

Inference of human transcription regulatory networks using deep sequencing data

Erik van Nimwegen

*Biozentrum, University of Basel,
and Swiss Institute of Bioinformatics*

What does “Inferring transcription regulatory networks” mean?

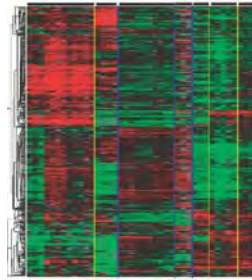
- For each TF, determine its cis-regulatory elements (binding sites) genome-wide.
- Determine which TFs are *active* under what conditions:
 - expression.
 - nuclear localization.
 - post-translational modifications.
 - anything that affects the TF's effect on its target genes.
- Determine time-dependent activities of TFs in dynamic processes such as cell cycle, developmental processes, etc.
- Determine the effect of each cis-regulatory element on the expression of the target gene.
- Determining the transcription regulatory logic of the cis-regulatory elements, i.e. mapping from TF binding configurations to effects on expression.

Ultimately we would like to be able to predict the expression dynamics of all genes essentially just from their DNA sequences

Typical high-throughput approaches

Gene expression data
(microarray)

clustering



Regulatory
"modules"

Examples:

Segal et al. Nat. Genet 2003

Beer and Tavazoie Cell 2004

Association

Over-
representation

Correlation

Pathways/
Functional
categories

Regulatory
motifs

TF expression
profiles

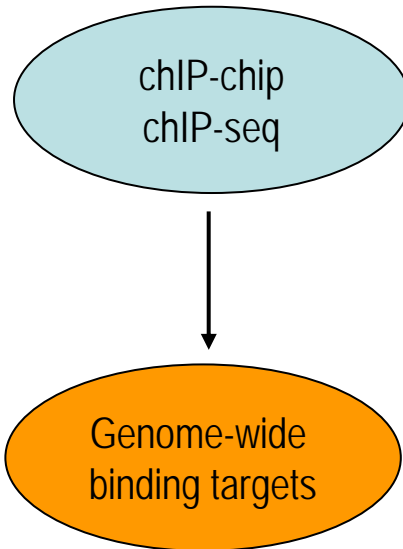
Benefits:

- One identifies regulatory *programs* i.e. cohorts of co-regulated genes in the process/condition under study.
- Relevant pathways identified.
- TFs/regulatory motifs are associated with the modules.

Disadvantages:

- Only some genes cluster, cluster boundaries are often unclear.
- Direct physical meaning often lacking.
- Gene expression profiles are not explained, but just classified.

Targeted high-throughput approaches



Examples:

Boyer et al. *Cell* 2005

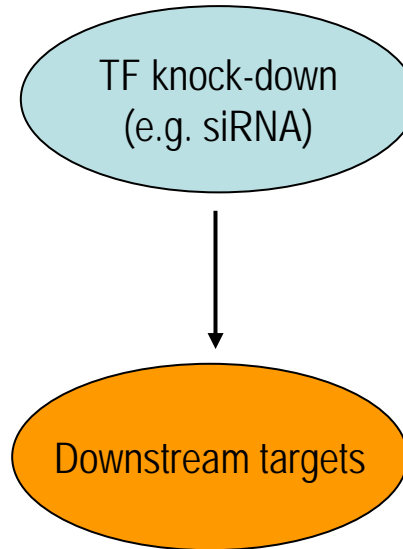
Jakobsen et al. *Genes & Dev.* 2007

Benefits:

- Infer direct molecular interactions.
- Genome-wide.

Disadvantages:

- Binding does not imply expression effects.



Examples:

Davidson et al. *Science* 2002

Imai et al. *Science* 2006

Benefits:

- Identify effects on expression.
- Genome-wide.

Disadvantages:

- Direct and indirect effects entangled..

- Labor intensive (one TF at a time)
- **Need to know the relevant TFs in advance**

Accelerating regulatory network reconstruction through computational prediction

- Real network reconstruction requires targeted and detailed experimental work.
- Provide analysis of high-throughput data that most efficiently tells *where to look*.

Develop a computational frame-work that:

- Uses easily produceable high-throughput data, e.g. **micro-array data**.
- Predict the **transcription regulators** that play a **key role** in the process under study (developmental time course, response to perturbations, disease versus healthy tissue).
- Predict how the regulators **change activity** (up-regulation, down-regulation, transient changes).
- Predict the **target gene sets** of the key regulators.
- Identify the **cis-regulatory elements** on the genome through which the regulators acts.

Linear models

- Explicitly predicting gene expression in terms of *activities* of the transcription factors, and the *response coefficients* of each gene to each transcription factor:

$$e_{gs} = \text{noise} + \tilde{c}_s + c_g + \sum_f R_{gf} A_{fs}$$

Expression of gene g in sample s

Basal gene expression

Response of gene g to factor f .

Activity of factor f in sample s

- Assumes a linear function. This is wrong but never a bad approximation when changes are not too large.
- The activities and response coefficients are inferred from the data and/or computational analysis.

Review: Bussemaker et al. *Annu Rev Biophys Biomol Struct* 2007

Linear models

- Explicitly predicting gene expression in terms of *activities* of the transcription factors, and the *response coefficients* of each gene to each transcription factor:

$$e_{gs} = \text{noise} + \tilde{c}_s + c_g + \sum_f R_{gf} A_{fs}$$

Response of gene
g to factor *f*.

We use DNA sequence analysis to predict transcription factor binding sites and estimate response coefficients in human genome-wide.

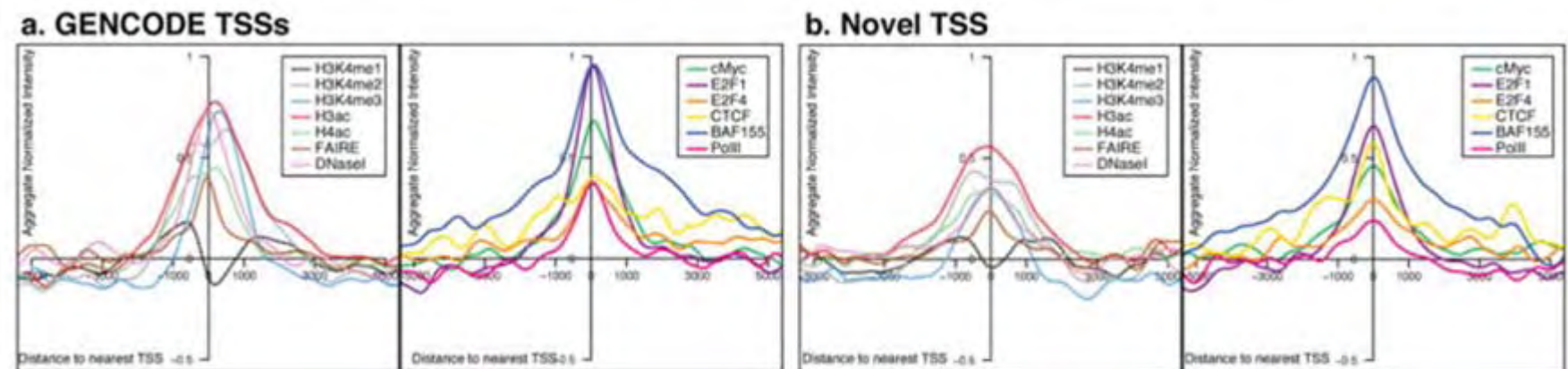
TFBS prediction in mammals: Focus on proximal promoters

Challenge:

- The intergenic regions in mammals are vast and functional sites can occur far from the gene.

However,

- Data from the ENCODE project suggests a large fraction of functional regulatory sites occurs near TSS. (*Nature*. **447**:799-816 2007)
- Regulatory sites thought to be distal often turn out to be alternative promoters.
- ChIP-chip for several TFs shows peaks at TSS:



We have a technology for mapping TSSs and their expression genome-wide.

Deep sequencing of 5' ends of mRNAs CAGE technology

Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.

Shiraki et al. *PNAS* **23** 15776-81 (2003)

Tag-based approaches for transcriptome research and genome annotation

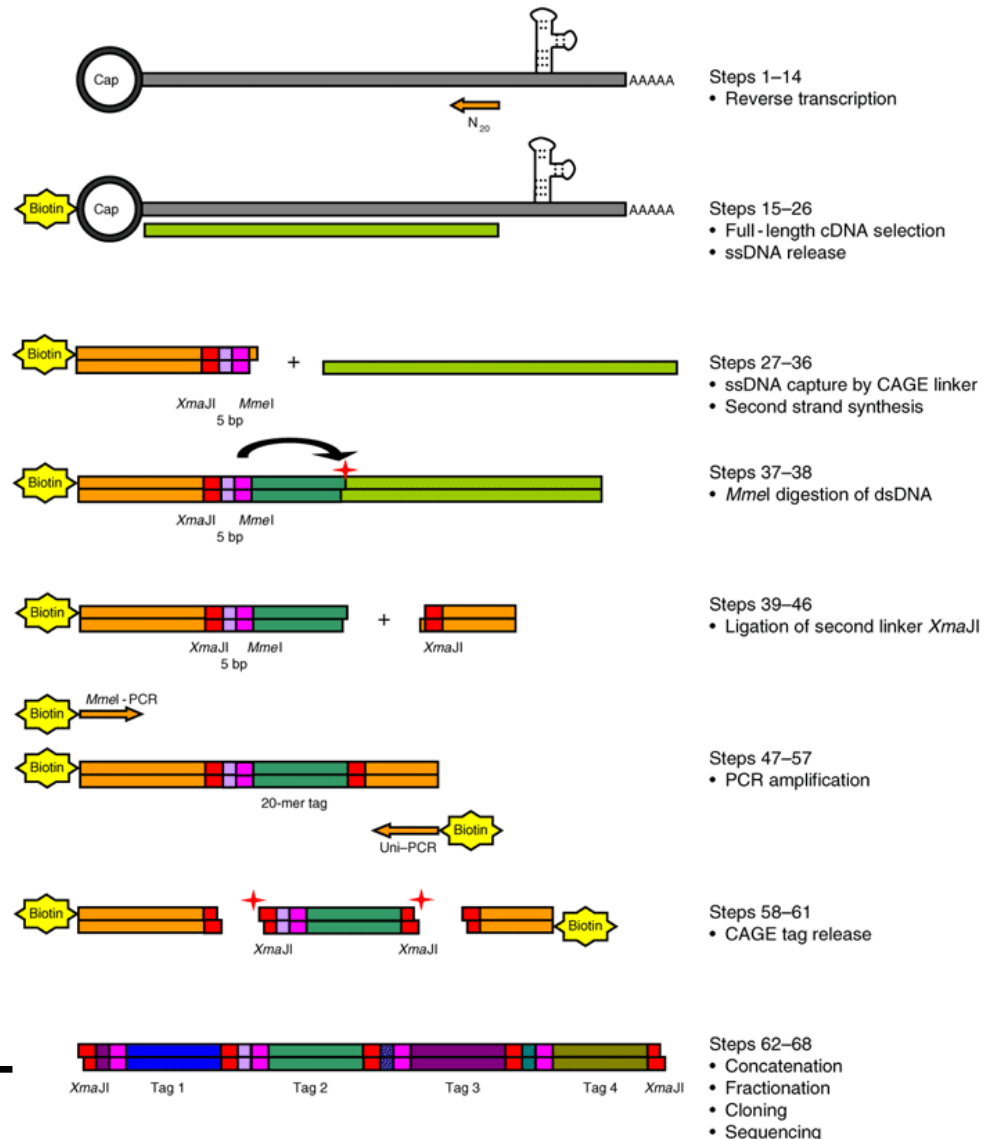
Harbers M, Carninci P.

