Joint inference and alignment of genome structures enables characterization of compartment-independent reorganization across cell types

Lila Rieber[1], Shaun Mahony[1,*]

[1]Department of Biochemistry and Molecular Biology and Center for Eukaryotic Gene Regulation, The Pennsylvania State University, University Park, PA 16802, USA

[*]To whom correspondence should be addressed: mahony@psu.edu

Abstract

Cell-type-specific chromosome conformation is correlated with differential gene regulation. Broad compartmentalization into two compartments (A & B) is proposed to be the main driver of cell-specific chromosome organization. However it is unclear what fraction of chromosome conformation changes between cell types and conditions is independent of changes in compartmentalization and whether any such compartment-independent reorganization is functionally important. We developed MultiMDS to jointly infer and align 3D chromosomal structures, thereby enabling a quantitative comparison of locus-specific changes across Hi-C datasets. We compared Hi-C datasets from yeast, which lack compartmentalization, grown with and without galactose. These comparisons confirmed known relocalizations as well as identifying additional examples. We also compared mammalian datasets across a variety of cell lines. We found a consistent enrichment for changes along the A/B compartment (nuclear interior/nuclear periphery) axis, even when comparing the same cell type from different individuals. Despite the prevalence of compartment changes, we consistently find compartment-independent relocalizations of loci that are within the A compartment in both compared cell types. Some such intra-compartment relocalizations involve loci that display enhancer-associated histone marks in one cell type and polycomb-associated histone marks in the other. MultiMDS thus enables a new way to compare chromosome conformations across two Hi-C datasets.

Availability: https://github.com/seqcode/multimds

Introduction

Chromosome conformation data have become available for diverse cell types, perturbations, and developmental stages. Comparing these datasets can highlight the relationship between three-dimensional structure and biological function. To date, such comparisons have primarily examined known chromosomal structures, such as A/B compartments, topologically associated domains (TADs), and loops. Though TADs are largely conserved across cell types and species (Dixon et al., 2012), extensive changes in compartmentalization (Dixon et al., 2015; Lin et al., 2012) and looping are detectable between datasets (Amano et al., 2009; Ferraiuolo et al., 2010; Rickman et al., 2012) and are correlated with differential gene expression. However, focusing on known types of changes means that other potentially functional differences remain

unexplored. For example, rearrangements within a compartment or TAD could be correlated with changes in gene regulation.

Current methods for comparing chromosomal conformation datasets (primarily from Hi-C experiments) can be classified as global, interaction-specific, or locus-specific. Some global methods calculate an overall similarity score for two Hi-C datasets, enabling clustering of experiments (Ursu et al., 2018; Yang et al., 2017). Global concordance in TAD boundaries can also be calculated (Zufferey et al., 2018). However, these methods cannot discover specific changes between datasets. Interaction-specific methods identify genomic locus pairs that significantly differ in their interaction frequency, which may indicate gain or loss of a chromatin loop (Heinz et al., 2010; Lun and Smyth, 2015; Paulsen et al., 2014; Stansfield et al., 2018), but cannot determine which locus in a pair is the driver of the change. Locus-specific methods identify changes in organization, such as a change in compartment score or gain or loss of a TAD boundary, that occur at a single locus. However, these methods are currently limited to providing 1-dimensional metrics based on changes in known forms of chromatin organization, preventing the discovery of novel types of structural change. There is currently no method for quantifying general locus-specific relocalizations between Hi-C datasets. Ideally, such a method would quantify the degree to which a given locus has changed position with respect to the rest of the genome, regardless of whether the relocalization was driven by changes in compartmentalization, TAD structure, looping, or a combination of several effects.

In a Hi-C dataset, each locus is represented as a vector of interaction frequencies with every other locus of the genome or chromosome. Because typical metrics for vector comparison, such as Pearson correlation, are biased by Hi-C distance decay, comparison of datasets is challenging (Yang et al., 2017). To reduce the dimensionality of the problem, the Hi-C dataset can be embedded in three dimensions, representing the population average chromosome structure. In principle, two 3D structures could be compared in order to assess the overall changes in chromatin organization across cell types or conditions. If the structures are comparable and correctly aligned, the degree to which a given locus has shifted position can be calculated as the Euclidean distance between the appropriate points on each structure. We have developed MultiMDS to simultaneously infer and align 3D structures from two Hi-C datasets. By applying our method to a number of mammalian and yeast Hi-C datasets, we identified examples of chromatin reorganization correlated with biological function, some of which confirm previous findings and some of which are potentially novel.

Methods

Datasets

We used published Hi-C datasets for yeast grown in glucose and galactose media (Kim et al., 2017), ENCODE cell lines (Rao et al., 2014), lymphoblastoid cell lines (LCLs) (Chaisson et al., 2018) made available through the 4D Nucleome Project (Dekker et al., 2017), wild-type and Brd2-knockout (KO) G1E cells (Hsu et al., 2017), HPC-7 cells (Wilson et al., 2016), wild-type and *Nipbl* conditional KO mouse hepatocytes (Schwarzer et al., 2017), and wild-type and

CTCF-depleted mouse embryonic stem cells (mESCs) (Nora et al., 2017). Autosomes were used for all analyses. K562 chr9 and chr22 were removed due to their translocations.

Yeast RNA-seq and Nup60 ChIP-seq data is from (Kim et al., 2017). *Saccharomyces uvarum* gene annotations are from (Scannell et al., 2011). Other ChIP-seq data is from ENCODE (ENCODE Project Consortium, 2012).

Algorithm

MultiMDS uses a novel joint multidimensional scaling (MDS) algorithm, which simultaneously embeds two datasets in a lower-dimensional space while minimizing the weighted sum of squared distances (SSD) between the embeddings. In addition to the weights for distance measurements used in standard MDS, which we refer to as distance weights, MultiMDS incorporates weights representing the expected similarity between the datasets, which we refer to as similarity weights. In our analysis, we used distance weights of 1 for all loci. We used equal similarity weights for all loci, equivalent to a single global parameter – referred to below as the penalty – which is selected empirically for each pairwise comparison (see below).

The algorithm minimizes stress, which is calculated as

$$\sigma(\boldsymbol{X_1}, \boldsymbol{X_2}) = \sum_{i<j} w_{ij}[(d_{ij}(\boldsymbol{X_1}) - \delta_{ij1})^2 + (d_{ij}(\boldsymbol{X_2}) - \delta_{ij2})^2] + \sum_i v_i d_i^2(\boldsymbol{X_1}, \boldsymbol{X_2})$$

where $\boldsymbol{X_a}$ is a configuration of points for dataset a, w_{ij} is the distance weight for points i and j, $d_{ij}(\boldsymbol{X_a})$ is the Euclidean distance between points i and j in $\boldsymbol{X_a}$, δ_{ija} is the dissimilarity between points i and j for dataset a, v_i is the similarity weight for point i, and $d_i(\boldsymbol{X_1}, \boldsymbol{X_2})$ is the Euclidean distance between point i in dataset 1 and point i in dataset 2.

We seek to find a compact expression for SSD, $\sum v_i d_i^2(\boldsymbol{X_1}, \boldsymbol{X_2})$, similar to (Borg and Groenen, 2005). Consider one weighted squared distance, $v_i d_i^2(\boldsymbol{X_1}, \boldsymbol{X_2})$. Let $\boldsymbol{x_1 a}$ be column a of the coordinate matrix $\boldsymbol{X_1}$. Let $\boldsymbol{e_i}$ be column i of the identity matrix $\boldsymbol{I}$.

$$v_i d_i^2(\boldsymbol{X_1}, \boldsymbol{X_2}) = v_i[\sum_{a=1}^{m} \boldsymbol{x_{1ia}}^2 - 2\sum_{a=1}^{m} \boldsymbol{x_{1ia}}\boldsymbol{x_{2ia}} + \sum_{a=1}^{m} \boldsymbol{x_{2ia}}^2]$$

$$= v_i[\sum_{a=1}^{m} \boldsymbol{x'_{1ia}}\boldsymbol{e_i}\boldsymbol{e'_i}\boldsymbol{x_{1ia}} - 2\sum_{a=1}^{m} \boldsymbol{x'_{1ia}}\boldsymbol{e_i}\boldsymbol{e'_i}\boldsymbol{x_{2ia}} + \sum_{a=1}^{m} \boldsymbol{x'_{2ia}}\boldsymbol{e_i}\boldsymbol{e'_i}\boldsymbol{x_{2ia}}]$$

$$= v_i[\sum_{a=1}^{m} \boldsymbol{x'_{1ia}}\boldsymbol{E_i}\boldsymbol{x_{1ia}} - 2\sum_{a=1}^{m} \boldsymbol{x'_{1ia}}\boldsymbol{E_i}\boldsymbol{x_{2ia}} + \sum_{a=1}^{m} \boldsymbol{x'_{2ia}}\boldsymbol{E_i}\boldsymbol{x_{2ia}}]$$

$$= v_i[\mathbf{Tr}\boldsymbol{X'_{1i}}\boldsymbol{E_i}\boldsymbol{X_{1i}} - 2\mathbf{Tr}\boldsymbol{X'_{1i}}\boldsymbol{E_i}\boldsymbol{X_{2i}} + \mathbf{Tr}\boldsymbol{X'_{2i}}\boldsymbol{E_i}\boldsymbol{X_{2i}}]$$

where m is the number of dimensions of the embedding (typically 3) and E_i is a matrix with $e_{ii} = 1$ and all other elements zero.

Summing over all elements gives

$$\mathbf{Tr}\boldsymbol{X_1'VX_1} - 2\mathbf{Tr}\boldsymbol{X_1'VX_2} + \mathbf{Tr}\boldsymbol{X_2'VX_2}$$

Combining with equation 8.27 in Borg & Groenen gives

$$\sigma(\boldsymbol{X_1}, \boldsymbol{X_2}) \leq \eta_{\delta 1}^2 + \mathbf{Tr}\boldsymbol{X_1'WX_1} - 2\mathbf{Tr}\boldsymbol{X_1'B(X_1)X_1} + \eta_{\delta 2}^2 + \mathbf{Tr}\boldsymbol{X_2'WX_2} - 2\mathbf{Tr}\boldsymbol{X_2'B(X_2)X_2} + \mathbf{Tr}\boldsymbol{X_1'VX_1} - 2\mathbf{Tr}\boldsymbol{X_1'VX_2} + \mathbf{Tr}\boldsymbol{X_2'VX_2} = \tau(\boldsymbol{X_1}, \boldsymbol{Z})$$

$\tau(\boldsymbol{X_1}, \boldsymbol{Z})$ achieves its minimum when $\nabla\tau(\boldsymbol{X_1}, \boldsymbol{Z}) = 0$. Holding $\boldsymbol{X_2}$ constant, we calculate the gradient and solve the system of linear equations as follows:

$$2\boldsymbol{WX_1} - 2\boldsymbol{B(Z)Z} + 2\boldsymbol{VX_1} - 2\boldsymbol{VX_2} = 0$$
$$(\boldsymbol{V} + \boldsymbol{W})\boldsymbol{X_1} = \boldsymbol{VX_2} + \boldsymbol{B(Z)Z}$$
$$\boldsymbol{X_1} = (\boldsymbol{V} + \boldsymbol{W})^+[\boldsymbol{VX_2} + \boldsymbol{B(Z)Z}]$$

This is the update formula for $\boldsymbol{X_1}$, where $(\boldsymbol{V} + \boldsymbol{W})^+$ is the Moore-Penrose inverse of $\boldsymbol{V} + \boldsymbol{W}$. The update formula for $\boldsymbol{X_2}$ is calculated similarly. The algorithm alternately updates $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ using the coordinates calculated at the previous step.

