UNIVERSITY OF CALIFORNIA
RIVERSIDE

Hierarchical Bayesian Modeling of Diploid Chromatin Contacts and Structures

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics & Bioinformatics

by

Tiantian Ye

December 2020

Dissertation Committee:

Dr. Wenxiu Ma, Chairperson
Dr. Tao Jiang
Dr. Thomas Girke

The Dissertation of Tiantian Ye is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

Firstly I would like to express my most sincere gratitude to my advisor Dr. Wenxiu Ma, for her invaluable guidance, support and encouragement in both my research and career. She helped me come up with this dissertation topic and has always been helpful and guided me throughout my research and the writing of this dissertation. I would like to thank Dr. Tao Jiang and Dr. Thomas Girke, who have been in my dissertation committee, for their insightful feedback and helpful suggestions, as well as their guidance throughout my studies. This work is supported by Grant DBI-1751317 from the National Science Foundation.

I would also like to thank Dr. Howard Juldelson, Dr. Frances Sladek, Dr. Thomas Girke, Dr. Shizhong Xu and Dr. Weixin Yao for serving on my qualifying exam committee, and their helpful suggestions for my dissertation proposal. I am grateful to Dr. Yijun Ruan for providing the phased ChIA-PET data used in chapter 5, and the referees for the constructive and valuable suggestions. In addition, I would like to thank all the members, former and present, in Dr. Ma's lab: Yangyang Hu, Luke Klein, Huiling Liu, Jinli Zhang, Li Ma, Jimmy Ni, Douglas Kirsher, and Jingfei Zhang, for their helpful feedback, discussion and suggestions. Last but not least, I would like to thank my parents for their love and support throughout my PhD study and my life in general.

The text of this dissertation, in part, is a reprint of the material as it appears in "ASHIC: hierarchical Bayesian modeling of diploid chromatin contacts and structures" (published on *Nucleic Acids Research*, October 19, 2020). The co-author Wenxiu Ma listed in that publication directed and supervised the research which forms the basis for this dissertation.

To my parents for all the love and support.

ABSTRACT OF THE DISSERTATION

Hierarchical Bayesian Modeling of Diploid Chromatin Contacts and Structures

by

Tiantian Ye

Doctor of Philosophy, Graduate Program in Genetics, Genomics & Bioinformatics
University of California, Riverside, December 2020
Dr. Wenxiu Ma, Chairperson

The recently developed Hi-C technique has been widely applied to map genome-wide chromatin interactions. However, current methods for analyzing diploid Hi-C data cannot fully distinguish between homologous chromosomes. Consequently, the existing diploid Hi-C analyses are based on sparse and inaccurate allele-specific contact matrices, which might lead to incorrect modeling of diploid genome architecture.

Here we present ASHIC (Allele-Specific diploid Hi-C modeling), a hierarchical Bayesian framework to model allele-specific chromatin organizations in diploid genomes. We developed two models under the Bayesian framework: the Poisson-multinomial (ASHIC-PM) model and the zero-inflated Poisson-multinomial (ASHIC-ZIPM) model. The proposed ASHIC methods impute allele-specific contact maps from diploid Hi-C data and simultaneously infer allelic 3D structures.

Through simulation studies, we demonstrated that ASHIC methods outperformed existing approaches, especially under low coverage and low SNP density conditions. Additionally, in the analyses of diploid Hi-C datasets in mouse and human, our ASHIC-ZIPM method produced fine-resolution diploid chromatin maps and 3D structures and provided insights into the allelic chromatin organizations and functions. To summarize, our work provides a statistically rigorous framework for investigating fine-scale allele-specific chromatin conformations.

The ASHIC software is publicly available at https://github.com/wmalab/ASHIC.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The three-dimensional (3D) organization of chromatin in the nucleus plays an essential role in gene regulation [5]. The recently developed chromosome conformation capture coupled with high-throughput sequencing (Hi-C) technique [6, 7, 8] and its variants [9, 10, 11] have been widely applied to map genome-wide chromatin interactions and to elucidate the principles of spatial genome architecture. The Hi-C experiment yields a genome-wide chromatin contact matrix; each entry (i, j) in the matrix represents the contact frequency between two loci i and j in the genome. The mapping and subsequent analyses of genome-wide Hi-C contact matrices in various organisms have demonstrated that the gene expression is tightly regulated by chromatin interactions at multiple scales ranging from active/inactive chromosomal compartments and sub-compartments [6, 10], to topologically associating domains (TADs) [2], and fine-scale chromatin loops [10, 9].

One hindrance of current Hi-C data analysis is the lack of allele-specific modeling for diploid genomes. Most mammalian genomes are diploid, in which the genome contains two sets of each chromosome—a maternal and a paternal copy. Hence, a chromatin contact observed between two genomic loci in the reference (haploid) genome may correspond to four distinct yet indistinguishable chromatin interactions in the diploid genome. For example, a chromatin contact mapped to a loci pair (i, j) on the same chromosome in the reference genome could be either an intra-chromosomal contact (m_i, m_j) on the maternal allele, or an intra-chromosomal contact (p_i, p_j) on the paternal allele, or inter-homologous contacts (m_i, p_j) or (p_i, m_j). However, the majority of existing Hi-C analyses on diploid genomes do not distinguish between homologous chromosomes. As a result, current analyses are based on an aggregated contact matrix generated with mixed signals of maternal and paternal chromatin contacts, which could result in the false identification of significant chromatin interactions and an inaccurate understanding of the diploid genome architecture. Therefore, statistical methods for rigorous and accurate modeling of diploid Hi-C data are needed to facilitate elucidation of the mechanisms of chromatin organization and gene regulation.

Recently, several methods have been developed to obtain allele-specific chromatin contact matrices and/or allelic 3D structures from diploid Hi-C data [12, 13, 10, 14, 15, 16, 17]. These methods use heterozygous single nucleotide polymorphisms (SNPs) to identify the allele identity of chromatin interactions. Specifically, a Hi-C contact is a mate pair with two read ends representing the two interacting chromatin fragments. If a read end overlaps with SNPs for which the allele identity can be determined, we term it an allele-certain read. For example, a read containing maternal-specific SNP(s) is assigned to the maternal allele; similarly a read containing paternal-specific SNP(s) is assigned to the paternal allele. In addition, reads without SNPs are allele-ambiguous reads. Based on the allele identity of the paired ends, we can then categorize diploid Hi-C contacts into three groups: both-end allele-certain contacts, one-end allele-ambiguous contacts, and both-end allele-ambiguous contacts.

Without a statistically rigorous allele inference method, many previous studies applied either an “allele-certain” or a “mate-rescue” strategy to reconstruct the allele-specific contact maps in diploid genomes. In the allele-certain approach, only both-end allele-certain contacts are used [10, 14]. However, the both-end allele-certain contacts only account for a small portion of the total chromatin contacts (Table 1.1). For example, in the Patski (BL6×*Spretus*) cell line of which the SNP density is approximately 1 per 75 bp, the proportion of both-end allele-certain contacts in a typical Hi-C dataset is about 35.6%. Whereas, in the human GM12878 cell line of which the SNP density is approximately 1 per 1700 bp, the both-end allele-certain proportion drops to 0.14%. Consequently, the diploid contact matrices obtained by such an allele-certain approach is often sparse and of low resolution.

	SNPs	SNP density	Total contacts	Both-end allele-certain	One-end allele-ambiguous	Both-end allele-ambiguous
Bonora et al. (BL6×Spretus)	35,441,735	1/75	365,294,454	35.62%	43.25%	21.14%
Rao et al. (GM12878)	1,787,252	1/1700	8,178,930,507	0.14%	5.73%	94.14%

Table 1.1: Proportion of allele-specific reads in published diploid Hi-C datasets.

To overcome the low-coverage issue of the allele-certain approach, several diploid Hi-C studies adopted a straightforward mate-rescue strategy to infer the allele identity of one-end allele-ambiguous contacts, i.e., the allele-ambiguous end of such contact is assigned to the same allele as its mate-end [13, 15, 18]. This mate-rescue method attempts to recover one-end allele-ambiguous contacts, which varies approximately from 5.7% (in the case of GM12878 cells) to 43.3% (in the case of Patski cells) of the total contacts (Table 1.1). However, one-end allele-ambiguous contacts are all assumed to be intra-chromosomal contacts in the results of the mate-rescue approach. Such false assumption would lead to inaccurate contact maps, especially in the regions where inter-chromosomal interactions are observed across chromosomal territories.

Since the mate-rescue method fails to infer inter-chromosomal interactions from one-end allele-ambiguous contacts, Tan et al. [16] proposed an iterative two-stage imputation algorithm Dip-C for modeling single-cell diploid Hi-C data. In the first imputation stage, one-end allele-ambiguous contacts are phased using an *ad hoc* voting procedure by their neighborhood on the contact matrix. In the second imputation stage, the assignment of allele-ambiguous contacts is refined by the 3D structures. The Dip-C method can be viewed as an advanced mate-rescue method, as it leverages additional information from both contact matrices and 3D structures to infer allele-ambiguous contacts. However, the Dip-C method is specifically designed for single-cell Hi-C data therefore may not adapt well to bulk Hi-C data. Moreover, Dip-C uses a deterministic voting strategy to assign allele-ambiguous contacts, which does not provide a probabilistic model of all possible allele origins.

One common drawback of the allele-certain and mate-rescue methods is that they do not utilize both-end allele-ambiguous contacts, which represent a substantial proportion of the total diploid contacts, ranging from 21.1% (Patski) to 94.1% (GM12878) (Table 1.1). Inferring the allele identity of both-end allele-ambiguous contacts remains a significant challenge. To date, few methods have been developed to address this problem. The Dip-C method [16] attempts to impute only inter-chromosomal rather than intra-chromosomal both-end allele-ambiguous contacts. Thus, it does not produce a fully imputed diploid contact map. In addition, our previously proposed Poisson-Gamma model [12] imputes both one-end and both-end allele-ambiguous contacts, and estimates the diploid contact matrices by an iterative expectation-maximization (EM) algorithm. However, the Poisson-Gamma method does not predict 3D structures nor use the structures to assist the assignment of allele-ambiguous contacts. As a result, it might not work robustly in fine-resolution analyses. Furthermore, Cauer et al. [17] developed diploid-PASTIS, an extension of the PASTIS model [19], to infer the diploid chromatin structures. Diploid Hi-C contacts are modeled as Poisson variables, and the optimal diploid structures are solved by maximizing the likelihood function with additional structural constraints. The diploid-PASTIS method is specifically designed to model diploid 3D structures, but does not infer allele-ambiguous contacts to impute diploid contact matrices.

To tackle the aforementioned challenges, we developed a hierarchical Bayesian framework for Allele-Specific diploid Hi-C modeling, named ASHIC. Briefly, allele-specific contact counts are modeled as Poisson-multinomial random variables (referred as the ASHIC-PM model) and diploid contact matrices and 3D structures are estimated via an EM algorithm. In addition, to overcome the sparsity issue of diploid Hi-C contact maps, we proposed a zero-inflated version of the ASHIC-PM method, namely the zero-inflated Poisson-multinomial model (in short, ASHIC-ZIPM). Both ASHIC models can completely dissect all diploid Hi-C contacts into allele-specific contact maps, while simultaneously reconstruct 3D homologous chromosomal structures. To the best of our knowledge, our ASHIC methods are the first methods that fully impute all allele-ambiguous contacts and infer both the diploid contact matrices and allelic 3D structures.

We thoroughly evaluated our methods through a series of simulation studies and demonstrated that our ASHIC methods outperformed the allele-certain and mate-rescue approaches in various settings of sequencing coverage, SNP density and homologous structural similarity. We also applied the ASHIC-ZIPM method to two published diploid Hi-C datasets [18, 10]. First, using the mouse Patski data [18], we successfully confirmed that the predicted diploid contact maps and 3D structures of the homologous X chromosomes exhibited distinct conformations, where the inactive X demonstrated the bipartite super-domains [12]. Furthermore, we studied fine-scale chromatin organizations of the imprinted *H19/Igf2* region at 10 kb resolution and revealed distinct parental-specific chromatin interactions anchored at *H19* and *Igf2*. With the fully imputed diploid contact matrices, we uncovered a maternal-specific sub-TAD at the *H19/Igf2* region. Second, using the human GM12878 data [10], we further confirmed the maternal-specific sub-TAD structure and parental-specific chromatin interactions at the human *H19/IGF2* imprinting locus. Our ASHIC-imputed allele-specific contacts maps were consistent with the previously published chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) results [4].

Chapter 2

Hierarchical Bayesian Models

To model diploid Hi-C data, we propose a hierarchical Bayesian modeling framework for imputing the allele-specific chromatin contacts and reconstructing the allelic 3D structures. Specifically, we model the generation of allele-specific contacts with either a Poisson-multinomial process (the ASHIC-PM model) or a zero-inflated Poisson-multinomial process (the ASHIC-ZIPM model) for the inference of diploid contact matrices and 3D structures. The ASHIC-ZIPM model is a zero-inflated version of ASHIC-PM and it explicitly accounts for the excessive zeros observed in Hi-C contact matrices.

2.1 Notations of Allele-Specific Chromatin Contacts

Let m and p denote a homologous chromosomal pair with same length n in a diploid genome. To construct the diploid Hi-C contact frequency matrix, we partition the chromosomes into fixed-size non-overlapping bins and count chromatin contacts observed between each bin pair. In the diploid setting, chromatin contacts between the bins i and

j can result from four distinct events: (m_i, m_j), (p_i, p_j), (m_i, p_j), or (p_i, m_j), where (m_i, m_j) and (p_i, p_j) are intra-chromosomal contacts on chromosome m and p, respectively, and (m_i, p_j) and (p_i, m_j) are inter-chromosomal contacts between the homologous chromosomes (Figure 2.1A). Therefore, the aggregated contact frequency T_{ij} between the bins i and j can be calculated as follows: $T_{ij} = \sum_\eta \sum_\theta T_{\eta_i \theta_j}$, where $T_{\eta_i \theta_j}$ is the unknown true allele-specific contact frequency between η_i (bin i on chromosome η) and θ_j (bin j on chromosome θ) that we aim to estimate, $\eta, \theta \in \{m, p\}$, $1 \leq i, j \leq n$ (Figure 2.1B).

Using heterozygous SNPs, we can classify single-end reads into three categories: reads containing allele-m-specific SNPs, reads containing allele-p-specific SNPs, and reads containing no SNPs. We refer to the first two categories as allele-certain reads while the last category as allele-ambiguous reads. Furthermore, since Hi-C contacts are paired-end reads, each end of the mated pair can either be allele-certain or allele-ambiguous.

Let $C_{\eta_i \theta_j}$ indicate the frequency of both-end allele-certain contacts between the bins η_i and θ_j. In addition, we specify $C_{\eta_i \theta_j^*}$ to be the contact frequency between η_i and θ_j where the allele identity of η_i is known but the allele identity of θ_j is unknown. In other words, one end of the Hi-C contact is from θ_j; however, the read does not overlap with any SNPs. Therefore the allele identity of θ_j remains unknown. Similarly, we use $C_{\eta_i^* \theta_j}$ when the allele identity of η_i is unknown and $C_{\eta_i^* \theta_j^*}$ when the allele identity of both ends are unknown.

Hence, the true allele-specific contact frequency $T_{\eta_i \theta_j}$ equals to the sum of the following four components:

$$T_{\eta_i \theta_j} = C_{\eta_i \theta_j} + C_{\eta_i \theta_j^*} + C_{\eta_i^* \theta_j} + C_{\eta_i^* \theta_j^*} \tag{2.1}$$

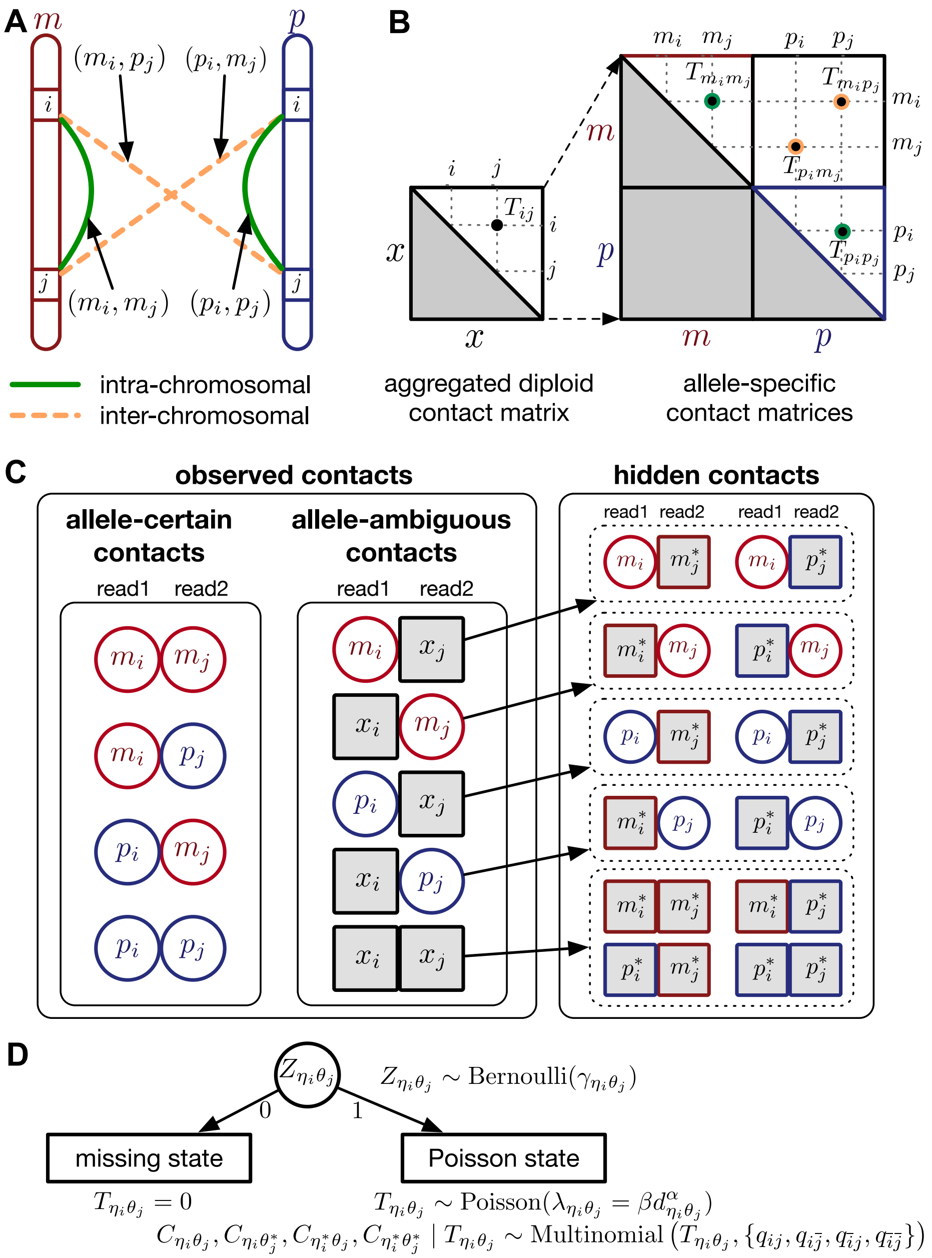
A
m
p
i
j
(m_i, p_j)
(p_i, m_j)
(m_i, m_j)
(p_i, p_j)
intra-chromosomal
inter-chromosomal
B
x
T_{ij}
aggregated diploid contact matrix
m_i m_j p_i p_j
$T_{m_i m_j}$
$T_{m_i p_j}$
$T_{p_i m_j}$
$T_{p_i p_j}$
allele-specific contact matrices
C
observed contacts
allele-certain contacts
allele-ambiguous contacts
hidden contacts
read1 read2
m_i m_j p_i p_j x_i x_j
m_i^* m_j^* p_i^* p_j^*
D
$Z_{\eta_i\theta_j}$
$Z_{\eta_i\theta_j} \sim \text{Bernoulli}(\gamma_{\eta_i\theta_j})$
0
1
missing state
Poisson state
$T_{\eta_i\theta_j} = 0$
$T_{\eta_i\theta_j} \sim \text{Poisson}(\lambda_{\eta_i\theta_j} = \beta d^{\alpha}_{\eta_i\theta_j})$
$C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}, C_{\eta_i^*\theta_j^*} \mid T_{\eta_i\theta_j} \sim \text{Multinomial}\left(T_{\eta_i\theta_j}, \{q_{ij}, q_{i\bar{j}}, q_{\bar{i}j}, q_{\bar{i}\bar{j}}\}\right)$

Figure 2.1: Overview of allele-specific modeling of diploid Hi-C data. (A) Diploid contact (i, j) is a combination of four distinct allele-specific contacts (m_i, m_j), (m_i, p_j), (p_i, m_j), and (p_i, p_j). (B) Reconstruction of allele-specific diploid contact matrix. (C) Observed allele-specific contacts between bins i and j can be decomposed into observed allele-certain contacts $\boldsymbol{C^O_{ij}}$, observed allele-ambiguous contacts $\boldsymbol{C^X_{ij}}$. We aim to decompose $\boldsymbol{C^X_{ij}}$ and infer the hidden contacts $\boldsymbol{C^H_{ij}}$, and impute the true allele-specific contacts $\boldsymbol{T_{ij}}$. (D) Illustration of the hierarchical Bayesian ASHIC-ZIPM model.

In diploid Hi-C data, we cannot directly observe $C_{\eta_i m_j^*}$ and $C_{\eta_i p_j^*}$ since the read end mapped to the bin j is allele-ambiguous and hence, it cannot be distinguished between m_j and p_j. As a result, the observed Hi-C contacts contain the following types of allele-ambiguous contacts:

$$\begin{aligned} C_{\eta_i x_j} &= C_{\eta_i m_j^*} + C_{\eta_i p_j^*} \\ C_{x_i \theta_j} &= C_{m_i^* \theta_j} + C_{p_i^* \theta_j} \\ C_{x_i x_j} &= C_{m_i^* m_j^*} + C_{m_i^* p_j^*} + C_{p_i^* m_j^*} + C_{p_i^* p_j^*} \end{aligned} \tag{2.2}$$

where x indicates that the allele identity of a read end is unknown. We refer to $C_{\eta_i x_j}$ and $C_{x_i \theta_j}$ as one-end allele-ambiguous contacts and $C_{x_i x_j}$ as both-end allele-ambiguous contacts (Figure 2.1C).

In summary, we define $\boldsymbol{C^O} = \{C_{\eta_i \theta_j}\}$ as the observed allele-specific contact frequencies, $\boldsymbol{C^X} = \{C_{\eta_i x_j}, C_{x_i \theta_j}, C_{x_i x_j}\}$ as the observed allele-ambiguous contact frequencies (LHS in eq. (2.2)), and $\boldsymbol{C^H} = \{C_{\eta_i \theta_j^*}, C_{\eta_i^* \theta_j}, C_{\eta_i^* \theta_j^*}\}$ as the unobserved (hidden) allele-specific contact frequencies (RHS in eq. (2.2)). Our goal is to decompose $\boldsymbol{C^X}$ and infer $\boldsymbol{C^H}$ in order to impute the true allele-specific frequencies $\boldsymbol{T} = \{T_{\eta_i \theta_j}\}$ by eq. (2.1) (Figure 2.1C).

2.2 Modeling True Allele-Specific Contact Frequencies

We adopt the coarse-grained polymer model [20] to represent the chromosomal structures. Each bin in the genome is represented as a bead in the 3D space, and each chromosome can be viewed as a chain of beads. Specifically, we denote $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$ to be the 3D coordinates of the homologous chromosomes m and p, respectively, where $\boldsymbol{X_m}, \boldsymbol{X_p} \in R^{3\times n}$. Let $\boldsymbol{x_{\eta_i}}$ and $\boldsymbol{x_{\theta_j}}$ to be the 3D coordinates of beads η_i and θ_j, respectively, where

$\eta, \theta \in \{m, p\}$. According to polymer physics [21, 19], the contact frequency $T_{\eta_i\theta_j}$ between η_i and θ_j is inversely correlated with their spatial distance $d_{\eta_i\theta_j}$, following a power-law decay function. That is, $T_{\eta_i\theta_j} \propto d^{\alpha}_{\eta_i\theta_j}$, where $\alpha < 0$ is the exponent of the distance-decay function, and $d_{\eta_i\theta_j}$ is the Euclidean distance between beads η_i and θ_j:

$$d_{\eta_i\theta_j} = \|\boldsymbol{x_{\eta_i}} - \boldsymbol{x_{\theta_j}}\|_2 = \sqrt{(x_{\eta_{i1}} - x_{\theta_{j1}})^2 + (x_{\eta_{i2}} - x_{\theta_{j2}})^2 + (x_{\eta_{i3}} - x_{\theta_{j3}})^2} \tag{2.3}$$

2.2.1 Poisson Model

Similar to the PASTIS method [19], we model the true allele-specific contact frequency $T_{\eta_i\theta_j}$ as a Poisson random variable:

$$T_{\eta_i\theta_j} \sim \text{Poisson}(\lambda_{\eta_i\theta_j} = \beta d^{\alpha}_{\eta_i\theta_j}), \tag{2.4}$$

where β is a scaling factor corresponding to the sequencing depth, $d_{\eta_i\theta_j}$ is the Euclidean distance between beads η_i and θ_j, and $\alpha < 0$ models the power-law decay rate.

2.2.2 Zero-Inflated Poisson Model

Definition 1 *Suppose $X \sim Poisson(\lambda)$ and Z is a binary variable. X and Z are independent. Let $Y = ZX$. Then, $Y|Z$ follows a zero-inflated Poisson (ZIP) distribution:*

$$Y|Z \sim \mathit{ZIP}(\lambda, Z).$$

The probability mass function (pmf) of the conditional distribution Y given Z is

$$f_{ZIP}(Y \mid Z) = \left[\mathbb{1}(Y = 0)\right]^{1-Z} \cdot \left[f_P(Y; \lambda)\right]^{Z},$$

where $f_P(\cdot)$ denotes the Poisson pmf, and $\mathbb{1}(A)$ is an indicator function defined as:

$$\mathbb{1}(A) := \begin{cases} 1, & \textit{if } A \textit{ is true}; \\ 0, & \textit{otherwise}. \end{cases}$$

Proposition 2 *Assume $Y|Z \sim ZIP(\lambda, Z)$. Then the conditional expectation of $Y|Z$ is*

$$\mathbb{E}(Y|Z) = Z\lambda.$$

Furthermore, to account for the excessive zeros in Hi-C contact matrices, we propose to use a zero-inflated Poisson (ZIP) distribution to model the contact counts (Figure 2.1D). We assume that $T_{\eta_i\theta_j}$ follows a ZIP distribution:

$$T_{\eta_i\theta_j} \mid Z_{\eta_i\theta_j} \sim \text{ZIP}(\lambda_{\eta_i\theta_j} = \beta d^{\alpha}_{\eta_i\theta_j}, Z_{\eta_i\theta_j}) \tag{2.5}$$

Different from the Poisson model, here we introduce $Z_{\eta_i\theta_j}$, a latent binary variable to indicate whether $T_{\eta_i\theta_j}$ is generated from the Poisson state ($Z_{\eta_i\theta_j} = 1$, $T_{\eta_i\theta_j} \sim$ Poisson$(\lambda_{\eta_i\theta_j})$) or the missing state ($Z_{\eta_i\theta_j} = 0$, $T_{\eta_i\theta_j} = 0$). Furthermore, we assume that $Z_{\eta_i\theta_j}$ follows a Bernoulli prior with a success probability $\gamma_{\eta_i\theta_j}$:

$$Z_{\eta_i\theta_j} \sim \text{Bernoulli}(\gamma_{\eta_i\theta_j}) \tag{2.6}$$

For intra-chromosomal contacts (where $\eta = \theta$), $\gamma_{\eta_i\theta_j}$ is a function of the corresponding genomic distance. For inter-chromosomal contacts ($\eta \neq \theta$), $\gamma_{\eta_i\theta_j}$ is set to a constant:

$$\gamma_{\eta_i\theta_j} = \begin{cases} \gamma_{|i-j|} = G(|i-j|), & \eta = \theta; \\ \gamma_{\text{inter}}, & \eta \neq \theta. \end{cases} \tag{2.7}$$

In other words, the true allele-specific contact frequency $T_{\eta_i\theta_j}$ is a mixture of two states. In the Poisson state (with probability $\gamma_{\eta_i\theta_j}$), $T_{\eta_i\theta_j}$ follows a Poisson distribution;

whereas in the missing state (with probability $1 - \gamma_{\eta_i\theta_j}$), $T_{\eta_i\theta_j} = 0$. The $\gamma_{\eta_i\theta_j}$ parameter acts as a weight between the Poisson and missing states. Then, the pmf of the conditional distribution $T_{\eta_i\theta_j}$ given $Z_{\eta_i\theta_j}$ can be written as

$$f(T_{\eta_i\theta_j} \mid Z_{\eta_i\theta_j}) = \left[\mathbb{1}\left(T_{\eta_i\theta_j} = 0\right)\right]^{1-Z_{\eta_i\theta_j}} \cdot \left[f_{\mathrm{P}}(T_{\eta_i\theta_j}; \lambda_{\eta_i\theta_j})\right]^{Z_{\eta_i\theta_j}}, \tag{2.8}$$

and the joint pmf of $T_{\eta_i\theta_j}$ and $Z_{\eta_i\theta_j}$ is

$$f(T_{\eta_i\theta_j}, Z_{\eta_i\theta_j}) = \left[(1 - \gamma_{\eta_i\theta_j})\mathbb{1}\left(T_{\eta_i\theta_j} = 0\right)\right]^{1-Z_{\eta_i\theta_j}} \left[\gamma_{\eta_i\theta_j} f_{\mathrm{P}}(T_{\eta_i\theta_j}; \lambda_{\eta_i\theta_j})\right]^{Z_{\eta_i\theta_j}}. \tag{2.9}$$

2.3 Modeling Allele-Identifiable Probability

As discussed in Section 2.1, we cannot directly observe the allele identity of all diploid Hi-C contacts. A higher SNP density results in a higher probability that our allele-aware mapping pipeline will be able to identify the true allele source of a read. We use q_i to denote the allele-identifiable probability of bin i in the genome, i.e., if a single-end read is mapped to bin i, the probability that the read overlaps with SNP(s) (and therefore can be distinguished between alleles m and p) is q_i. Consequently, assuming that bins i and j are independent, the probabilities that a paired-end contact between the bins i and j is both-ends allele-certain (q_{ij}), one-end allele-ambiguous at bin i ($q_{\bar{i}j}$), one-end allele-ambiguous at bin j ($q_{i\bar{j}}$), and both-end allele-ambiguous ($q_{\bar{i}\bar{j}}$) can be calculated as follows:

$$\begin{aligned} q_{ij} &= q_i q_j \\ q_{i\bar{j}} &= q_i(1 - q_j) \\ q_{\bar{i}j} &= (1 - q_i)q_j \\ q_{\bar{i}\bar{j}} &= (1 - q_i)(1 - q_j) \end{aligned} \tag{2.10}$$

2.4 Modeling Hidden Allele-Specific Contact Frequencies

Recall in eq. (2.1), the true allele-specific contact frequency $T_{\eta_i\theta_j}$ can be expressed as the sum of one observed allele-certain contact frequency $C_{\eta_i\theta_j}$ and three hidden allele-specific contact frequencies $C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}, C_{\eta_i^*\theta_j^*}$. We assume the generation of the decomposed allele-specific contact frequencies from the true allele-specific contact frequency follows a multinomial distribution in the ASHIC-PM model, and a zero-inflated multinomial (ZIM) distribution in the ASHIC-ZIPM model.

2.4.1 ASHIC-PM Model

In the Poisson-multinomial model, we assume that the decomposed allele-specific chromatin contact frequencies $C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}$, and $C_{\eta_i^*\theta_j^*}$ conditional on the true allele-specific contact frequency $T_{\eta_i\theta_j}$ follow a multinomial distribution with the probabilities $q_{ij}, q_{i\bar{j}}, q_{\bar{i}j}$, and $q_{\bar{i}\bar{j}}$. That is,

$$C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}, C_{\eta_i^*\theta_j^*} \mid T_{\eta_i\theta_j} \sim \text{Multinomial}\left(T_{\eta_i\theta_j}, \{q_{ij}, q_{i\bar{j}}, q_{\bar{i}j}, q_{\bar{i}\bar{j}}\}\right). \tag{2.11}$$

Proposition 3 *Given a multinomial experiment with X trials, where each possible outcome can occur with probabilities $p_1, p_2, \cdots, p_k$, suppose X follows a Poisson distribution with the parameter λ. Let X_i denote the number of occurrences of the i-th outcomes in X trials. Then, $X_1, X_2, \cdots, X_k$ follow mutually independent Poisson distributions with the parameters $p_1\lambda, p_2\lambda, \cdots, p_k\lambda$, respectively.*

Therefore, we can derive $C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}$, and $C_{\eta_i^*\theta_j^*}$ as mutually independent Poisson random variables:

$$\begin{aligned} C_{\eta_i\theta_j} &\sim \text{Poisson}(q_{ij}\lambda_{\eta_i\theta_j}), \\ C_{\eta_i\theta_j^*} &\sim \text{Poisson}(q_{i\bar{j}}\lambda_{\eta_i\theta_j}), \\ C_{\eta_i^*\theta_j} &\sim \text{Poisson}(q_{\bar{i}j}\lambda_{\eta_i\theta_j}), \\ C_{\eta_i^*\theta_j^*} &\sim \text{Poisson}(q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}). \end{aligned} \tag{2.12}$$

Proposition 4 *Let X_1, X_2 be independent Poisson random variables with*

$$X_1 \sim Poisson(\lambda_1),$$

$$X_2 \sim Poisson(\lambda_2).$$

Then, $X_1 + X_2 \sim Poisson(\lambda_1 + \lambda_2)$.

