Identifying topologically associating domains (TADs) in multiple resolutions

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

1 Program in Computational Biology and Bioinformatics,

2 Department of Molecular Biophysics and Biochemistry, and

3 Department of Computer Science,

Yale University, New Haven, CT 06520

Abstract

The advance of next-generation sequencing experiments like Hi-C has revealed that eukaryotic genomes are organized into structural units called topological associating domains (TADs). Nevertheless, it is clear from visual examination of the so-called chromosomal contact map that there are rich sub-structures within TADs. The fine structures may correspond to a multiple scale organization. Here, by deriving a background model that takes into account the effect of the differences in coverage as well as the bias introduced by genomic distance, we present a novel algorithm, MrTAD Finder, to identify TADs in multiple resolutions. MrTAD Finder is based on the concept network modularity, and the resolution is tuned by a single parameter. In a low resolution, larger TADs are found whereas in a high resolution, smaller TADs are identified as the nucleome is viewed on a finer scale. We further investigated various chromatin features such as histone modifications and transcription factors binding within TADs and near TAD boundaries. We found that TADs in different resolutions have different chromatin signatures, and their boundaries are established by different transcription factors. The observations suggest chromatin structures overall have multiple scales.

Keywords

3D genome organization, Hi-C, Topological associating domains (TADs), network modularity

Background

The packing of a linear eukaryotic genome within a cell nucleus is tight and highly organized [1][2][3]. The spatial organization of the genome determines the accessibility of certain genomic regions and thus regulates effective gene expression. In such an intricate 3D structure, one of the most important features is the so-called topologically associating domain (TAD) [4][5]. TADs refer to genomic regions that are highly self-interacting, meaning loci within a region interact often but interactions between different regions are less frequent. Although TAD emerges as a fundamental structural unit of a genome, there are a lot of un-resolved issues [6][7][8]. In particular, it has been suggested the existence of alternative domains [9]. More recently, it has been reported that TADs exhibit a certain hierarchical organization, meaning a TAD can be decomposed into sub-TADs or several TADs can form a bigger domain [10][11]. These studies resonates with the observation that chromatin features like histone modifications have multiple length scales [12][13], and suggests that the 3D structure of a genome could be viewed in multiple resolutions.

By mapping chromatin proximity in a genome-wide level, the Hi-C technology has emerged as a powerful technique to understand the 3D genome organization [14][15]. Results of a typical Hi-C experiment are usually summarized by a so-called chromosomal contact map [15]. 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 loci originated from bins *i* and *j*. 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 (or bins) is quantified by a weighted edge. In this paper, we use the concept of network modularity to identify TADs, which are essentially blocks along the diagonal of a contact map. The identification of modules, also known as community detection, is an important problem in network studies [16]. In its simplest form, it concerns with whether nodes of a given network can be divided into non-overlapping groups such that connections within groups are relatively dense while those between groups are sparse. By viewing a Hi-C contact map as a network, the highly spatially localized TADs immediately resemble densely connecting modules. Motivated by the

resemblance, we developed a method to identify TADs in multiple resolutions called MrTAD Finder (Mr stands for multiple resolutions). MrTAD Finder goes beyond a direct adaptation of community detection by taking into account the effect of genomic distance that are specific in the context of genome organization. We applied MrTAD Finder to various Hi-C datasets, arriving at TADs in different resolutions. Interestingly, TADs in different resolutions exhibit different chromatin signatures.

Results

Network modularity and the identification of TADs

The decomposition of modules refers to the problem of partitioning nodes of a given network into non-overlapping groups such that connections within groups are relatively dense while those between groups are sparse. The identification of modules pretty much follows the same rationale. A chromosome of interest is divided into different domains, such that the frequency of contacts between loci within domains are dense while interactions between domains are sparse. MrTAD Finder approaches the problem using the method of modularity maximization [16]. The essence is the so-called modularity function Q which is proportional to the sum $\sum_{i,j}\left(W_{ij}-\gamma E_{ij}\right)\delta_{\sigma_i\sigma_j}$. Here, W is an intra-chromosomal binned contact map whereas E is a null model. The function is maximized over all possible partitions of bins into a set of domains σ, with a particular choice of the so-called resolution parameter γ (see Figure 1A for a workflow). At the heart of the algorithm is the null model E. In the context of network modularity, the most common one is the so-called configuration model, in which $E_{ij} \propto k_i k_j$, meaning the degrees of nodes (its number of connections) are fixed to match those of the observed network but edges are randomly rewired [16]. Nevertheless, this simple model cannot be directly applied in the context of domain identification. It is because unlike conventional graphs in which the spatial location of nodes is not important, bins in the chromosome form a continuous structure. 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. Taking all into account, given a binned contact map W, say in 40kb, we define a new modularity function for TADs identification, with $E_{ij} = c_i^* c_j^* f(|i-j|)$. Here c_i^* is an

unknown associated with bin i, f represents the average number of contacts as a function of distance (in the unit of bin size), which can be estimated by aggregating separately the matrix elements corresponding to bins differed by each possible separation $d = |i - j|$ (see methods).

Given W, the elements in matrix E can be numerically calculated. For instance, in hES cell, in a size of 40kb, we found that in a whole genome level, f can be well fitted by a power-law function $f \sim d^{-\lambda}$ (see Figure 1B). The decay of contact frequency with respect to the genomic distance we observed is consistent with previous estimations done with slightly different binning strategies [15]. By estimating f from W, c_i^* and thus all the elements of the matrix E could further be found by solving a set of non-linear equations via a matrix iteration scheme (see methods). Figure 1A shows a particular example of contact map W in the hES cell and the corresponding null model E. As expected, the null model E exhibits a gradual decrease of contact frequency away from the diagonal. Figure 1C shows the enrichment of observed contacts W with respect to the null E under a Poisson model. The significantly interacting loci are mostly located near the diagonal. They are potentially reflecting regulatory interactions such as promoter-enhancer contacts. In general, c_i^* can be interpreted as an effective coverage of bin i, which is analogous to the degree of a node in a network setting (Figure S1).

Identifying TADs in multiple resolutions

We then applied MrTAD Finder to analyze Hi-C data of hES cell from ref. [4]. 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 contact map. Of particular interest is the choice of γ. As shown in Figure 2A, when γ increases, a large TAD is broken into a few small TADs. On the other hand, large TADs merge together to form even larger TADs as the value of γ is lowered. Therefore γ is referred as the resolution parameter that capture the fine structures in domains organization. Statistically speaking, γ is essentially quantifying to the proportion of the expected counts as compared to the observed counts. As γ increases, only elements close to the diagonal contribute positively to the modularity 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=2.25, there are about 2600 TADs in hES cells with a median size of roughly 1Mb. We then further compared the TADs identified at different resolutions by MrTAD Finder with TADs previously identified in ref. [4]. As quantified by the normalized mutual information (see methods for details), TADs identified by MrTAD Finder best match with TADs identified in ref. [4] when the resolution parameter is 2.9. In general, unless the resolution is sufficiently small ($\gamma < 1.5$), TADs called by MrTADFinder are quite consistent with the TADs called in in ref. [4] (see Figure 2D). Nevertheless, the introduction of the resolution parameter γ has broadened previous work on domains identification in the sense the algorithm used in ref. [4] focuses on a particular resolution instead. Furthermore, it is worthwhile to mention that a higher fraction of the genome is assigned to various TADs as compared to the case in ref. [4] (93% average over different resolutions as compared to 86%, see Figure S2).

Boundary signatures of TADs identified in different resolutions

We further investigated the TAD boundaries identified in different resolutions. First of all, we found the boundary signatures are generally consistent with the observations previously reported [4], 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, by identifying TADs in a variety of resolutions, we found the previously observed signatures change with respect to resolutions (Figure 3B, S3). 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 3C). Enrichment of histone marks like H3K36me3, 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 3C).

Apart from histone modifications, it is well known that certain transcription factors like CTCF plays an important role in the formation of TAD boundaries [17]. Though factors like CTCF have a tendency to bind close to the boundaries (Figure S4), it is not clear whether the enrichment is a reflection of direct involvement in boundary formation. In fact, we found that many

of the so-called HOT regions [18], genomic regions that are bound by extensive amount of transcription factors, are located very close to TAD boundaries (Figure 3D). To examine which factors are responsible for establishing the domain border in different resolutions, we employed a logistic regression model recently proposed by [19]. The model quantifies explicitly the influence of about 60 factors in classifying a set of borders identified by MrTAD Finder versus a set of random boundaries (see methods). In general, factors that are responsible for border formation are quite consistent across different resolutions (Figure 3E and S5). For instance, factors like CTCF, Rad21 and CHD7 are direct driver of border establishment and maintenance, whereas factors like MYC have a consistent negative effect. Overall, genomic features like the binding signals of a variety of transcription factors are quite successful in predicting the structural organization of chromatin (AUC=0.81, Figure S6) [20].

Chromatin signatures within TADs in different resolutions

Apart from the boundaries, we investigated various chromatin features along TADs in various resolutions (Figure 4A and S7). TADs identified in different resolutions are essentially different ways to segment a chromosome. By examining the location of peaks along TADs, we found histone marks like H3K4me3, H3K36me3, H3K27me3, H3K27ac, H3K9ac, H3K79me2 etc have peaks clustered near the two ends. The observation is generally true for TADs in all resolutions. Nevertheless, for low resolution like $\gamma = 0.5$, many histone marks are enriched at the middle, suggesting that adjacent TADs are merged. We further examined the peak density of different histone marks with respect to γ (Figure 4B). At the boundary regions, H3K4me3 has the highest peak density in low resolution, whereas marks like H3K36me3 and H3K27ac have the highest peak density in medium resolutions. Different characteristic marks are observed in the middle regions. We then looked at how annotated genomic regions are located along TADs. Figure 4C shows the fold enrichment of various Segway annotations [21] along TADs at different resolutions. Labels like Dnase3, Gen3 are in general less enriched at high resolutions, meaning such features are likely to be absent in the boundaries of high-resolution TADs. Nevertheless, there are labels like TSS, PromP that are more consistent with respect to different resolutions.

Based on the respective annotation genomic regions, we further divided TADs into different classes (Figure 4D), characterized by different expression levels (Figure 4E). The classes further echoed with annotation, for instance, the lowly expressed TADs are enriched with heterochromatin labels like QUI and CON (Figure 4F) [22]. We repeated the analysis using TADs in a higher resolution, but found that the clustering is quite stable (Figure S8).

What is the meaning of the multiple levels of segmentation? As TADs are basic units of chromosome organization, it is believed that long-range regulatory interactions are likely to be within TADs instead of between TADs. We identified the statistical significant contacts based on the Hi-C data [23] and examined how many of them appear within TADs. Of course, as the resolution increases, the absolute number of links within TADs decreases. Nevertheless, with respect to a null model in which TADs are shuffled, the enrichment of links within TADs actually increases for most chromosomes. In particular, in a few chromosomes, a characteristic resolution that maximizes the enrichment of within-TAD interactions exists (Figure S9). Besides its regulatory role, it has been demonstrated that TADs are stable units in the process of DNA replication [24]. Overlaying data from Hi-C and Repli-Seq experiment in IMR90 suggests that domains in different resolutions could correspond to sub-units that are replicated in a shorter time-scale (Figure S10).

Discussion

Here, we have introduced an intuitive algorithm to identify TADs based on Hi-C data. By introducing a single continuous parameter γ, we are able to further examine the rich structures or sub-structures stored in contact maps, and explore the organization of genome in multiple resolutions. The concept of resolution could further integrate the observed structures in different length scales such as compartment, domains, sub-domains etc.