MultiMDS also incorporates a small distance-decay prior to reduce noise. The corrected contact frequency is calculated as $c_{corrected} = c_{observed} * (1 - k) + c_{expected} * k$, where $c_{observed}$ is the observed contact frequency, $c_{expected}$ is the average contact frequency for loci pairs with the same linear separation, and k is a weight parameter. k is set to 0.05 by default.

Implementation

MultiMDS was implemented in python as a modification of miniMDS (Rieber and Mahony, 2017). The joint MDS algorithm was implemented using a modification of MDS from scikit-learn (Pedregosa et al., 2011). Distances were calculated as (contact frequency)$^{-1/4}$. (Wang et al., 2015)

Parameter selection

Similarity weights were selected by identifying the value at which increasing the similarity weight does not further increase reproducibility (see Fig. 1B).

Comparison to independent structural inference and alignment

To compare MultiMDS with an alternate approach that aligns independently estimated genome structures, we first performed structural inference using miniMDS on each Hi-C dataset independently. The two structures were then aligned using the Kabsch algorithm (Kabsch, 1976).

Compartment analysis

Compartment scores were calculated as described by (Lieberman-Aiden et al., 2009). Scores were normalized to range between -1 and 1. The active compartment was defined as the

compartment with greater enrichment for enhancers, promoters, and transcribed regions, as defined by ChromHMM 15-state annotation on ENCODE cell types (Ernst et al., 2011). The compartment axis for each chromosome was identified using linear support vector regression implemented in scikit-learn with default parameters (Pedregosa et al., 2011).

The fraction of relocalization occuring along each axis was normalized by dividing by the axis length, which was calculated as the average distance of each coordinate from the centroid along that axis.

Relocalization peak analysis
Peaks in relocalization magnitude were called using a continuous wavelet transform (Du et al., 2006), with peak widths ranging from 1 to 10. Relocalization peaks with absolute compartment score changes of less than 0.2 were defined as compartment-independent relocalizations. Compartment-independent relocalized bins were overlapped with ChIP-seq peaks and ChromHMM states using bedtools to calculate coverage (Quinlan and Hall, 2010), and p-values were calculated using an independent sample t-test. The ChromHMM states 4_Strong_Enhancer, 5_Strong_Enhancer, 6_Weak_Enhancer, and 7_Weak_Enhancer were classified as enhancers, and 12_Repressed was classified as repressed.

Differential ChIP-seq peaks
MACS2 (Zhang et al., 2008) was used to call broad peaks from Nup60 ChIP-seq data under glucose and galactose conditions separately. Glucose and galactose peaks were merged, and edgeR (Robinson et al., 2010) was run on the tag counts in each peak to identify differential enrichment between conditions. The biological coefficient of variation was estimated as 0.1.

Results

MultiMDS: a principled approach for comparing genome structures

We developed MultiMDS to quantify locus-specific relocalization between Hi-C datasets. MultiMDS takes as input two Hi-C contact matrices and outputs two aligned 3D structures (Fig. 1A). Relocalization is calculated as the locus-specific Euclidean distance between aligned structures. Structural inference is based on multidimensional scaling (MDS), which finds a low-dimensional embedding for a distance matrix (calculated from the contact matrix) that minimizes the difference between the input distances and the embedded distances. It is possible to align independently estimated MDS structures, for example (Kabsch, 1976). However, due to the inherent stochasticity of MDS, independent structural inference followed by alignment overestimates the difference between datasets and results in irreproducible output.

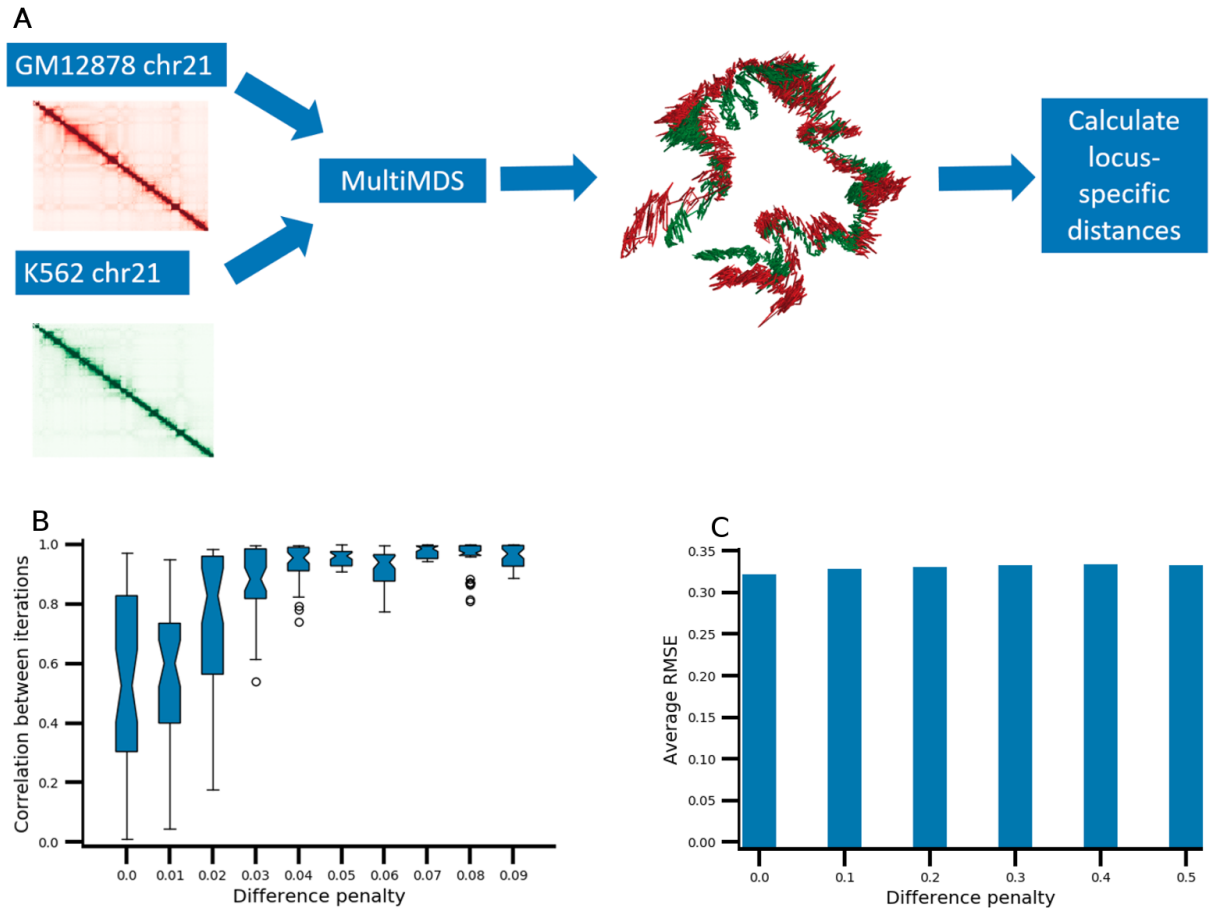


Figure 1. A) Example of MultiMDS applied to GM12878 and K562 chr21 datasets at 10kb resolution. **B)** Pairwise correlations between multiple iterations of MultiMDS run on GM12878 and K562 chr21, measured across a range of difference penalties. Zero penalty represents independent inference and alignment. **C)** Embedding error for MultiMDS run on GM12878 and K562 chr21, measured across a range of difference penalties.

To address this issue, MDS embedding and alignment is performed simultaneously on both datasets, while minimizing the difference between embeddings, weighted by an adjustable penalty. Even relatively small penalties suffice to improve the reproducibility of alignment, measured as the correlation between pairwise MultiMDS output across multiple iterations with the same input, to near perfect (Fig. 1B). This does not come at the cost of embedding accuracy, as the embedding error (measured as the root mean square difference between the input distance matrix and the embedding) for each dataset increases little even at relatively large penalties (Fig. 1C), suggesting that there are multiple structures that fit the data similarly well. MultiMDS provides the option to use the same partitioning algorithm as miniMDS to improve efficiency (Rieber and Mahony, 2017). Like miniMDS, MultiMDS is computationally efficient, with little increase in computational time relative to independent structure inference and alignment (Supplementary Fig. 1).

MultiMDS analysis detects known galactose-dependent genome rearrangements in yeast
Using MultiMDS, we compared the chromosome conformation of yeast grown with glucose to yeast grown with galactose (Kim et al., 2017). Because Hi-C was performed in hybrid yeast (*Saccharomyces cerevisiae* x *Saccharomyces uvarum*), it was possible to phase the data by homologs. Supplementary Fig. 2 shows the alignment of the *S. uvarum* chr13 structures from glucose and galactose conditions. The Has1-Tda1 locus, which had been shown to pair with its homolog upon galactose induction, is highlighted in red. The relocalization is visually apparent from the structures and is quantified in Fig. 2A. Though standard methods are unable to determine which homolog drives the pairing, our analysis suggests that only the *S. uvarum* locus relocalizes (Fig. 2A,B).

We also confirmed the relocalization of the Gal1-Gal7-Gal10 locus (Fig 2C,D). Gal3 and Gal4 also relocalize in the presence of galactose (Fig. 2E-H), though this had not been reported in the original paper. The Gal1-Gal7-Gal10 relocalization was stronger for the *S. uvarum* homolog, whereas the Gal3 relocalization only occurred for the *S. cerevisiae* homolog. The Gal4 relocalization occurred for both homologs.

On chr12, we observed that the rDNA genes create a boundary between the upstream and downstream genomic regions. Because this made chr12 difficult to align, we aligned the two parts of the chromosome separately (Supplementary Fig. 3). For the upstream region, which excluded the rDNA genes, we observed relocalization of Gal2, with similar magnitude between homologs (Fig. 2I,J). In the other region, the *S. cerevisiae* rDNA genes displayed the strongest relocalization of any locus in the genome (Fig. 2K,L), consistent with observations of the nucleolus changing conformation upon galactose induction (Harris et al., 2013). Relocalization at these genes cannot be observed with independent structure inference and alignment, demonstrating the importance of joint structural inference (Supplementary Fig. 4).

