Comparing 3D genome organization in multiple species using Phylo-HMRF

Yang Yang[1], Yang Zhang[1], Bing Ren[2], Jesse Dixon[3], and Jian Ma[1,*]

[1]Computational Biology Department, School of Computer Science,
Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2]Ludwig Institute for Cancer Research, Department of Cellular and Molecular Medicine,
Moores Cancer Center and Institute of Genomic Medicine, UCSD School of Medicine, La Jolla, CA 92093, USA
[3]Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[*]Correspondence: jianma@cs.cmu.edu

Abstract

Recent developments in whole-genome mapping approaches for the chromatin interactome (such as Hi-C) have offered new insights into 3D genome organization. However, our knowledge of the evolutionary patterns of 3D genome structures in mammalian species remains surprisingly limited. In particular, there are no existing phylogenetic-model based methods to analyze chromatin interactions as continuous features across different species. Here we develop a new probabilistic model, named phylogenetic hidden Markov random field (Phylo-HMRF), to identify evolutionary patterns of 3D genome structures based on multi-species Hi-C data by jointly utilizing spatial constraints among genomic loci and continuous-trait evolutionary models. The effectiveness of Phylo-HMRF is demonstrated in both simulation evaluation and application to real Hi-C data. We used Phylo-HMRF to uncover cross-species 3D genome patterns based on Hi-C data from the same cell type in four primate species (human, chimpanzee, bonobo, and gorilla). The identified evolutionary patterns of 3D genome organization correlate with features of genome structure and function, including long-range interactions, topologically-associating domains (TADs), and replication timing patterns. This work provides a new framework that utilizes general types of spatial constraints to identify evolutionary patterns of continuous genomic features and has the potential to reveal the evolutionary principles of 3D genome organization.

Introduction

In humans and other higher eukaryotes, chromosomes are folded and organized in three-dimensional (3D) space and different chromosomal loci interact with each other (Bonev and Cavalli, 2016; Rowley and Corces, 2018). Recent developments in whole-genome mapping approaches for the chromatin interactome such as Hi-C (Lieberman-Aiden et al., 2009; Rao et al., 2014) and ChIA-PET (Tang et al., 2015) have facilitated the identification of genome-wide chromatin organizations comprehensively, revealing important 3D genome features such as loops (Rao et al., 2014; Tang et al., 2015), topologically-associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012), and A/B compartments (Lieberman-Aiden et al., 2009). A limited number of attempts have been made to analyze these 3D genome features across different species. An earlier study using Hi-C showed that the positions of TADs were largely conserved between human and mouse within syntenic genomic regions (Dixon et al., 2012). Another study demonstrated that evolutionary changes in TAD structure correspond with the creation or elimination of CTCF binding sites using relatively low resolution Hi-C data from rhesus macaque, dog, rabbit,

and mouse (Rudan et al., 2015). More recently, TADs have been shown to have strong conservation in mammalian evolution with the TADs boundaries under potential negative selections against genome rearrangements (Lazar et al., 2018; Fudenberg and Pollard, 2019). These previous analyses pointed to the conservation and changes of 3D genome structure across different species, although a more comprehensive characterization of the detailed evolutionary patterns of 3D genome organization remains unclear. Additionally, as most of the initial comparative analysis of 3D genomes focused primarily on distantly related organisms, there is limited understanding of how 3D genome features may have evolved in closely related mammalian species, especially in recent primate evolution which is of particular interest to understand human specific and great-ape specific gene regulations.

On the algorithmic side, existing computational approaches for comparing 3D genome organization across multiple species have surprisingly limited capability. Importantly, multi-species functional genomic data from various high-throughput epigenomic assays (e.g., Hi-C, ChIP-seq, Repli-seq) are continuous traits in nature. However, such continuous signals are often converted to discrete values to identify distinctive feature patterns (e.g., presence or absence of TADs) for subsequent comparisons, which may cause dramatic loss of information of more subtle differences from the original data. Although methods have been developed to quantitatively analyze the strengths of chromatin interactions from Hi-C data, to the best of our knowledge, there are no existing phylogenetic-model based methods available to analyze Hi-C data as continuous signals across different species in a genome-wide manner to uncover evolutionary patterns of 3D genome organization.

We previously developed a method called phylogenetic hidden Markov Gaussian Processes (Phylo-HMGP) (Yang et al., 2018) to estimate evolutionary patterns given continuous functional genomic data (e.g., Repli-seq) along the genome from multiple species. Phylo-HMGP considers evolutionary affinities among species in a hidden Markov model (HMM), utilizing evolutionary constraints and also dependencies along one-dimensional (1D) genome coordinates. However, the HMMs, as used by Phylo-HMGP, are based on 1D Markov chains, which cannot be simply used to model generalized spatial dependencies (such as those reflected in Hi-C data) to consider the interactions between nodes in an arbitrary graph. Therefore, HMM-based methods cannot be directly applied to discovering patterns of higher-order chromatin interactions from Hi-C contact matrices, which consist of continuous measurements of contact frequencies between a pair of genomic loci.

Here, we develop a new probabilistic model, phylogenetic hidden Markov random field (Phylo-HMRF), which integrates the continuous-trait evolutionary constraints with the hidden Markov random field (HMRF) model, to capture evolutionary patterns of continuous genomic features across species by utilizing generalized spatial constraints. We demonstrated the advantage of Phylo-HMRF using simulation data. In addition, we applied Phylo-HMRF to a new Hi-C dataset from the same cell type (lymphoblastoid cell) in four primate species (human, chimpanzee, bonobo, and gorilla). Phylo-HMRF identified different evolutionary patterns of Hi-C contacts across the four species, including both conserved patterns and lineage-specific patterns. These patterns show strong correlations with long-range Hi-C interactions, TAD structures, and the evolutionary patterns of DNA replication timing in primate species. Phylo-HMRF offers an effective model to potentially reveal important evolutionary principles of 3D genome organization. The source code of Phylo-HMRF can be accessed at: https://github.com/ma-compbio/Phylo-HMRF.

Results

Overview of the Phylo-HMRF model

The overview of the Phylo-HMRF method is shown in Fig. 1. Our goal is to identify different evolutionary patterns of chromatin conformation from multi-species Hi-C data. The input contains the Hi-C

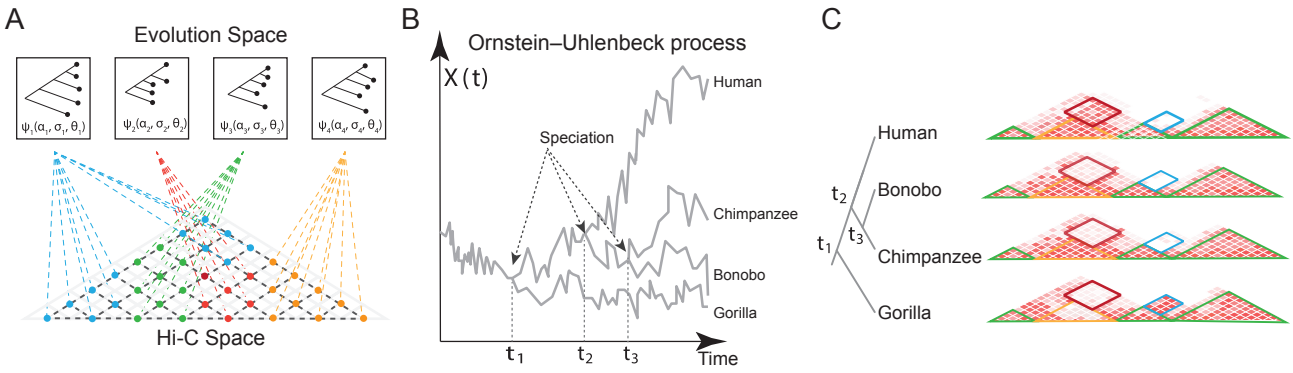

Figure 1: Overview of Phylo-HMRF. **(A)** Illustration of the possible evolutionary patterns of chromatin interaction. The Hi-C space is a combined multi-species Hi-C contact map, which integrates aligned Hi-C contact maps of each species. Each node represents the multi-species observations of Hi-C contact frequency between a pair of genomic loci, with a hidden state assigned. Nodes with the same color have the same hidden state and are associated with the same type of evolutionary pattern represented by a parameterized phylogenetic tree ψ_i. The parameters of ψ_i include the selection strengths α_i, Brownian motion intensities σ_i, and the optimal values θ_i based on the Ornstein-Uhlenback (OU) process assumption. **(B)** Illustration of the OU process over a phylogenetic tree with four observed species. Time axis represents the evolution history. $X(t)$ represents the trait at time t. The trajectories reflect the evolution of the continuous-trait features in different lineages, where the time points t_1, t_2, t_3 represent the speciation events. **(C)** A cartoon example of the possible evolutionary patterns (partitioned with different colors). Phylo-HMRF aims to identify evolutionary Hi-C contact patterns among four primate species in this work. The four Hi-C contact maps represent the observations from the four species, which are combined into one multi-species Hi-C map as the input to Phylo-HMRF, as shown in **(A)**. The phylogenetic tree of the four species in this study is on the left. The partitions with green borders are conserved Hi-C contact patterns. The partitions with red or blue borders represent lineage-specific Hi-C contact patterns.

