Centrum Zastosowań Matematyki

Warsztaty

Modelowanie matematyczne i współpraca interdyscyplinarna

26–28 września 2013 r.

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

„Człowiek - najlepsza inwestycja"

Projekt współfinansowany ze środków Unii Europejskiej
w ramach Europejskiego Funduszu Społecznego

Sieci Bayesowskie jako narzędzie bioinformatyka: Część I -wprowadzenie

Bartek Wilczyński

Institute of Informatics, University of Warsaw

Centrum Zastosowań Matematyki, Gdańsk, 2013

Plan of the lecture

- Bayesian Network (BN) models: examples, properties, limitations, applications...
- BN models and the data: conditional probability distributions, distribution factorization, fitting to data, making predictions
- BN topology: limitations, scoring functions
- Finding optimal BNs: problem statement, complexity issues and their causes,
- Effective solutions in special cases: Dynamic BNs and BN classifiers

Models in molecular biology

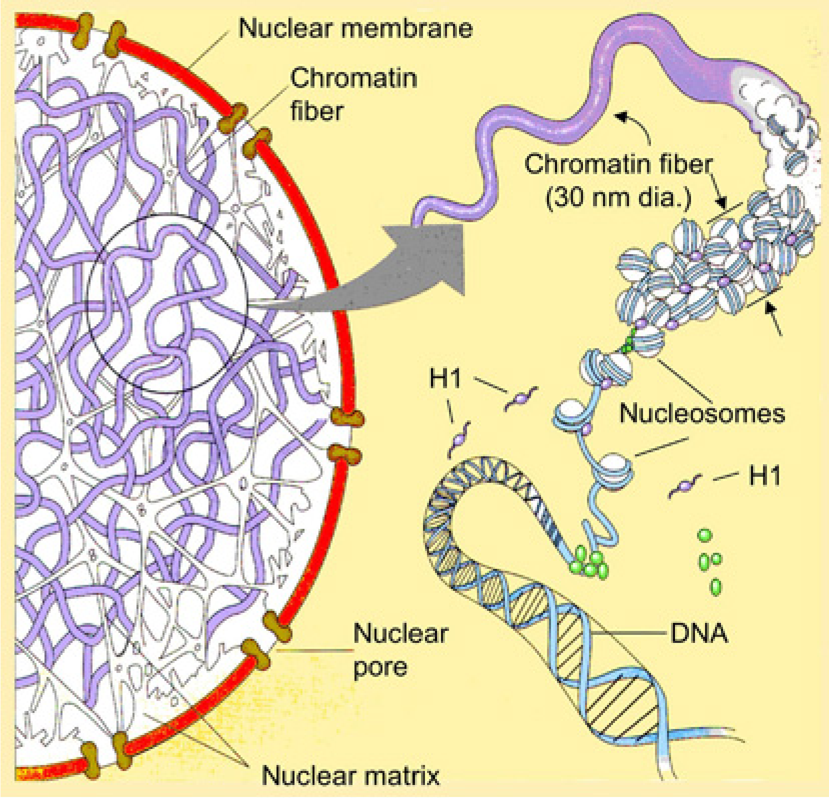
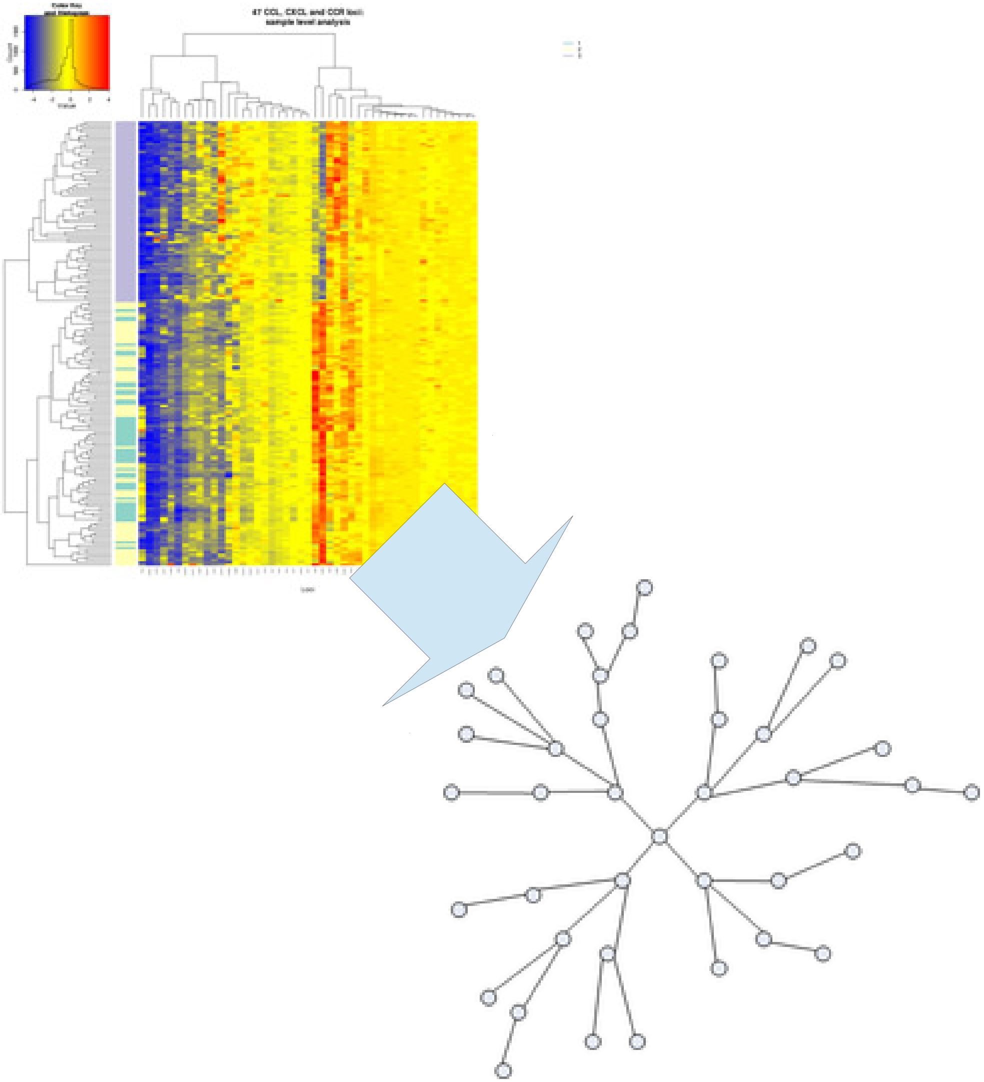


Fig. 1. Modifications of the histone components of nucleosomes help regulate DNA accessibility by promoting folding or unfolding of chromatin fibers, and by recruiting chromatin remodeling complexes and other factors to specific genomic loci.

- We are facing complex systems, with many parts unknown or unobservable
- Measurements, especially of the high-throughput kind, are noisy and indirect with respect to the actual processes
- We make thousands of observations and it is vital for us to have a more concise representation of the data that we are able to interpret
- A model needs not only to be simpler than the data, but it also needs to be more general than a single dataset
- The choice of your model will always depend on the process you are interested in

Finding relationships in multi-variate data

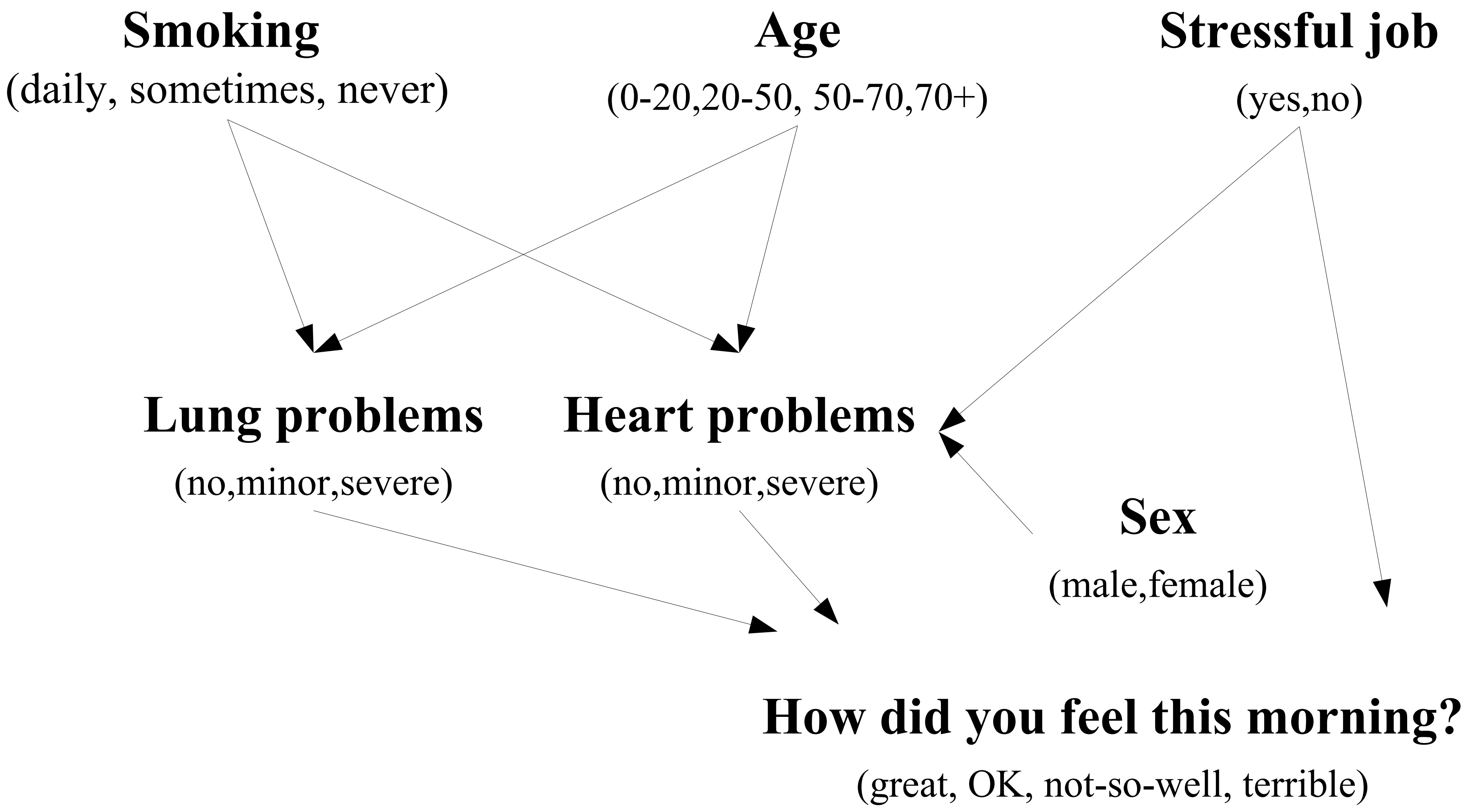


One of the most common scenarios in "omics" projects is the one in which we look for relationships between variables in a multivariate dataset. Many statistical (regression based or qualitative) methods compete in this broad field

Advantages of probabilistic models

- Natural handling of uncertainty, both at the data and model levels
- Straightforward treatment of missing values
- Possibility to incorporate unobservable (hidden) variables
- Natural (maximum likelihood) scoring of different models for the same dataset
- Prior distributions can be used to put human knowledge into the model

Example of a simple BN



Smoking
(daily, sometimes, never)

Age
(0-20,20-50, 50-70,70+)

Stressful job
(yes,no)

Lung problems
(no,minor,severe)

Heart problems
(no,minor,severe)

Sex
(male,female)

How did you feel this morning?
(great, OK, not-so-well, terrible)

Informal BN definition

- A directed acyclic graph
- with nodes representing random variables X={x_i}
- and edges between nodes representing dependencies (very roughly, we will discuss it later)
- Each edge is directed from a parent to a child, set of parents of x is denoted by Par(x)
- Each variable is associated with a value domain and a probability distribution P(x|Par(x))

How do observations look like?

Sex	Age	Smoking	Stress	Lung	Heart	Feel
M	0-20	never	N	No	no	great
F	70	sometimes	N	minor	no	OK
M	50-70	daily	Y	no	severe	Not-so-well
M	20-50	daily	N	no	minor	OK
F	70	never	N	no	minor	great
F	20-50	sometimes	Y	severe	minor	Not-so-well
F	20-50	never	Y	no	no	great
M	20-50	sometimes	N	minor	no	great
M	50-70	never	Y	severe	no	OK
F	0-20	never	N	no	severe	OK
M	20-50	daily	Y	no	no	OK
M	0-20	daily	N	no	no	Not-so-well
M	20-50	never	N	minor	no	OK
....	...	...	...	...	...	...

Conditional probability distributions

- For each variable we define the probabilities distribution in the respective domain conditional on the values of its parents
- It is assumed, that for each variable:

$$P(x|X) = P(x|Par(x))$$

Smoking	*Age*	Lung=No	Lung=minor	Lung=severe
never	*0-20*	0.999	0.0009	0.0001
sometimes	*0-20*	0.995	0.004	0.001
daily	*0-20*	0.99	0.005	0.005
never	*20-50*	0.99	0.005	0.005
sometimes	*20-50*	0.97	0.02	0.01
daily	*20-50*	0.9	0.07	0.03
....	…	…	…	…

Fitting CPDs to data

- Naturally, given enough observations, we can calculate the CPDs from contingency tables
- These contingency tables can be used to test the factorization

Sex\Age	0-20	20-50	50-70	70+
Male	55	43	42	25
Female	51	41	50	35

P(Age|Sex)

Sex\Age	0-20	20-50	50-70	70+
Male	0.33	0.26	0.25	0.15
Female	0.29	0.23	0.28	0.20

Now, can we find the best graph?

- Given a dataset with observations, we can try to find the “best” network topology (i.e. the best collection of parents' sets)
- In order to solve the problem we first need to formalize our objective function to score different graphs
- A score function usually can be written as a sum over variables:

$$Score = \sum_{i=1}^{N} S_{data}(x_i, Par(x_i), Data) + S_{graph}(x_i, Par(x_i))$$

Different approaches to BN score

- There are generally 3 main approaches to defining BN scores:
  - Bayesian statistics, e.g. BDe (Herskovits *et al.* '95)
  - Information Theoretic, e.g. MDL (Lam *et al.* '94)
  - Hypothesis testing, e.g. MMPC (Salehi *et al.* '10)
- There are also hybrid approaches, like the recent MIT (de Campos '06) approach that uses information theory and hypothesis testing

Minimum Description Length

- Graph score corresponds to the cost of encoding the CPDs:

$$g(\mathbf{Pa}) = |\mathbf{Pa}| \log n + \frac{\log N}{2}(k_X - 1) \prod_{Y \in \mathbf{Pa}} k_Y$$

- Data score corresponds to the cost of optimal encoding of the data, given the CPDs:

$$d(\mathbf{Pa}) = N \cdot H(X|\mathbf{Pa})$$

Bayesian Dirichlet equivalence

- We start from Bayes theorem:

$$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{G})P(\mathcal{D}|\mathcal{G}) = P(\mathcal{G}) \int P(\mathcal{D}|\mathcal{G}, \theta)P(\theta|\mathcal{G})d\theta$$

- To write the graph score:

$$g(|\mathbf{Pa}|) = |\mathbf{Pa}| \log \alpha^{-1}$$

- And the data score:

$$d(\mathbf{Pa}) = \log \left(\prod_{\mathbf{v} \in \mathcal{V}^{|\mathbf{Pa}|}} \frac{\Gamma(\sum_{v \in \mathcal{V}}(H_{v,\mathbf{v}} + N_{v,\mathbf{v}}))}{\Gamma(\sum_{v \in \mathcal{V}} H_{v,\mathbf{v}})} \prod_{v \in \mathcal{V}} \frac{\Gamma(H_{v,\mathbf{v}})}{\Gamma(H_{v,\mathbf{v}} + N_{v,\mathbf{v}})} \right)$$

How to find the “best” network?

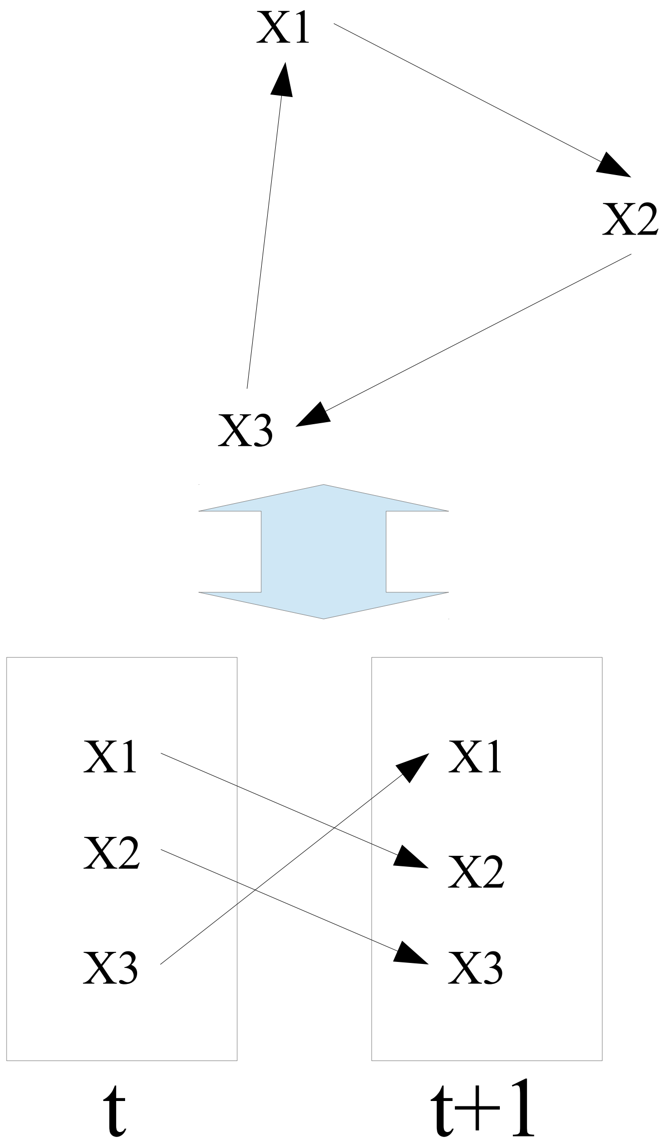
- We have two problems:
  1) There are exponentially many potential parent sets
  2) The desired network needs to be a DAG
- It was shown by Chickering ('96) that finding an optimal BN is NP-complete
- He has shown that the problem is NP-complete even if we limit size of Parent sets to 2
- So the acyclicity criterion (2) is enough to make it practically unfeasible to find best BNs

Is it always a problem?

- If we had a situation, where the acyclicity of the network is guaranteed by an external constraint, we would only need to worry about finding the optimal parent set
- There are several practical cases when it is true. To name two:
  - Dynamic BNs
  - Classification using BNs

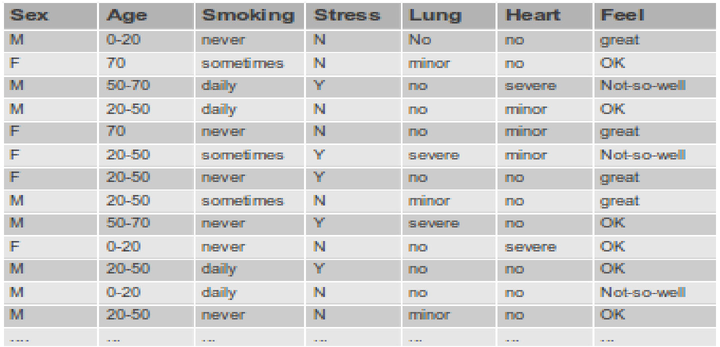
Dynamic Bayesian Networks

- Dynamic Bayesian Networks are an extension of the BN models to include temporal dependencies
- It is frequently used in a simple form, where only the dependencies between time points are allowed

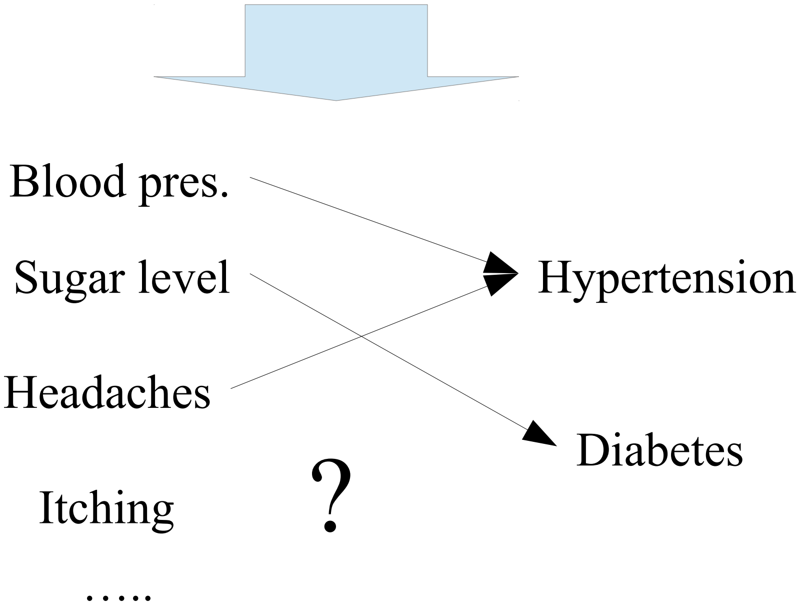


BNs for classification

- The problem of classification, we have a number of variables, with a specified subset of ***class variables***
- We are interested in models able to predict them from the other measured variables



Sex	Age	Smoking	Stress	Lung	Heart	Feel
M	0-20	never	N	No	no	great
F	70	sometimes	N	minor	no	OK
M	50-70	daily	Y	no	severe	Not-so-well
M	20-50	daily	N	no	minor	OK
F	70	never	N	no	minor	great
F	20-50	sometimes	Y	severe	minor	Not-so-well
F	20-50	never	Y	no	no	great
M	20-50	sometimes	N	minor	no	great
M	50-70	never	Y	severe	no	OK
F	0-20	never	N	no	severe	OK
M	20-50	daily	Y	no	no	OK
M	0-20	daily	N	no	no	Not-so-well
M	20-50	never	N	minor	no	OK
....	...	...	...	...	...	...



Blood pres.

Sugar level

Headaches

Itching

…..

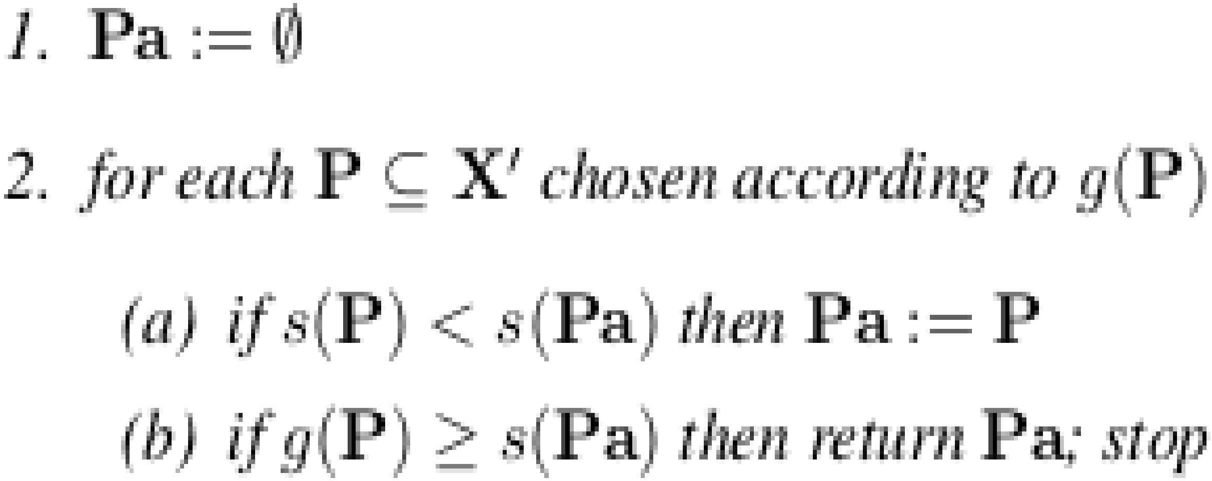
?

