

Simple statistical models for complex 3D genome data

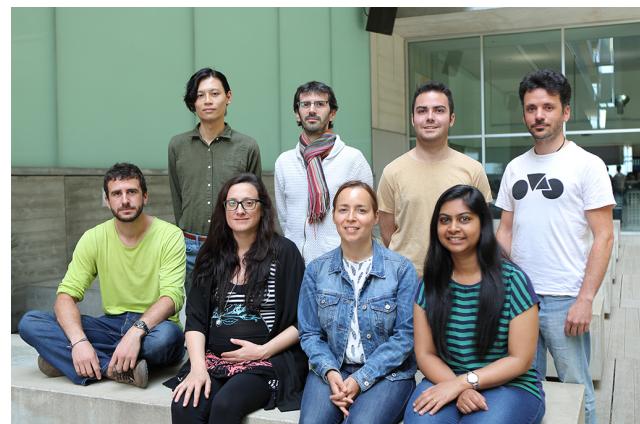
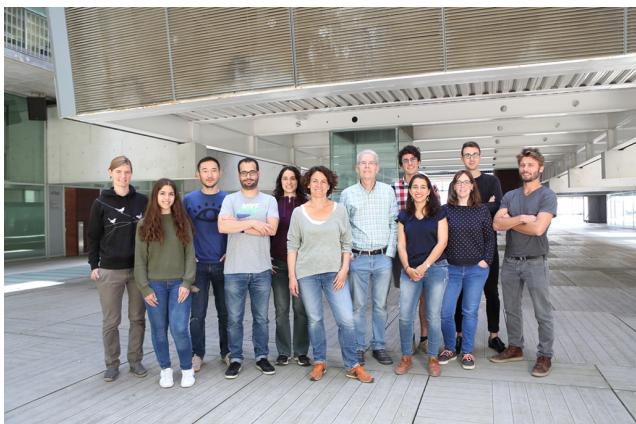
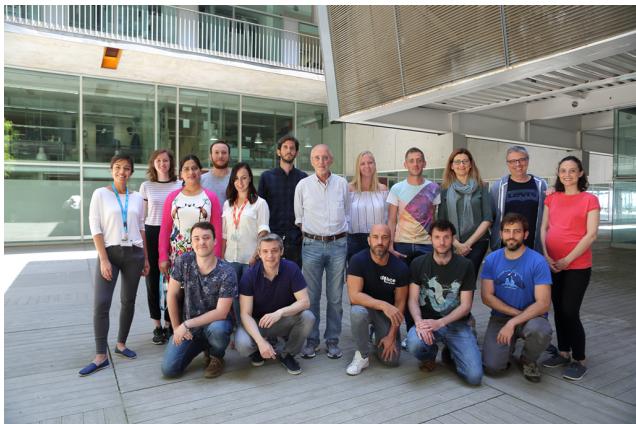
Barcelona, 1 February 2019

Enrique (Quique) Vidal

 enrique.vidal@crg.eu

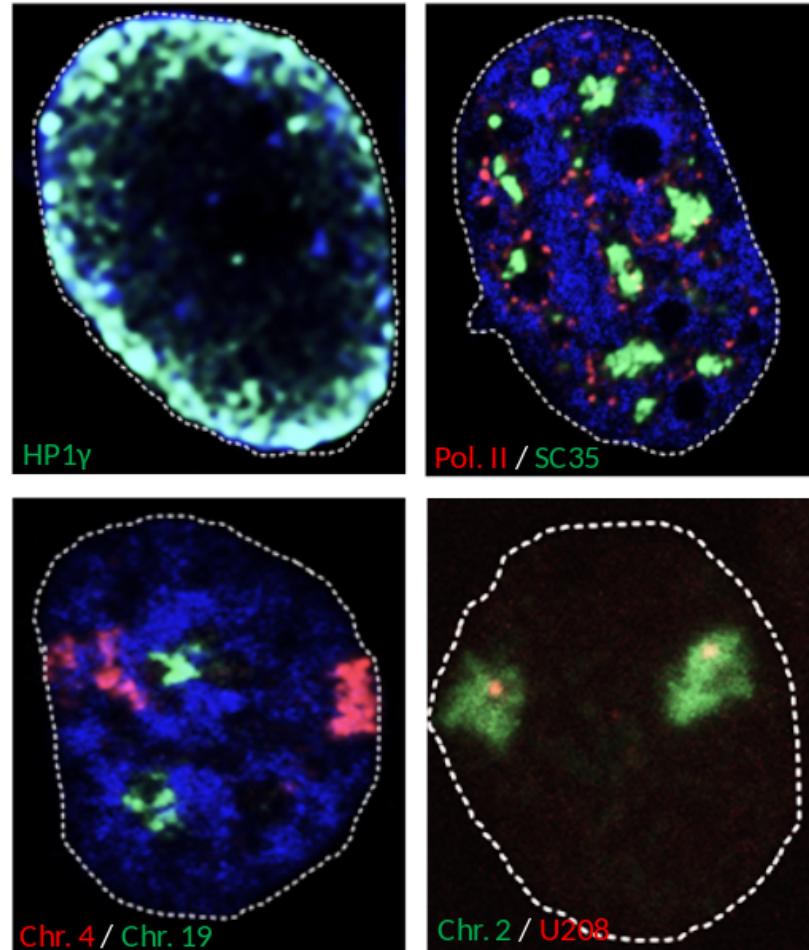
 [@qenvio](https://twitter.com/qenvio)  qenvio

4DGenome



Motivation

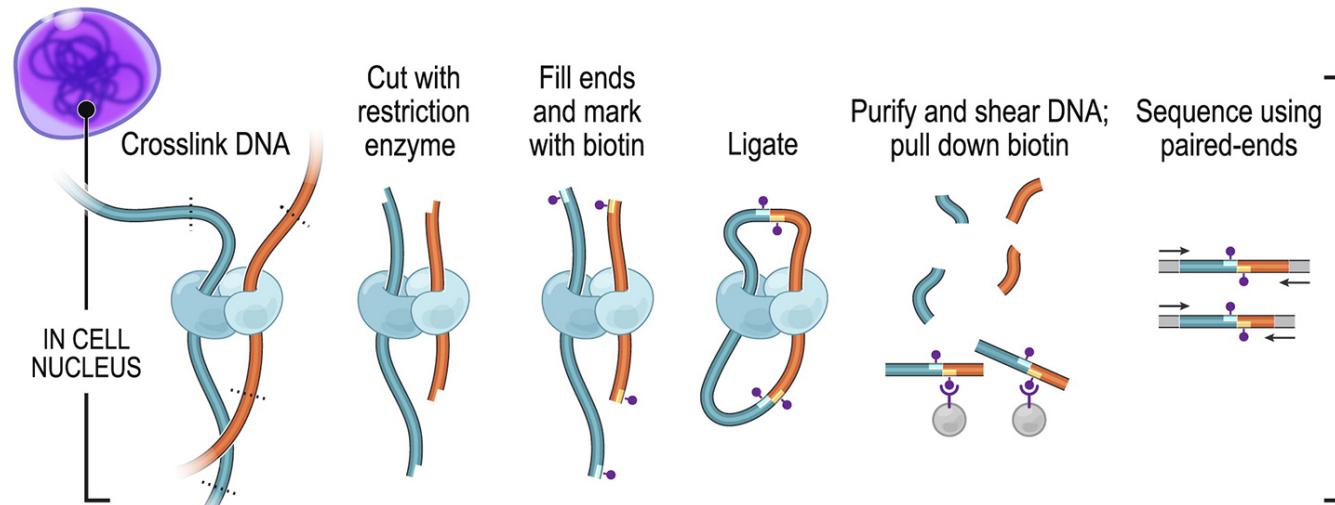
Genome structure is not random



Chromosome conformation capture

Dekker, J. et al. (2002)

Capturing chromosome conformation
Science



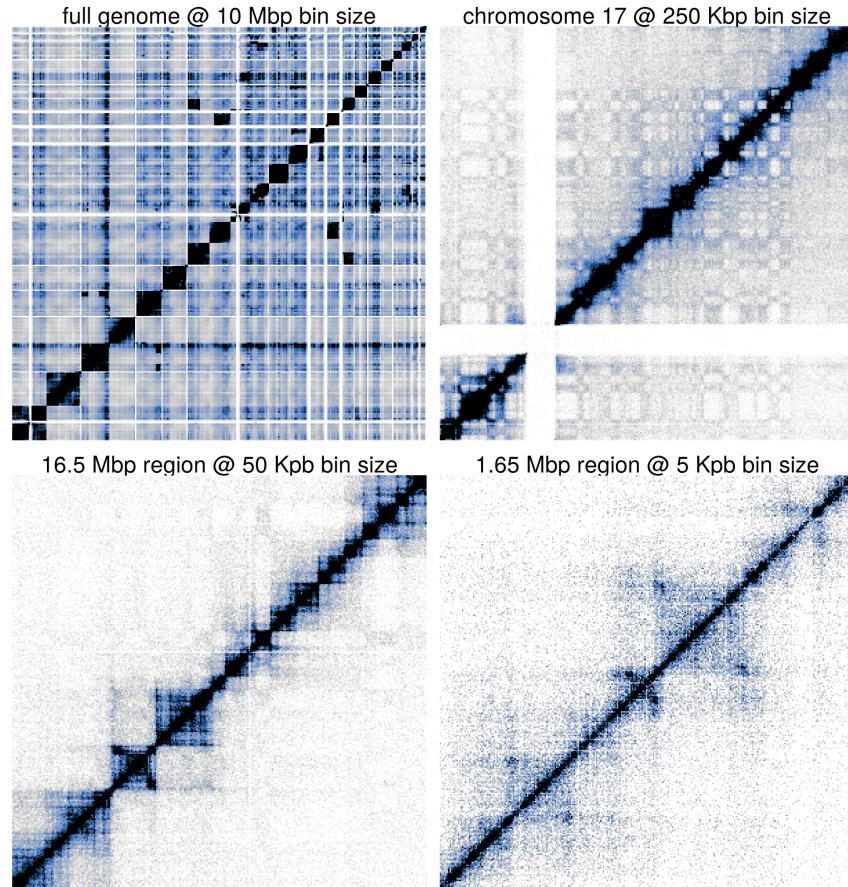
Lieberman-Aiden, E. et al. (2009)

Comprehensive mapping of long-range interactions reveals folding principles of the human genome

Science

What does Hi-C offer?

Territories, compartments, domains, loops



Biases

Genomic features affect Hi-C

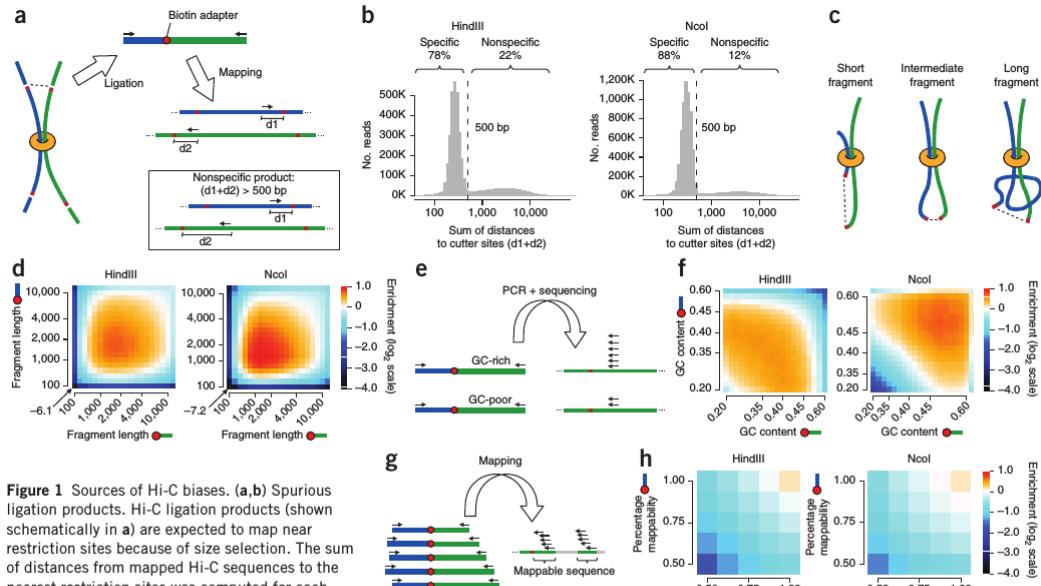


Figure 1 Sources of Hi-C biases. (a,b) Spurious ligation products (shown schematically in a) are expected to map near restriction sites because of size selection. The sum of distances from mapped Hi-C sequences to the nearest restriction sites was computed for each Hi-C paired read, and the distribution of distances was reconstructed (b). Two distinct populations

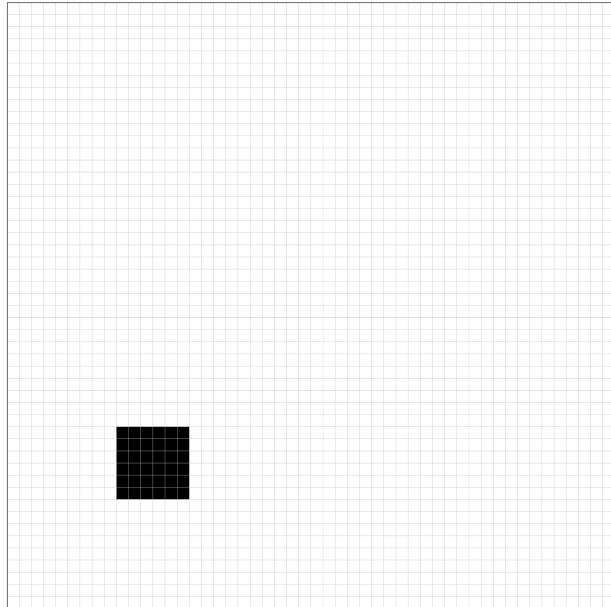
of reads are observed, one distributed as expected for normally ligated and size-selected products (HindIII 78%, Ncol 88%) and one including reads mapped farther away from restriction sites. (c,d) Fragment lengths and ligation efficiency. Restriction fragments of different lengths are shown schematically in c and can be hypothesized to affect crosslinking and ligation efficiency. The *trans* Hi-C coverage enrichment is defined as the ratio between the observed number of *trans* contacts and the total number of assayed fragment pairs. Shown are coverage enrichments for all of the fragment ends, binned into 20 equal-sized bins according to fragment length (x and y axes). Similar trends are observed for the HindIII and Ncol experiments. (e,f) Local GC content and Hi-C coverage. Ligation product processing and sequencing may be biased by GC content (e). *Trans*-contact enrichments (f) stratified according to the GC content of the 200 bp near the restriction fragment ends show intense and contrasting GC biases for the HindIII and Ncol experiments. (g,h) Effect of sequence uniqueness. Different fractions of uniquely mappable short tags are observed next to restriction sites (g). As shown in h, this has a direct empirical linear effect on Hi-C coverage.

