BIOZENTRUM
Universität Basel

SIB
Swiss Institute of Bioinformatics

Retracing transcription regulatory activities that control expression and chromatin dynamics

Erik van Nimwegen

Biozentrum, University of Basel,
and Swiss Institute of Bioinformatics

Transcription regulatory networks: Some imagination

Ultimate goal of reconstructing transcription regulatory networks:

• A physical model of the binding specificity of each transcription factor.

• Determined binding affinities genome-wide for each TF.

• Quantification of *TF activities* in given conditions of interest (expression, nuclear localization, post-translational modification, co-factor presence, etcetera.)

• A physical model that predicts genome-wide TF binding patterns in terms of sequence and TF concentrations.

• A physical model of the cis-regulatory logic, i.e. a quantitative mapping of TF binding configuration to quantitative effects on transcription of targets, and on local (and distal?) chromatin modifications.

• A dynamical model that integrates all this into a model of expression dynamics.

Transcription regulatory networks: Some reality

- Mammalian cells have wildly varying morphology, and behavior, but they all share the same DNA.

- We imagine that global gene expression patterns are key determinants of cellular state (to what extent there are clearly discrete states is still unclear).

- We imagine that gene expression state is controlled by the activities of TFs which are binding to sequence-specific binding of TFs to sites in the DNA (in a manner that may depend on and effect local chromatin state).

- In mammals there are about 2000 TFs that potentially play a role in controlling cell identity.

- For the large majority of cellular states and differentiation pathways we know next to nothing about which regulators are key in controlling expression and chromatin dynamics.

- We cannot do extensive genetic screens for every system.

BIOZENTRUM
Universität Basel

SIB
Swiss Institute of Bioinformatics

Accelerating regulatory network reconstruction

Provide automated analysis of high-throughput data
that tells where to invest detailed experimental effort.

Develop a computational frame-work that:

- Starts from genome-wide measurements of expression or chromatin state dynamics (micro-arrays, RNA-seq, ChIP-seq) for a system of study.

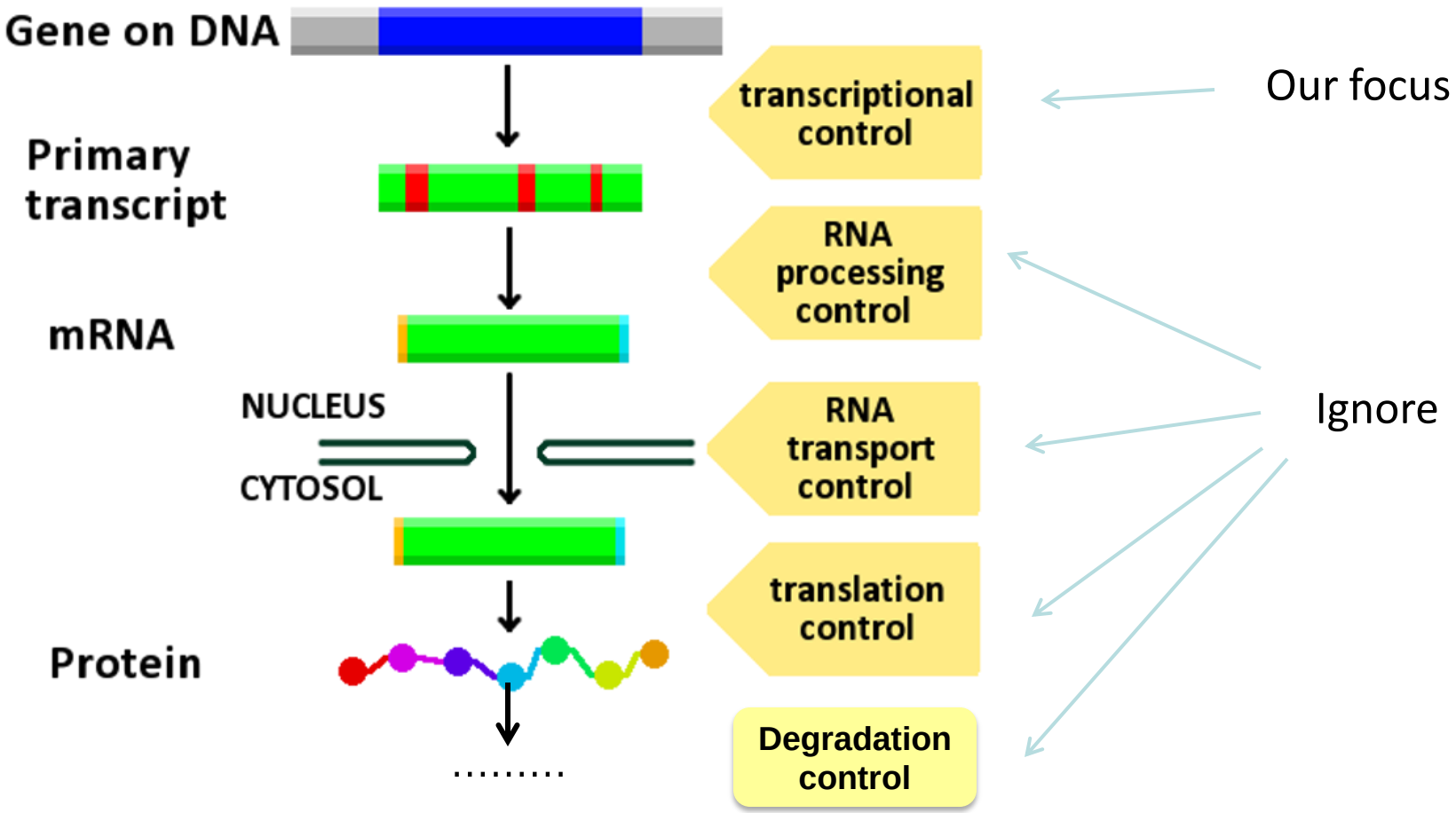
- Predicts the transcription regulators that play a key role in the process under study (developmental time course, response to perturbations, disease versus healthy tissue).

- Predicts how the regulators change activity (up-regulation, down-regulation, transient changes).

- Predicts the target gene sets of the key regulators.

- Predicts the cis-regulatory elements on the genome through which the regulators act.

Levels of gene expression control in eukaryotes



Regulation *of transcription initiation* is a key event in regulation of gene expression.

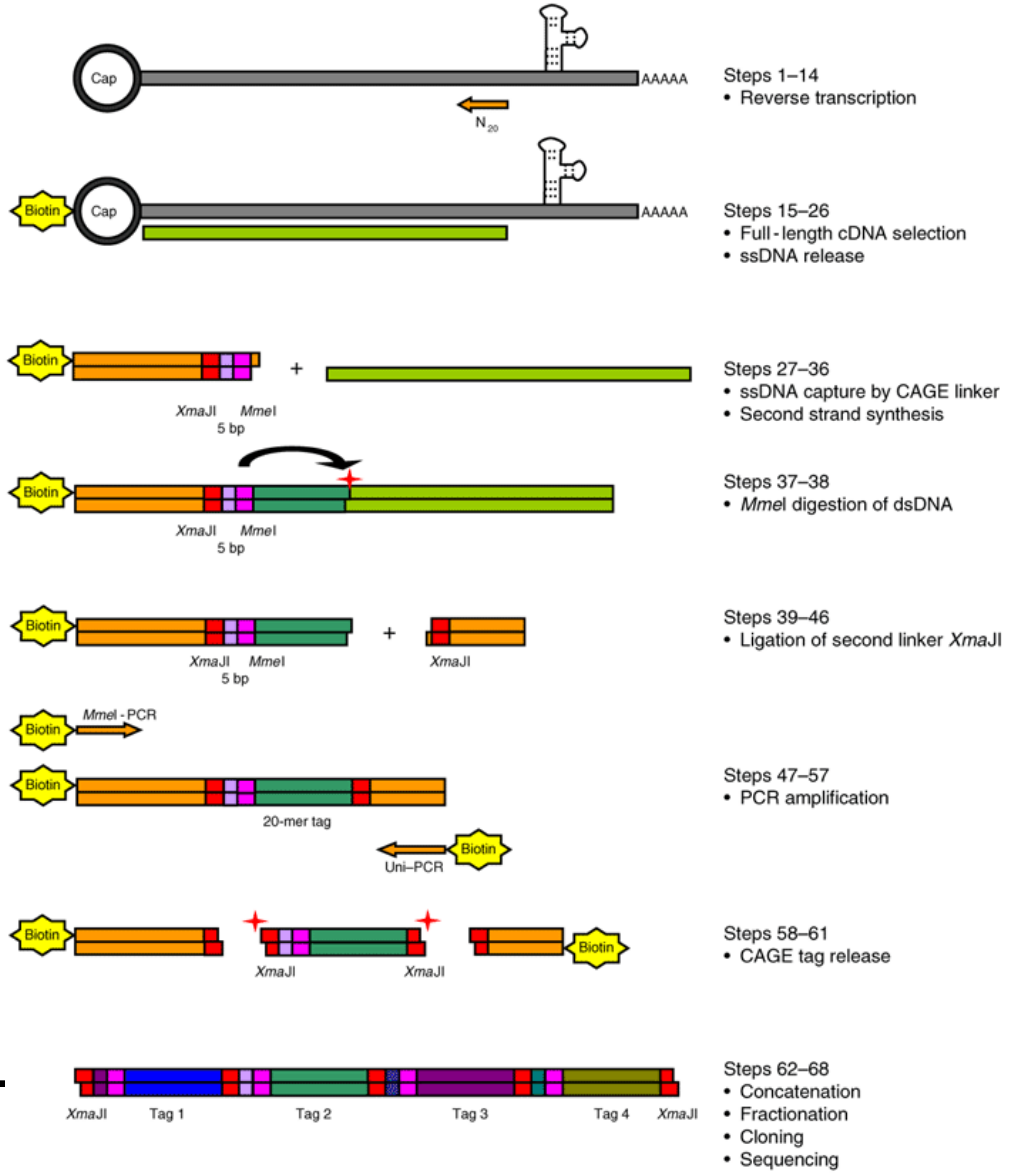
Question: Where is transcription initiated genome wide?