Nat Methods **2** 495-502 (2005)

Tagging mammalian transcriptome complexity

P. Carninci

Trends Genet **22** 501-10 (2006)

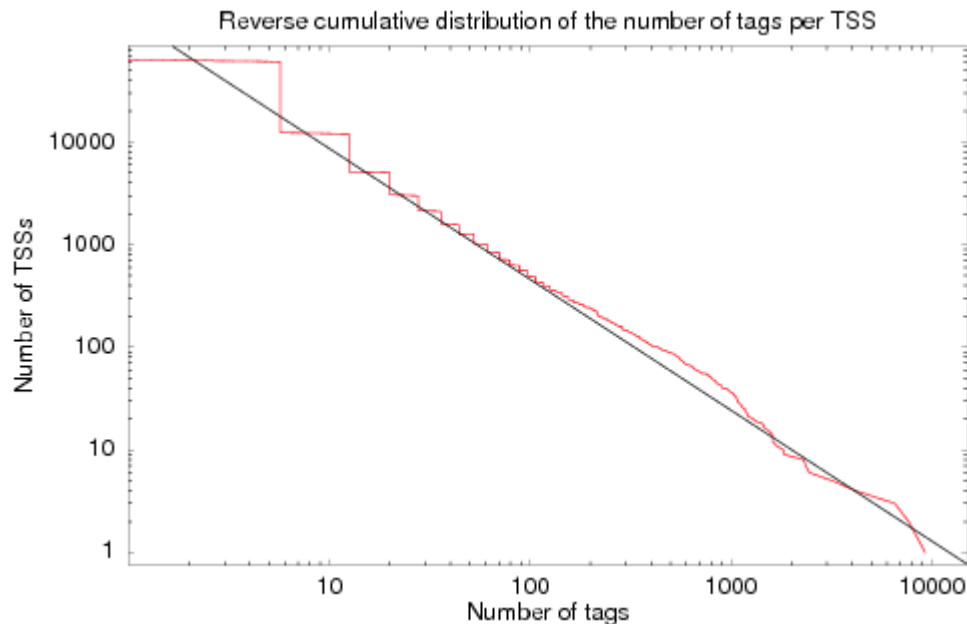


454/Solexa sequencing.
Mapping to the genome.

Deep sequencing of 5' ends of mRNAs

Number of samples with $> 10^5$ tags	56
Total number of mapped CAGE tags	25,469,648
Number of unique TSS positions	3,006,003

For any given sample the distribution of tags per TSS is a power-law:



The vast majority of TSSs have very low expression: **background transcription**.
The distribution can be used to **normalize** CAGE-tag counts across samples.

Expression noise can be modeled as *multiplicative noise*, followed by *Poisson sampling*.

x = true log-expression (per million).

n = raw number of tags.

t = normalized number of tags.

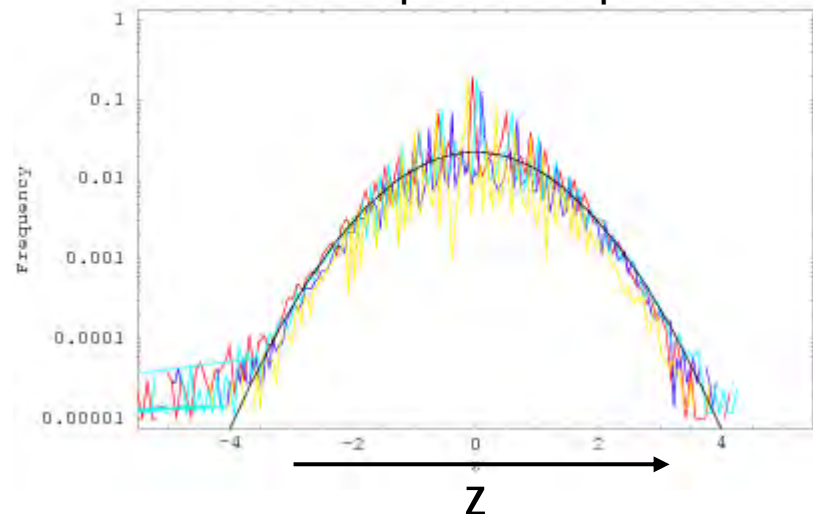
σ^2 = variance of the multiplicative noise.

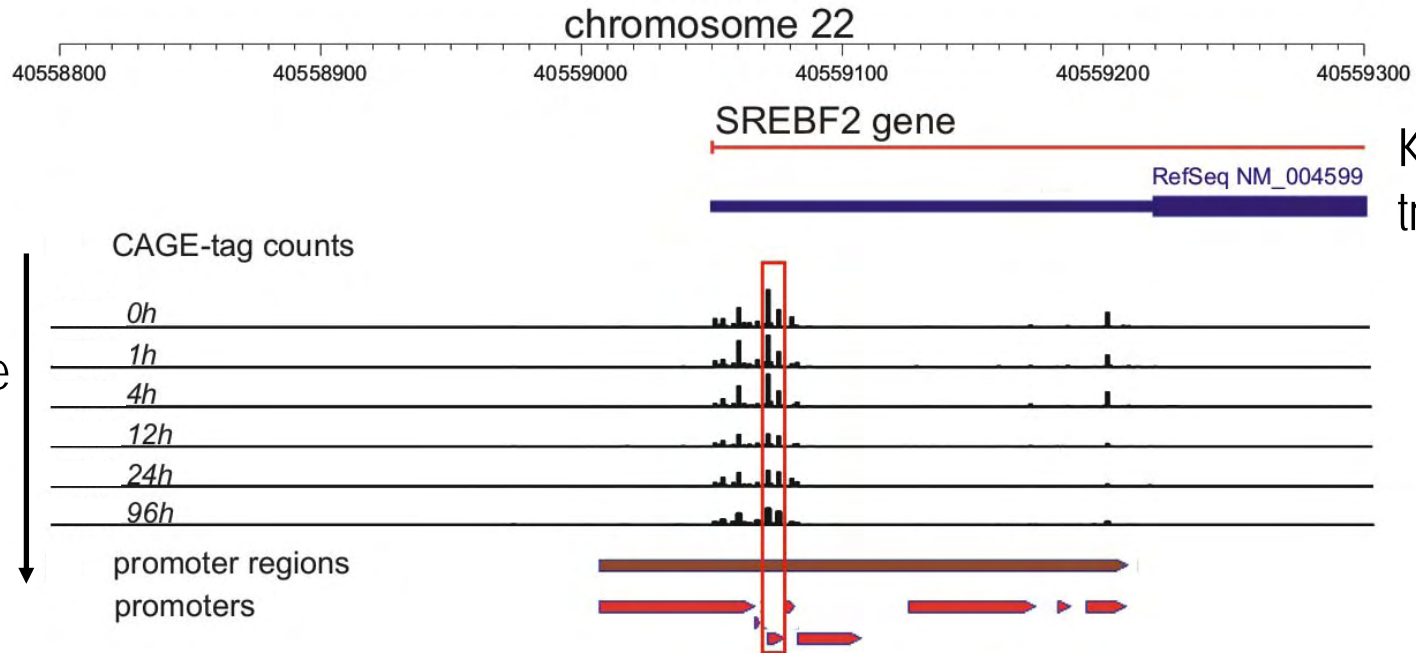
$$P(t \mid x, \sigma) = \frac{\exp\left(-\frac{1}{2} \frac{(\log(t) - x)^2}{\left(\sigma^2 + \frac{1}{n}\right)}\right)}{t \sqrt{2\pi \left(\sigma^2 + \frac{1}{n}\right)}}$$

Measure distribution of observed z-values for replicates.

$$z = \frac{\log(t_1) - \log(t_2)}{\sqrt{2\sigma^2 + \frac{1}{n_1} + \frac{1}{n_2}}}$$

Observed and predicted replicate noise





Known transcripts

Time course

What is a promoter?

Answer. A set of neighboring TSSs whose expression-profile is indistinguishable up to noise. We also cluster nearby promoters into promoter regions.

