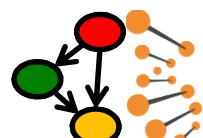


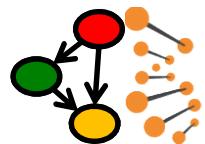
Lecture 14: Correlation vs. Conditional Dependence

Learning the human chromatin network
from all ENCODE ChIP-seq data



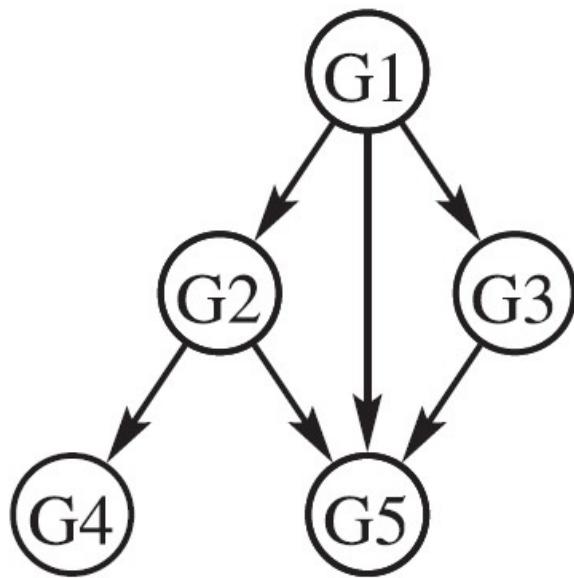
GENOME 560, Spring 2017

Scott Lundberg, CSE (slund1@cs.washington.edu)

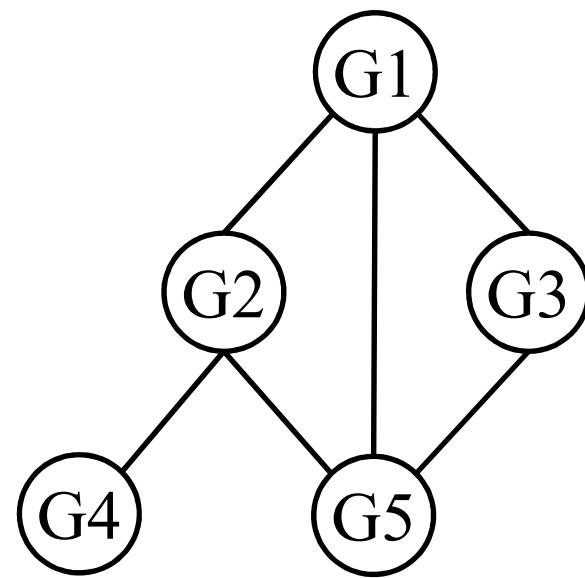


Directed vs. undirected graphical models

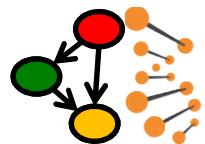
Directed graphical model
(Bayesian network)



Undirected graphical model
(Gaussian graphical model)

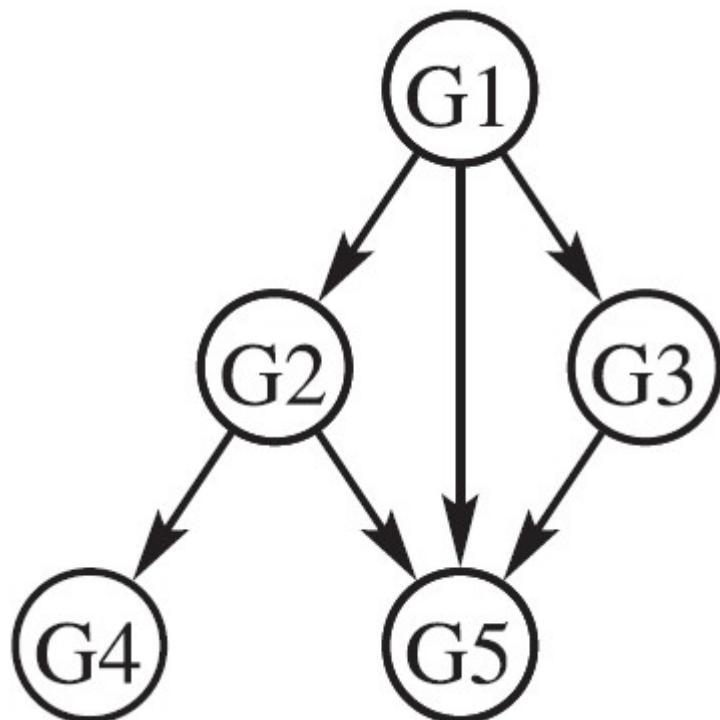


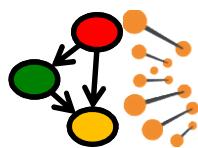
- Different conditional independence assumptions



Directed graphical models

- Probability distribution for a gene depends **only** on its regulators (parents) in the network.

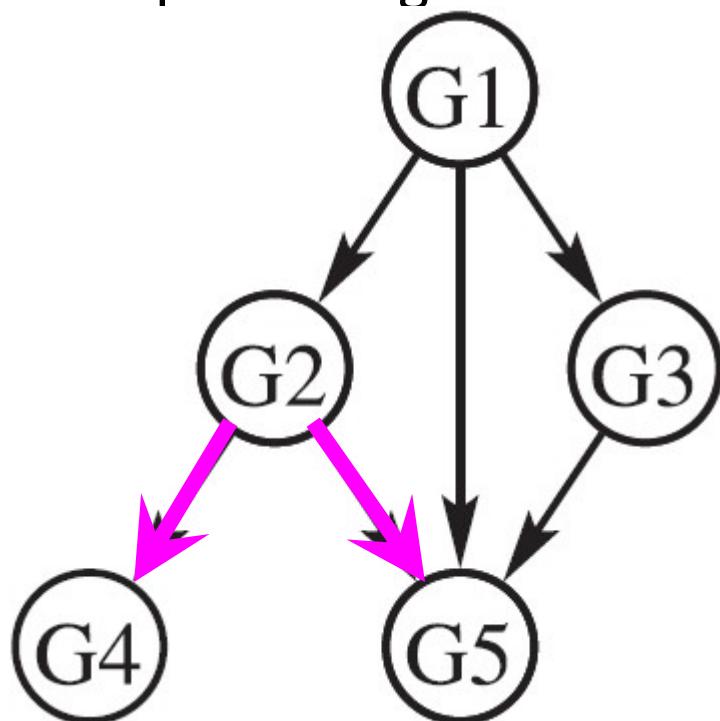


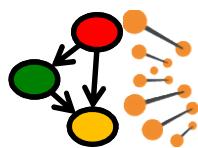


Review: Conditional Dependency Represented in Bayesian Networks

- The expression levels of G4 and G5 are related only because they share a common regulator G2.
- In mathematical term, G4 and G5 are conditionally independent given G2.

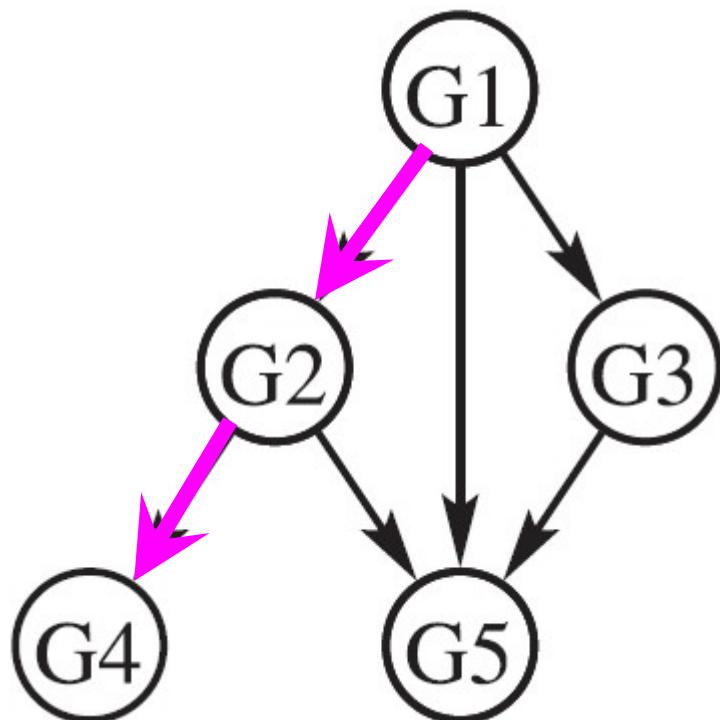
$$G4 \perp G5 \mid G2$$





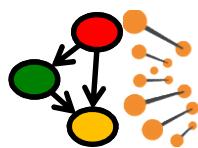
Review: Conditional Dependency Represented in Bayesian Networks

- The expression levels of G4 and G1 are related only because of gene G2.



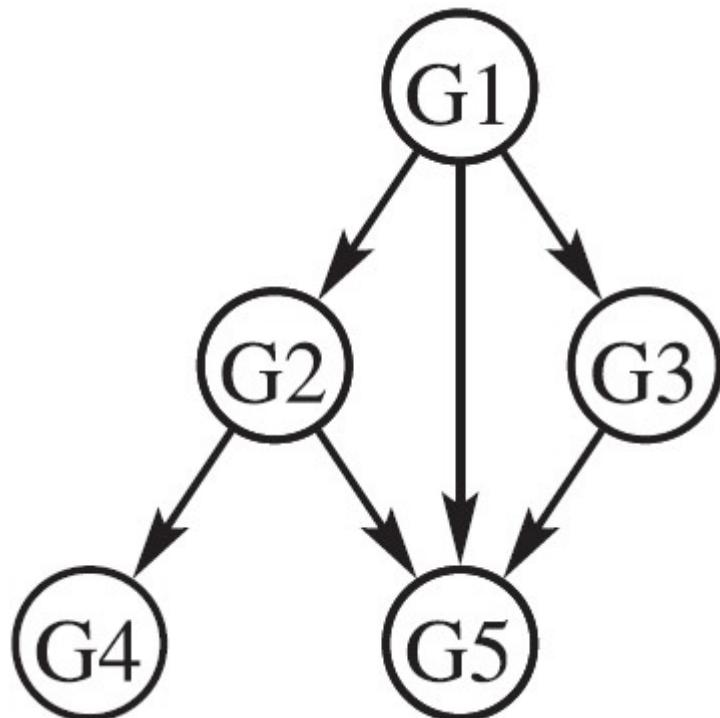
$$G4 \perp G5 \mid G2$$

$$\mathbf{G1 \perp G4 \mid G2}$$

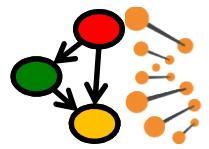


Review: Conditional Dependency Represented in Bayesian Networks

- $P(G1, G2, G3, G4, G5)$
= $P(G1) P(G2 | G1) P(G3 | G1) P(G4 | G2) P(G5 | G1, G2, G3)$



$$\begin{aligned} G4 \perp\!\!\!\perp G5 &| G2 \\ G1 \perp\!\!\!\perp G4 &| G2 \\ &\vdots \end{aligned}$$



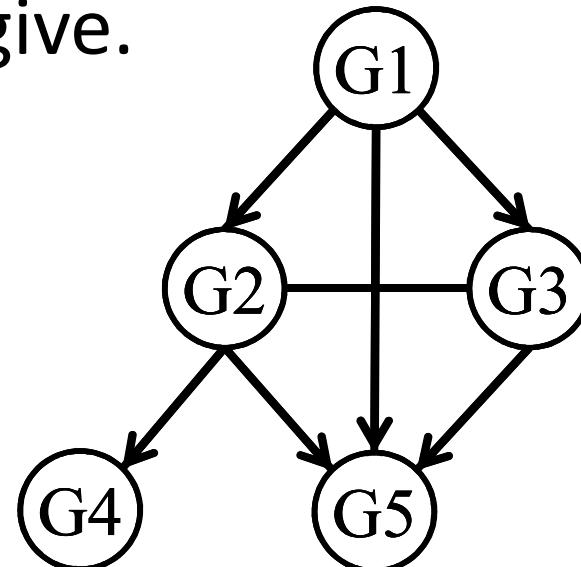
Convert to undirected model

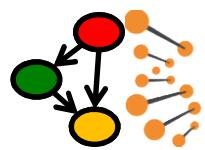
- Step 1: “Moralize” the graph (marry parents)
- Step 2: Remove arrow heads.
- This is the best approximation that an undirected model can give.

Pro: No concerns about cycles!

Pro: More stable inference.

Con: No “direction” of influence.

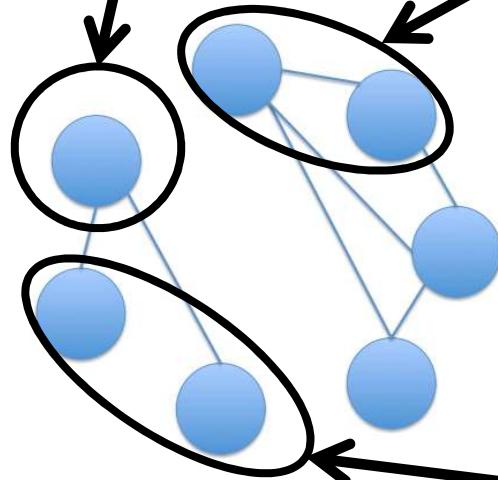




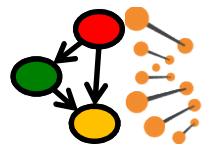
In undirected graphical models...

Each node represents a gene

Edge indicates 2 genes are conditionally dependent



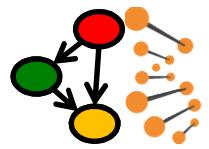
2 genes that are conditionally independent



Starbucks Example



Credit: Daniela Witten

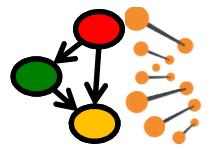


Coffeeshop data

Like Coffeeshops?
Like Coffee?
Like Tea?

Person 1
Person 2
.
.
.
.
.
.
.
.
Person 100

	Like Coffeeshops?	Like Coffee?	Like Tea?
Person 1	Y	Y	N
Person 2	N	N	N
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
Person 100	Y	N	Y

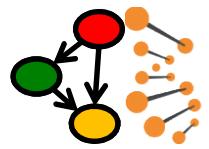


Correlation matrix

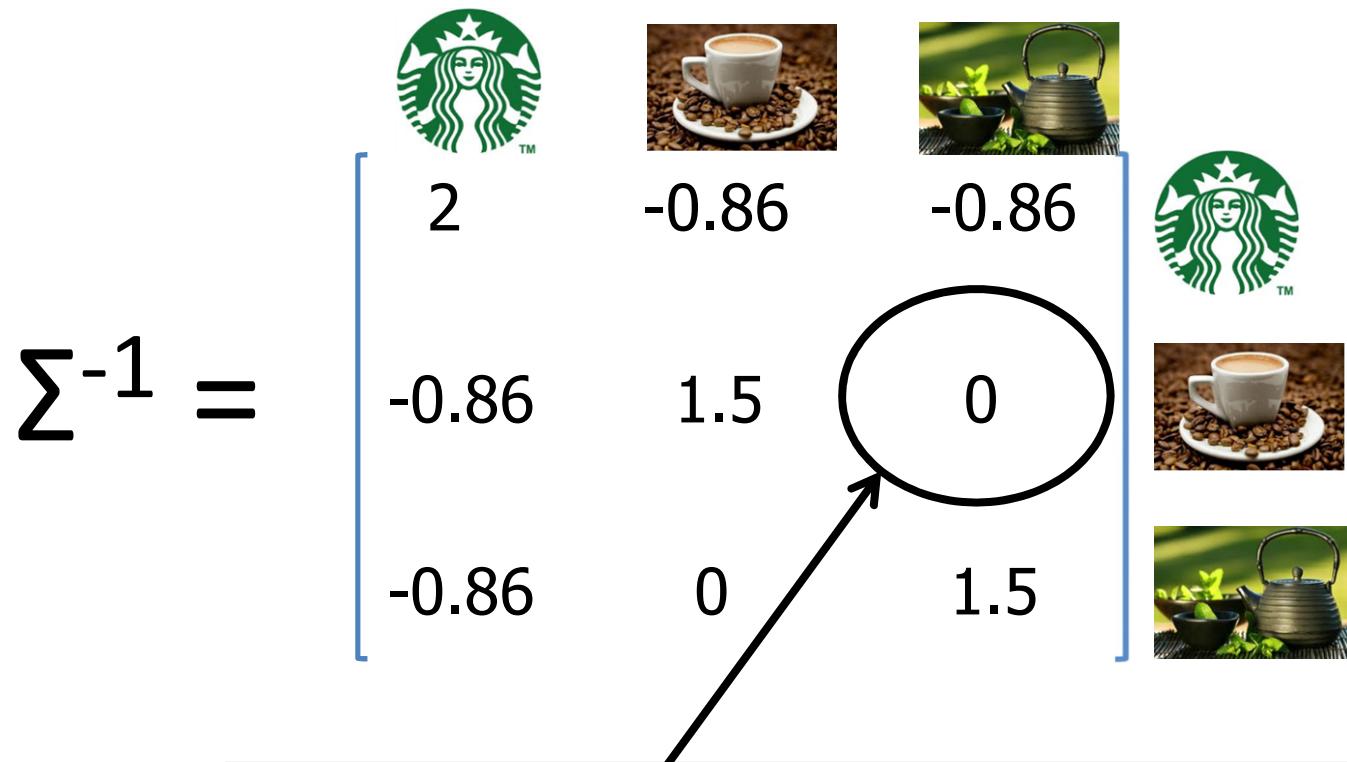
$$\Sigma = \begin{bmatrix} 1 & 0.58 & 0.58 \\ 0.58 & 1 & 0.33 \\ 0.58 & 0.33 & 1 \end{bmatrix}$$

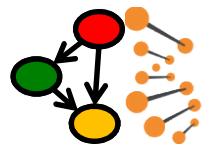
The correlation matrix is represented as a 3x3 grid. The diagonal elements are 1, indicating perfect correlation with themselves. The off-diagonal elements are 0.58 and 0.33, representing the correlation between the first and second variables, and the second and third variables respectively. To the left of the matrix, there is a blue bracket grouping the three rows together.