Deep sequencing of 5’ ends of mRNAs: deepCAGE technology

Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.
Shiraki et el. *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)



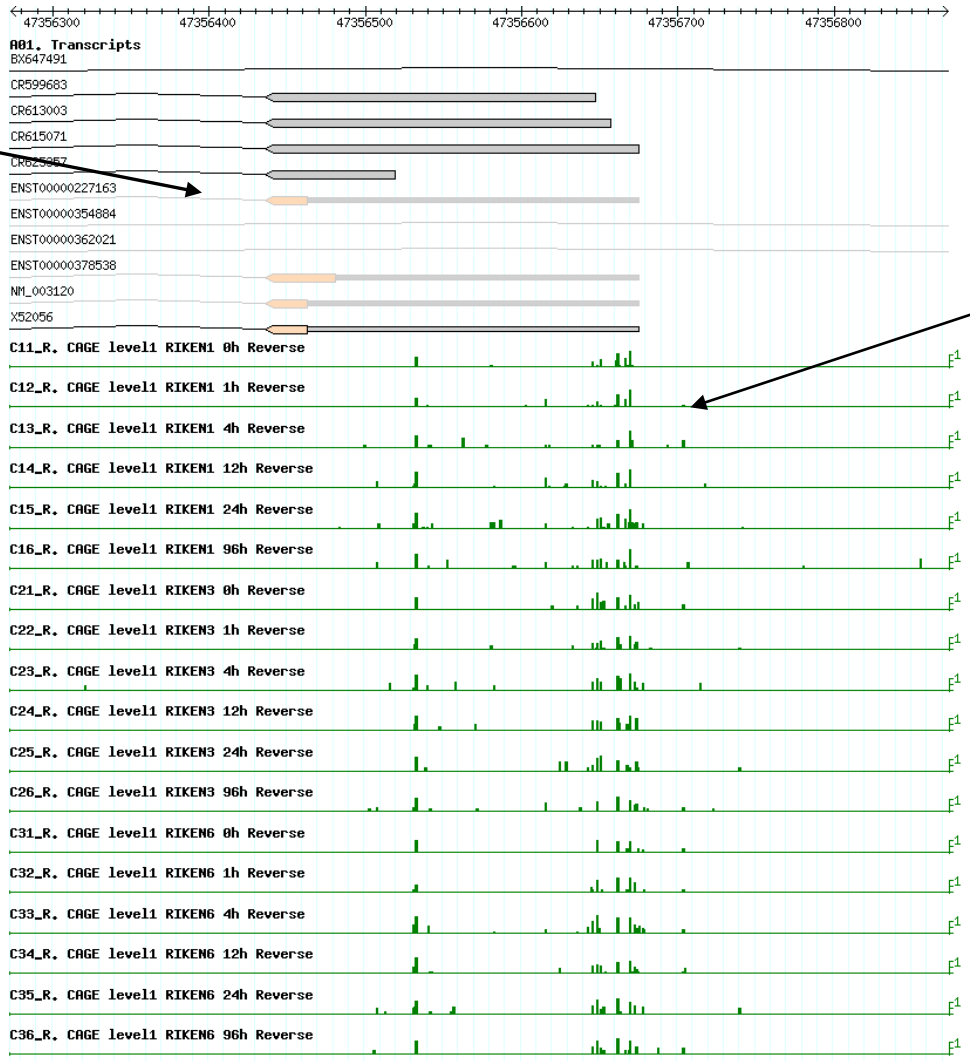
454/Solexa/Helicos sequencing.
Mapping to the genome.

Constructing `promoters'

Known transcripts

Tag counts at each position

Triplicate time course

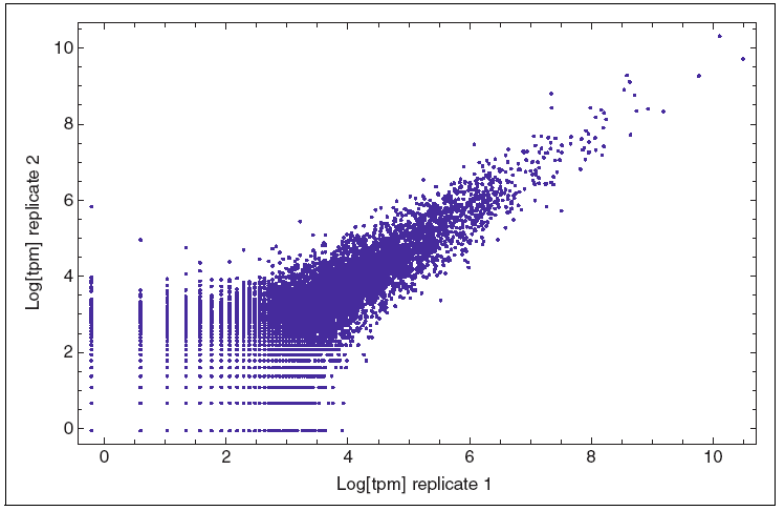


Transcription Start Cluster:

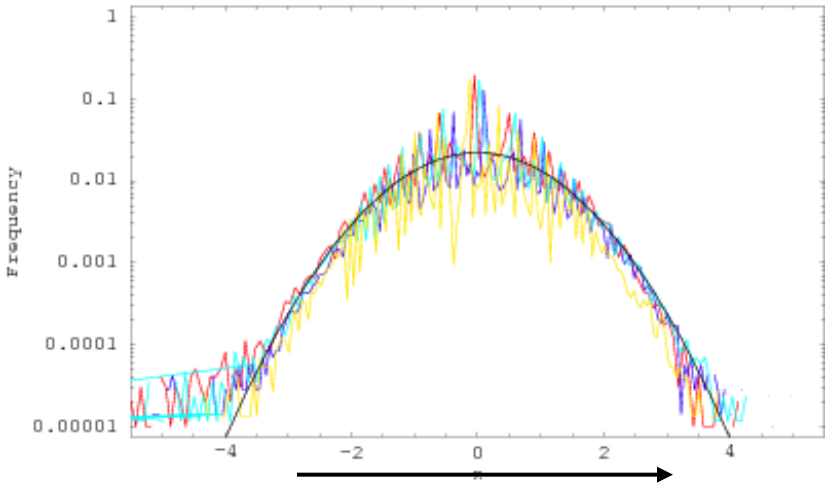
Cluster of neighboring TSSs that are co-expressed (within measurement noise).

Noise-model for CAGE expression data

Biological replicates:
log-tag count in one replicate versus another.



Observed and predicted replicate noise



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

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

$x = \text{true log-expression of the TSS.}$

$y = \text{log-expression in the sample}$

$$P(y \mid x, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-y}{\sigma}\right)^2\right]$$

$N = \text{total number of tags sequenced.}$

$t = \text{number of times TSS sequenced.}$

$$P(t \mid y) = \frac{(Ne^y)^t}{t!} e^{-Ne^y}$$

DeepCAGE data analysis: constructing human and mouse `promoteromes´

1: Genome Biol. 2009;10(7):R79. Epub 2009 Jul 22.

Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data.

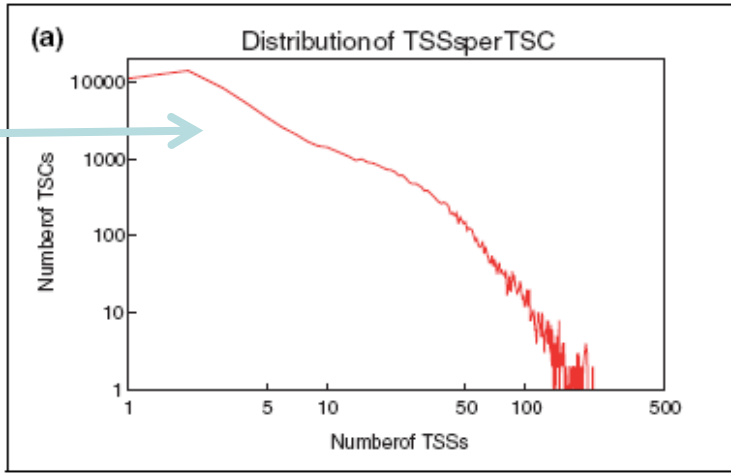
Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E.

Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Klingelbergstrasse 50/70, 4056-CH, Basel, Switzerland.

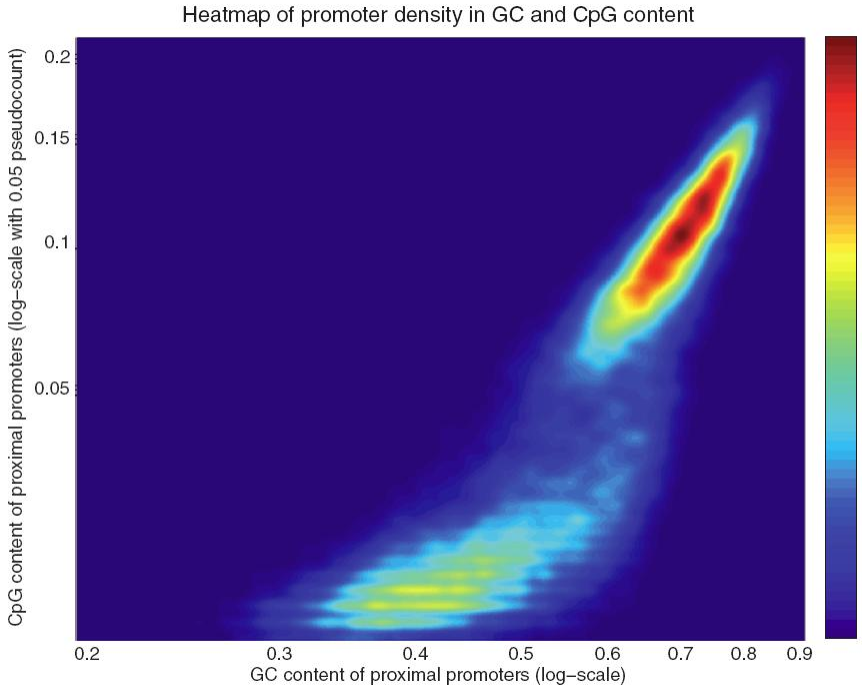
Basic promoterome stats:

Statistic	Human	Mouse
Number of samples	56	66
Number of TSSs in TSCs	860’823	608’474
Number of TSCs	74’273	77’286

There appear to be two regimes in the number of TSSs per TSCs.

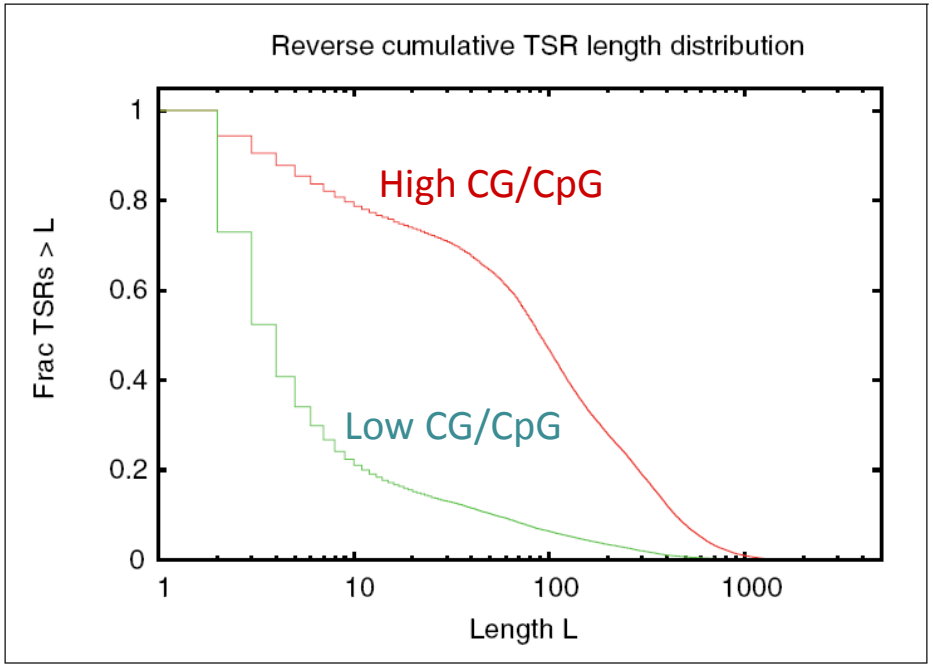


CpG and non-CpG promoters



Promoters can be separated into two classes: *high CG/CpG* and *low CG/CpG*.
CpG promoters are wide with many TSSs.
Non-CpG promoters are narrow with few TSSs.

These results confirm previous observations based on the FANTOM3 mouse data:



1: Nat Genet. 2006 Jun;38(6):626-35. Epub 2006 Apr 28.

Erratum in:
Nat Genet. 2007 Sep;39(9):1174.

