3D Genome Analysis with Galaxy

June 30
GCC 2014

Karen Reddy, Tonje G. Lien,
Morten Johansen, Jonas Paulsen

Training Day is Sponsored by

Doing the Branch, Release and Merge Waltz

Monday, June 30, 6:15pm
Salon A

http://bit.ly/gcc2014mergebof

We will focus on branching and release management with regard to existing instances which implement customized code within Galaxy. This may create huge challenges in the future, especially for instances in production which require a lot of maintenance and which run older versions of Galaxy. All Clouds and Clusters which require multiple extensions like:

- huge file management (upload, etc)
- authentication issues
- cluster/cloud connectivity
- And the customization of these issues is not easy and straightforward.

Galaxy End-User

Monday, June 30, 6:15pm
Multipurpose Room 324

bit.ly/gcc2014usersbof

This Birds-of-a-Feather session will serve as a forum for end-users of the Galaxy environment to share experiences and lessons learned, as well as address and discuss issues that hinder progress from the end-user perspective.

End-users of Galaxy who would like to share experiences (or listen to those of others) and developers interested in the perspective of the end-user should attend this BoF.

Break @ 3:00 - 3:30

Drinks and snacks will be available during the break, and in all Training Day Rooms after this workshop.

Network Options

Wifi

- eduroam
- hopkins
- GCC

    g@l@xycommittee

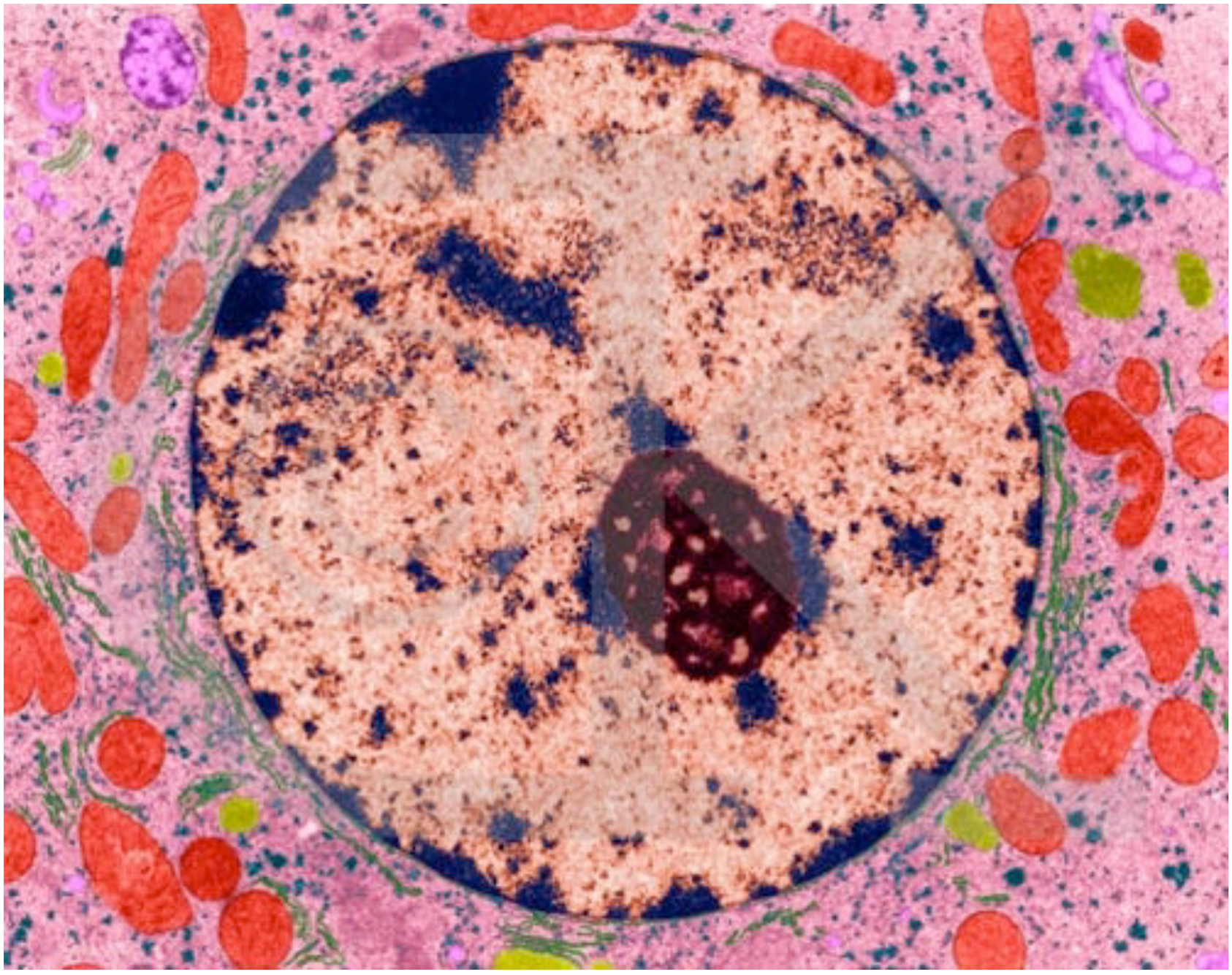
Ethernet: Just plug in.

Outline

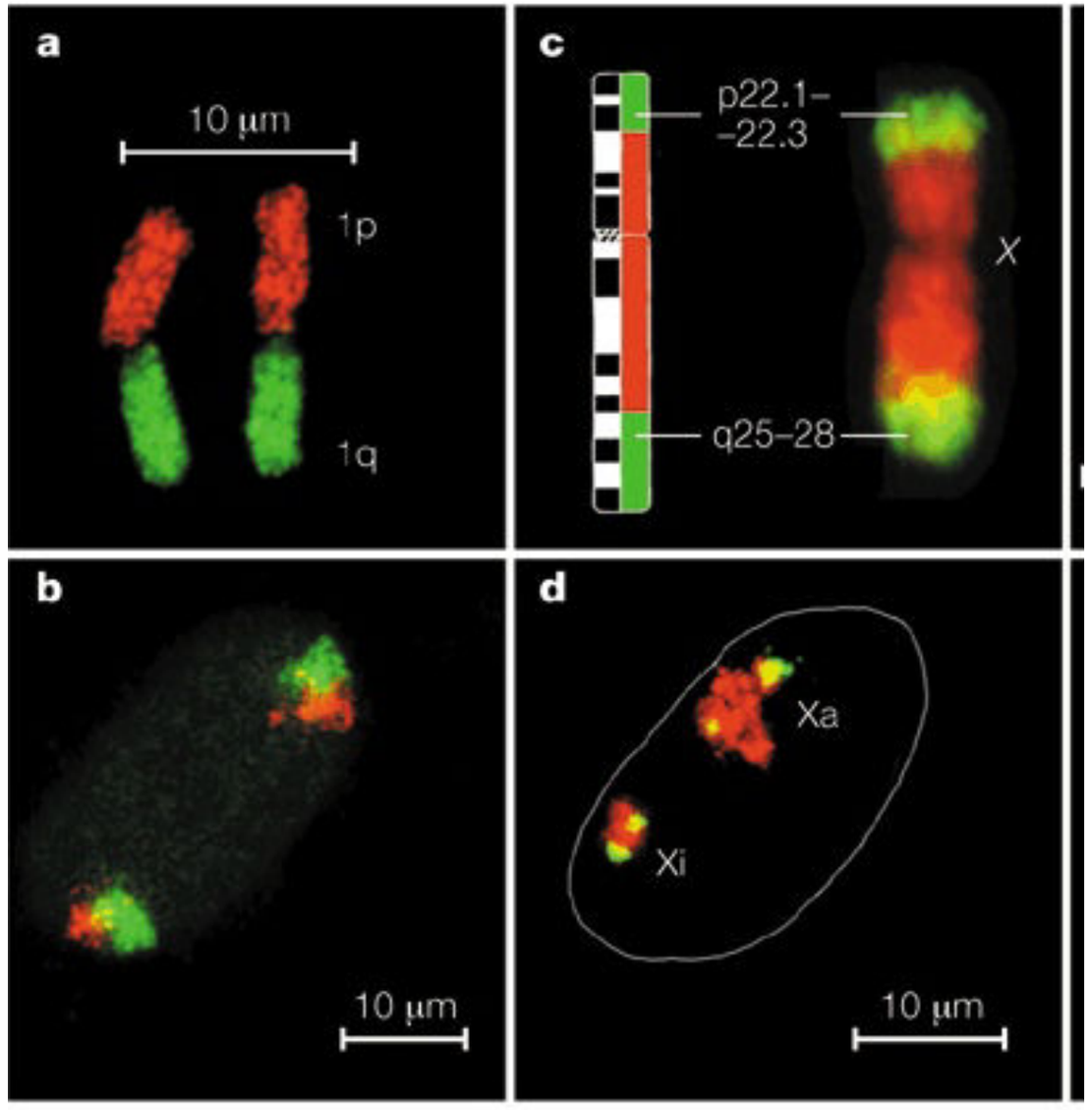
- Introduction to nuclear 3D organization
- Uploading and visualizing Hi-C
- Basics of data representation (genomic data types and tracks)
- Statistical analysis: descriptive and hypothesis testing
- Visualization of results
- 4 hands-on sessions

Nuclear Organization: regulation of the genome (?)

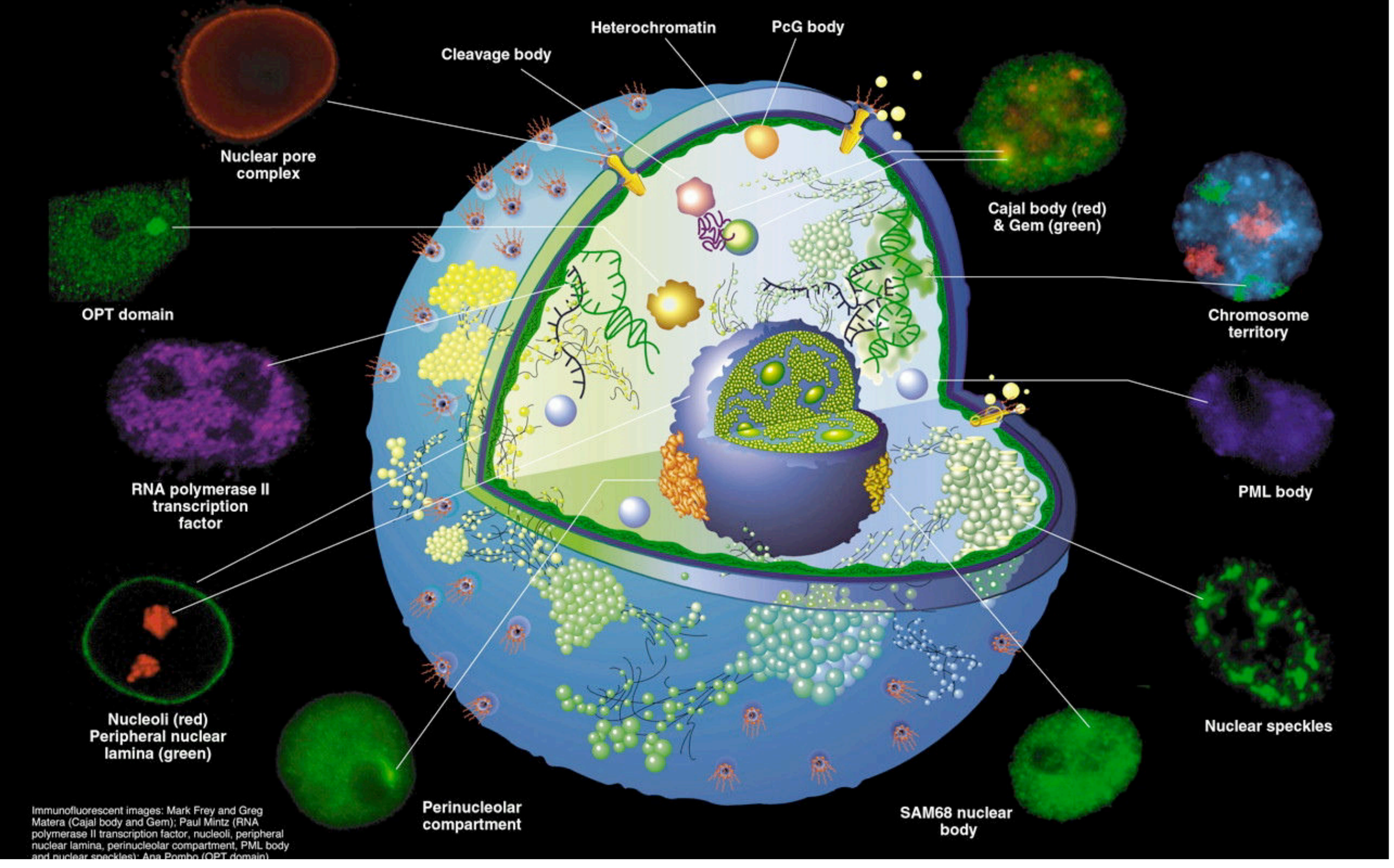
Heterochromatin and euchromatin



Chromosomes are organized



Cremer and Cremer, 2001

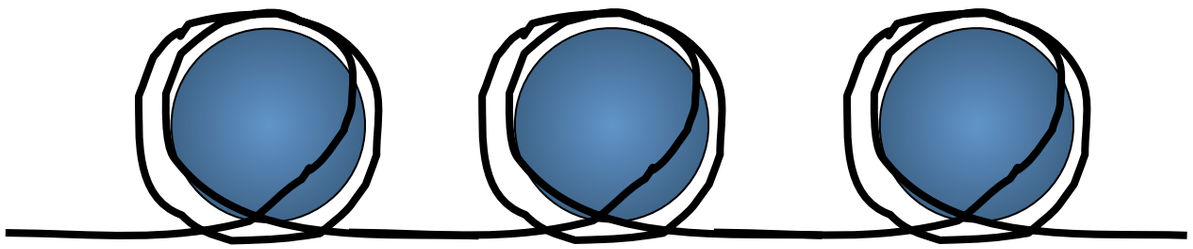


Processes (proteins) are organized

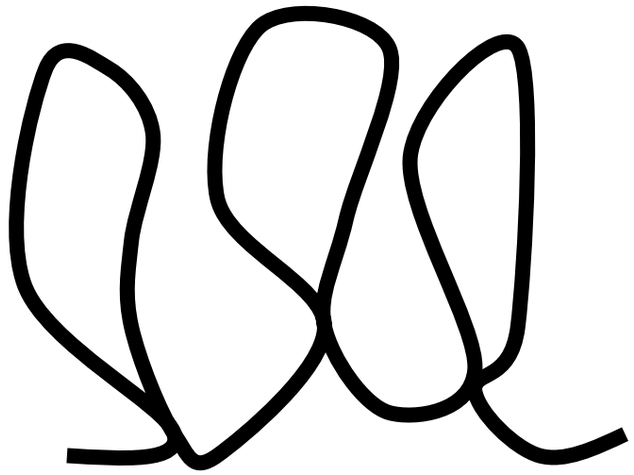
‘Nuclear domains’ Spector, Journal Cell Science

Factors Regulating Eukaryotic Gene Activity

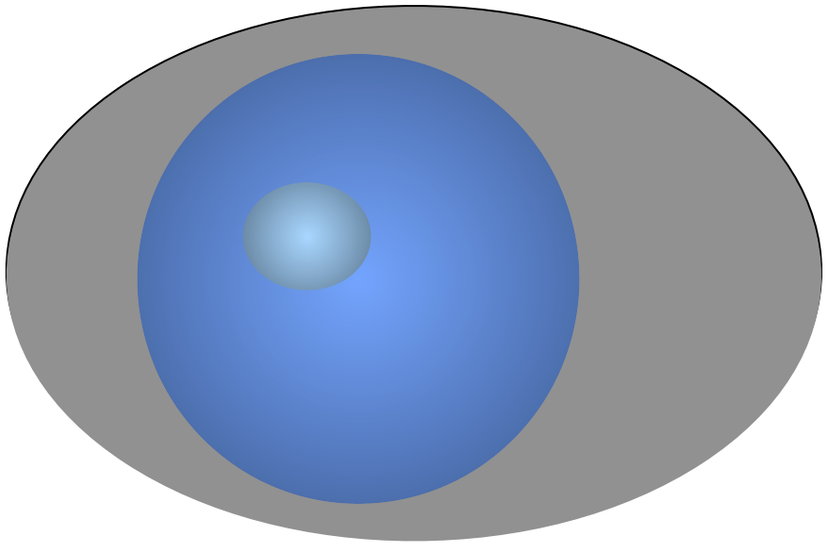
Cis-elements and *trans*-acting factors



Local chromatin structure



Higher-order
chromatin structure-
30nm fiber arranged on
Scaffold every 1-2 MB



Nuclear
compartmentalization

Understanding how the 3D genome functions requires analysis of disparate types of data ('tracks')

- Overview of data types
- Overview of scales/considerations

DamID

Dam-labeled Protein in Lamina

DpnI methyl specific digestion, purification and processing/labeling

Inner Nuclear Membrane

ChIP

Antibodies recognizing chromatin modification (e.g. H3K9Me2)

pull-down, chromatin isolation and processing/labeling

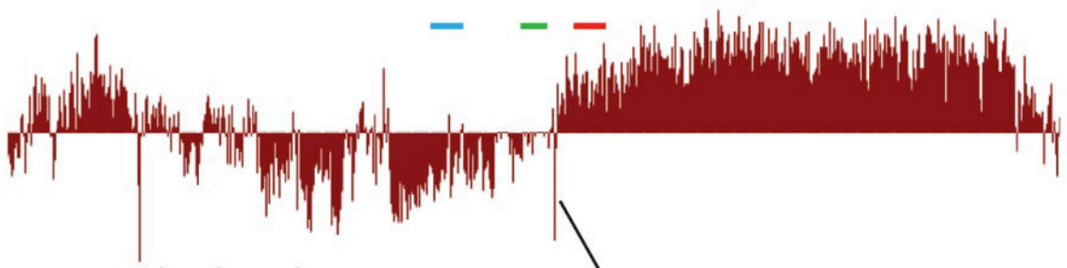
lamina associated region

active region of genome/gene

chromosomal segment

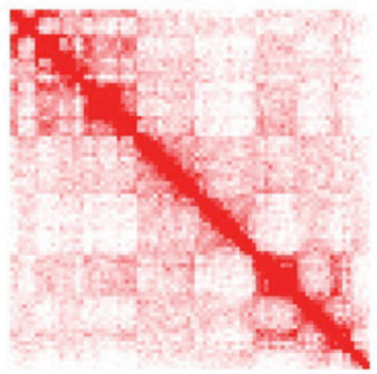
Hi-C

isolation of interacting fragments, deep sequencing and analysis



Map sequences back to the genome (array or deep sequencing), both techniques have resolution in the 1-2kb range, but information is a two dimensional representation of data.

