MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions

Koon-Kiu Yan[1,2] and Mark Gerstein[1,2,3*]

[1]Program in Computational Biology and Bioinformatics, [2]Department of Molecular Biophysics and Biochemistry, [3]Department of Computer Science, Yale University.

*Corresponding author

E-mail: pi@gersteinlab.org

Abstract

Genome-wide proximity ligation based assays such as Hi-C have revealed that eukaryotic genomes are organized into structural units called topologically associating domains (TADs). From a visual examination of the so-called chromosomal contact map, however, it is clear that the organization of the domains is not precisely defined. Instead, TADs exhibit various length scales, and in many cases nested organization can also be found. Here, by exploiting the resemblance between TADs in a chromosomal contact map and densely connected modules in a network, we formulate TAD identification as an optimization problem and propose an algorithm, MrTADFinder, to identify TADs from intra-chromosomal contact maps. MrTADFinder is based on the concept of modularity. A key component is to derive a background model for any given contact map, by numerically solving a set of matrix equations. The background model preserves the coverage of each genomic bin as well as the distance dependence of contact frequency for any pair of bins exhibited by the empirical map. Also, by introducing a tunable resolution parameter, MrTADFinder provides a self-consistent approach to identify TADs at different length scales, or resolutions. At a low resolution, larger TADs are found whereas, at a high resolution, smaller TADs are identified. We then apply MrTADFinder to identify TADs in various Hi-C datasets. The identified domains exhibit boundary signatures that are consistent with the earlier works. Moreover, by calling TADs at different resolutions, we observe that boundary signatures change with respect to the resolution, and different chromatin features may have different characteristic resolutions. We then report an enrichment of HOT regions near TAD boundaries and investigate the role of different transcription factors in determining domain borders at various resolutions. To further explore the interplay between domains organization and epigenomic features, we examine a distinctive pattern exhibited by the distribution of somatic mutations across boundaries. Overall, MrTADFinder provides a novel computational framework to explore the multi-scale structures stored in Hi-C contact maps.

Author Summary

The accommodation of the roughly 2m of DNA in the nuclei of mammalian cells results in an intricate structure, in which the topologically associating domains (TADs) formed by densely interacting genomic regions emerge as a fundamental structural unit. Identification of TADs is essential for understanding the role of 3D genome organization in gene regulation. By viewing the chromosomal contact map as a network, TADs correspond to the densely connected regions in the network. Motivated by this mapping, we propose a novel method, MrTADFinder, to identify TADs based on the concept of modularity in network science. Using MrTADFinder, we identify domains at various resolutions, and further explore the interplay between domains and other chromatin features like transcription factors binding and histone modifications at different resolutions. Overall, MrTADFinder provides a new computational framework to investigate the multiple length scales that are built inside the organization of the genome.

Introduction

The packing of a linear eukaryotic genome within a cell nucleus is dense and highly organized. Understanding the role of 3D genome in gene regulation is a major area of research [1][2][3][4]. Recently, genome-wide proximity ligation based assays such as Hi-C have provided insights into the complex structure by revealing various structural features regarding how a genome is organized [5][6][7]. Perhaps, one of the most important discoveries is the domain of self-interacting chromatin called topologically associating domain (TAD) [8][9]. Inside a TAD, genomic loci interact often but interactions between different TADs are less frequent. The TAD emerges as a fundamental structural unit of chromatin organization; it plays a significant role in mediating enhancer-promoter contacts and thus gene expression, and breaking or disruption of TADs suggested to diseases like cancers [10][11][12]. Therefore a deeper understanding of TADs from Hi-C data presents an important computational problem.

Results of a typical Hi-C experiment are usually summarized by a so-called chromosomal contact map [5]. By binning the genome into equally sized bins, the contact map is essentially a matrix whose element (i, j) reflects the population-averaged co-location frequencies of genomic loci originated from bins i and j. In this representation, TADs are displayed as blocks along the diagonal of a contact map [8][9]. Despite the fact that TADs are rather eye-catching in a contact map, computational identification is still challenging because of experimental factors such as noise and inadequate coverage. Moreover, it is apparent from a visual examination of the contact map that TADs exhibit various length scales: there are TADs that appear to be overlapping, and within many TADs, there are rich sub-structures.

Mathematically speaking, it is very natural to transform a contact matrix to a weighted network in which nodes are the genomic loci (or bins) whereas the interaction between two loci is quantified by a weighted edge. In network science, a widely studied problem is the identification of network modules, also known as community detection problem [13]. A module refers to a set

of nodes that are densely connected. In its simplest form, the community detection problem concerns with whether nodes of a given network can be divided into groups such that connections within groups are relatively dense while those between groups are sparse. Therefore, by viewing the chromatin interactions as a network, the highly spatially localized TADs immediately resemble densely connected modules. Motivated by the resemblance, we formulate the identification of TADs as a global optimization problem based on the observational contact map and a background model. As a network-based approach, our method goes beyond a direct adaptation of standard community detection algorithms. We introduce a novel background model that takes into account the effect of genomic distance, which is specific to the context of genome organization. The objective function is optimized using a heuristic algorithm that is efficient even if the size of the input contact map is large. Furthermore, by introducing a tuning parameter, our network approach can identify TADs at different resolutions. At a low resolution, larger TADs are found whereas, at a high resolution, smaller TADs are identified as the nucleome is viewed on a finer scale. In other words, the method can identify TADs at different length scales. We name our method MrTADFinder where the acronym Mr stands for multiple resolutions.

Results

A network modularity framework for TADs identification

The identification of modules in a network is formulated as a global optimization problem on the so-called modularity function over possible divisions of the network. Consider an unweighted network represented by an adjacency matrix A. For a particular division (i.e. a mapping from the set of all nodes to a set of modules), the modularity is defined as the fraction of edges within modules minus the expected fraction of such edges in a randomized null model of the network. Mathematically, the modularity is equal to

$$\frac{1}{2m}\sum_{i,j}\left(A_{ij}-\frac{k_i k_j}{2m}\right)\delta_{\sigma_i\sigma_j}. \tag{1}$$

Here, the summation goes over all possible pairs of nodes, the value of the Kronecker data $\delta_{\sigma_i\sigma_j}$ equals one if nodes i and j have the same label σ and zero otherwise, meaning only pairs of nodes within the same module are summed. In particular, m is the number of edges in the network whereas the expression $k_i k_j/2m$ represents the expected number of edges between i and j in a so-called configuration model. The configuration model is a randomized null model in which the degrees of nodes k_i are fixed to match those of the observed network, but edges are in other respects placed at random. High values of the modularity correspond to good partitions of a network into modules and similarly low values to bad partitions. Optimizing the modularity function leads us to the best partition over all possible partitions. More recently, a so-called resolution parameter γ has been incorporated in equation (1) to adjust the size of the resultant modules [14].

Following the network formalism, given a Hi-C contact map represented by a weighted matrix W, we define a similar objective function Q as

$$Q=\frac{1}{2N}\sum_{i,j}\left(W_{ij}-\gamma E_{ij}\right)\delta_{\sigma_i\sigma_j} \tag{2}$$

Here, i,j index the equally binned genomic loci. N is the total number of pair-end reads. E_{ij} is the expected number of contacts between locus i and locus j. γ is the resolution parameter that could be used to tune the size of resultant TADs. Very much similar to the network setting, the identification of TADs aims to partition the loci into domains such that Q is optimized. Nevertheless, it is important to emphasize two points. First, unlike the case in a network, the bins in a chromosome form a continuous chain and therefore genomic loci belonging to a TAD have to form a continuous segment. Second, simply because of the physical nature of chromosome, the

expected number of contacts between locus i and locus j depends on their genomic distance. Two loci that are close together in a 1-dimensional sense are expected to have a higher contact frequency as compared to two loci that are far apart. This point suggests that the null model E_{ij} in equation (2) has to be modified.

A novel null model of intra-chromosomal contact maps

Given an intra-chromosomal contact map W, the expected null model E is defined as

$$E_{ij} = \kappa_i^* \kappa_j^* f(|i-j|). \tag{3}$$

Here, f is the average number of contacts as a function of distance $d = |i-j|$. By considering all possible pairs of bins in W in terms of their distance apart and the contact frequency, we estimate f by local smoothing (see Methods). For intermediate values of d, f follows pretty well with a power-law function d^{-1} (see Figure S1), which is a well-known observation first reported in [5].