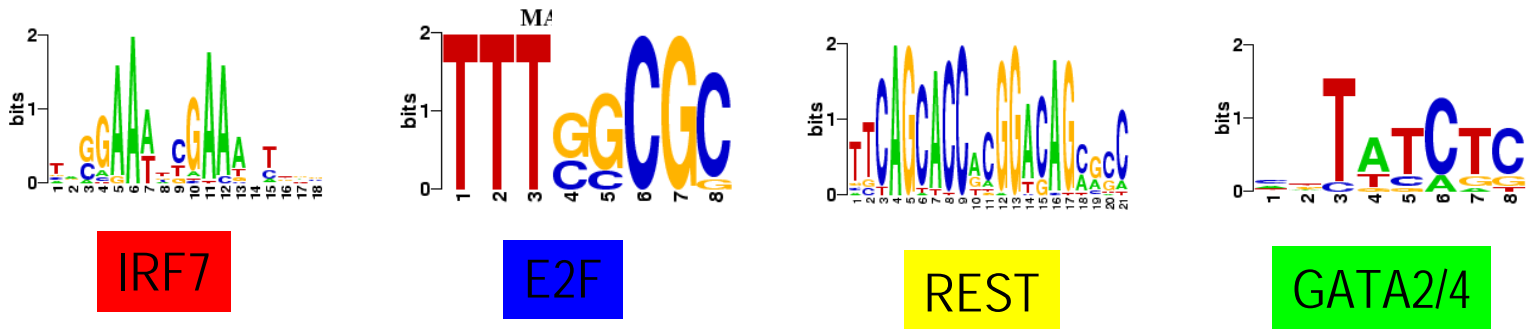
Number of promoter regions	43,164
Number of promoters	74,273
Number of TSSs in promoters	860,823
Total number of TSSs	3,006,003

Human promoterome

Predicting TFBSs in all proximal promoters

Input:

- 203 mammalian regulatory motifs (weight matrices) representing 551 human TFs.



- 43,164 proximal promoter regions (-300,+100) with respect to each TSS.
- Alignments with orthologous regions from other mammals.

CATTTCGCAGTGGCAAGGGACTGCCCTGGTCCCTGTGGAGC--GTCCCATTCGGTGACTTCCCACCAGCCCTTCCCCAGCGCCTCTGGAGGTCCAGACTGTCAGGTTGGAGCCTGGG
 CATTTCACAGTGGCAAGGGTCCGCCCTGGTCCCTGTGGAGG--GTCCAGTCGGTGACTTCCCAGCAGCCCTTCCCCAGTGCCTCTGGAGGTC--GACTGTC--GGTTGGAGCCTGG
 GAGGGGCGG---CTCGGGAGG-----CTGCGGACC--GGGCGAG--CGGGGGCG--GCG----GGGCGGCGGGGAGCCGGGCGGGGGCC-----TGCGGTTCGG--GCCTGG
 GATTGGCCGCGGCCAAGGACCCC-----TCCCTGGGGAGC--GTCCGGGTTCGGAGACT--CCCACTTGCCCTTCTCCAGCACCTCGTGAAGTCCGGAAGTGTACGGTTTG--GACTCG
 TATCTACAACAGCAAG--GA-----GTC--TG--GAAGCAAGTCCAAGT--GATGGA--TACAGCCATCACTTACC--GGGCCTCTGCTGGTTCGTGACTT-----

- The phylogenetic tree relating the species:
- ```

graph LR
 Root --- Dog
 Root --- Node1
 Node1 --- Cow
 Node1 --- Node2
 Node2 --- Mouse
 Node2 --- Node3
 Node3 --- Rhesus macaque
 Node3 --- Human

```

$$F_{n-1} \quad P(S_n | b, T)$$

|      |                                 |                                                           |
|------|---------------------------------|-----------------------------------------------------------|
| Scer | AAAAAATGAAAAATTCATGAGAAAAGAGTCA | GACATC-GAAACATACATAA--GTTGATATTC-CTTTGATATCG-----ACGACTA  |
| Spar | AAAAAATGAAAAATTCATGAGAAAAGAGTCA | GACATC-GAAACATACATAA--ATTGATATTC-CTTTAGCTTTT----AAAGACTA  |
| Smik | GAAAAACGAAAAATTCATG-GAAAAGAGTCA | ACCGTC-GAAACATACATAA--ACCGATATTT-CTTTAGCTTTTCGACAAAAATCTG |
| Sbay | GAAAAATAAAAAGTGATTG-GAAAAGAGTCA | GATCTCCAAAACATACATAATAACAGGTTTTTACATTAGCTTTT----GAAAACTA  |

$$F_{n-l} \quad P(S_{[n-l,l]} | w, T)$$

|      |                         |                                                                   |
|------|-------------------------|-------------------------------------------------------------------|
| Scer | AAAAAATGAAAAATTCATGAGAA | AAGAGTCAGACATC-GAAACATACATAA--GTTGATATTC-CTTTGATATCG-----ACGACTA  |
| Spar | AAAAAATGAAAAATTCATGAGAA | AAGAGTCAGACATC-GAAACATACATAA--ATTGATATTC-CTTTAGCTTTT----AAAGACTA  |
| Smik | GAAAAACGAAAAATTCATG-GAA | AAGAGTCAACCGTC-GAAACATACATAA--ACCGATATTT-CTTTAGCTTTTCGACAAAAATCTG |
| Sbay | GAAAAATAAAAAGTGATTG-GAA | AAGAGTCAGATCTCCAAAACATACATAATAACAGGTTTTTACATTAGCTTTT----GAAAACTA  |

$$F_{n-l} \quad \int P(S_{[n-l,l]} | w, T) P(w) dw$$

|      |                       |                                                                     |
|------|-----------------------|---------------------------------------------------------------------|
| Scer | AAAAAATGAAAAATTCATGAG | AAAAGAGTCAGACATC-GAAACATACATAA--GTTGATATTC-CTTTGATATCG-----ACGACTA  |
| Spar | AAAAAATGAAAAATTCATGAG | AAAAGAGTCAGACATC-GAAACATACATAA--ATTGATATTC-CTTTAGCTTTT----AAAGACTA  |
| Smik | GAAAAACGAAAAATTCATG-G | AAAAGAGTCAACCGTC-GAAACATACATAA--ACCGATATTT-CTTTAGCTTTTCGACAAAAATCTG |
| Sbay | GAAAAATAAAAAGTGATTG-G | AAAAGAGTCAGATCTCCAAAACATACATAATAACAGGTTTTTACATTAGCTTTT----GAAAACTA  |

**MotEvo:**

van Nimwegen, E.

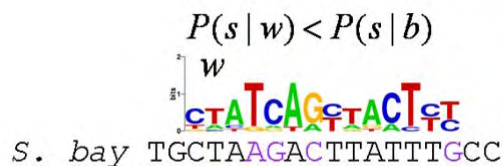
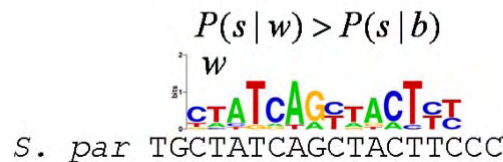
*BMC Bioinf* 8 Suppl 6, S4 (2007)

**MONKEY:**

Moses, A.M., Chiang, D.Y., Pollard, D.A.,  
Iyer, V.N. & Eisen, M.B.

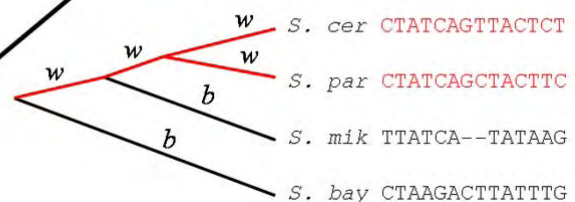
*Genome Biol* 5, R98 (2004).

WM probs single sequences



|        |                    |
|--------|--------------------|
| S. cer | TACTATCAGTTACTCTTC |
| S. par | TGCTATCAGCTACTTCCC |
| S. mik | TGTTATCA--TATAAGTA |
| S. bay | TGCTAAGACTTATTTGCC |

Selection pattern

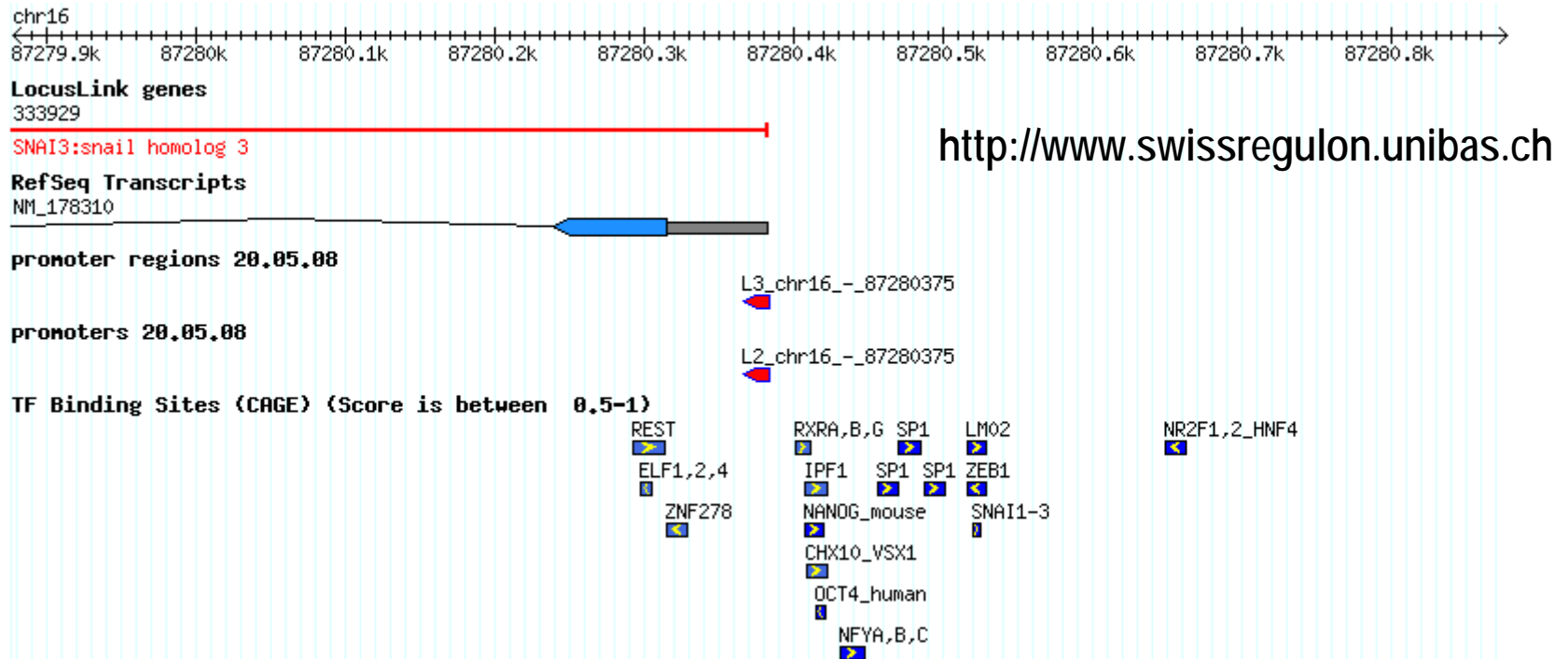






# Genome-wide annotation of regulatory sites

**Example:** Predicted TFBSs in the proximal promoter of the SNAI3 TF.



For each promoter  $p$  and motif  $m$  calculate the predicted number of functional sites

