Interrogating genome structure

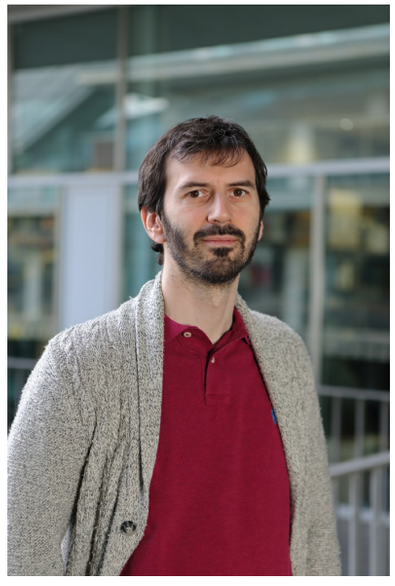
Reproducibility and biases

Münster, 12 July 2018

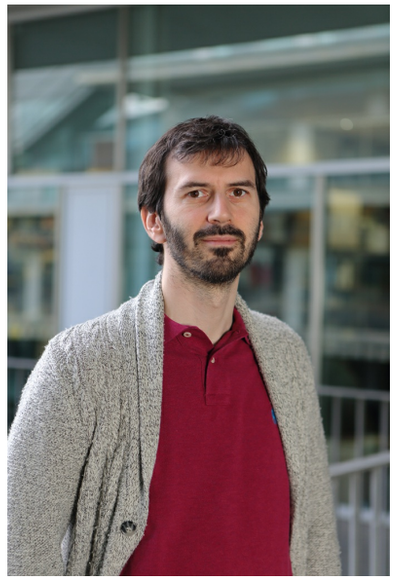
Enrique (Quique) Vidal
enrique.vidal@crg.eu

@qenvio qenvio





4DGenome

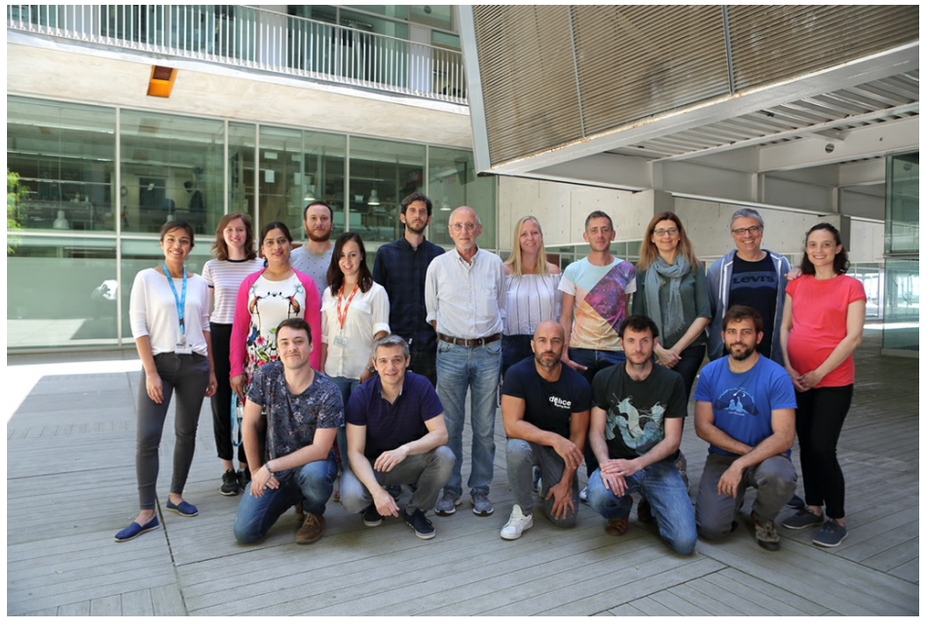
European Research Council

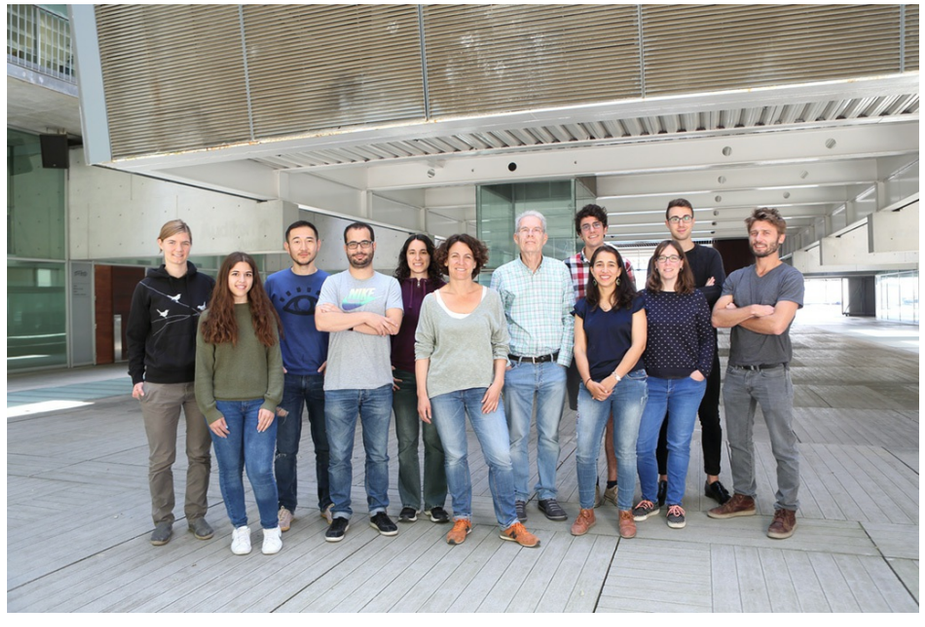
Centre for Genomic Regulation

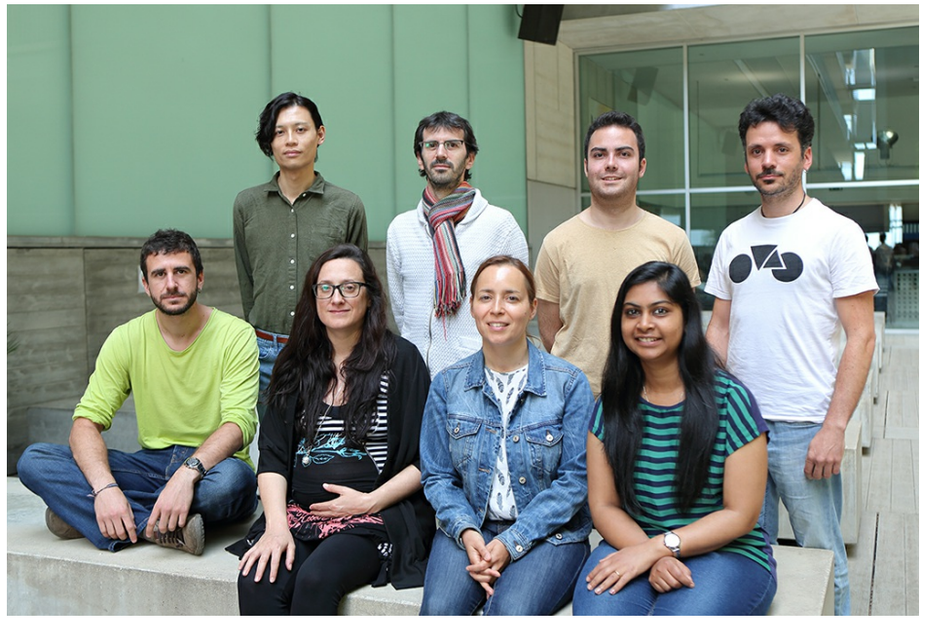
cnag

centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

4DGenome

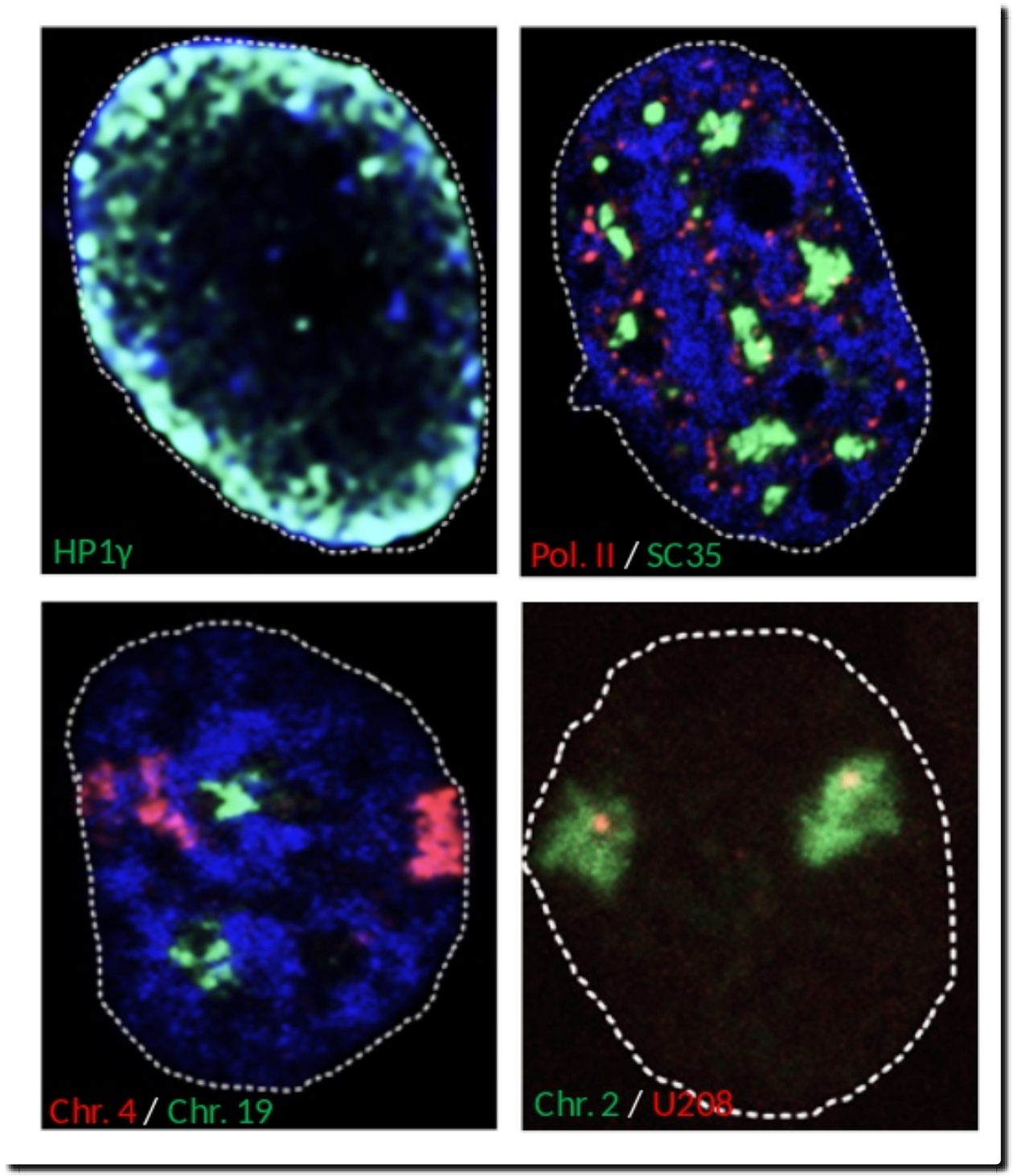






Motivation

Genome structure is not random



Chromosome conformation capture

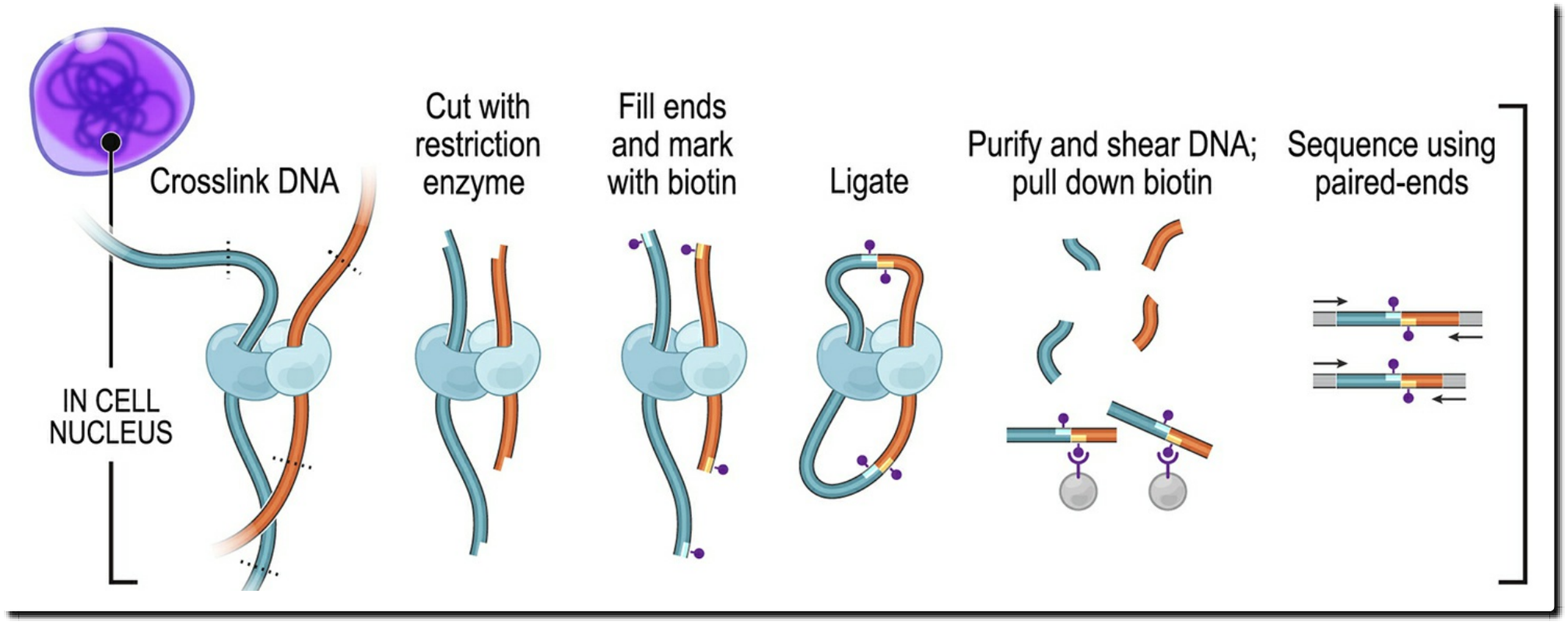
Dekker, J. et al. (2002)

Science

Chromosome conformation capture

Dekker, J. et al. (2002)

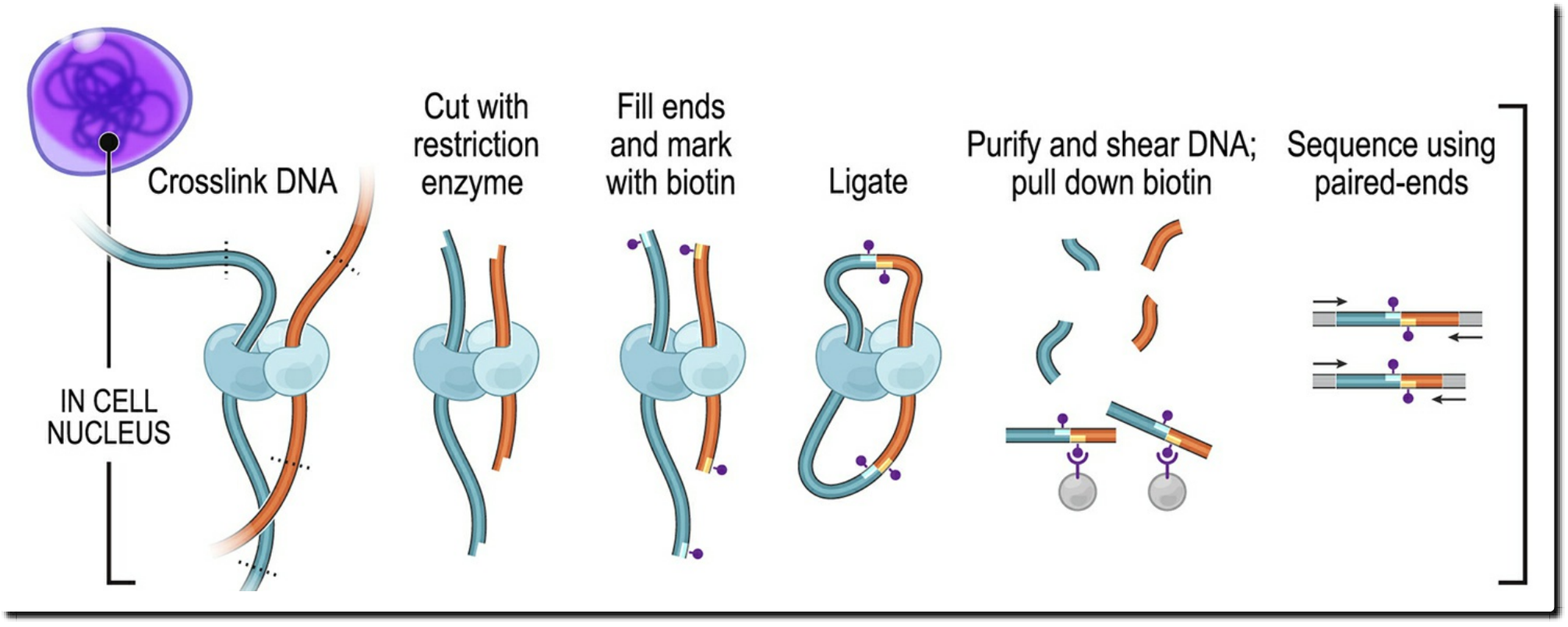
Science



Chromosome conformation capture

Dekker, J. et al. (2002)

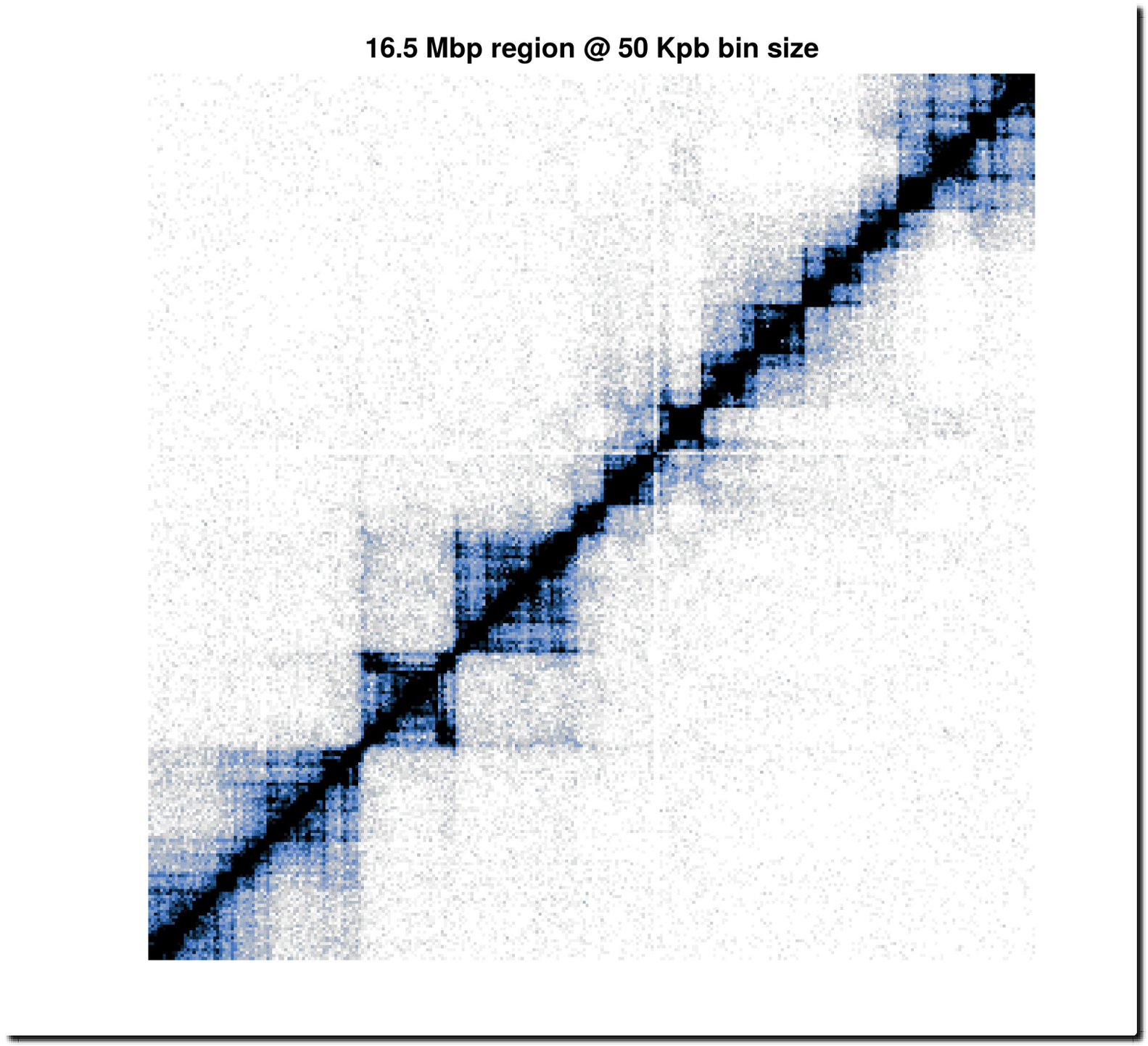
Science



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

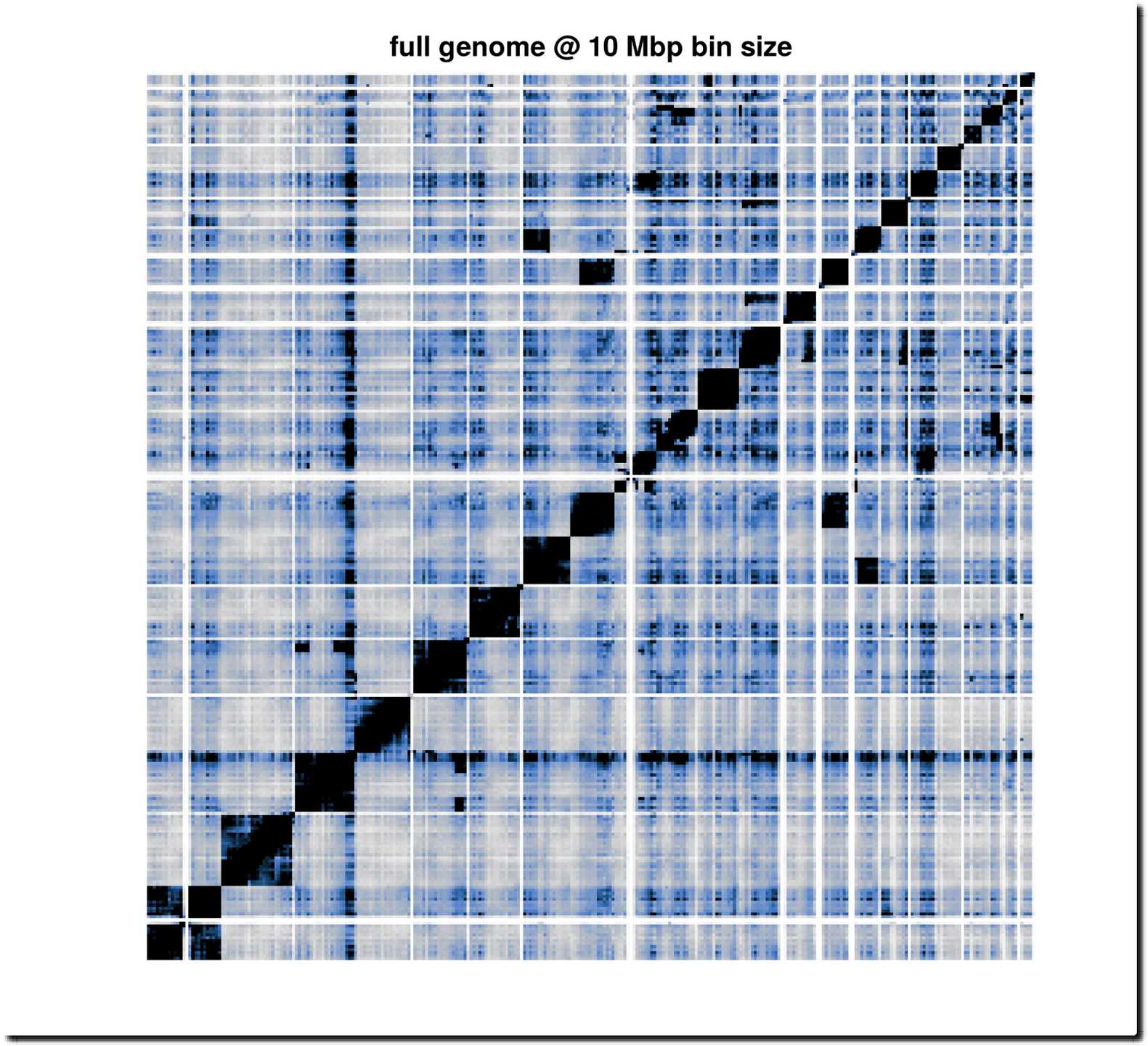
Science

Hi-C

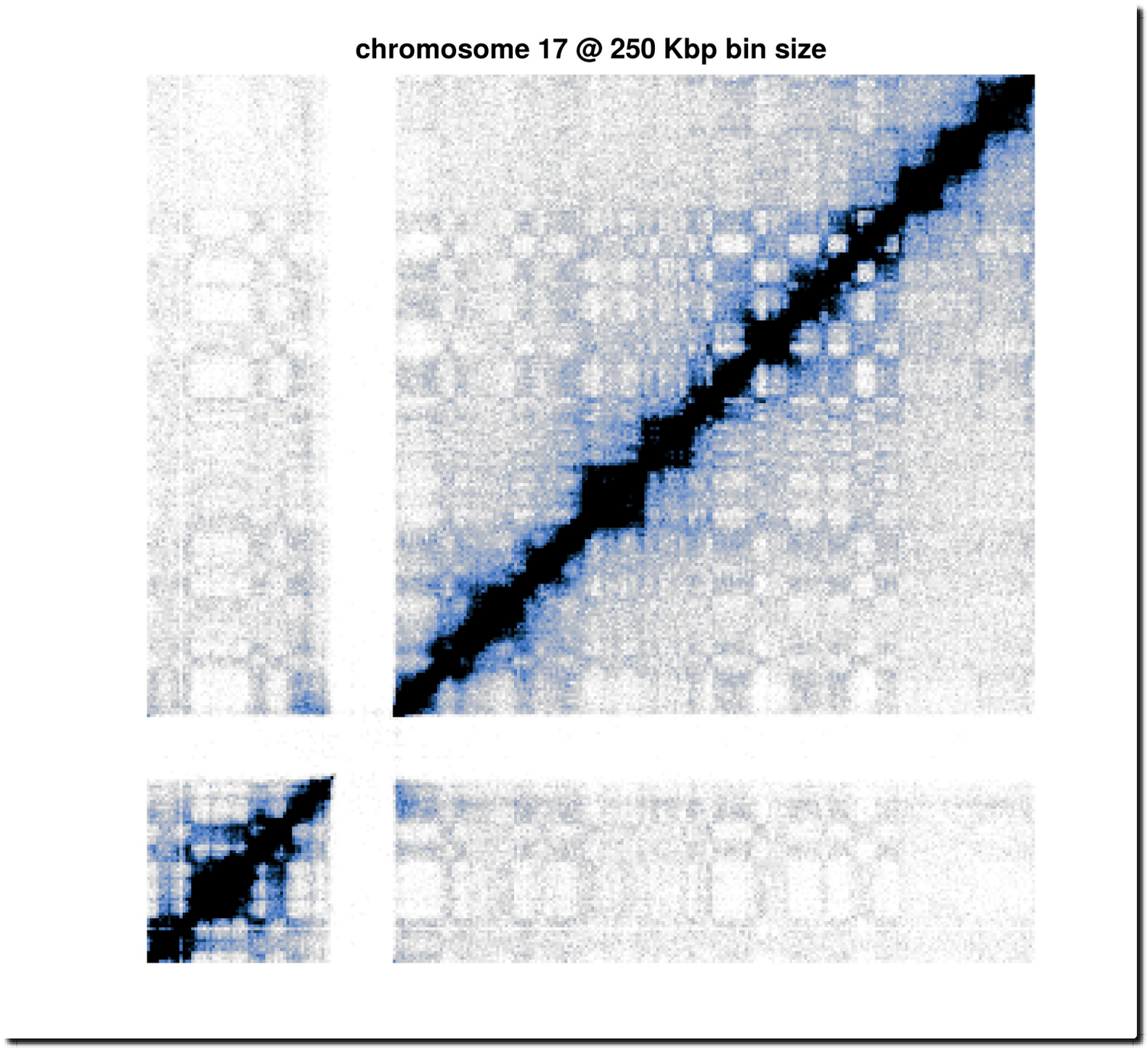


What can we get from Hi-C?