A few methods have already been proposed to identify TADs from Hi-C data [25]. One of the earliest methods is based on a 1D “directionality index” that captures whether contacts have an upstream/downstream bias [4], and later bias is exploited by the so-called arrowhead algorithm [26]. However, as intra-chromosomal interactions depend heavily on the distance

between interacting loci, it is important to have a proper normalization [25]. MrTAD Finder provides such a background model via matrix iteration. Furthermore, in terms of a more detailed picture of domains organization, the idea of continuous resolution used in MrTAD Finder is distinct in comparison with algorithms based on a bottom-up approach. Such a approach results in higher order domains that are organized in a tree structure [10][11].

MrTAD Finder is based on the idea of network modularity. Recent studies suggested a similar weighted network framework for Hi-C data and identified modules via spectral or other clustering methods [27][28]. Although a network perspective of chromosomal interactions has previously been proposed [29][30], a lot of widely studied concepts in networks have rarely been explored in the context of chromosomal organization. By facilitating the application of a variety of graph-theoretical tools, we believe that network algorithms will be useful for future analysis on the spatial organization of genome.

Materials and methods

Hi-C data and their pre-processing

The Hi-C data of human ES cells and IMR90 cells were generated by ref. [4], which was downloaded from GEO with accession number GSE35156. Raw reads were processed using Hi-C Pro [31], arriving at contact matrices in various bin sizes. In all analysis, the whole-genome contact map were 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 [33].

Chromatin Data

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

Expected null model for an observed intra-chromosomal contact map

Given an intra-chromosomal contact map W, the expected null model E is defined as $E_{ij} =$

$c_i^* c_j^* f(|i-j|)$. f, the average number of contacts as a function of distance d, is assumed to be a power-law function $d^{-\lambda}$. For each possible value of d, the corresponding matrix elements were aggregated. The power-law exponent γ was estimated using a maximum likelihood approach. As a null model, the resultant E matrix satisfies a set of constraints, namely

$$\sum_j E_{ij} = \sum_j W_{ij} = c_i \quad \forall i,$$

$$\sum_{ij} E_{ij} = \sum_{ij} W_{ij} = 2N.$$

The first equation means that for each bin i, the coverage c_i defined in the null model is the same as the coverage defined in the observed map. The second equation is a direct consequence of the first equation, where N is the number of reads mapped in the chromosome. As f can be estimated from the observed W, the only unknowns c_i^* can be solved numerically by an iterative matrix procedure. The procedure can be regarded as a generalization of a class of matrix balancing methods commonly used for normalizing Hi-C matrices [32][25].

The partition of domains

MrTAD Finder divides a chromosome into domains using the method of modularity maximization. The essence is the so-called modularity function Q, defined as

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

Here, W is an intra-chromosomal binned contact map, E is a null model, N is normalization constant and γ is the so-called resolution parameter. The value of the Kronecker data $\delta_{\sigma_i\sigma_j}$ equals one if nodes i and j have the same label and zero otherwise, meaning only pairs of bins within the same domain are summed. To maximize Q, MrTAD Finder employs a modified version of the widely used Louvain algorithm [34]. In a nutshell, the algorithm consists of two passes. The algorithm starts as every bin has its own label at the beginning. In the first pass, for each bin, the label was updated by either choosing the label of one of its neighboring bins or to remain unchanged based on whether or not the value of Q will be increased. When no more update is possible, the second pass is performed such that the adjacent bins with the same label were

merged to form a new contact matrix. The two passes are repeated iteratively until there is no increase of modularity is possible. The result is essentially a particular way to partition the chromosome, i.e. to identify a particular set of boundaries. As the result of the Louvain algorithm in general depends on the order of updates, multiple runs were performed to ensure robustness of the domains identified. A set of consistent boundaries was defined based on the so-called boundary score (the fraction of runs a location is identified as a boundary). By default, a cut-off of 0.9 was used (i.e. a boundary between two adjacent bins is defined as confident only if the two bins are called to belong to two different TADs in at least 9 out of 10 trials). TADs were defined as the regions partitioned by the set of consistent boundaries. Practically, if 10 trials are used to define the boundary score and thus the consensus domains, the results are highly robust (see Figure S11).

Quantifying the consistency between two sets of TADs

Given two sets of TADs, say in different cell lines, or called by different algorithm s, or called at two different resolutions, the consistency is quantified by the so-called normalized mutual information. Suppose X and Y are two random variables whose values x_i and y_i represent the TAD labels of bin i. The normalized mutual information MI_{norm} is defined as $MI_{norm} = \frac{2I(X;Y)}{H(X)+H(Y)}$, where H(X), H(Y) are the entropy of X and Y, and I(X;Y) is the mutual information quantifying to what extent the partition of TADs in X give the information on the partition of TADs in Y. 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.

Boundary signatures of chromatin features

Given the location of binding peaks of a transcription factors 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 uniformly 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 MrTAD Finder 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

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

here X represents all features, $\boldsymbol{\beta}$ is vector determining the coefficients of influence of 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.

Chromatin features within TADs

The number of peaks for each histone mark was counted in every 40kb bin along a TAD, and normalized by the average number of peaks in a bin as if peaks are uniformly distributed. For analysis in Figure 4A, because TADs are different in length, the peak density profiles are rescaled using the MATLAB function `imresize`. For the analysis based on Segway, we calculated the proportion of each class of annotation in each TAD, and performed normalization with respect to the corresponding proportion in the whole genome. Significant contacts are estimated using the tool Fit-Hi-C [23], in which FDR is set to be 0.01.

Software availability

Source code of MrTADFinder written in Julia can be downloaded from https://github.com/quantum-man/MrTADFinder.

Competing of Interest

The authors declare that they have no competing interests.

Authors' contributions

Acknowledgements

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. 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.

5. 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.

6. Cubeñas-Potts C, Corces VG. Topologically Associating Domains: An invariant framework or a dynamic scaffold? Nucleus. 2015;6:430–4.

7. Sexton T, Cavalli G. The Role of Chromosome Domains in Shaping the Functional Genome. Cell. 2015;160:1049–59.

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

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

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

11. 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.

12. 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.

13. Larson JL, Huttenhower C, Quackenbush J, Yuan G-C. A tiered hidden Markov model characterizes multi-scale chromatin states. Genomics. 2013;102:1–7.

14. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing Chromosome Conformation. Science. 2002;295:1306–11.

15. 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.

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

17. Moore BL, Aitken S, Semple CA. Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. Genome Biol. 2015;16:110.

18. 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.

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

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

21. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods. 2012;9:473–6.

22. Libbrecht MW, Ay F, Hoffman MM, Gilbert DM, Bilmes JA, Noble WS. Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. Genome Res. 2015;25:544–57.

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

24. 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.

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

26. 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.

27. 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.

28. 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.

29. Rajapakse I, Scalzo D, Tapscott SJ, Kosak ST, Groudine M. Networking the nucleus. Mol. Syst. Biol. [Internet]. 2010 [cited 2014 Apr 29];6. Available from: http://msb.embopress.org/content/6/1/395

30. 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.

31. 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.

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

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

Figure Legends

Figure 1. Workflow of MrTAD Finder

A) The input of MrTADFinder is an intra-chromosomal contact map W. A null model 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.

B) For a contact map W, given a particular distance d, interacting frequencies between loci that are separated by the distance d are aggregated. Loci that are close together on average have more contacts compared to loci that are far apart.

C) Enrichment of empirical contacts W with respect to the null model E under a Poisson model. For matrix element (i,j), a P-value is estimated by assuming W_{ij} is drawn from a Poisson distribution with mean E_{ij}. The significant interactions are located near the diagonal. The collection of significant interactions resembles the TAD structures.

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 by ref. [4].

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 called in ref. [4] is marked by the arrow.

D) Comparing TADs called by MrTADFinder with TADs called in ref. [4]. 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. The peak density is normalized by a null model in which peaks are uniformly distributed.

B) Histone modifications near the TAD boundary regions obtained in different resolutions.

C) Different histone marks show different level 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.

D) Enrichment of HOT (high-occupancy target) and XOT (extreme-occupancy target)regions near TAD boundary regions.

E) The most influential factors responsible for TAD boundary 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.

Figure 4. Chromatin signatures with TADs in different resolutions

A) Distribution of histone marks across TADs in different resolutions.

B) Peak density of various histone marks, near the TAD boundaries and near the middle regions, at different resolutions.

C) Distribution of various genome annotation labels along TADs at different resolutions. For each resolution, the resultant TADs are divided and scaled into 10 equal bins. Labels like Dnase are enriched near the boundary, but the enrichment decreases in high resolutions. Certain labels like TSS are more consistent in different resolutions.

D) Clustering of TADs ($\gamma \approx 1.0$) based on genome annotation, There are three classes of TADs (red, blue and green).

E) Average expression of the three classes of TADs.

F) Classes of TADs are signified by different labels: QUI (quiescent domains), CON ("constitutive heterochromatin), FAC ("facultative heterochromatin"), BRD (broad expression), SPC (specific expression).

Additional Files

Figure S1. Effective coverage c* the null model of MrTAD Finder. c* is highly correlated with the original coverage defined in raw Hi-C data.

Figure S2. Fraction of genome assigned to TADs with respect to resolution. In bin size of 40kb, over 92% of bins have TADs assigned. The red line shows the respective fraction in TADs called in Ref. [4] which does not incorporate the idea of resolution.

Figure S3. B) Histone modifications near the TAD boundary regions obtained in different resolutions (an expanded version of Figure 3B).

Figure S4. B) Peak density of CTCF near TAD boundaries obtained in different resolutions. The red line shows the same analysis using TADs called in Ref. [4].

Figure S5. The most influential factors responsible for TAD boundary formation at different resolutions (an expanded version of Figure 3E).

Figure S6. 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 S7. Distribution of histone marks across TADs in different resolutions (an expanded version of Figure 4A).

Figure S8. Clustering of TADs with $\gamma \approx 2.875$ based on genome annotation. Similar to Figure 4D, E, F, there are three classes of TADs (red, blue and green), characterized by different expression level and annotation labels: QUI (quiescent domains), CON (constitutive heterochromatin), FAC (facultative heterochromatin), BRD (broad expression), SPC (specific

expression). The order of labels in the heatmap is arranged so as to be the same as the one in Figure 4D.

Figure S9. Significant chromosomal links within TADs.
A) Fraction of significant chromosomal links within TADs in all chromosomes. The fraction decreases as resolution increases (the number of TADs increases). The analysis was repeated by shuffling TADs along the respective chromosomes (red).
B) Enrichment of significant links. The enrichment is the ratio between the fraction of significant links within real TADs versus the fraction of significant links within randomized TADs.

Figure S10. Relationship between TADs and DNA replication timing. TADs are called for IMR90 using different resolutions. Signals of Repli-Seq data in various stages of 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.

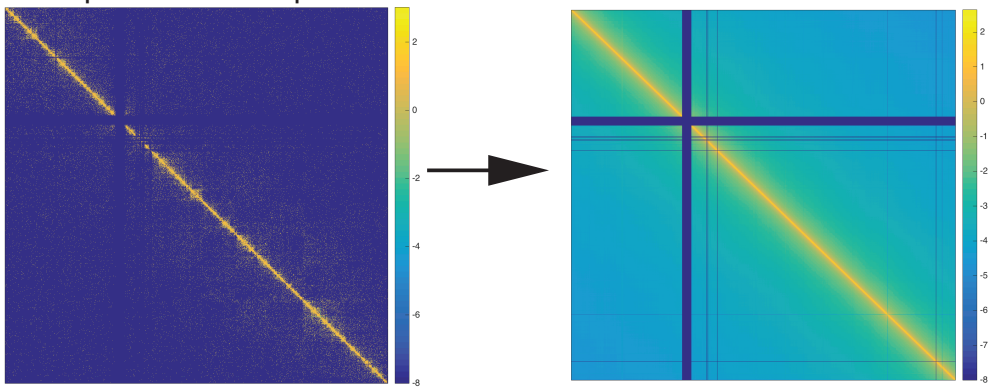
Figure S11. Robustness of domains. Using the default parameters (10 trials of the modified Louvain algorithm and a cut-off of 0.9), the normalized mutual information between two sets of called domains agrees extremely well (nMI=0.99).

