Bayesian Inference of Spatial Organizations of Chromosomes

Ming Hu[1], Ke Deng[1], Zhaohui Qin[2], Jesse Dixon[3,4,5], Siddarth Selvaraj[3,6], Jennifer Fang[3], Bing Ren[3,7] & Jun S. Liu[1*]

[1]Department of Statistics, Harvard University, Cambridge, Massachusetts, USA. [2]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA. [3]Ludwig Institute for Cancer Research, La Jolla, California, USA. [4]Medical Scientist Training Program, University of California, San Diego, La Jolla, California, USA. [5]Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, California, USA. [6]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, California, USA. [7]University of California, San Diego School of Medicine, Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, UCSD Moores Cancer Center, La Jolla, California, USA.

*To whom correspondence should be addressed. Email: jliu@stat.harvard.edu

Running head: Spatial Organizations of Chromosomes

Abstract

Knowledge of spatial chromosomal organizations is critical for the study of transcriptional regulation and other nuclear processes in the cell. Recently, chromosome conformation capture (3C) based technologies, such as Hi-C and TCC, have been developed to provide a genome-wide, three-dimensional (3D) view of chromatin organization. Appropriate methods for analyzing these data and fully characterizing the 3D chromosomal structure and its structural variation are still under development. Here we describe a novel Bayesian probabilistic approach, denoted as "Bayesian 3D constructor for Hi-C data" (BACH), to infer the consensus 3D chromosomal structure within the cell population. In addition, we describe a variant algorithm BACH-MIX to study the structural variation of chromatin within the cell population. Applying BACH and BACH-MIX to a high resolution Hi-C dataset generated from mouse embryonic stem cells, we found that most local genomic regions exhibit homogeneous 3D chromosomal structures, which can be reconstructed by the BACH algorithm. Using BACH, we constructed a model for the spatial arrangement of chromatin that reveals structural properties associated with euchromatic and heterochromatic regions in the genome. We observed strong associations between structural properties and several genomic and epigenetic features of the chromosome. Using BACH-MIX, we further found that the structural variation of chromatin is correlated with these genomic and epigenetic features. Our results demonstrate that BACH and BACH-MIX have the potential to provide new insights into the chromosomal architecture of mammalian cells.

Author Summary

Understanding how chromosomes fold provides insights into the complex relationship among chromatin structure, gene activity and the functional state of the cell. Recently, chromosome conformation capture based technologies, such as Hi-C and TCC, have been developed to provide a genome-wide, high resolution and three-dimensional (3D) view of chromatin organization. However, statistical methods for analyzing these data are still under development. Here we propose two Bayesian approaches, BACH and BACH-MIX, to infer the consensus 3D chromosomal structure and the structural variation of chromatin within the cell population. Applying BACH and BACH-MIX to a high resolution Hi-C dataset, we found that most local genomic regions exhibit homogeneous 3D chromosomal structures. Furthermore, spatial properties of 3D chromosomal structures and the structural variation of chromatin are associated with several genomic and epigenetic features. Noticeably, gene rich, accessible and early replicated genomic regions tend to be more elongated and exhibit higher structural variation than gene poor, inaccessible and late replicated genomic regions.

Introduction

The spatial organization of a genome plays an important role in gene regulation, DNA replication, epigenetic modification and maintenance of genome stability [1-5]. Understanding three-dimensional (3D) chromosomal structures and chromatin interactions is therefore essential for decoding and interpreting functions of the genome. Traditionally, the 3D organization of chromosomes has been studied by microscopic and cytogenic methods such as florescent in situ hybridization (FISH). Several FISH studies have shown that the global 3D chromosomal structures are highly dynamic [6-8]. However, due to the limitation of low throughput, low resolution FISH data, the 3D chromosomal structures at the fine scale are not fully understood. In particular, whether chromatin exhibits a consensus local 3D chromosomal structure is still under debate. More recently, higher throughput, higher resolution approaches based on chromosome conformation capture (3C) such as Hi-C [9] and TCC [10] allow genome-wide mapping of chromatin interactions. The chromatin interactions captured by Hi-C and TCC experiments, which are represented by the contact matrix in the original Hi-C study [9], provide an unprecedented opportunity for inferring 3D chromosomal structures at the fine resolution scale.

Much progress has been made in recent years to reconstruct 3D chromosomal structures from the Hi-C data. Bau and colleagues [11] translated the read counts in the contact matrix to spatial constraints of 3D chromosomal structures and used the software Integrated Modeling Platform (IMP) [12] to solve a constrained optimization problem. Duan et al. [13] devised a set of constraints for all loci of the genome, and solved a similar constrained optimization problem using an open-source software IPOPT [14]. Similar optimization-based approaches have also been used in studies of the fission yeast genome [15] and the human genome [10]. More recently, Rousseau et al. [16] developed a probabilistic model linking Hi-C data to spatial distances and designed a Markov-chain Monte Carlo-based method named MCMC5C. Different from the optimization-based approaches, MCMC5C models the uncertainties of spatial distances between two loci by assuming that the number of reads spanning those two loci follows a Gaussian distribution.

However, all these existing methods have several limitations: first of all, as pointed out in a recent work by Yaffe and Tanay [17], the raw data obtained from Hi-C experiments exhibit multiple layers of systematic biases, such as restriction enzyme cutting frequencies, GC content and sequence uniqueness. None of these existing methods take these systematic biases into consideration. In addition, the optimization-based methods are prone to be trapped in local modes due to the ultra-high dimensionality and the prohibitively large search space. MCMC5C suffers from the difficulty in estimating the Gaussian variance of each read count since the single Hi-C contact matrix does not provide enough information for variance estimation. Furthermore, except for MCMC5C, none of these existing methods comes with stand-alone software [16].

More important, all of these existing methods focus on reconstructing the consensus 3D chromosomal structures, while pay little attention to evaluating the magnitude of the structural variation of chromatin at different resolution scales. To quantify the structural variation of chromatin, the optimization-based methods usually require a large number of parallel runs, which is not directly interpretable and computationally intensive. Similarly, the Gaussian model in MCMC5C is derived from a consensus 3D chromosomal structure, which cannot be directly used to measure the structural variation of chromatin either.

Since the chromatin interactions captured by Hi-C experiments come from a cell population instead of a single cell, it is challenging to study the structural variation of chromatin from the Hi-C data. When the cell population consists of multiple sub-populations, each of which corresponds to a distinct 3D chromosomal structure, the Hi-C data can only be interpreted as a measurement of population averaged effect. The Hi-C data of mammalian genomes is further complicated by the fact that the two homologues of the same chromosome cannot be distinguished from each other without genotype information. Without fully characterizing the structural variation of chromatin within the cell population, the consensus 3D chromosomal structure inferred from the Hi-C data is not directly interpretable or even misleading.