$$N_{pm}$$



$$e_{ps} = \text{noise} + \tilde{c}_s + c_p + \sum_m N_{pm} A_{ms}$$

Expression of promoter  
 $p$  in sample  $s$

Basal promoter  
expression

Number of functional  
sites in promoter  $p$  for  
motif  $m$

Activity of motif  $m$   
in sample  $s$

Fitting activities, minimize:

$$\sum_p \left( e_{ps} - \sum_m N_{pm} A_{ms} - c_p - \tilde{c}_s \right)^2$$

SVD

$$A_{ms}^* \pm \delta A_{ms}$$

Significance of the motif:

$$z_m = \sqrt{\frac{1}{S} \sum_{s=1}^S \left( \frac{A_{ms}}{\delta A_{ms}} \right)^2}$$

Similar approach in yeast:

Nguyen DH, and P. D'haeseleer  
*Mol. Syst. Biol.* (2006)

doi:10.1038/msb4100054

Application to human:

Das, D., Nahle, Z. & Zhang, M.Q.  
*Mol Syst Biol* 2, 2006 0029 (2006).

# Human tissue atlas and cancer cell expression data

[Proc Natl Acad Sci U S A](#). 2004 Apr 20;101(16):6062-7. Epub 2004 Apr 9.

Related Arti  
l

**FREE** Full Text Article at  
[www.pnas.org](http://www.pnas.org)

**FREE** full text article  
in PubMed Central

**A gene atlas of the mouse and human protein-encoding transcriptomes.**

[Su AI](#), [Wiltshire T](#), [Batalov S](#), [Lapp H](#), [Ching KA](#), [Block D](#), [Zhang J](#), [Soden R](#), [Hayakawa M](#), [Kreiman G](#), [Cooke MP](#), [Walker JR](#), [Hogenesch JB](#).

The Genomics Institute of the Novartis Research Foundation, 10675 John J. Hopkins Drive, San Diego, CA 92121, USA.

## 79 human tissues, Affymetrix micro-array

☐ 1: [Mol Cancer Ther](#). 2007 Mar;6(3):820-32. Epub 2007 Mar 5.

**Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study.**

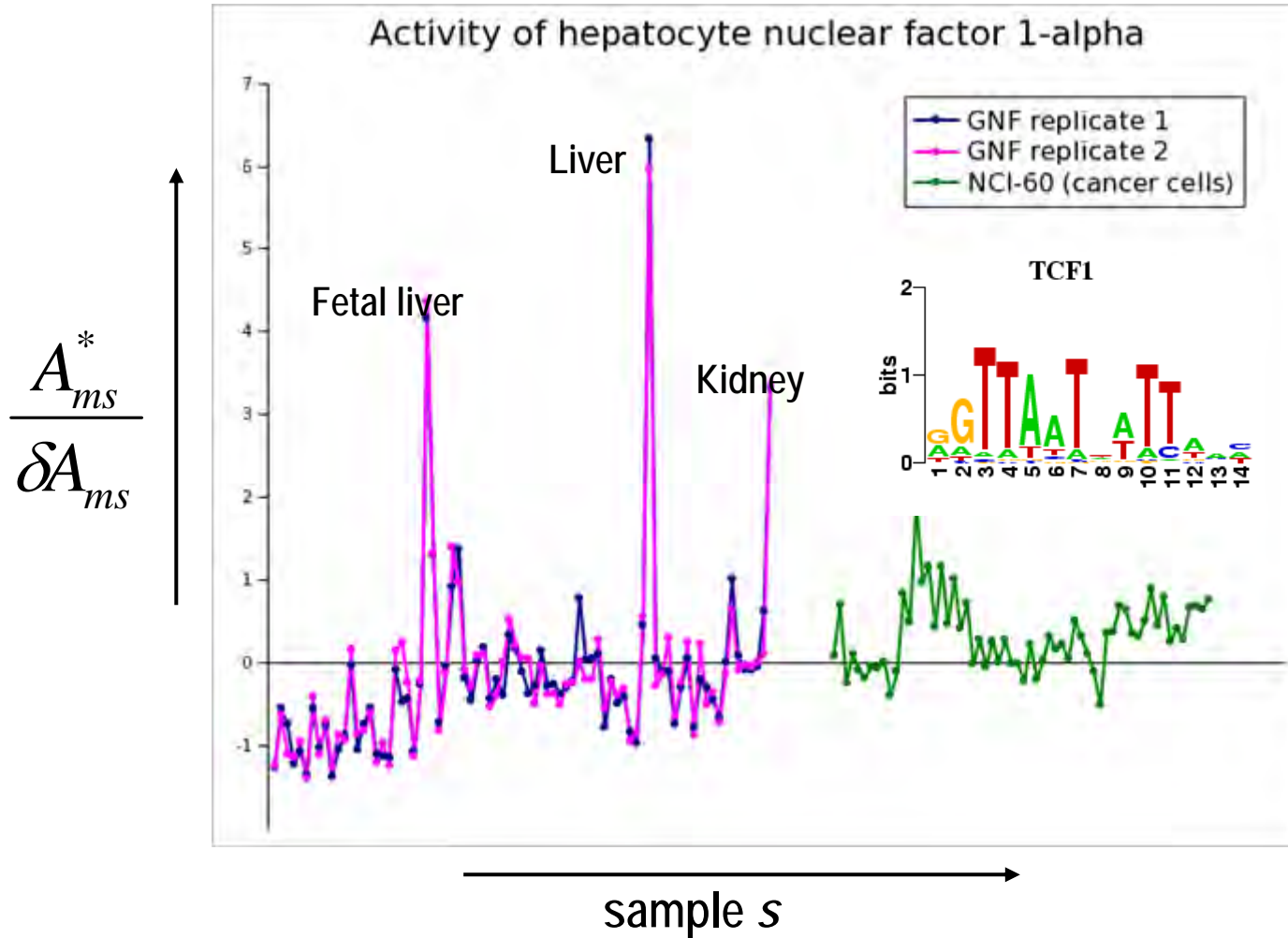
[Shankavaram UT](#), [Reinhold WC](#), [Nishizuka S](#), [Major S](#), [Morita D](#), [Chary KK](#), [Reimers MA](#), [Scherf U](#), [Kahn A](#), [Dolginow D](#), [Cossman J](#), [Kaldjian EP](#), [Scudiero DA](#), [Petricoin E](#), [Liotta L](#), [Lee JK](#), [Weinstein JN](#).

Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, Center for Cancer Research, National Cancer Institute/NIH, Bethesda, MD 20892, USA.

## 60 cancer cell lines, same Affymetrix micro-array

We associate probes with promoters and apply the same analysis to this data set.

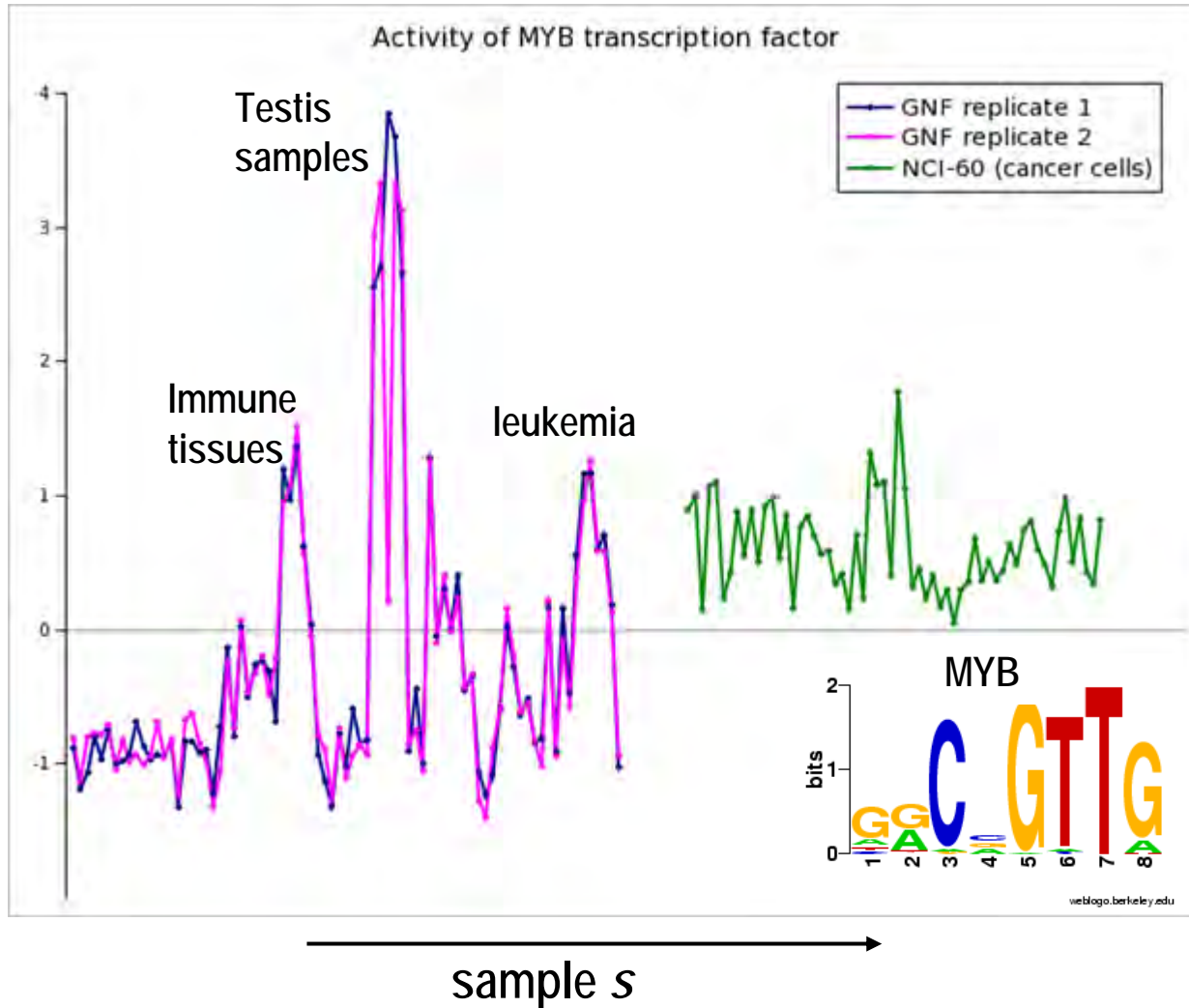
# In which samples is a given motif most active?