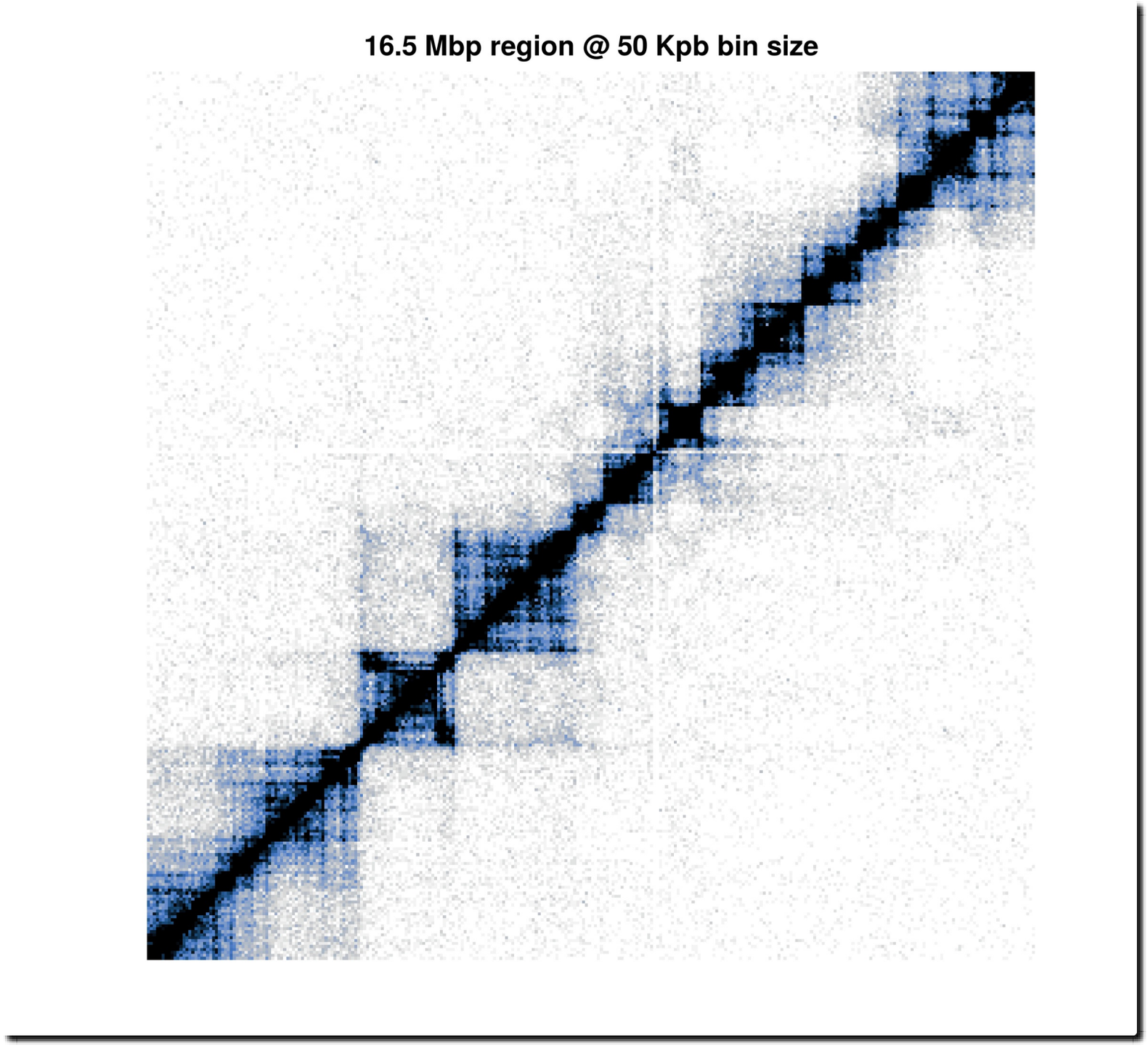
Territories



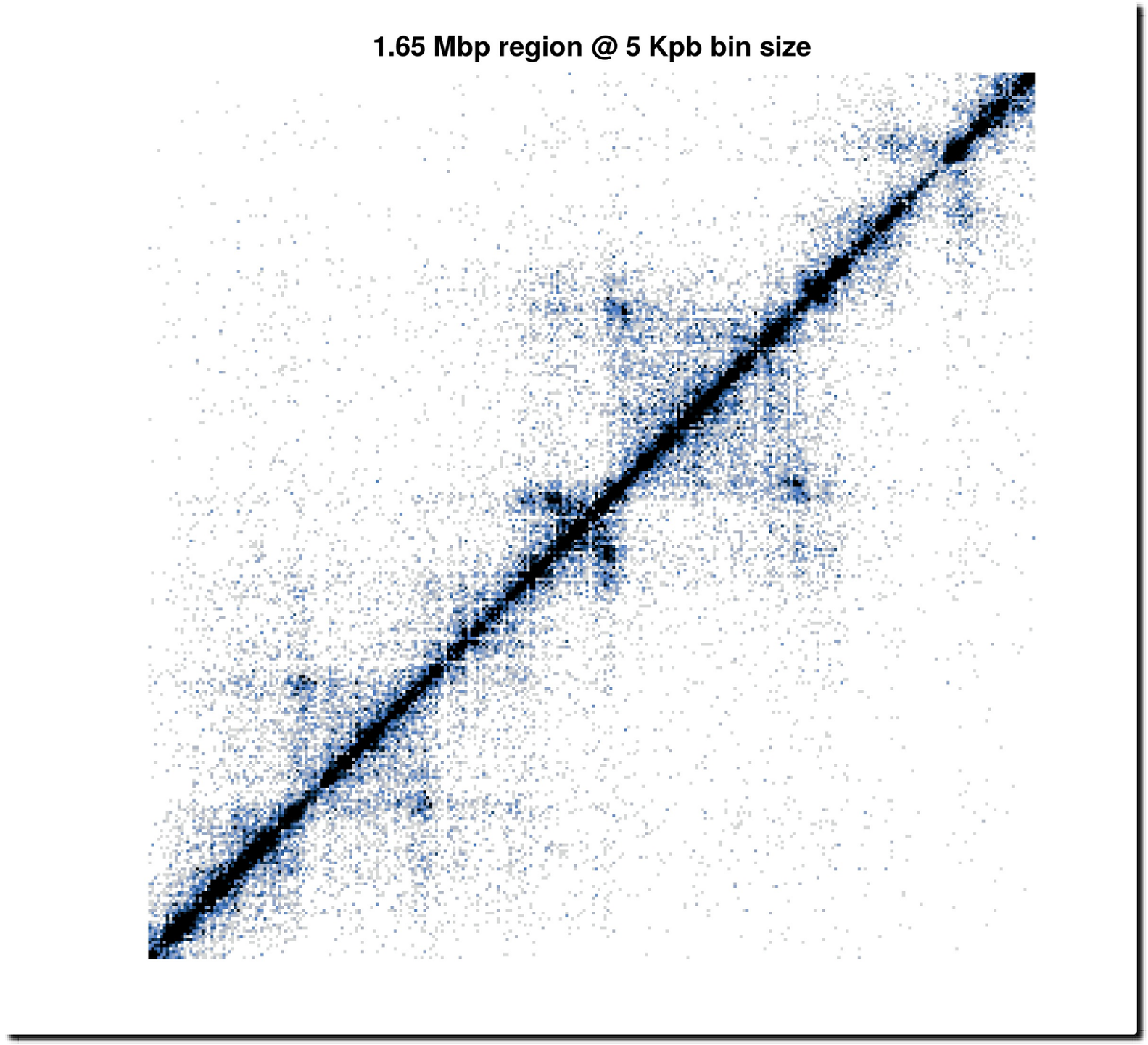
Compartments



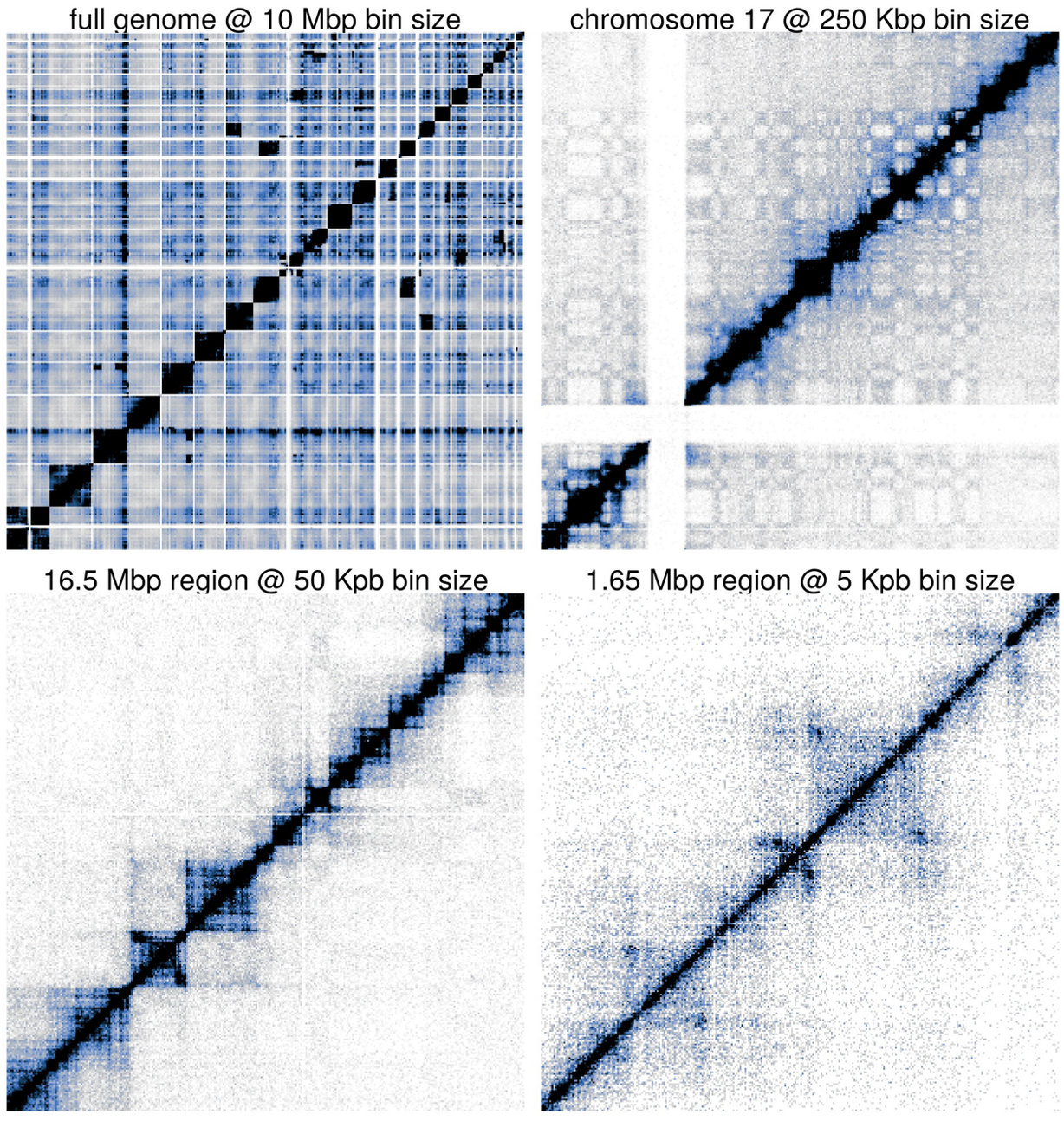
Domains (TADs)



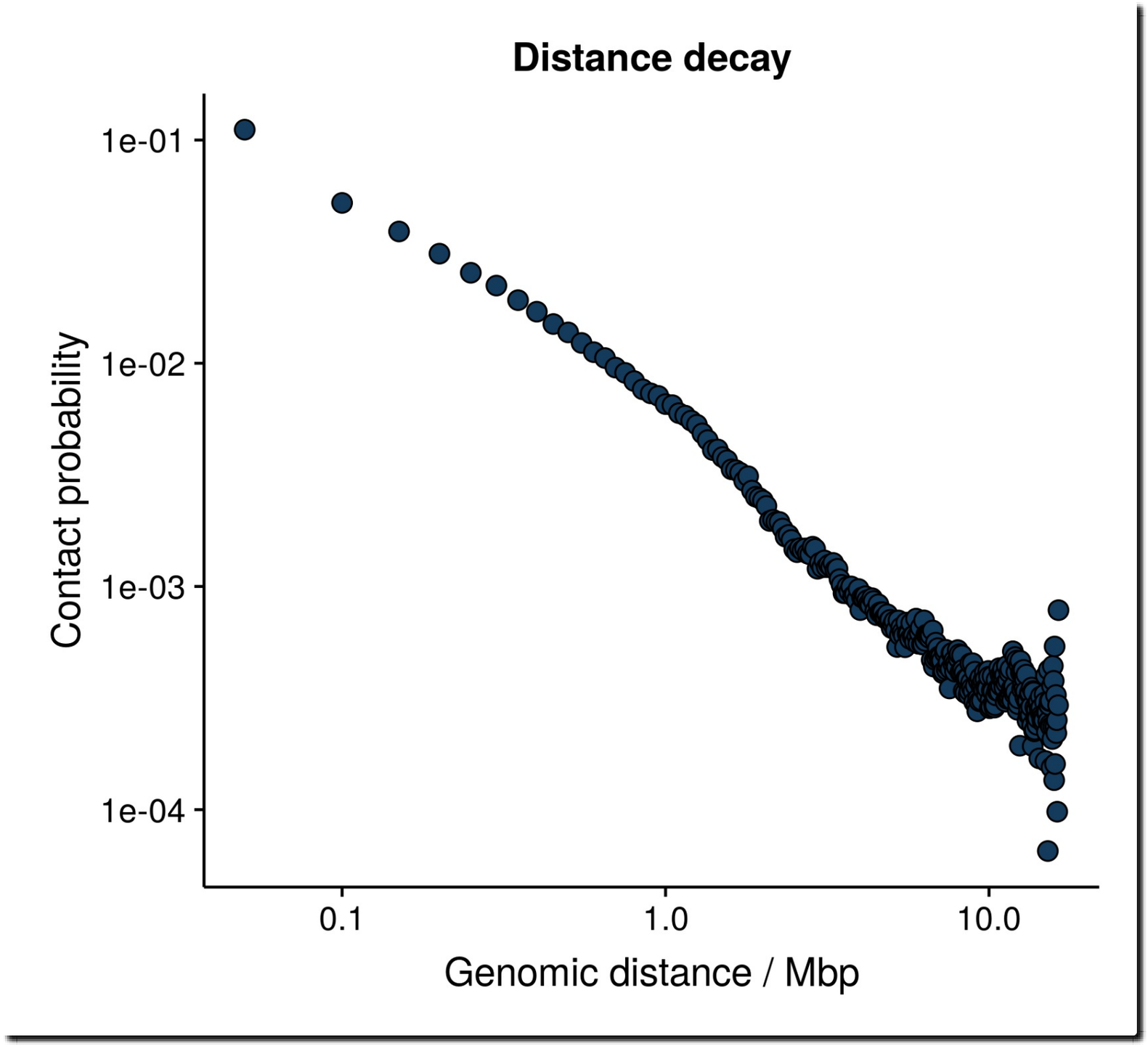
Loops



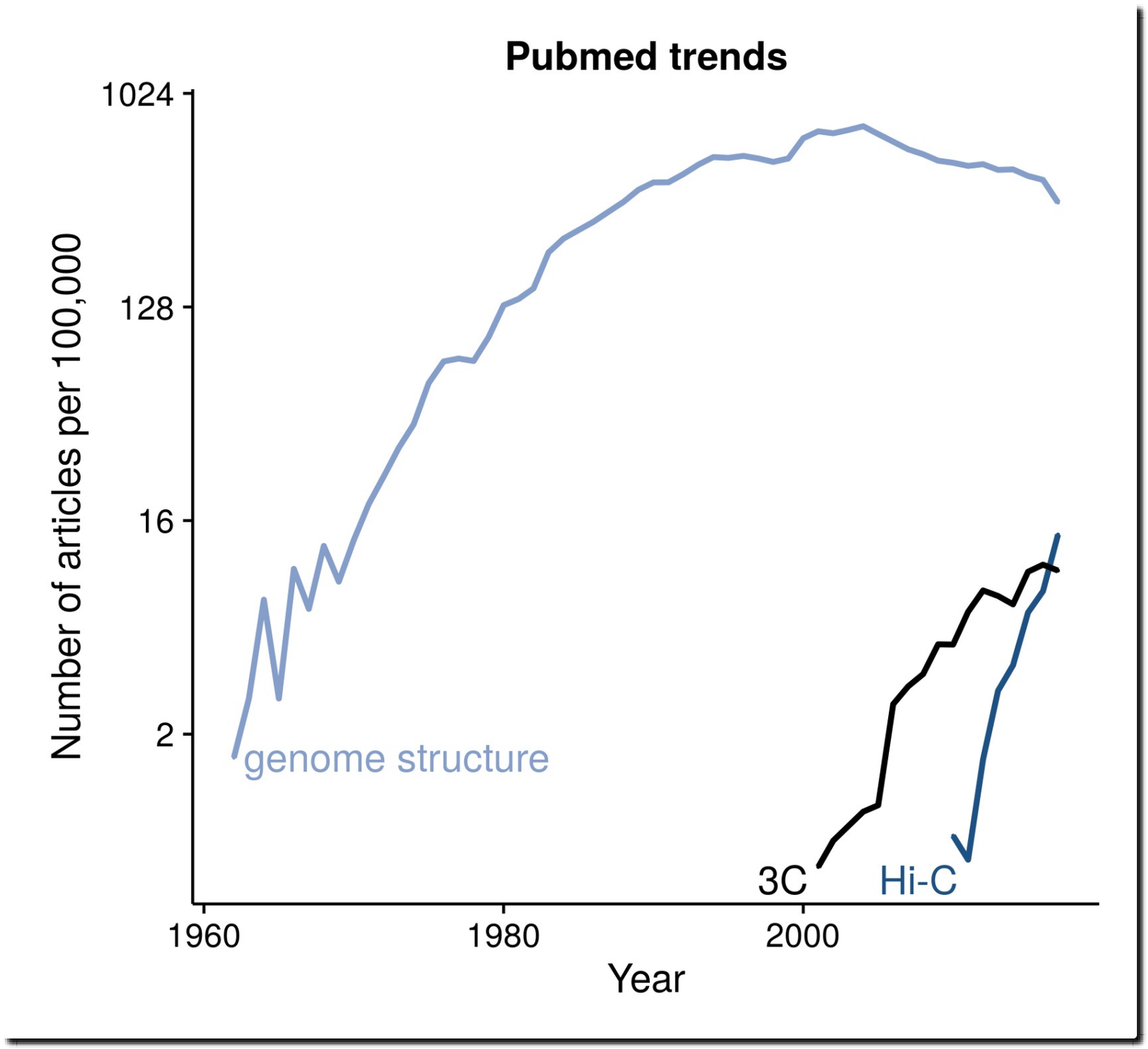
In a nutshell



Distance decay



Papers


Pubmed trends
1024
128
16
2
Number of articles per 100,000
genome structure
3C
Hi-C
1960
1980
2000
Year

Papers

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

Science

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

Nature genetics

Imakaev, M. et al. (2012)

Nature methods

Dixon, J. R. et al (2012)

Nature

Nora, E. P. et al. (2012)

Nature

Rao, S.S. et al. (2014)

Cell

Hug, B. H. (2017)

Cell

Stadhouders, R. and Vidal, E. et al (2018)

Naure genetics

Challenge 1

Systematic biases

Genomic features affect Hi-C

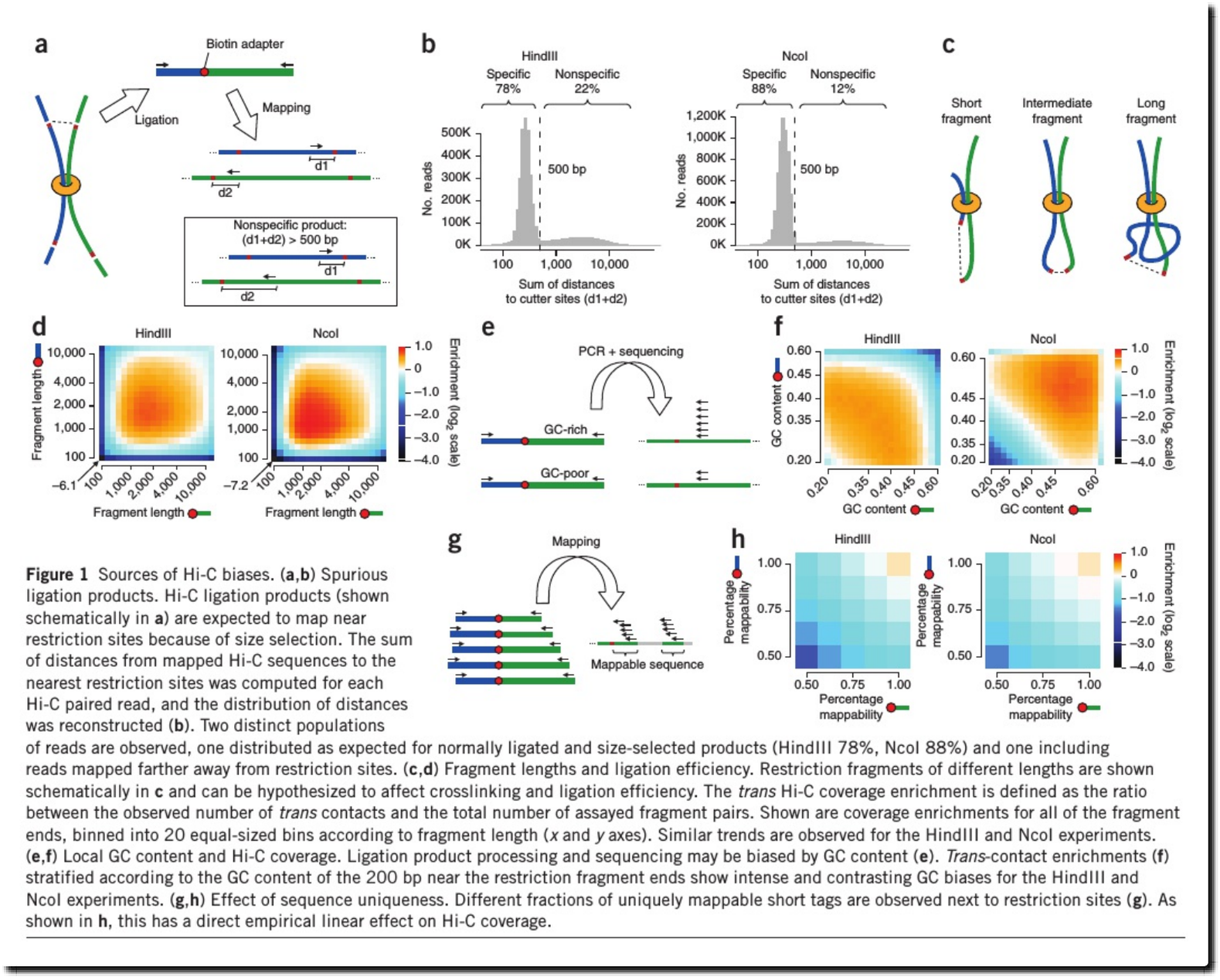
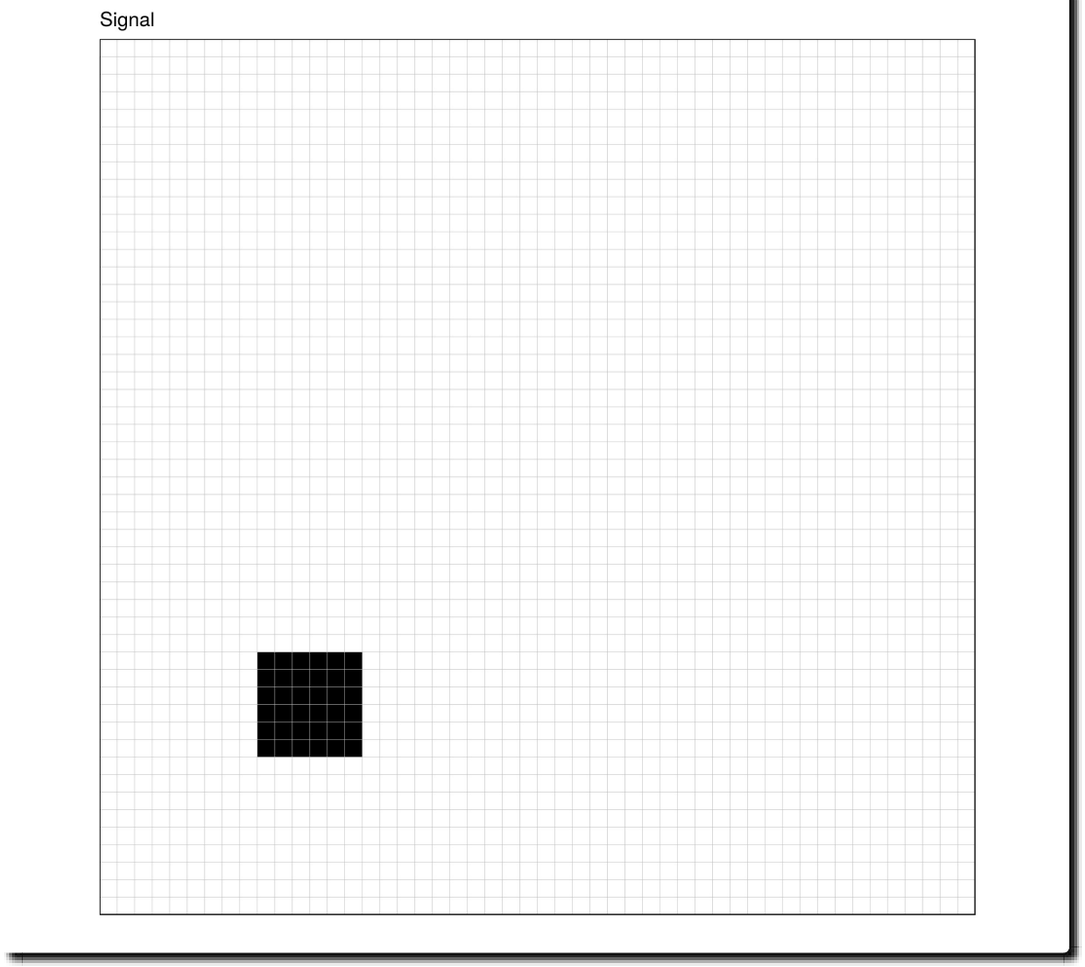


Figure 1 Sources of Hi-C biases. (**a**,**b**) Spurious ligation products. Hi-C ligation products (shown schematically in **a**) are expected to map near restriction sites because of size selection. The sum of distances from mapped Hi-C sequences to the nearest restriction sites was computed for each 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%, NcoI 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 NcoI 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 NcoI 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.

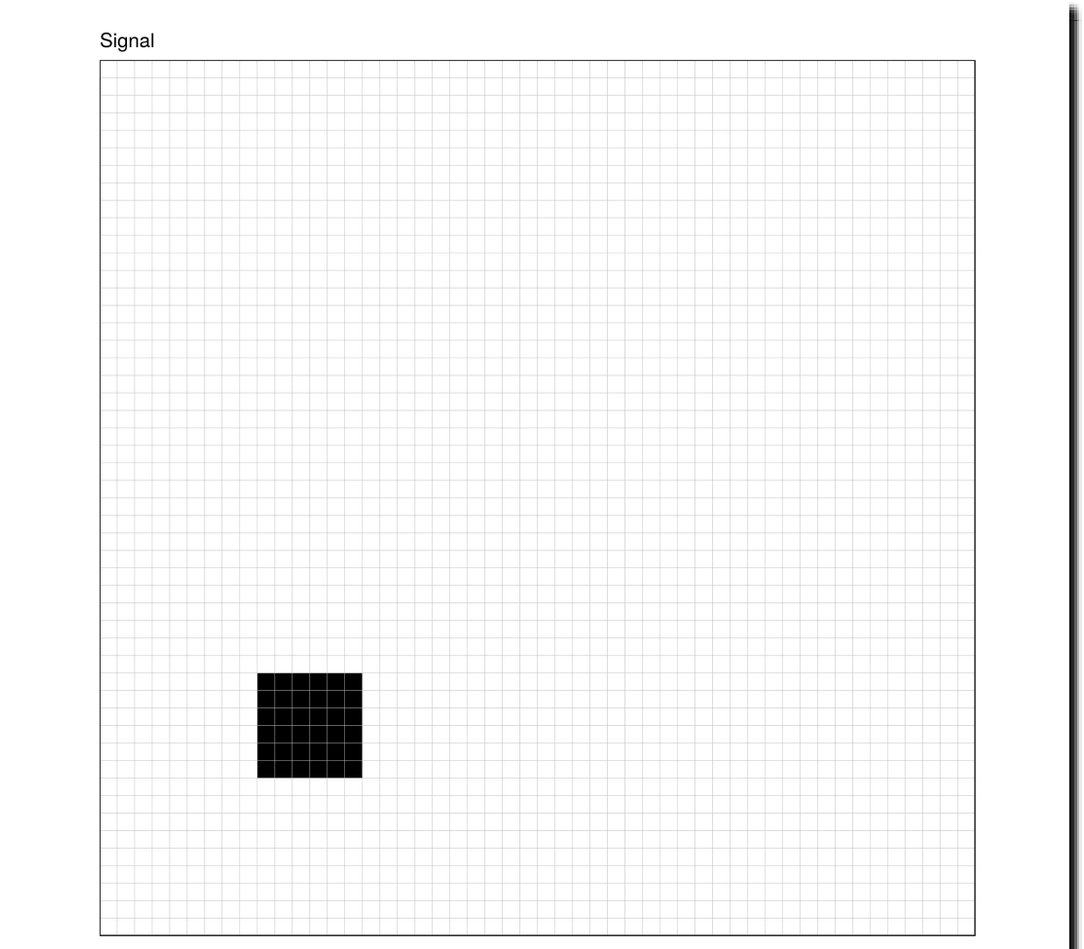
Yaffe, E. and Tanay, A. (2011)

