Klasyfikacja obszarów funkcjonalnych

Wykład dla biotechnologów

Bartek Wilczyński
bartek@mimuw.edu.pl

12.12.2013

Modyfikacje histonów a chromatyna

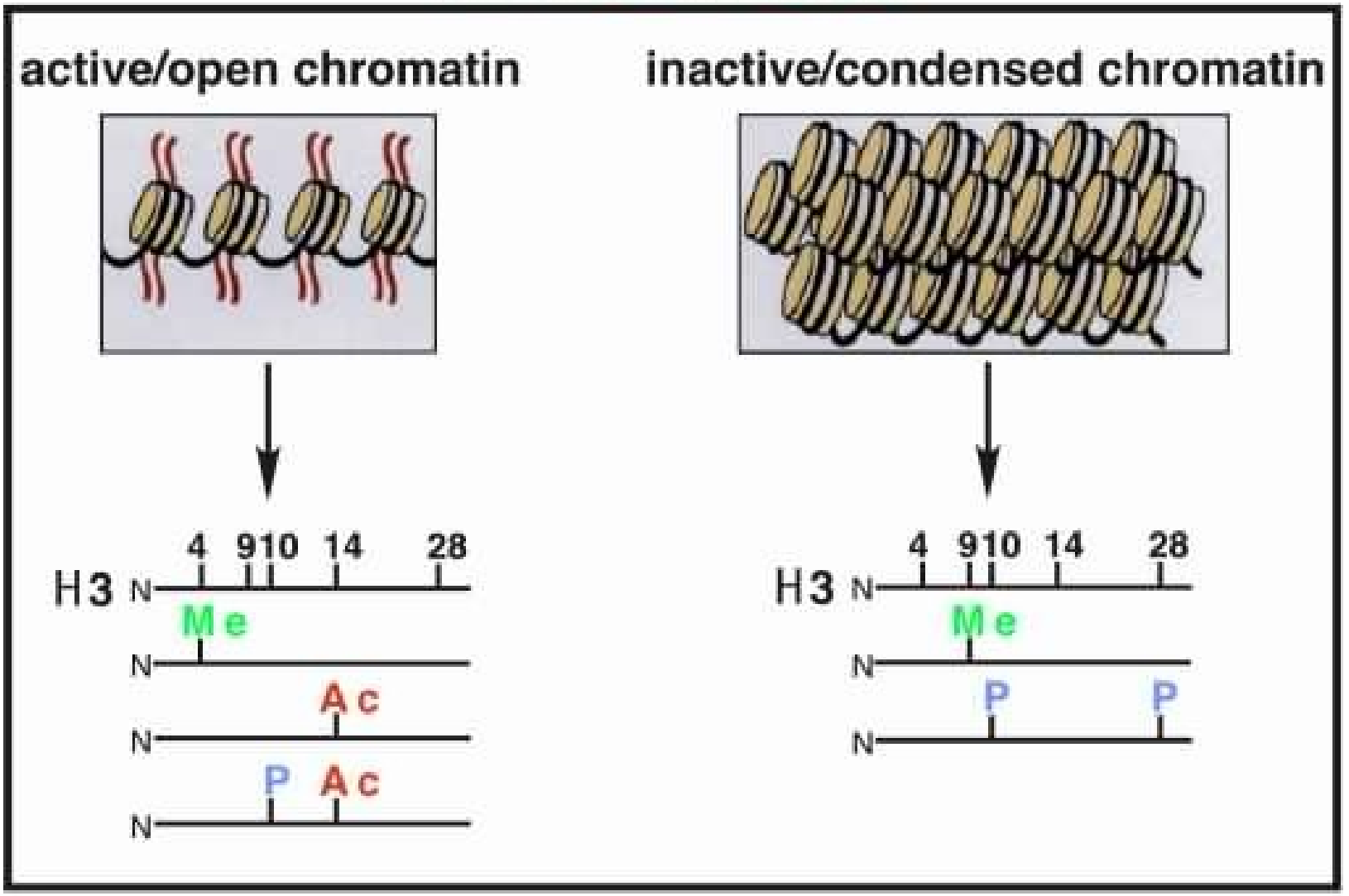
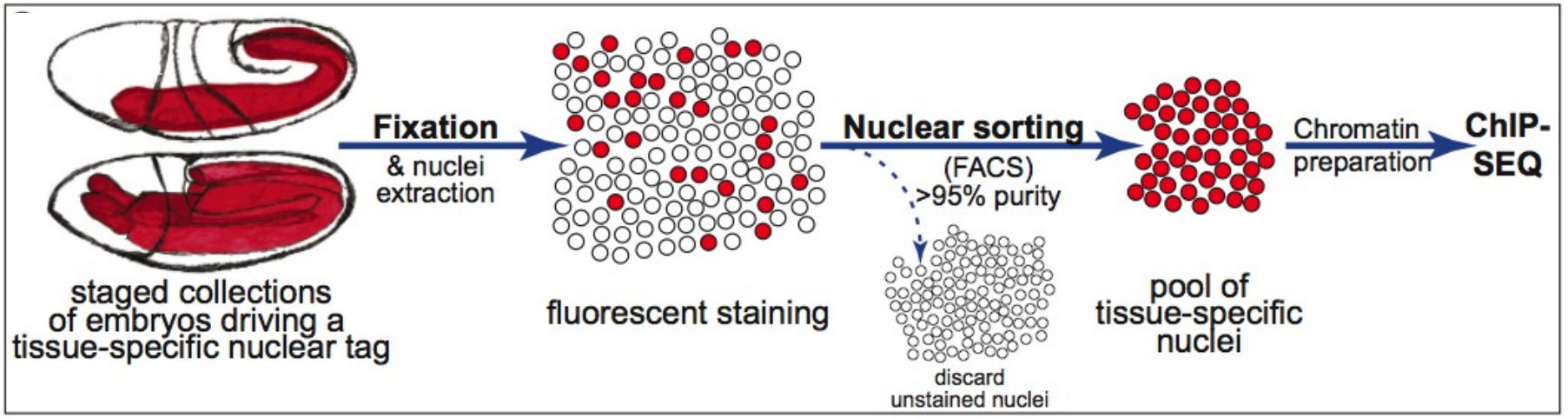
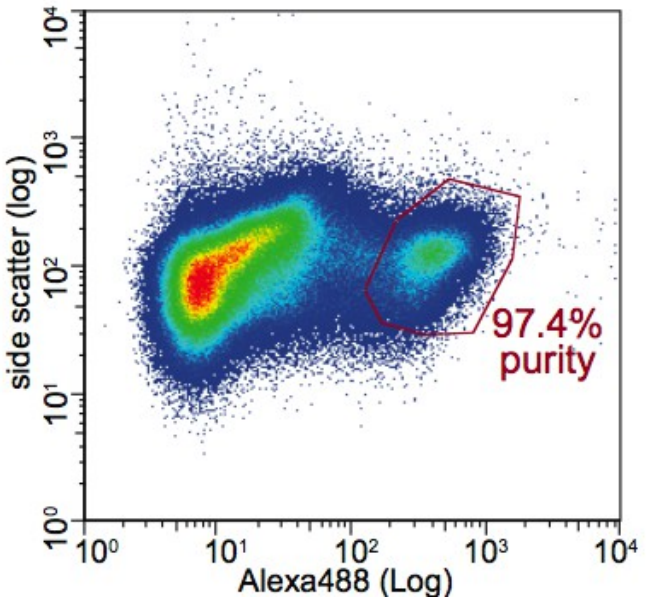
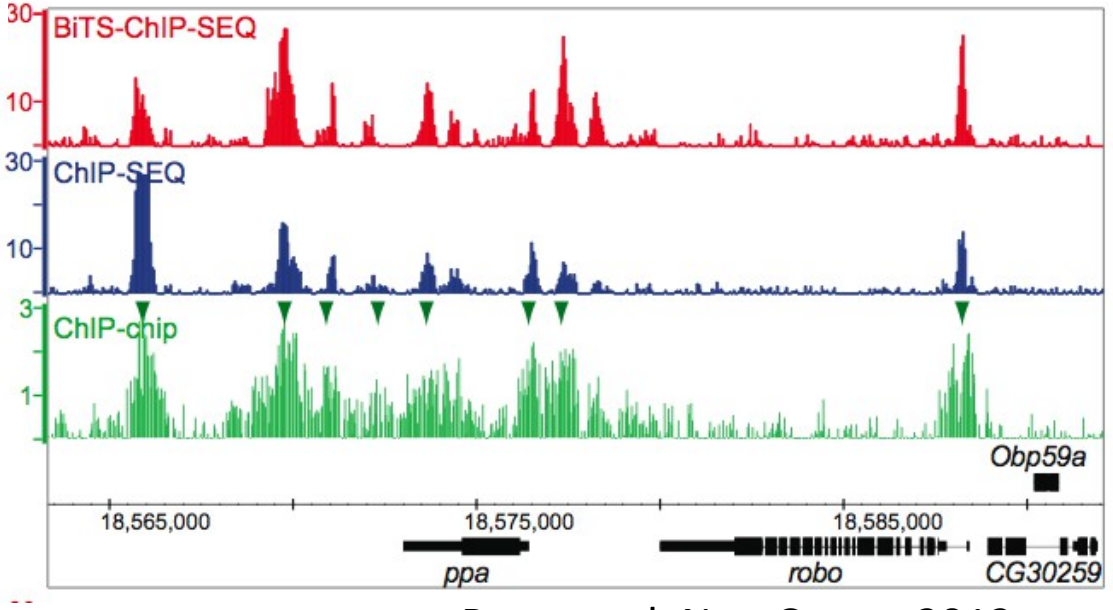


