Extracting Physical Characteristics of Open and Closed Chromatin Folding Domains Technical Guide

William Franz Lamberti[1,*] **and Chongzhi Zang**[1,**]

[1]University of Virginia, Center for Public Health Genomics, Charlottesville, VA 22903, USA
[*]william.f.lamberti@virginia.edu
[**]zang@virginia.edu

Introduction

This document provides an overview of EPICS and example data for 3D-EMISH and 3D-SIM. An overview of CD assignment is first provided. Then, extended details are provided on determining open (A) and closed (B) CDs. Examples using EPICS are provided for 3D-EMISH and then 3D-SIM.

Defining CD Assignment for Images

Below is our solution for reconstructing the chromatin folding structure from the raw image data while adhering to the concepts of explainability and interpretability. Using image operator notation to represent the image processing operators applied to the input data[1], we first smooth the image via

$$\{\mathbf{s}[\vec{x}]\} = \mathcal{S}\{\mathbf{t}[\vec{x}]\} \tag{1}$$

where $\mathcal{S}$ is the smoothing operator, $\{\mathbf{t}[\vec{x}]\}$ is the input image, and $\{\mathbf{s}[\vec{x}]\}$ is the resulting smoothed image. We then isolate the relevant signals using

$$\{\mathbf{i}[\vec{x}]\} = \mathbb{L}\{\mathbf{t}[\vec{x}]\} \tag{2}$$

where $\mathbb{L}$ identifies the relevant signals of interest and $\{\mathbf{i}[\vec{x}]\}$ is the set of images containing the isolated signals of interest. We then applied

$$\mathbf{m}[\vec{x}] = \mathbb{I}_B\mathbf{i}[\vec{x}] \tag{3}$$

where $\mathbb{I}_B$ interpolates the object using the other slices to construct the missing slices and $\mathbf{m}[\vec{x}]$ is the reconstructed object of interest from the given target signal image, $\mathbf{i}[\vec{x}]$. We then determined the CDs by:

$$\{\mathbf{d}[\vec{x}]\} = \mathbb{C}\mathbf{m}[\vec{x}], \tag{4}$$

where $\mathbb{C}$ determines the CDs from the input image and $\{\mathbf{d}[\vec{x}]\}$ are the resulting CDs.

We then collected a variety of explainable and interpretable metrics using the shorthand operator of $\mathbb{D}$:

$$\mathbb{D} = \mathbb{D}\{\mathbf{d}[\vec{x}]\}. \tag{5}$$

From these extracted metrics in our resulting matrix $\mathbb{D}$, we built a model that created rules for predicting whether a particular CD is open or closed. In other words,

$$f(\mathbb{D}_{i,}) = \begin{cases} A & \text{Rule 1} \\ B & \text{Rule 2} \end{cases}, \qquad \forall i. \tag{6}$$

Determining Rules for A-B Compartmentalization

Equation 6 was displayed in the manner presented to compare it to the typical Hi-C computational approach. However, we will expand that notation here to provide an explicit description of our analysis. For notation purposes, we have J batches such that $j \in \{1, 2, ..., J\}$. For the 3D-EMISH data, $J = 2$. For the FBI experiments, $J = 1$ since each experiment has a different treatment. Further, we have 19 features such that $q \in \{1, ..., 19\}$. Thus, $\forall q, j$, we perform

$$|\mathbb{D}_{,q,j}| = \frac{\mathbb{D}_{,q,j} - \hat{\mu}_{q,j}}{\hat{\sigma}_{q,j}}, \tag{7}$$

where $\hat{\mu}_{q,j}$ and $\hat{\sigma}_{q,j}$ are the sample mean and standard deviation of the q^{th} feature for the j^{th} batch. We do this to normalize the data and remove batch effects. We then perform the following to obtain the estimated A-B compartments:

$$\vec{d}_{,j} = \Re_2 |\mathbb{D}_{,q,j}| \tag{8}$$

where $\Re$ is the $k-$means clustering operator. In this case, we are performing k-means clustering with a known number of clusters of 2. The output is the estimated A-B compartments saved in an associated vector, $\vec{d}_{,j}$. Here, we inspected each cluster and determined which cluster is associated with open and closed for each batch.