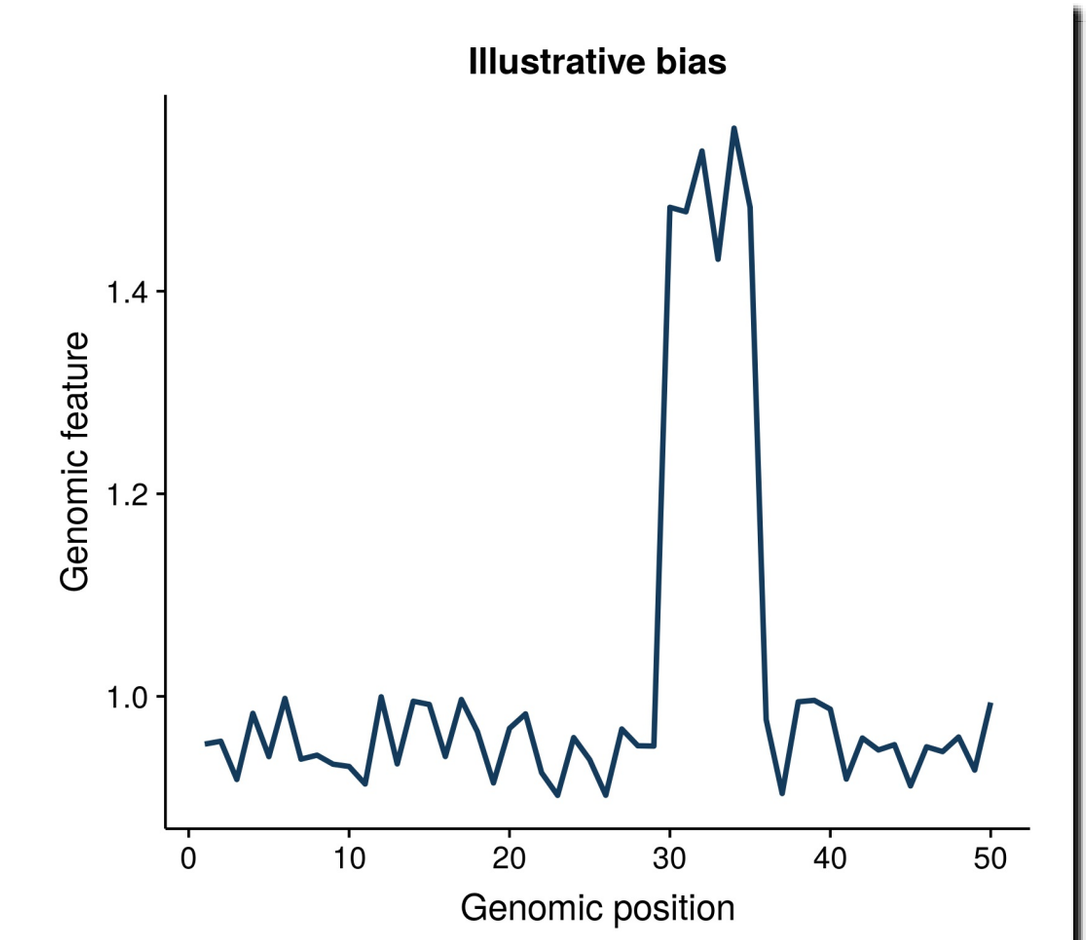
Nature genetics

Illustrating biases



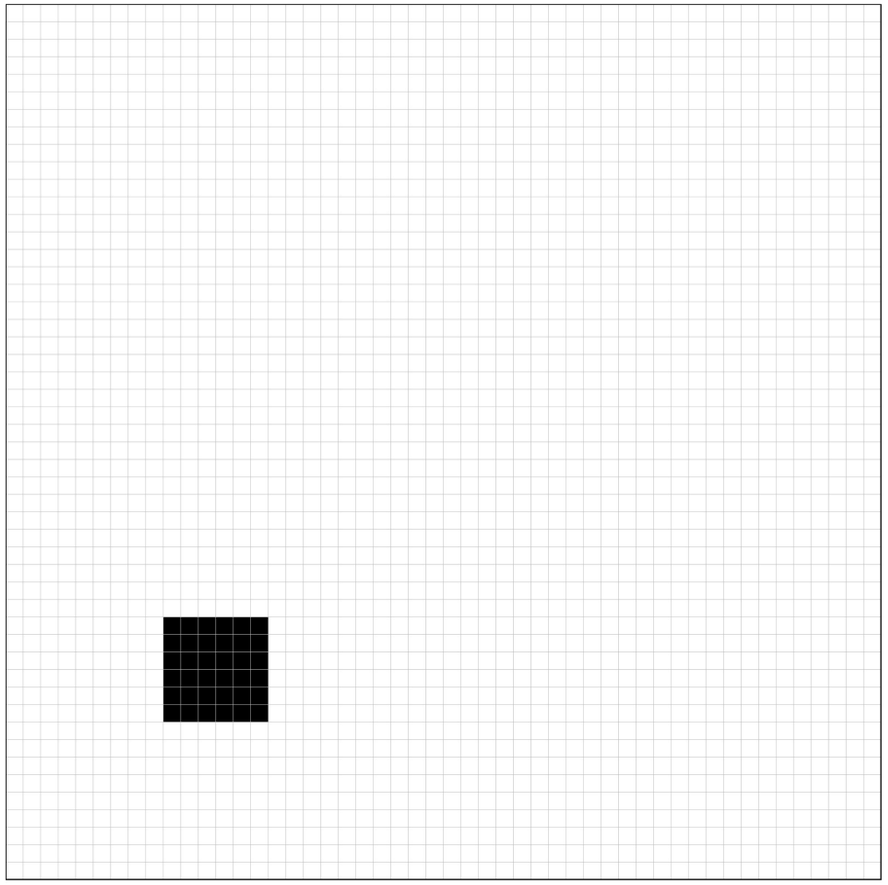
Illustrating biases



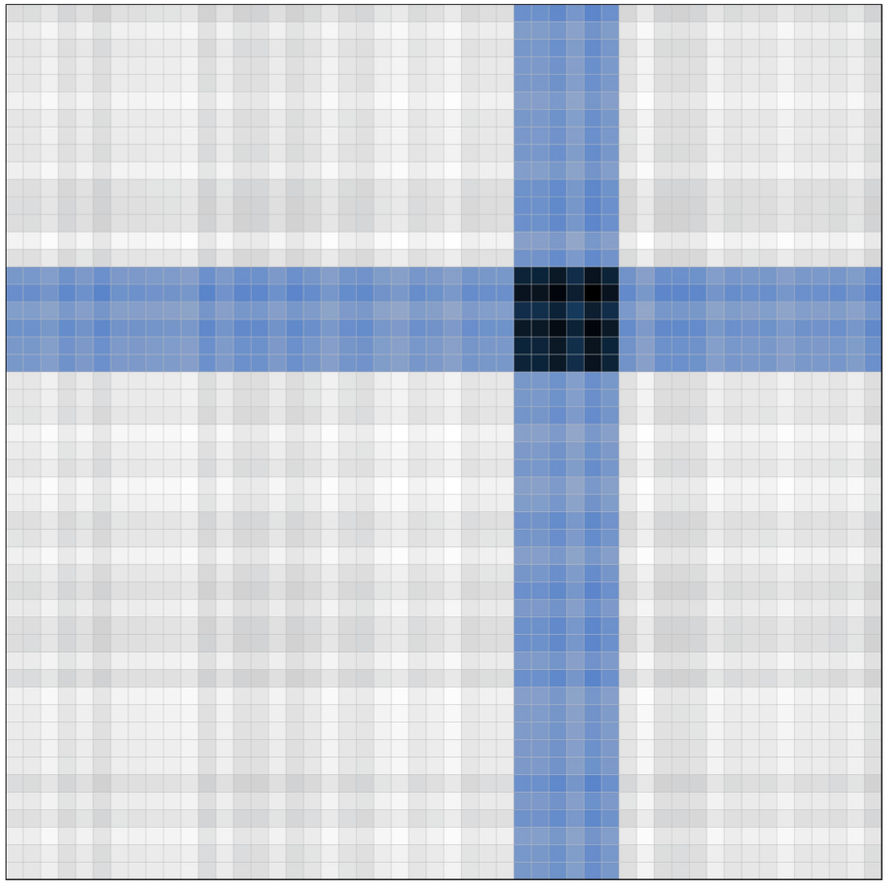


Illustrating biases

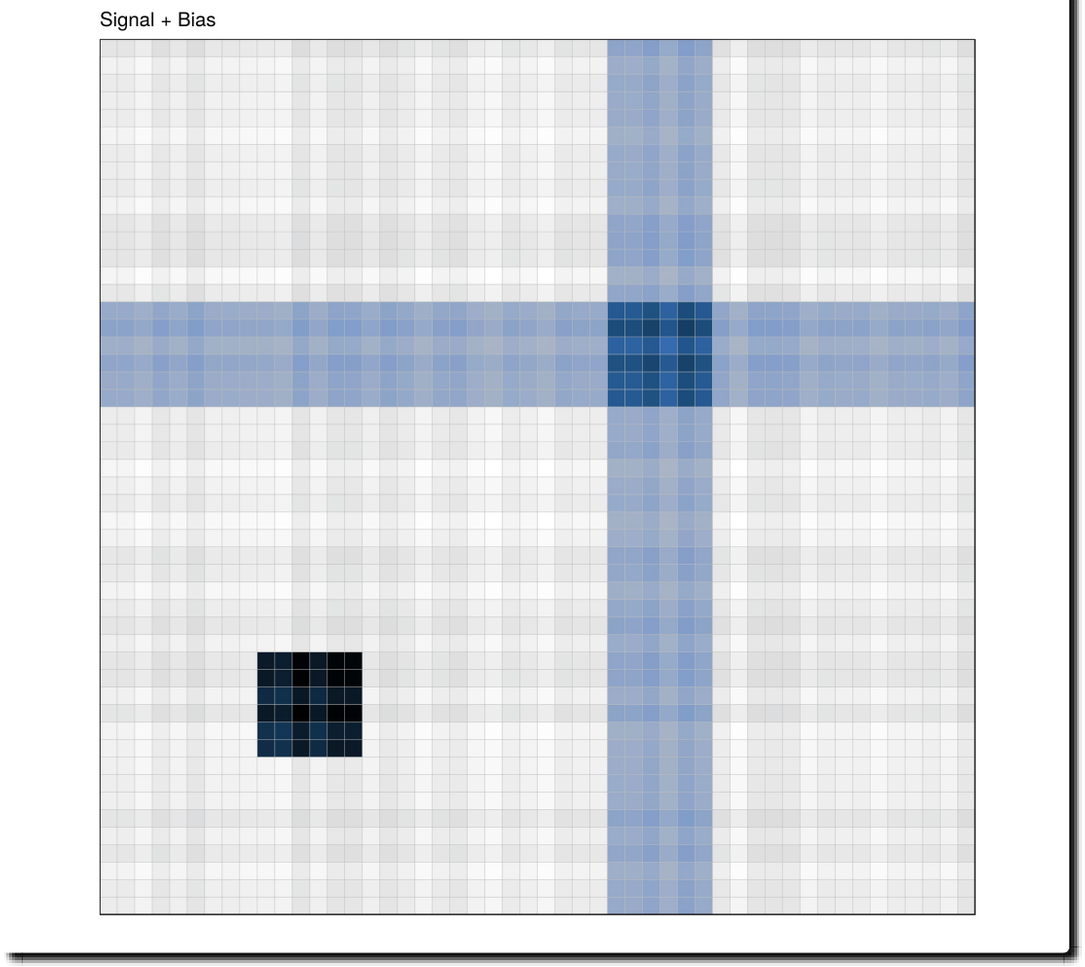
Signal



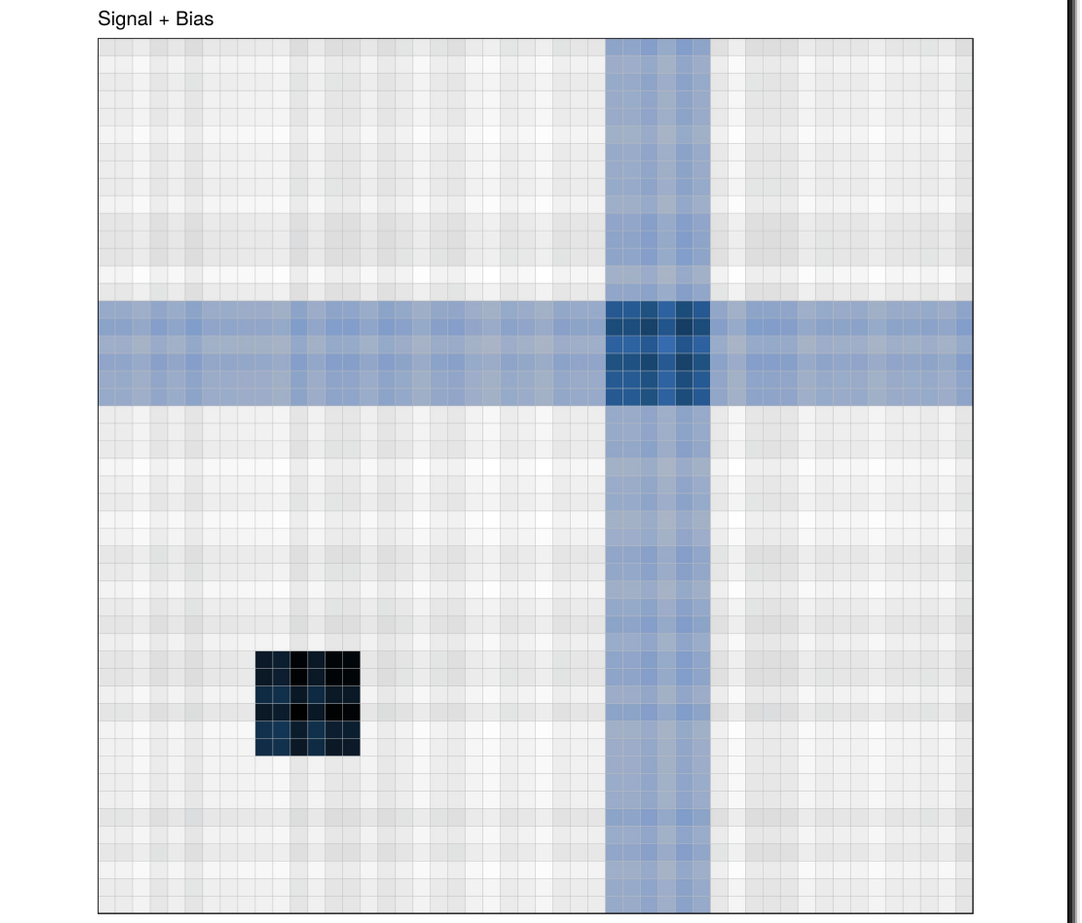
Bias

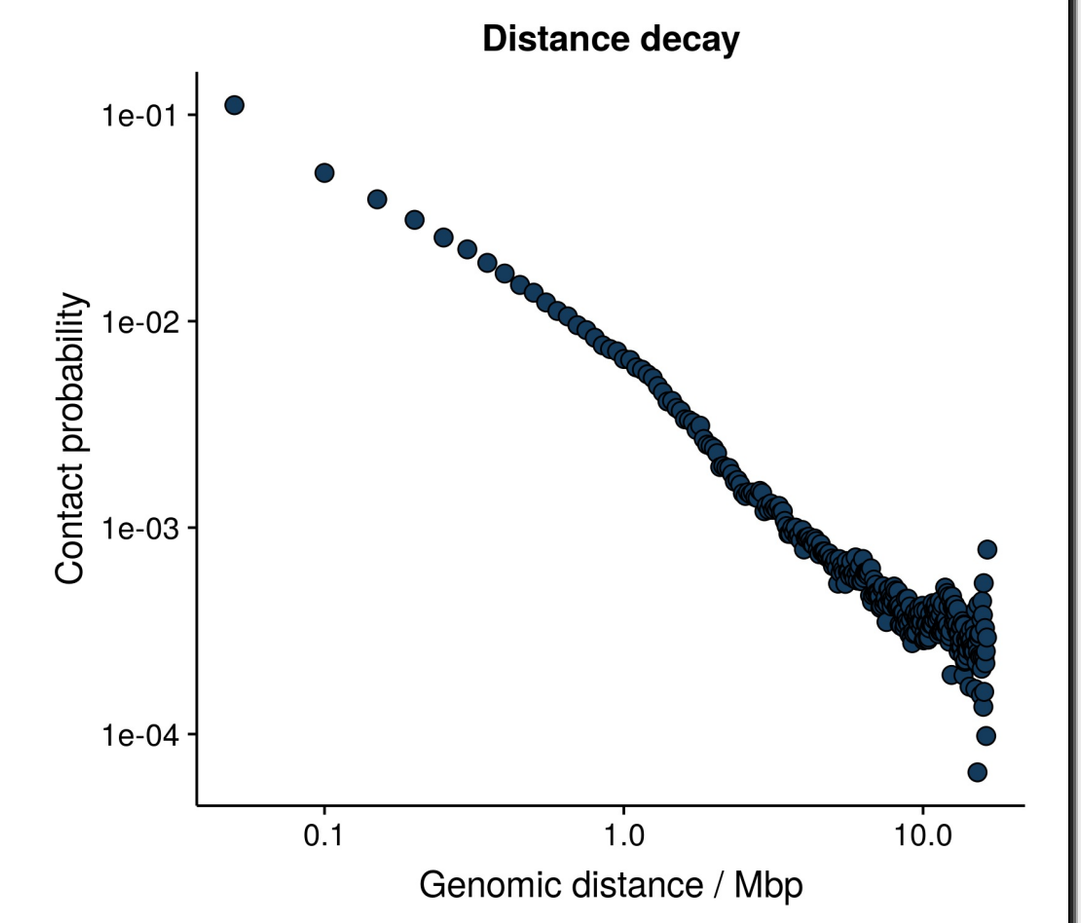


Illustrating biases



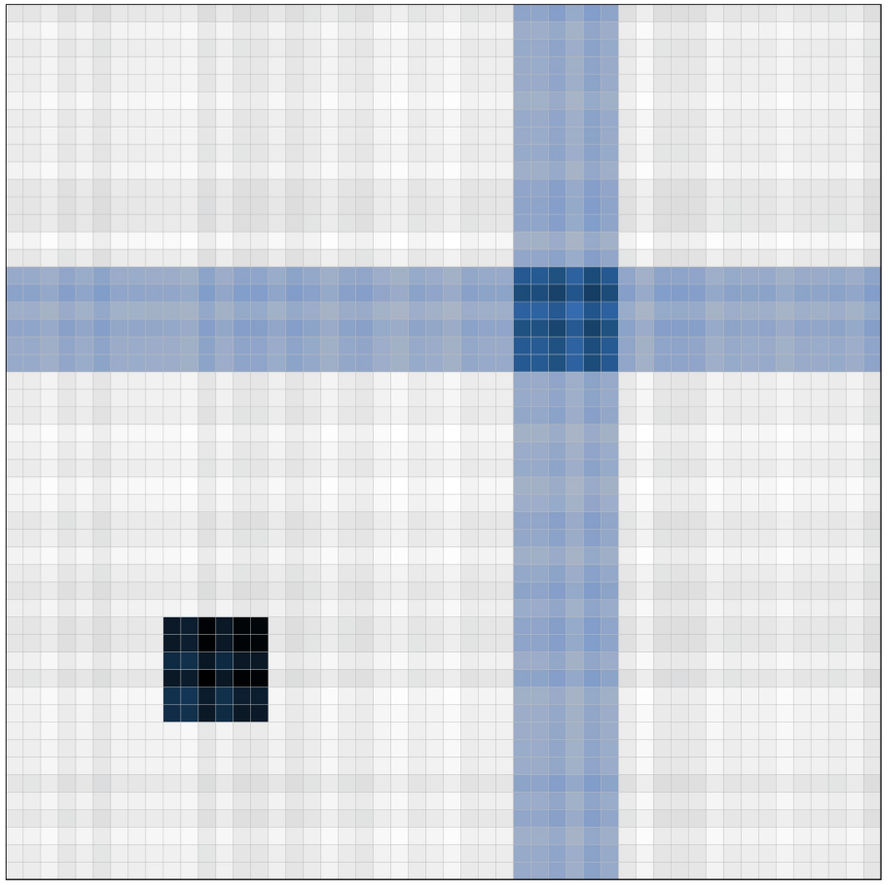
Illustrating biases



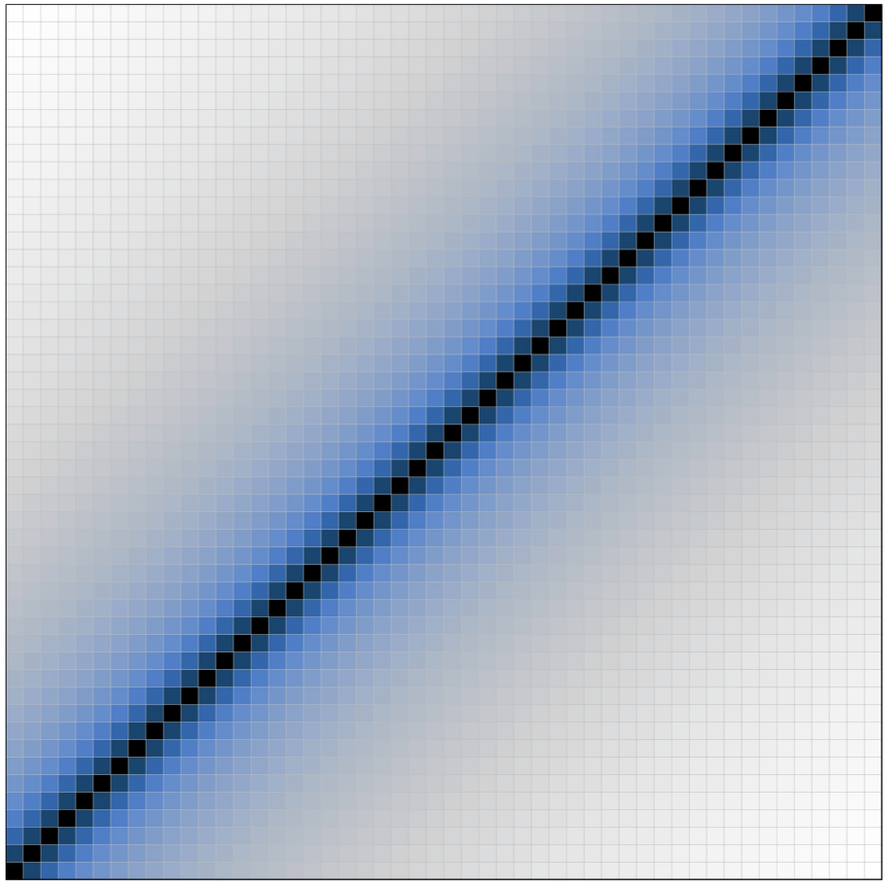


Illustrating biases

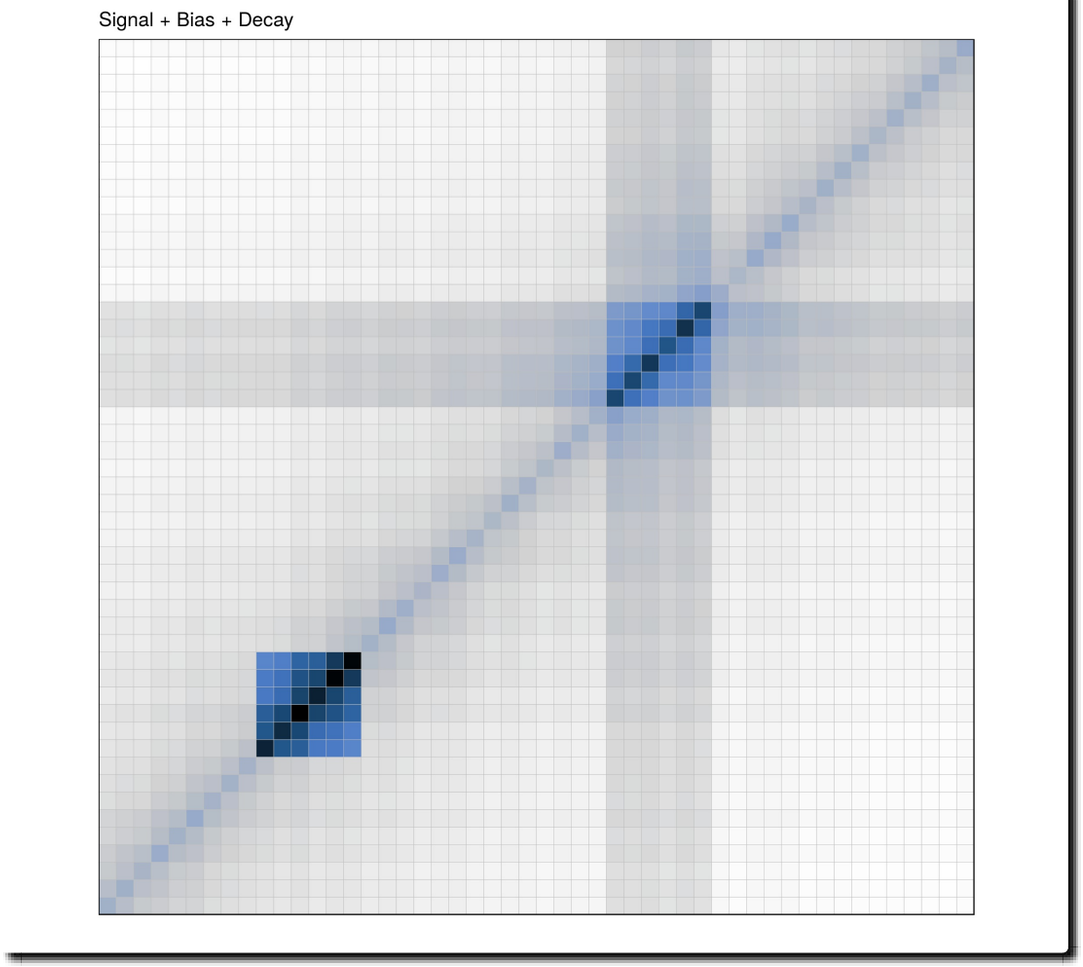
Signal + Bias



Decay

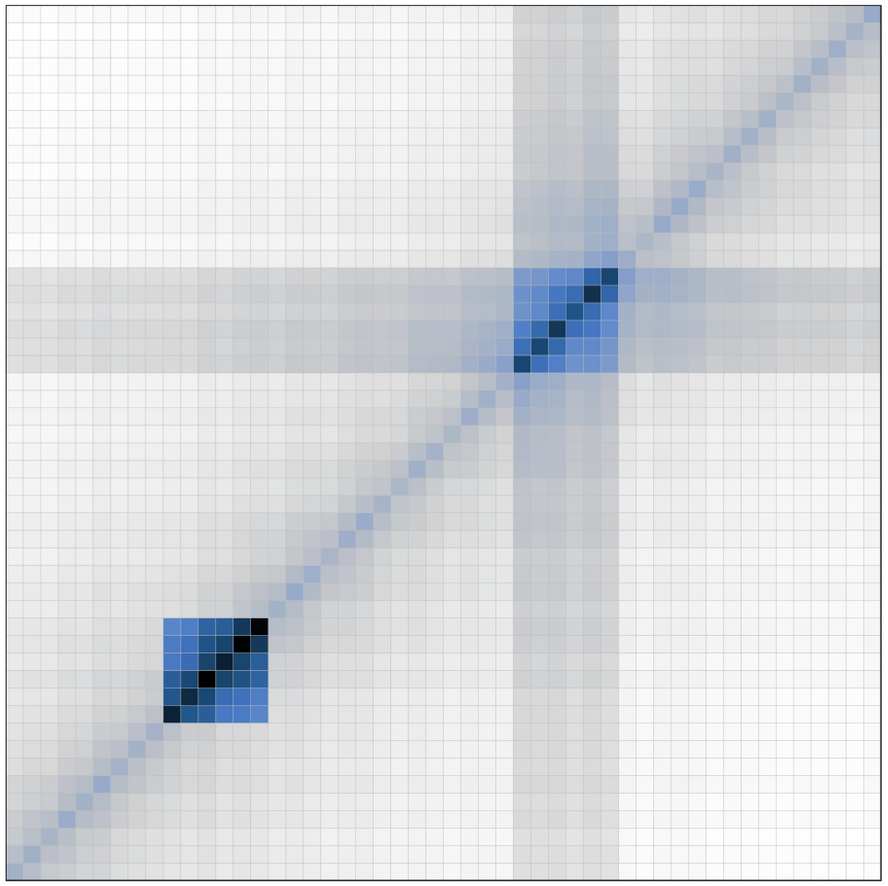


Illustrating biases

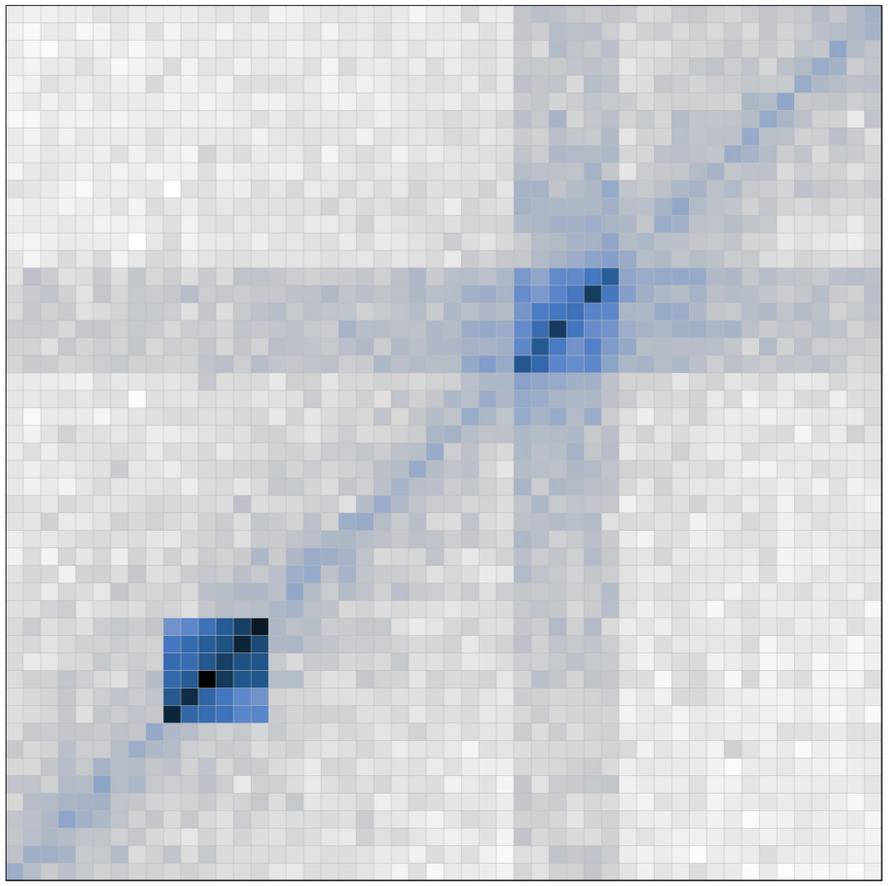


Illustrating biases

Signal + Bias + Decay

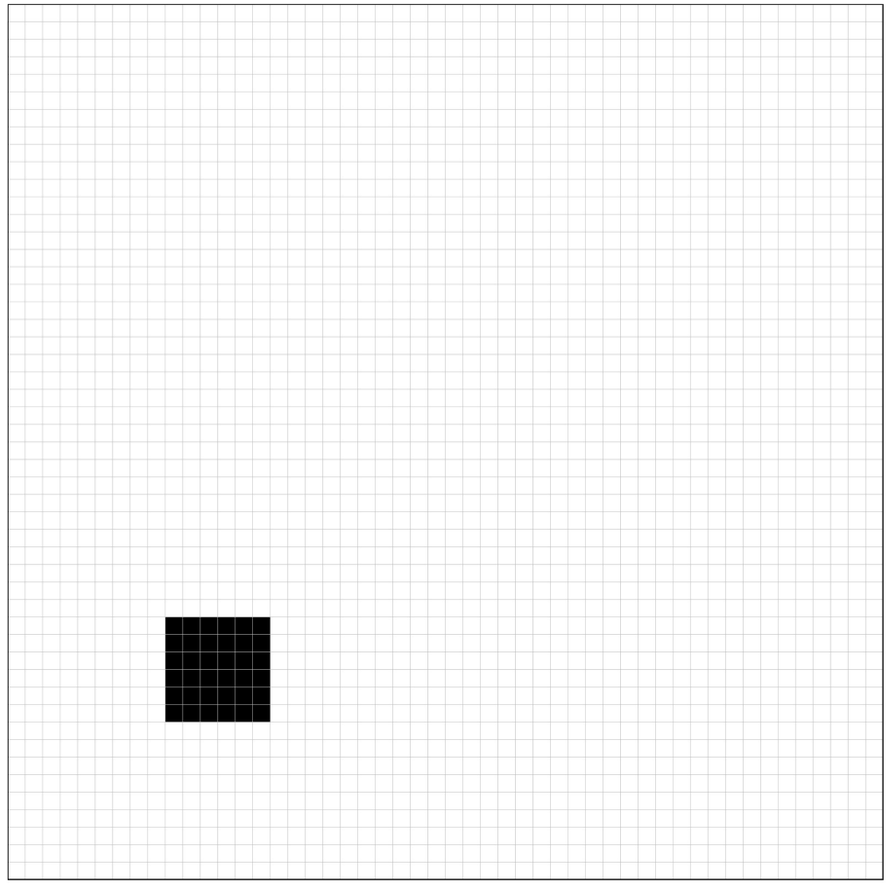


Signal + Bias + Decay + Noise

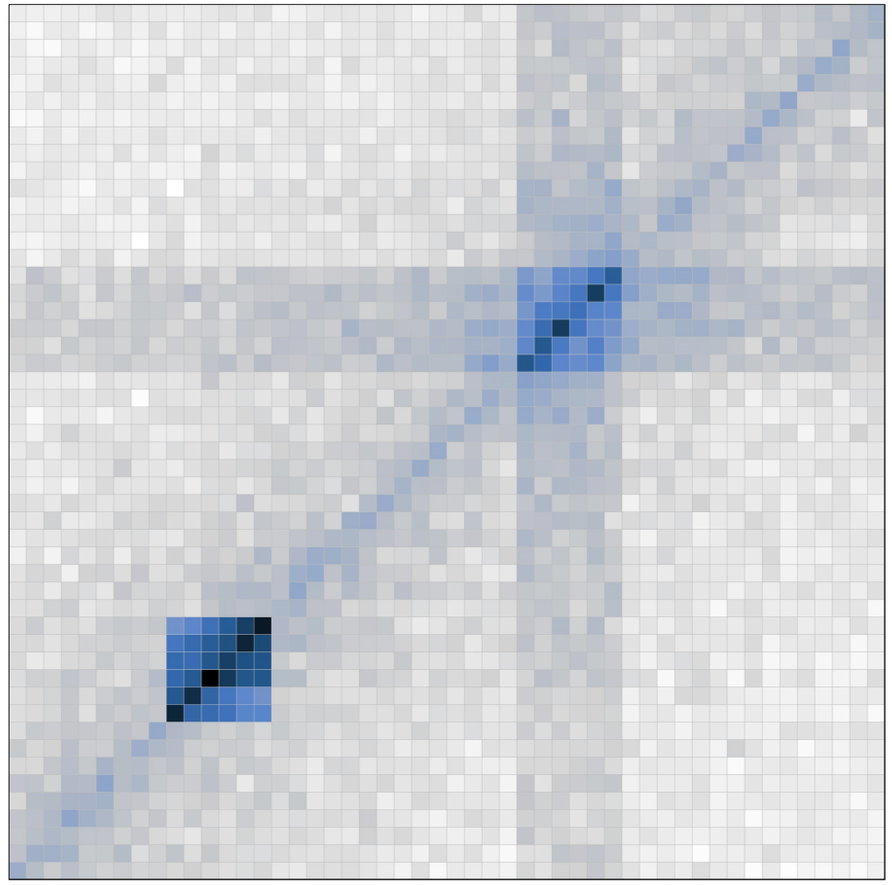


Illustrating biases

Signal



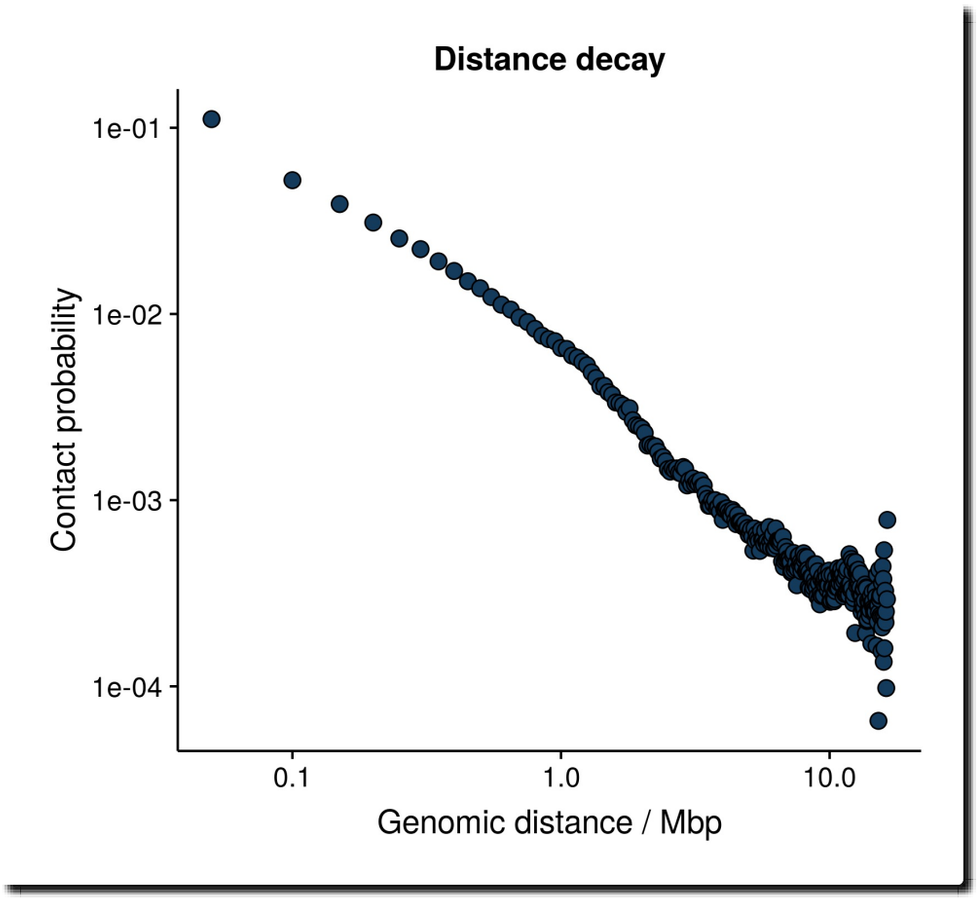
Signal + Bias + Decay + Noise



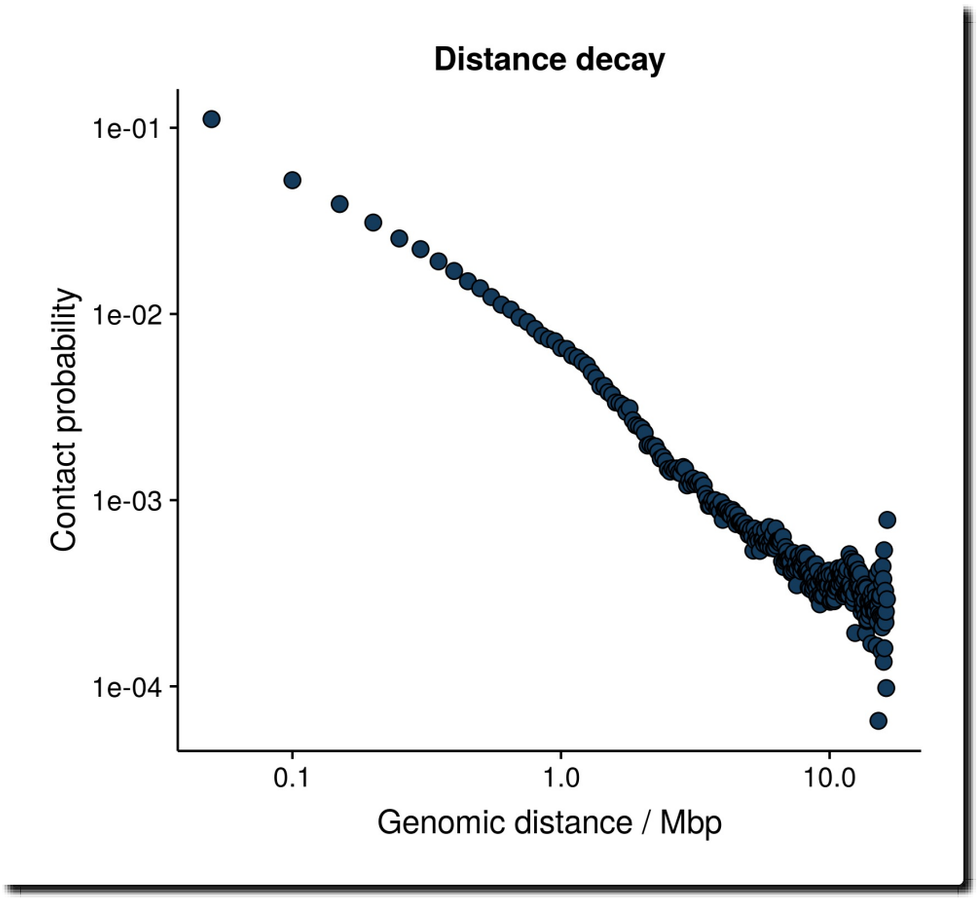
Challenge 2

Reproducibility

Decay drives Hi-C matrix comparison



