清华大学自动化系
Department of Automation

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

Qiao Liu, Hairong Lv and Rui Jiang

Ph.D. Candidate
Department of Automation, Tsinghua University,
Beijing, China
liu-q16@mails.tsinghua.edu.cn

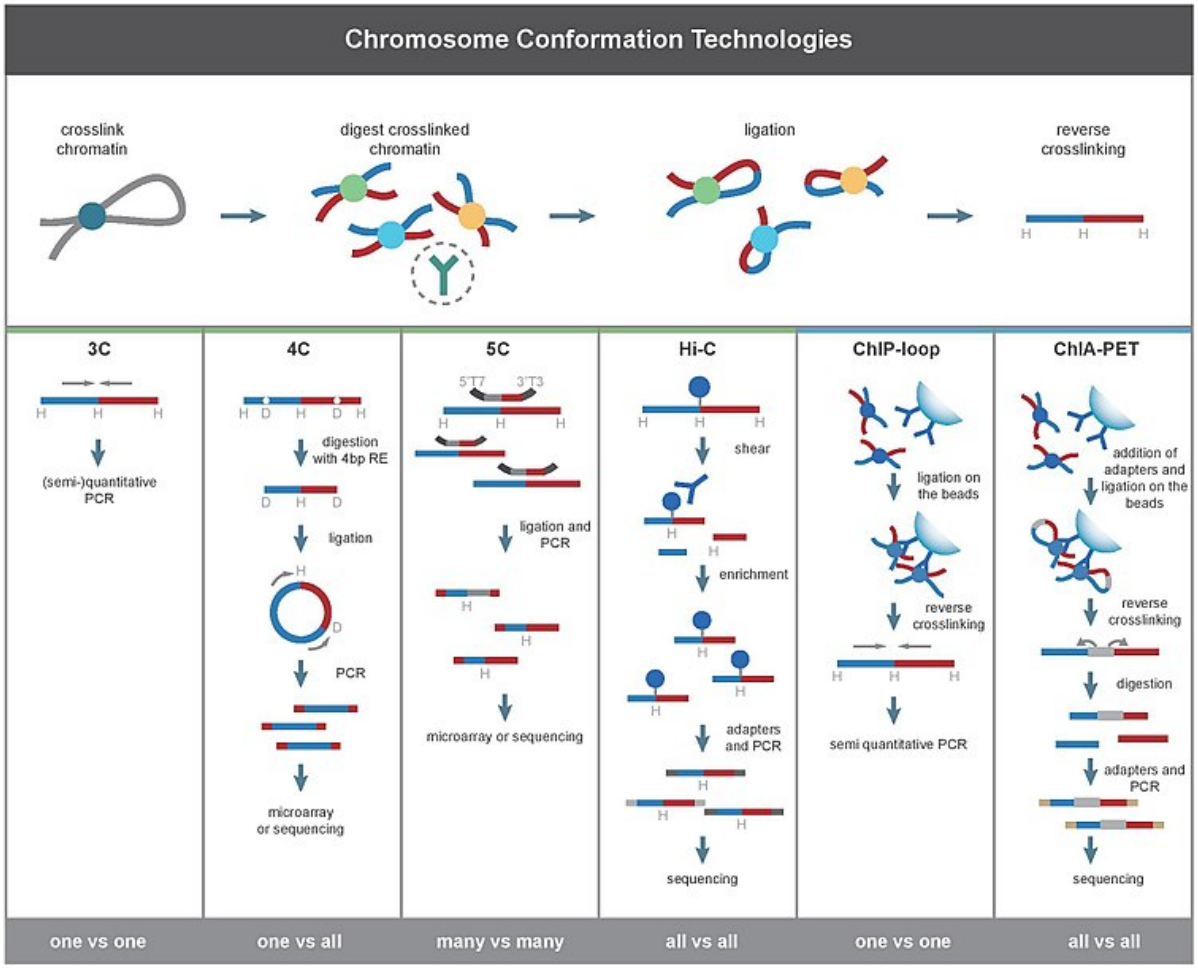
--ISMB/ECCB 2019

Outline

- Background
- Methods
- Results
- Conclusion

Development of Hi-C

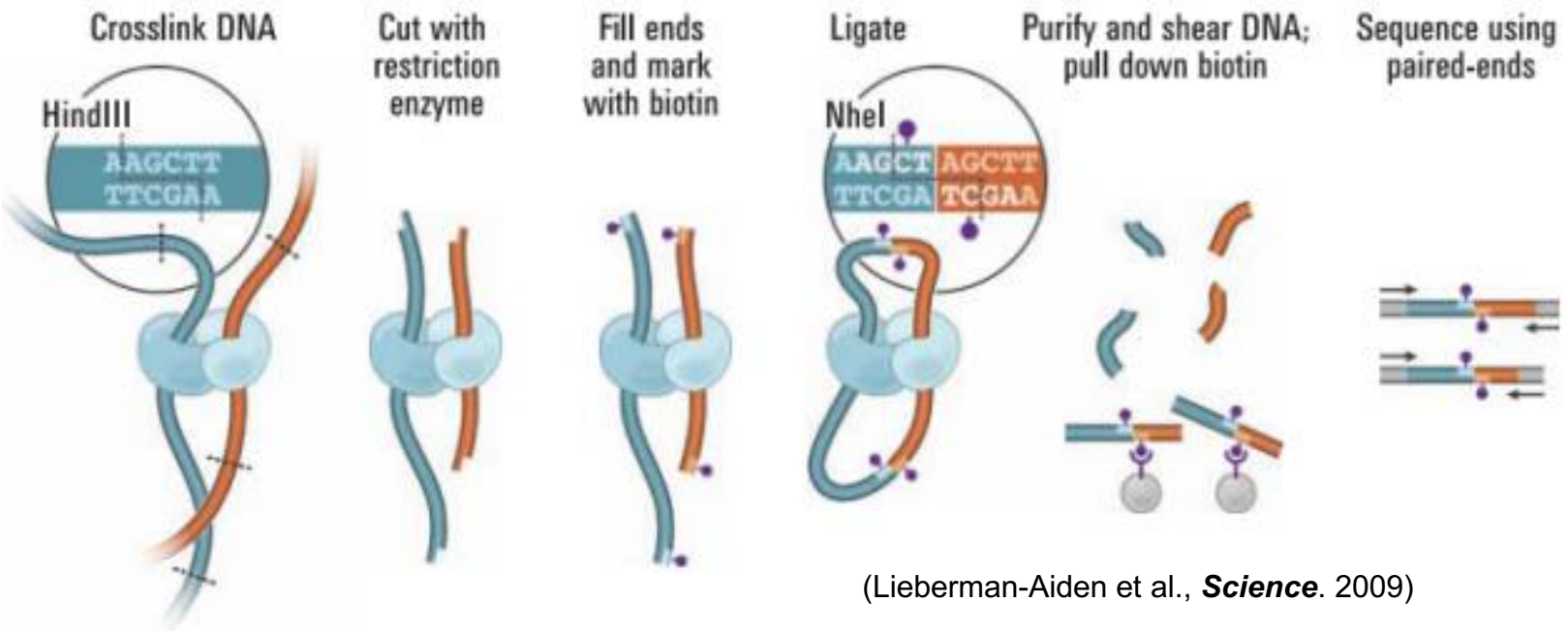
- Brief introduction to Hi-C data