Hypertension

Diabetes

Can we then find the solution?

Assuming that we have external constraints guaranteeing acyclicity of the network we can use the following greedy algorithm, for each variable independently:



```
1. $\mathbf{Pa} := \emptyset$
2. for each $\mathbf{P} \subseteq \mathbf{X}'$ chosen according to $g(\mathbf{P})$
    (a) if $s(\mathbf{P}) < s(\mathbf{Pa})$ then $\mathbf{Pa} := \mathbf{P}$
    (b) if $g(\mathbf{P}) \geq s(\mathbf{Pa})$ then return $\mathbf{Pa}$; stop
```

Dojer '06, Wilczynski&Dojer '09

But there are so many potential sets!

We are looking for an optimal sum of graph and data scores:

$$Score = \sum_{i=1}^{N} S_{data}(x_i, Par(x_i), Data) + S_{graph}(x_i, Par(x_i))$$

The graph score penalizes large parent sets. There is a limited number of parent sets with a score smaller than the optimal score. This leads to the bound on the complexity of the algorithm, in case of MDL it is:

$$\mathcal{O}(n^{\log_k N} N \log_k N)$$

Dojer '06, Wilczynski&Dojer '09

Software tools

- Banjo – an MCMC method from A. Hartemink group `www.cs.duke.edu/~amink/software/banjo/`
- BNfinder – Wilczynski&Dojer '09
- Bnlearn – an R package implementing the MMPC algorithm (Salehi et al '10)
- GlobalMIT – an implemmentation of Bnfinder greedy algorithm for the MIT score (Vinh *et al. '11)*

Summary

- BNs are a probabilistic modeling tool for multivariate data
- In general, it is NP-complete to find the optimal network (General solvers use heuristic approaches like MCMC)
- The problem lies in satisfying acyclicity of BN
- If you can avoid this problem, you can find optimal networks quickly

A few words of caution

- One needs substantial number of observations to calculate meaningful network
- The most obvious choice of the variable set is not always the best
- The fact that a network is “optimal” does not make it the “true” network
- Different scoring functions have their peculiarities, but should give you very similar “optimal” networks

Acknowledgments

- **Norbert Dojer**

- Pawel Bednarz
- Agnieszka Podsiadlo
- Joanna Giemza

- Post-doc positions available!

EMBO

INNOVATIVE
ECONOMY
NATIONAL COHESION STRATEGY

FNP
Fundacja na rzecz
Nauki Polskiej

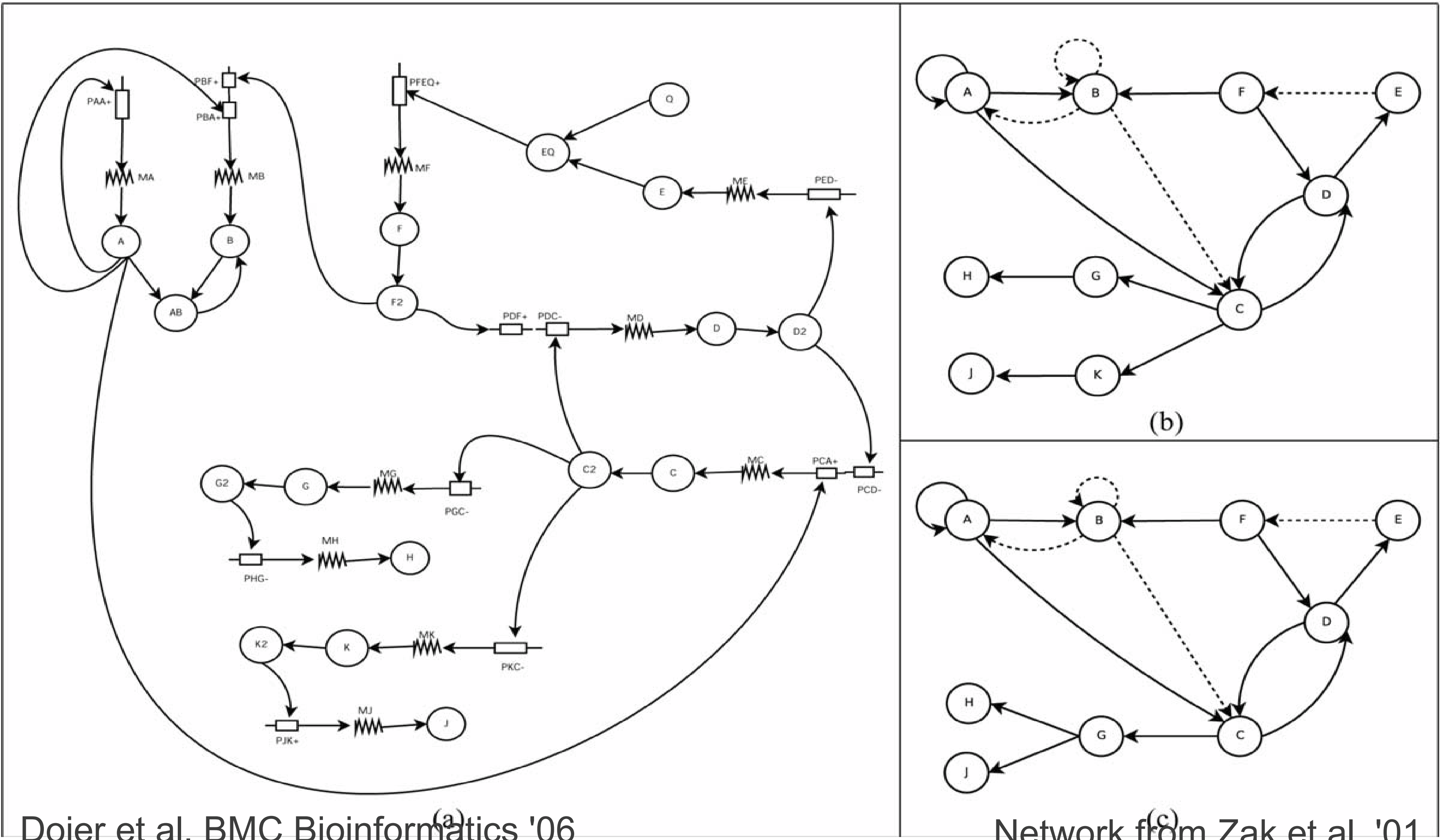
Sieci Bayesowskie jako narzędzie bioinformatyka: Część II - Zastosowania

Bartek Wilczyński

Institute of Informatics, University of Warsaw

Centrum Zastosowań Matematyki, Gdańsk,   2013

Regulatory Network Reconstruction

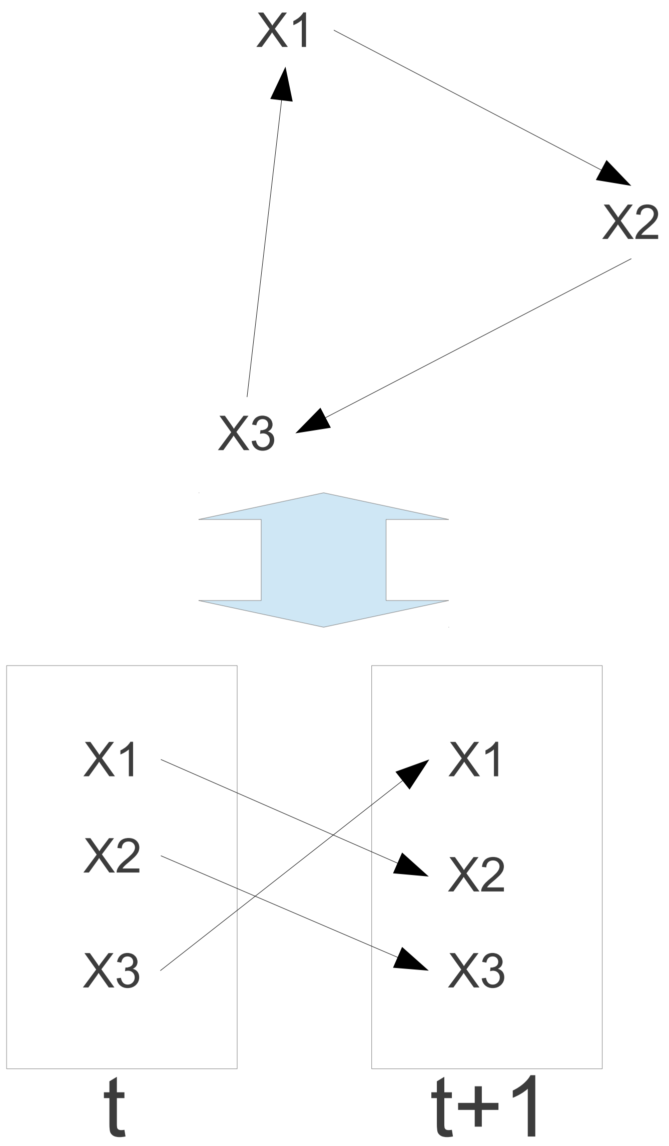


Dojer et al, BMC Bioinformatics '06

Network from Zak et al. '01

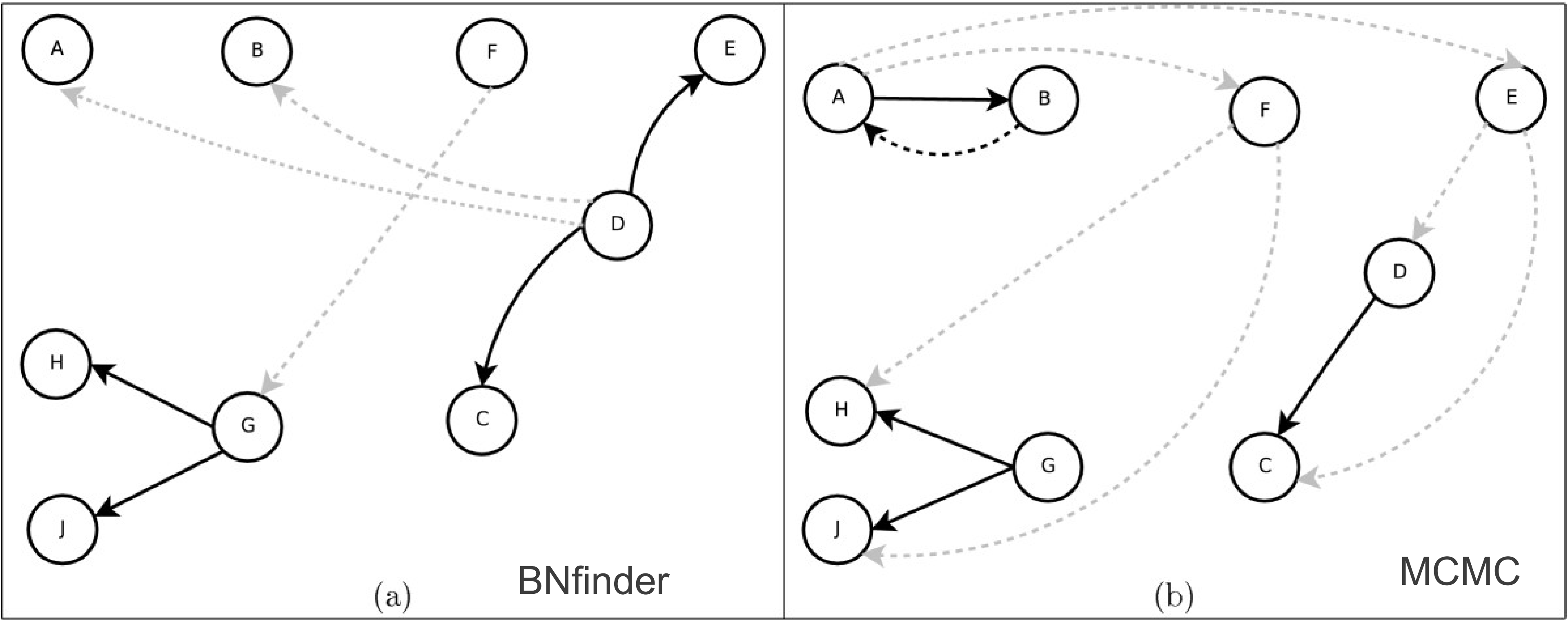
Dynamic Bayesian Networks

- Dynamic Bayesian Networks are an extension of the BN models to include temporal dependencies
- It is frequently used in a simple form, where only the dependencies between time points are allowed



DBNs reconstructed from „observational” expression data

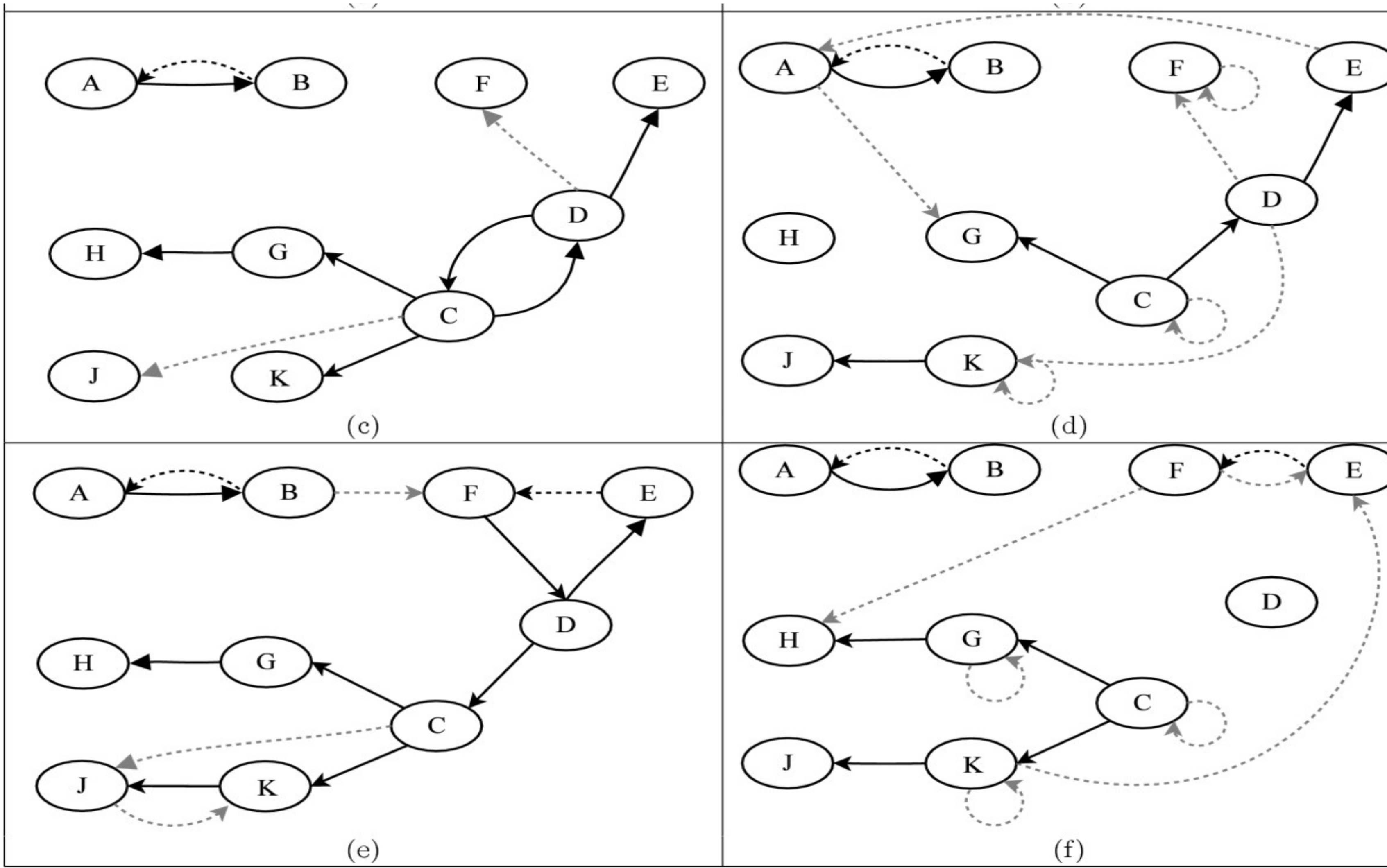
- 12 microarrays (simulated)
- Optimal discretization
- No knockouts



Networks reconstructed from data with knockouts: results vary depending on number of time-points

10x3 tp

10x12 tp



(c)

(d)

(e)

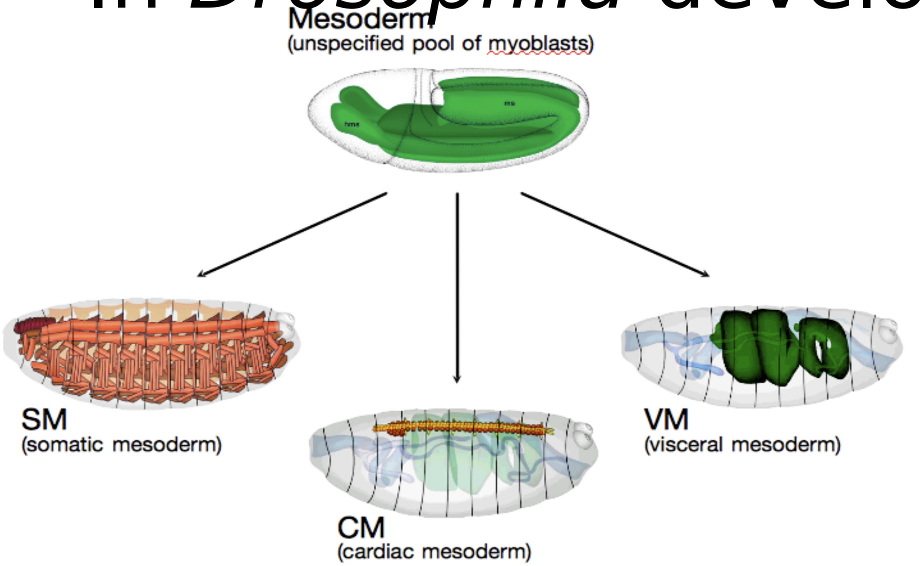
(f)

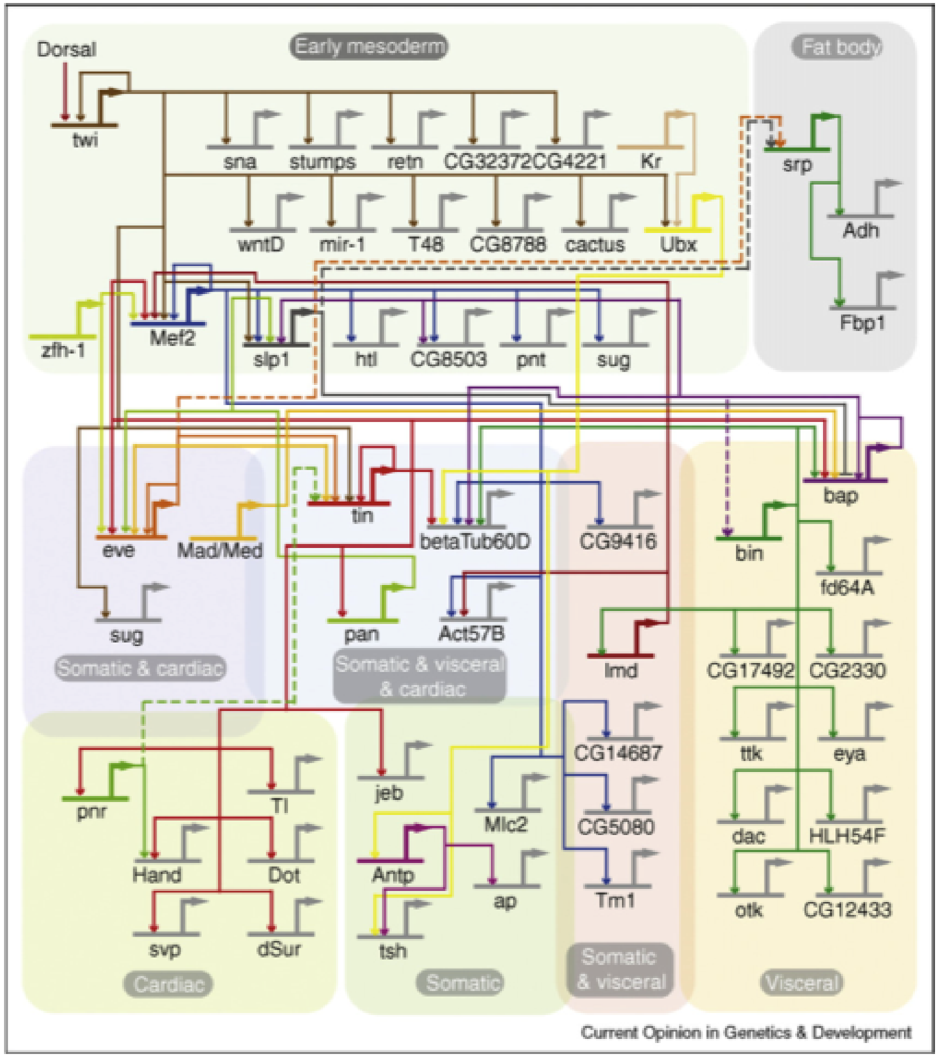
DBN reconstruction: summary

- The results you will get will vary depending on experimental conditions
- While significant, the number of faithfully reconstructed edges will fall short of 100%
- The network will most likely not correspond to a complete (even connected) model

