An Overview of HFFc6 and ESC Hi-C Data

Experiments

Two Tier-1 Cell Lines

Data is publicly available at the DCIC portal.

HFFc6: Human Foreskin Fibroblasts

https://data.4dnucleome.org/experiment-set-replicates/4DNES2R6PUEK/

ESC: Embryonic Stem Cells

https://data.4dnucleome.org/experiment-set-replicates/4DNES2M5JIGV/

Software

Mapping: Distiller *https://github.com/mirnylab/distiller-nf*

Format: Cooler *https://github.com/mirnylab/cooler*

Downstream Analysis: Cooltools *https://github.com/mirnylab/cooltools*

Visualization: HiGlass *https://github.com/hms-dbmi/higlass*

All other track files and code can be found at the Github repository:

https://github.com/dekkerlab/4dn_jawg

Overview

(billions)	**Raw Reads**	**Interactions**	**Cis**	**Trans**
HFFc6-combined	6.47	3.08	2.16	0.91
HFFc6-R1	3.27	1.59	1.08	0.50
HFFc6-R2	3.19	1.49	1.08	0.41
H1-ESC-combined	6.35	2.66	1.31	1.35
H1-ESC-R1	3.16	1.39	0.69	0.70
H1-ESC-R2	3.19	1.27	0.62	0.65

Compartments

Compartments were called using

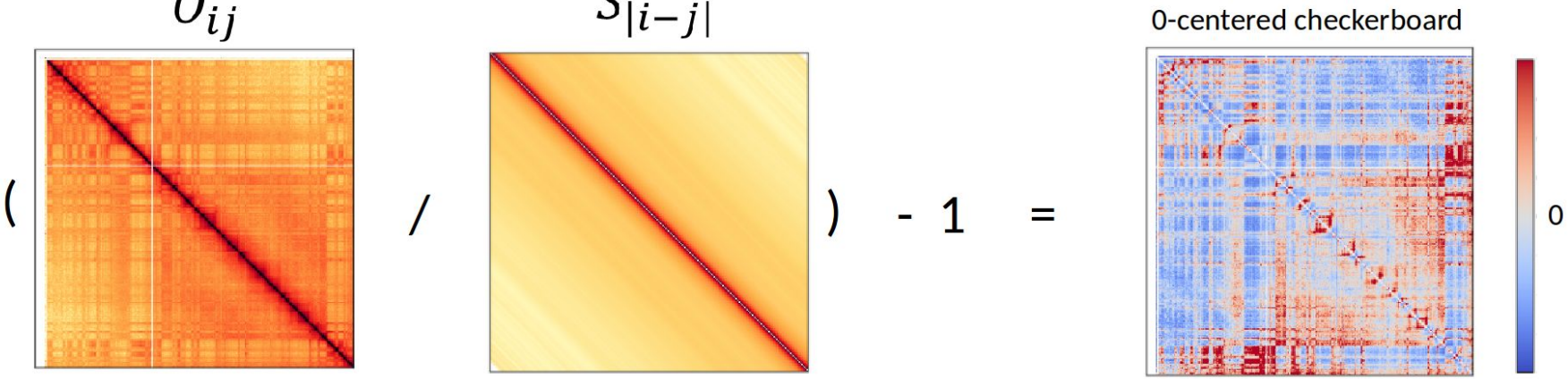
```
cooltools call_compartments
```

at 200 kb resolution, using eigenvector decomposition.

The eigenvectors were oriented based on gene density.

Eigenvector Decomposition

O_{ij} $S_{|i-j|}$



Def. of eigenvector decomposition says:

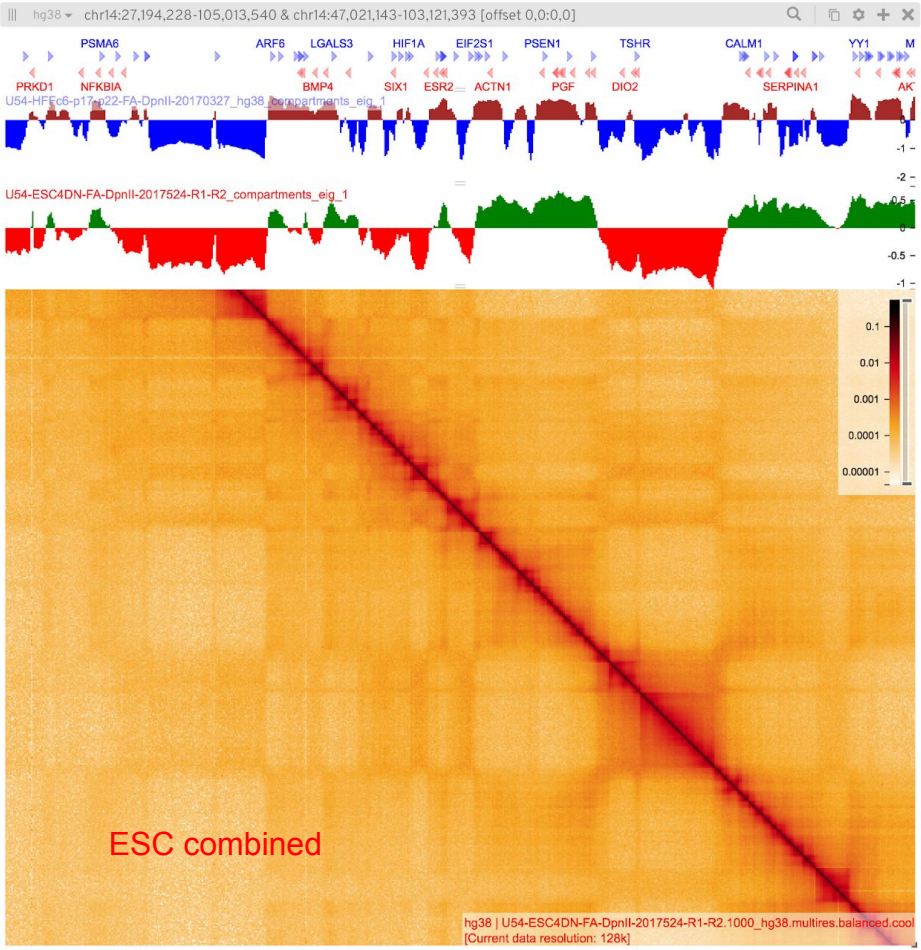
$A = Q\Lambda Q^{-1}$, where columns of Q are unit eigenvectors e_1, e_2, e_3,... ($\lambda_1 > \lambda_2 > \lambda_3 \ldots$)

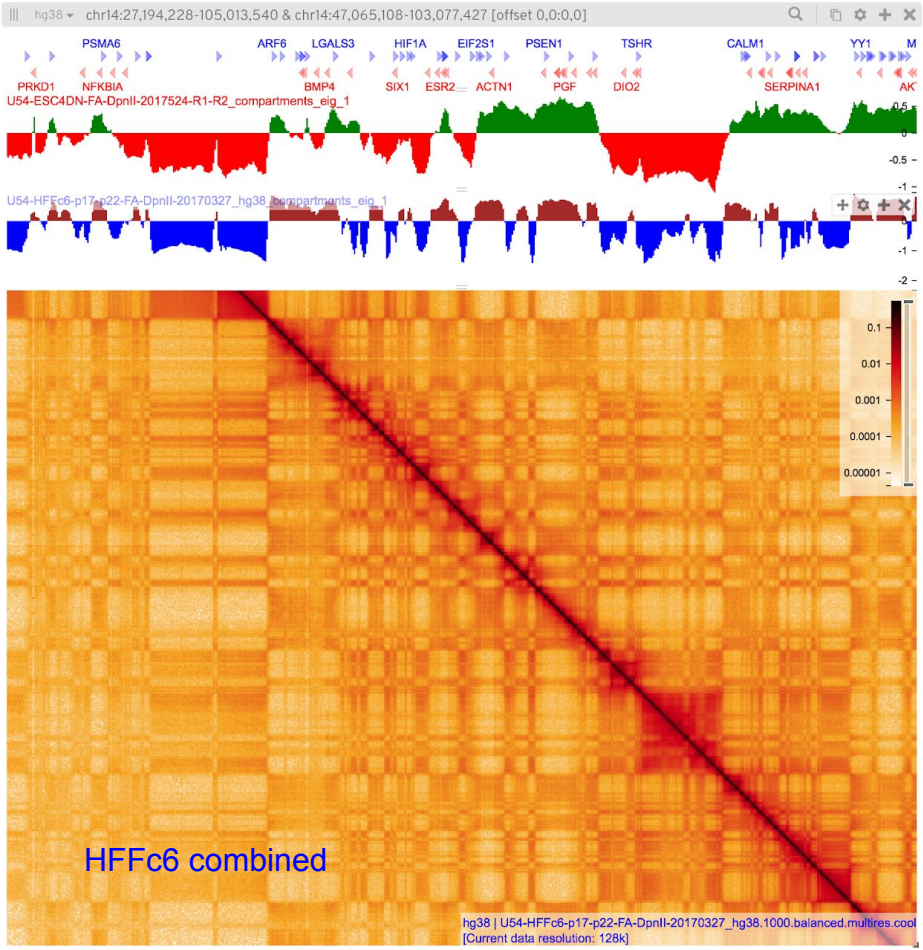
$A = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + \ldots$

$A \approx \lambda_1 e_1 e_1^T$ rank-1 approximation

Courtesy of Nezar Abdennur

Checkerboard ~ **Outer product** of eigenvector $\mathbf{e_1}$


hg38
chr14:27,194,228-105,013,540 & chr14:47,021,143-103,121,393 [offset 0,0:0,0]
PSMA6
ARF6
LGALS3
HIF1A
EIF2S1
PSEN1
TSHR
CALM1
YY1
M
PRKD1
NFKBIA
BMP4
SIX1
ESR2
ACTN1
PGF
DIO2
SERPINA1
AK
U54-HFFc6-p17-p22-FA-DpnII-20170327_hg38_compartments_eig_1
U54-ESC4DN-FA-DpnII-2017524-R1-R2_compartments_eig_1
0.1
0.01
0.001
0.0001
0.00001
ESC combined
hg38 | U54-ESC4DN-FA-DpnII-2017524-R1-R2.1000_hg38.multires.balanced.cool
[Current data resolution: 128k]


hg38
chr14:27,194,228-105,013,540 & chr14:47,065,108-103,077,427 [offset 0,0:0,0]
PSMA6
ARF6
LGALS3
HIF1A
EIF2S1
PSEN1
TSHR
CALM1
YY1
M
PRKD1
NFKBIA
BMP4
SIX1
ESR2
ACTN1
PGF
DIO2
SERPINA1
AK
U54-ESC4DN-FA-DpnII-2017524-R1-R2_compartments_eig_1
U54-HFFc6-p17-p22-FA-DpnII-20170327_hg38_compartments_eig_1
0.1
0.01
0.001
0.0001
0.00001
HFFc6 combined
hg38 | U54-HFFc6-p17-p22-FA-DpnII-20170327_hg38.1000.balanced.multires.cool
[Current data resolution: 128k]

The First Eigenvector Captures the Compartments

You can browse the compartment tracks on HiGlass:

http://higlass.dekkerlab.org/app/?config=DoSmbAetQjScbU0wuCkoYg

The tracks (of eigenvectors) are available at:

https://github.com/dekkerlab/4dn_jawg/tree/master/August_2_2018/compartments

Correlation of Compartments

	HFFc6-R1	HFFc6-R2	ESC-R1	ESC-R2
HFFc6-R1	1	0.95 (0.98)	0.69 (0.54)	0.69 (0.54)
HFFc6-R2		1	0.71 (0.56)	0.71 (0.56)
ESC-R1			1	0.98 (0.99)
ESC-R2				1

Black: We are looking at compartment type assignment

Two libraries have the same compartment / # Bins

Blue: Pearson correlation of eigenvectors

Insulation

We computed the insulation using two different models. In both cases, we used 10kb contact maps.