However, this is insufficient for determining clusters as we need to potentially interpret which cluster is open and which is closed. This introduces human subjectivity and error into the analysis. Further, the k-means clustering model does not provide any meaningful insight to which are the most important variables for discriminating the open and closed CDs from one another. Thus, we selected a LR model to describe open and closed CDs[2]. However, there are too many variables for this to be a truly interpretable and explainable model[3]. Thus, we used the least absolute shrinkage and selection operator (LASSO) to select the more important variables for our model[4–6]. Thus, we first modeled:

$$min_{\beta_0, \beta} \frac{1}{N} \sum_{i=1}^{N} l(\mathbb{D}_i, d_i, \alpha, \beta), \tag{9}$$

subject to $||\beta||_1 \leq \lambda$ where $||\cdot||_1$ is the L_1-norm and λ is a tuning parameter[6]. We selected the tuning parameter, λ, by using 10-folds cross validation on the training data. While the optimal λ was able to obtain a very high classification rate, it retained a larger number of variables. Thus, we selected the λ value within 1 standard error for the 3D-EMISH since it also has a very high classification rate, and has less variables retained in the models, and is the more prudent choice. Extended discussions on the choice of λ for the FBI data are provided in the Supplemental Material.

After we obtained the non-zero coefficients, we used those variables to model the following:

$$\log \frac{P(C = Closed | \mathbb{D} = x)}{P(C = Open | \mathbb{D} = x)} = \beta^T \mathbb{D}. \tag{10}$$

Image Processing for 3D-EMISH

The exact image processing steps differ from the more general framework provided in the main text in Equations 1 - 4. Specifically, we first apply

$$\mathbf{s}[\vec{x}] = \mathcal{S}_{0.1}(\mathbf{t}[\vec{x}] \otimes \mathbf{g}[\vec{x}]) \tag{11}$$

where $\mathbf{t}[\vec{x}]$ is the input image that contain a chromatin structure (CS), $\otimes$ is the convolution operator, $\mathbf{g}[\vec{x}]$ is the local minimum operator of a $2 \times 2 \times 2$ voxel, $\mathcal{S}_{0.1}$ is the Gaussian smoothing operator with a standard deviation of 0.1, and $\mathbf{s}[\vec{x}]$ is the resulting smoothed image. We first removed spurious noises using the local minimum using a $2 \times 2 \times 2$ window. We then further smoothed the image using a Gaussian operator to remove the block-like structure created by the previous operation. Examples of the input image and resulting image are provided in Figures 1 and 2, respectively.

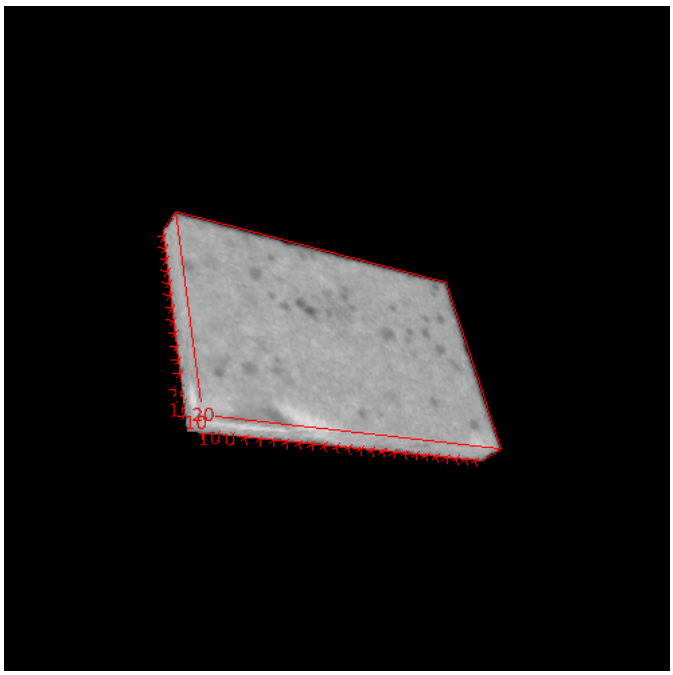

Figure 1: Example of input image for 3D-EMISH.

Next, we performed

$$\{\mathbf{o}[\vec{x}]\} = \Gamma_{\mathcal{O}}\{\mathbf{s}[\vec{x}]\} \tag{12}$$

where $\Gamma_{\mathcal{O}}$ applies Otsu thresholding to each slice of the input smoothed image-stack. This operation extracts the structure from the background without the need for human inputs. An example of this result is provided in Figure 3.