The images in the grid are arranged in a 3x3 pattern. The top row contains the Starbucks logo, a coffee cup on a saucer with coffee beans, and a traditional teapot and cups. The middle row contains a coffee cup on a saucer with coffee beans, and the bottom row contains a traditional teapot and cups. The images are repeated along the right edge of the matrix.



Inverse correlation matrix

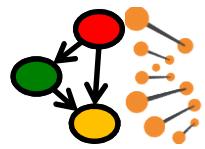




The network



Credit: Daniela Witten

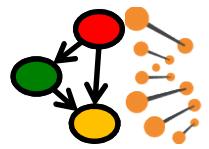


Correlation versus inverse correlation

- Every population correlation matrix Σ has an associated inverse correlation matrix Σ^{-1} .
- We might be tempted to estimate Σ in order to learn the network structure...

But this includes all indirect interactions!

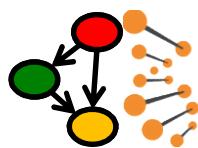
- Instead, we can estimate Σ^{-1} .
- What we really need: a good way to estimate Σ^{-1} !



Outline

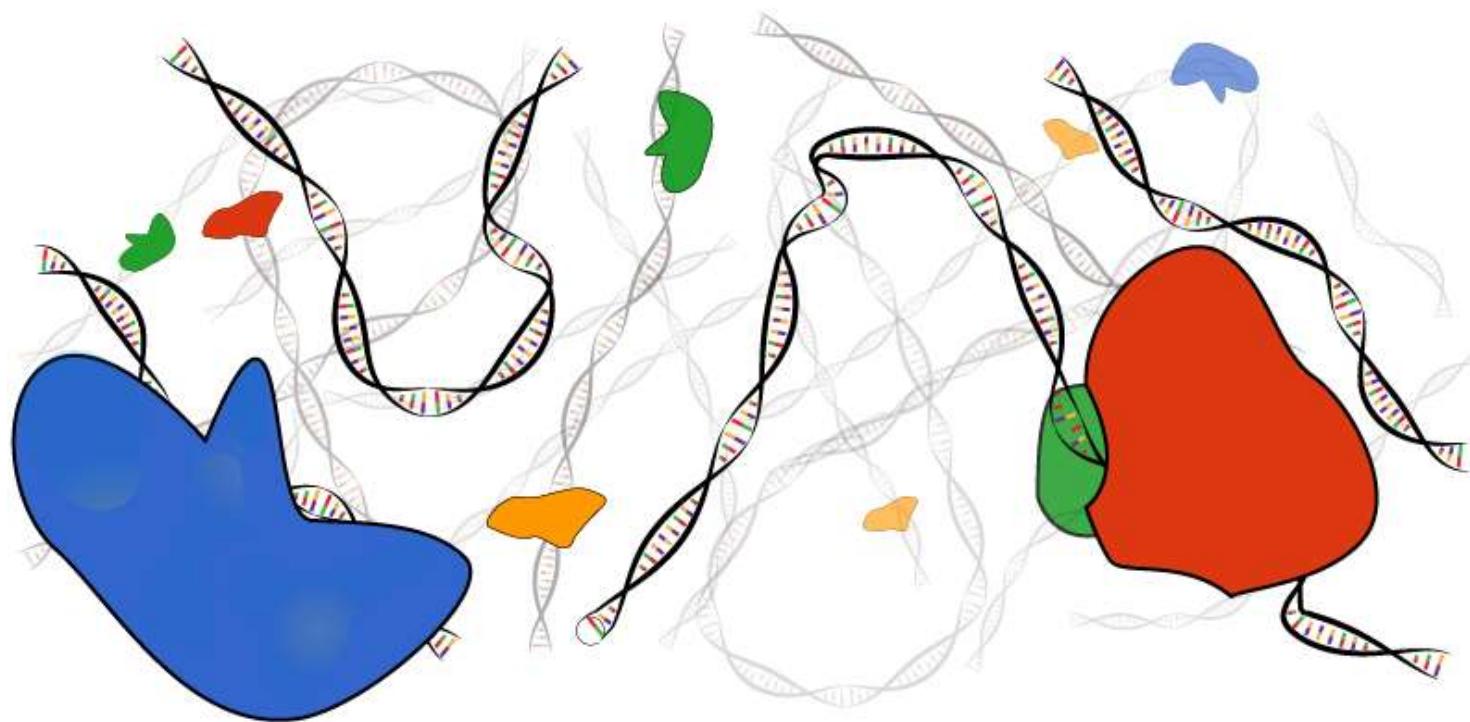
- Motivation
 - Learning a network of a large number of ChIP-seq datasets
- Key features of ChromNet
- Visualization tool
- Future directions

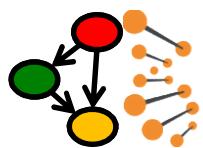




Regulatory factors co-localize in the genome to interact with each other.

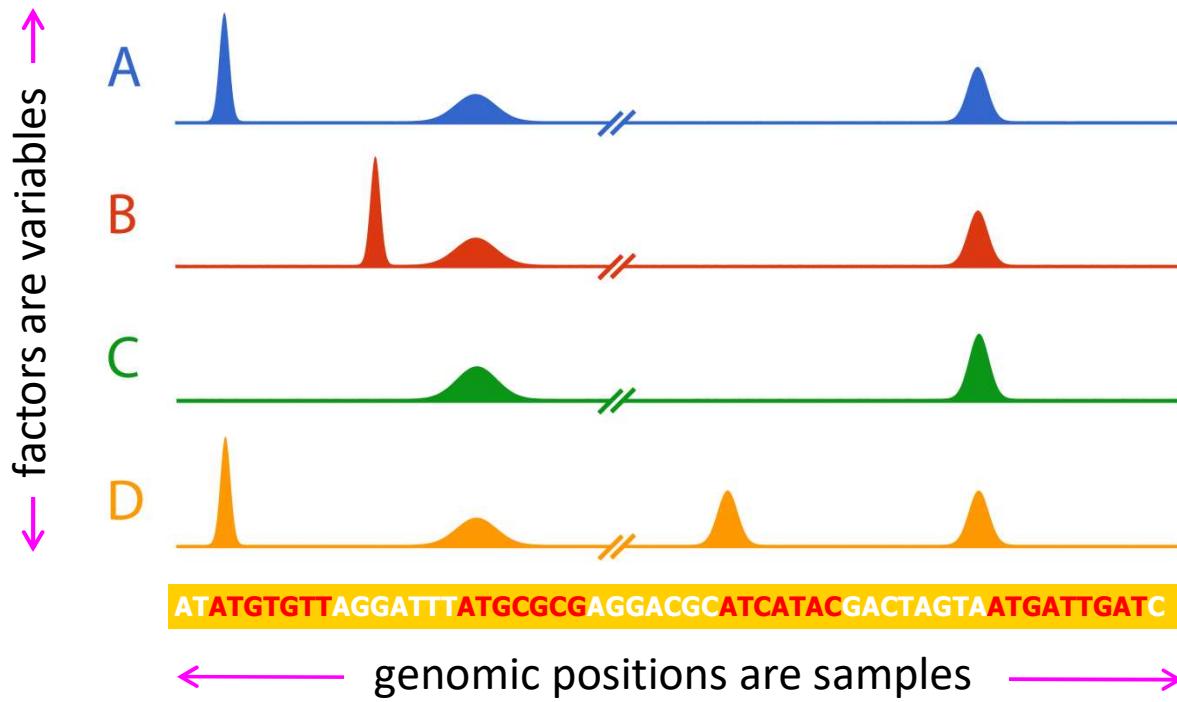
- Our goal is to learn this network of interactions, we call the *chromatin network**.



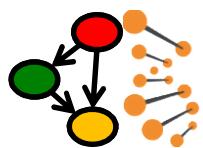


ChIP-seq data measure genome-wide localization of regulatory factors.

Consider the following simulated ChIP-seq data:

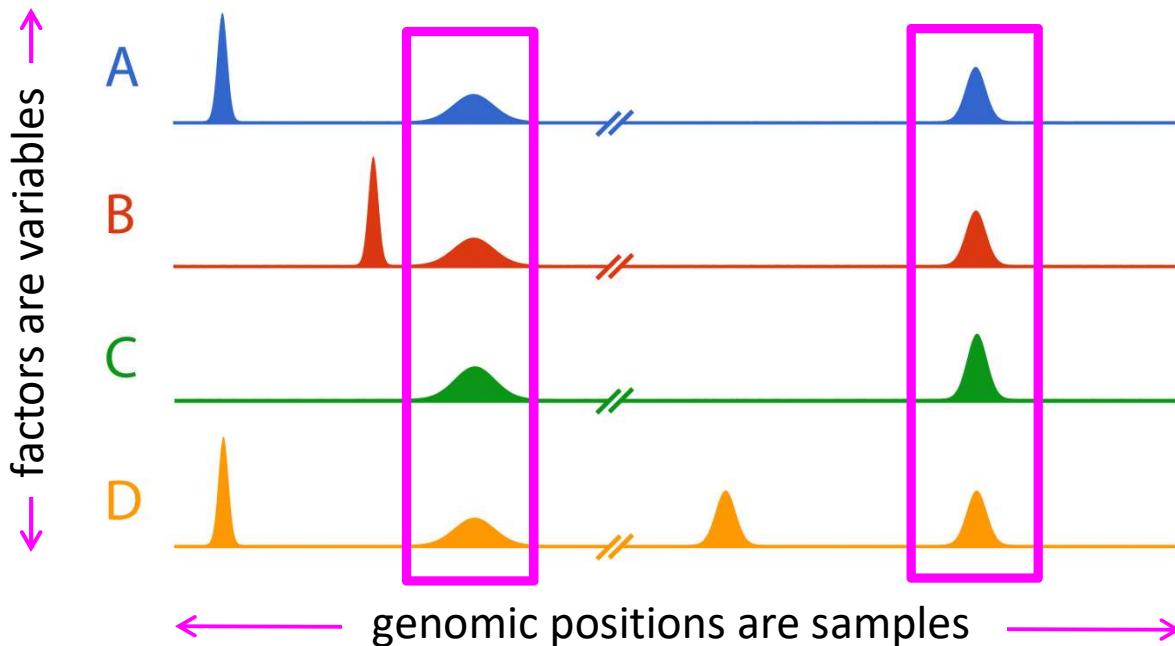


How would you infer which regulatory factors worked together?

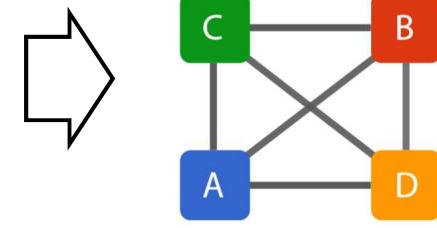


One idea is to see if the regulatory factors tend to show up together.

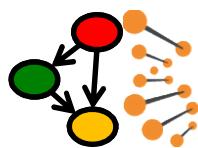
Consider the following simulated ChIP-seq data:



Co-occurrences

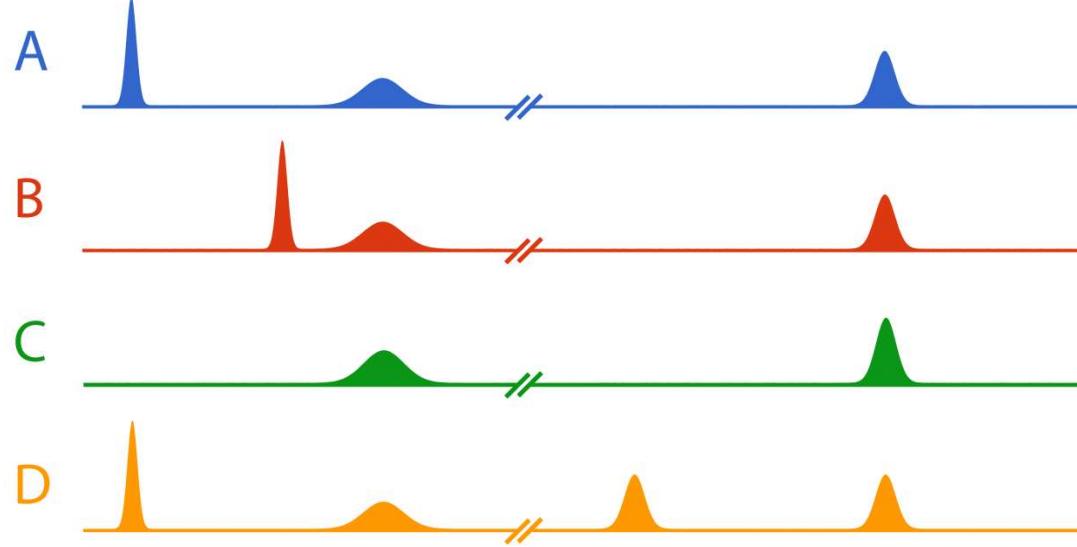


The problem is they *all* co-occur in a statistically significant manner!

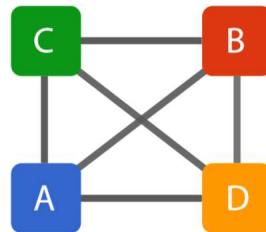


One idea is to see if the regulatory factors tend to show up together.

Consider the following simulated ChIP-seq data:

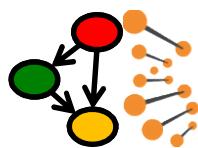


Co-occurrences



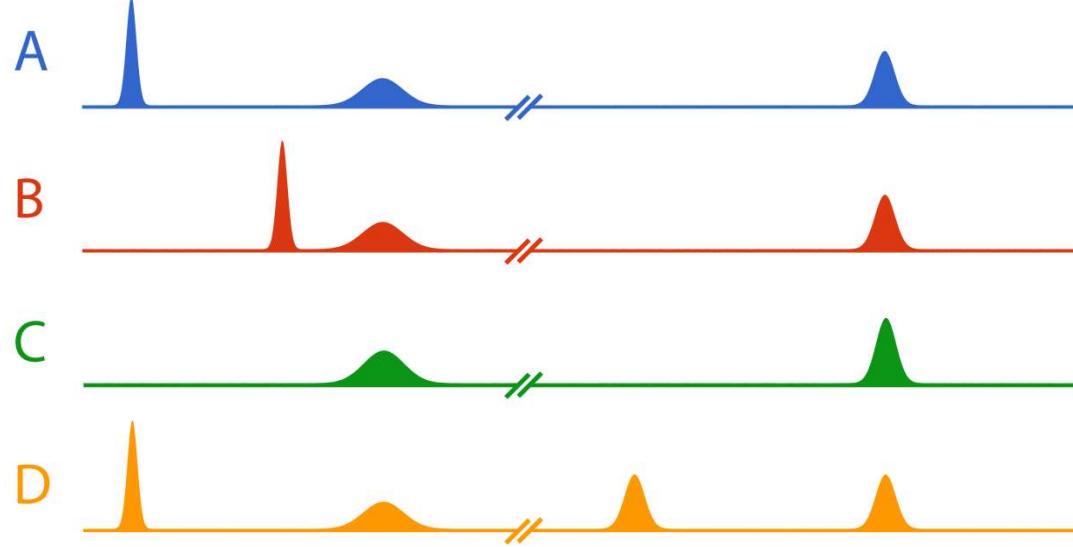
The problem is they *all* co-occur in a statistically significant manner!

- **Direct interaction:** *physical contact or functional coupling* that requires spatial proximity
- **Indirect interaction:** *Transitive effect* of direct interactions

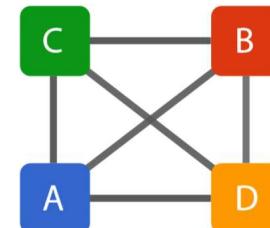


Here is a possible scenario of interactions among these 4 factors

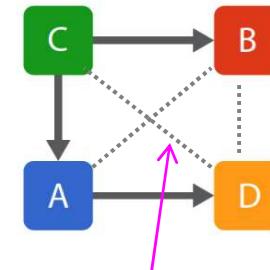
C recruits A and B, and A in turn recruits D



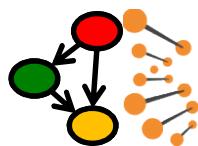
Co-occurrences



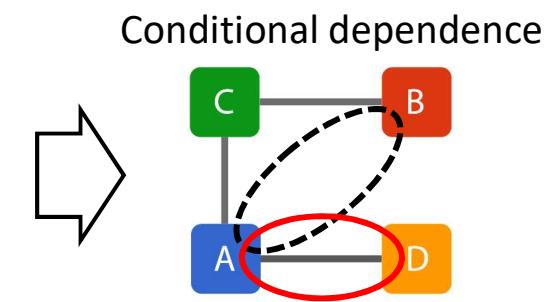
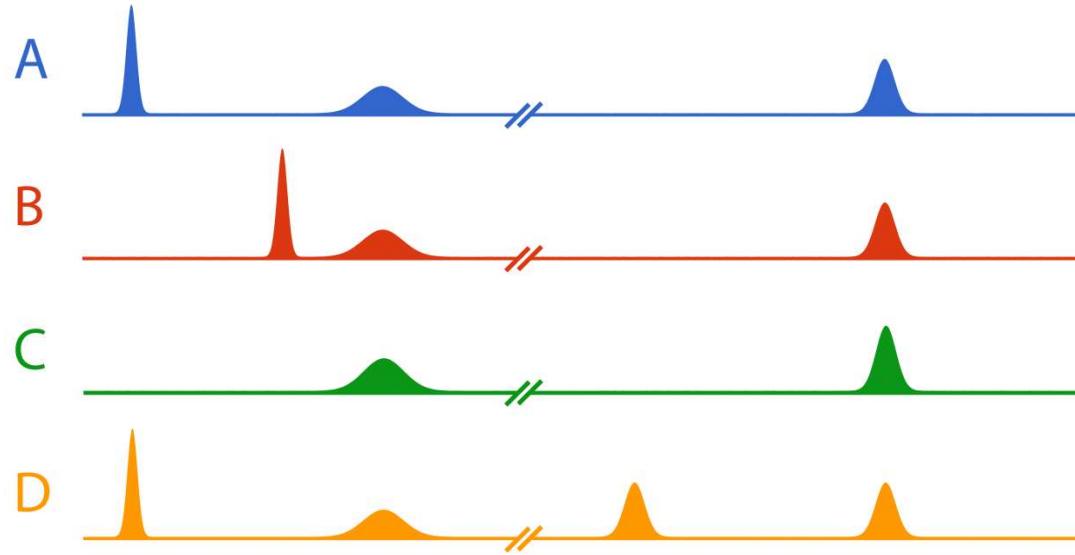
True interactions



indirect
interactions



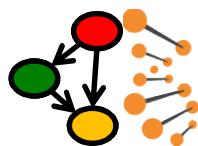
A better idea is to use conditional dependence to estimate direct interaction.



