HiCSR: A Hi-C Super-Resolution Framework for Producing Highly Realistic Contact Maps

Michael C. Dimmick
University of Toronto
Vector Institute
mdimmick@psi.toronto.edu

Leo J. Lee
University of Toronto
Vector Institute
ljlee@psi.toronto.edu

Brendan J. Frey
University of Toronto
Vector Institute
frey@psi.toronto.edu

1 Introduction

High-throughput chromosome conformation capture (Hi-C) [1] is one of the most popular methods to analyze the three-dimensional architecture of the genome. Using proximity-based ligation combined with high-throughput sequencing, the Hi-C method produces a heat map contact matrix where each value represents the interaction frequency between two loci on the genome. The analysis of these matrices has led to discoveries on the nature of chromatin folding [2, 3] and its role in gene regulation [4, 5], uncovering a new relationship between genome architecture and cellular functions.

Typical Hi-C matrices are generated at a resolution between 1 Kb and 1 Mb per pixel. In general, a linear increase in resolution requires a quadratic increase in sequencing depth [6], making high resolution Hi-C data costly to obtain. Single image super-resolution techniques have been very successful when applied to natural images [7, 8]. In the context of Hi-C, super-resolution provides a method to computationally increase the number of sequencing reads and therefore increase the resolution of the contact matrix. These methods have provided researchers with a means to generate a high resolution Hi-C data set with a significantly reduced number of sequencing reads.

Previous deep learning based Hi-C enhancement methods that optimized for a Mean Squared Error (MSE), such as HiCPlus [9] and HiCNN [10], suffer from a lack of high frequency content resulting in a blurred output. This is caused by an objective function which prefers solutions that are the pixel-wise average of many possible solutions that lie on the plausible image manifold [11, 12]. To avoid blurred predictions, hicGAN [13] and DeepHiC [14] were proposed. First, hicGAN replaced pixel-wise loss functions with a purely adversarial loss. However, this caused hicGAN predictions to miss details found in the true high resolution data. DeepHiC combined an adversarial loss, pixel-wise loss, and a perceptual loss derived from a VGG-16 loss network [15] trained on ImageNet [16]. However, the introduction of this perceptual loss caused unwanted image artifacts in DeepHiC's predictions not otherwise found in real Hi-C data due to the use of a loss network trained on a natural image dataset.

We therefore proposed a novel Hi-C Super-Resolution (HiCSR) framework capable of accurately recovering the fine details found in high resolution Hi-C contact maps. This was achieved using a novel loss function tailored to the Hi-C enhancement problem. HiCSR optimizes both an adversarial loss and feature reconstruction loss obtained from the latent representation of a denoising autoencoder (DAE) [17] pretrained to reconstruct high resolution Hi-C data. HiCSR was able to produce visually convincing and highly accurate Hi-C matrix enhancements given Hi-C data with 16 times fewer aligned reads. This was achieved while avoiding smoothed outputs caused by MSE, and image artifacts caused by perceptual losses developed for natural images.

2 Methodology

The dataset used to train and evaluate HiCSR was generated by randomly down-sampling the original aligned reads by a factor of 16 to simulate a low resolution Hi-C dataset. From both the original (high resolution) and down-sampled reads, two sets of intrachromosomal contact maps were generated. Both sets were normalized by sequence depth to remove model dependency on the total number of raw interactions. We defined the matrix $\mathbf{M}^c$ as the raw contact matrix of chromosome c, and performed a log transform on the contact matrices given by $\mathbf{X}^c = \log_2(1 + \mathbf{M}^c)$. Next, we applied a linear transform $2\mathbf{X}^c / \max_{i,j}\{\mathbf{X}^c_{i,j}\} - 1$, normalizing the matrices to the range $[-1, 1]$ for each chromosome. From these normalized contact matrices, we cropped overlapping 0.4×0.4 Mb sub-matrices from each contact map. Interactions with a genomic distance > 2 Mb (far from the matrix diagonal) were discarded. This was done as most meaningful interactions occur within Topologically Associating Domains (TADs), and the majority of TADs have a size < 1 Mb within the human genome [2].