A known liver-specific factor indeed shows highest activity in liver tissues.

# In which samples is a given motif most active?

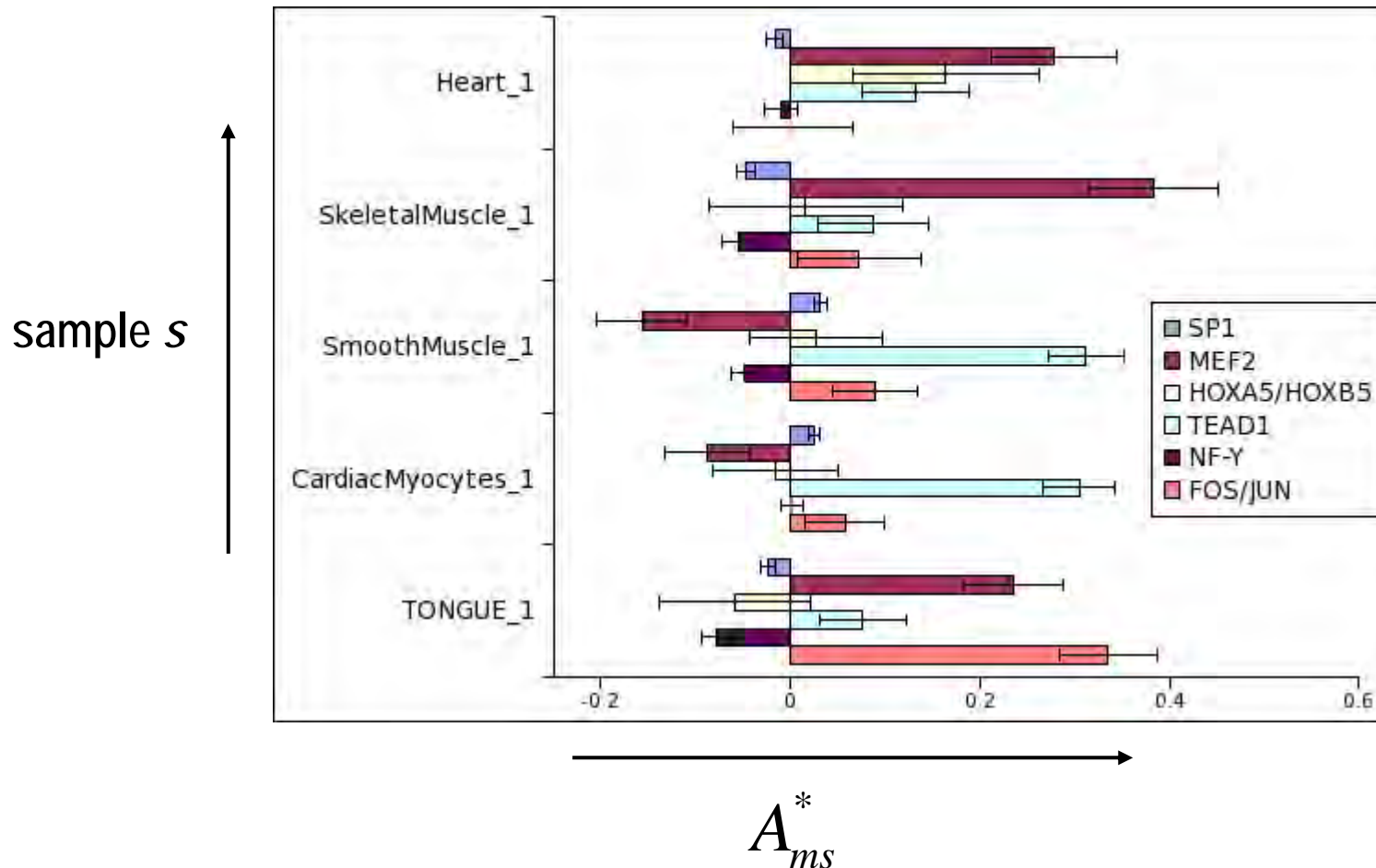
$$\frac{A_{ms}^*}{\delta A_{ms}}$$



MYB is high in testis. It is also up-regulated in *a//*NCI60 samples.

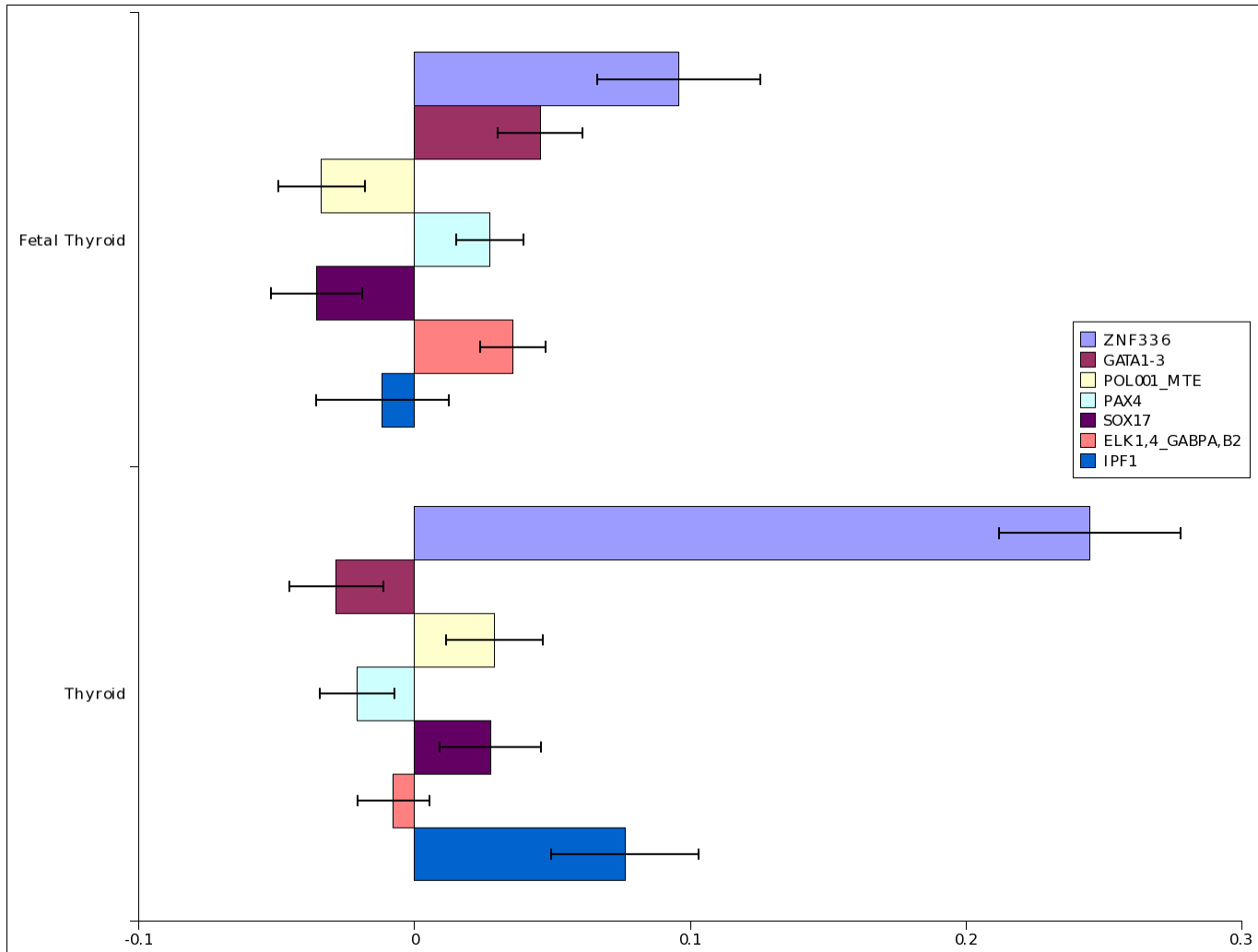
# Which motifs differentiate related tissues?

- We can focus in on a set of related tissues, e.g. **muscle tissues**, and determine which TFs vary most in activity across these tissues.



