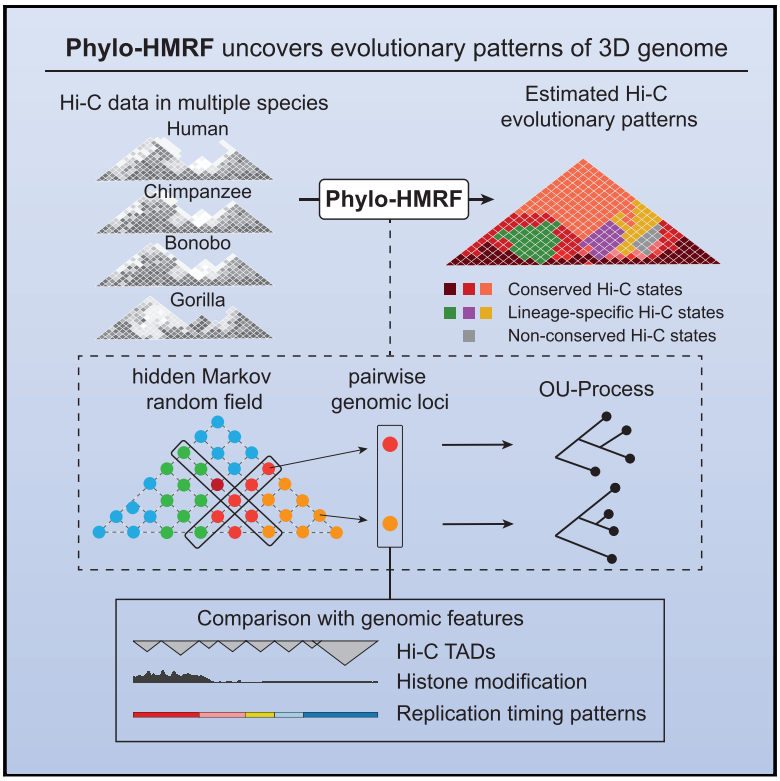
Report

Cell Systems

Comparing 3D Genome Organization in Multiple Species Using Phylo-HMRF

Graphical Abstract



Highlights

- Phylo-HMRF is a probabilistic model for comparing multi-species Hi-C data
- It uncovers evolutionary patterns of 3D genome organization
- We use Phylo-HMRF to analyze a Hi-C dataset from four primate species

Authors

Yang Yang, Yang Zhang, Bing Ren, Jesse R. Dixon, Jian Ma

Correspondence

jianma@cs.cmu.edu

In Brief

A new computational model called phylogenetic hidden Markov random field (Phylo-HMRF) allows us to reveal genome-wide evolutionary patterns of 3D genome organization based on multi-species Hi-C data.

Yang et al., 2019, Cell Systems *8*, 494–505
June 26, 2019
https://doi.org/10.1016/j.cels.2019.05.011

CellPress

OPEN ACCESS
CellPress

Cell Systems
Focus on RECOMB
Report

Comparing 3D Genome Organization in Multiple Species Using Phylo-HMRF

Yang Yang,[1] Yang Zhang,[1] Bing Ren,[2] Jesse R. Dixon,[3] and Jian Ma[1,4,*]
[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
[4]Lead Contact
*Correspondence: jianma@cs.cmu.edu
https://doi.org/10.1016/j.cels.2019.05.011

SUMMARY

Recent whole-genome mapping approaches for the chromatin interactome have offered new insights into 3D genome organization. However, our knowledge of the evolutionary patterns of 3D genome in mammals remains limited. In particular, there are no existing phylogenetic-model-based methods to analyze chromatin interactions as continuous features. Here, we develop phylogenetic hidden Markov random field (Phylo-HMRF) to identify evolutionary patterns of 3D genome based on multi-species Hi-C data by jointly utilizing spatial constraints among genomic loci and continuous-trait evolutionary models. 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 correlate with features of genome structure and function. This work provides a new framework to analyze multi-species continuous genomic features with spatial constraints and has the potential to help 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 (chromatin interaction analysis by paired-end tag sequencing) (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). Using relatively low-resolution Hi-C data from rhesus macaque, dog, rabbit, and mouse, another study demonstrated that evolutionary changes in the TAD structure correspond with the creation or elimination of the binding sites of CTCF (CCCTC binding-factor) (Vietri Rudan et al., 2015). More recently, TADs have been shown to have strong conservation in mammalian evolution, with the TAD 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 structure remains unclear. Additionally, as most of the initial comparative analysis of 3D genomes focused primarily on distantly related organisms, there is a 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, chromatin immunoprecipitation sequencing [ChIP-seq], and Repli-seq (replication timing by sequencing)) 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 a dramatic loss of information of more subtle differences across species 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; see Box 1) (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 spatial dependencies along

Box 1. Glossary

Markov random field (MRF): MRF is a set of random variables that have Markov property described by an undirected graph.
Markov property: A stochastic process has the Markov property if the conditional probability of future states of the process conditional on both the past and the present states depends only on the present state, not on the states that preceded the present state. In MRF, the Markov property is extended to random variables defined on a graph. For local Markov property, a variable is conditionally independent of all the other variables given its neighbors.
Hidden Markov random field (HMRF): The HMRF model is characterized by the hidden random field, the observable random field, and the conditional independence assumption. The hidden random field is a Markov random field with a state space.
Gaussian process: Gaussian process is a stochastic process, of which every finite collection of the random variables in the stochastic process has a multivariate Gaussian distribution, i.e., every finite linear combination of the random variables has Gaussian distribution.
Standard Brownian motion: A standard Brownian motion is a stochastic process $X = \{X_t : t \in [0, \infty)\}$ that has the following properties: (1) $X_0 = 0$ with probability 1; (2) The increment of X_t is independent of the past values; (3) The increment of X_t is normally distributed with mean 0 and the time interval of the increment as variance; (4) X has continuous paths.
Ornstein-Uhlenbeck process: The Ornstein-Uhlenbeck (OU) process is a Gaussian process that consists of two parts and models the Brownian motion under the effect of selection. The first part models the tendency toward the equilibrium around an optimal value. The second part is the Brownian motion. The parameters include the selection strength, the optimal value, and the Brownian motion intensity.
Expectation-Maximization (EM) algorithm: The EM algorithm uses iterative alternating Expectation-step (E-step) and Maximization-step (M-step) to find maximum likelihood or maximum a posteriori estimates of parameters of models with unobserved latent variables. The E-step computes the distribution of the latent variables and forms a function of the expectation of the log likelihood given current model parameter estimates. The M-step performs parameter estimation by maximizing the expectation of the log likelihood.

one-dimensional (1D) genome coordinates. However, the HMM, as used by Phylo-HMGP, is 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 each 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 (see Box 2). 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 cells) 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 other features of genome structure and function, such as TADs, A/B compartments, histone modifications, DNA replication timing (RT), and sequence properties. Phylo-HMRF offers an effective model to potentially help 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 Figure 1. Our goal is to identify different evolutionary patterns of chromatin conformation from multi-species Hi-C data. The input contains the Hi-C contact frequency data from each species. We first align the Hi-C sequencing reads of each species to the corresponding genome and then map Hi-C contacts in each species to the reference genome (human genome). We obtain a combined multi-species Hi-C contact map based on the reference genome as shown in Figure 1A, where each node in the map corresponds to multi-species contact frequencies between the corresponding pair of genomic loci (STAR Methods). We also assume that each node has a hidden state that represents the evolutionary pattern of Hi-C contacts between the corresponding paired 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 Figure 1B, Phylo-HMRF uses the Ornstein-Uhlenbeck (OU) process (Hansen, 1997; Butler and King, 2004; Hansen et al., 2008) as the continuous-trait evolutionary model. Figure 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 four species between specific paired genomic loci. We may also observe that strong Hi-C contacts only exist in some of the species between specific paired genomic loci. These different types of feature

Box 2. Primer

Comprehensive characterization of the detailed evolutionary patterns of 3D genome structure remains unclear. Existing computational approaches for comparing 3D genome organization across multiple species have limited capability without explicitly considering the continuous nature of the strengths of chromatin interactions. Here, we develop a new probabilistic model, phylogenetic hidden Markov random field (Phylo-HMRF), which integrates the continuous-trait evolutionary model with the hidden Markov random field (HMRF), to capture evolutionary patterns of chromatin interactions based on Hi-C across multiple species. Phylo-HMRF is a phylogenetic-model-based method to analyze Hi-C data as continuous signals across different species in a genome-wide manner to uncover evolutionary patterns of 3D genome organization. By applying Phylo-HMRF to Hi-C data from the same cell type (lymphoblastoid cells) in four primate species (human, chimpanzee, bonobo, and gorilla), we 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 other features of genome structure and function, such as topologically associating domains (TADs), A/B compartments, histone modifications, DNA replication timing, and sequence properties. Phylo-HMRF offers an effective model to potentially help reveal important evolutionary principles of 3D genome organization.

distributions across species are representative of the evolutionary patterns that we seek 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 dependencies that can be represented by the edges 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 contained 8 datasets. Suppose the samples in simulated datasets correspond to nodes in a graph $\mathcal{G}$. The samples thus represent features of node in $\mathcal{G}$. Similar to a Hi-C contact map, $\mathcal{G}$ has a 2D lattice structure of size $n \times n$, where each node is associated with a sample. The samples thus represent features of nodes in $\mathcal{G}$. 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 $\mathcal{G}$. We simulated the multi-species observations from multivariate Gaussian distributions with OU model parameters embedded. The details of the data simulation are in the STAR 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, to infer hidden states of the samples. Each method was run repeatedly for 10 times on each simulated dataset with different random initializations, given the number of hidden states $M = 10$. Additionally, we also included two image segmentation methods simple linear iterative clustering (SLIC) (Achanta et al., 2012) and Quick Shift (Vedaldi and Soatto, 2008) for comparison (STAR Methods). The two image segmentation methods were also run repeatedly for 10 times each. The average performance of the 10 results for each method was reported as the final performance 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 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) (STAR Methods).

The evaluation results are shown in Figure 2 and Table S1. We found that Phylo-HMRF outperforms 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 distributions, using Gaussian distributions alone in inference may not reveal the possible evolutionary dependencies between the species. In addition, both Phylo-HMRF and Gaussian-HMRF show advantages 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 segment 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 to assess if the advantage of Phylo-HMRF is consistent over varied simulation parameter settings (STAR Methods). The eight sets of simulated hidden states were shared between simulation studies I and II, while the observations were simulated with different parameter settings. We then applied Phylo-HMRF and the other methods to the datasets in simulation study II and evaluated the performance using the same procedure as we used in simulation study I. The evaluation results are in Figure S1. Again, we found that Phylo-HMRF consistently outperforms the other methods across all the datasets. Taken together, our simulation evaluation demonstrated that Phylo-HMRF is able to achieve improved accuracy consistently in estimating evolutionary patterns of Hi-C contacts in multiple species.

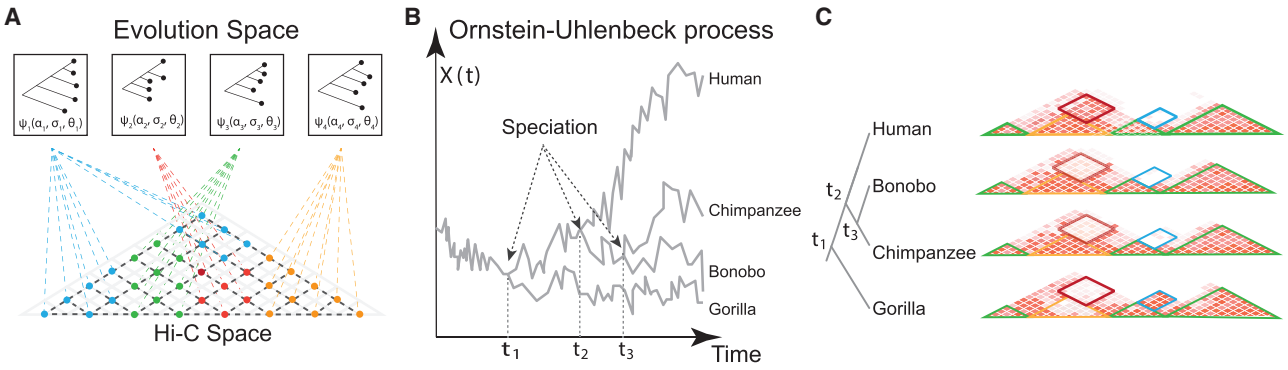


Figure 1. Overview of Phylo-HMRF

(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 paired 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 extant species. The x axis represents the evolutionary history in time. $X(t)$ on the y axis 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, and 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.

