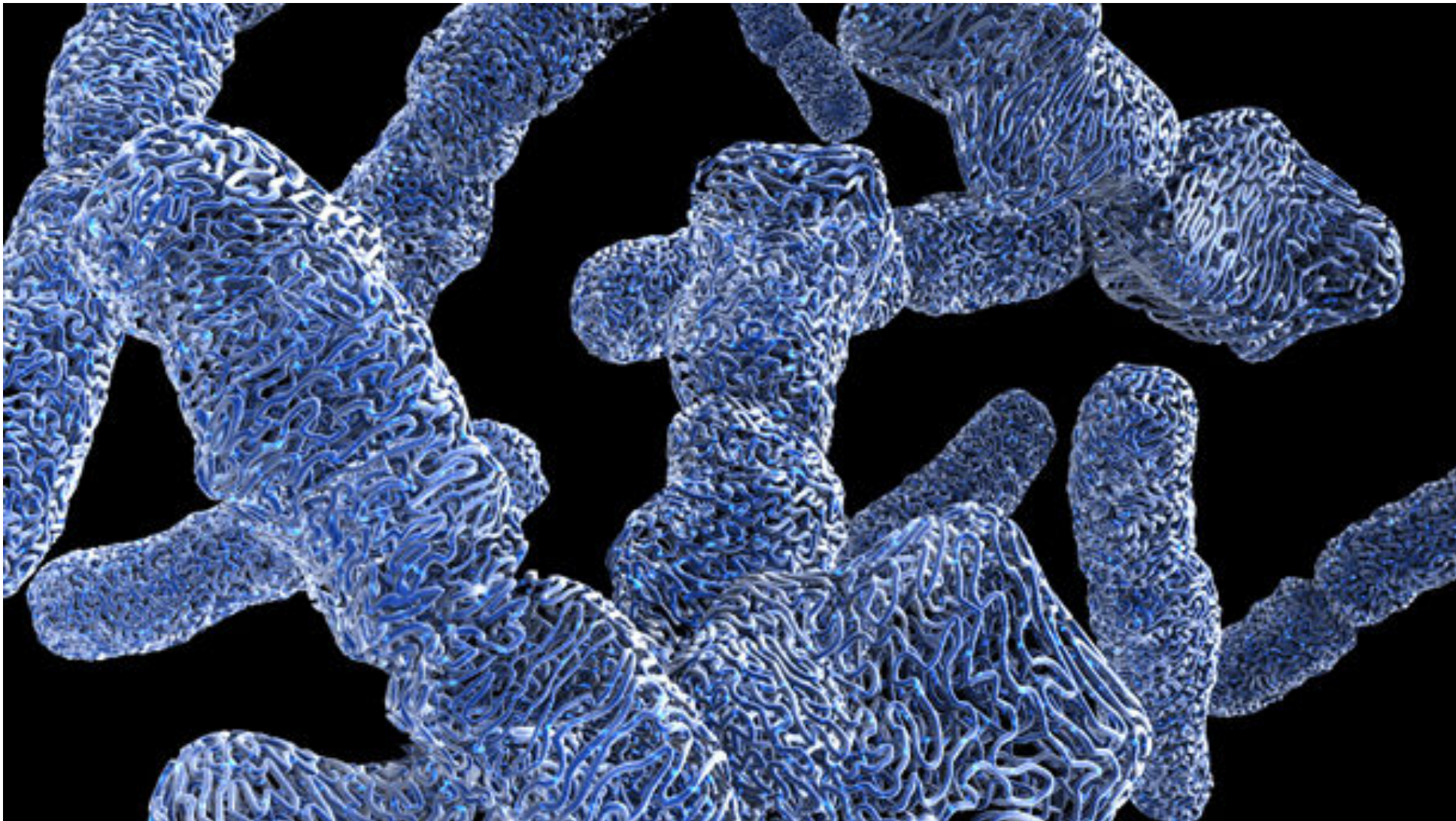
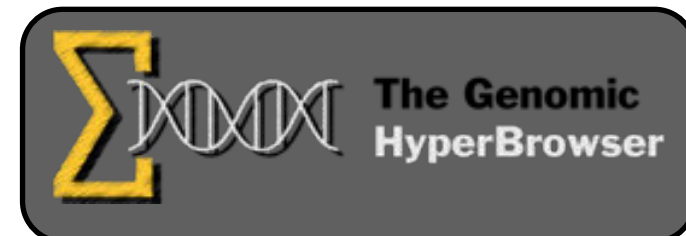


Analyzing 3D chromatin data in a Galaxy framework



Jonas Paulsen
(1 July 2013)



The team

Tonje G. Lien
Ingrid Glad
Lars Holden
Marit Holden
Ørnulf Borgan
Arnoldo Frigessi

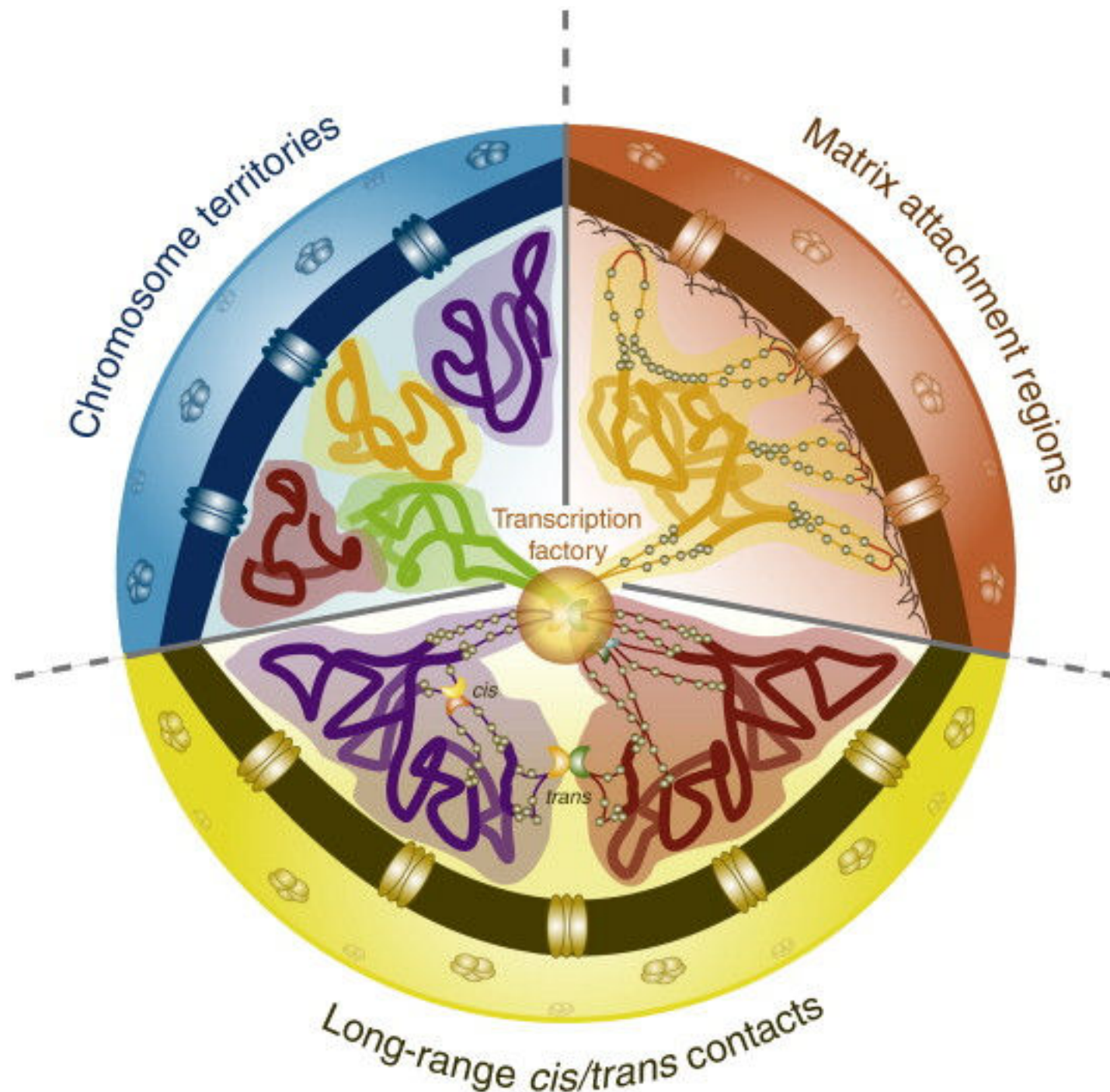
Eivind Hovig
Geir Kjetil Sandve
Kai Trengereid
Sveinung Gundersen
Jonas Paulsen



