

3-14-2019

TADKB:Family Classification and a Knowledge Base of Topologically Associating Domains

Tong Liu
University of Miami

Jacob Porter
University of Southern Mississippi

Chenguang Zhao
University of Southern Mississippi

Hao Zhu
University of Southern Mississippi

Nan Wang
New Jersey City University

See next page for additional authors

Follow this and additional works at: https://aquila.usm.edu/fac_pubs



Part of the [Genetics and Genomics Commons](#)

Recommended Citation

Liu, T., Porter, J., Zhao, C., Zhu, H., Wang, N., Sun, Z., Mo, Y., Wang, Z. (2019). TADKB:Family Classification and a Knowledge Base of Topologically Associating Domains. *BMC Genomics*, 20, 1-17.
Available at: https://aquila.usm.edu/fac_pubs/16260

This Article is brought to you for free and open access by The Aquila Digital Community. It has been accepted for inclusion in Faculty Publications by an authorized administrator of The Aquila Digital Community. For more information, please contact Joshua.Cromwell@usm.edu.

Authors

Tong Liu, Jacob Porter, Chenguang Zhao, Hao Zhu, Nan Wang, Zheng Sun, Yin-Yuan Mo, and Zheng Wang

DATABASE ARTICLE

Open Access



TADKB: Family classification and a knowledge base of topologically associating domains

Tong Liu¹, Jacob Porter², Chenguang Zhao², Hao Zhu², Nan Wang³, Zheng Sun⁴, Yin-Yuan Mo⁵ and Zheng Wang^{1*}

Abstract

Background: Topologically associating domains (TADs) are considered the structural and functional units of the genome. However, there is a lack of an integrated resource for TADs in the literature where researchers can obtain family classifications and detailed information about TADs.

Results: We built an online knowledge base TADKB integrating knowledge for TADs in eleven cell types of human and mouse. For each TAD, TADKB provides the predicted three-dimensional (3D) structures of chromosomes and TADs, and detailed annotations about the protein-coding genes and long non-coding RNAs (lncRNAs) existent in each TAD. Besides the 3D chromosomal structures inferred by population Hi-C, the single-cell haplotype-resolved chromosomal 3D structures of 17 GM12878 cells are also integrated in TADKB. A user can submit query gene/lncRNA ID/sequence to search for the TAD(s) that contain(s) the query gene or lncRNA. We also classified TADs into families.

To achieve that, we used the TM-scores between reconstructed 3D structures of TADs as structural similarities and the Pearson's correlation coefficients between the fold enrichment of chromatin states as functional similarities. All of the TADs in one cell type were clustered based on structural and functional similarities respectively using the spectral clustering algorithm with various predefined numbers of clusters. We have compared the overlapping TADs from structural and functional clusters and found that most of the TADs in the functional clusters with depleted chromatin states are clustered into one or two structural clusters. This novel finding indicates a connection between the 3D structures of TADs and their DNA functions in terms of chromatin states.

Conclusion: TADKB is available at <http://dna.cs.miami.edu/TADKB/>.

Keywords: Topologically associating domains, TADs, Family classification, Single-cell 3D genome structures, Long non-coding RNAs, lncRNAs

Background

Topologically associating domains (TADs) are DNA segments that are considered the structural and functional units of the mammalian genomes [1, 2]. The length of TADs varies from hundreds of kilobases up to a few million bases [1]. The boundaries of TADs are enriched with different factors [1], including the insulator binding protein CTCF and housekeeping genes. TADs pervade the whole genome, remain consistent across different

cell types, and are highly conserved between humans and mice [2]. Recently, TADs have been widely considered as the unit of chromosome organization [3] and being studied together with genes, CTCF, cohesion, and chromatin loops [2, 4, 5]. There are many methods that have been developed to detect topologically associating domains [1, 2, 6–13]. Most of them are based on the finding that the Hi-C contacts within a TAD are apparently more frequent and enriched than those between two different domains [1], which is the fundamental rule for defining domain locations in mammalian chromosomes.

* Correspondence: zheng.wang@miami.edu

¹Department of Computer Science, University of Miami, 1365 Memorial Drive, Coral Gables, FL 33124-4245, USA

Full list of author information is available at the end of the article



The Hi-C experiments [14] can capture the genome-wide proximate relationship between genomic locations based on millions of cells. The resolution of Hi-C experiments has been largely improved from originally 1 Mb in [14] to recently 1 kb in [2]. This high resolution makes it possible to detect enough Hi-C contacts within a TAD or detect genome-wide loops. For example, the study [2] identified about 10,000 loops, which often indicate promoter and enhancer interactions that is highly related to gene regulation. Studies also found that the loops usually are conserved between different cell types and species [2, 15].

The availability of high-resolution Hi-C contacts also makes it possible to reconstruct the three-dimensional (3D) structure of chromosomes. The Hi-C contact data indicate the proximate relationship between two genomic locations, with enough number of which computational algorithms can be used to construct a 3D structure that meets the Hi-C contacts. The early work conducted by Duan et al. [16] constructed the 3D structure of yeast genome based on 4C-related experiment (4C, a type of chromosome conformation capture experiment that was designed before the invention of Hi-C experiment). ChromSDE [17] uses semi-definite programming to construct 3D models, whereas Trieu et al. [18] applied optimization after obtaining the in-contact and not-in-contact relationships for bead pairs. PASTIS [19] uses metric multidimensional scaling to construct 3D structures, which at first calculates a wish distance between every pair of beads (a chromosome is evenly divided into beads with the same length). This wish distance is calculated directly from the number of Hi-C contacts by $d \sim c^{-1/3}$ (d is the wish distance; and c is the number of Hi-C contacts) so that higher number of Hi-C contacts indicate shorter wish distances. The multidimensional scaling algorithm tries to find a 3D structure that best meets all the wish distances.

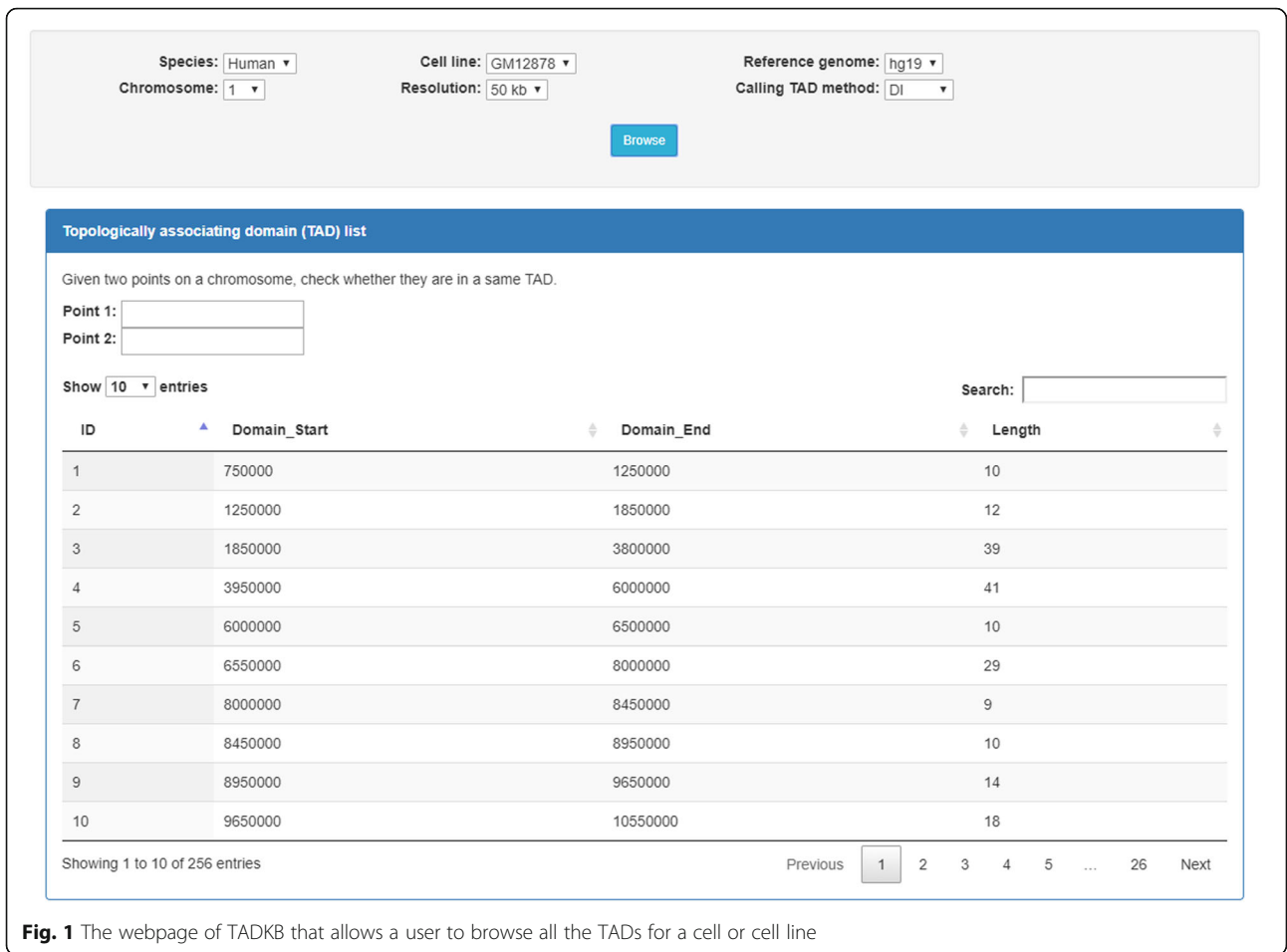
The converting formula $d \sim c^{-1/3}$ has a drawback, that is, when c is larger than 10 the converted distances are converged to a very small value. To overcome the drawback, instead of using the same parameter (1/3) for all Hi-C contacts we [20] defined a novel type of complex network based on Hi-C contacts and assigned a converting parameter for each pair of Hi-C contacts based on their affinity to the neighbors, from which we further inferred the wish distance for each bead pair. Based on the bead-pair specific wish distances, we reconstructed the 3D structures of chromosomes and TADs at the 40 kb resolution [20]. Although this technique was not used in TADKB, it is worth mentioning it for a broad review of the algorithms used to reconstruct genome 3D structures.