Edge indicates 2 factors are *conditionally dependent*.

2 factors that are not connected are conditionally independent.

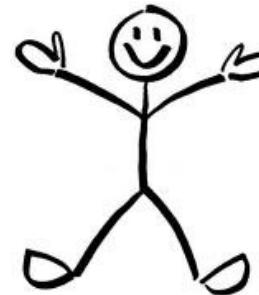
Conditional dependence represents a “direct” interaction.



Hypothetical example: What is conditional dependence relationship?

GS collaboration network

Su-In and Judit has a shared interest in proteomics

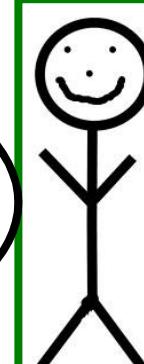


Su-In
Lee

Su-In and Tony are very interested in personalized oncology

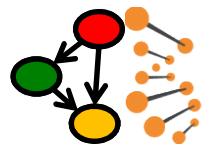


Judit
Villen

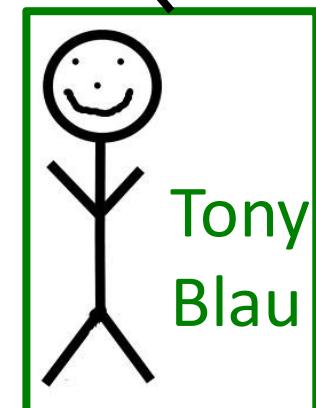


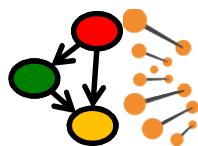
Tony
Blau

SHARED PREFERENCE CAN BE FULLY EXPLAINED BY Su-In

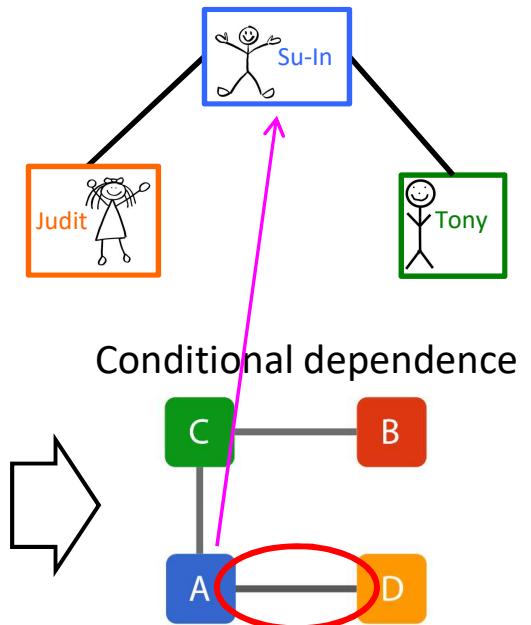
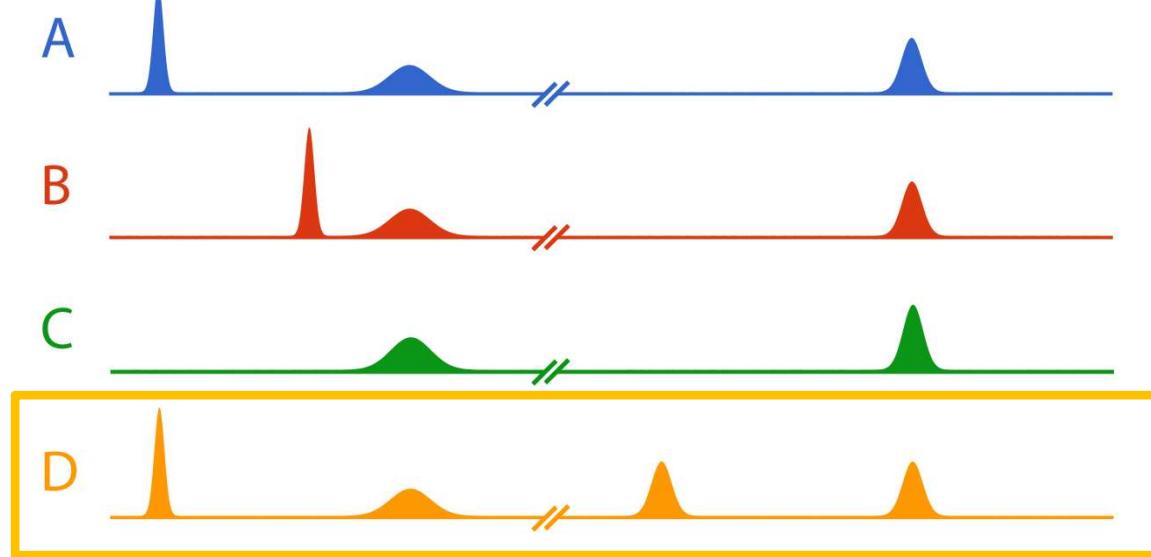


Conditional dependencies capture direct relationships





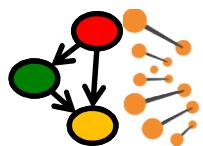
How can we capture conditional dependence relationships from data?



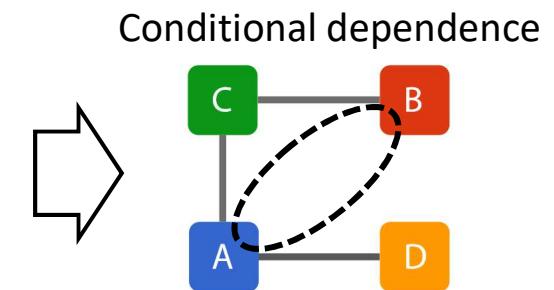
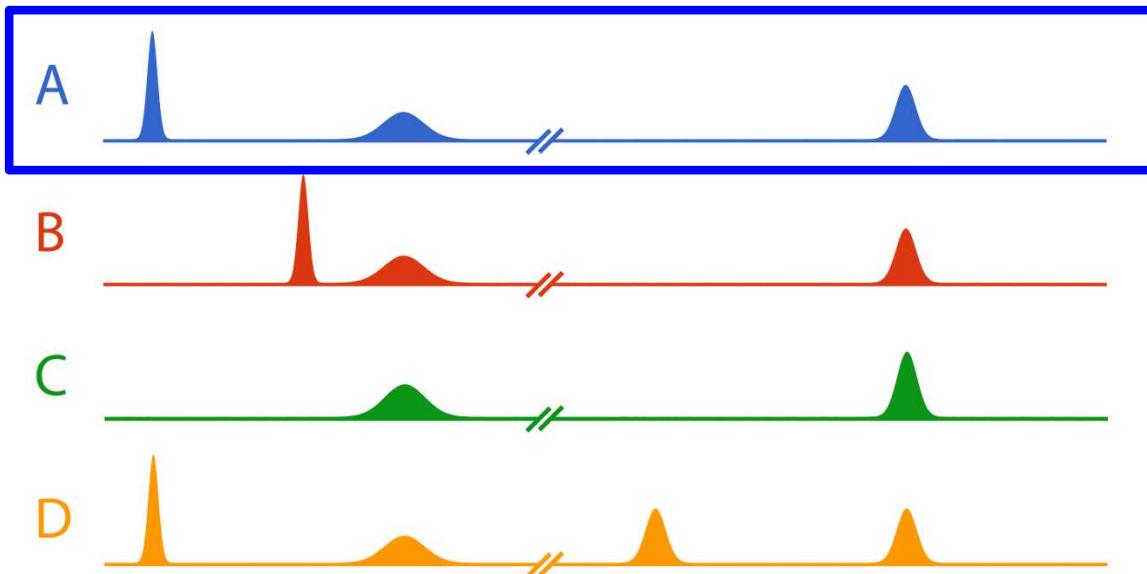
If we run a linear regression of D against all the other variables
which coefficients will be non-zero?

$$D = w_A A + w_B B + w_C C$$

B and C give no more information about D after accounting for A.



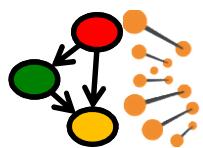
A better idea is to use conditional dependence to estimate direct interaction.



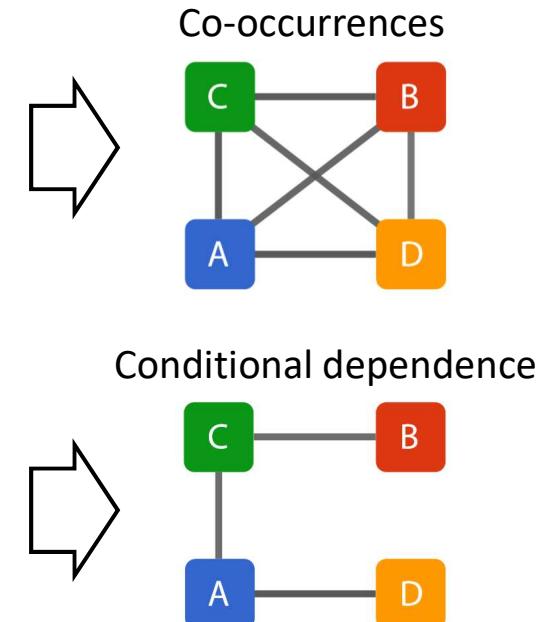
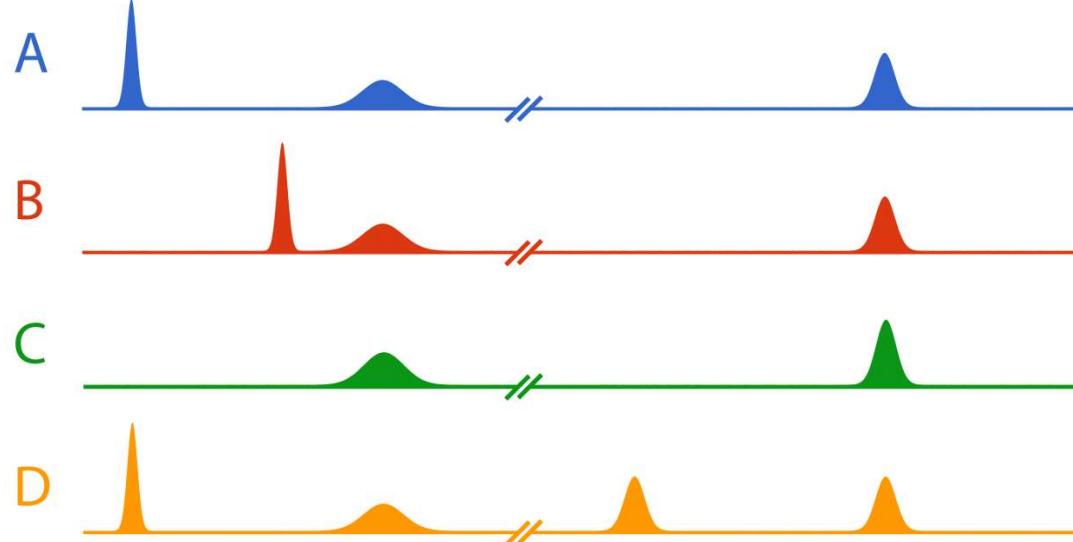
If we run a linear regression of A against all the other variables
which coefficients will be non-zero?

$$A = v_B B + v_C C + v_D D$$

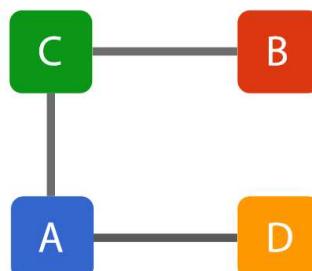
B gives no more information about A after accounting for C and D.



Conditional dependencies can better capture direct interactions

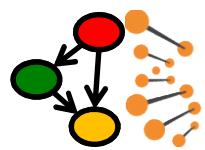


True interactions

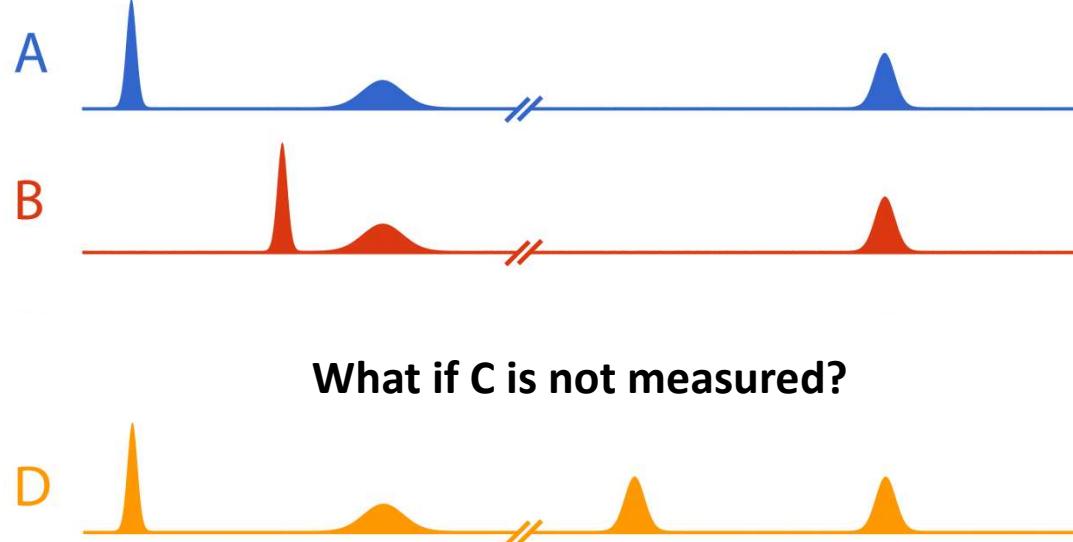


Direct interaction: physical contact or functional coupling that requires spatial proximity

Indirect interaction: Transitive effect of direct interactions

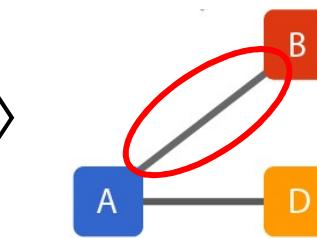
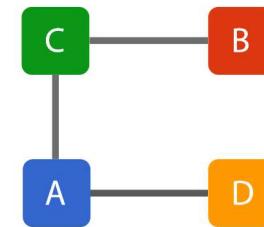


It is important to observe many factors.



What if C is not measured?

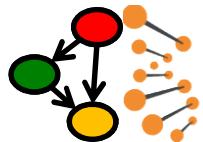
True interactions



Conditional dependence

Then the connection between A and B can no longer be explained by other variables.

Incorporating more ChIP-seq data sets would improve the network.



ChromNet – Learning a conditional dependence network from 1,451 ENCODE ChIP-seq datasets

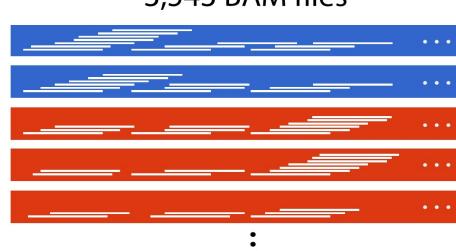
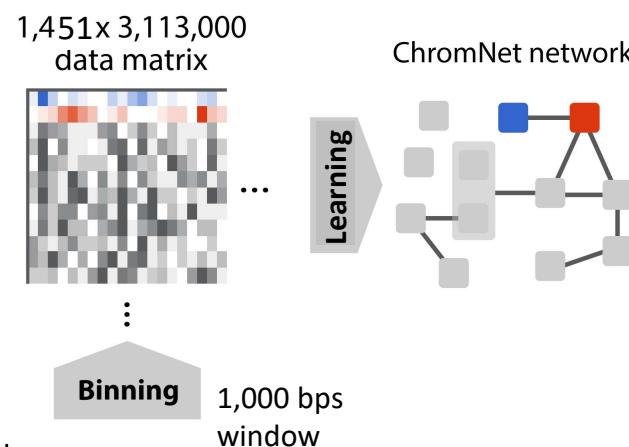
1,451 ChIP-seq datasets – 223 transcription factors and 14 histone marks from 105 cell types.

3,543 FASTQ files

AGGC...	AAAG...	GTAA...	...
CTAA...	AAAT...	GAAA...	...
TTGA...	AGTT...	CCGT...	...
ACGT...	CATC...	AGTC...	...
GAAG...	GAAC...	AGCC...	...

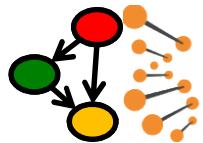
⋮

Bowtie2



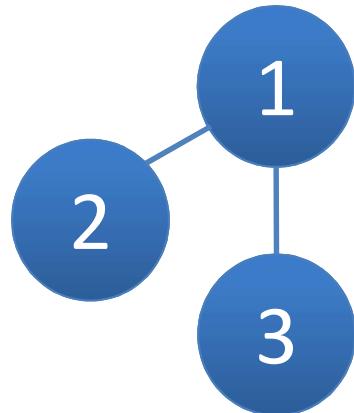
⋮

- Key motivations
 - The **more variables** are included, the more chance we have to **remove indirect interactions**.
 - The human genome has billions of positions, giving **statistical power** to infer conditional dependencies.
 - Integrating **ChIP-seq datasets from many cell types** into a **single network** provides unique advantages.