  - Derived from 3C (Chromosome Conformation Capture) experiment



(Schwartzman et al., ***Nat Rev Genet***. 2015)

Hi-C experiment

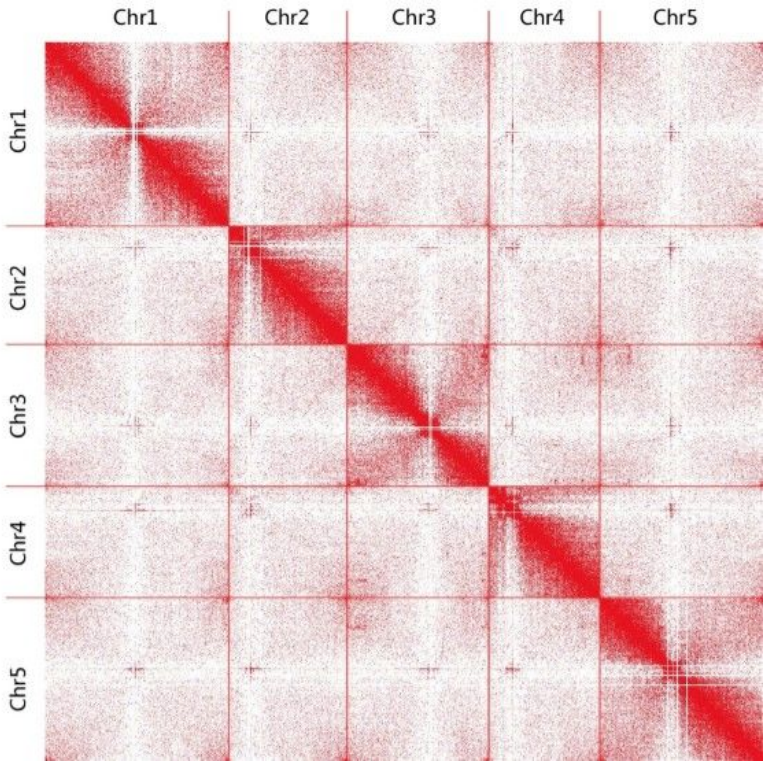
- Hi-C data reflect the genome-wide spatial contact of chromatin structure
- Hi-C data can be highly dynamic in different species, organs, tissues and cell types



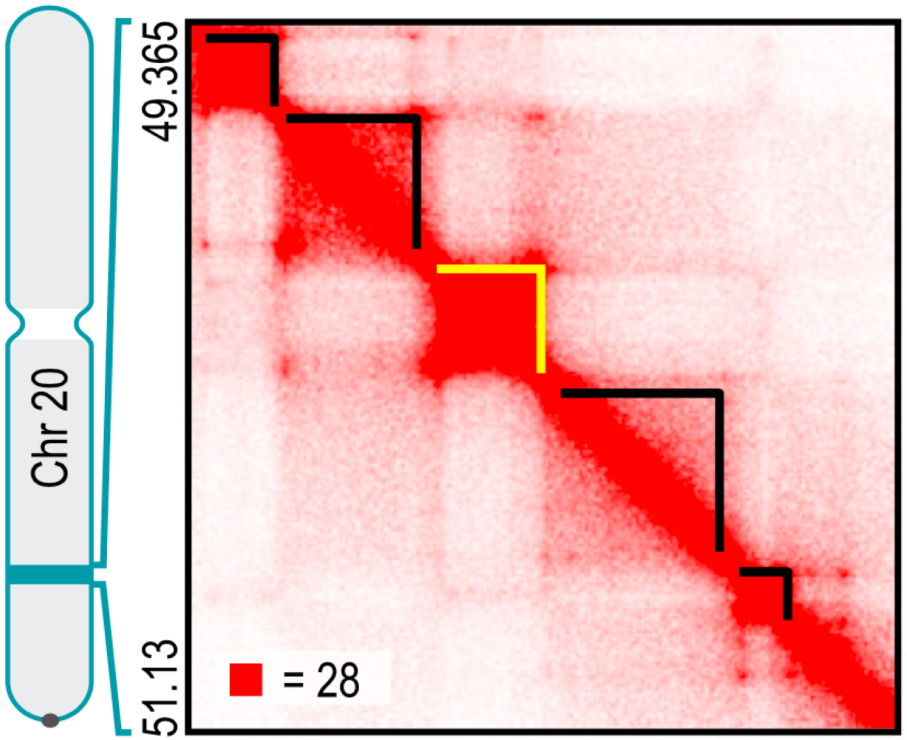
(Lieberman-Aiden et al., ***Science***. 2009)

