Reconstructing bayesian networks from genomics dataset
- a case study -

Carl Herrmann, Ashwini Sharma
Cancer Regulatory Genomics
Universität Heidelberg & DKFZ

Content of the presentation

- a brief introduction to the theory behind Bayesian networks
  (based on slides by Marco Scutari)

- a example of chromatin network reconstruction in Neuroblastoma

- Practical example / demo
  - Reconstruction the BN of T-cell signaling pathway (Sachs et al., 2005)
  - Reconstruction of the BN for chronic lymphocytic leukemia patients

- Presentation, data, R Markdown scripts can be found here:
  `https://github.com/crg-eilslabs/SEEDED`

Current challenges

RNA-seq

ATAC - seq

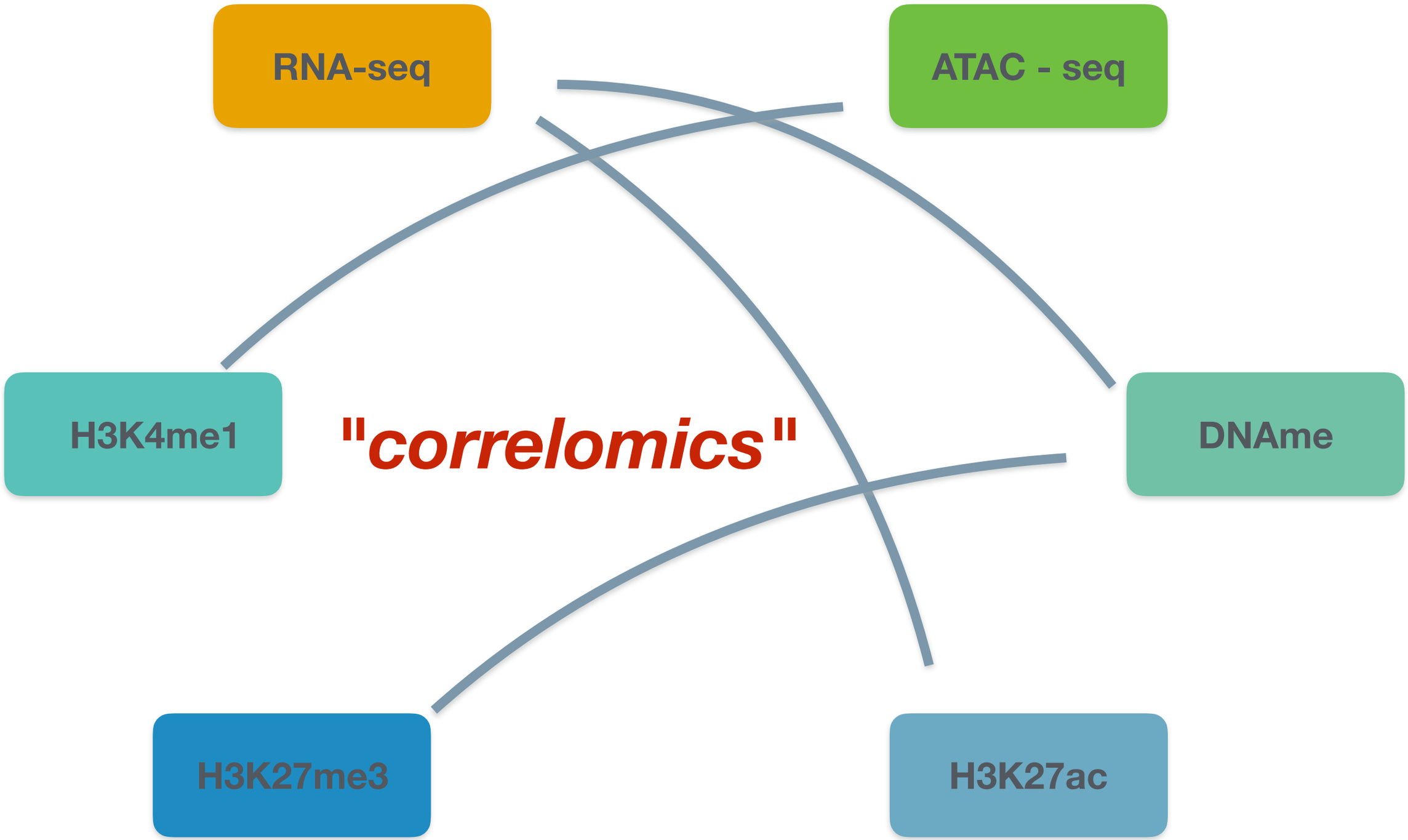
H3K4me1

DNAme

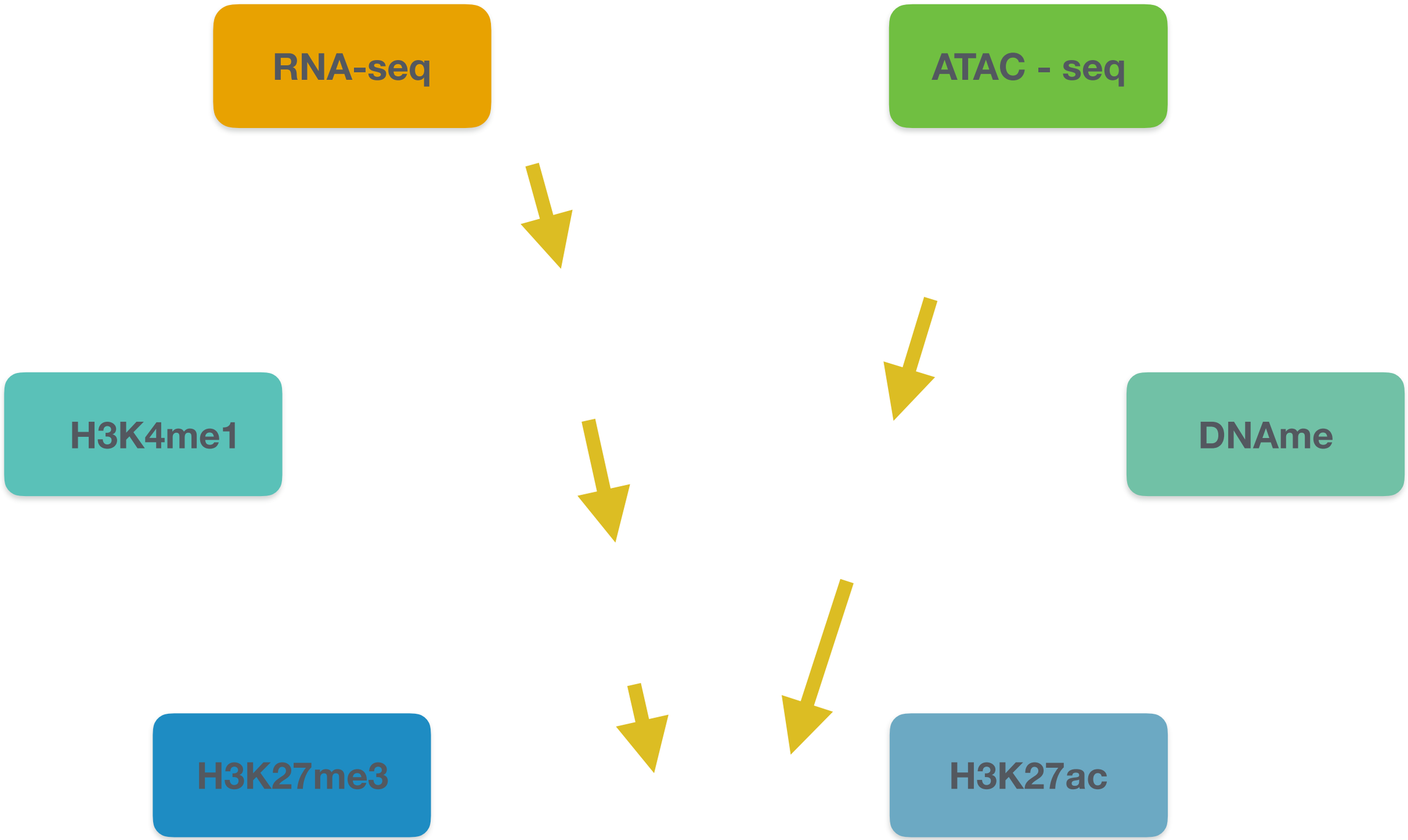
H3K27me3

H3K27ac

Current challenges



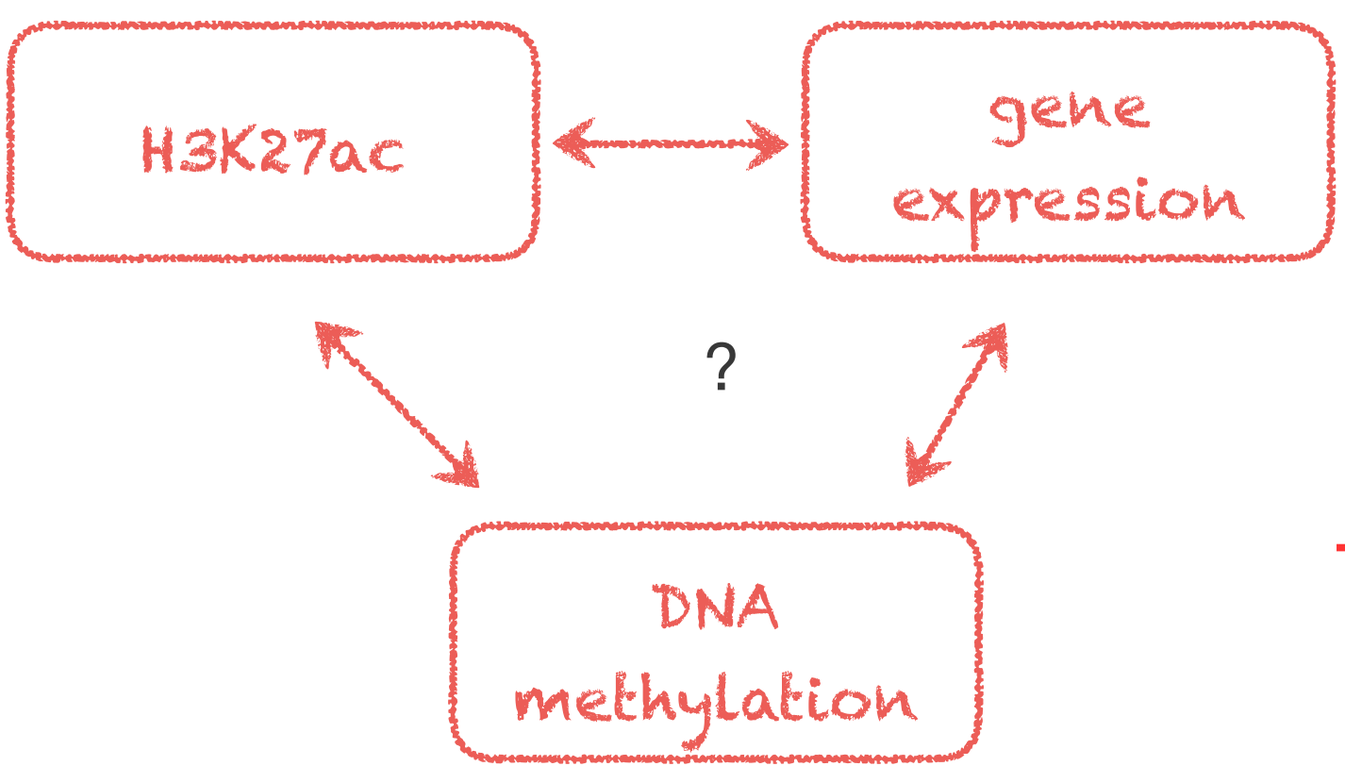
Current challenges



Modelling chromatin networks

- Most genomic analysis is based on correlation between features
  → ***"correlomics"***

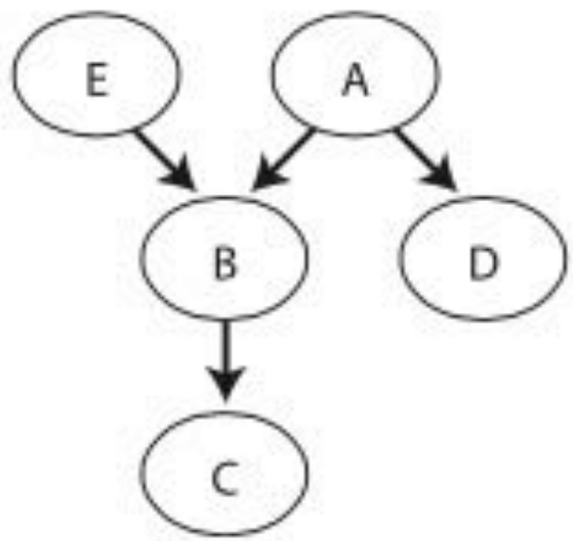
- Can we go beyond towards oriented/causal networks ??



→ ***Bayesian Networks***

What are bayesian networks ?