Yaffe, E. and Tanay, A. (2011)

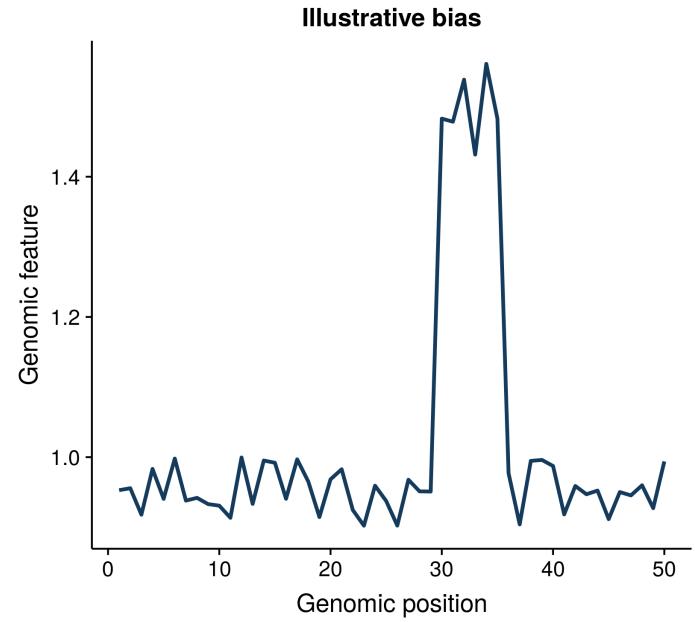
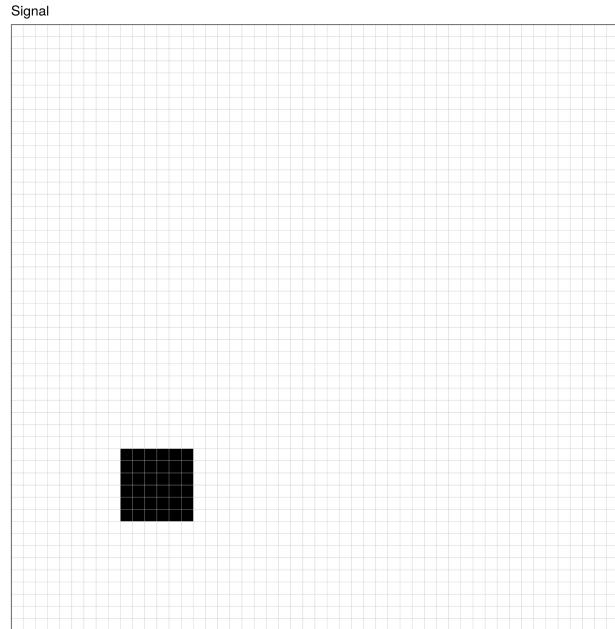
Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture
Nature genetics

Illustrating biases

Signal

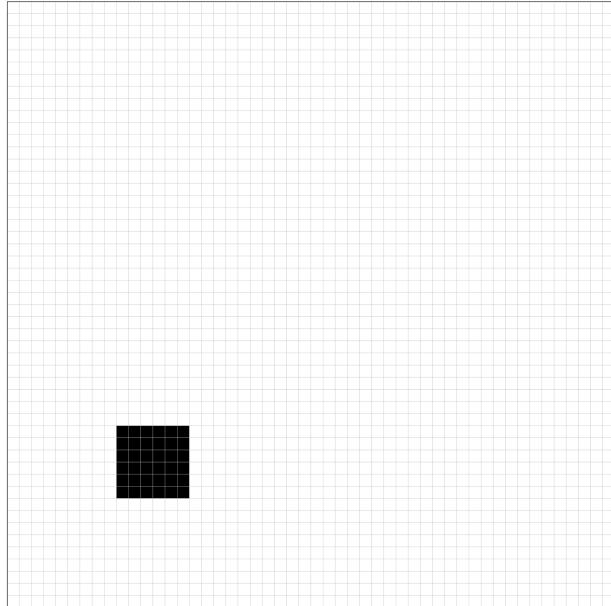


Illustrating biases

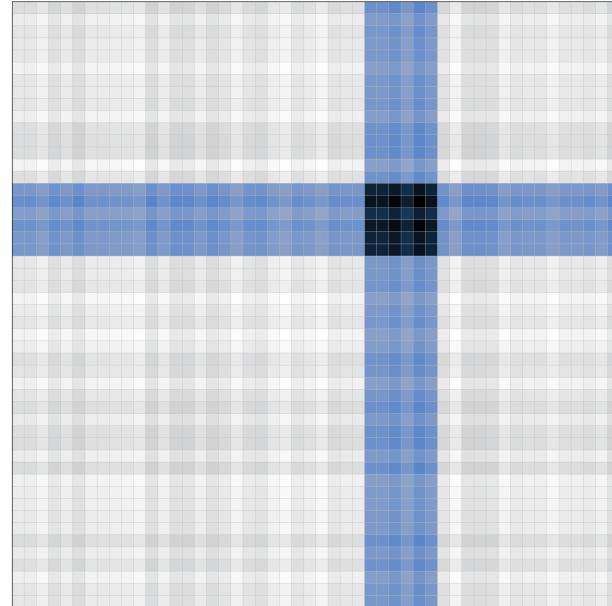


Illustrating biases

Signal

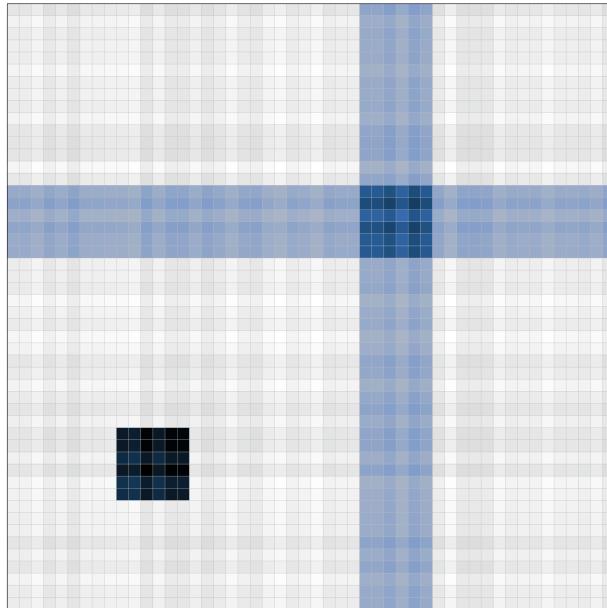


Bias

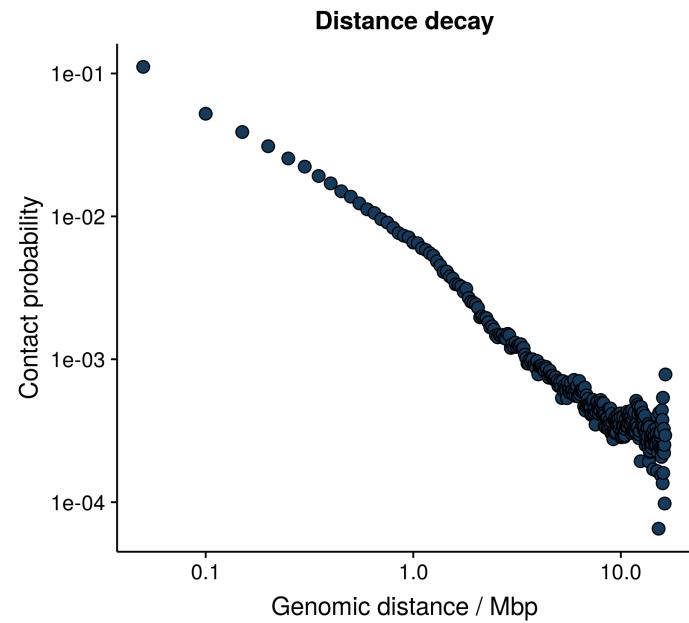
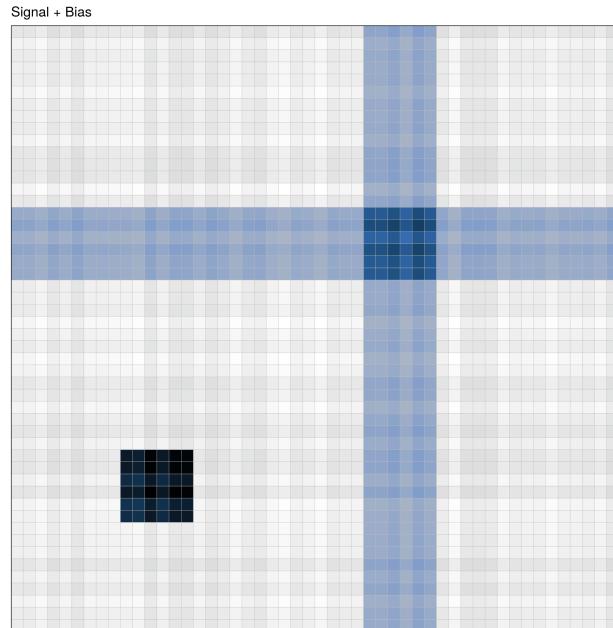