contact frequency data from each species. We align the Hi-C contacts in each species to the reference genome, and obtain a combined multi-species Hi-C contact map based on the reference genome as shown in Fig. 1A, where each node in the map corresponds to multi-species contact frequencies between the corresponding pair of genomic loci (Supplementary Methods). Hence each node is associated with a multi-dimensional feature as the multi-species observation. We also assume that each node has a hidden state that represents the evolutionary pattern of Hi-C contacts between the corresponding pair of genomic loci. Phylo-HMRF estimates the hidden state of each node by considering both spatial dependencies among nodes encoded by an HMRF and the evolutionary dependencies between species in the phylogeny. The continuous-trait evolutionary models are embedded into the HMRF. Therefore, each hidden state corresponds to an evolutionary model that is represented by a parameterized phylogenetic tree. The output of Phylo-HMRF contains the partition of the combined multi-species Hi-C contact map, where adjacent nodes with the same hidden state are in the same partition. These partitions reflect the distribution of different evolutionary patterns of Hi-C contact frequencies. As shown in Fig. 1B, Phylo-HMRF uses the Ornstein-Uhlenbeck (OU) process as the continuous-trait evolutionary model. Fig. 1C is an illustration of the possible Hi-C evolutionary patterns that Phylo-HMRF aims to uncover across four primate species in this work.

Note that an evolutionary pattern identified by Phylo-HMRF is associated with the conservation or variation of the feature of interest across different species. For example, in this study, we may observe conserved high Hi-C contacts in all the compared species between a specific pair of genomic loci. We may also observe that strong Hi-C contacts only exist in some of the species between a specific pair of genomic loci. These different types of feature distribution across species are representative of the evolutionary patterns that we want to identify as states. In addition, Phylo-HMRF provides a framework to utilize both general types of spatial dependencies among genomic loci and evolutionary relationships among species to identify evolutionary patterns from multi-species continuous-trait features. The general types of spatial dependencies refer to any type of dependencies that can be represented by the edge connections in a graph.

Performance evaluation of Phylo-HMRF using simulation

We evaluated the performance of Phylo-HMRF using simulations to demonstrate improvement in identifying 2D evolutionary feature patterns. We applied Phylo-HMRF to 16 simulated datasets in two sets of simulations, each of which contains 8 datasets. Suppose the samples in simulated datasets correspond to nodes in a graph $\boldsymbol{\mathcal{G}}$. Similar to a Hi-C contact map, $\boldsymbol{\mathcal{G}}$ has a 2D lattice structure of size $n \times n$, where each vertex is associated with a sample. The samples thus represent features of vertices in $\boldsymbol{\mathcal{G}}$. For example, a sample can represent the interaction intensities between the i-th locus and the j-th locus out of n genomic loci in multiple species, $(1 \leq i, j \leq n)$. We also assume that each sample has a class label, or hidden state. Each hidden state is associated with an emission probability function and determines the observed feature of the sample with this state. The hidden states of the samples are assumed to be from a Markov random field (MRF). Thus, the hidden state of a sample is spatially dependent on the hidden states of its neighbors in $\boldsymbol{\mathcal{G}}$. In the simulation evaluation, we first simulated the hidden state configurations of the samples by simulating an MRF through Gibbs sampling (Geman and Geman, 1984). We then simulated the multi-species observations from multi-variate Gaussian distributions with OU model parameters embedded. The details of the data simulation are in Supplementary Methods.

Based on the simulated datasets, we compared Phylo-HMRF with several other methods, including the Gaussian-HMRF method (Zhang et al., 2001), the Gaussian mixture model (GMM), and the *K*-means clustering method, to infer hidden states of the samples. Each method was run repeatedly for 10 times on each simulated dataset with different random initialization, given the number of hidden states $M = 10$. Additionally, we also included two image segmentation methods SLIC (Simple Linear Iterative Clustering) (Achanta et al., 2012) and Quick Shift (Vedaldi and Soatto, 2008) for comparison (Supplementary Methods). Both methods have been effectively used in image segmentation applications. The two methods were also run repeatedly for 10 times each. We adjusted the parameter configurations of each of the two methods such that the number of output segments is 10. To utilize the image segmentation methods for state estimation across species, we considered the graph $\boldsymbol{\mathcal{G}}$ as an image and considered the features of each species as one color channel of the image. The average performance of the 10 results for each method was reported as the final performance of the corresponding method with respect to different types of evaluation metrics. We evaluated the performance of each method by comparing the predicted states and ground truth hidden states, using evaluation metrics Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), Adjusted Rand Index (ARI), Precision, Recall, and F_1 score (Manning et al., 2008; Vinh et al., 2010) (Supplementary Methods).

The evaluation results are shown in Fig. 2 and Table S1. We found that Phylo-HMRF outforms all the other methods on different types of evaluation metrics in each simulated dataset in simulation study I. Phylo-HMRF consistently outperforms Gaussian-HMRF, demonstrating that encoding the evolution information can improve accuracy. Even though all the multi-species observations are simulated from Gaussian distribution, using Gaussian distribution alone in inference may not reveal the possible evolutionary dependencies between the species. In addition, both Phylo-HMRF and Gaussian-HMRF show advantage over GMM and *K*-means clustering, suggesting that encoding the spatial constraints is also crucial. Moreover, Phylo-HMRF outperforms the two image segmentation methods SLIC and Quick Shift in different simulated datasets. The image segmentation methods perform segmentation of the image representation of the cross-species data based on feature similarity and spatial proximity. Regions that belong to the same evolutionary pattern (e.g., conserved high in Hi-C contacts across species) can be assigned different labels if they are distant from each other in spatial location, which affects the accuracy and interpretability of hidden state estimation.

We further performed simulation study II, where we simulated another eight datasets with different parameter configurations, in order to assess if the advantage of Phylo-HMRF is consistent over varied

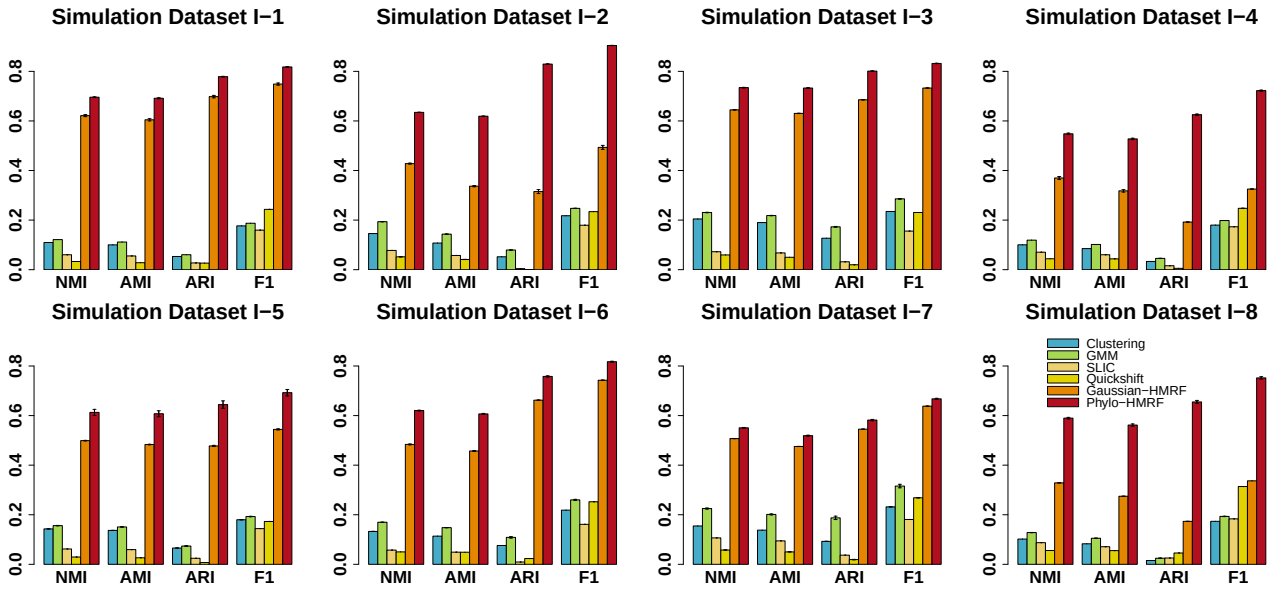

Figure 2: Performance evaluation of K-means Clustering, GMM, SLIC, Quick Shift, Gaussian-HMRF, and Phylo-HMRF on eight simulation datasets in simulation study I with respect to NMI (Normalized Mutual Information), AMI (Adjusted Mutual Information), ARI (Adjusted Rand Index), and F_1 score. The standard error of the results from 10 runs of each method is shown as the error bar, respectively.

