HiC2Self: self-supervised denoising for bulk and single-cell Hi-C contact maps

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

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

HiC2Self

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

- SVD reconstruction for signal enhancement
- Mask to achieve self-supervision
- Negative-binomial loss to run on raw counts

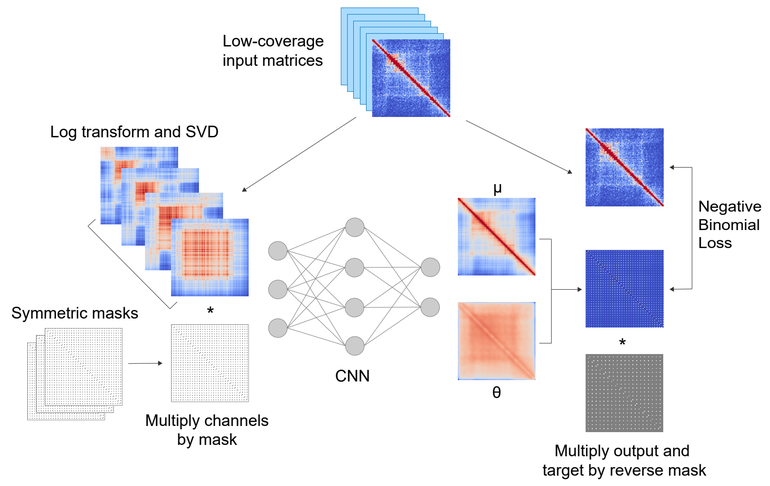

Figure 1:HiC2Self training framework

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 $\mathrm{J^C}$.

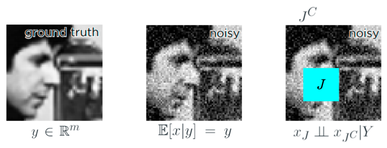

Figure 2:Input dimension partition and assumptions

Two assumptions:

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

$\mathcal{J}$-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

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

- 2E $\langle \mathbb{E}(\mathrm{f(x)} - \mathrm{x}), \mathbb{E}(\mathrm{x} - \mathrm{y}) \rangle$

can be simplified to

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

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 diagonally symmetric masks.

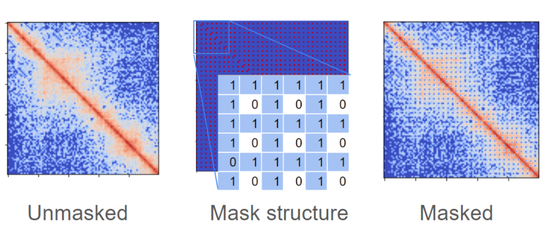

Figure 3:Mask designed for Hi-C contact map

Application with Bulk Hi-C Libraries

HiC2Self can denoise bulk Hi-C and Micro-C libraries, and recovery meaningful structures such as TADs and significant loops.

Application with bulk Hi-C library

- Hi-C library (GM12878) at 10kb resolution

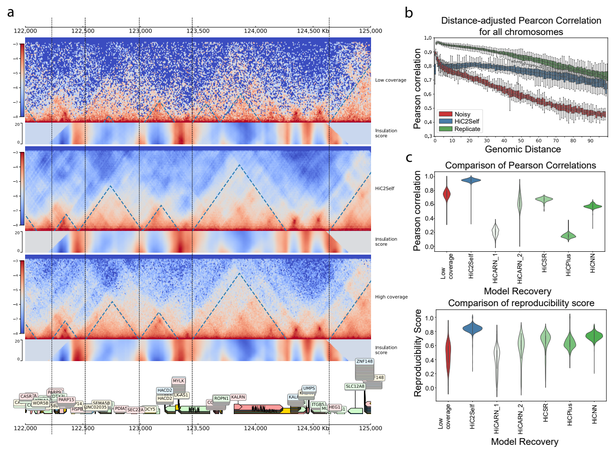

Figure 4:HiC2Self reconstruction of bulk Hi-C libraries.

(a) Top to bottom: low-coverage library (180.3 million reads), HiC2Self recovery, high-coverage library (3.5 billion reads). **(b)** Distance-adjusted Pearson correlation (HiC2Self in green). **(c)** Benchmark with supervised Hi-C denoising methods: HiCARN, HiCSR, HiCPlus, HiCNN.