Figure 4

Modyfikacje specyficzne dla tkanki

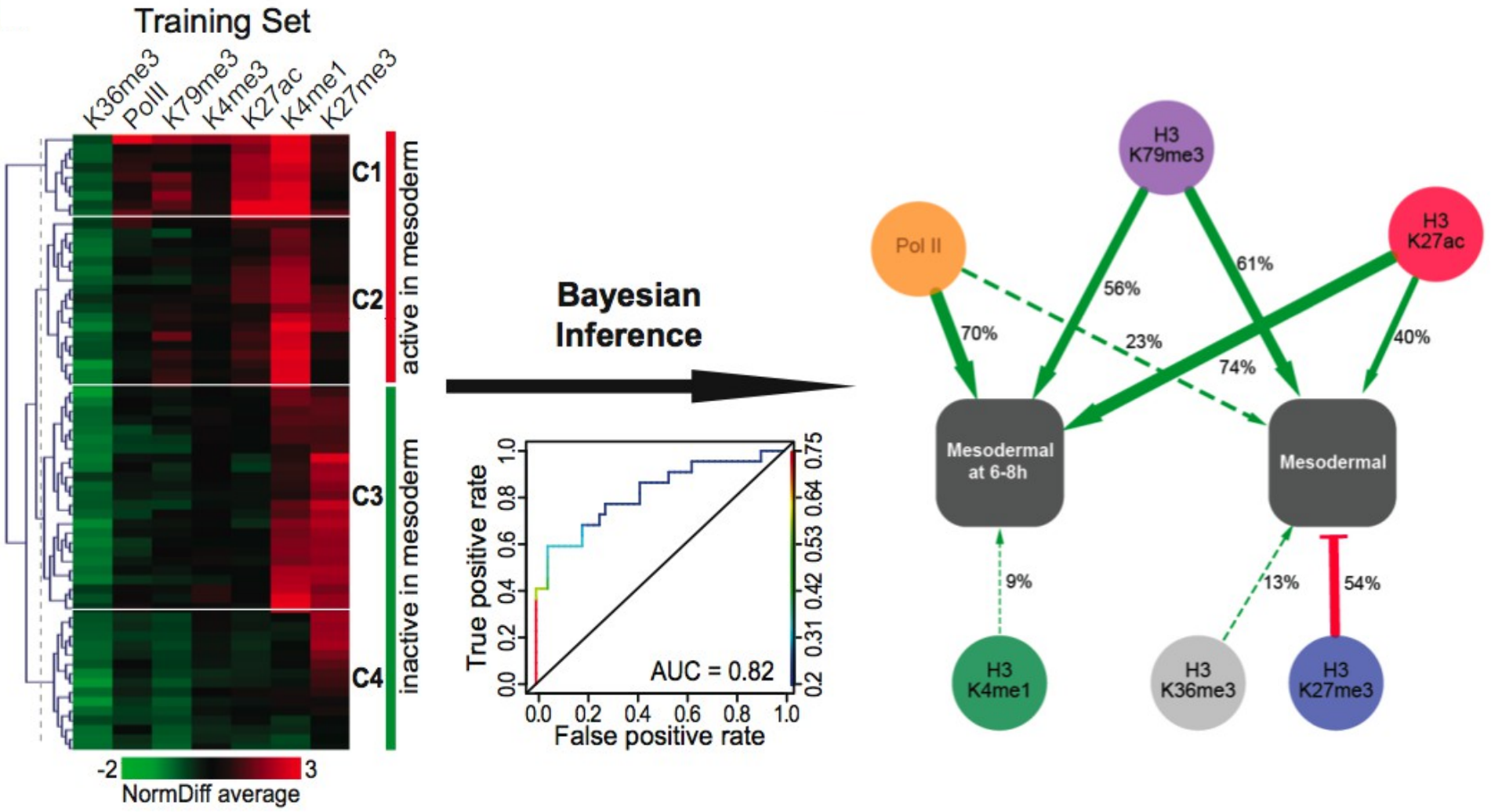






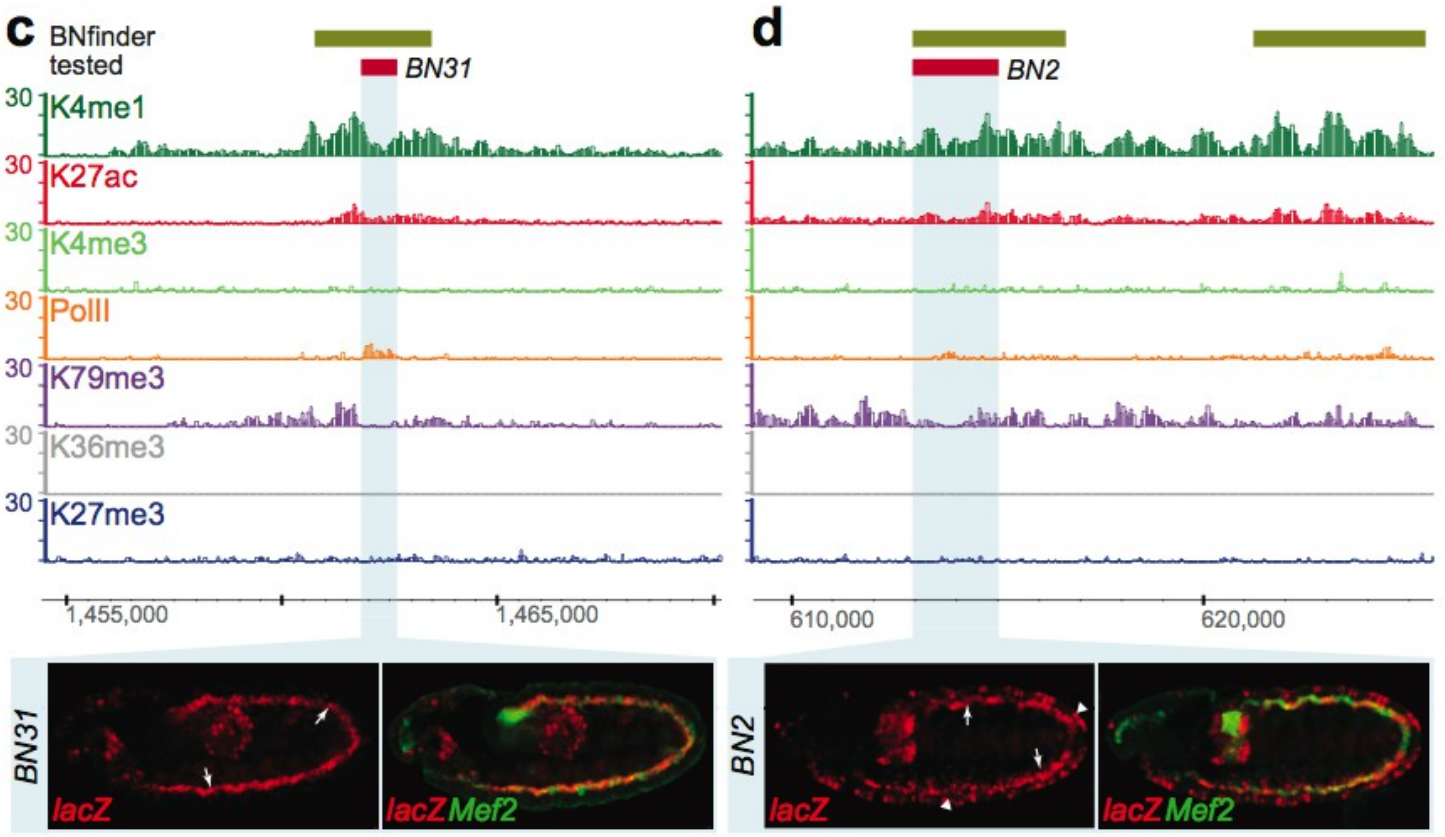
Bonn et al. Nat. Genet, 2012

Stworzenie klasyfikatora



Bonn et al. Nat. Genet, 2012

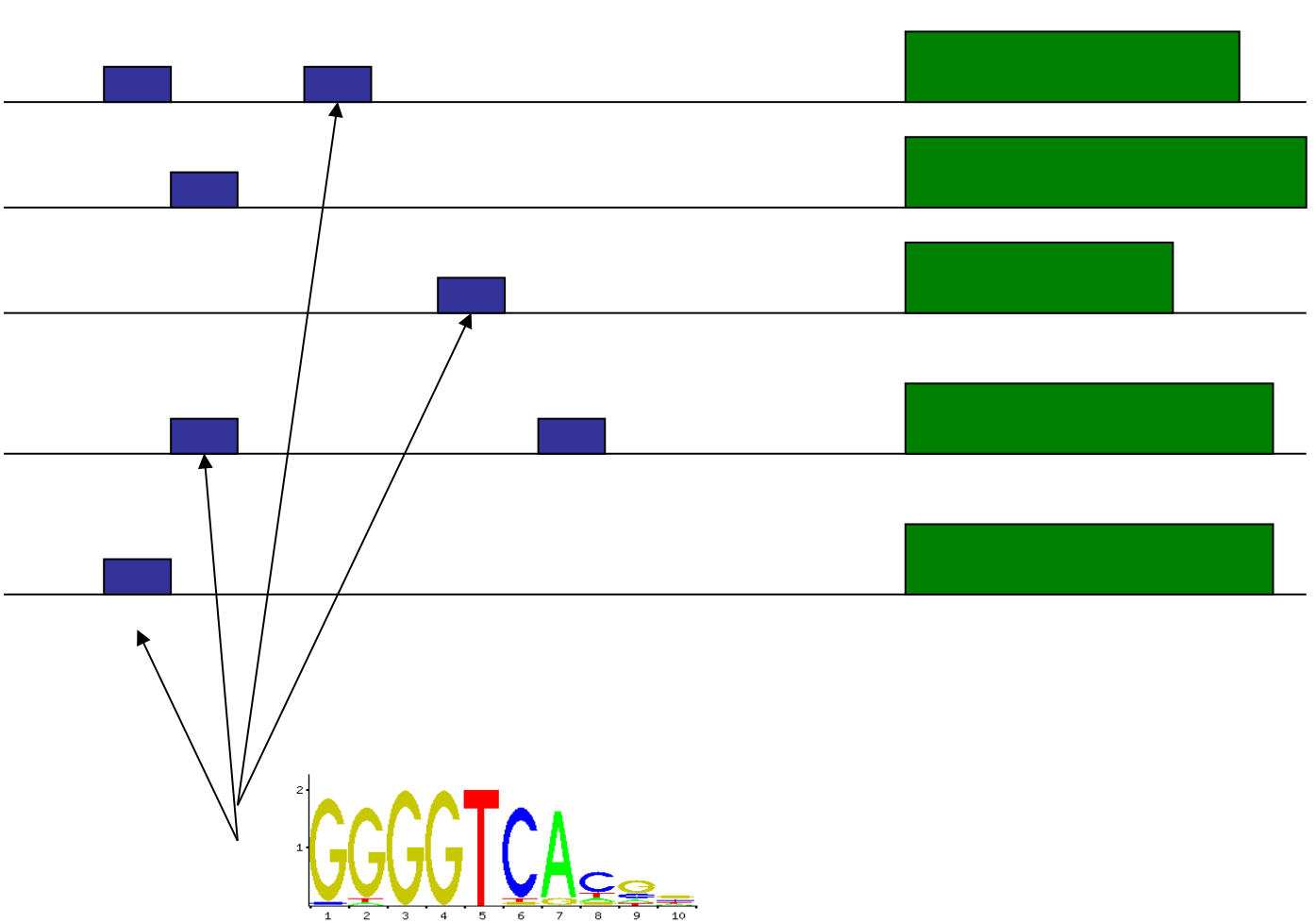
Weryfikacja eksperymentalna



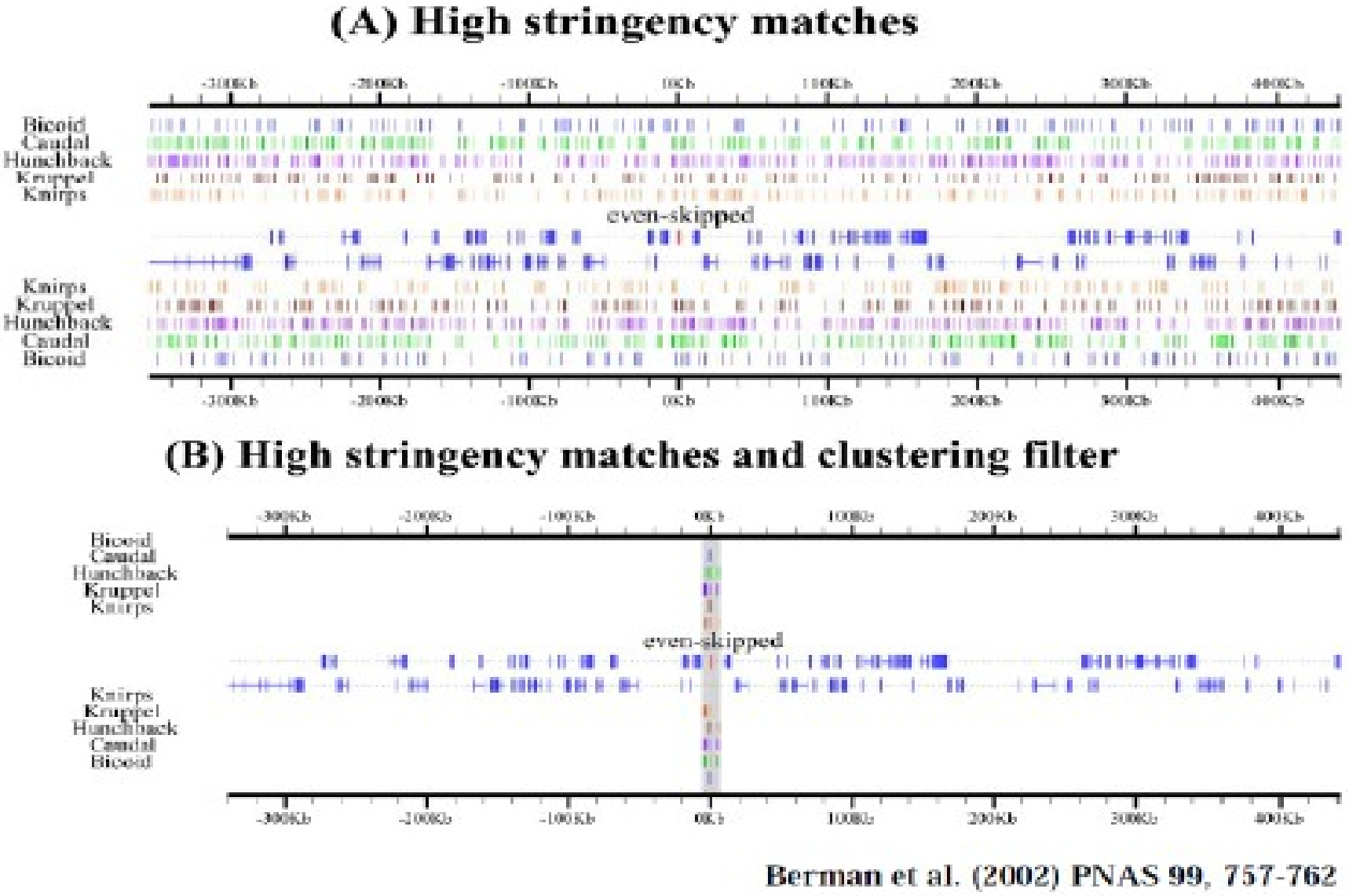
- 12 pozytywnych i 4 negatywne predykcje
- >90% prawidłowo! (1 pozytywna pomyłka)

