Supplemental Information (SI) for
Characterizing the 3D structure and dynamics of chromosomes and proteins in a common contact matrix framework
by Richard J. Lindsay, Bill Pham, Tongye Shen, and Rachel Patton McCord

I. Supplemental information for TAD imaging data analysis

First, we state the reason for using an individual cell's TAD Cartesian coordinates to display all the 3D chromatin TAD positions. Instead of an average structure, cell no. 1, numbered from 120 myofibroblast cells (Wang, S. et al. 2016), was selected for this purpose. As stated, only 47 of these 120 cells have complete Cartesian coordinates. Ideally, we should use an average of these 47 cells to closely represent the ensemble. However, one can obtain an average of meaningful Cartesian coordinates only if we can properly remove both overall translational and rotational degree of freedoms (DOFs). However, we could not properly separate internal DOFs with overall rotations using Cartesian PCA as chromosome 21 is extremely flexible. For instance, Chr21 of cell sample no. 105 is much larger in size compared to that of cell no. 95. Another evidence is that, in Cartesian PCA (with the overall translational motions already filtered), the top 3 eigenvalues are not much larger than the following eigenvalues. Thus, there is not a very good separation of internal DOFs from global motions.

We also displayed the I-PCA results of TAD imaging data analysis here: the top three I-PCs for Chr21 (Figure S1). The top two PCs for I-PCA from TAD data actually reveal certain differences in fluctuation compared to that of MI-PCA. The top eigenvector clearly displays an overall "breathing" motion in which corresponding contacts are being made and broken in sync. Only when one inspects I-PC3 can one see the connection with the dominant eigenvector of MI-PCA. This large difference between I-PCA and MI-PCA reflects the unique properties of analyzing distances between TADs measured by microscopy. In contact matrix analysis of Hi-C data or well-folded proteins, I-PCA and MI-PCA give fairly consistent results at the top PCs.

We also studied the robustness of E-PCA results due to the small sample size. Since we only have complete data for 47 cells, the noise level was initially a concern. Here, the level of convergence was checked by performing the calculations on an arbitrary split of the 47 cells into two groups. The first group contains the first 23 cells, while the second group contains the remaining 24 cells. A direct comparison between the corresponding elements of the top eigenvectors is shown here (Figure S2). The dot product of these normalized top eigenvectors is 0.834, demonstrating the robustness of the E-PCA analysis of this data. Additional corroborative evidence is shown from the consistency between this result (using TAD imagining data) and the result using Hi-C data (shown below in Section II of supplemental information).

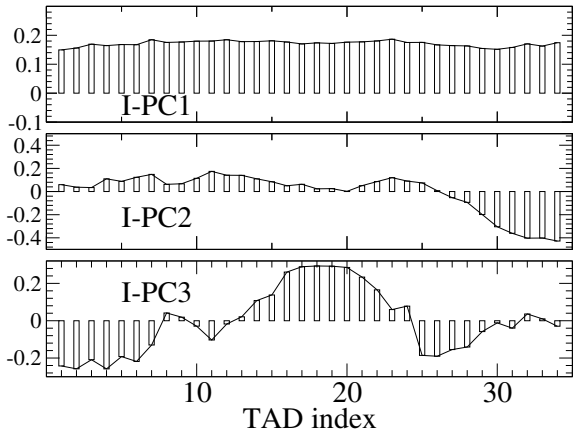

Fig. S 1: The top three eigenvectors of I-PCA from TAD imaging data of Chr21.

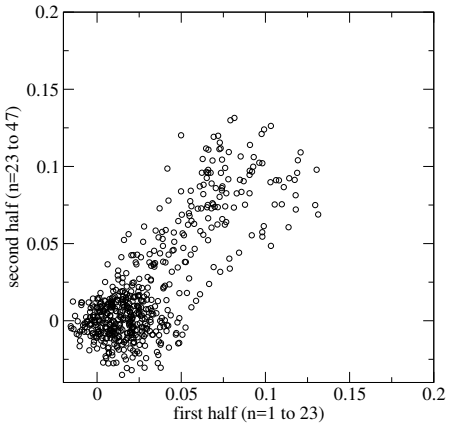

Fig. S 2: A direct comparison of E-PC1 components from two halves of TAD imaging data for Chr21.

