Supplementary Data

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, Pittsburgh, PA, 15213, USA*
[2] *Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA*

Supplementary Methods

Gaussian Network Model (GNM)

The GNM is an elastic network model developed for characterizing the collective dynamics of biomolecular structures using information on their inter-residue contact topology (1-3). Each node in the GNM is identified by a residue (n of them for a structure of n residues), and residue pairs i and j are connected by an elastic springs of force constant γ_{ij} provided that they are located within an interaction range (e.g. closer than a cutoff distance r_{cut}).

The major ingredient of the GNM is the n x n Kirchhoff matrix, $\mathbf{\Gamma}$, the off-diagonal elements of which account for the stiffness of interactions (see Eq 1). In the GNM, a uniform force constant $\gamma_{ij} = \gamma$ is adopted for all pairs such that $\mathbf{\Gamma}$ becomes equivalent to a connectivity matrix or Laplacian commonly used in graph theory, multiplied by the spring constant γ. The network is assumed to be at a global energy minimum under equilibrium conditions. The movements of the nodes away from their equilibrium positions entail a potential of the form

$$V_{GNM} = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbf{\Gamma}_{ij} \left(\Delta \boldsymbol{r}_i - \Delta \boldsymbol{r}_j\right)^2 = \tfrac{1}{2}\ [\Delta \mathbf{R}]^T\ \mathbf{\Gamma}\ [\Delta \mathbf{R}] \qquad \text{(S1)}$$

where $\Delta \boldsymbol{r}_i$ represents the displacement of node i with respect to its equilibrium position, $[\Delta \mathbf{R}]^T$ is the n-dimensional row vector $[\Delta \boldsymbol{r}_1\ \Delta \boldsymbol{r}_2\ \ldots . \Delta \boldsymbol{r}_n]$ of the displacement of all nodes, $[\Delta \mathbf{R}]$ is the corresponding column vector. Using Eq S1, the cross-correlations $< \Delta \boldsymbol{r}_i\ .\ \Delta \boldsymbol{r}_j >$ between the fluctuations of the nodes i and j can be evaluated as a thermodynamic average over all fluctuations, to obtain (1-3)

$$< \Delta \boldsymbol{r}_i\ .\ \Delta \boldsymbol{r}_j > = \ 3k_B T\ [\mathbf{\Gamma^{-1}}]_{ij} \qquad \text{(S2)}$$

Here k_B is the Boltzmann coefficient, T is the absolute temperature, and $[\mathbf{\Gamma^{-1}}]_{ij}$ and designates the ij^{th} element of the inverse of $\mathbf{\Gamma}$. The evaluation of the cross-correlation therefore requires the inversion of $\mathbf{\Gamma}$. But $\mathbf{\Gamma}$ is not invertible (its diagonal elements are evaluated as the negative sum of off-diagonal terms in the same row (see Eq 1). Instead, we evaluate the pseudoinverse. To this aim, we first diagonalize $\mathbf{\Gamma}$

and reconstruct it (or its pseudoinverse) after removing its zero eigenvalue. Diagonalization of $\boldsymbol{\Gamma}$ yields a unitary matrix, **U**, composed of the eigenvectors, $\boldsymbol{u}_k$, of $\boldsymbol{\Gamma}$, and a diagonal matrix $\boldsymbol{\Lambda}$ composed of the eigenvalues λ_k (with $\lambda_0 = 0$ and $\lambda_1 \leq \lambda_2 \ldots \leq \lambda_{n-1}$)

$$\boldsymbol{\Gamma} = \mathbf{U}\,\boldsymbol{\Lambda}\,\mathbf{U}^{\mathsf{T}} = [\boldsymbol{u}_0 \quad \boldsymbol{u}_1 \quad \cdots \quad \boldsymbol{u}_{n-1}] \begin{bmatrix} \lambda_0 & & & \\ & \lambda_1 & & \\ & & \ddots & \\ & & & \lambda_{n-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_0{}^T \\ \boldsymbol{u}_1{}^T \\ \vdots \\ \boldsymbol{u}_{n-1}{}^T \end{bmatrix} = \sum_{k=1}^{n-1} \lambda_k \, [\boldsymbol{u}_k \, \boldsymbol{u}_k{}^T] \tag{S3}$$

Here $\lambda_0 = 0$. The nonzero eigenvectors $\boldsymbol{u}_k$ $(1 \leq k \leq n\text{-}1)$ form an orthonormal basis set representing the collective displacements or fluctuations (of all nodes) along a given mode (*k*) axis, and the eigenvalue λ_k scales with the squared frequency of collective fluctuations along mode *k*. Using Eq S3, the pseudoinverse of $\boldsymbol{\Gamma}$ is obtained as

$$\boldsymbol{\Gamma}^{-1} = \mathbf{U}\,\boldsymbol{\Lambda}^{-1}\,\mathbf{U}^{\mathsf{T}} = \sum_{k=1}^{n-1} 1/\lambda_k \, [\boldsymbol{u}_k \, \boldsymbol{u}_k{}^T] \tag{S4}$$

which, substituted in Eq S2, yields the Eq 2 in the main text. The modes with smaller eigenvalue, or lower frequency describe the so-called *global* or *soft* motions of the macromolecule, which embody the entire structure. These modes are usually relevant to biological function. The other end of the spectrum refers to local motions described by *fast* modes. The zero eigenmode corresponds to the rigid translation of the system. If there are more than one zero eigenvalue, it usually indicates that the system contains disconnected regions.

Cross-correlations are organized in a *n* x *n covariance matrix*

$$\mathbf{C} \equiv \langle \Delta\mathbf{R}\, \Delta\mathbf{R}^T \rangle = 3k_B T \, [\boldsymbol{\Gamma}^{-1}] = \sum_{k=1}^{m} \frac{\boldsymbol{u}_k \, \boldsymbol{u}_k^T}{\lambda_k} \tag{S5}$$

$\mathbf{C}$ can be reconstructed using all modes ($m = n - 1$) or fewer modes ($m < n - 1$) with the following equation: Note that $\boldsymbol{u}_i$ is a $n \times 1$ column vector, therefore the dyadic product in the numerator is an $n \times n$ matrix. Usually a subset of slow modes provides a good representation of the full covariance. The diagonal elements of **C** represent the mean-square fluctuations (MSFs) in the positions of the nodes. The plot of MSFs as a function of node index yields the *mobility profile* of the structure.

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 unmapped regions are not constrained by other loci, so they may cause large fluctuations that obscure the signal from other regions. These extra zero-eigenvalue modes and unphysically large fluctuations were removed by discarding the unmapped regions. Note that the removal of the unmapped regions will not cause disconnections because the chromosomes are highly compact, so the loci next to the unmapped regions remained connected to the loci located at the other end of the region.

Hi-C Data Normalization

We tested three types of normalization methods applied to the Hi-C contact map: Vanilla-Coverage normalization (referred to as VCnorm), square-root Vanilla-Coverage normalization (referred to as sqrtVC) (4) and Knight-Ruiz normalization (referred to as KRnorm) (5). All three methods aim to eliminate the so-called "one-dimension bias" (6). We found that the GNM performed best on Hi-C maps normalized by VCnorm when benchmarked against experimental data (**Fig. S5-S7**). 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 KRnorm or sqrtVC 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 using KRnorm. This is because KRnorm ensures that every row and column sums to 1. As a consequence, all loci become almost equally constrained and the differences in their square fluctuations are suppressed.

In addition, computations with the three normalization methods were repeated at different resolutions, and VCnorm yielded the most robust agreement between theoretically predicted MSFs and experimentally observed accessibilities across all resolutions. Both KRnorm and sqrtVC showed poor correlations at high resolution (5kb) (**Fig. S6** and **S7**). Furthermore, VCnorm showed the expected improvement in correlation using increasing number of modes included in the analysis, while KRnorm or sqrtVC led to inconsistent results, even at 50kb resolution (**Fig. S6**). Due to the better performance across resolutions and numbers of modes, shown by agreement with experimental data, we chose VC normalized contact maps to perform further analyses.

Variation of Information (VI) metric

This metric is based in information theory, and measures the difference in information contained in two clusterings, or partitions, of a data set. If we consider each domain to be a cluster of nodes/points, this type of comparison becomes very natural. Formally, for two sets of clusters *C* and *C'*, VI is defined as follows:

$$VI(C, C') = H(C) + H(C') - 2I(C, C') \quad \text{(S6)}$$

where $H(C)$ represents the entropy of a set of clusters C, and $I(C,C')$ is the mutual information between the two partitions, given by

$$H(C) = -\sum_{k=1}^{K} P(k) \log P(k) \quad \text{(S7)}$$

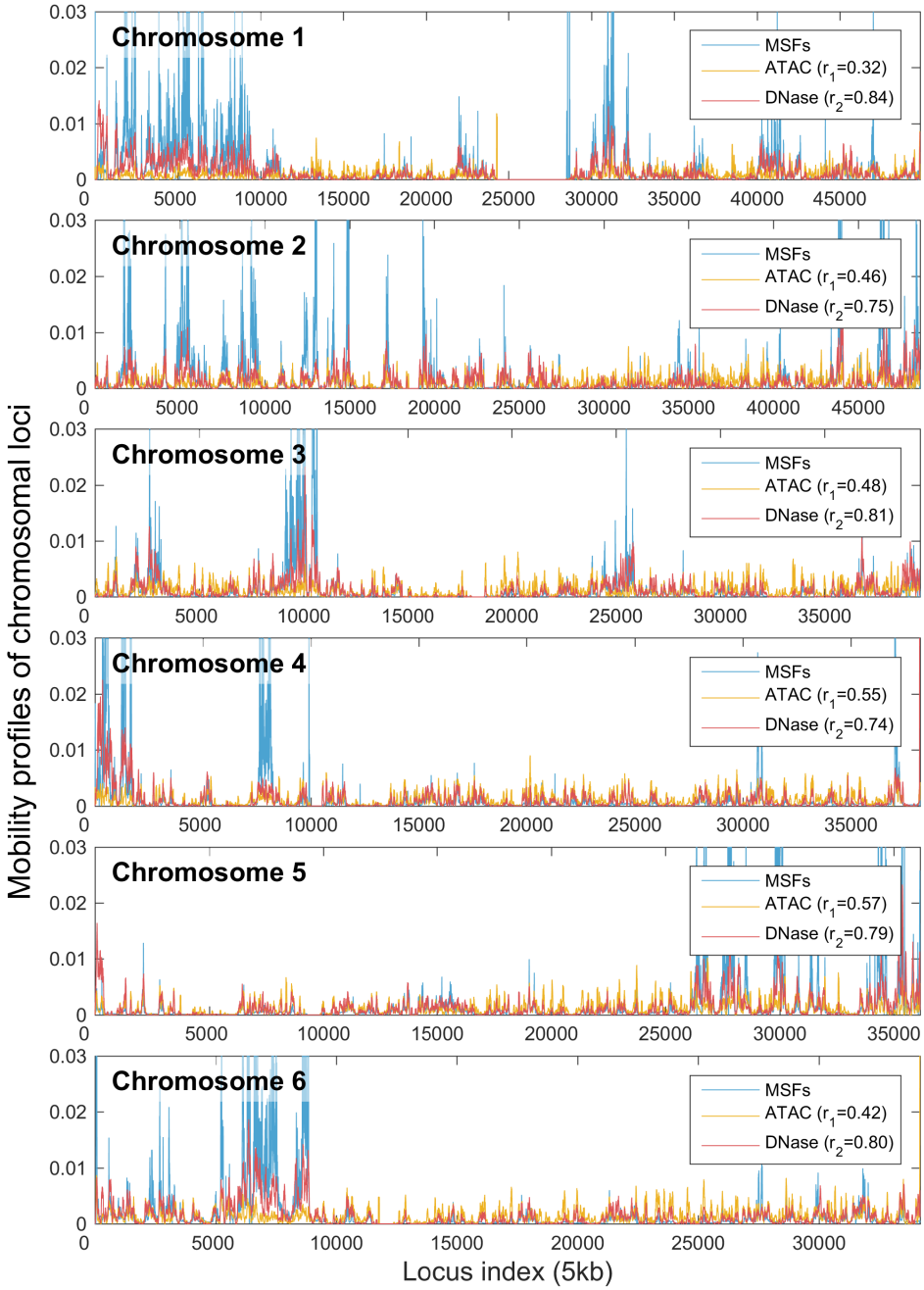
$$I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k,k')}{P(k)P(k')} \quad \text{(S8)}$$

where the probability of picking a node in cluster C_k, $P(k)$, is simply the number of points in that cluster divided by the total number of points in the data set. In this work, a "cluster" is the set of loci placed into the same domain or compartment.

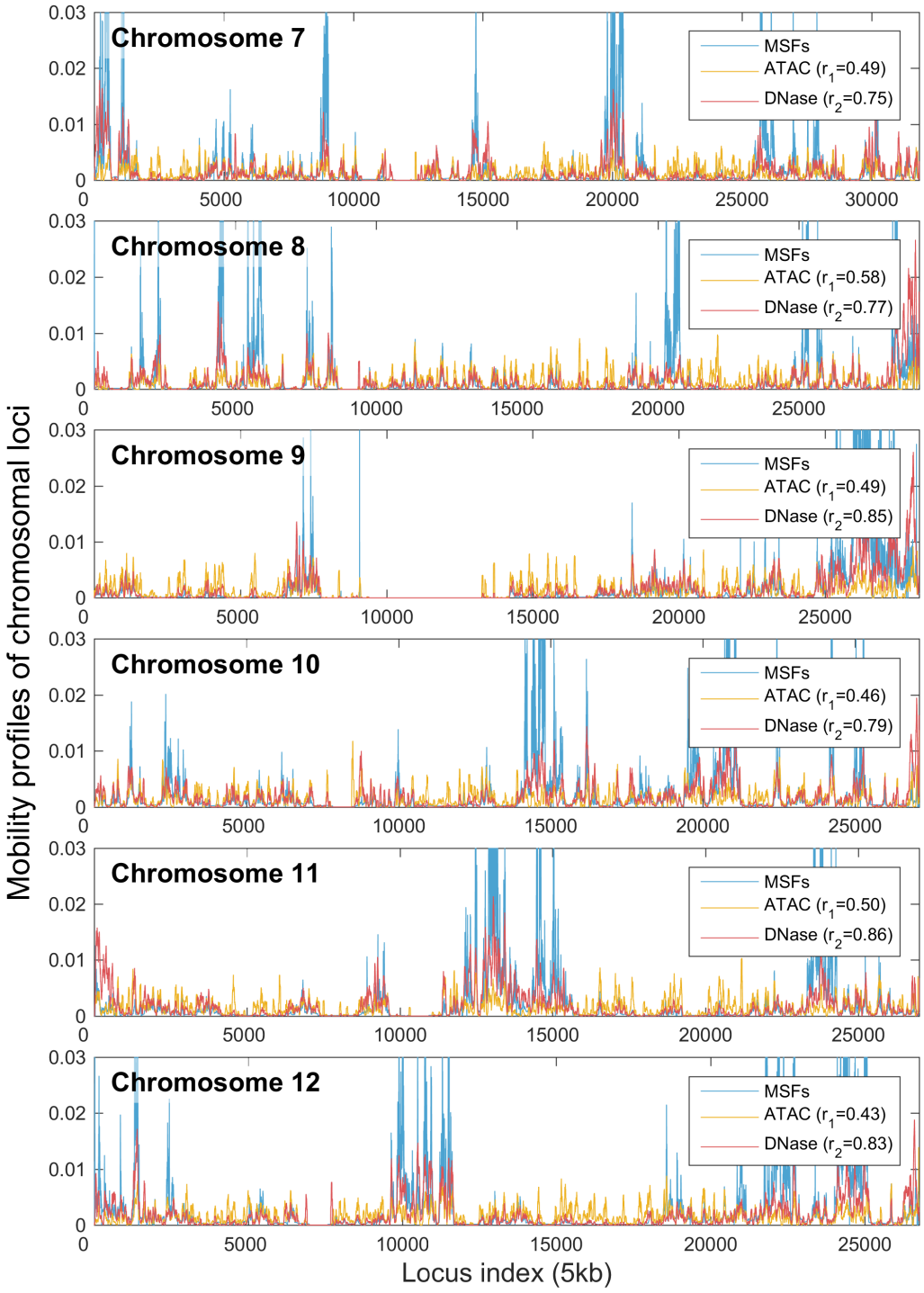
Note that this is a true metric on the space of clusterings; VI is commutative, satisfies the triangle inequality, and is always non-negative and equal to zero if and only if the two clusterings are identical.

More intuitively, VI is a measure of the amount of information that is lost and gained by changing from one clustering to another, without any assumptions placed on the clusterings themselves or how they were generated. More information can be found in (7).

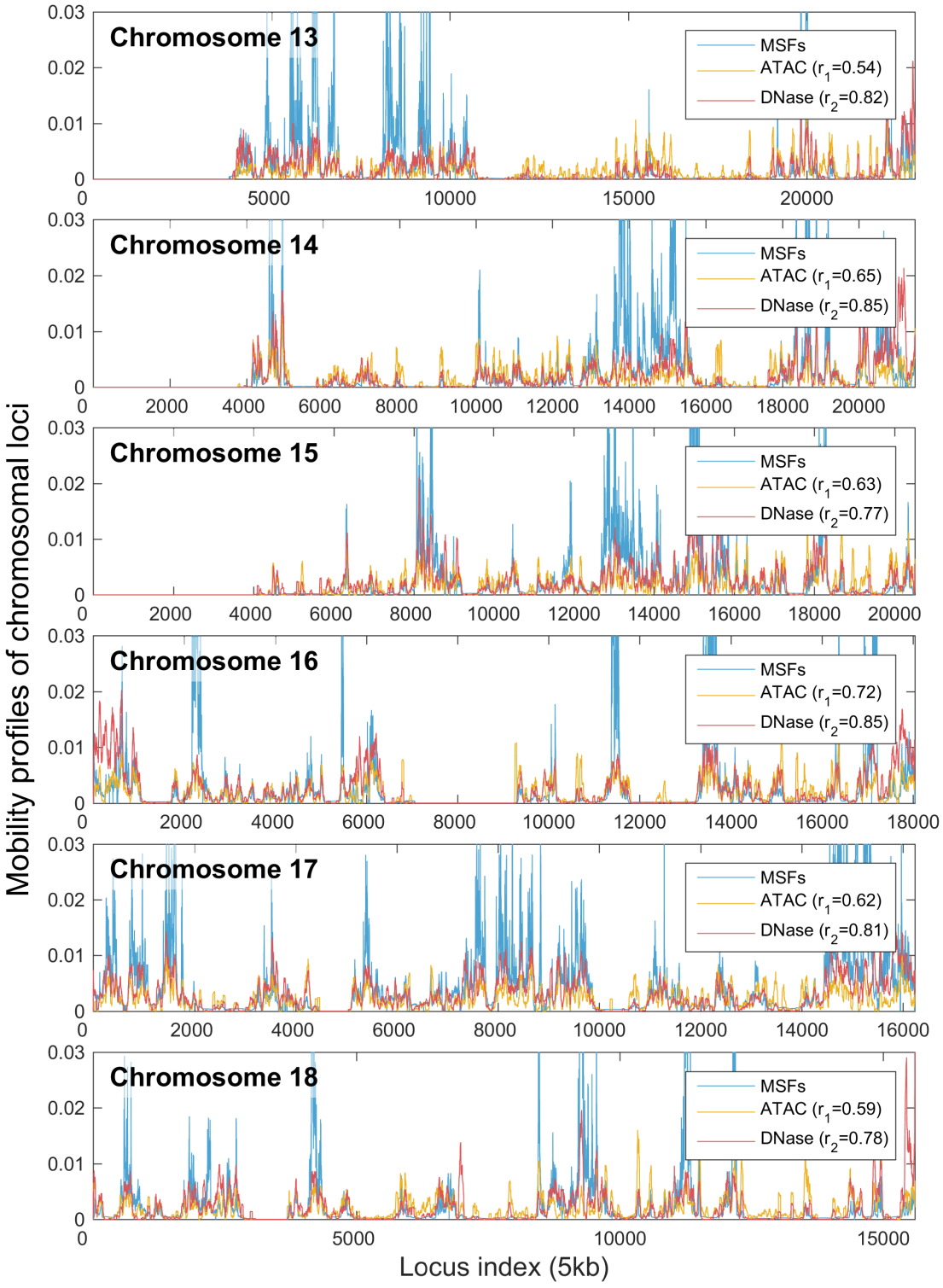
Supplementary Figures



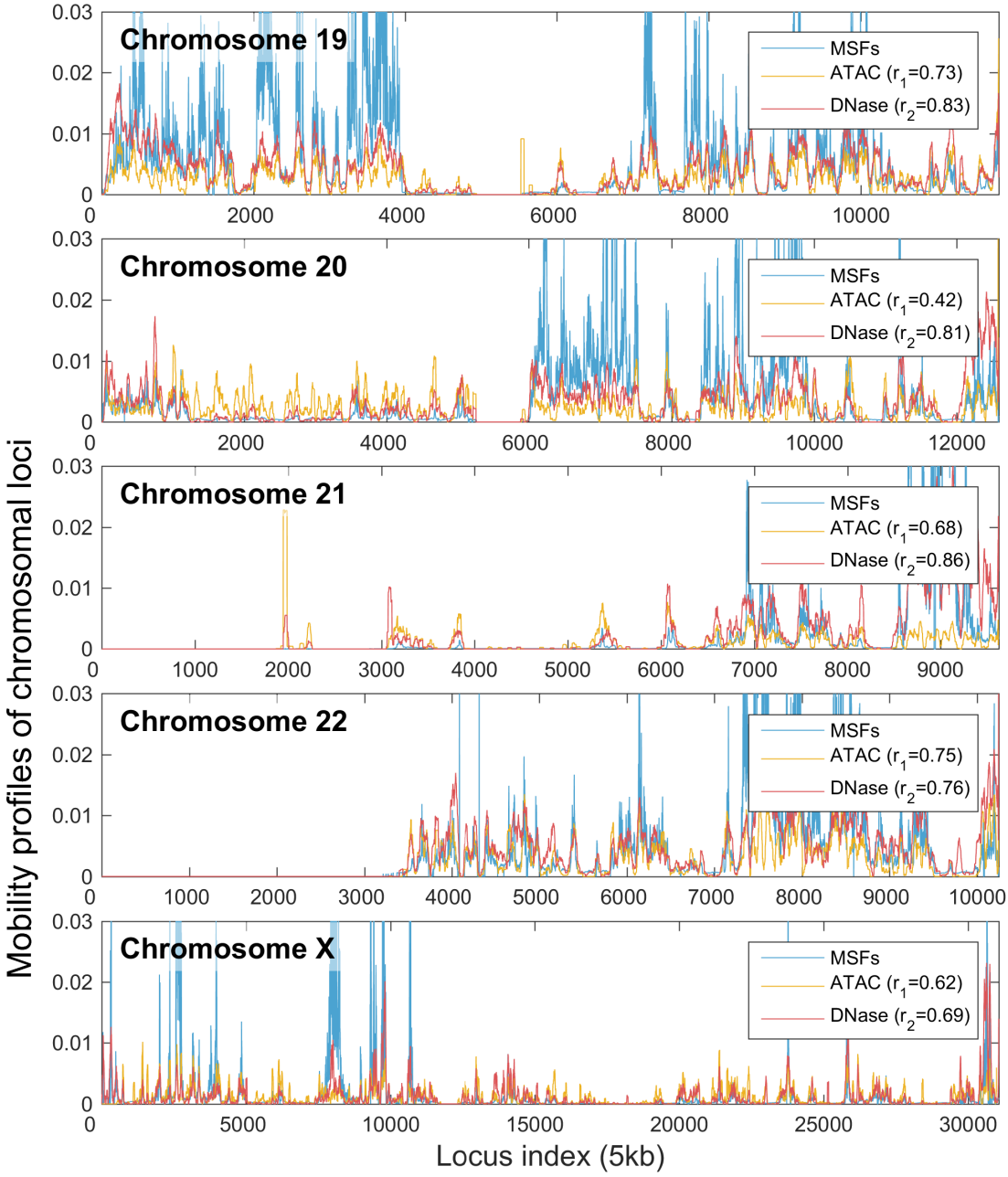
Supplementary Figure S1



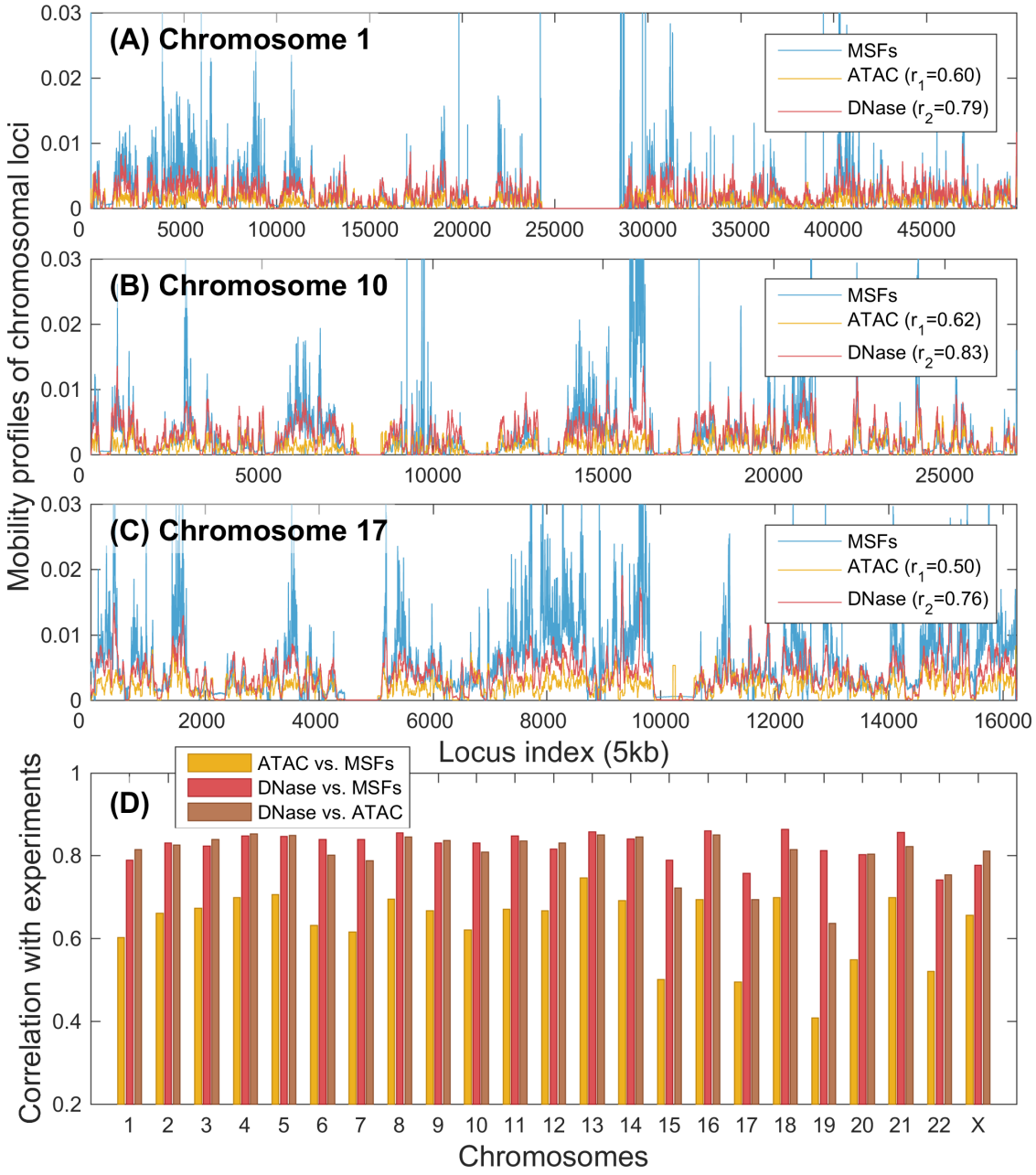
Supplementary Figure S1 (*continued*)



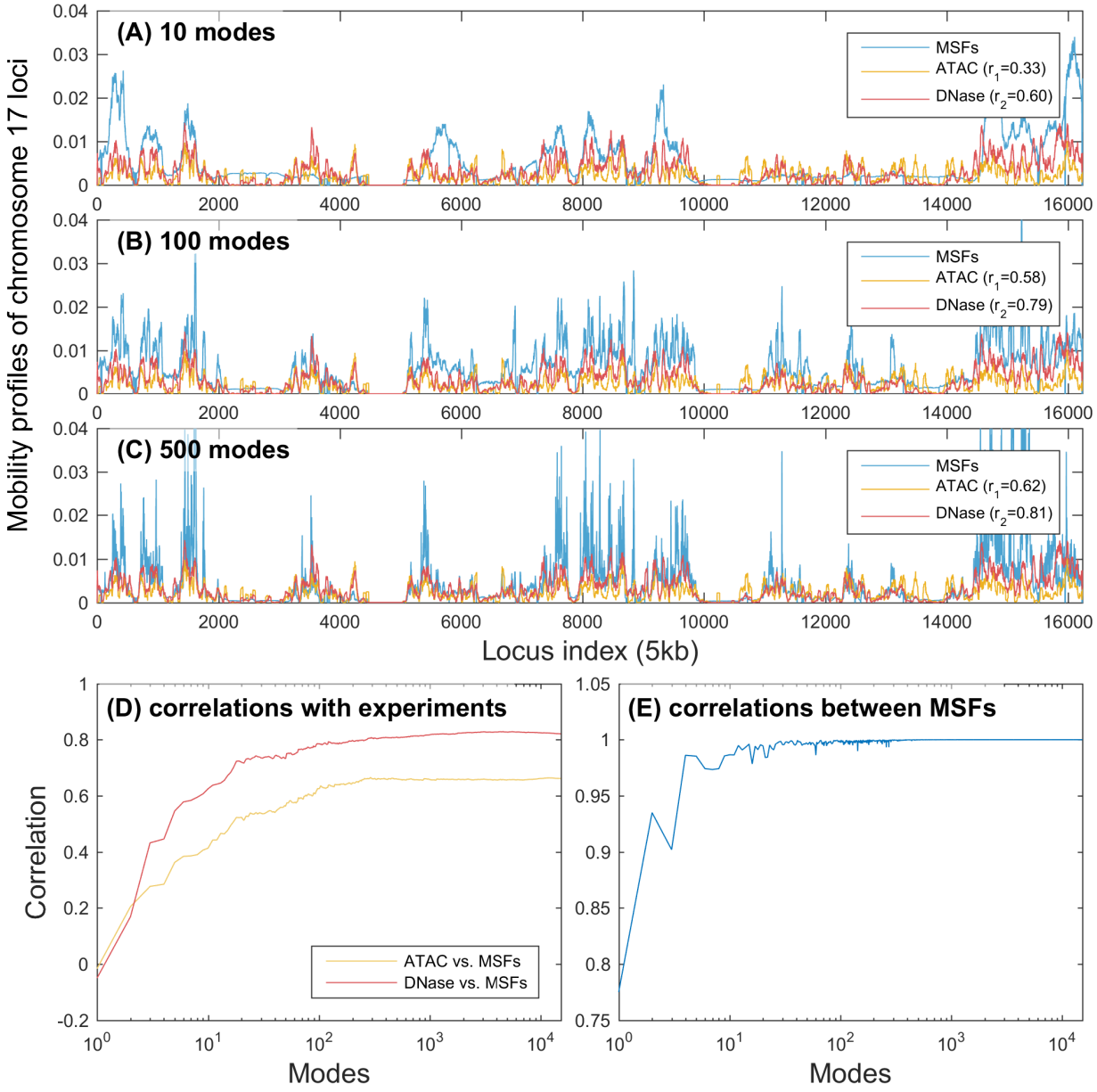
Supplementary Figure S1 (*continued*)



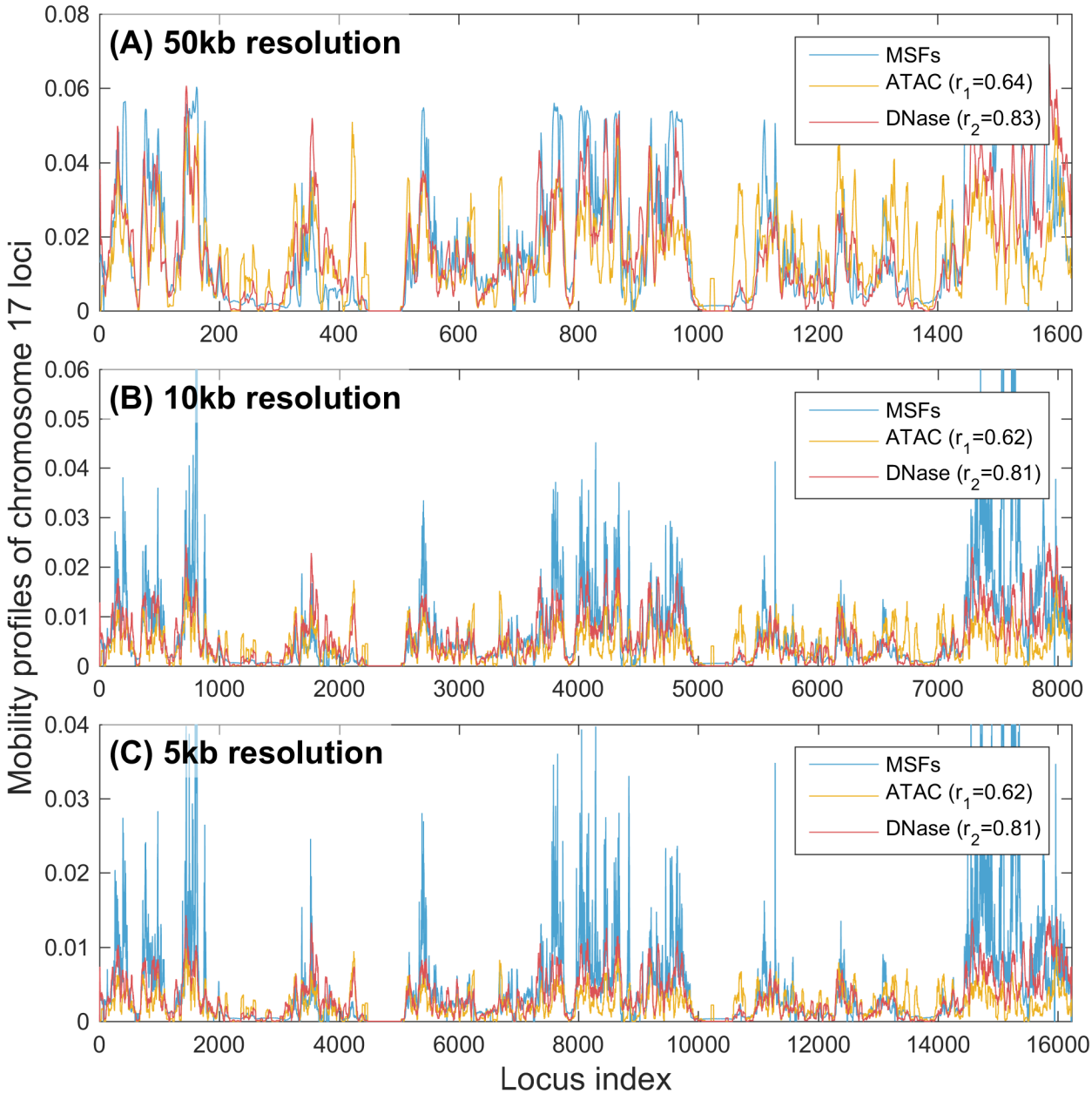
Supplementary Figure S1. GNM-predicted mobility profiles of gene loci compare favorably with accessibilities measured by ATAC and DNase-seq experiments in GM12878. Results are presented for all chromosomes. MSFs are based on 500 GNM modes (at the lowest frequency end of the spectrum) using Hi-C map at 5kb resolution obtained by Rao et al. for GM12878 (6). Spearman correlations between theoretical (MSFs) and experimental (ATAC (8) and DNase-seq (9)) data are shown for each chromosome in the corresponding legend box.



Supplementary Figure S2. GNM-predicted mobilities (MSFs) of chromosomal loci for IMR90 cells show good agreement with experimental data from chromatin accessibility experiments. (A) – (C) Mobility profiles obtained from GNM analysis of the equilibrium dynamics of chromosomes 1, 17, and X, respectively, shown in *blue*, are compared to the DNA accessibilities probed by ATAC-seq (*yellow*) and DNA-seq (*red*) experiments. GNM results are based on 500 slowest modes. r_1 is the Spearman correlations between GNM predictions and DNase-seq experiments; and r_2 is that between GNM and ATAC-seq. (D) Spearman correlations between theory and experiments for all chromosomes (red and yellow bars, as labeled). The Spearman correlation between the computed MSFs and experimental ATAC-seq data averaged over all chromosomes is 0.63 ± 0.08, and that between MSFs and DNase-seq data is 0.82 ± 0.03. For comparison, we also display the Spearman correlation between the two sets of experimental data (brown bars); the average in this case is 0.81 ± 0.06.

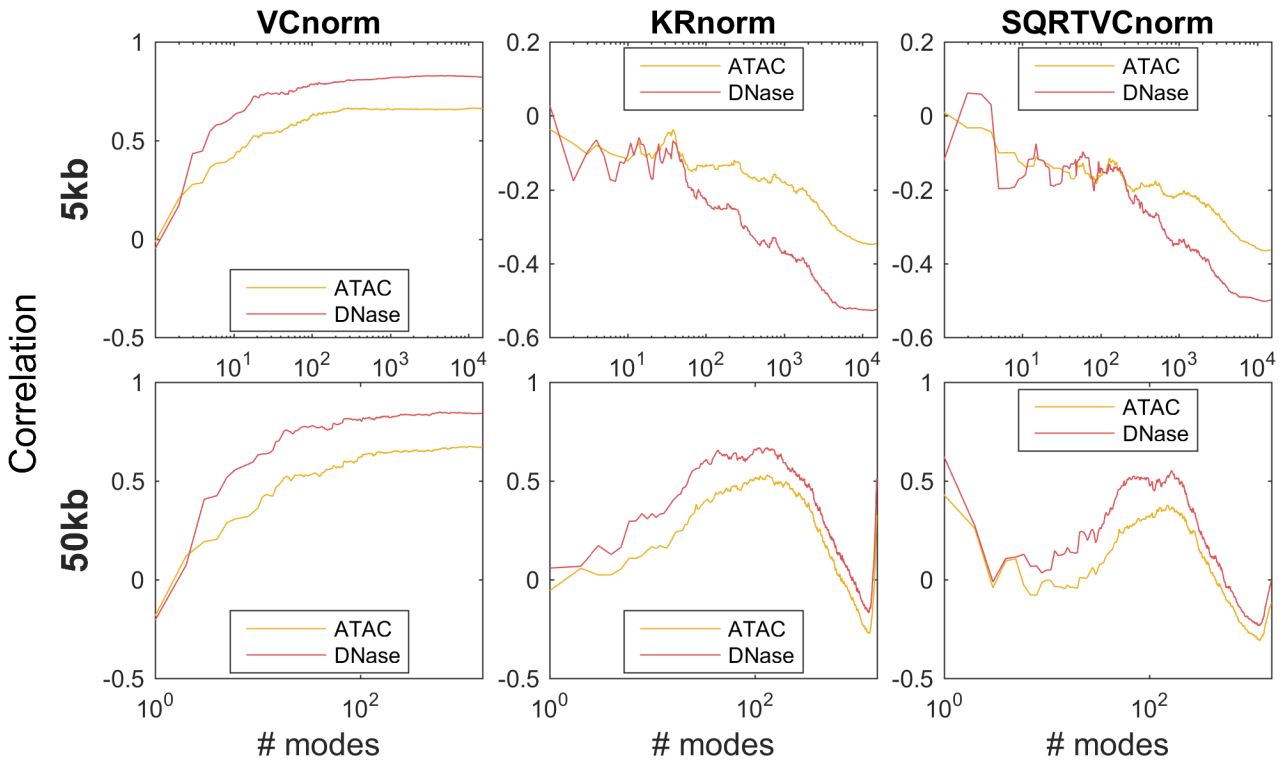


Supplementary Figure S3. GNM computations of mobility profiles using different subsets of modes show the robust convergence of results with a small subset of modes. Results are presented here for GM12878 chromosome 17, at 5kb resolution. **(A) – (C)** Comparisons between experimental data and computed MSF profiles obtained using 10, 100, and 500 GNM modes. **(D)** Spearman correlations between experimental and computationally predicted fluctuation/accessibility profiles obtained with different numbers of modes. **(E)** Spearman correlations between MSFs computed from slowest *i* modes and *i*+1 modes. Note that the abscissa is in logarithmic scale in panels D and E. The correlation levels off at around a few hundreds of modes, showing that the addition of higher modes does not practically change the predicted MSF profile, and a small subset of < 500 modes can be efficiently used for evaluating the MSFs.

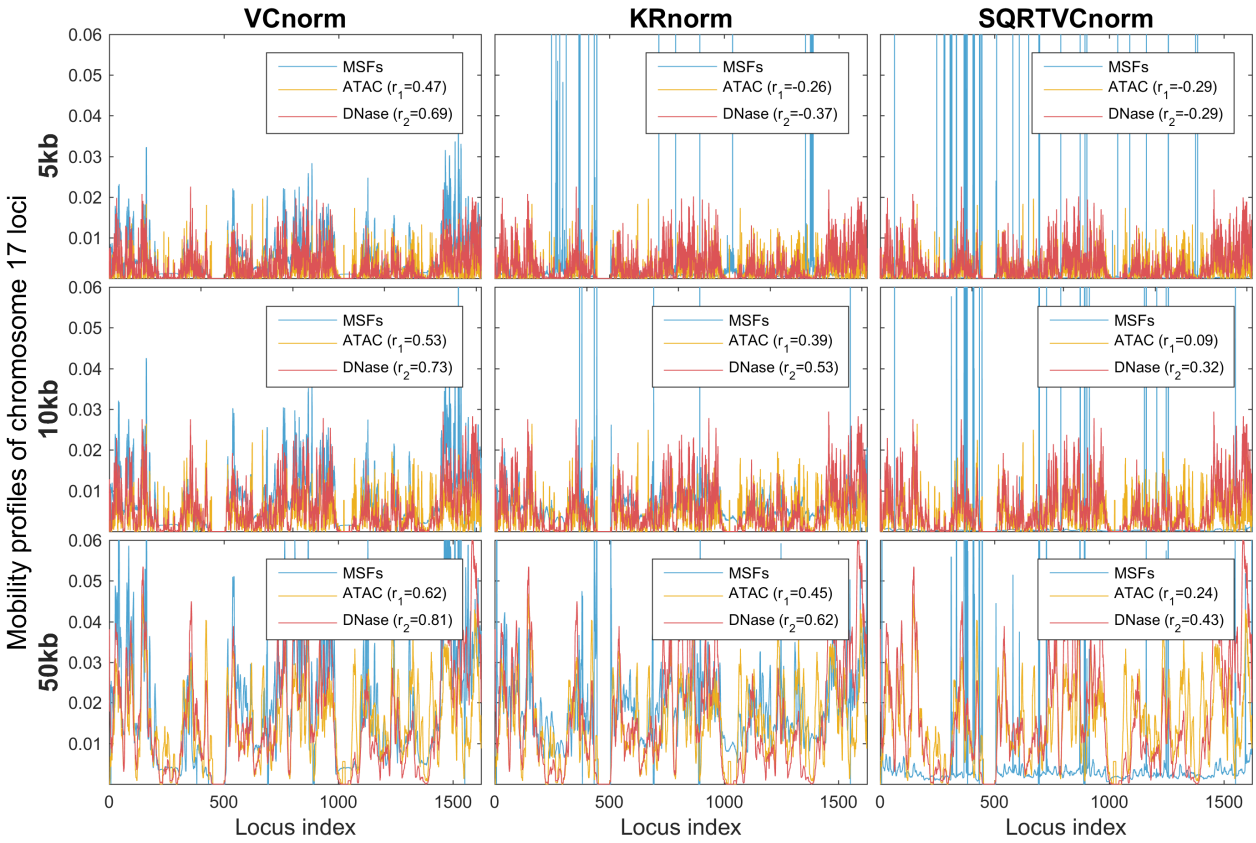


Supplementary Figure S4. Mobility profile of GM12878 chromosome 17 predicted by the GNM based on Hi-C maps at different resolutions. The three panels display the correlations between chromatin accessibility data (ATAC and DNase-seq) and GNM-predicted fluctuation profiles based on the Hi-C contact map 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.

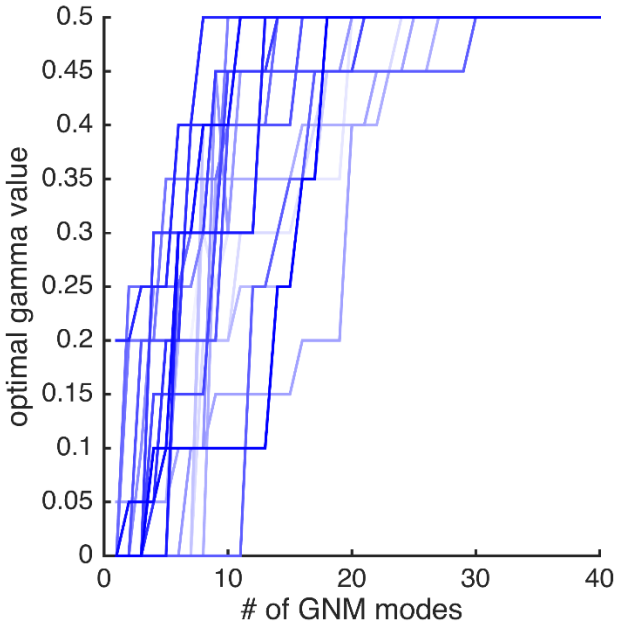
Supplementary Figure S5. Comparison of the MSFs obtained from different number of GNM modes (*rows*), and three different normalization methods (*columns*): Vanilla Coverage normalization (*left*), Knight-Ruiz normalization (*middle*), and square root Vanilla Coverage normalization (*right*). MSFs in this figure are calculated from Hi-C data at 50kb resolution for GM1287 chromosome 17.



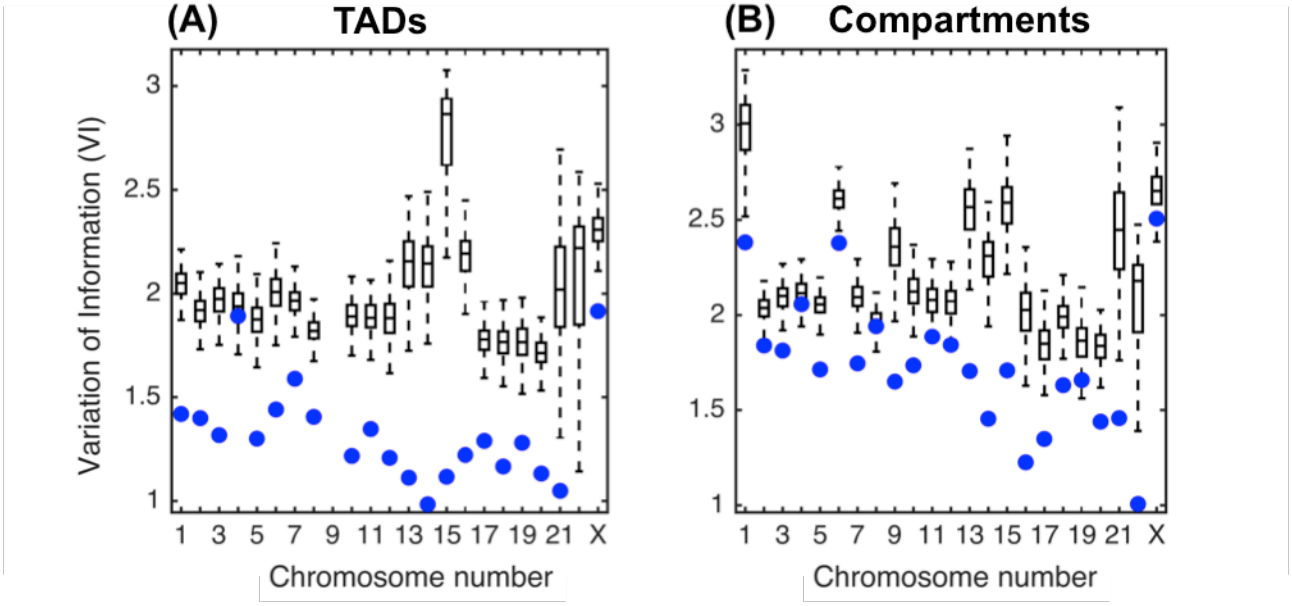
Supplementary Figure S6. The scanning of correlations between chromatin accessibility data from experiments and square fluctuations from theory calculated as a function of the number of modes included in the GNM analysis. The rows compare the correlations at different resolutions, and the columns compare those computed from three different normalization methods. Note the poor performance of KRnorm and SQRTVCnorm.



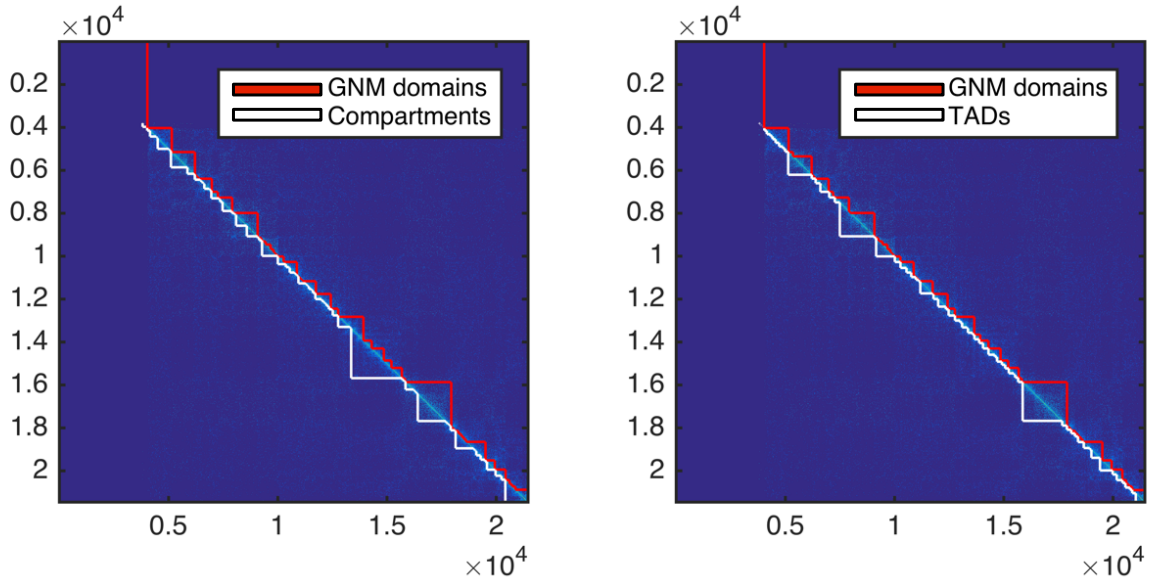
Supplementary Figure S7. Comparison of the MSFs obtained at different resolutions (*rows*) with accessibility data, using three different normalization methods (*columns*): Vanilla Coverage normalization (*left*), Knight-Ruiz normalization (*middle*), and square root Vanilla Coverage normalization (*right*). MSFs in this figure are calculated using m = 100 slowest modes.



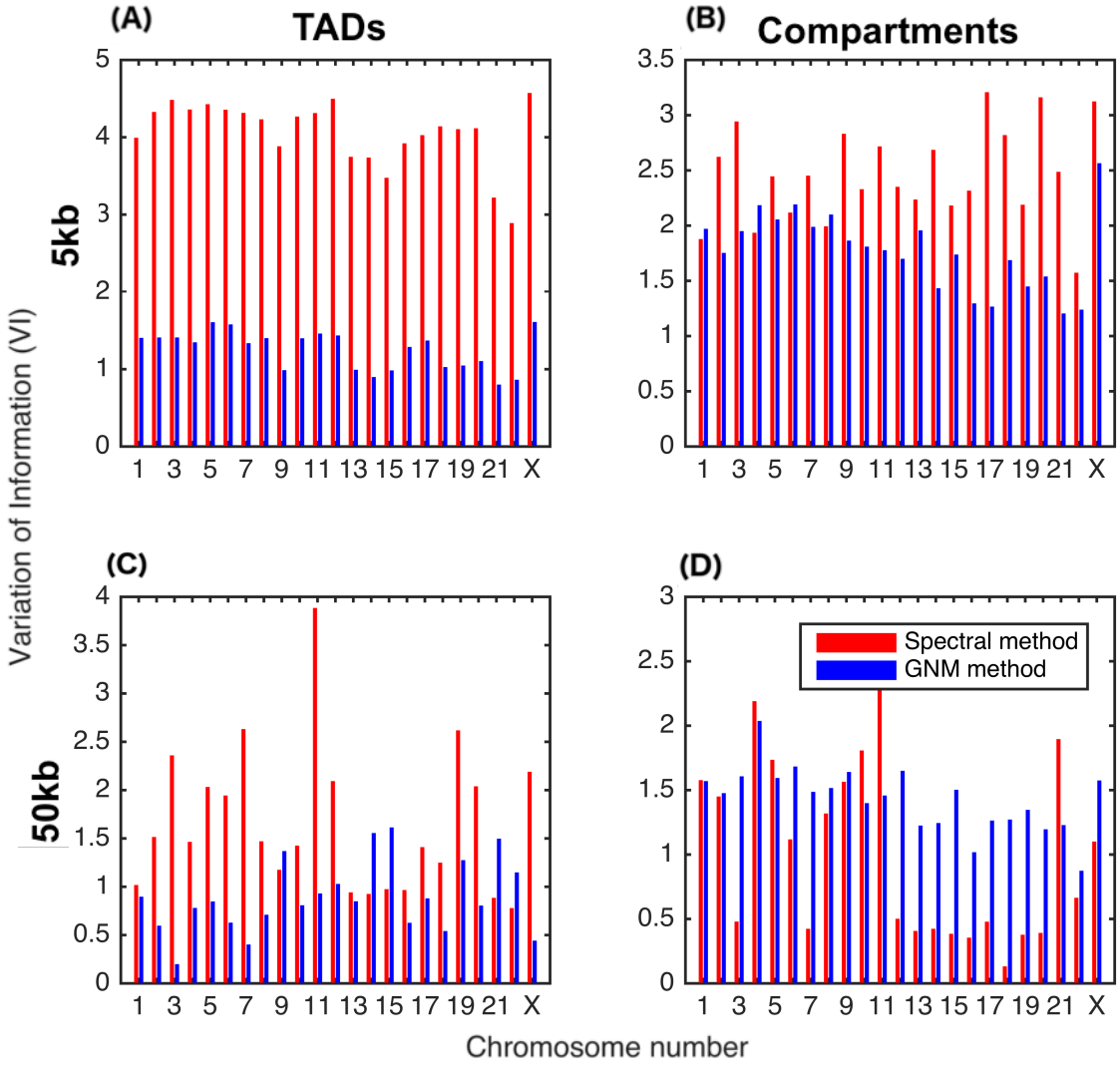
Supplementary Figure S8. Armatus gamma values as a function of GNM modes in GM12878 cells. The gamma values corresponding to the lowest VI value for each GNM mode increase monotonically with the number of modes used, showing that higher resolution domains can be found by using higher GNM modes, consistent with decreasing granularity of GNM-predicted substructures with increasing mode number.



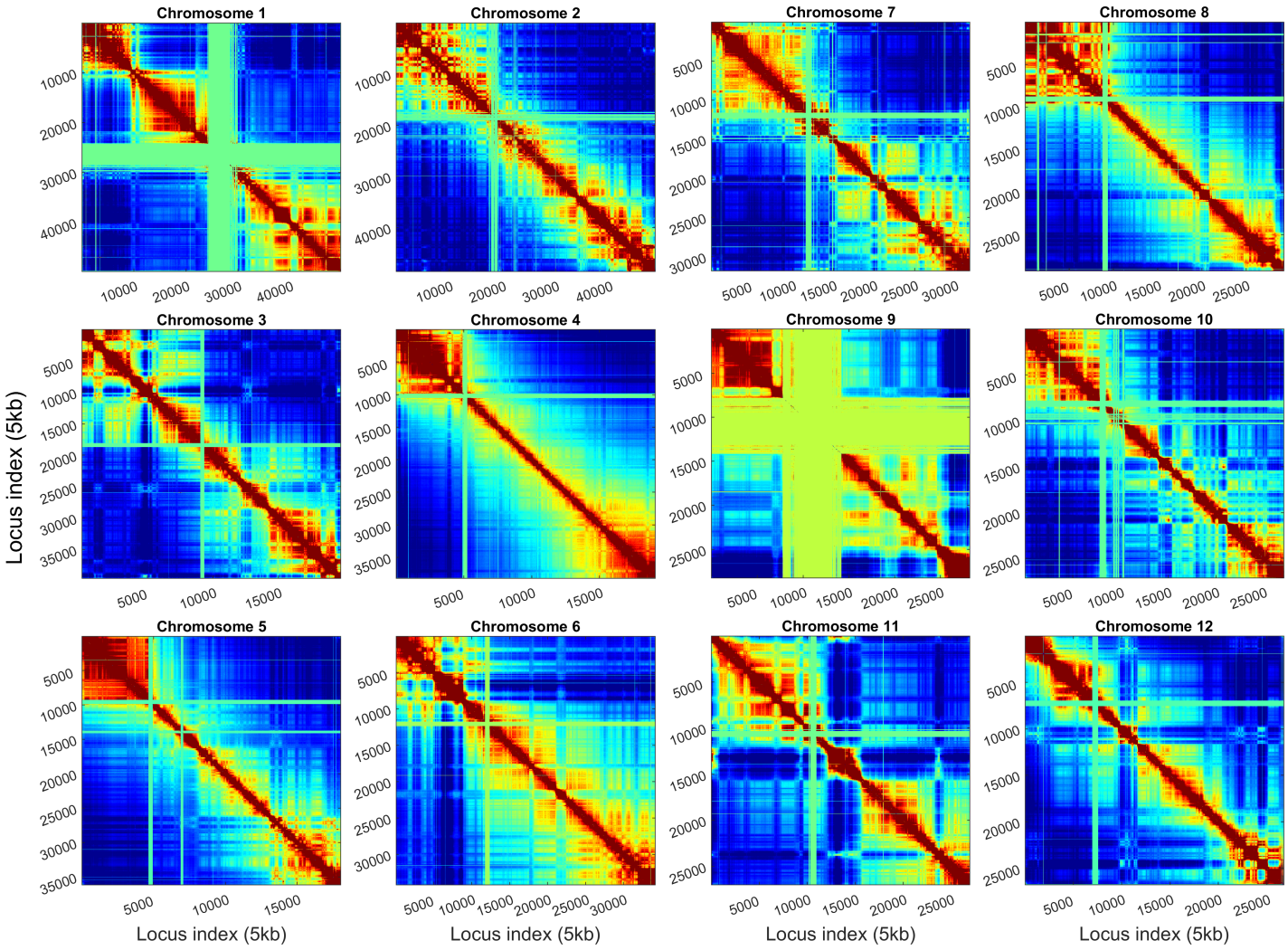
Supplementary Figure S9. Comparison of GNM domains to TADs and compartments for IMR90 cells. Variation of information (VI) measures for comparing GNM domains with (A) TADs and (B) compartments (lower VI indicates greater agreement). *Box plots* show the distribution of VI values obtained by randomly shuffling GNM domains and comparing to original TAD and compartment boundaries. *Blue dots* represent the VI value of the true GNM domains with TADs and compartments, respectively. Data was incomplete for computing TADs on chromosome 9, so this was left out. All comparisons were statistically significant except one chromosome with TADs, and three with compartments.



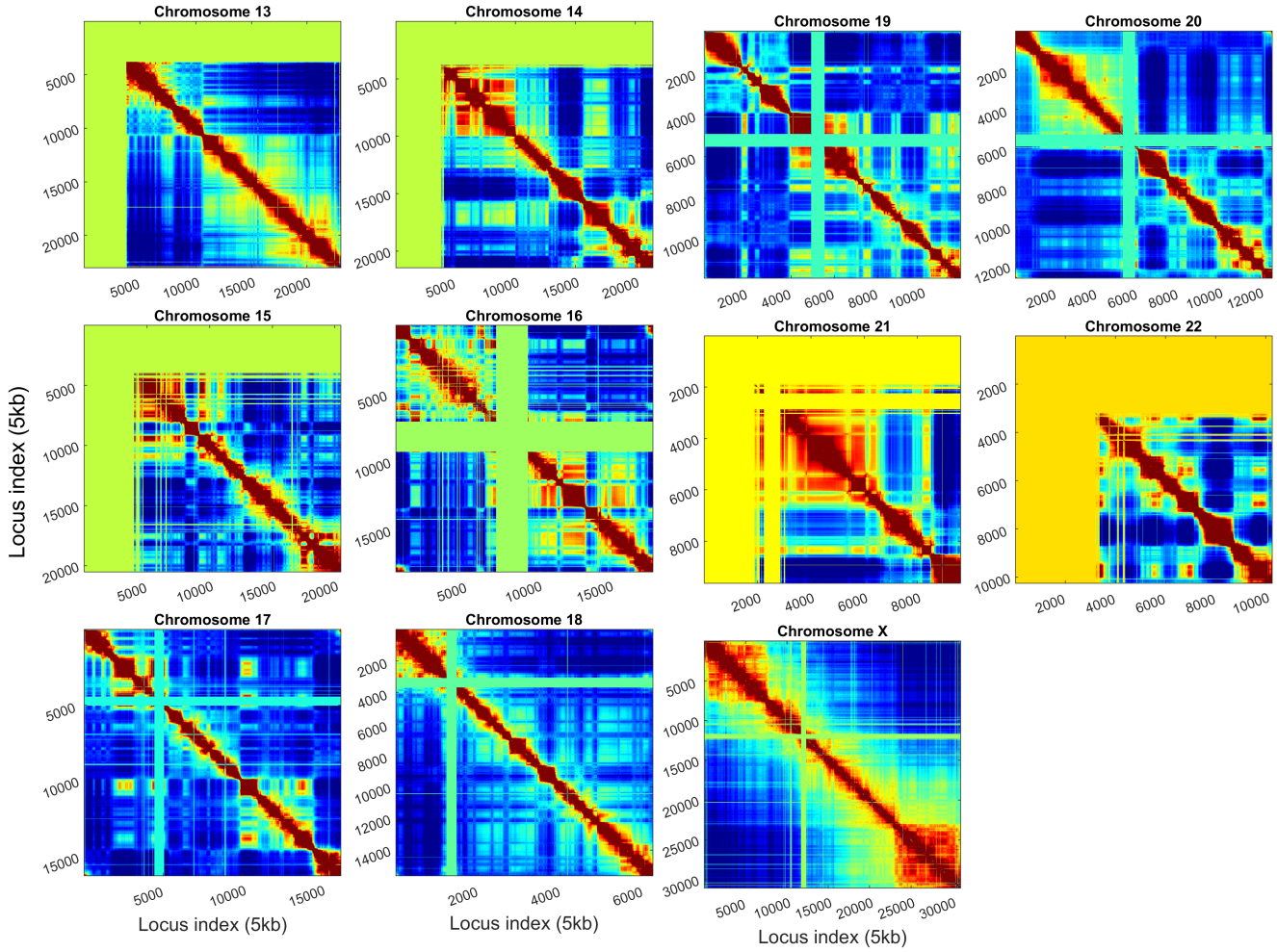
Supplementary Figure S10. Comparison of GNM domains with (A) compartments and (B) TADs for chromosome 14. In both panels, the background is a heat map of the Hi-C contact matrix for this chromosome, and the *red* and *white* lines represent the domains identified by the two indicated methods. The two axes represent the loci numbers. Data for compartments are from the work of Lieberman-Aiden et al. (4). TADs are computed using Armatus (10).



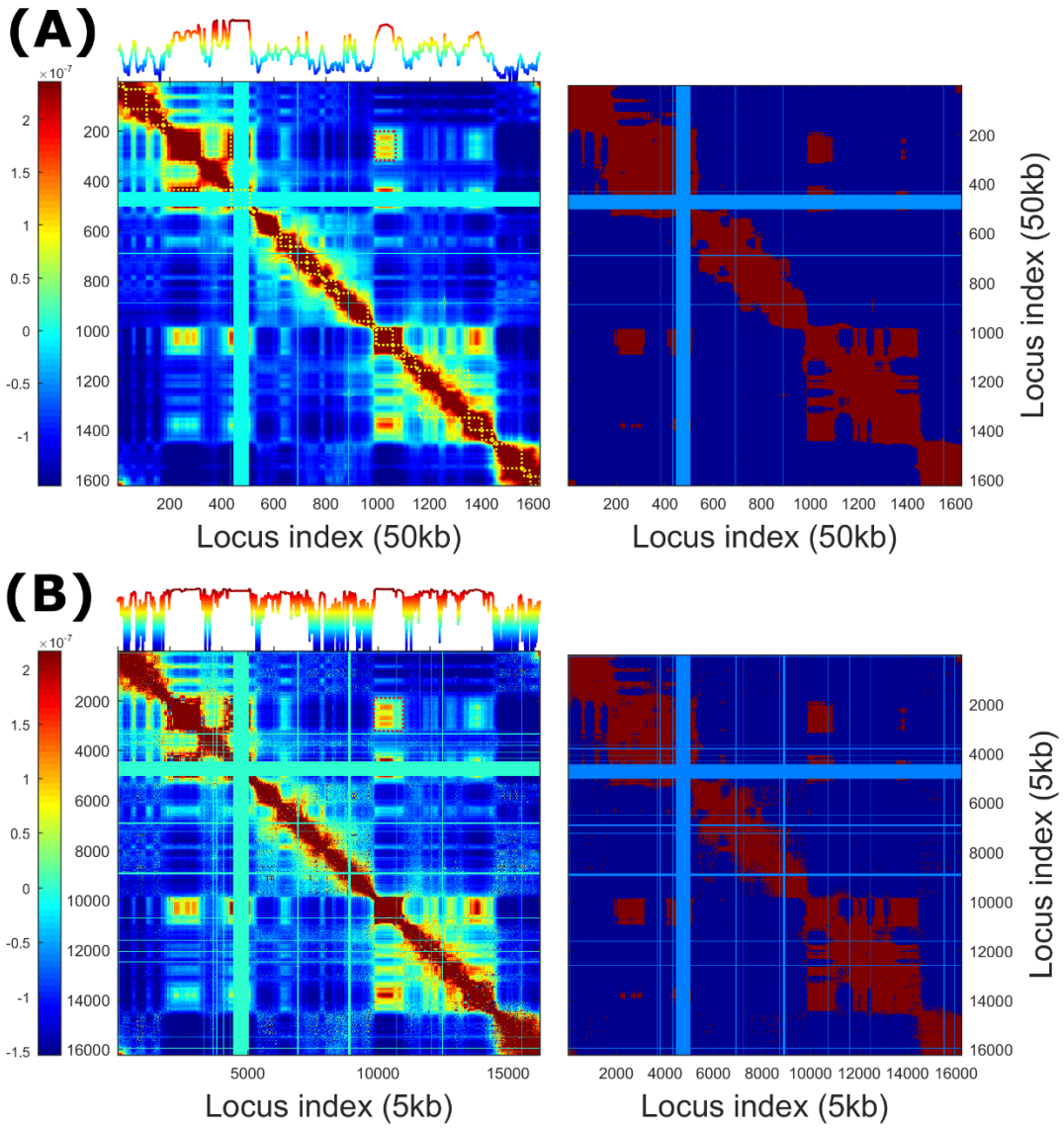
Supplementary Figure S11. Comparison of GNM-based method for finding TADs and compartments to the spectral method from (11) for GM12878. In all panels, *blue bars* represent the results from the GNM method and *red bars*, those from the spectral method. A lower variation of information (VI) value demonstrates better agreement. Compartments were calculated based on the method described in (4), and TADs were computed using the widely-used Armatus software (10). Armatus requires a resolution parameter γ, so the VI value shown for every comparison with TADs represents the lowest VI from comparing to TAD sets obtained from γ ranging from 0 to either 0.5 (for 5kb resolution) or 1 (for 100kb resolution), with a step size of 0.05. **(A)** Comparison to TADs at 5kb resolution. **(B)** Comparison to compartments at 5kb resolution. **(C)** Comparison to TADs at 100kb resolution. **(D)** Comparison to compartments at 100kb resolution.



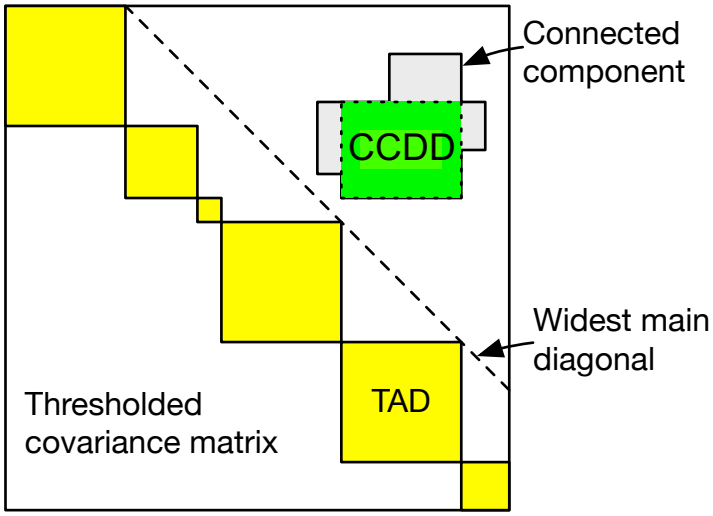
Supplementary Figure S12 (first part)



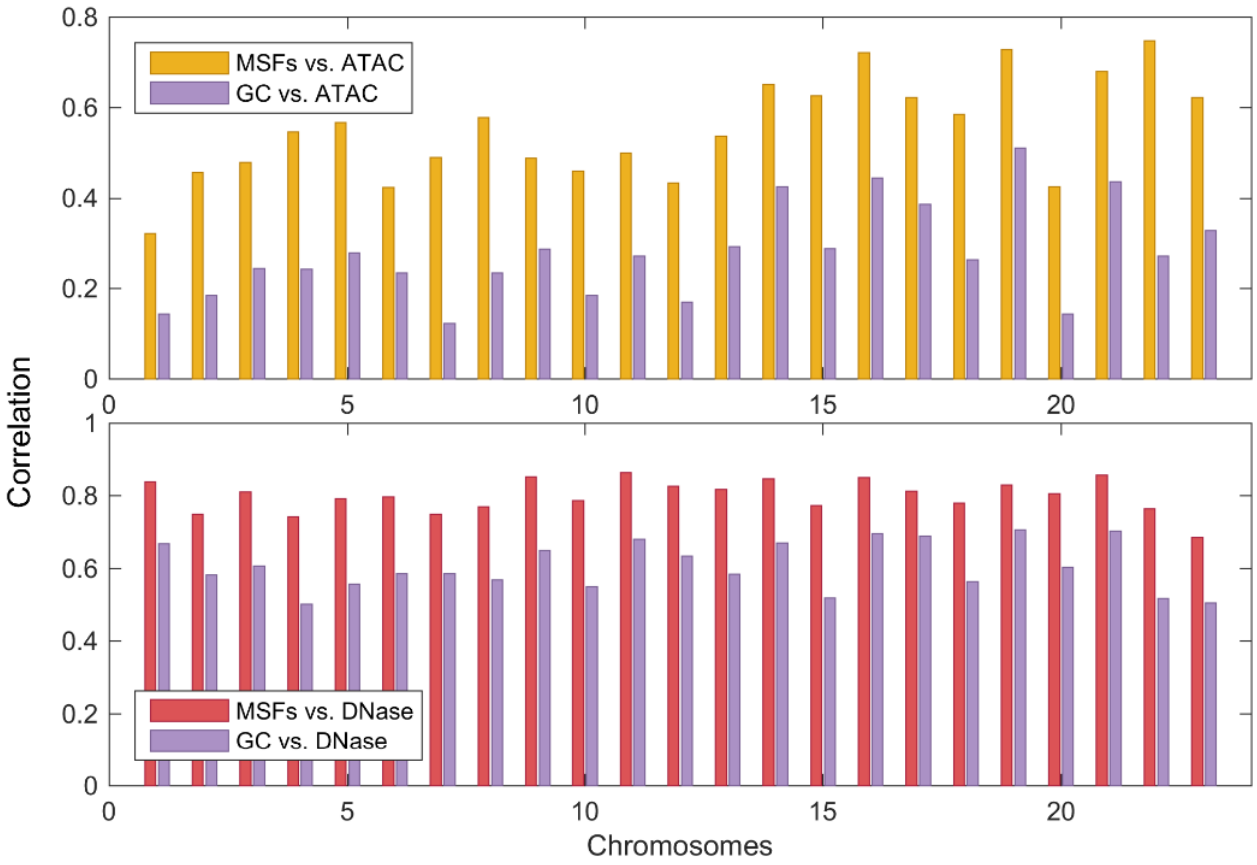
Supplementary Figure S12. Intra-chromosomal covariance of gene loci computed for all chromosomes at 5kb resolution in GM12878 cells. The entries in the map display the type and strength of correlations between the gene loci indicated along the two axes. The maps are color-coded from *dark red* the *dark 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.



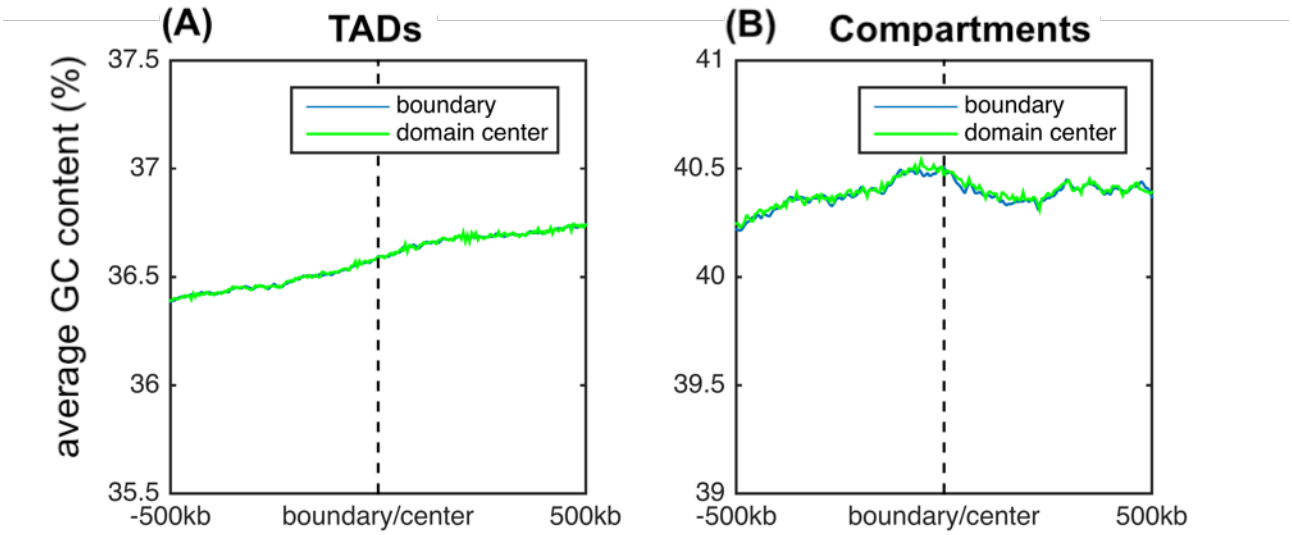
Supplementary Figure S13. Reproducibility of the covariance map computed for chromosome 17 using two different levels of resolution in GM12878. **(A)** Results at 50kb resolution computed using all GNM modes, **(B)** Results at 5kb resolution obtained with 500 slowest modes. The maps on the *right* of 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 CCDDs. 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.



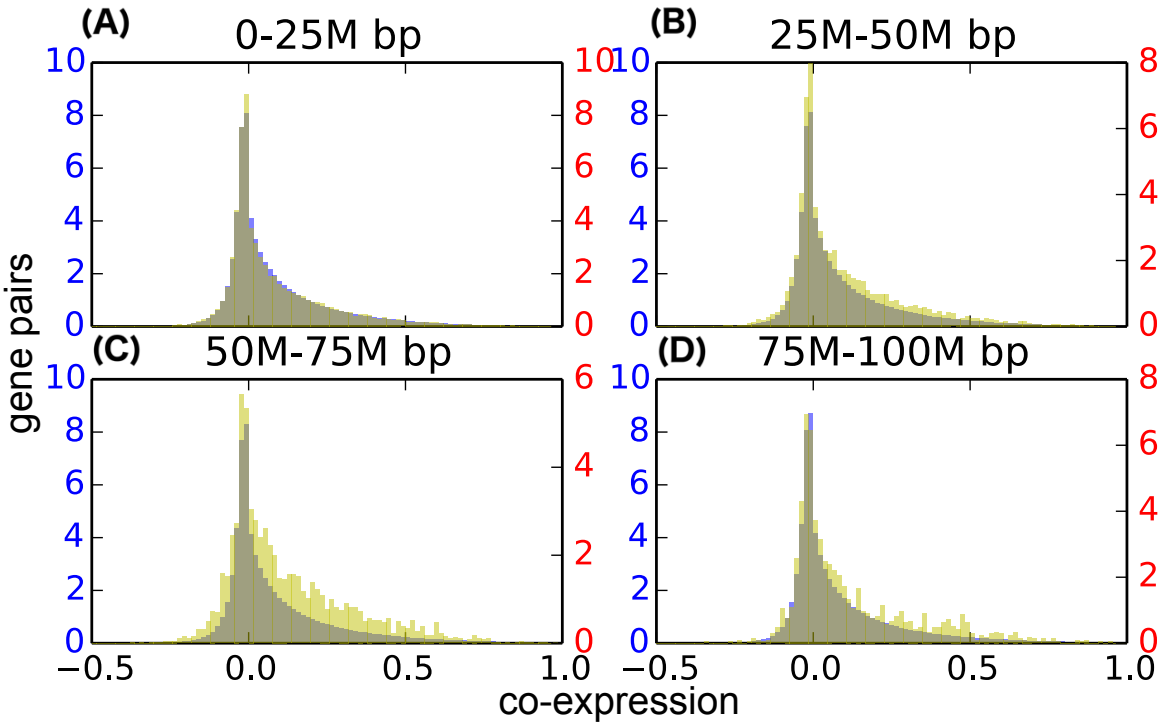
Supplementary Figure S14. Identification of cross-correlated distal domains (CCDDs). CCDDs are found by searching for connected components outside of the widest point of the main diagonal. The CCDD is then the rectangle of maximal area contained entirely within the connected component.



Supplementary Figure S15. Comparison of the level of agreement with experiments obtained by GNM (MSFs) and GC content profiles (GC), shown for 23 human chromosomes (GM12878 cells). Results are presented for two sets of accessibility data, ATAC-seq (*top*) and DNase-seq (*bottom*). GNM results are based on 500 slowest modes obtained from Hi-C data at 5kb. GNM consistently yields better agreement with experimental data.



Supplementary Figure S16. GC content as a function of position with respect to domain boundary/center does not exhibit any net (or stepwise) change at domain boundaries or centers. The same behavior is observed for TADs (A) and compartments (B). The green line represents the average GC content at and around boundaries of TADs **(A)** and compartments **(B)**. The blue line represents average GC content around the center of the structural domains. There is clearly no difference, so the GNM's ability to locate TADs and compartments cannot be attributed to the GC content covariate. (GM12878 cells)



Supplementary Figure S17. Co-expression enrichment of CCDDs is also present with bias-corrected RNA-seq. In each histogram, the yellow distribution represents gene pairs from CCDDs in GM12878 and the blue distribution represents background gene pairs. The only difference between this figure and Figure 5 from the main text is that RNA-seq values used here were calculated using Salmon's bias-correction. Results are qualitatively similar, except for potentially more co-expression here in the 25M-50M range. All are showing the normalized number of gene pairs with a particular Pearson expression correlation for gene pairs within a distance of (A) 0-25 million base pairs, (B) 25-50 million base pairs, (C) 50-75 million base pairs, and (D) 75-100 million base pairs. The more distant pairs (50-100 million base pairs apart) within the CCDDs show enriched expression correlations as compared to the background pairs. There were not enough gene pairs within CCDDs more than 100M base pairs apart to draw significant conclusions.

Supplementary 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

References

1. Bahar, I., Atilgan, A.R. and Erman, B. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, **2**, 173-181.
2. Bahar, I., Lezon, T.R., Yang, L.W. and Eyal, E. (2010) Global Dynamics of Proteins: Bridging Between Structure and Function. *Annu. Rev. Biophys.*, **39**, 23-42.
3. Haliloglu, T., Bahar, I. and Erman, B. (1997) Gaussian dynamics of folded proteins. *Phys. Rev. Lett.*, **79**, 3090-3093.
4. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. *et al.* (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, **326**, 289-293.
5. Knight, P.A. and Ruiz, D. (2013) A fast algorithm for matrix balancing. *Ima J Numer Anal*, **33**, 1029-1047.
6. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665-1680.
7. Meila, M. (2003) Comparing clusterings by the variation of information. *Lect. Notes. Artif. Int.*, **2777**, 173-187.
8. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213-+.
9. Song, L. and Crawford, G.E. (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, **2010**, pdb. prot5384.
10. Filippova, D., Patro, R., Duggal, G. and Kingsford, C. (2014) Identification of alternative topological domains in chromatin. *Algorithm Mol. Biol.*, **9**.
11. Chen, J., Hero, A.O. and Rajapakse, I. (2016) Spectral identification of topological domains. *Bioinformatics*, **32**, 2151-2158.