simulation parameter settings (Supplementary Methods). The eight sets of simulated hidden states are shared between simulation studies I and II, while the observable random fields are simulated with different parameter settings. We then applied Phylo-HMRF and the other methods to the datasets in simulation study II and evaluated performance using the same procedure as we used in simutation study I. The evaluation results are shown in Fig. S1. Again we found that Phylo-HMRF consistently outperforms the other methods across all datasets in simulation study II. Taken together, the simulation evaluation demonstrated that Phylo-HMRF is able to achieve improved accuracy consistently in estimating evolutionary patterns of Hi-C contacts in multiple species.

Phylo-HMRF identifies different Hi-C contact patterns across multiple primate species

We applied Phylo-HMRF to a Hi-C dataset from four primate species. We used the Hi-C data in GM12878 in human from the 4DN data portal. We generated Hi-C data from the lymphoblastoid cells of three other primate species, including chimpanzee, bonobo, and gorilla. There are 290M, 270M, 240M, and 290M mapped read pairs for the four primate species, respectively. The genome assemblies used for the four species are hg38, panTro5, panPan2, and gorGor4, respectively.

We ran Phylo-HMRF on all the syntenic regions on autosomes based on the human genome. We first identified 92 synteny blocks in 50 kb resolution (i.e., ignoring rearrangements smaller than 50 kb among the four species) using the method inferCARs (Ma et al., 2006), covering 92.64% of the sequenced regions in the human genome (Fig. S2, Supplementary Methods). For example, we identified 9 major synteny blocks on human chromosome 1 among the four species, covering 92.50% of human chromosome 1 (Fig. 3B). We then applied Phylo-HMRF to perform genome-wide evolutionary pattern estimation over the multiple synteny blocks across different chromosomes jointly, with each synteny block represented as a subgraph of $\mathcal{G}$. The number of states is set to be 30 based on estimation from the results of K-means clustering (Supplementary Methods, Fig. S3).

Phylo-HMRF identified both conserved and lineage-specific evolutionary patterns of Hi-C contact frequencies across the four primate species. We further categorized the 30 estimated hidden states into 13 groups which show higher-level distinctiveness of heterogeneous evolutionary patterns (Supplementary Methods). Four of the groups represent conserved or weakly conserved cross-species patterns in Hi-C

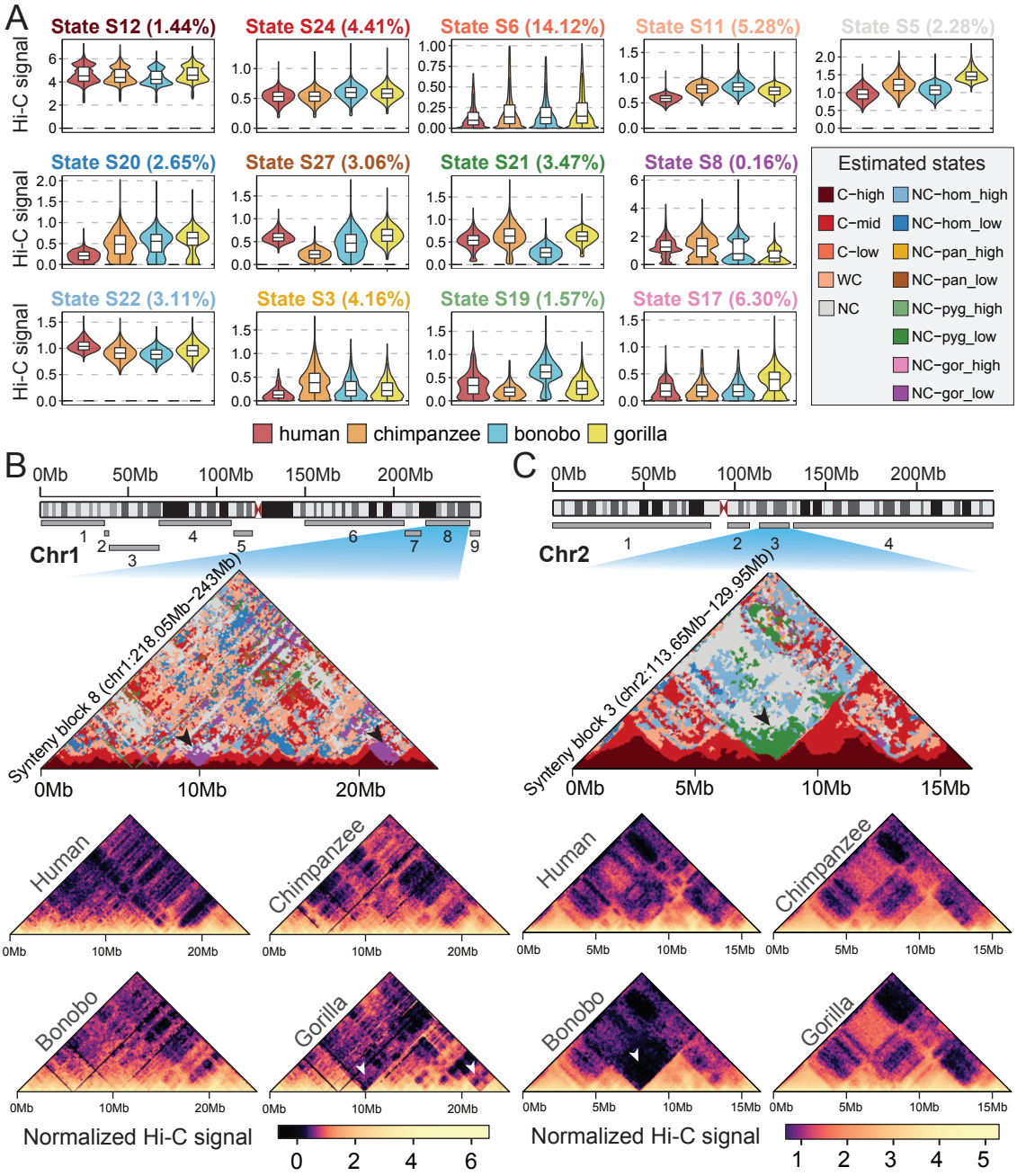

Figure 3: Evolutionary patterns of Hi-C contact frequency estimated by Phylo-HMRF. **(A)** Representative states from the 13 groups of evolutionary patterns. One state from each group is presented. The boxplots show the normalized cross-species Hi-C contact frequency distributions of the four species in the corresponding states with outliers removed. **(B)** Cross-species Hi-C contact frequency states identified in synteny block 8 on chromosome 1, in comparison with the Hi-C contact maps of the four primate species. First row: Locations of the nine identified synteny blocks on chromosome 1. Second row: Cross-species Hi-C contact frequency states identified by Phylo-HMRF. The black arrows point to two examples of identified gorilla-specific low Hi-C contact frequency state (NC-gor_low, purple color) in the combined Hi-C contact map. Third and fourth rows: Hi-C contact maps of the four primate species in this synteny block, with signal scale displayed at the bottom. Darker color in the Hi-C contact map represents lower contact frequency. The white arrows point to the corresponding locations of the two examples of identified gorilla-specific low contact state. **(C)** Cross-species Hi-C contact frequency states identified in synteny block 3 on chromosome 2. First row: Locations of the four identified synteny blocks on chromosome 2. Second row: Cross-species Hi-C contact frequency states identified by Phylo-HMRF. The black arrow points to one example of identified bonobo-specific low Hi-C contact state (NC-pyg_low, green color) in the combined Hi-C contact map. The white arrow points to the corresponding location of the example of identified bonobo-specific low state in the Hi-C contact maps of the four species.