Given a distance matrix, reconstructing a 3D structure can be considered as a dimensionality reduction problem.

Generally speaking, the methods to achieve that can be classified to linear (e.g., principal component analysis) and non-linear (e.g., multi-dimensional scaling [21] and t-distributed stochastic neighbor embedding [22]) methods. Non-linear methods are more complicated than the linear ones and can capture the non-linear relationships from the input data. Among most of the non-linear methods, t-distributed stochastic neighbor embedding (t-SNE) used Gaussian joint probabilities to represent affinities in the original space and Student's t-distributions to represent affinities in the embedded space [22]. It has been claimed in [22] that the t-SNE method has advantages such as being able to reveal the structures at different scales. Therefore, it can be used to capture and reconstruct local structures from single-cell Hi-C contact matrices [23, 24].

Long non-coding RNA (lncRNA) is defined as transcript of > 200 nucleotides that cannot be translated into protein. It has been found that > 74% of human genome is transcribed to RNA; however, only 2% of the transcripts are finally translated into proteins [25]. Therefore, non-coding RNAs take a large portion in human genome and have been considered as “junk”. It is until recently that more and more research has confirmed lncRNAs functions in gene expressions regulation [26, 27], epigenetic modification [28–30], and chromatin structures controlling [31]. For example, *Xist* is a lncRNA with gene locus located in the X-chromosome of mammal cells. Its important function is to inactivate one copy of X chromosome in female cells. Because every diploid wild-type female mammal cell has two copies of X chromosomes, in order to balance the amount of gene expressions or to perform “dosage compensation”, one of the X chromosomes in female is inactivated with highly compacted structure and silenced in terms of gene expression. This inactivation process is done by *Xist* lncRNAs that alter the 3D structure of X chromosome and eventually inactivate one copy of X chromosomes in female [32]. There are multiple databases for lncRNA such as NONCODE 2016 [33], LNCipedia 4.0 [34], and lncRNAdb 2.0 [35]. However, different lncRNA databases have different naming standards, which causes the problem that the same lncRNA has different IDs in different databases.

We built topologically associating domain knowledge base (TADKB), a knowledge base for TADs integrated with annotations of protein-coding genes and lncRNAs. TADKB defined TADs' families based on the common TADs shared in two types of clusters: (1) structural clusters based on 3D structural similarities; (2) chromatin-state clusters from the fold enrichment similarities of chromatin states. Moreover, TADKB unifies three lncRNA databases allowing users to cross-reference between them when they have different IDs for the same lncRNA.



Construction and content

TADKB provides the TADs called from eleven cell types: GM12878, HMEC, NHEK, IMR90, KBM7, K562, and HUVEC for human [2], and CH12-LX, ES, NPC, and CN for mouse [36]. The normalized Hi-C contact matrices were downloaded from the Gene Expression Omnibus (GEO) with ID GSE63525 for the first eight cell types at the resolutions of 50 kb and 10 kb and GEO GSE96107 for the last three cell types at the resolutions of 50 kb and 10 kb. The TAD locations for all of the cell types were detected using three different methods: (1) Directionality Index (DI) [1], Gaussian Mixture model And Proportion test (GMAP) [37], and Insulation Score (IS) [38]. For IS, we first combined the overlapping boundary regions and called domains between two successive boundaries. We also used two Hi-C variants: HiChIP [39] and SPRITE [40], and both the variants provided two cell lines' high-resolution chromatin contact data, including GM12878 and mES. The details of domain-detection results are shown in Additional file 1: Table S1. Hi-C data are normalized using KR [2, 41], whereas HiChIP and SPRITE data are normalized using Hi-Corrector [42] with 100 iterations. All TAD annotations described in Additional file 1: Table S1 can be downloaded from TADKB's download webpage.

Because the scale of Hi-C contacts widely varies and the contact-to-distance converting formula $d = (1/c)^{(1/3)}$ as defined in [19] is sensitive to the scale of the number of Hi-C contacts [20], we first rescaled the Hi-C contacts of each TAD to the range [1, 30] via linear transformation without considering missing Hi-C values. We then

used the formula $d = (1/c)^{(1/3)}$ to convert Hi-C contacts (c) into wish distances (d). We reconstructed each TAD's 3D structure using two manifold learning methods including metric multidimensional scaling (MDS) and t-distributed Stochastic Neighbor Embedding (t-SNE) [22] implemented in Scikit-learn [43] by reducing the dimensionality to three components. We found that the reconstructed 3D structures of TADs using t-SNE are very sensitive to two parameters (i.e., perplexity and learning rate). Therefore, we generated multiple 3D structures for each TAD using t-SNE with different configurations of the two parameters, superimposed these structures with the one predicted by MDS method [44], and selected the structure with the minimum root-mean-square deviation (RMSD) as the final structure from t-SNE.

We evaluated the reconstructed 3D structures using the correlation between exponent parameter (measuring the contact probability against genomic distances based on Hi-C contact maps, see definition in Additional file 1) and radius of gyration (measuring the compactness of reconstructed 3D structures) as described in our previous work [45]. Because a better reconstructed 3D structure should have a high consistency between the 2D structural characteristics represented by exponent parameter and the 3D compactness represented by radius of gyration, we calculated the correlations between all TADs' exponent parameters and radius of gyration for MDS- and t-SNE-inferred structures in GM12878. The Pearson's and Spearman correlation coefficients between contact-probability-based exponent parameters and MDS-based radius of

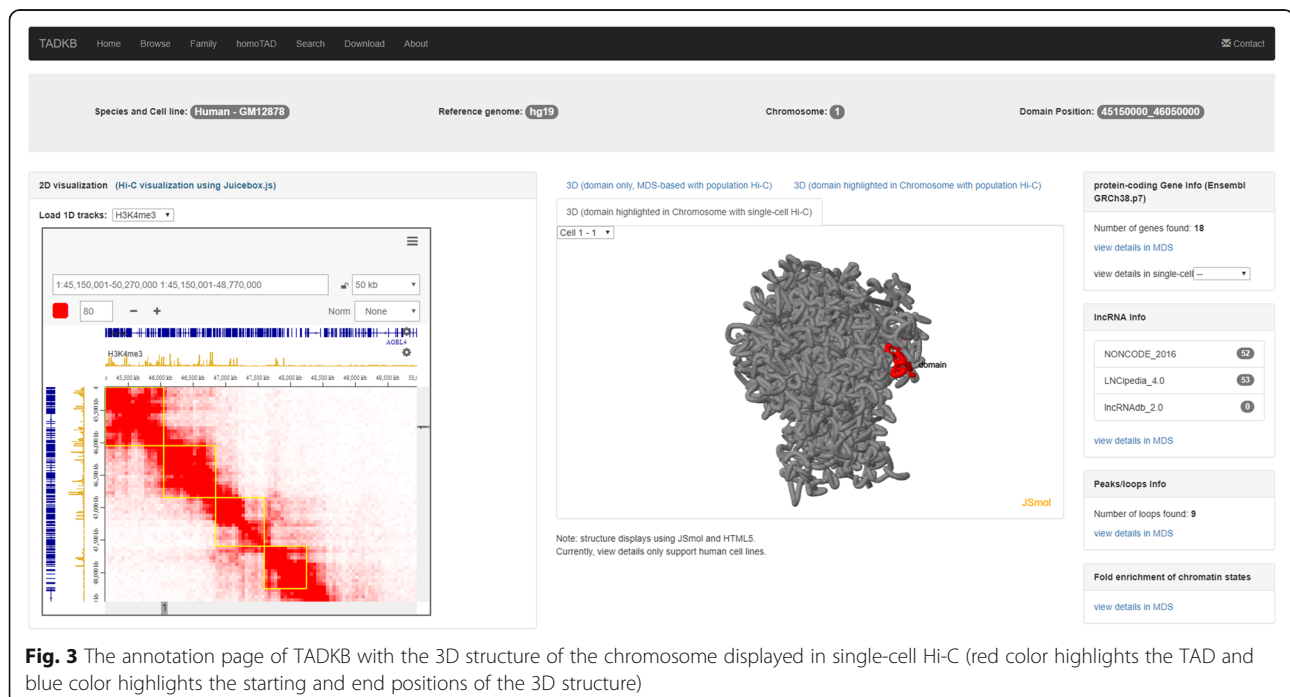


Fig. 3 The annotation page of TADKB with the 3D structure of the chromosome displayed in single-cell Hi-C (red color highlights the TAD and blue color highlights the starting and end positions of the 3D structure)

gyrations are -0.71 (P -Value $< 2.2e-16$) and -0.77 (P -Value $< 2.2e-16$), respectively, whereas the correlations between contact-probability-based exponent parameters and t-SNE-based radius of gyrations are -0.08 (P -Value = $8.2e-06$) and -0.02 (P -Value = 0.2487). Our evaluation results indicate that the structures inferred by MDS share higher consistency than the structures inferred by t-SNE. Therefore, we used MDS-based structures in the downstream analysis. The 3D structures of the chromosomes and TADs were inferred using the same method.