By eq. (2.2), eq. (2.12), and Proposition 4, we have

$$\begin{aligned} C_{\eta_i x_j} &\sim \text{Poisson}\left(q_{i\bar{j}}(\lambda_{\eta_i m_j} + \lambda_{\eta_i p_j})\right), \\ C_{x_i\theta_j} &\sim \text{Poisson}\left(q_{\bar{i}j}(\lambda_{m_i\theta_j} + \lambda_{p_i\theta_j})\right), \\ C_{x_i x_j} &\sim \text{Poisson}\left(q_{\bar{i}\bar{j}}(\lambda_{m_i m_j} + \lambda_{m_i p_j} + \lambda_{p_i m_j} + \lambda_{p_i p_j})\right), \end{aligned} \tag{2.13}$$

which implies that $C_{\eta_i x_j}, C_{x_i\theta_j}$, and $C_{x_i x_j}$ are mutually independent.

Proposition 5 *Let X_1 and X_2 be independent Poisson random variables with*

$$X_1 \sim Poisson(\lambda_1),$$

$$X_2 \sim Poisson(\lambda_2).$$

Then, $X_1 \mid X_1 + X_2 \sim Binomial\Big(X_1 + X_2, \frac{\lambda_1}{\lambda_1+\lambda_2}\Big)$.

Corollary 6 *Let $X_1, X_2, X_3,$ and X_4 be independent Poisson distribution with*

$$X_i \sim \textit{Poisson}(\lambda_i). \quad i = 1, 2, 3, 4$$

Then, $X_1 \mid (X_1 + X_2 + X_3 + X_4) \sim \textit{Binomial}\Big(X_1 + X_2 + X_3 + X_4, \frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}\Big).$

Therefore, we can derive the conditional distribution of the hidden allele-specific contact frequencies $\boldsymbol{C^H} = \{C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}, C_{\eta_i^*\theta_j^*}\}$, given the observed allele-ambiguous contact frequencies $\boldsymbol{C^X} = \{C_{\eta_i x_j}, C_{x_i\theta_j}, C_{x_i x_j}\}$, as follows:

$$\begin{aligned}
C_{\eta_i\theta_j^*} \mid C_{\eta_i x_j} &\sim \text{Binomial}\left(C_{\eta_i x_j}, \frac{\lambda_{\eta_i\theta_j}}{\lambda_{\eta_i m_j} + \lambda_{\eta_i p_j}}\right), \\
C_{\eta_i^*\theta_j} \mid C_{x_i\theta_j} &\sim \text{Binomial}\left(C_{x_i\theta_j}, \frac{\lambda_{\eta_i\theta_j}}{\lambda_{m_i\theta_j} + \lambda_{p_i\theta_j}}\right), \\
C_{\eta_i^*\theta_j^*} \mid C_{x_i x_j} &\sim \text{Binomial}\left(C_{x_i x_j}, \frac{\lambda_{\eta_i\theta_j}}{\lambda_{m_i m_j} + \lambda_{m_i p_j} + \lambda_{p_i m_j} + \lambda_{p_i p_j}}\right).
\end{aligned} \tag{2.14}$$

2.4.2 ASHIC-ZIPM Model

Definition 7 *Suppose $\boldsymbol{X} = \{X_1, \cdots, X_k, \cdots, X_K\}$ follows a multinomial distribution with the parameters n and $\boldsymbol{q} = \{p_1, \cdots, p_k, \cdots, p_K\}$. That is, $X \sim \textit{Multinomial}(n, \boldsymbol{q})$. Further, suppose that Z is a binary variable and that $\boldsymbol{X}$ and Z are independent. Let $\boldsymbol{Y} = Z\boldsymbol{X}$, that is, $Y_k = ZX_k$, for all k. Then, $\boldsymbol{Y}|Z$ follows a zero-inflated multinomial (ZIM) distribution:*

$$\boldsymbol{Y}|Z \sim \textit{ZIM}(n, \boldsymbol{q}, Z).$$

The pmf of the conditional distribution $\boldsymbol{Y}$ given Z is

$$f_{ZIM}(\boldsymbol{Y} \mid Z) = \left[\mathbb{1}(\boldsymbol{Y} = \boldsymbol{0})\right]^{1-Z} \cdot \left[f_M(\boldsymbol{Y}; n, \boldsymbol{q})\right]^Z, \tag{2.15}$$

where $f_M(\cdot)$ denotes the multinomial pmf.

In the zero-inflated Poisson-multinomial model, we assume that conditional on the true allele-specific contact frequency $T_{\eta_i\theta_j}$ and the Poisson state latent variable $Z_{\eta_i\theta_j}$, the decomposed allele-specific chromatin contact frequencies $C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}$, and $C_{\eta_i^*\theta_j^*}$ follow a zero-inflated multinomial (ZIM) model:

$$C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}, C_{\eta_i^*\theta_j^*} \mid T_{\eta_i\theta_j}, Z_{\eta_i\theta_j} \sim \mathrm{ZIM}\left(T_{\eta_i\theta_j}, \{q_{ij}, q_{i\bar{j}}, q_{\bar{i}j}, q_{\bar{i}\bar{j}}\}, Z_{\eta_i\theta_j}\right). \tag{2.16}$$

That is, $C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}, C_{\eta_i^*\theta_j^*} \mid T_{\eta_i\theta_j}, Z_{\eta_i\theta_j}$ has a mixture distribution. If $Z_{\eta_i\theta_j} = 1$, it reduces to the multinomial distribution as in eq. (2.11); otherwise, $C_{\eta_i\theta_j} = C_{\eta_i\theta_j^*} = C_{\eta_i^*\theta_j} = C_{\eta_i^*\theta_j^*} = 0$. By Proposition 3, we can derive $C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}$, and $C_{\eta_i^*\theta_j^*}$, conditional on $Z_{\eta_i\theta_j}$, as mutually independent ZIP random variables:

$$\begin{aligned}
C_{\eta_i\theta_j} \mid Z_{\eta_i\theta_j} &\sim \mathrm{ZIP}(q_{ij}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j}),\\
C_{\eta_i\theta_j^*} \mid Z_{\eta_i\theta_j} &\sim \mathrm{ZIP}(q_{i\bar{j}}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j}),\\
C_{\eta_i^*\theta_j} \mid Z_{\eta_i\theta_j} &\sim \mathrm{ZIP}(q_{\bar{i}j}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j}),\\
C_{\eta_i^*\theta_j^*} \mid Z_{\eta_i\theta_j} &\sim \mathrm{ZIP}(q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j}).
\end{aligned} \tag{2.17}$$

Therefore, we can factorize the joint distribution as follows:

$$\begin{aligned}
&f(C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}, C_{\eta_i^*\theta_j^*} \mid Z_{\eta_i\theta_j})\\
=&f_{\mathrm{ZIP}}(C_{\eta_i\theta_j}; q_{ij}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j}) f_{\mathrm{ZIP}}(C_{\eta_i\theta_j^*}; q_{i\bar{j}}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j}) f_{\mathrm{ZIP}}(C_{\eta_i^*\theta_j}; q_{\bar{i}j}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j})\cdot\\
&f_{\mathrm{ZIP}}(C_{\eta_i^*\theta_j^*}; q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j})\\
=&\left[f_{\mathrm{P}}(C_{\eta_i\theta_j}; q_{ij}\lambda_{\eta_i\theta_j}) f_{\mathrm{P}}(C_{\eta_i\theta_j^*}; q_{i\bar{j}}\lambda_{\eta_i\theta_j}) f_{\mathrm{P}}(C_{\eta_i^*\theta_j}; q_{\bar{i}j}\lambda_{\eta_i\theta_j}) f_{\mathrm{P}}(C_{\eta_i^*\theta_j^*}; q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j})\right]^{Z_{\eta_i\theta_j}}\cdot\\
&\left[\mathbb{1}\Big(C_{\eta_i\theta_j} = C_{\eta_i\theta_j^*} = C_{\eta_i^*\theta_j} = C_{\eta_i^*\theta_j^*} = 0\Big)\right]^{1-Z_{\eta_i\theta_j}}.
\end{aligned} \tag{2.18}$$

Note that $C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}$, and $C_{\eta_i^*\theta_j^*}$ are mutually conditionally independent on $Z_{\eta_i\theta_j}$ but not mutually independent. When $Z_{\eta_i\theta_j} = 1$, eq. (2.17) reduces to the special case in eq. (2.12), where $C_{\eta_i \theta_j}, C_{\eta_i \theta_j^*}, C_{\eta_i^* \theta_j}$, and $C_{\eta_i^* \theta_j^*}$ are mutually independent Poisson random variables.

Recall that in eq. (2.2), the observed allele-ambiguous contact frequency is expressed as the sum of the corresponding hidden allele-specific contact frequencies. Therefore, by Proposition 4, we can derive that the observed allele-ambiguous contact frequencies $C_{\eta_i x_j}$, $C_{x_i \theta_j}$, and $C_{x_i x_j}$ are ZIP random variables and mutually conditional independent given $\boldsymbol{Z}$:

$$C_{\eta_i x_j} \mid Z_{\eta_i m_j}, Z_{\eta_i p_j} \sim \text{ZIP}\left(q_{i\bar{j}} \sum_{\theta'} Z_{\eta_i \theta'_j} \lambda_{\eta_i \theta'_j},\ Z_{\eta_i x_j}\right), \tag{2.19}$$

where a new latent binary variable $Z_{\eta_i x_j}$ is introduced and defined as $Z_{\eta_i m_j}$ or $Z_{\eta_i p_j}$. That is, when $Z_{\eta_i m_j} = 1$ or $Z_{\eta_i p_j} = 1$, $Z_{\eta_i x_j} = 1$ and $C_{\eta_i x_j} \sim \text{Poisson}\left(q_{i\bar{j}}(Z_{\eta_i m_j} \lambda_{\eta_i m_j} + Z_{\eta_i p_j} \lambda_{\eta_i p_j})\right)$; otherwise, $Z_{\eta_i x_j} = 0$ and $C_{\eta_i x_j} = 0$. Similarly, we have

$$C_{x_i \theta_j} \mid Z_{m_i \theta_j}, Z_{p_i \theta_j} \sim \text{ZIP}\left(q_{\bar{i}j} \sum_{\eta'} Z_{\eta'_i \theta_j} \lambda_{\eta'_i \theta_j},\ Z_{x_i \theta_j}\right), \tag{2.20}$$

where $Z_{x_i \theta_j} := Z_{m_i \theta_j}$or $Z_{p_i \theta_j}$, and

$$C_{x_i x_j} \mid Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j} \sim \text{ZIP}\left(q_{\bar{i}\bar{j}} \sum_{\eta', \theta'} Z_{\eta'_i \theta'_j} \lambda_{\eta'_i \theta'_j},\ Z_{x_i x_j}\right), \tag{2.21}$$

where $Z_{x_i x_j} := Z_{m_i m_j}$or $Z_{m_i p_j}$or $Z_{p_i m_j}$or $Z_{p_i p_j}$.

Furthermore, by Corollary 6 and eq. (2.17), we can show that conditional on $\boldsymbol{Z} = \{Z_{\eta_i \theta_j}\}$ and the observed allele-ambiguous contact frequencies $C_{\eta_i x_j}, C_{x_i, \theta_j}$, and C_{x_i, x_j}, the hidden allele-specific contact frequencies $C_{\eta_i \theta_j^*}, C_{\eta_i^* \theta_j}$, and $C_{\eta_i^* \theta_j^*}$ are mutually independent zero-inflated binomial (ZIB) random variables:

$$C_{\eta_i \theta_j^*} \mid C_{\eta_i x_j}, Z_{\eta_i m_j}, Z_{\eta_i p_j} \sim \text{ZIB}\left(C_{\eta_i x_j}, \frac{\lambda_{\eta_i \theta_j}}{\sum_{\theta'} Z_{\eta_i \theta'_j} \lambda_{\eta_i \theta'_j}},\ Z_{\eta_i \theta_j}\right). \tag{2.22}$$

That is, $C_{\eta_i\theta_j^*} \mid C_{\eta_i x_j}, Z_{\eta_i m_j}, Z_{\eta_i p_j} \sim \text{Binomial}\left(C_{\eta_i x_j}, \frac{\lambda_{\eta_i\theta_j}}{Z_{\eta_i m_j}\lambda_{\eta_i m_j}+Z_{\eta_i p_j}\lambda_{\eta_i p_j}}\right)$ if $Z_{\eta_i\theta_j} = 1$; otherwise, $C_{\eta_i\theta_j^*} = 0$. Similarly, we have

$$C_{\eta_i^*\theta_j} \mid C_{x_i\theta_j}, Z_{m_i\theta_j}, Z_{p_i\theta_j} \sim \text{ZIB}\left(C_{x_i\theta_j}, \frac{\lambda_{\eta_i\theta_j}}{\sum_{\eta'} Z_{\eta_i'\theta_j}\lambda_{\eta_i'\theta_j}},\ Z_{\eta_i\theta_j}\right), \tag{2.23}$$

$$C_{\eta_i^*\theta_j^*} \mid C_{x_i x_j}, Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j} \sim \text{ZIB}\left(C_{x_i x_j}, \frac{\lambda_{\eta_i\theta_j}}{\sum_{\eta',\theta'} Z_{\eta_i'\theta_j'}\lambda_{\eta_i'\theta_j'}},\ Z_{\eta_i\theta_j}\right). \tag{2.24}$$

Definition 8 *Suppose $X \sim Binomial(n, p)$ and Z is a binary variable where X and Z are independent. Let $Y = ZX$. Then, $Y|Z$ follows a zero-inflated binomial (ZIB) distribution:*

$$Y|Z \sim ZIB\,(n, p, Z)\,.$$

The pmf of the conditional distribution Y given Z is

$$f_{ZIB}(Y \mid Z) = \left[\mathbb{1}(Y=0)\right]^{1-Z} \cdot \left[f_B(Y; n, p)\right]^{Z}, \tag{2.25}$$

where $f_B(\cdot)$ denotes the binomial pmf.

Proposition 9 *Assume $Y|Z \sim ZIB(n, p, Z)$. Then, the conditional expectation of $Y|Z$ is*

$$\mathbb{E}(Y|Z) = Znp.$$

Proposition 10 *Assume $Y|Z \sim ZIB(n, p, Z)$. Then, we have*

$$\mathbb{E}(ZY|Z) = \mathbb{E}(Y|Z) = Znp.$$

2.5 Model Summary

n		length of the structure
$\boldsymbol{X_m}$	$R^{3\times n}$	maternal structure
$\boldsymbol{X_p}$	$R^{3\times n}$	paternal structure
$\boldsymbol{x_{m_i}}$	R^3	i-th bead in structure $\boldsymbol{X_m}$
η	$\{m,p\}$	maternal or paternal indicator
θ	$\{m,p\}$	maternal or paternal indicator
$T_{\eta_i\theta_j}$		true allele-specific contact frequency between η_i and θ_j
$C_{\eta_i\theta_j}$		allele-specific contact frequency between η_i and θ_j where both η_i and θ_j are allele-identifiable
$C_{\eta_i^*\theta_j}$		allele-specific contact frequency between η_i and θ_j where only θ_j is allele-identifiable
$C_{\eta_i\theta_j^*}$		allele-specific contact frequency between η_i and θ_j where only η_i is allele-identifiable
$C_{\eta_i^*\theta_j^*}$		allele-specific contact frequency between η_i and θ_j where neither η_i or θ_j is allele-identifiable
$C_{x_i\theta_j}$		observed allele-ambiguous contact frequency between x_i (either m_i or p_i) and θ_j
$C_{\eta_ix_j}$		observed allele-ambiguous contact frequency between η_i and x_j (either m_j or p_j)
$C_{x_ix_j}$		observed allele-ambiguous contact frequency between x_i (either m_i or p_i) and x_j (either m_j or p_j)
$Z_{\eta_i\theta_j}$		Poisson state indicator for contacts between η_i and θ_j
$\gamma_{\eta_i\theta_j}$		Bernoulli prior for $Z_{\eta_i\theta_j}$
$d_{\eta_i\theta_j}$		spatial distance between $\boldsymbol{x_{\eta_i}}$ and $\boldsymbol{x_{\theta_j}}$
$\lambda_{\eta_i\theta_j}$		Poisson parameter
α		spatial distance decay exponent
β		scaling factor
$\boldsymbol{q}$	R^n	allele-identifiable probability

Table 2.1: Notation used in ASHIC models.

2.5.1 ASHIC-PM Model

In the ASHIC-PM model, we have

$$T_{\eta_i\theta_j} \sim \text{Poisson}(\lambda_{\eta_i\theta_j} = \beta d^{\alpha}_{\eta_i\theta_j}), \tag{2.4}$$

$$C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}, C_{\eta_i^*\theta_j^*} \mid T_{\eta_i\theta_j} \sim \text{Multinomial}\left(T_{\eta_i\theta_j}, \{q_{ij}, q_{i\bar{j}}, q_{\bar{i}j}, q_{\bar{i}\bar{j}}\}\right). \tag{2.11}$$

Based on the ASHIC-PM model, we can derive

$$\begin{aligned}
C_{\eta_i\theta_j} &\sim \text{Poisson}(q_{ij}\lambda_{\eta_i\theta_j}),\\
C_{\eta_i\theta_j^*} &\sim \text{Poisson}(q_{i\bar{j}}\lambda_{\eta_i\theta_j}),\\
C_{\eta_i^*\theta_j} &\sim \text{Poisson}(q_{\bar{i}j}\lambda_{\eta_i\theta_j}),\\
C_{\eta_i^*\theta_j^*} &\sim \text{Poisson}(q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}),
\end{aligned} \tag{2.12}$$

$$\begin{aligned}
C_{\eta_i x_j} &\sim \text{Poisson}\left(q_{i\bar{j}}(\lambda_{\eta_i m_j} + \lambda_{\eta_i p_j})\right),\\
C_{x_i\theta_j} &\sim \text{Poisson}\left(q_{\bar{i}j}(\lambda_{m_i\theta_j} + \lambda_{p_i\theta_j})\right),\\
C_{x_i x_j} &\sim \text{Poisson}\left(q_{\bar{i}\bar{j}}(\lambda_{m_i m_j} + \lambda_{m_i p_j} + \lambda_{p_i m_j} + \lambda_{p_i p_j})\right),
\end{aligned} \tag{2.13}$$

$$\begin{aligned}
C_{\eta_i\theta_j^*} \mid C_{\eta_i x_j} &\sim \text{Binomial}\left(C_{\eta_i x_j}, \frac{\lambda_{\eta_i\theta_j}}{\lambda_{\eta_i m_j} + \lambda_{\eta_i p_j}}\right),\\
C_{\eta_i^*\theta_j} \mid C_{x_i\theta_j} &\sim \text{Binomial}\left(C_{x_i\theta_j}, \frac{\lambda_{\eta_i\theta_j}}{\lambda_{m_i\theta_j} + \lambda_{p_i\theta_j}}\right),\\
C_{\eta_i^*\theta_j^*} \mid C_{x_i x_j} &\sim \text{Binomial}\left(C_{x_i x_j}, \frac{\lambda_{\eta_i\theta_j}}{\lambda_{m_i m_j} + \lambda_{m_i p_j} + \lambda_{p_i m_j} + \lambda_{p_i p_j}}\right).
\end{aligned} \tag{2.14}$$

2.5.2 ASHIC-ZIPM Model

In the ASHIC-ZIPM model, we have

$$Z_{\eta_i\theta_j} \sim \text{Bernoulli}(\gamma_{\eta_i\theta_j}), \tag{2.6}$$

$$T_{\eta_i\theta_j} \mid Z_{\eta_i\theta_j} \sim \mathrm{ZIP}(\lambda_{\eta_i\theta_j} = \beta d^{\alpha}_{\eta_i\theta_j}, Z_{\eta_i\theta_j}), \tag{2.5}$$

$$C_{\eta_i\theta_j}, C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}, C_{\eta_i^*\theta_j^*} \mid T_{\eta_i\theta_j}, Z_{\eta_i\theta_j} \sim \mathrm{ZIM}\left(T_{\eta_i\theta_j}, \{q_{ij}, q_{i\bar{j}}, q_{\bar{i}j}, q_{\bar{i}\bar{j}}\}, Z_{\eta_i\theta_j}\right). \tag{2.16}$$

Then, we can further derive

$$\begin{aligned} C_{\eta_i\theta_j} \mid Z_{\eta_i\theta_j} &\sim \mathrm{ZIP}(q_{ij}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j}), \\ C_{\eta_i\theta_j^*} \mid Z_{\eta_i\theta_j} &\sim \mathrm{ZIP}(q_{i\bar{j}}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j}), \\ C_{\eta_i^*\theta_j} \mid Z_{\eta_i\theta_j} &\sim \mathrm{ZIP}(q_{\bar{i}j}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j}), \\ C_{\eta_i^*\theta_j^*} \mid Z_{\eta_i\theta_j} &\sim \mathrm{ZIP}(q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j}), \end{aligned} \tag{2.17}$$

$$C_{\eta_i x_j} \mid Z_{\eta_i m_j}, Z_{\eta_i p_j} \sim \mathrm{ZIP}\left(q_{i\bar{j}} \sum_{\theta'} Z_{\eta_i\theta'_j}\lambda_{\eta_i\theta'_j},\ Z_{\eta_i x_j}\right), \tag{2.19}$$

$$C_{x_i\theta_j} \mid Z_{m_i\theta_j}, Z_{p_i\theta_j} \sim \mathrm{ZIP}\left(q_{\bar{i}j} \sum_{\eta'} Z_{\eta'_i\theta_j}\lambda_{\eta'_i\theta_j},\ Z_{x_i\theta_j}\right), \tag{2.20}$$

$$C_{x_i x_j} \mid Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j} \sim \mathrm{ZIP}\left(q_{\bar{i}\bar{j}} \sum_{\eta',\theta'} Z_{\eta'_i\theta'_j}\lambda_{\eta'_i\theta'_j},\ Z_{x_i x_j}\right), \tag{2.21}$$

$$C_{\eta_i\theta_j^*} \mid C_{\eta_i x_j}, Z_{\eta_i m_j}, Z_{\eta_i p_j} \sim \mathrm{ZIB}\left(C_{\eta_i x_j}, \frac{\lambda_{\eta_i\theta_j}}{\sum_{\theta'} Z_{\eta_i\theta'_j}\lambda_{\eta_i\theta'_j}},\ Z_{\eta_i\theta_j}\right), \tag{2.22}$$

$$C_{\eta_i^*\theta_j} \mid C_{x_i\theta_j}, Z_{m_i\theta_j}, Z_{p_i\theta_j} \sim \mathrm{ZIB}\left(C_{x_i\theta_j}, \frac{\lambda_{\eta_i\theta_j}}{\sum_{\eta'} Z_{\eta'_i\theta_j}\lambda_{\eta'_i\theta_j}},\ Z_{\eta_i\theta_j}\right), \tag{2.23}$$

$$C_{\eta_i^*\theta_j^*} \mid C_{x_i x_j}, Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j} \sim \mathrm{ZIB}\left(C_{x_i x_j}, \frac{\lambda_{\eta_i\theta_j}}{\sum_{\eta',\theta'} Z_{\eta'_i\theta'_j}\lambda_{\eta'_i\theta'_j}},\ Z_{\eta_i\theta_j}\right). \tag{2.24}$$

Note that when $\boldsymbol{Z} = \mathbf{1}$ (that is, $Z_{\eta_i\theta_j} = 1, \forall\eta,\theta \in \{m,p\}, \forall 1 \leq i,j \leq n$), the ASHIC-ZIPM model reduces to the ASHIC-PM model. Hence, the ASHIC-PM model is a special case of the ASHIC-ZIPM model.

2.6 Incorporating Bias Factors

It has been shown that the raw contact frequencies obtained from a real Hi-C experiment are affected by various types of systematic biases, including the length of restriction fragments, the GC content of ligation junctions, and the sequence uniqueness (mappability) [22]. These types of biases depend on the DNA sequences, thus can be affected by allelic sequence divergence such as SNPs, especially in high-SNP-density systems. In addition, depending on the choice of read mapping strategies, there may exist a mapping bias towards the reference allele. That is, reads containing the reference allele are more likely to be mapped. Therefore, we consider the bias factors to be allele-specific.

2.6.1 ASHIC-PM Bias-Incorporated Model

According to Imakaev et al. [23], the bias of observing Hi-C contacts between two chromatin loci η_i and θ_j can be factorized as the product of the bias factors b_{η_i} and b_{θ_j} of the contacting loci, respectively. Therefore, under the ASHIC-PM model, we have

$$T_{\eta_i\theta_j} \sim \text{Poisson}(b_{\eta_i} b_{\theta_j} \lambda_{\eta_i\theta_j}). \tag{2.26}$$

Consequently, the observed allele-certain, hidden allele-specific, and the observed allele-ambiguous contact frequencies follow the modified Poisson distributions:

$$\begin{aligned}
C_{\eta_i\theta_j} &\sim \text{Poisson}(q_{ij} b_{\eta_i} b_{\theta_j} \lambda_{\eta_i\theta_j}), \\
C_{\eta_i\theta_j^*} &\sim \text{Poisson}(q_{i\bar{j}} b_{\eta_i} b_{\theta_j} \lambda_{\eta_i\theta_j}), \\
C_{\eta_i^*\theta_j} &\sim \text{Poisson}(q_{\bar{i}j} b_{\eta_i} b_{\theta_j} \lambda_{\eta_i\theta_j}), \\
C_{\eta_i^*\theta_j^*} &\sim \text{Poisson}(q_{\bar{i}\bar{j}} b_{\eta_i} b_{\theta_j} \lambda_{\eta_i\theta_j}),
\end{aligned} \tag{2.27}$$

$$
\begin{aligned}
C_{\eta_i x_j} &\sim \text{Poisson}\left(q_{i\bar{j}} \sum_{\theta'} b_{\eta_i} b_{\theta'_j} \lambda_{\eta_i \theta'_j}\right), \\
C_{x_i \theta_j} &\sim \text{Poisson}\left(q_{\bar{i}j} \sum_{\eta'} b_{\eta'_i} b_{\theta_j} \lambda_{\eta'_i \theta_j}\right), \\
C_{x_i x_j} &\sim \text{Poisson}\left(q_{\bar{i}\bar{j}} \sum_{\eta',\theta'} b_{\eta'_i} b_{\theta'_j} \lambda_{\eta'_i \theta'_j}\right).
\end{aligned}
\tag{2.28}
$$

The conditional distributions of the hidden allele-specific contact frequencies given the observed allele-ambiguous contact frequencies follow:

$$
\begin{aligned}
C_{\eta_i \theta_j^*} \mid C_{\eta_i x_j} &\sim \text{Binomial}\left(C_{\eta_i x_j}, \frac{b_{\eta_i} b_{\theta_j} \lambda_{\eta_i \theta_j}}{\sum_{\theta'} b_{\eta_i} b_{\theta'_j} \lambda_{\eta_i \theta'_j}}\right), \\
C_{\eta_i^* \theta_j} \mid C_{x_i \theta_j} &\sim \text{Binomial}\left(C_{x_i \theta_j}, \frac{b_{\eta_i} b_{\theta_j} \lambda_{\eta_i \theta_j}}{\sum_{\eta'} b_{\eta'_i} b_{\theta_j} \lambda_{\eta'_i \theta_j}}\right), \\
C_{\eta_i^* \theta_j^*} \mid C_{x_i x_j} &\sim \text{Binomial}\left(C_{x_i x_j}, \frac{b_{\eta_i} b_{\theta_j} \lambda_{\eta_i \theta_j}}{\sum_{\eta',\theta'} b_{\eta'_i} b_{\theta'_j} \lambda_{\eta'_i \theta'_j}}\right).
\end{aligned}
\tag{2.29}
$$

2.6.2 ASHIC-ZIPM Bias-Incorporated Model

Similarly, under the ASHIC-ZIPM model, we have

$$
T_{\eta_i \theta_j} \mid Z_{\eta_i, \theta_j} \sim \text{ZIP}(b_{\eta_i} b_{\theta_j} \lambda_{\eta_i \theta_j}, Z_{\eta_i \theta_j}). \tag{2.30}
$$

Then, the observed allele-certain, hidden allele-specific, and the observed allele-ambiguous contact frequencies follow the modified ZIP distributions:

$$
\begin{aligned}
C_{\eta_i \theta_j} \mid Z_{\eta_i \theta_j} &\sim \text{ZIP}(q_{ij} b_{\eta_i} b_{\theta_j} \lambda_{\eta_i \theta_j}, Z_{\eta_i \theta_j}), \\
C_{\eta_i \theta_j^*} \mid Z_{\eta_i \theta_j} &\sim \text{ZIP}(q_{i\bar{j}} b_{\eta_i} b_{\theta_j} \lambda_{\eta_i \theta_j}, Z_{\eta_i \theta_j}), \\
C_{\eta_i^* \theta_j} \mid Z_{\eta_i \theta_j} &\sim \text{ZIP}(q_{\bar{i}j} b_{\eta_i} b_{\theta_j} \lambda_{\eta_i \theta_j}, Z_{\eta_i \theta_j}), \\
C_{\eta_i^* \theta_j^*} \mid Z_{\eta_i \theta_j} &\sim \text{ZIP}(q_{\bar{i}\bar{j}} b_{\eta_i} b_{\theta_j} \lambda_{\eta_i \theta_j}, Z_{\eta_i \theta_j}).
\end{aligned}
\tag{2.31}
$$

$$\begin{aligned}
C_{\eta_i x_j} \mid Z_{\eta_i m_j}, Z_{\eta_i p_j} &\sim \mathrm{ZIP}\left(q_{i\bar{j}} \sum_{\theta'} Z_{\eta_i \theta'_j} b_{\eta_i} b_{\theta'_j} \lambda_{\eta_i \theta'_j},\ Z_{\eta_i x_j} \right), \\
C_{x_i \theta_j} \mid Z_{m_i \theta_j}, Z_{p_i \theta_j} &\sim \mathrm{ZIP}\left(q_{\bar{i}j} \sum_{\eta'} Z_{\eta'_i \theta_j} b_{\eta'_i} b_{\theta_j} \lambda_{\eta'_i \theta_j},\ Z_{x_i \theta_j} \right), \qquad (2.32) \\
C_{x_i x_j} \mid Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j} &\sim \mathrm{ZIP}\left(q_{\bar{i}\bar{j}} \sum_{\eta', \theta'} Z_{\eta'_i \theta'_j} b_{\eta'_i} b_{\theta'_j} \lambda_{\eta'_i \theta'_j},\ Z_{x_i x_j} \right).
\end{aligned}$$

The conditional distributions of the hidden allele-specific contact frequencies given the observed allele-ambiguous contact frequencies and $\boldsymbol{Z}$ follow:

$$\begin{aligned}
C_{\eta_i \theta^*_j} \mid C_{\eta_i x_j}, Z_{\eta_i m_j}, Z_{\eta_i p_j} &\sim \mathrm{ZIB}\left(C_{\eta_i x_j}, \frac{b_{\eta_i} b_{\theta_j} \lambda_{\eta_i \theta_j}}{\sum_{\theta'} Z_{\eta_i \theta'_j} b_{\eta_i} b_{\theta'_j} \lambda_{\eta_i \theta'_j}},\ Z_{\eta_i \theta_j} \right), \\
C_{\eta^*_i \theta_j} \mid C_{x_i \theta_j}, Z_{m_i \theta_j}, Z_{p_i \theta_j} &\sim \mathrm{ZIB}\left(C_{x_i \theta_j}, \frac{b_{\eta_i} b_{\theta_j} \lambda_{\eta_i \theta_j}}{\sum_{\eta'} Z_{\eta'_i \theta_j} b_{\eta'_i} b_{\theta_j} \lambda_{\eta'_i \theta_j}},\ Z_{\eta_i \theta_j} \right), \\
C_{\eta^*_i \theta^*_j} \mid C_{x_i x_j}, Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j} &\sim \mathrm{ZIB}\left(C_{x_i x_j}, \frac{b_{\eta_i} b_{\theta_j} \lambda_{\eta_i \theta_j}}{\sum_{\eta', \theta'} Z_{\eta'_i \theta'_j} b_{\eta'_i} b_{\theta'_j} \lambda_{\eta'_i \theta'_j}},\ Z_{\eta_i \theta_j} \right).
\end{aligned}$$

$$(2.33)$$

Chapter 3

Inference via EM Algorithm

We design an EM algorithm to simultaneously infer 3D structures and estimate model parameters. Below, we describe our EM algorithm separately for the ASHIC-PM model and ASHIC-ZIPM model.

3.1 ASHIC-PM Model

In the ASHIC-PM model, the parameter space includes the homologous chromosome structures $\boldsymbol{X_m} \in R^{3\times n}$ and $\boldsymbol{X_p} \in R^{3\times n}$, the distance-decay exponent α, the scaling factor β, and the allele-identifiable probabilities $\boldsymbol{q} = \{q_k\}, 1 \leq k \leq n$.

Recall in eq. (2.4), the Poisson parameter $\lambda_{\eta_i\theta_j}$ is a function of α, β, $\boldsymbol{X_m}$, and $\boldsymbol{X_p}$. Here, we fix α and β to obtain a unique solution for $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$. Specifically, we use true estimate of α in simulations and set $\alpha = -3$ in real data. We set $\beta = 1$ in both cases. Note that $\alpha = -3$ is deduced from the following two relationships based on polymer physics studies: 1) $d \sim s^{B_1}$ between spatial distance d and genomic distance s, and 2) $c \sim s^{B_2}$

between contact count c and genomic distance s. As demonstrated by Lieberman-Aiden et al. [6], chromatin architecture follows a "fractal globule" model, in which $d \sim s^{1/3}$ and $c \sim s^{-1}$, hence $c \sim d^{-3}$. Also it has been shown that the fractal globule model is consistent with the observations in Hi-C data [6] and 3D-FISH data [24]. The parameter β only affects the scale of the estimated 3D structures. If we set β to a larger value, the estimated 3D structures will scale up; on the other hand, if we set β to a smaller value, the structures will shrink accordingly. Therefore, the choice of β does not affect the shape of the resulting structures.