Goals

- Create new statistical methods for use with 3D genome data
- Develop user-friendly tools for researchers

Why is this interesting?

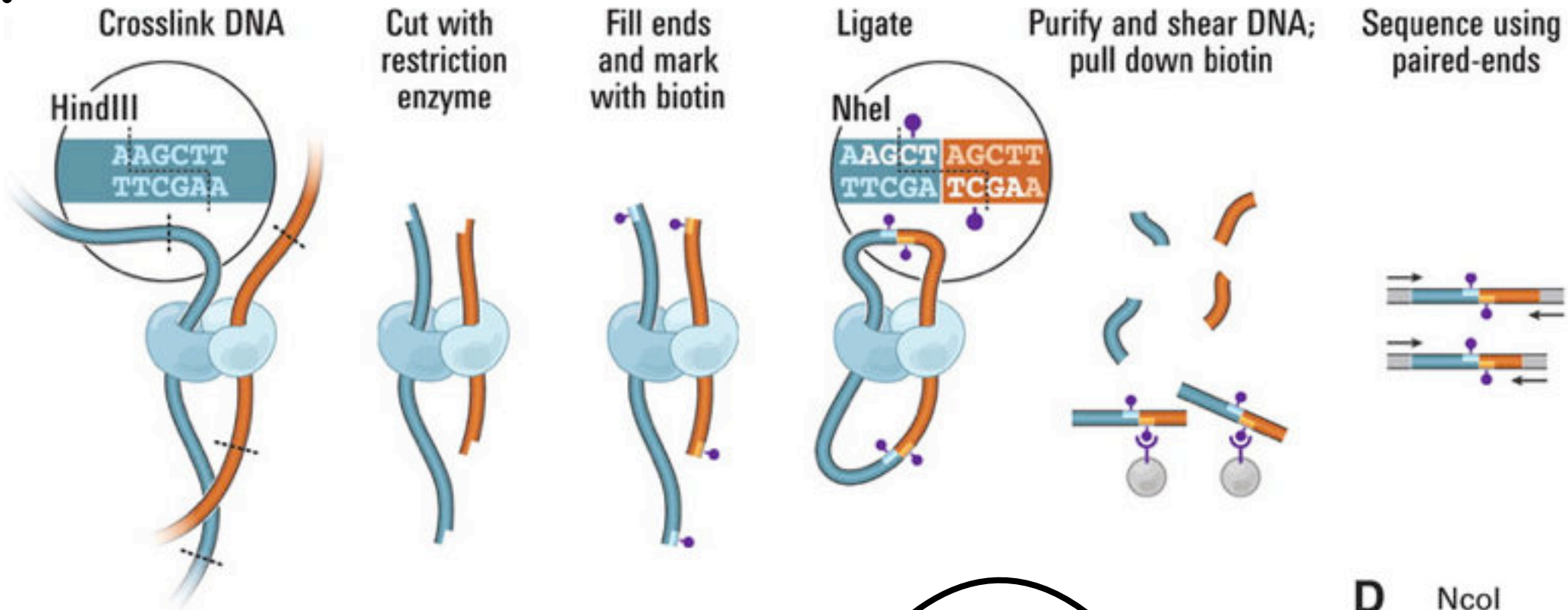


(Ethier et al. 2012)

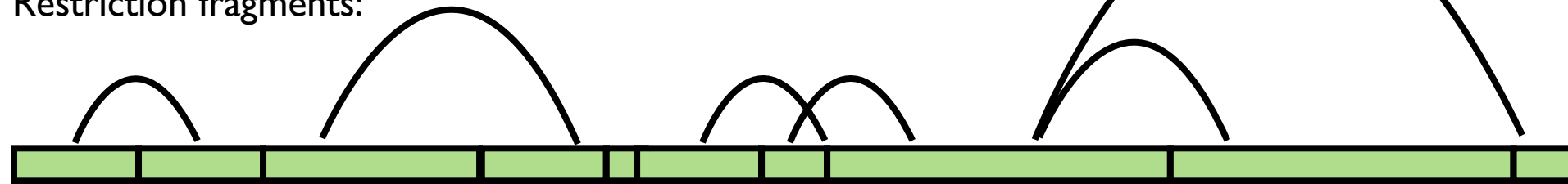
<http://www.sciencedirect.com/science/article/pii/S1874939911002227>

The data

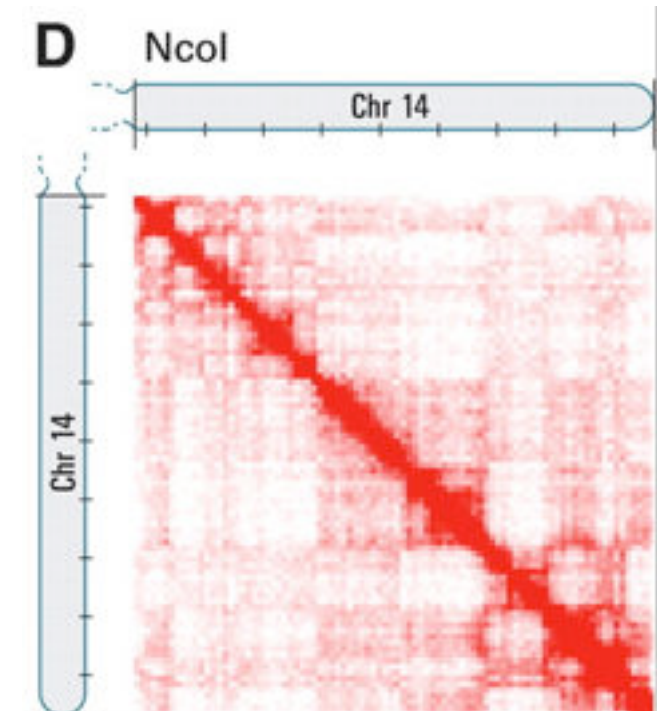
Hi-C:



Restriction fragments:



Equally-sized bins:



Different categories of questions

Different categories of questions

- “Query-set”

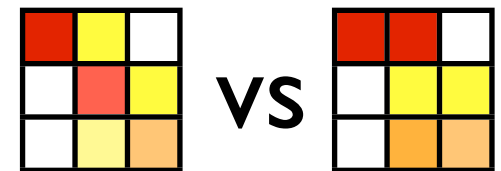


Different categories of questions

- “Query-set”



