# Identification of Functional Direct Interaction between Transcription Factors and their targets, a computational approach focused on lincRNA

Marika Catapano, Elena Grassi, Paolo Provero and Ivan Molineris

University of Turin
Computational Biology Unit

16 October 2015

Introduction

- Gene expression regulation
- ChIP-seq technology

Materials and Methods

- TBA and coexpression as FDI predictors

Results

- The importance of the interplay between TFs and lincRNAs

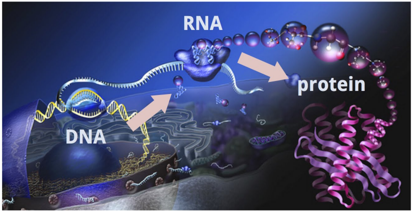
Conclusions

Future perspectives

# Transcriptional Regulation by TFs

**Transcription Factors (TF) are key proteins involved at this level**

- they are recruited by genomic regulatory region in a sequence dependent manner,
- allowing cells to precisely control the transcription rate.



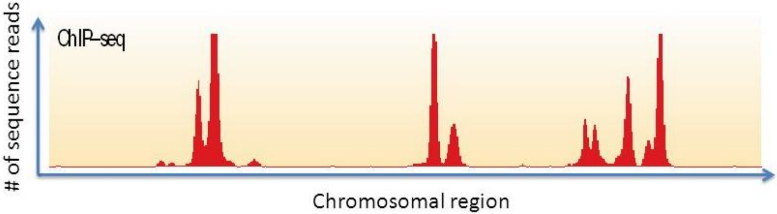
# ChIP-seq technology

ENCODE used ChIP-seq technologies to investigate binding regions.

- 4440 ChIP-seq experiments
- 141 TFs investigated [Homo Sapiens]

ChIPseq has some limits.
Among them:

- no discrimination between functional and non-functional binding.



Park.Nature Rewievs Genetics 2009.

# not only ChIP data..

**To further analyse TFs behaviour, ChIP data are often integrated with other observations.**

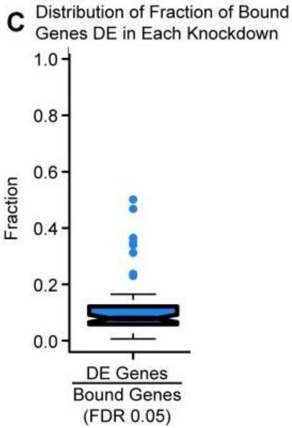
## Silencing

- integrating ChIP data with Differential expression under TFs perturbation (Cusanovich et al.)
  - TFs Knock-down
  - checking for differential expressed genes
  - ChIP data from ENCODE

# Binding does not imply functionality

In Cusanovich DA et al 's work:

- a median of 7.9% of the binding was apparently not functional.



# Aim of the work

The aim of the work is to develop a model:

- to identify direct interaction
  - discriminating between non-functional and direct functional interactions (FDIs)
  - using only generic data (genome sequence and global transcriptomic profiles) without the need of custom experiment.

**Gold standard FDI between a TF and a target gene:**

- Significant differential expression (FDR of 5%) of the targets after TF KD,
- Binding evidence of the TF in the target regulatory region.

# TBA and coexpression as FDI predictors

Our predictors:

Total Binding Affinity (TBA) of regulatory region

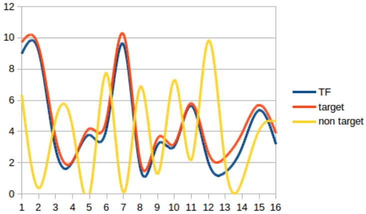
Measure of strength and preferential binding of TFs upon genomic regions.

Coexpression

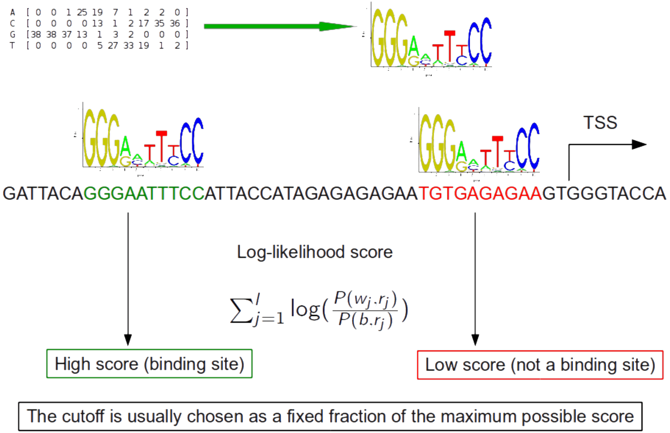
Correlation of gene expression data collected from 16 different cell lines (ENCODE).

# Coexpression

- Computed using Pearson's coefficient
- it measures the correlation of all TFs vs. all target genes.



# Transcription Factor binding



```
A  [ 0  0  1 25 19  7  1  2  2  0 ]
C  [ 0  0  0  0 13  1  2 17 35 36 ]
G  [38 38 37 13  1  3  2  0  0  0 ]
T  [ 0  0  0  0  5 27 33 19  1  2 ]
```

GATTACAGGGAATTTCCATTACCATAGAGAGAGAATGTGAGAGAAGTGGGTACCA

Log-likelihood score

$$\sum_{j=1}^{l} \log\left(\frac{P(w_j.r_j)}{P(b.r_j)}\right)$$

High score (binding site)

Low score (not a binding site)

The cutoff is usually chosen as a fixed fraction of the maximum possible score

- This approach assume that the binding of TFs acts through few strong binding sites.

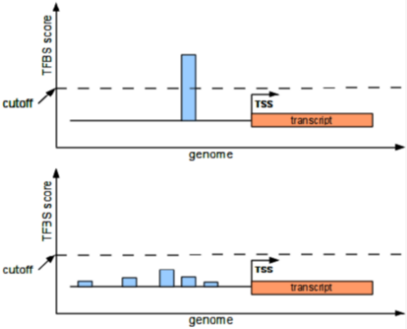
# Weak sites are important

TF binding is a thermodynamic process in which an important contribution is given by transient binding to weak sites

- Extensive low-affinity transcriptional interactions in the yeast genome. Tanay et al. Genome research (2006).
- Predicting expression patterns from regulatory sequence in Drosophila segmentation. Eran Segal et al. Nature (2008)

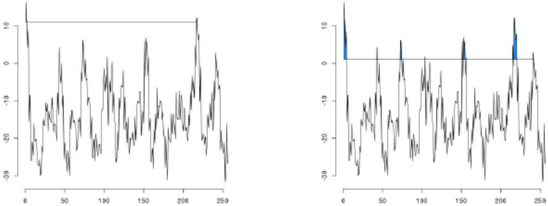
**Both strong and weak binding sites contribute, leading to high occupancy of the module DNA, and conferring robustness against mutation.**

- Approaches based on the definition of discrete binding sites inevitably miss such contribution.

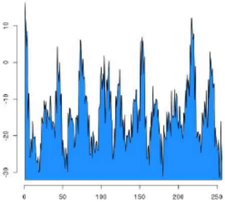


# TBA predicts binding better than cut-off-based methods

**Cut-off based methods loose sequence information, while TBA correctly takes it into account**



Cut-off based methods

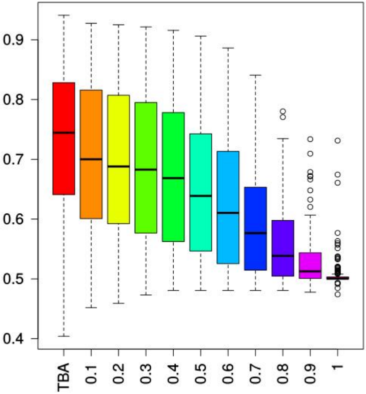


TBA

# TBA predicts binding better than cutoff-based methods

- We use Areas Under ROC curves to asses the predictive power of occupancy:
  - As positives we take the peaks of ChIP/seq experiments
  - As negatives we take the peaks of a control experiment (IgG mouse)
  - Compute the occupancy of all positives and all negatives at different cut-off
  - Compute the AUROC

⇒ Finally we study the dependence of the AUROC on the cut-off



[Grassi E. et al. unpublished]

- TBA is a better predictor of TF binding than methods based on the identification of discrete binding sites
- This confirms the relevance of transient binding to weak sites

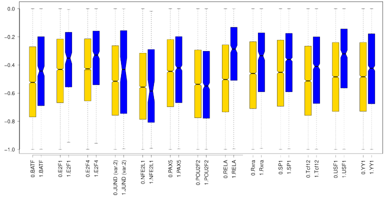
# TBA and coexpression as FDI predictors

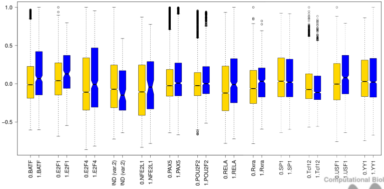
13 TFs in common with Cusanovich et al, for which PWMs were available.

