Computational Methods for Multi-Species Comparison of 3D Genome Organization and Function

Yang Yang

CMU-CB-20-103
December 2020

Computational Biology Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Jian Ma, Chair
Ziv Bar-Joseph
Anne-Ruxandra Carvunis
David Haussler

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Copyright © 2020 Yang Yang

This material is based upon work supported by the National Science Foundation under grant numbers DBI1619983 and IIS1619878, the National Human Genome Research Institute of the National Institutes of Health under award number R01HG007352, the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number U01HL145793, and the NIH National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under award number U54DK107965. The department specifically disclaims responsibility for any analyses, interpretations, or conclusions. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation and the National Institute of Health and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, donor or the U.S. Government.

Keywords: 3D genome organization, DNA replication timing, Ornstein-Uhlenbeck process, Gaussian process, hidden Markov model, hidden Markov random field, interpretable machine learning

Abstract

Recent developments in chromatin interaction mapping technologies have greatly advanced the study of higher-order genome organization in the three-dimensional (3D) cell nucleus, which is of vital importance to fundamental genome functions such as DNA replication timing (RT) and gene transcription. However, the principles underlying the 3D genome organization and function and the detailed patterns on how the 3D genome has changed in mammalian evolution remain largely unclear. To directly address these questions, in this Ph.D. dissertation, I developed new machine learning frameworks to advance the methodologies for the comparisons of 3D genome organization across multiple species and for unveiling critical information encoded in the genome that may regulate large-scale chromosome structure and function. First, I developed a new model named phylogenetic hidden Markov Gaussian processes (Phylo-HMGP) to simultaneously infer genome-wide heterogeneous evolutionary patterns of continuous-trait functional genomic features. Phylo-HMGP models both temporal dependencies across species and spatial dependencies along the genome. Real data application to a new RT dataset based on Repli-seq from five primate species demonstrated that Phylo-HMGP greatly refined our understanding of cross-species RT patterns. Next, I developed a new probabilistic model named Phylo-HMRF, a unique framework to compare multi-species 3D genome organizations based on Hi-C data. The method incorporates 3D spatial constraints with continuous-trait evolutionary models. Phylo-HMRF uncovered patterns of 3D genome evolution in primate species that show novel connections to other genome structural and functional features. Finally, I developed a generic interpretable machine learning framework named CONCERT to predict large-scale chromosome domain features with a focus on RT profile directly from genomic sequences. CONCERT enables the identification of genome-wide sequence elements that modulate the RT program by jointly performing predictive element estimation and long-range spatial dependency learning. Application of CONCERT to multiple human and mouse cell types demonstrated the effectiveness of the method. Taken together, the methods developed in this Ph.D. dissertation have established a series of new algorithmic formulations

for effective comparison and high-resolution characterization of evolutionary patterns of 3D genome organization, revealing genomic regions with conserved or species-specific structural and functional roles. The methods have the potential to provide critical insights into the regulatory principles of nuclear structures and the sequence determinants underlying the strongly intertwined nature of genome structure and function.

Acknowledgement

I would like to thank my advisor Jian Ma for his consistent support and guidance during my PhD study. Jian has been helpful and responsive in providing advice to me in my study and research. Jian has shown me the necessity of gaining domain knowledge from different fields to be engaged in interdisciplinary research.

I would like to thank Professor Ziv Bar-Joseph, Professor Anne-Ruxandra Carvunis, and Professor David Haussler for serving on my Thesis Committee, and for providing feedback and suggestions on my thesis research.

I would like to thank my collaborators: Professor David Gilbert, Professor Quanquan Gu, Dr. Takayo Sasaki, Professor Jesse Dixon, Professor Bing Ren, Julianna Crivello, and Professor Rachel J. O'Neill. I truly appreciate the collaborative opportunities during my PhD study.

I also appreciate very much the opportunities to attend different conferences and workshops during my PhD study, through which I broadened my horizons by communicating with many other researchers in computational biology and machine learning.

I would like to thank all of our present and former group members whom I have worked with. I would like to thank Yang Zhang, Yuchuan Wang, Ruochi Zhang, Shashank Singh, Benjamin Chidester, Dechao Tian, Tianming Zhou, Xiaopeng Zhu, Kyle Xiong, Wendy Yang, Weihua Pan, and Ashok Rajaraman. I am thankful that I learnt a lot from our group members, and I truly appreciate the generous help they gave to me.

I would like to thank my cousins Huan, Shasha, and Jun, who always care about me and give me encouragement. I would like to thank my father and mother, for their love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and related work

In human and other eukaryotes cells, chromosomes are folded and organized in the three-dimensional (3D) space in the cell nucleus and different chromosomal loci interact with each other [1–3]. Recent development in whole-genome mapping approaches for chromatin interactome such as Hi-C [1, 4], ChIA-PET [5], GAM [6], and SPRITE [7] has facilitated the identification of genome-wide chromatin organizations comprehensively, revealing important 3D genome features such as chromatin loops [4, 5], topologically associating domains (TADs) [8], and A/B compartments [1].

3D genome organization, including 3D chromatin structures and positioning of the chromosomes in the cell nucleus, is known to have important roles in gene regulation and genome functions [2]. 3D genome organization is closely related to vital genome functions such as DNA replication timing (RT) and transcription [9–11]. For example, chromatin loops may enable long-range nhancer-promoter interactions that are important for gene expression control [12]. Disruption of 3D genome organization may alter normal gene regulation and cause diseases [13–28]. Understanding the regulatory mechanism of 3D genome organization is important for understanding both the coordination of normal genome functions and the mechanisms of specific diseases that are related to the change of chromatin structures. However, our knowledge of the principles underlying 3D genome organization is surprisingly limited. A critical challenge is to fundamentally decode the

instructions encoded in our genome that govern genome organization and function. How to connect mutations in DNA sequences to the changes in genome organization and ultimately changes in phenotype is a fundamental problem in understanding the regulatory principles rooted in the DNA sequences and also a significantly under-explored area.

From another perspective, comparative genomics can provides key insights into the study on 3D genome. Comparative genomics focuses on the comparison of genomic or epigenomic features across different species, which provides us with a new temporal angle to study the relationships between genome functions and genomic or epigenomic features in the context of evolution. Algorithms in comparative genomics have been powerful to reveal potential molecular signatures for phenotypic differences between human and other species, and inform us about human genome functions [29–34]. With the availability of Hi-C data, there is the new direction on comparing genome organizations across different species. By identifying and analyzing conserved and non-conserved patterns of genome organization across species, we may gain more understanding on the relationship between genome organization and genome functions, and also how genome organization is regulated. For example, conserved genome organization patterns across different species may indicate the existence of important genome functions that have been preserved during evolution. Species-specific genome organization patterns are potentially correlated with species-specific gene regulatory functions, and may reveal clues on how underlying genomic sequence features are correlated with the specific patterns. Comparative study of 3D genome organization across species presents a new possible sub-area of the current comparative genomics studies, and may help us gain more advanced knowledge on human genome functions and related regulation mechanisms beyond the existing knowledge from the prior studies on the 1D genomic or epigenomic features. However, there have been very limited studies on the comparison of genome organization across different species [8, 35, 36].

3D genome structures are also known to be closely correlated with genome functions [2], including DNA replication timing. DNA replication, which duplicates the DNA on chromosomes during cell division, is one of the most fundamental functions of eukary-

otic cells. DNA replication timing represents the temporal order of DNA replication along the genome, which is highly regulated and robust [10, 11, 37]. Using Repli-seq [38] technology we can obtain DNA RT measurements, which can be processed as one-dimensional (1D) coninous genomic signals along the genome. Genome-wide RT maps have revealed replication profile domains that correlate with chromatin structure [39, 40] and higher order genome organization such as A/B compartments and TADs [9, 37, 41–43]. However, we still have limited understanding of how the RT program has evolved in recent primate species and its detailed connection to 3D genome evolution.

On the algorithmic side, comparative genomic algorithms for cross-species comparisons of 3D genome or the related functional genomic features have been under-explored. Multi-species functional genomic data from various high-throughput epigenomic assays (e.g., Hi-C, Repli-seq [38], ChIP-seq [44–46]) are continuous in nature. However, in prior studies, such continuous signals are usually discretized by selected thresholds or transformed into discrete values to identify distinctive feature patterns (e.g., presence or absence of TADs, presence or absence of peaks or domains) for subsequent comparisons [8, 35, 36, 47], which potentially cause dramatic loss of information of more subtle differences from the original data.

Also, there are new computational challenges in using the high-throughput sequencing data of 3D genome for comparative pattern recognition across species. The Hi-C data are usually 2D data presented by Hi-C contact maps (section 1.3.1, Figure 1.1B). The prior methods on comparative analysis of genomic features are mostly developed for the 1D genomic data, not applicable to 2D data. Moreover, for Hi-C data, there were no existing phylogenetic model-based methods available to analyze Hi-C data as continuous signals across different species to uncover genome-wide evolutionary patterns of 3D genome organization. For 1D functional genomic data that are correlated with 3D genome organization (e.g., Repli-seq data), there were also no phylogenetic model-based algorithms to compare continuous-trait functional genomic data across multiple species in a genome-wide manner. Furthermore, our knowledge on the DNA sequence-based regulatory principles of

genome organization and related genome functions such as DNA replication timing is still primitive [2, 3, 10].

This thesis is focused on method development to address the computational challenges of two tasks: (i) comparative pattern recognition based on continuous-trait functional genomic data and 3D genome organization data from different species; (ii) identification of DNA sequence elements that are potentially important for the regulation of genome organization and genome function patterns.

1.1.1 Continuous-trait functional genomic data comparison across multiple species

Multi-species functional genomic data (e.g., ChIP-seq data of transcription factors or histone marks, Repli-seq data) are highly informative for the comparative analysis of genome function conservation and differences between human and other mammalian species [48–51]. However, as we discussed, computational models are under-explored to fully model the continuous-trait functional genomic data in the context of multi-species comparisons. Continuous-trait models, which are key to the modeling of functional genomic signals, are gaining increasing attention in genome-wide comparative genomic studies [52, 53].

Several types of continuous-trait evolutionary models have been developed for comparison at individual loci. One basic model [54–56] assumes that continuous traits evolve by Brownian motion. This model has been extended to more complicated Gaussian processes such as the Ornstein-Uhlenbeck (OU) process [57–59]. However, the existing methods that use continuous-trait evolutionary models in comparative genomics either apply a single evolutionary model to signals of selected regions, or test different evolutionary model assumptions with prior knowledge at selected regions [50, 52, 53, 60, 61]. In other words, the continuous-trait evolutionary models have not been utilized to simultaneously estimate heterogeneous phylogenetic trees across different loci along the entire genome based on functional genomic data.

There are 1D functional genomic data (e.g., Repli-seq data and gene expression data) and 2D data (e.g., Hi-C data) which measure chromatin interactions of genome organiza-

tion. In this thesis we begin with comparative analysis of 1D functional genomic features that are closely correlated with 3D genome organization. As we have introduced, one important type of genome function that is connected with genome organization is the DNA RT program.

It is known that RT changes across half of the genome during cell differentiation and is altered in certain diseases [9, 62–67]. In addition, studies have shown conservation of RT between several eukaryotic species, including the conservation between human and mouse [41, 42, 63, 68, 69]. Microscopy studies revealed that chromosome regions with early and late RT have specific spatial localization preferences in the nucleus that are conserved in evolution [70]. However, we have limited understanding of how the RT program has evolved in mammals. Prior to our proposed work, to the best of our knowledge, there were no existing algorithms available to simultaneously infer heterogeneous continuous-trait evolutionary models along the entire genome, and there was no existing study to investigate the RT conservation and dynamics for more than two mammalian species beyond the human-mouse comparison.

1.1.2 Genome organization comparison across multiple species

In the existing studies, a limited number of attempts have been made to compare the 3D genome organization across different species. An earlier work using Hi-C showed that the positions of TADs were largely conserved between human and mouse within syntenic genomic regions [8]. Another study demonstrated that evolutionary changes in TAD structure correspond with the creation or elimination of CTCF binding sites using relatively low-resolution Hi-C data from rhesus macaque, dog, rabbit, and mouse [35]. More recently, TADs have been shown to have strong conservation in mammalian evolution with the TAD boundaries under potential negative selections against genome rearrangements [36, 71]. 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. Importantly, as most of the initial comparative analysis of 3D genomes focused primarily on distantly related or-

ganisms, there is limited understanding of how 3D genome features may have evolved in closely related mammalian species, especially in recent primate evolution which is of particular interest to understand human-specific and great ape-specific gene regulations.

The previous methods for genome organization comparison across species usually only utilize the discrete features from Hi-C data (e.g., comparing the presence or positions of TADs) [8, 36, 71, 72]. Although the strengths of chromatin interactions can be quantitatively analyzed from Hi-C data [1, 4], previous studies on comparing genome organization across species did not explicitly consider the continuous nature of chromatin interaction strength, and therefore are limited in the ability of fully utilizing the data of chromatin conformation capture to reveal detailed evolutionary patterns of genome organization. Also, the previous methods did not consider the phylogeny of the compared species. Prior to our work, to the best of our knowledge, there were no existing phylogenetic-model based methods available to analyze Hi-C data as continuous signals to compare 3D genome organization across different species in a genome-wide manner.

1.1.3 Identification of sequence elements that modulate DNA replication timing

3D genome organization is closely correlated with DNA replication timing. As we introduced, the RT program has a highly regulated genome-wide temporal pattern [10, 11, 37]. High-throughput Repli-seq together with 3D genome organization features derived from Hi-C have shown that active domains towards the nuclear interior generally have early RT patterns, which are also evolutionary conserved [10, 42]. However, the regulators of DNA RT in eukaryotic cells and the mechansims of its relationship with 3D genome organization have long been unclear, remaining open questions.

Different hypotheses exist on the potential regulation mechanisms of DNA RT, including both epigenetic regulation mechanisms and sequence-dependent regulation mechasisms [10]. Studies have showns that certain types of histone modifications such as H3K4me1, H3K4me2, H3K4me3, H3K20me1, H3K36me3, H3K9ac, and H3K27ac correlate with early RT [63]. But correlations may not suggest casual effects [73]. An explanation for

some of the correlations is that for histone modifications established quickly during chromatin maturation, their density detected by chromatin immunoprecipitation can be proportional to the copy number of each locus and thus match the RT profile [10]. Histone acetylation is found to be possibly involved in regulating early RT, but histone acetylation alone is not sufficient for the RT regulation [74, 75]. [You need to add conclusion sentences in your paragraphs when you write. It's unclear what this paragraph is all about. I don't think the discussion on histone acetylation etc. is that relevant to make your major point.]

For the hypotheses on sequence-dependent regulation mechanisms, different technologies have been developed or applied to generate genome-wide maps of RT initiation sites or active RT origins in mammalian cells with different types of sequence elements as sequencing targets [76–81]. The types of sequence features that are found to be associated with RT regulation include G-quadruplexes [76, 79, 82] and high GC density or asymmetry of nuecleotide distribution [83, 84]. However, it is observed that different features can be found at different RT origins [85]. The RT origin-specific sequence was only discovered in the budding yeast [86] and not identified in the other eukaryotes. Also, the relationship between RT origins and the RT program in mammalian cells is not explicit [9, 76]. It is inferred that any DNA sequence can be activated as RT origin in the proper context [87] and the RT is possibly regulated at the domain level instead of the origin level in mammalian cells. It was observed that factors bound nearby RT origins may also influence RT [64], suggesting the potential importance of genomic context for RT regulation. Though existing studies show evidence that sequence-dependent RT regulation mechanisms may exist, there are still no clearly revealed rules on the functions of DNA sequences for RT regulation.

There have been very limited discoveries of sequence elements that may regulate the RT program. Studies have shown evidence that specific types of long non-coding RNAs (lncRNAs) may have critical coordination roles for chromosome-wide RT patterns. One type is Xist, which is lncRNA expressed in only one X chromosome in female cells. Xist was found to be necessary for late replication of the inactive X chromosome in mice, but deletion of Xist in differentiated human cells resulted in even later replication [88, 89]. The

other two types are ASAR6 and ASAR15, which are ASARs (asynchronous replication and autosomal RNAs) that coat chromosome 6 and chromosome 15 in human, respectively. It was discovered that ASAR6 and ASAR15 are essential for RT coordination and condensation of the entire chromosome from which they are expressed, respectively [90, 91]. Deletion of ASAR6 and ASAR15 resulted in delay of replication on the corresponding chromosomes [90–92]. However, the RT regulation mechanisms of Xist and the ASARs remain to be elucidated.

Recently, Sima et al. [75] made a significant step forward in identifying specific *cis*-acting elements regulating early RT in mouse embryonic stem cells (mESCs). Specifically, based on CRISPR-mediated experiments on a number of genomic loci, several early replication control elements (ERCEs) that regulate early RT in the mESCs [75] have been experimentally validated. The identified ERCEs are observed to be enriched with active epigenetic marks that are representative of enhancers, and harbor binding sites of major pluripotency transcription factors. It was also found that cooperative interactions of multiple ERCEs in 3D space are required for early RT at local domains. However, it remains elusive whether there are sequence properties that are predictive of the entire RT program genome-wide across different cell types.

On the other side, computational models have also been developed to predict DNA RT using different types of features. For example, models have been developed to use combination of histone modifications and chromatin binding protein data to predict DNA RT in *Drosophila* [93]. Methods have also been developed to predict RT in human cells based on chromatin accessibility [39]. However, the relationship between sequence features and DNA RT modulation remains mostly unrevealed [10] and models that can accurately predict detailed genome-wide RT signals from sequences are not available.

In the past few years, there have been a number of machine learning based models that use DNA sequences to predict genomic functional patterns. DeepBind [94] and DeepSEA [95] are models that use convolution neural networks (CNNs) to generate features from the sequences to predict transcription factor (TF) binding sites. DanQ [96]

extends the model with LSTM (long-short term memory model) [97], to capture the spatial dependencies between genomic bases within a sequence. These model architectures have been extensively utilized in other recent methods. For example, studies have used sequence features to predict enhancers, transcription initiation sites and gene expression [98–102]. The similar models were also used to predict enhancer-promoter interactions [103] and protein contacts [104]. Most of the existing methods, however, generate feature representations from a given genomic locus itself, or a locus with extensions, without utilizing information of the neighboring loci in the larger-scale context, which is critical to the prediction over large-scale genomic domains such as RT domains. Some methods incorporate flanking regions [103] of a genomic sequence for feature extraction, using the extended sequence to generate features. However, a longer extended sequence increases computational cost and the number of model parameters to learn.

Very recently, methods have been developed to use dilated convolutions to learn longer range spatial dependencies in DNA sequences for the prediction of epigenetic profiles, gene expressions, and 3D genome organization [105–107], utilizing the advantage that dilated convolutions use repeated weights for adjacent positions to expand the receptive field width of convolution kernels without increasing model complexity. For long-range spatial dependency, the receptive field width with dilated convolutions [108] is still limited in size compared to the sequence length. Dilated convolutions use regularly spaced gaps to capture spatial dependency, which may lose information on more complicated dependency patterns. Also, the existing methods require extra evaluation steps for model interpretation after the model is trained, and were not designed for simultaneously identifying the potential sequence determinants. Therefore, it remains an unsolved problem to capture dependency between two arbitrary loci over a very large-scale domain, such as RT domains, and achieve high model interpretability.

1.2 Thesis goals

This PhD thesis is focused on addressing the gaps in methodology development in comparative genomics for higher order genome organization features to reveal new insights about 3D genome structure and function. The overarching goal is to develop new models and frameworks for multi-species comparison of continuous genomic data with a focus on Hi-C and Repli-seq data, and develop methods to identify sequence elements that may modulate genome organization and related genome function, uncovering potential regulatory principles of 3D genome organization and function. Overview of the method development in this thesis is illustrated in Figure 1.3. Specifically, the thesis goals are as follows.

- **Develop algorithms to analyze continuous multi-species functional genomic data**.

  We develop a new continuous-trait probabilistic model based on the integration of the Ornstein-Uhlenbeck (OU) processes and the hidden Markov model (HMM) for more accurate evolutionary state estimation using continous-trait functional genomic data from different species, exploiting both evolution affinities among species and spatial dependencies along the genome.

- **Develop algorithms to study the evolution of 3D genome organization**.

  We develop an evolutionary probabilistic model based on the OU process and hidden Markov random field (HMRF) for modeling 3D genome organization across multiple species, to identify higher order evolutionary patterns of the genome structure. The developed method models both the temporal dependencies across the species in the context of evolution and the spatial dependencies across genomic loci in the 3D space.

- **Develop algorithms to identify sequence features that modulate genome organization related genome function**.

  We develop an interpretable predictive model based method to simultaneously predict DNA RT profile from genomic sequence features and identify genome-wide se-

quence elements that modulate the DNA RT program. The method integrates two cooperative modules, the selector and the predictor, to perform importance estimation-based predictive genomic loci selection and learn long-range spatial dependencies across genomic loci for RT signal prediction jointly. The method is applicable to study other types of genome organization related functional genomic data over large-scale spatial domains.

1.3 Introduction to main relevant biological technologies

Here we will briefly introduce the main relevant biological technologies involved in this thesis work. The data we focused on include Hi-C data and Repli-seq data. We will introduce the Hi-C technology [1, 4, 109] and the Repli-seq technology [38]. The detailed data information and data processing procedures are described in the corresponding chapters.

1.3.1 Hi-C technology

Hi-C [1, 4, 109] is the extension of the 3C (Chromatin Conformation Capture) technology [110]. Hi-C can identify long-range chromatin interactions in an unbiased genome-wide manner. In the procedures of Hi-C, cells are first fixed with formaldehyde. The interacting chromatin loci are bound to each other by covalent DNA-protein cross-links. Next, the DNA is fragmented with a restriction enzyme, with the bound chromatin loci remain linked. The 5' overhangs are filled in with nucleotides and a biotinylated residue is incorporated. Next, blunt-end ligation is performed under dilute conditions which facilitates ligation events between cross-linked DNA fragments. A genome-wide library of ligation products is produced, which correspond to pairs of fragments that were originally in close spatial proximity to each other in the cell nucleus. For each ligation product, the site of junction is marked with biotin. The library is purified and sheared. The junctions are isolated with streptavidin beads and then analyzed using paired-end high-throughput sequencing. A catalog of interacting fragments is identified, and further analysis can be performed.

The Hi-C data can be further processed and visualized as a two-dimensional (2D) contact matrix (or Hi-C contact map), where each row and each column represents a genomic locus, and each entry corresponds to the contact frequency between two genomic loci. The contact frequency reflects the distance between two genomic loci in the 3D space in the cell nucleus. Two genomic loci that are close to each other are more likely to have higher contact frequency while two distant loci are more likely to have lower contact frequency. We can also generate Hi-C data for different species. Comparing chromatin structures across species can help us identify conserved and non-conserved genome organization patterns, which may further reveal important genome functions that are correlated with or regulated by the chromatin structures. The earlier studies were focused on chromatin structure comparison between human and mouse [8]. More recently, there were Hi-C data generated from other mammalian species [35, 72]. The Hi-C data have been lacking for closely related primate species. In the previous works, important specific chromatin structure features (e.g., TADs or chromatin loops) were first extracted from the Hi-C data and then compared between the species [8, 35, 72].

1.3.2 Repli-seq technology

Repli-seq [38] is the extension of the Repli-chip technology [111] for DNA replication timing measurement. The cell cycle in eukaryotes commonly consists of two main phases: interphase and mitosis. The interphase can be further divided into three sub-phases: G1 phase, G2 phase, and S phase. DNA is replicated during the S phase. In the procedures of E/L Repli-seq (Early/Late Repli-seq) [38], the cultured cells are first pulse-labeled with antibody BrdU to label nascent DNA. Cells are next fixed and sorted by flow cytometry based on their DNA context. DNA from the early S-phase and the late S-phase cells is purified and fragmented. Next, library construction and BrdU IP are performed in parallel. Adaptors are ligated to the purified DNA. The BrdU-labeled DNA is immunoprecipitated, followed by indexing the immunoprecipitated DNA. The indexed and pooled libraries are subsequently sequenced. The follow steps are performed to analyze the sequencing data: quality control, mapping reads to the reference genome, and calculating the log base 2 ratio

of normalized coverage from the early S fraction to the normalized coverage from the late S fraction in each genomic bin (log2 ratio early/late). The datasets to be compared are quantile-normalized and smoothed using the Loess smoothing method (locally weighted regression) [112, 113]. The processed data are 1D continuous signals at each genomic bin along the genome, of which higher value represents earlier replication and lower value represents later replication.

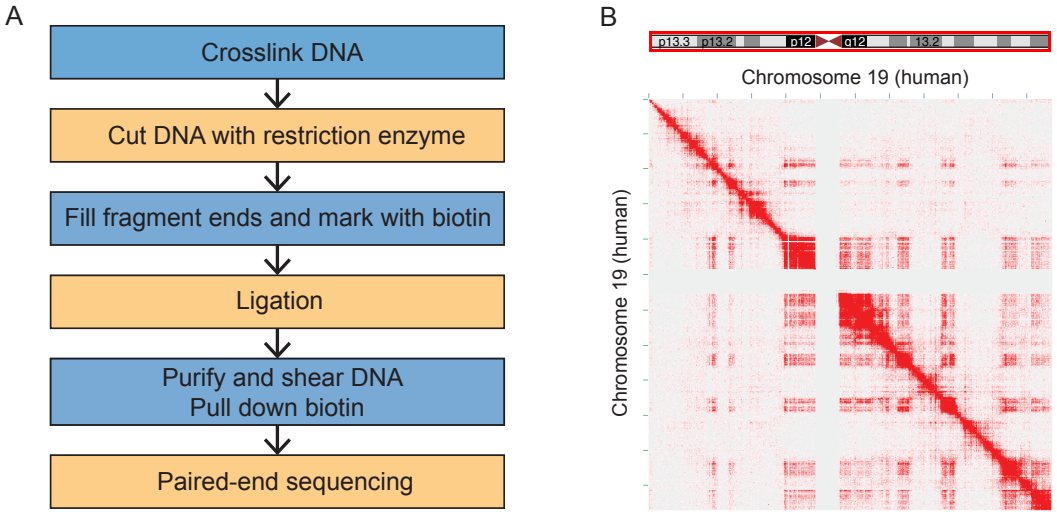

Figure 1.1: Overview of the Hi-C technology [1]. **(A)** The protocol of Hi-C procedures. First, chromatin segments in cells are cross-linked with formaldehyde. Second, chromatin is digested with a restriction enzyme. The ends of the cut chromatin segments are filled in with nucleotides and marked with biotin. Third, ligation is performed to create chimeric molecules. Fourth, DNA fragments are purified and sheared. Biotinylated junctions are isolated. Fifth, paired-end sequencing is performed to identify the junctions. **(B)** The intrachromosomal contact matrix produced by Hi-C on chromosome 19 of the GM12878 cell line in human. Each row and each column of the matrix corresponds to a genomic locus. The intensity value of the pixel represents the total number of reads sequenced over the junction of the two loci. A high intensity value represents close spatial proximity of two loci.

1.4 Structure of the thesis

First, we will introduce the Phylo-HMGP model for comparing continuous-trait functional genomic features across multiple species in Chapter 2. We will then introduce the Phylo-HMRF model for multi-species genome organization comparison using Hi-C data in Chap-

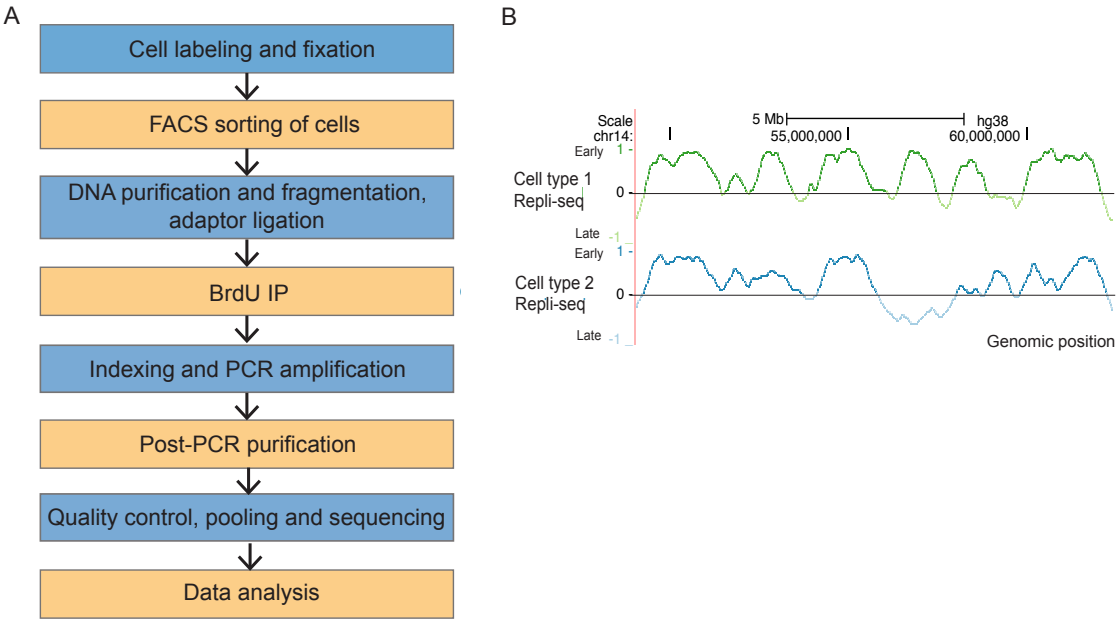

Figure 1.2: Overview of the Repli-seq technology [38]. **(A)** Procedures of the Repli-seq technology. The steps include labeling and fixation of cultured cells, FACS sample preparation and cell sorting, DNA preparation from FACS-sorted cells, DNA fragmentation, library construction, BrdU IP, indexing and amplification, DNA purification, quality control, pooling and sequencing, and Repli-seq data analysis. **(B)** Processed Repli-seq data after calculating log base 2 ratio of normalized read coverage of early versus late S-phase samples and performing data normalizations for compared datasets. The Repli-seq data of two cell types are shown.

ter 3. Next, in Chapter 4, we will explain the proposed CONCERT model for genome-wide DNA RT prediction from genomic sequences and identification of sequence elements that modulate the RT program. We will discuss conclusions and future work directions in Chapter 5.

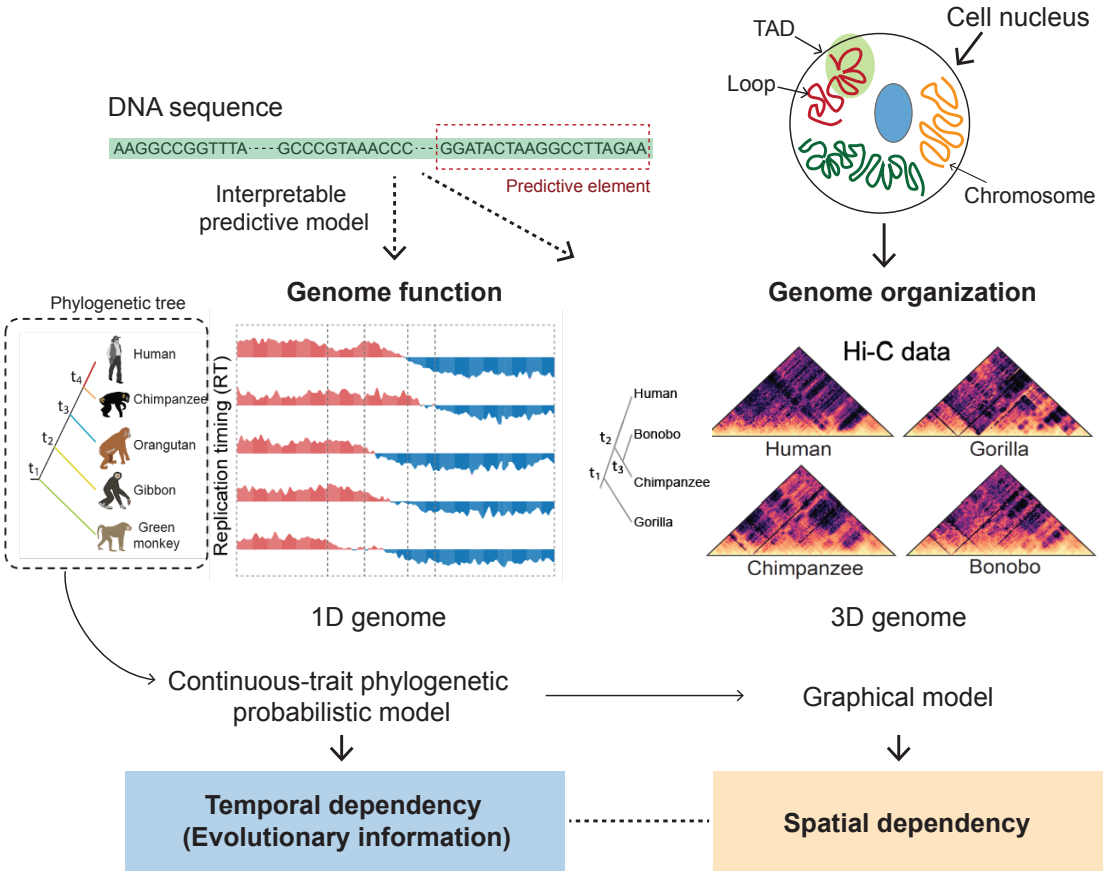

Figure 1.3: Overview of the method development. We use continuous-trait phylogenetic models to model the temporal dependency across species in the context of evolution. The temporal dependency represents the evolutionary relationships across species that are presented by the phylogenetic tree. We use different types of graphical models to model the spatial dependency across genomic loci along the 1D genome and in the 3D space, respectively. We embed the continuous-trait phylogenetic model into the graphical model, to integrate evolutionary relationships across species with the spatial distributions of genomic loci. The developed models perform evolutionary pattern identification across species for 1D functional genomic data and 3D genome organization data, respectively. We also develop an interpretable predictive model to simultaneously perform DNA RT signal prediction and identification of sequence elements that may modulate the DNA RT program.

Chapter 2

Continuous-trait probabilistic model to compare multi-species functional genomic data

2.1 Introduction

As we have introduced in Chapter 1, the computational methods for comparing continuous-trait functional genomic features across multiple species have been limited in the capability, lacking the development to address the computational challenges in comparing the continuous genomic signals using evolutionary information. We develop a new continuous-trait probabilistic model for more accurate evolutionary state estimation based on continuous-trait functional genomic data from different species. We call our model phylogenetic hidden Markov Gaussian processes (Phylo-HMGP). Our new method incorporates the evolutionary affinity among multiple species into the hidden Markov model (HMM) [114] for exploiting both temporal dependencies across species in the context of evolution and spatial dependencies along the genome in a continuous-trait model. Our Phylo-HMGP is fundamentally different from the existing models that are restricted to discrete representation of the studied traits [47, 115–119]. The proposed Phylo-HMGP model utilizes continuous-

trait evolutionary models with spatial constraints to more effectively study the genome-wide features across species. Our model is also flexible such that various continuous-trait evolutionary models or assumptions can be incorporated according to the actual problems to study.

Furthermore, we generated a new cross-species DNA replication timing (RT) dataset from the same cell type in five primate species (human, chimpanzee, orangutan, gibbon, and green monkey). We apply Phylo-HMGP to reveal genome-wide distributions of distinct evolutionary patterns of RT in the five primates. We found that constitutive early and constitutive late RT regions, as defined from human embryonic stem (ES) cell differentiation [9, 62], exhibit a strong correlation with the predicted conserved early RT and conserved late RT patterns. We also found distinct gene functions associated with different RT evolution patterns. In addition, the predicted RT patterns across species show correlations with other genomic and epigenomic features, including higher order genome organization, *cis*-regulatory elements, chromatin marks, and transposable elements. Our results from the comparative RT analysis in five primate species demonstrate the potential of our Phylo-HMGP model to help reveal regions with conserved or lineage-specific regulatory roles for the entire genome.

2.2 Overview of the phylogenetic hidden Markov Gaussian processes model

Here we first provide an overview of the proposed model (Figure 2.1). The details of the model are described in section 2.3 and section 2.4. Our model aims to estimate different evolutionary patterns from multi-species functional genomic signals. As illustrated in Figure 2.1C, the input contains the observed continuous-trait signals from orthologous genomic regions from multiple species. The output is a genome-wide partition where neighboring genomic segments have different predicted states of multi-species signals, reflecting different evolution patterns of the signals being considered.

We define a phylogenetic hidden Markov Gaussian processes (Phylo-HMGP) model as $\mathbf{h} = (S, \psi, A, \pi)$, where S is the set of states, ψ is the set of phylogenetic models, A is the state-transition probability matrix, and π represents the initial state probabilities, respectively. Suppose there are M hidden states. We have $S = \{s_1, \cdots, s_M\}$, $\psi = \{\psi_1, \cdots, \psi_M\}$, $A = \{a_{ij}\}$, $1 \leq i, j \leq M$, and $\pi = \{\pi_1, \cdots, \pi_M\}$. Figure 2.1A shows the state space where different states are associated with varied phylogenetic tree models. Each phylogenetic tree model is parameterized with the OU processes, an example of which is shown in Figure 2.1B. In the thesis, we focus on the Ornstein-Uhlenbeck (OU) process and apply it to analyze cross-species RT data. We also discuss and compare with Brownian motion process within the framework (section 2.5), which is also used as the Gaussian process for realizations of ψ_j to construct the emission probability distributions, $j = 1, \cdots, M$. ψ_j differs under different evolutionary models. Other Gaussian processes can also be embedded into the framework by alternative definitions of ψ_j. Gaussian process is a stochastic process, of which every finite collection of the random variables in the stochastic process has a multivariate Gaussian distribution. Specifically, every finite linear combination of the random variables follows a Gaussian distribution. Both OU process and Brownian motion are Gaussian processes.

Phylo-HMGP provides a generic framework to more effectively incorporate multi-species functional genomic data into the HMM for analyzing both temporal dependencies across species in the phylogeny and spatial dependencies along the entire genome in a continuous-trait model. The framework is flexible and the Gaussian processes embedded in the HMM can be specialized to different evolutionary models (e.g., in this work we focus on OU processes to construct the phylogenetic tree model). The source code of Phylo-HMGP can be accessed at: https://github.com/ma-compbio/Phylo-HMGP.

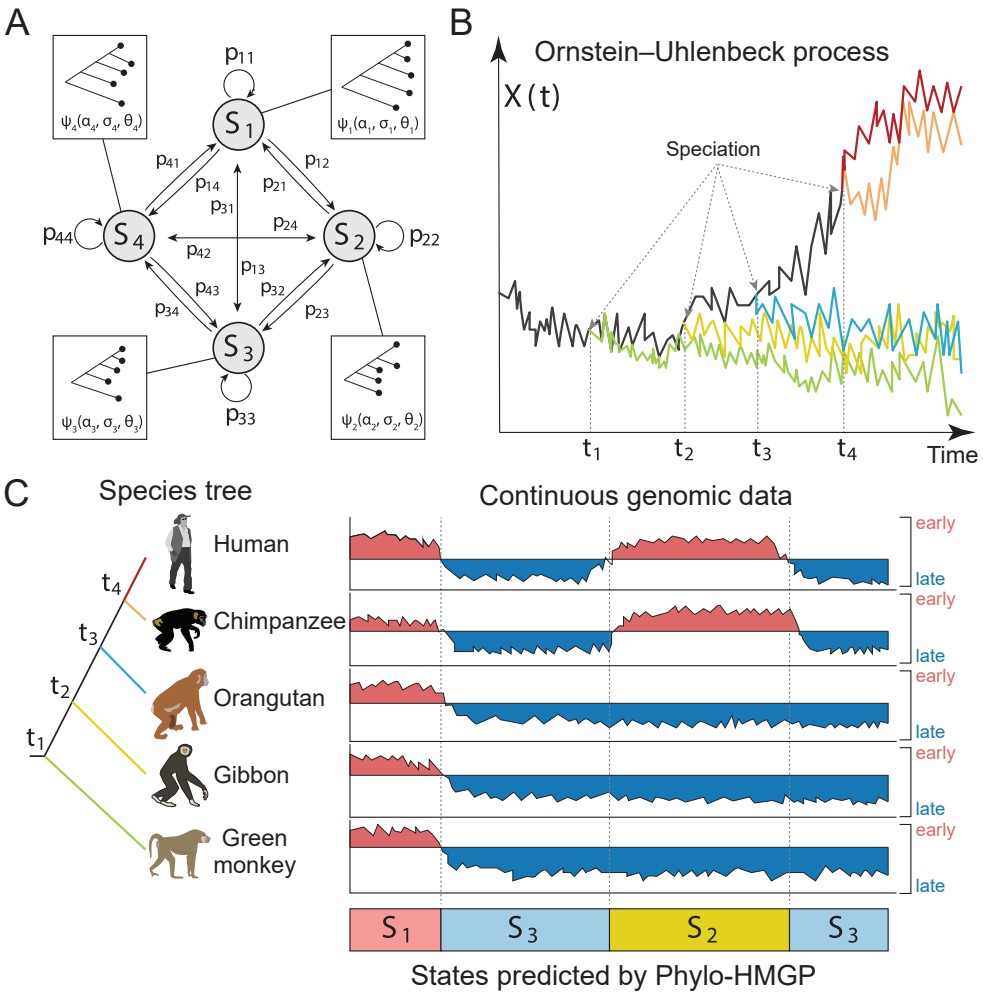

Figure 2.1: Overview of the Phylo-HMGP model. **(A)** Example of the state space and state transition probabilities of the Phylo-HMGP model associated with the continuous genomic data in **(C)**. S_i represents a hidden state. Each hidden state is determined by a phylogenetic model ψ_i, which is parameterized by the selective strengths α_i, Brownian motion intensities σ_i, and the optimal values θ_i of ancestor species and observed species on the corresponding phylogenetic tree. α_i, σ_i, and θ_i are all vectors. **(B)** Illustration of the Ornstein-Uhlenbeck (OU) processes along the species tree specified in **(C)**. $X(t)$ represents the continuous trait at time t. The trajectories of different colors along time correspond to the evolution of the continuous trait in different lineages specified by the corresponding colors in **(C)**, respectively. The time points t_1, t_2, t_3, and t_4 represent the speciation time points, which correspond to the speciation events shown in **(C)**. The observations of the five species also represent an example of state S_2 in **(C)**. **(C)** Simplified representation of input and output of the Phylo-HMGP model. The five tracks of continuous signals represent the observation from five species. S_i represents the underlying hidden states. Specifically, the example is the replication timing data, where 'early' and 'late' represent the early and late stages of replication timing, respectively. The species tree alongside the continuous data tracks shows the evolutionary relationships among the five species that are involved in the RT data comparison across species.

2.3 Ornstein-Uhlenbeck process assumption in Phylo-HMGP

We model the continuous traits with the Ornstein-Uhlenbeck processes in Phylo-HMGP, for which the Gaussian processes are specialized by the OU processes. The Phylo-HMGP model with OU process assumption is named Phylo-HMGP-OU. The OU process is characterized by the following stochastic differential equation [57, 58]:

$$dX_i(t) = \alpha[\theta_i - X_i(t)]dt + \sigma dB_i(t), \tag{2.1}$$

where $X_i(t)$ represents the observation of the i-th species at time point t, $B_i(t)$ is the Brownian motion, α, θ, and σ are parameters that represent the selection strength, the optimal value, and the Brownian motion intensity, respectively. where $X(t)$ represents the observation of X at time point t, and $B(t)$ represents the Wiener process. The Wiener process is also called the standard Brownian motion [120]. α, θ and σ are parameters that represent the selection strength, the optimal value and the fluctuation intensity of Brownian motion, respectively. Here we assume that $X(t)$ is a continuous variable. For example, X_i could be the ChIP-seq signal of a certain histone mark from a specific cell type at a specific locus in a species.

Under the assumption of the OU process, we can derive the expectation, the variance, and the covariance of the observations of species given the phylogenetic model ψ_j. The phylogenetic model is the combination of multiple OU processes that share parameters along common branches. Suppose that X_{p_i} is the trait value of the ancestor of the i-th species, and $X_{a_{ij}}$ is the trait value of the most recent common ancestor of the i-th and j-th species. We will prove the following properties:

$$\mathbb{E}(X_i) = \mathbb{E}(X_{p_i})e^{-\alpha_i t_{ip_i}} + \theta_i \left(1 - e^{-\alpha_i t_{ip_i}}\right), \tag{2.2}$$

$$\mathrm{Cov}(X_i,\ X_j) = \mathrm{Var}(X_{a_{ij}}) \exp\left(-\Sigma_{k\in l_{ij}}\alpha_k t_k - \Sigma_{k\in l_{ji}}\alpha_k t_k\right), \tag{2.3}$$

$$\mathrm{Var}(X_i) = \frac{\sigma_i^2}{2\alpha_i}\left(1 - e^{-2\alpha_i t_{ip_i}}\right) + \mathrm{Var}(X_{p_i})e^{-2\alpha_i t_{ip_i}}, \tag{2.4}$$

where t_{ip_i} is the length of the branch from node p_i to node i, and l_{ij} represents the set of the ancestor nodes of i and i itself after its divergence with j. p_i represents the parent node

of i. Since each node except the root node of the phylogenetic tree has and only has one parent node, without ambiguity, let $t_i = t_{ip_i}$ for any node i that is not the root node.

Let $X_0 = X(0)$, and $B_t = B(t)$. The solution to Equation 2.1 is

$$X(t) = X_0 e^{-\alpha t} + \theta(1 - e^{-\alpha t}) + \sigma \int_0^t e^{-\alpha(t-s)} dB_s. \tag{2.5}$$

$\int_0^t e^{-\alpha(t-s)} dB_s$ is Wiener integral. We have that

$$\int_0^t e^{-\alpha(t-s)} dB_s \sim \boldsymbol{\mathcal{N}}(0, \int_0^t e^{-2\alpha(t-s)} ds). \tag{2.6}$$

$\int_0^t e^{-2\alpha(t-s)} ds = \frac{1}{2\alpha}(1 - e^{-2\alpha t})$. Then we have

$$\sigma \int_0^t e^{-\alpha(t-s)} dB_s \sim \boldsymbol{\mathcal{N}}(0, \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t})). \tag{2.7}$$

Let $Z(t) = \sigma \int_0^t e^{-\alpha(t-s)} dB_s$. We have

$$X(t) = X_0 e^{-\alpha t} + \theta(1 - e^{-\alpha t}) + Z(t), \tag{2.8}$$

where $Z(t) \sim \boldsymbol{\mathcal{N}}(0, \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t}))$. $Z(t)$ is independent of X_0. Therefore,

$$X(t)|X_0 = x_0 \sim \boldsymbol{\mathcal{N}}(x_0 e^{-\alpha t} + \theta(1 - e^{-\alpha t}), \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t})). \tag{2.9}$$

We assume that the evolution of multiple species from a common ancestor follow the OU processes. The evolutionary relationships among multiple species form a tree structure, which is the evolutionary tree, or phylogenetic tree. We assume that each branch of the phylogenetic tree, which represents the evolution of a species from its most recent ancestor, is modeled by an OU process.

For the i-th branch, suppose the observation of the descendant species (child) is X_i and the observation of the ancestor species (parent) is X_{p_i}. Suppose the length of the i-the branch is t_i, which represents the evolution time from X_{p_i} to X_i. Based on Equation 2.8, with the variables $X(t)$, X_0, $Z(t)$ replaced by X_i, X_{p_i}, Z_i, respectively, we have that

$$X_i = X_{p_i} e^{-\alpha_i t_i} + \theta_i(1 - e^{-\alpha_i t_i}) + Z_i, \tag{2.10}$$

where $Z_i \sim \boldsymbol{\mathcal{N}}(0, \frac{\sigma_i^2}{2\alpha_i}(1 - e^{-2\alpha_i t_i}))$ and Z_i is independent of X_{p_i}. α_i and σ_i represent the selection strength and the fluctuation intensity of Brownian motion along the i-th branch, respectively. θ_i represent the optimal value of X_i in the OU process.

Then we have that

$$\mathbb{E}(X_i) = \mathbb{E}(X_{p_i})e^{-\alpha_i t_i} + \theta_i(1 - e^{-\alpha_i t_i}) + \mathbb{E}(Z_i) \tag{2.11}$$

$$= \mathbb{E}(X_{p_i})e^{-\alpha_i t_i} + \theta_i(1 - e^{-\alpha_i t_i}), \tag{2.12}$$

$$\mathrm{Var}(X_i) = \mathrm{Var}(X_{p_i}e^{-\alpha_i t_i}) + \mathrm{Var}(Z_i) \tag{2.13}$$

$$= \mathrm{Var}(X_{p_i})e^{-2\alpha_i t_i} + \frac{\sigma_i^2}{2\alpha_i}(1 - e^{-2\alpha_i t_i}). \tag{2.14}$$

Let V_T, V_B denote the set of indices of all the tree nodes and the set of indices of all the branches. The descendant node associated with the i-th branch is the i-th node. $\{Z_j\}_{j \in V_B}$ are independent of each other, and Z_j is independent of X_i for any $j \notin ancestor(i)$, where $ancestor(i)$ represents the set of all the ancestor nodes of node i.

For any $i \neq j$ and $i \notin ancestor(j)$ and $j \notin ancestor(i)$, we have

$$\mathrm{Cov}(X_i,\ X_j) = \mathrm{Cov}\left(X_{p_i}e^{-\alpha_i t_i} + \theta_i(1 - e^{-\alpha_i t_i}) + Z_i, X_{p_j}e^{-\alpha_j t_j} + \theta_j(1 - e^{-\alpha_j t_j}) + Z_j\right) \tag{2.15}$$

$$= \mathrm{Cov}(X_{p_i}, X_{p_j})e^{-\alpha_i t_i - \alpha_j t_j}. \tag{2.16}$$

Repeating the derivation above to calculate $\mathrm{Cov}(X_{p_i}, X_{p_j})$, we have that

$$\mathrm{Cov}(X_i,\ X_j) = \mathrm{Var}(X_{a_{ij}}) \exp\left(-\sum_{k \in l_{ij}} \alpha_k t_k - \sum_{k' \in l_{ji}} \alpha_{k'} t_{k'}\right), \tag{2.17}$$

where $X_{a_{ij}}$ represents the observation of the most recent common ancestor of species i and j, and l_{ij} represents the set of the ancestor nodes of species i and i itself after the divergence of species i with species j. For $k \in l_{ij}$, t_k corresponds to length of the branch from the parent of species k to species k.

If $j \in ancestor(i)$, then $i \notin ancestor(j)$, because there is no circle, we have

$$\mathrm{Cov}(X_i,\ X_j) = \mathrm{Cov}\left(X_{p_i}e^{-\alpha_i t_i} + \theta_i(1 - e^{-\alpha_i t_i}) + Z_i, X_j\right) \tag{2.18}$$

$$= \mathrm{Cov}(X_{p_i}, X_j)e^{-\alpha_i t_i}. \tag{2.19}$$

Repeating the derivation above to calculate $\mathrm{Cov}(X_{p_i}, X_j)$, we have that

$$\mathrm{Cov}(X_i,\ X_j) = \mathrm{Var}(X_j) \exp\left(-\sum_{k \in l_{ij}} \alpha_k t_k\right), \tag{2.20}$$

where l_{ij} represent the set of nodes on the path from node j (included) to node i (included). Equation 2.20 can also be written as Equation 2.17. The same rule applies to $i \in ancestor(j)$. Therefore, Equation 2.17 applies to any $i \neq j$.

Based on the assumptions of the OU process, the multi-species observations follow multivariate Gaussian distribution. As we show above, the expectation, variance, and covariance of the observations of species can be computed given the phylogenetic tree topolgy and the O-U process process parameters, which are consistent with the results given in [53].

Furthermore, we will show that the phylogenetic tree with OU process assumption for the continuous variables is a Gaussian directed acyclic graphich model (Gaussian DAG).

Suppose $\boldsymbol{\mathcal{G}}$ is a Directed Acyclic Graph (DAG) with vertices $\boldsymbol{\mathcal{V}} = (X_1, \cdots, X_d)$. Let $\boldsymbol{\mathcal{V}}$ also denote the set of indices of $X_1, \cdots, X_d$. Here we introduce the definition of Gaussian directed graphical models. Let Y be a continuous variable in a DAG with parents $X_1, ..., X_k$. Y has a linear Gaussian model of its parents if there are parameters $\beta_0, \beta_1, \cdots, \beta_k$ and σ^2 such that

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon, \quad \epsilon \sim \boldsymbol{\mathcal{N}}(0, \sigma^2). \tag{2.21}$$

Here $\boldsymbol{\mathcal{N}}(0, \sigma^2)$ represents Gaussian distribution with mean value 0 and variance σ^2. A directed graphical model is a Gaussian directed graphical model if all the variables are continuous, and every variable and its parents follow a linear Gaussian model [121].

For a phylogenetic tree, $\boldsymbol{\mathcal{G}}$ is of tree topology, where each node except the root node is attached to a branch as the child node of its parent node. In Equation 2.10, let

$$\beta_{i,1} = e^{-\alpha_i t_i}, \tag{2.22}$$

$$\beta_{i,0} = \theta_i (1 - e^{-\alpha_i t_i}), \tag{2.23}$$

$$\eta_i^2 = \frac{\sigma_i^2}{2\alpha_i}(1 - e^{-2\alpha_i t_i}). \tag{2.24}$$

We have that

$$X_i = X_{p_i}\beta_{i,1} + \beta_{i,0} + Z_i, \tag{2.25}$$

where $Z_i \sim \boldsymbol{\mathcal{N}}(0, \eta_i^2)$. Suppose X_0 is the observation of the root node. Assume $X_0 \sim \boldsymbol{\mathcal{N}}(\beta_0, \eta_0^2)$. Then $X_0 = \beta_0 + Z_0$, where $Z_0 \sim \boldsymbol{\mathcal{N}}(0, \eta_0^2)$. Therefore, in the phylogenetic tree with OU process assumption where the variables are continuous variables, X_i has a linear Gaussian model of its parent for any $i \in \boldsymbol{\mathcal{V}}$. Therefore, the phylogenetic tree with OU process assumption for the continuous variables is a Gaussian DAG. Here the continuous variable may refer to a type of continuous-trait genomic feature.