- Networks of **dependencies**
  - A and E are independent
  - B depends both on A and E
  - C depends on B



- Not necessarily **causal** networks !
  - direction of edges is not always well defined

- Need interventional data (perturbations) to turn a BN into a causal network
  - B is the direct and unique cause of C

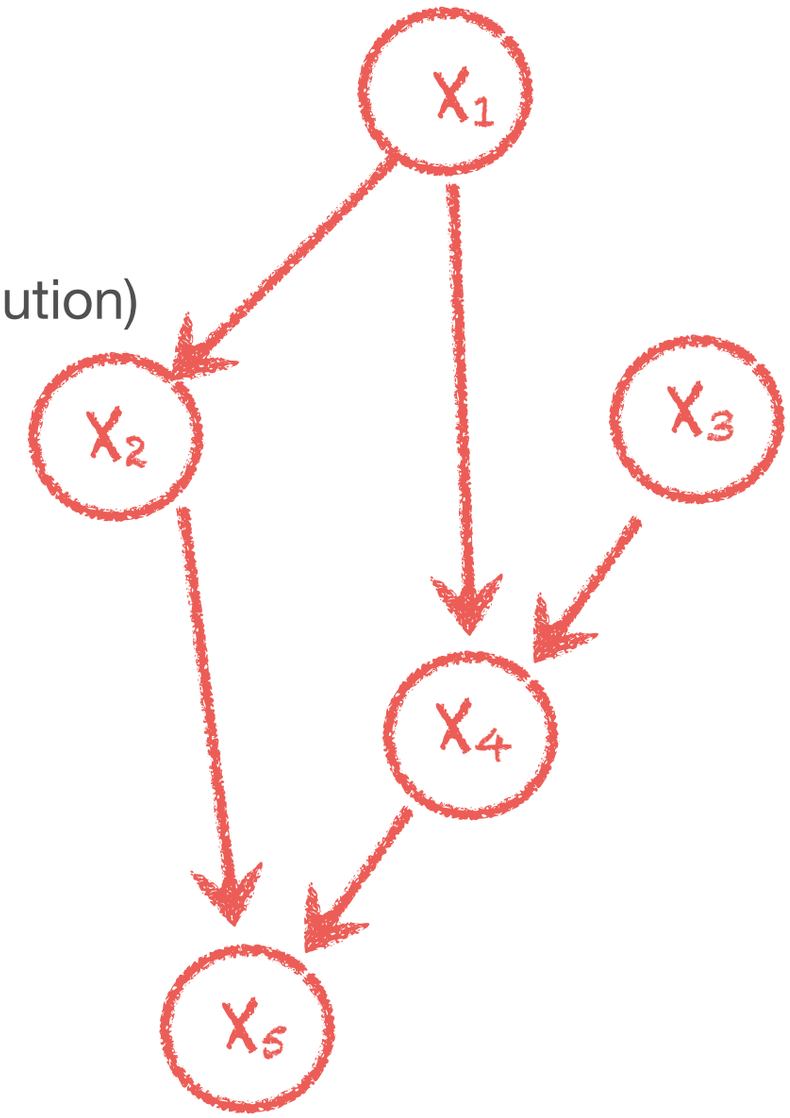
What are bayesian networks ?

- **Directed acyclic graph** $G = (V,A)$ with nodes V and edges A
- each node v_i is a random variable X_i
- random variable X_i can be
  - **discrete** (multinomial variables high/mid/low)
  - **continuous** (modelled as multivariate gaussian distribution)
- we want to compute the joint probability $P(X_1,X_2,\ldots X_n)$

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|X_{j\neq i})$$
$$= \prod_{i=1}^{n} P(X_i|pa(X_i))$$

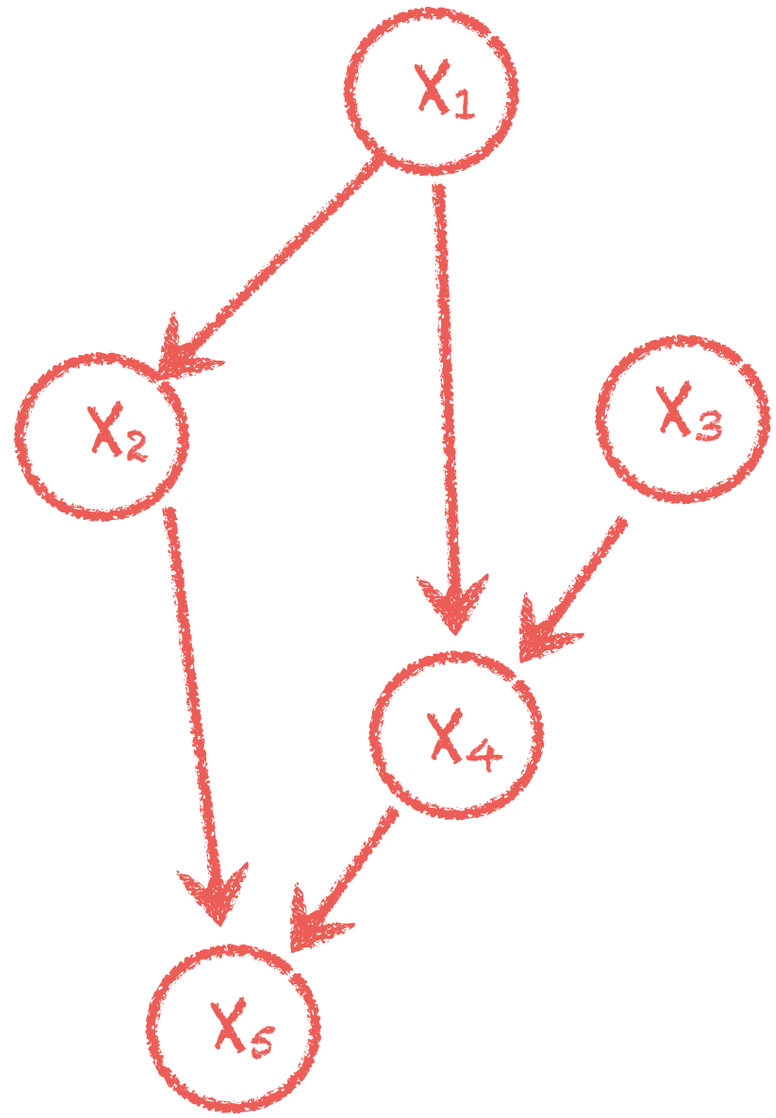
parents of X_i

[Friedman, 2004,Friedman et al., 2000]



Conditional (in)dependence

- conditional (in)dependence: knowing the state of some nodes makes others nodes (in)dependent
- example:
  - X_3 has an indirect effect on X_5; but knowing the state of X_4 makes X_5 and X_3 independent (if I know X_4, then X_5 does not give me additional information
    → ***X_4 d*-separates X_3 and X_5**
  - X_1 and X_3 are independent; but knowing the state of X_1 AND X_4 gives me additional information on X_3



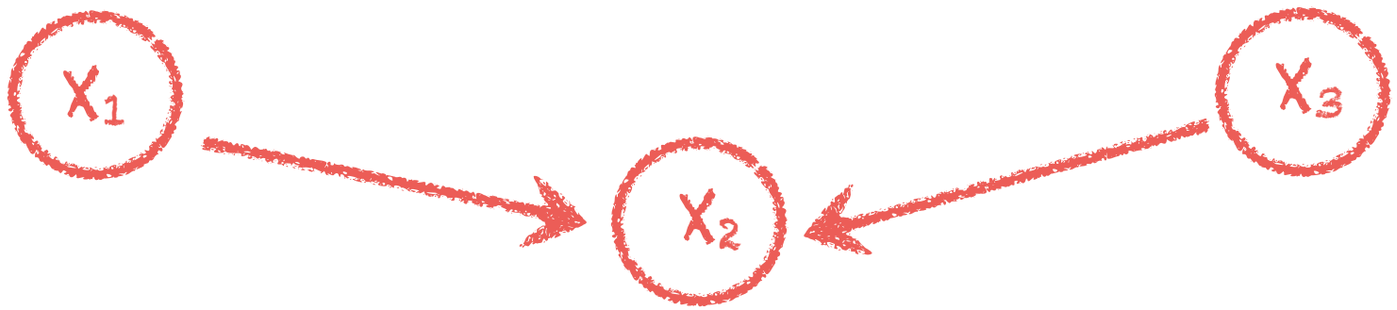
[Friedman, 2004,Friedman et al., 2000]

Conditional (in)dependence

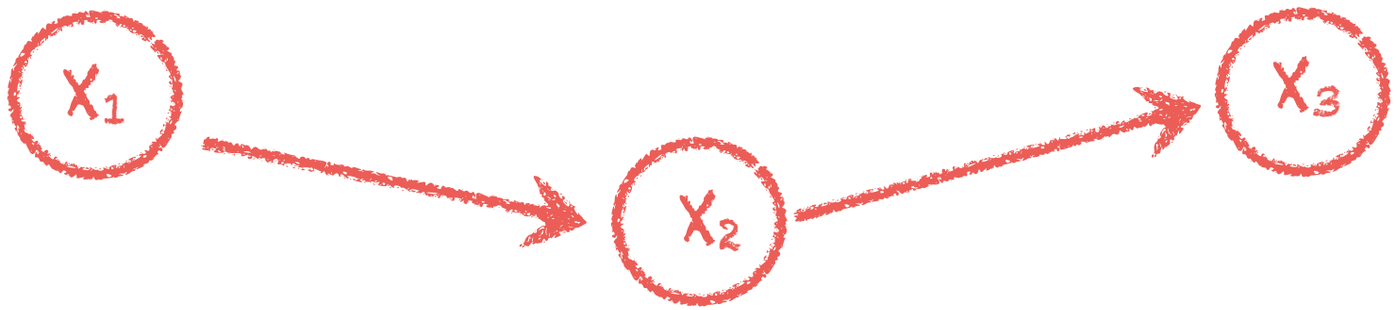
- high earnings: wheather was probably good
- if harvest was low: knowing that earning was high does not give me more information on the weather condition
- the middle node **d-separates** the 2 external ones

- sportiness and academic achievements are independent
- but if I know that someone was admitted to college and is very sporty, this lowers the probability of high academic achievements.
- ***"v-structure"***

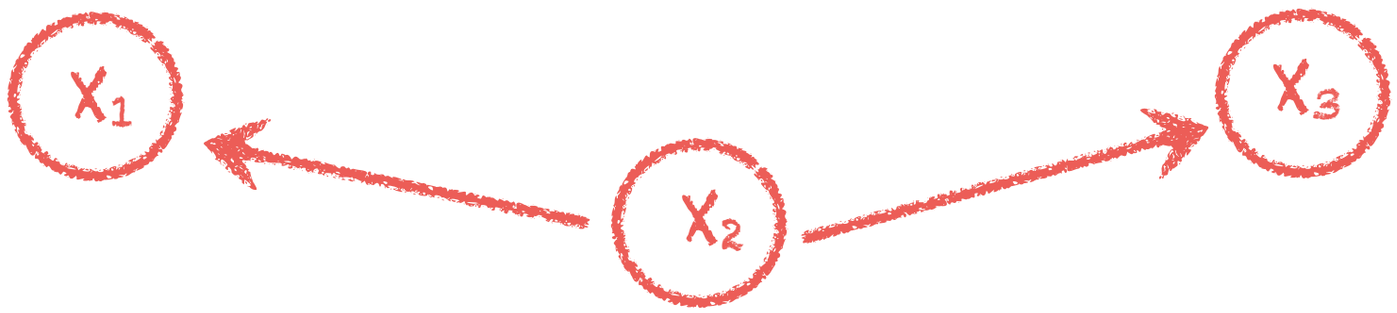
Equivalence



$$P(X_1, X_2, X_3) = P(X_2|X_1, X_3)P(X_1)P(X_3)$$



$$P(X_1, X_2, X_3) = P(X_3|X_2)P(X_2|X_1)P(X_1)$$



$$P(X_1, X_2, X_3) = P(X_3|X_2)P(X_1|X_2)P(X_2)$$

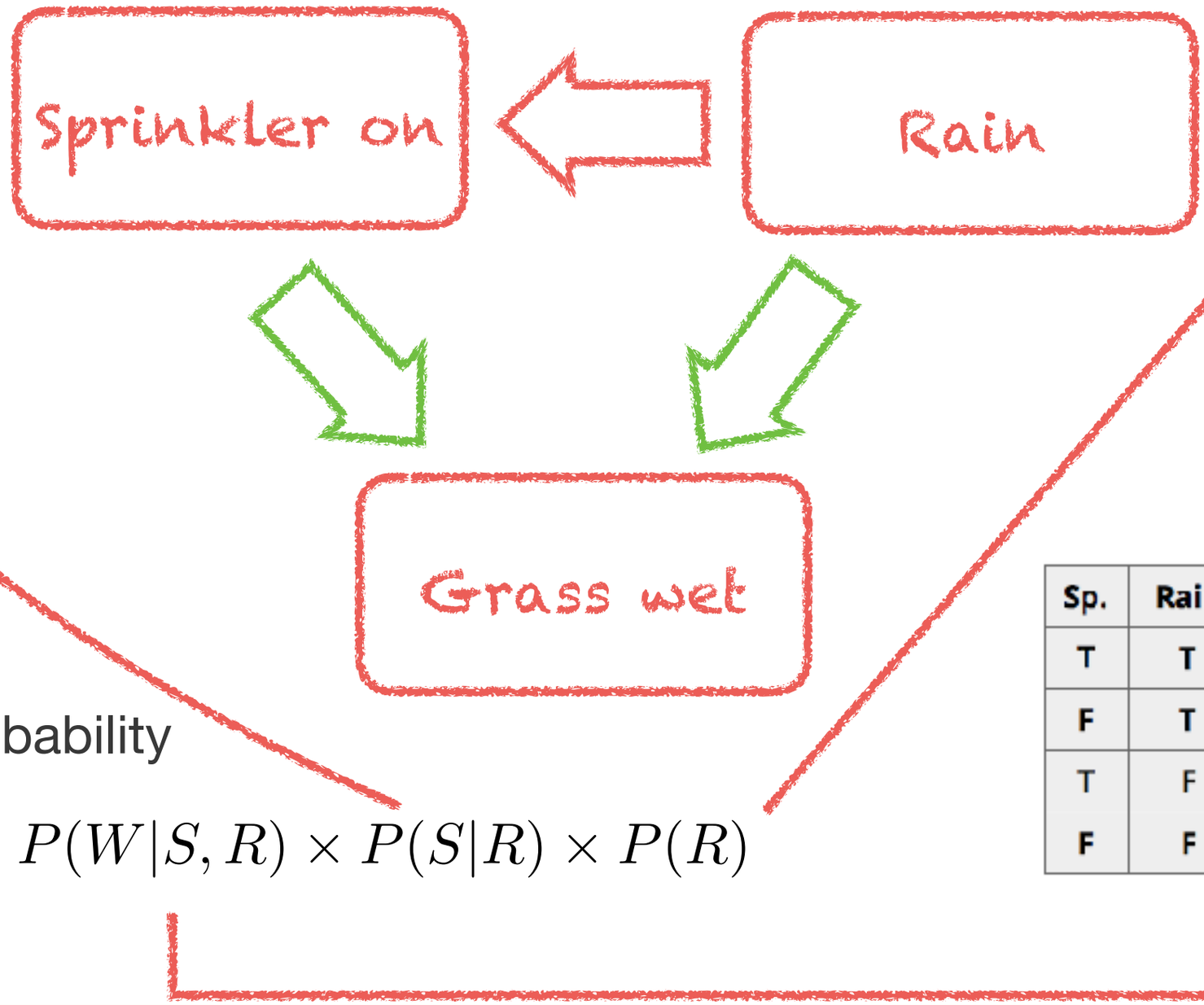
These 2 networks have same probability → *equivalence*