Diamond Insulation: Average interactions in a fixed size window around the diagonal as described in [Crane et. al, Nature, 2015.]

In our case, the window size is 100kb and the contact map has 10kb resolution.

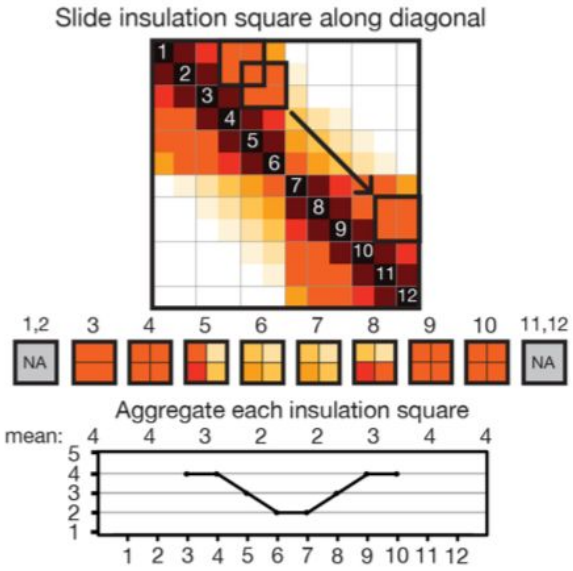


figure from Crane et. al., Nature 2015

Insulation

Diamond insulation score was computed using cooltools

```
cooltools diamond_insulation
```

The insulation tracks are available at

https://github.com/dekkerlab/4dn_jawg/tree/master/August_2_2018/insulation/diamond-insulation

Insulation Potential

Developed and implemented by Noam Kaplan.

$$
\begin{aligned}
P_{\text{observed}}(i,j) &= \text{Observed contact frequency} \\
P_{\text{distance}}(i,j) &= \text{Interaction probabilty given by genomic distance} \\
c_k &= \text{Insulation potential}
\end{aligned}
$$

Assume $i < j+1$; $0 \leq c_k \leq 1$. Then,

$$P_{\text{observed}}(i,j) = \frac{1}{Z} P_{\text{distance}}(i,j) \prod_{k=i+1}^{j-1} c_k$$

$c_k = 0 \rightarrow$ complete insulation,
$c_k = 1 \rightarrow$ No insulation

This model gives us the insulation potential for each bin.

Noam Kaplan
Technion

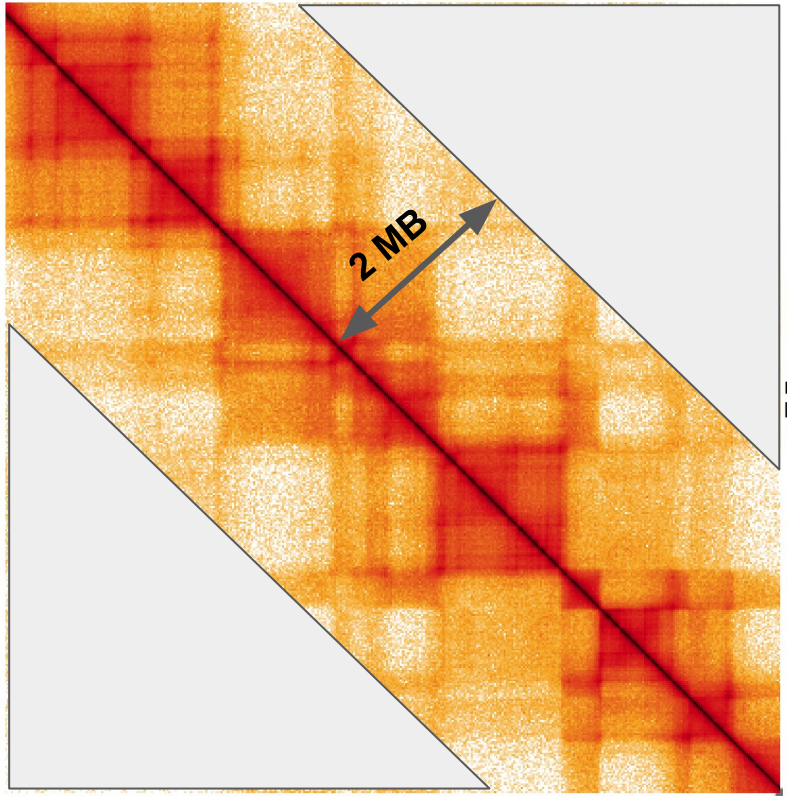
https://kaplanlab.technion.ac.il/

Insulation Potential

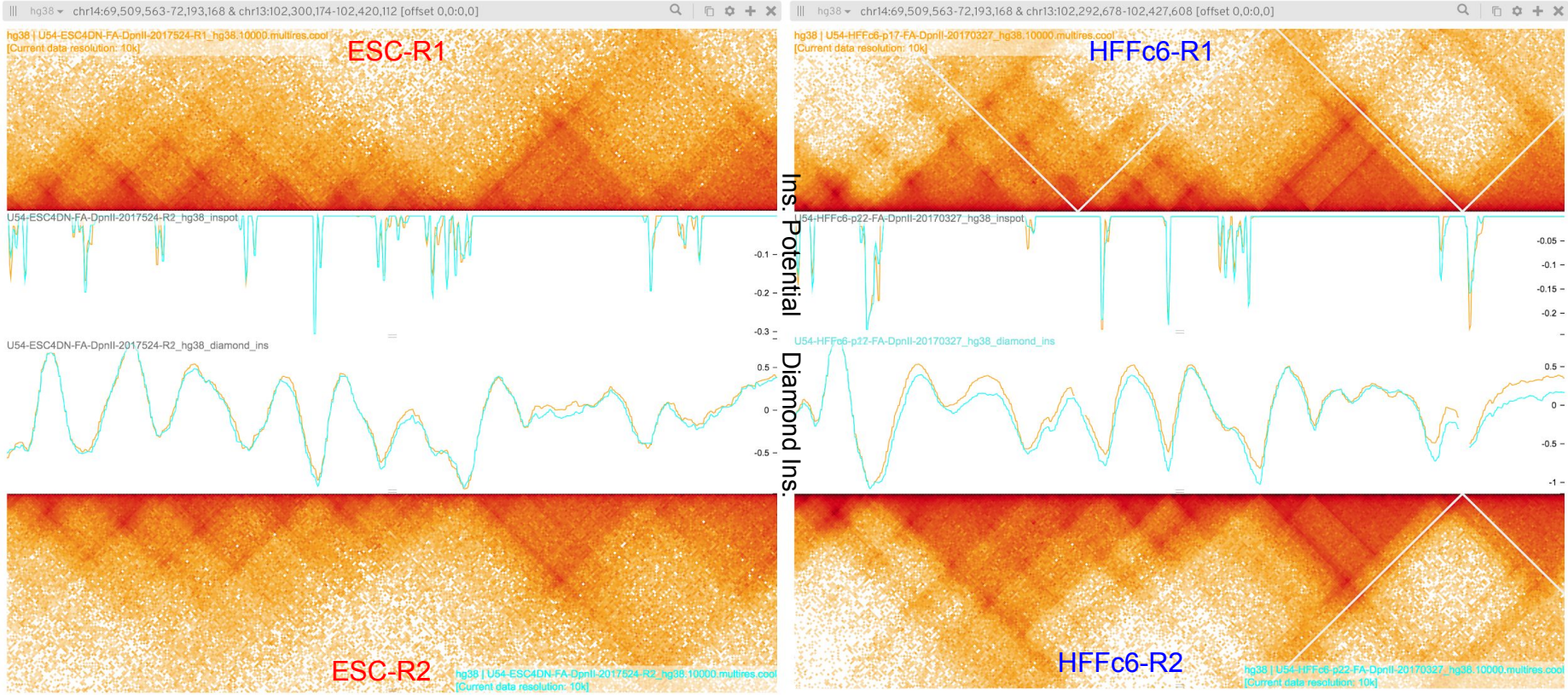
Heuristically, we consider 2 MB around the diagonal to compute the insulation potential. The insulation potential is computed from the balanced 10 KB contact maps.

The insulation potential tracks are at:

https://github.com/dekkerlab/4dn_jawg/tree/master/August_2_2018/insulation/insulation-potential



