Deconvolution Of Ensemble Chromatin Interaction Data Reveals The Latent Mixing Structures In Cell Subpopulations

Emre Sefer, Geet Duggal, and Carl Kingsford

Computational Biology Department, School of Computer Science, Carnegie Mellon University
{esefer,geet,carlk}@cs.cmu.edu

Abstract. Chromosome conformation capture (3C) experiments provide a window into the spatial packing of a genome in three dimensions within the cell. This structure has been shown to be highly correlated with gene regulation, cancer mutations, and other genomic functions. However, 3C provides mixed measurements on a population of typically millions of cells, each with a different genome structure due to the fluidity of the genome and differing cell states. Here, we present several algorithms to deconvolve these measured 3C matrices into estimations of the contact matrices for each subpopulation of cells and relative densities of each subpopulation. We formulate the problem as that of choosing matrices and densities that minimize the Frobenius distance between the observed 3C matrix and the weighted sum of the estimated subpopulation matrices. Results on *HeLa* 5C and mouse and bacteria Hi-C data demonstrate the methods' effectiveness. We also show that domain boundaries from deconvolved matrices are often more enriched or depleted for regulatory chromatin markers when compared to boundaries from convolved matrices.

1. Introduction

The spatial organization of the genome as it is packed into the cell is closely linked to its function. Chromatin loops as well as locally clustered topological domains [6] play a role in long-range transcriptional regulation [11, 1] and the progression of cancer [10]. For instance, the impact of the long-range interacting gene clusters in the conformation of HOXA cluster is better understood in the context of the genome's three-dimensional relationships [23]. Loci of mutations that affect expression of genomically far-away genes (eQTLs) are statistically significantly closer in 3D to their regulated genes than expected by a stringent null model [7], indicating that 3D contacts play a widespread role in gene regulation. Measuring and modeling the three-dimensional shape of eukaryotic and prokaryotic genomes is thus essential to obtain a more complete understanding of how genomes function.

A class of recently introduced experimental techniques called chromosome conformation capture (3C) allows for the measurement of pairwise genomic contacts at much higher resolutions than FISH microscopy experiments [5]. These

techniques cross-link spatially close fragments of the genome within a population of millions of cells and use high-throughput sequencing to determine which fragments were cross linked together. Since the development of the original 3C method, a number of enhancements to the protocol such as 3C, 4C, 5C, Hi-C, and TCC, have been introduced [26, 17, 15, 22]. Genome-wide interactions from Hi-C experiments, for example, can be analyzed at fragment lengths as low as 10kb [14], though resolutions of 20-40kb are more common. Here, for simplicity, we refer to all 3C-like techniques as 3C. All of these methods result in a matrix $\mathbf{F} : V \times V \rightarrow \mathbb{R}_0^+$ where $V = \{1, 2, \ldots, n\}$ is the set of genome fragments and where $F_{i,j}$ is the number of times genome fragment i was observed in close proximity to fragment j within the assayed population of cells. Under the assumption that these contact events will be more common for spatially close pairs as shown in [28], the counts can be converted into spatial distances. The count matrix $\mathbf{F}$ or its associated distance matrix are then analyzed in the context of long-range gene regulation or used to produce three dimensional models of the genome [30].

A challenge with 3C data is that it is collected over a population of cells. The genome structures within these cells vary since (1) They exist at different points in time within a particular phase of the cell cycle, (2) They may be associated with different methylation and therefore heterochromatin formations [2], and (3) Chromatin itself can fluidly take on different three-dimensional forms. Analysis of the combined matrix $\mathbf{F}$ therefore may be misleading.

We tackle the problem of extracting the genome contact map of each subpopulation of cells from the combined, ensemble matrix $\mathbf{F}$. A subpopulation represents cells with similar interaction matrices and can model cells in distinct subphases in the cell cycle (e.g. early G1 vs. late G1), cells that are undergoing different gene expression programs, or cells that are in different stochastic structural states. We present a method to deconvolve the observed $\mathbf{F}$ into a collection of biologically-plausible, unobserved subpopulation matrices $\mathbf{F^i}$ such that

$$\mathbf{F} \approx \sum_i \lambda_i \mathbf{F^i}, \tag{1}$$

where λ_i are the relative abundances (densities) of cells in each subpopulation (class) i. This is the *3C Deconvolution Problem* (*3CDE*), which we show to be NP-hard whether λ_i are in $\mathbb{R}$ or $\mathbb{N}$.

To solve this problem, we assume that the interaction matrix $\mathbf{F^i}$ of each class is composed of *nonoverlapping* topological domains that are highly self-interacting consecutive genomic intervals. Such topological domains have been widely observed and are a natural unit of genome structure [6, 3]. We model these domains here using a particular type of quasi-clique, allowing for missing interactions within a densely interacting domain. The algorithm supports the use of prior knowledge of topological domain structure as estimated from the ensemble matrix $\mathbf{F}$ or through other means that inform the choice of domains that appear in each $\mathbf{F^i}$. We explore two variants of our algorithm: one called *3CDEint* in which the class densities λ_i are required to be integers and one called *3CDEfrac* in which they are not. The integer case is appropriate when

the matrix $\mathbf{F}$ contains unnormalized counts, while the real-valued version is appropriate when $\mathbf{F}$ has been normalized to account for experiment bias [32].

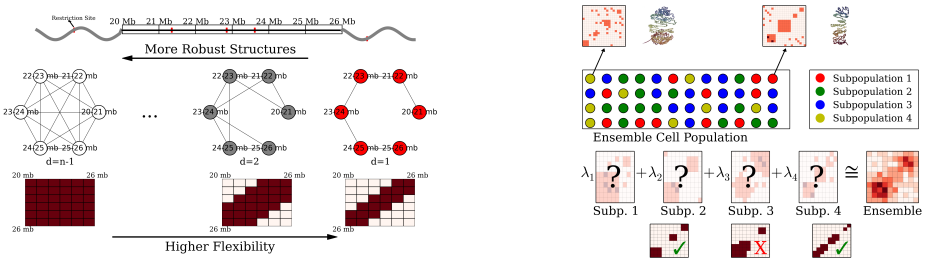
Both *3CDEint* and *3CDEfrac* solve *3CDE* in an iterative two-step fashion that alternates between optimizing the matrices $\mathbf{F^i}$ (Step 1 in Sec. 2.3) and then optimizing the densities λ_i (Step 2 in Sec. 2.4). We show that each step can be solved near-optimally. These two steps use non-monotone supermodular optimization and SDP relaxations, respectively. For smaller problem instances, we develop optimal methods *3CDEint*-opt and *3CDEfrac*-opt based on Quadratic Integer Programming that allow us to compare our approximate solutions of *3CDEint* and *3CDEfrac* to the true optimal solutions.

