hicGAN infers super resolution Hi-C data with generative adversarial networks

Qiao Liu[1], Hairong Lv[2] and Rui Jiang[1,*]

[1]Ministry of Education Key Laboratory of Bioinformatics; Bioinformatics Division, Beijing National Research Center for Information Science and Technology; Department of Automation, Tsinghua University, Beijing 100084, China.

*To whom correspondence should be addressed.

Contents

Supplementary Notes

Supplementary Note. S1. The limitations of HiCPlus and comparison to hicGAN.

In our paper, the primary purpose of our work is to enhancing resolution of the low resolution Hi-C data to high resolution Hi-C data, which is essentially an image-enhancing problem. HiCPlus [1] is the only previous work so far that applies a convolutional neural network (CNN) for enhancing the resolution of Hi-C data by minimizing the mean squared error (MSE) between generated Hi-C data and real high resolution Hi-C data. However, it still has three major limitations.

First, HiCPlus takes MSE, one of the widely used pixel-wise measurements, as the objective function. Previous studies about generative models, especially in vision, have already demonstrated that using L_2 loss function, such as MSE, tends to yield blurry images [2-5]. The Hi-C samples generated by HiCPlus were indeed blurry compared to the high resolution Hi-C data in the original paper, which may result in losing some importance structure information. The MSE is not recommended as an ideal objective function in the task of image synthesis and image super-resolution [2-5]. To avoid this, our hicGAN model does not optimize any pixel-wise measurement such as MSE. Instead, we introduced another discriminator network to help discern the generated Hi-C data from real high resolution Hi-C data. Hi-C images generated by hicGAN are be more realistic compared to Hi-C images generated by HiCPlus.

Second, the network architecture of HiCPlus is a convolutional neural network that contains three convolutional layers and a fully-connected layer. The input of HiCPlus will go through three convolutional layers and then be flattened as a fixed dimensional vector. Due to the existence of the fully-connected layer, the input size is fixed in both training and test processes. If one needs to enhance the resolution of Hi-C data within a relatively large genomic region, HiCPlus has to divide the large genomic region into small patches and enhance each patch respectively. Then the enhanced patches need to be reconstructed again. It is not user-friendly at all. The generator network of our hicGAN model is a fully convolutional network without any fully-connected layer. Our model has the ability to enhance any size of the insufficient sequenced Hi-C sample, which is much convenient for an enhancing task.

Third, also the most important is that HiCPlus applied no normalization to the raw Hi-C raw contacts count. So it is quite sensitive to the sequencing depth of the Hi-C data. For example, when low resolution Hi-C data in the training process and the low resolution Hi-C data in the test process have different sequencing depth, HiCPlus typically performs badly. This severely restricts the generalization and wide use of HiCPlus. In our model hicGAN, we eliminated the effect of sequencing depth by designing a normalization procedure in details. The sequencing depths of different cell type vary a lot, our hicGAN still achieves superior performance in the cross-cell-type experiments.

Although HiCPlus is the pioneer work for enhancing the resolution of Hi-C data with a computational framework, it still contains some major limitations. Our hicGAN model overcomes the above limitations, thus can be considered as a new tool for processing Hi-C data.

Reference

[1]. Zhang, Yan, et al. "Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus." *Nature communications* 9.1 (2018): 750.

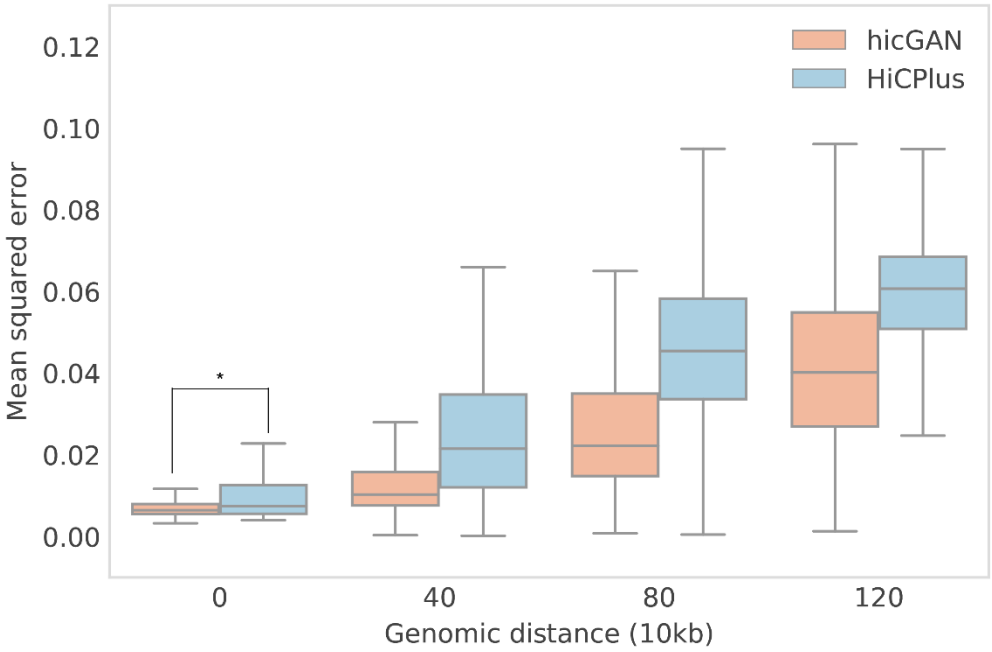
[2]. Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." *arXiv preprint arXiv*:1511.05440 (2015).

