Reproducibility analysis for 3D data (work in progress)

Oana Ursu

Kundaje/Snyder labs

Department of Genetics, Stanford School of Medicine

In a nutshell

Given 2 contact maps, measure their **reproducibility**

How do we define reproducibility?

Similarity at multiple scales => Wavelet analysis

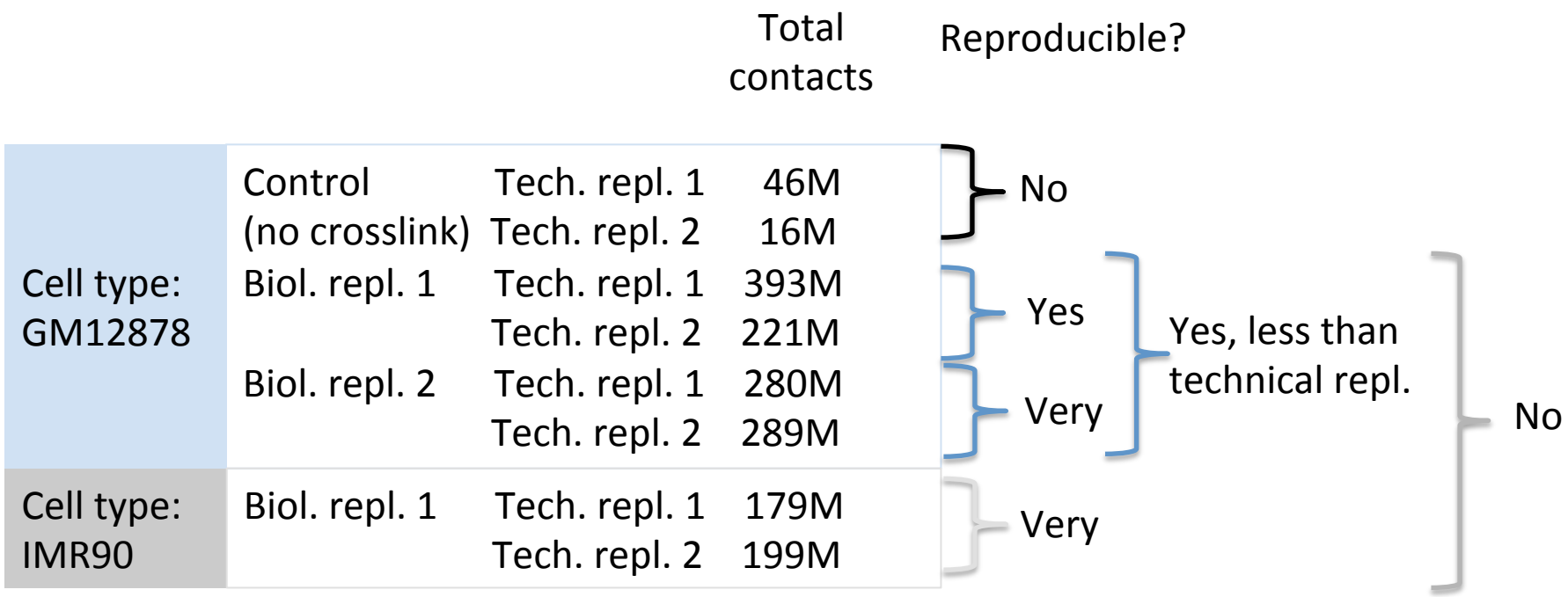
2D: compartments, domains, loops

1D: anchors

Similarity of smoothed networks => Smooth the contact maps using diffusion

Since HiC is a sparse sample of the underlying contacts, consider 2 contact maps similar even when the individual contacts they detect are not identical, as long as the overall structure is preserved

Datasets used



Data from Rao et al., Cell, 2014

Data processing

Start with matrix of observed counts at high resolution

- this presentation: 10kb, but the plan is to use 1kb
- this presentation: chr21, but the plan is to run genome-wide

Normalization: coverage

- divide each entry by the row sum and column sum
- rationale 1: analysis can begin at high resolution (for which ICE does not converge)
- rationale 2: use minimal processing for quick computation of reproducibility
- note: analysis of post-ICE inputs also available (see supplementary slides)

Resulting contact map is input to reproducibility analysis

Future

Instead of count data, use observed/expected (to remove distance-dependence)

ICE: Imakaev et al., Nature Methods, 2012

Overview

Given 2 contact maps, measure their **reproducibility**

How do we define reproducibility?

Similarity at multiple scales => Wavelet analysis

2D: compartments, domains, loops

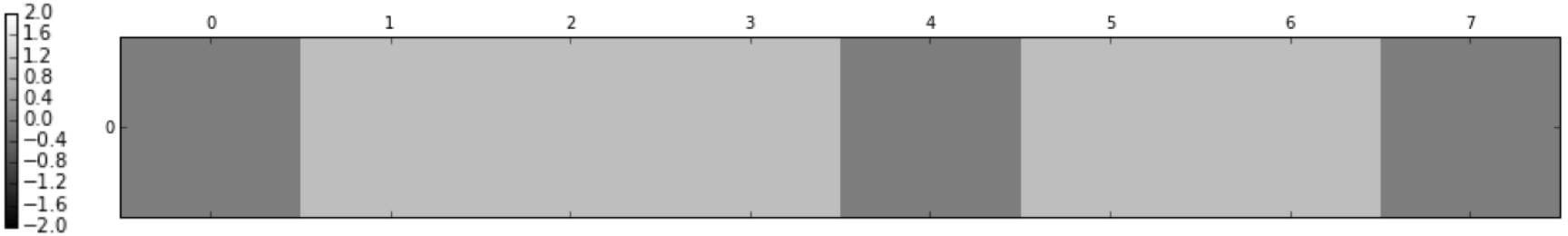
1D: anchors

Similarity of smoothed networks => Smooth the contact maps using diffusion

Since HiC is a sparse sample of the underlying contacts, consider 2 contact maps similar even when the individual contacts they detect are not identical, as long as the overall structure is preserved

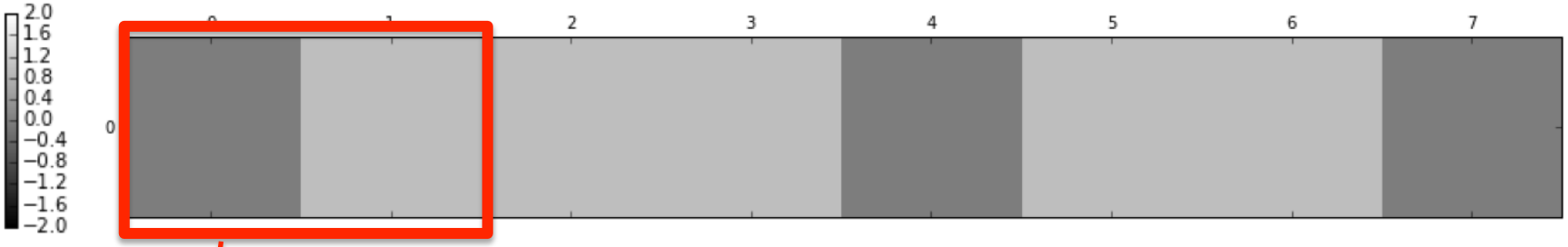
Intro to Haar wavelets (1D)

Consider a vector of values: [0,1,1,1,0,1,1,0].

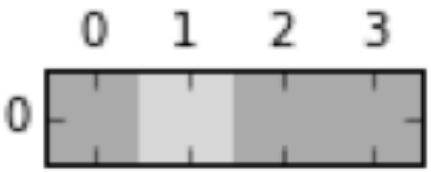


Intro to wavelets (1D)

Consider a vector of values: [0,1,1,1,0,1,1,0].
Compute the means of successive pairs of entries => approximation coefficients

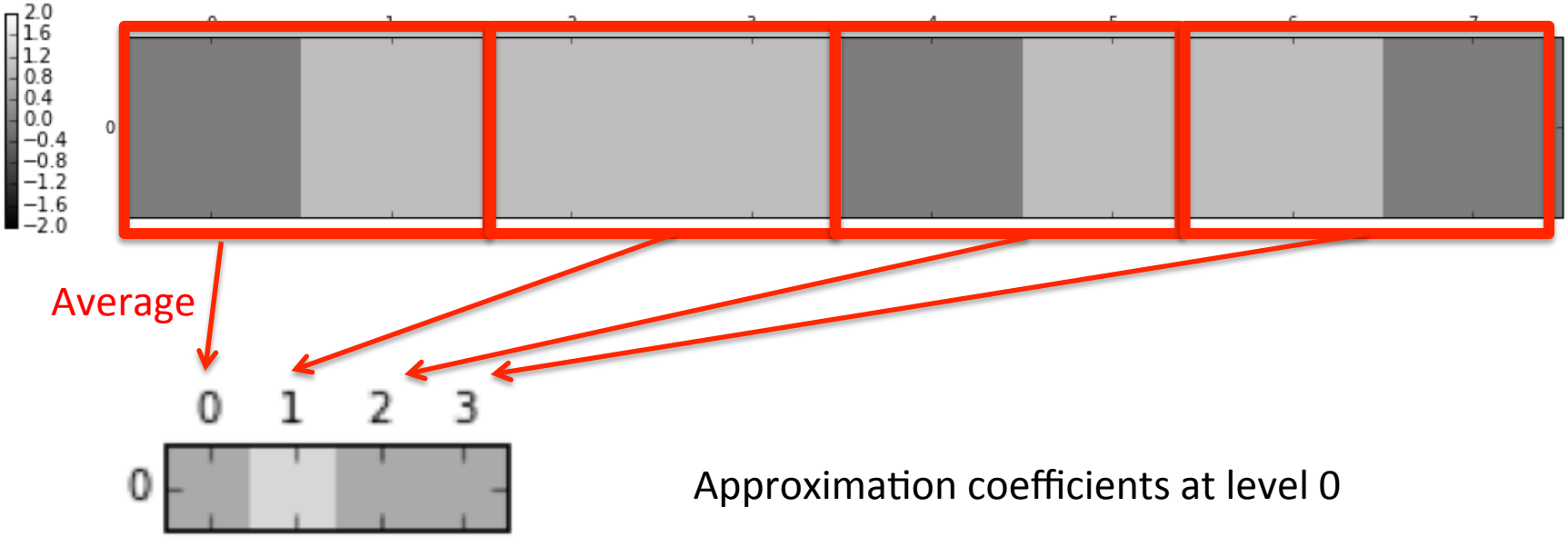


Average



Intro to wavelets (1D)

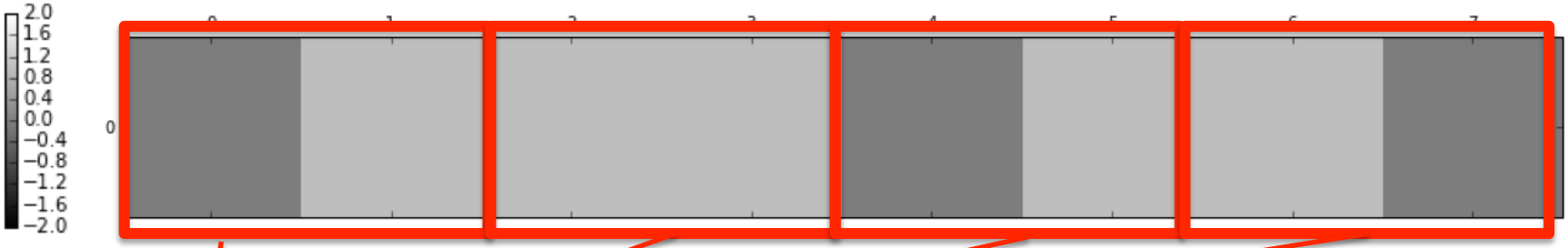
Consider a vector of values: [0,1,1,1,0,1,1,0].
Compute the means of successive pairs of entries => approximation coefficients



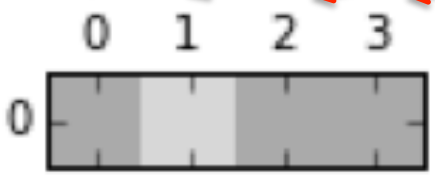
Approximation coefficients at level 0

Intro to wavelets (1D)

Consider a vector of values: [0,1,1,1,0,1,1,0].
Compute the means of successive pairs of entries => approximation coefficients