Insulation of Biological Replicates



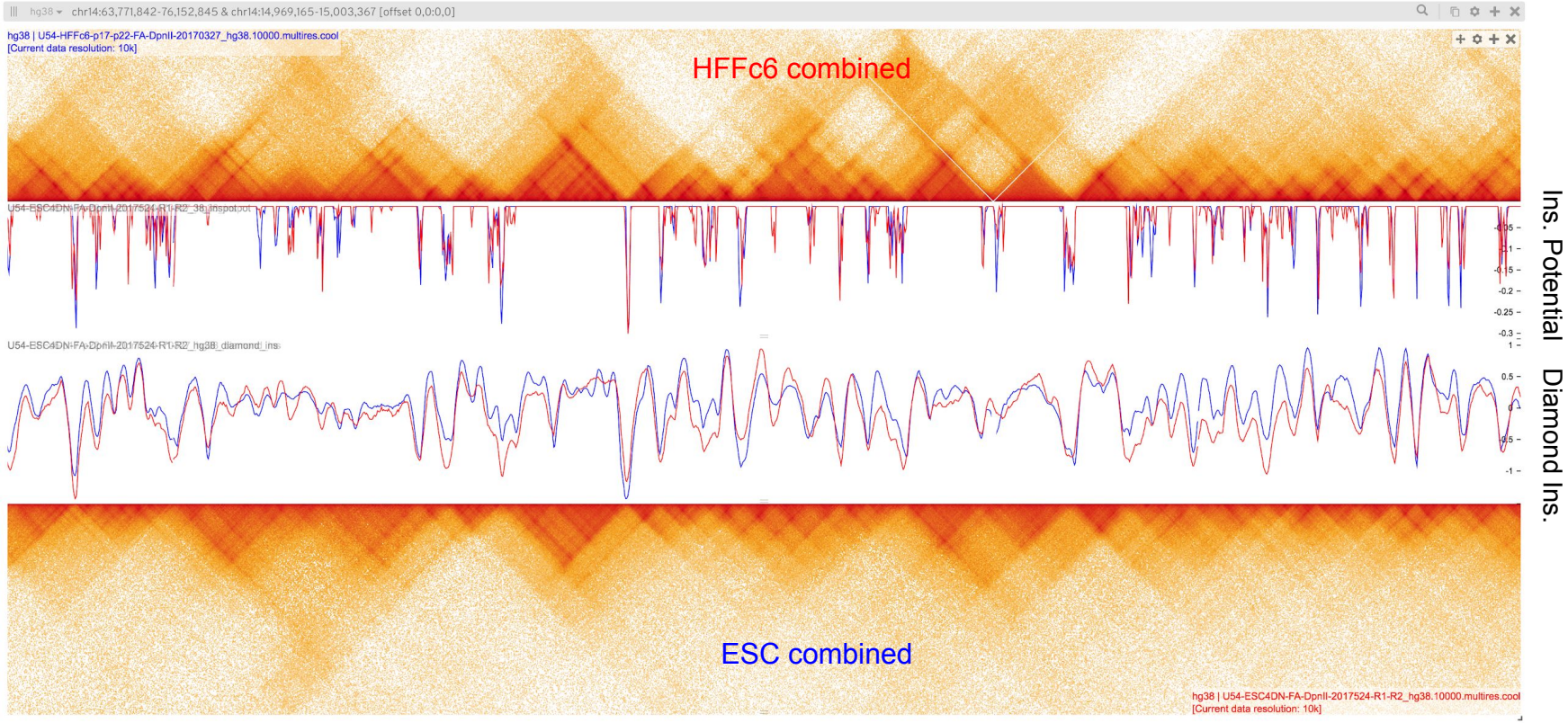
http://higlass.dekkerlab.org/app/?config=EaCbq-YZQPuBhPpIme7pKg

Pearson Correlation of Insulation Scores

	HFFc6-R1	HFFc6-R2	ESC-R1	ESC-R2
HFFc6-R1	1	0.99 (0.93)	0.95 (0.77)	0.95 (0.77)
HFFc6-R2		1	0.95 (0.68)	0.94 (0.68)
ESC-R1			1	0.98 (0.98)
ESC-R2				1

Black: Insulation Potential
Blue: Diamond Insulation

Insulation: ESC vs HFFc6



http://higlass.dekkerlab.org/app/?config=BcQ0JWxPTQeaSMJotRimgw

Available on HiGlass

Insulation Profiles of Biological Replicates:

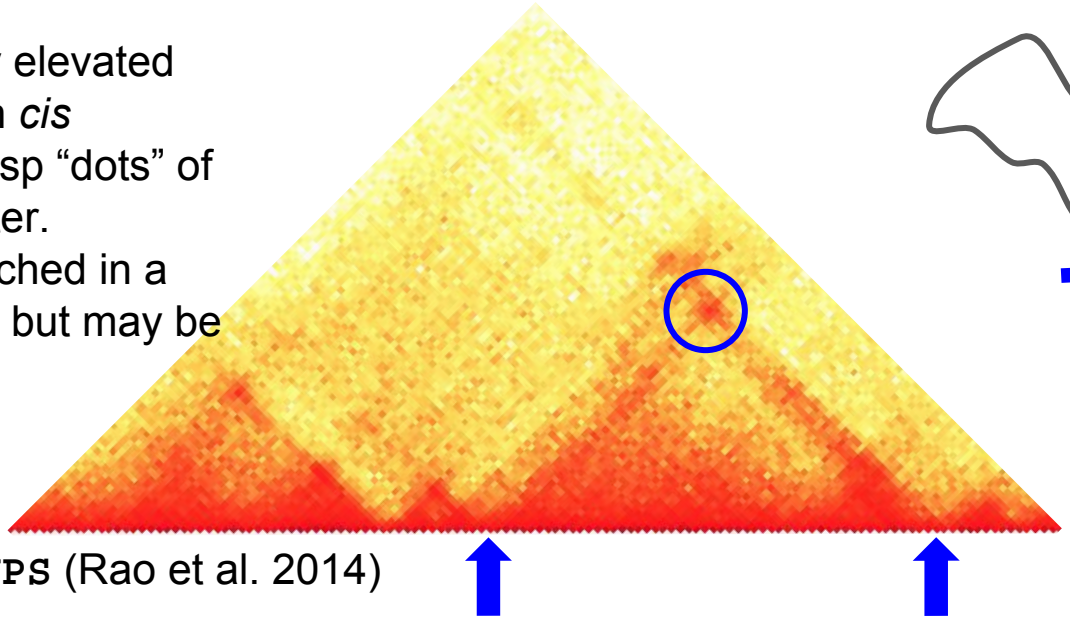
http://higlass.dekkerlab.org/app/?config=EaCbq-YZQPuBhPpIme7pKg

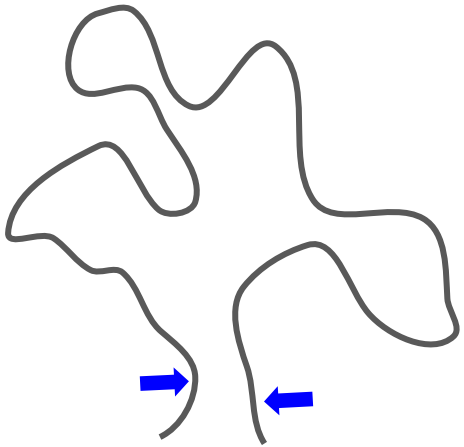
Insulation Profiles of Combined Replicates:

http://higlass.dekkerlab.org/app/?config=fVfm6kgOSN-1TIhFBblplw

Dot Calling

- searching for locally elevated contact frequency in *cis*
- manifest itself as crisp “dots” of ~20-30 kb in diameter.
- subset of loops enriched in a **population** of cells, but may be highly dynamic



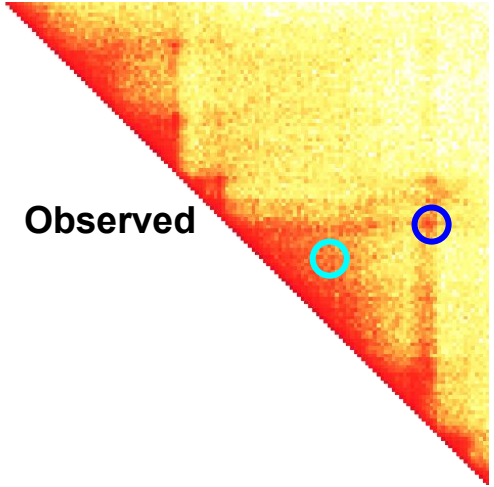


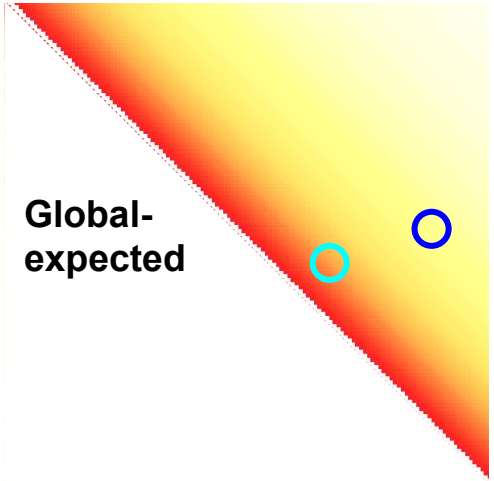
- existing tool: `HiCCUPS` (Rao et al. 2014)
- Need a new tool:
  - to work with the `.cool` files
  - works on a CPU only
  - embrace Python ecosystem
  - enable parameter-tweaking for a wider audience

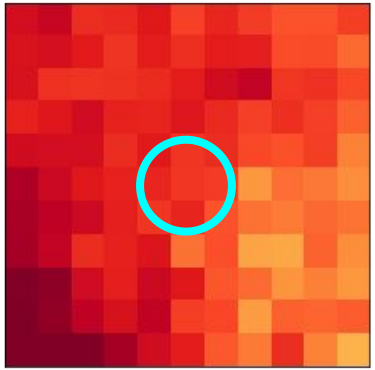
https://www.youtube.com/watch?v=8FW6gOx5lPI&index=1&list=PLnV-JMzdy0ZPP7_W7xJEyqYpu75Yvq7fj

Global-expected is not sufficient

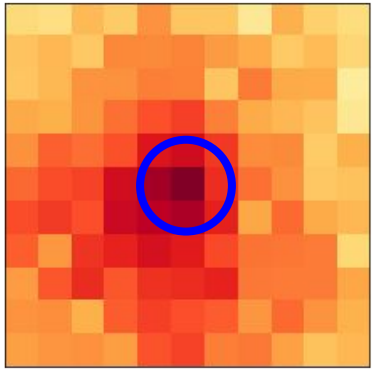
Pixels enriched over the global expected (distance decay) emphasize dots, TAD, stripes/flames, etc.