From the diploid Hi-C data, we can directly observe the allele-certain contacts $\boldsymbol{C^O} = \{C_{\eta_i\theta_j}\}$ and the allele-ambiguous contact frequencies $\boldsymbol{C^X} = \{C_{\eta_i x_j}, C_{x_i\theta_j}, C_{x_i x_j}\}$. The unobserved latent variables are the hidden allele-specific contact frequencies $\boldsymbol{C^H} = \{C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}, C_{\eta_i^*\theta_j^*}\}$ according to eq. (2.2).

The goal of the EM algorithm is to find the maximum likelihood estimate (MLE) of the model parameters, reconstruct the allelic 3D structures $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$, and decompose $\boldsymbol{C^X}$ and infer $\boldsymbol{C^H}$ to impute the true allele-specific contact frequencies $\boldsymbol{T} = \{T_{\eta_i\theta_j}\}$ from eq. (2.1).

The complete likelihood of the observed data $\{\boldsymbol{C^O},\ \boldsymbol{C^X}\}$ and the unobserved latent data $\boldsymbol{C^H}$ is

$$\mathcal{L}_c = \mathcal{L}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q} \mid \boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{C^H}\right) = p\left(\boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{C^H} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right). \tag{3.1}$$

The marginal likelihood of the observed data $\boldsymbol{C^O}$ and $\boldsymbol{C^X}$ is

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q} \mid \boldsymbol{C^O}, \boldsymbol{C^X}\right) &= p\left(\boldsymbol{C^O}, \boldsymbol{C^X} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) \\ &= \sum_{\boldsymbol{C^H}} p\left(\boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{C^H} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right).\end{aligned} \tag{3.2}$$

To solve the MLE of the marginal likelihood of observed data $\{\boldsymbol{C^O}, \boldsymbol{C^X}\}$, we propose an EM algorithm which applies the following two steps iteratively:

- Expectation step (E-step):

$$\mathcal{Q} = \mathcal{Q}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}; \boldsymbol{X_m}^{(t)}, \boldsymbol{X_p}^{(t)}, \boldsymbol{q}^{(t)}\right) = \mathbb{E}_{\boldsymbol{C^H} \mid \boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{X_m}^{(t)}, \boldsymbol{X_p}^{(t)}, \boldsymbol{q}^{(t)}} \left(\log \mathcal{L}_c\right).$$

- Maximization step (M-step):

$$\boldsymbol{X_m}^{(t+1)}, \boldsymbol{X_p}^{(t+1)}, \boldsymbol{q}^{(t+1)} = \underset{\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}}{\arg\max}\, \mathcal{Q}.$$

3.1.1 E-step

We can factorize the complete likelihood function in eq. (3.1) as

$$\begin{aligned}
\mathcal{L}_c =& \mathcal{L}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q} \mid \boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{C^H}\right) \\
=& p\left(\boldsymbol{C^O} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) p\left(\boldsymbol{C^H} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) p\left(\boldsymbol{C^X} \mid \boldsymbol{C^H}\right) \\
=& \prod_{i<j} \underbrace{p\left(\boldsymbol{C^O_{ij}} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right)}_{\mathcal{L}^O} \underbrace{p\left(\boldsymbol{C^H_{ij}} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right)}_{\mathcal{L}^H} \underbrace{p\left(\boldsymbol{C^X_{ij}} \mid \boldsymbol{C^H_{ij}}\right)}_{\mathcal{L}^X}.
\end{aligned} \tag{3.3}$$

From eq. (2.12), we can write out

$$\begin{aligned}
\mathcal{L}^O =& p\left(\boldsymbol{C^O_{ij}} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) \\
=& \prod_{\eta,\theta} p\left(C_{\eta_i\theta_j} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) \\
=& \prod_{\eta,\theta} f_{\mathrm{P}}(C_{\eta_i\theta_j}; q_{ij}\lambda_{\eta_i\theta_j}) \\
=& \prod_{\eta,\theta} \frac{\left(q_{ij}\lambda_{\eta_i\theta_j}\right)^{C_{\eta_i\theta_j}} e^{-q_{ij}\lambda_{\eta_i\theta_j}}}{C_{\eta_i\theta_j}!},
\end{aligned} \tag{3.4}$$

$$
\begin{aligned}
\mathcal{L}^H =& p\left(\boldsymbol{C_{ij}^H} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) \\
=& \prod_{\eta,\theta} p\left(C_{\eta_i\theta_j^*} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) p\left(C_{\eta_i^*\theta_j} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) p\left(C_{\eta_i^*\theta_j^*} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) \\
=& \prod_{\eta,\theta} f_{\mathrm{P}}(C_{\eta_i\theta_j^*}; q_{i\bar{j}}\lambda_{\eta_i\theta_j}) f_{\mathrm{P}}(C_{\eta_i^*\theta_j}; q_{\bar{i}j}\lambda_{\eta_i\theta_j}) f_{\mathrm{P}}(C_{\eta_i^*\theta_j^*}; q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}) \\
=& \prod_{\eta,\theta} \frac{\left(q_{i\bar{j}}\lambda_{\eta_i\theta_j}\right)^{C_{\eta_i\theta_j^*}} e^{-q_{i\bar{j}}\lambda_{\eta_i\theta_j}}}{C_{\eta_i\theta_j^*}!} \frac{\left(q_{\bar{i}j}\lambda_{\eta_i\theta_j}\right)^{C_{\eta_i^*\theta_j}} e^{-q_{\bar{i}j}\lambda_{\eta_i\theta_j}}}{C_{\eta_i^*\theta_j}!} \frac{\left(q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}\right)^{C_{\eta_i^*\theta_j^*}} e^{-q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}}}{C_{\eta_i^*\theta_j^*}!}.
\end{aligned}
\tag{3.5}
$$

Furthermore from eq. (2.2), we have

$$
\begin{aligned}
\mathcal{L}^X =& p\left(\boldsymbol{C_{ij}^X} \mid \boldsymbol{C_{ij}^H}\right) \\
=& \left[\prod_{\eta} p\left(C_{\eta_i x_j} \mid \boldsymbol{C_{ij}^H}\right)\right] \left[\prod_{\theta} p(C_{x_i\theta_j} \mid \boldsymbol{C_{ij}^H})\right] p(C_{x_i x_j} \mid \boldsymbol{C_{ij}^H}) \\
=& \left[\prod_{\eta} \mathbb{1}\Big(C_{\eta_i x_j} = C_{\eta_i m_j^*} + C_{\eta_i p_j^*}\Big)\right] \left[\prod_{\theta} \mathbb{1}\left(C_{x_i\theta_j} = C_{m_i^*\theta_j} + C_{p_i^*\theta_j}\right)\right] \cdot \\
& \mathbb{1}\Big(C_{x_i x_j} = C_{m_i^* m_j^*} + C_{m_i^* p_j^*} + C_{p_i^* m_j^*} + C_{p_i^* p_j^*}\Big).
\end{aligned}
\tag{3.6}
$$

Moreover, the complete log-likelihood function can be written as

$$
\log \mathcal{L}_c = \sum_{i<j} \left(\log \mathcal{L}^O + \log \mathcal{L}^H + \log \mathcal{L}^X\right). \tag{3.7}
$$

In the E-step, we need to calculate the conditional expectation of the complete log-likelihood function given the observed data and the current parameter estimation, denoted by $\mathbb{E}_{\boldsymbol{C^H}|\boldsymbol{C^O},\boldsymbol{C^X},\boldsymbol{X_m^{(t)}},\boldsymbol{X_p^{(t)}},\boldsymbol{q}^{(t)}}\left(\log \mathcal{L}_c\right)$.

In short, we use $\mathbb{E}_{\boldsymbol{C^H}|\bullet}\left(\right)$ to replace $\mathbb{E}_{\boldsymbol{C^H}|\boldsymbol{C^O},\boldsymbol{C^X},\boldsymbol{X_m^{(t)}},\boldsymbol{X_p^{(t)}},\boldsymbol{q}^{(t)}}\left(\right)$. We can show that

$$
\begin{aligned}
\mathcal{Q} =& \mathcal{Q}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}; \boldsymbol{X_m^{(t)}}, \boldsymbol{X_p^{(t)}}, \boldsymbol{q}^{(t)}\right) \\
=& \mathbb{E}_{\boldsymbol{C^H}|\bullet}\left(\log \mathcal{L}_c\right) \\
=& \sum_{i<j} \left(\log \mathcal{L}^O + \mathbb{E}_{\boldsymbol{C^H}|\bullet}\left(\log \mathcal{L}^H\right) + \mathbb{E}_{\boldsymbol{C^H}|\bullet}\left(\log \mathcal{L}^X\right)\right).
\end{aligned}
\tag{3.8}
$$

First, we can write

$$\log \mathcal{L}^O = \sum_{\eta,\theta} \left(C_{\eta_i\theta_j} \log q_{ij} + C_{\eta_i\theta_j} \log \lambda_{\eta_i\theta_j} - q_{ij}\lambda_{\eta_i\theta_j} \right) + c. \tag{3.9}$$

Second, we have

$$\begin{aligned}\log \mathcal{L}^H = \sum_{\eta,\theta} \Big[& \Big(C_{\eta_i\theta_j^*} \log q_{i\bar{j}} + C_{\eta_i\theta_j^*} \log \lambda_{\eta_i\theta_j} - q_{i\bar{j}}\lambda_{\eta_i\theta_j} \Big) + \\ & \left(C_{\eta_i^*\theta_j} \log q_{\bar{i}j} + C_{\eta_i^*\theta_j} \log \lambda_{\eta_i\theta_j} - q_{\bar{i}j}\lambda_{\eta_i\theta_j} \right) + \\ & \Big(C_{\eta_i^*\theta_j^*} \log q_{\bar{i}\bar{j}} + C_{\eta_i^*\theta_j^*} \log \lambda_{\eta_i\theta_j} - q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j} \Big) \Big] + c.\end{aligned} \tag{3.10}$$

Recall that in eq. (2.14), at the $(t+1)$-th iteration, we can compute the conditional expectation of $C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}$, and $C_{\eta_i^*\theta_j^*}$ given the observed data and the parameters estimated from the t-th iteration. Specifically, we define

$$\begin{aligned}\mathbb{C}_{\eta_i\theta_j^*} &:= \mathbb{E}\left(C_{\eta_i\theta_j^*} \mid C_{\eta_i x_j}; \lambda^{(t)}_{\eta_i\theta_j}, \lambda^{(t)}_{\eta_i\tilde{\theta}_j} \right) = \frac{\lambda^{(t)}_{\eta_i\theta_j}}{\lambda^{(t)}_{\eta_i\theta_j} + \lambda^{(t)}_{\eta_i\tilde{\theta}_j}} C_{\eta_i x_j}, \\ \mathbb{C}_{\eta_i^*\theta_j} &:= \mathbb{E}\left(C_{\eta_i^*\theta_j} \mid C_{x_i\theta_j}; \lambda^{(t)}_{\eta_i\theta_j}, \lambda^{(t)}_{\tilde{\eta}_i\theta_j} \right) = \frac{\lambda^{(t)}_{\eta_i\theta_j}}{\lambda^{(t)}_{\eta_i\theta_j} + \lambda^{(t)}_{\tilde{\eta}_i\theta_j}} C_{x_i\theta_j}, \\ \mathbb{C}_{\eta_i^*\theta_j^*} &:= \mathbb{E}\left(C_{\eta_i^*\theta_j^*} \mid C_{x_i x_j}; \lambda^{(t)}_{\eta_i\theta_j}, \lambda^{(t)}_{\eta_i\tilde{\theta}_j}, \lambda^{(t)}_{\tilde{\eta}_i\theta_j}, \lambda^{(t)}_{\tilde{\eta}_i\tilde{\theta}_j} \right) = \frac{\lambda^{(t)}_{\eta_i\theta_j}}{\lambda^{(t)}_{\eta_i\theta_j} + \lambda^{(t)}_{\eta_i\tilde{\theta}_j} + \lambda^{(t)}_{\tilde{\eta}_i\theta_j} + \lambda^{(t)}_{\tilde{\eta}_i\tilde{\theta}_j}} C_{x_i x_j},\end{aligned} \tag{3.11}$$

where $\tilde{\theta}$ is the opposite allele of θ ($\tilde{\theta} = m$ if $\theta = p$; $\tilde{\theta} = p$ if $\theta = m$), $\tilde{\eta}$ is the opposite allele of η ($\tilde{\eta} = m$ if $\eta = p$; $\tilde{\eta} = p$ if $\eta = m$), and $\lambda^{(t)}_{\eta_i\theta_j} = \beta \left(d^{(t)}_{\eta_i\theta_j} \right)^{\alpha}$.

Consequently, we have

$$\begin{aligned} C_{\eta_i x_j} =& \mathbb{C}_{\eta_i m_j^*} + \mathbb{C}_{\eta_i p_j^*}, \\ C_{x_i\theta_j} =& \mathbb{C}_{m_i^*\theta_j} + \mathbb{C}_{p_i^*\theta_j}, \\ C_{x_i x_j} =& \mathbb{C}_{m_i^* m_j^*} + \mathbb{C}_{m_i^* p_j^*} + \mathbb{C}_{p_i^* m_j^*} + \mathbb{C}_{p_i^* p_j^*}, \end{aligned} \tag{3.12}$$

From eq. (3.10) and eq. (3.11), we can show that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{C^H}|\bullet}\left(\log \mathcal{L}^H\right) \\
=&\sum_{\eta,\theta}\Big[\Big(\mathbb{C}_{\eta_i\theta_j^*}\log q_{i\bar{j}}+\mathbb{C}_{\eta_i\theta_j^*}\log\lambda_{\eta_i\theta_j}-q_{i\bar{j}}\lambda_{\eta_i\theta_j}\Big)+ \\
&\left(\mathbb{C}_{\eta_i^*\theta_j}\log q_{\bar{i}j}+\mathbb{C}_{\eta_i^*\theta_j}\log\lambda_{\eta_i\theta_j}-q_{\bar{i}j}\lambda_{\eta_i\theta_j}\right)+ \\
&\Big(\mathbb{C}_{\eta_i^*\theta_j^*}\log q_{\bar{i}\bar{j}}+\mathbb{C}_{\eta_i^*\theta_j^*}\log\lambda_{\eta_i\theta_j}-q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}\Big)\Big]+c.
\end{aligned} \tag{3.13}$$

Furthermore, from eq. (3.12), we have

$$\mathbb{E}_{\boldsymbol{C^H}|\bullet}\left(\log \mathcal{L}^X\right)=0. \tag{3.14}$$

Taken together, we have

$$\begin{aligned}
\mathcal{Q}=&\sum_{i<j}\sum_{\eta,\theta}\Big[C_{\eta_i\theta_j}\log q_{ij}+\mathbb{C}_{\eta_i\theta_j^*}\log q_{i\bar{j}}+\mathbb{C}_{\eta_i^*\theta_j}\log q_{\bar{i}j}+\mathbb{C}_{\eta_i^*\theta_j^*}\log q_{\bar{i}\bar{j}} \\
&+\Big(C_{\eta_i\theta_j}+\mathbb{C}_{\eta_i\theta_j^*}+\mathbb{C}_{\eta_i^*\theta_j}+\mathbb{C}_{\eta_i^*\theta_j^*}\Big)\log\lambda_{\eta_i\theta_j}-\lambda_{\eta_i\theta_j}\Big]+c \\
=&\sum_{i<j}\sum_{\eta,\theta}\Big[\Big(C_{\eta_i\theta_j}+\mathbb{C}_{\eta_i\theta_j^*}\Big)\log q_i+\left(C_{\eta_i\theta_j}+\mathbb{C}_{\eta_i^*\theta_j}\right)\log q_j \\
&+\Big(\mathbb{C}_{\eta_i^*\theta_j}+\mathbb{C}_{\eta_i^*\theta_j^*}\Big)\log\left(1-q_i\right)+\Big(\mathbb{C}_{\eta_i\theta_j^*}+\mathbb{C}_{\eta_i^*\theta_j^*}\Big)\log\left(1-q_j\right) \\
&+\Big(C_{\eta_i\theta_j}+\mathbb{C}_{\eta_i\theta_j^*}+\mathbb{C}_{\eta_i^*\theta_j}+\mathbb{C}_{\eta_i^*\theta_j^*}\Big)\log\lambda_{\eta_i\theta_j}-\lambda_{\eta_i\theta_j}\Big]+c,
\end{aligned} \tag{3.15}$$

where $\lambda_{\eta_i\theta_j}=\beta\left(d_{\eta_i\theta_j}\right)^{\alpha}$.

3.1.2 M-step

In the M-step, we aim to find the optimal solution of $\boldsymbol{X_m},\boldsymbol{X_p},\boldsymbol{q}$ that maximizes the conditional expectation of the complete log-likelihood:

$$\boldsymbol{X_m}^{(t+1)},\boldsymbol{X_p}^{(t+1)},\boldsymbol{q}^{(t+1)}=\underset{\boldsymbol{X_m},\boldsymbol{X_p},\boldsymbol{q}}{\arg\max}\,\mathcal{Q},$$

where $\mathcal{Q} = \mathcal{Q}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}; \boldsymbol{X_m^{(t)}}, \boldsymbol{X_p^{(t)}}, \boldsymbol{q}^{(t)}\right) = \mathbb{E}_{\boldsymbol{C^H}|\bullet}\left(\log \mathcal{L}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q} \mid \boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{C^H}\right)\right)$. By setting the partial derivative of $\mathcal{Q}$ with respect to q_k to 0 (i.e., $\frac{\partial \mathcal{Q}}{\partial q_k} = 0$) for $k = 1, \cdots, n$, we can obtain a closed-form solution of $\boldsymbol{q}$:

$$\begin{aligned}
q_k^{(t+1)} =& \frac{S_1(k)}{S_1(k) + S_2(k)}, \\
S_1(k) =& \sum_{\eta,\theta}\left[\sum_{i<k}\left(C_{\eta_i\theta_k} + \mathbb{C}_{\eta_i^*\theta_k}\right) + \sum_{j>k}\left(C_{\eta_k\theta_j} + \mathbb{C}_{\eta_k\theta_j^*}\right)\right], \\
S_2(k) =& \sum_{\eta,\theta}\left[\sum_{i<k}\left(\mathbb{C}_{\eta_i\theta_k^*} + \mathbb{C}_{\eta_i^*\theta_k^*}\right) + \sum_{j>k}\left(\mathbb{C}_{\eta_k^*\theta_j} + \mathbb{C}_{\eta_k^*\theta_j^*}\right)\right].
\end{aligned} \tag{3.16}$$

Note that $S_1(k)$ can be regarded as the sum of the estimated allele-specific contact frequency between bin k and other bins where the read end at bin k can be distinguished between alleles m and p; $S_2(k)$ can be regarded as the sum of estimated allele-specific contact frequency between bin k and other bins where the read end at bin k are not allele-identifiable. Based on eq. (3.12), we can simplify eq. (3.16) as follows:

$$\begin{aligned}
S_1(k) =& \sum_{i<k}\left(\sum_{\eta,\theta} C_{\eta_i\theta_k} + \sum_{\theta} C_{x_i\theta_k}\right) + \sum_{j>k}\left(\sum_{\eta,\theta} C_{\eta_k\theta_j} + \sum_{\eta} C_{\eta_k x_j}\right), \\
S_2(k) =& \sum_{i<k}\left(\sum_{\eta} C_{\eta_i x_k} + C_{x_i x_k}\right) + \sum_{j>k}\left(\sum_{\theta} C_{x_k\theta_j} + C_{x_k x_j}\right).
\end{aligned} \tag{3.17}$$

Note that $S_1(k)$ and $S_2(k)$ contain only observed values; therefore, we only need to estimate $\boldsymbol{q}$ once at the beginning.

There is no closed-form solution for the 3D structures $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$. Therefore, we use a nonlinear optimizer, the L-BFGS-B algorithm [25] (`fmin_l_bfgs_b` from the `SciPy` package), to iteratively update $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$. The L-BFGS-B algorithm only needs the derivative of a function to determine the direction of steepest descent and obtain an estimate of the Hessian matrix in each iteration.

For instance, the derivative of the expectation of log-likelihood with respect to $\boldsymbol{X_m}$ is:

$$\frac{\partial \mathcal{Q}}{\partial \boldsymbol{X_m}} = \begin{bmatrix} \frac{\partial \mathcal{Q}}{\partial x_{m_{11}}} & \cdots & \frac{\partial \mathcal{Q}}{\partial x_{m_{n1}}} \\ \frac{\partial \mathcal{Q}}{\partial x_{m_{12}}} & \cdots & \frac{\partial \mathcal{Q}}{\partial x_{m_{n2}}} \\ \frac{\partial \mathcal{Q}}{\partial x_{m_{13}}} & \cdots & \frac{\partial \mathcal{Q}}{\partial x_{m_{n3}}} \end{bmatrix}. \tag{3.18}$$

From eq. (3.15), we have

$$\frac{\partial \mathcal{Q}}{\partial x_{m_{k1}}} = \sum_{i<j} \sum_{\eta,\theta} \alpha \left(\frac{\mathbb{T}_{\eta_i\theta_j}}{d_{\eta_i\theta_j}} - \beta d_{\eta_i\theta_j}^{\alpha-1} \right) \frac{\partial d_{\eta_i\theta_j}}{\partial x_{m_{k1}}}. \quad 1 \le k \le n \tag{3.19}$$

Here, we define $\mathbb{T}_{\eta_i\theta_j}$ as the estimated allele-specific contact frequency between η_i and θ_j:

$$\mathbb{T}_{\eta_i\theta_j} := C_{\eta_i\theta_j} + \mathbb{C}_{\eta_i\theta_j^*} + \mathbb{C}_{\eta_i^*\theta_j} + \mathbb{C}_{\eta_i^*\theta_j^*}. \tag{3.20}$$

From eq. (2.3), we have

$$\frac{\partial d_{\eta_i\theta_j}}{\partial x_{m_{k1}}} = \begin{cases} \frac{x_{\eta_{i1}} - x_{\theta_{j1}}}{d_{\eta_i\theta_j}}, & \text{if } \eta = m, i = k; \\ \frac{x_{\theta_{j1}} - x_{\eta_{i1}}}{d_{\eta_i\theta_j}}, & \text{if } \theta = m, j = k; \\ 0, & \text{otherwise.} \end{cases} \tag{3.21}$$

Therefore, we have

$$\frac{\partial \mathcal{Q}}{\partial x_{m_{k1}}} = \sum_{j \ne k} \sum_{\theta} \alpha \left(\frac{\mathbb{T}_{m_k\theta_j}}{d_{m_k\theta_j}} - \beta d_{m_k\theta_j}^{\alpha-1} \right) \frac{x_{m_{k1}} - x_{\theta_{j1}}}{d_{m_k\theta_j}}. \tag{3.22}$$

Similarly, we can write out the partial derivatives with respect to $x_{m_{k2}}$ and $x_{m_{k3}}$.

```
Input: observed allele-certain contact counts $\boldsymbol{C^O}$ and allele-ambiguous
       contact counts $\boldsymbol{C^X}$.
Output: estimated total allele-specific contact matrix $\mathbb{T}$, allelic structures
        $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$, and allele-identifiable probabilities $\boldsymbol{q}$.
Compute the allele-identifiable probabilities $\boldsymbol{q}$ with eq. (3.16);
Initialize the allelic structures $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$;
while not converge do
    ⊡ E-step
    for $\eta, \theta \in \{m, p\}$, $1 \leq i < j \leq n$ do
        Update hidden allele-specific counts $\mathbb{C}_{\eta_i\theta_j^*}$, $\mathbb{C}_{\eta_i^*\theta_j}$, $\mathbb{C}_{\eta_i^*\theta_j^*}$ with eq. (3.11);
        Update total allele-specific counts $\mathbb{T}_{\eta_i\theta_j}$ with eq. (3.20);
    end
    ⊡ M-step
    Update allelic structures $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$ by eq. (3.22) with current $\mathbb{T}$;
end
return $\mathbb{T}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}$
```

Algorithm 1: EM algorithm for the ASHIC-PM model

3.2 ASHIC-ZIPM Model

In the ASHIC-ZIPM model, the parameter space contains the homologous chromosome structures $\boldsymbol{X_m} \in R^{3\times n}$ and $\boldsymbol{X_p} \in R^{3\times n}$, the distance-decay exponent α, the scaling factor β, the hyper parameter $\boldsymbol{\gamma} = \{\gamma_{\eta_i\theta_j}\}$ (for the Bernoulli prior distribution of the Poisson state latent variables $\boldsymbol{Z} = \{Z_{\eta_i\theta_j}\}$), and the allele-identifiable probabilities $\boldsymbol{q} = \{q_k\}$, $1 \leq k \leq n$. Note that in eq. (2.5), the ZIP parameter $\lambda_{\eta_i\theta_j}$ is a function of α, β, $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$. Similar to the ASHIC-PM model, we fix α and β in order to obtain a unique solution for $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$.

From the diploid Hi-C data, we can directly observe the allele-certain contacts $\boldsymbol{C^O} = \{C_{\eta_i\theta_j}\}$ and the allele-ambiguous contact frequencies $\boldsymbol{C^X} = \{C_{\eta_i x_j}, C_{x_i\theta_j}, C_{x_i x_j}\}$. The unobserved latent variables include the hidden allele-specific contact frequencies $\boldsymbol{C^H} = \{C_{\eta_i\theta_j^*}, C_{\eta_i^*\theta_j}, C_{\eta_i^*\theta_j^*}\}$ as defined in eq. (2.2) and the Poisson state latent variables $\boldsymbol{Z}$. The goal of the EM algorithm is to find the MLE of the model parameters, reconstruct the allelic 3D structures $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$, and decompose $\boldsymbol{C^X}$ and infer $\boldsymbol{C^H}$ to impute the true allele-specific frequencies $\boldsymbol{T} = \{T_{\eta_i\theta_j}\}$ from eq. (2.1).

The complete likelihood of the observed data $\{\boldsymbol{C^O},\ \boldsymbol{C^X}\}$ and the unobserved latent variables $\{\boldsymbol{C^H},\ \boldsymbol{Z}\}$ is

$$\begin{aligned}\mathcal{L}_c =& \mathcal{L}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{\gamma}, \boldsymbol{q} \mid \boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{C^H}, \boldsymbol{Z}\right) \\ &= p\left(\boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{C^H}, \boldsymbol{Z} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \gamma, \boldsymbol{q}\right) \\ &= p\left(\boldsymbol{Z} \mid \gamma\right) p\left(\boldsymbol{C^O} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) p\left(\boldsymbol{C^H} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) p\left(\boldsymbol{C^X} \mid \boldsymbol{C^H}\right).\end{aligned} \tag{3.23}$$

The marginal likelihood of the observed data $\boldsymbol{C^O}$ and $\boldsymbol{C^X}$ is:

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \gamma, \boldsymbol{q} \mid \boldsymbol{C^O}, \boldsymbol{C^X}\right) &= P\left(\boldsymbol{C^O}, \boldsymbol{C^X} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \gamma, \boldsymbol{q}\right) \\ &= \sum_{\boldsymbol{C^H}} \sum_{\boldsymbol{Z}} P\left(\boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{C^H}, \boldsymbol{Z} \mid \boldsymbol{X_m}, \boldsymbol{X_p}, \gamma, \boldsymbol{q}\right).\end{aligned} \tag{3.24}$$

To solve the MLE of the marginal likelihood of observed data $\{\boldsymbol{C^O}, \boldsymbol{C^X}\}$, we propose the EM algorithm, which iteratively applies the following two steps:

- Expectation step (E-step):

$$\mathcal{Q} = \mathcal{Q}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \gamma, \boldsymbol{q}; \boldsymbol{X_m^{(t)}}, \boldsymbol{X_p^{(t)}}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{q}^{(t)}\right) = \mathbb{E}_{\boldsymbol{C^H}, \boldsymbol{Z} \mid \boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{X_m^{(t)}}, \boldsymbol{X_p^{(t)}}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{q}^{(t)}}\left(\log \mathcal{L}_c\right).$$

- Maximization step (M-step):

$$\boldsymbol{X_m^{(t+1)}}, \boldsymbol{X_p^{(t+1)}}, \boldsymbol{\gamma}^{(t+1)}, \boldsymbol{q}^{(t+1)} = \underset{\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{\gamma}, \boldsymbol{q}}{\arg\max}\ \mathcal{Q}.$$

3.2.1 E-step

We can factorize the complete likelihood function in eq. (3.23) as

$$\begin{aligned}\mathcal{L}_c =& \mathcal{L}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \gamma, \boldsymbol{q} \mid \boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{C^H}, \boldsymbol{Z}\right) \\ =& p\left(\boldsymbol{Z} \mid \gamma\right) p\left(\boldsymbol{C^O} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) p\left(\boldsymbol{C^H} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) p\left(\boldsymbol{C^X} \mid \boldsymbol{C^H}\right) \\ =& \prod_{i<j} \underbrace{p\left(\boldsymbol{Z_{ij}} \mid \gamma\right)}_{\mathcal{L}^Z} \underbrace{p\left(\boldsymbol{C^O_{ij}} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right)}_{\mathcal{L}^O} \underbrace{p\left(\boldsymbol{C^H_{ij}} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right)}_{\mathcal{L}^H} \underbrace{p\left(\boldsymbol{C^X_{ij}} \mid \boldsymbol{C^H_{ij}}\right)}_{\mathcal{L}^X},\end{aligned} \tag{3.25}$$

where

$$\begin{aligned}\mathcal{L}^Z =& p\left(\boldsymbol{Z_{ij}} \mid \gamma\right) \\ =& \prod_{\eta,\theta} p\left(Z_{\eta_i\theta_j} \mid \gamma_{\eta_i\theta_j}\right) \\ =& \prod_{\eta,\theta} \left(\gamma_{\eta_i\theta_j}\right)^{Z_{\eta_i\theta_j}} \left(1 - \gamma_{\eta_i\theta_j}\right)^{1 - Z_{\eta_i\theta_j}}.\end{aligned} \tag{3.26}$$

From eq. (2.17), we can write out:

$$\begin{aligned}
\mathcal{L}^O =& p\left(\boldsymbol{C^O_{ij}} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) \\
=& \prod_{\eta,\theta} p\left(C_{\eta_i\theta_j} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) \\
=& \prod_{\eta,\theta} f_{\text{ZIP}}(C_{\eta_i\theta_j}; q_{ij}\lambda_{\eta_i\theta_j}, Z_{\eta_i\theta_j}) \\
=& \prod_{\eta,\theta} \left[f_{\text{P}}(C_{\eta_i\theta_j}; q_{ij}\lambda_{\eta_i\theta_j})\right]^{Z_{\eta_i\theta_j}} \left[\mathbb{1}\left(C_{\eta_i\theta_j} = 0\right)\right]^{1-Z_{\eta_i\theta_j}}
\end{aligned} \tag{3.27}$$

and

$$\begin{aligned}
\mathcal{L}^H =& p\left(\boldsymbol{C^H_{ij}} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) \\
=& \prod_{\eta,\theta} p\left(C_{\eta_i\theta^*_j} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) p\left(C_{\eta^*_i\theta_j} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) p\left(C_{\eta^*_i\theta^*_j} \mid \boldsymbol{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}\right) \\
=& \prod_{\eta,\theta} f_{\text{ZIP}}(C_{\eta_i\theta^*_j}; q_{i\bar{j}}\lambda_{\eta_i\theta_j}) f_{\text{ZIP}}(C_{\eta^*_i\theta_j}; q_{\bar{i}j}\lambda_{\eta_i\theta_j}) f_{\text{ZIP}}(C_{\eta^*_i\theta^*_j}; q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}).
\end{aligned} \tag{3.28}$$

Furthermore, from eq. (2.2), we have the same expression of $\mathcal{L}^X$ as in the ASHIC-PM model:

$$\begin{aligned}
\mathcal{L}^X =& p\left(\boldsymbol{C^X_{ij}} \mid \boldsymbol{C^H_{ij}}\right) \\
=& \left[\prod_{\eta} p\left(C_{\eta_i x_j} \mid \boldsymbol{C^H_{ij}}\right)\right] \left[\prod_{\theta} p(C_{x_i\theta_j} \mid \boldsymbol{C^H_{ij}})\right] p(C_{x_i x_j} \mid \boldsymbol{C^H_{ij}}) \\
=& \left[\prod_{\eta} \mathbb{1}\left(C_{\eta_i x_j} = C_{\eta_i m^*_j} + C_{\eta_i p^*_j}\right)\right] \left[\prod_{\theta} \mathbb{1}\left(C_{x_i\theta_j} = C_{m^*_i\theta_j} + C_{p^*_i\theta_j}\right)\right] \cdot \\
& \mathbb{1}\left(C_{x_i x_j} = C_{m^*_i m^*_j} + C_{m^*_i p^*_j} + C_{p^*_i m^*_j} + C_{p^*_i p^*_j}\right).
\end{aligned} \tag{3.29}$$

Moreover, the complete log-likelihood function can be written as

$$\log \mathcal{L}_c = \sum_{i<j} \left(\log \mathcal{L}^Z + \log \mathcal{L}^O + \log \mathcal{L}^H + \log \mathcal{L}^X\right). \tag{3.30}$$

In the E-step, we need to calculate the conditional expectation of the complete log-likelihood function given the observed data and the current parameter estimation, denoted by $\mathbb{E}_{\boldsymbol{C^H},\boldsymbol{Z}|\boldsymbol{C^O},\boldsymbol{C^X},\boldsymbol{X}_{\boldsymbol{m}}^{(t)},\boldsymbol{X}_{\boldsymbol{p}}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{q}^{(t)}}\left(\log \mathcal{L}_c\right)$. In short, we use $\mathbb{E}_{\boldsymbol{C^H},\boldsymbol{Z}|\bullet}()$ to replace $\mathbb{E}_{\boldsymbol{C^H},\boldsymbol{Z}|\boldsymbol{C^O},\boldsymbol{C^X},\boldsymbol{X}_{\boldsymbol{m}}^{(t)},\boldsymbol{X}_{\boldsymbol{p}}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{q}^{(t)}}()$. We can show that:

$$\begin{aligned}\mathcal{Q} =&\mathcal{Q}\left(\boldsymbol{X}_{\boldsymbol{m}}, \boldsymbol{X}_{\boldsymbol{p}}, \boldsymbol{\gamma}, \boldsymbol{q}; \boldsymbol{X}_{\boldsymbol{m}}^{(t)}, \boldsymbol{X}_{\boldsymbol{p}}^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{q}^{(t)}\right)\\ =&\mathbb{E}_{\boldsymbol{C^H},\boldsymbol{Z}|\bullet}\left(\log \mathcal{L}_c\right) \\ =&\sum_{i<j}\left[\mathbb{E}_{\boldsymbol{Z}|\bullet}\left(\log \mathcal{L}^Z\right) + \mathbb{E}_{\boldsymbol{Z}|\bullet}\left(\log \mathcal{L}^O\right) + \mathbb{E}_{\boldsymbol{C^H},\boldsymbol{Z}|\bullet}\left(\log \mathcal{L}^H\right) + \mathbb{E}_{\boldsymbol{C^H}|\bullet}\left(\log \mathcal{L}^X\right)\right].\end{aligned} \tag{3.31}$$

In order to compute $\mathcal{Q}$, we first need to calculate the conditional expectations of $Z_{\eta_i\theta_j}, Z_{\eta_i\theta_j}C_{\eta_i\theta_j^*}, Z_{\eta_i\theta_j}C_{\eta_i^*\theta_j}$, and $Z_{\eta_i\theta_j}C_{\eta_i^*\theta_j^*}$, given the observed data and current estimations of parameters.

