Supplemental Information

Chromosomal dynamics predicted by an elastic network model explains genome-wide accessibility and long-range couplings

Natalie Sauerwald,[1,*] She Zhang,[2,*] Carl Kingsford[1] and Ivet Bahar[2,+]

[1] *Computational Biology Department, School of Computer Science, Carnegie Mellon University, and* [2]*Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, 3064 Pittsburgh, PA 15213*

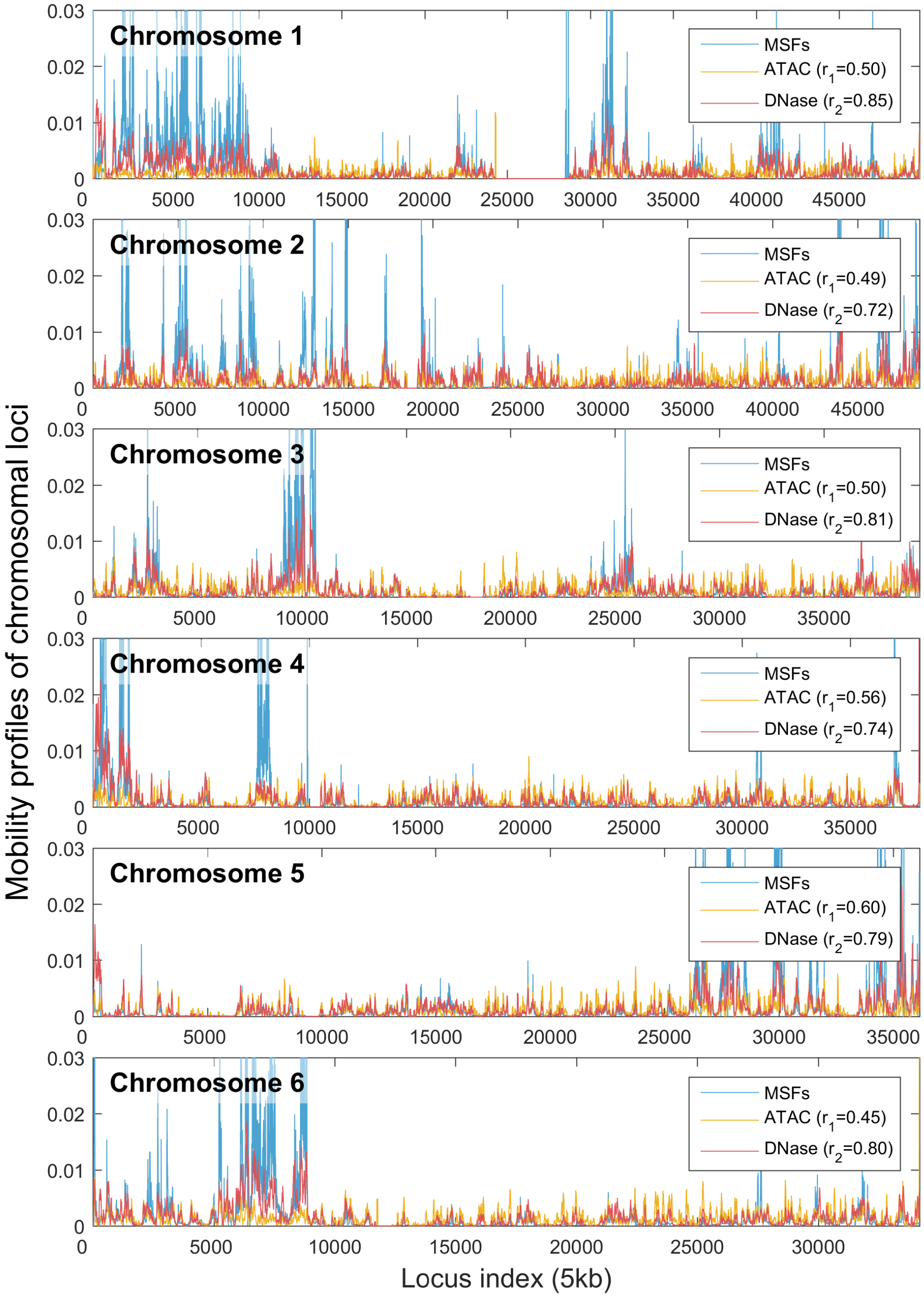

Figure S1

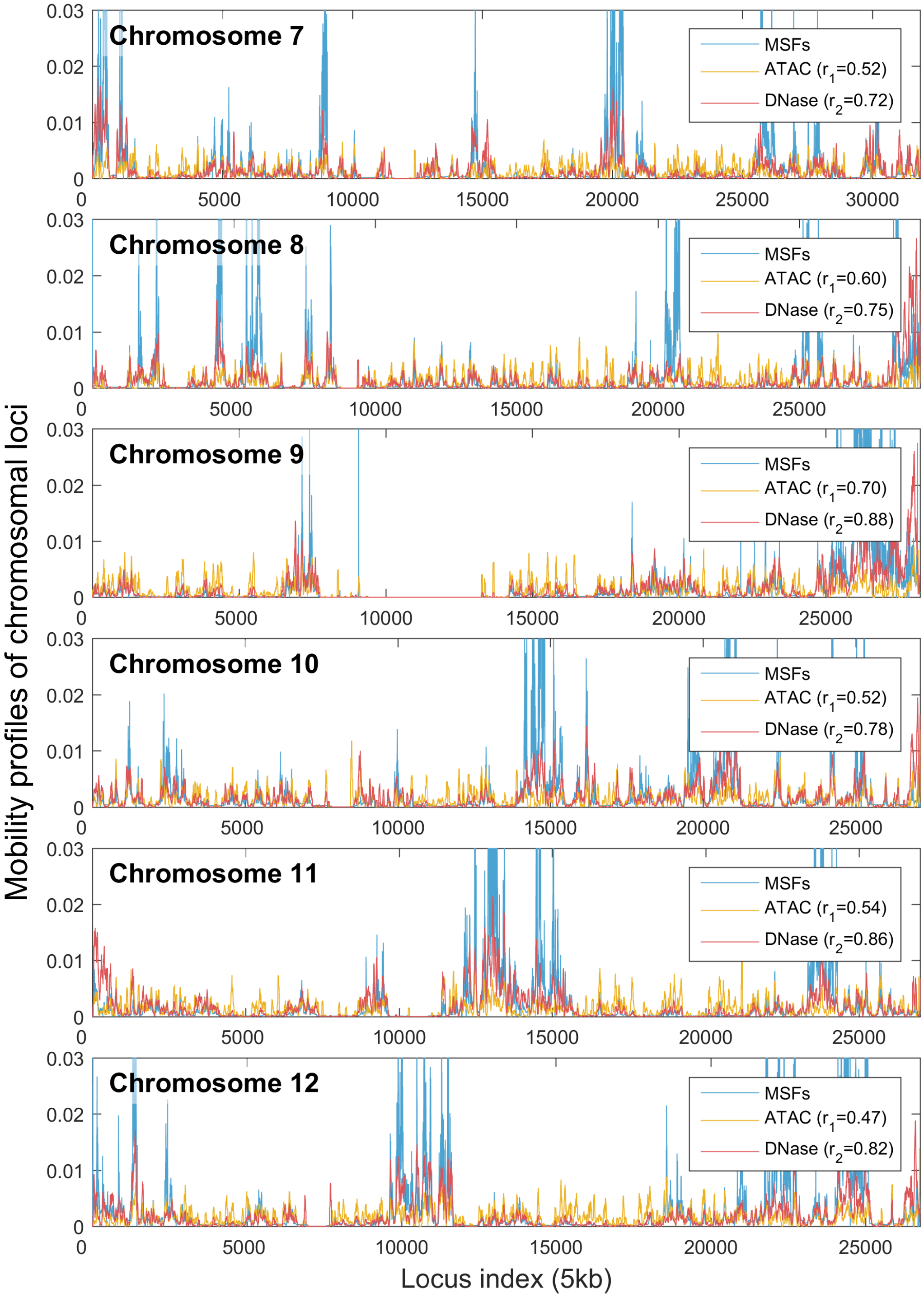

Figure S1 (*continued*)