Next, we applied

$$\mathbf{i}[\vec{x}] = \mathcal{I}_{(1)} \rhd_1 \mathbf{o}[\vec{x}] \tag{13}$$

where $\mathcal{I}_{(1)}$ extracts the largest object from the image-stack and $\rhd_1$ erodes the isolated object 1 times. We first eroded the image isolate spurious aspects of the chromatin folding structure. We then isolated the largest object from the surrounding noise in the image. An example of this result is provided in Figure 4.

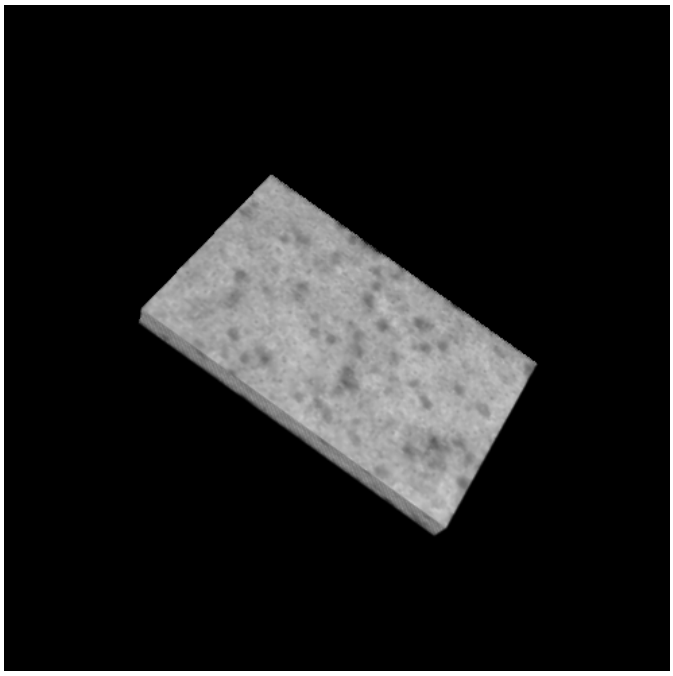

Figure 2: Example of smoothed 3D-EMISH image.

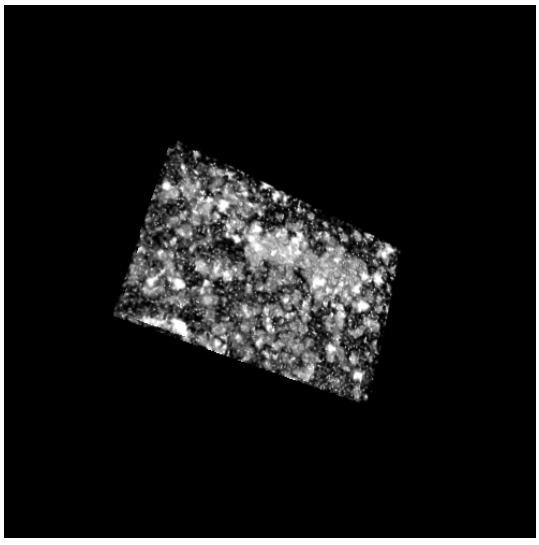

Figure 3: Example of isolated signals from 3D-EMISH example image.

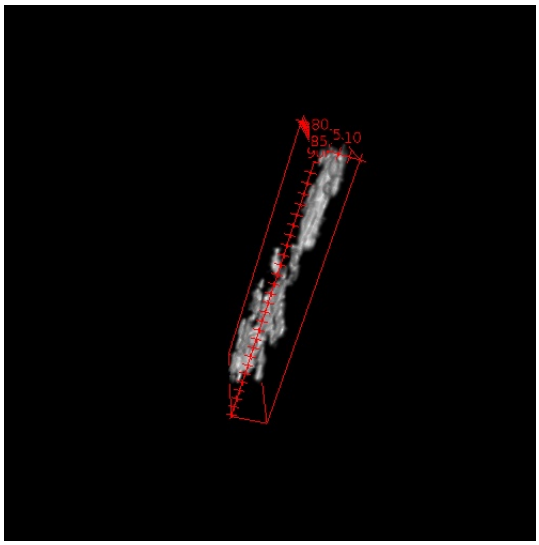

Figure 4: Example of largest isolated object from 3D-EMISH example image.