[3]. Pathak, Deepak, et al. "Context encoders: Feature learning by inpainting." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

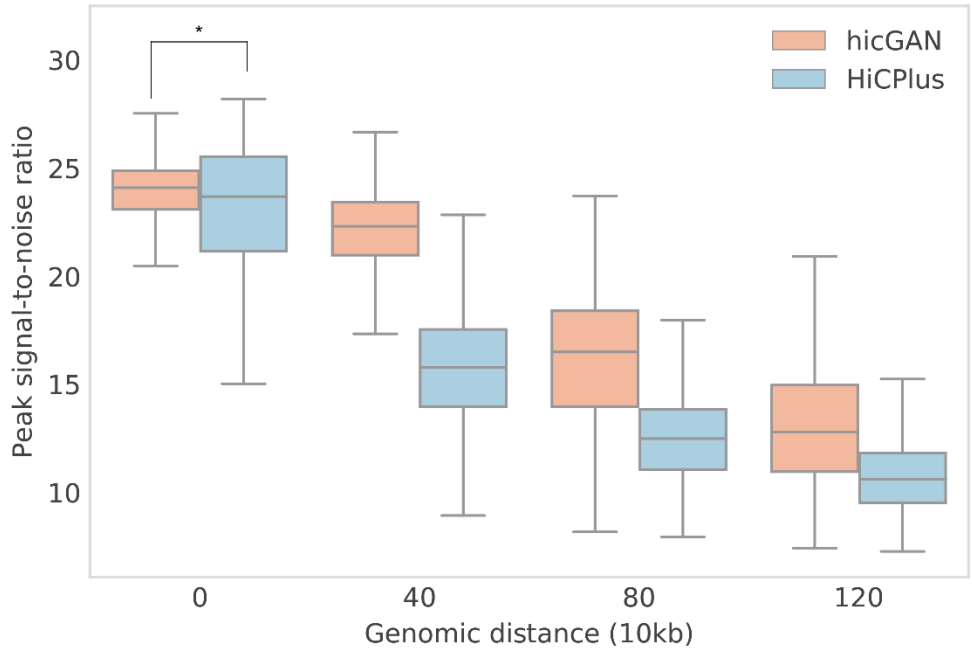
[4]. Zhang, Richard, Phillip Isola, and Alexei A. Efros. "Colorful image colorization." *European Conference on Computer Vision*. Springer, Cham, 2016.