We show that our estimated deconvoluted matrices and topological domain structures are very similar to those derived from ground truth single cell data [20] as well domain structures in particular cell phases [21]. We also show that domain boundaries from deconvolved matrices are often more enriched or depleted for regulatory chromatin markers H3K4me3, H3K36me3, H3K9me3 and CTCF when compared to boundaries from convolved matrices. The deconvolved domain substructures we produce may therefore be more useful in analyses of long-range regulation with respect to chromatin structure, and our methods can be used as way to simultaneously find domains while determining population substructures.

1.1 Related Work

Most existing methods for finding domains within 3C matrices [9, 6] and for embedding 3C matrices in 3D space [17, 33] treat 3C interaction data as a single unit ignoring the fact that it is an ensemble over millions of cells. Although none of the existing methods explicitly solve the deconvolution problem, some [24, 12, 15, 9] find multiple 3D embeddings or multiple domain decompositions. For example, Rousseau et al. (2001) [24] develop an MCMC sampling technique *MCMC5C*, and Hu et al. (2013) [12] develop *BACHMIX* that optimizes likelihood over a mixture model to find multiple embeddings. Neither of these methods considers the additive affects of interactions. Another method discussed in Kalhor et al. [15] generates a population of structures by restricting the number of times each interaction is involved in a solution, which may mimic the deconvolution to a certain extent but ignores the domain structure of the genome. *Armatus* [9] finds multiple optimal and near-optimal domain decompositions at multiple scales by optimizing a density-like objective. None of these methods determine domain substructures or population densities of these substructures.

On the experimental side, two recent Hi-C modifications try to limit the effect of cell-to-cell variations. Nagano et al. (2013) [20] carry out experiments on single cells that come at a higher experimental cost and produce lower-resolution interaction matrices. Another modification measures the interactions at a particular cell phase by arresting the population of cells at that phase by thymidine and nocodazole. However, these chemicals may disrupt the original genome structure [21, 16]. Since single cell 3C data [20] is so recent, we provide the first comparison of deconvoluted structures to real single cell matrices.



(a) (b)

Fig. 1: (a) *d-bandwidth-quasi-clique* (*d-BQC*). (b) Given the ensemble matrix, we infer the mixing matrices in terms of *BQC*s and the densities λ's without letting *BQC*s overlap in each subpopulation.

1.2 The Deconvolution Problem (*3CDE*)

We want to estimate the interaction matrices $\mathbf{F^i}$ of the subpopulations. Without additional constraints, deconvolution is under-constrained because an infinite number of matrices can explain the ensemble data equally well. However, we can exploit the fact that a 3C interaction matrix is (1) fairly dense around the diagonal due to the abundance of short-range interactions even being sparse overall, and (2) composed of topological domains that are highly self-interacting, non-overlapping genomic intervals that are the building blocks of genome [6, 3].

We encode these assumptions by modeling topological domains as *bandwidth-quasi-cliques* (*BQC*s) to allow domain structures to be locally dense while not requiring all interactions to exist. A *d-BQC* is a defined by a genomic subrange $[s_p, e_p]$ where there is an interaction between every pair of fragments that are separated by at most d fragments, resulting in a banded pattern of interactions. Figure 1a shows a *BQC* for a 6-loci domain at 1 mb resolution. Let $l_{\min}$ and $l_{\max}$ be minimum and maximum possible domain sizes ($l_{\min} \leq e_p - s_p + 1 \leq l_{\max}$). There are $e_p - s_p$ possible *BQC*s for a domain p covering the range $[s_p, e_p]$, so total number of *BQC*s over all domains is $\sum_{l=l_{\min}}^{l_{\max}} (n-l+1)(l-1) = O\left(n(l_{\max} - l_{\min})^2\right)$, where n is the number of fragments.

We assume that the observed ensemble matrix $\mathbf{F}$ is sum of binary interaction matrices ($\{\mathbf{F^1}, \mathbf{F^2}, \ldots, \mathbf{F^k}\}$), each multiplied by their densities ($\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$). We further assume that each $\mathbf{F^i}$ is composed of non-overlapping *BQC*s. Finally, we assume that the number of classes k is given or it can be easily estimated as each subpopulation is a meaningful distinct unit such as different phase of the cell cycle. Let $I = \{1, \ldots, k\}$ be the set of class labels. Figure 1b illustrates *3CDE*, which is defined formally below:

*Problem 1 (*3CDE*).* We are given an ensemble interaction matrix $\mathbf{F}$, a number of classes k, and (optionally) a set of prior domains P_c. For each class i, we want to choose a set of nonoverlapping *bandwidth-quasi-cliques* and density λ_i such that the squared Frobenius norm of the difference between $\mathbf{F}$ and the sum of the matrices $\mathbf{F^i}$ derived from the chosen *bandwidth-quasi-cliques* is minimized.