Visualization of Hi-C

- Hi-C data can be represented as a symmetric matrix
  - All values are non-negative
  - Sparse and distributed around the diagonal
- The resolution of Hi-C data is defined as the bin size for constructing an appropriate Hi-C matrix



(Rao et al., ***Cell***. 2014)



(Rao et al., ***Cell***. 2014)

Motivation

- High resolution Hi-C data require deeper sequencing depth and cost more money
- The resolution of most Hi-C datasets are coarse due to sequencing cost
- The Hi-C resolution directly affects the results of downstream analysis such as predicting enhancer-promoter interactions or identifying TAD boundaries
- Therefore, it is urgent to develop a computational approach to infer high-resolution Hi-C data

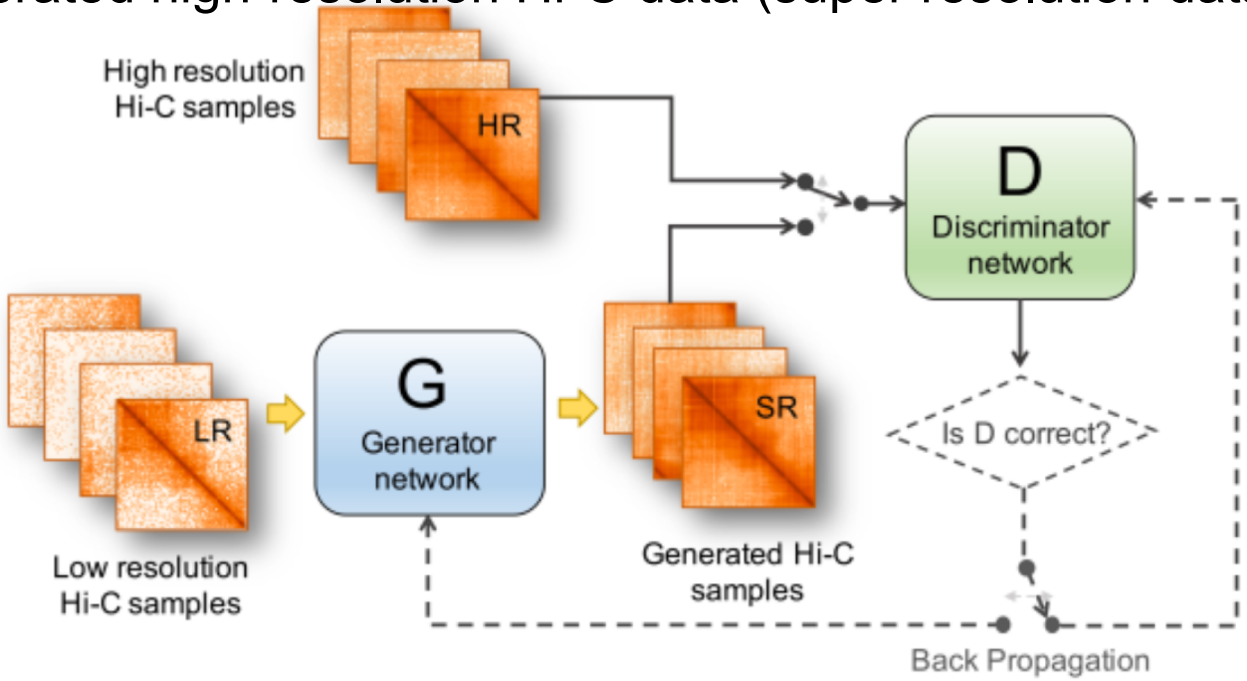
Behind hicGAN

- hicGAN was built based on a generative adversarial network (GAN, Goodfellow et al., 2014)
- GAN was first proposed for generating realistic images
- hicGAN was inspired from image super resolution task