Key Challenge I – Large data

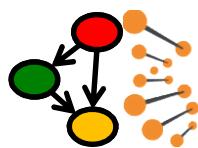
- Learning a network among thousands of ChIP-seq datasets based on millions of samples ($1,451 \times 3,000,000$) is **highly computationally intensive**.
- **Our approach:** Compute the inverse covariance matrix Σ^{-1} .
 - A zero element $(\Sigma^{-1})_{ij} = 0$ means that the i -th and j -th variables are conditionally independent to each other (Gaussian*, binary **)



$$\Sigma^{-1} = \begin{bmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0 \\ 0.2 & 0 & 1 \end{bmatrix}$$

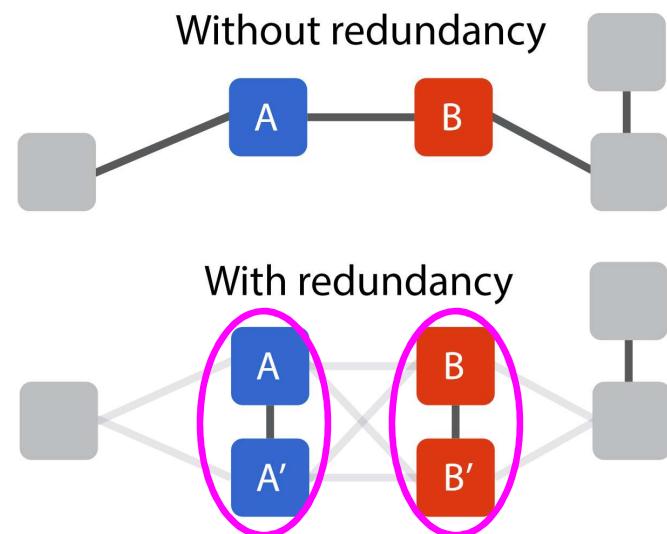
* Lauritzen, *Graphical Models*, 1996

** Loh, Wainwright et al. *The Annals of Statistics*, 2013



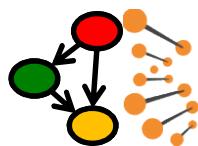
Key Challenge II – High redundancy in data

- Conditional dependence has been applied before to smaller datasets (e.g. 23 histone marks*).
- Large networks contain much more redundancy.
 - Same factors measured in different labs, conditions or cell types.
 - Some factors are functionally very much related.
- When **variables are highly correlated** each other, they explain each other and disconnect from other nodes.
- **Solution:** We developed the **group graphical model****.



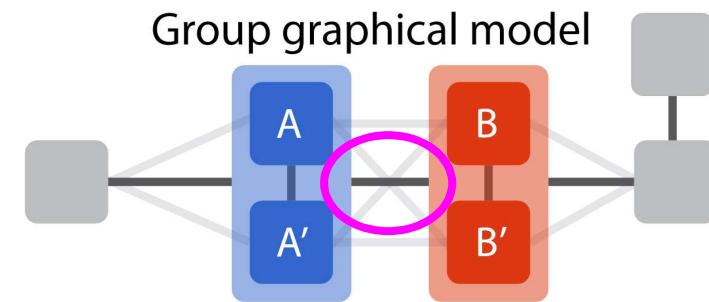
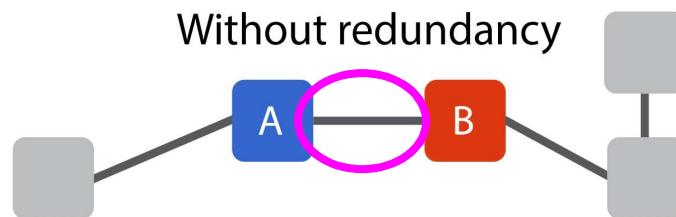
** Lundberg et al. Accepted to *Genome Biology*

* Lasserre et al., *PLoS Computational Biology*, 2013



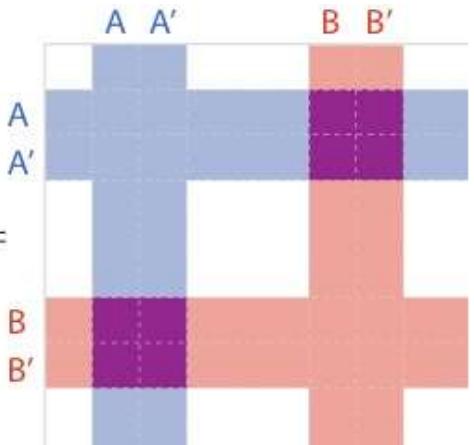
To address this problem, we developed the group graphical model*

- In a group graphical model, we want to **preserve the magnitude of the edge between (A, A') and (B, B')** in the model without redundancy.

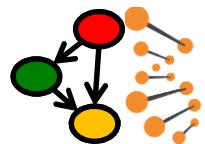


- We prove that this can be estimated by summing the entries in the Σ^{-1} *:

$$G_{[A,A'][B,B']} = \Sigma_{AB}^{-1} + \Sigma_{AB'}^{-1} + \Sigma_{A'B}^{-1} + \Sigma_{A'B'}^{-1} \quad \Sigma^{-1} =$$



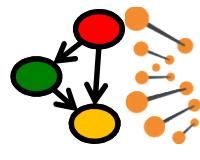
* Lundberg et al. Accepted to *Genome Biology*



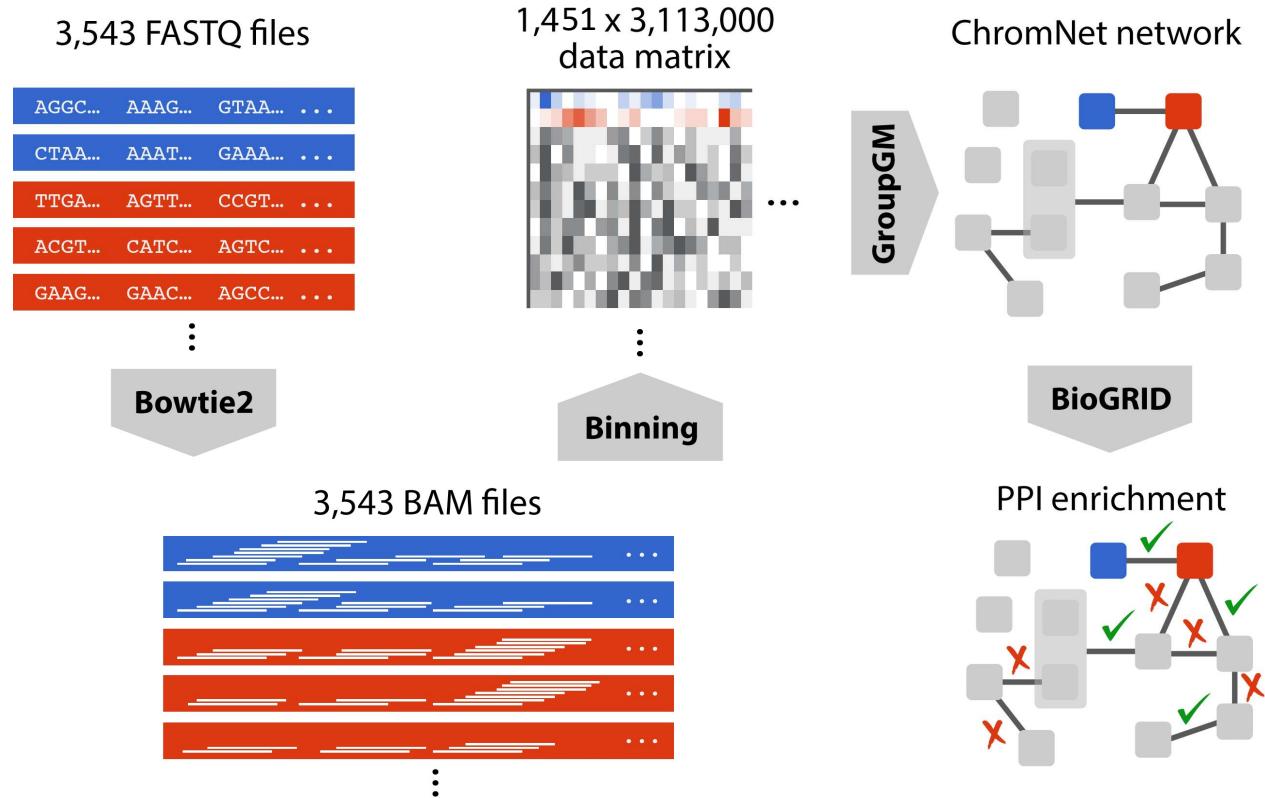
Outline

- Motivation
- Key features of ChromNet
 - Identifying conditional dependence
 - Group-based modeling
- Learning a joint model of all cell types
- Identifying context-specificity
- Visualization tool
- Discussion

Group graphical
model (GroupGM)



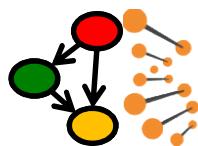
Integrating 1,451 ChIP-seq datasets into a group graphical model



Evaluation

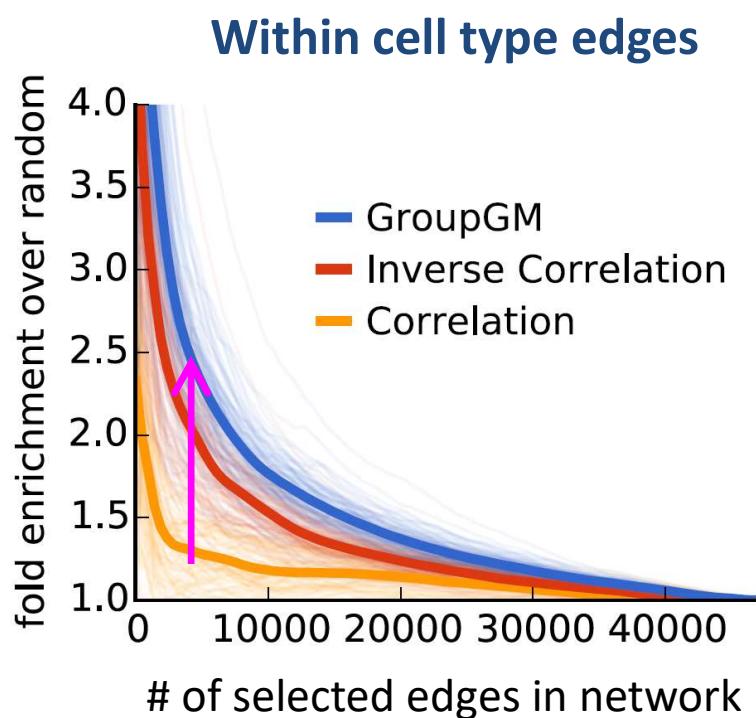
Comparison with

- Correlation network (co-occurrences)
- Inverse correlation Σ^{-1} (conditional dependence without group modeling)

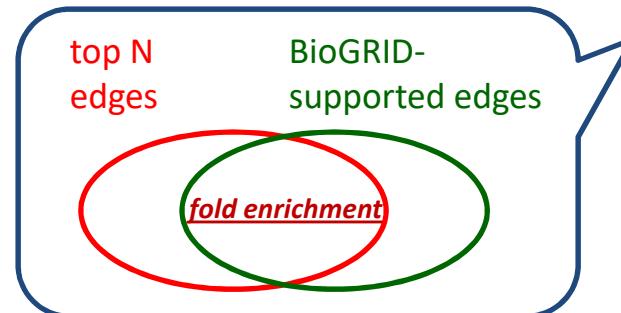


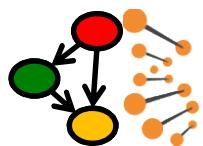
Evaluating how well the network estimate reveals known interactions

- The human chromatin network is not fully understood.
- As a “silver standard”, we use previously **known physical interactions between proteins (PPIs)** annotated in BioGRID.



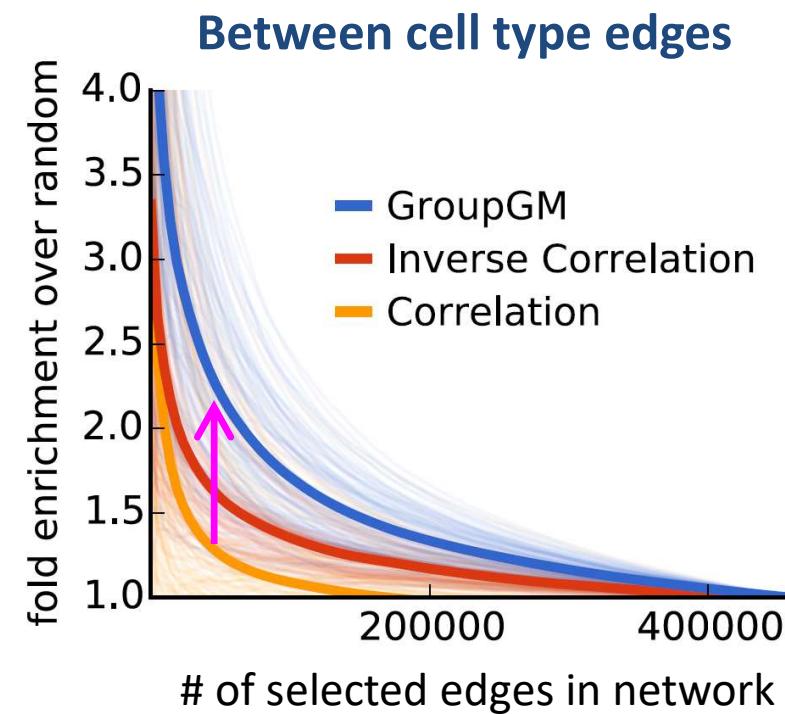
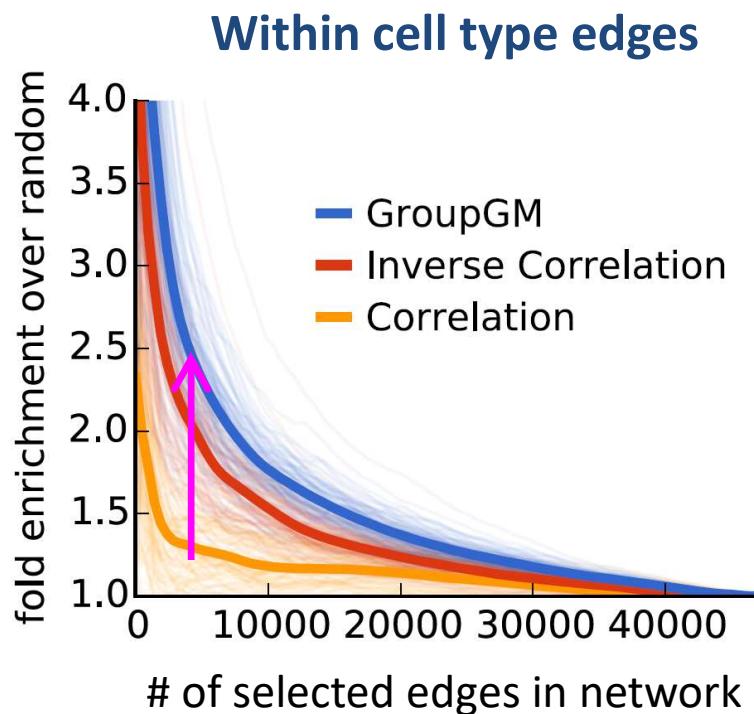
- Consider **N strongly weighted edges** (x-axis).
- Compute the **fold enrichment of BioGRID-supported edges** over random N edges (y-axis).

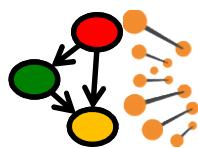




Evaluating how well the network estimate reveal known interactions

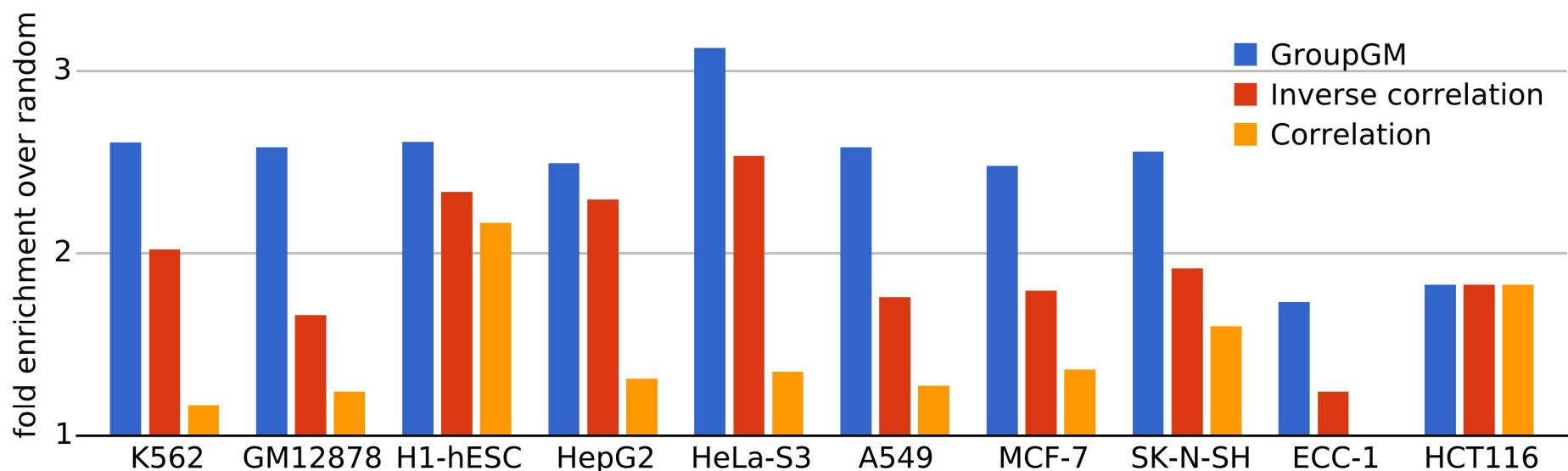
- The human chromatin network is not fully understood.
- As a “silver standard”, we use previously **known physical interactions (PPIs)** annotated in BioGRID.

