HiC2Self: Self-supervised Hi-C contact map denoising

Rui Yang
Sloan Kettering Institute
ruy4001@med.cornell.edu

Alireza Karbalayghareh
Sloan Kettering Institute
karbalaa@mskcc.org

Christina Leslie
Sloan Kettering Institute
cleslie@cbio.mskcc.org

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. We found our self-supervised method was competitive with or outperformed existing supervised Hi-C denoising algorithms while providing greater ease of use.

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 (1). 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.

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 (2) and HiCNN (3) use convolutional neural networks to predict high coverage 2D contact maps from low coverage or downsampled contact maps in the same cell type. hicGAN (4), DeepHiC (5) and HiCSR (6) 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 (7), with a mask structure and negative-binomial loss function designed for Hi-C raw count matrices.

2 Method

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.

Self-supervision framework Noise2Self (7) 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

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

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 training framework is shown in Figure 1A.

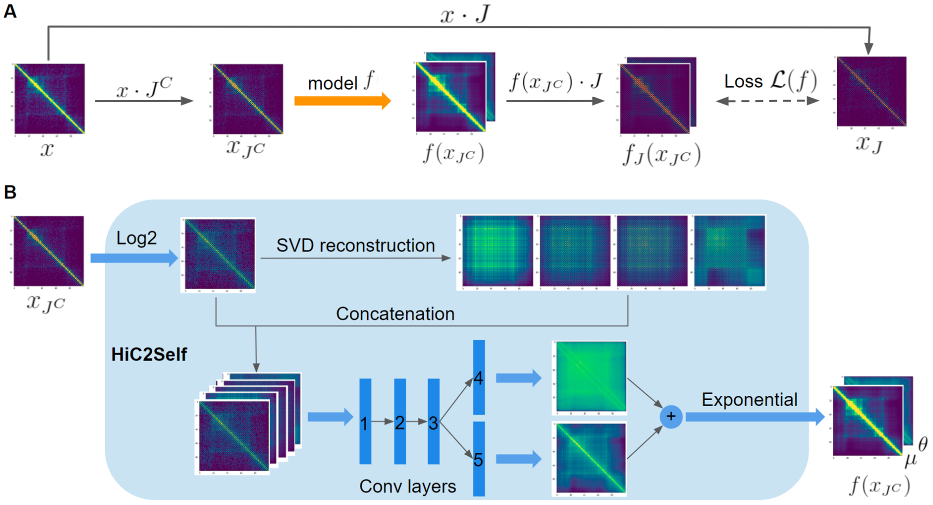

Figure 1: Training framework and model architecture

Model architecture HiC2Self uses a simple convolutional neural network (CNN), as shown in Figure 1B. 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 signal extracted from low-coverage submatrices, 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 (see Table 1). 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.

Layer	Type	Filter size	input dimension	output dimension	input channels	output channels	Activation function
1	Convolution	5×5	100×100	100×100	5	64	ReLU
2	Convolution	5×5	100×100	100×100	64	64	ReLU
3	Convolution	5×5	100×100	100×100	64	32	ReLU
4	Convolution	3×3	100×100	100×100	32	1	Exponential
5	Convolution	3×3	100×100	100×100	32	1	Exponential

Table 1: Structure of convolutional layers

Loss function Inspired by the deep count autoencoder (DCA) model for single cell data (8), 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

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

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.

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 (9) for downstream analysis.

3 Experiments and Results

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 (10)). 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 (9). Equal-sized (100×100) submatrices were extracted along the diagonal from intra-chromosomal low-coverage Hi-C contact maps to train the model.

Denoising on normalized data In order to validate our model framework and compare with previously published methods, we first trained our model (with necessary changes) 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 use Pearson correlation (per genomic distance) with high coverage data as the metric for evaluation and found comparable performance to hicGAN (Figure 2).

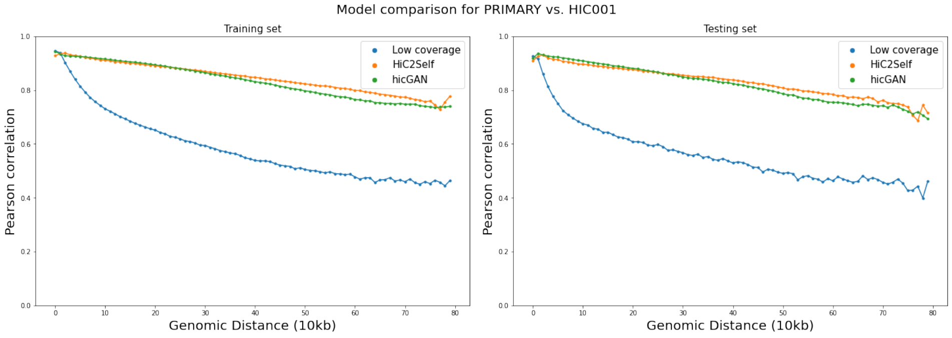

Figure 2: Performance on log-transformed data

Whole chromosome prediction Given the competitive performance on normalized data, we next trained our model with negative binomial loss on raw count data and produced predicted high coverage raw count contact maps. We generated denoised predictions within 1Mb distance from diagonal for chromosome 18 and multiplied by a scaling factor of 10 to increase the count range. The result was saved into .hic format and visualized using Juicebox (9) (Figure 3, color scale for the low-coverage matrix is 1/10 of the scale for denoised and high coverage matrices.)