contact frequency, which are conserved high in Hi-C contact frequency (C-high), conserved middle-level (C-mid), conserved low (C-low), and weakly conserved middle-level (WC). The four groups cover 51.14% of all the nodes in the cross-species Hi-C maps of the synteny blocks. Nine of the groups corre-

spond to non-conserved evolutionary patterns in Hi-C contacts, where eight exhibit lineage-specific patterns. Specifically, the nine groups are human-specific high in Hi-C contact frequency (NC-hom_high), human-specific low (NC-hom_low), chimpanzee-specific high (NC-pan_high), chimpanzee-specific low (NC-pan_low), bonobo-specific high (NC-pyg_high), bonobo-specific low (NC-pyg_low), gorilla-specific high (NC-gor_high), gorilla-specific low (NC-gor_low) and non-conserved (NC). Examples of the representative estimated states from each of the 13 groups are shown in Fig. 3A. Hi-C contact frequency distributions of multiple species in all 30 estimated states are shown in Fig. S4.

In Fig. 3B and Fig. 3C, as examples we show the estimated states in synteny block 8 on chromosome 1 and in synteny block 3 on chromosome 2, along with the input Hi-C contact maps of the four species. The rotated upper triangular matrix as shown in the second row of Fig. 3B or Fig. 3C represents the estimated hidden states of the graph $\mathcal{G}$ of the HMRF in Hi-C data comparison in the corresponding synteny block, which is of the same size as the multi-species Hi-C contact maps in this synteny block. Each vertex in $\mathcal{G}$ corresponds to a pair of genomic loci in the Hi-C contact map. The hidden state configuration is visualized as an image, where different colors represent different estimated hidden states. Adjacent nodes that are assigned to the same hidden state form a contiguous segment in the image. Therefore, based on the estimated states, $\mathcal{G}$ is partitioned, reflecting different cross-species Hi-C contact patterns. In Fig. 3B, we found that there are two gorilla-specific low Hi-C contact patterns near the diagonal area, which are colored in purple. These two regions correspond to the gorilla-specific low Hi-C states that appear in the state-distance plots of synteny block 8 on chromosome 1 in Fig. S5. In Fig. 3C, we observed that there is a bonobo-specific low Hi-C contact frequency pattern detected near the diagonal area, which is colored in green. By comparing the estimated hidden states to the corresponding Hi-C contact maps of the four species, we found that the estimated states accurately reflect what can be observed in Hi-C contact maps in different species.

Next we compared the distributions of evolutionary patterns of Hi-C contacts over changing distances between a pair of genomic loci in each synteny block. We consider that Hi-C contacts over short genomic loci distances are local Hi-C contacts (genomic loci distance $<$3Mb), and Hi-C contacts over large distances represent longer-range contacts. Short genomic loci distances correspond to an area near the diagonal of the Hi-C contact map. The state-distance plots across the synteny blocks on all the autosomes are shown in Fig. 4A. We observed that C-high, C-mid, and WC states are the predominant states around the diagonal area when the genomic loci distance is within the range of around 2.5 Mb (black arrow in Fig. 4A). This suggests that the majority of the local Hi-C contact patterns and the associated genome structures are likely to be conserved across difference species. At large genomic loci distances, which corresponds to the off-diagonal area in the Hi-C contact map, the C-low and non-conserved states have much larger percentages as expected. We also found that the distribution over genomic loci distance varies across different lineage-specific states. For single synteny blocks, the state-distance plots of the major synteny blocks on chromosome 1 and chromosome 2 are shown in Fig. S5 and Fig. S6 as examples. Overall, we found that there are similar enrichment patterns of states within a short distance range across the synteny blocks, while different blocks also exhibit varied trends of how evolutionary patterns are distributed at different genomic loci distances. We also observed that the lineage-specific states are distributed unevenly among the synteny blocks, showing occurrences either in local Hi-C contacts or long-range Hi-C contacts.

Together these results demonstrate the effectiveness of Phylo-HMRF to identify evolutionary patterns of Hi-C contacts across different species in a phylogeny.

Hi-C evolutionary patterns correlate with DNA replication timing patterns

We next compared the predicted states from Phylo-HMRF with other features of genome structure and function. We previously reported evolutionary patterns of DNA replication timing (RT) using Repli-seq

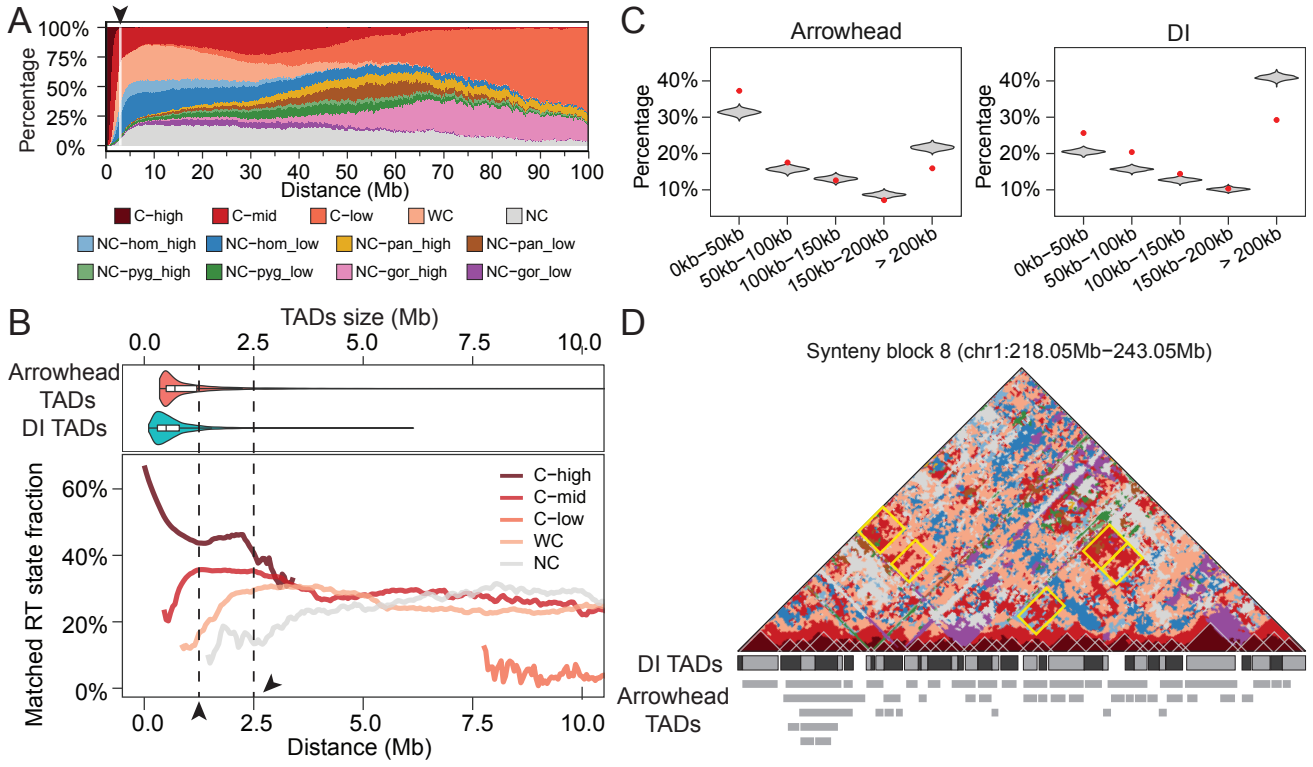

Figure 4: Comparison between the evolutionary states of Hi-C contacts estimated by Phylo-HMRF and other features of genome structure and function. **(A)** Global state-distance plots show the enrichment of different evolutionary patterns at different distances between genomic loci across synteny blocks on all autosomes. The black arrow points to the distance range around 2.5Mb where C-high and C-mid are the predominant states. **(B)** The percentage of paired genomic loci that is conserved in RT in five Hi-C contact evolutionary pattern groups. The top panel shows the distributions of TAD sizes, aligned with the axis of the genomic loci distance. The first black arrow points to the position of the first observed changing point on the Matched-RT state fraction curve for the C-high state and for the C-mid state. The second black arrow points to the position of 2.5MB, where the trend change is observed in the state-distance plot as shown in **(A)**. **(C)** Distributions of distance between the boundaries of all the identified local conserved high Hi-C contact patterns and the nearest TAD boundaries for both Arrowhead TADs and DI TADs. **(D)** Comparison between the boundaries of local conserved high Hi-C contact patterns and the TAD boundaries in synteny block 8 on chromosome 1. Several potential long-range conserved TAD interaction patterns are shown with yellow solid lines as borders.