We then applied

$$\mathbf{m}[\vec{x}] = \mathbb{I}_B \mathbf{i}[\vec{x}] \tag{14}$$

where $\mathbb{I}_B$ interpolates the object using the other slices to construct the missing slices. We performed this step to provide an estimate of the chromatin physical structure in physical space. An example of this result is provided in Figure 5.

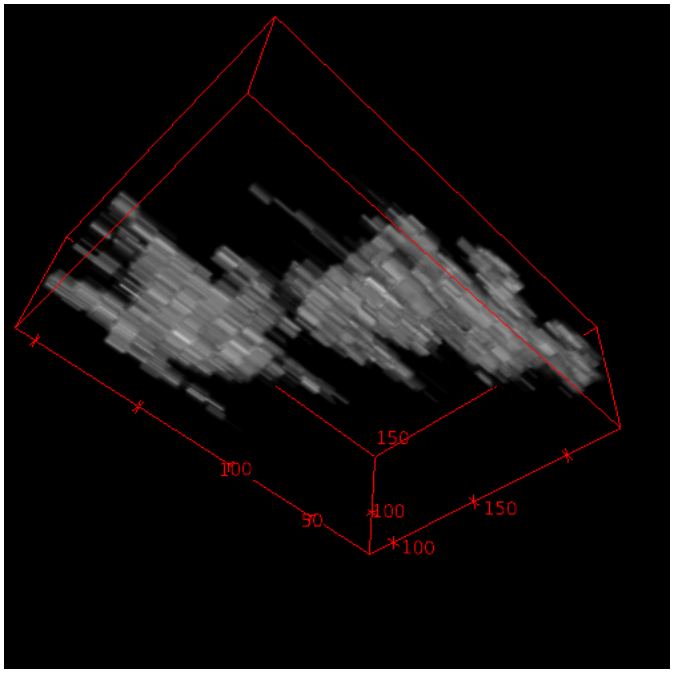

Figure 5: Example of reconstructed object of interest from 3D-EMISH data.

To determine the number of CDs, we then need to apply the mask to the original image. We do this by

$$\mathbf{c}[\vec{x}] = \mathbf{t}[\vec{x}] \times \mathbf{i}[\vec{x}] \tag{15}$$

$$\vec{c} = \mathfrak{K}_{K,10} \mathbb{P}_{0.60,1} \mathbf{c}[\vec{x}] \tag{16}$$

where $\mathbb{P}_{0.60,1}$ finds the local max peaks of the structure from the input image while ignoring those intensities less than the 60^{th} percentile of intensity values of the structure while requiring a distance of 1 voxel for each peak. The next image operator, $\mathfrak{K}_{K,10}$ performs k-means clustering on the potential peaks to identify the centers of each domain with automatic cluster determination. We consider up to 10 potential clusters. This outputs k centers in a vector $\vec{c}$. These operators were performed to find the center of each CD in the CFS. We then needed to extract the CDs in the correct units. Thus, we performed:

$$\{\mathbf{p}[\vec{x}]\} = \mathfrak{K}_{K,\vec{c}} \mathbb{P}_{0.60,1} \mathbf{c}[\vec{x}] \tag{17}$$

$$\mathbf{d}[\vec{x}] = \mathbb{I}_B \mathbf{p}[\vec{x}] \tag{18}$$

where we first predicted which domain each core of the structure in $\mathbf{i}[\vec{x}]$ belonged to in Equation 17. We then interpolated in Equation 18 as was done previously in Equation 14. An example of the discovered CDs is provided in Figure 6.