Bonn et al. Nat. Genet, 2012

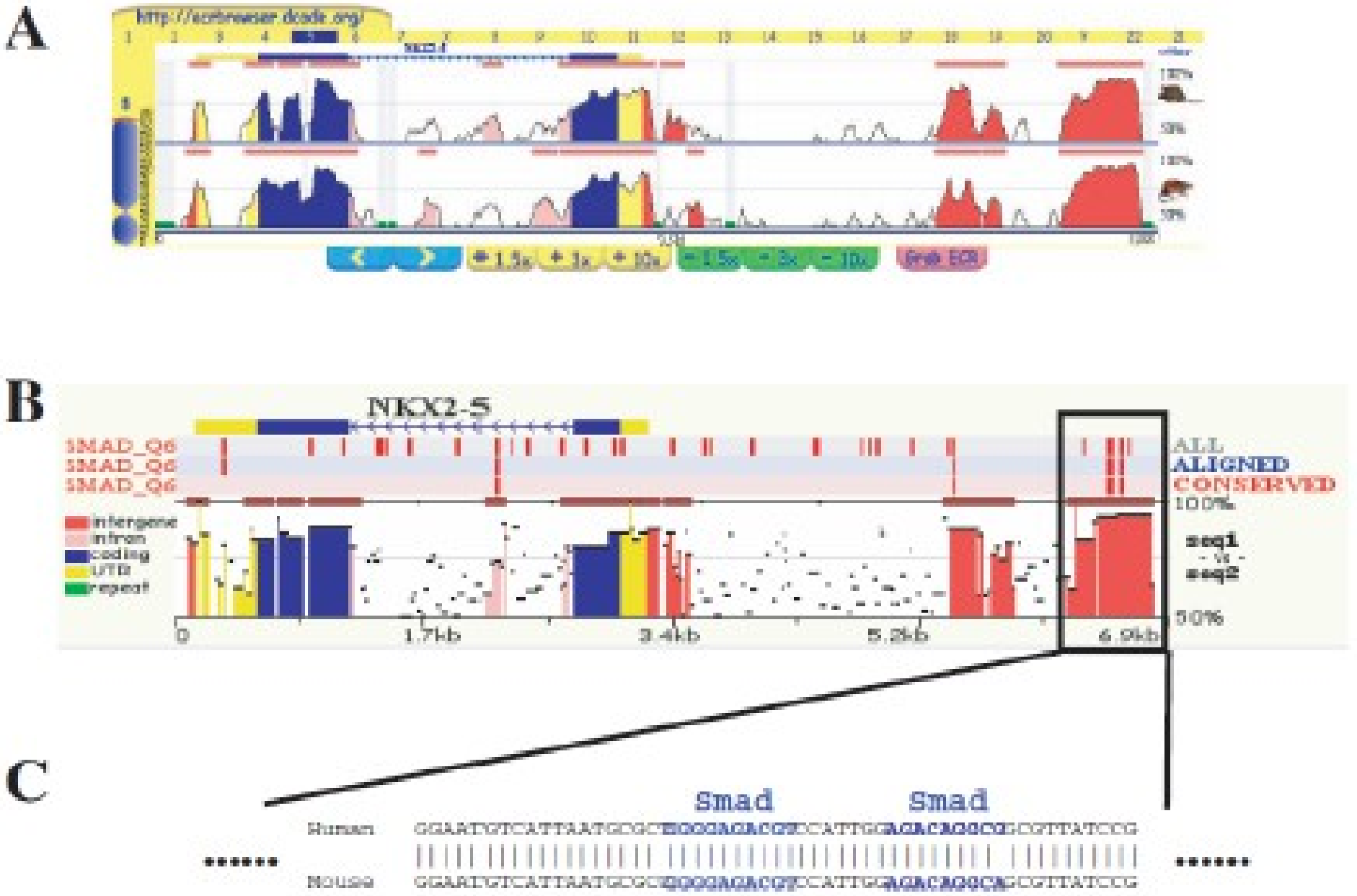
Znajdowanie nowych motywów



Szukanie Enhancerów przy pomocy motywów

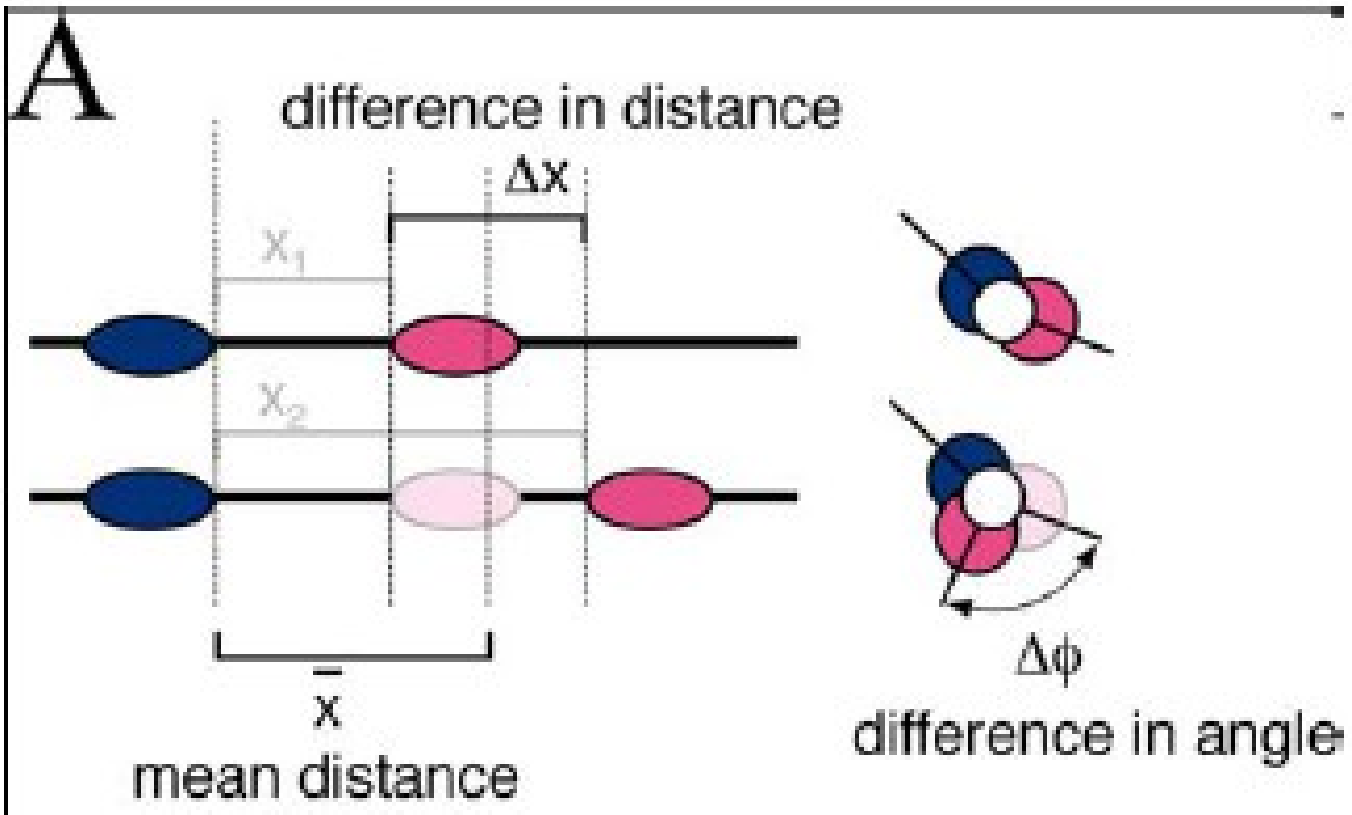


Berman et al. (2002) PNAS 99, 757-762



Loots, Ovcharenko, NAR 2004

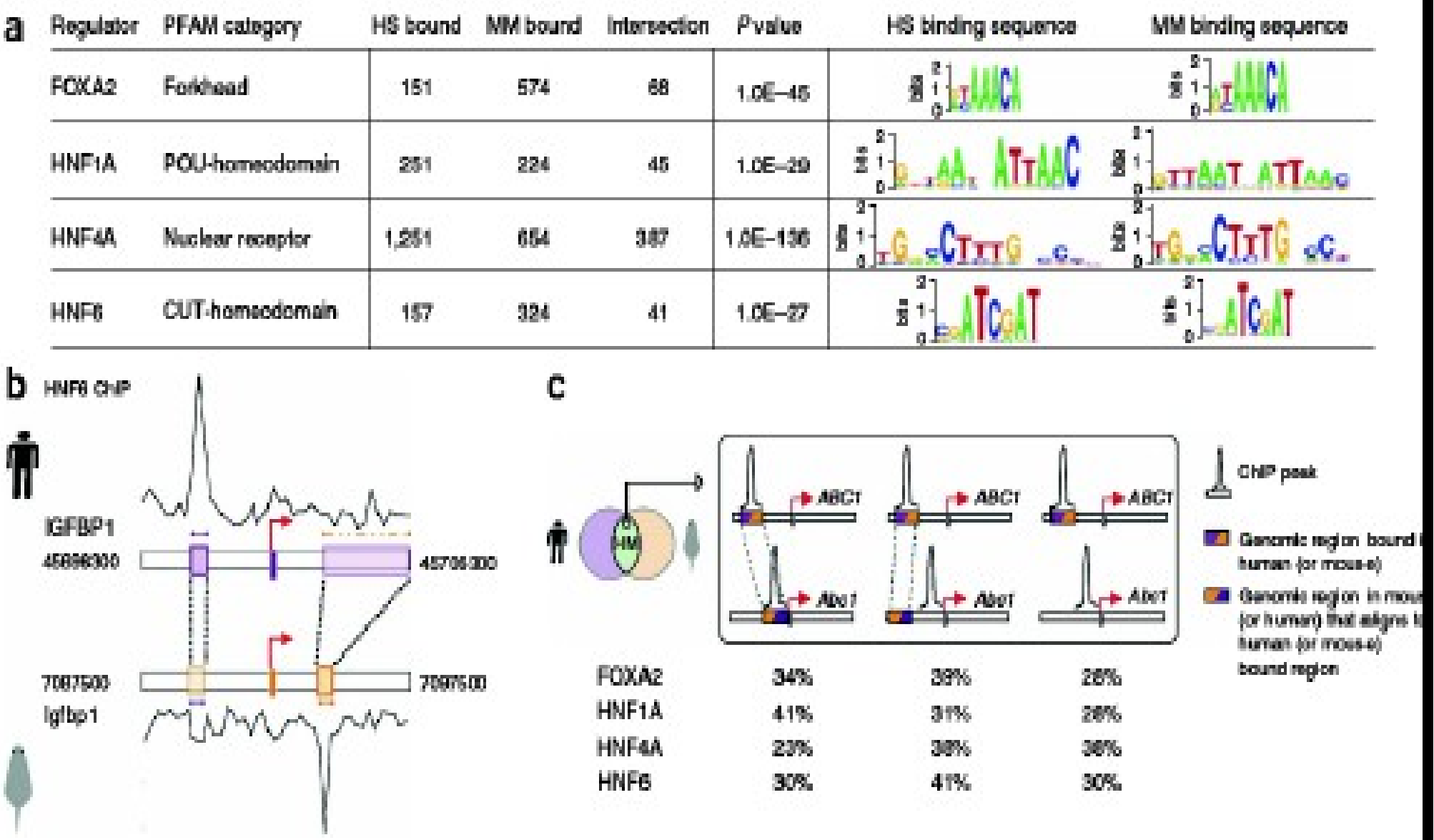
The EEL method *(Hallikas et al. Cell, 2006)* aligns binding sites from promoter regions of homologous genes using a special function:



$$\text{Score} = \lambda \Delta G_T - \mu \bar{x} - \frac{\nu \Delta x^2 + \xi \Delta \phi^2}{2\bar{x}}$$