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 Rao et al. (2014). We generated Hi-C data from the lymphoblastoid cells of three other primate species, including chimpanzee, bonobo, and gorilla (STAR Methods). There are 290 M, 270 M, 240 M, and 290 M mapped read pairs for the four primate species, respectively. We ran Phylo-HMRF on all the syntenic regions on the autosomes using the human genome as the reference. 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 (Figure S2; STAR Methods). For example, we identified 9 major synteny blocks on human chromosome 1 among the four species, covering 92.50% of human chromosome 1 (Figure 3B). We then applied Phylo-HMRF to elucidate genome-wide evolutionary patterns 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 using the Elbow method (Thorndike, 1953; Goutte et al., 1999) (STAR Methods; Figure S3). Specifically, we observed how the sum of squared errors (SSE) of each clustering result changed with respect to different choices of K and chose the number of states in a range where the contribution of a larger K to a smaller SSE experienced relatively sharp decrease and tended to be small.

Phylo-HMRF identified both conserved and lineage-specific evolutionary patterns of Hi-C contact frequencies across the four primate species. For the convenience of presentation of the analysis results, we further categorized the 30 estimated hidden states into 13 groups that show higher-level distinctiveness of heterogeneous evolutionary patterns (STAR Methods). Four of the groups represent conserved or weakly conserved cross-species patterns in Hi-C 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. In the conserved states, the four species have consistently high or low or middle-level Hi-C contact frequency signals. Nine of the groups correspond to non-conserved (NC) 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 NC. Representative estimated states from each of the 13 groups are shown in Figure 3A. Hi-C contact frequency distributions of multiple species in all the 30 estimated states are shown in Figure S4.

In Figures 3B and 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 illustrated in the second panel of Figures 3B and 3C, represents the estimated hidden states of the graph $\mathcal{G}$ of the HMRF in Hi-C data comparison in the corresponding synteny block. Each node 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 2D segment in the image. Therefore,

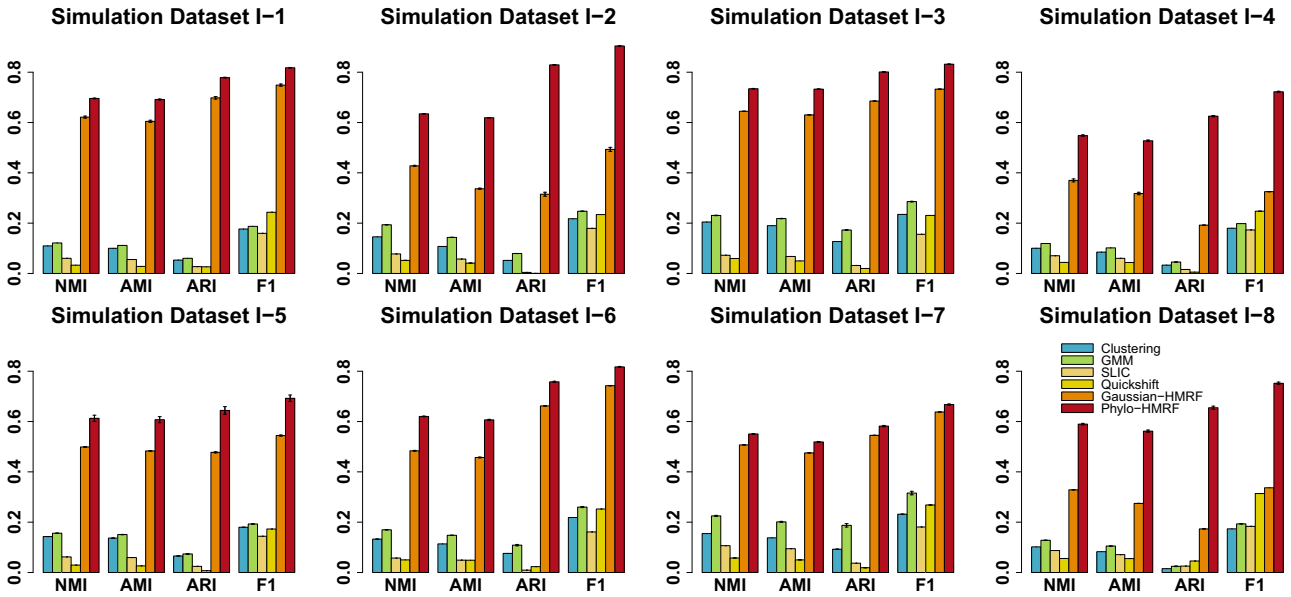


Figure 2. Evaluation Using Simulated Datasets

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 normalized mutual information (NMI), adjusted mutual information (AMI), adjusted Rand index (ARI), and F_1 score. The standard error of the results from 10 runs of each method is shown as the error bar. See also Table S1 and Figure S1.

based on the estimated states, $\mathcal{S}$ is partitioned, reflecting different cross-species Hi-C contact patterns. In Figure 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 Figure S5. In Figure 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 distance 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 < 3 Mb), and Hi-C contacts over large distances represent longer-range contacts. Short genomic loci distances correspond to areas near the diagonal of the Hi-C contact map. The state-distance plots across the synteny blocks on all the autosomes are shown in Figure 4A. We observed that C-high, C-mid, and WC states are the predominant states for the short-range contacts (black arrow in Figure 4A). This suggests that the majority of the local Hi-C contact patterns and the associated genome structures are likely to be conserved across different species. At larger genomic loci distances, which correspond to the off-diagonal areas in the Hi-C contact map, the C-low and NC states have much higher 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 Figures S5 and 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 observed that the lineage-specific states are distributed unevenly among the synteny blocks, showing occurrences either in local Hi-C contacts or in long-range Hi-C contacts.

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

Hi-C Evolutionary Patterns Correlate with Replication Timing and Histone Modifications

We next compared the predicted states from Phylo-HMRF with other features of genome structure and function. Earlier studies have shown that DNA RT is closely correlated with genome organization (Rhind and Gilbert, 2013; Pope et al., 2014). We previously reported evolutionary patterns of DNA RT using Repli-seq data 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 NC. For each pair of genomic loci with estimated Hi-C states, we examined the RT evolutionary 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 NC 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 (0–10 Mb). The results are shown in Figure 4B. Notably, it is clear that C-high, C-mid, and WC Hi-C

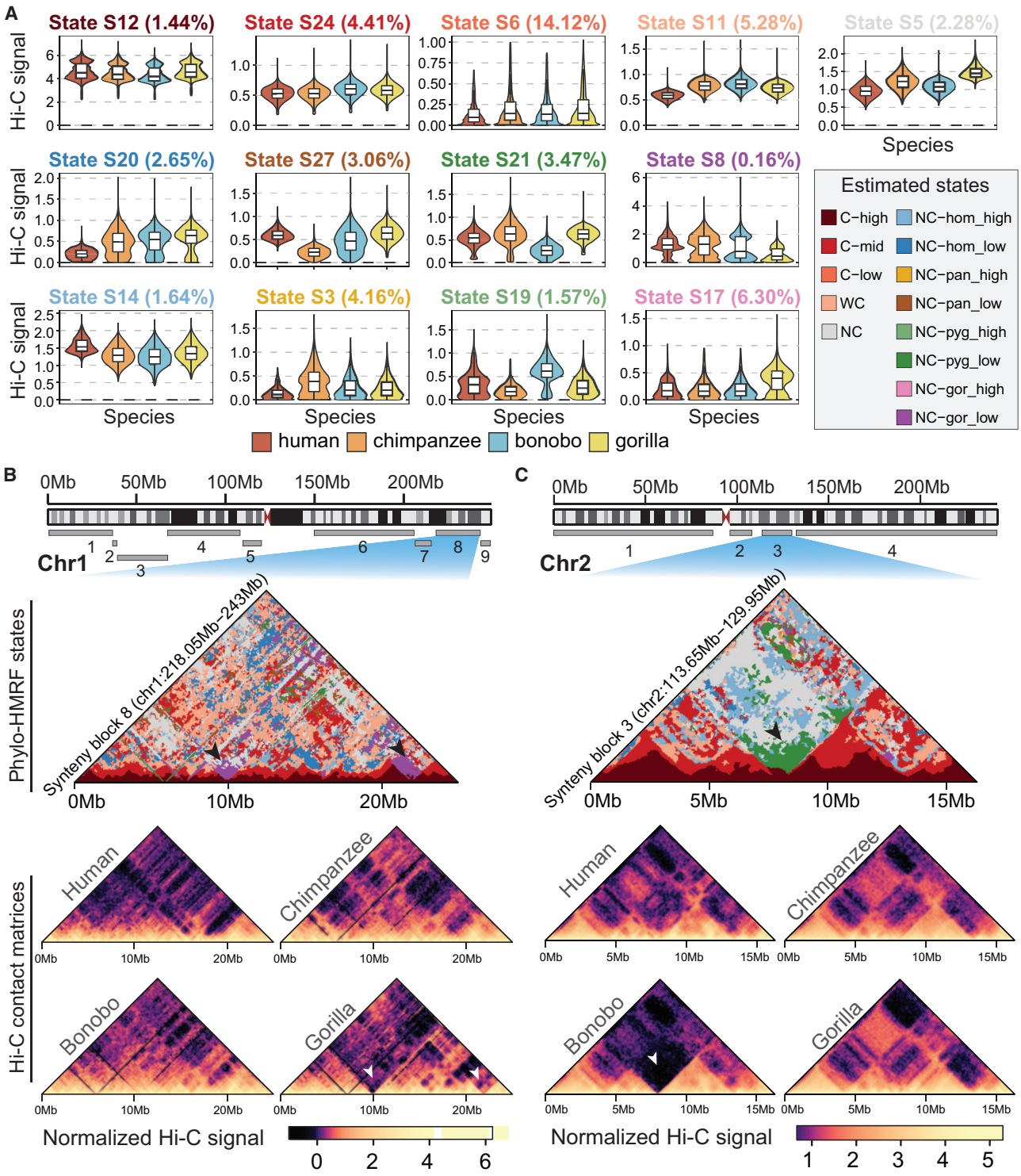


Figure 3. Evolutionary Patterns of Hi-C Contact Frequency Estimated by Phylo-HMRF

(A) Representative states from the 13 groups of evolutionary patterns. One state from each group is presented. The box plots show the normalized cross-species Hi-C contact frequency distributions of the four species in the corresponding states with outliers removed. The violin plot outlines illustrate the kernel probability density of the data.

(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. Top panel: Locations of the nine identified synteny blocks on chromosome 1. Middle panel (Phylo-HMRF states): 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. Bottom panel (Hi-C contact matrices): Hi-C contact maps of the four primate species in this synteny block, with signal

(legend continued on next page)

contact patterns have higher enrichment of genome contacts with conserved RT patterns than the NC states over most of the distance range. 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 where both loci have consistent conserved RT patterns across species. We further explored potential connections between the features and the curves observed in Figure 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) (named Arrowhead TADs and DI TADs, respectively). The average sizes of Arrowhead TADs and DI TADs are around 1 Mb and 0.6 Mb, respectively. As shown in Figure 4B, there is a changing point around 1 Mb on the distance axis on the matched-RT state fraction curve for both the C-high state and C-mid state, which approximately matches the average size of TADs. To further assess if the shapes and differences of the curves occur by chance, we randomly shuffled the RT evolutionary states along the genome and plotted the matched-RT state fraction curves for different groups of estimated Hi-C contact states based on the shuffled RT states using the same procedure as described above (STAR Methods). Specifically, the RT evolutionary states were estimated at the resolution of 6 Kb (Yang et al., 2018). We merged adjacent genomic bins with the same RT evolutionary states into segments and performed random shuffle of the segments. The curves based on the shuffled RT states are shown in Figure S7. We found that the fractions of conserved-in-RT paired genomic loci in different Hi-C contact states based on the shuffled RT states are similar to each other over most of the genomic loci distance range, not exhibiting the characteristics and diversities of curves for different Hi-C contact states found in Figure 4B. These observations suggest that the identified evolutionary patterns of local high Hi-C contacts and the evolutionary patterns of RT states may be constrained by the TAD structures.

Next, we examined the histone modification composition of paired genomic loci using the ChIP-seq data for 11 histone modifications in GM12878 from the ENCODE project (ENCODE Project Consortium, 2012). We hypothesize that paired genomic loci assigned to conserved Hi-C states inferred by Phylo-HMRF may have more similar histone modification signals than those in NC Hi-C states. To test this, we computed the percentage of paired genomic loci that have more similar histone modification signal strength than expected in the C-high, C-mid, WC, C-low, and NC Hi-C states estimated by Phylo-HMRF over a range of different distances (0–10 Mb). Specifically, for each type of histone modification, we calculated the absolute differences of quantiles derived from the signal strength (reads per million mapped reads) at paired genomic loci from a specific estimated state. We then compared the observed changes of quantiles in each state with respect to the background distribution calculated based on paired genomic loci randomly chosen from the entire synteny blocks (STAR Methods). Interestingly, even in the same range of 1D genome distance, signals from paired genomic loci for all the 11 histone modifications exhibit stronger similarities if those paired loci are annotated as conserved states than NC states as shown in Figures S8 and 4C. Additionally, the percentage of similar signals between paired loci is particularly high in the C-high state for genomic loci that are less than 4 Mb apart, which is similar to the observations in the comparison between estimated Hi-C evolutionary states and estimated evolutionary RT states. Overall, these results suggest that the conserved high Hi-C contact states are strongly correlated with genomic loci pairs that have similar histone modifications.