Decay drives Hi-C matrix comparison



- Correct by expected decay **O / E**
- Compute by distance **SCC**

Yang, T. et al. (2017)

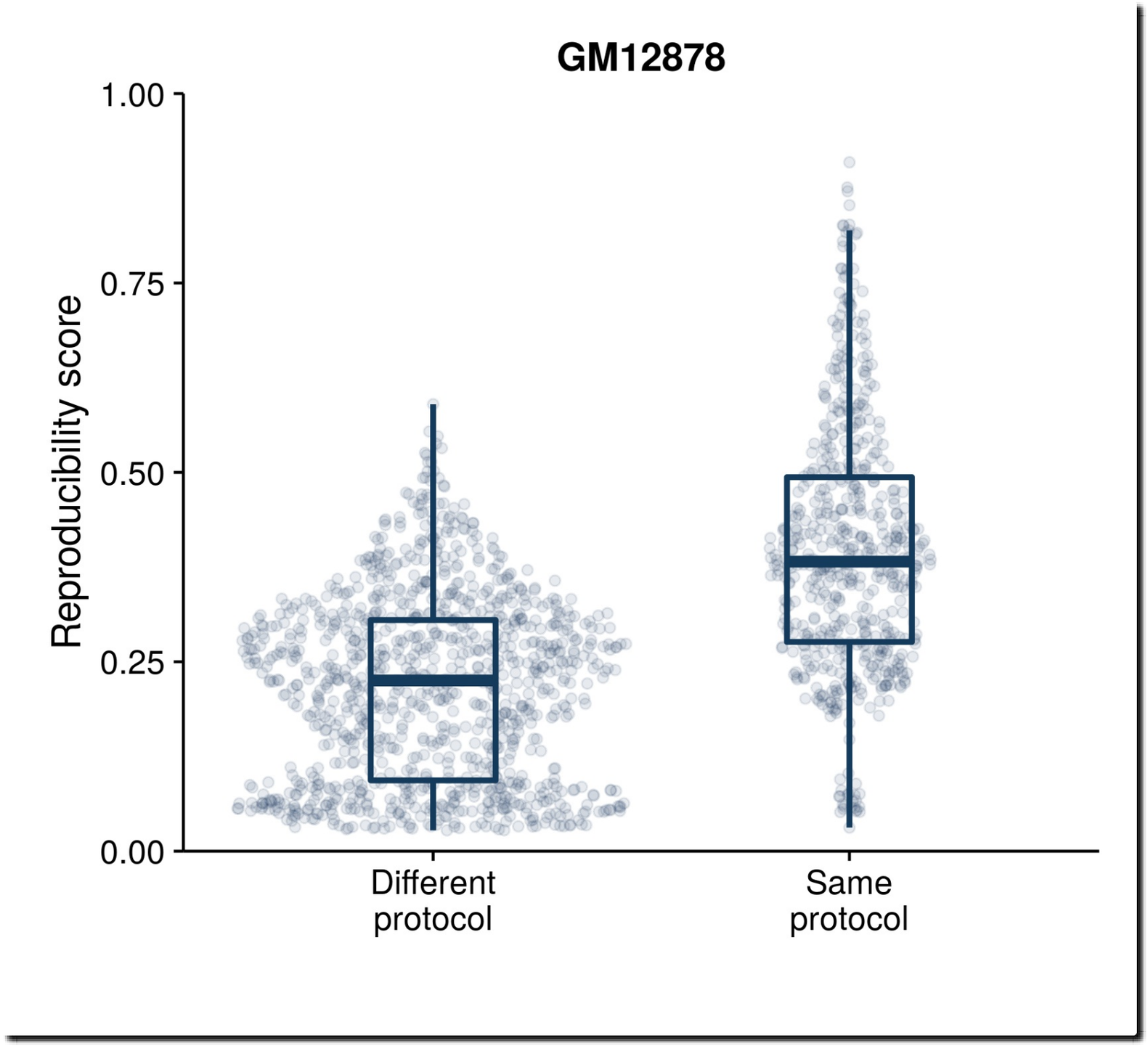
Genome research

- Spectral decomposition
  Reproducibility score

Yan, K. K. et al. (2017)

Bioinformatics

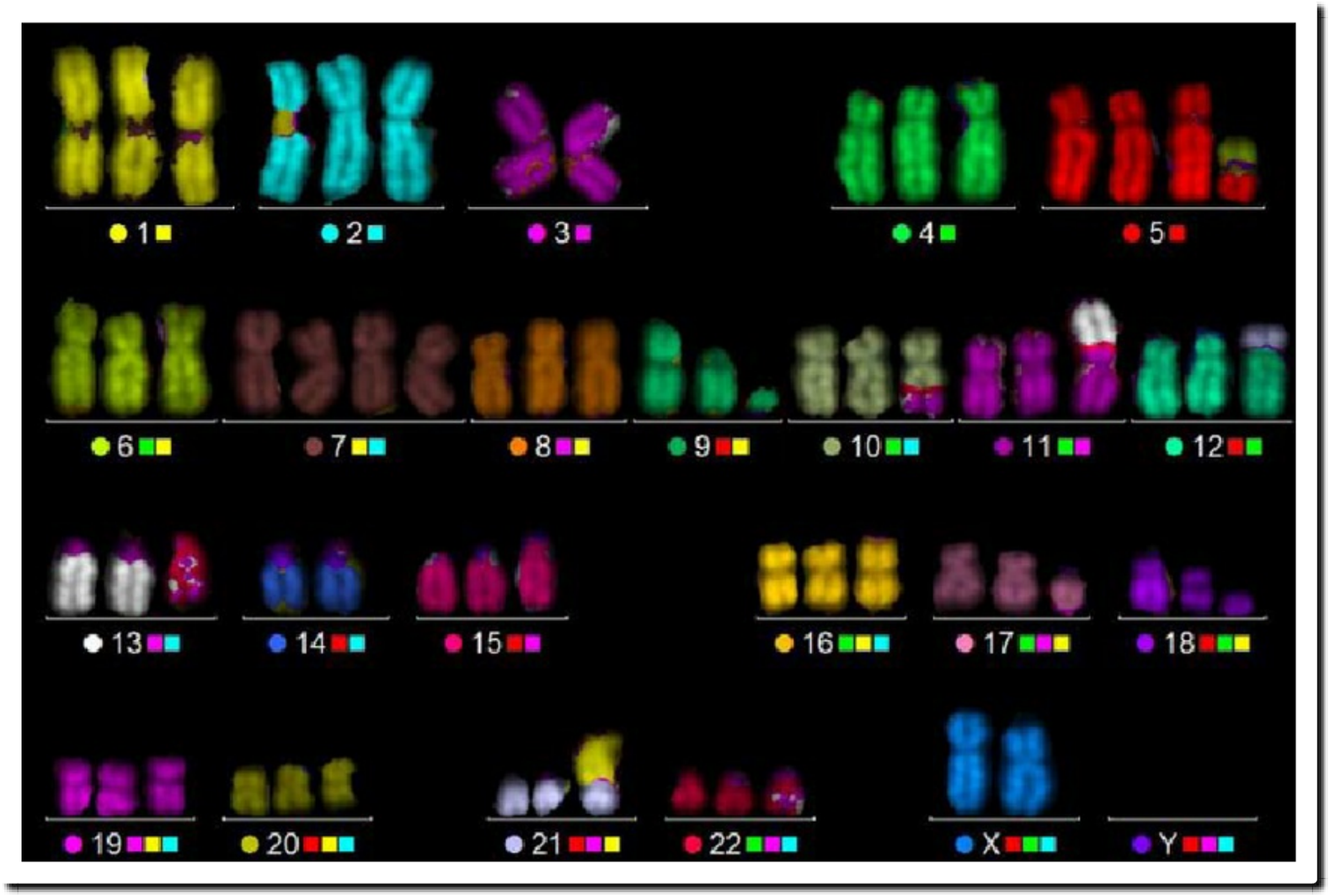
Reproducibility



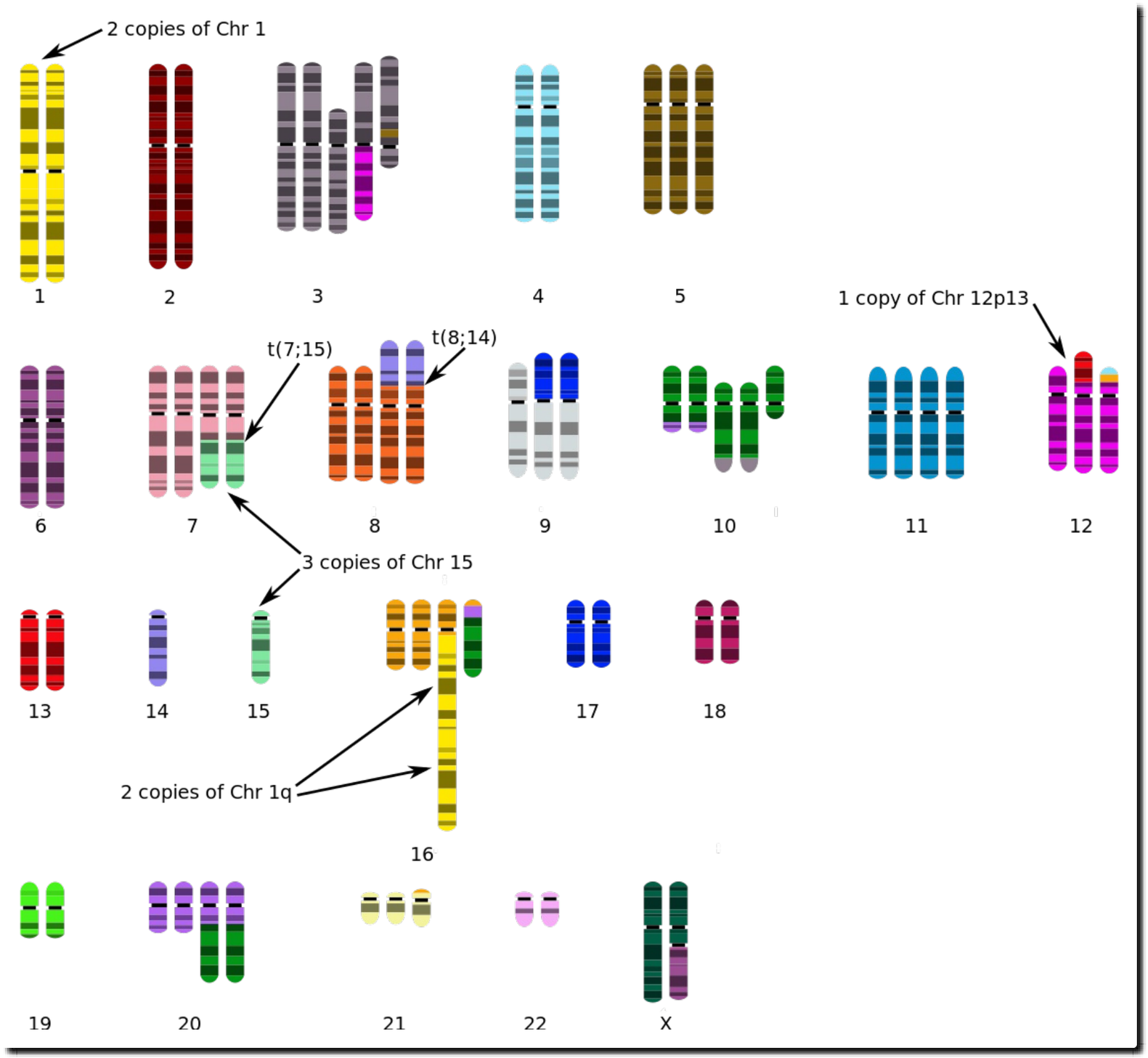
Challenge 3

Aberrant karyotypes

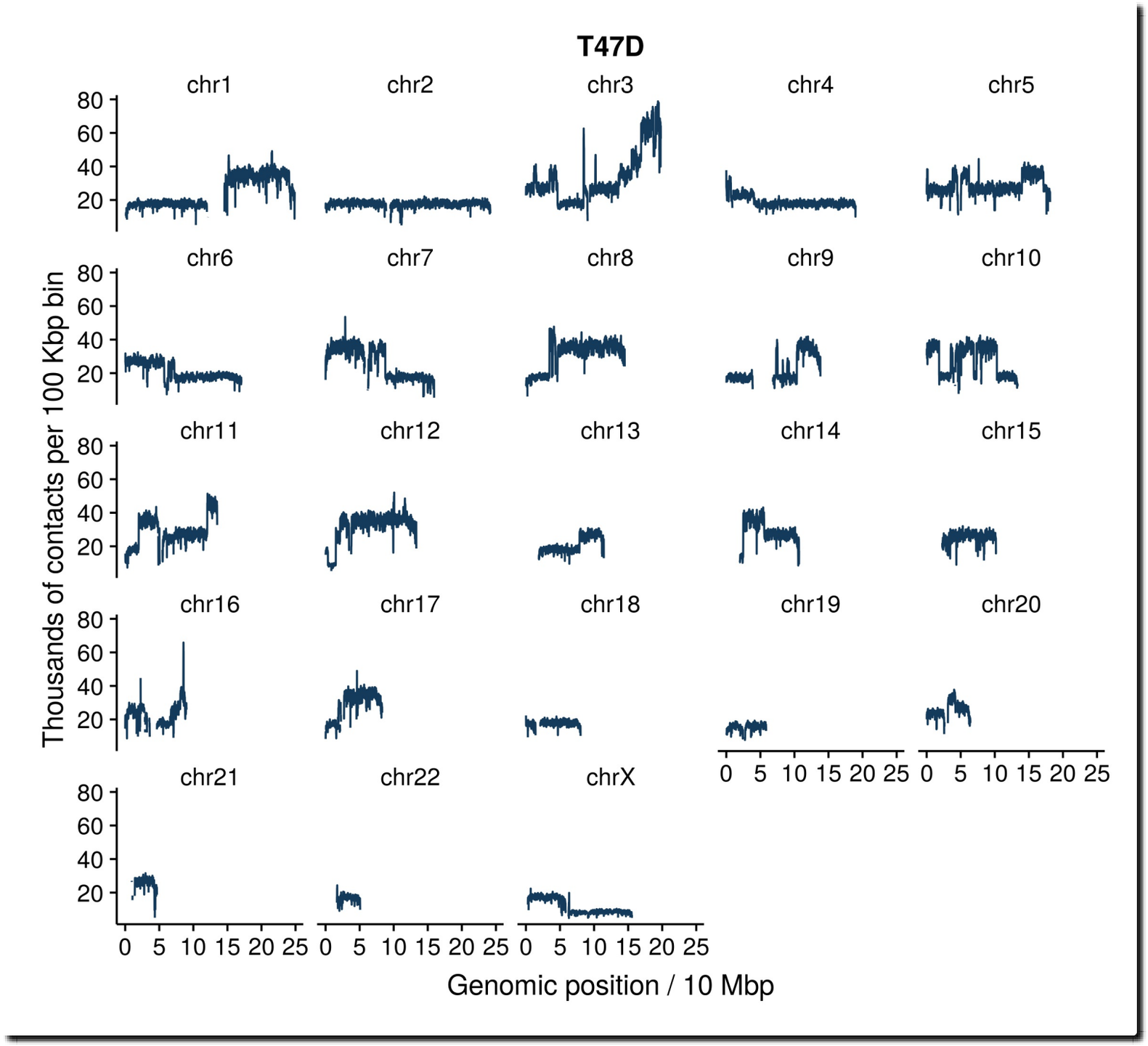
Cancer cell line: K562



Cancer cell line: T47D



Cancer cell line: T47D



Previous approaches

HiCNorm

Explicit model of biases

Regression model

All matrix entries

Hu, M. et al. (2012)

Bioinformatics

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

Nature genetics

HiCNorm

Explicit model of biases

Regression model

All matrix entries

Hu, M. et al. (2012)

Bioinformatics

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

Nature genetics

ICE

Implicit correction

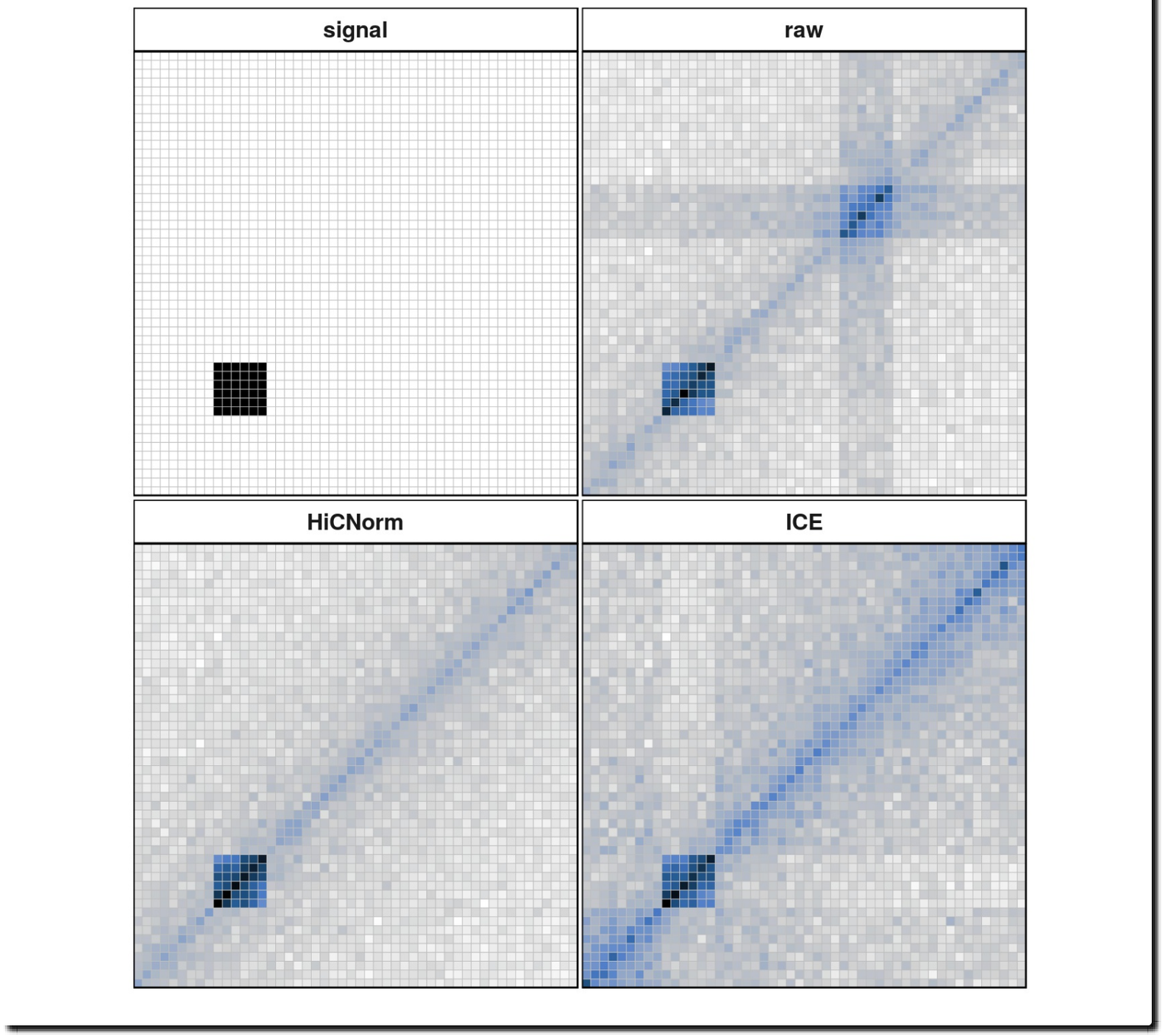
Matrix balancing

Equal visibility

Imakaev, M. et al. (2012)