As a null model, the resultant E matrix satisfies a set of constraints, namely

$$\begin{aligned} \sum_j E_{ij} &= \sum_j W_{ij} = c_i \quad \forall i, \\ \sum_{ij} E_{ij} &= \sum_{ij} W_{ij} = 2N. \end{aligned} \tag{4}$$

The first equation means that the coverage c_i, i.e. the total number of reads (one end of pair-end reads) mapped to bin i, defined in the observed map is the same as the coverage defined in the null model. The second equation is a direct consequence of the first equation, where N is the total number of pair-end reads mapped to the chromosome. As f has been estimated from the observed W, we can numerically solve all the unknowns κ_i^* in the system of matrix equations (see Methods). Mathematically, κ_i^* can be regarded as an effective coverage because of the correlation between κ_i^* and the coverage c_i is extremely high (r=0.95, Figure S2). In comparison with equation (1), κ_i^* is conceptually analogous to the degree k_i. As shown in Figure 1, given a particular matrix W, the contact frequency of the resultant null model E are the highest in the

diagonal and decrease gradually away from the diagonal. With W and E, for any given resolution parameter γ, we employ a modified Louvain algorithm to optimize Q (see Methods for details). To ensure robustness, multiple runs of the modified Louvain algorithm are performed (Figure 1 and Methods). It is important to emphasize that the conventional Louvain algorithm used in network analysis [15] cannot be directly used because chromatin domains are continuous segments.

Identifying TADs in multiple resolutions

As a demonstration, we applied MrTADFinder to analyze Hi-C data of hES cell from [8]. Figure 2A shows a particular snapshot of the contact map (for chromosome 10) and its alignment with the identified TADs. In general, the TADs displayed agree well with the apparent block structures in the contact map. Of particular interest is the choice of γ that capture various length scales in domains organization. As shown in Figure 2A, when γ increases, a large TAD breaks into a few small TADs. On the other hand, a few large TADs merge together to form an even larger TAD as the value of γ is lowered. Statistically speaking, γ quantifies to what extent do we accept the enrichment of empirical contact frequency over the expectation. As γ increases, only matrix elements close to the diagonal contribute positively to the objective function. Therefore, in general, the size of TADs decreases (see Figure 2B) and the number of TADs increases (see Figure 2C). For example, when γ=1.0, there are about 1000 TADs in hES cells with a median size of 3Mb. When γ=2.25, the number of TADs increases to 2600 and the median size is roughly 1Mb.

We then further compared the TADs identified at different resolutions by MrTADFinder with TADs identified by a previous method. As quantified by the normalized mutual information (see Methods for details), TADs identified by MrTADFinder best match with TADs identified in [8] when the resolution parameter is 2.9. In general, unless the resolution is sufficiently small

($\gamma < 1.5$), the two methods are quite consistent (see Figure 2D). Nevertheless, the introduction of the resolution parameter γ opens an extra dimension in domain identification in a sense the algorithm used in [8] focuses on a particular resolution instead.

Chromatin signatures near TAD boundaries identified in various resolutions

The interplay between 3D genome organization and various chromatin features has widely been investigated since some of the first Hi-C experiments were reported [5][8][9]. By identifying TADs and their boundaries using MrTADFinder, we found the boundary signatures that are consistent with the observations previously reported [8], for instance, the enrichment of active promoter mark H3K4me3 or active enhancer mark H3K27ac, as well as the depletion of transcriptional repression mark like H3K9me3 (Figure 3A). Nevertheless, though the appearance of quite many sharp peaks at the boundaries is rather obvious, no general pattern emerges by aligning a variety of chromatin features with TADs (Figure 2A). To better understand the relationship between domains organization and different chromatin features, we further examined the chromatin features near different sets of boundaries that were identified in different resolutions. We found that in general, the enrichment of peak density at boundary decreases as resolution increases, indicating that various chromatin features appear in the boundaries of low-resolution TADs do not appear in high-resolution TADs (Figure 3A). More specifically, the enrichment of histone marks like H3K36me3 and H3K4me3 exhibits a monotonic drop whereas certain marks exhibit characteristic resolutions. For instance, the enrichment of mark H3K27me3 remains high up to a resolution of $\gamma = 2.5$ (Figure 3B).

Apart from histone modifications, it is well known that certain transcription factor binding sites are enriched near the boundary regions of TADs [8]. Instead of looking at individual factors, we further explored the location of the so-called HOT regions and XOT regions on TADs. High-occupancy target (HOT) regions and extreme-occupancy target (XOT) regions are

genomic regions that are bound by an extensive amount of transcription factors [16]. As expected, we found a strong enrichment of HOT regions and an even stronger enrichment of XOT regions near TAD boundaries in hES cells (Figure 3C). The observation is, in general, true for all tested resolutions. The observation agrees with the idea that HOT regions are very accessible regions in open chromatin. Nevertheless, it is still widely unknown if transcription factors bind to HOT regions simply because of thermodynamics, or the binding will result in important biological consequences.

Motivated by the observation that many factors tend to bind to the boundary regions, we further examine which factors are responsible for establishing the domain border, and more interestingly for borders in different resolutions. To do so, we employed a logistic regression model recently proposed by [17]. The model classifies a set of boundaries identified by MrTADFinder versus a set of random boundaries by integrating the binding signals of 60 transcription factors (see Methods for details). Generally speaking, the model is quite successful (AUC=0.81, Figure S3). The result is consistent with an early work based on histone modifications [18]. An interesting observation is, the predicting power of the model decreases as the resolution increases. The regression model further quantifies explicitly the influence of each of the transcription factors. In general, factors that are responsible for border formation are quite consistent across different resolutions (Figure 3D). For instance, factors like CTCF, Rad21 and CHD7 are direct drivers of border establishment and maintenance, whereas factors like MYC have a consistently negative effect. Nevertheless, the relative importance of factors does change with resolutions. For instance, Rad21 has a higher predictive power in classifying high-resolution domains in compared with classifying low-resolution domains.

TAD boundaries and mutational burden

We have examined the interplay between domains organization and chromatin features. Recently, it has been reported that epigenomic features shape the mutational landscape of cancer [19]. Motivated by this linkage, we further investigated the occurrence of somatic mutations near the boundaries. More specifically, we mapped the somatic mutations obtained from breast cancer samples to the TAD boundaries we identified in MCF7 cells (see Methods). In a given resolution, there are 85 boundary regions identified on chromosome 10. The regions can be clustered into 3 groups based on the positional distribution of somatic mutations (see Figure S4). As in shown in Figure 4, two of the clusters exhibit a step-function behavior (blue and red) in which the abrupt transition essentially happens at the boundary. For boundary regions in the remaining cluster, the mutational burden exhibits no difference across the TAD boundaries. Because of the close relationship between TADs and replication-timing domains [20], the observation resonates with a well-known observation that genomic regions with a high mutational burden are replicated at a later stage during DNA-replication [21]. As shown in the inset, using Repli-seq data in S1 phase, the upstream regions of the boundaries found in the blue cluster have a high mutation rate but a low Repli-seq signal, meaning they are indeed replicated at a later stage during replication. On the contrary, the upstream regions of the boundaries found in the red cluster are replicated at an early stage and therefore exhibit a low mutation rate.

Motivated by the relationship between TADs and DNA replication, we overlaid TADs in different resolutions with data from Repli-seq experiment (Figure S5). We observed that TADs identified in different resolutions match with the Repli-seq data in different stages of a cell cycle. For instance, while a TAD identified in a low resolution does not replicate at an early phase, say S1, its sub-structures identified in a higher resolution correspond to two separate peaks at later stages, say S2 and S3 (Figure S5).

Comparison with existing methods based on CTCF enrichment

There are quite a few existing methods on identifying TADs using Hi-C data. Dixon *et al*. identified TADs based on the so-called directionality index using Hi-C data in hES cell and found an enrichment of CTCF binding sites at the boundary regions [8]. Since then the enrichment of chromatin features has been used as a benchmark for various TAD calling algorithms [22][23][24]. As a comparison, we performed the same analysis using TADs based on MrTADFinder. As shown in Figure 5, both methods exhibit a similar pattern. In fact, as reported in [22][23][24], the enrichment pattern of CTCF binding peaks is qualitatively the same for all the proposed methods. By repeating the analysis in different resolutions, we observed that the level of enrichment depends on the resolution (Figure 5, Figure S6). At a low resolution, i.e. for larger TADs, the enrichment signal is stronger, and the signal tends to extend over a longer distance from the boundary. At a higher resolution, the signal is weaker and confined to near the boundary. In general, Figure 5 suggests that boundaries identified in lower resolutions are more likely to be bound by CTCFs. From a biological standpoint, as a boundary identified in a lower resolution separates two large domains, the results may bring insights on how to mediate chromatin loops at different length scales via an important architectural protein [25][26]. As the level of CTCF enrichment might be the consequence of different chromatin length scales, it might not be fair to use it directly for benchmarking the performance of different algorithms.

Robustness and performance of MrTADFinder

Because of the stochastic nature of the modified Louvain algorithm, we explored the robustness of MrTADFinder. In the current setting based on multiple runs of the modified Louvain procedure, we found the results of two independent callings highly robust. In fact, the normalized mutual information is higher than 0.99 (see Figure S7).

MrTADFinder is implemented in Julia. Julia programmers can import MrTADFinder as a library for calling various functions. It can also be run in command line if Julia and the required packages are installed. The performance of MrTADFinder, in general, depends on the size of the

input contact map. We have tested the performance using the contact maps of GM12878 cell generated by the Aiden lab [27]. The performance is reasonable. For instance, for chromosome 10, in a bin-size of 25kb (i.e. a contact map 5400 by 5400), the time required to arrive at all TADs with 10 runs of Louvain algorithm is about 20 minutes on a laptop with 2.8GHz Intel Core i7 and 16Gb of RAM. The time required is only 6 minutes if the bin size is 50kb.

