

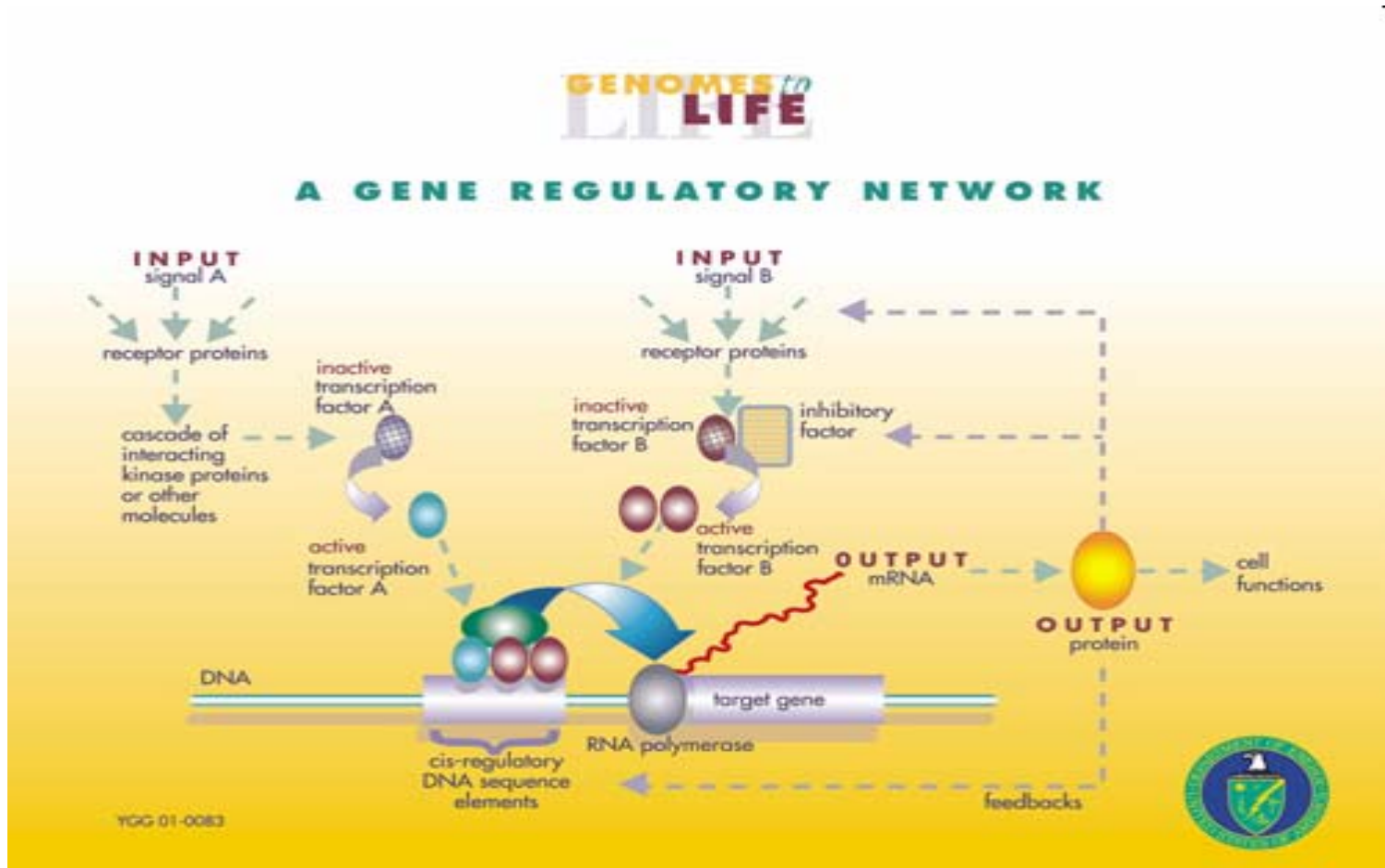


**Max-Planck-Institute for Molecular Genetics**  
**Department Vertebrate Genomics, Prof. Dr. Lehrach**  
**Bioinformatics Group, Dr. Herwig**  
**Lukas Chávez**