Illustrating biases

Signal + Bias

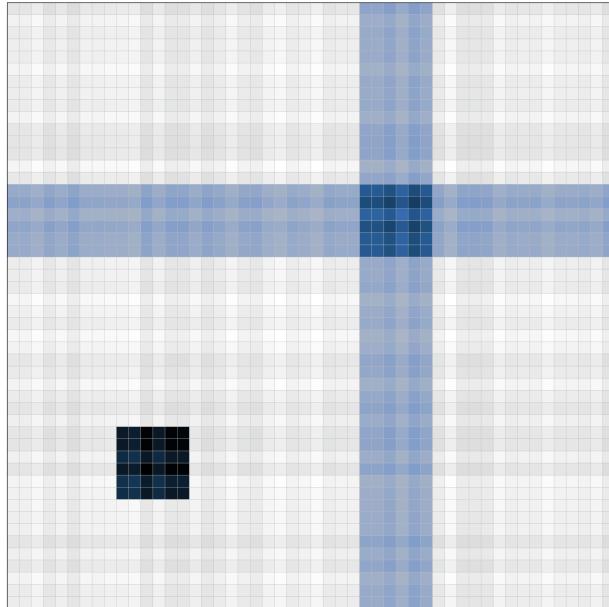


Illustrating biases

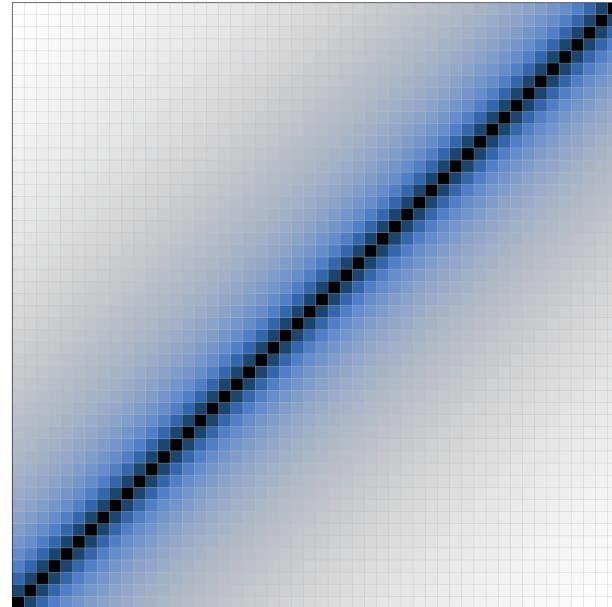


Illustrating biases

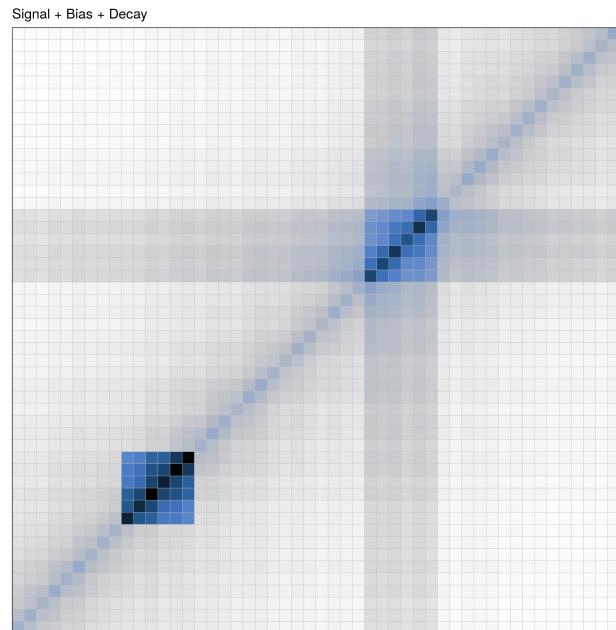
Signal + Bias



Decay

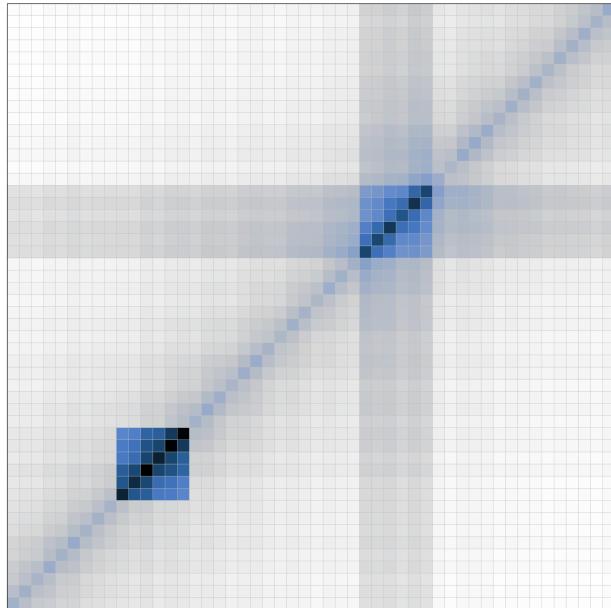


Illustrating biases

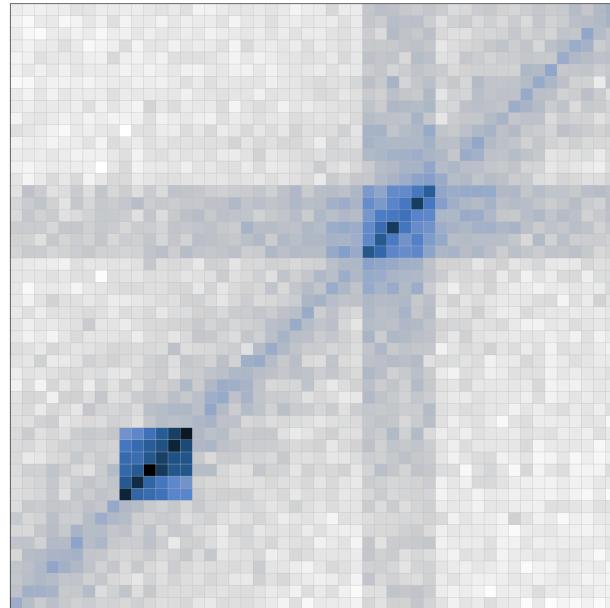


Illustrating biases

Signal + Bias + Decay

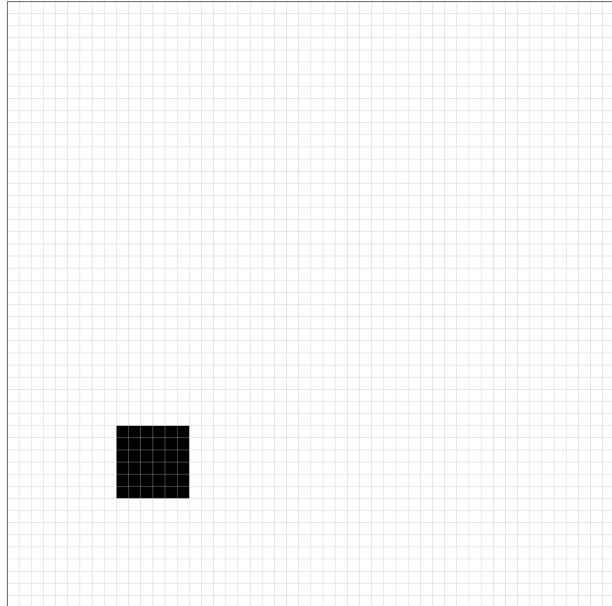


Signal + Bias + Decay + Noise

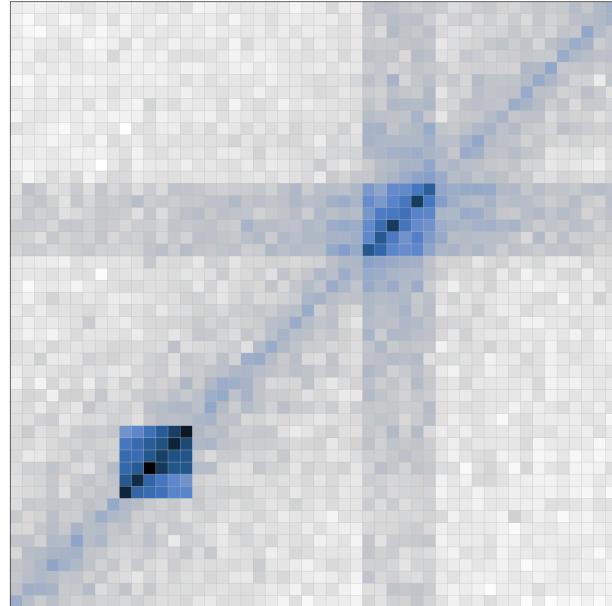


Illustrating biases

Signal



Signal + Bias + Decay + Noise



Previous approaches

HiCNorm

Explicit model of biases

Regression model

All matrix entries

ICE

Implicit correction

Matrix balancing

Equal visibility

Yaffe, E. and Tanay, A. (2011)

Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture

Nature genetics

Hu, M. et al. (2012)

HiCNorm: removing biases in Hi-C data via Poisson regression

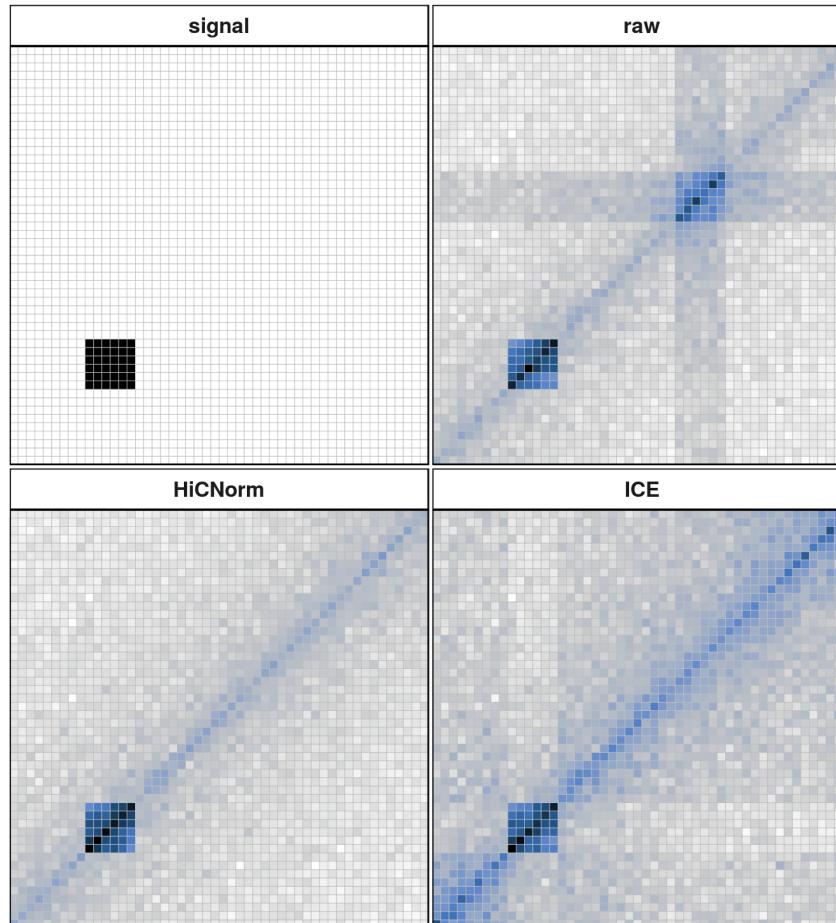
Bioinformatics

Imakaev, M. et al. (2012)

Iterative correction of Hi-C data reveals hallmarks of chromosome organization

Nature methods

In action



HiCNorm

✓ fewer artifacts

✗ very slow

✗ no high resolution

ICE

✗ more artifacts

✓ fast

✓ high resolution

HiCNorm

ICE

✓ fewer artifacts

✗ more artifacts

✗ very slow

✓ fast

✗ no high resolution

✓ high resolution

✗ aberrant karyotypes

Wish list

Wish list

♡ Suitable aberrant karyotypes

Wish list

♡ Suitable aberrant karyotypes

♡ As good as existing

Wish list

♡ Suitable aberrant karyotypes

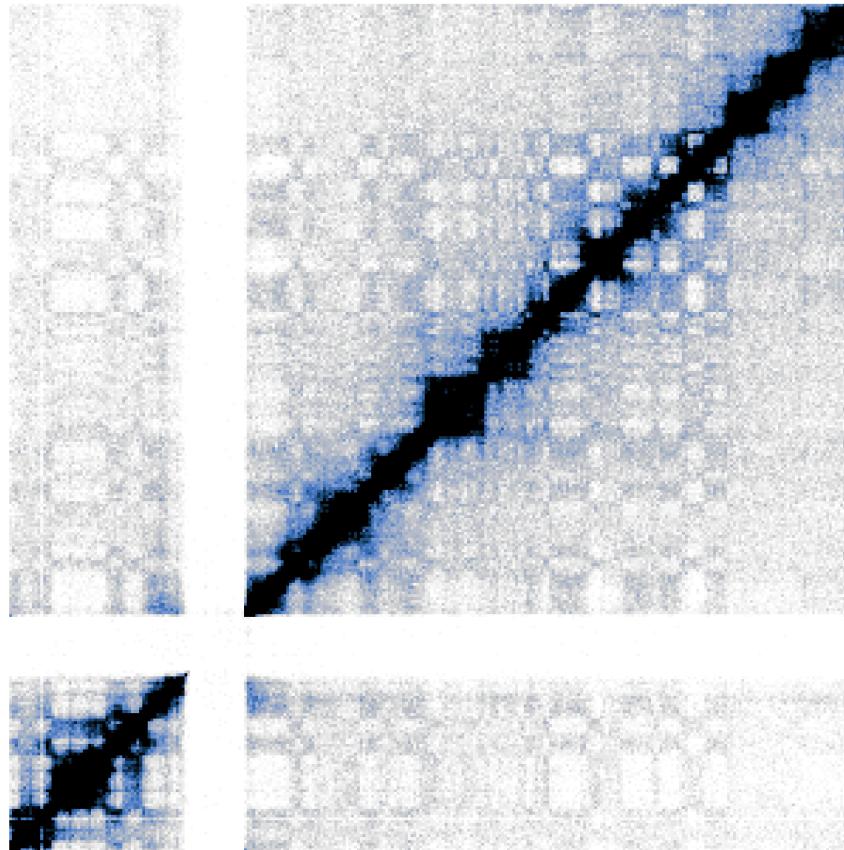
♡ As good as existing

♡ Fast and scalable (high resolution)