Optimization using dynamic programming

Despite the similarity between equations (1) and (2), network modules are rather arbitrary collections of nodes, but domains are continuous segments along the chromosome. In fact, the total number of possible partitions for a chromosome is much smaller than the total number of ways to divide a network into modules. As a result, while the optimization of equation (1) is an NP-hard problem, the optimization of (2) can be quite efficiently solved using dynamic programming (see Methods and Figure S8). It is instructive to explore this avenue because quite some algorithms for identifying TADs are based on dynamic programming but with different objective functions [22][23][24].

The time complexity of this dynamic programming algorithm is in order of $O(n^3)$, where n is the size of the contact map. Given the time complexity, finding the optimal partition using a bin size of 40kb is quite impractical. Therefore, though the connection between identifying TADs and problems like finding RNA secondary structure is of theoretical interest, MrTADFinder is developed based on the modified Louvain algorithm. Nevertheless, we have implemented the dynamic programming approach and performed a comparison with the heuristic. Using a contact map of hES cell (chromosome 1) with a bin size of 500kb, we found the sub-optimal partitions based on our modified Louvain algorithm are very close to the optimal partition based on dynamic programming. The normalized mutual information between optimal and sub-optimal values is 0.977±0.007.

Discussion

In this paper, we have introduced an algorithm to identify TADs from Hi-C data and performed several analyses to show the biological significance of the TADs identified. In particular, by introducing a single continuous parameter γ, we can further examine domains organization and its interplay with a variety of chromatin features in multiple resolutions. It is important to emphasize that the idea of resolution we introduced in MrTADFinder is different from some other usages of the same term in Hi-C analysis. From an experimental standpoint, the resolution of a Hi-C experiment refers to the average fragment size as digested by restriction enzymes (~4kb to ~1kb) [5][27] or more recently by micrococcal nuclease (~150bp) [28]. Regarding the construction of contact maps, the term resolution has been used to refer to the bin size, where the proper choice usually depends on the number of reads in the stage of data processing. Both usages are primarily technical. What we mean by resolution, however, refers to the multiple length scales built inside the organization of the genome. It is well known that there are structures in different length scales such as compartment, domains, and sub-domains [29], and chromatin features like histone marks exhibit multiple length scales [30]. The concept of resolution introduced here points to the integration of these structures and enables one to explore the rich structures hidden in contact maps.

A novel contribution of this work is the derivation of an expected model for any intra-chromosomal contact map by solving a system of matrix equations. The null model preserves the coverage of each genomic bin as well as the distance dependence of contact frequencies in the observed map. As such features of contact maps are involved in most computational analysis of Hi-C data, apart from the identification of TADs, the expected model can be used for applications like finding compartments [5] and identifying potential enhancer-target linkages [31]. Mathematically, the expected matrix is solved by an iterative procedure. The procedure can be

regarded as a generalization of a class of matrix balancing methods used for normalizing Hi-C matrices [32], as the later is merely a different set of matrix equations. However, it is important to emphasize that the so-called ICE algorithm aims to remove bias in the contact map, whereas our method aims to generate a background model. While MrTADFinder focuses on intra-chromosomal interactions, recent studies employ various clustering methods to identify inter-chromosomal clusters using Hi-C contact frequency [33][34]. It is worthwhile to point out that similar expected models used in this study can also be derived for inter-chromosomal interactions to better separate signal and noise.

Several methods have been developed for identifying TADs from Hi-C data [35]. One of the earliest methods is based on the so-called directionality index, a 1D statistic measuring whether the contacts have an upstream or downstream bias [8], and later the bias is exploited by the so-called arrowhead algorithm [27]. Later algorithms exploit the block diagonal nature of TADs in a contact map [22][36]. However, the dependence of intra-chromosomal interactions and genomic distance is not explicitly modeled. The algorithm TADtree does model the distance dependence but does not take into account both the genomic distance and the effects of coverage in a compact mathematical formalism [23]. The algorithm TADtree, and more recent efforts, namely Matryoshka [24] and metaTAD [37] aim to investigate the hierarchical organization of TADs based on a tree structure. Indeed, merging smaller TADs at the lower level of the hierarchy results at larger TADs similar to the TADs obtained by MrTADFinder at a low resolution. Nevertheless, MrTADFinder does not impose a hierarchical organization. The probabilistic nature of Louvain algorithm enables the definition of TAD boundaries in a probabilistic fashion, and therefore a possibility to define overlapping TADs. To a certain extent, the idea of continuous resolution used in MrTADFinder is distinct in comparison with algorithms based on a bottom-up approach, but similar in spirit to Ref. [22] .

MrTADFinder is motivated by the community detection problem in network studies. Although a network perspective of chromosomal interactions has previously been proposed

[38][39], a lot of widely studied concepts in networks have rarely been explored in the context of chromosomal organization. A network representation is arguably more flexible than a simple matrix representation, for instance, transcription factors binding and histone modifications can be easily incorporated into the network, forming a decorated network. Moreover, one could extend the framework by concatenating multiple Hi-C contact maps to form a multi-layer network. The same idea has been used for cross-species transcriptomic analysis [40]. By facilitating the application of a variety of graph-theoretical tools, we believe that network algorithms will be useful for future studies on the spatial organization of the genome.

Materials and methods

Hi-C data and their pre-processing

The Hi-C data of human ES cells and IMR90 cells were reported in Ref. [8]. Hi-C data in MCF cells were reported in Ref. [41]. Data in GM12878 were reported in [27]. Raw reads were processed using Hi-C Pro [42], arriving at contact matrices in various bin sizes. In all analysis, the whole-genome contact map was iteratively corrected for uniform coverage [32]. Intra-chromosomal contact maps were then extracted from the whole-genome contact map of bin size 40kb for downstream analysis. Contact maps were all generated by the tool HiCPlotter [43].

Chromatin Data

All chromatin data, including histone modifications, transcription factors binding, expression, replication timing, were downloaded from the ENCODE data portal.

Deriving a background model for any given intra-chromosomal contact map

The average number of contacts as a function of genomic distance can be estimated by considering all elements in matrix W. A local smoothing approach similar to the method used in

[44] was employed. The window size equals to 1% of the data.

Equation (3) and (4) can be rewritten in the form

$$\sum_j \kappa_i^* \kappa_j^* f(|i-j|) = c_i \quad \forall i. \tag{5}$$

The system of non-linear equation is similar to the matrix balance approach used in [32]. As the aim of [32] is to remove bias, the coverage c_i is the same for all bin i and f is replaced by the original empirical map. Nevertheless, the unknowns κ_i^* can be used by a similar iterative procedure as proposed in [32].

Heuristic procedures for optimizing Q

To optimize the objective function Q, we employ a modified version of Louvain algorithm [15], which is widely used in identifying modules in networks. In a nutshell, the algorithm consists of two steps. The algorithm starts as every bin has its own label, and the label will end up as an identifier for the module it belongs. In the first step, for each bin, we update its label by either choosing the label of one of its two neighboring bins or by remaining unchanged based on whether or not the value of Q will be increased. There will be multiple rounds of updates in this step. For each round of update, we go through all the bins once, but the order is random. The updating procedure will be repeated for multiple rounds until no more update is possible. We will then perform the second step such that the bins with the same labels will be locked together, in a sense their labels will only be updated in a synchronized fashion. It is worthwhile to mention that the updating procedure in the first step makes sure bins with the same labels form a continuous segment. Once the bins are locked to form super-bins, the first step will be performed again but in the level of super-bins. The two steps will be repeated iteratively until no increase of modularity is possible.

The output of the modified Louvain algorithm is essentially a particular partition of the entire chromosome. As the result of the algorithm, in general, depends on the order of updates,

multiple runs are performed to probe the fuzziness of the assignment. As the chromosome is binned into n equally sized bins, we examine, say after 10 trials, how likely the border between bin i and bin $i+1$ is indeed a domain boundary, i.e. bin i and bin $i+1$ are called to belong to two different TADs by the modified Louvain algorithm. We then naturally define a boundary score for each of the n+1 borders as the fraction of trials in which a border is called as a boundary. To define a set of consensus boundaries, we choose a cut-off of 0.9. In other words, the border between two adjacent bins is defined as a confident boundary only if they are called to belong to two different domains in at least 9 out of 10 trials. The final output of MrTADFinder is a set of consensus TADs defined as regions between the consensus domains

The boundary score assigned to each border is not merely an immediate but serves as a proxy of the degree of insulation. A border with a high boundary score is more effective in forbidding the contacts between its left and right regions.