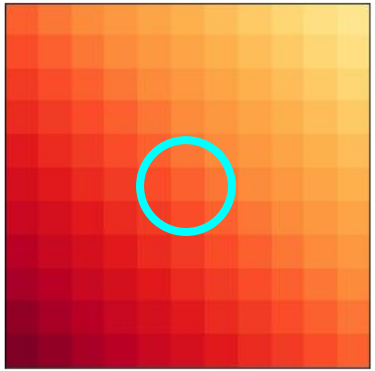




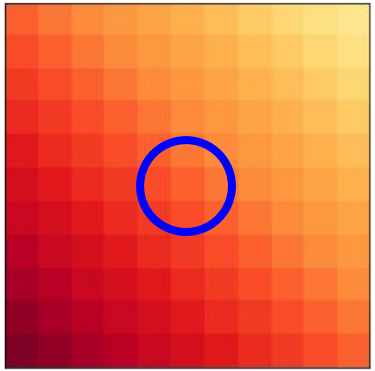


Observed



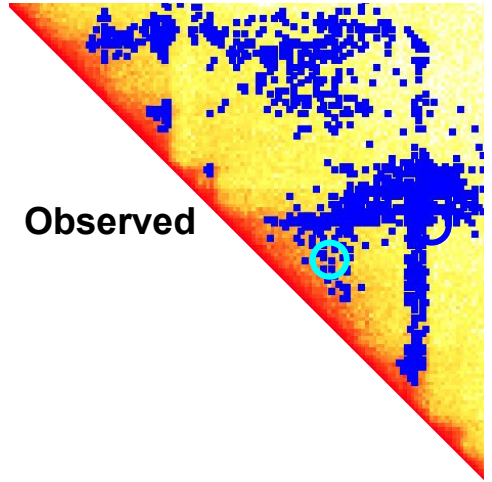


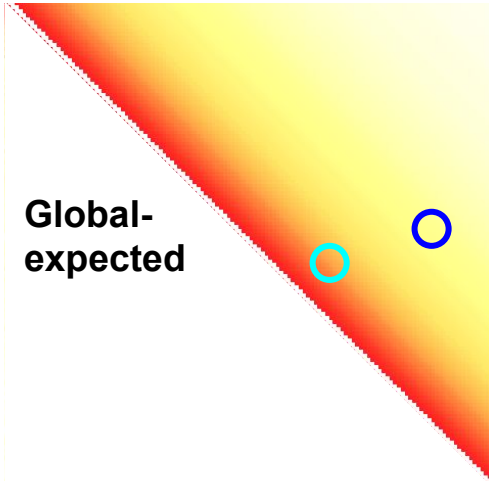
Global-
expected

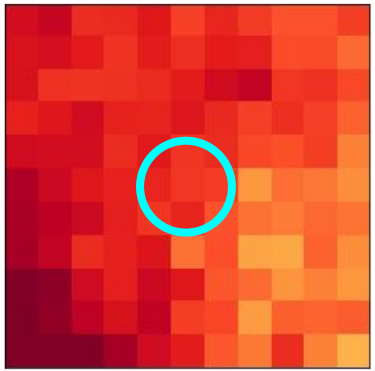


Global-expected is not sufficient

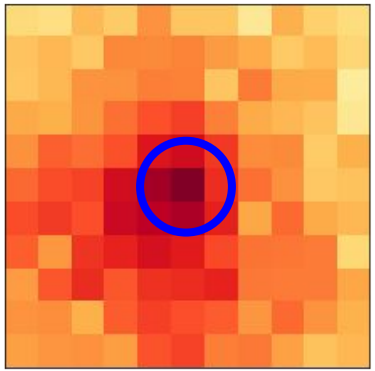
Pixels enriched over the global expected (distance decay) emphasize dots, TAD, stripes/flames, etc.

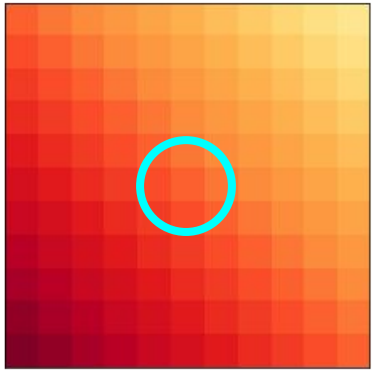




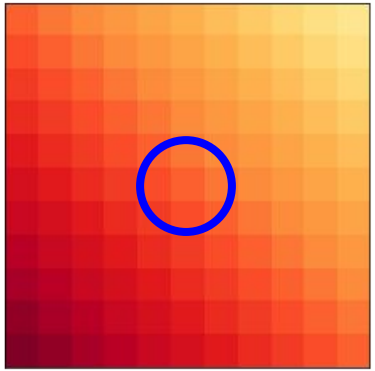


Observed



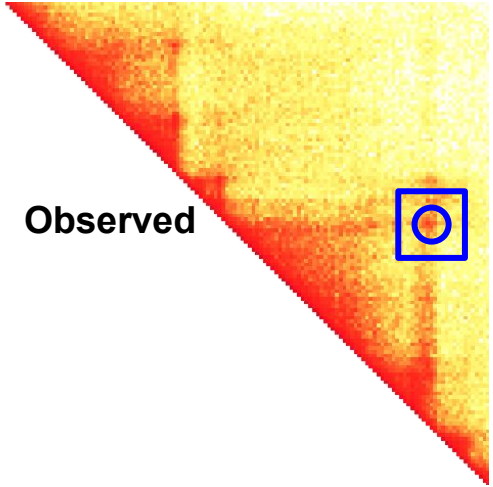


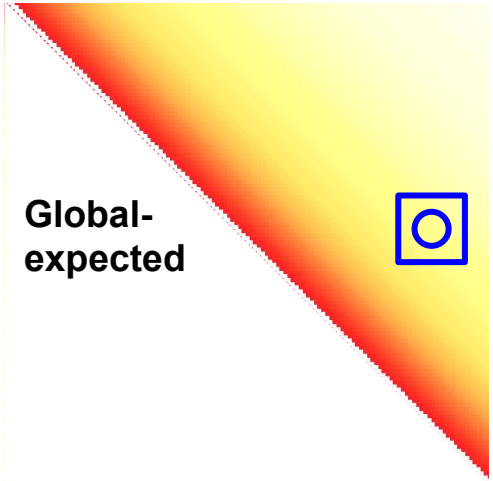
Global-
expected

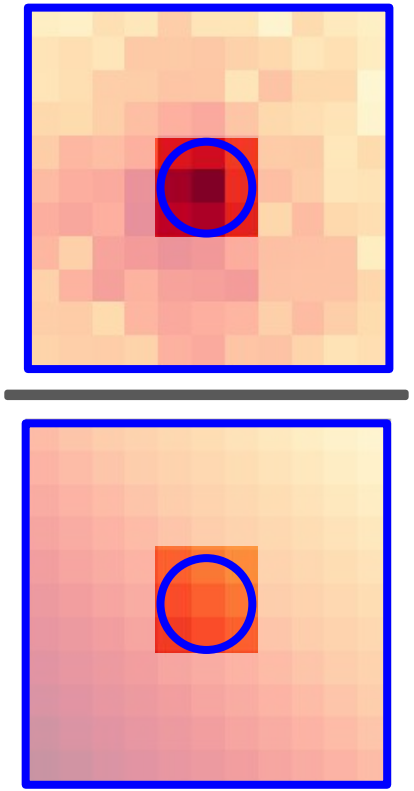


HiCCUPS uses local enrichment (Rao et al 2014)

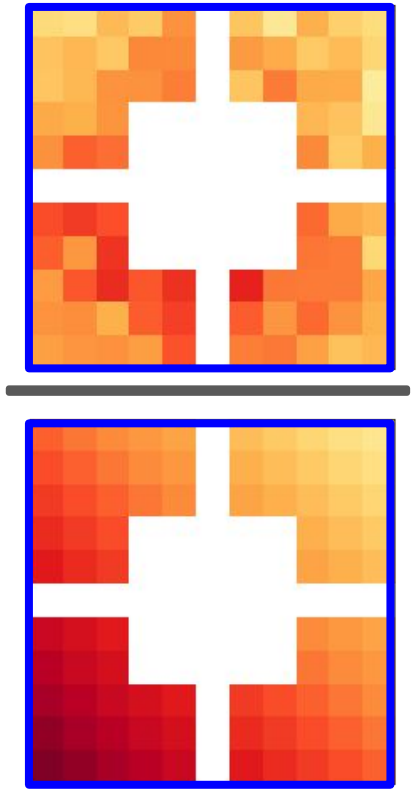
An enrichment of pixels surrounding a “dot” must be much smaller than the enrichment of the “dot” itself.





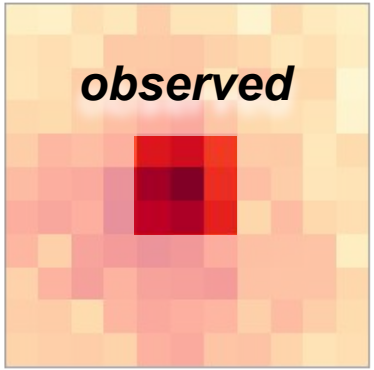


VS



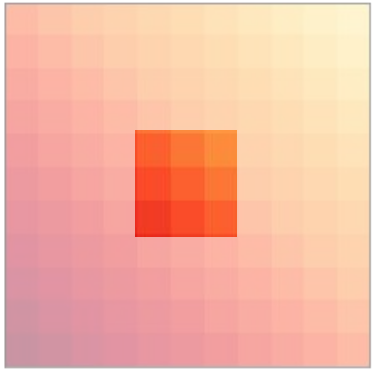
Reimplementing `HiCCUPS` (Rao et al 2014)

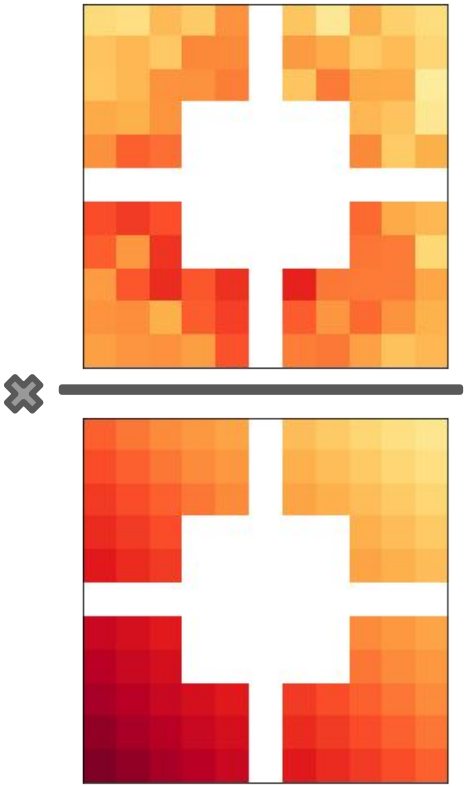
- `HiCCUPS` compares an ***observed*** with ***locally adjusted expected***, using Poisson distribution (after unbalancing)
- Hypothesis are grouped by their loc. adjusted expected (λ-chunking) and significantly enriched pixels are determined using BH-FDR[(1)]
- Subsequent pixel clustering and enrichment filtering yields the final list of called peaks.



