BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–22

Supplementary Material for Spectral Identification of Topological Domains

Jie Chen, Alfred O. Hero III, Indika Rajapakse

1 USER GUIDE FOR THE MATLAB TOOLBOX

We implement the algorithm with Matlab and publish this set of code as a public available toolbox. Researchers can use this code to identify topological domains with the proposed algorithms. The toolbox can be downloaded at http://www.jie-chen.com/codes.html. The files in the toolbox are listed in Table S1.

Table S1. Files in the toolbox folder

File name	File type	Description
TAD_Laplace	script (function)	The main code of the proposed algorithms
Draw_TADs	script (function)	Display the Hi-C matrix with identified topological domains
sumDiag	script (function)	Compute the diagonal summations for the matrix
Test_Algorithm	script	Script that call the functions to illustrate the algorithms
HiC_chr22_100kb	data	Matlab data file that stores the Hi-C matrix of chromosome 22 at 100kb
TADs_DP	folder	Dynamic programming algorithm for comparison. See and run Run_Test_Script.m in the folder for testing it
TADs_HMM	folder	Hidden Markov model algorithm for comparison. See and run Run_Test_Script.m in the folder for testing it

1.1 Function interface

1.1.1 TAD_Laplace This is the main function to perform the proposed algorithms. The function takes the inputs and gives the output as follows:

TAD_Boundaries = TAD_Laplace(H0, thr, ms, MERG)

where

- H0: The input Hi-C matrix. Requirement: H0 should be square and symmetric.
- thr: The Fiedler value threshold λ_{thr} in the paper. Default value: 0.8.
- ms: Minimum splitting size. The algorithm will not split a block whose size is not larger than ms. Default value: 2.
- MERG: Single bin merging flag. The algorithm may result in over segmentation of domains into single bins. Activate this option (set MERG = 1) if users intend to merge these single bins to their adjacent domains. These bins will be merged with their upstream or downstream bins depending on which has more contacts. Default value: 0.

The output "TAD_Boundaries " contains the positions of domain boundaries.

1.1.2 Draw_TADs This function displays the Hi-C matrix and identified topological domains. It takes the inputs as follows:

Draw_TADs(Ht, TAD_Boundaries, CLIM)

where

- Ht: Transformed Hi-C matrix.
- Boundaries: Boundaries of topological domains.
- CLIM = [CLOW CHIGH] can specify the scaling for the called Matlab function *imagesc*.

1.2 Examples

Before calling the function, let us load the example data in the folder by

```
load HiC_chr22_100kb
```

The data is then loaded in the variable H.

We then remove the unmappable regions by discarding the rows and columns without reads:

```
idx = (sum(H)~=0);
H = H(idx,idx);
```

Function calling example 1: only the matrix is provided for the function. All parameters will use the default values.

```
TAD_Boundaries = TAD_Laplace(H);
```

Function calling example 2: we provide the parameters to the function.

```
TAD_Boundaries = TAD_Laplace(H, 0.9, 1, 1);
```

where we perform the algorithm on H with Fiedler vector threshold $\lambda_{\mathrm{thr}} = 0.9$, minimum splitting size 2 bins, and emerging single bin domain to neighboring domains.

Finally, we display results by

```
Draw_TADs(log(H), TAD_Boundaries, [-3,6]);
```

The file "Test_Algorithm" shows a complete example.

2 CELL CULTURE AND HI-C INFORMATION OF FIBROBLASTS

2.1 Cell Culture and Crosslinking of Chromatin

Human foreskin fibroblasts from a normal karyotyped male individual (Cat # CRL-2522, ATCC, Manassas, VA) were propagated in growth medium, composed of MEM medium (cat #11095-098, Life Technologies, Grand Island, NY), 10% fetal bovine serum (FBS, cat #10082-147, Life Technologies), 1 X non-essential amino acids (NEAA, cat #11140-050, Life Technologies), and 1 X antibiotic/antimycotic (cat #15240-062, Life Technologies).

To prepare cultures for time series Hi-C and RNA-seq experiments, we trypsinized fibroblasts propagated in T225 flasks with 0.25% trypsin (Cat #25200-056, Life Technologies). Dissociated cells were plated in 150mm cell culture plates for Hi-C experiments, and in 6-well plates for RNA-seq experiments. Cells were cultured in growth medium for 36 hours before proceeding to cell cycle synchronization. To synchronize the cell cycle, the cells were incubated in serum-free MEM medium supplemented with 1 X NEAA and 1 X Antibiotic/antimycotic for 48 hours. Under this condition, it has been reported that more than 95% of the cells are in G0/G1 phase.

We also synchronized the circadian clock in the cell cycle-synchronized cultures with dexamethasone treatment for one hour. At the 48th hour of cell cycle synchronization, we added dexamethasone to the cultures at a final concentration of 100nM for the treatment group. For the control group, we added ethanol to cultures at a final concentration of 0.001% (dexamethasone solvent concentration in the treatment cultures). At the end of circadian clock synchronization, cells were rinsed twice with PBS, fed with growth medium, and time point zero was set.

Base line samples from dexamethasone treated plates and control plates were taken at the end of the 1-hour circadian clock synchronization without feeding of growth medium. Thereafter, time series sampling of cells treated with dexamethasone was performed at 8-hour intervals for a total of 56 hours. All samples for Hi-C experiments were cross linked in situ. Total RNA was extracted directly from 6-well plates (see RNA isolation and RNA-seq).

Approximately 25 X 10^6 cells were cross linked with 1% formaldehyde (Cat #BP531-25, Fisher Scientific, Pittsburgh, PA) in serum free-medium for 10 min at room temperature, and then quenched with glycine (Cat #G8898-500g, Sigma-Aldrich, St. Louis, MO) added to a final concentration of 0.125 M. Cross linked cells were harvested and flash frozen in liquid nitrogen, and then stored at -80^{o}C until the construction of Hi-C libraries.

2.2 Generation of Hi-C Libraries

We adopted the methods for Hi-C library construction from Belton et al. (Methods 58:268, 2012). For each Hi-C library, approximately 25 X 10^6 cells were resuspended in 1mL ice-cold lysis buffer, consisting of 10mM Tris-HCl, 10mM NaCl, 0.2% Igpel (Cat # 8896-50mL, Sigma-Aldrich), and 10 mL protease inhibitor cocktail (Cat # P8340-1ml, Sigma-Aldrich). All resuspended cells were incubated on ice for 15 min. Cells were homogenized in a Dounce homogenizer on ice with pestle A, and the lysate was transfer to a 1.7mL tube. Cells were collected by spinning for 5 minutes at 2,000xg, and then washed twice in 500 μL of ice cold 1x NEB buffer 2. Cells were distributed between

4 individual 1.7 ml centrifuge tubes (50 μL per tube). Chromatins in each tube were digested with 400u of restriction enzyme HindIII (Cat # R0104M, New England BioLabs, Ipswich, MA) in 1x NEB buffer 2 at 37^{o}C overnight on a spin wheel.

After HindIII digestion, restriction site overhanging ends were filled and labeled with biotin using DNA polymerase I large (Klenow) fragment (Cat # M0210L, New England BioLabs) in a reaction containing dATP, dGTP, dTTP, and biotin-14-dCTP (Cat # 19518-018, Life Technologies) in each of the 4 HindIII digestion tube. DNA fragments labeled biotin-14-dCTP from each of the 4 tubes were ligated at 16^{o}C for 4 hours in an 8.23 mL reaction containing 1x ligation buffer, 1% Triton X-100 (Cat # T8787-250ML, Sigma-Aldrich), 1 mg/ml bovine serum albumin (BSA)(Cat # BP9706-100, Fisher Scientific), 10 mM ATP (Cat # A9187-1g, Sigma-Aldrich), and 50u T4 DNA ligase (Cat # 15224-025, Life Technologies). Reverse cross linking was performed in two steps. First, 50 μL of 10 mg/ml proteinase K (Cat # 25530-015, Life Technologies) were added to each ligation reaction tube and incubated at 65^{o}C for 4 hours. Then, another 50 μL of proteinase K were added to each tube and continued incubating at 65^{o}C overnight. Next, DNA was extracted with saturated phenol : chloroform (1:1) (Cat # 1100631, Fisher Scientific), and desalted by using AMICON® Ultra Centrifugal Filter Unit (Cat # UFC503024, Millipore, Billerica, MA) with 1 x TE buffer. The final volume of desalted DNA was adjusted to 100 μL in 1 x TE buffer.

Removal of Biotin from un-ligated ends was carried out in 8 individual reactions each of 50 μL containing 5 μg of Hi-C DNA, 1 mg/ml BSA,1X NEB buffer 2, 25 nM dATP, 25 nM dGTP, and 15 u T4 DNA polymerase at 20^{o}C for 4 hours. The Hi-C DNA was then pooled in a single tube, purified with single phenol extraction, and precipitated by ethanol. The DNA was re-dissolved in 105 μL of water, and transferred to a microTUBE AFA tube (Cat # 520045, Covaris, Woburn, Massachusetts). DNA fragmentation was performed in a Sonicator (Covaris S2, Covaris). The DNA fragments of size 200–400 bp were recovered with Agencourt AMPure XP mixture (Cat # A63880, Beckman Coulter, Indianapolis IN) following the manufacturer's protocols.

DNA fragment ends were repaired in a 70 μL reaction containing 1 X ligation buffer (Cat # B0202, New England BioLabs), 0.25 mM of dNTP mixture, 7.5 u of T4 DNA polymerase (Cat # M0203L, New England BioLabs), 25 u of T4 polynucleotide kinase (Cat # M0201S, New England BioLabs), 2.5 u of DNA polymerase I large fragment at 20^{o}C for 30 min. The reaction is purified with a MinElute column (Cat 28204, Qiagen, Valencia, CA). The DNA is eluted in 32 μL of elusion buffer for A-tailing, which was performed in a 50 μL reaction containing purified DNA (5 μg), 1 X NEB buffer 2, 0.2 mM dATP, 15 u Klenow fragment (3'$\rightarrow$5' exo-) (Cat # M0212L, New England BioLabs). The reaction was incubated at 37^{o}C for 30 min, then at 65^{o}C for 20 min to inactivate Klenow (exo-).

For Streptavidin pull-down of biotinylated Hi-C ligation products, the biotinylated Hi-C ligation products are mixed with MyOne C1 streptavidin bead solution (Cat # 65001, Life Technologies) for binding of biotinylated Hi-C fragments. Non-specifically binding DNA was removed by washing with 1 X binding buffer (5 mM Tris-HCl (pH8.0), 0.5 mM EDTA, and 1 M NaCl), then with 1 X T4 Ligation buffer (Cat # 46300-018, Life Technologies). The DNA-bound beads were resuspended in 38.75 μL of 1 X ligation buffer for adapter ligation.

Illumina adapter ligation was performed in a 50 μL reaction by adding to the DNA-bound beads suspension of 1 X T4 ligation buffer, 90 pM of Illumina paired end adapter, 3 u of T4 DNA ligase (Cat # 15224-025, Life Technologies). The reaction was incubated at room temperature for 2 hours. The beads were reclaimed, and the supernatant discarded. The beads were washed twice in Tween Wash Buffer (5 mM Tris-HCl (pH 8.0), 0.5 mM EDTA, 1 M NaCl, 0.05% Tween 20), and once in 1 X binding buffer (5 mM Tris-HCl (pH 8.0), 0.5 mM EDTA, and 1 M NaCl), and twice in 1 X NEB buffer 2. After the last wash, the beads were resuspended in 20 μL of 1X NEB buffer 2.

The Hi-C DNA sample was amplified by 15 PCR cycles (optimized in the log amplification phase) for Illumina HiSeq sequencing. Each PCR reaction in 25 μL, 1.5 μL of Bead-bound Hi-C DNA, 0.35 μL of PE primer 1.0, 0.35 μL of PE primer 2.0, 0.2 μL of 25mM dNTP, 2.5 μL of 10X PfuUltra buffer, 19.6 μL of H_2O, and 0.5 μL of PfuUltra Fusion DNA polymerase. The PCR cycling parameters were 98^{o}C for 30 seconds, followed by 15 cycles at 98^{o}C for 10 seconds, 65^{o}C for 30 seconds, and 72^{o}C for 30 seconds, and a final extension at 72^{o}C for 7 minutes. PCR products pooled from the supernatant of multiple reactions were subjected to AMpure XP beads purification to remove primer dimers. A standard quality control (QC) procedure was performed on the purified PCR products (Hi-C library). Each Hi-C library passed the QC procedure and was then sequenced in a single lane of a flow cell on a HiSeq 2000 sequencer to generate paired-end sequence reads at 100 bases per end read.

2.3 Generation of Hi-C Matrices

We standardized a pipeline to process Hi-C sequence data at the University Bioinformatics Core facilities. With this pipeline, raw sequence reads were processed with FastQC (`http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/`) for data quality control. Paired-end reads with excellent quality were mapped to the reference human genome (HG19) using Bowtie2 with default parameter settings and the "–very-sensitive-local" preset option, which produced a SAM formatted file for each member of the read pair (R1 and R2). HOMER (`http://homer.salk.edu/homer/interactions/`) was used to develop the contact matrix with "makeTagDirectory" with the "–tbp 1" setting, and with "analyzeHiC" with the "-raw" and "-res 1000000" settings to produce the raw contact matrix at 1MB resolution.