Let $\tilde{\alpha}_i = \alpha_i t_i$, $\tilde{\sigma}_i^2 = \sigma_i^2 t_i$, $i \in \boldsymbol{\mathcal{V}}$. We have

$$\tilde{\alpha}_i = \alpha_i t_i = -\log \beta_{i,1}, \tag{2.26}$$

$$\theta_i = \frac{\beta_{i,0}}{1-\beta_{i,1}}, \tag{2.27}$$

$$\tilde{\sigma}_i^2 = \sigma_i^2 t_i = \frac{\eta_i^2(2\alpha_i t_i)}{1-e^{-2\alpha_i t_i}} = \frac{2\tilde{\alpha}_i \eta_i^2}{1-\beta_{i,1}^2} = -\frac{2\eta_i^2 \log \beta_{i,1}}{1-\beta_{i,1}^2}, \tag{2.28}$$

where $\beta_{i,1} \in (0,1)$. Therefore, the OU model parameters are functions of the Guassian DAG model parameters for the phylogenetic tree with OU processes.

2.4 Ornstein-Uhlenbeck processes with spatial dependencies

2.4.1 Overall framework of Phylo-HMGP with OU processes

As defined in section 2.2, a Phylo-HMGP model is reprented as $\mathbf{h} = (S, \psi, A, \pi)$, where S, ψ, A, and π denote the set of states, the set of phylogenetic models, the state-transition probability matrix, and the initial state probabilities, respectively. In the Phylo-HMGP model with OU process, ψ_j is defined as: $\psi_j = (\theta_j, \alpha_j, \sigma_j, \tau_j, \beta_j)$, $j = 1, \cdots, M$, where θ_j, α_j, σ_j denote the OU process parameters of the j-th state, respectively. τ_j, β_j represent the topology of the phylogenetic tree and the branch lengths, respectively. We allow varied selection strength and Brownian motion intensity along each branch and varied optimal values at interior nodes or leaf nodes. Suppose there are r branches. We have $\theta_j \in \mathbb{R}^{r+1}$,

$\alpha_j, \sigma_j \in \mathbb{R}^r$. Suppose $\mathbf{x} = (x_1, \cdots, x_N)$ are observations of consecutive regions along a sequence of length N, and $\mathbf{y} = (y_1, \cdots, y_N)$ are the underlying hidden states, respectively. Each observation x_i is a multi-dimensional vector of the trait values of the compared species with respect to a certain type of functional genomic feature for an orthologous genomic region. Suppose there are d species, which correspond to the d leaf nodes in the phylogenetic tree. We have $x_i \in \mathbb{R}^d$, $y_i \in \{1, \cdots, M\}$, $i = 1, \cdots, N$. The hidden state y_i indicates a specific phylogenetic model ψ_j from which the observation x_i is generated. $\{\psi_j\}_{j=1}^M$ represent different evolutionary patterns of the genomic features across the multiple species. For example, one phylogenetic model ψ_i may represent conserved evolution of the feature across species, while another model ψ_j may represent strong selection strength along one lineage that results in a lineage-specific pattern. Given the input of multi-species functional genomic signals over a range of regions, which can be processed into the observations $\mathbf{x}$, we are trying to infer the underlying evolutionary patterns and predict the evolutionary states $\mathbf{y}$ through model parameter estimation. Each y_i represents an evolutionary pattern, parameterized by the inferred phylogenetic model ψ_{y_i}. The output includes the estimated model parameters $\hat{\mathbf{h}}$ and predicted states $\hat{\mathbf{y}}$. Note that our Phylo-HMGP-OU is different from the HMMSDE methods that use a temporal HMM to simulate a single OU process [122]. Phylo-HMGP-OU embeds phylogenetic models constructed by complex of OU processes into a spatial HMM to utilize both temporal and spatial dependencies between variables.

2.4.2 Parameter estimation

Let Θ be the model parameters. The joint probability of the observations $\mathbf{x}$ and states $\mathbf{y}$ is $p(\mathbf{x}, \mathbf{y}|\Theta) = \pi_{y_0} \prod_{i=1}^N a_{y_{i-1}, y_i} p(x_i|y_i, \Theta)$ [123]. We use Expectation-Maximization (EM) algorithm [124] for parameter estimation. Suppose Θ^g is the current estimate of model parameters. The EM algorithm computes the expectation of the complete-data log

likelihood, which is defined as the Q function $Q(\Theta, \Theta^g)$:

$$Q(\Theta, \Theta^g) = \mathbb{E}[\log p(\mathbf{x}, \mathbf{y}|\Theta)|\mathbf{x}, \Theta^g] = \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{x}, \mathbf{y}|\Theta^g) \log p(\mathbf{x}, \mathbf{y}|\Theta), \tag{2.29}$$

where $\boldsymbol{\mathcal{S}}_N$ is the set of all state sequences of length N. We have

$$\log p(\mathbf{x}, \mathbf{y}|\Theta)) = \log(p(\mathbf{y}|\Theta)p(\mathbf{x}|\mathbf{y}, \Theta)) \tag{2.30}$$

$$= \log p(\mathbf{y}|\Theta) + \log p(\mathbf{x}|\mathbf{y}, \Theta)) \tag{2.31}$$

$$= \log \pi_{y_0} + \sum_{i=1}^{N} \log a_{y_{i-1}, y_i} + \sum_{i=1}^{N} \log p(x_i|y_i). \tag{2.32}$$

Then we have

$$Q(\Theta, \Theta^g) = \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{x}, \mathbf{y}|\Theta^g) \log \pi_{y_0} + \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{x}, \mathbf{y}|\Theta^g) \sum_{i=1}^{N} \log a_{y_{i-1}, y_i} \tag{2.33}$$

$$+ \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{x}, \mathbf{y}|\Theta^g) \sum_{i=1}^{N} \log p(x_i|y_i). \tag{2.34}$$

The model parameters π, A and ψ can be updated separately in the Maximization-step (M-step), which correspond to the three parts of $Q(\Theta, \Theta^g)$, respectively. The parameters of the Ornstein-Uhlenbeck (OU) model are involved in $p(x_i|y_i)$ of the third term of $Q(\Theta, \Theta^g)$.

Define that $q_{-i} = (y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_N)$. We represent the third part as:

$$\sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{x}, \mathbf{y}|\Theta^g) \sum_{i=1}^{N} \log p(x_i|y_i)$$

$$= \sum_{i=1}^{N} \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{x}, \mathbf{y}|\Theta^g) \log p(x_i|y_i) \tag{2.35}$$

$$= \sum_{i=1}^{N} \sum_{l=1}^{M} \sum_{y_{-i} \in \boldsymbol{\mathcal{S}}_{N-1}} p(\mathbf{x}, y_1, \cdots y_{i-1}, y_i = l, y_{i+1}, \cdots, y_N|\Theta^g) \log p(x_i|y_i = l) \tag{2.36}$$

$$= \sum_{l=1}^{M} \sum_{i=1}^{N} p(\mathbf{x}, y_i = l|\Theta^g) \log p(x_i|y_i = l). \tag{2.37}$$

Here $p(\mathbf{x}, y_i = l|\Theta^g)$ can be computed using forward-backward algorithm [123, 125]. Assume that continuous-trait variables follow multivariate Gaussian distributions. For a given

state l, we have

$$\log p(x|\mu_\Theta^{(l)}, \Sigma_\Theta^{(l)}) \propto -\frac{1}{2}\log|\Sigma_\Theta^{(l)}| - \frac{1}{2}(x-\mu_\Theta^{(l)})^T[\Sigma_\Theta^{(l)}]^{-1}(x-\mu_\Theta^{(l)}) \tag{2.38}$$

The underlying phylogenetic model ψ_l is embedded into $\Sigma_\Theta^{(l)}$ and $\mu_\Theta^{(l)}$ by Eq. (2.2)-(2.4). Then the negative expected log likelihood of state l is:

$$L(\Theta^{(l)}) = \frac{1}{2}\log|\Sigma_\Theta^{(l)}| \sum_{i=1}^{N} p\left(\mathbf{x}, y_i = l|\Theta^g\right) \tag{2.39}$$

$$+ \frac{1}{2}\sum_{i=1}^{N}\left(x_i - \mu_\Theta^{(l)}\right)^T[\Sigma_\Theta^{(l)}]^{-1}\left(x_i - \mu_\Theta^{(l)}\right)p(\mathbf{x}, y_i = l|\Theta^g). \tag{2.40}$$

Therefore, multiplied by $2/N$, the third part of the negative Q function with respect to a given state l can be represented as:

$$\tilde{L}(\Theta^{(l)}) = \frac{1}{N}\log|\Sigma_\Theta^{(l)}| \sum_{i=1}^{N} w_i^{(l)} + \mathrm{tr}\left([\Sigma_\Theta^{(l)}]^{-1}\tilde{S}_\Theta^{(l)}\right), \tag{2.41}$$

where

$$w_i^{(l)} = p(\mathbf{x}, y_i = l|\Theta^g) \tag{2.42}$$

$$\tilde{S}_\Theta^{(l)} = \frac{1}{N}\sum_{i=1}^{N} w_i^{(l)}\left(x_i - \mu_\Theta^{(l)}\right)\left(x_i - \mu_\Theta^{(l)}\right)^T. \tag{2.43}$$

Here $\Theta^{(l)}$ represents the phylogenetic model parameters associated with state l. We have $\Theta^{(l)} = \{\theta_l, \alpha_l, \sigma_l, \tau_l, \beta_l\}, l = 1, \cdots, M$.

We perform parameter estimation for each of the possible states. We assume τ_l is given. β_l can be combined in effect to α_l and σ_l. In practice, if the real branch lengths are unknown, we perform the transformation that

$$\tilde{\alpha}_v = \alpha_v\beta_v, \tag{2.44}$$

$$\tilde{\sigma}_v^2 = \sigma_v^2\beta_v, \tag{2.45}$$

where β_v represents the length of branch from the parent of node v to node v. Using this approach the branch lengths are incorporated into $\{\alpha_l, \sigma_l\}$. Then $\Theta^{(l)} = \{\theta_l, \tilde{\alpha}_l, \tilde{\sigma}_l\}$. The

objective function for parameter estimation for a given state l is

$$\min_{\Theta^{(l)}} \frac{1}{N} \log |\Sigma_\Theta^{(l)}| \sum_{i=1}^{N} w_i^{(l)} + \text{tr}\left([\Sigma_\Theta^{(l)}]^{-1} \tilde{S}_\Theta^{(l)} \right). \tag{2.46}$$

A challenge is that there are approximately two times more model parameters than the feature dimension for each state. We apply ℓ_2-norm regularization to the parameters $\Theta^{(l)}$. In each M-step, the objective function of a given state l is defined as:

$$\min_{\Theta^{(l)}} \frac{1}{N} \log |\Sigma_\Theta^{(l)}| \sum_{i=1}^{N} w_i^{(l)} + \text{tr}\left([\Sigma_\Theta^{(l)}]^{-1} \tilde{S}_\Theta^{(l)} \right) + \lambda \|\Theta^{(l)}\|_2^2, \tag{2.47}$$

where $w_i^{(l)}$ and $\tilde{S}_\Theta^{(l)}$ are defined as above. We define $\lambda = \lambda_0/\sqrt{N}$, and tune λ_0 based on a fixed simulation dataset. We estimated the range of λ_0 that can improve the performance of Phylo-HMGP-OU (see section 2.4.5 and Figure 2.2). Accordingly, we applied the same λ_0 to all the simulation datasets and the real data as a fixed coefficient, without tuning λ_0 on each dataset specially, in order to avoid overfitting of λ_0 on a particular dataset.

From the first two parts of $Q(\Theta, \Theta^g)$ we can update the estimates of $\boldsymbol{\pi}$ and A accordingly. Let $A = \{a_{kl}\}$, where $a_{kl} = p(y_i = l | y_{i-1} = k)$, $k, l = 1, \cdots, M$. We have:

$$\pi_l = \frac{p(\mathbf{x}, y_0 = l | \Theta^g)}{p(\mathbf{x} | \Theta^g)}, \tag{2.48}$$

$$a_{kl} = \frac{\sum_{i=1}^{N} p(\mathbf{x}, y_{i-1} = k, y_i = l | \Theta^g)}{\sum_{i=1}^{N} p(\mathbf{x}, y_{i-1} = k | \Theta^g)}. \tag{2.49}$$

Therefore, in each E-step, given the present estimated model parameters Θ^g, we compute $p(\mathbf{x}, y_i = l | \Theta^g)$ and $p(\mathbf{x}, y_{i-1} = k, y_i = l | \Theta^g)$ using the forward-backward algorithm [123, 125](section 2.4.4), $k, l = 1, \cdots, M$. In each M-step, we solve the maximum expected likelihood estimation problem to update the parameters π, A, and $\{\psi_j\}_{j=1}^M$. Given the estimated model parameters, we can predict a most likely sequence of hidden states $\hat{\mathbf{y}}$ using the Viterbi algorithm [126].

Note that the existing discrete-trait Phylo-HMMs [115, 116] can also be represented as $\mathbf{h} = (S, \psi, A, \pi)$, where the evolutionary model ψ_j is defined according to the substitution process with respect to an alphabet Σ_j of discrete characters, e.g., $\Sigma_j = \{A, C, G, T\}$

for nucleotides. In the discrete-trait Phylo-HMMs, ψ_j is defined as $\psi_j = \{Q_j, b_j, \tau_j, \beta_j\}$, where Q_j is the substitution rate matrix, b_j is the vector of the background character frequencies, τ_j is the tree topology, and β_j represents the branch lengths. This realization of ψ_j is limited to the discrete characters, where transition probabilities between two characters can be computed to model evolution of characters, e.g., the HKY85 model [127]. For the continuous traits, we use continuous-trait evolutionary model assumptions to define ψ_j.

2.4.3 Initialization of the Expectation-Maximization algorithm in Phylo-HMGP

Phylo-HMGP uses the Expectation-Maximization (EM) algorithm for parameter estimation. The EM algorithm seeks local minima and the results of EM algorithm are influenced by initializations. We designed different ways for parameter initialization. The first approach is to estimate OU model parameters initially based on the primitive state estimation results from K-means clustering. We perform model estimation for each cluster separately as single-state estimation, and use the estimates as initial model parameter values for the EM algorithm.

The second approach is to generate initial values randomly. There are three types of parameters in the OU model for a single state, which are optimal values θ, selection strength α, and Brownian motion intensity σ. We sample random variables from uniform distributions for the initial values of θ, α, and σ, respectively.

The third approach is to use a linear combination of the initial parameter values obtained from the first approach and the second approach. We estimate the initial parameter values as $\Theta_0 = w_1\Theta_1 + (1 - w_1)\Theta_2$, where Θ_1 and Θ_2 are parameter estimates from the first and second approaches. By changing the initial weight w_1, we have different initialization schemes. Based on the performance with respect to varied w_1 in simulation study I, we observed that Phylo-HMGP is not very sensitive to initialization on four datasets, while on the other datasets the performance is improved as w_1 increases within a range. Given $w_1 \in [0.2, 1.0]$, the performance of Phylo-HMGP on each simulated dataset is comparable to the best performance it can achieve on the corresponding dataset. For performance

comparison with other methods in the simulation study, we fixed $w_1 = 0.8$ for all the datasets to prevent overfitting on a particular dataset. The initialization weight w_1 is an input parameter to the implemented program and can be adjusted within $[0, 1]$ by the user's choice.

2.4.4 Forward-backward algorithm in the Expectation-step

For the objective function defined in 2.47, we compute $w_i^{(l)} = p(\mathbf{x}, y_i = l|\Theta^g)$ in the E-step of the EM algorithm using the forward-backward algorithm and current model parameter estimates Θ^g. Let $\mathbf{x} = (x_1, \cdots, x_T)$ be the observation sequence. We define:

$$\alpha_l(t) = p(x_1, x_2, \cdots, x_t, y_t = l|\Theta^g), \tag{2.50}$$

and

$$\beta_l(t) = p(x_{t+1}, x_{t+2}, \cdots, x_T|y_t = l, \Theta^g). \tag{2.51}$$

According to the forward-backward algorithm, we have:

$$p(\mathbf{x}, y_t = l|\Theta^g) = \alpha_l(t)\beta_l(t). \tag{2.52}$$

Both $\alpha_j(t)$ and $\beta_j(t)$ can be computed recursively. Let $\pi_l = p(y_1 = l)$ be the initial state distribution, $l = 1, \cdots, M$. The forward procedure to compute $\alpha_l(t)$ is as follows.

$$\alpha_l(1) = \pi_l p(x_1|y_1 = l), \tag{2.53}$$

$$\alpha_l(t+1) = [\sum_{j=1}^{M} \alpha_j(t)a_{lj}]p(x_{t+1}|y_{t+1} = l), \tag{2.54}$$

$$p(\mathbf{x}|\Theta) = \sum_{l=1}^{M} \alpha_l(T). \tag{2.55}$$

The backward procedure to compute $\beta_l(t)$ is as follows:

$$\beta_l(T) = 1, \tag{2.56}$$

$$\beta_l(t) = \sum_{j=1}^{M} a_{lj}y(x_{t+1}|y_{t+1} = j)\beta_j(t+1), \tag{2.57}$$

$$p(\mathbf{x}|\Theta) = \sum_{l=1}^{M} \beta_l(1)\pi_l p(x_1|y_1 = l). \tag{2.58}$$

We also update the transition probability between any two states. Define that $\epsilon_{kl}(t) = p(y_t = k, y_{t+1} = l|\mathbf{x}, \Theta)$. We have:

$$\epsilon_{kl}(t) = \frac{p(y_t = k, y_{t+1} = l, \mathbf{x}|\Theta)}{p(\mathbf{x}|\Theta)} = \frac{\alpha_k(t)a_{kl}p(x_{t+1}|y_{t+1} = l)\beta_l(t+1)}{\sum_{k=1}^{M}\sum_{l=1}^{M}\alpha_k(t)a_{kl}p(x_{t+1}|y_{t+1} = l)\beta_l(t+1)}. \tag{2.59}$$

Equivalent to Eq. (2.49), the transition matrix can be updated as:

$$a_{kl} = \frac{\sum_{t=1}^{T-1}\epsilon_{kl}(t)}{\sum_{t=1}^{T-1}p(y_t = k|\mathbf{x}, \Theta)}, k, l = 1, \cdots, M, k \neq l. \tag{2.60}$$

With $p(\mathbf{x}, y_i = l|\Theta^g)$ and $p(\mathbf{x}, y_{i-1} = k, y_i = l|\Theta^g)$ computed in each E-step, $k, l = 1, \cdots, M$, we update the parameters π, A, and $\{\psi_j\}_{j=1}^{M}$ in each M-step.

2.4.5 Estimation of the regularization coefficient in Phylo-HMGP

For the objective function defined in Formula (2.47), we define $\lambda = \lambda_0/\sqrt{N}$, where N is the sample size, and we observe how performance of Phylo-HMGP-OU changes with respect to λ_0 based on a fixed simulation dataset (simulation dataset I-1), in order to estimate a range of λ_0 in which the performance of Phylo-HMGP-OU can be improved with the l_2-norm regularization. We tuned λ_0 from 0 to 5, with the step size of 0.5, and compared the performance of the model with respect to the different choices of λ_0. We found that the model with $\lambda_0 \in [3.0, 5.0]$ reaches relatively higher F_1 score than the other choices of λ_0 on this dataset (Figure 2.2). We selected $\lambda_0 = 4.0$ and applied it to all the simulation datasets and the RT data as a fixed coefficient, without tuning λ_0 on each dataset specially, in order to avoid overfitting of λ_0 on a particular dataset. We also repeated the experiment on dataset I-1 and observed how the performance of Phylo-HMGP-OU changes with λ_0 on the other datasets in simulation study I. We found that the performance of Phylo-HMGP-OU is not sensitive to λ_0 ranging in [3.0,5.0] on most of the simulation datasets (I-1,I-3,I-4,I-5,I-6). Phylo-HMGP-OU still reaches comparable performance to the highest performance it can achieve on dataset I-2. We only used the performance resulted from $\lambda_0 = 4.0$ on all the simulation datasets for performance evaluation and comparison.

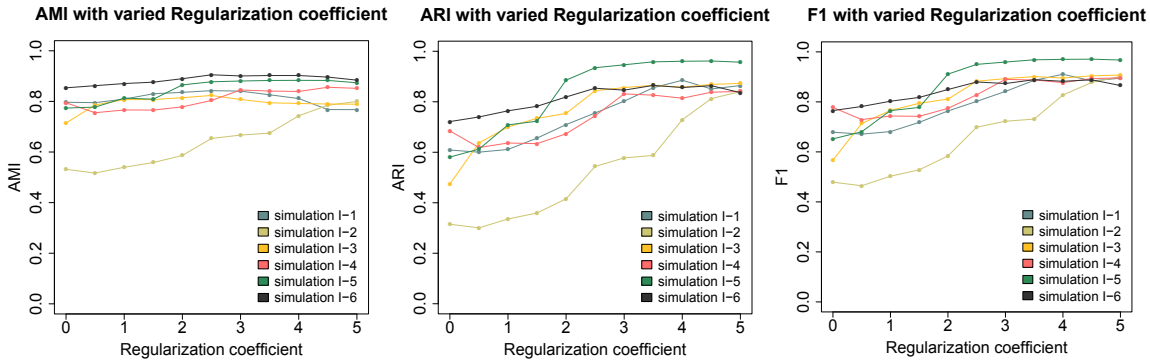

Figure 2.2: Performance evaluation on AMI, ARI, and F_1 score in simulation study I with respect to varied l_2-norm regularization coefficient λ_0.

2.5 Brownian motion in Phylo-HMGP

The proposed Phylo-HMGP framework is flexible. Any type of Gaussian process can be used to realize the continuous-trait evolutionary model according to the assumption used. For more comprehensive method evaluation of the proposed framework, we also developed the Phylo-HMGP-Brownian Motion (Phylo-HMGP-BM) method, where the embedded continuous-trait model is the Brownian motion model. A standard Brownian motion is a stochastic process $X = \{X_t : t \in [0, \infty)\}$ that has the following properties: (i) $X_0 = 0$ with probability 1; (ii) The increment of X_t is independent of the past values; (iii) The increment of X_t is normally distributed with mean 0 and the time interval of the increment as variance; (iv) X has continuous paths.

Phylo-HMGP-BM is also built from $\mathbf{h} = (S, \psi, A, \pi)$. For Phylo-HMGP-BM, ψ_j is defined as $\psi_j = (\mu_j, \tau_j, \beta_j, \lambda_j)$, $j = 1, \cdots, M$, where μ_j denotes the mean values of leaf nodes, and τ_j, β_j, λ_j denote the phylogenetic tree topology, the branch lengths, and the evolution rates on branches, respectively. Under the Brownian motion assumption, the covariance between observations of two species depends on the depth of their nearest common ancestor in the phylogenetic tree. The covariance matrix based on the BM model can therefore be presented as a linear combination of covariance matrices [128].

Suppose r is the number of branches of the phylogenetic tree, and d is the number of leaf nodes (i.e., the number of observed species). We have $\lambda_j \in \mathbb{R}^r$, $\beta_j \in \mathbb{R}^r$, and $\mu_j \in \mathbb{R}^d$,

$j = 1, \cdots, M$. We number the branches with $1, \cdots, r$. For any vector f, let $f(k)$ be the k-th element of f. Let $v_j \in \mathbb{R}^r$, and $v_j(k) = \lambda_j(k) \cdot \beta_j(k)$, $k = 1, \cdots, r$, which reflects the combined effect of branch length and evolution rate along each branch. Without loss of generality, suppose $v \in \mathbb{R}^r$ is the transformed evolution rate vector for an arbitrary state. Then $v(k)$ represents the transformed evolution rate on the k-th branch. Suppose X_i is the observation of a species. Based on the model of Brownian motion, the mean value of X_i is identical to that of the observation of its ancestor and the variance of X_i is proportional to the evolution time from its ancestor. We have:

$$\mathbb{E}[X_i] = \mathbb{E}[X_p], \tag{2.61}$$

$$\mathrm{Var}(X_i) = \sum_{k \in S_a(i)} v(k), \tag{2.62}$$

$$\mathrm{Cov}(X_i, X_j) = \sum_{k \in S_a(i,j)} v(k), \tag{2.63}$$

where t_i represents the branch length from the nearest ancestor of species i to species i, and $S_a(i, j)$ represents the set of common of ancestors of species i and j. The covariance matrix based on the Brownian motion model can therefore be presented as [128]:

$$\Sigma_v = G_0 + \sum_{k=1}^{r} v(k) G_k, \tag{2.64}$$

where G_k is a binary matrix representing contribution of a specific branch to the covariance matrix. Suppose $\mathbf{x} = (x_1, \cdots, x_N)$ are observations of consecutive genome regions along a sequence of length N, and $\mathbf{y} = (y_1, \cdots, y_N)$ are the corresponding hidden states. Similar to Phylo-HMGP-OU, we use EM algorithm for parameter estimation. We define the Q function $Q(\Theta, \Theta^g)$ in the same way as Eq. (2.33), which is the expectation of the complete-data log likelihood function, but with different realization of $p(\mathbf{x}, \mathbf{y}|\Theta^g)$ according to the assumption of the Brownian Motion model. Here Θ represents the model parameters and Θ^g is the current estimate of the parameters. Based on the original Brownian motion model, the expectation of observation of descendant species is always identical to that of its ancestor. We have $\mathbb{E}(X_i) = \mathbb{E}(X_0)$ under this assumption, where X_0 corresponds to the most

remote ancestor. However, we can observe shift of the mean value of the phenotype in real world problems [129, 130]. Therefore we relax this constraint on the expectation of the observations, using a weaker assumption that allows the expectation to be shifted on branches. Then we considers the phenotype expectation of each species as model parameters, allowing the expectation to vary between species.

We compute the third part of the negative Q function with respect to each state, i.e., the negative expected log likelihood of each state (multiplied by $2/N$), which is denoted by $\tilde{L}(\Theta^{(l)})$, $l = 1, \cdots, M$. Here $\Theta^{(l)}$ represents the model parameters associated with state l. We have $\Theta^{(l)} = \{v_l, \mu_l\}$. We minimize $\tilde{L}(\Theta^{(l)})$ to estimate $\Theta^{(l)}$. Accordingly, the objective function of a given state l is:

$$\min_{v_l, \mu_l} \frac{1}{N} \log |\Sigma_{v,\mu}^{(l)}| \sum_{i=1}^{N} w_i^{(l)} + \mathrm{tr}\big([\Sigma_{v,\mu}^{(l)}]^{-1} \tilde{S}_{\mu}^{(l)}\big), \tag{2.65}$$

where $w_i^{(l)} = p(\mathbf{x}, y_i = l|\Theta^g)$, and $\tilde{S}_{\mu}^{(l)} = \frac{1}{N} \sum_{i=1}^{N} w_i^{(l)} (x_i - \mu_l)(x_i - \mu_l)^T$, $l = 1, \cdots, M$. Using EM algorithm, in each E-step, given the estimated parameters Θ^g, we compute $p(\mathbf{x}, y_i = l|\Theta^g)$ using the forward-backward algorithm, $l = 1, \cdots, M$. In each M-step, we solve the maximum expected likelihood estimation problem to update the parameters associated with each state. Different optimization algorithms can be applied to solve the optimization problem. Let $v_{l,k} = v_l(k)$, $k = 1, \cdots, r$. For the Phylo-HMGP-BM, the gradient with respect to $v_{l,k}$ can be computed explicitly [128] and we implemented the gradient descent method based on the derived gradient as an alternative optimization approach:

$$\frac{\partial \tilde{L}(\Theta^{(l)})}{\partial v_{l,k}} = \nabla_{G_k} \tilde{L}(\Theta^{(l)}) = \Sigma_{i=1}^{N} w_i^{(l)} \mathrm{tr}(G_k [\Sigma_{v,\mu}^{(l)}]^{-1}) - N \mathrm{tr}(\tilde{S}_{\mu}^{(l)} [\Sigma_{v,\mu}^{(l)}]^{-1} G_k [\Sigma_{v,\mu}^{(l)}]^{-1}). \tag{2.66}$$

2.6 Initial estimation of the state number for comparing RT data across species

To apply Phylo-HMGP to the replication timing data, we first estimated the possible number of states using K-means clustering. We performed K-means clustering to the datasets with an increasing cluster number K, computed the Sum of Squared Error (SSE) of each clustering result, and observed how SSE changed with respect to K. We estimated the state number to be approximately 20-40 based on the K-means clustering results, as the decreasing rate of SSE with respect to the increasing K slows down in this range (Figure 2.3). Based on the observation the 'elbow point' of the SSE curve is around 30, and small fluctuation of the state number around 30 dose not present significant change of the reduction of the SSE decreasing rate compared to the state number of 30. We therefore set the state number to be 30.

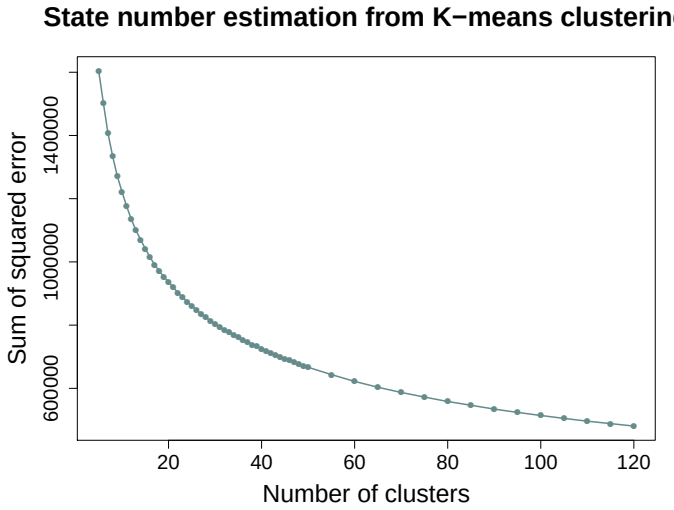

Figure 2.3: The change of Sum of Squared Error (SSE) with respect to an increased cluster number in K-means clustering on RT data. The state number was estimated to be between 20 and 40 based on the results from K-means clustering.

2.7 Data preparation and processing

2.7.1 Experimental model and subject details

The five primate species included in the RT data comparison are Homo Sapiens (Human), Pan troglodytes (Common Chimpanzee), Pongo pygmaeus (Bornean Orangutan), Nomascus leucogenys (Northern White Cheeked Gibbon), and Cercopithecus aethiops (Green Monkey). We used GM12878 cell line from human. The GM12878 cell line is a lymphoblastoid cell line established from EBV (Epstein-Barr Virus)-transformed B-lymphocytes from a female donor. The GM12878 cell line was obtained from the Coriell Cell Repositories of Coriell Institute for Medical Research. We used lymphoblastoid cell lines from the other four non-human primate species, each of which is from one biological individual. The lymphoblastoid cell line of each of the species was derived from B-lymphoctyes by EBV transformation. The cells of Common Chimpanzee (abbreviated as Chimpanzee) are male. The cells of Bornean Orangutan (abbreviated as Orangutan) are male. The cell lines of Chimpanzee and Orangutan have been used in [131]. The cells of Northern White Cheeked Gibbon (abbreviated as Gibbon) are male. The cells of Green Monkey are female and the cell line was obtained from the Coriell Cell Repositories. For each species, we only used autosomes for data analysis and excluded data from the sex chromosomes. The key resources used are shown in Table 2.1.

2.7.2 Repli-seq data processing

We generated Repli-seq data from the lymphoblastoid cell line of each of the five primate species. We performed quality control of the Repli-seq reads using the software FastQC (available from http://www.bioinformatics.babra-ham.ac.uk/projects/fastqc) and removed adapter sequences using the FASTX-Toolkit (http://hannonlab.csh-l.edu/fastx_toolkit) for data preprocessing. To obtain RT signals in orthologous genome regions across the multiple species, we collected the RT signal values for each 6kb bin of human genome and its orthologous regions in each of the other species if RT measurements are available. Specif-

ically, first, we mapped the preprocessed sequencing reads to the genome assemblies of hg19 (Human), panTro4 (Chimpanzee), ponAbe2 (Orangutan), nomLeu3 (Gibbon), and chlSab2 (Green Monkey), respectively, using Bowtie2 [132]. The genome assemblies were downloaded from the UCSC genome browser [133–137]. Second, we used human genome (hg19) as the reference and divided the reference genome into 6kb bins. We then aligned each bin in human genome to each of the other species with reciprocal mapping using liftOver [138] to obtain the orthologous regions. Third, for each species, we calculated Repli-seq read count within a given genomic window (an orthologous region) in early and late phases of RT, respectively, normalized by the total read count in early or late RT phase on the whole genome accordingly. The RT signal in each orthologous region is defined as the base 2 logarithm ratio of read count per million reads between the early and late phases of RT within this region.

For each species, we identify each sequence of consecutive bins without RT signals as a gap. The bin size (6kb in human) is much smaller than the scale of the RT signals (the replication domain is typically at the scale of 400-800kb [42]). We assume that RT does not change sharply at a small size gap if the gap is between both early RT signals or both late RT signals. We then performed data imputation for gaps smaller than 48kb using nearest neighbor imputation. In this way we can reduce missing data and have more continuous segments where cross-species observations are available. More specifically, if the RT signals on both sides of a gap smaller than 48kb are both early RT signals or both late RT signals with difference smaller than 1/3, we assign to each bin in the gap the RT signal of a signal-available bin that is nearest to this bin. We then used the software HMMSeg [139] to perform wavelet smoothing [140] of the RT signals in each species, using the window size of 24kb.

Next, we found the orthologous regions where the RT signals across five species are all available. We then performed data normalization of the signals of each species in the regions. We observed that the different species have varied RT scales around [-5,5]. We performed feature scaling to scale non-negative RT signals (primarily early RT) in each

species to [0,5] and scale non-positive RT signals (primarily late RT) in each species to [-5,0]. We formed the normalized RT signals in orthologous regions across five species into a five-dimensional feature vector and assigned it to the corresponding reference 6kb bin in human genome as a sample. We excluded the orthologous regions on the sex chromosome and only used autosomes of each species for data analysis. We obtained 419,754 samples in the orthologous regions across species.

2.8 RT state prediction and RT state grouping

We applied Phylo-HMGP-OU to multi-species Repli-seq data (there are five primate species in our study) to perform state estimation, with the state number set to be 30. We used $\lambda_0 = 4.0$ for l_2-norm regularization and $w_1 = 0.2$ for parameter initialization (see section 2.4.3). We repeated the estimation 10 times with different initializations. We choose the result with the highest objective function value for further analysis.

We classified the 30 RT states predicted by Phylo-HMM-OU into 5 RT groups, namely, conserved early (noted as E), conserved late (L), weakly conserved early (WE), weakly conserved late (WL), and non-conserved (NC). If the majority ($>98\%$) of the regions in a state share the pattern that all of the five species consistently have positive RT signals (early in RT), we assign this state to the conserved early (E) group. If a state does not satisfy this criteria, but instead satisfy that at least four species are consistently early in RT in more than 90% regions of this state, we assign this state to the weakly conserved early (WE) group. We assign states to the L and WL groups in a similar way accordingly. The remaining states are assigned to the NC group.

2.9 Approaches for the simulation studies

2.9.1 Data simulation for the simulation studies

We performed performance evaluation of the develop method in simulation studies. We used two types of models for data simulation, corresponding to Simulation Study I (SS-I) and Simulation Study II (SS-II). Each study consists of six synthetic datasets. In SS-I, for each dataset, samples were generated from an HMM with 10 states and with multivariate Gaussian distribution as the emission probability distribution. The Gaussian distribution of each state follows a different OU model on the same phylogenetic tree topology. The OU model parameters of each state were randomly generated with non-negative constraints of selection strength $\{\alpha_j\}_{j=1}^{M}$ and Brownian motion intensity $\{\sigma_j\}_{j=1}^{M}$ (M is the state number). Phylogenetic trees with four leaf nodes and with five leaf nodes were used as tree topologies for parameter simulation, each used for three datasets. The transition probability matrix of the HMM was randomly generated with the assumption that self-transition probability of a state is the dominant probability as compared to probabilities of transitions to other states.

In SS-II, samples were generated from a Gaussian mixture model instead of an HMM. For each dataset, samples were generated based on a mixture model with 10 states where Gaussian distributions are the emission probability distributions. We defined a transition probability matrix between the 10 states as we defined in SS-I, and computed the equilibrium probability distribution of the 10 states from the transition probability matrix. We then divided the genome into continuous segments of varied lengths. Each segment represents a series of samples that share the same state, e.g., adjacent fixed-size bins of the same state on the genome. We randomly sampled the segment length from a truncated Normal distribution by which the length is non-negative. The state of each segment was drawn randomly and independently from the computed equilibrium probability distribution of the 10 states. Parameters of the Gaussian distribution of each state were shared between two corresponding datasets in SS-I and SS-II. For example, simulation dataset I-1 (dataset 1 in SS-I) and simulation dataset II-1 (dataset 1 in SS-II) are assigned with the same set of Gaussian dis-

tributions for 10 states. However, different assumptions (HMM and non-HMM models) were used to simulate the two types of datasets.

In both simulation study I (SS-I) and II (SS-II), phylogenetic trees with four leaf nodes and with five leaf nodes were used as tree topologies. Datasets with even-number (I-2, I-4, I-6, II-2, II-4, and II-6) were based on the same topology of five leaf nodes, which is identical to the topology of the species tree specified in Figure 2.1C and is also the topology used in the RT data study. Datasets with odd-number (I-1, I-3, I-5, II-1, II-3, and II-5) were based on the same topology of four leaf nodes, which is identical to the topology of the subtree of the species tree specified in Figure 2.1C that contains the primate species human, chimpanzee, orangutan, and gibbon. 50,000 samples were generated for each dataset.

The emission probability distribution of each state in each dataset is Gaussian distribution, parameterized by a multivariate OU model $\psi_j = (\theta_j, \alpha_j, \sigma_j)$, where j is the index of the state, and θ_j, α_j, σ_j represent the optimal value vector, the selection strengths and the Brownian motion intensities along the branches, respectively. The selection strength $\alpha_{j,k}$ and Brownian motion intensity $\sigma^2_{j,k}$ along each branch are each randomly and independently sampled from the uniform distribution $Unif[0, 2]$, $k = 1, \cdots, d$ (d is the number of branches). The optimal value $\theta_{j,l}$ ($l = 1, \cdots, d + 1$) of each node is randomly and independently sampled from a Normal distribution $\boldsymbol{\mathcal{N}}(0, 2)$. In the transition probability matrix A of one dataset in SS-I, the self-transition probability of state j is defined as $a_{jj} = a_0 + (1 - a_0) \times p$, where p is randomly sampled from uniform distribution $Unif[0, 1]$ and a_0 is set to be 0.7. The transition probabilities of state j to other states are first randomly sampled from uniform distribution $Unif[0, 1]$ and then normalized to be summed to $1 - a_{jj}$. In SS-II, the fragment length (the number of continuous bins of the same state) is sampled from the a truncated Normal distribution $\boldsymbol{\mathcal{N}}(50, 30)$ with the minimal fragment length to be 5. We first sampled a transition probability matrix $\tilde{A}$ in the same way as in SS-I. Then we estimated the equilibrium probability distribution $\tilde{\pi}$ of the states from $\tilde{A}$ based on $\tilde{\pi} = \tilde{\pi}\tilde{A}$. We sampled the state of each fragment from $\tilde{\pi}$ randomly and independently.

We calculated the Davies-Bouldin Index (DBI) [141] for each dataset in SS-I and SS-

II, to estimate the difficulty in state prediction in different datasets. DBI can be used to measure how discriminative is each cluster (state) compared to the others. A high DBI represents that the states have large variances within themselves while the state-to-state distances are small, making it difficult to distinguish the states. The DBIs for the six datasets in SS-I are 2.3127, 2.0770, 1.9706, 1.3127, 1.5045 and 1.4623, respectively. The DBIs for datasets in SS-II are 2.2677, 2.0608, 1.9597, 1.3116, 1.4987, and 1.4864, respectively. We found that datasets I-1, I-2, II-1, and II-2 have relatively higher DBIs.

2.9.2 Performance evaluation in the simulation studies

We used Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), Precision, Recall, and F_1 score [142, 143] for performance evaluation in the simulation studies. Suppose $X = \{x_1, \cdots, x_N\}$ is the set of samples. Suppose $\Omega = \{\omega_1, \cdots, \omega_K\}$ is the set of predicted states which represents a partition of S into K states, and $C = \{c_1, \cdots, c_M\}$ is the ground truth set of states. Let $I(\Omega, C)$ be the mutual information between Ω and C, and $NMI(\Omega, C)$ be the normalized mutual information. We have

$$I(\Omega; C) = \sum_{k=1}^{K} \sum_{j=1}^{M} P(\omega_k, c_j) \log \frac{P(\omega_k, c_j)}{P(\omega_k) P(c_j)}, \tag{2.67}$$

$$NMI(\Omega; C) = \frac{I(\Omega; C)}{[H(\Omega) + H(C)]/2}, \tag{2.68}$$

where $H(\Omega)$ and $H(C)$ represent the entropies of Ω and C, respectively. The entropy is defined as $H(\Omega) = -\sum_{k=1}^{K} P(\omega_k) \log P(\omega_k)$. $P(\omega_k)$, $P(c_j)$, and $P(\omega_k, c_j)$ represent the probabilities that a sample is in state ω_k, in state c_j, and in both ω_k and c_j, respectively. The maximum likelihood estimates of $P(\omega_k)$, $P(c_j)$, and $P(\omega_k, c_j)$ are $|\omega_k|/N$, $|c_j|/N$, and $|\omega_k \cap c_j|/N$, respectively, where $|\omega_k|$ denotes the size of ω_k and N is the number of samples.

Adjusted Mutual Information (AMI) is an adjustment of the mutual information to cor-

rect the effect of agreement between two partitions that is solely due to chance. We have

$$AMI(\Omega; C) = \frac{I(\Omega; C) - \mathbb{E}[I(\Omega; C)]}{\max\{H(\Omega), H(C)\} - \mathbb{E}[I(\Omega; C)]}, \tag{2.69}$$

where $\mathbb{E}(I(\Omega; C))$ represents the expectation of $I(\Omega; C)$, which can be estimated using Ω and C [143].

The Rand Index (RI) [142] is another metric to compare two partitions, which is defined as

$$RI = \frac{TP + TN}{TP + FP + FN + TN}. \tag{2.70}$$

TP (true positive) represents the number of pairs of samples in X that are in the same subset in Ω and also in the same subset in C. FP (false positive) is the number of pairs of samples in X that are in the same subset in Ω but in different subsets in C. FN (false negative) is the number of pairs of samples in X that are in different subsets in Ω but in the same subset in C. TN (true negative) is the number of pairs of samples in X that are in different subsets in Ω and also in different subsets in C.

The Adjusted Rand Index (ARI) corrects the Rand Index for the effect of agreement that is solely due to chance between partitions. ARI is defined as

$$ARI = \frac{RI - \mathbb{E}[RI]}{\max\{RI\} - \mathbb{E}[RI]}, \tag{2.71}$$

where $\mathbb{E}(RI)$ represents the expectation of RI.

Precision, Recall, and F_1 score are defined as

$$Precision = \frac{TP}{TP + FP}, \tag{2.72}$$

$$Recall = \frac{TP}{TP + FN}, \tag{2.73}$$

$$F_1 = \frac{2Precision \times Recall}{Precision + Recall}. \tag{2.74}$$

For the compared methods, we used the GaussianMixture function and the KMeans function in the scikit-learn library [144] to implement the Gaussian Mixture Model (GMM) method and the K-means clustering method, respectively. We used the hmmlearn library

(https://github.com/hmmlearn/) to implement the Gaussian-HMM method. Each compared method is repeated 10 times with different initializations and guaranteed convergence each time. Specifically, the parameter initializations of the Gaussian-HMM method and the GMM method were based on K-means clustering. In each experiment, the best result from 10 randomly-initialized K-means clustering results (based on the clustering evaluation criteria used in hmmlearn or scikit-learn, respectively) was selected for estimating the initial parameters of Gaussian-HMM or GMM, respectively. 10 random initializations were also used for K-means clustering in each experiment, and the clustering with the best performance was chosen as the result. 10 experiments were repeated for the compared methods as well as Phylo-HMGP, and the average of the 10 runs was reported as the final performance of the corresponding method.

2.10 Results

2.10.1 Simulation study demonstrates the robustness of Phylo-HMGP

To explore whether incorporating evolutionary temporal constraints into the HMM can improve the accuracy of identifying different evolutionary patterns, we applied our method to 12 synthetic datasets in two types of simulation studies. We assessed the performance based on Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), Precision, Recall, and F_1 score [142, 143] by comparing the predicated states with the ground truth states (see section 2.9.2 We used HMM to generate the samples in simulation study I (SS-I), while simulation study II (SS-II) did not use HMM and was instead based on a Gaussian Mixture Model (GMM). Both SS-I and SS-II contained six synthetic datasets (sample size = 50,000 each), respectively. Detailed descriptions of the simulated datasets are in section 2.9.1.

We compared Phylo-HMGP-OU and Phylo-HMGP-BM with the Gaussian-HMM method, the GMM method, and the K-means clustering method in both SS-I and SS-II. For each dataset, we ran each method 10 times. Each method was started from different initializa-

tions and given the state number as 10. We reported the average performance of the 10 runs as the final performance of the respective method. We applied the same regularization parameter to Phylo-HMGP-OU on all of the 12 datasets, without tuning the parameter specifically on each dataset. The results show that Phylo-HMGP-OU outperforms the other methods on AMI, ARI, and F_1 score on all of the six datasets in SS-I (Figure 2.4A and Table. 2.3). In particular, Phylo-HMGP-OU shows significant advantage in reaching higher ARI on average in different datasets, as compared to the other methods. In SS-II (Figure 2.4B and Table. 4.4), the performance of Phylo-HMGP-OU decreases occasionally (SS-II-1 and II-2) as compared to its performance in SS-I. However, Phylo-HMGP-OU still outperforms the other methods in five of the six datasets. Phylo-HMGP-BM reaches the highest performance on SS-I-1, while Phylo-HMGP-OU maintains comparable performance to Phylo-HMGP-BM on this dataset. These simulation results strongly suggest that Phylo-HMGP-OU can achieve robust performance even when the data are simulated from a non-HMM model such as the Gaussian mixture model. Note that in the rest of the Results section, we use "Phylo-HMGP" to refer to "Phylo-HMGP-OU".

2.10.2 Phylo-HMGP reveals genome-wide RT patterns across primate species

Next, we applied the Phylo-HMGP method to study different evolutionary patterns of RT in mammalian genomes. We generated genome-wide RT maps based on Repli-seq [38] in lymphoblastoid cells from five primate species, including human, chimpanzee, orangutan, gibbon, and green monkey. See Methods section for the details on how we processed the data. We then applied Phylo-HMGP to this multi-species RT dataset. We set the state number as 30 based on estimation from K-means clustering (see section 2.6 and Figure 2.3). We identified both conserved and lineage-specific states with differences in RT patterns across species. Here we classified the 30 states into five groups: conserved early (denoted as E), conserved late (L), weakly conserved early (WE), weakly conserved late (WL), and non-conserved (NC) (see section 2.8). In the E group, all five species have early RT. In the WE group, four species have early RT. We assign states to the L group and the WL group

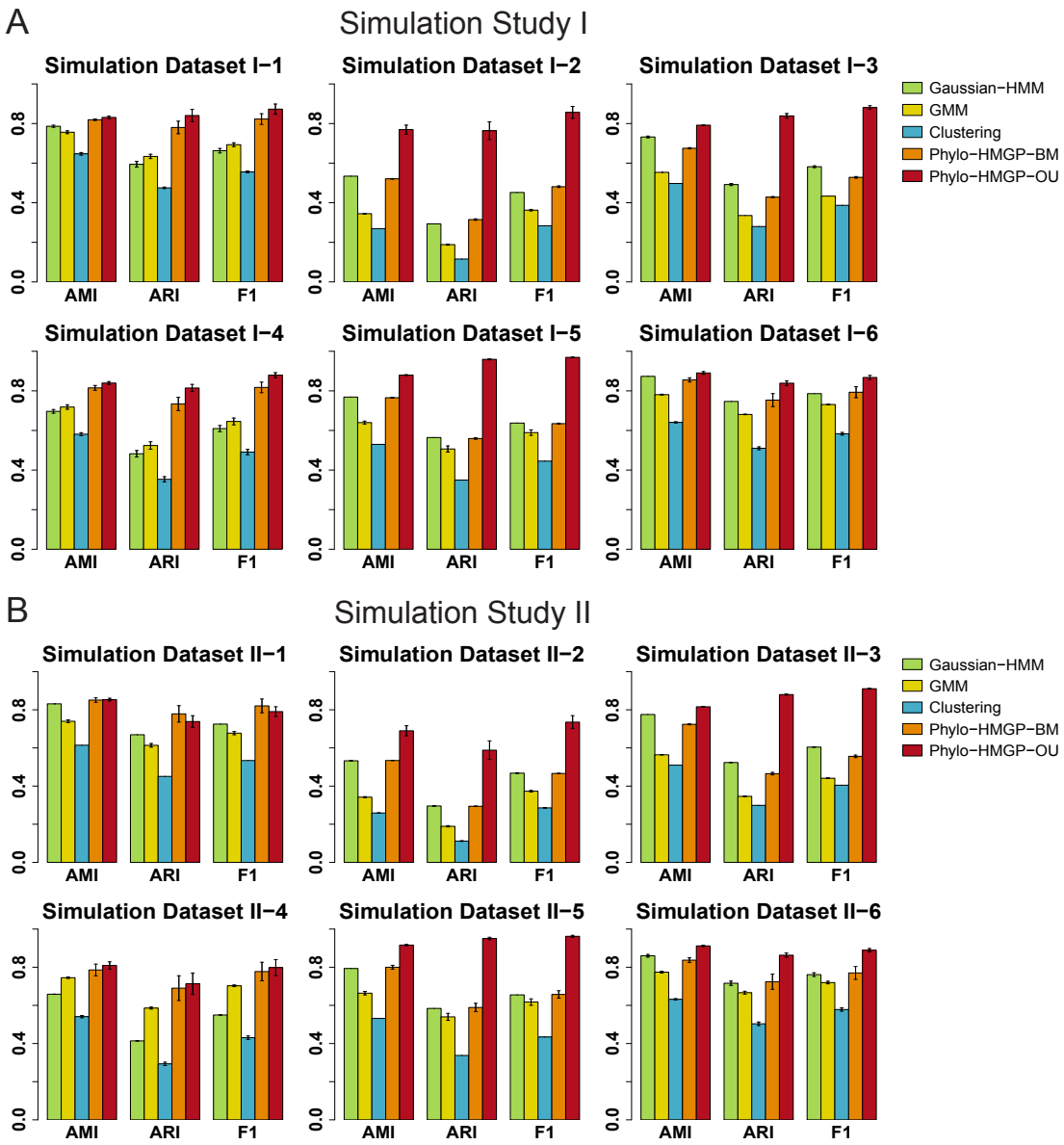

Figure 2.4: Prediction performance evaluation using simulated datasets. **(A)** Evaluation of Gaussian-HMM, GMM, K-means Clustering, Phylo-HMGP-BM, and Phylo-HMGP-OU on six simulation datasets in Simulation Study I in terms of AMI (Adjusted Mutual Information), ARI (Adjusted Rand Index), and F_1 score. **(B)** Evaluation of Gaussian-HMM, GMM, K-means Clustering, Phylo-HMGP-BM, and Phylo-HMGP-OU on six simulation datasets in Simulation Study II in terms of AMI, ARI, and F_1 score. In both **(A)** and **(B)**, the standard error of the results of 10 repeated runs for each method is also shown as the error bar.

similarly. The remaining states are assigned to the NC group.

The representative RT signal patterns of the 30 predicted states are shown in Figure 2.5A, with examples of the states and groups shown in Figure 2.5B and Figure 2.5D.

Distributions of RT signals of the five species in each of the 30 states are shown in Figure 2.6, including other lineage-specific patterns, conserved patterns, or divergent patterns. States 1-8 are conserved early or conserved late states of RT, making up approximately 47.7% of the whole genome. States 9-18 display different lineage-specific RT patterns. State 9 (Figure 2.5B) and 10 represent human-chimpanzee (hominini) specific patterns of early RT and late RT, respectively. State 11 shows human-chimpanzee-orangutan (hominid) specific early RT. States 12-18 reflect single-lineage specific patterns, where one species differs from all the other species.

Phylo-HMGP estimated the transition probabilities between the 30 predicted states (Figure 2.7). We noticed that overall the transition probabilities are higher within the E group and L group. Phylo-HMGP also simultaneously estimated the model parameters of selection strength, Brownian motion intensity, and optimal values of the phylogenetic model associated with each state (Figures 2.8 and 2.9). We found that the estimated parameters correspond very well to the lineage-specific RT patterns. For example, for state 9 and 10, the human-chimpanzee specific states, the estimated strongest selection strength happened on the branch leading to human and chimpanzee, and strong Brownian motion intensity is also estimated for human and chimpanzee. We observed similar correlations for other states. We also compared Phylo-HMGP with the other methods on an evaluation dataset constructed from the RT data and found that Phylo-HMGP outperforms other methods (see section 2.10.3 and Figure 2.10).

2.10.3 Evaluation of Phylo-HMGP in comparsion with other methods on RT data

We also compared Phylo-HMGP-OU with Gaussian-HMM method, GMM method, K-means clustering method, and Phylo-HMGP-BM on the Repli-seq dataset, based on the average performance from 10 repeated runs of each method. We have applied each method to the RT data for state prediction, with state number set to be 30, as estimated in section 2.6. However, there is no available ground truth for the RT data. For evaluation

purpose, we constructed an evaluation state set. Specifically, we discretized the signals of each species into 5 levels, and identified 12 possible selected representative states (10 possible lineage-specific states and two conserved states) from all the combinations of the 5 levels in orthologous regions across the species. We fit a Gaussian-HMM with five hidden states independently for each species, with each state representing a discretized level of the RT signal values. High signal values (level 1 and 2) and low signal values (level 3 and 4) correspond to early phase and late phase in RT, respectively. For example, human early state represents early RT only in human and non-early RT in the other four species at the orthologous regions. The 10 possible lineage-specific states identified from discrete levels of RT signals are human early/late, chimpanzee early/late, orangutan early/late, gibbon early/late, and green monkey early/late, respectively. The two identified conserved states are conserved RT early and late, respectively. The 12 selected states cover around 60% of all the orthologous regions with cross-species RT signals. We constrained the evaluation of different methods to the regions where the 12 selected states are present. Within these regions, we used the 12 selected states as a known partition, and evaluated the relevance of the prediction of each method to this partition, using the evaluation metrics AMI, NMI, RI, and F_1 score (Figure 2.10).