Modelling transcriptional networks

In *Drosophila* development

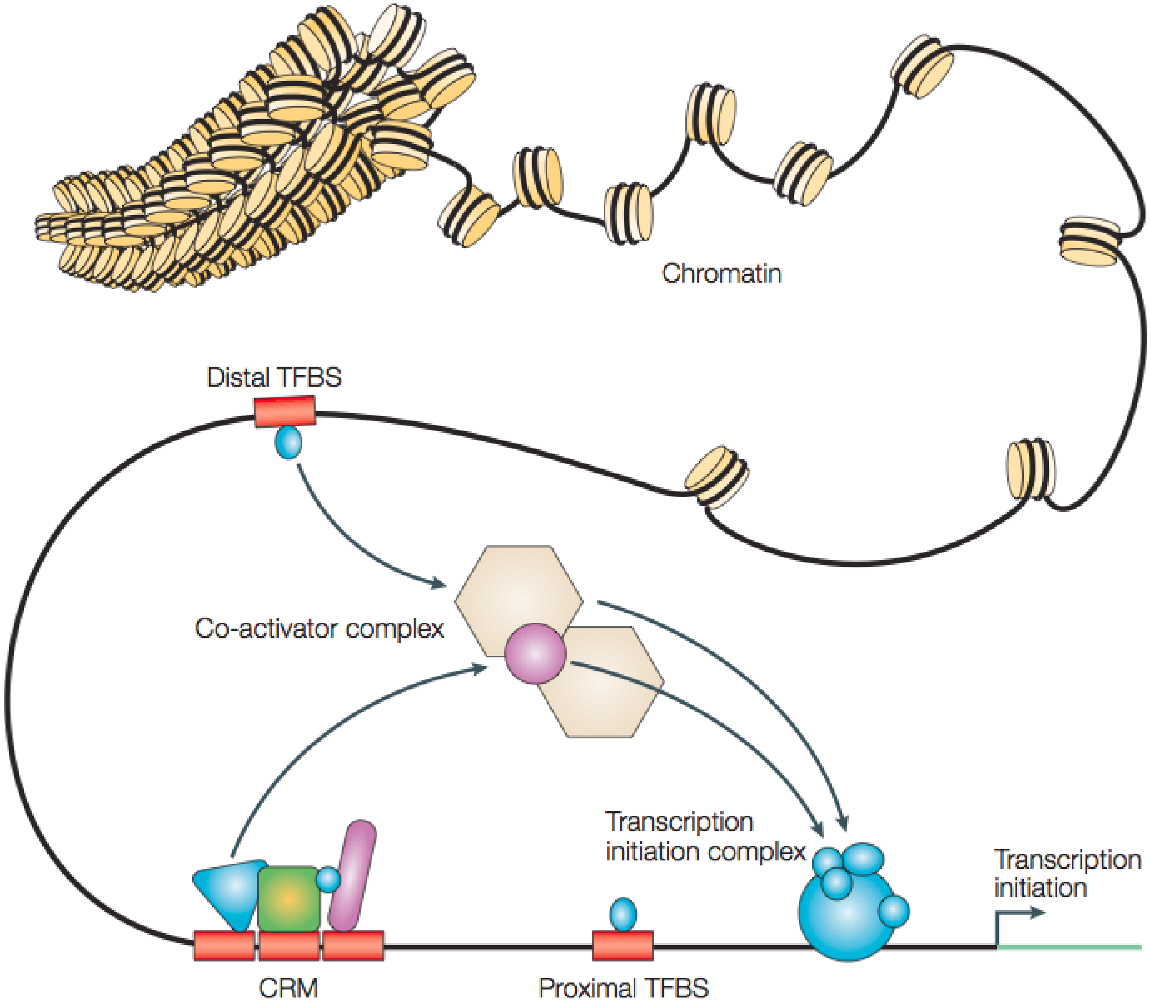


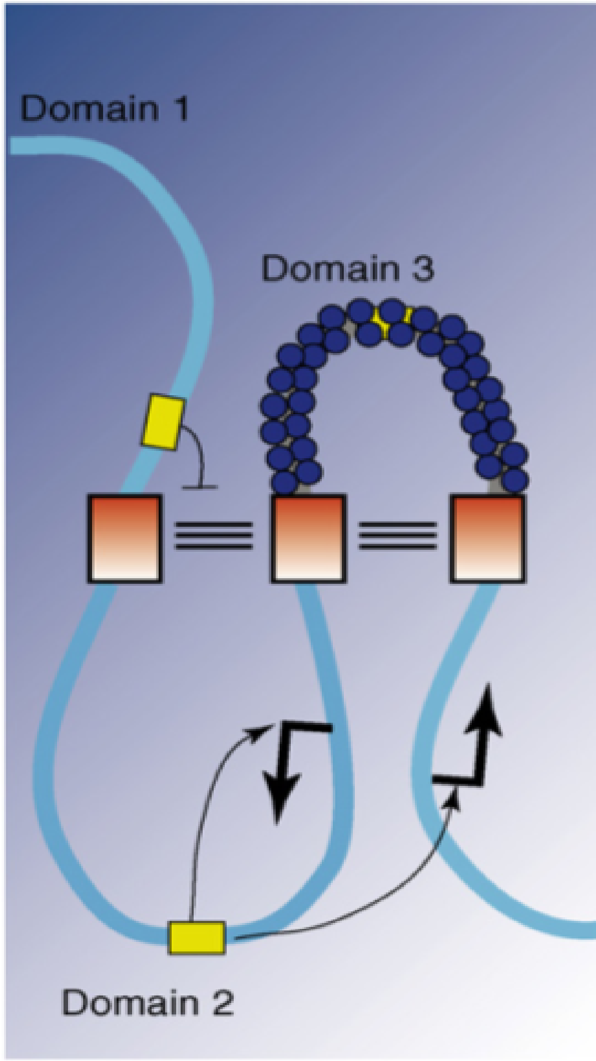


- Transcriptional networks as wiring diagrams
- Complexity on two different levels: many genes with many inputs

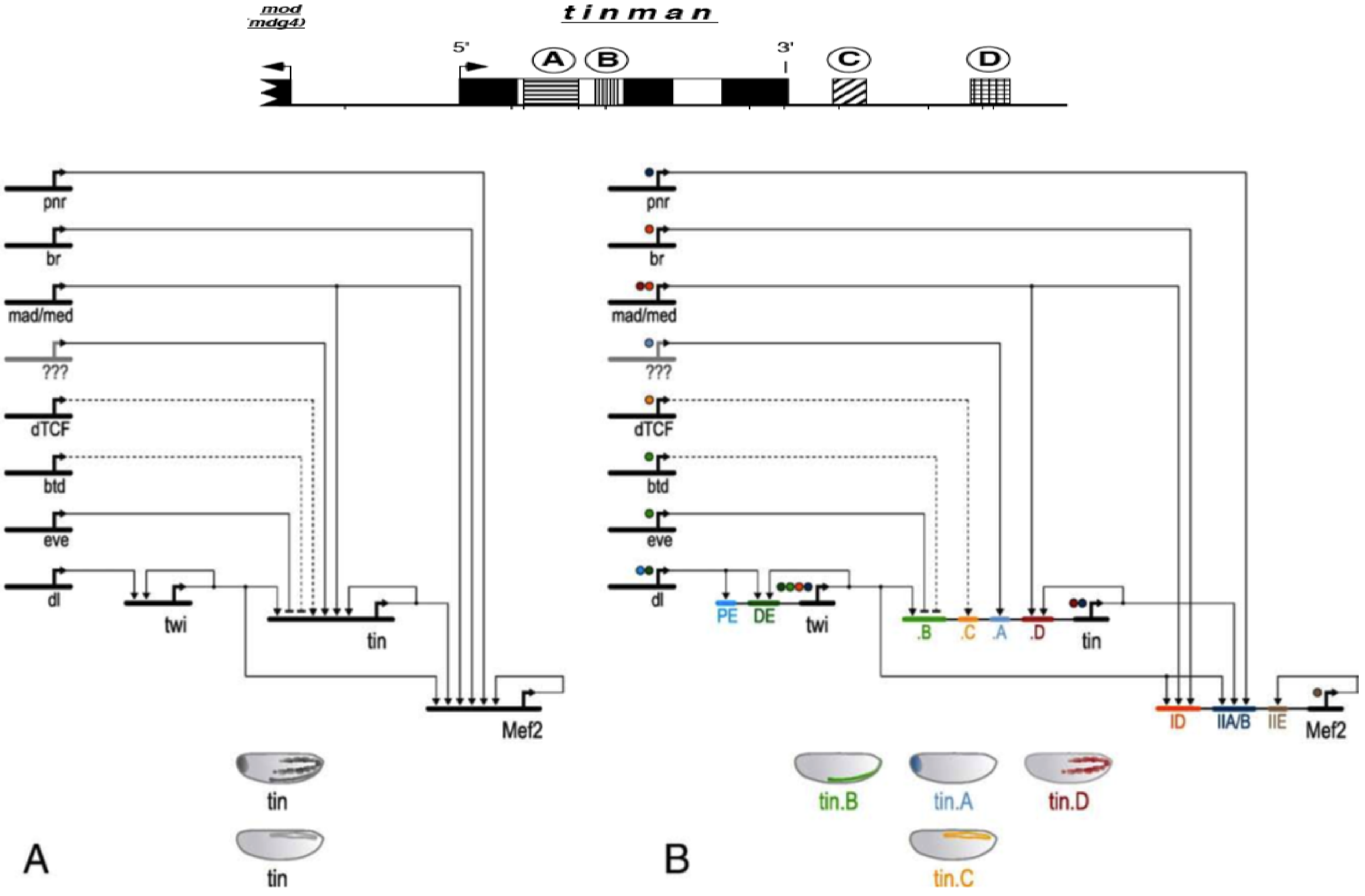
Bonn & Furlong, 2009

Transcription regulation



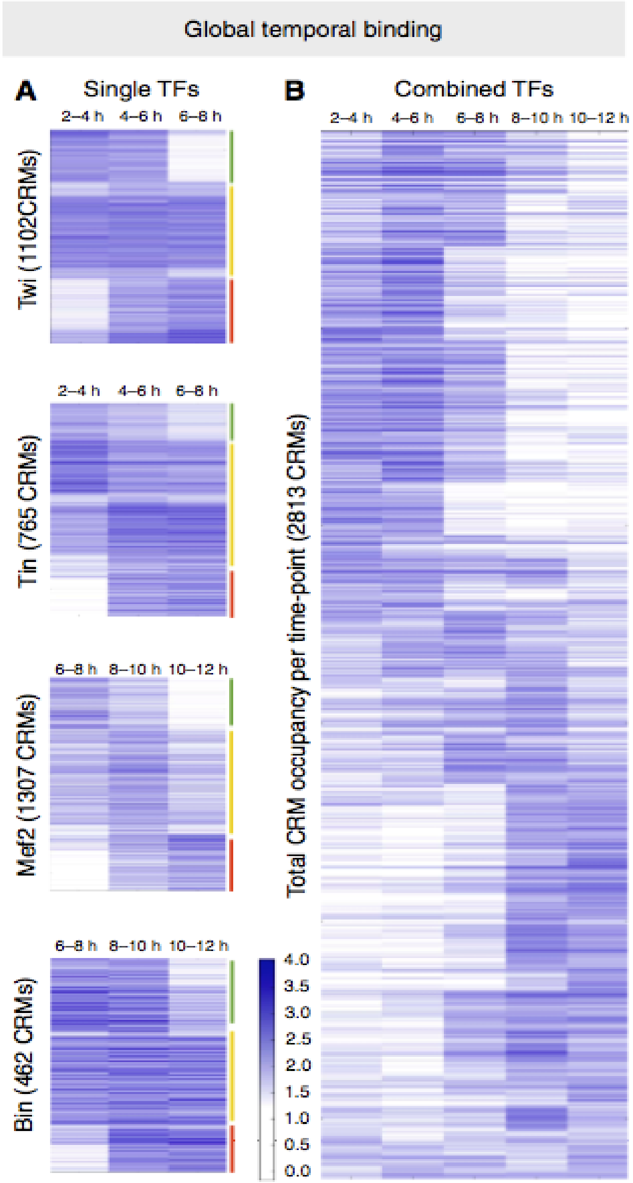


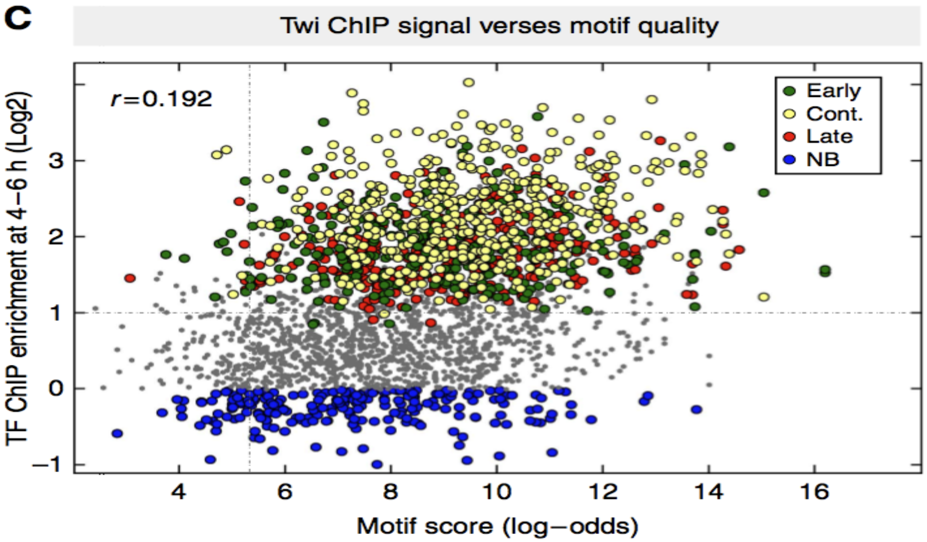
Genes *integrate* action of multiple enhancers

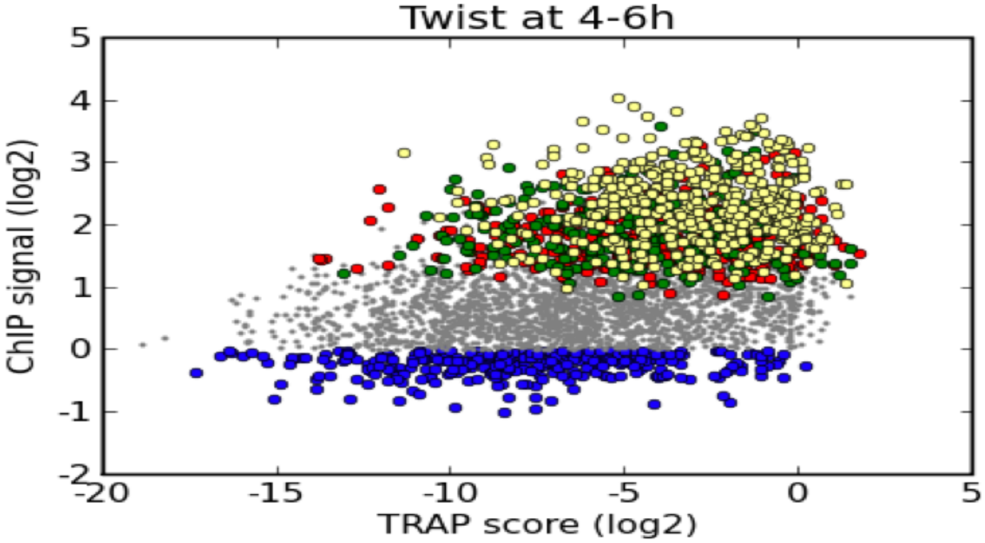


Wilczynski & Furlong, Dev. Biol, 2010

Temporal binding is unpredictable...

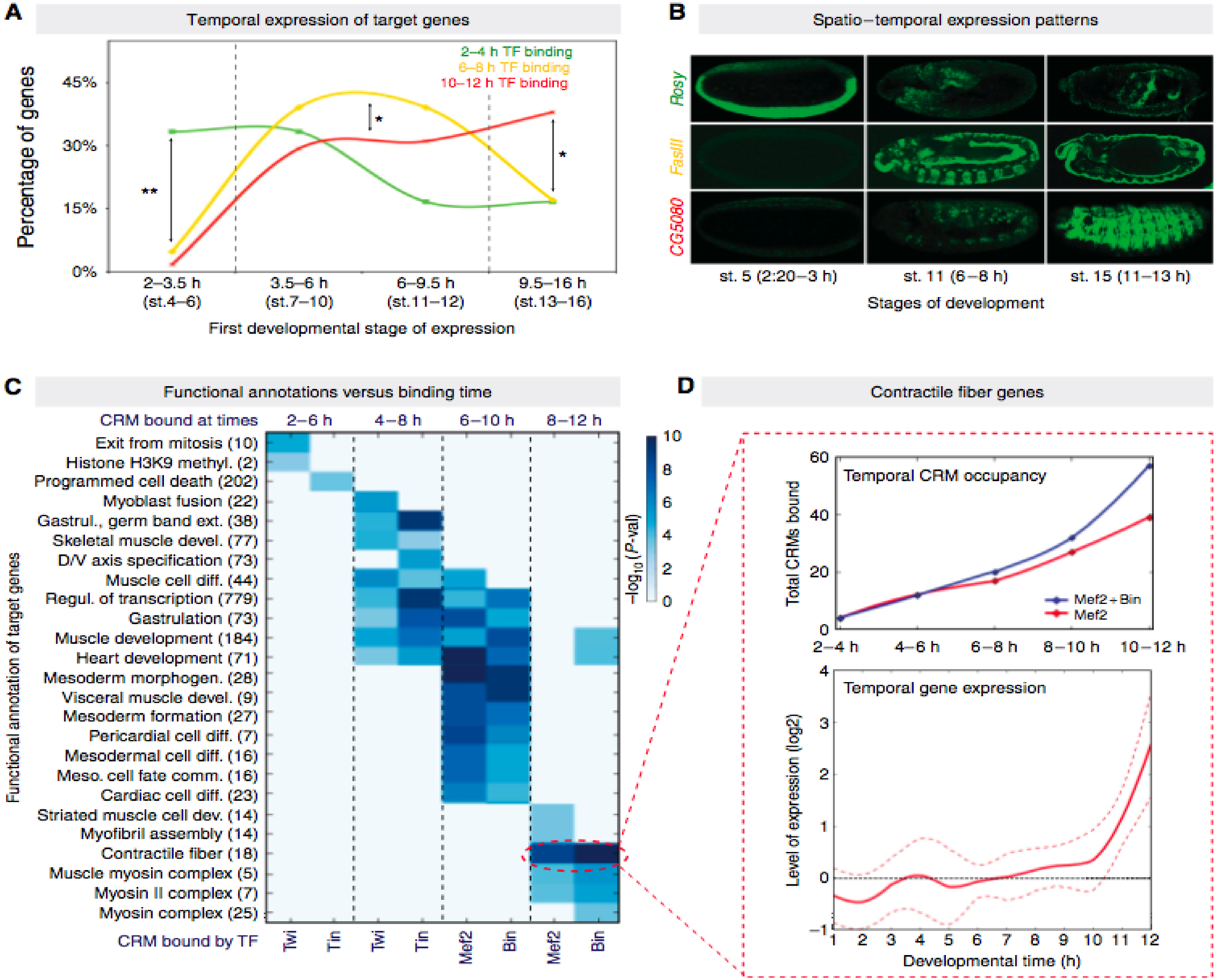






Wilczynski & Furlong, MSB, 2010

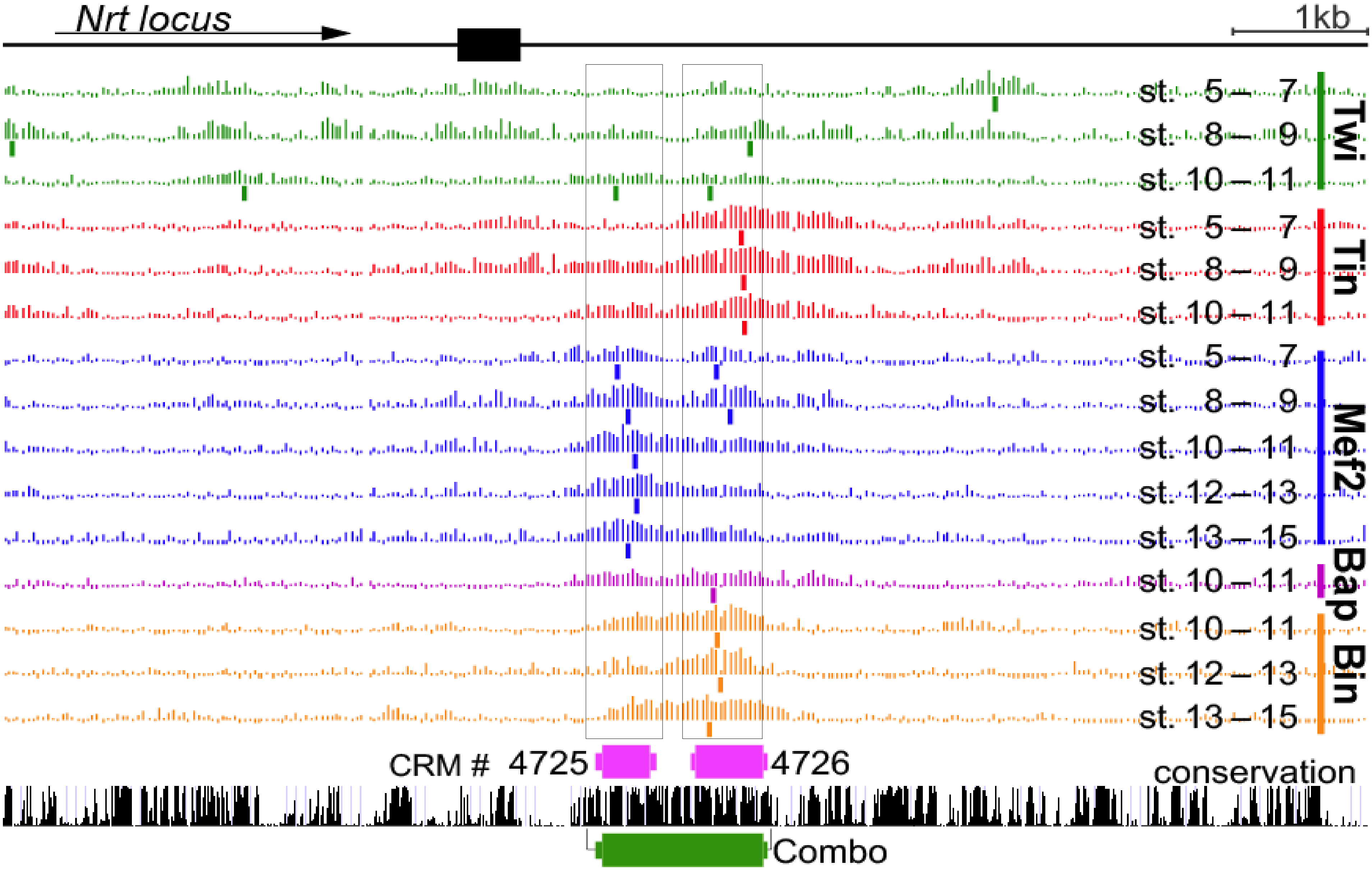
...yet reflects developmental function.