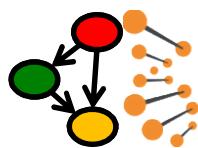




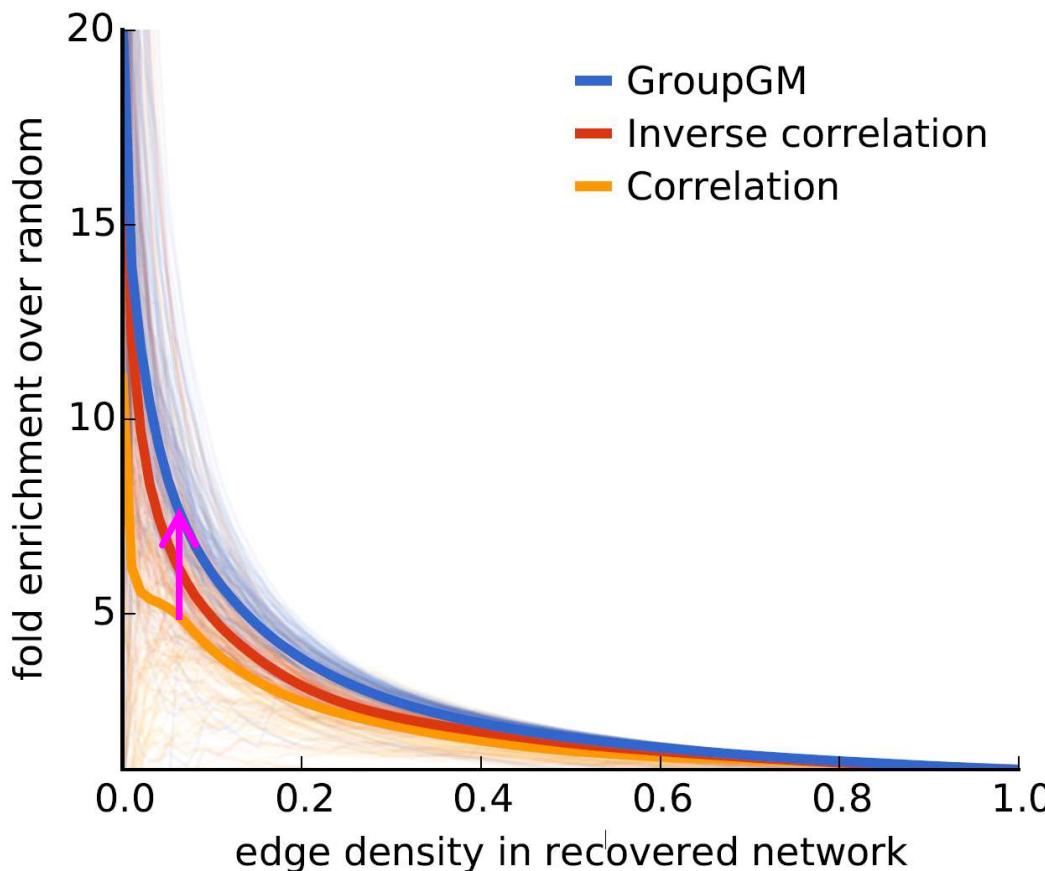
GroupGM is consistently better than alternative methods across cell types

- We measure the fold enrichment across different cell types, for a fixed N value (= total number of BioGRID-supported edges).



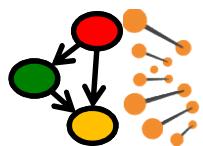


Chromnet reveals known TF-histone marks interactions



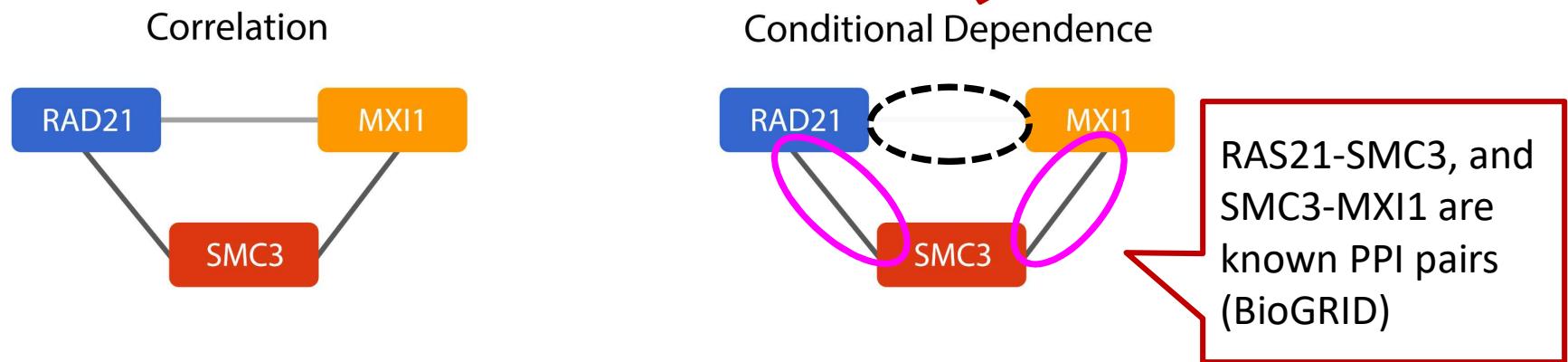
- We used known histone mark/writer combinations* to validate TF/histone mark edges
- GroupGM is better than the alternative methods

* Histome database: Khare et al. *Nucleic Acids Research* (2012)



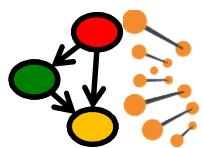
Importance of conditional dependence: SMC3 separates RAD21 and MXI1

A previous study identified 200 potential interactors of RAD21. MXI1 was not one of them*.

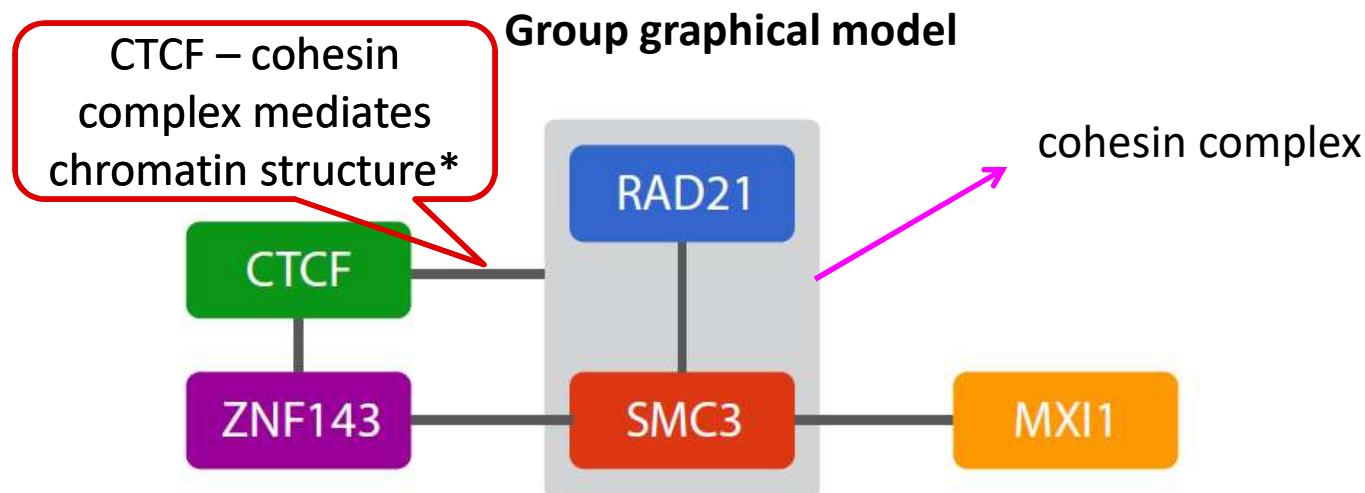


HeLa-S3 cervical carcinoma cells

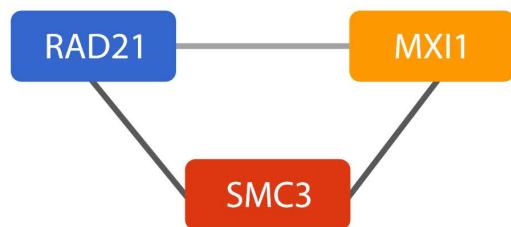
* Panigrahi et al., “A cohesin-RAD21 interactome”. *Biochemical Journal* 442.3 (2012)



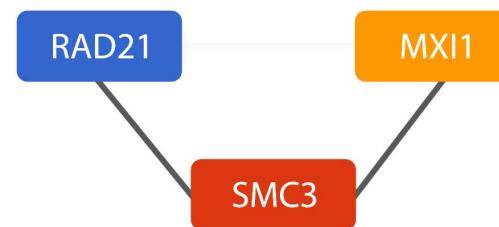
Here is what the group graphical model captured



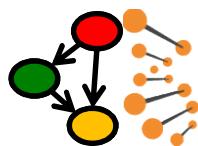
Correlation



Conditional Dependence

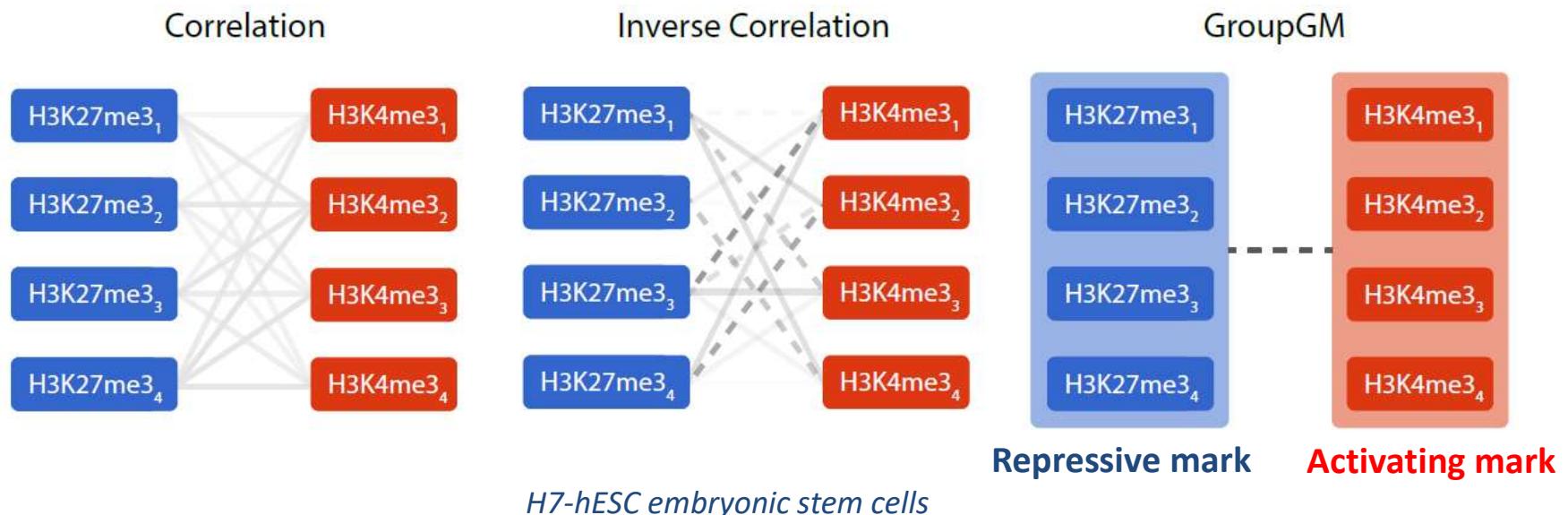


* Parelho et al. "Cohesins functionally associate with CTCF on mammalian chromosome arms". *Cell* (2008)

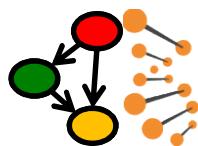


Recovering a connection between H3K27me3 and H3K4me3

- Typically the problem of redundancy is avoided by either averaging or removing redundant variables.
 - However, such choices require knowledge of both the data and the scale of the interactions being considered.



H3K28me3 and H3K4me3 were measured at UW at different time points of differentiation.

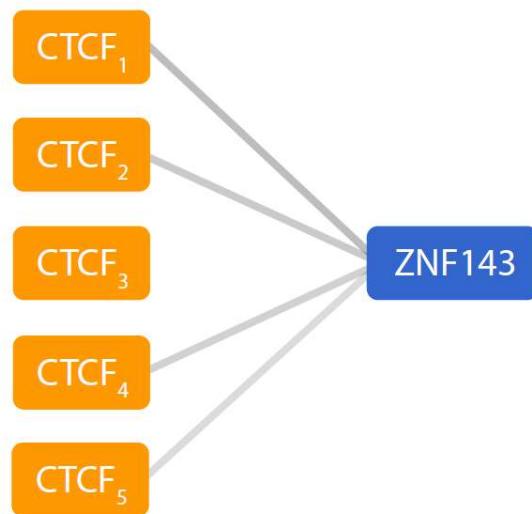


Another example to show how GroupGM mitigates problems with highly correlated variables

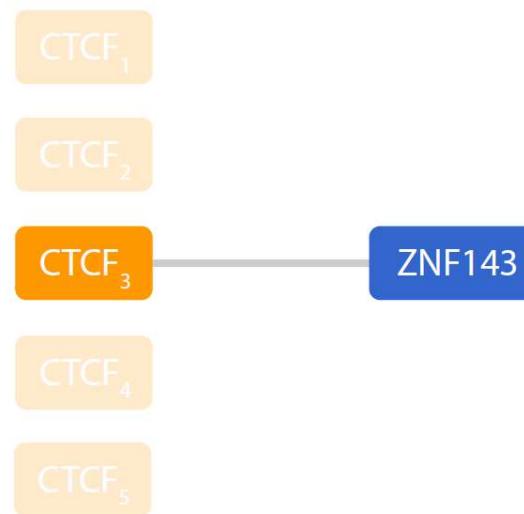
- ZNF143 – CTCF connections in K562

standard conditional
dependence model

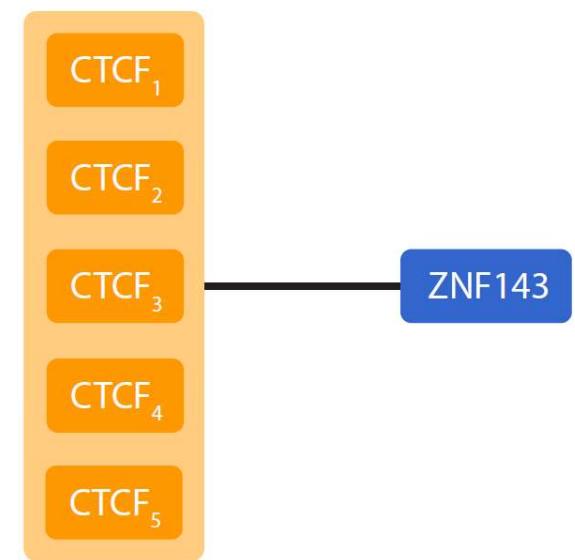
Inverse correlation
(all datasets)

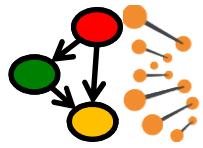


Inverse correlation
(4 datasets removed)



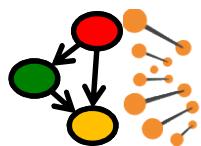
Group graphical model
(all datasets)





Outline of Talk

- Motivation
 - Increased number of possible interactions
 - Direct comparison across cell types when conditioned on the same global set of datasets
 - Extraction of global pattern of the network
- Learning a joint model of all cell types
- Identifying context-specificity
- Visualization tool
- Discussion

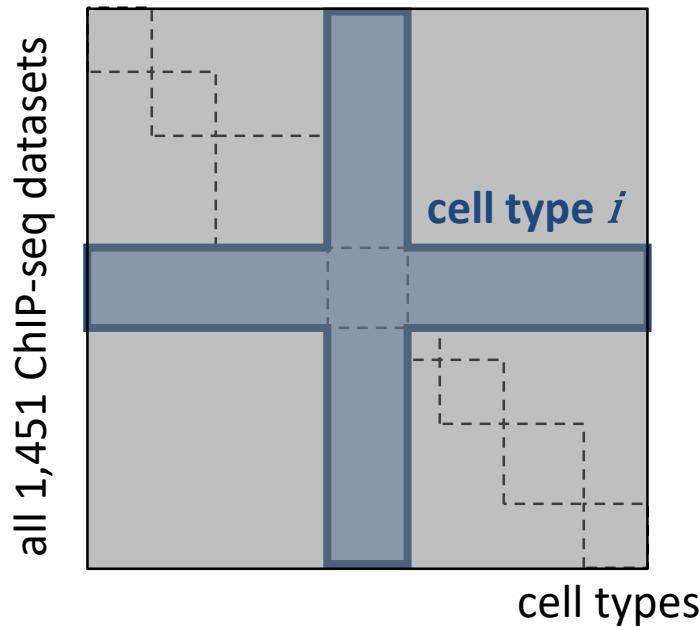


The value of learning a joint network of ChIP-seq datasets from multiple cell types

network estimates (Σ^{-1})