Yeast genes, including Gal1-Gal7-Gal10, relocalize to the nuclear periphery upon activation, where they interact with the nuclear pore components, including Nup60 (Kim et al., 2017). Differential ChIP-seq enrichment of Nup60 in the presence of galactose relative to glucose had been qualitatively observed in Gal1-Gal7-Gal10, though not Has1-Tda1 (Kim et al., 2017). Using peak calling, we observed Nup60 peaks throughout the gene bodies of Gal1-7-10 (Supplementary Figs. 5A, 6A), Gal2 (Supplementary Figs. 5B, 6A), and Tda1 but not Has1 (Supplementary Figs. 5C, 6A). Differential enrichment was also found in the downstream portion of Gal3 (Supplementary Figs. 5D, 6A) and near the transcription start site of Gal4 (Supplementary Figs. 5E, 6A). As a negative control we found that Hxt1, a glucose transporter, lost Nup60 binding in the presence of galactose (Supplementary Figs. 5F, 6A). Tda1 and all relocalized Gal genes were upregulated, and Has1 and Hxt1 were downregulated (Supplementary Fig. 6B).

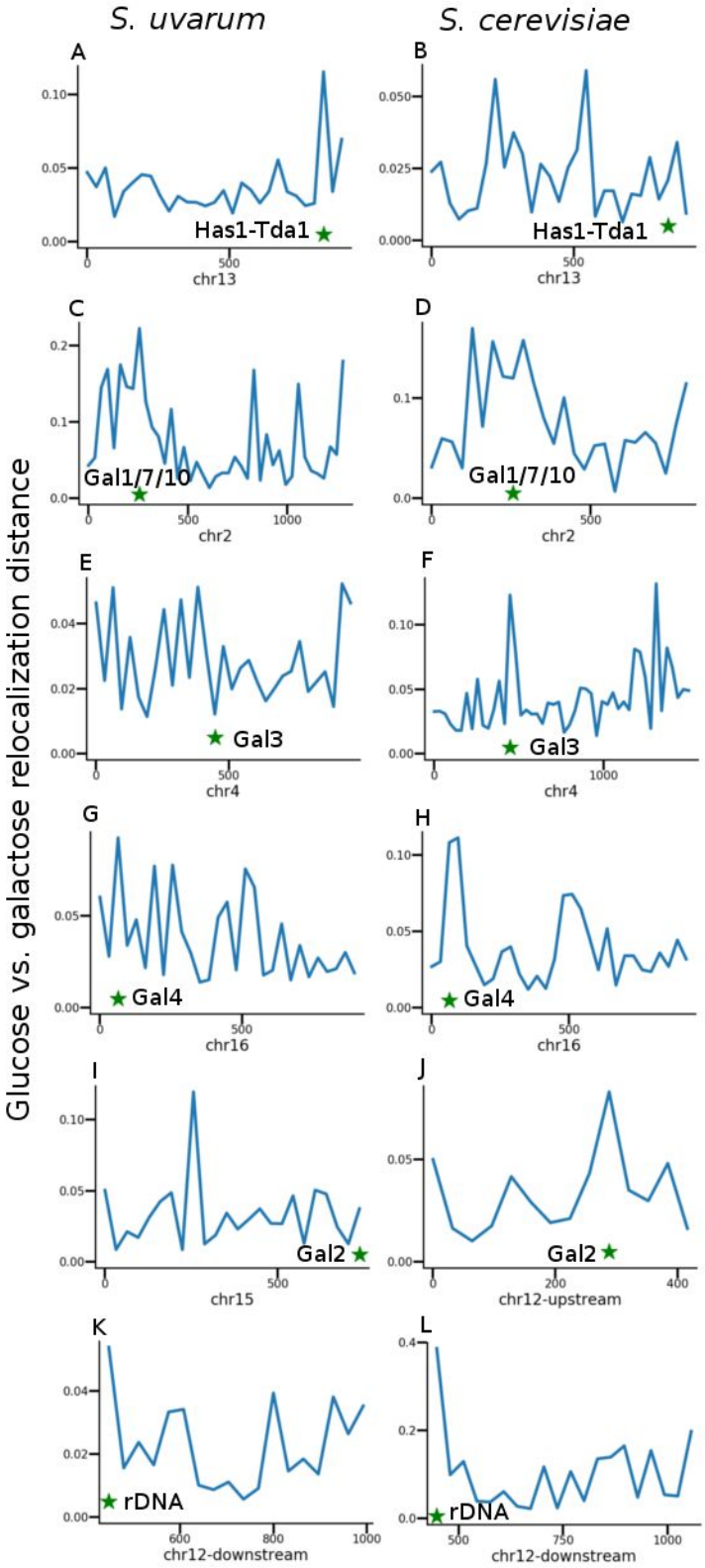


Figure 2. Relocalization of *S. cerevisiae* x *S. uvarum* hybrid yeast loci between glucose and galactose conditions, separated by homolog, for selected chromosomes.

Inter-compartment rearrangements dominate mammalian cell-specific differences in genome structure

We used A/B compartment changes to validate MultiMDS in mammalian datasets at 100kb resolution. Compartment score is a measure of the position of a locus along an axis between the nuclear interior (A compartment) and periphery (B compartment) (Luperchio et al., 2017) (Supplementary Fig. 7). Thus compartment changes are relocalizations, though not all relocalizations are compartment changes. For example, the antiviral genes Mx1 and Mx2 have a weaker A compartment score in K562 relative to GM12878, associated with a loss of activity, and can be observed relocalizing along the compartment axis between these cell types (Supplementary Fig. 8). Compartment score is calculated as the first principal component of the contact frequency correlation matrix and does not involve structural inference, so it represents a relatively independent validation. We calculated the degree to which compartment changes between GM12878 and K562 are correlated with relocalization along a single axis in MultiMDS output structures. Though there are no other metrics specifically designed to quantify locus-specific changes, a simple approach is to calculate the Pearson or Spearman correlation between the rows of the two contact frequency matrices, so these metrics were used for comparison. Relocalization measured by MultiMDS had a higher correlation with compartment changes for all chromosomes (Fig. 3).

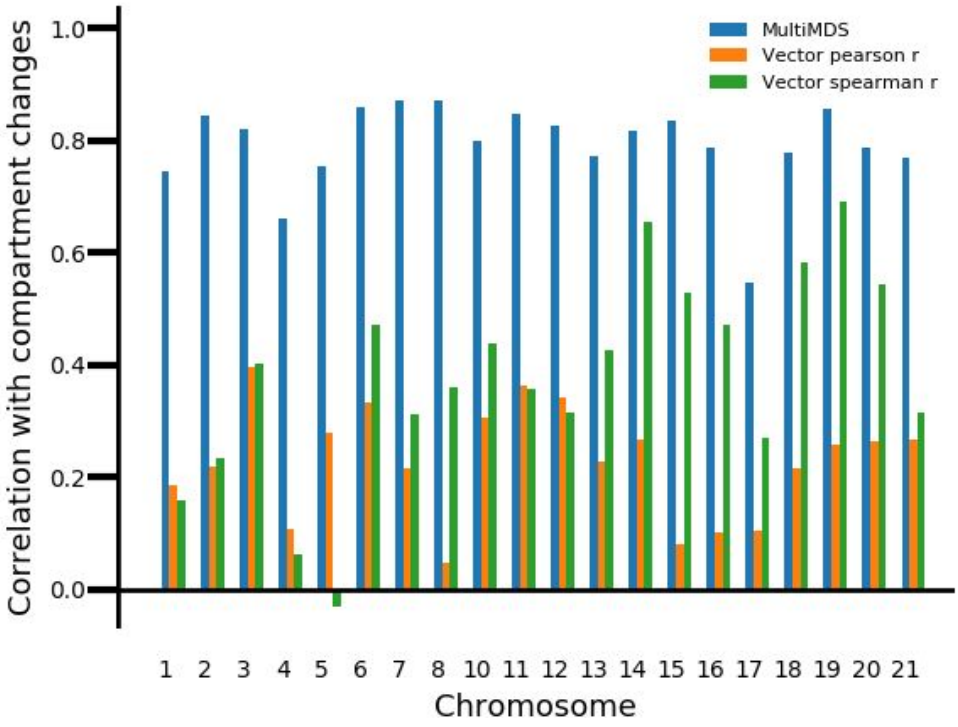


Figure 3. Correlation of compartment changes with locus-specific changes measured by Pearson correlation, Spearman correlation, or MultiMDS relocalization along the compartment SVR axis.

We found that the compartment axis is enriched for relocalization in pairwise comparisons of ENCODE cell types (GM12878, K562, KBM7, HUVEC, HMEC, and NHEK), even when normalized to axis length ($p = 2.8 \times 10^{-149}$) (Fig. 4A). Comparisons of mouse cell types (G1E, HPC-7, mESC, and hepatocyte) revealed a similar magnitude of compartment enrichment, indicating that this effect is not species-specific ($p = 10^{-41}$) (Fig. 4B). Compartment changes are also enriched when comparing lymphoblastoid cell lines (LCLs) from different individuals ($p = 3.4 \times 10^{-36}$) (Fig. 4C). Though the total change between these cell lines is small (Fig. 4E), the degree of compartment enrichment is similar to that seen in intra-cell-type comparisons. Compartment enrichment was not seen when comparing GM12878 replicates ($p = 0.08$) (Supplementary fig. 9), and compartment changes were in fact depleted in mouse cell type replicates ($p = 3 \times 10^{-9}$) (Fig. 4D). The enrichment of compartment changes cannot be observed using independent structure inference and alignment (Supplementary Fig. 10).

Various architectural proteins organize chromatin through distinct mechanisms. CTCF (Nora et al., 2017) and Brd2 (Hsu et al., 2017) enforce TAD boundaries, and their loss does not affect compartmentalization. Cohesin appears to oppose compartmentalization by forming TADs, and its loss causes a finer-grained compartmentalization to appear (Schwarzer et al., 2017). We confirmed the relationships of these proteins to compartmentalization by comparing wild-type Hi-C datasets to their corresponding depletions. Compartment changes were enriched in the cohesin depletion versus control ($p < 10^{-4}$) (Fig. 5A), were not enriched in the Brd2 depletion ($p = 0.32$) (Fig. 5B), and were slightly underrepresented in the CTCF depletion ($p < 0.05$) (Fig. 5C).