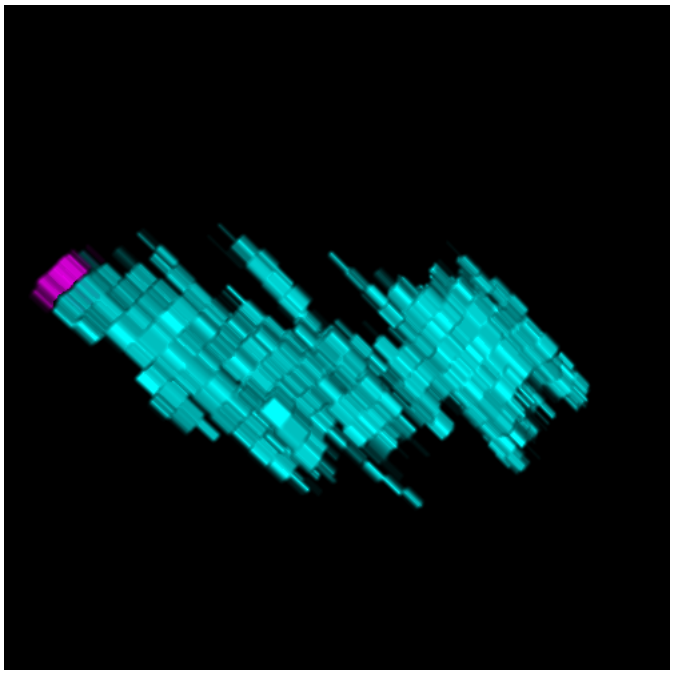

Figure 6: Example of CDs identified from example input 3D-EMISH data. Each different color represents a unique CD.

Image Processing for Fluorescence Immunostaining

There are multiple targets in the colon cancer cell image, this differs from the data obtained from 3D-EMISH data. The 3D-EMISH images had a single target in each image. Thus, some of the image processing steps need to be altered for us to collect the metrics.

Since the fluorescence images have three channels, we first need to relevant channels. To that end, we first apply

$$\mathbf{r}[\vec{x}] = \begin{Bmatrix} 1 \\ \varnothing \\ \varnothing \end{Bmatrix} \mathbf{t}[\vec{x}] \tag{19}$$

$$\mathbf{b}[\vec{x}] = \begin{Bmatrix} \varnothing \\ \varnothing \\ 1 \end{Bmatrix} \mathbf{t}[\vec{x}] \tag{20}$$

where we extracted the blue channel for the DAPI stained intensities and red channel for H3K27me3 intensities. $\mathbf{r}[\vec{x}]$ and $\mathbf{b}[\vec{x}]$ are shown together in Figure 7. The DAPI channel will be used as a mask to help remove the spurious signals outside of the nucleus. We then applied Equation 11 to smooth $\mathbf{b}[\vec{x}]$ and obtain $\mathbf{s}[\vec{x}]$ as seen in Figure 8.