Hi-C Evolutionary Patterns Show Correlation with A/B Compartments and TADs

From Hi-C data, it has been revealed that at megabase resolution chromatin is segregated into two compartments: A and B (Lieberman-Aiden et al., 2009). Compartment A regions contain largely open and active chromatin and compartment B regions are typically transcriptionally more repressed. We compared the Hi-C evolutionary patterns in three different types of interactions between compartments: A-A interaction, B-B interaction, and A-B interaction. We calculated the percentage of conserved states in the interacting area of pairwise genomic bins in the multi-species Hi-C contact map for each type of compartment interactions. We specifically considered two cases: (1) a pair of 50 kb genomic bins that are in the same TAD and (2) a pair of 50 kb genomic bins that are in different TADs. We observed that for pairwise bins in two different TADs, the interactions of genomic loci coming from the same type of compartments have a higher percentage of conserved states than those from different types of compartments over varied distance between paired genome loci. Notably, pairwise bins both from B compartment (i.e., B-B interactions) have the highest fraction of conserved states (Figures 4D and S9). A recent study based on imaging has shown that the contact frequencies of paired genomic loci in the B compartment are higher than the A-B and A-A pairs (Finn et al., 2019). Our results provide the evolutionary context to support this observation for interacting loci within and across A/B compartments.

TADs are important higher-order genome organization features revealed by Hi-C (Dixon et al., 2012). We next 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 along the diagonal, we used squares with varied sizes that match the segments to detect the block patterns (STAR Methods). We identified 2,793 block patterns on the diagonals of the Hi-C contact maps of all the synteny blocks on all autosomes. We compared the boundaries of the diagonal blocks

scale displayed at the bottom. The 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. Top panel: Locations of the four identified synteny blocks on chromosome 2. Middle panel: 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 in the bottom panel. See also Figures S2–S4.

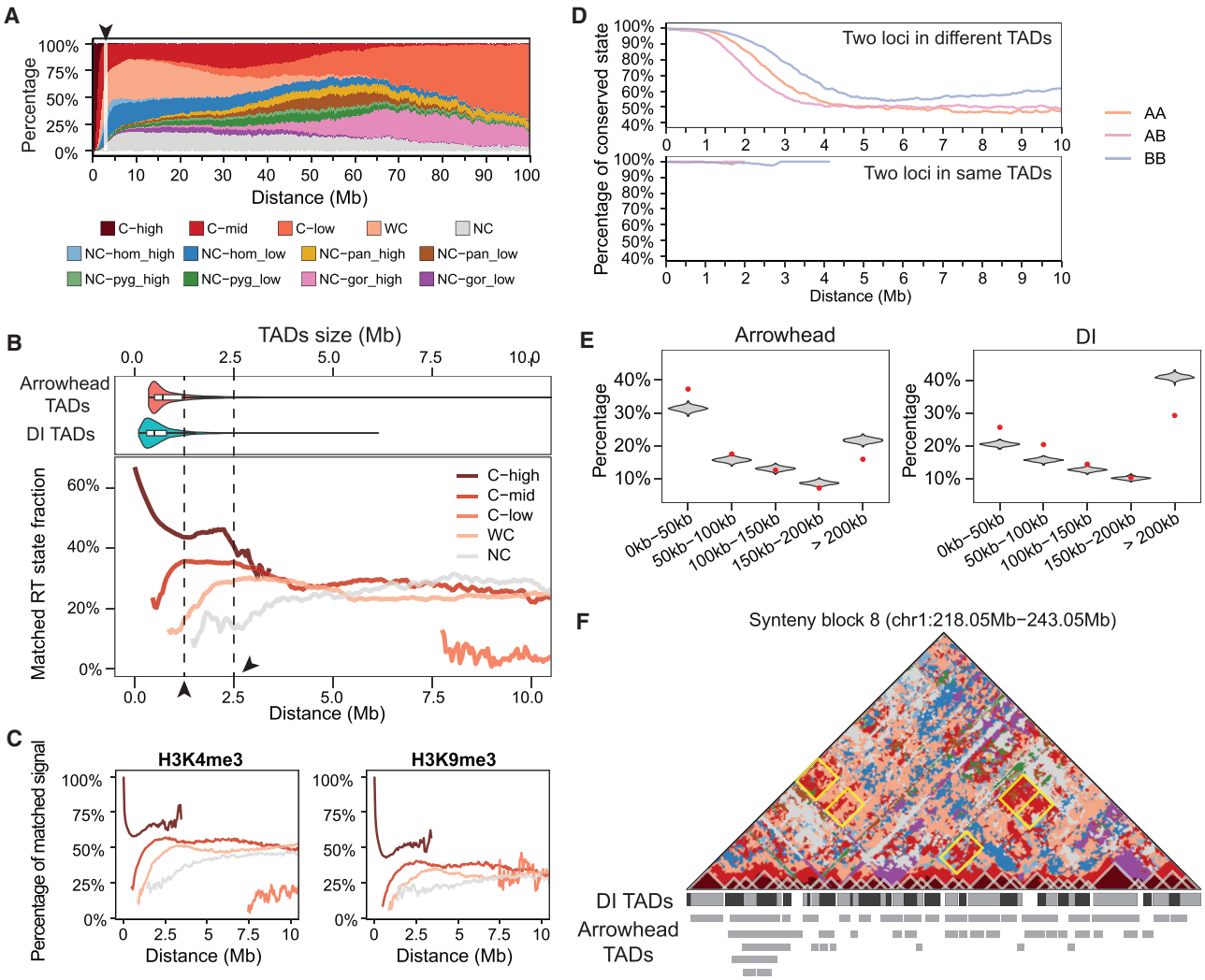


Figure 4. Comparison between the Evolutionary States of Hi-C Contacts Estimated by Phylo-HMRF and Other Features of Genome Structure and Function

(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 2.5 Mb, where C-high and C-mid are the predominant states.

(B) The percentage of pairwise genomic loci that are conserved in RT in five estimated 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 the C-mid state. The second black arrow points to the position of 2.5 Mb, where the trend change is observed in the state-distance plot, as shown in (A).

(C) The percentage of pairwise genomic loci that have a similar signal strength of a given type of histone modification in five estimated Hi-C evolutionary pattern groups. The histone modifications shown are H3K4me3 and H3K9me3.

(D) The percentage of conserved states in the interacting area of pairwise genomic loci in the multi-species Hi-C contact map for each type of A/B compartment interactions. The first panel shows the percentage of conserved states in the interacting area from different TADs. The second panel shows the percentage of conserved states in the interacting area in the same TADs.

(E) Distributions of the distance between the boundaries of the identified local conserved high Hi-C contact patterns and the nearest TAD boundaries for both Arrowhead TADs and DI TADs. The red dots are the percentages of the distance between the identified local conserved high Hi-C contact pattern boundaries and the nearest Arrowhead or DI TAD boundaries in different distance ranges. The density plots correspond to the empirical distributions of the distance between boundaries of local contact patterns and the nearest TADs, with the identified local contact patterns randomly shuffled (STAR Methods).

(F) Comparison between the boundaries of identified local conserved high Hi-C contact patterns (blocks with white borders) and the TAD boundaries in synteny block 8 on chromosome 1. Several long-range conserved TAD interaction patterns are shown with solid yellow lines as borders.

See also Figures S5–S11.

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 and calculated the percentages of the distances in five ranges (Figure 4E; STAR 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 the relatively small distance (e.g., [0, 50 Kb] and (50 Kb, 100 Kb]) than expected (Figure 4E). 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 expected percentage 36.08% observed from the empirical distance distribution (empirical p value < 2e−03, STAR Methods). 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 the 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 human chromosome 1 from the Hi-C evolutionary states estimated by Phylo-HMRF. We observed that the block boundaries show high consistency with the TAD boundaries (Figure 4F). 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. In addition, we found C-mid and WC states in the off-diagonal area of the Hi-C contact map that potentially correspond to the long-range interactions between two TADs that are conserved across species. Examples of the potentially conserved long-range TAD interactions are shown in Figure 4F (highlighted in yellow borders).

Furthermore, we sought to explore whether there are connections between DNA sequence features and the Hi-C contact evolutionary patterns identified by Phylo-HMRF. We analyzed the enrichment of different transposable element (TE) families in each estimated Hi-C contact state across species (e.g., PIF-Harbinger, hAT-Tag1) (STAR Methods; Figure S10). We then specifically assessed the potential correlations between the evolutionary patterns on TADs and sequence properties, in particular, TEs and transcription-factor-binding sites (TFBSs), by characterizing TADs into two groups based on whether a TAD is involved in conserved long-range TAD-TAD interactions (STAR Methods). We detected TE families and transcription factor (TF) motifs that show distinct enrichment patterns in the different groups of TADs (STAR Methods; Figure S11). This analysis suggests that the evolutionary patterns identified by Phylo-HMRF have the potential to reveal patterns of sequence properties in forming Hi-C contacts and long-range TAD interactions, although additional work is needed to delineate the roles of such sequence features in specific loci and their functional significance.

Taken together, our results suggest that A/B compartments and 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. Such evolutionary patterns also pave the way for the next stage in identifying potential sequence determinants for the formation of 3D genome structures.

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 the analysis of 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 a strong correlation with other types of features for genome structure and function, such as TADs, A/B compartments, DNA RT, and histone modifications. We identified conserved long-range interacting TADs based on the Hi-C contact evolutionary states estimated by Phylo-HMRF and discovered TEs and TF motif features that are correlated with the conserved long-range interacting TADs. From a methodology standpoint, Phylo-HMRF is a flexible framework that can 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 an effective method to uncover detailed evolutionary patterns of 3D genome organization based on multi-species Hi-C dataset.

There are several aspects where our method can be improved. First, model selection methods such as the utilization of the AIC and BIC criteria (Akaike, 1998; Schwarz, 1978) may help select the number of states more efficiently. 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 increases 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 mis-alignments 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 even more general. Finally, currently, we primarily depend

on manual inspection to assign estimated states into different groups to facilitate analysis. It will therefore be useful to develop a systematic approach to group the estimated states automatically.

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., 2019; 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). In this work, we made an initial attempt to explore the global correlations between sequence features and the evolutionary patterns of 3D genome organization features. However, our current analysis is still limited in its scope (only on transposons and TF motifs) and the ability to establish mechanistic characterization. Nevertheless, although future work is needed to develop integrative models to simultaneously consider both higher-order genome organization evolution and sequence level changes, our Phylo-HMRF model developed in this work has the potential to serve as a generic analytic framework to reveal different evolutionary patterns of chromatin interaction and their connections to the evolution of genome sequence and function.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2019.05.011.

ACKNOWLEDGMENTS

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

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 14, 2019
Accepted: May 22, 2019
Published: June 19, 2019

WEB RESOURCES

Phylo-HMRF, https://github.com/ma-compbio/Phylo-HMRF

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. *34*, 2274–2282.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Selected Papers of Hirotugu Akaike (Springer), pp. 199–213.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300.

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nat. Rev. Genet. *17*, 661.

Boykov, Y., and Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans. Pattern Anal. Mach. Intell. *26*, 1124–1137.

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Machine Intell. *23*, 1222–1239.

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

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

Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D., et al. (2018). The UCSC Genome Browser database: 2018 update. Nucleic Acids Res. *46*, D762–D769.

Celeux, G., Forbes, F., and Peyrard, N. (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. Pattern Recognit. *36*, 131–144.

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

Chou, H.H., Hayakawa, T., Diaz, S., Krings, M., Indriati, E., Leakey, M., Paabo, S., Satta, Y., Takahata, N., and Varki, A. (2002). Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. Proc. Natl. Acad. Sci. USA *99*, 11736–11741.

Choudhary, M.N., Friedman, R.Z., Wang, J.T., Jang, H.S., Zhuo, X., and Wang, T. (2018). Co-opted transposons help perpetuate conserved higher-order chromosomal structures. bioRxiv. https://doi.org/10.1101/485342.

Comaniciu, D., and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. *24*, 603–619.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B *39*, 1–38.

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.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. *3*, 95–98.

Durrett, R. (2019). Probability: Theory and Examples, *vol. 49* (Cambridge University Press).

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

Felsenstein, J. (1985). Phylogenies and the comparative method. Am. Nat. *125*, 1–15.

Finn, E.H., Pegoraro, G., Brandão, H.B., Valton, A.L., Oomen, M.E., Dekker, J., Mirny, L., and Misteli, T. (2019). Extensive heterogeneity and intrinsic variation in spatial genome organization. Cell *176*, 1502–1515.

Freckleton, R.P. (2012). Fast likelihood calculations for comparative analyses. Methods Ecol. Evol. *3*, 940–947.

Friedman, N., Ninio, M., Pe'er, I., and Pupko, T. (2002). A structural EM algorithm for phylogenetic inference. J. Comput. Biol. *9*, 331–353.