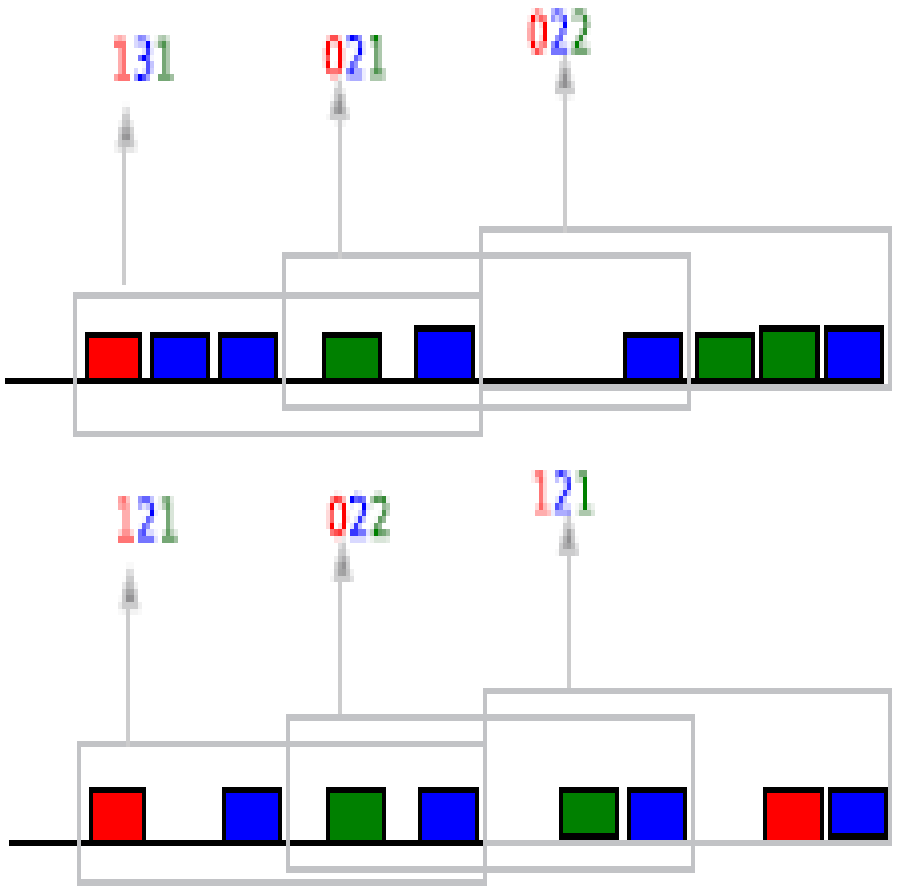
relative weights

affinity clustering conservation



Odom et al. Nature Genetics, 2007, Similar results for Drosophila, Li et al, Genome Biol. 2007

$$S(P,P') = |P \cap P'| - \alpha \cdot |P \otimes P'| - \beta$$



$$P = p_1, p_2, p_3 = 131, 021, 022$$

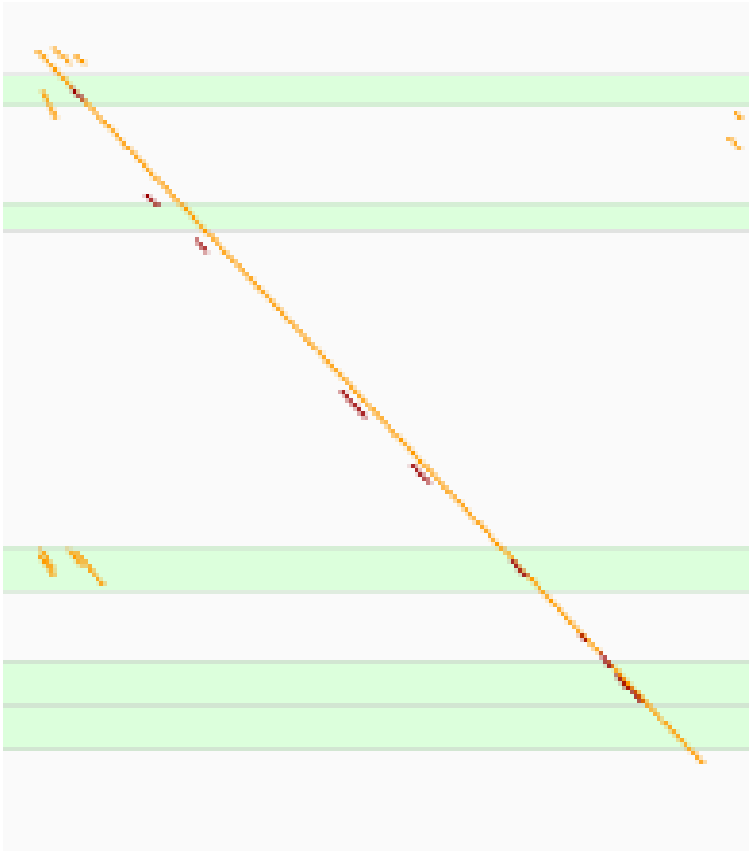
$$P' = p'_1, p'_2, p'_3 = 121, 022, 121$$

$$S(P,P') =$$

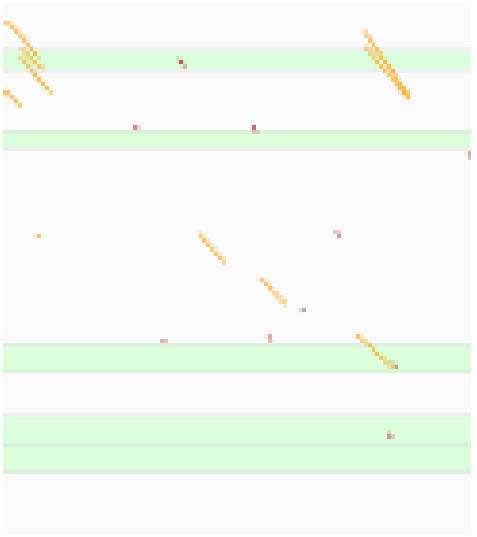
$$S(p_1,p'_1) + S(p_2,p'_2) + S(p_3,p'_3)$$