Quantifying the consistency between two sets of TADs

Given two sets of TADs, say in different cell lines, or called by different algorithms, we employ the so-called normalized mutual information to quantify the consistency. Suppose X and Y are two random variables whose values x_i and y_i represent the corresponding domain labels of bin i. The normalized mutual information $\mathrm{MI}_{\mathrm{norm}}$ is defined as

$$MI_{norm} = \frac{2I(X;Y)}{H(X)+H(Y)}, \tag{6}$$

here $H(X), H(Y)$ are the entropy of X and Y, and $I(X;Y)$ is the mutual information quantifying to what extent the domain labels in X predict the labels in Y. A normalized form of mutual information is used here to make sure the value lies between 0 and 1 for comparison. To have a fair comparison, bins that are not assigned to any TADs in both sets of partitions are not counted. If two sets of partitions are identical, the value of normalized mutual information is 1.

Chromatin signatures within TADs in different resolutions

Given the location of binding peaks of a transcription factor or a histone mark, the peak density near TAD boundaries was estimated by considering for all boundaries the region from upstream 600kb to downstream 600kb. The regions were aligned, and the number of peaks was summed accordingly. To calculate the enrichment, the number of peaks was normalized by the expected number of peaks in a particular region under a null model that peaks are randomly distributed in the genome.

The influence of individual transcription factors on the formation of domain borders was formulated as a classification problem. For a particular resolution, the set of boundaries called by MrTADFinder was used as a positive set whereas a set of random boundaries obtained by swapping the TADs along the genome was chosen as the negative set. The signal values of 60 transcription factors are used as features for classification. The combined effect of all features was modeled the logistic function

$$f\left(X,(\beta_0,\boldsymbol{\beta})\right) = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{exp}(-\beta_0+\boldsymbol{\beta X})}, \tag{7}$$

here X represents all features; $\boldsymbol{\beta}$ is a vector determining the coefficients of influence for all features and β_o is a bias parameter. Using the training set, a likelihood function was defined. An optimal $\boldsymbol{\beta}$ was inferred by optimizing the likelihood function using gradient descent with L1-regularization. To have a more accurate estimate, 10-fold cross-validation was performed, and the calculation was done with multiple negative training sets.

Somatic mutations

The set of somatic mutations were downloaded from the data portal of the International Cancer

Genome Consortium (ICGC). The mutations were called the breast cancer samples of 676 donors. The samples were sequenced in a whole-genome level. Breast cancer samples were used in this analysis to match the Hi-C data of MCF7 cell.

Dynamic programming

The idea is to extensively enumerate all the possible partitions of the chromosome. In a nutshell, a binned chromosome can be considered as a sequence $(1, 2, \cdots, n-1, n)$. Rather than partitioning the whole sequence at a first place, we look for the optimal partition for all the possible sub-sequences starting from sub-sequences with length 1. Let us denote the optimal value of modularity Q for a sequence $a_1 a_2 \ldots a_{l-1} a_l$ as $optQ(a_1 a_2 \ldots a_{l-1} a_l)$. The value is the maximum of the following l possibilities:

$$\begin{gathered} optO(a_1) + optO(a_2 \ldots a_{l-1} a_l), \\ optO(a_1 a_2) + optO(a_3 \ldots a_{l-1} a_l), \\ \vdots \\ optO(a_1 a_2 a_{3\ldots} a_{l-1}) + optO(a_l), \\ \sum_{ij} Q_{ij}. \end{gathered} \tag{8}$$

Suppose the maximum is the sum $optO(a_1 a_2 \ldots a_r) + optO(a_{r+1} \ldots a_{l-1} a_l)$, where $1 \leq r < l$. The sum corresponds to the case that the optimal partition of $a_1 a_2 \cdots a_l$ is a combination of the optimal partitions of $a_1 a_2 \cdots a_r$ and $a_{r+1} \ldots a_{l-1} a_l$ (see Figure S7). It is not necessary that $a_1 a_2 \cdots a_r$ forms a single domain. The key is that the expression $optQ(a_1 a_2 \ldots a_{l-1} a_l)$ can be found recursively because all possibilities depend on the optimal values of sub-sequences shorter than l. The last summation in (4) sums Q over all positions from a_1 to a_l, meaning the l bins belong to the same domain. Once the value of $optQ(a_1 a_2 \ldots a_{n-1} a_n)$ is found, we can trace back the actual partition for the whole chromosome. The procedure is analogous to the Nussinov

algorithm in finding the optimal secondary structure of RNA [45].

Acknowledgments

We want to thank the 3D Nucleome subgroup in the ENCODE consortium for discussion and early data processing. KKY acknowledges Shaoke Lou, Anurag Sethi, Joel Rozowsky and Arif Harmanci for feedback and discussion. KKY acknowledges Timur Galeev and Jonathan Warrell for critical reading on an earlier version of the manuscript. KKY acknowledges Nezar Abdennur from the Mirny lab for useful insights. This work was supported by the HPC facilities operated by, and the staff of, the Yale Center for Research Computing.

Author Contributions. Conceived, designed and performed the study: KKY, with input from MG. Wrote the paper: KKY, MG.

Funding. We thank the support by NIH award U41 HG007000.

Competing Interests. The authors declare no conflict of interest.

References

1. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. 2013;14:390–403.

2. Risca VI, Greenleaf WJ. Unraveling the 3D genome: genomics tools for multiscale exploration. Trends Genet. 2015;31:357–72.

3. Rowley MJ, Corces VG. The three-dimensional genome: principles and roles of long-distance interactions. Curr. Opin. Cell Biol. 2016;40:8–14.

4. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat. Rev. Genet. 2016;17:661–78.

5. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009;326:289–93.

6. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. 2011;30:90–8.

7. Fullwood MJ, Ruan Y. ChIP-based methods for the identification of long-range chromatin interactions. J. Cell. Biochem. 2009;107:30–9.

8. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.

9. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell. 2012;148:458–72.

10. Dekker J, Heard E. Structural and functional diversity of Topologically Associating Domains. FEBS Lett. 2015;589:2877–84.

11. Valton A-L, Dekker J. TAD disruption as oncogenic driver. Curr. Opin. Genet. Dev. 2016;36:34–40.

12. Lupiáñez DG, Spielmann M, Mundlos S. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. Trends Genet. 2016;32:225–37.

13. Newman MEJ. Modularity and Community Structure in Networks. Proc. Natl. Acad. Sci. 2006;103:8577–82.

14. Fortunato S, Barthélemy M. Resolution limit in community detection. Proc. Natl. Acad. Sci. 2007;104:36–41.

15. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008;2008:P10008.

16. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, et al. Comparative analysis of regulatory information and circuits across distant species. Nature. 2014;512:453–6.

17. Mourad R, Cuvier O. Computational Identification of Genomic Features That Influence 3D Chromatin Domain Formation. PLOS Comput Biol. 2016;12:e1004908.

18. Huang J, Marco E, Pinello L, Yuan G-C. Predicting chromatin organization using histone marks. Genome Biol. 2015;16:162.

19. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518:360–4.

20. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, et al. Topologically associating domains are stable units of replication-timing regulation. Nature. 2014;515:402–5.

21. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–8.

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

23. Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. Bioinformatics. 2015;btv485.

24. Malik LI, Patro R. Rich chromatin structure prediction from Hi-C data. bioRxiv. 2015;32953.

25. Ong C-T, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. 2014;15:234–46.

26. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell. 2015;163:1-17.

27. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014;159:1665–80.

28. Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell. 2015;162:108–19.

29. Bouwman BA, Laat W de. Getting the genome in shape: the formation of loops, domains and compartments. Genome Biol. 2015;16:154.

30. Harmanci A, Rozowsky J, Gerstein M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. Genome Biol. 2014;15:474.

31. Ay F, Bailey TL, Noble WS. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. 2014;24:999–1011.

32. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods. 2012;9:999–1003.

33. Fotuhi Siahpirani A, Ay F, Roy S. A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions. Genome Biol. 2016;17:114.

34. Dai C, Li W, Tjong H, Hao S, Zhou Y, Li Q, et al. Mining 3D genome structure populations identifies major factors governing the stability of regulatory communities. Nat. Commun. 2016;7:11549.

35. Ay F, Noble WS. Analysis methods for studying the 3D architecture of the genome. Genome Biol. 2015;16:183.

36. Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics. 2014;30:i386–92.

37. Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol. Syst. Biol. 2015;11:852–852.

38. Rajapakse I, Scalzo D, Tapscott SJ, Kosak ST, Groudine M. Networking the nucleus. Mol. Syst. Biol. 2010. 6: 395.

39. Kruse K, Sewitz S, Babu MM. A complex network framework for unbiased statistical analyses of DNA–DNA contact maps. Nucleic Acids Res. 2013;41:701–10.

40. Yan K-K, Wang D, Rozowsky J, Zheng H, Cheng C, Gerstein M. OrthoClust: an orthology-based network framework for clustering data across multiple species. Genome Biol. 2014;15:R100.

41. Barutcu et al_Genome Biology_2015_Chromatin interaction analysis reveals changes in small chromosome and telomere.pdf.

42. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16:259.

43. Akdemir KC, Chin L. HiCPlotter integrates genomic data with interaction matrices. Genome Biol. 2015;16:198.

44. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature. 2015;523:240–4.