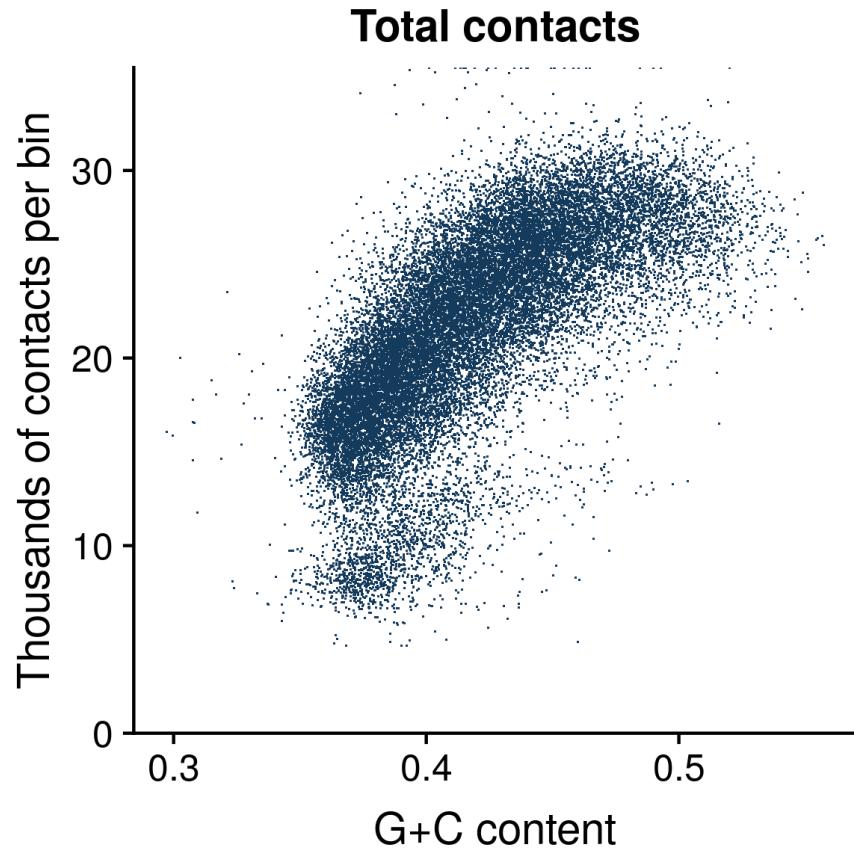
2D → 1D

2D → 1D

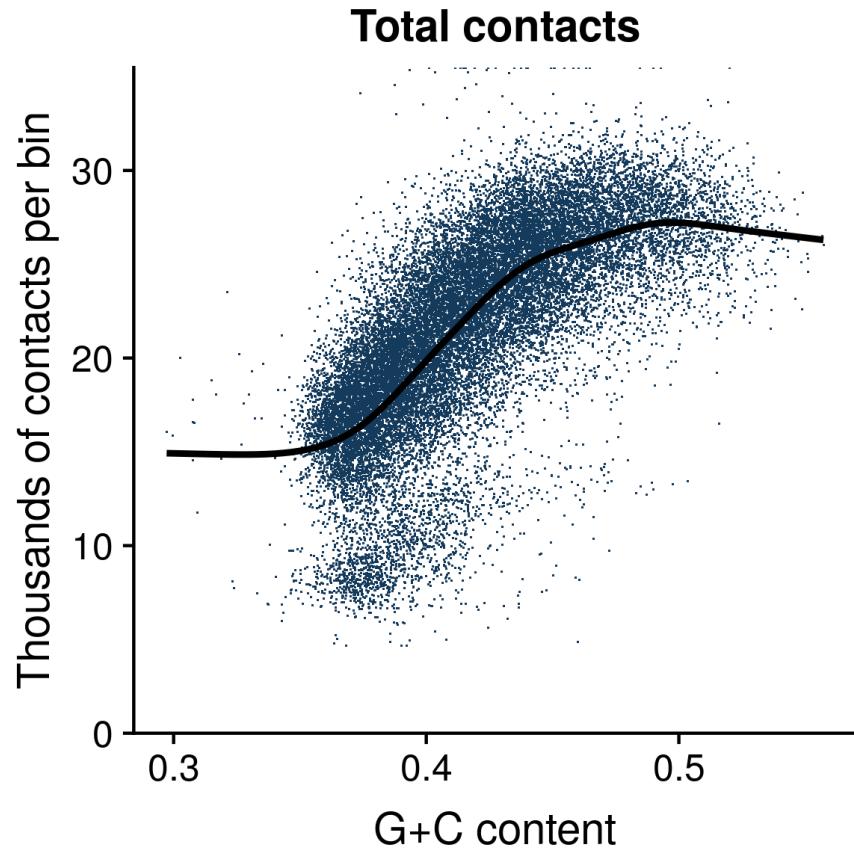
chromosome 17 @ 250 Kbp bin size



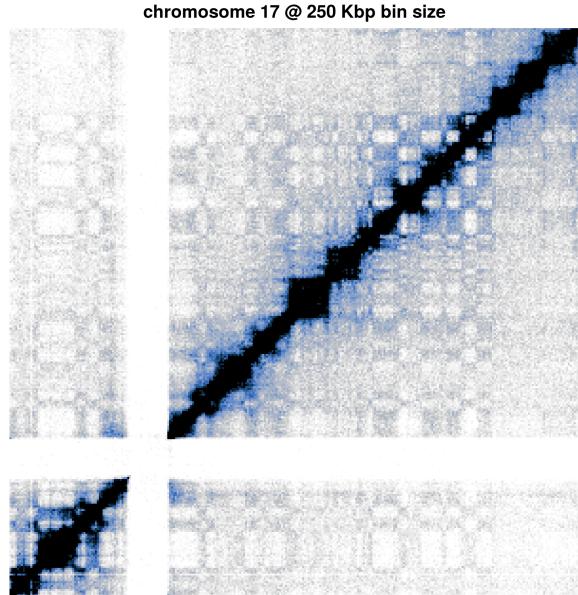
2D → 1D vs. genomic features



2D → 1D vs. genomic features



OneD model



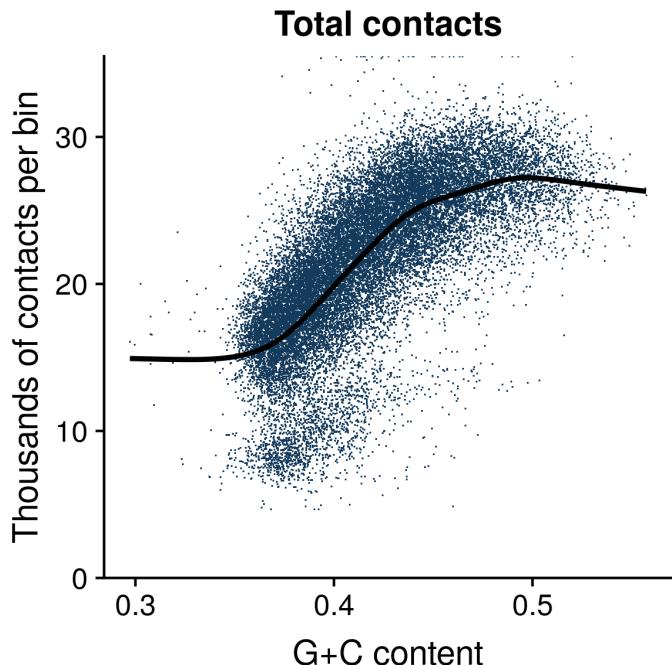
$$t_i = \sum_j^n x_{i,j} \sim NB(\lambda_i, \theta)$$

$$\log(\lambda_i) \propto \sum_k f_k(x_k)$$

$$\lambda'_i = \frac{\lambda_i}{\sum_i^n \lambda_j / n}$$

$$\hat{x}_{i,j} = \frac{x_{i,j}}{\sqrt{\lambda'_i \lambda'_j}}$$

OneD model



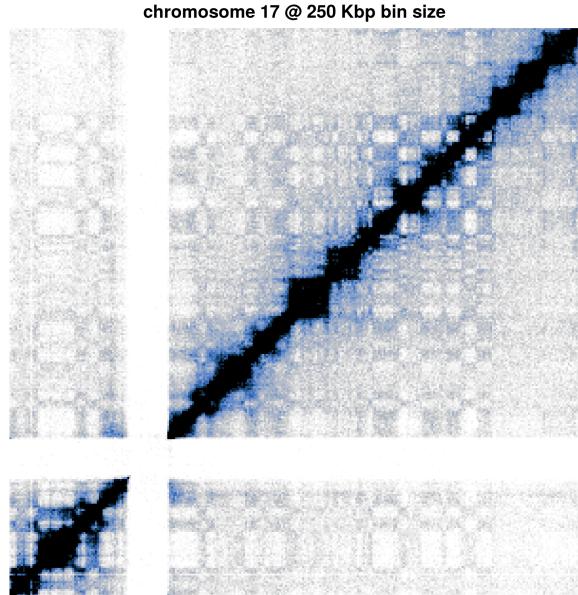
$$t_i = \sum_j^n x_{i,j} \sim NB(\lambda_i, \theta)$$

$$\log(\lambda_i) \propto \sum_k f_k(x_k)$$

$$\lambda'_i = \frac{\lambda_i}{\sum_i^n \lambda_j / n}$$

$$\hat{x}_{i,j} = \frac{x_{i,j}}{\sqrt{\lambda'_i \lambda'_j}}$$

OneD model



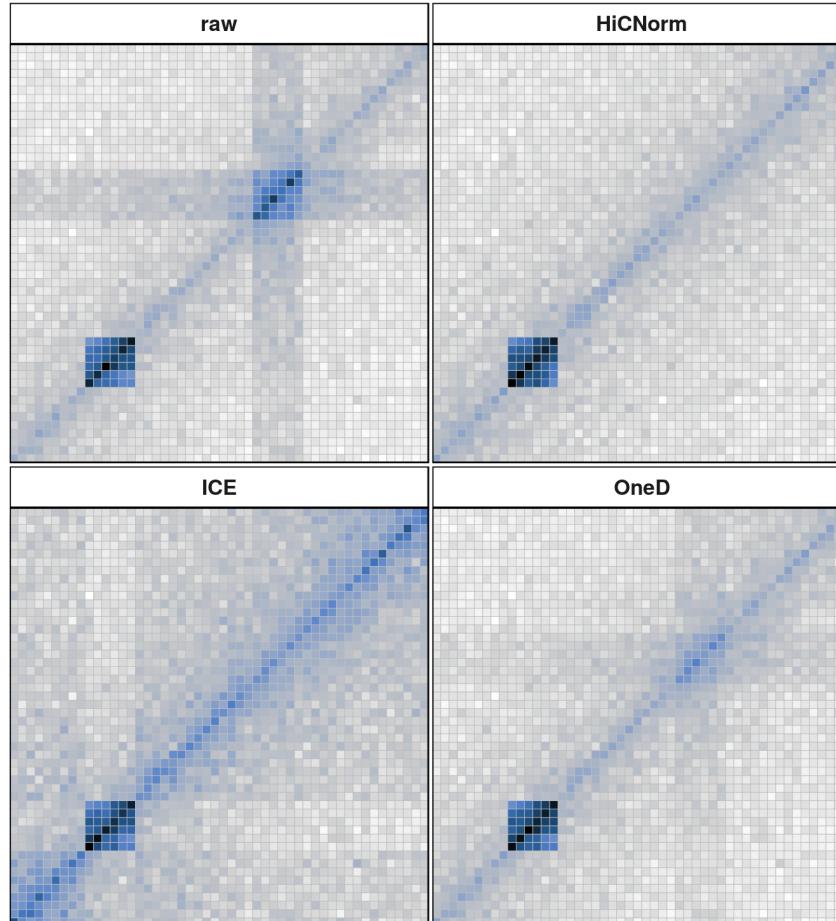
$$t_i = \sum_j^n x_{i,j} \sim NB(\lambda_i, \theta)$$

$$\log(\lambda_i) \propto \sum_k f_k(x_k)$$

$$\lambda'_i = \frac{\lambda_i}{\sum_i^n \lambda_j / n}$$

$$\hat{x}_{i,j} = \frac{x_{i,j}}{\sqrt{\lambda'_i \lambda'_j}}$$

OneD in action



OneD performance

Benchmark strategy

Compare methods

Previous, **OneD** and doing nothing (raw)

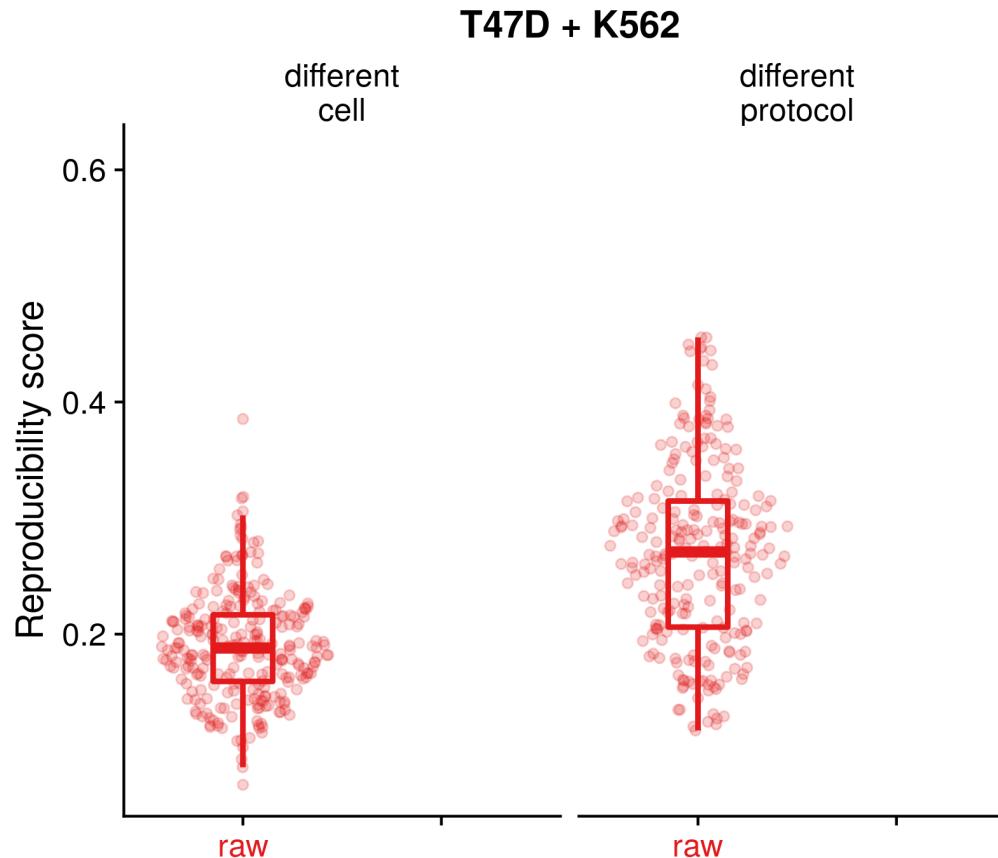
Undesired variability

Different protocols (restriction enzyme, in-situ / diluted, lab, etc ...)