Average

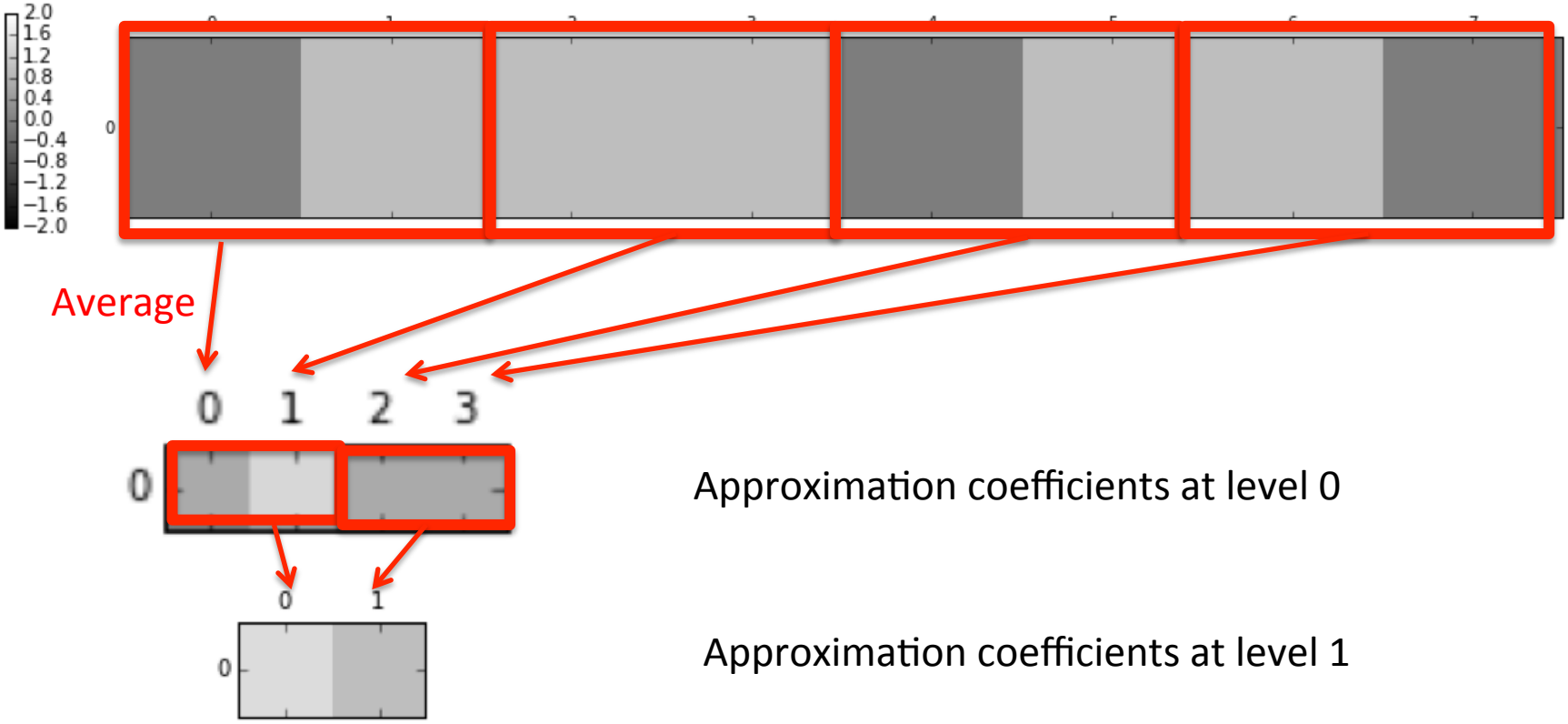


Approximation coefficients at level 0

Repeat!

Intro to wavelets (1D)

Consider a vector of values: [0,1,1,1,0,1,1,0].
Compute the means of successive pairs of entries => approximation coefficients

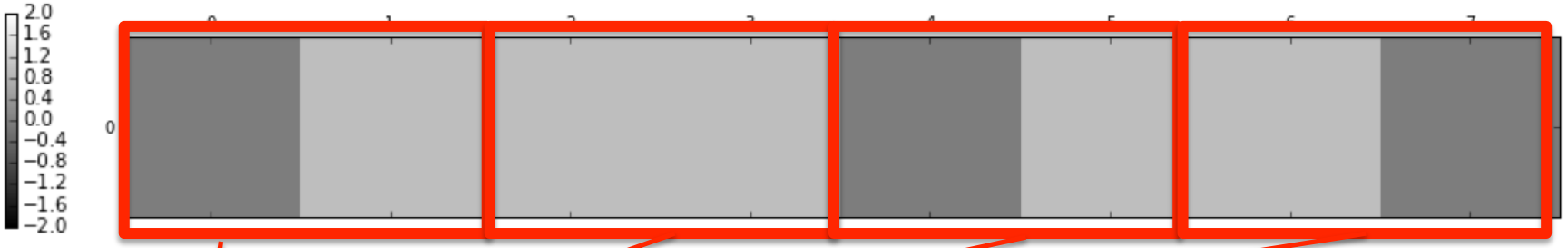


Approximation coefficients at level 0

Approximation coefficients at level 1

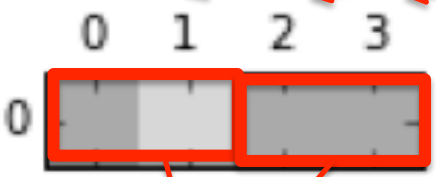
Intro to wavelets (1D)

Consider a vector of values: [0,1,1,1,0,1,1,0].
Compute the means of successive pairs of entries => approximation coefficients

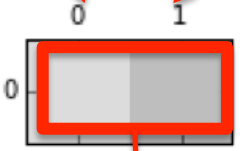


Average

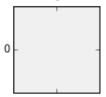
Approximation coefficients smooth the vector



Approximation coefficients at level 0



Approximation coefficients at level 1



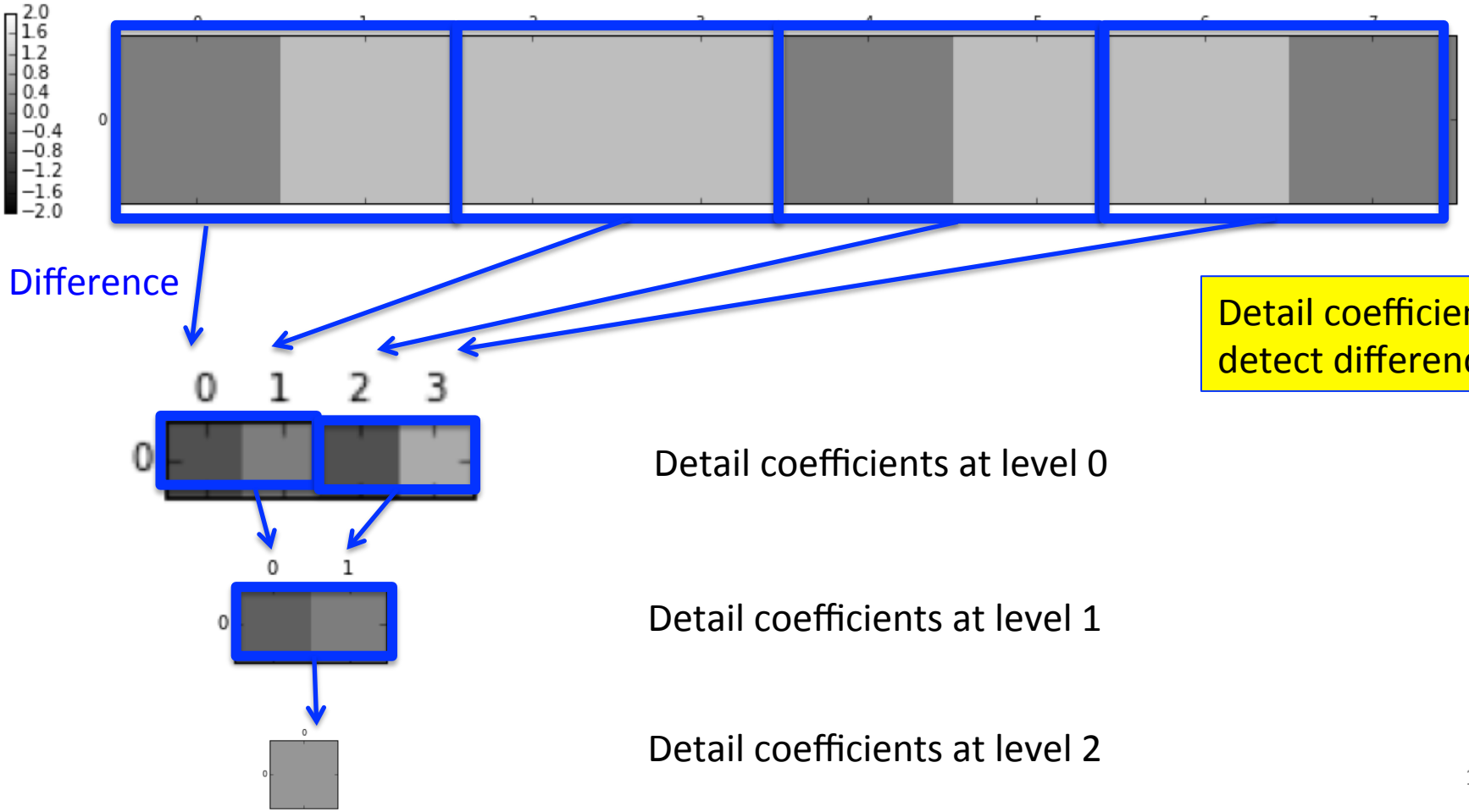
Approximation coefficients at level 2

Intro to wavelets (1D)

Similarly, we can compute differences between pairs of entries => detail coefficients

Intro to wavelets (1D)

Similarly, we can compute differences between pairs of entries => detail coefficients



Intro to wavelets (2D)

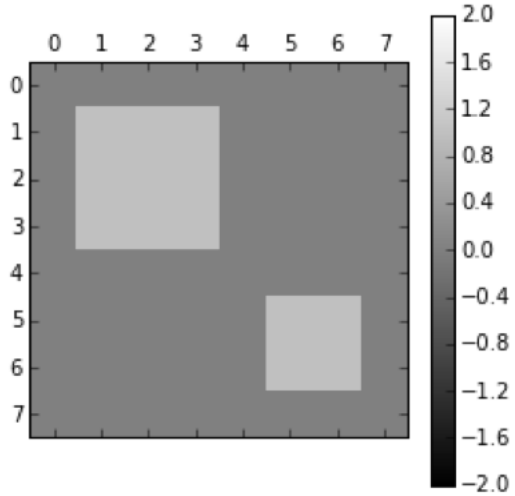
We can perform wavelet decomposition in 2D as well!