Nature methods

In action



HiCNorm

✓ fewer artifacts

✗ very slow

✗ no high resolution

ICE

✗ more artifacts

✓ fast

✓ high resolution

HiCNorm

✓ fewer artifacts

✗ very slow

✗ no high resolution

ICE

✗ more artifacts

✓ fast

✓ 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

♡ Not worse on normal karyotypes

Wish list

♡ Suitable aberrant karyotypes

♡ As good as existing

♡ Not worse on normal karyotypes

♡ Fast

Wish list

♡ Suitable aberrant karyotypes

♡ As good as existing

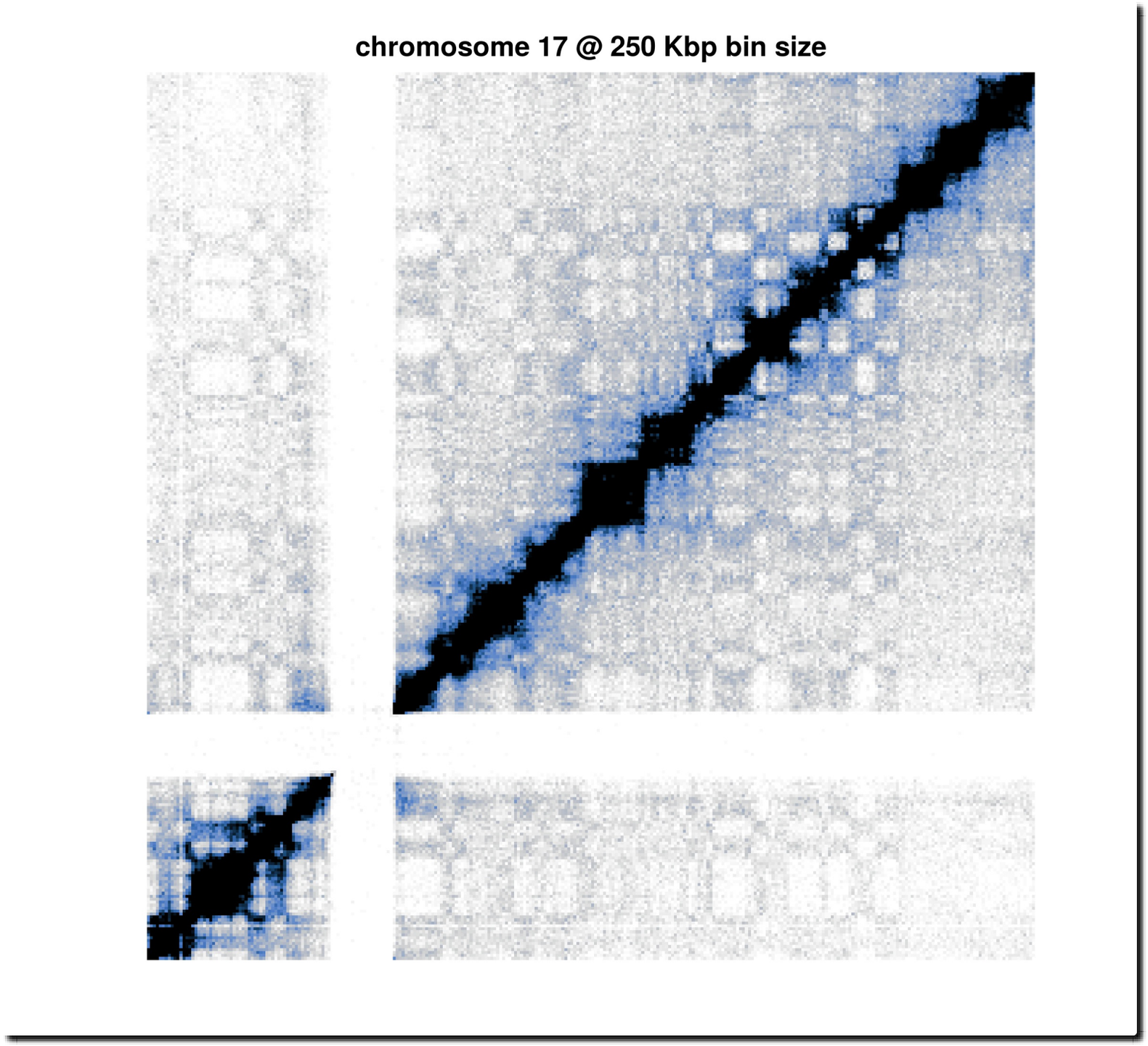
♡ Not worse on normal karyotypes

♡ Fast

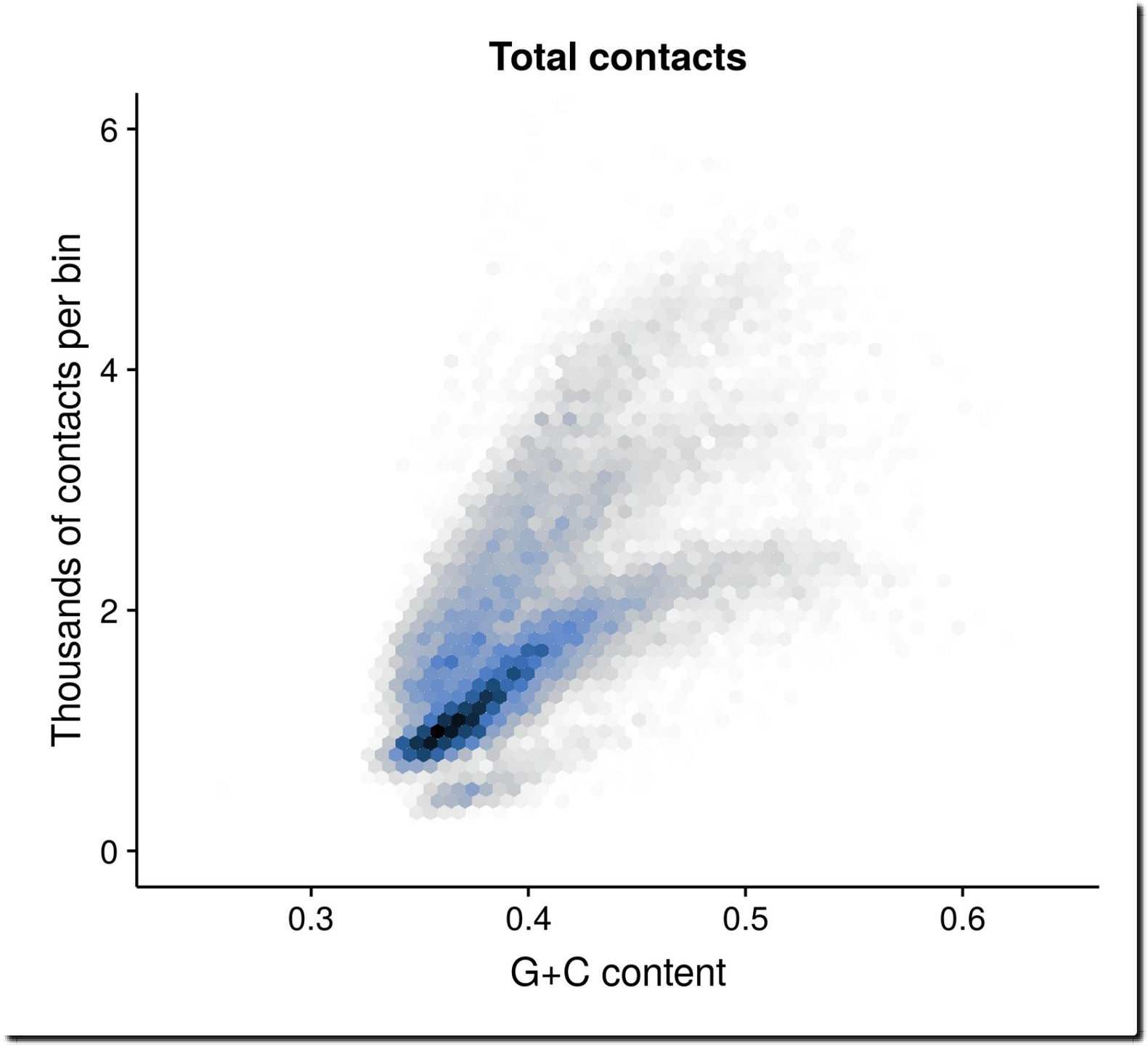
♡ Usable at high resolution

2D → 1D

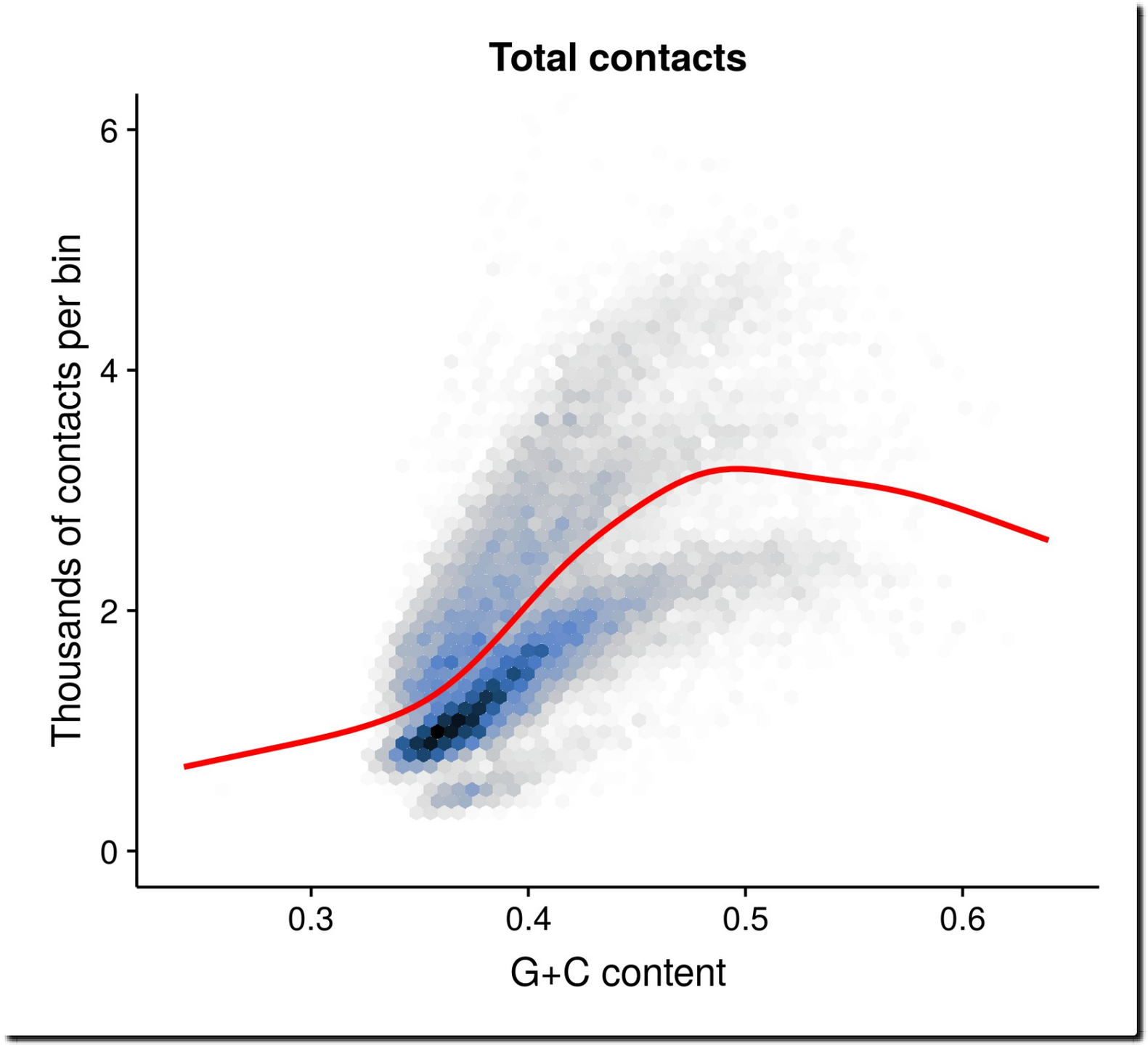
2D → 1D



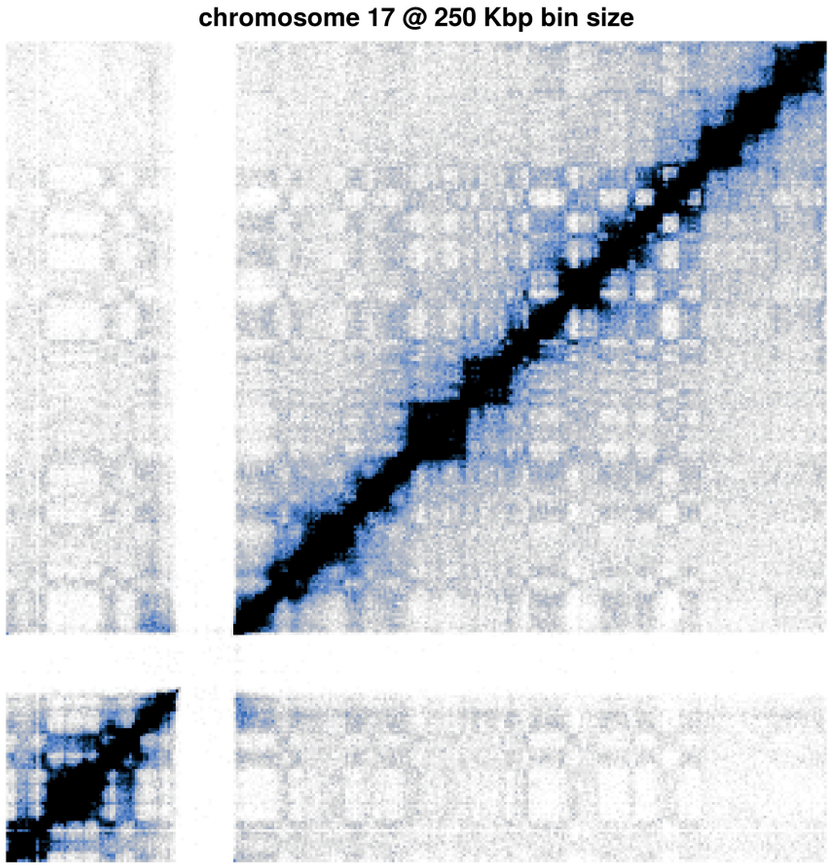
Totals vs. Genomic features



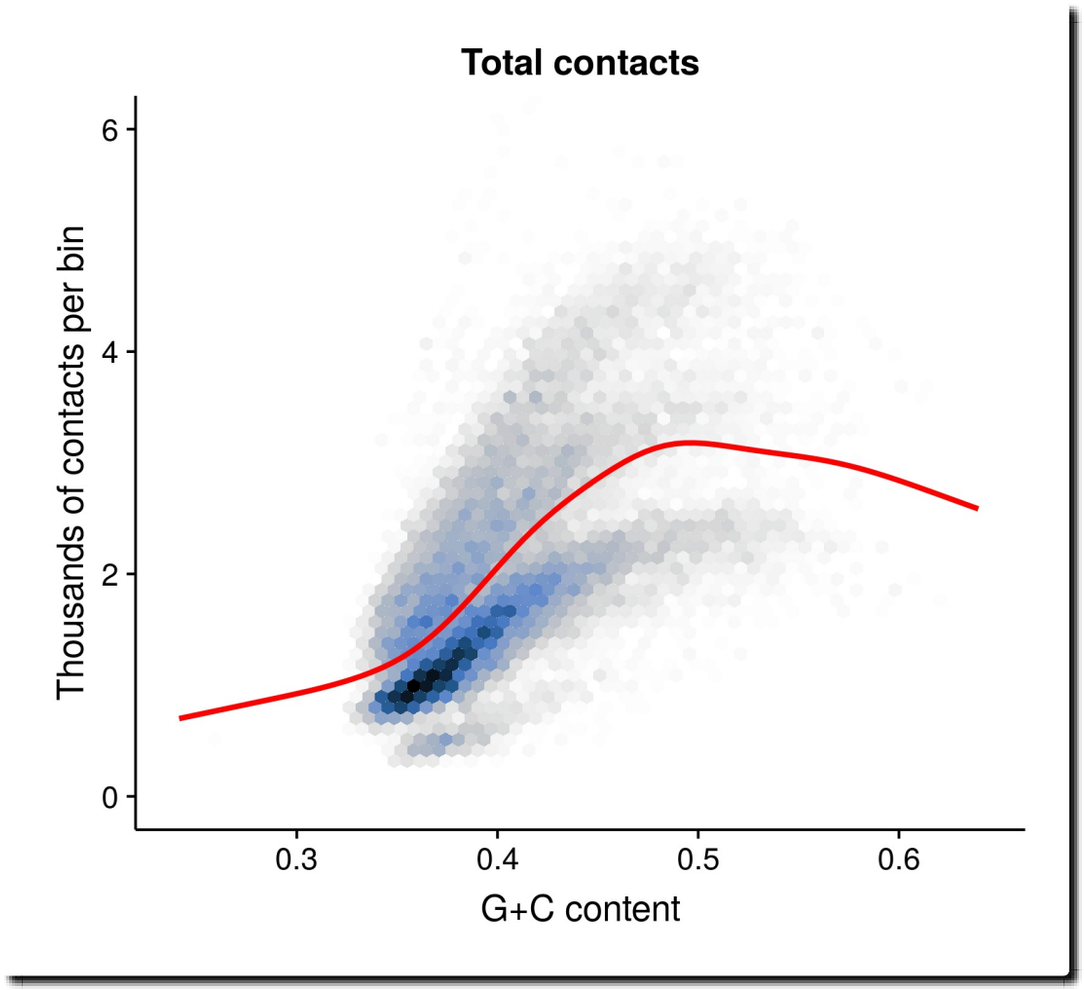
Non-linear relationship



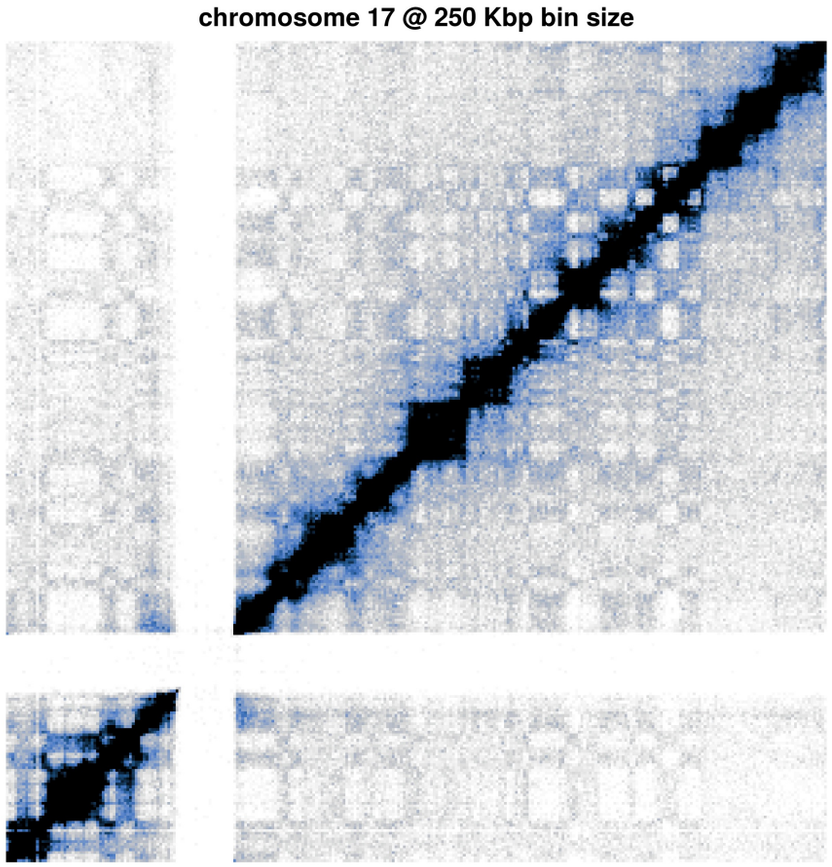
OneD model



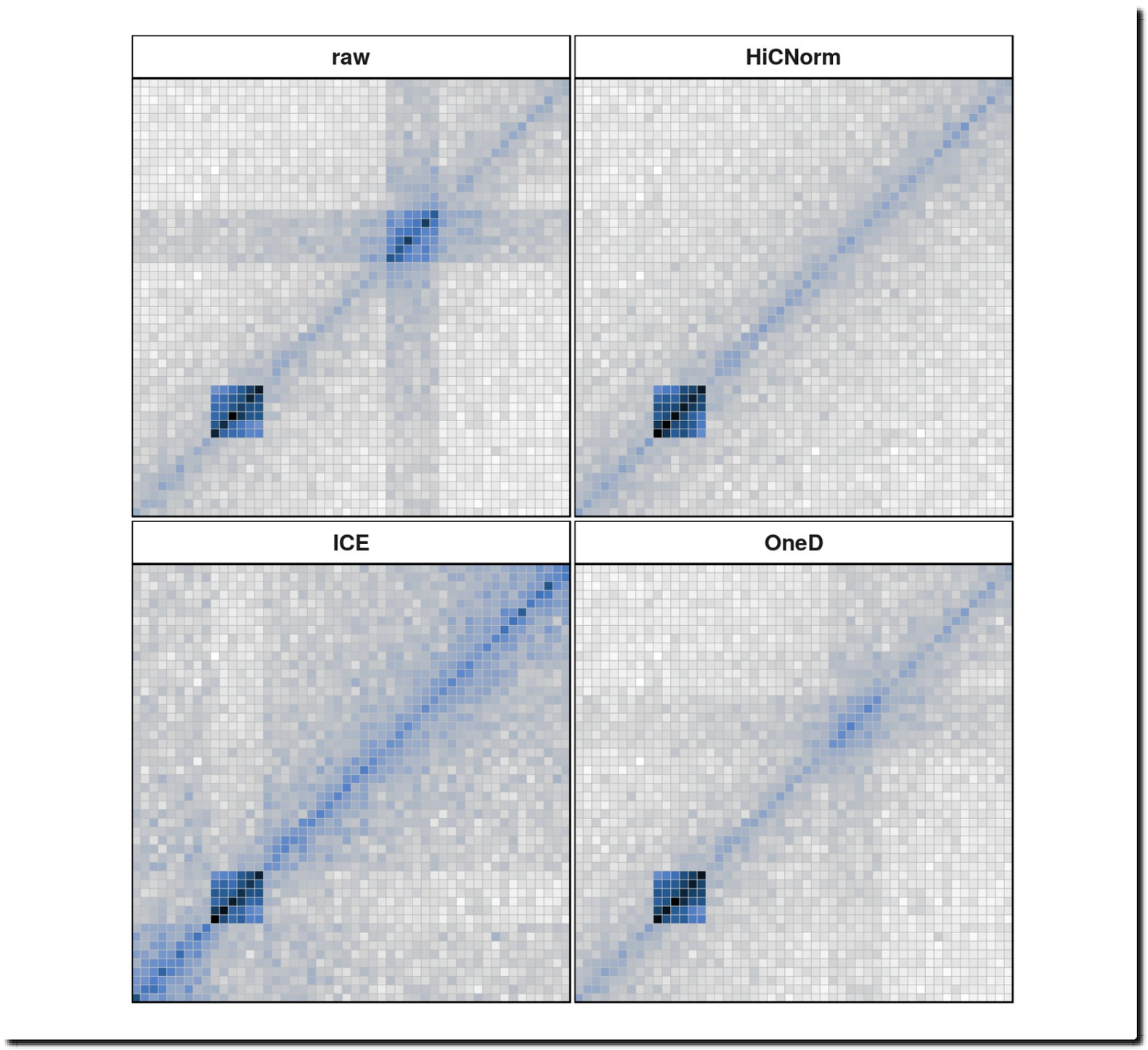
OneD model


Total contacts
Thousands of contacts per bin
6
4
2
0
0.3
0.4
0.5
0.6
G+C content

OneD model



In action



OneD performance

Benchmark strategy

Different methods

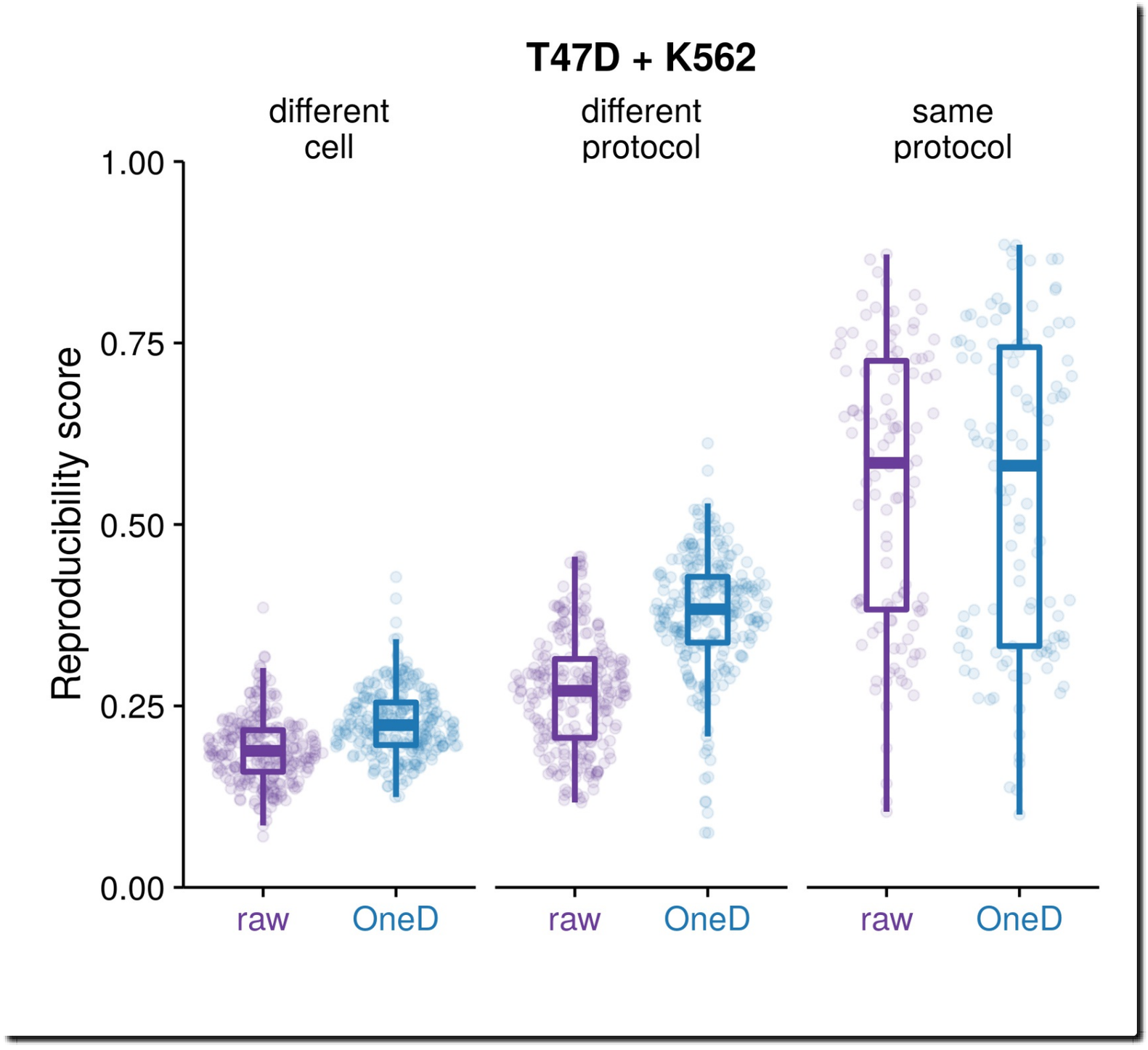
Including doing nothing (raw)

Different protocols

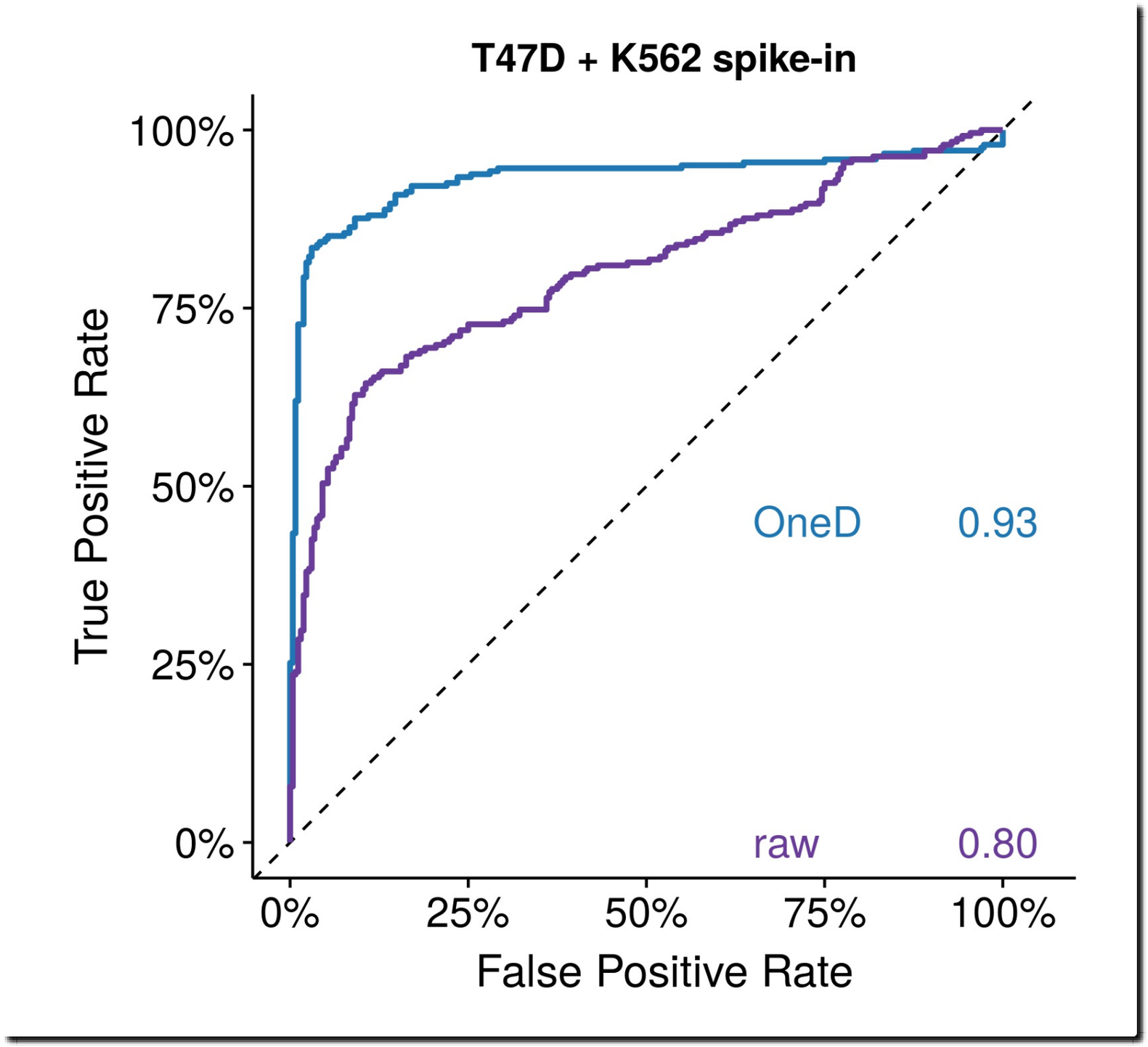
Restriction enzyme, in-situ / diluted, lab, etc ...