- Difference between treatments

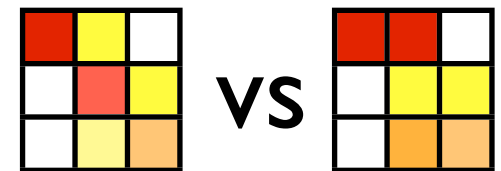


Different categories of questions

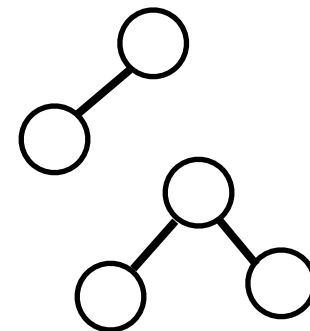
- “Query-set”



- Difference between treatments

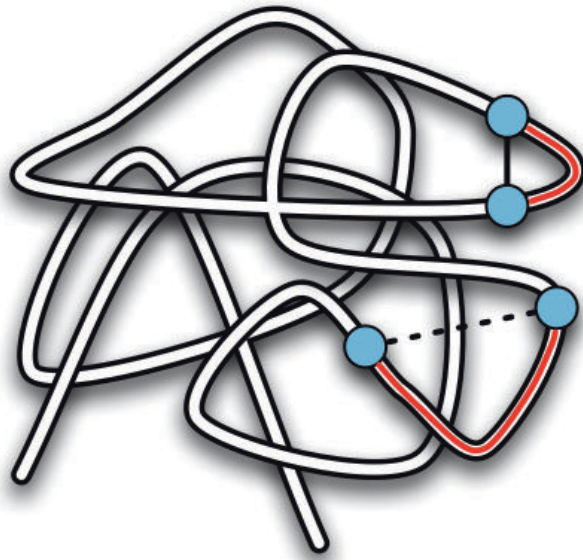


- Descriptive statistics

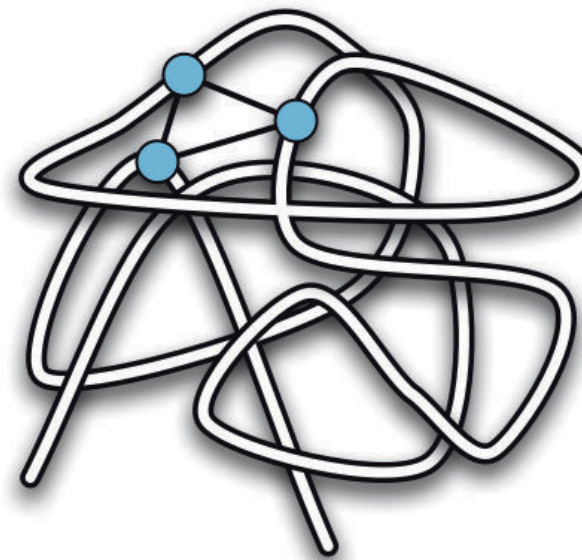


Complex data

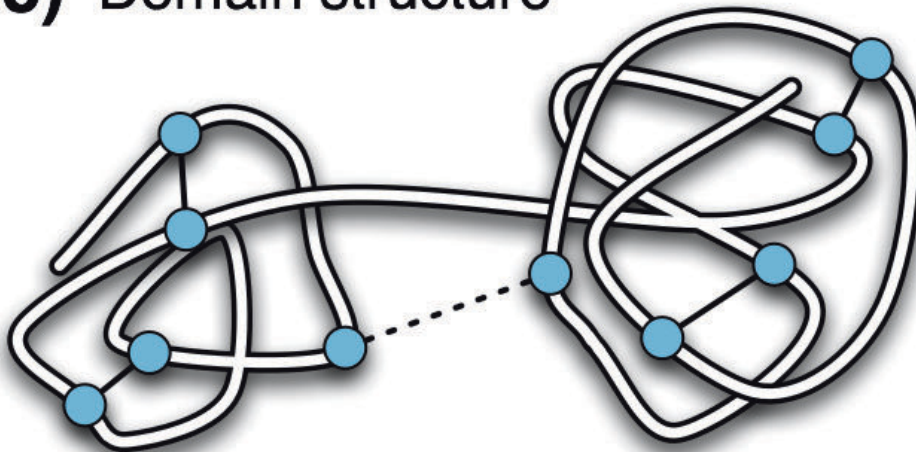
a) Sequence-based distance



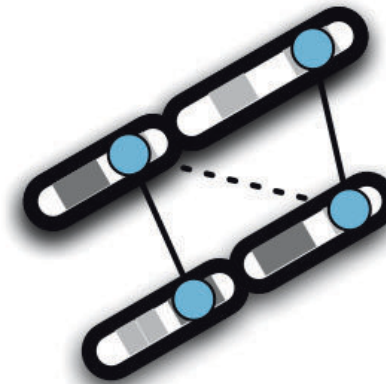
b) Transitivity relations



c) Domain structure

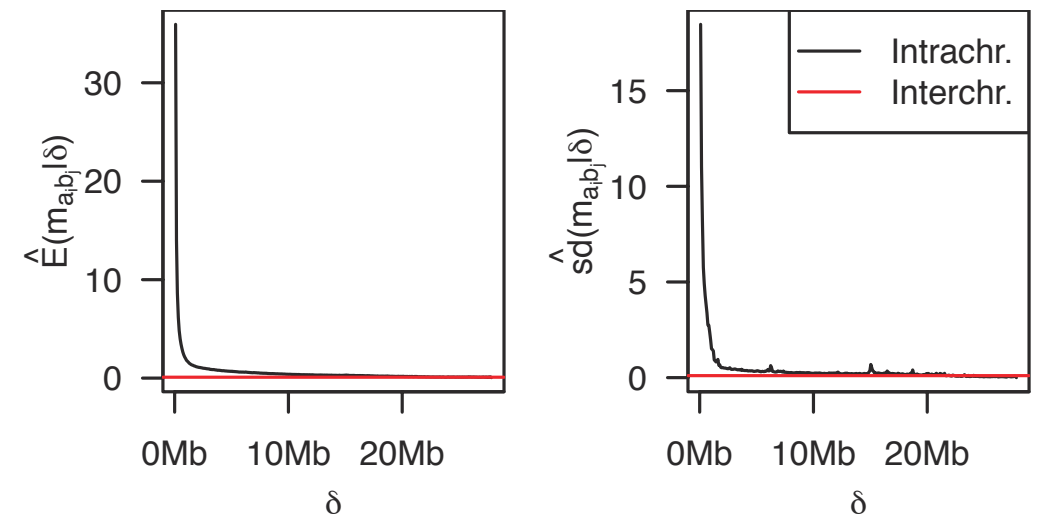


d) Regional preferences



Strategy

- Subtract background signal from the data
- Permutation test
- Dependencies are taken into account in the permutation



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

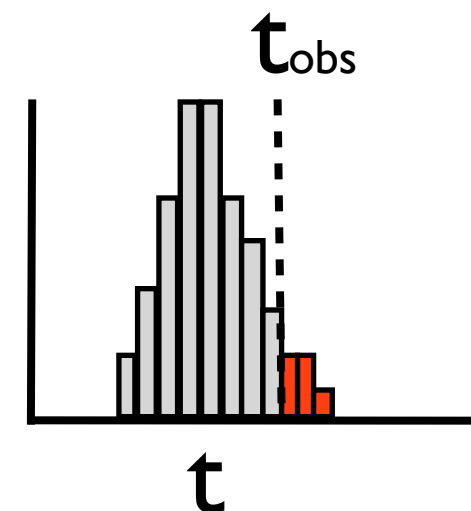
$$t = \frac{1}{M} \sum_{a_i, b_j \in Q} m_{a_i b_j}^*$$