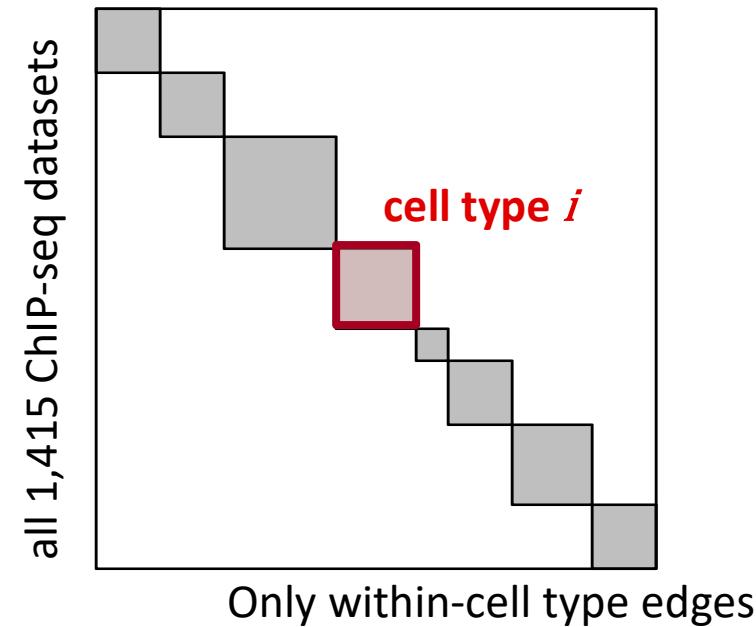
joint model

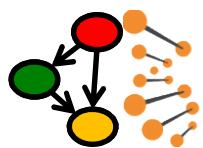
all 1,451 ChIP-seq datasets



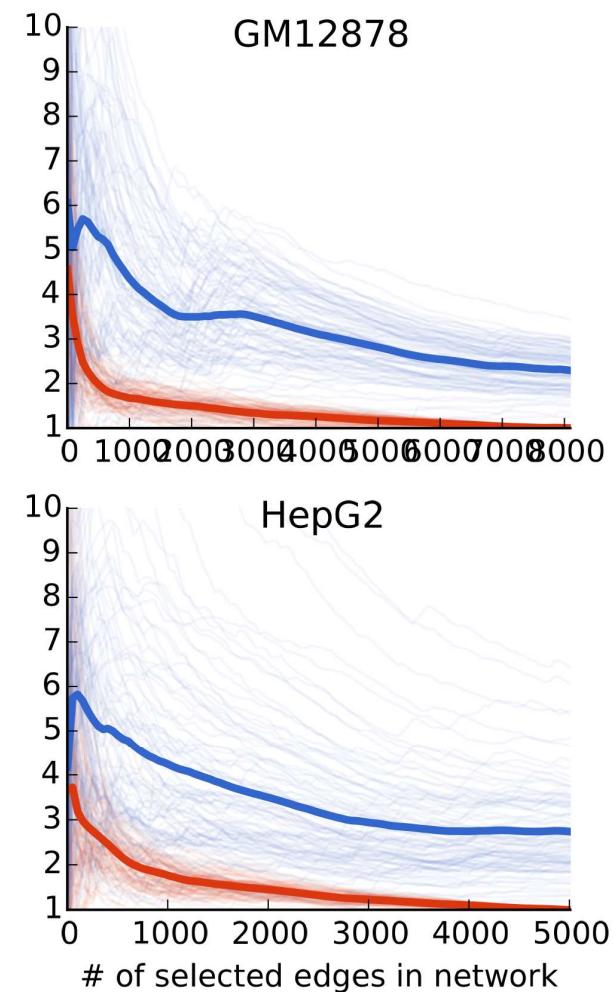
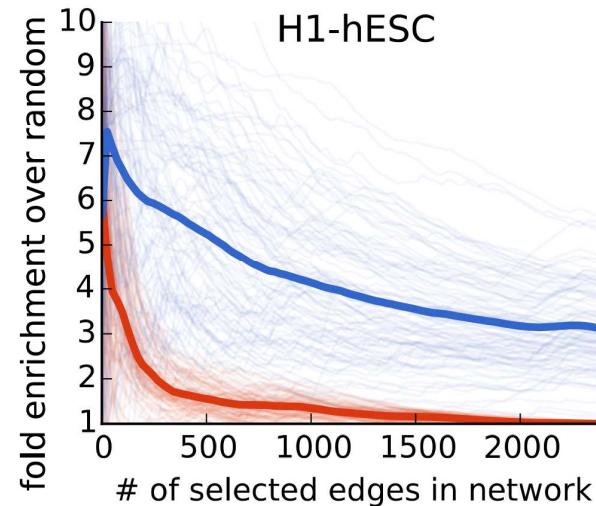
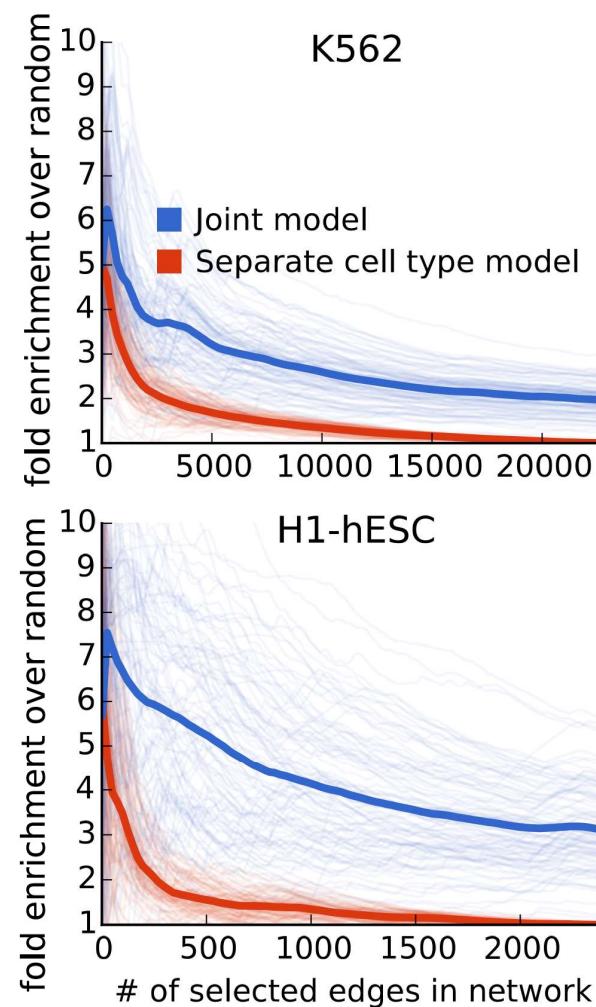
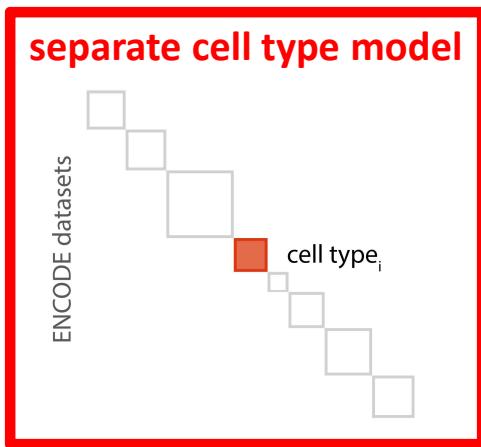
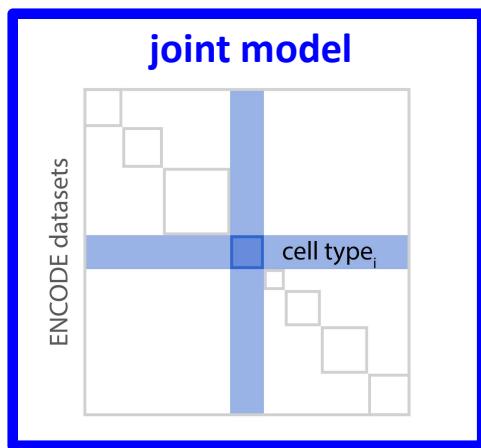
separate cell type model

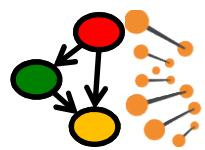
all 1,415 ChIP-seq datasets



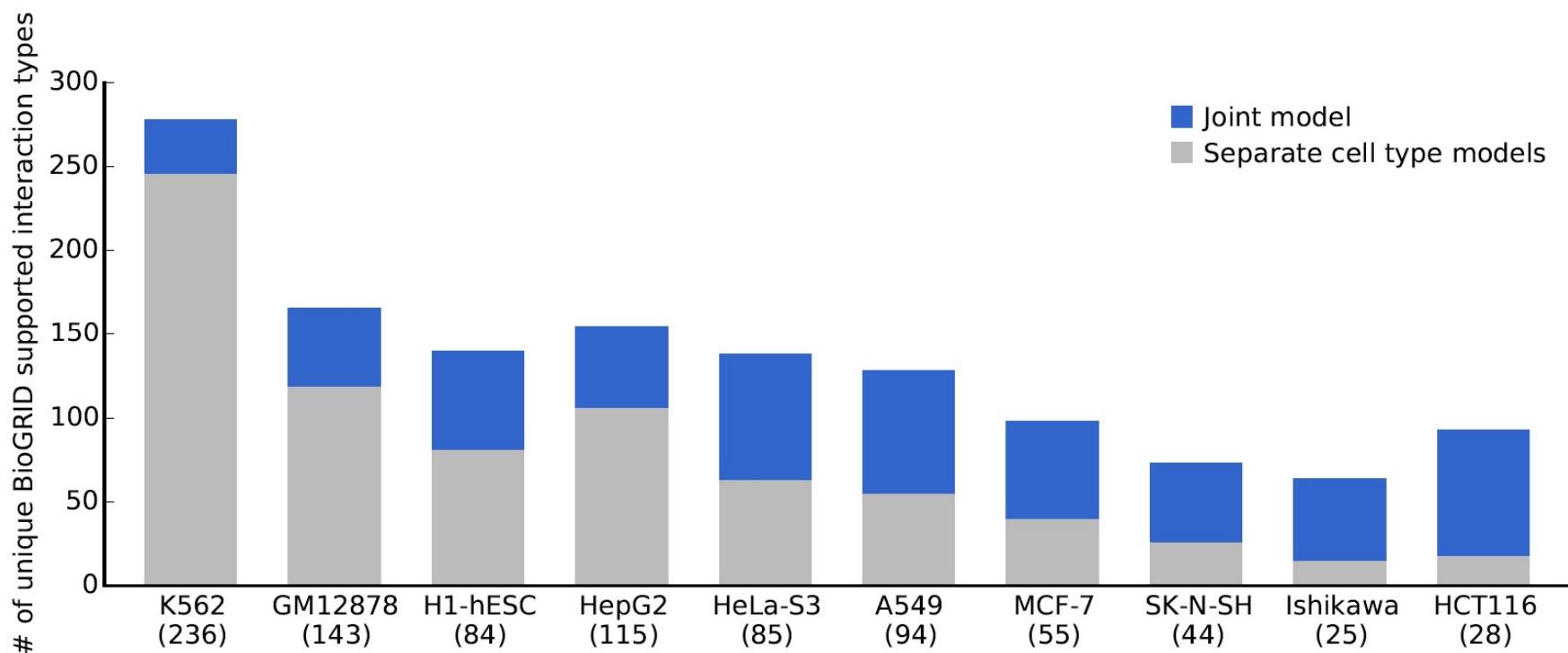


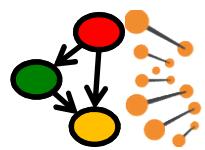
Comparison between joint model and separate cell type-model



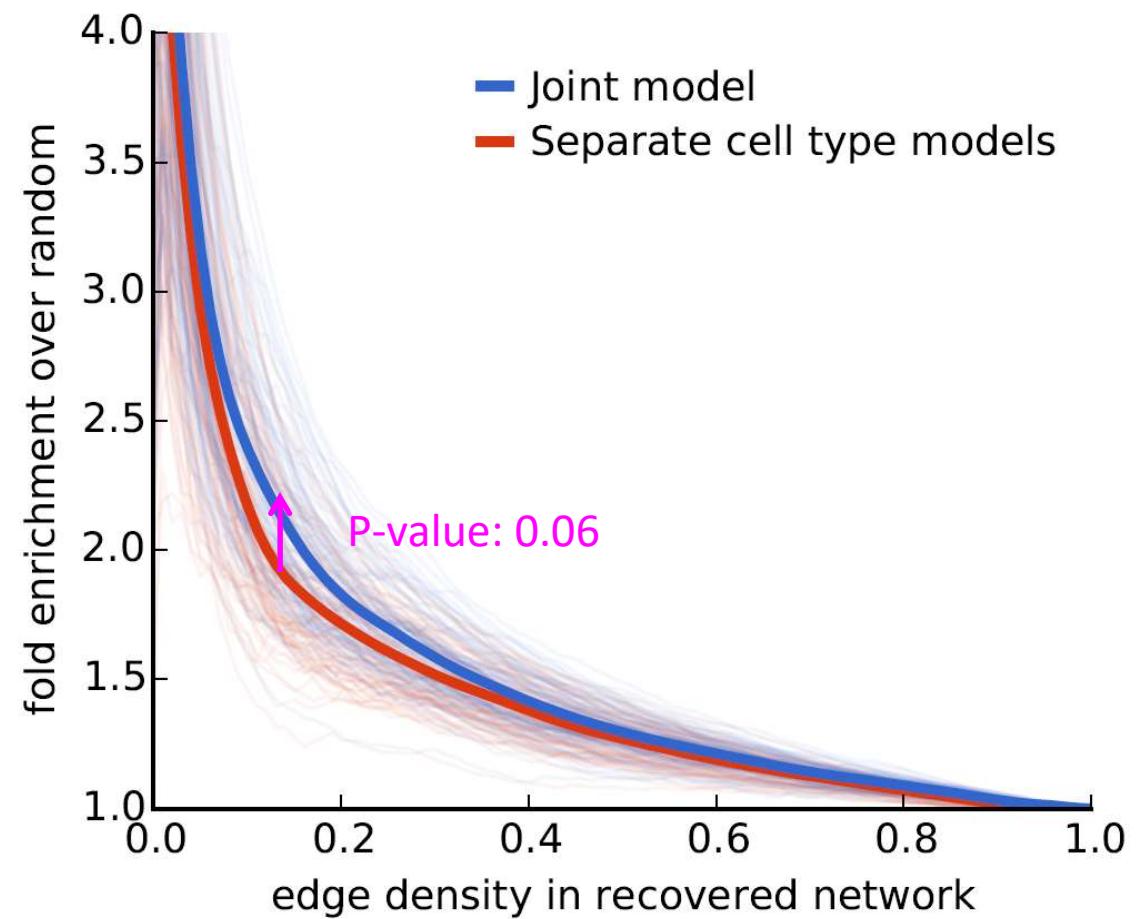
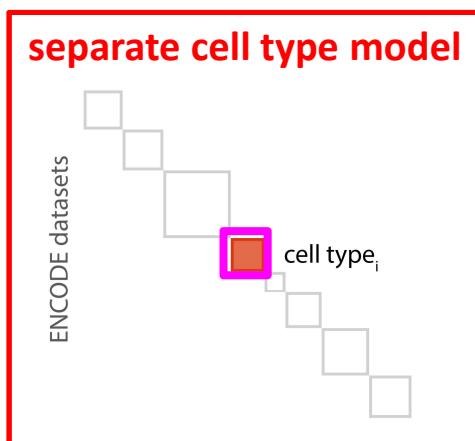
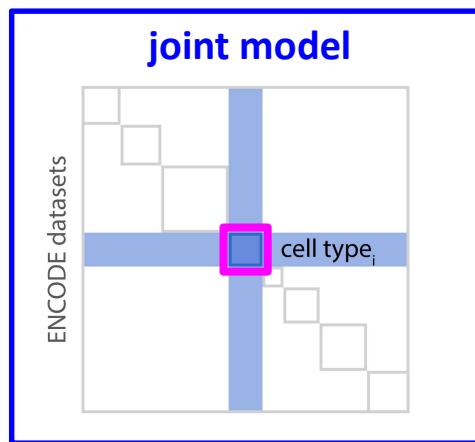


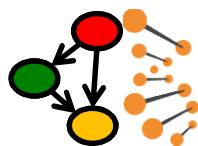
Increased number of unique factor-factor interaction types





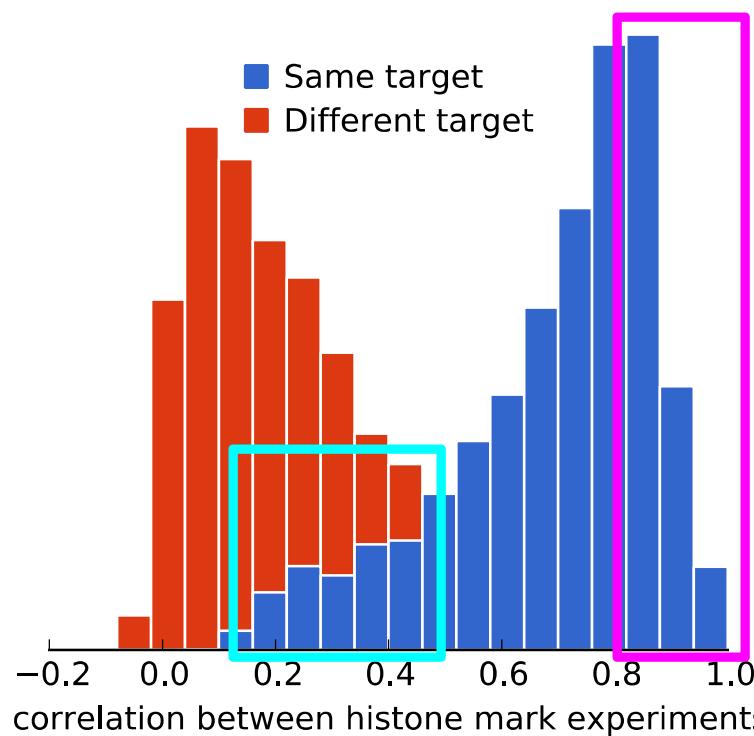
Focusing on only within cell type edges...