[5]. Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

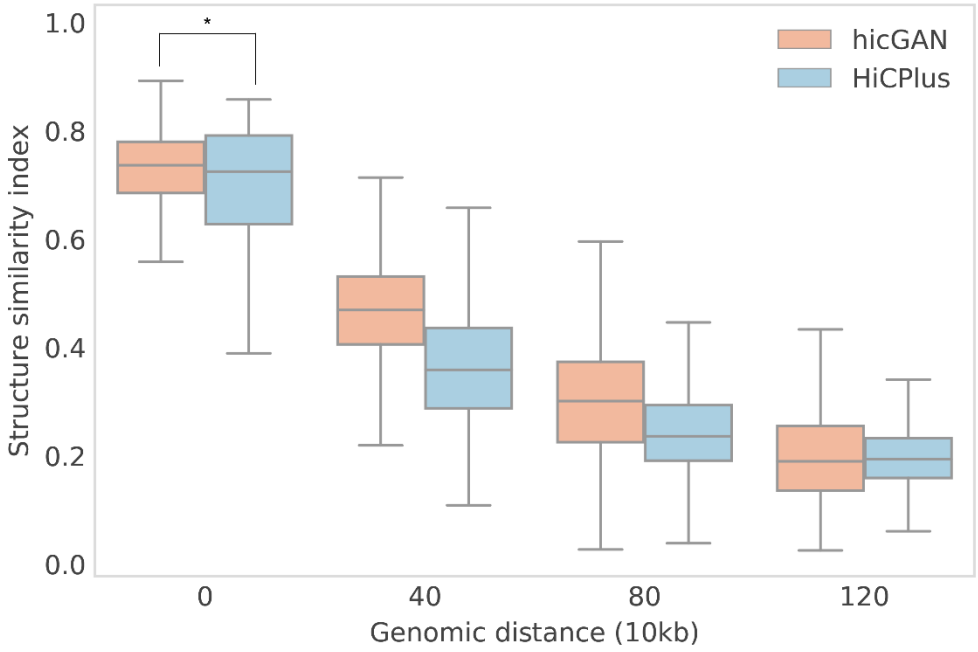
Supplementary Figures



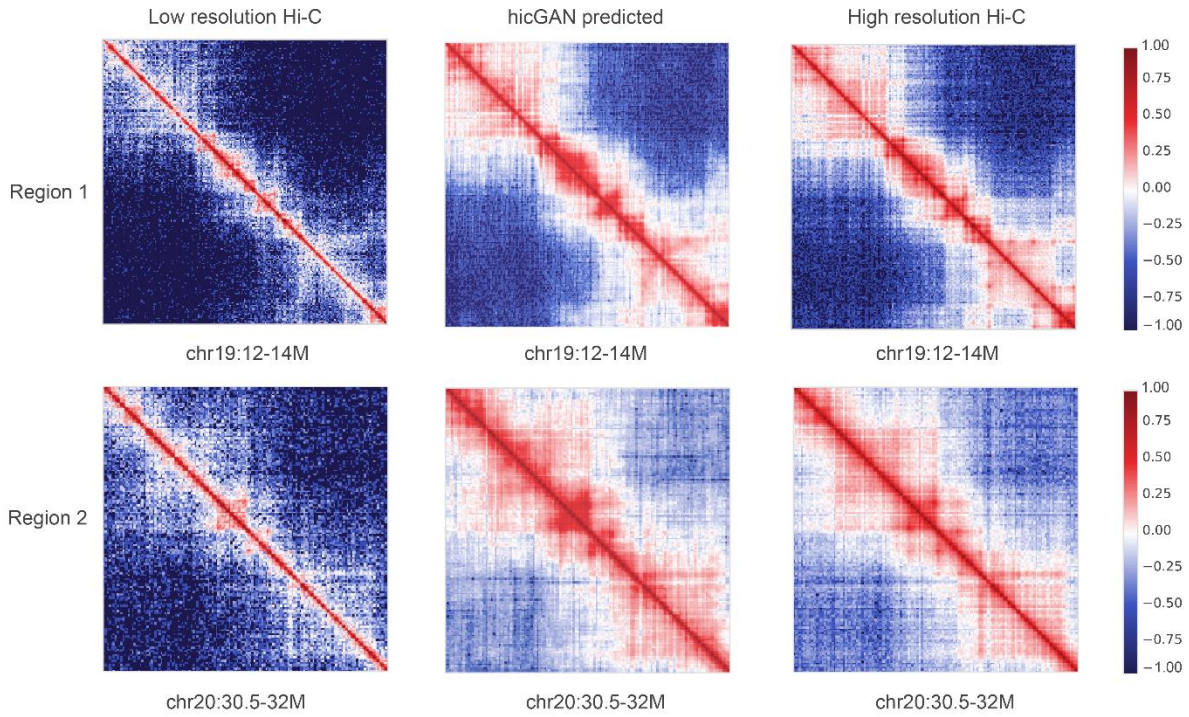
Supplementary Fig. S1. The performance of hicGAN and HiCPlus under different genomic distance considering the MSE measurement. At the genomic distance of 0 (diagonal Hi-C samples), hicGAN achieves an average MSE of 0.0078, compared to 0.0144 of HiCPlus (*One-sided Mann-Whitney U test, p-values=1.11×10^{-11}). Our hicGAN model outperforms HiCPlus by a significantly larger margin at a further genomic distance.