# **Integrative Analysis of Gene Regulatory Networks**

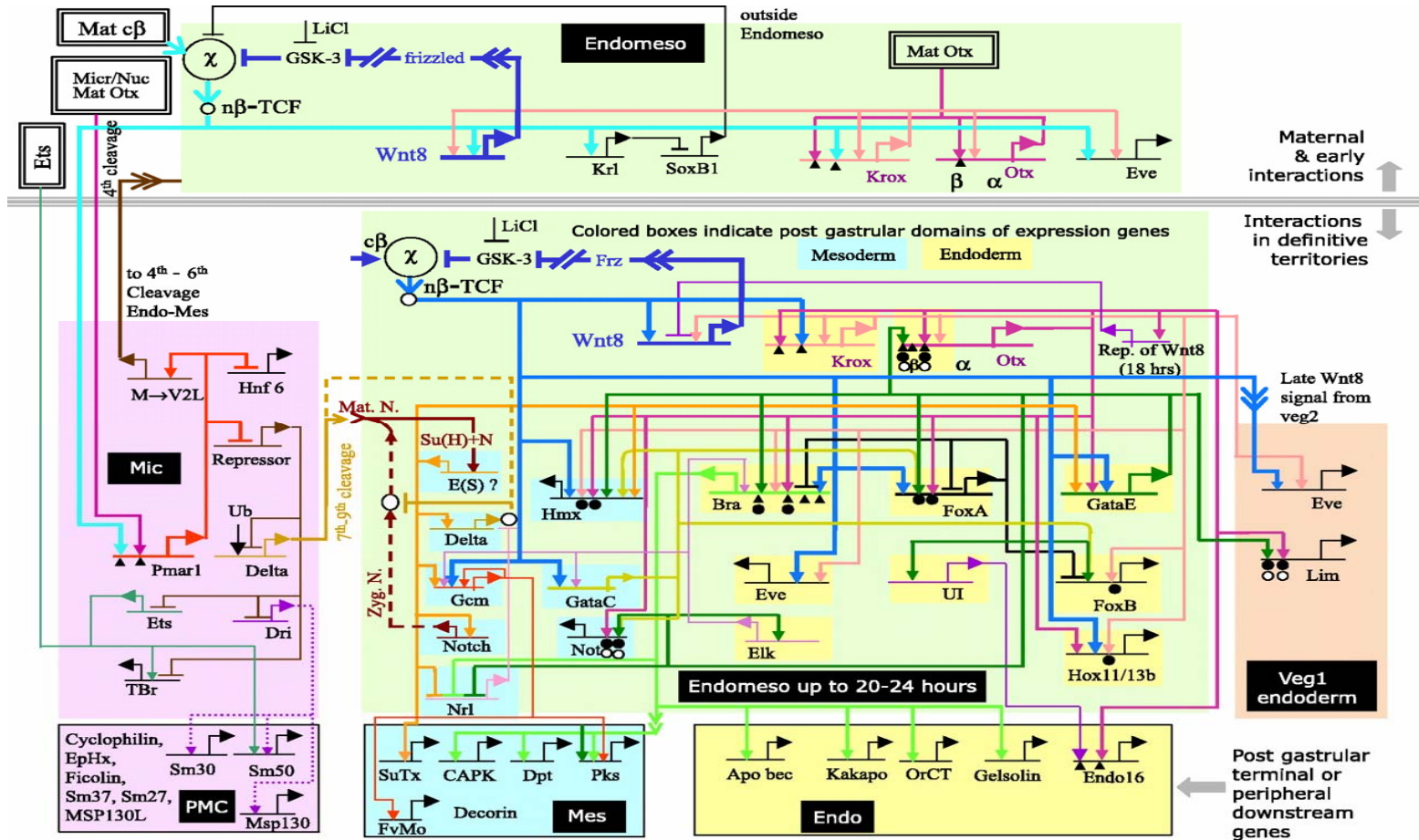


# 1. Gene regulatory networks



[http://en.wikipedia.org/wiki/File:Gene\\_Regulatory\\_Network.jpg](http://en.wikipedia.org/wiki/File:Gene_Regulatory_Network.jpg)

# 1. Gene regulatory networks (2)



Davidson et al., *A genomic regulatory network for development*. Science, CA Inst Tech, Pasadena, 2002, 295, 1669-1678