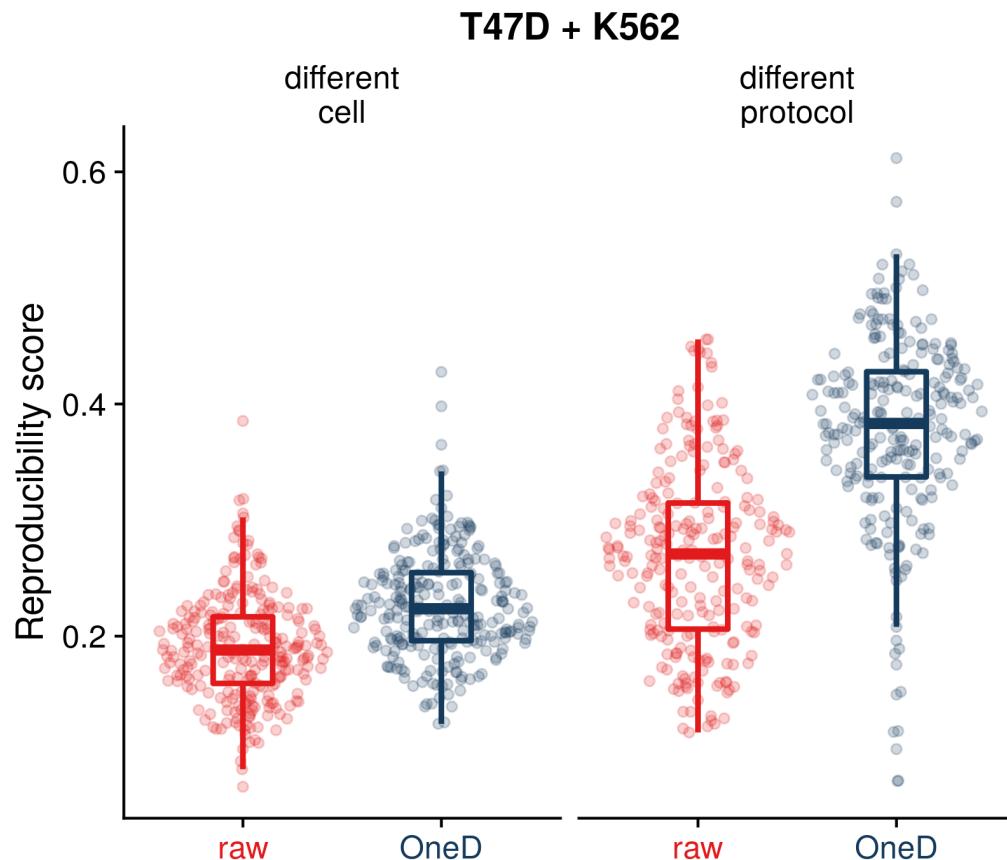
Relevant variability

Different cell types

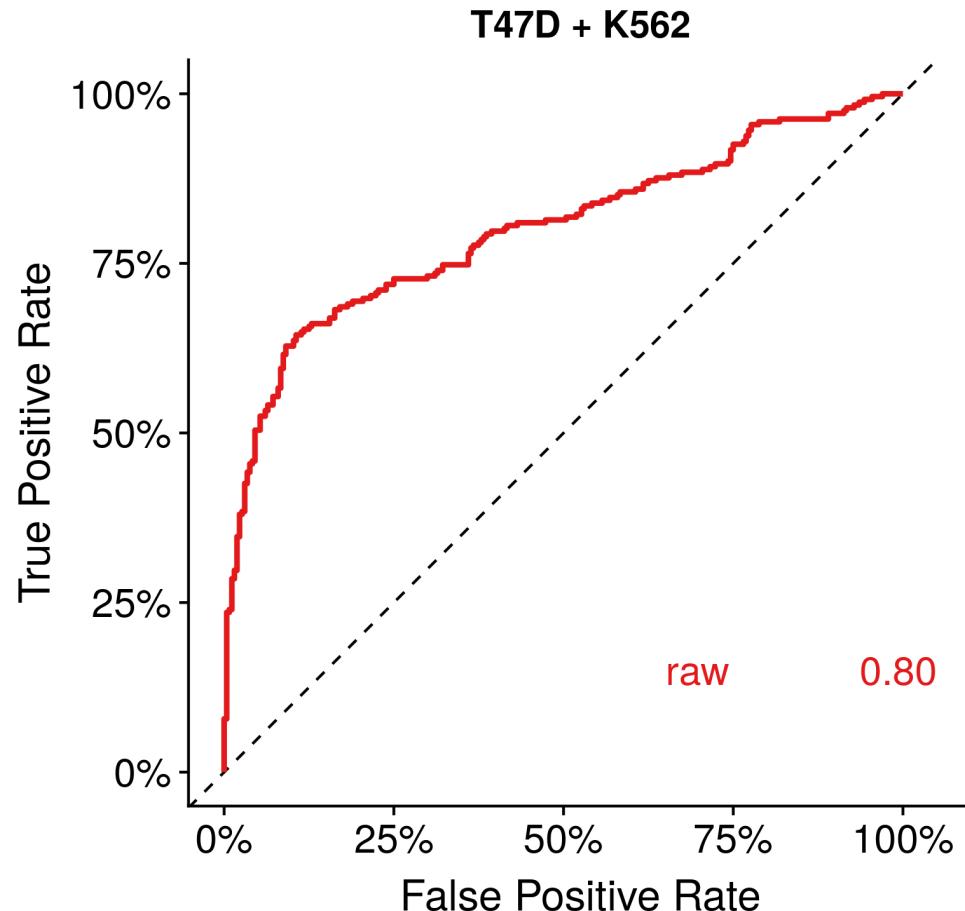
Pair-wise comparison



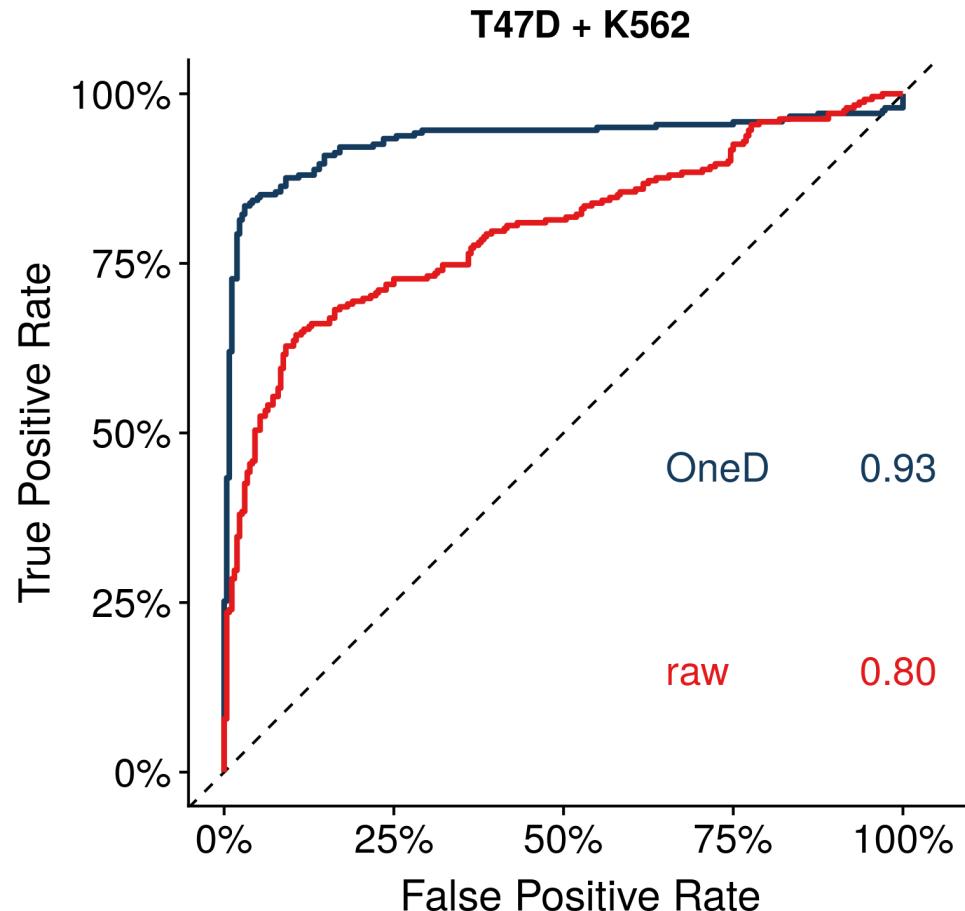
Pair-wise comparison



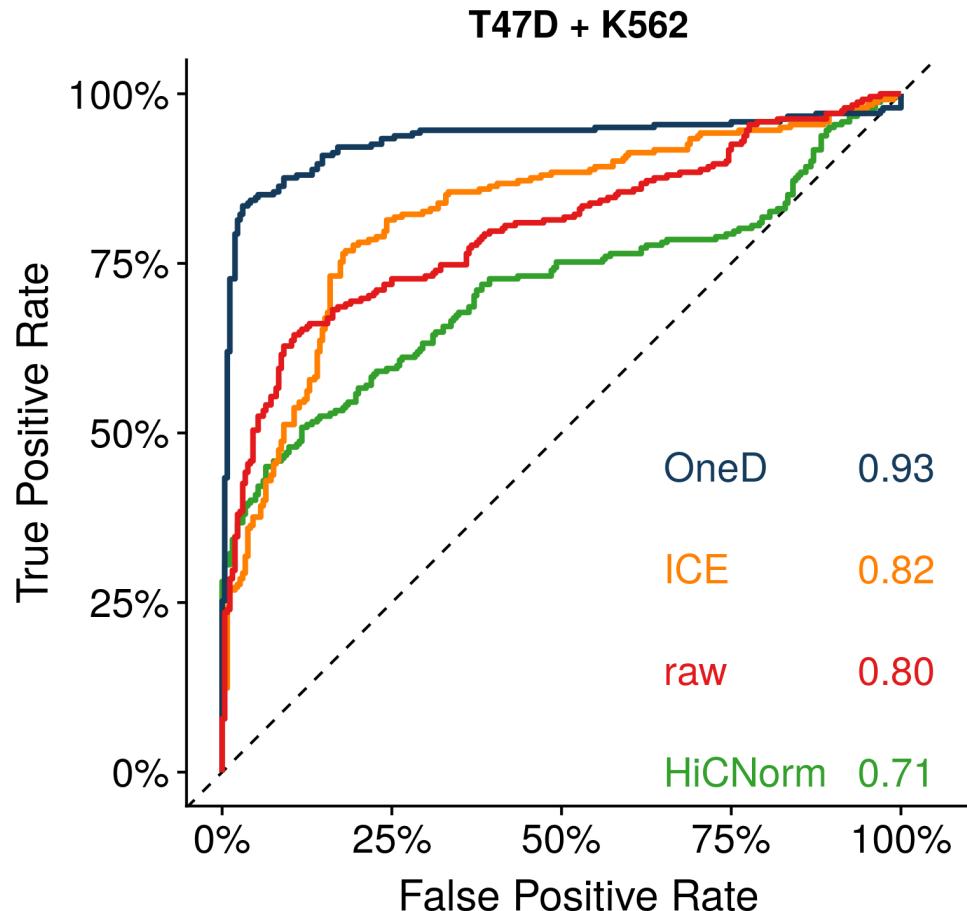
ROC and AUC



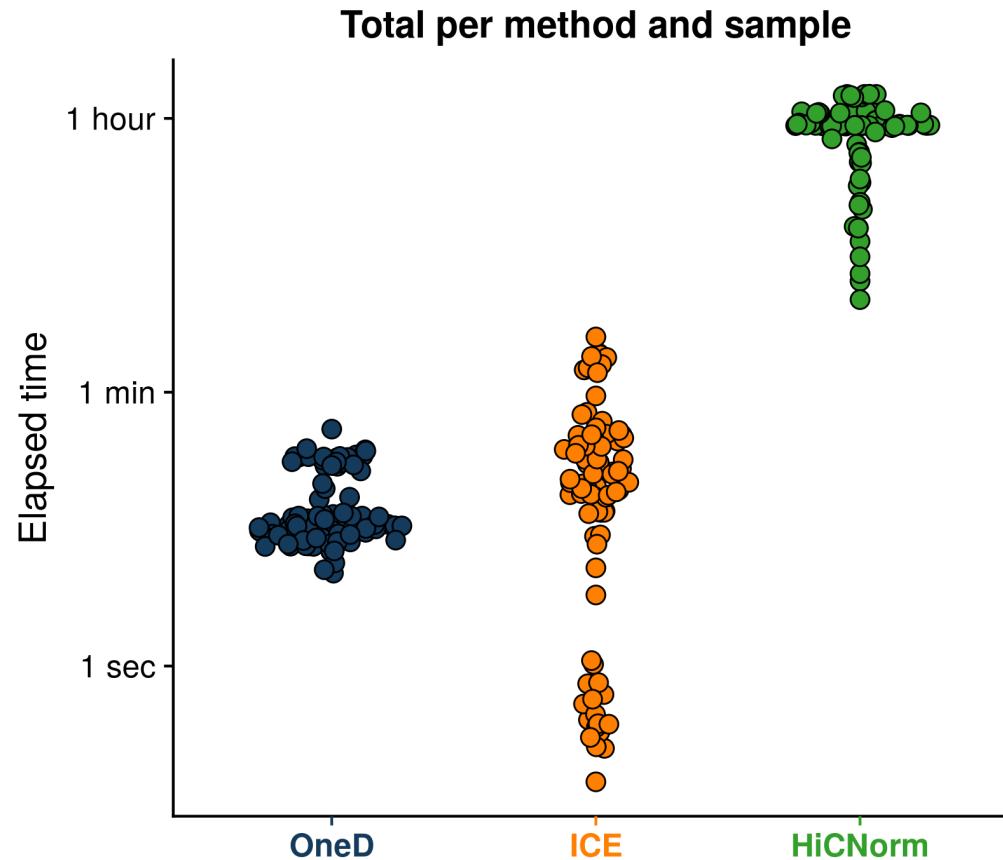
ROC and AUC



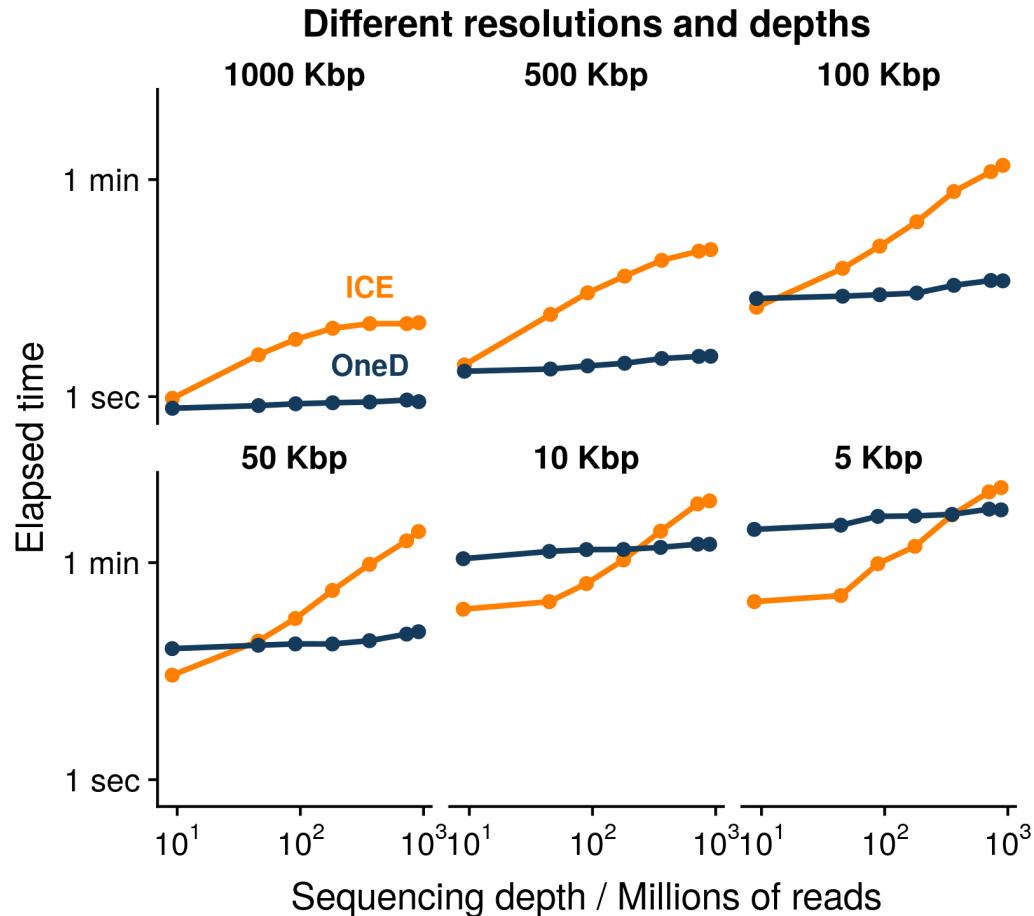
ROC and AUC



Speed and resolution



Speed and resolution



Wish list

- ♡ Suitable aberrant karyotypes ✓
- ♡ ~~As good as~~ Better than existing ✓
- ♡ Fast and scalable (high resolution) ✓

OneD availability

R package

<https://github.com/qenvio/dryhic>

Paper

Vidal, E. et al. (2018)

OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes

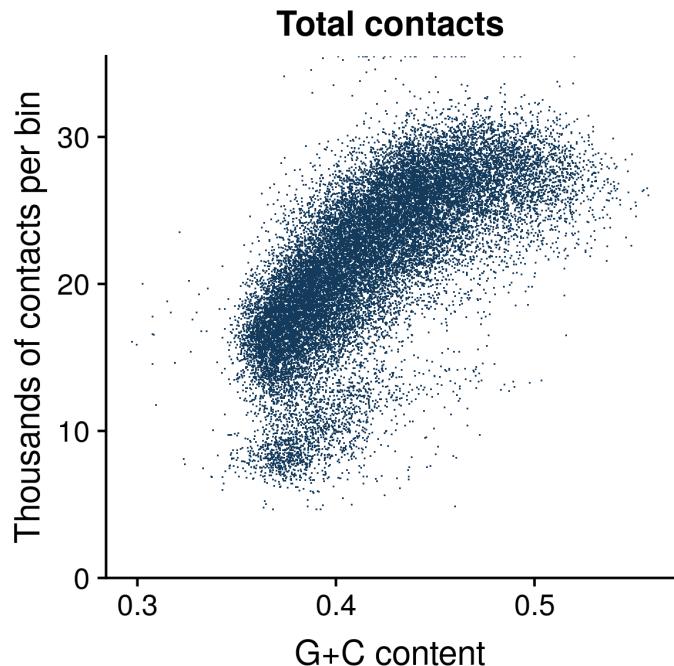
Nucleic acids research

<https://doi.org/10.1093/nar/gky064>

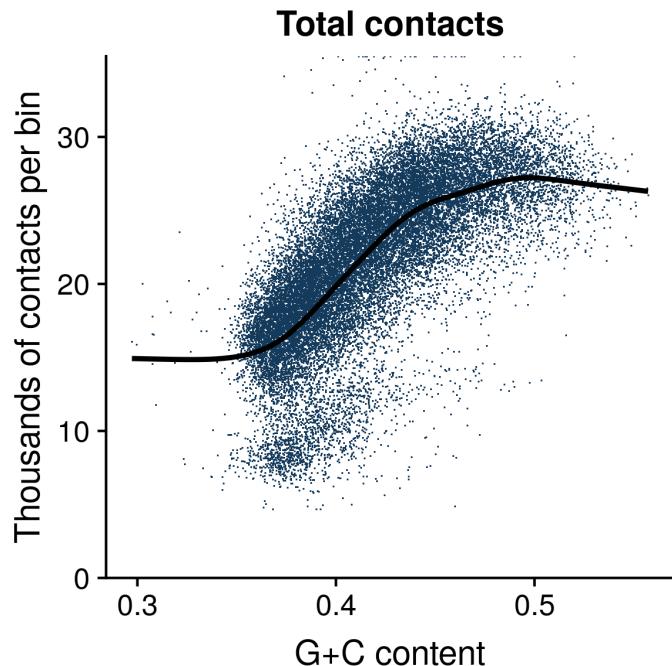
Exploiting the model

Less is more

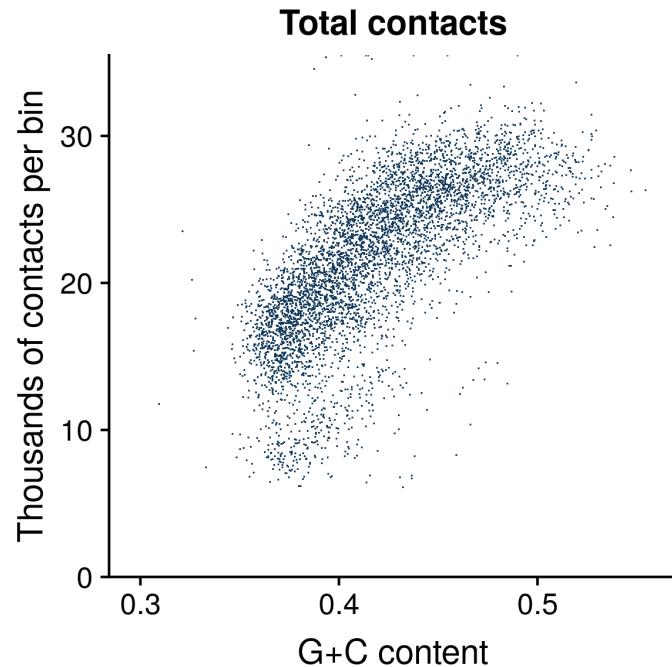
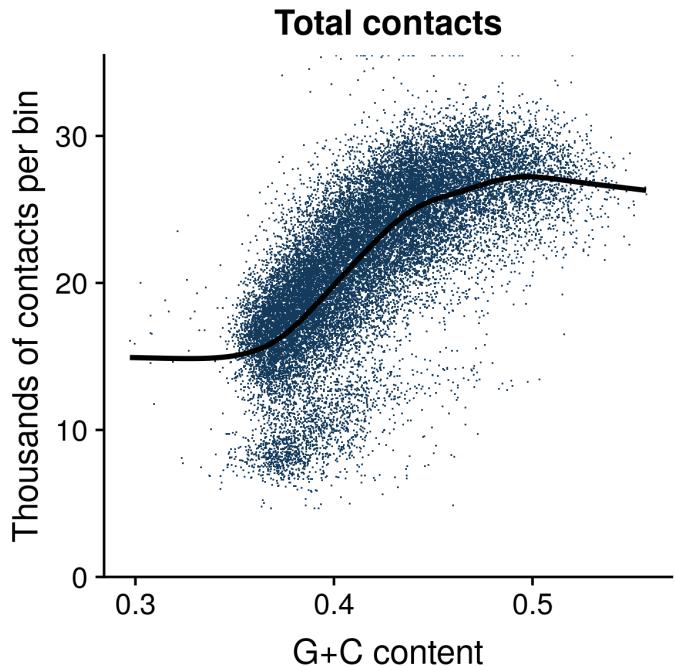
Use only a subset



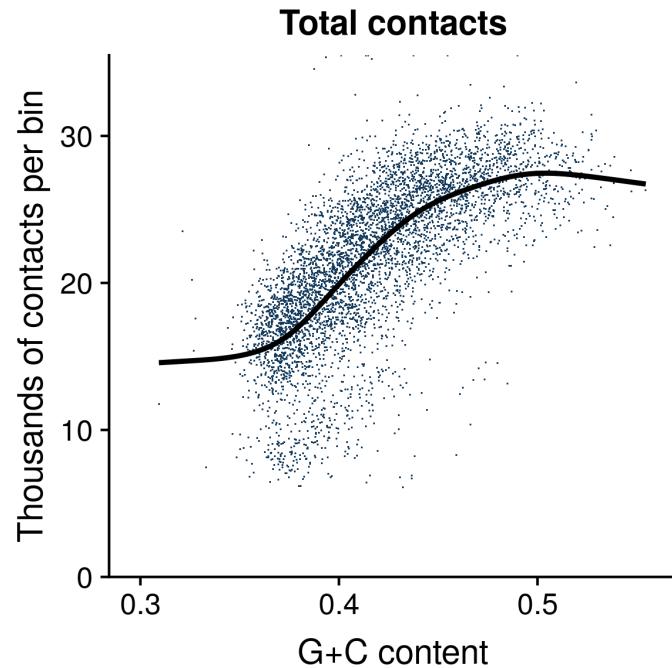
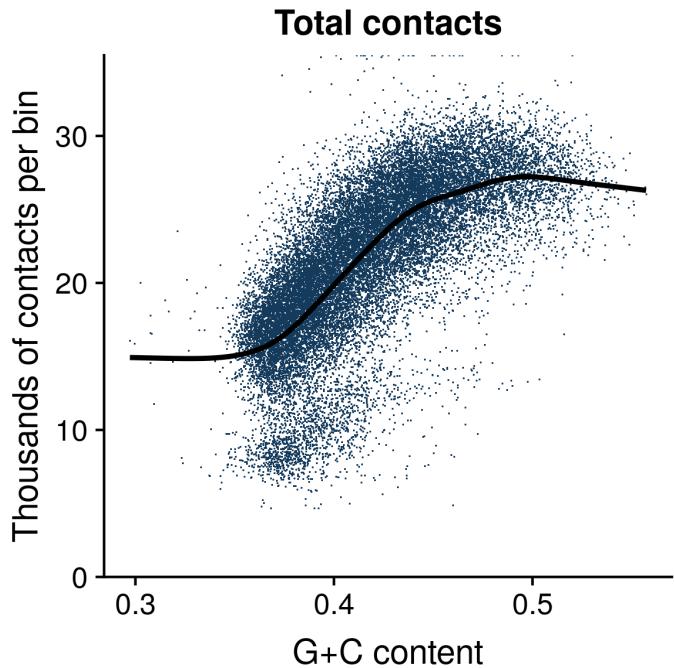
Use only a subset



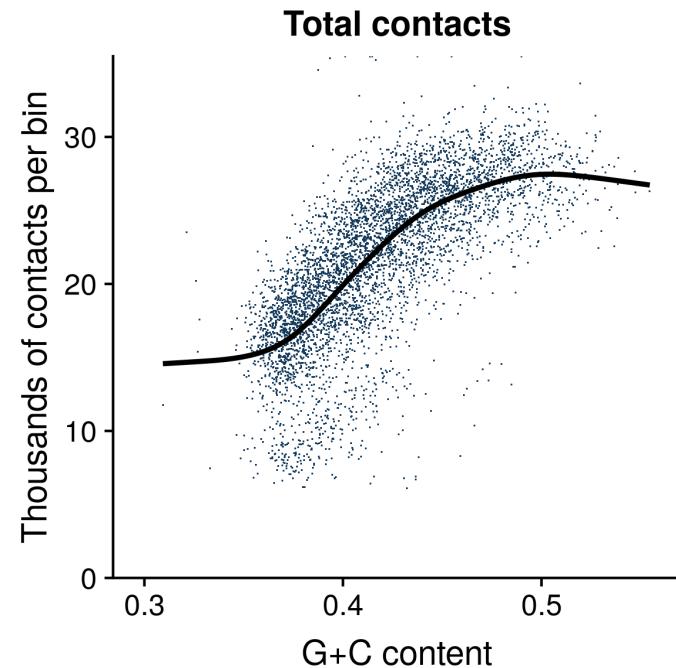
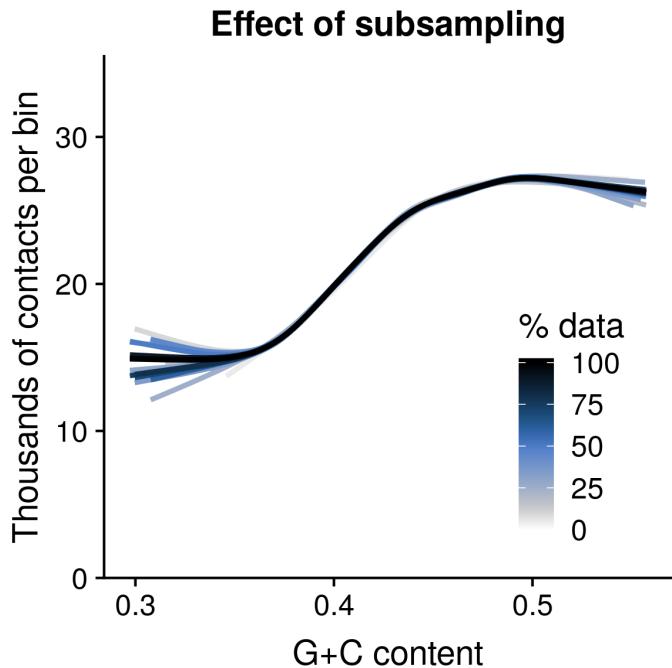
Use only a subset



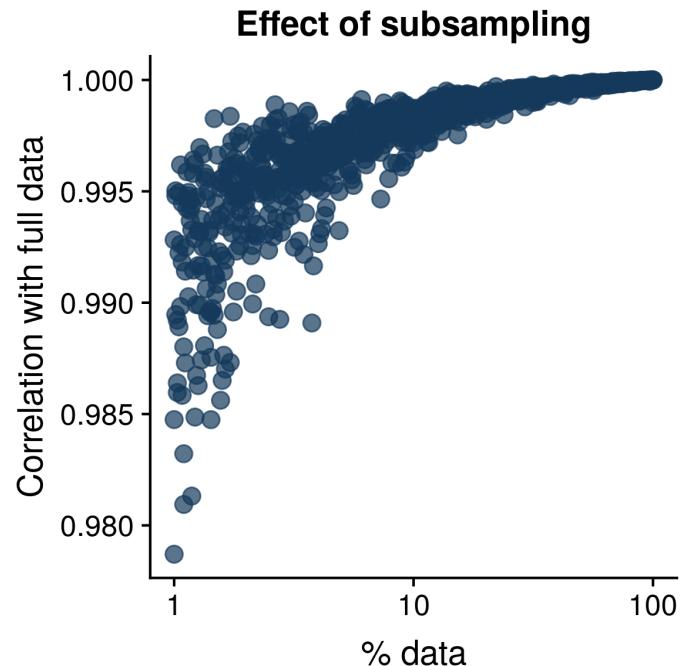
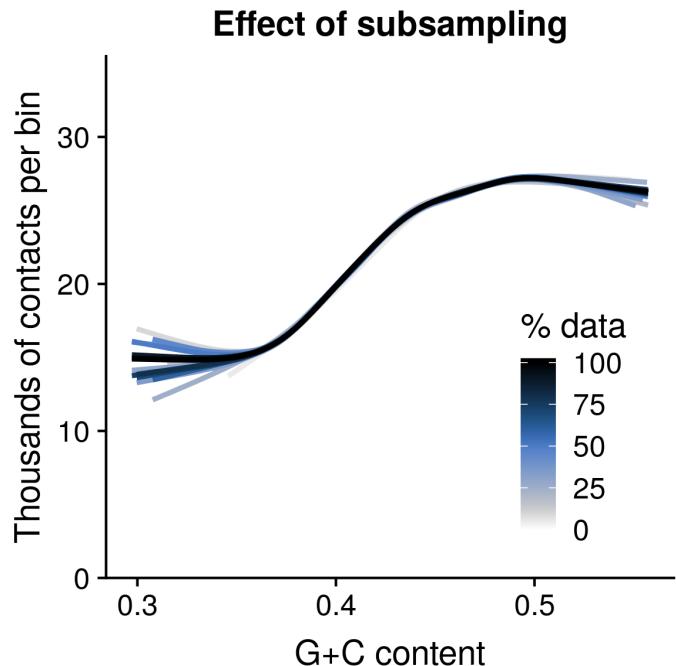
Use only a subset



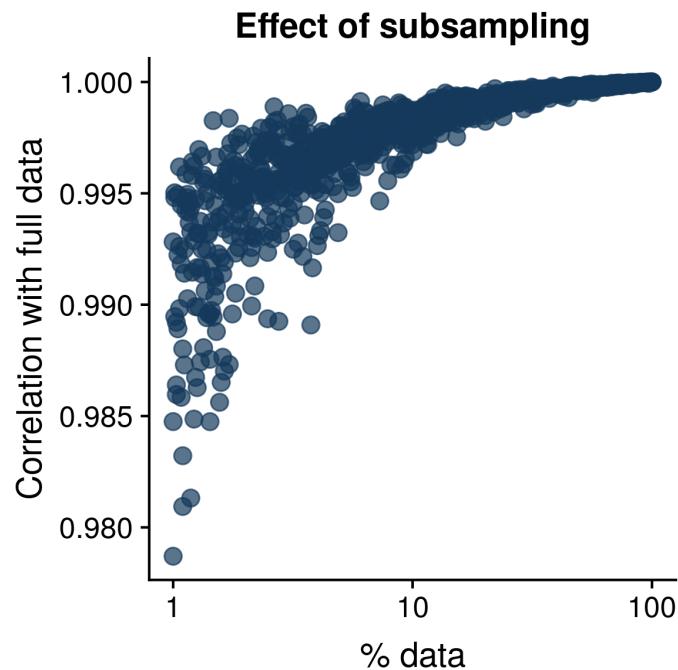
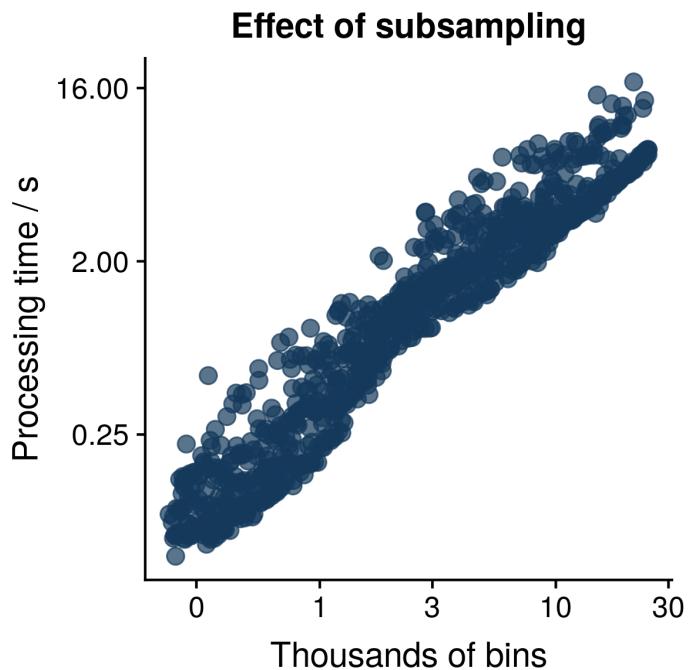
Use only a subset