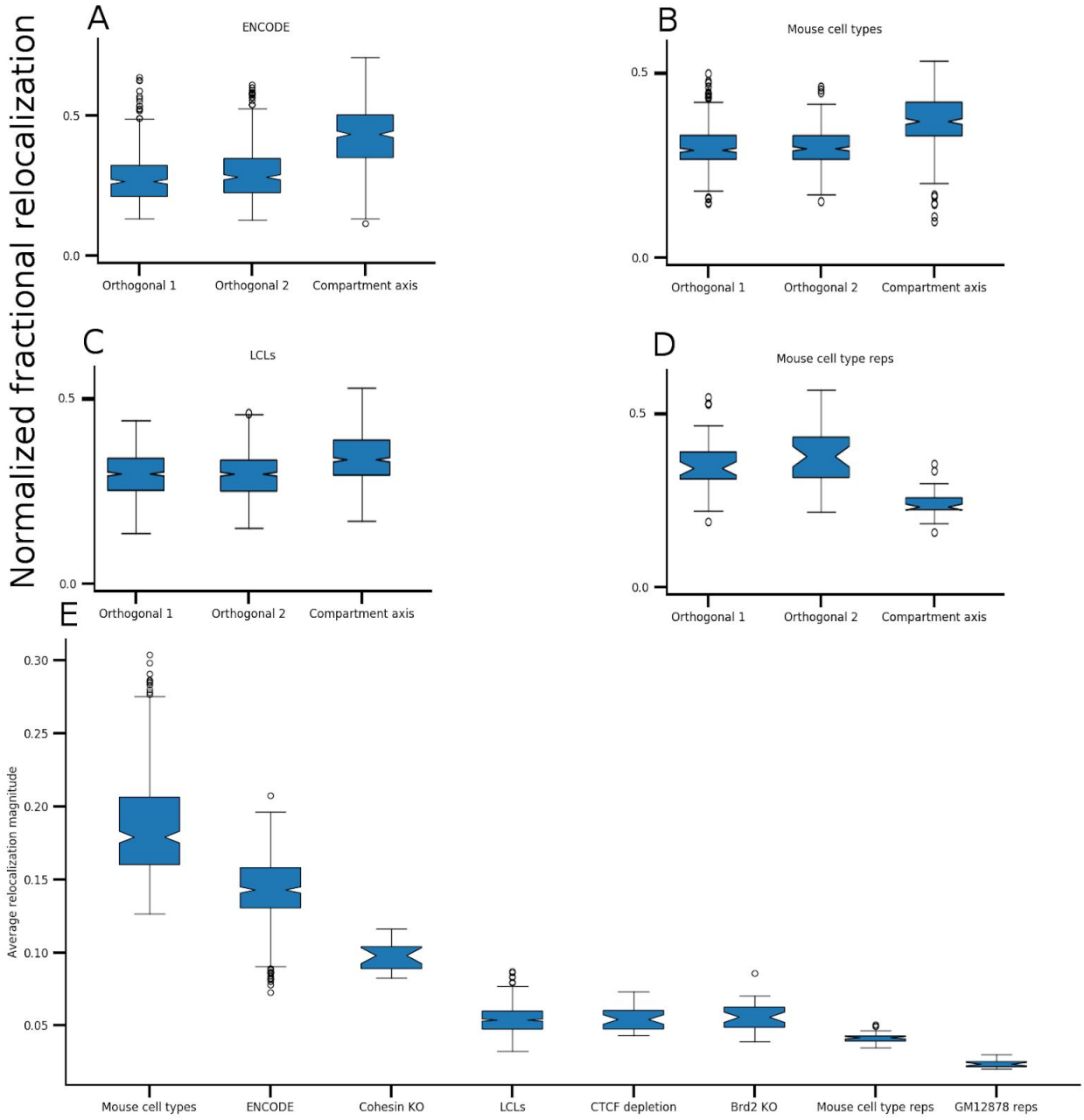


Figure 4. Fraction of total relocalization distance along each axis for all chromosomes for comparisons of **A)** ENCODE cell lines (GM12878, K562, KBM7, HUVEC, HMEC, and NHEK), **B)** mouse cell types (G1E, HPC-7, mESC, and hepatocyte), **C)** LCLs, and **D)** mouse cell type replicates. Total magnitude of change along all axes is shown in **E**.

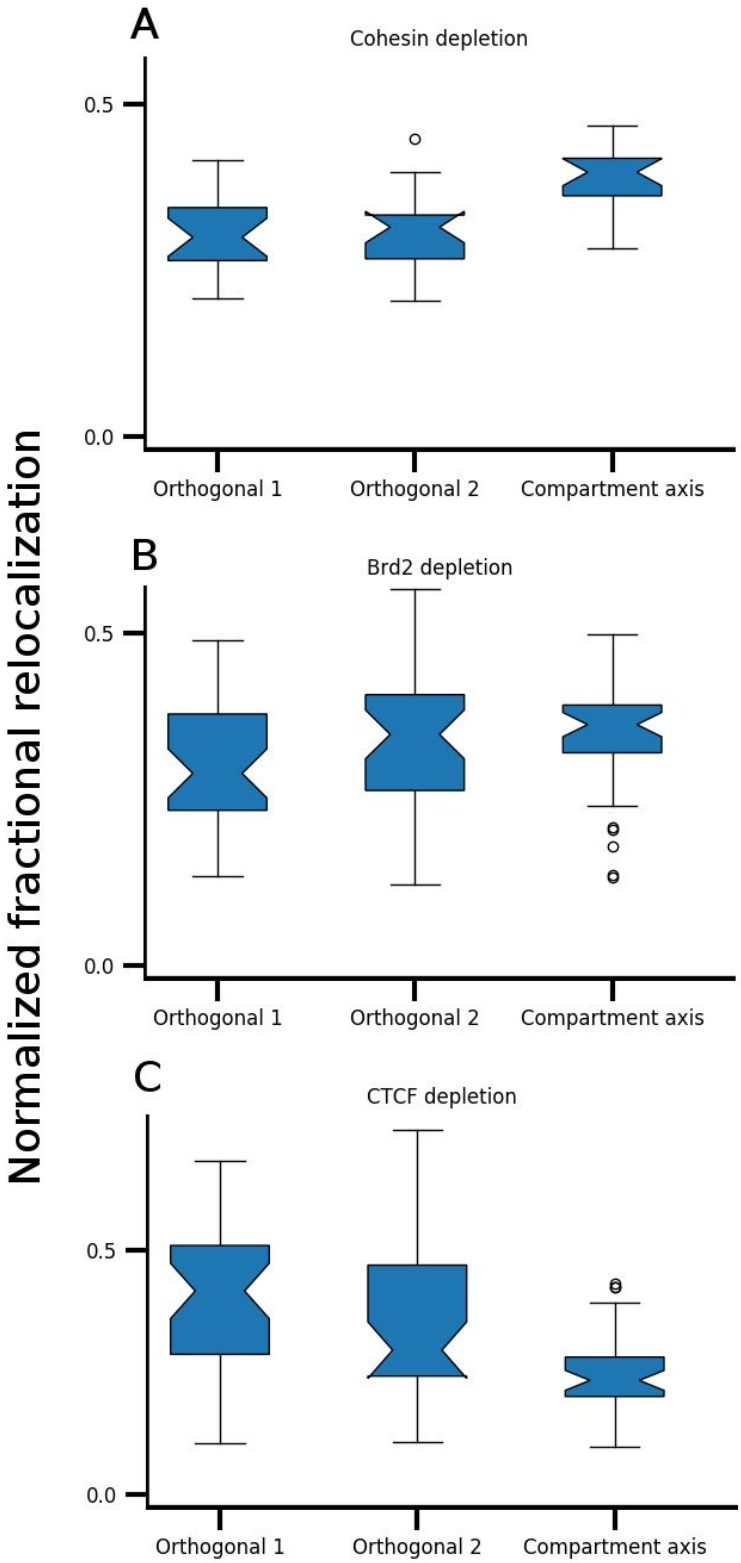


Figure 5. Fraction of total relocalization distance along each axis for all chromosomes upon depletion of architectural proteins: **A)** cohesin, **B)** Brd2, and **C)** CTCF.

MultiMDS detects intra-compartment rearrangements across cell types

Next we focused on relocalizations that occur independently of compartment score changes in GM12878 compared to K562. We called peaks in relocalization magnitude for 10kb bins and identified those with little change in compartment score (Supplementary Fig. 11). Though most relocalization peaks overlap compartment changes, we noted that few compartment change peaks occur without relocalization peaks, as we would expect. Because the B compartment has little biological signal, we focused on peaks that were within the A compartment in both cell types. Relative to the background A compartment, these peaks were enriched for H3K27ac ($p = 1.8 \times 10^{-8}$), H3K4me1 ($p = 9.4 \times 10^{-8}$), H3K4me3 ($p = 1.5 \times 10^{-6}$), H3K9ac ($p = 1.9 \times 10^{-5}$), H2AZ ($p = 1.1 \times 10^{-8}$), H3K4me2 ($p = 3.1 \times 10^{-6}$), and ChromHMM enhancer states ($p = 1.3 \times 10^{-12}$) in GM12878, and the polycomb-associated marks H3K27me3 ($p = 2.4 \times 10^{-4}$) and EZH2 ($p = 4.8 \times 10^{-4}$), and the ChromHMM repressed state in K562 ($p = 3.5 \times 10^{-4}$) (Fig. 6). This was not due to the peaks having different compartment scores or changes in compartment score relative to the background (Supplementary Fig. 12). The enrichments cannot be observed using independent MDS (Supplementary Fig. 13).

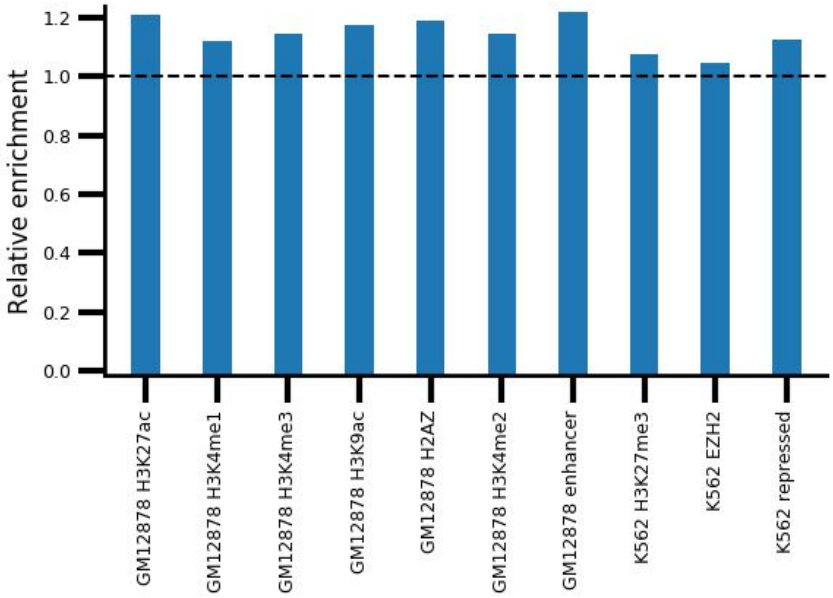


Figure 6. Enrichment of selected chromatin marks in A compartment loci that relocalize without changing compartment score relative to those that do not relocalize.