$$P(X_1|X_2)P(X_2) = P(X_2|X_1)P(X_1)$$

Discrete Bayesian Network

Rain

Sprinkler	T	F
T	0.01	0.99
F	0.4	0.6



Rain

T	F
0.2	0.8

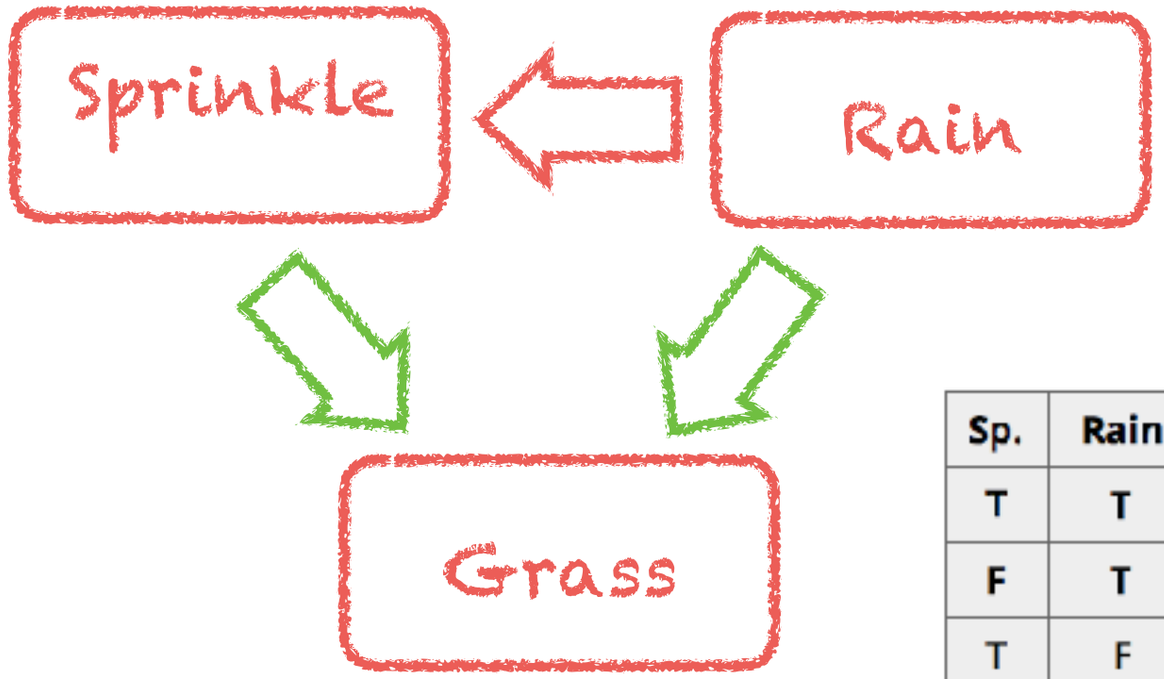
Joint probability

$$P(W, S, R) = P(W|S, R) \times P(S|R) \times P(R)$$

Grass wet

Sp.	Rain	T	F
T	T	0.99	0.01
F	T	0.8	0.2
T	F	0.9	0.1
F	F	0	1

Discrete Bayesian Network



Rain

	T	F
0.2	0.8	

Sprinkler

	T	F
T	0.01	0.99
F	0.4	0.6

Grass wet

Sp.	Rain	T	F
T	T	0.99	0.01
F	T	0.8	0.2
T	F	0.9	0.1
F	F	0	1

Joint probability:

$$P(W, S, R) = P(W|S, R) \times P(S|R) \times P(R)$$

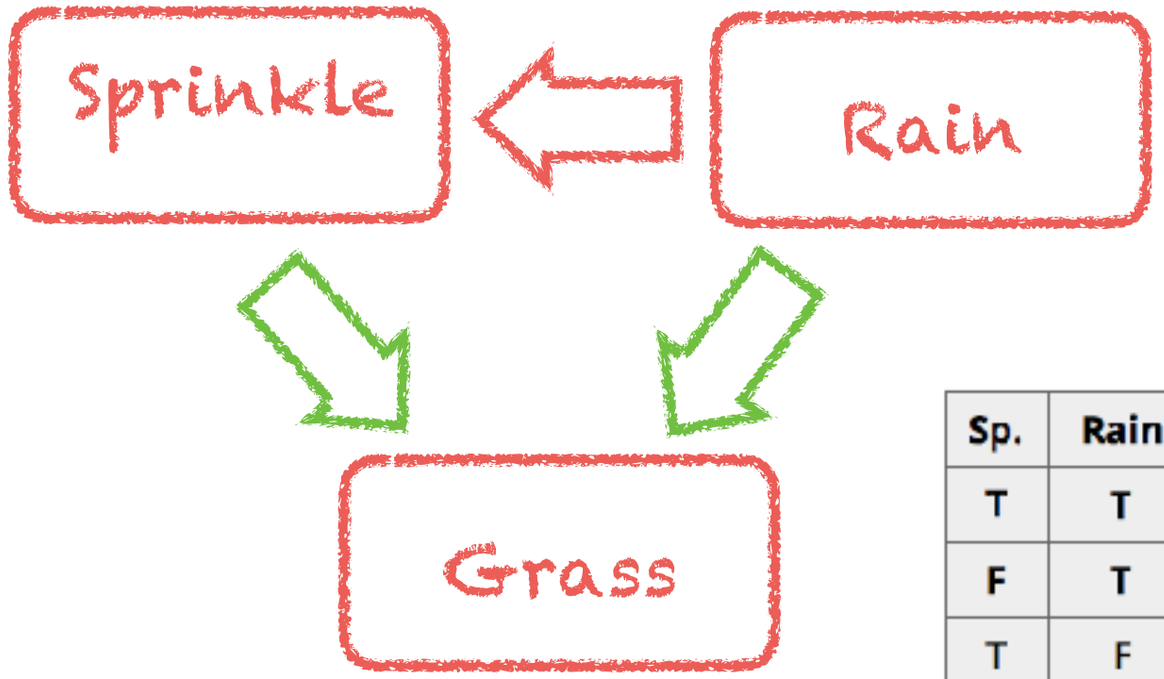
What is the probability that the grass is wet, the sprinkler on and it rains ?

$$\begin{aligned} P(W = 1, S = 1, R = 1) &= P(W = 1|S = 1, R = 1)P(S = 1|R = 1)P(R = 1) \\ &= 0.99 \times 0.01 \times 0.2 \\ &= 0.00198 \end{aligned}$$

Inference

Rain

	T	F
0.2	0.8	



Sprinkler / **Rain**

	T	F
T	0.01	0.99
F	0.4	0.6

Grass wet

Sp.	Rain	T	F
T	T	0.99	0.01
F	T	0.8	0.2
T	F	0.9	0.1
F	F	0	1

What is the probability that the grass is wet given that the sprinkler is on ?

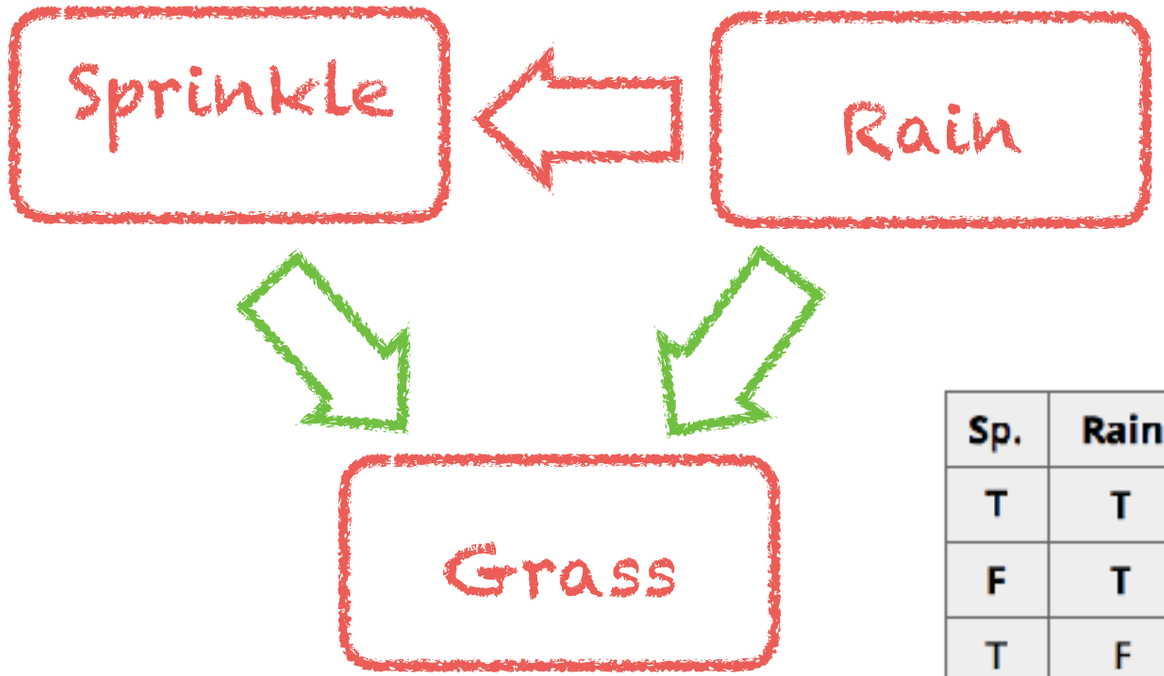
→ **sum over marginal variable rain (unobserved variable)**

$$
\begin{aligned}
P(W=1|S=1) &= \frac{1}{P(S=1)} \sum_{r \in 0,1} P(W=1, S=1, r) \\
&= \sum_{r \in 0,1} P(W=1|S=1,r) P(S=1|r) P(r) \\
&= 0.9 \times 0.99 \times 0.8 + 0.99 \times 0.01 \times 0.2 \\
&= 0.714
\end{aligned}
$$

Inference

Rain

	T	F
T	0.2	0.8



Sprinkler

	Rain T	Rain F
T	0.01	0.99
F	0.4	0.6

Grass wet

Sp.	Rain	T	F
T	T	0.99	0.01
F	T	0.8	0.2
T	F	0.9	0.1
F	F	0	1

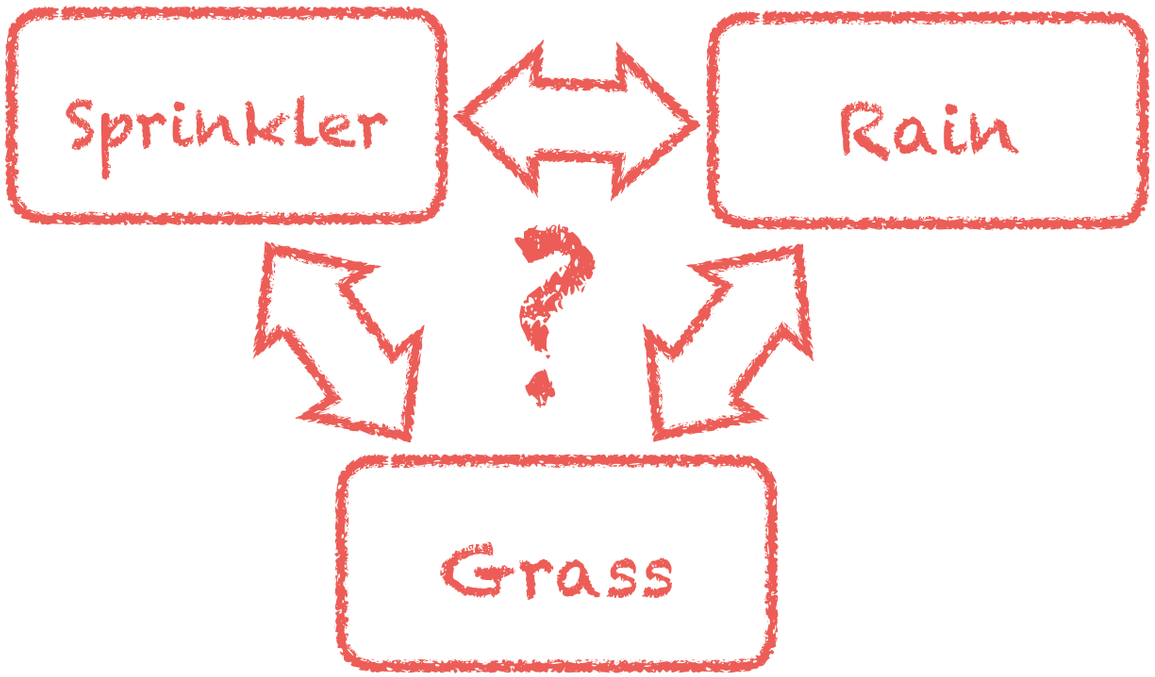
What is the probability that it rains, given that I observed wet gras ?