VS

locally - adjusted expected =



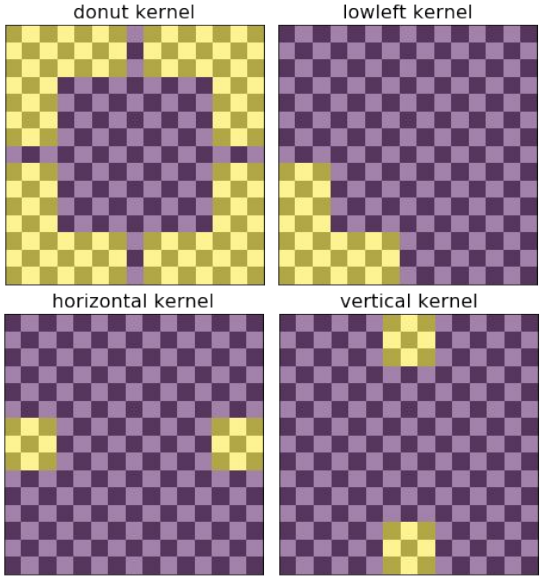


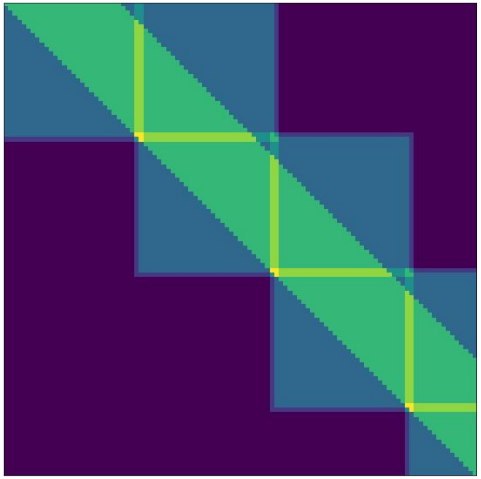
(1) http://multithreaded.stitchfix.com/blog/2015/10/15/multiple-hypothesis-testing/

- All of this steps have been reimplemented in `cooltools` already

Some details of the implementation

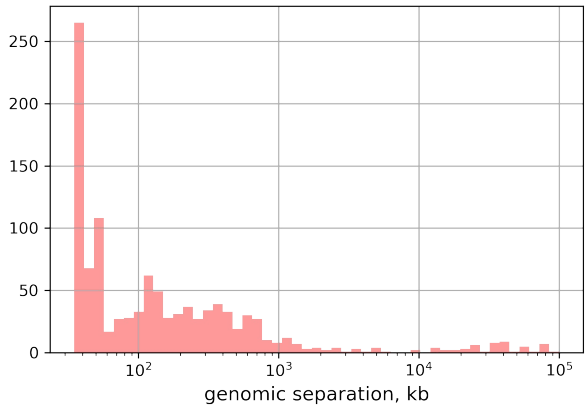
- Defining local surrounding of a pixel
  - **`HiCCUPS`** is using 4 to emphasize dots:
- **`HiCCUPS`** is using dynamic filter resizing
  - “shirking” near diagonal
  - enlarging in noisy areas (hi-res, distant dots)
  - not implemented in **`cooltools`** yet
- **`cooltools`** reimplementation of **`HiCCUPS`** relies heavily on Python ecosystem (pandas, numpy, scipy.ndimage. etc.)
- uses **`.cool`** as an input
- Does not require specialized hardware, aka GPUs, using matrix tiling to parallelize on a modern multi-core CPUs.
- Explores arbitrary region around the diagonal - presented data used 10mb cutoff.
- Takes ~30 minutes to process GM12878 (Rao et al 2014) dataset at 5kb and 10kb resolutions and combine the results using 12 CPU cores.
- **`cooltools call_dots --help`**
  - https://github.com/dekkerlab/cooltools/tree/new-dotfinder

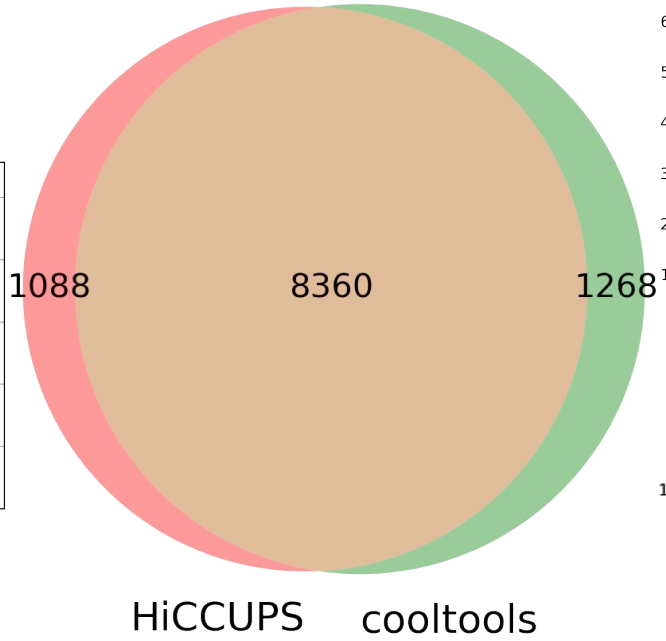


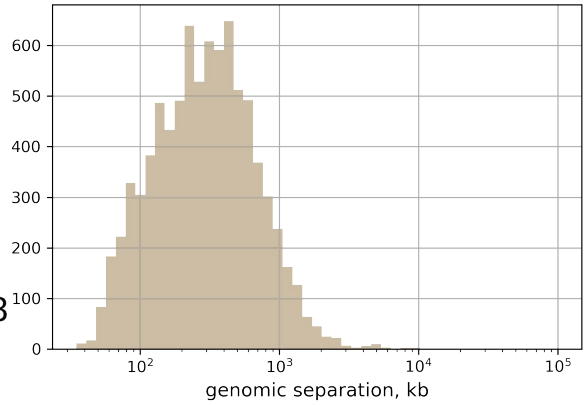


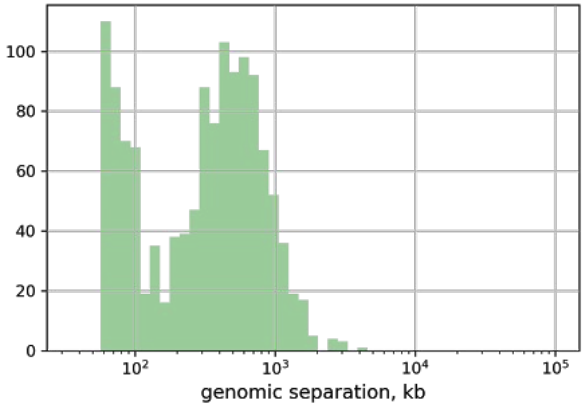
`HiCCUPS` (Rao et al 2014)

Method validation using GM12878 data (Rao et al 2014)





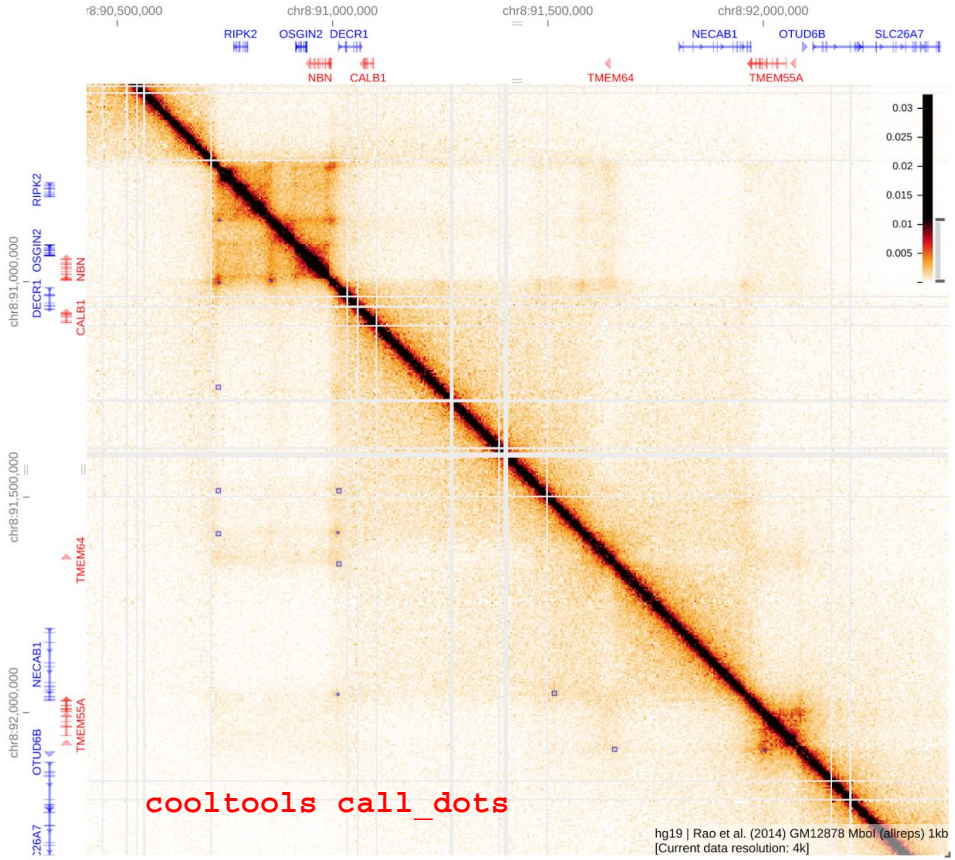


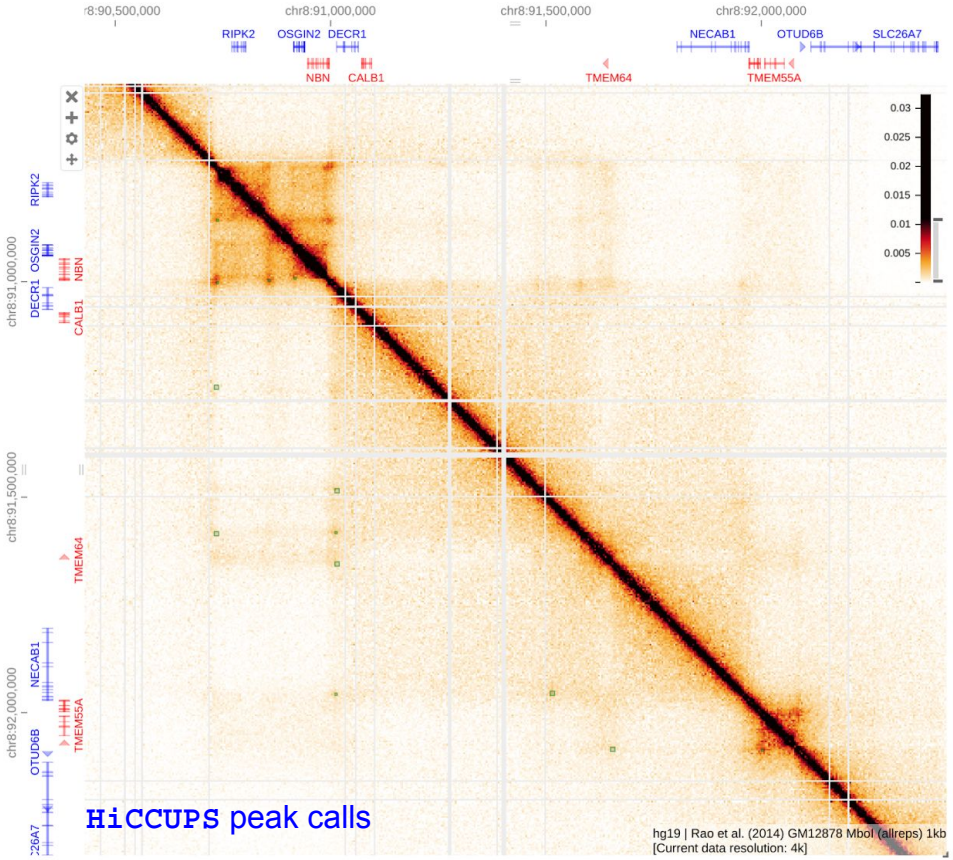


- 5kb and 10kb calls combined.
- 10mb cutoff for peak calls
- 15kb cutoff for the overlap.