A.

input: contact map W

null model E

$$E_{ij} = c_i^* c_j^* f(|i-j|)$$



Choose a particular resolution γ
Optimize Q over all possible partitions

$$Q = \frac{1}{2N} \sum_{ij} (W_{ij} - \gamma E_{ij}) \delta_{\sigma_i \sigma_j}$$

γ: resolution parameter

Multiple runs to define boundary scores for all adjacient bins

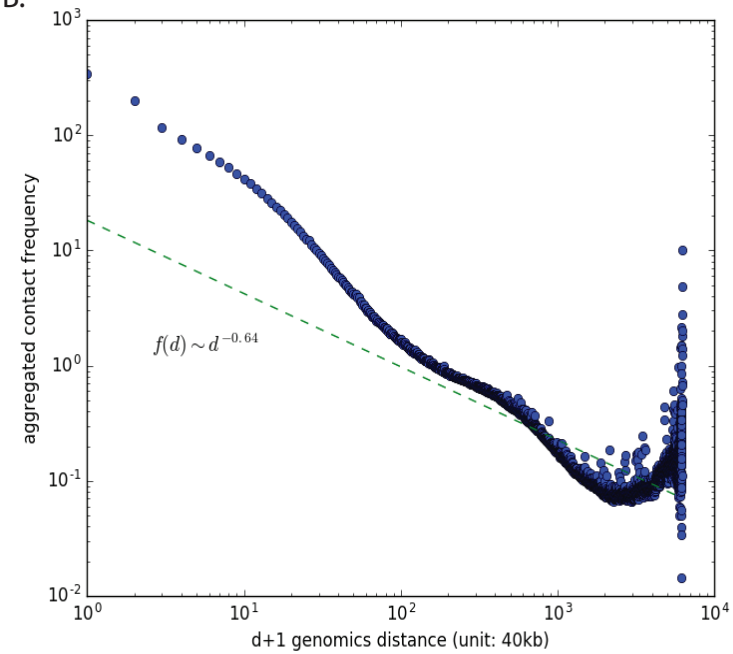
consensus boundaries based on the boundary scores

consensus domains

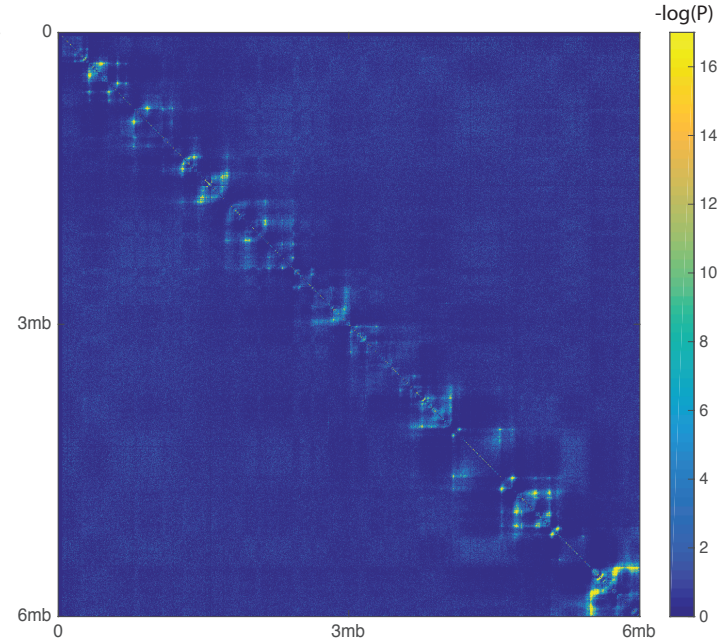
output: TADs

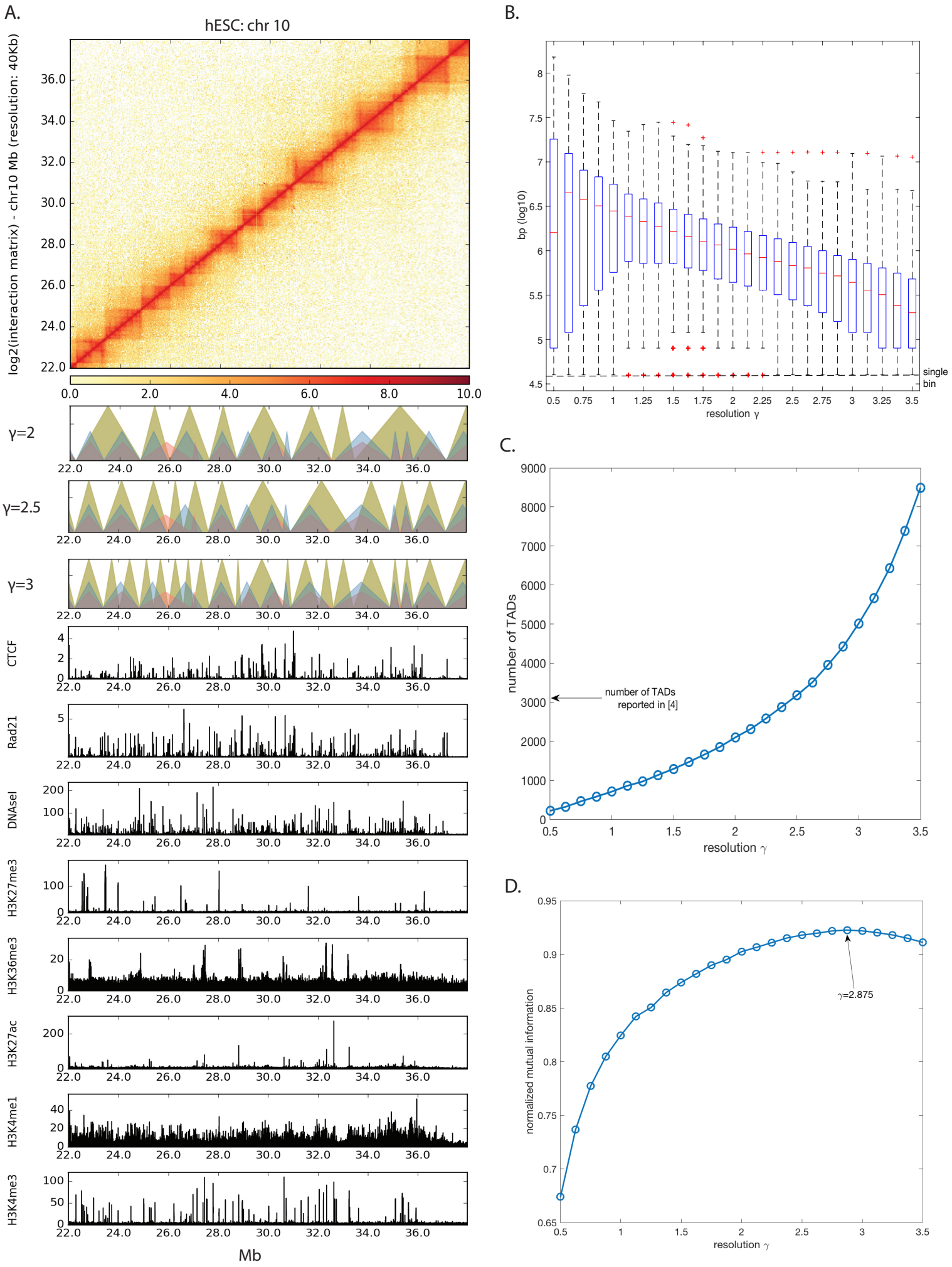
Row	chr	domain_st	domain_ed
1	"chr10"	40001	1120000
2	"chr10"	1120001	3240000
3	"chr10"	3240001	4840000
4	"chr10"	4840001	5680000
5	"chr10"	5680001	5760000
6	"chr10"	5760001	5920000
7	"chr10"	5920001	6000000
8	"chr10"	6000001	7560000
9	"chr10"	7560001	9360000
10	"chr10"	9360001	11520000

B.

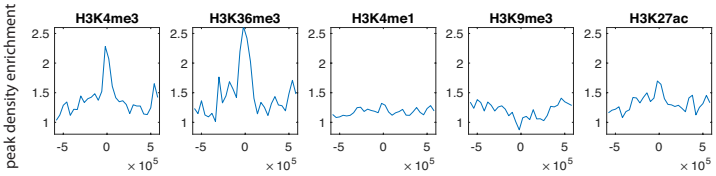


C.

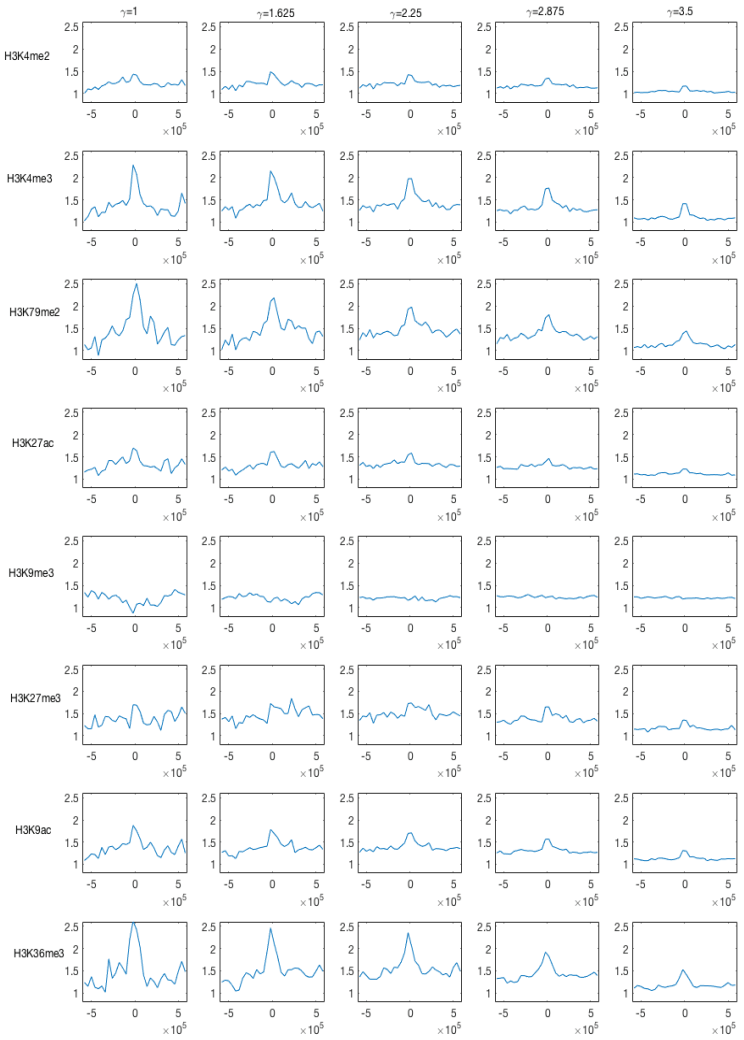



A.
hESC: chr 10
log2(interaction matrix) - chr10 Mb (resolution: 40Kb)
γ=2
γ=2.5
γ=3
CTCF
Rad21
DNAseI
H3K27me3
H3K36me3
H3K27ac
H3K4me1
H3K4me3
Mb
B.
bp (log10)
single bin
resolution γ
C.
number of TADs
number of TADs reported in [4]
resolution γ
D.
normalized mutual information
γ=2.875
resolution γ