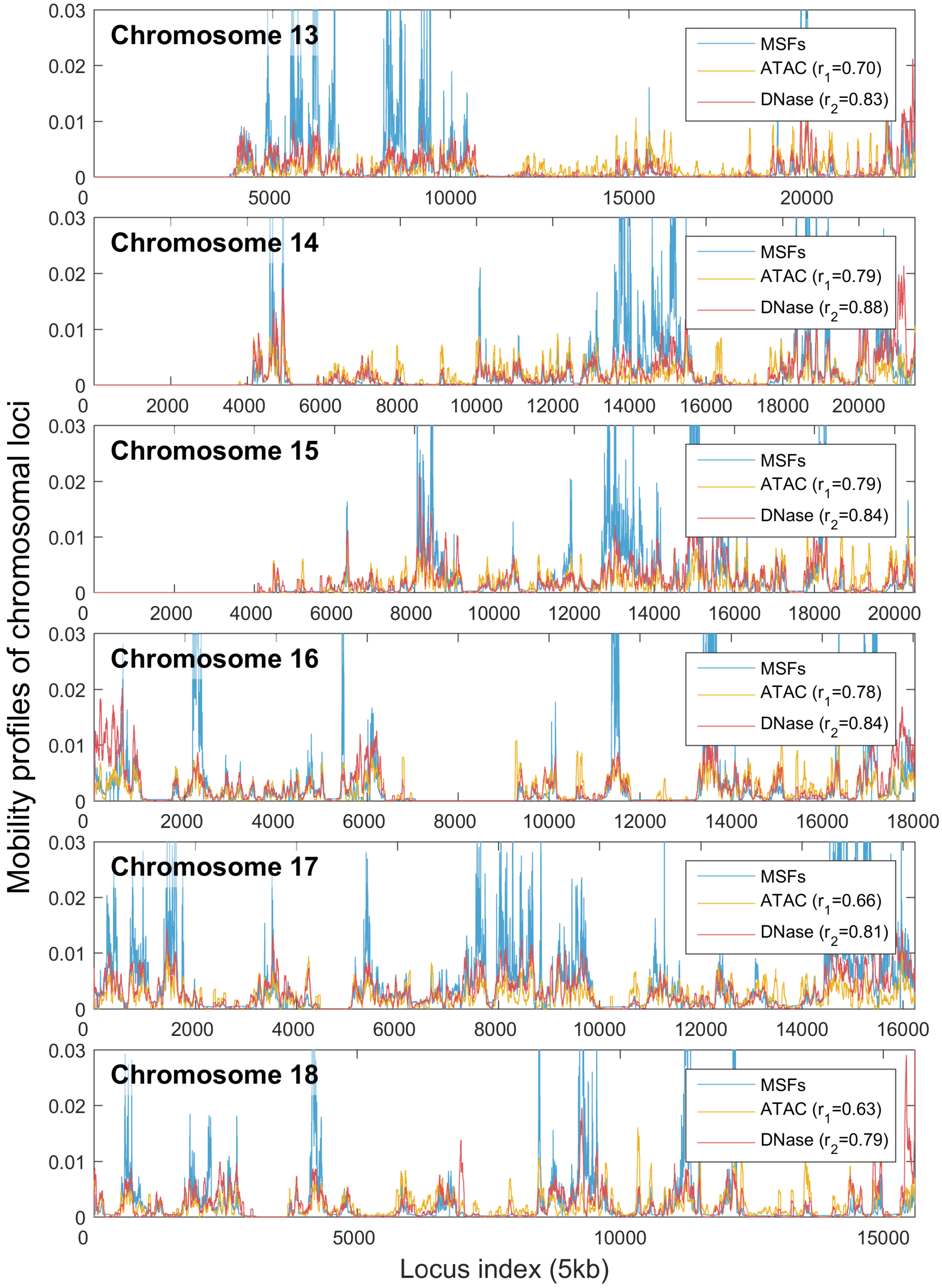

Figure S1 (*continued*)