Conditional expectation of $Z_{\eta_i\theta_j}$

First, we can compute the conditional expectation of $Z_{\eta_i\theta_j}$, which is the same as the posterior probability of $Z_{\eta_i\theta_j}$, given the observed data observed $= \{C_{\eta_i\theta_j}, C_{\eta_i x_j}, C_{x_i\theta_j}, C_{x_i x_j}\}$ as shown below:

$$\begin{aligned}\mathbb{Z}_{\eta_i\theta_j} :=& \mathbb{E}\left(Z_{\eta_i\theta_j} \mid \text{observed}\right)\\ =&P\left(Z_{\eta_i\theta_j} = 1 \mid \text{observed}\right)\\ =&f\left(Z_{\eta_i\theta_j} \mid \text{observed}\right).\end{aligned} \tag{3.32}$$

Take $Z_{m_im_j}$ as an example:

$$f\left(Z_{m_im_j} \mid \text{observed}\right) = \sum_{Z_{m_ip_j}, Z_{p_im_j}, Z_{p_ip_j}} f\left(Z_{m_im_j}, Z_{m_ip_j}, Z_{p_im_j}, Z_{p_ip_j} \mid \text{observed}\right). \tag{3.33}$$

The posterior joint distribution of Z can be obtained using Bayes' theorem:

$$\begin{aligned}&f(Z_{m_im_j}, Z_{m_ip_j}, Z_{p_im_j}, Z_{p_ip_j} \mid \text{observed})\\ =&\frac{f(\text{observed} \mid Z_{m_im_j}, Z_{m_ip_j}, Z_{p_im_j}, Z_{p_ip_j})f(Z_{m_im_j}, Z_{m_ip_j}, Z_{p_im_j}, Z_{p_ip_j})}{f(\text{observed})},\end{aligned} \tag{3.34}$$

where $f(Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j})$ equals the product of the prior densities:

$$\begin{aligned}
&f(Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j}) \\
=&f(Z_{m_i m_j}) f(Z_{m_i p_j}) f(Z_{p_i m_j}) f(Z_{p_i p_j}) \\
=&\prod_{\eta,\theta} (\gamma_{\eta_i \theta_j})^{Z_{\eta_i \theta_j}} (1-\gamma_{\eta_i \theta_j})^{1-Z_{\eta_i \theta_j}} \\
=&\left(\gamma_{|i-j|}\right)^{Z_{m_i m_j}+Z_{p_i p_j}} \left(1-\gamma_{|i-j|}\right)^{2-Z_{m_i m_j}-Z_{p_i p_j}} \cdot (\gamma_{\text{inter}})^{Z_{m_i p_j}+Z_{p_i m_j}} (1-\gamma_{\text{inter}})^{2-Z_{m_i p_j}-Z_{p_i m_j}}
\end{aligned} \tag{3.35}$$

Additionally, we have:

$$\begin{aligned}
&f\,(\text{observed}) \\
=&\sum_{Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j}} f\left(\text{observed} \mid Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j}\right) \cdot f(Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j})
\end{aligned} \tag{3.36}$$

For a given pair of loci (i, j), we have 9 observed values in total: $C_{m_i m_j}$, $C_{m_i p_j}$, $C_{p_i m_j}$, $C_{p_i p_j}$, $C_{m_i x_j}$, $C_{p_i x_j}$, $C_{x_i m_j}$, $C_{x_i p_j}$, and $C_{x_i x_j}$, which are conditionally independent given $Z_{m_i m_j}$, $Z_{m_i p_j}$, $Z_{p_i m_j}$, and $Z_{p_i p_j}$. Therefore, we have:

$$\begin{aligned}
&f\left(\text{observed} \mid Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j}\right) \\
=&f(C_{m_i m_j}, C_{m_i p_j}, C_{p_i m_j}, C_{p_i p_j}, C_{m_i x_j}, C_{p_i x_j}, C_{x_i m_j}, C_{x_i p_j}, C_{x_i x_j} \mid Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j}) \\
=&\left[\prod_{\eta,\theta} f(C_{\eta_i \theta_j} \mid Z_{\eta_i \theta_j})\right] \cdot \left[\prod_{\eta} f(C_{\eta_i x_j} \mid Z_{\eta_i m_j}, Z_{\eta_i p_j})\right] \cdot \left[\prod_{\theta} f(C_{x_i \theta_j} \mid Z_{m_i \theta_j}, Z_{p_i \theta_j})\right] \cdot \\
&f(C_{x_i x_j} \mid Z_{m_i m_j}, Z_{m_i p_j}, Z_{p_i m_j}, Z_{p_i p_j}),
\end{aligned} \tag{3.37}$$

where the conditional densities are defined in eq. (2.17)-(2.21). By considering eqs. (3.32)-(3.37) together, we compute the value of $\mathbb{Z}_{\eta_i \theta_j}$.

Conditional expectations of $Z_{\eta_i\theta_j}C_{\eta_i\theta_j^*}, Z_{\eta_i\theta_j}C_{\eta_i^*\theta_j},$ and $Z_{\eta_i\theta_j}C_{\eta_i^*\theta_j^*}$

From eq. (2.22) and Proposition 10, at the $(t+1)$-th iteration, we can compute the conditional expectations of $Z_{\eta_i\theta_j}C_{\eta_i\theta_j^*}$, $Z_{\eta_i\theta_j}C_{\eta_i^*\theta_j}$, and $Z_{\eta_i\theta_j}C_{\eta_i^*\theta_j^*}$ given the observed data and the parameters estimated from the t-th iteration:

$$\begin{aligned}
\mathbb{ZC}_{\eta_i\theta_j^*} :=& \mathbb{E}\left(Z_{\eta_i\theta_j}C_{\eta_i\theta_j^*} \mid \text{observed}; \lambda^{(t)}_{\eta_i\theta_j}, \lambda^{(t)}_{\eta_i\tilde{\theta}_j}\right) \\
=& \mathbb{E}\left[\mathbb{E}\left(Z_{\eta_i\theta_j}C_{\eta_i\theta_j^*} \mid Z_{\eta_i\theta_j}, Z_{\eta_i\tilde{\theta}_j}, \text{observed}; \lambda^{(t)}_{\eta_i\theta_j}, \lambda^{(t)}_{\eta_i\tilde{\theta}_j}\right) \mid \text{observed}\right] \\
=& \mathbb{E}\left[\mathbb{E}\left(C_{\eta_i\theta_j^*} \mid C_{\eta_i x_j}, Z_{\eta_i\theta_j}, Z_{\eta_i\tilde{\theta}_j}, \text{observed}; \lambda^{(t)}_{\eta_i\theta_j}, \lambda^{(t)}_{\eta_i\tilde{\theta}_j}\right) \mid \text{observed}\right] \\
=& \mathbb{E}\left[\frac{Z_{\eta_i\theta_j}\lambda^{(t)}_{\eta_i\theta_j}}{Z_{\eta_i\theta_j}\lambda^{(t)}_{\eta_i\theta_j} + Z_{\eta_i\tilde{\theta}_j}\lambda^{(t)}_{\eta_i\tilde{\theta}_j}} C_{\eta_i x_j} \mid \text{observed}\right] \qquad (3.38) \\
=& C_{\eta_i x_j}\left[\frac{\lambda^{(t)}_{\eta_i\theta_j}}{\lambda^{(t)}_{\eta_i\theta_j} + \lambda^{(t)}_{\eta_i\tilde{\theta}_j}} f\left(Z_{\eta_i\theta_j} = 1, Z_{\eta_i\tilde{\theta}_j} = 1 \mid \text{observed}\right) + \right. \\
& \left. f\left(Z_{\eta_i\theta_j} = 1, Z_{\eta_i\tilde{\theta}_j} = 0 \mid \text{observed}\right)\right],
\end{aligned}$$

where $\tilde{\theta}$ is the opposite allele of θ ($\tilde{\theta} = m$ if $\theta = p$; $\tilde{\theta} = p$ if $\theta = m$), $\lambda^{(t)}_{\eta_i\theta_j} = \beta\left(d^{(t)}_{\eta_i\theta_j}\right)^{\alpha}$, and $f\left(Z_{\eta_i\theta_j}, Z_{\eta_i\tilde{\theta}_j} \mid \text{observed}\right) = \sum_{Z_{\tilde{\eta}_i\theta_j}, Z_{\tilde{\eta}_i\tilde{\theta}_j}} f\left(Z_{\eta_i\theta_j}, Z_{\eta_i\tilde{\theta}_j}, Z_{\tilde{\eta}_i\theta_j}, Z_{\tilde{\eta}_i\tilde{\theta}_j} \mid \text{observed}\right)$.

Similarly, we can show that:

$$\begin{aligned}
\mathbb{ZC}_{\eta_i^*\theta_j} :=& \mathbb{E}\left(Z_{\eta_i\theta_j}C_{\eta_i^*\theta_j} \mid \text{observed}; \lambda^{(t)}_{\eta_i\theta_j}, \lambda^{(t)}_{\tilde{\eta}_i\theta_j}\right) \\
=& C_{x_i\theta_j}\left[\frac{\lambda^{(t)}_{\eta_i\theta_j}}{\lambda^{(t)}_{\eta_i\theta_j} + \lambda^{(t)}_{\tilde{\eta}_i\theta_j}} f\left(Z_{\eta_i\theta_j} = 1, Z_{\tilde{\eta}_i\theta_j} = 1 \mid \text{observed}\right) + \right. \qquad (3.39) \\
& \left. f\left(Z_{\eta_i\theta_j} = 1, Z_{\tilde{\eta}_i\theta_j} = 0 \mid \text{observed}\right)\right],
\end{aligned}$$

where $\tilde{\eta}$ is the opposite allele of η ($\tilde{\eta} = m$ if $\eta = p$; $\tilde{\eta} = p$ if $\eta = m$).

Furthermore, we have:

$$\begin{aligned}\mathbb{ZC}_{\eta_i^*\theta_j^*} :=& \mathbb{E}\left(Z_{\eta_i\theta_j}C_{\eta_i^*\theta_j^*} \mid \text{observed}; \lambda^{(t)}_{\eta_i\theta_j}, \lambda^{(t)}_{\eta_i\tilde{\theta}_j}, \lambda^{(t)}_{\tilde{\eta}_i\theta_j}, \lambda^{(t)}_{\tilde{\eta}_i\tilde{\theta}_j}\right)\\ =& C_{x_ix_j}\Bigg[\sum_{Z_{\eta_i\tilde{\theta}_j}, Z_{\tilde{\eta}_i\theta_j}, Z_{\tilde{\eta}_i\tilde{\theta}_j}} \frac{\lambda_{\eta_i\theta_j}}{\lambda_{\eta_i\theta_j} + Z_{\eta_i\tilde{\theta}_j}\lambda_{\eta_i\tilde{\theta}_j} + Z_{\tilde{\eta}_i\theta_j}\lambda_{\tilde{\eta}_i\theta_j} + Z_{\tilde{\eta}_i\tilde{\theta}_j}\lambda_{\tilde{\eta}_i\tilde{\theta}_j}}\cdot \\ & f\left(Z_{\eta_i\theta_j} = 1, Z_{\eta_i\tilde{\theta}_j}, Z_{\tilde{\eta}_i\theta_j}, Z_{\tilde{\eta}_i\tilde{\theta}_j} \mid \text{observed}\right)\Bigg].\end{aligned} \tag{3.40}$$

Consequently, we have

$$\begin{aligned} C_{\eta_ix_j} =& \mathbb{ZC}_{\eta_im_j^*} + \mathbb{ZC}_{\eta_ip_j^*},\\ C_{x_i\theta_j} =& \mathbb{ZC}_{m_i^*\theta_j} + \mathbb{ZC}_{p_i^*\theta_j},\\ C_{x_ix_j} =& \mathbb{ZC}_{m_i^*m_j^*} + \mathbb{ZC}_{m_i^*p_j^*} + \mathbb{ZC}_{p_i^*m_j^*} + \mathbb{ZC}_{p_i^*p_j^*}.\end{aligned} \tag{3.41}$$

Furthermore, we have

$$C_{\eta_i\theta_j} = \mathbb{Z}_{\eta_i\theta_j}C_{\eta_i\theta_j}. \tag{3.42}$$

The above is true because of the following:

When $C_{\eta_i\theta_j} > 0$, $\mathbb{Z}_{\eta_i\theta_j} = \mathbb{E}(Z_{\eta_i\theta_j} \mid \text{observed}) = 1$. Therefore, $\mathbb{Z}_{\eta_i\theta_j}C_{\eta_i\theta_j} = C_{\eta_i\theta_j}$.

When $C_{\eta_i\theta_j} = 0$, LHS = RHS = 0.

Conditional expectation of the complete log-likelihood

Recall eq. (3.31):

$$\mathcal{Q} = \sum_{i<j}\left[\mathbb{E}_{\boldsymbol{Z}|\bullet}\left(\log\mathcal{L}^Z\right) + \mathbb{E}_{\boldsymbol{Z}|\bullet}\left(\log\mathcal{L}^O\right) + \mathbb{E}_{\boldsymbol{C^H},\boldsymbol{Z}|\bullet}\left(\log\mathcal{L}^H\right) + \mathbb{E}_{\boldsymbol{C^H}|\bullet}\left(\log\mathcal{L}^X\right)\right]. \tag{3.31}$$

First, we have

$$\log\mathcal{L}^Z = \sum_{\eta,\theta}\left[Z_{\eta_i\theta_j}\log\gamma_{\eta_i\theta_j} + (1 - Z_{\eta_i\theta_j})\log(1 - \gamma_{\eta_i\theta_j})\right]. \tag{3.43}$$

Using the conditional expectation $\mathbb{Z}_{\eta_i\theta_j}$, we can compute $\mathbb{E}_{\boldsymbol{Z}|\bullet}\left(\log\mathcal{L}^Z\right)$ as

$$\mathbb{E}_{\boldsymbol{Z}|\bullet}\left(\log\mathcal{L}^Z\right) = \sum_{\eta,\theta}\left[\mathbb{Z}_{\eta_i\theta_j}\log\gamma_{\eta_i\theta_j} + (1-\mathbb{Z}_{\eta_i\theta_j})\log(1-\gamma_{\eta_i\theta_j})\right]. \tag{3.44}$$

Second, we have

$$\log\mathcal{L}^O = \sum_{\eta,\theta} Z_{\eta_i\theta_j}\left(C_{\eta_i\theta_j}\log q_{ij} + C_{\eta_i\theta_j}\log\lambda_{\eta_i\theta_j} - q_{ij}\lambda_{\eta_i\theta_j}\right) + c, \tag{3.45}$$

$$\mathbb{E}_{\boldsymbol{Z}|\bullet}\left(\log\mathcal{L}^O\right) = \sum_{\eta,\theta}\mathbb{Z}_{\eta_i\theta_j}\left(C_{\eta_i\theta_j}\log q_{ij} + C_{\eta_i\theta_j}\log\lambda_{\eta_i\theta_j} - q_{ij}\lambda_{\eta_i\theta_j}\right) + c. \tag{3.46}$$

Third, we have

$$\begin{aligned}\log\mathcal{L}^H = \sum_{\eta,\theta} Z_{\eta_i\theta_j}\Big[&\Big(C_{\eta_i\theta_j^*}\log q_{i\bar{j}} + C_{\eta_i\theta_j^*}\log\lambda_{\eta_i\theta_j} - q_{i\bar{j}}\lambda_{\eta_i\theta_j}\Big) + \\ &\left(C_{\eta_i^*\theta_j}\log q_{\bar{i}j} + C_{\eta_i^*\theta_j}\log\lambda_{\eta_i\theta_j} - q_{\bar{i}j}\lambda_{\eta_i\theta_j}\right) + \\ &\Big(C_{\eta_i^*\theta_j^*}\log q_{\bar{i}\bar{j}} + C_{\eta_i^*\theta_j^*}\log\lambda_{\eta_i\theta_j} - q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}\Big)\Big] + c.\end{aligned} \tag{3.47}$$

From eq. (3.47) and eqs. (3.38)-(3.40), we can show that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{C^H},\boldsymbol{Z}|\bullet}\left(\log\mathcal{L}^H\right)\\ =&\sum_{\eta,\theta}\Big[\Big(\mathbb{ZC}_{\eta_i\theta_j^*}\log q_{i\bar{j}} + \mathbb{ZC}_{\eta_i\theta_j^*}\log\lambda_{\eta_i\theta_j} - \mathbb{Z}_{\eta_i\theta_j}q_{i\bar{j}}\lambda_{\eta_i\theta_j}\Big) + \\ &\left(\mathbb{ZC}_{\eta_i^*\theta_j}\log q_{\bar{i}j} + \mathbb{ZC}_{\eta_i^*\theta_j}\log\lambda_{\eta_i\theta_j} - \mathbb{Z}_{\eta_i\theta_j}q_{\bar{i}j}\lambda_{\eta_i\theta_j}\right) + \\ &\Big(\mathbb{ZC}_{\eta_i^*\theta_j^*}\log q_{\bar{i}\bar{j}} + \mathbb{ZC}_{\eta_i^*\theta_j^*}\log\lambda_{\eta_i\theta_j} - \mathbb{Z}_{\eta_i\theta_j}q_{\bar{i}\bar{j}}\lambda_{\eta_i\theta_j}\Big)\Big] + c.\end{aligned} \tag{3.48}$$

Lastly, from eq. (3.41), we have

$$\mathbb{E}_{\boldsymbol{C^H},\boldsymbol{Z}|\bullet}\left(\log\mathcal{L}^X\right) = 0. \tag{3.49}$$

3.2.2 M-Step

Assume that at the t-th iteration, the current estimations of parameters are $\boldsymbol{X_m}^{(t)}, \boldsymbol{X_p}^{(t)}, \boldsymbol{\gamma}^{(t)}$, and $\boldsymbol{q}^{(t)}$. In the M-step, we aim to find the optimal solutions of $\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{\gamma}$, and $\boldsymbol{q}$ that maximize the conditional expectation of the complete log-likelihood function:

$$\boldsymbol{X_m}^{(t+1)}, \boldsymbol{X_p}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}, \boldsymbol{q}^{(t+1)} = \underset{\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{\gamma}, \boldsymbol{q}}{\arg\max}\ \mathcal{Q},$$

where

$$\begin{aligned}\mathcal{Q} &= \mathcal{Q}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{\gamma}, \boldsymbol{q}; \boldsymbol{X_m}^{(t)}, \boldsymbol{X_p}^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{q}^{(t)}\right) \\ &= \mathbb{E}_{\boldsymbol{C^H}, \boldsymbol{Z} \mid \bullet}\left(\log \mathcal{L}\left(\boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{\gamma}, \boldsymbol{q} \mid \boldsymbol{C^O}, \boldsymbol{C^X}, \boldsymbol{C^H}, \boldsymbol{Z}\right)\right)\end{aligned}$$

From eq. (3.31), we can divide $\mathcal{Q}$ into three parts:

$$\mathcal{Q} = \mathcal{Q}(\boldsymbol{X_m}, \boldsymbol{X_p}) + \mathcal{Q}(\boldsymbol{q}) + \mathcal{Q}(\gamma) + c, \tag{3.50}$$

where $\mathcal{Q}(\boldsymbol{X_m}, \boldsymbol{X_p})$, $\mathcal{Q}(\boldsymbol{q})$, and $\mathcal{Q}(\boldsymbol{\gamma})$ can be expressed as follows.

$$\mathcal{Q}(\boldsymbol{X_m}, \boldsymbol{X_p}) = \sum_{i<j} \sum_{\eta,\theta} \left[(\mathbb{Z}_{\eta_i\theta_j} C_{\eta_i\theta_j} + \mathbb{ZC}_{\eta_i\theta_j^*} + \mathbb{ZC}_{\eta_i^*\theta_j} + \mathbb{ZC}_{\eta_i^*\theta_j^*}) \log \lambda_{\eta_i\theta_j} - \mathbb{Z}_{\eta_i\theta_j} \lambda_{\eta_i\theta_j} \right]. \tag{3.51}$$

Define $\mathbb{ZT}_{\eta_i\theta_j}$ as the estimated total allele-specific contact frequency between η_i and θ_j:

$$\begin{aligned}\mathbb{ZT}_{\eta_i\theta_j} :=& \mathbb{Z}_{\eta_i\theta_j} C_{\eta_i\theta_j} + \mathbb{ZC}_{\eta_i\theta_j^*} + \mathbb{ZC}_{\eta_i^*\theta_j} + \mathbb{ZC}_{\eta_i^*\theta_j^*} \\ =& C_{\eta_i\theta_j} + \mathbb{ZC}_{\eta_i\theta_j^*} + \mathbb{ZC}_{\eta_i^*\theta_j} + \mathbb{ZC}_{\eta_i^*\theta_j^*}.\end{aligned} \tag{3.52}$$

Thus, we have

$$\mathcal{Q}(\boldsymbol{X_m}, \boldsymbol{X_p}) = \sum_{i<j} \sum_{\eta,\theta} \left(\mathbb{ZT}_{\eta_i\theta_j} \log \lambda_{\eta_i\theta_j} - \mathbb{Z}_{\eta_i\theta_j} \lambda_{\eta_i\theta_j} \right), \tag{3.53}$$

and

$$\begin{aligned}\mathcal{Q}(\boldsymbol{q}) =& \sum_{i<j} \sum_{\eta,\theta} \Big[(\mathbb{Z}_{\eta_i\theta_j} C_{\eta_i\theta_j} + \mathbb{ZC}_{\eta_i\theta_j^*}) \log q_i + (\mathbb{Z}_{\eta_i\theta_j} C_{\eta_i\theta_j} + \mathbb{ZC}_{\eta_i^*\theta_j}) \log q_j \\ &+ (\mathbb{ZC}_{\eta_i^*\theta_j} + \mathbb{ZC}_{\eta_i^*\theta_j^*}) \log(1 - q_i) + (\mathbb{ZC}_{\eta_i\theta_j^*} + \mathbb{ZC}_{\eta_i^*\theta_j^*}) \log(1 - q_j) \Big],\end{aligned} \tag{3.54}$$

and

$$\mathcal{Q}(\boldsymbol{\gamma}) = \sum_{i<j} \sum_{\eta,\theta} \left[\mathbb{Z}_{\eta_i\theta_j} \log \gamma_{\eta_i\theta_j} + \left(1 - \mathbb{Z}_{\eta_i\theta_j}\right) \log(1 - \gamma_{\eta_i\theta_j}) \right]. \tag{3.55}$$

Therefore, we can update $\boldsymbol{X}, \boldsymbol{q}$, and $\boldsymbol{\gamma}$ separately.

Solve $\boldsymbol{q}$

To update $\boldsymbol{q}$, set

$$\frac{\partial \mathcal{Q}}{\partial q_k} = 0 \tag{3.56}$$

for all $k = 1, \cdots, n$.

We can obtain a closed-form solution of $\boldsymbol{q}$ as follows:

$$q_k^{(t+1)} = \frac{S_1(k)}{S_1(k) + S_2(k)}, \tag{3.57}$$

where

$$\begin{aligned} S_1(k) = &\sum_{\eta,\theta} \left[\sum_{i<k} \left(\mathbb{Z}_{\eta_i\theta_k} C_{\eta_i\theta_k} + \mathbb{ZC}_{\eta_i^*\theta_k} \right) + \sum_{j>k} \left(\mathbb{Z}_{\eta_k\theta_j} C_{\eta_k\theta_j} + \mathbb{ZC}_{\eta_k\theta_j^*} \right) \right], \\ S_2(k) = &\sum_{\eta,\theta} \left[\sum_{i<k} \left(\mathbb{ZC}_{\eta_i\theta_k^*} + \mathbb{ZC}_{\eta_i^*\theta_k^*} \right) + \sum_{j>k} \left(\mathbb{ZC}_{\eta_k^*\theta_j} + \mathbb{ZC}_{\eta_k^*\theta_j^*} \right) \right]. \end{aligned} \tag{3.58}$$

From eq. (3.41), $S_1(k)$ and $S_2(k)$ can be simplified as follows: note that the resulting equations (eq. (3.59)) are exactly the same as eq. (3.16) in the ASHIC-PM case.

$$\begin{aligned} S_1(k) = &\sum_{i<k} \left(\sum_{\eta,\theta} C_{\eta_i\theta_k} + \sum_{\theta} C_{x_i\theta_k} \right) + \sum_{j>k} \left(\sum_{\eta,\theta} C_{\eta_k\theta_j} + \sum_{\eta} C_{\eta_k x_j} \right), \\ S_2(k) = &\sum_{i<k} \left(\sum_{\eta} C_{\eta_i x_k} + C_{x_i x_k} \right) + \sum_{j>k} \left(\sum_{\theta} C_{x_k\theta_j} + C_{x_k x_j} \right). \end{aligned} \tag{3.59}$$

Because $S_1(k)$ and $S_2(k)$ contain only observed values, we only need to estimate $\boldsymbol{q}$ once at the beginning.

Solve γ

From eq. (2.7), we assume that intra-chromosomal contact counts ($T_{\eta_i\theta_j}$, where $\eta = \theta$) with the same genomic distance $k = |i - j|$ share the same hyper parameter γ_k.

On the other hand, inter-chromosomal contact counts (i.e. $\eta \neq \theta$) share the same hyper parameter γ_{inter}.

To update γ_k, set

$$\frac{\partial \mathcal{Q}(\boldsymbol{\gamma})}{\partial \gamma_k} = 0 \tag{3.60}$$

for all $k = 1, \cdots, n-1$.

Similarly for γ_{inter}, we set

$$\frac{\partial \mathcal{Q}(\boldsymbol{\gamma})}{\partial \gamma_{\text{inter}}} = 0. \tag{3.61}$$

We can obtain a closed-form solution of γ_k and γ_{inter} as follows:

$$\gamma_k^{(t+1)} = \frac{\sum_{j-i=k} \sum_{\eta=\theta} \mathbb{Z}_{\eta_i \theta_j}}{\sum_{j-i=k} \sum_{\eta=\theta} 1}, \tag{3.62}$$

$$\gamma_{\text{inter}}^{(t+1)} = \frac{\sum_{i<j} \sum_{\eta \neq \theta} \mathbb{Z}_{\eta_i \theta_j}}{\sum_{i<j} \sum_{\eta \neq \theta} 1}. \tag{3.63}$$

Solve $\boldsymbol{X}$

There is no closed-form solution for the 3D structures $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$. Therefore, we use the L-BFGS-B algorithm [25] to find the optimal solution. To update $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$, we take the gradients of $\mathcal{Q}(\boldsymbol{X_m}, \boldsymbol{X_p})$ with respect to $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$ and set them all to 0.

$$\frac{\partial \mathcal{Q}}{\partial \boldsymbol{X_m}} = \begin{bmatrix} \frac{\partial \mathcal{Q}}{\partial x_{m_{11}}} & \cdots & \frac{\partial \mathcal{Q}}{\partial x_{m_{n1}}} \\ \frac{\partial \mathcal{Q}}{\partial x_{m_{12}}} & \cdots & \frac{\partial \mathcal{Q}}{\partial x_{m_{n2}}} \\ \frac{\partial \mathcal{Q}}{\partial x_{m_{13}}} & \cdots & \frac{\partial \mathcal{Q}}{\partial x_{m_{n3}}} \end{bmatrix} \tag{3.64}$$

Take the partial derivative with respect to $x_{m_{k1}}$ as an example:

$$\frac{\partial \mathcal{Q}(\boldsymbol{X_m}, \boldsymbol{X_p})}{\partial x_{m_{k1}}} = \sum_{i<j} \sum_{\eta,\theta} \alpha \left(\frac{\mathbb{Z}\mathbb{T}_{\eta_i \theta_j}}{d_{\eta_i \theta_j}} - \mathbb{Z}_{\eta_i \theta_j} \beta d_{\eta_i \theta_j}^{\alpha-1} \right) \frac{\partial d_{\eta_i \theta_j}}{\partial x_{m_{k1}}}, \tag{3.65}$$

where

$$\frac{\partial d_{\eta_i\theta_j}}{\partial x_{m_{k1}}} = \begin{cases} \frac{x_{\eta_{i1}} - x_{\theta_{j1}}}{d_{\eta_i\theta_j}}, & \text{if } \eta = m, i = k; \\ \frac{x_{\theta_{j1}} - x_{\eta_{i1}}}{d_{\eta_i\theta_j}}, & \text{if } \theta = m, j = k; \\ 0, & \text{otherwise.} \end{cases} \tag{3.66}$$

Therefore, we have

$$\begin{aligned} \frac{\partial \mathcal{Q}(\boldsymbol{X_m}, \boldsymbol{X_p})}{\partial x_{m_{k1}}} = & \sum_{i<k} \sum_{\eta} \alpha \left(\frac{\mathbb{Z}\mathbb{T}_{\eta_i m_k}}{d_{\eta_i m_k}} - \mathbb{Z}_{\eta_i m_k} \beta d_{\eta_i m_k}^{\alpha-1} \right) \frac{x_{m_{k1}} - x_{\eta_{i1}}}{d_{\eta_i m_k}} + \\ & \sum_{j>k} \sum_{\theta} \alpha \left(\frac{\mathbb{Z}\mathbb{T}_{m_k\theta_j}}{d_{m_k\theta_j}} - \mathbb{Z}_{m_k\theta_j} \beta d_{m_k\theta_j}^{\alpha-1} \right) \frac{x_{m_{k1}} - x_{\theta_{j1}}}{d_{m_k\theta_j}}. \end{aligned} \tag{3.67}$$

Similar to the process in the ASHIC-PM case, we update $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$ using the numerical optimizer `fmin_l_bfgs_b` from `SciPy`. We initialize each iteration with the current estimates of $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$, and we use the derivative to determine the direction of steepest descent and step size in the line search.

The 3D coordinates of each initial structure were randomly sampled from a unit cube. We first applied the multidimensional scaling (MDS) method [26, 19] to obtain a draft structure, and then used the draft structure as the starting point for the ASHIC models.

```
Input: observed allele-certain contact counts $\boldsymbol{C^O}$ and allele-ambiguous
       contact counts $\boldsymbol{C^X}$.
Output: estimated total allele-specific contact matrix $\mathbb{ZT}$, Poisson state
        latent variables $\mathbb{Z}$, allelic structures $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$, allele-identifiable
        probabilities $\boldsymbol{q}$, and Bernoulli priors $\boldsymbol{\gamma}$.
Compute the allele-identifiable probabilities $\boldsymbol{q}$ with eq. (3.59);
Initialize the Bernoulli priors $\boldsymbol{\gamma}$;
Initialize the allelic structures $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$;
while not converge do
    ⊡ E-step
    for $\eta, \theta \in \{m, p\}$, $1 \leq i < j \leq n$ do
        Update the posteriors $\mathbb{Z}_{\eta_i\theta_j}, \mathbb{ZC}_{\eta_i\theta_j^*}$, $\mathbb{ZC}_{\eta_i^*\theta_j}$, and $\mathbb{ZC}_{\eta_i^*\theta_j^*}$ with
          eqs. (3.32),(3.38)-(3.40);
        Update the total alelle-specific contacts $\mathbb{ZT}_{\eta_i\theta_j}$ with eq. (3.52);
    end
    ⊡ M-step
    Update the Bernoulli priors $\boldsymbol{\gamma}$ with eq. (3.62) and (3.63) using current $\mathbb{Z}$;
    Update allelic structures $\boldsymbol{X_m}$, $\boldsymbol{X_p}$ by eq. (3.67) with current $\mathbb{ZT}$ and $\mathbb{Z}$;
end
return $\mathbb{ZT}, \mathbb{Z}, \boldsymbol{X_m}, \boldsymbol{X_p}, \boldsymbol{q}, \boldsymbol{\gamma}$
```

Algorithm 2: EM algorithm for the ASHIC-ZIPM model

3.3 Bias-Incorporated Variant

In Section 2.6.1, we discussed the bias-incorporated variant of ASHIC models. Inference on the bias-incorporated models via EM algorithm is similar with a few modifications on the E-step and M-step.