Comment in:
Nat Genet. 2006 Jun;38(6):608-9.

Genome-wide analysis of mammalian promoter architecture and evolution.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engström PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y.

Genome Exploration Research Group, RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan.

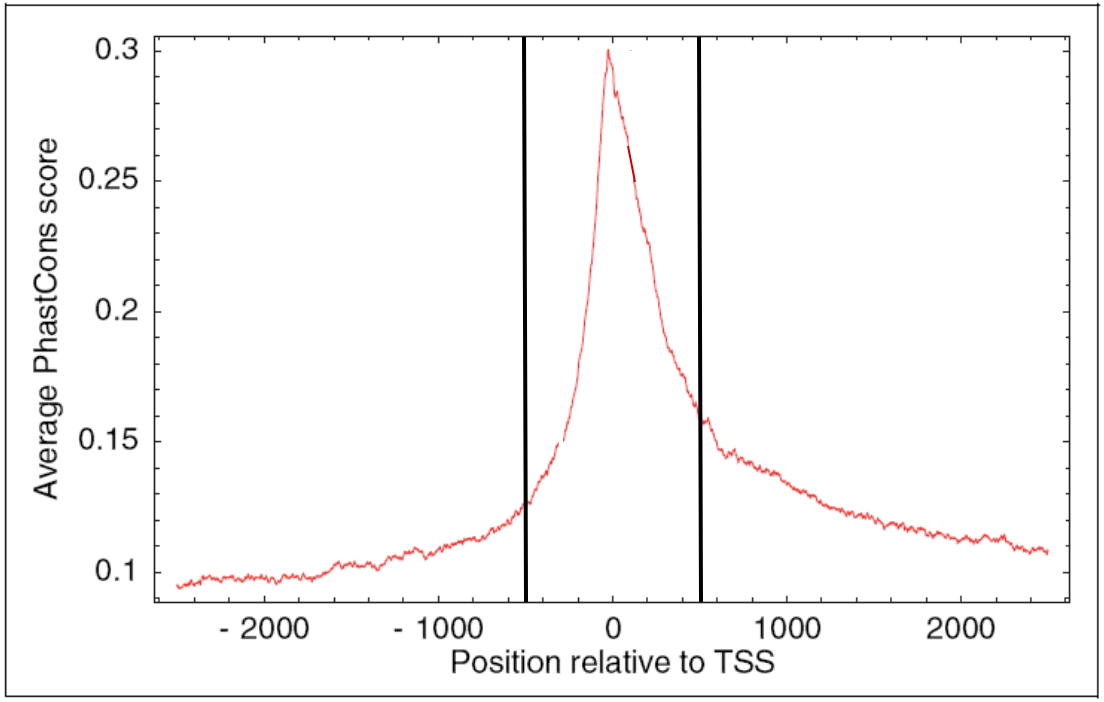
Conservation statistics around TSCs: Proximal promoters

1: Genome Res. 2005 Aug;15(8):1034-50. Epub 2005 Jul 15.

Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D.

Center for Biomolecular Science and Engineering, University of California, Santa Cruz, Santa Cruz, California 95064, USA. acs@soe.ucsc.edu



PhastCons: probabilistically classifies positions as conserved/not-conserved based on vertebrate genome alignments.

- There is a narrow peak of conservation around TSCs.
- This indicates that conserved DNA signals around TSS are concentrated in a small region.
- We choose (-500,+500) as a conservative estimate of the *proximal promoter*.

Prediction of transcription factor binding sites

• The basic hypothesis is that transcription regulatory patterns are ultimately driven by regulatory sites in the DNA that are recognized by transcription factors.

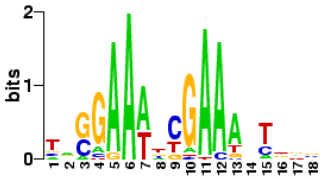
• The next step in analyzing regulation of gene expression is thus to identify regulatory sites genome-wide.

• For the moment we will **focus on proximal promoter regions only**.
Main reason: we have no reliable way of identifying relevant enhancers genome-wide. We hope sites for the relevant factors are often at the promoters as well.

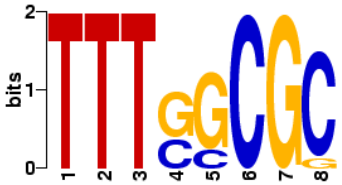
Predicting TFBSs in all proximal promoters

Input:

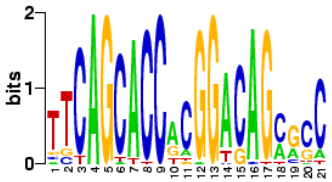
- A curated collection of 189 mammalian **regulatory motifs** (weight matrices) representing 340 human TFs.



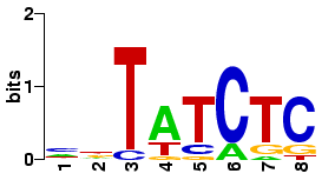
IRF7



E2F



REST



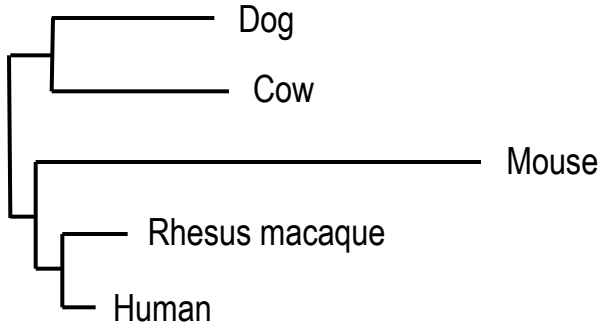
GATA2/4

All **proximal promoter regions** (-500,+500) with respect to each TSS.

- Multiple alignments with **orthologous regions** from 6 other mammals.

```
CATTCGCAGTGGCAAGGGACTGCCCTGGTCCCTGTGGAGC-GTCCCATTCGGTGACTTCCCACCAGCCCTTCCCCAGCGCCTCTGGAGGTCCAGACTGTCAGGTTGGAGCCTGGG
CATTCACAGTGGCAAGGGTCCGCCCTGGTCCCTGTGGAGG--GTCCCAGTCGGTGACTTCCCGCCAGCCCTTCCCCAGTGCCTCTGGAGGTC--GACTGTC-GGTTGGAGCCTGG
GAGGGGCGG---CTCGGGAGG---------CCTGCGGACC--GGGCGAG-CGGGGGCG-GCG----GGGCGGCGGGGGAGCCGGGCGGGGGCC------TGCGGTCGG-GCCTGG
GATTGGCCGCGGCCAAGGACCCC-----TCCCTGGGGAGC--GTCCGGGTCGGAGACT-CCCACTTGCCCTTCTCCAGCACCTCGTGAAGTCCGGACTGTACGGTTTG-GACTCG
TATCTACAACAGCAAG-GA--------GTC--TG-GAAGCAAGTCCAAGT-GATGGA-TACAGCCATCACTTACC--GGGCCTCTGCTGGTCGTGACTT----------------
```

- The phylogenetic tree relating the species:



MotEvo Algorithm

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

```
Scer AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATC-GAAACATACATAA--GTTGATATTC-CTTTGATATCG-----ACGACTA
Spar AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATC-GAAACATACATAA--ATTGATATTC-CTTTAGCTTTT----AAAGACTA
Smik GAAAAACGAAAAATTCATG-GAAAAGAGTCAACCGTC-GAAACATACATAA--ACCGATATTT-CTTTAGCTTTCGACAAAAATCTG
Sbay GAAAAATAAAAAGTGATTG-GAAAAGAGTCAGATCTCCAAAACATACATAATAACAGGTTTTTACATTAGCTTTT----GAAAACTA
```

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

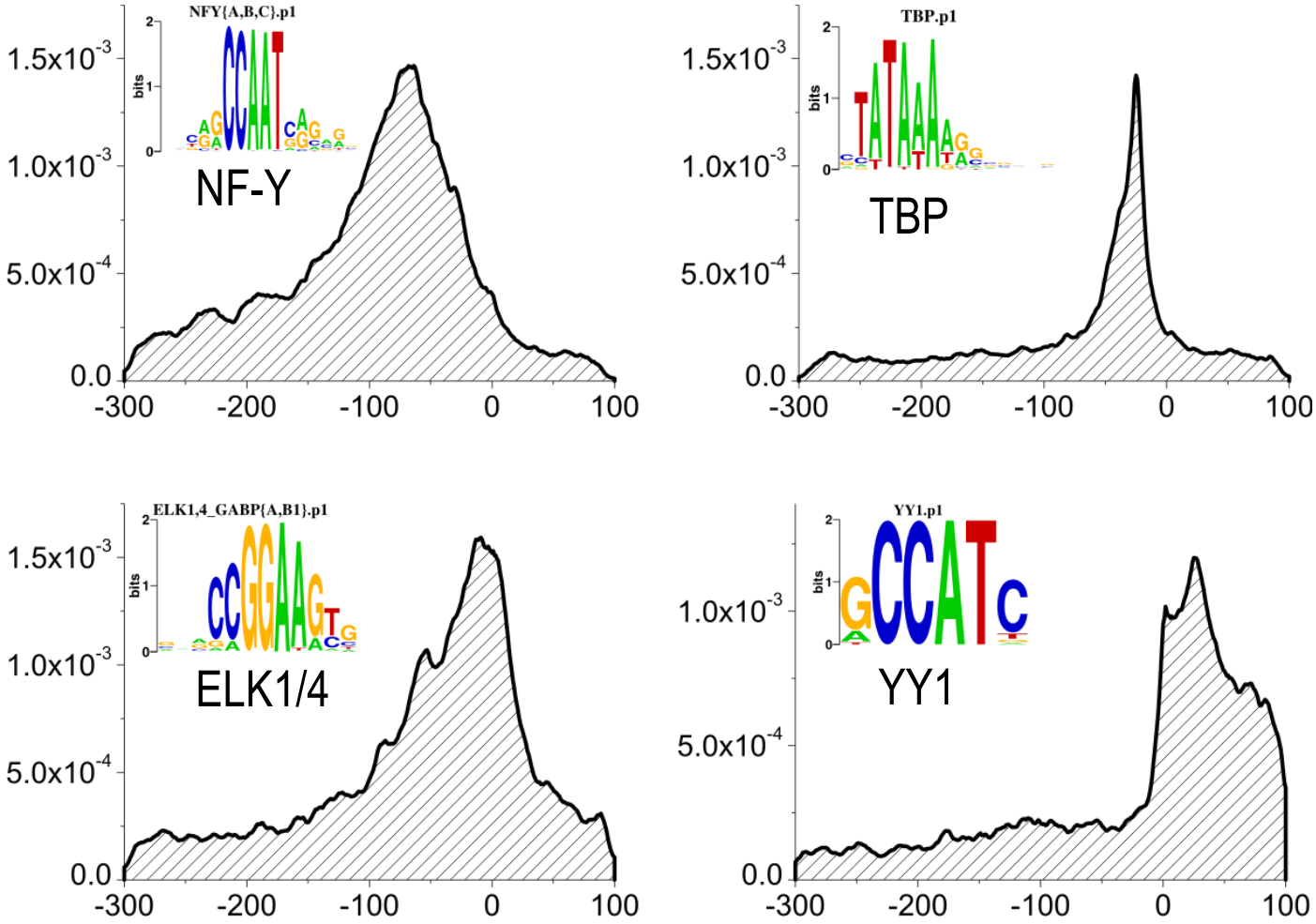
```
Scer AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATC-GAAACATACATAA--GTTGATATTC-CTTTGATATCG-----ACGACTA
Spar AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATC-GAAACATACATAA--ATTGATATTC-CTTTAGCTTTT----AAAGACTA
Smik GAAAAACGAAAAATTCATG-GAAAAGAGTCAACCGTC-GAAACATACATAA--ACCGATATTT-CTTTAGCTTTCGACAAAAATCTG
Sbay GAAAAATAAAAAGTGATTG-GAAAAGAGTCAGATCTCCAAAACATACATAATAACAGGTTTTTACATTAGCTTTT----GAAAACTA
```

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

```
Scer AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATC-GAAACATACATAA--GTTGATATTC-CTTTGATATCG-----ACGACTA
Spar AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATC-GAAACATACATAA--ATTGATATTC-CTTTAGCTTTT----AAAGACTA
Smik GAAAAACGAAAAATTCATG-GAAAAGAGTCAACCGTC-GAAACATACATAA--ACCGATATTT-CTTTAGCTTTCGACAAAAATCTG
Sbay GAAAAATAAAAAGTGATTG-GAAAAGAGTCAGATCTCCAAAACATACATAATAACAGGTTTTTACATTAGCTTTT----GAAAACTA
```

- MotEvo considers all possible ways into which the multiple alignment can be partitioned into:
  - Background columns: Evolve neutrally.
  - Binding sites for one of the motifs: Constrained by the WM.
  - Unknown functional elements: Unknown constraints.
- Forward/backward algorithm to sum over possible configurations.
-
- *Posterior probability* for site occurrence at each position and each motif.