Image Super Resolution using GAN

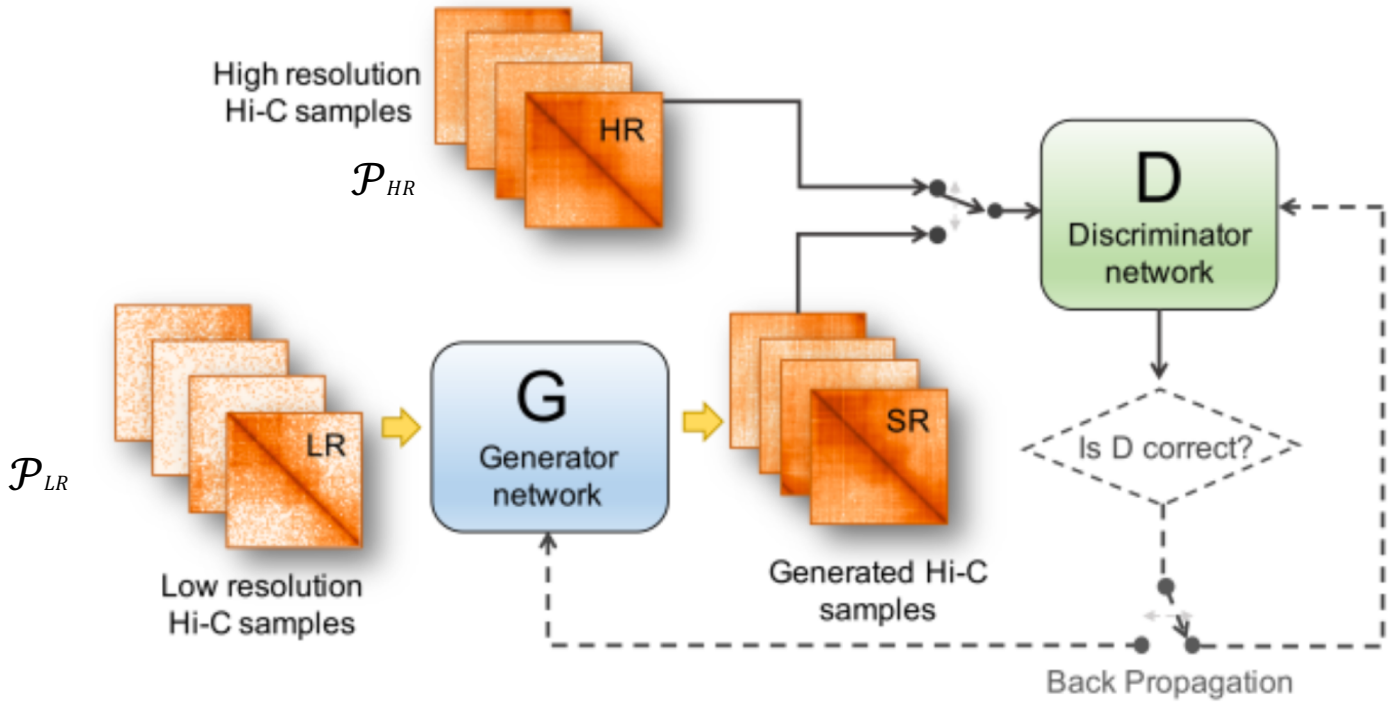
Model overview

- hicGAN is a GAN network combined with a generator (G) and a discriminator (D)
- Generator: generate super-resolution Hi-C data with low-resolution Hi-C data
- Discriminator: judge whether it is real high-resolution Hi-C data or generated high-resolution Hi-C data (super resolution data)



Low resolution Hi-C data can be obtained by randomly downsampling the sequencing reads by e.g. 1/16

Model overview

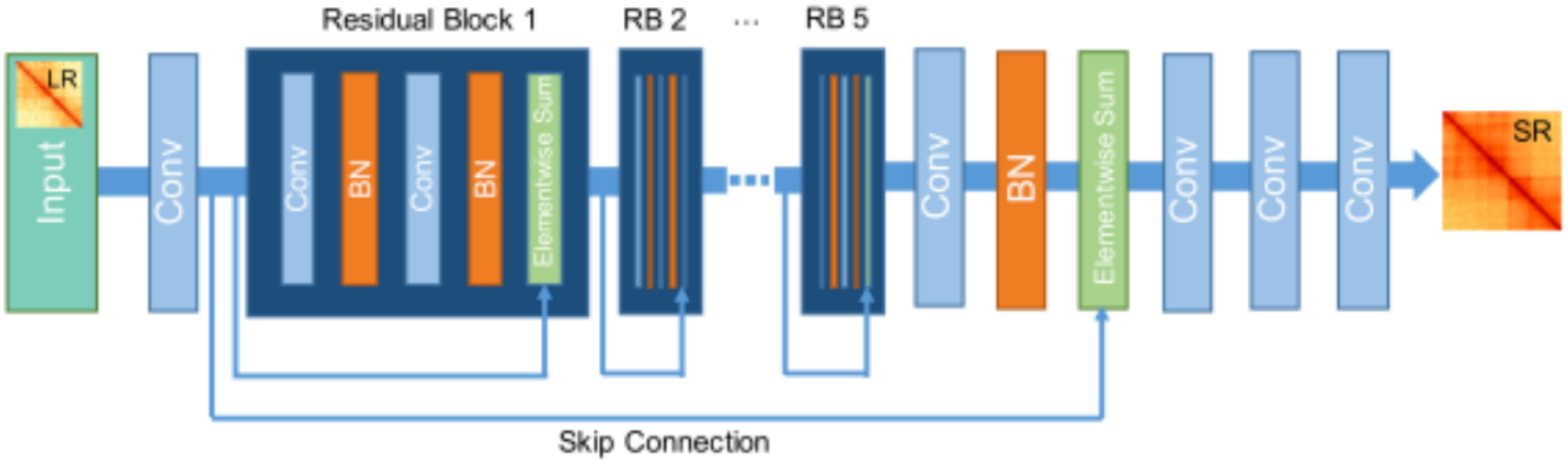


- We define generator network as $G_\theta(\cdot)$, parametrized by θ
- We define discriminator as $D_\omega(\cdot)$, parametrized by ω
- Min-max problem is defined as the following

$$\min_{\theta}\max_{\omega}\mathbb{E}_{\boldsymbol{x}_{HR}\sim\mathcal{P}_{HR}}\left[\log\left(D_\omega(\boldsymbol{x}_{HR})\right)\right]+\mathbb{E}_{\boldsymbol{x}_{LR}\sim\mathcal{P}_{LR}}\left[\log\left(1-D_\omega(G_\theta(\boldsymbol{x}_{LR}))\right)\right]$$