Widely used in image processing

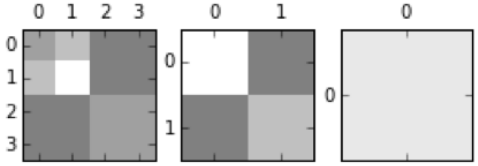
- approximation coefficients capture smoothed versions of an image
- detail coefficients capture edges, corners

Example in 2D

Start with a contact map (with 2 TADs)

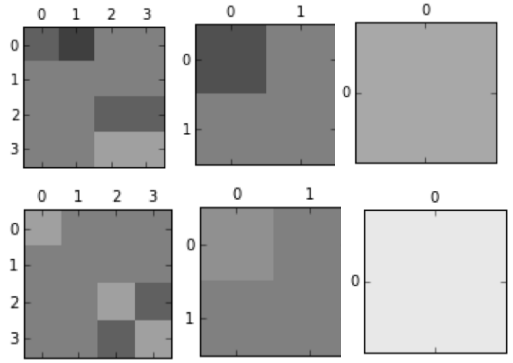


Approximation coefficients



Rows: approx
Cols: approx

Detail coefficients



Horizontal/ vertical

Rows: approx
Cols: detail

Diagonal

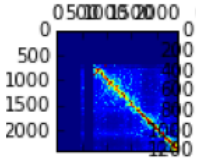
Rows: detail
Cols: detail

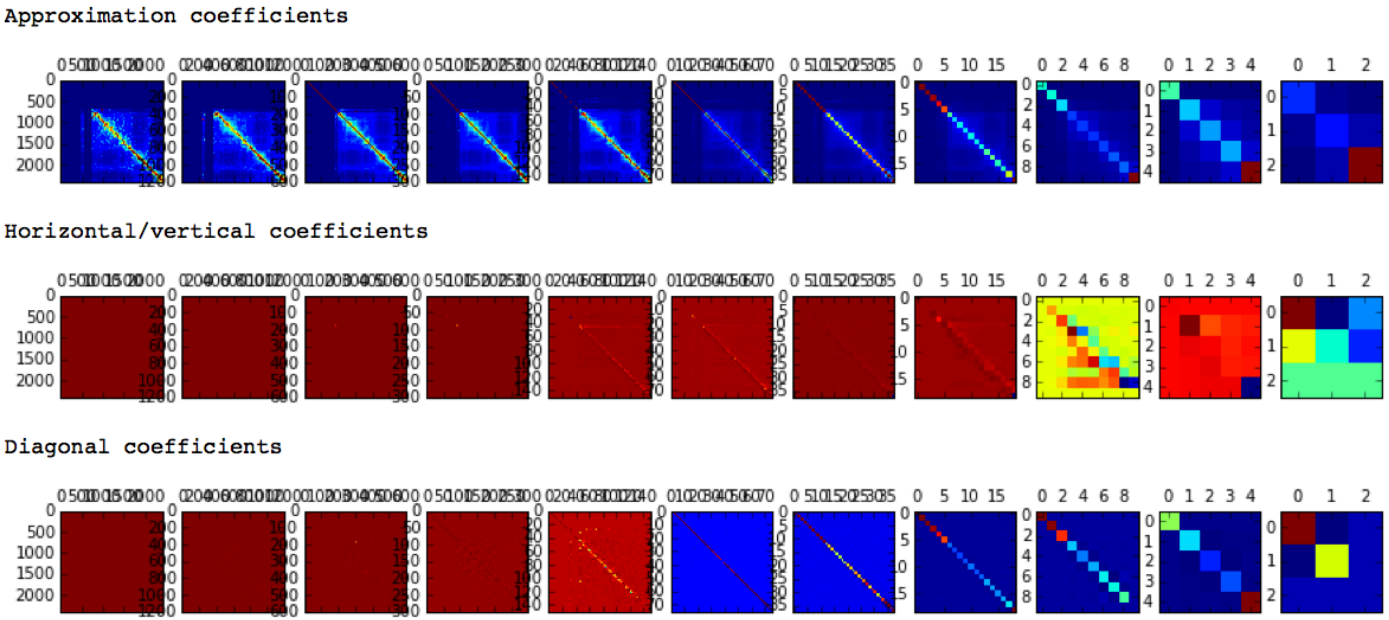
Levels ----->

Strategy for comparing contact maps using wavelets

Contact map => wavelet coefficients

Contact map





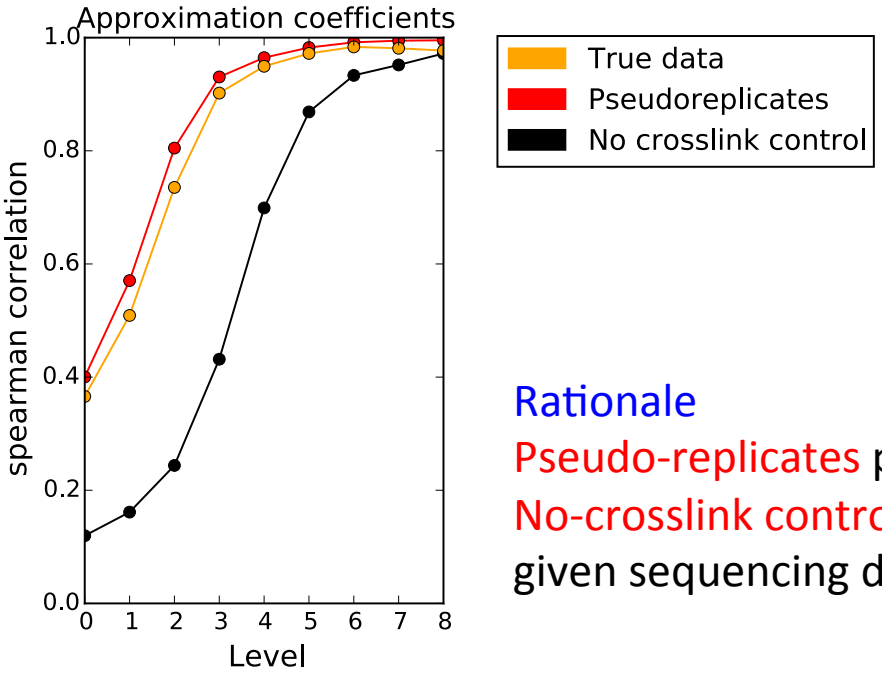
Levels ----->

Strategy for comparing contact maps using wavelets

Contact map => wavelet coefficients

At each level, compute the correlation between wavelet coefficients for the 2 samples

=> Compute an AUC (AUC/total area), so [0,1]



Rationale

Pseudo-replicates provide an upper bound for reproducibility

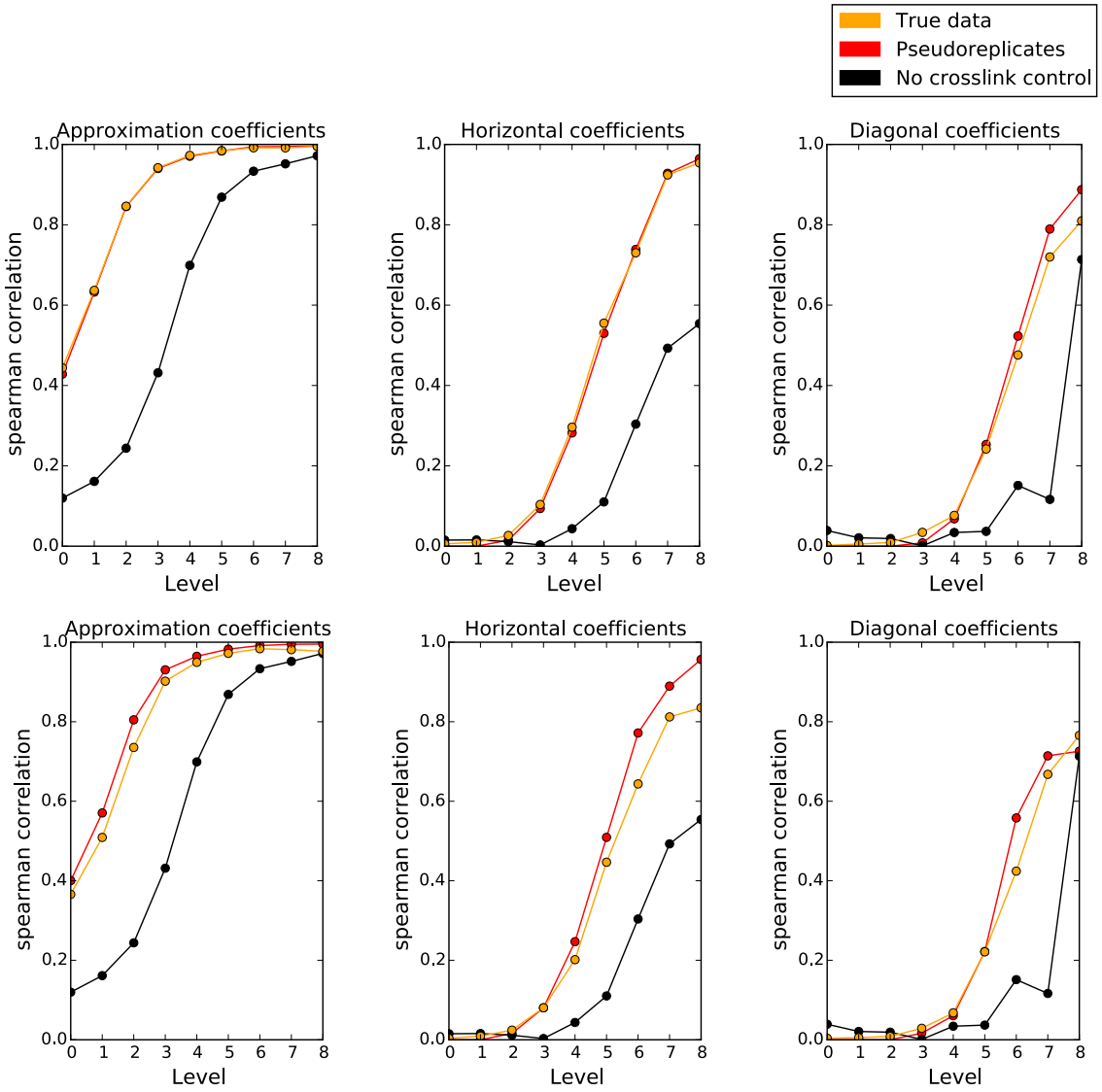
No-crosslink control is a reference for low reproducibility at a given sequencing depth

Highlights (1)
Technical replicates > biological replicates

Reproducibility

Technical replicates
Biological replicate 2
Techrep 1 vs Techrep 2

Biological replicates
Biological replicate 1
vs
Biological replicate 2



Highlights (2)
Same cell type > different cell types

Reproducibility

Same cell type

Biological replicate 1 (GM12878)

vs

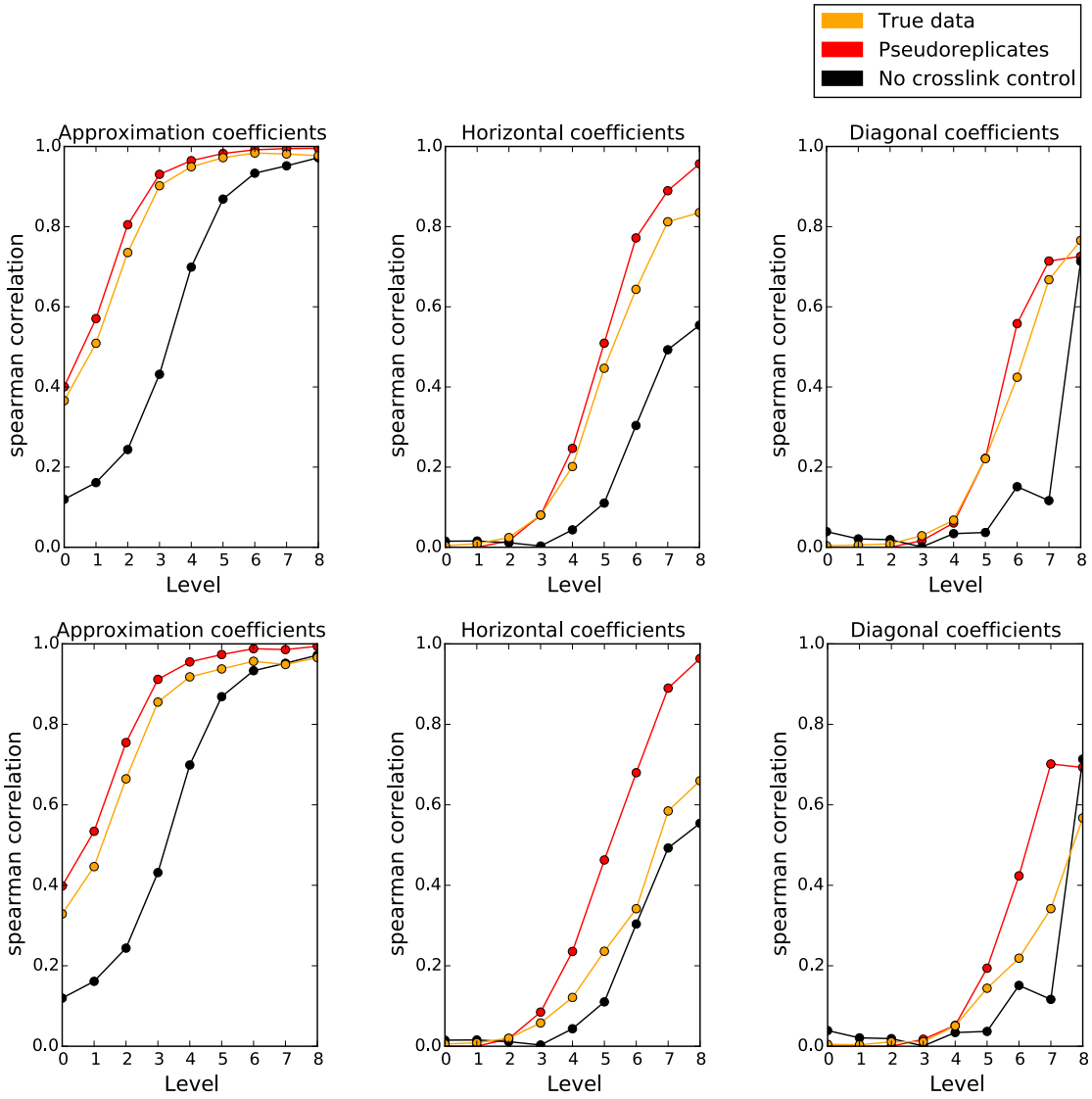
Biological replicate 2 (GM12878)