http://higlass.dekkerlab.org/app/?config=T_euGvJVS_29YhcriWTULg

Heatmap sneak-peak, use HiGlass web-link below to browse interactively

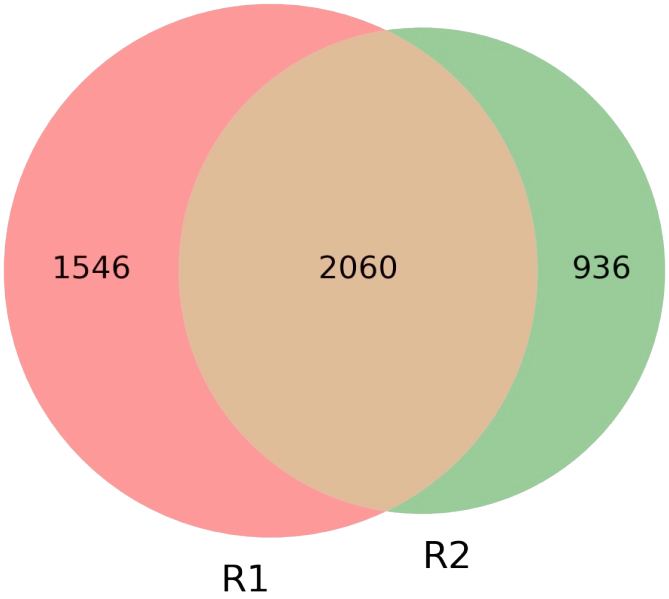


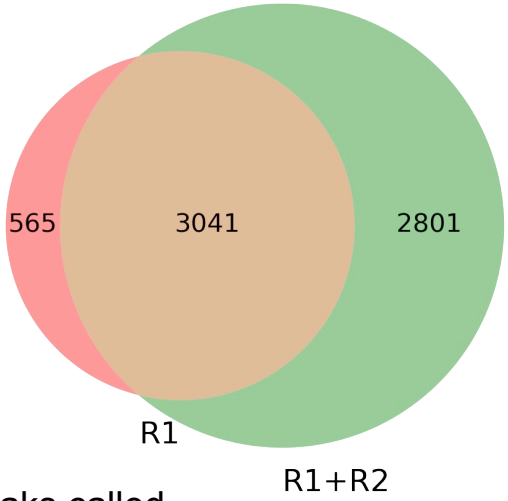


http://higlass.dekkerlab.org/app/?config=T_euGvJVS_29YhcriWTULg

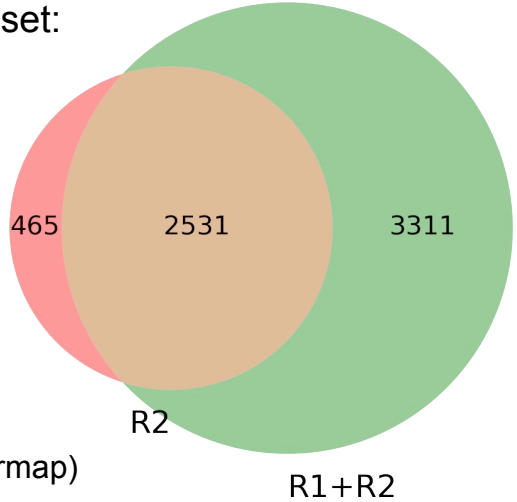
Dot Calling: ESC

Reproducibility: biological replicates R1 and R2





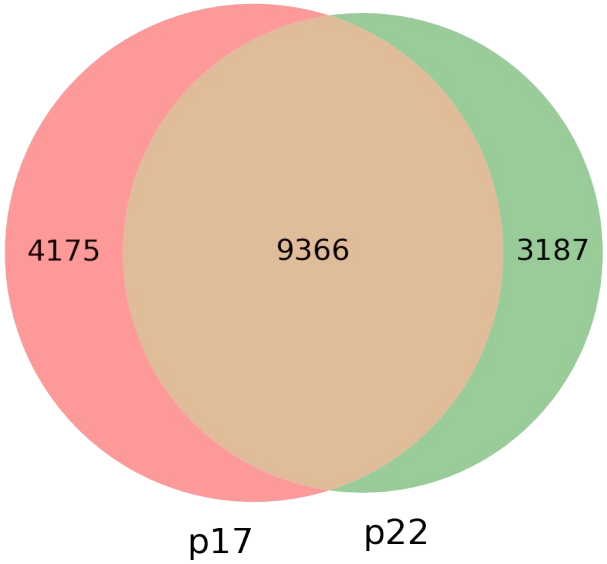
R1+R2 - list of peaks called on combined dataset: 5842 total

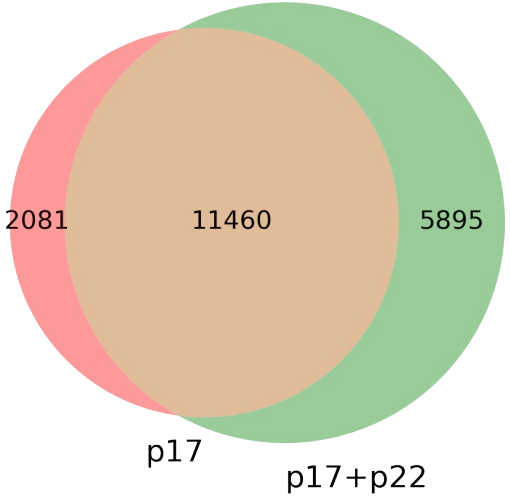


http://higlass.dekkerlab.org/app/?config=P9V4pIBnSLq1Po0v28n_JA
http://higlass.dekkerlab.org/app/?config=DBdzNRNiRASY2N1ImWgB-g (adjusted colormap)