W

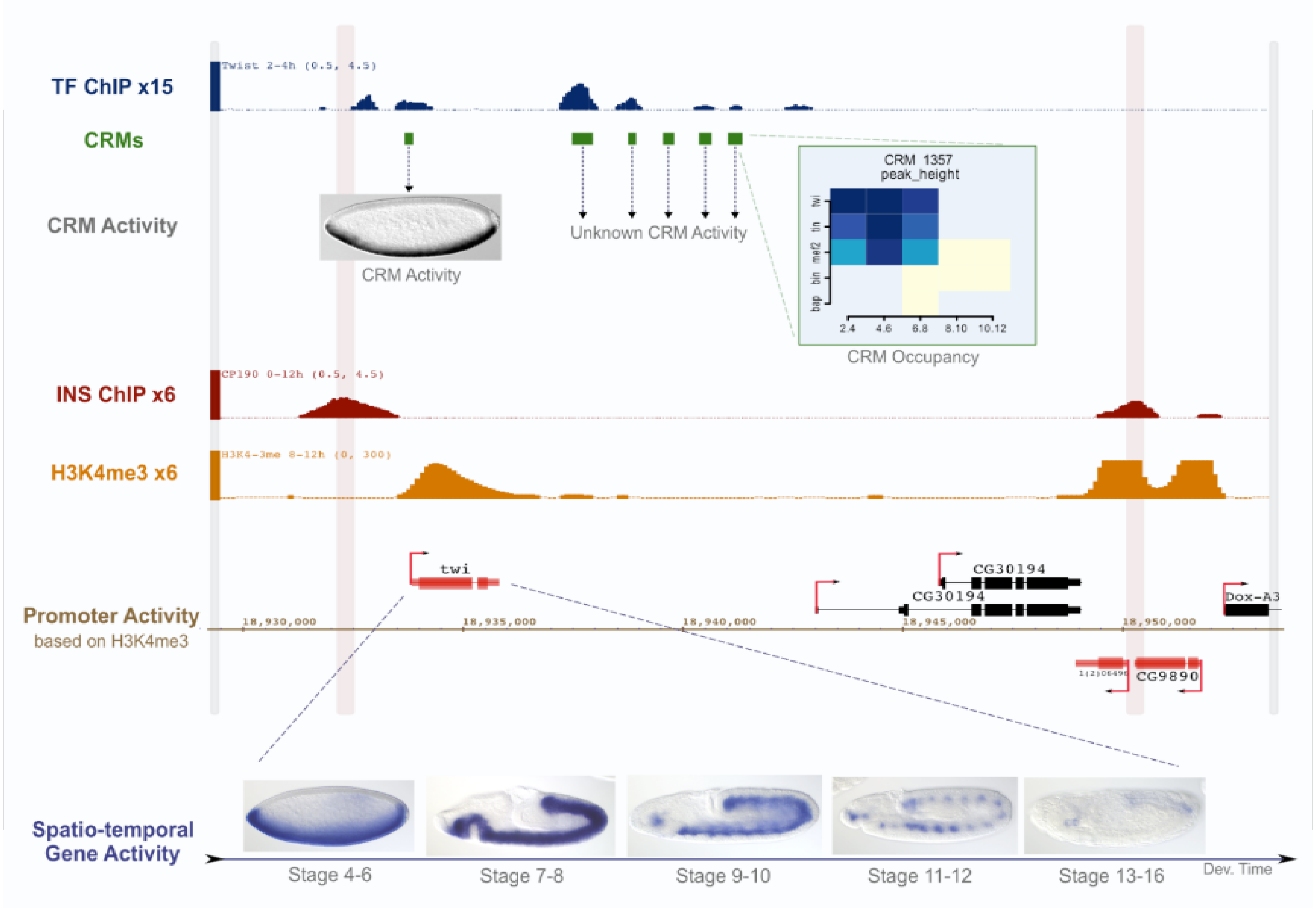
Wilczynski & Furlong, MSB, 2010

Mesoderm CRM atlas

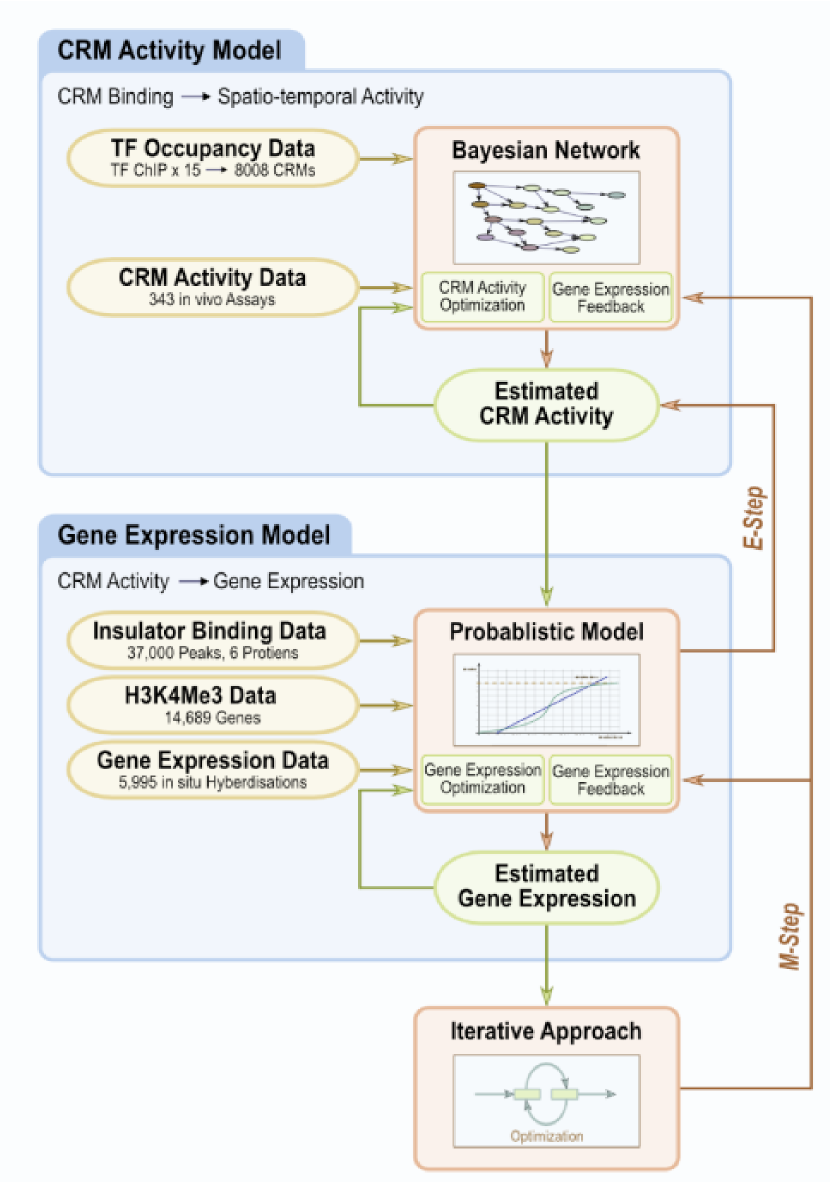


Zinzen, Gagneur, Girardot et al. 2009

Multiple layers of data in transcriptional regulation

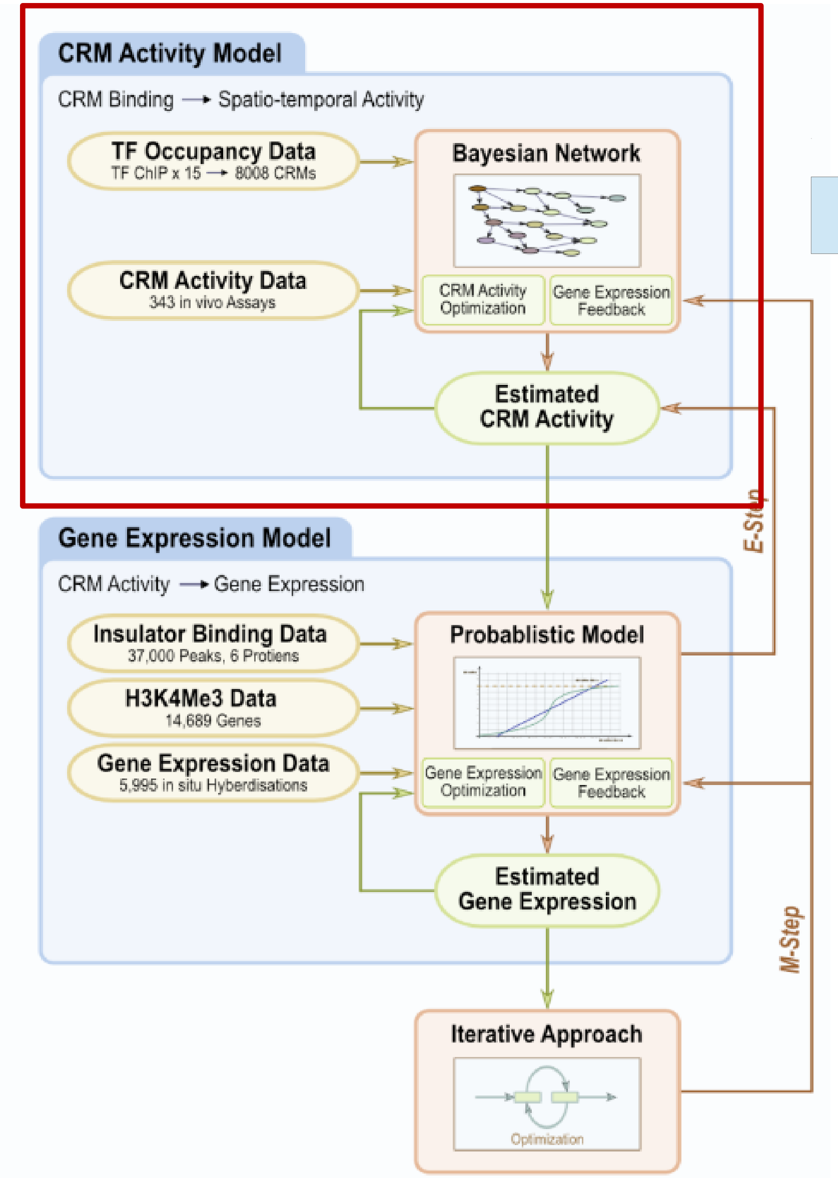


3 layer model of gene regulation



- 8008 enhancers compiled from 15 ChIP experiments (almost 20k binding peaks)
- Activity data for ~140 enhancers divided into
  - 3 tissues (MESO, VM, SM)
  - 5 stages (4-6,7-8,9-10,1112,13-16)
- Gene expression data for 5082 genes from the BDGP database

Model structure and optimization



CRM Activity Model

CRM Binding → Spatio-temporal Activity

TF Occupancy Data
TF ChIP x 15 → 8008 CRMs

CRM Activity Data
343 in vivo Assays

Bayesian Network

CRM Activity Optimization

Gene Expression Feedback

Estimated CRM Activity

E-Step

Gene Expression Model

CRM Activity → Gene Expression

Insulator Binding Data
37,000 Peaks, 6 Protiens

H3K4Me3 Data
14,689 Genes

Gene Expression Data
5,995 in situ Hyberdisations

Probablistic Model

Gene Expression Optimization

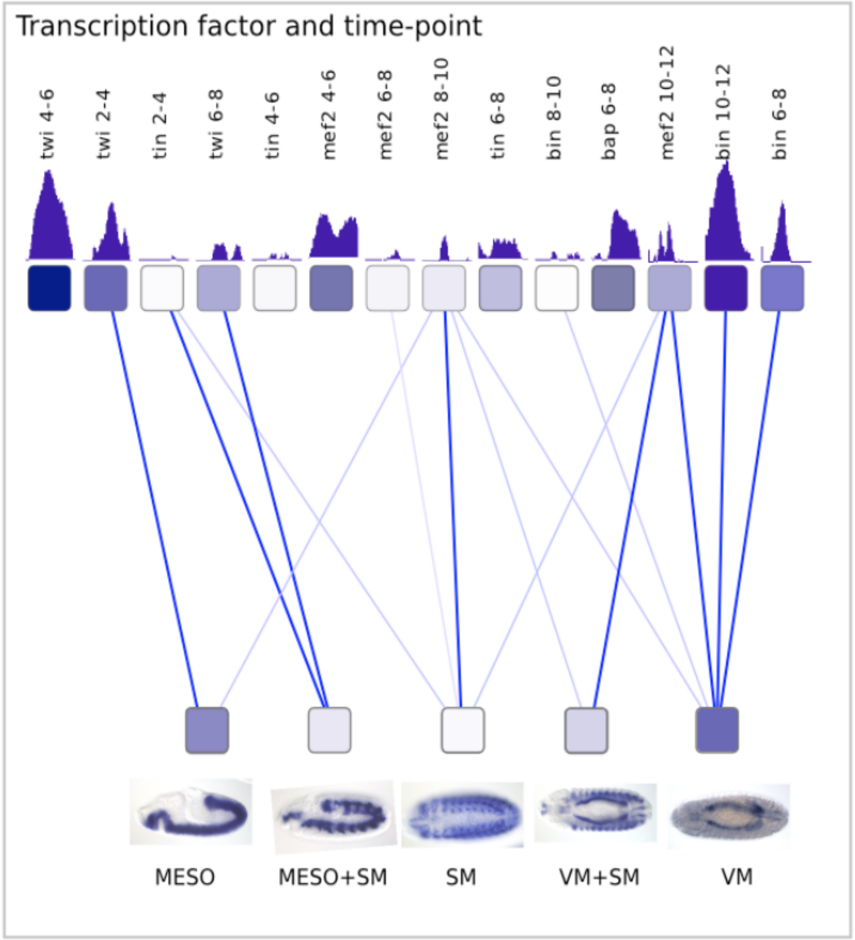
Gene Expression Feedback

Estimated Gene Expression

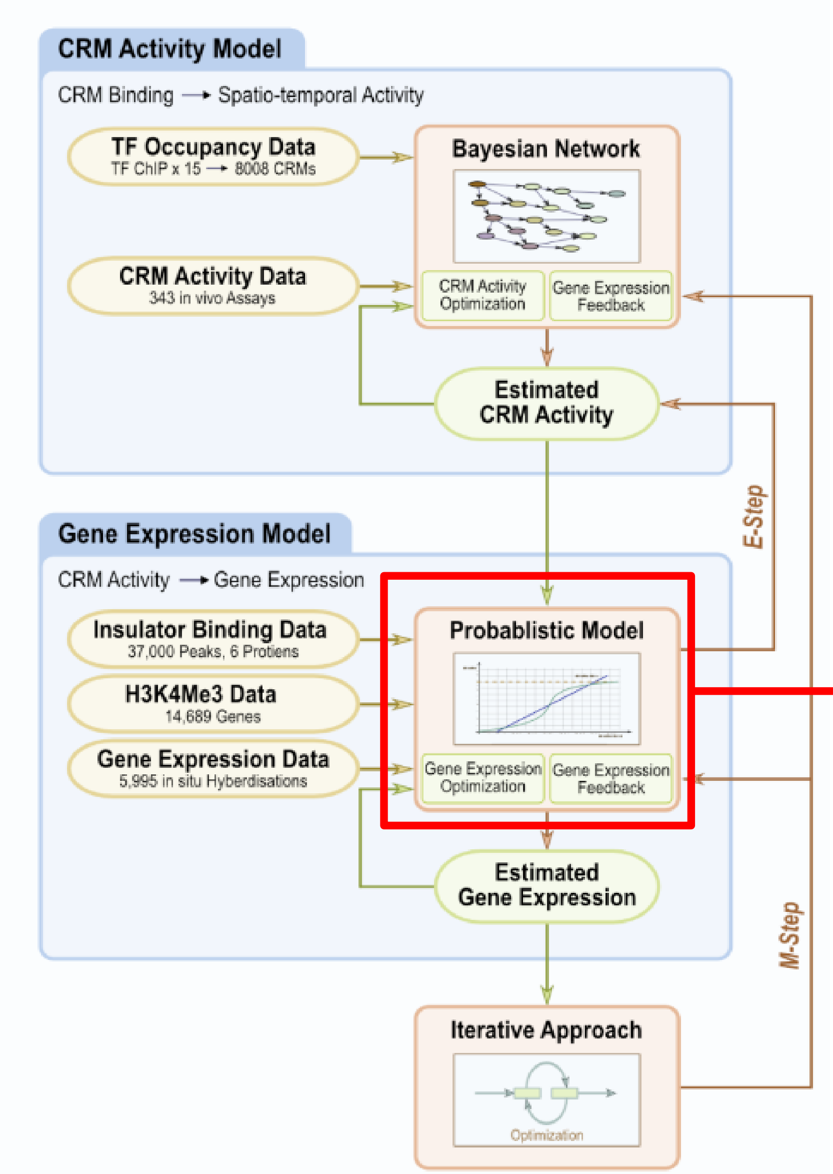
M-Step

Iterative Approach

Optimization



Model structure and optimization



CRM Activity Model

CRM Binding → Spatio-temporal Activity

TF Occupancy Data
TF ChIP x 15 → 8008 CRMs

CRM Activity Data
343 in vivo Assays

Bayesian Network

CRM Activity Optimization

Gene Expression Feedback

Estimated CRM Activity

Gene Expression Model

CRM Activity → Gene Expression

Insulator Binding Data
37,000 Peaks, 6 Protiens

H3K4Me3 Data
14,689 Genes

Gene Expression Data
5,995 in situ Hyberdisations

Probablistic Model

Gene Expression Optimization

Gene Expression Feedback

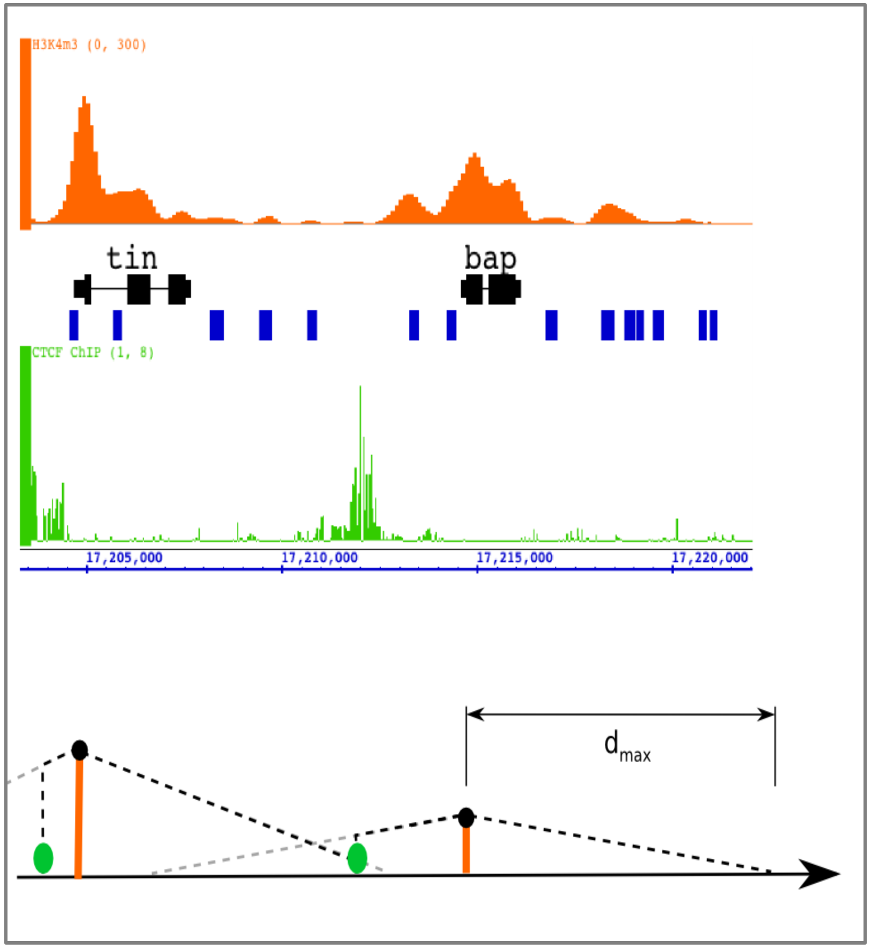
Estimated Gene Expression

Iterative Approach

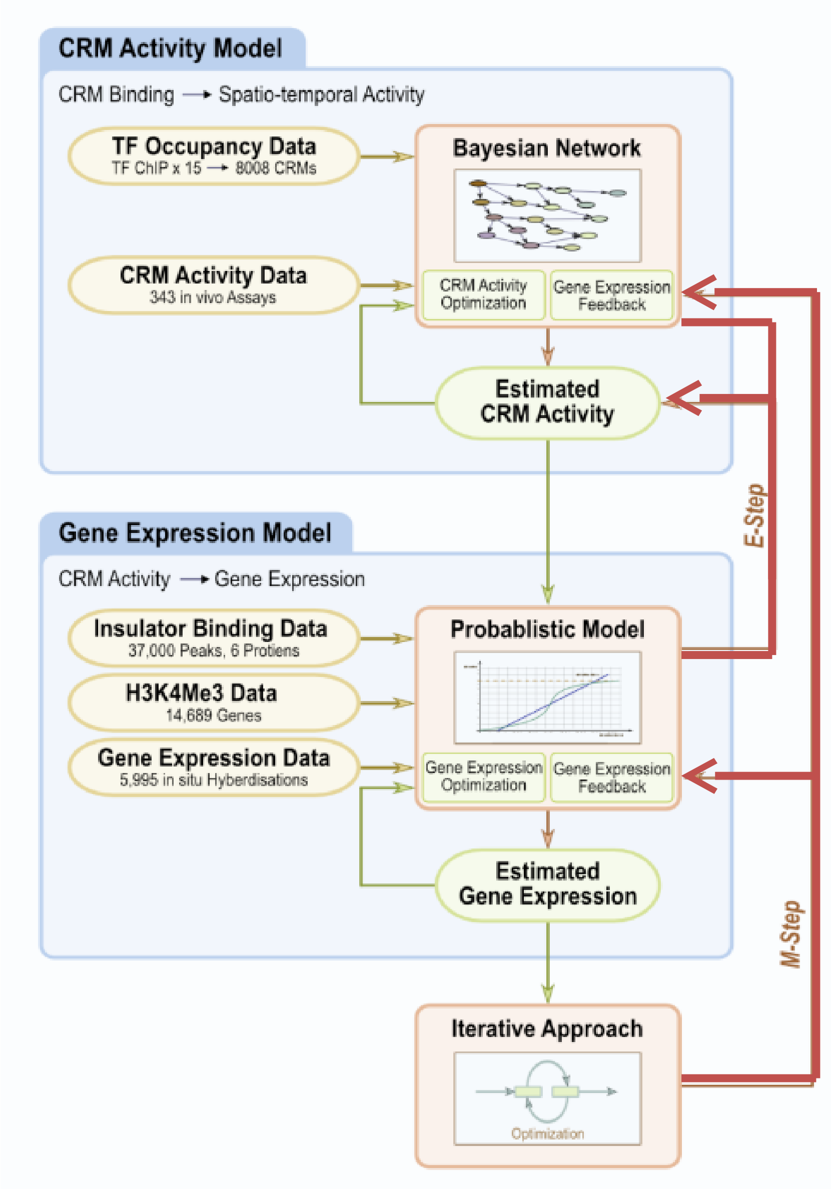
Optimization

E-Step

M-Step

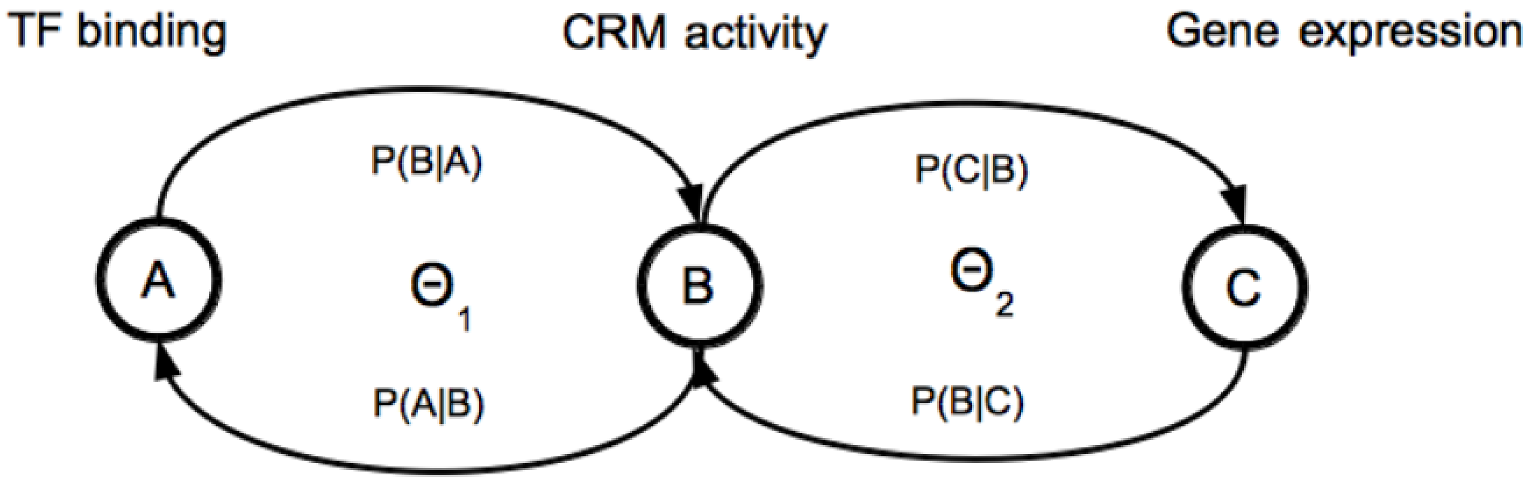


Expectation-Maximization approach



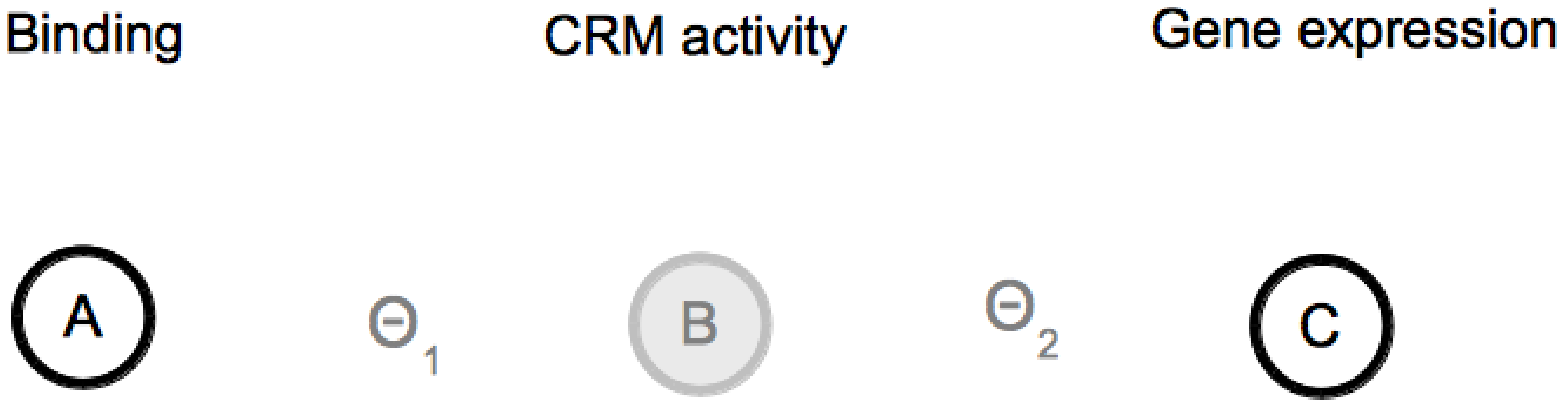
- CRM activity is in a large part a hidden variable for us
- With a probabilistic model we can use iterative strategy to find the optimal model combining the information from both TF binding and gene expression levels
- In the E-step we estimate the CRM activity and then in the M-step we maximize the BN and promoter models