Ultra-high resolution with Micro-C

- Micro-C mES data at 1kb resolution
- HLA region on chr6
- HiC2Self runs on raw counts

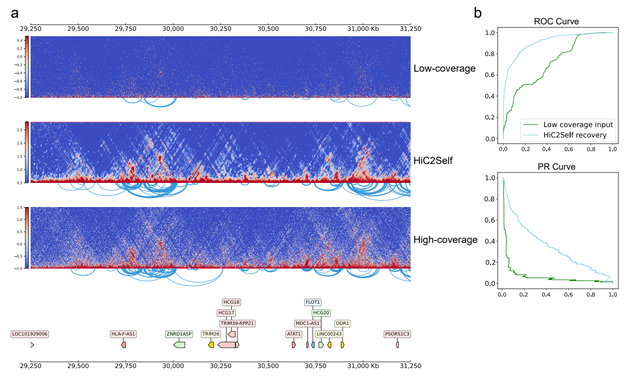

Figure 5:HiC2Self reconstruction of ultra-high resolution Micro-C library at 1kb resolution.

(a) Top to bottom: low-coverage library (188.8million reads), HiC2Self reconstruction, high-coverage library (3.4 billion reads, unseen by the model). Arcs show significant interactions called from each map. **(b)** ROC curve of the significant interactions.

Application with Single-Cell Hi-C Library

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).

Cell-cluster specific structures at pseudo-bulk level.

- 50-cell pseudo-bulk at 10kb resolution
- Cells pre-clustered by methylation profile
- 4Mb region around *RORB* gene, related to schizophrenia during human brain development.

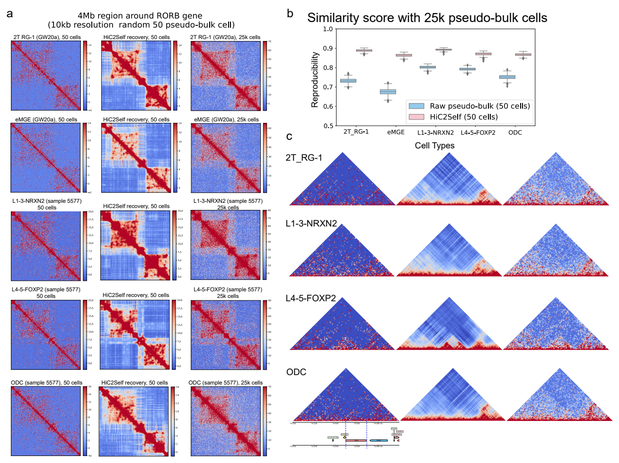

Figure 6:HiC2Self application with 50-cell pseudo-bulk Hi-C.

(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. **(b)** Reproducibility score (Ursu, 2018) for 50-cell pseudo-bulk (blue), and HiC2Self recovery (pink) vs. 25k pseudo-bulk. **(c)** Zoomed-in examination of the *RORB* locus, and HiC2Self recovers an unique structure at L4-5-FOXP2 cell stage.

Examine cell-specific 3D structures around genes.

- Single-cell Hi-C at 50kb resolution
- 20Mb region around *DLG2* gene
- Only one cell type (Exc-CA cells)

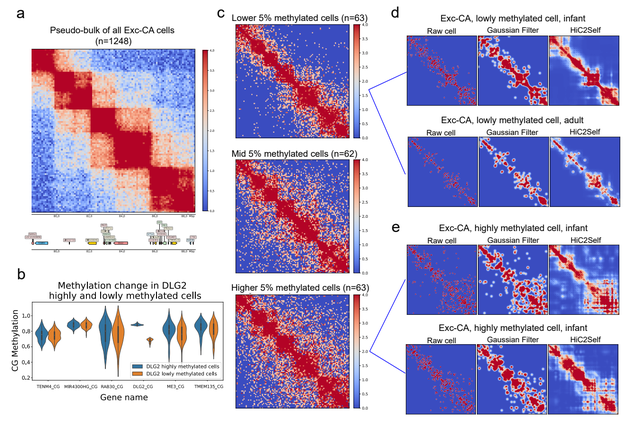

Figure 7:HiC2Self application with single-cell Hi-C.

(a) Pseudo-bulk of Exc-CA cells, 20Mb around *DLG2* gene at 50kb resolution. **(b)** Methylation around *DLG2* and nearby genes. **(c)** Pseudo-bulk of top 5%, middle 5% and bottom 5% methylated cells of *DLG2* gene. **(d)** HiC2Self recovery of two random cells sampled from the lower 5% methylated cells. **(e)** HiC2Self recovery of two random cells sampled from top 5% methylated cells.

Reference and Github

References

- Noise2Self. Batson Royer, 2019.
- Reproducibility score (GenomeDISCO). Ursu, 2018.
- sn-m3C-seq. Heffel, 2022.

QR Code

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

github