$$= (4-\alpha-\beta)+(3-\alpha-\beta)+(3-2\alpha-\beta$$

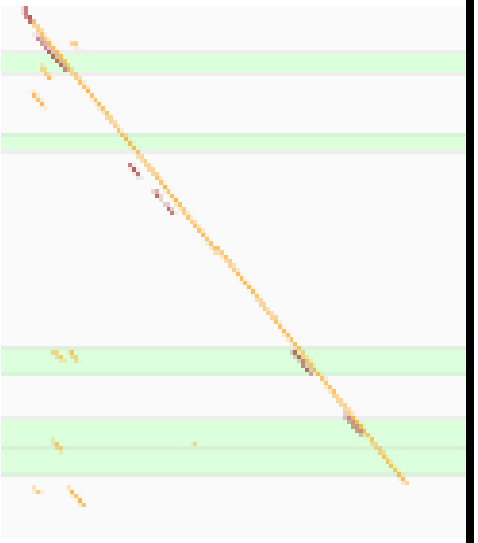
$$= 9 - 4\alpha - 3\beta$$



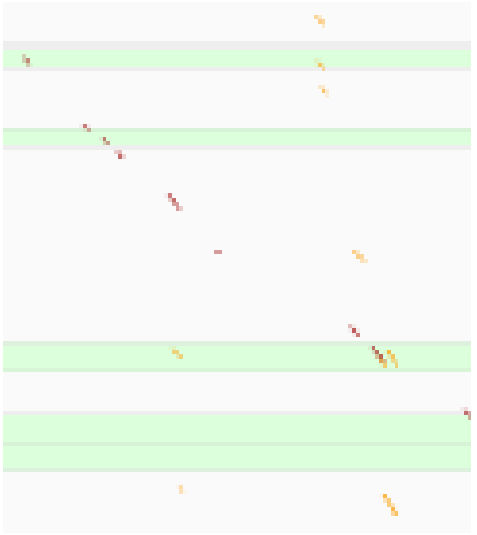
a) *D. pseudoobscura*



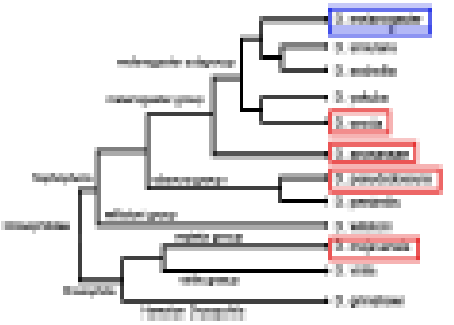
b) *D. erecta*



c) *D. ananassae*

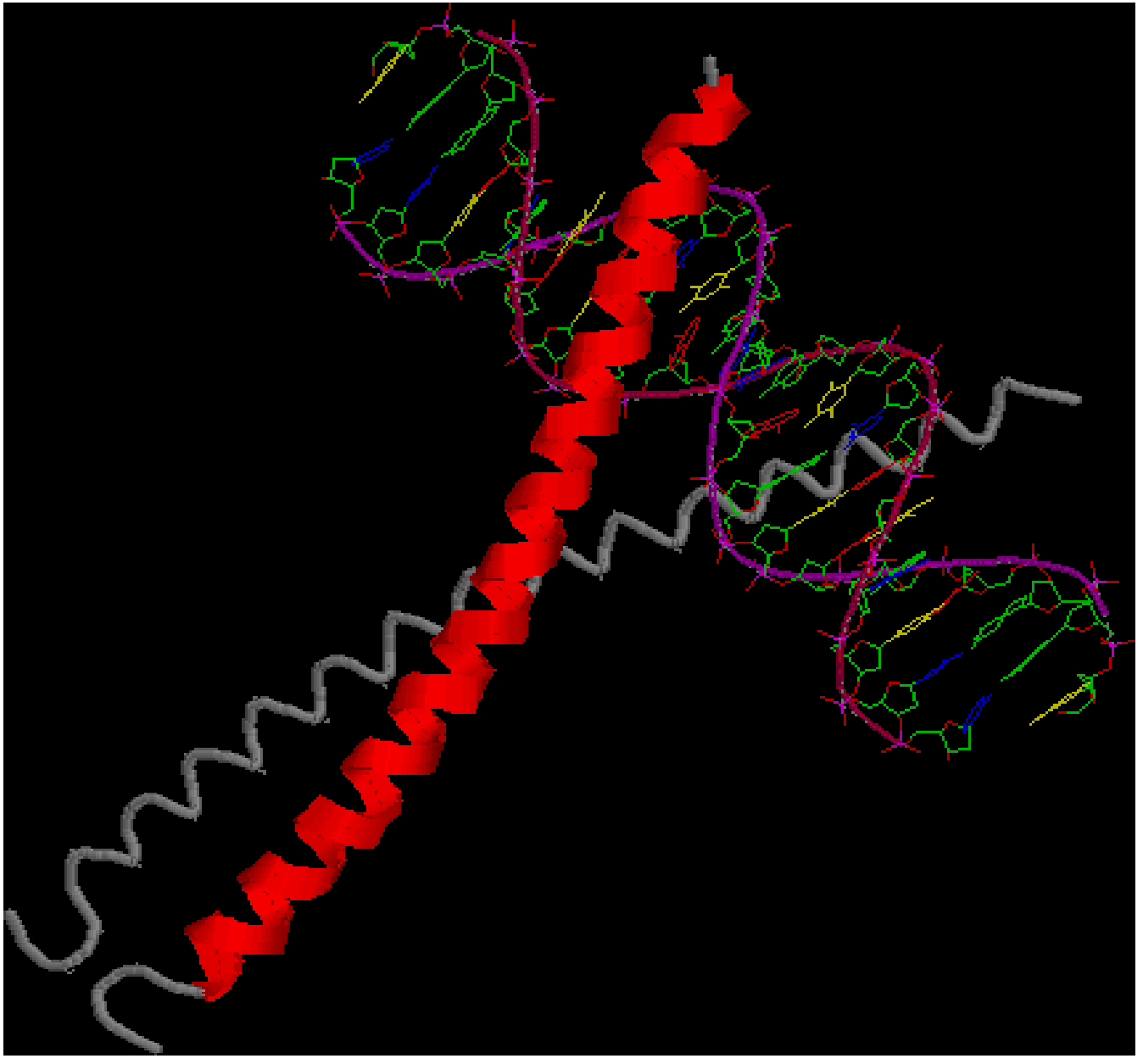


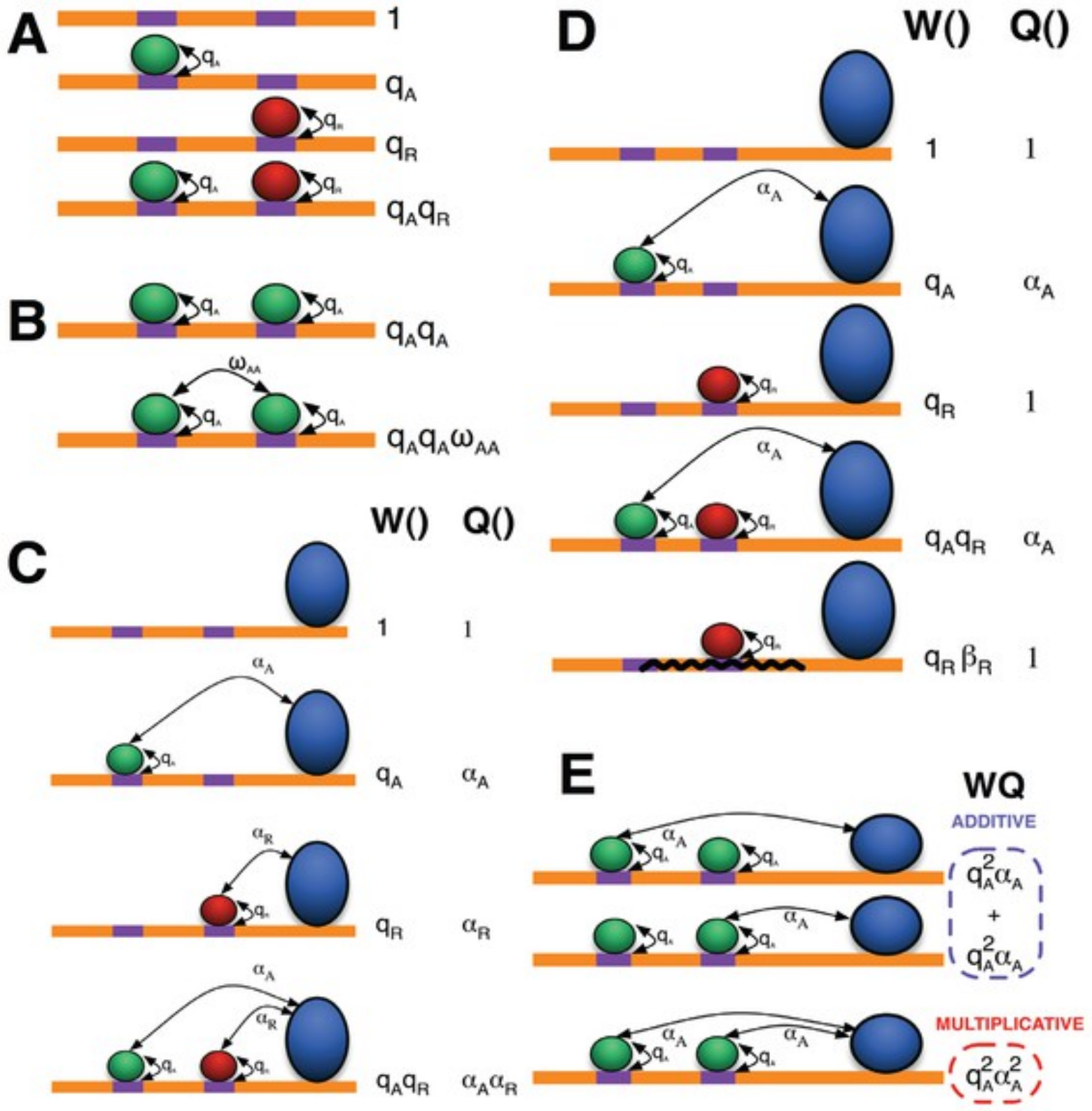
d) *D. mojavensis*



e) phylogenetic tree

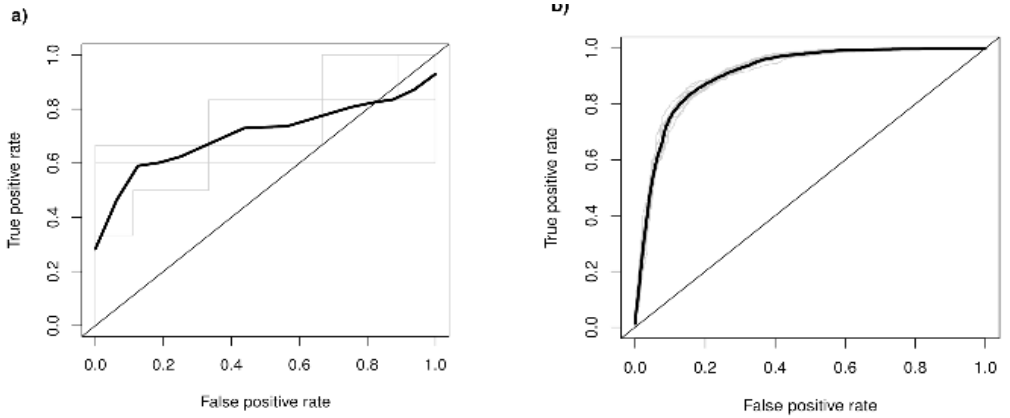
Wiązanie czynników transkrypcyjnych




A
1
q_A
q_R
$q_A q_R$
B
$q_A q_A$
ω_{AA}
$q_A q_A \omega_{AA}$
C
W()
Q()
1
1
q_A
α_A
q_R
α_R
$q_A q_R$
$\alpha_A \alpha_R$
D
W()
Q()
1
1
q_A
α_A
q_R
1
$q_A q_R$
α_A
$q_R \beta_R$
1
E
WQ
ADDITIVE
$q_A^2 \alpha_A$
+
$q_A^2 \alpha_A$
MULTIPLICATIVE
$q_A^2 \alpha_A^2$

We can add the sequence information

- First, we increase the size of the training dataset to use non-curated CRM predictions



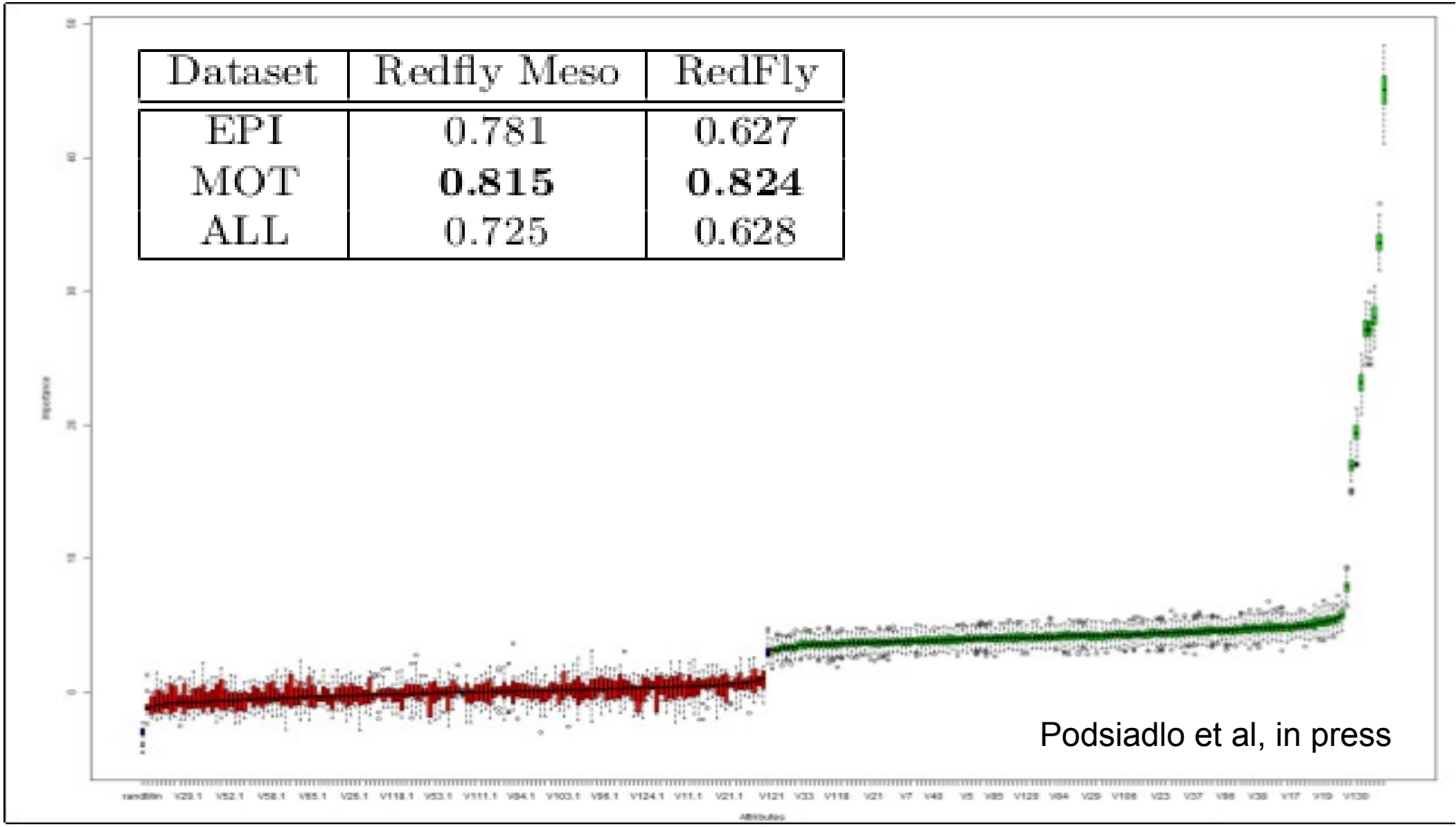
- Then, we add in data on motif affinity for JASPAR motifs

Dataset	BNFinder	SVM	RF
EPI	**0.9**	0.88	0.86
MOT	0.5	**0.89**	0.87
ALL	0.93	0.97	**0.98**