2. Approximate 3C Deconvolution Methods

2.1 Mathematical Formulation and Hardness

We formulate the *3CDE* problem using a three-part objective that (1) minimizes squared Frobenius norm of the difference between observed convolved matrix and convolution of the deconvolved matrices, (2) maximizes the quality of domains defined by their *BQC*s, and (3) maximizes the overlap with a prior set of candidate domains P_c if available. Formally, given minimum and maximum domain sizes $l_{\min}$ and $l_{\max}$, let $P = \{[s_p, e_p] \mid s_p \in 1, \ldots, n,\, e_p \in s_p + l_{\min} - 1, \ldots, \min(n, s_p + l_{\max} - 1)\}$ be the set of possible domains, and $M : V \to 2^P$ be a function that maps each 3C fragment to the set of domains to which it could belong:

$$M(v) = \{p \mid \forall p = [s_p, e_p] \in P,\, s_p \leq v \leq e_p\}$$

Define $G_q = (V_q, E_q)$ to be the *BQC* intersection graph where

$$V_q = \{(p, d) \mid p \in P,\, d \in 1, \ldots, l_p - 1\} \qquad \text{(the set of possible } BQC\text{s)} \tag{2}$$

$$E_q = \{((p_i, d), (p_j, t)) \mid (p_i, d), (p_j, t) \in V_q^2,\, i \neq j,\, p_i \cap p_j \neq \emptyset\} \tag{3}$$

A pair (p, d) represents a *BQC* by its domain and bandwidth d and l_p is the number of fragments in domain p. We can express *3CDE* as:

$$\min \overbrace{\sum_{(u,v)\in V^2} \left(F_{u,v} - \left(\sum_{i\in I} \lambda_i \Big(\sum_{p\in M(u)\cap M(v)} \sum_{d=|u-v|}^{l_p-1} x_{pdi} \Big) \right) \right)^2}^{\left\| \mathbf{F} - \sum_{i\in I} \lambda_i \mathbf{F^i} \right\|_F^2} + \underbrace{\sum_{i\in I} \sum_{(p,d)\in V_q} w_{pd}(1 - x_{pdi})}_{\text{Domain Weakness}} + \underbrace{\lambda^p \sum_{i\in I} \sum_{p\in P_c} \sum_{d\in 1,\ldots,l_p-1} (1 - x_{pdi})}_{\text{Distance From Prior}} \tag{4}$$

$$\text{s.t.} \quad x_{pdi} + x_{rti} \leq 1, \qquad \forall\, ((p, d), (r, t)) \in E_q,\, \forall i \in I \tag{5}$$

$$x_{pdi} \in \{0, 1\}, \qquad \forall (p, d) \in V_q,\, \forall i \in I \tag{6}$$

where $x_{pdi} = 1$ if *d-BQC* of interval p is assigned to class i. Here, d ranges from $|u - v|$ to $l_p - 1$ for each entry (u, v) since *d-BQC* of p can correspond to matrix entries up to d away from the diagonal. Eqns. (5) ensures each $\mathbf{F^i}$ is made up of nonoverlapping *BQC*s. We penalize for selecting less dense (weaker) *BQC*s where w_{pd} defines the quality of *d-BQC* of p. We also reward larger overlaps with the prior candidate domains P_c from domain finders, such as *Armatus*, by minimizing the distance from the prior domains where λ^p is weight of the prior.

3CDE has two variants depending on the class densities: (1) *3CDEint* where λ_i are integers, and (2) *3CDEfrac* where λ_i can take any nonnegative values (useful for normalized $\mathbf{F}$). *3CDE* is NP-complete whether λ_i are in $\mathbb{R}$ or $\mathbb{N}$ as proven in Theorem 1, and *3CDEint* can be solved exactly in pseudo-polynomial time by dynamic programming as in Theorem 2. However, this approach is impractical, and prohibitively slow for large n, k, and $F_{max} = \max\{F_{i,j}\}$.

Theorem 1. 3CDE *is NP-complete.*

Theorem 2. 3CDEint *can be solved exactly in* $O(kn^{4k-1}F_{max}^k)$ *time.*

2.2 Practical Approximate Methods

Due to hardness of *3CDE*, we design the approximate methods *3CDEfrac* and *3CDEint* for integer and real-valued class densities respectively. Both methods are similar, so we explain *3CDEint* in detail and discuss the differences between *3CDEfrac* from *3CDEint* in the last subsection. Let $S = \{0, 1, \ldots, F_{max}\}$ be the set of integer subpopulation densities where $F_{max} = \max\{F_{i,j}\}$, and we define $y_{is} = 1$ if subpopulation i's density is s. Program (4)–(6) can be expressed as constrained minimization of the biset function $Q(X, Y)$ as in Program (7)–(11):

$$\min Q(X,Y) = \sum_{(u,v)\in V^2} \left(F_{u,v} - \left(\sum_{i\in I} \sum_{s\in S} s y_{is} \Big(\sum_{p\in M(u)\cap M(v)} \sum_{d=|u-v|}^{l_p-1} x_{pdi} \Big) \right) \right)^2 + \sum_{i\in I} \sum_{(p,d)\in V_q} w_{pd}^c (1 - x_{pdi}) \tag{7}$$

$$\text{s.t} \quad x_{pdi} + x_{rti} \leq 1, \quad \forall\, ((p,d),(r,t)) \in E_q,\, \forall i \in I \tag{8}$$

$$\sum_{s\in S} y_{is} = 1, \quad \forall i \in I \tag{9}$$

$$x_{pdi} \in \{0,1\}, \quad \forall (p,d) \in V_q,\, \forall i \in I \tag{10}$$

$$y_{is} \in \{0,1\}, \quad \forall i \in I,\, \forall s \in S \tag{11}$$

where $w_{pd}^c = w_{pd} + \lambda^p$ is the combined domain prior and robustness weight. The nonoverlapping BQC constraints (8) depend only on X, and (9) ensures a single density assignment for each class. We solve Program (7)–(11) iteratively in two steps starting with unit class densities. We describe these two steps with their approximation guarantees in detail below. Intuitively, the first step tries to find the best BQC assignments X given the class densities Y, while the second step tries to find the best Y given X. These steps are iterated until convergence.