We used our in-house tool named SCL (manuscript submitted) to reconstruct the 3D structures of chromosomes based on single-cell Hi-C data. The single-cell haplotype-resolved chromosomal 3D structures at 40 kb resolution of 17 GM12878 cells were generated based on the single-cell Hi-C data released from [46]. For the chromosomes 10 and 19 of cell 1, chromosomes 1, 2, 4, and 11 of cell 4, all chromosomes of cell 8, and

chromosome 6 of cell 10, the raw single-cell Hi-C contacts (file name *.raw.con.txt.gz) were used to infer their 3D structures. For all other chromosomes and cells, the single-cell Hi-C contact after imputation were used (file name *.impute3.round4.con.txt.gz). All single-cell Hi-C data were downloaded from [46].

After obtaining the reconstructed 3D structures, we used 3D structure alignment tools to compare the structural similarity between any given two TADs. In this study, we used TM-align [47] to superimpose two TADs' structures and obtained the TM-score as the structural similarity score normalized by the length of the smaller TAD. Therefore, given the reconstructed 3D structures of all TADs in a genome we used TM-score to generate a structural similarity matrix.

We next used chromatin-state annotation [48] to explore the chromatin-state similarity between any two TADs. We downloaded the 25-state annotations from

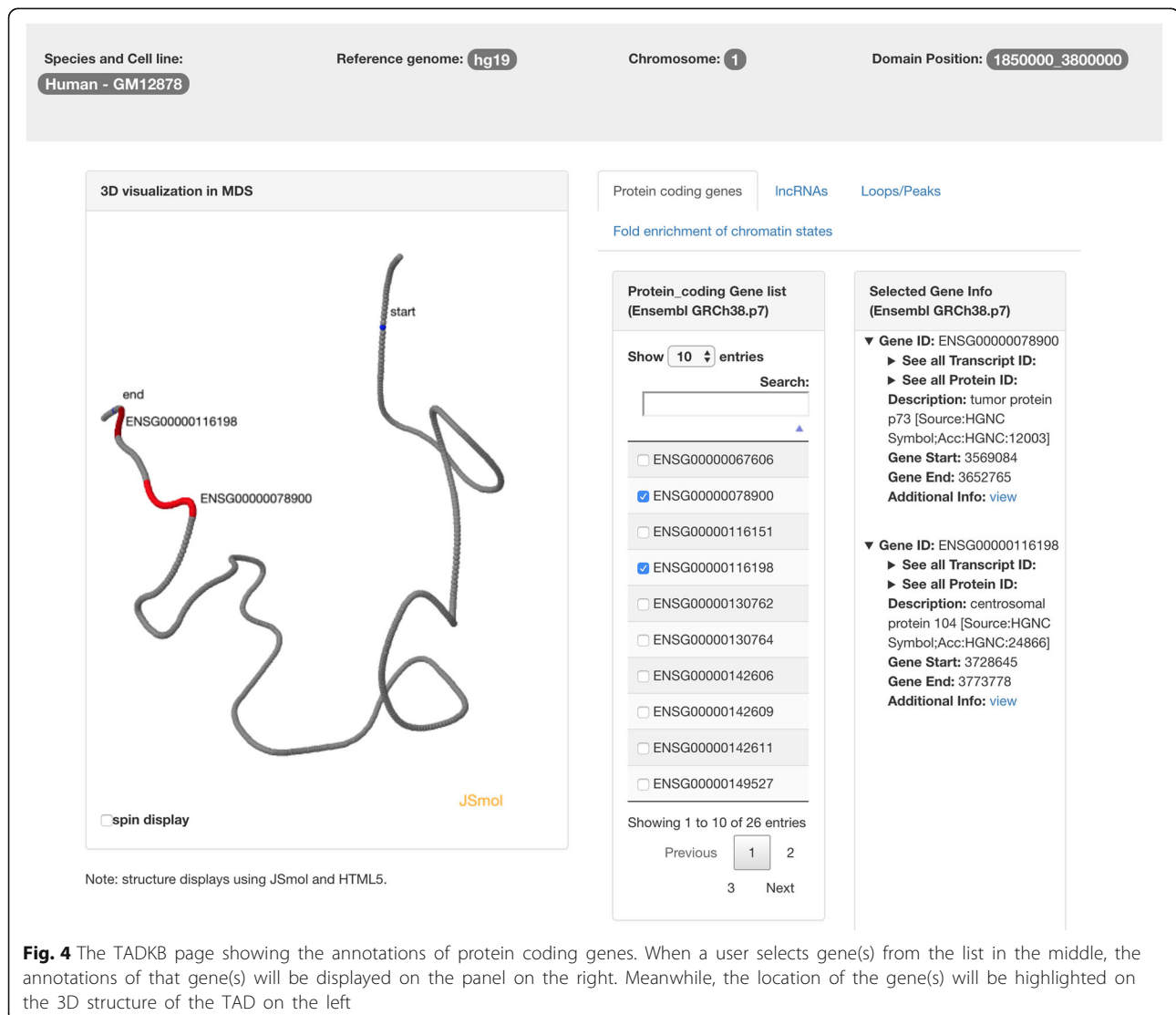


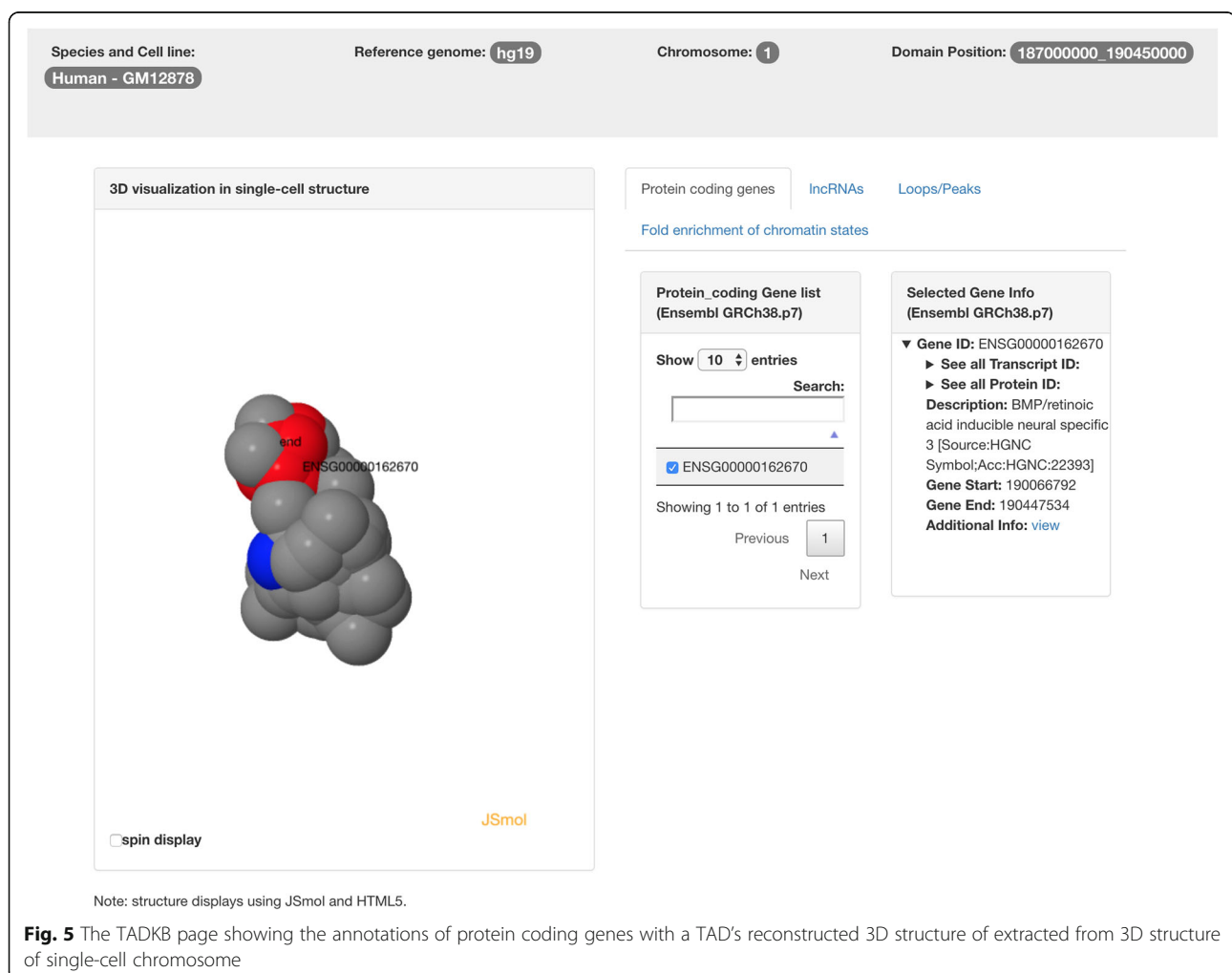
Fig. 4 The TADKB page showing the annotations of protein coding genes. When a user selects gene(s) from the list in the middle, the annotations of that gene(s) will be displayed on the panel on the right. Meanwhile, the location of the gene(s) will be highlighted on the 3D structure of the TAD on the left