Using hierarchical clustering of ChIP-seq coverage, we identified distinct subsets of relocalization peaks (Fig. 7). A large minority of peaks are highly enriched for active marks. While some of these peaks are active in both GM12878 and K562, some lose active marks in K562 and gain H3K27me3. Other peaks have H3K27me3 in only one cell type but lack active marks in the other cell type.

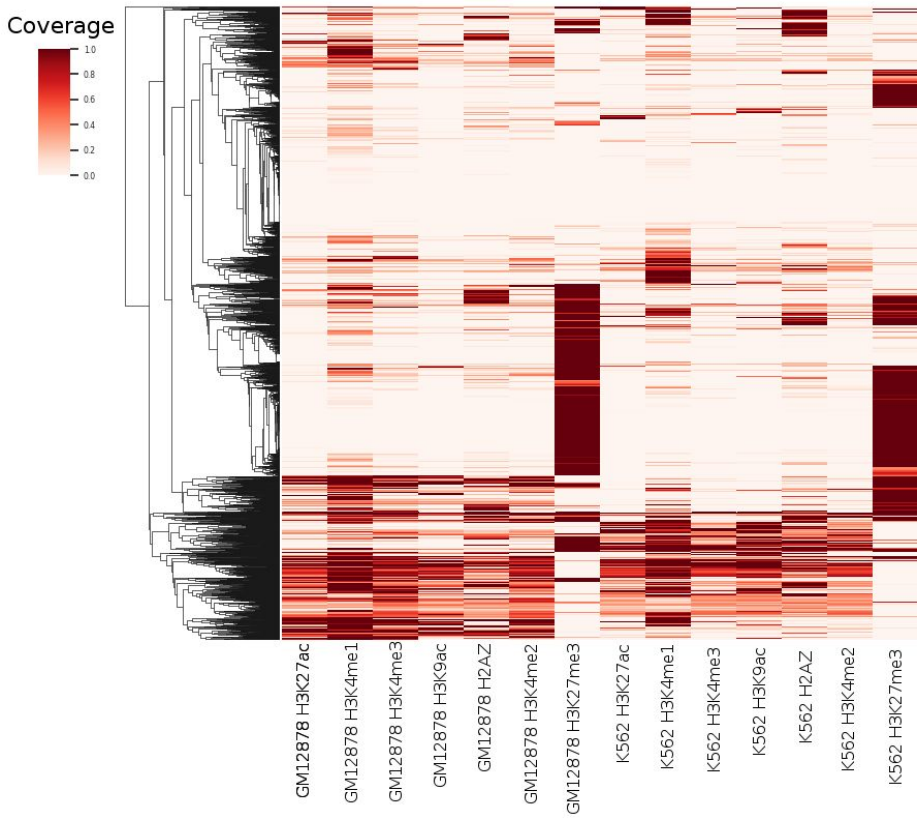


Figure 7. Hierarchical clustering of ChIP-seq peak coverage in relocalization peaks.

Discussion

MultiMDS is a computationally efficient tool for quantifying locus-specific changes between Hi-C datasets. It captures A/B compartment score changes better than alternative methods and quantifies the contribution of these changes to global differences in Hi-C data. At the same time, MultiMDS is able to go further than previous types of Hi-C analysis by identifying changes not based on differential compartmentalization, TADs, or looping. MultiMDS also differs from differential loop calling tools because it can identify the specific locus that drives the loop formation. We demonstrated this by showing that a single homolog drives galactose-inducible homolog pairing at the Has1-Tda1 locus.

Because of the strong correlation between regulatory activity and compartmentalization, changes in compartment score have dominated previous comparisons of Hi-C data. To explore compartment-independent changes, we used Hi-C data from yeast, which lack compartmentalization. Examples of loci that change conformation upon galactose induction, such as Gal1-Gal7-Gal10 (Kim et al., 2017) and the nucleolus (Harris et al., 2013), had been previously identified, but this had not been done systematically. We confirmed these examples and also demonstrated that Gal2, Gal3, and Gal4 relocalize. All the relocalized Gal genes are upregulated and gain Nup60 peaks, suggesting that the effect occurs via nuclear pore association. It had originally been reported that Has1-Tda1 does not gain Nup60 binding, based on visual inspection, but a quantitative analysis revealed a Nup60 peak at Tda1 upon galactose

induction. This is consistent with the upregulation of Tda1 and its role in glucose metabolism (Kim et al., 2017).

Due to the strong effect of compartmentalization on long-range interactions, it is challenging to remove the effect of compartment score changes, which are strongly associated with changes in gene regulation, from other changes in genome structure. The contribution of compartment changes to overall structural changes was enriched in both mouse and human cell types and even in LCLs from different individuals. While a significant amount of compartment switching occurred between cell types, differences between LCLs were mainly driven by a weakening or strengthening of existing compartmentalization. As expected, a large fraction of changes after cohesin depletion are driven by compartment changes, specifically a gain in compartmentalization strength, while Brd2 depletion and CTCF depletion do not significantly affect compartmentalization.

A-compartment loci that relocalize between GM12878 and K562 but do not overlap with compartment score changes were enriched for various ChIP-seq peaks, even relative to the background A compartment, which is generally high in ChIP-seq signal. Some of these loci were enriched for active marks in GM12878 (H3K27ac, H3K4me1, H3K4me3, H3K9ac, H2AZ, and H3K4me2), which appear to represent enhancers, but were enriched for repressive H3K27me3 and EZH2 in K562, suggesting asymmetric changes in the structure/regulation relationship between these cell types. It has been shown that distance from nuclear speckles in the A compartment is independent of the nuclear lamina compartment axis and is correlated with superenhancers and H3K4me3, H3K9ac, and CTCF peaks (Chen et al., 2018). Thus the compartment-independent relocalizations we identified may represent changes in nuclear speckle association. Other relocalization peaks have cell-type-specific H3K27me3 without active marks, which may be associated with polycomb hubs that organize the 3D genome (Denholtz et al., 2013; Schoenfelder et al., 2015).

MultiMDS is a user-friendly tool that provides both visual and quantitative metrics of relocalization, as well as a built-in method for quantifying the contribution of the compartment axis to global changes, which could be used, for example, to determine the role of additional architectural proteins in compartmentalization. Our preliminary results showing the correlation of functional features with compartment-independent relocalizations hints at the existence of novel forms of nuclear organization, which can be further explored using MultiMDS. The number of possible pairwise comparisons and biological questions will increase exponentially as more Hi-C datasets are produced.

Funding
This material is based upon work supported by the National Science Foundation under Graduate Research Fellowship Program Grant No. DGE1255832 (to LR) and ABI Innovation Grant No. DBI1564466 (to SM). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work is also supported by the National Institutes of Health under grants NIGMS R01GM121613 and NIGMS R01GM125722 (to SM).

Conflict of Interest: none declared.

Acknowledgements
The authors thank the members of the Center for Eukaryotic Gene Regulation at Penn State for helpful feedback and discussions.

References

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. Dev. Cell *16*, 47–57.

Borg, I., and Groenen, P.J.F. (2005). Modern Multidimensional Scaling: Theory and Applications (Springer Science & Business Media).

Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O., Guo, L., Collins, R.L., et al. (2018). Multi-platform discovery of haplotype-resolved structural variation in human genomes. BioRxiv 193144.

Chen, Y., Zhang, Y., Wang, Y., Zhang, L., Brinkman, E.K., Adam, S.A., Goldman, R., Steensel, B. van, Ma, J., and Belmont, A.S. (2018). Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. J Cell Biol *217*, 4025–4048.

Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'Shea, C.C., Park, P.J., Ren, B., et al. (2017). The 4D nucleome project. Nature *549*, 219–226.

Denholtz, M., Bonora, G., Chronis, C., Splinter, E., de Laat, W., Ernst, J., Pellegrini, M., and Plath, K. (2013). Long-Range Chromatin Contacts in Embryonic Stem Cells Reveal a Role for Pluripotency Factors and Polycomb Proteins in Genome Organization. Cell Stem Cell *13*, 602–616.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Du, P., Kibbe, W.A., and Lin, S.M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics *22*, 2059–2065.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Ferraiuolo, M.A., Rousseau, M., Miyamoto, C., Shenker, S., Wang, X.Q.D., Nadler, M., Blanchette, M., and Dostie, J. (2010). The three-dimensional architecture of Hox cluster silencing. Nucleic Acids Res. *38*, 7472–7484.

Harris, B., Bose, T., Lee, K.K., Wang, F., Lu, S., Ross, R.T., Zhang, Y., French, S.L., Beyer, A.L., Slaughter, B.D., et al. (2013). Cohesion promotes nucleolar structure and function. Mol. Biol. Cell *25*, 337–346.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010).

Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell *38*, 576–589.

Hsu, S.C., Gilgenast, T.G., Bartman, C.R., Edwards, C.R., Stonestrom, A.J., Huang, P., Emerson, D.J., Evans, P., Werner, M.T., Keller, C.A., et al. (2017). The BET Protein BRD2 Cooperates with CTCF to Enforce Transcriptional and Architectural Boundaries. Mol. Cell *66*, 102-116.e7.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A *32*, 922–923.

Kim, S., Liachko, I., Brickner, D.G., Cook, K., Noble, W.S., Brickner, J.H., Shendure, J., and Dunham, M.J. (2017). The dynamic three-dimensional organization of the diploid yeast genome. ELife *6*, e23623.

Lieberman-Aiden, E., Berkum, N.L. van, 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.

Lin, Y.C., Benner, C., Mansson, R., Heinz, S., Miyazaki, K., Miyazaki, M., Chandra, V., Bossen, C., Glass, C.K., and Murre, C. (2012). Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. Nat. Immunol. *13*, 1196–1204.

Lun, A.T.L., and Smyth, G.K. (2015). diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics *16*, 258.

Luperchio, T.R., Sauria, M.E., Wong, X., Gaillard, M.-C., Tsang, P., Pekrun, K., Ach, R.A., Yamada, N.A., Taylor, J., and Reddy, K. (2017). Chromosome Conformation Paints Reveal The Role Of Lamina Association In Genome Organization And Regulation. BioRxiv 122226.

Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930-944.e22.

Paulsen, J., Sandve, G.K., Gundersen, S., Lien, T.G., Trengereid, K., and Hovig, E. (2014). HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. Bioinformatics *30*, 1620–1622.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rao, S.S.P., 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.

Rickman, D.S., Soong, T.D., Moss, B., Mosquera, J.M., Dlabal, J., Terry, S., MacDonald, T.Y., Tripodi, J., Bunting, K., Najfeld, V., et al. (2012). Oncogene-mediated alterations in chromatin conformation. Proc. Natl. Acad. Sci. U. S. A. *109*, 9083–9088.