2.3 Step 1: Non-monotone Supermodular Optimization for Estimating Mixing Matrices

When the class densities Y are given, (9) disappears, and the objective is slightly modified as in:

$$\min Q(X|Y) = \sum_{(u,v)\in V^2} \left(F_{u,v} - \left(\sum_{i,s\in Y} s \Big(\sum_{p\in M(u)\cap M(v)} \sum_{d=|u-v|}^{l_p-1} x_{pdi} \Big) \right) \right)^2 + \sum_{i\in I} \sum_{(p,d)\in V_q} w_{pd}^c (1 - x_{pdi}) \tag{12}$$

This is *Minimum Non-monotone Supermodular Independent Set in the Interval Graph* defined by the BQC intersection graph G_q since objective (12) is non-monotone supermodular. We solve its fractional relaxation optimally, round the fractional solution via $(1, e^{-1})$-balanced contention resolution scheme by [8] 100 times, and return the minimum solution. This scheme gives $\frac{1}{e} + (1 - \frac{1}{e})\overline{Q}$ approximation guarantee as in Lemma 1 where $\overline{Q} = \frac{Q(\emptyset,\emptyset)+\epsilon}{k \min_{(p,d)}(w^c_{pd})+\epsilon}$ for arbitrarily small constant $\epsilon > 0$. This bound is also preserved up to an additive error for large matrices which weights are usually estimated by sampling [27]. Each rounding step is defined as follows: For each class i, we choose a BQC with probability $1 - e^{-x_{pdi}}$ to put into the solution R. After sampling, we mark the BQC represented by x_{pdi} for deletion if there is a different BQC in R that intersects the starting point of p. After removing all marked BQCs from R, we return independent set R as a solution.

We can achieve similar approximation bound by transforming the program into *Set Cover* where (1) we replace every x_{pdi} with $\hat{x}_{pdi} = 1 - x_{pdi}$, (2) define a variable for each quadratic term, and (3) introduce extra covering constraints to enforce the quadratic costs when none of its linear terms are added. This *Set Cover* can be solved by greedy method which runs faster for large matrices.

Lemma 1. *Step 1 can be approximated to a factor* $\frac{1}{e} + (1 - \frac{1}{e})\overline{Q}$.

2.4 Step 2: SDP Relaxation of Binary Least Squares for Density Assignment

Given BQC assignments X, (8) disappears, and the resulting program is a binary quadratic program under the assignment constraints (9). However, the size of this program is linear in terms of F_{max} which may be arbitrarily large. To efficiently estimate the class densities, we express the program more compactly by defining a variable for every $s \in S' = \{2^d \mid d \in 0, 1, \ldots, \lfloor \log(F_{max}) \rfloor\}$. This modification also removes (9) without losing any expressiveness since we can express any density up to F_{max} as a sum of subset of S'. The resulting problem is:

$$\min Q(Y|X) = \sum_{(u,v)\in V^2} \left(F_{u,v} - \left(\sum_{i\in I} m_{ui}\, m_{vi} \sum_{s\in S'} s y_{is} \right) \right)^2 + \text{Constant} \quad (13)$$

$$= \sum_{i\in I, s\in S'} \sum_{j\in I, t\in S'} st \Big(\sum_{(u,v)\in V^2} m_{ui} m_{vj} \Big) y_{is} y_{jt} - 2 \sum_{i\in I, s\in S'} s \Big(\sum_{(u,v)\in V^2} F_{u,v} m_{ui} m_{vi} \Big) y_{is}$$

where binary $y_{is} = 1$ if s is part of class i's density, m_{ui} is an indicator for whether u is assigned to a BQC in class i that is known from given X, and $\sum_{s\in S'} s y_{is}$ is the density of class i. Optimizing (13) is NP-hard via reduction from *PARTITION* [31]. To solve it efficiently, we turn our $\{0, 1\}$ quadratic program into homogenous $\{\pm 1\}$ quadratic program by replacing every y_{is} with $(1 + y'_{is})/2$ where $y'_{is} \in \{\pm 1\}$, and then by substituting $y'_{is} = r y''_{is}$ where $r \in \{\pm 1\}$. The

resulting boolean program can be rewritten as:

$$\min_{Y} \quad \mathbf{y''TAy''} - 2\mathbf{b^T}r\mathbf{y''} + \|\mathbf{b}\|^2 \tag{14}$$

$$\text{s.t.} \quad {y''}^2_{is} = 1, \quad i \in 1, \ldots, k\,, s \in S' \tag{15}$$

$$r^2 = 1 \tag{16}$$

where $\mathbf{A}$ is the matrix of quadratic coefficients in (13) modified by the transformation above, $\mathbf{b}$ is the modified vector of linear coefficients in (13), and $\mathbf{y''}$ is a $k|S'|$ length vector. We relax this quadratically constrained quadratic program into the following semidefinite program (SDP):

$$\mathbf{Y}^* = \arg\min_{\mathbf{Y''}} Tr(\hat{\mathbf{A}}\mathbf{Y''}) \tag{17}$$

$$\text{s.t} \quad {Y''}^2_{t,t} = 1, \quad t \in 1, \ldots, k|S^{'}| + 1 \tag{18}$$

$$\mathbf{Y''} \succeq 0 \tag{19}$$

where $\mathbf{Y''} = [\mathbf{y}^{\mathbf{T}^{''}}, r]^T\,[\mathbf{y''}, r]$ is positive-semidefinite matrix, and $\hat{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & -\mathbf{b} \\ -\mathbf{b^T} & \|\mathbf{b}\|^2 \end{bmatrix}$. After solving this SDP optimally, we run the following rounding procedure based on Gaussian sampling [18]: We generate a set of random vectors ξ_l, $l \in 1, \ldots, L = 100$ from multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{Y}^*)$, quantize each of them into a binary vector $\hat{y}_l = \text{sign}(\xi_l) \in \{\pm 1\}^{k|S'|+1}$, and obtain a solution by $\hat{y} = \min_{l \in 1,\ldots,L} \hat{y}_l^T \mathbf{A} \hat{y}_l$. This procedure gives $\frac{2}{\pi} + (1 - \frac{2}{\pi})\overline{Q}$ approximation guarantee for Step 2 as proven in Lemma 2.

Lemma 2. *Step 2 can be approximated to a factor* $\frac{2}{\pi} + (1 - \frac{2}{\pi})\overline{Q}$.

2.5 The case of real-valued densities: *3CDEfrac*

We modify only Step 2 of *3CDEint* for nonnegative, real-valued class densities. Let y_i be the variable for class i's density, *3CDEfrac*'s second step optimally solves the following convex quadratic program:

$$\min_{Y} \sum_{i \in I} \sum_{j \in I} \Big(\sum_{(u,v) \in V^2} m_{ui} m_{vj} \Big) y_i y_j \; - 2 \sum_{i \in I} \Big(\sum_{(u,v) \in V^2} F_{uv} m_{ui} m_{vi} \Big) y_i \tag{20}$$

$$y_i \geq 0, \quad i \in I \tag{21}$$

3. Exact 3C Deconvolution Methods

For smaller problem instances, we develop optimal methods *3CDEint*-opt and *3CDEfrac*-opt based on convex Quadratic Integer Programming (QIP). *3CDEint*-

opt can be expressed as in Program (22)–(27):

$$\min \sum_{(u,v)\in V^2} \left(F_{u,v} - \left(\sum_{p\in M(u)\cap M(v)} \sum_{d=|u-v|}^{l_p-1} \sum_{i\in I} y_{pdi} \right) \right)^2 - \sum_{i\in I} \sum_{(p,d)\in V_q} w^c_{pd} x_{pdi} \tag{22}$$

$$\text{s.t } x_{pdi} + x_{rti} \leq 1, \quad \forall\, ((p,d),(r,t)) \in E_q, \forall i \in I \tag{23}$$

$$y_{pdi} \leq F_{max}\, x_{pdi}, \quad \forall (p,d) \in V_q, \forall i \in I \tag{24}$$

$$|y_{pdi} - y_{rti}| \leq F_{max}(2 - x_{pdi} - x_{rti}), \quad \forall\, ((p,d),(r,t)) \notin E_q, \forall i \in I \tag{25}$$

$$x_{pdi} \in \{0,1\}, \quad \forall (p,d) \in V_q, \forall i \in I \tag{26}$$

$$y_{pdi} \in \{0,1,\ldots,F_{max}\}, \quad \forall (p,d) \in V_q, \forall i \in I \tag{27}$$

where binary $x_{pdi} = 1$ if *d-BQC* of domain p is assigned to class i, and integer y_{pdi} is its density. Objective (22) is convex as shown previously, and overlapping *BQC*s cannot coexist in the same class according to (23). (24) ensures that density of *d-BQC* of domain p in class i is 0 if not used in i, and if assigned, its density is at most F_{max}. Lastly, (25) ensures that all *BQC*s of the same class have the same density. When the class densities are real-valued, we propose *3CDEfrac*-opt by relaxing the integer density constraints (27) in Program (22)–(27) which turns it into convex Mixed Integer Quadratic Program (MIQP).

4. Results

4.1 Implementation

We implemented our methods using CPLEX [13] to solve LP, ILP and convex quadratic programs, and SDPT3 [29] to solve SDP relaxations. We use the public implementations of *Armatus* [9] and *MCMC5C* [24] for comparison, and implemented the 3C normalization method by [32]. Code, datasets and theorem proofs can be found at `http://www.cs.cmu.edu/~ckingsf/research/3cde`. The approximate methods are reasonably fast: *3CDEint* and *3CDEfrac* can deconvolve *CD4*$^+$ interaction matrices in less than 15 minutes on a laptop with 2.5Ghz processor and 8Gb Ram when $l_{\max} = 25$. They typically converge in fewer than 5 iterations. Our methods can also deconvolve larger 20-40 kbp resolution matrices under 30 minutes by restricting $l_{\max} = 50$ as topological domains are usually megabase-sized.

4.2 Evaluating Performance

We evaluate deconvolution methods in the few cases where small, synchronized populations were assayed with 3C methods. Nagano et al. [20] performed Hi-C on 10 single mouse cells, Naumova et al. [21] performed Hi-C on several populations HeLa cells, each synchronized to a specific phase of the cell cycle, and Le et al. [16] performed Hi-C on populations of *Caulobacter* cells, also synchronized to

various phases of the cell cycle. In each of these experiments, we have more-than-usual confidence that the assayed cells represent a single, unmixed population of structures. To simulate a more typical population of cells with mixture, we sum together the individual matrices from each of these experiments to obtain a synthetic ensemble matrix **F** that we then attempt to deconvolve into its constituent components (the matrices from the single cell or synchronized experiments).

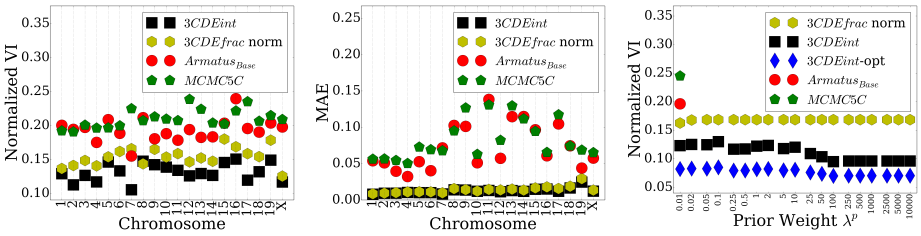
We measure the agreement between our estimated subpopulation contact matrices and the true contact matrices (single cell or synchronized cell cycle) using two metrics: the normalized mean absolute error (MAE) and the normalized Variation of Information (NVI) [19]. Let $T_p = \{T_p^1, \ldots, T_p^k\}$ and $E_p = \{E_p^1, \ldots, E_p^k\}$ be the set of true and estimated domain partitions respectively, and **T** and **E** be the set of associated interaction matrices. To estimate either metric (MAE or VI), we perform a minimum-weight bipartite perfect matching between **T** and **E** where the edges are weighted by the value of the metric (VI or MAE) and the value of the agreement between **T** and **E** is the average value of the minimum perfect matching. In the case of VI, this metric measures agreement between clusterings (here partitions of fragments into domains and non-domains). Since the true domain partitions are unknown, we use the consensus *Armatus* domains computed on each known subpopulation as the truth. In both measures, lower score means better performance.

We compare our methods with greedy baseline $Armatus_{Base}$ and *MCMC5C* [24]. In $Armatus_{Base}$, we add the domains from the top-k *Armatus* decompositions into a set. For each class, we shuffle the set, and iterate through half of the set by assigning a domain from this set unless it intersects with the currently-assigned domains. We repeat this procedure 10000 times to estimate the distribution of the scores. Using domains from *Armatus* equips $Armatus_{Base}$ with domains that appear in the convoluted data set, and it is therefore a more conservative comparison to our methods. We present the mean $Armatus_{Base}$ score, and estimate P-values of our results from this distribution to test for the significance. We also estimate the matrices of k embeddings via inverse frequency-distance mapping in *MCMC5C*. When estimating the marker distribution, we define a domain boundary as a region extended to left and right of the exact boundary by half of the resolution since this reflects the uncertainty in its position due to binning. Unless otherwise noted, we use an exponential kernel for *BQC* quality, and assume no prior domain knowledge.

4.3 Deconvolution of Single Mouse $CD4^+$ Interaction Matrices

We apply our method and the baseline methods to the $CD4^+$ interaction dataset at 250 kbp resolution by providing them with the sum of the matrices from the 10 experiments in which 3C contacts were estimated on single mouse $CD4^+$ cells. We compare the estimated subpopulation matrices using this summed matrix as input to the original single cell matrices. Performance is shown in Figures 2a–2b.

3CDEint and *3CDEfrac* nearly always perform the best in identifying contact matrices that match the single cell matrices. Even though $Armatus_{Base}$ greedily assigns domains to the classes, mean $Armatus_{Base}$ performs better than



(a) Chromosome-wise NVI (b) Chromosome-wise MAE (c) Chro. 17 NVI

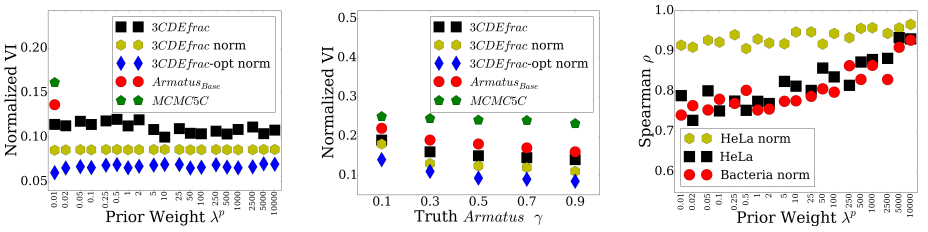
Fig. 2: Chromosome-wise deconvolution performance of $CD4^+$ dataset in terms of (a) Normalized VI, (b) Mean Absolute Error (MAE). (c) Performance on the 17th chromosome for various prior weights λ^p.

MCMC5C in Figure 2a for most of the chromosomes. *3CDEfrac* over normalized data [32] may perform worse than $Armatus_{Base}$ because $CD4^+$ data is an ensemble over only 10 cells rather than millions of cells as in traditional 3C experiments. We observe similar performance trend in terms of the metric MAE as in Figure 2b. Normalization does not decrease the performance as it did for normalized VI in Figure 2a. *3CDEint* performs significantly better than $Armatus_{Base}$ on all chromosomes ($p < 0.05$) in terms of both metrics since variance of the distribution of $Armatus_{Base}$ scores is low even though the mean scores are close to ours. In general, lower matrix error scores show the quality of the deconvolution in estimating the mixing matrices.

We examine the performance of chromosome 17 as the domain prior weight λ is increased (Figure 2c). The prior weight seems to have little effect on the overall performance, though *3CDEfrac* over normalized data is more robust to different prior weights. Chromosome 17 is small enough that we can use *3CDEint*-opt to find the true optimum of our objective (blue diamonds in Figure 2c). This shows that our heuristics are achieving close to the optimum value.

4.4 Temporal Deconvolution of Interphase Populations in *HeLa* and *Caulobacter* Cells

We deconvolve the sum of measured matrices of the 21st chromosome of *HeLa* cells at 250 kbp resolution using data from Naumova et al. [21]. Here, each subpopulation represents cells at a particular phase of the cell cycle, and so we are deconvolving along the temporal dimension. Figure 3a shows the performance for several choices of prior. Again, we match the true matrices better than either a greedy approach or sampling approach (*MCMC5C*). All the methods perform better in *HeLa* cells than $CD4^+$ cells as shown in Figure 2c. Unlike in $CD4^+$, normalization improves the deconvolution performance as well as making the performance of both approximate *3CDEfrac* and exact *3CDEfrac*-opt less dependent on the prior weight. This performance stability shows that we may obtain true domain decompositions without strong reliance on prior data.



(a) *HeLa* (b) *Caulobacter Crescentus* (c) Densities of The Phases

Fig. 3: (a) Deconvolution performance on *HeLa* dataset by increasing prior weight λ^p in terms of NVI. (b) Performance on prokaryotic bacteria dataset vs. *Armatus* γ in terms of NVI. (c) Performance of *3CDEfrac* in estimating the densities of the cell cycle phases on eukaryotic *HeLa* and prokaryotic *Caulobacter* datasets in terms of Spearman's correlation ρ by increasing λ^p.

3CDEfrac and *3CDEfrac*-opt also outperform the competing methods in terms of average error per matrix entry: *3CDEfrac* without a domain prior can achieve MAE of 0.004, whereas *MCMC5C* achieves almost 8-fold more MAE, 0.03.

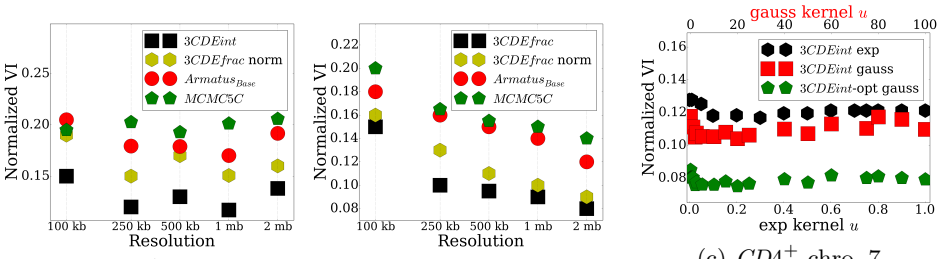
We performed a similar experiment for the bacterium *Caulobacter* where Le et al. (2013) provide cell-cycle-phase-specific Hi-C matrices. Figure 3b reports these results using the NVI metric as the resolution of the ground truth domains was varied. While ground truth matrices are known in these experiments, the true domain decomposition is estimated computationally via a topological domain finder *Armatus*. This program has a parameter γ that controls the domain sizes, with larger γ corresponding to smaller domains. As γ increases, all methods perform better, however, the ranking of the methods in terms of performance is same regardless of γ. We observe similar performance trend on *HeLa* dataset as well. This shows both that we can deconvolve bacterial Hi-C experiments and that the performance is robust to the scale at which we define the true domains.

Our methods also estimate the densities of the mixing cell cycle phases quite accurately on *HeLa* and *Caulobacter* if densities of the 4 cell cycle phases (early G1, mid G1, S, M) are assumed to be proportional to their durations. Figure 3c plots the Spearman's ρ correlation between estimated and true densities at 250 kbp for both datasets. We often achieve correlations over 0.75. Existing methods do not provide any estimate of the densities of the subpopulations.

4.5 Effect of Resolution and Robustness Prior

The deconvolution methods developed here work well at various 3C resolutions. When we binned the input 3C matrices at increasing intervals, increasing the resolution leads to larger, more detailed interaction matrices, which usually decreases performance somewhat (Figure (4a)–(4b)). The performance decreases monotonically on *HeLa* dataset by increasing resolution, but the score trend is non-monotonic in $CD4^+$ cells due to its smaller population size with more influ-

ential outliers. However, the *3CDEfrac* and *3CDEint* methods still outperform the other methods. This is likely due in part to the definition of *BQC*s, which can properly model long-range, out-of-domain interactions in the higher resolution matrices. The choice of the kernel for the robustness prior also seems to have relatively little effect on performance as shown in Figure (4c) or the 7th $CD4^+$ chromosome. We obtain similar results for 21st *HeLa* chromosome.



(a) $CD4^+$ chro. 4 (b) *HeLa*

(A) 3C Resolution

(c) $CD4^+$ chro. 7

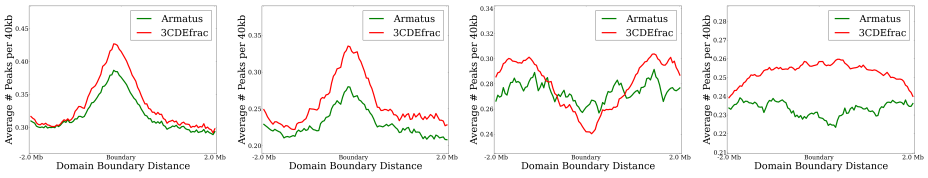
(B) Robustness Prior

Fig. 4: Effect of 3C resolution on the performance in (a) 4th $CD4^+$ chromosome, (b) *HeLa* cells, and the effect of weighting kernel of the robustness prior in (c) $CD4^+$ chromosome 7.

4.6 Distribution of Epigenetic Markers Relative To Deconvolved Domains

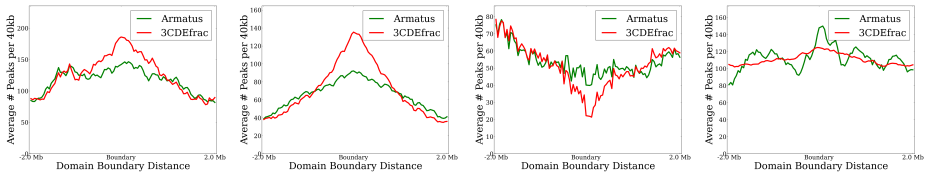
Epigenetic markers are distributed differently in the genome depending on its conformation, and domain organization of the genome is correlated to a certain extent with their distribution. For instance, H3K4me3 and CTCF binding sites are enriched in the domain boundaries due to their insulator roles. We calculate the distribution of several such markers near the domain boundaries as identified within the subpopulation matrices (Figure 5). Each subfigure in Figure 5 plots the average number of markers in 40 kb bins for $+/-$ 2 Mb from all the estimated domain boundaries that occur within some estimated subpopulation matrix. For *Armatus* domain, we estimate the average number of markers over top-k decompositions for multiple γ between 0.1 and 0.9 ($k = 4$ for *HeLa*, and $k = 10$ for $CD4^+$). We obtain histone markers from ChIP-Seq experiments [25, 4] for $CD4^+$ cells, from [2] for *HeLa* cells, and add CTCF sites from CTCFBSDB [34].

Overall, the relationship between histone markers and our domain boundaries are consistent with the experimentally-characterized different roles of the epigenetic markers [2]. Barrier-like histones H3K4me3, H3K27ac, and CTCF are more enriched in the deconvolved domain boundaries than *Armatus* boundaries in both species, whereas non-promoter-associated represssor H3K9me3 is more



(a) H4K4me3 (b) H3K27ac (c) H3K9me3 (d) H3K4me1

(A) *CD4*$^+$ cells



(e) H3K4me3 (f) CTCF (g) H3K9me3 (h) H3K4me1

(B) *HeLa* cells

Fig. 5: Distribution of several markers around the domain boundaries in A) *CD4*$^+$, (B) *HeLa* cells. Red and green lines represent *3CDEfrac* and *Armatus* respectively in all plots.

depleted in the deconvolved domain boundaries. This greater enrichment and depletion of the histones near the deconvolved domain boundaries, in accordance with the experimental results, show the improvement in extracting biologically-plausible domains from the ensemble data achieved by deconvolution.

5. Conclusion

We formulate the novel 3C deconvolution problem to estimate classes of contact matrices and their densities in the ensemble chromatin interaction data. We prove its hardness and design optimal and near-optimal methods that are practical on real data. Experimental results on mouse, *HeLa*, and bacteria datasets demonstrate that our methods outperform related methods in unmixing convoluted interaction matrices of prokaryotes and eukaryotes as well as in estimating the mixing densities without any biological prior. Our methods solve the previously unsolved problem of 3C experiments efficiently, and they return biologically meaningful domains supporting their alternative use as domain finders.