Fudenberg, G., and Pollard, K.S. (2019). Chromatin features constrain structural variation across evolutionary timescales. Proc. Natl. Acad. Sci. USA *116*, 2175–2180.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. *6*, 721–741.

Geman, S., and Graffigne, C. (1986). Markov random field image models and their applications to computer vision. In Proceedings of the International Congress of Mathematicians, *vol. 1* (American Mathematical Society), p. 2.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F.Å., and Hansen, L.K. (1999). On clustering fMRI time series. NeuroImage *9*, 298–310.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Hansen, T.F. (1997). Stabilizing selection and the comparative analysis of adaptation. Evolution *51*, 1341–1351.

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

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC genome browser database: update 2006. Nucleic Acids Res. *34*, D590–D598.

Hoerl, A.E., and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics *12*, 55–67.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S.R., Tan, G., et al. (2017). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. *46*, D260–D266.

Kolmogorov, V., and Zabih, R. (2004). What energy functions can be minimizedvia graph cuts? IEEE Trans. Pattern Anal. Mach. Intell. *26*, 147–159.

Lazar, N.H., Nevonen, K.A., O'Connell, B., McCann, C., O'Neill, R.J., Green, R.E., Meyer, T.J., Okhovat, M., and Carbone, L. (2018). Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. Genome Res. *28*, 983–997.

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

Ma, J., Zhang, L., Suh, B.B., Raney, B.J., Burhans, R.C., Kent, W.J., Blanchette, M., Haussler, D., and Miller, W. (2006). Reconstructing contiguous regions of an ancestral genome. Genome Res. *16*, 1557–1565.

MacKay, D.J. (2003). Information Theory, Inference and Learning Algorithms (Cambridge University Press).

Manning, C.D., Raghavan, P., and Schütze, H. (2008). Introduction to Information Retrieval (Cambridge University Press).

Naval-Sánchez, M., Potier, D., Hulselmans, G., Christiaens, V., and Aerts, S. (2015). Identification of lineage-specific cis-regulatory modules associated with variation in transcription factor binding and chromatin activity using Ornstein–Uhlenbeck models. Mol. Biol. Evol. *32*, 2441–2455.

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

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

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.

Pele, O., and Werman, M. (2010). The quadratic-chi histogram distance family. In European Conference on Computer Vision (Springer), pp. 749–762.

Perona, P., and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. *12*, 629–639.

Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S., Canfield, T.K., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402–405.

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

Rhind, N., and Gilbert, D.M. (2013). DNA replication timing. Cold Spring Harb. Perspect. Biol. *5*, a010132.

Rohlfs, R.V., Harrigan, P., and Nielsen, R. (2014). Modeling gene expression evolution with an extended Ornstein–Uhlenbeck process accounting for within-species variation. Mol. Biol. Evol. *31*, 201–211.

Rowley, M.J., and Corces, V.G. (2018). Organizational principles of 3D genome architecture. Nat. Rev. Genet. *19*, 789–800.

Schwarz, G. (1978). Estimating the dimension of a model. Ann. Stat. *6*, 461–464.

Sima, J., Chakraborty, A., Dileep, V., Michalski, M., Klein, K.N., Holcomb, N.P., Turner, J.L., Paulsen, M.T., Rivera-Mulia, J.C., Trevilla-Garcia, C., et al. (2019). Identifying cis elements for spatiotemporal control of mammalian DNA replication. Cell *176*, 816–830.

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

Thorndike, R.L. (1953). Who belongs in the family? Psychometrika *18*, 267–276.

Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., and Yu, T.; scikit-image contributors (2014). scikit-image: image processing in Python. PeerJ *2*, e453.

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

Veksler, O. (1999). Efficient graph-based energy minimization methods in computer vision, PhD thesis (Cornell University).

Vietri Rudan, M.V., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. *10*, 1297–1309.

Vinh, N.X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. *11*, 2837–2854.

Yang, Y., Gu, Q., Zhang, Y., Sasaki, T., Crivello, J., O'Neill, R.J., Gilbert, D.M., and Ma, J. (2018). Continuous-trait probabilistic model for comparing multi-species functional genomic data. Cell Syst. *7*, 208–218.

Zhang, J. (1992). The mean field theory in EM procedures for Markov random fields. IEEE Trans. Signal Process. *40*, 2570–2583.

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

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging *20*, 45–57.

Zhang, Y., Li, T., Preissl, S., Grinstein, J., Farah, E., Destici, E., Lee, A.Y., Chee, S., Qiu, Y., Ma, K., et al. (2019). 3D chromatin architecture remodeling during human cardiomyocyte differentiation reveals A role of HERV-H in demarcating chromatin domains. bioRxiv. https://doi.org/10.1101/485961.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
RPMI 1640	Lonza	BW12702F12
37% Formaldehyde	Fisher	F79500
Critical Commercial Assays		
TruSeq DNA PCR-Free Low Throughput Library Prep kit	Illumina	20015962
Deposited Data		
Hi-C data of chimpanzee, bonobo, gorilla	This paper	GEO: GSE128800
Hi-C data of GM12878 cell line	Rao et al., 2014	GEO: GSE63525
Human genome hg38	International Human Genome Sequencing Consortium	http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips
Chimpanzee genome panTro5	The Chimpanzee Genome Sequencing Consortium	http://hgdownload.soe.ucsc.edu/goldenPath/panTro5/bigZips
Bonobo genome panPan2	Max-Planck Institute for Evolutionary Anthropology	http://hgdownload.soe.ucsc.edu/goldenPath/panPan2/bigZips
Gorilla genome gorGor4	Wellcome Trust Sanger Institute, European Bioinformatics Institute	http://hgdownload.soe.ucsc.edu/goldenPath/gorGor4/bigZips
ChIP-seq data of H2AFZ (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF762TRA
ChIP-seq data of H2AFZ (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF848PUT
ChIP-seq data of H3K27ac (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF804NCH
ChIP-seq data of H3K27ac (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF948GTC
ChIP-seq data of H3K27me3 (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF231DJN
ChIP-seq data of H3K27me3 (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF175YYN
ChIP-seq data of H3K36me3 (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF958QVX
ChIP-seq data of H3K36me3 (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF460TXJ
ChIP-seq data of H3K4me1 (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF153KPG
ChIP-seq data of H3K4me1 (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF815TLX
ChIP-seq data of H3K4me2 (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF803ROB
ChIP-seq data of H3K4me2 (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF128WUO
ChIP-seq data of H3K4me3 (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF019VEK
ChIP-seq data of H3K4me3 (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF278QPY
ChIP-seq data of H3K79me2 (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF676NDU
ChIP-seq data of H3K79me2 (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF231YZJ

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ChIP-seq data of H3K9ac (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF737GSB
ChIP-seq data of H3K9ac (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF424IMO
ChIP-seq data of H3K9me3 (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF663EWP
ChIP-seq data of H3K9me3 (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF758GUH
ChIP-seq data of H3K9me3 (replicate 3) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF370XAS
ChIP-seq data of H4K20me1 (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF880XJW
ChIP-seq data of H4K20me1 (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF937PBY
RNA-seq data (replicate 1) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF212CQQ
RNA-seq data (replicate 2) in GM12878	ENCODE Project Consortium	https://www.encodeproject.org/files/ENCFF350QZU
Experimental Models: Cell Lines		
Homo Sapiens (Human): lymphoblastoid cell line GM12878	Rao et al., 2014	Cat# GM12878; RRID: CVCL_7526
Pan troglodytes (Common Chimpanzee): lymphoblastoid cell line	Ajit Varki (Chou et al., 2002)	N/A
Pan Paniscus (Bonobo): lymphoblastoid cell line Kidogo	Ajit Varki (Chou et al., 2002)	N/A
Gorilla gorilla (Gorilla): lymphoblastoid cell line	Ajit Varki (Chou et al., 2002)	N/A
Software and Algorithms		
Phylo-HMRF	This paper	https://github.com/ma-compbio/Phylo-HMRF
Juicer	Durand et al., 2016	https://github.com/aidenlab/juicer
liftOver	Hinrichs et al., 2006	https://genome.ucsc.edu/cgi-bin/hgLiftOver
inferCars	Ma et al., 2006	http://www.bx.psu.edu/miller_lab/car
scikit-learn	Pedregosa et al., 2011	http://scikit-learn.org/stable

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and algorithms should be directed to and will be fulfilled by the Lead Contact Jian Ma (jianma@cs.cmu.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The four primate species included in this study are Homo Sapiens (Human), Pan troglodytes (Common Chimpanzee), Pan Paniscus (Bonobo), and Gorilla gorilla (Gorilla). For human we used the GM12878 cell line, a lymphoblastoid cell line which is established from EBV (Epstein-Barr Virus)-transformed B-lymphocytes from a female donor. For the three non-human primate species, we used lymphoblastoid cell lines from the corresponding species. The lymphoblastoid cell lines of Common Chimpanzee (abbreviated as Chimpanzee), Bonobo, and Gorilla were kindly provided by Dr. Ajit Varki (University of California at San Diego, La Jolla, CA, USA) and have been used in (Chou et al., 2002). Each of the cell lines is established from EBV-transformed B cells from one biological individual of the corresponding species. The cells of Chimpanzee and Bonobo are from male. The cells of Gorilla are from female. The cells of the three species were grown as suspension cultures in RPMI-1640 media supplemented with 15% fetal bovine serum and Penicillin/Streptomycin at 37°C. We only used autosomes for analysis for each of the species.

METHOD DETAILS

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 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ 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 $\mathcal{V}$ as the set of indices of the nodes in $\mathcal{G}$, i.e., $\mathcal{V} = \{1, \cdots, N\}$. The i-th node is associated with a random variable $X_i \in \mathbb{R}^d$ representing the multi-species observations on this node, $i \in \mathcal{V}$. 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 $\mathcal{G}$.

Using a hidden Markov random field (HMRF) model, we assume that each node in $\mathcal{G}$ is also associated with a random variable $Y_i \in S = \{1, \cdots, M\}$, representing the unknown hidden state of this node, $i \in \mathcal{V}$. S is the set of hidden states. We assume $Y = \{Y_i\}_{i\in\mathcal{V}}$ to be an MRF. For each configuration of Y, X_i follows a conditional probability distribution $p(x_i|y_i)$, which is the emission probability distribution, and $X = \{X_i\}_{i\in\mathcal{V}}$ 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 the same hidden states.

Phylo-HMRF estimates the evolutionary patterns by inferring the hidden states $\mathbf{y} = \{y_i\}_{i\in\mathcal{V}}$ from the observations $\mathbf{x} = \{x_i\}_{i\in\mathcal{V}}$, using the assumption that there are spatial dependencies between adjacent nodes in the graph $\mathcal{G}$. 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 $\mathcal{G}$ is:

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

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:

$$E(\mathbf{y}|\mathbf{x}, \Theta) = \sum_{i\in\mathcal{V}} U(x_i, y_i, \Theta) + \sum_{(i,j)\in\mathcal{E}} V(y_i, y_j; x_i, x_j), \quad \text{(Equation 2)}$$

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 $f(y_i, y_j; x_i, x_j) \propto \exp(-V(y_i, y_j; x_i, x_j))$. 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:

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

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 Gaussian process (MacKay, 2003) 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., 2014; 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 $\widehat{X}_i$, the OU process can be represented as the following (Hansen, 1997; Butler and King, 2004):

$$d(\widehat{X}_i(t)) = \alpha[\theta_i(t) - \widehat{X}_i(t)]dt + \sigma dB_i(t), \quad \text{(Equation 4)}$$

where $\widehat{X}_i(t)$ represents the observation of $\widehat{X}_i$ at time point t, $B_i(t)$ represents the standard Brownian motion (Durrett, 2019), 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 multi-species observations follow multivariate Gaussian distribution, and the expectation, variance, and covariance of the observations of species can be computed given the phylogenetic tree. Let X_i, X_j denote the observed traits of the i-th species (species i) and the j-th species (species j), respectively. Suppose that X_p is the trait of the ancestor of species i, and X_a is the trait of the most recent common ancestor of species i and species j. We have (Butler and King, 2004; Rohlfs et al., 2014):

$$\mathbb{E}(X_i) = \mathbb{E}(X_p)e^{-\alpha_i t_{ip}} + \theta_i\left(1 - e^{-\alpha_i t_{ip}}\right), \quad \text{(Equation 5)}$$

$$\mathrm{Cov}(X_i, X_j) = \mathrm{Var}(X_a)\exp\Big(-\sum_{k\in l_{ij}}\alpha_k t_k - \sum_{k\in l_{ji}}\alpha_k t_k\Big), \quad \text{(Equation 6)}$$

$$\mathrm{Var}(X_i) = \frac{\sigma_i^2}{2\alpha_i}\left(1 - e^{-2\alpha_i t_{ip}}\right) + \mathrm{Var}(X_p)e^{-2\alpha_i t_{ip}}, \quad \text{(Equation 7)}$$

where t_{ip}, t_k represent evolution time along the corresponding branches in the phylogenetic tree, respectively. l_{ij} represents the set of the ancestor nodes of species i and i itself after the divergence of species i with species j. Specifically, t_{ip} corresponds to length of the branch from the parent of species i to species i. For $k \in l_{ij}$, t_k corresponds to length of the branch from the parent of species k to species k. 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 State 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 an 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 (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: $Q(\Theta,\Theta^g) = \mathbb{E}[\log p(\mathbf{x},\mathbf{y}|\Theta)|\mathbf{x},\Theta^g]$, where $\mathbf{x}$, $\mathbf{y}$ represent the observations and the hidden states, respectively. 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 later section):

$$Q(\Theta,\Theta^g) = \sum_{l=1}^{M}\sum_{i\in\mathscr{V}} p(y_i = l|x_i,\Theta^g)\log p(x_i|y_i = l,\Theta) + \sum_{l=1}^{M}\sum_{i\in\mathscr{V}} p(y_i = l|x_i,\Theta^g)\log p\left(y_i = l\middle|y^g_{\mathscr{N}_i},\Theta\right), \quad \text{(Equation 8)}$$

where the two parts of the Q function encode the unary potential and the pairwise potential of the HMRF of $\mathscr{G}$, 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:

$$p(y_i = l|x_i,\Theta^g) = \frac{p(x_i|y_i = l,\Theta^g)p\left(y_i = l\middle|y^g_{\mathscr{N}_i}\right)}{\sum_{l=1}^{M}p(x_i|y_i = l,\Theta^g)p\left(y_i = l\middle|y^g_{\mathscr{N}_i}\right)}, \quad \text{(Equation 9)}$$