Different cell types

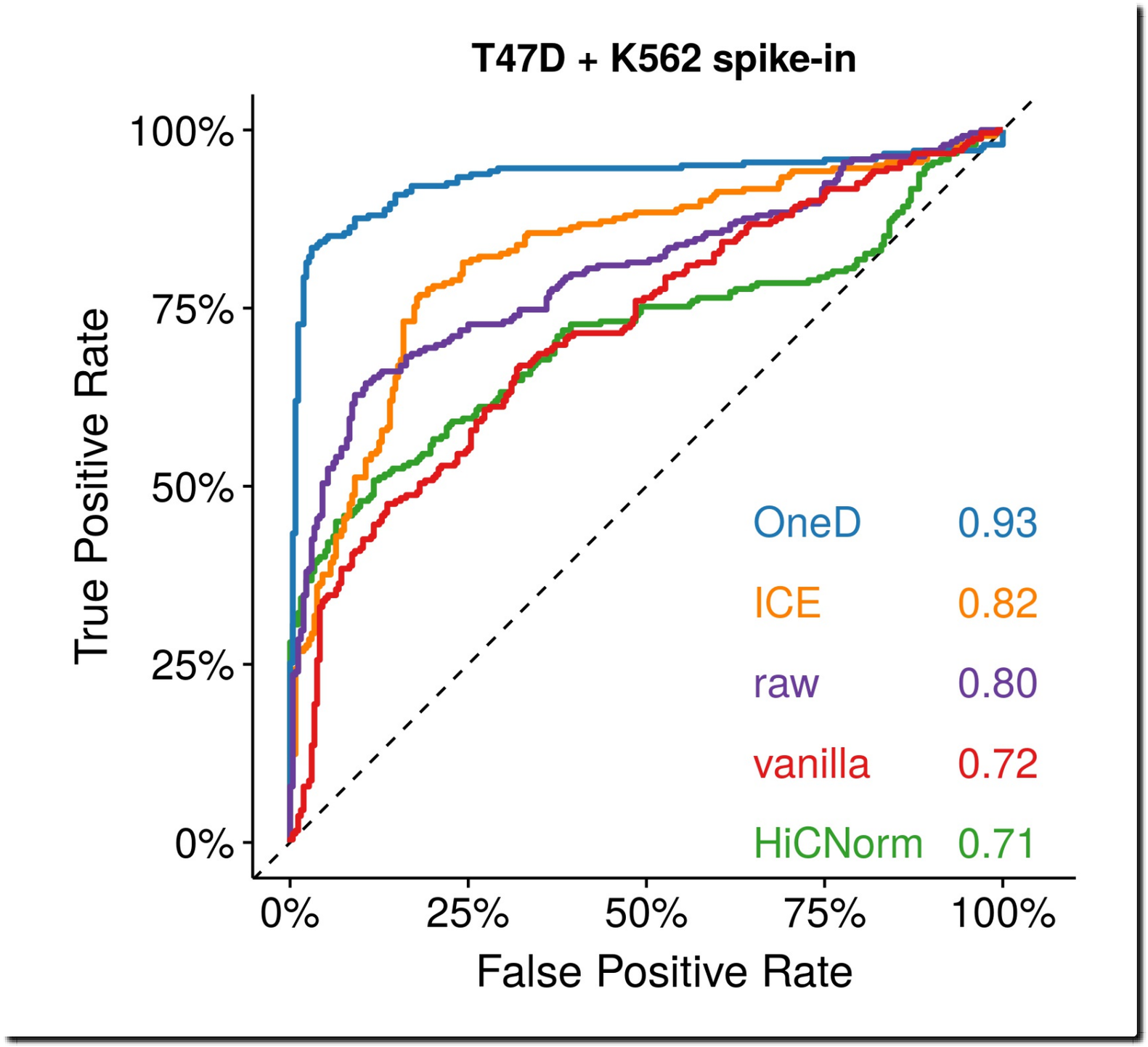
Pair-wise comparison



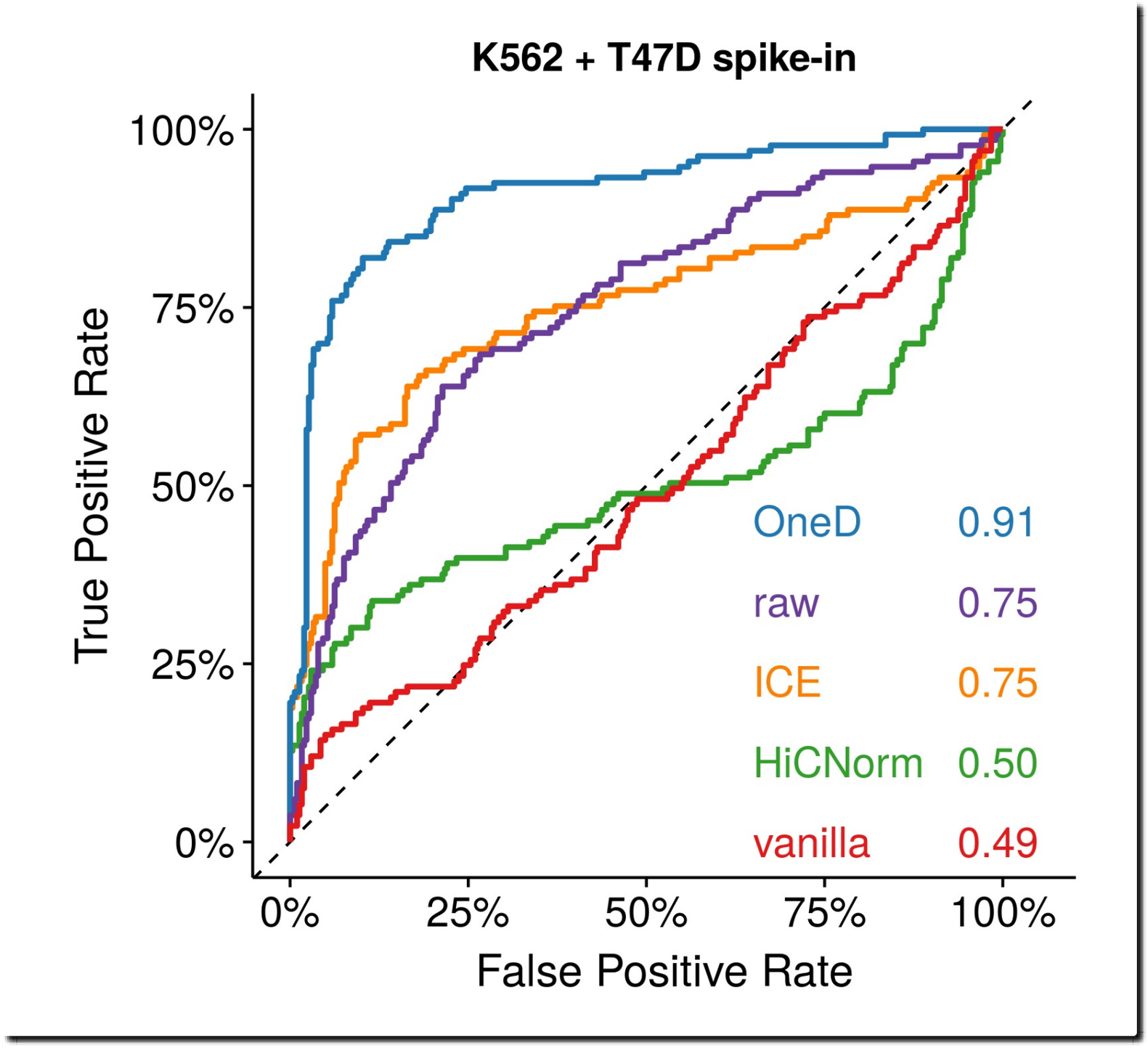
ROC and AUC


T47D + K562 spike-in
100%
75%
50%
25%
0%
True Positive Rate
0%
25%
50%
75%
100%
False Positive Rate
OneD 0.93
raw 0.80

ROC and AUC


T47D + K562 spike-in
True Positive Rate
100%
75%
50%
25%
0%
0%
25%
50%
75%
100%
False Positive Rate
OneD 0.93
ICE 0.82
raw 0.80
vanilla 0.72
HiCNorm 0.71

ROC and AUC


K562 + T47D spike-in
100%
75%
50%
25%
0%
True Positive Rate
0%
25%
50%
75%
100%
False Positive Rate
OneD 0.91
raw 0.75
ICE 0.75
HiCNorm 0.50
vanilla 0.49

Wish list

♡ Suitable aberrant karyotypes ✓

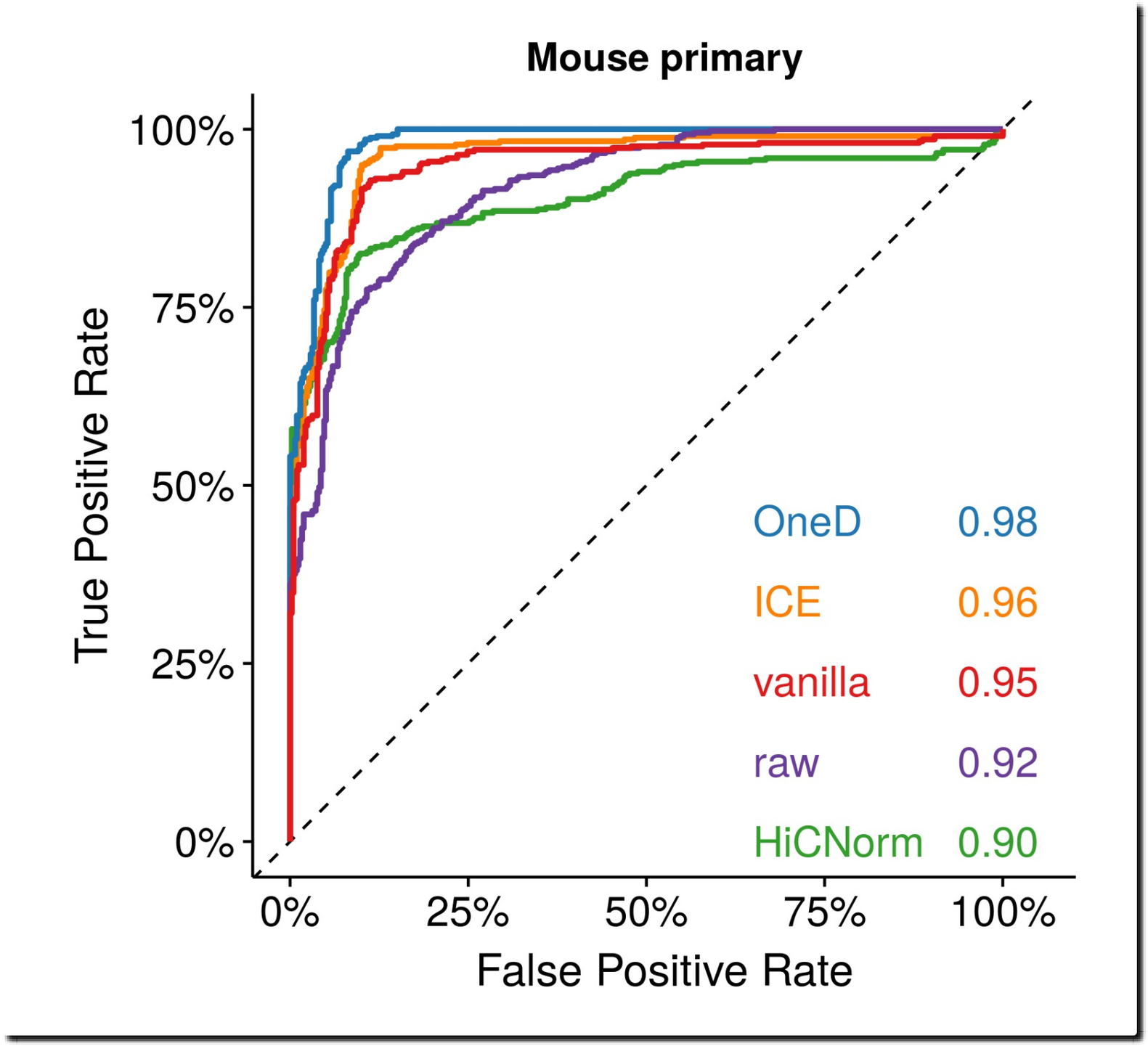
♡ ~~As good as~~ Better than existing ✓

♡ Not worse on normal karyotypes

♡ Fast

♡ Usable at high resolution

ROC and AUC (diploid)



Wish list

♡ Suitable aberrant karyotypes ✓

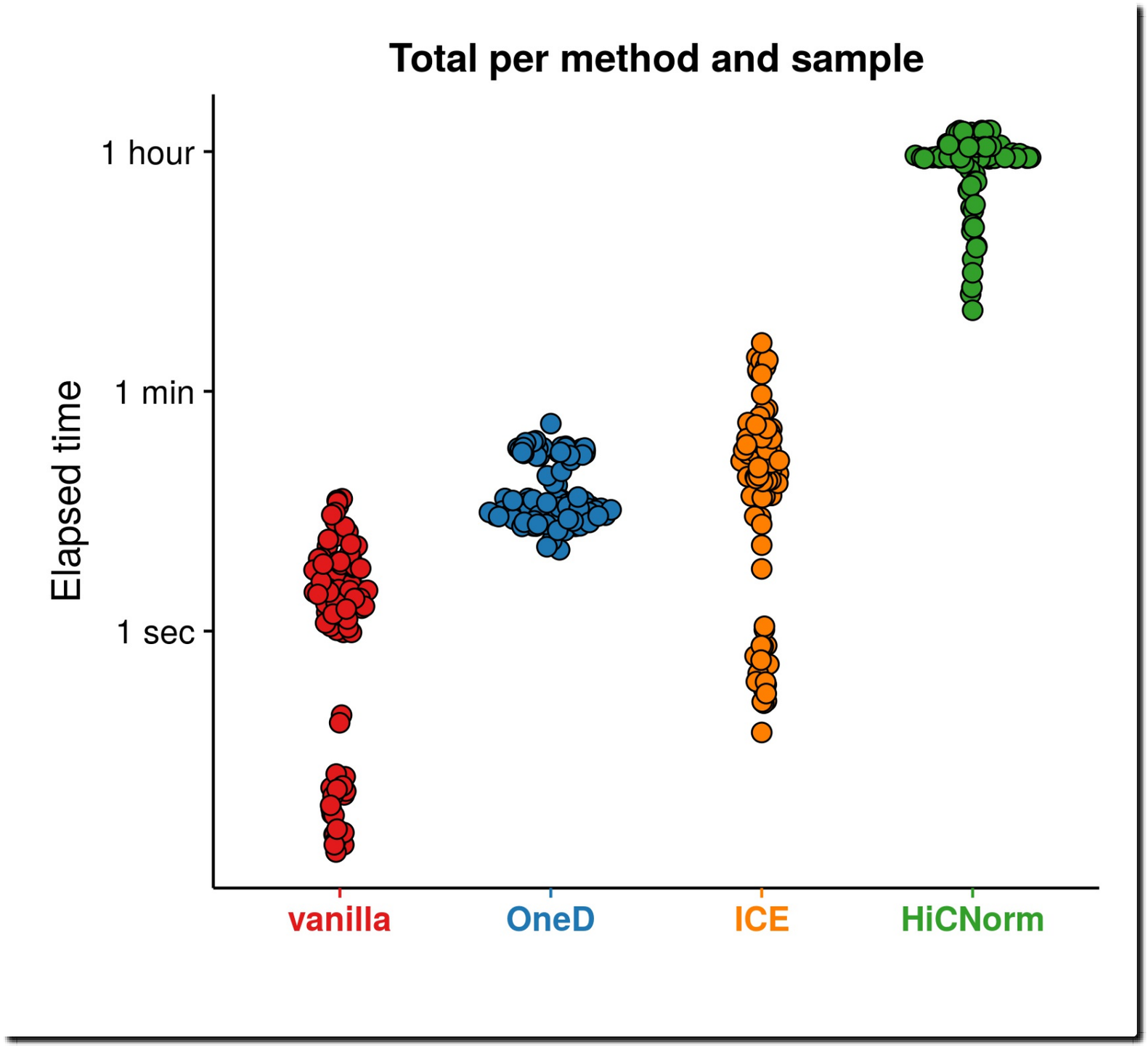
♡ Better than existing ✓

♡ Not worse on normal karyotypes ✓

♡ Fast

♡ Usable at high resolution

Speed


Total per method and sample
1 hour
1 min
1 sec
Elapsed time
vanilla
OneD
ICE
HiCNorm

Wish list

♡ Suitable aberrant karyotypes ✓

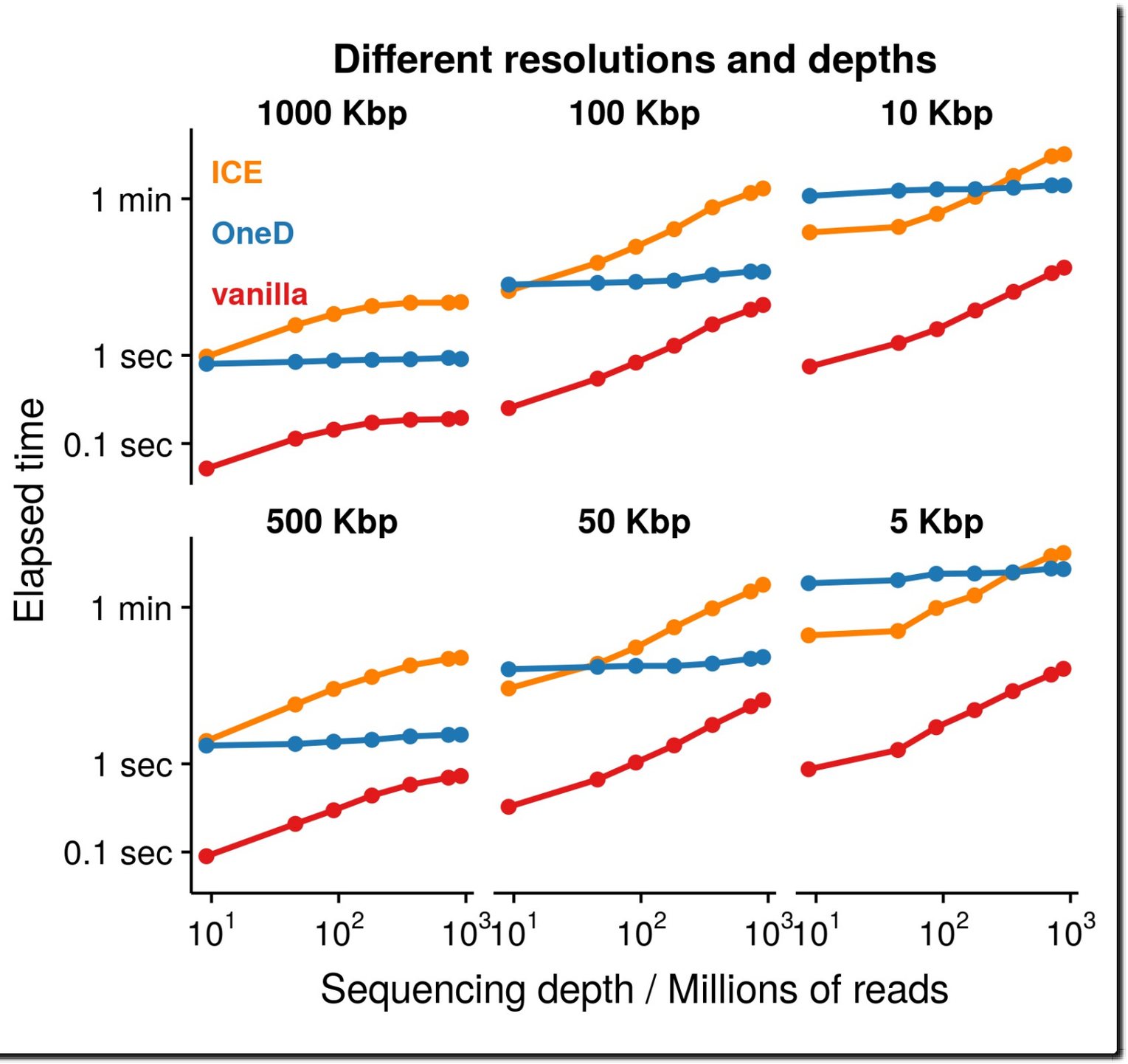
♡ Better than existing ✓

♡ Not worse on normal karyotypes ✓

♡ Fast ✓

♡ Usable at high resolution

Resolution


Different resolutions and depths
1000 Kbp
100 Kbp
10 Kbp
500 Kbp
50 Kbp
5 Kbp
ICE
OneD
vanilla
1 min
1 sec
0.1 sec
Elapsed time
10^1
10^2
10^3
Sequencing depth / Millions of reads

Wish list

♡ Suitable aberrant karyotypes ✓

♡ Better than existing ✓

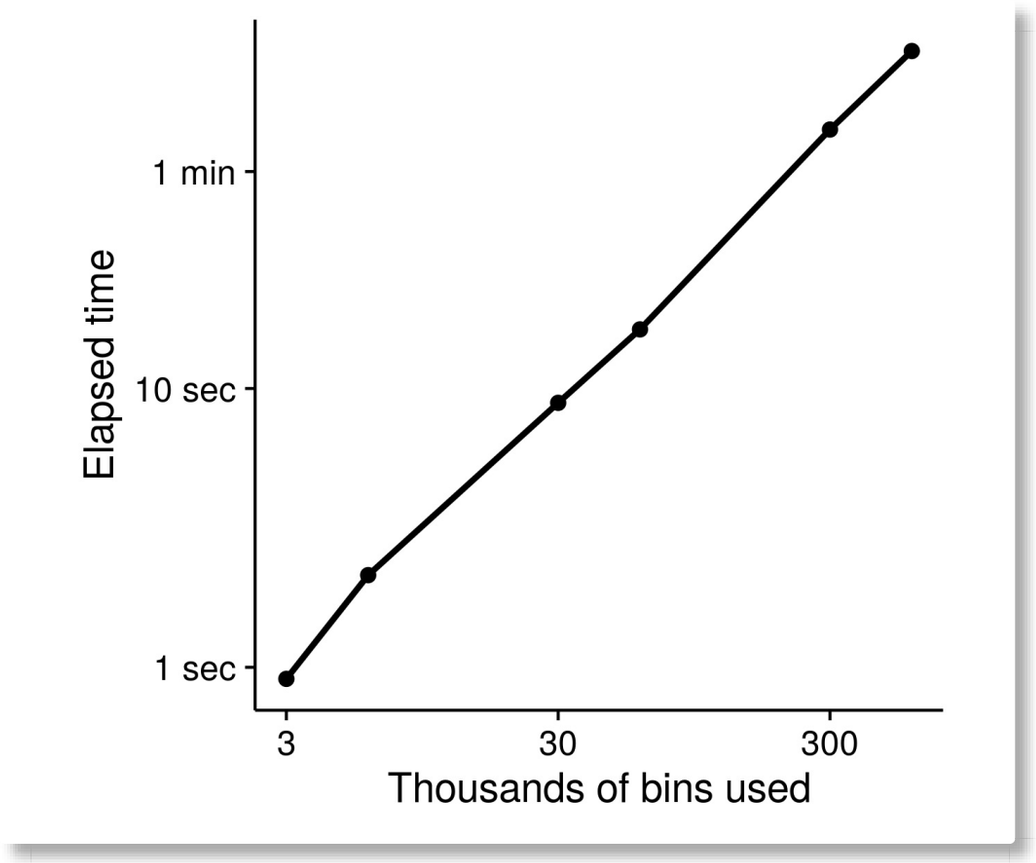
♡ Not worse on normal karyotypes ✓

♡ Fast ✓

♡ Usable at high resolution ✓

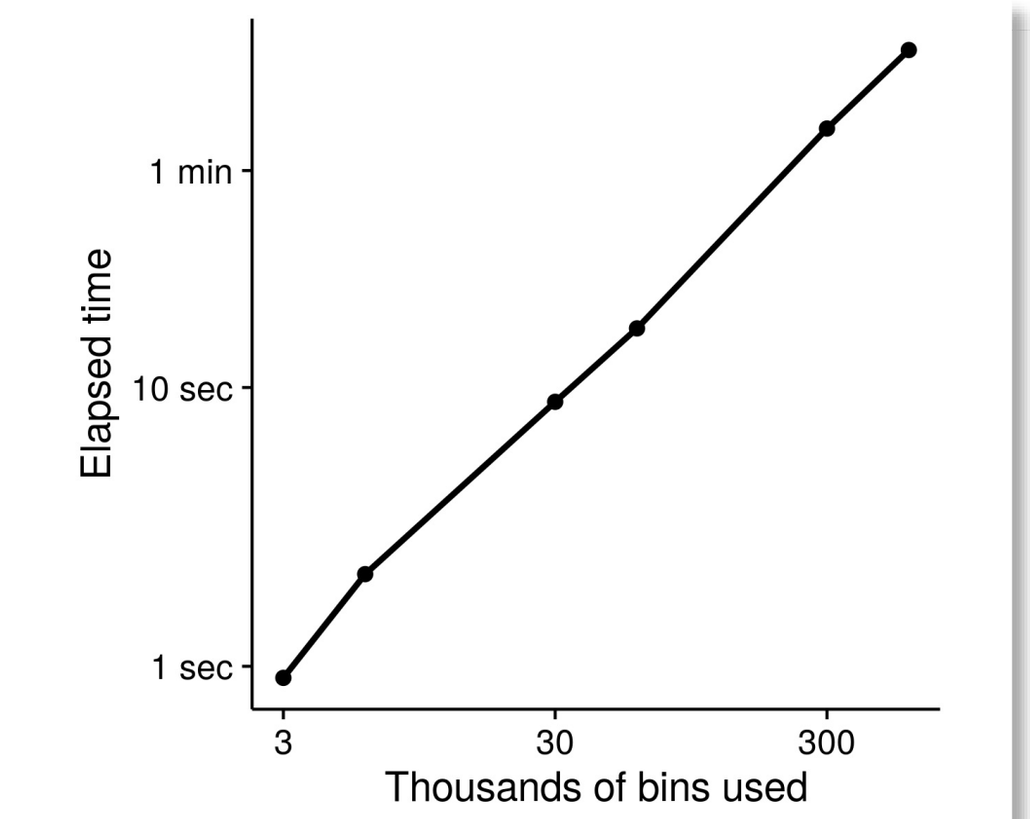
Resolution (with some tricks)

Resolution (with some tricks)

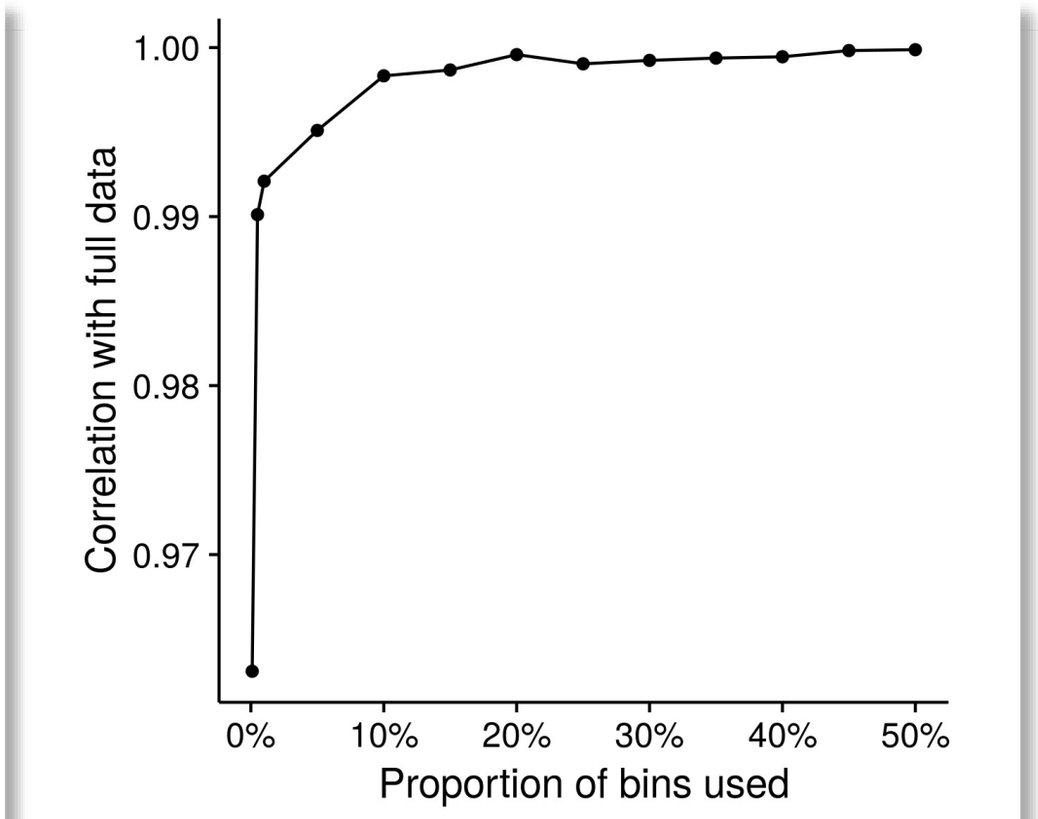


Less data, less time

Resolution (with some tricks)



Less data, less time



Less data, same performance

Summary

Summary

Hi-C can interrogate genome structure

Hi-C experiments have biases (including copy number)

Summary

Hi-C can interrogate genome structure

Hi-C experiments have biases (including copy number)

OneD removes biases increasing reproducibility

OneD is fast and suitable at high resolutions

OneD availability

R package

https://github.com/qenvio/dryhic

Paper

Vidal, E. et al. (2018)