Podsiadlo et al, in press

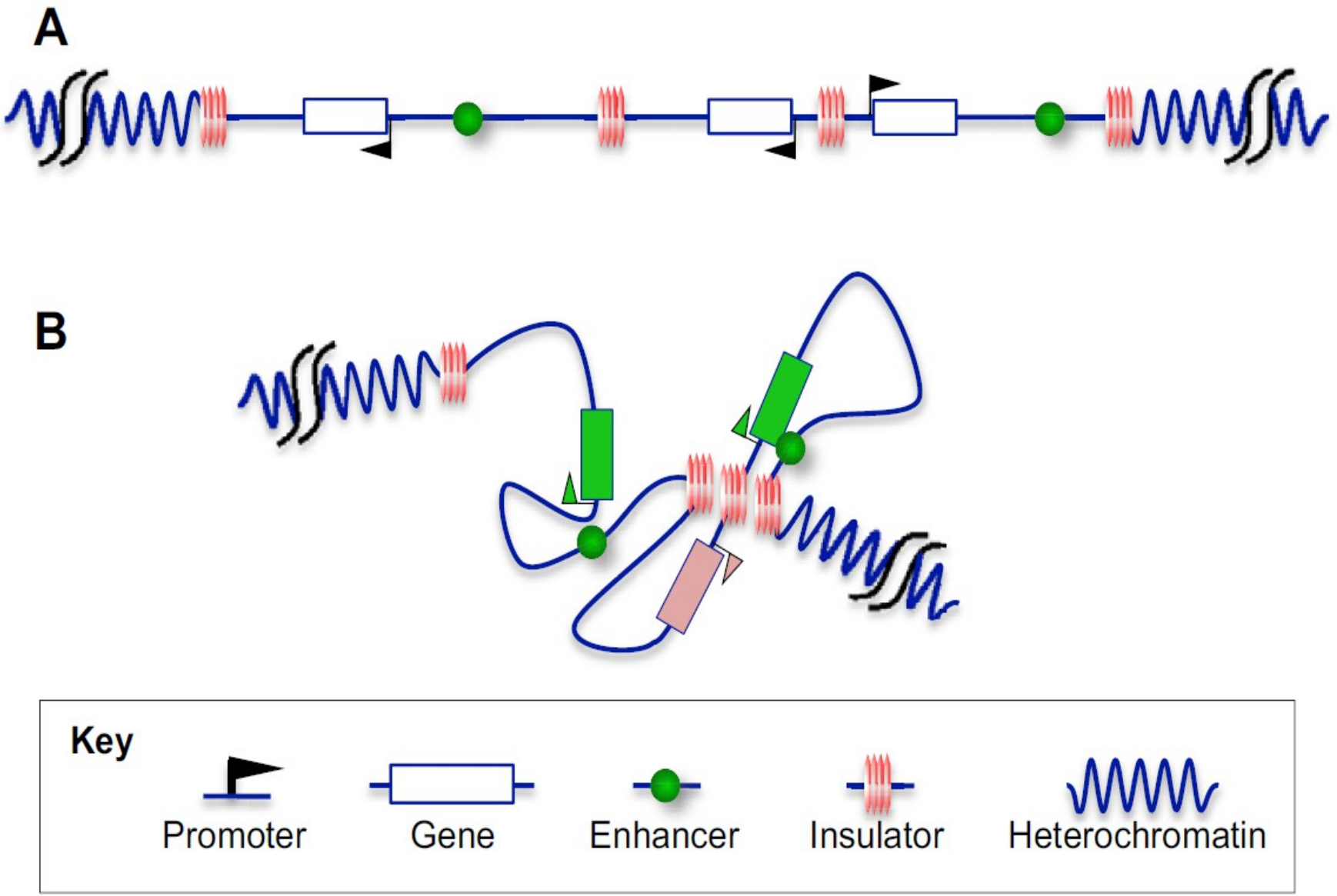
Importance of features is not consistent with generality

Dataset	Redfly Meso	RedFly
EPI	0.781	0.627
MOT	**0.815**	**0.824**
ALL	0.725	0.628