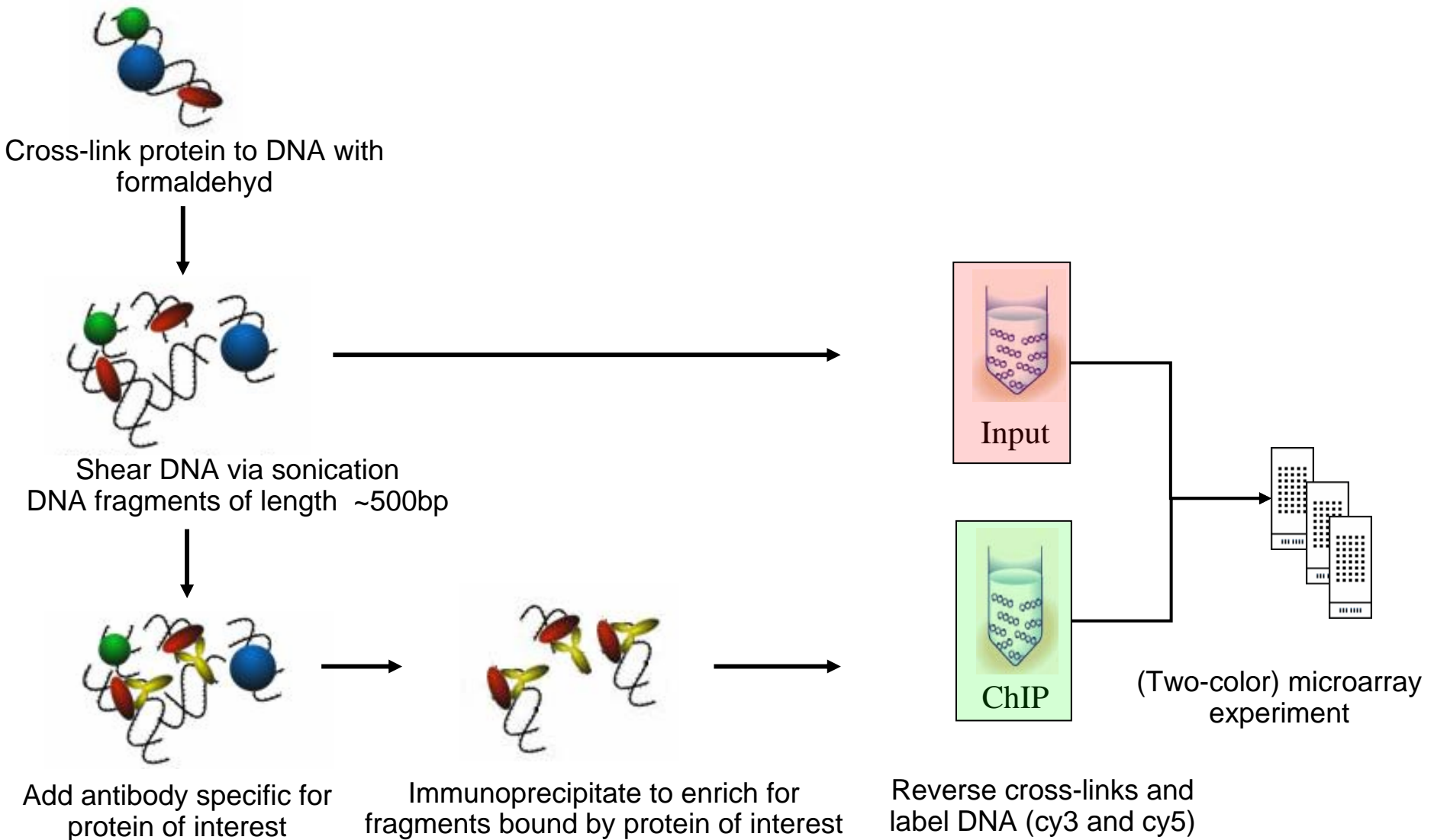


# 1. Gene regulatory networks (3)

- **Reconstruction of transcriptional networks** based on
  - **ChIP-Chip** experiments: direct protein-DNA interactions
  - **RNAi knock-down** followed by microarray experiments: downstream regulation of gene expression (direct and indirect gene targets)
  - **Sequence based** analysis of transcription factor binding sites: direct targets and denovo motif discovery

