HiC2Self: Self-supervised Hi-C contact map denoising

Rui Yang[1] **Alireza Karbalayghareh**[1] **Christina Leslie**[1]

Abstract

We propose HiC2Self, a self-supervised method for denoising Hi-C contact maps that needs only low coverage data for training and imputes high coverage interaction count data that can be used for downstream analyses. Using a self-denoising framework based on Noise2Self, we designed a unique mask structure tailored for Hi-C contact maps and adopted a negative binomial loss function in order to directly process the raw count matrix without additional normalization or recovery steps. By training on multiple resolutions simultaneously, HiC2Self is able to capture global and local contact structures. We find our self-supervised method is competitive with or outperformed existing supervised Hi-C denoising algorithms while providing greater ease of use, as well as having the potential to be applied to single-cell Hi-C data.

1. Introduction

Hi-C is a genome-wide chromatin conformation capture assay that is used to study 3D genomic organization. Hi-C paired-end sequencing data produces a contact matrix between genomic bins that reveals principles of chromatin folding at resolutions, such as A/B compartments when data is binned at megabase scale and topologically associating domains (TADs) for 10-50kb bins (Szabo et al., 2019). Intra-chromosomal Hi-C contact maps are usually visualized by a symmetric heatmap, where x and y coordinates indicate genomic locations along the chromosome, and each pixel shows the strength of chromatin interaction (normalized read count) between the corresponding bins. High-resolution Hi-C contact maps require generation of multiple replicate libraries and extremely high sequencing coverage (1-2B reads), incurring considerable costs. Contact maps generated from libraries with only shallow sequencing have high noise due to sparsity.

[1]Computational Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, US. Correspondence to: Christina Leslie <cleslie@cbio.mskcc.org>.

Preliminary work. Under review by *the 2023 ICML Workshop on Computational Biology*. Do not distribute.

Given the success of deep learning technology for image denoising and super-resolution, several groups have designed supervised deep learning models to "denoise" Hi-C contact maps. HiCPlus (Zhang et al., 2018) and HiCNN (Song et al., 2018) use convolutional neural networks to predict high coverage 2D contact maps from low coverage or downsampled contact maps in the same cell type. hicGAN (Liu et al., 2019), DeepHiC (Hong et al., 2020) and HiCSR (Dimmick et al., 2020) all use generative adversarial networks (GAN) to impute high resolution data, with DeepHiC and HiCSR employing loss functions specifically tailored to Hi-C data. These supervised approaches all require paired low-/high-coverage Hi-C data to train the model, which can then be applied to other cell types where only low-coverage data are available. Existing approaches also normalize and preprocess Hi-C input data to fit the training framework, which typically requires an additional post-prediction recovery procedure to reconstruct a genome-wide matrix for downstream analysis.

In this study, we present HiC2Self, a self-supervised Hi-C denoising model that only requires low-coverage Hi-C data for training and can be applied directly to raw count matrices without normalization steps. The self-supervision framework is based on Noise2Self (Batson & Royer, 2019), with a mask structure and negative-binomial loss function designed for Hi-C raw count matrices.

2. Method

2.1. Data Preparation

High coverage Hi-C data sets are generated by sequencing multiple libraries and aggregating read counts across libraries. To obtain low-coverage Hi-C training data, we generated a contact map from a single library and evaluated performance against the aggregated multi-library map. Intra-chromosomal Hi-C raw count contact maps were generated without normalization. For each chromosome in the low-coverage dataset, we further extracted equal-sized square submatrices along the diagonal, representing genomic interactions up to 1Mb in linear distance. These symmetric submatrices X are used as the training set for our model.

2.2. Self-supervised framework

Noise2Self (Batson & Royer, 2019) is a self-supervised denoising framework that uses $\mathcal{J}$-invariant functions f, where $\mathcal{J}$ represents a partition of the input data dimensions m into subsets, and we consider a subset $J \in \mathcal{J}$ and its complement J^C. Given an unseen clean signal $y \in \mathbb{R}^m$, we assume that x is a mean-zero noisy observation, where $\mathbb{E}[x|y] = y$. For any fixed subset J, we further assume that a noisy observation on subdimension x_J is independent of the one on its complement x_{J^C} given y. With these two assumptions, a function $f : \mathbb{R}^m \to \mathbb{R}^m$ is defined as $\mathcal{J}$-invariant if $f(x)_J$ is independent of x_J for every $J \in \mathcal{J}$.