of multiple primate species (Yang et al., 2018). We identified 5 groups of RT evolutionary states, which are conserved early in RT (E), weakly conserved early (WE), conserved late (L), weakly conserved late (WL), and non-conserved (NC). For each pair of genomic loci with estimated Hi-C states, we examined the RT state composition of the corresponding two genomic loci. If the paired genomic loci share similar conserved RT states, i.e., both are E/WE or both are L/WL, we annotated this pair as conserved in RT (C), otherwise we annotated it as non-conserved in RT. We then computed the percentage of contact loci that have the conserved RT states in the C-high, C-mid, WC, C-low, and NC Hi-C states identified by Phylo-HMRF over a range of different distances (0-10Mb). The results are shown in Fig. 4B. Notably, it is clear that C-high, C-mid, and WC Hi-C contacts patterns have higher enrichment of genome contacts with conserved RT patterns than the NC group over most of the distance ranges. The percentage is particularly high in the C-high state for genomic loci that are less than 4 Mb apart. This suggests that the conserved high Hi-C contact states are strongly correlated with those genomic loci pairs that both have consistent conserved RT patterns across species. We further explored potential connections between the features and the curves observed in Fig. 4B with known chromatin structure patterns such as TADs. We considered the TADs in GM12878 in human called using the Arrowhead method (Rao et al., 2014) and the Directionality Index (DI) method (Dixon et al., 2012), which are named Arrowhead TAD and DI TAD, respectively. The average sizes of Arrowhead TADs and DI TADs are around 1Mb and 0.6Mb, respectively. As shown in Fig. 4B, there is a changing point around 1MB in the distance axis on the

Matched-RT state fraction curve both for the C-high state and for the C-mid state, which approximately matches the average size of TADs. This implies that the identified evolutionary patterns of local high Hi-C contacts and the evolutionary patterns of RT states may be constrained by the TADs structures.

Hi-C evolutionary patterns correlate with TADs

TADs are important higher-order genome organization features revealed by Hi-C (Dixon et al., 2012). Here we focus on the diagonal of the Hi-C contact maps that reflect local Hi-C contact patterns across species. For segments of estimated Hi-C states on the diagonal, we used windows that could match the segments to detect the block patterns (Supplementary Methods). We identified 2,793 block patterns on the diagonals of the Hi-C contact maps of all the synteny blocks on all chromosomes.

We compared the boundaries of the diagonal blocks detected from the states predicted by Phylo-HMRF with TADs boundaries called using Arrowhead and DI, respectively. For each boundary of every identified diagonal block, we calculated the distance between the block boundary and the nearest TAD boundary. We then computed the percentages of the diagonal blocks whose distance to the nearest TAD boundaries fall in different distance ranges. Since the Hi-C contact frequency was measured at a resolution of 50Kb in this study, the boundaries of the diagonal blocks detected from the estimated states are all at 50Kb resolution. The distance between a diagonal block boundary and a nearest TAD boundary is calculated in increments of 50Kb (one genome bin) accordingly. We also estimated the empirical distributions of the distances between boundaries of a possible diagonal block and a nearest TAD by randomly shuffling the identified diagonal blocks (Supplementary Methods).

We observed that the distance between the boundaries of the identified diagonal blocks and the nearest TADs are significantly more enriched in the distance intervals that represent relatively small distances (e.g., [0,50Kb] and (50Kb,100Kb]) than expected (Fig. 4C). Specifically, 55.60% of the identified diagonal block boundaries are matched by an Arrowhead TAD boundary with the distance less than 2 bins, which is significantly higher than the corresponding expected percentage 36.08% observed from the empirical distribution (empirical p-value$<$2e-03). Similarly, the percentages of the identified diagonal block boundaries that are matched by a DI TAD boundary within 1 bin or 2 bins are significantly higher than the expected percentages (empirical p-value$<$1e-03). In contrast, the percentage of the diagonal block boundaries with distance to a nearest TAD boundary larger than 4 bins are significantly smaller than expected (empirical p-value$<$1e-03).

For example, we identified 34 blocks along the diagonal of the Hi-C map of synteny block 8 on chromosome 1 from the estimated Hi-C evolutionary states based on Phylo-HMRF. The identified blocks are all C-high states, which are shown with white borders in Fig. 4D. We observed that the block boundaries show high consistency with the TAD boundaries (Fig. 4D). Specifically, 70.77% of the block boundaries match an Arrowhead TAD boundary or a DI TAD boundary within 2 bins. The capability of Phylo-HMRF in detecting TAD boundaries without using a TAD calling algorithm implies that TADs are an important type of units of genome organization evolution. The result also reflects the accuracy of Phylo-HMRF in estimating Hi-C evolutionary patterns.

Furthermore, we observed C-mid and WC states in off-diagonal area of the Hi-C contact map, which potentially correspond to the long-range interactions between two TADs that are conserved across species. Five examples of the potentially conserved long-range TADs interactions are shown in Fig. 4D (highlighted in yellow borders).

Taken together, our results suggest that TADs are important 3D genome organization features in genome evolution in primate species. In addition, the evolutionary changes of intra-TAD interactions (i.e., local contacts) and inter-TAD interactions (i.e., long-range contacts) can be uncovered effectively by Phylo-HMRF.

Discussion

In this work, we developed Phylo-HMRF, a continuous-trait probabilistic model that provides a new framework to utilize spatial dependencies among genomic loci in 3D space to identify evolutionary patterns of Hi-C contacts across different species in a phylogeny. We applied Phylo-HMRF to analyze Hi-C data from the lymphoblastoid cells in four primate species (human, chimpanzee, bonobo, and gorilla). Phylo-HMRF is able to identify different genome-wide cross-species Hi-C contact patterns, including conserved and lineage-specific patterns in both local interactions and long-range interactions. The identified evolutionary patterns of 3D genome structure have strong correlation with other types of genomic structural and functional features such as TADs and DNA replication timing. From a methodology standpoint, Phylo-HMRF is a flexible framework that can also be applied to other types of multi-species continuous-trait features where there are 2D or 3D spatial dependencies for the features among the genomic loci. Overall, through a proof-of-principle application, we demonstrate that Phylo-HMRF is a promising new method to effectively uncover detailed evolutionary patterns of 3D genome organization based on Hi-C across multiple species.

There are several aspects where our method can be improved. First, model selection methods such as the utilization of the AIC and BIC criterion (Akaike, 1973; Schwarz et al., 1978) may help select the number of states more automatically. Second, Phylo-HMRF has only been applied to synteny blocks across species at the moment and does not explicitly model the chromatin conformation differences due to large-scale genome rearrangements in evolution. It will be an important next step to model genome rearrangements and genome organization evolution in an integrative manner. Third, to study a larger number of more distantly related species, we may face several challenges. As the number of model parameters and feature dimensions increase linearly with the tree size, both the computation demand increases and the model is exposed to a higher possibility of local minima and overfitting especially for a smaller sample size of multi-species feature observations. There will also be more potential misalignments among genomes of distantly related species, resulting in fewer available samples. It will be useful to incorporate more efficient parameter regularization which is compatible with the evolutionary models in the optimization part of Phylo-HMRF, and to develop imputation methods for the missing observations in multi-species genomic data, especially in large-scale phylogenetic trees. Fourth, we assume that all the phylogenetic trees associated with different hidden states share the same topology in our current Phylo-HMRF model. Incorporating inference of varied tree topologies (Friedman et al., 2002) will make Phylo-HMRF much more general.

To fully understand 3D genome organization evolution, it will be crucial to explore the underlying mechanisms of the different evolutionary patterns in 3D genome structure across species, e.g., in concert with the evolution of particular types of DNA sequence features that play key roles in the formation and maintenance of genome architecture and function (Sima et al., 2018; Choudhary et al., 2018; Zhang et al., 2019), which may in turn inform us about the principles of 3D genome organization. For example, we previously showed that more conserved CTCF motifs in mammalian evolution (considering motif turnover) are more likely to be involved in CTCF mediated chromatin loops (Zhang et al., 2018). Our Phylo-HMRF model has the potential to serve as a generic analytic framework to connect different evolutionary patterns of chromatin interaction with the evolution of genome sequence and function.

Methods

Overall framework of Phylo-HMRF for cross-species comparison of Hi-C data

We assume that a two-dimensional Hi-C contact map is given in each species, where each entry of the map represents the contact frequency between the corresponding two genomic loci. We use the human genome as the reference and align the contact pairs of genomic loci of the other species to the human genome. As a consequence, Hi-C contact maps of the other species are equivalently aligned to the human genome to be comparable. In this study, we compare the multi-species Hi-C contact frequencies in the syntenic regions genome-wide (synteny blocks were identified based on inferCARs (Ma et al., 2006)), in order to focus on the 3D genome changes that are not resulted from large-scale genome rearrangements. We then obtain a multi-species contact map $\mathbf{I} \in \mathbb{R}^{n\times n\times d}$, where n is the number of loci in the studied region on the reference genome, and d is the number of the species. $\mathbf{I}$ can be represented by a graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$, where $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\mathcal{E}}$ represent the set of nodes and the set of edges, respectively. Each node corresponds to a position in $\mathbf{I}$, i.e., the contact between a pair of genomic loci. The number of nodes is $N = n \times n$. We also denote $\boldsymbol{\mathcal{V}}$ as the set of indices of the nodes in $\boldsymbol{\mathcal{G}}$, i.e., $\boldsymbol{\mathcal{V}} = \{1, \cdots, N\}$. The i-th node is associated with a random variable $X_i \in \mathbb{R}^d$ representing the multi-species observations on this node, $i \in \boldsymbol{\mathcal{V}}$. The k-th element of X_i ($k = 1, \cdots, d$) is the aligned contact frequency measurement of the k-th species between the corresponding two genomic loci. If two positions in the multi-species contact map are adjacent, there is an edge between the corresponding nodes in $\boldsymbol{\mathcal{G}}$.

Using an HMRF model, we assume that each node in $\boldsymbol{\mathcal{G}}$ is also associated with a random variable $Y_i \in S = \{1, \cdots, M\}$, representing the unknown hidden state of this node, $i \in \boldsymbol{\mathcal{V}}$. S is the set of hidden states. For each configuration of Y, X_i follows a conditional probability distribution $p(x_i|y_i)$, which is the emission probability distribution, and $X = \{X_i\}_{i\in\boldsymbol{\mathcal{V}}}$ is the observable random field or emitted random field. The hidden state Y_i reflects different evolutionary patterns of chromatin contact frequency across species, e.g., some regions in $\mathbf{I}$ may exhibit conserved high (or low) contact frequency across species, while some may have lineage-specific high or low contact frequency. The spatial information is embedded in the MRF with the constraints on the hidden states of neighboring nodes. The neighboring nodes are expected to be more likely to have similar hidden states.

Phylo-HMRF estimates the evolutionary patterns by inferring the hidden states $\mathbf{y} = \{y_i\}_{i\in\boldsymbol{\mathcal{V}}}$ from the observations $\mathbf{x} = \{x_i\}_{i\in\boldsymbol{\mathcal{V}}}$, using the assumption that there are spatial dependencies between adjacent nodes in the graph $\boldsymbol{\mathcal{G}}$. In Phylo-HMRF, each hidden state $Y_i = l$ is associated with a phylogenetic model ψ_l. We therefore define the Phylo-HMRF model as $\mathbf{h} = (S, \psi, \beta)$, where S is the set of states, ψ is the set of phylogenetic models associated with the states, and β contains the pairwise potential parameters, respectively. Suppose $X^{(l)} = (X_1^{(l)}, \cdots, X_d^{(l)})$ represent the values of leaf nodes of the phylogenetic tree associated with the l-th phylogenetic model ψ_l, $l = 1, \cdots, M$. The emission probability of each state is $p(X|\psi_l)$, which is determined by the phylogenetic model underlying this state. The joint probability of the graph $\boldsymbol{\mathcal{G}}$ is:

$$p(\mathbf{y}, \mathbf{x}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = \frac{1}{Z}\prod_{i\in\boldsymbol{\mathcal{V}}} p(x_i|y_i) \prod_{(i,j)\in\boldsymbol{\mathcal{E}}} f(y_i, y_j; x_i, x_j) \tag{1}$$

where Z is the normalization constant, $p(x_i|y_i)$ is the emission probability function, which measures the probability that the local observation is generated from a certain hidden state, and $f(\cdot)$ is the compatibility function which measures the consistency of hidden states between the neighboring nodes. The joint probability can be transformed into the energy function by taking the negative logarithm of the joint

probability. In this work, given the observations across species, the energy function can be defined as:

$$E(\mathbf{y}|\mathbf{x},\Theta) = \sum_{i\in\boldsymbol{\mathcal{V}}} U(x_i, y_i, \Theta) + \sum_{ij\in\boldsymbol{\mathcal{E}}} V(y_i, y_j; x_i, x_j), \tag{2}$$

where $U(x_i, y_i)$ is the unary potential function encoding local compatibility between observations and hidden states with model parameters Θ, and $V(y_i, y_j)$ is the pairwise potential function encoding neighborhood information, respectively. We have that $f(y_i, y_j; x_i, x_j) \propto \exp\left(-V(y_i, y_j; x_i, x_j)\right)$. If we take into consideration the effect of the difference between features of neighboring nodes on the pairwise potential, $V(y_i, y_j; x_i, x_j)$ and the compatibility function $f(\cdot)$ will depend not only on the labels of the neighboring nodes, but also on their features or observations. We minimize the energy function to estimate the hidden states $\mathbf{y}$. By minimizing the energy function we maximize the joint probability of the graph equivalently. As the model parameters Θ are unknown, we estimate $\mathbf{y}$ and Θ simultaneously. The objective function is:

$$\{\mathbf{y}^*, \Theta^*\} = \arg\min_{\mathbf{y},\Theta} E(\mathbf{y}|\mathbf{x},\Theta). \tag{3}$$

Phylo-HMRF model with Ornstein-Uhlenbeck process

Ornstein-Uhlenbeck process assumptions

In Phylo-HMRF, we model the continuous traits with the Ornstein-Uhlenbeck (OU) process. The OU process is a stochastic Gaussian process that extends the Brownian motion (Felsenstein, 1985; Pagel, 1999; Freckleton, 2012) with the trend towards equilibrium around optimal values (Hansen, 1997; Butler and King, 2004; Hansen et al., 2008). The OU process has been recently used to model the evolution of genomic features (Rohlfs et al., 2013; Brawand et al., 2011; Naval-Sánchez et al., 2015; Chen et al., 2019; Yang et al., 2018). In our previous work (Yang et al., 2018), we found that the OU process has clear advantages in performance as compared to the simpler Brownian motion model. Therefore, we utilize the OU processes to realize the phylogenetic models in Phylo-HMRF.

For the observation of a lineage $\hat{X}_i$, the OU process can be represented as the following (Hansen, 1997; Butler and King, 2004):

$$d(\hat{X}_i(t)) = \alpha[\theta_i(t) - \hat{X}_i(t)]dt + \sigma dB_i(t), \tag{4}$$

where $\hat{X}_i(t)$ represents the observation of $\hat{X}_i$ at time point t, $B_i(t)$ represents the Brownian motion, and α, θ_i and σ are parameters that represent the selection strength, the optimal value and the fluctuation intensity of Brownian motion, respectively.