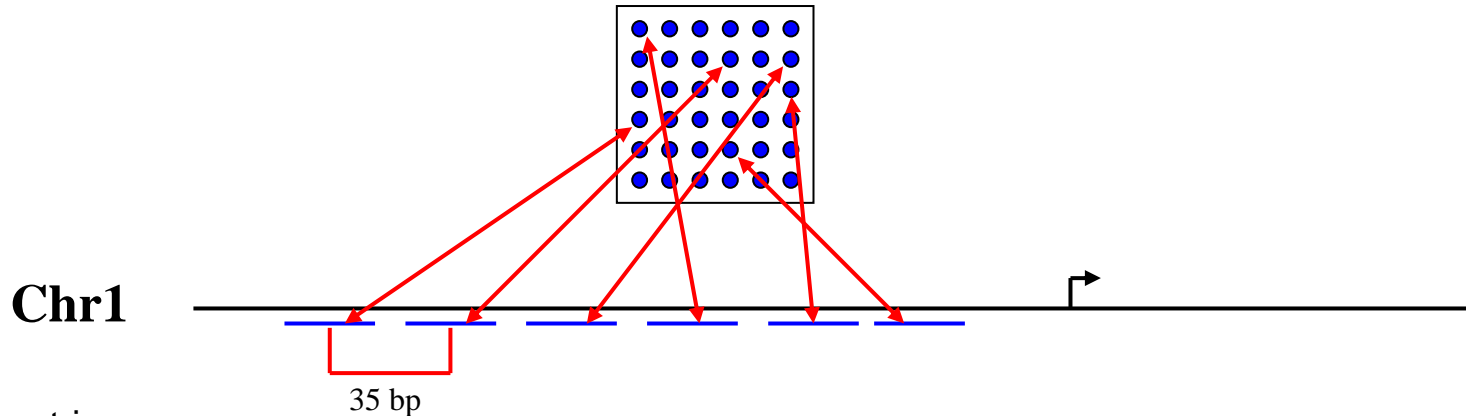


# 2.1. ChIP-on-Chip: Chromatin Immuno-Precipitation





## 2.2. ChIP-on-Chip: Tiling Arrays



### • Affymetrix

#### • **Human Promotor Array:**

- 4.6 Mio 25-mer oligos with a 35bp probe spacing
- 10kb coverage (-7.5kb upstream to 2.5kb downstream) of ~25,500 human promoters
- probe design based on the NCBI build 34

#### • **Human Tiling 2.0R Array Set:**

- 45 Mio 25-mer oligos distributed over 7 arrays

### • Roche NimbleGen

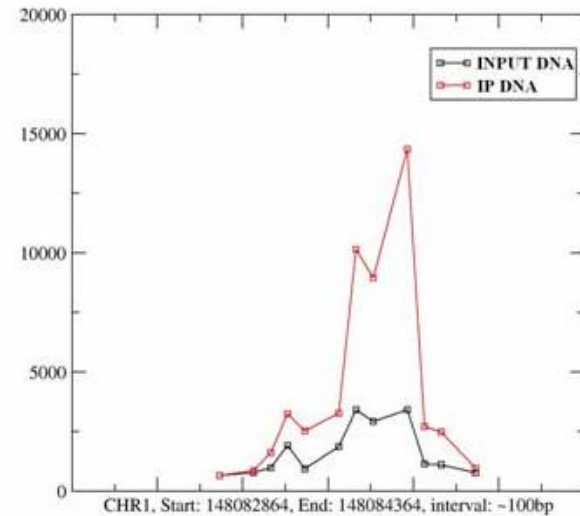
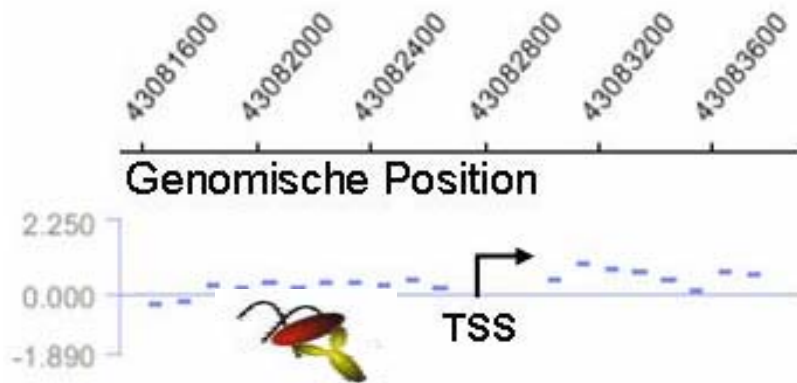
#### • **Whole Genome Sets**