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

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

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

where Z is the normalization constant. We can adopt different definitions of the pairwise potential $V(y_i, y_j)$ (see later section). The definition we use takes into consideration the difference between features of the adjacent nodes in imposing the penalty on inconsistent states of the neighbors:

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

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

Let $L(\Theta^{(l)}) = -\sum_{i\in\mathcal{T}} w_i^{(l)} \log p(x_i|y_i = l, \Theta)$, $w_i^{(l)} = p(y_i = l|x_i, \Theta^g)$. 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 derived as (details in later section):

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

where $\tilde{S}_\Theta^{(l)} = \frac{1}{N}\sum_{i\in\mathcal{T}} w_i^{(l)}(x_i - \mu_\Theta^{(l)})(x_i - \mu_\Theta^{(l)})^T$, $w_i^{(l)}$ is defined as above, $\Theta^{(l)}$ represents the phylogenetic model parameters associated with hidden state l, and λ is the regularization coefficient of the l_2-norm regularization (Hoerl and Kennard, 1970) that is used to reduce overfitting of the model. tr(A) represents the trace of a matrix A. $\Sigma_\Theta^{(l)}$, $\mu_\Theta^{(l)}$ represent the covariance matrix and the mean vector of the multivariate Gaussian distribution associated with the phylogenetic model of state l, respectively. Using OU processes, 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 OU 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 $\tilde{\alpha}_{l,v} = \alpha_{l,v}b_{l,v}$, $\tilde{\sigma}^2_{l,v} = \sigma^2_{l,v}b_{l,v}$ to present the combined effect of the branch length and the selection or Brownian motion parameters along this branch. Here $b_{l,v}$ 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 are 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) (see later section). 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:

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

   where $U(\mathbf{x}|\mathbf{y})$ and $U(\mathbf{y})$ are the total unary potential and total pairwise potential of $\mathcal{G}$, respectively.
3. *Calculate Posterior Probability Distribution*. In each Expectation-step (E-step), given the current estimated model parameters Θ^g and the estimated hidden state configuration in the previous step, we compute $p(y_i = l|x_i, \Theta^g)$ using Equations 9 and 10.
4. *Estimate Model Parameters by Solving the Optimization Problem with OU Models Embedded*. In each M-step, we solve the optimization problem in Formula 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 (see later section). Graph Cuts algorithms seek 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 problems and α-expansion or α-β swap algorithms can be used (Boykov et al., 2001; Veksler, 1999). The solution is an approximate solution in the multi-labeling problem that has been shown to be strongly probably able to reach a local minima (Boykov et al., 2001).

Detailed Description of Phylo-HMRF with OU Process

In Phylo-HMRF, we embed the OU model into the emission probability function of the HMRF model. Suppose Θ and Θ^g represent the model parameters and the current model parameter estimates, respectively. We use the EM algorithm to maximize the expectation of the complete-data log likelihood, which is the Q function $Q(\Theta, \Theta^g)$. We have:

$$Q(\Theta, \Theta^g) = \mathbb{E}[\log p(\mathbf{x}, \mathbf{y}|\Theta)|\mathbf{x}, \Theta^g] = \sum_{\mathbf{y}\in\mathcal{S}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g)\log p(\mathbf{x}, \mathbf{y}|\Theta), \quad \text{(Equation 14)}$$

where $\mathbf{x}$ are the observations, $\mathbf{y}$ are the hidden states, and $\mathcal{S}_N$ is the set of all the possible state configurations of size N. N is the sample size. We have:

$$Q(\Theta, \Theta^g) = \sum_{\mathbf{y}\in\mathcal{S}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g)[\log p(\mathbf{x}|\mathbf{y}, \Theta) + \log p(\mathbf{y}|\Theta)], \quad \text{(Equation 15)}$$

where $p(\mathbf{y}|\Theta)$ represents the probability of a hidden state configuration over the whole graph $\mathcal{G}$. Using pseudo-likelihood approximation (Geman and Graffigne, 1986), we can approximate $p(\mathbf{y}|\Theta)$ with:

$$p(\mathbf{y}|\Theta) = \prod_{i\in\mathcal{V}} p\left(y_i|y_{\mathcal{N}_i}, \Theta\right). \quad \text{(Equation 16)}$$

where $\mathcal{N}_i$ denote the set of neighboring nodes of the node i. Then we have:

$$\begin{aligned}
Q(\Theta, \Theta^g) &= \sum_{\mathbf{y}\in\mathcal{S}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g)[\log p(\mathbf{x}|\mathbf{y}, \Theta) + \log p(\mathbf{y}|\Theta)] \\
&= \sum_{\mathbf{y}\in\mathcal{S}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g)\left[\sum_{i\in\mathcal{V}} \log p(x_i|y_i, \Theta) + \sum_{i\in\mathcal{V}} \log p\left(y_i|y_{\mathcal{N}_i}, \Theta\right)\right] \\
&= \sum_{\mathbf{y}\in\mathcal{S}_N}\sum_{i\in\mathcal{V}} p(\mathbf{y}|\mathbf{x}, \Theta^g)\log p(x_i|y_i, \Theta) + \sum_{\mathbf{y}\in\mathcal{S}_N}\sum_{i\in\mathcal{V}} p(\mathbf{y}|\mathbf{x}, \Theta^g)\log p\left(y_i|y_{\mathcal{N}_i}\right) \\
&= \sum_{i\in\mathcal{V}}\sum_{l=1}^{M}\sum_{q_{-i}\in\mathcal{S}_{N-1}} p(y_i = l, y_{-i}|\mathbf{x}, \Theta^g)\log p(x_i|y_i = l, \Theta) + \sum_{\mathbf{y}\in\mathcal{S}_N}\sum_{i\in\mathcal{V}} p(\mathbf{y}|\mathbf{x}, \Theta^g)\log p\left(y_i|y_{\mathcal{N}_i}\right) \\
&= \sum_{l=1}^{M}\sum_{i\in\mathcal{V}} p(y_i = l|\mathbf{x}, \Theta^g)\log p(x_i|y_i = l, \Theta) + \sum_{i\in\mathcal{V}}\sum_{\mathbf{y}\in\mathcal{S}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g)\log p\left(y_i|y_{\mathcal{N}_i}\right).
\end{aligned} \quad \text{(Equation 17)}$$

Let $x_{\mathcal{N}_i}$ be the neighbors of x_i, and $y_{\mathcal{N}_i}$ be the state configuration of $x_{\mathcal{N}_i}$. Here y_{-i} represents the hidden states of all the nodes other than node i.

Calculating $\sum_{i\in\mathcal{V}}\sum_{\mathbf{y}\in\mathcal{S}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g)\log p(y_i|y_{\mathcal{N}_i})$ requires computing $p(\mathbf{y}|\mathbf{x}, \Theta^g)$ and $\log p(y_i|y_{\mathcal{N}_i})$ over all the possible configurations $\mathbf{y}\in\mathcal{S}_N$, which is computationally intractable. By mean-field approximation (Celeux et al., 2003; Zhang, 1992), we can use estimated hidden states $y^g_{\mathcal{N}_i}$ from the previous iteration of the HMRF-EM algorithm to approximate $p(y_i|y_{\mathcal{N}_i})$, which can simplify the computation of the Q function. Then we have:

$$\sum_{i\in\mathcal{V}}\sum_{\mathbf{y}\in\mathcal{S}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g)\log p\left(y_i|y^g_{\mathcal{N}_i}\right) = \sum_{i\in\mathcal{V}}\sum_{l=1}^{M}\sum_{q_{-i}\in\mathcal{S}_{N-1}} p(y_i = l, y_{-i}|\mathbf{x}, \Theta^g)\log p\left(y_i = l|y^g_{\mathcal{N}_i}\right) = \sum_{l=1}^{M}\sum_{i\in\mathcal{V}} p(y_i = l|x_i, \Theta^g)\log p\left(y_i = l|y^g_{\mathcal{N}_i}, \Theta\right). \quad \text{(Equation 18)}$$

Therefore, from Equations 17 and 18, we have the Q function in Equation 8. The parameters of the OU model are embedded in the term $p(x_i|y_i)$ of $Q(\Theta, \Theta^g)$. The second part of the Q function encodes the pairwise potentials and does not include the OU model parameters. Therefore, the first and second parts of the Q function can be estimated separately.

Based on the model of the OU process, we assume the observations of observed species (leaf nodes in the phylogenetic tree) follow multivariate Gaussian distribution. We have:

$$\begin{aligned}
p(x_i|y_i = l, \Theta) &= p\left(x_i|\Theta^{(l)}\right) \\
&= \frac{1}{(2\pi)^{d/2}\left|\Sigma^{(l)}_\Theta\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(x_i - \mu^{(l)}_\Theta\right)^T\left[\Sigma^{(l)}_\Theta\right]^{-1}\left(x_i - \mu^{(l)}_\Theta\right)\right\},
\end{aligned} \quad \text{(Equation 19)}$$

$$\log p\left(x|\Theta^{(l)}\right) \propto -\frac{1}{2}\log\left|\Sigma^{(l)}_\Theta\right| - \frac{1}{2}\left(x_i - \mu^{(l)}_\Theta\right)^T\left[\Sigma^{(l)}_\Theta\right]^{-1}\left(x_i - \mu^{(l)}_\Theta\right), \quad \text{(Equation 20)}$$

where $\Theta^{(l)}$ represent the OU model parameters associated with the l-th state and d is the number of the observed species, $l \in \{1, \cdots, M\}$. The underlying phylogenetic model ψ_l is embedded into the covariance matrix $\Sigma^{(l)}_\Theta$ and the mean vector $\mu^{(l)}_\Theta$ according to Equations 5, 6, and 7. We calculate $p(x_i|y_i = l, \Theta^g)$ in the same way as shown in Equation 19 by replacing Θ

with Θ^g, where Θ^g denotes the current model parameter estimates. Let $w_i^{(l)} = p(y_i = l|x_i, \Theta^g)$. $p(y_i = l|x_i, \Theta^g)$ is calculated using Equation 9. We have:

$$Q(\Theta, \Theta^g) = \sum_{l=1}^{M} \sum_{i \in \mathscr{V}} w_i^{(l)} \log p(x_i|y_i = l, \Theta) + \sum_{l=1}^{M} \sum_{i \in \mathscr{V}} w_i^{(l)} \log p\left(y_i = l \middle| y^g_{\mathscr{N}_i}\right) \quad \text{(Equation 21)}$$

$$= \sum_{l=1}^{M} \sum_{i \in \mathscr{V}} w_i^{(l)} \left[-\frac{1}{2} \log \left|\Sigma_\Theta^{(l)}\right| - \frac{1}{2} \left(x_i - \mu_\Theta^{(l)}\right)^T \left[\Sigma_\Theta^{(l)}\right]^{-1} \left(x_i - \mu_\Theta^{(l)}\right) + \log p\left(y_i = l \middle| y^g_{\mathscr{N}_i}\right) \right] + C, \quad \text{(Equation 22)}$$

where C is a constant.

We perform parameter estimation for each of the possible states.