45. Nussinov R, Pieczenik G, Griggs J, Kleitman D. Algorithms for Loop Matchings. SIAM J. Appl. Math. 1978;35:68–82.

Figure Captions

Figure 1: Overview of MrTADFinder. The input of MrTADFinder is an intra-chromosomal contact map *W*. A null model *E* is obtained from W. Given a particular resolution γ; the chromosome is partitioned probabilistically in a way such that the objective function Q is maximized. A boundary score is defined after multiple trials for all adjacent bins. Adjacent bins that are robustly assigned to two different TADs form a consensus boundary. The output of MrTADFinder is a set of consensus domains bound by the consensus domains.

Figure 2. Identification of TADs in multiple resolutions. A) A part of the contact map of the chromosome 10 in hES cell. The greenish triangles below represent TADs called by MrTADFinder in three different resolutions. The TADs called agree well visually with the contact map. The blue triangles and red triangles represent TADs called in human ES cells and human IMR90 cells respectively as reported in [8]. The distribution of peaks for a variety of chromatin features is aligned below. B) The size of TADs called in different resolutions. The median TADs size decreases from 3 Mbp to 300 kbp as the resolution increases from 0.75 to 3.5. C) The number of TADs increases as the resolution increases. When γ=2.25, there are about 2600 TADs in hES cells with a median size of roughly 1Mb. The median size goes down to 300kb when the resolution increases to 3.5. The number of TADs identified in [8] is marked by the arrow. D) Comparing TADs called by MrTADFinder with TADs called in [8]. Two algorithms agree the most in a particular resolution ($\gamma \approx 2.875$).

Figure 3. Boundary signatures in different resolutions. A) Histone modifications near the TAD boundary regions obtained in various resolutions. The peak density is obtained by counting the number of peaks in every 40kb bin, and normalized by a null model in which peaks are randomly distributed. B) Different histone marks show different levels of enrichment near TAD boundaries

at different resolutions. Despite a general decreasing trend, the signal of certain marks likes H3K27me3 remains flat until a very high resolution. C) Enrichment of HOT (high-occupancy target) and XOT (extreme-occupancy target) regions near TAD boundaries in hES cell. Boundaries are identified by MrTADFinder at a resolution $\gamma = 2.75$. The y-axis is normalized by a null model that peaks are randomly distributed in along the chromosome. D) The most influential factors responsible for TAD boundaries formation at different resolutions. Factors with a positive coefficient have a direct effect on border establishment or maintenance, whereas factors like MYC has a negative effect. The factors are sorted by corresponding P-values and only the significant factors are displayed.

Figure 4 Mutational burdens across TAD boundaries. The 3 clusters of boundary regions exhibit distinct patterns in terms of mutational burden. For blue and red clusters, the area marks the first and the third quartiles. For the green cluster, only the mean values at different positions are shown for clarity. The inset shows the average Repli-seq signal for the red and blue clusters.

Figure 5: Enrichment of CTCF peaks near TAD boundaries at two different resolutions. The red line shows the same analysis using TADs reported in [8].

Software availability

The source code can be downloaded at https://github.com/gersteinlab/MrTADFinder.

Supporting Information

Figure S1: Dependence of contact frequency and genomic distance. The analysis was performed using the contact map of the chromosome 1 of MCF7, binned in 250kb sized bins. The red line

$f(d)$ is the average contact frequency as a function of distance d obtained by smoothing. The green line shows a power-law function d^{-1}.

Figure S2. Effective coverage κ_i^* of loci is highly correlated with the coverage c_i.

Figure S3. Using transcription factors binding signals for predicting TAD boundaries. For each resolution, a logistic regression model based on transcription factors binding signals was trained to classify the TAD boundaries versus a set of random boundaries. The error bars were estimated by repeating the analysis using an ensemble of random boundaries. The performance (AUC and ACC) decreases as the resolution increases.

Figure S4 (A) TAD boundary regions (±600kb of boundary) of MCF7 (chromosome 10) are clustered based on the position distribution of mutations along the region. The regions are clustered into 3 groups (blue, red, green). (B) The 3 clusters of boundary regions exhibit distinct patterns regarding mutational burden.

Figure S5. The relationship between TADs and DNA replication timing. TADs are identified for IMR90 using different resolutions. Signals of Repli-seq data in various stages of a cell cycle and a part of the contact map of the chromosome 10 are displayed. The TADs match visually well with the replication timing signals. The middle TAD identified in $\gamma = 1$ does not replicate at S1, its sub-units identified in $\gamma = 1.25$ replicate in S2 and S3.as shown by the peaks in the Repli-seq signal.

Figure S6. Enrichment of CTCF peaks near TAD boundaries at two different resolutions. The red line shows the same analysis using TADs reported in [8]. This figure is an extension of Figure 5.

Figure S7. Robustness of MrTADFinder. Histogram for pairs of independently called TADs. Using the default parameters (10 trials of the modified Louvain algorithm and a cut-off of 0.9), the normalized mutual information between two sets of called domains agrees extremely well (nMI=0.99).

Figure S8: Identifying TADs by dynamic programming. The optimal value of Q for a chromosome segment running from i to j is stored in M_{ij}. The values of all elements in M can be enumerated using dynamic programming, starting from fragments of length 1 where $M_{ii} = Q_{ii}$. There are different ways to divide a fragment of length l (gray lines). Suppose the optimal way is marked by the red line, then $M_{1l} = M_{1r} + M_{rl}$.