Although the global 3D chromosomal structure is indeed quite dynamic within the cell population, the local 3D chromosomal structure could be homogeneous. A recent study [18] on a high resolution Hi-C dataset has discovered that mammalian genomes are composed of thousands of 1 MB long, revolutionary conservative topological domains, which serve as individual units of genome function. These findings motivate the hypothesis that each topological domain may share a consensus 3D chromosomal structure in order to keep its conservative functional forms. For local genomic regions where this hypothesis holds true, the mixture of cell population and the ambiguity of two homologue chromosomes will not be the major barrier for 3D modeling based on Hi-C data anymore.

In this work, we test the hypothesis of consensus 3D structure at the topological domain scale from the computational point of view via a rigorous statistical analysis of Hi-C data. To achieve this goal, we propose two integrated probabilistic approaches called BACH (which is the short name for "**Ba**yesian 3D **C**onstructor for **H**i-C data") and BACH-MIX. It should be noted that our approach is broadly analogous to inferential structure determination (ISD) [19], a Bayesian approach developed to study macromolecular structure. In the BACH algorithm, we assume that the local genomic region (i.e., a topological domain) of interest exhibits a consensus 3D chromosomal structure within the cell population, and employ efficient Markov chain Monte Carlo (MCMC) computational algorithms to infer the underlying consensus 3D chromosomal structure. In the BACH-MIX algorithm, we assume that the genomic region of interest consists of multiple distinct 3D chromosomal structures, and explicitly model the structural variation of chromatin using a mixture component model. By comparing the goodness of fit of BACH and BACH-MIX for the same genomic region via statistical model selection principles, we provide a quantitative approach to evaluate the structural variation of chromatin for any given local genomic region.

Applying BACH and BACH-MIX to a high resolution Hi-C dataset, we found that BACH, instead of BACH-MIX, is preferred in about half of topological domains. For the topological domains in which BACH-MIX fits data better than BACH, most of them contain one dominant sub-population, and the 3D chromosomal structure of the dominant sub-population can be reconstructed by the BACH algorithm. These results suggest that most topological domains exhibit homogeneous 3D chromosomal structures within the cell population. We also found that geometrical properties of these topological domains, particularly the shape, are associated with several genomic and epigenetic features. Furthermore, we found significantly lower structural variation at domain center regions than at domain boundary regions. Additionally, the structural variation of chromatin is also related with several genomic and epigenetic features. Most important, our BACH and BACH-MIX

algorithms provide a solid statistical framework to infer the consensus 3D chromosomal structure and quantify the structural variation of chromatin within the cell population.

Results

The BACH algorithm

The BACH algorithm takes the chromosomal contact matrix generated by Hi-C or TCC experiments and local genomic features [17] (restriction enzyme cutting frequencies, GC content and sequence uniqueness) as input, and produces, via MCMC computation, the posterior distribution of 3D chromosomal structures (Methods). In the BACH algorithm, we assume that there exists a consensus 3D chromosomal structure within the cell population (this assumption will be relaxed later in the BACH-MIX algorithm). Furthermore, we assume that the number of sequencing reads spanning two genomic loci follows a Poisson distribution, where the Poisson rate is negatively associated with the corresponding spatial distance between them and is also affected by a few other factors. BACH can be used to reconstruct consensus 3D chromosomal structures from the Hi-C contact matrix, and infer the uncertainties of the spatial distance between any two genomic loci from the corresponding posterior distribution. Simulation studies have shown that the BACH algorithm works well under the posited model (Supplementary Information).

Compared to other published methods, BACH has the following advantages: (1) It integrates the correction of known systematic biases associated with Hi-C data, such as restriction enzyme cutting frequencies, GC content and sequence uniqueness [17]; (2) It utilizes a Poisson model that better fits the count data generated from Hi-C experiments than the Gaussian model used in MCMC5C, and performs more robustly when applied to several experimental datasets (see the following RESULTS section for validation); (3) It employs advanced MCMC techniques, such as Sequential Monte Carlo and Hybrid Monte Carlo (see Supplementary Information for details), that significantly improve the efficiency in exploring the vast space of possible models [20].

The BACH-MIX algorithm

In the BACH algorithm, we assume that chromosomal regions of interest exhibit a consensus 3D chromosomal structure within the cell population. However, this assumption may not be true, because chromosomal regions may exist in multiple inter-convertible configurations. To test the consensus 3D chromosomal structure assumption and study the structural variation of chromatin within the cell population, we propose a variant algorithm called BACH-MIX (Methods). In BACH-MIX, we assume that the genomic region of interest can be divided into two adjacent sub-regions, each with a consensus 3D chromosomal structure, but the relative position between the two adjacent sub-regions can vary within the cell population. BACH-MIX models the uncertainty of the relative position between the two adjacent sub-regions by a mixture component model, where each component corresponds to one specific relative position. The weight of each component represents the proportion of that component within the cell population. Clearly, BACH is a special case of BACH-MIX, where the cell population consists of a unique component. Applying BACH and BACH-MIX to the same Hi-C contact map and comparing the goodness of fit of two models via statistical principles of model selection, we provide a quantitative way to test the consensus 3D chromosomal structure assumption.

BACH-MIX contains two types of parameters: the parameters to determine the local 3D chromosomal structures of the two adjacent sub-regions, and the parameters to determine the relative position between the two adjacent sub-regions. In practice, the local 3D chromosomal structures of the two adjacent sub-regions

can be estimated by applying BACH twice separately, each to the contact map of one sub-region. The main computation in BACH-MIX is to estimate the parameters corresponding to the relative position between the two adjacent sub-regions.

The relative position between the two adjacent sub-regions can be represented by a rotation matrix with three Euler angles [21]. We also take into account mirror symmetry structures that cannot be explained by rotations. To simplify the computation, we discretize the range of each Euler angle into four bins of equal sizes, and approximate the collection of distinct 3D chromosomal structures within the cell population by 128 (i.e., 4^{3}) relative positions between two adjacent sub-regions. The BACH-MIX algorithm takes 3D chromosomal structures BACH predicted for two adjacent sub-regions and the corresponding local genomic features [17] (restriction enzyme cutting frequencies, GC content and sequence uniqueness) as input, and produces the posterior distribution of the spatial arrangement of the two sub-regions, quantified by the proportion of each of the 128 orientations between the two. Simulation studies have shown that the BACH-MIX algorithm works well under the posited model (Supplementary Information).

In practice, the majority of the 128 relative positions between the two adjacent sub-regions are ineffective in terms of having low proportions. To overcome the over-fitting caused by these ineffective relative positions, we adopt a two-step procedure: first, we apply the full BACH-MIX model with 128 relative positions to estimate the proportion for each of them; second, we remove the ineffective relative positions whose proportion is less than 1%, and re-estimate the proportion for the effective relative positions.