2.4 RNA Isolation and RNA-seq

We used TRIzol® Reagent (Cat # 15596-018, Life Technologies) to extract RNA from cells grown in 6-well plates. All extracted RNA was treated with RNAse-free DNaseI (Cat # 79254, Qiagen), then submitted to the University of Michigan Bioinformaics Core lab for library construction and RNA sequencing (RNA-seq) on the Illumina Hiseq-2000 platform. Single-end 50-base sequence reads were generated at a multiplex of 4 per sequencing lane.

Table S2. Identified TAD numbers

Proposed with $\lambda = 0.8$		Proposed with $\lambda = 0.9$		HMM with 100kb data		HMM with 40kb data (Dixon)	
chr	number	chr	number	chr	number	chr	number
1	300	1	388	1	44	1	237
2	272	2	359	2	84	2	187
3	281	3	319	3	69	3	159
4	227	4	265	4	69	4	133
5	233	5	267	5	61	5	145
6	207	6	263	6	39	6	131
7	193	7	242	7	53	7	119
8	202	8	242	8	39	8	111
9	220	9	240	9	45	9	129
10	179	10	227	10	31	10	100
11	165	11	207	11	40	11	104
12	170	12	202	12	44	12	103
13	122	13	146	13	32	13	62
14	107	14	128	14	31	14	66
15	89	15	123	15	8	15	68
16	103	16	136	16	18	16	88
17	97	17	139	17	18	17	81
18	88	18	106	18	23	18	55
19	74	19	104	19	15	19	64
20	80	20	92	20	12	20	60
21	49	21	64	21	9	21	34
22	34	22	45	22	18	22	27
Total	3492	Total	4305	Total	802	Total	2241

During RNAseq data processing, we checked the raw reads with FastQC (version 0.10.0), to identify potential quality problems in the reads data (eg. low quality scores, over-represented sequences, inappropriate GC content). We used Tophat (version 2.0.9) and Bowtie (version 2.1.0.0) to align the reads to the reference transcriptome (HG19). We used default parameter settings for alignment, with the exception of: "–b2-very-sensitive" telling the software to spend extra time searching for valid alignments, as well as "–no-coverage-search" and "–no-novel-juncs" to limit the search to known transcripts. We then performed a second round of quality assessment using FastQC on the aligned reads. Data was found to be of excellent quality overall. We used Cufflinks/Cuffdiff (version 2.1.1) for expression quantification and differential expression analysis, using UCSC hg19.fa as the reference genome and UCSC hg19.gtf as the reference transcriptome. For the CuffDiff analysis, we used parameter settings: "–multi-read-correct" to adjust expression (FPKM) calculations for reads that map in more than one locus, as well as "–compatible-hits-norm" and "–upper-quartile-norm" for normalization of expression calculations across samples. We used a locally developed Perl script to format the Cufflinks output.

3 IDENTIFIED NUMBER OF TADS

We report the identified number of TADs via the proposed method and the HMM method in Table S2. It can be found that the proposed methods identified more TADs than the HMM method. In fact, even the first splitting step that segments the chromosome by overall organization leads to 3101 domains. This suggests that it is reasonable that the total number of TADs is larger than this number. The coordinates of the identified TAD boundaries are reported in Table S3.

4 SUPPLEMENTAL MATERIALS: COMPARISON BETWEEN THE FIEDLER VECTOR AND THE LEADING EIGENVECTOR FOR THE FIRST LAYER SEGMENTATION.

At the first step of our algorithm, the Fiedler vector is computed to segment the Hi-C matrix into two compartments. This operation is similar to the A/B segmentation by Liebermann-Aiden. Both of the leading eigenvector and Fiedler vector characterize the structure of the graph, while the latter is a relaxed solution to the normalized graph segmentation problem. They do provide similar results, and the latter is more widely used in the machine learning community. A comparison of their calculation steps is given by Fig. 1. The correlation between the signed version of these two vectors are given by Fig. 2(a). It can be observed that they have high similarities with high correlation values. Further, the comparison of their correlations with the RNAseq counts is shown in Fig. 2(b) (using the same way as for Fig. 4 in the main document). Higher correlation values are observed for the Fiedler vector for most chromosomes.

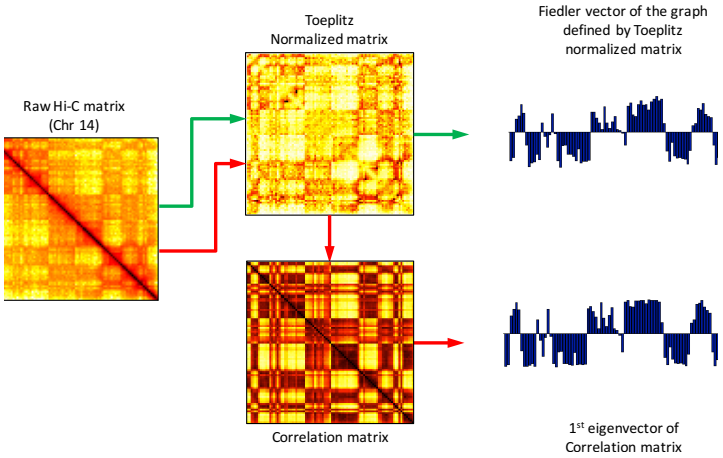
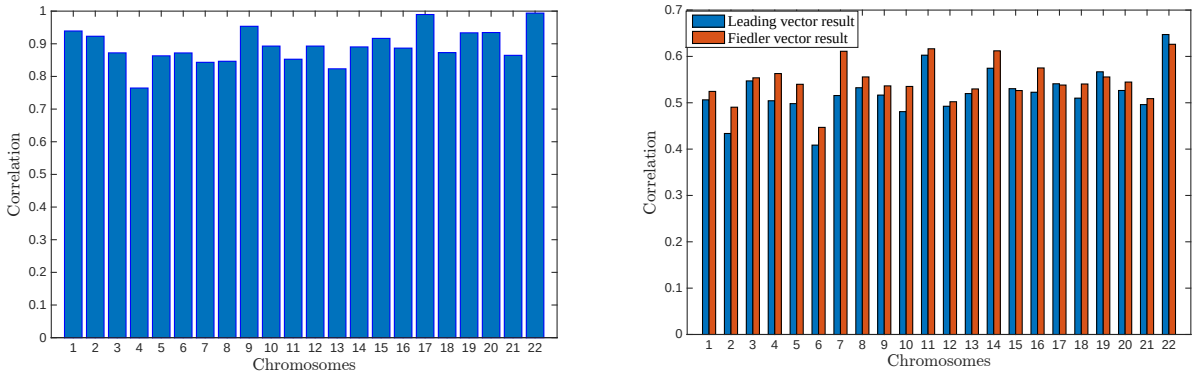


Fig. S1. This figure shows how our method of chromosome partitioning via the Fiedler vector compares to previously published methods from (Lieberman-Aiden et al. 2009). Green arrows represent our method, red arrows represent Lieberman-Aiden et al method. In both methods the raw Hi-C Matrix is Toeplitz normalized to highlight long distance Hi-C contacts. (Lieberman-Aiden et al. 2009) then creates a correlation matrix based on the Toeplitz normalized matrix. The first eigen vectors of this matrix acts to partition the chromosome into two compartments. Our method partitions the chromosome directly from the Toeplitz normalized matrix by calculating the Fiedler vectors of this matrix. The end results from each method are comparable).



(a) (b)

Fig. S2. (a) The correlations between the signed Fiedler vectors and the signed leading eigenvectors over chromosomes. (b) The correlation between Fiedler vector and RNAseq versus that computed from the leading eigenvector.

5 SUPPLEMENTAL FIGURE: ILLUSTRATION OF THE HIERARCHICAL DOMAIN IDENTIFICATION

In Fig. S3, we illustrate the of the hierarchical identification process via regions of chromosome 22.

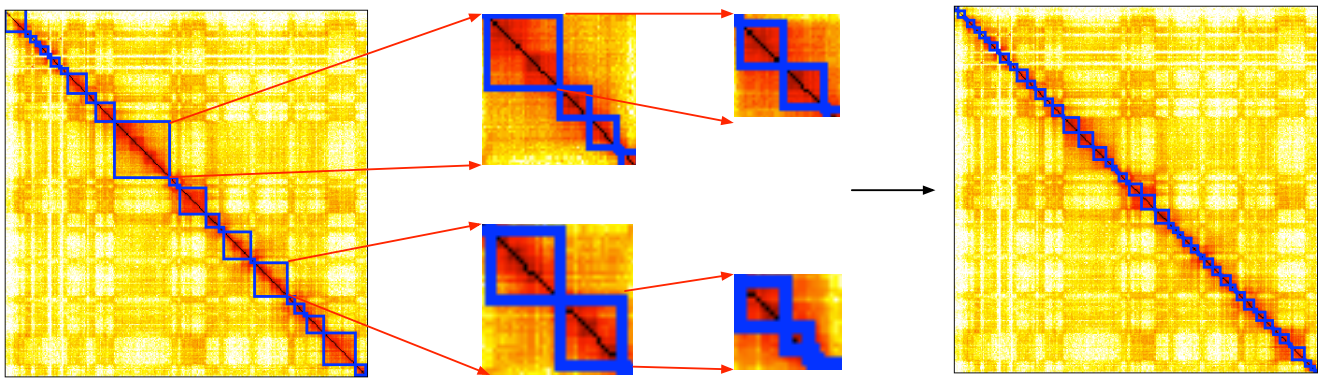


Fig. S3. Illustration of the hierarchical identification process with regions of chromosome 22.

6 SUPPLEMENTAL FIGURES: ILLUSTRATION OF THE IDENTIFIED DOMAINS

In Fig. S4–S25, We illustrate the identified TADs with the proposed algorithm using $\lambda_{\mathrm{thr}} = 0.8$ and $\lambda_{\mathrm{thr}} = 0.9$. The TAD size distribution and Fiedler number distribution are also provided.

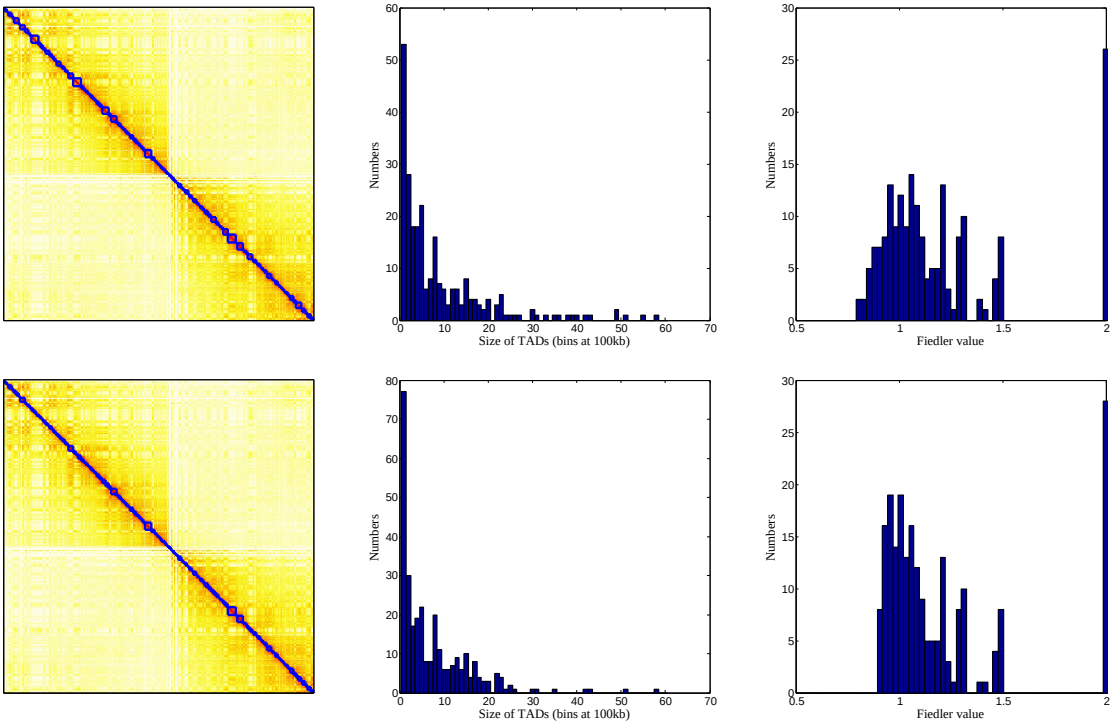


Fig. S4. Identified topological domains for chromosome 1. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

