HiC2Self: Self-Supervised Hi-C Contact Map Denoising

Rui Yang [1,2] Alireza Karbalayghareh [1] Christina Leslie [1]

[1]Memorial Sloan Kettering Cancer Center [2]Tri-Institutional Computational Biology Program

Abstract

Hi-C is a genome-wide chromatin conformation capture assay that is used to study 3D genomic organization. High-resolution Hi-C conact maps require generation of Semultiple replicate libraries and extremely high sequencing coverage (1-2B reads), incurring considerable costs. 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.

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 $\mathrm{J} \in \mathcal{J}$ and its complement $\mathrm{J^C}$.

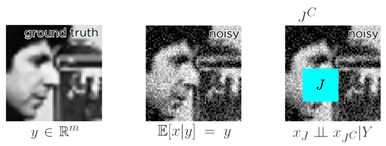

Figure 1:Input dimension partition and assumptions

Two assumptions:

- Zero-mean noise: $\mathbb{E}[\mathrm{x}|\mathrm{y}] = \mathrm{y}$
- Conditional independence: $\mathrm{x_J} \perp\!\!\!\perp \mathrm{x_{J^C}}|\mathrm{Y}$

$\mathcal{J}$-invariant function $\mathrm{f} : \mathbb{R}^\mathrm{m} \to \mathbb{R}^\mathrm{m}$, $\mathrm{f(x)_J} \perp\!\!\!\perp \mathrm{x_J}$ for every $\mathrm{J} \in \mathcal{J}$.

With this $\mathcal{J}$-invariant function f, the ordinary denoising loss function defined as

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

could be simplified to

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

Mask design The upper and lower tri-angle of Hi-C contact maps are symmetric. Masks are used for $\mathcal{J}$-invariant partition. To satisfy the independence assumption, we designed diagonal-symmetric masks.

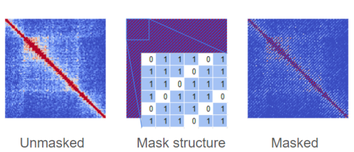

Figure 2:Mask designed for Hi-C contact map

HiC2Self

HiC2Self HiC2Self is a self-supervised Hi-C contact map denoising method stemmed from Noise2Self framework. It consists of four components:

- SVD reconstruction for signal enhancement
- mask to achieve self-supervision
- multi-resolution training to capture global and local structures
- negative binomial loss to run on raw counts

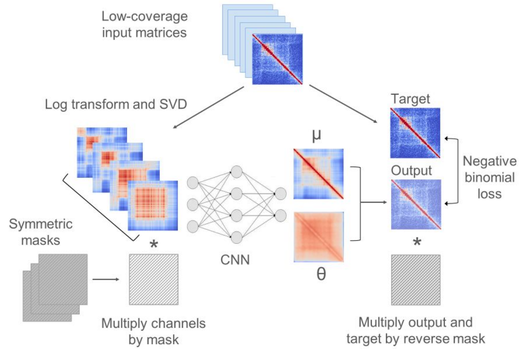

Figure 3:HiC2Self training framework

Application with Bulk Hi-C Libraries

Application with bulk library HiC2Self effectively reconstructs 3D structures from low-coverage Hi-C libraries.

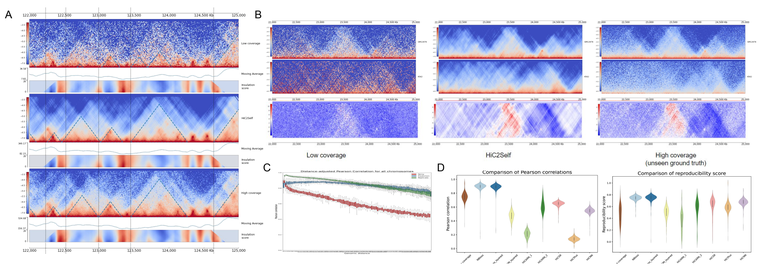

Figure 4:HiC2Self reconstruction of bulk Hi-C libraries.

(A) Visual comparison of low-coverage input library (top), HiC2Self reconstruction (middle), high-coverage library that never been seen by the model (bottom). **(B)** HiC2Self can reconstruct cell-type specific 3D structures. Top to bottom: low-coverage libraries (left) for GM12878, K562, difference between the two cell types. HiC2Self recovery (middle), and unseen high-coverage libraries (right). **(C)** Distance-stratified Pearson correlation with high-coverage library. **(D)** We benchmarked HiC2Self on a new dataset with several supervised models on a completely held-out chromosome (chromosome 4), and use Pearson correlation and Reproducibility score (Ursu, 2018) as evaluation metrics. HiC2Self outperformed supervised models on new dataset.

Application with Single-Cell Hi-C Libraries

Multi-modality single-cell assay has been a fast developing field. For example, sn-m3C-seq, a single-cell methylation and Hi-C co-assay, has been developed and used to study human brain development (Heffel, 2022).

Examine cell-type specific structures at pseudo-bulk level. We focused on studying cell-type specific 3D structure around RORB gene, an important gene locus related to schizophrenia during human brain development. Cells are clustered and annotated based on the methylation profile, and we randomly pick 50 cells from each cell type to create pseudo-bulk Hi-C libraries at 10kb resolution. With HiC2Self, we are able to examine the cell-type specific 3D structure at good resolution for rare cell types.

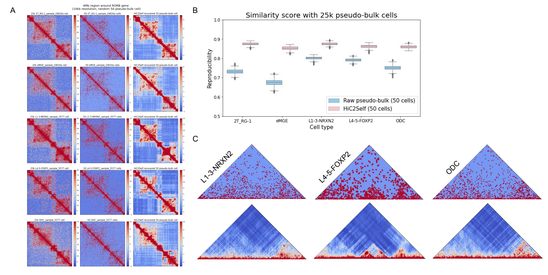

Figure 5:HiC2Self application with 50-cell pseudo-bulk Hi-C.

(A) (Left) 25k cell pseudo-bulk Hi-C at 10kb resolution, 4 mega-base around RORB gene locus. (Middle) 50 cell pseudo-bulk Hi-C. (Right) HiC2Self recovery on the 50-cell pseudo-bulk level. **(B)** Reproducibility score (Ursu, 2018) for 50-cell pseudo-bulk vs. 25k pseudo-bulk (blue), and HiC2Self recovery vs. 25 pseudo-bulk (pink). **(C)** Zoomed-in examination of the RORB locus, and observed unique structure at L4-5-FOXP2 cell stage.

Examine cell-specific 3D structures. We further extended HiC2Self to single-cell Hi-C reconstruction (LRIG1 gene locus, 50kb resolution).

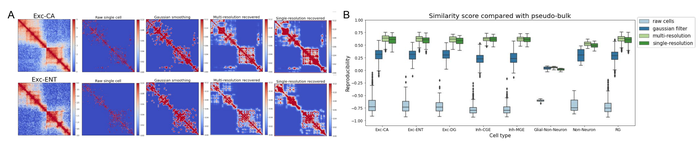

Figure 6:HiC2Self application with single-cell Hi-C.

(A) Two example cells. Left to right: pseudo-bulk of the cell type, raw single cell, gaussian smoothing of the cell, HiC2Self recovery using multi-resolution training, HiC2Self recovery using single-resolution training. **(B)** Reproducibility score (Ursu, 2018) of the raw single cell and recoveries compared with pseudo-bulk.

Github and Abstract

github abstract