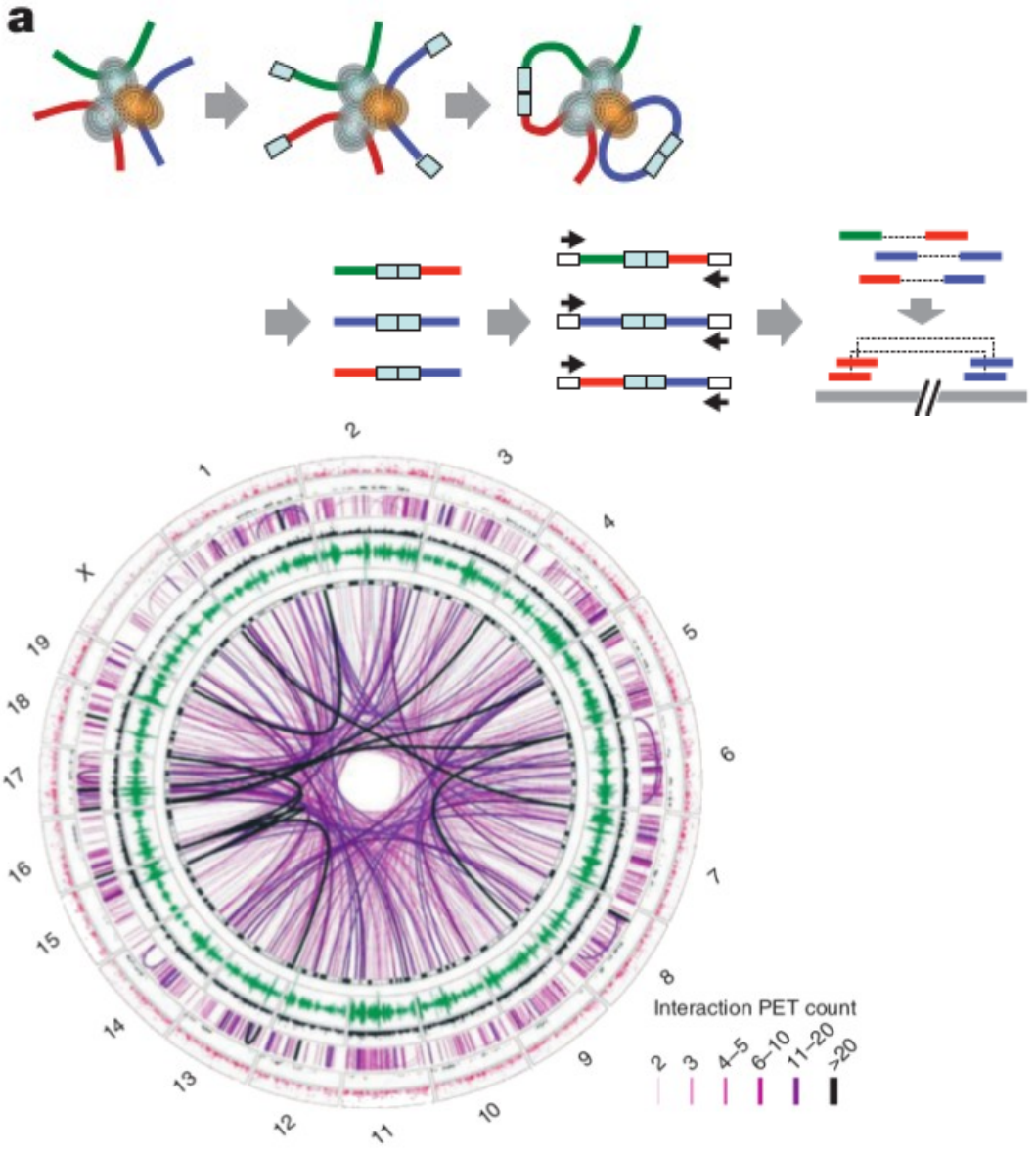
Podsiadlo et al, in press

Insulator looping creates regulatory domains



Herold et al. Development 2012

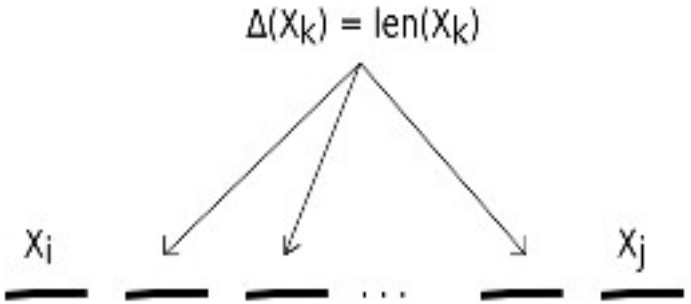
Enter chromosome interaction data



- 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 domains and interactions

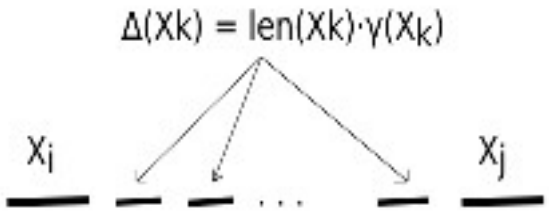
Data from Handoko et al. 2011

Basic Model

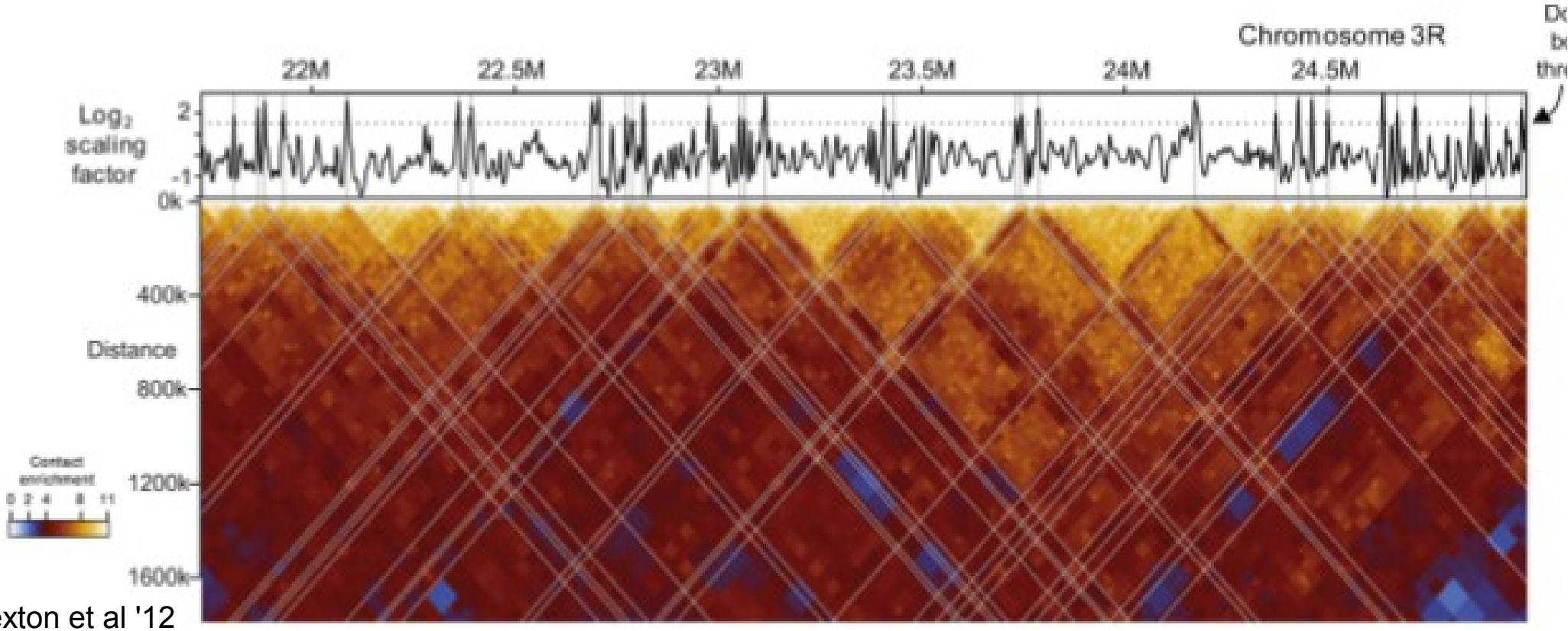


$$P(contact(X,Y)) = f(\sum_{i<k<j} len(X_k))$$

Distance scaling model

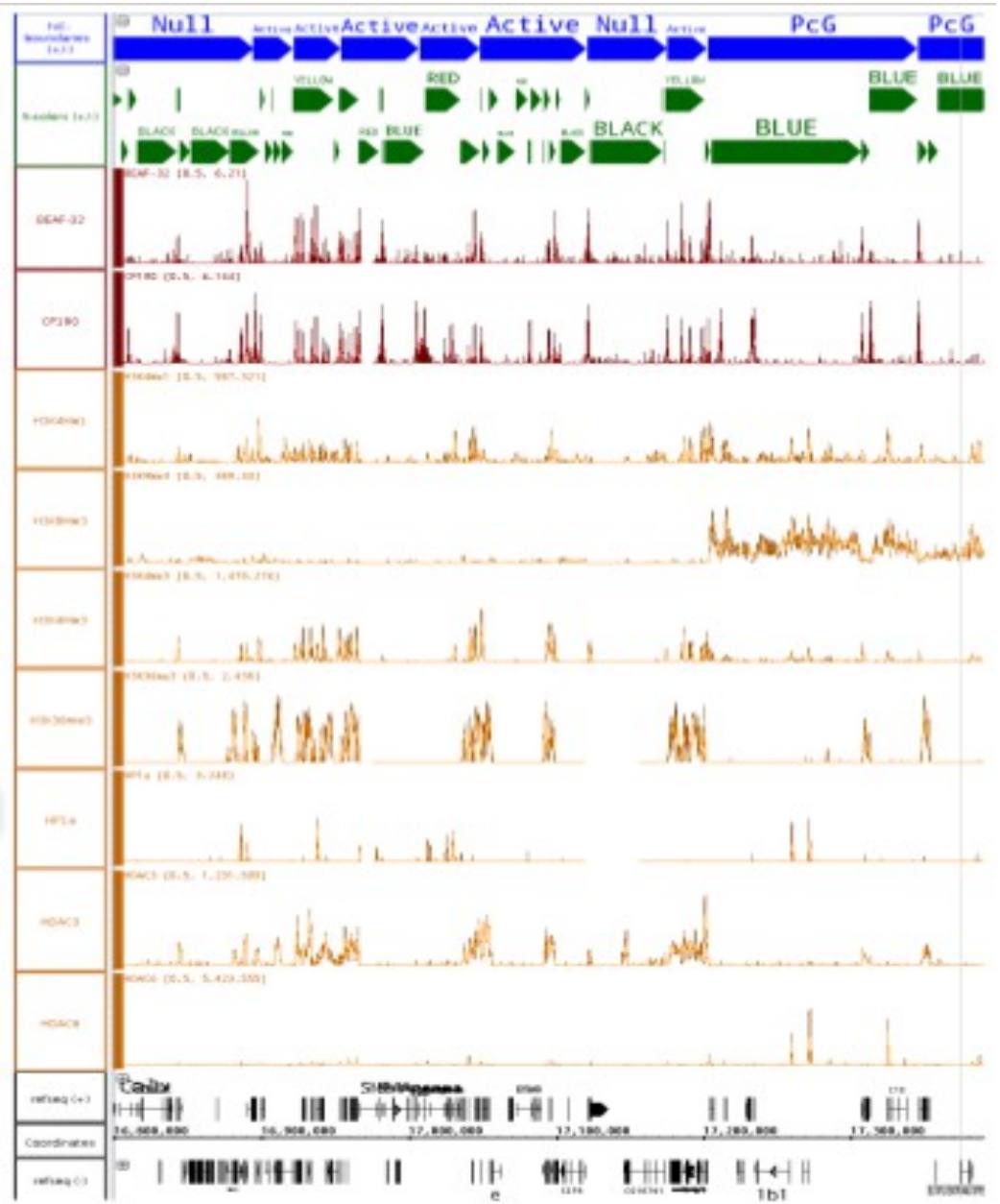


$$P(contact(X,Y)) = f(\sum_{i<k<j} len(X_k) \cdot \gamma_k)$$



Sexton et al '12

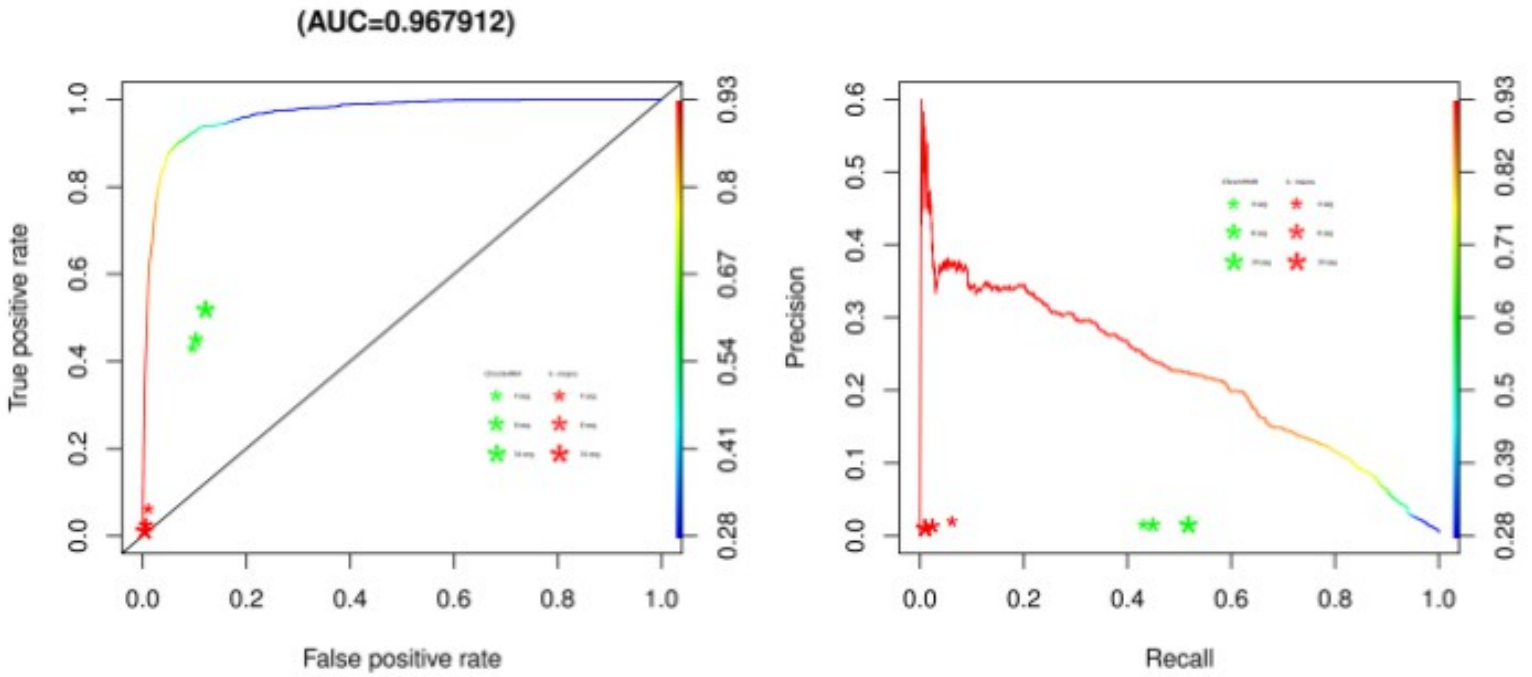
Predicting boundary elements from modEncode data



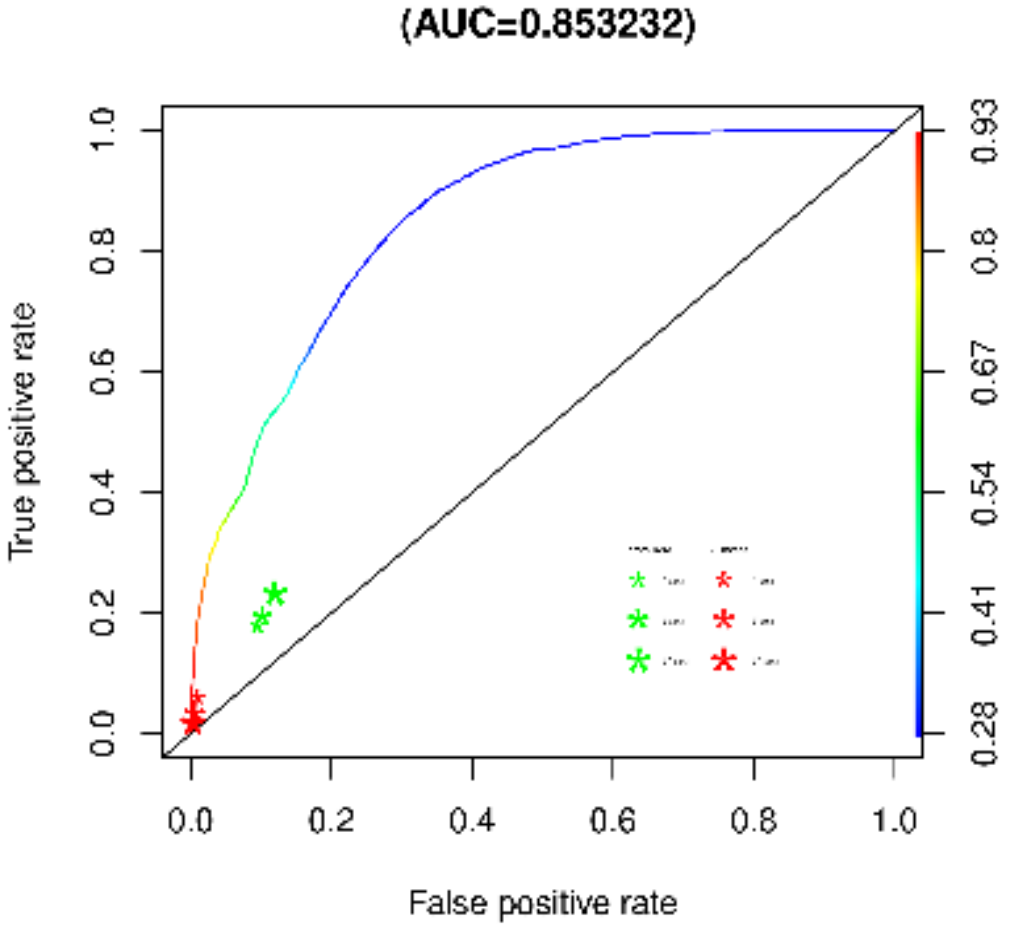
- We can use all chromatin IP data available in modENCODE for late embryos and try to predict domain boundaries
- We will use the Hi-C derived data as our main interaction set and Dam-ID derived data as additional validation

BN classifier can predict boundaries

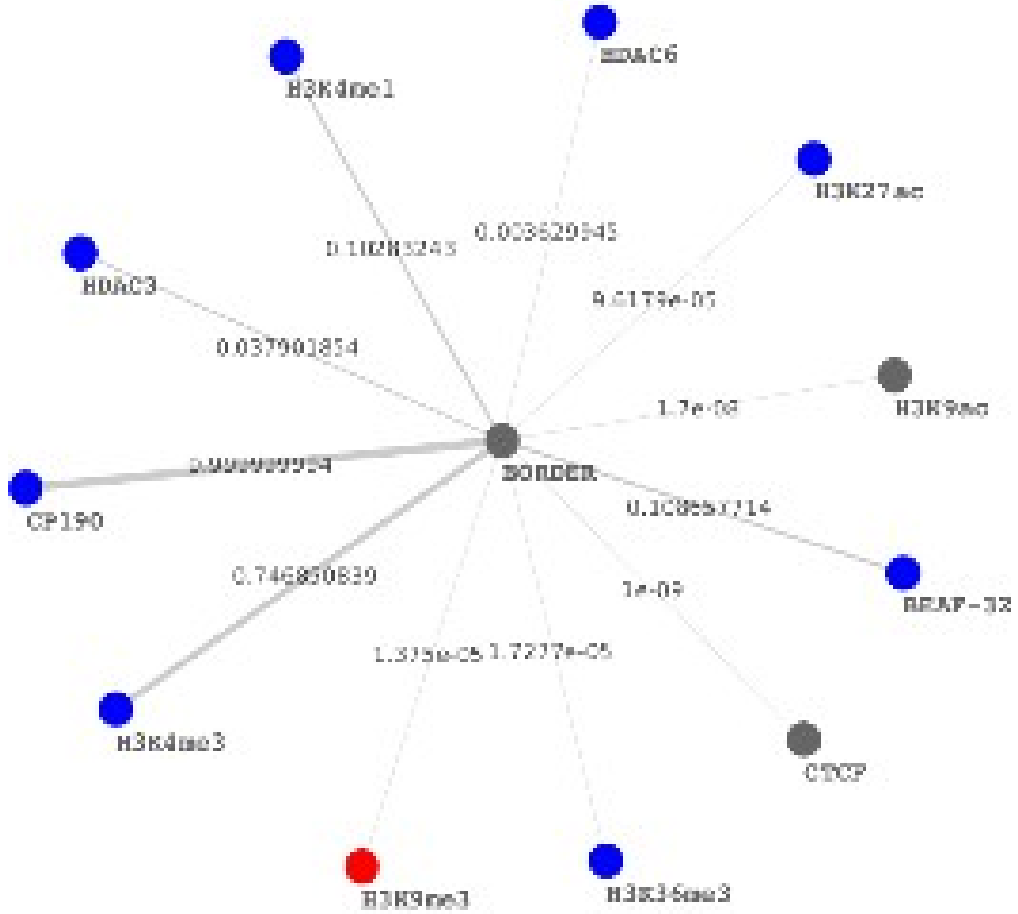
- Using BN classifiers trained on modENCODE data, we can predict position of boundary elements
- This method outperforms HMMs and clustering of histone modification data