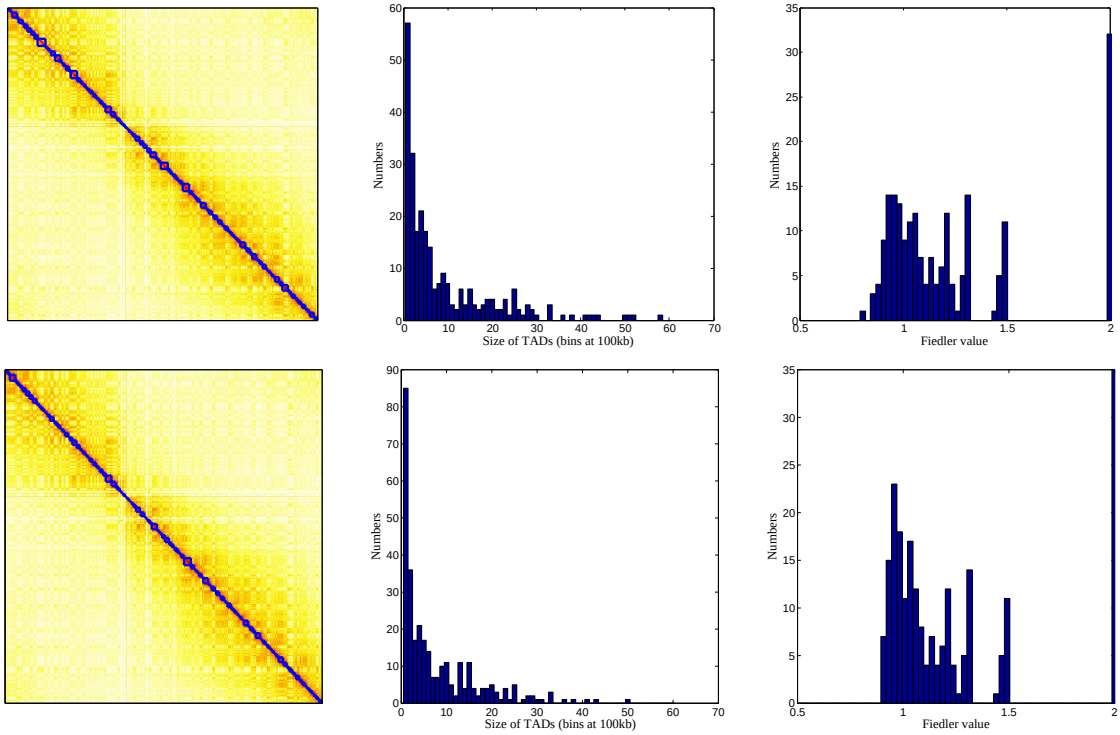


Fig. S5. Identified topological domains for chromosome 2. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

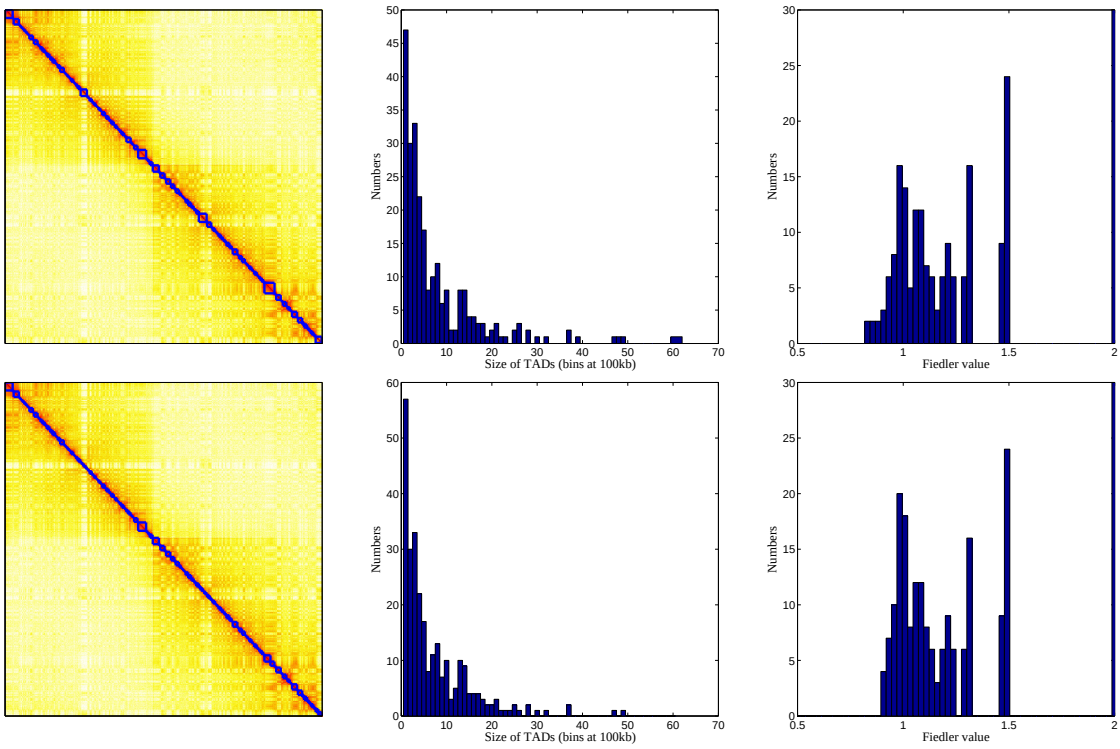


Fig. S6. Identified topological domains for chromosome 3. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

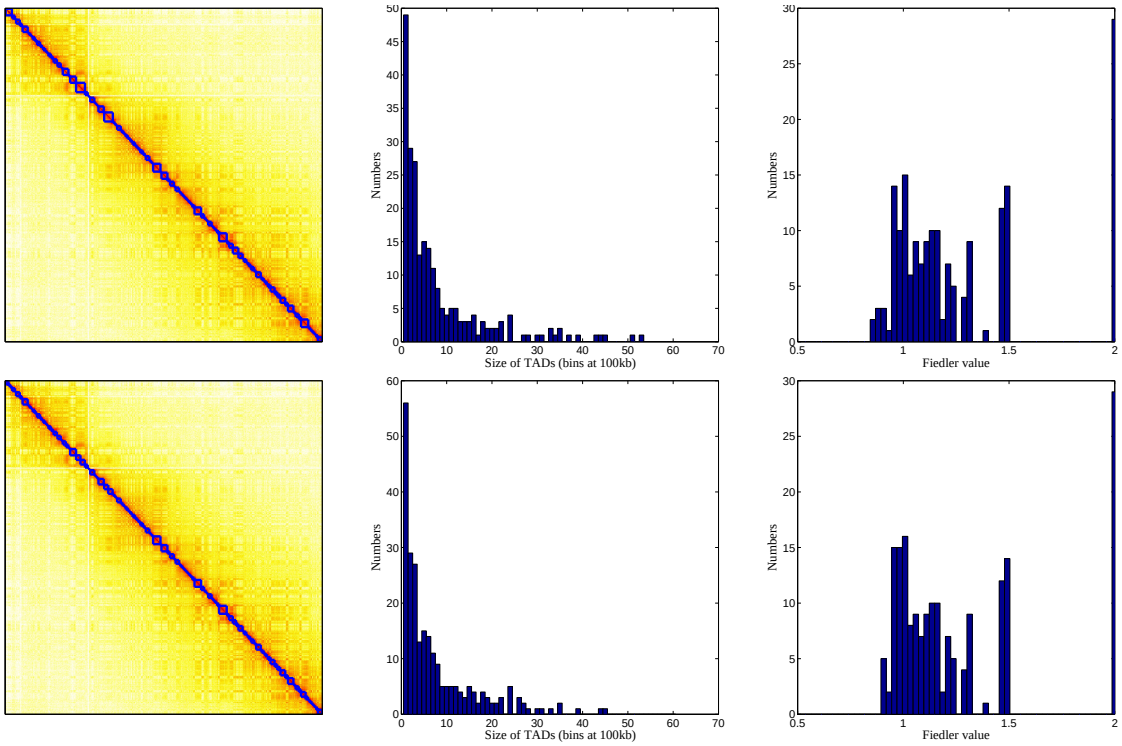


Fig. S7. Identified topological domains for chromosome 4. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

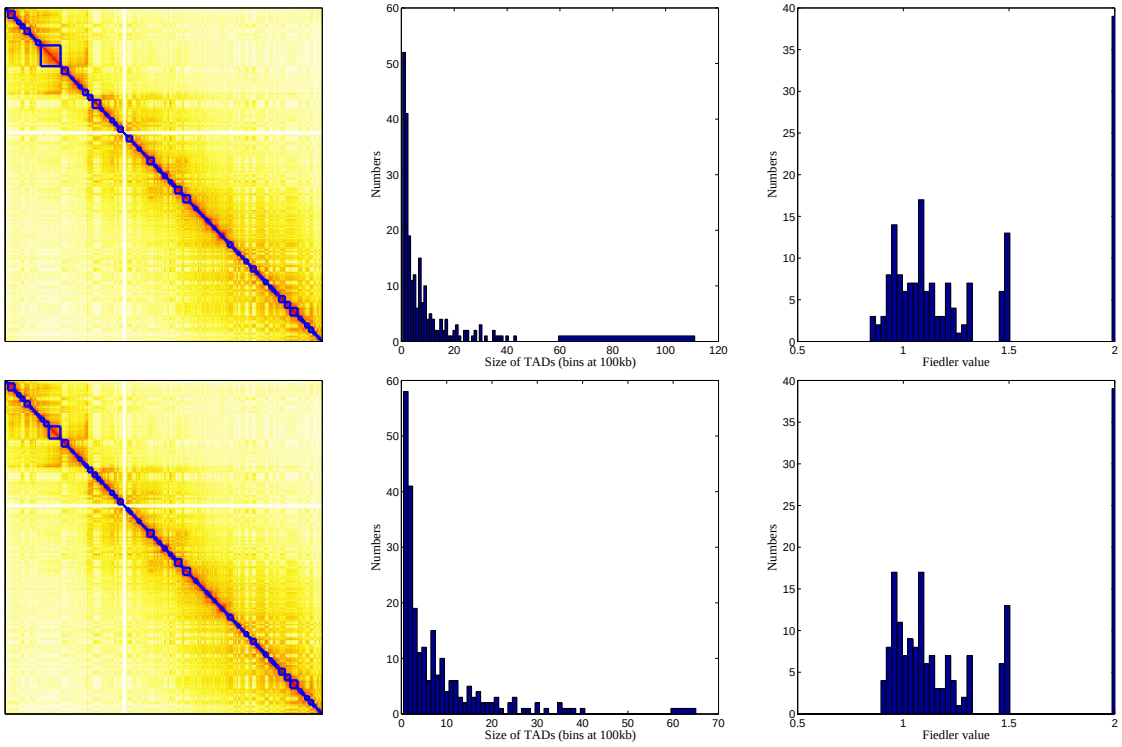


Fig. S8. Identified topological domains for chromosome 5. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

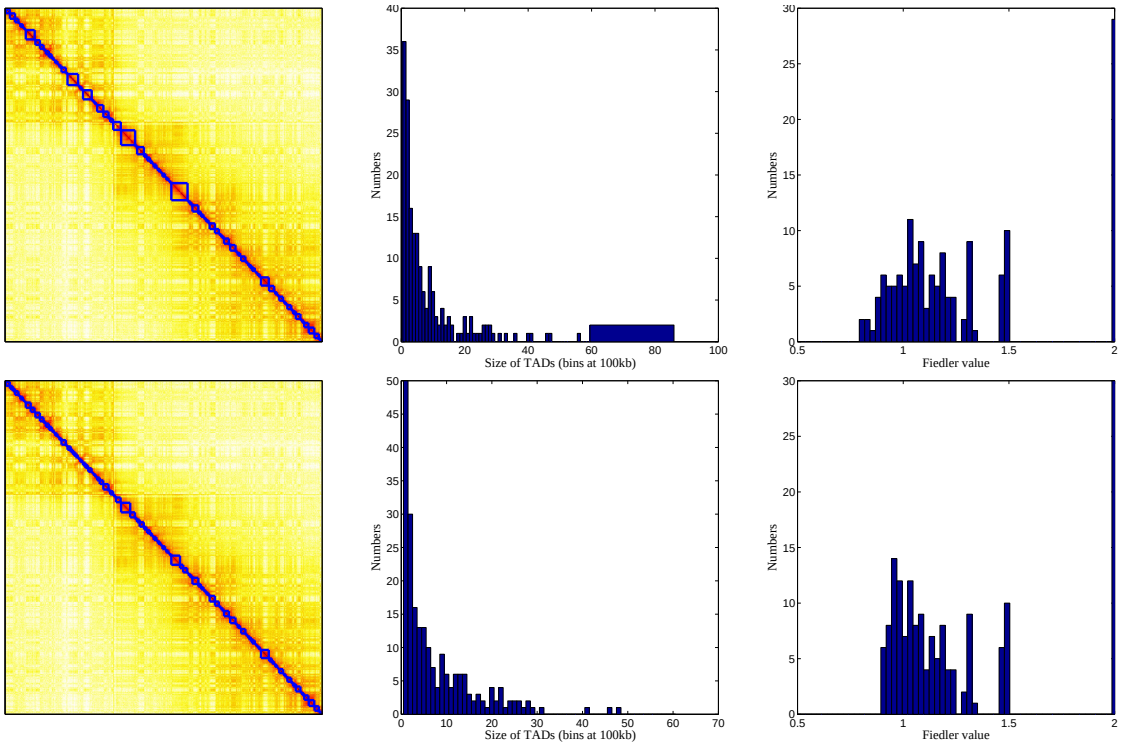


Fig. S9. Identified topological domains for chromosome 6. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