Paulsen et al. NAR (2013):

5164–5174 Nucleic Acids Research, 2013, Vol. 41, No. 10
doi:10.1093/nar/gkt1227 Published online 9 April 2013

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

Jonas Paulsen¹, Tonje G. Lien², Geir Kjetil Sandve^{3,4}, Lars Holden⁵, Ørnulf Borgan², Ingrid K. Glad² and Eivind Hovig^{1,3,6,*}



Effect size - enrichment score

- Ratio of observed over expected interaction frequencies
- Exp is estimated using two components: sequenced based distance and the signal coming from domain etc. properties

$$\bar{m}_Q = \sum_{a_i, b_j \in Q} w_\delta \cdot m_{a_i b_j} = \frac{1}{\sum 1/\hat{\sigma}_\delta^2} \sum_{a_i, b_j \in Q} \frac{1}{\hat{\sigma}_\delta^2} m_{a_i b_j}.$$

$$Exp = \bar{\hat{E}}_Q + \frac{1}{R} \sum_{r=1}^R B_{Q_r}.$$

$$\bar{\hat{E}}_Q = \sum_{a_i, b_j \in Q} w_\delta \cdot \hat{E}(m|\delta) = \frac{1}{\sum 1/\hat{\sigma}_\delta^2} \sum_{a_i, b_j \in Q} \frac{1}{\hat{\sigma}_\delta^2} \hat{E}(m|\delta),$$

$$B_{Q_r} = \bar{m}_{Q_r} - \bar{\hat{E}}_{Q_r},$$

$$\underline{S = \bar{m}_Q / Exp},$$



The Genomic HyperBrowser

v1.6 (powered by Galaxy)

[Analyze Data](#)
[Workflow](#)
[Shared Data](#)
[Visualization](#)
[Admin](#)
[Help](#)
[User](#)
Using 1.4 Gb

Tools

Options

HYPERBROWSER ANALYSIS

Statistical analysis of tracks

- Analyze genomic tracks

Visual analysis of tracks

Specialized analysis of tracks

Text-based analysis interface

3D ANALYSIS

3D tools

- Convert from category BED to linked GTrack
- Convert from two category BED to case/control linked GTrack
- Analyze spatial colocalization of track elements (in 3D)
- Enrichment of colocalization of track elements (in 3D)
- colocalization between two point tracks tool
- Find significant difference between two replicated 3d datasets

HYPERBROWSER TRACK PROCESSING

HyperBrowser track repository

Customize tracks

Generate tracks

Format and convert tracks

Export and import tracks

GTrack tools

ARTICLE/DOMAIN-SPECIFIC TOOLS

The differential disease regulome

MCDDR

Transcription factor analysis

Gene tools

HYPERBROWSER INTERNAL TOOLS

Admin of genomes and tracks

Development tools

Assorted tools

STANDARD GALAXY TOOLS

Get Data



The Genomic HyperBrowser

History

Options

Trans-splicing 4.2 Mb

55: HyperBrowser: 'List of non-adjusted and adjusted 3D contact frequencies' on 'K562-all-1M (Inter- and intrachromosomal)' vs 'Linked fusion genes (6)'

54: HyperBrowser: 'List of non-adjusted and adjusted 3D contact frequencies' on 'IMR90-all-1M (Inter- and intrachromosomal)' vs 'Linked fusion genes (6)'

53: HyperBrowser: 'List of non-adjusted and adjusted 3D contact frequencies' on 'hESC-all-1M (Inter- and intrachromosomal)' vs 'Linked fusion genes (6)'

52: HyperBrowser: 'List of non-adjusted and adjusted 3D contact frequencies' on 'GM06990-all-1M (Inter- and intrachromosomal)' vs 'Linked fusion genes (6)'

Notice

This is an specialized version of The Genomic HyperBrowser providing analysis of 3D co-localization of genomic elements. The system is currently in development. For other uses, we recommend the stable [3D co-localization version of the HyperBrowser](#). The user and history database is not the same between the different versions.

If you have a *genomic track*, this is the place to analyze it!

To analyze a track, simply:

- Click [Statistical analysis of tracks: Analyze genomic tracks](#) in the left-hand menu.
- Select tracks from your Galaxy history or browse our collection. (To load a track to your history, click [Get data: Upload file](#))
- Select the analysis you are interested in:
 - any property of a single track
 - any relation between a pair of tracks

For help using the system:

- Click [The Genomic Hyperbrowser: Help](#) in the left-hand menu.
- Or, look through the following screencasts: (further screencasts are available from the help menu)





What is the Genomic HyperBrowser?

Getting started

Interface overview

Previously published version:

<http://hyperbrowser.uio.no/3d-coloc>

The Genomic HyperBrowser v1.6 (powered by Galaxy)

Analyze Data Workflow Shared Data Visualization Admin Help User Using 1.4 Gb

Tools Options

HYPERBROWSER ANALYSIS

Statistical analysis of tracks

- Analyze genomic tracks

Visual analysis of tracks

Specialized analysis of tracks

Text-based analysis interface

3D ANALYSIS

3D tools

- Convert from category BED to linked GTrack
- Convert from two category BED to case/control linked GTrack
- Analyze spatial colocalization of track elements (in 3D)
- Enrichment of colocalization of track elements (in 3D)
- colocalization between two point tracks tool
- Find significant difference between two replicated 3d datasets

HYPERBROWSER TRACK PROCESSING

HyperBrowser track repository

Customize tracks

Generate tracks

Format and convert tracks

Export and import tracks

GTrack tools

ARTICLE/DOMAIN-SPECIFIC TOOLS

The differential disease regulome

MCFDR

Transcription factor analysis

Gene tools

HYPERBROWSER INTERNAL TOOLS

Admin of genomes and tracks

Development tools

Assorted tools

STANDARD GALAXY TOOLS

Get Data

The Genomic HyperBrowser

Notice

This is a specialized version of The Genomic HyperBrowser providing analysis of 3D co-localization of genomic elements. The system is currently in development. For other uses, we recommend the stable [3D co-localization version of the HyperBrowser](#). The user and history database is not the same between the different versions.

If you have a genomic track, this is the place to analyze it!