Cross cell type

Biological replicate 1 (GM12878)

vs

Biological replicate 1 (IMR90)



Reproducibility statistics

$AUC_{true} - AUC_{no\ crosslink}$

⇒ Reproducible contact maps will be very different from the no-crosslink

$AUC_{true} / AUC_{pseudoreplicates}$

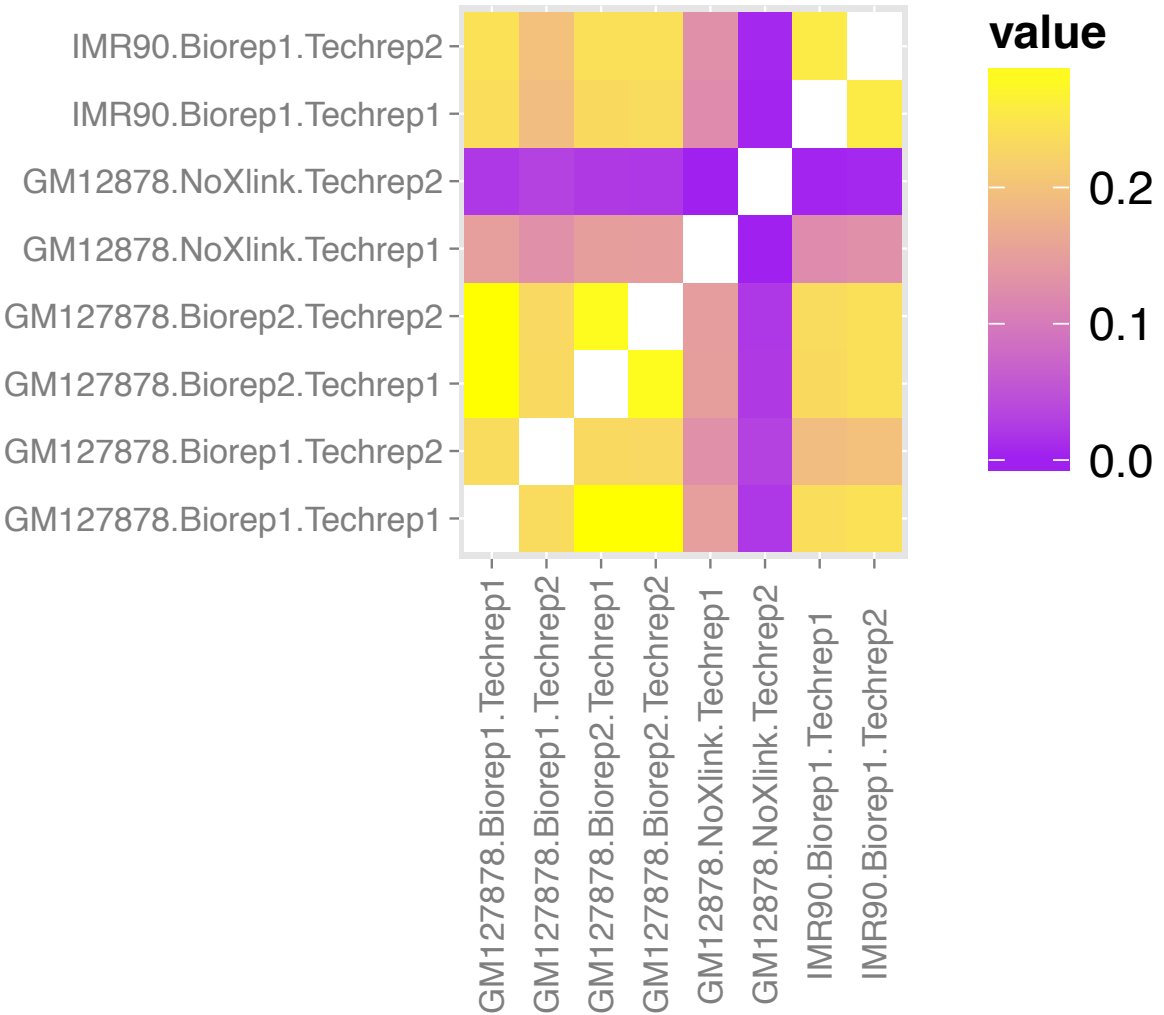
⇒ Reproducible contact maps will be close to the reproducibility of pseudoreplicates

Minimum level at which a threshold correlation is attained

=> Reproducible contact maps will achieve high correlations earlier

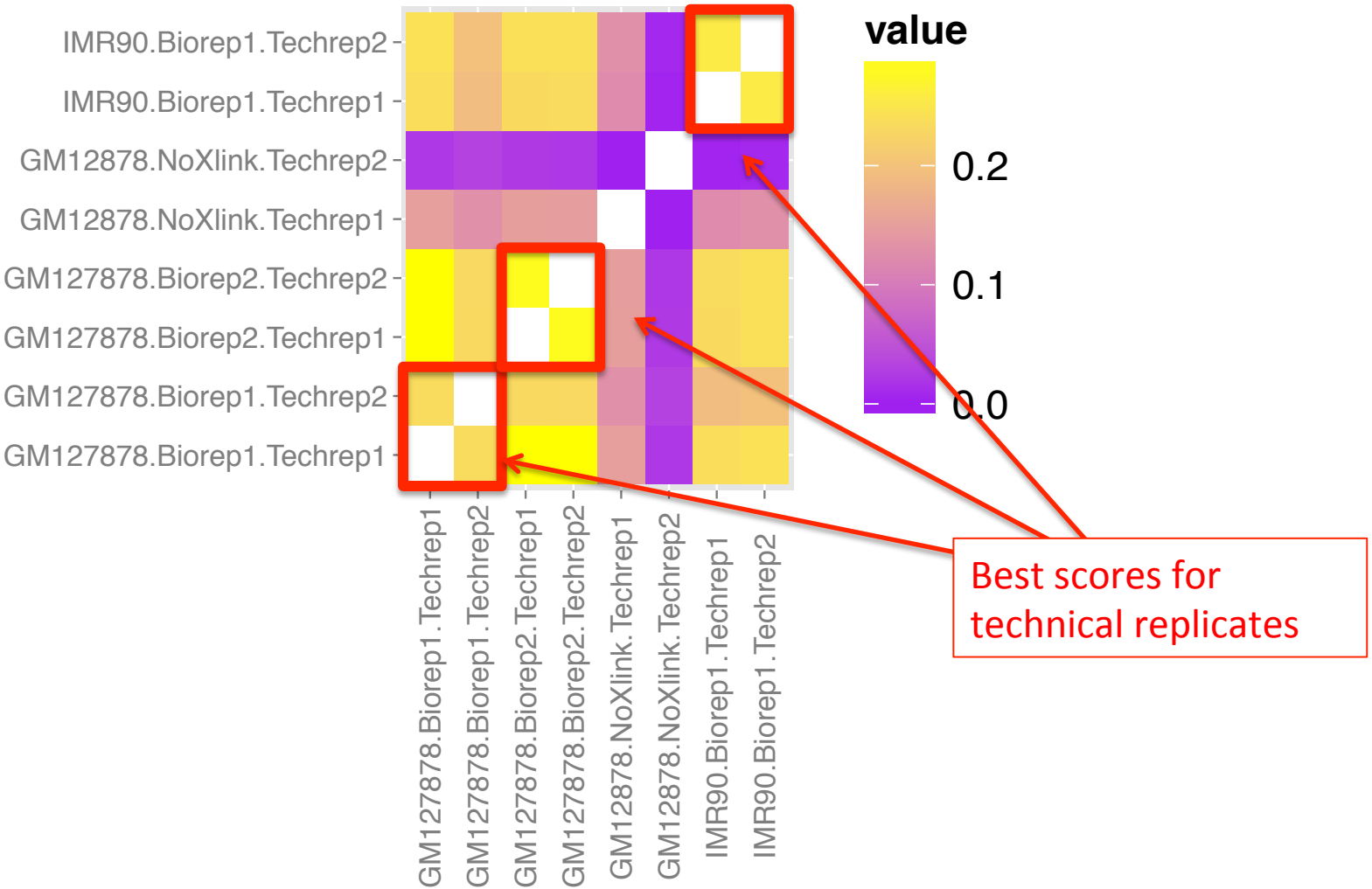
AUC[true] – AUC[no-crosslink]

(approximation coefficients)



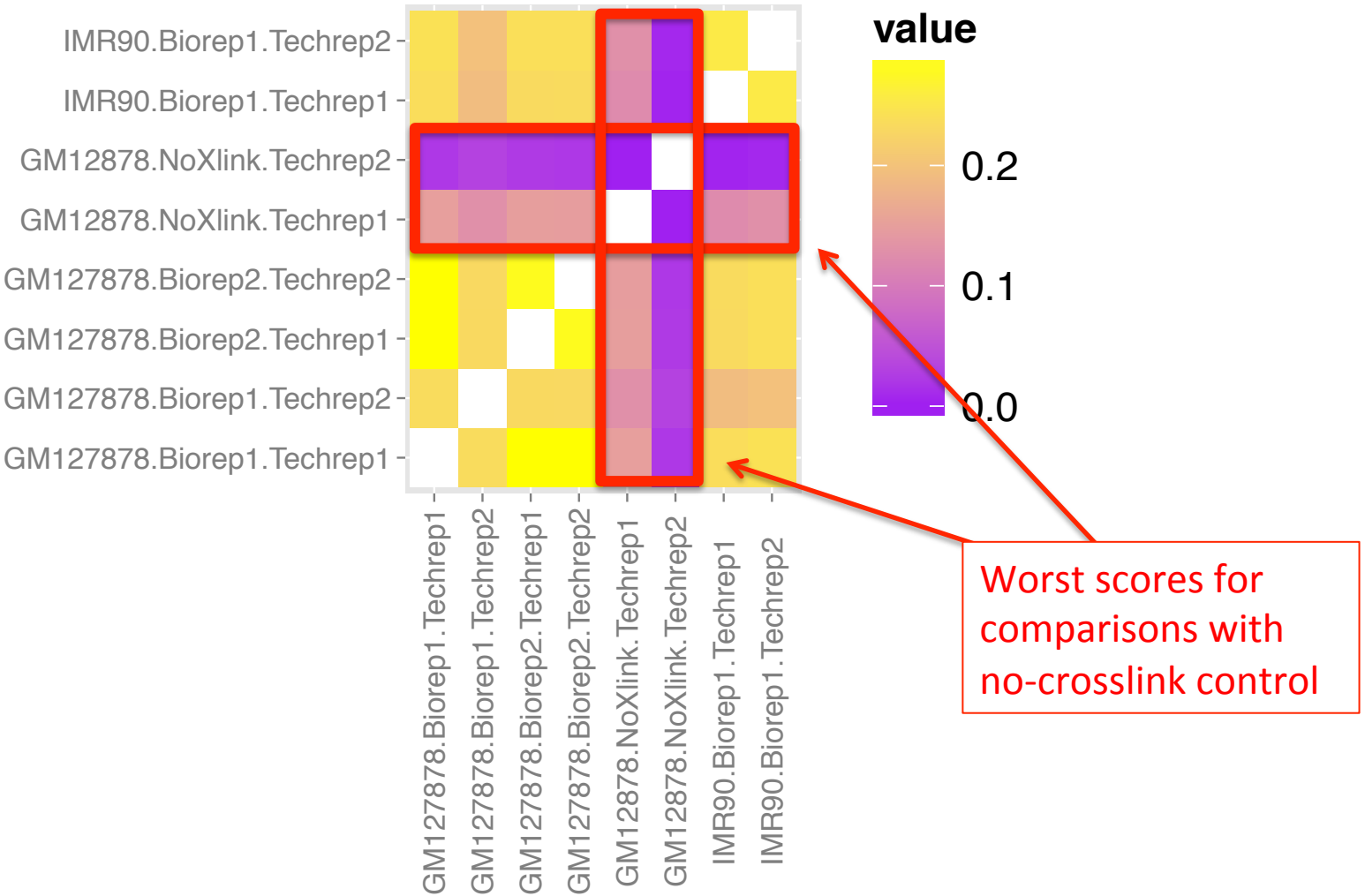
AUC[true] – AUC[no-crosslink]