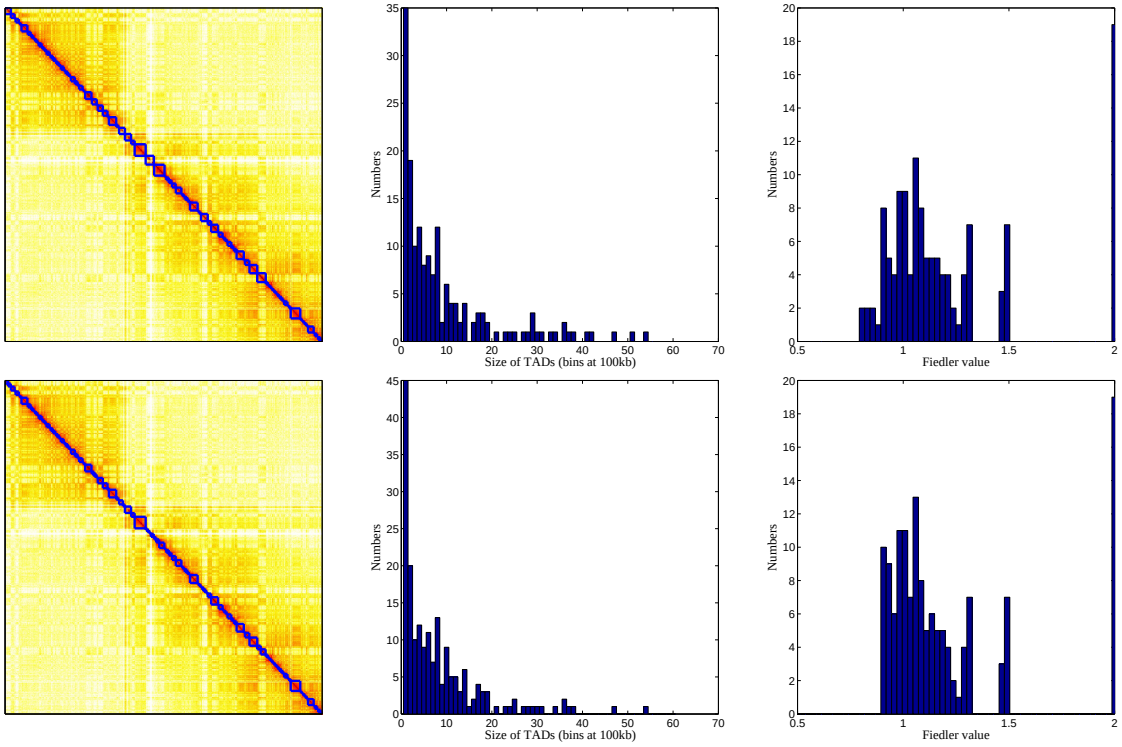


Fig. S10. Identified topological domains for chromosome 7. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

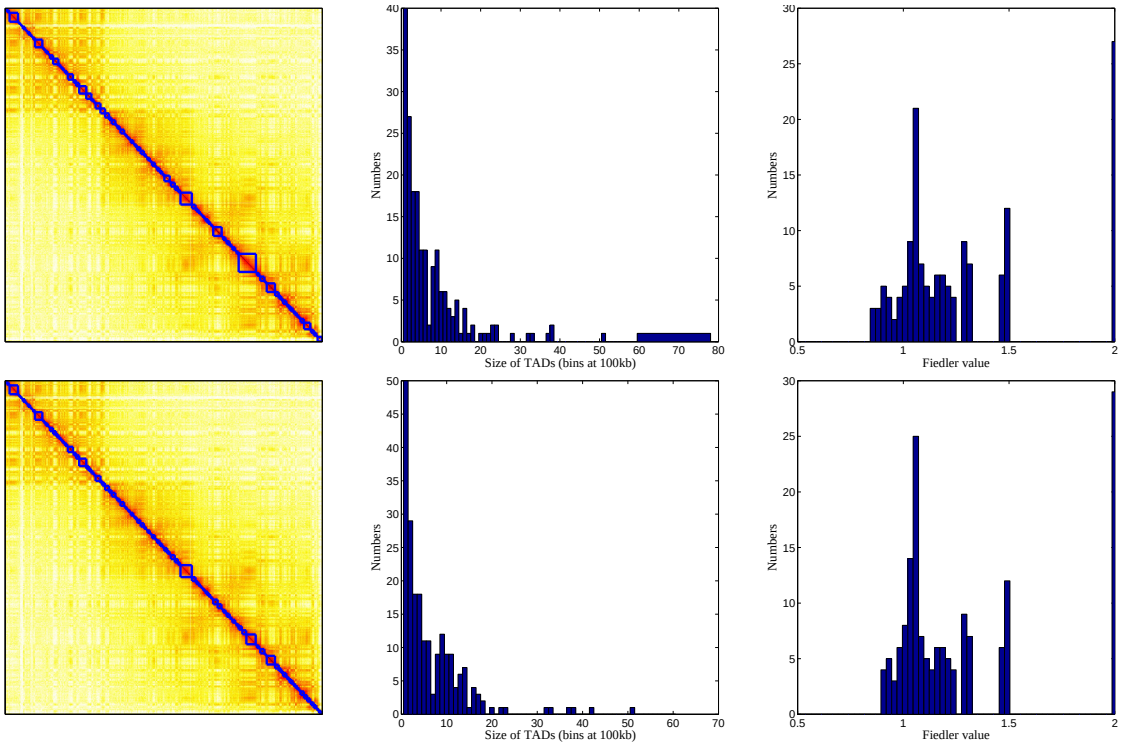


Fig. S11. Identified topological domains for chromosome 8. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

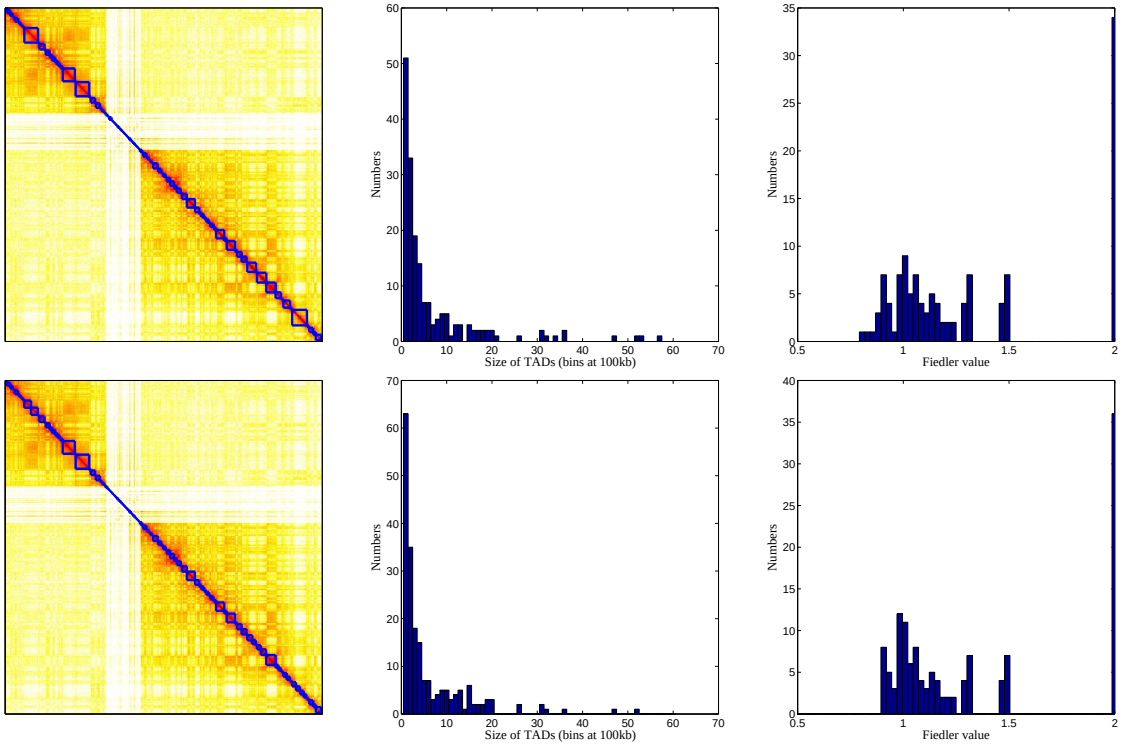


Fig. S12. Identified topological domains for chromosome 9. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

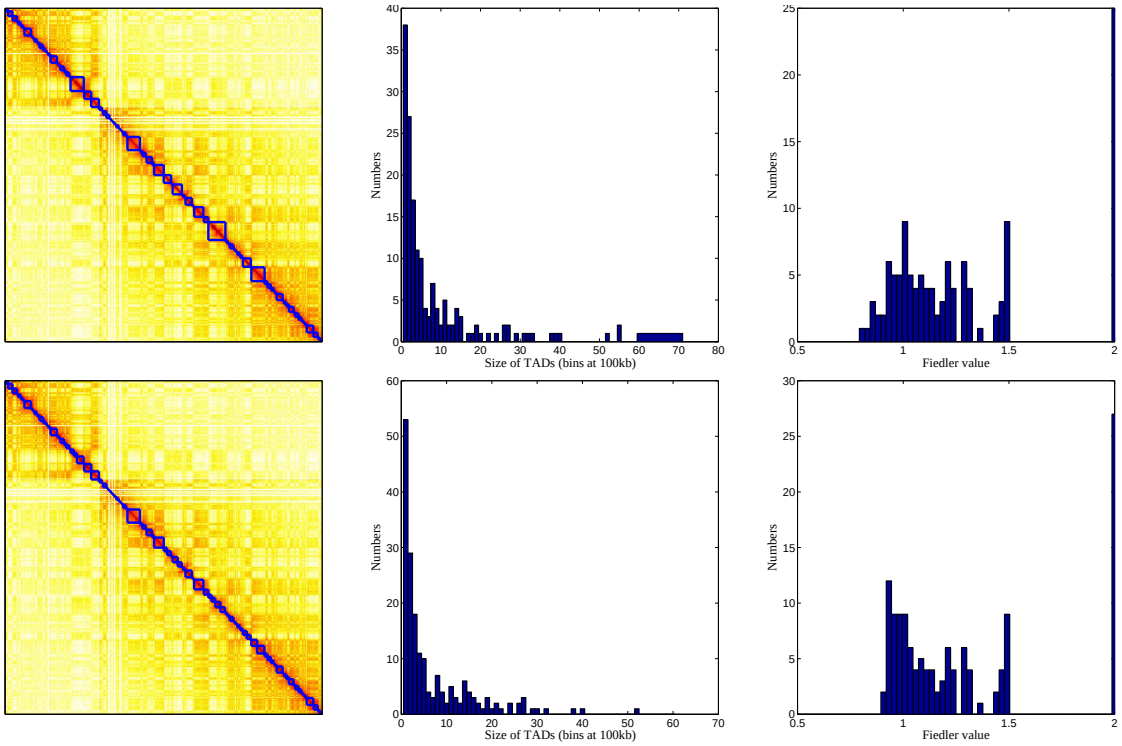


Fig. S13. Identified topological domains for chromosome 10. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

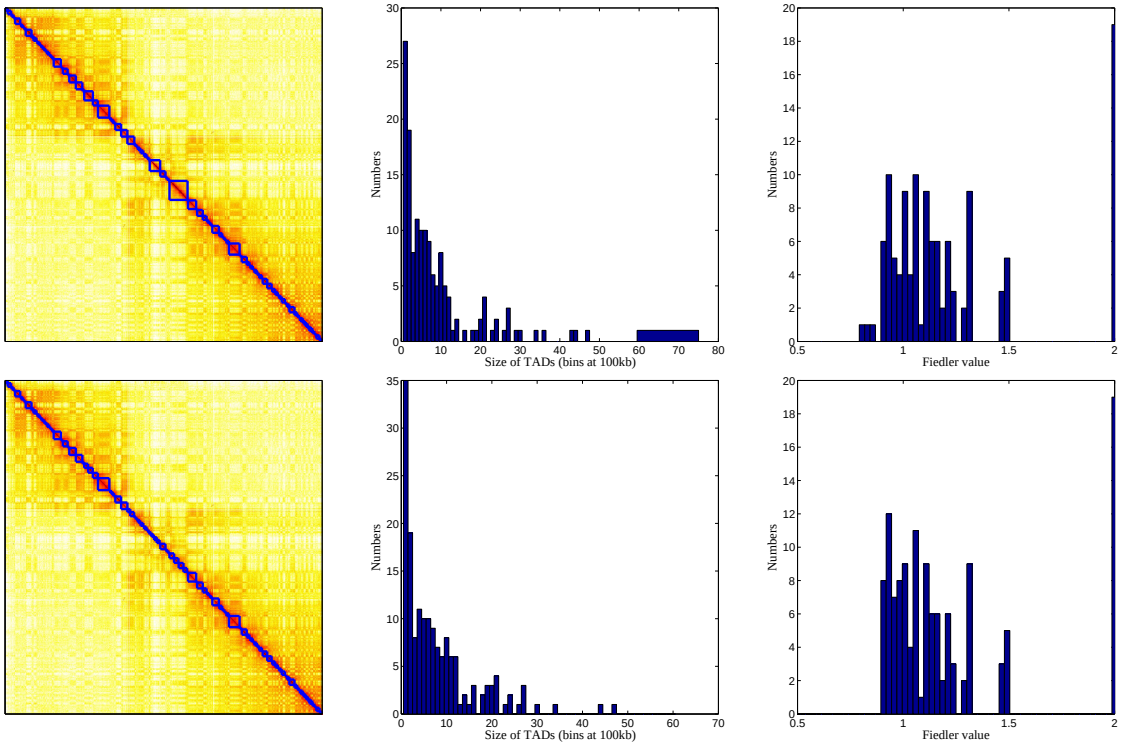


Fig. S14. Identified topological domains for chromosome 11. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