Gene expression prediction



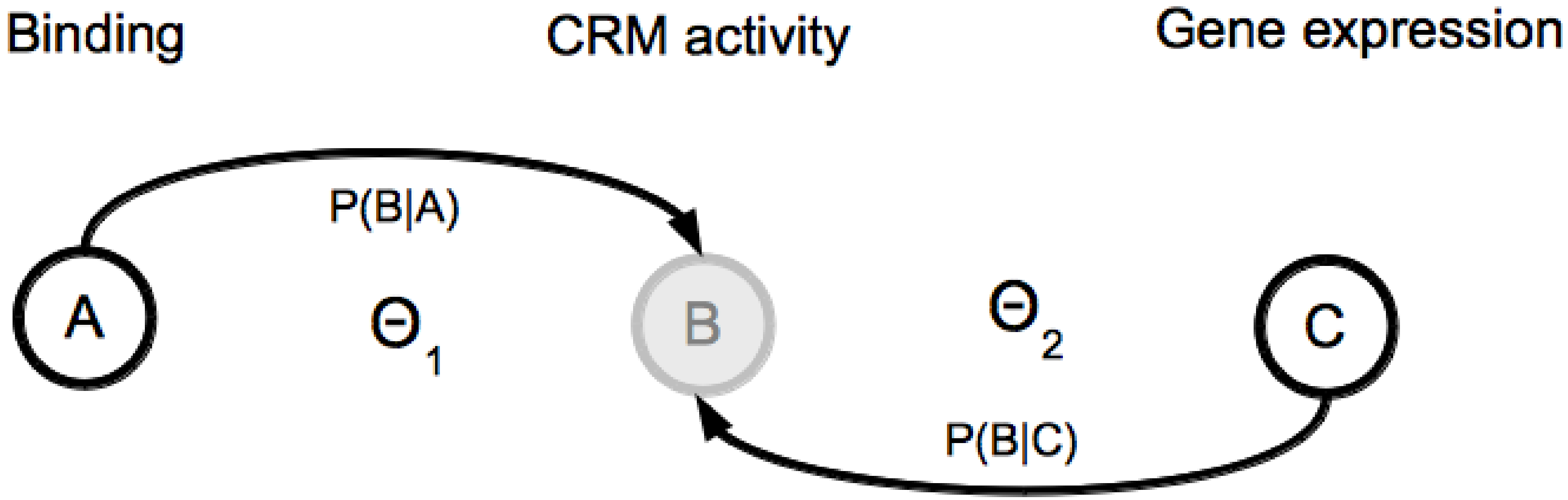
- Using known enhancers from CAD and in-situ annotations from BDGP
- We are trying to train a model predicting expression of target genes in different tissues
- Target gene assignment is key to linking CRM activity with expression of Targets

EM algorithm (1)



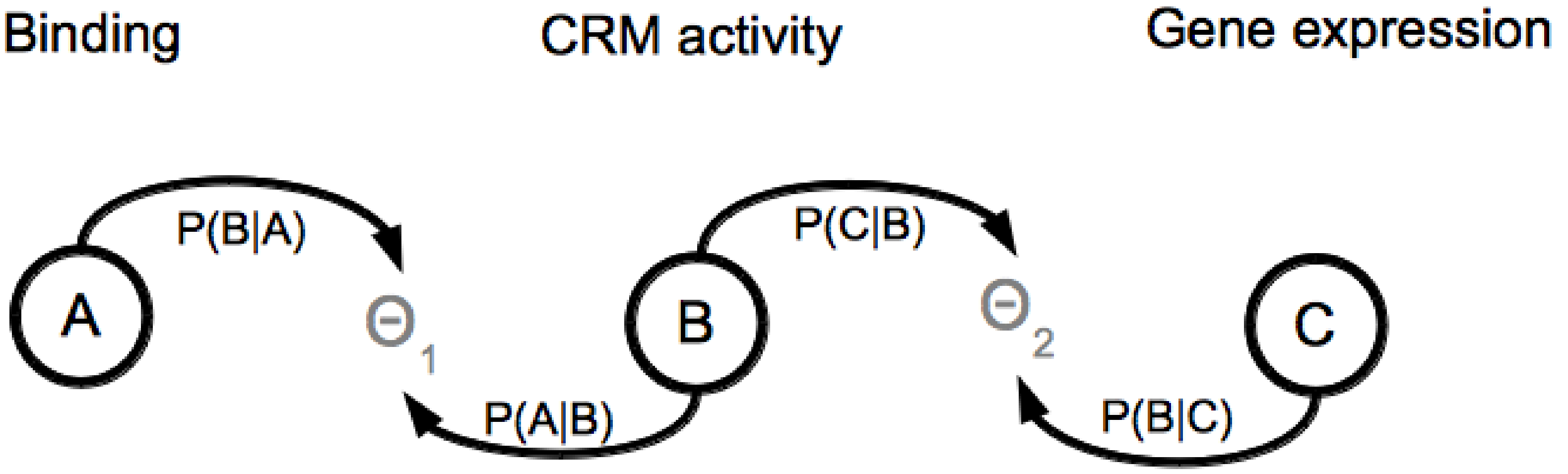
STEP 0: initialize parameter values to some (usually informed) guess
In our case: learn Θ_1 from CAD and set Θ_2 to maximal value.

EM algorithm (2)



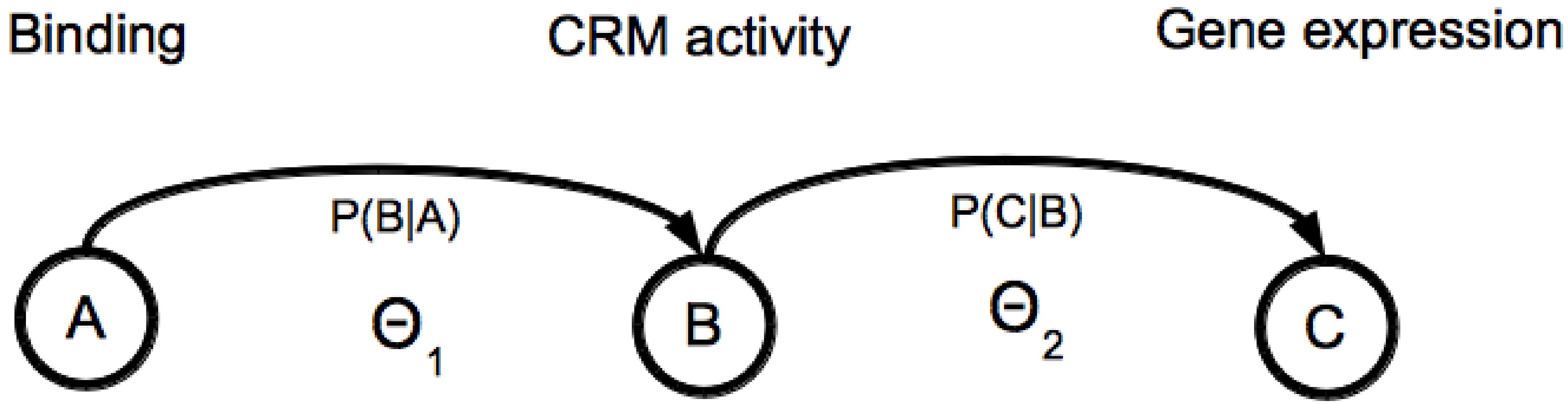
Once we have parameter values, we can calculate expected probabilities for hidden variables. That's E-STEP (expectation).

EM algorithm (3)



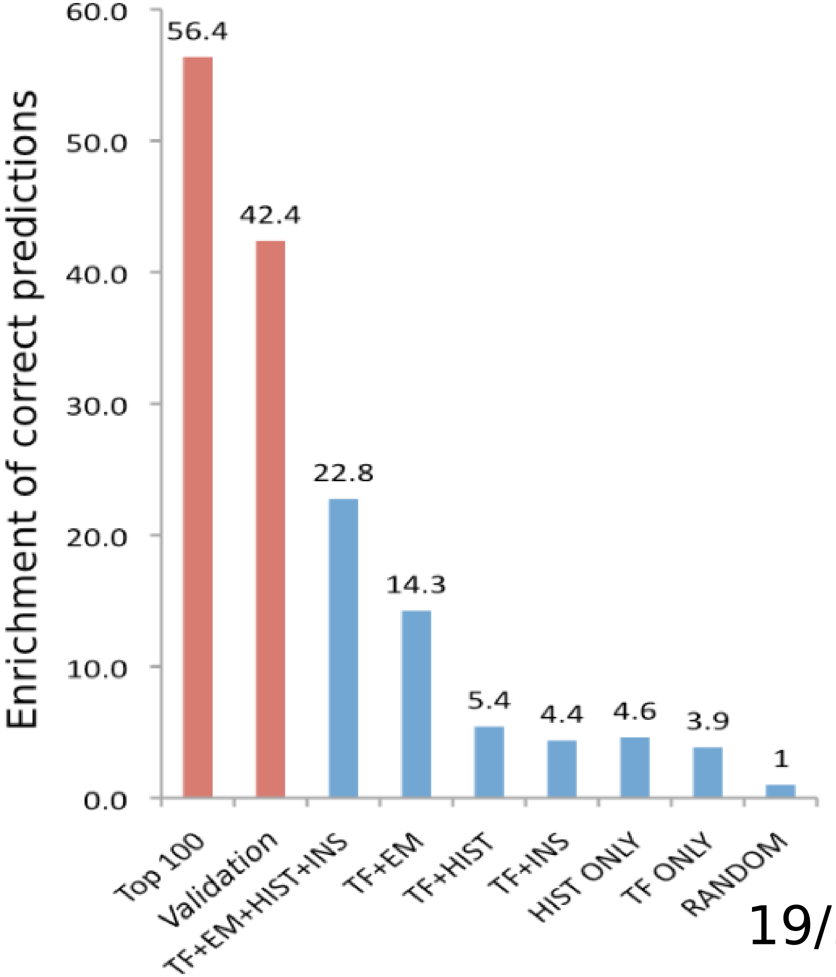
Once we have the estimated probabilities of hidden variables, we can find new set of parameters maximizing the likelihood function given our current expectation.
That's M-STEP (maximization)

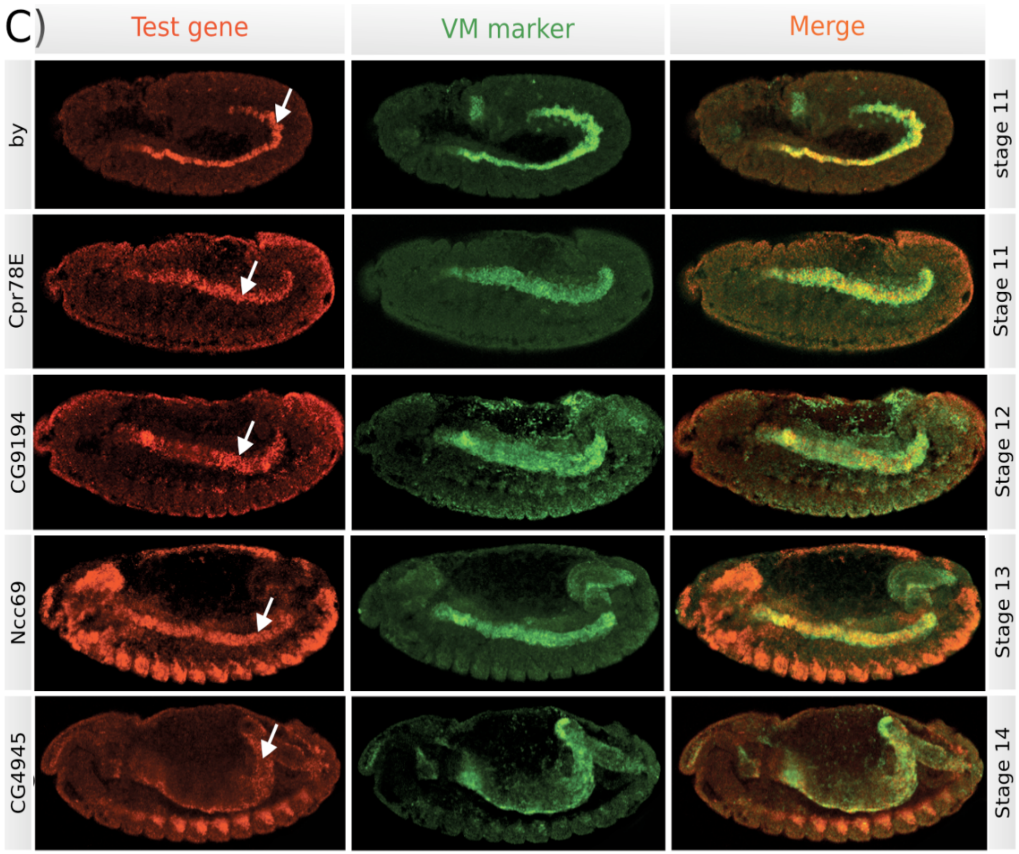
EM algorithm (4)



We can iterate these two steps until we reach convergence.
In the end, we have a full model which allows us to “predict” C from A.

Validation of predictions by in-situ hybrydization





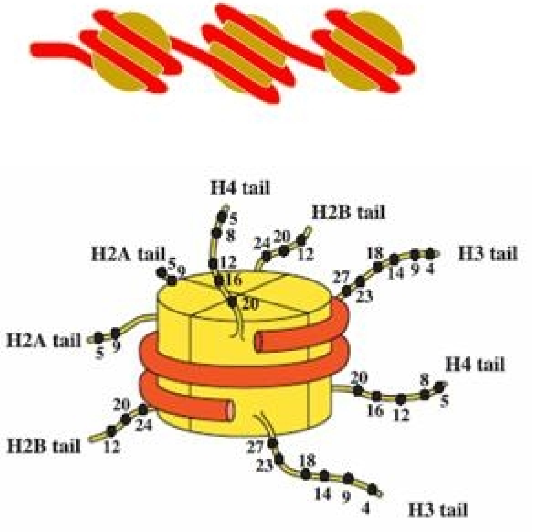
19/20 correct stage, 10/20 correct tissue

Wilczynski et al. submitted

Summary

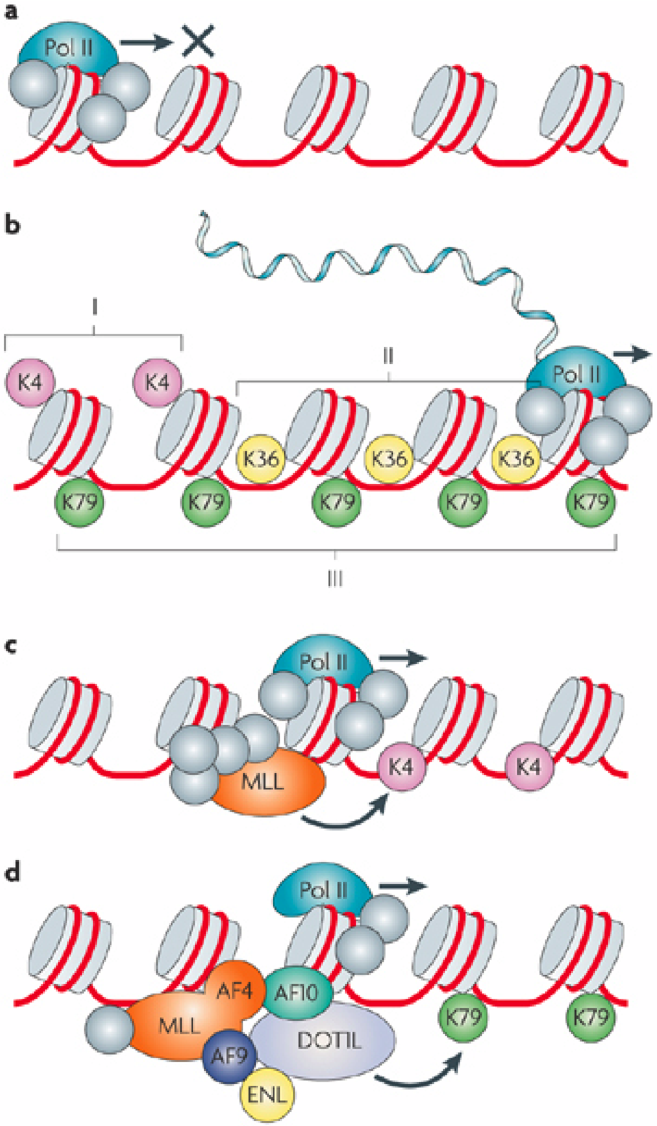
- Integrated model gives better predictions of gene expression and enhancer activity
- We see indication of enhancer sharing among developmental genes
- Very long range interactions do happen in Drosophila (typical locus is 50-100kb =~ 10 genes)
- We can find „missing players” that should be specific to related tissues

Histone modifications and their role in transcription



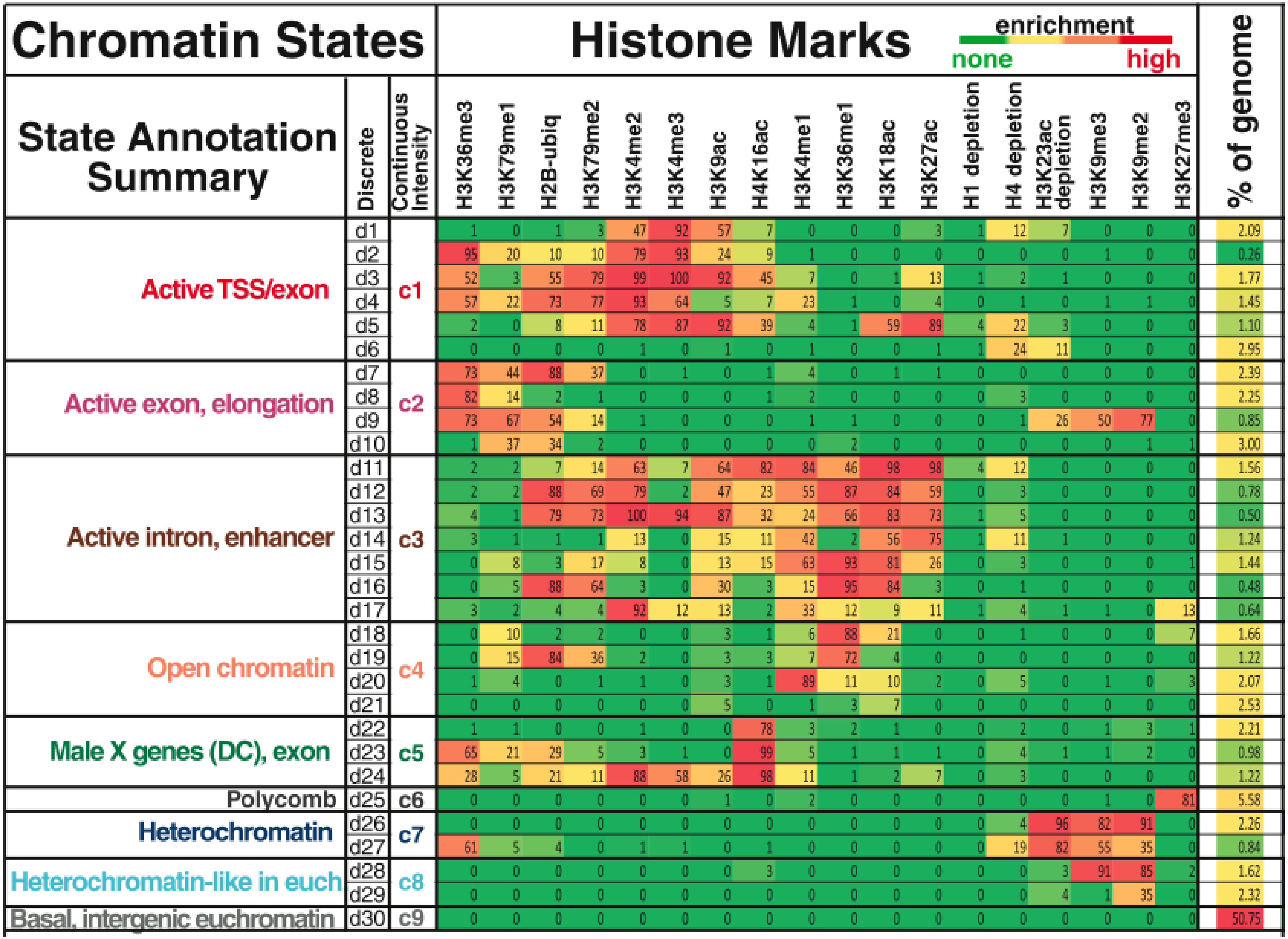
Ac	acetylation
Me	methylation
Ub	ubiquination
SU	sumoylation
P	phosphorylation

The figure illustrates nucleosome models and major posttranslational modifications which play essential roles in gene expression regulation and disease processes



Nature Reviews | Cancer

Modencode - Histone code?