- 10 array set with 21 Mio 50-70mer probes and a 100bp interval

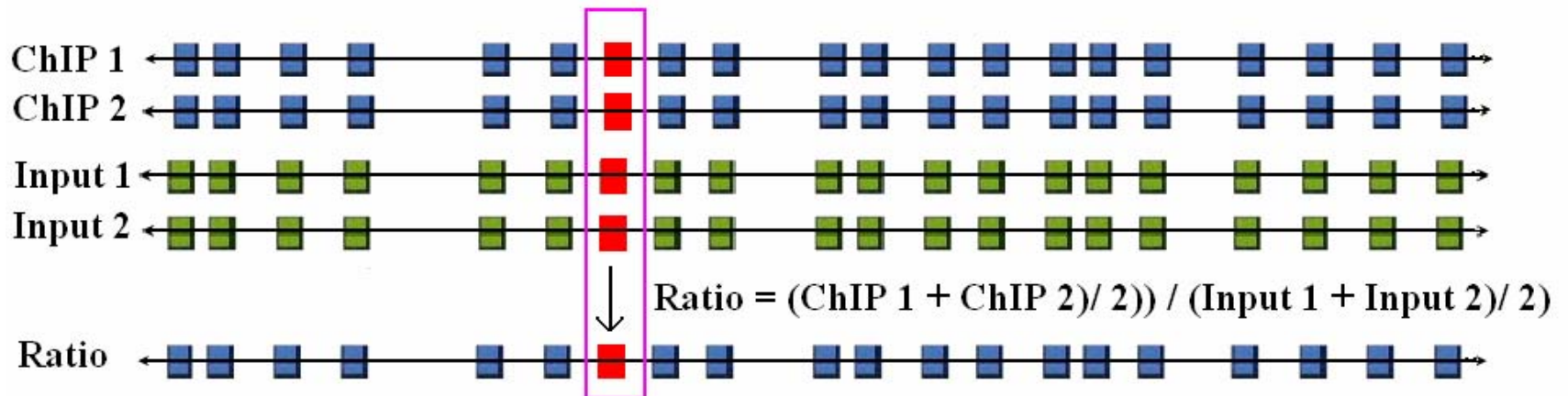
#### • **Promotor Sets:** different sets available