To analyze a track, simply:

1. Click [Statistical analysis of tracks: Analyze genomic tracks](#) in the left-hand menu.
2. Select tracks from your Galaxy history or browse our collection. (To load a track to your history, click [Get data: Upload file](#))
3. Select the analysis you are interested in:
 - any property of a single track
 - any relation between a pair of tracks

For help using the system:

1. Click [The Genomic Hyperbrowser: Help](#) in the left-hand menu.
2. Or, look through the following screencasts: (further screencasts are available from the help menu)

What is the Genomic HyperBrowser? Getting started Interface overview

History Options

Trans-splicing 4.2 Mb

S5: HyperBrowser: 'List of non-adjusted and adjusted 3D contact frequencies' on 'K562-all-1M (Inter- and intrachromosomal)' vs 'Linked fusion genes (6)'

S4: HyperBrowser: 'List of non-adjusted and adjusted 3D contact frequencies' on 'IMR90-all-1M (Inter- and intrachromosomal)' vs 'Linked fusion genes (6)'

9.6 Kb
format: html, database: hg19
Info: Using all chromosomes of genome build "hg19" as bins

HTML file

S3: HyperBrowser: 'List of non-adjusted and adjusted 3D contact frequencies' on 'hESC-all-1M (Inter- and intrachromosomal)' vs 'Linked fusion genes (6)'

9.6 Kb
format: html, database: hg19
Info: Using all chromosome arms of genome build "hg19" as bins

HTML file

S2: HyperBrowser: 'List of non-adjusted and adjusted 3D contact frequencies' on 'GM06990-all-1M (Inter- and intrachromosomal)' vs 'Linked fusion genes (6)'

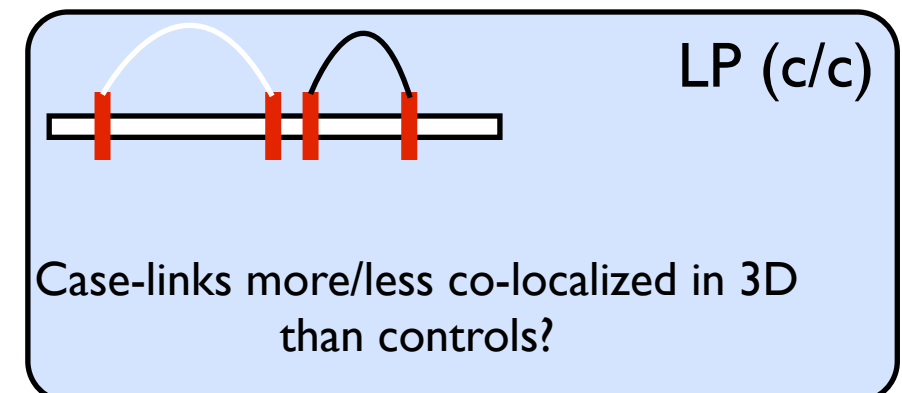
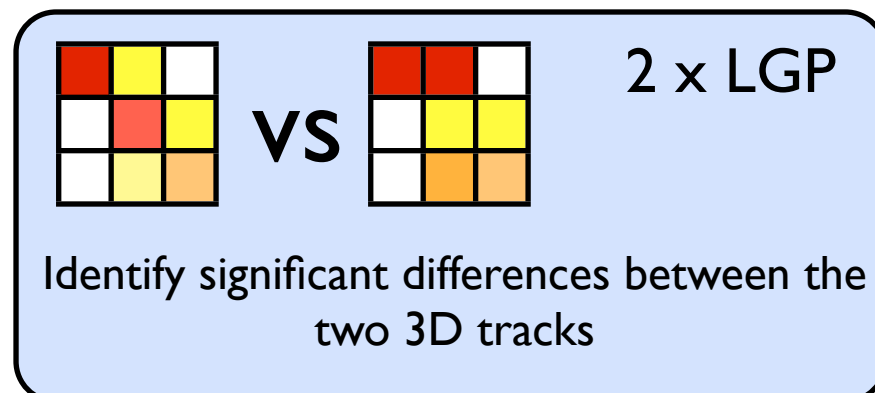
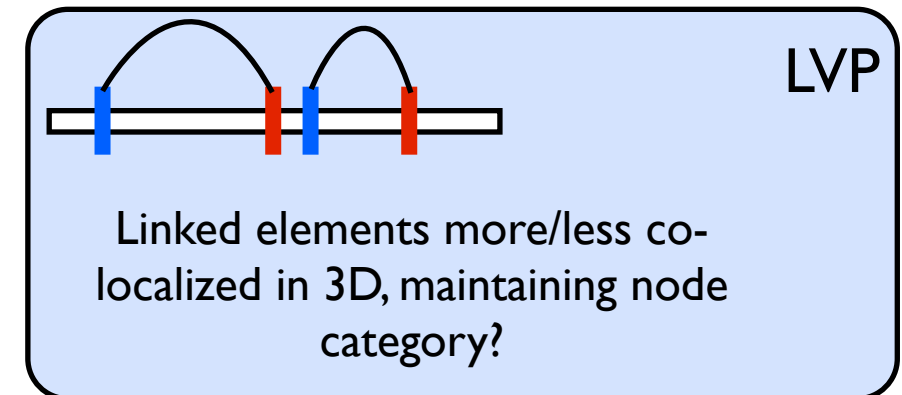
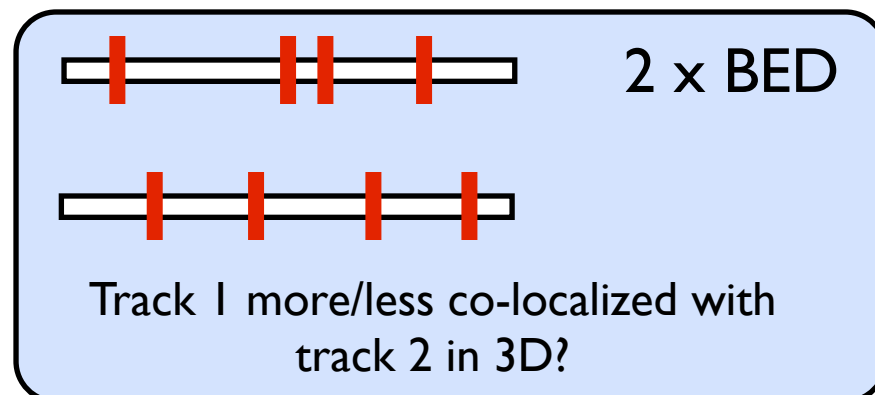
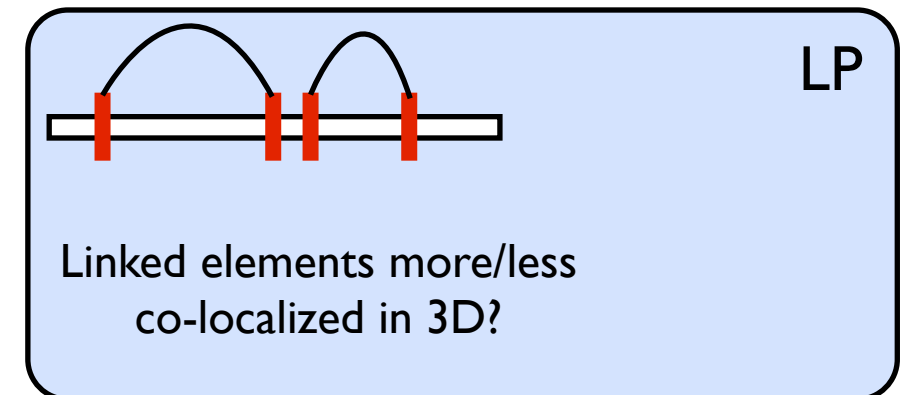
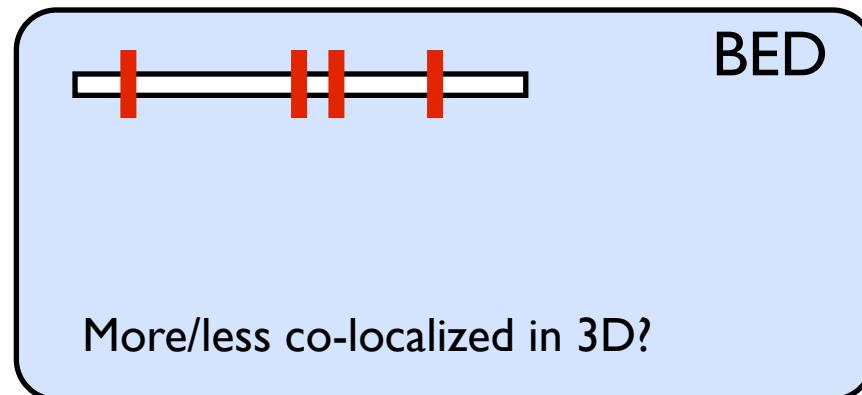
9.6 Kb
format: html, database: hg19
Info: Using all chromosomes of genome build "hg19" as bins