(approximation coefficients)



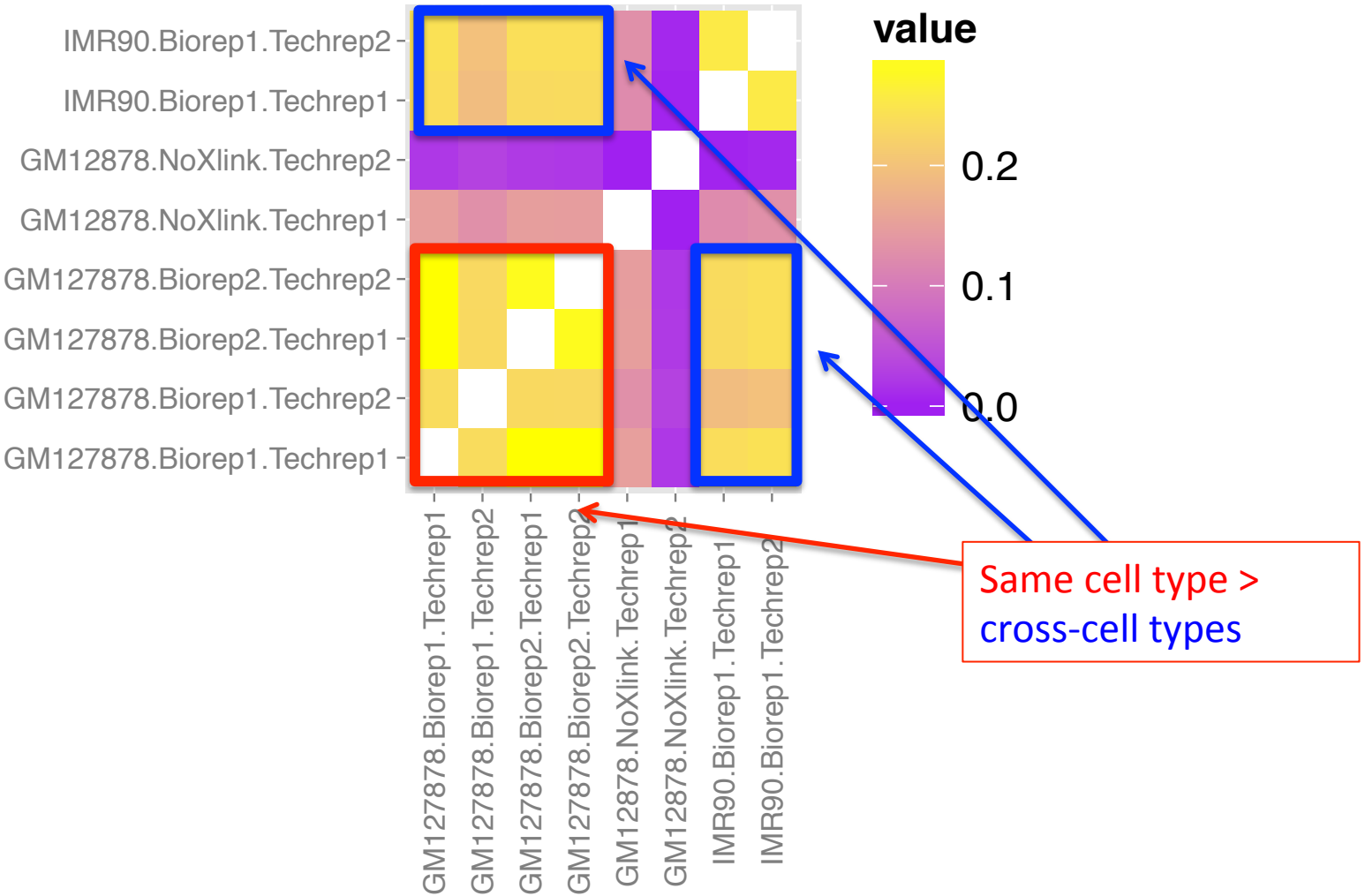
AUC[true] – AUC[no-crosslink]

(approximation coefficients)



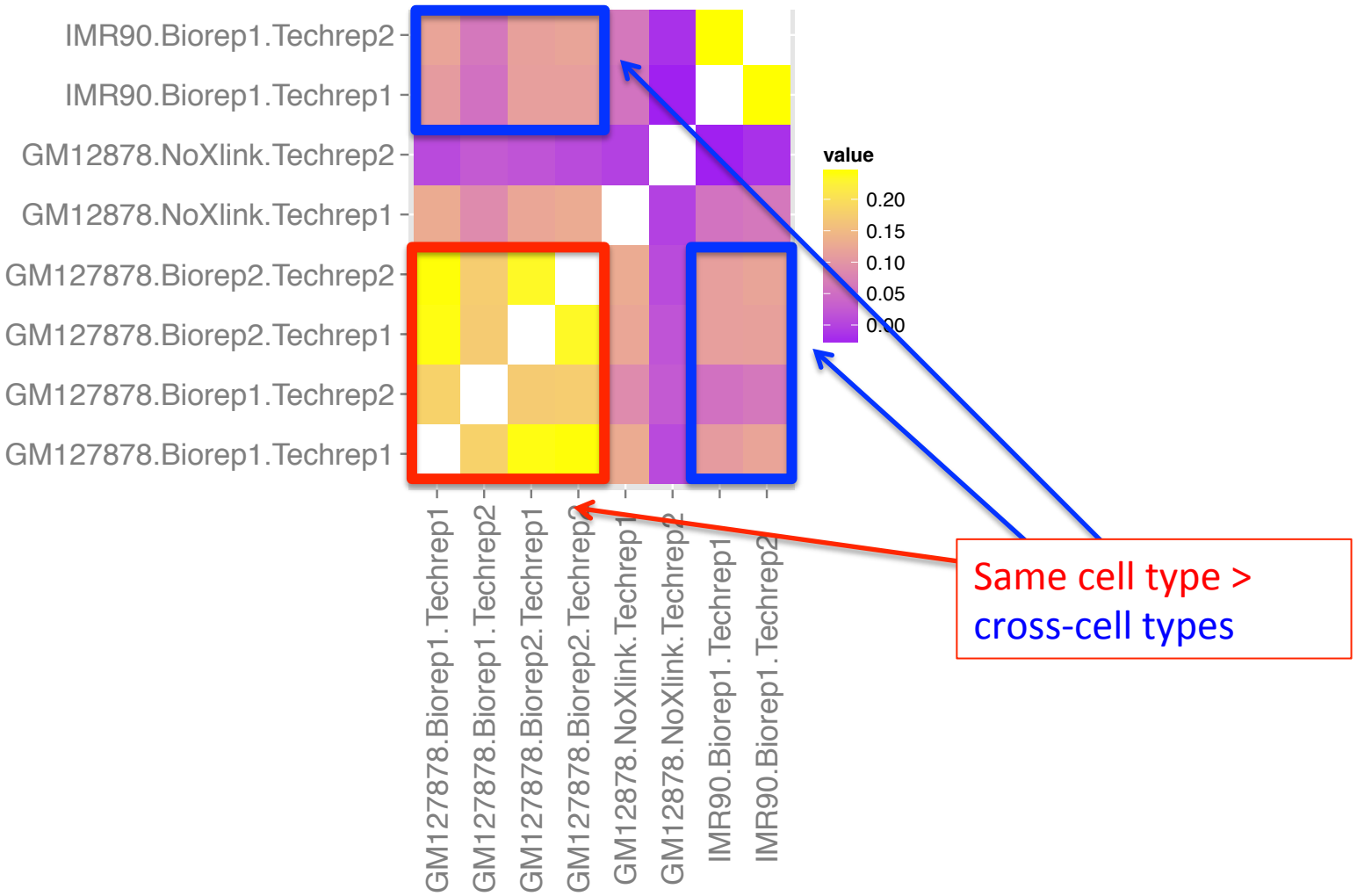
AUC[true] – AUC[no-crosslink]

(approximation coefficients)



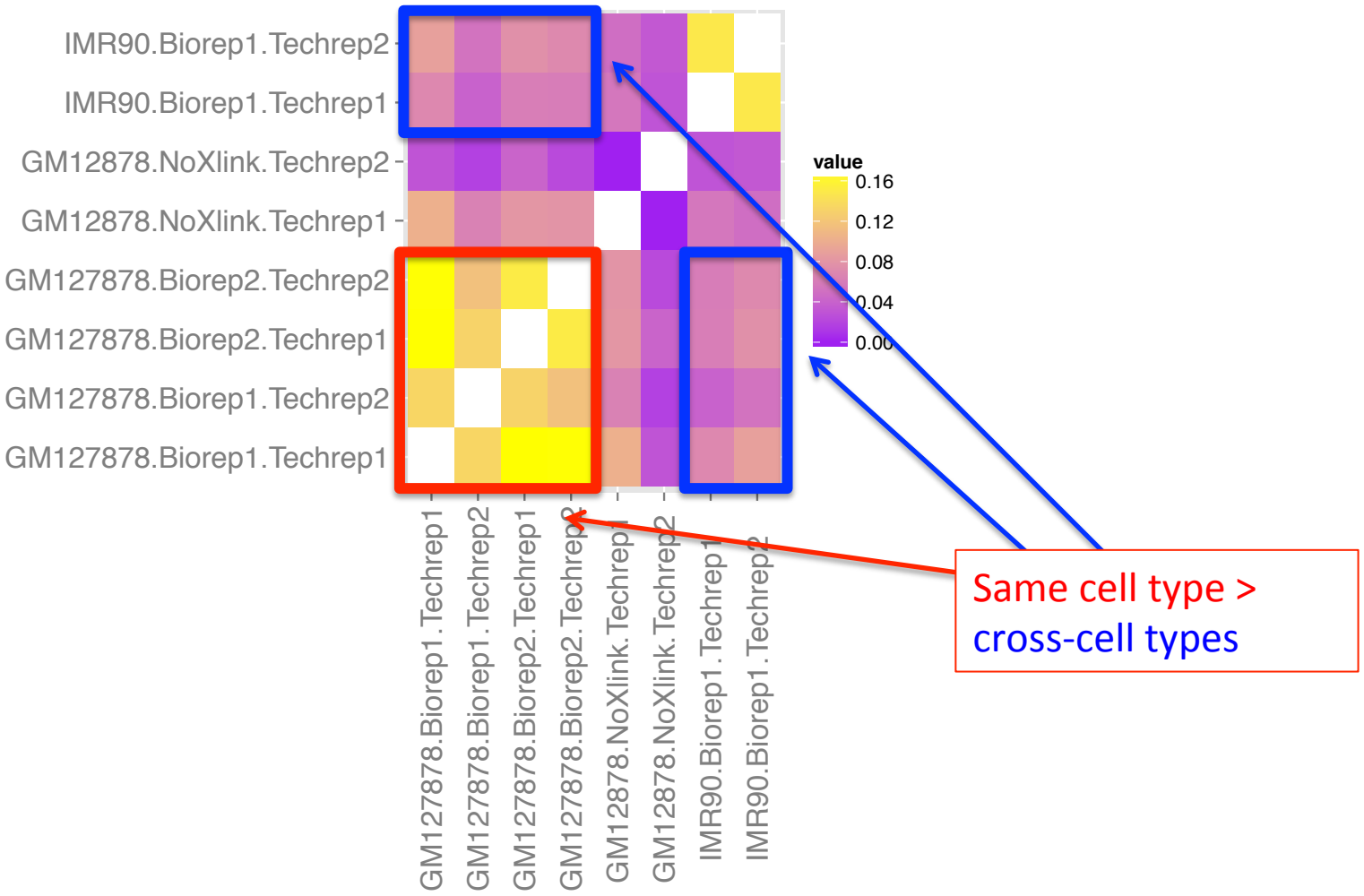
AUC[true] – AUC[no-crosslink]