Verify *in situ*

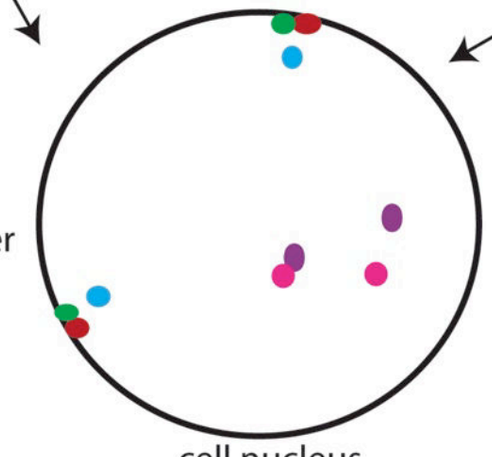


inter- or intra-chromosomal interaction map

Verify locus interaction *in situ*

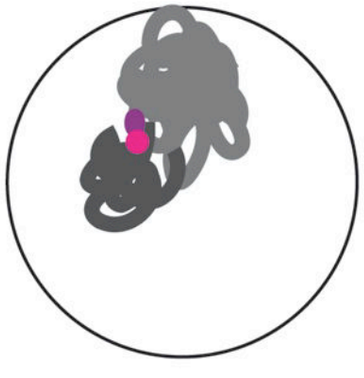
generate predicted *in situ* interactions and chromosome position algorithmically.

3D-FISH has a resolution on the order of 200kb, relatively low, but output maintains 3D information.



cell nucleus

predicted interacting loci



DNA probe hybridizing outside of LAD/LOCK (far from border)

DNA probe hybridizing outside of LAD/LOCK (near border)

DNA probe hybridizing to interior of LAD/LOCK

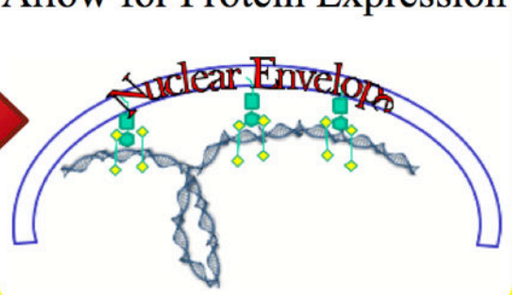
DNA Adenine Methylase Identification
DamID

Methylation Genomic DNA Double Stranded Adaptor

Create Dam-Protein X Fusion Vectors

Dam-Protein X

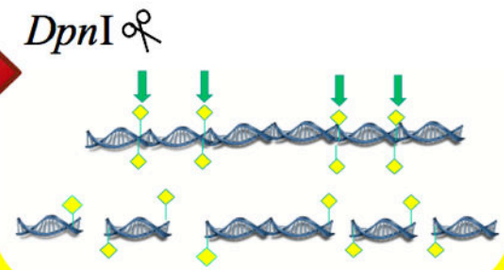
Virally Transduce Cells and Allow for Protein Expression



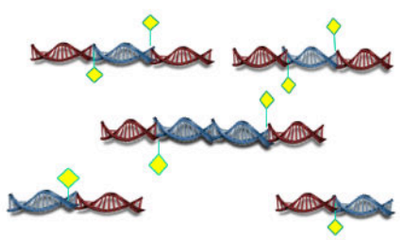
Prepare Genomic DNA

*Dpn*I Digest gDNA (Cuts methylated GATCs)

*Dpn*I

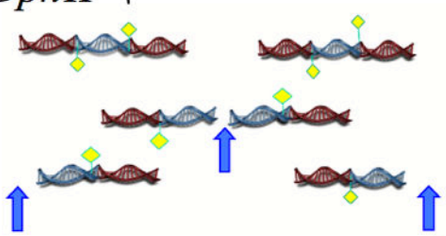


Ligate on Adaptors

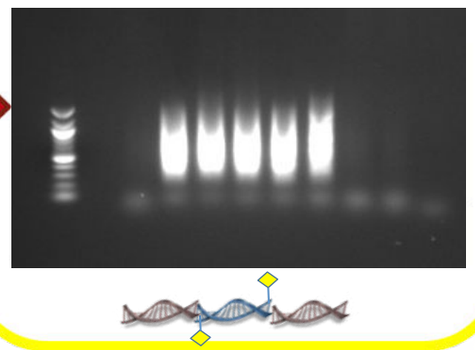


*Dpn*II Digest DNA (Cuts unmethylated GATCs)

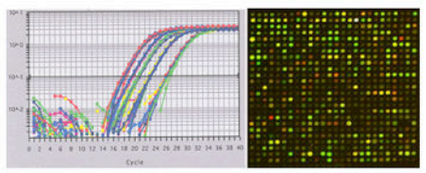
*Dpn*II



PCR Amplification

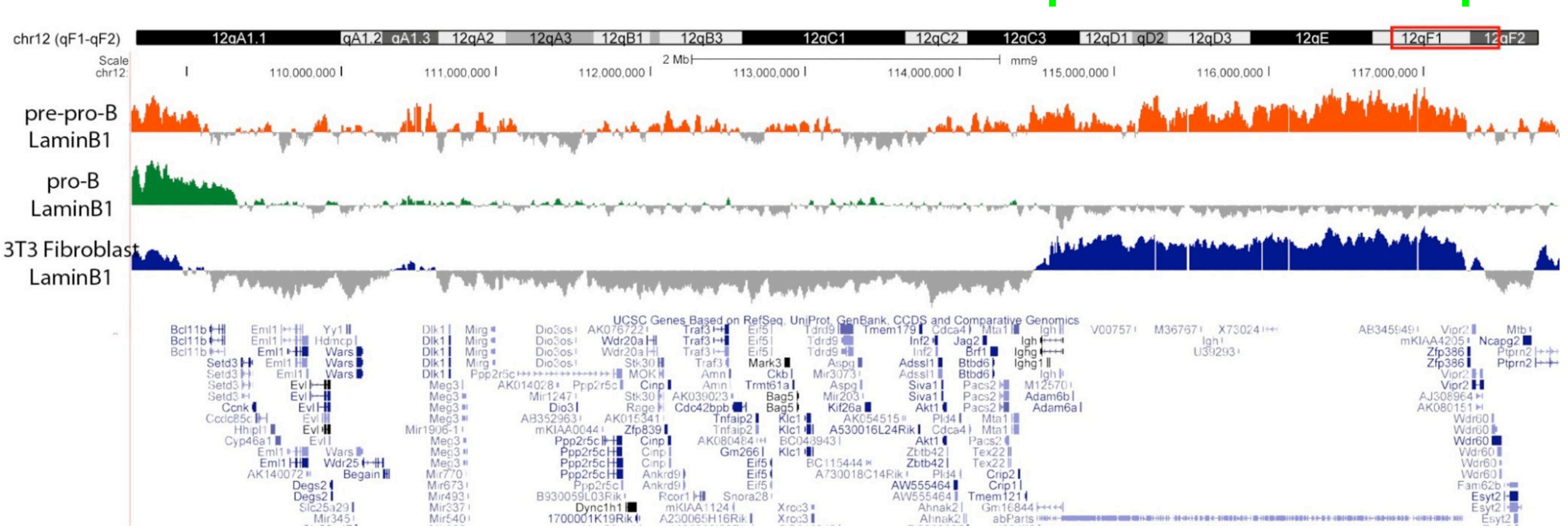


Analyze DNA by:
- q-PCR
- Microarrays



LADs are dynamic across cell types

Igh Locus

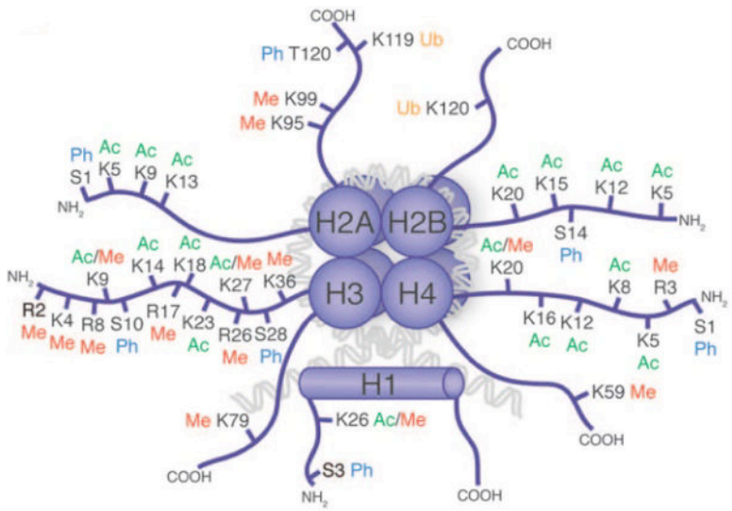


Immunoglobulin Heavy Chain (Igh) locus and surrounding region

Histones have modifiable N-terminal tails that influence gene activity

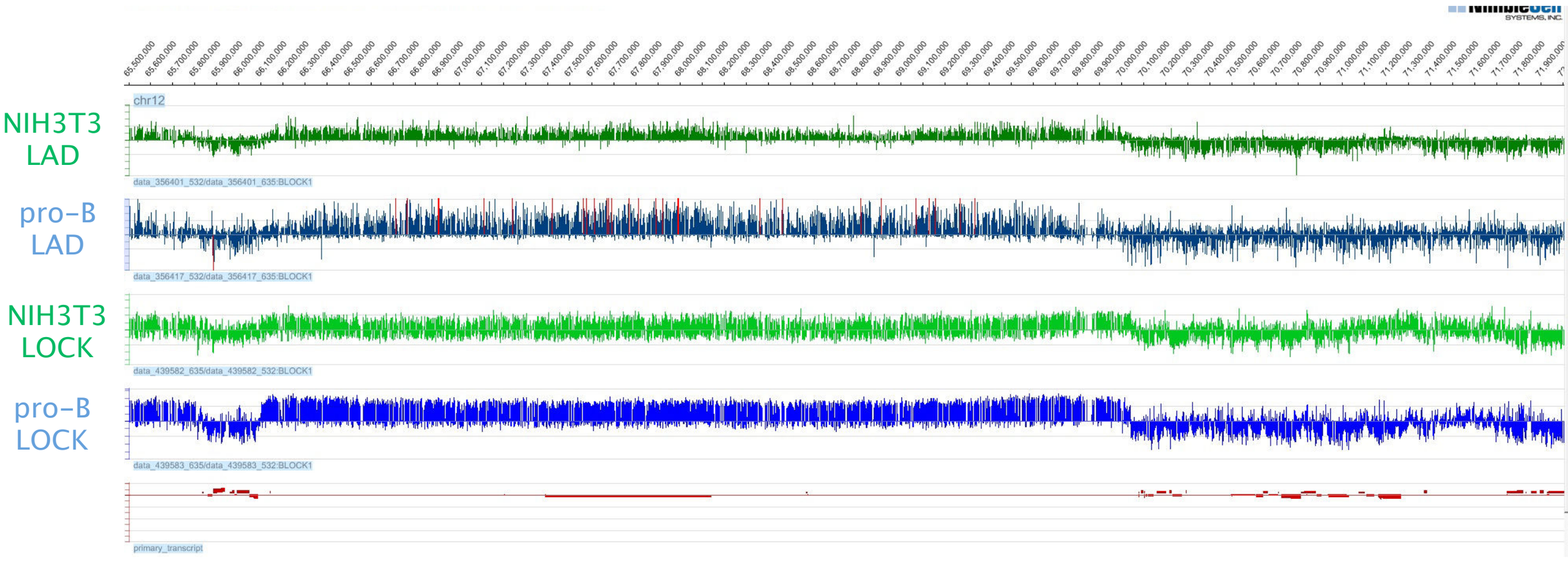
Nucleosomes can harbor post translational modifications that cause condensation and repression (heterochromatin) – Often lysine methylation

Alternatively, lysine acetlylation can lead to transcription upregulation and/or elongation

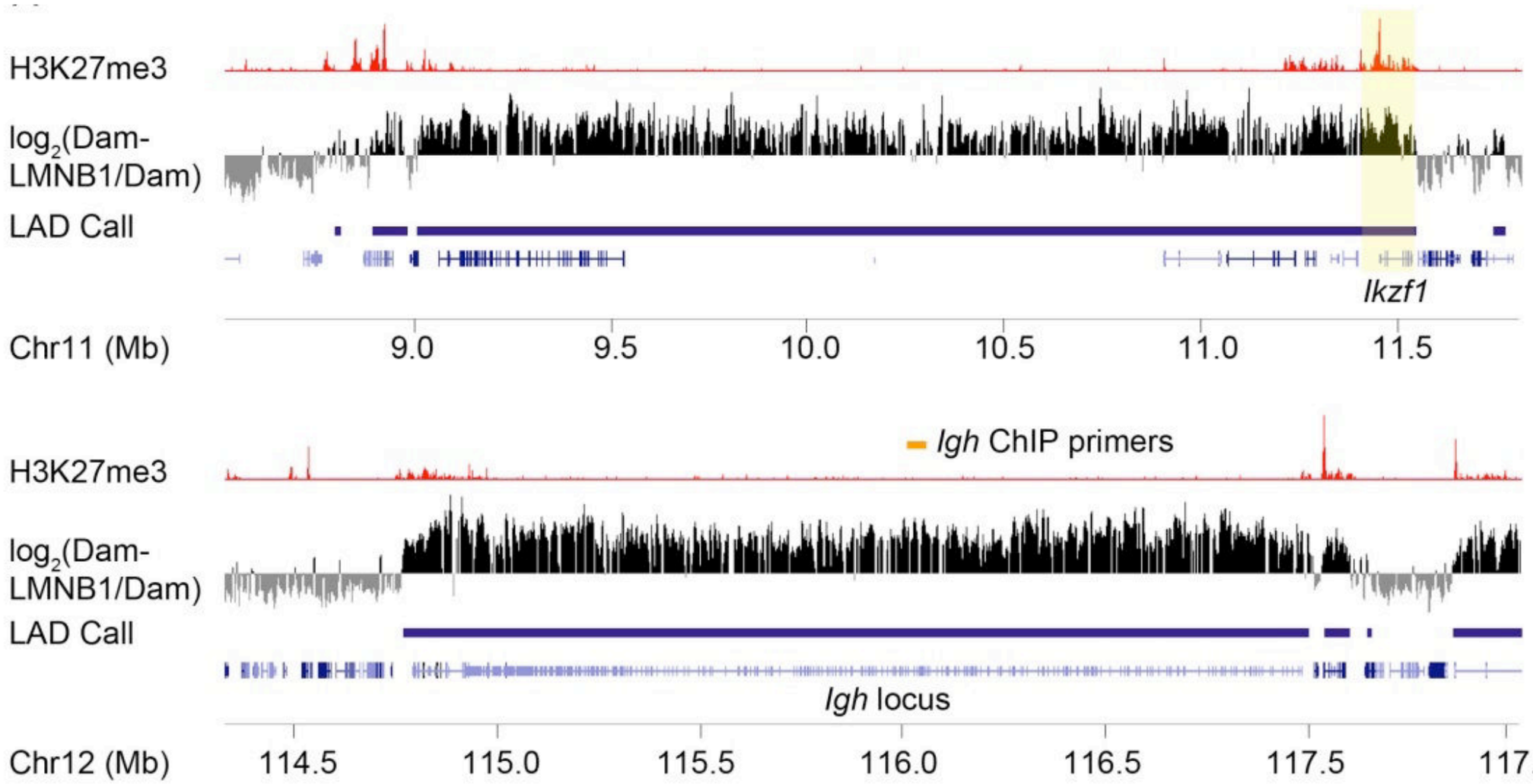


These modifications are added by specific enzymes and can act as a histone code for other molecules