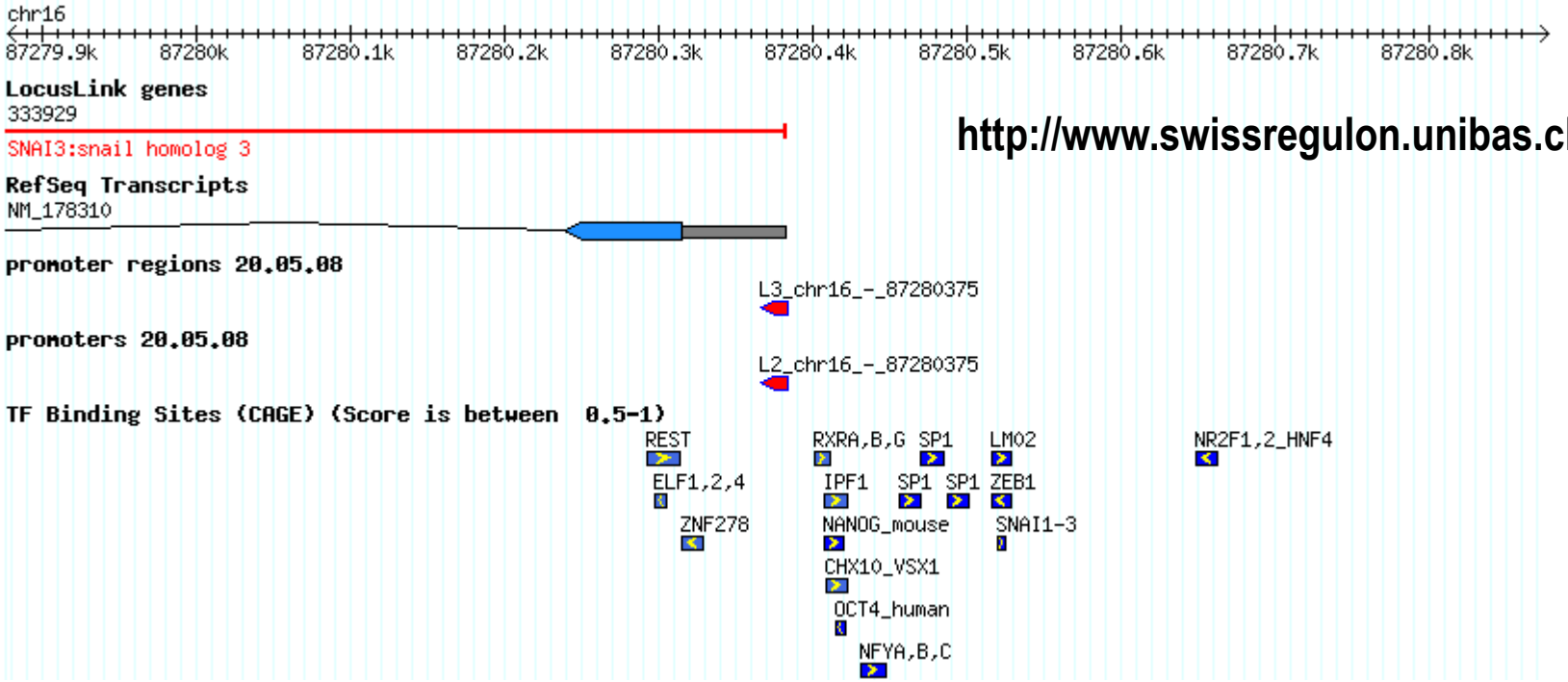
TFs have distinct positional preferences with respect to TSS



The positional preferences of each motif are incorporated in the site predictions (the positional profile is iteratively estimated and used as a prior).

Genome-wide annotation of regulatory sites in proximal promoters

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



http://www.swissregulon.unibas.ch

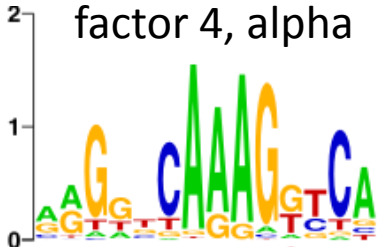
Summarizing the TFBS predictions:

For each promoter *p* and motif *m* we sum the posteriors of the predicted sites to obtain a matrix of site-counts: N_{pm}

Polymorphisms in sites

HNF4A – hepatochte nuclear factor 4, alpha

z=5.6



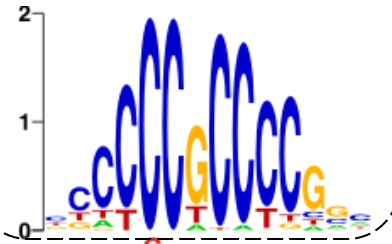
FOXA2 →

GAAACACGCCGGCCTGAATATCAGAG[A/G]CAAATCTCAGCCTCCCAACCGTCGGCCGCTGCTAGAGGGG

18% G, 82% A

SP1

z=8.4



TNFRSF1B →

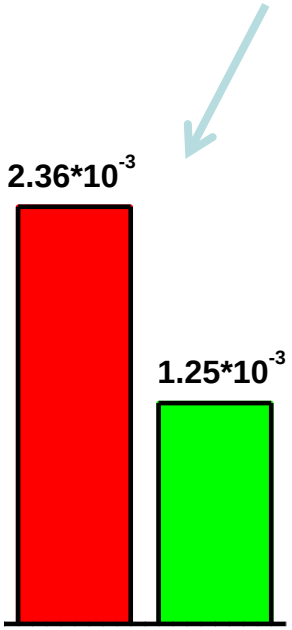
TGGCCCCGCCC[C/T]GCCCCGCCGCCTGGCCT ⟷ 50 bp ⟷ CAGCGGAGCCTGGAGAGAAGGCGCTGGGCTGCGAGGG

96% C, 4% T

Polymorphisms in sites

SNPs are biased to conserve binding site affinity

SNP density is lower at sites compared to flanking positions.



BIOZENTRUM
Universität Basel

SIB
Swiss Institute of Bioinformatics

Leveraging the genome-wide binding site predictions

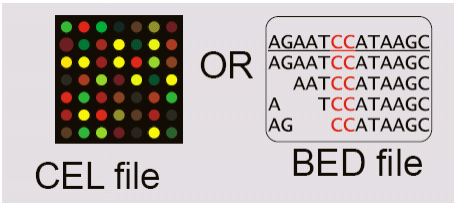
- We have the positions of transcription initiation clusters genome-wide.
- We have extracted the proximal promoters around each promoter.
- We have predicted regulatory sites for a large number of motifs (TFs): N_{pm}

- We will use the predictions to analyze genome-wide dynamics of expression and chromatin state.

- We first focus on analyzing expression dynamics.

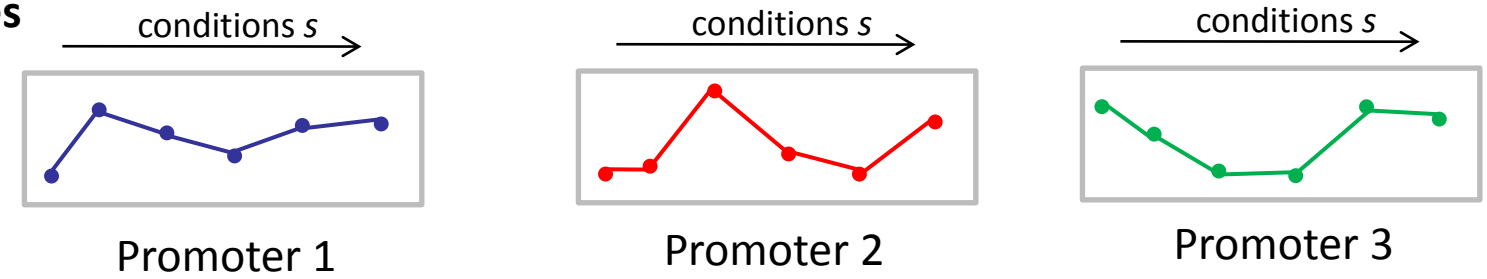
Quantifying expression data

Input:



Expression profiles

e_{ps}



Motif Activity Response Analysis

$$e_{ps} = \text{noise} + \sum_m N_{pm} A_{ms}$$

Expression of promoter *p* in sample *s*

Number of functional sites in promoter p for motif m

Activity of motif *m* in sample *s*

Linear model: Expression is weighted sum of the *activities* of all TFs that have binding sites in the promoter.

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

Bayesian inference of the motif activities (simple using SVD):

$$P(A_{ms} \mid e, N) \propto \exp\left(-\frac{1}{2}\left(\frac{A_{ms} - A^*_{ms}}{\delta A_{ms}}\right)^2\right)$$

Significance of motif *m*:

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

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.

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.