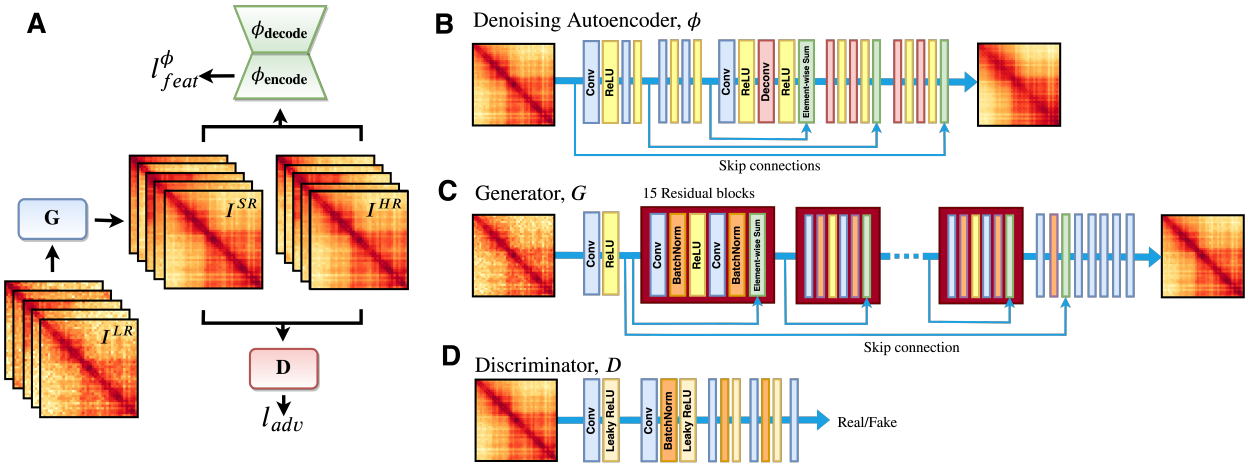

Figure 1: Overview of HiCSR. (**A**) During training, Generator G produces super-resolution outputs $I^{SR} = G(I^{LR})$ given a low resolution input, I^{LR}. Then, Discriminator D classifies true high resolution data I^{HR} and enhanced data $G(I^{LR})$ as either true high resolution data or an output from G. The discriminator's performance on the classification task produces an adversarial loss, l_{adv}. The inputs to D are also passed through the DAE, where feature reconstruction loss l^{ϕ}_{feat} is computed. (**B**) DAE consists of 5 convolutional layers followed by 5 deconvolutional layers with skip connections every other layer. (**C**) G network consists of 15 residual blocks with a skip connection, and is used to enhance low resolution Hi-C data to produce I^{SR}. (**D**) D network classifies inputs as real high resolution or enhanced Hi-C data.

An overview of the HiCSR framework can be seen in Figure 1. We employed an adversarial training procedure [18] with a generator network G and discriminator network D working in opposition. The discriminator network was trained to differentiate between real high resolution Hi-C samples and outputs produced by the generator network. In turn, the generator network attempted to produce enhanced Hi-C data which fooled the discriminator network into accepting its predictions as true high resolution data. We used the standard generator loss over all training samples:

$$l_{adv} = \sum_{n=1}^{N} -\log D(G(I_n^{LR})) \tag{1}$$

A DAE was then trained with symmetric skip connections using a fully convolutional image restoration architecture [19] which learned to reconstruct a clean sample from a noise corrupted high resolution input. High resolution Hi-C inputs I^{HR} were corrupted with Gaussian noise: $\tilde{I}^{HR} = I^{HR} + \eta Z$ where $Z \sim \mathcal{N}(0, 1)$ and η is a noise corruption factor. The DAE was trained to reconstruct the noise free input by minimizing MSE between the original matrix and the DAE's output.

Following previous work outlining the use of loss networks [8, 20], we used the encoder of the DAE as a loss network ϕ in place of a pixel-wise loss, such as MSE. We computed a feature reconstruction loss l^{ϕ}_{feat} which measured the similarity between feature maps of shape $C_j \times W_j \times H_j$ for each of the $j = 1, \ldots, N$ encoder layers between two separate DAE inputs. Specifically, the feature reconstruction loss was computed as the sum of the squared normalized Euclidean distances between the pre-activation feature representations of the high resolution matrix I^{HR}, and enhanced matrix $G(I^{LR})$ across all layers of the encoder network:

$$l^{\phi}_{feat} = \sum_{j=1}^{N} \frac{1}{C_j W_j H_j} \|\phi_j(I^{HR}) - \phi_j(G(I^{LR}))\|_2^2 \tag{2}$$

The total objective function of the generator to be minimized was then given by $l_{total} = \lambda_a l_{adv} + \lambda_f l^{\phi}_{feat}$, where λ_a and λ_f are scaling constants. Each aspect of the loss function focuses on a specific and desirable aspect of enhancement. The feature reconstruction loss ensures that HiCSR enhanced Hi-C data shares accurate feature representations with true high resolution Hi-C data. The adversarial loss ensures that the generator favours outputs which lie on the true high resolution Hi-C data manifold, encouraging visually convincing solutions. The minimization of the combined losses ensures that HiCSR generates enhanced Hi-C data that is both accurate and visually convincing.