Rieber, L., and Mahony, S. (2017). miniMDS: 3D structural inference from high-resolution Hi-C data. Bioinformatics *33*, i261–i266.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Scannell, D.R., Zill, O.A., Rokas, A., Payen, C., Dunham, M.J., Eisen, M.B., Rine, J., Johnston, M., and Hittinger, C.T. (2011). The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 Genes Genomes Genet. *1*, 11–25.

Schoenfelder, S., Sugar, R., Dimond, A., Javierre, B.-M., Armstrong, H., Mifsud, B., Dimitrova, E., Matheson, L., Tavares-Cadete, F., Furlan-Magaril, M., et al. (2015). Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nat. Genet. *47*, 1179–1186.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., H. Haering, C., Mirny, L., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51–56.

Stansfield, J.C., Cresswell, K.G., Vladimirov, V.I., and Dozmorov, M.G. (2018). HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. BMC Bioinformatics *19*, 279.

Ursu, O., Boley, N., Taranova, M., Wang, Y.X.R., Yardimci, G.G., Stafford Noble, W., and Kundaje, A. (2018). GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. Bioinformatics 2701–2707.

Wang, S., Xu, J., and Zeng, J. (2015). Inferential modeling of 3D chromatin structure. Nucleic Acids Res. *43*, e54.

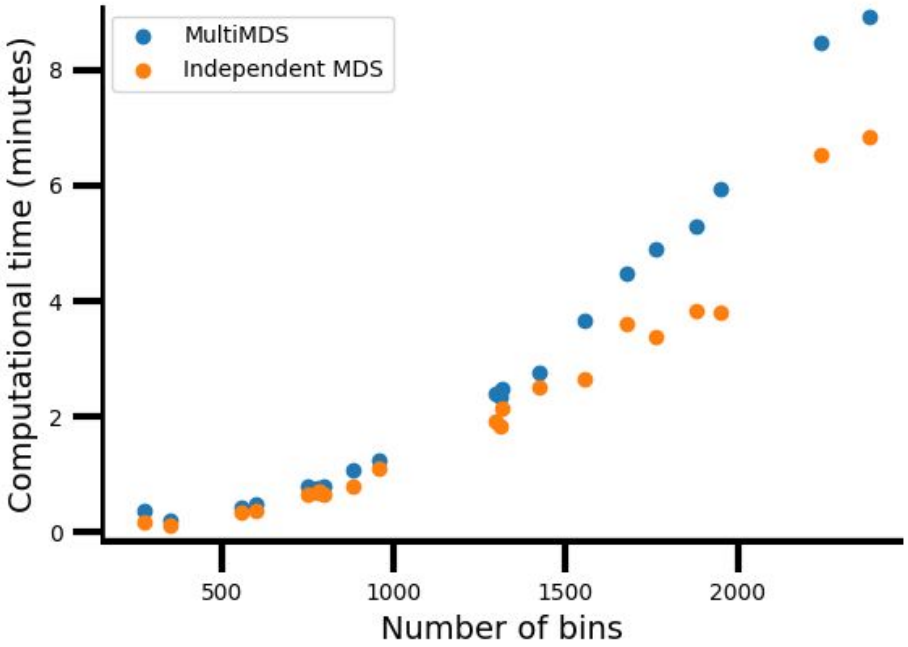
Wilson, N.K., Schoenfelder, S., Hannah, R., Castillo, M.S., Schütte, J., Ladopoulos, V., Mitchelmore, J., Goode, D.K., Calero-Nieto, F.J., Moignard, V., et al. (2016). Integrated genome-scale analysis of the transcriptional regulatory landscape in a blood stem/progenitor cell model. Blood e12–e23.

Yang, T., Zhang, F., Yardımcı, G.G., Song, F., Hardison, R.C., Noble, W.S., Yue, F., and Li, Q. (2017). HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res. *27*, 1939–1949.

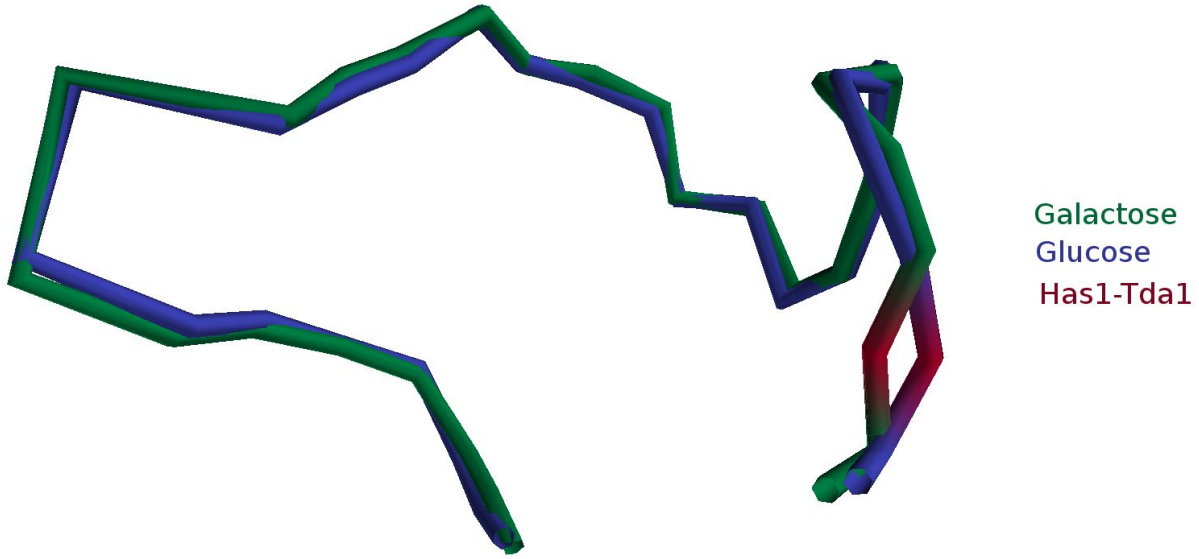
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zufferey, M., Tavernari, D., Oricchio, E., and Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. Genome Biol. *19*, 217.