Large chromatin signatures/regions

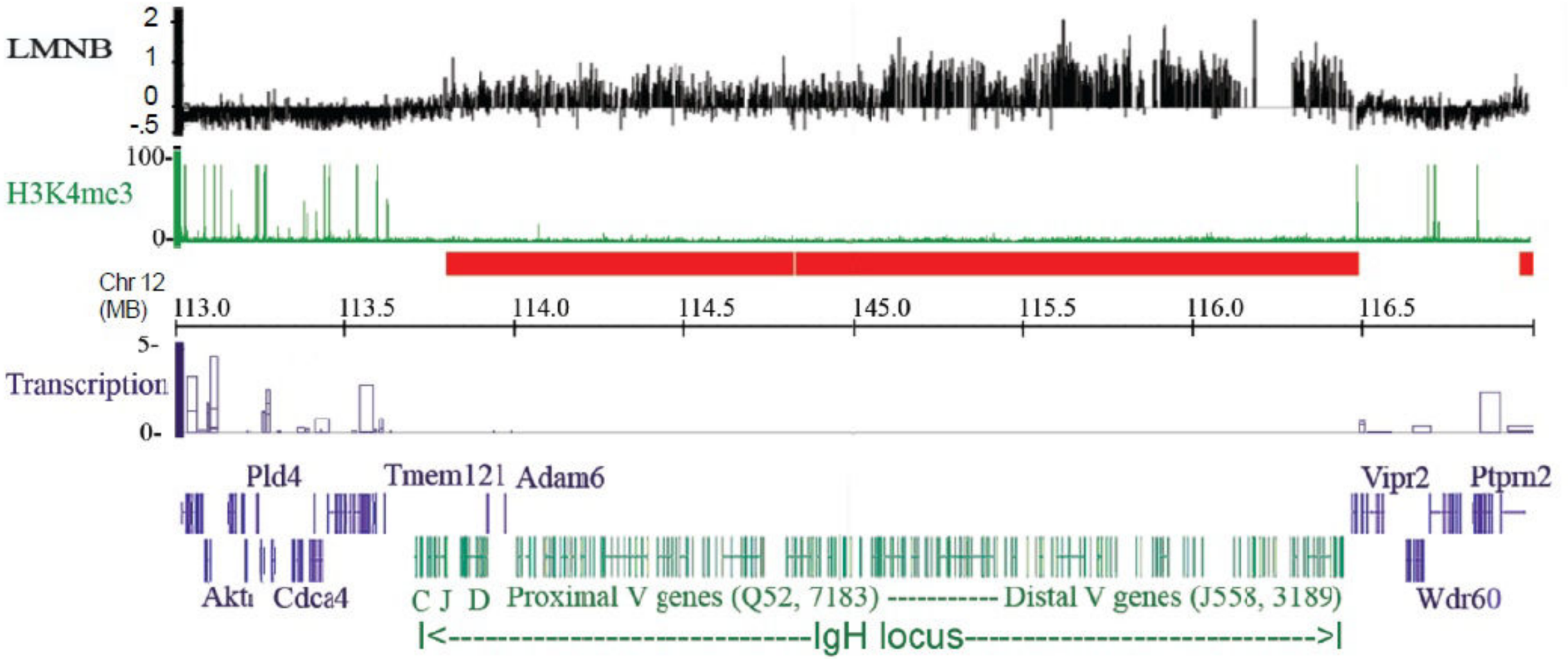


Chromatin States: large-ish (medium)



Harr et al, under revision JCB

Chromatin and transcription state-smaller regions

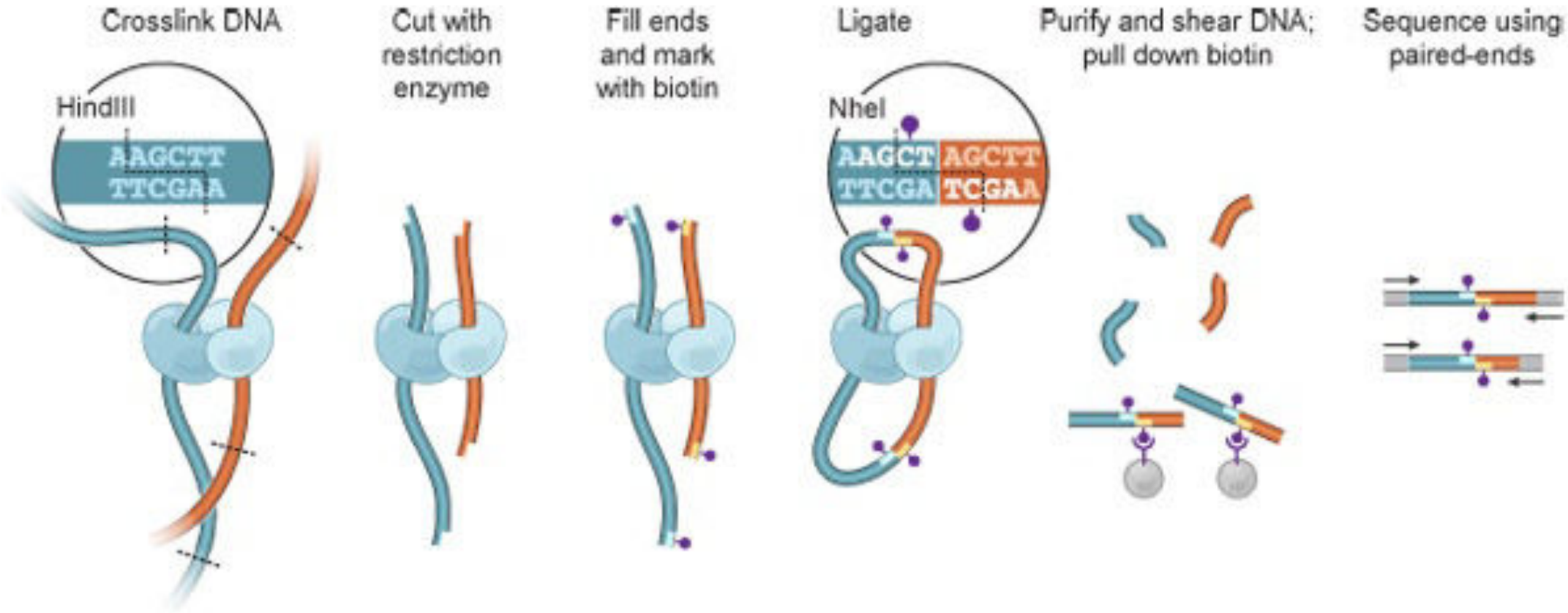


Zullo et al., Cell (2012)

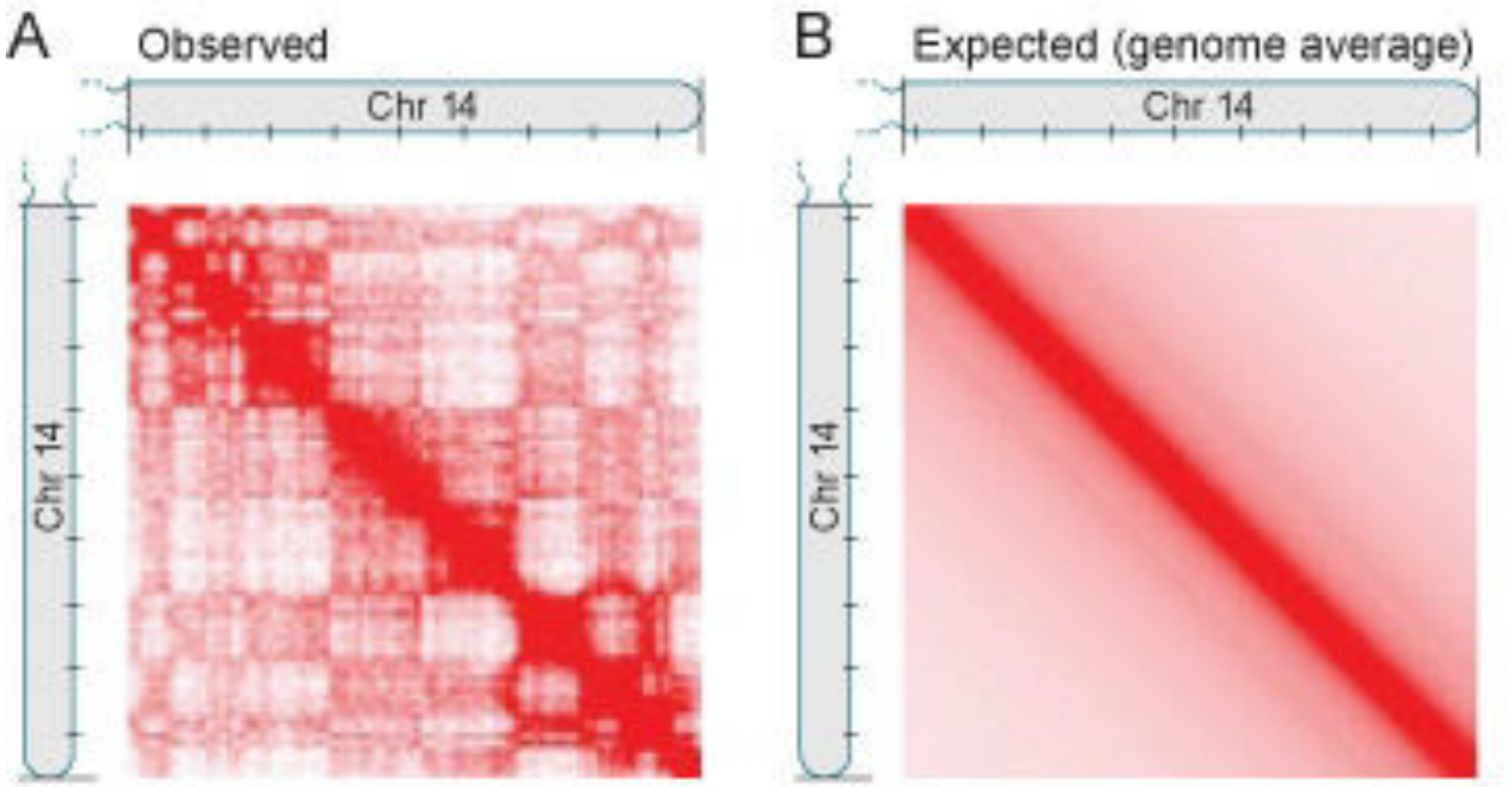
Other data types

- Nucleotide skewing (GC rich/AT rich)
- LINE, SINE or other elements
- Cytobands
- DNA breakpoints
- Gene rich/gene poor
- Conservation
- Non-coding transcripts

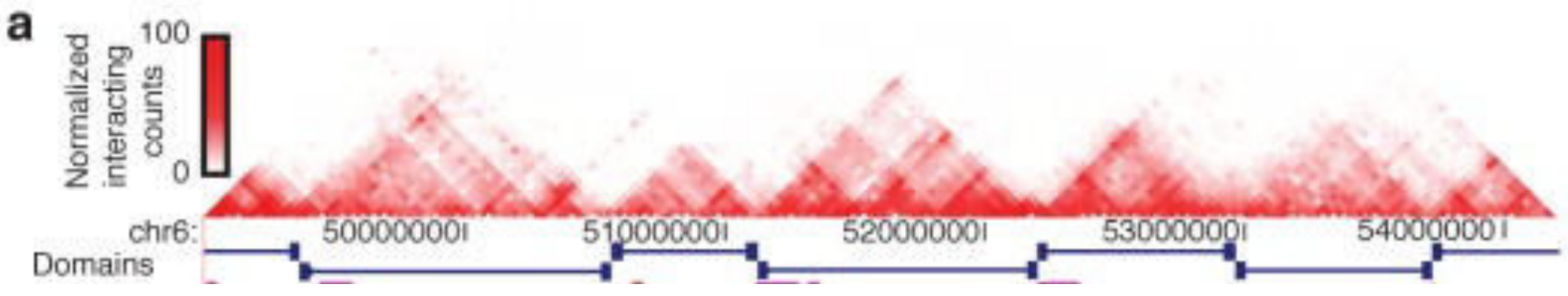
Chromosome conformation Capture (Hi C)



Hi-C: A Method to Study the Three-dimensional Architecture of Genomes.
Nynke L. van Berkum, Erez Lieberman-Aiden, Louise Williams, Maxim Imakaev, Andreas Gnirke, Leonid A. Mirny, **Job Dekker,** Eric S. Lander

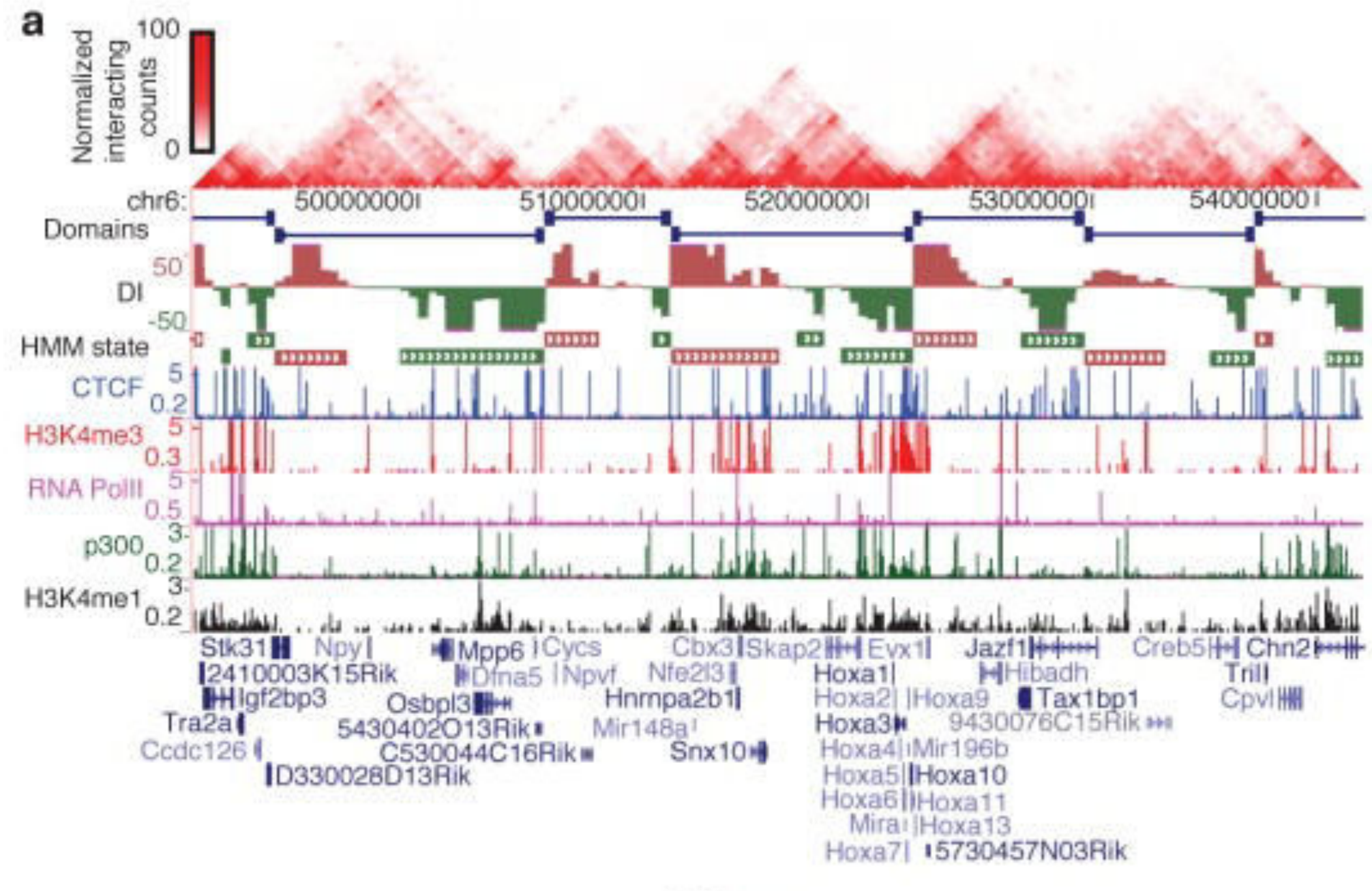


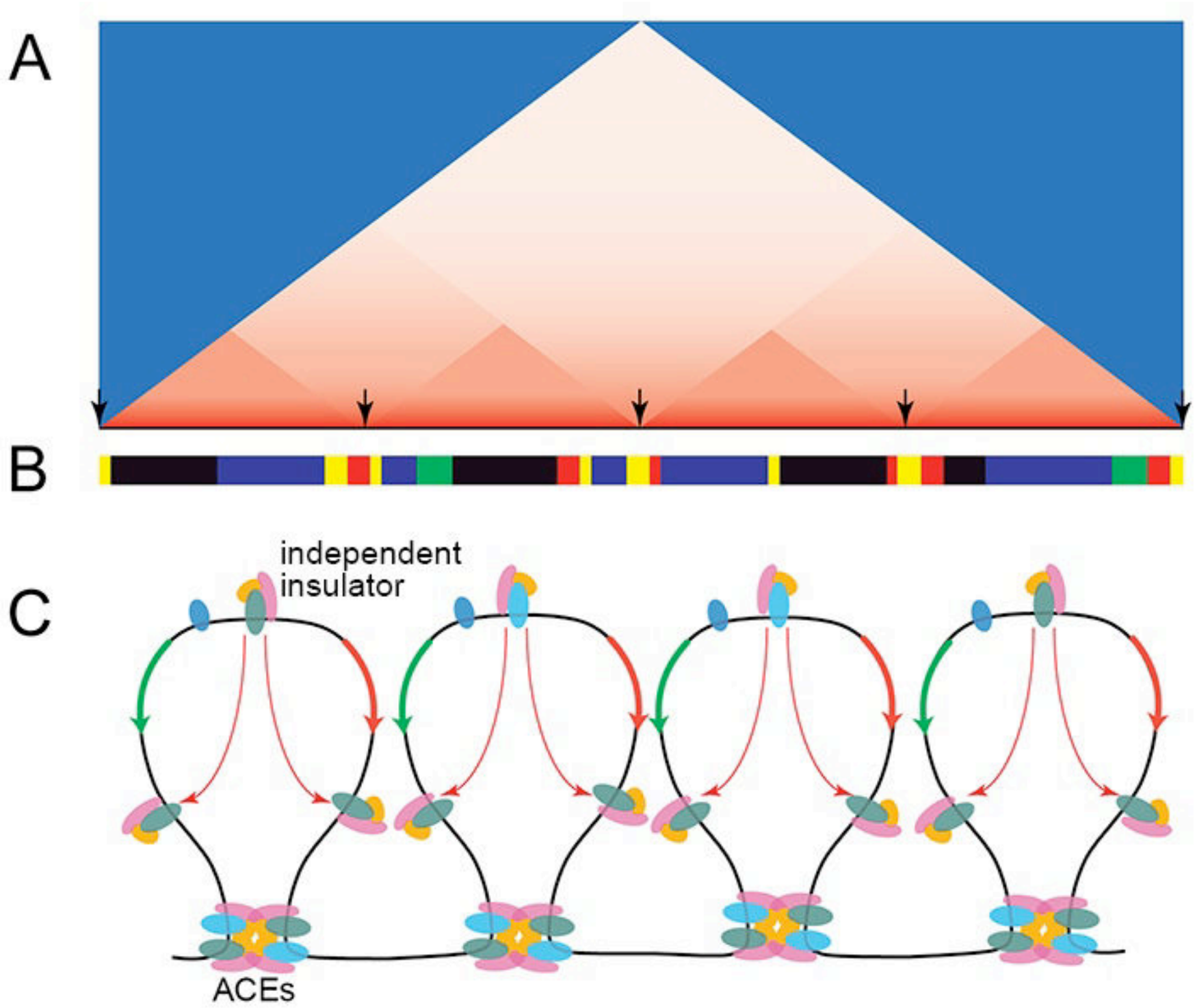
Dekker Lab



Dixon et al., 2012

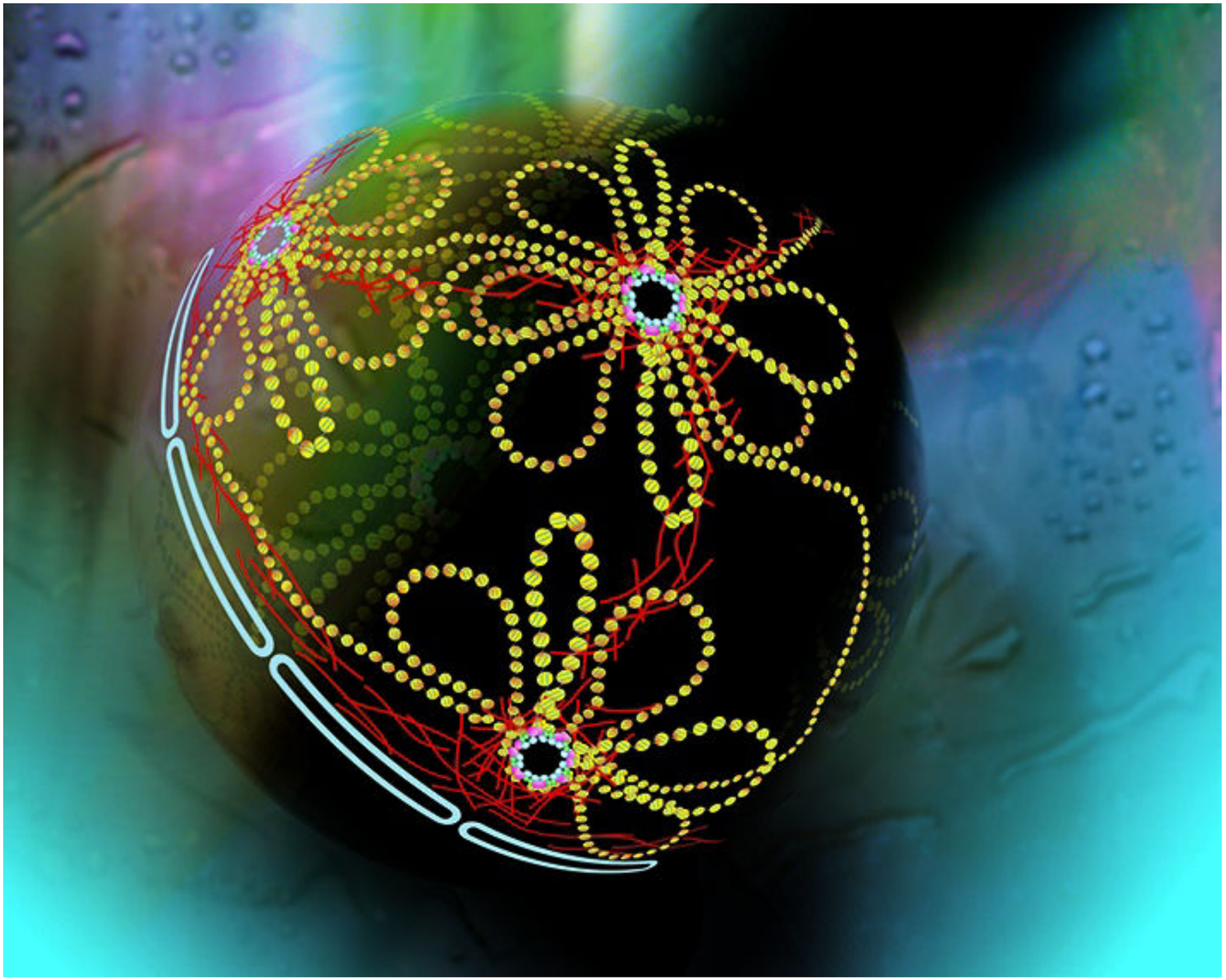
Putting it all together….





Phillips-Cremens et al., 2013

Chromosomes organized into loops AND scaffolded at the nuclear periphery



Hands-on sessions

- Step by step recipe using Galaxy Pages
- If you're stuck - ask us or a neighbor
- All steps will be shown after you have tried on your own

Technical notes

- Do not submit large jobs on your own during the course
- We are many people submitting jobs at the same time!
- Might be some queueing of runs or slow GUI
- Might be some DB operational errors (just refresh)

Create account

- Go to http://**hyperbrowser.uio.no/3d**
- Create an account

Hands on: Uploading and visualizing Hi-C data

- Upload a file containing Hi-C data
  - chr6 in RWPE1 cell-line
- Convert to GTrack format
- Visualize data as a heat map

Hands on 1: Uploading and visualizing Hi-C data

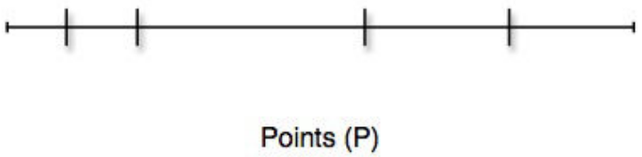
Go to:

https://hyperbrowser.uio.no/3d/u/gcc2014/p/visualization

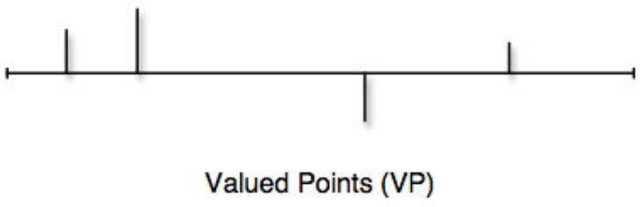
Demo

Analysis of genomic data in HyperBrowser

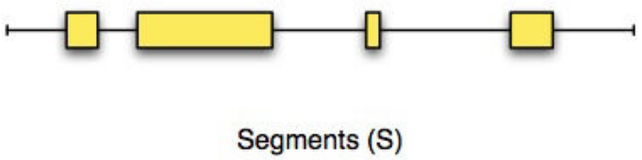
Genomic data: 7 basic track types



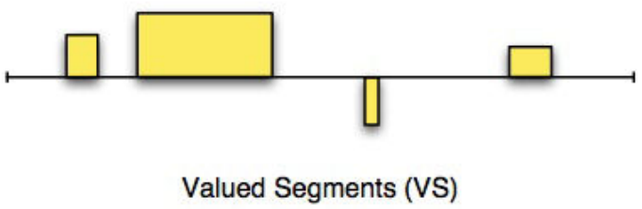
SNPs, TFs



GWAS

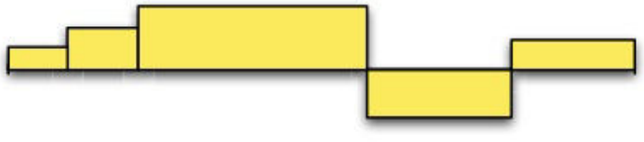


Genes,
histone modifications

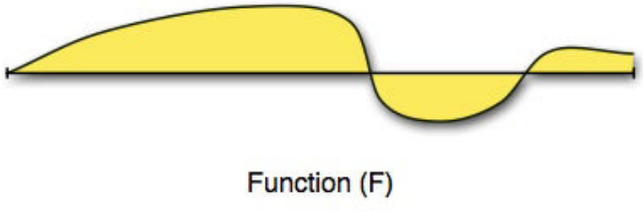


Gene expression

Cytobands,
restriction fragments
chromatin states

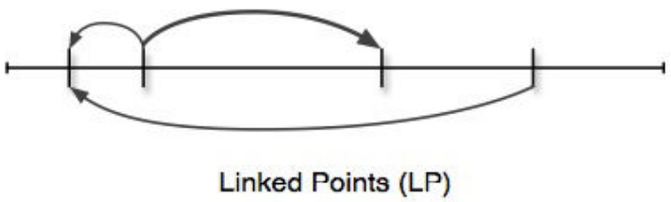


GC content

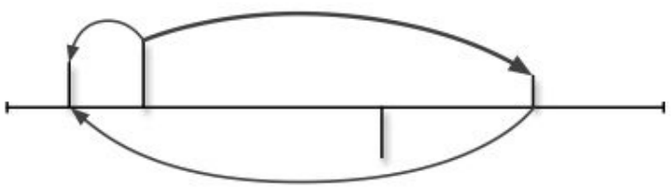


Melting
temperature

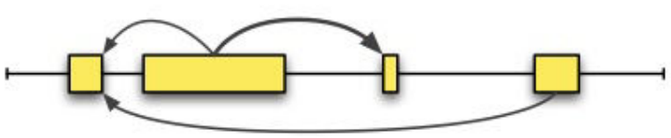
Gundersen et al. (2011)



Linked Points (LP)

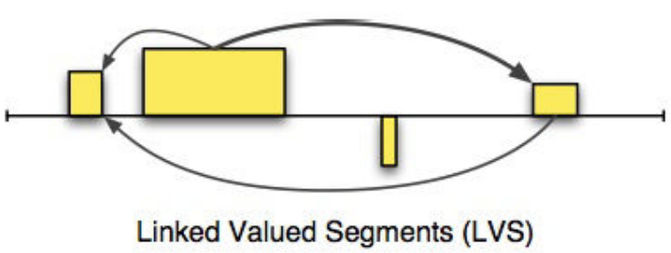


Linked Valued Points (LVP)

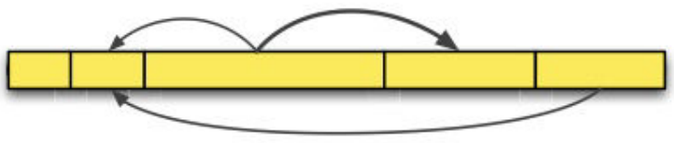


ChIA-PET

Linked Segments (LS)

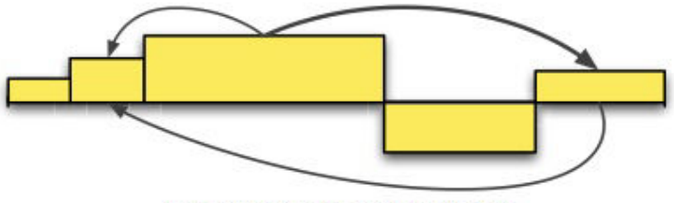


Linked Valued Segments (LVS)

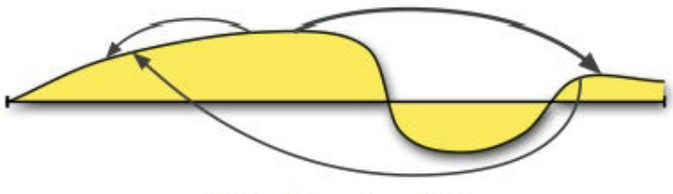


5C, Hi-C

Linked Genome Partition (LGP)



Linked Step Function (LSF)



Linked Function (LF)

Linked track types

Gundersen et al. (2011)

Combination of track types

- Combinations of track types give rise to structured statistical questions

	Points	**Segments**	**Function**	**Valued Points**	**Valued Segments**
Points	Different frequencies?	Located inside?	Higher values at locations?		Located in highly valued segments?
	Segments	Overlap?	Higher values inside?		
		Function	Correlated?		
			Valued Points	Nearby values similar?	Categories differentially located in targets?
				Valued Segments	

Sandve et al. (2010 & 2013)

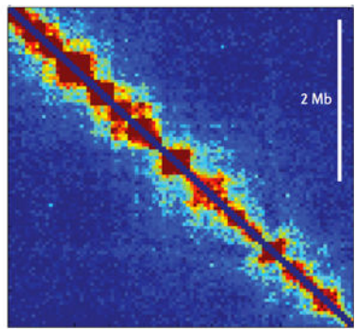
Track1 type	Track2 type	Statistical investigation	Description
F	F	Correlated?	Are the values of track1 and track2 more positively correlated than expected by chance?
P	F	Higher values at locations?	Are the values of track2 higher at the points of track1, than what is expected by chance?
S	F	Higher values inside?	Are the values of track2 higher inside the segments of track1, than what is expected by chance?
P	VS	Located in segments with high values?	Does the number of track1-points that fall in track2-segments depend on the value of track2-segments?
S	VP	Higher values inside segments?	Do the points of track2 that occur inside segments of track1 have higher values than points occurring outside the segments of track1?
VP	VP	Nearby values similar?	When track1-points and track2-point are nearby each other, are the values more similar than expected by chance?
P	VS (c/c)	Located in case segments	Does the number of track1-points that fall in track2-segments depend on whether the track2-segments are marked as case or control?
VS (c/c)	S	Preferential overlap?	Are the segments of track1 marked as case overlapping unexpectedly more with the segments of track2 than the segments of track1 marked as control?
VP (cat)	VS (cat)	Category pairs differentially co-located?	Which categories of track1-points fall more inside which categories of track2-segments?
LGP	P	Co-localized in 3D?	Are the points of track2 closer in 3D (as defined by track1) than expected by chance?
P	P	Different frequencies?	Where is the relative frequency of points of track1 different from the relative frequency of points of track2, more than expected by chance?
P	P	Located nearby?	Are the points of track1 closer to the points of track2 than expected by chance?
P	S	Located inside?	Are the points of track1 falling inside the segments of track2, more than expected by chance?
P	S	Located non-uniformly inside?	Do the points of track1 tend to accumulate more toward the borders of the segments of track2?
P	S	Located nearby?	Are the points of track1 closer to the segments of track2 than expected by chance?
S	S	Similar segments?	Are track1-segments similar (in position and length) to track2-segments, more than expected by chance?
S	S	Overlap?	Are the segments of track1 overlapping the segments of track2, more than expected by chance?
S	S	Located nearby?	Are the segments of track1 closer to the segments of track2 than expected by chance?

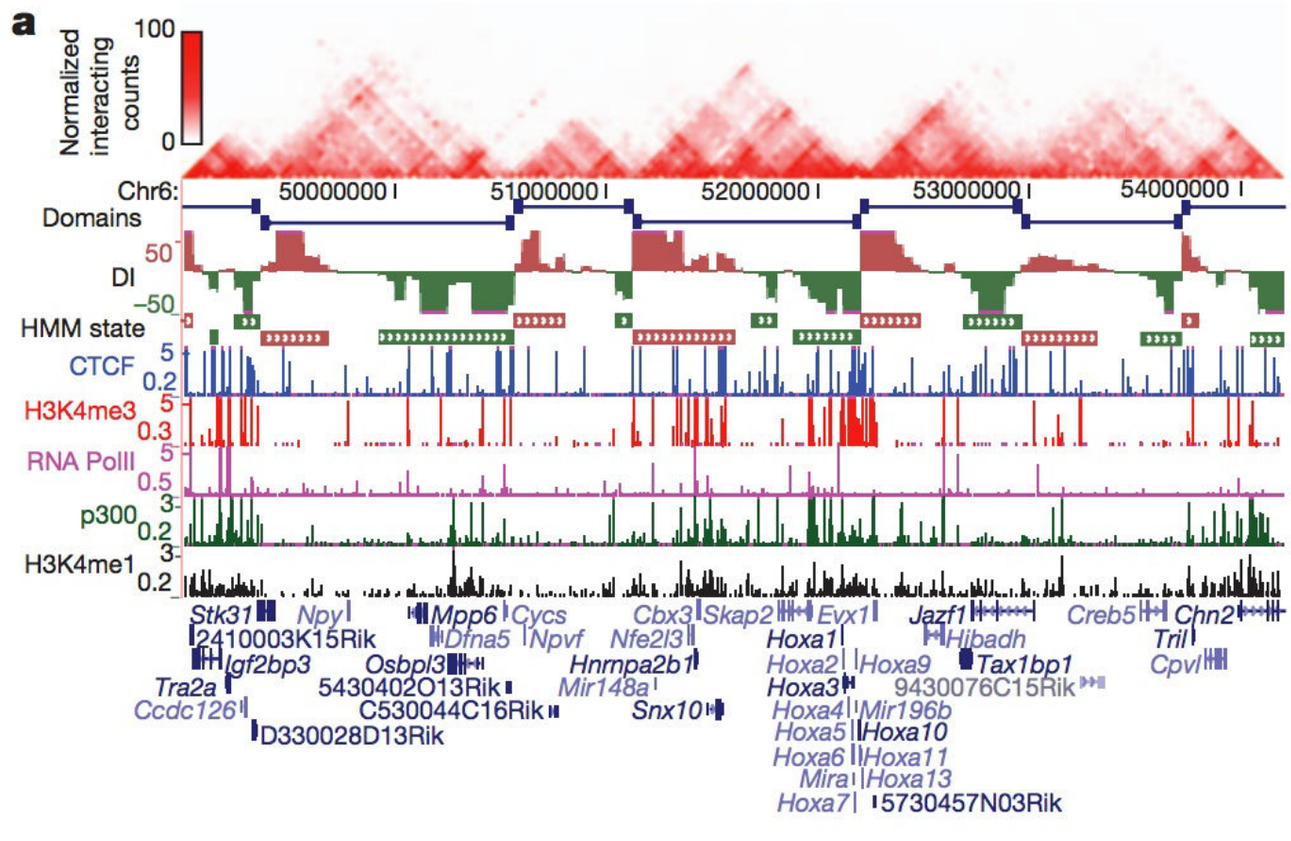
Track1 type	Track2 type	Statistical investigation	Description
F	F	Correlated?	Are the values of track1 and track2 more positively correlated than expected by chance?
P	F	Higher values at locations?	Are the values of track2 higher at the points of track1, than what is expected by chance?
S	F	Higher values inside?	Are the values of track2 higher inside the segments of track1, than what is expected by chance?
P	VS	Located in segments with high values?	Does the number of track1-points that fall in track2-segments depend on the value of track2-segments?
S	VP	Higher values inside segments?	Do the points of track2 that occur inside segments of track1 have higher values than points occurring outside the segments of track1?
VP	VP	Nearby values similar?	When track1-points and track2-point are nearby each other, are the values more similar than expected by chance?
P	VS (c/c)	Located in case segments	Does the number of track1-points that fall in track2-segments depend on whether the track2-segments are marked as case or control?
VS (c/c)	S	Preferential overlap?	Are the segments of track1 marked as case overlapping unexpectedly more with the segments of track2 than the segments of track1 marked as control?
VP (cat)	VS (cat)	Category pairs differentially co-located?	Which categories of track1-points fall more inside which categories of track2-segments?
LGP	P	Co-localized in 3D?	Are the points of track2 closer in 3D (as defined by track1) than expected by chance?
P	P	Different frequencies?	Where is the relative frequency of points of track1 different from the relative frequency of points of track2, more than expected by chance?
P	P	Located nearby?	Are the points of track1 closer to the points of track2 than expected by chance?
P	S	Located inside?	Are the points of track1 falling inside the segments of track2, more than expected by chance?
P	S	Located non-uniformly inside?	Do the points of track1 tend to accumulate more toward the borders of the segments of track2?
P	S	Located nearby?	Are the points of track1 closer to the segments of track2 than expected by chance?
S	S	Similar segments?	Are track1-segments similar (in position and length) to track2-segments, more than expected by chance?
S	S	Overlap?	Are the segments of track1 overlapping the segments of track2, more than expected by chance?
S	S	Located nearby?	Are the segments of track1 closer to the segments of track2 than expected by chance?

Hands on 2: are insulators/CTCF found at the borders of TADs?

Are insulators/CTCF found at the borders of TADs?

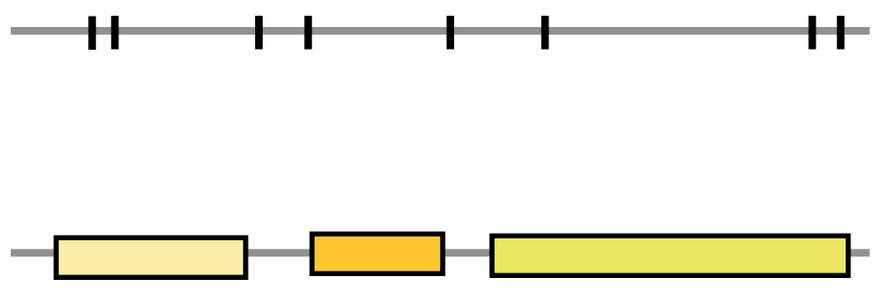
Dixon et al. (2012)





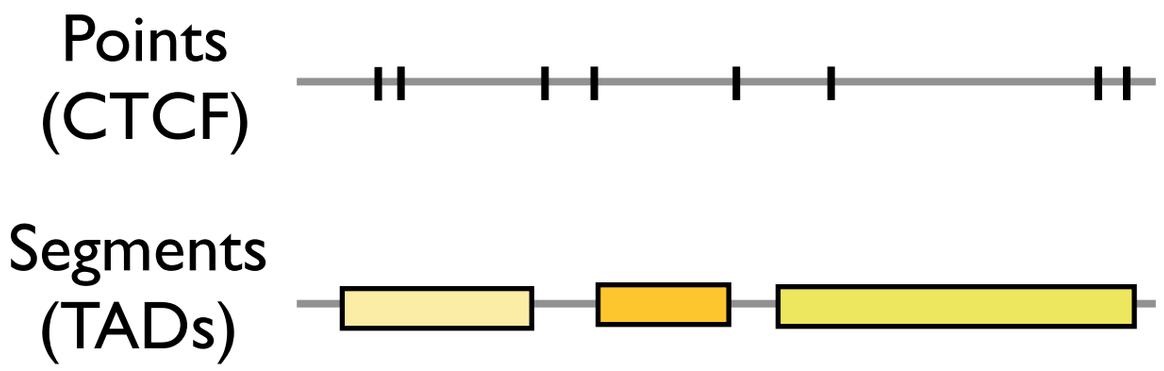
Points (CTCF)

Segments (TADs)



Hands on 2: are insulators/CTCF found at the borders of TADs?

- Visualize enrichment of CTCF at borders of TADs
- Test whether the enrichment is statistically significant



Hands on 2: are insulators/CTCF found at the borders of TADs?

Go to:

- hyperbrowser.uio.no/3d/**u/gcc2014/p/insulators**

What is hypothesis testing?

- H0: null hypothesis (a neutral baseline)

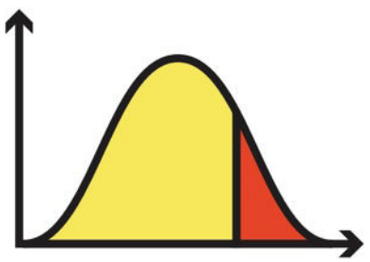
  H1: alternative hypothesis (what you really want to show)

- Test statistic: T

- P value: How likely is our observation (or more extreme), given H0

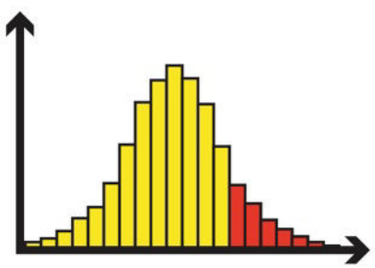
Distribution of T is known: (analytic test)

p = P(T > t_obs)



Distribution of T is unknown: (permutation test)

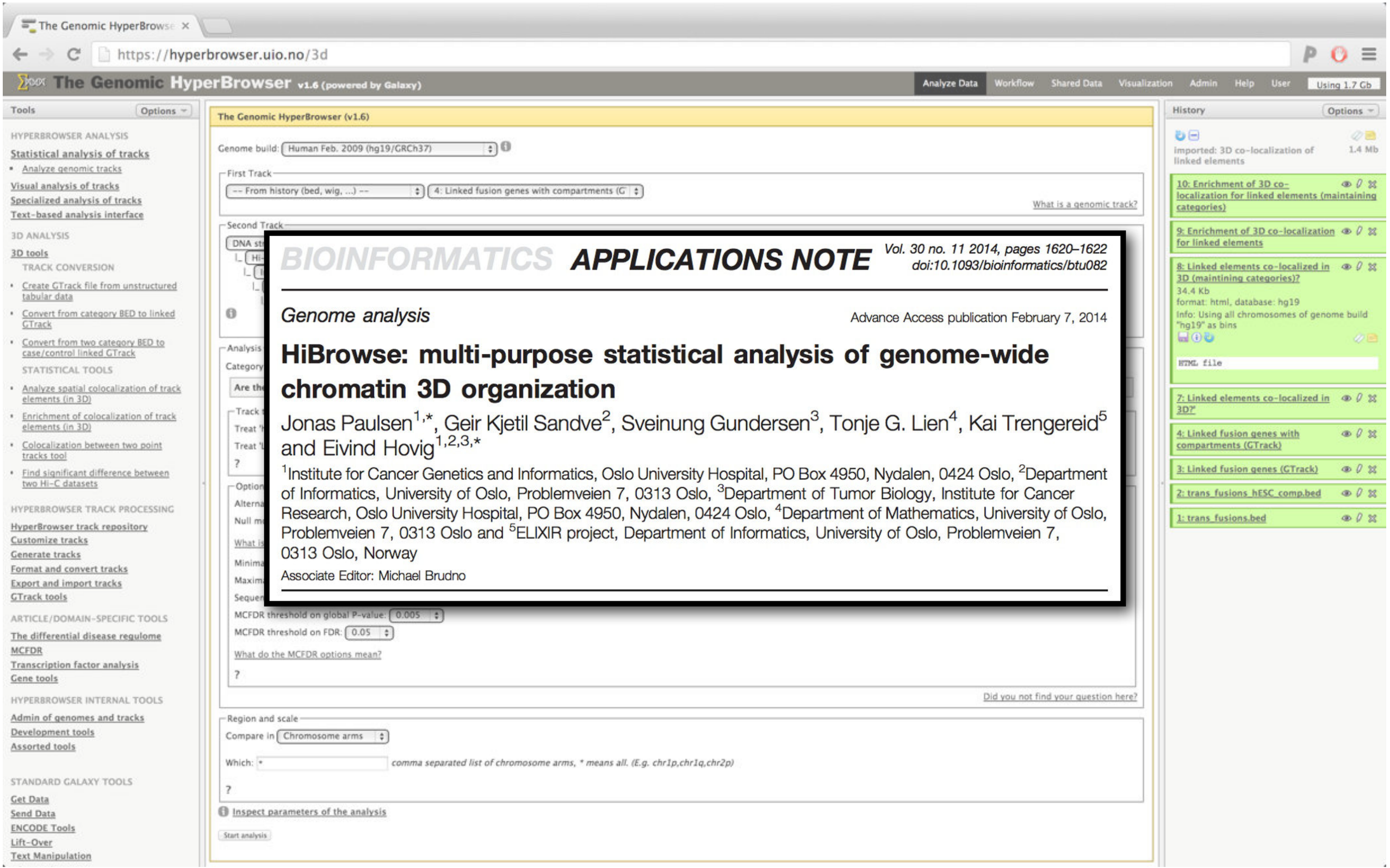
$$p = \frac{\sum_{r=1}^{R} I(t_r \geq t_{obs}) + 1}{R + 1}$$



- Observation unlikely -> low p value -> reject H0, left with H1

Analysis of Hi-C data in HyperBrowser

HiBrowse



BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 11 2014, pages 1620–1622
doi:10.1093/bioinformatics/btu082

Genome analysis

Advance Access publication February 7, 2014

HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization

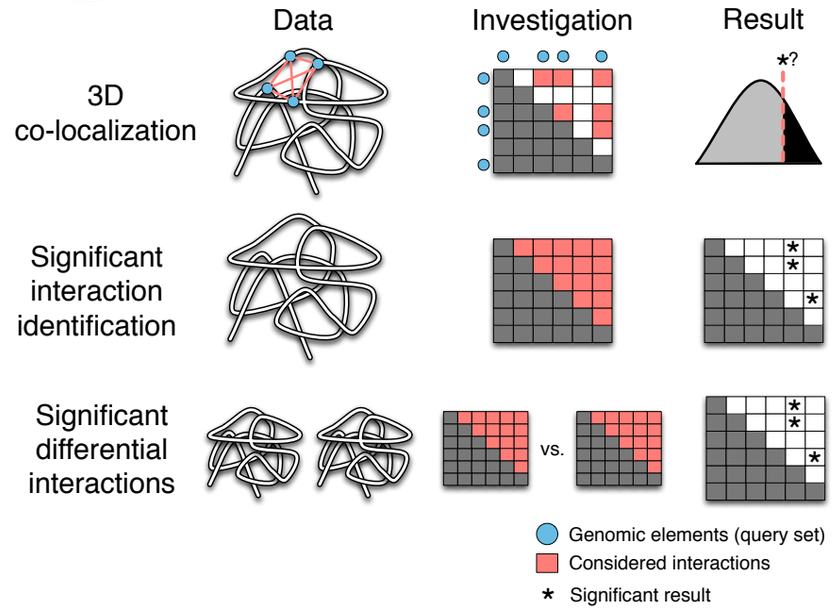
Jonas Paulsen[1,*], Geir Kjetil Sandve[2], Sveinung Gundersen[3], Tonje G. Lien[4], Kai Trengereid[5] and Eivind Hovig[1,2,3,*]

[1]Institute for Cancer Genetics and Informatics, Oslo University Hospital, PO Box 4950, Nydalen, 0424 Oslo, [2]Department of Informatics, University of Oslo, Problemveien 7, 0313 Oslo, [3]Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, PO Box 4950, Nydalen, 0424 Oslo, [4]Department of Mathematics, University of Oslo, Problemveien 7, 0313 Oslo and [5]ELIXIR project, Department of Informatics, University of Oslo, Problemveien 7, 0313 Oslo, Norway

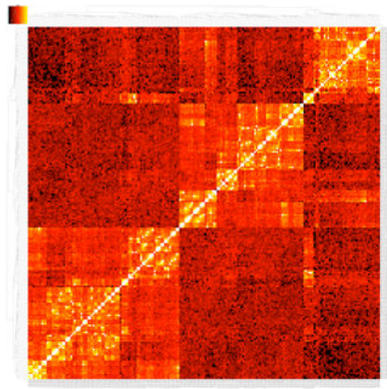
Associate Editor: Michael Brudno

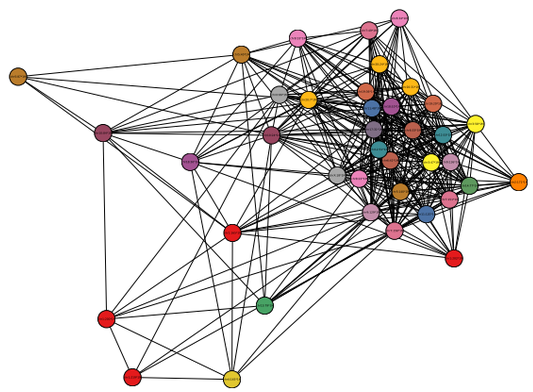
HiBrowse

- Hypothesis testing



- Descriptive/visualization analysis

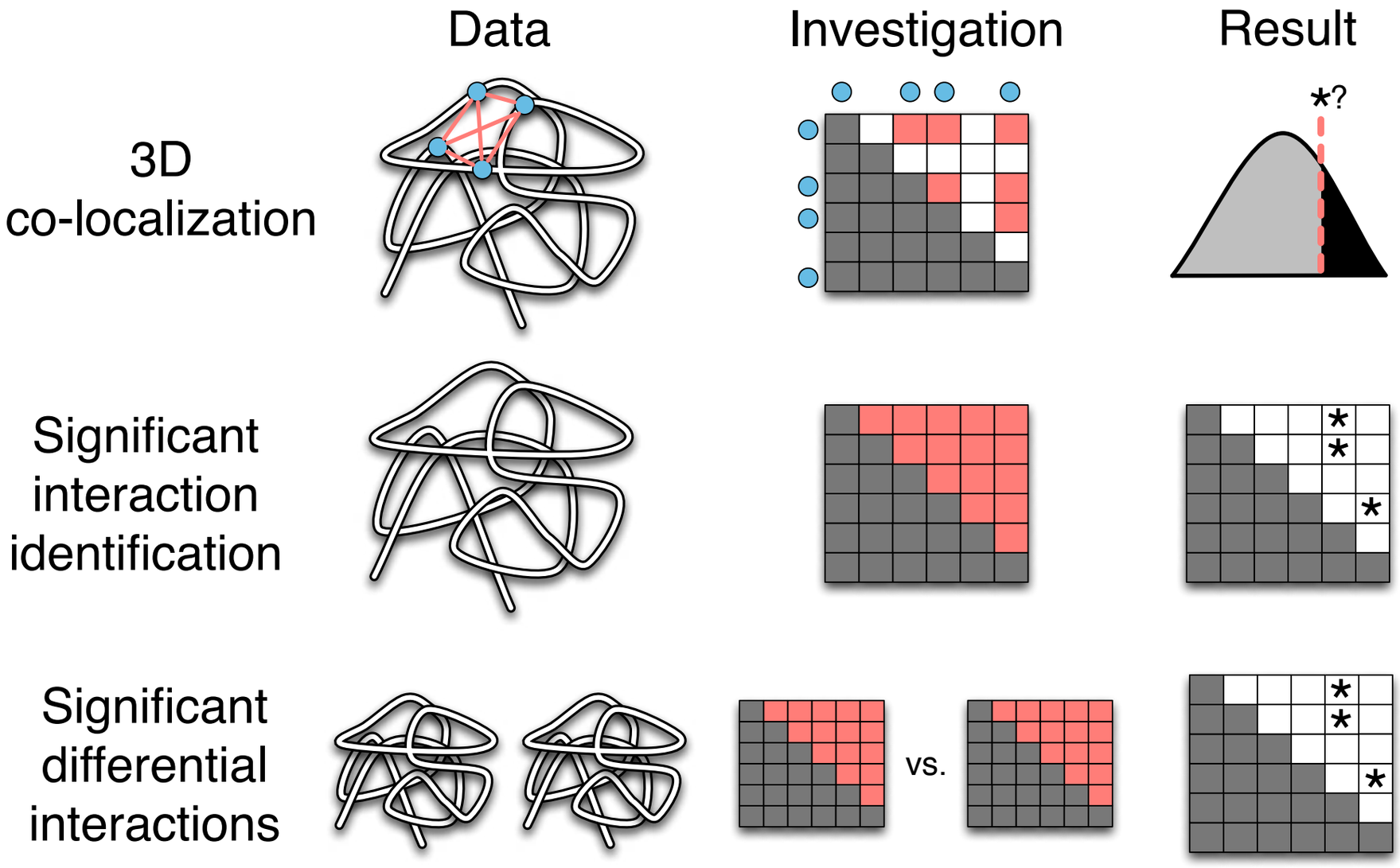




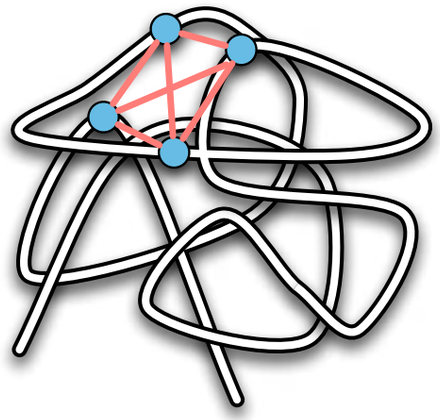
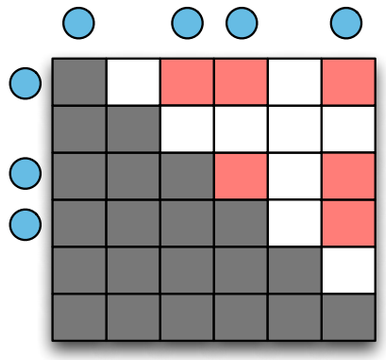
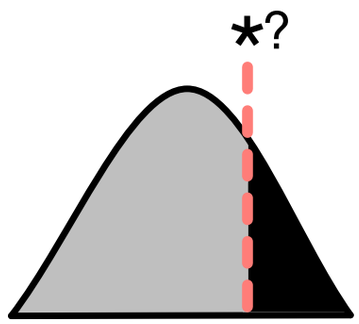
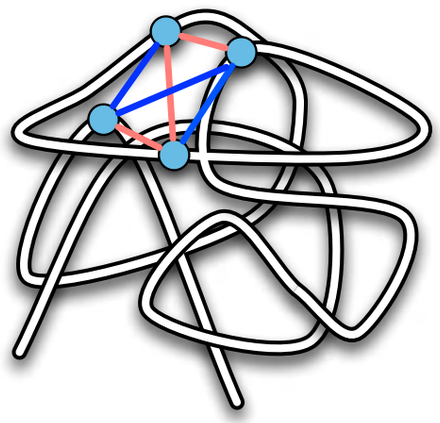
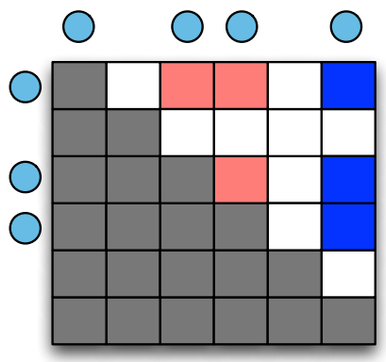
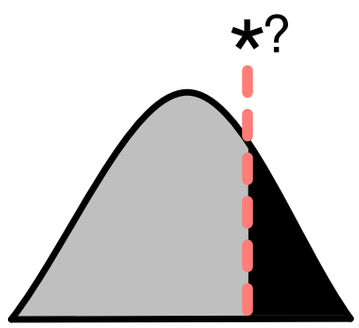
- Database of publicly available Hi-C datasets

Species	Cell-line/tissue	Treatment	Bin-size(s)	Ref
Human	GM06990		100k, 200k, 500k, 1M	[2]
Human	K562		100k, 200k, 500k, 1M	[2]
Human	GM12878		100k, 200k, 500k, 1M	[57]
Human	hESC		100k, 200k, 500k, 1M	[4]
Human	IMR90		100k, 200k, 500k, 1M	[4]
Human	RWPE1	ERG	200k, 500k, 1M	[28]
Human	RWPE1	GFP	200k, 500k, 1M	[28]
Mouse	mESC		100k, 200k, 500k, 1M	[4]
Mouse	cortex		100k, 200k, 500k, 1M	[4]
Mouse	pre-pro-B		100k, 200k, 500k, 1M	[30]
Mouse	pro-B		100k, 200k, 500k, 1M	[30]
D. melanogaster	embryo		10k, 20k, 40k, 80k, 160k	[58]
A. thaliana	Col	WT	200k, 500k, 1M	[31]
A. thaliana	Col	atmorc6-1	200k, 500k, 1M	[31]

Statistical analysis of Hi-C data



Types of 3D co-localization

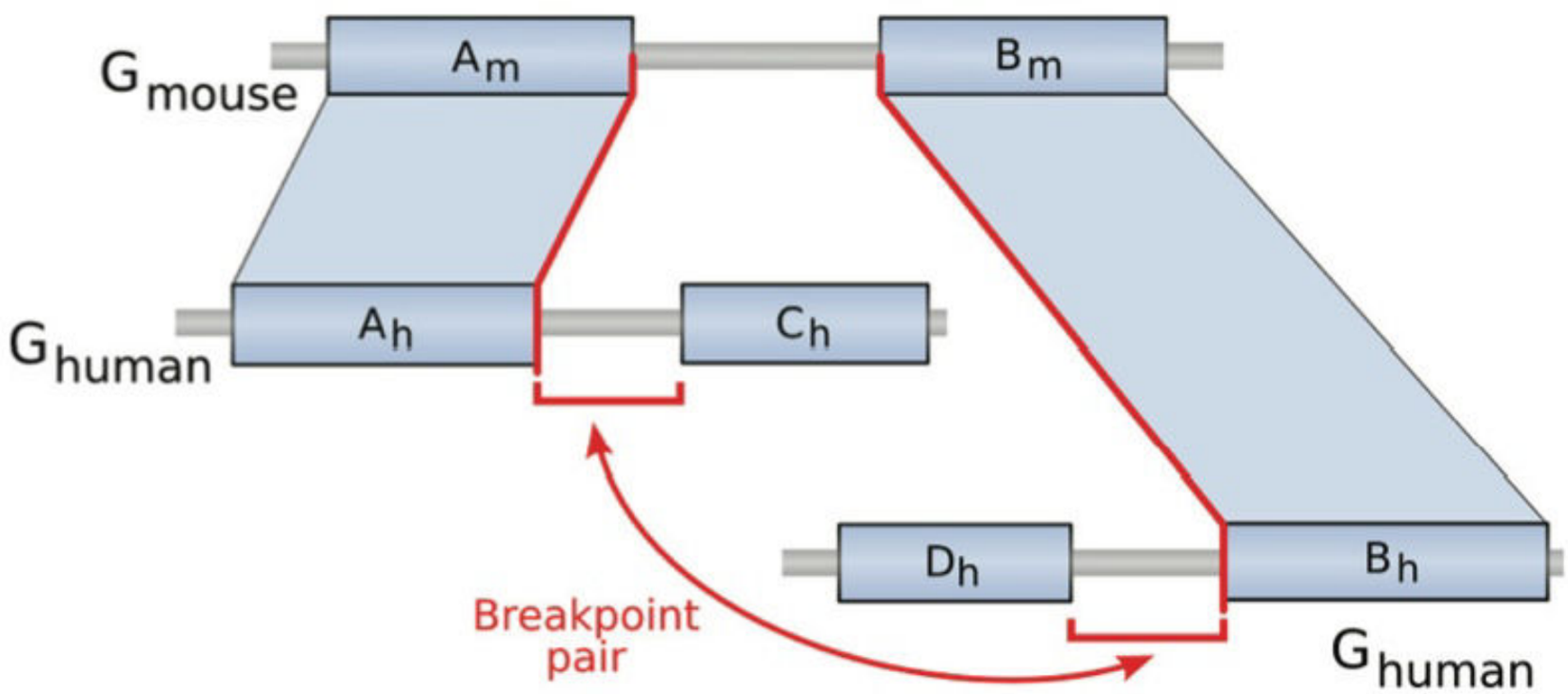
	Data	Investigation	Result
All-vs-all			
Focused			

Hands-on:
Are evolutionary breakpoints between human and mouse co-localized in 3D?

- Try to replicate some of the findings in Véron et al. (2011)

```
chr1 59012767 59120353 pair1
chr1 67610044 67630390 pair1
chr1 59012767 59120353 pair2
chr9 27300620 27325091 pair2
chr1 93835340 93905158 pair3
chr4 10001    533317   pair3
```

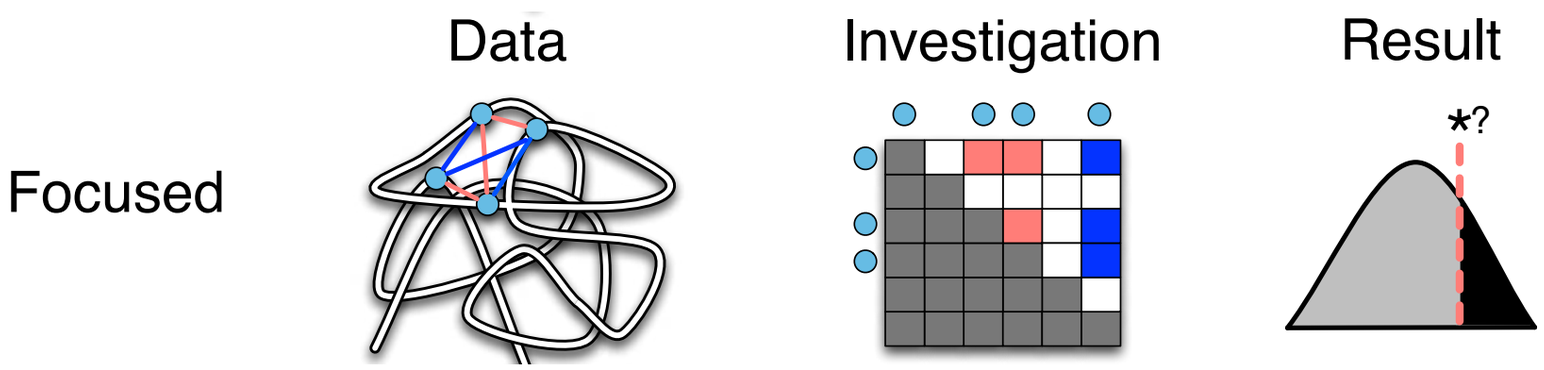
•••



Hands-on:
Are evolutionary breakpoints between human and mouse co-localized in 3D?

- Upload data from Véron et al. (2011)
- Convert to linked GTrack
- Hypothesis testing of 3D co-localization of linked elements

- GM06990 was used in Véron et al (2011).
- Try out same analysis, also on different cell-lines



Hands-on 3:

Are evolutionary breakpoints between human and mouse co-localized in 3D?

Go to:

- https://hyperbrowser.uio.no/3d/**u/gcc2014/p/breakpoints**

Demo

What was seen in Véron et al. (2011)?

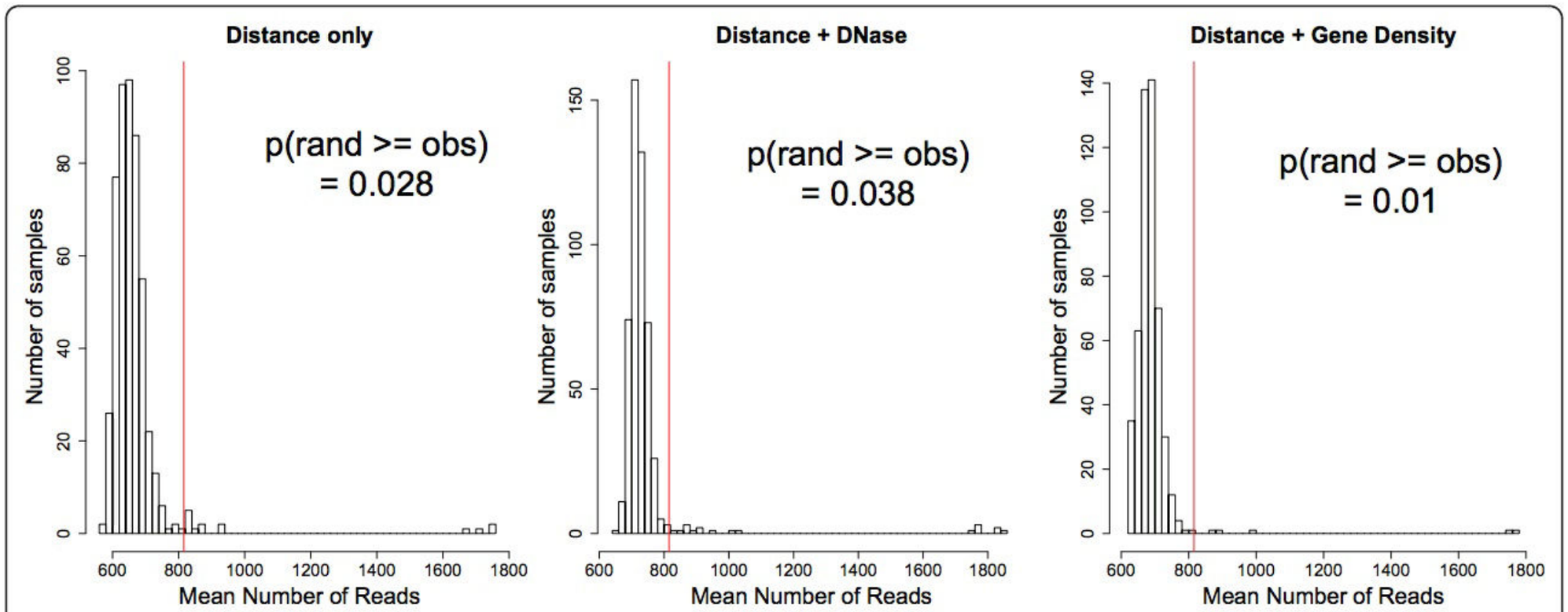


Figure 3 Locus pairs containing a breakpoint pair have more reads than expected. Histogram of values obtained by sampling 500 times the pairwise read count data, using pairs without breakpoints but at the same distance (left), same distance and same DNaseI sensitivity (middle), same distance and same gene density (right) as those in the breakpoint sets. The red vertical bar show the value for the actual breakpoint set.

Figure from Véron et al. (2011)

More details on hypothesis testing of 3D co-localization

The test statistic

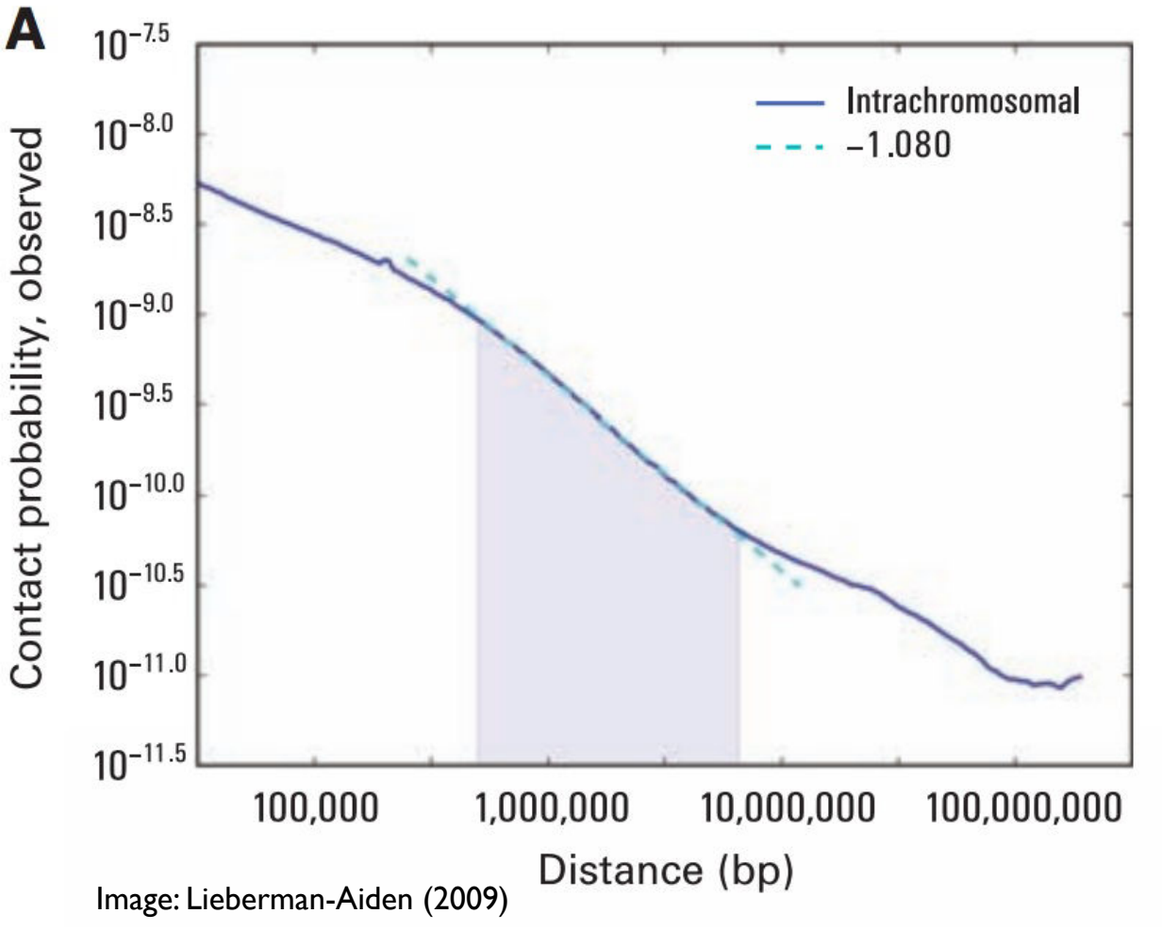
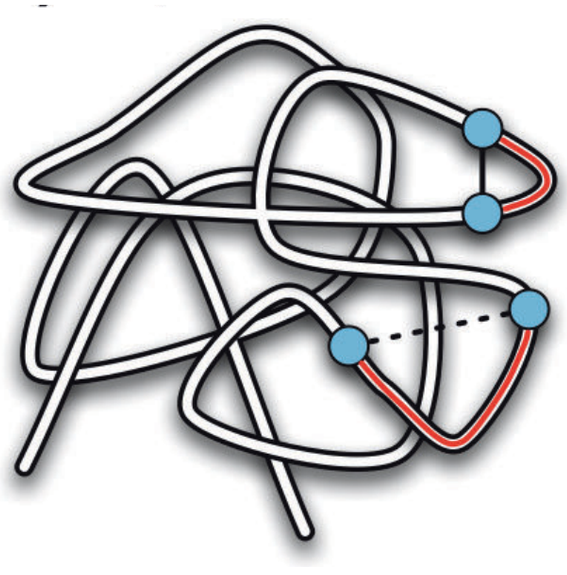


Image: Lieberman-Aiden (2009)



5164–5174 *Nucleic Acids Research, 2013, Vol. 41, No. 10* *Published online 9 April 2013*
doi:10.1093/nar/gkt227

Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements

Jonas Paulsen[1], Tonje G. Lien[2], Geir Kjetil Sandve[3,4], Lars Holden[5], Ørnulf Borgan[2], Ingrid K. Glad[2] and Eivind Hovig[1,3,6,*]

[1]Section for Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, PO Box 4950, Nydalen, N-0424 Oslo, Norway, [2]Department of Mathematics, University of Oslo, PO Box 1053, Blindern, 0316 Oslo, Norway, [3]Department of Informatics, University of Oslo, PO Box 1080, Blindern, 0316 Oslo, Norway, [4]Centre for Cancer Biomedicine, Faculty of Medicine, University of Oslo, PO Box 4950, Nydalen, 0424 Oslo, Norway, [5]Statistics for Innovation, Norwegian Computing Center, 0314 Oslo, Norway and [6]Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, PO Box 4950, Nydalen, N-0424 Oslo, Norway

Genomic distance correction

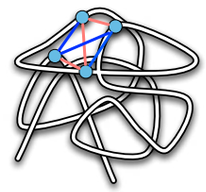
$$m^*_{a_i b_j} = \frac{m_{a_i b_j} - \hat{E}(m|\delta)}{\widehat{sd}(m|\delta)}$$

Test statistic

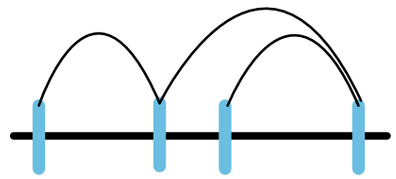
$$t = \frac{1}{M} \sum_{a_i, b_j \in Q} m^*_{a_i b_j}$$

Permutation test

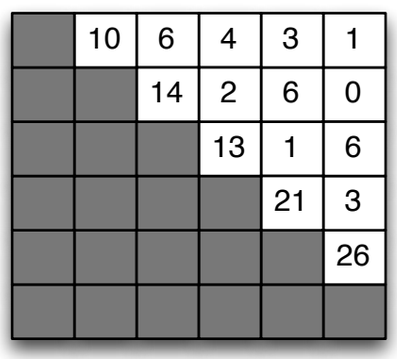
Focused

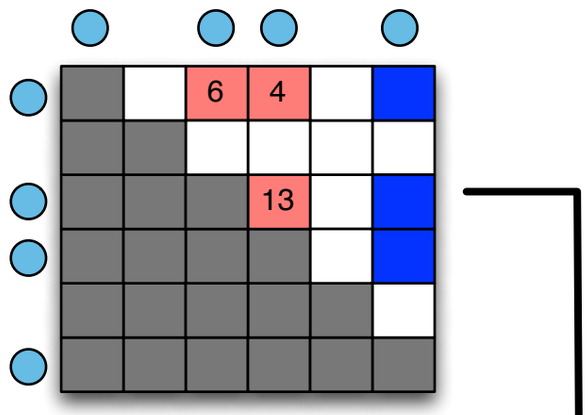


Track 1: Linked Points

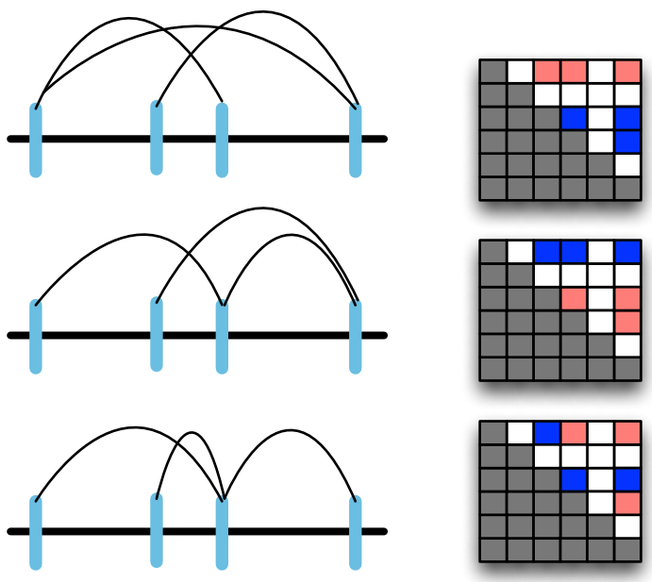


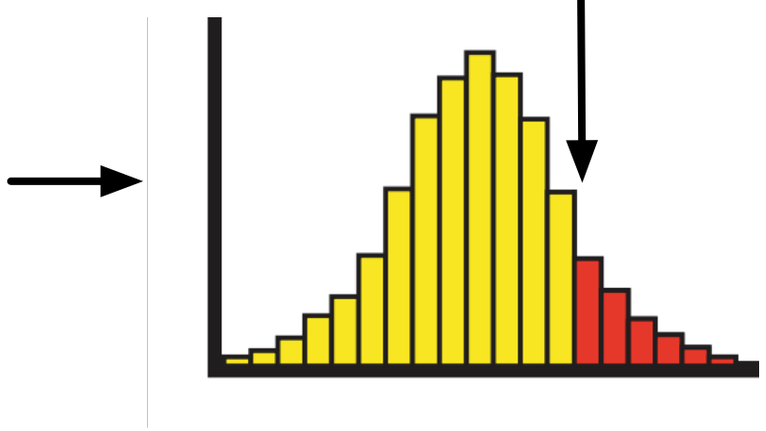
Track 2: Hi-C





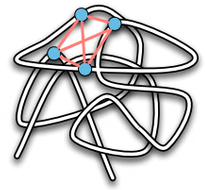
Permutation:



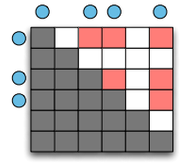


Permutation test

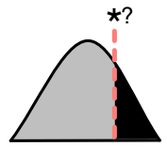
All-vs-all



Data



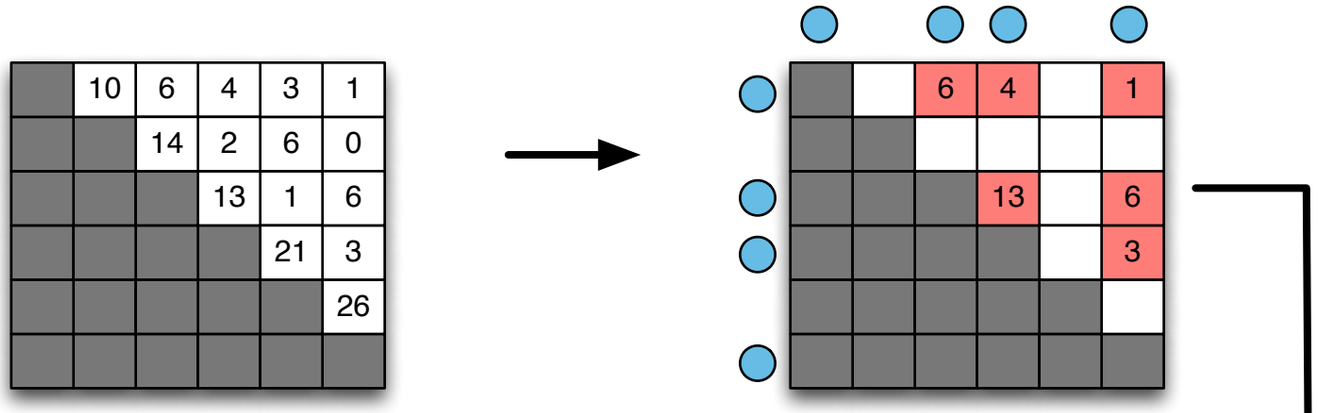
Investigation



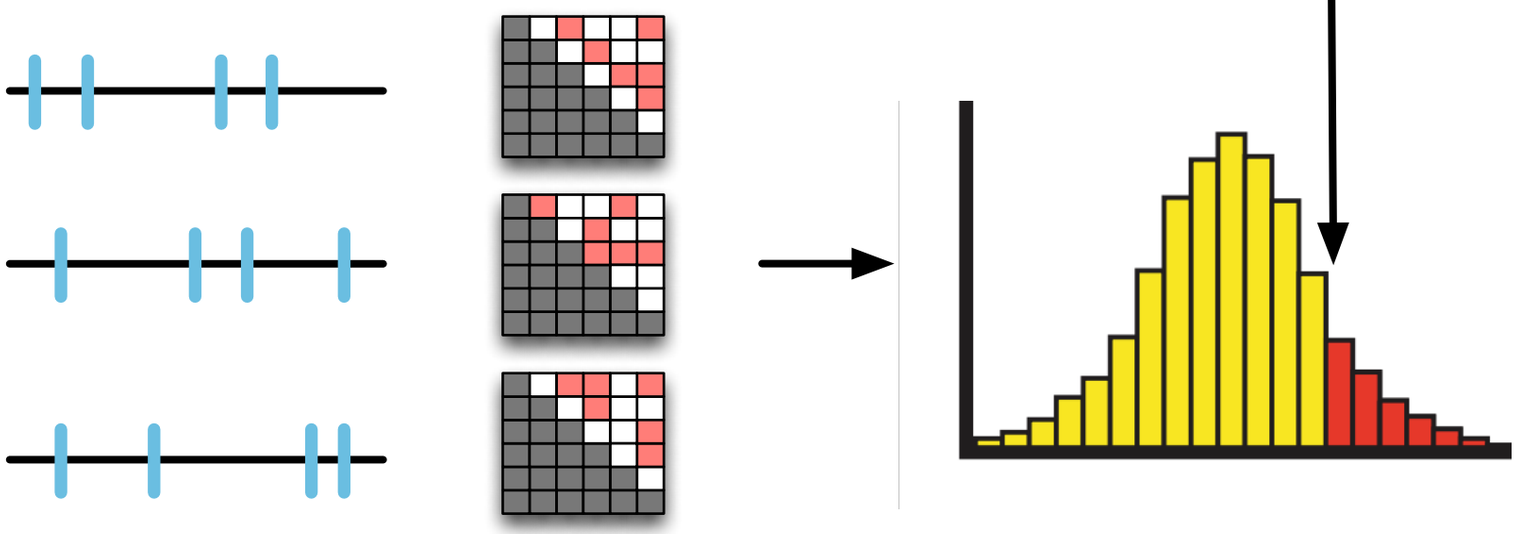
Result

Track 1: Points

Track 2: Hi-C

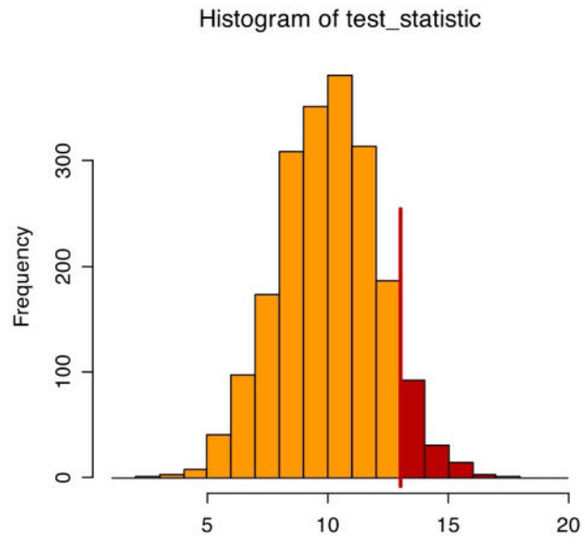


Permutation:



Descriptive statistic

A P-value in itself is not enough:



Enrichment score:

indicates the amount of 3D co-localization

Ratio of observed co-localization compared to expected.

=1 : Neither more nor less
>1 : More than expected
<1 : Less than expected

e.g. 1.1 = 10% higher

5164–5174 *Nucleic Acids Research, 2013, Vol. 41, No. 10* *Published online 9 April 2013*
doi:10.1093/nar/gkt227

Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements

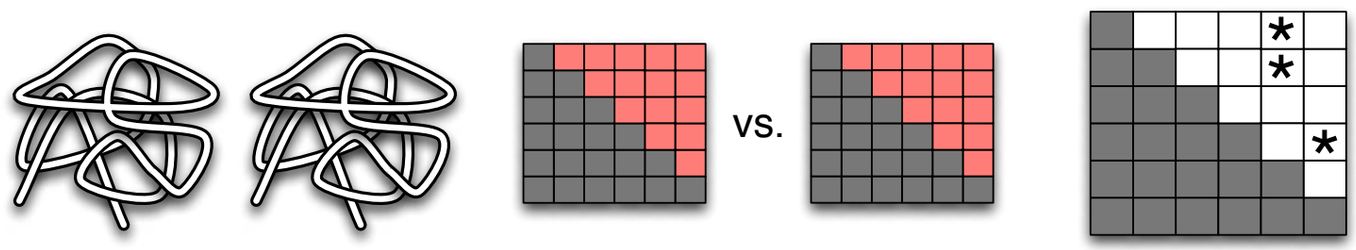
Jonas Paulsen[1], Tonje G. Lien[2], Geir Kjetil Sandve[3,4], Lars Holden[5], Ørnulf Borgan[2], Ingrid K. Glad[2] and Eivind Hovig[1,3,6,*]

[1]Section for Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, PO Box 4950, Nydalen, N-0424 Oslo, Norway, [2]Department of Mathematics, University of Oslo, PO Box 1053, Blindern, 0316 Oslo, Norway, [3]Department of Informatics, University of Oslo, PO Box 1080, Blindern, 0316 Oslo, Norway, [4]Centre for Cancer Biomedicine, Faculty of Medicine, University of Oslo, PO Box 4950, Nydalen, 0424 Oslo, Norway, [5]Statistics for Innovation, Norwegian Computing Center, 0314 Oslo, Norway and [6]Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, PO Box 4950, Nydalen, N-0424 Oslo, Norway

Identification of differential Hi-C interactions

- Two treatments
- Biological/technical replicates
- Where are the differences between the two treatments?
- "Similar" to differential gene expression
- Analytical test: Assume interaction counts follow a negative binomial distribution over replicates

Significant differential interactions



Hands-on:
Finding differential 3D interactions between normal and prostate cancer cells

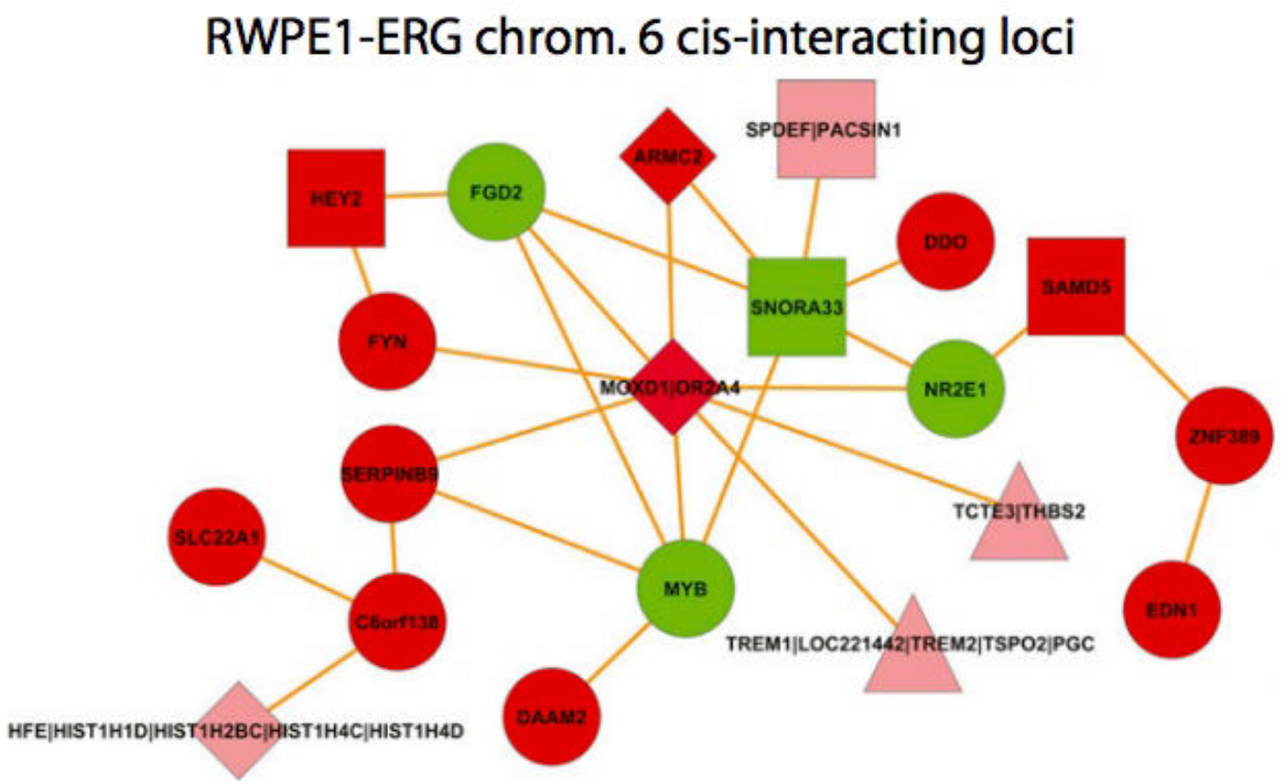
- Try to replicate some of the findings in Rickman et al. (2012)
    - Upload data from Rickman et al. (2012) (four biological replicates for each sample)
    - Identify significant differences between Hi-C datasets from prostate cancer cells (RWPE1-ERG) and normal prostate cells (RWPE1-GFP)
    - Visualize results using network and circos plots
    - chr6

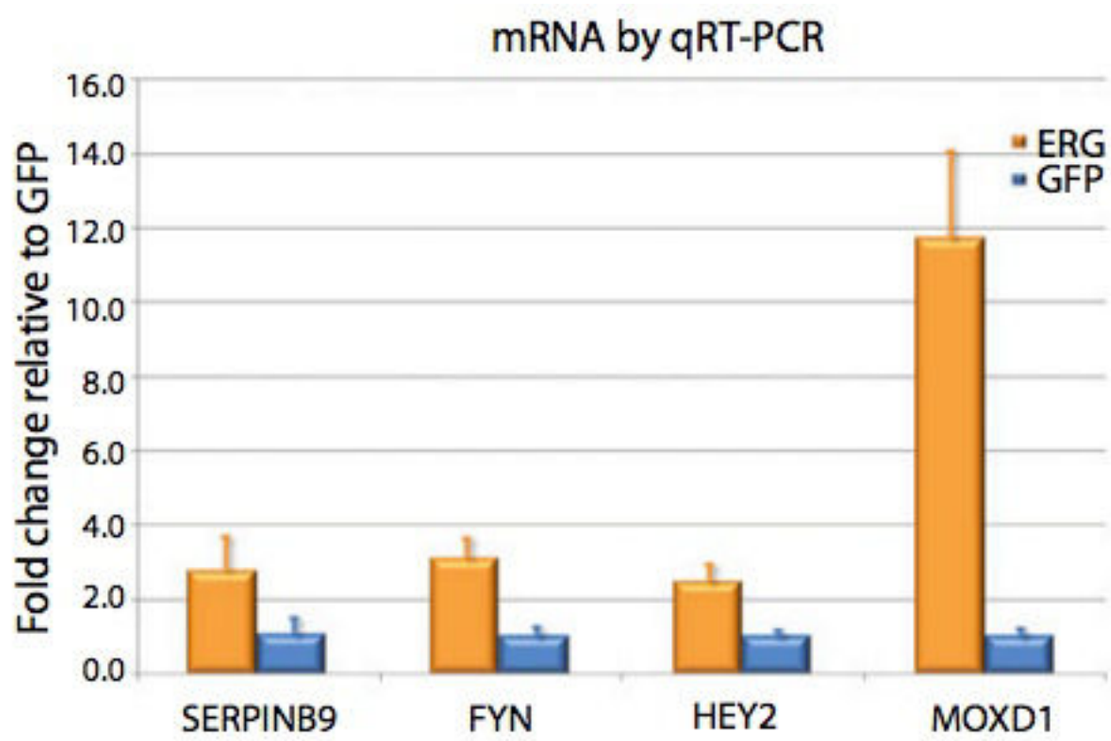
Hands-on 4:
Finding differential 3D interactions
between normal and prostate cancer cells

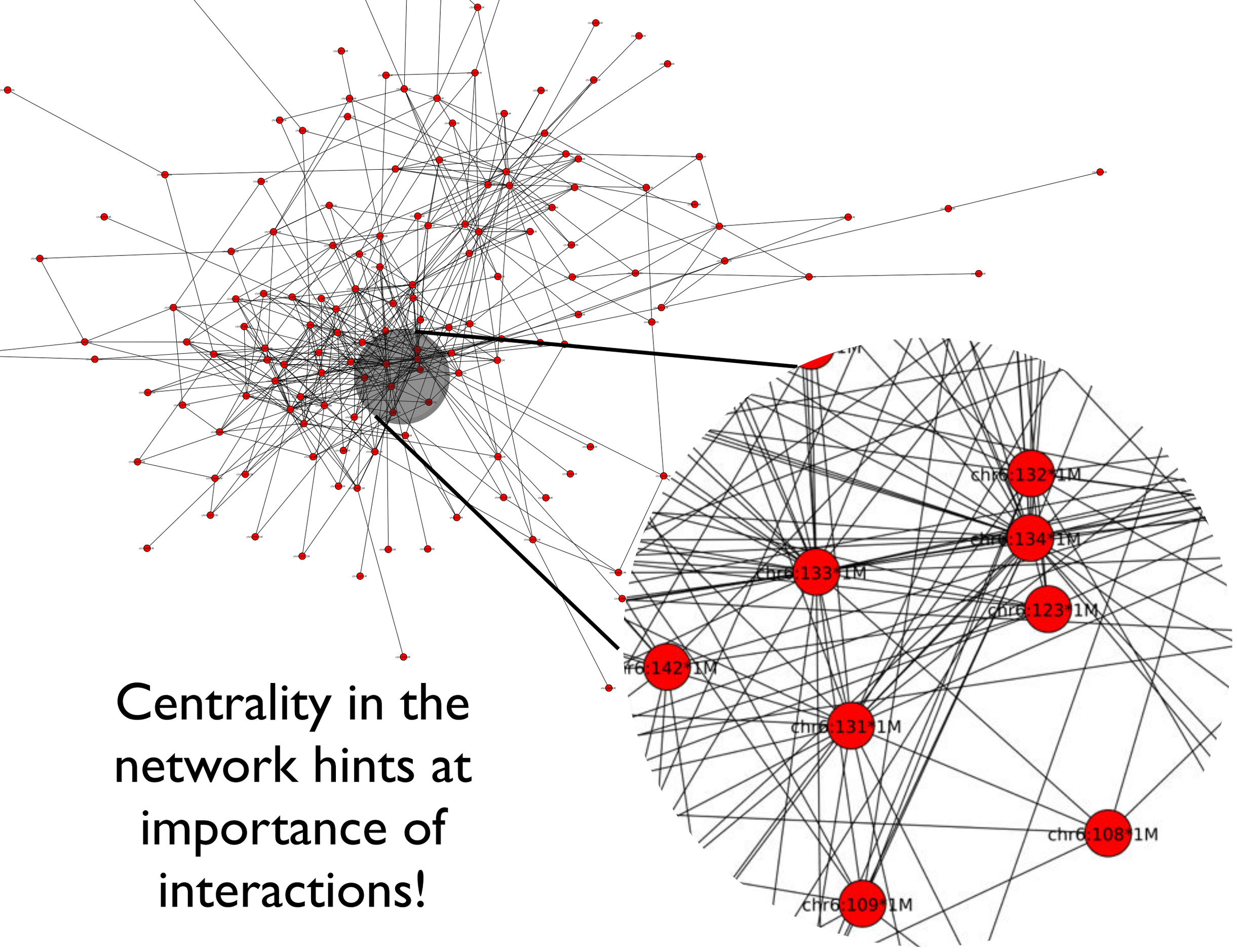
Go to:

- https://hyperbrowser.uio.no/3d/**u/gcc2014/p/differential**

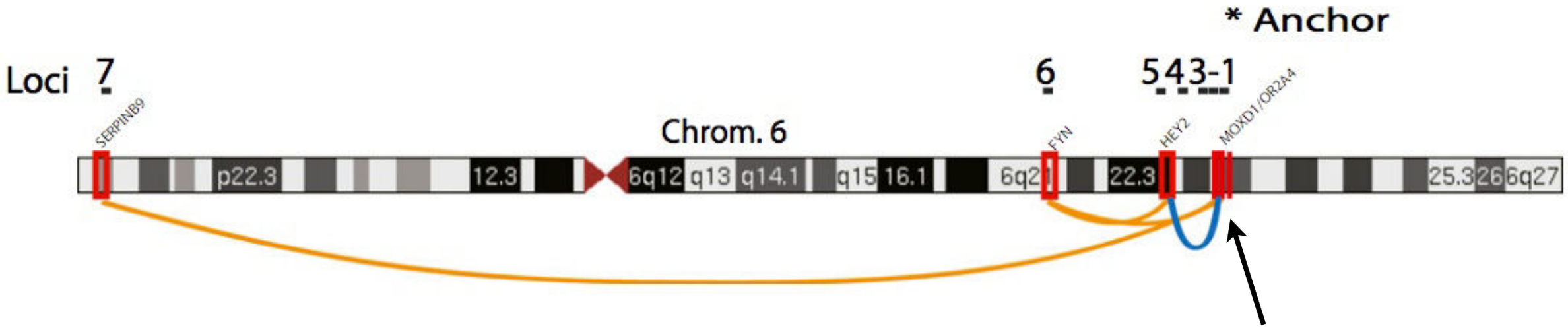
What was seen in Rickman et al. (2012)?



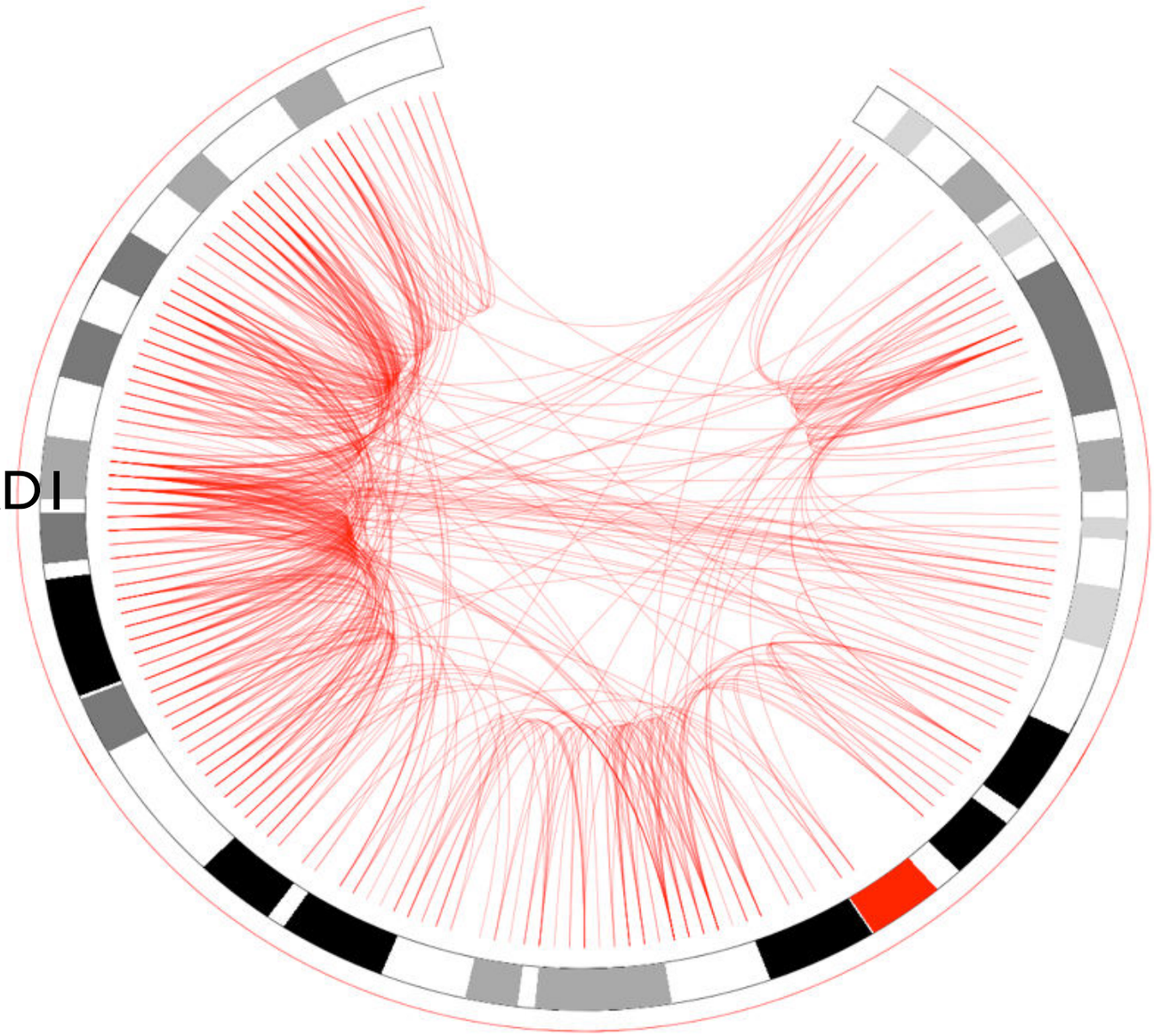




Centrality in the network hints at importance of interactions!



chr6:133*1M



MOXD1

Conclusions

What have we learned?

- Genomic 3D organization
- Track types + combination of pairs of tracks
- hyperbrowser.uio.no/3d
  - Uploading + visualization of Hi-C data
  - 3D co-localization
  - Differential interactions
  - Descriptive/visualization analysis

Reproducibility

- All the analyses that we did are all documented in your histories - and can easily scrutinized, changed, re-run.