Nucleic acids research

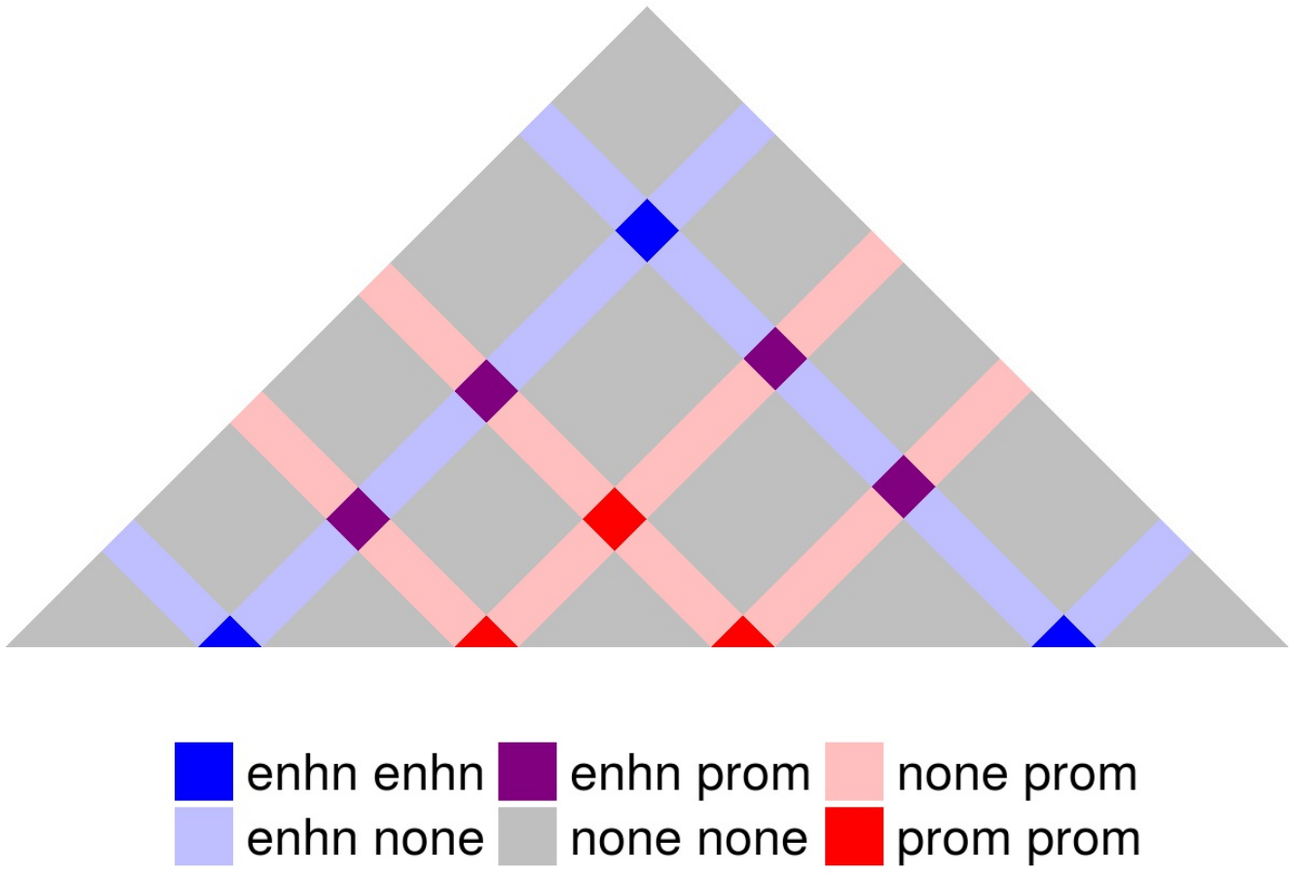
https://doi.org/10.1093/nar/gky064

The inner life of TADs

Genome-wide enrichment of contacts between regions

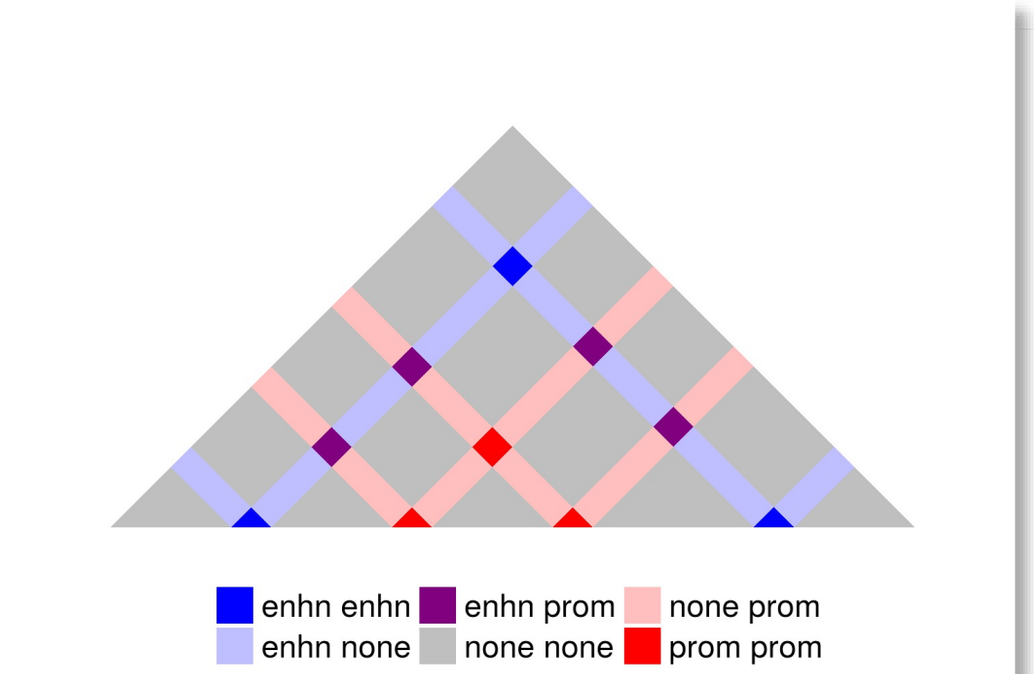
(i.e. promoters and enhancers)

Types of bins inside a TAD

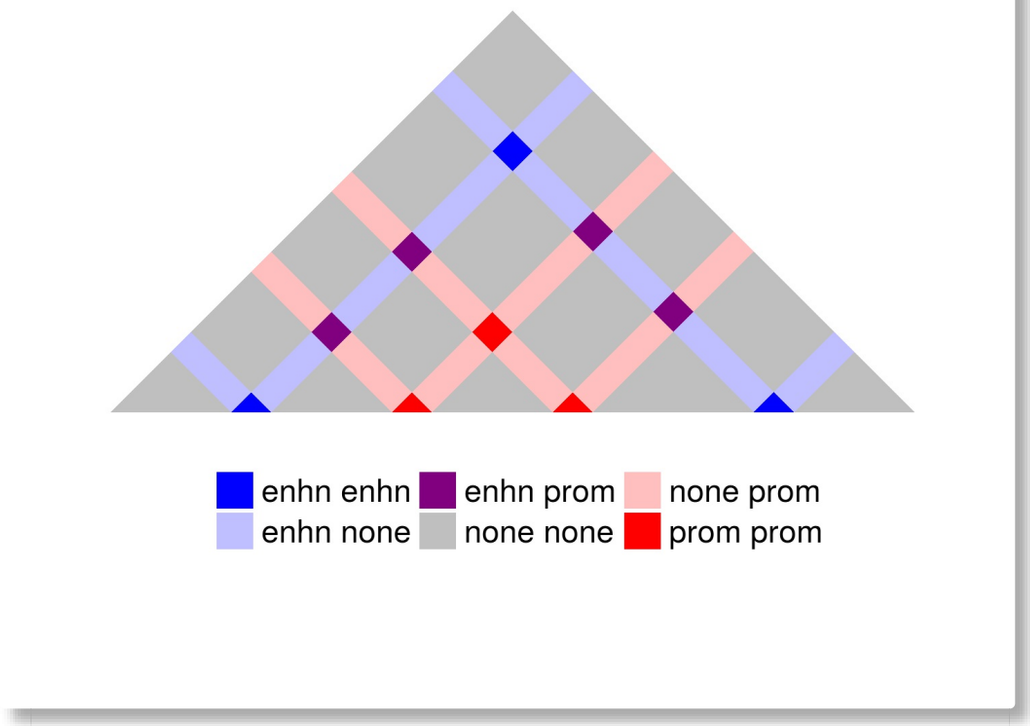


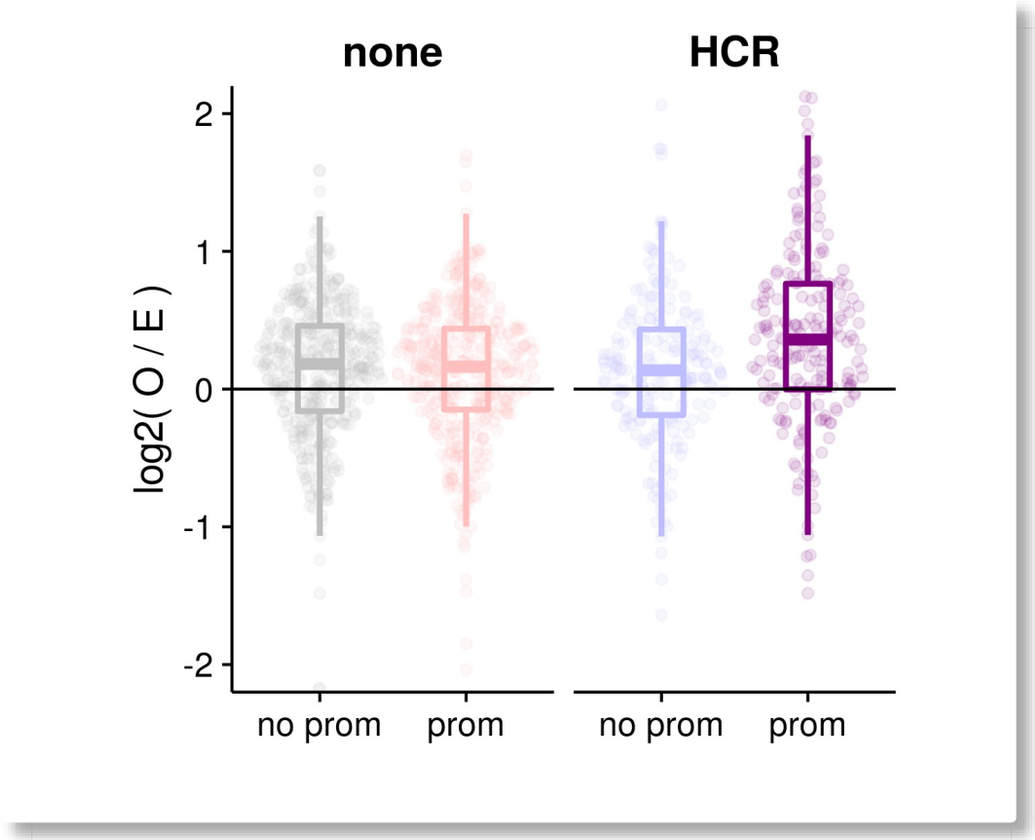
Example: HCR and H3K4me3

Example: HCR and H3K4me3



Example: HCR and H3K4me3

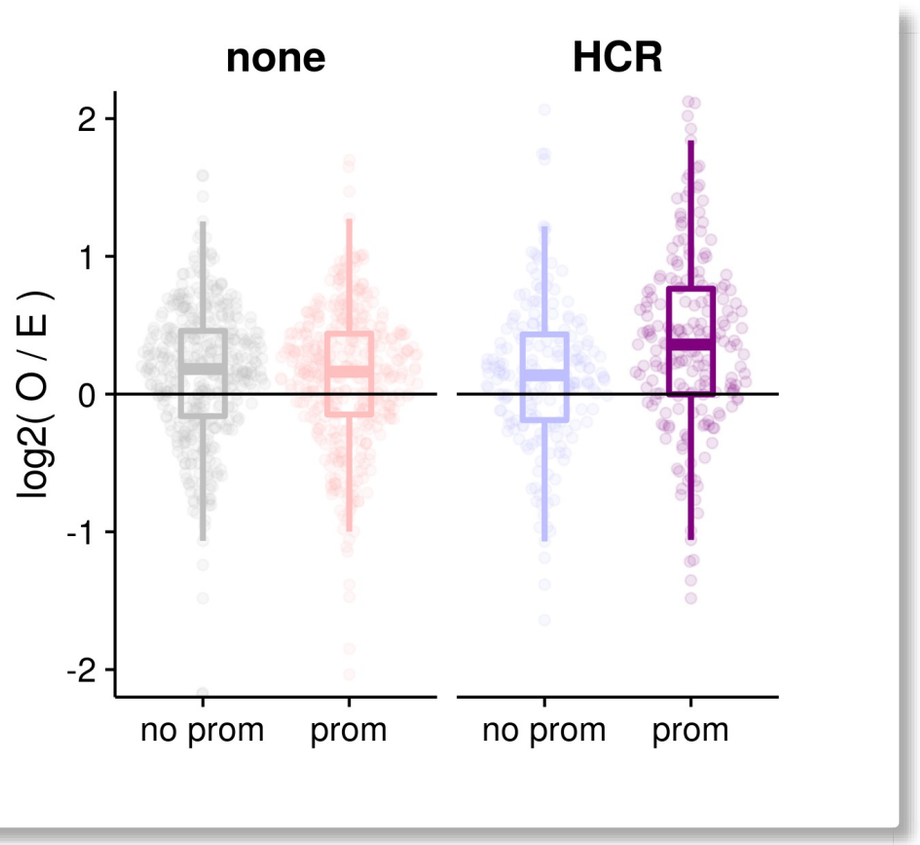




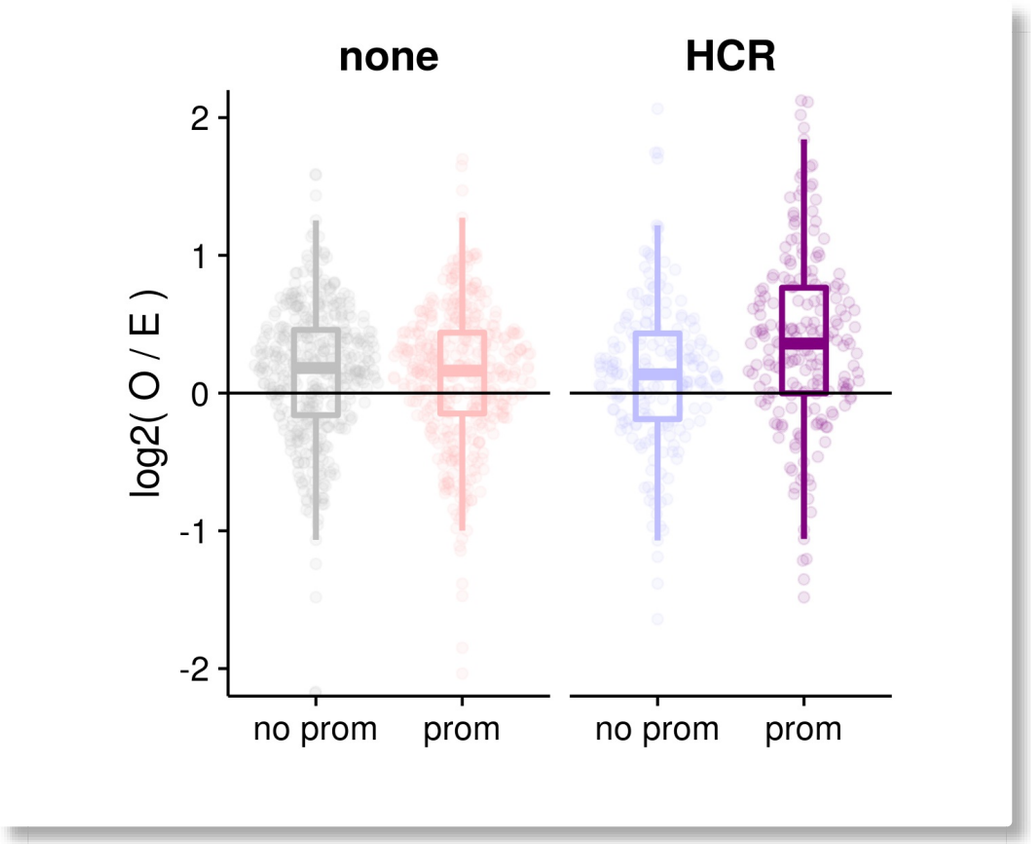
Le Dily, F. and Vidal, E. et al. (2017)