For multi-species observations, based on the model assumptions of the OU process, the expectation, variance, and covariance of the observations of species can be computed given the phylogenetic tree. Suppose that X_p is the ancestor of species X_i, and X_a is the common ancestor of species X_i and X_j. We have (Butler and King, 2004; Rohlfs et al., 2013):

$$\mathbb{E}(X_i) = \mathbb{E}(X_p)e^{-\alpha_i t_{ip}} + \theta_i(1 - e^{-\alpha_i t_{ip}}), \tag{5}$$

$$\mathrm{Cov}(X_i,\ X_j) = \mathrm{Var}(X_a)\exp\left(-\Sigma_{k\in l_{ij}}\alpha_k t_k - \Sigma_{k\in l_{ji}}\alpha_k t_k\right), \tag{6}$$

$$\mathrm{Var}(X_i) = \frac{\sigma_i^2}{2\alpha_i}(1 - e^{-2\alpha_i t_{ip}}) + \mathrm{Var}(X_p)e^{-2\alpha_i t_{ip}}. \tag{7}$$

where t_{ip}, t_k represent evolution time along corresponding branches in the phylogenetic tree, respectively. In the Phylo-HMRF model $\mathbf{h} = (S, \psi, \beta)$, ψ_l is defined as $\psi_l = (\theta_l, \alpha_l, \sigma_l, \tau_l, b_l)$, $1 \le l \le M$, where

θ_l, α_l, σ_l denote the optimal values, the selection strengths, and the Brownian motion intensities of the corresponding OU model, respectively, and τ_l, b_l represent the topology of the phylogenetic tree, the branch lengths, respectively. M is the number of states. We assume that the phylogenetic tree topology is identical across different states. For the phylogenetic tree of a hidden state, we allow varied selection strengths and Brownian motion intensities along different branches and varied optimal values at the interior nodes or leaf nodes. Thus each branch is assigned a selection strength and a Brownian motion intensity, and each node is assigned an optimal value as parameters. Suppose there are r branches in the tree. We have $\theta_l \in \mathbb{R}^{r+1}$, $\alpha_l, \sigma_l \in \mathbb{R}_+^r$, where values in α_l and σ_l are non-negative. According to the actual problem studied, ψ_l can be specialized to different evolutionary models. We focus on the OU processes in this work.

Model parameter estimation and hidden states inference

We use the Expectation-Maximization (EM) algorithm (Dempster et al., 1977; Zhang et al., 2001) for parameter estimation in our model. Zhang et al. (2001) developed the HMRF-EM algorithm where EM is adapted to estimate a HMRF model with several justified assumptions and approximations, including the pseudo-likelihood assumption (Geman and Graffigne, 1986) and mean-field approximation (Celeux et al., 2003; Zhang, 1992). The original HMRF-EM algorithm uses the multivariate Gaussian distribution as the emission probability function of a hidden state. The main difference is that in our method we use the OU processes to model the emission probability in the HMRF. Also, we utilize the Graph Cuts algorithm (Adelson and Bergen, 1991; Boykov et al., 2001) for hidden state estimation given estimates of model parameters.

Let Θ be the model parameters. Suppose Θ^g is the current estimate of model parameters. The EM algorithm aims to maximize the expectation of the complete-data log likelihood, which is defined as the Q function: $Q(\Theta, \Theta^g) = \mathbb{E}[\log p(\mathbf{x}, \mathbf{y}|\Theta)|\mathbf{x}, \Theta^g]$. Using pseudo-likelihood approximation (Geman and Graffigne, 1986) and mean-field approximation (Celeux et al., 2003; Zhang, 1992), the Q function is derived as (details in Supplementary Methods):

$$Q(\Theta, \Theta^g) = \sum_{l=1}^{M} \sum_{i \in \boldsymbol{\mathcal{V}}} p(y_i = l|x_i, \Theta^g) \log p(x_i|y_i = l, \Theta) + \sum_{i \in \boldsymbol{\mathcal{V}}} \sum_{l=1}^{M} p(y_i = l|x_i, \Theta^g) \log p(y_i = l|y^g_{\boldsymbol{\mathcal{N}}_i}, \Theta), \tag{8}$$

where the two parts of the Q function encode the unary potential and the pairwise potential of the HMRF of $\boldsymbol{\mathcal{G}}$, respectively. $p(y_i = l|x_i, \Theta^g)$ is posterior probability of each sample assigned to a hidden state given the current model parameter estimates. Using the Markov property of HMRF (Zhang et al., 2001), we have:

$$p(y_i = l|x_i, \Theta^g) = \frac{p(x_i|y_i = l, \Theta^g) p(y_i = l|y^g_{\boldsymbol{\mathcal{N}}_i})}{\sum_{l=1}^{M} p(x_i|y_i = l, \Theta^g) p(y_i = l|y^g_{\boldsymbol{\mathcal{N}}_i})}, \tag{9}$$

where $\boldsymbol{\mathcal{N}}_i$ denotes the set of nodes that are neighbors of node i in $\boldsymbol{\mathcal{G}}$. We calculate $p(x_i|y_i, \Theta^g)$ based on the OU process assumption (Supplementary Methods).

Let $V(y_i, y_j)$ be the pairwise potential on a pair of adjacent nodes (y_i, y_j). We have:

$$p(y_i = l|y^g_{\boldsymbol{\mathcal{N}}_i}) = \frac{1}{Z} \exp\left(- \sum_{j \in \boldsymbol{\mathcal{N}}_i} V(l, y^g_j) \right), \tag{10}$$

where Z is the normalization constant. We can adopt different definitions of the pairwise potential $V(y_i, y_j)$ (Supplementary Methods). The definition we use takes into consideration the difference be-

tween features of the adjacent vertices in imposing the penalty on inconsistent states of the neighbors:

$$V(y_i, y_j) = \beta_0 I(y_i \neq y_j) \exp\left(-\beta_1 \frac{\|x_i - x_j\|_2^2}{\|x_i\|_2 \|x_j\|_2}\right), \tag{11}$$

where β_0, β_1 are predefined adjustable regularization coefficients. Based on the definition of the pairwise potential, $p(y_i = l | y^g_{\mathcal{N}_i})$ does not depend on the OU model parameters.

Let $L(\Theta^{(l)}) = -\sum_{i \in \boldsymbol{\mathcal{V}}} w_i^{(l)} \log p(x_i | y_i = l, \Theta)$, $w_i^{(l)} = p(y_i = l | x_i, \Theta^g)$. We perform parameter estimation for each of the possible states. In each Maximization-step (M-step), the objective function of a given state l is defined as (details in Supplementary Methods):

$$\min_{\Theta^{(l)}} \frac{1}{N} \log |\Sigma_\Theta^{(l)}| \sum_{i \in \boldsymbol{\mathcal{V}}} w_i^{(l)} + \operatorname{tr}\left([\Sigma_\Theta^{(l)}]^{-1} \tilde{S}_\Theta^{(l)}\right) + \lambda \|\Theta^{(l)}\|_2^2, \tag{12}$$

where $\tilde{S}_\Theta^{(l)} = \frac{1}{N} \sum_{i \in \boldsymbol{\mathcal{V}}} w_i^{(l)} \left(x_i - \mu_\Theta^{(l)}\right)\left(x_i - \mu_\Theta^{(l)}\right)^T$, $w_i^{(l)}$ is defined as above, $\Theta^{(l)}$ represents the phylogenetic model parameters associated with state l, and λ is the regularization coefficient of the l_2-norm regularization that is used to reduce overfitting of model (Supplementary Methods). We have that $\Theta^{(l)} = \{\theta_l, \alpha_l, \sigma_l\}$, where θ_l, α_l, σ_l represent the optimal values, the selection strengths, and the Brownian motion intensities of the phylogenetic tree model associated with hidden state l, respectively. As described previously, the OU model of state l is $\psi_l = (\theta_l, \alpha_l, \sigma_l, \tau_l, b_l)$, $1 \leq l \leq M$. We assume that τ_l is given. If the branch lengths b_l are unknown, we perform the transformation that $\tilde{\alpha}_{l,v} = \alpha_{l,v}\beta_{l,v}$, $\tilde{\sigma}^2_{l,v} = \sigma^2_{l,v}\beta_{l,v}$ to present the combined effect of the branch length and the selection or Brownian motion parameters along this branch. Here $\beta_{l,v}$ represents the length of the branch from the parent of node v to node v in the phylogenetic tree of state l. Then $\Theta^{(l)} = \{\theta_l, \tilde{\alpha}_l, \tilde{\sigma}_l\}$, where $\tilde{\alpha}_l$, $\tilde{\sigma}_l$ are the transformed selection strengths and the transformed Brownian motion intensities, respectively.

In Phylo-HMRF, the overall steps of the OU-model embedded HMRF-EM algorithm is as follows.

