A binary segmentation method for detecting topological domains in Hi-C data

N. Raveendran [a] **and G. Sofronov** [a]

[a] *School of Mathematical and Physical Sciences, Macquarie University, Sydney, Australia*
Email: nishanthi.raveendran@mq.edu.au

Abstract: The three-dimensional (3D) architecture of chromosomes in nuclear space plays an important role in studying gene expression and regulation in cell biology. In particular, chromosome conformation capture (3C) techniques are used to study the spatial structure of chromosomes. Many such methods have been developed in the last two decades. Among them, Hi-C is a technology that uses a deep sequencing approach to detect the 3D spatial organization of a genome. That is, Hi-C allows us to evaluate spatial proximity between any pair of loci along the genome. This results in an interaction matrix with the frequency of interactions between genomic loci that physically interact in the nucleus. Highly self-interacting regions appear in the contact map of such interaction matrices. These regions are called topological domains and they play an important role in regulating gene expression and other genomic functions. Thus detecting such topological domains will provide new insights on chromosomal conformation in better understanding of cell functioning and various diseases.

The topological domains centered along a diagonal region in contact maps are more likely to exist and prominent in data. In this study, we focus on detecting such domains, and we approach this problem as a two-dimensional segmentation problem. To solve this segmentation problem, we propose an algorithm based on the binary segmentation method, a well-known recursive partitioning technique used in change point detection problems. Our numerical experiments illustrate the usefulness of this approach. We obtain estimates for the number of diagonal blocks and their boundaries in an artificially generated data matrix and compare the results of these estimates to those obtained with the HiCSeg R package. We conclude that binary segmentation method works well in identifying such domains with easy implementation and a low computational cost.

Keywords: *Hi-C data, two-dimensional segmentation, binary segmentation method*

1 INTRODUCTION

The three-dimensional organization of chromosomes in the nucleus plays an important role in studying gene expression and regulation features (Dixon et al. (2012)). In cell biology, chromosome conformation capture (3C) is used to study the spatial structure of chromosomes. Several chromosome conformation capture technologies have been developed in past decades. Among them, Hi-C is a technology which uses a deep sequencing approach to detect the 3D spatial organization of whole genomes. Hi-C provides the frequency of interactions between genomic loci that physically interact in the nucleus in the format of an interaction matrix (Lieberman-Aiden et al. (2009)). There are strong patterns of regions with high interaction frequency in the contact map of the interaction matrix which are called topological domains (Lévy-Leduc et al. (2014)). Since these topological domains are highly related to cell functioning, identifying these regions with Hi-C matrix became an important topic. In the Hi-C maps, higher intensity occurs due to both cis- and trans-interacting regions (Fraser et al. (2009)). Typically, cis-interacting regions are more likely to exist and prominent in the data. These higher-intensity regions appear as square blocks centered along a diagonal in contact maps, for example, see Figure 1. Thus, most existing methods to identify topological domains focus on the diagonal region than the off-diagonal region.

A variety of methods have been developed in the literature to identify topological domains. The first method was presented in Dixon et al. (2012), where two-dimensional data was converted into a one-dimensional index, called the directionality index (DI). Then a regular Hidden Markov Model (HMM) is used to detect the change of frequency between the upstream and downstream regions of topological domains. A method based on the distance-scaling factor is presented in Sexton et al. (2012) to detect the interacting regions across the diagonal of the contact map, where the boundaries are determined using the maximum value in distance-scaling factors. In Filippova et al. (2014), the authors presented a method based on dynamic programming, called “Armatus”, where they used a fixed resolution to find the optimal and near-optimal solutions and then aggregated it with different resolutions. Lévy-Leduc et al. (2014) also proposed a dynamic programming method called “HiCseg” to estimate diagonal domains and their boundaries for both raw and normalized Hi-C data within a maximum likelihood framework. Zhou (2017) presented an algorithm based on the binary segmentation method to identify topological domains in both diagonal and off-diagonal regions with a hierarchical overlapping structure. The algorithm initially generates a large number of raw blocks as potential domains. These raw blocks were then filtered using a simple filter step to obtain a subset, and the final domains were estimated using optimization and cross-validation techniques. However, the computational cost of the algorithm was found to be high.

In this paper, we approach this problem as a two-dimensional (2D) spatial segmentation problem where detecting domains of diagonal blocks in a symmetric matrix can be seen as a particular 2D segmentation task. Spatial segmentation algorithms are widely used in many areas, including spatial epidemiology (Gangnon & Clayton (2000)), ecology (see, for example, Beckage et al. (2007), López et al. (2010), Raveendran & Sofronov (2017)), climatology (Tripathi & Govindaraju (2009)) and economic applications (Arbia et al. (2008), Cai et al. (2016)). For example, similar spatial segmentation problems were recently studied by Raveendran & Sofronov (2019, 2021) to identify homogeneous spatial domains in lattice data. To solve this two-dimensional segmentation problem, we propose to use the multiple change point detection methodology, which is commonly used to detect change points and their locations in linear data arising in a wide range of applications such as genomics, economics, climatology, and bioinformatics (Evans et al. (2011), Polushina & Sofronov (2011, 2013, 2016), Priyadarshana & Sofronov (2012, 2014), Sofronov et al. (2009)). Several methods exist to detect such change points, ranging from exact approaches such as segment neighbourhood search, optimal partitioning and pruned exact linear time, to approximate approaches such as binary segmentation methods. For a general overview of change point methods and their applications, see Chen & Gupta (2012), Killick & Eckley (2014).

In this study, to detect non-overlapping diagonal domains in Hi-C data, we propose an algorithm based on the binary segmentation method, a well-studied and widely cited method used within the change point literature. This method originates from the work of Scott & Knott (1974) and Sen & Srivastava (1975) and it was further studied and proposed as circular binary segmentation for analysing the DNA copy number data by Olshen et al. (2004) and as a wild binary segmentation approach by Fryzlewicz (2014). This method is also used in two-dimensional segmentation problems, for example, see Raveendran & Sofronov (2017) in which the binary segmentation method was used to segment spatial binary data observed over a two-dimensional lattice. This method offers several benefits, including simplicity and ease of implementation, while also resulting in significant computational cost savings.

The rest of the paper is organized as follows. In Section 2, we define a statistical model for normalized Hi-C data. We provide a detailed description of the proposed method in Section 3. The results of our analysis are given in Section 4. We conclude the paper with brief remarks in Section 5.

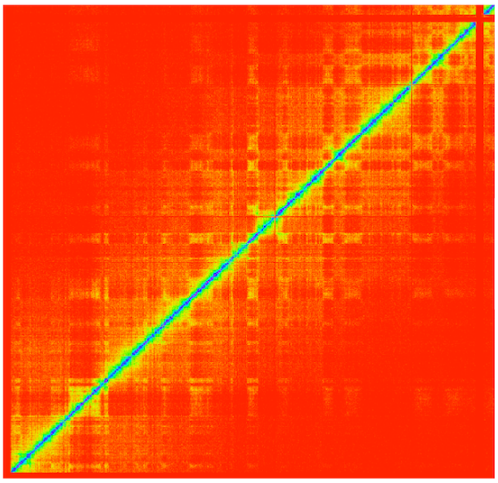

Figure 1. An example of Hi-C contact map. Image from Zhou (2017).

2 STATISTICAL MODEL

Here we define our statistical model. We assume that the interaction matrix X with dimensions $n \times n$ is formed by K potential non-overlapping diagonal domains, which have rectangular shapes, X is a symmetric matrix ($X' = X$) in which $X_{i,j}, 1 \leq i \leq j \leq n$, represents an intensity of the interactions between positions i and j in the nuclear space. Typically, both raw and normalised Hi-C interaction matrices are being used in the literature. Normalized Hi-C data follow a Normal distribution where raw data, which is often count data, can be modelled by Poisson or negative binomial distributions (Lévy-Leduc et al. (2014)). In this study, we assume that all intensities are independent random variables following a Normal distribution. Then X can be written as

$$X_{i,j} \sim \mathsf{N}(\cdot; \mu_{i,j}, \sigma^2),$$

where the matrix of means $\mu_{i,j}$, $1 \leq i \leq j \leq n$, represents the diagonal matrix and σ^2 is assumed to be a constant and does not depend on the values of i or j. In this study, we utilize a block-diagonal model which assumes that the observations are realizations of random variables having their mean which changes along the diagonal blocks but which is constant outside the blocks. Figure 2 shows a simple illustration of the proposed model.

Let $D_k, k = 1, 2, \ldots, K$, indicate a set of all block diagonal domains to be identified with $t_0, t_1, \ldots, t_K$ being the boundaries of the block domains, $t_0 = 1$ and $t_K = n + 1$ (see Figure 2). The area outside of all these block diagonal domains is denoted by H. The k-th domain D_k and the area H can be written as follows

$$D_k = \{(i, j) : t_{k-1} \leq i \leq j \leq t_k - 1\},$$

$$H = \{(i, j) : 1 \leq i \leq j \leq n\} \cap (\cup D_k)^c.$$

The parameters μ_{ij} for the entire data can then be written as

$$\mu_{i,j} = \begin{cases} \mu_k & \text{if } (i,j) \in D_k, k = 1, 2, \ldots, K, \\ \mu' & \text{if } (i,j) \in H. \end{cases}$$

The objective is to estimate the number of diagonal domains K and their boundaries $(t_k)_{0 \leq k \leq K}$. Once we could obtain block boundaries in the case where the number of domains is known, estimating an optimal number of domains K^* would be considered a model selection problem. Here we use the maximum likelihood approach. For a set of domains D_k with their boundaries $(t_k)_{0\le k\le K}$ and parameters $(\mu_k)_{1\le k\le K}$ and μ', the log-likelihood of the data satisfying above assumptions can be written as

$$\begin{aligned} l(X) &= \sum_{1\le i\le j\le n} \log \mathsf{N}(X_{i,j};\mu_{ij},\sigma^2) \\ &= \sum_{k=1}^{K}\sum_{(i,j)\in D_k} \log \mathsf{N}(X_{i,j};\mu_k,\sigma^2) + \sum_{(i,j)\in H} \log \mathsf{N}(X_{i,j};\mu',\sigma^2), \end{aligned}$$

where D_k and H represent diagonal domains and outside area, respectively. A similar model was studied in the context of a dynamic programming framework; for more details, see Lévy-Leduc et al. (2014).

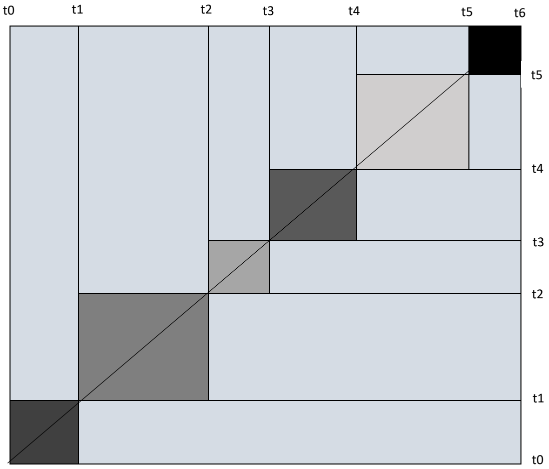

Figure 2. An example of a matrix with $K = 6$ block diagonal domains.

3 METHODOLOGY

In this section, we describe our proposed algorithm to estimate the number of domains and their boundaries. We approach our problem as a change point detection problem which is commonly used in statistics to detect abrupt changes and their locations. In this study, we propose a binary segmentation algorithm which is a well-known recursive partitioning tool used in change point detection literature and it leads to simple solutions for such problems. It was first introduced by Scott & Knott (1974) in the context of cluster analysis. The method starts with searching a single change point for entire data through the cumulative sum (CUSUM) procedure or the likelihood ratio test. If a change point is detected, the segment is then split into two subsegments. The same procedure is then continued on subsegments until a certain stopping criterion is met.

In our problem, the intensity of interaction along the row and column of the interaction matrix changes abruptly on domain boundaries, where the change point locations are the indices that are used to make horizontal and vertical boundaries of such domains. To solve this two-dimensional segmentation problem, we use a modified version of the binary segmentation algorithm proposed by Raveendran & Sofronov (2017), where the algorithm is used to detect homogeneous domains in two-dimensional lattice data. We use this algorithm to detect homogeneous domains along the diagonal of matrix X with the model defined in Section 2. Figure 3 illustrates an example of the first two iterations of the proposed algorithm.

One of the good features of this algorithm is that it detects the number of domains and their locations simultaneously and ends when no more change points are detected, but sometimes this leads to the overestimation of the total number of change points. To avoid this issue, one can see the task of estimating the number of domains as a model selection problem. In this study, we use popular model selection criteria such as AIC

(Akaike (1974)), BIC (Schwarz (1978)) and mBIC (modified BIC, defined for change-point problems) (Chen et al. (2006)) to estimate the number of domains. The AIC, BIC and mBIC for our model can be described as below

$$\text{AIC} = -2\text{log}L(\hat{\theta}_p) + 2p,$$

$$\text{BIC} = -2\text{log}L(\hat{\theta}_p) + p\log m,$$

$$\text{mBIC} = -2\,\text{log}L(\hat{\theta}_p) + 2(p+1)\log m,$$

where $\log L(\hat{\theta}_p)$ is the maximum likelihood for the model with p parameters, and $m = n \times n$ is the number of data points. A model that minimizes a criterion is considered to be the most appropriate model.

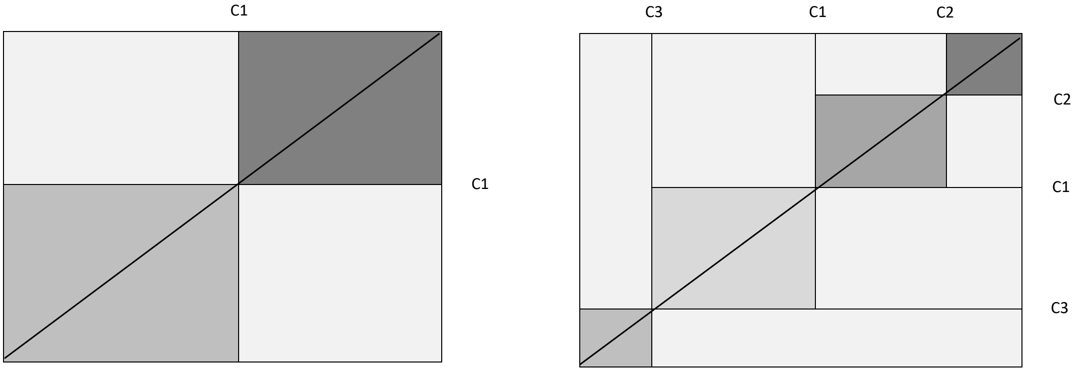

Figure 3. An example of first two iterations of the proposed algorithm. The left figure shows the first iteration starting with the first change point at C1 which identifies two diagonal domains. The right figure shows the second iteration with each domain identified in iteration 1 being divided into two domains with the change points C2 and C3, respectively. Note that, since it is a symmetric matrix, the detected change points in the i-th row and the i-th column are the same.

4 NUMERICAL RESULTS AND CONCLUSION

In this section, we discuss an example with artificially generated data to illustrate the usefulness of the proposed algorithm. We generated a non-overlapping diagonal interaction matrix of size $n = 200$ according to the model described in Section 2. Following Lévy-Leduc et al. (2014), we use specific parameter values which are derived from the interaction matrix of Chromosome 19 of the mouse cortex (see Dixon et al. (2012)). The resulting matrix has six diagonal block domains with the following parameter values: $\mu_1 = 2.87$, $\mu_2 = 4.85$, $\mu_3 = 7.92$, $\mu_4 = 4.33$, $\mu_5 = 11.99$, $\mu_6 = 3.67$, $\mu' = 0.09$ and $\sigma = 0.67$ with change points at $(20, 70, 100, 150, 180)$. Figure 4 shows the correctly identified diagonal domains and their boundaries for the generated data obtained by the proposed algorithm. It is clear from Figure 5 that the values of AIC, BIC and mBIC are lowest for the case when the number of domains is equal to 6. We also compared our results with the output obtained from the HiCSeg R package developed by Lévy-Leduc et al. (2014), which produced the same answer as ours, that is, detected 6 diagonal blocks with the exact same boundary locations. This shows that the proposed algorithm can perform well in identifying such domains in a two-dimensional segmentation setting.

In this paper, we proposed a simple and low computational cost algorithm to detect topological domains in Hi-C data matrix. In particular, we focus on identifying self-interacting blocks centered along the diagonal in the interaction matrix. We proposed a binary segmentation algorithm to detect such domains in two-dimensional segmentation framework. One of the key challenges in detecting such domains is the computational burden to fully take advantage of the Hi-C technology with high resolution. In such a scenario, the binary segmentation method is a fast algorithm and saves lots of computational time and it can be implemented with the computational complexity $O(n \log n)$. Our methodology could easily be extended to improve the algorithmic efficiency of our method and the modelling of the data in a way to detect both diagonal and off-diagonal regions with a hierarchical overlapping structure. This will provide more insight into 3D spatial chromosomal conformation in a better understanding of cell functioning.

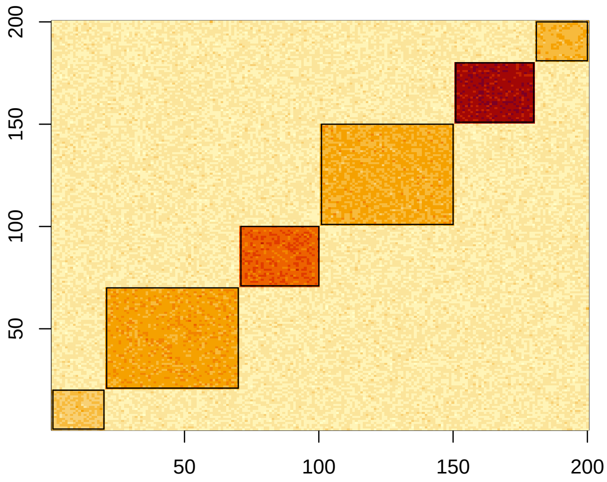

Figure 4. Estimated diagonal blocks and their boundaries for the generated data matrix.

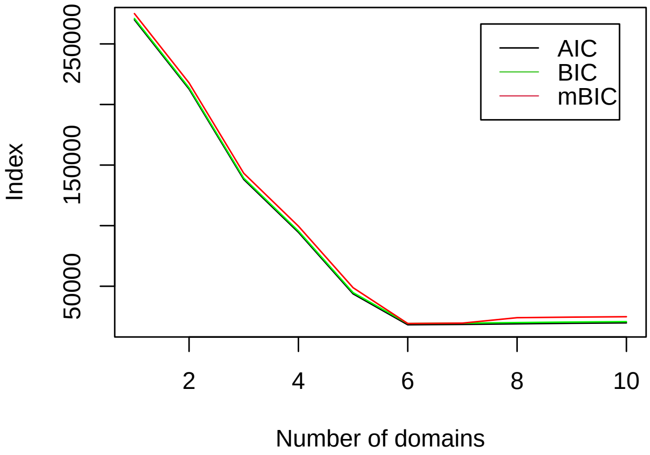

Figure 5. Number of diagonal blocks for the generated data matrix.

References

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Arbia, G., Espa, G. & Quah, D. (2008), 'A class of spatial econometric methods in the empirical analysis of clusters of firms in the space', *Empirical Economics* **34**, 81–103.

Beckage, B., Joseph, L., Belisle, P., Wolfson, D. B. & Platt, W. J. (2007), 'Bayesian change-point analyses in ecology', *New Phytologist* **174**(2), 456–467.

Cai, F., Le-Khac, N.-A. & Kechadi, T. (2016), 'Clustering approaches for financial data analysis: a survey', *arXiv preprint arXiv:1609.08520* .

Chen, J. & Gupta, A. K. (2012), *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*, Springer.

Chen, J., Gupta, A. & Pan, J. (2006), 'Information criterion and change point problem for regular models', *Sankhyā: The Indian Journal of Statistics* **68**(2), 252–282.

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

Evans, G. E., Sofronov, G. Y., Keith, J. M., Kroese, D. P. et al. (2011), 'Estimating change-points in biological sequences via the Cross-Entropy method', *Annals of Operations Research* **189**(1), 155.

Filippova, D., Patro, R., Duggal, G. & Kingsford, C. (2014), 'Identification of alternative topological domains in chromatin', *Algorithms for Molecular Biology* **9**, 1–11.

Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M. A., Hayashizaki, Y., Blanchette, M. & Dostie, J. (2009), 'Chromatin conformation signatures of cellular differentiation', *Genome biology* **10**(4), 1–18.

Fryzlewicz, P. (2014), 'Wild binary segmentation for multiple change-point detection', *The Annals of Statistics* **42**(6), 2243–2281.

Gangnon, R. E. & Clayton, M. K. (2000), 'Bayesian detection and modeling of spatial disease clustering', *Biometrics* **56**(3), 922–935.

Killick, R. & Eckley, I. (2014), 'changepoint: An R package for changepoint analysis', *Journal of statistical software* **58**(3), 1–19.

Lévy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. (2014), 'Two-dimensional segmentation for analyzing Hi-C data', *Bioinformatics* **30**(17), i386–i392.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O. et al. (2009), 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome', *science* **326**(5950), 289–293.

López, I., Gámez, M., Garay, J., Standovár, T. & Varga, Z. (2010), 'Application of change-point problem to the detection of plant patches', *Acta biotheoretica* **58**, 51–63.

Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. (2004), 'Circular binary segmentation for the analysis of array-based DNA copy number data', *Biostatistics* **5**(4), 557–572.

Polushina, T. & Sofronov, G. (2011), Change-point detection in biological sequences via genetic algorithm, *in* '2011 IEEE Congress of Evolutionary Computation (CEC)', IEEE, pp. 1966–1971.

Polushina, T. & Sofronov, G. (2016), A Cross-Entropy method for change-point detection in four-letter DNA sequences, *in* '2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)', IEEE, pp. 1–6.

Polushina, T. V. & Sofronov, G. Y. (2013), A hybrid genetic algorithm for change-point detection in binary biomolecular sequences, *in* 'Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2013)', pp. 1–8.

Priyadarshana, M. & Sofronov, G. (2012), A modified Cross Entropy method for detecting multiple change points in DNA count data, *in* '2012 IEEE Congress on Evolutionary Computation', IEEE, pp. 1–8.

Priyadarshana, W. & Sofronov, G. (2014), 'Multiple break-points detection in array CGH data via the Cross-Entropy method', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **12**(2), 487–498.

Raveendran, N. & Sofronov, G. (2017), 'Binary segmentation methods for identifying boundaries of spatial domains', *2017 Federated Conference on Computer Science and Information Systems* **8**, 95–102.

Raveendran, N. & Sofronov, G. (2019), Identifying clusters in spatial data via sequential importance sampling, *in* 'Recent Advances in Computational Optimization: Results of the Workshop on Computational Optimization WCO 2017', Springer, pp. 175–189.

Raveendran, N. & Sofronov, G. (2021), 'A Markov chain Monte Carlo algorithm for spatial segmentation', *Information* **12**(2), 58.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Scott, A. J. & Knott, M. (1974), 'A cluster analysis method for grouping means in the analysis of variance', *Biometrics* **30**(3), 507–512.

Sen, A. & Srivastava, M. S. (1975), 'On tests for detecting change in mean', *The Annals of Statistics* **3**, 98–108.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. & Cavalli, G. (2012), 'Three-dimensional folding and functional organization principles of the drosophila genome', *Cell* **148**(3), 458–472.

Sofronov, G. Y., Evans, G. E., Keith, J. M. & Kroese, D. P. (2009), 'Identifying change-points in biological sequences via sequential importance sampling', *Environmental Modeling & Assessment* **14**, 577–584.

Tripathi, S. & Govindaraju, R. S. (2009), Change detection in rainfall and temperature patterns over India, *in* 'Proceedings of the 3rd International Workshop on Knowledge Discovery from Sensor Data', pp. 133–141.

Zhou, Y. (2017), Statistical Methods to Detect Hierarchical Topological Domains in Chromatin, PhD thesis, Johns Hopkins University.