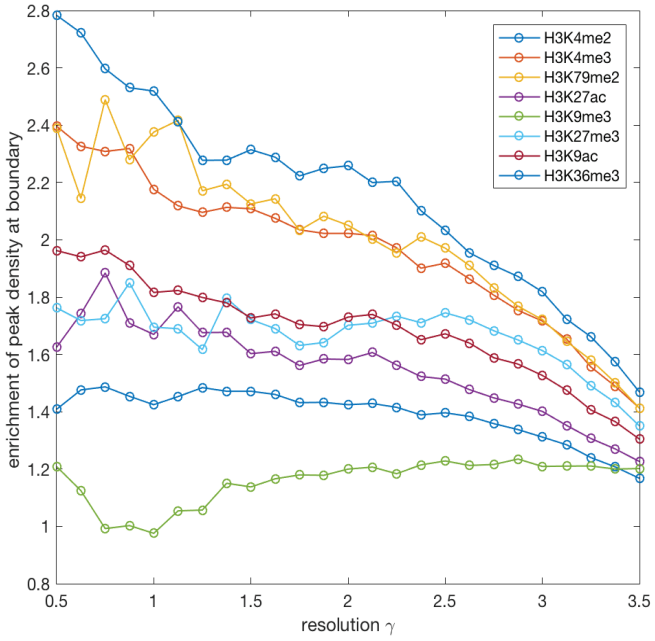
A.


peak density enrichment
H3K4me3
H3K36me3
H3K4me1
H3K9me3
H3K27ac

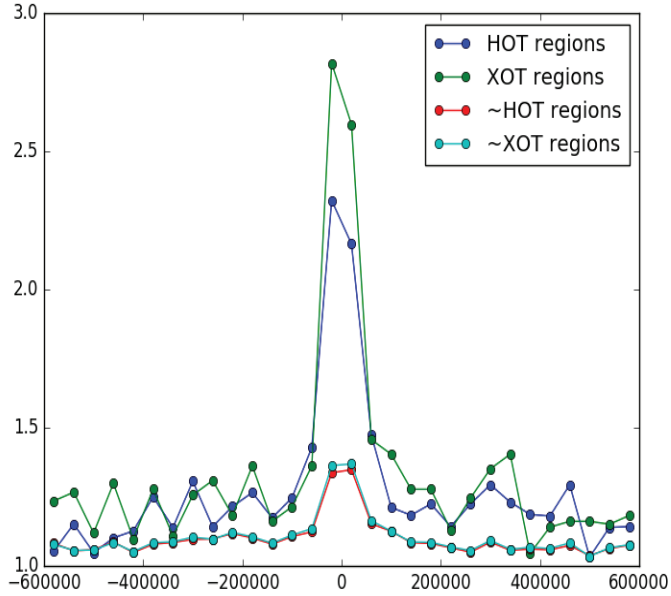
B.


γ=1
γ=1.625
γ=2.25
γ=2.875
γ=3.5
H3K4me2
H3K4me3
H3K79me2
H3K27ac
H3K9me3
H3K27me3
H3K9ac
H3K36me3

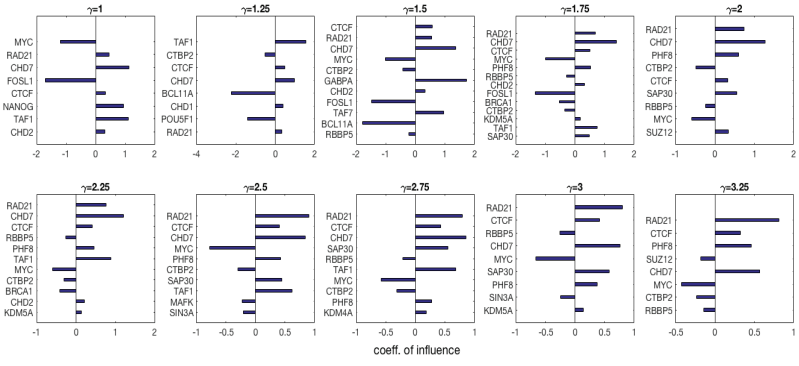
C.

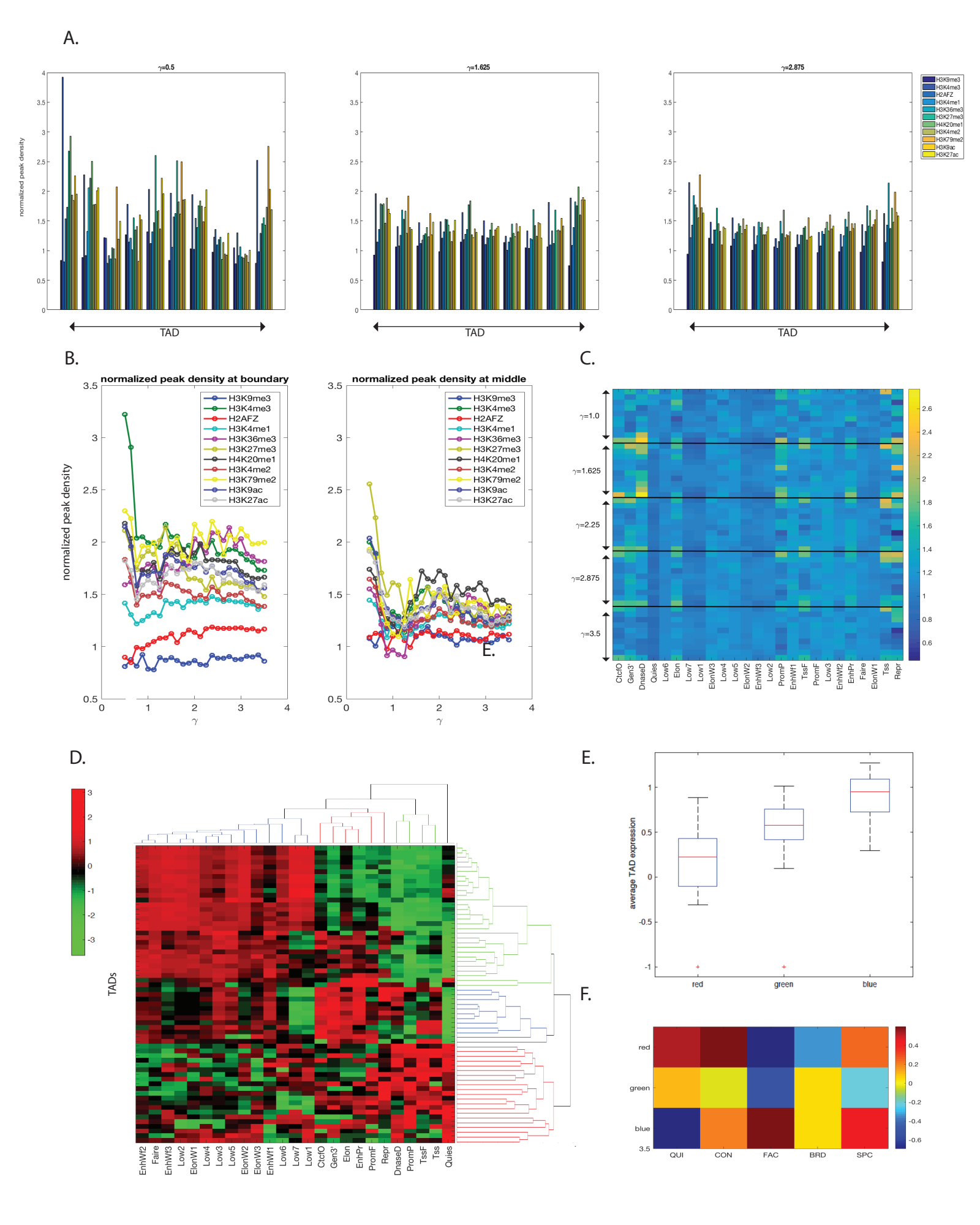

H3K4me2
H3K4me3
H3K79me2
H3K27ac
H3K9me3
H3K27me3
H3K9ac
H3K36me3
enrichment of peak density at boundary
resolution γ

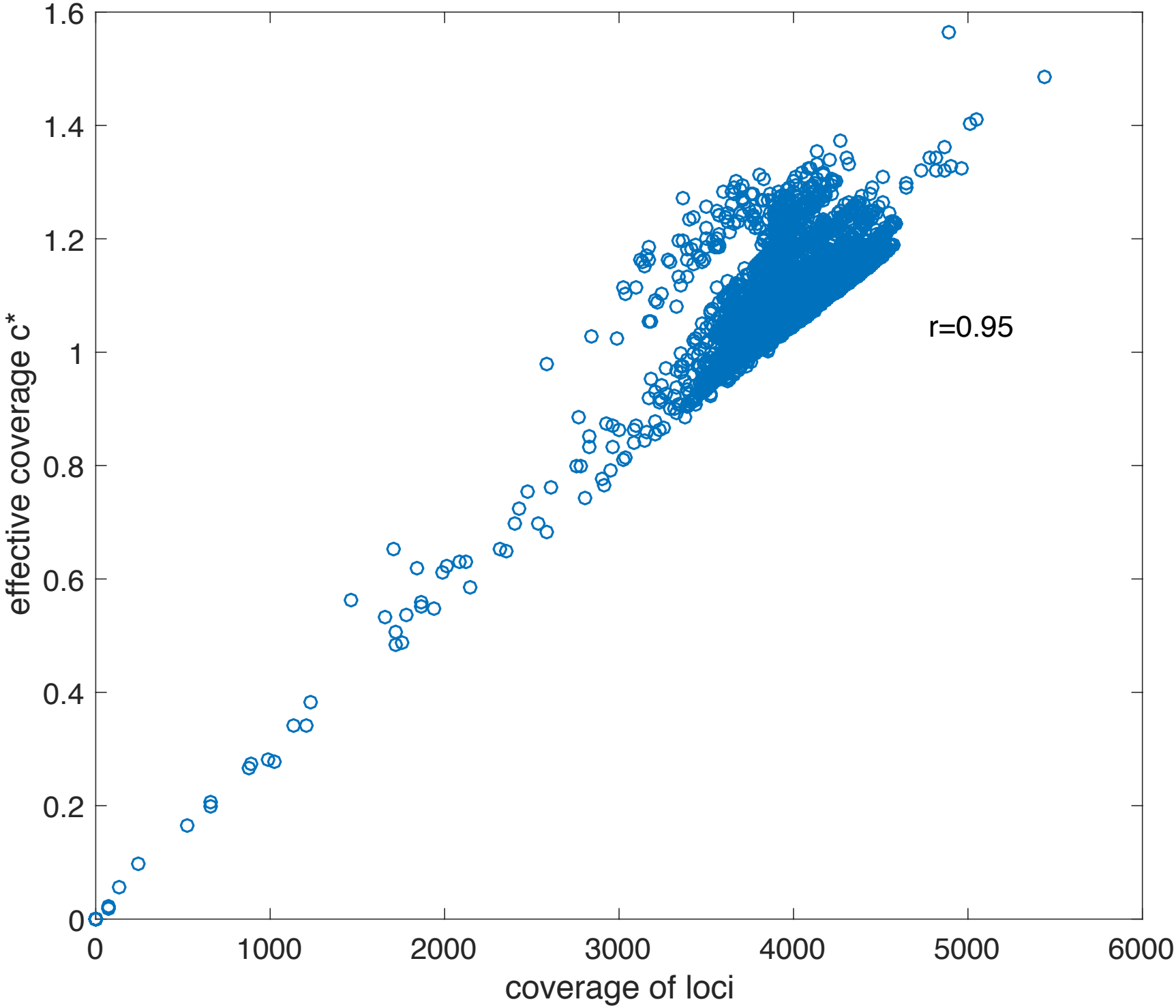
D.