Predictions make sense, and the model brings new information

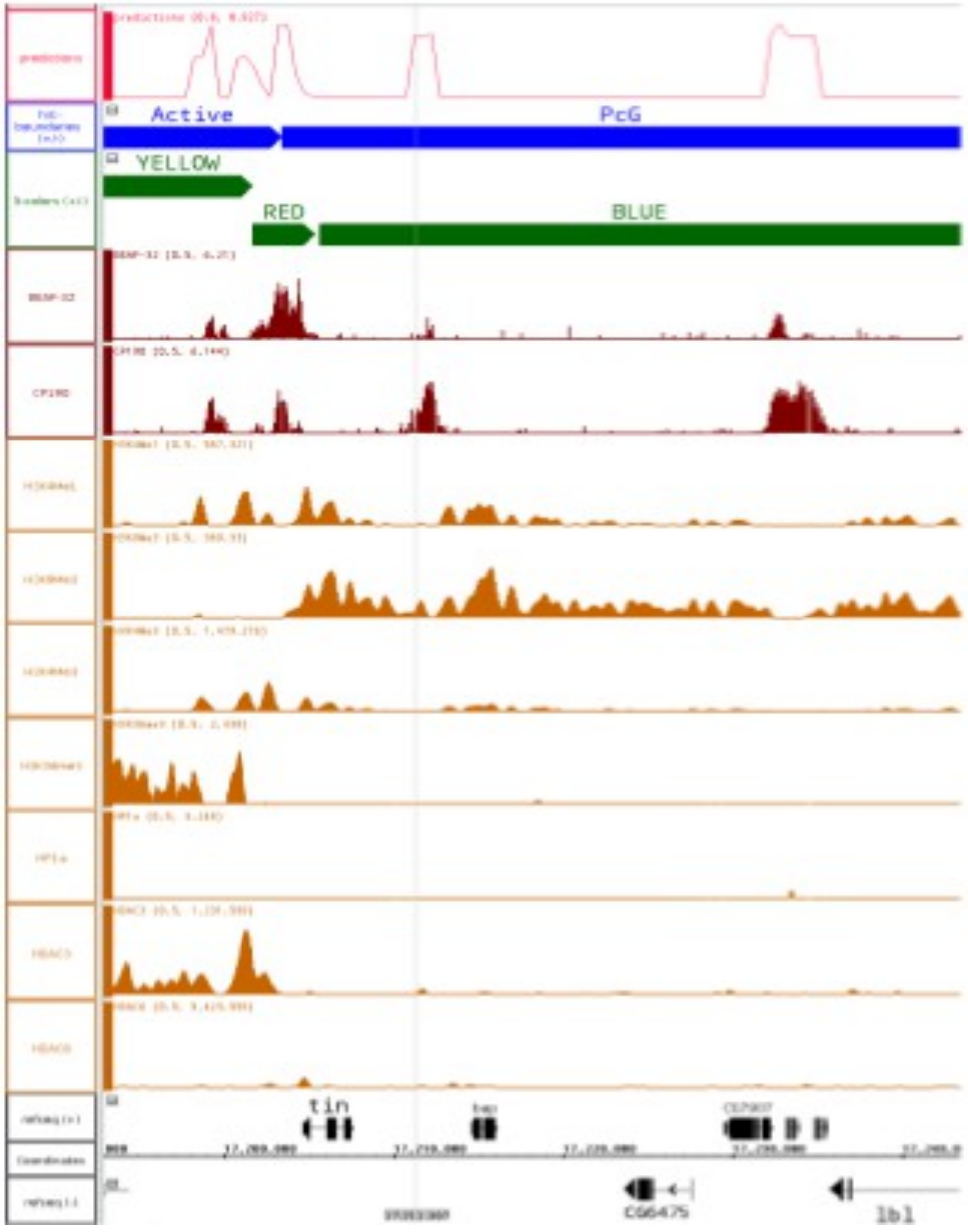


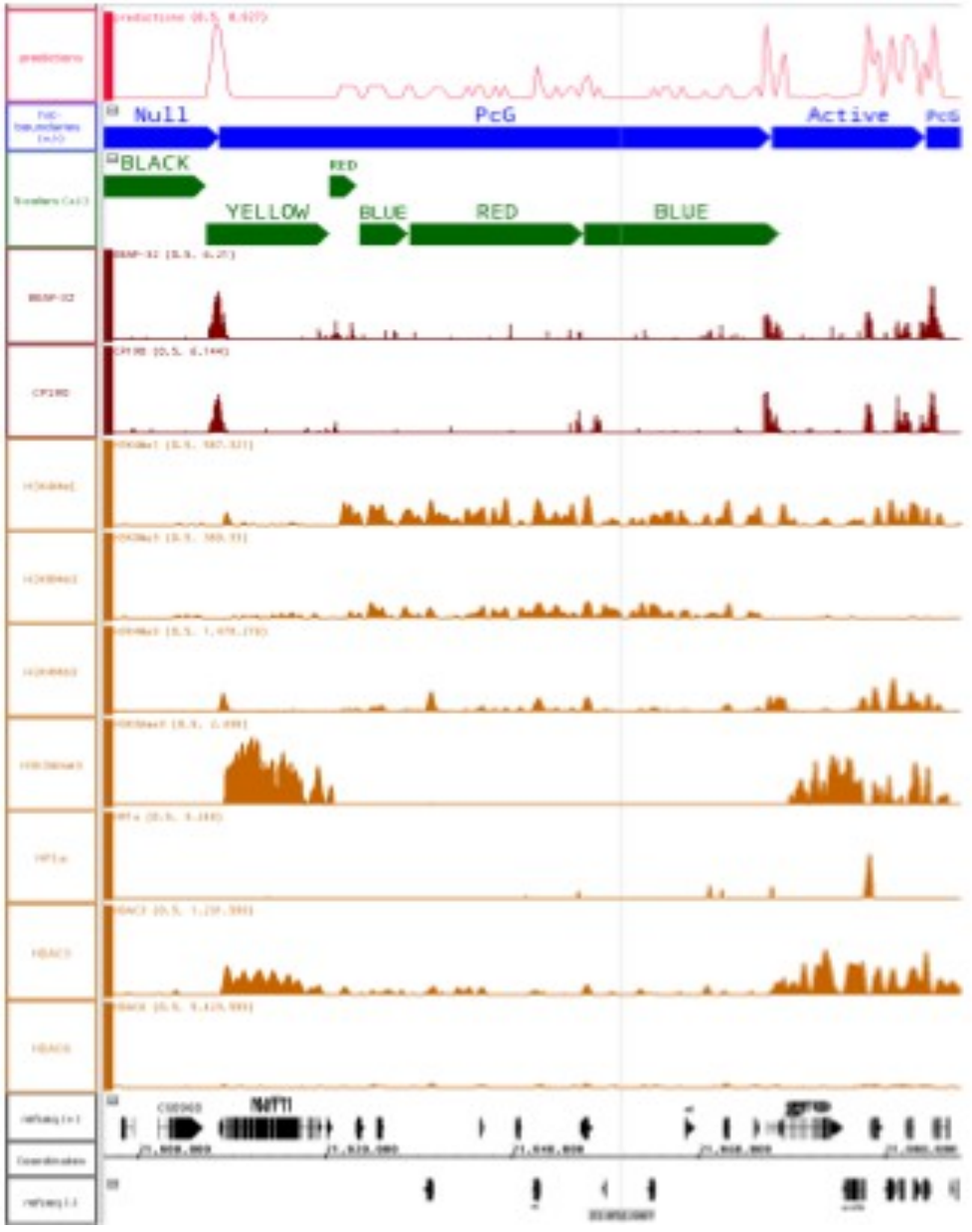
DamID validation results



Impact of signals

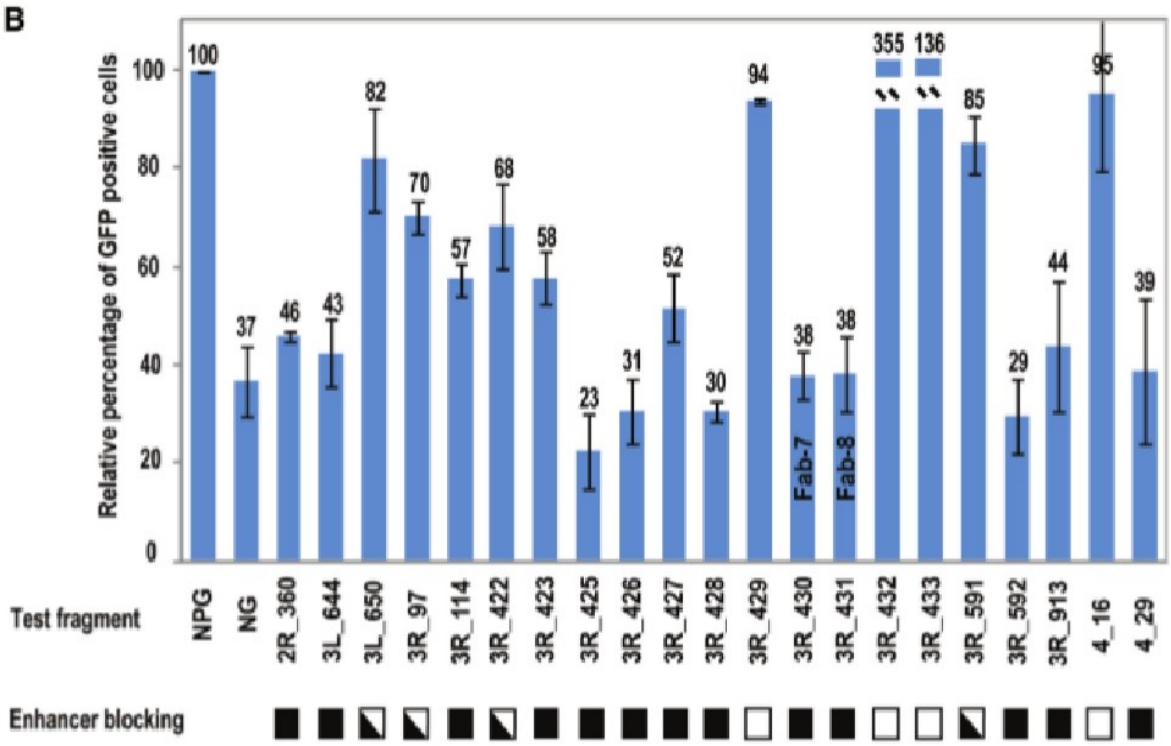
Some predictions are unexpected





Assaying insulator activity of predictions

- K. Mishra grroup used transgenic reporter assay to test their predictions from cdBEST method
- Our method predicts 14 out of 17 positive cases



Srinivasan et al, NAR 2012

Podsumowanie

- Używając zarówno cech sekwencyjnych jak i modyfikacji histonów można dobrze przewidywać lokalizację miejsc funkcjonalnych DNA
- W przypadku cech sekwencyjnych może nam pomóc porównanie spokrewnionych gatunków
- Ta metoda działa nie tylko dla enhancerów ale też dla miejsc izolatorowych