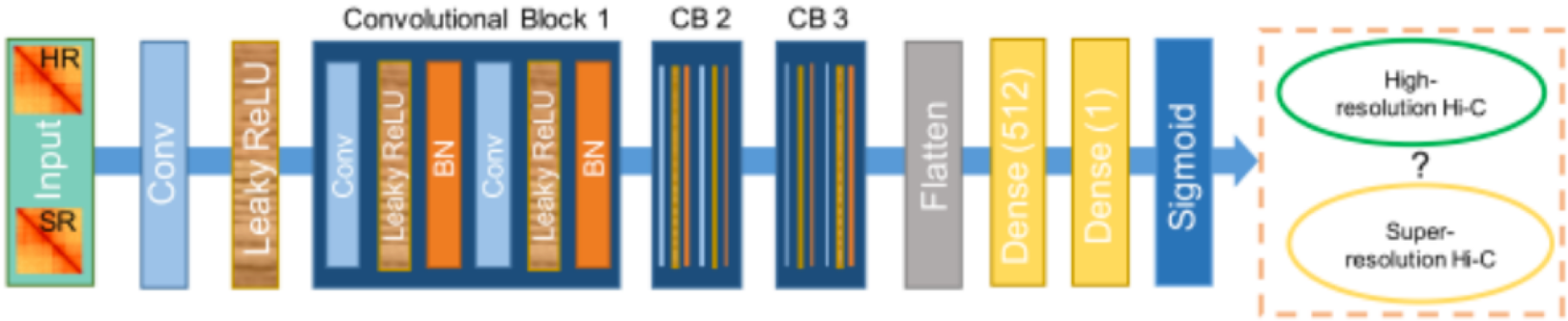
Model architecture

- Generator: generate super-resolution Hi-C data with low-resolution Hi-C data
- Generator network adopts a novel dual-stream residual architecture
- Note that Generator network is a fully convolutional network



Model architecture

- Discriminator: judge whether it is real high-resolution Hi-C data or generated high-resolution Hi-C data (super resolution data)
- The discriminator network is a deep convolutional neural network



Model training

- The adversarial training of hicGAN model
- In the training process, parameters of D and G are iteratively updated

Algorithm Adversarial training of hicGAN

Require: θ_0 for initial parameters of generator network, ω_0 for initial parameters of discriminator network, batch size m and learning rate α.

While θ has not converged, **do**

Sample $\{x_{HR}^{(i)}\}_{i=1}^{m} \sim \mathcal{P}_{HR}$ as a batch high resolution Hi-C data

Sample $\{x_{LR}^{(i)}\}_{i=1}^{m} \sim \mathcal{P}_{LR}$ as a batch low resolution Hi-C data

$$g_\omega \leftarrow \nabla_\omega \frac{1}{m} \sum_{i=1}^{m} [\log\left(D_\omega\left(x_{HR}^{(i)}\right)\right) + \log\left(1 - D_\omega\left(G_\theta\left(x_{LR}^{(i)}\right)\right)\right)]$$

$$\omega \leftarrow \omega + \alpha \cdot \mathrm{Adam}(\omega, g_\omega)$$

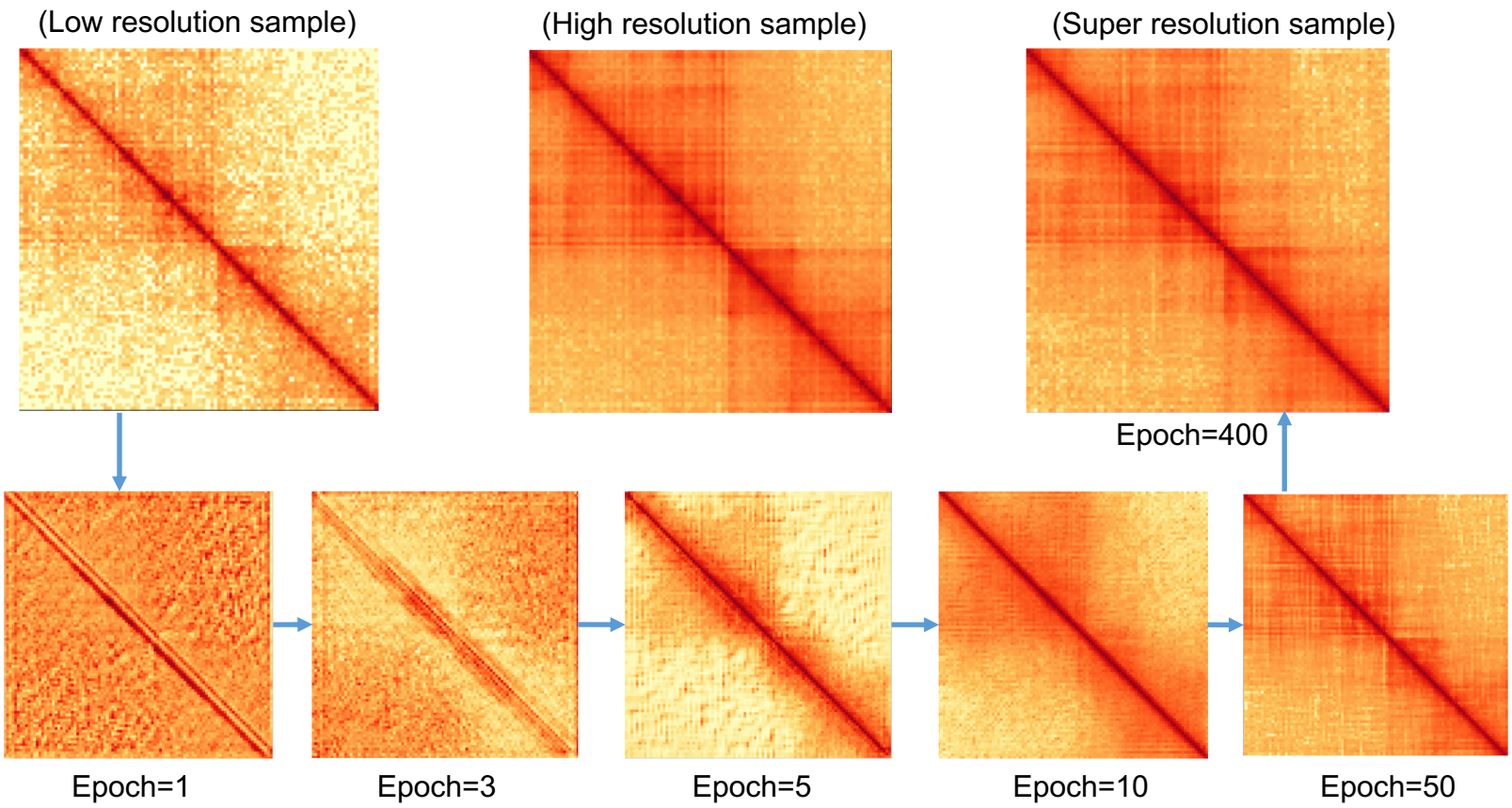
$$g_\theta \leftarrow \nabla_\theta \frac{1}{m} \sum_{i=1}^{m} \log(1 - D_\omega\left(G_\theta\left(x_{LR}^{(i)}\right)\right))$$