As each method predicted 30 states, which is a finer partition than 12 states, the evaluation measures are generally lower than those in the simulation studies. For example, regions in one state in the 12-state partition may be predicted to be in different states in the 30-state partition, which affects the F_1 score by reducing the Recall and also affects the other metrics. Also, the selected states used for comparison were estimated using predictions from Gaussian-HMM in each species, which would favor Gaussian-HMM and GMM. The regions where 12 selected states are present are less continuous than the whole genome regions and have weaker spatial dependencies between regions, which again would favor GMM. However, Phylo-HMGP-OU still outperforms the other methods in each of the four evaluation metrics. Phylo-HMGP-BM ranks second in performance. Even though the 12-state partition is not exactly ground truth, it nevertheless demonstrates that Phylo-HMGP

outperforms the other methods in the RT data application, which is consistent with the results from the simulation studies.

2.10.4 RT evolution patterns correlate with A/B compartments and histone marks

Analysis based on Hi-C data has shown that the genome can be divided into two compartments called A/B compartments [1], with at least five subcompartments, namely A1, A2, B1, B2 and B3, which have different genomic and epigenomic properties [4]. A1 and A2 subcompartments both show early RT, with the difference that replications in A2 regions finish later than A1. B2 and B3 subcompartments show late RT, while replications in B1 happen in the middle of S-phase [4]. We used the subcompartment definitions in the human lymphoblastoid cell line GM12878 from [4] and calculated the enrichment of the five subcompartments in the 30 predicted RT states. We observed that different predicted RT evolution patterns show distinct enrichments of the subcompartments. For example, the predicted RT states in the E group (state 1-4) show the strongest correlation with A1 or A2, while the predicted RT states in the L group are enriched with B2 and B3. The majority of the states in the NC group are most enriched with A2 or B1. States in the WE group and WL group are enriched with A2/B1, and B2/B3, respectively.

We next compared the enrichments of different histone marks and CTCF binding site within each RT state. Figure 2.5A panel 3 shows the enrichment distributions of histone marks and CTCF binding site across the five predicted RT groups. These distributions are consistent with the epigenomic feature patterns of the subcompartments that are enriched in the corresponding states. We found that RT states in the E group show strong positive correlation with active histone marks (e.g., H3K27ac, H3K36me3, and H3K4me1) and the CTCF binding sites. On the contrary, RT states in the L/WL groups show distinct depletion of these histone marks and the CTCF binding sites. The majority of predicted states in the NC group instead exhibit variations in the enrichments of different types of histone marks.

Among the NC states, state 9 is identified as a human-chimpanzee specific early RT

state (Figure 2.5B). It displays a unique pattern of histone mark enrichment, showing the strongest correlation with H2A.Z (p-values<1e-07) as compared to other predicted states. Recent studies have reported that H2A.Z is progressively enriched towards early RT loci [145]. Another state with interesting features is state 4, a conserved early state. The RT is significantly early in human in state 4, similar to other states in the E group. All of the other states in the E group (state 1-3) are strongly correlated with A1 subcompartment. State 4, however, is enriched with A2 subcompartment and is more positively correlated with H3K9me3, which generally has stronger enrichment in A2 than A1 [4]. Therefore, state 4 represents a distinct state in the E group. These results demonstrate that Phylo-HMGP has the sensitivity to distinguish within similar evolutionary patterns of RT.

2.10.5 Different RT evolution patterns reflect different functions

Previous studies have shown that different genomic regions have different levels of cell type specificity for RT, including constitutively early, constitutively late, and more dynamic across different cell types [9, 62]. We compared the states from Phylo-HMGP with the constitutive and developmental RT patterns discovered during ES cell differentiation [9], including constitutively early (CE), constitutively late (CL), developmentally regulated (D), and undetermined. We found that overall the constitutively early or constitutively late RT regions in the human genome have high consistency with the strongly conserved RT evolution patterns (Figure 2.5C). The findings are consistent with previous observations in human-mouse RT comparison [62].

Among the CE regions that are also covered in the cross-species RT comparisons by Phylo-HMGP, 99.45% of the regions are assigned to the states of conserved early or weakly conserved early (p-value<2.2e-16). Also, 86.94% of the CL regions in human are within states of conserved late or weakly conserved late (p-value<2.2e-16). In contrast, the D regions show more diverse patterns across the five RT groups predicted by Phylo-HMGP. This also suggests that the RT regions in lymphoblastoid cells with similar RT profile across different cell types are highly likely to be conserved in primates. However, a significant

fraction of conserved RT regions in primates also shows cell-type specific RT patterns in human. We performed Gene Ontology (GO) analysis for the conserved RT early regions with respect to the constitutive/non-constitutive RT patterns using DAVID [146], and found clear differences in gene functions (Table 2.4). We further performed GO analysis for the lineage-specific RT states (see Figure 2.11A and Table 2.5). We found that genes associated with different states have different functions and biological processes. For example, the hominini-specific early RT state (state 9) is enriched with genes having sensory functions. These analyses suggest that genomic regions with different RT evolution patterns may contain genes with distinct functions.

2.10.6 Boundaries of RT evolution patterns correlate with TAD boundaries

Earlier studies discovered that TADs defined from Hi-C data have high correspondence with replication domains [9, 42]. We next asked whether the states found by Phylo-HMGP correlate with the boundaries of TADs. We used the TADs called by two methods, Directionality Index (DI) [8] and Arrowhead [4]. We named the TADs as DI TADs and Arrowhead TADs, the median lengths of which are 440kb and 185kb, respectively. For each boundary of a TAD, we calculated the distance between the TAD boundary and the nearest state boundary from Phylo-HMGP. Specifically, to filter the TADs that are far away from any predicted states, we extended each boundary of a TAD with 30kb and used the states that overlap with the extended TAD to calculate the boundary distance. We then calculated the percentages of boundary distances that fall into four distance intervals respectively. The first interval is [0,12kb]. The remaining three intervals are determined by the empirical distance distribution obtained from TAD shuffling, and equally cover the distances that are larger than 12kb. We shuffled the TADs 1000 times by randomly relocating them along the genome. We calculated and merged the boundary distances of each shuffle of TADs to form the empirical boundary distance distribution. Furthermore, for each shuffle of TADs, we computed the percentage of boundary distances that fall into each distance interval to form empirical distributions for each interval.

We found that the boundary distances between the DI TADs and the predicted RT states are significantly more enriched in the interval [0,12kb] than expected (Figure 2.11B, empirical p-value$<$1e-03). The percentage drops in the intervals that correspond to increased boundary distances. The percentage is significantly lower than expected in the fourth interval that covers the largest distances (empirical p-value$<$1e-03). The comparison based on Arrowhead TADs show similar results. This analysis demonstrates the correlation between the boundaries of RT evolution states and the TAD boundaries.

2.10.7 RT evolution patterns have enrichment of different transposable elements

It is known that RT correlates with certain transposable element (TE) families, e.g., the early RT regions are typically enriched with SINE elements [37]. We next looked at the connection between RT evolution patterns and the involvement of TEs based on RepeatMasker annotation. We obtained the RepeatMasker annotations for each of the five primate species from the UCSC Genome Browser [147]. For the TE families shared among the five primate species, we calculated the fold change of their enrichment in the orthologous regions of each species in each state (Figure 2.11C). We found that there exist distinct patterns of TE enrichment across different RT states and RT groups. Alu elements are strongly involved in conserved early RT states and depleted in conserved late RT states across the five species, with a clear changing correlation with RT across the five RT groups. On the contrary, L1 and LTR elements ERVL and ERV1 correlate negatively with early RT but positively with late RT. TEs in the LTR class and DNA class generally have more diversity in their distributions over states in the WE, WL, and NC groups. We also found that the repetitive sequence elements srpRNA, scRNA, and snRNA (based on RepeatMasker annotations) have a strong positive correlation with conserved early RT and negative correlation with conserved late RT (p-value$<$1e-04), having a similar enrichment pattern to Alu in the E, WL, and L groups. Although some of these correlations (such as those with srpRNA, scRNA, and snRNA) have not been reported before and further investigations are needed,

this nevertheless demonstrates the potential of our method to provide new insights into the impact of sequence evolution on DNA replication timing.

2.10.8 Lineage-specific early RT regions harbor unique TFBS

We then asked whether there are specific transcription factor binding sites (TFBS) that are enriched in regions with specific types of RT evolution patterns. We used FIMO [148] to perform motif scanning in the orthologous open chromatin regions of each species, using 635 position weight matrices (PWMs) of TF binding motifs from the JASPAR 2016 core vertebrate motif database [149]. Specifically, we identified open chromatin regions in human genome as DNase-seq peak regions with +/-250bp extension, using DNase-seq data of the GM12878 cells downloaded from the ENCODE annotation data in the UCSC genome browser [150]. We used the liftOver tool [138] to project the identified open chromatin regions in human genome to genomes of the other primate species, obtaining orthologous open chromatin regions in the other species. In each orthologous region, we computed the motif frequency for each of the PWMs within the open chromatin area for each species (p-value$<$1e-04 required for each motif). We then normalized the frequency by the open chromatin area size within this orthologous region.

To identify TF binding motifs that may be lineage-specifically enriched in predicted lineage-specific RT states, we used two types of tests and selected motifs that can pass both tests. First, within each lineage-specific RT state, we performed binomial tests to find the motifs that are significantly more enriched in the RT-specific species than expected (p-value$<$0.05). Second, for a motif that passes the binomial test in a lineage-specific RT state, we calculated the fold change of its motif frequency within the RT-specific species compared to the other species. To estimate the empirical p-value, we randomly sampled the same number of regions as the lineage-specific RT state from the whole genome for 2000 times, and calculated the same type of fold change to form the empirical distribution. We selected the motifs that have empirical p-values$<$0.05.

We identified sets of motifs that show lineage-specific enrichment in the lineage-specific

early RT states (Figure 2.11D). We checked whether the TFs in the lineage-specific RT states that involve human are expressed and found that the majority of them are expressed. Specifically, we analyzed the expressions (in human) of those TFs with enriched motifs in the lineage-specific states that involve human (i.e., state 9 and state 11). The gene expression data were obtained from the ENCODE Project [151] (ENCODE Data Coordination Center accession: ENCSR000AEC; GEO accession: GSE78550). We found that 24 out of the 28 TFs (86%) associated with state 9 and 11 have FPKM greater than 0.01 (10/13 for state 9 and 14/15 for state 11). If we use the lower bound of the 95% credible interval for the FPKM greater than 0.1 as the threshold, 20 out of the 28 TFs (71%) associated with state 9 and 11 are expressed (10/13 for state 9 and 10/15 for state 11). For the TFs with low or no expression, we further searched for the most similar binding motifs using TOMTOM [152]. We found that all the TFs for the matching motifs are expressed. Therefore, if highly similar motifs are also considered, all the identified motifs correspond to expressed TFs.

Also, we found that the identified lineage-specific enriched TF binding motifs vary in different states. However, there are still a number of TF binding motifs (or motifs with similar PWMs) shared between different states. For example, FOXC1 is significantly enriched in human and chimpanzee in the hominini-specific state (state 9), and also enriched in green monkey in the green monkey-specific state (state 18). Interestingly, many of corresponding TFs associated with species-specific early RT are from the FOX family (e.g., FOXC1, FOXO3, and FOXD1), the ELF family (e.g., ELF1 and ELF3), and the ETV family (e.g., ETV3 and ETV6). TFs of the FOX family are known regulators in B cells [153] (lymphoblastoid cells are B cells) and FOXO3 was previously found to be crucial for regulating cell cycle progression through its binding partnership with DNA replication factor Cdt1 [154]. Many of the other identified TFs are also known regulators in B cells, such as EBF1, IRF8, RUNX2, and POU5F1 [153]. Although these findings need further studies to evaluate the functional significance of the corresponding TFs in lineage-specific biology, our analysis points to the direction that connects lineage-specific changes in *cis*-regulatory

elements with lineage-specific changes in RT.

2.10.9 Evaluation based on *cis*-regulatory module evolution

In addition to the real data application on the Repli-seq data, we also applied the models to predict different states of *cis*-regulatory module (CRM) evolution along the genome using features only from DNA sequences. We focused on a recent dataset for promoters and enhancers marked by H3K4me3 and H3K27ac in vertebrate liver cells [48]. We used four species, including human (hg19), macaque (rheMac2), marmoset (calJac3), and mouse (mm10). We used hg19 as the reference and divide it into 5 kb bins. For each of the orthologous regions, we used the method Cluster-Buster [155] to compute a CRM score for presence of homotypic motif clusters within this region of the respective species, using a selected collection of 382 position weight matrices of TF binding motifs from the JASPAR 2016 core vertebrate motif database [149]. We only used expressed TFs in liver cell based on gene expression data of human liver from GSE61260 [156]. We computed CRM sores for the 286,287 orthologous regions across the four species. We applied Phylo-HMGP to perform state prediction along the genome, with the state number set to be 16. Here we assumed that the calculated CRM scores are associated with the activities of regulatory elements (e.g., enhancers or promoters). The ChIP-seq data, which can be used to identify and validate the existence of regulatory elements such as enhancers or promoters, were used to prepare benchmarks to evaluate the performance of the proposed model Phylo-HMGP in discovering different CRM patterns across species.

We used the peak regions called from ChIP-seq data of histone modification H3K27ac and H3K4me3 [48] to evaluate the different states estimated by Phylo-HMGP. For enhancer-evolution associated state prediction, we segmented the reference genome into different benchmark states based on the species-specific distribution of H3K27ac peaks. We then compared the states predicted by Phylo-HMGP-OU with the benchmark states, in comparison with the results from Gaussian-HMM, K-means clustering, and Phylo-HMGP-BM. We also performed state evaluation using the H3K4me3 dataset. The results are shown in

Figure 2.12. Phylo-HMGP-OU achieved the highest RI and F_1 score among the different methods in the four experiments. Although the overall accuracy of using the CRM score for predicting enhancer/promoter activities is not high and it remains an open problem to more accurately predict regulatory region activities from genome sequence, our evaluation again demonstrates the general utility and advantage of Phylo-HMGP.

2.11 Discussion

We developed Phylo-HMGP, which is a new continuous-trait probabilistic model for more accurate genome-wide evolutionary state estimation based on features from different species using functional genomic signals. The Phylo-HMGP model establishes a new integrated framework to utilize the continuous-trait evolutionary model with spatial constraints to more effectively study the heterogeneous evolutionary feature patterns encoded in the genome-wide functional genomic datasets across multiple species. Both simulation studies and real data application demonstrate the advantage of Phylo-HMGP as compared to other methods. Importantly, we generated a new cross-species RT dataset from the same cell type in five primate species (human, chimpanzee, orangutan, gibbon, and green monkey) to study RT evolution patterns in primates for the first time using Phylo-HMGP. Our results from the comparative RT analysis demonstrate the potential of the model to help reveal regions with conserved or lineage-specific regulatory roles for the entire genome.

There are a number of areas that our model can be further improved. For Phylo-HMGP, the number of model parameters increases linearly with the number of species. There can be many local minima in parameter estimation for large scale evolutionary trees. Therefore, both more effective parameter constraints in accordance with the tree structure and more effective optimization methods need to be developed. Also, hierarchical state estimation methods can be developed to to group similar predicted patterns for state prediction refinement. In addition, the current Phylo-HMGP assumes that all the phylogenetic tree models have the same tree topolopy. But in certain application domains this may not be accurate.

Therefore, it would be useful to improve the model by incorporating inference of alternative tree topologies [157]. Furthermore, we need to improve the interpretation of the estimated model parameters of the evolutionary models associated with the predicted states, to gain deeper understanding of the evolutionary mechanisms underlying the different functional genomic feature patterns.

Genetic variation can contribute to differences in RT [158–160]. Our current study has the limitation that it does not specifically consider the impact of intra-species variation on RT evolutionary patterns we identified. We did, however, compare the RT variant loci (among different individuals) identified in human lymphoblastoid cells [158] with the cross-species RT evolution states we found. We observed that the RT variations among individuals are distributed sparsely on the genome, with a small percentage of the whole genome and of each predicted RT evolution state. This suggests that the impact of the intra-species variation on RT patterns across different species we found is likely to be very minor. However, it would be an important methodological improvement to model both the inter-species differences and intra-species variations when population level functional genomic data are available for various cell types in different species.

We believe that Phylo-HMGP provides a generic framework to more precisely capture the evolutionary history of functional genomic signals across different species. In addition to the cross-species RT comparisons, we also applied Phylo-HMGP to predict the evolution of *cis*-regulatory modules and demonstrated the advantage and the generic utility of our new method (see section 2.10.9 and Figure 2.12). From the application to the RT data, we found that different RT evolution patterns predicted by Phylo-HMGP correlate with RT patterns across different cell types and various other genomic and epigenomic features, including higher-order genome organization features, *cis*-regulatory elements, transposons, and gene functions. Such insights from comparative functional genomic analysis may in turn help interpret the impact of sequence evolution on genome organization and function. One important future direction would be to develop more integrated models to holistically consider sequence features (from mutations and small insertions/deletions to large-scale

genome rearrangements) and functional genomic signals across multiple species.

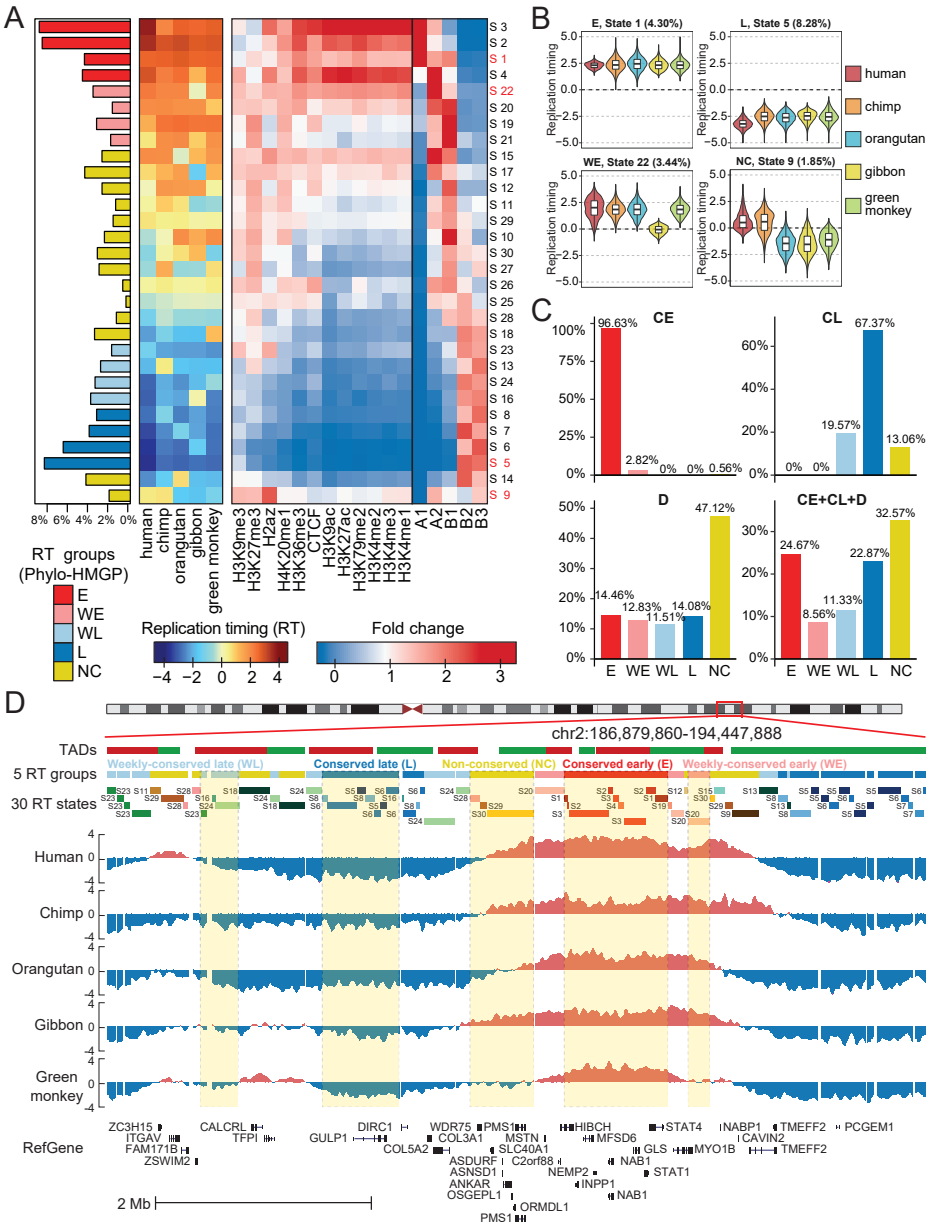

Figure 2.5: RT evolution patterns identified by Phylo-HMGP. **(A)** Panel 1 (leftmost): Proportions of the 30 RT states on the entire genome. The RT states are categorized into 5 groups: conserved early (E), weakly conserved early (WE), weakly conserved late (WL), conserved late (L), and other stages (NC), respectively. Panel 2: Patterns of the 30 states. Each row of the matrix corresponds to the state at the same row in Panel 1, and columns are species. Each entry represents the median of the RT signals of the corresponding species in the associated state. Panel 3: Enrichment of different types of histone marks and CTCF binding site (higher fold change represents higher enrichment). Panel 4: Enrichment of subcompartment A1, A2, B1, B2, and B3. **(B)** Four examples of RT signal distributions in states with different patterns (State 1: E; State 5: L; State 22: WE; State 9: NC with human-chimpanzee specific early RT). **(C)** Comparison of predicted RT patterns with the constitutively early/late RT regions identified across cell types. **(D)** Examples of different RT states and RT groups in five species predicted by Phylo-HMGP. TADs called by the Directionality Index method are shown at the top.

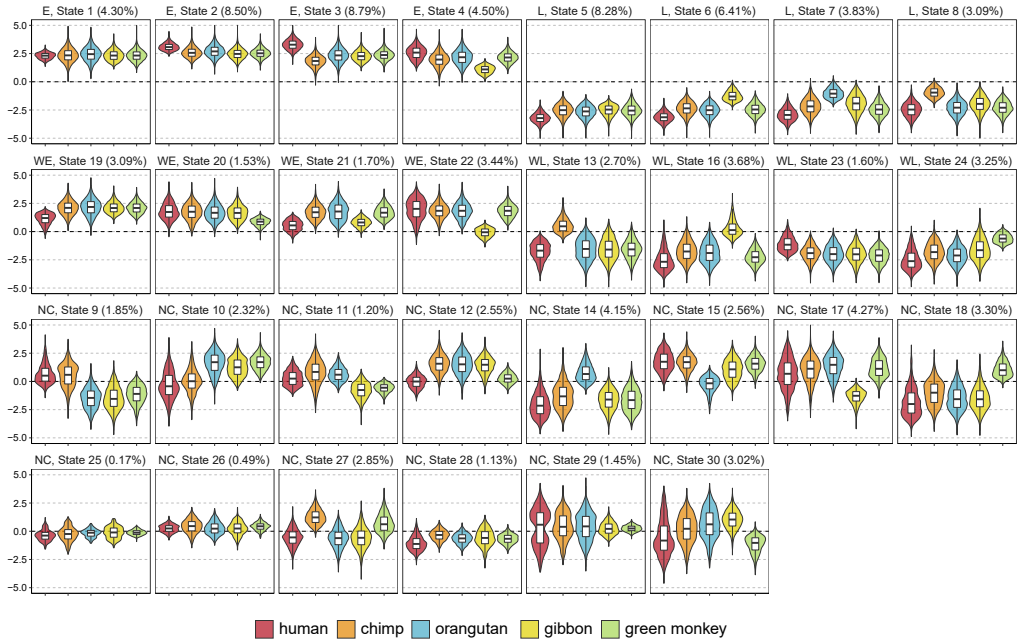

Figure 2.6: Different patterns of replication timing (RT) across five primate species predicted by Phylo-HMGP-OU for 30 states. The y-axis represents the RT signal value. Box plots of the RT signal distributions of the five primate species in each predicted state are shown. The percentage of the number of regions in each predicted state and the RT group label are also shown in the title of each plot.

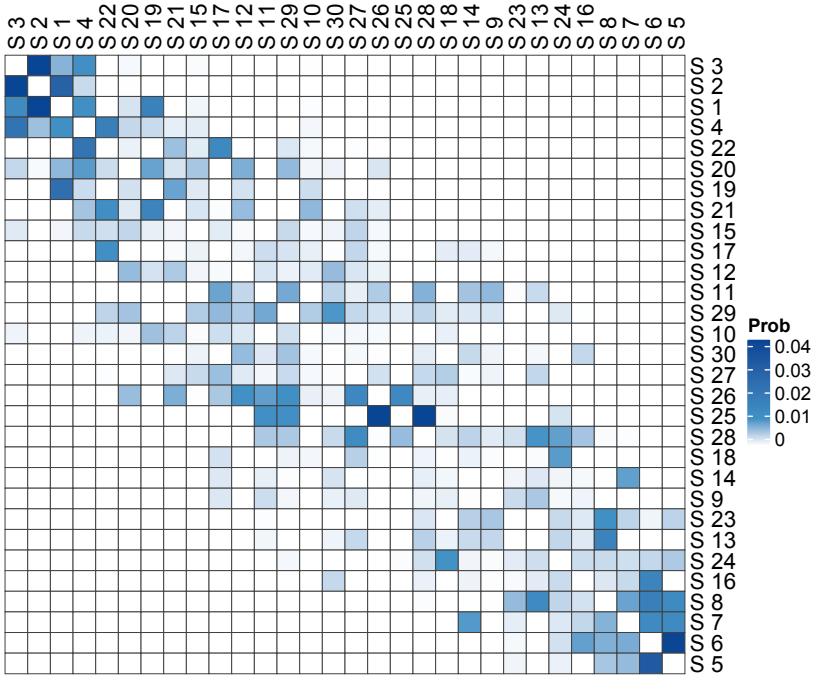

Figure 2.7: Transition probability matrix of the 30 states estimated by Phylo-HMGP-OU. The self-transition probability is not shown (set to be blank) to illustrate the probabilities of transitions to other states more significantly.

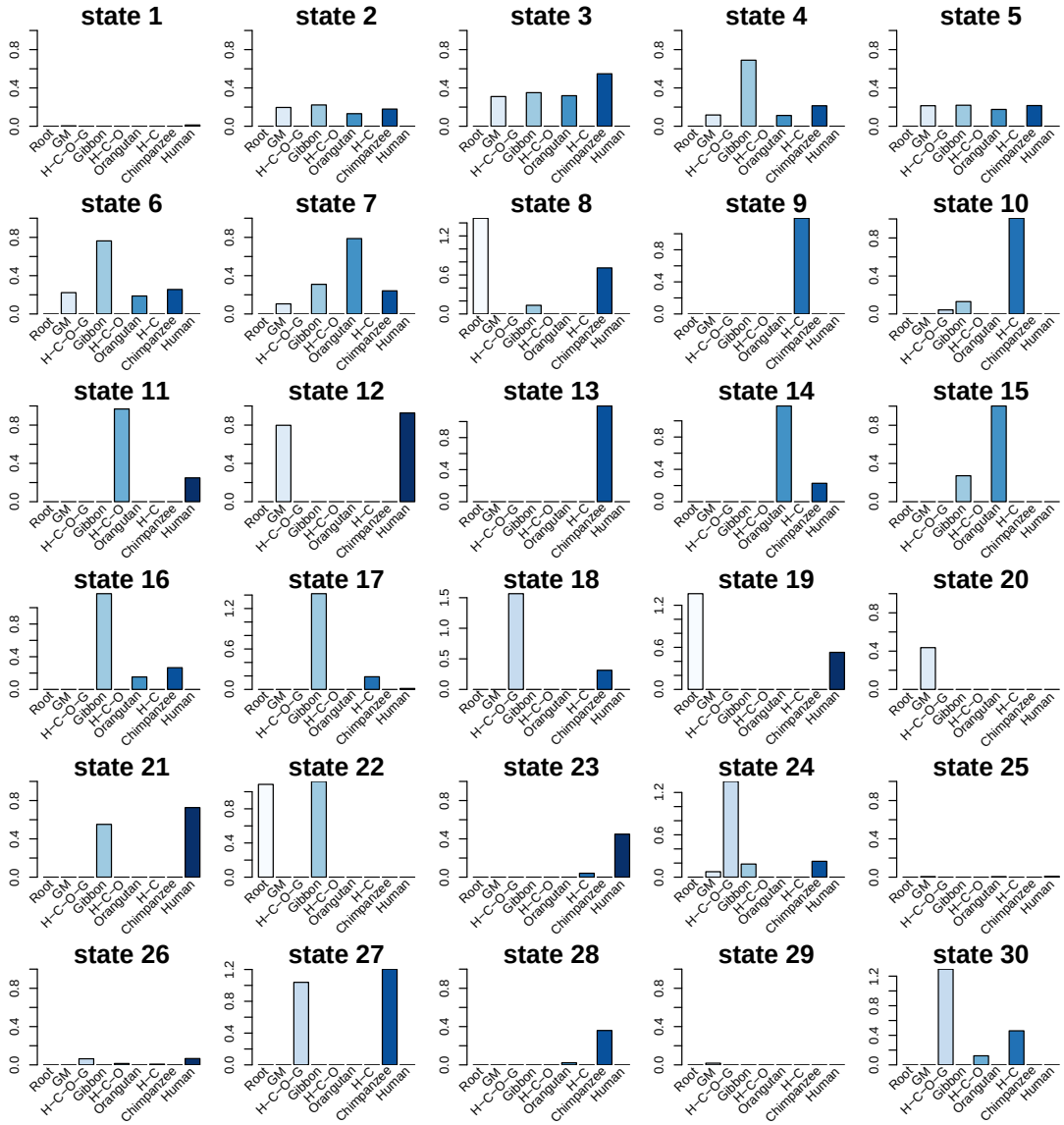

Figure 2.8: Estimated selection strength along each branch of the phylogenetic tree in the 30 states predicted by Phylo-HMGP-OU. 'GM' stands for Green Monkey. Each column corresponds to the branch connecting the nearest ancestor of the species specified by the species name to the species. Root stands for the branch connecting the remote root node ancestor with the nearest common ancestor of green monkey and human. H-C-O-G, H-C-O, and H-C represent the branches leading to the clade of human, chimpanzee, orangutan, and gibbon, the clade of human, chimpanzee, and orangutan, and the clade of human and chimpanzee, respectively.

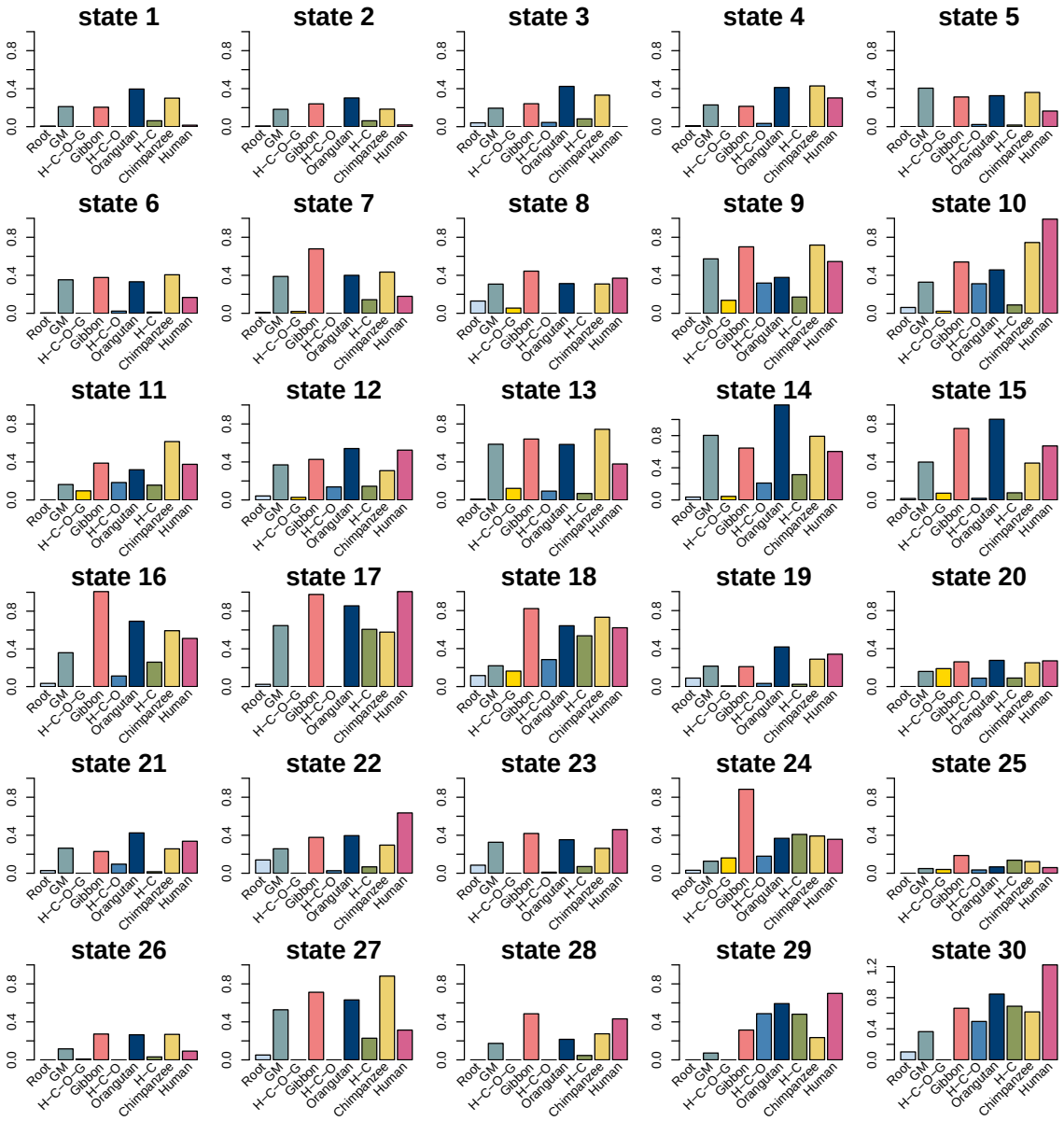

Figure 2.9: Estimated Brownian motion intensity along each branch of the phylogenetic tree in the 30 states predicted by Phylo-HMGP-OU. 'GM' stands for Green Monkey. Each column corresponds to the branch connecting the nearest ancestor of the species specified by the species name to the species. Root stands for the branch connecting the remote root node ancestor with the nearest common ancestor of green monkey and human. H-C-O-G, H-C-O, and H-C represent the branches leading to the clade of human, chimpanzee, orangutan, and gibbon, the clade of human, chimpanzee, and orangutan, and the clade of human and chimpanzee, respectively.

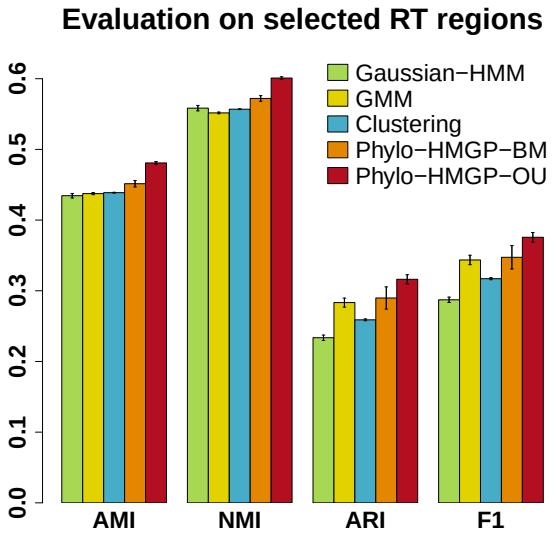

Figure 2.10: Performance evaluation of different methods on RT data. Evaluation of Gaussian-HMM, GMM, K-means Clustering, Phylo-HMGP-BM, and Phylo-HMGP-OU on the RT dataset in terms of AMI, NMI, ARI and F_1 score in the genomic regions of 12 different states (including 10 lineage-specific states and two conserved states) identified from comparison of discretized single-species observations. The standard error of the results of 10 repeated runs for each method is shown as the error bar.

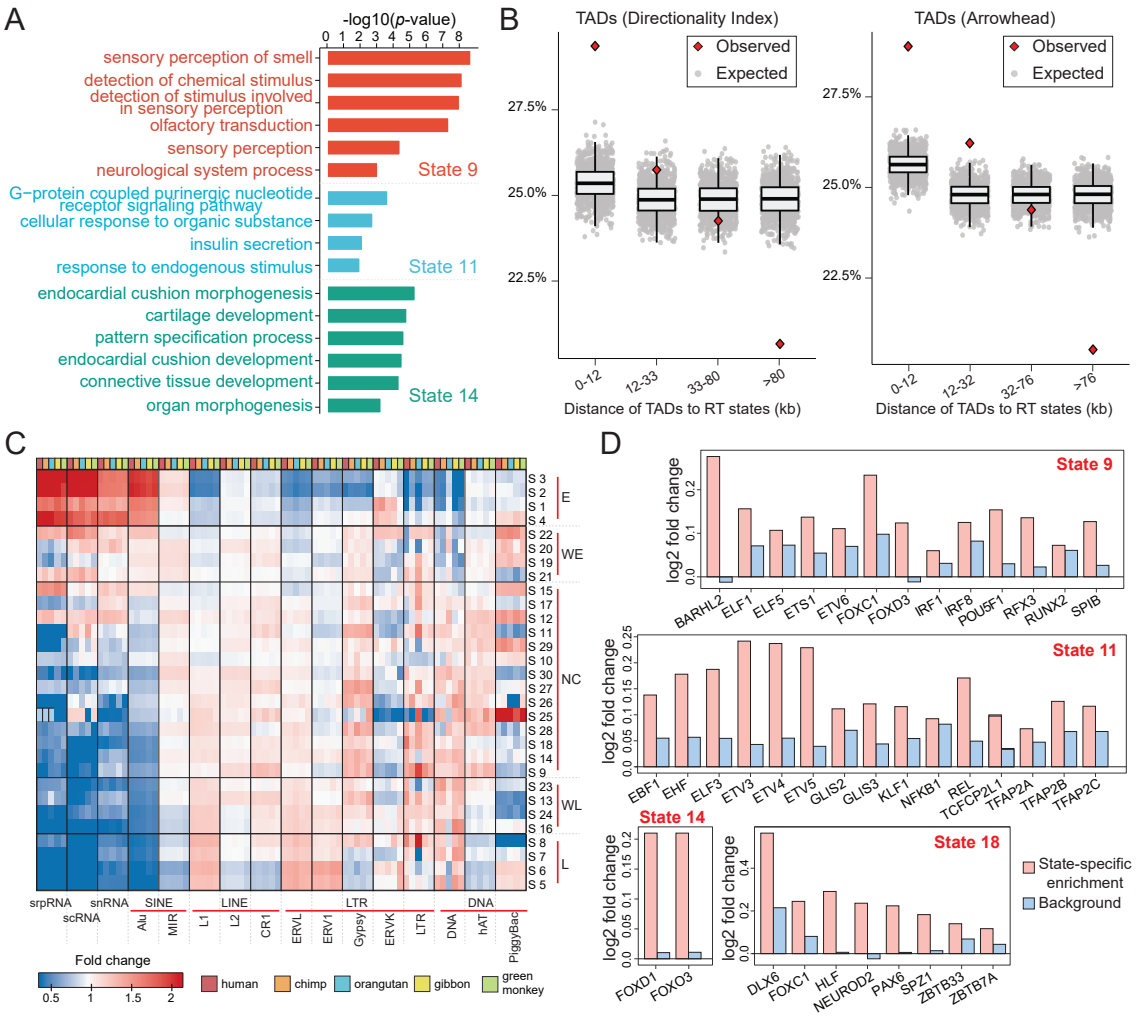

Figure 2.11: Comparisons between the RT evolution patterns and other genomic features. **(A)** Example gene ontology (GO) analysis results of state 9, state 11, and state 14. **(B)** Percentages of the distances between TAD boundaries and boundaries of predicted states in different intervals. The expected distances are calculated based on randomly shuffled TADs. Two types of TADs from different methods are used, namely TADs called by the Directionality Index method and TADs called by Arrowhead. **(C)** Transposable element enrichment in different RT states. **(D)** Motif enrichment in different lineage-specific RT states. State 9: human-chimpanzee specific early RT. State 11: human-chimpanzee-orangutan specific early RT. State 14: orangutan specific early RT. State 18: green monkey specific early RT. The GO analysis results of the other lineage-specific RT states are included in Table 2.5.

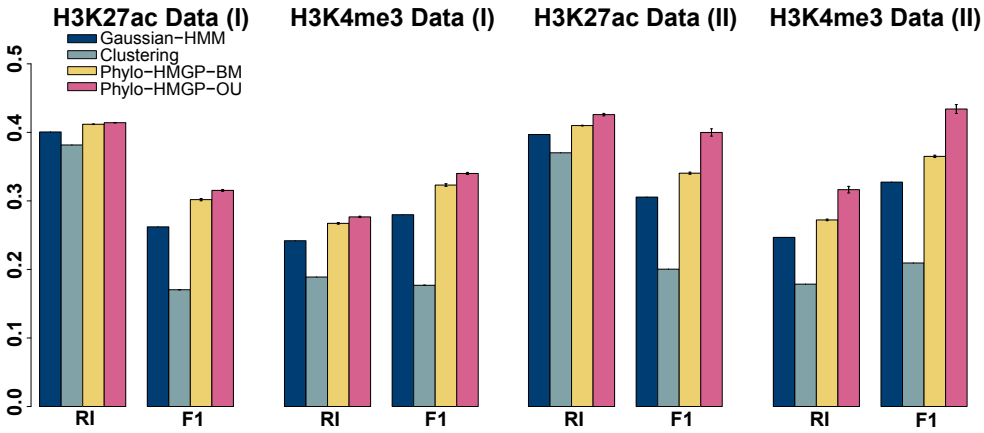

Figure 2.12: Performance evaluation of different methods based on *cis*-regulatory module evolution. Evaluation of Gaussian-HMM, K-means Clustering, Phylo-HMGP-BM, and Phylo-HMGP-OU on H3K27ac and H3K4me3 ChIP-seq datasets in terms of RI (Rand Index), and F_1 score. Experiments (I) represent they are perforemd for the four mammal species. Experiments (II) represent they are performed for the three primate species human, macaque, and marmoset. The standard error of the results of 10 repeated runs for each method is shown as the error bar.

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-BrdU antibody	BD biosciences	Cat#555627; RRID:AB_395993
Anti-mouse IgG	Sigma-Aldrich	Cat#M7023; RRID:AB_260634
Deposited Data		
Repli-seq data of five primate species	This part of study	GEO: GSE111733
Experimental Models: Cell Lines		
Human: lymphoblastoid GM12878	Coriell Cell Repositories	Cat#GM12878; RRID:CVCL_7526
Pan troglodytes (Common Chimpanzee): lymphoblastoid cell line	E. Eichler and M. Ventura (Johnson et al., 2006)	PTR
Pongo pygmaeus lymphoblastoid cell line (Bornean Orangutan):	E. Eichler and M. Ventura (Johnson et al., 2006)	PPY
Nomascus leucogenys (Northern White Cheeked Gibbon): lymphoblastoid cell line	L. Carbone	NLE
Cercopithecus aethiops (Green Monkey): lymphoblastoid cell line	Coriell Cell Repositories	Cat#PR01205; RRID:CVCL_2Y01
Software and Algorithms		
Phylo-HMGP	This part of study	https://github.com/ma-compbio/Phylo-HMGP
FastQC	web portal	http://www.bioinformatics.babraham.ac.uk/projects/fastqc
FASTX-Toolkit	web portal	http://hannonlab.cshl.edu/fastx_toolkit
Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
liftOver	Hinrichs et al.,2006	https://genome.ucsc.edu/cgi-bin/hgLiftOver
HMMSeg	Day et al.,2007	https://noble.gs.washington.edu/proj/hmmseg
Other		
Heat-inactivated FBS	Seradigm	Premium Grade HI FBS 1500-500H Lot #: 035B15
Repli-seq protocol	Marchal et al., 2018	N/A

Table 2.1: Table of key resources used in the study of comparing continuous-trait functional genomic data across multiple species using Phylo-HMGP.

Simulation	Method	AMI	NMI	ARI	Precision	Recall	F_1
Dataset I-1	Gaussian-HMM	0.7863	0.8216	0.5941	0.8206	0.5569	0.6629
Dataset I-1	GMM	0.7562	0.7985	0.6340	**0.8903**	0.5679	0.6932
Dataset I-1	Clustering	0.6471	0.6926	0.4752	0.7539	0.4412	0.5558
Dataset I-1	Phylo-HMM-BM	0.8190	0.8494	0.7801	0.8496	0.8122	0.8226
Dataset I-1	Phylo-HMM-OU ($\lambda_0 = 0$)	0.7966	0.8183	0.6088	0.7973	0.5918	0.6786
Dataset I-1	Phylo-HMM-OU ($\lambda_0 = 4.0$)	**0.8309**	**0.8733**	**0.8405**	0.8568	**0.8986**	**0.8728**
Dataset I-2	Gaussian-HMM	0.5342	0.6901	0.2931	0.9666	0.2948	0.4519
Dataset I-2	GMM	0.3446	0.4508	0.1884	0.8114	0.2335	0.3625
Dataset I-2	Clustering	0.2692	0.3599	0.1157	0.7173	0.1768	0.2837
Dataset I-2	Phylo-HMM-BM	0.5208	0.6621	0.3150	0.9531	0.3218	0.4810
Dataset I-2	Phylo-HMM-OU ($\lambda_0 = 0$)	0.5315	0.6782	0.3149	0.9602	0.3191	0.4790
Dataset I-2	Phylo-HMM-OU ($\lambda_0 = 4.0$)	**0.7691**	**0.8229**	**0.7638**	**0.9811**	**0.7718**	**0.8565**
Dataset I-3	Gaussian-HMM	0.7323	0.7912	0.4919	0.8314	0.4476	0.5819
Dataset I-3	GMM	0.5536	0.6200	0.3354	0.7442	0.3057	0.4334
Dataset I-3	Clustering	0.4973	0.5561	0.2798	0.6580	0.2741	0.3870
Dataset I-3	Phylo-HMM-BM	0.6752	0.7277	0.4282	0.7646	0.4045	0.5283
Dataset I-3	Phylo-HMM-OU ($\lambda_0 = 0$)	0.7146	0.7698	0.4741	0.8088	0.4372	0.5674
Dataset I-3	Phylo-HMM-OU ($\lambda_0 = 4.0$)	**0.8309**	**0.8733**	**0.8405**	**0.8568**	**0.8986**	**0.8728**
Dataset I-4	Gaussian-HMM	0.6965	0.7967	0.4824	0.9585	0.4482	0.6095
Dataset I-4	GMM	0.7185	0.8106	0.5241	0.9737	0.4849	0.6457
Dataset I-4	Clustering	0.5812	0.6810	0.3535	0.8928	0.3395	0.4911
Dataset I-4	Phylo-HMM-BM	0.8151	0.8644	0.7342	**0.9756**	0.7112	0.8174
Dataset I-4	Phylo-HMM-OU ($\lambda_0 = 0$)	0.7946	0.8544	0.6846	0.9736	0.6589	0.7776
Dataset I-4	Phylo-HMM-OU ($\lambda_0 = 4.0$)	**0.8396**	**0.8607**	**0.8146**	0.9687	**0.8071**	**0.8792**
Dataset I-5	Gaussian-HMM	0.7688	0.8331	0.5644	0.9091	0.4908	0.6374
Dataset I-5	GMM	0.6398	0.6949	0.5065	0.8366	0.4552	0.5893
Dataset I-5	Clustering	0.5298	0.5906	0.3494	0.7227	0.3218	0.4453
Dataset I-5	Phylo-HMM-BM	0.7654	0.8243	0.5594	0.8908	0.4929	0.6346
Dataset I-5	Phylo-HMM-OU ($\lambda_0 = 0$)	0.7738	0.8370	0.5809	0.9180	0.5057	0.6521
Dataset I-5	Phylo-HMM-OU ($\lambda_0 = 4.0$)	**0.8798**	**0.9187**	**0.9594**	**0.9524**	**0.9867**	**0.9692**
Dataset I-6	Gaussian-HMM	0.8736	**0.9141**	0.7466	**0.9642**	0.6635	0.7861
Dataset I-6	GMM	0.7810	0.8157	0.6810	0.8884	0.6215	0.7312
Dataset I-6	Clustering	0.6409	0.6785	0.5101	0.7457	0.4791	0.5833
Dataset I-6	Phylo-HMM-BM	0.8554	0.8772	0.7534	0.9052	0.7133	0.7932
Dataset I-6	Phylo-HMM-OU ($\lambda_0 = 0$)	0.8538	0.8891	0.7216	0.9325	0.6490	0.7652
Dataset I-6	Phylo-HMM-OU ($\lambda_0 = 4.0$)	**0.8906**	0.8958	**0.8390**	0.9224	**0.8200**	**0.8678**

Table 2.2: Performance evaluation in simulation study I. Performance evaluation of Gaussian-HMM, GMM (Gaussian Mixture Model), K-means Clustering, Phylo-HMGP-BM, Phylo-HMGP-OU ($\lambda_0 = 0$), and Phylo-HMGP-OU ($\lambda_0 = 4.0$) on six simulated datasets in simulation study I with respect to AMI (Adjusted Mutual Information), NMI (Normalized Mutual Information), ARI (Adjusted Rand Index), Precision, Recall, and F_1 score. Each method is repeated 10 times with different initializations on each simulation dataset. The average performance from the 10 repeated runs of each method is presented. The best performance of the compared methods is in bold font.

Simulation	Method	AMI	NMI	ARI	Precision	Recall	F_1
Dataset II-1	Gaussian-HMM	0.8314	0.8765	0.6693	**0.9244**	0.5965	0.7251
Dataset II-1	GMM	0.7399	0.7823	0.6138	0.8800	0.5511	0.6776
Dataset II-1	Clustering	0.6144	0.6623	0.4509	0.7579	0.4124	0.5341
Dataset II-1	Phylo-HMM-BM	0.8511	**0.8796**	**0.7786**	0.8785	**0.7842**	**0.8205**
Dataset II-1	Phylo-HMM-OU	**0.8534**	0.8706	0.7385	0.8319	0.7591	0.7906
Dataset II-2	Gaussian-HMM	0.5328	0.6863	0.2957	0.9550	0.3099	0.4680
Dataset II-2	GMM	0.3421	0.4477	0.1893	0.8206	0.2416	0.3733
Dataset II-2	Clustering	0.2590	0.3481	0.1119	0.7278	0.1780	0.2860
Dataset II-2	Phylo-HMM-BM	0.5343	0.6877	0.2944	0.9550	0.3085	0.4663
Dataset II-2	Phylo-HMM-OU	**0.6901**	**0.7833**	**0.5883**	**0.9736**	**0.6033**	**0.7353**
Dataset II-3	Gaussian-HMM	0.7754	0.8307	0.5236	**0.8469**	0.4700	0.6045
Dataset II-3	GMM	0.5641	0.6248	0.3463	0.7225	0.3176	0.4412
Dataset II-3	Clustering	0.5098	0.5606	0.2991	0.6385	0.2963	0.4048
Dataset II-3	Phylo-HMM-BM	0.7239	0.7727	0.4652	0.7793	0.4338	0.5561
Dataset II-3	Phylo-HMM-OU	**0.8159**	**0.8952**	**0.8797**	**0.8469**	**0.9848**	**0.9105**
Dataset II-4	Gaussian-HMM	0.6585	0.7623	0.4139	0.9318	0.3898	0.5497
Dataset II-4	GMM	0.7451	0.8262	0.5859	**0.9778**	0.5486	0.7027
Dataset II-4	Clustering	0.5411	0.6463	0.2933	0.8678	0.2868	0.4307
Dataset II-4	Phylo-HMM-BM	0.7857	0.8411	0.6904	0.9543	0.6812	0.7775
Dataset II-4	Phylo-HMM-OU	**0.8092**	**0.8556**	**0.7135**	0.9583	**0.7045**	**0.7981**
Dataset II-5	Gaussian-HMM	0.7932	0.8529	0.5842	0.9078	0.5125	0.6551
Dataset II-5	GMM	0.6640	0.7155	0.5399	0.8595	0.4827	0.6173
Dataset II-5	Clustering	0.5317	0.5894	0.3375	0.7028	0.3142	0.4343
Dataset II-5	Phylo-HMM-BM	0.7999	0.8595	0.5891	0.9223	0.5139	0.6573
Dataset II-5	Phylo-HMM-OU	**0.9162**	**0.9454**	**0.9504**	**0.9555**	**0.9694**	**0.9622**
Dataset II-6	Gaussian-HMM	0.8600	0.8969	0.7162	**0.9411**	0.6396	0.7613
Dataset II-6	GMM	0.7742	0.8118	0.6665	0.8887	0.6050	0.7197
Dataset II-6	Clustering	0.6319	0.6733	0.5036	0.7554	0.4689	0.5784
Dataset II-6	Phylo-HMM-BM	0.8371	0.8612	0.7241	0.8788	0.6944	0.7701
Dataset II-6	Phylo-HMM-OU	**0.9117**	**0.9150**	**0.8632**	0.9314	**0.8506**	**0.8888**

Table 2.3: Performance evaluation in simulation study II. Performance evaluation of Gaussian-HMM, GMM, K-means Clustering, Phylo-HMGP-BM, and Phylo-HMGP-OU ($\lambda_0 = 4.0$) on six simulated datasets in simulation study II with respect to AMI, NMI, ARI, Precision, Recall, and F_1 score. Each method is repeated 10 times with different initializations on each simulation dataset. The average performance from the 10 repeated runs of each method is presented. The best performance of the compared methods is in bold font.