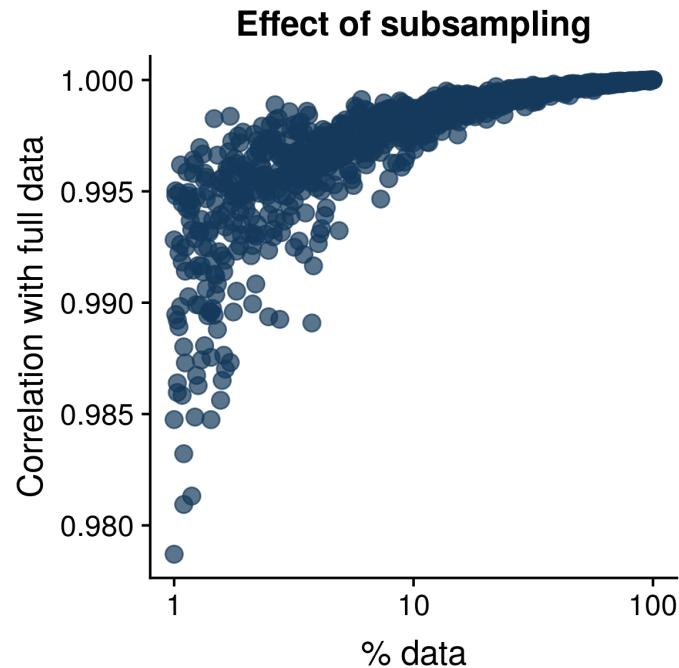
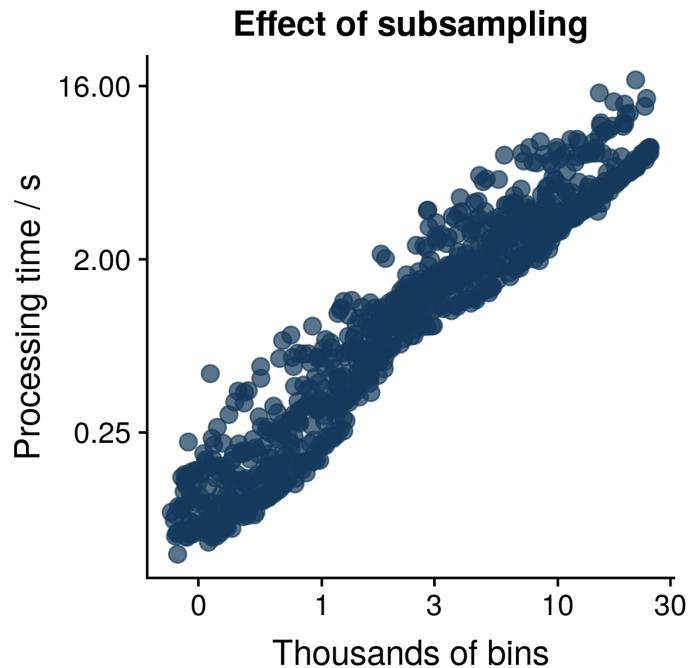
Use only a subset



Use only a subset



Use only a subset

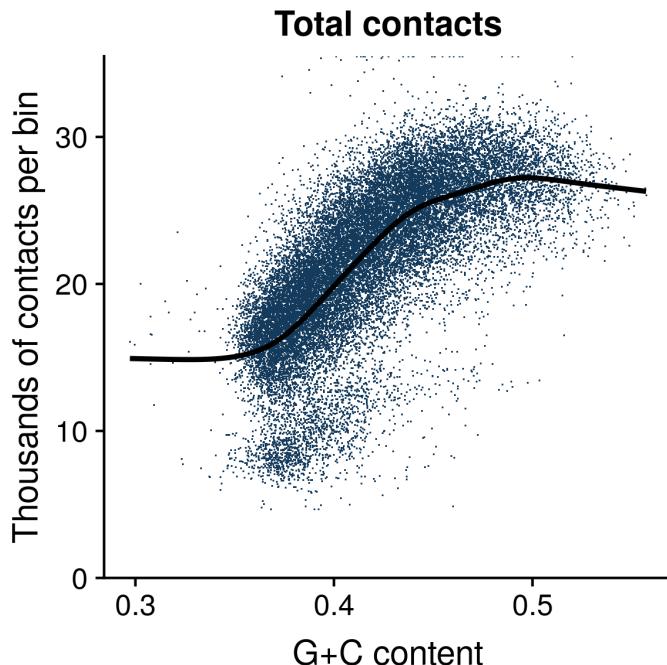


Incomplete designs ✓

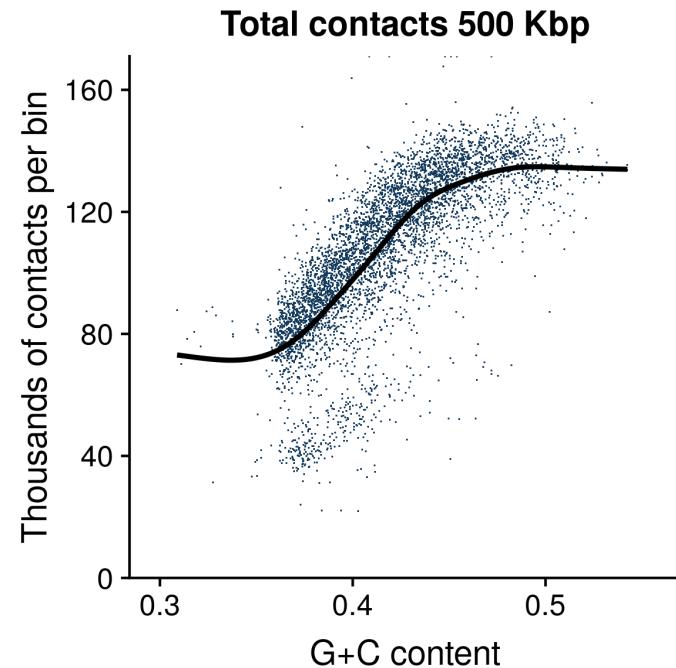
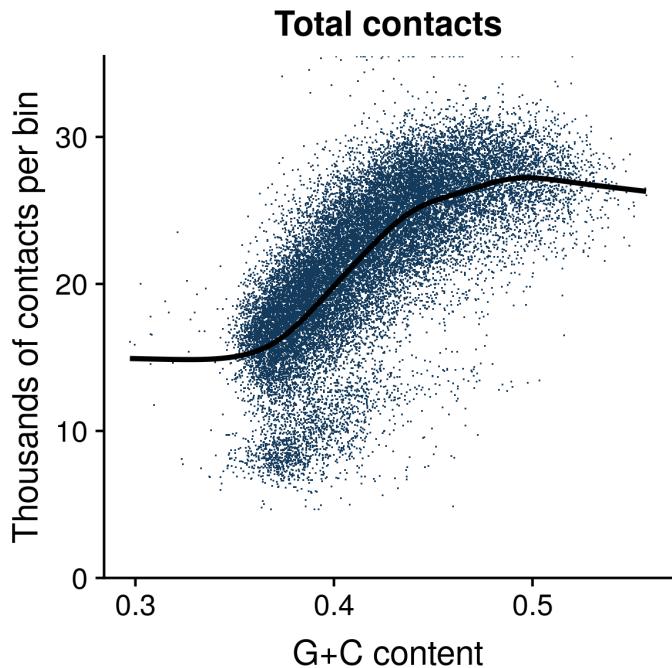
Exploiting the model

Resolution-free

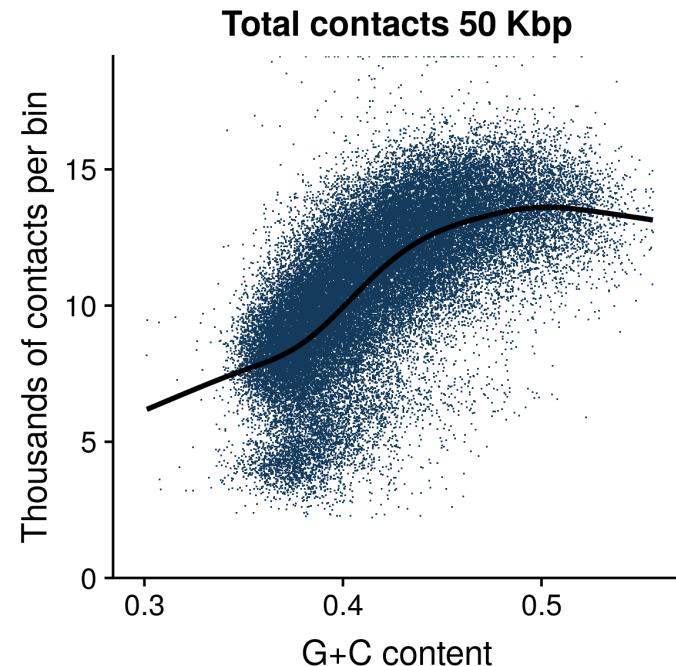
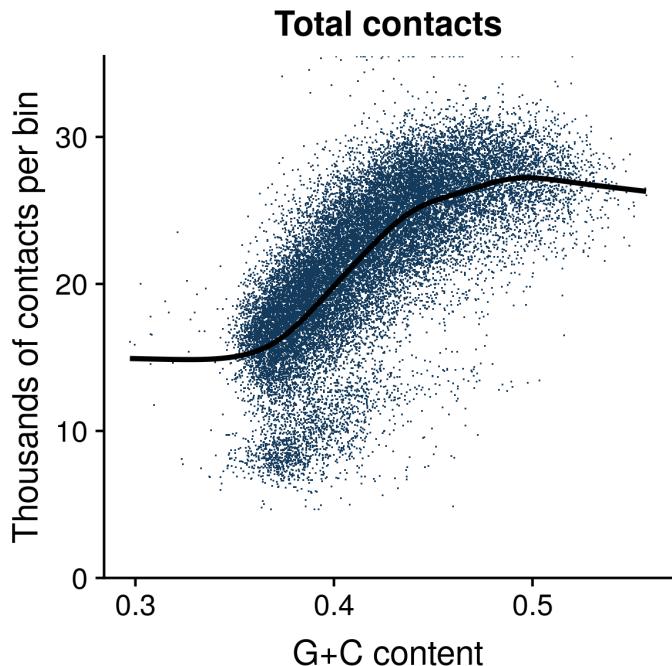
Fit once, project everywhere



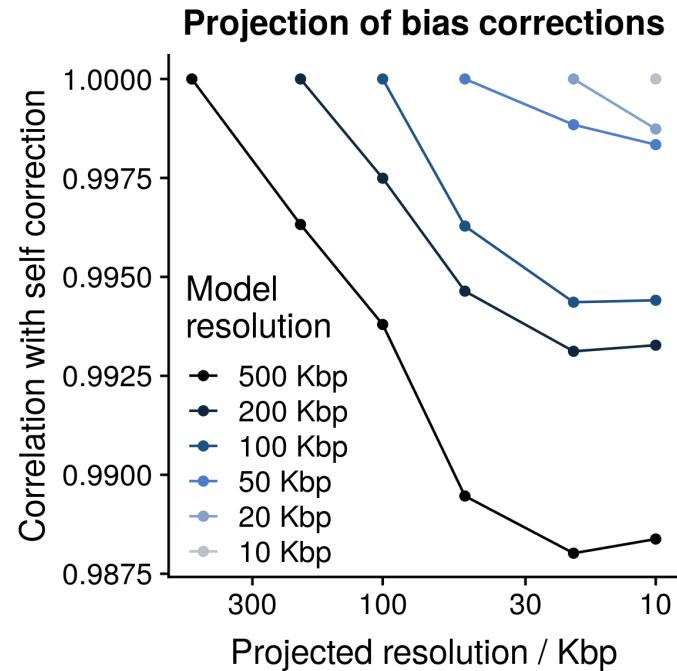
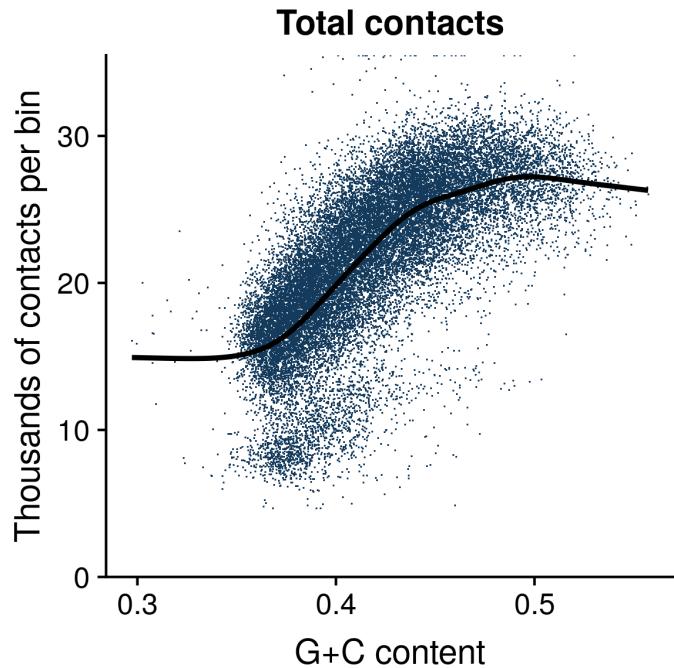
Fit once, project everywhere