3 Experiments

The performance of HiCSR was evaluated on the GM12878 cell line using paired-end Hi-C reads downloaded from the Gene Expression Omnibus (GEO) database (accession GSE63525) [3]. The reads were processed into low and high resolution contact maps using the Hi-C processing pipeline, HiC-Pro [21] with default settings. The high resolution matrix was generated with all available paired end reads, and the low resolution matrix using $1/16^{\text{th}}$ the original paired end reads. The contact matrices were then cropped as described in Section 2 to create a dataset of low resolution sub-matrices and their corresponding high resolution counterparts for each chromosome. The dataset was then split such that chromosomes 1-16 were used for training, 17, 18 for validation, and 19-22 for testing.

For all comparisons, pretrained models for both DeepHiC and hicGAN provided by the authors were used. As HiCPlus and HiCNN prescribe no normalization and were sensitive to sequence depth, they were retrained according to the training methods described in their respective publications.

Table 1: Performance on test chromosomes for different loss functions and state-of-the-art MSE and adversarial based enhancement methods compared to HiCSR

	LR	HiCNN	hicGAN	HiCSR	HR
MAE	2.323	0.433	0.554	0.389	0
MSE	7.484	0.321	0.567	0.352	0
PSNR (dB)	12.96	26.64	24.20	26.24	∞

Both the generator/discriminator pair, as well as the DAE network, were trained on 70484 sub-matrices of size 40×40. The DAE was trained over 600 epochs using the Adam optimizer [23] with a batch size of 256, a learning rate of 5×10^{-3}, and a noise corruption factor of $\eta = 0.1$. After training the autoencoder, the generator and discriminator training was done in an alternating fashion over 500 epochs using the Adam optimizer with a batch size of 64, and a learning rate of 10^{-5}. The LeakyReLU activation used in the discriminator was implemented with $\alpha = 0.2$. Scaling factors λ_a and λ_f were chosen through cross-validation as 0.1 and 1 respectively. Once trained, Hi-C super-resolution predictions were made with the generator alone, and the discriminator and DAE were discarded.

On the test chromosomes, we compared HiCSR's predictions to the low resolution Hi-C contact map (LR), and the state-of-the-art MSE and adversarial based enhancement methods. Metrics used for the comparison were Mean Absolute Error (MAE), MSE, and Peak Signal to Noise Ratio (PSNR) and results are summarized in Table 1. We found that HiCSR outperformed all models in MAE, and adversarial methods in MSE and PSNR. Despite a worse MSE performance compared to HiCNN, HiCSR produces a highly realistic enhanced contact map capable of reproducing high frequency content found in true high resolution Hi-C data. Visual examples comparing a suite of Hi-C enhancements are included in Figure 2, along with Insulation Scores [24] and called TAD boundary annotations. The aforementioned image artifacts produced by the DeepHiC method can be seen in the comparison. We found that all enhancement methods were highly successful at recovering true TAD boundaries, and that there was a high correlation ($r > 0.90$) between true high resolution Hi-C data and enhanced predictions from all models.

We then examined the performance of HiCSR as a function of genomic distance (Figure 3 A,B). Specifically, we computed the MSE and Person Correlation Coefficient (PCC) as a function of genomic distance on the test chromosomes and found that HiCSR showed a clear advantage over the previous state-of-the-art adversarial based model on all