The ordinary denoising loss function is defined as

$$\begin{aligned}\mathcal{L}_f &= \mathbb{E}_{x,y}||f(x) - y||^2 \\ &= \mathbb{E}_x||f(x) - x||^2 + ||x - y||^2 - 2\langle f(x) - x, x - y\rangle\end{aligned}$$

which is the sum of a self-supervised loss and the variance of the noise. With a J-invariant function f and the previous assumptions, this simplifies to

$$\mathcal{L}(f) = \sum_{J \in \mathcal{J}} \mathbb{E}||f_J(x_{J^C}) - x_J||^2$$

so that the denoising function f can be optimized using only noisy observations x.

The $\mathcal{J}$-invariance property is realized using masks. We denote the masked area as x_J and the unmasked area as x_{J^C}. Given the symmetric nature of Hi-C contact maps and the requirement that $x_J \perp\!\!\!\perp x_{J^C}|y$, we designed masks that are symmetric with respect to the diagonal.

The self-supervised training scheme is shown in Figure 1A.

2.3. Multi-resolution training

As introduced in Section 1, Hi-C contact maps can be binned into different resolutions, which provides diverse insights of chromatin folding structures. Therefore, training HiC2Self on multiple resolutions simultaneously to capture both local and global information of 3D structure has become a natural choice. Define function $g : \mathbb{R}^m \to \mathbb{R}^{\frac{m}{2}}$ as resolution degradation. For each resolution, HiC2Self is trained to minimize the loss between $g_J(x)$ and $g_J(f(x_{J^C}))$. We trained HiC2Self at four different resolutions simultaneously.

By incorporating multiple resolutions into the training process, HiC2Self can effectively capture the intricate details of local chromatin interactions as well as considering the broader, global genomic associations. This approach greatly improves the signal recovery ability of HiC2Self, especially on sparse single-cell Hi-C data. Multi-resolution training approach is shown in Figure 1B.

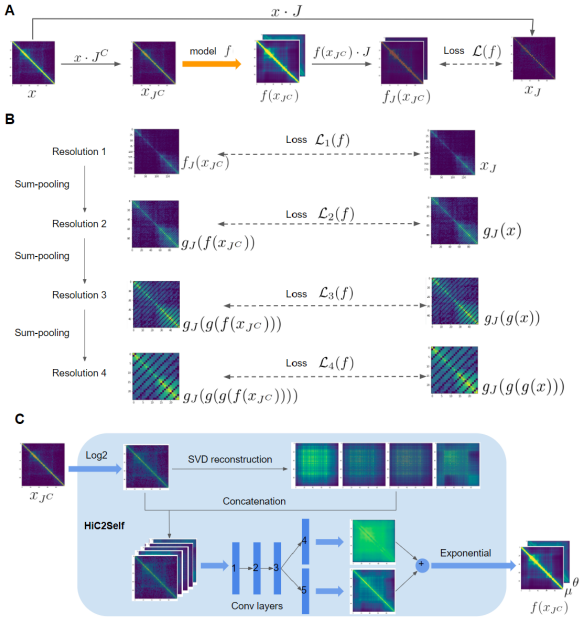

Figure 1. Training framework and model architecture

2.4. Model architecture

HiC2Self uses a simple convolutional neural network (CNN), as shown in Figure 1C. Within the model, raw count input matrices X were first log2-transformed ($X' = log2(X_{J^C} + 1)$) in order to guarantee numerical stability for subsequent steps.

Singular value decomposition (SVD) and low-rank reconstruction is a classic approach for 2D image compression and denoising. In order to enhance the low-rank structures extracted from low-coverage submatrices in the log2-transformed space, we performed SVD on the log2-transformed matrices $X' = U\Sigma U^T$, generated reconstructions $X'_k = \sum_{i=1}^{k} u_i \Sigma_i u_i^T$ using the top k eigenvectors, $k \in [1, 4]$, and concatenated these matrices with X' as additional input channels for the CNN.

The convolutional part of the model consists of five equal-sized convolutional layers, where each of the first three layers is followed by ReLU activation functions. An exponential function was used as the activation function for layer 4 and 5 in order to transform output values back into raw count space.

2.5. Loss function

2.5.1. Denoising bulk Hi-C

Inspired by the deep count autoencoder (DCA) model for single cell data (Eraslan et al., 2019), we used a negative binomial loss for the raw count matrices to train our model. We assume that count from each bin (x_{ij}) of the contact map

X follows a negative binomial distribution with parameters μ_{ij} and θ_{ij}, $x_{ij} \sim NB(\mu_{ij}, \theta_{ij})$. The loss function is defined as

$$\begin{aligned}&\mathcal{L}(f)\\&= -logL_{NB}\\&= \sum(log\Gamma(x+1) + log\Gamma(\theta) - log\Gamma(x+\theta) +\\&\theta log(\frac{\mu+\theta}{\theta}) + xlog(\frac{\mu+\theta}{\mu}))\end{aligned}$$

As shown in Figure 1B, HiC2Self outputs two channels, corresponding to μ and θ in the loss function above. We use μ_{ij}, the expected value for each bin x_{ij}, as the predicted value for our denoising results.