→ **integrate over marginal variable sprinkler**

$$
\begin{aligned}
P(R=1|W=1) &= \frac{1}{P(W=1)} \sum_{s \in 0,1} P(W=1, R=1, s) \\
&= \sum_{s \in 0,1} P(W=1|R=1,s) P(s|R=1) P(R=1) \\
&= 0.99 \times 0.01 \times 0.2 + 0.8 \times 0.4 \times 0.2 \\
&= 0.066
\end{aligned}
$$

Structure learning

What is the most likely network given the observed data ?



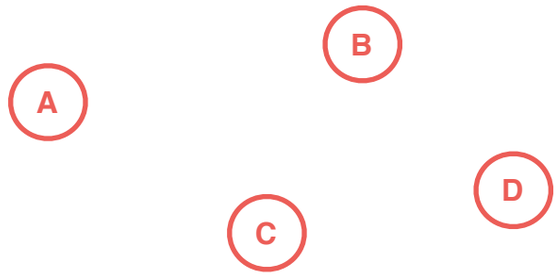
Day	Rain	Spr.	Grass wet ?
1	yes	no	no
2	no	no	no
3	no	yes	yes
4	yes	no	yes
5	...	...	...

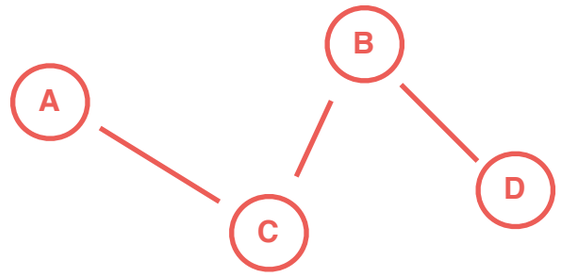
Important assumption : all observations are sampled from the same random variable !

However, (unobserved) confounding variables could violate this assumption (influence of seasons ?)

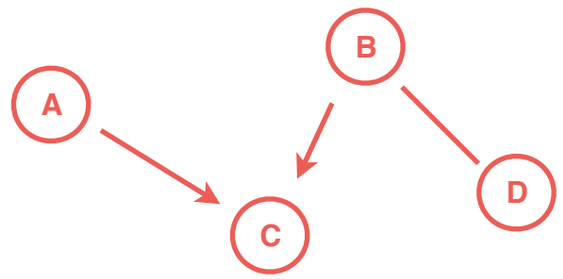
Methods for structure learning

- 2 classes of methods
- ***"constrained based methods"***
  - identify the pairs of nodes which **cannot be made conditional independent**
  - relate these nodes by undirected edge
  - if C does not *d*-separate A and B, then form a *v*-structure
  - apply heuristics to (possibly) direct the still undirected arcs (if A → B and B - C , then orient B→ C)



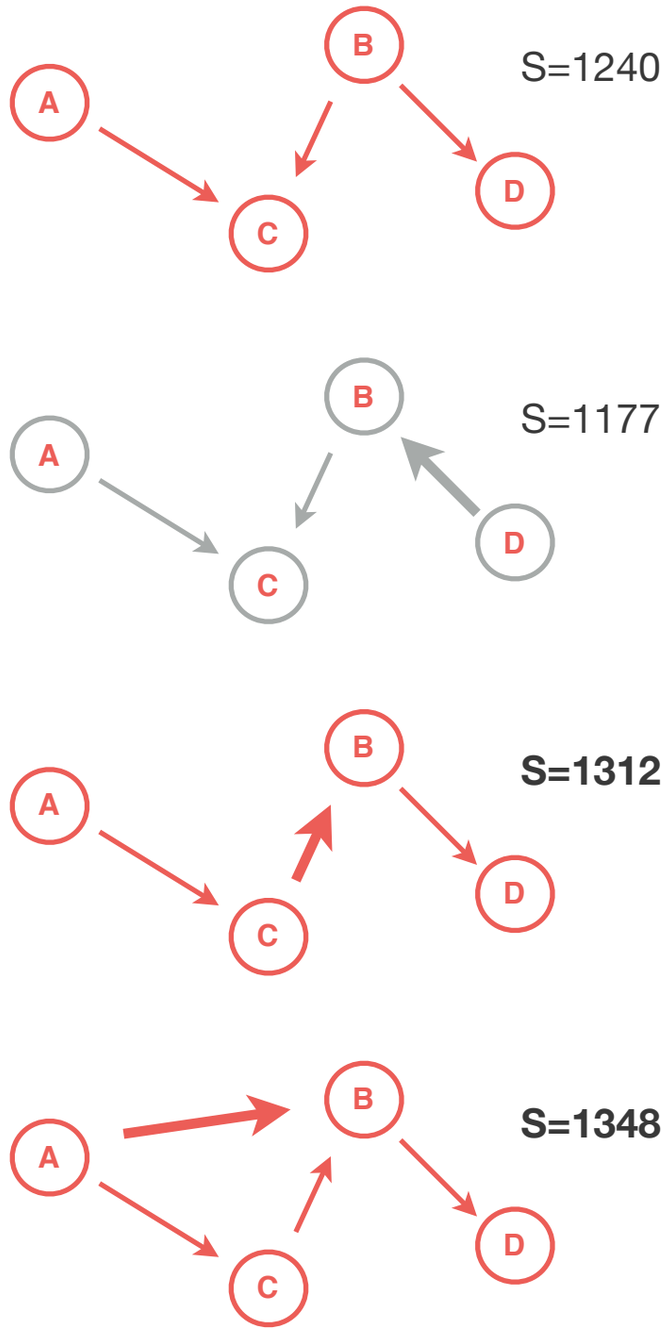


- this algorithm can be refined to handle the exponential complexity of the procedure ! (detection of Markov blanket)



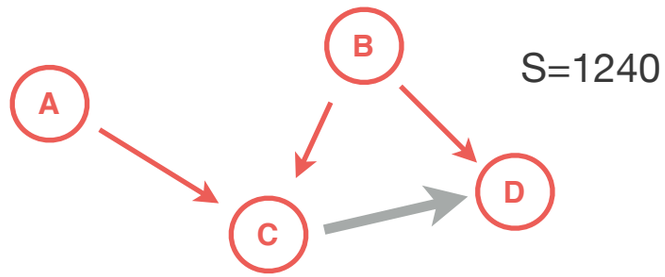
Methods for structure learning

- 2 classes of methods
- ***"score based methods"***
  - assign a likelihood score to each possible network
  - select the network with the highest likelihood score
- heuristics needed to handle the exponential number of possible networks !
  - hill-climbing : modify the current network slightly and check if this improves the score
  - improvements to **avoid local optima**:
    - *tabu search*: allow search to proceed around local optimum, **avoiding previous tested solutions**
    - *simulated annealing*: several random initialization, allow steps which degrade the score

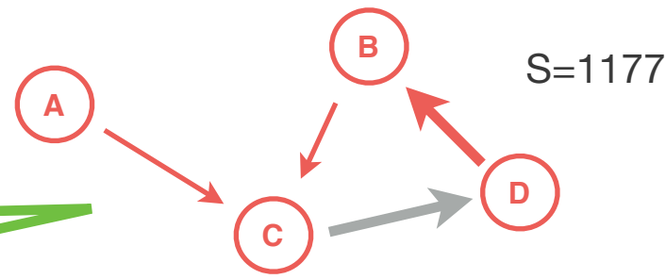


Methods for structure learning

- some edges can be
  - **whitelisted**: they should be present in the final network, based on literature evidence, etc…
  - **blacklisted**: these should NOT appear in the network (unrealistic relations)

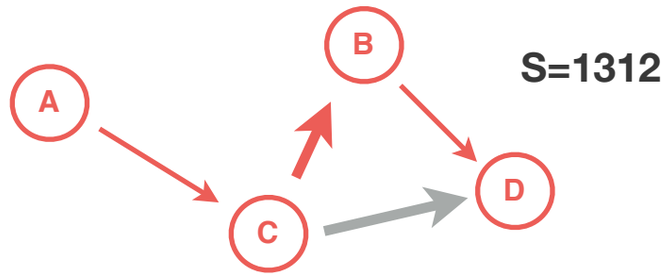


S=1240

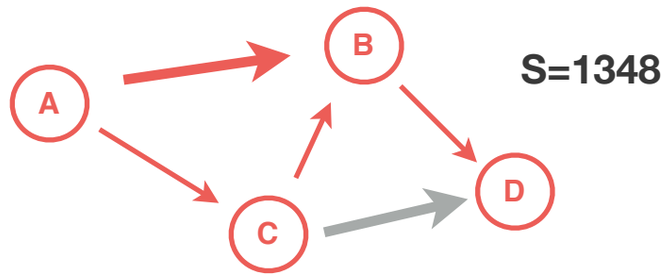


S=1177

this move is not possible as it would create a cycle B → C → D



S=1312



S=1348

Genomics application

DNA methylation

Gene expression

H3K27ac

H3K4me1

H3K4me3

H3K36me3

H3K9me3

H3K27me3

MYCN

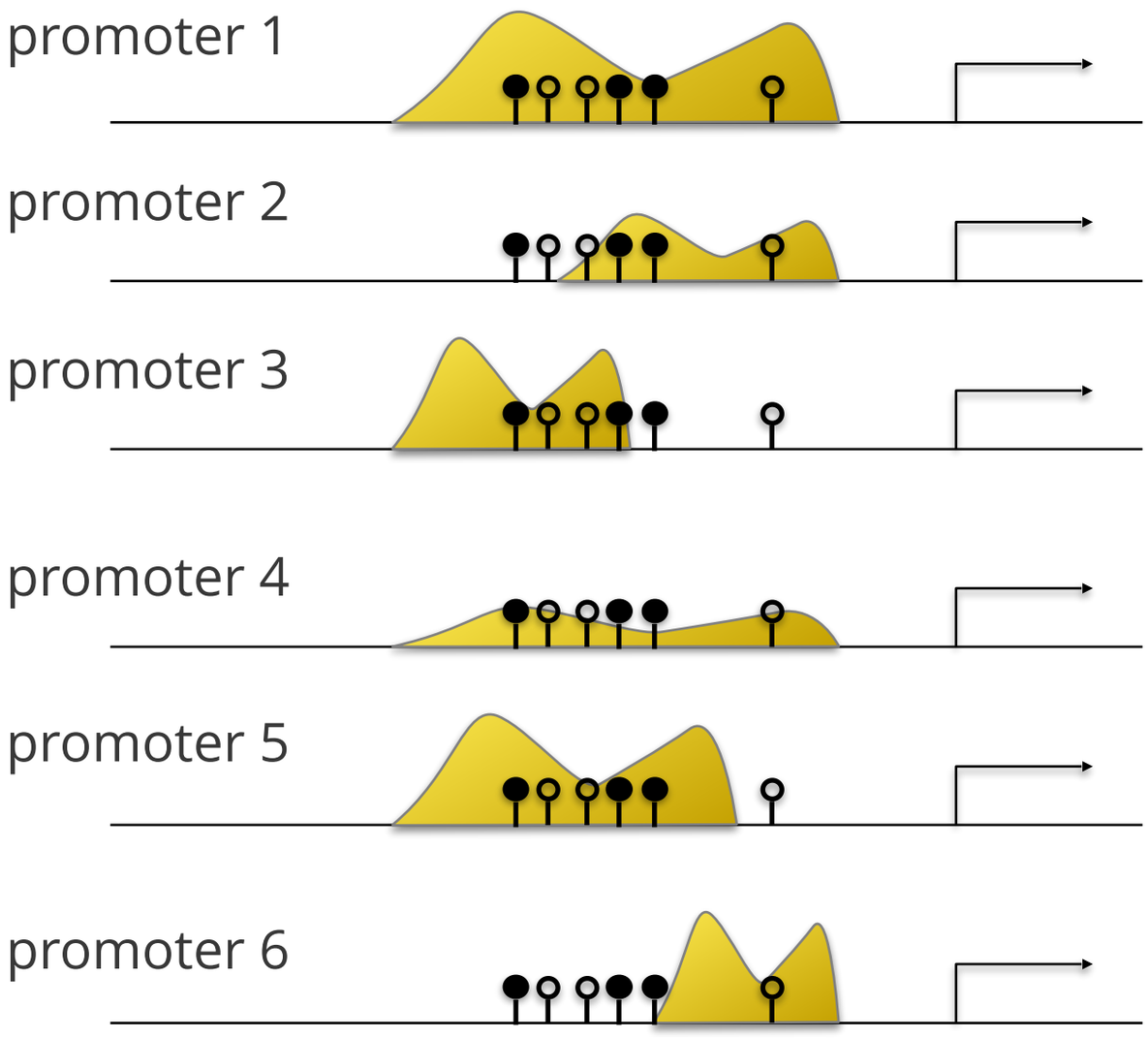
EZH2

DNMT1

DNMT3

- Various neuroblastoma cell lines
- normal conditions / treated (inhibition)
- state at gene promoters represent the observations of the random variables