Most topological domains exhibit homogeneous 3D chromosomal structures

We applied BACH and BACH-MIX to a dataset recently generated in our lab [18] from a mouse embryonic stem cell (mESC) line. The dataset includes 476 million reads obtained from two biological replicates processed with the use of the restriction enzyme HindIII (referred to as the HindIII sample); and 237 million reads in another biological replicate processed with the use of the restriction enzyme NcoI (referred to as the NcoI sample). To the best of our knowledge, this dataset provides the highest sequencing depth of a mammalian genome to date. Previous analysis of this dataset showed that the mouse genome is composed of 2,200 topological domains characterized by high frequencies of intra-domain interactions but infrequent inter-domain interactions [18].

We conducted a genome-wide analysis by applying BACH and BACH-MIX to this ultra-high resolution mESC Hi-C dataset. Both BACH and BACH-MIX were applied to the 40 KB resolution Hi-C contact matrices. In the preprocessing procedure, we filtered out 300 topological domains whose length is less than 400 KB or do not contain known mouse gene (13.64% out of total 2,200 domains). We also filtered out a subset of 40 KB genomic loci within each topological domain according to restriction enzyme cutting frequencies (number of restriction enzyme cutting site <= 5), GC content (<= 0.3) and sequence uniqueness (mappability score <= 0.8) (Figure S1), and created the 40 KB resolution Hi-C contact matrix for each topological domain. We then applied BACH to each of these 1,900 topological domains to infer its 3D chromosomal structure.

To validate the spatial distances inferred by the BACH algorithm, we compared the spatial distances BACH predicted (referred to as the BACH distances) to the spatial distances measured by FISH [22] (referred to as the FISH distances). In the HindIII sample, the Pearson correlation coefficient between the BACH distances and the FISH distances is 0.88 (95% credible interval is [0.83, 0.92]). In the NcoI sample, the Pearson correlation coefficient between the BACH distances and the FISH distances is 0.83 (95% credible interval is [0.67, 0.93]). These results suggest that the spatial distances BACH predicted are consistent with the spatial

distances measured by FISH (Supplementary Information and Figure S2).

As a comparison, we applied MCMC5C and obtained the corresponding predictions of spatial distances (referred to as the MCMC5C distances). The Pearson correlation coefficients between the MCMC5C distances and the FISH distances are 0.79 and 0.11 in the HindIII sample and the NcoI sample, respectively, which are much worse than those of the BACH's results (z-test p-values < 2.4e-5). Since MCMC5C does not remove systematic biases (restriction enzyme cutting frequencies, GC content and sequence uniqueness), we also applied a modified BACH algorithm without bias correction (denoted as BACH-SUB) and obtained the corresponding predictions of spatial distances (referred to as the BACH-SUB distances). The Pearson correlation coefficients between the BACH-SUB distances and the FISH distance are 0.87 (95% credible interval is [0.81, 0.92]) and 0.18 (95% credible interval is [0.02, 0.30]) in the HindIII sample and the NcoI sample, respectively. BACH-SUB significantly outperforms MCMC5C in the HindIII sample (z-test p-value = 0.0004), and is comparable with MCMC5C in the NcoI sample (z-test p-value = 0.1669). All these results suggest that the Poisson model used in the BACH algorithm fits the count data generated by the Hi-C experiment better than the Gaussian model used in MCMC5C. In addition, integrating the correction of known systematic biases [17] further improves the reproducibility for the predicted spatial distances.

In the previous analysis, we obtained the 3D chromosomal structure predicted by BACH for each topological domain. Next, we equally divided each topological domain into two sub-regions, and applied BACH-MIX to infer the spatial arrangement of the two sub-regions. We evaluated the goodness of fit of the BACH model and the BACH-MIX model for each of these 1,900 topological domains in terms of Akaike information criterion (AIC) [23], which penalizes the log-likelihood of a model with the number of parameters in the model. Smaller AIC indicates better model fitting. In the HindIII sample, BACH achieved smaller AIC than BACH-MIX in 945 out of 1,900 (49.74%) topological domains. For the rest 955 topological domains where BACH-MIX fits data better than BACH, 551 topological domains have one dominant spatial arrangement of the two sub-regions with proportion larger than 80%. In 547 out of these 551 topological domains, the dominant 3D chromosomal structure can be captured by BACH. Therefore, BACH can reconstruct the consensus structure or the dominant structure in 1,492 topological domains (78.53% of 1,900 topological domains). We obtained consistent results in the NcoI sample. In the NcoI sample, BACH achieved smaller AIC than BACH-MIX in 1,199 out of 1,900 (63.11%) topological domains. For the rest 701 topological domains where BACH-MIX fits data better than BACH, 446 topological domains have one dominant spatial arrangement of the two sub-regions with proportion larger than 80%. In 442 out of these 446 topological domains, the dominant 3D chromosomal structure can be captured by BACH. Therefore, BACH can reconstruct the consensus structure or the dominant structure in 1,641 topological domains (86.37% of 1,900 topological domains). These results suggest around half topological domains exhibit consensus 3D chromosomal structures. For topological domains in which the consensus 3D chromosomal structure assumption is invalid, more than half of them contain one dominant sub-population with mixture proportion larger than 80%, and the 3D chromosomal structure corresponding to the dominant sub-population can be reconstructed by the BACH algorithm.

Structural properties of topological domains correlate with genomic and epigenetic features

In the following analysis, we focus on 1,342 (the overlap between 1,492 topological domains in the HindIII sample and 1,641 topological domains in the NcoI sample, 70.63% out of 1,900) topological domains in which BACH can reconstruct the consensus 3D chromosomal structure or the 3D chromosomal structure of the dominant sub-population in both HindIII sample and NcoI sample. To summarize the structural properties of topological domains, we approximated each 3D chromosomal structure BACH predicted (40 KB resolution) by a cylinder, and computed the ratio between its height and diameter, abbreviated as *HD ratio*

(Methods). Domains with higher HD ratios are more elongated. HD ratios of the structures inferred from the HindIII sample and the NcoI sample are highly reproducible (Pearson correlation coefficients = 0.78, p-value < 2.2e-16).

To evaluate the relationship between structural properties of chromatin (measured by HD ratio) and its functional forms at the topological domain scale, we collected genomic and epigenetic features for each topological domain, including gene density (UCSC reference genome mm9), gene expression [24], five histone modification marks (H3K36me3 [25], H3K27me3 [26], H3K4me3 [24], H3K9me3 [27] and H4K20me3 [26]), RNA polymerase II [24], chromatin accessibility [28], genome-nuclear lamina interaction [29] and DNA replication time [30]. By computing the correlation between the HD ratio and each of the genomic and epigenetic features, we found that the HD ratio is positively associated with gene density, gene expression, transcription elongation histone modification mark H3K36me3, repressive histone modification mark H3K27me3, promoter mark H3K4me3, RNA polymerase II, accessible chromatin and early replicated genomic regions, and negatively associated with heterochromatin marks H3K9me3, H4K20me3 and lamina associated domains (Table S1). These correlations are similar computed based on either the HindIII sample or the NcoI sample. Two illustrative examples are shown in Figure 1 and Table S2. Consistent with other existing biological evidences, these results demonstrate that the gene rich, actively transcribed, accessible, and early replicated chromatin tends to be more elongated than the gene poor, lowly transcribed, inaccessible and late replicated chromatin, which is consistent with previous FISH experiments [31].