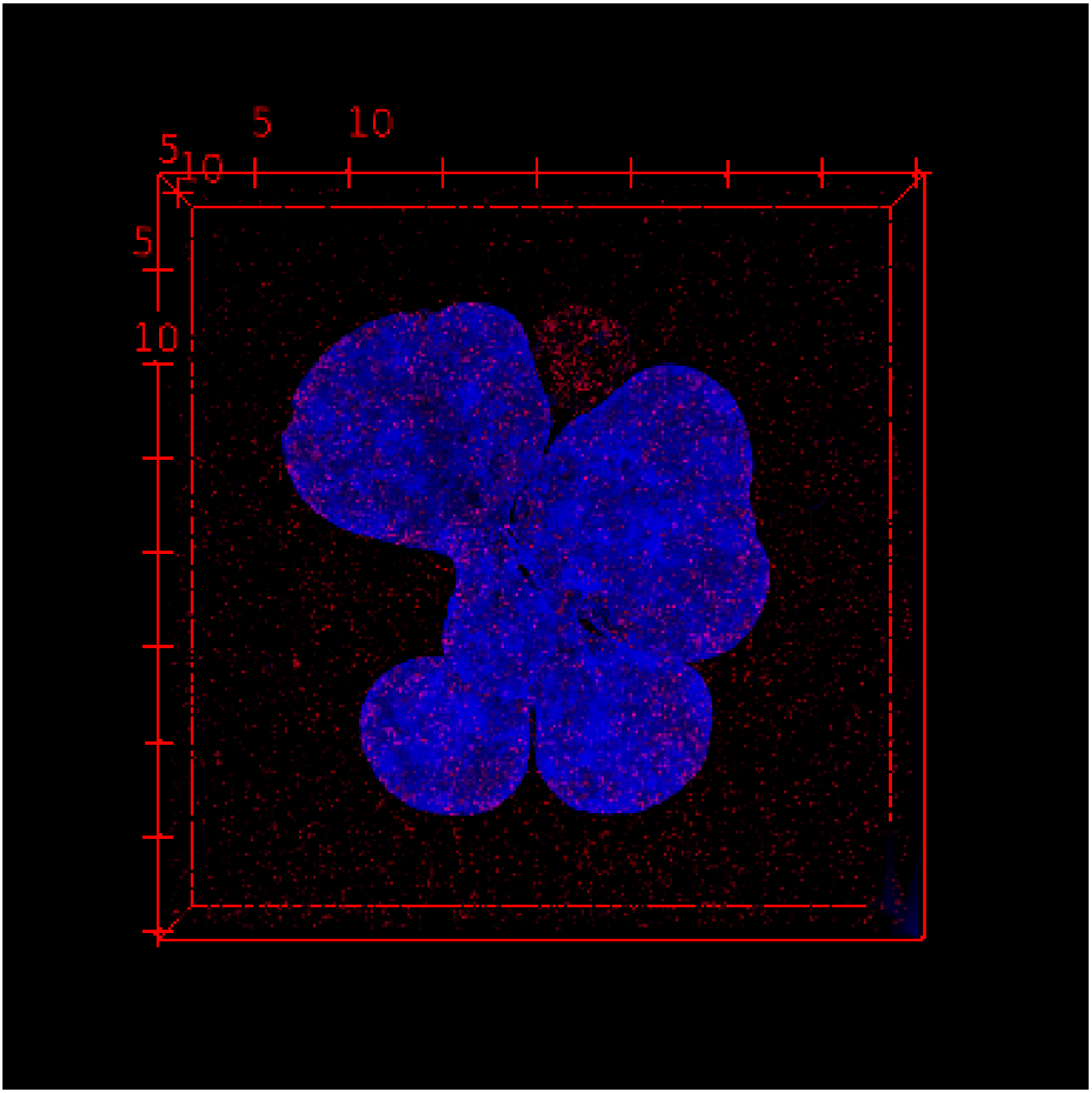

Figure 7: The DAPI and H3K27me3 FBI raw image data for the 30 hour treatment. Additional changes were made from the raw data in the presented image to help aid in visualization purposes of this manuscript.

We then applied

$$\mathbf{n}[\vec{x}] = \mathcal{B} \triangleright_5 \mathcal{B}\mathcal{I}_{(1)}\Gamma_{Q_{0.80}>}\mathbf{s}[\vec{x}] \tag{21}$$

where Q is the quantile operator, $\mathcal{B}$ is the binary fill hole operator, and $\triangleright$ is the erosion operator in order to extract mask of nucleus to extract only relevant H3K27me3 signals. The quantile operator was applied to ensure that only the strongest signals were retrained from the original image. The isolation, binary fill

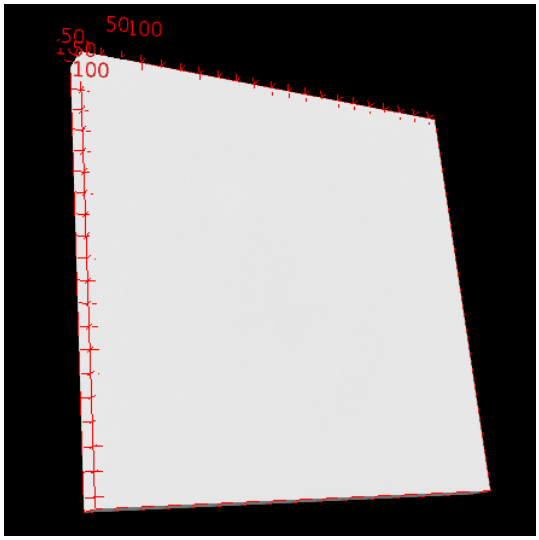

Figure 8: Smoothed image of the DAPI stained image from the FBI example. The input image was the example image from Figure 7.

holes, and erosion operators were applied to help identify the nucleus while also filling in the holes missed by the DAPI staining to ensure that every part of the nucleus is extracted. An example is shown in Figure 9