Learning Network Structures



DNAme	K27ac
0.57	128.8
0.45	75.2
0.89	98.3
0.21	21.3
0.18	86.2
0.41	67.3

DNAme	K27ac
mid	5
mid	3
high	4
low	2
low	4
mid	3

3 states

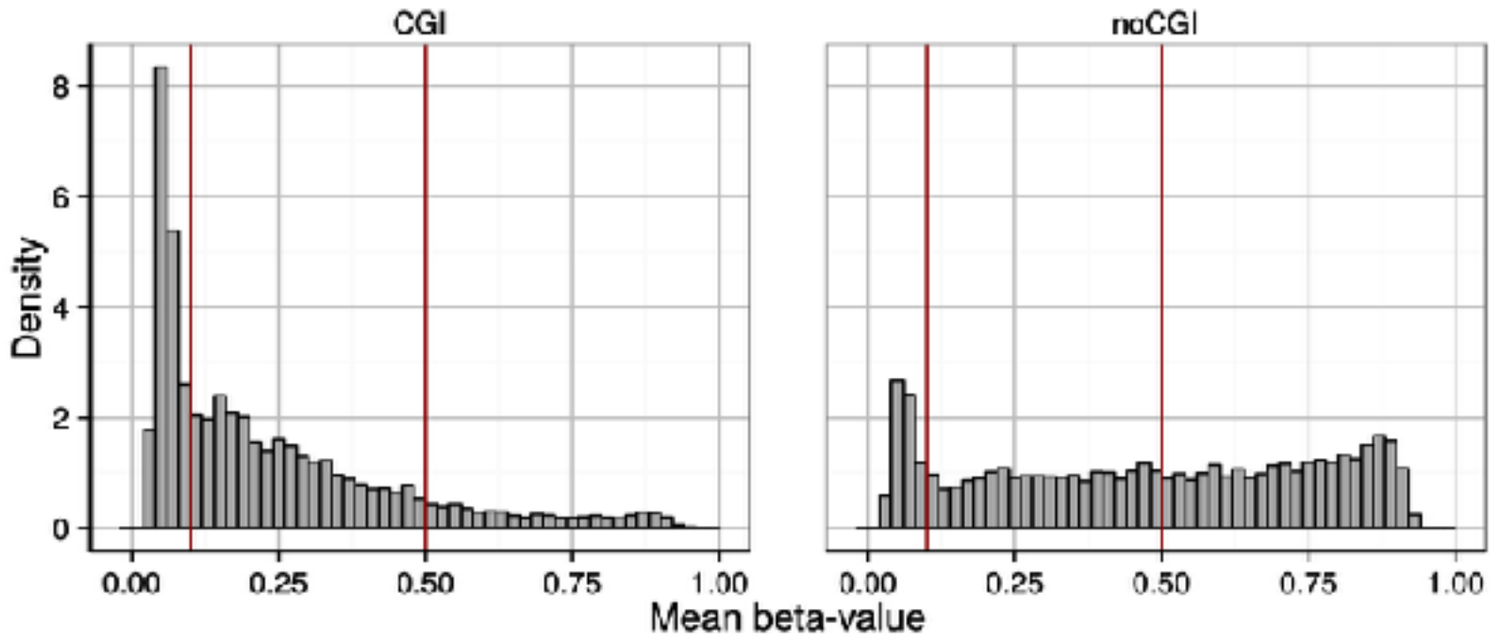
5 states

Discretisation strategies

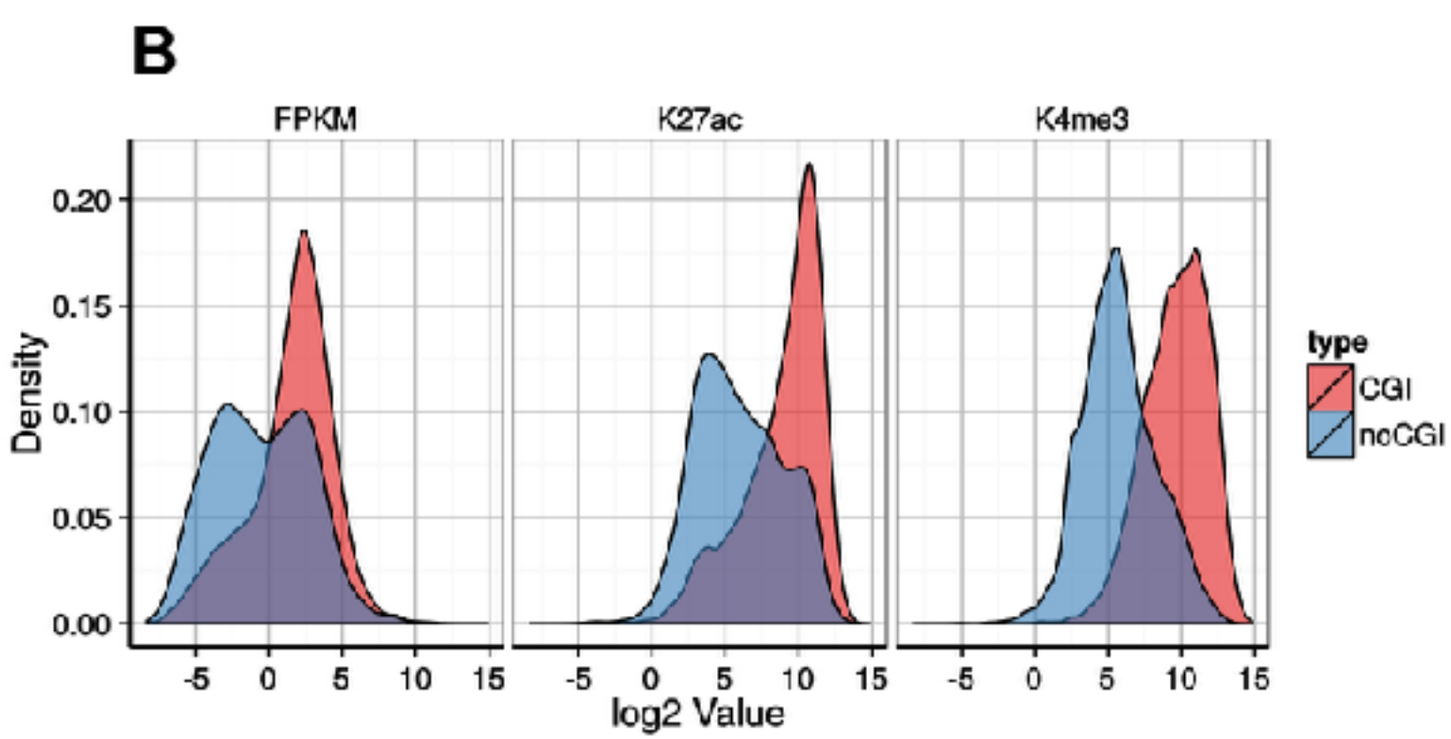
- Continuous data can be discretised to circumvent the requirement for specific distributions (normal distribution)
- several discretisation strategies
  - *naive discretisation*
    → define bins according to external evidence (low / mid / high)
  - *quantile-based discretisation*
    → equally balanced levels
  - *k-means based discretisation*
    → automatic definition of number of levels [Ckmeans.1d.dp, Wang et al. 2011]
  - *mutual information preserving discretisation*
    → quantile-discretisation, then merging of levels such as to maintain the mutual information structure [Hartemink, 2005]

Different promoters have different distributions

- CpG-island overlapping
- non CpG-island overlapping



This might indicate that the observations correspond to different random variables



Learning Network Structures

- Akaike Information Criterion (*AIC*) / Bayesian information criterion (BIC)

$$BIC = \sum_{i=1}^{n} \log P(X_i | pa(X_i)) - \frac{d}{2} \log(n)$$

Likelihood of data, given learned parameters

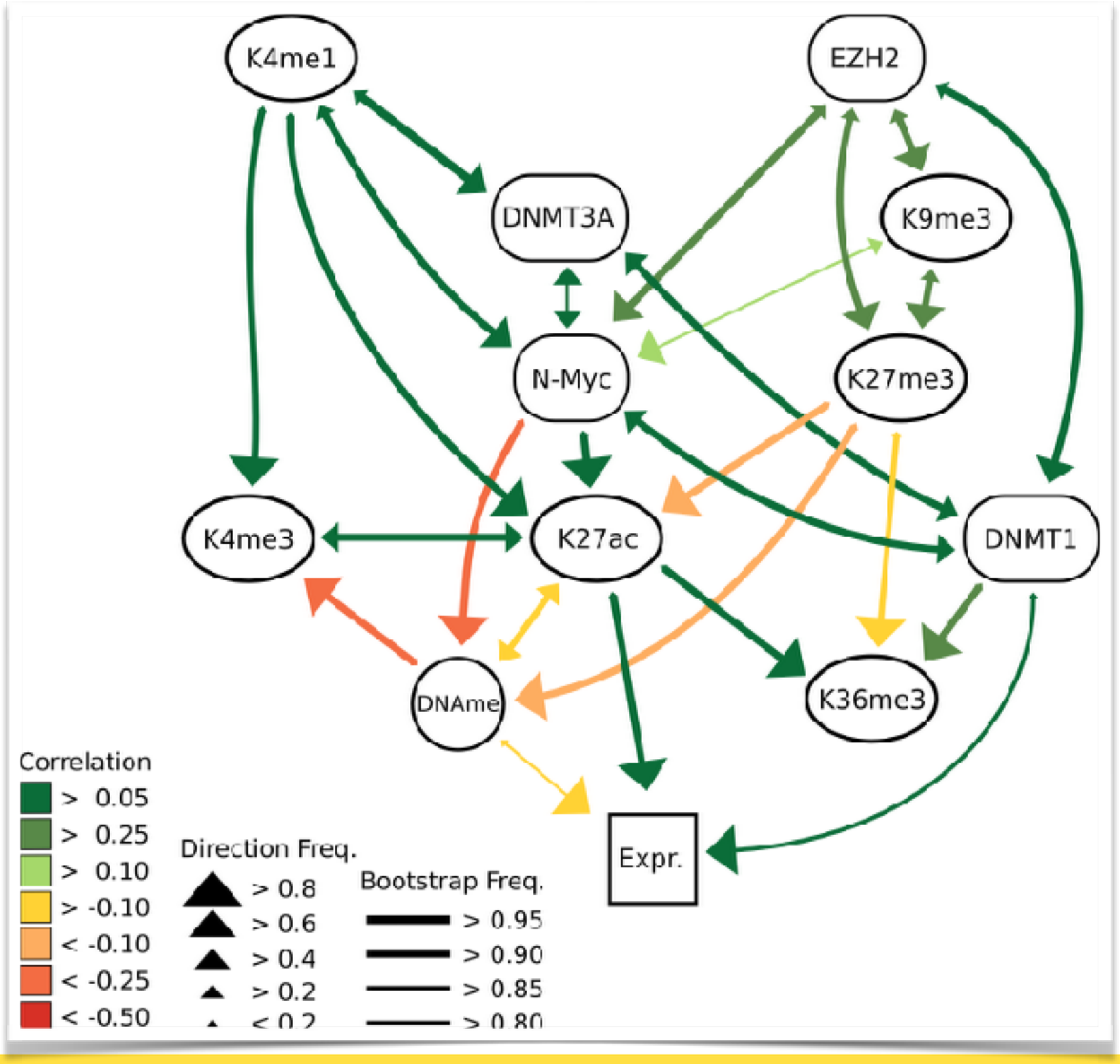
d = number of parameters, complexity penalty

- **TABU search**
  - Optimize score by adding / removing / redirecting arcs
  - Prohibit the last n changes ("memory effect")
  - Try additional m steps when hitting a maximum (avoid local optima)

- **Bootstrapping** :
  - randomly sample a subset of observations *N* times
  - average the network
  - determine strength of edge / direction as the proportion of observations

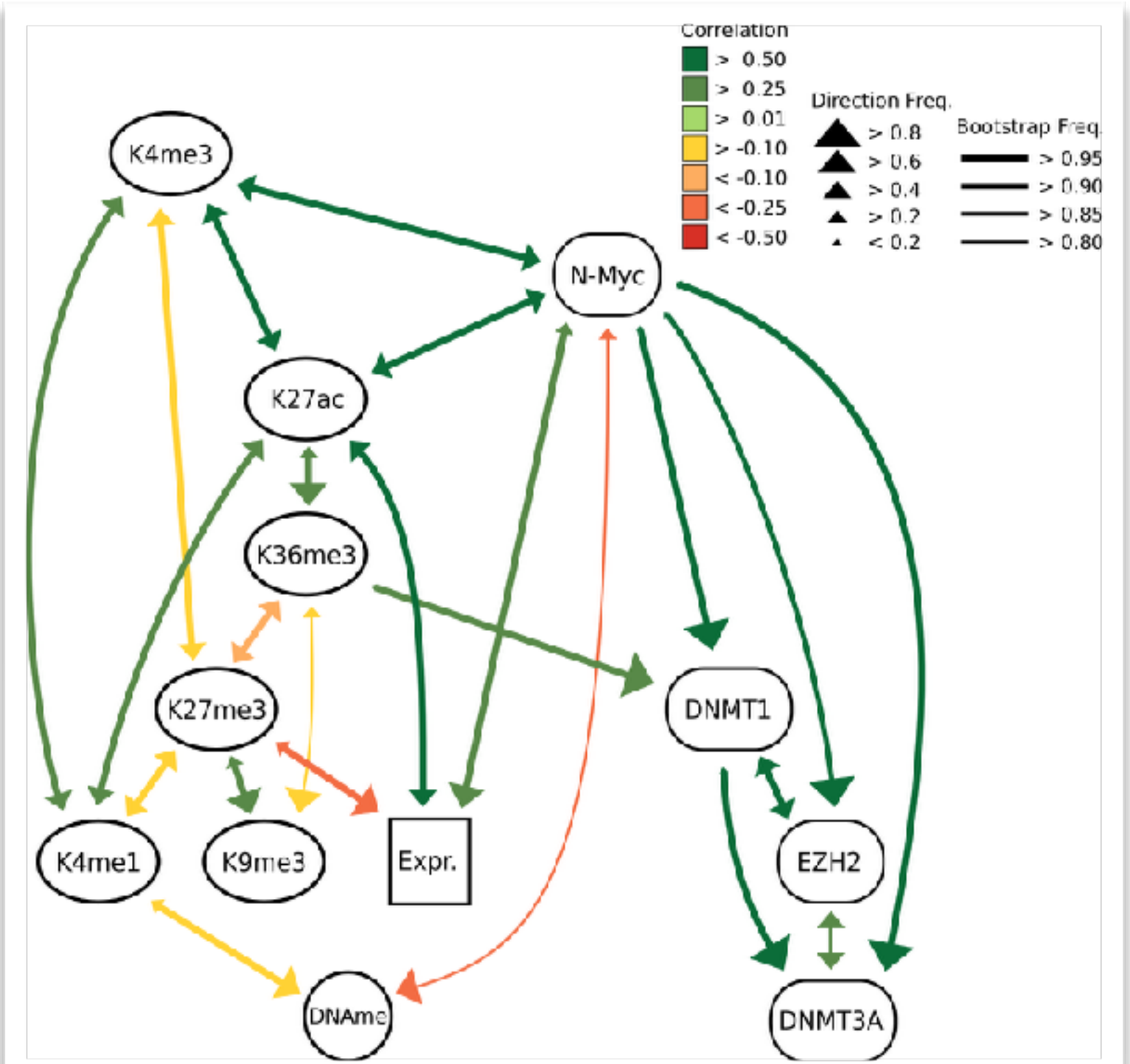
Promoter BN

- non-CGI Promoters
- (*n* = 5139)