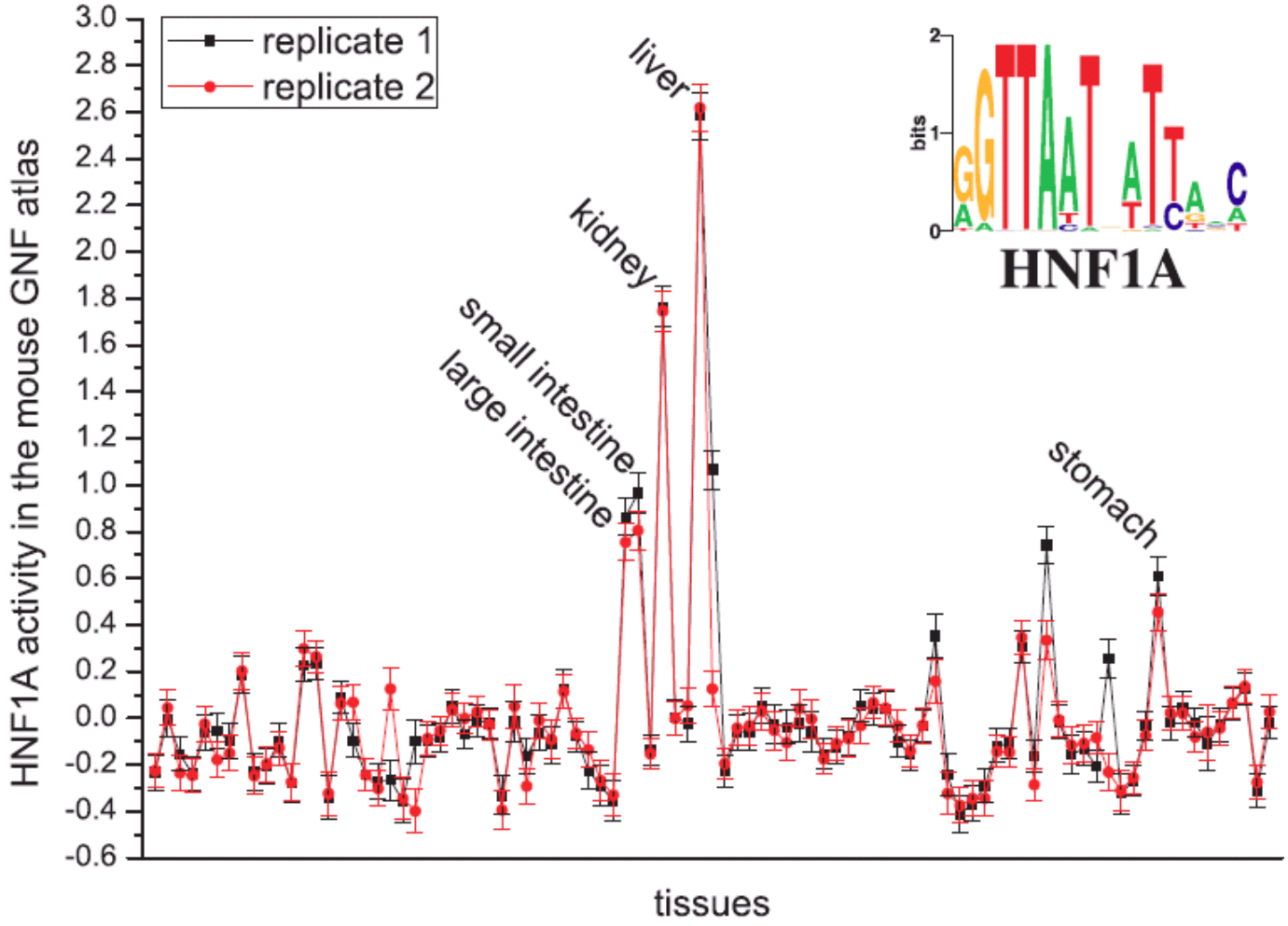
Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study.

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.

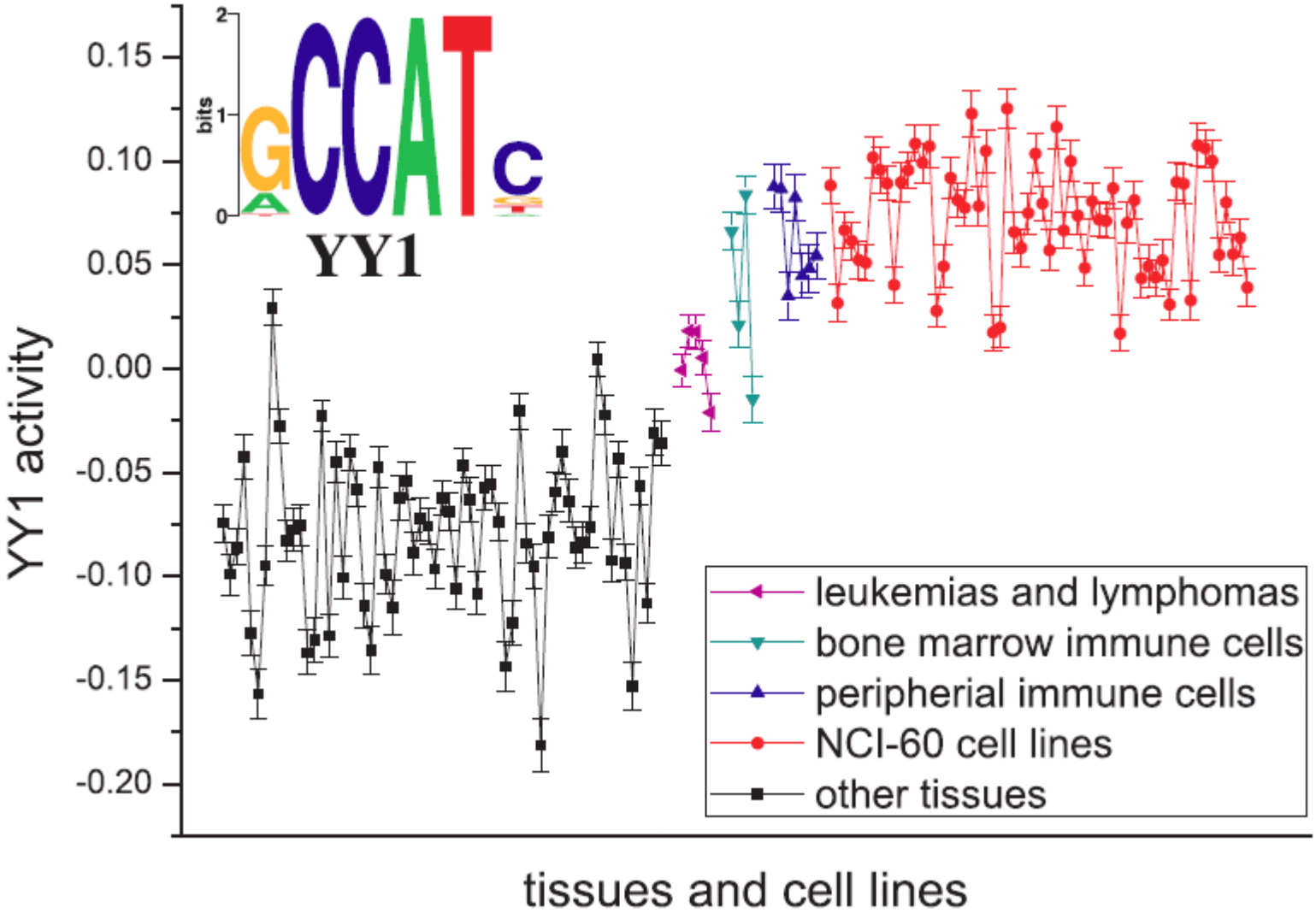
60 cancer cell lines, same Affymetrix micro-array

In which samples is a given motif most active?

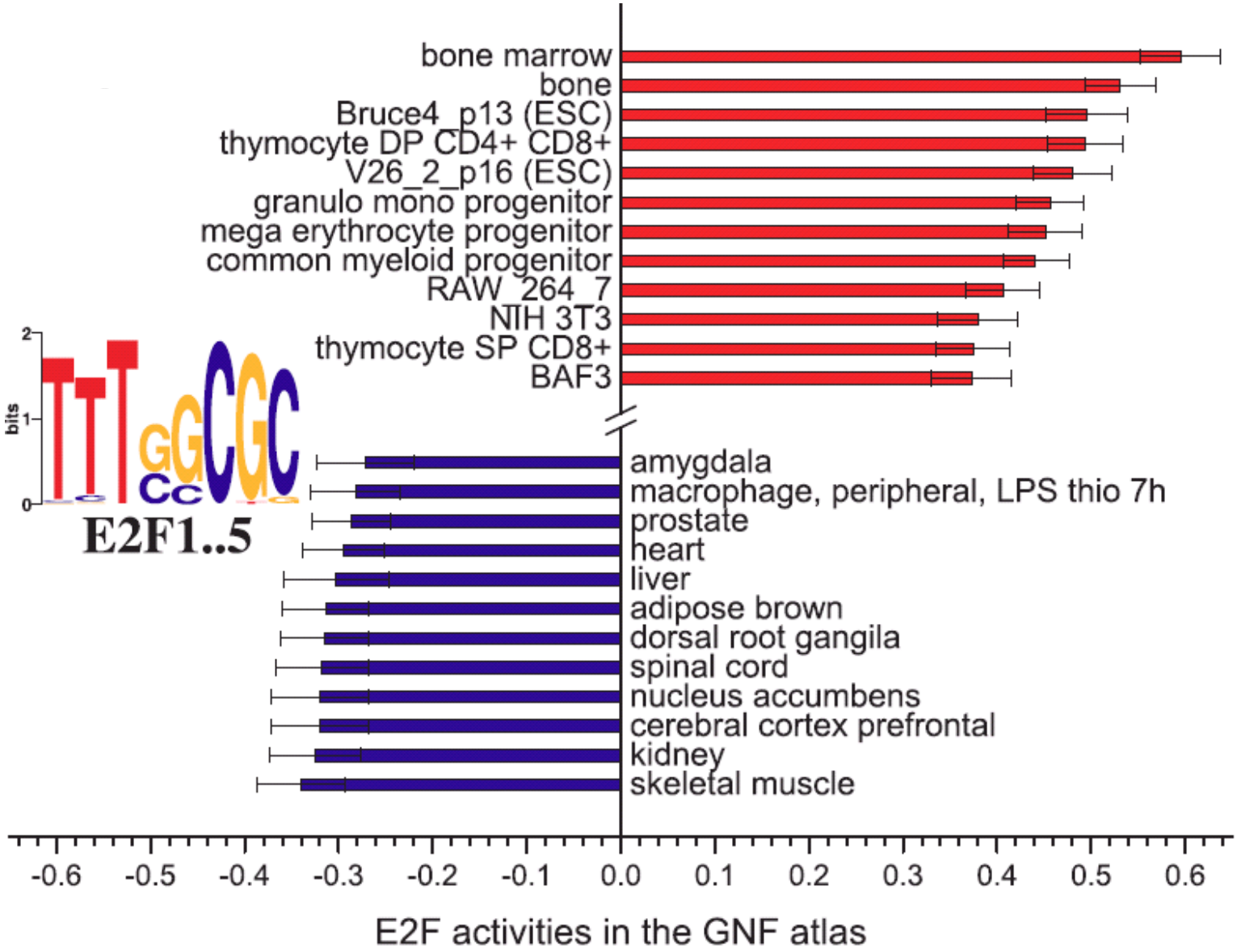


A known liver-specific factor indeed shows highest activity in liver tissues.
Other tissues in which HNF1a is known to be active are also recovered.