HTML file

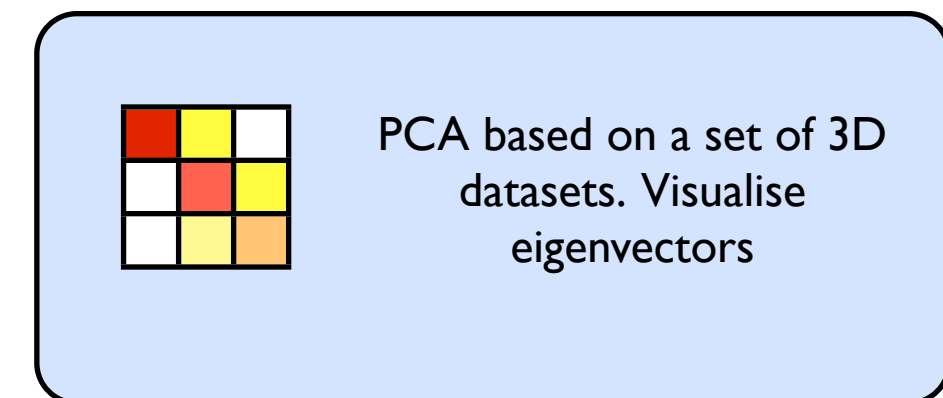
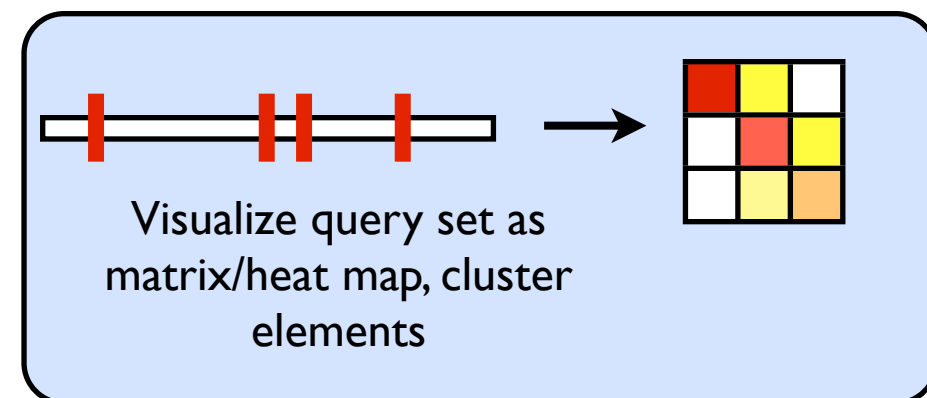
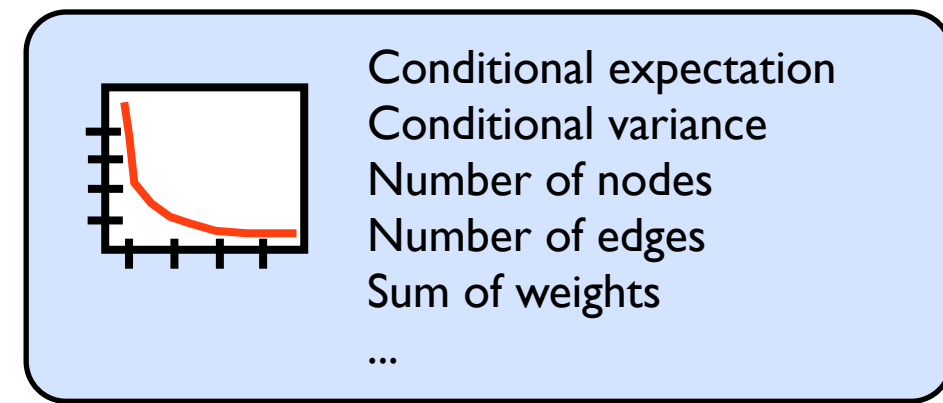
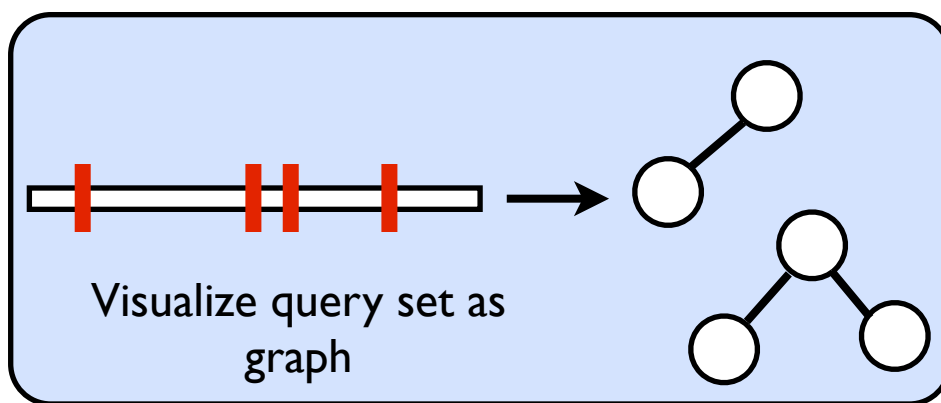
Previously published version:

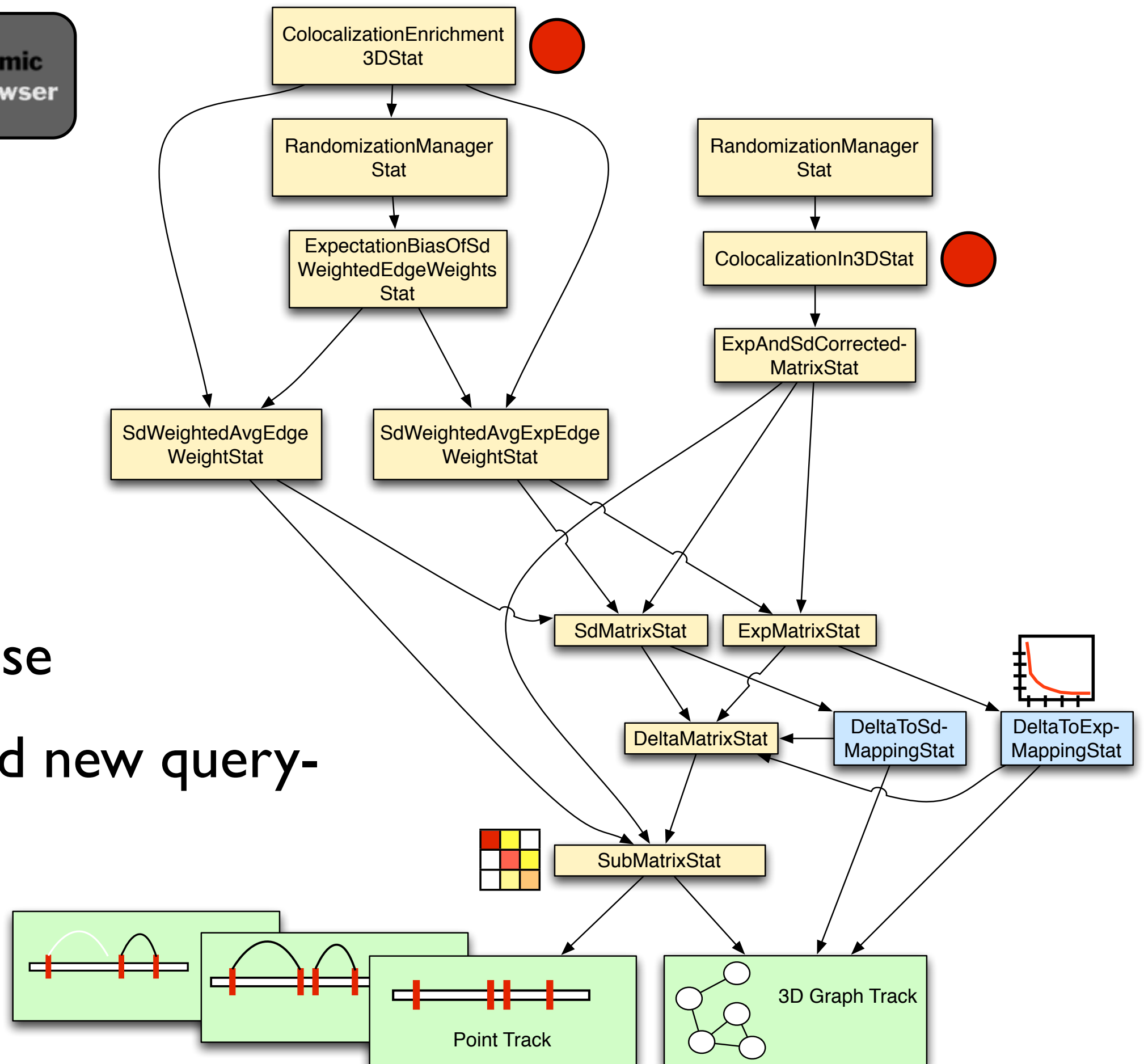
<http://hyperbrowser.uio.no/3d-coloc>

Available hypothesis tests



Descriptive statistics

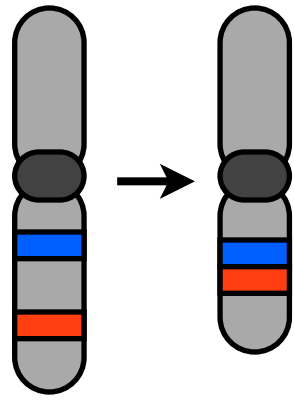




- Modular
- Code re-use
- Easy to add new query-tracks

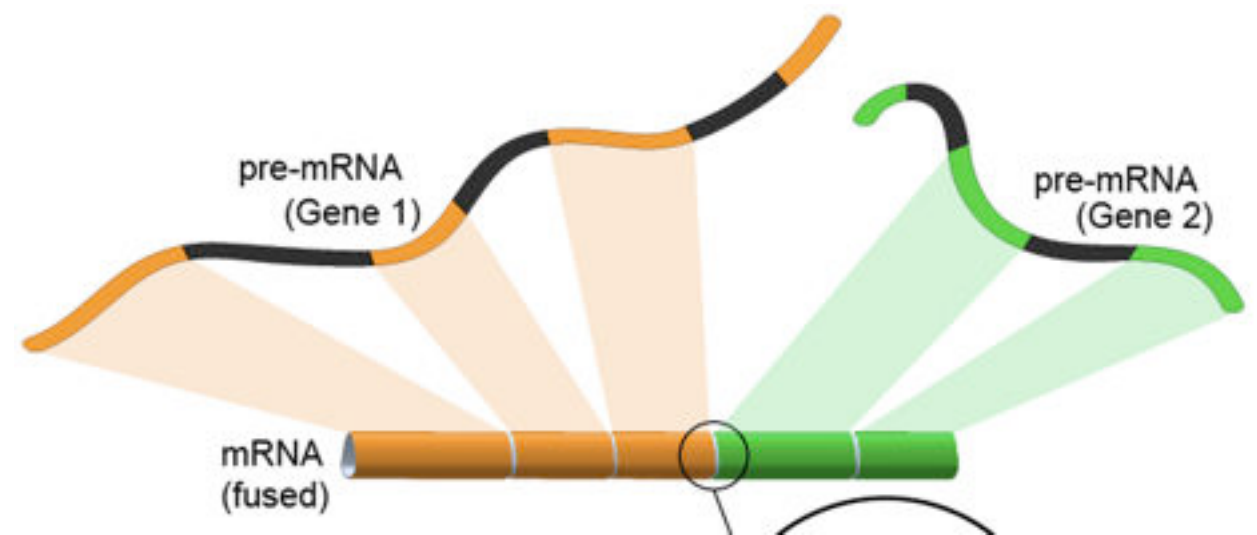
Example: Fusion transcripts

Structural



- Translocation
- Inversion
- Deletion

Post-transcriptional



- Read-through
- Trans-splicing

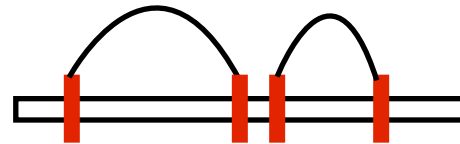
Example: Fusion transcripts

BED-file

chr1	151	152	A
chr1	121	122	A
chr3	312	313	B
chr3	613	614	B
chr1	151	152	C
chr5	521	522	C
	.		
	.		
	.		

Conversion tool

Linked Elements



chr1	151	152	0	1;4
chr1	121	122	1	0
chr3	312	313	2	3
chr3	613	614	3	2
chr5	521	522	4	1
	.			
	.			
	.			



Linked elements co-localized in 3D?

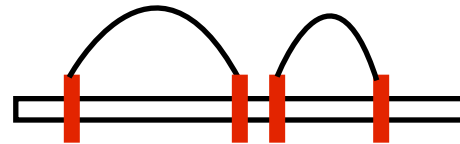
Example: Fusion transcripts

BED-file

```
chr1 151 152 A
chr1 121 122 A
chr3 312 313 B
chr3 613 614 B
chr1 151 152 C
chr5 521 522 C
.
```

Conversion tool

Linked Elements



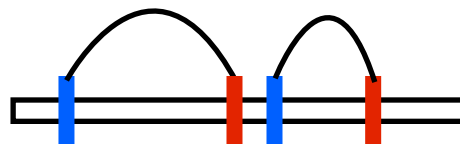
```
chr1 151 152 0 1;4
chr1 121 122 1 0
chr3 312 313 2 3
chr3 613 614 3 2
chr5 521 522 4 1
.
```



Linked elements co-localized in 3D?

```
chr1 151 152 A,0
chr1 121 122 A,1
chr3 312 313 B,1
chr3 613 614 B,0
chr1 151 152 C,0
chr5 521 522 C,1
.
```

Conversion tool



```
chr1 151 152 0 0 1;4
chr1 121 122 1 1 0
chr3 312 313 1 2 3
chr3 613 614 0 3 2
chr5 521 522 1 4 1
.
```




Linked elements co-localized in 3D,
maintaining
categories?