$$\theta \leftarrow \theta + \alpha \cdot \mathrm{Adam}(\theta, g_\theta)$$

end while

An training example

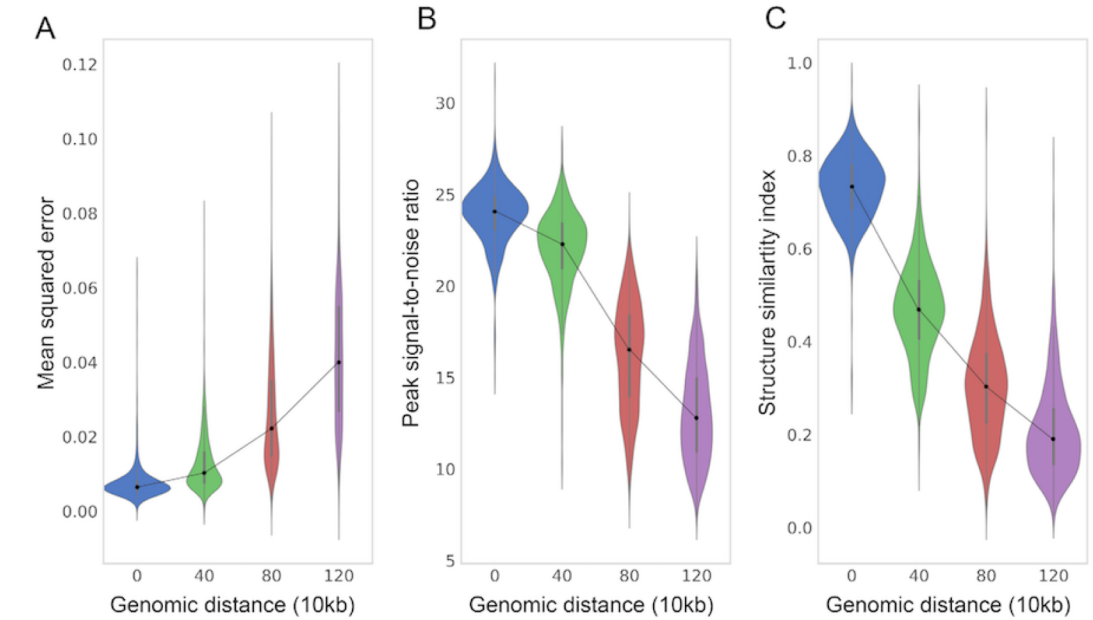
- Experiment: GM12878 Chr1-16 for training, Chr17-22 for testing



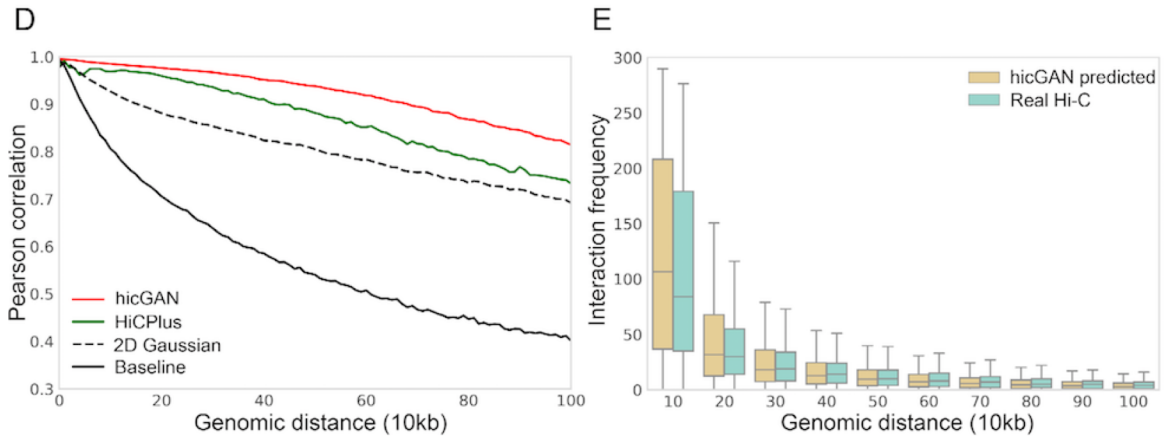
Test Genomic region Chr17:70M-71M from GM12878 cell line