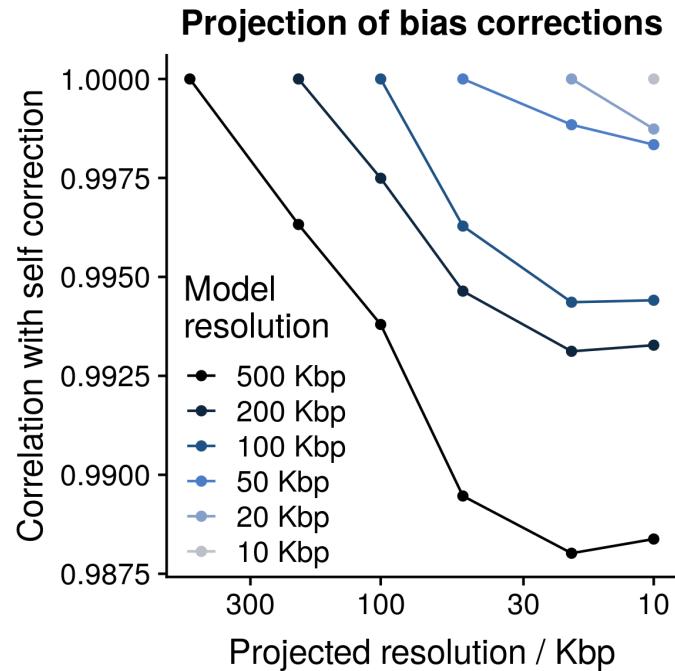
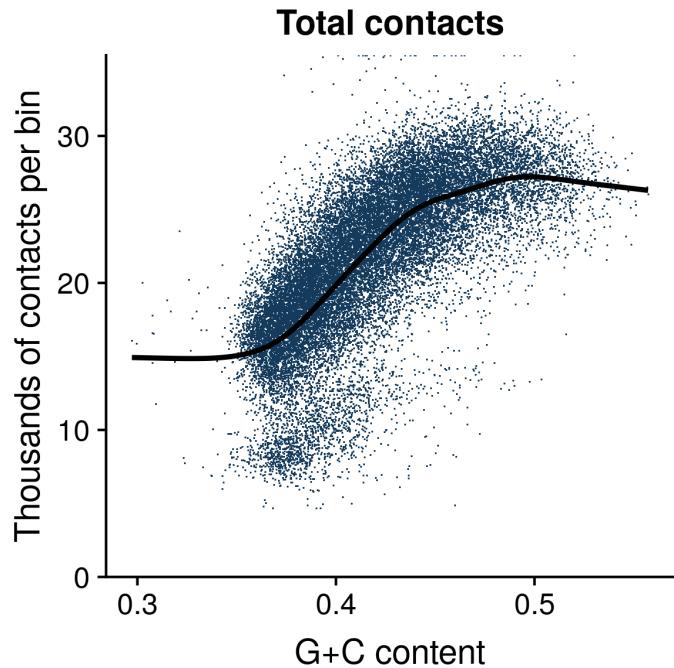
Fit once, project everywhere



Fit once, project everywhere



Fit once, project everywhere

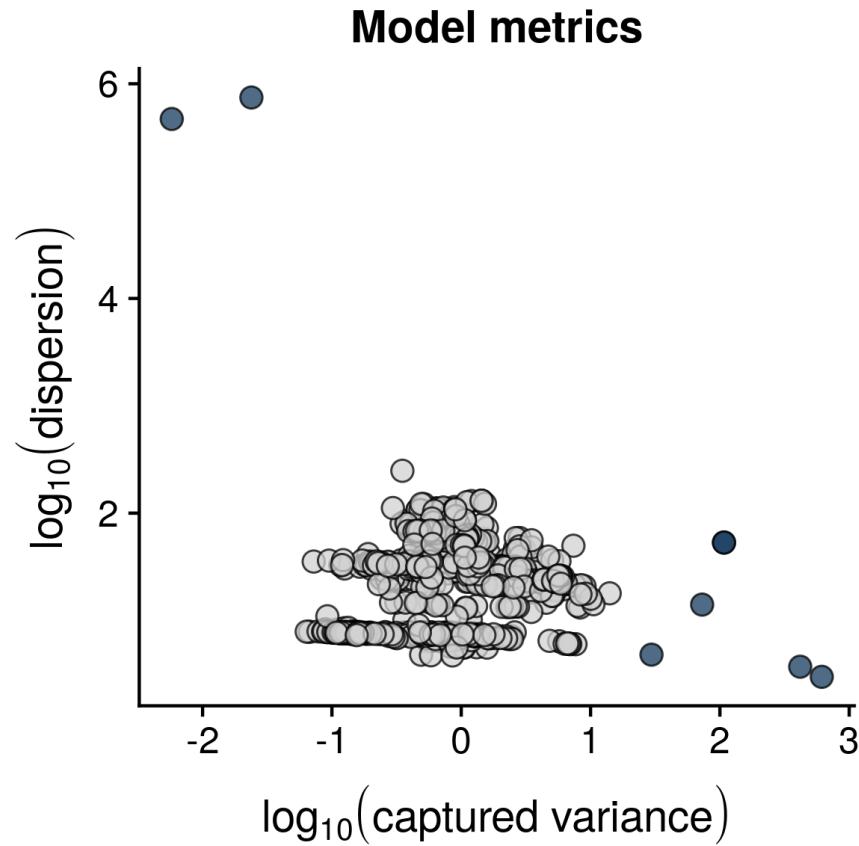


Effortless bias removal at high resolution ✓

Exploiting the model

Quality control

Model metrics as quality control



Grand summary

Grand summary

Hi-C data present biases

Grand summary

Hi-C data present biases

Use explicit modeling to reduce them

Grand summary

Hi-C data present biases

Use explicit modeling to reduce them

Exploit the model!

Grand summary

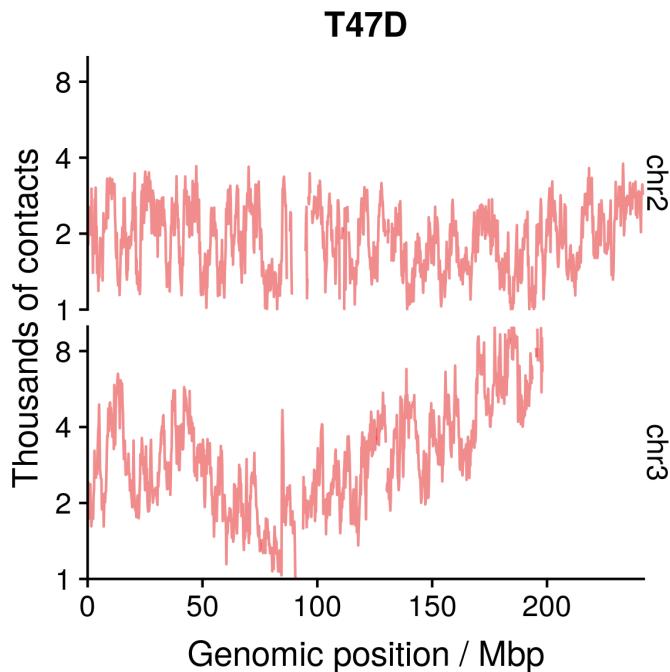
Hi-C data present biases

Use explicit modeling to reduce them

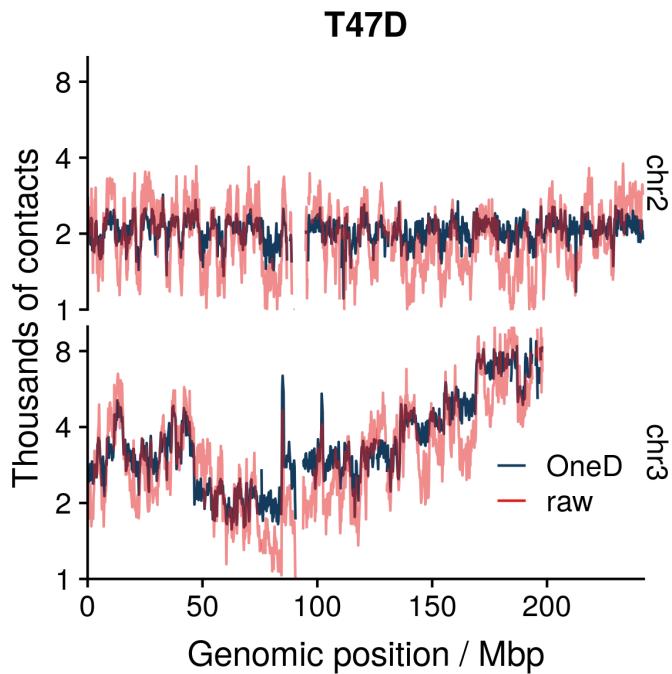
Exploit the model!

OneD:  [qenvio/dryhic](#)

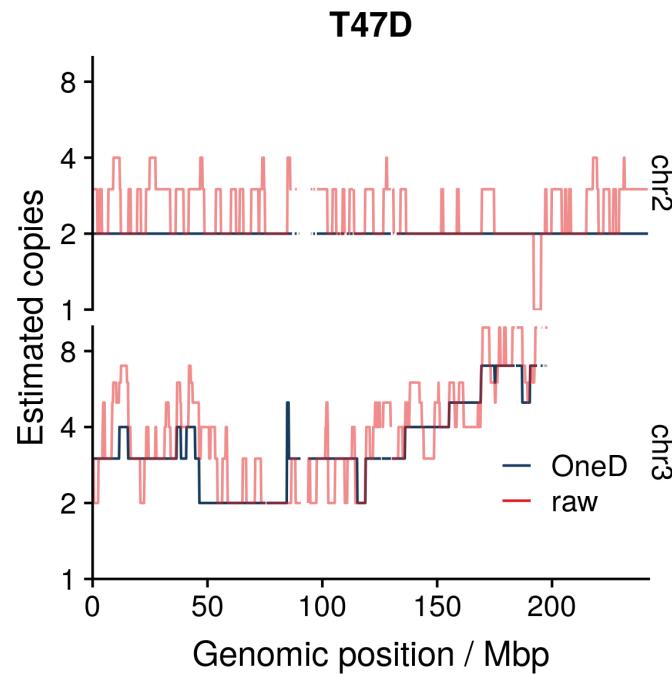
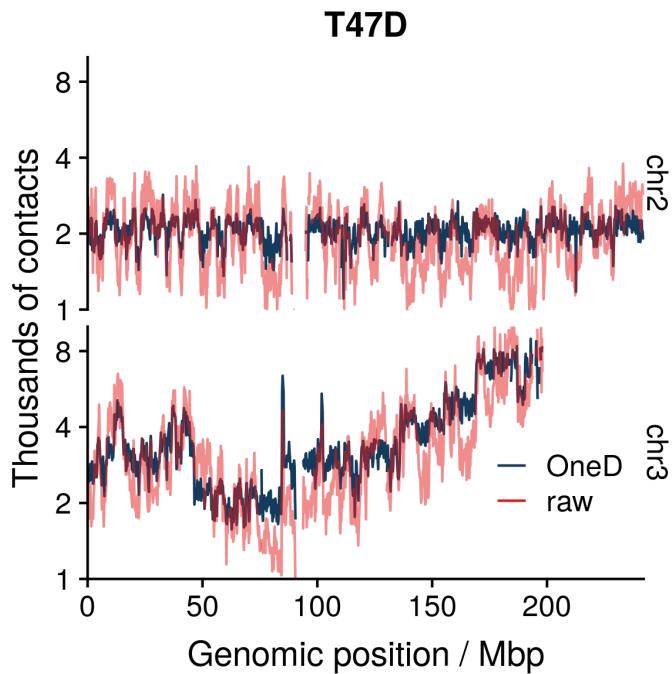
Total contacts → copy number



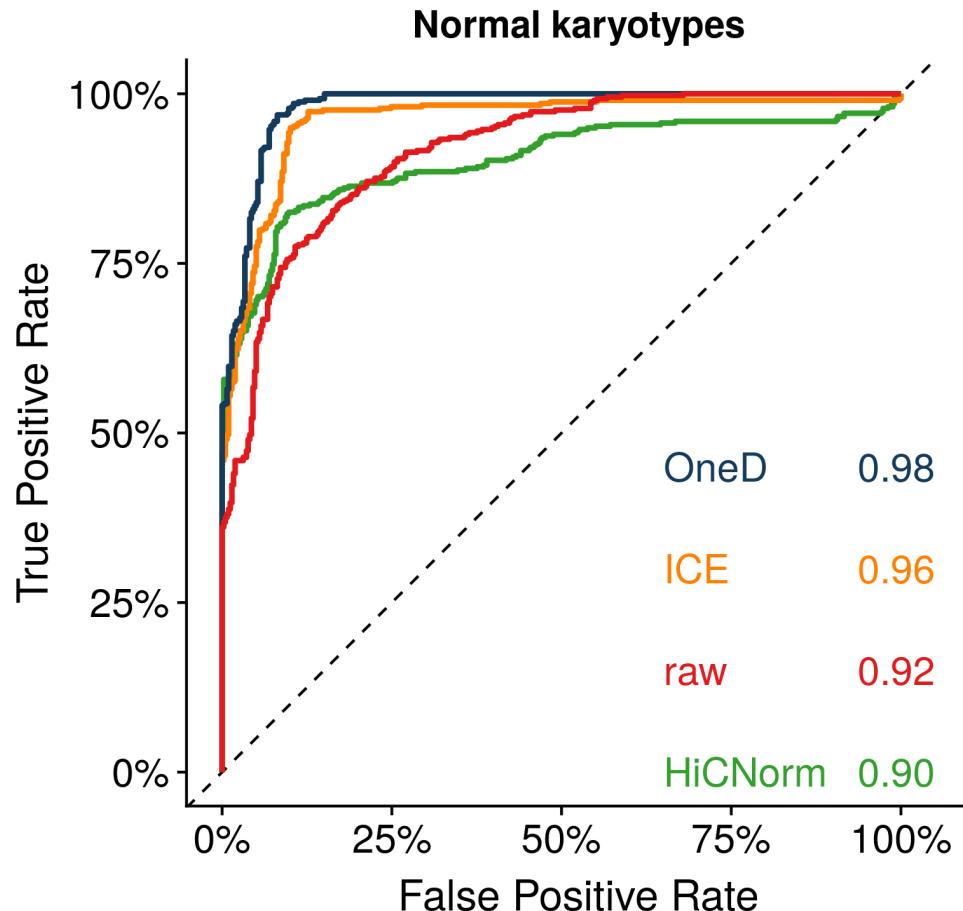
Total contacts → copy number



Total contacts → copy number



ROC and AUC normal karyotypes



PRC and AUC