The original Hi-C study [9] has shown the relationship between chromatin interactions and the genomic and epigenetic features. Lieberman-Aiden et al. [9] demonstrated by applying the principle component analysis (PCA) method to the Hi-C data that two compartments (compartment A and compartment B) in the mammalian genome can be obtained, where compartment A is strongly associated with open, accessible, and actively transcribed chromatin [32]. Following their strategy, we also applied the PCA method to the mESC Hi-C dataset [18] and obtained the compartment label (A or B) for each topological domain. While interesting, the compartment label (A or B) does not provide an intuitive visual interpretation of the data and cannot reflect the systematic biases and noise sources in the data discussed previously. We further evaluate the structural properties of topological domains (measured by HD ratios) in two different compartments. We found that topological domains in compartment A are significantly more elongated than those in compartment B. In the HindIII sample, mean HD ratios for domains in compartment A and compartment B are 1.73 and 1.33, respectively (two sample t-test p-value < 2.2e-16). Similarly, in the NcoI sample, mean HD ratios for domains in compartment A and compartment B are 1.72 and 1.24, respectively (p-value < 2.2e-16). Two illustrative examples are shown in Figure 1 and Table S2. These results suggest that the HD ratio obtained in the BACH algorithm provides an intuitive visual interpretation of the Hi-C data.

Structural variations of topological domains correlate with genomic and epigenetic features

We further study the structural variation of chromatin within the cell population. We first select3d 562 topological domains with size larger than 1 MB, and applied BACH and BACH-MIX to the 1 MB region around the center of each selected domain center region. Additionally, we used 985 domain boundaries with size shorter than 40 KB as the control group, and applied BACH and BACH-MIX to the 1 MB region around each selected domain boundary region. We equally divided each 1 MB genomic region (domain center / boundary region) into two 500 KB adjacent sub-regions with fixed 3D chromosomal structures BACH predicted, and then inferred the spatial arrangements between the two. Both BACH and BACH-MIX were applied to the 40 KB resolution Hi-C contact matrices.

Among all the possible spatial arrangements of two adjacent genomic regions, we define the effective

structures as those with their posterior mean proportions greater than 5%, and report the number of effective structures at each locus. A locus with a smaller number of effective structures exhibits lower structural variation than a locus with a larger number of effective structures. In the HindIII sample, the average number of effective structure is 1.84 for the domain center regions, and 2.23 for the domain boundary regions (Figure S3A, two sample t-test p-value = 6.6e-13). Similarly, in the NcoI sample, the average number of effective structures is 1.72 for the domain center regions, and 2.02 for the domain boundary regions (Figure S3B, two sample t-test p-value = 1.2e-8). We changed the threshold for the effective structure to 10% and 1%, and observed consistent results (Figure S3 and Table S3). These results suggest that domain center regions exhibit lower structural variation than domain boundary regions.

Figure 2 shows two illustrative examples in the HindIII sample, one for the domain center region (Chromosome 2, 117,580,000~118,580,000, Figure 2A), and one for the domain boundary region (Chromosome 1, 135,540,000~136,540,000, Figure 2B). Under threshold 5%, BACH-MIX identified one effective structure for the domain center region with proportion 98% (Figure 2C), and three effective structures for the domain boundary region, with proportions 45% (Figure 2D), 31% (Figure 2E) and 24% (Figure 2F), respectively.

Next, we evaluated the relationship between the structural variation of topological domains and its functional forms. We divided the 562 selected domain center regions into two groups, regions with high structural variation (i.e., containing multiple effective structures, threshold = 5%) and regions with low structural variation (i.e., containing one effective structure, threshold = 5%), and compared the genomic and epigenetic features between these two groups (Table S4). We observed significant enrichment of gene density, transcription elongation histone modification mark H3K36me3, repressive histone modification mark H3K27me3, promoter mark H3K4me3, RNA polymerase II, accessible chromatin and early replicated genomic regions in regions with high structural variation, and significant enrichment of heterochromatin marks H3K9me3, H4K20me3 and genome-nuclear lamina interaction in regions with low structural variation. Noticeably, we did not observe significant association between gene expression and the structural variation. These results suggest that gene rich, accessible and early replicated chromatins are more likely to exhibit multiple structural configurations than gene poor, inaccessible and late replicated chromatins.

3D chromosomal structure model for the whole chromosome

Due to the structural variation of chromatin within the cell population, inferring 3D structure for the whole chromosome is challenging and difficult to interpret. However, after taking a close look at the whole chromosome Hi-C contact matrix, we found that for some long chromosomes, the cell population contains one dominant sub-population, and the 3D chromosomal structure of this dominant sub-population can be reconstructed by the BACH algorithm.

Our reasoning goes as follows. Assume that the cell population consists of several sub-populations, each of which exhibits a distinct 3D chromosomal structure. If there does exist a dominant sub-population (referred to as the dominant structure *)* within the cell population (e.g., with a proportion larger than 80%), then the 3D structure BACH predicted (referred to as *)* should be quite close to the structure of . If we subtract a large proportion (for example 50%) of read counts corresponding to *from the original Hi-C contact map,* and run BACH one more time on the residual contact map, we expect to get a similar 3D structure from the second run of BACH, because *is still* the 3D chromosomal structure of the dominant sub-population within the cell population with respect to the residual contact map. On the other hand, if there is not such a dominant sub-population in the cell population, the two 3D chromosomal structures BACH predicted in the two stages, and , should be very different from each other. Thus, the similarity between *and (*which can be measured by the

RMSD) serves as a measurement of homogeneity of the cell population. A smaller RMSD indicates a more homogeneous cell population.

In practice, we adopted the following two-step procedure: in the first step, we applied BACH to the input Hi-C contact matrix and obtained the first BACH predicted 3D chromosomal structure and the expected Hi-C contact matrix. In the second step, we defined the residual matrix as the difference between the input Hi-C contact matrix and half of the expected Hi-C contact matrix. We applied BACH to the residual matrix and obtained the second BACH predicted 3D chromosomal structure. We conducted a series of simulation studies to evaluate the performance for the above two-step procedure when the input Hi-C contact matrix is generated from a mixture population. The simulation results (Supplementary Information, Figure S4 and Table S5) are consistent to our expectation: when the mixture population contains one dominant sub-population (proportion >= 80%), the two 3D chromosomal structures BACH predicted in the two stages, *and , show high similarity (low RMSD), and are both close to the 3D chromosomal structure of the dominant sub-population.* In contrast, when the mixture population contains two sub-populations with comparable proportions (the dominant sub-population with proportion <= 70%), *and show high discrepancy, and both are different from either of the two underlying simulated 3D chromosomal structures.*