(horizontal coefficients)

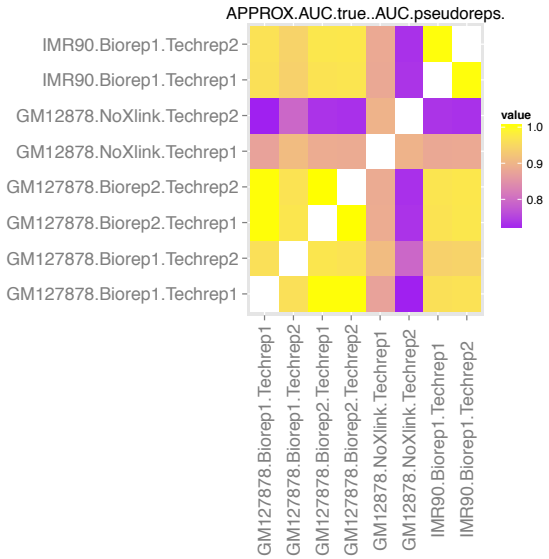


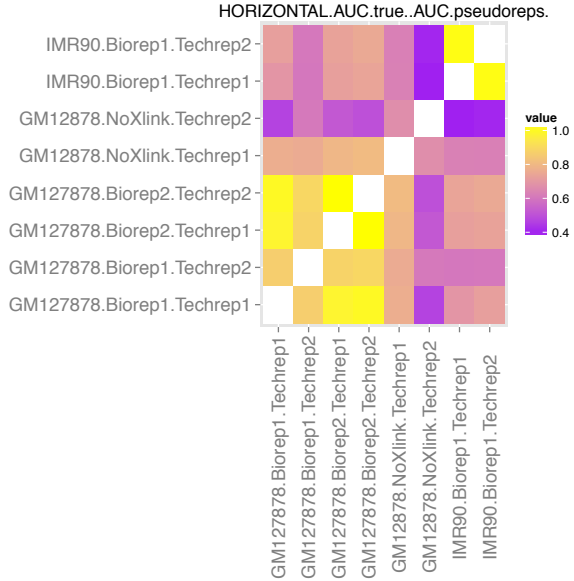
AUC[true] – AUC[no-crosslink]

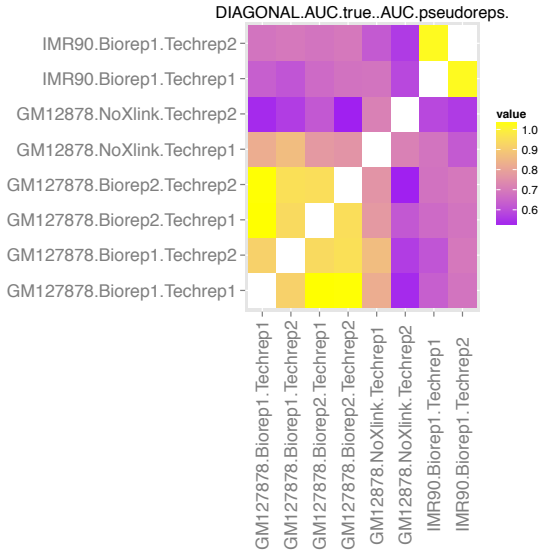
(diagonal coefficients)



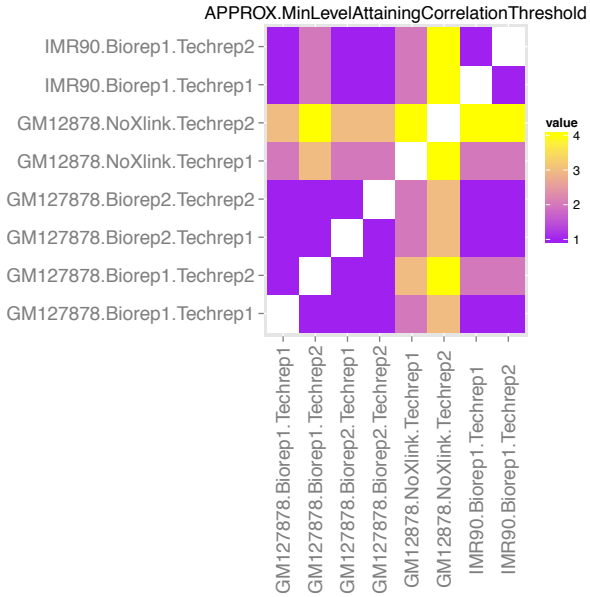
AUC[true] / AUC[pseudoreplicates]

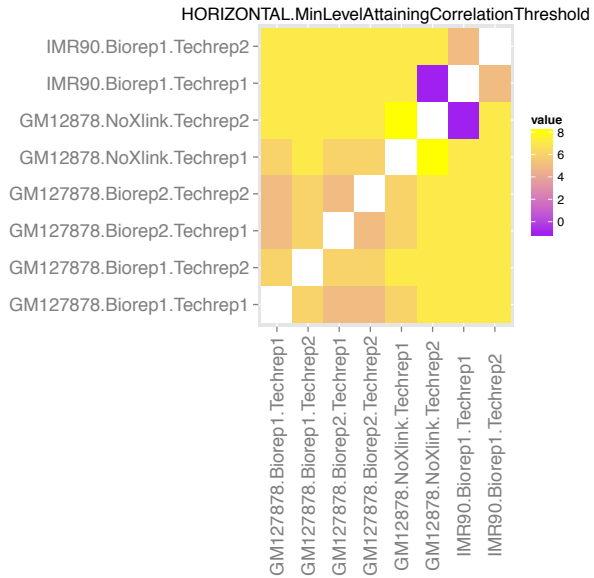


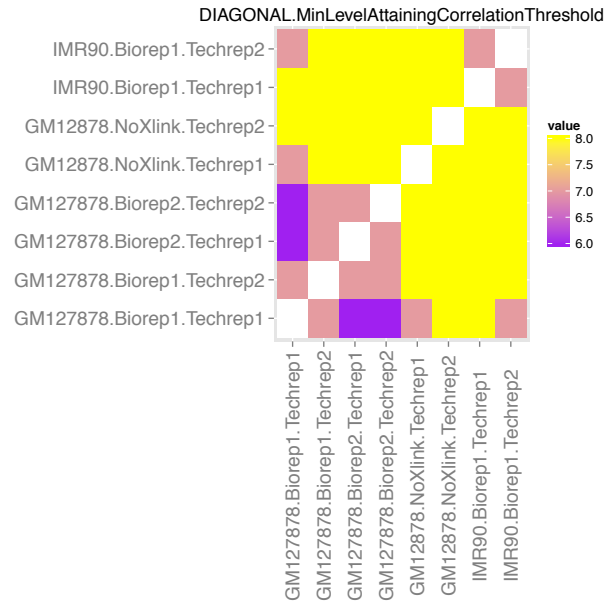




Level at which correlation >= 0.5





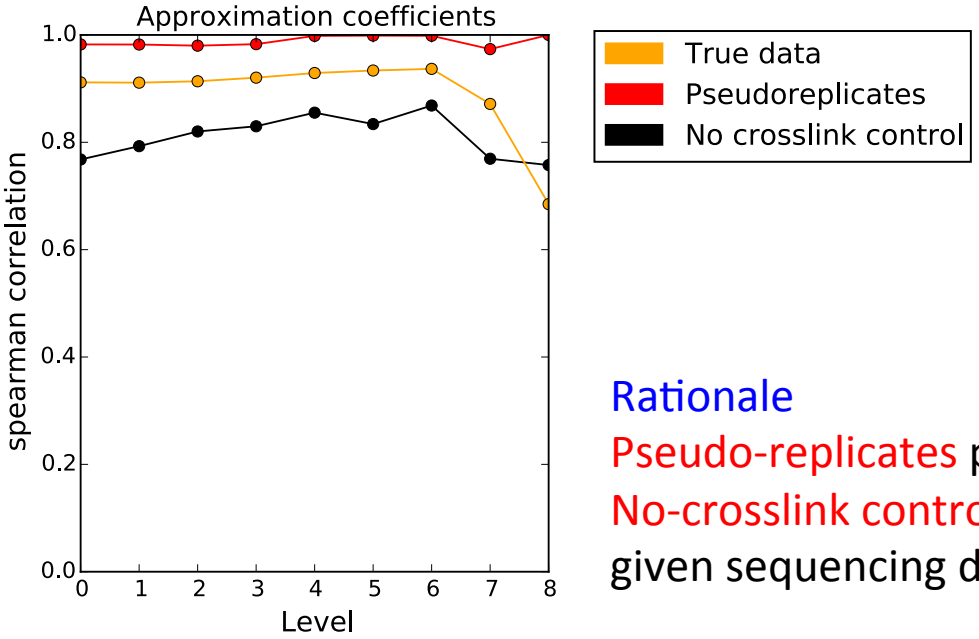


Applying the same strategy in 1D to check reproducibility of anchors

Contact map => row sums => wavelet coefficients in 1D

At each level, compute the correlation between wavelet coefficients for the 2 samples

=> Compute an AUC (AUC/total area), so [0,1]