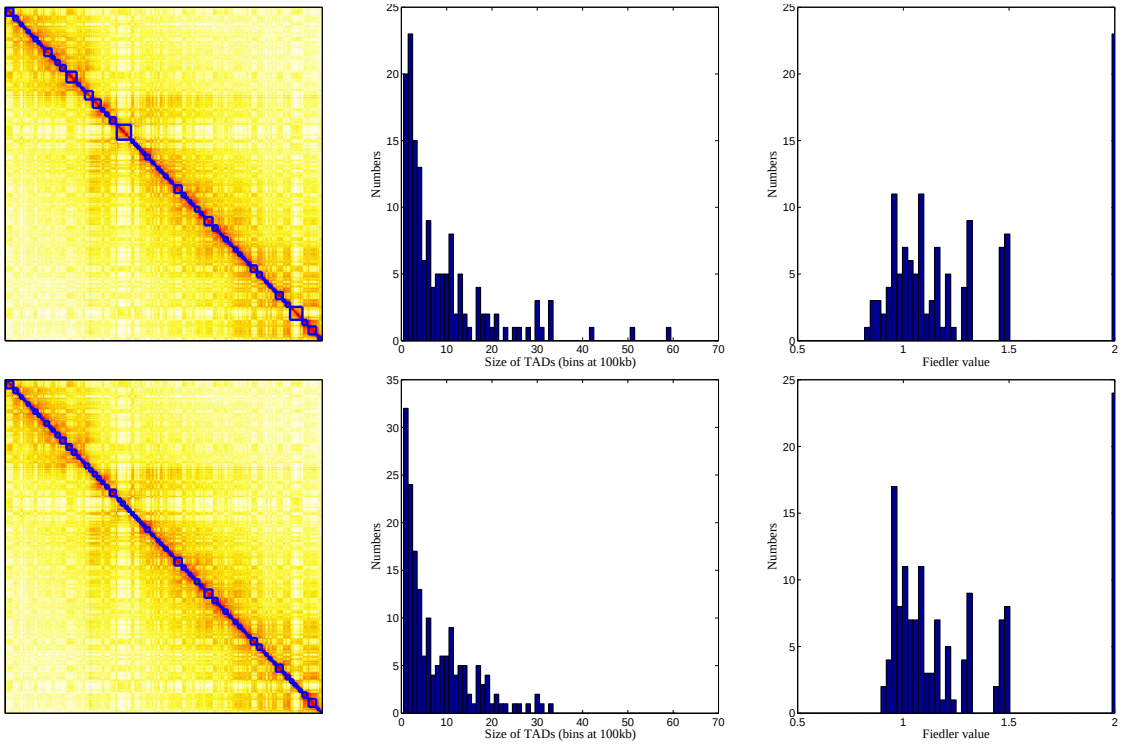


Fig. S15. Identified topological domains for chromosome 12. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

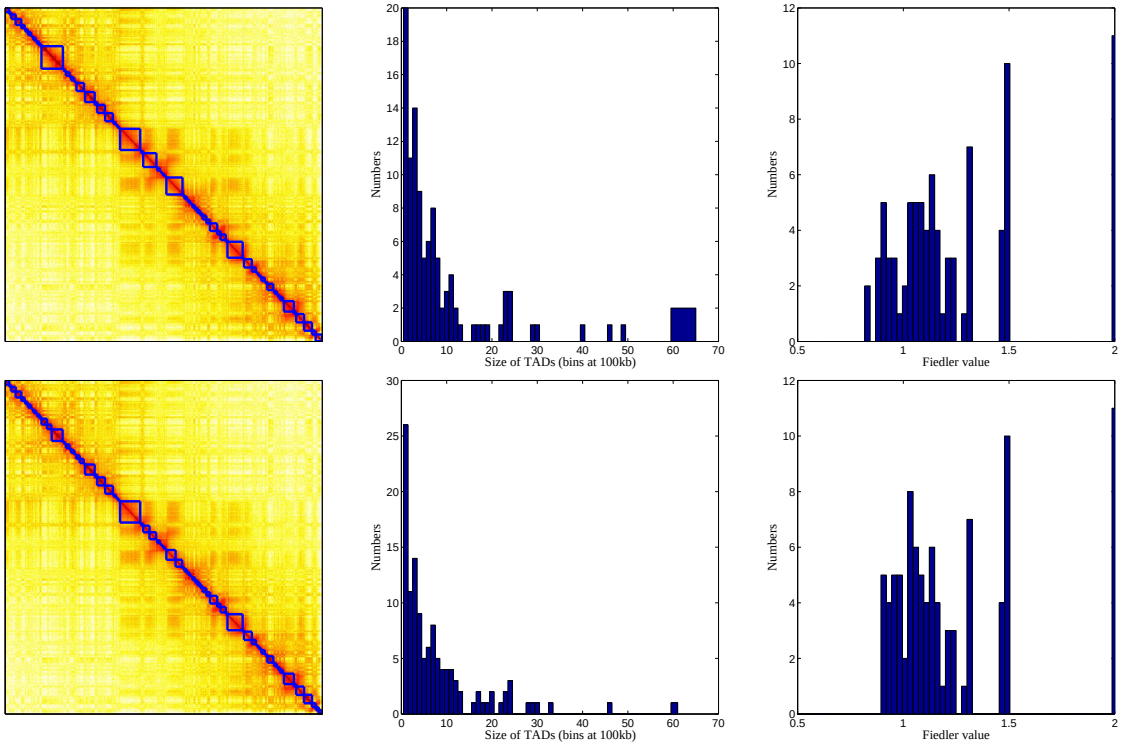


Fig. S16. Identified topological domains for chromosome 13. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

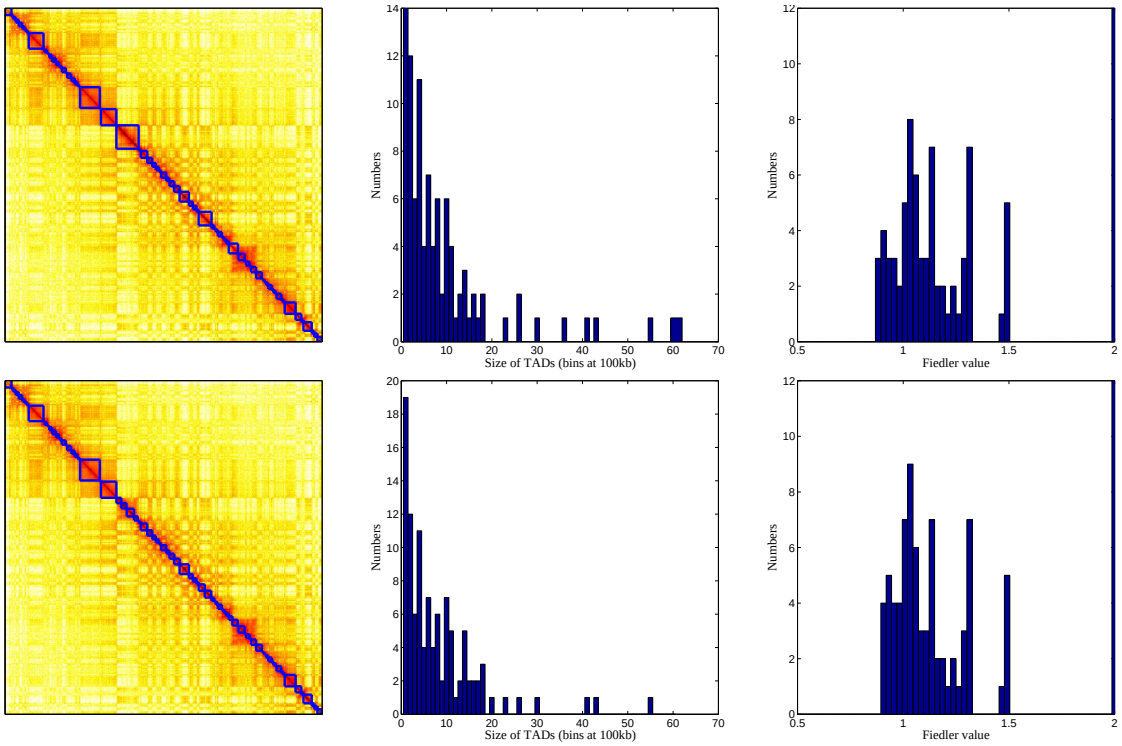


Fig. S17. Identified topological domains for chromosome 14. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

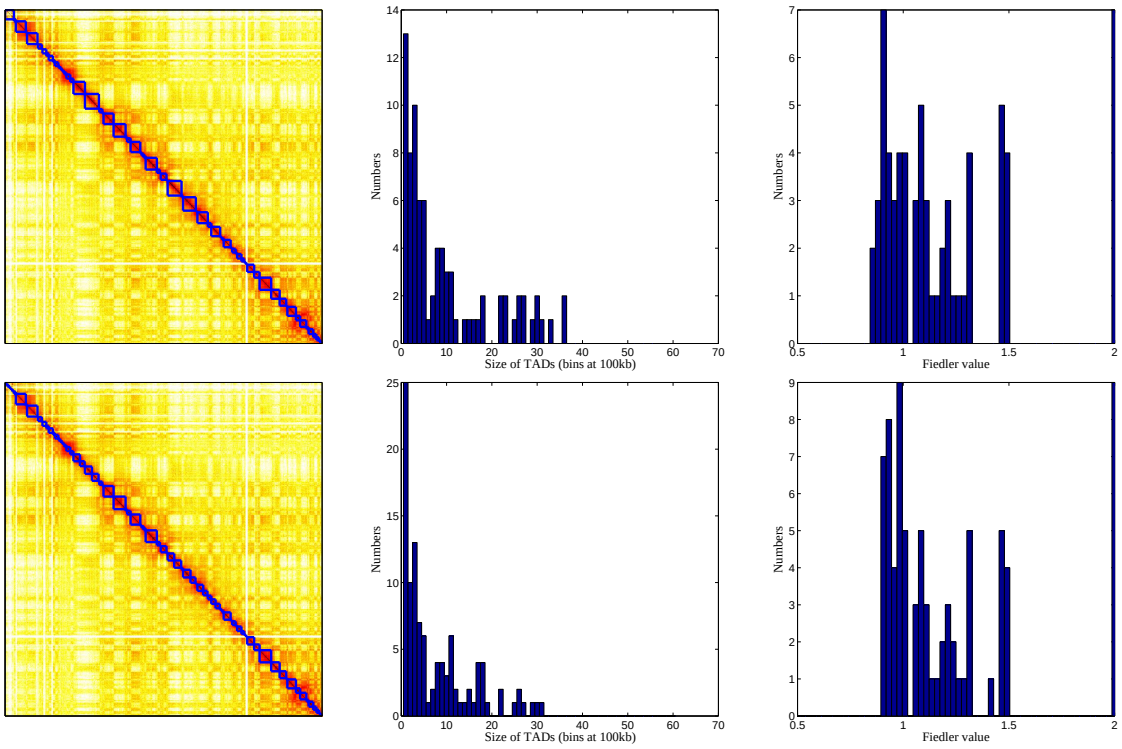


Fig. S18. Identified topological domains for chromosome 15. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