Motif activities associated with disease



A regulatory motif associated with cell proliferation



BIOZENTRUM
Universität Basel

SIB
Swiss Institute of Bioinformatics

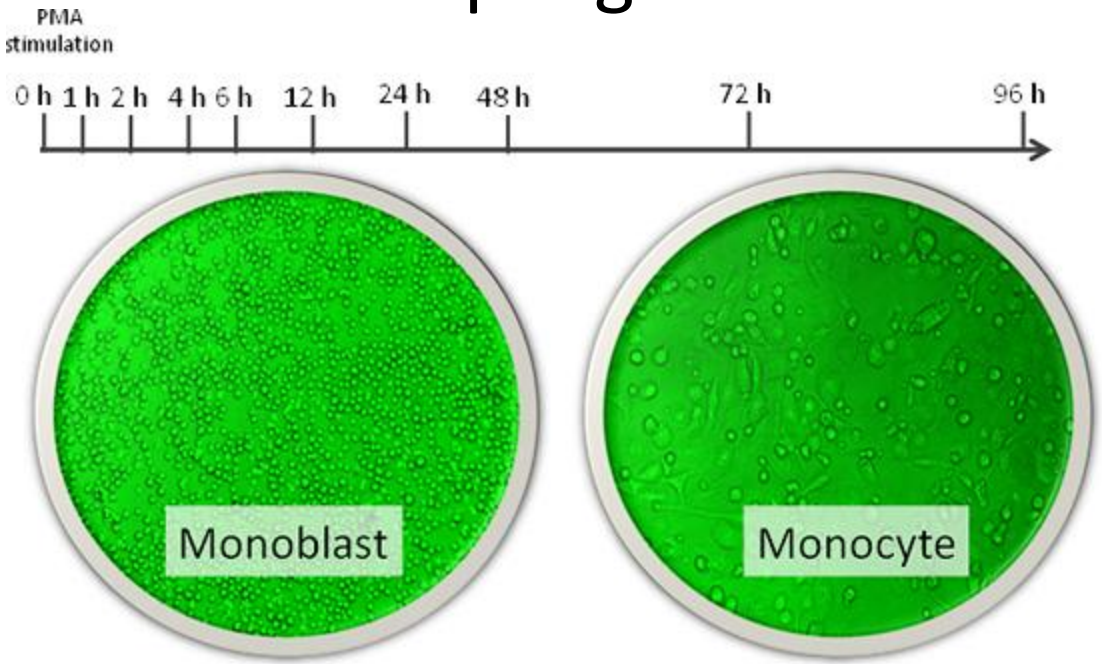
FANTOM4

Nature Genetics **41**, 553 - 562 (2009)
Published online: 19 April 2009 | doi:10.1038/ng.375

The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line

The FANTOM Consortium & Riken Omics Science Center[1]

THP-1 cells differentiating into a macrophage like state



Cells stop proliferating and become adherent

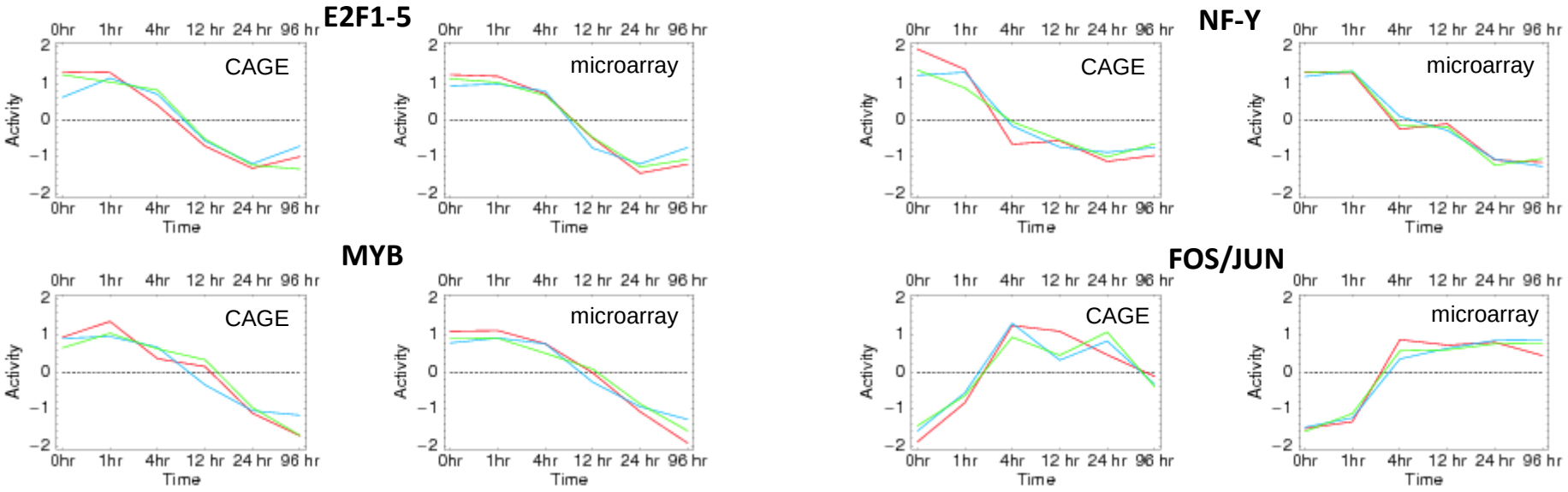
- Expression was obtained (in triplicate) at 6 time points:

1. Before PMA treatment.
2. After **1** hour of PMA treatment.
3. After **4** hours of PMA treatment.
4. After **12** hours of PMA treatment.
5. After **24** hours of PMA treatment.
6. After **96** hours (4 days) of PMA treatment.

- Genome wide expression was measured using both DeepCAGE and micro-arrays.
- DeepCAGE samples had between 1 and 2 million mapped TSSs each.

Motif Activities

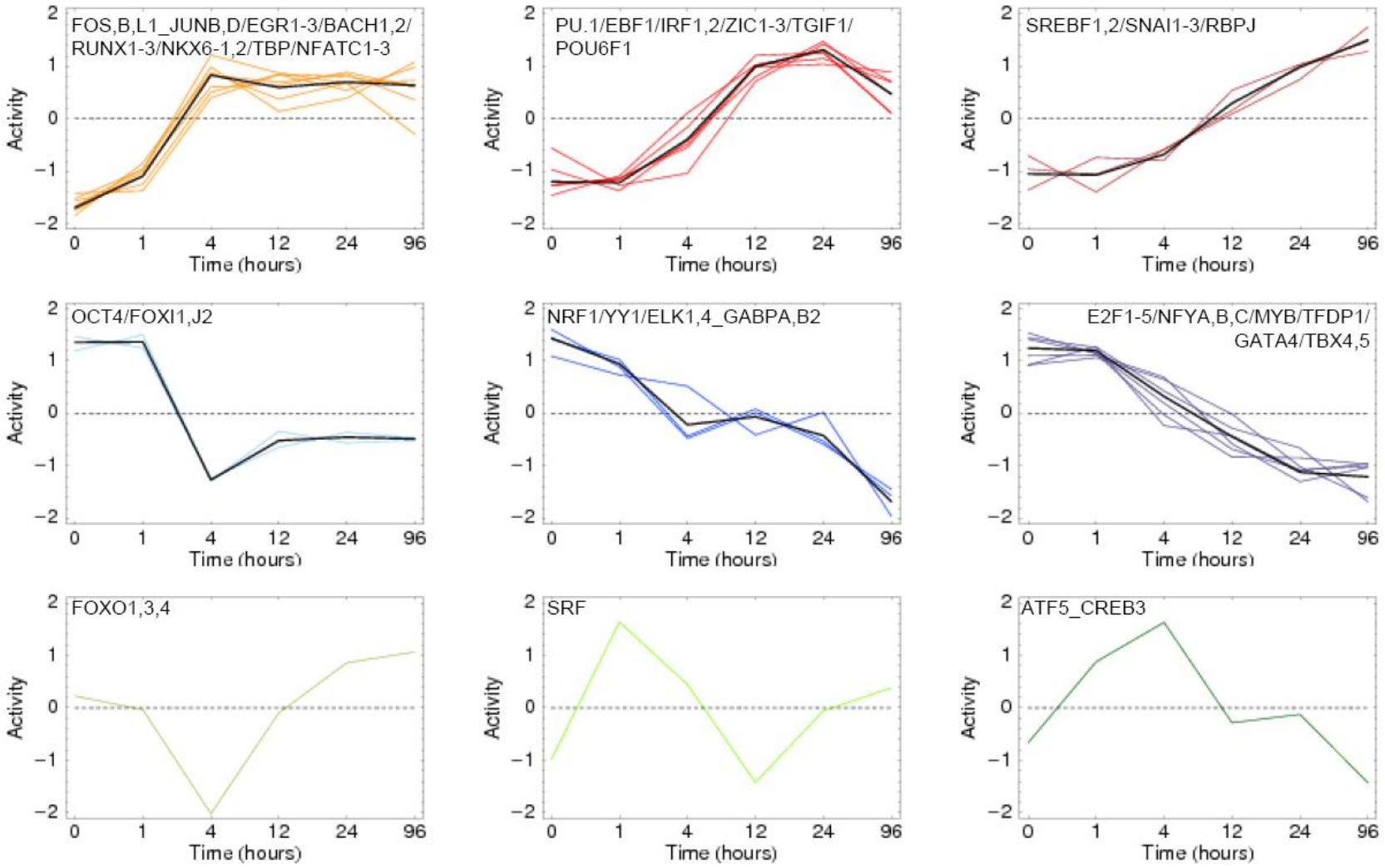
Inferred activity profiles of the top 4 motifs



Different colors correspond to the three biological replicates.

- Activity profiles are significantly *more reproducible* then the expression data of individual genes.
- This is because activity is inferred from the global behavior of *all promoters.*

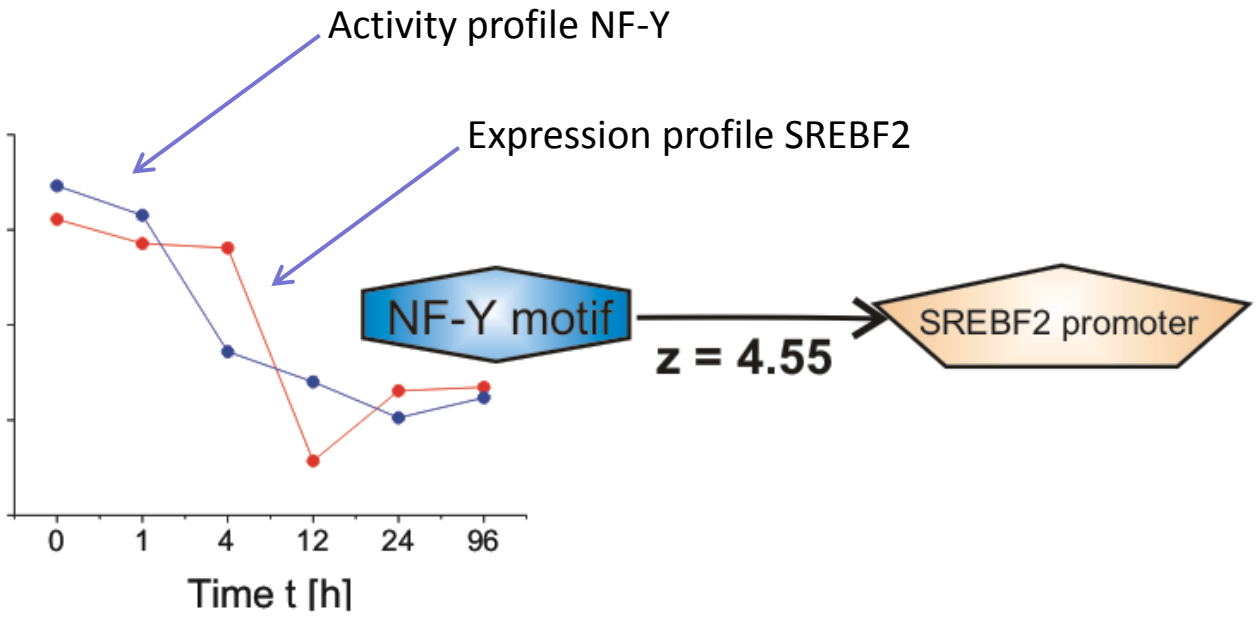
Core motif activity profile clusters



- Activity of each motif at each time point is characterized by Gaussian with given mean and variance.
- For any set of motifs we can calculate the probability of the `data' assuming all activity profiles are the same.