Let $L(\Theta^{(l)}) = -\sum_{i \in \mathscr{V}} w_i^{(l)} \log p(x_i|y_i = l, \Theta)$. We have:

$$L\left(\Theta^{(l)}\right) = \frac{1}{2} \log \left|\Sigma_\Theta^{(l)}\right| \sum_{i \in \mathscr{V}} w_i^{(l)} + \frac{1}{2} \sum_{i \in \mathscr{V}} \left(x_i - \mu_\Theta^{(l)}\right)^T \left[\Sigma_\Theta^{(l)}\right]^{-1} \left(x_i - \mu_\Theta^{(l)}\right) w_i^{(l)}. \quad \text{(Equation 23)}$$

Therefore, the first part of the negative Q function with respect to a given state l can be represented as:

$$\tilde{L}\left(\Theta^{(l)}\right) = \frac{1}{N} \log \left|\Sigma_\Theta^{(l)}\right| \sum_{i \in \mathscr{V}} w_i^{(l)} + \mathrm{tr}\left(\left[\Sigma_\Theta^{(l)}\right]^{-1} \tilde{S}_\Theta^{(l)}\right), \quad \text{(Equation 24)}$$

where $\tilde{S}_\Theta^{(l)} = \frac{1}{N} \sum_{i \in \mathscr{V}} w_i^{(l)} \left(x_i - \mu_\Theta^{(l)}\right)\left(x_i - \mu_\Theta^{(l)}\right)^T$, and $\Theta^{(l)}$ represents the phylogenetic model parameters associated with state l. We assume that the phylogenetic tree topology τ_l is given. The branch lengths b_l can be combined in effect to α_l and σ_l. As we allow varied selection strengths and Brownian motion intensities of the OU model along each branch of the phylogenetic tree, and allow varied optimal values on each tree node, there are many OU model parameters to estimate, which may result in overfitting of the model if the sample size is not large enough. We apply ℓ_2-norm regularization to the parameters $\Theta^{(l)}$ to reduce model overfitting by adding the regularization term $\lambda \|\Theta^{(l)}\|_2^2$. In each M-step, the objective function of a given state l is defined in Formula 12, where λ is the regularization coefficient. We define $\lambda = \lambda_0/\sqrt{N}$. λ_0 can be predefined. We choose $\lambda_0 = 4.0$ in both the simulation study and real data study. The same regularization coefficient was adopted in (Yang et al., 2018) and the value of $\lambda_0 = 4.0$ was used. We found that the performance of the model was not sensitive to the choice of λ_0 within a range. We therefore use the same choice of λ_0 in Phylo-HMRF.

For the second part of the Q function, we can have different definitions of the pairwise potential $V(y_i, y_j)$. We consider two definitions. The first definition is:

$$V\left(y_i, y_j\right) = \beta_0 I\left(y_i \neq y_j\right), \quad \text{(Equation 25)}$$

where β_0 is a predefined adjustable regularization coefficient, which can also be considered as pairwise potential parameter. The second definition is shown in Equation 11. In Phylo-HMRF we mainly use the second definition. The results in the simulation evaluation and in the real data application were obtained by Phylo-HMRF using the second definition. The pairwise potential coefficients β_0 and β_1 in Equation 11 can either be estimated as parameters or predefined. In many applications the pairwise potential coefficients are often estimated through a number of trials and predefined. The pairwise potential coefficients can be chosen such that the pairwise potential is at the same scale of the unary potential. We choose $\beta_0 \in [1, 3]$ and $\beta_1 \in [0.1, 0.5]$ in the simulation evaluation and the real data application based on empirical observations from a simulation dataset.

Model Initialization in Phylo-HMRF

In the OU-model embedded HMRF-EM algorithm, we need to initialize the model parameters. We follow the similar approaches in Yang et al. (2018) for parameter initialization in the EM algorithm. For the first approach, we perform K-means clustering on the samples. We assign a hidden state to the samples in the same cluster. For each cluster, we estimate the OU model parameters by maximum likelihood estimation (MLE). The objective function of the MLE problem is similar to that defined in Formula 12. The difference is that we set $w_i^{(l)} = 1$, and change $i \in \mathscr{V}$ to the constraint $i \in \mathscr{C}_l$, where $\mathscr{C}_l$ represents the set of the nodes that are assigned to state l by the K-means clustering result. The estimated model parameters are used as initialization of the OU model parameters for each hidden state. The second approach is to randomly sample the parameter values from predefined uniform distributions.

For the third approach, we use a linear combination of parameters obtained from the first and second approaches for parameter initialization. The initial parameter values are chosen as $\Theta_0 = w_1\Theta_1 + (1 - w_1)\Theta_2$, where $w_1 \in [0, 1]$, Θ_1 and Θ_2 are parameter estimates from the first and second approaches, respectively. In practice, we used the third approach.

HMRF-EM Algoirthm and Graph Cuts Algorithm Used in Phylo-HMRF

In Phylo-HMRF, given the current estimated model parameters, we use the Graph Cuts algorithm (Boykov et al., 2001; Boykov and Kolmogorov, 2004) to estimate the hidden states in step 2 of the HMRF-EM algorithm. In step 2, we seek an approximate solution to

the energy minimization problem as defined in Equation 13. We have also shown that minimizing the energy is equivalent to maximizing the joint probability.

We find the solution $\{\mathbf{y}^*, \Theta^*\} = \text{argmax}_{\mathbf{y},\Theta} E(\mathbf{y}|\mathbf{x}, \Theta)$ approximately by alternatively performing

$$\mathbf{y}^* = \underset{\mathbf{y}}{\text{argmin}} E(\mathbf{y}|\mathbf{x}, \Theta^g), \tag{Equation 26}$$

and

$$\Theta^* = \underset{\Theta}{\text{argmax}} \mathbb{E}[\log p(\mathbf{x}, \mathbf{y}|\Theta)|\mathbf{x}, \Theta^g], \tag{Equation 27}$$

where Θ^g is the current estimates of the model parameters. We use $\mathbf{y}^*$ in computing $\mathbb{E}[\log p(\mathbf{x}, \mathbf{y}|\Theta)|\mathbf{x}, \Theta^g]$ with the mean-field approximation.

We use the Graph Cuts algorithm for the first stage (Equation 26) and use the EM algorithm for the second stage (Equation 27). The energy minimization problem for MRF (Equation 26) is known to be NP-hard (Veksler, 1999). Graph Cuts algorithms can effectively seek approximate solutions to the energy minimization problem. We define the unary cost and the pairwise cost of the graph $\mathcal{G}$ of the HMRF to utilize the Graph Cuts algorithm. The unary cost corresponds to the unary potential, which is:

$$U(x_i|y_i, \Theta^g) \propto -\log(p(x_i|y_i, \Theta^g)). \tag{Equation 28}$$

The pairwise cost corresponds to the pairwise potential. We compute the edge weights in $\mathcal{G}$ by calculating:

$$\overline{w}_{ij} = \exp\left(-\beta_1 \frac{||x_i - x_j||_2^2}{||x_i||_2 ||x_j||_2}\right), \tag{Equation 29}$$

where β_1 is an coefficient, and we use a pairwise state transition cost matrix $\overline{V} \in \mathbb{R}^{M \times M}$, where M is the number of hidden states and $\overline{V}_{ij}$ represents the penalty on $y_j \neq y_i$ for a directed edge $(i \leftarrow j)$. In our problem, $\mathcal{G}$ is an undirected graph, and we simplify $\overline{V}$ as $\overline{V}_{ij} = \beta_0, i, j = 1, \cdots, M$. However, in more complicated problem settings, we can realize $\overline{V}$ with varied elements $\overline{V}_{ij}$ and estimate the elements as model parameters. The calculation of the pairwise cost and edge weights is based on the second definition of the pairwise potential (Equation 11). We use the GCO library to perform the Graph Cuts algorithm (Boykov and Kolmogorov, 2004; Boykov et al., 2001; Kolmogorov and Zabih, 2004).

Approach to Generating the Simulated Datasets

In the simulation evaluation, we suppose that the samples in simulated dataset correspond to nodes in a graph $\mathcal{G}$. The graph $\mathcal{G}$ has 2D lattice structure of size $n \times n$, where each node is associated with a sample. Each node can be assigned 2D coordinates based on its position in the graph. Let $\mathcal{N}_i$ denote the set of neighboring nodes of the node i, i.e., the nodes that are connected to node i in $\mathcal{G}$. We use 8-connected neighborhood system. Suppose the node i has coordinates (c_{i_1}, c_{i_2}). Then the nodes with coordinates $(c_{i_1}, c_{i_2} \pm 1)$, $(c_{i_1} \pm 1, c_{i_2})$, $(c_{i_1} - 1, c_{i_2} \pm 1)$, and $(c_{i_1} + 1, c_{i_2} \pm 1)$ are the neighbors of node i.

In simulation study I, for each simulated dataset, we first generate a configuration of the hidden states of the samples by simulating an MRF through Gibbs sampling (Geman and Geman, 1984). We assume that the hidden state of each sample is associated with an emission probability function and the hidden states are from an MRF. We use the Markov property:

$$p(y_i|y_{-i}) = p(y_i|y_{\mathcal{N}_i}), \tag{Equation 30}$$

where y_{-i} represents the hidden states of all the nodes other than the i-th node, i.e., $y_{-i} = \{y_j, j \in \mathcal{V}, j \neq i\}$. $y_{\mathcal{N}_i}$ represents the hidden states of the neighbors of node i. We randomly initialize the hidden state configuration of the $N = n \times n$ samples at time step $t = 0$. The hidden state $y_i^{(t)}$ of sample i at time step t is sampled from the probabilistic distribution $p(y_i|y_{\mathcal{N}_i}^{(t-1)}), y_i \in S = \{1, \cdots, M\}$. $p(y_i = l|y_{\mathcal{N}_i}^{(t-1)})$ is calculated using Equation 10, $l \in S$. We use the first definition of pairwise potential and use $\beta_0 = 2$ (Equation 25). We repeat this sampling process until the maximum of time steps to take T is reached. We use $\mathbf{y}^{(T)} = \{y_i^{(T)}\}_{i \in \mathcal{V}}$ as the hidden states of the samples. We then simulate observations of the samples based on the hidden states using the emission probability functions $p(x_i|y_i, \theta_{y_i})$, where θ_{y_i} represents parameters of the emission probability function of hidden state y_i. We assume that the emission probability function of each hidden state is a multivariate Gaussian distribution. Suppose the observations are $x_i \in \mathbb{R}^d, i \in \mathcal{V}$. Let d be the number of species. x_i then represents the multi-species observations. We assume that each of the Gaussian distributions is associated with a different OU model. For each hidden state, we randomly sample the OU model parameters selection strength and Brownian motion intensity on each branch from uniform distribution Unif[0, 1], and sample the optimal values from normal distribution $\mathcal{N}(0.5, 0.25)$. We then use OU model parameters to calculate the Gaussian distribution parameters $\theta_l, l \in S$, and simulate samples based on the hidden states and the corresponding multivariate Gaussian distributions $p(x_i|y_i, \theta_{y_i})$. We use $n = 500$, $N = 250{,}000$, $d = 4$, $M = 10$, and use the same topology of the phylogenetic tree as we use in real data analysis for the OU models that are associated with the Gaussian distributions. In simulation study II, we use the same eight sets of hidden states simulated in simulation study I, but we simulate OU model parameters with different parameter settings. We randomly sample the OU model parameters selection strength and Brownian motion intensity on each branch from uniform distribution Unif[0, 1.5] and sample the optimal values from normal distribution $\mathcal{N}(1, 0.5)$. We also use $n = 500$, $N = 250{,}000$, $d = 4$, $M = 10$.

We assume that the data simulation process is hidden from us. When we applied Phylo-HMRF to the simulated datasets, we used the second definition of the pairwise potential by considering the feature difference of adjacent nodes (Equation 11), and used $\beta_0 = 1$, $\beta_1 = 0.1$. Therefore, the parameter settings we used to implement HMRF-EM are different from the simulation parameter settings, which can better evaluate whether the model has robust capability. We also tried varied parameters $\beta_0 = 1.5$, and $\beta_0 = 2$ and tested the performance of Phylo-HMRF. We found that Phylo-HMRF still maintains higher accuracy than the other methods and the performance is improved by a moderate level, demonstrating the robustness of Phylo-HMRF. We only report the results obtained with $\beta_0 = 1$ in the performance evaluation.

Cell Culture and Hi-C Data Generation

For the species chimpanzee, bonobo and gorilla, the lymphoblastoid cells were grown as suspension cultures in RPMI-1640 media supplemented with 15% fetal bovine serum and Penicillin/Streptomycin at 37°C. Cells were fixed in suspension culture at density of 1 million cells/mL with 1% formaldehyde for 10 minutes at room temperature. Fixation was quenched with glycine, and cells were harvested by centrifugation, washed with 1X DPBS, and frozen in aliquots of 5 million cells until ready for further processing.

Hi-C libraries were prepared using the original dilution Hi-C protocol (Lieberman-Aiden et al., 2009) and using the HindIII restriction enzyme. In brief, nuclei were isolated by performing gentle cell lysis on ice followed by 10 strokes with a Dounce homogenizer. Nuclei were washed in lysis buffer and resuspended in NEB Buffer 2, and digested with 400U of HindIII restriction enzyme overnight. On the following day, the restriction enzyme was heat inactivated and ends were filled in with nucleotides including a biotin-14-dCTP and blunt end ligated for 4 hours. The samples were then subject to overnight proteinase K treatment and reverse cross-linking. DNA was purified using phenol:chloroform extraction and ethanol precipitation. Libraries were prepared using Illumina TruSeq adaptors, and biotin containing fragments were isolated using streptavidin coated beads. Libraries were amplified using on-bead PCR amplification and purified using SPRI beads. Paired-end sequencing was performed to generate the Hi-C data for each species.