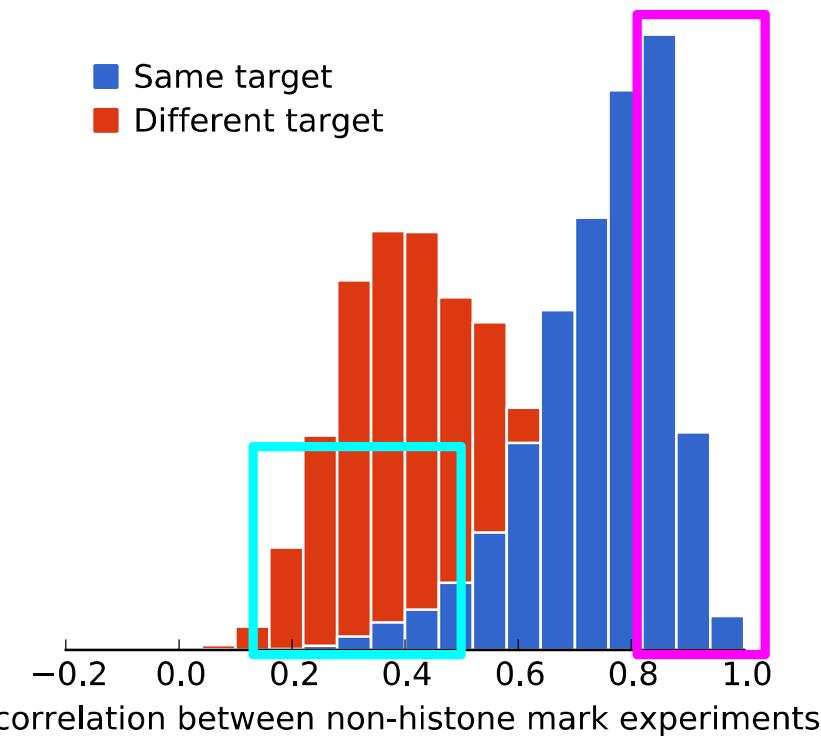


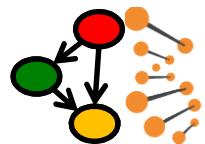
Correlation of the same factor between different cell types

Correlations between the same histone marks in different cell types

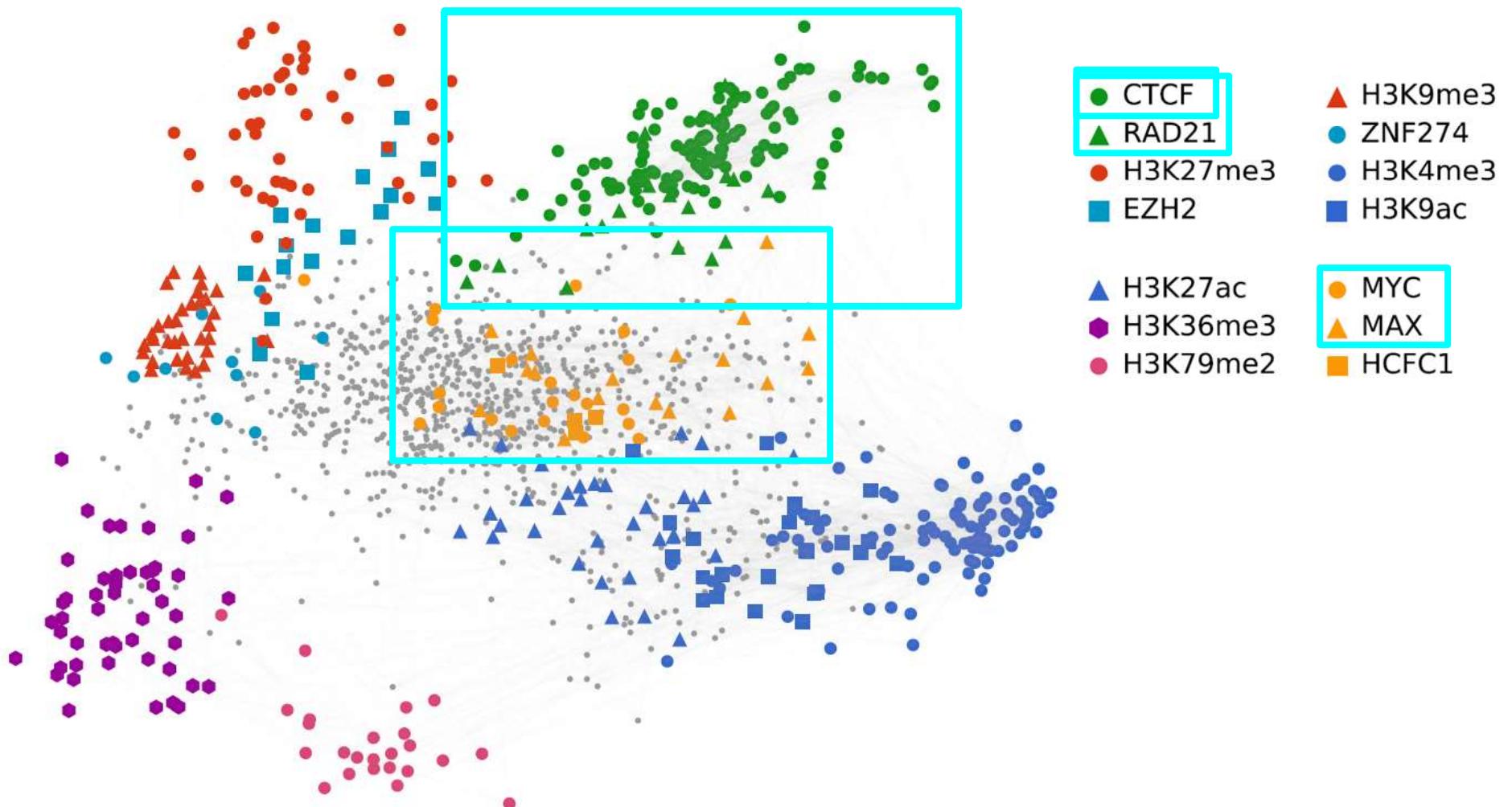


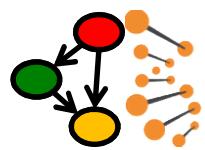
Correlations between the same transcription factors





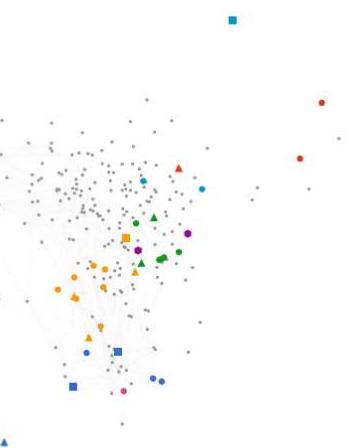
Global pattern of the entire network



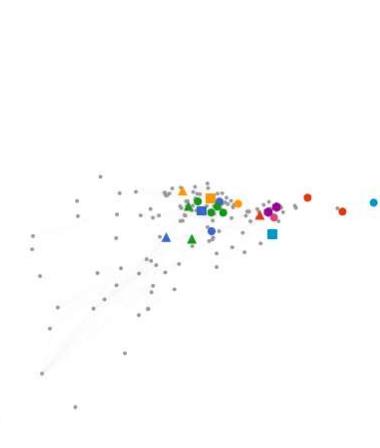


Global pattern in individual cell type networks

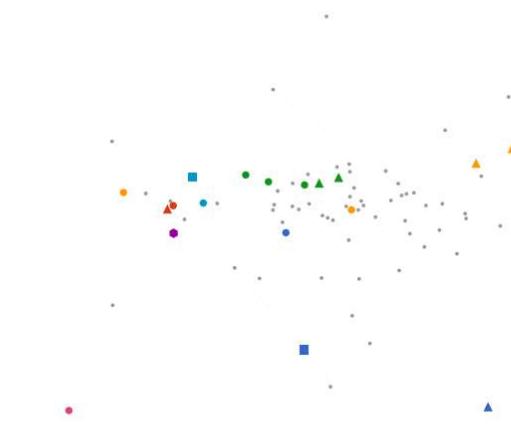
K562



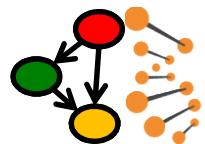
GM12878



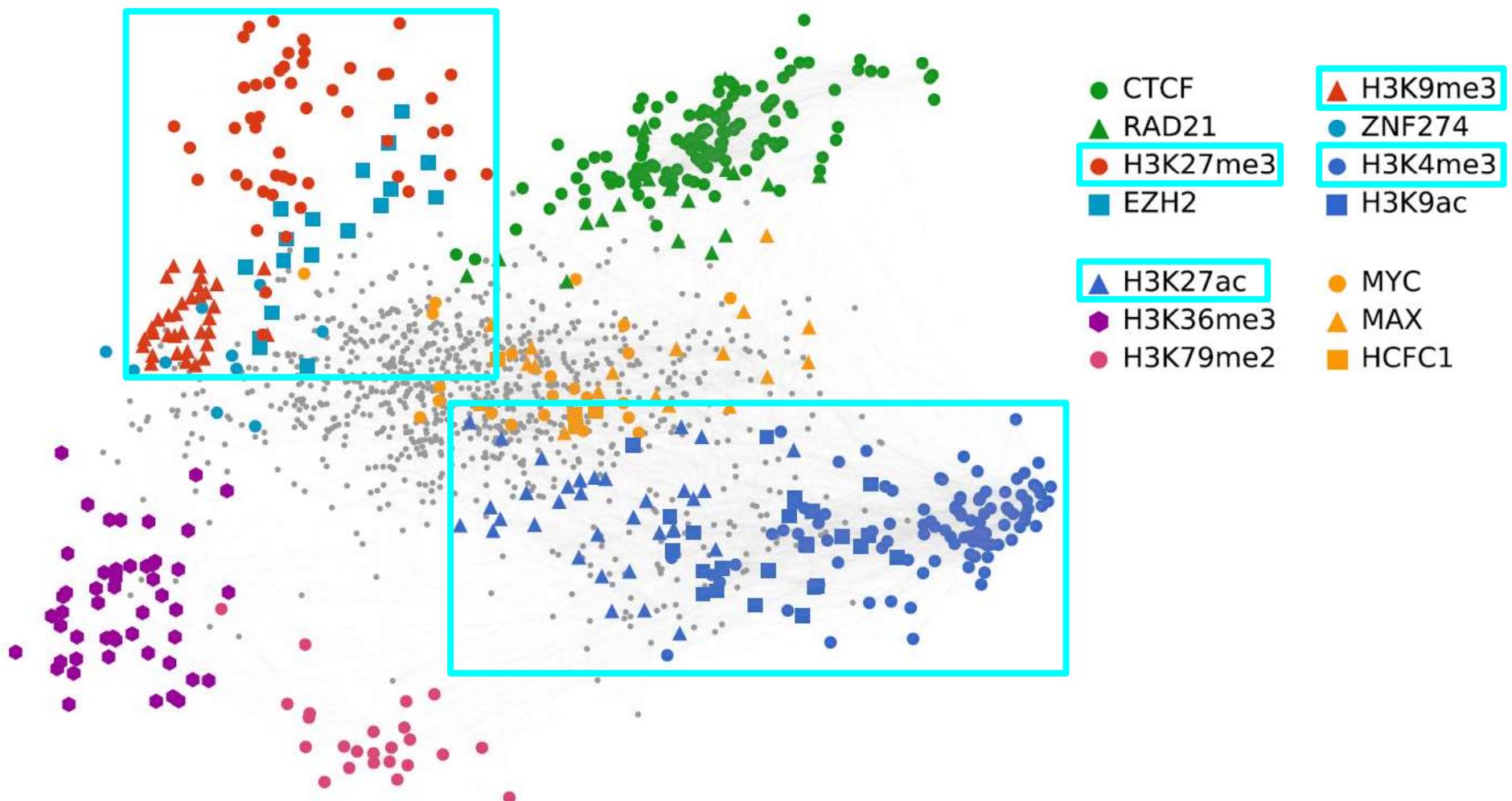
H1-hESC

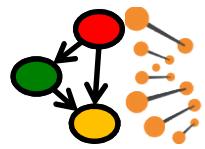


- | | | | |
|------------|-----------|------------|---------|
| ● CTCF | ▲ H3K9me3 | ▲ H3K27ac | ● MYC |
| ▲ RAD21 | ● ZNF274 | ● H3K36me3 | ▲ MAX |
| ● H3K27me3 | ● H3K4me3 | ● H3K79me2 | ■ HCFC1 |
| ■ EZH2 | ■ H3K9ac | | |



Global pattern of the entire network

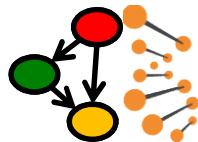




Outline

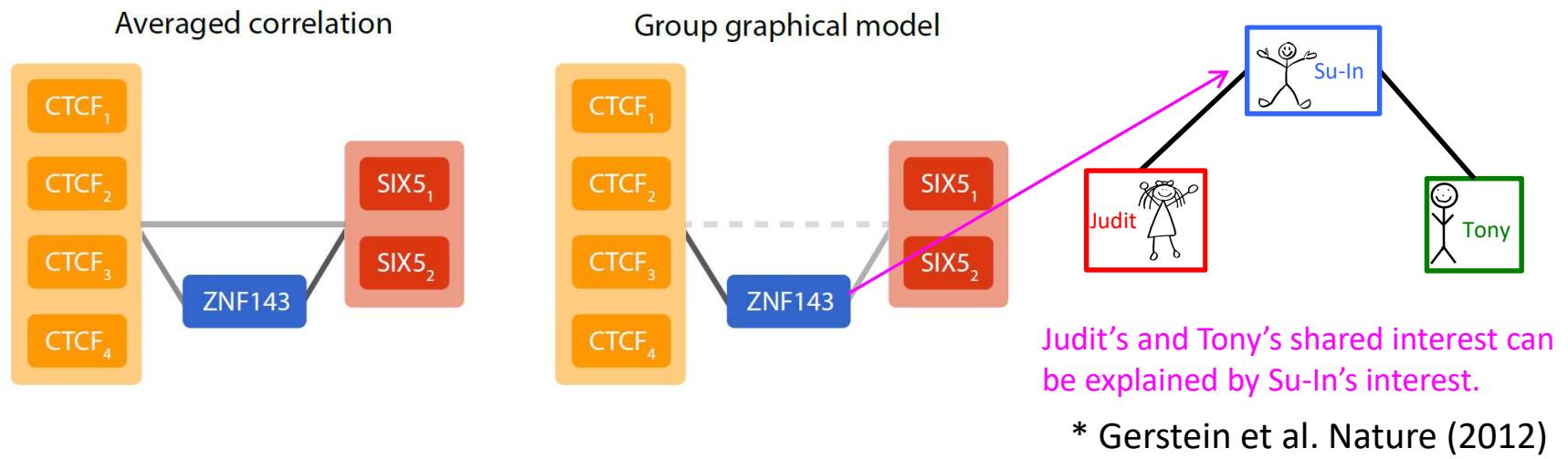
- Motivation
- Key features of ChromNet
 - Identifying conditional dependence
 - Group-based modeling
 - Learning a joint model of all cell types
 - Identifying context-specificity
- Visualization tool
- Discussion

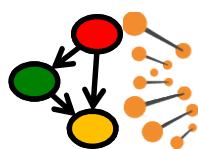




How context-specificity is encoded in a conditional dependence network?

- Transcription factors co-associate in a combinatorial and context-specific fashion.
 - For example, CTCF and SIX5 closely co-associate specifically when ZNF143 is also present.*
- This means that CTCF and SIX5 are *conditionally independent (not connected with each other)* given ZNF143
- ***CTCF – SIX5 relationship is context specific and ZNF is the context.***

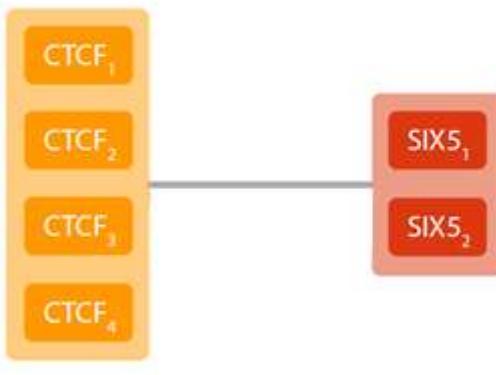




What if the mediator is not captured?

- It is important to understand the genomic context in which any given edge occurs.
 - We developed a new efficient method to estimate the contribution of the subset of data for each position to each network edge.
 - Say that ZNF143 is not present.

Group graphical model without ZNF143



Measuring the
difference in the
 $\Sigma^{-1}(\Sigma^{-1})_{ij}$

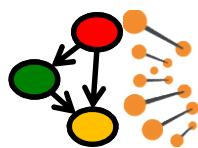
$$\begin{aligned}
\Sigma &= (D^{-1}(\Sigma - uu^T)D^{-1})^{-1} \\
&= D(\Sigma - uu^T)^{-1}D \\
&= D(\Sigma + uBu^T)^{-1}D \\
&= D(\Sigma^{-1} - \Sigma^{-1}u(B^{-1} + u^T\Sigma^{-1}u)^{-1}u^T\Sigma^{-1})D \\
&= D(\Sigma^{-1} - \Sigma^{-1}u(-1 + u^T\Sigma^{-1}u)^{-1}u^T\Sigma^{-1})D \\
&= D(\Sigma^{-1} - v(-1 + u^Tv)^{-1}v^T)D \\
&= D(\Sigma^{-1} - \frac{vv^T}{u^Tv - 1})D
\end{aligned}$$

estimated contribution of each position to CTCF – SIX5 edge



ATATGTGTTAGGATTATGCGCGAGGACGCATCATACTACTAGTAATGATTGATC

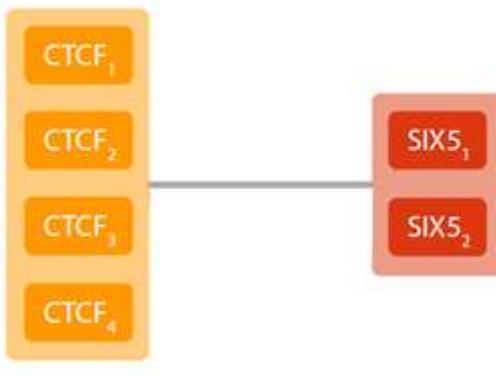
← genomic positions are samples →



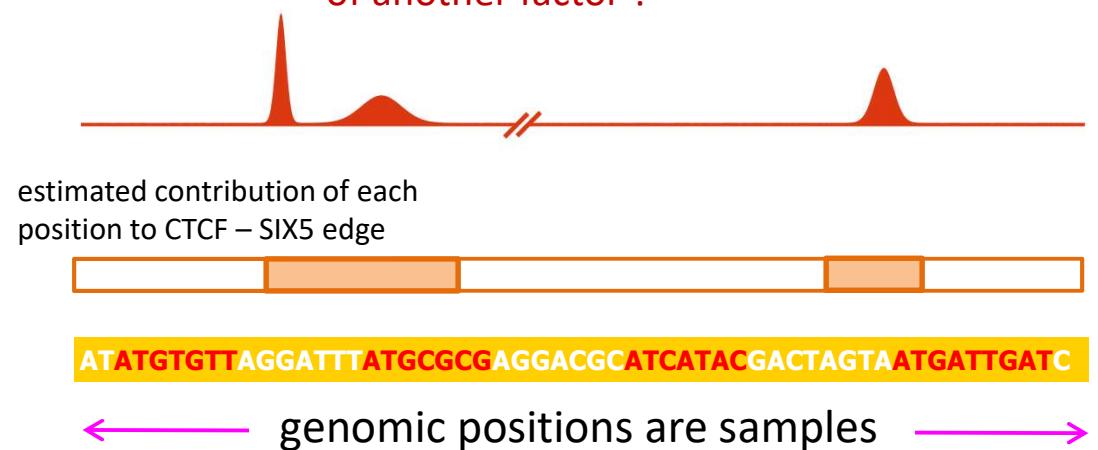
What if the mediator is not captured?