Evaluation of super resolution Hi-C data

- Experiment: GM12878 Chr1-16 for training, Chr17-22 for testing
- Metrics: MSE, PSNR, SSIM and Pearson's correlation

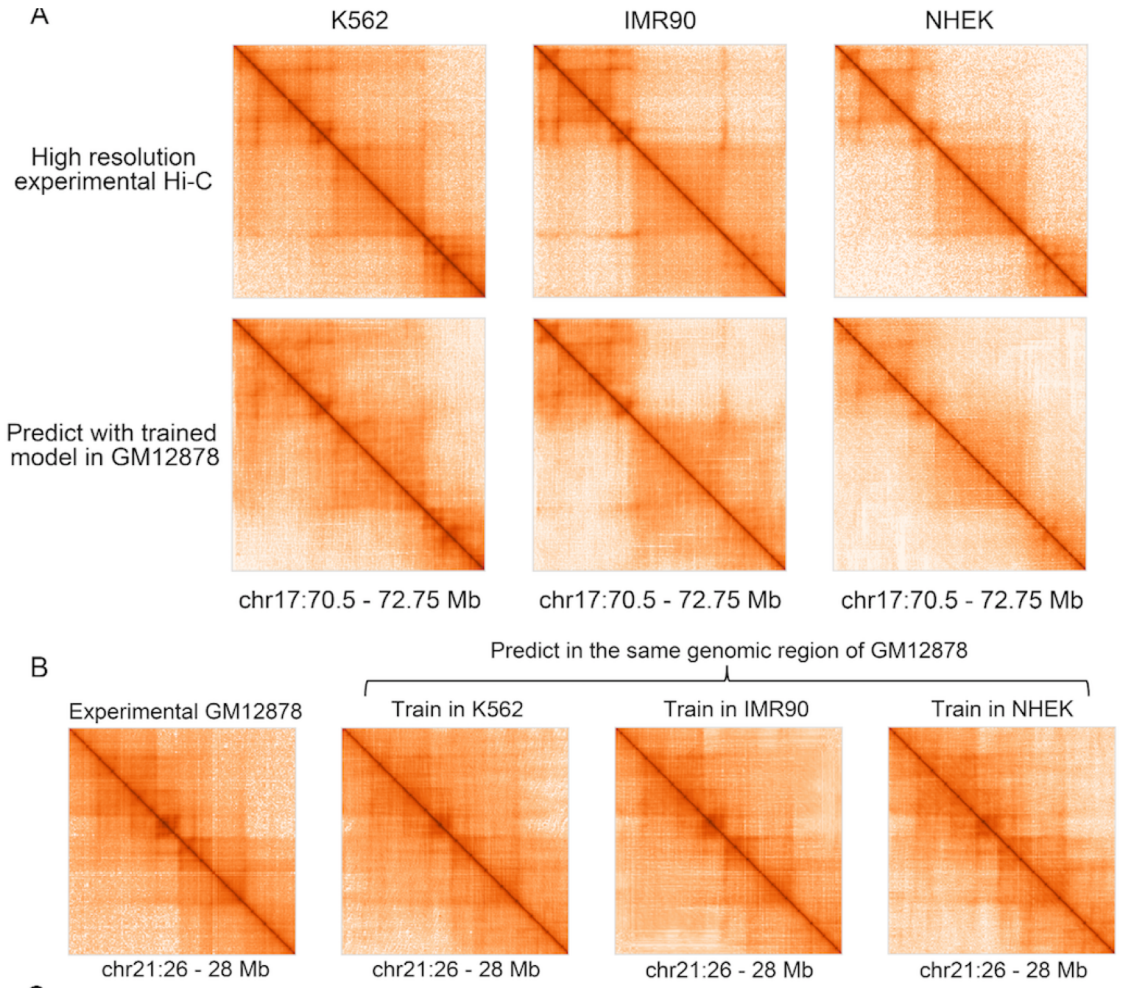


MSE: Mean Square Error
PSNR: Peak signal-to-noise ratio
SSIM: Structure Similarity Index

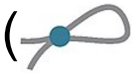


Cross cell type prediction

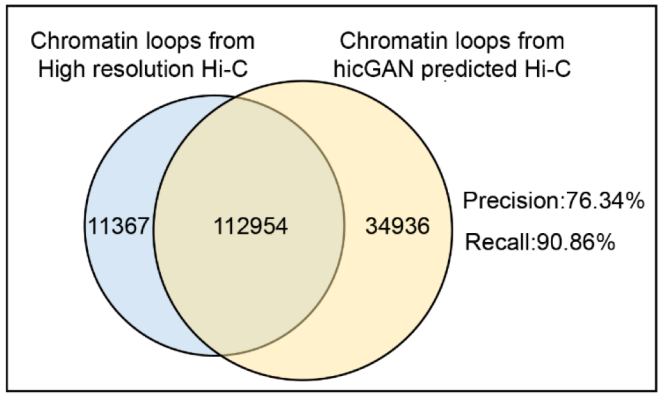
- Experiment settings
  - Train in cell type A, test in cell type B, C and D
  - Train in B, C and D, respectively, test in A