1. *Initialize the model parameter*. We perform K-means clustering on the samples. The clustering results are used to assign initial hidden states to the samples. For each cluster, we estimate the OU model parameters using Maximum Likelihood Estimation (MLE) (Supplementary Methods). The estimated model parameters are used as initialization of the OU model parameters for each hidden state.
2. *Estimate the hidden states given current parameter estimates*. We seek an approximate solution to the optimization problem

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} \{U(\mathbf{x}|\mathbf{y}) + U(\mathbf{y})\}, \tag{13}$$

   where $U(\mathbf{x}|\mathbf{y})$ and $U(\mathbf{y})$ are the total unary potential and total pairwise potential of $\boldsymbol{\mathcal{G}}$, respectively.
3. *Calculate posterior probability distribution*. In each Expectation-step (E-step), given the current estimated model parameters Θ^g and the estimated state configuration in the previous step, we compute $p(y_i = l | x_i, \Theta^g)$ using Eq. (9, 10).
4. *Estimate model parameters by solving the optimization problem with OU models embedded*. In each M-step, we solve the MLE problem in (12) to update the parameters $\{\psi_l\}_{l=1}^M$.
5. *Repeat step 2-4 until convergence is reached or the maximum number of iterations is reached*.

In Phylo-HMRF, given the current estimated model parameters, we use the Graph Cuts algorithm to estimate the hidden states in Step 2 (Supplementary Methods). Graph Cuts algorithms seek to approximate solutions to an energy minimization problem by solving a max-flow/min-cut problem in a

graph (Boykov et al., 2001; Boykov and Kolmogorov, 2004). Graph Cuts algorithms have been effectively used in image segmentation applications. Studies have shown that for binary image segmentation, finding a min-cut is equivalent to finding the maximum of posterior $p(\mathbf{y}|\mathbf{x})$ (Boykov et al., 2001). For multiple labels, the multi-labeling problem can be converted to a sequence of binary-labeling problem by α-expansion or α-β swap algorithms (Boykov et al., 2001). The solution is an approximate solution in the multi-labeling problem that has been shown to be a strongly probable local minima (Boykov et al., 2001).

Acknowledgement

This work was supported in part by National Institutes of Health grant R01HG007352 (J.M.), National Institutes of Health Common Fund 4D Nucleome Program grants U54DK107965 (J.M.) and U54DK107977 (B.R.), National Institutes of Health Director's Early Independence Award DP5OD023071 (J.D.), and National Science Foundation grant 1717205 (J.M.). The authors would like to thank members of Jian Ma's laboratory for helpful comments to improve the manuscript.

Author Contributions

Conceptualization, J.M.; Methodology, Y.Y., J.M.; Software, Y.Y.; Resources, B.R., J.D.; Visualization, Y.Z.; Investigation, Y.Y., Y.Z., B.R., J.D., and J.M.; Writing – Original Draft, Y.Y., and J.M.; Writing – Review & Editing, Y.Y., Y.Z., B.R., J.D., and J.M.; Funding Acquisition, B.R., J.D., J.M.

Declaration of Interests

The authors declare no competing interests.

References

R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34 (11):2274–2282, 2012.

E. H. Adelson and J. R. Bergen. The plenoptic function and the elements of early vision. *Computational Models of Visual Processing*, pages 3–20, 1991.

H. Akaike. Information theory and an extension of the maximum likelihood principle. pages 267–281. Proceeding of Second International Symposium on Information Theory, 1973.

B. Bonev and G. Cavalli. Organization and function of the 3d genome. *Nature Reviews Genetics*, 17 (11):661, 2016.

Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9): 1124–1137, 2004.

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

D. Brawand, M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, et al. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, 2011.

M. A. Butler and A. A. King. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist*, 164(6):683–695, 2004.

G. Celeux, F. Forbes, and N. Peyrard. Em procedures using mean field-like approximations for markov model-based image segmentation. *Pattern Recognition*, 36(1):131–144, 2003.

J. Chen, R. Swofford, J. Johnson, B. B. Cummings, N. Rogel, K. Lindblad-Toh, W. Haerty, F. Di Palma, and A. Regev. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Research*, 29(1):53–63, 2019.

M. N. Choudhary, R. Z. Friedman, J. T. Wang, H. S. Jang, X. Zhuo, and T. Wang. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *bioRxiv*, page 485342, 2018.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (methodological)*, pages 1–38, 1977.

J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398): 376–380, 2012.

J. Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15, 1985.

R. P. Freckleton. Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution*, 3(5):940–947, 2012.

N. Friedman, M. Ninio, I. Pe'er, and T. Pupko. A structural em algorithm for phylogenetic inference. *Journal of Computational Biology*, 9(2):331–353, 2002.

G. Fudenberg and K. S. Pollard. Chromatin features constrain structural variation across evolutionary timescales. *Proceedings of the National Academy of Sciences*, 116(6):2175–2180, 2019.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

S. Geman and C. Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, volume 1, page 2. Berkeley, CA, 1986.

T. F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, pages 1341–

1351, 1997.

T. F. Hansen, J. Pienaar, and S. H. Orzack. A comparative method for studying adaptation to a randomly evolving environment. *Evolution*, 62(8):1965–1977, 2008.

N. H. Lazar, K. A. Nevonen, B. O'Connell, C. McCann, R. J. O'Neill, R. E. Green, T. J. Meyer, M. Okhovat, and L. Carbone. Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Research*, 2018.

E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

J. Ma, L. Zhang, B. B. Suh, B. J. Raney, R. C. Burhans, W. J. Kent, M. Blanchette, D. Haussler, and W. Miller. Reconstructing contiguous regions of an ancestral genome. *Genome Research*, 16(11): 000–000, 2006.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

M. Naval-Sánchez, D. Potier, G. Hulselmans, V. Christiaens, and S. Aerts. Identification of lineage-specific cis-regulatory modules associated with variation in transcription factor binding and chromatin activity using ornstein–uhlenbeck models. *Molecular Biology and Evolution*, 32(9):2441–2455, 2015.

E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381, 2012.

M. Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877, 1999.

S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

R. V. Rohlfs, P. Harrigan, and R. Nielsen. Modeling gene expression evolution with an extended ornstein–uhlenbeck process accounting for within-species variation. *Molecular Biology and Evolution*, 31(1):201–211, 2013.

M. J. Rowley and V. G. Corces. Organizational principles of 3d genome architecture. *Nature Reviews Genetics*, page 1, 2018.

M. V. Rudan, C. Barrington, S. Henderson, C. Ernst, D. T. Odom, A. Tanay, and S. Hadjur. Comparative hi-c reveals that ctcf underlies evolution of chromosomal domain architecture. *Cell Reports*, 10(8): 1297–1309, 2015.

G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

J. Sima, A. Chakraborty, V. Dileep, M. Michalski, K. N. Klein, N. P. Holcomb, J. L. Turner, M. T. Paulsen, J. C. Rivera-Mulia, C. Trevilla-Garcia, et al. Identifying cis elements for spatiotemporal control of mammalian dna replication. *Cell*, 2018.

Z. Tang, O. J. Luo, X. Li, M. Zheng, J. J. Zhu, P. Szalaj, P. Trzaskoma, A. Magalska, J. Wlodarczyk, B. Ruszczycki, et al. Ctcf-mediated human 3d genome architecture reveals chromatin topology for transcription. *Cell*, 163(7):1611–1627, 2015.

A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *European Conference on Computer Vision*, pages 705–718. Springer, 2008.

N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct): 2837–2854, 2010.

Y. Yang, Q. Gu, Y. Zhang, T. Sasaki, J. Crivello, R. J. O'Neill, D. M. Gilbert, and J. Ma. Continuous-trait

probabilistic model for comparing multi-species functional genomic data. *Cell Systems*, 7:208–218, 2018.

J. Zhang. The mean field theory in em procedures for markov random fields. *IEEE Transactions on signal processing*, 40(10):2570–2583, 1992.

R. Zhang, Y. Wang, Y. Yang, Y. Zhang, and J. Ma. Predicting ctcf-mediated chromatin loops using ctcf-mp. *Bioinformatics*, 34(13):i133–i141, 2018.

Y. Zhang, M. Brady, and S. Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20 (1):45–57, 2001.

Y. Zhang, T. Li, S. Preissl, J. Grinstein, E. Farah, E. Destici, A. Y. Lee, S. Chee, Y. Qiu, K. Ma, et al. 3d chromatin architecture remodeling during human cardiomyocyte differentiation reveals a role of herv-h in demarcating chromatin domains. *bioRxiv*, page 485961, 2019.