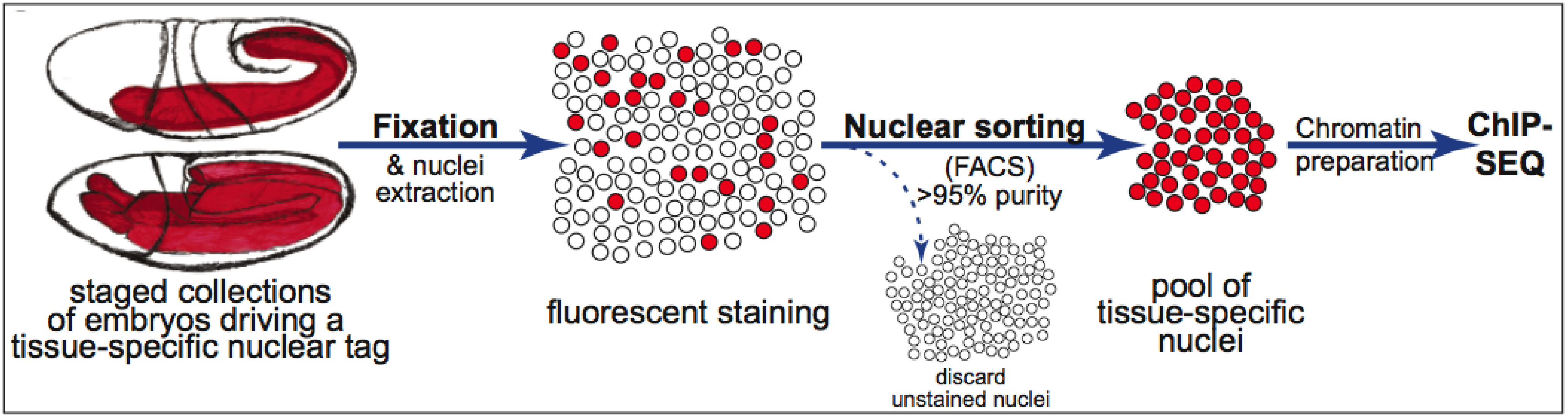


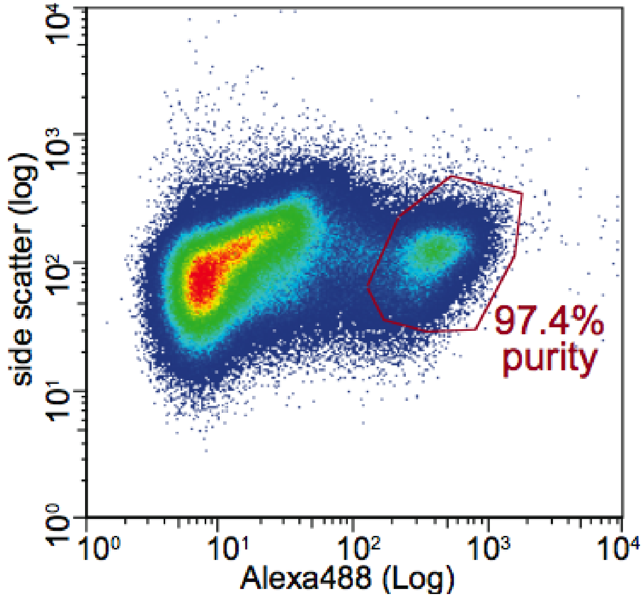
enrichment: none – high

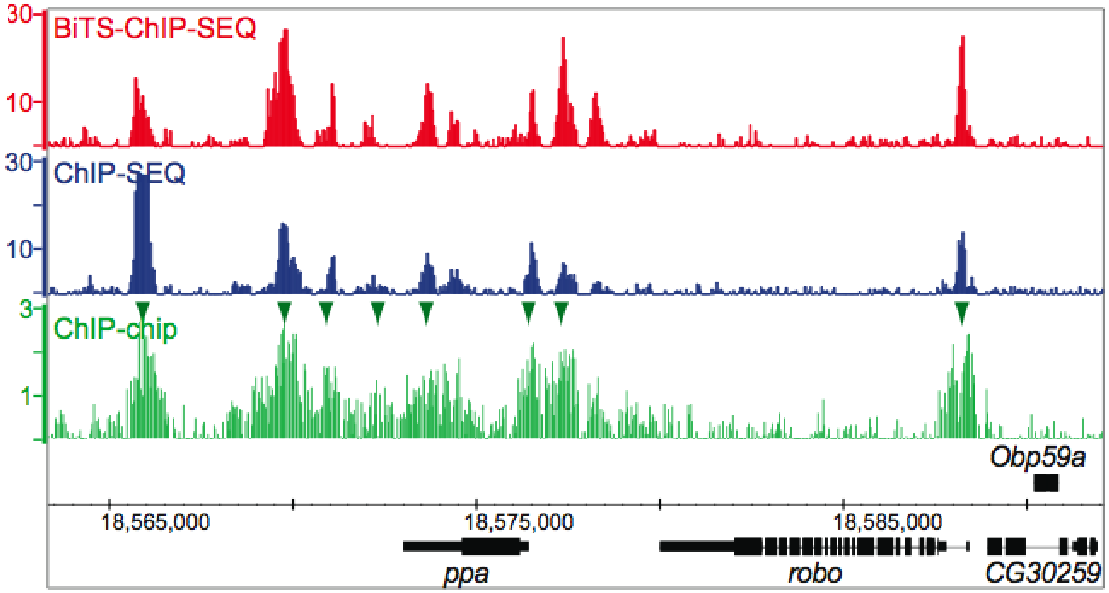
Chromatin States			Histone Marks																		% of genome
State Annotation Summary	Discrete	Continuous Intensity	H3K36me3	H3K79me1	H2B-ubiq	H3K79me2	H3K4me2	H3K4me3	H3K9ac	H4K16ac	H3K4me1	H3K36me1	H3K18ac	H3K27ac	H1 depletion	H4 depletion	H3K23ac depletion	H3K9me3	H3K9me2	H3K27me3	
Active TSS/exon	d1	c1	1	0	1	3	47	92	57	7	0	0	0	3	1	12	7	0	0	0	2.09
	d2		95	20	10	10	79	93	24	9	1	0	0	0	0	0	0	1	0	0	0.26
	d3		52	3	55	79	99	100	92	45	7	0	1	13	1	2	1	0	0	0	1.77
	d4		57	22	73	77	93	64	5	7	23	1	0	4	0	1	0	1	1	0	1.45
	d5		2	0	8	11	78	87	92	39	4	1	59	89	4	22	3	0	0	0	1.10
	d6		0	0	0	0	1	0	1	0	1	0	0	1	1	24	11	0	0	0	2.95
Active exon, elongation	d7	c2	73	44	88	37	0	1	0	1	4	0	1	1	0	0	0	0	0	0	2.39
	d8		82	14	2	1	0	0	0	1	2	0	0	0	0	3	0	0	0	0	2.25
	d9		73	67	54	14	1	0	0	0	1	1	0	0	0	1	26	50	77	0	0.85
	d10		1	37	34	2	0	0	0	0	0	2	0	0	0	0	0	0	1	1	3.00
Active intron, enhancer	d11	c3	2	2	7	14	63	7	64	82	84	46	98	98	4	12	0	0	0	0	1.56
	d12		2	2	88	69	79	2	47	23	55	87	84	59	0	3	0	0	0	0	0.78
	d13		4	1	79	73	100	94	87	32	24	66	83	73	1	5	0	0	0	0	0.50
	d14		3	1	1	1	13	0	15	11	42	2	56	75	1	11	1	0	0	0	1.24
	d15		0	8	3	17	8	0	13	15	63	93	81	26	0	3	0	0	0	1	1.44
	d16		0	5	88	64	3	0	30	3	15	95	84	3	0	1	0	0	0	0	0.48
	d17		3	2	4	4	92	12	13	2	33	12	9	11	1	4	1	1	0	13	0.64
Open chromatin	d18	c4	0	10	2	2	0	0	3	1	6	88	21	0	0	1	0	0	0	7	1.66
	d19		0	15	84	36	2	0	3	3	7	72	4	0	0	0	0	0	0	0	1.22
	d20		1	4	0	1	1	0	3	1	89	11	10	2	0	5	0	1	0	3	2.07
	d21		0	0	0	0	0	0	5	0	1	3	7	0	0	0	0	0	0	0	2.53
Male X genes (DC), exon	d22	c5	1	1	0	0	1	0	0	78	3	2	1	0	0	2	0	1	3	1	2.21
	d23		65	21	29	5	3	1	0	99	5	1	1	1	0	4	1	1	2	0	0.98
	d24		28	5	21	11	88	58	26	98	11	1	2	7	0	3	0	0	0	0	1.22
Polycomb	d25	c6	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	1	0	81	5.58
Heterochromatin	d26	c7	0	0	0	0	0	0	0	0	0	0	0	0	0	4	96	82	91	0	2.26
	d27		61	5	4	0	1	1	0	1	0	0	0	0	0	19	82	55	35	0	0.84
Heterochromatin-like in euch	d28	c8	0	0	0	0	0	0	0	3	0	0	0	0	0	0	3	91	85	2	1.62
	d29		0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1	35	0	2.32
Basal, intergenic euchromatin	d30	c9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50.75

Modencode, Roy et al, Science 2010

Getting a clear picture - BiTS-Chip

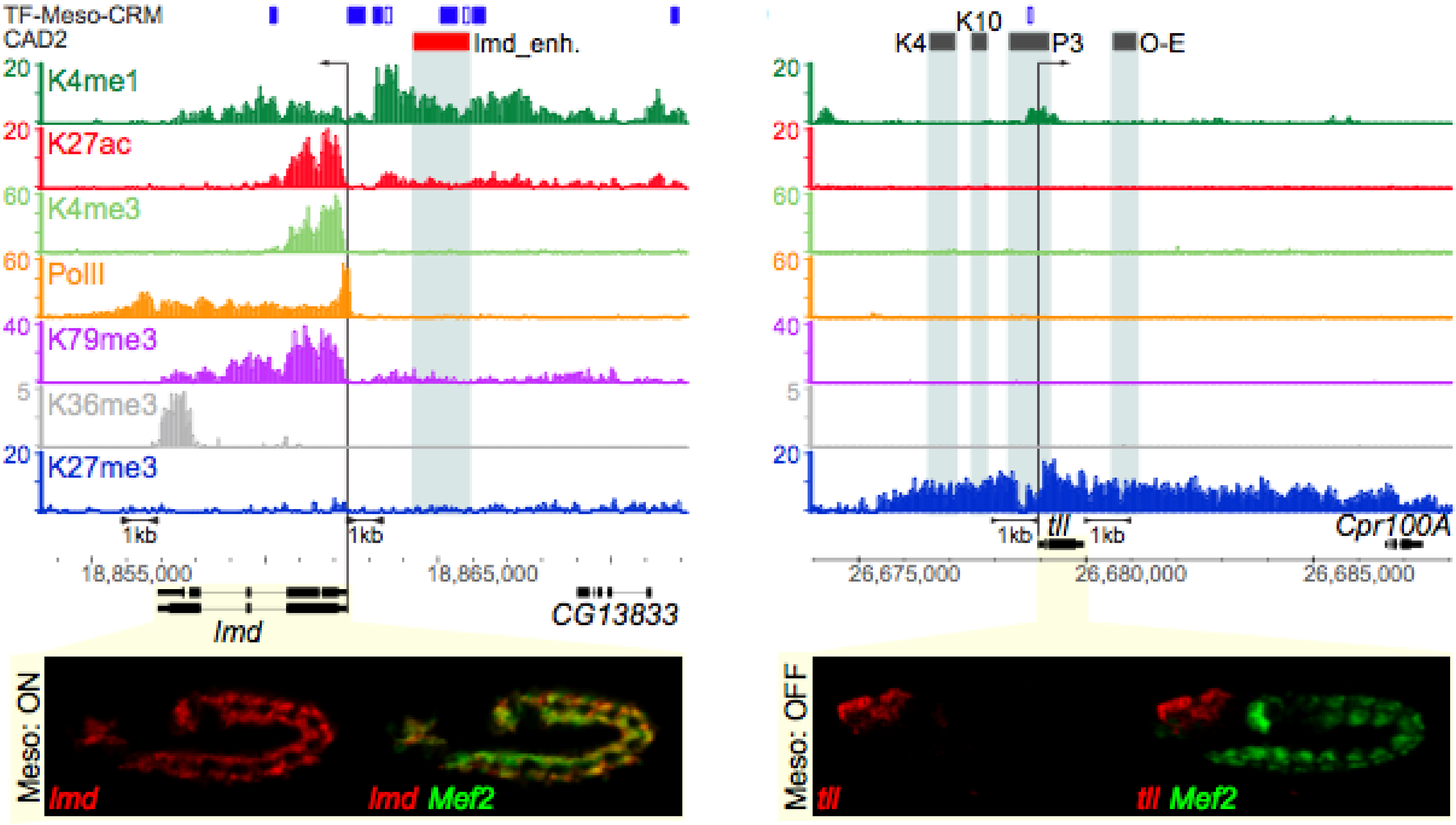






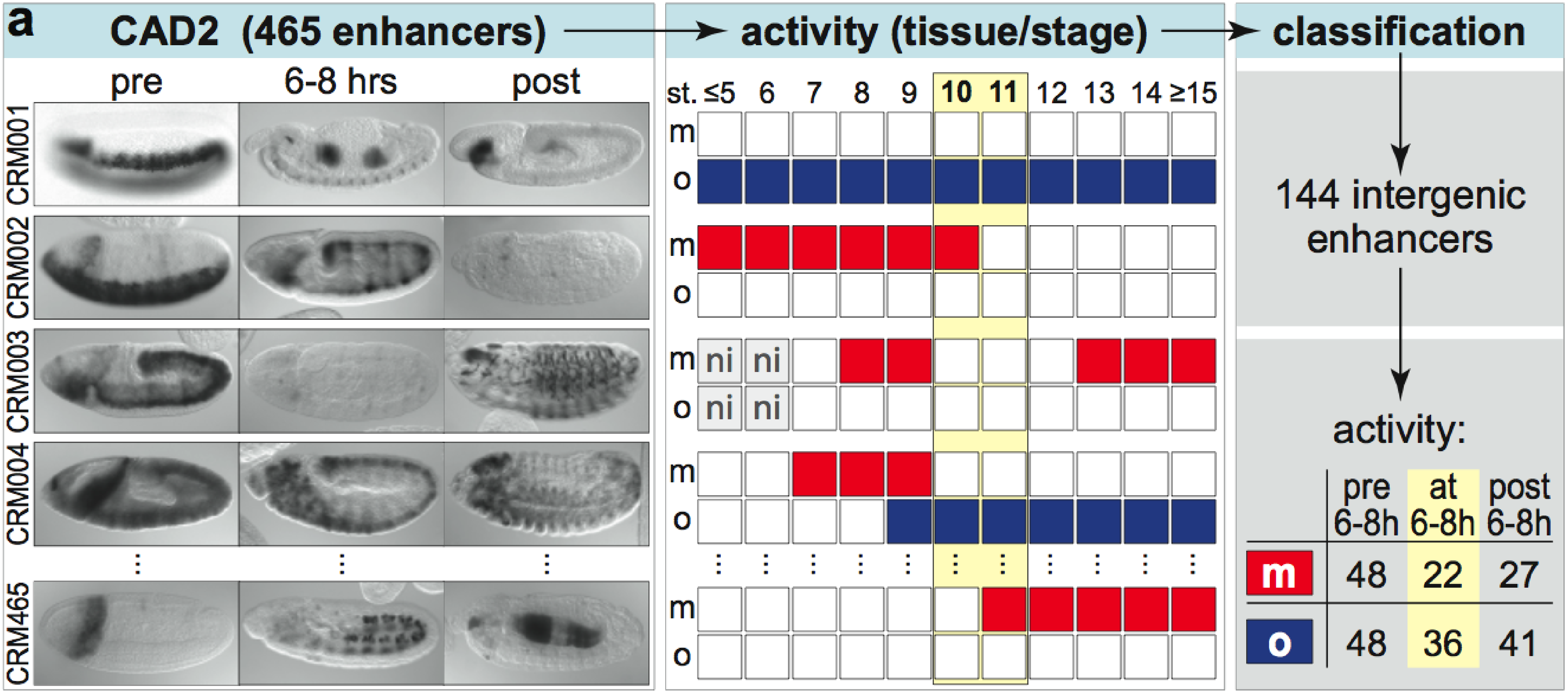
Bonn et al. Nat. Genet, 2012

What do we see in regions with known genes



Bonn et al. Nat. Genet, 201

Can we extract patterns for „known” enhancers?

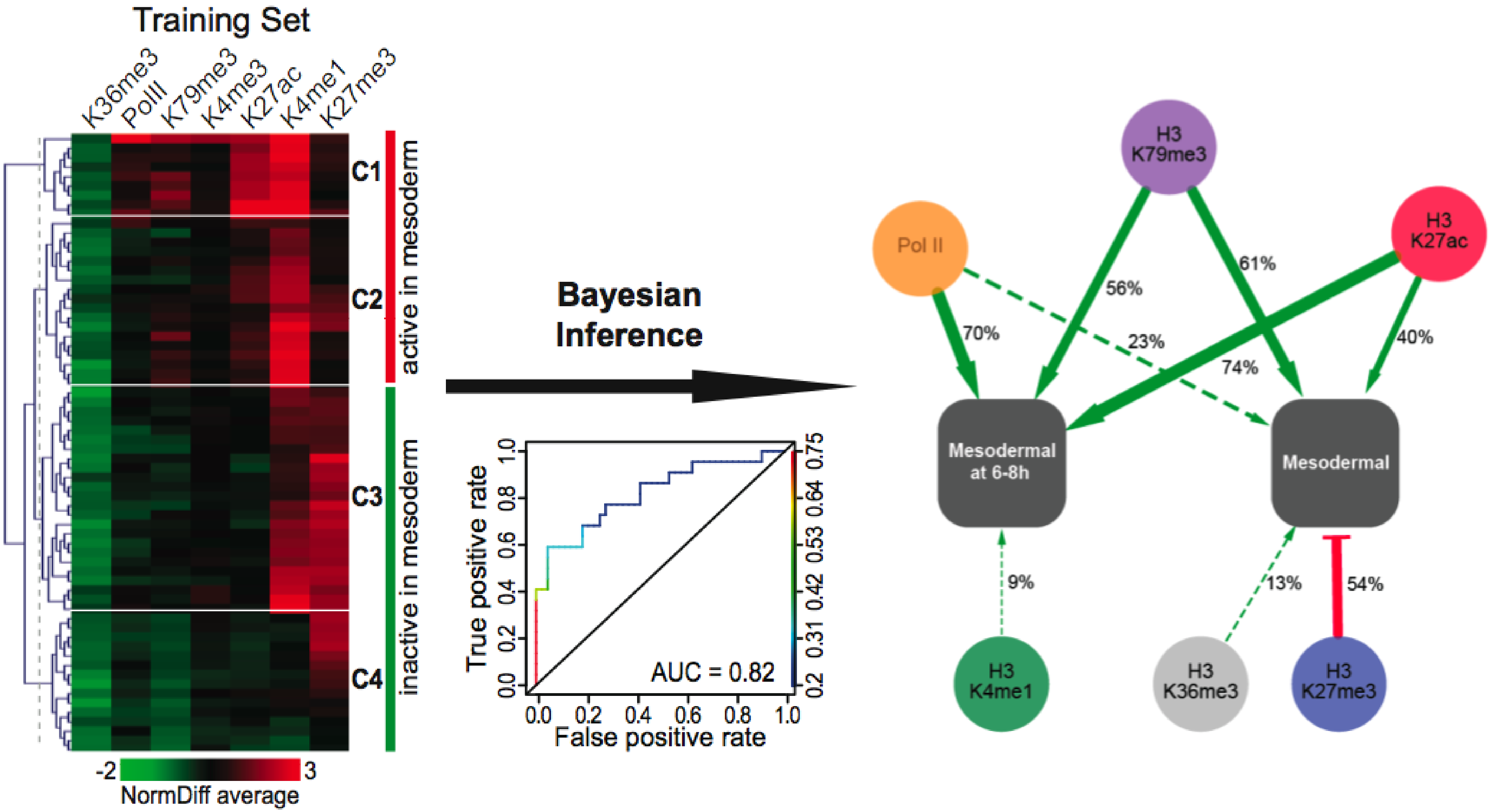


st.	≤5	6	7	8	9	10	11	12	13	14	≥15
m											
o	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
m	🟥	🟥	🟥	🟥	🟥	🟥					
o											
m	ni	ni		🟥	🟥				🟥	🟥	🟥
o	ni	ni									
m			🟥	🟥	🟥						
o					🟦	🟦	🟦	🟦	🟦	🟦	🟦
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
m							🟥	🟥	🟥	🟥	🟥
o											

	pre 6-8h	at 6-8h	post 6-8h
m	48	22	27
o	48	36	41

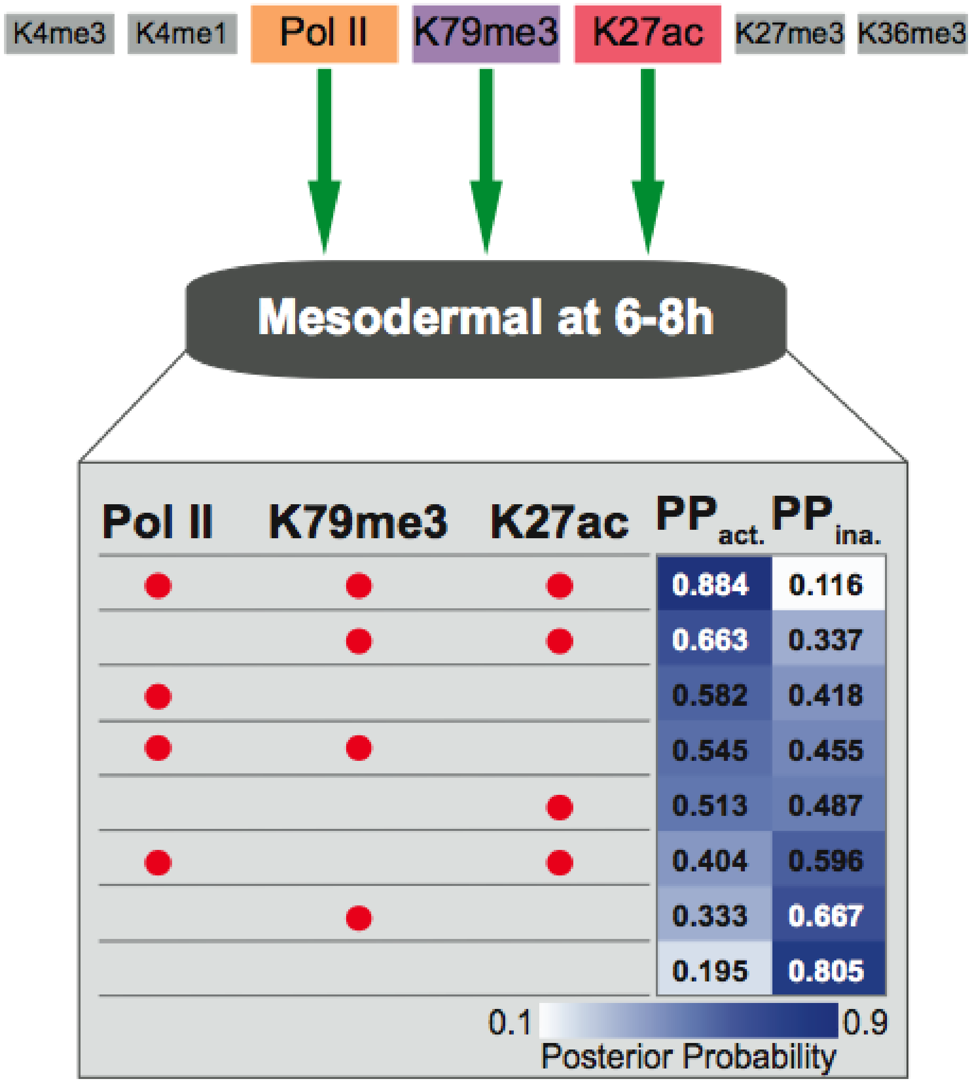
Bonn et al. Nat. Genet, 201

Can we model the relationship between histone marks and activity?