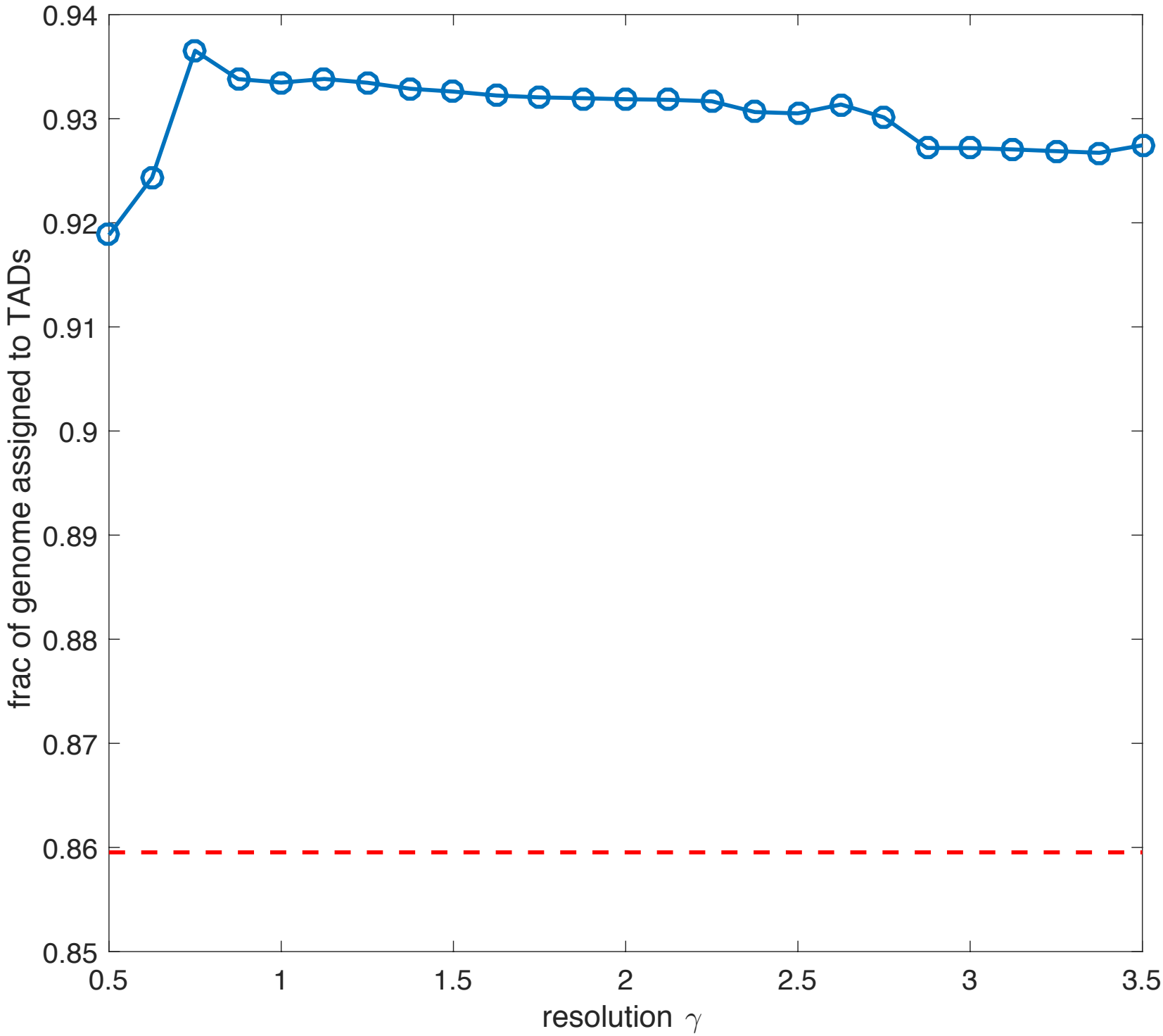

HOT regions
XOT regions
~HOT regions
~XOT regions

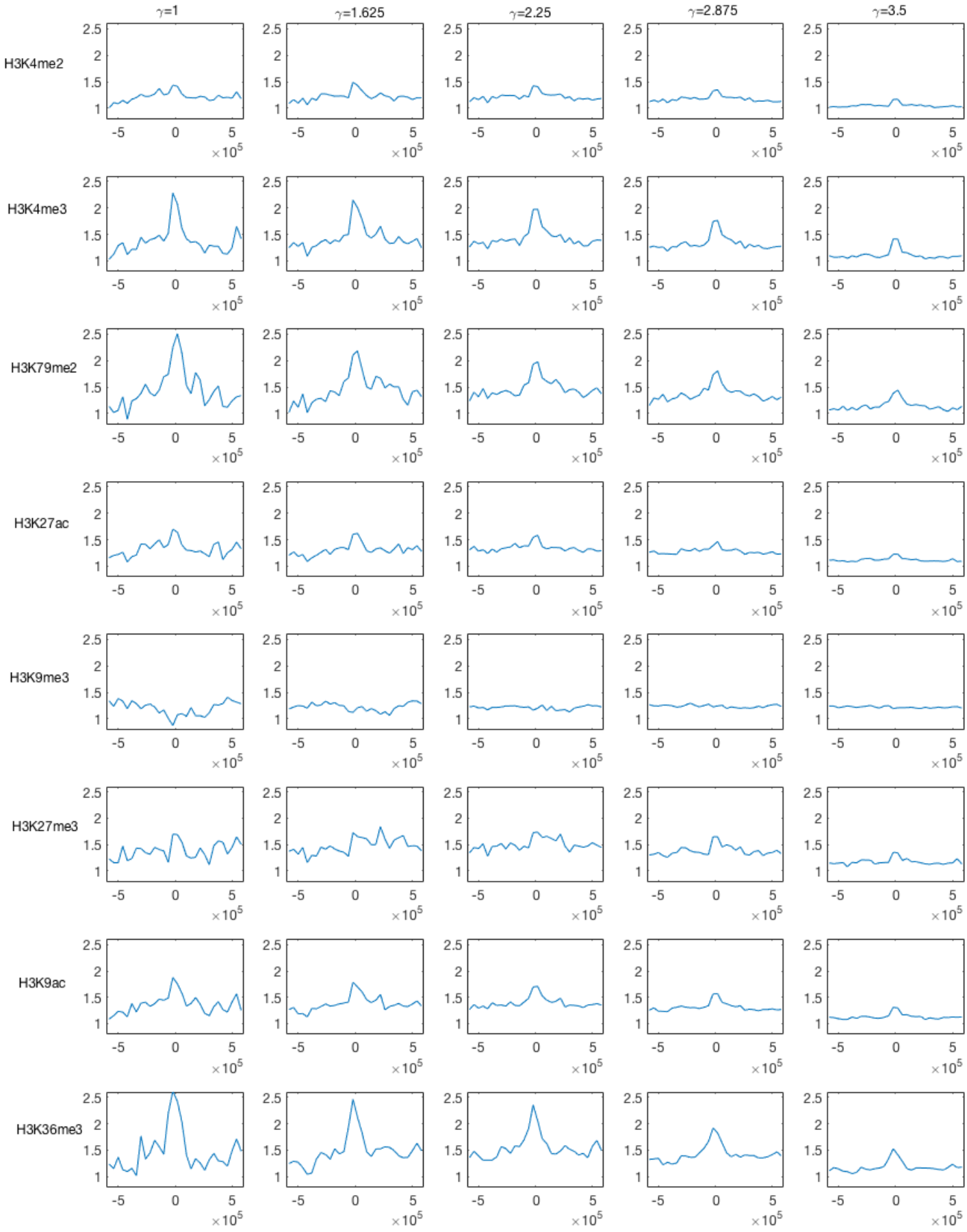
E.

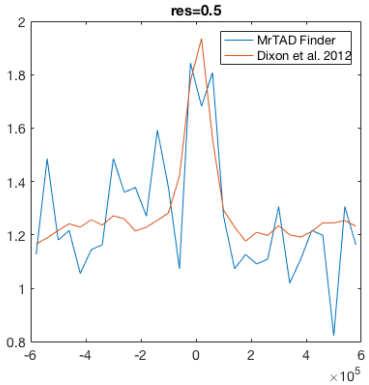

γ=1
γ=1.25
γ=1.5
γ=1.75
γ=2
γ=2.25
γ=2.5
γ=2.75
γ=3
γ=3.25
coeff. of influence

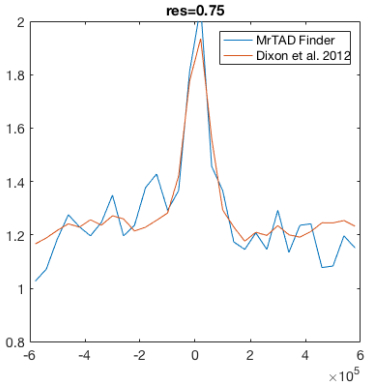

A.
γ=0.5
γ=1.625
γ=2.875
normalized peak density
TAD
H3K9me3
H3K4me3
H2AFZ
H3K4me1
H3K36me3
H3K27me3
H4K20me1
H3K4me2
H3K79me2
H3K9ac
H3K27ac
B.
normalized peak density at boundary
normalized peak density at middle
γ
C.
γ=1.0
γ=1.625
γ=2.25
γ=2.875
γ=3.5
CtcfO
Gen3'
DnaseD
Ques
Low6
Elon
Low7
Low1
ElonW3
Low4
Low5
ElonW2
EnhW3
Low2
PromP
EnhW1
TssF
PromF
Low3
EnhW2
EnhPr
Faire
ElonW1
Tss
Repr
D.
TADs
EnhWf2
Faire
EnhWf3
Low2
ElonW1
Low4
Low3
Low5
ElonW2
ElonW3
EnhWf1
Low6
Low7
Low1
CtcfO
Gen3'
Elon
EnhPr
PromF
Repr
DnaseD
PromP
TssF
Tss
Ques
E.
average TAD expression
red
green
blue
F.
red
green
blue
QUI
CON
FAC
BRD
SPC

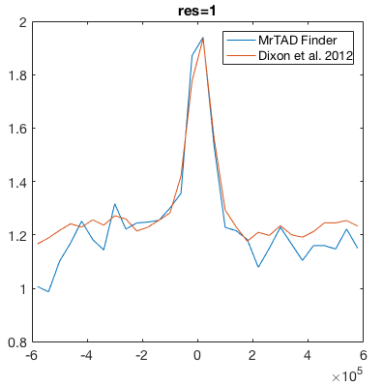

1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
effective coverage c*
0
1000
2000
3000
4000
5000
6000
coverage of loci
r=0.95

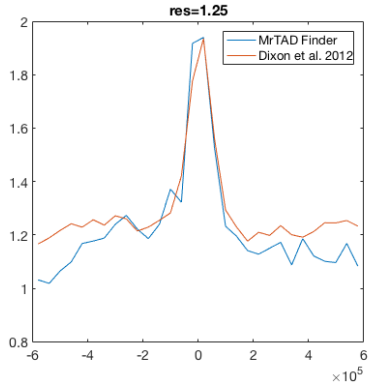

0.94
0.93
0.92
0.91
0.9
0.89
0.88
0.87
0.86
0.85
frac of genome assigned to TADs
0.5
1
1.5
2
2.5
3
3.5
resolution γ

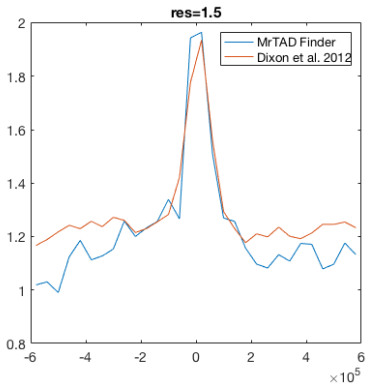

γ=1
γ=1.625
γ=2.25
γ=2.875
γ=3.5
H3K4me2
H3K4me3
H3K79me2
H3K27ac
H3K9me3
H3K27me3
H3K9ac
H3K36me3
2.5
2
1.5
1
-5
0
5
×10^5