Conserved RT early	GO term/Pathway	Count	Fold enrichment	p-value
Constitutive RT early	intracellular transport	431	1.2	6.4e-06
	amide biosynthetic process	214	1.3	1.5e-05
	posttranscriptional regulation of gene expression	149	1.4	2.0e-05
	cellular amide metabolic process	277	1.2	2.0e-05
	peptide biosynthetic process	195	1.3	2.4e-05
	peptide metabolic process	231	1.3	2.5e-05
	mRNA metabolic process	193	1.2	5.8e-05
	translation	187	1.3	5.9e-05
	microtubule-based process	179	1.3	1.0e-04
	clathrin-mediated endocytosis	22	2.1	3.0e-04
	regulation of vascular permeability	18	2.3	3.9e-04
	mitochondrion organization	183	1.2	7.2e-04
Non-constitutive RT early	immune response	540	1.3	7.9e-013
	regulation of immune response	327	1.3	2.7e-010
	defense response	520	1.2	5.5e-08
	symbiosis, encompassing mutualism through parasitism	376	1.2	5.7e-08
	interspecies interaction between organisms	376	1.2	5.7e-08
	viral process	362	1.2	1.2e-07
	multi-organism cellular process	364	1.2	1.7e-07
	immune effector process	257	1.3	2.6e-07
	immune response-activating signal transduction	176	1.4	7.9e-07
	activation of immune response	192	1.3	1.1e-06
	immune response-regulating signaling pathway	186	1.3	1.2e-06
	regulation of immune system process	465	1.2	2.5e-06

Table 2.4: Gene ontology (GO) terms or pathways that show significant correlation with regions that are both constitutive RT early and conserved RT early, and regions that are conserved RT early but not constitutive RT early. We used DAVID [146] to perform the gene ontology analysis. The column Count represents the number of genes found to be associated with the corresponding GO term/pathway in the queried regions. The conserved early RT regions are based on state prediction by Phylo-HMGP. We observed that genes enriched in the regions that are both constitutive early and conserved early are mainly involved in basic biological functions and processes that are shared between different cell types. Genes associated with the regions that are not constitutive early but conserved early are involved in the cell type specific functions of the lymphoblastoid cells, such as the immune response functions.

Predicted state	GO term/Pathway	Count	Fold enrichment	p-value
State 9	sensory perception of smell	16	8.5	1.9e-09
	detection of chemical stimulus	17	6.6	6.3e-09
	detection of stimulus involved in sensory perception	17	6.4	9.2e-09
	olfactory transduction	16	6.0	4.1e-08
	sensory perception	21	2.9	3.6e-05
	neurological system process	23	2.2	8.2e-04
State 10	regulation of nucleic acid-templated transcription	95	1.5	1.4e-05
	regulation of RNA biosynthetic process	95	1.5	1.9e-05
	regulation of nucleobase-containing compound metabolic process	101	1.4	1.4e-04
	aromatic compound biosynthetic process	107	1.4	1.5e-04
	heterocycle biosynthetic process	106	1.4	2.1e-04
	regulation of nitrogen compound metabolic process	104	1.3	5.8e-04
	cellular macromolecule biosynthetic process	114	1.3	1.1e-03
State 11	G-protein coupled purinergic nucleotide receptor signaling pathway	4	33.2	2.0e-04
	cellular response to organic substance	32	1.8	1.6e-03
	insulin secretion	6	5.0	6.7e-03
	response to endogenous stimulus	23	1.8	9.6e-03
State 14	endocardial cushion morphogenesis	7	15.6	4.4e-06
	cartilage development	13	4.9	1.4e-05
	endocardial cushion development	7	11.6	2.7e-05
	connective tissue development	14	4.1	4.1e-05
	pattern specification process	18	2.8	2.1e-04
	organ morphogenesis	29	2.0	5.1e-04
	enzyme linked receptor protein signaling pathway	27	2.0	1.1e-03
	circulatory system development	27	2.0	1.3e-03
	cardiovascular system development	27	2.0	1.3e-03
State 16	cell adhesion	27	2.2	1.4e-04
	regulation of nervous system development	15	2.6	1.5e-03
	regulation of neurogenesis	13	2.6	4.3e-03
	regulation of cellular component organization	29	1.7	4.5e-03
	neuron projection morphogenesis	11	2.8	6.1e-03
	neuron differentiation	18	2.0	8.3e-03
State 18	regulation of locomotion	32	2.6	1.4e-06
	regulation of cell motility	31	2.7	1.6e-06
	regulation of cell migration	29	2.7	3.5e-06
	regulation of cell differentiation	45	1.9	4.0e-05
	epithelium development	33	1.9	3.4e-04
	cell proliferation	43	1.5	6.6e-03
	regulation of anatomical structure morphogenesis	28	1.7	7.8e-03
	movement of cell or subcellular component	42	1.5	8.3e-03

Table 2.5: Example gene ontology (GO) terms or pathways that show significant correlation with lineage-specific RT states. State 9: human-chimpanzee specific early RT state; State 10: human-chimpanzee specific late RT state; State 11: human-chimpanzee-orangutan specific early RT state; State 14: orangutan specific late RT state; State 16: gibbon specific late RT state; State 18: green monkey specific late RT state. We use DAVID [146] to perform the gene ontology analysis.

Chapter 3

Phylo-HMRF for multi-species genome organization comparison

3.1 Introduction

3D genome organization, including 3D genome structure and spatial positioning of the chromosomes within the cell nucleus, is closely correlated with important genome functions and has critical roles in the modulation of genome functions [2]. As we introduced in Chapter 1, the evolution of 3D genome organization across species, especially across closely related mammalian species, has not been well explored, and the existing computational approaches for comparing genome organization across multiple species have limited capability. The comparison of genome organization across multiple species utilizing the high-throughput sequencing data such as Hi-C data presents new computational challenges.

As described in Chapter 2, we previously developed a method called phylogenetic hidden Markov Gaussian processes (Phylo-HMGP) [161] 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 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.

Comprehensive characterization of the detailed evolutionary patterns of 3D genome structure remains unclear. As described in Chapter 1, prior computational approaches for comparing genome organization across species did not explicitly consider the continuous nature of the chromatin interaction strengths, and are therefore limited in the ability of fully utilizing the Hi-C data to reveal detailed evolutionary patterns of genome organization. Here, we develop a new probabilistic model, phylogenetic hidden Markov random field (Phylo-HMRF), which integrates the continuous-trait evolutionary constraints with the hidden Markov random field (HMRF) model, to capture evolutionary patterns of continuous genomic features across species by utilizing generalized spatial constraints. We demonstrated the advantage of Phylo-HMRF using simulation data. In addition, we applied Phylo-HMRF to a new Hi-C dataset from the same cell type (lymphoblastoid 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, as well as 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.

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

The overview of the Phylo-HMRF method is shown in Figure 3.1. 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 3.1A, where each node in the map corresponds to multi-species contact frequencies between the corresponding pair of genomic loci. 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. An evolutionary pattern identified by Phylo-HMRF is associated with the conservation or variation of the feature of interest across different species. For example, 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 distributions across species are representative of the evolutionary patterns that we seek to identify as states. 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 3.1B, Phylo-HMRF uses the Ornstein-Uhlenbeck (OU) process [57–59] as the continuous-trait evolutionary model. Figure 3.1C is an illustration of the possible Hi-C evolutionary pat-

terns that Phylo-HMRF aims to uncover across four primate species in this work.

Furthermore, 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.

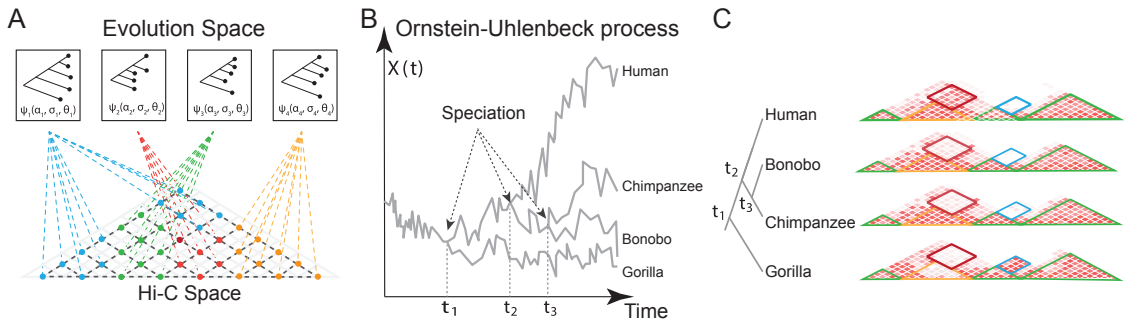

Figure 3.1: Overview of the Phylo-HMRF model. **(A)** Illustration of the possible evolutionary patterns of chromatin interaction. The Hi-C space is a combined multi-species Hi-C contact map, which integrates aligned Hi-C contact maps of each species. Each node represents the multi-species observations 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, 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 involved in the Hi-C data comparison 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.

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. We compare the multi-species Hi-C contact frequencies in the syntenic regions genome-

wide (synteny blocks were identified based on inferCARs [162]), in order to focus on the 3D genome changes that are not resulted from large-scale genome rearrangements. We then obtain a multi-species contact map $\mathbf{I} \in \mathbb{R}^{n \times n \times d}$, where n is the number of loci in the studied region on the reference genome, and d is the number of the species. $\mathbf{I}$ can be represented by a graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$, where $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\mathcal{E}}$ represent the set of nodes and the set of edges, respectively. Each node corresponds to a position in $\mathbf{I}$, i.e., the contact between a pair of genomic loci. The number of nodes is $N = n \times n$. We also denote $\boldsymbol{\mathcal{V}}$ as the set of indices of the nodes in $\boldsymbol{\mathcal{G}}$, i.e., $\boldsymbol{\mathcal{V}} = \{1, \cdots, N\}$. The i-th node is associated with a random variable $X_i \in \mathbb{R}^d$ representing the multi-species observations on this node, $i \in \boldsymbol{\mathcal{V}}$. The k-th element of X_i ($k = 1, \cdots, d$) is the aligned contact frequency measurement of the k-th species between the corresponding two genomic loci. If two positions in the multi-species contact map are adjacent, there is an edge between the corresponding nodes in $\boldsymbol{\mathcal{G}}$.

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

Phylo-HMRF estimates the evolutionary patterns by inferring the hidden states $\mathbf{y} = \{y_i\}_{i \in \boldsymbol{\mathcal{V}}}$ from the observations $\mathbf{x} = \{x_i\}_{i \in \boldsymbol{\mathcal{V}}}$, using the assumption that there are spatial dependencies between adjacent nodes in the graph $\boldsymbol{\mathcal{G}}$. In Phylo-HMRF, each hidden state $Y_i = l$ is associated with a phylogenetic model ψ_l. We therefore define the Phylo-HMRF model as $\mathbf{h} = (S, \psi, \beta)$, where S is the set of states, ψ is the set of phylogenetic models

associated with the states, and β contains the pairwise potential parameters, respectively. Suppose $X^{(l)} = (X_1^{(l)}, \cdots, X_d^{(l)})$ represent the values of leaf nodes of the phylogenetic tree associated with the l-th phylogenetic model ψ_l, $l = 1, \cdots, M$. The emission probability of each state is $p(X|\psi_l)$, which is determined by the phylogenetic model underlying this state. The joint probability of the graph $\boldsymbol{\mathcal{G}}$ is:

$$p(\mathbf{y}, \mathbf{x}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = \frac{1}{Z} \prod_{i \in \boldsymbol{\mathcal{V}}} p(x_i|y_i) \prod_{(i,j) \in \boldsymbol{\mathcal{E}}} f(y_i, y_j; x_i, x_j), \tag{3.1}$$

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

$$E(\mathbf{y}|\mathbf{x}, \Theta) = \sum_{i \in \boldsymbol{\mathcal{V}}} U(x_i, y_i; \Theta) + \sum_{(i,j) \in \boldsymbol{\mathcal{E}}} V(y_i, y_j; x_i, x_j), \tag{3.2}$$

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

$$\{\mathbf{y}^*, \Theta^*\} = \arg\min_{\mathbf{y}, \Theta} E(\mathbf{y}|\mathbf{x}, \Theta). \tag{3.3}$$

3.3 Ornstein-Uhlenbeck process assumption in Phylo-HMRF

In Phylo-HMRF, we model the continuous traits with the Ornstein-Uhlenbeck (OU) process. The OU process is a Gaussian process [163] that extends the Brownian motion [54–56] with the trend towards equilibrium around optimal values [57–59]. The OU process has been recently used to model the evolution of genomic features [50, 52, 53, 61, 161]. In our previous work [161], we found that the OU process has clear advantages in performance as compared to the simpler Brownian motion model. Therefore, we utilize the OU processes to realize the phylogenetic models in Phylo-HMRF.

For the observation of a lineage $\hat{X}_i$, the OU process can be represented as the following [57, 58]:

$$d(\hat{X}_i(t)) = \alpha[\theta_i - \hat{X}_i(t)]dt + \sigma dB_i(t), \tag{3.4}$$

where $\hat{X}_i(t)$ represents the observation of $\hat{X}_i$ at time point t, $B_i(t)$ represents the standard Brownian motion [120], 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_i} is the trait of the ancestor of species i, and $X_{a_{ij}}$ is the trait of the most recent common ancestor of species i and species j. As shown in Chapter 2, we have [53, 58]:

$$\mathbb{E}(X_i) = \mathbb{E}(X_{p_i})e^{-\alpha_i t_{ip_i}} + \theta_i \left(1 - e^{-\alpha_i t_{ip_i}}\right), \tag{3.5}$$

$$\mathrm{Cov}(X_i,\ X_j) = \mathrm{Var}(X_{a_{ij}}) \exp\left(-\Sigma_{k\in l_{ij}}\alpha_k t_k - \Sigma_{k\in l_{ji}}\alpha_k t_k\right), \tag{3.6}$$

$$\mathrm{Var}(X_i) = \frac{\sigma_i^2}{2\alpha_i}\left(1 - e^{-2\alpha_i t_{ip_i}}\right) + \mathrm{Var}(X_{p_i})e^{-2\alpha_i t_{ip_i}}, \tag{3.7}$$

where t_{ip_i}, 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_i} corresponds to length of the branch from the parent of species i to species i. For any $k \in l_{ij}$, 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 $\psi_l = (\theta_l, \alpha_l, \sigma_l, \tau_l, b_l)$, $l = 1, \cdots, M$, 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.

3.4 Phylo-HMRF model with Ornstein-Uhlenbeck processes

In Phylo-HMRF, we embed the OU model into the emission probability function of the HMRF model. We use the Expectation-Maximization (EM) algorithm [124, 164] for model parameter estimation.

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:

$$Q(\Theta, \Theta^g) = \mathbb{E}[\log p(\mathbf{x}, \mathbf{y}|\Theta)|\mathbf{x}, \Theta^g] = \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g) \log p(\mathbf{x}, \mathbf{y}|\Theta), \tag{3.8}$$

where $\mathbf{x}$ are the observations, $\mathbf{y}$ are the hidden states, and $\boldsymbol{\mathcal{S}}_N$ is the set of all the possible

state configurations of size N. N is the sample size. We have:

$$Q(\Theta, \Theta^g) = \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g)[\log p(\mathbf{x}|\mathbf{y}, \Theta) + \log p(\mathbf{y}|\Theta)], \tag{3.9}$$

where $p(\mathbf{y}|\Theta)$ represents the probability of a hidden state configuration over the whole graph $\boldsymbol{\mathcal{G}}$. Using pseudo-likelihood approximation [165], we can approximate $p(\mathbf{y}|\Theta)$ with:

$$p(\mathbf{y}|\Theta) = \prod_{i \in \boldsymbol{\mathcal{V}}} p(y_i|y_{\boldsymbol{\mathcal{N}}_i}, \Theta). \tag{3.10}$$

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

$$Q(\Theta, \Theta^g) = \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g) \left[\log p(\mathbf{x}|\mathbf{y}, \Theta) + \log p(\mathbf{y}|\Theta)\right] \tag{3.11}$$

$$= \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g) \left[\sum_{i \in \boldsymbol{\mathcal{V}}} \log p(x_i|y_i, \Theta) + \sum_{i \in \boldsymbol{\mathcal{V}}} \log p(y_i|y_{\boldsymbol{\mathcal{N}}_i}, \Theta)\right] \tag{3.12}$$

$$= \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} \sum_{i \in \boldsymbol{\mathcal{V}}} p(\mathbf{y}|\mathbf{x}, \Theta^g) \log p(x_i|y_i, \Theta) + \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} \sum_{i \in \boldsymbol{\mathcal{V}}} p(\mathbf{y}|\mathbf{x}, \Theta^g) \log p(y_i|y_{\boldsymbol{\mathcal{N}}_i}) \tag{3.13}$$

$$= \sum_{i \in \boldsymbol{\mathcal{V}}} \sum_{l=1}^{M} \sum_{y_{-i} \in \boldsymbol{\mathcal{S}}_{N-1}} p(y_i = l, y_{-i}|\mathbf{x}, \Theta^g) \log p(x_i|y_i = l, \Theta) + \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} \sum_{i \in \boldsymbol{\mathcal{V}}} p(\mathbf{y}|\mathbf{x}, \Theta^g) \log p(y_i|y_{\boldsymbol{\mathcal{N}}_i}) \tag{3.14}$$

$$= \sum_{l=1}^{M} \sum_{i \in \boldsymbol{\mathcal{V}}} p(y_i = l|\mathbf{x}, \Theta^g) \log p(x_i|y_i = l, \Theta) + \sum_{i \in \boldsymbol{\mathcal{V}}} \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g) \log p(y_i|y_{\boldsymbol{\mathcal{N}}_i}). \tag{3.15}$$

Let $x_{\boldsymbol{\mathcal{N}}_i}$ be the neighbors of x_i, and $y_{\boldsymbol{\mathcal{N}}_i}$ be the state configuration of $x_{\boldsymbol{\mathcal{N}}_i}$. To calculate $\sum_{i \in \boldsymbol{\mathcal{V}}} \sum_{\mathbf{y} \in \boldsymbol{\mathcal{S}}_N} p(\mathbf{y}|\mathbf{x}, \Theta^g) \log p(y_i|y_{\boldsymbol{\mathcal{N}}_i})$, we need to compute $p(\mathbf{y}|\mathbf{x}, \Theta^g)$ and $\log p(y_i|y_{\boldsymbol{\mathcal{N}}_i})$ over all the possible configurations $\mathbf{y} \in \boldsymbol{\mathcal{S}}_N$, which is computationally intractable. By mean-field approximation [166, 167], we can use estimated hidden states $y^g_{\boldsymbol{\mathcal{N}}_i}$ from the previous iteration of the HMRF-EM algorithm to approximately estimate $p(y_i|y_{\boldsymbol{\mathcal{N}}_i})$, which

can simplify the computation of the Q function. Using the mean-field approximation, we have

$$\sum_{i\in\boldsymbol{\mathcal{V}}}\sum_{\mathbf{y}\in\boldsymbol{\mathcal{S}}_N} p(\mathbf{y}|\mathbf{x},\Theta^g)\log p(y_i|y^g_{\boldsymbol{\mathcal{N}}_i}) = \sum_{i\in\boldsymbol{\mathcal{V}}}\sum_{l=1}^{M}\sum_{y_{-i}\in\boldsymbol{\mathcal{S}}_{N-1}} p(y_i=l,y_{-i}|\mathbf{x},\Theta^g)\log p(y_i=l|y^g_{\boldsymbol{\mathcal{N}}_i}) \tag{3.16}$$

$$= \sum_{l=1}^{M}\sum_{i\in\boldsymbol{\mathcal{V}}} p(y_i=l|x_i,\Theta^g)\log p(y_i=l|y^g_{\boldsymbol{\mathcal{N}}_i},\Theta). \tag{3.17}$$

Therefore, from Equations 3.15 and 3.17, we have the Q function in Equation 3.18.

$$\begin{aligned} Q(\Theta,\Theta^g) =& \sum_{l=1}^{M}\sum_{i\in\boldsymbol{\mathcal{V}}} p(y_i=l|x_i,\Theta^g)\log p(x_i|y_i=l,\Theta) \\ &+ \sum_{l=1}^{M}\sum_{i\in\boldsymbol{\mathcal{V}}} p(y_i=l|x_i,\Theta^g)\log p(y_i=l|y^g_{\boldsymbol{\mathcal{N}}_i},\Theta), \end{aligned} \tag{3.18}$$

where the two parts of the Q function encode the unary potential and the pairwise potential of the HMRF of $\boldsymbol{\mathcal{G}}$, respectively. The parameters of the OU model are embedded in the terms $p(x_i|y_i=l,\Theta)$ of $Q(\Theta,\Theta^g)$, $l\in\{1,\cdots,M\}$. The second part of the Q function encodes the pairwise potentials and does not include the OU model parameters.

Here $p(y_i=l|x_i,\Theta^g)$ is posterior probability of each sample assigned to a hidden state given the current model parameter estimates. Using the Markov property of HMRF [164], we have

$$p(y_i=l|x_i,\Theta^g) = \frac{p(x_i|y_i=l,\Theta^g)p(y_i=l|y^g_{\boldsymbol{\mathcal{N}}_i})}{\sum_{l=1}^{M} p(x_i|y_i=l,\Theta^g)p(y_i=l|y^g_{\boldsymbol{\mathcal{N}}_i})}, \tag{3.19}$$

where $\boldsymbol{\mathcal{N}}_i$ denotes the set of nodes that are neighbors of node i in $\boldsymbol{\mathcal{G}}$. We calculate $p(x_i|y_i,\Theta^g)$ based on the OU process assumption. Specifically, we assume the observations of observed species (leaf nodes in the phylogenetic tree) follow multivariate Gaussian distribution parameterized by the OU model parameters. We have:

$$p(x_i|y_i = l, \Theta) = p(x_i|\Theta^{(l)})$$
$$= \frac{1}{(2\pi)^{d/2}|\Sigma_\Theta^{(l)}|^{1/2}} \exp\left\{-\frac{1}{2}\left(x_i - \mu_\Theta^{(l)}\right)^T [\Sigma_\Theta^{(l)}]^{-1}\left(x_i - \mu_\Theta^{(l)}\right)\right\}, \quad (3.20)$$
$$\log p(x|\Theta^{(l)}) \propto -\frac{1}{2}\log|\Sigma_\Theta^{(l)}| - \frac{1}{2}\left(x_i - \mu_\Theta^{(l)}\right)^T [\Sigma_\Theta^{(l)}]^{-1}\left(x_i - \mu_\Theta^{(l)}\right), \quad (3.21)$$

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 $\Sigma_\Theta^{(l)}$ and the mean vector $\mu_\Theta^{(l)}$ according to Equations 3.5, 3.6, and 3.7. We calculate $p(x_i|y_i = l, \Theta^g)$ in the same way as shown in Equation 3.20 by replacing Θ with Θ^g, where Θ^g denotes the current model parameter estimates.

Let

$$w_i^{(l)} = p(y_i = l|x_i, \Theta^g). \quad (3.22)$$

$p(y_i = l|x_i, \Theta^g)$ is calculated using Equation 3.19. We have

$$Q(\Theta, \Theta^g) = \sum_{l=1}^{M}\sum_{i\in\boldsymbol{\mathcal{V}}} w_i^{(l)} \log p(x_i|y_i = l, \Theta) + \sum_{l=1}^{M}\sum_{i\in\boldsymbol{\mathcal{V}}} w_i^{(l)} \log p\left(y_i = l|y^g_{\boldsymbol{\mathcal{N}}_i}\right) \quad (3.23)$$
$$= \sum_{l=1}^{M}\sum_{i\in\boldsymbol{\mathcal{V}}} w_i^{(l)} \left[-\frac{1}{2}\log|\Sigma_\Theta^{(l)}| - \frac{1}{2}\left(x_i - \mu_\Theta^{(l)}\right)^T [\Sigma_\Theta^{(l)}]^{-1}\left(x_i - \mu_\Theta^{(l)}\right)\right]$$
$$+ \sum_{l=1}^{M}\sum_{i\in\boldsymbol{\mathcal{V}}} w_i^{(l)} \log p\left(y_i = l|y^g_{\boldsymbol{\mathcal{N}}_i}\right) + C, \quad (3.24)$$

where C is a constant.

We perform parameter estimation for each of the possible states.

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

$$L(\Theta^{(l)}) = \frac{1}{2}\log|\Sigma_\Theta^{(l)}| \sum_{i\in\boldsymbol{\mathcal{V}}} w_i^{(l)} + \frac{1}{2}\sum_{i\in\boldsymbol{\mathcal{V}}}\left(x_i - \mu_\Theta^{(l)}\right)^T [\Sigma_\Theta^{(l)}]^{-1}\left(x_i - \mu_\Theta^{(l)}\right) w_i^{(l)}. \quad (3.25)$$

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

$$\tilde{L}(\Theta^{(l)}) = \frac{1}{N}\log|\Sigma_\Theta^{(l)}| \sum_{i\in\boldsymbol{\mathcal{V}}} w_i^{(l)} + \mathrm{tr}\left([\Sigma_\Theta^{(l)}]^{-1}\tilde{S}_\Theta^{(l)}\right), \quad (3.26)$$

where

$$\tilde{S}_{\Theta}^{(l)} = \frac{1}{N} \sum_{i \in \boldsymbol{\mathcal{V}}} w_i^{(l)} \left(x_i - \mu_{\Theta}^{(l)} \right) \left(x_i - \mu_{\Theta}^{(l)} \right)^T, \tag{3.27}$$

and $\Theta^{(l)}$ represents the phylogenetic model parameters associated with state l. $\text{tr}(A)$ 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.

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 $\lambda \|\Theta^{(l)}\|_2^2$.

Let $L(\Theta^{(l)}) = -\sum_{i \in \boldsymbol{\mathcal{V}}} w_i^{(l)} \log p(x_i | y_i = l, \Theta)$, $w_i^{(l)} = p(y_i = l | x_i, \Theta^g)$. In each Maximization-step (M-step) of the EM algorithm, the objective function of a given state l is derived as

$$\min_{\Theta^{(l)}} \frac{1}{N} \log |\Sigma_{\Theta}^{(l)}| \sum_{i \in \boldsymbol{\mathcal{V}}} w_i^{(l)} + \text{tr} \left([\Sigma_{\Theta}^{(l)}]^{-1} \tilde{S}_{\Theta}^{(l)} \right) + \lambda \|\Theta^{(l)}\|_2^2, \tag{3.28}$$

where $\tilde{S}_{\Theta}^{(l)}$, $w_i^{(l)}$, $\Theta^{(l)}$ are defined as above, and λ is the regularization coefficient of the l_2-norm regularization [168].

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 $\psi_l = (\theta_l, \alpha_l, \sigma_l, \tau_l, b_l)$, $1 = 1, \cdots, M$. We assume that τ_l is given. If the branch lengths b_l are unknown, we perform the transformation that $\tilde{\alpha}_{l,v} = \alpha_{l,v} b_{l,v}$, $\tilde{\sigma}_{l,v}^2 = \sigma_{l,v}^2 b_{l,v}$ 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.

For 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 [161] 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 as shown in Equation 3.18, let $V(y_i, y_j)$ be the pairwise potential on a pair of adjacent nodes (y_i, y_j), we have

$$p(y_i = l|y^g_{\boldsymbol{\mathcal{N}}_i}) = \frac{1}{Z} \exp\left(-\sum_{j \in \boldsymbol{\mathcal{N}}_i} V(l, y^g_j)\right), \tag{3.29}$$

where Z is the normalization constant. We can adopt different definitions of the pairwise potential $V(y_i, y_j)$. We consider two definitions. The first definition is

$$V(y_i, y_j) = \beta_0 I(y_i \neq y_j), \tag{3.30}$$

where β_0 is a predefined adjustable regularization coefficient, which can also be considered as pairwise potential parameter.

For the second definition, we use takes into consideration the difference between features of the adjacent nodes in imposing the penalty on inconsistent states of the neighbors. The second definition is

$$V(y_i, y_j) = \beta_0 I(y_i \neq y_j) \exp\left(-\beta_1 \frac{\|x_i - x_j\|_2^2}{\|x_i\|_2 \|x_j\|_2}\right), \tag{3.31}$$

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

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 3.31 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.

3.5 Model inference

3.5.1 Model parameter estimation and hidden state inference

We use the Expectation-Maximization (EM) algorithm [124, 164] for parameter estimation in our model. Zhang et al. [164] 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 [165] and mean-field approximation [166, 167]. 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 [169] for hidden state estimation given estimates of model parameters.

The EM algorithm aims to maximize the expectation of the complete-data log likelihood, which is defined as the Q function in Equation 3.8: $Q(\Theta, \Theta^g) = \mathbb{E}[\log p(\mathbf{x}, \mathbf{y}|\Theta)|\mathbf{x}, \Theta^g]$, where $\mathbf{x}$, $\mathbf{y}$ represent the observations and the hidden states, respectively. Θ, Θ^g represent the model parameters, and the current estimate of model parameters, respectively. Using pseudo-likelihood approximation [165], mean-field approximation [166, 167], and l_2-norm regularization, the objective function for optimizing the OU model parameters of a given state l in each Maximization-step (M-step) is shown in Formula 3.28. We perform parameter estimation for each of the possible states.

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 section 3.5.3). The estimated model parameters are used as initialization of the OU model parameters for each hidden state.

2. *Estimate the hidden states given current parameter estimates*. We seek an approximate solution to the optimization problem:

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}\in\boldsymbol{\mathcal{S}}_N} \{U(\mathbf{x}|\mathbf{y}) + U(\mathbf{y})\}, \tag{3.32}$$

where $U(\mathbf{x}|\mathbf{y})$ and $U(\mathbf{y})$ are the total unary potential and total pairwise potential of $\boldsymbol{\mathcal{G}}$, respectively.

3. *Calculate posterior probability distribution*. In each Expectation-step (E-step), given the current estimated model parameters Θ^g and the estimated hidden state configuration in the previous step, we compute $p(y_i = l|x_i, \Theta^g)$ using Equations 3.19,3.29.

4. *Estimate model parameters by solving the optimization problem with OU models embedded*. In each M-step, we solve the optimization problem in Formula 3.28 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. Graph Cuts algorithms seek approximate solutions to an energy minimization problem by solving a max-flow/min-cut problem in a graph [169, 170]. 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 $p(\mathbf{y}|\mathbf{x})$ [169]. 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 [169, 171]. 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 [169].

3.5.2 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 [169, 170] 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 3.32. We have also shown that minimizing the energy is equivalent to maximizing the joint probability.

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

$$\mathbf{y}^* = \arg\min_{\mathbf{y}} E(\mathbf{y}|\mathbf{x}, \Theta^g), \tag{3.33}$$

and

$$\Theta^* = \arg\max_{\Theta} \mathbb{E}\left[\log p(\mathbf{x}, \mathbf{y}|\Theta)|\mathbf{x}, \Theta^g\right], \tag{3.34}$$

where Θ^g is the current estimates of the model parameters.

We use the Graph Cuts algorithm for the first stage (Equation 3.33) and use the EM algorithm for the second stage (Equation 3.34). We use the solution $\mathbf{y}^*$ from the first stage to compute $\mathbb{E}[\log p(\mathbf{x}, \mathbf{y}|\Theta)|\mathbf{x}, \Theta^g]$ in the second stage with the mean-field approximation [166, 167]. The energy minimization problem for MRF (Equation 3.33) is known to be NP-hard [171]. 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 $\boldsymbol{\mathcal{G}}$ of the HMRF to utilize the Graph Cuts algorithm. The unary cost corresponds to the unary potential, which is:

$$U(x_i|y_i, \Theta^g) \propto -\log(p(x_i|y_i, \Theta^g). \tag{3.35}$$

The pairwise cost corresponds to the pairwise potential. We calculate the edge weights in $\boldsymbol{\mathcal{G}}$ as

$$\bar{w}_{ij} = \exp\left(-\beta_1 \frac{\|x_i - x_j\|_2^2}{\|x_i\|_2 \|x_j\|_2}\right), \tag{3.36}$$

where β_1 is an coefficient, and we use a pairwise state transition cost matrix $\bar{V} \in \mathbb{R}^{M\times M}$, where M is the number of hidden states and $\bar{V}_{ij}$ represents the penalty on $y_j \neq y_i$ for

a directed edge ($i \leftarrow j$). In our problem, $\boldsymbol{\mathcal{G}}$ is an undirected graph, and we simplify $\bar{V}$ as $\bar{V}_{ij} = \beta_0$, $i, j = 1, \cdots, M$. However, in more complicated problem settings, we can realize $\bar{V}$ with varied elements $\bar{V}_{ij}$ 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 3.31). We use the GCO library to perform the Graph Cuts algorithm [169, 170, 172].

3.5.3 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. [161] 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 3.28. The difference is that we set $w_i^{(l)} = 1$, and change $i \in \boldsymbol{\mathcal{V}}$ to the constraint $i \in \boldsymbol{\mathcal{C}}_l$, where $\boldsymbol{\mathcal{C}}_l$ 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.

3.6 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 [144] 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 3.2). 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.

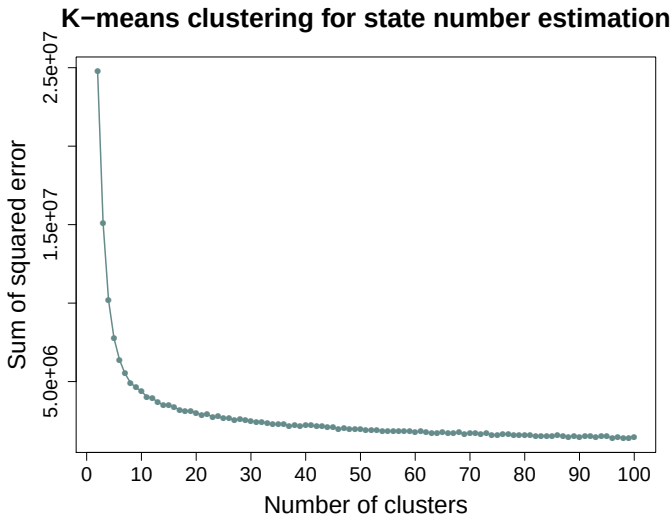

Figure 3.2: The change of Sum of Squared Errors (SSE) with respect to an increased number of clusters in K-means clustering on the cross-species Hi-C data. Using the Elbow method, we estimate the Elbow point (the point where the contribution of a larger K to a smaller SSE decreases evidently and tends to be small) to be in the range 15-30, and we choose the number of state to be 30 accordingly.

3.7 Data preparation and processing

3.7.1 Experimental model and subject details

The four primate species included in the Hi-C data comparison are Homo Sapiens (Human), Pan troglodytes (Common Chimpanzee), Pan Paniscus (Bonobo), and Troglodytes 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 [173]. 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. We only used autosomes for analysis for each of the species. The key resources used are shown in Table 3.1.

3.7.2 Cross-species Hi-C data processing

We used the Hi-C data from the lymphoblastoid cells in human (GM12878) from Rao et al. [4]) 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 [174] 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 [138]. 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 the Hi-C data comparison across species).

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 [4, 174]. 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 [175], 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 align the Hi-C contact maps of the four species in each synteny block to obtain a combined multi-species Hi-C contact map. Each node in the multi-species Hi-C contact map is associated with a feature vector, the entries of which correspond 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 $\hat{x} = \log(1 + x)$ 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. 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 [162]. 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.

3.8 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 [176] 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.

3.9 Approaches for the simulation studies

3.9.1 Approaches to generating the simulated datasets

We first evaluated the performance of the developed method in simulation studies. In the simulation evaluation, we suppose that the samples in simulated dataset correspond to nodes in a graph $\boldsymbol{\mathcal{G}}$. The graph $\boldsymbol{\mathcal{G}}$ 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 $\boldsymbol{\mathcal{N}}_i$ denote the set of neighboring nodes of the node i, i.e., the nodes that are connected to node i in $\boldsymbol{\mathcal{G}}$. 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 [177]. 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:

$$p(y_i|y_{-i}) = p(y_i|y_{\boldsymbol{\mathcal{N}}_i}), \tag{3.37}$$

where y_{-i} represents the hidden states of all the nodes other than the i-th node, i.e., $y_{-i} = \{y_j, j \in \boldsymbol{\mathcal{V}}, j \neq i\}$. $y_{\boldsymbol{\mathcal{N}}_i}$ 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 $(t+1)$ is sampled from the probabilistic distribution $p(y_i|y_{\boldsymbol{\mathcal{N}}_i}^{(t-1)})$, $y_i \in S = \{1, \cdots, M\}$. $p(y_i = l|y_{\boldsymbol{\mathcal{N}}_i}^{(t-1)})$ is calculated using Equation 3.29, $l \in S$. We use the first definition of pairwise potential and use $\beta_0 = 2$ (Equation 3.30). We repeat this sampling process until the maximum of time steps to take T is reached. We use $\mathbf{y}^{(T)} = \{y_i^{(T)}\}_{i \in \boldsymbol{\mathcal{V}}}$ 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 $x_i \in \mathbb{R}^d$, $i \in \boldsymbol{\mathcal{V}}$. 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 $\boldsymbol{\mathcal{N}}(0.5, 0.25)$. We then use OU model parameters to calculate the Gaussian distribution parameters θ_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 $\boldsymbol{\mathcal{N}}(1, 0.5)$. We also use $n = 500$, $N = 250000$, $d = 4$, $M = 10$.

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 3.31), 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.

3.9.2 Performance evaluation in the simulation studies

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 [142, 143]. These metrics compare two partitions of a set. The metrics were also defined and used in Chapter 2. 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:

$$NMI(\Omega; C) = \frac{I(\Omega; C)}{[H(\Omega) + H(C)]/2}, \tag{3.38}$$

where $H(\Omega)$ and $H(C)$ are the entropies of Ω and C, respectively.

We have $H(\Omega) = -\sum_{k=1}^{K} P(\omega_k) \log P(\omega_k)$, $H(C) = -\sum_{j=1}^{M} P(c_j) \log P(c_j)$. $P(\omega_k)$ 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 is defined as:

$$AMI(\Omega; C) = \frac{I(\Omega; C) - \mathbb{E}[I(\Omega; C)]}{\max\{H(\Omega), H(C)\} - \mathbb{E}[I(\Omega; C)]}, \tag{3.39}$$

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 [143].

The Rand Index (RI) [142] is defined as:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}, \tag{3.40}$$

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 is defined as:

$$ARI = \frac{RI - \mathbb{E}[RI]}{\max\{RI\} - \mathbb{E}[RI]}, \tag{3.41}$$

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

Precision, Recall, and F_1 score are defined as:

$$Precision = \frac{TP}{TP + FP}, \tag{3.42}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3.43}$$

$$F_1 = \frac{2Precision \times Recall}{Precision + Recall}. \tag{3.44}$$

3.9.3 Other methods compared in the simulation evaluation

In the simulation studies, for performance evaluation on state estimation, we compared Phylo-HMRF with the Gaussian-HMRF method [164], the Gaussian Mixture Model (GMM), the *K*-means clustering method, and two image segmentation methods SLIC [178] and Quick Shift [179]. 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 the color information and the spatial coordinates over an image. Quick Shift is approximation of the Mean Shift algorithm [180] with kernel methods utilized, performing mode seeking in segmenting an image. We use the scikit-learn library [144] for implementation of the GMM method and the K-means clustering algorithm. We use the scikit-image library [181] which includes the 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 in the scikit-image library, 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.

3.10 Quantification and analysis methods

3.10.1 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 50Kb, 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 50Kb (one bin) accordingly. We compute the percentages of the distance in five distance intervals, which are 0-50Kb, 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.

3.10.2 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 [161] 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 3.11). 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.

3.10.3 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 3.8F. 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).

3.10.4 Estimating the 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 genome 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. 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.

3.11 Results

3.11.1 Performance evaluation of Phylo-HMRF in the simulation studies

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 (simulation studies I and II), each of which contains 8 datasets. Suppose the samples in simulated datasets correspond to nodes in a graph $\boldsymbol{\mathcal{G}}$. The samples thus represent features of node in $\boldsymbol{\mathcal{G}}$. Similar to a Hi-C contact map, $\boldsymbol{\mathcal{G}}$ has a 2D lattice structure of size $n \times n$, where each node is associated with a sample. The samples thus represent features of nodes in $\boldsymbol{\mathcal{G}}$. For example, a sample can represent the interaction intensities between the i-th locus and the j-th locus out of n genomic loci in multiple species ($1 \leq i, j \leq n$). We also assume that each sample has a class label or

hidden state. Each hidden state is associated with an emission probability function and determines the observed feature of the sample with this state. The hidden states of the samples are assumed to be from a Markov random field (MRF). Thus, the hidden state of a sample is spatially dependent on the hidden states of its neighbors in $\mathcal{G}$. We simulated the multi-species observations from multivariate Gaussian distributions with OU model parameters embedded. The details of the data simulation approaches are described in section 3.9.1.

We compared Phylo-HMRF with several other methods on the simulated datasets, including the Gaussian-HMRF method [164], the Gaussian mixture model (GMM), the *K*-means clustering, and two image segmentation methods SLIC (Simple Linear Iterative Clustering) [178] and Quick Shift [179] (section 3.9.3), 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 $M = 10$. 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 [142, 143].

The evaluation results are shown in Figure 3.3 and Table 3.2. 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 sim-

ilarity 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. 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 performance using the same procedure as we used in simulation study I. The evaluation results are shown in Figure 3.4. 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.

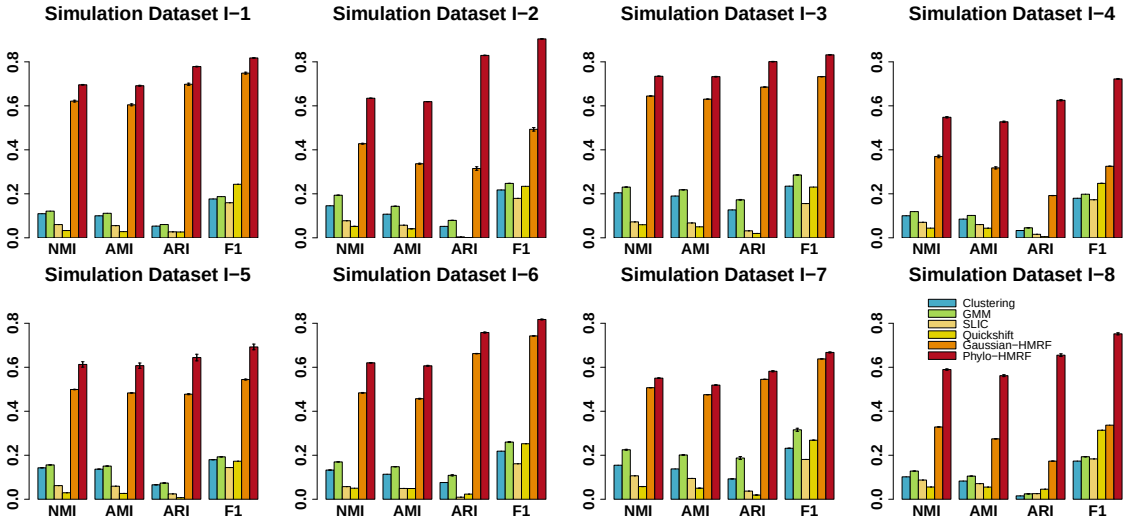

Figure 3.3: Performance evaluation of different methods in simulation study I. Performance evaluation of K-means clustering, GMM, SLIC, Quick Shift, Gaussian-HMRF, and Phylo-HMRF on eight simulation datasets in simulation study I with respect to NMI (Normalized Mutual Information), AMI (Adjusted Mutual Information), ARI (Adjusted Rand Index), and F_1 score. The standard error of the results from 10 runs of each method is shown as the error bar, respectively.

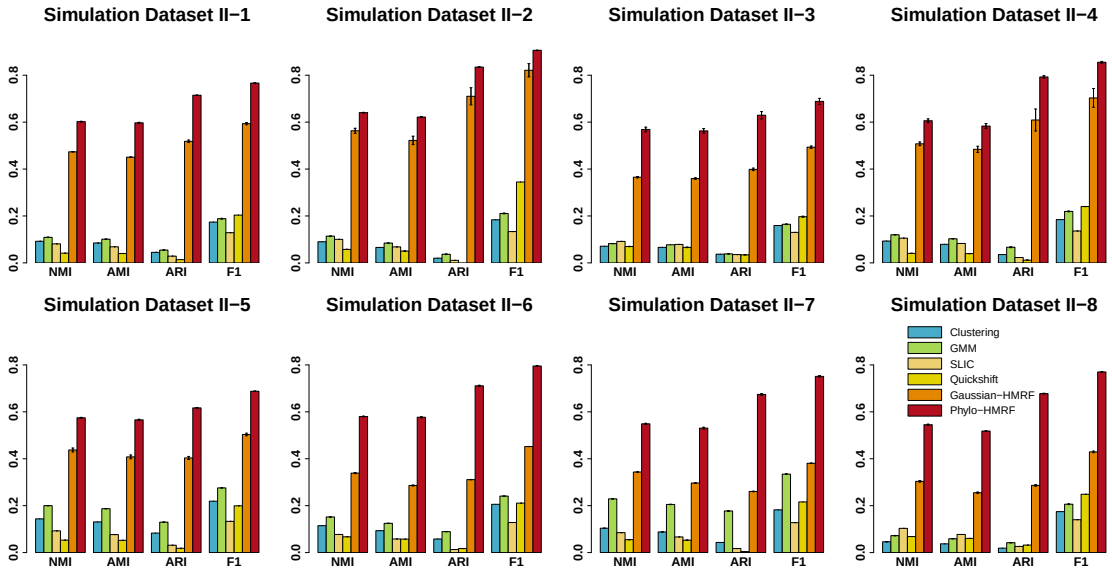

Figure 3.4: Performance evaluation of different methods in simulation study II. Performance evaluation of K-means Clustering, GMM, SLIC, Quickshift, Gaussian-HMRF, and Phylo-HMRF on eight simulation datasets in simulation study II with respect to NMI, AMI, ARI, and F_1 score. The standard error of the results from 10 runs of each method is shown as the error bar, respectively.

3.11.2 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. [4]. We generated Hi-C data from the lymphoblastoid cells of three other primate species, including chimpanzee, bonobo, and gorilla (see section 3.7.2). There are 290M, 270M, 240M, and 290M 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 50kb resolution (i.e., ignoring rearrangements smaller than 50kb among the four species) using the method inferCARs [162], covering 92.64% of the sequenced regions in the human genome (Figure 3.5, section 3.7.2). For example, we identified 9 major synteny blocks on human chromosome 1 among the four species, covering 92.50% of human chromosome 1 (Figure 3.6B). We then applied Phylo-HMRF to elucidating 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 [182, 183] (section 3.6, Figure 3.2). 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 which show higher-level distinctiveness of heterogeneous evolutionary patterns (section 3.8). 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 evolutionary patterns in Hi-C contacts, where eight exhibit lineage-specific patterns. Specifically, the nine groups are human-specific high in Hi-C contact frequency (NC-hom_high), human-specific low (NC-hom_low), chimpanzee-specific high (NC-pan_high), chimpanzee-specific low (NC-pan_low), bonobo-specific high (NC-pyg_high), bonobo-specific low (NC-pyg_low), gorilla-specific high (NC-gor_high), gorilla-specific low (NC-gor_low), and non-conserved (NC). Representative estimated states from each of the 13 groups are shown in Figure 3.6A. Hi-C contact frequency distributions of each species in all the 30 estimated states are shown in Figure 3.7.

In Figure 3.6B-C, we show the estimated states in synteny block 8 on chromosome 1 and in synteny block 3 on chromosome 2 as examples, along with the input Hi-C contact maps of the four species. The rotated upper triangular matrix as illustrated in the second panel of Figure 3.6B-C 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, based on the estimated states, $\mathcal{G}$ is partitioned, reflecting different cross-species Hi-C contact patterns. In Figure 3.6B, 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 3.9. In Figure 3.6C, 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 $<$3Mb), 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 3.8A. We observed that C-high, C-mid, and WC states are the predominant states for the short-range contacts (black arrow in Figure 3.8A). 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 non-conserved 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 Figure 3.9 and Figure 3.10 as examples, respectively. 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 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.

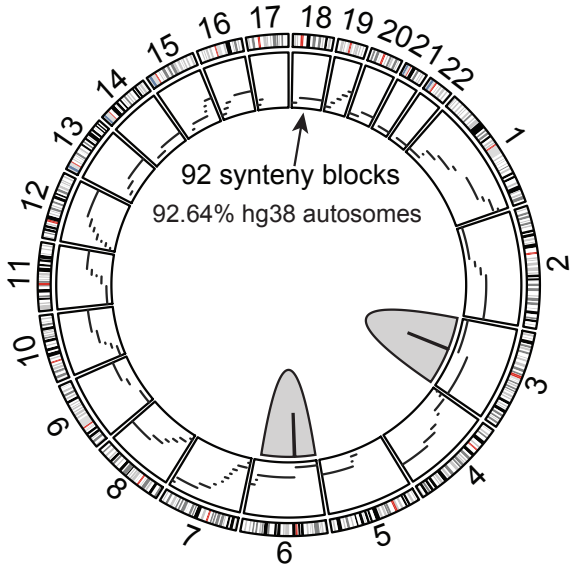

Figure 3.5: Distribution of the identified synteny blocks on the 22 autosomes of the human genome. For chromosome 3 and chromosome 6, we divide the large size synteny blocks into two parts each according to the chromosome arms. The synteny blocks cover 92.64% of human chromosomes 1-22.

3.11.3 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 replication timing (RT) is closely correlated with genome organization [37, 42]. We previously reported evolutionary patterns of DNA replication timing using Repli-seq data of multiple primate species [161]. We identified 5 groups of RT evolutionary states, which are conserved early in RT (E),

weakly conserved early (WE), conserved late (L), weakly conserved late (WL), and non-conserved (NC). For each pair of genomic loci with estimated Hi-C states, we examined the RT 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 non-conserved in RT. We then computed the percentage of contact loci that have the conserved RT states in the C-high, C-mid, WC, C-low, and NC Hi-C states identified by Phylo-HMRF over a range of different distances (0-10Mb). The results are shown in Figure 3.8B. It is observed that C-high, C-mid, and WC Hi-C 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 4Mb 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 3.8B with known chromatin structure patterns such as TADs. We considered the TADs in GM12878 in human called using the Arrowhead method [4] and the Directionality Index (DI) method [8] (named Arrowhead TADs and DI TADs, respectively). The average sizes of Arrowhead TADs and DI TADs are around 1Mb and 0.6Mb, respectively. As shown in Figure 3.8B, there is a changing point around 1Mb 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 plot 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 (section 3.10.2). Specifically, the RT evolutionary states were estimated at the resolution of 6Kb [161]. 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 3.117. 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 3.8B. 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 [151]. 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 non-conserved 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 (0-10Mb). 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 (section 3.10.4). 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 non-conserved states as shown in Figure 3.12 and Figure 3.8C. Additionally, the percentage of similar signals between paired loci is particularly high in the C-high state for genomic loci that are less than 4Mb 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.

3.11.4 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 [1]. 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: (i) a pair of 50kb genomic bins that are in the same TAD; (ii) a pair of 50kb 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. We observe that pairwise bins both from B compartment (i.e., B-B interactions) have the highest fraction of conserved states (Figure 3.8D, Figure 3.13). 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 [184]. 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 [8]. 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 (section 3.10.1). 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 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 3.8E, section 3.10.1).

We observed that the distance between the boundaries of the identified diagonal blocks and the nearest TADs are significantly more enriched in the distance intervals that represent relatively small distance (e.g., [0,50Kb] and (50Kb,100Kb]) than expected (Figure 3.8E). 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$<$2e-03, section 3.10.1). Similarly, the percentages of the identified diagonal block boundaries that are matched by a DI TAD boundary within 1 bin or 2 bins are significantly higher than the expected percentages (empirical p-value$<$1e-03). In contrast, the percentage of the diagonal block boundaries with distance to a nearest TAD boundary larger than 4 bins are significantly smaller than expected (empirical p-value$<$1e-03).

For example, we identified 34 blocks along the diagonal of the Hi-C map of synteny block 8 on 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 3.8F). 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 off-diagonal area of the Hi-C contact map, which potentially correspond to the long-range interactions between two TADs that are conserved across species. Examples of the potentially conserved long-range TAD interactions are shown in Figure 3.8F, which are highlighted with 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 TE families in each estimated Hi-C contact state across species (e.g., PIF-Harbinger, hAT-Tag1) (section 3.11.5, Figure 3.1510). We then

specifically assessed the potential correlations between the evolutionary patterns on TADs and sequence properties, in particular, transposable elements (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 (section 3.10.3). We detected TE families and TF motifs that show distinct enrichment patterns in the different groups of TADs (section 3.10.3, Figure 3.15). 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.

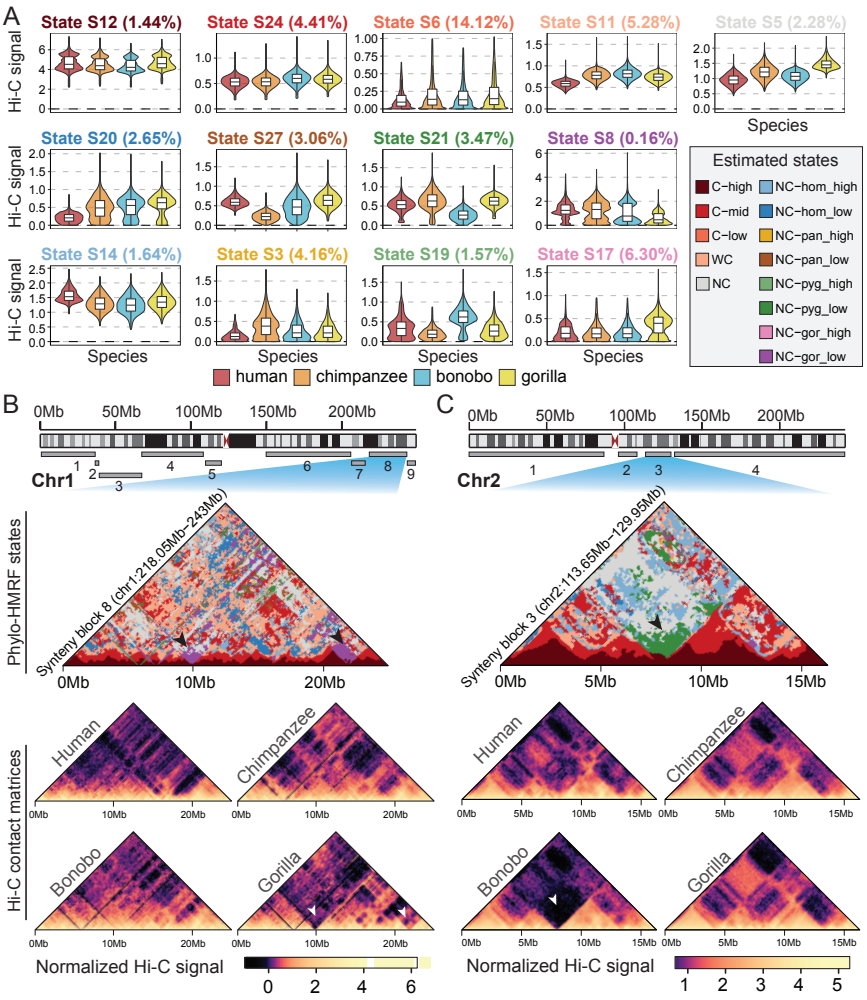

Figure 3.6: Evolutionary patterns of Hi-C contact frequency estimated by Phylo-HMRF. **(A)** Representative states from the 13 groups of evolutionary patterns. One state from each group is presented. The box plots show the normalized cross-species Hi-C contact frequency distributions of the four species 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 scale displayed at the bottom. Darker color in the Hi-C contact map represents lower contact frequency. The white arrows point to the corresponding locations of the two examples of identified gorilla-specific low contact state. **(C)** Cross-species Hi-C contact frequency states identified in synteny block 3 on chromosome 2. 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.