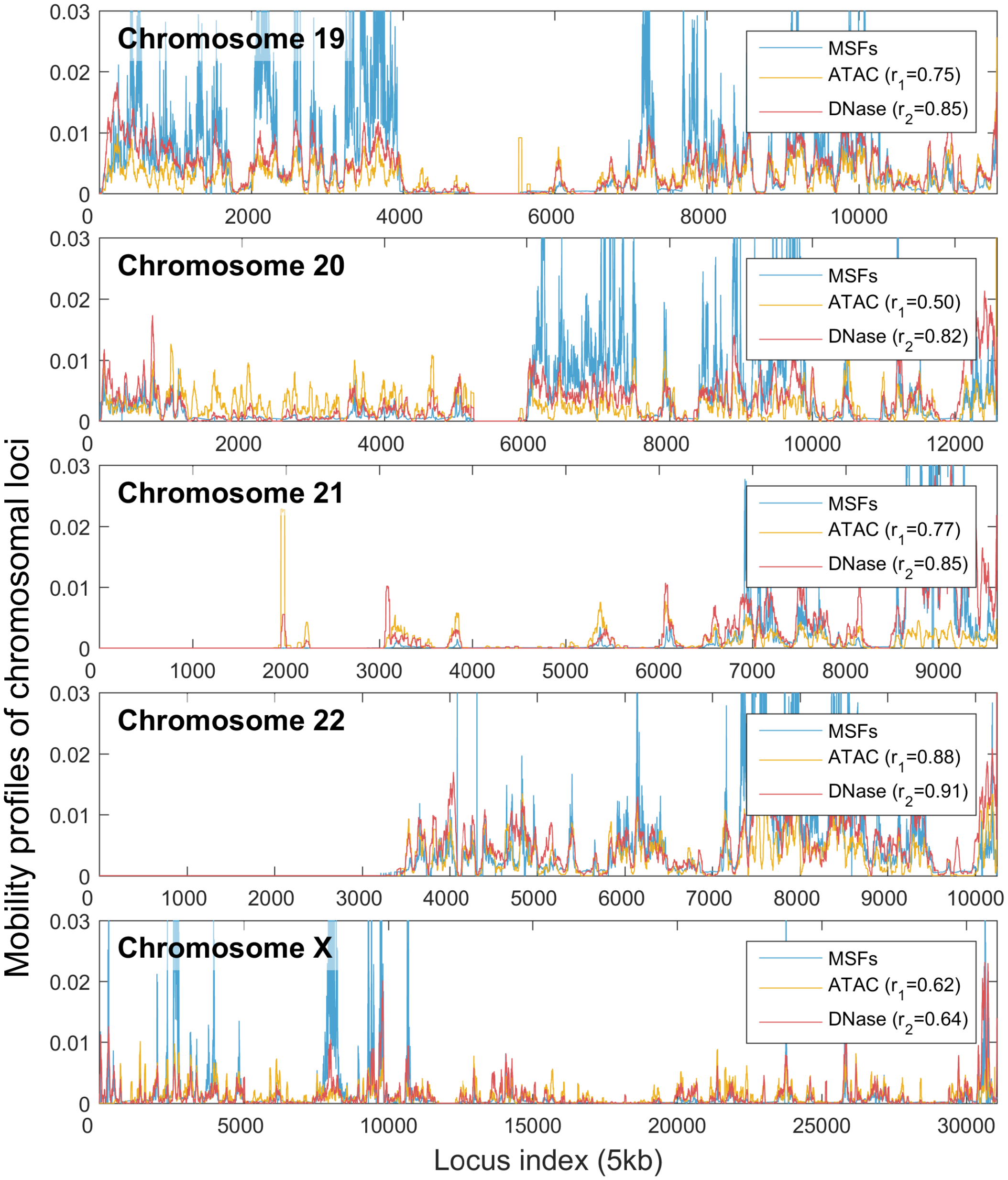

Figure S1. Comparison of GNM-predicted MSFs of gene loci with accessibility measured by ATAC and DNase-seq for all chromosomes. MSFs are calculated by using 500 GNM modes (at the lowest frequency end of the spectrum) based on Hi-C map at 5kb resolution obtained by Rao et al. for GM12878 (1). Spearman correlations between theoretical (MSFs) and experimental (ATAC (2) and DNase-seq (3)) data are shown for each chromosome in the corresponding legend box.

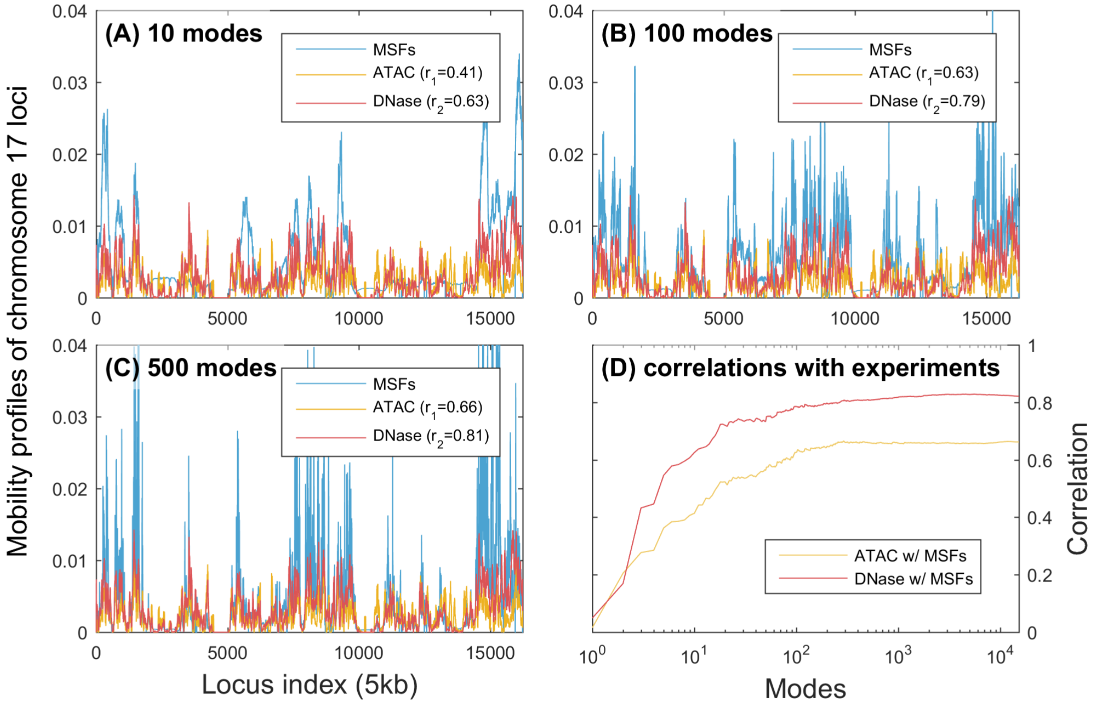

Figure S2. Contributions of different subsets of modes to the mobility profile of chromosome 17. (A) – (C) Comparisons between experimental data and computed MSF profiles obtained using 10, 100, and 500 GNM modes in the computations. **(D)** Spearman correlations between experimental and computationally predicted fluctuation/accessibility profiles obtained with different numbers of modes. Note that the abscissa is in logarithmic scale. The correlation levels off at around a few hundreds of modes.

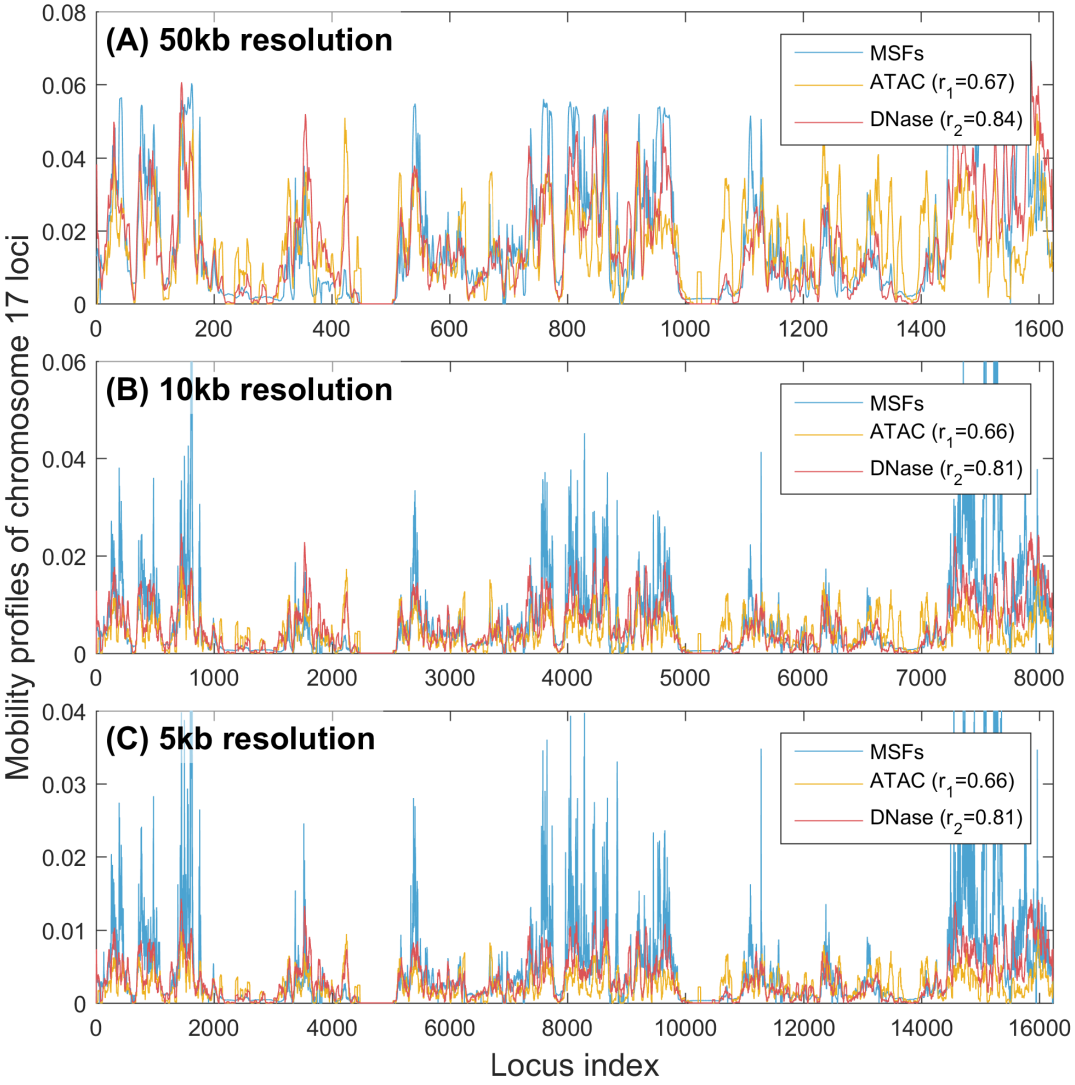

Figure S3. Mobility profile of chromosome 17 predicted by the GNM based on Hi-C maps at different resolutions. The three panels display the correlations between chromatin accessibility data (ATAC and DNase-seq) and GNM-predicted fluctuation profiles based on the Hi-C contact map for chromosome 17 at **(A)** 50kb, **(B)** 10kb, and **(C)** 5kb resolution. GNM results are computed using 500 lowest-frequency modes. The level of agreement between computational predictions and experimental observations is insensitive to the resolution of experimental data.

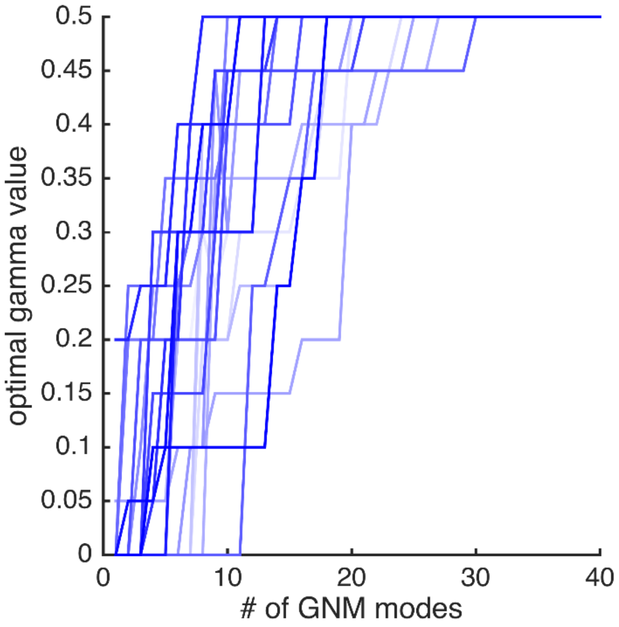

Figure S4. Armatus gamma values as a function of GNM modes. The gamma values corresponding to the lowest VI value for each GNM mode increase monotonically with the number of modes used, showing that higher resolution domains can be found by using higher GNM modes, consistent with decreasing granularity of GNM-predicted substructures with increasing mode number.

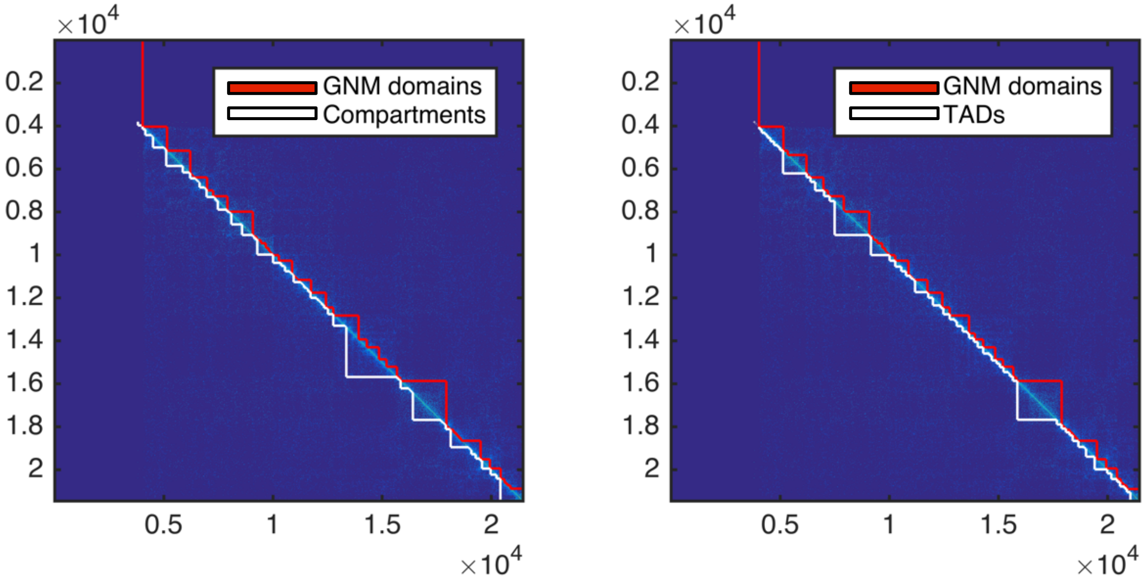

Figure S5. Comparison of GNM domains with (A) compartments and (B) TADs for chromosome 14. In both panels, the background is a heat map of the Hi-C contact matrix for this chromosome, and the *red* and *white* lines represent the domains identified by the two indicated methods. The two axes represent the loci numbers. Data for compartments are from the work of Lieberman-Aiden et al. (4). TADs are computed using Armatus (5).

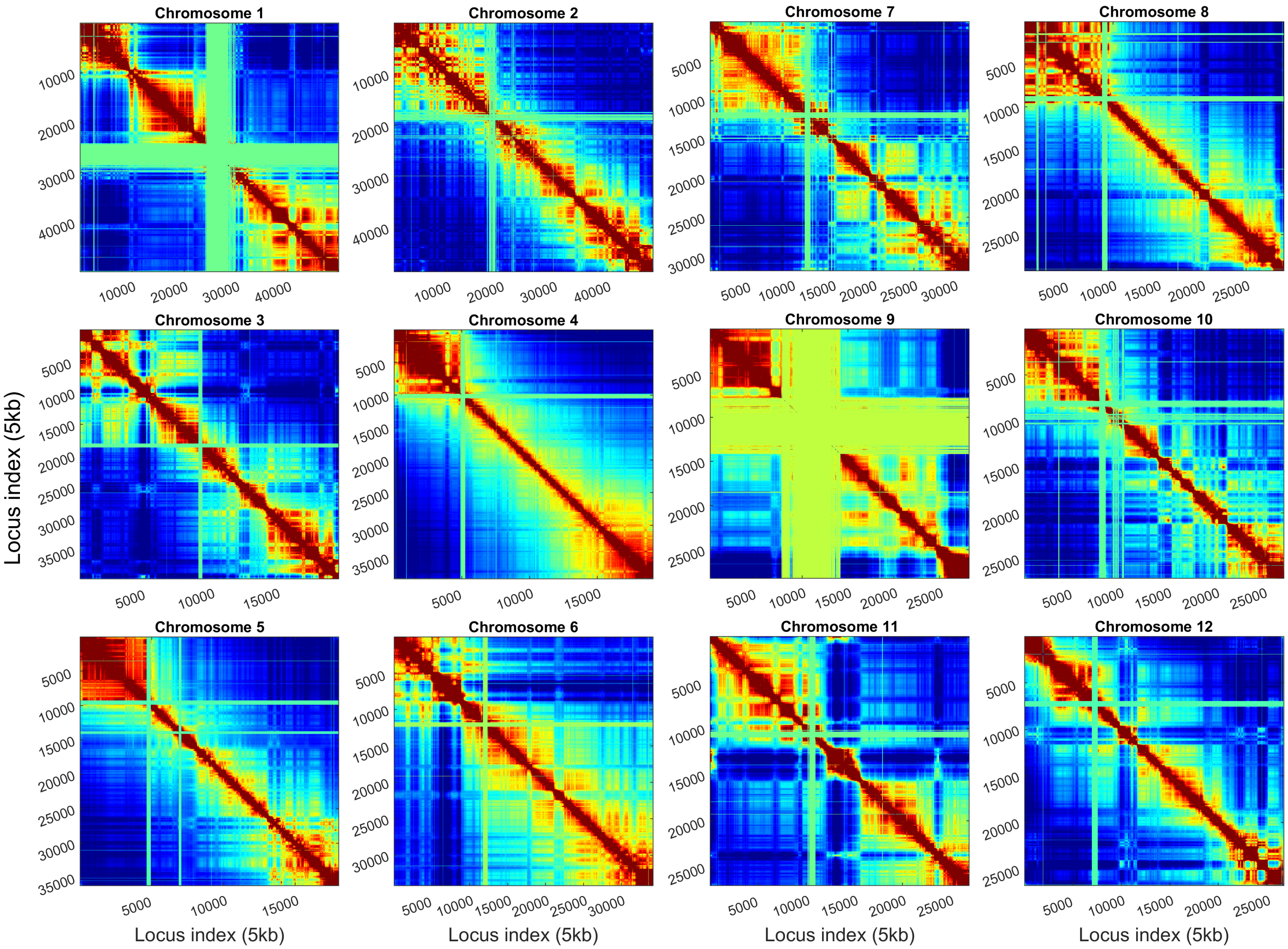

Figure S6 (first part)

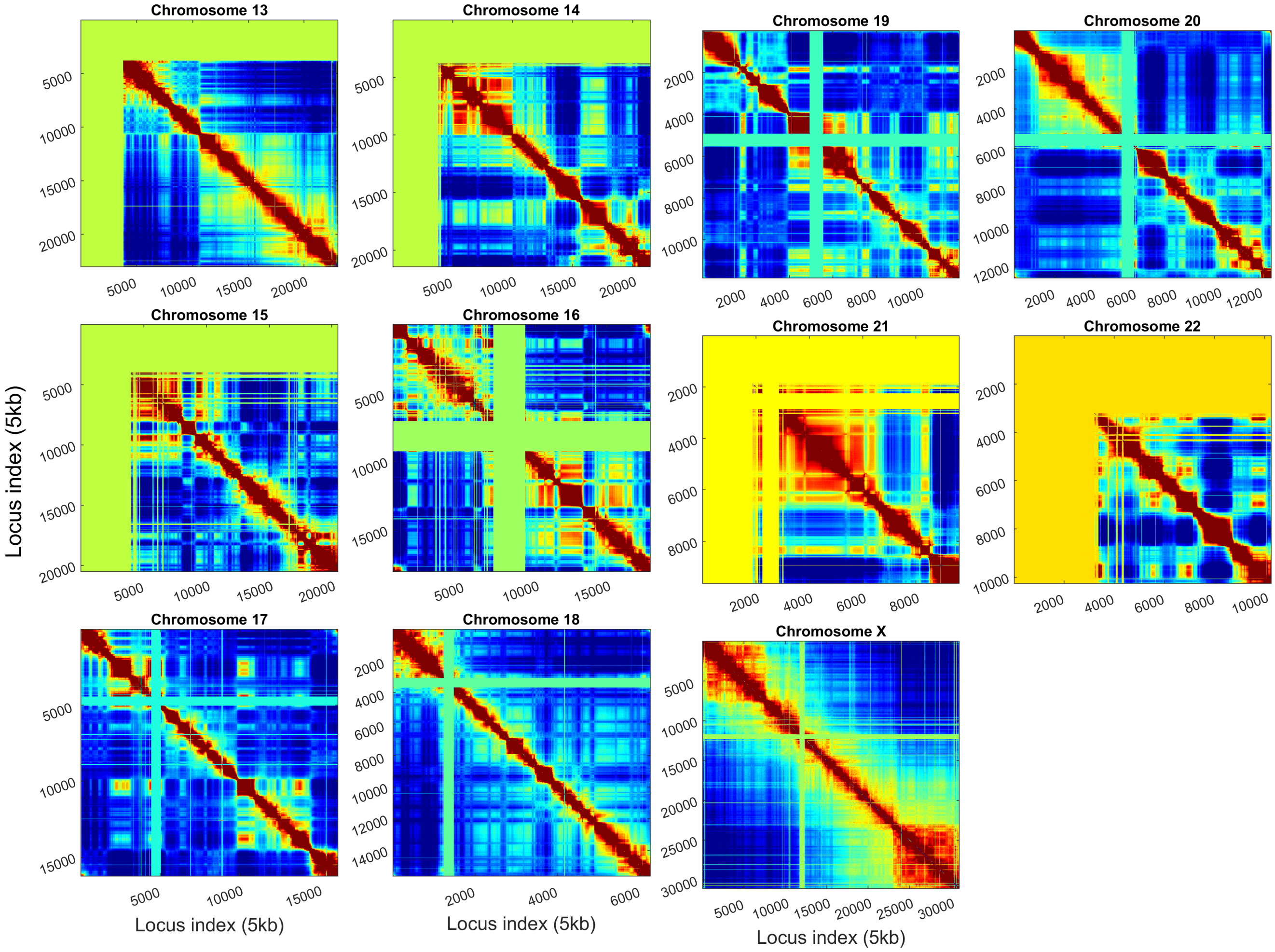

Figure S6. Intra-chromosomal covariance of gene loci computed for all chromosomes at 5kb resolution. The entries in the map display the type and strength of correlations between the gene loci indicated along the two axes. The maps are color-coded from *dark red* the *dark blue*, with *dark red* indicating the gene loci pairs that show the strongest cross-correlations in their spatial movements (same direction, same sense movements in space), and *dark blue* regions refer to gene pairs undergoing anticorrelated movements (same direction, opposite sense). *Green/yellow bands* refer to regions that lack Hi-C contact data. The *red blocks* along the diagonal are indicative of highly coupled clusters of loci. Results are obtained using all GNM modes for the individual chromosomes.

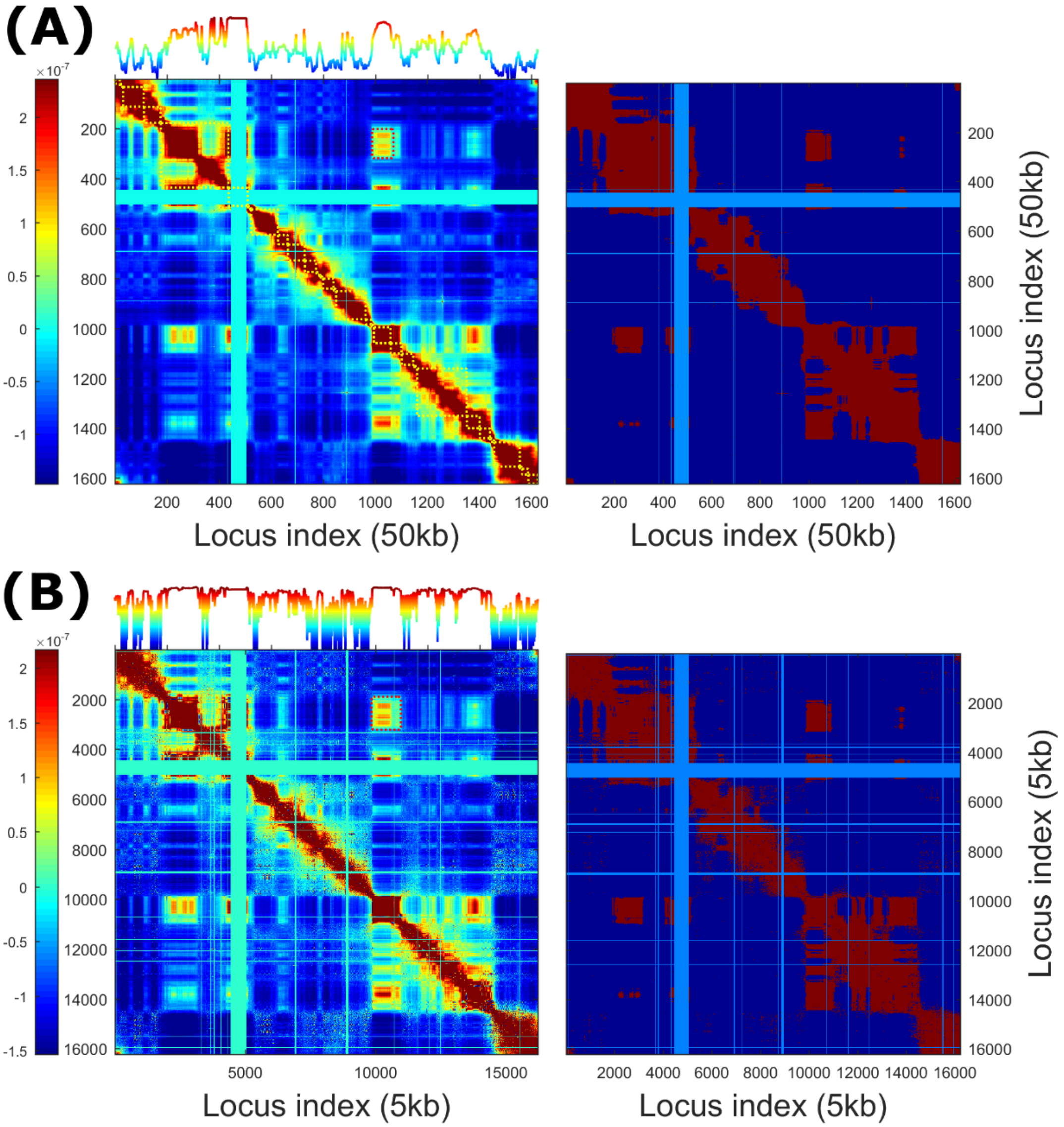

Figure S7. Reproducibility of the covariance map computed for chromosome 17 using two different levels of resolution. **(A)** Results at 50kb resolution computed using all GNM modes, **(B)** Results at 5kb resolution obtained with 500 slowest modes. The maps on the *right* of the covariance maps show the sign of the covariance. *Red* indicates positive, *blue* indicates negative. Most of the positively correlated gene loci are contiguous along the chromosome, except for a few off-diagonal islands which correspond to CCDs. The curve along the upper abscissa represent the average covariance of corresponding loci (averaged over its correlations with all other loci). Maxima indicate gene loci that are engaged in strong couplings with other loci.

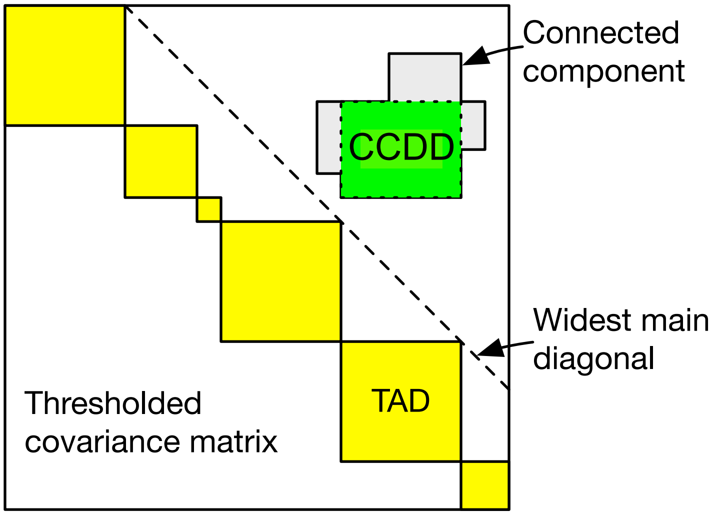

Figure S8. Identification of cross-correlated distal domains (CCDDs). CCDDs are found by searching for connected components outside of the widest point of the main diagonal. The CCDD is then the rectangle of maximal area contained entirely within the connected component.

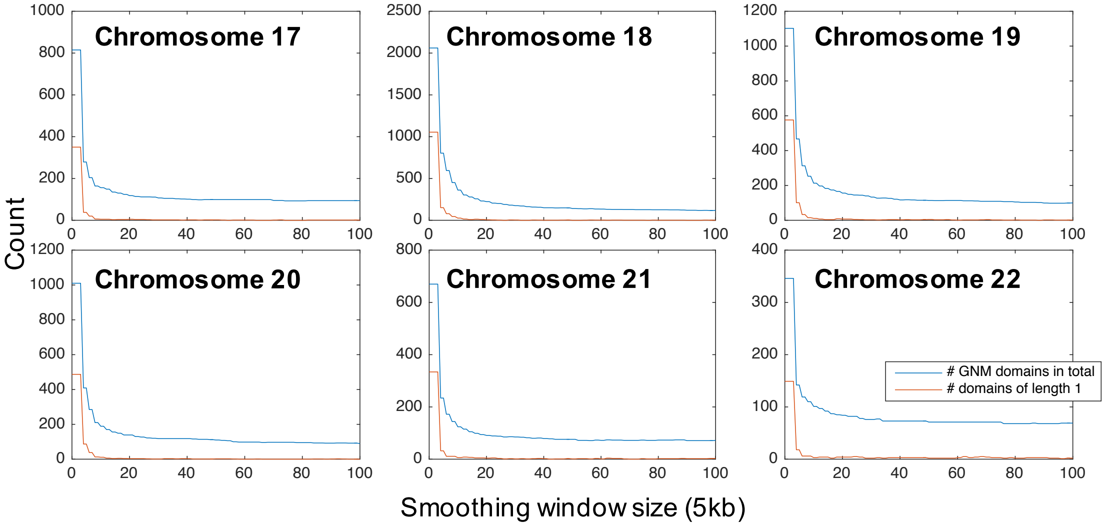

Figure S9. The number of GNM domains in total and of length one varied by smoothing window size. A drastic decrease of domains of length one can be seen around 10-20. This critical value is consistent on all chromosomes (only 17-22 are shown here). The smoothing window was chosen to be 16 for eigenvectors obtained from Hi-C data at 5kb resolution.

Methods

Gaussian Network Model

Gaussian network model (GNM) is a bead-and-spring representation of biological macromolecules(6-8). A common representation of protein is that each residue is a node, and two residues is connected by a spring if their spatial distance is close. To adapt the modeling of chromatin structure, the nodes represent loci and the strength of the springs indicate the intensity of interactions.

The major component of GNM is the **Kirchhoff matrix** (**Γ**, also termed as Laplacian matrix) representing the connectivity of loci. The matrix can be easily constructed from Hi-C contact map **M**, where each entry $\mathbf{M}_{ij}$ is the number of interactions between two locus *i* and locus *j*:

$$\mathbf{\Gamma}_{ij} = \begin{cases} -\mathbf{M}_{ij}, & (i \neq j) \\ -\sum_{k,\, k \neq i} \mathbf{\Gamma}_{ik}, & (i = j) \end{cases}$$

Intuitively, the off-diagonal elements each are the negative of the value at the corresponding position in Hi-C map, and the diagonal elements are the negative summation of the row or column where the element is located in the Kirchhoff matrix. In this way, unlike GNM on proteins, we use non-uniform force constants for chromosomes.

The diagonalization of the Kirchhoff matrix results in **eigenvectors and eigenvalues**:

$$\mathbf{\Gamma} = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathrm{T}}$$

Suppose n is the number of nodes in the system, then **V** is a unitary matrix where each column is an eigenvector $\boldsymbol{u}_i$ $(0 < i \leq n)$. **D** is a diagonal matrix of eigenvalues, λ_1, λ_2, ..., λ_n, usually ordered ascendingly. An eigenvector and its eigenvalue is a GNM mode, of which the pattern of the motion of each node will be manifested by the eigenvector, and the eigenvalue is squared vibrational frequency of the corresponding mode. Therefore, the modes with small eigenvalue are slow modes with low frequency but large amplitude given conserved energy. These slow modes describe global and collective motions of the macromolecule, which are usually more biologically meaningful than local motions described by fast modes. In addition, in the case of GNM, there is always one zero eigenvalue corresponding to the rigid translation of the entire system. If there are more than one zero eigenvalue, it usually indicates the system contains disconnected regions.

Another important output of GNM is the **covariance matrix**. It can be proved that the inverse of Kirchhoff matrix is proportional to the covariance matrix of displacements of interacting nodes:

$$\langle \Delta\mathbf{R} \cdot \Delta\mathbf{R}^{T} \rangle = \frac{3k_B T}{\gamma}(\mathbf{\Gamma}^{-1})$$

Where k_B is the Boltzmann constant, T is the temperature, and γ is the spring constant. The covariance matrix can also be reconstructed by using all the modes $(m = n)$ or fewer modes $(m < n)$ with the following equation:

$$\langle \Delta\mathbf{R} \cdot \Delta\mathbf{R}^{T} \rangle = \sum_{i=1}^{m} \frac{\boldsymbol{u}_i \cdot \boldsymbol{u}_i^{T}}{\lambda_i}$$

Note that $\boldsymbol{u}_i$ is a $n \times 1$ column vector, therefore the product in the numerator is an $n \times n$ matrix. Usually the covariance matrix calculated from the slow modes shares most features with the full covariance matrix.

Square fluctuations are the diagonal values of the covariance matrix (i.e. the variances of the displacements of loci). Their values indicate the mobility of loci: higher the square fluctuation, higher the mobility.

Removal of Unmapped Regions

In the Hi-C map there are regions where no cross-linked DNA fragments can be mapped. These unmapped regions are isolated from the system, and their existence may lead to multiple zero-eigenvalue modes. These modes correspond to the rigid translation of isolated regions which will not cause conformational change, therefore they are biologically meaningless. In addition, these unmapped regions are not constrained by any other loci, so they may cause large fluctuations that flatten the signal from other regions. These extra zero-eigenvalue modes and unphysically large fluctuations can be effectively removed by simply discarding the unmapped regions. Note that the removal of the unmapped regions will not cause disconnections because chromosomes are highly compact, so the loci next to the unmapped regions are still connected to the loci located on the other end of the region.

Normalization

Two types of normalization methods were applied to the Hi-C contact map: Vanilla-Coverage normalization (referred as VCnorm) (4) and Knight-Ruiz normalization (referred as KRnorm) (9). Both methods aim to eliminate the so-called “one-dimension bias” (1). However, we found that GNM performed much better on Hi-C map normalized by VCnorm. Not only are the correlations with the chromatin accessibility lower, but also the square fluctuations become flatter and flatter by adding more modes in the calculation when KR normalization has been applied on the contact map. In the extreme case, when all the modes are used, the square fluctuations become almost completely flat along the chromosome. This is because KRnorm ensures that every row and column sums to 1. By setting the same sum for each row or column, all the loci will be constrained by the approximately same number of springs. Consequently, the mobilities measured by square fluctuations calculated based on all the modes will be similar for all the loci. Due to the better performance and fit with the theory of GNM, we chose VC normalized contact maps to perform further analyses.

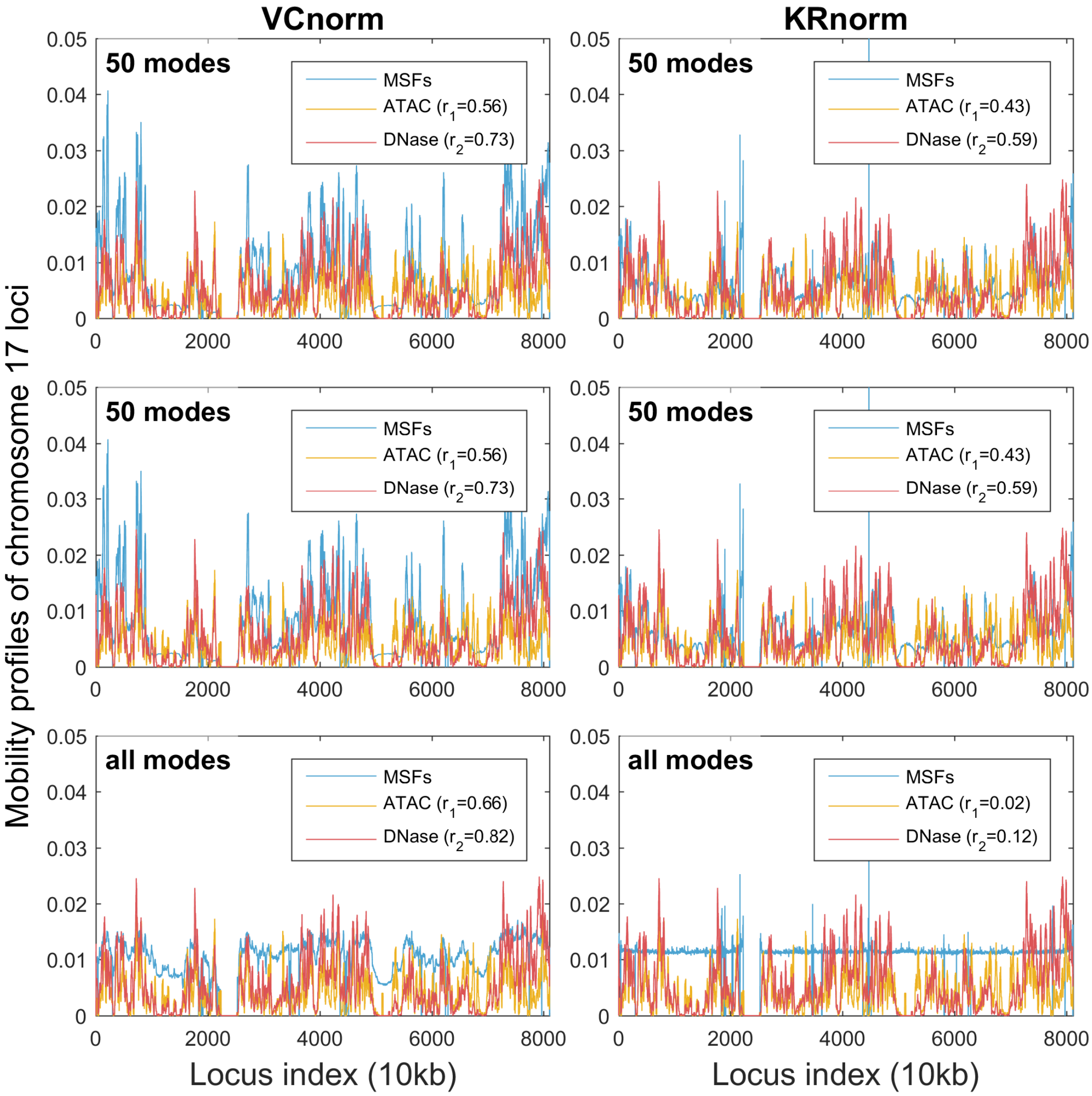

Figure S12. Comparison of the results obtained with two different normalization methods: Vanilla Coverage normalization (left) and Knight-Ruiz (right) normalization.

Table S1. List of Sequence Read Run (SRR) IDs for all 212 RNA-seq experiments from the Sequence Read Archive used in co-expression calculations.(*)

SRR038295	SRR038448	SRR038449	SRR065510	SRR065514
SRR065515	SRR065532	SRR089332	SRR089333	SRR1024156
SRR1024157	SRR1066622	SRR1066623	SRR1066624	SRR1066625
SRR1066626	SRR1066627	SRR1066628	SRR1066629	SRR1066630
SRR1066631	SRR1066632	SRR1066633	SRR1066634	SRR1066635
SRR1066636	SRR1066637	SRR1066638	SRR1066639	SRR1066640
SRR1066641	SRR1153470	SRR1163655	SRR1293901	SRR1293902
SRR1803196	SRR1803197	SRR1803198	SRR1909074	SRR1909076
SRR1909078	SRR1909107	SRR1909108	SRR1909113	SRR1983907
SRR1983908	SRR1983909	SRR2192704	SRR2192705	SRR2192706
SRR2192707	SRR2192708	SRR2192709	SRR2192710	SRR2192711
SRR2192712	SRR2192713	SRR306998	SRR306999	SRR307000
SRR307001	SRR307002	SRR307003	SRR307004	SRR307005
SRR307006	SRR307007	SRR307008	SRR307009	SRR307010
SRR307011	SRR307012	SRR307897	SRR307898	SRR307899
SRR307900	SRR307921	SRR307922	SRR315297	SRR315298
SRR317058	SRR317059	SRR317060	SRR317061	SRR3191739
SRR3191740	SRR3191773	SRR3191774	SRR3191775	SRR3191776
SRR3191777	SRR3191778	SRR3191779	SRR3191849	SRR3192069
SRR3192132	SRR3192133	SRR3192134	SRR3192135	SRR3192136
SRR3192137	SRR3192138	SRR3192139	SRR3192140	SRR3192218
SRR3192396	SRR3192397	SRR3192398	SRR3192399	SRR3192400
SRR3192401	SRR3192402	SRR3192403	SRR3192406	SRR3192407
SRR3192657	SRR3192658	SRR363871	SRR390498	SRR390507
SRR390508	SRR390509	SRR390510	SRR390511	SRR390512
SRR390513	SRR390514	SRR390517	SRR390542	SRR390543
SRR390544	SRR390545	SRR521447	SRR521448	SRR521449
SRR521450	SRR521451	SRR521452	SRR521453	SRR521454
SRR521455	SRR521456	SRR521466	SRR521467	SRR521510
SRR521511	SRR521512	SRR527657	SRR527658	SRR527677
SRR527678	SRR530637	SRR530638	SRR545687	SRR545688
SRR549363	SRR549364	SRR576703	SRR764776	SRR764777
SRR764778	SRR764779	SRR764780	SRR764781	SRR764782
SRR764783	SRR764784	SRR764785	SRR764786	SRR764787
SRR764788	SRR764789	SRR764790	SRR764791	SRR764792
SRR764793	SRR764794	SRR764795	SRR764796	SRR764797
SRR764798	SRR764799	SRR764800	SRR764801	SRR764802
SRR764803	SRR764804	SRR764805	SRR764806	SRR764807
SRR764808	SRR764809	SRR764810	SRR764811	SRR764812

SRR764813	SRR764814	SRR764815	SRR764816	SRR764817
SRR768411	SRR768412	SRR972706	SRR972707	SRR972712
SRR972713	SRR972714	SRR972715	SRR972716	SRR972717
SRR975411	SRR975412			

(*) the data can be found at http://www.ncbi.nlm.nih.gov/sra

1. Rao SS, *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665-1680.
2. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, & Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10(12):1213-+.
3. Song L & Crawford GE (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols* 2010(2):pdb. prot5384.
4. Lieberman-Aiden E*, et al.* (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326(5950):289-293.
5. Filippova D, Patro R, Duggal G, & Kingsford C (2014) Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology* 9.
6. Bahar I, Atilgan AR, & Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & Design* 2(3):173-181.
7. Bahar I, Lezon TR, Yang LW, & Eyal E (2010) Global Dynamics of Proteins: Bridging Between Structure and Function. *Annual Review of Biophysics, Vol 39* 39:23-42.
8. Haliloglu T, Bahar I, & Erman B (1997) Gaussian dynamics of folded proteins. *Physical Review Letters* 79(16):3090-3093.
9. Knight PA & Ruiz D (2013) A fast algorithm for matrix balancing. *Ima Journal of Numerical Analysis* 33(3):1029-1047.