**Robust and High-Resolution Study of Chromatin Spatial Organization via Joint Modeling of Single Cell Imaging and Hi-C Sequencing Data**

Ying Sun, Eric Van Buren, Quan Sun, Yuchen Yang and Yun Li

**Abstract**

*The abstract consists of a single paragraph of 250 words or fewer. It should clearly convey the conceptual advance and significance of the work to a broad readership. In particular, the abstract should contain a brief background of the question, a description of the results without extensive experimental detail, and a brief summarization of the significance of the findings. References should not be cited in the abstract.*

The two most commonly-used methods to study chromatin spatial organization are

molecular-biology-based sequencing techniques and microscopy-based imaging

techniques. Sequencing techniques such as Hi-C are high-throughput and genome-wide, but lack the ability to capture chromatin structure at high resolution and detect difference between sub-populations within a heterogenous sample. By contrast, single cell imaging techniques such as ORCA reveal spatial organization in single cells, thus empowering profiling chromatin architecture in different cell types. Previous peak calling methods such as Fit-HiC and HMRFBayesHiC have been developed to identify biologically meaning interactions in a single datatype. It is exigent to refine our understanding of chromatin spatial organization by borrowing information from both imaging and sequencing data. Here, we leverage the complementary nature of sequencing and imaging data via a Cauchy combination test based joint model, to improve the detection of long-range pairwise chromatin interactions. To evaluate the joint model, we first tested the internal reproducibility of the model using ORCA single cell imaging and Hi-C data from BX-C region on drosophila 3R chromosome. Our results demonstrate improvement of internal consistency of detecting chromosomal interactions with controlled type-I error. We further compared our results with significant interactions identified by CHIA-DROP, and found the most significant joint-model peaks overlap with more known interactions than those from the traditional union approach.

**Introduction**

The manner by which eukaryotic genomes are packaged into nuclei while maintaining

crucial functions remains a fundamental mystery in biology. In each cell, around 2

meters of DNA are densely packed into chromatins, which further loop and fold to fit into

the nucleus of diameter <10 μm. Such looping and folding also allow distantly located genomic regions to interact 3D space1. Many studies have shown that these long-range chromosomal contacts have functional importance in regulating gene expression. For example, Smemo et al. (2016) provided compelling experimental evidences to show that obesity-associated variants within gene FTO regulates the functional gene IRX3 ~XX kb away (rather than FTO itself)2. Therefore, studying chromatin architecture will improve our understanding of the molecular underpinnings of biological processes underlying health and diseases.

Genome-wide chromosome conformation capture techniques such as Hi-C3,4 have been widely used to study chromatin spatial organization. Hi-C data can be summarized into contact matrix of all possible pairwise interactions between ligated fragments genome-wide. Numerous methods have been developed to call peaks from the Hi-C pairwise contact matrix. These methods seek to determine if the observed contact frequency is significantly higher than the null value which is expected from chromatin random collision in 3D space.

Although many methods exist for peak calling, we choose to focus on a Bayesian method utilizing Hidden Markov Random Fields (HMRF) first proposed in Xu et al. (2015) to call peaks from Hi-C data. HMRFs are an extension of Hidden Markov Models (HMMs) to two dimensions by creating an underlying Markov random field. The Markov property explicitly accounts for the dependency between neighboring genetic loci. Unlike many competitor methods, information is thus borrowed from neighboring regions to improve peak calling.

Optical reconstruction of chromatin architecture (ORCA) is an imaging technology that traces the DNA path in single cells. Primary probes with unique barcodes are placed at equal distance apart on a genomic region of interest and relative 3D position of each barcode is captured by microscopic imaging. ORCA reconstructs DNA trajectory at genomic resolution as high as 2 kilobases. This method is also compatible with simultaneous imaging of mRNA and nascent RNA in each cell, which makes differential analysis of chromosome architecture in different cell types possible.