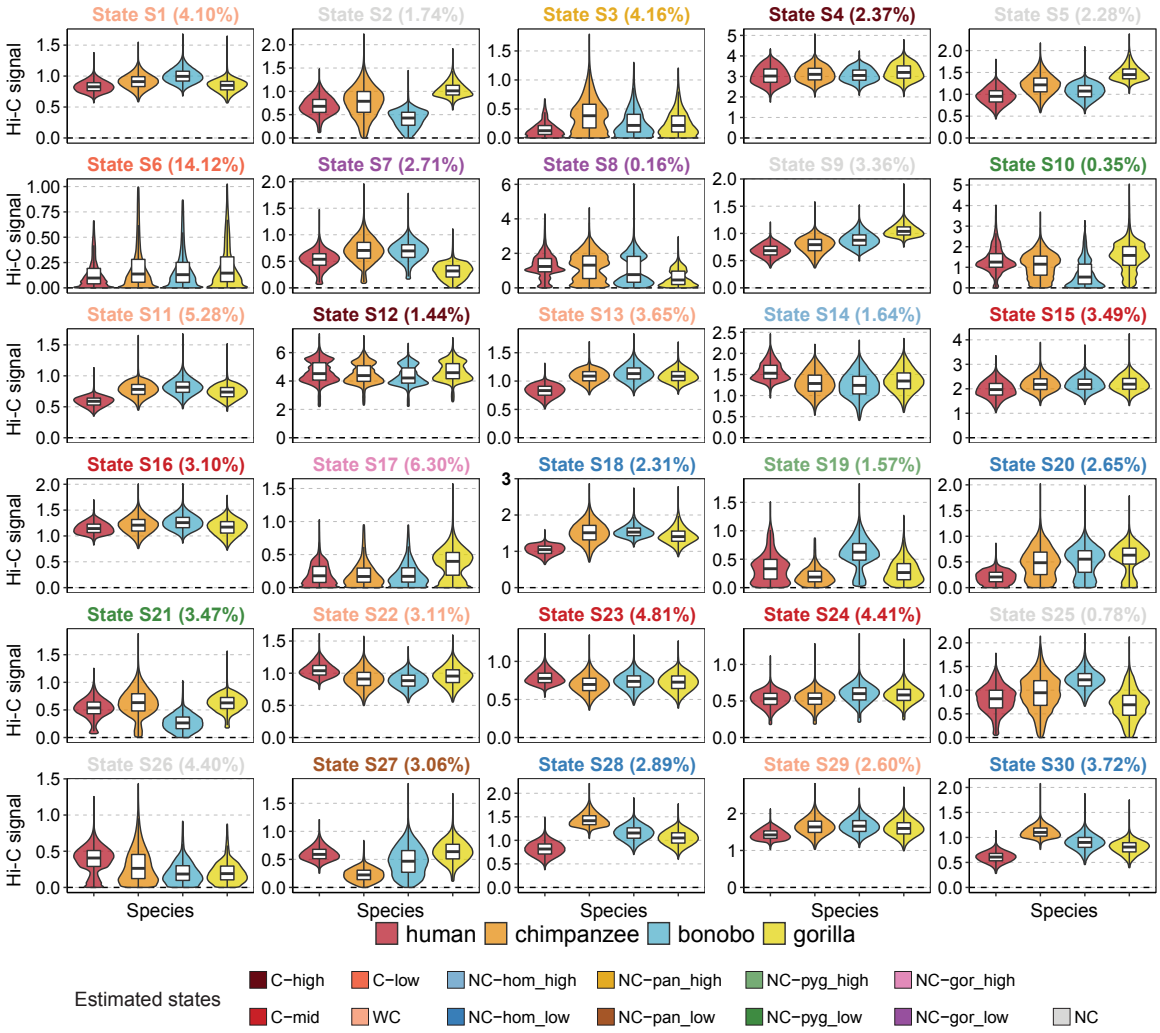

Figure 3.7: Hi-C evolutionary patterns identified by Phylo-HMRF in all the major synteny blocks on all autosomes across four primate species. Each boxplot shows the normalized cross-species Hi-C contact frequency distributions in the corresponding state. The color of the boxplot title shows the evolutionary group assignment of the corresponding state.

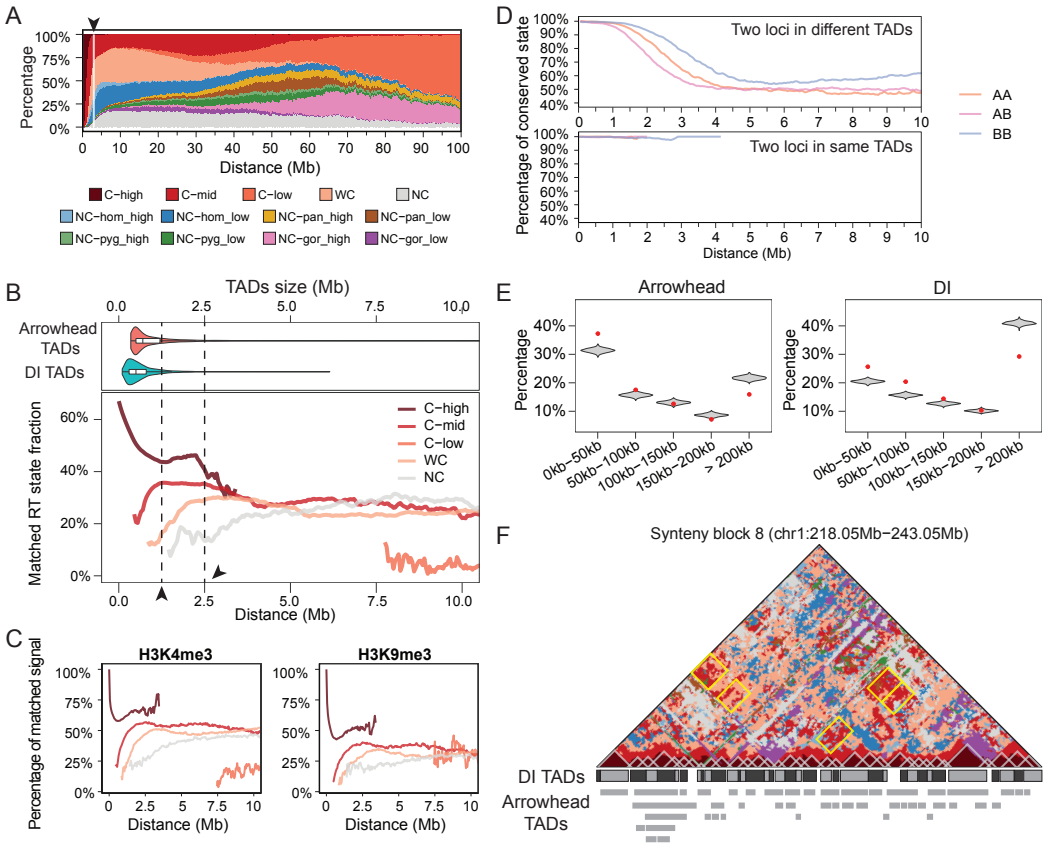

Figure 3.8: Comparison between the evolutionary states of Hi-C contacts estimated by Phylo-HMRF and other features of genome structure and function. **(A)** Global state-distance plots show the enrichment of different evolutionary patterns at different distances between genomic loci across synteny blocks on all autosomes. The black arrow points to the distance range around 2.5Mb where C-high and C-mid are the predominant states. **(B)** The percentage of 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 for the C-mid state. The second black arrow points to the position of 2.5Mb, where the trend change is observed in the state-distance plot as shown in **(A)**. **(C)** The percentage of pairwise genomic loci that have 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 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 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 distance between boundaries of local contact patterns and the nearest TADs, with the identified local contact patterns randomly shuffled. **(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. Examples of long-range conserved TAD interaction patterns are shown with yellow solid lines as borders.

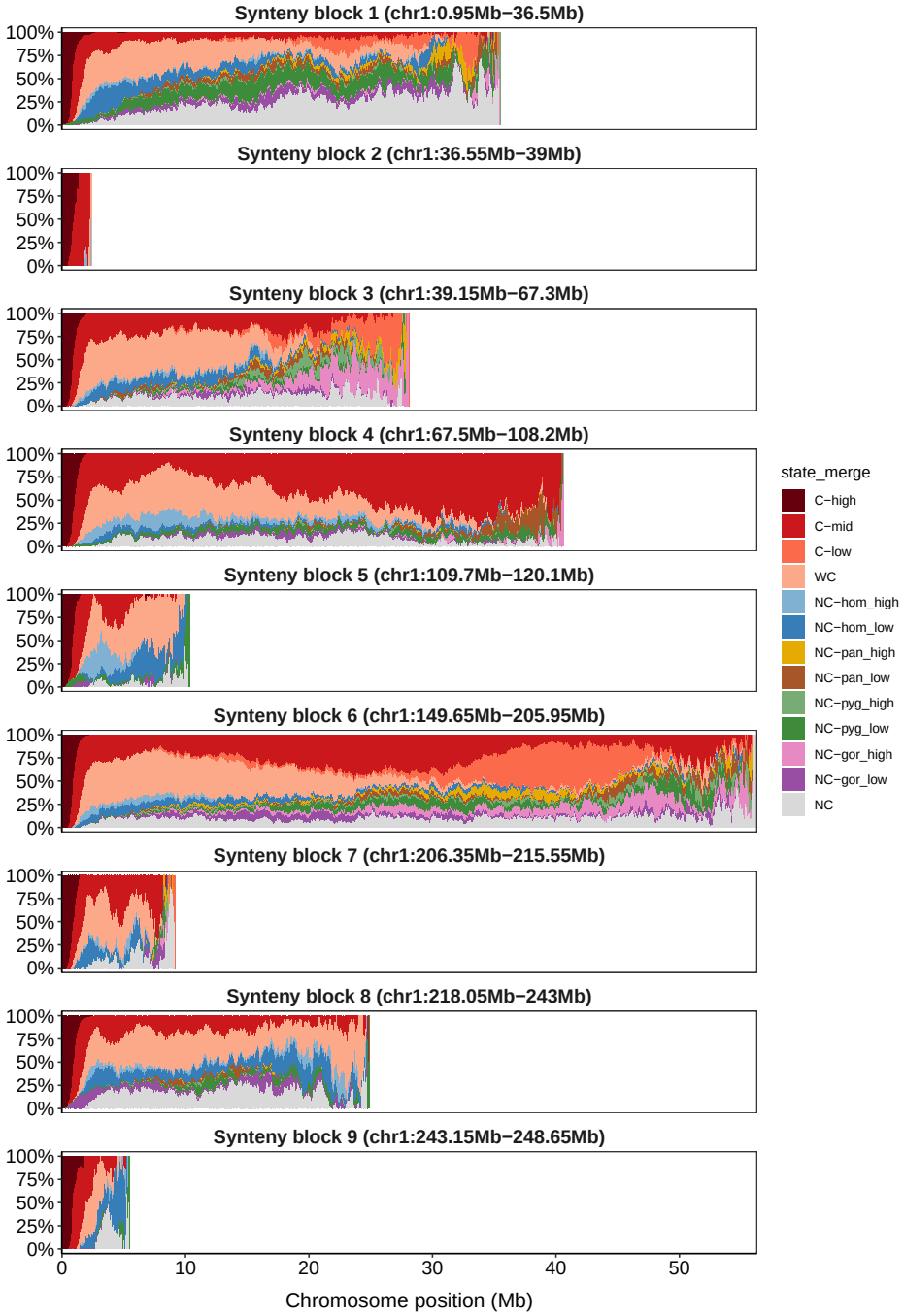

Figure 3.9: Distributions of predicted Hi-C contact evolutionary states over changing distance between paired genomic loci in each synteny block on human chromosome 1.

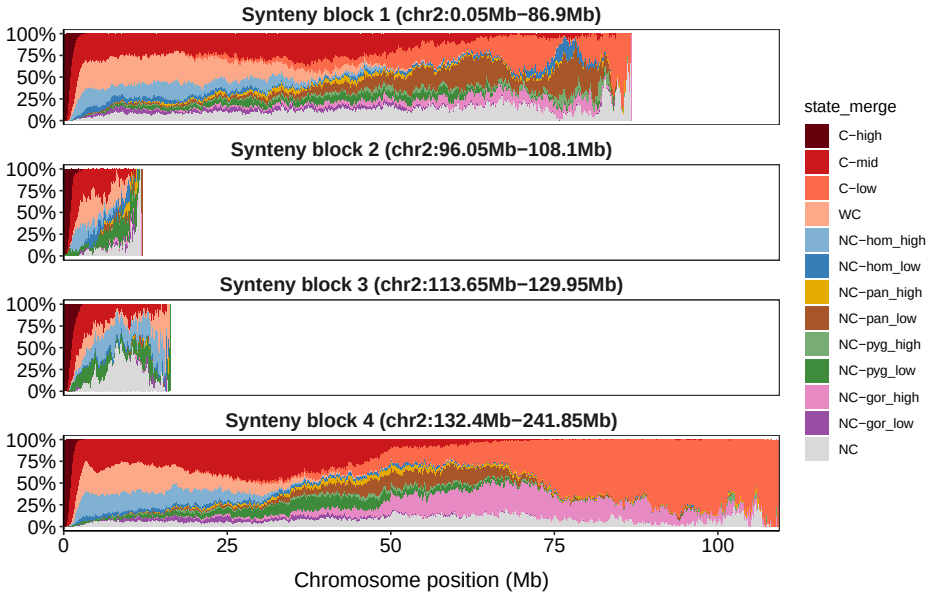

Figure 3.10: Distributions of predicted Hi-C contact evolutionary states over changing distance between paired genomic loci in each synteny block on human chromosome 2.

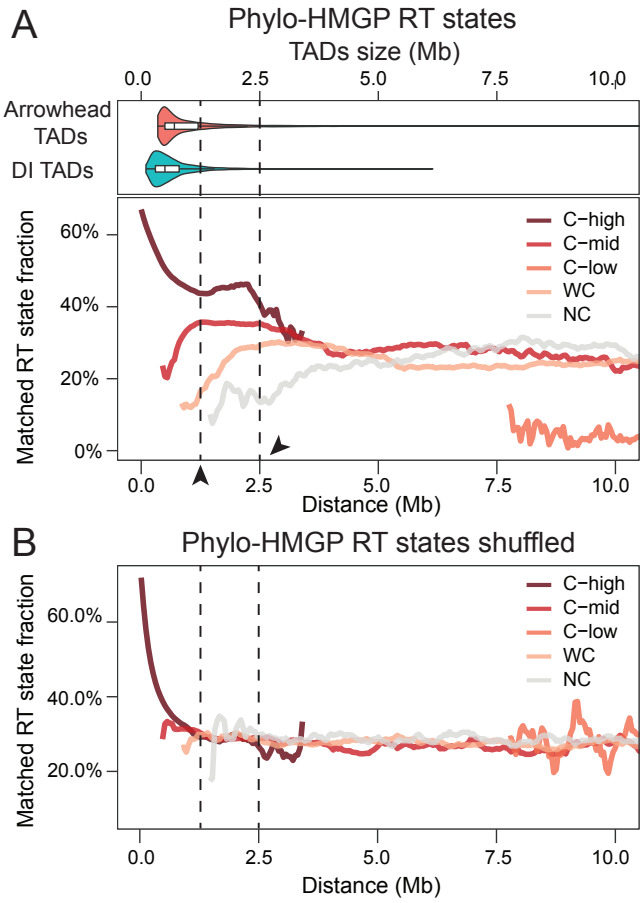

Figure 3.11: Fractions of conserved-in-RT paired genomic loci in different estimated Hi-C contact evolutionary patterns based on shuffled RT states in comparison with the fraction curves based on original predicted RT states. **(A)** Fractions of conserved-in-RT paired genomic loci in different groups of estimated Hi-C contact states based on the RT states identified by Phylo-HMGP (note that this is the same figure in Figure 3.8B for comparison with (B). **(B)** Fractions of conserved-in-RT paired genomic loci in different groups of estimated Hi-C contact states based on shuffled RT states. RT states were predicted at the resolution of 6Kb in Phylo-HMGP (each genomic bin is 6Kb in size). Adjacent genomic bins with the same RT states were merged into segments and random shuffle was performed on the segments in each chromosome individually to obtain the shuffled RT states.

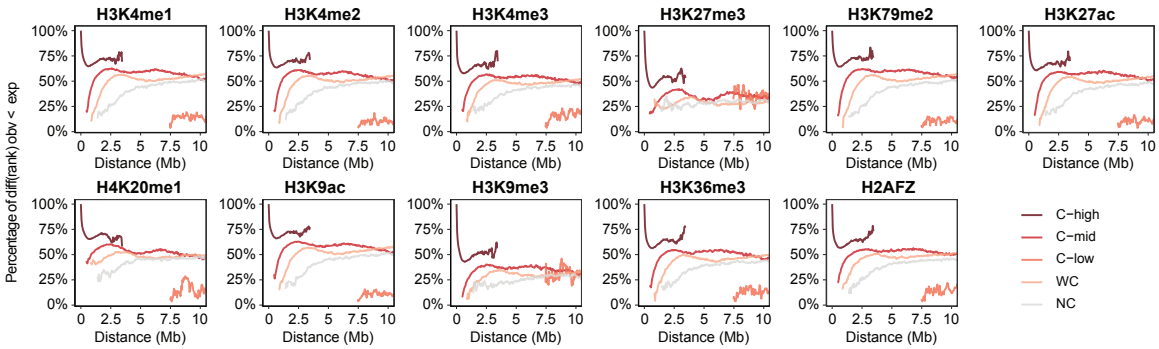

Figure 3.12: The percentage of paired genomic loci that have similar signal strength of a specific type of histone modification in the five estimated Hi-C contact evolutionary pattern groups. Each subplot corresponds to a type of histone modification. There are 11 types of histone modifications included in analysis. The histone modification types are shown in the corresponding sub-plot titles.

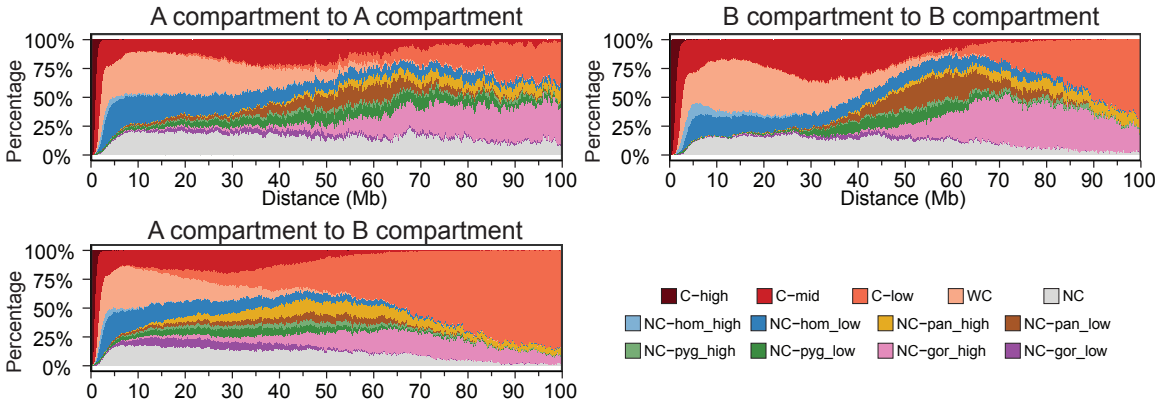

Figure 3.13: Distributions of predicted Hi-C contact evolutionary states over changing distance between paired genomic loci from the A/B compartments.

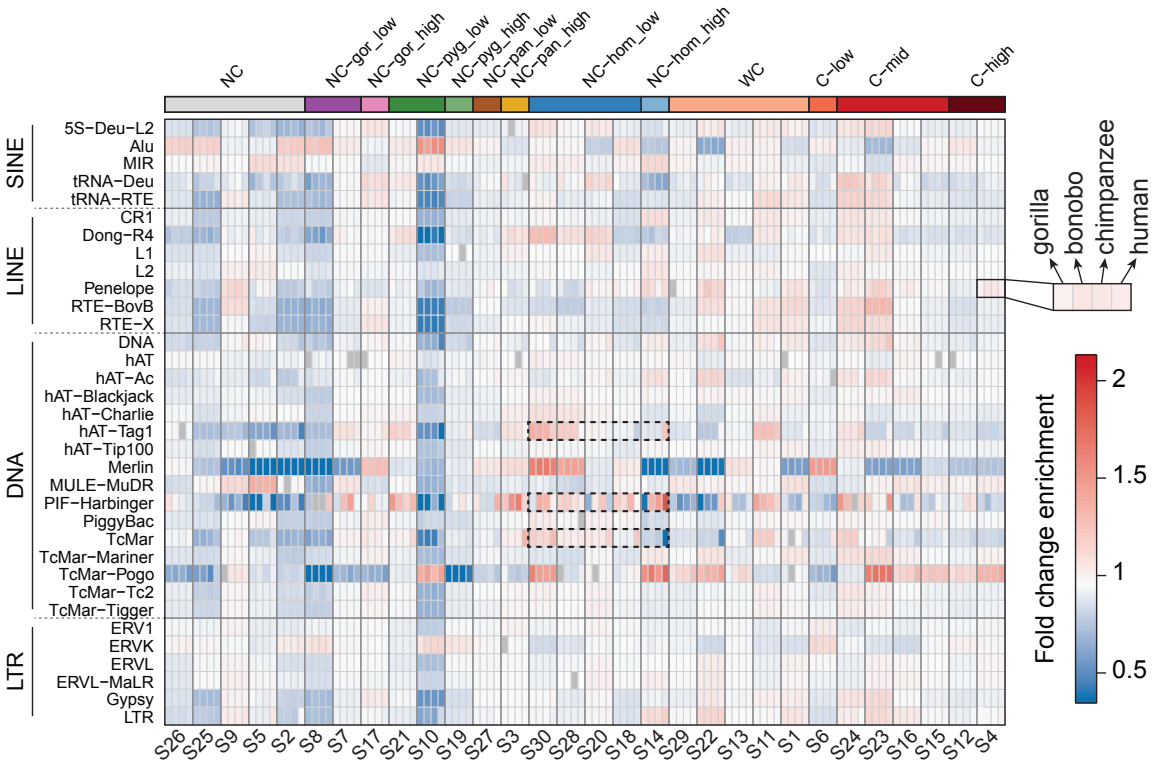

Figure 3.14: Fold change of TE enrichment in Hi-C contact evolutionary states identified by Phylo-HMRF. Each row in the matrix corresponds to one TE family. The names of the TE families are shown on the left, grouped by TE classes. Every four columns correspond to TE enrichment patterns of the four primate species in one state. The four columns are in the order of gorilla, bonobo, chimpanzee, and human, respectively. We call the associated four columns a combined column. The state names corresponding to the combined columns are shown below the matrix. The states are arranged according to their group assignment. The color bar on top of the matrix shows the group assignment. Each element of the matrix represents the enrichment fold change of the corresponding TE family in the corresponding species in the corresponding state. The color bar on the right shows the scale of the fold change values. The matrix elements with FDR$>$0.01 are masked with gray colors. The examples of human-specific TE enrichment patterns in human-specific Hi-C contact states are shown with dashed-line borders in the matrix.

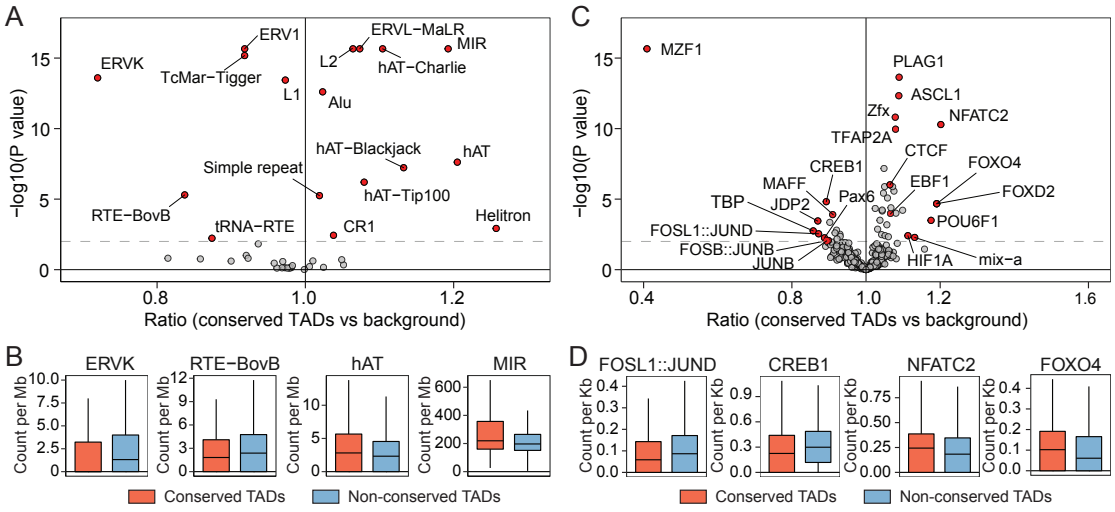

Figure 3.15: TE enrichment and TF binding motif enrichment in conserved long-range interacting TADs. **(A)** Fold change of the normalized TE enrichment in conserved long-range interacting TADs (abbreviated as conserved TADs). The x-axis shows the fold change of conserved TADs versus the background. The y-axis shows the negative log10 of the p-value of the significance that a TE is enriched in the conserved TADs (or not enriched), in comparison with the enrichment of this TE in the background. The dashed line shows the threshold of p-value 0.01. Each dot represents a TE family. The red dots are the TE families significantly enriched (or not enriched) in the conserved TADs. **(B)** The box plots of distributions of single-TAD normalized TE enrichment in the two classes of TADs for four examples of TE families. **(C)** Fold change of the normalized TF binding motif enrichment in conserved TADs. The x-axis shows the fold change of conserved TADs versus the background. The y-axis shows the negative log10 of the p-value of the significance that a TF binding motif is enriched (or not enriched) in the conserved TADs. Each dot refers to a motif. The red dots are examples of the motifs that are significantly enriched (or not enriched) in the conserved TADs. **(D)** The box plots of distributions of single-TAD normalized motif enrichment in the two classes of TADs for four examples of TF binding motifs.

3.11.5 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. Transposable elements (TE) are known to be associated with genome organization [37, 161, 185]. 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 [147] 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 3.14), 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 [186]).

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 Figure 3.15A-B. 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. [185] 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 human genome as GM12878 DNase-seq peak regions (downloaded from the ENCODE project [151]) with +/-250bp extension. We used the software FIMO [148] to scan motifs based on the 579 PWMs of TF binding motifs downloaded from the JASPAR TF binding profile database [187] (with p-value$<$1e-04 as the cutoff) in the open chromatin regions on the human genome. We only retain the motif scanning results for the 345 TFs that are expressed in GM12878 (with FPKM$>$0.1; the RNA-seq datasets used for gene expression analysis are shown in Table 3.1). 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 (Figure 3.15C-D). 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 the conserved and non-conserved TADs.

3.12 Discussion

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 strong correlation with other types

of features for genome structure and function, such as TADs, A/B compartments, DNA replication timing, 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 [188, 189] 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 increase linearly with the tree size, both the computation demand increases and the model is exposed to a higher possibility of local minima and overfitting especially for a smaller sample size of multi-species feature observations. There will also be more potential 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 will make Phylo-HMRF

even more general. We propose to employ the structural EM algorithm [157] for simultaneous polygenetic tree topology inference and parameter estimation. 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 [75, 185, 190], 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 [191]. Based on the Hi-C contact evolutionary states identified with Phylo-HMRF, we made the 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 transposable elements and TF binding 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 has the potential to serve as a generic analytic framework to reveal different evolutionary patterns of chromatin interactions and their connections to the evolution of genome sequence and function.

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 part of study	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
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		
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	Ajit Varki (Chou et al.,2002)	N/A
Troglodytes Gorilla (Gorilla): lymphoblastoid cell line	Ajit Varki (Chou et al.,2002)	N/A
Software and Algorithms		
Phylo-HMRF	This part of study	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

Table 3.1: Table of key resources used in the study of multi-species genome organization comparison using Phylo-HMRF.

Simulation	Method	NMI	AMI	ARI	Precision	Recall	F_1
Dataset I-1	Clustering	0.1096	0.1001	0.0532	0.2445	0.1384	0.1767
Dataset I-1	GMM	0.1213	0.1115	0.0603	0.2502	0.1497	0.1873
Dataset I-1	SLIC	0.0602	0.0554	0.0272	0.2117	0.1280	0.1595
Dataset I-1	Quickshift	0.033	0.0281	0.0265	0.1972	0.3173	0.2433
Dataset I-1	Gaussian-HMRF	0.6214	0.6046	0.6981	0.8068	0.6985	0.7487
Dataset I-1	Phylo-HMRF	**0.6957**	**0.6915**	**0.7786**	**0.8373**	**0.7989**	**0.8177**
Dataset I-2	Clustering	0.1457	0.1074	0.0520	0.5626	0.1349	0.2176
Dataset I-2	GMM	0.1936	0.1434	0.0793	0.6118	0.155	0.2474
Dataset I-2	SLIC	0.0776	0.0573	0.0042	0.4604	0.1113	0.1793
Dataset I-2	Quickshift	0.0519	0.0412	0.0002	0.4519	0.1579	0.2341
Dataset I-2	Gaussian-HMRF	0.4279	0.3368	0.3150	0.8547	0.3470	0.4933
Dataset I-2	Phylo-HMRF	**0.6346**	**0.6191**	**0.8293**	**0.9293**	**0.8807**	**0.9043**
Dataset I-3	Clustering	0.2046	0.1900	0.1269	0.2938	0.1956	0.2348
Dataset I-3	GMM	0.2306	0.2183	0.1728	0.3237	0.2558	0.2857
Dataset I-3	SLIC	0.0724	0.0676	0.0318	0.1881	0.1329	0.1557
Dataset I-3	Quickshift	0.0595	0.0500	0.0198	0.1658	0.3791	0.2307
Dataset I-3	Gaussian-HMRF	0.6447	0.6304	0.6854	0.7534	0.7131	0.7327
Dataset I-3	Phylo-HMRF	**0.7344**	**0.7327**	**0.8010**	**0.8288**	**0.8350**	**0.8319**
Dataset I-4	Clustering	0.1003	0.0850	0.0332	0.3526	0.1207	0.1798
Dataset I-4	GMM	0.1191	0.1017	0.0455	0.3685	0.1355	0.1982
Dataset I-4	SLIC	0.0705	0.0602	0.0159	0.3224	0.1183	0.1731
Dataset I-4	Quickshift	0.044	0.0437	0.0054	0.3025	0.2095	0.2475
Dataset I-4	Gaussian-HMRF	0.3699	0.3179	0.1921	0.5872	0.2249	0.3252
Dataset I-4	Phylo-HMRF	**0.5480**	**0.5275**	**0.6253**	**0.8297**	**0.6396**	**0.7223**
Dataset I-5	Clustering	0.1429	0.1368	0.0657	0.2061	0.1594	0.1798
Dataset I-5	GMM	0.1561	0.1508	0.0737	0.2100	0.1782	0.1928
Dataset I-5	SLIC	0.0620	0.0593	0.0244	0.1644	0.1282	0.1441
Dataset I-5	Quickshift	0.0295	0.0264	0.0073	0.1451	0.2138	0.1729
Dataset I-5	Gaussian-HMRF	0.4990	0.4833	0.4776	0.5968	0.5004	0.5443
Dataset I-5	Phylo-HMRF	**0.6129**	**0.6074**	**0.6442**	**0.7160**	**0.6706**	**0.6924**
Dataset I-6	Clustering	0.1327	0.1137	0.0760	0.3647	0.1561	0.2186
Dataset I-6	GMM	0.1697	0.1479	0.1087	0.3944	0.1939	0.2600
Dataset I-6	SLIC	0.0574	0.0491	0.0093	0.2707	0.1148	0.1613
Dataset I-6	Quickshift	0.0502	0.0488	0.0233	0.2770	0.2318	0.2524
Dataset I-6	Gaussian-HMRF	0.4837	0.4572	0.6622	0.8029	0.6907	0.7426
Dataset I-6	Phylo-HMRF	**0.6203**	**0.6067**	**0.7580**	**0.8582**	**0.7802**	**0.8173**
Dataset I-7	Clustering	0.1547	0.1379	0.0926	0.3506	0.1732	0.2318
Dataset I-7	GMM	0.2250	0.2013	0.1874	0.4632	0.2397	0.3157
Dataset I-7	SLIC	0.1066	0.0946	0.0370	0.2809	0.1334	0.1809
Dataset I-7	Quickshift	0.0577	0.0502	0.0191	0.2451	0.2965	0.2684
Dataset I-7	Gaussian-HMRF	0.5072	0.4754	0.5452	0.7254	0.5694	0.6380
Dataset I-7	Phylo-HMRF	**0.5506**	**0.5190**	**0.5819**	**0.7561**	**0.5975**	**0.6675**
Dataset I-8	Clustering	0.1019	0.0828	0.0153	0.3396	0.1165	0.1735
Dataset I-8	GMM	0.1281	0.1052	0.0246	0.3514	0.1333	0.1933
Dataset I-8	SLIC	0.0873	0.0710	0.0255	0.3558	0.1236	0.1834
Dataset I-8	Quickshift	0.0557	0.0553	0.0457	0.3513	0.2835	0.3138
Dataset I-8	Gaussian-HMRF	0.3286	0.2746	0.1733	0.5354	0.2457	0.3368
Dataset I-8	Phylo-HMRF	**0.5897**	**0.5618**	**0.6550**	**0.8405**	**0.6805**	**0.7520**

Table 3.2: Performance evaluation in simulation study I. Performance evaluation of K-means Clustering, GMM, SLIC, Quick Shift, Guassian-HMRF, and Phylo-HMRF on eight simulated datasets in simulation study I using evaluation measurements NMI (Normalized Mutual Information), AMI (Adjusted Mutual Information), ARI (Adjusted Rand Index), Precision, Recall, and F_1 score. Each method is repeated 10 times on each simulation dataset with different initializations if applicable. The average performance from the 10 repeated runs of each method is presented. The highest performance of the compared methods is in bold font.

Chapter 4

Genome-wide prediction of DNA replication timing from genomic sequences

4.1 Introduction

In Chapter 2 and Chapter 3, we discussed the two proposed models Phylo-HMGP and Phylo-HMRF, which aim at identifying evolutionary patterns across species from the 3D genome organization and genome function data. Here, we turn to the exploration of sequence-dependent regulation mechanism of 3D genome organization and function, identifying sequence elements that may modulate genome organization related genome function, with a focus on studying DNA RT program. In Chapter 1, we have known that there are both epigenetic regulation mechanism and sequence-dependent regulation mechanism hypotheses for the highly regulated temporal patterns of DNA RT. However, both regulation mechanisms are not well understood, without known rules on the regulatory principles of DNA RT in higher eukaryotic cells. In this thesis we focus on the sequence-dependent regulation mechanism hypothesis. Specifically, we develop an interpretable context-of-sequences-aware model for RT prediction, named CONCERT (CONtext-of-sequenCEs for Replication

Timing), using DNA sequence features only. CONCERT unifies (i) modeling of long-range spatial dependencies across different genomic loci and (ii) detection of a subset of genomic loci that are predictive of the studied genomic signal over large-scale spatial domains. Applications of CONCERT to nine human cell lines and mESCs demonstrate that the model reaches high RT prediction accuracy. Crucially, our method can identify sequences that are potentially important for shaping RT program. Our results suggest that dependencies between RT signals and DNA sequences may only exist for a limited number of genomic regions in the whole genome, which is consistent with the finding from the experimentally validated loci of ERCEs [75]. In addition, we generate the landscape of important sequence loci for predicting RT patterns across multiple cell types. CONCERT provides a general framework to predict sequence determinants for large-scale genomic signals. The source code of CONCERT can be accessed at: https://github.com/ma-compbio/CONCERT.

4.2 CONCERT – Intepretable RT prediction using sequence features with context

4.2.1 Overall framework of CONCERT

We develop an interpretable context-attentive model, CONCERT, to simultaneously identify predictive genomic loci that are potentially important for modulating DNA RT and predict genome-wide RT profiles from DNA sequences features. Our CONCERT model is structured with two functionally cooperative components: (1) Selector and (2) Predictor, which are trained jointly within one framework in modeling long-range spatial dependencies across genomic loci, detecting predictive genomic loci, and learning context-aware feature presentations of genomic sequences. The selector is targeted at estimating which genomic loci are of potential importance in modulating the RT profiles, approximately selecting a set of predictive genomic loci by importance estimation-based subset sampling. Leveraging sequence importance estimation from the selector, the predictor performs selective learning of spatial dependencies across genomic loci, to make prediction of RT

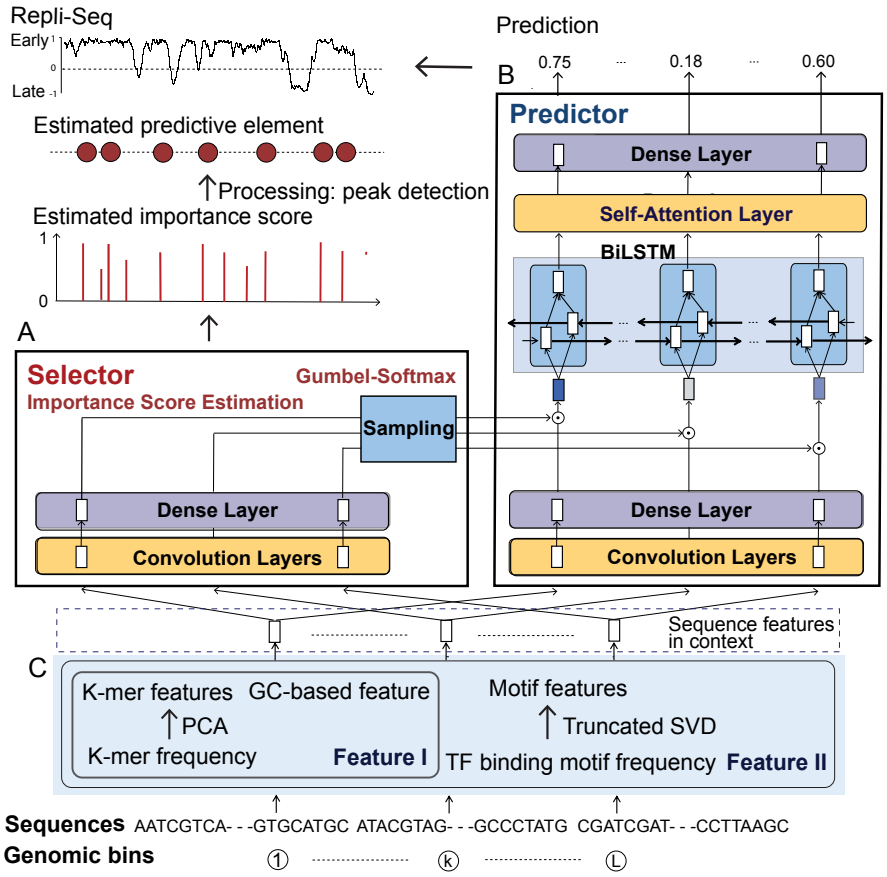

Figure 4.1: Overview of the CONCERT model. The detailed model architecture is shown in Figure 4.2. **(A)** The selector module. The selector uses Gumbel-Softmax trick to perform importance estimation-based genomic loci sampling. The input to the selector are sequence features of genomic loci in a given context window. The selector is connected to the predictor. The output of the selector to the predictor are normalized estimatd importance scores of genomic loci within a context window (with Gumbel-Softmax trick applied). The output of the selector to the users are estimated importance scores of each genomic locus in the context window (without adding random variables from the Gumbel Distribution). We further perform processing of the original estimated importance scores by filtering and local peak detection to identify a set of predictive elements (section 4.8). **(B)** The predictor module. The predictor utilizes the sequence importance estimation from the selector to perform selective learning of spatial dependencies across genomic loci with the BiLSTM layer and the self-attention mechanism. The input to the predictor are the sequence features of genomic loci in a given context window and the importance scores of each locus estimated by the selector. The output of the predictor are predicted RT signals of genomic loci in the given context window. **(C)** The input layer, where input sequence features are extracted from each genomic locus in a given context window. The input feature types include K-mer frequency-based features with dimension reduction, GC profile-based features, and transcription factor (TF) binding motif frequency-based features with dimension reduction. We combination of different types of features. The combination of K-mer frequency-based features and GC-based features is named Feature I. The combination of the three feature types is named Feature II. We primarily use the prediction results from Feature I in analysis.

signals over large-scale spatial domains. The overview of the CONCERT model is shown in Figure 4.1. The input to the model contains DNA sequences of the genomic loci within each context window. The output includes both predicted genomic signals and locus-wise estimated importance scores. The estimated scores can be further processed to delineate predictive sequence elements that are important for the genomic signal prediction. Our method embeds the interpretation mechanism into predictive model training, not requiring extra steps for model explanation. The model is also capable of handling different sizes of context without the increase of model parameters, making it scalable to the context size of genomic loci.

For interpretable RT prediction, we aim to maximize the mutual information between the selected input sequences and the corresponding RT signals, adopting the model interpretation mechanism introduced in [192]. In [192] instance-wise model explanation is performed as a separate step given a trained model. Two rounds of model training are performed, as firstly a task-specific predictive model is trained and secondly an explainer model is trained to interpret the predictive model. Training of the predictive model in the first step did not involve the explainer. The CONCERT model integrates the selector module with the predictor module, to simultaneously predict the genomic signals and estimates the context-related importance of each genomic locus for the prediction. The two modules share information between each other. Training is performed in one round, with the selector and the predictor learned jointly.

Given the context window of size L centered at a genomic locus, let $X = (X_1, \cdots, X_L) \in \mathbb{R}^{L \times d}$, $Y = (Y_1, \cdots, Y_L) \in \mathbb{R}^{L \times c}$ denote the sequence features of genomic loci in the context window and the corresponding RT signals, where $X_k \in \mathbb{R}^d$, $Y_k \in \mathbb{R}$ represent the feature vector and the RT signal of the k-th genomic locus within the context window, respectively. Here d, c denote the dimension of the feature vector for each locus and the number of types of genomic signals, respectively. Let $f_\theta : X \in \mathbb{R}^{L \times d} \rightarrow Y \in \mathbb{R}^{L \times c}$ be the function mapping features associated with a context window to the corresponding signals, with θ as the parameters. We have $c = 1$ in our study. The proposed model is applicable to

data of multiple genomic signal types with $c > 1$. Suppose we select a subset of loci as the predictive loci from a context window. Let $\boldsymbol{\mathcal{V}}^{(L)} = \{\nu \in \mathbb{R}^L | \nu_k \in \{0, 1\}, k = 1, \cdots, L\}$. Let $S = (S_1, \cdots, S_L) \in \boldsymbol{\mathcal{V}}^{(L)}$ represent the selection of predictive genomic loci. S is a L-dimensional binary value vector, with $S_k = 1$ if the k-th genomic locus is selected and $S_k = 0$ otherwise.

Let X_S be the representation of X with only the features of selected genomic loci retained. Let $\odot$ represent element-wise multiplication. Suppose each element corresponds to a genomic locus. We have $X_S = X \odot S \in \mathbb{R}^{L \times d}$. Specifically, let $(X_S)_k$ denote the feature representation of the k-th locus in X_S. If $S_k = 0$, we set $(X_S)_k$ to be a zero-value feature vector. We have $(X_S)_k = \mathbf{0} \in \mathbb{R}^d$ if $S_k = 0$; $(X_S)_k = X_k$ if $S_k = 1$.

Given X_S, the mutual information between X_S and Y is

$$I(X_S, Y) = \mathbb{E}\Big[\log \frac{p(X_S, Y)}{p(X_S)p(Y)}\Big] \tag{4.1}$$

$$= \mathbb{E}\Big[\log \frac{p(Y|X_S)}{p(Y)}\Big] \tag{4.2}$$

$$= \mathbb{E}\Big[\log p(Y|X_S)\Big] + \text{Constant} \tag{4.3}$$

Suppose $\phi_\beta : X \in \mathbb{R}^{L \times d} \to S \in \mathbb{R}^L$ is the function of the selector, mapping X to S, with β as the parameters. The objective function is

$$\max_{\theta, \beta} \mathbb{E}\Big[\log p_\theta(Y|X_S)], \tag{4.4}$$

$$S = \phi_\beta(X), \tag{4.5}$$

which can be rewritten as

$$\max_{\theta, \beta} \mathbb{E}\Big[\log p_\theta(Y|X \odot \phi_\beta(X))]. \tag{4.6}$$

Suppose the number of genomic loci along the genome is N_g and the number of context windows is N_c. We assume that each genomic locus is the center of a context window. In this case, we have $N_g = N_c$. Specifically, suppose there are N consecutive genomic loci along the genome without gaps, which are sequentially indexed with $1, \cdots, N$ in the order

of their genomic coordinates. There are N context windows centered on each genomic locus respectively, indexed with $1, \cdots, N$ correspondingly. Suppose the flanking region size on each side of a locus is l (with genomic locus as the unit). For the i-th locus, the $(i-l)$-th - $(i+l)$-th loci are included in the corresponding i-th context window. We have $L = 2 \times l + 1$. For $i < l$ or $i > N - l$, there are not enough genomic loci to fill in the context window with the i-th locus as the center. There are blank positions in the boundary region of the window. In this case, we use the first genomic locus (if $i < l$) or the N-th locus (if $i > N - l$) to fill in the blank positions of the context window. In real data, there may be missing values or regions not mapped by enough sequencing reads. The genomic loci without measured genomic signals are gaps between the loci with signals. We denote the loci with signals as the sample loci. We use a threshold of the gap (denoted by thr_gap) to divide the sample loci into sequence fragments. Two sample loci with distance larger than thr_gap along the 1D genome on each chromosome are assigned to different sequence fragments. We then delineate the context windows for each locus in each fragment as described above. Let $\boldsymbol{\mathcal{I}}_N = \{1, \cdots, N\}$. Let $x_i \in \mathbb{R}^{L \times d}$, $y_i \in \mathbb{R}^{L \times c}$ be the sequence features and signals of the i-th context window, respectively, $i \in \boldsymbol{\mathcal{I}}_L$. Let $\tilde{x}_k$, $\tilde{y}_k$ be the feature vector and the signal of the k-th genomic locus, respectively, $k \in \boldsymbol{\mathcal{I}}_N$. We have $x_i = (\tilde{x}_{i-l}, \cdots, \tilde{x}_i, \cdots, \tilde{x}_{i+l})$, $y_i = (\tilde{y}_{i-l}, \cdots, \tilde{y}_i, \cdots, \tilde{y}_{i+l})$. Here x_i represents a context window consisting of the sequence features of L associated genomic loci, while $\tilde{x}_k$ represents a single loucs. The context windows are overlapping. One genomic locus is involved in at most L context windows, with the relative positions of the locus in the associated context windows ranging from 1 to L. Accordingly, in general the i-th locus can be in context with the $(i-2l)$-th locus to the $(i+2l)$-th locus.

By using overlapping context windows, we involve the feature representation of a genomic locus in different contexts which are composed of varied combinations of genomic loci, extending the context scope of a single locus from $2l$ loci to at most $4l$ loci and utilizing more diverse context information to capture potential spatial dependencies across different subsets of loci. The overlapping context window can be formed by applying a

sliding window along the genome with stride $d = 1$. If we use $1 < d < L$, we have $N_c < N_g$. In this case, not every genomic locus is the center of a context window, but every genomic locus is covered by at least one context window. Each context window is overlapped with at least one another context window. The learned model can still generate importance score estimates and RT signal predictions for each locus. In the prediction stage, we use the trained model to predict RT signals $\hat{y}_i \in \mathbb{R}^{L \times c}$ and estimate locus-wise importance scores $\alpha_i \in \mathbb{R} \in \mathbb{R}^L$ for the loci in each context window. For each locus that is involved in multiple context windows, we use the average of all the RT predictions for this locus in the associated context windows as the predicted RT signal, and we choose the importance score estimated in the context window where the locus is the center. In the case that $1 < d < L$ and $N_c < N_g$, we use the importance score averaged from the estimated scores for this locus in all the associated context windows.

4.2.2 The selector module in CONCERT

With the selector module we aim to select a subset of genomic loci that are predictive of the RT signals from the context window. We begin with the assumption that we only select one predictive genomic locus from the L loci in the context window. Suppose a selected locus is represented by a discrete random variable $\hat{s}_i \in \boldsymbol{\mathcal{V}}^{(L,1)}$. We approximate $\hat{s}_i$ with category probabilities $s_i = (s_{i,1}, s_{i,2}, \cdots, s_{i,L}) \in \mathbb{R}^L$. Each category represents selecting the corresponding locus as the predictive locus. We use the Gumbel-Softmax method [193] to approximate a categorical distribution with a Concrete distribution which is differentiable.

First, we add randomness to the probability of selecting each locus using random variables G_k, $k = 1, \cdots, L$. $\{G_k\}_{k=1}^L$ are independently sampled from a Gumbel distribution. Let

$$G_k = -\log(-\log U_k),\ U_k \sim \text{Uniform}(0, 1), \tag{4.7}$$

where $\text{Uniform}(0, 1)$ represents uniform distribution on $[0, 1]$. We have $G_k \sim \text{Gumbel}(0, 1)$. The use of random perturbation facilitates learning robust selection that is less sensitive to

noise.

Suppose $\alpha_1, \cdots, \alpha_L$ are unnormalized category probabilities, $\alpha_k > 0$, $k = 1, \cdots, L$. If we use the Gumbel-Max trick [193], we select the $\tilde{k}$-th genomic locus as the predictive locus with the criterion

$$\tilde{k} = \arg\max_k \{\log \alpha_k + G_k\}_{k=1}^{L}. \tag{4.8}$$

The objective function is not differentiable everywhere in its domain. We then employ the Gumbel-Softmax trick which was developed for continuous relaxation of the $\arg\max$ function [193]. Using the Gumbel-Softmax trick, the probability of selecting locus k is

$$Z_k = \frac{\exp((\log \alpha_k + G_k)/\tau)}{\sum_{k=1}^{L} \exp((\log \alpha_k + G_k)/\tau)}, \tag{4.9}$$

where τ is a parameter. The softmax function approaches the $\arg\max$ function as τ approaches zero, with the orders of $\{\log \alpha_k + G_k\}$ preserved [193]. We denote the distribution of $Z = (Z_1, \cdots, Z_L) \in \mathbb{R}^L$ as a Concrete distribution parameterized by (α, τ), where $\alpha = (\alpha_1, \cdots, \alpha_L)$. We have $Z \sim \text{Concrete}(\alpha, \tau)$.

To sample a subset of m predictive genomic loci from the L loci, we sample a vector $\tilde{z}$ from $\boldsymbol{\mathcal{V}}^{(L,m)}$. We adopt the strategy proposed in [192] to use continuous relaxation for the approximation of sampling $\tilde{z}$. First, we sample m Concrete random vectors $z^{(1)}, \cdots, z^{(m)}$ independently from the distribution $\text{Concrete}(\alpha, \tau)$. Second, for each genomic locus k, we take the maximal value at the k-th dimension of the sampled random vectors as the estimated probability of selecting locus k.

Let $\tilde{S} = (\tilde{S}_1, \cdots, \tilde{S}_L)$, where $\tilde{S}_k$ represents the probability of selecting locus k. $\tilde{S}$ is a continuous approximation of the discrete vector $\tilde{z}$. As defined, $X = (X_1, \cdots, X_L)$ represents the sequence features in a context window. We have

$$Z^{(j)} \sim \text{Concrete}(\alpha, \tau), j = 1, \cdots, m, \tag{4.10}$$

$$\alpha = \tilde{\phi}_\beta(X), \tag{4.11}$$

$$\tilde{S}_k = \max\{Z_k^{(1)}, \cdots, Z_k^{(m)}\}, k = 1, \cdots, L, \tag{4.12}$$

where $\tilde{\phi}_\beta(\cdot)$ is the function mapping the sequence features to $\alpha = (\alpha_1, \cdots, \alpha_L)$ in the selector module, with parameters β. We interpret α_k as the estimated importance score of the k-th locus.

Let G denote the set of random variables $\{\{G_k^{(j)}\}_{k=1}^{L}\}_{j=1}^{m}$ sampled from the Gumbel distribution. We have $\tilde{S} = \phi_\beta(X, G)$. Suppose the number of genomic loci is N. As each locus is the center of a context window, the number of context windows is N. Using regularization on model complexity to reduce model over-fitting, the objective function is

$$\min_{\theta,\beta} \; -\mathbb{E}_{\theta,\beta}\big(\log p_\theta(Y|X, \phi_\beta(X, \epsilon))\big) + \lambda_1\Omega_\theta + \eta_1\Omega_\beta, \tag{4.13}$$

where $\epsilon = \{\epsilon^{(j)} \in \mathbb{R}^L\}_{j=1}^{m}$ consists of random variables independently sampled for $x \in \mathbb{R}^{L\times d}$. We have $\epsilon^{(j)} = (\epsilon_1^{(j)}, \cdots, \epsilon_L^{(j)})$, where $\epsilon_k^{(j)} \sim$ Gumbel$(0, 1)$, $k = 1, \cdots, L$, $j = 1, \cdots, m$. Ω_θ, Ω_β represent model complexities of the predictor and the selector, respectively, with λ_1, η_1 as corresponding regularization coefficients. Specifically, we use l_2 norm of model parameters as measurements of Ω_θ and Ω_β. In the implementation, the selector module consists of one dense layer for preliminary feature transformation, a convolution layer for utilizing limited local context information, followed by two subsequent dense layers and an activation layer to generate estimates of α, and a Gumbel distribution based random variable sampler (section 4.4, Figure 4.2).