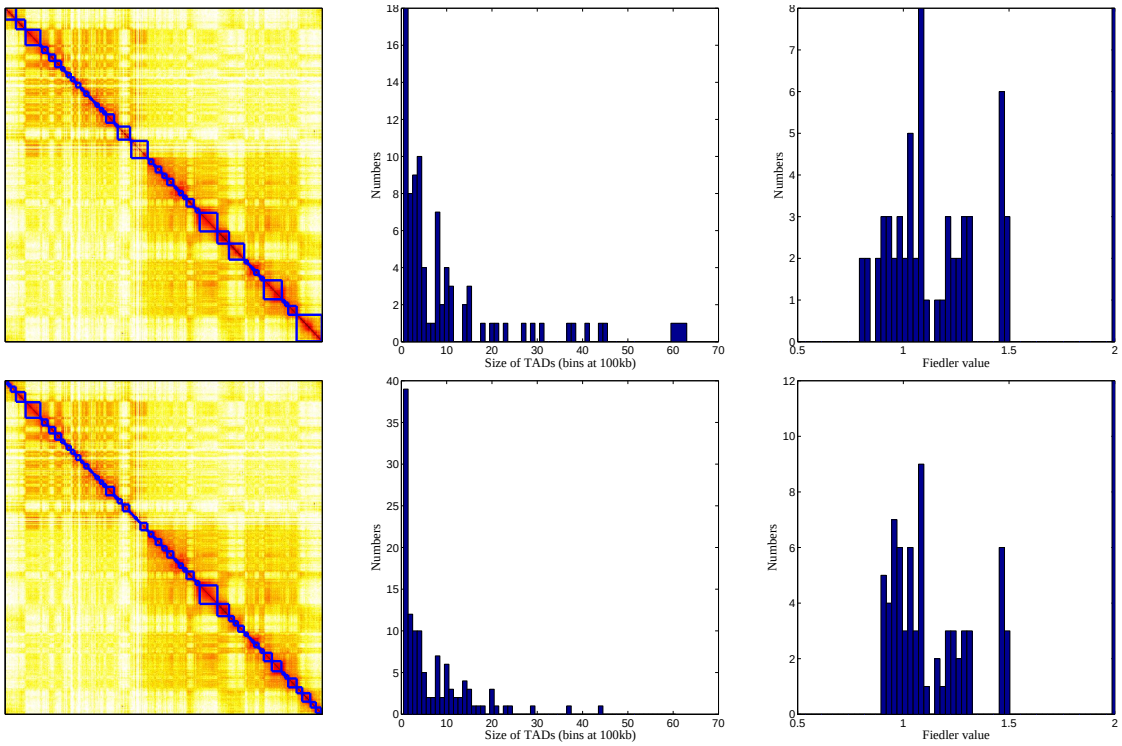


Fig. S19. Identified topological domains for chromosome 16. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

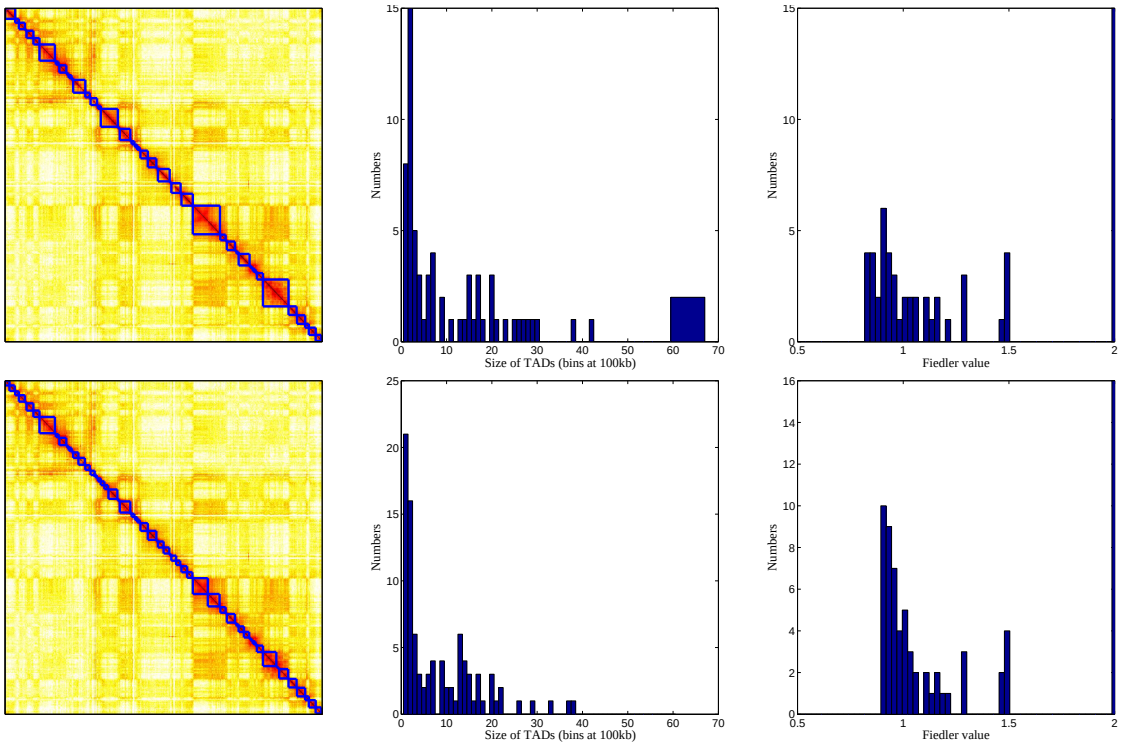


Fig. S20. Identified topological domains for chromosome 17. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

7 SUPPLEMENTAL FIGURES: ILLUSTRATION OF THE IDENTIFIED DOMAINS

In Fig. S4–S25, We illustrate the identified TADs with the proposed algorithm using $\lambda_{\mathrm{thr}} = 0.8$ and $\lambda_{\mathrm{thr}} = 0.9$. The TAD size distribution and Fiedler number distribution are also provided.

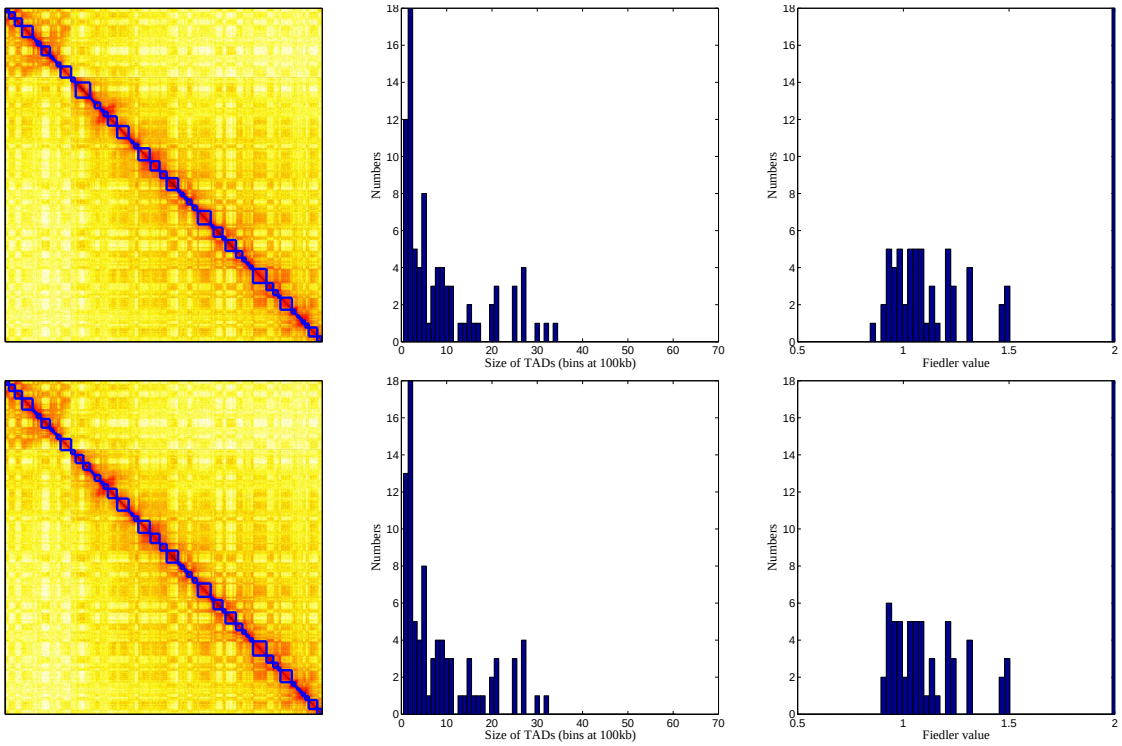


Fig. S21. Identified topological domains for chromosome 18. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

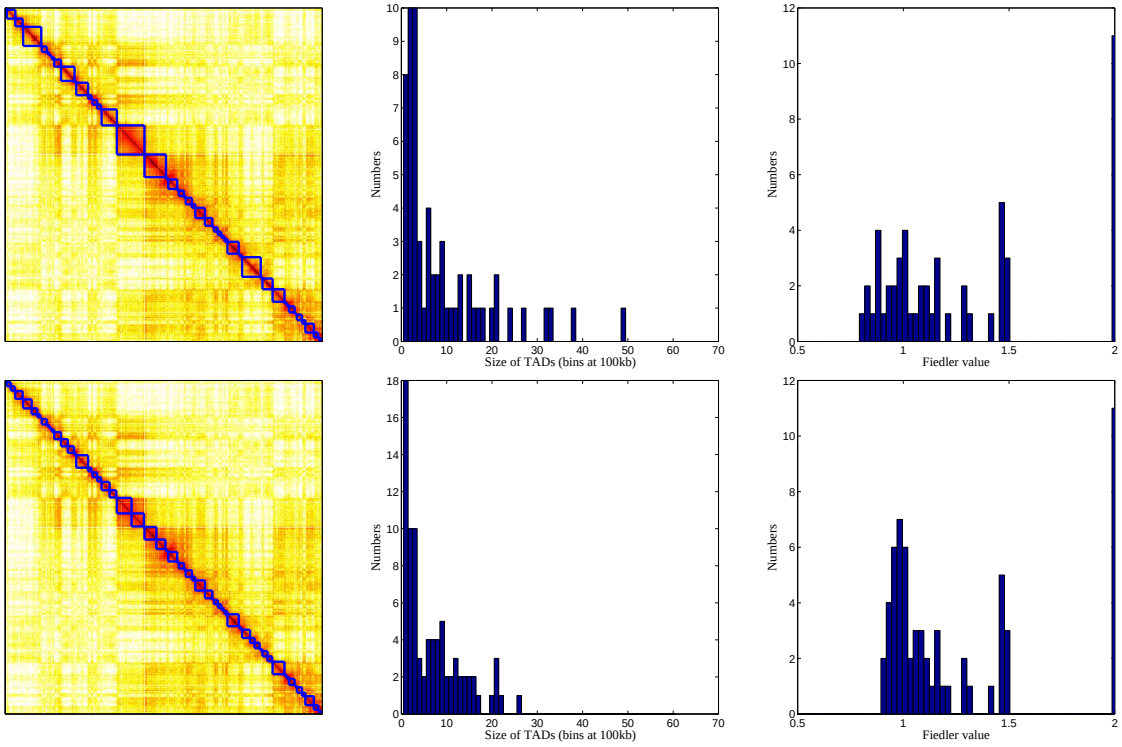


Fig. S22. Identified topological domains for chromosome 19. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

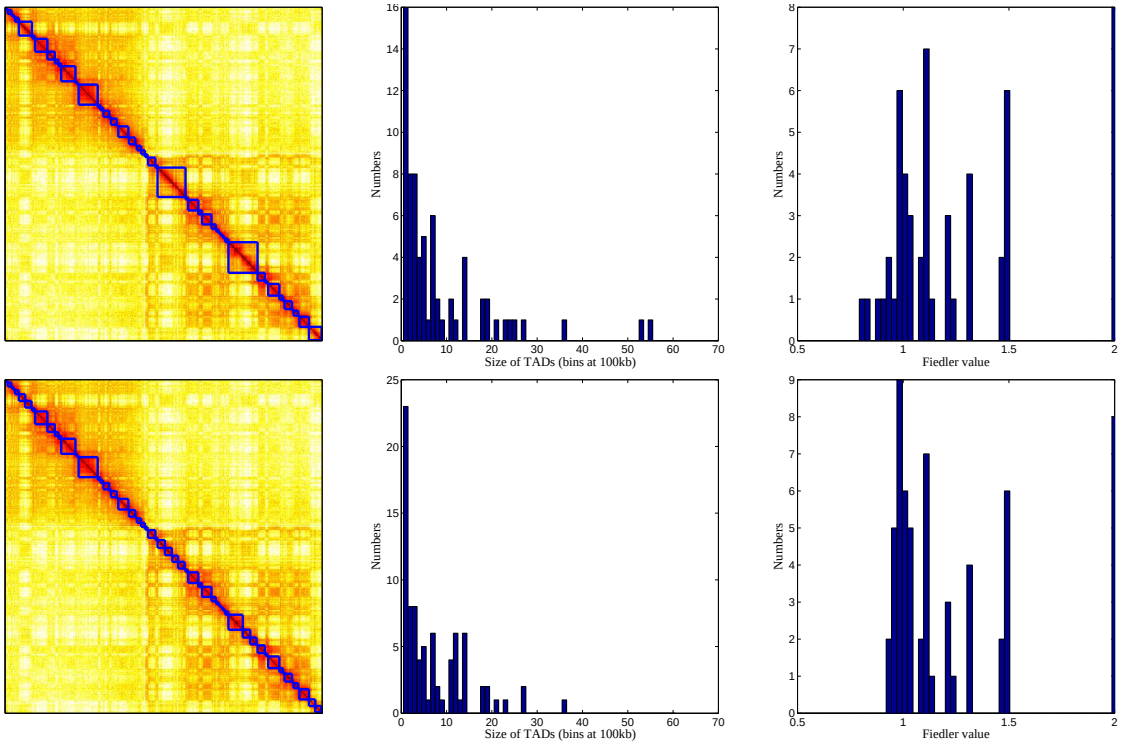


Fig. S23. Identified topological domains for chromosome 20. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