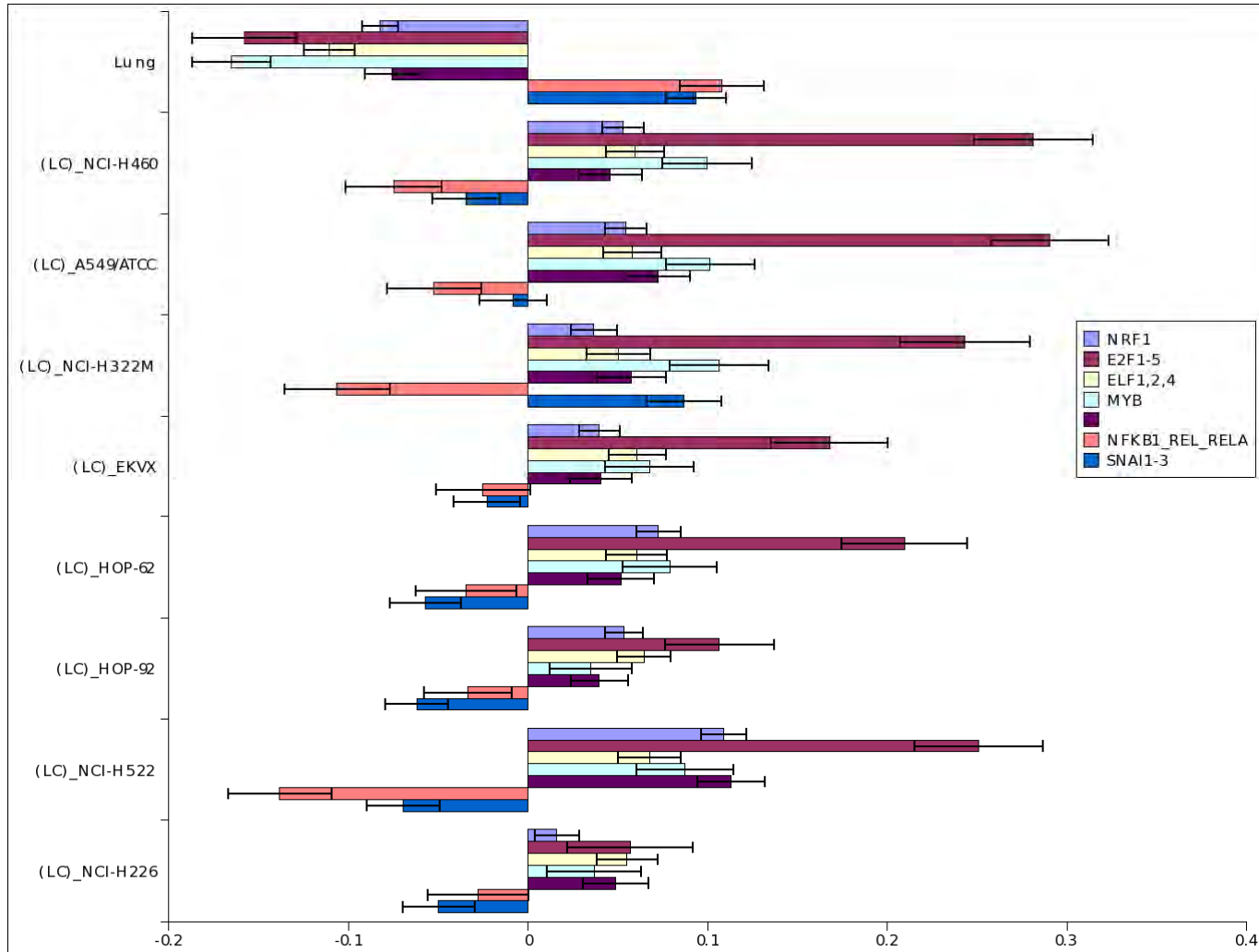
# Which motifs change in development of a tissue?

## Fetal thyroid and thyroid



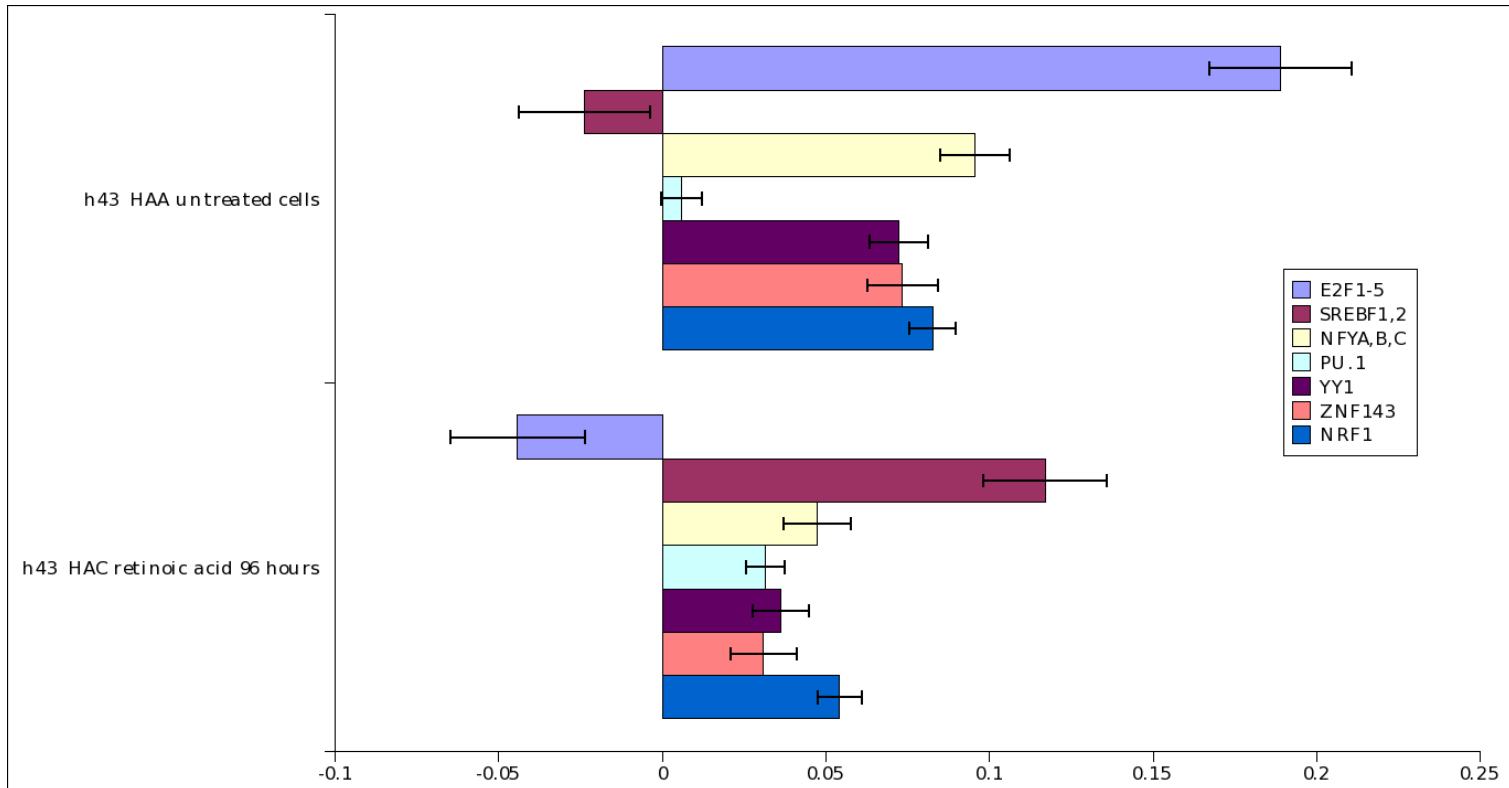
# Which motifs differentiate healthy from tumor tissues?

## Lung and lung tumors



# Which motifs change activity under a perturbation?

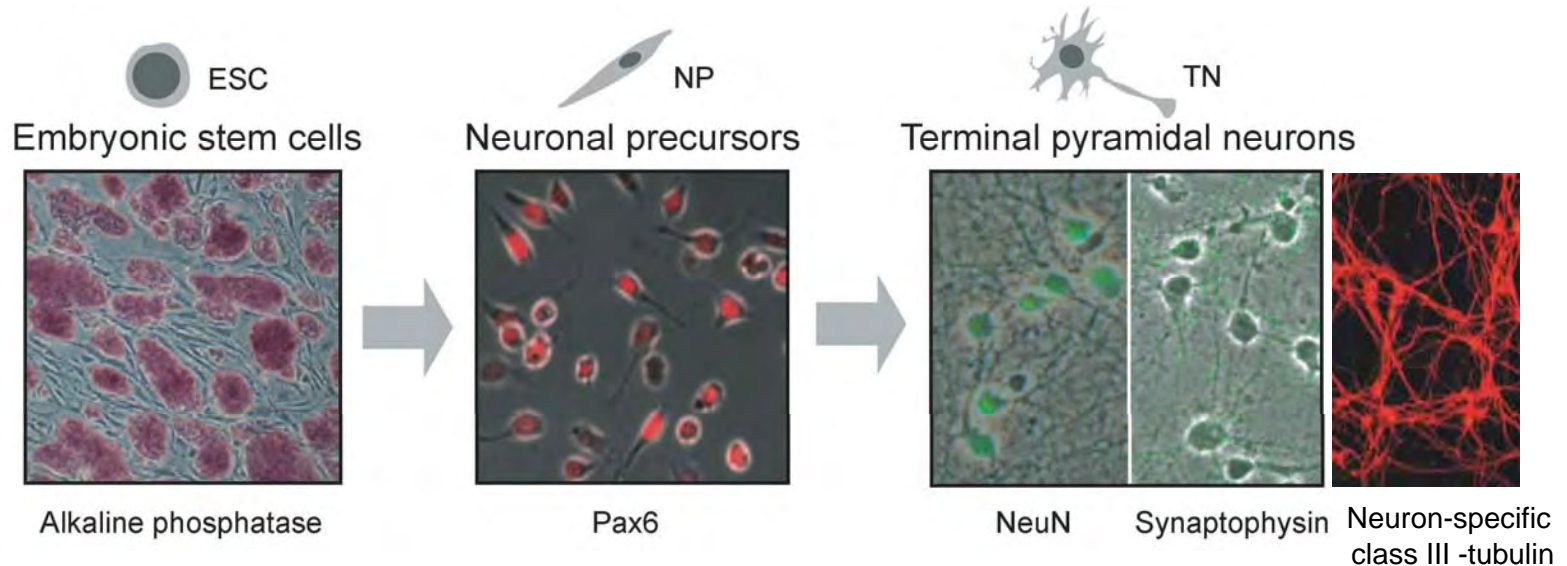
Monocytes before and after treatment with retinoic acid