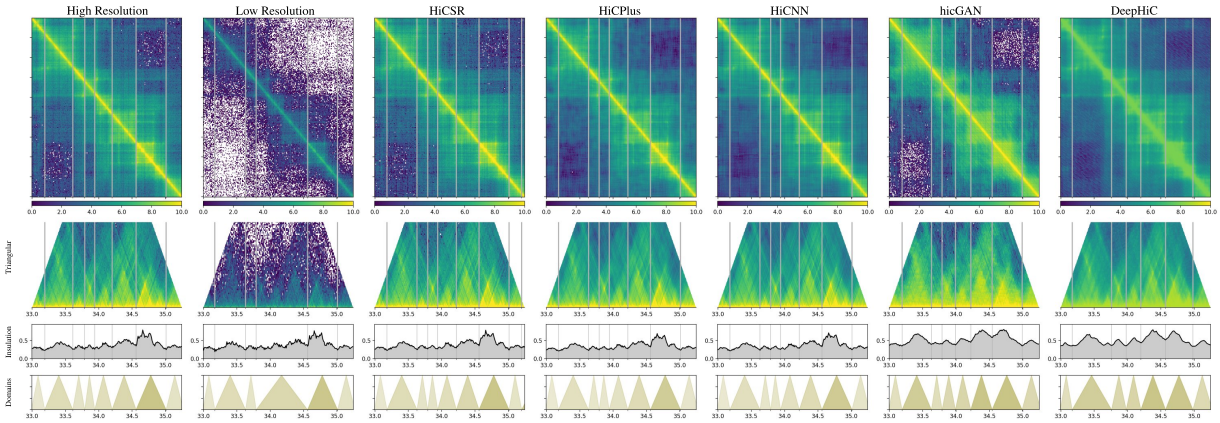

Figure 2: Log scaled super-resolution comparisons of chr21:33 - 35.25 Mb at 10 Kb resolution with Insulation Score and TAD annotations computed and plotted with HiCPlotter [22]. From left to right: original high resolution, 16x down-sampled, HiCSR, HiCPlus, HiCNN, hicGAN, and DeepHiC.

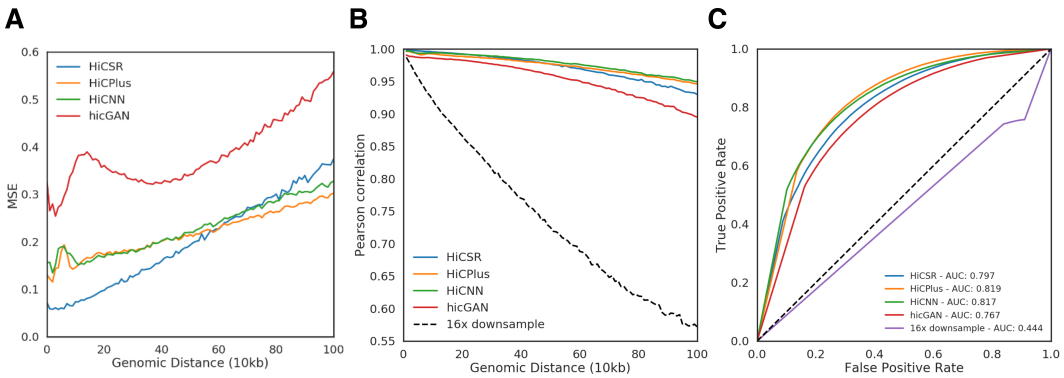

Figure 3: Model comparisons, HiCSR in blue. (**A**) MSE evaluated at a range of genomic distances. (**B**) PCC across genomic distance between real high resolution Hi-C data and enhanced matrices. (**C**) ROC curve for identifying statistically significant contacts found in high resolution Hi-C data averaged over ten classification runs.

metrics. Additionally, at low genomic distances HiCSR outperformed all previously proposed models on all measured metrics. This is a particularly useful property as the majority of low noise, and significant interactions occur at low genomic distances. We excluded DeepHiC from metric based comparisons, as its error at low genomic distances is magnitudes larger than other models.

We also compared each model's ability to recover statistically significant chromosomal interactions found in the high resolution contact matrices using Fit-Hi-C [25]. A binary classification problem was formulated by calling significant interactions with a cutoff of $q < 0.05$ within a genomic distance of 2 Mb. As there are few statistically significant contacts relative to the total size of the matrix, we randomly sub-sampled insignificant contacts to match the number of statistically significant contacts to make a balanced dataset. The Receiver Operator Characteristic (ROC) curve was plotted and the Area Under the Curve (AUC) was computed as shown in Figure 3 (C). We found that HiCSR achieved an AUC score of 0.797, outperforming the current best adversarial based model hicGAN, which achieved a score of 0.767.

4 Discussions and future work

Our findings show that the HiCSR framework is capable of producing accurate and visually convincing high resolution Hi-C contact maps from low resolution data using 16 times fewer sequencing reads. Our method leveraged the strengths of all previous deep leaning methods and improved them by introducing a feature reconstruction loss developed specifically for the super-resolution of Hi-C data. In this way, HiCSR outperforms all previous models for low genomic distances, and adversarial based methods at all genomic distances through increased accuracy and a reduction in optimization artifacts caused by a perceptual loss from natural images.

Through this work we demonstrated the efficacy of domain specific loss networks for biological problems that are formulated as non-natural images. We showed that network losses trained on natural images do not always transfer to problems outside their intended domain. This was observed in DeepHiC's method where a loss network from a VGG-16 image classifier introduced undesirable image artifacts, and produced patterns not found in the true data generating distribution. We showed that the use of a pretrained, domain specific DAE as a loss network is a viable substitute to standard perceptual loss functions for problems outside the natural image domain.