#### • **Custom tiling arrays:** define your own regions of interest

## 2.3. ChIP-on-Chip: Data Analysis- Ratios

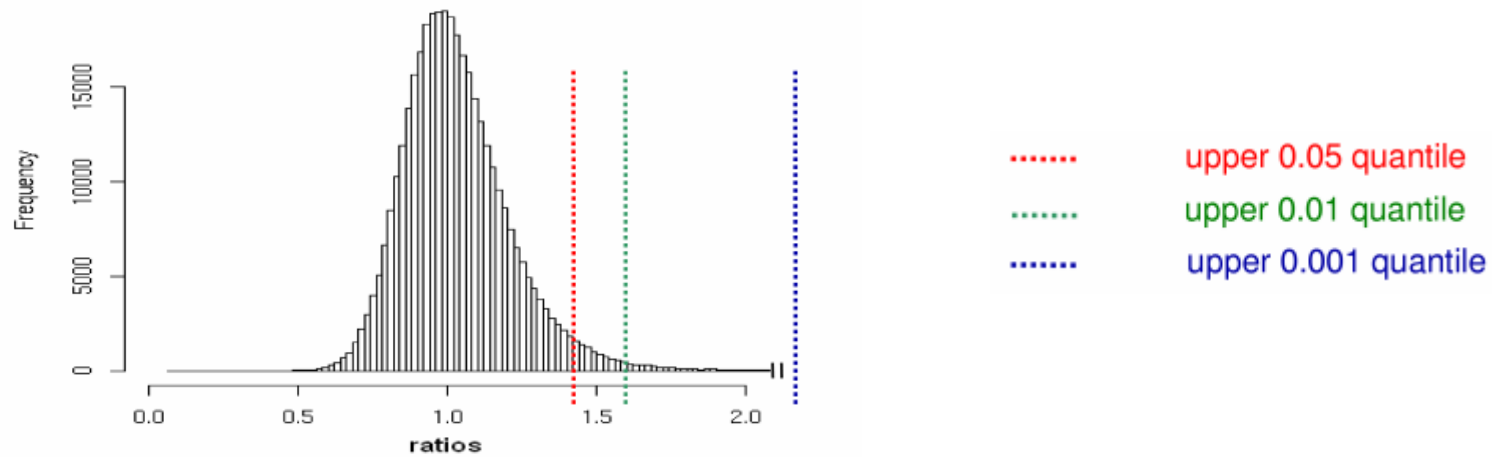


Calculate ratios (ChIP/Input):

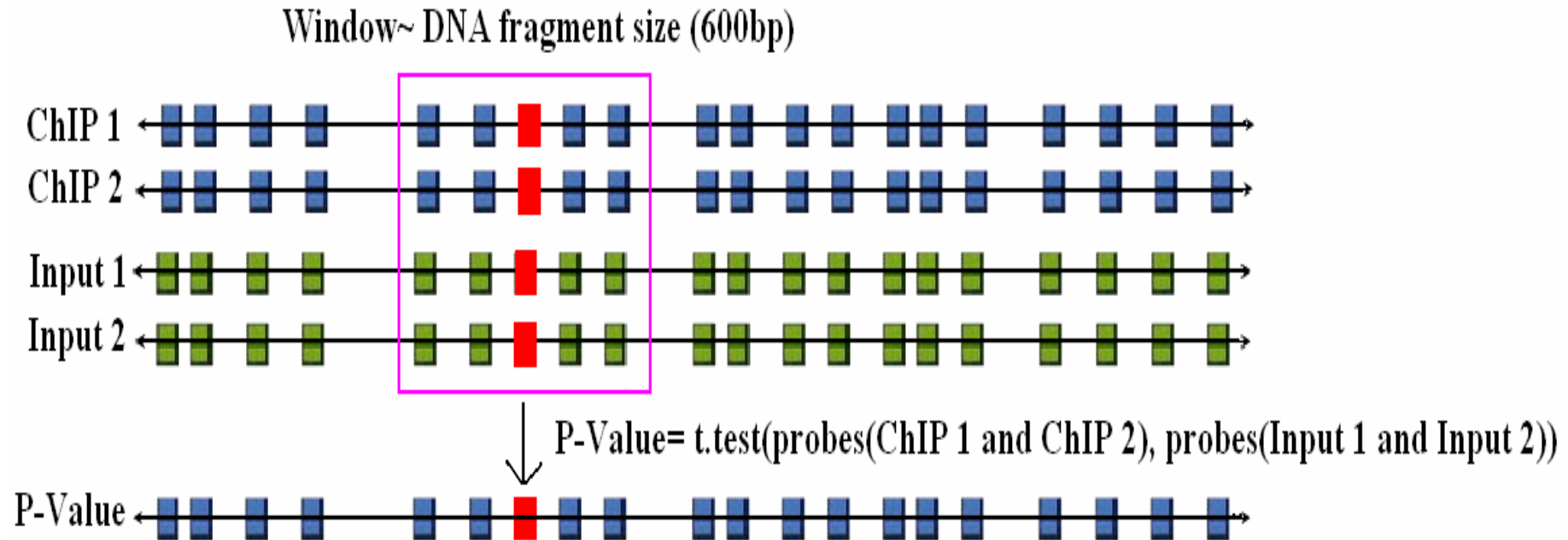




## 2.3. ChIP-on-Chip: Data Analysis- Ratios (2)



## 2.4. ChIP-on-Chip: Data Analysis- P-Values

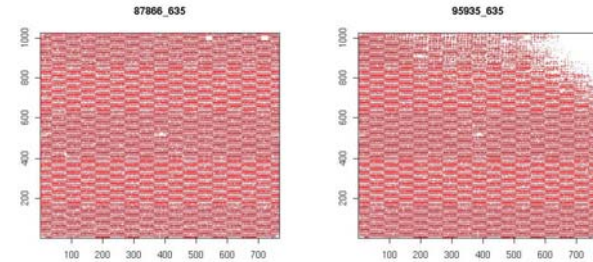


- For each oligo calculate a p-value by
  - considering all oligonucleotides within the defined window and
  - apply a **statistical test** (t-test, wilcoxon etc.)
  - subsequent **interval analysis** on p-values