Acknowledgements. This research is funded in part by the Gordon and Betty Moore Foundations Data-Driven Discovery Initiative through Grant GBMF4554 to Carl Kingsford. It is partially funded by the US National Science Foundation (CCF-1256087, CCF-1319998) and the US National Institutes of Health (R21HG006913, R01HG007104). C.K. received support as an Alfred P. Sloan Research Fellow.

References

1. Ay, F., Bunnik, E.M., Varoquaux, N., Bol, S.M., Prudhomme, J., Vert, J.P., Noble, W.S., Le Roch, K.G.: Three-dimensional modeling of the p. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Research 24(6), 974–988 (2014)
2. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., Zhao, K.: High-resolution profiling of histone methylations in the human genome. Cell 129(4), 823 – 837 (2007)
3. Bickmore, W.A., van Steensel, B.: Genome architecture: domain organization of interphase chromosomes. Cell 152(6), 1270–1284 (2013)
4. Deaton, A.M., Webb, S., Kerr, A.R., Illingworth, R.S., Guy, J., Andrews, R., Bird, A.: Cell type-specific DNA methylation at intragenic CpG islands in the immune system. Genome Research 21(7), 1074–1086 (2011)
5. Dekker, J., Marti-Renom, M.A., Mirny, L.A.: Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews Genetics 14(6), 390–403 (2013)
6. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485(7398), 376–380 (2012)
7. Duggal, G., Wang, H., Kingsford, C.: Higher-order chromatin domains link eQTLs with the expression of far-away genes. Nucleic Acids Research 42(1), 87–96 (2014)
8. Feldman, M., Naor, J., Schwartz, R.: A unified continuous greedy algorithm for submodular maximization. In: Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on. pp. 570–579. IEEE (2011)
9. Filippova, D., Patro, R., Duggal, G., Kingsford, C.: Identification of alternative topological domains in chromatin. Algorithms for Molecular Biology 9(1), 14 (2014)
10. Fudenberg, G., Getz, G., Meyerson, M., Mirny, L.A.: High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. Nature biotechnology 29(12), 1109–1113 (2011)
11. Gorkin, D., Leung, D., Ren, B.: The 3d genome in transcriptional regulation and pluripotency. Cell Stem Cell 14(6), 762 – 775 (2014)
12. Hu, M., Deng, K., Qin, Z., Dixon, J., Selvaraj, S., Fang, J., Ren, B., Liu, J.S.: Bayesian inference of spatial organizations of chromosomes. PLoS computational biology 9(1), e1002893 (2013)
13. ILOG, Inc: ILOG CPLEX: High-performance software for mathematical programming and optimization (2006), see `http://www.ilog.com/products/cplex/`
14. Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A., Ren, B.: A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503(7475), 290–294 (2013)
15. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., Chen, L.: Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nature Biotechnology 30(1), 90–98 (2012)
16. Le, T.B.K., Imakaev, M.V., Mirny, L.A., Laub, M.T.: High-resolution mapping of the spatial organization of a bacterial chromosome. Science 342(6159), 731–734 (2013)
17. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al.: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science 326(5950), 289–293 (2009)

18. Luo, Z.Q., Ma, W.K., So, A.C., Ye, Y., Zhang, S.: Semidefinite relaxation of quadratic optimization problems. Signal Processing Magazine, IEEE 27(3), 20–34 (2010)
19. Meilă, M.: Comparing clusterings—an information based distance. J. Multivar. Anal. 98(5), 873–895 (2007)
20. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., Fraser, P.: Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502(7469), 59–64 (2013)
21. Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B.R., Mirny, L.A., Dekker, J.: Organization of the mitotic chromosome. Science 342(6161), 948–953 (2013)
22. Noble, W.S., jun Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A.: A three-dimensional model of the yeast genome. Nature 465(7296), 363–367 (2010)
23. Rousseau, M., Crutchley, J.L., Miura, H., Suderman, M., Blanchette, M., Dostie, J.: Hox in motion: tracking HoxA cluster conformation during differentiation. Nucleic Acids Research 42(3), 1524–1540 (2014)
24. Rousseau, M., Fraser, J., Ferraiuolo, M., Dostie, J., Blanchette, M.: Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling. BMC Bioinformatics 12(1), 1–16 (2011)
25. Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., Ren, B.: A map of the cis-regulatory sequences in the mouse genome. Nature 488(7409), 116–120 (2012)
26. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., de Laat, W.: Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nature Genetics 38(11), 1348–1354 (2006)
27. Svitkina, Z., Fleischer, L.: Submodular approximation: Sampling-based algorithms and lower bounds. SIAM Journal on Computing 40(6), 1715–1737 (2011)
28. Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J.R., Wickramasinghe, P., Lee, M., Fu, Z., Noma, K.i.: Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Research 38(22), 8164–8177 (2010)
29. Tütüncü, R.H., Toh, K.C., Todd, M.J.: Solving semidefinite-quadratic-linear programs using SDPT3. Mathematical programming 95(2), 189–217 (2003)
30. Varoquaux, N., Ay, F., Noble, W.S., Vert, J.P.: A statistical approach for inferring the 3d structure of the genome. Bioinformatics 30(12), i26–i33 (2014)
31. Verdú, S.: Computational complexity of optimum multiuser detection. Algorithmica 4(1-4), 303–312 (1989)
32. Yaffe, E., Tanay, A.: Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature Genetics 43(11), 1059–1065 (2011)
33. Zhang, Z., Li, G., Toh, K.C., Sung, W.K.: Inference of spatial organizations of chromosomes using semi-definite embedding approach and hi-c data. In: Research in Computational Molecular Biology. pp. 317–332. Springer (2013)
34. Ziebarth, J.D., Bhattacharya, A., Cui, Y.: CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. Nucleic Acids Research 41(D1), D188–D194 (2013)