the roadmap epigenomics project [49] for six cell types including GM12878, HMEC, HUVEC, IMR90, K562, and NHEK. The 25 states are (1) active TSS, (2) promoter upstream TSS, (3) promoter downstream TSS 1, (4) promoter downstream TSS 2, (5) transcribed-5' preferential, (6) strong transcription, (7) transcribed-3' preferential, (8) weak transcription, (9) transcribed & regulatory (Prom/Enh), (10) transcribed 5' preferential and Enh, (11) transcribed 3' preferential and Enh, (12) transcribed and weak Enhancer, (13) active enhancer 1, (14) active enhancer 2, (15) active enhancer flank, (16) weak enhancer 1, (17) weak enhancer 2, (18) primary H3K27ac possible Enhancer, (19) primary DNase, (20) ZNF genes & repeats, (21) heterochromatin, (22) poised promoter, (23) bivalent promoter, (24) repressed polycomb, and (25) quiescent/low. For each TAD in each of the six cell types with available chromatin-state annotations, we computed its fold enrichment of each state using the OverlapEnrichment function in ChromHMM [48]. Given any two TADs in a cell type, we calculated the Pearson's correlation coefficient between their fold

enrichment values and treated the absolute value of the correlation as the chromatin-state similarity score. In this way, we generated a functional similarity matrix for each cell type.

After that, we clustered TADs based on their similarities at the structural and chromatin-state aspects. We used Spectral Clustering [50] implemented in Scikit-learn [43] as the clustering algorithm as it outperforms the other algorithms (e.g., Affinity Propagation [51]) when dealing with non-convex clusters.

We downloaded protein-coding gene annotations from Ensembl [52] and lncRNA annotations from NONCODE 2016 [33], LNCipedia 4.0 [34], and lncRNAdb 2.0 [35]. Since we use hg19 and mm9 as reference genomes when identifying domain locations, gene data that are inconsistent with the two reference genomes are first converted using liftOver [53] to hg19 human or mm9 mouse genome coordinates. We mapped genes onto TADs for each of the eleven cell types by comparing their genomic positions. For example, if a lncRNA's genomic position has an overlap with a TAD's genomic positions (i.e., start and



end positions), then we labeled this lncRNA to belong to this TAD. The sequence search function was implemented based on BLAST [54].

Utility and discussion

Overview

TADKB has the following main components: browse, family view, acrossCells, search, and download. Detailed description of each component will be presented as follows.

Browsing component

The browse component allows users to select species, cells or cell lines, reference genomes, chromosomes,

resolutions, and domain-caller methods. After a user makes the selection, all the TADs that meet the criteria will be displayed in a list as shown in Fig. 1. The TADs are listed with their starting positions in the chromosome. The ID, start genomic position, end genomic position, and length for each TAD will be displayed. Given two points on a chromosome, TADKB can check whether the two points are in a same TAD.

Once the user clicks one TAD, the main information page of that TAD will be displayed as shown in Fig. 2. This information page contains the Hi-C 2D visualization along with TAD annotations and 1D tracks (gene and various histone modifications from roadmap epigenomics project [49]) via Juicebox.js [55], the

The screenshot displays the TADKB interface for a specific TAD. At the top, the user has selected 'Human - GM12878' as the species and cell line, 'hg19' as the reference genome, '1' as the chromosome, and '1850000_3800000' as the domain position. The main content is divided into two primary sections: a 3D visualization and a list of lncRNAs.

3D visualization in MDS: A 3D plot shows the TAD structure with 'start' and 'end' points. Two lncRNAs are highlighted: NONHSAG000146.2 and NONHSAG000141.2. The visualization is powered by JSmol.

lncRNA list: A list of lncRNAs is shown, with 10 entries displayed. The selected lncRNA, NONHSAG000141.2, is checked. The list includes various NONHSAG IDs.

selected lncRNA info: Two lncRNAs are detailed:

- NONHSAG000141.2:** lncRNA Start: 2138421, lncRNA End: 2139208, Predicted functions: NA, Binding Protein and Class (lncRNATOR): NA. Alternative lncRNAs (LNCipedia): lnc-SKI-4:1.
- NONHSAG000146.2:** lncRNA Start: 2235084, lncRNA End: 2235807, Predicted functions: NA, Binding Protein and Class (lncRNATOR): NA. Alternative lncRNAs (LNCipedia): lnc-RER1-4:1.

Navigation and search options are provided, including a search bar and pagination for the lncRNA list (showing 1 to 10 of 274 entries).

Fig. 6 The TADKB page showing the annotations of lncRNAs. Three major lncRNA databases NONCODE, LNCipedia, and lncRNAdb are integrated. Different IDs from different lncRNA databases will be unified. The locations of the selected lncRNA(s) will be highlighted on the 3D structure of the TAD on the left

reconstructed 3D structures (MDS-based) of the selected TAD, the 3D structure of its chromosome with the selected TAD highlighted (need to click the corresponding tab), the 3D structure of its chromosome in single cells with the selected TAD highlighted (currently only structures for GM12878 are available), the numbers of protein coding genes, the lncRNAs (NONCODE, LNCipedia, and lncRNAdb) existent in the selected TAD, and the loops or peaks detected in the selected TAD which usually indicate promoter-enhancer interactions.

When a user clicks the tab of 3D structure of the chromosome, the 3D structure of the chromosome will be displayed with the selected TAD highlighted. Figure 3 shows an example page of single-cell chromosomal 3D structure. This function allows users to know the 3D location of the selected TAD in the chromosome.

When a user clicks the panel for protein coding gene information, a new page will be displayed as shown in Fig. 4 for MDS-based 3D structure of TAD using population Hi-C and Fig. 5 for single-cell structure of TAD

using single-cell Hi-C. The user can select the coding gene(s) of interest, which will be highlighted in the 3D structure of the TAD. In this way, the user can know whether two genes are spatially proximate. The annotations of selected coding gene(s) will be automatically listed in the panel on the right, which contains: the gene ID in Ensembl, all the transcript IDs, all the protein IDs, description, gene start position, gene end position, and additional information. Once the user clicks additional information, he/she will be redirected to the annotation page on Ensembl.

Once the user clicks the lncRNAs page, all the lncRNAs defined in NONCODE, LNCipedia, and lncRNAdb will be listed. Similarly, when a user selects any lncRNA(s), the annotations will be displayed in the panel on the right as shown in Fig. 6. For each lncRNA, TADKB provides the information of lncRNA start and end locations, predicted functions, binding protein and class predicted by lncRNator [56], exons number, transcripts, and links to the three major lncRNA databases for more details. An important feature of TADKB is that



Fig. 7 The TADKB page showing the loops or peaks. Loops in DNA can indicate the enhancer-promoter interaction

it combines the three different databases for lncRNAs. These three databases have their own scheme of assigning IDs to lncRNAs, which causes inconvenience for biologists to cross-reference the definitions in these databases. In TADKB, the definitions or IDs for the same lncRNA will be combined. The ID from another lncRNA database(s) will be shown in the “Alternative lncRNAs” drop list on the panel on the right. Figure 6 shows the example of a lncRNA in NONCODE that is also overlapped with a lncRNA definition in LNCipedia.

When the user clicks the Loops/Peaks tab, all the peaks will be displayed as shown in Fig. 7. Loops or peaks can indicate enhancer-promoter interactions. The selected peaks will be highlighted in the 3D structure of the TAD. If the user also highlighted coding gene(s) or lncRNA(s) previously, he/she can see whether a peak existed between genes or lncRNAs.

Under the Fold enrichment of chromatin states tab, users can see the fold enrichment of each chromatin