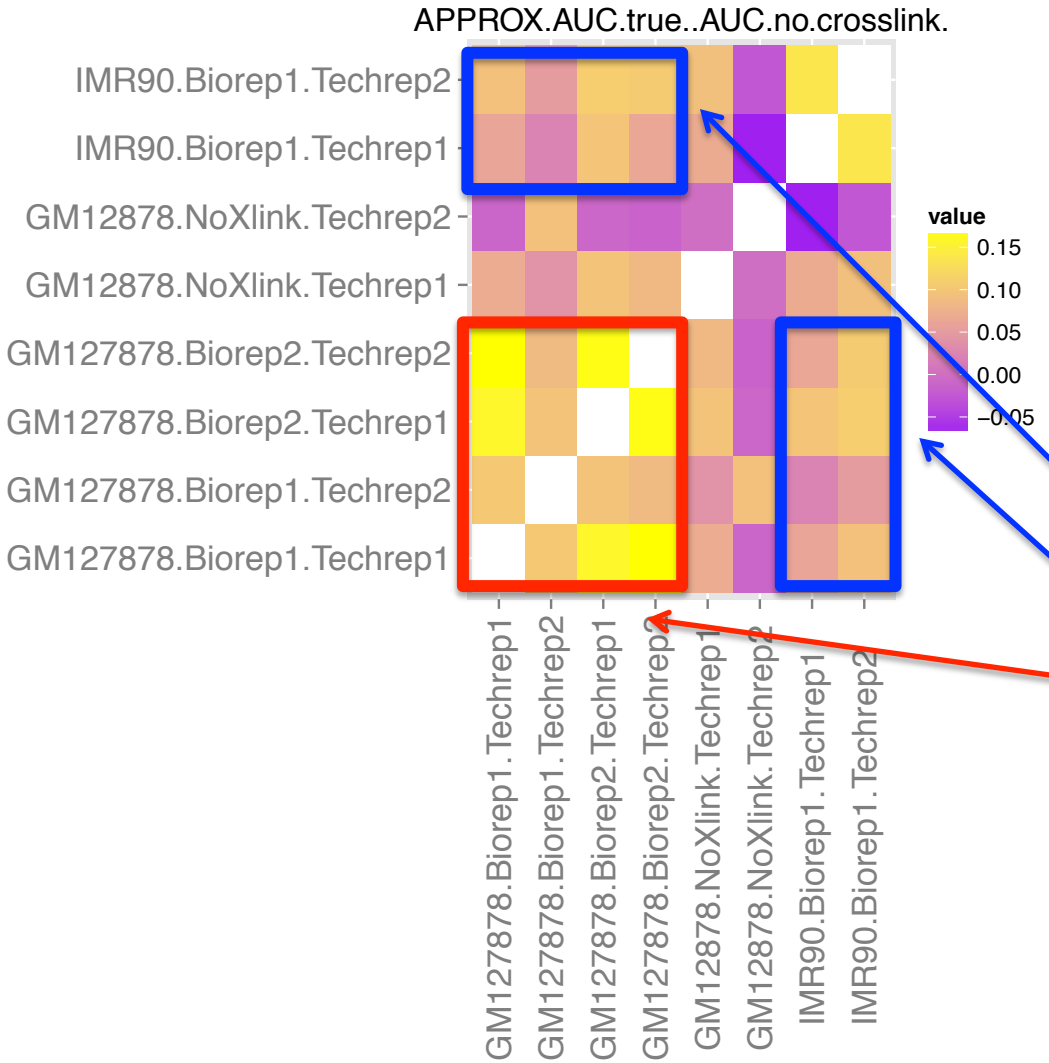
Rationale

Pseudo-replicates provide an upper bound for reproducibility

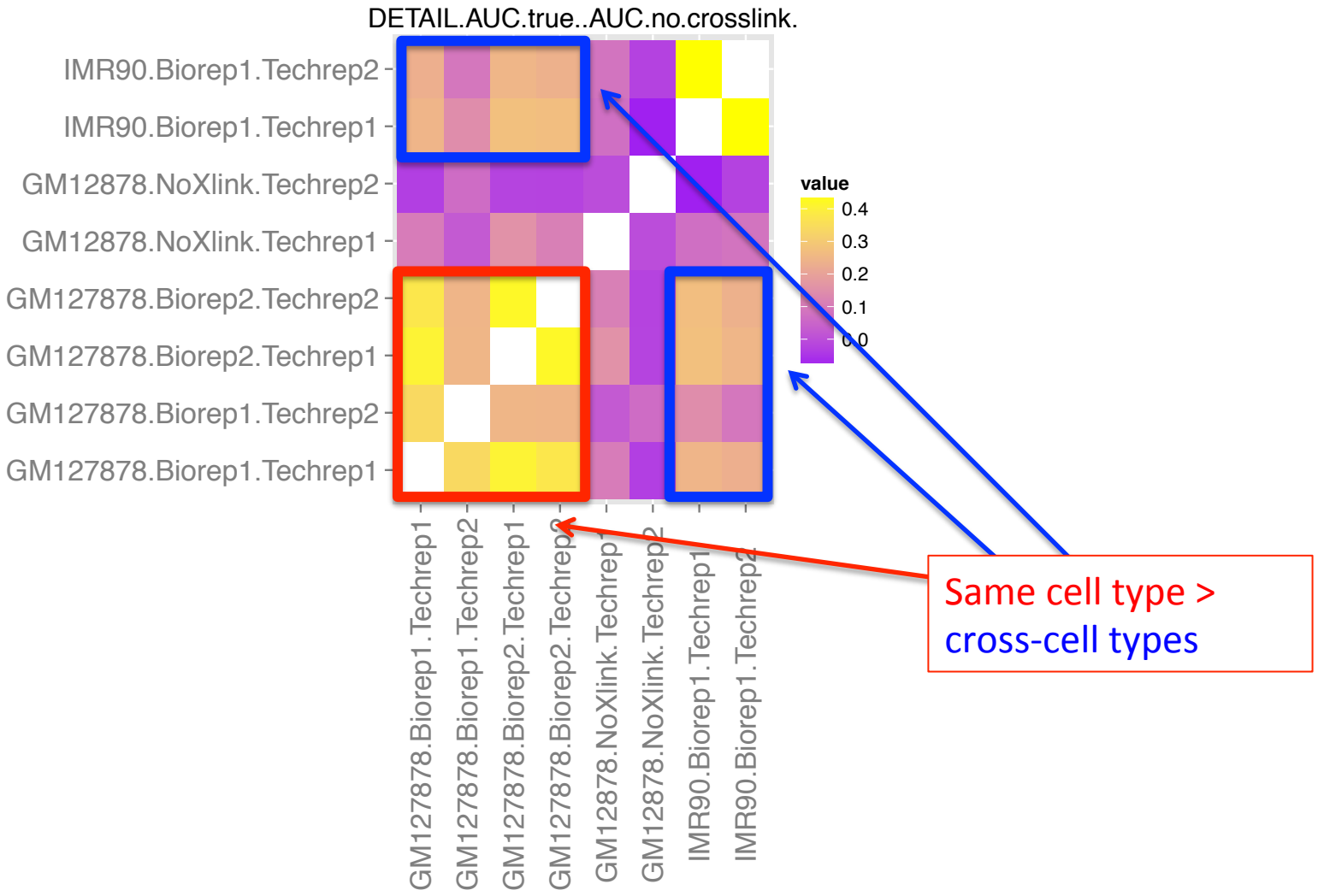
No-crosslink control is a reference for low reproducibility at a given sequencing depth

AUC[true] – AUC[no-crosslink] in 1D

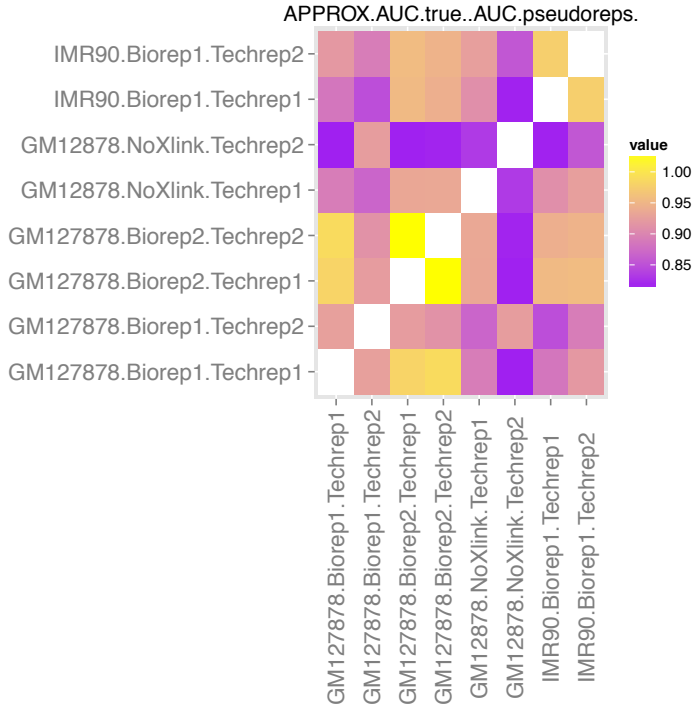
(approximation coefficients)

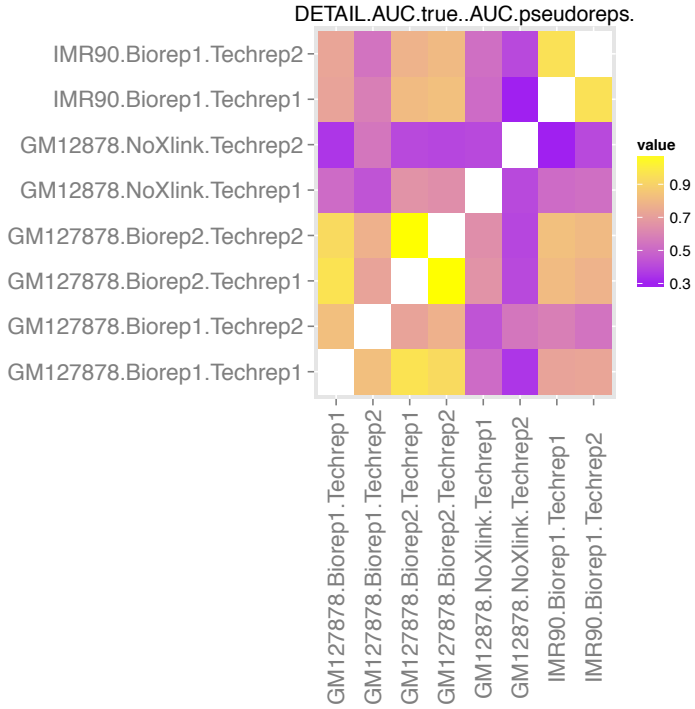


AUC[true] – AUC[no-crosslink] in 1D
(detail coefficients)

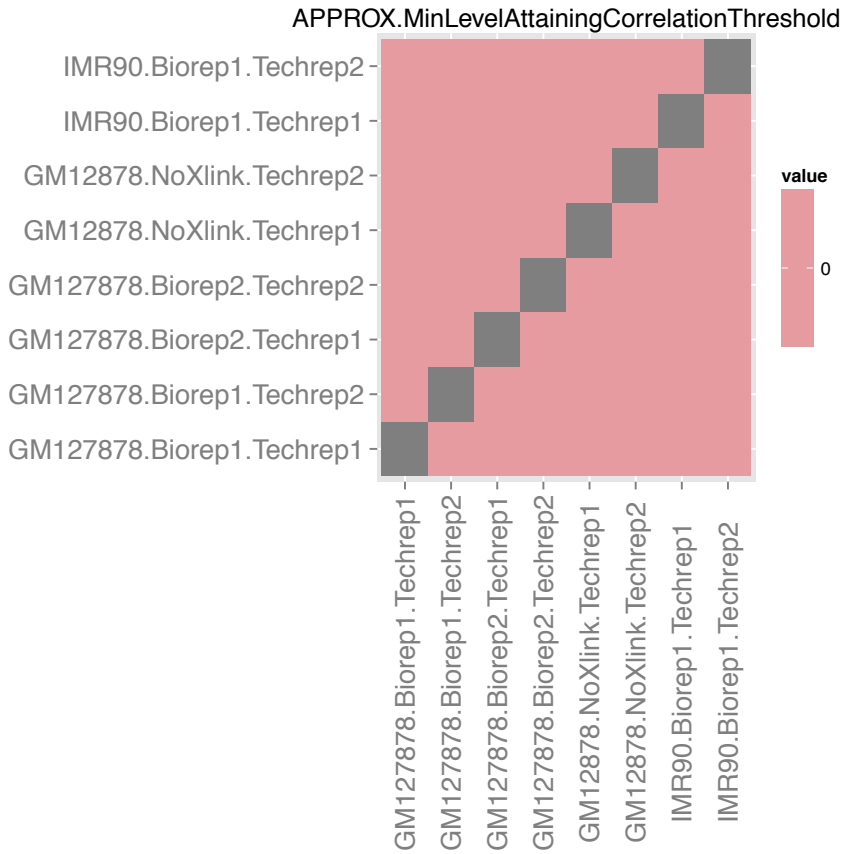


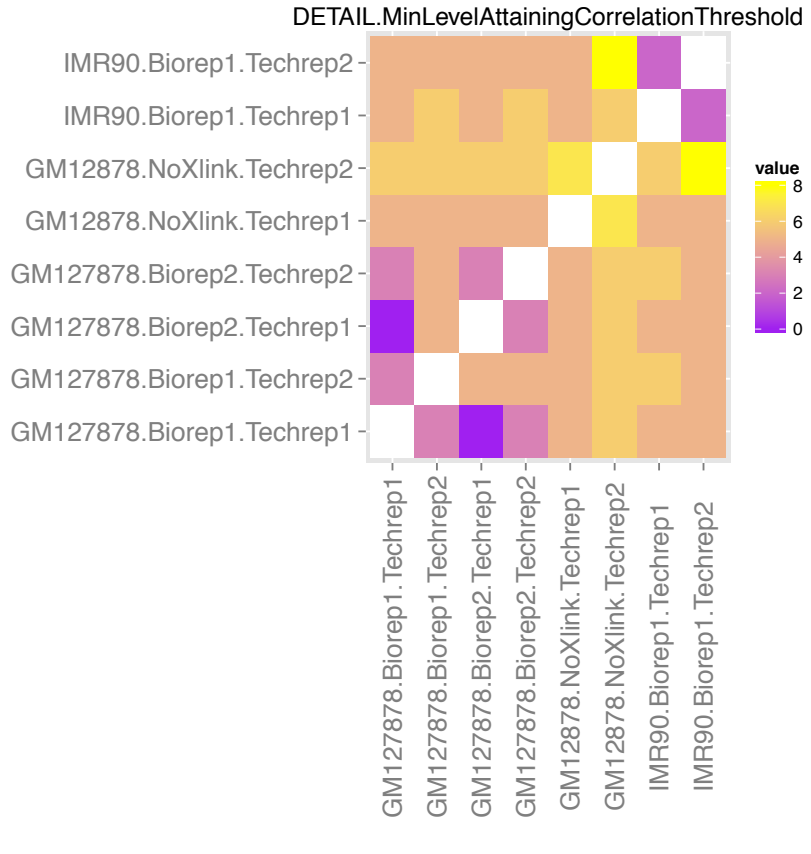
AUC[true] / AUC[pseudoreplicates] in 1D





Level at which correlation >= 0.5 in 1D





Overview

Given 2 contact maps, measure their **reproducibility**

How do we define reproducibility?