Prediction of regulated target promoters

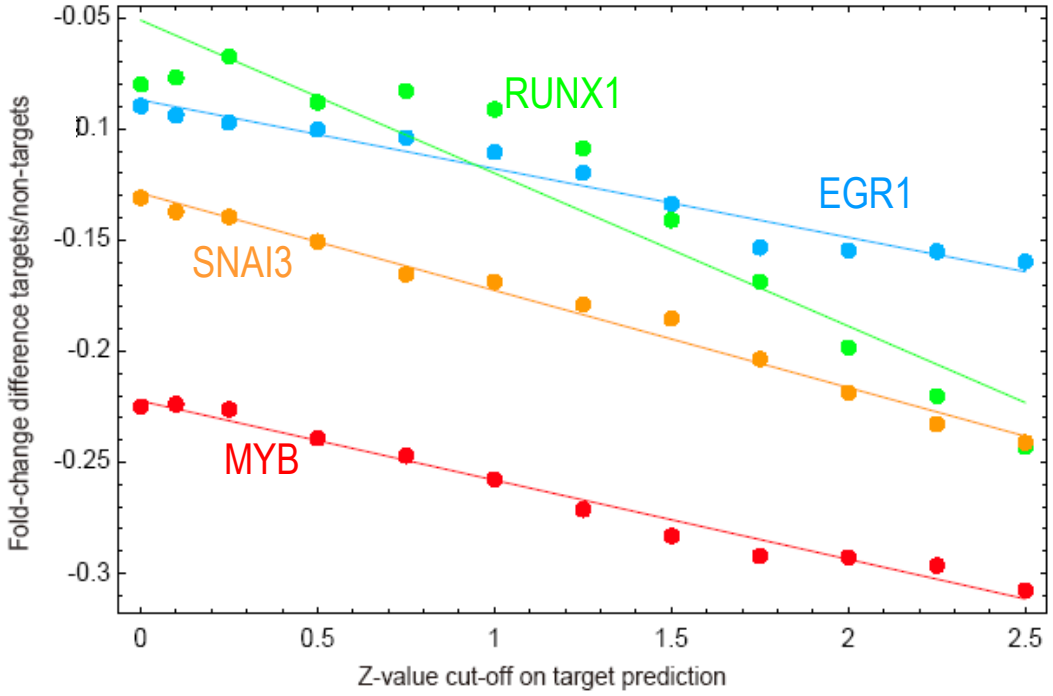
- For each motif go through list of all promoters with predicted TFBSs $N_{pm} > 0$
- Investigate the correlation between *expression profile of the promoter* and *activity profile of the motif*.



The z-value quantifies the strength of the correlation.

Validating predicted targets with siRNA TF knock down.

- Knock down TF associated with the motif using siRNA.
- Use micro-array (triplicate) to determine average fold-change siRNA vs. mock transfection.



- **Predicted targets are down-regulated relative to non-target genes.**
- **Higher confidence targets respond more strongly to siRNA knock-down.**

Core regulatory network

Edge confirmed by literature

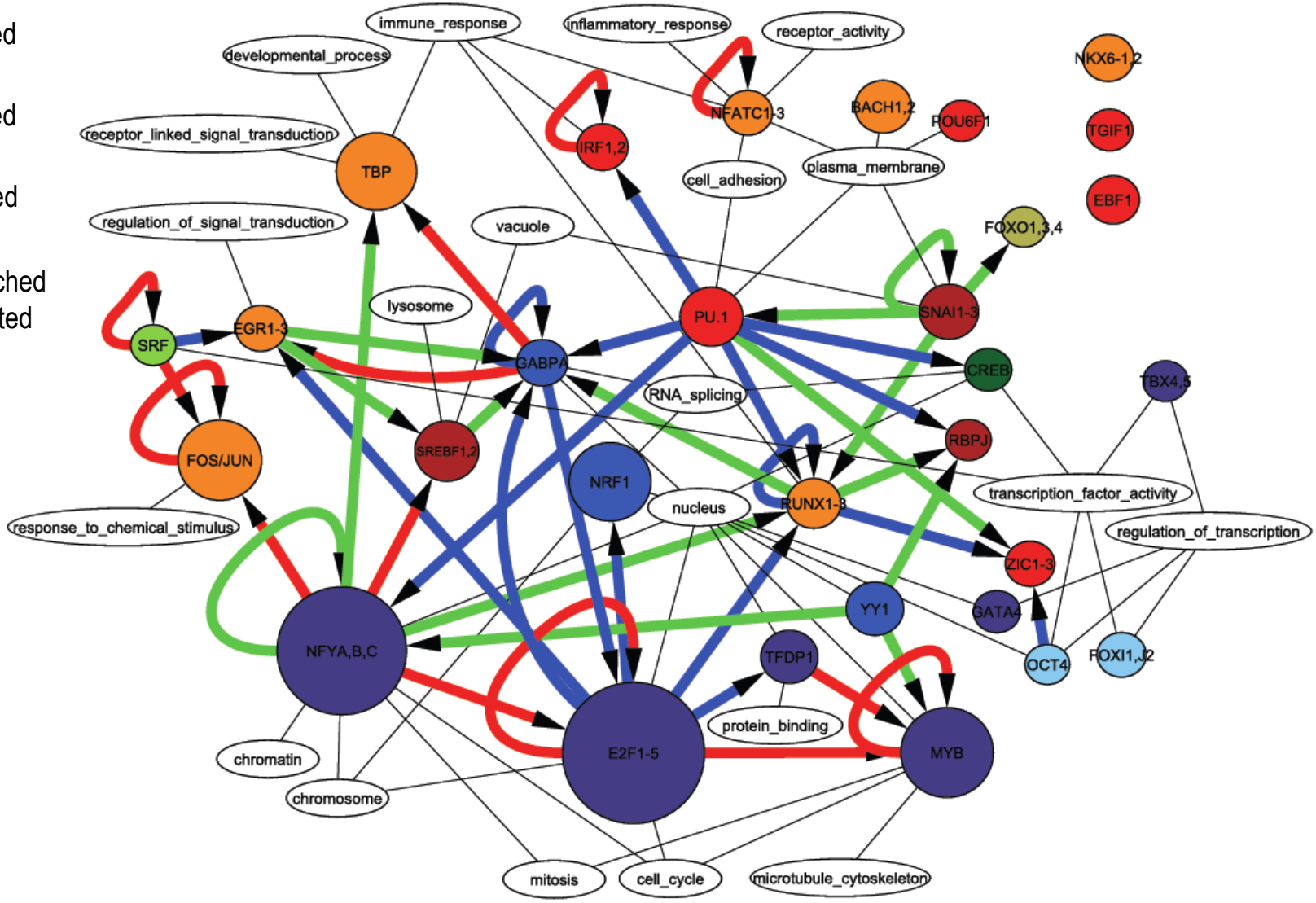
Edge confirmed by ChIP-chip

Edge confirmed by siRNA

GO-term enriched among predicted targets

Size node = Z-value activity profile

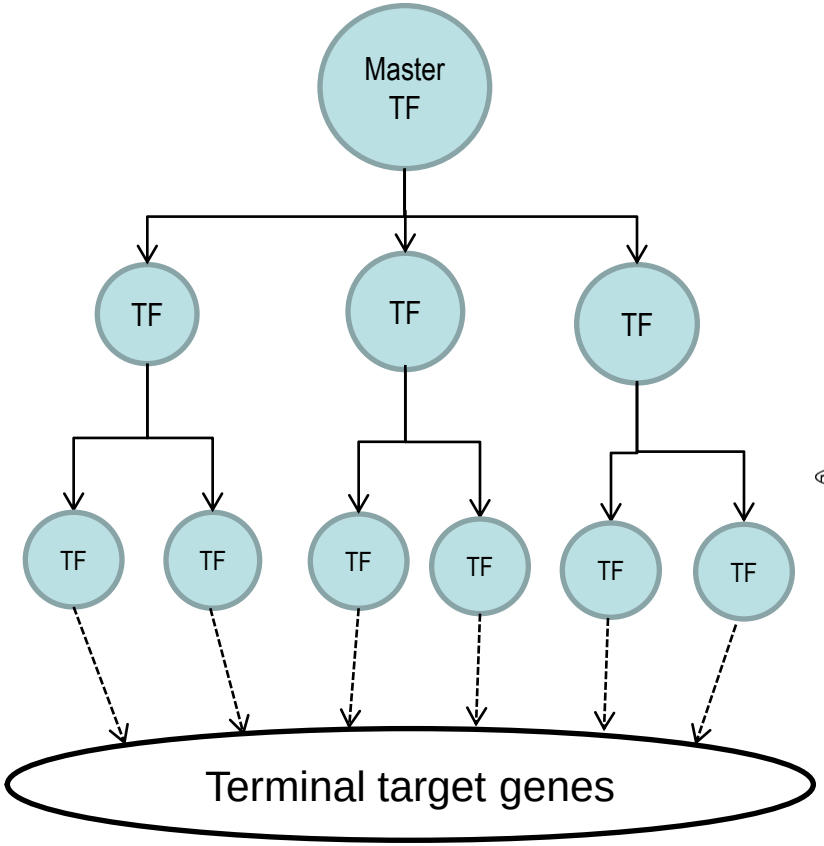
Color node = Membership activity cluster



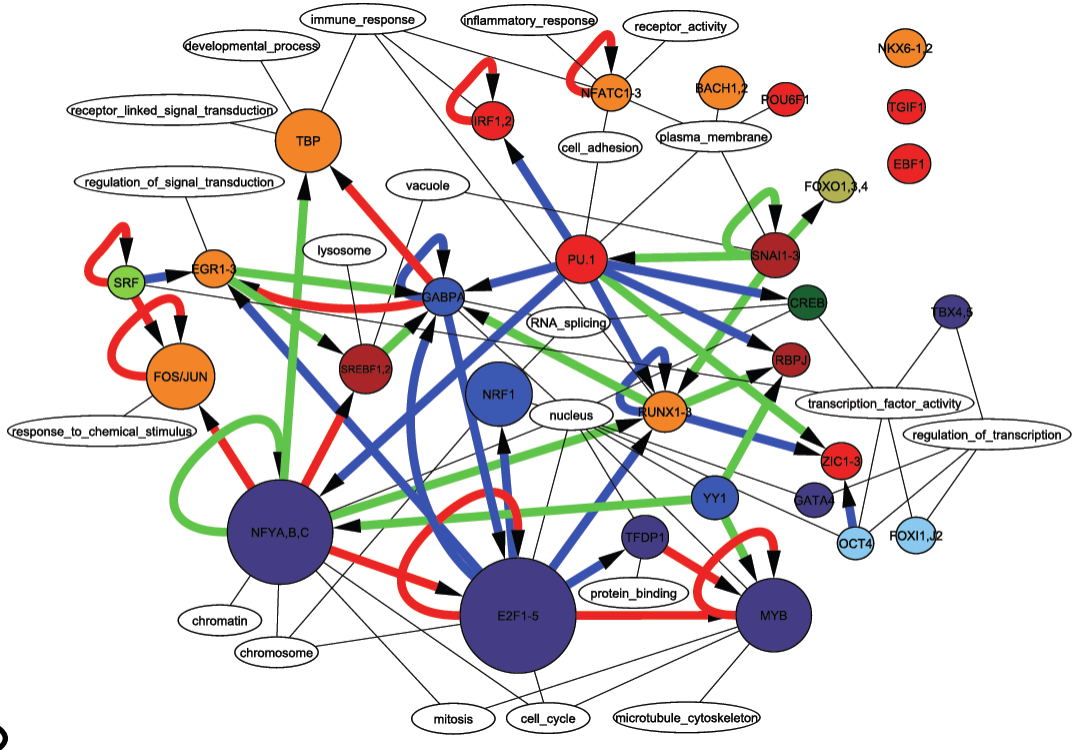
Of all 199 predicted edges (z>1.5) between the 30 core motifs, 55 had independent experimental support.