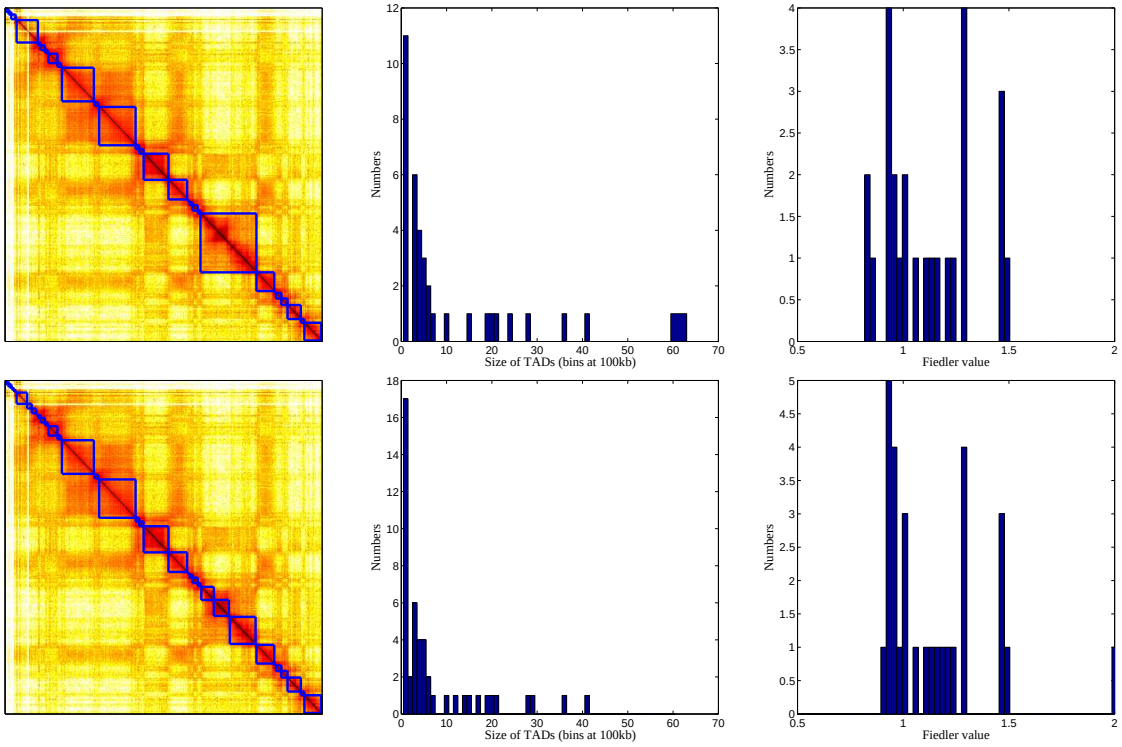


Fig. S24. Identified topological domains for chromosome 21. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

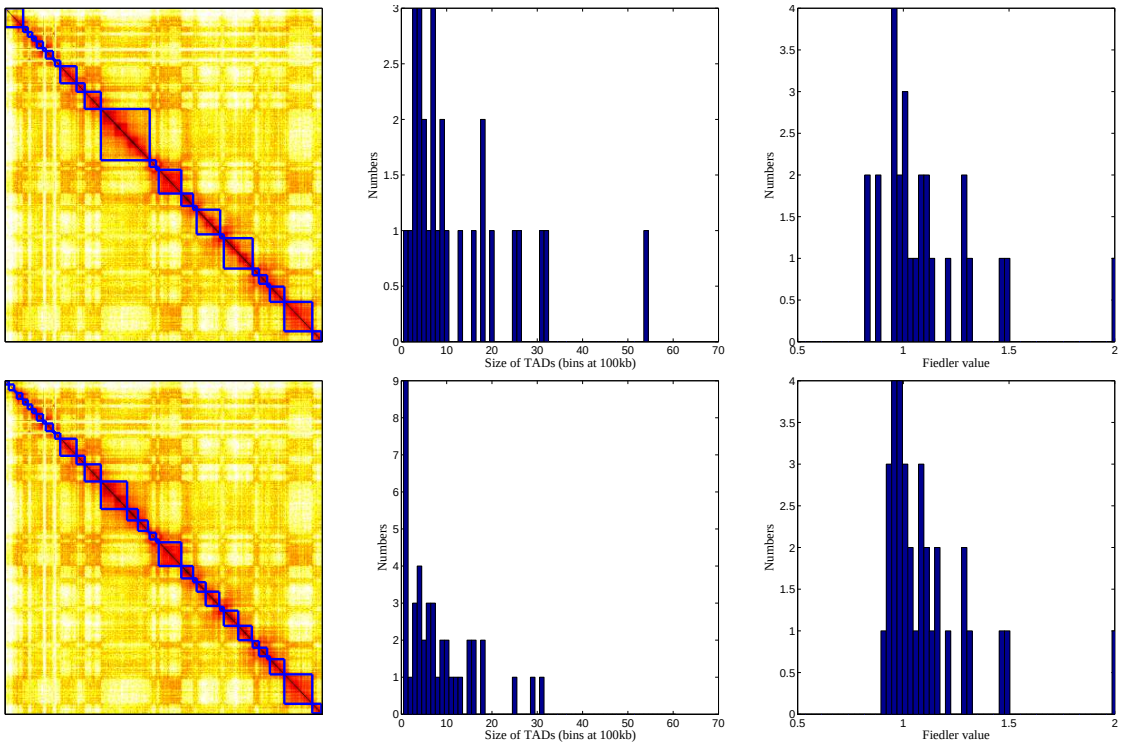


Fig. S25. Identified topological domains for chromosome 22. First row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.8$. Second row: results of Algorithm 1 with $\lambda_{\mathrm{thr}} = 0.9$. First column: identified domains. Second column: distribution of identified domain sizes. Third column: distribution of Fiedler values of the identified domains.

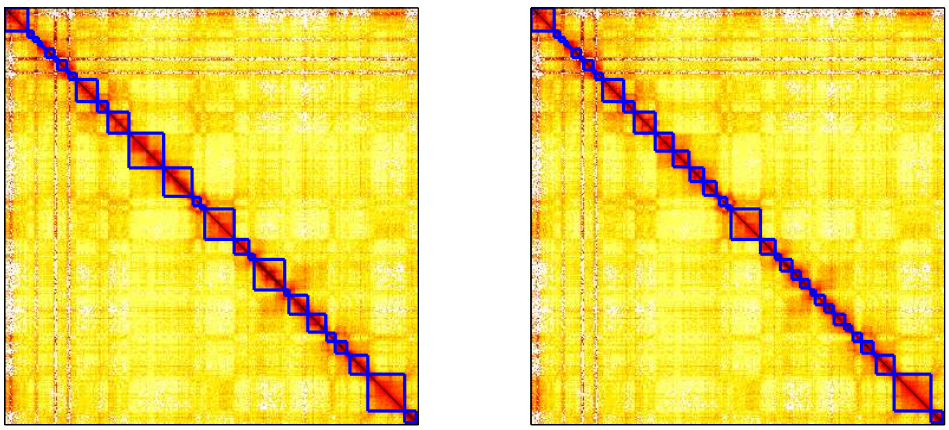


Fig. S26. Identified topological domains for chromosome 22 after depth normalization. Left: $\lambda_{\mathrm{thr}} = 0.8$. Right: $\lambda_{\mathrm{thr}} = 0.9$.

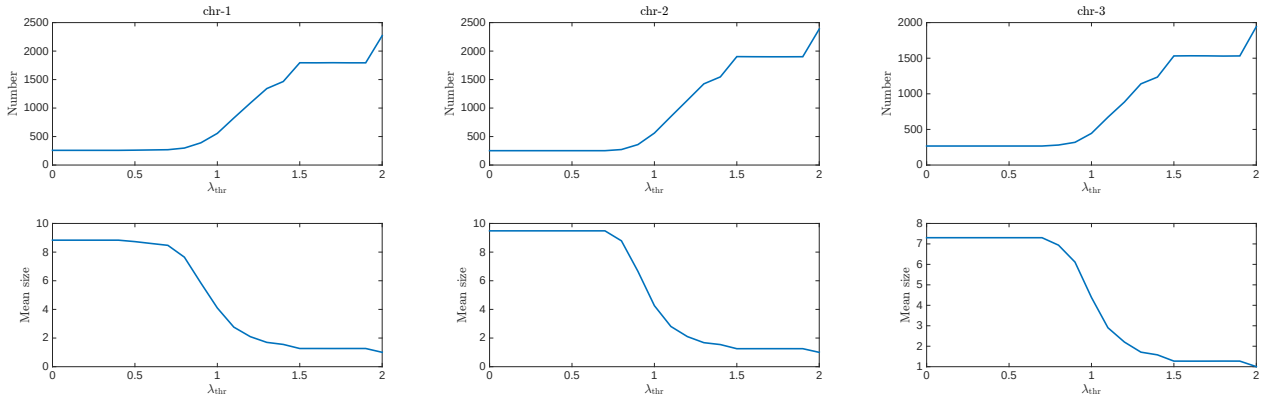


Fig. S27. The number of identified TADs (top) and mean TAD size (bottom) on Chromosome 1 to 3 (left to right) versus the Fiedler number threshold.

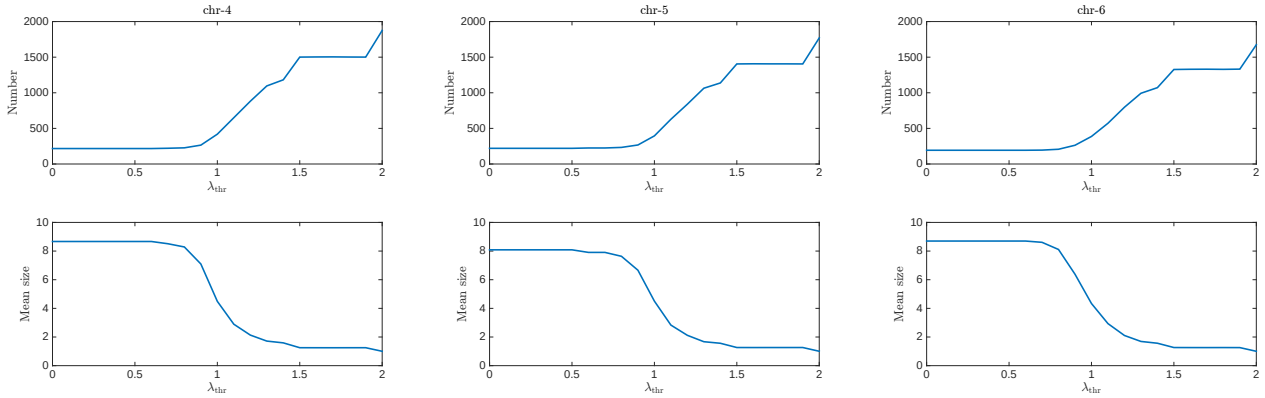


Fig. S28. The number of identified TADs (top) and mean TAD size (bottom) on Chromosome 4 to 6 (left to right) versus the Fiedler number threshold.

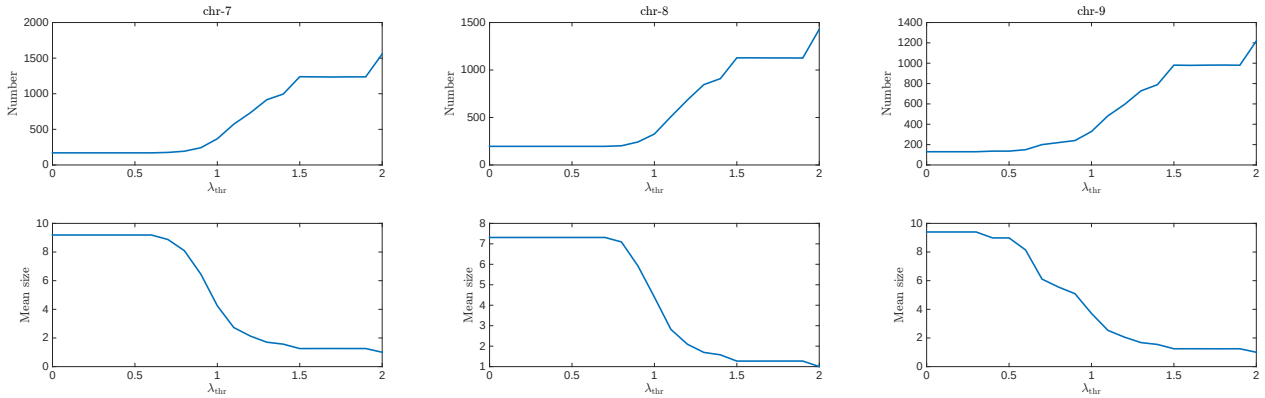


Fig. S29. The number of identified TADs (top) and mean TAD size (bottom) on Chromosome 7 to 9 (left to right) versus the Fiedler number threshold.