Figure 1

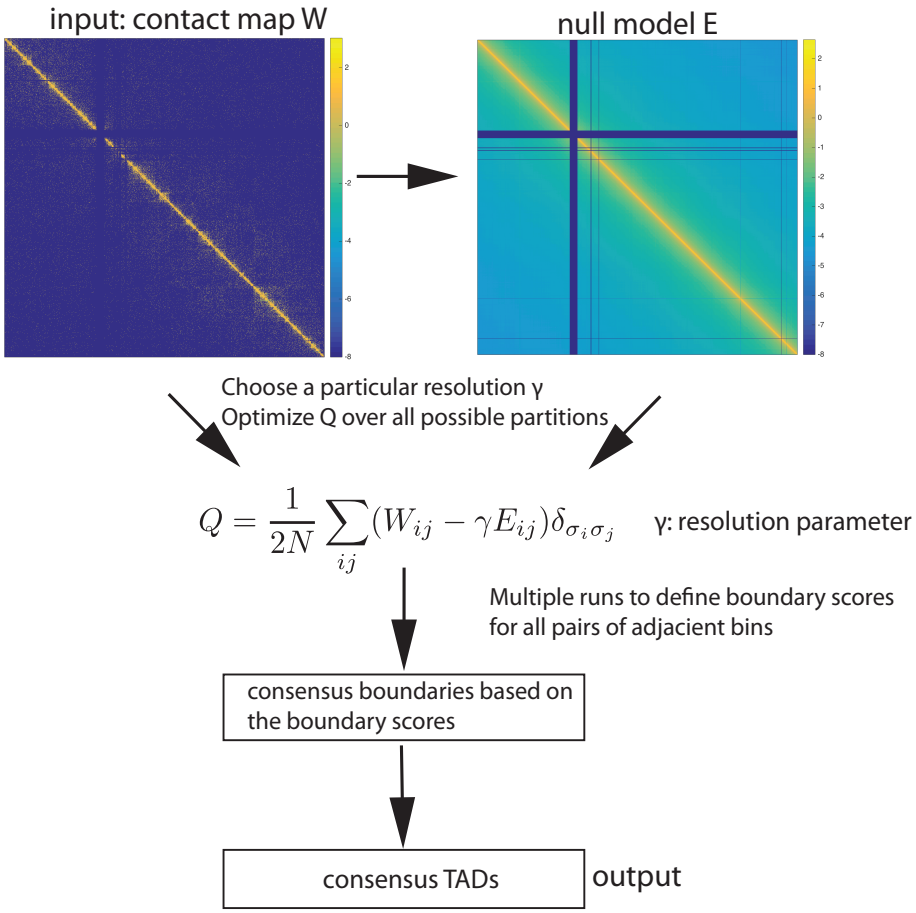

Figure 2

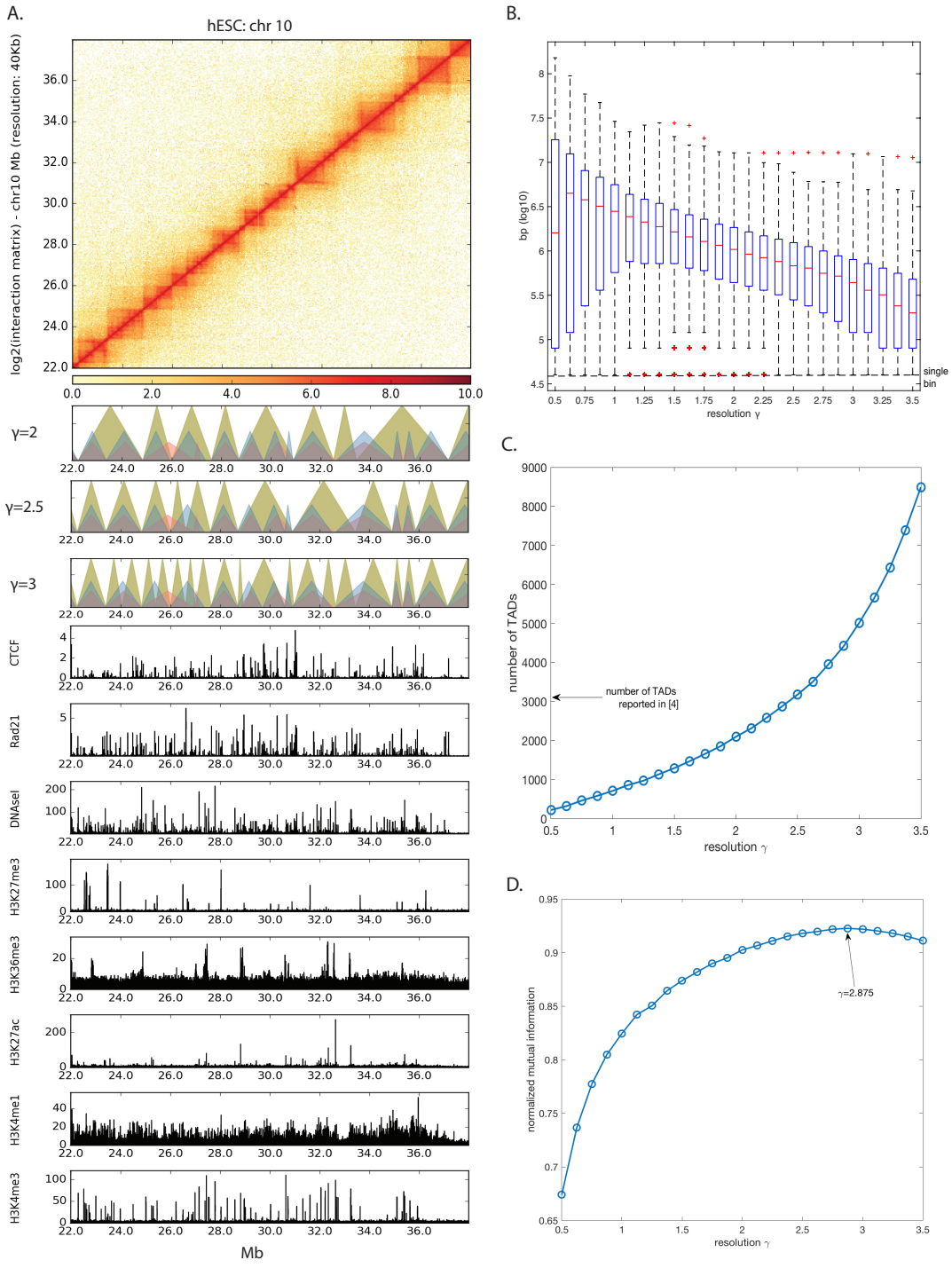

Figure 3

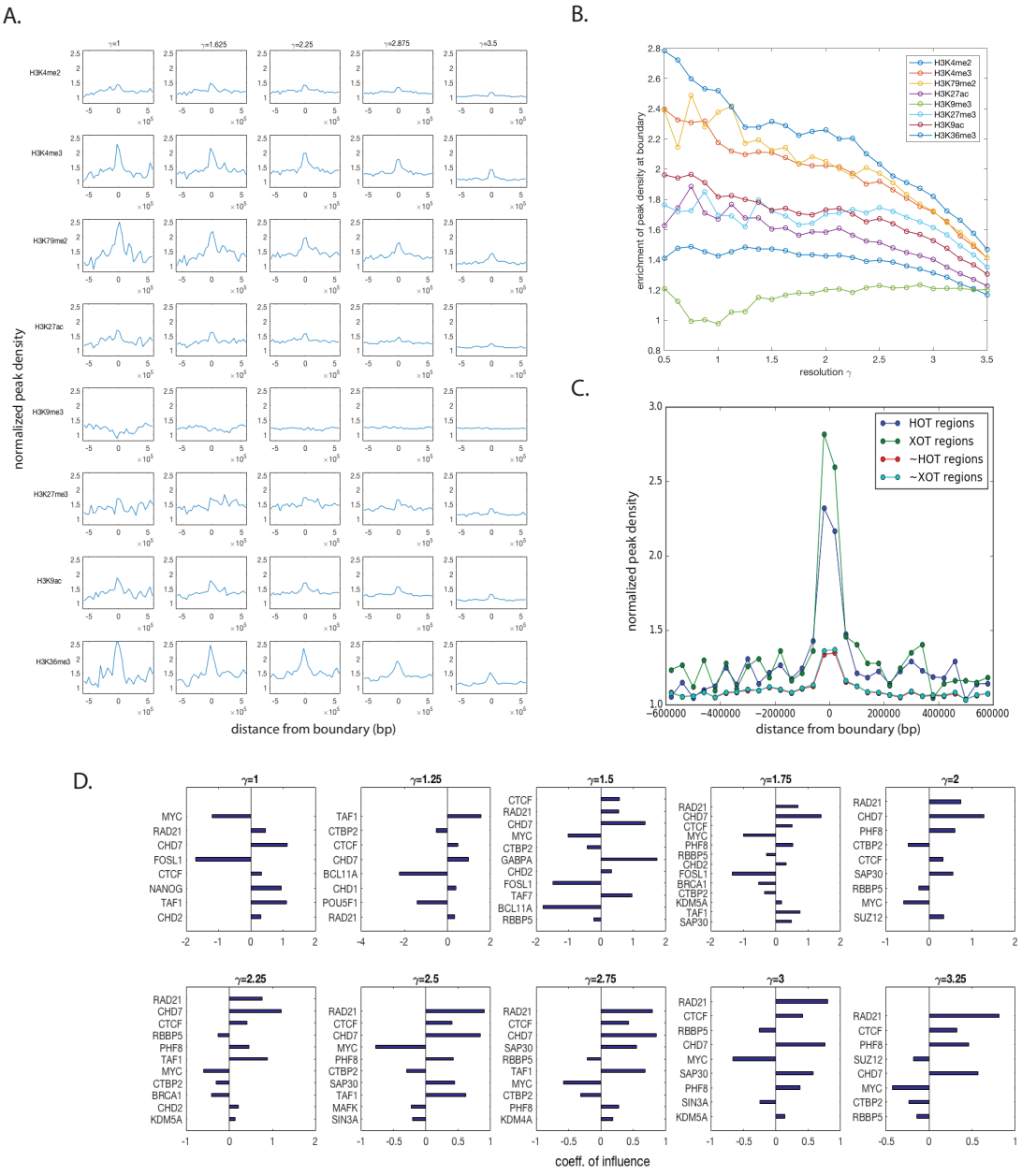

Figure4

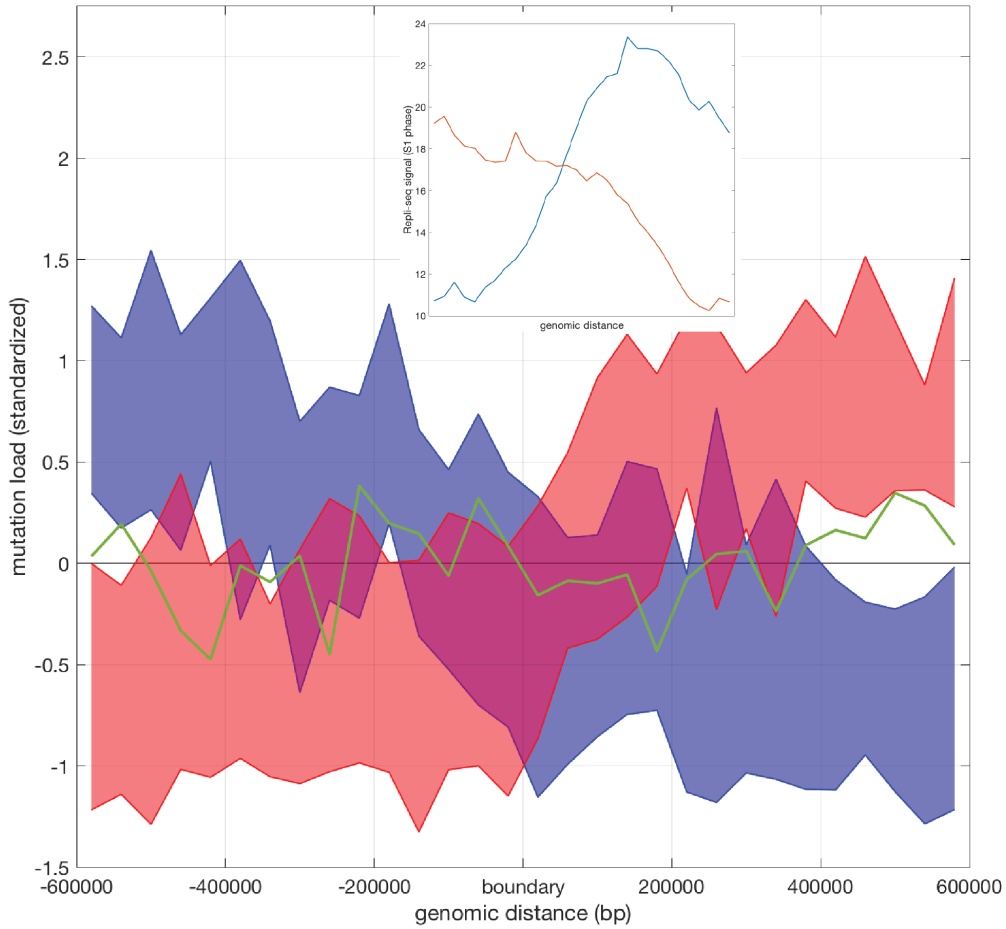

Figure 5