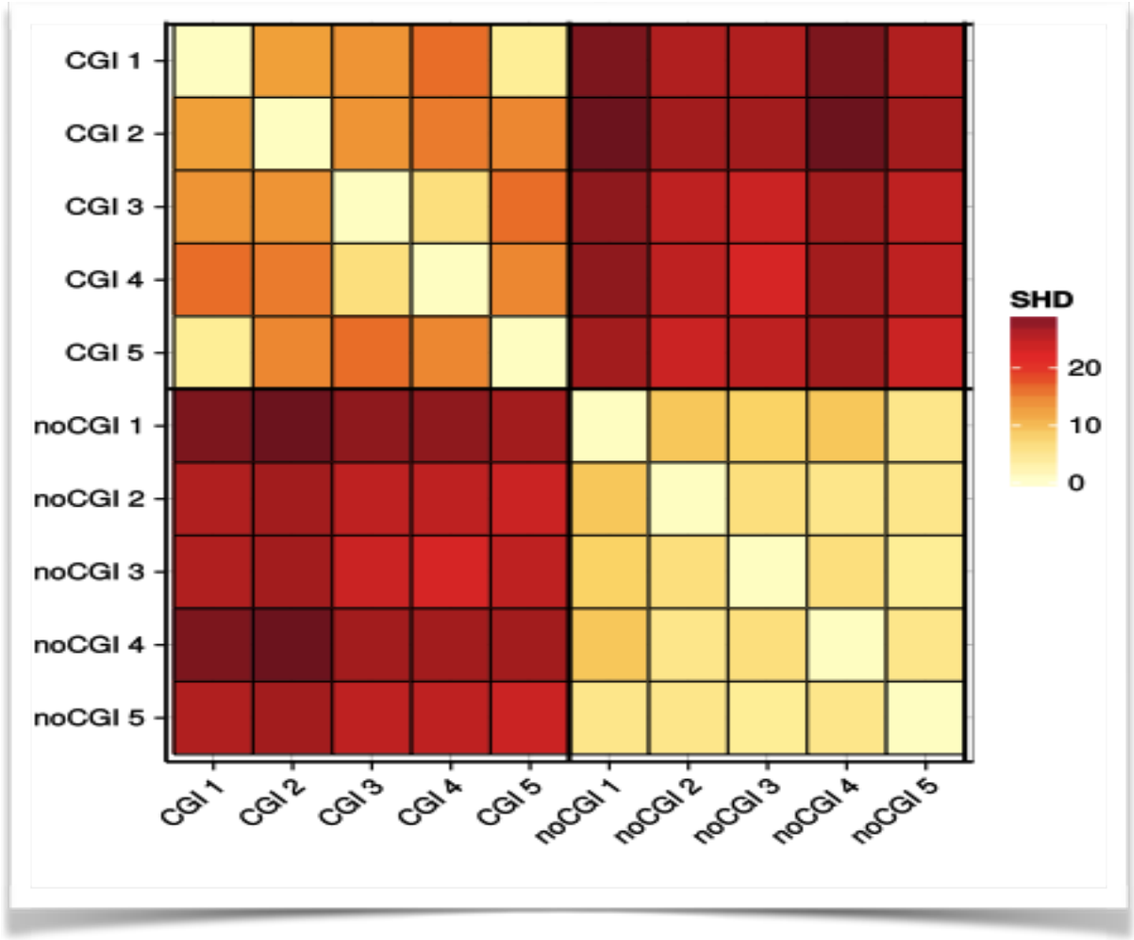


Promoter BN

- CGI Promoters
- (*n* = 8906)



Robustness of BN ?

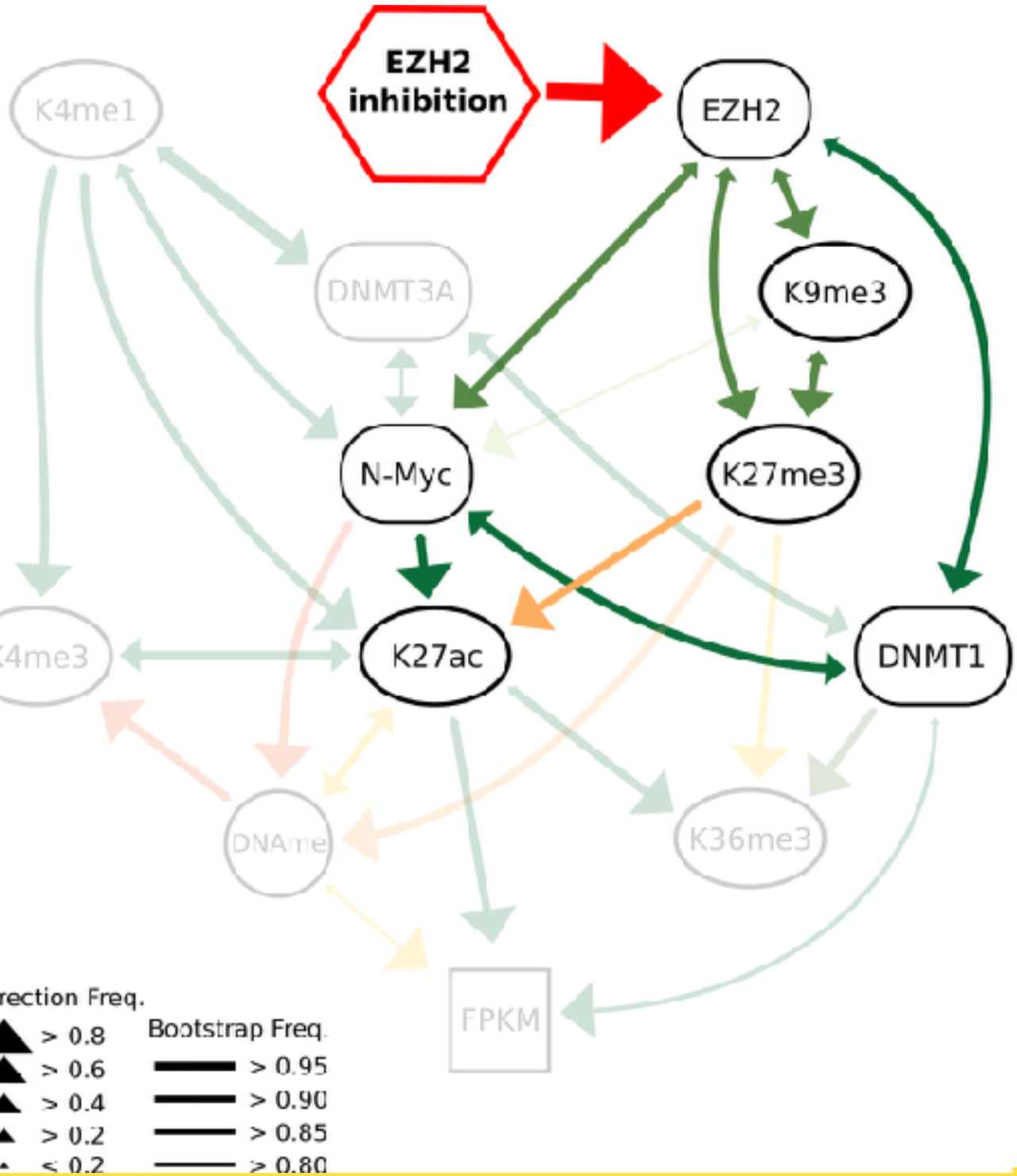


- Compare several instances of the CGI/non-CGI networks
- Structural Hamming distance = number of operations needed to turn one network into another
- Weight operations using the weight of each arc

Bootstrapped CGI / non-CGI promoter BN are robust

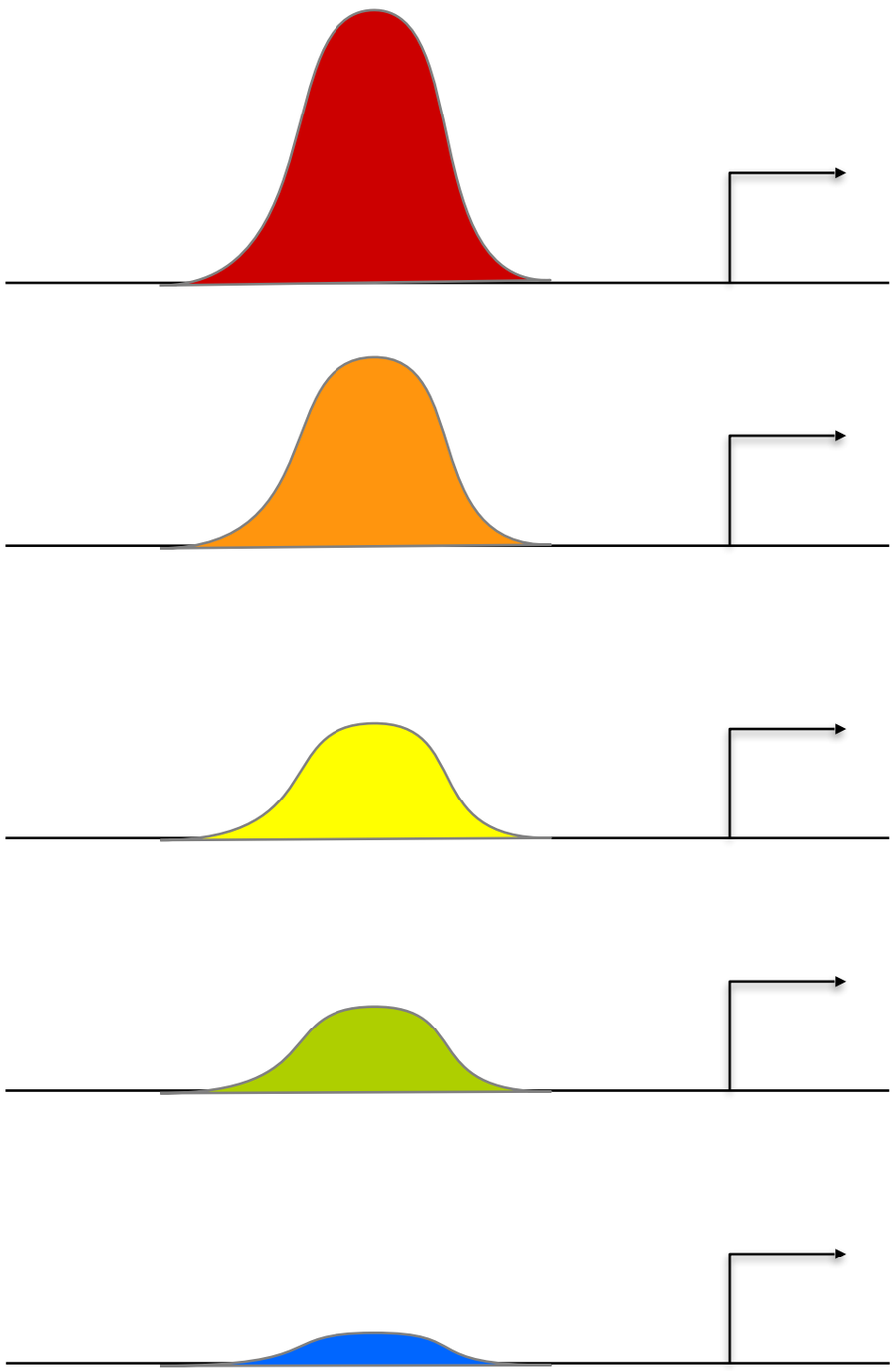
Predicting Interventions

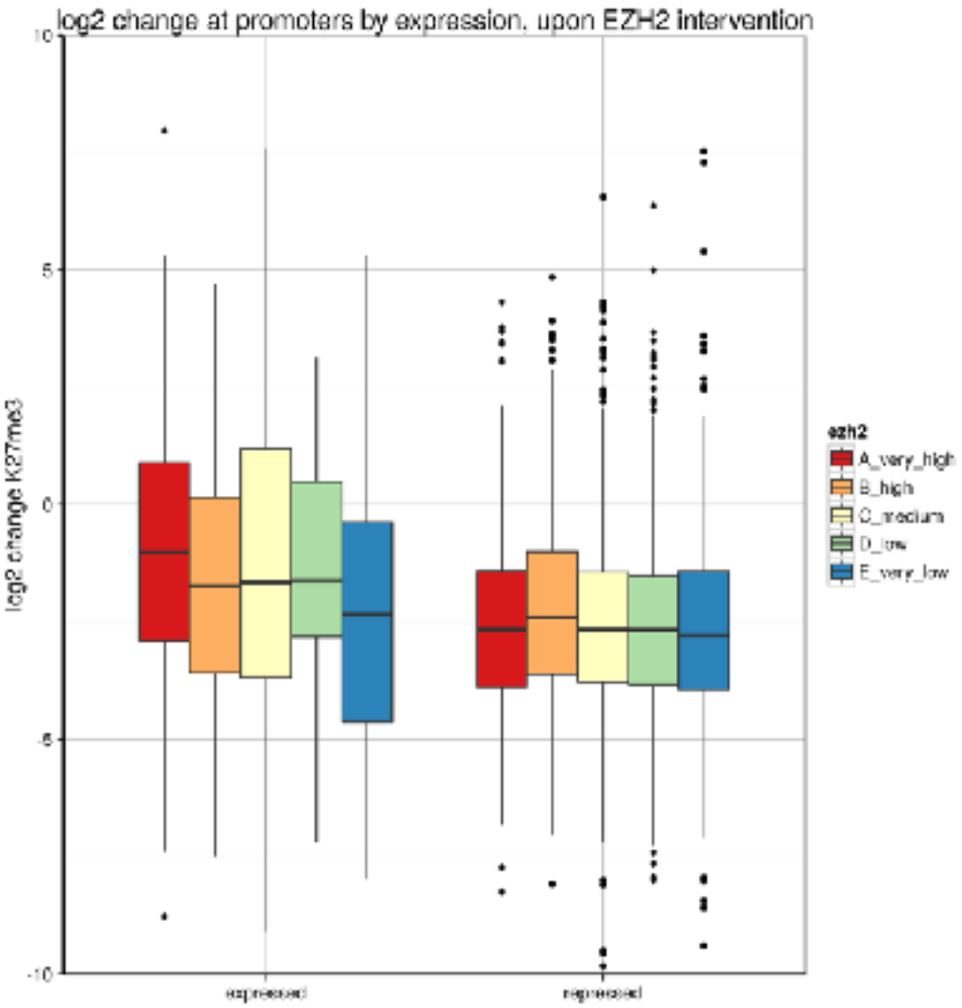
- Small molecule EZH2 inhibitor
- Histone mark ChIP-seq,
- RNA-seq (control vs. treated)
- Same NB cell line Be(2)-C