4.2.3 The predictor module in CONCERT

The predictor module consists of four sequentially connected parts. The first part is a preliminary feature transformation sub-module, with an interface from the selector. The sub-module consists of one convolution layer and two dense layers, performing local transformations of the locus-wise input features from a context window to generate an intermediate feature representation for each locus. The output of the sub-module is further weighted by the importance scores $s_i = \phi_\beta(x_i, \epsilon_i) = (s_{i,1}, \cdots, s_{i,L}) \in \mathbb{R}^L$ estimated by the selector for predictive genomic loci selection. The weighted feature is $\tilde{x}_i = f_\theta^{(1)}(x_i) \odot \phi_\beta(x_i, \epsilon_i)$, where $f_\theta^{(1)}(\cdot)$ is the mapping function of the sub-module and $\odot$ represents element-wise multiplication.

The second part is a BiLSTM (Bi-directional Long Short-Term Memory) layer [194], which enables bi-directional context information sharing across genomic loci (section **??**). Recurrent neural networks (RNNs) have been widely used for modeling sequential data in diverse applications [195–197]. LSTM [97] and GRU [198] are two representative variants of RNNs, using cell memory and gate mechanisms to overcome the limitation that standard RNNs cannot learn long-range dependencies (interactions between time points or positions that are a number of steps apart) in practice. The standard LSTM is featured with the cell state, the *input* gate, the *forget* gate, the *output* gate, and the hidden state:

$$\tilde{c}_t = \sigma_c(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \tag{4.14}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{4.15}$$

$$h_t = o_t \odot \sigma_h(c_t), \tag{4.16}$$

$$i_t = \sigma_i(x_t W_{xi} + h_{t-1}W_{hi} + b_i), \tag{4.17}$$

$$f_t = \sigma_f(x_t W_{xf} + h_{t-1}W_{hf} + b_f), \tag{4.18}$$

$$o_t = \sigma_o(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \tag{4.19}$$

where i_t, f_t, o_t represent the *input* gate, the *forget* gate, and the *output* gate for the i-th position, respectively. σ_i, σ_f, σ_o, σ_c, and σ_h denote nonlinear activation functions. x_t, c_t, h_t represent the input feature, the cell state, and the hidden state at the i-th position, respectively. LSTM uses cell state c_t as memory to pass information through the series of data. The cell state c_t is updated with a fraction of the previous cell state c_{t-1} and a fraction of the candidate cell state $\tilde{c}_t$, which is a nonlinear function of the input x_t and the previous hidden output state h_{t-1}. The fractions are controlled by the *forget* gate and the *input* gate i_t, respectively. $\{W_{hi}, W_{hf}, W_{ho}, W_{xi}, W_{xf}, W_{xo}\}$ are weight parameters.

BiLSTM overcomes the limitation of the standard LSTM that information flow is only allowed in one direction. BiLSTM splits the hidden states of a standard LSTM into forward and backward states, forming two non-interfering parallel LSTM layers with bi-directional information flow. The output of BiLSTM is concatenation of the forward and backward states, producing a context-aware feature representation for each locus. We utilize BiLSTM

to model the bi-directional spatial dependencies across a sequence of genomic loci.

The third part is the self-attention layer, which is employed to capture dependencies between two arbitrary loci within a context window [199]. With BiLSTM, information sharing is constrained by the spatial order of the loci. The shared information vanishes as the distance between loci increases, making it still challenging to model dependencies between distant loci. For each specific genomic locus l (noted as a target locus, with feature representation x_l), the self-attention layer estimates an attention value of each locus l' (including the target locus) in the context window, measuring the importance of $x_{l'}$ for predicting the signal of x_l. The attention value is calculated using the feature interaction between each locus and the target locus. The features of all the loci in the context are weighted by their target-specific attention values and averaged to generate an updated context-aware feature vector of the target locus. Specifically, the self-attention mechanism used in the model is:

$$h_{l,l'} = \sigma_h(W_l x_l + W_x x_{l'} + b_l), \tag{4.20}$$

$$e_{l,l'} = \sigma_a(W_a h_{l,l'} + b_a), \tag{4.21}$$

$$a_{l,l'} = \frac{e_{l,l'}}{\sum_{k \in \boldsymbol{\mathcal{C}}_l} e_{k,l'}}, \tag{4.22}$$

$$\tilde{x}_l = \sum_{l' \in \boldsymbol{\mathcal{C}}_l} a_{l,l'} x_{l'}. \tag{4.23}$$

where $\boldsymbol{\mathcal{C}}$ represents the index set of genomic loci in the context window. σ_h, σ_a denote nonlinear activation functions. $\{W_l, W_x, b_l, W_a, b_a\}$ are model parameters. For a context window of size L, the estimated attention values can be represented by a matrix $A_L = \{a_{l,l'}\}_{l,l' \in \boldsymbol{\mathcal{C}}} \in \mathbb{R}^{L \times L}$, where $a_{l,l'}$ represents the attention value locus l' receives from locus l. The self-attention layer selectively aggregates the outputs from the BiLSTM layer to model distance-insensitive spatial dependencies across genomic loci.

The fourth part is a fully-connected layer, mapping the features of each locus to a continuous value as the predicted RT signal. We use Keras/Tensorflow [200, 201] to implement CONCERT. The architecture details and hyperparameters of the model used in the experiments are described in section 4.4.

4.3 Feature representation of genomic loci

4.3.1 Feature engineering with specific sequence patterns

With prior knowledge of the possible connections between DNA sequence and RT signals, we extract different types of features from each genomic locus. We concatenate different types of features to form a feature vector of the corresponding genomic bin, which is used as the input layer of the model. The predefined features include the transformed K-mer frequency features, the GC profile based features, the optional transcription factor (TF) binding motif features, and the optional conservation score based features (section 4.3). Let $x_i^{(kmer)}$, $x_i^{(gc)}$, $x_i^{(motif)}$, and $x_i^{(phyloP)}$ be the corresponding four types of features of the i-th genomic locus, respectively, $i = 1, \cdots, N$. We primarily use the K-mer frequency features and in our study.

Feature representation based on K-mer frequency

First, we extract K-mer frequency features from the DNA sequence of each genomic locus. A DNA subsequence of length K is denoted as K-mer. There are four types of nucleotides: A, G, C, T. Therefore, there are 4^K different types of K-mers for a specific K. For a specific K, we count the occurrences of each type of K-mer in the sequence of each locus, normalized by the sequence length. The normalized K-mer frequency feature is 4^K-dimensional. The feature dimensionality increases exponentially as K increases. Let $\tilde{x}_i^{(kmer,K)}$ be the original frequency-based feature vector for a specific K. For the balance between prediction performance and computation cost, we choose $K = 5$ and $K = 6$ to form concatenated feature vector $\tilde{x}_i^{(kmer)} = (\tilde{x}_i^{(kmer,5)}, \tilde{x}_i^{(kmer,6)}) \in \mathbb{R}^{5120}$. We extract K-mer based features from the DNA sequence of each genomic locus. There are four types of nucleotides: A, G, C, T. Therefore, there are 4^K different types of K-mers for a specific K. For a given sequence of a locus and a specific K, the number of occurrences of each type of K-mer in this sequence, normalized by the sequence length, forms a 4^K-dimensional feature vector. Let $\tilde{x}_i^{(kmer)}$ be the K-mer based feature vector of the i-th locus, and $\tilde{x}_i^{(kmer,K)}$ be the feature vector for a specific K. We choose $K = 5$ and $K = 6$, and concatenate

$\tilde{x}_i^{(kmer,5)}$ and $\tilde{x}_i^{(kmer,6)}$ for the i-th locus. We have $\tilde{x}_i^{(kmer)} = (\tilde{x}_i^{(kmer,5)}, \tilde{x}_i^{(kmerr,6)}) \in \mathbb{R}^{5120}$. Since the K-mer frequency based feature vector is high-dimensional, we used different approaches to reduce the feature dimensionality, and evaluated the prediction performance on validation data using the dimension-reduced features. The methods we tested for dimension reduction include PCA (Principal Component Analysis) [202], Kernel PCA [203], Sparse PCA [204], ICA (Independent Component Analysis) [205], Mini-Batch Dictionary Learning [206], Truncated SVD (singular value decomposition) [207], and Autoencoder [208]. We used the scikit-learn library [144] for implementation of the feature dimension reduction methods.

Specifically, we pooled the training samples and test samples, and performed feature dimension reduction on the pooled data using each of the methods above, without using the label information of the samples. We compared the RT prediction performance based on features from different dimension reduction methods. We choose PCA as the dimension reduction method for the K-mer frequency features and choose the reduced dimension to be $d_{kmer} = 50$ based on evaluation of the prediction performance and computation efficiency. For the PCA method, we use the top d_{kmer} components with the largest variances on the transformed coordinates in the new feature space after dimension reduction as the feature values for each sample.

Feature representation using GC profile

GC content is known to correlate with RT profiles [209]. Early replications were found to be enriched in GC-rich regions. We extract two types of GC-based features from the sequence of each genomic locus, which are GC content (denoted by $f^{(gc)}$) and GC skew (denoted by $f^{(skew)}$), respectively. Let $n(\alpha)$ be the number of nucleotide α in a given DNA sequence, $\alpha \in \{A, G, C, T\}$. For a given genomic locus, we have

$$f_i^{(gc)} = \frac{n(G) + n(C)}{n(A) + n(G) + n(C) + n(T)}, \tag{4.24}$$

$$f_i^{(skew)} = \frac{n(G) - n(C)}{n(G) + n(C)}. \tag{4.25}$$

$f^{(gc)} \in [0, 1]$, $f^{(skew)} \in [-1, 1]$. Let $x_i^{(gc)}$, $f_i^{(gc)}$, $f_i^{(skew)}$ be the GC-based feature vector, the

GC content feature, and GC skew feature of the i-th genomic locus, respectively. We have $x_i^{(GC)} = (f_i^{(gc)}, f_i^{(skew)})$.

Feature representation using transcription factor binding motifs

We use TF binding motif frequency as the fourth type of feature for each genomic locus. Let $x_i^{(motif)}$ be the TF binding motif-based feature of the i-th locus. We used the software FIMO [148] and 769 position weight matrices (PWMs) of TF binding motifs from the HOCOMOCO v11 database [210] to perform motif scanning on human genome hg38. In each bin, we computed the motif frequency for each PWM within the bin (p-value$<$1e-05 required for each motif). We then normalized the frequency by the genomic bin size, resulting in a 769-dimensional feature vector. We also performed dimension reduction on the motif features. Since the normalized motif frequency features are sparse and the Truncated SVD method [207, 211] is appropriate for transformation of sparse features, we use the Truncated SVD method for dimension reduction of the motif features, with the new feature dimension to be $d_{motif} = 50$.

Feature representation using phyloP scores

We download the phyloP (phylogenetic p-values) scores on human genome from the UCSC Genome Browser [212]. The phyloP scores are conservation scores calculated by phyloP [213] from the PHAST package [214] based on multiple sequence alignments of 99 vertebrate genomes to the human genome hg38. We call the phyloP scores 100-way phyloP scores. A 100-way phyloP score is assigned per base pair in most regions of the human genome. For each genomic bin, we extract a feature vector from the phyloP scores in this bin by calculating different types of statistics of the scores. Suppose $c_i = \{c_{i,1}, \cdots, c_{i,N_i}\}$ are the phyloP scores in the genomic bin of the i-th sample, where N_i is the number of phyloP scores in this bin. First, we calculate the average, median, maximum and minimum of the scores in c_i. Second, we calculate the distribution of the scores in different intervals. We observe that the phyloP scores range from -20 to 10. We divide the range [-20,10] into 15 evenly spaced non-overlapping intervals with length 2 each. The intervals are $[-20, -18), [-18, -16) \cdots, [6, 8), [8, 10]$. We calculate the frequency

of phyloP scores in each interval, normalized by the total number of phyloP scores in c_i. Let $x_{i,1}^{(phyloP)}, x_{i,2}^{(phyloP)}, x_{i,3}^{(phyloP)}, x_{i,4}^{(phyloP)}$ and $x_{i,5}^{(phyloP)}, \cdots, x_{i,19}^{(phyloP)}$ be the four statistics from the first step and the 15 normalized frequencies from the second step, respectively. Let $x_i^{(phyloP)}$ bet the feature vector based on phyloP scores for the i-th locus. We have $x_i^{(phyloP)} = (x_{i,1}^{(phyloP)}, \cdots, x_{i,19}^{(phyloP)})$.

4.3.2 Feature representation learning from local genomic sequences

We also develop a model variant which incorporates a local-level sub-model combining convolution nerual networks (CNN) with BiLSTM to learn local-context aware feature representations from locus-wise sequences (section 4.5), as an alternative to the CONCERT model with pre-engineered feature representations. With the CNN-BiLSTM sub-model integrated, the model is re-formulated into a framework with a two-level hierarchical structure, providing a local-global multi-resolution scheme. At the first level, the sub-model aims to capture spatial dependencies within the sequence of each locus, functioning as a local feature extraction component. At the second level, the model learns dependencies across different genomic loci over a larger-scale domain. The CONCERT-hierarchical structure shows similar RT prediction accuracy to the described basic CONCERT model in human cell lines (Table 4.3). We focus on the basic CONCERT model in the result analysis. Still, the hierarchical structure provides an option to enable the flexibility of learning long-range spatial dependencies while remaining attentive to local sequence feature patterns.

4.4 Model architecture and hyperparameters

We implemented the model using Keras/Tensorflow. The architecture of the proposed model as implemented by Keras/Tensorflow is shown in Figure 4.2.

For both the selector and the predictor modules, each of the dense layers and convolution layers is followed by a batch normalization layer [215] and an activation layer. The dense layer is applied to the locus-wise feature representations. The convolution layer is

applied to the series of feature representations of the loci associated with the given context window. The main components of the selector module are as follows.

1. A dense layer for preliminary feature transformation, followed by batch normalization and ReLU (rectified linear unit) activation. Number of hidden units: 50.
2. A 1D (one-dimensional) convolution layer to utilize local context information for importance estimation of each locus, followed by batch normalization and ReLU activation. Number of filters: 50. Kernel size:3. Stride: 1. Padding method: 'same' (the left and right sides of the input are padded with zero values evenly such that the output has the same width dimension as the input).
3. Dense layer for feature transformation, followed by batch normalization and ReLU activation. Number of hidden units: 25.
4. Dense layer for estimating importance scores, followed by batch normalization and ReLU or sigmoid or linear activation. Different activation functions used correspond to different model variants. Number of hidden units: 1.
5. A random variable sampler that samples random variables from Gumbel distribution Gumbel(0,1). Sampling are repeated k times independently for each context window. We choose $k = 10$. The temperature parameter $\tau = 0.1$.
6. An element-wise Add layer to add the random variables from Gumbel(0,1) to the estimated importance scores. For each locus, there are k values corresponding to the k samplings. The maximal value of the k results are retained for the corresponding locus. We choose $k = 10$ and the temperature parameter $\tau = 0.1$.
7. Softmax layer following the element-wise Add layer to generate importance score estimates.

The main components of the predictor module are as follows.

1. A dense layer for preliminary feature transformation, followed by batch normalization and ReLU activation. Number of hidden units: 50.
2. A 1D convolution layer, followed by batch normalization and ReLU activation. Number of filters: 50. Kernel size:3. Stride: 1. Padding method: 'same'.

3. Dense layer for feature transformation, followed by batch normalization and ReLU activation. Number of hidden units: 50.
4. An element-wise multiplication layer to multiply the feature representations of each locus with the importance scores estimated by the selector using the Gumbel-Softmax method.
5. BiLSTM layer. Number of hidden units in each single LSTM layer: 32. Output dimension: 64. Layer normalization applied. Activation function: tanh.
6. Self-attention layer. Number of hidden units: 50. Activation function: tanh.
7. Dense layer, followed by batch normalization and sigmoid or tanh activation to predict the genomic signals. Number of hidden units: 1.

We use two types of loss functions: mean squared error (MSE) or log hyperbolic cosine loss (log-cosh loss). Log-cosh loss is the logarithm of the hyperbolic cosine of the difference between the predicted and ground truth values. Let $y \in \mathbb{R}^d$, $\hat{y} \in \mathbb{R}^d$ denote the ground truth and predicted values, respectively. The log-cosh loss is

$$L(y, \hat{y}) = \sum_{i=1}^{d} \log(\cosh(y_i - \hat{y}_i)) \tag{4.26}$$

$$= \sum_{i=1}^{d} \log(\frac{1}{2}(e^{(y_i - \hat{y}_i)} + e^{-(y_i - \hat{y}_i)})). \tag{4.27}$$

For $\delta_i = y_i - \hat{y}_i$, $\log(\cosh \delta_i)$ approximates $\frac{1}{2}\delta_i^2$ for small $|\delta_i|$ and approximates $|\delta_i| - \log 2$ for large $|\delta_i|$. Therefore, the log-cosh loss is similar to MSE but it is not very sensitive to the outliers in the data. RT prediction evaluations with the two types of loss functions show that the prediction performance from the two functions is similar on the Repli-seq data in our study. We use the Adam optimizer for model parameter estimation. The batch size is 512. The learning rate is 0.0005. We choose the the l_1 regularization coefficients (applied to both selector and predictor modules) to zero. We choose the the l_2 regularization coefficients (applied to both selector and predictor modules) to 1e-04 or 1e-05. The parameter configurations were selected based on searching through a range of combinations of different parameters with evaluation on the validation data. We use early stopping where

training is stopped if no performance gain is observed on the validation data for consecutive 7 epochs.

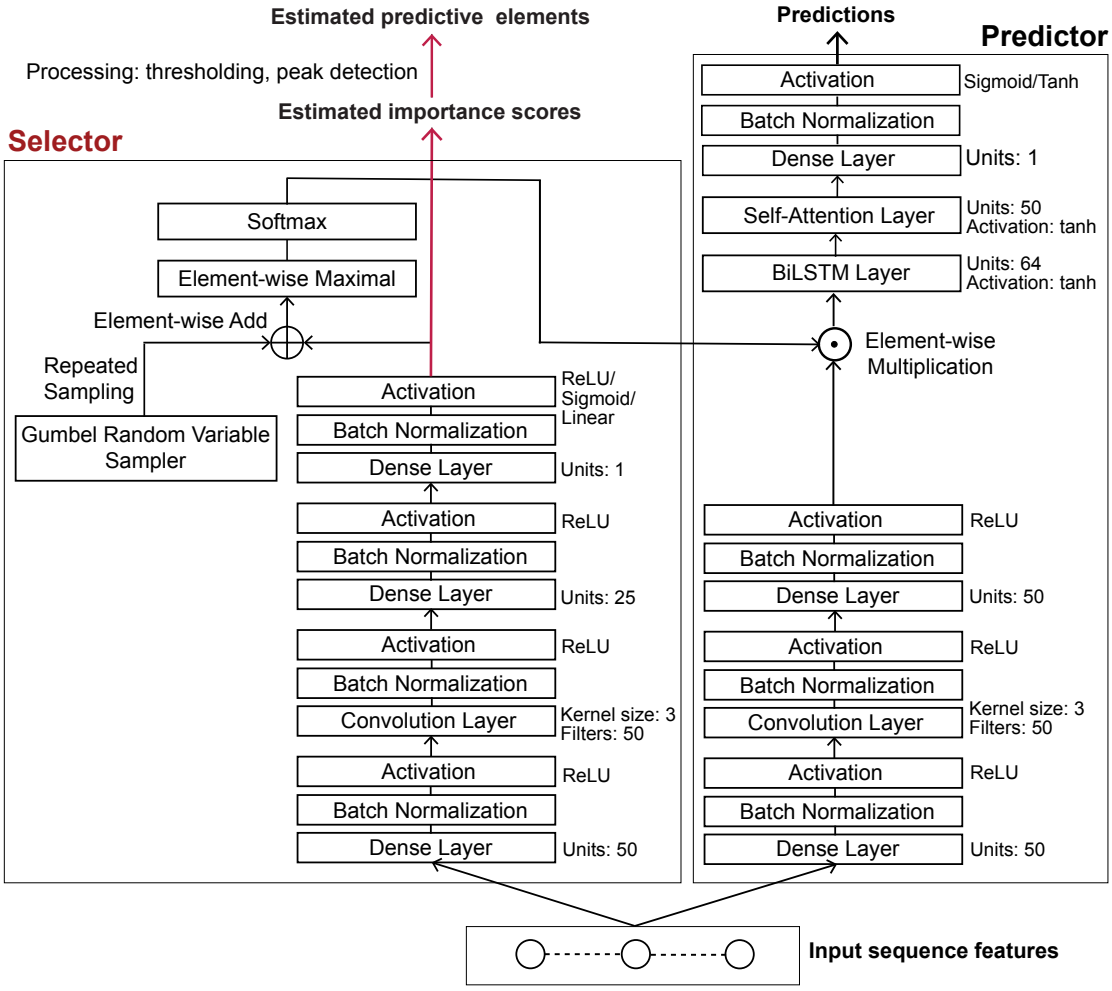

Figure 4.2: The architecture of the CONCERT model. The two primary modules are the selector and the predictor. The model is implemented using Keras/Tensorflow [200, 201]. The parameter configurations used for each layer or component in practice are shown beside the corresponding layer or component.

4.5 Hierarchical structure to learn feature representation from genomic sequences

We also propose a hierarchical structure of CONCERT (denoted as CONCERT-Hierarchical) as an alternative to the basic CONCERT model to learn feature representation from locus-wise genomic sequences. The architecture of CONCERT-Hierarchical consists of two levels. The first level is a CNN-BiLSTM sub-model applied to each locus in the context window, to perform local feature representation learning of locus-wise sequences, by learning spatial dependencies across different positions of the sequence of each locus. The second level is the same as the basic CONCERT model without the input layer for pre-engineered feature representation, which consists of the selector module and the predictor module. The input to the first level is one-hot encoded feature vectors of the sequences from each locus in each context window. The output from the first level is used as input to the second level, to replace the original pre-engineered input layer. Suppose there are L genomic loci in each context window, with the genomic bin size of w. The input feature vector of each locus is a binary-valued matrix of dimensions $4 \times w$, where each row corresponds to the nucleotide type $\alpha \in \{A, C, G, T\}$ and each column represents a base in the sequence of the corresponding locus. For each column, the row corresponding to the nucleotide at this position is set to value 1, and the other rows are zeros. The input feature of each context window is $(4 \times w \times L)$ dimensional. The structure of the CNN-BiLSTM sub-model is as follows:

1. Three sequentially connected convolution layers, each followed by a max pooling layer, a batch normalization layer, and ReLU activation. The parameter configurations of the three convolution layers and the max pooling layers are as follows:

   Convolution layer 1: number of filters: 128, kernel size: 10, stride: 5, padding method: "valid" (input is not padded on each side);

   Max pooling layer 1: pooling size: 10, stride for pooling: 10;

   Convolution layer 2: number of filters: 32, kernel size: 7, stride: 1, padding: "valid";

Max pooling layer 2: pooling size: 5, stride: 5;

Convolution layer 3: number of filters: 32, kernel size: 5, stride: 1, padding: "valid";

Max pooling layer 3: pooling size: 2, stride: 2.

2. BiLSTM layer. Number of hidden units in each single LSTM layer: 32. Output dimension: 64. Layer normalization applied. Activation function: tanh.
3. Dense layer, followed by batch normalization and ReLU activation. Number of hidden units: $d = 50$.
4. Max pooling layer or concatenation layer. If max pooling layer is used, max pooling is applied to the outputs of the dense layer for the loci in the context window to form a d-dimensional feature vector. If concatenation is used, the outputs of the dense layer for each locus in the context window are concatenated to form a feature vector of dimension $w \times d$.

4.6 Data collection and processing

We collected genome-wide RT maps based on Repli-seq data [38] in 9 human cell lines GM12878, H1-hESC, H9-hESC, HCT116, HEK293, K562, IMR90, RPE-hTERT, and U2OS. We divided the genomes into consecutive non-overlapping 5Kb bins (genomic loci). We followed the scripts on https://github.com/4dn-dcic/repli-seq-pipeline for Repli-seq analysis. We performed quality control of the Repli-seq reads using FastQC [216] and removed adapter sequences using Cutadapt [217]. We calculated the RT signal values for each non-overlapping 5Kb bin of human genome. Specifically, first, we mapped the processed sequencing reads to the human genome assemblies of hg38, using BWA [218]. The genome assemblies were downloaded from the UCSC Genome Browser [133, 134]. We divided the human genome into non-overlapping 5Kb bins. We then calculated Repli-seq read count within a given genomic window (a genomic bin) in early and late phases of RT, respectively, normalized by the total mapped read count in early or late RT phase on the whole genome accordingly. The RT signal in each bin is defined as the base 2 logarithm

ratio of read count per million reads between the early and late phases of RT within this region. Lastly, the $\log 2$ ratio signal is quantile normalized and smoothed using loess-smooth method.

The RT data in mouse we used are from the cas/129 hybrid mESCs [75], which are derived from the cross between M. casteneus (CAST/Ei) and M. musculus (129/sv). SNP calls for CAST/Ei and 129/sv genomes were downloaded from Sanger Institute (ftp://ftp-mouse.sanger.ac.uk/REL-1505-SNPs_Indels/). First we used bowtie2 to map reads to a mouse genome (mm10) that all SNP positions are masked by the ambiguity base 'N'. Then we used SNPsplit [219] to parse the mapped reads to the corresponding allele. Reads that do not overlap with SNPs were discarded. Reads overlapping with SNPs in the two genomes were sorted and processed in each allele individually. We followed the same scripts (https://github.com/4dn-dcic/repli-seq-pipeline) used in human RT analysis. The mouse RT signals in each allele are calculated in 5Kb resolution.

4.7 RT prediction in mESCs and human cell lines

We performed cross-chromosome model training and prediction. We use two subsets of chromosomes which are non-overlapping for training. We use the chromosomes in each subset for training and make predictions with the trained model on the left-out chromosomes which are not included in the training subset. For each cell type in human, we used chromosome 1 to chromosome 16 as alternating training and test data. We used chromosome 17 to chromosome 22 as reserved test data. We only included autosomes in the analysis. We choose the context of +/-50 bins (bin size: 5Kb) for each genomic locus to form a context window centered at this locus. Therefore, the context size of each locus is 505Kb (101 bins included). Specifically, for each cell type, first, we used the odd-numbered chromosomes of chromosome 1 to chromosome 15 (chromosome 1, chromosome 3, $\cdots$, chromosome 15) for model training, and predicted RT signals on the left-out 14 autosomes with the trained model. Next, we used the even-number chromosomes of chromosome 2

to chromosome 16 (chromosome 2, chromosome 4, $\cdots$, chromosome 16) for training, and predicted RT signals on the left-out autosomes. We split the samples on the chromosomes used for training into training data and validation data, with the ratio 9:1. Specifically, for each chromosome in the training set, we select the first 90% of genomic bins for training based on the genomic coordinates, and use the last 10% of genomic bins as validation data. With the training and prediction scheme as described, we obtained RT signal predictions on all the autosomes. For RT prediction in mESCs, for each of the two alleles (mutant allele and WT allele), we first used the odd-numbered chromosomes of chromosome 1 to chromosome 15 as the training data. We used the trained model to predict RT signals and estimate locus-wise importance scores on the left-out 11 autosomes, including chromosome 16 which differs in the presence of the three identified ERCEs (a,b,c) between the two alleles, and chromosome 8, which differs in the presence of the two identified ERCEs (d,e) between the two alleles. Next, we used the even-numbered chromosomes of chromosome 2 to chromosome 16 for model training, and performed RT prediction and importance score estimation on the left-out autosomes. We combine the RT predictions and importance score estimates on each chromosome for genome-wide performance evaluation and analysis.

4.8 Processing estimated importance scores for predictive element identification

We generated a set of annotations of the potentially predictive genomic loci based on the locus-wise importance scores estimated by our method on the Repli-Seq data of each cell type. We normalized the estimated importance scores on each chromosome to the scale of [0,1], based on the rank of the score of each locus among all the loci on the corresponding chromosome. The original score not below $x\%$ of all the scores is transformed to be $x \in [0, 1]$. We then filtered the genomic loci to select a subset of loci as potentially predictive genomic loci. First, we performed peak calling on the importance scores along each chromosome using the peak finding function in the scikit-learn library [144], with con-

straints on both the peak strength and the minimal distance between two adjacent peaks, to suppress local non-maximal scores. Specifically , we require the normalized estimated score of a peak to be above 0.90 and the distance between two peaks to be at least 5 bins (25Kb). Let S_1 denote the set of loci with identified local peaks of estimated scores. Second, we select genomic loci with the normalized score above 0.975, denoted as set S_2. The union of S_1 and S_2 is the selected subset of predictive genomic loci. We merged selected loci that are not more than 25Kb (5 bins) apart into one element. The remaining isolated loci each represents an element. We divide the selected elements into early RT group (denoted by S_E) and late RT group (denoted by S_L) based on the early or late RT domains they reside in, respectively. We focus on analyzing elements in S_E.

4.9 Predicting RT signals using sequence features without context

We use three methods as baseline methods to predict RT signals from DNA sequence features without modeling the long-range spatial dependencies across genomic loci (section 4.9). The second method is XGBboost regression (XGBR) [220]. XGBboost regression utilizes Gradient Tree Boosting [221] to fit regression models for continuous value variables. The second method is random forest (RF), which performs predictive model learning based on an ensemble of decision trees[cite]. The third method is linear regression (noted as LR), which performs linear mapping from the sequence feature of each locus to the RT signal. The fourth method is a locus-wise deep neural network model (noted as DNN-local). DNN-Local is similar to the predictor of CONCERT, except that the convolution layer is replaced by a dense layer, the BiLSTM layer is replaced by two connected dense layers each with the same number of hidden units as the single LSTM layer of the BiLSTM, and there is no self-attention layer. The structure of DNN-Local is as follows. The main components of the predictor module are as follows.

1. Three sequentially connected dense layers, each followed by batch normalization and

ReLU activation. Number of hidden units: 50.

2. Two dense layers, each followed by batch normalization and ReLU activation. Number of hidden units: 32.
3. Dense layer, followed by batch normalization and sigmoid or tanh activation to predict the genomic signals. Number of hidden units: 1.

The four methods use feature representation of each locus to predict the corresponding genomic signal, without modeling spatial dependencies across a series of genomic loci in the larger-scale genomic region.

4.10 Quantification and analysis methods

4.10.1 RT prediction performance evaluation

The evaluation metrics we use for comparing the predicted RT signals with the ground truth RT signals are the Pearson correlation coefficient (PCC, or Pearson's r) [222], the Spearman's rank correlation coefficient (Spearman's ρ) [223], the explained variance [224], and the coefficient of determination (R^2 sore) [225].

Let $\hat{Y}$, Y denote the predicted RT signals and the ground truth RT signals, respectively. For predictions on a sample of N genomic loci, let $\hat{y}$, y denote the predicted signals and the ground truth signals, respectively. we have $\hat{y} = \{\hat{y}_1, \cdots, \hat{y}_N\}$, $y = \{y_1, \cdots, y_N\}$.

The Pearson correlation coefficient is defined as

$$\rho_{Y,\hat{Y}} = \frac{\mathbb{E}(Y\hat{Y}) - \mathbb{E}(Y)\mathbb{E}(\hat{Y})}{\sqrt{\mathrm{Var}(Y)}\sqrt{\mathrm{Var}(Y)}}, \tag{4.28}$$

where Var represents the variance of a random variable.

The Spearman's rank correlation coefficient (denoted as r_s) is defined as the Pearson correlation coefficient between two rank variables. Let rg_Y, $rg_{\hat{Y}}$ be the rank variables converted from Y, $\hat{Y}$, respectively. For a sample $y = \{y_1, \cdots, y_N\}$, we have $rg_Y = \{rg_{y_1}, \cdots, rg_{y_i}\}$, where rg_{y_i} is an integer representing the rank of y_i in the set of $\{y_1, \cdots, y_N\}$ based on the predefined order of the values, $i = 1, \cdots, N$. The same

definition applies to $rg_{\hat{y}_i}$. The Spearman's rank correlation coefficient is defined as

$$r_s(Y, \hat{Y}) = \rho_{rg_Y, rg_{\hat{Y}}}, \tag{4.29}$$

where ρ represents the Pearson correlation coefficient between two random variables.

The Pearson correlation coefficient and the Spearman's rank correlation coefficient are both in the range of [-1,1], with the higher value represents the better prediction performance in our study.

The explained variance is defined as

$$\text{explained\_variance}(Y, \hat{Y}) = 1 - \frac{\text{Var}(Y - \hat{Y})}{\text{Var}(Y)}, \tag{4.30}$$

The R^2 score is estimated as

$$R^2(y, \hat{y}) = 1 - \frac{\sum_i^N (y_i - \hat{y}_i)^2}{\sum_i^N (y_i - \bar{y}))}, \tag{4.31}$$

where $\bar{y} = \frac{1}{N} \sum_i^N y_i$. Both of the explained variance and the R^2 score are in the range of $(-\infty, 1]$, with the higher value representing the better prediction performance.

For RT prediction, genomic loci with original RT signal above the threshold zero is labeled with 1 (early replication) and otherwise labeled with 0 (late replication). For evaluation of early/late RT classification performance, we use the metrics AUROC (area under receiver operating characteristic curve) and AUPR (area under precision-recall curve). We normalize the predicted RT signals to the scale of [0,1], with the higher value corresponding to be earlier replicating. By setting varying thresholds to determine the predicted signal to be early or late RT, we calculate AUROC and AUPR accordingly.

4.10.2 Estimated importance score comparison between ERCE and non-ERCE regions

We compared the estimated importance scores between the ERCEs and the non-ERCE genomic regions in the mESC 129/CAST hybrid cells. We only included predicted ERCEs on the autosomes in analysis. For each ERCE, we randomly sample 200 genomic regions

of the same length as the ERCE and without overlapping with any ERCE from the genome. An ERCE ranges from 50Kb to 200Kb in length, containing 10-40 genomic bins. For each ERCE or each randomly sampled element, we calculate both the maximal value and the mean value of the estimated importance scores of genomic bins in this ERCE or sampled region. We merged the maximal values (or mean values) of estimated importance scores associated with each ERCE. We then merged the maximal values (or mean values) of estimated importance scores associated with each randomly sampled region, to form the background importance score distribution. We compared the distribution of estimated maximal-value (or mean-value) importance scores of ERCE in comparison with the background distribution. We observed that ERCEs have higher maximal-value (or mean-value) per-region importance scores than the randomly sampled genomic regions. To map mESC ERCEs to human genome, we use liftOver [138] with the minimal remapping ratio of 0.5 (mapping between mm10 and hg38) and require the regions mapped to the human genome can be mapped back to the original ERCE, with the region re-mapped to the mouse genome overlapping with the original ERCE.

4.10.3 Evaluating estimated importance scores of genomic loci with *cis*-regulatory elements and TF binding sites

To evaluate estimated importance scores of genomic loci with cCREs in open chromatin regions, we downloaded peak region annotations of DNase-seq data in each of the five cell lines from the ENCODE Project database [151]. We extended each peak by +/-250bp. Among all the genomic loci that are overlapping with the extended DNase-seq peak regions (noted as open chromatin regions), we compared the distributions of estimated importance scores between the loci that contains at least one cCRE and the loci that does not overlap with a cCRE. We use the importance score distribution of the non-cCRE genomic loci as the background distribution. To evaluate estimated imprudence scores in genomic loci with binding sites of specific TFs or proteins (e.g., EP300,POLR2A), we used the peak regions identified from the ChIP-seq data of the corresponding TFs/proteins as the binding site an-

notations, which are sourced from the ENCODE project database [151]. For each specific type of TF/protein, we identified the genomic loci with binding sites of the corresponding TF/protein. We compared the distributions of estimated importance scores between the genomic loci with the binding sites and the background genomic loci in the open chromatin regions without binding sites of the corresponding TF/protein. We performed both Kolmogorov–Smirnov test [226] and Mann–Whitney U test [227] between the two distributions of the estimated scores in the genomic loci with the features and the background loci to identify TFs with enrichment in genomic loci of higher estimated importance scores.

4.11 Results

We applied CONCERT to predict RT signals and identify predictive sequence elements in mouse embryonic stem cells (mESCs) and in 9 human cell lines GM12878, H1-hESC, H9-hESC, HCT116, HEK293, K562, IMR90, RPE-hTERT, and U2OS. We only included autosomal chromosomes in the analysis. We collected genome-wide RT maps based on Repli-seq data [38] in each cell line (data processing procedures described in section 4.6). Two-fraction tepli-seq data in human cell types are from 4DN data portal [228]. Repli-seq data in mESCs are sourced from GEO:GSE114139 [75]. We used genome assembly hg38 for human and genome assembly mm10 for mouse. We divided the genomes into consecutive non-overlapping 5Kb bins (genomic loci). The processed RT signal at each locus is a continuous value, with the higher value representing the earlier replication.

4.11.1 Predicting DNA RT profiles in multiple human cell types

We performed cross-chromosome RT prediction in the 9 human cell types by using two non-overlapping sets of training chromosomes to perform two-fold modeling training and prediction (section 4.7), obtaining genome-wide RT predictions and importance score estimates. For performance evaluation, we calculated the Pearson correlation coefficient (PCC, or Pearson's r) [222], the Spearman's rank correlation coefficient (Spearman's ρ) [223], the

explained variance [224], and the coefficient of determination (R^2 sore) [225] between the predicted and ground truth RT signals. We included four methods XGBoost Regression (XGBR) [220], Random Forest (RF) [229], Linear Regression (LR) [230], and the local-sequence focused DNN model (DNN-Local) (section 4.9) for performance comparison.

We observed that our method reaches high prediction accuracy in most of the cell types, consistently outperforming the compared methods in each cell type with respect to different evaluation metrics (Figure 4.3A, Figure 4.4, Table 4.1). With our method, the median PCC and median Spearman's ρ between predicted and real RT signals in 9 cell types are 0.81 and 0.80, respectively. Our method exhibits evident performance increases of 0.16-0.18, 0.17-0.19, 0.16-0.30, 0.19-0.29 in PCC, Spearman's ρ, explained variance and R^2 score over the compared methods in different cell types. We found that the performance of CONCERT varies across cell types, with relatively high prediction accuracy in HCT116 and HEK293 (above 0.85 in both PCC and Spearman's ρ) while lower accuracy in IMR-90 and RPE-hTERT, suggesting that there may exist cell-type specific dependencies between DNA sequences and RT patterns. The performance variation is similarly observed on the compared methods, The examples of predicted RT signals in comparison with real signals on chromosome 14 are visualized in Figure 4.3B. Notably, our method has the capability of capturing not only the similar RT patterns shared by different cell types, but also cell type-specific patterns.

Furthermore, we evaluated the performance of CONCERT using different choices of context sizes (Figure 4.3C, Figure 4.6). For each cell type, we observed that the prediction accuracy increases as the context size expands, indicating that long-range contexts are informative for RT prediction. The performance increase slows down as the context size exceeds +/-150Kb (30 bins). We also analyzed the performance from different input feature combinations (Table 4.4). We found that prediction performance with motif features as additional input (Feature II) is similar to the performance without motif features as input (Feature I) (Table 4.4). We observed that using GC content alone is not sufficient for RT prediction (Table 4.4).

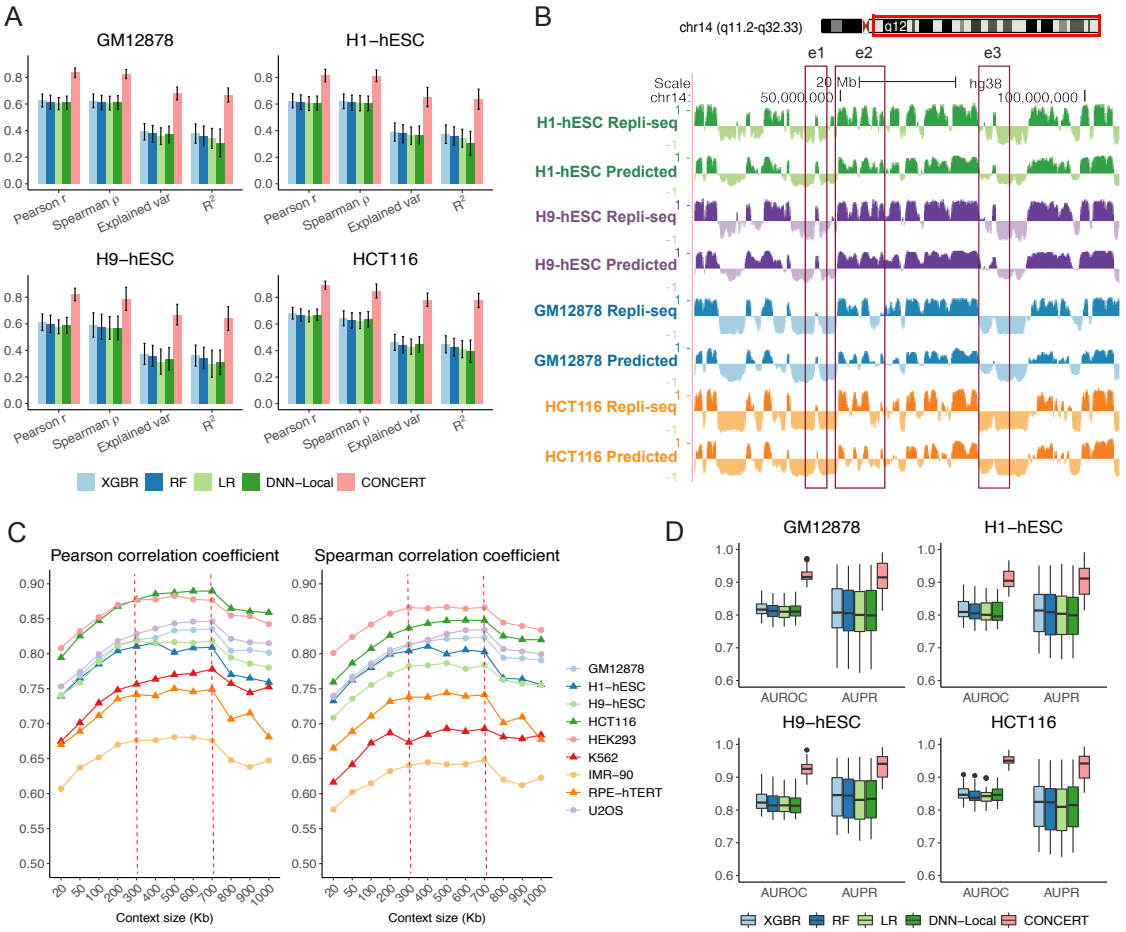

Figure 4.3: Performance evaluation of RT prediction and sequence importance score estimation in human cell lines. **(A)** Performance evaluation and comparison of RT prediction using different methods in cell lines GM12878, H1-hESC, H9-hESC, and HCT116. The error bar shows the standard deviation of performance across 22 autosomes in each cell line. Performance evaluation in the nine human cell lines are shown in Figure 4.4. Performance evaluation on 22 autosomes in H1-hESC cell line are shown in Table 4.2. **(B)** RT predictions on the long arm of chromosome 14 in cell lines GM12878, H1-hESC, H9-hESC, and HCT116. For each cell type, the first track shows the real Repli-seq data, and the second track shows the predicted RT profile. Three example regions with RT signal variations across cell types are marked with red color boxes (noted as $e1$, $e2$, and $e3$). The $e2$ region also contains the example region shown in Figure 4.7l. **(C)** The change of RT prediction performance measured by Pearson correlation coefficient and Spearman's rank correlation coefficient with respect to the change of the context size surrounding each genomic locus in the nine human cell lines. The change of RT prediction performance measured by explained variance and R^2 score with respect to the change of the context size in the nine human cell lines are shown in Figure 4.6. **(D)** RT classification performance across chromosomes measured by AUROC and AUPR in cell lines GM12878, H1-hESC, H9-hESC, and HCT116. Performance evaluation in the nine human cell lines are shown in Figure 4.5.

Additionally, we evaluated the propose method on the classification performance of predicting early or late replication. Each genomic bin is assigned a binary label corresponding

to early or late replication based on the RT signal at this locus. Genomic bins with original RT signals above the cutoff of 0 are labeled as early replication, and otherwise labeled as late replication. We normalized the predicted RT signals to the scale of [0,1], with higher value corresponding to earlier replication. We calculate the AUROC (area under receiver operating characteristic curve) and AUPR (area under precision-recall curve) of the predictions in comparison with the ground truth early/late labels across 22 autosomes. Our method reaches 0.81-0.95 in AUROC, 0.72-0.96 in AUPR in 9 human cell types, outperforming the compared methods (Figure 4.3D, Figure 4.5). In a recent study [231], the method GEEK (Gene Expression Embedding Framework) provided feature embeddings of biological objects from heterogeneous networks based on integration of multiple sources of network information. The learned embeddings were demonstrated useful for modeling gene expression levels, and were also applied to early/late RT classification in four human cell lines GM12878, K562, HUVEC, and NHEK, achieving cross-chromosome median AUROC of 0.8955 and below 0.85 in cell lines GM12878 and K562, respectively. The accuracy GEEK achieved on RT classification through integrating a variety of informative biological network data provides a reference to assess the RT prediction performance of our method. Our method reached median AUROC of 0.9215 and 0.9253 across chromosomes in GM12878 and K562, respectively, using only sequence features as input.

Together, these observations demonstrate the effectiveness of CONCERT in learning spatial dependencies across genomic loci with long-range context information to predict RT profiles from DNA sequences.

4.11.2 Predicting DNA RT profiles in mouse ESCs and evaluation with ERCEs

Next, we applied the developed method to RT signal prediction in the mouse embryonic stem cells (mESCs), and evaluated the predicted RT-related importance of genomic loci using annotations from the recent discovery of early replication control elements (ERCEs) [75]. In the recent study [75], multiple *cis*-regulatory elements with crucial roles in controlling early replication, named early replication control elements, were identified in mESCs

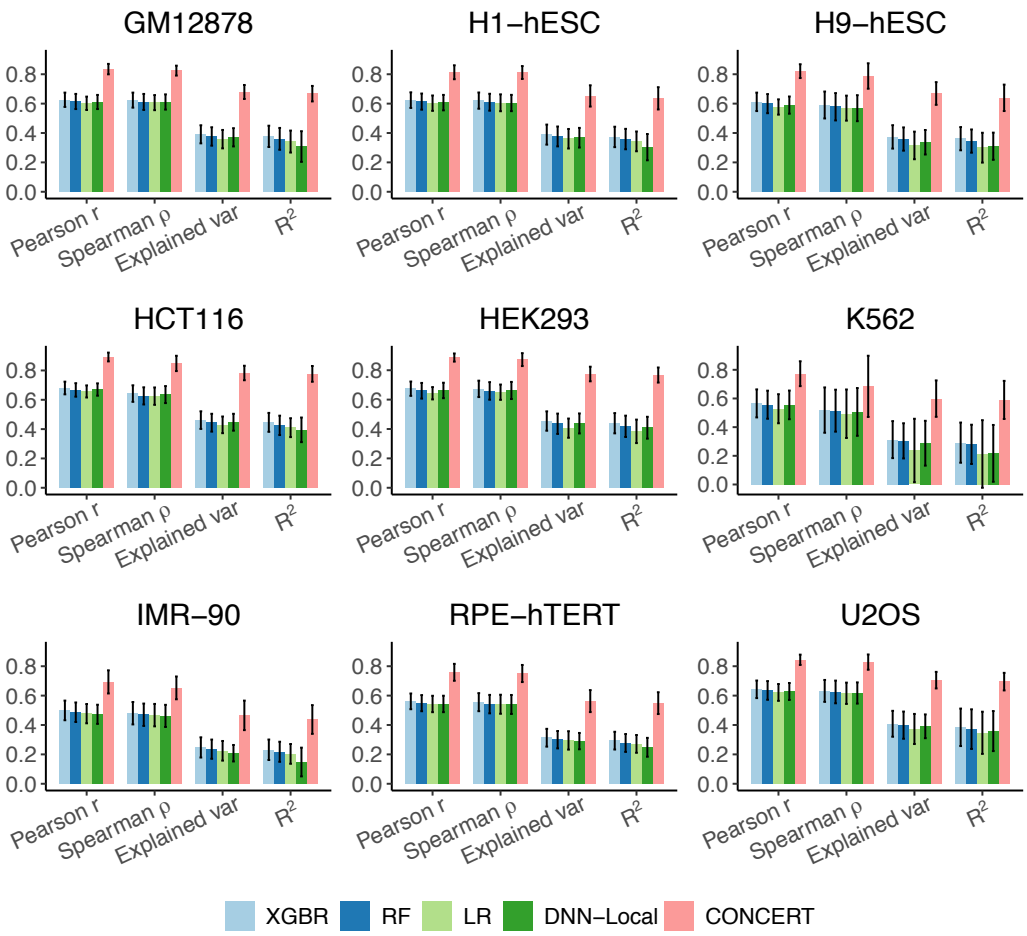

Figure 4.4: RT prediction performance of CONCERT in nine human cell lines. The evaluation metrics are Pearson correlation coefficient (Pearson's r, noted as Pearson r in the axis label), Spearman's rank correlation coefficient (Spearman's ρ, noted as Spearman ρ in the axis label), explained variance, and R^2 score. The error bar shows the standard deviation of performance across 22 autosomes in each cell line.

through a series of CRISPR-mediated deletion or inversion experiments. Specifically, CRISPR-mediated deletions of segments of varied lengths and combinations at a pluripotency-associated replication domain, the murine *Dppa2*/4 domain (containing three active genes *Dppa4*, *Dppa2*,and *Morc1/Morc*) on chromosome 16 in mESCs, led to the discovery of three cooperative ERCEs within the domain. Targeted deletion of the three elements

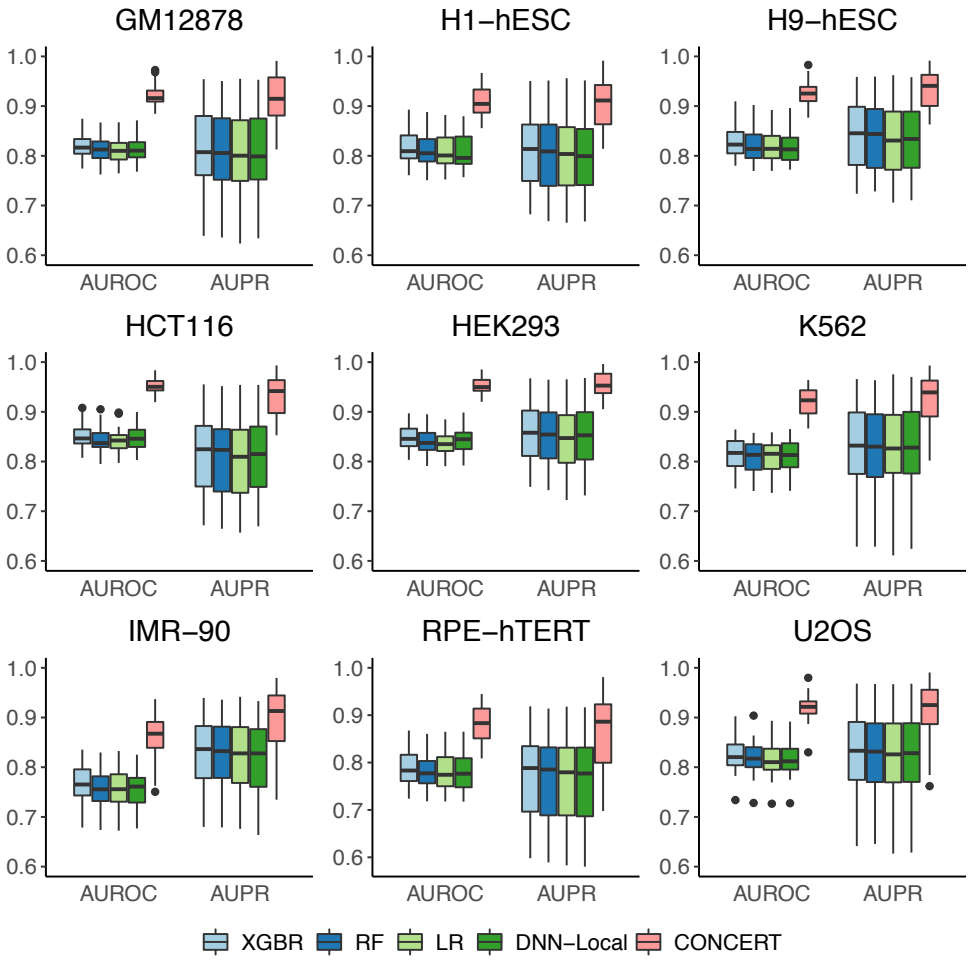

Figure 4.5: RT classification performance of CONCERT in nine human cell lines. The boxplots show the RT classification performance measured by AUROC and AUPR across 22 autosomes in nine human cell lines.

caused a switch from early to late replication of the domain. The three ERCEs (denoted as sites a, b, c, Figure 4.7B) are featured with the enrichment of active epigenetic marks that are typically representative of enhancers, including DNase1 hypersensitivity (HS), P300, H3K27ac, H3K4me1, H3K4me3, binding sites of MED1, and binding sites of major pluripotency transcription factors (TF) POU5F1 (also known as OCT4), SOX2, and NANOG. One ERCE (site a) also corresponds to a super-enhancer in mESCs [232, 233].

Furthermore, 1835 ERCEs were predicted genome-wide based on the presence of spe-

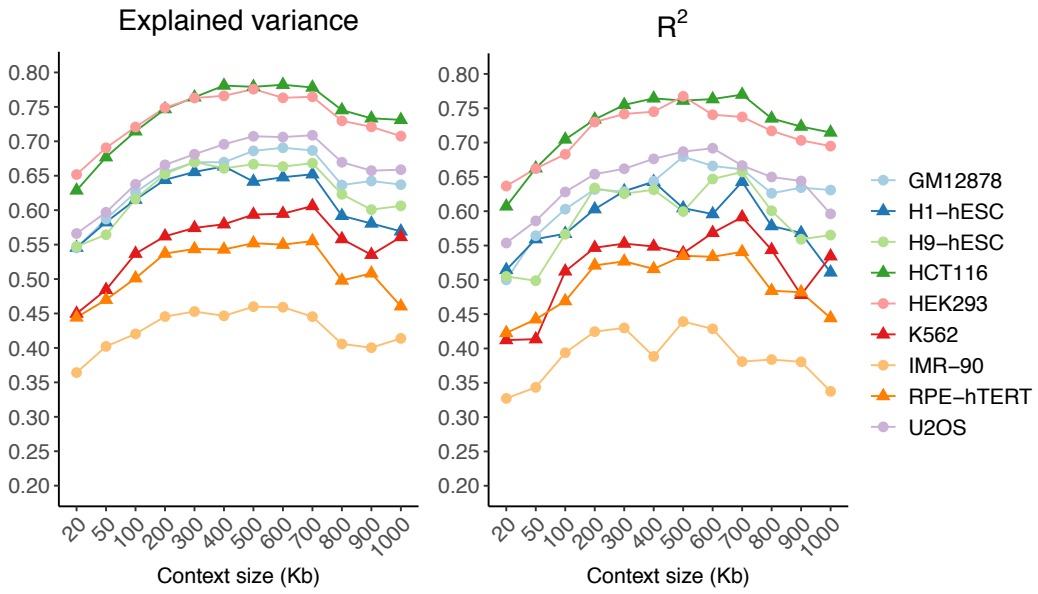

Figure 4.6: The change of RT prediction performance evaluated by explained variance and R^2 score with respect to the change of context size surrounding each genomic locus in nine human cell lines.

cific properties shared with the three experimentally identified ERCEs in the *Dppa2/4* domain on chromosome 16 in mESCs [75]. The predicted ERCEs are located at the sites which display significant CTCF-independent chromatin interactions, overlap with P300 and OSN (OCT4,SOX2,and NANOG) binding sites, and reside in inter-LAD (lamina-associating domain) regions. CRISPR-mediated deletions performed at the *Zfp42/Rex1* domain (containing three active genes *Trim1*, *Trim2*, and *Zfp42/Rex1*) on chromosome 8 further validated the existence of two ERCEs (denoted as sites d and e, shown in Figure 4.7B) predicted within the domain. Targeted deletions of both ERCEs caused a switch from early to late replication of the domain.