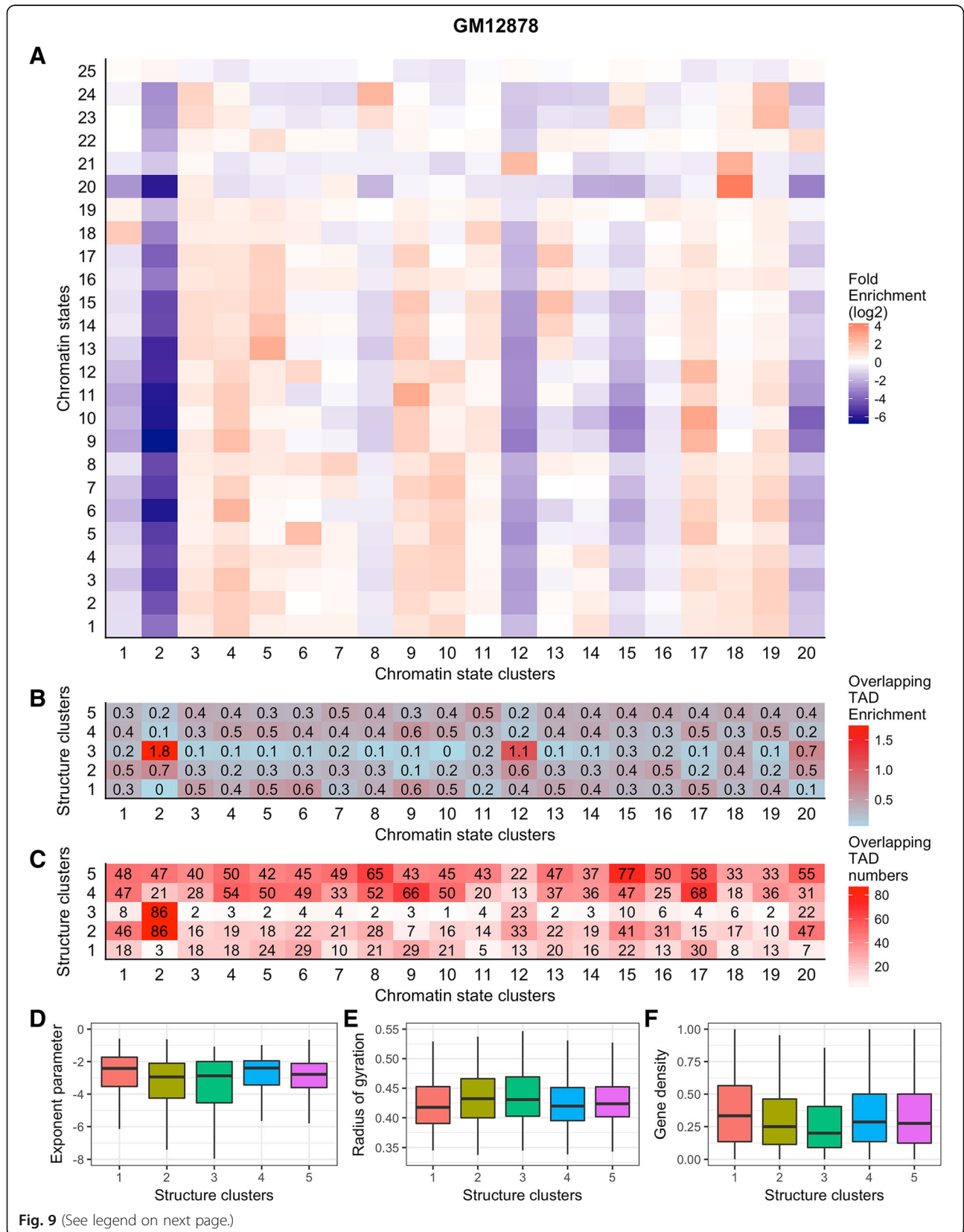
state as shown in Fig. 8. Rows with red color indicate that fold enrichment of that state is larger than one (i.e., enriched for the state), whereas blue color highlights the depleted chromatin states.

TAD family component

As described in the construction and content section, we used spectral clustering algorithm to cluster the TADs in a cell type based on their structural and chromatin-state similarities. Since spectral clustering needs the number of clusters as input, we predefined three numbers of clusters (i.e., 10, 20, and 30) for chromatin-state clustering with Pearson’s correlation between two TADs’ fold enrichments of chromatin states as similarity, and predefined four numbers of clusters (i.e., 2, 3, 5, and 10) for structural clustering with TM-score between two TADs’ MDS-inferred 3D structures as similarity. After obtaining the chromatin-state clusters, we gathered all TADs in a same cluster,



Fig. 8 The TADKB page showing the fold enrichment of chromatin states. Red color indicates fold enrichment larger than 1, otherwise blue color



(See figure on previous page.)

Fig. 9 a Each chromatin-state cluster's fold enrichment of 25 states. **b** The normalized overlapping TAD enrichment between chromatin-state clusters and structural clusters. **c** The original overlapping TAD numbers between chromatin-state clusters and structural clusters. **d, e,** and **f** The distribution of exponent parameters, radius of gyration, and gene density of the TADs in structural clusters. The cell type is GM12878 and the predefined numbers of chromatin-state and structural clusters are 20 and 5, respectively. Gene density is calculated by normalizing the number of protein-coding genes found within a TAD by the TAD's number of bins

computed their fold enrichment for each chromatin state, and found that each cluster has a unique state enrichment pattern. We also found that some clusters are apparently enriched with most of the states (log2 of fold enrichment larger than zero), whereas some other clusters are heavily depleted of chromatin states (log2 of fold enrichment less than zero). An example for GM12878 with the number of chromatin-state clusters equal to 20 can be found in Fig. 9(a), which shows that there are at least three clusters apparently depleted of chromatin states (i.e., clusters 2, 12, and 20).

We compared the clusters of chromatin states and 3D structures and found that there are overlapping TADs, that is, the same TADs were found in both types of

clusters. An example shown in Fig. 9(c) has the numbers of chromatin-state and structural clusters equal to 20 and 5, respectively. We then normalized the number of overlapping TADs by the sizes of the two types of clusters to obtain the overlapping TAD enrichment, which is insensitive to the size of clusters. For example, the number of overlapping TADs between the structural cluster number 1 and chromatin-state cluster number 1 is 18 (see Fig. 9(c)); This value 18 was divided by 167 (the size of chromatin-state cluster number 1) and further divided by 338 (the size of structural cluster number 1), which results in 0.00031 (times 1000 for better visualization); and the final value is 0.3 (see the value in the left-bottom in Fig. 9(b)).

Topologically associating domain (TAD) Families

Human GM12878 select one of cell types

30 select number of clusters based on Pearson correlation between two TADs' chromatin state fold enrichment

10 select number of clusters based on TMscores of two TADs' MDS-based 3D structures

Browse

Topologically associating domain (TAD) families

Show 10 entries Search:

Family ID	Species	Number of TADs	score for TADs' average chromatin state enrichment
1	Human	16	0.080
2	Human	11	0.080
3	Human	16	0.080
4	Human	22	0.080
5	Human	8	0.080
6	Human	28	0.080
7	Human	10	0.080
8	Human	5	0.080
9	Human	9	0.880
10	Human	5	0.880

Showing 1 to 10 of 222 entries Previous 1 2 3 4 5 ... 23 Next

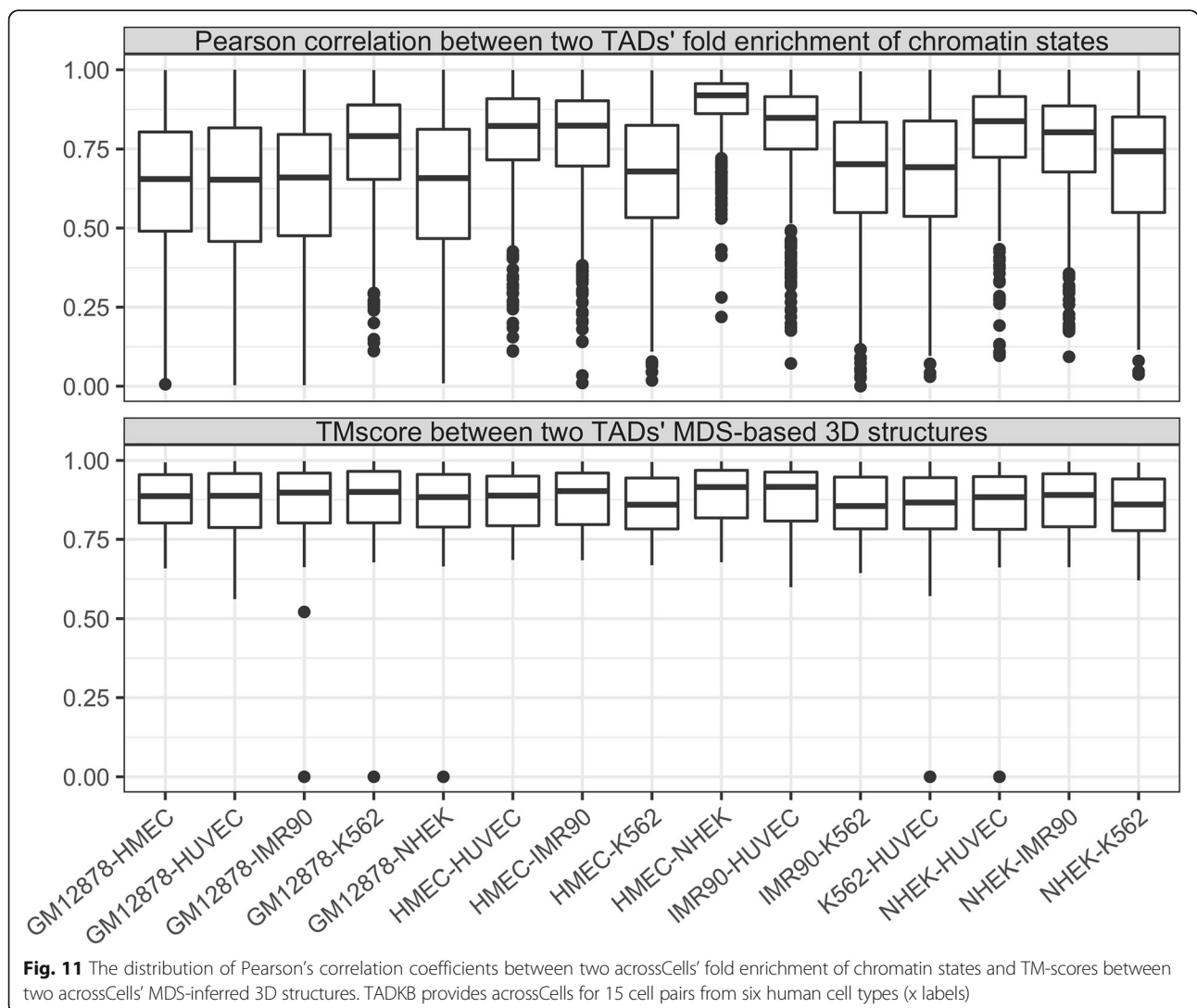
Fig. 10 The family browsing page of TADKB listing all the families of a species. In this example, the families of human GM12878 are listed

From Fig. 9(a) and (b), we observed that most of the TADs in the chromatin-state clusters that are depleted of chromatin states (e.g., 2, 12, and 20) can be found in the second and third structural clusters, especially in the third structural cluster. We tested all the possible number-of-cluster combination configurations (i.e., select one from 10, 20, and 30 as the number of chromatin-state clusters, and select one from 2, 3, 5, and 10 as the number of structural clusters) for TADs detected by DI at 50 kb resolution of the six human cell types, including GM12878, HMEC, HUVEC, IMR90, K562, and NHEK ($6 \times 3 \times 4 = 72$ heat-maps; all can be downloaded from the TADKB website) and observed the same patterns, that is, most of the TADs in the chromatin-state clusters that are depleted of chromatin states can be found in one or two structural clusters, indicating that this observation does not occur by accident. This observation may provide a novel way to

connect TADs' 3D structures with DNA functions indicated by chromatin states.