- blue cases → true FDI
- yellow cases → no direct regulation

TBA values

Coexpression values





Computational Biology Unit

Molecular Biotechnology Center
Università di Torino

# Validation purpose: Logistic model development

**In order to integrate the information carried by TBA and Coexpression**

$$\log\left(\frac{p_1}{1-p_2}\right) = \alpha + \underset{\substack{\uparrow \\ \text{TBA}}}{\beta_1 X_1} + \underset{\substack{\uparrow \\ \text{coexpression}}}{\beta_2 X_2}$$

- independent logistic model for each TF to predict functional direct interaction (FDI) between a TF and a putative target gene.
- gold standard: Functional direct interaction (FDI).

# Fitting the model

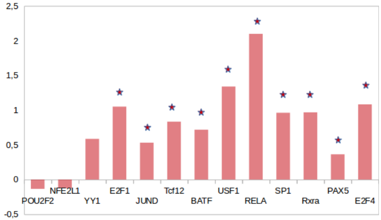
The models were significant in 12 cases with a Wilcoxon Pvalue $\leq 0.05$

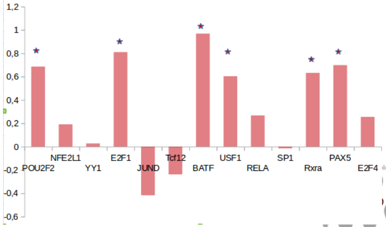
## TBA coefficients

- TBA showed to be a good predictor for the majority of TFs analysed.

## Coexpression coefficients

- Coexpression significantly contribute in 6 out of 13 cases for FDI predictions.





# FDI PREDICTION

We fitted a model, using all 13 TFs together, that can be generalized to any TFs for which we have PWMs information.

We made the model able to compute a probability of FDI for each TF and putative target.

**Final result:**

list of putative FDI regarding 208 TFs reported in the Jaspar database, including TF-lincRNA interactions.

| BIOTYPE | n of putative FDI target |
|---|---:|
| protein coding | 14641 |
| lincRNA | 7420 |
| antisense | 3148 |
| processed pseudogene | 1192 |
| unprocessed pseudogene | 1904 |

cut-off=60%

# TF-lincRNA FDI

LincRNAs are involved in many mechanisms.
Among them they have a crucial role in gene regulation and chromatin remodelling.

**TFs and lincRNAs are two important and diverse classes of regulatory molecules.**

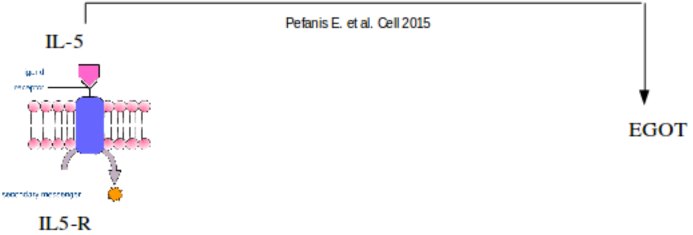
**Shedding light on the functional link between the two is crucial to understand the cell regulatory network.**

# AN EXAMPLE

**EGOT**
highly expressed in human bone marrow and in mature eosinophils, its expression can be induced by IL-5

→TFT is regulated by MAPK signalling
→MAPK takes role also in eosinophil differentiation pathway under IL-5R activation

**TFT**
(T brachyury transcription factor)
embryonic nuclear transcription factor that binds to a specific DNA element, the palindromic T-site



IL-5

Pefanis E. et al. Cell 2015

EGOT

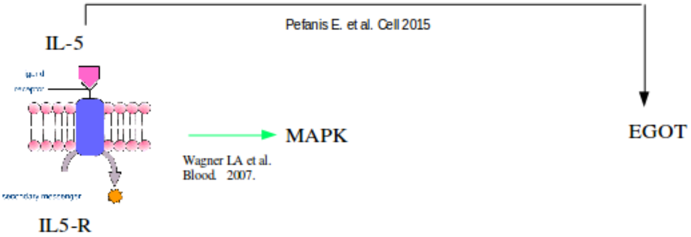
IL5-R

# AN EXAMPLE

**EGOT**
highly expressed in human bone marrow and in mature eosinophils, its expression can be induced by IL-5

→TFT is regulated by MAPK signalling
→MAPK takes role also in eosinophil differentiation pathway under IL-5R activation

**TFT**
(T brachyury transcription factor)
embryonic nuclear transcription factor that binds to a specific DNA element, the palindromic T-site



Pefanis E. et al. Cell 2015

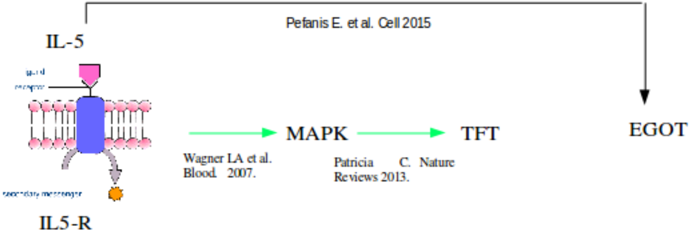
Wagner LA et al. Blood. 2007.