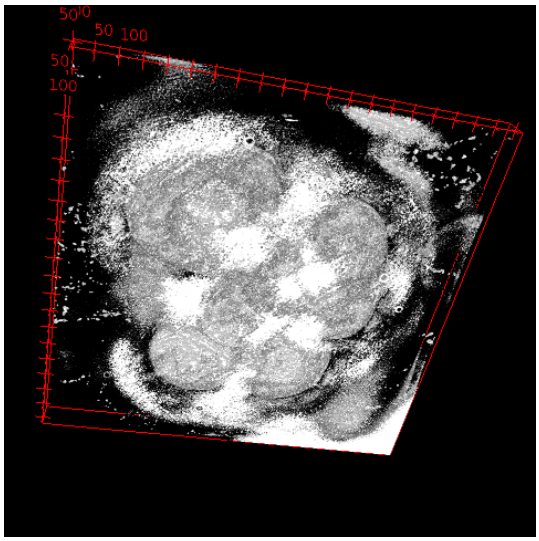

Figure 9: Example of identified mask from the DAPI stained image.

We then need to extract the relevant H3K27me3 signals. To that end, we apply:

$$\mathbf{r'}[\vec{x}] = \mathbf{n}[\vec{x}] \times \Gamma_{>59}\mathbb{H}_{R,0,255}\mathbf{r}[\vec{x}] \tag{22}$$

where $\mathbb{H}$ rescales the intensities from 0 to 255 in order to rescale and isolate the relevant signals. This results in an image of the mask of each CD in a single image as seen in Figure 10.

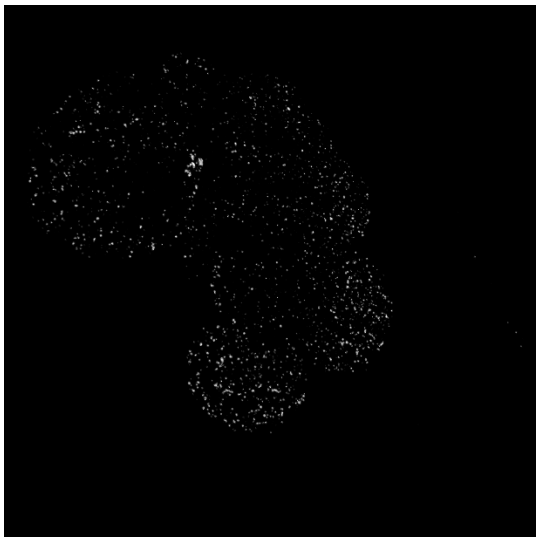

Figure 10: Example of the extracted signals of interest from the FBI image only within the nucleus.

To extract the relevant image masks for the individual CDs, we then apply

$$\{\mathbf{s}_i[\vec{x}]\} = \mathcal{I}\mathbf{r'}[\vec{x}] \tag{23}$$

which provides the set of individual CD masks such that $i \in \{1, 2, ..., n\}$. We then apply Equation 18 to $\mathbf{r}[\vec{x}] \times \{\mathbf{s}_i[\vec{x}]\}$ and $\{\mathbf{s}_i[\vec{x}]\}, \forall i$ to interpolate the image to ensure that each voxel is approximately the same in the $x-y-z$ directions. Figure 11 provides an example of all the CDs extracted from the example input image.

Figure 11: All of the CDs extracted from the example input FBI image. Each different grayscale intensity represents a unique CD.

We then extract the shape and intensity metric using Equation 5. Lastly, we then apply Equations 7 - 10 as in the usual case to classify the open and closed objects. However, we only have one experiment or batch in this case. Conversely, we have three different treatments. Thus, we will build three separate models for each of the three cells: the control, the 6 hour treatment, and the 30 hour treatment.

References

1. Kinser, J. M. *Image Operators: Image Processing in Python* (CRC Press, Boca Raton, FL, 2018), 1st edn.

2. Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression* (John Wiley & Sons, Incorporated, New York, UNITED STATES, 2013).

3. Lamberti, W. F. An Overview of Explainable and Interpretable Artificial Intelligence. In *AI Assurance: Towards Valid, Explainable, Fair, and Ethical AI* (Elsevier, 2022).

4. James, G., Witten, D., Hastie, T. & Tibshirani, R. (eds.) *An introduction to statistical learning: with applications in R*. No. 103 in Springer texts in statistics (Springer, New York, 2013). OCLC: ocn828488009.

5. Hastie, T., Robert, T. & Jerome, F. *Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2017), 2nd (corrected 12th printing) edn.

6. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288 (1994).