Conclusions: Master regulators

Regulatory cascade with master



Recurrent network of mutually regulating TFs

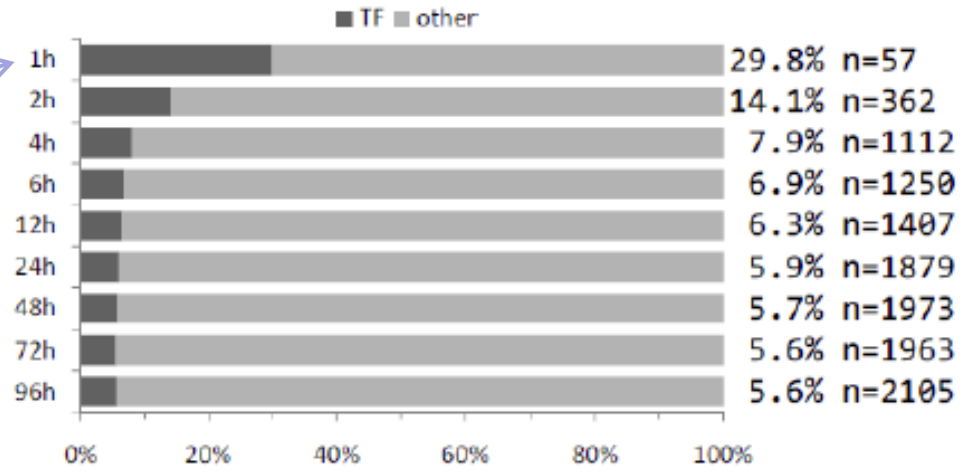


- No pre-defined direction of the regulatory flow and many feed-backs.
- No single `master' switch that controls the entire process.
- High connectivity between the regulators.

TFs without known motifs

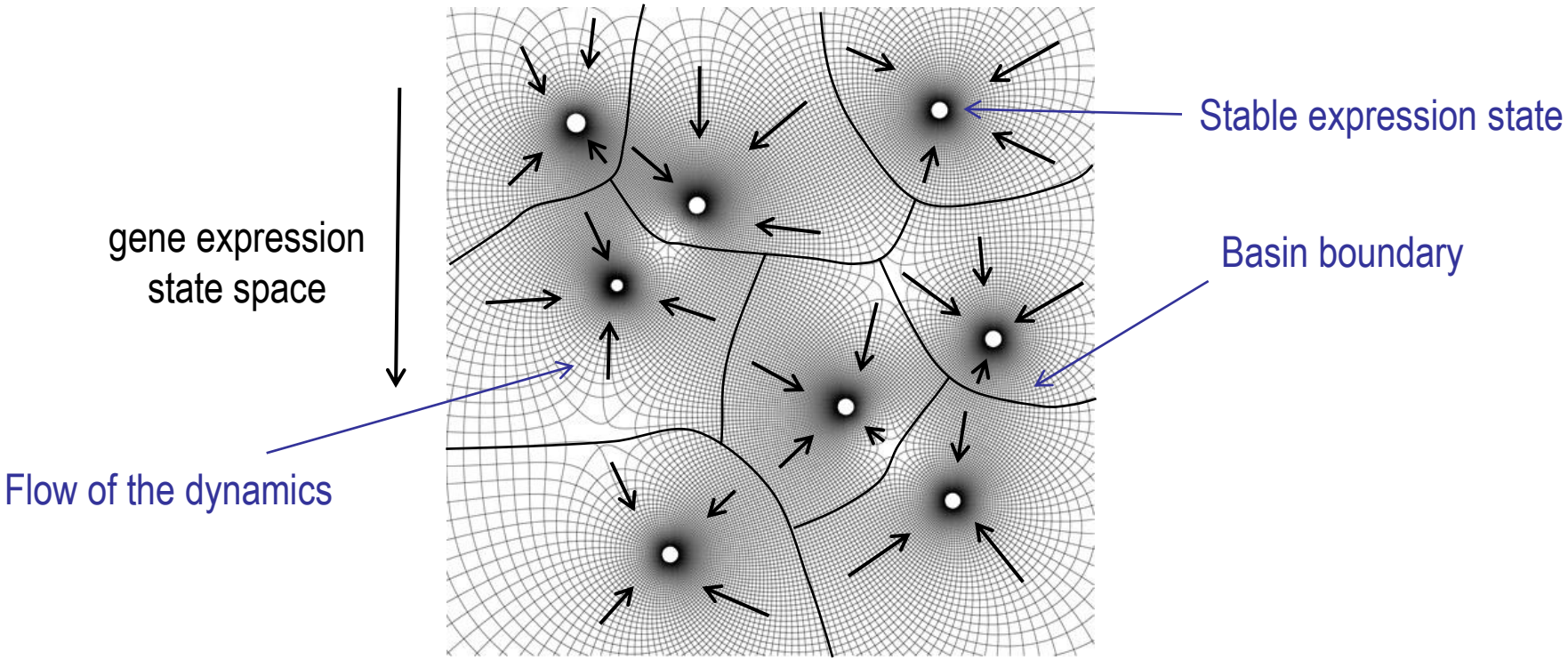
Description	Count	Regulating core motifs
Expressed TFs	610	455 without known motif
Statically expressed	408	none
Down-regulated	34	OCT4, GATA4, NF-Y
Up-regulated	64	SNAI, TBP
Transiently changing	110	SRF

A large fraction of genes that change expression at 1 hour are TFs.



- A large fraction of genes that change expression at 1 hour are TFs.
- Predicted regulators of these early response TFs are: SRF, FOSL2, TBP
- All these are *known* PMA-responsive genes.

Conclusions:
Stable cellular states as dynamic attractors



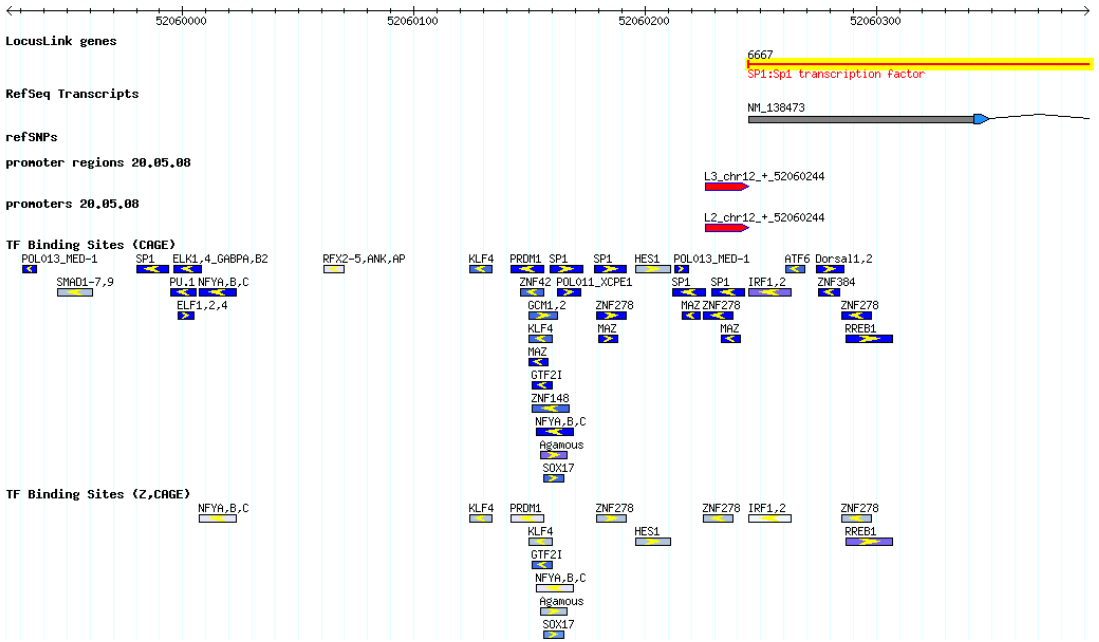
- Many TFs mutually regulating each other *create* stable expression states.
- Move between attractors involves transiently active TFs (e.g. SRF) that target mainly other TFs.
- The state space is *high-dimensional*: there are many different ways to move between different stable states.
- Many places in state space are likely `no-go' areas: cell death.

SwissRegulon and MARA

SwissRegulon

(http://www.swissregulon.unibas.ch)

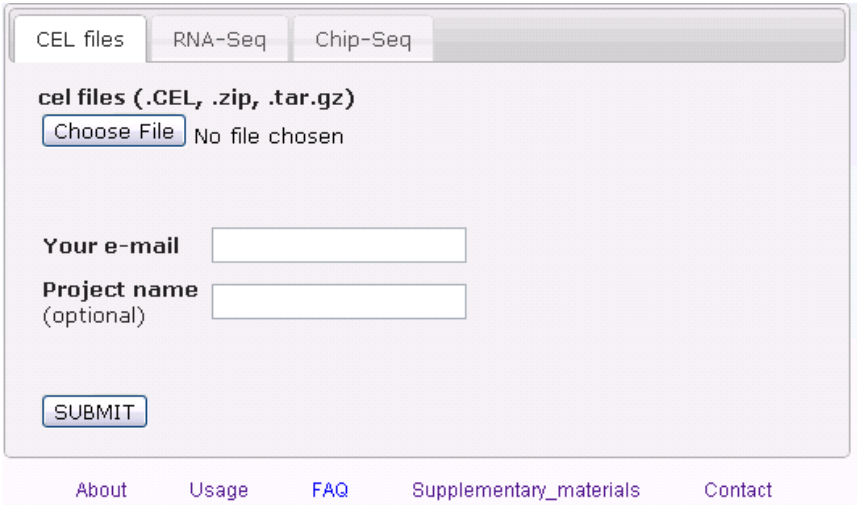
- Human and mouse *promoteromes*.
- Regulatory motif weight matrices.
- Predicted regulatory sites genome-wide.



MARA

(http://www.swissregulon.unibas.ch/mara)

- Upload your own data.
- Automated Motif Activity Response Analysis:
  - Significant motifs affecting expression.
  - Motif activities.
  - Target promoter lists.
  - Associated GO-terms.



Acknowledgments

Biozentrum

Piotr Balwierz
MARA
promoteromes

Phil Arnold
MotEvo/
epigenetic
modifications

Mikhail Pachkov
SwissRegulon site
MARA webserver

Omics Science Center
RIKEN Institute, Yokohama, Japan

Yoshihide Hayashizaki

Harukazu Suzuki

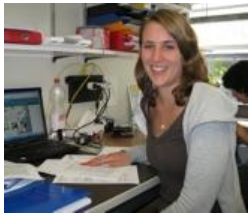
Alistair Forrest

FANTOM4
GNP-ECW

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

Dirk Schübeler

Fabio Mohn



Anne Schöler

CELL PLASTICITY
Systems Biology of Cell Differentiation

http://www.cellplasticity.org/organization.asp

SystemsX.ch
The Swiss Initiative in Systems Biology