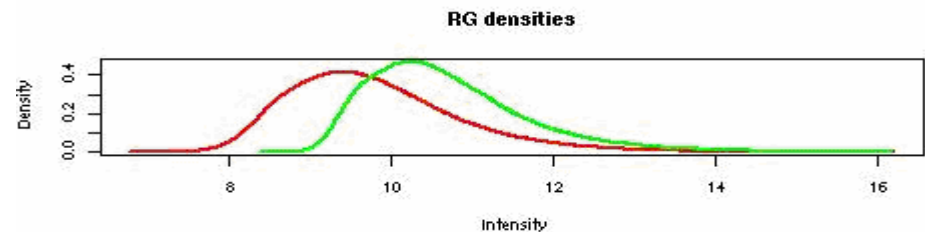
## 2.5. ChIP-on-Chip Data Analysis using Bioconductor

### General features

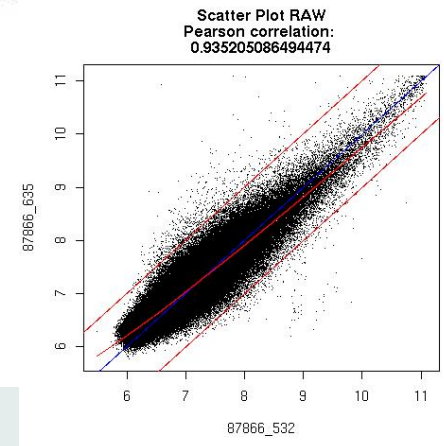
- Mapping of oligos to a reference genome (e.g. for updated genomes) using the *posToProbeAnno()* function
- Array image reconstruction using the *image()* function



- Density distribution of the single-channel densities using the *plotDensities()* function (see the limma package)



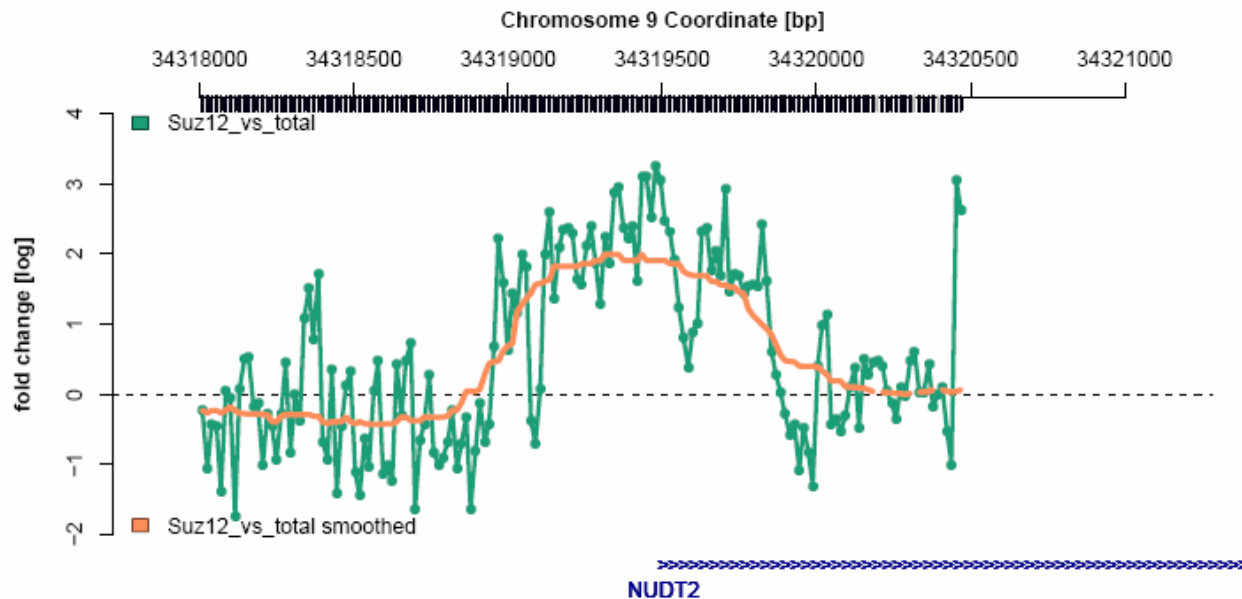
- Many kinds of normalization methods available
- Calculate correlation coefficients, create scatter plots, MA plots etc.