We applied the same two-step procedure to the real Hi-C data to generate 3D chromosomal structure for each mouse chromosome by treating each topological domain as an individual unit. Figure S5 lists the alignment of two 3D chromosomal structures BACH predicted in the two stages, and *, from 20 mouse chromosomes in both HindIII sample and NcoI sample.* To quantify the magnitude of the discrepancy between *and for each chromosome with given size, we used* the random walk scheme (Supplementary Information) to simulate two 3D chromosomal structures with the same size, and then calculated the RMSD between them. We repeated the simulation 1,000 times to obtain the null distribution of RMSD. Figure S6 lists the null distribution of RMSD for different chromosomes. Based on the null distribution of RMSD, we calculated the p-value of RMSD(*,) for each chromosome (Table S6*). We found that in long chromosomes (chromosome 1~14 and chromosome X), *and are similar (*RMSD(*,) is small, p-value <= 0.05), which indicates these long chromosomes may* exhibit a dominant sub-population within the cell population. In contrast, in short chromosomes (chromosome 15~19), *and are different (*RMSD(*,) is large, p-value > 0.05), which indicates these short chromosomes may* exhibit multiple distinct sub-populations with comparable mixture proportions within the cell population. These results are consistent in both HindIII sample and NcoI sample.

To control the chromosome size, we conducted the following control experiments. We first zoomed in the Hi-C contact matrix of each chromosome by equally splitting one topological domain into two sub-domains, and then treated each sub-domain as an individual unit. In addition, we zoomed out the Hi-C contact matrix of each chromosome by merging two adjacent topological domains into one super-domain, and then treated each super-domain as an individual unit. We applied the previous two-step procedure to the zoom-in and zoom-out Hi-C contact matrices, and reported the RMSD(*,)* and corresponding p-value in Table S7 and Table S8. We found consistent results in zoom-in and zoom-out Hi-C contact matrices: RMSD(*,) is small* (p-value <= 0.05) in most of long chromosomes, and large (p-value > 0.05) in most of short chromosomes. We also aligned the 3D chromosomal structure inferred from the zoom-in and zoom-out Hi-C contact matrices to the 3D chromosomal structure inferred from the original Hi-C contact matrices, and found that the 3D chromosomal structures inferred at different resolution scales show high level of similarity (Table S9 and Table S10). Furthermore, we equally split each chromosome into two halves, and applied the previous two-step procedure to each half chromosome, treating each topological domain as an individual unit. We found that RMSD(*,) is large (p-value > 0.05) in most chromosomes (*Table S11). In most chromosomes, the 3D chromosomal structure inferred from the half chromosome aligned well with the 3D chromosomal structure

inferred from the whole chromosome (Table S12). All these results suggest that long chromosomes may exhibit a dominant sub-population within the cell population, and short chromosomes may exhibit multiple distinct sub-populations with comparable mixture proportions within the cell population. These conclusions are consistent at different resolution scales, and are not affected by the size of input Hi-C contact matrix. Furthermore, as suggested in the simulation studies, the BACH algorithm is able to reconstruct the 3D chromosomal structure of the dominant sub-population for long chromosomes. In the following analysis, we focus on the 3D chromosomal structure at the whole chromosome scale for the long chromosomes (chromosome 1~14 and chromosome X).

We used BACH, BACH-SUB and MCMC5C to generate spatial models of each long chromosome (chromosome 1~14 and chromosome X) by treating each topological domain as an individual unit (Figure S7). The 3D chromosomal structures predicted by BACH from the HindIII sample and NcoI sample achieved significantly higher reproducibility (measured by the normalized root mean square deviations, i.e., RMSD, Methods) than those predicted by MCMC5C (paired t-test p-value = 1.4e-7). Similarly, BACH-SUB also achieved significantly higher reproducibility (measured by the normalized root mean square deviations, i.e., RMSD, Methods) than those predicted by MCMC5C (paired t-test p-value = 0.0465). We also conducted the same analysis using a published human Hi-C dataset [9] and found that BACH and BACH-SUB consistently outperformed MCMC5C (data not shown). Since other published 3D reconstruction methods do not provide stand-alone software, we were not able to conduct similar comparative studies for them. The significant improvement of BACH over MCMC5C is due to that BACH explicitly integrates the correction of known systematic biases [17]. Without biases correction, BACH-SUB still outperforms MCMC5C, which indicates that the Poisson model used in BACH fits the count data better that the Gaussian model used in MCMC5C.

Structural properties of long chromosomes correlate with genomic and epigenetic features

We applied the BACH algorithm to the whole chromosome Hi-C contact matrix, and obtained the predicted 3D chromosomal structures for long chromosomes. We first investigate how compartment A and compartment B are distributed spatially in the whole chromosome model. Among all 2,200 topological domains, 1,026 domains belong to compartment A, 752 domains belong to compartment B, and the rest 422 domains, referred to as straddle domains, contain genomic regions from both compartment A and compartment B. For each 3D chromosomal structure BACH predicted, we fitted a plane through the straddle domains using the least square method, and then counted the numbers of topological domains belonging to compartment A and compartment B, respectively, at each side of the fitted plane, which can be denoted by a two-by-two contingency table. Fisher's exact test was used to measure the magnitude of spatial separations between two compartments. Among the selected 15 mouse chromosomes (chromosome 1~14 and chromosome X), we found that the compartment label (A or B) of topological domains is significantly associated with the spatial location of topological domains relative to the fitted plane (on the left side or on the right side) in 14 chromosomes in both HindIII sample and NcoI sample (Table S13). As shown in Figure 3A, topological domains with the same compartment label tend to locate on the same side of the structure, consistent with their interaction frequencies, and the observation that compartment B tends to be associated with nuclear membrane [33,34].

We further study how genomic and epigenetic features are distributed spatially in the whole chromosome model. Similar to the previous analysis for compartment labels (A or B), we conducted the same analysis for each of the eleven genomic and epigenetic features in consideration (Table S13). We used 33rd and 67th percentiles as the thresholds and divided all 2,200 topological domains into three groups: domains with low value, with median value, and with high value of the feature. For each 3D chromosomal structure BACH predicted, we fitted a plane through domains with median value of the feature using the least square method.

Next, we used the Fisher's exact test p-value to measure the magnitude of association between the group label (low value group or high value group) and spatial location of topological domains relative to the fitted plane (on the left side or on the right side). Table S13 lists the number of chromosomes with significant spatial separation patterns for each genomic and epigenetic feature in both HindIII sample and NcoI sample (threshold for Fisher's exact test p-value is 0.05). We observed that the gene density, transcription elongation histone modification mark H3K36me3, repressive histone modification mark H3K27me3, promoter mark H3K4me3, RNA polymerase II, chromatin accessibility, DNA replication time, heterochromatin marks H3K9me3 and H4K20me3 and genome-nuclear lamina interaction of topological domains are significantly associated with the spatial location of topological domains relative to the fitted plane (on the left side or on the right side) among more than ten chromosomes (Table S13 and Figure 3B ~ Figure 3L).

Discussion

In this paper, we describe BACH and BACH-MIX, two Bayesian statistical models, to study 3D chromosomal structures and the structural variation of chromatin from the Hi-C data. The benefits of using a probabilistic approach are two-folds: first, rigorous statistical inference can be carried out to properly remove systematic biases; second, sequencing depth variations can be explicitly modeled by the embedded Poisson distribution. Our results demonstrate that BACH is more reproducible and more accurate than an existing algorithm (MCMC5C). Application of BACH to a recently published Hi-C dataset from the mouse ES cells revealed structural properties of the mammalian chromosomes. Specifically, we found that the geometric shape of topological domains is strongly correlated with several genomic and epigenetic features, for example, the gene rich, actively transcribed, accessible and early replicated chromatin tends to be more elongated than the gene poor, lowly transcribed, inaccessible and late replicated chromatin. Furthermore, by using a variant BACH-MIX algorithm, we found that the structural variation of chromatin is also related with several genomic and epigenetic features.

There are several issues that we have not addressed in this paper, such as the biophysical properties of chromatin fiber [35,36] and the low sequencing depth of inter-chromosomal chromatin interactions. In principle, the biophysical properties can be accommodated directly in our Bayesian model as spatial constraints through an informative prior on spatial distances. With more experimental work and additional data, the BACH and BACH-MIX algorithms can be applied to study the spatial arrangement of multiple chromosomes simultaneously. With the rapid accumulation of high throughput genome-wide chromatin interaction data, the BACH and BACH-MIX algorithms could be valuable tools for understanding higher order chromatin architecture of mammalian cells.

Methods

The BACH algorithm

To reconstruct the underlying 3D chromosomal structure, we develop the following probabilistic model, similar to the "beads-and-string" model (Figure S8) that has been used extensively in chemistry. The genomic region of interest is divided into

consecutive, disjoint loci of equal size

, and each locus

is represented by a bead in the 3D space, whose location is given by the Cartesian coordinates

. The Euclidean distance

between loci

and

represents the spatial distance between these two loci

and

:

.

Under this representation, reconstructing the 3D chromosomal structure is equivalent to placing these beads in the 3D space, i.e., specifying the Cartesian coordinates

of these loci.

Let

be the

symmetric contact matrix generated by the Hi-C experiment, where each entry

represents the number of paired-end reads spanning two loci

and

. The variation of

can be explained by several factors. Lieberman-Aiden et al. [9] first reported the negative association between the number of paired-end reads spanning two loci (

) and the corresponding spatial distance (

). Recently, Yaffe and Tanay [17] identified some systematic biases, including restriction enzyme cutting frequencies, GC content and sequence uniqueness of fragment ends, which substantially affect Hi-C data. Taking all these unique features into consideration, we propose the following Poisson model.

Let

,

and

represent the number of fragment ends within locus

, the mean GC content of fragment ends within locus

, and the mean mappability score of fragment ends within locus

, respectively [17]. We assume that the off-diagonal count

in the contact matrix

follows a Poisson distribution with rate

, where:

In this model,

measures the magnitude of negative association (

) between

and

.

,

and

are the coefficients for the enzyme effect, GC content effect and mappability effect, respectively. Let

(

matrix) represent the Cartesian coordinates of the

loci of interest, and let

be the collection of all nuisance parameters. The joint likelihood is of the form:

We adopt a fully Bayesian approach with non-informative priors for all model parameters, and obtain the following joint posterior distribution:

Due to the high dimensionality of the parameter space, designing an efficient computational tool to draw samples from

is essential for the statistical inference of our model. To achieve this goal, we propose a three-stage statistical inference procedure (Figure S9). First we assign initial values for the nuisance parameters using a Poisson regression approach [37]. We then use sequential importance sampling (SIS) [38] to generate an initial 3D chromosomal structure. At the end, we apply Gibbs sampler [39] with hybrid Monte Carlo [20,40] and

adaptive rejection sampling (ARS) [41] to further refine the 3D chromosomal structure and the nuisance parameters. More details of three-stage statistical inference procedure can be found in Supplementary Information.

HD ratio

Let

represent the Cartesian coordinates of the genomic region

with

loci, where

. First we shift the genomic region

such that its weight center is at the original point

. We then conduct the principle component analysis on the

by 3 matrix

, and rotate matrix

to matrix

,

, such that the x-axis is the direction of the first principle component (the one explains most variability) and the y-axis and the z-axis are the directions of the second and the third principle components, respectively. We use a cylinder to approximate the 3D chromosomal structure of the genomic region

. The height of the cylinder is defined as the difference between the 90% quantile of

and the 10% quantile of

. The radius of the cylinder is defined as two times the median of

. We further define HD ratio of the genomic region

as the ratio between the height of the cylinder and the diameter of the cylinder, and then normalized by the size of genomic region

. By the definition, genomic regions with higher HD ratios are more elongated.

The BACH-MIX algorithm

We propose the BACH-MIX algorithm to study the relative position of two adjacent genomic regions. Here we assume that each genomic region exhibits a unique consensus 3D chromosomal structure, but the relative

position of two adjacent genomic regions has certain level of flexibility, and varies according to a probabilistic distribution. More precisely, let

and

represent the 3D chromosomal structures of two adjacent genomic region

and

, respectively, where

. The relative position of the genomic region

with respect to the genomic region

is determined by three Euler angles [21]

and an index

for mirror symmetry. Let

be the collection of these four parameters, and define the rotation matrix

and the mirror symmetry matrix

as:

The relative position of the genomic region

with respect to the genomic region

, denoted

, can be calculated by:

Therefore, each

corresponds to a 3D chromosomal structure of two adjacent genomic regions

and

, and a probabilistic distribution

defines a mixture of distinct relative positions between the two adjacent genomic regions

and

. To further simplify the statistical inference problem on

, we discretize the four dimensional space of

, and use a multinomial distribution

to approximate

.

Let

be the

by

dimensional contact matrix for inter-region chromatin interactions, where

represent the number of reads spanning the

th locus in the genomic region

and the

th locus in the genomic region

. We assume that

follow Poisson distribution with rate

, where

Here

is the Poisson offset for the relative position

, which is proportional to

. The statistical inference problem on the multinomial distribution

is equivalent to infer

.

is the spatial distance between the

th locus in the genomic region

and the

th locus in the genomic region

with rotation matrix

and mirror symmetry matrix

.

,

and

are local genomic features which follow the previous definitions. The joint likelihood is of form:

We adopt a fully Bayesian approach with non-informative priors for all model parameters, and obtain the following joint posterior distribution:

We use hybrid Monte Carlo to jointly update the parameters

(Figure S9). The first order partial derivatives with respect to

is of the form:

Normalized root mean square deviation (RMSD)

Assuming

and

are the Cartesian coordinates of two genomic regions

and

, respectively, where

and

. We first remove the scaling effect by a regression procedure. Let

and

be the Euclidean distance between loci

and

in

and

, respectively. We regress

against

and obtain the slope

. Define

, where

. Assume

has the singular value decomposition

, and then the optimal rotation matrix

can minimize the sum of square error

[42]. The normalized RMSD is defined as:

Empirically, normalized RMSD less than 0.1 indicates high similarity, normalized RMSD between 0.1 and 0.2 indicates moderate similarity, while normalized RMSD larger than 0.2 indicates low similarity.

Model implementation

Under the default setting of BACH, we draw 100 3D chromosomal structures at each step of sequential importance sampling. We further enrich each 3D chromosomal structure ten times when we implement the rejection control technique. In the Gibbs sampler of BACH and BACH-MIX, we run three parallel chains with 5,000 MCMC iterations in each chain. The first 1,000 samples are dropped as the burn-in stage, and then every 50^{th} sample in the last 4,000 samples are used for the posterior inference. We use the Gelman-Rubin statistic [39] to measure the mixing of three parallel chains. Empirically, the Gelman-Rubin statistics less than 1.1 indicates that three parallel chains converge to the same posterior distribution.

Computation time

The computation time of BACH and BACH-MIX depends on the number of MCMC iterations and the number of loci in the genomic region of interest. All MCMC calculations are conducted on computing nodes in Harvard Linux cluster "Odyssey", each with dual Xeon E5410 2.3GHz quad core processors and 32GB RAM. Under the default setting, BACH takes 81 seconds to predict a 3D chromosomal structure with 25 loci; BACH-MIX takes 22 minutes to predict the proportion of 128 distinct 3D chromosomal structures for two 13 loci adjacent genomic regions. The computation time increases almost quadratically with the number of loci in the genomic region of interest.

URL

BACH and BACH-MIX can be freely downloaded at http://www.fas.harvard.edu/~junliu/BACH/.

Funding

This work was supported by the Ludwig Institute for Cancer Research (B.R), US National Institutes of Health grants R01HG005119 (Z.Q), R01HG003991 (B.R) and 5R01GM080625 (J.S.L).

Author Contributions

Conceived and performed the analysis: M.H, K.D, Z.Q, J.D, S.S, J.F, B.R and J.S.L. Wrote the paper: M.H, K.D, Z.Q and J.S.L.

Competing Interests

The authors have declared no competing interests exist.

References

1. Dekker J (2008) Gene regulation in the third dimension. Science 319: 1793-1794.

2. Fraser P, Bickmore W (2007) Nuclear organization of the genome and the potential for gene regulation. Nature 447: 413-417.

3. Miele A, Dekker J (2008) Long-range chromosomal interactions and gene regulation. Mol Biosyst 4: 1046-1057.

4. Misteli T (2007) Beyond the sequence: cellular organization of genome function. Cell 128: 787-800.

5. Misteli T (2004) Spatial positioning; a new dimension in genome function. Cell 119: 153-156.

6. Gasser SM (2002) Visualizing chromatin dynamics in interphase nuclei. Science 296: 1412-1416.

7. Lanctot C, Cheutin T, Cremer M, Cavalli G, Cremer T (2007) Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nat Rev Genet 8: 104-115.

8. Gerlich D, Beaudouin J, Kalbfuss B, Daigle N, Eils R, et al. (2003) Global chromosome positions are transmitted through mitosis in mammalian cells. Cell 112: 751-764.

9. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289-293.

10. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2011) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol.

11. Bau D, Sanyal A, Lajoie BR, Capriotti E, Byron M, et al. (2011) The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol 18: 107-114.

12. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, et al. (2007) Determining the architectures of macromolecular assemblies. Nature 450: 683-694.

13. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, et al. (2010) A three-dimensional model of the yeast genome. Nature 465: 363-367.

14. Wachter A, Biegler LT (2006) On the Implementation of an Interior-Point Filter Line-Search Algorithm for Large-Scale Nonlinear Programming. Mathematical Programming 106: 25-27.

15. Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, et al. (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Res 38: 8164-8177.

16. Rousseau M, Fraser J, Ferraiuolo MA, Dostie J, Blanchette M (2011) Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. BMC Bioinformatics 12: 414.

17. Yaffe E, Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet 43: 1059-1065.

18. Dixon J, Selvaraj S, Yue F, Kim A, Li Y, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature In press.

19. Rieping W, Habeck M, Nilges M (2005) Inferential structure determination. Science 309: 303-306.

20. Liu J (2001) Monte Carlo Strategies in scientific computing. New York: Springer-Verlag.

21. Beard DA, Schlick T (2001) Computational modeling predicts the structure and dynamics of chromatin fiber. Structure 9: 105-114.

22. Eskeland R, Leeb M, Grimes GR, Kress C, Boyle S, et al. (2010) Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. Mol Cell 38: 452-464.

23. Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic

Control 19: 716-723.

24. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, et al. (2012) A map of the cis-regulatory sequences in the mouse genome. Nature.

25. Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, et al. (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell 134: 521-533.

26. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

27. Bilodeau S, Kagey MH, Frampton GM, Rahl PB, Young RA (2009) SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes Dev 23: 2484-2489.

28. Schnetz MP, Handoko L, Akhtar-Zaidi B, Bartels CF, Pereira CF, et al. (2010) CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression. PLoS Genet 6: e1001023.

29. Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman S, Solovei I, et al. (2010) Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. Mol Cell 38: 603-613.

30. Hiratani I, Ryba T, Itoh M, Rathjen J, Kulik M, et al. (2009) Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res 20: 155-169.

31. Goetze S, Mateos-Langerak J, Gierman HJ, de Leeuw W, Giromus O, et al. (2007) The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. Mol Cell Biol 27: 4475-4487.

32. Zhang Y, McCord RP, Ho YJ, Lajoie BR, Hildebrand DG, et al. (2012) Spatial Organization of the Mouse Genome and Its Role in Recurrent Chromosomal Translocations. Cell.

33. Mekhail K, Moazed D (2010) The nuclear envelope in genome organization, expression and stability. Nat Rev Mol Cell Biol 11: 317-328.

34. van Steensel B, Dekker J (2010) Genomics tools for unraveling chromosome architecture. Nat Biotechnol 28: 1089-1095.

35. Bystricky K, Heun P, Gehlen L, Langowski J, Gasser SM (2004) Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. Proc Natl Acad Sci U S A 101: 16495-16500.

36. Amzallag A, Vaillant C, Jacob M, Unser M, Bednar J, et al. (2006) 3D reconstruction and comparison of shapes of DNA minicircles observed by cryo-electron microscopy. Nucleic Acids Res 34: e125.

37. McCullagh P, Nelder JA (1989) Generalized linear models: Chapman & Hall/CRC.

38. Liu JS, Chen R (1998) Sequential Monte-Carlo Methods For Dynamic-Systems. Journal of the American Statistical Association 93: 1032-1044.

39. Gelman A, Carlin JB, Stern HS, Rubin DB (1995) Bayesian data analysis. London: Chapman & Hall. xix, 526 p.

40. Duane S, Kennedy AD, Pendleton BJ, Roweth D (1987) Hybrid Monte-Carlo. Physics Letters B 195: 216-222.

41. Gilks W, Wild P (1992) Adaptive rejection sampling for Gibbs sampling. Applied Statistics 41: 337-348.

42. Arun KS, Huang TS, Blostein SD (1987) Least-squares fitting of two 3-d point sets. IEEE Trans Pattern Anal Mach Intell 9: 698-700.

Figure 1. Two illustrative examples of 3D models for two topological domains using BACH. Two illustrative examples in the HindIII sample: one for a more elongated 1 MB domain (chromosome 18, 33,960,000~34,960,000) belonging to compartment A, the other for a less elongated 1 MB domain (chromosome 7, 62,040,000~63,040,000) belonging to compartment B. In Figure 1B and Figure 1D, each sphere represents a 40 KB genomic region. All spheres are of equal size. In Figure 1B and Figure 1D, the x axis is direction of the first principle component. The diameters of two fitted cylinders (grey) are set to be one. The height of the fitted cylinder in Figure 1B is 1.89 times larger than that in Figure 1D. The rank in descending order among the selected 1,900 domains was used to measure the relative magnitudes of genomic and epigenetic features (Table S2). The more elongated 1 MB domain has a high gene density, high gene expression, high H3K36me3, high H3K4me3, high RNA polymerase II, high chromatin accessibility, early DNA replication time, low H3K9me3, low H4K20me3 and low genome-nuclear lamina interaction. The 3D chromosomal structure BACH predicted for this domain (Figure 1B) has a high HD ratio (HD ratio = 2.16, rank = 171). The less elongated 1 MB domain has a low gene density, low gene expression, low H3K36me3, low H3K4me3, low RNA polymerase II, low chromatin accessibility, late DNA replication time, high H3K9me3, high H4K20me3 and high genome-nuclear lamina interaction. The 3D chromosomal structure BACH predicted for this domain (Figure 1D) has a low HD ratio (HD ratio = 1.14, rank = 1,313). The more elongated 1 MB domain has median H3K27me3 signal, while the less elongated 1MB domain has low H3K27me3 signal. These results can be partially explained by the weak correlation between the HD ratio and H3K27me3 (Table S3**,** Pearson correlation coefficients = 0.18, p-value = 4.4e-16). They are also consistent with the results in the human Hi-C study demonstrating weak enrichment of H3K27me3 in compartment A [9].

(A) 40 KB resolution Hi-C contact matrix of a more elongated domain belonging to compartment A. The color scheme is proportional to Log2 read counts.

(B) The 3D chromosomal structure BACH predicted for the domain described in Figure 1A. HD ratio = 2.16.

(C) 40 KB resolution Hi-C contact matrix of a less elongated domain belonging to compartment B. The color scheme is proportional to Log2 read counts.

(D) The 3D chromosomal structure BACH predicted for the domain described in Figure 1C. HD ratio = 1.14.

A. **C.**

B. **D.**

Figure 2. Domain center regions exhibit lower structural variation than domain boundary regions. Two illustrative examples in the HindIII sample: one for the domain center region (Chromosome 2, 117,580,000~118,580,000) with low structural variation, the other for the domain boundary region (Chromosome 1, 135,540,000~136,540,000) with high structural variation**.** In Figure 2C ~ Figure 2F, each sphere represents a 40 KB genomic region. All spheres are of equal size.

(A) 40 KB resolution Hi-C contact maps of a 1 MB domain center region in the HindIII sample. The color scheme is proportional to Log2 read counts.

(B) 40 KB resolution Hi-C contact maps of a 1 MB domain boundary region in the HindIII sample. The color scheme is proportional to Log2 read counts.

(C) The effective structure BACH-MIX predicted (proportion = 0.98) for the domain center region. Red spheres and lines represent the bottom left region in Figure 2A, blue spheres and lines represent the top right region in Figure 2A.

(D) The first effective structure BACH-MIX predicted (proportion = 0.45) for the domain boundary region. Red spheres and lines represent the bottom left region in Figure 2B, blue spheres and lines represent the top right region in Figure 2B.

(E) The second effective structure BACH-MIX predicted (proportion = 0.31) for the domain boundary region. Red spheres and lines represent the bottom left region in Figure 2B, purple spheres and lines represent the top right region in Figure 2B.

(F) The third effective structure BACH-MIX predicted (proportion = 0.24) for the domain boundary region. Red spheres and lines represent the bottom left region in Figure 2B, green spheres and lines represent the top right region in Figure 2B.

A. **B.**

C. **D.**

E. **F.**

Figure 3. Spatial organization of genomic and epigenetic features. We used the 3D chromosomal structure BACH predicted for chromosome 2 in the HindIII sample as an illustrative example. In Figure 2A ~ Figure 2L, each sphere represent a topological domain. The volume of each sphere is proportional to the genomic size of the corresponding topological domain. In Figure 2A, the red, white and blue colors represent topological domains belonging to compartment A, straddle region and compartment B, respectively. Topological domains with the same compartment label tend to locate on the same side of the structure. In Figure 2B ~ Figure 2L, the red, white and blue colors represent topological domains with high value of features, median value of features and low value of features, respectively. The color scheme is proportional to the magnitude of the continuous measurement of genetic and epigenetic features. We also report the odds ratio (OR) of the two by two contingency table and the p-value of Fisher's exact test.

(A) Spatial organization of compartment label. OR = 39.20, p-value = 4.4e-16.

(B) Spatial organization of gene density. OR = 13.21, p-value = 2.2e-8.

(C) Spatial organization of gene expression. OR = 4.00, p-value = 0.0012.

(D) Spatial organization of chromatin accessibility. OR = 26.88, p-value = 5.9e-12.

(E) Spatial organization of genome-nuclear lamina interaction. OR = 40.00, p-value = 4.9e-13.

(F) Spatial organization of DNA replication time. OR = 32.00, p-value = 1.1e-10.

(G) Spatial organization of H3K36me3. OR = 10.91, p-value = 1.0e-7.

(H) Spatial organization of H3K27me3. OR = 2.17, p-value = 0.0706.

(I) Spatial organization of H3K4me3. OR = 24.43, p-value = 2.1e-11.

(J) Spatial organization of H3K9me3. OR = 15.71, p-value = 6.7e-8.

(K) Spatial organization of H4K20me3. OR = 45.10, p-value = 1.0e-13.

(L) Spatial organization of RNA polymerase II. OR = 5.47, p-value = 0.0001.

A. Compartment label	**B**. Gene density	**C.** Gene expression
D. Chromatin accessibility	**E.** Lamina interaction	**F.** DNA replication time
G. H3K36me3	**H.** H3K27me3	**I.** H3K4me3

J. H3K9me3	**K.** H4K20me3	**L.** RNA polymerase II