Similarity at multiple scales => Wavelet analysis

2D: compartments, domains, loops

1D: anchors

Similarity of smoothed networks => Smooth the contact maps using diffusion

Since HiC is a sparse sample of the underlying contacts, consider 2 contact maps similar even when the individual contacts they detect are not identical, as long as the overall structure is preserved

Strategy for comparing contact maps following diffusion

Contact map + graph diffusion => cleverly smoothed and de-noised contact map

An example

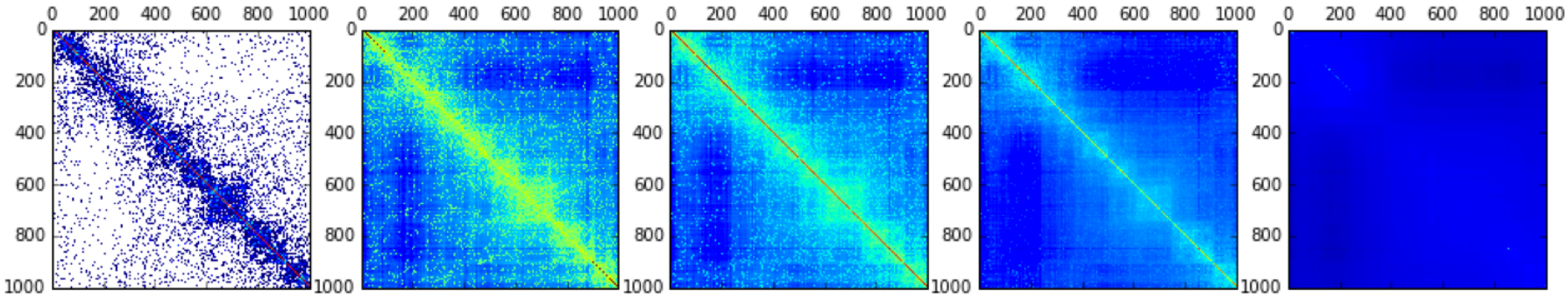
Original contact map (log10)

Diffusion Beta=0.001

Diffusion Beta=0.005

Diffusion Beta=0.01

Diffusion Beta=0.03



Repeat wavelet comparison

Work in progress

Identifying optimal resolution = level at which threshold correlation is reached

Compute reproducibility across all ENCODE datasets

Targeted wavelet analysis: e.g. 2Mb windows around the genome, to compare high resolution differences

3D organization differences between conditions using the wavelet framework

3D graph completion using diffusion kernels: Identify optimal level of diffusion by maximizing intra-TAD reads/inter-TAD reads, where TADs are defined as communities in the diffused graph

Additional reproducibility metrics based on comparing graphlet distributions

(see supplementary slides)

Thanks

Advisors
Prof. Anshul Kundaje
Prof. Michael Snyder

Kundaje lab, especially:
Nathan Boley
Maryna Taranova
Chuan-Sheng Foo
Avanti Shrikumar

Adam Rubin
Bo Wang
Rachel Wang

Snyder Lab

Questions?

oursu@stanford.edu

Supplementary slides

Comparing graphlet distributions

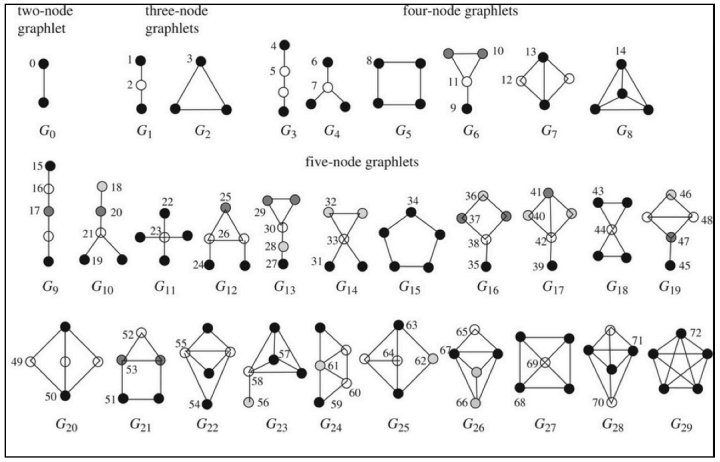


Figure. Schematic of graphlets. From https://parasol.tamu.edu/dreu2013/OLeary/ .

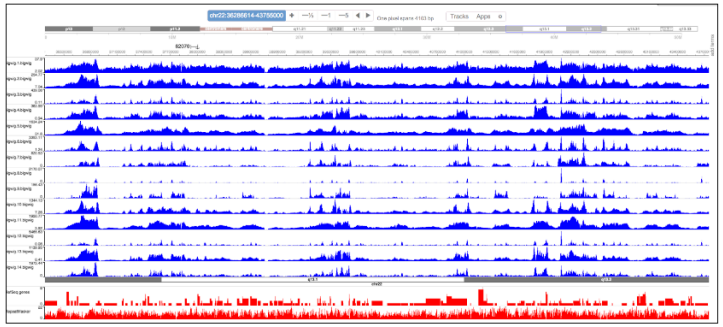


Figure. Each signal track is a graphlet node orbit. I am plotting the counts for that orbit across nodes on chr22. Note that numbers in this graph are orbit +1. e.g. orbit 0 is defined here as bigwig.1.bigwig..Json for visualizing all my tracks:
http://mitra.stanford.edu/kundaje/leepc12/web_portal_cache/1502682563.json

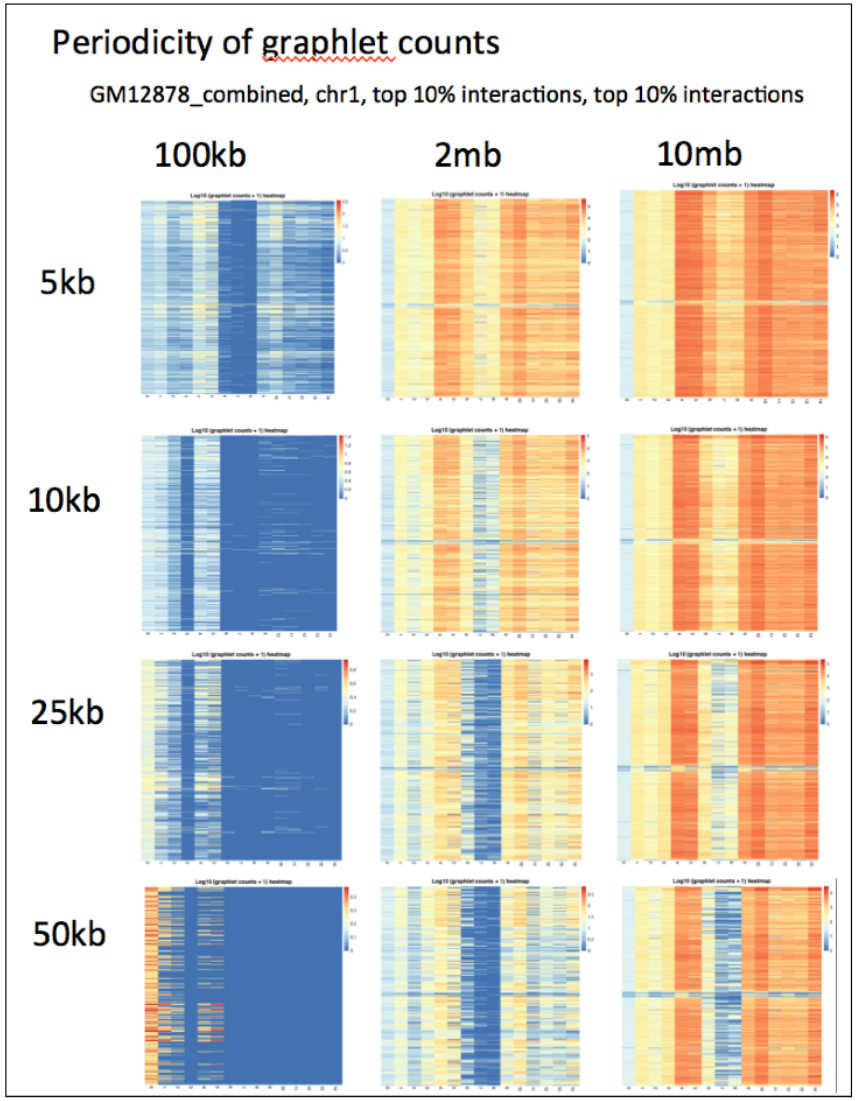


Figure. Rows = nodes (bins of resolution size), columns=graphlet orbits. Plotted is the graphlet count for each orbit across each node.

Post-ICE analysis

Reproducibility

Biological replicate 2
(Comparison of tech rep)
Similar sequencing depths

Biological replicate 1
(Comparison of tech rep)
Different sequencing depths

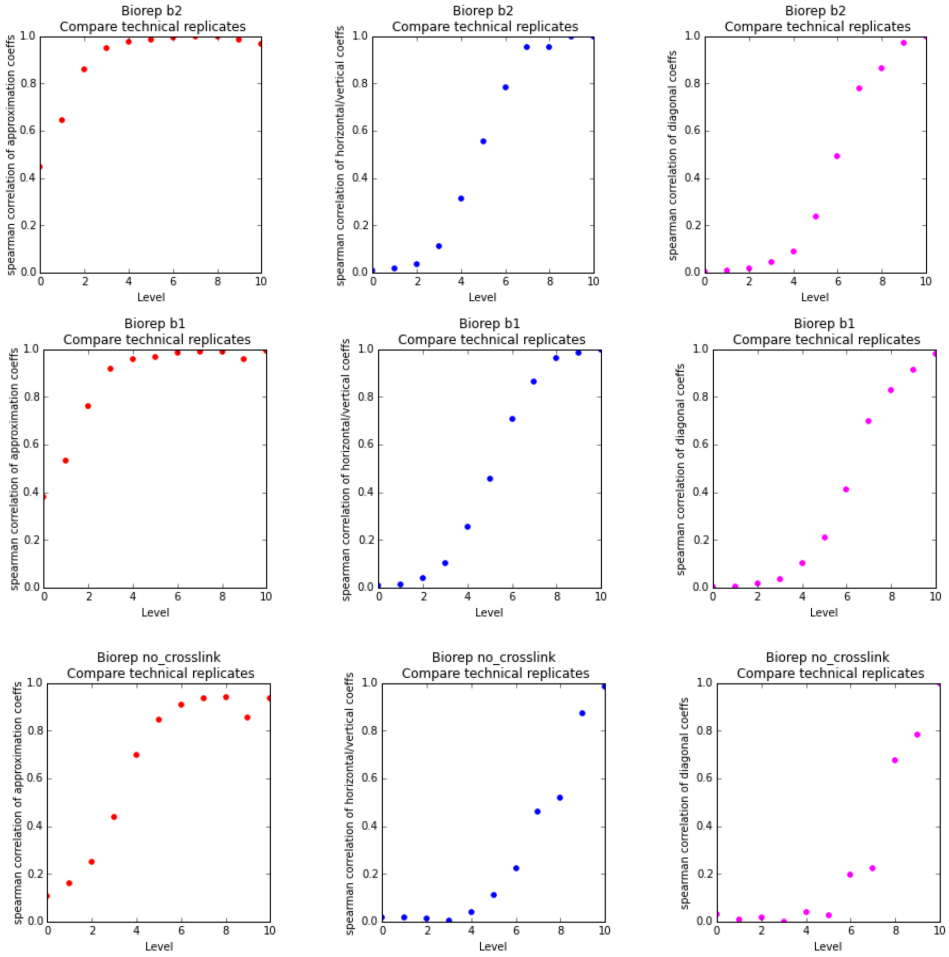
No-crosslinking control
(Comparison of tech rep)

Spearman correlation among tech replicates for:

Aproximation coefficients

Horizontal coefficients

Diagonal coefficients



Post-ICE analysis

Reproducibility

Biorep2 (t1) vs biorep 2(t2)
Technical replicates

Biorep2 (t1) vs biorep 1(t2)
Biological replicates,
similar sequencing depth

Biorep2 (t1) vs no-crosslink
control

Spearman correlation among tech replicates for:

Aproximation coefficients	Horizontal coefficients	Diagonal coefficients