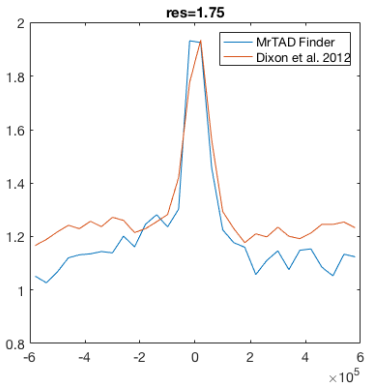

res=0.5
MrTAD Finder
Dixon et al. 2012
×10^5

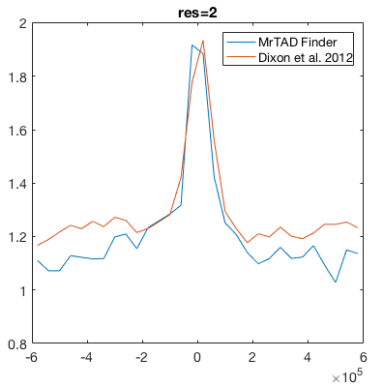

res=0.75
MrTAD Finder
Dixon et al. 2012
×10^5

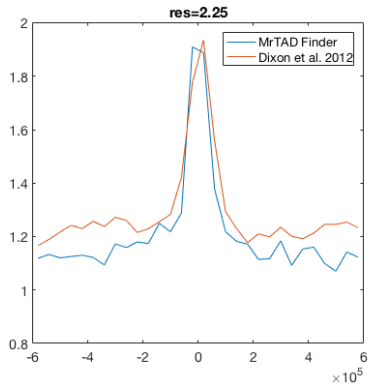

res=1
MrTAD Finder
Dixon et al. 2012
×10^5

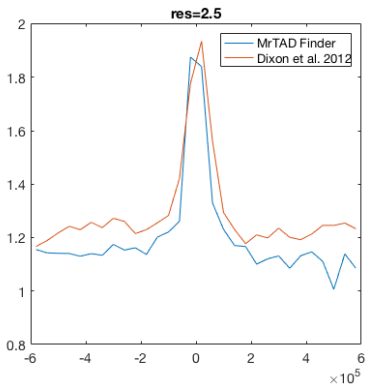

res=1.25
MrTAD Finder
Dixon et al. 2012
×10^5

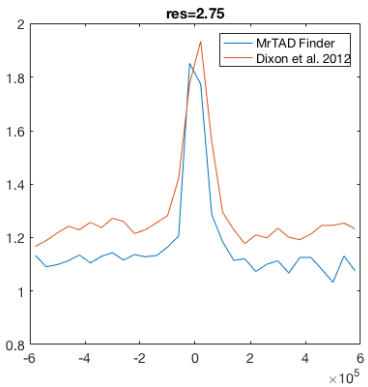

res=1.5
MrTAD Finder
Dixon et al. 2012
×10^5

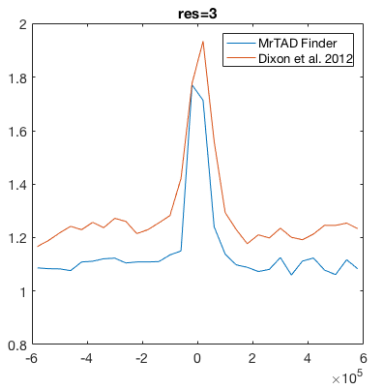

res=1.75
MrTAD Finder
Dixon et al. 2012
×10^5

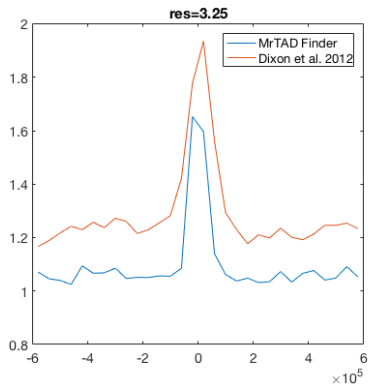

res=2
MrTAD Finder
Dixon et al. 2012
×10^5

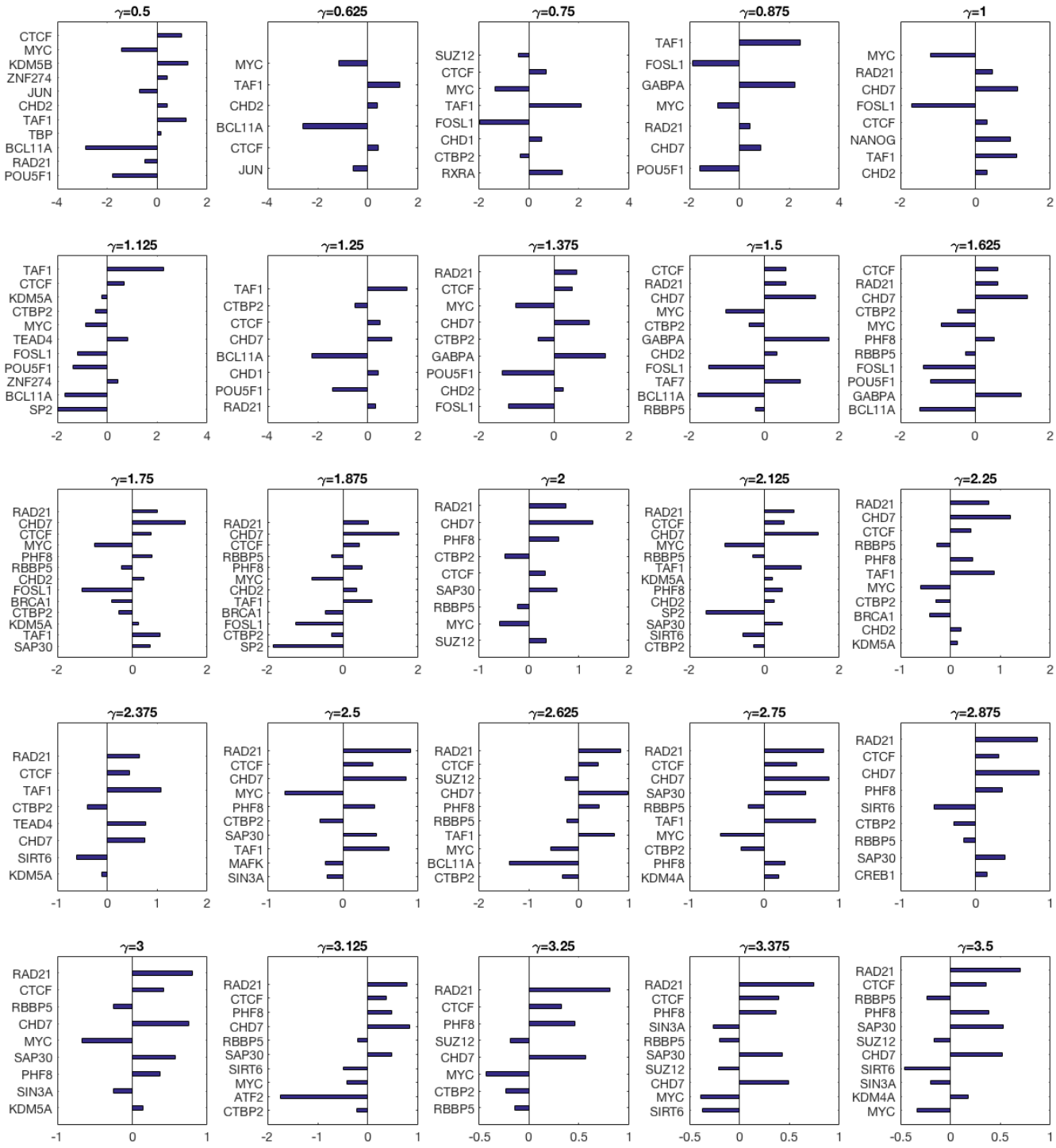

res=2.25
MrTAD Finder
Dixon et al. 2012
×10^5

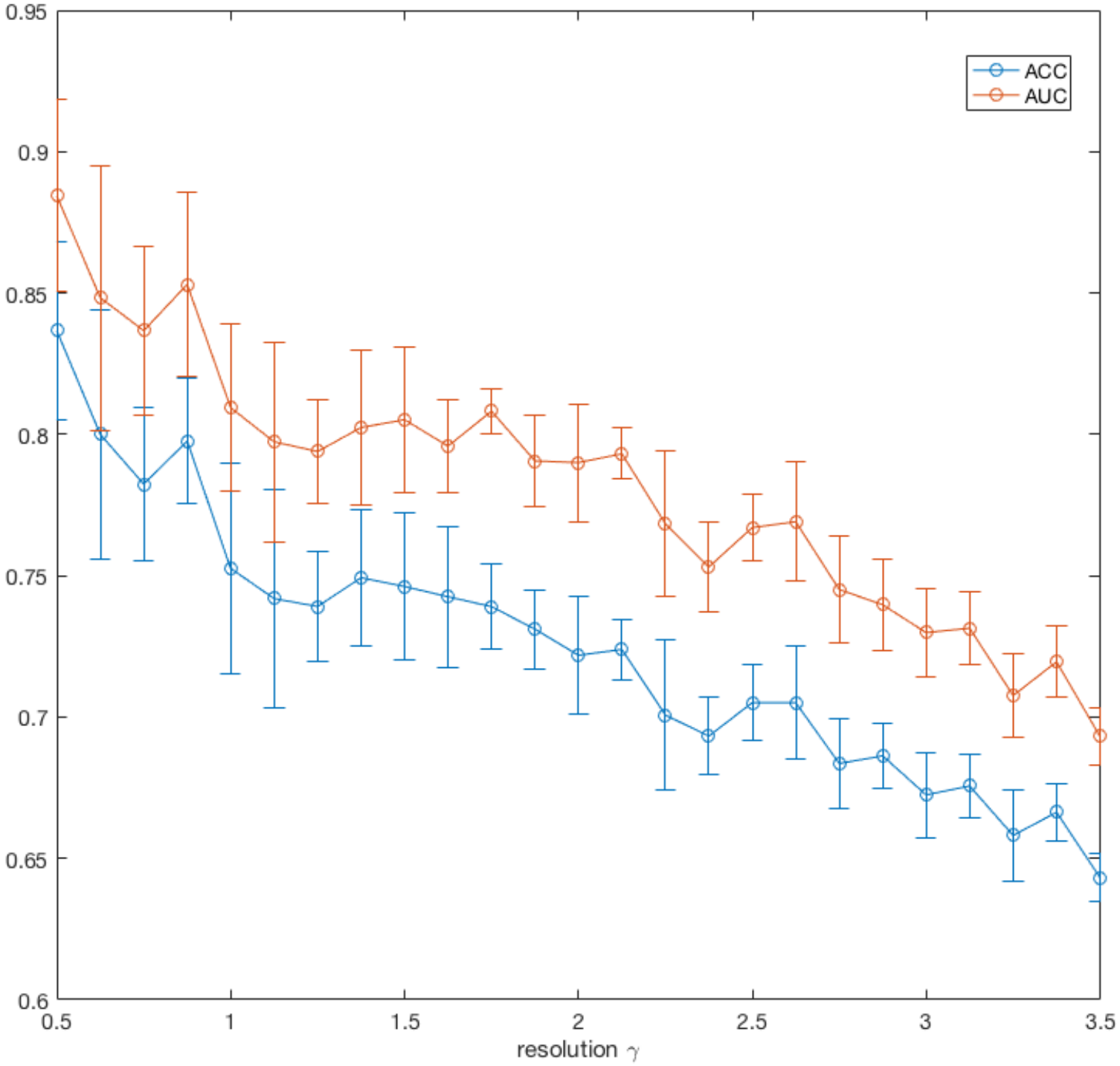

res=2.5
MrTAD Finder
Dixon et al. 2012
×10^5

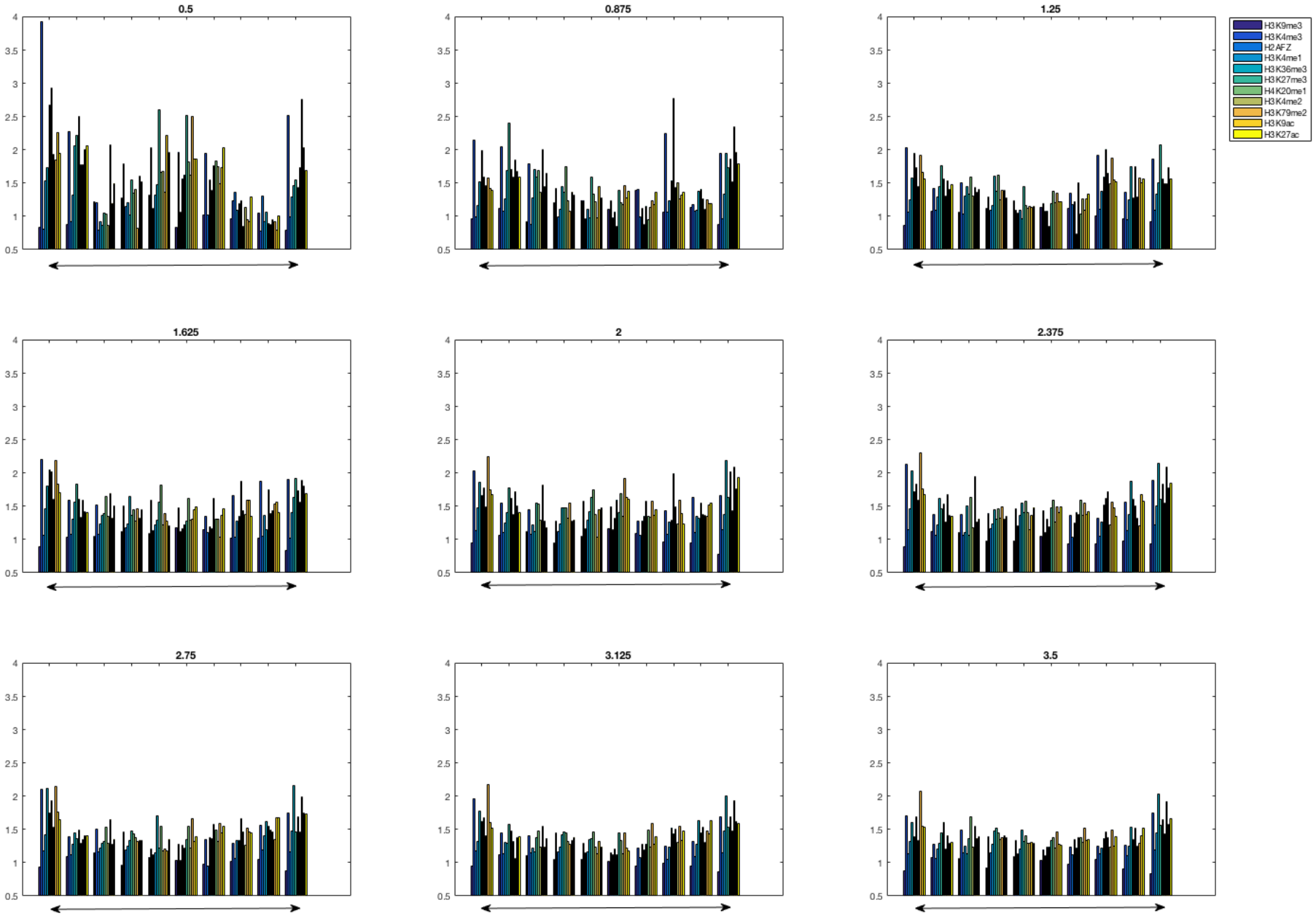

res=2.75
MrTAD Finder
Dixon et al. 2012
×10^5

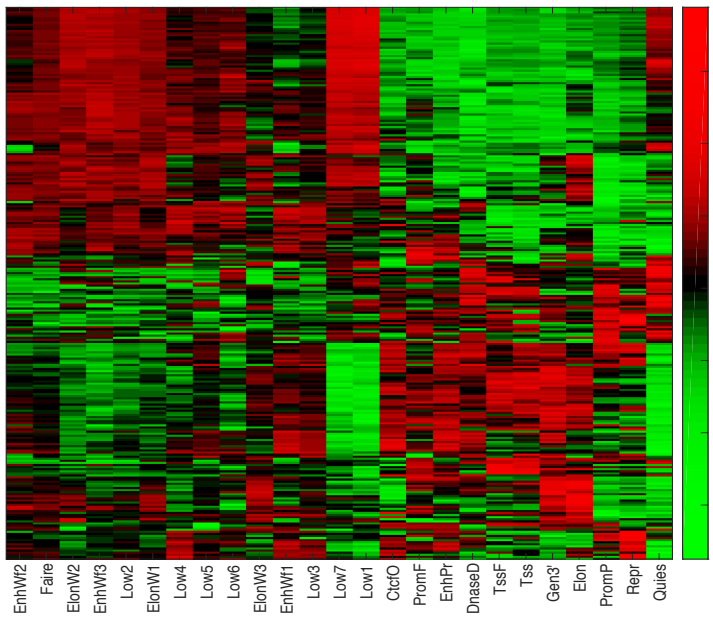

res=3
MrTAD Finder
Dixon et al. 2012
×10^5

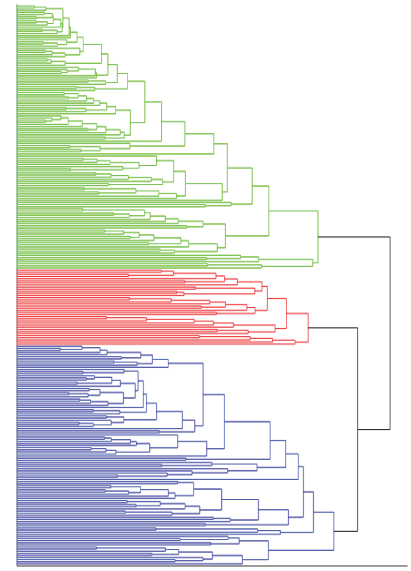

res=3.25
MrTAD Finder
Dixon et al. 2012
×10^5

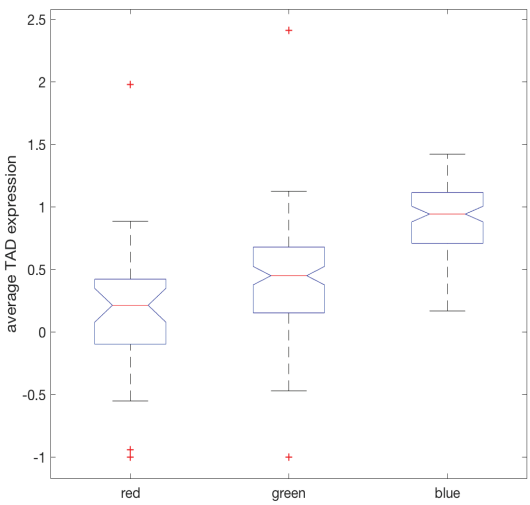

γ=0.5
CTCF
MYC
KDM5B
ZNF274
JUN
CHD2
TAF1
TBP
BCL11A
RAD21
POU5F1
-4 -2 0 2
γ=0.625
MYC
TAF1
CHD2
BCL11A
CTCF
JUN
-4 -2 0 2
γ=0.75
SUZ12
CTCF
MYC
TAF1
FOSL1
CHD1
CTBP2
RXRA
-2 0 2 4
γ=0.875
TAF1
FOSL1
GABPA
MYC
RAD21
CHD7
POU5F1
-2 0 2 4
γ=1
MYC
RAD21
CHD7
FOSL1
CTCF
NANOG
TAF1
CHD2
-2 0 2
γ=1.125
TAF1
CTCF
KDM5A
CTBP2
MYC
TEAD4
FOSL1
POU5F1
ZNF274
BCL11A
SP2
-2 0 2 4
γ=1.25
TAF1