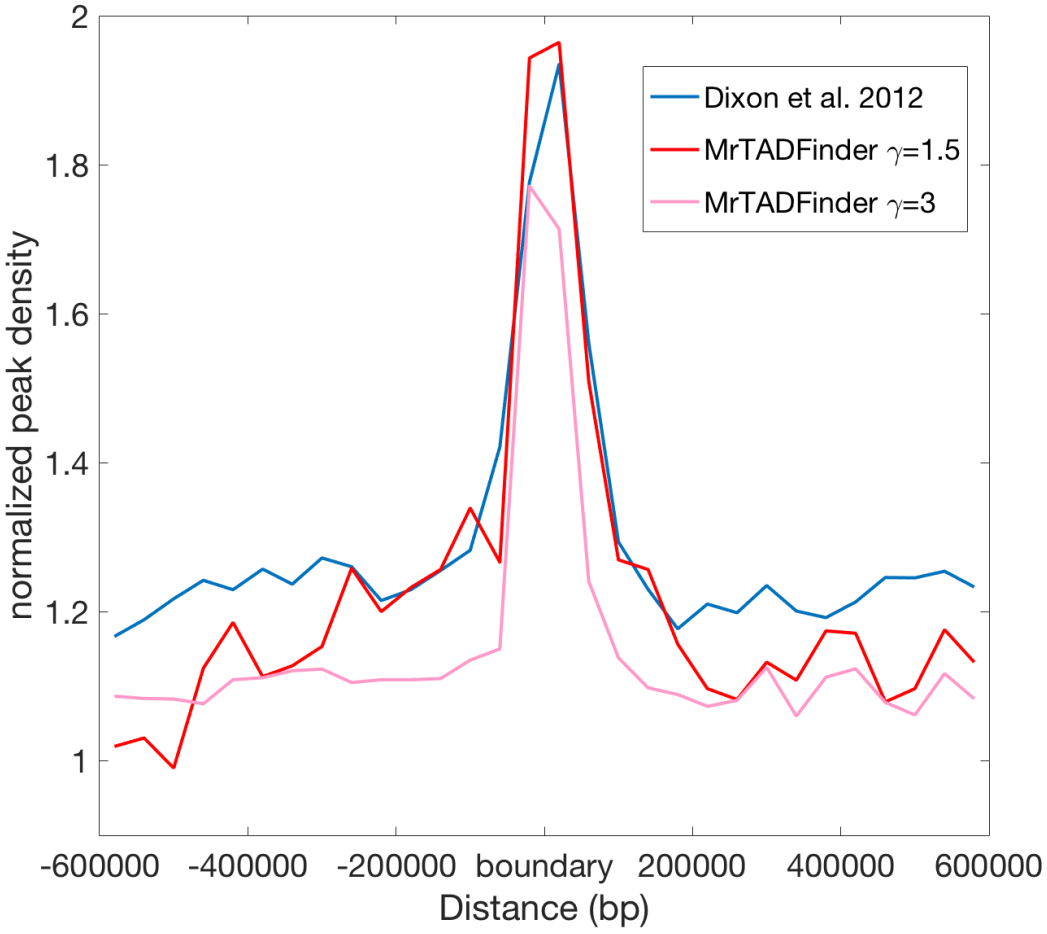

Figure S1

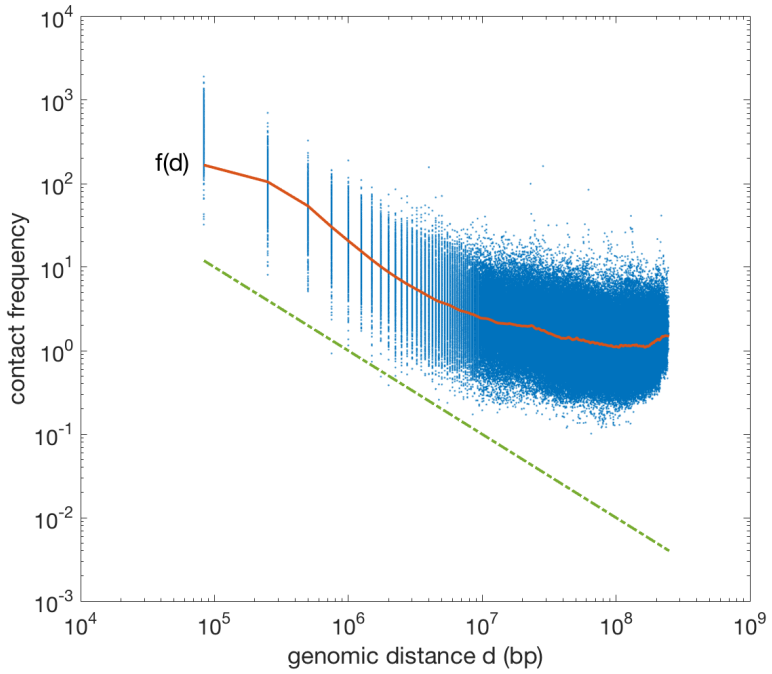

Figure S2

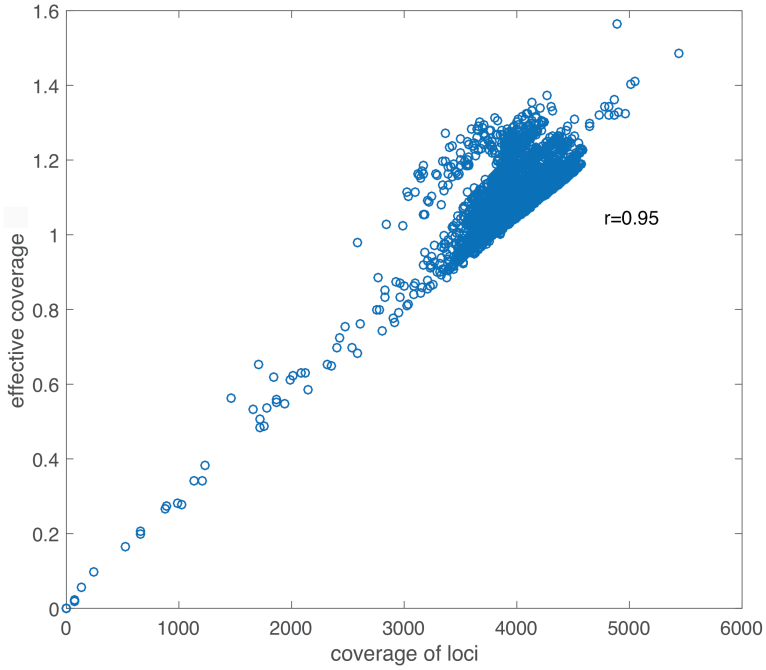

Figure S3

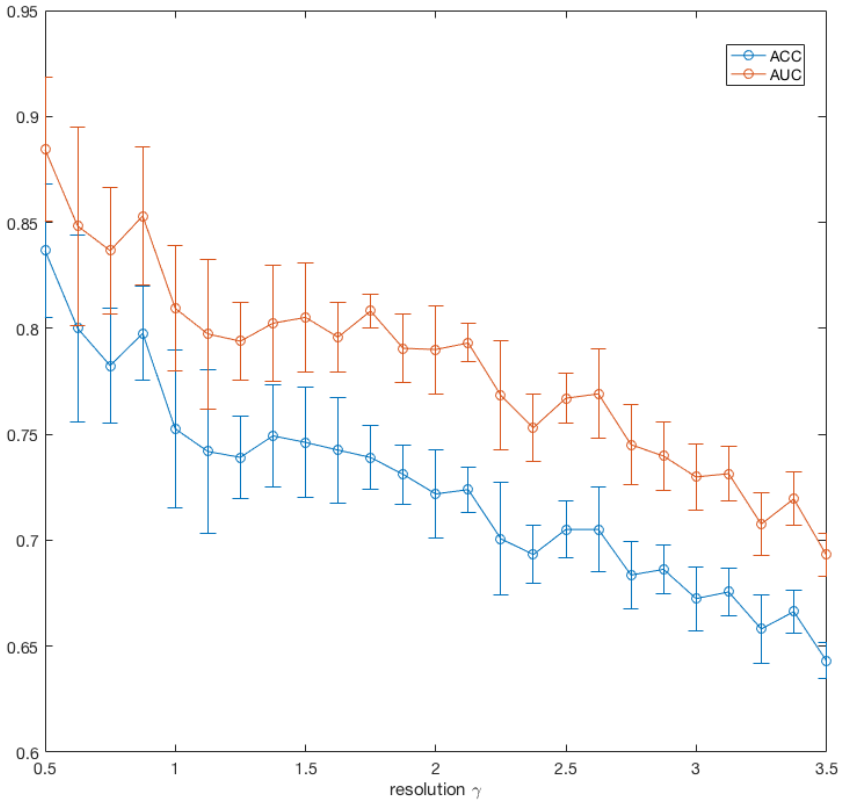

Figure S4