In the ASHIC-PM model, we can update the expectation of the marginal log-likelihood (eq. (3.15)) as follows:

$$\begin{aligned}\mathcal{Q} =& \sum_{i<j}\sum_{\eta,\theta}\Big[\left(C_{\eta_i\theta_j} + \mathbb{C}_{\eta_i\theta_j^*}\right)\log q_i + \left(C_{\eta_i\theta_j} + \mathbb{C}_{\eta_i^*\theta_j}\right)\log q_j \\ &+ \left(\mathbb{C}_{\eta_i^*\theta_j} + \mathbb{C}_{\eta_i^*\theta_j^*}\right)\log\left(1-q_i\right) + \left(\mathbb{C}_{\eta_i\theta_j^*} + \mathbb{C}_{\eta_i^*\theta_j^*}\right)\log\left(1-q_j\right) \\ &+ \left(C_{\eta_i\theta_j} + \mathbb{C}_{\eta_i\theta_j^*} + \mathbb{C}_{\eta_i^*\theta_j} + \mathbb{C}_{\eta_i^*\theta_j^*}\right)\left(\log\lambda_{\eta_i\theta_j} + \log b_{\eta_i} + \log b_{\theta_j}\right) - b_{\eta_i}b_{\theta_j}\lambda_{\eta_i\theta_j}\Big] + c.\end{aligned} \tag{3.68}$$

In the E-step, we need to update eq. (3.11) as

$$\begin{aligned}\mathbb{C}_{\eta_i\theta_j^*} &= \frac{b_{\eta_i}b_{\theta_j}\lambda_{\eta_i\theta_j}^{(t)}}{\sum_{\theta'} b_{\eta_i}b_{\theta_j'}\lambda_{\eta_i\theta_j'}^{(t)}}C_{\eta_i x_j}, \\ \mathbb{C}_{\eta_i^*\theta_j} &= \frac{b_{\eta_i}b_{\theta_j}\lambda_{\eta_i\theta_j}^{(t)}}{\sum_{\eta'} b_{\eta_i'}b_{\theta_j}\lambda_{\eta_i'\theta_j}^{(t)}}C_{x_i\theta_j}, \\ \mathbb{C}_{\eta_i^*\theta_j^*} &= \frac{b_{\eta_i}b_{\theta_j}\lambda_{\eta_i\theta_j}^{(t)}}{\sum_{\eta',\theta'} b_{\eta_i'}b_{\theta_j'}\lambda_{\eta_i'\theta_j'}^{(t)}}C_{x_i x_j}.\end{aligned} \tag{3.69}$$

In the M-step, we need to plug the bias factors into eq. (3.22) as

$$\frac{\partial\mathcal{Q}}{\partial x_{m_{k1}}} = \sum_{j\neq k}\sum_{\theta}\alpha\left(\frac{\mathbb{T}_{m_k\theta_j}}{d_{m_k\theta_j}} - b_{m_k}b_{\theta_j}\beta d_{m_k\theta_j}^{\alpha-1}\right)\frac{x_{m_{k1}} - x_{\theta_{j1}}}{d_{m_k\theta_j}}. \tag{3.70}$$

Consequently, we can apply similar modifications to the log-likelihood function, E-step, and M-step in the ASHIC-ZIPM model.

We can initialize the true allele-specific contact frequencies by treating observed allele-certain contact frequencies $\boldsymbol{C^O}$ as Poisson parameters, then initialize the bias factors

b_{η_i} for $1 \leq i \leq n, \eta \in \{m, p\}$ using ICE [23]. Draft allele-specific contact matrices are then obtained through EM iterations. The final bias factors are estimated on the draft matrices, following which we refine the estimation of allele-specific contact matrices and structures through another round of EM iteration with the bias-incorporated model, as described above.

3.4 Inter-Homologous Optimization

Homologous chromosomes are often organized into separate chromosome territories [16, 17], which leads to an excessive number of zeros in inter-chromosomal contacts. These zeros are indeed "true" zeros caused by the long distance of inter-chromosomal contacts. However, such excessive sparsity could result in the underestimation of Poisson state probability γ_{inter}; in such a case, fewer "true" zeros will be included in the Poisson likelihood (a lower constraint in the inter-chromosomal distance). After the iterations of the EM algorithm, inter-chromosomal distances between homologous pair become progressively shorter, and in the E-step, more ambiguous contacts will be assigned as inter-chromosomal, leading to the inaccurate estimation of intra-chromosomal contacts.

To overcome the excessive sparsity problem in inter-chromosomal contacts, we make the following modifications to our EM algorithm. Recall that in the M-step, we optimize $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$ jointly to maximize $\mathcal{Q}(\boldsymbol{X_m}, \boldsymbol{X_p})$. Alternatively, we can divide the likelihood function $\mathcal{Q}(\boldsymbol{X_m}, \boldsymbol{X_p})$ into three parts: two intra-chromosomal parts $\mathcal{Q}_{\text{intra}}(\boldsymbol{X_m})$ and $\mathcal{Q}_{\text{intra}}(\boldsymbol{X_p})$, as well as one inter-chromosomal part $\mathcal{Q}_{\text{inter}}(\boldsymbol{X_m}, \boldsymbol{X_p})$. Accordingly, we can update the structures with respect to the three parts separately.

$$
\begin{aligned}
&\mathcal{Q}(\boldsymbol{X_m}, \boldsymbol{X_p}) \\
=&\sum_{i<j}\sum_{\eta,\theta}\left(\mathbb{ZT}_{\eta_i\theta_j}\log\lambda_{\eta_i\theta_j} - \mathbb{Z}_{\eta_i\theta_j}\lambda_{\eta_i\theta_j}\right) \\
=&\underbrace{\sum_{i<j}\sum_{\eta=\theta=m}\left(\mathbb{ZT}_{\eta_i\theta_j}\log\lambda_{\eta_i\theta_j} - \mathbb{Z}_{\eta_i\theta_j}\lambda_{\eta_i\theta_j}\right)}_{\mathcal{Q}_{\text{intra}}(\boldsymbol{X_m})} + \underbrace{\sum_{i<j}\sum_{\eta=\theta=p}\left(\mathbb{ZT}_{\eta_i\theta_j}\log\lambda_{\eta_i\theta_j} - \mathbb{Z}_{\eta_i\theta_j}\lambda_{\eta_i\theta_j}\right)}_{\mathcal{Q}_{\text{intra}}(\boldsymbol{X_p})} + \\
&\underbrace{\sum_{i<j}\sum_{\eta\neq\theta}\left(\mathbb{ZT}_{\eta_i\theta_j}\log\lambda_{\eta_i\theta_j} - \mathbb{Z}_{\eta_i\theta_j}\lambda_{\eta_i\theta_j}\right)}_{\mathcal{Q}_{\text{inter}}(\boldsymbol{X_m},\boldsymbol{X_p})}.
\end{aligned}
\tag{3.71}
$$

First, during the intra-chromosomal optimization, we consider only the intra-chromosomal contacts and update the two homologous structures separately to maximize their intra-chromosomal likelihood functions, $\mathcal{Q}_{\text{intra}}(\boldsymbol{X_m})$ and $\mathcal{Q}_{\text{intra}}(\boldsymbol{X_p})$:

$$
\begin{aligned}
\boldsymbol{X_m}^{(t+1)} &= \underset{\boldsymbol{X_m}}{\arg\max}\, \mathcal{Q}_{\text{intra}}(\boldsymbol{X_m}), \\
\boldsymbol{X_p}^{(t+1)} &= \underset{\boldsymbol{X_p}}{\arg\max}\, \mathcal{Q}_{\text{intra}}(\boldsymbol{X_p}).
\end{aligned}
\tag{3.72}
$$

To update the maternal structure $\boldsymbol{X_m}$, we calculate the gradient of the $\mathcal{Q}_{\text{intra}}(\boldsymbol{X_m})$ part in eq. (3.71) with respect to $\boldsymbol{X_m}$, following which we update $\boldsymbol{X_m}$ iteratively using the L-BFGS-B algorithm [25] provided in the `SciPy` package [27]. Specifically, we calculate

$$
\frac{\partial \mathcal{Q}_{\text{intra}}(\boldsymbol{X_m})}{\partial x_{m_{k1}}} = \sum_{i<j}\alpha\left(\frac{\mathbb{ZT}_{m_im_j}}{d_{m_im_j}} - \mathbb{Z}_{m_im_j}\beta d_{m_im_j}^{\alpha-1}\right)\frac{\partial d_{m_im_j}}{\partial x_{m_{k1}}}. \tag{3.73}
$$

The paternal structure $\boldsymbol{X_p}$ can be updated in a similar manner.

After the individual homologous chromosomal structures are optimized, we update the relative position of these two structures to maximize the inter-chromosomal likelihood function. Instead of updating the coordinates of two structures directly with respect to the inter-chromosomal contacts, we solve a simplified problem of finding the optimal rotation matrix $\boldsymbol{R} \in R^{3\times 3}$ and translation vector $\boldsymbol{v} \in R^3$ between the two structures $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$:

$$\boldsymbol{R}, \boldsymbol{v} = \underset{\boldsymbol{R},\boldsymbol{v}}{\arg\max}\, \mathcal{Q}_{\text{inter}}(\boldsymbol{X_m}, \boldsymbol{X_p}). \tag{3.74}$$

Let $\boldsymbol{x_{m_i}}$ and $\boldsymbol{x_{p_i}}$ be the coordinates of the i-th bin on the maternal and paternal chromosome, respectively. After applying rotation and translation transformations to the paternal structure, the inter-chromosomal distance between the i-th bin on the maternal chromosome and the j-th bin on the paternal chromosome becomes

$$d_{m_i p_j} = \left\| \boldsymbol{x_{m_i}} - \boldsymbol{R}(\boldsymbol{x_{p_j}} - \boldsymbol{v}) \right\|_2,$$

while the intra-chromosomal distances remain invariant.

We find the optimal solution of $\boldsymbol{R}$ and $\boldsymbol{v}$ as described below. Specifically, we first estimate the length of $\boldsymbol{v}$ (i.e., $\|\boldsymbol{v}\|_2$, the distance between two homologs) through MLE on a simplified inter-chromosomal likelihood function:

$$\begin{aligned}
\mathcal{Q}_{\text{inter}}(\boldsymbol{X_m}, \boldsymbol{X_p}) &= \sum_{i<j} \sum_{\eta \neq \theta} \left(\mathbb{ZT}_{\eta_i\theta_j} \log \lambda_{\eta_i\theta_j} - \mathbb{Z}_{\eta_i\theta_j} \lambda_{\eta_i\theta_j} \right) \\
&= \sum_{i,j} \left[\mathbb{ZT}_{m_i p_j} \alpha \log d_{m_i p_j} - \mathbb{Z}_{m_i p_j} \beta \left(d_{m_i p_j} \right)^{\alpha} \right] + c \qquad (3.75) \\
&= \sum_{i,j} \left[\mathbb{ZT}_{m_i p_j} \alpha \log d_{\text{inter}} - \mathbb{Z}_{m_i p_j} \beta d_{\text{inter}}^{\alpha} \right] + c.
\end{aligned}$$

The above simplification is performed by assuming that all inter-chromosomal bin pairs share the same inter-chromosomal distance d_{inter} instead of $d_{m_i p_j}$. Then, the length of $\boldsymbol{v}$ is set to be the maximum likelihood estimate of d_{inter}:

$$\|\boldsymbol{v}\|_2 = \hat{d}_{\text{inter}}^{\text{MLE}} = \left(\frac{\sum_{i,j} \mathbb{ZT}_{m_i p_j}}{\sum_{i,j} \mathbb{Z}_{m_i p_j} \beta} \right)^{1/\alpha}. \tag{3.76}$$

We then find the optimized $\boldsymbol{R}$ and $\boldsymbol{v}$ with the constraint on $\|\boldsymbol{v}\|_2$ to maximize the inter-chromosomal likelihood function $\mathcal{Q}_{\text{inter}}(\boldsymbol{X_m}, \boldsymbol{X_p})$ via the L-BFGS-B algorithm [25].

Another benefit of splitting the intra-chromosomal and inter-chromosomal likelihood is that it expedites the structure optimization process. The original optimization has a problem size of $2n \times 3$ (n loci in three dimensions for each homologous chromosome). We divide it into two (intra-chromosomal) subproblems, each with a size of $n \times 3$, and a trivial (inter-chromosomal) optimization with 6 unknowns (three Euler angles and a translation vector in three dimensions). Furthermore, we can run the two intra-chromosomal subroutines parallelly to increase the optimization speed further.

Chapter 4

Simulation Analyses

4.1 Simulation Settings

4.1.1 Homologous X Chromosome Structures

First, We considered the human homologous X chromosomes as the ground truth and simulated diploid Hi-C datasets as described below. We assumed that the allele-specific chromatin contact frequencies follow the ASHIC-ZIPM model. The true model parameters α_m, α_p, β, $\boldsymbol{\gamma}$, and $\boldsymbol{q}$ were estimated from two published datasets on human GM12878 cells: the predicted X chromosome structures ($\boldsymbol{X_m}, \boldsymbol{X_p} \in R^{3\times n}$) from single-cell Hi-C data by Tan et al. [16], and the allele-specific contact matrices C_m and C_p from *in situ* bulk Hi-C data by Rao et al. [10], both at 100 kb resolution (Table 4.1).

At the default setting, we generated 10 simulated allele-specific Hi-C datasets with the scale factor $\beta = 100\%\hat{\beta}$ and the average allele-identifiable probability $\overline{q} = 0.5$. Subsequently, we kept other parameters fixed and generated 10 additional datasets for each of the decreased β values ($50\%\hat{\beta}$, $20\%\hat{\beta}$, and $10\%\hat{\beta}$), and another 10 datasets for each of the decreased $\bar{q}$ values (0.25, 0.1, and 0.05). In total, 70 diploid Hi-C datasets were generated in this simulation study. For each simulated dataset, we ran 10 random initializations and chose the result with the highest observed log-likelihood for performance evaluation and subsequent analyses.

Reference	Type of data	GEO accession	Notes
Rao et al. (2014)	in situ (bulk) Hi-C	GSE63525	bulk diploid Hi-C data in human lymphoblastoid GM12878 cells
Tan et al. (2018)	single-cell structures	GSE117876 GSM3271347	single-cell X chromosomes structure (GM12878, Cell 1) predicted from single-cell Hi-C datasets at 100-kb resolution
Bonora et al. (2018)	in situ (bulk) DNase Hi-C	GSE107282 GSM2863686	bulk diploid Hi-C data on wild-type (WT) patski (BL6×Spretus) cells
Bonora et al. (2018)	CTCF ChIP-seq	GSE107282 GSM2863715	CTCF ChIP-seq data in WT patski (BL6×Spretus) cells
Tang et al. (2015)	CTCF ChIA-PET	GSE72816 GSM1872886	CTCF ChIA-PET data in human lymphoblastoid GM12878 cells
Tang et al. (2015)	RNA PolII ChIA-PET	GSE72816 GSM1872887	RNA PolII ChIA-PET data in human lymphoblastoid GM12878 cells

Table 4.1: Published datasets used in this study.

4.1.2 Identical Chromosome Structures

To study the effect of structural differences on the performance of our methods, we deployed a challenging simulation setting where we simulated diploid Hi-C datasets using two identical chromosome structures. Briefly, we duplicated the paternal (inactive) X chromosome structure $\boldsymbol{X_p}$ predicted by Tan et al. [16] as the pseudo-maternal structure. Then we used the two identical chromatin structures as the ground truth and simulated diploid Hi-C datasets in a similar manner as previously described.

The relative position of these two identical structures was determined by a reversed structural superposition procedure. Using the original homologous structures $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$, we calculated the optimal translation vector $\boldsymbol{v}$ and rotation matrix $\boldsymbol{R}$ using the Kabsch algorithm [28], such that the root-mean-square deviation (RMSD) between $\widetilde{\boldsymbol{X_m}} = \boldsymbol{R}(\boldsymbol{X_m} - \boldsymbol{v})$ and $\boldsymbol{X_p}$ was minimized. Then we duplicated $\boldsymbol{X_p}$ and reversed the superposition of $\boldsymbol{X_p}$ by $\widetilde{\boldsymbol{X_p}} = \boldsymbol{R}^{-1}\boldsymbol{X_p} + \boldsymbol{v}$. The resulting identical structures $\widetilde{\boldsymbol{X_p}}$ and $\boldsymbol{X_p}$ was served as the pseudo-homologous chromosome structures in which the relative position between $\widetilde{\boldsymbol{X_p}}$ and $\boldsymbol{X_p}$ remained approximately the same as the original homologous pair $\boldsymbol{X_m}$ and $\boldsymbol{X_p}$.

4.1.3 Simulation Parameter Estimation

We assumed that the maternal and paternal X chromosomes share the same Poisson state priors ($\boldsymbol{\gamma}$) and scale factor (β), but might possess different exponents (α_m and α_p) of the spatial distance decay effect. To estimate α_m and α_p, a two-step curve fitting procedure was applied to the maternal and paternal data separately. Consider the example of the maternal chromosome. First, we fit an exponential function for the relationship between spatial distance (d) and genomic distance (s) as $d \propto s^{B_{m1}}$. Specifically, the spatial distance between the bins i and j was calculated as their Euclidean distance on the maternal structure $\boldsymbol{X_m}$. For each genomic distance s_l from 1 to n, we calculated the average maternal spatial distance $\overline{d_l}$ among all bin pairs with genomic distance s_l. This resulted in n data points: $(s_1, \overline{d_1}), \ldots, (s_n, \overline{d_n})$. We applied the `curve_fit` function from `SciPy` package [27] in Python to fit the curve and estimate B_{m1}. Second, the relationship between the contact frequency (C_m) and the genomic distance (s) was fitted in a similar way as $C_m \propto s^{B_{m2}}$. From these two curve fitting steps, the relationship between the observed contact frequency and the underlying spatial distance could be deduced as $C_m \propto d^{\alpha_m}$, where $\alpha_m = \frac{B_{m2}}{B_{m1}}$. The paternal parameter counterpart α_p was estimated similarly. The empirically derived exponents are $\alpha_m = -3.02$ and $\alpha_p = -3.15$; both are close to the theoretical value of -3.

We estimated the scale factor β as described below. Unlike the ASHIC-PM model, the MLE of β does not have a closed-form solution under the ASHIC-ZIPM model. To simplify the estimation, we approximated each contact frequency $C_{m_i m_j}$ or $C_{p_i p_j}$ as an independent Poisson random variable with parameter in the form of $\beta(d_{m_i m_j})^{\alpha_m}$ or $\beta(d_{p_i p_j})^{\alpha_p}$, and calculated the β estimate as $\hat{\beta} = \frac{\sum_{i<j}\left(C_{m_i m_j} + C_{p_i p_j}\right)}{\sum_{i<j}\left((d_{m_i m_j})^{\alpha_m} + (d_{p_i p_j})^{\alpha_p}\right)}$. To evaluate the performance of our methods at lower sequencing coverage, we gradually decreased the value of β from $100\%\hat{\beta}$ to $50\%\hat{\beta}$, $20\%\hat{\beta}$, and $10\%\hat{\beta}$ to generate simulation datasets.

To estimate the Poisson state priors $\boldsymbol{\gamma}$, we performed the following spline fitting procedure. First, we assumed that the contact frequencies with the same genomic distance were i.i.d. ZIP random variables, then obtained the maximum likelihood estimate ($\hat{\gamma}$) at each genomic distance using `GenericLikelihoodMode` from `statsmodels` package [29] in

Python. Since the variability of $\hat{\gamma}$ increases with an increase in the genomic distance, we bin all $\hat{\gamma}$ into $K = 200$ genomic distance intervals with equal space in the log scale (using `logspace` function from `NumPy` [30]) to facilitate a smooth fitting. We calculated the average genomic distance ($\bar{s}_k$), and the average γ estimate ($\bar{\gamma}_k$) for each interval k, followed by fitting a cubic spline using the points $(\bar{s}_1, \bar{\gamma}_1), \ldots, (\bar{s}_K, \bar{\gamma}_K)$. Finally, we performed an anti-tonic regression on the spline to ensure that the generated $\boldsymbol{\gamma}$ estimates decrease monotonically.

The allele-identifiable probabilities $\boldsymbol{q}$ depend on the SNP density and vary among different cell systems. For example, in the mouse Patski cells, $\overline{q}$ is around 0.6, i.e., the chance that a 70-bp read overlaps with any SNP is 60% on average. Whereas in human GM12878 cells the $\overline{q}$ is around 0.04. To simulate the allele-identifiable probabilities, we first generated $\boldsymbol{q}$ randomly from a Beta distribution: $q_i \sim \texttt{Beta}(2, 2)$, where $\overline{q} = 0.5$ that mimics the mouse Patski cells.

To evaluate the performance of our methods at a lower SNP density, we gradually decreased the average q value from 0.5 to 0.25, 0.1, and 0.05 by using positively skewed Beta distributions. For instance, we used $q_i \sim \texttt{Beta}(2, 38)$ such that $\overline{q} = 0.05$ mimics the SNP density in GM12878 cells.

4.2 Convergence and Running Time

The convergence of the EM algorithm is defined as the relative increase of log-likelihood between two consecutive iterations is less than 10^{-4}. We tested our ASHIC software using a single core on an Intel E5-2683v4 processor with 8GB memory allocation. In a typical simulation setting, two X chromosomes were partitioned into 3000+ bins at

100 kb resolution. With the default sequencing coverage ($\beta = 100\%\hat{\beta}$) and SNP density ($\overline{q} = 0.5$) setting, both ASHIC-ZIPM and ASHIC-PM converged within 20 iterations (2 h). Lower coverage or lower SNP density requires more iterations. For example, when $\overline{q}$ reduced to 0.05, the EM algorithm of ASHIC-ZIPM took about 50 iterations (8 h) to converge. When β decreased to $10\%\hat{\beta}$, the EM algorithm of ASHIC-ZIPM underwent about 90 iterations (20 h) to converge.

4.3 Evaluation Metrics

Recovery rate (RR): We used recovery rate (RR) to measure the proportion of allele-specific contacts recovered by each method. For each method specifically, the recovery rate is defined as follows: Theoretically, both ASHIC-PM and ASHIC-ZIPM can recover 100% of allele-specific contacts. Recall $\mathbb{T}_{\eta_i\theta_j}$ and $\mathbb{ZT}_{\eta_i\theta_j}$ are the imputed estimates of $T_{\eta_i\theta_j}$ by ASHIC-PM and ASHIC-ZIP, respectively.

$$\begin{aligned}
\text{RR(ASHIC-PM)} =& \frac{\sum_{i<j}\sum_{\eta,\theta}\mathbb{T}_{\eta_i\theta_j}}{\sum_{i<j}\sum_{\eta,\theta}T_{\eta_i\theta_j}} \\
\text{RR(ASHIC-ZIPM)} =& \frac{\sum_{i<j}\sum_{\eta,\theta}\mathbb{ZT}_{\eta_i\theta_j}}{\sum_{i<j}\sum_{\eta,\theta}T_{\eta_i\theta_j}} \\
\text{RR(allele-certain)} =& \frac{\sum_{i<j}\sum_{\eta,\theta}C_{\eta_i\theta_j}}{\sum_{i<j}\sum_{\eta,\theta}T_{\eta_i\theta_j}} \\
\text{RR(mate-rescue)} =& \frac{\sum_{i<j}\left(\sum_{\eta,\theta}C_{\eta_i\theta_j} + \sum_{\eta}C_{\eta_i x_j} + \sum_{\theta}C_{x_i\theta_j}\right)}{\sum_{i<j}\sum_{\eta,\theta}T_{\eta_i\theta_j}}
\end{aligned} \tag{4.1}$$

Imputation error rate (IER): Within the recovered intra-chromosomal contacts, we further computed the imputation error rate (IER) for each method as defined below.

$$\begin{aligned}\text{IER(ASHIC-PM)} =& \frac{\sum_{i<j}\sum_{\eta}\left|\mathbb{T}_{\eta_i\eta_j} - T_{\eta_i\eta_j}\right|}{\sum_{i<j}\sum_{\eta} T_{\eta_i\eta_j}} \\ \text{IER(ASHIC-ZIPM)} =& \frac{\sum_{i<j}\sum_{\eta}\left|\mathbb{Z}\mathbb{T}_{\eta_i\eta_j} - T_{\eta_i\eta_j}\right|}{\sum_{i<j}\sum_{\eta} T_{\eta_i\eta_j}}\end{aligned} \tag{4.2}$$

Note that the allele-certain method uses only both-end allele-certain contacts and does not impute any allele-ambiguous contacts, resulting in a zero IER score.

Stratum adjusted correlation coefficient (SCC): To measure the similarity of the imputed contact matrix and the true contact matrix, we computed the stratum adjusted correlation coefficient (SCC) using `HiCRep` package [31]. The range of the genomic distance was set to be 0–5 Mb, and the smoothing window size was set as $h = 0$. The SCC values ranged from -1 to 1, where higher values indicate higher similarity between two Hi-C contact matrices. In the original `HiCRep` paper [31], the authors reported typical SCC values of pseudo-replicates, biological replicates and non-replicates. Pseudo-replicates have the highest SCC values from 0.96 to 0.98; biological replicates have a wider range of SCC values from 0.87 to 0.98; whereas non-replicates have the lowest SCC values that are typically smaller than 0.8.

Distance error rate (DER): Given the estimated chromosomal structures $\hat{\boldsymbol{X}}_{\boldsymbol{m}}, \hat{\boldsymbol{X}}_{\boldsymbol{p}}$ from our ASHIC-PM or ASHIC-ZIPM model, we measured their similarity with the ground truth allelic structures $\boldsymbol{X}_{\boldsymbol{m}}, \boldsymbol{X}_{\boldsymbol{p}}$ using the distance error rate (DER). Since the estimated structure and the true structure might be at different scales, we first re-scaled each structure by dividing the coordinates by the scale of the structure. The scale s of a 3D structure $\boldsymbol{X} \in R^{3\times n}$ is defined as the root mean square distance from the points to the structure centroid, i.e., $s = \sqrt{\frac{1}{n}\sum_i \|\boldsymbol{x_i} - \overline{\boldsymbol{x}}\|_2^2}$, where $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_i \boldsymbol{x_i}$ is the centroid. After re-scaling, each structure resulted with the scale equaling 1.

We then calculated the intra-chromosomal Euclidean distance $\widehat{d}_{\eta_i\eta_j}$ between loci i and j on the re-scaled estimated structure $\hat{\boldsymbol{X}}_{\boldsymbol{\eta}}$ ($\eta \in \{m, p\}$), and $d_{\eta_i\eta_j}$ between loci i and j on the re-scaled ground truth structure $\boldsymbol{X}_{\boldsymbol{\eta}}$. The distance error rate is defined as

$$\text{DER} = \frac{\sum_{i<j}\sum_{\eta}\left|\widehat{d}_{\eta_i\eta_j} - d_{\eta_i\eta_j}\right|}{\sum_{i<j}\sum_{\eta} d_{\eta_i\eta_j}} \tag{4.3}$$

Homologous distance error rate (HDER): In the simulation experiments of two identical homologous structures, we calculated the homologous distance error rate (HDER) between the two homologous structures. Since the two estimated homologous structures are generated with the same scaling factor β, there was no need for re-scaling. The homologous distance error rate is calculated as

$$\text{HDER} = \frac{\sum_{i<j}\left|\widehat{d}_{m_im_j} - \widehat{d}_{p_ip_j}\right|}{\sum_{i<j}\widehat{d}_{p_ip_j}} \tag{4.4}$$

Recall, Precision, and F_1 score of significant chromatin interactions: To further evaluate the imputed contact matrices, we called significant interactions on the imputed chromatin contact matrices using Fit-Hi-C [1] and compared with significant interactions called from the ground true matrices. We applied Fit-Hi-C to the allele-specific chromatin contacts between genomic distance of 300 kb and 5 Mb, and filtered significant interactions with q-value $< 10^{-6}$.

Significant interactions called on the imputed maternal and paternal contact matrices were denoted as $\widehat{\text{SI}}_m$ and $\widehat{\text{SI}}_p$, and the significant interactions called on the ground true maternal and paternal matrices were denoted as SI_m and SI_p, respectively.

For each method, we calculated the recall, precision, and F_1 score as follows:

$$\begin{aligned}\text{Recall} =& \frac{\left|\widehat{\text{SI}} \cap \text{SI}\right|}{|\text{SI}|} \\ \text{Precision} =& \frac{\left|\widehat{\text{SI}} \cap \text{SI}\right|}{\left|\widehat{\text{SI}}\right|} \\ F_1 =& \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}\end{aligned} \tag{4.5}$$

Note that when the ground true maternal and paternal structures are different, we define the true set to be the allele-specific interactions that are unique on the maternal or paternal allele. That is, $\text{SI} = \text{SI}_{m \cup p} - \text{SI}_{m \cap p}$, where $\text{SI}_{m \cup p} = \text{SI}_m \cup \text{SI}_p, \text{SI}_{m \cap p} = \text{SI}_m \cap \text{SI}_p$. Similarly, we have $\widehat{\text{SI}} = \widehat{\text{SI}}_{m \cup p} - \widehat{\text{SI}}_{m \cap p}$, where $\widehat{\text{SI}}_{m \cup p} = \widehat{\text{SI}}_m \cup \widehat{\text{SI}}_p, \widehat{\text{SI}}_{m \cap p} = \widehat{\text{SI}}_m \cap \widehat{\text{SI}}_p$

On the other hand, when the ground true maternal and paternal structures were identical, we defined the true set as the common interactions shared between the maternal and paternal interactions. That is, $\text{SI} = \text{SI}_{m \cap p} = \text{SI}_m \cap \text{SI}_p$. Similarly, we have $\widehat{\text{SI}} = \widehat{\text{SI}}_{m \cap p} = \widehat{\text{SI}}_m \cap \widehat{\text{SI}}_p$.

4.4 Human Homologous X Chromosome Structures

4.4.1 Default Simulation Setting

We first evaluated the performance of the proposed ASHIC methods on simulated diploid Hi-C datasets of the homologous X chromosomes in human GM12878 cells. Of the two X chromosomes, the active X chromosome (denoted as Xa) is the maternal copy and the inactive X chromosome (denoted as Xi) is the paternal copy. We considered the 3D structures of Xa and Xi published by Tan et al. [16] as the ground truth and generated 10 simulated diploid Hi-C datasets at 100-kb resolution. Each simulated dataset contained two intra-chromosomal contact matrices, one for Xa and one for Xi, as well as one inter-chromosomal contact matrix between Xa and Xi.

We compared our ASHIC-ZIPM and ASHIC-PM methods with two commonly used approaches for analyzing diploid Hi-C data. The first approach is the allele-certain method that uses only both-end allele-certain contacts [10, 14]. The second approach is the mate-rescue method that combines both-end allele-certain contacts with one-end allele-ambiguous contacts by assigning the allele-ambiguous read-end to the same allele as the allele-certain mate-end [13, 15, 18].

To evaluate the imputation of diploid Hi-C contact maps, we first calculated the proportion of allele-specific contacts recovered by each method (Tables 4.2,4.3). At the default sequencing coverage ($\beta = 100\%\hat{\beta}$) and SNP density ($\overline{q} = 0.5$) setting, the allele-certain and mate-rescue approaches recovered evidently smaller proportion of diploid chromatin contacts (25.65% and 75.55%, respectively) compared to the ASHIC-ZIPM and ASHIC-PM methods that were able to recover all one-end and both-end allele-ambiguous reads, thereby achieving 100% full recovery rate.

Average allele-identifiable	Both-end	One-end	Both-end
probability $\bar{q}$	allele-certain	allele-ambiguous	allele-ambiguous
0.05	0.25%	9.57%	90.18%
0.10	1.02%	18.14%	80.84%
0.25	6.50%	38.01%	55.49%
0.50	25.65%	49.89%	24.45%

Table 4.2: Proportion of allele-certain and allele-ambiguous contacts of simulated data.

Average allele-identifiable probability $\bar{q}$	ASHIC-ZIPM	ASHIC-PM	Allele-certain	Mate-rescue
0.05	100.00%	100.00%	0.25%	9.82%
0.10	100.00%	100.00%	1.02%	19.16%
0.25	100.00%	100.00%	6.50%	44.51%
0.50	100.00%	100.00%	25.65%	75.55%

Table 4.3: Recovery rate of diploid Hi-C methods of simulated data.

Next, we sought to assess the accuracy of the imputed allele-specific contact matrices. Recent studies have demonstrated that the genomic distance dependence and sequencing depth have confounding effects on measuring the similarity between Hi-C contact matrices [31]. To account for these confounding factors, we computed the stratum adjusted correlation coefficient (SCC) using the `HiCRep` package [31] to measure the similarity between the imputed contact matrices and true matrices (Figure 4.1A). We observed that the imputed diploid matrices obtained by ASHIC-ZIPM and ASHIC-PM had near-perfect SCC values of 0.9997 and 0.9996, respectively; whereas mate-rescue and allele-certain methods demonstrated lower SCC values of 0.9733 and 0.8100, respectively. ASHIC-ZIPM showed a significantly higher SCC values than ASHIC-PM (p-value $= 2.53 \times 10^{-3}$, one-sided paired Wilcoxon signed-rank test). In addition, ASHIC-ZIPM performed significantly better than the allele-certain and mate-rescue methods (p-values $= 2.53 \times 10^{-3}$, one-sided paired Wilcoxon signed-rank tests). Note that $p = 2.53 \times 10^{-3}$ is the smallest possible p-value given the sample size.