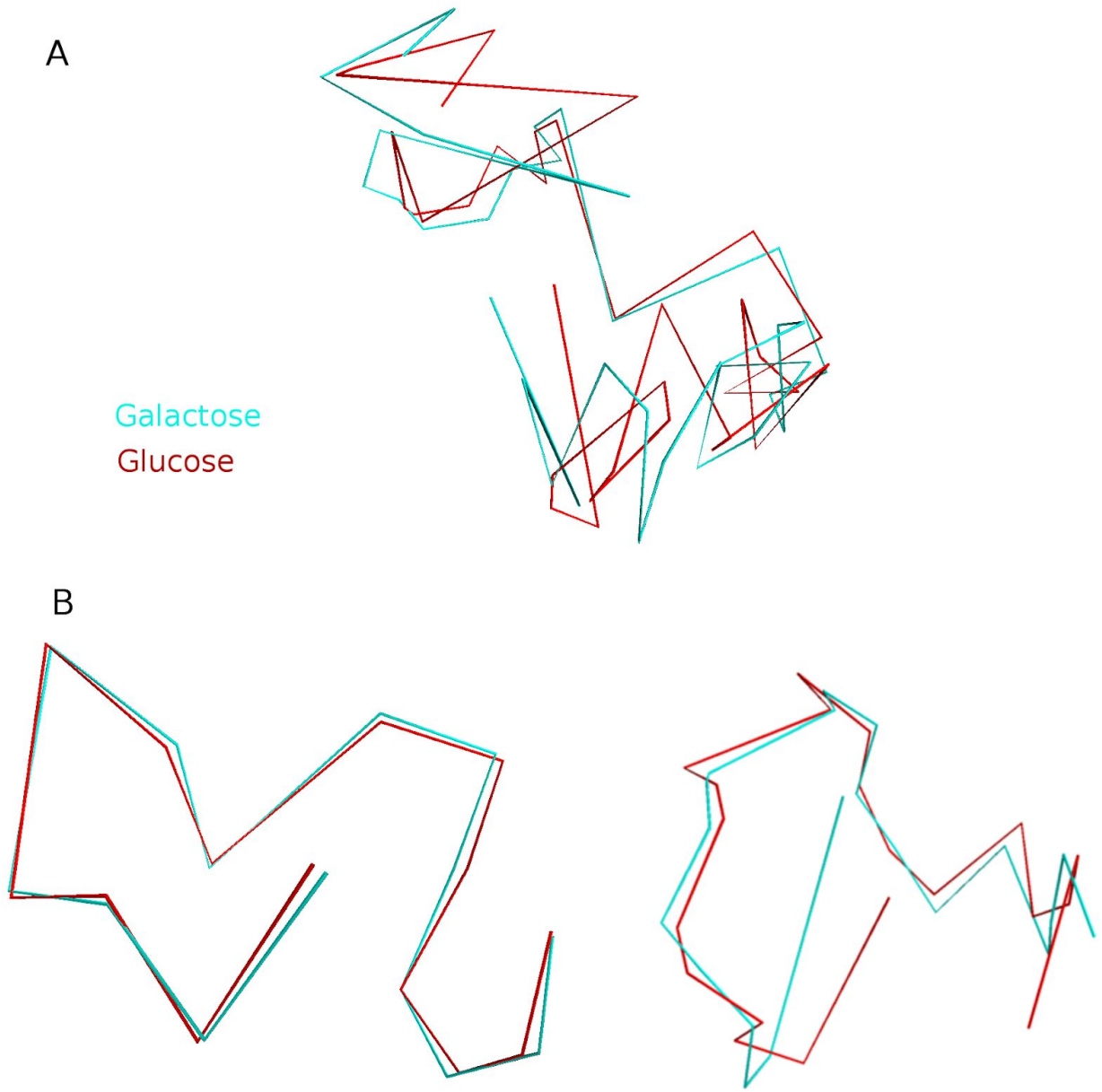
Supplementary Figures



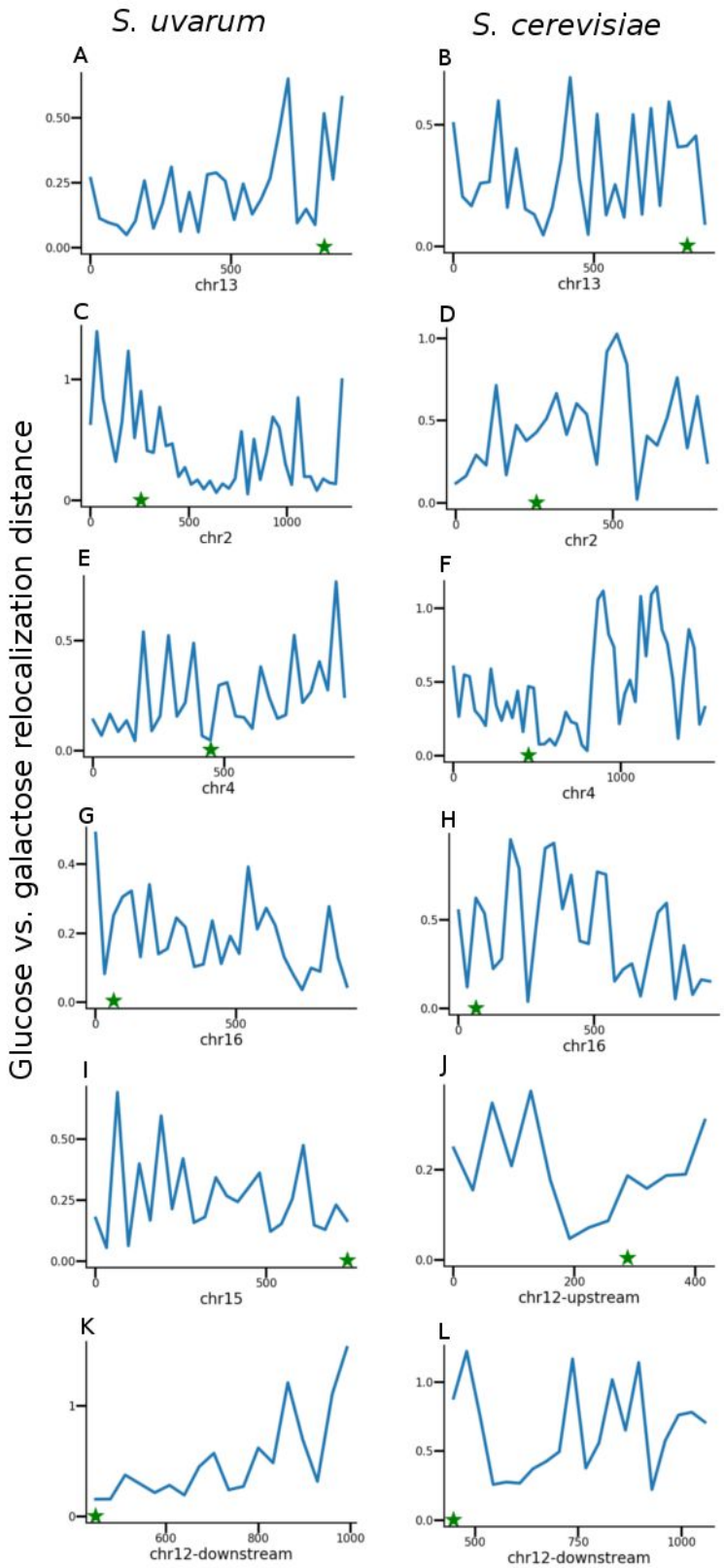
Supplementary Figure 1. Computational time for datasets of various sizes, measured for MultiMDS and for independent MDS structural inference followed by alignment.



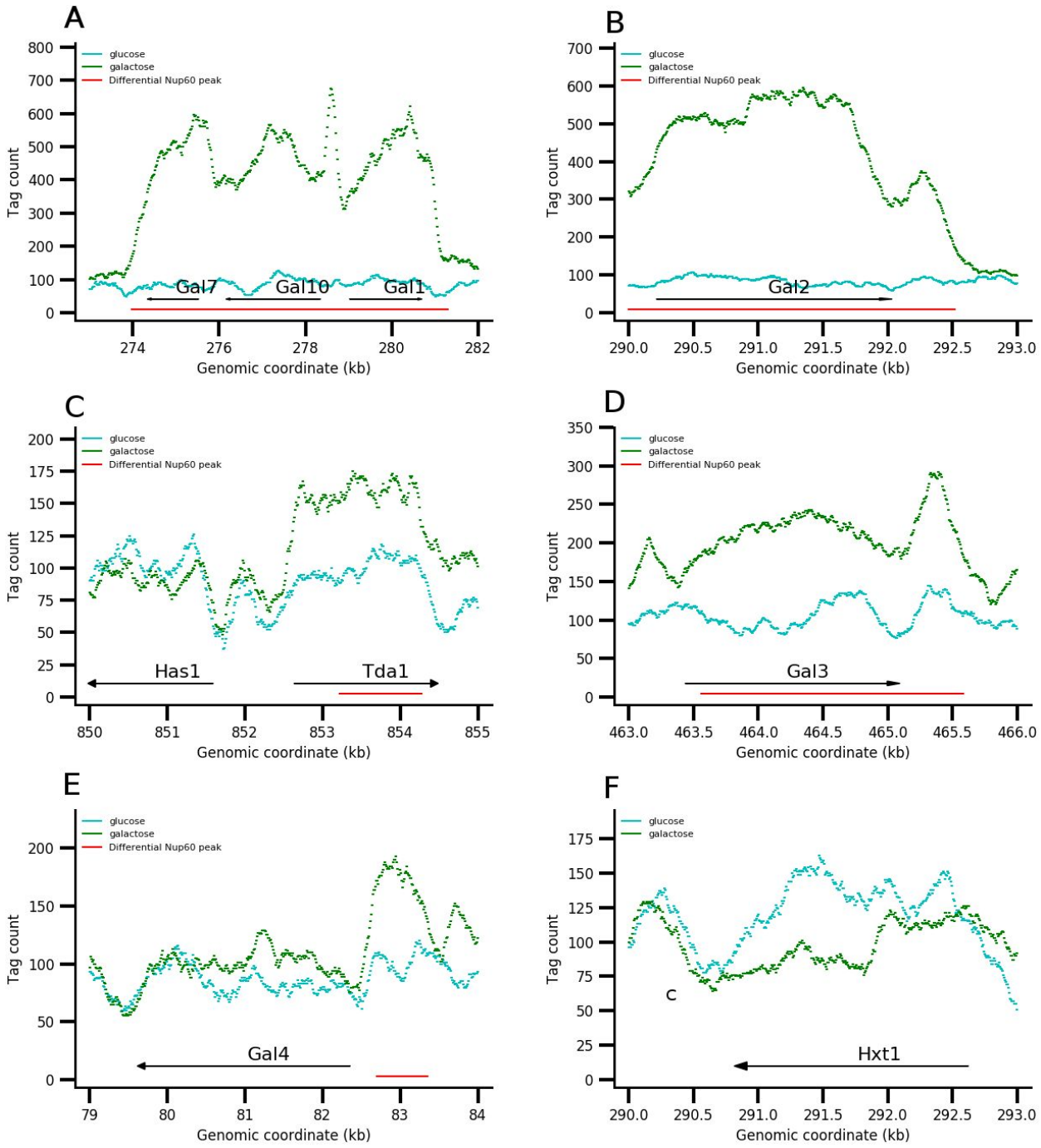
Supplementary Figure 2. *S. uvarum* chr13 aligned structures for glucose (blue) and galactose (green) conditions. Has1-Tda1 locus is highlighted in red.



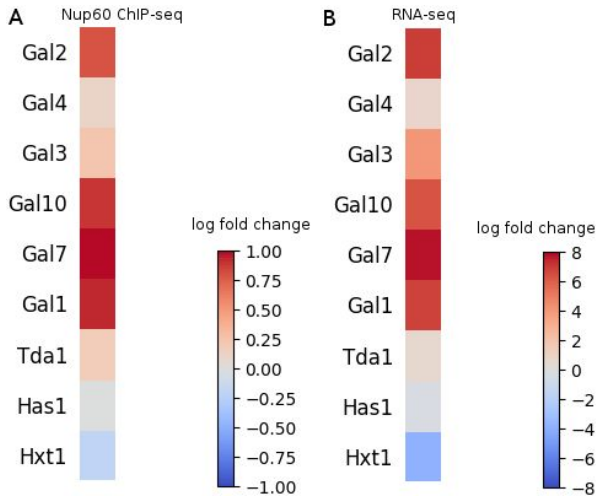
Supplementary Figure 3. chr12 *S. cerevisiae* glucose (red) and galactose (cyan) alignment for **A)** full structures and **B)** split structures.



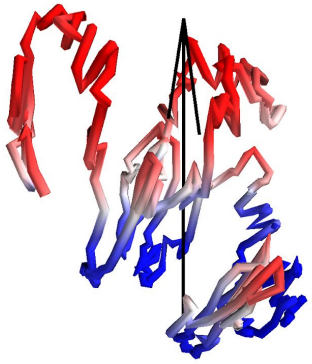
Supplementary Figure 4. Locus-specific relocalization calculated from independent structural inference and alignment.



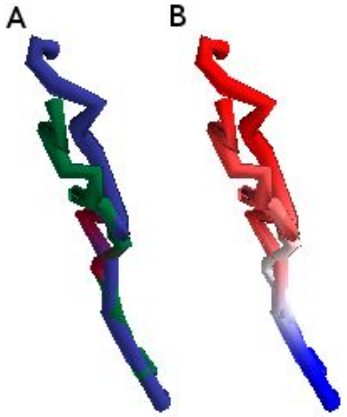
Supplementary Figure 5. Nup60 ChIP-seq tag counts in glucose and galactose conditions at **A)** Gal1-Gal7-Gal10, **B)** Gal2, **C)** Has1-Tda1, **D)** Gal3, **E)** Gal4, and **F)** Hxt1.



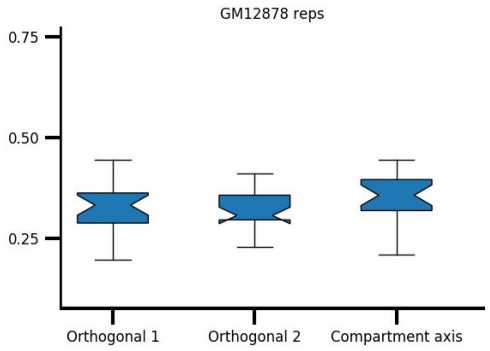
Supplementary Figure 6. Log fold change of **A)** Nup60 ChIP-seq tag counts and **B)** RNA-seq counts at selected genes in galactose relative to glucose.