We plotted the distributions of exponent parameters and radius of gyrations of the mutual TADs overlapped in (1) the structural cluster number 3 and the chromatin-state cluster number 2 (86 mutual TADs), and (2) the structural cluster number 3 and the chromatin-state cluster number 12 (23 mutual TADs) (Fig. 9(d) and (e)) and found that compared with the other mutual sets, the TADs in these two mutual sets have smaller exponent parameters and larger radius of gyrations, which may indicate that these TADs have a less compacted 3D structure and they all have depleted chromatin-state enrichment. We also plotted the gene density distribution (Fig. 9f), showing that the TADs in these two mutual sets have apparently smaller gene density.

We next explored whether our observations are resulted from heterochromatins or gene desert. First, we



downloaded the gap table for hg19 from UCSC genome Table Browser, compared the gaps of heterochromatins and centromeres with the 2773 TADs from GM12878, and found that (1) only 15 TADs (see Additional file 1: Table S2 for details of the 15 TADs) are overlapped with some heterochromatin or centromere regions; (2) only three out of 15 TADs belong to the two structural clusters (clusters 2 and 3 in Fig. 9) with depleted chromatin state enrichment. Therefore, we think our observations are not related to heterochromatins or centromeres. Second, from Fig. 9 we can observe that most of the TADs have positive gene densities, indicating that most of the TADs do not belong to gene desert. Therefore, we think our observations may not be related to gene desert neither.

We listed the chromosomes, coordinates, exponent parameters, and radius of gyrations of the TADs in the overlapping sets between (1) the chromatin-state cluster number 2 and the structural cluster number 3 (Additional file 1: Table S3), (2) the chromatin-state cluster number 12 and the structural cluster number 3 (Additional file 1: Table S4), and (3) the chromatin-state

cluster number 20 and the structural cluster number 3 (Additional file 1: Table S5). We gathered the coding-genes existent in the TADs in these three sets and run a GO enrichment test using AmiGO2 (<http://amigo.geneontology.org/rte>). The enriched GO terms in biological process ontology (BPO), cellular component ontology (CCO), and molecular function ontology (MFO) are also listed in the caption of corresponding Additional file 1.

TADKB defines the overlapping/common TADs between chromatin-state and structural clusters as a family. In this way, both structural and chromatin-state features of TADs are considered when grouping TADs into families. Each family comes with a score assigned to that specific family. For the families constructed based on chromatin states, the score is the percentage of positive values in fold enrichment of each chromatin state (log2). Notice that a smaller score (e.g., <0.5) indicates that on average the corresponding chromatin-state cluster/family is depleted of chromatin states. An example of the details of one family is shown in Fig. 10. We next calculated the average Hi-C heat map for each family. Since the sizes of TADs (number of bins) in each family

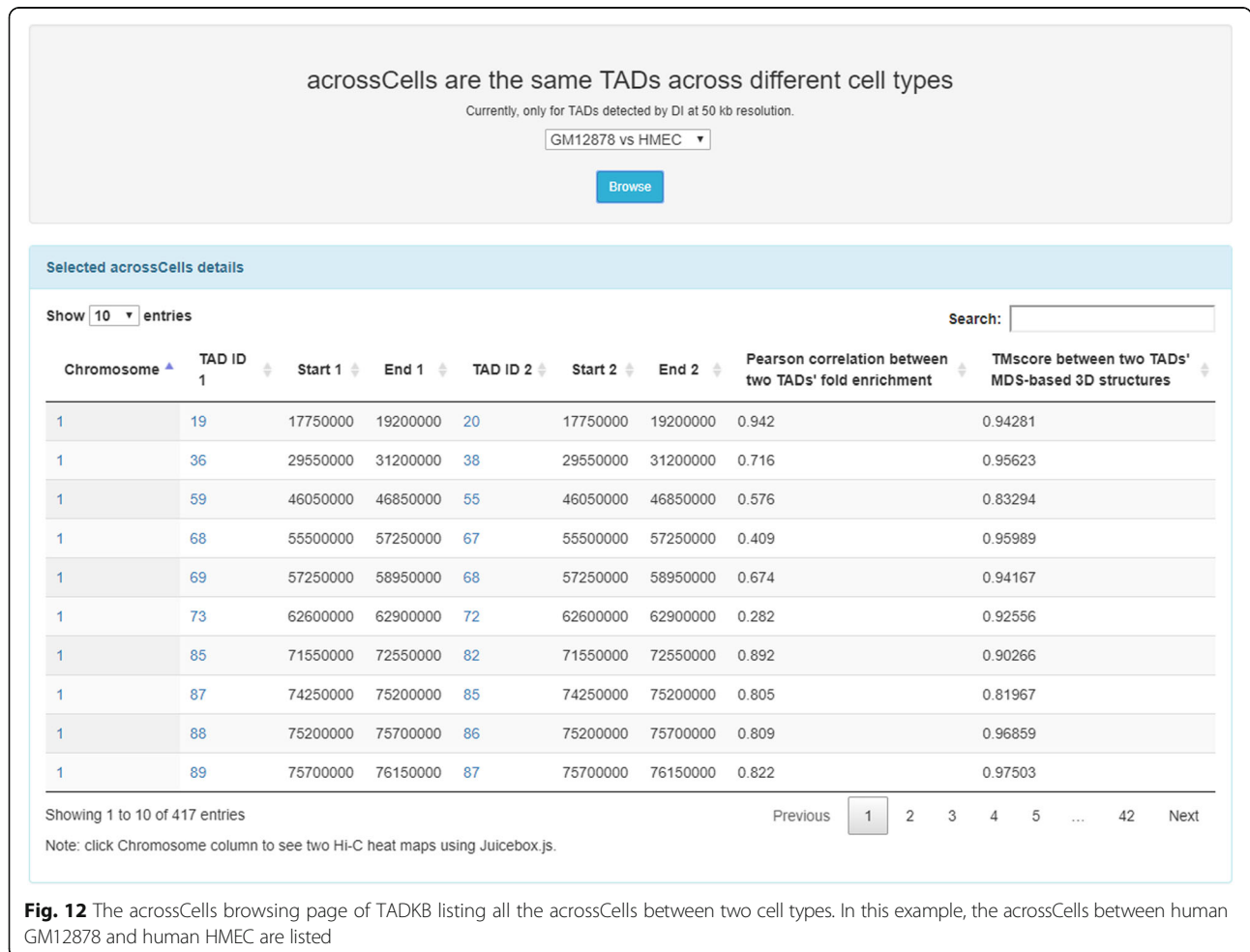


Fig. 12 The acrossCells browsing page of TADKB listing all the acrossCells between two cell types. In this example, the acrossCells between human GM12878 and human HMEC are listed

vary, we cannot directly calculate the average Hi-C matrix for a family. Therefore, for each TAD we extracted a 30×30 Hi-C submatrix from the large matrix of the whole chromosome by evenly extending the sizes of smaller TADs (< 30 bins) and evenly reducing the sizes of bigger TADs (> 30 bins). After that, all TADs' Hi-C matrices were with the same sizes (30×30) and we then calculated the average Hi-C heatmaps. The average heatmap (log2 scale) can be found under *Table of family members* on the Family webpage. An example of the average heatmaps of three families can be found in Additional file 1: Figure S1, which shows that they have different patterns in terms of their average Hi-C contact matrices.

acrossCells component

We defined acrossCells as the set of two TADs from the same species that are in different cell lines but exist in the same chromosome and with the same coordinates (start and end positions in the chromosome). We provided acrossCells of 15 cell-pairs among the six cell types in human with available chromatin-state annotations. For each pair of acrossCells, we computed the Pearson's correlation coefficient between their fold enrichment of 25 chromatin states and the TM-score between their MDS-inferred reconstructed 3D structures. The distribution of these

two similarity measures can be found in Fig. 11, which shows that the acrossCells found in HMEC and NHEK have very similar enrichment pattern of chromatin states, and acrossCells always have very high TM-scores. An example of the TADKB webpage showing the acrossCells between GM12878 and HMEC can be found in Fig. 12. Users can also browse the two dynamic Hi-C heatmaps side by side for TAD pairs in acrossCells by clicking the chromosome column in the selected acrossCells TADs table. An example figure can be found in Additional file 1: Figure S2.

TAD search component

A user can submit gene (protein coding gene or lncRNA) names, IDs, or query DNA sequences; and TADKB will search against in-house sequence sets and provide matched genes and their associated TADs. Figure 13 shows the web page of searching. If there are hits found, TADKB will display the TADs that contains the hit sequences as shown in Fig. 14. A user can then further click one of the TADs and then browse detailed information of it.