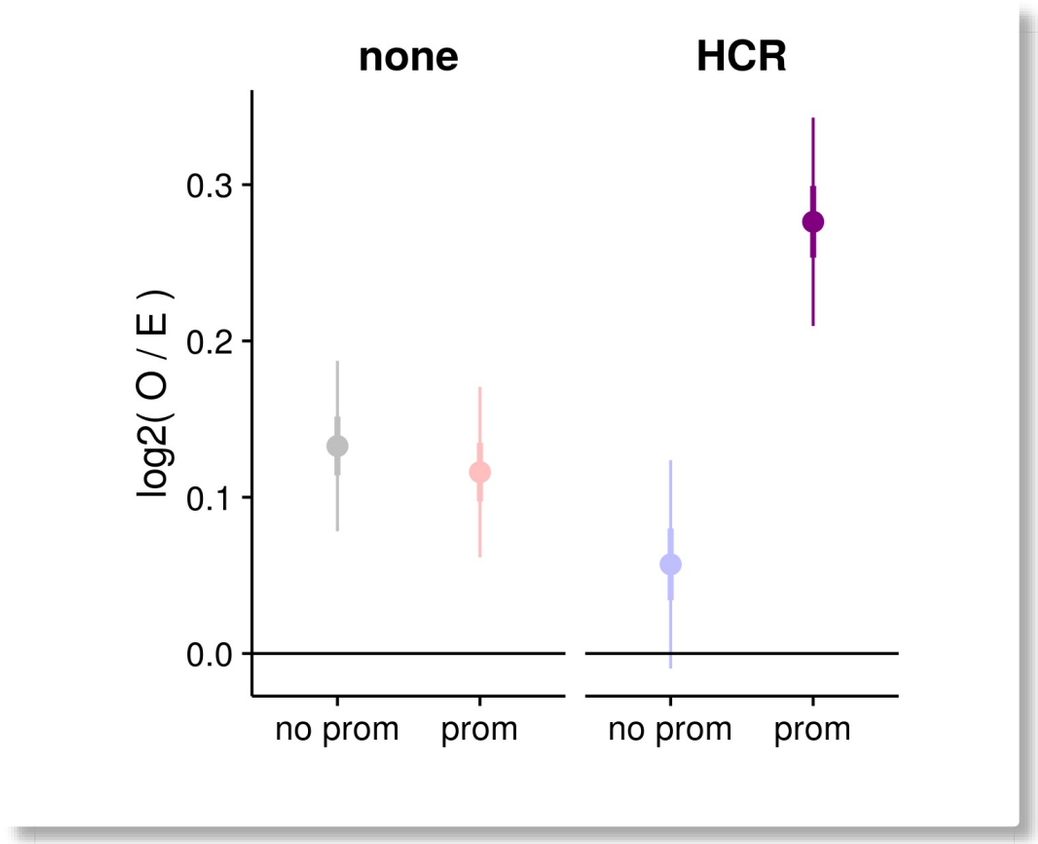
bioRxiv
https://doi.org/10.1101/233874

Summarizing via linear mixed model

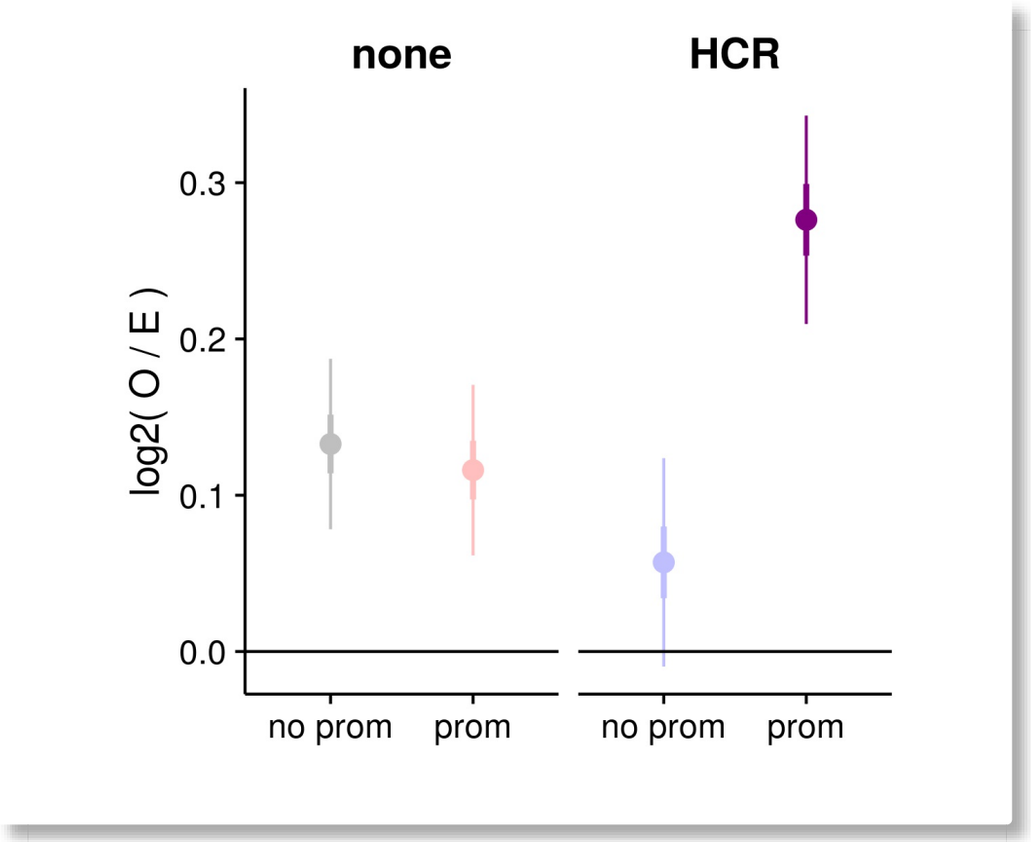


Summarizing via linear mixed model

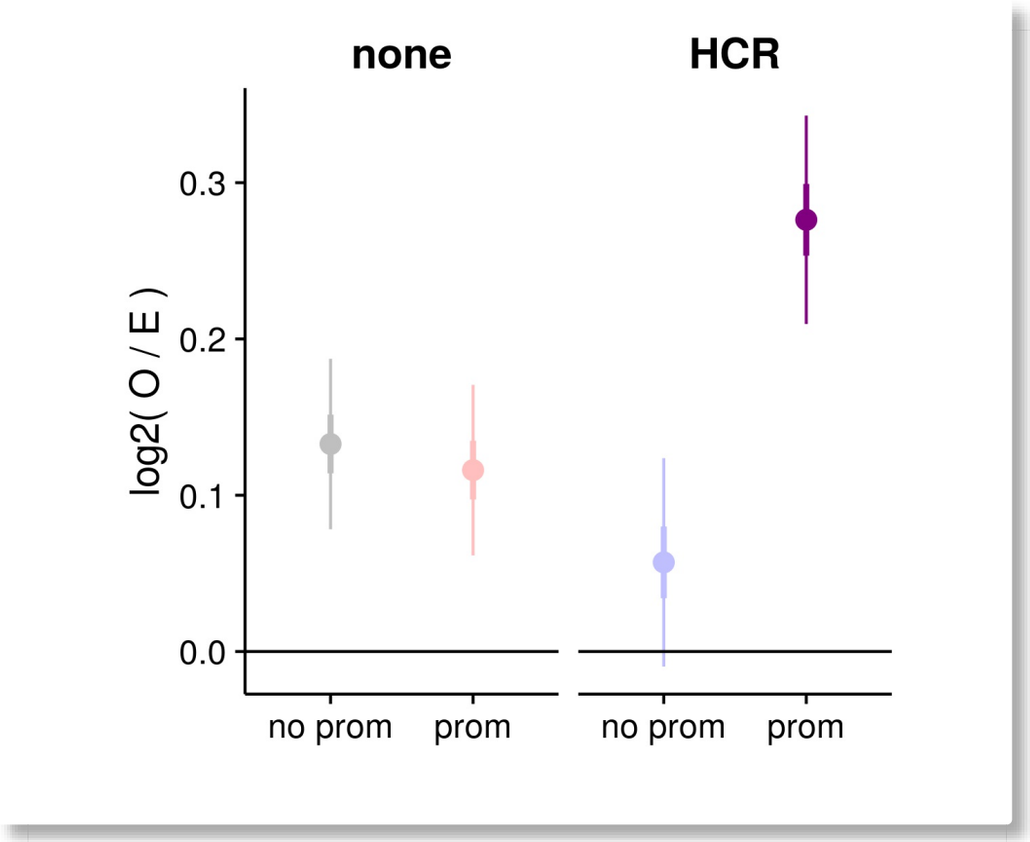


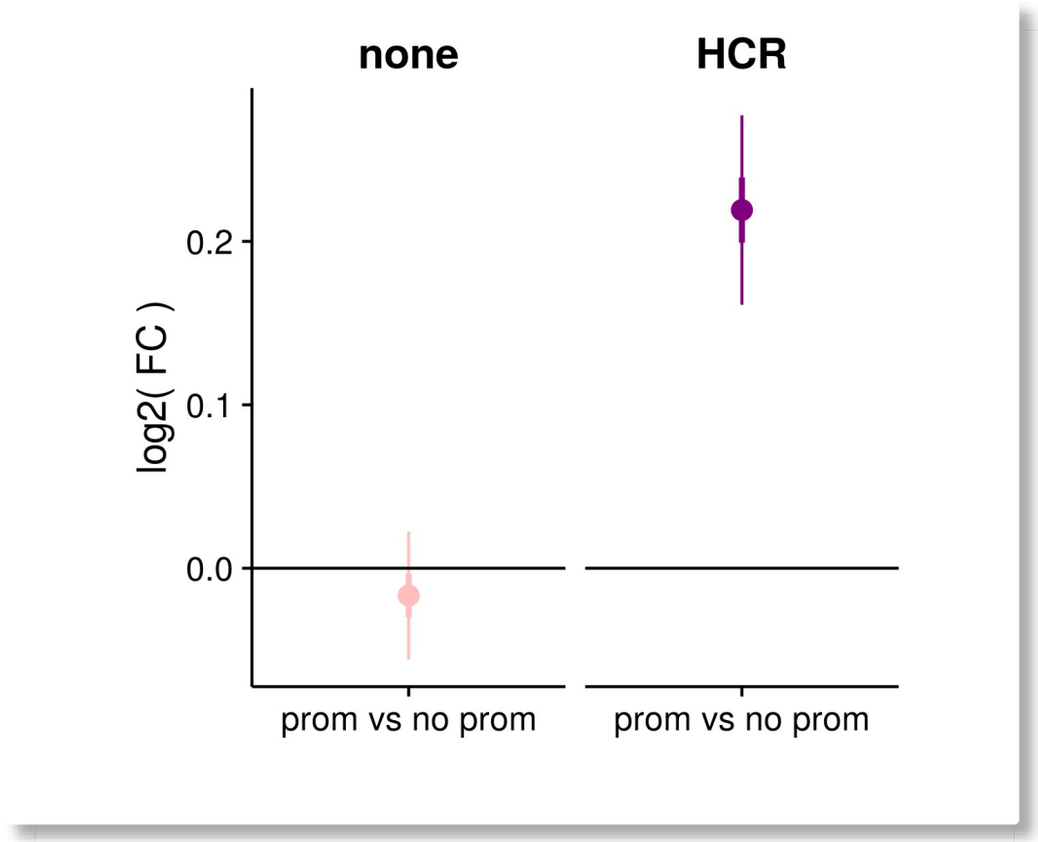


Simple comparison

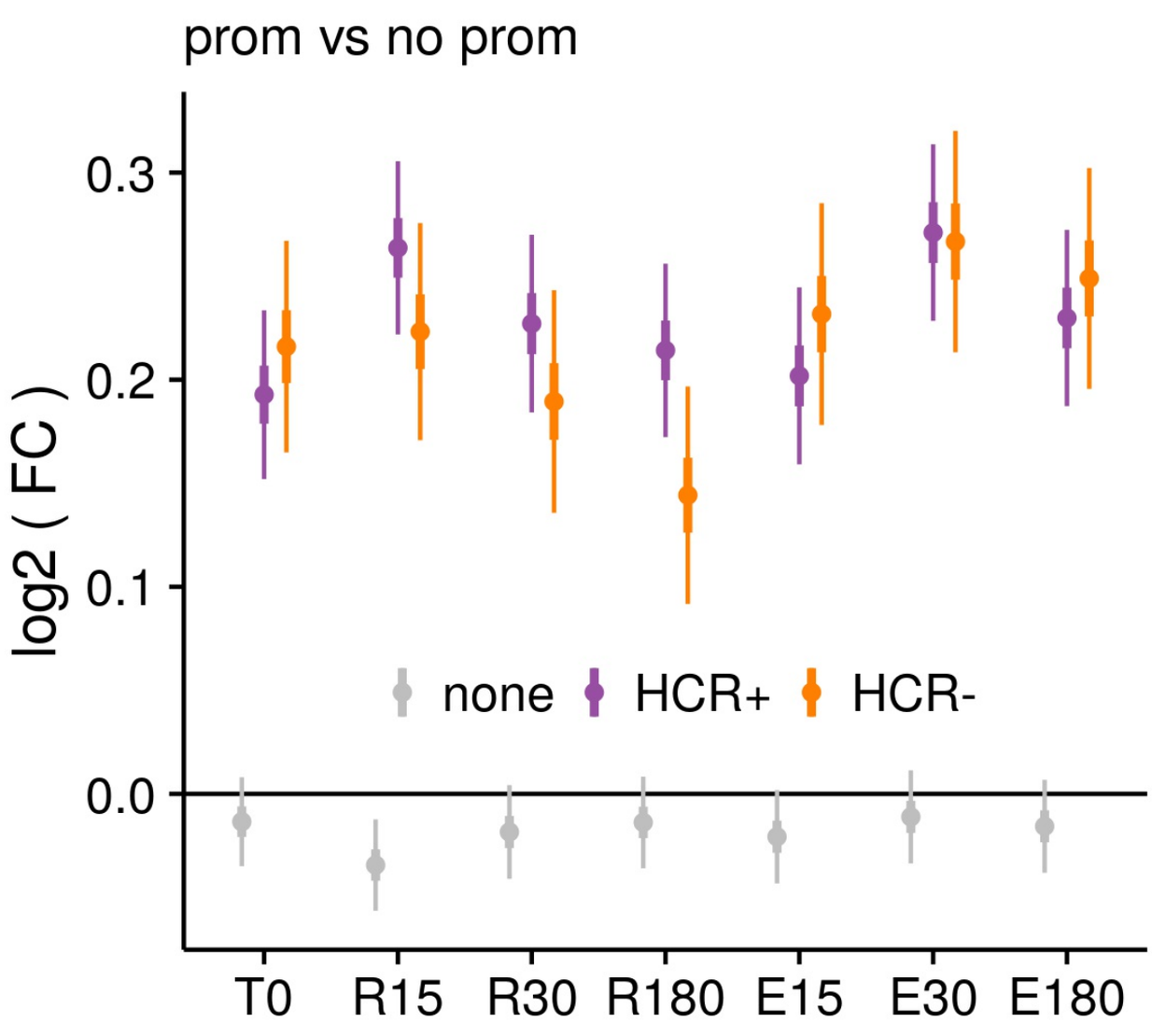


Simple comparison

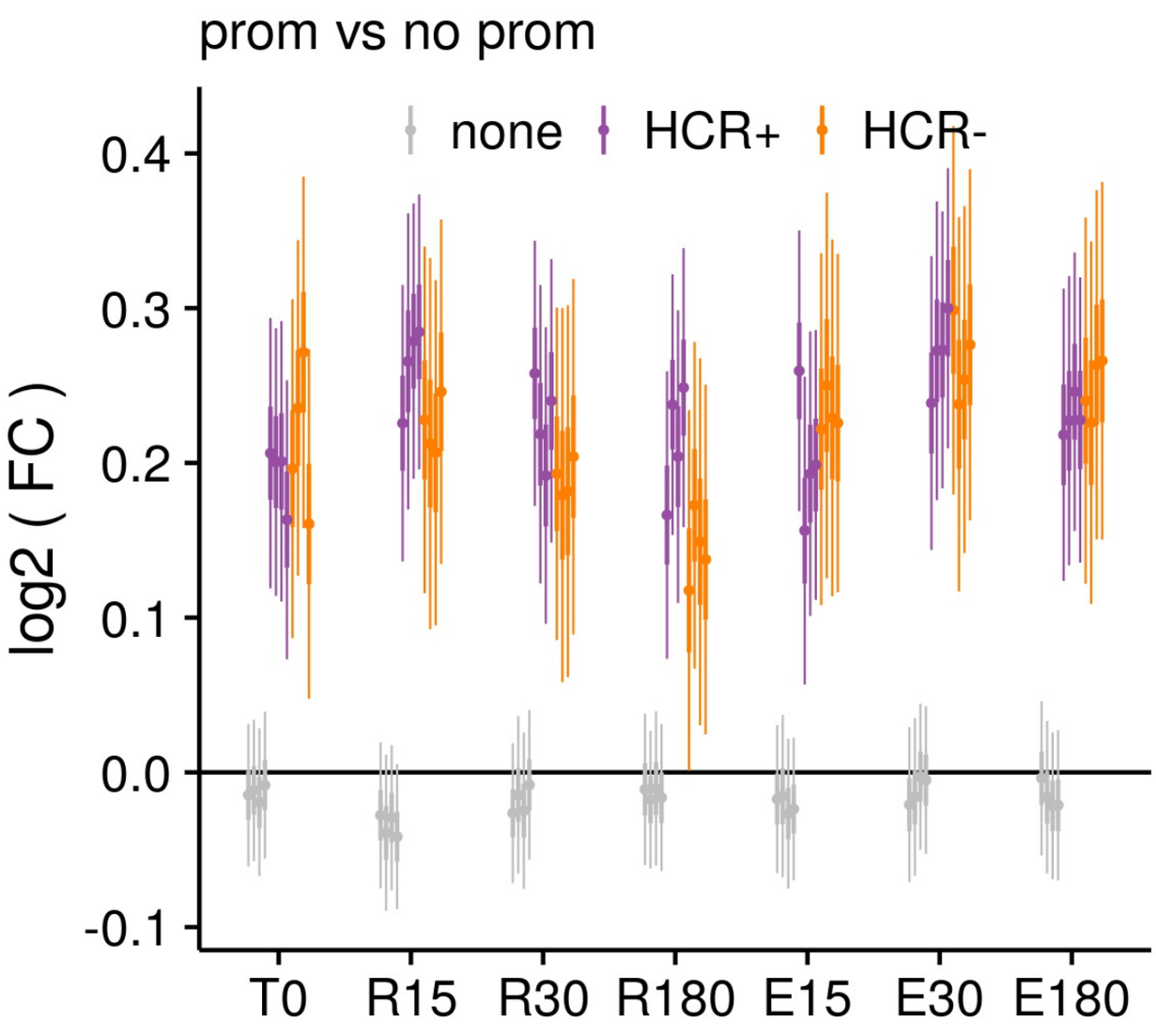




More samples

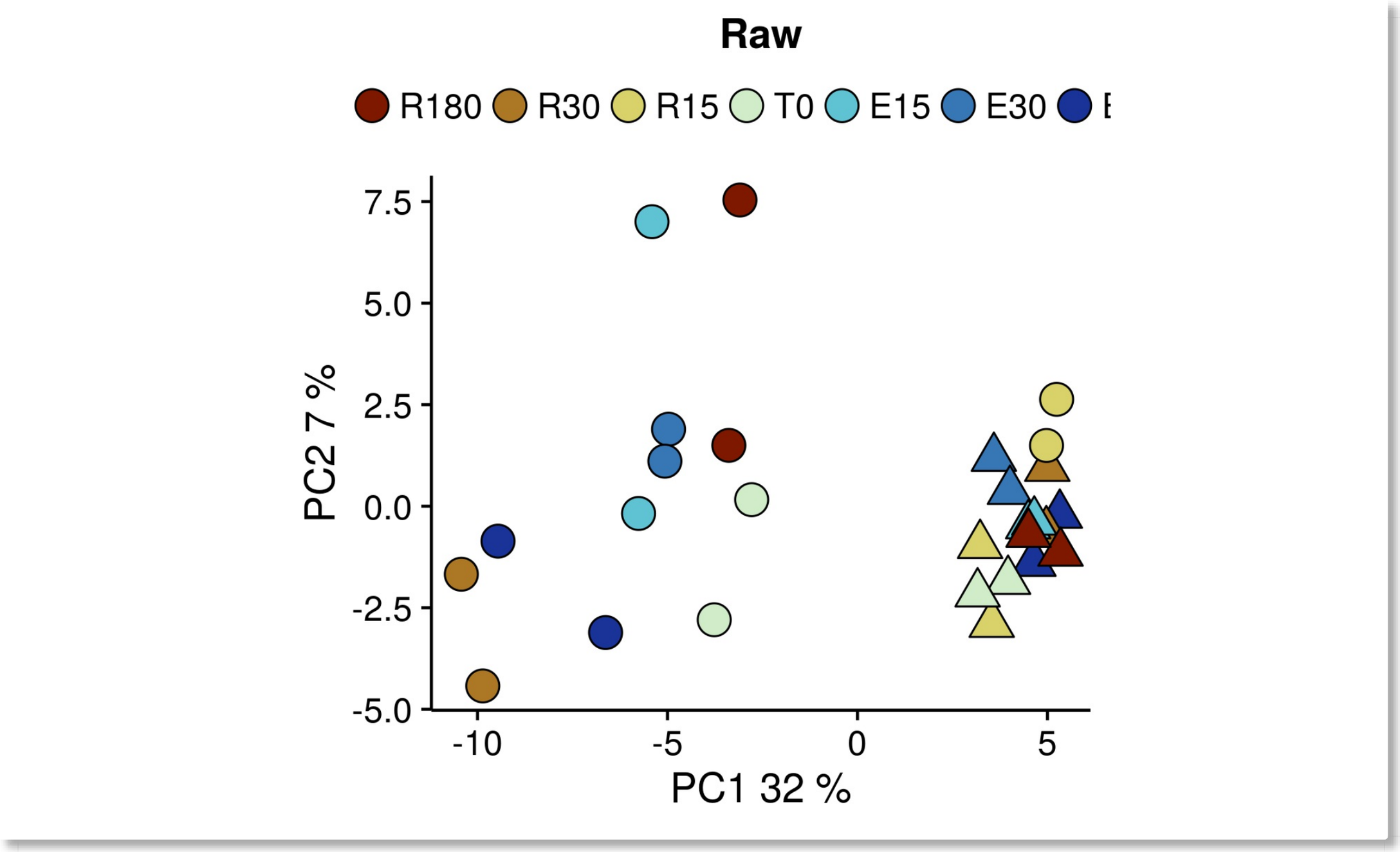


Even more samples (replicates!)

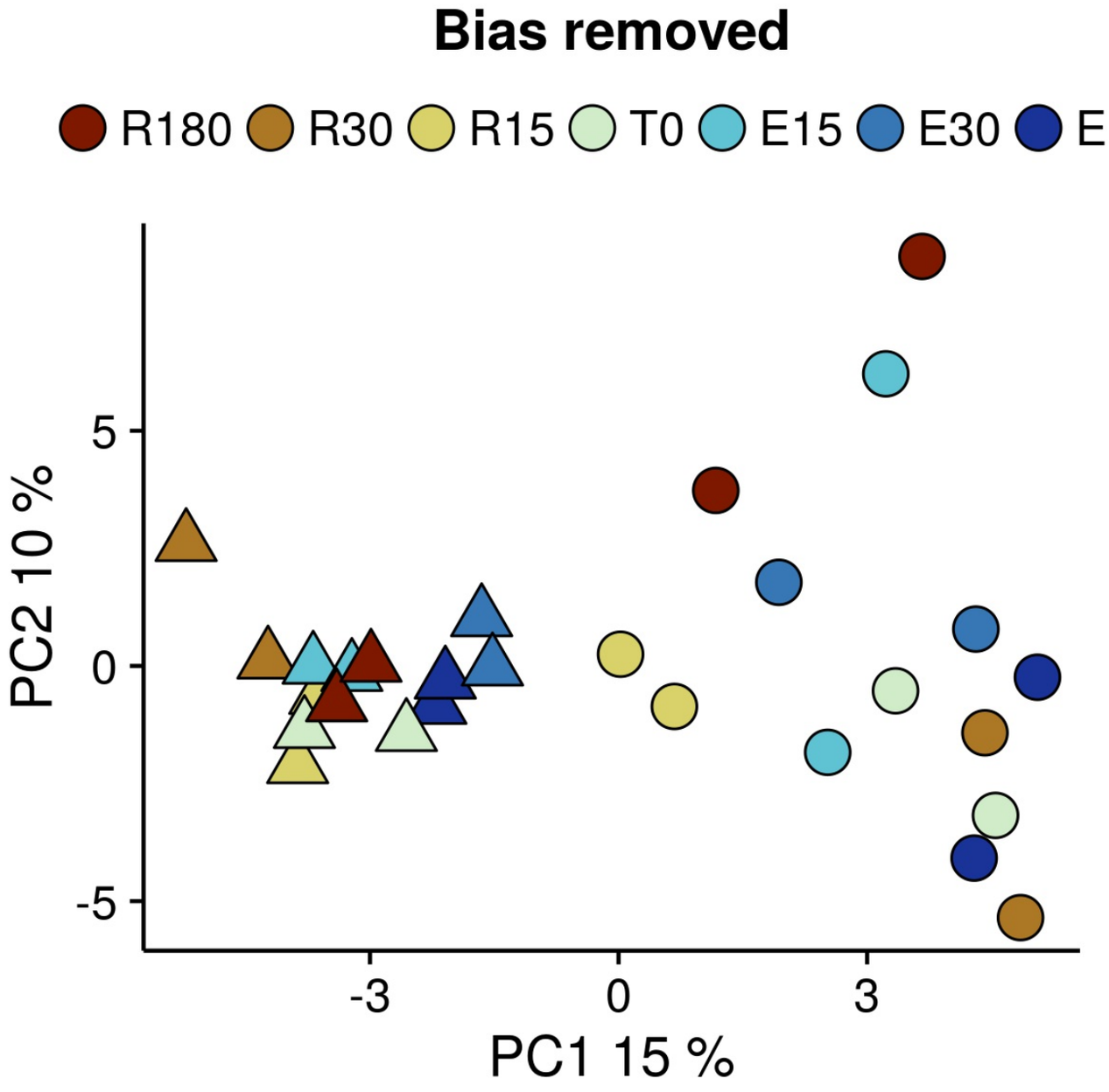


Do we have batch effect?

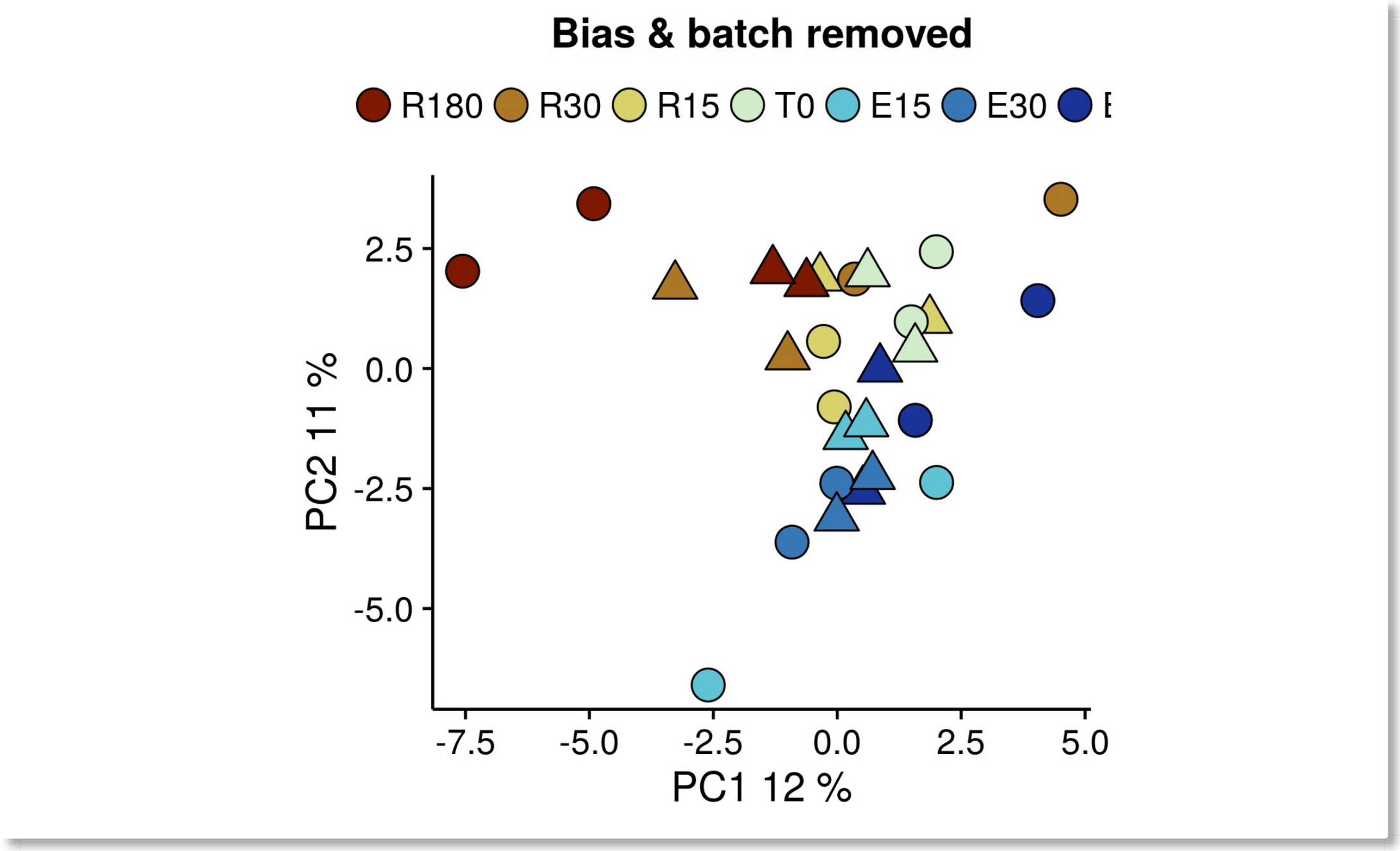
Do we have batch effect?



Removing biases helps ...

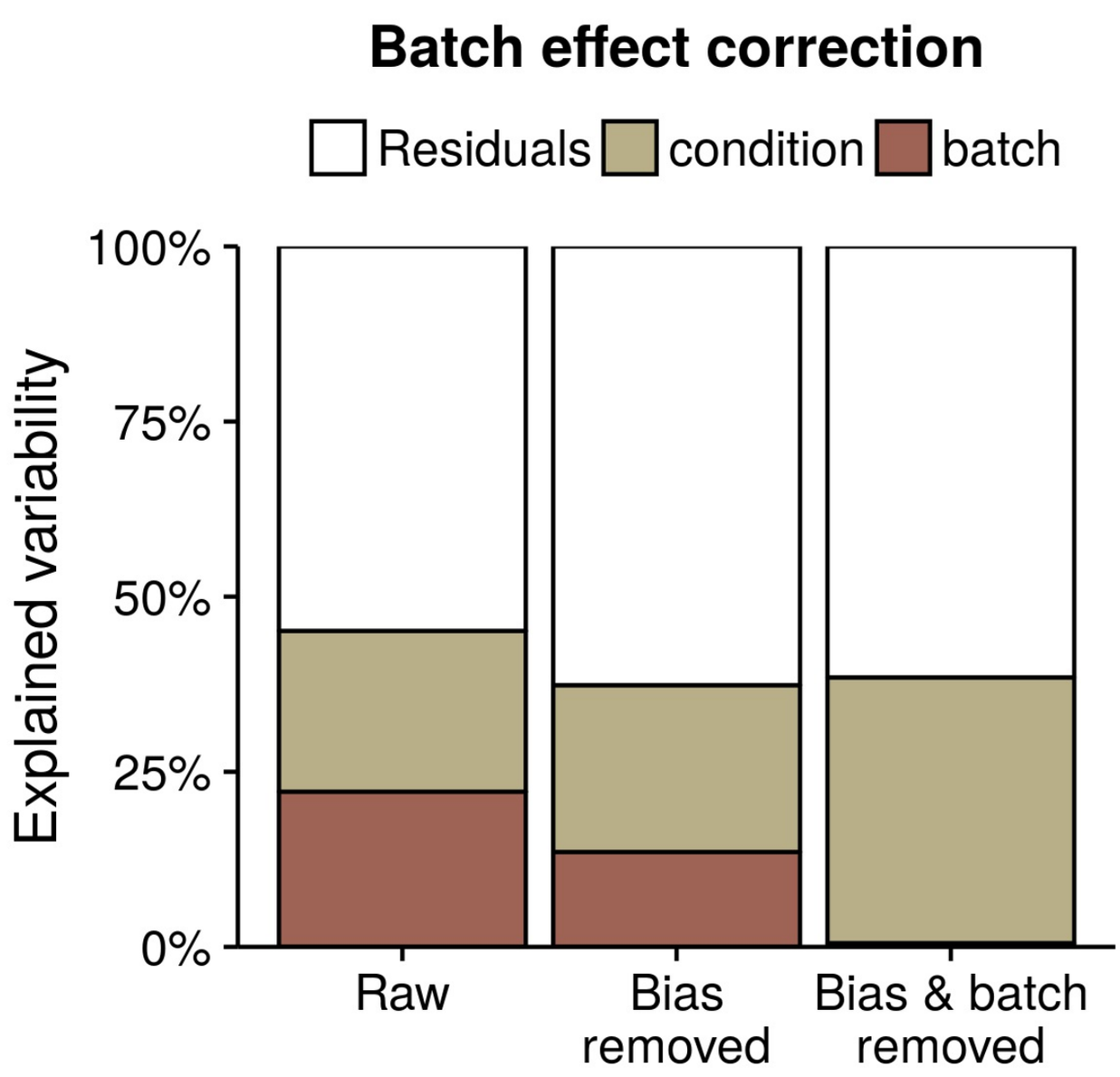


… and removing batch effect is even better



Total variance explained

Total variance explained



Grand summary

Grand summary

Genome structure plays a role and can be measured

Grand summary

Genome structure plays a role and can be measured

Increasing production of Hi-C data

Grand summary

Genome structure plays a role and can be measured

Increasing production of Hi-C data

Methodological challenges to analyze and integrate Hi-C data

Grand summary

Genome structure plays a role and can be measured

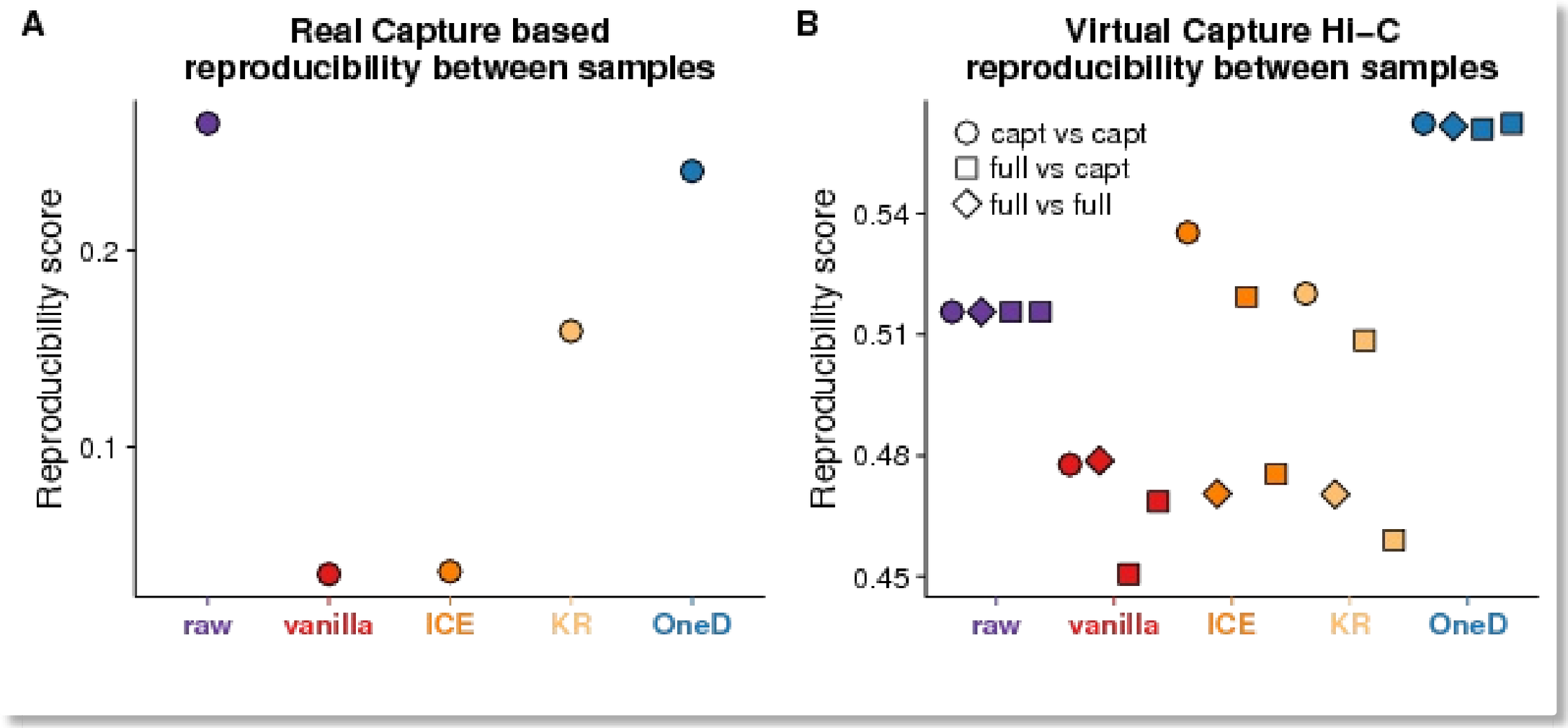
Increasing production of Hi-C data

Methodological challenges to analyze and integrate Hi-C data

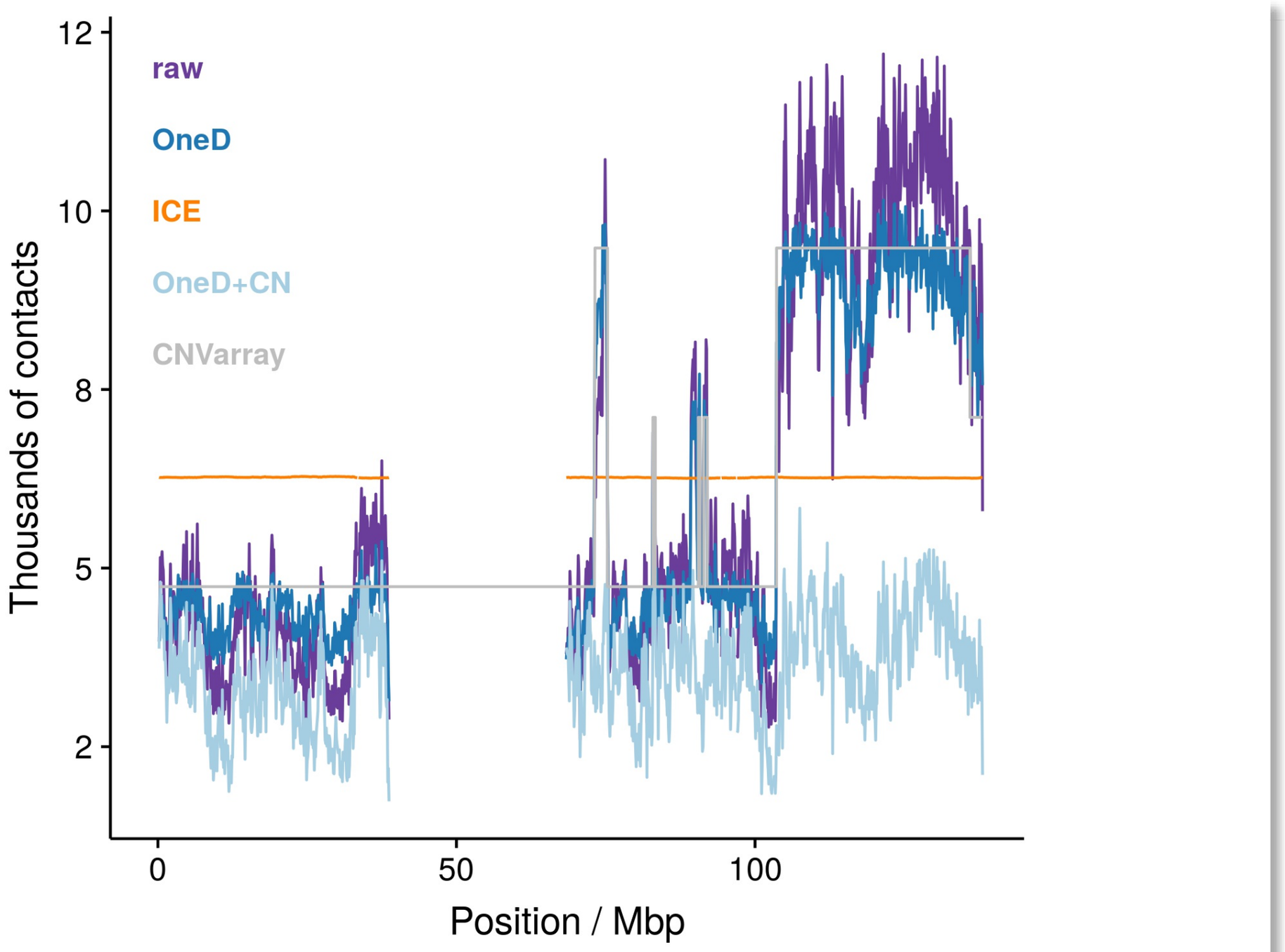
OneD: qenvio/dryhic

OneD extras

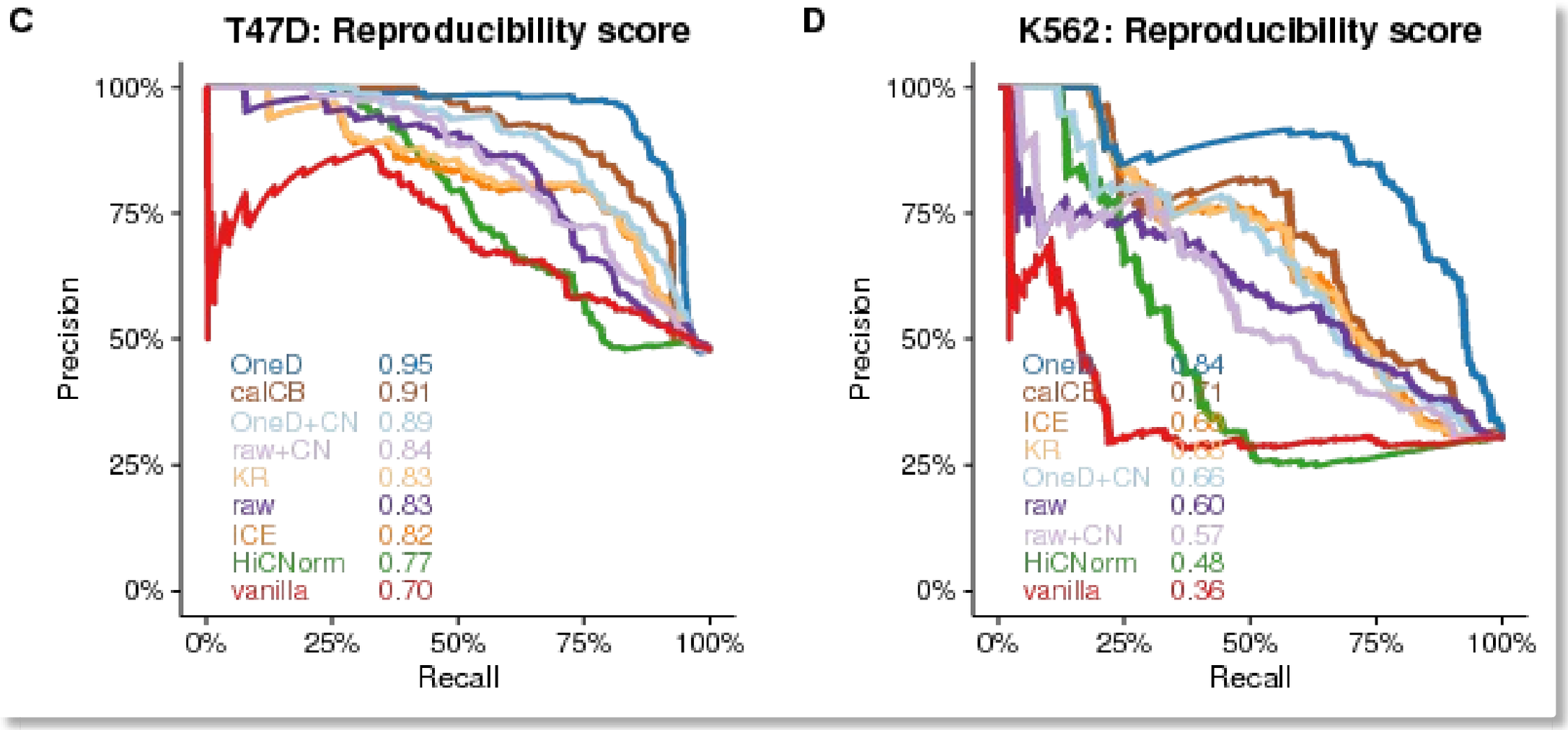
Incomplete designs



Copy number estimation



Precision-Recall



Precision-Recall