There are many ways in which this work can be further explored and improved. An investigation into additional domain specific loss terms, such as arrowhead matrix [3] reconstruction, could improve model performance and ensure that the enhanced matrices share important problem specific properties with true high resolution Hi-C data. Additionally, further exploration aims to evaluate the effect of larger down-sample ratios on Hi-C enhancement, as well as the ability for HiCSR to perform super-resolution on cell lines previously unseen in the training data. Finally, we plan to improve upon this work by expanding the suite of similarity metrics used to compare performance to include Hi-C reproducibility methods such as HiCRep [26] and HiC-Spector [27] to gain further insight into both the strengths and weaknesses of the HiCSR framework.

References

[1] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, no. 5950, pp. 289–293, 2009.

[2] J. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, pp. 376–80, 2012.

[3] S. Rao, M. H Huntley, N. Durand, E. K Stamenova, I. Bochkov, J. Robinson, A. L Sanborn, I. Machol, A. Omer, E. S Lander, and E. Lieberman Aiden, "A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, 2014.

[4] M. Franke, D. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jerković, W.-L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, and S. Mundlos, "Formation of new chromatin domains determines pathogenicity of genomic duplications," *Nature*, vol. 538, pp. 265–269, 2016.

[5] D. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, and S. Mundlos, "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions," *Cell*, vol. 161, 2015.

[6] A. Schmitt, M. Hu, and B. Ren, "Genome-wide mapping and analysis of chromosome architecture," *Nature Reviews Molecular Cell Biology*, vol. 17, 2016.

[7] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016.

[8] J. Johnson, A. Alahi, and F. F. Li, "Perceptual losses for real-time style transfer and super-resolution," 2016.

[9] Y. Zhang, L. An, J. Xu, B. Zhang, W. J. Zheng, M. Hu, J. Tang, and F. Yue, "Enhancing hi-c data resolution with deep convolutional neural network hicplus," *Nature Communications*, vol. 9, no. 1, p. 750, 2018.

[10] T. Liu and Z. Wang, "HiCNN: a very deep convolutional neural network to better enhance the resolution of Hi-C data," *Bioinformatics*, 2019.

[11] M. Mathieu, C. Couprie, and Y. Lecun, "Deep multi-scale video prediction beyond mean square error," 2015.

[12] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016.

[13] Q. Liu, H. Lv, and R. Jiang, "hicGAN infers super resolution Hi-C data with generative adversarial networks," *Bioinformatics*, vol. 35, no. 14, pp. 99–107, 2019.

[14] H. Hong, S. Jiang, H. Li, C. Quan, C. Zhao, R. Li, W. Li, G. Du, X. Yin, Y. Huang, C. Li, H. Chen, and X. Bo, "Deephic: A generative adversarial network for enhancing hi-c data resolution," *bioRxiv*, 2019.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.

[17] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," pp. 1096–1103, 2008.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," pp. 2672–2680, 2014.

[19] X. Mao, C. Shen, and Y. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections," *CoRR*, vol. abs/1606.08921, 2016.

[20] Mao, "Single image super-resolution via perceptual loss guided by denoising auto-encoder," *SpringerLink*, 2018.

[21] N. Servant, N. Varoquaux, B. R. Lajoie, E. Viara, C.-J. Chen, J.-P. Vert, E. Heard, J. Dekker, and E. Barillot, "Hic-pro: an optimized and flexible pipeline for hi-c data processing," *Genome Biology*, vol. 16, no. 1, p. 259, 2015.

[22] K. C. Akdemir and L. Chin, "Hicplotter integrates genomic data with interaction matrices," *Genome Biology*, vol. 16, no. 1, p. 198, 2015.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2014.

[24] E. Crane, Q. Bian, R. P. McCord, B. R. Lajoie, B. S. Wheeler, E. J. Ralston, S. Uzawa, J. Dekker, and B. J. Meyer, "Condensin-driven remodelling of x chromosome topology during dosage compensation," *Nature*, 2015.

[25] F. Ay, T. Bailey, and W. Stafford Noble, "Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts," *Genome research*, vol. 24, 2014.

[26] T. Yang, F. Zhang, G. Yardimci, F. Song, R. Hardison, W. Noble, F. Yue, and Q. Li, "Hicrep: assessing the reproducibility of hi-c data using a stratum- adjusted correlation coefficient," *Genome Research*, vol. 27, p. gr.220640.117, 2017.

[27] K.-K. Yan, G. G. Yardımcı, C. Yan, W. S. Noble, and M. Gerstein, "HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps," *Bioinformatics*, vol. 33, no. 14, pp. 2199–2201, 2017.