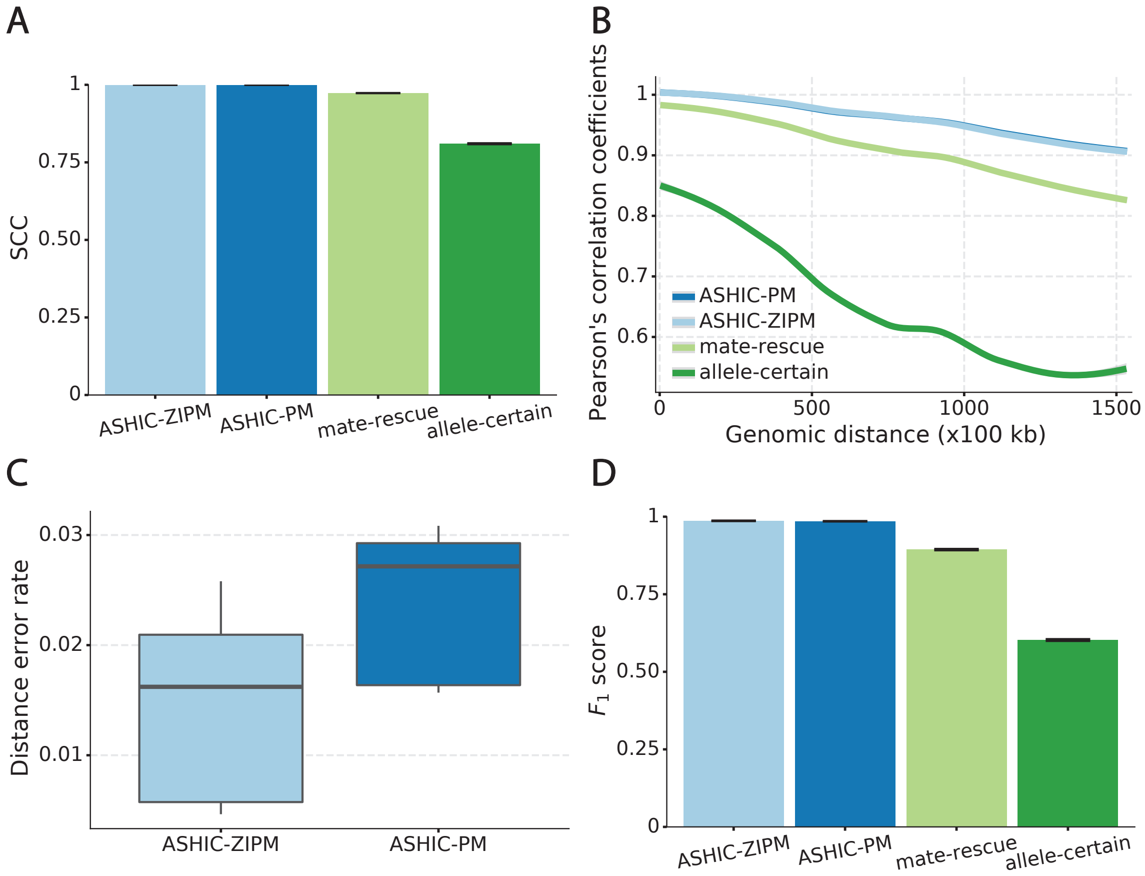

Figure 4.1: Evaluation on simulated homologous X chromosome (Xa/Xi) data. (A) Stratum-adjusted correlation coefficients (SCCs) and (B) and Pearson's correlation coefficients (PCCs) between the imputed diploid contact matrices and the true contact matrices. The PCC curves are smoothened using the locally weighted LOESS method. (C) Distance error rates between the predicted allelic 3D structures and the true structures. (D) F_1 scores of the identified allele-specific chromatin interactions.

The SCC statistic is a weighted average of the Pearson's correlation coefficients (PCCs) across different genomic distances [31]. To breakdown the effect of genomic distance, we computed the PCCs between the imputed contact matrices and the true matrices at different genomic distances (Figure 4.1B). As expected, the PCC values decreased as the genomic distance increased for all four methods. We observed that the ASHIC-ZIPM and ASHIC-PM methods demonstrated similar PCC values across all genomic distances.

In addition, the ASHIC-ZIPM and ASHIC-PM methods outperformed the allele-certain and mate-rescue approaches by large margin, especially at large genomic distances. Taken together, the SCC and PCC results showed that our ASHIC methods can accurately impute allele-specific contact matrices. Moreover, the imputation accuracy outperformed the allele-certain and mate-rescue approaches, especially for long-range contacts.

In addition to imputing diploid Hi-C contact matrices, the ASHIC-ZIPM and ASHIC-PM methods also predict allele-specific 3D structures. To evaluate the accuracy of the predicted allelic structures, we calculated the distance error rates between the predicted structures and the ground truth (Figure 4.1C). We observed that ASHIC-ZIPM yielded significantly lower distance error rates and thereby, more accurate allelic 3D structures than those obtained by ASHIC-PM (p-value $= 2.53 \times 10^{-3}$, one-sided paired Wilcoxon signed-rank test).

Furthermore, we investigated whether the imputed diploid contact matrices can facilitate the detection of allele-specific chromatin interactions. First, we called significant interactions using the Fit-Hi-C package [1] on the true diploid contact matrices. We subsequently defined the maternal-specific interactions as the interactions that were called only from the true maternal matrix but not from the paternal matrix. The paternal-specific interactions were defined accordingly. The final set of true allele-specific interactions was defined as the union of both monoallelic sets, which contained 9061.5 interactions on average (Table 4.4, $\beta = 100\%\hat{\beta}$). Following the same procedure, we then identified the allele-specific interactions from the imputed diploid contact matrices resulting from the four methods, separately. We evaluated the identified allele-specific interactions from each method using three metrics: precision, recall, and their harmonic mean F_1 score (Figure 4.1D, Figure 4.2A,B, $\beta = 100\%\hat{\beta}$). ASHIC-ZIPM and ASHIC-PM maintained the highest F_1 scores of 0.9867 and 0.9853, respectively. In addition, ASHIC-ZIPM significantly outperformed mate-rescue ($F_1 = 0.8940$) and allele-certain ($F_1 = 0.6024$) in terms of the F_1 scores (p-values $= 2.53 \times 10^{-3}$, one-sided paired Wilcoxon signed-rank tests). The low F_1 scores of the mate-rescue and allele-certain methods were primarily contributed by their low recall rates (Figure 4.2A,B, $\beta = 100\%\hat{\beta}$), which was a result of their low recovery rates of allele-ambiguous contacts (Table 4.3, $\bar{q} = 0.5$).

Sequencing coverage β	Bi-allelic	Maternal-specific	Paternal-specific	True set
10%	24.1	1040.1	1096.3	2136.4
20%	69.2	1998.9	2003.0	4001.9
50%	226.9	3464.4	3412.5	6876.9
100%	424.1	4700.9	4360.6	9061.5

Table 4.4: Number of allele-specific interactions in homologous X chromosome (Xa/Xi) simulations. Chromatin interactions are called using Fit-Hi-C [1] on true maternal (Xa) and paternal (Xi) contact matrices separately. Maternal-specific interaction set contains interactions called only from maternal contact matrix but not from paternal contact matrix. Paternal-specific interaction set is defined in a similar way. The bi-allelic interaction set contains common interactions called from both maternal and paternal contact matrices. The true set is defined as the union of maternal-specific interaction set and paternal-specific interaction set.

Collectively, our comparisons have demonstrated that the proposed ASHIC-ZIPM and ASHIC-PM methods outperformed the existing mate-rescue and allele-certain approaches with respect to the recovery rate of allele-ambiguous contacts, the accuracy of imputed diploid contact matrices and predicted allelic 3D structures, and the ability to

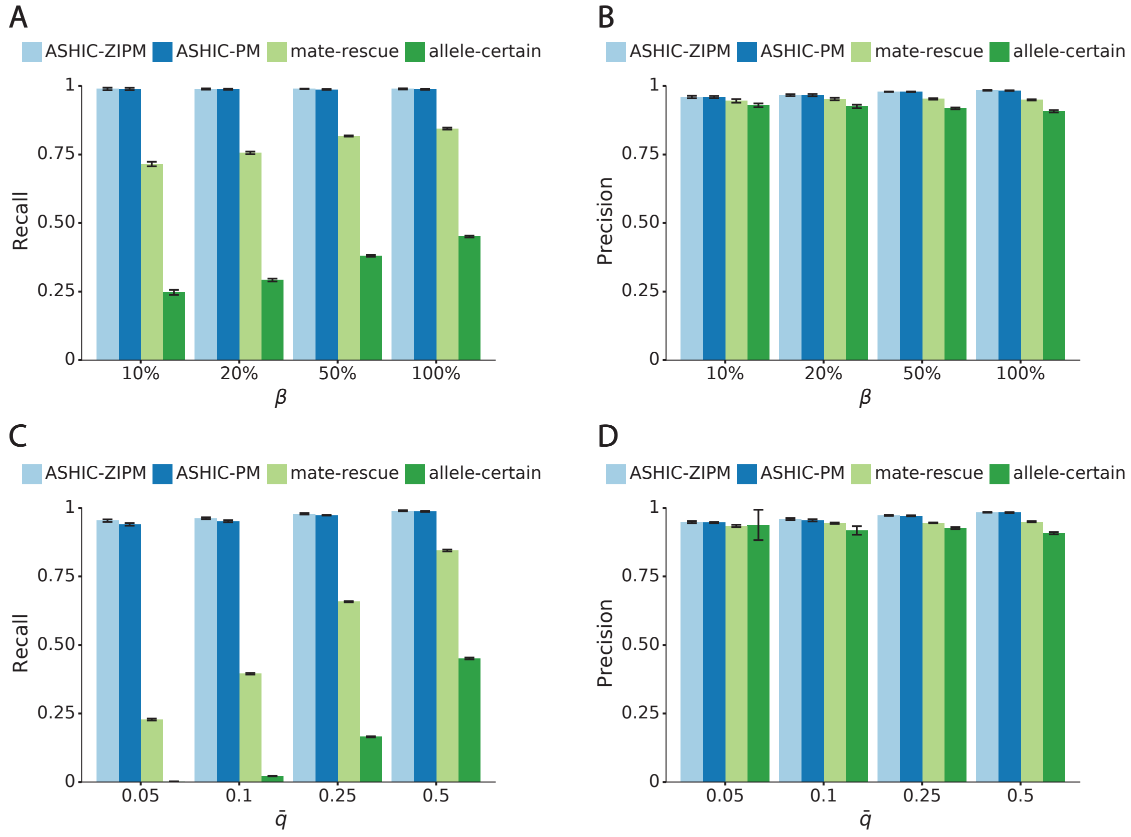

Figure 4.2: ASHIC-ZIPM-imputed diploid contact maps show high recall and precision of allele-specific chromatin interactions on the homologous X chromosome simulation data. (A) Recall and (B) Precision of the identified allele-specific interactions at different sequencing coverage levels. (C) Recall and (D) Precision of the identified allele-specific interactions at different SNP density levels.

facilitate the detection of allele-specific chromatin loops. In addition, ASHIC-ZIPM demonstrated a better performance overall than that of ASHIC-PM, especially in the prediction of allelic 3D structures. To further evaluate the performance of these methods under different circumstances, we conducted a series of additional simulation experiments by adjusting three major factors: sequencing coverage, SNP density, and homologous structural similarity.

4.4.2 Low Sequencing Coverage Data

The sequencing coverage of Hi-C contact matrices is a major factor that can affect the performance of the diploid Hi-C methods. An observed zero entry in the Hi-C contact matrix can be either a "true" zero as a result of no physical contact between the pair of chromatin fragments, or a "missing" zero as a result of insufficient sequencing coverage. Lower sequencing depth of Hi-C experiments yields lower-coverage and sparse contact matrices that containing excessive "missing" zeros. As a result, it becomes more challenging to distinguish the "true" zeros from the "missing" zeros.

While generating the simulation datasets, the scale factor β controls the coverage of simulated contact matrices. We estimated $\hat{\beta}$ from the published Hi-C data by Rao et al. [10] (Section 4.1.3). At the default $\beta = 100\%\hat{\beta}$ setting, the simulated Hi-C map contained about 4.9 million contacts from the homologous X chromosomes. To evaluate the performance of our methods on lower-coverage data, we fixed the SNP density $\overline{q} = 0.5$ and gradually decreased the value of β from $100\%\hat{\beta}$ to $50\%\hat{\beta}$, $20\%\hat{\beta}$, and $10\%\hat{\beta}$, resulting in 2.5 million, 1.0 million, and 0.5 million contacts, respectively. We then repeated the assessments of the ASHIC-ZIPM, ASHIC-PM, mate-rescue, and allele-certain methods with these low-coverage simulation datasets.

As shown in Figure 4.3A, ASHIC-ZIPM and ASHIC-PM maintained the highest SCC values across all coverage levels. When the sequencing coverage decreased from $100\%\hat{\beta}$ to $10\%\hat{\beta}$, the SCC values for both methods only dropped by 0.28%. On the other hand, when sequencing coverage lowered, the SCC values decreased evidently for mate-rescue and allele-certain by 1.80% and 10.38%, respectively. These results suggested that our

ASHIC methods can robustly and accurately infer allele-specific contact matrices under low sequencing coverage conditions.

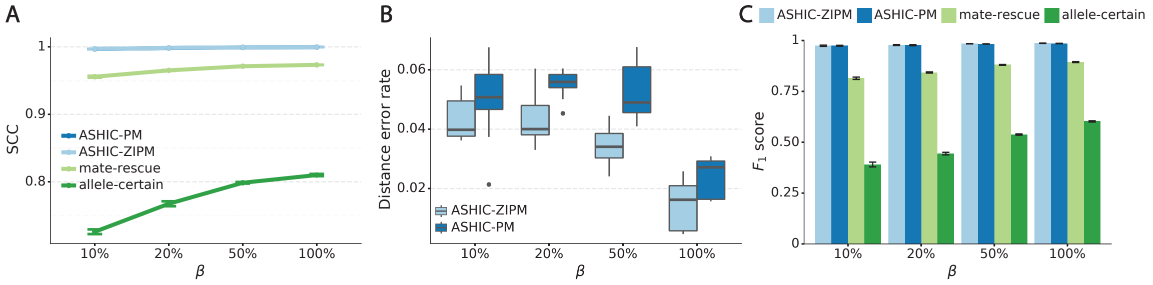

Figure 4.3: ASHIC-ZIPM accurately imputes diploid contact maps and 3D structures on low-coverage Xa/Xi simulation data. (A) SCCs between the imputed diploid contact matrices and the true contact matrices, (B) Distance error rates between the predicted allelic 3D structures and the true structures, and (C) F_1 scores of the identified allele-specific chromatin interactions at different sequencing coverage β levels.

Additionally, we observed that ASHIC-ZIPM produced more accurate 3D structures with smaller distance error rates than those produced by ASHIC-PM across all sequencing coverage levels (Figure 4.3B). The improvements of the distance error rates were significant at coverage levels $100\%\hat{\beta}$, $50\%\hat{\beta}$ and $20\%\hat{\beta}$ (p-values $= 2.53 \times 10^{-3}, 2.53 \times 10^{-3}, 6.26 \times 10^{-3}$, respectively, one-sided paired Wilcoxon signed-rank tests).

When the sequencing coverage decreased from $100\%\hat{\beta}$ to $10\%\hat{\beta}$, the true set of allele-specific interactions decreased from 9061.5 to 2136.4 interactions (Table 4.4). As shown in Figure 4.3C, when the coverage decreased from $100\%\hat{\beta}$ to $10\%\hat{\beta}$, the ability of the allele-certain method to detect allele-specific interactions was highly impacted as its F_1 scores dropped by 35.17% from 0.6024 to 0.3906. The decrease of F_1 score for mate-rescue was less severe, about 8.90% from 0.8940 to 0.8144. The ASHIC methods consistently delivered robust results against coverage changes (ASHIC-ZIPM: $\Delta F_1 = 1.26\%$, ASHIC-PM:

$\Delta F_1 = 1.14\%$), and maintained high F_1 score even at the lowest $10\%\hat{\beta}$ level (ASHIC-ZIPM: 0.9743, ASHIC-PM: 0.9740). The decay in F_1 scores for the allele-certain and mate-rescue methods was primarily contributed by their low recall rates (Figure 4.2A,B).

Taken together, our results demonstrated that the ASHIC methods significantly outperformed other methods in low sequencing coverage conditions, resulted in more accurately imputed matrices and benefited the detection of allele-specific interactions on low-coverage data.

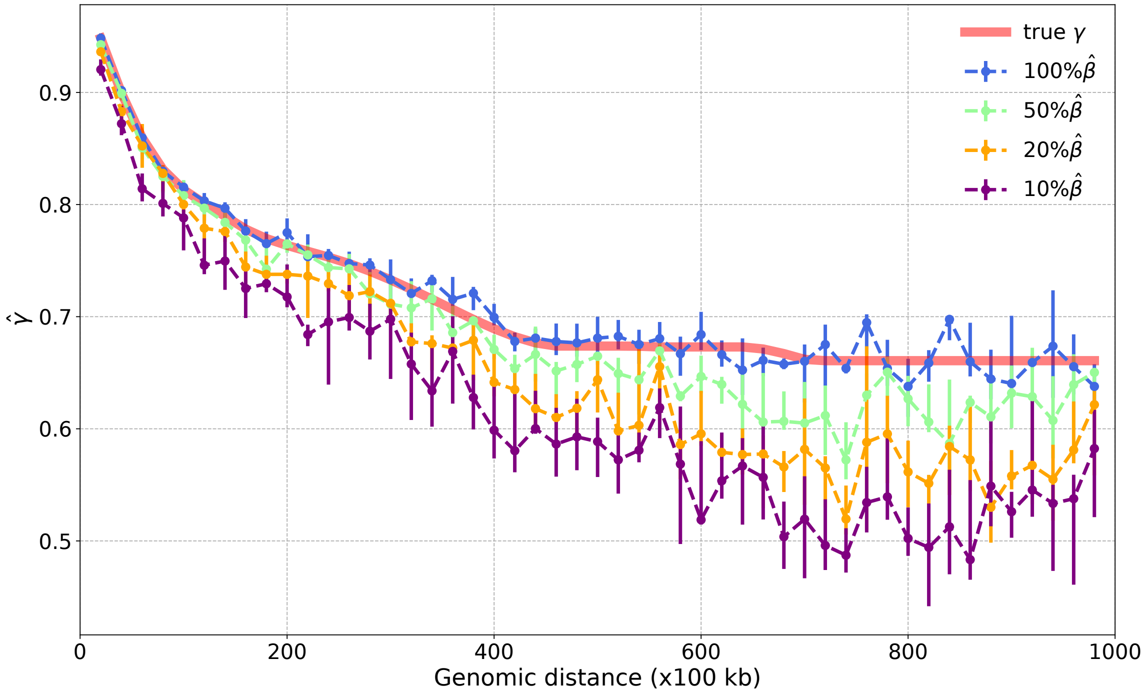

Figure 4.4: ASHIC-ZIPM adjusts estimated $\boldsymbol{\gamma}$ for low-coverage homologous X chromosome simulation data to account for additional missing zeros. The estimated $\boldsymbol{\gamma}$ of each genomic distance within 100 Mb are drawn for every 2 Mb interval. Each dot represents the median value of γ estimated from the 10 simulated datasets. The bottom and top error bars represent the first and third quartiles of estimated γ at that genomic distance.

In particular, we observed that ASHIC-ZIPM had better performance than ASHIC-PM under low coverage conditions. This is owing to the fact that in our ASHIC-ZIPM model, the Poisson state probabilities $\boldsymbol{\gamma}$ act as weights between the "true" and "missing" zeros. When the sequencing coverage lowered, the observed diploid matrices contained additional "missing" zeros. The zero-inflated model explicitly adjusted the estimation of $\boldsymbol{\gamma}$ to model these "missing" zeros, thereby achieving better model fitting results. Consistent with our expectations, the estimated values of $\boldsymbol{\gamma}$ became smaller as coverage decreased, which demonstrated its ability to account for the additional "missing" zeros (Figure 4.4).

4.4.3 Low SNP Density Data

In addition to the sequencing coverage, the SNP density is another major factor affecting the performance of the diploid Hi-C methods. The SNP density varies across different species and cell lines. For example, the F1 mouse cross (BL6×*Spretus*) has a relatively high SNP density of approximately 1 SNP per 75 bp. On average, a 70-bp sequence read has a 60% chance overlapping with SNP(s), thus being allele-identifiable. Whereas the GM12878 cell line has a low SNP density about 1 for every 1700 bp, which is corresponding to an average allele-identifiable probability of 0.04 (Table 1.1).

To evaluate the performance of our methods on low-SNP-density data, we fixed the coverage level at $100\%\hat{\beta}$ and then gradually decreased $\overline{q}$, the average allele-identifiable probability, from 0.5 which mimics the BL6×*Spretus* cross, to 0.25, 0.1, and 0.05, where the smallest value mimics the GM12878 cells.

When the SNP density was low, fewer both-end allele-certain contacts but higher number of one-end allele-ambiguous and both-end allele-ambiguous contacts were observed. Consequently, as the average allele-identifiable probability $\overline{q}$ decreased from 0.5 to 0.05, the recovery rates dropped dramatically from 25.65% to 0.25% for allele-certain and from

75.55% to 9.82% for mate-rescue (Table 4.3). In contrast, our ASHIC methods were able to recover all allele-ambiguous reads at the lowest $\overline{q} = 0.05$ setting. Among the recovered contacts, 15.95% for ASHIC-ZIPM and 17.60% for ASHIC-PM were incorrectly imputed (Table 4.5, $\beta = 100\%\hat{\beta}, \bar{q} = 0.05$).

Sequencing coverage β	Average allele-identifiable probability $\bar{q}$	ASHIC-ZIPM	ASHIC-PM
10%	0.50	7.88%	8.12%
20%	0.50	7.06%	7.42%
50%	0.50	5.88%	6.43%
100%	0.50	4.84%	5.48%
100%	0.25	9.55%	10.68%
100%	0.10	13.85%	15.50%
100%	0.05	15.95%	17.60%

Table 4.5: Imputation error rates in homologous X chromosome (Xa/Xi) simulations.

Consistent with the high recovery rates and low imputation error rates, the SCC values also demonstrated robust and accurate imputation of diploid contact matrices by the ASHIC methods at low SNP density settings (Figure 4.5A). When the average allele-identifiable probability $\overline{q}$ decreased from 0.5 to 0.05, the SCC values dropped significantly from 0.8100 to 0.3959 for allele-certain and from 0.9733 to 0.8719 for mate-rescue, respectively. In contrast, the SCC values remained high at 0.9941 and 0.9922 for ASHIC-ZIPM and ASHIC-PM, respectively, at the lowest $\overline{q} = 0.05$ setting. Moreover, ASHIC-ZIPM significantly outperformed ASHIC-PM at the lowest SNP density level (p-value $= 8.30 \times 10^{-3}$, one-sided paired Wilcoxon signed-rank test). The difference between our ASHIC methods

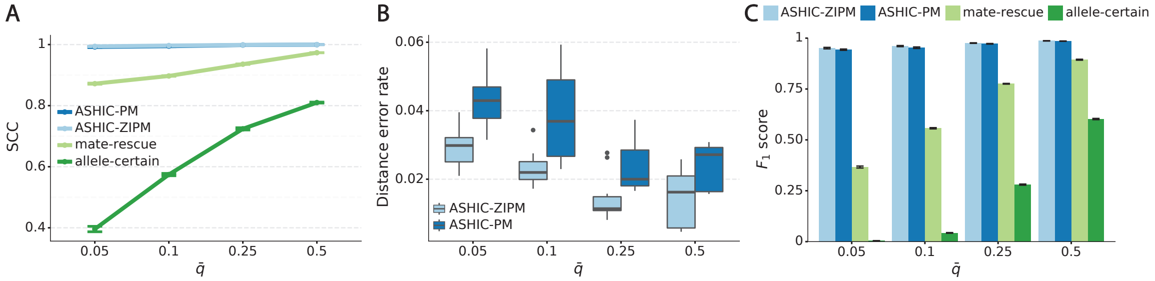

Figure 4.5: ASHIC-ZIPM accurately imputes diploid contact maps and 3D structures on low-SNP-density Xa/Xi simulation data. (A) SCCs between the imputed diploid contact matrices and the true contact matrices, (B) Distance error rates between the predicted allelic 3D structures and the true structures, and (C) F_1 scores of the identified allele-specific chromatin interactions at different SNP density $\overline{q}$ levels.

and other methods was also observed on the PCC plot at the lowest SNP density, particularly for long genomic distances (Figure 4.6A). Furthermore, when comparing the predicted allelic 3D structures with the ground truth, ASHIC-ZIPM outperformed ASHIC-PM significantly at all SNP density levels (p-values $= 2.53 \times 10^{-3}, 4.67 \times 10^{-3}, 3.46 \times 10^{-3}, 2.53 \times 10^{-3}$, for $\overline{q} = 0.5, 0.25, 0.1, 0.05$, respectively, one-sided paired Wilcoxon signed-rank tests) (Figure 4.5B).

Next, we questioned whether the ability to detect allele-specific chromatin interactions was impacted by low SNP density levels. Adjusting the average allele-identifiable probability did not affect the underlying true diploid contact matrices. As a result, the true set of allele-specific interactions remained the same at different SNP density settings (Table 4.4, $\beta = 100\%\hat{\beta}$). As shown in Figure 4.5C, low SNP density severely impacted the allele-certain and mate-rescue methods. The F_1 scores of allele-certain dropped from 0.6024 to 0.0039, recovering only 17.8 out of 9061.5 true allele-specific interactions. Similarly, the F_1 score of mate-rescue dropped from 0.8940 to 0.3666. In contrast, when SNP density

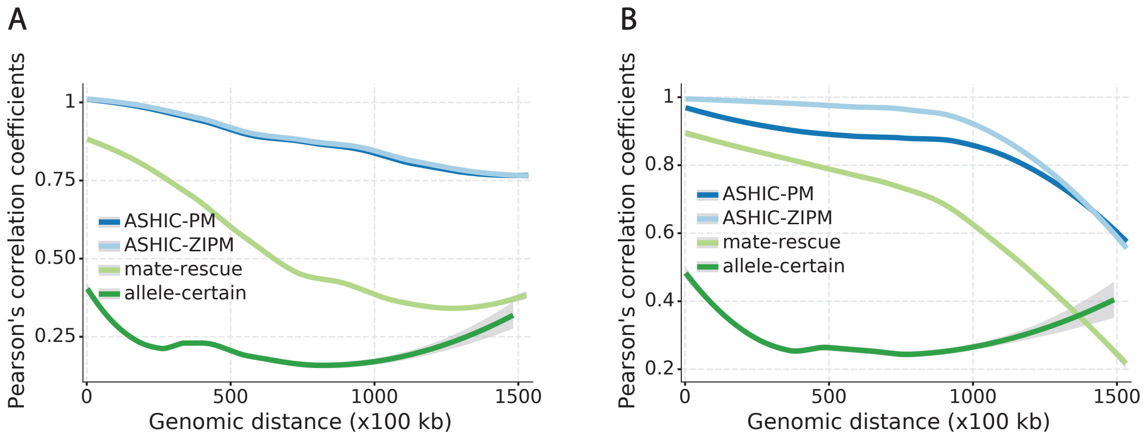

Figure 4.6: ASHIC-ZIPM accurately imputes diploid contact maps on low-SNP-density simulation data. Pearson's correlation coefficients (PCCs) between the imputed diploid contact matrices and the true contact matrices at the lowest SNP density ($\overline{q} = 0.05$) for the homologous X chromosome (A) and the identical-homolog (B) simulation data. The PCC curves are smoothened using the locally weighted LOESS method.

lowered, the F_1 score of our methods decreased only slightly—3.62% for ASHIC-ZIPM and 4.26% for ASHIC-PM. In addition, our ASHIC methods outperformed the other methods by a notable margin. We observed that decreasing SNP density increased the margin between ASHIC-ZIPM and other methods. Taken together, our results demonstrated that the ASHIC-ZIPM method significantly exceeded other methods with high robustness in low SNP density situations.

4.5 Identical Homologous Chromosomal Structures

In the aforementioned simulation settings, we took the homologous X chromosomes in GM12878 cells as the ground truth, where Xa and Xi have drastically dissimilar structures. Unlike the X chromosomes, homologous autosomes often have similar 3D shapes. Imputing diploid Hi-C contact matrices and allelic structures from homologs with similar structures is a more challenging problem than the one from homologs with different structures. To evaluate our methods in such situation, we duplicated the paternal/Xi structure as the pseudo-maternal structure to build an identical homologous structure pair (see Section 4.1.2). We then generated simulation datasets and evaluated our methods at different coverage and SNP density settings, similarly as previously described.

4.5.1 Low Sequencing Coverage Data

As demonstrated in previous homologous structure simulations, our ASHIC methods maintained high accuracy of imputed diploid contact matrices at low sequencing coverage settings (Figure 4.7A). The SCC values were all above 0.9949 for ASHIC-ZIPM and above 0.9938 for ASHIC-PM at various sequencing coverage levels. On the other hand, the SCC values of mate-rescue demonstrated a minor decline from 0.9778 to 0.9664 when the coverage decreased from $100\%\hat{\beta}$ to $10\%\hat{\beta}$. The allele-certain method was the most impacted, as its SCC values declined by 7.46% from 0.8362 to 0.7738 when the coverage level dropped from $100\%\hat{\beta}$ to $10\%\hat{\beta}$.

We then evaluated the accuracy of the allelic 3D structures predicted by our ASHIC methods. Overall, ASHIC-ZIPM generated more accurate structures with smaller distance error rates than the ones predicted by ASHIC-PM across all coverage levels (Figure 4.7B). The improvements were significant at $100\%\hat{\beta}$, $50\%\hat{\beta}$, and $10\%\hat{\beta}$ levels (p-values $= 2.53 \times 10^{-3}, 1.42 \times 10^{-2}, 2.53 \times 10^{-3}$, one-sided paired Wilcoxon signed-rank tests).

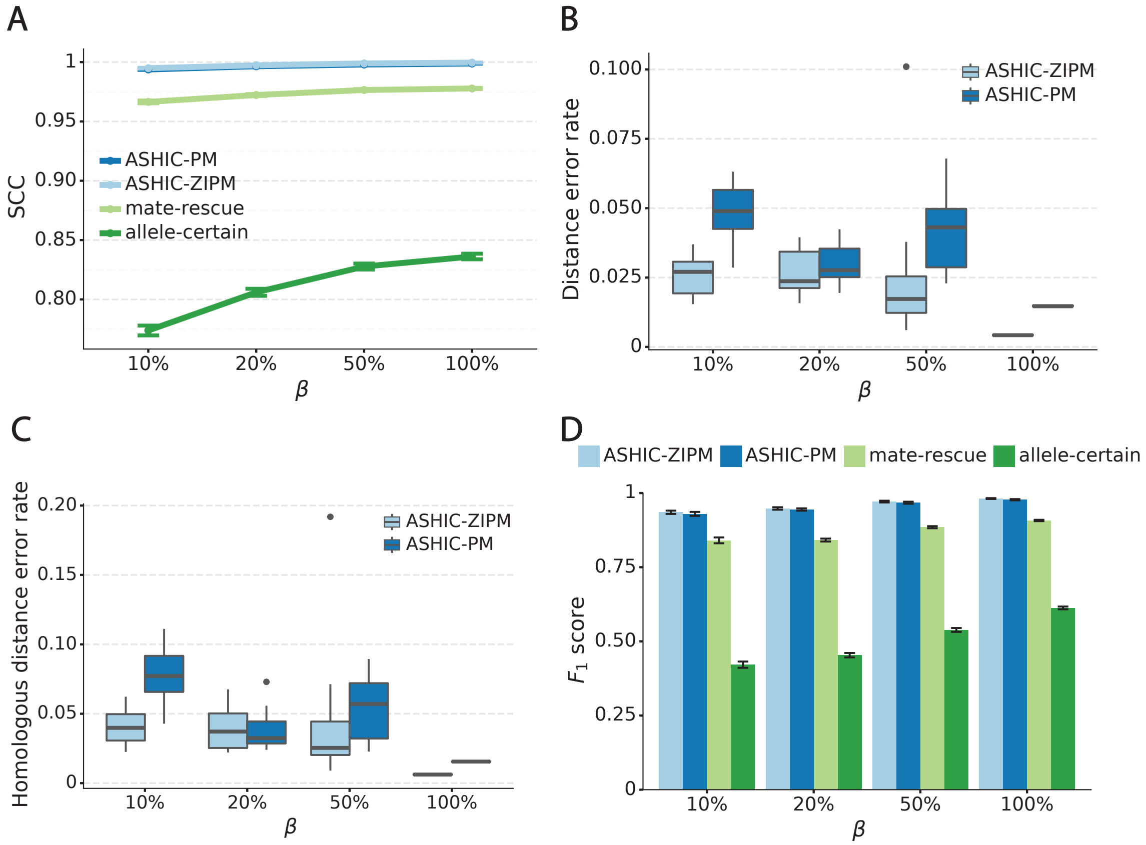

Figure 4.7: ASHIC-ZIPM accurately imputes diploid contact maps and 3D structures on low-coverage identical-homolog simulation data. (A) SCCs between the imputed diploid contact matrices and the true contact matrices, (B) Distance error rates between the predicted allelic 3D structures and the true structures, (C) Homologous distance error rates between the predicted maternal and paternal 3D structures, and (D) F_1 scores of the identified bi-allelic interactions at various sequencing coverage β levels.

In addition to comparing the predicted allelic structures against the ground truth structures, we further calculated the homologous distance error rate between the predicted maternal and paternal structures (Figure 4.7C). For both ASHIC-ZIPM and ASHIC-PM methods, the average homologous distance error rates were smaller than 0.08, suggesting that both models produced homologous structures with very similar shapes. Furthermore, the ASHIC-ZIPM model had significantly lower homologous distance error rates than

ASHIC-PM, at sequencing coverage $100\%\hat{\beta}$ and $10\%\hat{\beta}$ levels (p-values $= 2.53 \times 10^{-3}$, one-sided paired Wilcoxon signed-rank tests). These results further confirmed that ASHIC-ZIPM predicted more accurate allelic 3D structures than the structures predicted by ASHIC-PM.

Next, we investigated the effects of low sequencing coverage on the detection of chromatin interactions when the homologous structures were identical. Similar to the case of different homologous structures, we applied Fit-Hi-C [1] to call significant interactions on the two allele-specific contact matrices separately. Given that the two ground truth homologous structures were identical, we defined the true integration set as the bi-allelic interactions shared by both maternal and paternal chromosomes (see Section 4.3).