CTBP2
CTCF
CHD7
BCL11A
CHD1
POU5F1
RAD21
-4 -2 0 2
γ=1.375
RAD21
CTCF
MYC
CHD7
CTBP2
GABPA
POU5F1
CHD2
FOSL1
-2 0 2
γ=1.5
CTCF
RAD21
CHD7
MYC
CTBP2
GABPA
CHD2
FOSL1
TAF7
BCL11A
RBBP5
-2 0 2
γ=1.625
CTCF
RAD21
CHD7
CTBP2
MYC
PHF8
RBBP5
FOSL1
POU5F1
GABPA
BCL11A
-2 0 2
γ=1.75
RAD21
CHD7
CTCF
MYC
PHF8
RBBP5
CHD2
FOSL1
BRCA1
CTBP2
KDM5A
TAF1
SAP30
-2 0 2
γ=1.875
RAD21
CHD7
CTCF
RBBP5
PHF8
MYC
CHD2
TAF1
BRCA1
FOSL1
CTBP2
SP2
-2 0 2
γ=2
RAD21
CHD7
PHF8
CTBP2
CTCF
SAP30
RBBP5
MYC
SUZ12
-1 0 1 2
γ=2.125
RAD21
CTCF
CHD7
MYC
RBBP5
TAF1
KDM5A
PHF8
CHD2
SP2
SAP30
SIRT6
CTBP2
-2 0 2
γ=2.25
RAD21
CHD7
CTCF
RBBP5
PHF8
TAF1
MYC
CTBP2
BRCA1
CHD2
KDM5A
-1 0 1 2
γ=2.375
RAD21
CTCF
TAF1
CTBP2
TEAD4
CHD7
SIRT6
KDM5A
-1 0 1 2
γ=2.5
RAD21
CTCF
CHD7
MYC
PHF8
CTBP2
SAP30
TAF1
MAFK
SIN3A
-1 0 1
γ=2.625
RAD21
CTCF
SUZ12
CHD7
PHF8
RBBP5
TAF1
MYC
BCL11A
CTBP2
-2 -1 0 1
γ=2.75
RAD21
CTCF
CHD7
SAP30
RBBP5
TAF1
MYC
CTBP2
PHF8
KDM4A
-1 0 1
γ=2.875
RAD21
CTCF
CHD7
PHF8
SIRT6
CTBP2
RBBP5
SAP30
CREB1
-1 0 1
γ=3
RAD21
CTCF
RBBP5
CHD7
MYC
SAP30
PHF8
SIN3A
KDM5A
-1 0 1
γ=3.125
RAD21
CTCF
PHF8
CHD7
RBBP5
SAP30
SIRT6
MYC
ATF2
CTBP2
-2 -1 0 1
γ=3.25
RAD21
CTCF
PHF8
SUZ12
CHD7
MYC
CTBP2
RBBP5
-0.5 0 0.5 1
γ=3.375
RAD21
CTCF
PHF8
SIN3A
RBBP5
SAP30
SUZ12
CHD7
MYC
SIRT6
-0.5 0 0.5 1
γ=3.5
RAD21
CTCF
RBBP5
PHF8
SAP30
SUZ12
CHD7
SIRT6
SIN3A
KDM4A
MYC
-0.5 0 0.5 1

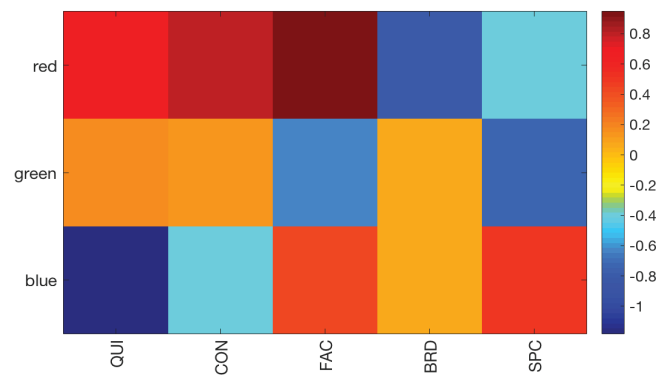

ACC
AUC
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.5
1
1.5
2
2.5
3
3.5
resolution γ