# Example Application

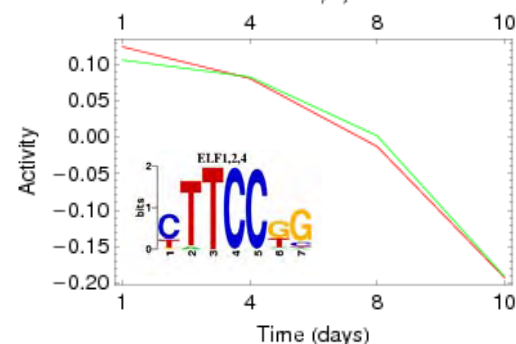
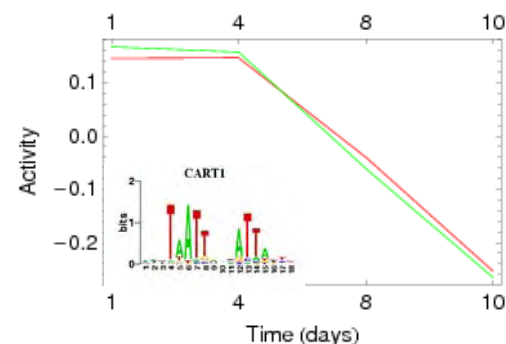
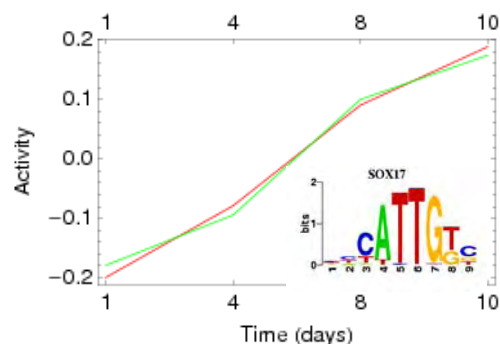
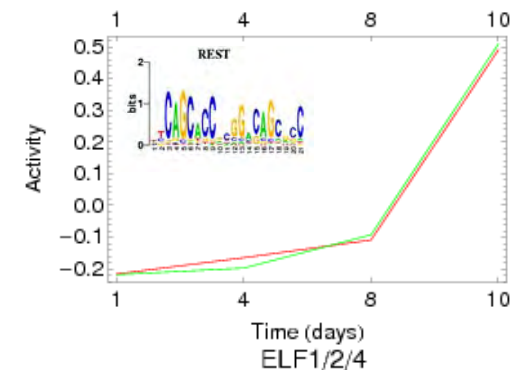
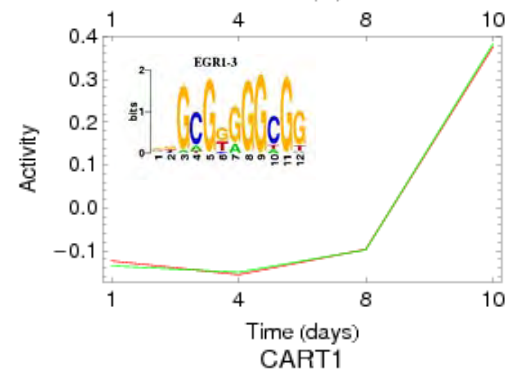
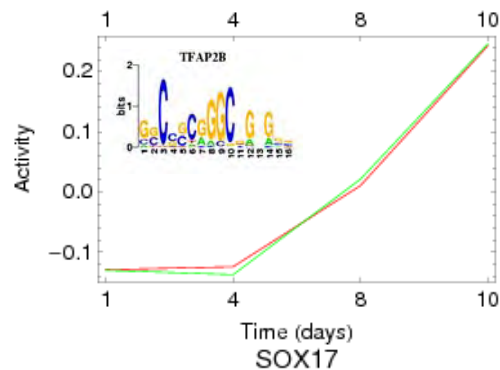
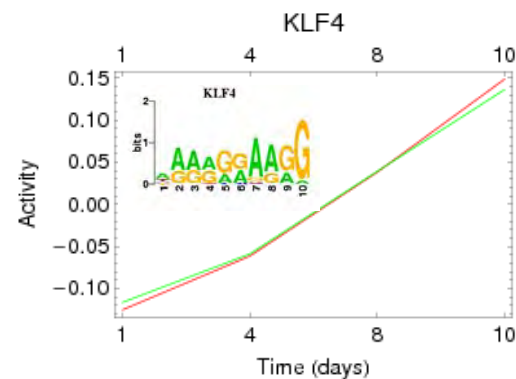
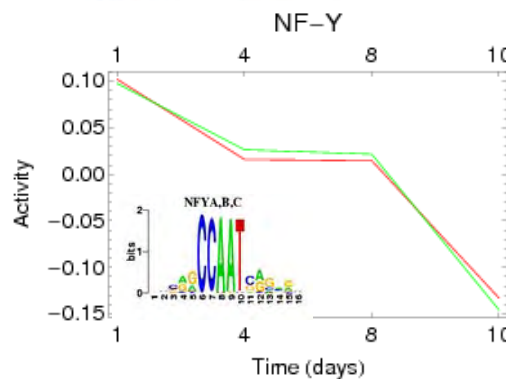
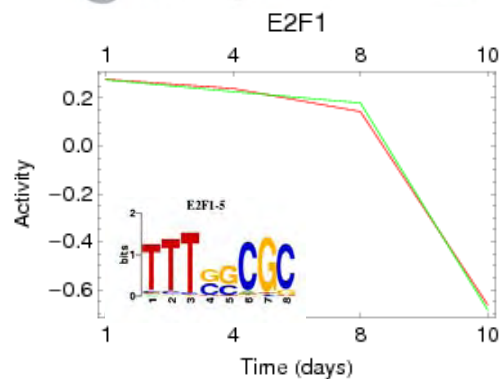
epigenetic reprogramming during terminal neuronal differentiation of murine stem cells *in vitro*



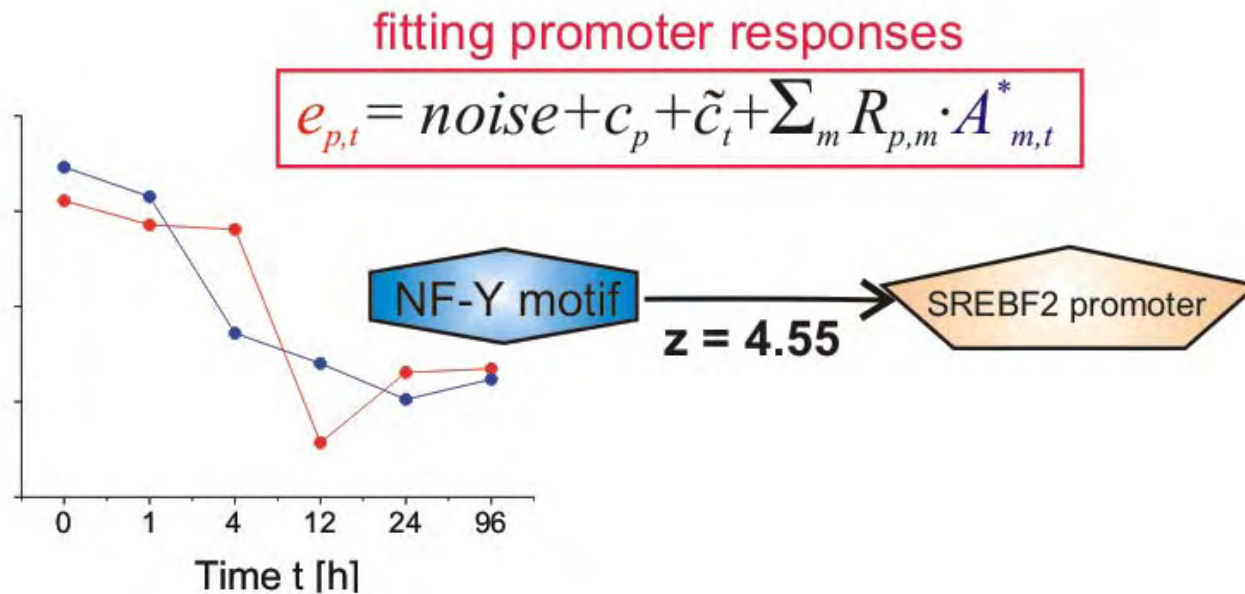
- Micro-array expression data at 4 time points (ESC, early NP, late NP, TN) in duplicate.
- Nimblegen human promoter chips.
- chIP-chip for methylated DNA, Polymerase II, H3K4me, and H3K27me (3 time points).