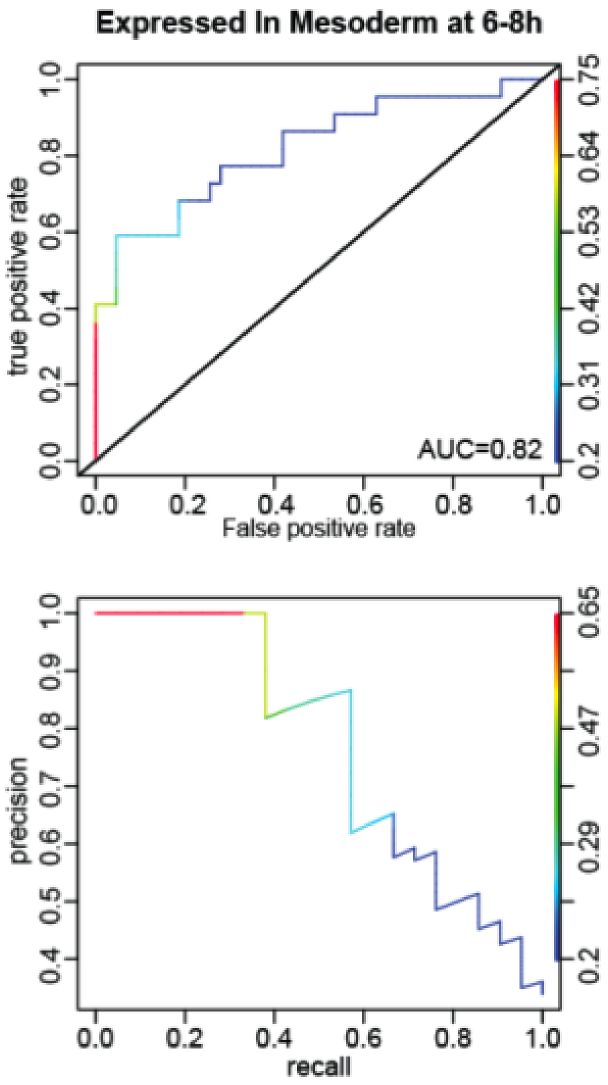


Bonn et al. Nat. Genet, 201

How does the network work?

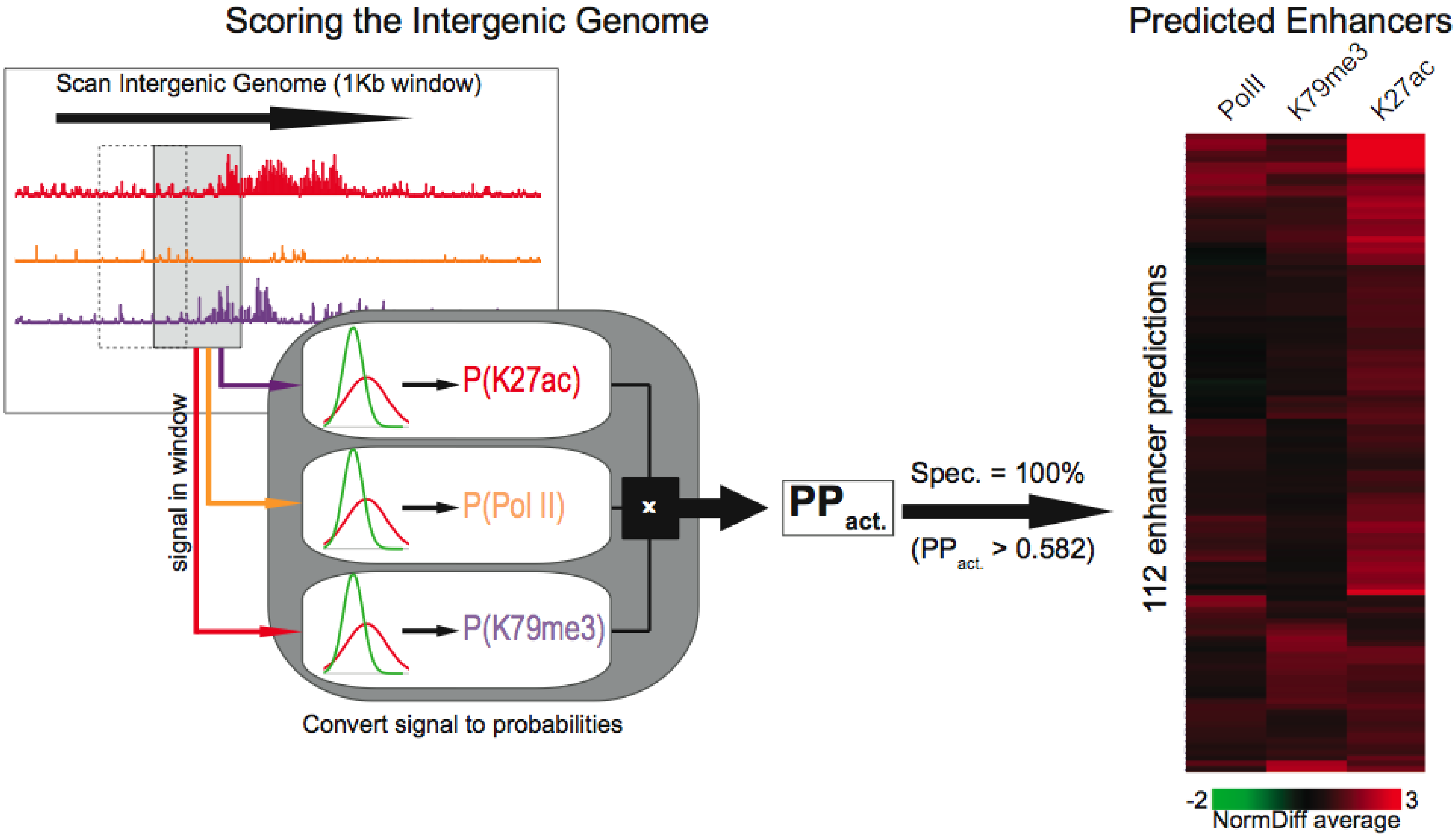


Pol II	K79me3	K27ac	$PP_{act.}$	$PP_{ina.}$
●	●	●	0.884	0.116
	●	●	0.663	0.337
●			0.582	0.418
●	●		0.545	0.455
		●	0.513	0.487
●		●	0.404	0.596
	●		0.333	0.667
			0.195	0.805



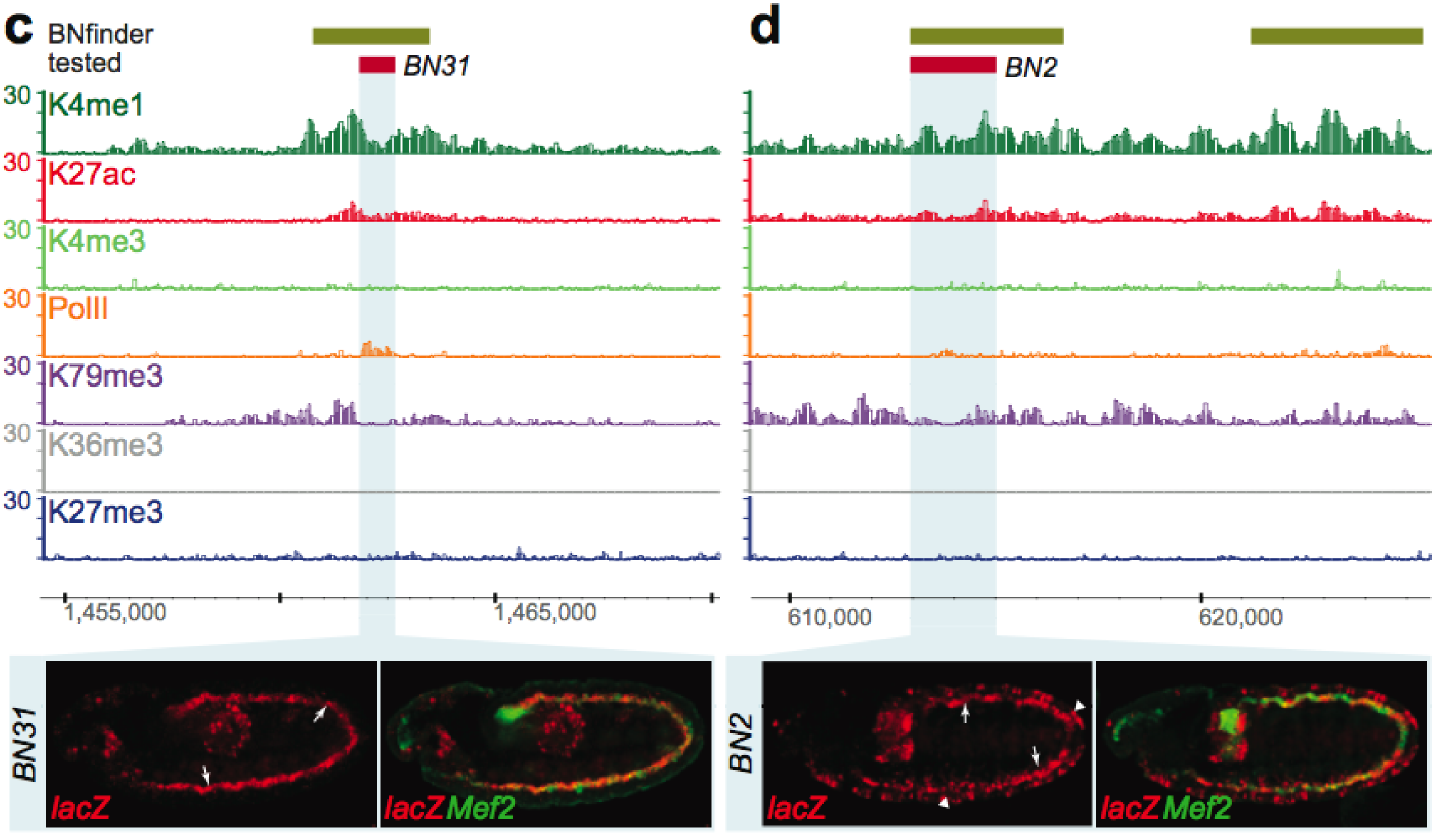
Bonn et al. Nat. Genet, 201

How can it be useful?



Bonn et al. Nat. Genet, 201

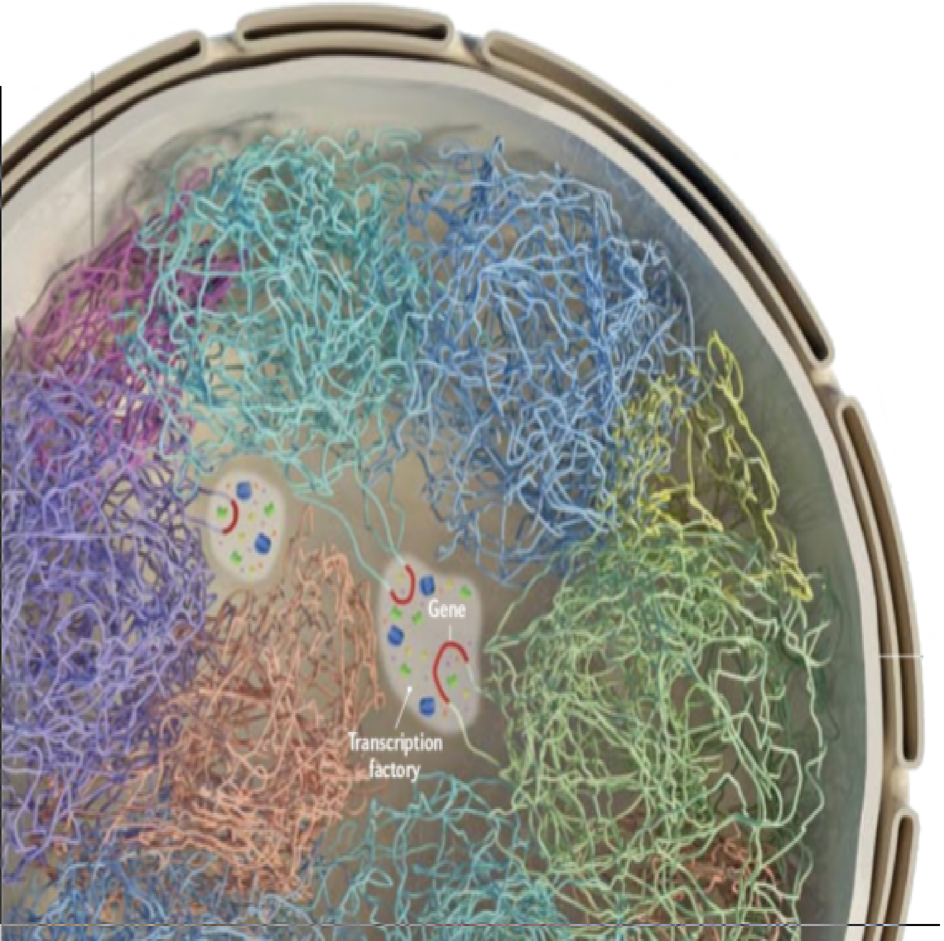
Does it work?



- 12 positive and 4 negative predictions tested
- >90% success!

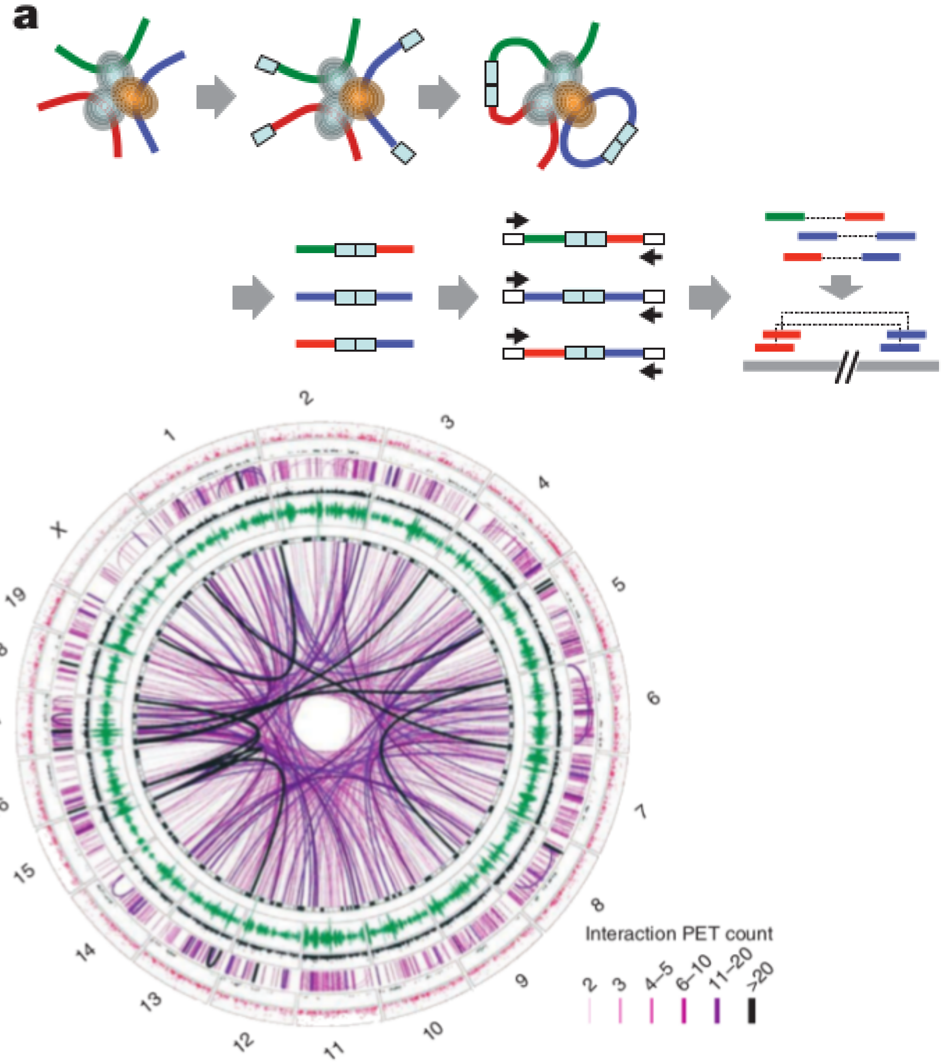
Bonn et al. Nat. Genet, 201

New projects: Insulators and their interactions



- Chromosomes are packed in the nucleus in a non-random fashion
- Chromosomal interactions are mediated by proteins, including insulators

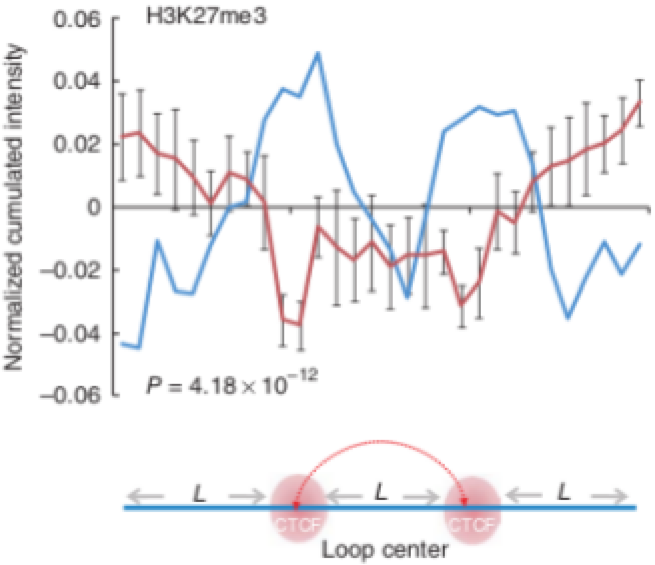
Playing with chromosome interaction data



- Hi-C and ChIA-Pet protocols allow for measurement of chromatin interactions
- Getting this type of data is still difficult
- We can use computational methods to predict interactions

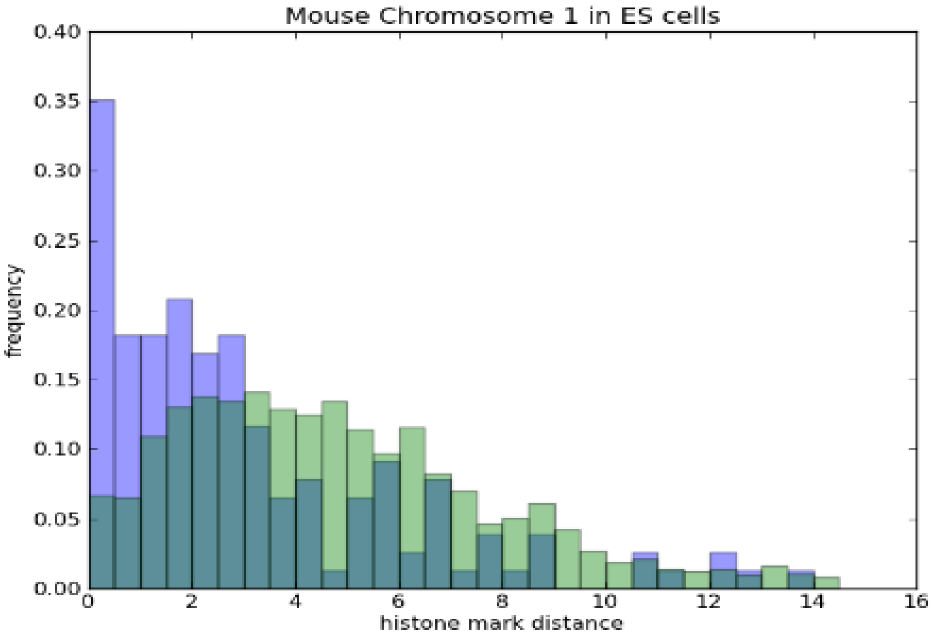
Data from Handoko et al. 2011

Using histone modification similarity to find likely pairs



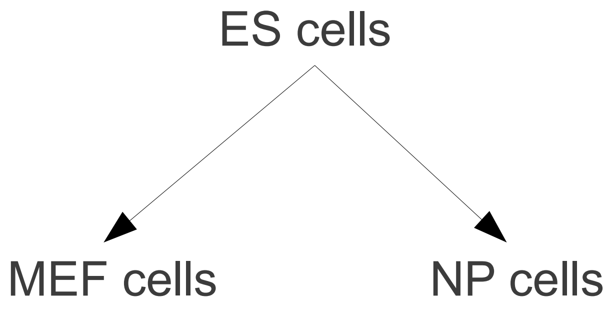
If we define a reasonable measure, we can see this signal on a genome-wide scale to be significant

Histone modification profiles match between interacting anchor points



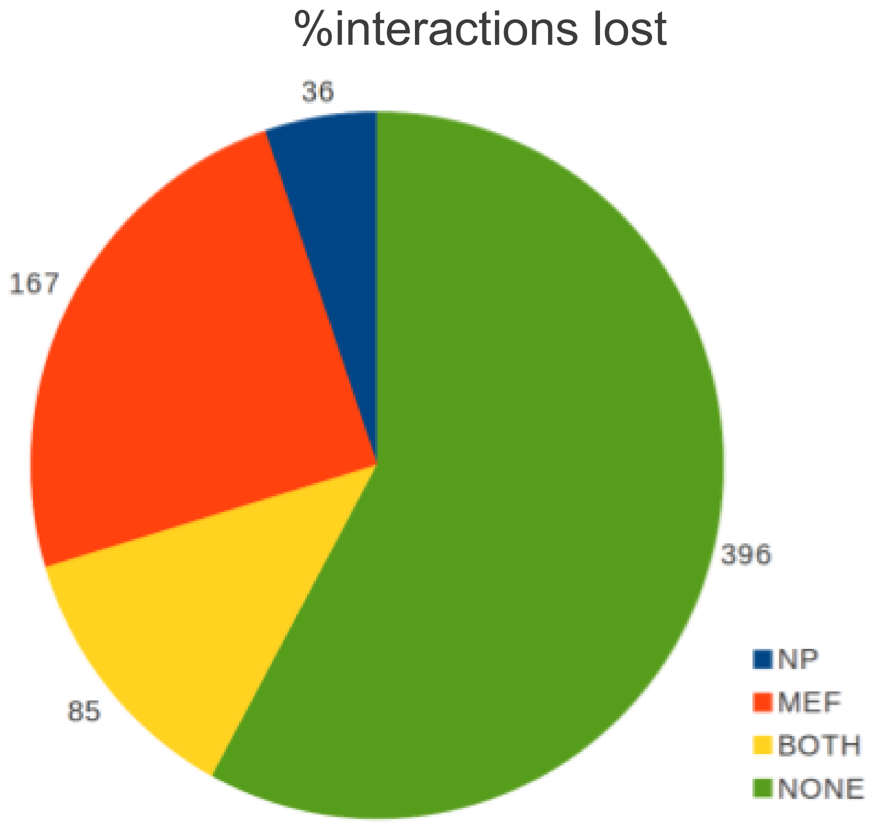
Making predictions

- We can use MCMC algorithm to search the (HUGE!) space of possible interaction ensembles (not real 3D conformations)
- We can get accuracy of ~50% on a small set of interactions
- That is significant, but not quite satisfactory
- Predicting interaction dynamics is feasible