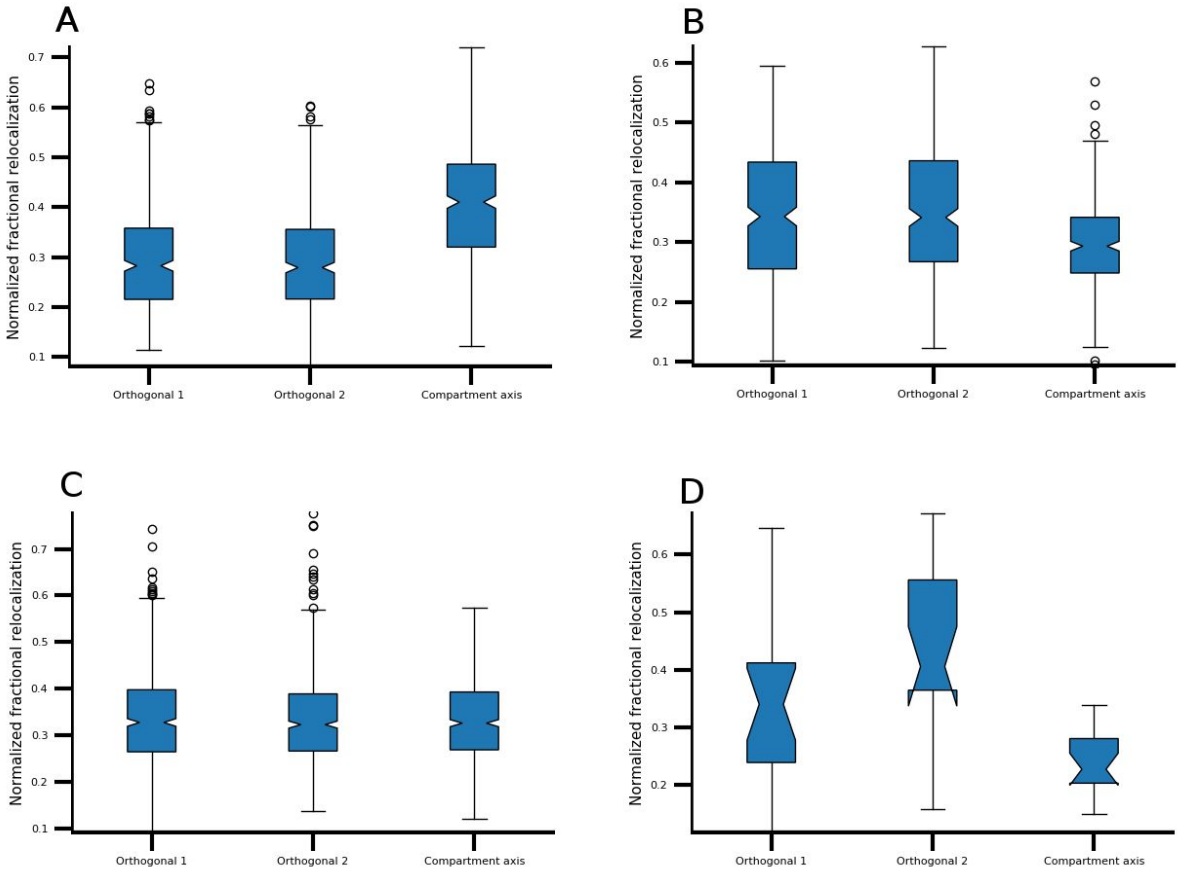
Supplementary Figure 7. Compartment score plotted on GM12878 chr21 at 100kb resolution. Positive scores (corresponding to A compartment) are plotted in reds, and negative scores (corresponding to B compartment) are plotted in blues. The structure has been rotated so that the SVR axis (arrow) aligns with the z axis. Pericentromeric regions have been removed.



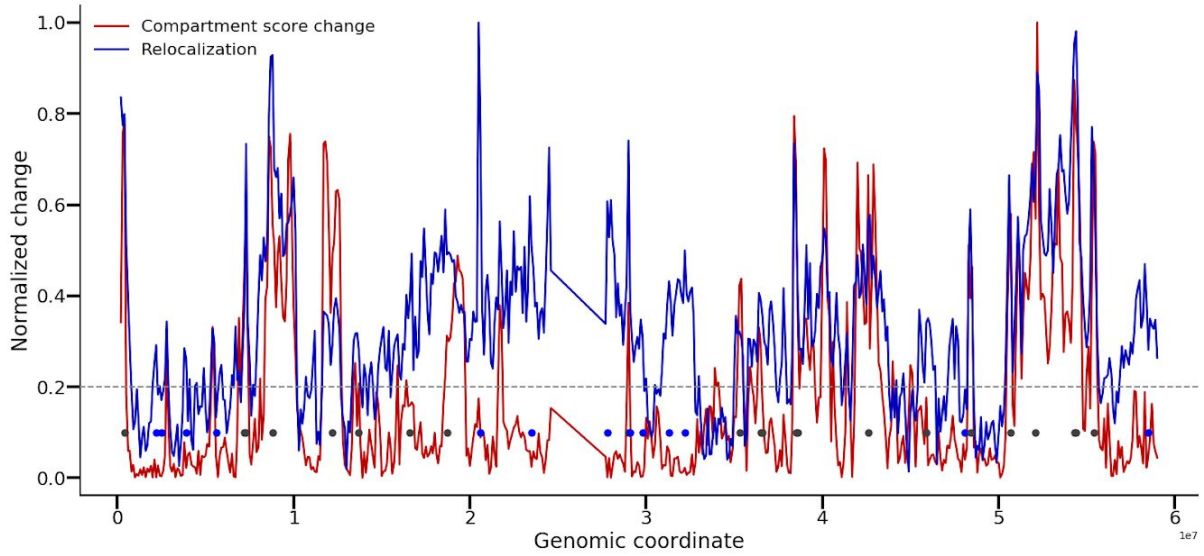
Supplementary Figure 8. chr21:41900000-43900000 aligned to the compartment axis (see Supplementary Fig. 7). **A)** Blue: GM12878. Green: K562: Red: Mx1/Mx2. **B)** Compartment scores. Positive scores are plotted in reds, and negative scores are plotted in blues.



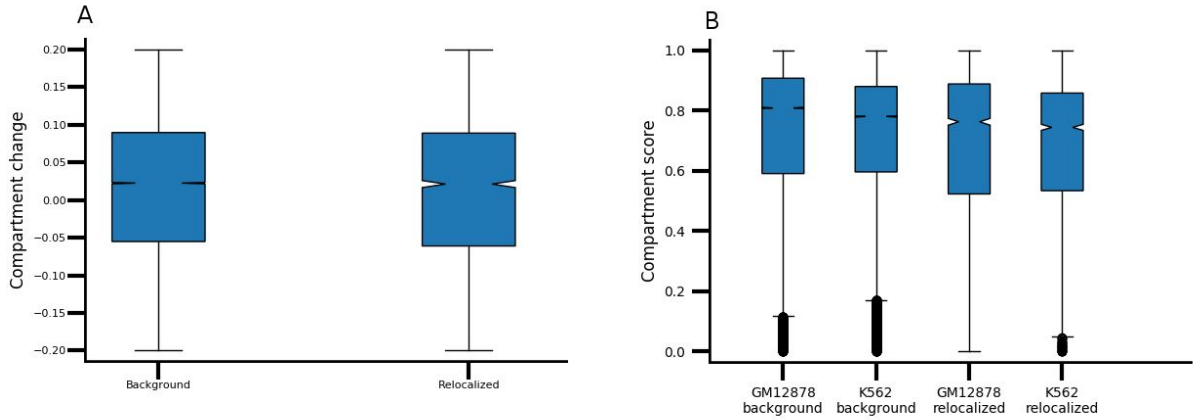
Supplementary figure 9. Fraction of total relocalization distance along each axis for all chromosomes for GM12878 replicates.



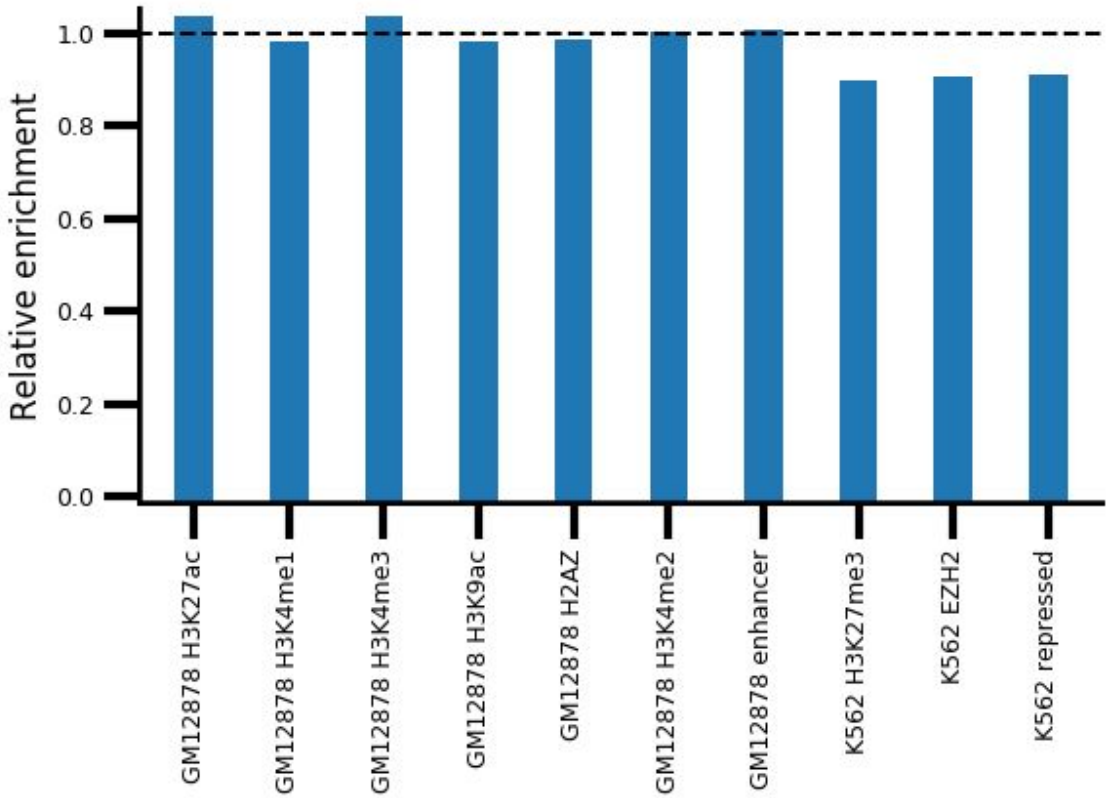
Supplementary Figure 10. Fraction of total relocalization distance along each axis for all chromosomes, calculated from independent structural inference and alignment, for **A)** ENCODE cell lines, **B)** mouse cell types, **C)** LCLs, and **D)** cohesin KO.



Supplementary Figure 11. Relocalization and absolute compartment score changes between GM12878 and K562 chr19, normalized to 1. For ease of visualization, 100kb resolution is shown. Gray dashed line is the threshold for compartment-independent relocalizations. Gray dots are compartment-dependent relocalizations, and blue dots are compartment-independent relocalizations.



Supplementary Figure 12. Validation that compartment-independent relocalizations are not different from background in compartmentalization. **A)** Change in compartment score in 10kb A compartment bins that did not significantly relocalize and compartment-independent relocalizations. **B)** Compartment scores for both cell types in 10kb A compartment bins that did not significantly relocalize and compartment-independent relocalizations.



Supplementary Figure 13. Enrichment of selected chromatin marks in A compartment loci that relocalize without changing compartment score relative to those that do not relocalize, calculated from independent structural inference and alignment.