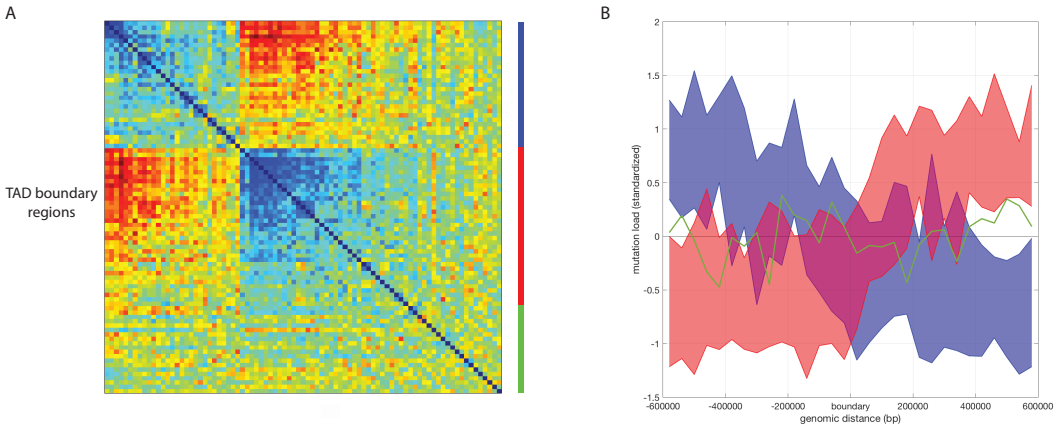

Figure S5

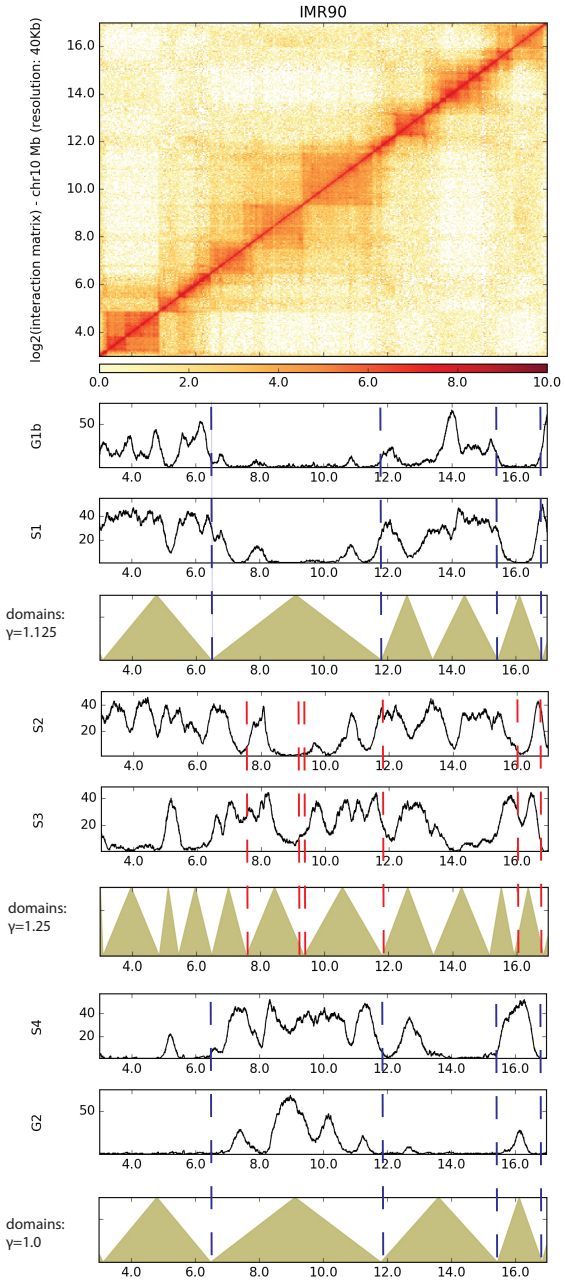

Figure S6

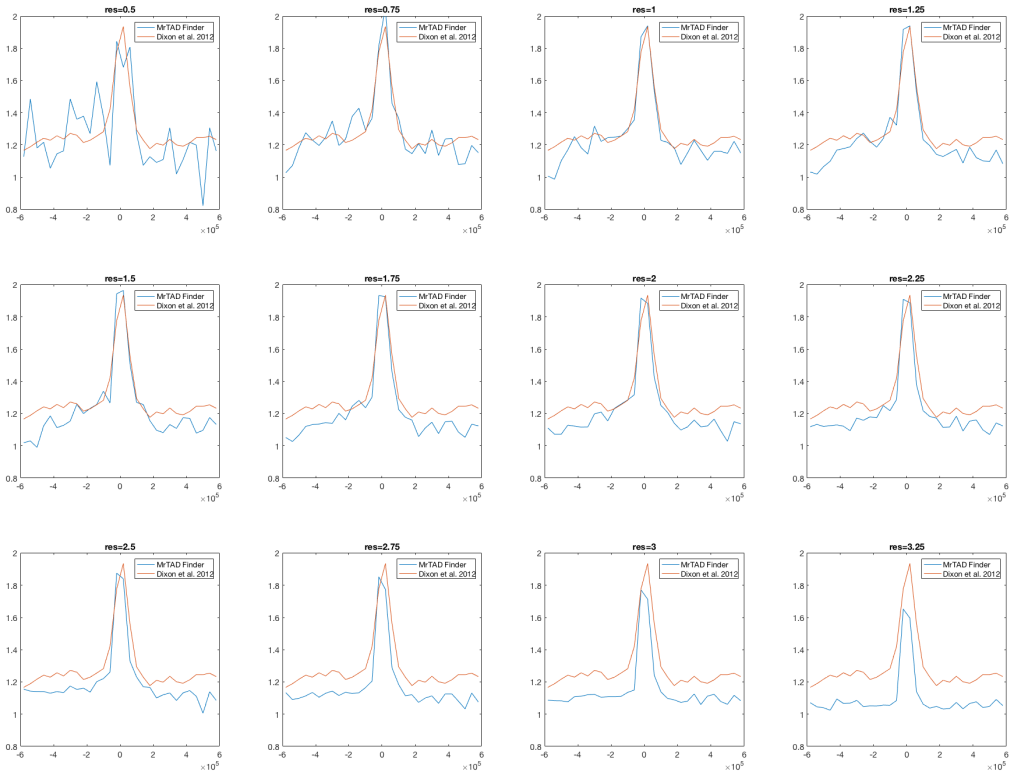

Figure S7

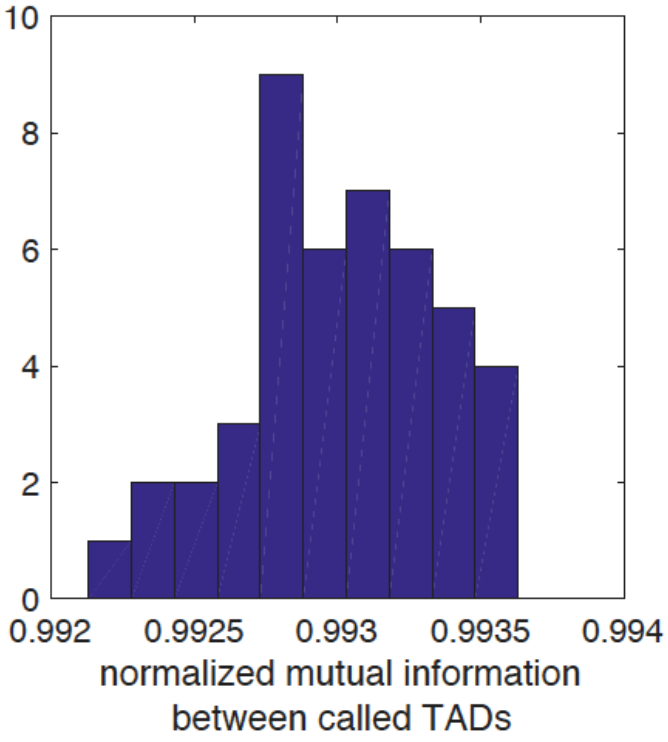

Figure S8