Cross-Species Hi-C Data Processing

We used the Hi-C data from the lymphoblastoid cells in human (GM12878) from Rao et al. (2014) and generated Hi-C data in lymphoblastoid cells in chimpanzee, bonobo, and gorilla. The genome assemblies used for the four species are hg38, panTro5, panPan2, and gorGor4, respectively. The genome assemblies were downloaded from the UCSC genome browser. We used Juicer (Durand et al., 2016) to process the Hi-C sequencing reads of each of the three species to obtain the Hi-C contact pairs based on the corresponding genome assembly. Each Hi-C contact pair is a pair of reads that are mapped to two genomic loci based on the corresponding genome assembly, representing chromatin contact between these two genomic loci. For the human GM12878 data, the Hi-C contacts file resulted from merging and processing all replicates has much higher coverage than the data for the other primate species, affecting the comparability of the Hi-C contact maps between human and the other species. We therefore performed random sampling of merged and processed data of five technical replicates of the same biological replicate in GM12878 cells to obtain approximately 2.9×10^8 Hi-C contact pairs in human, comparable to the other species. We obtained approximately 2.9×10^8, 2.7×10^8, 2.4×10^8, and 2.9×10^8 Hi-C contact pairs for human, chimpanzee, bonobo, and gorilla, respectively.

Next, we aligned the Hi-C contact pairs of the non-human species to the human genome. We mapped the aligned loci of the two ends of a contact pair in the Hi-C data of the non-human species from the original genome assembly to the human genome with reciprocal mapping using the tool liftOver (Hinrichs et al., 2006). With this conversion, the Hi-C contact maps of the four species that are computed from the aligned Hi-C contacts are all based on the human genome coordinates and therefore can be directly compared in the synteny blocks across species. The synteny blocks are the genome regions where the order of genome loci is preserved and there are no chromosome rearrangements greater than a certain resolution (50 kb in this study).

We used Juicer Tools to extract the Hi-C contact maps of each species at the resolution of 50Kb from the .hic files of the corresponding species, performing normalization by Knight-Ruiz matrix balancing (Rao et al., 2014; Durand et al., 2016). For the Hi-C contact map of each species in each synteny block, we perform two-step filtering to interpolate the missing values and smooth the signals. In the first step we use the median filter for interpolation of possible missing values. For each node without signal value in the Hi-C contact map, we use the median of Hi-C contact frequencies of the 8-connected neighbors as the value assigned to the node. Median filter has the characteristic to preserve edges in image. In the second step, we apply an anisotropic diffusion filter (Perona and Malik, 1990), which is an edge-preserving filter, to the whole Hi-C contact map in this synteny block to smooth the signals while maintaining the edge features, which correspond to more rapid changes of Hi-C contact frequencies. After preprocessing the Hi-C map of each species, we then align the Hi-C contact maps of the four species in each synteny block to obtain a combined multi-species Hi-C contact map, where each node in the map corresponds to Hi-C contact frequencies between the corresponding pair of genome loci in the four species. Hence each node is associated with a multi-dimensional feature vector as the multi-species observation. As the scales of Hi-C contact frequencies in different species are different, we normalize the Hi-C contact frequencies of each species to the same scale over all the autosomes using feature scaling. We then perform the $\widehat{x} = \log(1 + x)$ transformation to the normalized Hi-C contact signals of each species.

We focus on the comparison of Hi-C data in synteny blocks across species in this study. Therefore, we use the Hi-C contact signals within the synteny blocks as input to Phylo-HMRF, which are the subgraphs of the combined multi-species Hi-C contact map of each chromosome. The multi-species Hi-C contact map of each synteny block is symmetric. We therefore only keep the upper triangular part of the Hi-C contact map, and consider each entry as a sample.

Each sample is a multi-dimensional feature vector, where each dimension represents the Hi-C contact frequency of the corresponding species between the corresponding paired genomic loci specified by the coordinates of the entry in the Hi-C contact map. Hence, there are $N(N+1)/2$ samples for a synteny block of size N. We originally identified 90 synteny blocks in 50kb resolution on the autosomes based on inferCARs (Ma et al., 2006). There are two large size synteny blocks on chromosome 3 and chromosome 6, which exceeds 150Mb and 190Mb each. We then divide the two large synteny blocks into two parts each according to the two chromosome arms, respectively. For each divided synteny block, we still consider the interactions between the two subregions. Overall, we have 92 synteny blocks ($\geq$2.5Mb in size each) identified in the autosomes with 30,154,205 samples.

Initial Estimation of the Number of States for Phylo-HMRF

We estimated the possible number of hidden states using the K-means clustering before applying Phylo-HMRF to the cross-species Hi-C data. We performed K-means clustering using the scikit-learn library (Pedregosa et al., 2011) on the cross-species Hi-C data with the cluster number K increased from 2 to 100. We computed the Sum of Squared Errors (SSE) of each clustering result, and observed how SSE changed with respect to the different choices of K by plotting the SSE-K curve (Figure S3). We found that the decreasing rate of SSE with respect to the increasing K slows down in the range of 15-30. Small fluctuation of the number of states around 30 does not result in significant reduction of SSE with respect to the increase of K. We therefore set the number of hidden states to be 30.

State Estimation on the Hi-C Data by Phylo-HMRF

The differences of Hi-C contact frequencies across species can be either resulted from genome rearrangements or other types of genome evolution. In this work, we specifically focus on changes within synteny blocks. For all the autosomes in the human genome, we run Phylo-HMRF jointly on the multiple synteny blocks of the chromosomes and identified possible different evolutionary patterns of the Hi-C contact frequencies across species in a genome-wide manner. When applying Phylo-HMRF to the Hi-C data, we use the second definition of the pairwise potential by considering the feature difference of adjacent nodes, and use $\beta_0 = 3$, $\beta_1 = 0.1$.

We applied Phylo-HMRF to the multi-species Hi-C data to predict 30 hidden states. We further categorize the 30 estimated hidden states into 13 groups, as described in the Results section. Based on the Hi-C contact frequency distributions in the four species in each estimated state, we identify the states with distinctively higher or lower Hi-C contact frequency values than other states in all the four species consistently as the C-high and C-low states, respectively. Specifically, for the C-high or C-low states, the median of chi-squared distances (Pele and Werman, 2010) between the Hi-C contact frequency signals of each pair of species is significantly smaller than expected by chance (empirical p *value*<5e−04) and smaller than the non-conserved states. Also, in the C-high states, Hi-C signals of at least 95% of the samples in each species are consistently larger than the 95% quantile of the Hi-C signal values in the corresponding species. In the C-low states, Hi-C signals of at least 90% of the samples in each species are consistently smaller than the 25% quantile of the Hi-C signal values in the corresponding species. For each species, we categorize the Hi-C signals into $n = 20$ equally spaced intervals by calculating the 5%-95% quantiles. For a given state, suppose $u = (u_1, \cdots, u_n)$, $v = (v_1, \cdots, v_n)$ represent the percentages of samples in each interval for two compared species, respectively. We calculate the chi-squared distance $\chi^2_d(u, v) = \frac{1}{2}\sum_{i=1}^{n}\frac{(u_i - v_i)^2}{u_i + v_i}$, which is a measure of similarity between two distributions. To estimate the empirical distribution, we calculate chi-squared distance between two randomly sampled distributions $\hat{u}$, $\hat{v}$ each time and repeat the process to calculate 10^5 distances. For the other states, the states showing similar feature distributions of Hi-C contact frequency in the four species are annotated as C-mid and WC. The rest states are annotated as non-conserved (NC) states and we further identify the lineage-specific states where one species shows divergence in feature distribution from the other species.

After applying Phylo-HMRF to the cross-species Hi-C data for hidden state estimation results, we obtained segmentation of the cross-species Hi-C contact map in synteny blocks, where each node is assigned a label that represents the estimated hidden state. Neighboring nodes with the same hidden state form a local segment. The segmentation results in synteny blocks can be visualized as color images. We then perform simple post-processing of the segmentation results to obtain more smoothed segmentation that facilitates downstream analysis.

For the segmentation resulted from state estimation in each synteny block of a chromosome, we consider it as an image and first find all the connected components in this image. Each connected component can be considered as a segment of the cross-species Hi-C contact map. Nodes in a connected component have the same estimated states. If the size of the connected component (i.e., the number of nodes in the segment) is smaller than a threshold, we query the states of all the external nodes in local neighborhood to any node in the component and use the most frequent observed state of the external neighbors to reassign states for this component. We set the threshold to be 10, and use window size of 5 to define local neighborhood surrounding a node. We applied this post-processing step once and obtained slightly smoothed segmentation results.

QUANTIFICATION AND STATISTICAL ANALYSIS

Performance Metrics in the Simulation Evaluation

We evaluated the accuracy of Phylo-HMRF in estimating hidden states by comparing the predicted states with the ground truth states in the simulation evaluation, 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). These metrics compare

two partitions of a set. Suppose $\mathbf{x} = \{x_1, \cdots, x_N\}$ are the samples. Suppose $\Omega = \{\omega_1, \cdots, \omega_K\}$ and $C = \{c_1, \cdots, c_M\}$ are the predicted partition of the samples and the ground truth partition of the samples, respectively. The mutual information (MI) between Ω and C is $I(\Omega; C) = \sum_{k=1}^{K}\sum_{j=1}^{M} P(\omega_k, c_j)\log\frac{P(\omega_k, c_j)}{P(\omega_k)P(c_j)}$. The NMI between Ω and C is:

$$NMI(\Omega; C) = \frac{I(\Omega; C)}{[H(\Omega) + H(C)]/2}, \quad \text{(Equation 31)}$$

where $H(\Omega)$ and $H(C)$ are the entropies of Ω and C, respectively. $H(\Omega) = -\sum_{k=1}^{K} P(\omega_k)\log P(\omega_k)$, $H(C) = -\sum_{j=1}^{M} P(c_j)\log P(c_j)$. $P(\omega_k)$ represents the probability that a sample is in partition ω_k. The maximum likelihood estimate of $P(\omega_k)$ is $|\omega_k|/N$, where $|\omega_k|$ denotes the size of ω_k and N is the sample size.

AMI corrects MI by removing the effect of agreement between two partitions that is due to chance. AMI is defined as:

$$AMI(\Omega; C) = \frac{I(\Omega; C) - \mathbb{E}[I(\Omega; C)]}{\max\{H(\Omega), H(C)\} - \mathbb{E}[I(\Omega; C)]}, \quad \text{(Equation 32)}$$

where $\mathbb{E}[I(\Omega; C)]$ represents the expectation of $I(\Omega; C)$. $\mathbb{E}(I(\Omega; C))$ can be estimated based on Ω and C (Vinh et al., 2010).

The Rand Index (RI) (Manning et al., 2008) also compares two partitions, which is defined as:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}, \quad \text{(Equation 33)}$$

where TP (true positive), FP (false positive), FN (false negative), and TN (true negative) represent the number of sample pairs that are in the same subset in Ω and also in the same subset in C, the number of sample pairs that are in the same subset in Ω but in different subsets in C, the number of sample pairs that are in different subsets in Ω but in the same subset in C, and the number of sample pairs that are in different subsets in Ω and also in different subsets in C, respectively.

ARI corrects RI by removing the effect of agreement between partitions that is due to chance:

$$ARI = \frac{RI - \mathbb{E}[RI]}{\max\{RI\} - \mathbb{E}[RI]}, \quad \text{(Equation 34)}$$

where $\mathbb{E}[RI]$ represents the expectation of RI.

Precision, Recall, and F_1 score are defined as:

$$Precision = \frac{TP}{TP + FP}, \quad \text{(Equation 35)}$$

$$Recall = \frac{TP}{TP + FN}, \quad \text{(Equation 36)}$$

$$F_1 = \frac{2Precision \times Recall}{Precision + Recall}. \quad \text{(Equation 37)}$$

Other methods compared in the simulation evaluation

In the simulation evaluation, we compared Phylo-HMRF with the Gaussian-HMRF method (Zhang et al., 2001), the Gaussian Mixture Model (GMM), the K-means clustering method, and two image segmentation methods SLIC (Achanta et al., 2012) and Quick Shift (Vedaldi and Soatto, 2008) in state estimation. To utilize the image segmentation methods, we consider the combined multi-species Hi-C contact map as an image, and consider the features of each species as one color channel of the image. We normalize the features of each species to be in the range [0,1] accordingly, which is the scale of a color channel, to prepare the input for SLIC and Quick Shift. For the segmentation results, we consider segments with the same label as the same state. SLIC performs K-means clustering in the joint space of color information and spatial coordinates over an image. Quick Shift is approximation of the Mean Shift algorithm (Comaniciu and Meer, 2002) with kernel methods utilized, performing mode seeking in segmenting an image. We use the scikit-learn library (Pedregosa et al., 2011) for implementation of the GMM method and the K-means clustering. We use the scikit-image library (Van der Walt et al., 2014) that includes implementation of the SLIC and Quick Shift algorithms. For the methods Gaussian-HMRF, GMM, and K-means clustering, we set the number of states to be 10, respectively, which is the number of ground truth states. For the two image segmentation methods implemented by scikit-image, there are no input arguments to set the exact number of output labels of the segments. We adjust the parameter configurations of each of the two methods such that the number of output labels is approximately 10 and comparable to the state estimation results of the other methods. For SLIC, we use an argument to set the approximate number of labels and adjust the other parameters to have the number of output labels approximating 10. For Quick Shift, there is no argument to set an exact or approximate number of output labels. We then tune the input parameters to have the number of output labels approximating 10.