We sought to explore whether ERCEs can be identified by our method computationally using sequence features only. We used the same mESC RT data in CAST/129 hybrid cells from [75]. There are two alleles, named mutant allele, with targeted ERCE deletions, and WT (wild type) allele, which share similar RT profiles except for the ERCE deleted regions. The similar feature extraction and model training procedures applied to RT prediction in human cell types were applied to mESCs (section 4.7). We observed that our method reaches high RT prediction accuracy in mESCs, with PCCs 0.90, 0.91, Spearman's ρ 0.90, 0.89, explained variances 0.80, 0.79, and R^2 scores 0.80, 0.79 on the mutant and WT alleles, respectively (Figure 4.7A), consistently outperforming other methods. Next, we identified predictive elements based on the estimated importance scores. We normalized the estimated scores to the scale [0,1], with the original score not below $x\%$ of all the scores on the corresponding chromosome transformed to $x/100$. We then filtered the genomic loci to select a subset of loci with distinguishable prominence by detecting local peaks of scores (section 4.8). Selected loci within specific distance ($\leq$5 bins apart) are merged into one element. Each selected predictive genomic locus or merged element is considered as a predictive element. Here we focus on analyzing the identified predictive elements in early RT domains.

Our method identified 7 predictive elements on the *Dppa*2/4 domain on chromosome 16 in mESCs, of which 3 elements directly overlap with ERCEs a, b, c and 2 elements overlap with b and c with +/-5Kb (1bin) extensions (Figure 4.8). Our method identified 3 predictive elements on the *Zfp42/Rex1* domain on chromosome 8 (Figure 4.7B), of which two elements directly overlap with ERCEs d and e. Each of the five ERCEs that were experimentally validated by CRISPR-mediated deletions in [75] is matched by one or more predictive elements detected by our method. CONCERT also predicted local switch from early RT to relatively later RT with each of the two sets of ERCEs (a,b,c and d,e) deleted, consistent with the experiment results in [75] (Figure 4.8, Figure 4.7B).

Furthermore, we investigated genome-wide matches between the identified predictive elements and the predicted ERCEs. We only included predicted ERCEs on the autosomes in

analysis. There are 1798 ERCEs predicted on autosomes (50Kb$\sim$200Kb in length) by [75]. We found that 1419 of the 1798 predicted ERCEs with +/-2 bin extensions (78.92%, empirical p-value$<$1e-02, Figure 4.7C) overlap with one or more predictive elements identified by CONCERT. There are 1092 ERCEs directly matched by identified predictive elements without extensions. The predictive elements identified by our method also have shorter distance to the nearest ERCEs and higher overlapping percentage with the matched ERCEs in comparison with the background distribution where the elements are randomly shuffled (Figure 4.7D-E, p-value$<$2.2e-16 with Kolmogorov-Smirnov test, or K-S test [226]). We also evaluated the estimated importance scores of ERCEs in comparison with randomly sampled non-ERCE sequences. We calculate both the maximal and mean values of the estimated importance scores of genomic bins in each ERCE or randomly sampled region. We observed that ERCEs on average show much higher region-wise maximal scores and mean scores, both in early RT domains and genome-wide (Figure 4.7F, p-value$<$2.2e-16 with K-S test). These observations confirm the effectiveness of our method in predicting sequence elements that may have regulatory functions for RT.

Evaluating estimated importance scores with mESC ERCEs mapped to human genome

The ERCEs were only predicted in mESCs [75]. We mapped the predicted mouse ERCEs to the human genome, and evaluated the importance scores estimated by our method of the mapped ERCE regions in H1-hESC cell line. We used liftOver [138] to map each predicted ERCE on mouse autosomes to human autosomes with reciprocal mapping (section 4.10.2), obtaining 446 mapped ERCEs. We then compared the distribution of region-wise maximal (or mean) estimated importance scores of the mapped ERCEs with the background distribution based on randomly sampled regions in H1-hESC cell line, using the same comparison approach for mouse ERCEs as described above. We observed that the ERCEs mapped to human genome exhibit higher estimated importance scores in distribution as compared with the background regions in both early RT domains and genome-wide (p-values$<$1e-22, Fig-

ure 4.7G). The mapped ERCEs with co-binding sites of two or more pluripotency TFs also show higher importance scores in distribution than the background regions (p-value$<$1e-10 genome-wide, p-value$<$1e-05 in early RT domains, Figure 4.7H). Furthermore, we found that 71.08% of mapped ERCEs with +/-2 bin extensions can be matched by predictive elements identified by our method (empirical p-value$<$1e-02).

One example of ERCE mapped to the human genome that can be identified by our method is shown in Figure. 4.7I. This mapped ERCE is located in a cell type-specific early RT domain on chromosome 14 of human genome. The corresponding local domain shows early RT in H1-hESC and H9-hESC, and late RT in the other cell types including GM12878 and HCT116. The mapped ERCE contains a co-binding site of the pluripotency TFs POU5F1, SOX2, and NANOG in H1-hESC. Our method identifies predictive elements that overlap with this mapped ERCE in the corresponding region in H1-hESC and H9-hESC. Specifically, the identified predictive elements are co-located with the pluripotency TF co-binding site, and found to be cell type-specific, as they are not predicted in the other cell types such as GM12878 and HCT116. Also, our method predicts early RT only in H1-hESC and H9-hESC on the corresponding domain, and predicts later RT for the other cell types.

We further observed that the OTX2 gene, which encodes the transcription factor OTX2, is co-located with the pluripotency TF co-binding site and the predictive elements identified at this location (Figure. 4.7I). Based on gene expression profiles of H1-hESC, GM12878, and K562 cell lines that are downloaded from the ENCODE project database [151], we observed that OTX2 gene is specifically expressed in H1-hESC with TPM$>$1 (TPM: transcripts per million, which represents for every one million RNA molecules in the RNA-seq sample, how many are from the specific gene/transcript), while it is not expressed or has low expression in the other cell types GM12878 and K562 (TPM$<$0.01). The OTX2 gene has leading critical roles in the early stages of embryonic development and ESC differentiation, with the encoded protein involved in the development of brain and sensory organs including the eyes [235–238]. Mutations of OTX2 may cause development delay, severe or-

gan dysfunctions and structural abnormalities [239–241]. Furthermore, based on enhancer annotations in different human cell lines that are downloaded from the EnhancerAtlas 2.0 database [242], we observed that there are cell type-specific enhancers at the location of the OTX2 gene in H1-hESC and H9-hESC, while there are no enhancers identified in the same region in GM12878 and HCT116. The enhancers near the promoter region of the OTX2 gene in H1-hESC and H9-hESC are co-located with the pluripotency TF co-binding site and overlapped with the RT predictive loci identified by our method in the corresponding two cell lines. The observations suggest the potential connections underlying the OTX2 gene, the co-bound pluripotency TFs, the regulatory elements, and the ESC-specific early RT pattern. The original ERCE of this mapped ERCE is also overlapped with predictive genomic loci identified by our method in the mESCs. We infer that the mapped ERCE at this analyzed location is highly likely to be an ESC-specific functional ERCE in human.

These observations demonstrate the effectiveness of our method in identifying potential predictive sequence elements for RT modulation in human cell lines.

4.11.3 Repetitive element enrichment in identified predictive genomic loci

Next, we explored potential connections between repetitive elements (REs) and genomic loci of high estimated importance (Figure 4.9A, Figure 4.10). It is known that RT is correlated with certain types of transposable elements (TEs). For example, early RT regions are typically enriched with SINEs [37]. TEs are one type of repetitive elements (REs). However, it is still unclear how important TEs or other types of REs are in modulating global RT program. We retrieved human genome RE annotations from the UCSC Genome Browser [212]. The REs we used for analysis mainly consist of the TEs (including the classes SINE, LINE, LTR, and DNA), the retroposons, and the small RNA molecule sequences. For each human cell type, we classified the genomic loci into $K = 20$ groups based on importance score ranking. Each of the $K = 20$ groups represents an importance level, with the k-th group consists of genomic loci with importance scores ranking between top $5(k-1)\%$ and top $5k\%$ on the corresponding chromosome. We calculated

the fold change of enrichment for each RE family in each group in comparison with the average coverage across all groups (section 4.10.2). First, we restricted the comparison to early RT domains. Only genomic loci with early RT are retained for comparison. We observed that Alu elements of the SINE class are significantly more enriched in genomic loci with high estimated importance (p-value$<$1e-07 for genomic loci with top 15% estimated importance), which is consistently observed across different cell types (Figure 4.9A). For each cell type, the enrichment of Alu across different groups of genomic loci shows strong positive correlation with the corresponding estimated importance level, displaying a clear trend where the enrichment increases as the importance level ascends. The observations show that even within the early RT regions, the different enrichment patterns of Alu elements with respect to the correlation with RT can be distinguished by our method based on estimating sequence importance for RT prediction. In contrast, L1 from the LINE class, and ERVL, ERVK, LTR from the LTR class show negative correlations between enrichment and estimated importance levels. Genome-wide comparison of RE enrichment across estimated sequence importance levels show similar results (Figure 4.10). The observations provide support that Alu elements may be engaged in modulating RT.

Moreover, we observed that scRNA, snRNA, srpRNA, and SVA elements also exhibit strong positive correlations between enrichment and estimated importance of genomic loci (p-value$<$1e-07). The enrichment patterns are similar to the patterns observed on Alu elements and consistent across different cell types, including GM12878. Previous studies on RT comparison in the lymphoblastoid cell type across multiple primate species showed that scRNA, snRNA, and srpRNA have strong positive correlations with early RT profiles that are conserved across species [161]. Here the observations that scRNA, snRNA, and srpRNA also exhibit stronger enrichment in genomic regions with higher estimated importance scores are consistent with the previous studies, indicating the potential roles of scRNA, snRNA, and srpRNA in modulating RT. Overall, the analysis of RE enrichment based on estimated sequence importance suggests the potential of CONCERT in prioritizing experimental characterizations of possible sequence determinants of RT.

4.11.4 Evaluating estimated importance scores with candidate *cis*-regulatory elements and TFs

Next, we evaluated estimated importance scores of genomic loci containing candidate *cis*-Regulatory Elements (cCREs) within the open chromatin regions in human cell types. The cCRE annotations were retrieved from the SCREEN (Search Candidate cis-Regulatory Elements by ENCODE) database [243], which provides registry of cCREs derived from various types of data in the ENCODE Project [151]. There are mainly four groups of cCREs, including active and poised enhancer-like elements (distal and proximal), active promoter-like elements, likely poised elements with DNase and H3K4me3 marks, and CTCF-only elements. The cCREs are labeled with dELS (distal enhancer-like signatures), pELS (proximal ELS), PLS (promoter-like signatures), DNase-H3K4me3, and CTCF-only, respectively. Within the open chromatin regions (identified using DNase-seq data, section 4.10.3), for each type of cCREs, we compared the distributions of the estimated importance between the genomic loci containing the specific type of cCREs and the loci depleted of any cCRE.

We observed that genomic loci with pELS, PLS, or DNase-H3K4me3 exhibit significantly higher estimated importance scores on average than the background non-cCRE loci, consistently observed in cell lines GM12878 and H1-hESC (Figure 4.9B-C, Figure 4.11, p-value$<$1e-04). The observations suggest the potential connections between specific cCREs (pELS, PLS, and DNase-H3K4me3) and RT regulation. In contrast, the estimated score distribution with CTCF-only signatures does not significantly differ from the background distribution, consistent with findings in [75] that CTCF proteins are dispensable for RT.

We also analyzed the estimated importance distribution of genomic loci with different transcription factor (TF) binding motifs (section 4.10.3). We identified a number of TF motifs enriched in loci of higher importance scores in H1-hESC, including TAF1, ATF3, KLF4, EGR1, and BACH1. KLF4 is known to be important for transcription in the pluripotent state, and TAF1 is involved in transcription initiation and enhancers [244]. We further examined the estimated scores of loci with binding sites of these TFs and several other proteins critical for the pluripotency. The TF/protein binding sites are identified from

peaks based on the corresponding ChIP-seq data downloaded from the ENCODE project database [151]. We observed enrichment of these proteins in estimated RT-predictive loci (Figure 4.9D). The analysis suggests that BACH1 may be a potential RT-relevant TF, whose role in cell-cycle control was revealed in [245, 246] but its RT-modulating function has yet to be validated.

The observations suggest the potential functions of specific *cis*-regulatory elements and transcription factors for RT modulation. Our method provides an approach to detect the potential RT-related *cis*-regulatory elements and transcription factors through identification of RT predictive sequence elements.

4.11.5 Identifying RT predictive genomic loci across different cell types

With the cooperation of the selector and predictor modules, the CONCERT model identifies the genomic loci with potential importance in RT signal prediction in different cell lines.

To identify cell type-specific RT predictive loci, first, we identify predictive loci of higher confidence based on predictions from multiple experiments using different model variants. We employed 4 model variants, and generated 18 sets of genome-wide RT predictions and RT predictive loci predictions using the model variants in each of the nine human cell lines. The mode variants mainly differ in the activation function used in the selector module and the loss function in the predictor module (model architecture details are described in section 4.4). Specifically, we use sigmoid, ReLU, or linear function as the activation function applied to the output of the dense layer for importance score estimation in the selector module. Additionally, we tested another activation function tanh. The model variant with tanh as the activation function in the selector module has similar RT prediction performance to the other model variants. The model variants include: model I, with sigmoid activation in the selector module and MSE as loss function in the predictor module;model II, with ReLU activation in the selector and MSE as loss function; model III, with linear activation in the selector and logcosh as loss function; model IV, with tanh activation in the selector and logcosh as loss function. The four model variants show comparable RT

prediction performance based on cross-chromosome RT prediction performance evaluation in each cell line. There are other model variants with alternative combinations of model configurations, which also exhibit similar RT prediction accuracy. Here we use the four variants to identify the predictive loci of higher reliability that can be detected with different model configurations. We then generated 18 sets of predictions with the model variants. The default input features of each genomic bin consist of the K-mer frequency features with dimension reduction transformation and the GC-content based features. We classified the predictions into 5 categories, which are as follows: (i) two sets of predictions by model I in two repeated experiments, with the context window size of 505Kb (flanking region: +/-50bin) and chromosome 1-16 used for two-fold training; (ii) two sets of predictions by model II in two repeated experiments, with the context window size of 505Kb; (iii) two sets of predictions by model III in two repeated experiments, with the context window size of 505Kb (flanking region: +/-50bin) and TF binding motif features as additional input; (iv) six sets of predictions by model III, with the context window size increasing from 505Kb to 1005Kb (flanking region: +/-50bin, +/-60bin, . . . , +/-100bin); (v) six sets of predictions by model IV, with the context window size increasing from 505Kb to 1005Kb.

If a genomic locus can be identified as an RT predictive locus in at least 9 experiments from at least three different categories, without flanking regions used as tolerance for position offset in different sets of predictions, we select the locus to the set of higher-reliability candidate predictive loci in the corresponding cell type. The resulted set of selected loci is generally smaller in size compared to the set of predictive loci identified from a single experiment, but it is more likely to preserve the loci with higher fidelity in RT predictiveness.

To select cell type-specific predictive genomic loci, we use flanking region of +/-1 bin (5Kb) for each predicted predictive locus in each cell type. If two predictive loci that are identified in two different cell types are not more than 1 bin apart, we consider the two loci as probably identified in both cell types and not cell type specific. We then identified sets of cell type-specific predictive loci, of which each locus is only predicted with high confidence in the corresponding cell type and not selected in the other cell types. We also merged the

predictions in H1-hESC and H9-hESC cell lines as the predictions in hESC cells. The number of estimated cell type-specific predictive genomic loci in each cell type is shown in Table 4.5. Next we identified predictive genomic loci that are shared by different cell types. We identified 3228 genomic loci that are selected as predictive loci across different predictions in at least five of the nine human cell types. The numbers of identified predictive loci that are shared by a specific number of cell types are shown in Table 4.6. For each cell type, the number of identified predictive loci in this cell type that are also identified in at least another four cell types is shown in Table 4.7. The numbers of identified predictive loci in H1-hESC or H9-hESC cell line that are also identified in different numbers of other cell types are shown in Table 4.8.

One example of identified cell type-specific predictive genomic loci is shown in Figure 4.3I as described in section 4.11.2, where there are cell type-specific predictive loci identified in a cell type-specific early RT region in H1-hESC and H9-hESC that overlap with a predicted mESC ERCE mapped to the human genome. H1-hESC and H9-hESC both exhibit early RT in the local region. The other cell lines including GM12878 and HCT116 are of late RT in this region, and did not show predictive elements identified at this location. The cross cell type comparison suggests that there exist both sequence elements with conserved and cell type-specific functional importance in modulating RT across different cell lines.

Specifically, we identified 3566 ESC-specific predictive genomic loci in human that are only predicted in H1-hESC or H9-hESC-hESC. We performed Gene Ontology (GO) analysis for the identified human ESC-specific RT predictive genomic loci using the software GREAT [247] (Figure 4.12). We observed that the genes found to be significantly associated with the corresponding identified loci mostly have functions for biological processes that are related to the development of tissues, organs and systems, which are important processes in embryonic development. The observations suggest that the identified hESC-specific predictive genomic loci for RT are potentially in close connections with specific stages of embryonic development, which are related to the functions of the H1-hESC and

H9-hESC cells. The observations also suggest that RT programs in specific cell types are possibly regulated to be in accordance with the functions of the corresponding cell type.

Furthermore, we compared gene expression levels between identified predictive loci and background loci in cell lines H1-hESC, GM12878, and K562, using the gene expression data downloaded from the ENCODE project database [151]. We identified cell type-specific expressed genes in proximity to the predicted RT predictive genomic loci, including cell type-specific predictive loci. For each of the compared cell types H1-hESC, GM12878, and K562, if a gene has TPM$>$1 in this cell type and TPM$<$1 in the other cell types, we consider it as a cell type-specific expressed gene. For the remaining genes with TPM$>$1 in this cell type, if the gene expression level measured by TPM is at least 2 times the gene expression level in each of the other cell types, we consider it as a gene with cell type-specific increased expression.

We observed that both cell type-specific expressed genes and genes with cell type-specific increased expression are more enriched in the identified RT predictive genomic loci than the background loci in each cell type, and also show higher enrichment in the identified cell type-specific predictive loci, both in early RT regions and genome-wide (Figure). We performed GO analysis of the cell type-specific expressed genes or genes with increased expression that are co-located with the identified cell type-specific RT predictive genomic loci in H1-hESC using the software DAVID [146]. Many of the specific genes are found to be associated with biological processes of system, organ, and tissue development, which are partially consistent with the GO analysis we performed before for the RT predictive loci using GREAT and related to the functions of hESCs. Examples of identified biological processes are shown in Table 4.9. The observations suggest the close connections between RT regulation and transcription and cell type-specific functions.

4.12 Discussion

We developed a new method called CONCERT to predict DNA replication timing profiles and identify sequence elements that may modulate the RT programs using DNA sequence features only. The prediction results of applying the model to nine cell lines in human demonstrate that CONCERT reaches high quality prediction performance in the majority of the studied cell types, and outperforms the baseline methods which do not consider the spatial dependencies across genomic loci. With two collaborating modules, the selector and the predictor, CONCERT performs estimation of predictive sequence elements, selective learning of spatial dependencies across genome loci, and modeling the dependencies between underlying DNA sequences and RT signals, utilizing long-range context information. The results reveal that GC content alone is not sufficient to etimated the RT signals from the sequences. There are more refined levels of information encoded in the DNA sequences that are involved in the potential sequence-dependent regulation mechanism of DNA replication. The results additionally show variation in the potential dependencies between DNA sequences and DNA RT signals in different cell types. Additionally, we found a number of genomic loci that are potentially important for predicting RT profile and are shared by different cell lines using the predictions from CONCERT. Our method also reaches high RT prediction accuracy in mESCs. For predictive sequence element identification in mESCs, each of the five early replication control elements (ERCEs) in mESCs that were experimentally validated through CRISPR-mediated deletions in [75] can be identified by our method. Our method provides a generic framework to predict large-scale genomic features and discover the underlying predictive sequence elements that may modulate RT program using solely DNA sequence features.

There are a number of possible improvements for CONCERT. First, the accuracy of the estimated importance scores as measurement of the importance of a genomic locus in predicting RT signals need to be further evaluated. Second, we need further analysis into what sequence features are underlying the identified potentially predictive genomic loci that are

relevant to the regulation mechanism of DNA replication, and how the sequence features function in the regulation mechanism. Third, we will study the possible cooperative functions of identified predictive elements, as revealed in [75], where several ERCEs are found to cooperate as a group to regulate early RT in the local domain. Such analysis would also pave the way for future experimental validation through genome engineering. Furthermore, we can include other types of epigenetic features into the model to reflect cell-type specific characteristics of DNA replication timing regulation, and improve performance of the model in RT prediction and identification of possible RT regulators. Nevertheless, CONCERT has already shown strong promise to help uncover important sequence level properties that potentially modulate the DNA replication timing program.

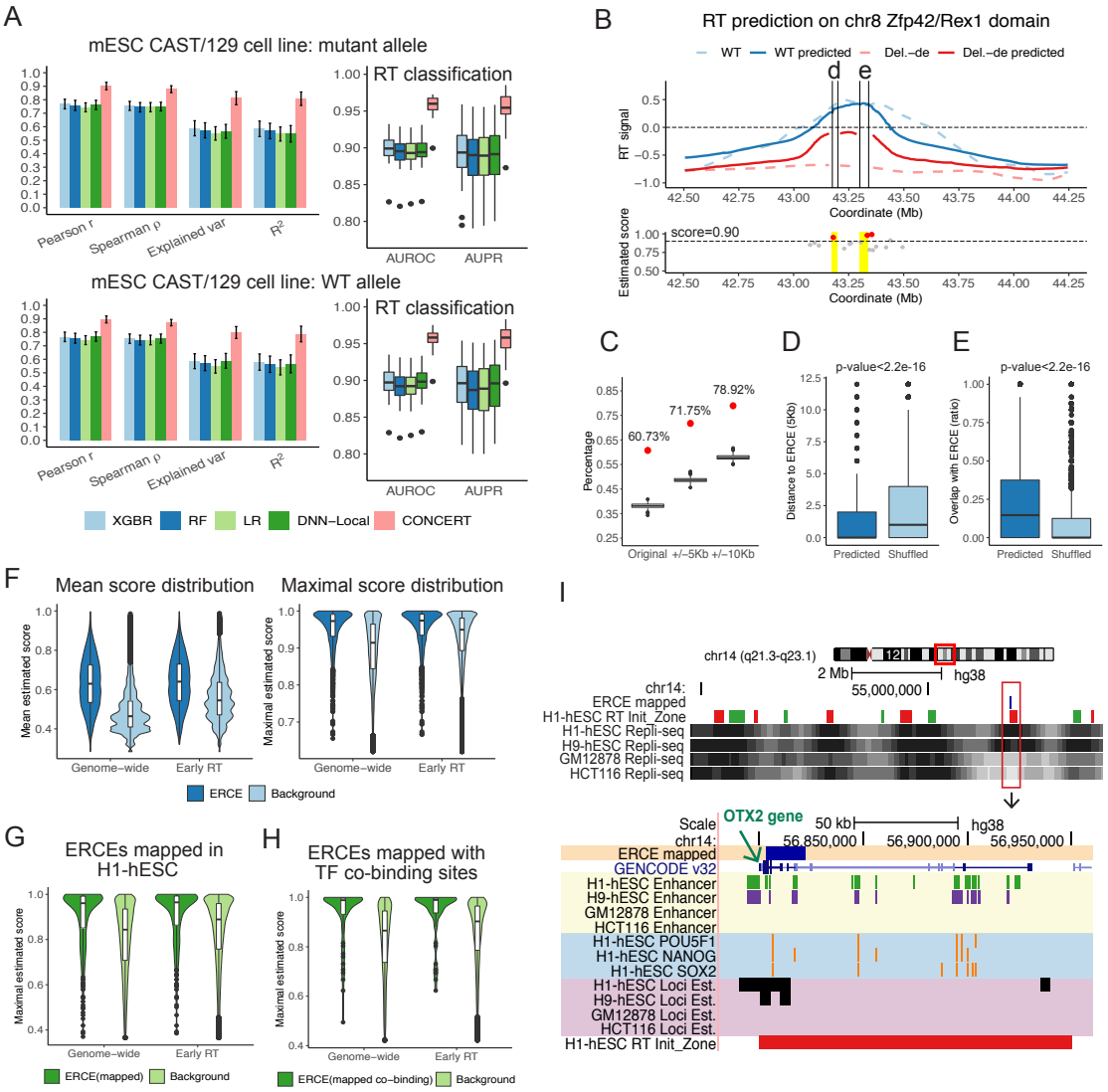

Figure 4.7: Performance evaluation of RT prediction and importance score estimation in mESCs. **(A)** (Left) Evaluation of RT prediction on the mutant and WT alleles of mESC CAST/129 by comparing to different methods. (Right) Early/late RT classification performance on the two alleles of mESC CAST/129 using different methods. **(B)** RT prediction with respect to presence of ERCEs on Zfp42/Rex1 domain on chr8 in mESCs and identified genomic loci with high estimated importance scores in the domain. Normalized estimated importance scores below 0.5 are filtered. **(C)** Percentage of the predicted ERCEs that overlap with the identified predictive sequence elements, with varying extension sizes of the original ERCEs. **(D)** Distribution of the distance between identified predictive elements and the nearest ERCEs, in comparison with background distribution where the predicted sequence importance scores are shuffled. **(E)** Distribution of overlapping ratio between the identified predictive sequence elements and the matched ERCEs. **(F)** Distribution of the estimated importance scores of ERCE regions in comparison with background distribution in non-ERCE regions. **(G)** Distribution of the estimated importance scores of ERCE regions mapped to human genome in comparison with the background distribution in H1-hESC. **(H)** Importance score distribution of ERCE regions mapped to human genome with co-binding sites of POU5F1, SOX2, and NANOG in H1-hESC. **(I)** Example of RT prediction and predictive element identification on chr14:54,295,000-58,750,000. Our method identified cell type-specific predictive elements in cell lines H1-hESC and H9-hESC, which overlap with a predicted mESC ERCE [75] mapped to the human genome. The predictive elements overlap with a co-binding site of POU5F1, SOX2, and NANOG in cell line H1-hESC. Loci Est. represents estimated predictive elements. H1-hESC RT initiation zone annotations are from [234].

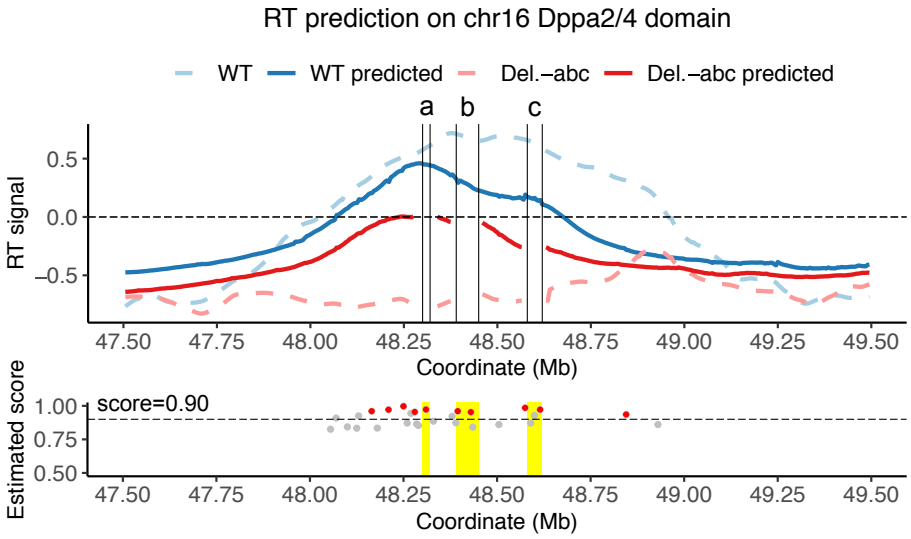

Figure 4.8: RT prediction on the Dppa2/4 domain on chromosome 16 in mESCs [75]. The first sub-figure shows the predicted signals on WT allele (without deletions) and the mutant allele (with ERCE a,b,c deleted). The second sub-figure shows the normalized estimated importance scores and identified genomic loci. Normalized estimated scores below 0.50 are filtered. Identified predictive genomic loci are marked with red color. The yellow rectangles show the positions of ERCE a,b,c.

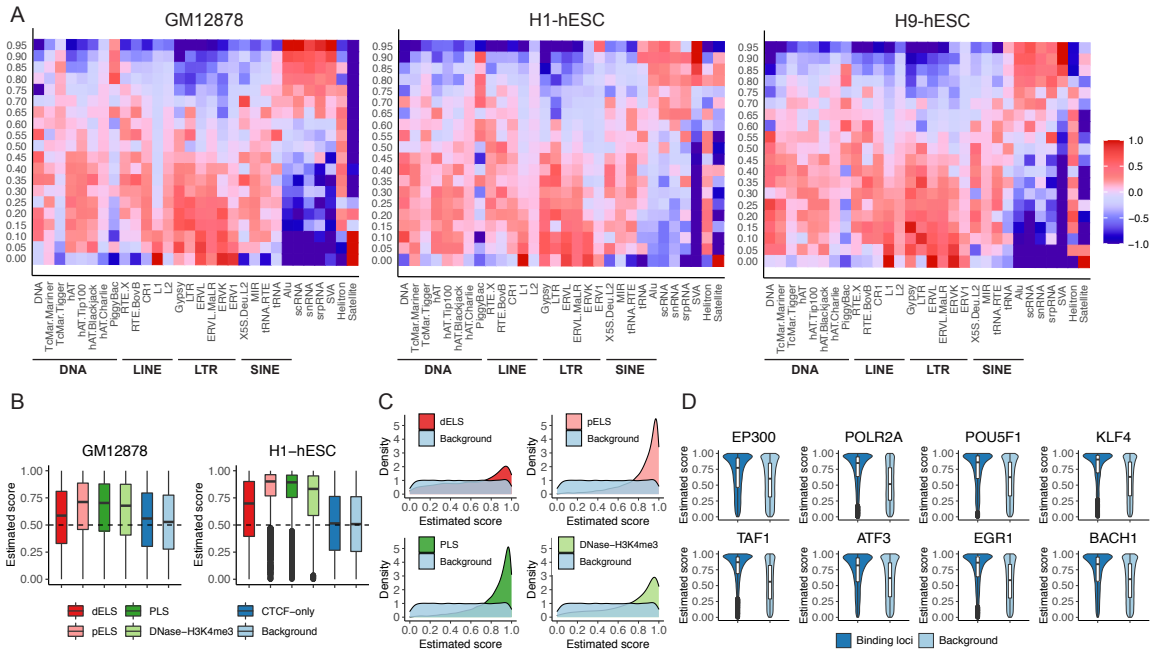

Figure 4.9: Comparisons with repetitive elements (REs) and transcription factor (TF) binding sites. **(A)** RE enrichment in genomic loci at different sequence importance levels in GM12878, H1-hESC, and H9-hESC in early RT regions. **(B)** Distribution of importance scores in genomic loci with candidate CREs in open chromatins in GM12878 and H1-hESC. **(C)** Density distribution of importance scores in genomic loci with dELS, pELS, PLS, and DNase-H3K4me3 in the open chromatin regions in H1-hESC, in comparison with non-CRE background distribution. **(D)** Examples of TFs or proteins showing enrichment in genomic loci of higher importance scores in open chromatin regions in H1-hESC.

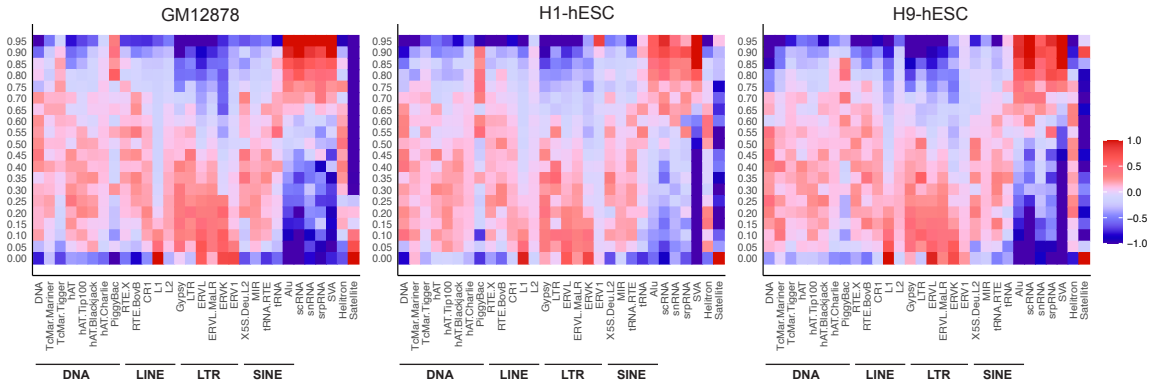

Figure 4.10: RE enrichment fold change of genomic loci at different estimated sequence importance levels genome-wide in cell lines GM12878, H1-hESC, and H9-hESC.

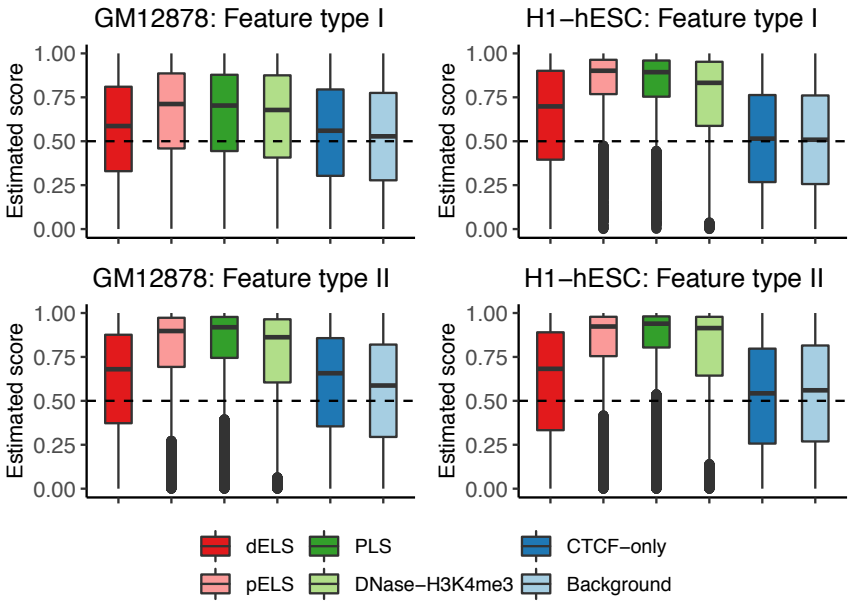

Figure 4.11: Distribution of estimated importance scores in genomic loci with candidate CREs in the open chromatin regions in cell lines GM12878 and H1-hESC with respect to different types of input features used for prediction. Feature type I: input feature consists of K-mer frequency features and GC-based features. Feature type II: input feature consists of K-mer frequency features, GC-based features, and TF binding motif frequency features.

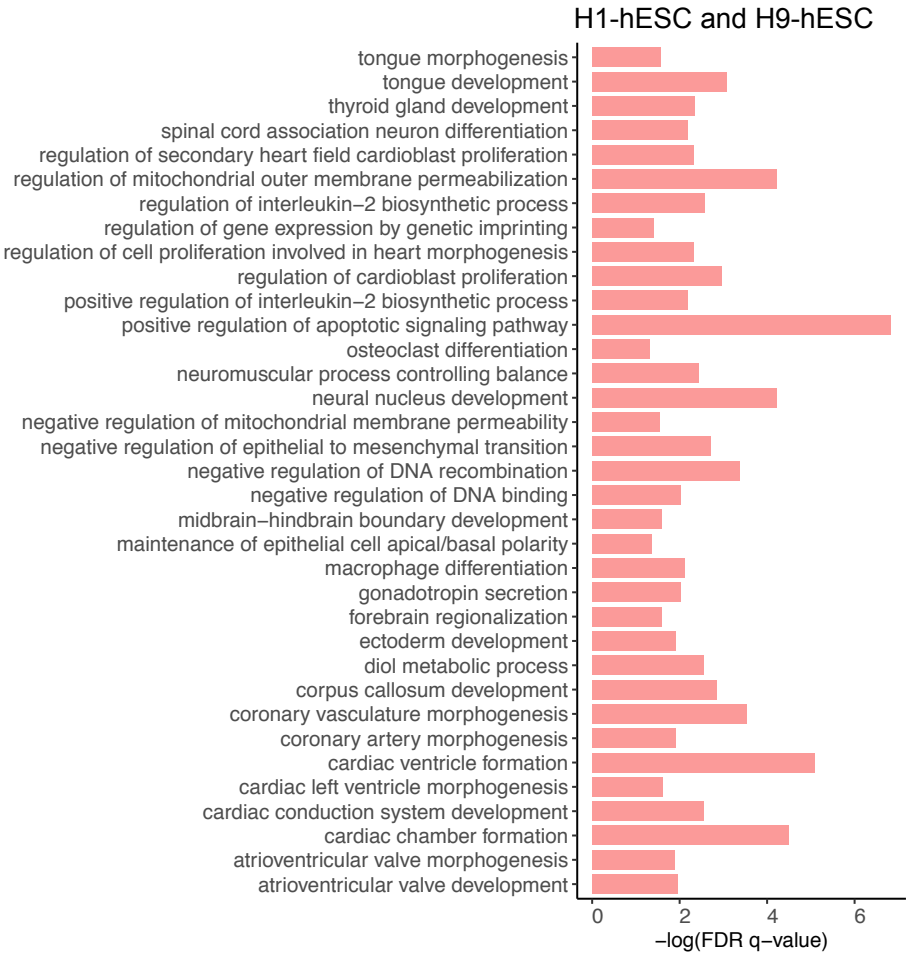

Figure 4.12: Gene ontology (GO) analysis result of the identified hESC-specific predictive genomic loci. GO analysis is performed using the software GREAT [247] for the two hESC cell types H1-hESC and H9-hESC. The vertical axis shows the list of biological processes that are found to be significantly associated with the identified hESC-specific predictive genomic loci. The horizontal axis shows the log base 10 value of q-value which controls the false discovery rate (FDR) of each identified associated biological process based on the binomial test performed by GREAT. The q-values of the identified associated biological processes are smaller than 0.05.

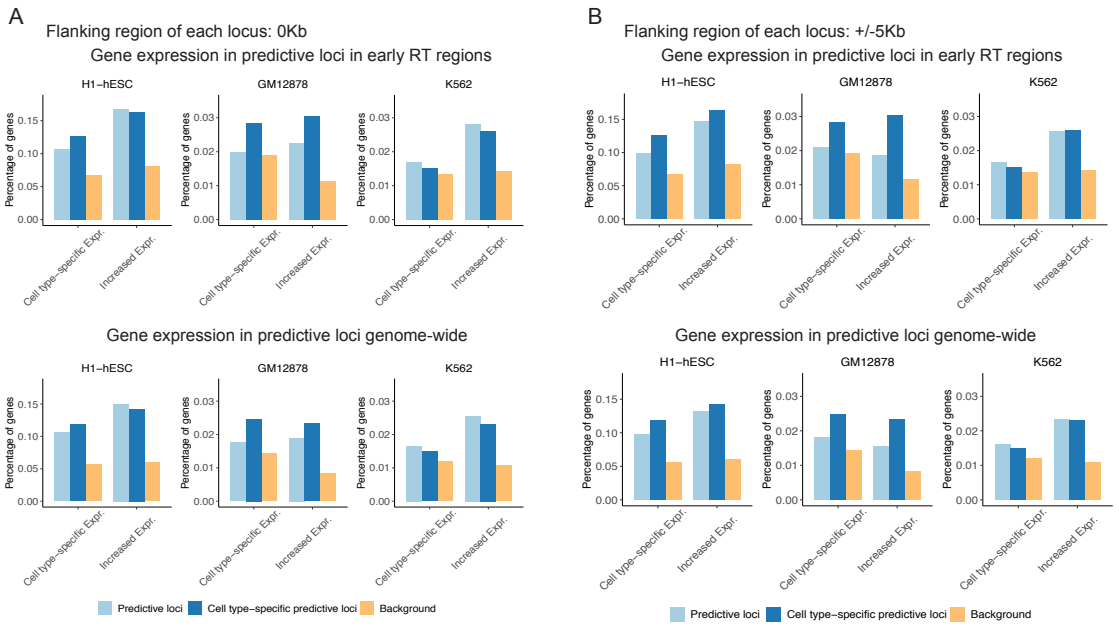

Figure 4.13: Gene expression level comparison between the identified predictive loci and the background loci in cell lines H1-hESC, GM12878, and K562. **(A)** The percentage of genes with cell type-specific expression (abbreviated as cell type-specific expr.) or cell type-specific increased expression (abbreviated as increased expr.) in the identified predictive genomic loci, identified cell type-specific predictive loci, and background loci in H1-hESC, GM12878, and K562 in the early RT regions (upper panel) and genome-wide (lower panel), respectively. The background loci represent all the genomic loci in the early RT regions or genome-wide, respectively. **(B)** The percentage of genes with cell type-specific expression or cell type-specific increased expression in proximity to the identified predictive genomic loci, identified cell type-specific predictive loci, and background loci in H1-hESC, GM12878, and K562 in the early RT regions (upper panel) and genome-wide (lower panel), respectively. Each genomic locus is extended by flanking regions of +/-1 bin (5Kb). The genes overlapping the genomic locus with flanking regions are considered to be in proximity to the corresponding locus.

Cell type	Method	PCC	Spearman's ρ	Explained variance	R^2 score
GM12878	XGBR	0.6556	0.6573	0.4298	0.4297
	RF	0.6443	0.6455	0.4148	0.4148
	LR	0.6311	0.6387	0.3983	0.3983
	DNN-Local	0.6362	0.6404	0.4039	0.3629
	CONCERT	**0.8392**	**0.8412**	**0.6975**	**0.6919**
H1-hESC	XGBR	0.6422	0.6427	0.4123	0.4114
	RF	0.6332	0.6329	0.4007	0.4000
	LR	0.6211	0.6268	0.3856	0.3846
	DNN-Local	0.5890	0.5863	0.3398	0.3382
	CONCERT	**0.8084**	**0.8109**	**0.6500**	**0.6486**
H9-hESC	XGBR	0.6418	0.6449	0.4119	0.4114
	RF	0.6313	0.6337	0.3984	0.3981
	LR	0.6046	0.6232	0.3651	0.3645
	DNN-Local	0.6205	0.6246	0.3813	0.3662
	CONCERT	**0.8168**	**0.8123**	**0.6671**	**0.6611**
HCT116	XGBR	0.7101	0.6699	0.5042	0.5040
	RF	0.6979	0.6546	0.4869	0.4865
	LR	0.6879	0.6515	0.4731	0.4728
	DNN-Local	0.6984	0.6600	0.4875	0.4512
	CONCERT	**0.8951**	**0.8630**	**0.7977**	**0.7973**
HEK293	XGBR	0.7011	0.7083	0.4912	0.4912
	RF	0.6881	0.6952	0.4734	0.4734
	LR	0.6667	0.6851	0.4433	0.4433
	DNN-Local	0.6802	0.6890	0.4617	0.4601
	CONCERT	**0.8910**	**0.8935**	**0.7914**	**0.7901**
K562	XGBR	0.6251	0.6137	0.3908	0.3903
	RF	0.6154	0.6045	0.3786	0.3780
	LR	0.5882	0.5928	0.3455	0.3448
	DNN-Local	0.6095	0.5989	0.3685	0.3307
	CONCERT	**0.8148**	**0.7946**	**0.6630**	**0.6605**
IMR90	XGBR	0.5102	0.4985	0.2600	0.2597
	RF	0.4985	0.4877	0.2484	0.2482
	LR	0.4847	0.4831	0.2343	0.2340
	DNN-Local	0.4701	0.4703	0.2181	0.1873
	CONCERT	**0.6779**	**0.6547**	**0.4589**	**0.4589**
RPE-hTERT	XGBR	0.5704	0.5643	0.3251	0.3223
	RF	0.5582	0.5512	0.3115	0.3088
	LR	0.5525	0.5496	0.3049	0.3020
	DNN-Local	0.5302	0.5247	0.2809	0.2802
	CONCERT	**0.7519**	**0.7501**	**0.5608**	**0.5530**
U2OS	XGBR	0.6892	0.6880	0.4751	0.4751
	RF	0.6826	0.6811	0.4658	0.4658
	LR	0.6682	0.6712	0.4464	0.4464
	DNN-Local	0.6754	0.6740	0.4558	0.4545
	CONCERT	**0.8611**	**0.8611**	**0.7393**	**0.7378**

Table 4.1: Performance evaluation of RT prediction using different methods in nine human cell lines. Performance evaluation is based on Pearson correlation coefficient (PCC), Spearman's rank correlation coefficient (Spearmans' ρ), explained variance and R^2 score between predicted RT signals and real RT signals across 22 autosomes in nine human cell types using different methods for prediction. The highest performance of the compared methods is in bold font.

Chromosome	PCC	Spearman's ρ	Chromosome	PCC	Spearman's ρ
chr1	0.8482	0.8534	chr12	0.8303	0.8322
chr2	0.7786	0.7801	chr13	0.7789	0.7771
chr3	0.8014	0.8041	chr14	0.8538	0.8523
chr4	0.7943	0.7844	chr15	0.8278	0.8303
chr5	0.7715	0.7761	chr16	0.8693	0.8782
chr6	0.8042	0.8038	chr17	0.8951	0.8641
chr7	0.6991	0.7076	chr18	0.8193	0.8142
chr8	0.7493	0.7349	chr19	0.7646	0.8126
chr9	0.8073	0.8128	chr20	0.8161	0.8312
chr10	0.7859	0.7883	chr21	0.8765	0.8717
chr11	0.8579	0.8638	chr22	0.8674	0.7845

Table 4.2: RT prediction performance of CONCERT on each autosomal chromosome in H1-hESC cell line. Performance evaluation is based on Pearson correlation coefficients (PCC) and Spearman's rank correlation coefficient (Spearman's ρ) between predicted RT signals and real RT signals on each autosome in H1-hESC cell line.

Model type	PCC	Spearman's ρ	Explained variance	R^2 score
CONCERT-basic	0.8084	0.8109	0.6500	0.6486
CONCERT-hierarchical	0.7850	0.7911	0.6220	0.6336

Table 4.3: Performance evaluation of RT prediction in H1-hESC cell line using different model structures of CONCERT. Performance evaluation is based on Pearson correlation coefficient (PCC), Spearman's rank correlation coefficient (Spearman's ρ, explained variance and R^2 score between predicted RT signals and real RT signals across 22 autosomes in cell line H1-hESC using the basic CONCERT model with predefined feature representation Feature I (combination of Kmer frequency features and GC-based features) and the CONCERT-hierarchical model.

Input feature type	PCC	Explained variance	R^2 score
Feature I	0.8084	0.6500	0.6486
Feature II	0.7827	0.5970	0.5652
GC-based feature	0.7285	0.5158	0.4700
Feature I+phyloP score	0.8259	0.6662	0.6570

Table 4.4: RT prediction performance evaluation of CONCERT in H1-hESC cell line using different types of features. Performance evaluation is based on Pearson correlation coefficient (PCC), explained variance and R^2 score between predicted RT signals and real RT signals across 22 autosomes in H1-hESC cell line using different types of features with the basic CONCERT model. Feature I: combination of Kmer frequency features, GC-based features, Feature II: combination of Kmer frequency features, GC-based features, and TF binding motif features, Feature I+ phyloP score: combination of Kmer frequency features, GC-based features, and phyloP score features.

Cell type	GM12878	H1-hESC	H9-hESC	H1/H9	HCT116
Number of identified predictive loci	3334	1547	860	3566	761
Cell type	HEK293	K562	IMR-90	RPE-hTERT	U2OS
Number of identified predictive loci	1768	4278	218	2002	3911

Table 4.5: The number of identified cell type-specific predictive loci. Each column shows the number of predictive loci that are only identified in one of the nine human cell types with flanking region size of +/-1 bin (bin size: 5Kb). H1/H9 corresponds to the number of predictive genomic loci that are only identified in H1-hESC or H9-hESC cell line.

Cell type	GM12878	H1-hESC	H9-hESC	H1/H9	HCT116
Number of identified predictive loci	1966	2731	3032	1826	2684
Cell type	HEK293	K562	IMR-90	RPE-hTERT	U2OS
Number of identified predictive loci	2963	1946	1118	1182	674

Table 4.6: The number of identified predictive loci in each cell type that are also identified in at least another four cell types without including flanking regions.

Number of cell types	2	3	4	5	6	7	8	9
Number of commonly identified predictive loci	9120	4373	2768	1788	893	398	129	20

Table 4.7: The number of identified RT predictive genomic loci shared by a specific number of cell types. Predicted predictive loci in H1-hESC and H9-hESC cell lines are not merged.

Number of cell types (including H1-hESC and H9-hESC)	3	4	5	6	7	8	9
Number of identified predictive loci	4055	3039	2162	1172	480	152	22

Table 4.8: The number of predictive genomic loci identified in H1-hESC or H9-hESC cell line that are also identified in different numbers of other cell types. Predicted predictive loci in H1-hESC and H9-hESC cell lines are merged. The column of number of cell types equal to k ($k = 3, \cdots, 9$) corresponds to the number of predictive loci identified in H1-hESC or H9-hESC cell line and in another $(k-2)$ cell lines.

Cell type	GO biological process	p-value
H1-hESC	positive regulation of synapse assembly	5.2e-05
	neuron migration	1.4e-04
	synapse assembly	4.9e-04
	regulation of calcium ion-dependent exocytosis	5.2e-04
	calcium ion-regulated exocytosis of neurotransmitter	7.8e-04
	negative regulation of cell proliferation	1.2e-03
	axon guidance	1.6e-03
	regulation of metanephric nephron tubule epithelial cell differentiation	1.7e-03
	heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	2.0e-03
	positive regulation of neuron projection development	2.0e-03
	vesicle fusion	3.7e-03
	brain development	4.4e-03
	negative regulation of catenin import into nucleus	4.9e-03
	homophilic cell adhesion via plasma membrane adhesion molecules	7.2e-03
	positive regulation of canonical Wnt signaling pathway	9.5e-03
	vocalization behavior	9.6e-03
	calcium ion transport into cytosol	9.6e-03
	positive regulation of neuron differentiation	9.8e-03
	nervous system development	1.2e-02
	positive regulation of dendrite extension	1.6e-02
	negative regulation of cell migration	1.9e-02
	determination of left/right symmetry	2.1e-02
	regulation of cardiac conduction	2.2e-02
	learning	2.3e-02
	cell adhesion	2.6e-02
	negative regulation of protein phosphorylation	2.8e-02
	negative regulation of angiogenesis	2.9e-02
	cell morphogenesis	2.9e-02
	negative regulation of Ras protein signal transduction	2.9e-02
	negative regulation of mitotic cell cycle	3.1e-02
	regulation of exocytosis	3.1e-02
	hippo signaling	3.4e-02
	neuromuscular junction development	3.6e-02

Table 4.9: Example biological processes that show correlation with cell type-specific expressed genes in predicted cell type-specific RT predictive genomic loci in H1-hESC cell line. We use DAVID [146] to perform the gene ontology (GO) analysis.

Chapter 5

Conclusions

In this thesis, we developed several methods for multi-species comparison of 3D genome organization and function, and genome-wide prediction of sequence elements that modulate specific types of genome function. The developed methods are aimed at addressing the computational challenges for comparing various types of functional genomic data from multiple species and identifying underlying sequence elements that may have functions in modulating genome organization and function. The applications of the proposed methods to real data reveal evolutionary patterns of 3D genome organization and DNA replication timing across multiple primate species, with identifications of genomic and epigenetic features that are correlated with the predicted evolutionary patterns. The methods have the potential to provide new insights that may advance our understanding of the regulation mechanisms of the human genome organization and function, and how they have evolved. There are multiple directions for the future work based on this thesis, which present the possibility to extend the work to more advanced methodology development and broader applications.

5.1 Summary of the methods developed in the thesis

5.1.1 Continuous-trait probabilistic model for multi-species functional genomic data comparison

A large amount of multi-species functional genomic data from high-throughput assays are becoming available to help understand the molecular mechanisms for phenotypic diversity across species. However, continuous-trait probabilistic models, which are key to such comparative analysis, remain under-explored. There was major methodological deficiency of computational methods in comparing genomic features across different species. Most of the genomic data are originally continuous signal values. All prior studies typically discretize the continuous signals into ordinal and discrete properties before performing comparison [47, 115]. Such discretization causes information loss and potential bias in the follow-up analysis. To compare continuous-trait functional genomic data across species, we developed a new model, called phylogenetic hidden Markov Gaussian processes (Phylo-HMGP), to simultaneously infer heterogeneous evolutionary states of functional genomic features in a genome-wide manner. Phylo-HMGP incorporates the Ornstein-Uhlenbeck processes [58] into the hidden Markov model, to exploit both temporal dependencies across species encoded in the evolutionary tree and the spatial dependencies across genomic loci. Both simulation studies and real data applications demonstrate the effectiveness of Phylo-HMGP. Importantly, we applied Phylo-HMGP to analyze a new cross-species DNA replication timing (RT) dataset from the same cell type in five primate species (human, chimpanzee, orangutan, gibbon, and green monkey). We demonstrate that Phylo-HMGP enables discovery of genomic regions with distinct evolutionary patterns of RT. Our method provides the first generic framework for comparative analysis of multi-species continuous functional genomic signals to help reveal regions with conserved or lineage-specific regulatory roles.