2.5.2. Application on single-cell Hi-C

Due to the highly sparse nature of single-cell data, employing a likelihood-based loss function poses challenges when training HiC2Self. To address this issue, we incorporated an alternative training function by applying structural similarity (SSIM) loss during the training of HiC2Self on single-cell Hi-C. SSIM is a perceptual loss metric that measures the structural similarity between two images. It takes into account information about luminance, contrast, and structural similarity, making it potentially more robust to sparse targets.

2.6. Genome-wide prediction

HiC2Self produces denoised results as raw counts, which can easily be assembled into a whole-chromosome prediction. To do this, we extracted submatrices along the diagonal, consecutively striding by one bin each time. Denoised results were generated for each submatrix, and predicted counts for overlapping submatrices were averaged. The resulting predicted high coverage results were saved as a .hic file using Juicer tools (Durand et al., 2016) for downstream analysis.

3. Experiments and Results

3.1. Data Preparation

Bulk Hi-C data. HiC2Self was trained and evaluated on real low- and high-coverage Hi-C data as described above. Low-/high-coverage raw count matrices for the ENCODE GM12878 cell line were downloaded from GEO (GSE63525 (Rao et al., 2014)). A single low-coverage library (experiment HIC001) with 2.5M reads was used as low-coverage data to train the model, and pooled primary libraries with 3.5B reads (low/high ratio = $1/18$) was used as high-coverage Hi-C data to evaluate model performance. Raw count data were downloaded in .hic format and further binned at 10kb resolution matrix using Juicer (Durand et al., 2016). Equal-sized (100×100) submatrices were extracted along the diagonal from intra-chromosomal low-coverage Hi-C contact maps to train the model.

Single-cell Hi-C data. Single-cell Hi-C contact maps were downloaded from GEO (GSE49262 (Nagano et al., 2013)). Contact maps of chromosome 1 from each cell are binned at 1Mb resolution, and equal-sized (198×198) submatrices were used for training.

3.2. Denoising on bulk Hi-C

We first evaluated the prediction performance on bulk Hi-C using the low-coverage library (2.5M reads), and the distance-adjusted Pearson correlation of each chromosome with unseen high-coverage library (3.5B reads) is shown in Figure 2A. HiC2Self was trained with only low-coverage contact maps for each chromosome. Both low- and high-coverage maps are binned into 10kb resolution, and up to 1M distance from the diagonal are recovered using HiC2Self. Bar plot shows the mean and standard deviation of Pearson correlation from each genomic distance of chromosome 1-22. Red bars show the correlation between low- and high-coverage maps; green bars show the correlation of another high-coverage biological replicate (1.8B reads) with the ground truth (3.5B reads), and blue bars show the performance of HiC2Self recovered maps. Using a single library with less than 0.1% reads, HiC2Self is able to recovery comparable performance to biological library.

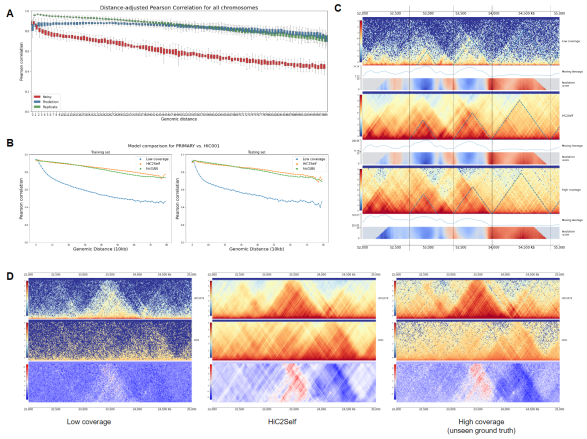

Figure 2. HiC2Self performance on bulk Hi-C

In order to validate our model framework and compare with previously published methods, we also trained our model using mean squared error on normalized data (log2-transformation followed by min/max rescaling to produce values between -1 and 1). The supervised model hicGAN was trained on 5,000 submatrices extracted from paired low-/high-coverage Hi-C data, with chromosome 3, 8, 12

held out for testing. We again use Pearson correlation per genomic distance with high-coverage data as the metric for evaluation and found comparable performance to hicGAN (Figure 2B).

We also use TAD caller TopDom (Shin et al., 2016) to evaluate the performance of recovered maps on downstream analysis. Figure 2C shows the comparison of low-coverage map (top), HiC2Self recovered map (middle) and high coverage-map (bottom). Dashed blue lines are showing the TAD calls from each contact map. HiC2Self provides consistent results with the high-coverage library.