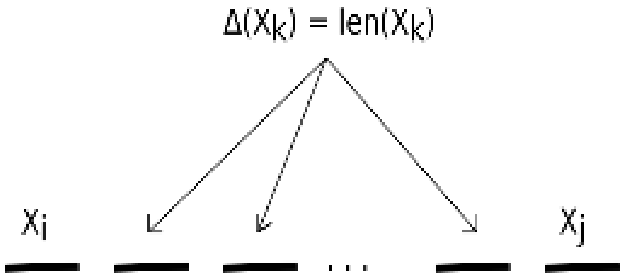

ES cells

MEF cells

NP cells

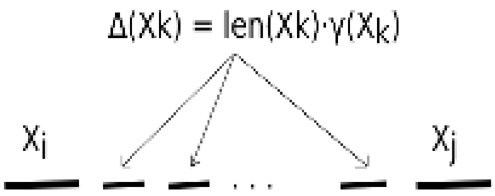


Basic Model

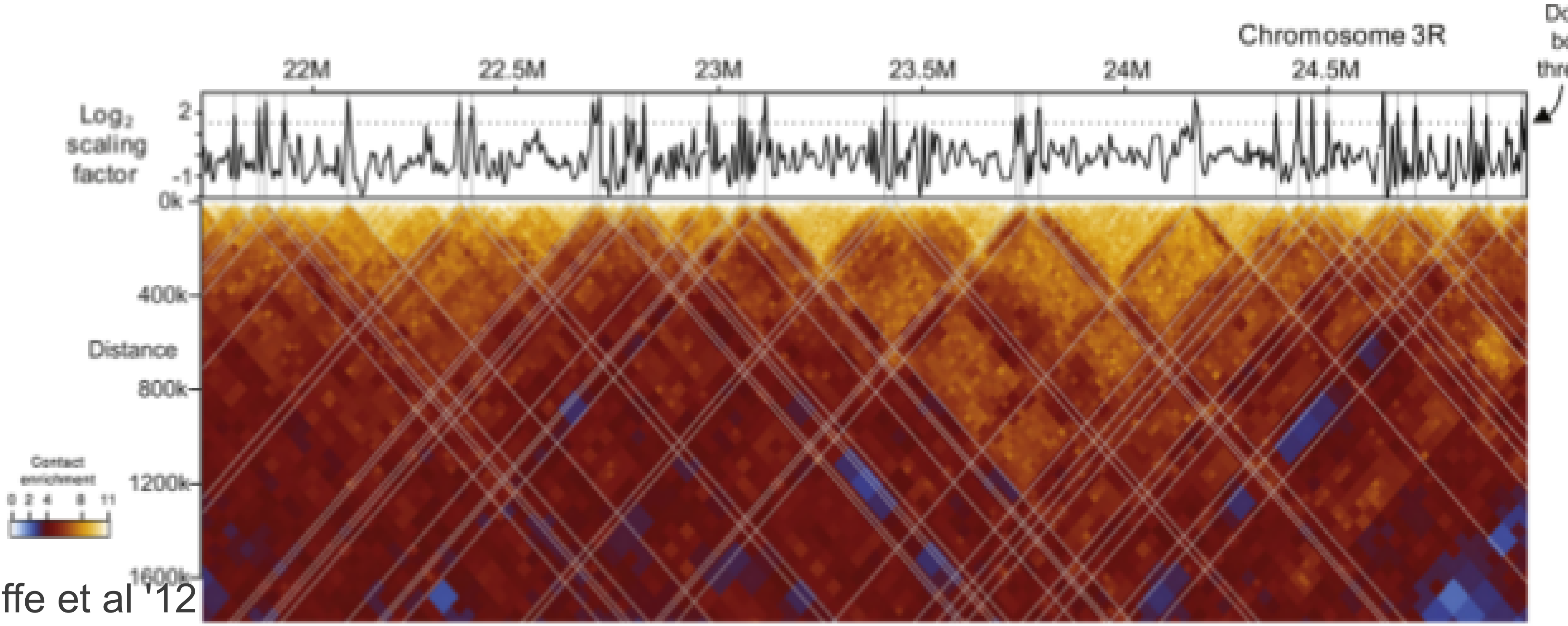


$$P(contact(X,Y)) = f(\sum_{i<k<j} len(X_k))$$

Distance scaling model

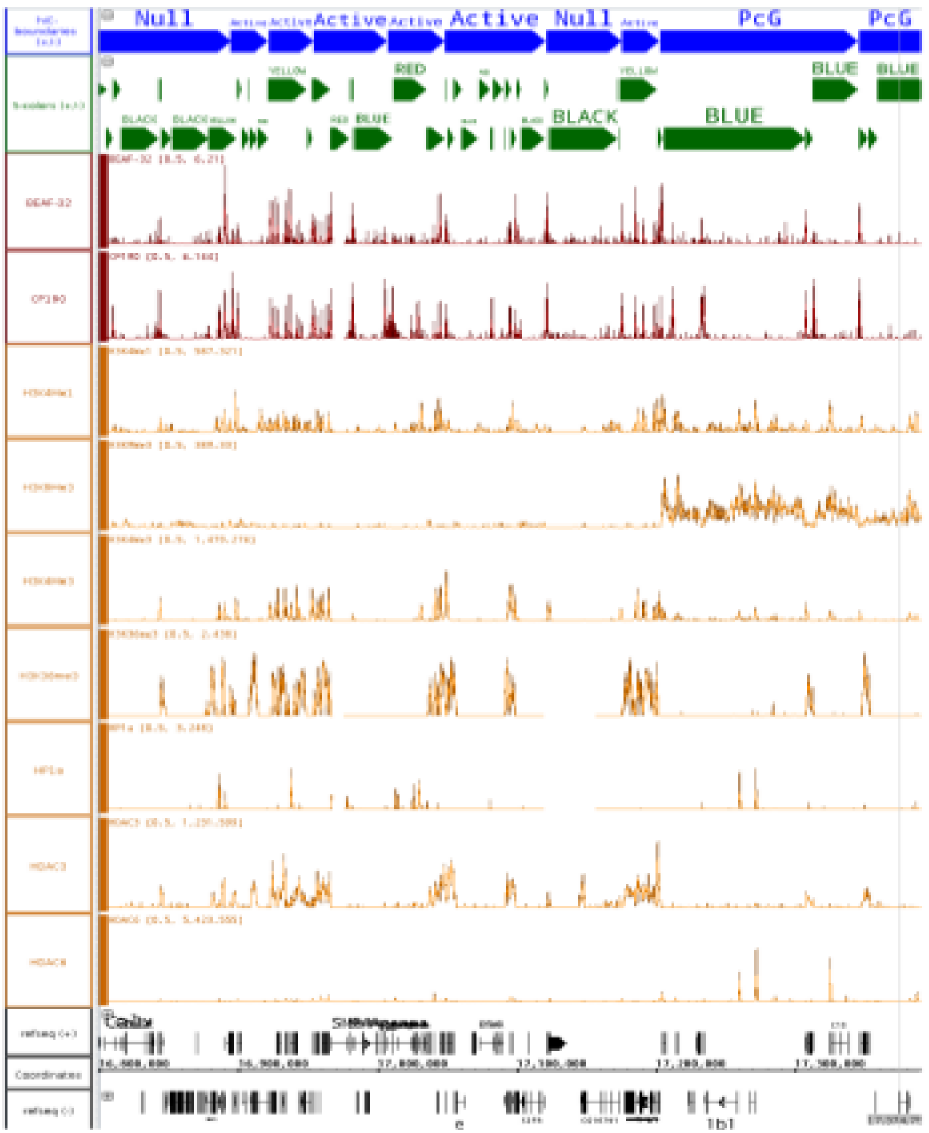


$$P(contact(X,Y)) = f(\sum_{i<k<j} len(X_k) \cdot \gamma_k)$$



Yaffe et al '12

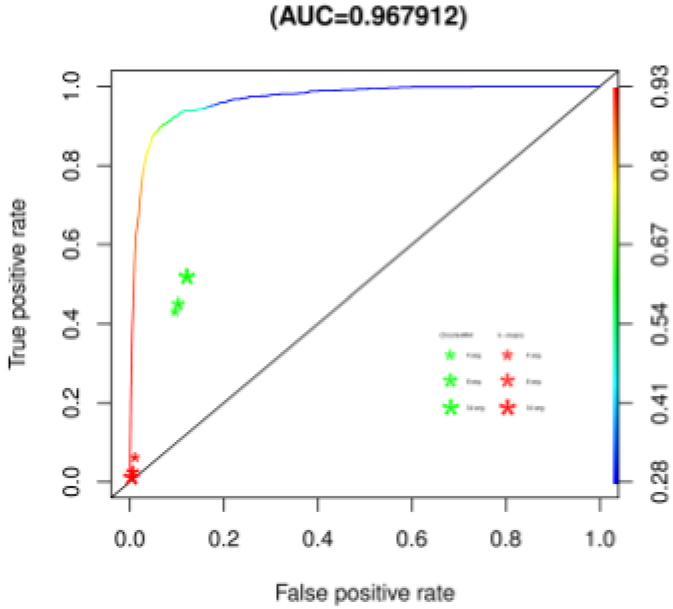
Predicting boundary elements from modEncode data

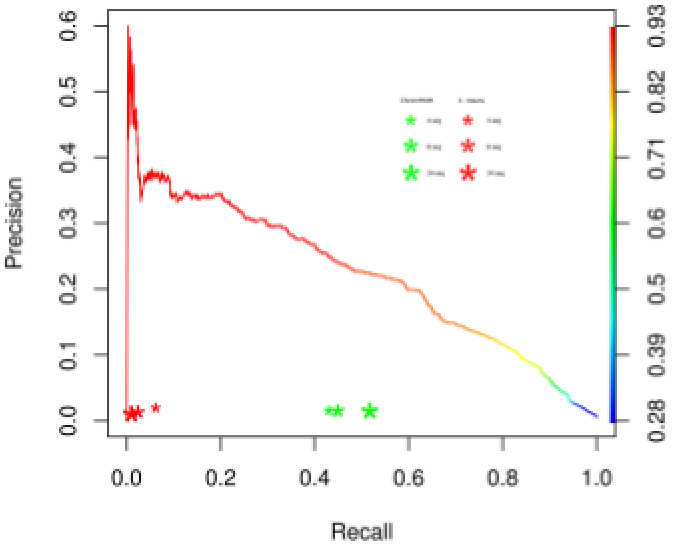


- We can use all chromatin IP data available in modENCODE for late embryos and try to predict domain boundaries
- We will use the Hi-C derived data as our main interaction set and Dam-ID derived data as additional validation

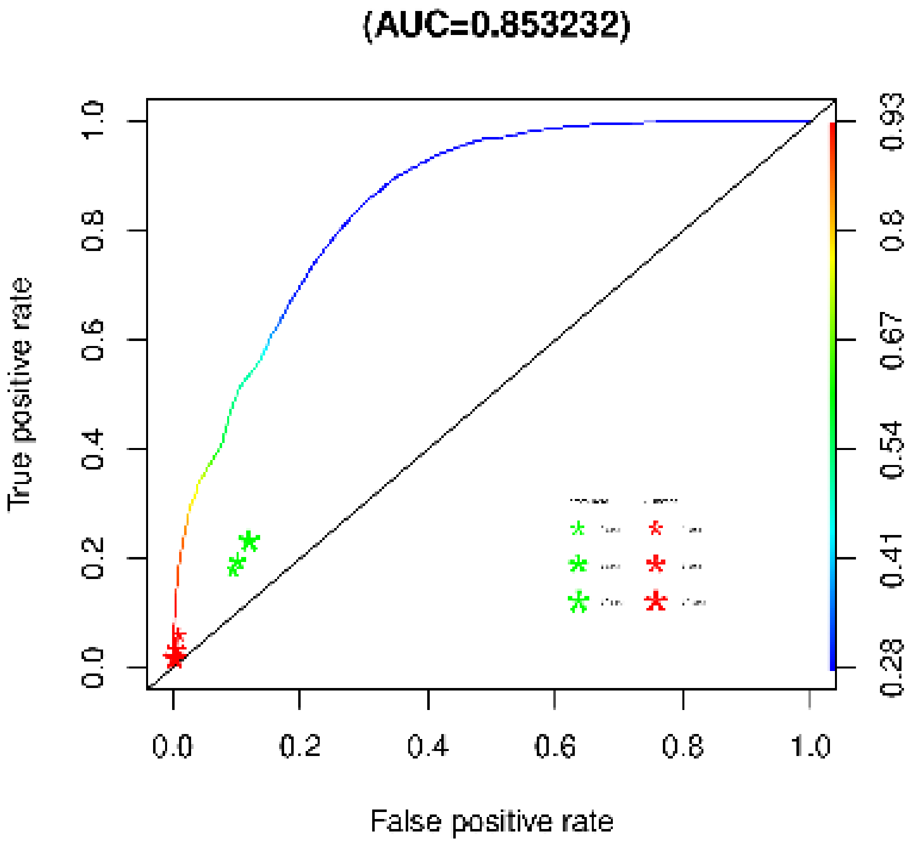
BN classifier can predict boundaries

- Using BN classifiers trained on modENCODE data, we can predict position of boundary elements
- This method outperforms HMMs and clustering of histone modification data

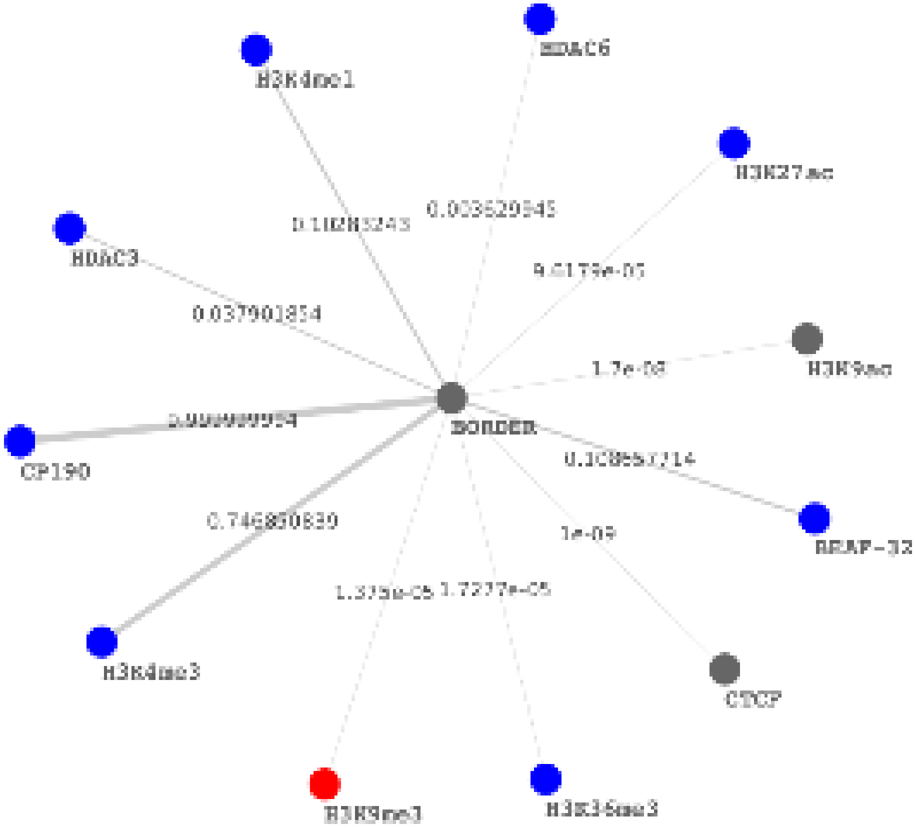




Predictions make sense, and the model brings new information

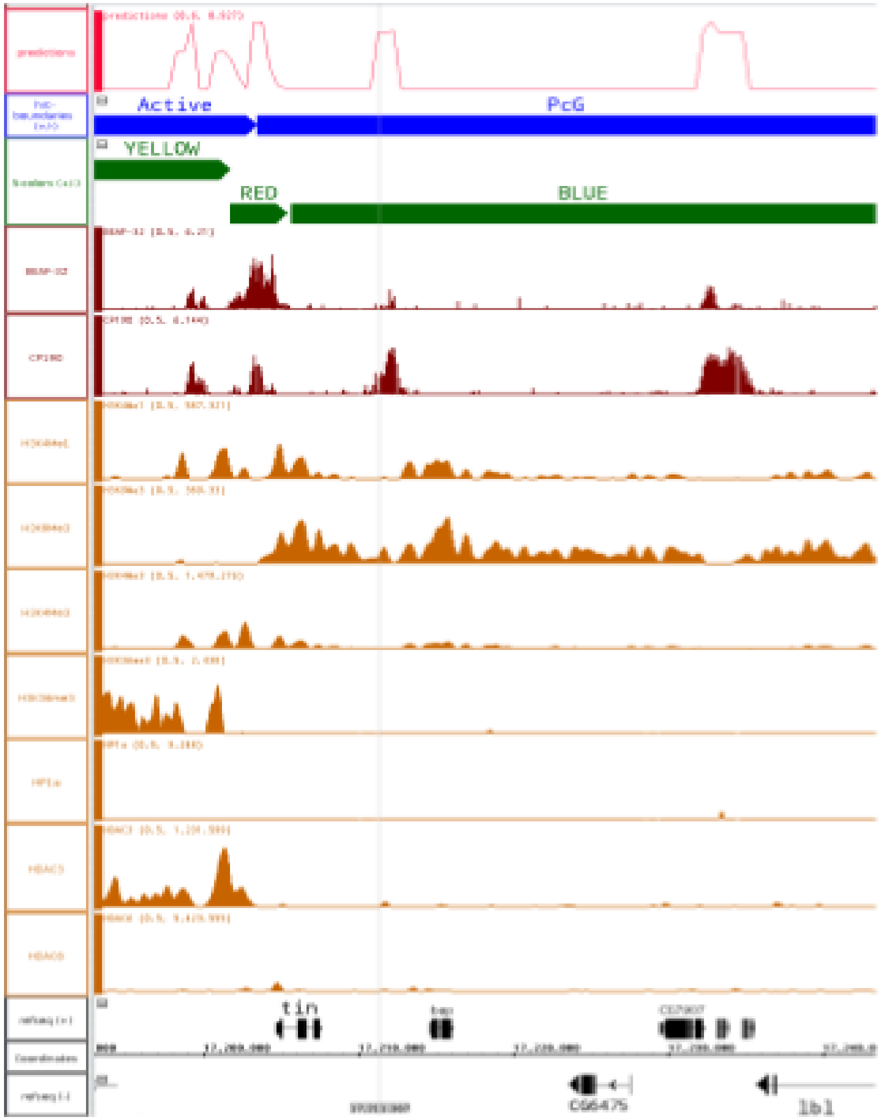


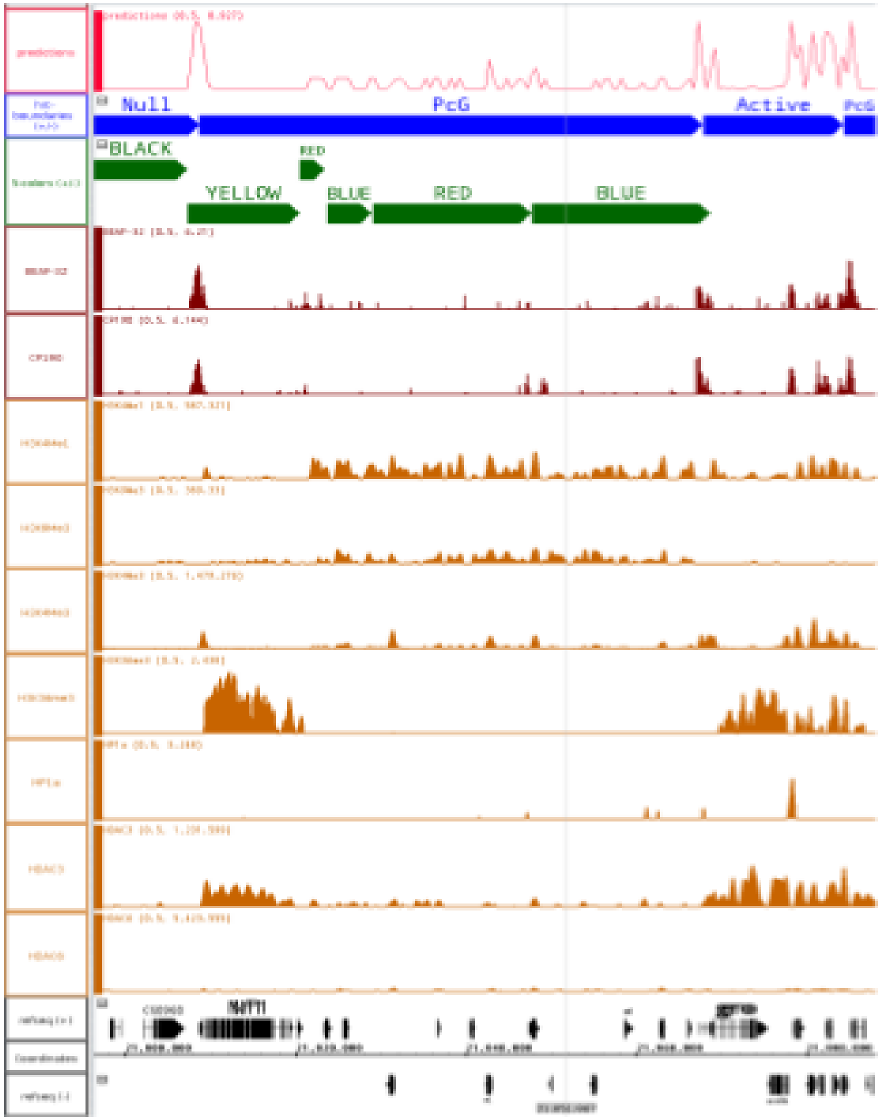
DamID validation results



Impact of signals

Some predictions are unexpected





Acknowledgments

- **Eileen Furlong**
- Ya-Hsin Liu
- Lucia Ciglar
- Zhen-Xuan Yeo

- Charles Girardot
- Robert Zinzen
- Stefan Bonn

Łukasz Bieniasz-Krzywiec

Paweł Bednarz

Post-doc positions available

INNOVATIVE ECONOMY
NATIONAL COHESION STRATEGY

FNP
Foundation for Polish Science

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND

EMBO