II. Supplemental information for Hi-C data analysis

For the E-PCA analysis in Figure 4, with permission of the European Genome-phenome Archive, we first downloaded and raw sequencing reads from Hi-C experiments on 8 blood cell types published in Javierre et al., 2016. To this set, we added Hi-C data on GM12878 cells from experiments performed in our own lab. We mapped, filtered, binned, and iteratively corrected the Hi-C data as previously described (Imakaev et al., 2012). To explore structural variations at the length scale of compartment organization, we used 250 kb binned contact matrices for downstream MI-PCA, I-PCA, and E-PCA analyses. A 17 Mb section of Chr10 was chosen as an example region meeting the following criteria: 1) it is a long stretch of the genome with no highly repetitive bins that needed to be excluded from the analysis, 2) it contains numerous clearly defined A/B compartment switches, 3) it is from a chromosome that is of intermediate length and gene density. To remove the dominant effects of the strong Hi-C diagonal, variables were scaled to unit variance during E-PCA analyses. To match standard MI-PCA analyses of Hi-C data, these analyses were performed on contact matrices that had been normalized for the expected number of random contacts at each genomic distance, as described previously (Crane et al., 2015) (Figure S3a). We note that the resulting first principal component, typically descriptive of the A/B compartmentalization of the chromosome structure, was very similar regardless of whether we used the mean contact matrix across the 9 cell types or included all contact matrices in the I-PCA analysis (Figure S3b). As with the E-PCA analyses on the raw contact matrices presented in Figure 4, E-PCA analysis on the genomic distance normalized matrices also produced principal components that segregate cells according to their place along the blood cell lineage (Figure S4a), with the matrices representing the first two PCs quite similar to those obtained from unnormalized contact matrices, just reversed in their importance (Figure S4b).

Due to the different resolution of the 250 kb binned contact data and the 34 TADs imaged by Wang et al., the results from E-PCA analysis of the 250 kb binned Chr21 Hi-C matrices were not directly comparable to the TAD E-PCA analysis. However, when we summed the Hi-C contacts into the same genomic distance bins used for that imaging data, then normalized each bin to account for their varying sizes, we found that the E-PCA results for the 9 cell type Hi-C data were similar to those obtained from the imaging data (Figure S5). In particular, the first PC reflects a separation of two major domains as the main motion as in the TAD imaging data. This suggests that the motions of the chromosome within and between single cells of the same cell type may be related to the changes in chromosome folding between cell types.