5.1.2 Phylo-HMRF for multi-species comparison of genome organization

With the development of Hi-C technology and the availability of multi-species Hi-C data, the comparison of multi-species 3D genome organization presents a new challenge in comparative genomics. We developed a phylogenetic hidden Markov random field model (Phylo-HMRF) to identify evolutionary patterns of 3D genome based on multi-species Hi-C data by jointly utilizing spatial constraints among genomic loci in the 3D space 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. The developed method provides the first computational framework to compare genome organization across multiple species using Hi-C data as continuous signals. The method also provides a new framework to analyze multi-species continuous genomic features with spatial constraints in the 3D space and has the potential to help reveal the evolutionary principles of 3D genome organization.

5.1.3 Genome-wide prediction of sequence elements that modulate RT

DNA replication timing in eukaryotic cells is strongly correlated with 3D genome organization. Proper RT control is of vital importance to maintain the composition of the epigenome and gene transcription. However, our understanding of the genomic sequence determinants regulating DNA replication timing remains limited. A major algorithmic challenge is to delineate a series of potential sequence determinants in shaping the RT programs over large-scale genomic domains. We develop a new method, named CONCERT, to simultaneously predict RT from sequence features and identify genomic sequence elements that modulate RT in a genome-wide manner. CONCERT integrates two functionally cooperative modules, a selector, which performs importance estimation-based subset sampling of the genomic sequences to detect predictive elements, and a predictor, which incorporates the bi-directional recurrent neural networks and the self-attention mechanism to perform selective learning of long-range spatial dependencies across genomic loci. We applied

CONCERT to predict RT and identify potential functional elements for RT modulation in mouse embryonic stem cells and multiple human cell types. CONCERT provides a generic interpretable machine learning framework for predicting large-scale functional genomic profiles based on sequence features and provides new insights into the potential sequence determinants of the RT program.

5.2 Future work

There are still a number of limitations of the developed methods in this thesis. New methods or method extensions based on the current models can be developed to address the existing unresolved challenges and the new challenges presented by more various types of genome organization data and functional genomic data. We will discuss different directions of potential future work, which include: (i) integrating different types of genome organization and genome function features for comprehensive evolutionary pattern identification across species (section 5.2.1 and section 5.2.2); (ii) increasing the scalabilty of the current methods for applications to larger-scale datasets (section 5.2.3); (iii) modeling both inter-species and intra-species variations to improve the accuracy in evolutionary pattern identification (section 5.2.4); (iv) extending the developed regulatory sequence element identification method towards specific applications (section 5.2.5 and section 5.2.6).

5.2.1 Integrating multiple feature types for comparative genomic pattern identification

The developed methods Phylo-HMGP and Phylo-HMRF provide frameworks to identify evolutionary patterns using a single type of continuous genomic feature. In our study, the studied genomic data are Repli-seq data and Hi-C data, respectively. With increasing availability of high-throughput sequencing data of 3D genome organization and function, there can be data of multiple feature types that are correlated with 3D genome organization for the comparison across species, such as Repli-seq data, RNA-seq data, TSA-seq data [248],

and DamID data [249]. For example, TSA-seq technology [248] estimates the cytological distance of chromosome loci genome-wide relative to specific nuclear compartments, providing measurement of spatial positioning of chromosomes in the cell nucleolus. DamID technology [249] provides genome-wide mapping of in vivo targets of chromatin proteins in eukaryotic cells and therefore provides another type of measurement of 3D genome organization. It would be important to extend the framework of Phylo-HMGP to involve multiple types of genomic features, to perform multi-variate feature comparison across different species along the genome. However, how to model multiple genomic features where there are inter-dependencies among the different types of features, in accordance with the evolutionary model and the spatial constraints of genomic loci, presents a new computational challenge. It will be useful to develop a multi-variate phylogenetic probabilistic model with spatial dependencies to study the evolutionary patterns of multiple continuous genomic features from related species. Specifically, we can incorporate the coupled hidden Markov model [250] to model the interactions among different types of genomic features. First, we have multiple Phylo-HMGP models corresponding to different types of genomic features, with one Phylo-HMGP model associated with one type of feature, respectively. Each individual Phylo-HMGP model has their own set of states. There are interactions among the different Phylo-HMGP models. The state of one model at the i-th genomic locus depends on the states of all the models at the $(i-1)$-th genomic locus. With this cross-model dependency, we can estimate the states for the different Phylo-HMGP models jointly. Next, we will explore representative combinations of states across different genomic features. Therefore, there are both state estimation results for each individual type of genomic feature, in the context that feature interactions are considered, and combinatory states estimated for the multiple genomic features.

For further improvement of the present methods, it will also be useful to include chromosome rearrangements as input features for genome organization comparison across species.

Through integration of multiple feature types, the developed methods can take advantage of the increasing availability of different types of genome organization data and func-

tional genomic data, to enable more comprehensive depiction of the evolutionary patterns of genome organization and related genome function across different species.

5.2.2 Integrating 1D functional genomic features with Hi-C data for genome organization comparison

It will also be useful to integrate 1D genomic features with Hi-C data for genome organization comparison cross species. The proposed methods Phylo-HMGP and Phylo-HMRF are applicable to 1D functional genomic data (e.g., Repli-seq data), and Hi-C data, which are presented as 2D contact matrices, respectively. How to integrate the two types of data which have different dimensions into one framework for comparison presents an another computational challenge. One approach is to transform the Hi-C data into 1D data by calling A/B compartments to capture chromatin compartment patterns encoded in the Hi-C data, and align the transformed 1D data with the other 1D functional genomic features (e.g., Repli-seq data). The disadvantage of the approach is that it causes loss of information from the original Hi-C data. The second approach is to transform the 1D functional genomic data to 2D data. We can use Repli-seq data as the example. We will generate a contact matrix corresponding to the Hi-C contact matrix, where each entry corresponds to concatenation of RT signals at the interacting paired genomic loci. We then align this RT-encoded contact matrix with the Hi-C contact matrix for multi-variate feature comparison across species utilizing the 3D spatial constraints. Specifically, I propose to use coupled-HMRF model [251]. The contact matrix with with RT concatenation features and the Hi-C contact matrix are each modeled with a Phylo-HMRF, each with a set of hidden states. Phylo-HMRF models are coupled. The state at each node (representing a pair of interacting chromatin loci) in one Phylo-HMRF is not only dependent on the states of its neighbors in the model of the same feature type, but also dependent on the states of its neighbors in the coupled HMRF model. The hidden states of the two models are inferred jointly with one objective function. We then use combinations of estimated hidden states from each model for interpretation of evolutionary patterns.

The developed method will enable utilizing Hi-C data and related 1D functional genomic data simultaneously to identify evolutionary patterns of genome organization that are closely coupled with evolution of specific genome functions.

5.2.3 Application to large-scale phylogenetic trees

The Phylo-HMGP and Phylo-HMRF models have only been applied to genomic feature comparison across a small number of closely related primate species. How to apply the proposed methods to larger-scale phylogenetic trees to perform functional genomic feature comparison across more distant species presents is another direction we can explore. There are several challenges. First, there will be more missing alignments between genomes of distantly related species, resulting in fewer available samples (each sample corresponds to an orthologous genomic locus across species). Second, as the number of model parameters increases linearly with the tree size, the model complexity increases if more species are included for comparison. The computation cost increases and the model is more easily to be over-fitted, especially if the sample size is small. To address the challenges, first, we will use imputation methods to impute part of the missing values which occur in a small scale that are due to mis-alignment of sequences between species. Second, we will explore efficient parameter regularization methods which are compatible with the evolutionary models to reduce model over-fitting. Furthermore, we can employ a hierarchical inference approach, where we divide the phylogenetic tree into clades, which are smaller-scale subtrees, and perform model inference on the sub-trees. We then aggregate the evolutionary state estimation results from the sub-trees to estimate genome-wide evolutionary patterns.

Moreover, it will also be useful to extend the methods to simultaneously infer the phylogenetic tree topology and the evolutionary states, as we have discussed in section 2.11.

The improved methods will enable us to perform genomic feature comparison across a larger number of species with the available data and gain broader-scope understanding on the evolution of genome organization or related genome functions, which may help us identify important regulatory mechanism of genome organization that are conserved across

different species or species-specific regulatory mechanisms.

5.2.4 Modeling intra-species variations

Genetic variations can contribute to differences in DNA replication timing among individuals [158–160]. Studies have also revealed that there is heterogeneity and intrinsic variation of genome organization across individual cells [184]. Phylo-HMGP and Phylo-HMRF have the limitation that the methods do not specifically consider the impact of intra-species variation on the evolutionary pattern identification. Since we only have one replicate from each species for analysis in our study, we have not extended the methods to account for intra-species variations, which may be caused by the genetic variations across different individuals. It would be an important methodological improvement to model the intra-species variations. Assume we have the specific type of functional genomic data or Hi-C data of multiple individuals from the same species, we propose to model the observation of an individual with a Gaussian distribution, with both the mean value and variance of the individual-wise distribution accounting for the intra-species variations. The average of observations from different individuals of the same species follows a Gaussian distribution, which can be parameterized by the OU process. In this way, the parameters of the species-level distribution will be both functions of the OU parameters and functions of the individual-level distribution parameters, which encode the intra-species variations. However, this assumption largely increases the number of parameters to learn. Alternatively, we will assume the parameters of individual distributions are sampled from a distribution, which is dependent on the species-level OU process.

The method will enable us to capture both inter-species differences and intra-species variations for more accurate evolutionary pattern identification when population-level multi-species genomic data are available.

5.2.5 Predicting genome-wide sequence elements that modulate 3D genome organization

The CONCERT model is now limited to predicting 1D functional genomic signals (e.g., Repli-seq data) and identifying sequence elements with potential modulation functions. It would be useful to improve the CONCERT model to apply to prediction of genome-wide sequence elements that modulate 3D genome organization, by extending the framework to predict 2D Hi-C contact maps and identifying predictive elements from both ends of chromatin interactions. For this purpose, the CONCERT model will be changed into the combination of two selector modules and a modified predictor module. For each pair of genomic loci in a chromatin interaction, each of the two selectors attend to one sequence of the paired loci to perform importance score estimation for predictive element selection. The input sequence features from the paired loci are each weighted by the corresponding estimated importance scores and concatenated as input to the predictor for genomic signal prediction. In this way, we perform interaction-wise importance score estimation for the sequence of each genomic locus, to consider possible feature interactions between a pair of loci. To model the spatial dependencies across genomic loci in the 3D space, instead of using BiLSTM which models the 1D spatial dependencies, we will use 2D CNNs (convolution neural networks) in the predictor module.

The new method will facilitate effective identification of the sequence elements that may modulate genome organization, for which we still have very limited knowledge.

5.2.6 Incorporating epigenetic features to study cell type-specific regulation mechanism of genome organization and function

Furthermore, we can include epigenetic features into the CONCERT model to study cell-type specific characteristics of DNA replication timing regulation and genome organization regulation, and improve the prediction accuracy. For example, studies have revealed that chromatin accessibility is critical for RT prediction [39]. Chromatin accessibility varies across different cell types and may also change during the cell differentiation process. In-

corporating chromatin accessibility measurements (e.g., DNase-seq data [252], ATAC-seq data [253]) may facilitate more accurate prediction of specific genomic features such as DNA RT, and may also reveal possible interactions between the sequence features and the chromatin accessibility. We will also include other types of epgenetic features, for example, histone modifications with experimental evidence in potential RT modulation [74, 75, 254] and TF binding sites based on ChIP-seq data into the modeling and prediction.

Integrating DNA sequence features with epigenetic features for genomic signal prediction and regulatory element identification may help us gain understanding on how sequence elements and potential epigenetic regulators function cooperatively to regulate genome organization and genome functions in different cell types.

5.3 Summary

In this thesis, the method development is focused on two related aspects: (i) comparative pattern recognition based on continuous-trait functional genomic data and 3D genome organization data from different species; (ii) identification of DNA sequence elements that may be involved in the potential sequence-dependent regulation mechanisms of the genome organization and genome function patterns. The methods developed for (i) are mainly focused on addressing two connected computational challenges: (1) modeling different types of genomic features across species as continuous traits for comparison, without the need of data discretization; (2) modeling both the evolutionary dependencies across species and the spatial dependency of genomic loci along the 1D genome or on the 3D genome. The method developed for (ii) is an attempt to connect the observed functional genomic feature patterns to the underlying DNA sequences, which is mainly aimed at addressing the challenge of how to effectively identify a small set of sequence elements with potential modulation functions from the large-scale spatial domains for a specific type of functional genomic feature of which the regulatory principles are unclear.

The methods in this thesis build up a set of computational frameworks for compara-

tive pattern recognition and the study on sequence-dependent regulation mechanisms of genome functions. The different directions of the future work, as discussed in section 5.2, suggest the possibilities to further improve the current computational frameworks with new method development. The methods, with possible improvement from the future work, not only provide us with efficient approaches to analyzing multi-species genome organization and functional genomic data, but also have the potential to advance our understanding on the regulatory principles of genome organization and related genome functions.

The methods developed in Chapter 2 (Phylo-HMGP) and in Chapter 3 (Phylo-HMRF) with related real data applications were presented in [161] and [255], respectively.

Bibliography

[1] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[2] Boyan Bonev and Giacomo Cavalli. Organization and function of the 3D genome. *Nature Reviews Genetics*, 17(11):661, 2016.

[3] M. J. Rowley and V. G. Corces. Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, 19(12):789–800, Dec 2018.

[4] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[5] Zhonghui Tang, Oscar Junhong Luo, Xingwang Li, Meizhen Zheng, Jacqueline Jufen Zhu, Przemyslaw Szalaj, Pawel Trzaskoma, Adriana Magalska, Jakub Wlodarczyk, Blazej Ruszczycki, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, 163(7):1611–1627, 2015.

[6] Robert A Beagrie, Antonio Scialdone, Markus Schueler, Dorothee CA Kraemer, Mita Chotalia, Sheila Q Xie, Mariano Barbieri, Inês de Santiago, Liron-Mark Lavitas, Miguel R Branco, et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646):519–524, 2017.

[7] Sofia A Quinodoz, Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt,

Elizabeth Detmar, Mason M Lai, Alexander A Shishkin, Prashant Bhat, Yodai Takei, et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*, 174(3):744–757, 2018.

[8] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[9] Vishnu Dileep, Ferhat Ay, Jiao Sima, Daniel L Vera, William S Noble, and David M Gilbert. Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program. *Genome Research*, 25(8):1104–1113, 2015.

[10] Claire Marchal, Jiao Sima, and David M Gilbert. Control of DNA replication timing in the 3D genome. *Nature Reviews Molecular Cell Biology*, pages 1–17, 2019.

[11] Jian Ma and Zhijun Duan. Replication timing becomes intertwined with 3D genome organization. *Cell*, 176(4):681–684, 2019.

[12] Stefan Schoenfelder and Peter Fraser. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, page 1, 2019.

[13] Tom Misteli. Higher-order genome organization in human disease. *Cold Spring Harbor Perspectives in Biology*, 2(8):a000794, 2010.

[14] Peter Hugo Lodewijk Krijger and Wouter De Laat. Regulation of disease-associated gene expression in the 3D genome. *Nature Reviews Molecular Cell Biology*, 17(12): 771, 2016.

[15] Hui Zheng and Wei Xie. The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology*, page 1, 2019.

[16] Angela Sparago, Flavia Cerrato, Maria Vernucci, Giovanni Battista Ferrero, Margherita Cirillo Silengo, and Andrea Riccio. Microdeletions in the human H19 DMR result in loss of IGF2 imprinting and Beckwith-Wiedemann syndrome. *Nature Genetics*, 36(9):958–960, 2004.

[17] Randell T Libby, Katharine A Hagerman, Victor V Pineda, Rachel Lau, Diane H

Cho, Sandy L Baccam, Michelle M Axford, John D Cleary, James M Moore, Bryce L Sopher, et al. CTCF cis-regulates trinucleotide repeat instability in an epigenetic manner: a novel basis for mutational hot spot determination. *PLoS Genet*, 4 (11):e1000257, 2008.

[18] Thomas Eggermann, Katja Eggermann, and Nadine Schönherr. Growth retardation versus overgrowth: Silver-Russell syndrome is genetically opposite to Beckwith-Wiedemann syndrome. *Trends in Genetics*, 24(4):195–204, 2008.

[19] Kerstin S Wendt, Keisuke Yoshida, Takehiko Itoh, Masashige Bando, Birgit Koch, Erika Schirghuber, Shuichi Tsutsumi, Genta Nagae, Ko Ishihara, Tsuyoshi Mishiro, et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, 451(7180):796–801, 2008.

[20] Vincent Dion and John H Wilson. Instability and chromatin structure of expanded trinucleotide repeats. *Trends in Genetics*, 25(7):288–297, 2009.

[21] Michael Witcher and Beverly M Emerson. Epigenetic silencing of the p16INK4a tumor suppressor is associated with loss of CTCF binding and a chromatin boundary. *Molecular Cell*, 34(3):271–284, 2009.

[22] Sabina Benko, Christopher T Gordon, Delphine Mallet, Rajini Sreenivasan, Christel Thauvin-Robinet, Atle Brendehaug, Sophie Thomas, Ove Bruland, Michel David, Marc Nicolino, et al. Disruption of a long distance regulatory region upstream of SOX9 in isolated disorders of sex development. *Journal of Medical Genetics*, 48 (12):825–830, 2011.

[23] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[24] Alejandro Medrano-Fernández and Angel Barco. Nuclear organization and 3D chromatin architecture in cognition and neuropsychiatric disorders. *Molecular Brain*, 9 (1):1–12, 2016.

[25] Darío G Lupiáñez, Malte Spielmann, and Stefan Mundlos. Breaking tads: how alterations of chromatin domains result in disease. *Trends in Genetics*, 32(4):225–237, 2016.

[26] Manuel Rosa-Garrido, Douglas J Chapski, Anthony D Schmitt, Todd H Kimball, Elaheh Karbassi, Emma Monte, Enrique Balderas, Matteo Pellegrini, Tsai-Ting Shih, Elizabeth Soehalim, et al. High-resolution mapping of chromatin conformation in cardiac myocytes reveals structural remodeling of the epigenome in heart failure. *Circulation*, 136(17):1613–1625, 2017.

[27] Liron Davis, Itay Onn, and Evan Elliott. The emerging roles for the chromatin structure regulators ctcf and cohesin in neurodevelopment and behavior. *Cellular and Molecular Life Sciences*, 75(7):1205–1214, 2018.

[28] Elphège P Nora, Anton Goloborodko, Anne-Laure Valton, Johan H Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A Mirny, and Benoit G Bruneau. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, 169(5):930–944, 2017.

[29] Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241, 2003.

[30] Dario Boffelli, Jon McAuliffe, Dmitriy Ovcharenko, Keith D Lewis, Ivan Ovcharenko, Lior Pachter, and Edward M Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611): 1391–1394, 2003.

[31] Xiaohui Xie, Jun Lu, EJ Kulbokas, Todd R Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434 (7031):338–345, 2005.

[32] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F Lin, Brian J Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, et al. A

high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.

[33] Jessica Alföldi and Kerstin Lindblad-Toh. Comparative genomics as a tool to understand evolution and disease. *Genome Research*, 23(7):1063–1068, 2013.

[34] Jeffrey Rogers and Richard A Gibbs. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nature Reviews Genetics*, 15(5):347–359, 2014.

[35] Matteo Vietri Rudan, Christopher Barrington, Stephen Henderson, Christina Ernst, Duncan T Odom, Amos Tanay, and Suzana Hadjur. Comparative hi-c reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Reports*, 10 (8):1297–1309, 2015.

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

[37] Nicholas Rhind and David M Gilbert. DNA replication timing. *Cold Spring Harbor Perspectives in Biology*, 5(8):a010132, 2013.

[38] Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq, author=Marchal, Claire and Sasaki, Takayo and Vera, Daniel and Wilson, Korey and Sima, Jiao and Rivera-Mulia, Juan Carlos and Trevilla-García, Claudia and Nogues, Coralin and Nafie, Ebtesam and Gilbert, David M. *Nature Protocols*, 13(5):819–839, 2018.

[39] Yevgeniy Gindin, Manuel S Valenzuela, Mirit I Aladjem, Paul S Meltzer, and Sven Bilke. A chromatin structure-based model accurately predicts DNA replication timing in human cells. *Molecular Systems Biology*, 10(3):722, 2014.

[40] Federico Comoglio, Tommy Schlumpf, Virginia Schmid, Remo Rohs, Christian Beisel, and Renato Paro. High-resolution profiling of drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell Reports*, 11(5):821–834, 2015.

[41] Tyrone Ryba, Ichiro Hiratani, Junjie Lu, Mari Itoh, Michael Kulik, Jinfeng Zhang, Thomas C Schulz, Allan J Robins, Stephen Dalton, and David M Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Research*, 20(6):761–770, 2010.

[42] Benjamin D Pope, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L Vera, Yanli Wang, R Scott Hansen, Theresa K Canfield, et al. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405, 2014.

[43] Irina Solovei, Katharina Thanisch, and Yana Feodorova. How to rule the nucleus: divide et impera. *Current Opinion in Cell Biology*, 40:47–59, 2016.

[44] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

[45] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, 2007.

[46] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, 4(8):651–657, 2007.

[47] Jianghan Qu, Emily Hodges, Antoine Molaro, Pascal Gagneux, Matthew D Dean, Gregory J Hannon, and Andrew D Smith. Evolutionary expansion of DNA hypomethylation in the mammalian germline genome. *Genome Research*, 28(2):145–158, 2018.

[48] Diego Villar, Camille Berthelot, Sarah Aldridge, Tim F Rayner, Margus Lukk, Miguel Pignatelli, Thomas J Park, Robert Deaville, Jonathan T Erichsen, Anna J Jasinska, et al. Enhancer evolution across 20 mammalian species. *Cell*, 160(3):

554–566, 2015.

[49] Justin Cotney, Jing Leng, Jun Yin, Steven K Reilly, Laura E DeMare, Deena Emera, Albert E Ayoub, Pasko Rakic, and James P Noonan. The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell*, 154(1):185–196, 2013.

[50] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, et al. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, 2011.

[51] Michal Levin, Leon Anavy, Alison G Cole, Eitan Winter, Natalia Mostov, Sally Khair, Naftalie Senderovich, Ekaterina Kovalev, David H Silver, Martin Feder, et al. The mid-developmental transition and the evolution of animal body plans. *Nature*, 531(7596):637–641, 2016.

[52] Marina Naval-Sánchez, Delphine Potier, Gert Hulselmans, Valerie Christiaens, and Stein Aerts. Identification of lineage-specific cis-regulatory modules associated with variation in transcription factor binding and chromatin activity using ornstein–uhlenbeck models. *Molecular Biology and Evolution*, 32(9):2441–2455, 2015.

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

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

[55] Mark Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401 (6756):877–884, 1999.

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

[57] Thomas F Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351, 1997.

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

[59] Thomas F Hansen, Jason Pienaar, and Steven Hecht Orzack. A comparative method for studying adaptation to a randomly evolving environment. *Evolution*, 62(8):1965–1977, 2008.

[60] Alexei J Drummond, Marc A Suchard, Dong Xie, and Andrew Rambaut. Bayesian phylogenetics with beauti and the beast 1.7. *Molecular Biology and Evolution*, 29 (8):1969–1973, 2012.

[61] Jenny Chen, Ross Swofford, Jeremy Johnson, Beryl B Cummings, Noga Rogel, Kerstin Lindblad-Toh, Wilfried Haerty, Federica Di Palma, and Aviv Regev. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Research*, 29(1):53–63, 2019.

[62] Tyrone Ryba, Ichiro Hiratani, Takayo Sasaki, Dana Battaglia, Michael Kulik, Jinfeng Zhang, Stephen Dalton, and David M Gilbert. Replication timing: a fingerprint for cell identity and pluripotency. *PLoS Computational Biology*, 7(10):e1002225, 2011.

[63] Feng Yue, Yong Cheng, Alessandra Breschi, Jeff Vierstra, Weisheng Wu, Tyrone Ryba, Richard Sandstrom, Zhihai Ma, Carrie Davis, Benjamin D Pope, et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515 (7527):355–364, 2014.

[64] Juan Carlos Rivera-Mulia, Quinton Buckley, Takayo Sasaki, Jared Zimmerman, Ruth A Didier, Kristopher Nazor, Jeanne F Loring, Zheng Lian, Sherman Weissman, Allan J Robins, et al. Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. *Genome Research*, 25 (8):1091–1103, 2015.

[65] Tyrone Ryba, Dana Battaglia, Bill H Chang, James W Shirley, Quinton Buckley, Benjamin D Pope, Meenakshi Devidas, Brian J Druker, and David M Gilbert. Ab-

normal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. *Genome Research*, 22(10):1833–1844, 2012.

[66] Takayo Sasaki, Juan Carlos Rivera-Mulia, Daniel Vera, Jared Zimmerman, Sunny Das, Michelle Padget, Naoto Nakamichi, Bill H Chang, Jeff Tyner, Brian J Druker, et al. Stability of patient-specific features of altered DNA replication timing in xenografts of primary human acute lymphoblastic leukemia. *Experimental Hematology*, 51:71–82, 2017.

[67] Juan Carlos Rivera-Mulia, Romain Desprat, Claudia Trevilla-Garcia, Daniela Cornacchia, Hélène Schwerer, Takayo Sasaki, Jiao Sima, Tyler Fells, Lorenz Studer, Jean-Marc Lemaitre, et al. DNA replication timing alterations identify common markers between distinct progeroid diseases. *Proceedings of the National Academy of Sciences*, 114(51):E10972–E10980, 2017.

[68] Eitan Yaffe, Shlomit Farkash-Amar, Andreas Polten, Zohar Yakhini, Amos Tanay, and Itamar Simon. Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genetics*, 6(7):e1001011, 2010.

[69] Shlomit Farkash-Amar and Itamar Simon. Genome-wide analysis of the replication program in mammals. *Chromosome research*, 18(1):115–125, 2010.

[70] Irina Solovei, Moritz Kreysing, Christian Lanctôt, Süleyman Kösem, Leo Peichl, Thomas Cremer, Jochen Guck, and Boris Joffe. Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. *Cell*, 137(2):356–368, 2009.

[71] Nathan H Lazar, Kimberly A Nevonen, Brendan O'Connell, Christine McCann, Rachel J O'Neill, Richard E Green, Thomas J Meyer, Mariam Okhovat, and Lucia Carbone. Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Research*, 28(7):983–997, 2018.

[72] Sylvain Foissac, Sarah Djebali, Kylie Munyard, Nathalie Vialaneix, Andrea Rau, Kevin Muret, Diane Esquerré, Matthias Zytnicki, Thomas Derrien, Philippe Bardou, et al. Multi-species annotation of transcriptome and chromatin structure in domesti-

cated animals. *BMC biology*, 17(1):1–25, 2019.

[73] Tomoki Yokochi, Kristina Poduch, Tyrone Ryba, Junjie Lu, Ichiro Hiratani, Makoto Tachibana, Yoichi Shinkai, and David M Gilbert. G9a selectively represses a class of late-replicating genes at the nuclear periphery. *Proceedings of the National Academy of Sciences*, 106(46):19363–19368, 2009.

[74] Kazumasa Yoshida, Julien Bacal, Damien Desmarais, Ismaël Padioleau, Olga Tsaponina, Andrei Chabes, Véronique Pantesco, Emeric Dubois, Hugues Parrinello, Magdalena Skrzypczak, et al. The histone deacetylases sir2 and rpd3 act on ribosomal DNA to control the replication program in budding yeast. *Molecular Cell*, 54 (4):691–697, 2014.

[75] Jiao Sima, Abhijit Chakraborty, Vishnu Dileep, Marco Michalski, Kyle N Klein, Nicolas P Holcomb, Jesse L Turner, Michelle T Paulsen, Juan Carlos Rivera-Mulia, Claudia Trevilla-Garcia, et al. Identifying cis elements for spatiotemporal control of mammalian DNA replication. *Cell*, 176(4):816–830, 2019.

[76] Emilie Besnard, Amélie Babled, Laure Lapasset, Ollivier Milhavet, Hugues Parrinello, Christelle Dantec, Jean-Michel Marin, and Jean-Marc Lemaitre. Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nature Structural & Molecular Biology*, 19(8):837, 2012.

[77] Larry D Mesner, Veena Valsakumar, Marcin Cieślik, Rebecca Pickin, Joyce L Hamlin, and Stefan Bekiranov. Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early-and late-firing origins. *Genome Research*, 23(11):1774–1788, 2013.

[78] Nataliya Petryk, Malik Kahli, Yves d'Aubenton Carafa, Yan Jaszczyszyn, Yimin Shen, Maud Silvain, Claude Thermes, Chun-Long Chen, and Olivier Hyrien. Replication landscape of the human genome. *Nature Communications*, 7(1):1–13, 2016.

[79] Alexander R Langley, Stefan Gräf, James C Smith, and Torsten Krude. Genome-wide identification and characterisation of human DNA replication origins by initia-

tion site sequencing (ini-seq). *Nucleic Acids Research*, 44(21):10230–10247, 2016.

[80] Gaetano Ivan Dellino, Davide Cittaro, Rossana Piccioni, Lucilla Luzi, Stefania Banfi, Simona Segalla, Matteo Cesaroni, Ramiro Mendoza-Maldonado, Mauro Giacca, and Pier Giuseppe Pelicci. Genome-wide mapping of human DNA-replication origins: levels of transcription at orc1 sites regulate origin selection and replication timing. *Genome Research*, 23(1):1–11, 2013.

[81] Benoit Miotto, Zhe Ji, and Kevin Struhl. Selectivity of orc binding sites and the relation to replication timing, fragile sites, and deletions in cancers. *Proceedings of the National Academy of Sciences*, 113(33):E4810–E4819, 2016.

[82] Christelle Cayrou, Benoit Ballester, Isabelle Peiffer, Romain Fenouil, Philippe Coulombe, Jean-Christophe Andrau, Jacques van Helden, and Marcel Méchali. The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Research*, 25(12):1873–1885, 2015.

[83] Jean-Charles Cadoret, Françoise Meisch, Vahideh Hassan-Zadeh, Isabelle Luyten, Claire Guillet, Laurent Duret, Hadi Quesneville, and Marie-Noëlle Prioleau. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proceedings of the National Academy of Sciences*, 105(41):15837–15842, 2008.

[84] Franck Picard, Jean-Charles Cadoret, Benjamin Audit, Alain Arneodo, Adriana Alberti, Christophe Battail, Laurent Duret, and Marie-Noelle Prioleau. The spatiotemporal program of DNA replication is associated with specific combinations of chromatin marks in human cells. *PLoS Genet*, 10(5):e1004282, 2014.

[85] Boris Bartholdy, Rituparna Mukhopadhyay, Julien Lajugie, Mirit I Aladjem, and Eric E Bouhassira. Allele-specific analysis of DNA replication origins in mammalian cells. *Nature Communications*, 6(1):1–12, 2015.

[86] DT Stinchcomb, K Struhl, and RW Davis. Isolation and characterisation of a yeast chromosomal replicator. *Nature*, 282(5734):39–43, 1979.

[87] David M Gilbert. Making sense of eukaryotic DNA replication origins. *Science*, 294

(5540):96–100, 2001.

[88] Silvia Diaz-Perez, Yan Ouyang, Vanessa Perez, Roxanna Cisneros, Moira Regelson, and York Marahrens. The element (s) at the nontranscribed xist locus of the active x chromosome controls chromosomal replication timing in the mouse. *Genetics*, 171 (2):663–672, 2005.

[89] Silvia V Diaz-Perez, David O Ferguson, Chen Wang, Gyorgyi Csankovszki, Chengming Wang, Shih-Chang Tsai, Devkanya Dutta, Vanessa Perez, SunMin Kim, C Daniel Eller, et al. A deletion at the mouse xist gene exposes trans-effects that alter the heterochromatin of the inactive x chromosome and the replication time and dna stability of both x chromosomes. *Genetics*, 174(3):1115–1133, 2006.

[90] Nathan Donley, Eric P Stoffregen, Leslie Smith, Christina Montagna, and Mathew J Thayer. Asynchronous replication, mono-allelic expression, and long range cis-effects of ASAR6. *PLoS Genet*, 9(4):e1003423, 2013.

[91] Nathan Donley, Leslie Smith, and Mathew J Thayer. ASAR15, a cis-acting locus that controls chromosome-wide replication timing and stability of human chromosome 15. *PLoS Genet*, 11(1):e1004923, 2015.

[92] Eric P Stoffregen, Nathan Donley, Daniel Stauffer, Leslie Smith, and Mathew J Thayer. An autosomal locus that controls chromosome-wide replication timing and mono-allelic expression. *Human molecular genetics*, 20(12):2366–2378, 2011.

[93] Federico Comoglio and Renato Paro. Combinatorial modeling of chromatin features quantitatively predicts DNA replication timing in drosophila. *PLoS Computational Biology*, 10(1):e1003419, 2014.

[94] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831, 2015.

[95] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931, 2015.

[96] Daniel Quang and Xiaohui Xie. Danq: a hybrid convolutional and recurrent deep

neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 44(11):e107–e107, 2016.

[97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[98] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999, 2016.

[99] Sai Zhang, Hailin Hu, Tao Jiang, Lei Zhang, and Jianyang Zeng. Titer: predicting translation initiation sites by deep learning. *Bioinformatics*, 33(14):i234–i242, 2017.

[100] Josh T Cuperus, Benjamin Groves, Anna Kuchina, Alexander B Rosenberg, Nebojsa Jojic, Stanley Fields, and Georg Seelig. Deep learning of the regulatory grammar of yeast 5 untranslated regions from 500,000 random sequences. *Genome Research*, 27 (12):2015–2024, 2017.

[101] Bite Yang, Feng Liu, Chao Ren, Zhangyi Ouyang, Ziwei Xie, Xiaochen Bo, and Wenjie Shu. Biren: predicting enhancers with a deep-learning-based model using the DNA sequence alone. *Bioinformatics*, 33(13):1930–1936, 2017.

[102] Jacob D Washburn, Maria Katherine Mejia-Guerra, Guillaume Ramstein, Karl A Kremling, Ravi Valluru, Edward S Buckler, and Hai Wang. Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proceedings of the National Academy of Sciences*, 116(12):5542–5549, 2019.

[103] Shashank Singh, Yang Yang, Barnabás Póczos, and Jian Ma. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quantitative Biology*, pages 1–16, 2019.

[104] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Computational Biology*, 13(1):e1005324, 2017.

[105] David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y

McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, 2018.

[106] Geoff Fudenberg, David R Kelley, and Katherine S Pollard. Predicting 3D genome folding from DNA sequence with akita. *Nature Methods*, pages 1–7, 2020.

[107] Ron Schwessinger, Matthew Gosden, Damien Downes, Richard C Brown, A Marieke Oudelaar, Jelena Telenius, Yee Whye Teh, Gerton Lunter, and Jim R Hughes. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nature Methods*, pages 1–7, 2020.

[108] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[109] Nynke L Van Berkum, Erez Lieberman-Aiden, Louise Williams, Maxim Imakaev, Andreas Gnirke, Leonid A Mirny, Job Dekker, and Eric S Lander. Hi-c: a method to study the three-dimensional architecture of genomes. *JoVE (Journal of Visualized Experiments)*, (39):e1869, 2010.

[110] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, 2002.

[111] Tyrone Ryba, Dana Battaglia, Benjamin D Pope, Ichiro Hiratani, and David M Gilbert. Genome-scale analysis of replication timing: from bench to bioinformatics. *Nature Protocols*, 6(6):870–895, 2011.

[112] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[113] William S Cleveland and Susan J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610, 1988.

[114] Lawrence Rabiner and B Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[115] Adam Siepel and David Haussler. Phylogenetic hidden markov models. In *Statistical*

Methods in Molecular Evolution, R. Nielsen, ed. (New York, USA: Springer),, pages 325–351. 2005.

[116] Asger Hobolth, Ole F Christensen, Thomas Mailund, and Mikkel H Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS Genetics*, 3(2):e7, 2007.

[117] Kevin J Liu, Jingxuan Dai, Kathy Truong, Ying Song, Michael H Kohn, and Luay Nakhleh. An hmm-based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Computational Biology*, 10(6):e1003649, 2014.

[118] Jens Ledet Jensen and Anne-Mette Krabbe Pedersen. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Advances in Applied Probability*, 32(2):499–517, 2000.

[119] Gerton Lunter and Jotun Hein. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, 20(suppl_1):i216–i223, 2004.

[120] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[121] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[122] Evelyn Dittmer. *Hidden Markov Models with time-continuous output behavior*. PhD thesis, Freie Universität Berlin, Berlin, Germany, 2009.

[123] Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, Technical Report:ICSI TR–97–021, 1998.

[124] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 1–38, 1977.

[125] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[126] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–

269, 1967.

[127] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[128] Piotr Zwiernik, Caroline Uhler, and Donald Richards. Maximum likelihood estimation for linear gaussian covariance models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1269–1292, 2017.

[129] Gavin H Thomas, Shai Meiri, and Albert B Phillimore. Body size diversification in anolis: novel environment and island effects. *Evolution*, 63(8):2017–2030, 2009.

[130] Gavin H Thomas, Robert P Freckleton, and Tamas Székely. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. *Proceedings of the Royal Society of London B: Biological Sciences*, 273 (1594):1619–1624, 2006.

[131] Matthew E Johnson, Ze Cheng, V Anne Morrison, Steven Scherer, Mario Ventura, Richard A Gibbs, Eric D Green, and Evan E Eichler. Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proceedings of the National Academy of Sciences*, 103(47):17626–17631, 2006.

[132] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[133] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome Research*, 12(6):996–1006, 2002.

[134] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[135] The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437:69–87, 2005.

[136] Devin P Locke, LaDeana W Hillier, Wesley C Warren, Kim C Worley, Lynne V

Nazareth, Donna M Muzny, Shiaw-Pyng Yang, Zhengyuan Wang, Asif T Chinwalla, Pat Minx, et al. Comparative and demographic analysis of orang-utan genomes. *Nature*, 469(7331):529–533, 2011.

[137] Lucia Carbone, R Alan Harris, Sante Gnerre, Krishna R Veeramah, Belen Lorente-Galdos, John Huddleston, Thomas J Meyer, Javier Herrero, Christian Roos, Bronwen Aken, et al. Gibbon genome and the fast karyotype evolution of small apes. *Nature*, 513(7517):195–201, 2014.

[138] Angela S Hinrichs, Donna Karolchik, Robert Baertsch, Galt P Barber, Gill Bejerano, Hiram Clawson, Mark Diekhans, Terrence S Furey, Rachel A Harte, Fan Hsu, et al. The ucsc genome browser database: update 2006. *Nucleic Acids Research*, 34 (suppl_1):D590–D598, 2006.

[139] Nathan Day, Andrew Hemmaplardh, Robert E Thurman, John A Stamatoyannopoulos, and William S Noble. Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11):1424–1426, 2007.

[140] Donald B Percival and Andrew T Walden. *Wavelet methods for time series analysis (Cambridge, UK: Cambridge University Press)*. 2006.

[141] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(2):224–227, 1979.

[142] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval (Cambridge, UK: Cambridge University Press)*.

[143] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[144] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[145] Qian Du, Saul A Bert, Nicola J Armstrong, C Elizabeth Caldon, Jenny Z Song,

Shalima S Nair, Cathryn M Gould, Phuc-Loi Luu, Timothy Peters, Amanda Khoury, et al. Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nature Communications*, 10(1):1–15, 2019.

[146] Da Wei Huang, Brad T Sherman, Qina Tan, Jack R Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9): R183, 2007.

[147] Jonathan Casper, Ann S Zweig, Chris Villarreal, Cath Tyner, Matthew L Speir, Kate R Rosenbloom, Brian J Raney, Christopher M Lee, Brian T Lee, Donna Karolchik, et al. The ucsc genome browser database: 2018 update. *Nucleic Acids Research*, 46(D1):D762–D769, 2017.

[148] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[149] Anthony Mathelier, Oriol Fornes, David J Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, et al. Jaspar 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–D115, 2016.

[150] Kate R Rosenbloom, Cricket A Sloan, Venkat S Malladi, Timothy R Dreszer, Katrina Learned, Vanessa M Kirkup, Matthew C Wong, Morgan Maddren, Ruihua Fang, Steven G Heitner, et al. Encode data in the ucsc genome browser: year 5 update. *Nucleic Acids Research*, 41(D1):D56–D63, 2012.

[151] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[152] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, 2007.

[153] Elisa Laurenti, Sergei Doulatov, Sasan Zandi, Ian Plumb, Jing Chen, Craig April, Jian-Bing Fan, and John E Dick. The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nature Immunology*, 14(7):756–763, 2013.

[154] Yiru Zhang, Yuqian Xing, Lei Zhang, Yang Mei, Kazuo Yamamoto, Tak W Mak, and Han You. Regulation of cell cycle progression by forkhead transcription factor foxo3 through its binding partner DNA replication factor cdt1. *Proceedings of the National Academy of Sciences*, 109(15):5717–5722, 2012.

[155] Martin C Frith, Michael C Li, and Zhiping Weng. Cluster-buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Research*, 31(13):3666–3668, 2003.

[156] Steve Horvath, Wiebke Erhart, Mario Brosch, Ole Ammerpohl, Witigo von Schönfels, Markus Ahrens, Nils Heits, Jordana T Bell, Pei-Chien Tsai, Tim D Spector, et al. Obesity accelerates epigenetic aging of human liver. *Proceedings of the National Academy of Sciences*, 111(43):15538–15543, 2014.

[157] Nir Friedman, Matan Ninio, Itsik Pe'er, and Tal Pupko. A structural EM algorithm for phylogenetic inference. *Journal of Computational Biology*, 9(2):331–353, 2002.

[158] Amnon Koren, Robert E Handsaker, Nolan Kamitaki, Rosa Karlić, Sulagna Ghosh, Paz Polak, Kevin Eggan, and Steven A McCarroll. Genetic variation in human DNA replication timing. *Cell*, 159(5):1015–1026, 2014.

[159] Rituparna Mukhopadhyay, Julien Lajugie, Nicolas Fourel, Ari Selzer, Michael Schizas, Boris Bartholdy, Jessica Mar, Chii Mei Lin, Melvenia M Martin, Michael Ryan, et al. Allele-specific genome-wide profiling in human primary erythroblasts reveal replication program organization. *PLoS genetics*, 10(5):e1004319, 2014.

[160] Juan Carlos Rivera-Mulia, Andrew Dimond, Daniel Vera, Claudia Trevilla-Garcia, Takayo Sasaki, Jared Zimmerman, Catherine Dupont, Joost Gribnau, Peter Fraser, and David M Gilbert. Allele-specific control of replication timing and genome organization during development. *Genome Research*, pages gr–232561, 2018.

[161] Yang Yang, Quanquan Gu, Yang Zhang, Takayo Sasaki, Julianna Crivello, Rachel J O'Neill, David M Gilbert, and Jian Ma. Continuous-trait probabilistic model for comparing multi-species functional genomic data. *Cell Systems*, 7:208–218, 2018.

[162] Jian Ma, Louxin Zhang, Bernard B Suh, Brian J Raney, Richard C Burhans, W James Kent, Mathieu Blanchette, David Haussler, and Webb Miller. Reconstructing contiguous regions of an ancestral genome. *Genome Research*, 16(12):1557–1565, 2006.

[163] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[164] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

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

[166] Gilles Celeux, Florence Forbes, and Nathalie Peyrard. EM procedures using mean field-like approximations for markov model-based image segmentation. *Pattern Recognition*, 36(1):131–144, 2003.

[167] Jun Zhang. The mean field theory in EM procedures for markov random fields. *IEEE Transactions on Signal Processing*, 40(10):2570–2583, 1992.

[168] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[169] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

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

[171] Olga Veksler. *Efficient graph-based energy minimization methods in computer vision*. PhD thesis, Cornell University, USA, 1999.

[172] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized-via graph cuts? *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 26 (2):147–159, 2004.

[173] Hsun-Hua Chou, Toshiyuki Hayakawa, Sandra Diaz, Matthias Krings, Etty Indriati, Meave Leakey, Svante Paabo, Yoko Satta, Naoyuki Takahata, and Ajit Varki. Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proceedings of the National Academy of Sciences*, 99 (18):11736–11741, 2002.

[174] juicer provides a one-click system for analyzing loop-resolution hi-c experiments.

[175] Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7): 629–639, 1990.

[176] Ofir Pele and Michael Werman. The quadratic-chi histogram distance family. In *European Conference on Computer Vision*, pages 749–762. Springer, 2010.

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

[178] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[179] Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *European Conference on Computer Vision*, pages 705–718. Springer, 2008.

[180] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[181] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[182] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[183] Cyril Goutte, Peter Toft, Egill Rostrup, Finn Å Nielsen, and Lars Kai Hansen. On clustering fMRI time series. *NeuroImage*, 9(3):298–310, 1999.

[184] Elizabeth H Finn, Gianluca Pegoraro, Hugo B Brandão, Anne-Laure Valton, Marlies E Oomen, Job Dekker, Leonid Mirny, and Tom Misteli. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell*, 176(6):1502–1515, 2019.

[185] Yanxiao Zhang, Ting Li, Sebastian Preissl, Maria Luisa Amaral, Jonathan D Grinstein, Elie N Farah, Eugin Destici, Yunjiang Qiu, Rong Hu, Ah Young Lee, et al. Transcriptionally active herv-h retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics*, 51(9):1380–1388, 2019.

[186] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[187] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Cheneby, Shubhada R Kulkarni, Ge Tan, et al. Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46(D1): D260–D266, 2017.

[188] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[189] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[190] Mayank NK Choudhary, Ryan Z Friedman, Julia T Wang, Hyo Sik Jang, Xiaoyu Zhuo, and Ting Wang. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *Genome Biology*, 21(1):1–14, 2020.

[191] Ruochi Zhang, Yuchuan Wang, Yang Yang, Yang Zhang, and Jian Ma. Predicting CTCF-mediated chromatin loops using CTCF-MP. *Bioinformatics*, 34(13):i133–i141, 2018.

[192] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. Learning to explain: An information-theoretic perspective on model interpretation. *arXiv preprint arXiv:1802.07814*, 2018.

[193] C Maddison, A Mnih, and Y Teh. The concrete distribution: A continuous relaxation of discrete random variables. International Conference on Learning Representations, 2017.

[194] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[195] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2009.

[196] Xiangang Li and Xihong Wu. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4520–4524. IEEE, 2015.

[197] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[198] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint*

arXiv:1406.1078, 2014.

[199] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[200] François Chollet. keras. https://github.com/fchollet/keras, 2015.

[201] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org.

[202] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901.

[203] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[204] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[205] Te-Won Lee. Independent component analysis. In *Independent Component Analysis*, pages 27–66. Springer, 1998.

[206] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Confer-*

ence on Machine Learning, pages 689–696. ACM, 2009.

[207] Per Christian Hansen. The truncatedsvd as a method for regularization. *BIT Numerical Mathematics*, 27(4):534–553, 1987.

[208] Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

[209] Kathryn Woodfine, Heike Fiegler, David M Beare, John E Collins, Owen T McCann, Bryan D Young, Silvana Debernardi, Richard Mott, Ian Dunham, and Nigel P Carter. Replication timing of the human genome. *Human Molecular Genetics*, 13(2):191–202, 2004.

[210] Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259, 2017.

[211] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[212] Maximilian Haeussler, Ann S Zweig, Cath Tyner, Matthew L Speir, Kate R Rosenbloom, Brian J Raney, Christopher M Lee, Brian T Lee, Angie S Hinrichs, Jairo Navarro Gonzalez, et al. The ucsc genome browser database: 2019 update. *Nucleic Acids Research*, 47(D1):D853–D858, 2019.

[213] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

[214] Melissa J Hubisz, Katherine S Pollard, and Adam Siepel. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in bioinformatics*, 12(1): 41–51, 2011.

[215] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep net-

work training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[216] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

[217] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, 17(1):10–12, 2011.

[218] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[219] Felix Krueger and Simon R Andrews. SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Research*, 5, 2016.

[220] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*, 2016.

[221] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[222] David Freedman, Robert Pisani, and Roger Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[223] Ann Lehman, Norm O'Rourke, Larry Hatcher, and Edward Stepanski. *JMP for basic univariate and multivariate statistics: methods for researchers and social scientists*. Sas Institute, 2013.

[224] Kevin E O'Grady. Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, 92(3):766, 1982.

[225] O Heinisch. Steel, RGD, and JH Torrie: Principles and Procedures of Statistics.(With special Reference to the Biological Sciences.) McGraw-Hill Book Company, New York, Toronto, London 1960, 481 S., 15 Abb.; 81 s 6 d. *Biometrische Zeitschrift*, 4(3):207–208, 1962.

[226] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[227] Henry B Mann and Donald R Whitney. On a test of whether one of two random vari-

ables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60, 1947.

[228] Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O?shea, Peter J Park, Bing Ren, et al. The 4D nucleome project. *Nature*, 549(7671):219, 2017.

[229] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[230] Alvin C Rencher and William F Christensen. Chapter 10, multivariate regression–section 10.1, introduction. *Methods of multivariate analysis, Wiley Series in Probability and Statistics*, 709:19, 2012.

[231] Qin Cao, Zhenghao Zhang, Alexander Xi Fu, Qiong Wu, Tin-Lap Lee, Eric Lo, Alfred SL Cheng, Chao Cheng, Danny Leung, and Kevin Y Yip. A unified framework for integrative study of heterogeneous gene regulatory mechanisms. *Nature Machine Intelligence*, pages 1–10, 2020.

[232] Jill M Dowen, Zi Peng Fan, Denes Hnisz, Gang Ren, Brian J Abraham, Lyndon N Zhang, Abraham S Weintraub, Jurian Schuijers, Tong Ihn Lee, Keji Zhao, et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, 159(2):374–387, 2014.

[233] Aziz Khan and Xuegong Zhang. dbsuper: a database of super-enhancers in mouse and human genome. *Nucleic Acids Research*, 44(D1):D164–D171, 2016.

[234] Peiyao A Zhao, Takayo Sasaki, and David M Gilbert. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biology*, 21(1):1–20, 2020.

[235] Dario Acampora, Luca G Di Giovannantonio, and Antonio Simeone. Otx2 is an intrinsic determinant of the embryonic stem cell state and is required for transition to a stable epiblast stem cell condition. *Development*, 140(1):43–55, 2013.

[236] Christa Buecker, Rajini Srinivasan, Zhixiang Wu, Eliezer Calo, Dario Acampora, Tiago Faial, Antonio Simeone, Minjia Tan, Tomasz Swigut, and Joanna Wysocka.

Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell stem cell*, 14(6):838–853, 2014.

[237] Sayaka Sugiyama, Alain Prochiantz, and Takao K Hensch. From brain formation to plasticity: insights on Otx2 homeoprotein. *Development, growth & differentiation*, 51(3):369–377, 2009.

[238] Chi Kin Ip, Nicolas Fossat, Vanessa Jones, Thomas Lamonerie, and Patrick PL Tam. Head formation: OTX2 regulates Dkk1 and Lhx1 activity in the anterior mesendoderm. *Development*, 141(20):3859–3867, 2014.

[239] Amit S Verma and David R FitzPatrick. Anophthalmia and microphthalmia. *Orphanet journal of rare diseases*, 2(1):47, 2007.

[240] KF Schilter, Adele Schneider, Tanya Bardakjian, J-F Soucy, Rebecca C Tyler, Linda M Reis, and Elena V Semina. OTX2 microphthalmia syndrome: four novel mutations and delineation of a phenotype. *Clinical genetics*, 79(2):158–168, 2011.

[241] Dina Zielinski, Barak Markus, Mona Sheikh, Melissa Gymrek, Clement Chu, Marta Zaks, Balaji Srinivasan, Jodi D Hoffman, Dror Aizenbud, and Yaniv Erlich. Otx2 duplication is implicated in hemifacial microsomia. *PloS one*, 9(5):e96788, 2014.

[242] Tianshun Gao and Jiang Qian. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Research*, 48(D1):D58–D64, 2020.

[243] Jill E Moore, Michael J Purcaro, Henry E Pratt, Charles B Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, Rajinder Kaul, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

[244] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, 2009.

[245] Zihua Gong, Ja-Eun Kim, Charles Chung Yun Leung, JN Mark Glover, and Junjie

Chen. BACH1/FANCJ acts with TopBP1 and participates early in DNA replication checkpoint control. *Molecular Cell*, 37(3):438–446, 2010.

[246] Charles Chung Yun Leung, Zihua Gong, Junjie Chen, and JN Mark Glover. Molecular basis of BACH1/FANCJ recognition by TopBP1 in DNA replication checkpoint control. *Journal of Biological Chemistry*, 286(6):4292–4301, 2011.

[247] Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. Great improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5):495–501, 2010.

[248] Yu Chen, Yang Zhang, Yuchuan Wang, Liguo Zhang, Eva K Brinkman, Stephen A Adam, Robert Goldman, Bas Van Steensel, Jian Ma, and Andrew S Belmont. Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *Journal of Cell Biology*, 217(11):4025–4048, 2018.

[249] Bas van Steensel and Steven Henikoff. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nature Biotechnology*, 18 (4):424–428, 2000.

[250] Matthew Brand, Nuria Oliver, and Alex Pentland. Coupled hidden markov models for complex action recognition. In *CVPR*, pages 994–999. IEEE, 1997.

[251] Baoyuan Wu, Bao-Gang Hu, and Qiang Ji. A coupled hidden markov random field model for simultaneous face clustering and tracking in videos. *Pattern Recognition*, 64:361–373, 2017.

[252] Lingyun Song and Gregory E Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):pdb–prot5384, 2010.

[253] Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109(1):21–29, 2015.

[254] Muhammad Shoaib, David Walter, Peter J Gillespie, Fanny Izard, Birthe Fahrenkrog, David Lleres, Mads Lerdrup, Jens Vilstrup Johansen, Klaus Hansen,

Eric Julien, et al. Histone H4K20 methylation mediated chromatin compaction threshold ensures genome integrity by limiting DNA replication licensing. *Nature Communications*, 9(1):1–11, 2018.

[255] Yang Yang, Yang Zhang, Bing Ren, Jesse R Dixon, and Jian Ma. Comparing 3d genome organization in multiple species using phylo-hmrf. *Cell Systems*, 8(6):494–505, 2019.