Alignment between boundaries of identified local-contact block patterns and TADs

For segments of estimated Hi-C evolutionary patterns along the diagonal of the Hi-C contact map of a synteny block, we use windows (squares) that can match the segments to identify the local-contact block patterns. A segment is a continuous 2D region with the same estimated state. Specifically, we use a sliding window with changeable size to find possible matches to the segments on the diagonal. With a window of the lower bound size located at a starting position along the diagonal, we first find the dominant estimated state within this window. If there is no dominant state, we move the window to the next position by one bin. The dominant state is defined as the state with the percentage exceeding a threshold and with the highest percentage within the window. If there is a dominant state, we then increase the window size from the lower bound gradually until the percentage of the dominant state within the window decreases or the dominant state changes or disappears. If a window has the highest percentage of the dominant state, and the percentage reaches the threshold (we use 0.95), we identify it as a local-contact block. We then reset the window to the lower bound size and move it to the next position by a stride that is half of the previously identified block size. We repeat these steps until we scan all the estimated states along the diagonal of the Hi-C contact map. Since the Hi-C contact frequency was measured at a resolution of 50 Kb 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 the nearest TAD boundary is calculated in increments of 50 Kb (one bin) accordingly. We compute the percentages of the distance in five distance intervals, which are 0–50 Kb, 50–100Kb, 100-150Kb, 150-200Kb, and >200Kb, respectively.

In addition, we estimate the empirical distributions of the distance between boundaries of a possible diagonal block and the nearest TAD by randomly shuffling the identified diagonal blocks. For each synteny block, we shuffle the identified local-contact block patterns on the diagonal of the multi-species Hi-C contact map 1,000 times by randomly relocating them within this synteny block. For each boundary of each randomly relocated diagonal block in a shuffle, we calculate the distance between the block boundary and the nearest TAD boundary of a specific type of TAD (i.e., Arrowhead TAD or DI TAD). For each shuffle, we then compute the percentages of the distances between the diagonal block boundaries and the corresponding nearest TAD boundaries in the five distance intervals. We merge the percentages from each shuffle of diagonal block patterns as an empirical distribution for each distance interval.

Analysis of the Connection between Estimated Hi-C and RT Evolutionary Patterns

To compare the conservation of Hi-C states with replication timing (RT) conservation, we examined RT evolutionary states identified in our previous paper (Yang et al., 2018) using Repli-seq datasets from analogous lymphoblastoid cells across primate species. In that paper, we classified the evolutionary patterns of RT into five distinct categories reflecting the different levels of conservation, i.e., conserved early in RT (E), weakly conserved early (WE), conserved late (L), weakly conserved late (WL), and non-conserved (NC). With this data, we are able to test whether the evolutionary patterns of Hi-C data are correlated with RT evolutionary patterns. To do that, for each pair of genomic loci, we consider it has a matched RT state if the pairwise genomic loci present similar RT evolutionary patterns (i.e., both are E/WE or both are L/WL). The fraction of matched RT states thus indicates the concordance between conservation of RT and conservation of interactions of pairwise genomic loci. We remove the distance confounding factors by comparing the fraction of matched RT states across different Hi-C states over a range of different distances (0-10Mb). To further confirm our observation is not due to the randomness of RT evolutionary state calls, we attempt to repeat this analysis by using randomly shuffled RT evolutionary states. In each chromosome, we first merge adjacent RT evolutionary states (in a resolution of 6kb) into longer segments. We then randomly shuffle the labels of the RT evolutionary states in each chromosome so that the global distribution of RT evolutionary patterns remains the same. Finally, we repeat plotting the distributions of pairwise genomic loci with matched RT states over a range of different distances (0-10Mb) (Figure S7). We observed that the matched-RT state curves in different Hi-C contact states based on the shuffled RT states did not exhibit the trends, changing points, and diversities of the curves for different Hi-C contact states based on the original predicted RT states.

Estimating Background Distribution of Histone Modification Similarity

To compare the histone modification composition for each pair of genomic loci with estimated Hi-C states, we computed the percentage of paired genomic loci that have more similar histone modification signal strength than expected in the C-high, C-mid, WC, C-low, and NC states over a range of different distances (0-10Mb). In order to get the background distributions of quantile changes between paired genomic loci, we randomly select 1,000 pairwise bins on the genome and calculate the absolute value of quantile differences of the histone modification signal strengths for each pair. We repeat this process for 20 times and the combined dataset is considered as the background distribution. Finally, we compare the observed difference of quantiles between pairwise bins with the background distribution. In this study, we consider that two genomic loci tend to have more similar histone modification signals if the differences of quantiles are smaller than the median value calculated from the background distribution.

Detecting Conserved Long-Range Interacting TADs

In order to study the evolutionary patterns on TADs and related genomic and epigenomic features, we classified the TADs into two groups based on the whether a TAD is involved in conserved long-range TAD-TAD interactions. We showed examples of long-range TAD-TAD interactions that are conserved across species in the synteny block 8 on chromosome 1 of human in Figure 4F. We identified the conserved long-range TAD-TAD interactions in all the synteny blocks across 22 autosomes in human based on the Hi-C evolutionary contact states estimated by Phylo-HMRF. We use the DI TAD annotations in GM12878 cell line. For each pair of DI

TADs in the same synteny region on human genome that are at least 3Mb apart, we calculate the percentages of different estimated Hi-C contact states in the block of the 2D multi-species Hi-C contact map which corresponds to possible interactions between the two TADs. The block is a 2D region in the Hi-C contact map, where each node represents the Hi-C contact frequency between a genome loci in one TAD and a genome loci in another TAD. For each pair of TADs that are at least 3Mb apart, if the total of percentages of C-high states (S4, S12) and C-middle states with relatively higher Hi-C contact frequencies (S15, S16) in the block exceeds 50%, and the state with the highest percentage is one of the four states, we consider it as conserved long-range TAD-TAD interaction. There are 3,541 TADs that are located in the synteny regions of 22 automosomes in human. We detected 3,365 pairs of TADs that are associated with conserved long-range interactions. We observed that for 44.46%, 41.84%, and 13.70% of the 3,365 TAD pairs with conserved long-range interactions, the distance between the paired TADs are in the ranges 3-5Mb, 5-10Mb, and larger than 10Mb, respectively. To define the long-range interacting TADs, we consider the conserved long-range TAD-TAD interactions between TADs that are more than 10Mb apart in 1D genome distance. We label a TAD as conserved long-range interacting TAD (hereafter abbreviated as conserved TAD for simplicity in related discussions) if it is involved in conserved long-range interaction with a TAD that is more than 10Mb away. Otherwise a TAD is labeled as non-conserved long-range interacting TAD (hereafter abbreviated as non-conserved TAD).

Analysis of Sequence Features in Evolutionary Patterns of TADs

We sought to explore the potential connections between transposable elements (TE) with different Hi-C contact evolutionary states estimated by Phylo-HMRF. TEs are known to be associated with genome organization (Rhind and Gilbert, 2013; Yang et al., 2018; Zhang et al., 2019). We first analyzed the enrichment of different TE families in each estimated Hi-C contact state across species. We use the RepeatMasker annotations of each of the four species human, chimpanzee, bonobo and gorilla retrieved from the UCSC Genome Browser (Casper et al., 2018) to compute the TE enrichment in each species in each estimated state. The positions of TEs on the genome of non-human species are mapped to the human genome using liftOver. For each state, we calculate the coverage of a TE family in all the paired genomic loci of the state in each species. We then normalize the TE family coverage by the total size of all the paired genomic loci of this state for each species, which is denoted as normalized TE enrichment of the corresponding species in this state. We also calculate the background TE enrichment for each species by normalizing the total coverage of each TE family of each species in all the paired genomic loci of all the states with the total size of all the paired genomic loci of all the states. We compute the fold change of the normalized TE enrichment of each species in each state with respect to the corresponding background enrichment. We use chi-squared test to assess whether a TE enrichment is lineage-specific in a lineage-specific state. We identified several TE families that show human-specific enrichment patterns in the identified human-specific state (Figure S10), including PIF-Harbinger, hAT-Tag1, and TcMAr. For example, the TE family PIF-Harbinger has significantly higher enrichment in human while lower enrichment in the other species in the human-specific high state S14 (NC-hom_high) (FDR (False Discovery Rate) <0.01. FDR is calculated by adjusting *p value* from the chi-squared test using Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995)).

Next, we examined the potential connections between TEs and evolutionary patterns in TADs. As we described, we classify TADs into conserved long-range interacting TADs (abbreviated as conserved TADs) or non-conserved long-range interacting TADs (abbreviated as non-conserved TADs) according to whether a TAD is involved in conserved long-range TAD-TAD interactions. For each TE family in human, we calculate their occurrence frequencies in the two classes of TADs we have defined. For single TADs, the occurrence frequency of a TE family in each TAD is normalized by the length of the TAD. For each TAD class, the total occurrence frequency of a TE family in this class of TADs is normalized by the total length of this class of TADs, which is denoted as normalized TE enrichment for this TAD class. We calculate the fold change of the normalized TE enrichment in each TAD class with respect to the background enrichment. The background enrichment is calculated by normalizing the total occurrence frequency of a TE family in all the TADs with the total length of the TADs. We then use chi-squared test to assess whether a TE family is enriched in the conserved TADs. We found that multiple TE families show distinct enrichment patterns in the conserved TADs compared to the background. 10 TE families are significantly more enriched in the conserved TADs (p *value*<0.01), and 6 TE families are significantly less enriched in the conserved TADs (p *value*<0.01), respectively, as shown in Figures S11A and S11B. For example, MIR, hAT, and Helitron are among the more enriched TE families, and ERV1, ERVK, and RTE-BovB are among the less enriched TE families. Zhang et al. (2019) revealed that HERV-H, which is a subfamily of ERV1, has important roles in forming human-specific TADs specifically in human pluripotent stem cells, suggesting that they are likely to be less enriched in conserved TADs.

Additionally, we sought to explore whether there are connections between TADs with different evolutionary patterns and the transcription factor binding sites (TFBSs). We identify open chromatin regions in the human genome as GM12878 DNase-seq peak regions (downloaded from the ENCODE project (ENCODE Project Consortium, 2012)) with +/-250-bp extension. We used FIMO (Grant et al., 2011) to scan motifs based on the 579 PWMs of TF binding motifs from JASPAR (Khan et al., 2017) (with p *value*<1e−04 as cutoff) in the open chromatin regions in human. Here we only retain the motif scanning results for the 345 TFs that are expressed in GM12878 (with FPKM>0.1; the RNA-seq datasets used for gene expression analysis are shown in the Key Resources Table). We compute the frequency of each TF binding motif within the open chromatin area in each TAD. We then normalize the frequency by the open chromatin area size within this TAD, which is denoted as normalized TF motif enrichment in this TAD. We also calculate the fold change of the normalized motif enrichment in each TAD class with respect to the background enrichment. The normalized motif enrichment in each TAD class is calculated by normalizing occurrence frequency of a motif in the open chromatin regions in this class of TADs with the total length of open chromatin regions in this class of TADs. The background enrichment is calculated by

normalizing the total occurrence frequency of a motif in the open chromatin regions in all the TADs with the total length of open chromatin regions in all the TADs. We use chi-squared test to assess if a motif is enriched in the conserved TADs. We identified a set of TF binding motifs that show distinct enrichment patterns in the conserved TADs compared to the background (Figures S11C and S11D). We found multiple TF binding motifs that are significantly more enriched in the conserved TADs (p *value*<0.01). For example, the motifs of ASCL1, NFATC2, FOXD2, and FOXO4 are among the more enriched motifs. We also found TF binding motifs that are significantly less enriched in the conserved TADs (p *value*<0.01), such as motifs of FOSL1::JUND, MAFF, and CREB1. These results show that there are differences in terms of TF binding motif enrichment in conserved and non-conserved TADs.

DATA AND SOFTWARE AVAILABILITY

The accession number of the Hi-C datasets of chimpanzee, bonobo, and gorilla reported in this paper is GEO: GSE128800. The source code of Phylo-HMRF can be accessed at: https://github.com/ma-compbio/Phylo-HMRF.