Changes H3K27me3

EZH2 signal at the promoter



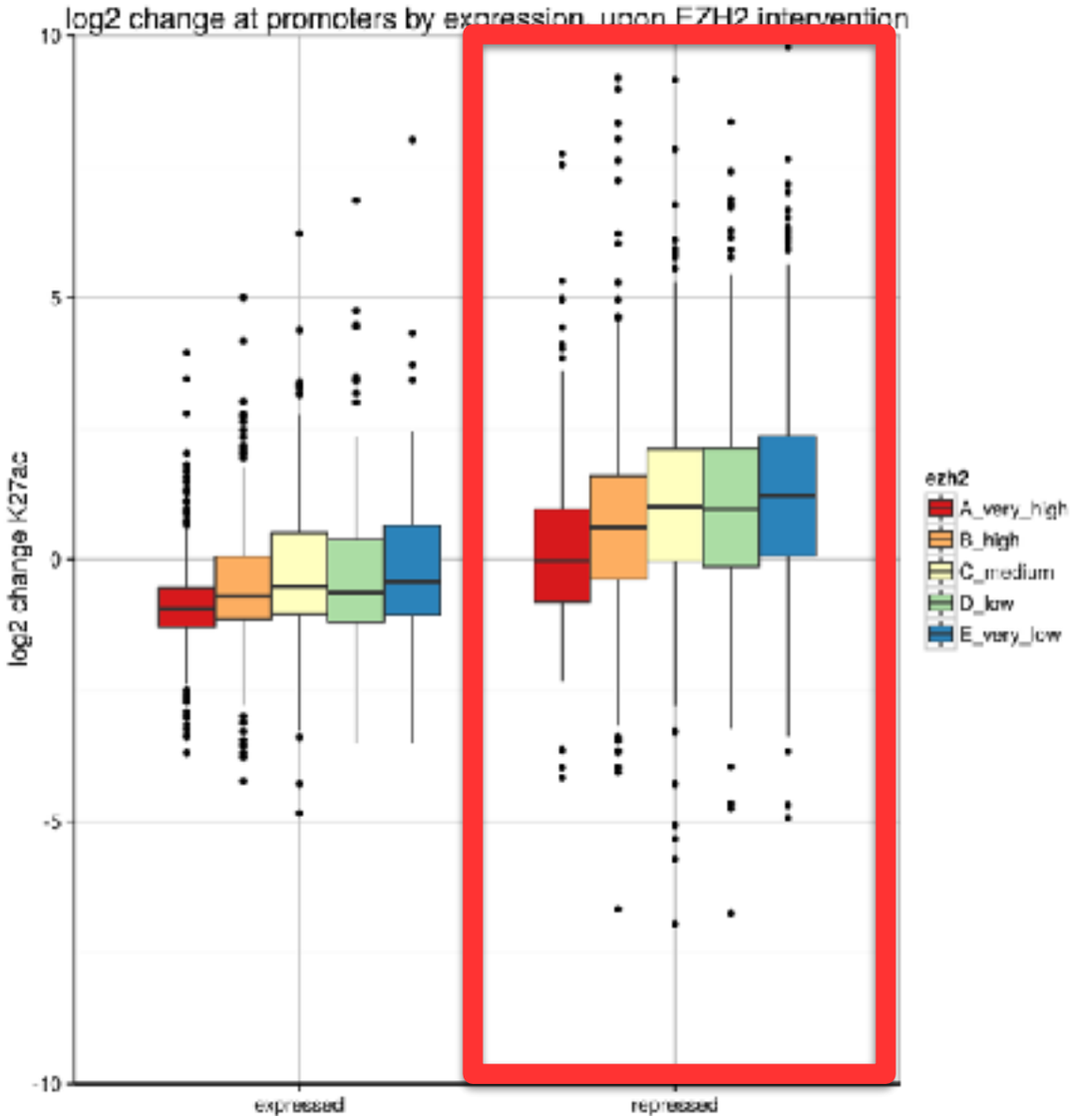


$$\log_2\left(\frac{inhihition}{control}\right)$$

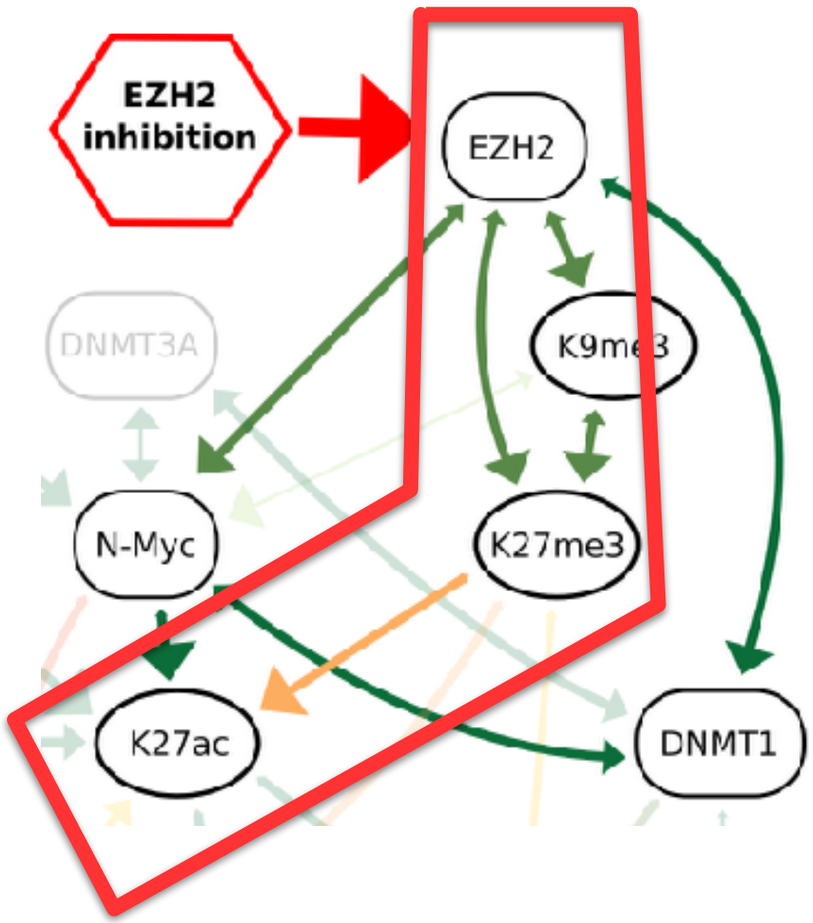
As expected, EZH2 inhibition reduces the overall H3K27me3 signal at the gene promoters

Changes H3K27ac

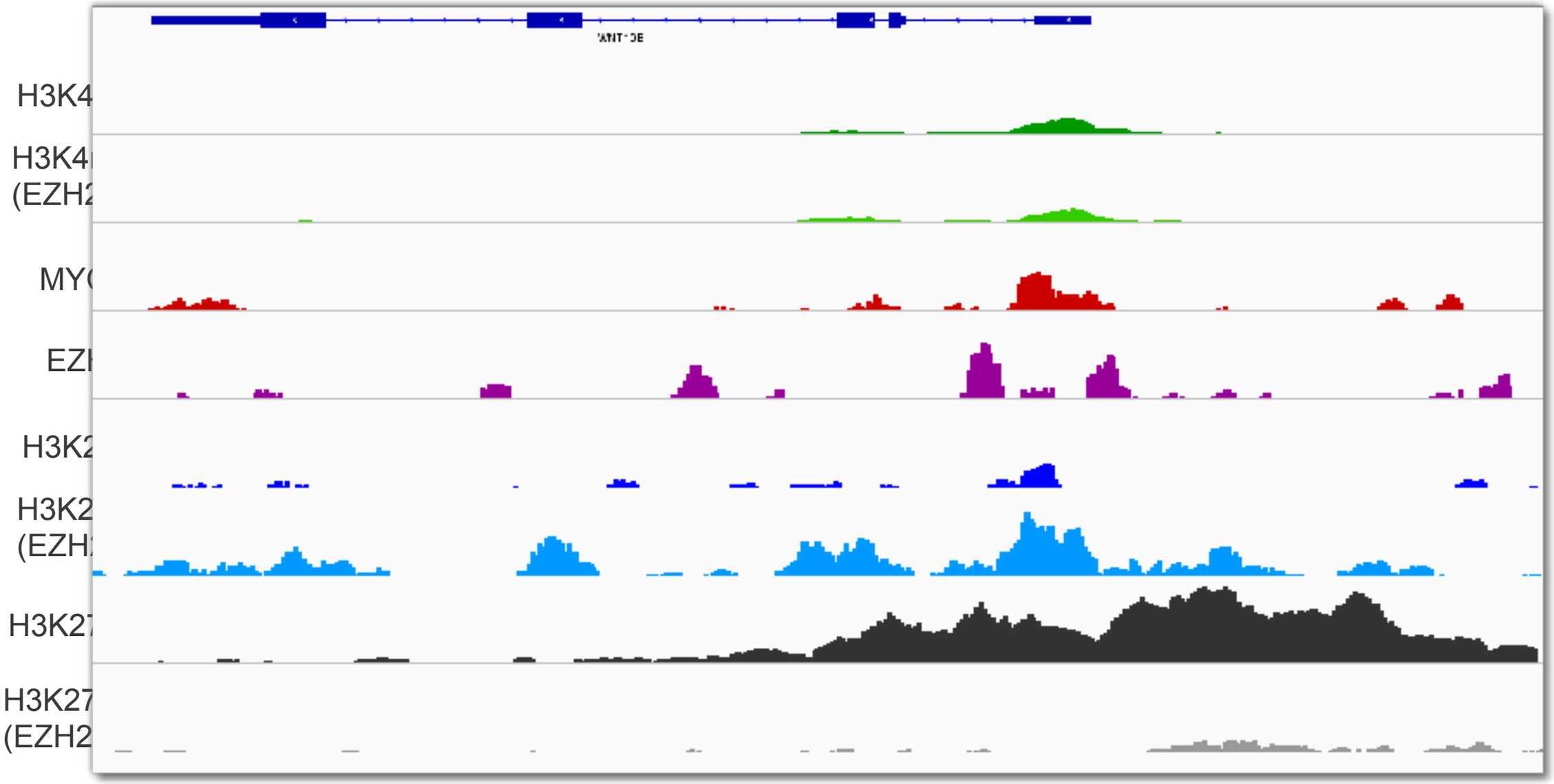
- **Increase of K27ac** at the promoter of repressed genes