Convert from categorical BED file to linked GTrack

Select a specific genome?

Select Categorical BED file:

 [Corresponding batch command line:](#)

The Genomic HyperBrowser (v1.6)

Genome build: 

First Track



[What is a genomic track?](#)

Second Track

Analysis

Category: ?

Are the points linked by edges in 'Linked fusion genes (6)' closer in 3D (as defined by 'GM06990-all-1M (Inter- and intrachromosomal)') than expected by chance?

Track type

Treat 'GM06990-all-1M (Inter- and intrachromosomal)' as:

Treat 'Linked fusion genes (6)' as:

?

Options

Alternative hypothesis:

Null model:

[What is a null model?](#)

Minimal number of MC samples:

Maximal number of MC samples:

Sequential MC threshold (m):

MCFDR threshold on global P-value:

MCFDR threshold on FDR:

Result output:

[What do the MCFDR options mean?](#)

?

[Did you not find your question here?](#)

You asked:

Are the points linked by edges in 'Linked fusion genes' closer in 3D (as defined by 'GM06990-all-1M (Inter- and intrachromosomal)') than expected by chance?

Simplistic answer:

Yes – the data suggests this (p-value: 0.009901)

Precise answer:

The p-value is 0.009901 for the test

H0: The points of track 2 are located independently in 3D, as defined by track 1

vs

H1: The points of track 2 are located closer in 3D, as defined by track 1

Low p-values are evidence against H0.

The test was also performed for each bin separately, resulting in 0 significant bins out of 22, at 10% FDR* (2 bins excluded from FDR-analysis due to lacking p-values).

Please note that both the effect size and the p-value should be considered in order to assess the practical significance of a result.

* False Discovery Rate: The expected proportion of false positive results among the significant bins is no more than 10%.

P-values were computed under the **null model** defined by the following preservation and randomization rules:

Preserve 3D graph (T1) and Query graph (T2), randomize IDs in T2

The **test statistic** used is:

Main result of analysis

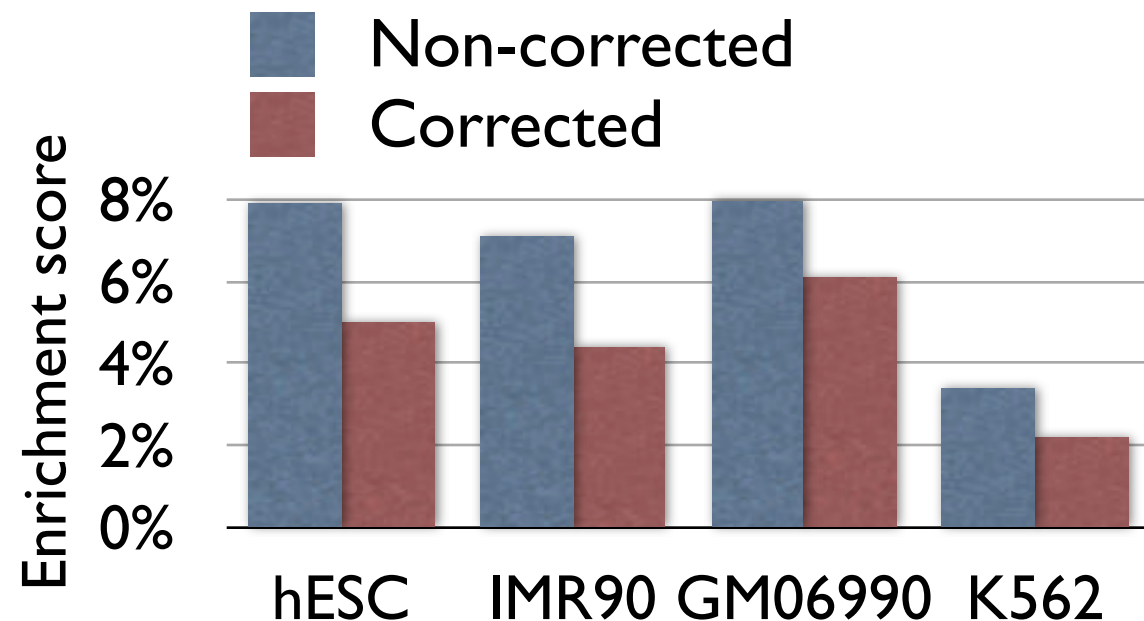
The value of the test statistic is 0.3773.

The p-values may be subject to further parameter choices, which are listed in the [run description](#).

[See full details](#) of the results in table form.

Results

- All 4 cell lines $P < 0.01$, even when correcting for domain architecture
- Enrichment goes down when correcting for chromatin domains, but is still present
- Some trans-spliced are proximal in several cell-lines



Top gene pairs:

Position1	Position2	Gene1	Gene2	GM06990	hESC	IMR90	K562
chr11:3*1M	chr9:135*1M	NUP214	INS-IGF2				
chr11:66*1M	chr16:90*1M	SPG7;RPL13	MALAT1				
chr11:66*1M	chr19:2*1M	MBD3	MALAT1				
chr11:66*1M	chr19:3*1M	NEAT1	DOT1L				
chr11:66*1M	chr19:5*1M	MAP2K2	MALAT1				
chr13:30*1M	chr9:36*1M	TMEM8B	MTUS2				
chr16:3*1M	chr11:65*1M	SRRM2	ATG2A				
chr16:3*1M	chr17:8*1M	TRAF7	EIF4A1				
chr16:30*1M	chr16:90*1M	CPNE7	BOLA2				
chr16:31*1M	chr19:5*1M	ORAI3	DPP9				
chr17:41*1M	chr16:2*1M	IFT140	ATP6V0A1				
chr17:41*1M	chr16:3*1M	PDPK1	CNTNAP1				
chr17:41*1M	chr19:3*1M	PSME3	LMNB2				
chr17:5*1M	chr16:30*1M	PFN1	BOLA2				
chr17:81*1M	chr11:66*1M	TBCD	NEAT1				
chr17:81*1M	chr19:4*1M	EEF2	CSNK1D				
chr17:81*1M	chr9:132*1M	SPTAN1;GOLGA2	GPS1;FN3KRP				
chr19:11*1M	chr11:66*1M	MALAT1	DNMT1;DNM2				
chr19:18*1M	chr17:74*1M	WBP2	UNC13A				
chr19:2*1M	chr1:2*1M	SSU72	CNN2				
chr19:46*1M	chr16:90*1M	RPL13	CLPTM1				
chr19:46*1M	chr9:132*1M	CERCAM	CD3EAP				
chr19:50*1M	chr17:74*1M	WBP2	FTL				
chr19:50*1M	chr6:32*1M	PPP1R15A	HSPA1A				
chr21:48*1M	chr16:16*1M	NDE1	LSS				
chr21:48*1M	chr16:2*1M	COL6A2	CACNA1H				
chr21:48*1M	chr19:3*1M	DOT1L	COL6A1				
chr3:50*1M	chr11:66*1M	NEAT1	DAG1				
chr3:50*1M	chr19:51*1M	QARS	AP2A1				
chr7:2*1M	chr19:2*1M	MAD1L1	DAZAP1				
chr7:45*1M	chr7:7*1M	EIF2AK1	DBNL				

Summary

- 3D structure of chromatin makes statistical models of data challenging
- We develop a range of tools, implemented in a Galaxy-framework (Hyperbrowser)
- Relatively simple to do complex analysis
- <http://hyperbrowser.uio.no/3d-coloc/>