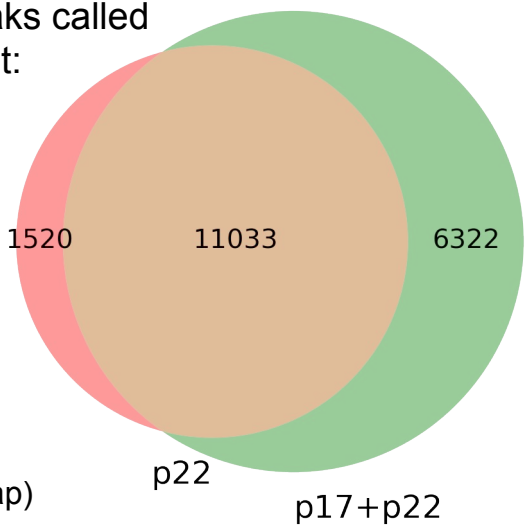
Dot Calling: HFFc6

Reproducibility: biological replicates p17 and p22



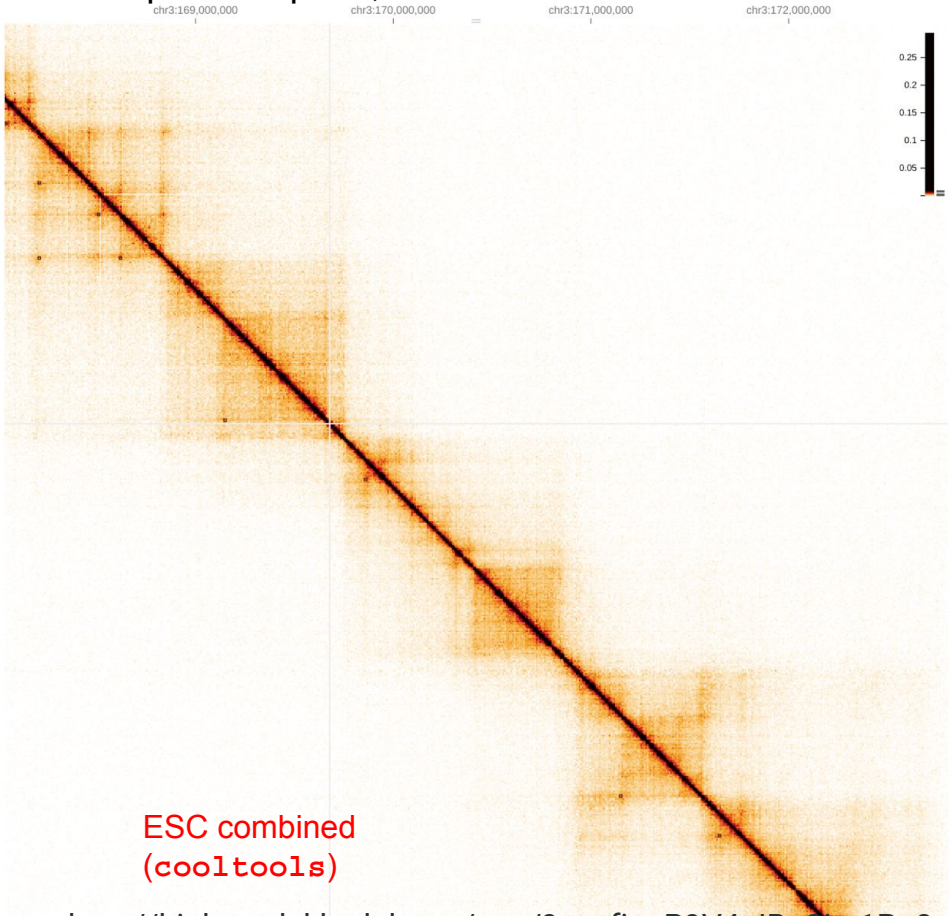


p17+p22 - list of peaks called on combined dataset: 17355 total

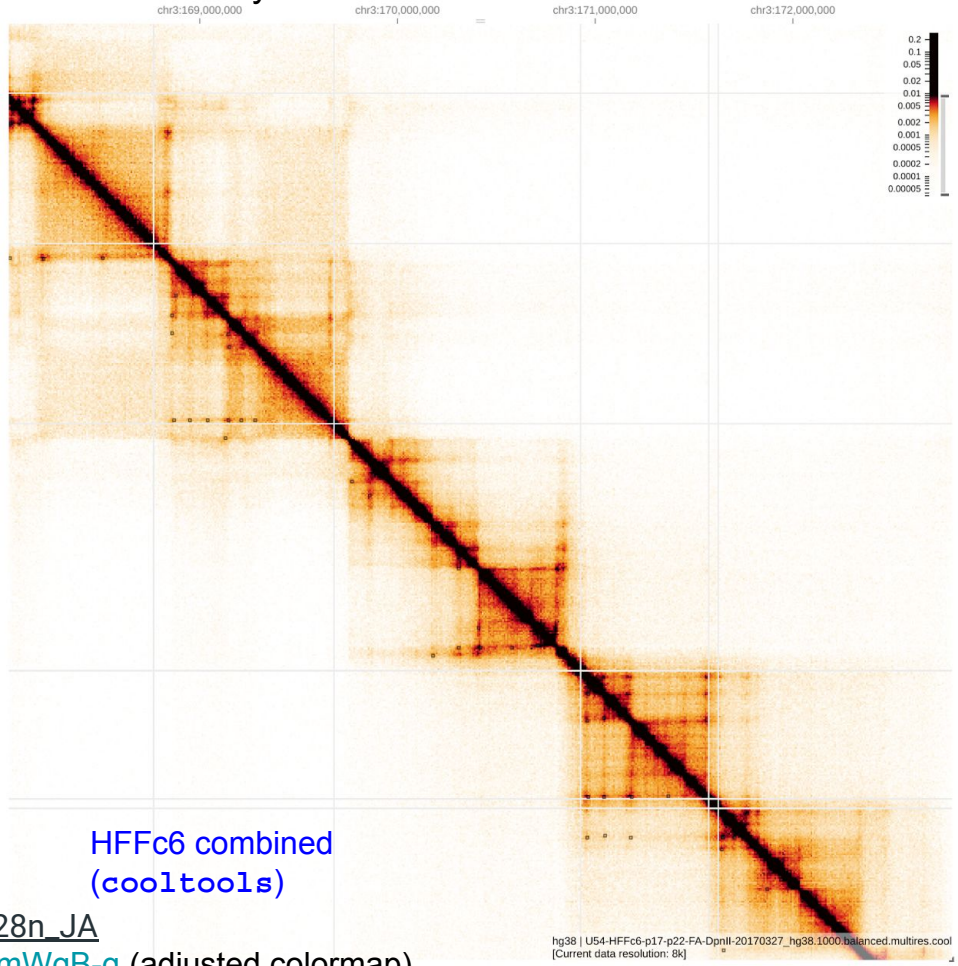


http://higlass.dekkerlab.org/app/?config=P9V4pIBnSLq1Po0v28n_JA
http://higlass.dekkerlab.org/app/?config=DBdzNRNiRASY2N1ImWgB-g (adjusted colormap)

Heatmap sneak-peak, use HiGlass web-link below to browse interactively



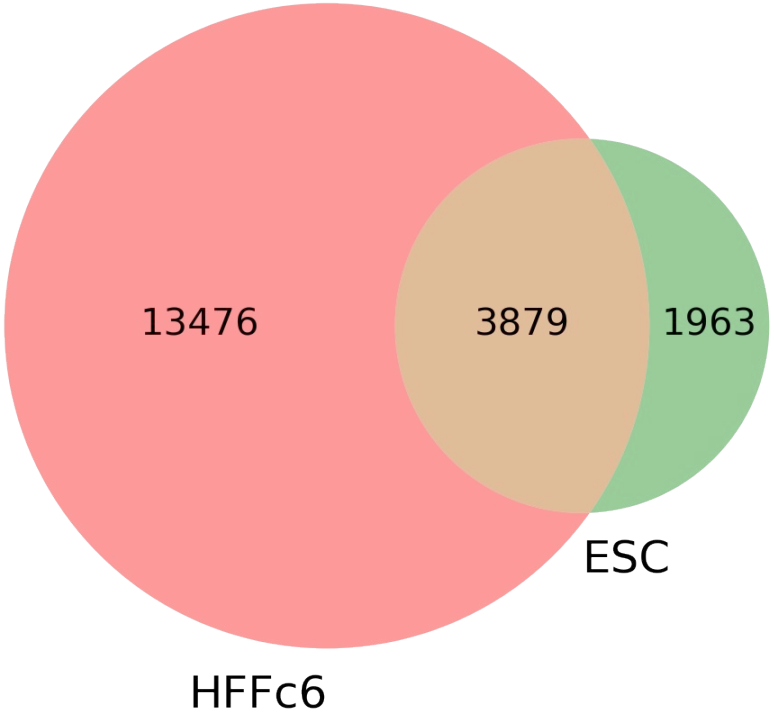
ESC combined
(cooltools)



HFFc6 combined
(cooltools)

http://higlass.dekkerlab.org/app/?config=P9V4plBnSLg1Po0v28n_JA
http://higlass.dekkerlab.org/app/?config=DBdzNRNiRASY2N1ImWgB-g (adjusted colormap)

Dot Calling: HFFc6 vs ESC



http://higlass.dekkerlab.org/app/?config=P9V4pIBnSLq1Po0v28n_JA
http://higlass.dekkerlab.org/app/?config=DBdzNRNiRASY2N1ImWgB-g (adjusted colormap)

Summary

In HFFc6 and ESC Hi-C experiments, now we have publicly available

1) Compartment tracks in the form of eigenvectors
2) Diamond Insulation and Insulation Potential Tracks
3) Dot Calls

These tracks can be downloaded from https://github.com/dekkerlab/4dn_jawg

These tracks can also be visualized in our HiGlass browser http://higlass.dekkerlab.org/

Raw sequencing data, metadata and contact maps of these experiments are available at the 4DN DCIC data portal.

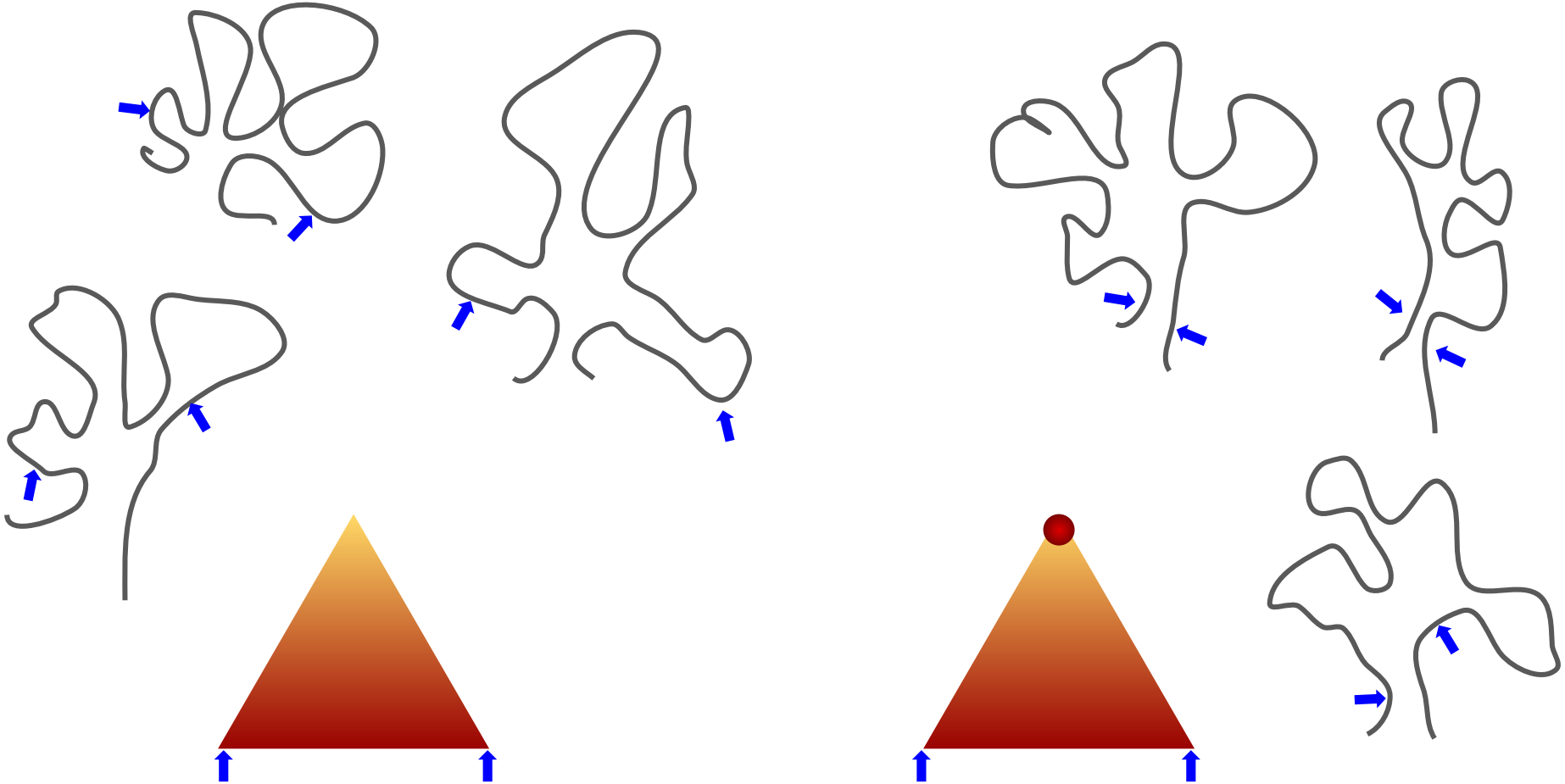
Future Directions

- Comparison of Diamond Insulation and Insulation Potential
- Reproducibility metrics, (for example HiCRep)
- Integrating other data: CTCF track, Chip-Seq or Repli-Seq Tracks
- Dot Calling: address dynamic filter-size, intra vs GW balancing effects, etc.
- Suggestions ?

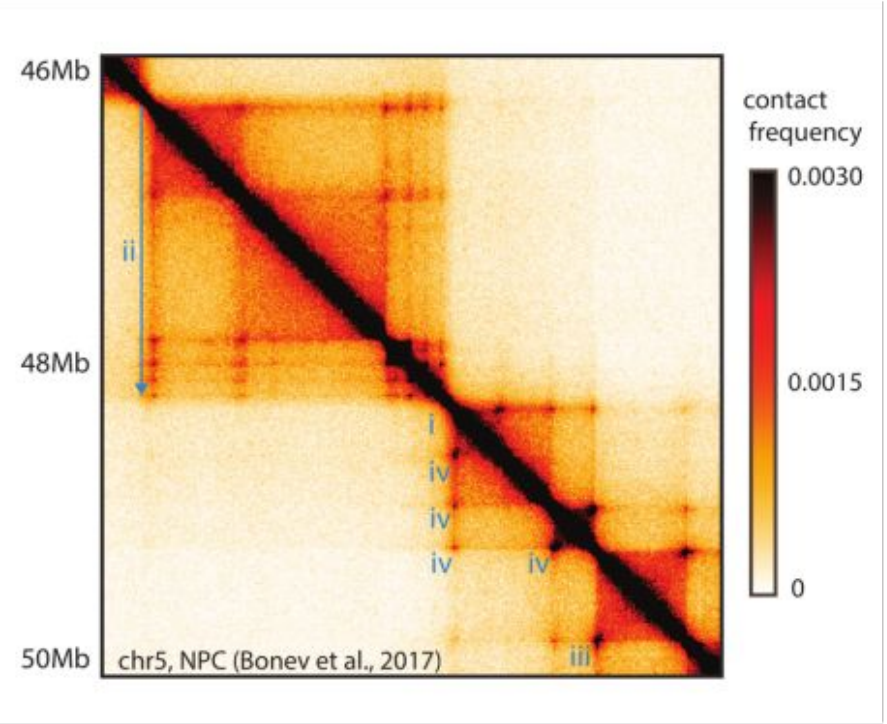
Thank you!