# AN EXAMPLE

**EGOT**
highly expressed in human bone marrow and in mature eosinophils, its expression can be induced by IL-5

→TFT is regulated by MAPK signalling
→MAPK takes role also in eosinophil differentiation pathway under IL-5R activation

**TFT**
(T brachyury transcription factor)
embryonic nuclear transcription factor that binds to a specific DNA element, the palindromic T-site



Pefanis E. et al. Cell 2015

Wagner LA et al. Blood 2007.

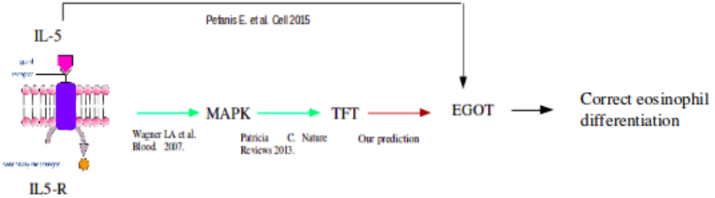
Patricia C. Nature Reviews 2013.

# AN EXAMPLE

**EGOT**
highly expressed in human bone marrow and in mature eosinophils, its expression can be induced by IL-5

→TFT is regulated by MAPK signalling
→MAPK takes role also in eosinophil differentiation pathway under IL-5R activation

**TFT**
(T brachyury transcription factor)
embryonic nuclear transcription factor that binds to a specific DNA element, the palindromic T-site



Petanis E. et al. Cell 2015

IL-5

MAPK

TFT

EGOT

Correct eosinophil differentiation

Wagner LA et al. Blood 2007.

Patricia C. Nature Reviews 2013.

Our prediction

IL5-R

# Conclusions I

- Integration of information by TBA and correlation in gene expression is a powerful predictor of FDI between TFs and their target genes.

# Repressor and activator TFs

Distinguishing between activating and repressing FDI by fitting two independent logistic models for each TF to predict:

- activating FDI,
- repressive FDI.

if the gene is upregulated after TF KD we used the data as negative example to fit the repressor model, if the gene was downregulated we used the data as positive example to fit the avtivator model. In this way we allowed each TF to act as activator on certain genes and as repressor on others.

# Repressive and activating TFs

ACTIVATOR

TBA coefficients showed to be significant for 9 TF for the activator model. Coexpression coefficients showed to be significant for 4 cases.

REPRESSOR

TBA showed to be significant in 5 cases for the repressor model. The coefficient associated to the coexpression was significant for 5 TFs.

**Coexpression coefficient was positive also for the significant repressing FDIs, where anticoexpression was expected.**

# Conclusion II

A possible explanation of this apparent paradox could be due to:

- coexpression that come from very different cell lines,
- such a coarse grained coexpression was always positively correlated with FDI indicating that TFs and genes need to be expressed in the same context to interact, even if the TF acts as a repressor.

We also tested tissue specific coexpression getting similar results.

# Future perspectives

The overall performance of the model is encouraging, but still not satisfactory (AUC 0.56).

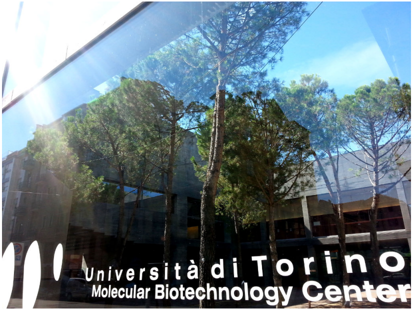
Some improvements could be made to the model:

- more sophisticated measure of coexpression,
- including composite effects of TFs,
- using additional predictors
    - epigenomic data
- make the model able to identify TF specific activity (repressive or activating)
    - time course expression data

# Acknowledgement

I would like to thanks

- Paolo Provero
- Ivan Molineris
- Elena Grassi



and the Computational Biology Unit at University of Turin, MBC-Molecular Biotechnology Center.

Ugo Ala, Davide Marnetto, Umberto Perron, Elisa Mariella, Simona Baghai, Federica Mantica, Elisa Reale and Stefano Gigliotti.

Computational Biology Unit CBU@MBC Molecular Biotechnology Center Università di Torino