We again used Pearson correlation by genomic distance to evaluate model performance on log2 transformed counts. For comparison, we downloaded another independent high-coverage pooled library GM12878 replicate with 3B reads. The correlation by genomic distance in Figure 3B show slightly better correlation than the biological replicate data.

We also ran HiC-DC+ (12) to call significant 3D interactions ($qvalue \leq 0.05$) on chromosome 18 (Figure 3C) and obtained good overlap with interactions identified on high-resolution Hi-C data.

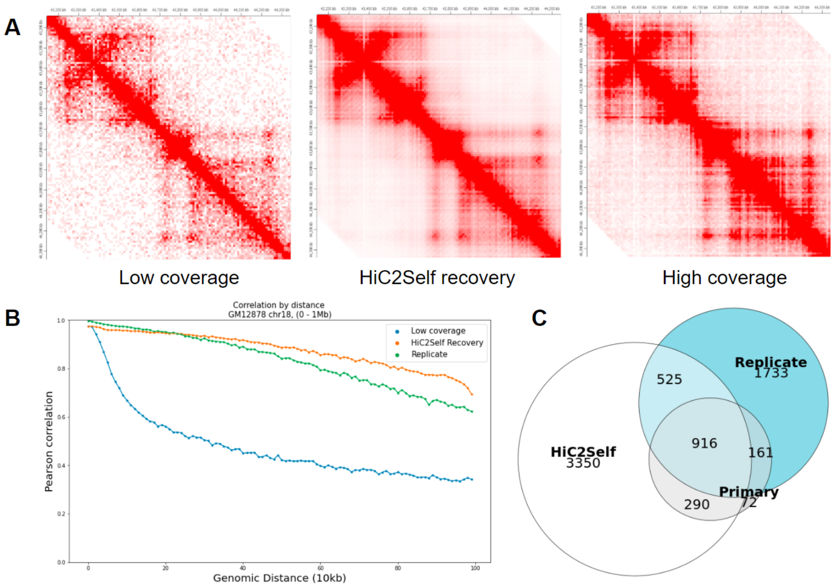

Figure 3: Performance on raw count data

4 Discussion

In this study, we developed HiC2Self, a self-supervised Hi-C contact map denoising model that achieves comparable performance with supervised Hi-C denoising methods without the requirement to train on paired low- and high-coverage data sets. Importantly, the model trains on unnormalized raw count data and produces high-coverage contact maps in count space, facilitating downstream analyses using Hi-C tools such as TAD and interaction callers.

We compared HiC2Self (with necessary changes) with existing supervised methods for denoising normalized Hi-C contact matrices and also assessed the usefulness of denoised read count contact matrices for downstream analyses. Interestingly, we found that adding SVD reconstructions of low-coverage matrices as input channels led to improved performance, and indeed our self-supervised model was competitive with a state-of-the-art supervised denoising method. Potentially our SVD reconstruction channels might improve supervised approaches as well. In additional experiments (not described above), we found that the generalizability of supervised models depended strongly on matching the low-/high sequencing coverage relationship in training data to the test data. A mismatch between training and test sequencing coverage scenarios led to poor performance, suggesting some inflexibility in the model for generalization. Our self-supervised model avoids this challenge of generalization and showed robust performance across data sets. In the raw count space comparison, our model recovered a majority of the significant interactions identified by high-resolution Hi-C data, showing its capacity as a valid denoising tool for downstream analysis. We will continue working on the evaluation of model performance and analysis of results in future work.

References

[1] Szabo, Q., Bantignies, F., Cavalli, G., et al. Principles of genome folding into topologically associating domains. Science Advances. 2019;5(4):eaaw1668. doi:10.1126/sciadv.aaw1668

[2] Zhang, Y., An, L., Xu, J., et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. Nature Communications. 2018;9:750. doi:10.1038/s41467-018-03113-2

[3] Liu, T., Wang, Z., et al. HiCNN: a very deep convolutional neural network to better enhance the resolution of Hi-C data. Bioinformatics. 2019;35(21):4222-4228. doi:10.1093/bioinformatics/btz251

[4] Liu, Q., Lv, H., Jiang, R. hicGAN infers super resolution Hi-C data with generative adversarial networks. Bioinformatics. 2019;35(14):i99-i107. doi:10.1093/bioinformatics/btz317

[5] Hong, H., Jiang, S., Li, H., et al. DeepHiC: A generative adversarial network for enhancing Hi-C data resolution. PLoS Computational Biology. 2020;16(2):e1007287. doi:10.1371/journal.pcbi.1007287

[6] Dimmick, M., Lee, L., Frey, B. HiCSR: a Hi-C super-resolution framework for producing highly realistic contact maps. bioRxiv - Genomics. 2020. doi:10.1101/2020.02.24.961714

[7] Batson, J., Royer, L. Noise2Self: Blind Denoising by Self-Supervision. arXiv - cs.CV. 2019;1901.11365. arXiv:1901.11365

[8] Eraslan, G., Simon, L.M., Mircea, M. et al. Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications. 2019;10(390). doi:10.1038/s41467-018-07931-2

[9] Durand, N., Shamim, M., Machol, I., et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Systems. 2016;3(1):95-98. doi:10.1016/j.cels.2016.07.002

[10] Rao, S., Huntley, M., Durand, N., et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-80. doi: 10.1016/j.cell.2014.11.021

[11] Larsson, J. eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. R package. 2020.

[12] Sahin, M., Wong, W., Zhan, Y., Van Deynze, K., Koche, R., Leslie, C.S. HiC-DC+: systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. In preparation.