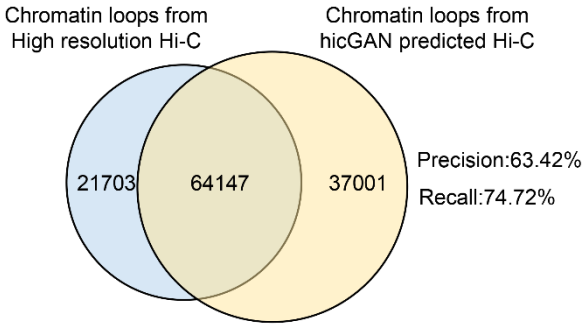
Supplementary Fig. S2. The performance of hicGAN and HiCPlus under different genomic distance considering the PSNR measurement. At the genomic distance of 0 (diagonal Hi-C samples), hicGAN achieves an average PSNR of 23.89 dB, compared to 22.92 dB of HiCPlus (*One-sided Mann-Whitney U test, *p*-values=0.0012). Our hicGAN model outperforms HiCPlus by a larger margin at a further genomic distance.



Supplementary Fig. S3. The performance of hicGAN and HiCPlus under different genomic distance considering the SSIM measurement. At the genomic distance of 0 (diagonal Hi-C samples), hicGAN achieves an average SSIM of 0.731, compared to 0.694 of HiCPlus (*One-sided Mann-Whitney U test, p-values=1.52×10^{-4}). Our hicGAN model outperforms HiCPlus by a larger margin at a further genomic distance.



Supplementary Fig. S4. We showed two examples of the Hi-C data predicted by hicGAN. We extracted two genomic regions (chr19:12-14M and chr20:30.5-32M) from down-sampled low resolution Hi-C data (left), Hi-C data predicted by hicGAN model (middle), high resolution Hi-C data (right), respectively. The Hi-C matrices predicted by hicGAN are similar to high resolution Hi-C matrices.



Supplementary Fig. S5. The Venn plot of the significant chromatin loops from high resolution Hi-C data and Hi-C data predicted by hicGAN model in K562 cell type using Fit-Hi-C software with a strict threshold (q-value<1e-06). Note that low resolution Hi-C data can only recover 3.73% of chromatin loops from high resolution Hi-C data.

Supplementary Tables

Supplementary Table. S1. The detailed hyperparameters of generator network of hicGAN model. The input and the output have exactly the same size. Each inner element-wise sum in a residual block (RB) will take a summation of the RB's input and the output of the RB's second convolutional layer in an element-wise manner. The outer element-wise sum will take a summation of the output of the first convolutional layer and the output of the previous layer in an element-wise manner. W and h are both set to 40 in the training process. No limitations on W and h in the test process. The filter size 3 is fixed by hyperparameter tuning.

Layer		Output shape	Configuration
Input		w × h × 1	
Convolution		w × h × 64	3 × 3 conv, stride=1,num=64
Residual Block 1	Convolution	w × h × 64	3 × 3 conv, stride=1,num=64
			Batch normalization
	Convolution	w × h × 64	3 × 3 conv, stride=1,num=64
			Batch normalization
	Element-wise Sum	w × h × 64	
Residual Block 2		w × h × 64	The same as above
…		…	…
Residual Block 5		w × h × 64	The same as above
Convolution		w × h × 64	3 × 3 conv, stride=1,num=64
			Batch normalization
Element-wise Sum		w × h × 64	
Convolution		w × h × 128	3 × 3 conv, stride=1,num=128
Convolution		w × h × 256	3 × 3 conv, stride=1,num=256
Convolution		w × h × 1	3 × 3 conv, stride=1,num=1

Supplementary Table. S2. The detailed hyperparameters of discriminator network of hicGAN model. It contains three convolutional blocks (CBs). The size will reduce by half after entering each CB. Then it is flattened before going through two fully-connected layers. The output is a single value which denotes the probability that the input sample is from real high resolution Hi-C data. The filter size 3 is fixed by hyperparameter tuning.

Layer		Output shape	Configuration
Input		$w \times h \times 1$	
Convolution		$w \times h \times 64$	3×3 conv, stride=1,num=64
			Leaky ReLu
Convolutional Block 1	Convolution	$w/2 \times h/2 \times 64$	3×3 conv, stride=2,num=64
			Leaky ReLu
			Batch normalization
	Convolution	$w/2 \times h/2 \times 64$	3×3 conv, stride=1,num=64
			Leaky ReLu
			Batch normalization
Convolutional Block 2		$w/4 \times h/4 \times 64$	The same as above
Convolutional Block 3		$w/8 \times h/8 \times 64$	The same as above
Flatten		$(w/8 \times h/8 \times 64) \times 1$	Flattened as a vector
Dense		512	Fully-connected, num=512
Dense		1	Fully-connected, num=1
			Sigmoid function

Supplementary Table. S3. Information of Hi-C datasets across four cell types used in our experiments. Each cell type contains multiple experiments, we pooled the aligned sequencing reads from each experiment together before data preprocessing.