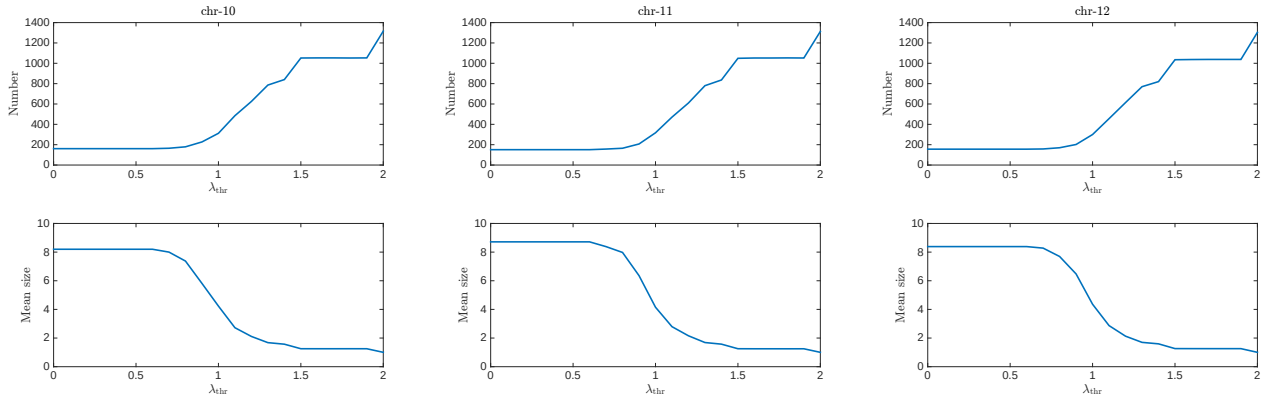


Fig. S30. The number of identified TADs (top) and mean TAD size (bottom) on Chromosome 10 to 12 (left to right) versus the Fiedler number threshold.

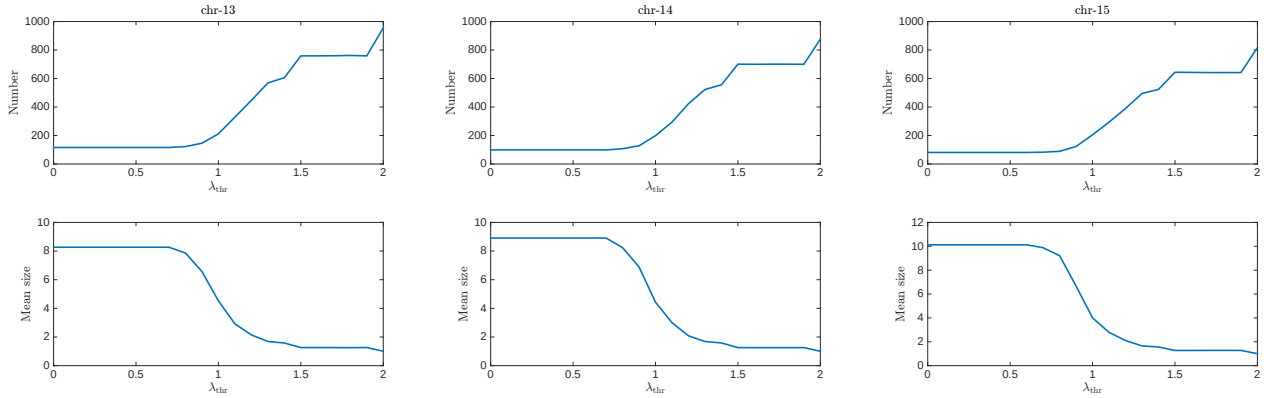


Fig. S31. The number of identified TADs (top) and mean TAD size (bottom) on Chromosome 13 to 15 (left to right) versus the Fiedler number threshold.

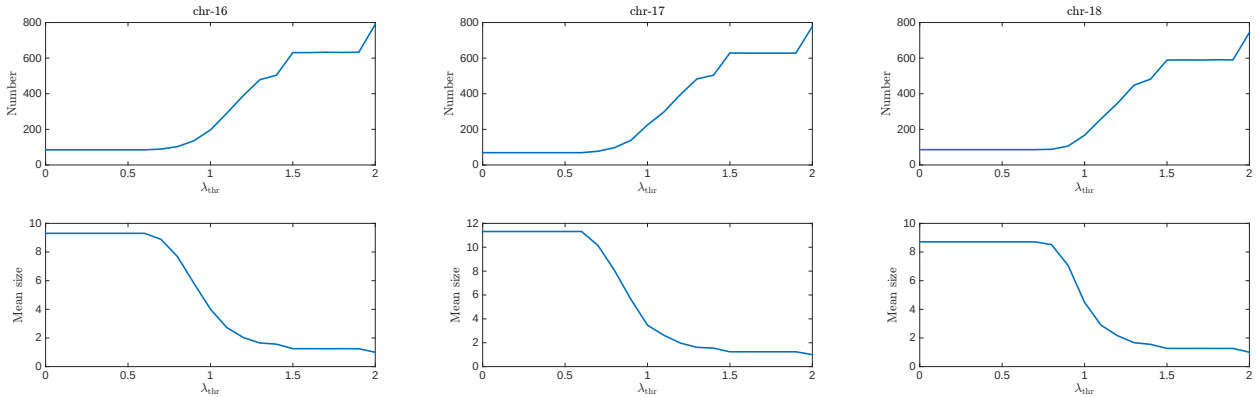


Fig. S32. The number of identified TADs (top) and mean TAD size (bottom) on Chromosome 16 to 18 (left to right) versus the Fiedler number threshold.

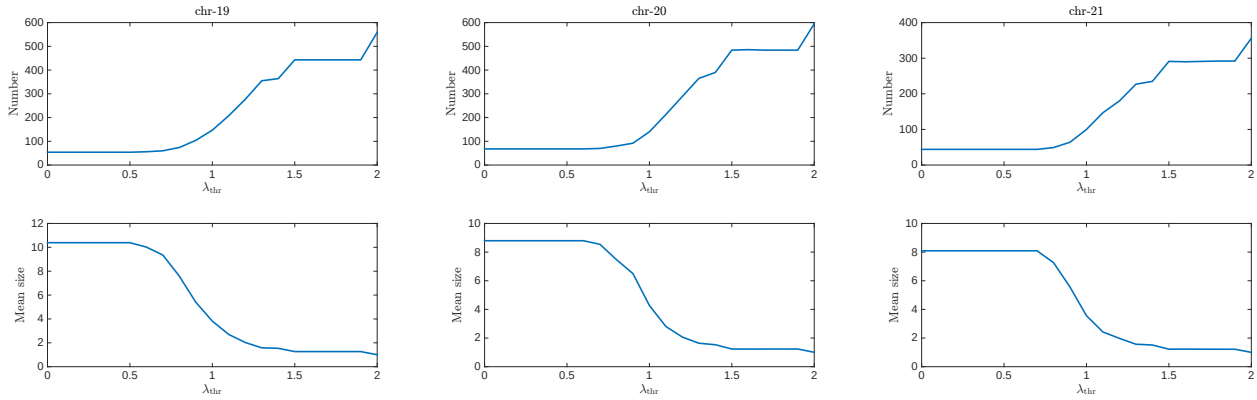


Fig. S33. The number of identified TADs (top) and mean TAD size (bottom) on Chromosome 19 to 21 (left to right) versus the Fiedler number threshold.

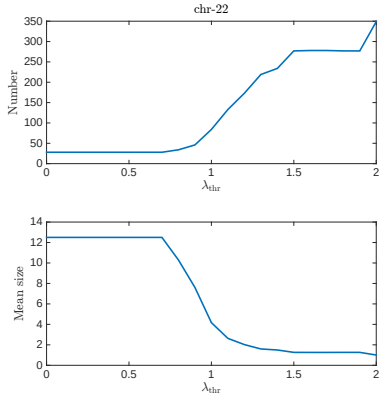


Fig. S34. The number of identified TADs (top) and mean TAD size (bottom) on Chromosome 22 versus the Fiedler number threshold.

Algorithm 1: Identification of TADs via graph Laplacian (2)

Parameters: Fiedler number threshold $\lambda_{\rm thr}$.
Pre-processing: For a given chromosome, set a mask matrix $\mathbf{M}$, and get the matrix $\overline{\overline{\mathbf{H}}}_{\rm M}$ by (1).
Algorithm:
Calculate the Fiedler vector of the matrix $\overline{\overline{\mathbf{H}}}_{\rm M}$:

$$\lambda_2, \boldsymbol{v} \leftarrow {\rm Fv}(\overline{\overline{\mathbf{H}}}_{\rm M}) \tag{2}$$

Repeat the above step for each domain defined by $\boldsymbol{v}$ until the obtained sub-domain has a Fiedler number larger than the threshold, or its size reaches the lower bound.

8 VARIANCE OF RNASEQ COUNTS WITHIN TADS

Besides Fig. 4 which illustrates the consistency between expression level and first layer domains, we illustrate the mean variance of RNAseq counts (log scale) within domains (for each Fiedler value threshold, we calculate the variance of log-scaled RNA-Seq reads in each TAD, and average them to get an average value) versus the Fiedler number threshold (Fig. 35). It is shown that finer TADs leads to more consistent transcription level (smaller variance) within the domains. This observation is consistent with the expected structure and function relation of the TADs.

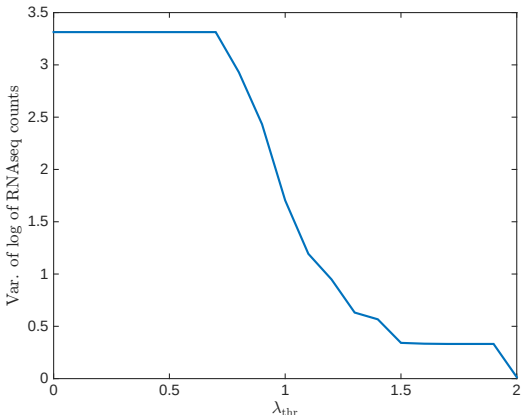


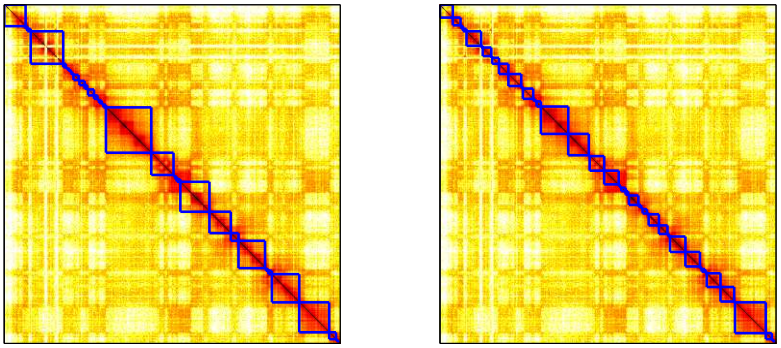
Fig. S35. Mean variance of RNAseq counts (log scale) within domains versus the Fiedler number threshold.

9 ALTERNATIVE METHOD WITHOUT USING $\overline{\overline{\mathbf{H}}}$

Step 1 in the Algorithm uses matrix $\mathbf{H}_{\rm N}$ and differs from the followed subsequent steps that use $\overline{\overline{\mathbf{H}}}$. Starting with $\overline{\overline{\mathbf{H}}}$ can be beneficial by considering the long-range contact patterns related with the nuclei compartment structures, although our objective is to identify local structures. Further, using $\mathbf{H}_{\rm N}$ in the first step directly segments the chromosome to a number of first layer TADs. This reduces the number of recursions of the algorithm. Alternatively, instead of using the Toeplitz normalized Hi-C matrix in the first step, we can only recursively apply step 2 of the Algorithm 1 on a masked matrix $\overline{\overline{\mathbf{H}}}$ namely,

$$\overline{\overline{\mathbf{H}}}_{\rm M} = \overline{\overline{\mathbf{H}}} \otimes \mathbf{M} \tag{1}$$

where $\mathbf{M}$ is a mask matrix to emphasize the diagonal part of the matrix. It can be a binary band matrix such that $[\mathbf{M}]_{ij} = 1$ for $|i-j| < \ell_B$, otherwise $[\mathbf{M}]_{ij}$ =0, or a fading matrix such that $[\mathbf{M}]_{ij} = \frac{1}{|i-j|^\gamma}$. Either the parameter ℓ_B and γ define the desired range that the domains are considered. The full algorithm is summarized in Algorithm 1 in this file. Illustrative results with Chromosome 22 are shown in Fig. S36.



(a) $\lambda_{\mathrm{thr}} = 0.8$ (b) $\lambda_{\mathrm{thr}} = 0.9$

Fig. S36. Illustration of estimated topological domains on chromosome 22 by Algorithm 1 in this supplemental material.