Since HiC2Self is a self-supervised denoising method, it could be easily applied to different cell types without generalization problems. Figure 2D shows a comparison of GM12878 (top row in each panel), K562 (middle row), and the absolute difference between the two cell types (bottom row). Low-coverage data are shown on the left, with HiC2Self recovery in the middle, and high-coverage on the right. We can see from the absolute difference map that HiC2Self can differentiates well for different cell types.

3.3. Denoising on single-cell Hi-C

In additional to bulk Hi-C libraries, we also explored the application of HiC2Self on single-cell Hi-C data. Figure 3 shows the single-cell contact map of chromosome 1 from an example cell. From left right: 1Mb resolution, sum-pooling to 2Mb resolution, further sum-pooling to 4Mb resolution, and 8Mb resolution. Figure 3A shows the raw contact map of the example cell at four resolutions, and Figure 3B shows the corresponding HiC2Self recovered maps.

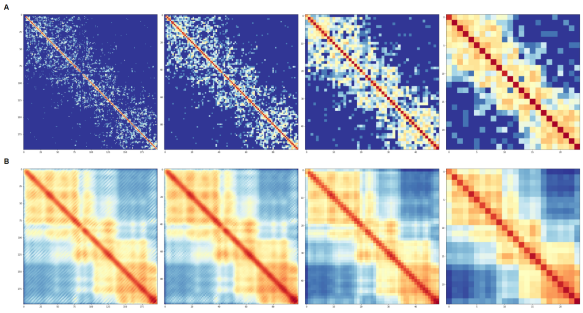

Figure 3. HiC2Self recovery of single-cell Hi-C

By training with multiple resolution, HiC2Self is able to capture both local and global structures from the single-cell contact map, and recovery signals.

Model availability

Data preparation pipeline and model scripts are available at github.com/ruy204/HiC2Self.

References

Batson, J. D. and Royer, L. A. Noise2self: Blind denoising by self-supervision. *ArXiv*, abs/1901.11365, 2019.

Dimmick, M. C., Lee, L. J., and Frey, B. J. Hicsr: a hi-c super-resolution framework for producing highly realistic contact maps. *bioRxiv*, 2020. doi: 10.1101/2020.02.24.961714. URL https://www.biorxiv.org/content/early/2020/07/07/2020.02.24.961714.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., and Aiden, E. L. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell Systems*, 3(1):95–98, Jul 2016. ISSN 24054712. doi: 10.1016/j.cels.2016.07.002. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471216302198.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10 (1):390, January 2019. ISSN 2041-1723. doi: 10.1038/s41467-018-07931-2. URL https://www.nature.com/articles/s41467-018-07931-2.

Hong, H., Jiang, S., Li, H., Du, G., Sun, Y., Tao, H., Quan, C., Zhao, C., Li, R., Li, W., Yin, X., Huang, Y., Li, C., Chen, H., and Bo, X. Deephic: A generative adversarial network for enhancing hi-c data resolution. *PLOS Computational Biology*, 16(2):1–25, 02 2020. doi: 10.1371/journal.pcbi.1007287. URL https://doi.org/10.1371/journal.pcbi.1007287.

Liu, Q., Lv, H., and Jiang, R. hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics*, 35(14):i99–i107, 07 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz317. URL https://doi.org/10.1093/bioinformatics/btz317.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, Oct 2013. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature12593. URL https://www.nature.com/articles/nature12593.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, Dec 2014. ISSN 00928674. doi: 10.1016/j.cell.2014.11.021. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867414014974.

Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F., and Zhou, X. J. Topdom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Research*, 44(7): e70, Apr 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1505. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4838359/.

Song, T.-A., Chowdhury, S. R., Kim, K., Gong, K., Fakhri, G. E., Li, Q., and Dutta, J. Super-resolution pet using a very deep convolutional neural network. In *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*, pp. 1–2, 2018. doi: 10.1109/NSSMIC.2018.8824683.

Szabo, Q., Bantignies, F., and Cavalli, G. Principles of genome folding into topologically associating domains. *Science Advances*, 5(4):eaaw1668, April 2019. ISSN 2375-2548. doi: 10.1126/sciadv.aaw1668. URL https://www.science.org/doi/10.1126/sciadv.aaw1668.

Zhang, Y., An, L., Xu, J., Zhang, B., Zheng, W. J., Hu, M., Tang, J., and Yue, F. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nature Communications*, 9(1):750, February 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03113-2. URL https://www.nature.com/articles/s41467-018-03113-2.