Collaboration with Dirk Schubeler, FMI, Basel

# Activities of the most significant motifs



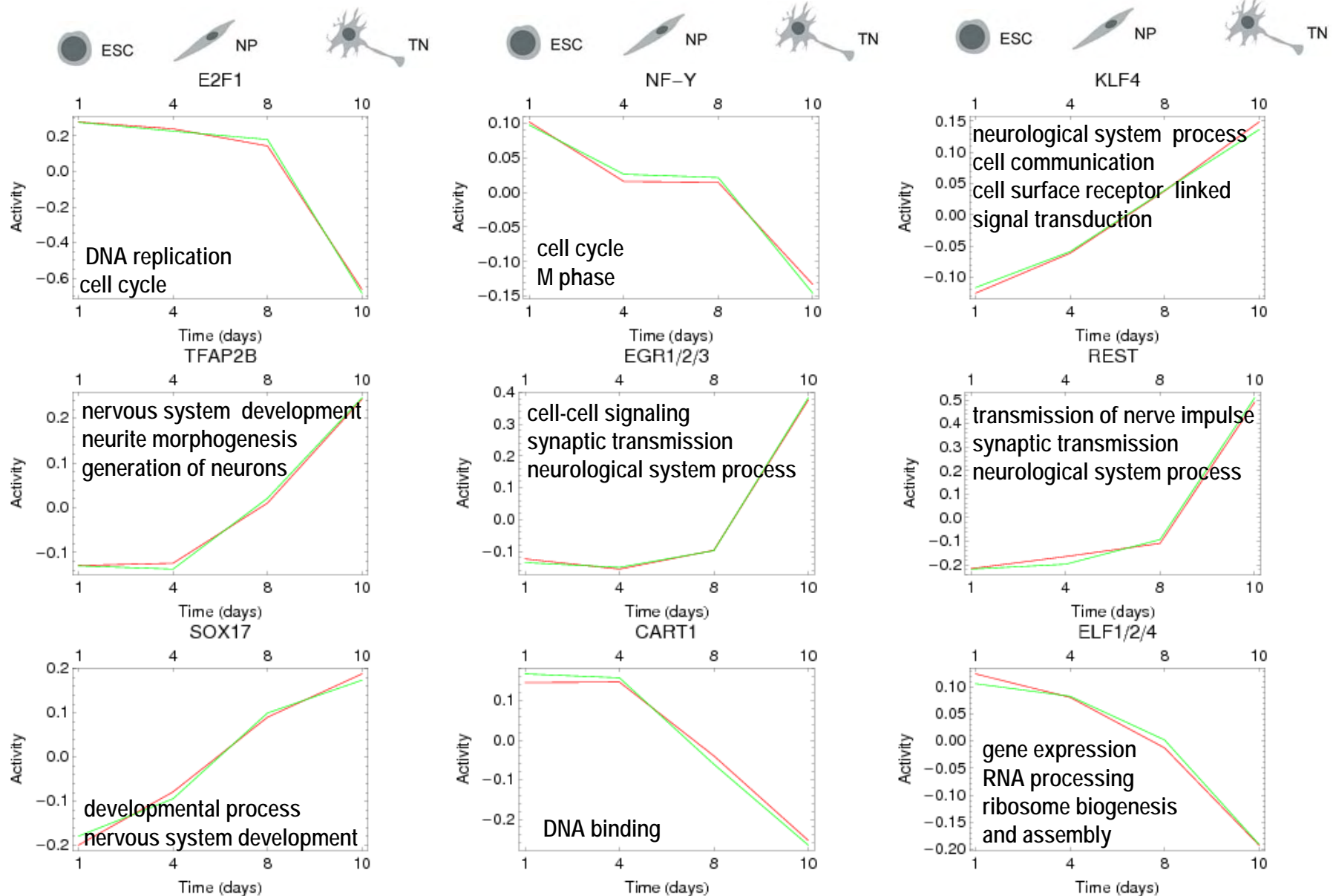
- For each motif go through list of all promoters with predicted TFBSs  $N_{pm} > 0$
- Investigate the correlation between *expression profile of the promoter* and *activity profile of the motif*.



Our final predictions of regulatory targets of each motif obey

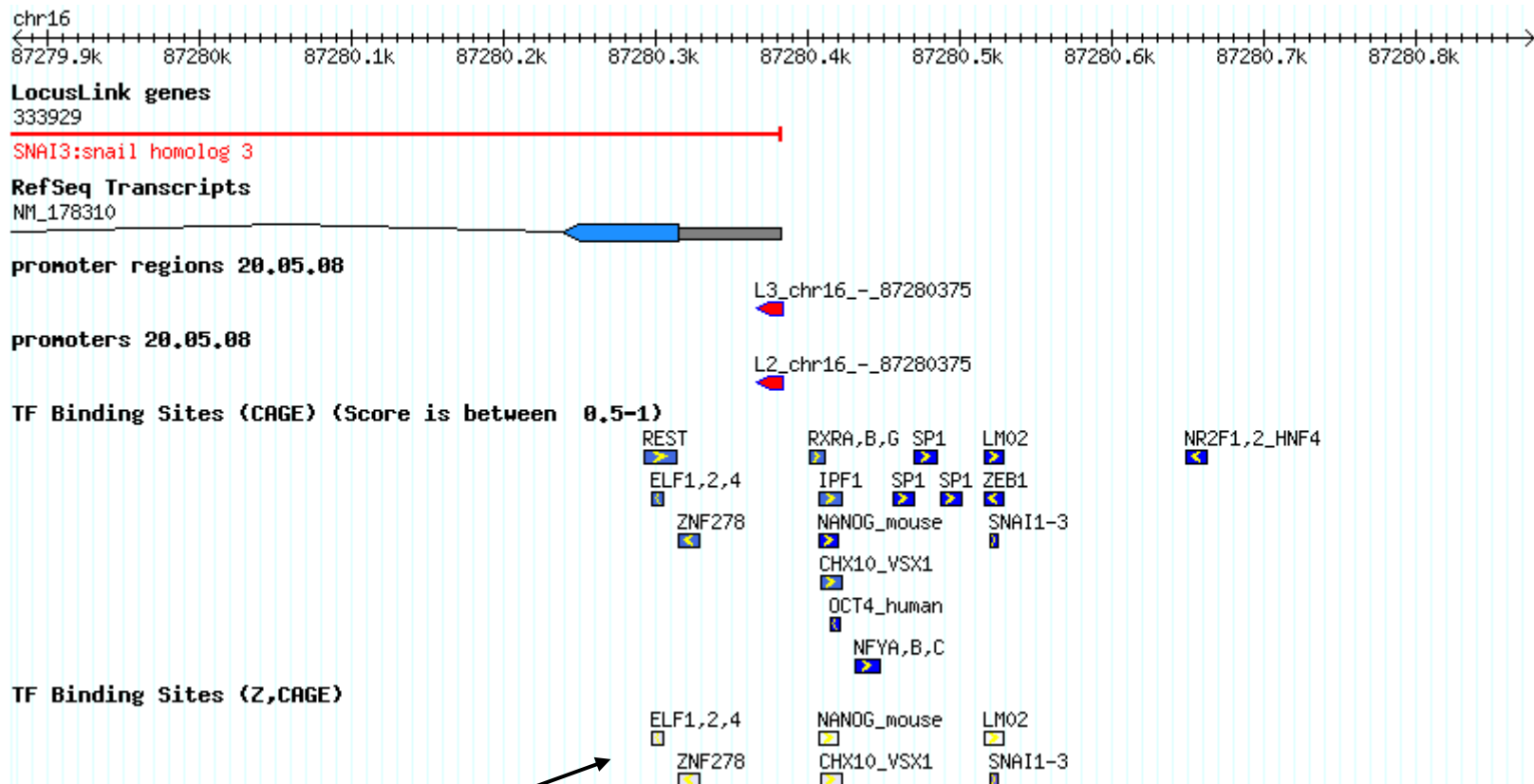
- The promoter has a predicted TFBS for the motif.
- The TFBS shows conservation and correct positioning w.r.t. TSS.
- The expression of the promoter significantly correlates with the activity profile of the motif.

# Targets of the most significant motifs: Association with Gene Ontology categories



# Predicted effects of expression of regulatory sites

**Example:** Predicted TFBSs in the proximal promoter of the SNAI3 TF.



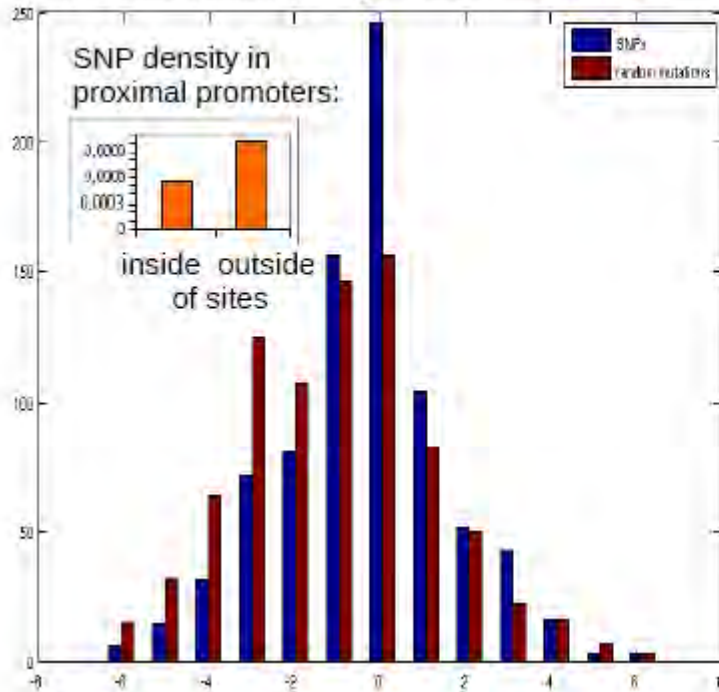
Z-values quantify correlation between motif activity and target expression.



# SNPs predicted to contribute to expression variation in humans

- We intersect the predicted TFBSs genome-wide with SNPs.
- **SNP-density** in TFBSs is almost **a factor 2 smaller** than in flanking regions (in proximal promoter).
- The **effect on WM-score** of the SNPs in TFBSs is clearly **lower** than effects of random mutations.

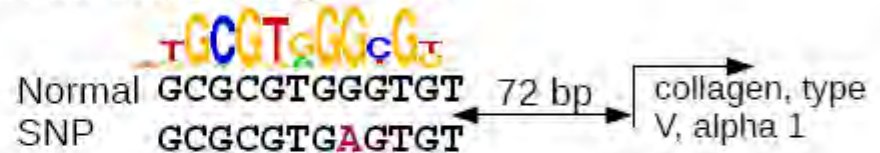
**A** TFBS score change under mutation



**B**

Egr-1 site in collagen, type V, alpha 1 gene promoter

TFBS score change: -6.68



PU.1 site in MAZ promoter

TFBS score change: -4.81



# Acknowledgments

## Biozentrum



## Omics Science Center

## RIKEN Institute, Yokohama, Japan



Yoshihide Hayashizaki



Harukazu Suzuki



Piero Carninci



Alistair Forrest



Carsten Daub



Friedrich Miescher Institute  
for Biomedical Research  
Part of the Novartis Research Foundation



**DKBW**  
Département Klinisch-Biologische Wissenschaften  
Universität Basel



Dirk Schübeler



Gerhard Christofori