There is substantial evidence which suggests that the physical distance between two genetic loci will be correlated with the contact frequency in 3D between the two regions. Loci closer in physical distance would naturally be assumed to be closer in 3D distance, however recent evidence has been suggested physical distance and 3D distance cannot be assumed as perfect substitutes for one another. Further, high-resolution single-cell imaging data can refine our understanding of chromosomal structure on a genomic region of interest and make predictions on other regions.

Our goal is to leverage the advantages of both data types via combining their estimated peak probabilities produce consensus peak probability estimates. This task must be undertaken with caution because the peak probability estimates from the two data sources are necessarily correlated. The nature of this dependency is unknown. In the paper, we develop joint model utilizing Cauchy combination test, which was recently articulated to combine dependent p-values Liu and Xie (2020). Our joint model does not require any information about the underlying correlation structure in the data while being able to account for the inherent correlation between their posterior probabilities. The model is also well-suited for a strong-sparsity setting.

**Material and Methods**

**Modelling the Null in Single Cell Imaging Data**

We first develop an algorithm for calling significant pairwise contact from single cell imaging data. Single cell imaging data can be converted to contact frequency matrix by categorizing a contact based on an observed distance between probes of 150nm or less, as was done in Mateo et al. (2019). Doing so, we can produce analogous heatmaps as those obtained for Hi-C. This is in a sense a loss of information, because the distance matters only to the extent of its relation to the threshold of 150nm.

To leverage the full potential of single cell data, we treat relative distance between pairs of genetic loci as a continuous variable. Our null model is based on a two-step smooth spline regression of 3D spatial distance of a pair of loci on their corresponding linear genomic distance.

Suppose we have an independent sample of n observations from some continuous bivariate distribution , and consider the nonparametric regression model

where , is some unknown “smooth” function, and  are error terms with mean zero and variance . The goal is to estimate the unknown function , from the data.

To estimate function , we minimize a penalized least square function

where  is a penalty term, and is the smoothing parameter that controls the effect the penalty. denotes the  derivative of , and is the space of functions with square integrable derivative. After an initial spline is fitted to the imaging data, we removed outliers that are 1.5 standard deviation less than the mean to eliminate top data points that are are closer in 3D distance than expected. With the remaining dataset, we fit a second smooth spline using the same method as described above and obtain a final null distance model for single cell imaging data.

**Single Cell Imaging Peak Calling**

After fitting a second spline, we perform one-tailed t-test on each pair of loci. Let represent the p-value for chromatin interaction between bin and bin from a specific

Hi-C peak calling method. Consider the null hypothesis that a loci pair is compatible with random chromatin looping . Define the test statistics as

where is the sample mean, is the null value, is the standard deviation and is the number of observations for each specific pair of loci. After one-tailed t-test we obtain , which then can be combined with Hi-C peak probabilities.

**Cauchy Combination Test**

The joint model is based on the aggregated Cauchy combination test to combine a set of p values . We use a linear combination of transformed *p*-values with non-negative weights:

where is the individual *p*-value and is the non-negative weight such that . When only one *p*-value is considered (), it is straightforward to show that follows a Cauchy distribution (location parameter, scale parameter ) under the null hypothesis that is uniformly distributed between 0 and 1.

**Step-by-step Workflow**

To implement the joint model, we first process the single cell imaging data. For technical

reasons, the coordinates of some probes will be missing. Therefore, we first perform linear interpolation to fill in coordinates of missing probes. Naturally, such approach involves some loss of precision, and could be refined by supervised methods in future work. Let represent the *p*-value for chromatin interaction between bin and bin from a specific Hi-C peak calling method.

**Results**

**Discussion**

**References**

1. Ramani, V. (2016). Understanding Spatial Genome Organization: Methods and Insights. Understanding Spatial Genome Organization: Methods and Insights, 14(1), 7-12.
2. Smemo, S., Tena, J. J., Kim, K., Gamazon, E. R., Sakabe, N. J., G mez-Mar n, C., . . .N brega, M. A. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature, 507(7492), 371-375. doi:10.1038/nature13138
3. 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).
4. Van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. J. Vis. Exp.