$$\log_2\left(\frac{inhihition}{control}\right)$$

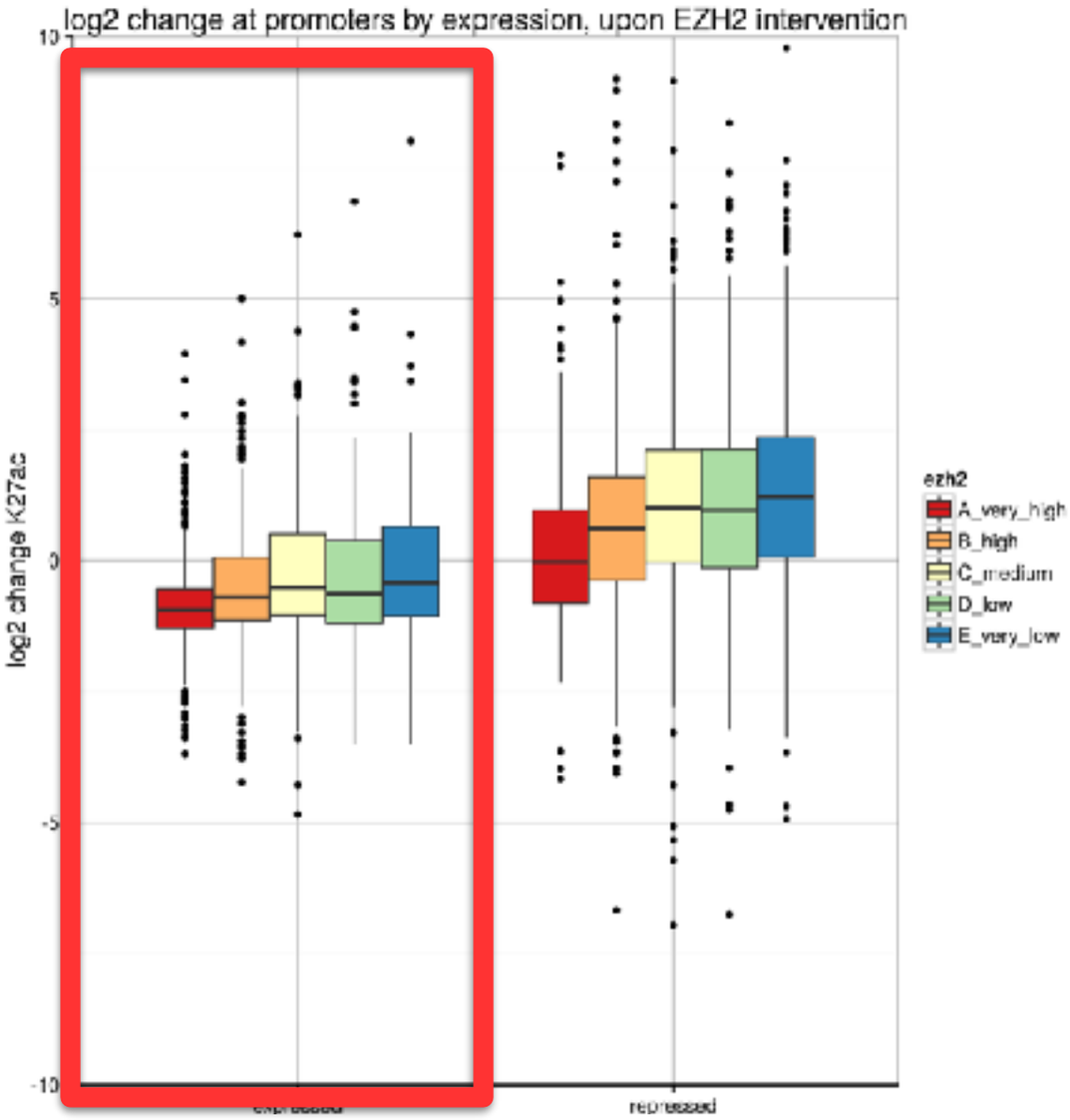


Changes in H3K27ac upon EZH2 inhibition

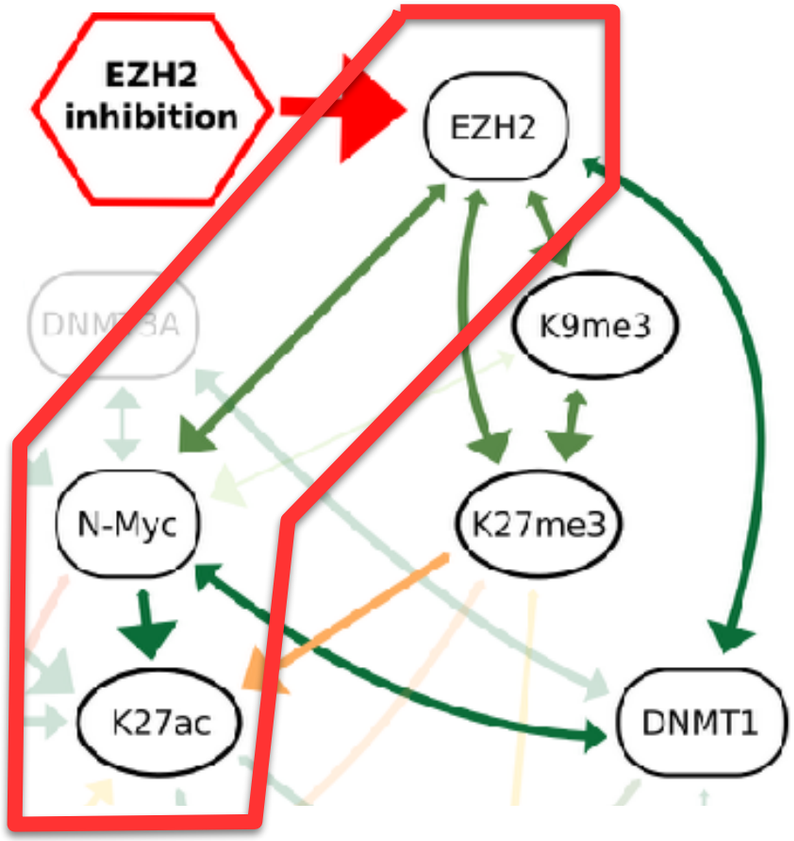


Changes H3K27ac

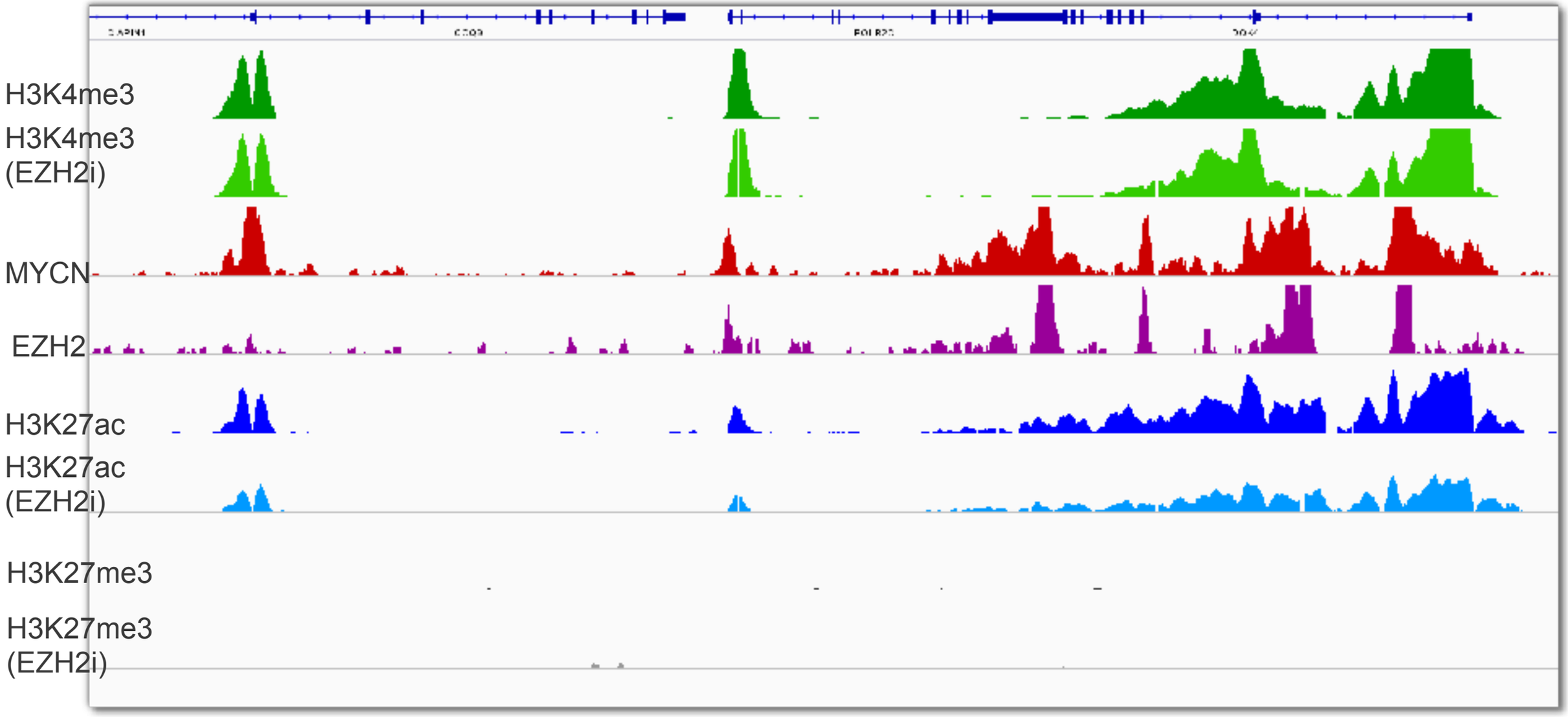
- EZH2 inhibition leads to a **reduction of K27ac** at the promoter of expressed genes



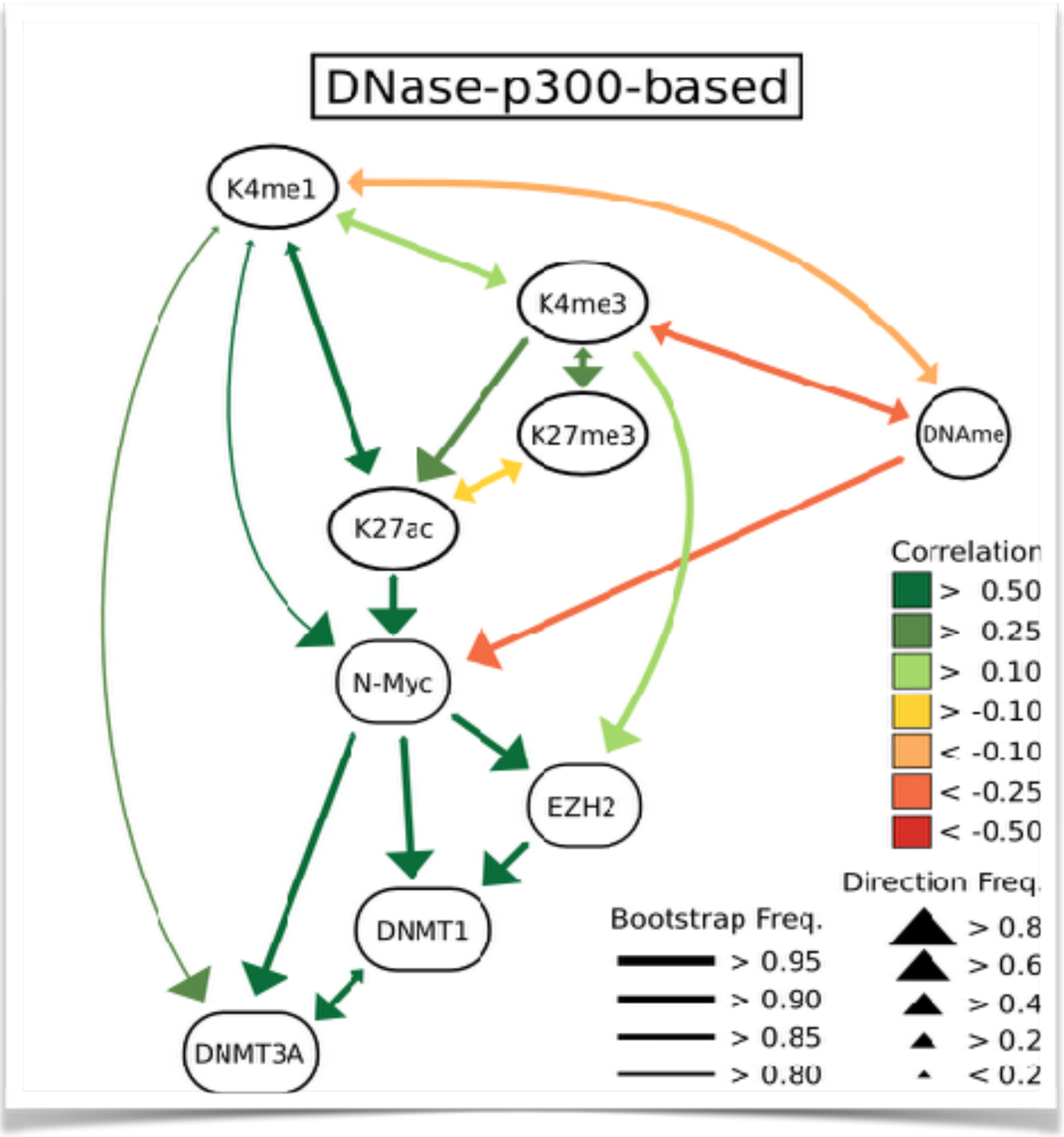
$$\log_2\left(\frac{inhihition}{control}\right)$$



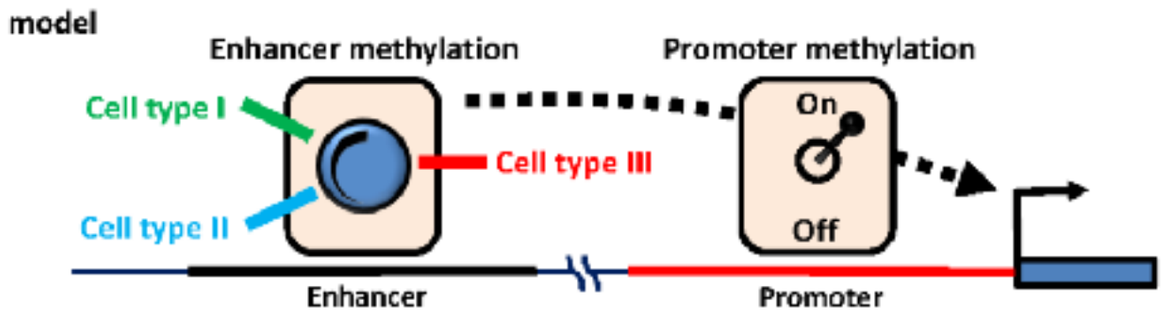
Changes in H3K27ac upon EZH2 inhibition



Enhancer networks



- Enhancers defined using DNAse/p300 in matching tissues / cell lines
- DNA methylation appears to play a more "active" role, compared to promoter networks



[Aran et al, Genome Biol. 2013]

Acknowledgements

- **Cancer Regulatory Genomics**
  - **Ashwin Sharma (BN)**
  - **Ron Schwessinger (BN)**
  - Calvin Chan
  - Qi Wang
  - Andres Quinterò
- **Neuroblastoma Genetics**
  - Frank Westermann
  - Daniel Dreidax (BN, data)
  - Moritz Gartlgruber (BN, data)