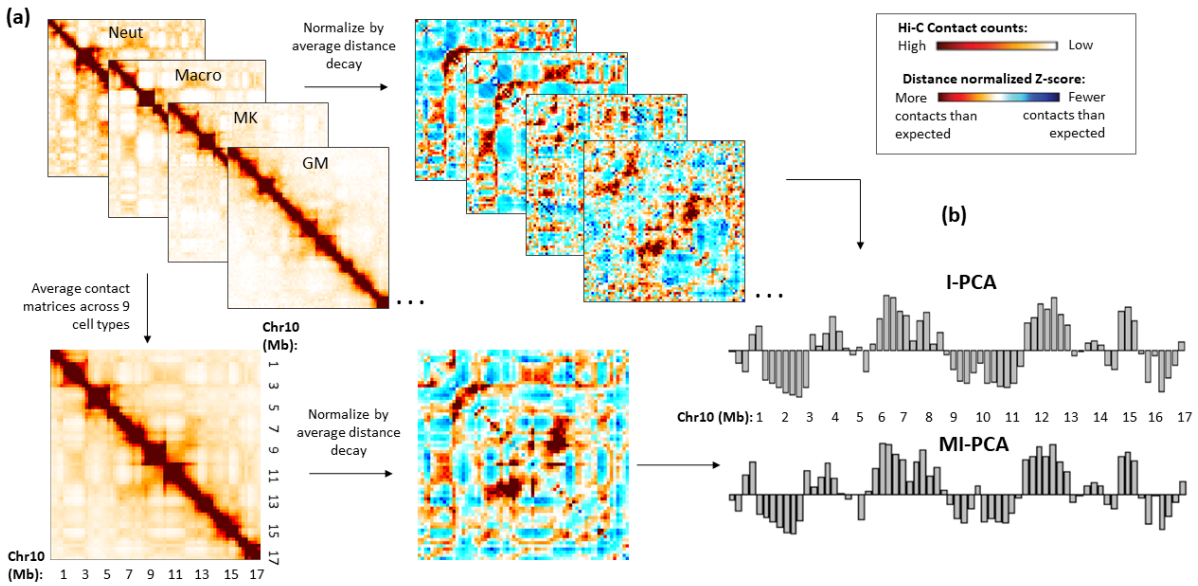

Fig. S 3: Comparing MI-PCA with I-PCA using Hi-C data of Chr10:1-17mb from 9 types of blood cells. (a) Hi-C data are pre-processed starting from individual cell type contact matrices (top left) by either directly calculating the contact distance decay normalized z-Score matrices for each cell type (top right) or by first averaging all contact count matrices (lower left) and then finding the distance normalized z-Score matrix (lower right). (b) The entire set of individual cell type zScore matrices are used for I-PCA (top) while the averaged distance normalized matrix is used for MI-PCA (bottom). The components of eigenvector PC-1 plotted along Chr10 for MI-PCA and I-PCA show a similar pattern.

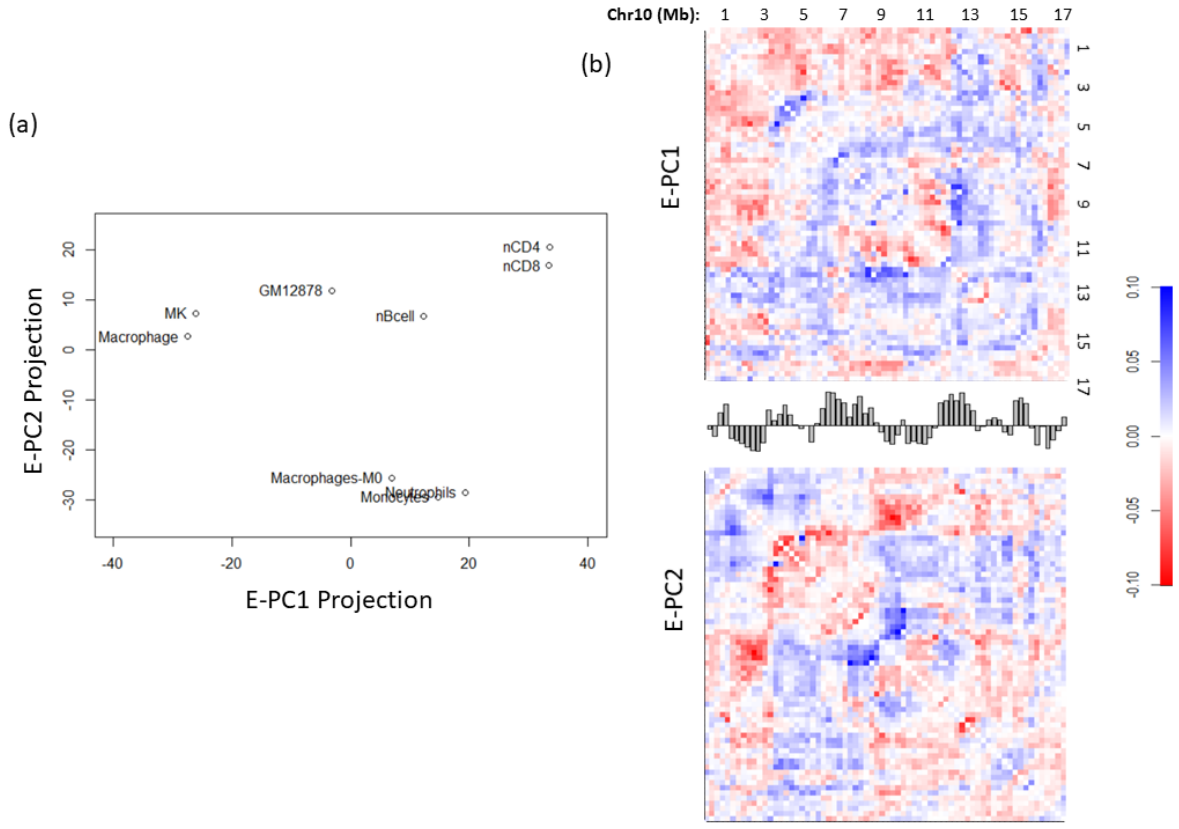

Fig. S 4: E-PCA on 9 blood cell types using distance normalized z-Score matrices (as shown in Fig S3a). (a) Plotting the projection of each cell type onto E-PC1 and E-PC2 from distance-normalized Hi-C data separates cell types according to cell lineage as was observed with non-distance normalized E-PCA results. (b) E-PCA displacement matrices are similar to those calculated for non-distance normalized E-PCA results (Fig 4c), except that E-PC1 here is more similar to E-PC2 for the non-normalized data and vice-versa.

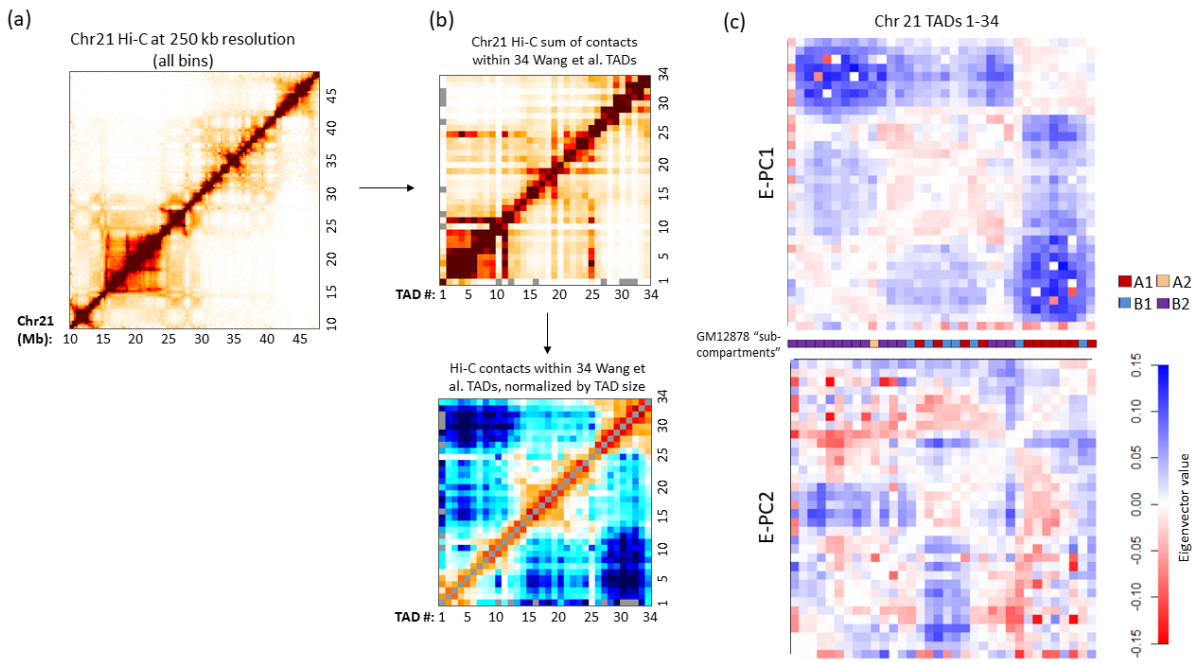

Fig. S 5: E-PCA of Chr21 Hi-C data across 9 blood cell types, aggregated for comparison with TAD imaging data. (a) Example Hi-C contact matrix from one of the 9 cell types, binned at 250 kb as for previous Chr10 analysis. (b) Hi-C contacts are summed into bins that match the TAD coordinates used for imaging in IMR90 cells (top) and then normalized to correct for varying genomic size of each TAD. Color scales as in Fig S3. (c) Normalized data for each of 9 blood cell types was used as input to E-PCA, and the displacement matrices of the first two PCs are shown here (comparable results to TAD imaging E-PCA on this same region). Sub-compartment classifications (Rao et al.) for GM12878 are shown for reference.

III. Supplemental information for protein simulation data analysis

Presented here is a direct comparison of the elements of the top eigenvectors (PC1 and PC2) obtained from several methods of I-PCA. The top eigenvectors of implicit contact analyses (I-PCA, MI-PCA, and one-frame version of MI-PCA/I-PCA) turned out to be nearly identical for a highly structured biopolymer, such as the folded protein signaling complex in this case (Figure S6). Only when we inspected the second dominant mode, PC2, specifically for the RXR:CAR complex, we can see that part of the complex (CAR) shows noticeable differences between I-PCA and MI-PCA (Figure S7). The highly consistent results between different MI-PCA and I-PCA (as well as between results of RXR:TR and RXR:CAR) indicate that the I-PCA method locates the consensus features of the ensemble. This forms a sharp contrast with the calculations of E-PCA which had revealed drastic differences between RXR:TR and RXR:CAR.

In addition to protein conformations obtained from all-atom simulation, we also present here the I-PCA results obtained from an NMR ensemble, as shown in Figure S8. Total 50 models of hen lysozyme (PDB ID: 1E8L) were used to construct the contact matrices. The top two I-PCs are compared with the corresponding 1-frame I-PCA (the first model). Again, we see from Figure S8 that I-PCA results are similar to instant frame I-PCA.

Using data from RXR:TR (N=493), we demonstrate the I-PCA projection onto the top two PCs here (Figure S9 a). Total 200 ns of simulation data with a snapshot frequency of every 50 ps, we have T=4,000. Thus, total 493 x 4,000 points are shown (Figure S9 a). The snapshot average version of the result (total 4,000 points) is shown in Figure S8 b. Using mean contact matrix obtained from RXR:CAR (N=476) as a test system, we demonstrate MI-PCA projection here. The results of the 476 rows of the mean matrix are projected onto the top two PCs, shown in a conventional plot (PC1 vs PC2, Figure S9 c) and as a function of the row index (Figure S9 d). It is interesting to point out that the projection is quite aligned with eigenvector itself, as demonstrated using the ratio of PC1's projection over the elements of the eigenvector PC1 (Figure S9 e). This may be related to the nature of MI-PCA.

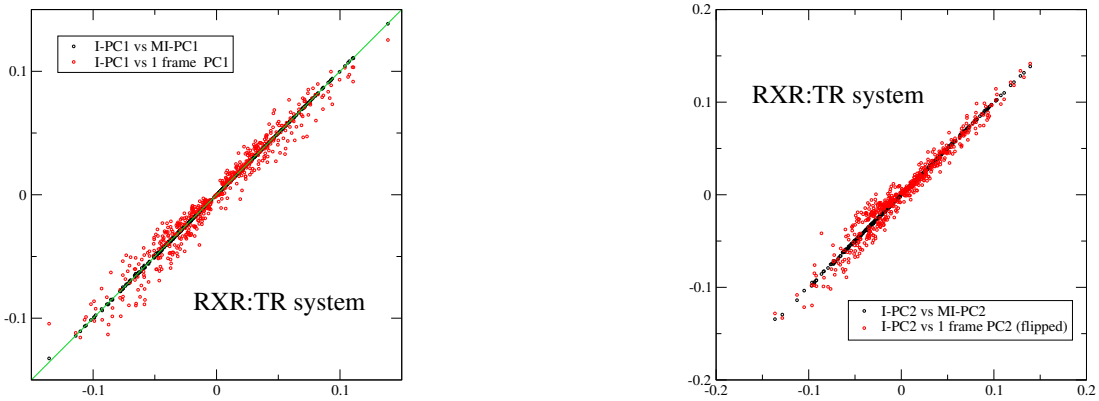

Fig. S 6: The comparison of top eigenvectors using I-PCA, MI-PCA, and 1-frame PCA (MI-PCA style with a single frame) for the RXR:TR system.

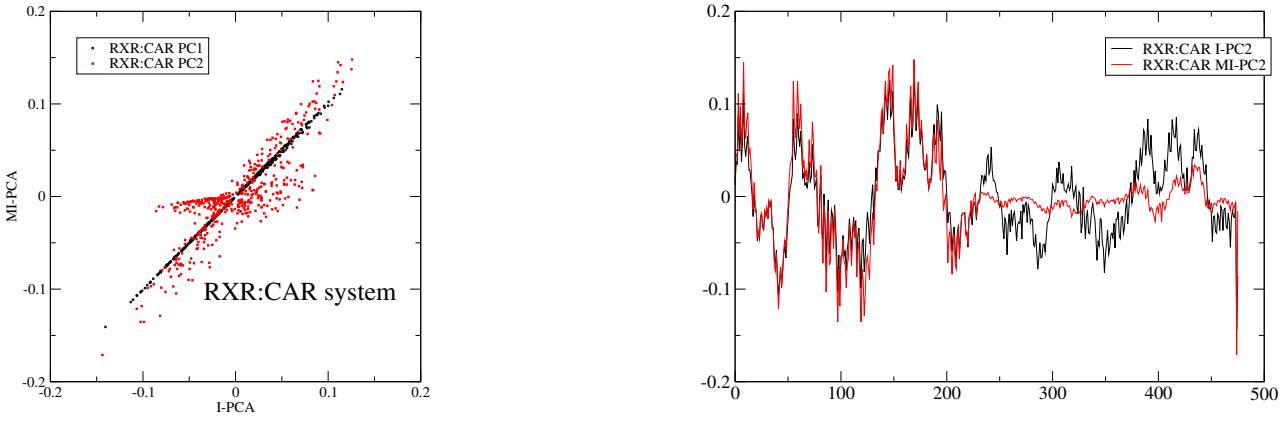

Fig. S 7: The comparison of top eigenvectors using I-PCA and MI-PCA for the RXR:CAR system.

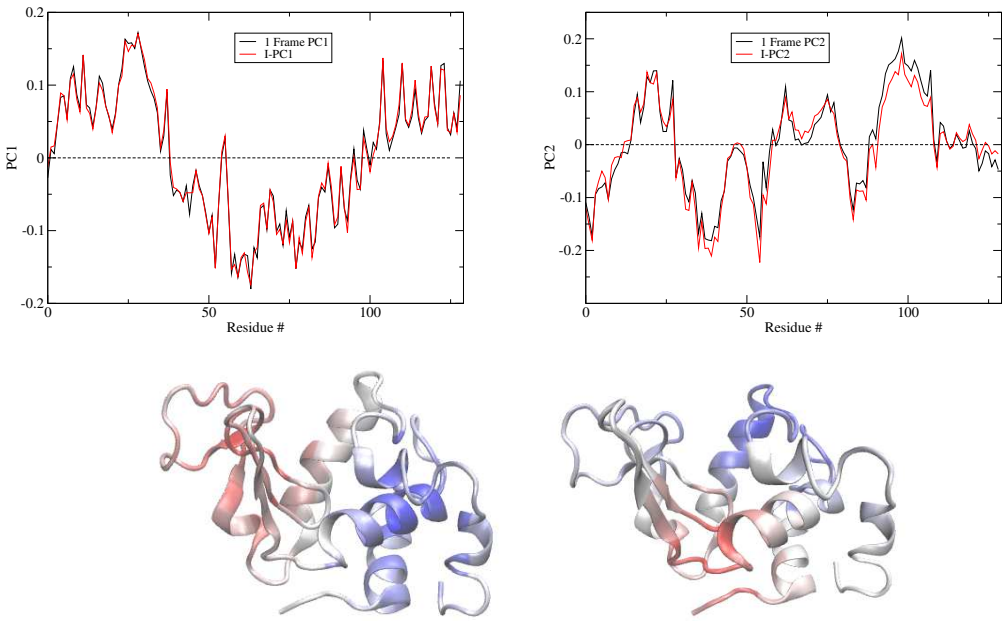

Fig. S 8: The I-PC1 and 1-frame PC1 of an NMR structure ensemble for hen lysozyme are shown in (left panel). The corresponding PC2s are shown in (right panel), with 1-frame PC2 flipped for a better comparison. The corresponding 3D rendering of I-PCs are shown with color-coded values (red to white to blue for negative to zero to positive).

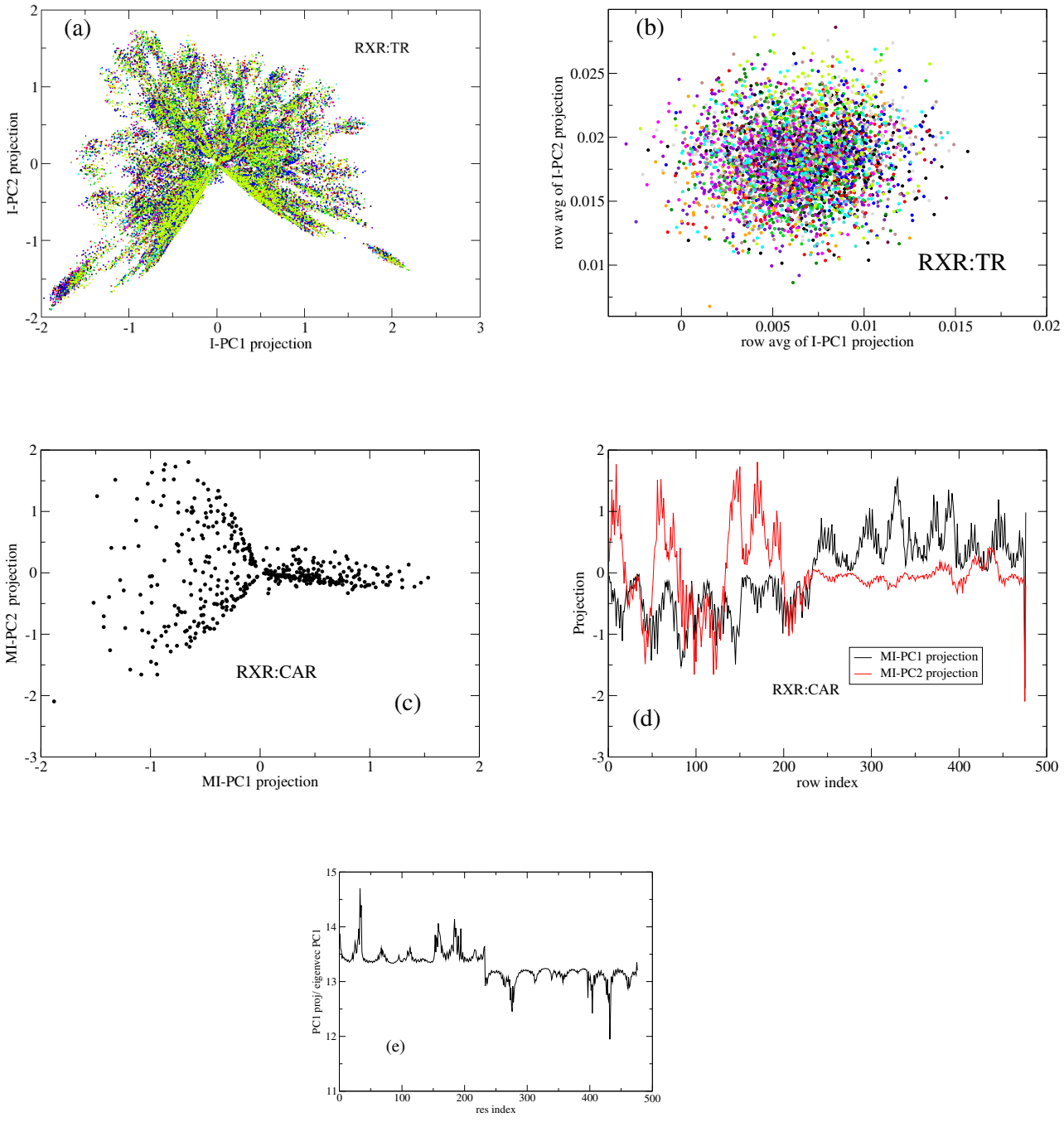

Fig. S 9: The I-PC projection for the RXR:TR system is shown in (a) and (b), with colors indicating time blocks. The PC projection of the mean contact matrix for the RXR:CAR system is shown in (c-e).

References:

Crane E, et al. (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 523:240.

Imakaev M, et al. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Meth 9(10):999-1003.

Javierre BM, et al. (2016) Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 167(5):1369-1384.e1319.

Wang S, et al. (2016) Spatial organization of chromatin domains and compartments in single chromosomes. Science 353:598-602.