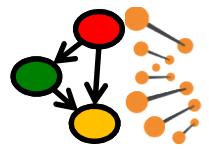
- It is important to understand the genomic context in which any given edge occurs.
- We developed a new efficient method to estimate the contribution of the subset of data for each position to each network edge.
 - Say that ZNF143 is not present.

Group graphical model without ZNF143

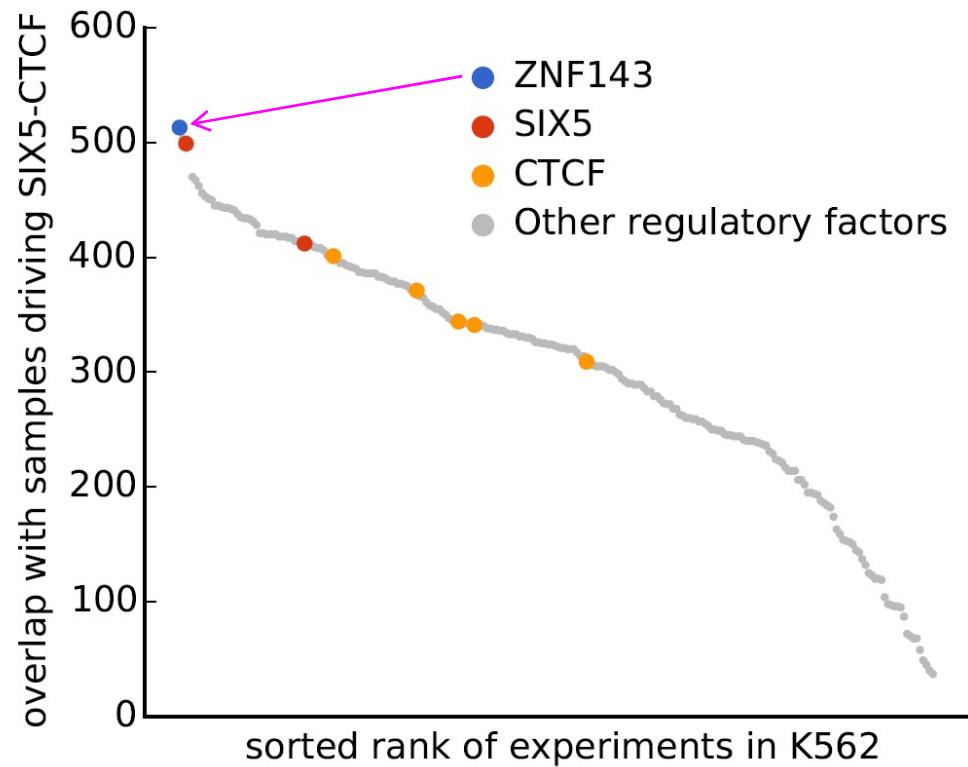


Does the estimated amount of contribution overlap with localization of another factor ?

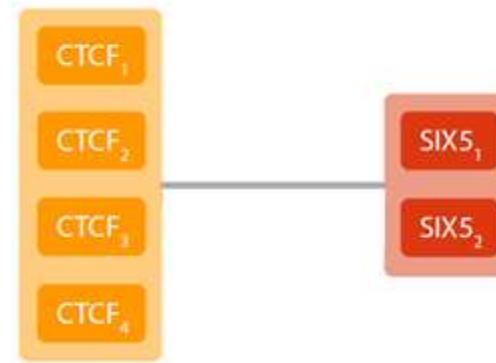




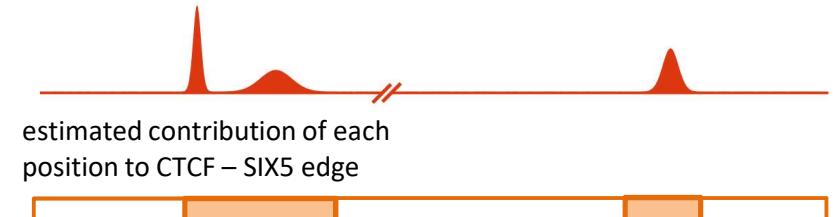
ZNF143 is captured as the context

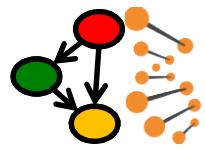


Group graphical model without ZNF143



Does the estimated amount of contribution overlap with localization of another factor ?

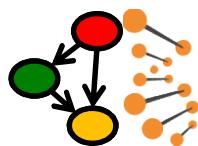




Outline

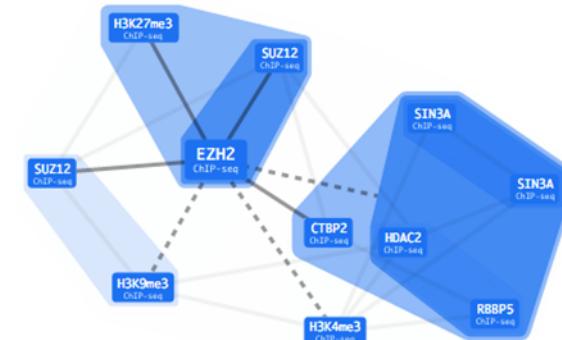
- Motivation
- Key features of ChromNet
 - Identifying conditional dependence
 - Group-based modeling
 - Learning a joint model of all cell types
 - Identifying context-specificity
- Visualization tool
- Discussion





Simultaneously visualizing groups and network edges

- Easy access to the network information is important.
 - Interactive tool to navigate the human chromatin network
 - <http://chromnet.cs.washington.edu/#/>
- The group graphical model perspective requires a unique visual encoding.
- This encoding must **simultaneously display both clusters and networks**.
- Ideally it should be both dynamic and interactive in order to allow exploration.





Cell Types and Treatments

H1-hESC

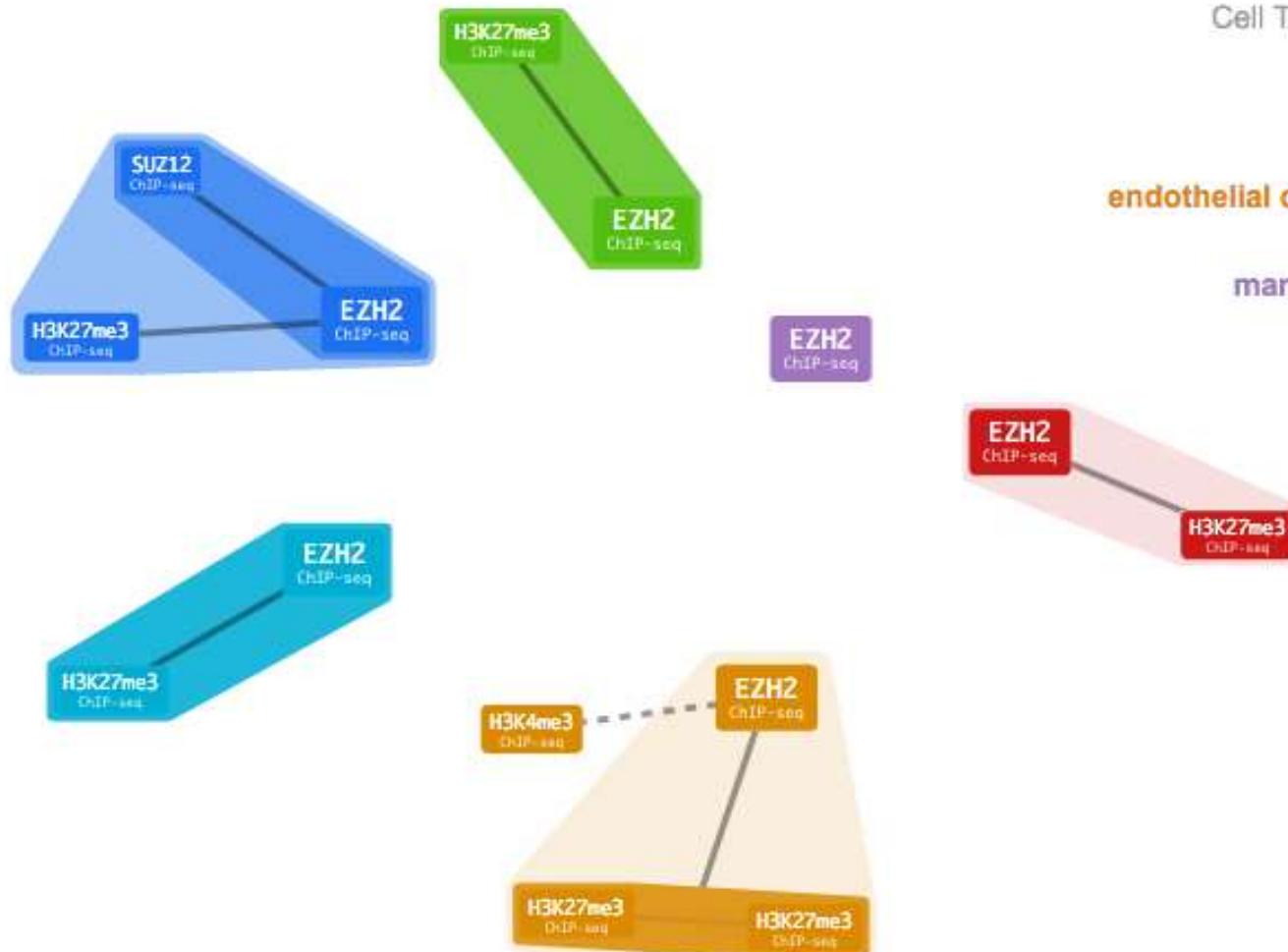
HeLa-S3

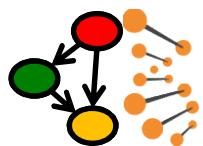
HepG2

endothelial cell of umbilical vein

keratinocyte

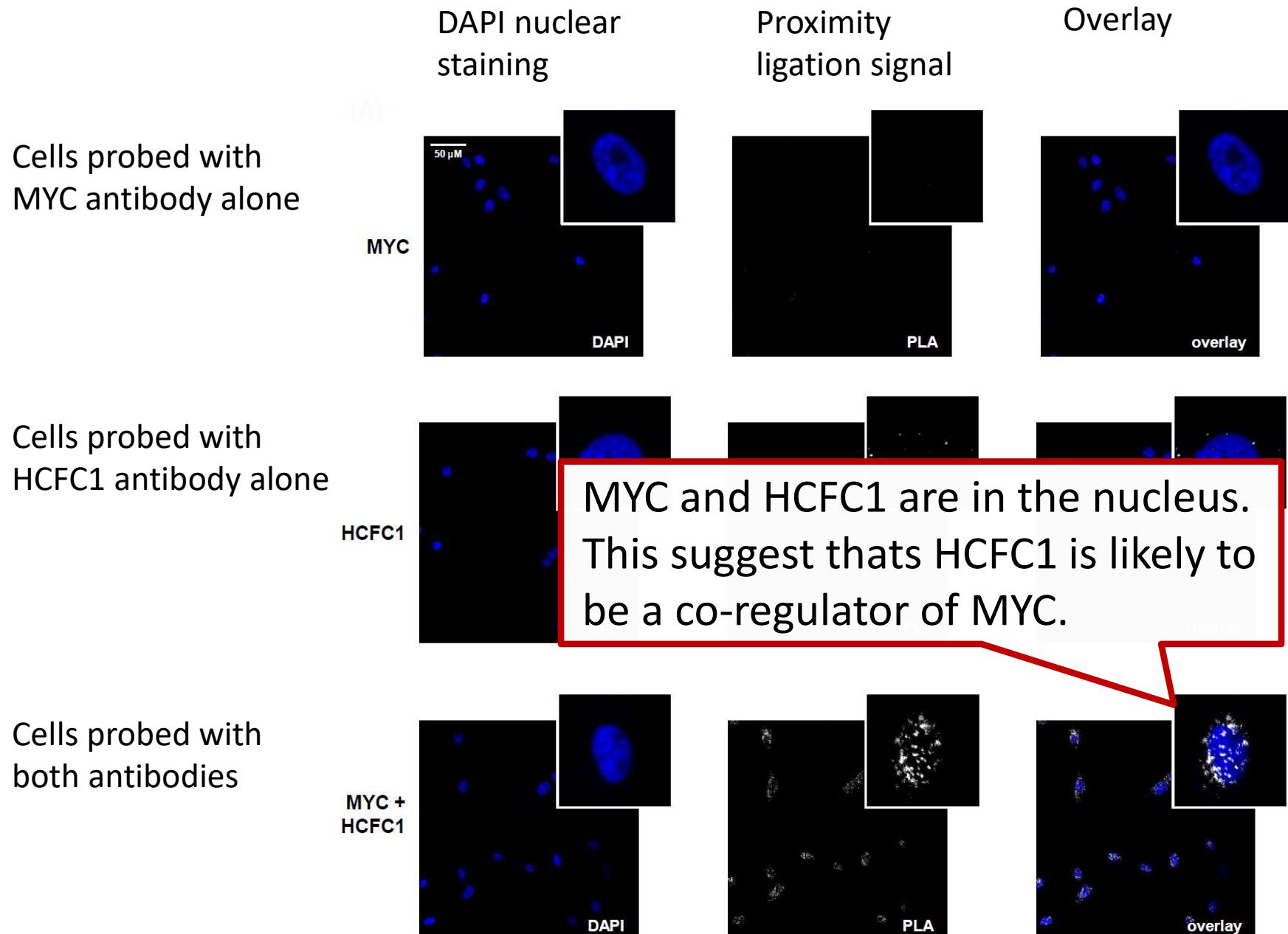
mammary epithelial cell

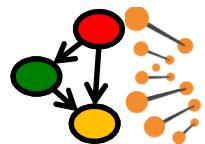




An example of a novel interaction MYC-HCFC1 and wet-lab validation

- One of the strongest edges that connect proteins **not known to interact with each other** is MYC-HCFC1 (Host Cell Factor C1)
 - **MYC:** cell growth, cell cycle progression, **oncogenesis**
 - **HCFC1:** cell cycle progression, **oncogenesis**
- To validate the potential novel MYC-HCFC1 interaction, we performed a **proximity ligation assay (PLA)** in MCF10A mammary epithelial cells.
- When two proteins that are probed with specific antibodies are within **close proximity of each other**, **fluorescence signals are produced** that are measured and quantified using fluorescence microscopy.





More information is available

- Preprint is available on bioRxiv
- The ChromNet browser
 - <http://chromnet.cs.washington.edu/>



Learning the human chromatin network from all ENCODE ChIP-seq data

Scott M. Lundberg¹, William B. Tu^{2,3},
Brian Raught^{2,3}, Linda Z. Penn^{2,3}, Michael M. Hoffman^{2,3,4}, Su-In Lee^{1,5}

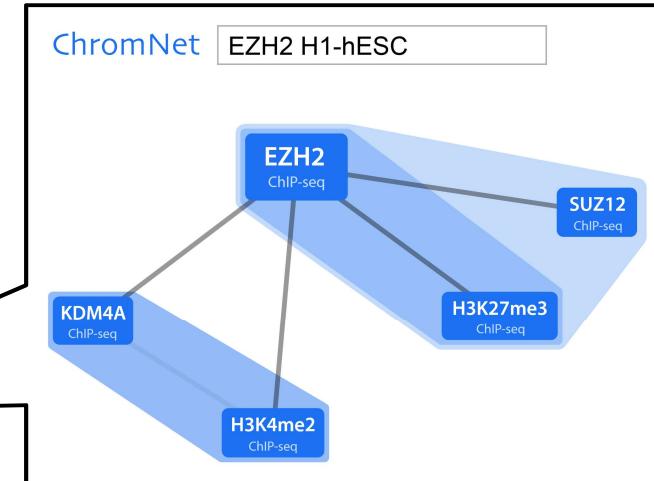
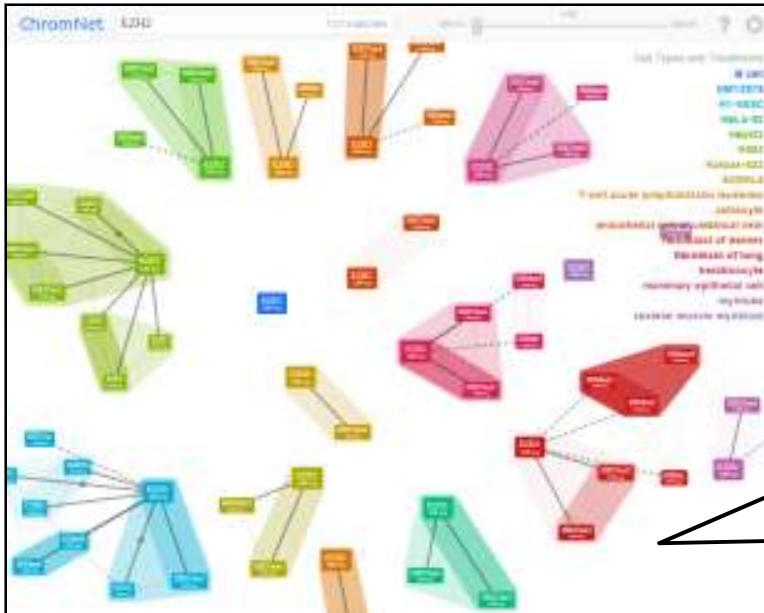
¹ Department of Computer Science and Engineering, University of Washington

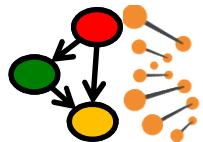
² Department of Medical Biophysics, University of Toronto

³ Princess Margaret Cancer Centre

⁴ Department of Computer Science, University of Toronto

⁵ Department of Genome Sciences, University of Washington





Acknowledgements

UW CSE

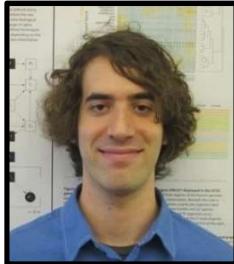
Genome Sciences



Scott Lundberg



Bill Noble



Michael Hoffman



Linda Penn



William Tu



Brian Raught

computational biologist

experimental biologists

Funding:

