \sum_{k}^{m}\left[\frac{1}{l\left(f_C(k)\right)}\left(\sum_{i}^{n(k)} l(o[i])\right)^2\right] \end{aligned} \tag{22}$$

We can further upperbound 22 by selecting minimum $l(f_C(k))$:

1. $\min(l(f_C(k))) = l(o[i])$ for atomic segments,
2. $\min(l(f_C(k))) = \sum_i^{n(k)} l(o[i])$ for divisible segments.

And we rewrite 22:

$$\begin{aligned}
&\sum_i^{n_r} \frac{l(o[i])^2}{l\left(f_C(k)\right)} + \sum_k^m \left[\frac{1}{l\left(f_C(k)\right)} \left(\sum_i^{n(k)} l(o[i]) \right)^2 \right] \\
&\leq \sum_i^{n_r} \frac{l(o[i])^2}{l\left(o[i]\right)} + \sum_k^m \left[\frac{1}{\sum_i^{n(k)} l(o[i])} \left(\sum_i^{n(k)} l(o[i]) \right)^2 \right] \\
&= \sum_i^{n_r} l(o[i]) + \sum_k^m \left[\sum_i^{n(k)} l(o[i]) \right] \\
&= \sum_i^{n_r} l(o[i]) + \sum_i^{n-n_r} l(o[i]) = N
\end{aligned} \tag{23}$$

Which is what we wanted to prove. □

Supporting information

S1 Fig. Technical and biological replicates distance distribution of real Hi-C datasets by chromosome using different metrics.

S2 Fig. Comparison of p-values between technical and biological replicates distance distribution of different chromosomes using different metrics.

S3 Fig. Relationship between gene expression fold change and local BP score or local MI score of all genes for different pairs of cell types.

S4 Fig. Relationship between gene expression fold change and local BP score or local MI score of all genes assigned to 30 quantiles for different pairs of cell types. Shown is mean gene expression fold change for each quantile.

S5 Fig. Relationship between gene expression fold change and local BP score or local MI score of all genes assigned to 30 quantiles for different pairs of cell types. Shown is median gene expression fold change vs median local score for each quantile as well as their Pearson corelation.

S6 Fig. Corelation between median quantile gene expression fold change and median quantile local score for different pairs of cell types.

S7 Fig. P value of corelation between median quantile gene expression fold change and median quantile local score for different pairs of cell types.

S8 Fig. Relationship between methylation fold change and local BP score or local MI score of all TADs for different pairs of cell types.

S9 Fig. Relationship between methylation fold change and local BP score or local MI score of all TADs assigned to 30 quantiles for different pairs of cell types. Shown is mean methylation fold change for each quantile.

S10 Fig. Relationship between methylation fold change and local BP score or local MI score of all TADs assigned to 30 quantiles for different pairs of cell types. Shown is median methylation fold change vs median local score for each quantile as well as their Pearson corelation.

S11 Fig. Corelation between median quantile methylation fold change and median quantile local score for different pairs of cell types.

S12 Fig. P value of corelation between median quantile methylation fold change and median quantile local score for different pairs of cell types.

S1 Table. Corelation between median quantile gene expression fold change and median quantile local score for different pairs of cell types.

S2 Table. Corelation between median quantile methylation fold change and median quantile local score for different pairs of cell types.

Acknowledgments

This work was supported by the Polish National Science Centre grant decision No. [DEC 2015/16/W/NZ2/00314].

References

1. Dekker, J. and Rippe, K. and Dekker, M. and Kleckner, N. Capturing chromosome conformation. Science 2002 Feb;295(5558):1306–1311

2. Lieberman-Aiden, E. and van Berkum, N. L. and Williams, L. and Imakaev, M. and Ragoczy, T. and Telling, A. and Amit, I. and Lajoie, B. R. and Sabo, P. J. and Dorschner, M. O. and Sandstrom, R. and Bernstein, B. and Bender, M. A. and Groudine, M. and Gnirke, A. and Stamatoyannopoulos, J. and Mirny, L. A. and Lander, E. S. and Dekker, J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009 Oct;326(5950):289–293

3. Zhao, Z. and Tavoosidana, G. and Sjolinder, M. and Gondor, A. and Mariano, P. and Wang, S. and Kanduri, C. and Lezcano, M. and Sandhu, K. S. and Singh, U. and Pant, V. and Tiwari, V. and Kurukuti, S. and Ohlsson, R. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet. 2006 Nov;38(11):1341–1347

4. Simonis, M. and Klous, P. and Splinter, E. and Moshkin, Y. and Willemsen, R. and de Wit, E. and van Steensel, B. and de Laat, W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet. 2006 Nov;38(11):1348–1354

5. Dostie, J. and Richmond, T. A. and Arnaout, R. A. and Selzer, R. R. and Lee, W. L. and Honan, T. A. and Rubio, E. D. and Krumm, A. and Lamb, J. and Nusbaum, C. and Green, R. D. and Dekker, J. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006 Oct;16(10):1299–1309

6. Dixon, J. R. and Selvaraj, S. and Yue, F. and Kim, A. and Li, Y. and Shen, Y. and Hu, M. and Liu, J. S. and Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012 Apr;485(7398):376–380

7. Flyamer, I. M. and Gassler, J. and Imakaev, M. and Brandao, H. B. and Ulianov, S. V. and Abdennur, N. and Razin, S. V. and Mirny, L. A. and Tachibana-Konwalski, K. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature 2017 Apr;544(7648):110–114

8. Andrey, G. and Montavon, T. and Mascrez, B. and Gonzalez, F. and Noordermeer, D. and Leleu, M. and Trono, D. and Spitz, F. and Duboule, D. A switch between topological domains underlies HoxD genes collinearity in mouse limbs. Science 2013 Jun;340(6137):1234167–10

9. Nora, E. P. and Lajoie, B. R. and Schulz, E. G. and Giorgetti, L. and Okamoto, I. and Servant, N. and Piolot, T. and van Berkum, N. L. and Meisig, J. and Sedat, J. and Gribnau, J. and Barillot, E. and Bluthgen, N. and Dekker, J. and Heard, E. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 2012 Apr;485(7398):381–385

10. Le Dily, F. and Bau, D. and Pohl, A. and Vicent, G. P. and Serra, F. and Soronellas, D. and Castellano, G. and Wright, R. H. and Ballare, C. and Filion, G. and Marti-Renom, M. A. and Beato, M. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. 2014 Oct;28(19):2151–2162

11. Dixon, J. R. and Jung, I. and Selvaraj, S. and Shen, Y. and Antosiewicz-Bourget, J. E. and Lee, A. Y. and Ye, Z. and Kim, A. and Rajagopal, N. and Xie, W. and Diao, Y. and Liang, J. and Zhao, H. and Lobanenkov, V. V. and Ecker, J. R. and Thomson, J. A. and Ren, B. Chromatin architecture reorganization during stem cell differentiation. Nature 2015 Feb;518(7539):331-336

12. Chandra, T. and Ewels, P. A. and Schoenfelder, S. and Furlan-Magaril, M. and Wingett, S. W. and Kirschner, K. and Thuret, J. Y. and Andrews, S. and Fraser, P. and Reik, W. Global reorganization of the nuclear landscape in senescent cells. Cell Rep 2015 Feb;10(4):471–

13. Barutcu, A. R. and Lajoie, B. R. and McCord, R. P. and Tye, C. E. and Hong, D. and Messier, T. L. and Browne, G. and van Wijnen, A. J. and Lian, J. B. and Stein, J. L. and Dekker, J. and Imbalzano, A. N. and Stein, G. S. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. Genome Biol. 2015 Sep;16:214–

14. Fraser, J. and Ferrai, C. and Chiariello, A. M. and Schueler, M. and Rito, T. and Laudanno, G. and Barbieri, M. and Moore, B. L. and Kraemer, D. C. and Aitken, S. and Xie, S. Q. and Morris, K. J. and Itoh, M. and Kawaji, H. and Jaeger, I. and Hayashizaki, Y. and Carninci, P. and Forrest, A. R. and Semple, C. A. and Dostie, J. and Pombo, A. and Nicodemi, M. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol. Syst. Biol. 2015 Dec;11(12):852–

15. Filippova, D. and Patro, R. and Duggal, G. and Kingsford, C. Identification of alternative topological domains in chromatin. Algorithms Mol. Biol. 2014 Apr;9:14–

16. Imakaev, M. and Fudenberg, G. and McCord, R. P. and Naumova, N. and Goloborodko, A. and Lajoie, B. R. and Dekker, J. and Mirny, L. A. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods 2012 Oct;9(10):999–1003

17. Xie, W. and Schultz, M. D. and Lister, R. and Hou, Z. and Rajagopal, N. and Ray, P. and Whitaker, J. W. and Tian, S. and Hawkins, R. D. and Leung, D. and Yang, H. and Wang, T. and Lee, A. Y. and Swanson, S. A. and Zhang, J. and Zhu, Y. and Kim, A. and Nery, J. R. and Urich, M. A. and Kuan, S. and Yen, C. A. and Klugman, S. and Yu, P. and Suknuntha, K. and Propson, N. E. and Chen, H. and Edsall, L. E. and Wagner, U. and Li, Y. and Ye, Z. and Kulkarni, A. and Xuan, Z. and Chung, W. Y. and Chi, N. C. and Antosiewicz-Bourget, J. E. and Slukvin, I. and Stewart, R. and Zhang, M. Q. and Wang, W. and Thomson, J. A. and Ecker, J. R. and Ren, B. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell 2013 May;153(5):1134–1148

18. Robinson, M. D. and McCarthy D. J. and Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010 Jan;26(1):139–140

19. Wilczynski, B. and Furlong, E. E. Challenges for modeling global gene regulatory networks during development: insights from Drosophila Dev. Biol. 2010 Apr;340(2):161–169

20. Song, L. and Zhang, Z. and Grasfeder, L. L. and Boyle, A. P. and Giresi, P. G. and Lee, B. K. and Sheffield, N. C. and Graf, S. and Huss, M. and Keefe, D. and Liu, Z. and London, D. and McDaniell, R. M. and Shibata, Y. and Showers, K. A. and Simon, J. M. and Vales, T. and Wang, T. and Winter, D. and Zhang, Z. and Clarke, N. D. and Birney, E. and Iyer, V. R. and Crawford, G. E. and Lieb, J. D. and Furey, T. S. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 2011 Oct;21(10):1757–1767

21. Rao, S. S. and Huntley, M. H. and Durand, N. C. and Stamenova, E. K. and Bochkov, I. D. and Robinson, J. T. and Sanborn, A. L. and Machol, I. and Omer, A. D. and Lander, E. S. and Aiden, E. L. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014 Dec;159(7):1665–1680

22. Nagano, T. and Lubling, Y. and Stevens, T. J. and Schoenfelder, S. and Yaffe, E. and Dean, W. and Laue, E. D. and Tanay, A. and Fraser, P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 2013 Oct;502(7469):59–64

23. Nagano, T. and Lubling, Y. and Varnai, C. and Dudley, C. and Leung, W. and Baran, Y. and Mendelson Cohen, N. and Wingett, S. and Fraser, P. and Tanay, A. Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature 2017 Jul;547(7661):61–67

24. Beagrie, R. A. and Scialdone, A. and Schueler, M. and Kraemer, D. C. and Chotalia, M. and Xie, S. Q. and Barbieri, M. and de Santiago, I. and Lavitas, L. M. and Branco, M. R. and Fraser, J. and Dostie, J. and Game, L. and Dillon, N. and Edwards, P. A. and Nicodemi, M. and Pombo, A. Complex multi-enhancer contacts captured by genome architecture mapping Nature 2017 Mar;543(7646):519–524

25. Yaffe, E. and Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. 2011 Oct;43(11):1059–1065

26. Dekker, J. and Heard, E. Structural and functional diversity of Topologically Associating Domains. FEBS Lett. 2015 Oct;589(20 Pt A):2877–2884

27. Crane, E. and Bian, Q. and McCord, R. P. and Lajoie, B. R. and Wheeler, B. S. and Ralston, E. J. and Uzawa, S. and Dekker, J. and Meyer, B. J. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 2015 Jul;523(7559):240–244

28. Feng, S. and Cokus, S. J. and Schubert, V. and Zhai, J. and Pellegrini, M. and Jacobsen, S. E. Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. Mol. Cell 2014 Sep;55(5):694–707