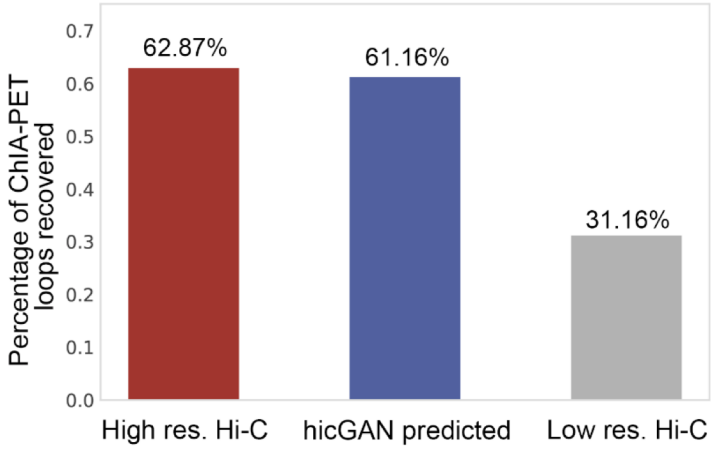
hicGAN recovers meaningful chromatin loops

- Chromatin loops () are identified by Fit-HiC software (Ferhat et al., ***Genome Res***, 2014)

A

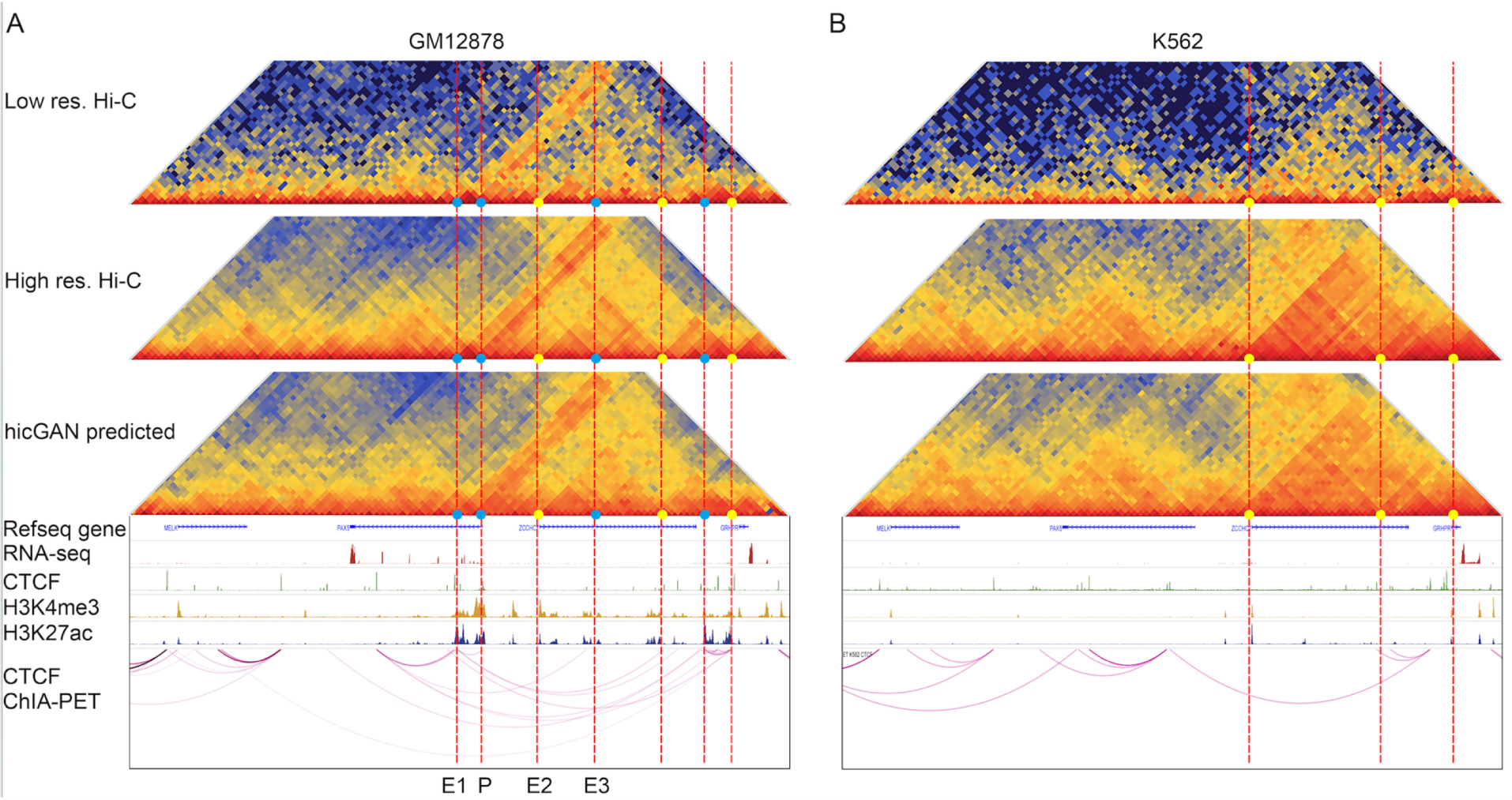


B



hicGAN identifies cell type specific domains

- Three types of Hi-C data extracted from a differential genomic region (chr9:36.5–37.5M) between GM12878 and K562 cell type



Conclusions

- We proposed hicGAN, an open-sourced framework, for inferring high resolution Hi-C data from low resolution Hi-C data with GAN
- hicGAN effectively enhances the resolution of low resolution Hi-C data by generating matrices that are highly similar to high resolution Hi-C matrices
- hicGAN helps identifies meaningful chromatin loops and cell-type specific domain boundaries
- hicGAN provides a fascinating insight into disclosing complex mechanism underlying the conformation of chromatin structure

Acknowledgement

Prof. Rui Jiang

Jiang lab members

- I would like to thank all members in Jiang lab for their helpful discussion
- I would like to thank ISCB for providing the travel fellowship

Thank you !