Sequencing coverage β	Bi-allelic (true set)	Maternal-specific	Paternal-specific
10%	759.9	365.7	356.5
20%	1559.8	517.7	520.3
50%	3020.6	601.2	592.9
100%	4103.1	679.1	657.5

Table 4.6: Number of allele-specific interactions in duplicate X chromosome (Xi/Xi) simulations. Chromatin interactions are called using Fit-Hi-C [1] on true maternal (Xi) and paternal (Xi) contact matrices separately. Maternal-specific interaction set contains interactions called only from maternal contact matrix but not from paternal contact matrix. Paternal-specific interaction set is defined in a similar way. The bi-allelic interaction set contains common interactions called from both maternal and paternal contact matrices. The true set is defined as the bi-allelic set.

When the coverage dropped from $100\%\hat{\beta}$ to $10\%\hat{\beta}$, the number of interactions in the true set decreased by 81.48% from 4103.1 to 759.9 (Table 4.6). As shown in Figure 4.7D, the allele-certain method was the most impacted by the sequencing coverage changes, where its F_1 scores decreased by 31.23% from 0.6127 to 0.4214 as the coverage dropped from $100\%\hat{\beta}$ to $10\%\hat{\beta}$. The F_1 score of mate-rescue decreased to a less extend, by 7.37% from 0.9075 to 0.8406. Whereas our ASHIC-ZIPM and ASHIC-PM methods demonstrated consistent high F_1 scores of 0.9351 and 0.9296, respectively, even under the lowest coverage $10\%\hat{\beta}$ setting.

4.5.2 Low SNP Density Data

When the SNP density lowered, we observed an overall decreasing trend in the SCC values for all four methods (Figure 4.8A). The allele-certain and mate-rescue methods were greatly impacted by the low SNP density. When the average allele-identifiable probability $\overline{q}$ decreased from 0.5 to 0.05, the SCC values dropped significantly by 46.52% from 0.8362 to 0.4472 for allele-certain and by 9.16% from 0.9778 to 0.8883 for mate-rescue.

Again, our ASHIC methods maintained robustly high accuracy of the imputed contact matrices; the SCC values decreased only by 0.45% from 0.9996 to 0.9950 for ASHIC-ZIPM and by 2.38% from 0.9988 to 0.9750 for ASHIC-PM when $\overline{q}$ decreased from 0.5 to 0.05. The visible difference between ASHIC-ZIPM and ASHIC-PM at the lowest SNP density level $\overline{q} = 0.05$ was also supported by the PCC measures, where ASHIC-ZIPM outperformed ASHIC-PM by an evidently large margin of PCCs within genomic distance of 100 Mb (Figure 4.6B).

In terms of structural accuracy, ASHIC-ZIPM also outperformed ASHIC-PM with significantly smaller distance error rates across all SNP density levels (p-values $= 2.53\times10^{-3}$, one-sided paired Wilcoxon signed-rank tests) (Figure 4.8B). Furthermore, the allelic structures predicted by ASHIC-ZIPM demonstrated significantly smaller homologous distance

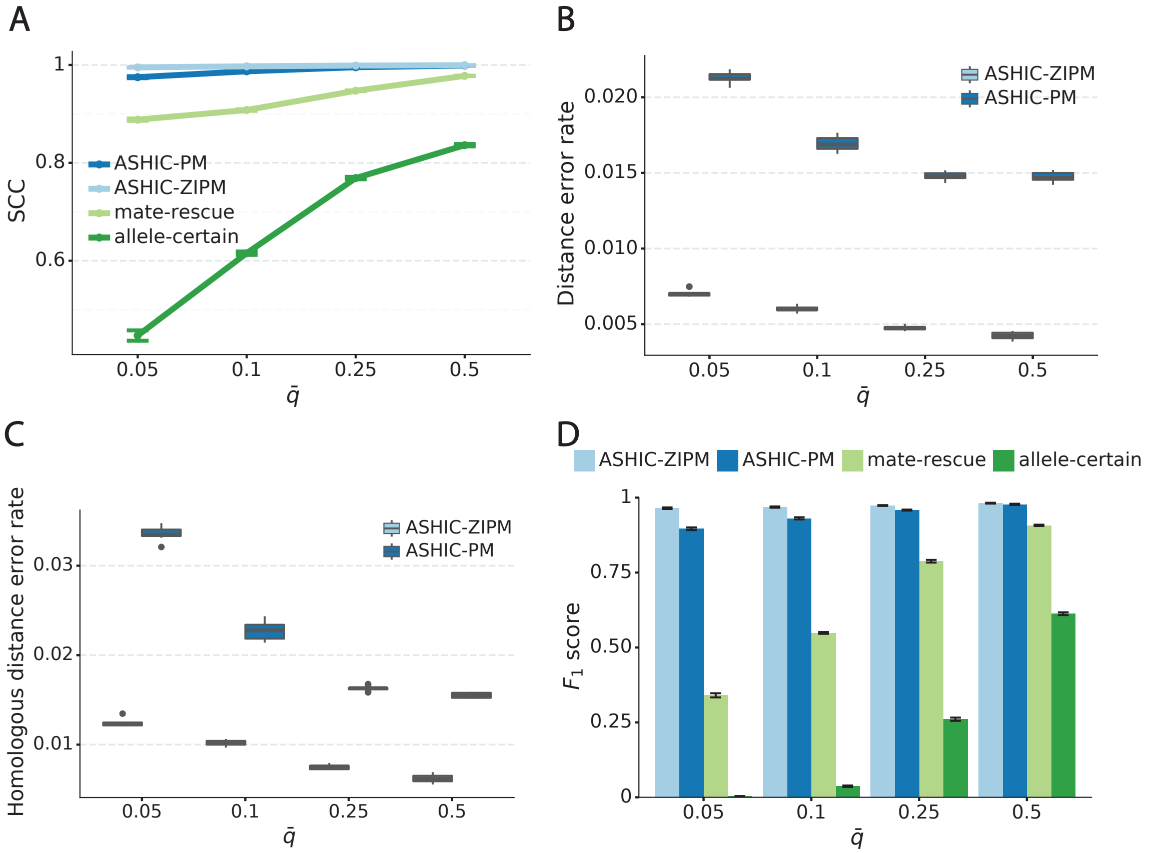

Figure 4.8: ASHIC-ZIPM accurately imputes diploid contact maps and 3D structures on low-SNP-density identical-homolog simulation data. (A) SCCs between the imputed diploid contact matrices and the true contact matrices, (B) Distance error rates between the predicted allelic 3D structures and the true structures, (C) Homologous distance error rates between the predicted maternal and paternal 3D structures, and (D) F_1 scores of the identified bi-allelic interactions at different SNP density $\bar{q}$ levels.

error rates than the ones predicted by ASHIC-PM (p-values $= 2.53 \times 10^{-3}$, at all four $\bar{q}$ levels, one-sided paired Wilcoxon signed-rank tests) (Figure 4.8C).

In addition to achieving the highest imputation accuracy of the diploid contact matrices and 3D structures, ASHIC-ZIPM also demonstrated the best performance with respect to the detection of biallelic chromatin interactions under low SNP density conditions (Figure 4.8D). When the average allele-identifiable probability $\bar{q}$ decreased from 0.5

to 0.05, the F_1 values dropped by 99.43% for allele-certain, 62.49% for mate-rescue, and 8.32% for ASHIC-PM. The ASHIC-ZIPM model showed the smallest decline in F_1 scores, merely 1.70% from 0.9814 to 0.9647. Moreover, we observed that ASHIC-ZIPM significantly outperformed all other methods by a large margin across all SNP density levels (p-values $= 2.53 \times 10^{-3}$, one-sided paired Wilcoxon signed-rank tests)

Taken together, we demonstrated that our ASHIC methods significantly outperformed the allele-certain and mate-rescue methods under low SNP density conditions when the homologous structures have identical shapes. In addition, ASHIC-ZIPM evidently outperformed the ASHIC-PM model by a large margin, especially at the lowest SNP density level.

Chapter 5

Real Data Analyses

We used two diploid Hi-C datasets in our study (Table 4.1). First, allelic mapping results of the wild-type Patski Hi-C dataset published by Bonora et al. [18] were downloaded from GEO (GSE107282). Second, the raw sequencing reads of the GM12878 Hi-C dataset published by Rao et al. [10] were downloaded from GEO (GSE63525) and the allele-specific mapping was performed using HiC-Pro [32]. Briefly, HiC-Pro aligned reads to a masked reference genome where all SNP sites are N-masked. Then reads overlap with SNP sites were assigned to either maternal or paternal allele based on the nucleotide at the SNP position. Reads that do not overlap with any SNPs were labeled as allele-ambiguous. Reads with conflicting allele assignment or unexpected allele at SNP sites were discarded. For each genomic region, we ran 20 random initializations with the ASHIC-ZIPM model and chose the one with the highest likelihood for subsequent analyses.

5.1 Bipartite Structure of Mouse Inactive X Chromosome

The X chromosomes in mammalian females is a representative example of homologous structural difference. In contrast to males having only one X chromosome, the females have two X chromosomes. To compensate for the dosage imbalance of X-linked genes between females and males, one X chromosomes in female cells is randomly silenced through the X chromosome inactivation (XCI) mechanism [33]. To study the structural differences between the active X (Xa) and inactive X (Xi) chromosomes, we applied ASHIC-ZIPM to a published diploid Hi-C data generated from wild-type Patski (BL6×*Spretus*) cells [18]. The Patski cell line has completely skewed XCI such that the maternal BL6 X is always inactive while the paternal *Spretus* X is always active. Several Hi-C studies conducted on the Patski cells have demonstrated that the maternal Xi and paternal Xa chromosomes exhibit distinct morphology and chromatin contact profiles [12, 18]. Specifically, Xi shows a clear bipartite structure, where the entire chromosome is densely packed into two superdomains. The hinge region between the two superdomains contains the macrosatellite repeat locus *Dxz4* and represents a nucleolus-associated domain [12, 14, 18, 10].

To study the bipartite organization of Xi, we applied our ASHIC-ZIPM model to the Patski Hi-C data and reconstructed the diploid contact maps and 3D structures of Xa and Xi at various resolutions (500 kb, 100 kb and 50 kb). As shown in Figure 5.1A, the contact map of Xa demonstrated a clear plaid pattern representing the alternating A/B compartments. In contrast, Xi was clearly separated into two superdomains by a hinge region containing *Dxz4*. We observed frequent intra-superdomain contacts but sparse inter-superdomain contacts on Xi.

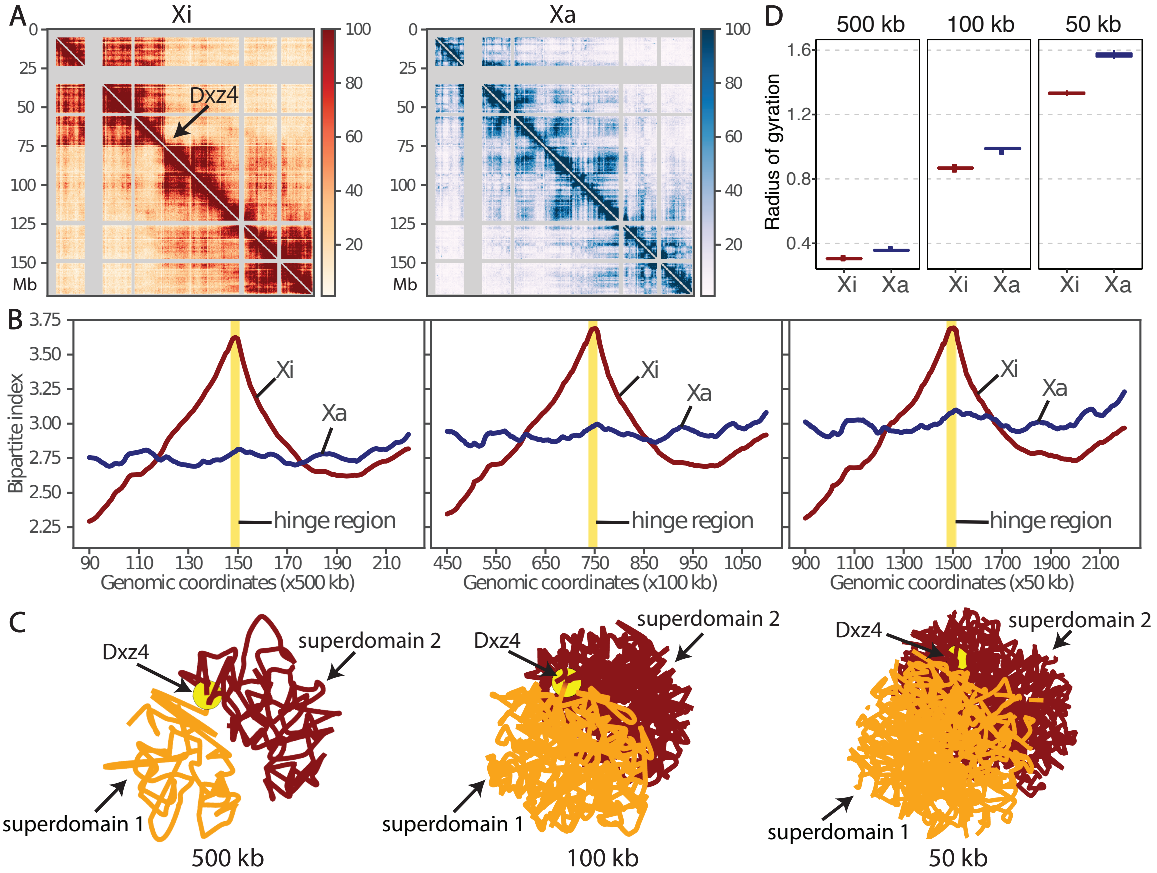

Figure 5.1: Bipartite organization of the inactive X chromosome in mouse Patski cells. (A) ASHIC-ZIPM-imputed allele-specific Hi-C contact matrices of Xi and Xa are shown at 500 kb resolution. The Xi shows a bipartite structure of two superdomains connected by a hinge region (*Dxz4*), indicated by an arrow. Gray strips indicate low mappability regions. (B) Chromosome-wise bipartite index (BI) values for Xi (brown) and Xa (blue) at 500 kb (left), 100 kb (middle), and 50 kb (right) resolutions. The Xi curve shows an evident peak at the hinge region (yellow). (C) The Xi structures predicted by ASHIC-ZIPM at 500 kb, 100 kb, and 50 kb resolutions. The first superdomain (centromeric region) is shown in orange, and the second superdomain (distal region) is shown in brown. The hinge region (*Dxz4*) is marked by a yellow ball. The 3D structures are interpolated and smoothed by the Akima interpolator in `SciPy`. (D) Box plots of the radius of gyration for the Xi (brown) and Xa (blue) structures at 500 kb, 100 kb, and 50 kb resolutions.

To measure how well the two superdomains are separated by the hinge region on the inactive X (Xi) chromosome, we calculated the bipartite index (BI) score [12] for each locus k on chromosomes $\eta \in \{\mathrm{Xi}, \mathrm{Xa}\}$ for both X chromosomes (Figure 5.1B) using the imputed contact matrices by ASHIC-ZIPM. A higher BI score of bin η_k indicated greater bipartite separation at the particular hinge region.

$$\mathrm{BI}(\eta_k) = \frac{\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\mathbb{ZT}_{\eta_i\eta_j}}{k^2} + \frac{\sum_{i=k+1}^{n}\sum_{j=k+1}^{n}\mathbb{ZT}_{\eta_i\eta_j}}{(n-k)^2}}{2\frac{\sum_{i=1}^{k}\sum_{j=k+1}^{n}\mathbb{ZT}_{\eta_i\eta_j}}{k(n-k)}} \tag{5.1}$$

At all three resolutions, we observed an evident BI peak at the hinge region (*Dxz4*) on Xi, confirming the existence of bipartite organization on Xi. In contrast, the BI values were rather flat across the entire Xa, indicating the absence of bipartite structure. These observations demonstrated that our ASHIC-ZIPM method can produce robust and consistent diploid contact maps across different resolutions.

In addition to the existence of two superdomains in the Xi contact map, we also observed that the predicted Xi structures preserved the bipartite conformation across all three resolutions (Figure 5.1C). The two superdomains were clearly separated in space, as each superdomain occupied half of the sphere and there were minimal interactions between them. In addition, the hinge region (*Dxz4*) connecting the two superdomains was located towards the periphery of the Xi structure, which is consistent with previous DNA-FISH results [12]. While the previously published Xa and Xi structures were at 1 Mb [12] and 500 kb [17] resolutions, our method produced chromosomal structures at 50 kb resolution and successfully confirmed the bipartite organization of Xi.

With regards to the overall morphology of the chromosomal structures, we observed that Xi exhibited a more condensed structure than Xa, which is consistent with the fact that Xi is almost entirely silenced.

In particular, to measure the compactness of the estimated X chromosome structures, we calculated the radius of gyration (R_g) [34] (Figure 5.1D), which is defined as the root mean square distance of the points to the centroid, same as the scale of the structure. A smaller R_g value indicates a more compact 3D structure. Across all three resolutions, Xi consistently showed a significantly lower R_g value than Xa, indicating that Xi was more tightly packed (p-values $= 4.43 \times 10^{-5}$, one-sided paired Wilcoxon signed-rank tests).

To assess the reproducibility of the inferred allelic contact maps and 3D structures, we randomly split the X chromosome data into two pseudo-replicates and performed ASHIC-ZIPM analysis on each one separately. At 500 kb resolution, the imputed allelic contact matrices were highly similar with SCC values of 0.9632 (Xi) and 0.9691 (Xa) between the two pseudo-replicates (Figure 5.2). Additionally, the allelic 3D structures estimated from the pseudo-replicates were well aligned with similar global architecture. Moreover, similar results at 100 kb resolution further confirmed the reproducibility of the ASHIC method (Figure 5.3).

Collectively, the results obtained on the Patski Hi-C data demonstrated that our ASHIC-ZIPM method can accurately and robustly detect distinct allele-specific chromatin organizations of Xa and Xi at fine resolution.

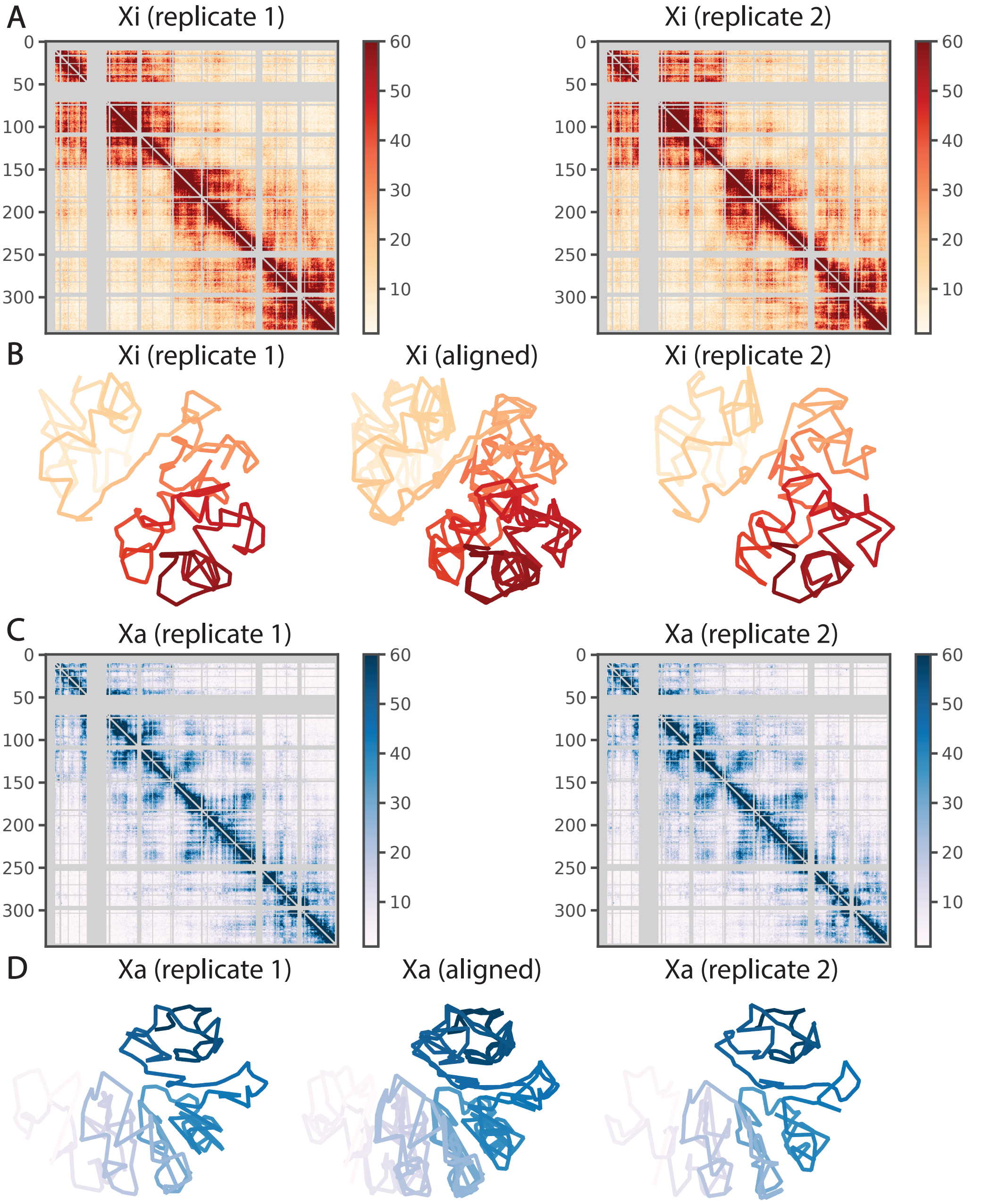

Figure 5.2: Comparison of ASHIC-ZIPM-imputed allele-specific contact matrices and 3D structures of Xi and Xa between two pseudo-replicates at 500 kb resolution. (A) Heatmaps of imputed Xi contact matrices from pseudo-replicate 1 (left) and 2 (right). (B) Predicted 3D structures of Xi from pseudo-replicate 1 (left) and 2 (right), and the optimal alignment between them (middle). (C) Heatmaps of imputed Xa contact matrices from pseudo-replicate 1 (left) and 2 (right). (D) Predicted 3D structures of Xa from pseudo-replicate 1 (left) and 2 (right), and the optimal alignment between them (middle).

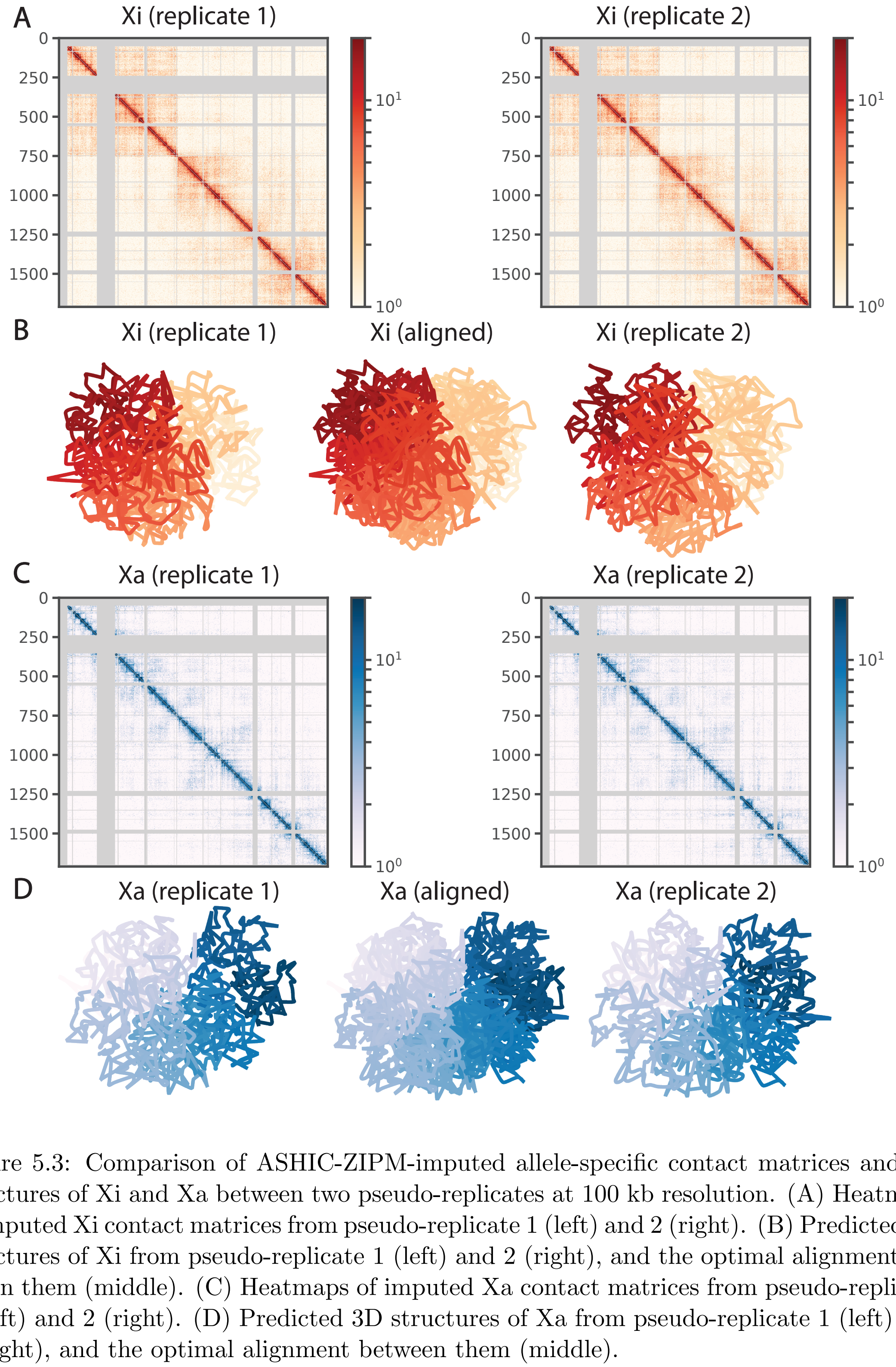

Figure 5.3: Comparison of ASHIC-ZIPM-imputed allele-specific contact matrices and 3D structures of Xi and Xa between two pseudo-replicates at 100 kb resolution. (A) Heatmaps of imputed Xi contact matrices from pseudo-replicate 1 (left) and 2 (right). (B) Predicted 3D structures of Xi from pseudo-replicate 1 (left) and 2 (right), and the optimal alignment between them (middle). (C) Heatmaps of imputed Xa contact matrices from pseudo-replicate 1 (left) and 2 (right). (D) Predicted 3D structures of Xa from pseudo-replicate 1 (left) and 2 (right), and the optimal alignment between them (middle).

5.2 Mouse *H19/Igf2* Imprinting Region

Imprinting is an epigenetic mechanism that causes a subset of genes to express exclusively on one allele in diploid cells. The expression of imprinted genes is controlled by parental-specific epigenetic modifications, such as DNA methylation, at the imprinting control regions. One well-studied example is the *H19/Igf2* imprinting region. In the mouse genome, the paternally expressed *Igf2* gene is located approximately 80 kb upstream (telomeric side) from the long non-coding RNA *H19* that is expressed only on the maternal allele. These two genes demonstrate opposite allele-specific expression yet share a common set of enhancers located downstream of *H19* [35, 36, 37]. It has been shown that the parent-specific expression pattern of *H19* and *Igf2* is controlled by the H19 differentially methylated region (H19-DMR) located 2 kb upstream from *H19* [38]. The H19-DMR is methylated only on the paternal allele, and therefore exhibits methylation-sensitive CCCTC-binding factor (CTCF) binding. On the maternal allele, the unmethylated H19-DMR recruits CTCF bindings and therefore blocking the interactions between the enhancers and *Igf2*. As a result, *Igf2* remains unexpressed, while *H19* can still access the enhancers and thus is activated. Whereas on the paternal allele, the methylated H19-DMR inhibits CTCF bindings. Consequently, *Igf2* can access the enhancers and being activated; while the *H19* silencing is likely caused by spreading of methylation from H19-DMR [39].

It has been widely speculated that CTCF attains enhancer-blocking insulation function via the formation of chromatin loops [40]. Using diploid Hi-C contact maps of human GM12878 cells at 25 kb resolution, Rao et al. [10] examined the *H19/IGF2* imprinting region and identified parental-specific chromatin loops between the *H19/IGF2* cluster and a distal region which was referred to as the H19/Igf2 Distal Anchor Domain (HIDAD). The HIDAD-*H19* loop was present exclusively on maternal allele; in contrast, the HIDAD-*IGF2* loop appeared only on the paternal allele. Additionally, Llères et al. [41] performed a diploid 4C-seq study on the mouse ESCs and showed that H19-DMR interacted significantly more with the mouse homologue of HIDAD (mHIDAD) on maternal allele compared to the interactions on the paternal allele. They subsequently performed 3D DNA-FISH experiments and confirmed that the distances between mHIDAD and *H19* were significantly shorter on the maternal allele than the distances on the paternal allele.

Although the aforementioned 4C-seq study [41] and several other 3C studies [42, 43, 44] have been conducted in the *H19/Igf2* imprinting region, diploid Hi-C studies are still restricted to a rather coarse resolution due to the limitations of low SNP density and insufficient sequencing coverage. To bridge this gap and provide a holistic view of chromatin structures on the imprinted *H19/Igf2* region, we applied our ASHIC-ZIPM method to the published diploid Hi-C data in mouse Patski cells [18], and generated fine-scale allele-specific contact maps and 3D structures of a 5-Mb region (chr7: 140–145 Mbp) around the *H19/Igf2* imprinting region at 10 kb resolution.

First, we constructed a differential contact map using log-fold-change values between the imputed maternal and paternal contacts (Figure 5.4A). Along with the contact map, we also visualized the allelic CTCF ChIP-seq data [18]. Consistent with previous studies [45, 46], we observed a clear maternal-specific CTCF binding at the H19-DMR locus. Additionally, a few bi-allelic CTCF binding clusters were observed at mHIDAD, near the *Syt8* and *Lsp1* genes, and at the telomeric side of *Igf2*. As shown in Figure 5.4A, the contacts between mHIDAD and *H19* were enriched on the maternal allele (box 1), whereas the contacts between mHIDAD and *Igf2* were enriched on the paternal allele (box 2). In addition to the contacts between mHIDAD and *H19/Igf2*, *H19* and *Igf2* demonstrated differential contact preferences to the bi-allelic CTCF clusters near *Syt8* and *Lsp1* (boxes 3 and 4). To further characterize the parental-specific chromatin interactions, we identified chromatin loops with genomic distance of 30–500 kb from the imputed allelic contact maps using Fit-Hi-C [1] with a strict FDR threshold (q-value $< 10^{-5}$). The identified chromatin loops were mostly anchored to the CTCF binding clusters (Figure 5.4A). We further categorized these chromatin loops into bi-allelic loops that were shared between the two alleles, or monoallelic loops that are either maternal-specific or paternal-specific. Consistent with the differential contact map, chromatin loops anchored at *H19* and *Igf2* were primarily parental-specific. We observed a distinct pattern of maternal-specific chromatin loops between mHIDAD and *H19* and paternal-specific chromatin loops between mHIDAD and *Igf2*. Besides mHIDAD, the region containing bi-allelic CTCF binding clusters near the *Syt8* and *Lsp1* genes also demonstrated parental-specific chromatin interactions with *H19* and *Igf2*. Specifically, these CTCF clusters interacted preferentially with *H19* on the maternal allele and with *Igf2* on the paternal allele. These observations are consistent with the previous 4C-seq results in mouse ESCs [41].

Besides the differential contact map, we also examined the allele-specific chromatin conformations using the predicted allelic 3D structures (Figure 5.4B). The overall chromatin organizations of the *H19/Igf2* imprinting region appeared to be similar between the two alleles. However, the relative spatial position among mHIDAD, *H19*, and *Igf2* demonstrated parental-specific differences. From the 3D structures, we observed that mHIDAD was spatially close to *H19* on the maternal allele, presumably forming a chromatin loop. In addition, we observed that *Igf2* was much closer to mHIDAD on paternal structure than on the maternal structure.

For the quantitative comparison, we calculated the pairwise Euclidean distances of mHIDAD, *H19*, and *Igf2* on the maternal and paternal structures predicted by ASHIC-ZIPM from 20 random initializations. As shown in Figure 5.4C, the distance between mHIDAD and *H19* was significantly smaller on the maternal structure than that on the paternal structure (p-value $= 4.43 \times 10^{-5}$, one-sided Wilcoxon paired signed-rank test), which is consistent with the previous DNA-FISH data [41]. In contrast, the distance between mHIDAD and *Igf2* was significantly larger on maternal allele (p-value $= 4.43 \times 10^{-5}$, one-sided Wilcoxon paired signed-rank test), which is consistent with the observation of paternal-specific HIDAD-*IGF2* loop in human GM12878 cells [10]. No significant difference of the distance between *H19* and *Igf2* was detected on our predicted allelic structures. These observations demonstrated that our method can stably predict fine-scale 3D structures that reflect the distinct parental-specific chromatin conformations.