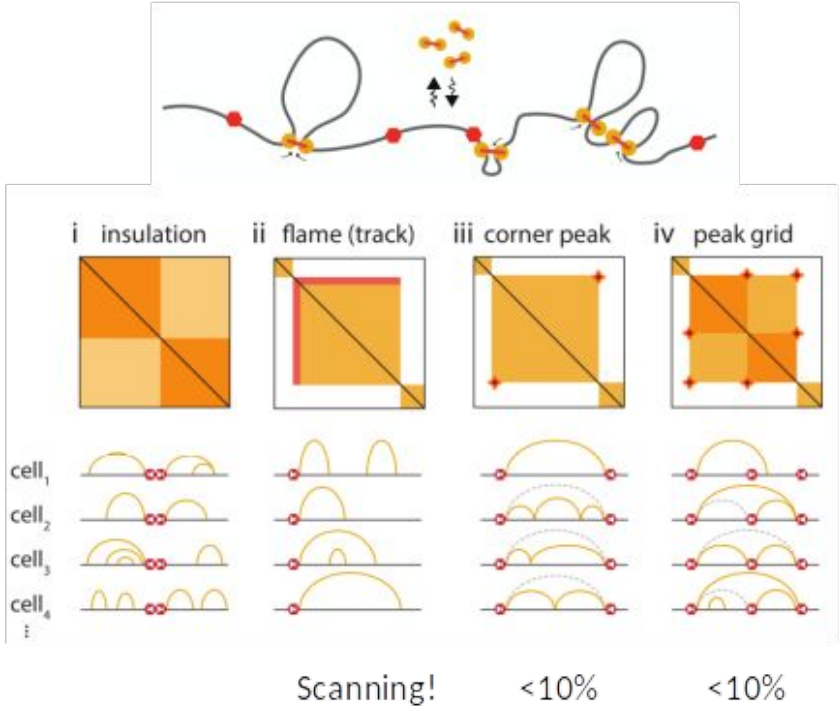
Backup slides

Dots (looping interactions) vs a loop ...



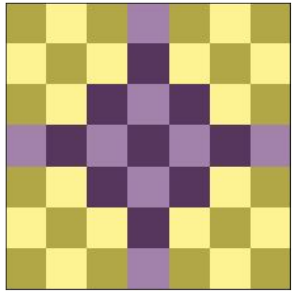
TAD anatomy

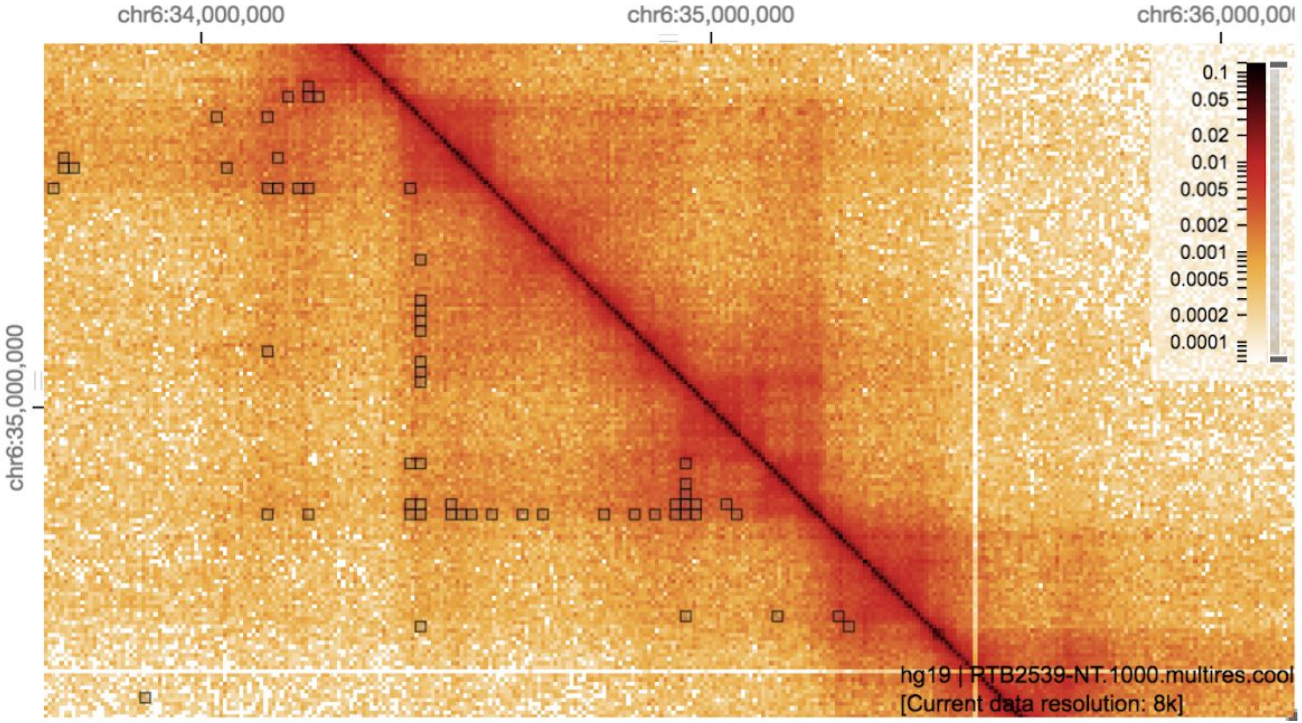


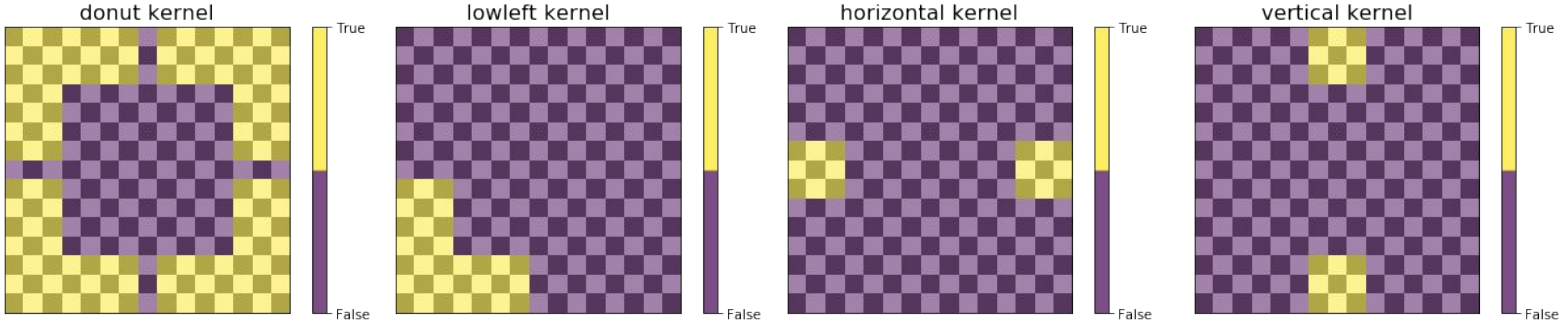
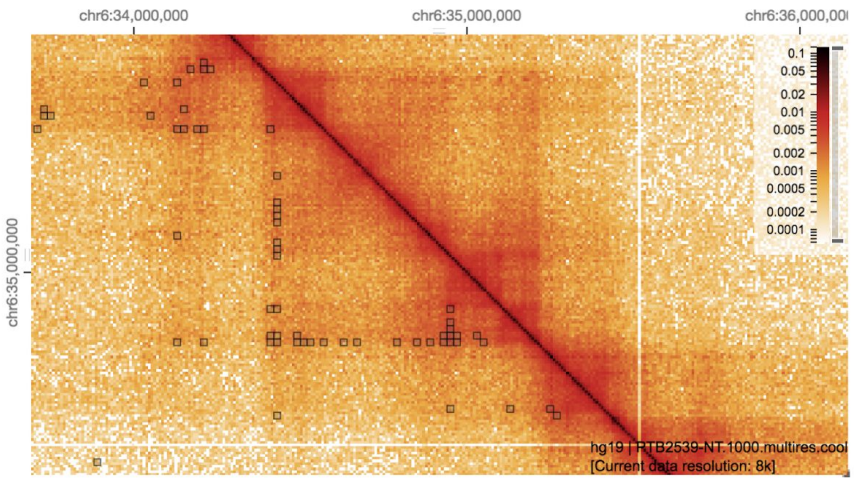
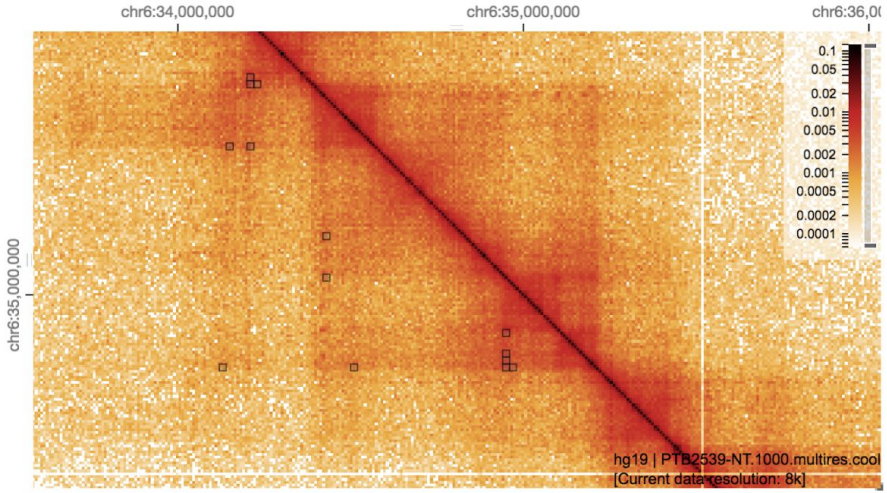


Fudenberg, Abdennur, et al, 2018. [Cold Spring Harbor Symposia on Quantitative Biology]

Donut alone is not a sufficient filter for “dots”





TAD borders, corners, flares etc… while we are looking for DOTS

Additional filters are used to emphasize “dots”