Downloading component

The download component allows users to download 90 TAD annotation files as described in Additional file 1:

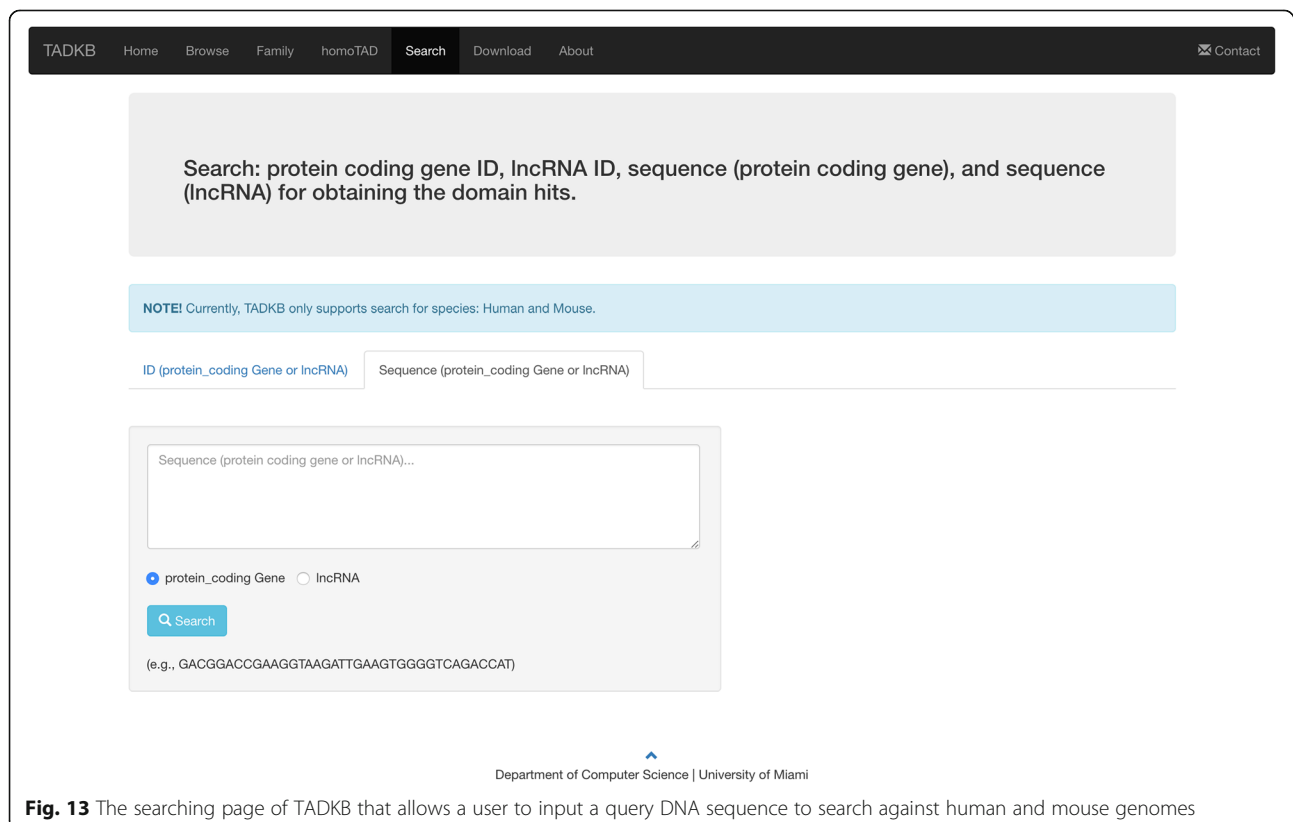


Fig. 13 The searching page of TADKB that allows a user to input a query DNA sequence to search against human and mouse genomes

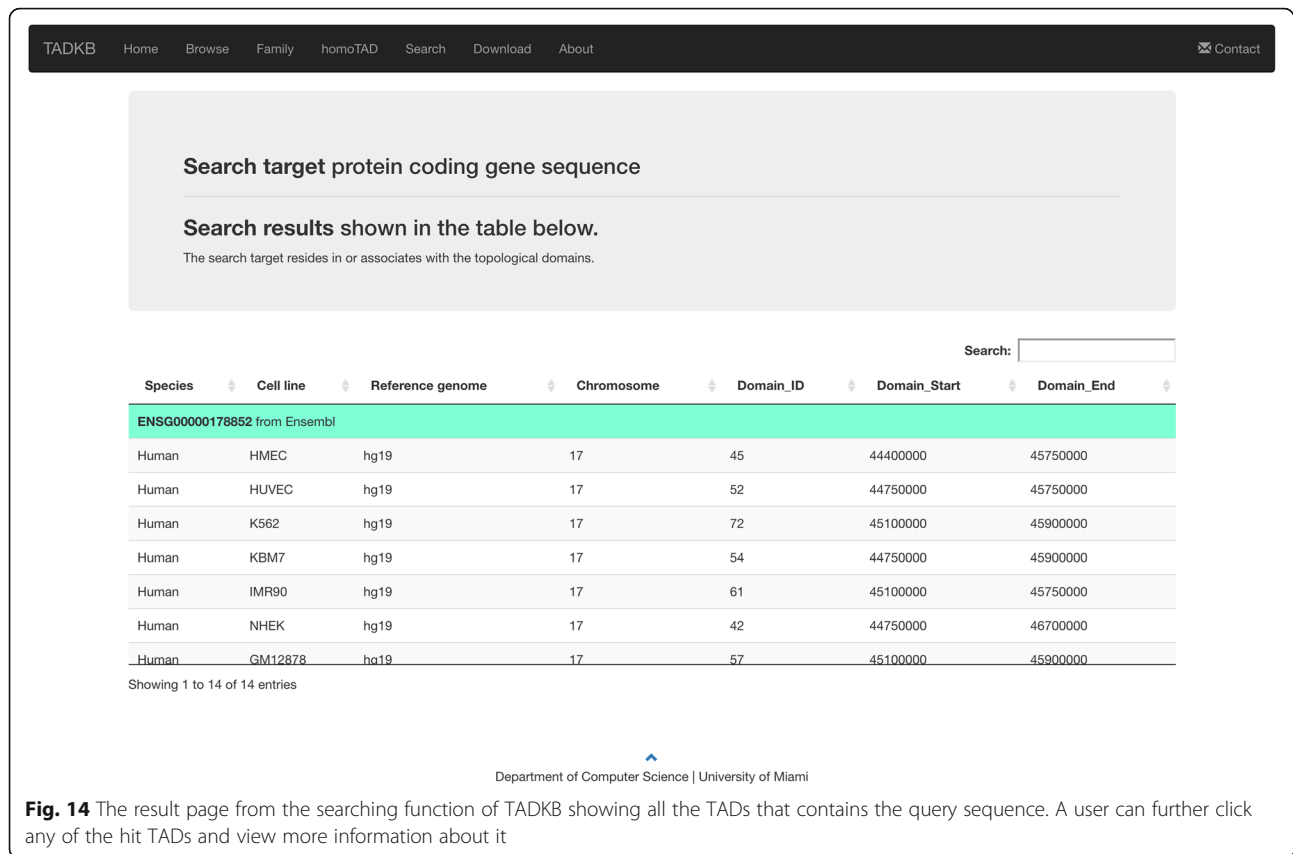


Table S1 and 72 heatmaps about the overlapping TAD analysis between chromatin-state and structural clusters.

Conclusion

TADKB, a database for topologically associating domains, has been built that integrates the 2D and 3D structures of TAD, 3D structure of chromosome, annotations of coding genes and lncRNAs, loops or peaks, and family classification of TADs. TADKB allows users to view the genomic locations of coding gene, lncRNAs, and loops on the 3D structure of TAD. It also integrates three major lncRNA databases so that the different IDs from different lncRNA databases can be unified. The TAD families in TADKB are defined as the overlapping TADs found in chromatin-state and structural clusters. We also found that most of the TADs in depleted chromatin-state clusters also exist in one or two structural clusters; and these TADs mostly have smaller exponent parameter and larger radius of gyration. TADKB provides a convenient searching function so that based on a query DNA sequence the TADs that contains the hits of the query sequence will be outputted. Based on the other TADs within the same family of the hit TAD(s), more annotations may be provided for the query sequence. The role that lncRNA plays in forming up chromosome 3D structures is not yet clear or

determined. However, lncRNAs have been eventually found to be playing an important role in either assisting or regulating many important DNA functions although lncRNAs had been originally considered not functioning at all in the genome. Therefore, the annotations of lncRNAs are also integrated into TADKB.