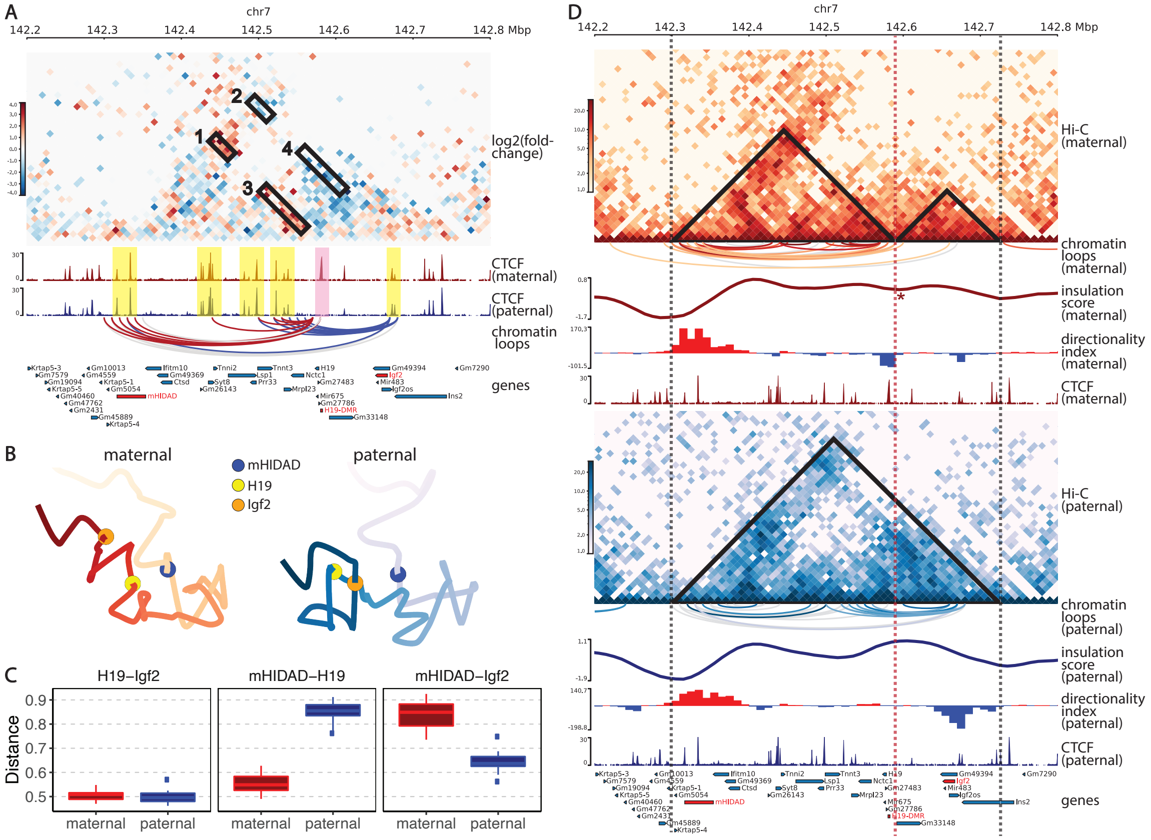

Figure 5.4: Allele-specific chromatin organizations of the *H19/Igf2* imprinting region in mouse Patski cells. (A) Differential contact map between the ASHIC-ZIPM-imputed maternal and paternal contacts at 10 kb resolution. Contact counts are normalized separately on each allele to account for the potential mapping bias towards the reference genome. The red vs blue color key indicates maternal vs paternal enrichment. Four allelicly enriched chromatin interacting regions are labeled in boxes 1–4. Maternal-specific CTCF peak (pink) and bi-allelic CTCF binding clusters (yellow) are highlighted. Chromatin loops are called using Fit-Hi-C [1] and categorized into maternal-specific (red), paternal-specific (blue), and bi-allelic (gray). Only loops anchored at *H19* or *Igf2* are displayed. (B) Allelic 3D structures of the *H19/Igf2* imprinting region predicted by ASHIC-ZIPM at 10 kb resolution. The maternal (red) and paternal (blue) structures are overall similar, but the relative spatial positions of mHIDAD (blue), *H19* (yellow), and *Igf2* (orange) are evidently different. (C) Box plots of pairwise Euclidean distances between *H19*-*Igf2* (left), mHIDAD-*H19* (middle), and mHIDAD-*Igf2* (right). (D) Allelic Hi-C contact maps at 10 kb resolution (top panel: maternal allele, red color key; bottom panel: paternal allele, blue color key). Maternal-specific (red), paternal-specific (blue), and bi-allelic (gray) chromatin loops are called using Fit-Hi-C [1]. A local minimum of the insulation score (IS) is marked by an asterisk. Positive and negative directionality index (DI) values [2] are shown in red and blue, respectively. (Sub-)TAD domains derived from IS and DI measures are labeled as triangles on the contact maps, and dashed lines indicate (sub-)TAD boundaries. Panels (A) and (D) are drawn using `pyGenomeTracks` [3].

5.2.1 Maternal-Specific Sub-TAD at Mouse *H19/Igf2* Locus

In addition to the formation of chromatin loops, CTCF also participates in the establishment of higher-order chromatin structures such as topologically-associating domains (TADs). TADs are sub-megabase genomic regions containing frequent local chromatin interactions, whereas TAD boundaries result in physical insulation between neighboring domains [2]. It has been observed that CTCF bindings are often enriched at TAD boundaries and play an important role in TAD formation [2, 10]. Since the genome is organized in a hierarchical manner, smaller domains called sub-TADs are often observed within the large TADs. Unlike TADs that are mostly invariant between cell types, sub-TADs are more variable and play a pivotal role in mediating cell-type-specific gene regulation [47, 48]. Based on the presence of monoallelic CTCF bindings at H19-DMR, Llères et al. [41] proposed a novel parental-specific sub-TAD model for the regulation of imprinting at *H19/Igf2* locus. Supported by allelic 4C-seq and DNA-FISH data, they speculated that several bi-allelic CTCF binding sites form a first layer of TAD on both alleles. In addition, the maternal-specific CTCF binding around H19-DMR hijacks the first layer of TAD and consequently creates an additional layer of sub-TAD on the maternal allele.

To verify this hypothesis, we calculated the insulation score (IS) [49] and directionality index (DI) [2] using `TADtool` [50] to search for possible (sub-)TAD boundaries around the *H19/Igf2* imprinting region. Overall we observed similar IS values on both alleles, except at the H19-DMR locus (Figure 5.4D, Figure 5.5). Specifically, we observed a local minimum of IS values at H19-DMR only on the maternal allele indicating a potential presence of a sub-TAD boundary at H19-DMR. Consistently, the DI values suggested similar (sub-)TAD pattern (Figure 5.4D). We observed strong positive DIs at mHIDAD on both alleles, indicating that mHIDAD is highly biased towards interacting with its downstream loci and serves as a starting position of a TAD. On the other hand, the telomeric-side flanking region of *Igf2* demonstrated negative DIs on both alleles, indicating a likely ending boundary of a TAD. Furthermore, a negative DI region around H19-DMR appeared only on the maternal allele, suggesting H19-DMR has a higher tendency to interact with its upstream loci, possibly indicating an ending position for a maternal-specific sub-TAD.

Both the IS and DI measurements suggested that *H19/Igf2* is embedded within a TAD demarcated by two main boundaries: one near mHIDAD and the other one at the telomeric side of *Igf2*. The locations of the two boundaries were in good agreement between both alleles. However, the (sub-)TAD organization within this TAD region undergoes drastic parental-specific changes. Specifically, we observed a sub-TAD boundary at H19-DMR locus exclusively on the maternal allele. The TAD and sub-TAD boundaries mentioned above were all located at CTCF binding clusters. We further examined the allelic chromatin loops within this imprinting region (Figure 5.4D). On the maternal allele, chromatin loops were mostly confined to the mHIDAD-*H19* sub-TAD. Whereas on the paternal allele, we observed several chromatin loops connecting the centromeric side of H19-DMR with *Igf2*, indicating the absence of insulation at H19-DMR. These observations of allelic chromatin loops are consistent with the parental-specific (sub-)TAD structures.

Taken together, these results supported the hypothesis that the maternal-specific CTCF binding at H19-DMR forms a chromatin loop with the CTCF binding sites at mHIDAD. This mHIDAD-*H19* loop creates an additional layer of sub-TAD inside the original

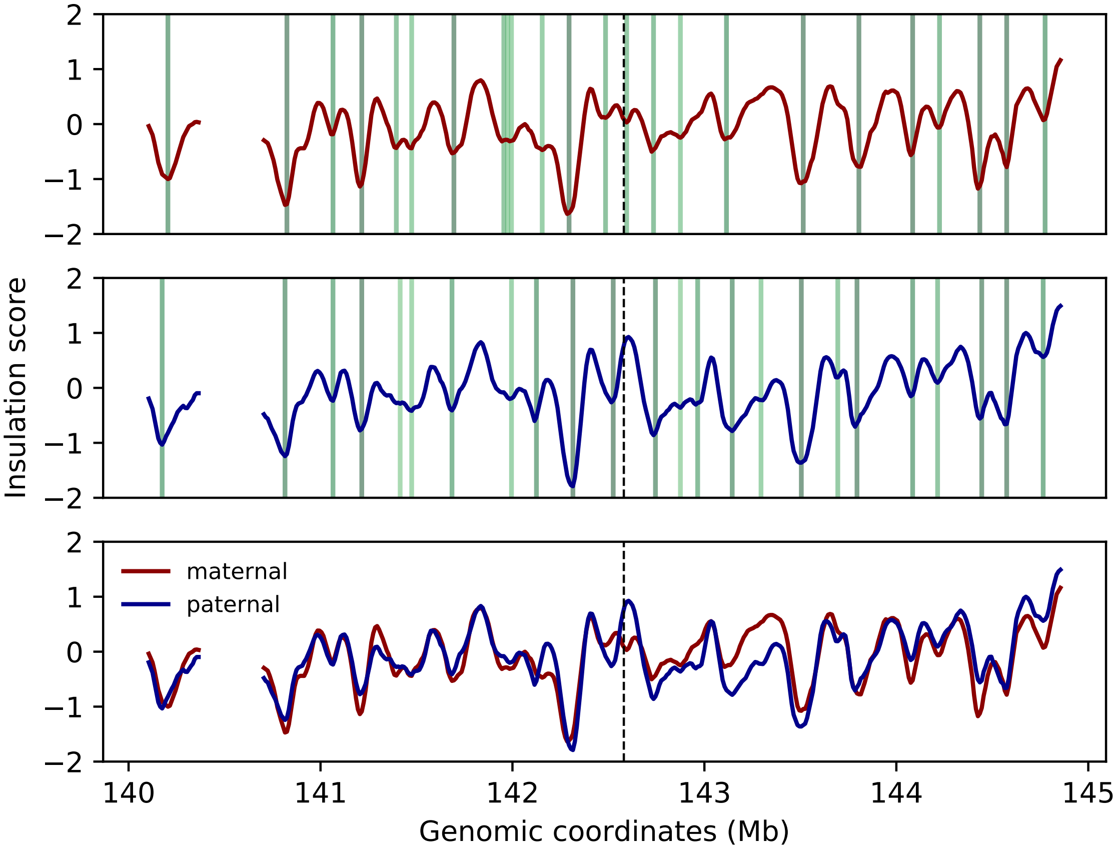

Figure 5.5: Insulation scores of the *H19/Igf2* imprinting region show potential sub-TAD boundary at the H19-DMR locus on maternal chromosome. Insulation scores of chr7:140–145 Mb on maternal chromosome (top) and paternal chromosome (middle). Green lines are the identified (sub)TAD boundaries. Black dashed lines are at the H19-DMR locus. Insulation scores of the maternal and paternal chromosome are drawn together in the bottom panel for better visual comparison.

mHIDAD-*Igf2* TAD. The maternal-specific mHIDAD-*H19* sub-TAD organization mediates the insulation between the centromeric side of H19-DMR and *Igf2*, and thereby leading to the silencing of *Igf2* on the maternal allele.

5.3 Allelic Chromatin Contacts in Human GM12878 Cells

Besides Hi-C, ChIA-PET is another popular technique for detecting genome-wide chromatin interactions [51]. ChIA-PET incorporates chromatin immunoprecipitation-based enrichment and focuses on the mapping of chromatin interactions mediated by a specific protein of interest. Applying an advanced long-read ChIA-PET strategy, Tang et al. [4] comprehensively mapped the functional chromatin interactions mediated by CTCF and RNA polymerase II (RNAPII) with haplotype specificity in human cell lines. To further assess our method, we applied ASHIC-ZIPM to the published Hi-C data in human GM12878 cells [10], and compared the imputed allelic chromatin maps with the phased ChIA-PET data published by Tang et al. [4].

We first looked at a 4-Mb region (chr11: 1–5 Mbp) around the H19/IGF2 imprinting locus and generated allelic contact maps and structures at 10 kb resolution (Figure 5.6A). Compared to the 25-kb-resolution mate-rescued Hi-C maps reported by Rao et al. [10], our ASHIC-imputed allelic contact maps showed much higher coverage and finer interaction patterns. Similar to the mouse *H19/Igf2* region, the human *H19/IGF2* imprinting region also exhibited a maternal-specific sub-TAD organization. The sub-TAD boundary located at H19-DMR and was enriched with maternal-specific CTCF bindings. In addition, we observed maternal-specific chromatin contacts between H19-DMR and several loci (including HIDAD) at the telomeric side (red boxes), which was in high correspondence with the maternal-biased ChIA-PET loops mediated by CTCF. On the paternal allele, we observed enriched chromatin contacts between *IGF2* and the aforementioned telomeric-side loci (blue boxes), which was consistent with our observations with the mouse *Igf2* homolog. We did not observe the corresponding paternal-biased CTCF ChIA-PET loops, probably due to the absence of SNPs at the *IGF2* locus.

In addition to CTCF-mediated parental-specific chromatin loops, our approach also revealed RNAPII-mediated allelic chromatin interactions. For example, we studied another 4-Mb region (chr12: 8–12 Mbp) containing the *LOC374443*, *CLEC2D*, and *CLECL1* multi-gene complex. Previously, Tang et al. [4] discovered paternally biased RNAPII-mediated interactions between this paternally expressed multi-gene complex and its distal enhancer (300 kb apart). Consistently, our ASHIC-imputed allelic contact maps showed paternal-enriched long-range contacts (blue box) between the distal enhancer and the promoters of the three genes, as shown in Figure 5.6B.

Collectively, these results demonstrated that our ASHIC method is capable of imputing diploid chromatin maps in low-SNP-density cells such as GM12878 and the ASHIC-imputed allelic contacts are in high correspondence with the phased ChIA-PET data.

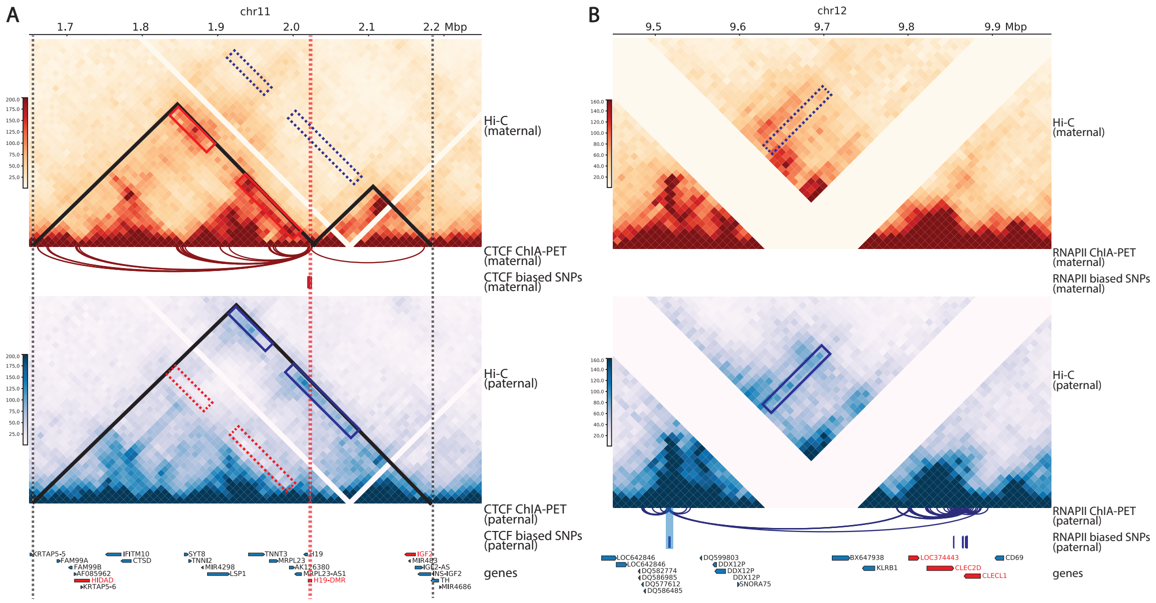

Figure 5.6: Allele-specific Hi-C chromatin maps and ChIA-PET loops in human GM12878 cells. ASHIC-imputed allelic contact maps are shown at 10 kb resolution (top panel: maternal allele, red color key; bottom panel: paternal allele, blue color key). Phased ChIA-PET loops and SNPs with haplotype-biased ChIA-PET bindings are obtained from Tang et al. [4]. (A) *H19/IGF2* imprinting region. Maternal-enriched and paternal-enriched chromatin interacting regions are labeled in red and blue boxes, respectively. Vertical dashed lines indicate (sub-)TAD boundaries. (B) Allelic long-range enhancer-promoter interactions at *LOC374443*, *CLEC2D*, and *CLECL1* genes. Blue box indicates the paternal-enriched chromatin interacting region. The distal enhancer associated with paternal-biased RNAPII-mediated ChIA-PET loops is highlighted in blue. Both panels are drawn using `pyGenomeTracks` [3].

Chapter 6

Conclusions

In this work, we proposed a hierarchical Bayesian framework for imputing allele-specific contacts and reconstructing allelic 3D structures from diploid Hi-C data. We developed two models under this Bayesian framework: ASHIC-PM and ASHIC-ZIPM. To the best of our knowledge, our ASHIC methods are the first methods that produce fully decomposed diploid Hi-C contact matrices as well as the allelic 3D structures.

Unlike the existing allele-certain and mate-rescue approaches, our ASHIC methods utilize all diploid Hi-C contacts, including both-end allele-ambiguous contacts. As a result, ASHIC methods exceeded the allele-certain and mate-rescue methods, in terms of producing more accurate diploid matrices and structures as well as facilitating better detection of allele-specific chromatin interactions. We also conducted a series of simulation experiments and evaluated how the performance of our methods was impacted by various factors, including sequencing coverage, SNP density, and homologous structural similarity. Overall, our models significantly outperformed other methods, especially under low sequencing coverage and low SNP density conditions. The ability of the ASHIC methods in inferring allele-ambiguous contacts at low-SNP-density setting is critical for analyses in diploid human cells such as GM12878, where the existing mate-rescue method [10] was only able to rescue 5.86% of total diploid contacts (Table 1.1).

In our simulation studies, we did not compare the ASHIC methods with the recently published Dip-C method by Tan et al. [16] as their method was specifically designed for single-cell Hi-C data. Another reason was that Dip-C does not impute intra-chromosomal both-end allele-ambiguous contacts. Therefore we expect that its performance would be close to the mate-rescue method. In addition, our earlier work of the Poisson-Gamma model [12] imputes diploid contact counts based on genomic distances rather than spatial distances, and therefore is not computationally stable on fine-resolution (such as 100 kb) or low-coverage Hi-C data. Lastly, the newly developed diploid-PASTIS method by Cauer et al. [17] predicts only the allelic 3D structures rather than the diploid contact matrices. Therefore, we did not evaluate the diploid-PASTIS method in our simulations as most of our evaluation metrics were based on imputed contact matrices.

The main advantage of the ASHIC-ZIPM model over the ASHIC-PM model is that ASHIC-ZIPM explicitly accounts for the excessive zeros in Hi-C matrices, by modeling the probabilities whether each observed zero count is a “true” zero or a “missing” zero. As a result, we observed that the ASHIC-ZIPM model consistently outperformed the ASHIC-PM model in all simulation settings. While the performance of the two models were often similar, the improvements of ASHIC-ZIPM over ASHIC-PM became more evident when the SNP density decreased. In addition, the differences between the ASHIC-ZIPM and

ASHIC-PM models were particularly noticeable under the more challenging simulation setting of identical homologous structures. This is owing to the fact that when SNP density was low, only few allele-certain contacts were observed. The ASHIC-PM model uses the allele-certain contacts to initialize the EM algorithm and treats all zeros as “true” zeros, thereby producing less optimal results. In contrast, ASHIC-ZIPM explicitly adjusts the weights between “true” and “missing” zeros and thereby archiving more accurate models.

Hi-C contact counts could be over-dispersed, thus a Negative Binomial (NB) model may provide a better fit than a Poisson model. However, our ASHIC models leverage on two nice properties of the Poisson distribution: the outcomes from a Poisson-multinomial hierarchical model are Poisson variables; and the sum of Poisson variables is also a Poisson variable. If we adapt a NB model, we will no longer have such a neat and tractable hierarchical model and as a result the model fitting will become computationally expensive. In addition, we would like to point out that the ZIP model can account for over-dispersion to some extent by fitting a mixture of Poisson state and the zero (missing) state. Furthermore, the ASHIC methods use the spatial distance rather than the genomic distance between the contacting pair as the Poisson or ZIP parameter, therefore could be less impacted by the over-dispersion.

We demonstrated the applications of our ASHIC-ZIPM method in the mouse Patski cells and in the human GM12878 cells. Previous studies predicted allelic X chromosome structures at 1 Mb [12] and 500 kb [17] resolutions. In contrast, our method utilized all diploid contacts and produced finer-scale allelic structures of the entire X chromosomes at 50-kb resolution. Our results further confirmed the existence of the bipartite structure of Xi. The ability to impute all allele-ambiguous contacts is particularly important when zooming into local imprinting regions. Since imprinting regions are often small, fine-resolution allelic contact maps and 3D structures are required for an in-depth study. With our ASHIC-ZIPM model, we produced the first 10-kb-resolution diploid Hi-C contact maps of the mouse *H19/Igf2* imprinting region, and revealed the existence of the maternal-specific sub-TAD organization at H19-DMR. This sub-TAD formation creates an insulation between *H19* and *Igf2* that likely prevents the activation of *Igf2* on the maternal allele. Our study of the human *H19/IGF2* imprinting region further confirmed this parental-specific chromatin organization. Furthermore, the ASHIC-imputed diploid Hi-C maps offered an informative view of the (sub-)TAD organizations on the imprinting region, whereas the previous 4C-seq study [41] was restricted to only few anchor regions.

Currently, only a few limitations can be attributed to our ASHIC methods. First, our methods provide chromosome-wide modeling of diploid Hi-C data. One possible future extension is to build a genome-wide model by incorporating an additional estimation step in the EM algorithm to model the relative position of multiple homologous chromosomes. We could further parallelize the optimization procedures for each homologous chromosome pair to speed up the genome-wide modeling. Second, our model is specifically designed for diploid genomes. Extending the model to polyploid or aneuploid genomes remains a challenging problem. Third, the computational efficiency of our EM algorithm, especially the structure estimation step, could be further improved. One possible solution is to adapt an iterative modeling strategy similar to [16, 34], starting with coarse-resolution modeling then through interpolation to gradually refine the structures to finer resolutions. Lastly, it is possible to incorporate biological replicates into our models to improve reproducibility. For example, we can optimize either a single structure or an ensemble of structures to simultaneously find the best fit for all supplied biological replicates.

Bibliography

[1] F. Ay, T. L. Bailey, and W. S. Noble. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome research*, 24(6):999–1011, 2014.

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

[3] F. Ramírez, V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Grüning, J. Villaveces, B. Habermann, A. Akhtar, and T. Manke. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature communications*, 9(1):1–15, 2018.

[4] 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.

[5] J. Dekker. Gene regulation in the third dimension. *Science*, 319(5871):1793–1794, 2008.

[6] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[7] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367, 2010.

[8] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, 30(1):90–98, 2012.

[9] W. Ma, F. Ay, C. Lee, G. Gulsoy, X. Deng, S. Cook, J. Hesson, C. Cavanaugh, C. B. Ware, A. Krumm, J. Shendure, C. A. Blau, C. M. Disteche, W. S. Noble, and Z. Duan.

Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of lincRNA genes in human cells. *Nature methods*, 12(1):71–78, 2015.

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

[11] V. Ramani, D. A. Cusanovich, R. J. Hause, W. Ma, R. Qiu, X. Deng, C. A. Blau, C. M. Disteche, W. S. Noble, J. Shendure, and Z. Duan. Mapping 3D genome architecture through in situ DNase Hi-C. *Nature protocols*, 11(11):2104–2121, 2016.

[12] X. Deng, W. Ma, V. Ramani, A. Hill, F. Yang, F. Ay, J. B. Berletch, C. A. Blau, J. Shendure, Z. Duan, W. S. Noble, and C. M. Disteche. Bipartite structure of the inactive mouse X chromosome. *Genome biology*, 16(1):152, 2015.

[13] L. Giorgetti, B. R. Lajoie, A. C. Carter, M. Attia, Y. Zhan, J. Xu, C. J. Chen, N. Kaplan, H. Y. Chang, E. Heard, and J. Dekker. Structural organization of the inactive X chromosome in the mouse. *Nature*, 535(7613):575–579, 2016.

[14] E. M. Darrow, M. H. Huntley, O. Dudchenko, E. K. Stamenova, N. C. Durand, Z. Sun, S. C. Huang, A. L. Sanborn, I. Machol, M. Shamim, A. P. Seberg, E. S. Lander, B. P. Chadwick, and E. L. Aiden. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences*, 113(31):E4504–E4512, 2016.

[15] Z. Du, H. Zheng, B. Huang, R. Ma, J. Wu, X. Zhang, J. He, Y. Xiang, Q. Wang, Y. Li, J. Ma, X. Zhang, K. Zhang, Y. Wang, M. Q. Zhang, J. Gao, J. R. Dixon, X. Wang, J. Zeng, and W. Xie. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature*, 547(7662):232–235, 2017.

[16] L. Tan, D. Xing, C. H. Chang, H. Li, and X. S. Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405):924–928, 2018.

[17] A. G. Cauer, G. Yardimci, J. P. Vert, N. Varoquaux, and W. S. Noble. Inferring diploid 3D chromatin structures from Hi-C data. In *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[18] G. Bonora, X. Deng, H. Fang, V. Ramani, R. Qiu, J. B. Berletch, G. N. Filippova, Z. Duan, J. Shendure, W. S. Noble, and C. M. Disteche. Orientation-dependent Dxz4 contacts shape the 3D structure of the inactive X chromosome. *Nature communications*, 9(1):1445, 2018.

[19] N. Varoquaux, F. Ay, W. S. Noble, and J. P. Vert. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, 30(12):i26–i33, 2014.

[20] S. Wang, J. Xu, and J. Zeng. Inferential modeling of 3D chromatin structure. *Nucleic acids research*, 43(8):e54–e54, 2015.

[21] J. Dekker, M. A. Marti-Renom, and L. A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6):390–403, 2013.

[22] Eitan Yaffe and Amos Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43(11):1059, 2011.

[23] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999, 2012.

[24] Julio Mateos-Langerak, Manfred Bohn, Wim de Leeuw, Osdilly Giromus, Erik MM Manders, Pernette J Verschure, Mireille HG Indemans, Hinco J Gierman, Dieter W Heermann, Roel Van Driel, et al. Spatially confined folding of chromatin in the interphase nucleus. *Proceedings of the National Academy of Sciences*, 106(10):3812–3817, 2009.

[25] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[26] Joseph B Kruskal. *Multidimensional scaling*. Number 11. Sage, 1978.

[27] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[28] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

[29] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[30] T. E. Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[31] T. Yang, F. Zhang, G. G. Yardımcı, F. Song, R. C. Hardison, W. S. Noble, F. Yue, and Q. Li. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome research*, 27(11):1939–1949, 2017.

[32] Nicolas Servant, Nelle Varoquaux, Bryan R Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. Hic-pro: an optimized and flexible pipeline for hi-c data processing. *Genome biology*, 16(1):259, 2015.

[33] M. F. Lyon. Gene action in the X-chromosome of the mouse (Mus musculus L.). *nature*, 190(4773):372–373, 1961.

[34] T. J. Stevens, D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, M. Leeb, K. J. Wohlfahrt, W. Boucher, A. O'Shaughnessy-Kirwan, J. Cramard, A. J. Faure, M. Ralser, E. Blanco, L. Morey, M. Sansó, M. G. S. Palayret, B. Lehner, L. D. Croce, A. Wutz, B. Hendrich, D. Klenerman, and E. D. Laue. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648):59–64, 2017.

[35] H. Yoo-Warren, V. Pachnis, R. S. Ingram, and S. M. Tilghman. Two regulatory domains flank the mouse H19 gene. *Molecular and Cellular Biology*, 8(11):4707–4715, 1988.

[36] P. A. Leighton, J. R. Saam, R. S. Ingram, C. L. Stewart, and S. M. Tilghman. An enhancer deletion affects both H19 and Igf2 expression. *Genes & development*, 9(17):2079–2089, 1995.

[37] K. Ishihara, N. Hatano, H. Furuumi, R. Kato, T. Iwaki, K. Miura, Y. Jinno, and H. Sasaki. Comparative genomic sequencing identifies novel tissue-specific enhancers and sequence elements for methylation-sensitive factors implicated in Igf2/H19 imprinting. *Genome research*, 10(5):664–671, 2000.

[38] J. L. Thorvaldsen, K. L. Duran, and M. S. Bartolomei. Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2. *Genes & development*, 12(23):3693–3702, 1998.

[39] D. P. Barlow and M. S. Bartolomei. Genomic imprinting in mammals. *Cold Spring Harbor perspectives in biology*, 6(2):a018382, 2014.

[40] M. Merkenschlager and D. T. Odom. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell*, 152(6):1285–1297, 2013.

[41] D. Llères, B. Moindrot, R. Pathak, V. Piras, M. Matelot, B. Pignard, A. Marchand, M. Poncelet, A. Perrin, V. Tellier, R. Feil, and D. Noordermeer. CTCF modulates allele-specific sub-TAD organization and imprinted gene activity at the mouse Dlk1-Dio3 and Igf2-H19 domains. *Genome Biology*, 20(1):1–17, 2019.

[42] A. Murrell, S. Heeson, and W. Reik. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nature genetics*, 36(8):889–893, 2004.

[43] F. Court, M. Baniol, H. Hagege, J. S. Petit, M. N. Lelay-Taha, F. Carbonell, M. Weber, G. Cathala, and T. Forne. Long-range chromatin interactions at the mouse Igf2/H19 locus reveal a novel paternally expressed long non-coding RNA. *Nucleic acids research*, 39(14):5893–5906, 2011.

[44] S. Kurukuti, V. K. Tiwari, G. Tavoosidana, E. Pugacheva, A. Murrell, Z. Zhao, V. Lobanenkov, W. Reik, and R. Ohlsson. CTCF binding at the H19 imprinting control

region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proceedings of the national academy of sciences*, 103(28):10684–10689, 2006.

[45] A. C. Bell and G. Felsenfeld. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, 405(6785):482–485, 2000.

[46] A. T. Hark, C. J. Schoenherr, D. J. Katz, R. S. Ingram, J. M. Levorse, and S. M. Tilghman. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, 405(6785):486–489, 2000.

[47] J. E. Phillips-Cremins, M. E. G. Sauria, A. Sanyal, T. I. Gerasimova, B.R. Lajoie, J. S. K. Bell, C. T. Ong, T. A. Hookway, C. Guo, Y. Sun, M. J. Bland, W. Wagstaff, S. Dalton, T. C. McDevitt, R. Sen, J. Dekker, J. Taylor, and V. G. Corces. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, 153(6):1281–1295, 2013.

[48] C. Cubeñas-Potts and V. G. Corces. Topologically associating domains: an invariant framework or a dynamic scaffold? *Nucleus*, 6(6):430–434, 2015.

[49] E. Crane, Q. Bian, R. P. McCord, B. R. Lajoie, B. S. Wheeler, E. J. Ralston, S. Uzawa, J. Dekker, and B. J. Meyer. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 523(7559):240–244, 2015.

[50] K. Kruse, C. B. Hug, B. Hernández-Rodríguez, and J. M. Vaquerizas. TADtool: visual parameter identification for TAD-calling algorithms. *Bioinformatics*, 32(20):3190–3192, 2016.

[51] Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, et al. An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009.

[52] H. Akima. A new method of interpolation and smooth curve fitting based on local procedures. *Journal of the ACM (JACM)*, 17(4):589–602, 1970.

[53] J. A. Beagan and J. E. Phillips-Cremins. On the existence and functionality of topologically associating domains. *Nature Genetics*, pages 1–9, 2020.

[54] S. Carstens, M. Nilges, and M. Habeck. Inferential structure determination of chromosomes from single-cell Hi-C data. *PLoS computational biology*, 12(12):e1005292, 2016.

[55] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *science*, 295(5558):1306–1311, 2002.

[56] J. Dekker and T. Misteli. Long-range chromatin interactions. *Cold Spring Harbor perspectives in biology*, 7(10):a019356, 2015.

[57] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[58] S. Kadauke and G. A. Blobel. Chromatin loops in gene regulation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1789(1):17–25, 2009.

[59] C. R. Kaffer, M. Srivastava, K. Y. Park, E. Ives, S. Hsieh, J. Batlle, A. Grinberg, S. P. Huang, and K. Pfeifer. A transcriptional insulator at the imprinted H19/Igf2 locus. *Genes & Development*, 14(15):1908–1919, 2000.

[60] J. Kim, W. D. Frey, K. Sharma, S. Ghimire, R. Teruyama, and L. Stubbs. Allele-specific enhancer interaction at the Peg3 imprinted domain. *PloS one*, 14(10), 2019.

[61] B. R. Lajoie, J. Dekker, and N. Kaplan. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods*, 72:65–75, 2015.

[62] L. A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome research*, 19(1):37–51, 2011.

[63] I. K. Nolis, D. J. McKay, E. Mantouvalou, S. Lomvardas, M. Merika, and D. Thanos. Transcription factors mediate long-range enhancer–promoter interactions. *Proceedings of the National Academy of Sciences*, 106(48):20222–20227, 2009.

[64] J. C. Rivera-Mulia, A. Dimond, D. Vera, C. Trevilla-Garcia, T. Sasaki, J. Zimmerman, C. Dupont, J. Gribnau, P. Fraser, and D. M. Gilbert. Allele-specific control of replication timing and genome organization during development. *Genome research*, 28(6):800–811, 2018.

[65] Q. Szabo, F. Bantignies, and G. Cavalli. Principles of genome folding into topologically associating domains. *Science advances*, 5(4):eaaw1668, 2019.