Cell type		GEO accession	Raw contacts
GM12878 (total contacts: 2,634,479,637)	Experiment 1	GSM1551550	148,358,011
	Experiment 2	GSM1551551	232,985,572
	Experiment 3	GSM1551552	393,983,023
	Experiment 4	GSM1551553	130,829,245
	Experiment 5	GSM1551554	247,790,014
	Experiment 6	GSM1551555	126,776,812
	Experiment 7	GSM1551556	148,770,483
	Experiment 8	GSM1551557	162,374,252
	Experiment 9	GSM1551558	94,570,176
	Experiment 10	GSM1551559	46,510,588
	Experiment 11	GSM1551560	45,743,329
	Experiment 12	GSM1551561	144,362,539
	Experiment 13	GSM1551562	60,768,266
	Experiment 14	GSM1551563	221,733,196
	Experiment 15	GSM1551564	99,684,024
	Experiment 16	GSM1551565	98,965,718
	Experiment 17	GSM1551566	126,146,087
	Experiment 18	GSM1551567	104,128,302
K562 (total contacts: 932,208,867)	Experiment 1	GSM1551618	310,243,422
	Experiment 2	GSM1551619	389,060,442
	Experiment 3	GSM1551620	64,453,847
	Experiment 4	GSM1551621	52,564,216
	Experiment 5	GSM1551622	50,840,830
	Experiment 6	GSM1551623	65,046,110
IMR90 (total contacts: 1,136,673,301)	Experiment 1	GSM1551599	179,593,204
	Experiment 2	GSM1551600	199,657,173
	Experiment 3	GSM1551601	21,641,031
	Experiment 4	GSM1551602	94,002,885
	Experiment 5	GSM1551603	190,666,866
	Experiment 6	GSM1551604	213,069,892
	Experiment 7	GSM1551605	238,042,250
NHEK (total contacts: 664,899,299)	Experiment 1	GSM1551614	171,515,191
	Experiment 2	GSM1551615	196,697,990
	Experiment 3	GSM1551616	296,686,118

Supplementary Table. S4. *p*-values of two statistic tests between different methods. All the hypothesis tests are one-sided and performed based on the Pearson correlation coefficient (PCC) under different genomic distance (from 0 to 1Mb, 100 points in total). In the first test, we count the number of cell lines that our method hicGAN outperforms a HiCPlus and then perform a binomial exact test with the alternative hypothesis that the probability that our method outperforms the baseline is greater than 0.5. In the second test, we apply a Mann-Whitney U test with the alternative hypothesis that the PCCs achieved by our method have a positive shift when compared with those of HiCPlus.

Statistical test	hicGAN vs HiCPlus
Exact binomial test	2.2×10^{-16}
Mann-Whitney U test	7.9×10^{-07}

Supplementary Table. S5. *p*-values of two statistic tests between hicGAN models trained in different cell types and 2D Gaussian. All the hypothesis tests are one-sided and performed based on the Pearson correlation coefficient (PCC) under different genomic distance (from 0 to 1Mb, 100 points in total). In the first test, we count the number of cell lines that hicGAN outperforms 2D Gaussian and then perform a binomial exact test with the alternative hypothesis that the probability that hicGAN outperforms 2D Gaussian is greater than 0.5. In the second test, we apply a Mann-Whitney U test with the alternative hypothesis that the PCCs achieved by hicGAN have a positive shift when compared with those of 2D Gaussian. Note that the hicGAN models were all trained in GM12878. The cross-cell-type prediction performances only decreased slightly.

Statistical test	GM12878	K562	IMR90	NHEK
Exact binomial test	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
Mann-Whitney U test	8.6×10^{-20}	1.2×10^{-18}	7.5×10^{-16}	8.2×10^{-19}

Supplementary Table. S6. *p*-values of two statistic tests between hicGAN models trained in assembled cell types and single cell type (K562). All the hypothesis tests are one-sided and performed based on the mean squared error (MSE) under different genomic distance (from 0 to 1Mb, 100 points in total).

Statistical test	assembled cell types vs single cell type
Exact binomial test	3.7×10^{-08}
Mann-Whitney U test	1.5×10^{-05}