Additional file

Additional file 1: Figure S1. The 30 × 30 average Hi-C contact matrices calculated from all members in a TAD family. The family is from GM12878 with 20 predefined chromatin-state clusters and five predefined structural clusters. **Figure S2.** An example of two Hi-C heat maps for TAD pairs in acrossCells. **Table S1.** Number of TADs detected by three different domain-caller methods at the resolutions of 50 kb and 10 kb for Hi-C, HiChIP, and SPRITE data. **Table S2.** The details of the 15 TADs overlapped with heterochromatins or centromeres. **Table S3.** The TADs are from the overlapping between the second chromatin-state cluster and the third structural cluster in Fig. 8. The genes from the following TADs are enriched for GO terms: GO:0044278 'cell wall disruption in other organism', GO:1905874 'regulation of postsynaptic density organization', GO:1905606 'regulation of presynapse assembly', and GO:1905606 'regulation of presynapse assembly from BPO; GO:0016342 'catenin complex', GO:0099061 'integral component of postsynaptic density membrane', and GO:0005913 'cell-cell adherens junction' from CCO; GO:0005004 'GPI-linked ephrin receptor activity', GO:0005003 'ephrin receptor activity', and GO:0030594 'neurotransmitter receptor activity' from MFO. The overrepresentation test was generated with PANTHER. **Table S4.** The TADs are from the overlapping between the twelfth chromatin-state cluster and the third structural cluster in Fig. 8. The genes from the following TADs are enriched for GO terms: GO:0050911

'detection of chemical stimulus involved in sensory perception of smell', GO:0050907 'detection of chemical stimulus involved in sensory perception', and GO:0007608 'sensory perception of smell' from BPO; GO:0005886 'plasma membrane' and GO:0016021 'integral component of membrane' from CCO; and GO:0005549 'odorant binding', GO:0004984 'olfactory receptor activity', and GO:0004930 'G protein-coupled receptor activity' from MFO. The overrepresentation test was generated with PANTHER. (PDF 332 kb)

Abbreviations

2D: Two dimensional; 3D: Three dimensional; lncRNA: Long non-coding RNA; TAD: Topologically associating domain

Acknowledgements

Not applicable.

Funding

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R15GM120650 to ZW and start-up funding from the University of Miami to ZW. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the University of Miami.

Availability of data and materials

TADKB can be freely accessed at <http://dna.cs.miami.edu/TADKB/>.

Authors' contributions

TL generated most of the knowledge in the website and designed and built the website. TL acquired most of the data, analyzed and interpreted the data, and generated the figures in the manuscript. JP participated in the design and building of the database at the early stage of development. CZ acquired the data related to protein functions, conducted functional analysis of TADs, and wrote related parts of the manuscript. HZ acquired the single-cell Hi-C data, generated the three-dimensional structures of chromosomes based on single-cell Hi-C data, and wrote related part of the manuscript. TL and ZW drafted the manuscript. NW, ZS, and YM provided scientific advices and participated in the design of the database. All authors edited the manuscript. ZW conceived and advised the research. All authors have read and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Computer Science, University of Miami, 1365 Memorial Drive, Coral Gables, FL 33124-4245, USA. ²School of Computing Sciences and Computer Engineering, University of Southern Mississippi, 118 College Drive, Hattiesburg, MS 39406, USA. ³Department of Computer Science, New Jersey City University, 2039 Kennedy Blvd, Jersey City, NJ 07305, USA. ⁴Department of Electrical and Computer Engineering, California Baptist University, 3739 Adams Street, Riverside, CA 92504, USA. ⁵Department of Pharmacology and Toxicology, University of Mississippi Medical Center, 2500 N State St, Jackson, MS 39216, USA.

Received: 11 September 2018 Accepted: 21 February 2019

Published online: 14 March 2019

References

- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485(7398):376–80.
- Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;159(7):1665–80.
- Dixon JR, Gorkin DU, Ren B. Chromatin domains: the unit of chromosome organization. *Mol Cell*. 2016;62(5):668–80.
- Zuin J, Dixon JR, van der Reijden MI, Ye Z, Kolovos P, Brouwer RW, van de Coput MP, van de Werken HJ, Knoch TA, van Ucken WF. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci*. 2014;111(3):996–1001.
- Rudan MV, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjir S. Comparative hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep*. 2015;10(8):1297–309.
- Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*. 2012;148(3):458–72.
- Chen Y, Wang Y, Xuan Z, Chen M, Zhang MQ. De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. *Nucleic Acids Res*. 2016;44(11):e106.
- Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. *Algorithms Mol Biol*. 2014;9(1):14.
- Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing hi-C data. *Bioinformatics*. 2014;30(17):i386–92.
- 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(4):544–57.
- Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, Ong C-T, Hookway TA, Guo C, Sun Y. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*. 2013;153(6):1281–95.
- Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res*. 2015;44(7):e70.
- Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. *Bioinformatics*. 2015;32(11):1601–9.
- Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326(5950):289–93.
- Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J, Wingett SW, Várnai C, Thiecke MJ. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*. 2016;167(5):1369–84 e1319.
- Duan Z, Andronescu M, Schutz K, Mcllwin S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. A three-dimensional model of the yeast genome. *Nature*. 2010;465(7296):363–7.
- Zhang Z, Li G, Toh K-C, Sung W-K. 3D chromosome modeling with semi-definite programming and hi-C data. *J Comput Biol*. 2013;20(11):831–46.
- Trieu T, Cheng J. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Res*. 2014;42(7):e52.
- Varoquaux N, Ay F, Noble WS, Vert J-P. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*. 2014;30(12):i26–33.
- Liu T, Wang Z. Reconstructing high-resolution chromosome three-dimensional structures by hi-C complex networks. *BMC Bioinformatics*. 2018;19(Suppl 17):496.
- Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*. 1964;29(1):1–27.
- Lvd M, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9(Nov):2579–605.
- Ramani V, Deng X, Qiu R, Gunderson KL, Steemers FJ, Disteche CM, Noble WS, Duan Z, Shendure J. Massively multiplex single-cell hi-C. *Nat Methods*. 2017;14(3):263–6.

24. Liu T, Wang Z. sChICNorm: a software package to eliminate systematic biases in single-cell hi-C data. *Bioinformatics*. 2018;34(6):1046–7.
25. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F. Landscape of transcription in human cells. *Nature*. 2012;489(7414):101–8.
26. Monnier P, Martinet C, Pontis J, Stancheva I, Ait-Si-Ali S, Dandolo L. H19 lncRNA controls gene expression of the imprinted gene network by recruiting MBD1. *Proc Natl Acad Sci*. 2013;110(51):20693–8.
27. Schuldt A. Gene expression: an ncRNA relocation package. *Nat Rev Mol Cell Biol*. 2011;13(1):1–1.
28. Lee JT. Epigenetic regulation by long noncoding RNAs. *Science*. 2012; 338(6113):1435–9.
29. Morlando M, Ballarino M, Fatica A, Bozzoni I. The role of long noncoding RNAs in the epigenetic control of gene expression. *ChemMedChem*. 2014; 9(3):505–10.
30. Maia BM, Rocha RM, Calin GA. Clinical significance of the interaction between non-coding RNAs and the epigenetics machinery. *Epigenetics*. 2014;9(1):75–80.
31. Magistri M, Faghihi MA, St Laurent G III, Wahlestedt C. Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts. *Trends Genet*. 2012;28(8):389–96.
32. Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science*. 2013;341(6147):1237973.
33. Zhao Y, Li H, Fang S, Kang Y, Hao Y, Li Z, Bu D, Sun N, Zhang MQ, Chen R. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res*. 2015. <https://doi.org/10.1093/nar/gkv1252>.
34. Volders P-J, Verheggen K, Menschaert G, Vandepoel K, Martens L, Vandesompele J, Mestdagh P. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res*. 2015;43(D1):D174–80.
35. Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. lncRNADB v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res*. 2015;43(Database issue):D168–73.
36. Bonev B, Cohen NM, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot J-P, Tanay A. Multiscale 3D genome rewiring during mouse neural development. *Cell*. 2017;171(3):557–72 e524.
37. Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian mixture model and proportion test. *Nat Commun*. 2017;8(1):535.
38. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*. 2015;523(7559):240.
39. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods*. 2016;13(11):919.
40. Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, Lai MM, Shishkin AA, Bhat P, Takei Y et al. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell*. 2018;174(3):744–57 e724.
41. Knight PA, Ruiz D. A fast algorithm for matrix balancing. *IMA J Numer Anal*. 2013;33(3):1029–47.
42. Li W, Gong K, Li Q, Alber F, Zhou XJ. Hi-corrector: a fast, scalable and memory-efficient package for normalizing large-scale hi-C data. *Bioinformatics*. 2015;31(6):960–2.
43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12(Oct):2825–30.
44. Kabach W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr Sect A: Cryst Phys, Diffr, Theor Gen Crystallogr*. 1978;34(5):827–8.
45. Liu T, Wang Z. Measuring the three-dimensional structural properties of topologically associating domains. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): 2018: IEEE; 2018: 21–28.
46. Tan L, Xing D, Chang C-H, Li H, Xie XS. Three-dimensional genome structures of single diploid human cells. *Science*. 2018;361(6405):924–8.
47. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 2005;33(7):2302–9.
48. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 2012;9(3):215.
49. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317.
50. Shi J, Malik J. Normalized cuts and image segmentation. *IEEE T Pattern Anal*. 2000;22(8):888–905.
51. Frey BJ, Dueck D. Clustering by passing messages between data points. *Science*. 2007;315(5814):972–6.
52. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L. Ensembl 2016. *Nucleic Acids Res*. 2015; 44(D1):D710–6.
53. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F. The UCSC genome browser database: update 2006. *Nucleic Acids Res*. 2006;34(suppl 1):D590–8.
54. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25(17):3389–402.
55. Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst*. 2018;6(2):256–8 e251.
56. Park C, Yu N, Choi I, Kim W, Lee S. lncRNAtor: a comprehensive resource for functional investigation of long noncoding RNAs. *Bioinformatics*. 2014. <https://doi.org/10.1093/bioinformatics/btu325>.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