0.5
0.875
1.25
1.625
2
2.375
2.75
3.125
3.5
H3K9me3
H3K4me3
H2AFZ
H3K4me1
H3K36me3
H3K27me3
H4K20me1
H3K4me2
H3K79me2
H3K9ac
H3K27ac

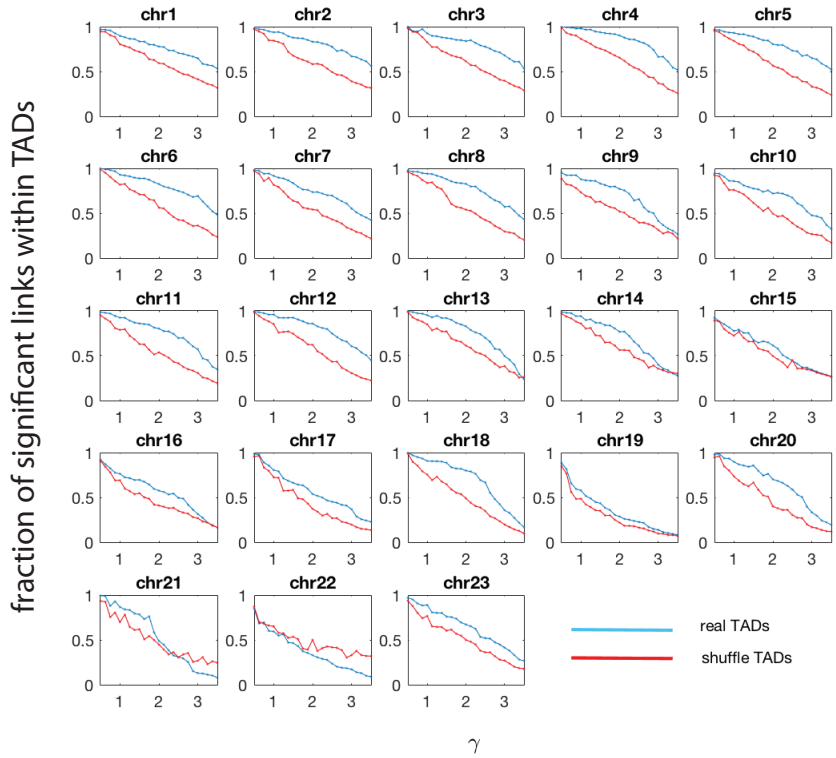

EnhWf2
Faire
ElonW2
EnhWf3
Low2
ElonW1
Low4
Low5
Low6
ElonW3
EnhWf1
Low3
Low7
Low1
CtcfO
PromF
EnhPr
DnaseD
TssF
Tss
Gen3'
Elon
PromP
Repr
Quies



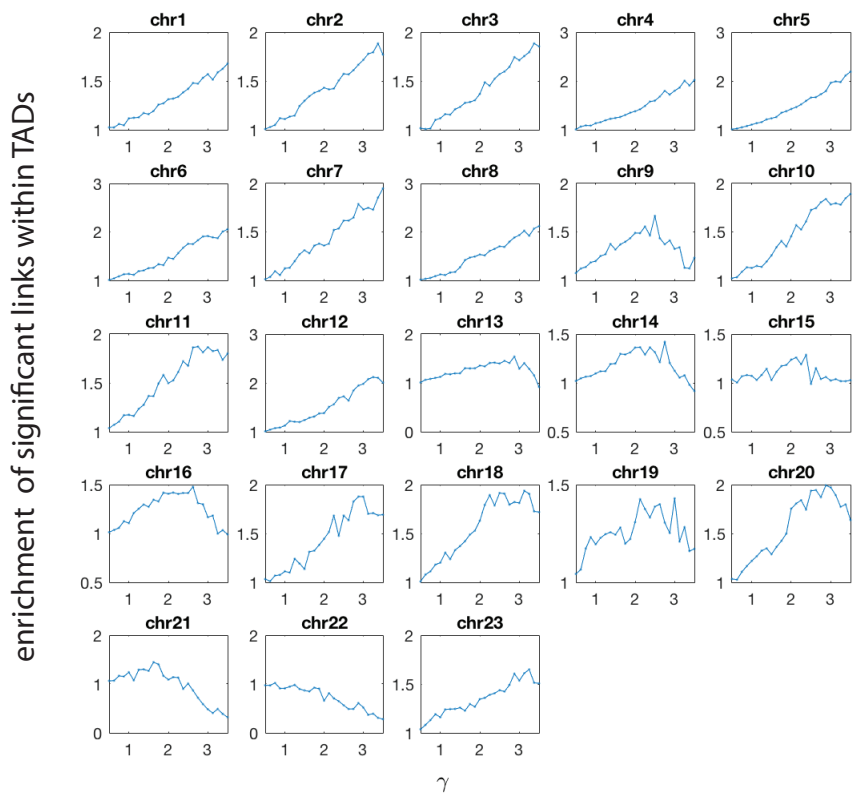

average TAD expression
2.5
2
1.5
1
0.5
0
-0.5
-1
red
green
blue

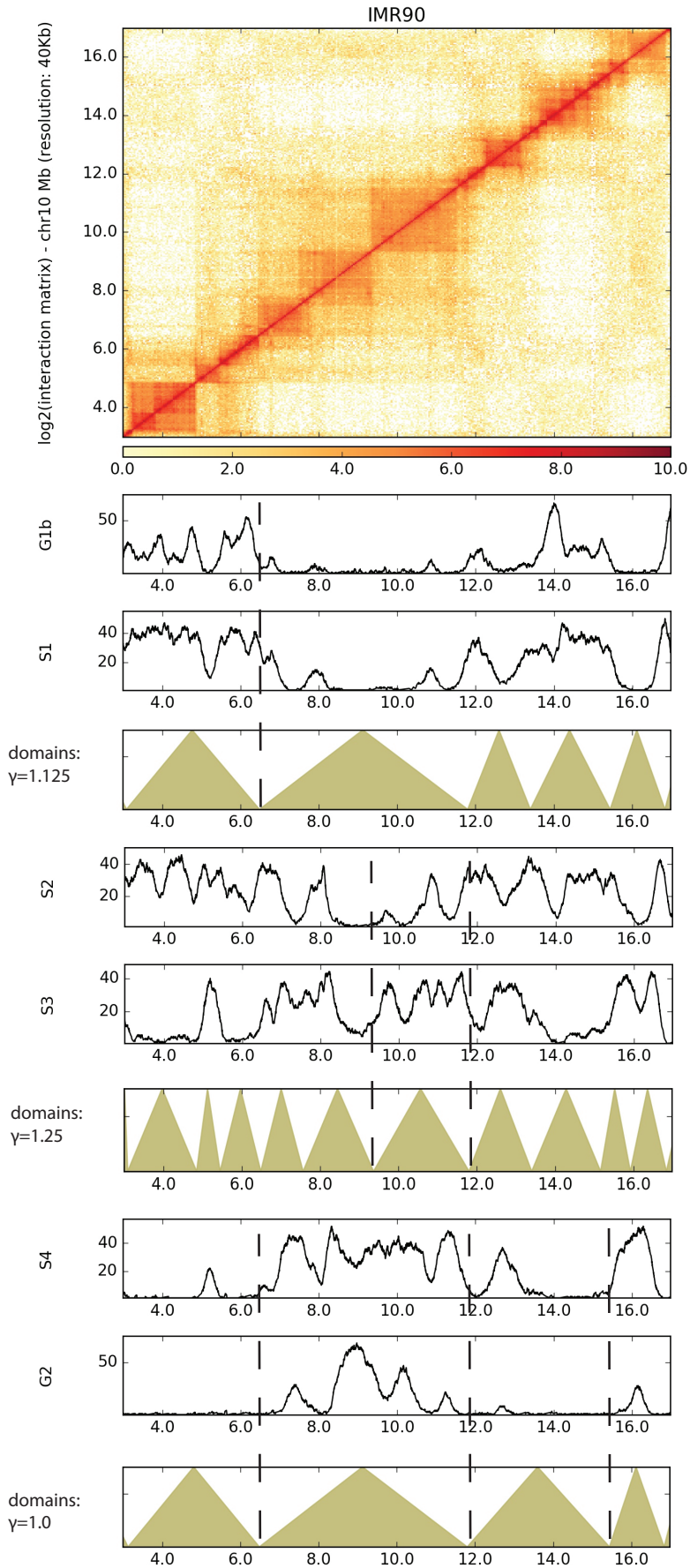

red
green
blue
QUI
CON
FAC
BRD
SPC
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

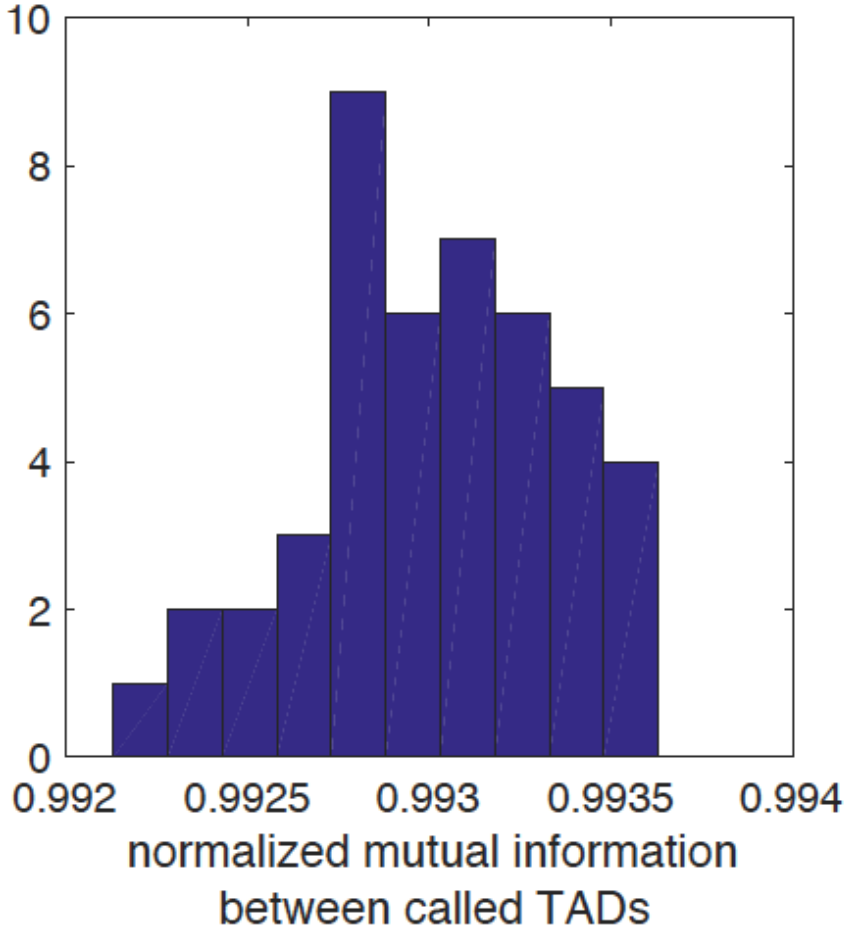
A


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chr23
real TADs
shuffle TADs
fraction of significant links within TADs
γ

B


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chr23
enrichment of significant links within TADs
γ


IMR90
log2(interaction matrix) - chr10 Mb (resolution: 40Kb)
G1b
S1
domains:
γ=1.125
S2
S3
domains:
γ=1.25
S4
G2
domains:
γ=1.0


10
8
6
4
2
0
0.992
0.9925
0.993
0.9935
0.994
normalized mutual information
between called TADs