## 2.5. ChIP-on-Chip Data Analysis using Bioconductor (2)

### Ringo package

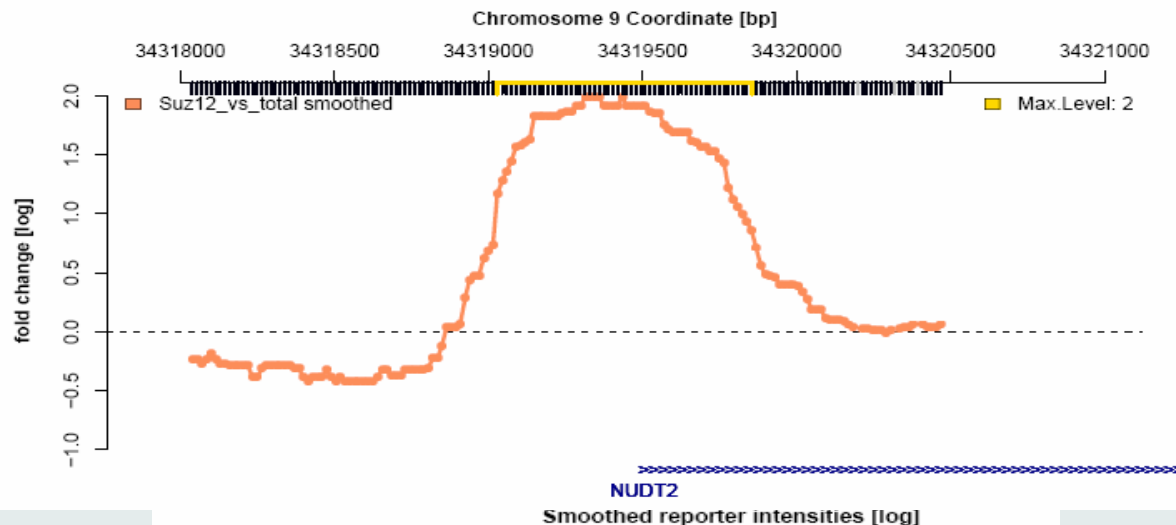
- `source("http://bioconductor.org/biocLite.R"); biocLite("Ringo")`
- Visualize intensities along the chromosome
- Smoothing of probe intensities: oligos respond differently to hybridized material caused by e.g. different probe GC content, melting temperature, and secondary structure
  - 800bp window slides along the chromosome
  - replaces the intensity of a probe by the median of all probes within this window



## 2.5. ChIP-on-Chip Data Analysis using Bioconductor (3)

### Ringo package (2)

- Finding ChIP-enriched regions on smoothed  $\log_2(\text{ChIP}/\text{Input})$  ratios
  - define a window (e.g. 500bp)
  - require at least 3 oligos within the window
  - each ratio is above a threshold  $t_0$ .
- Provides a method for estimation of  $t_0$  considering the *null-distribution*
  - *Null distribution*: Distribution of smoothed ratios within non-binding regions
- Output enriched regions



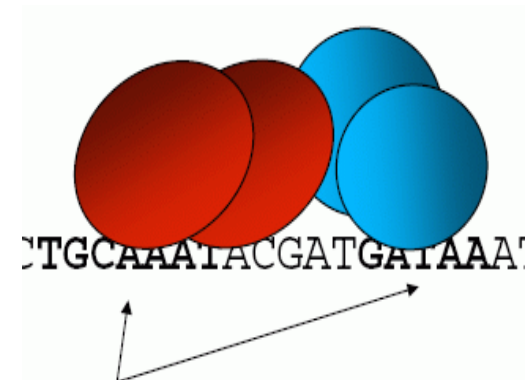
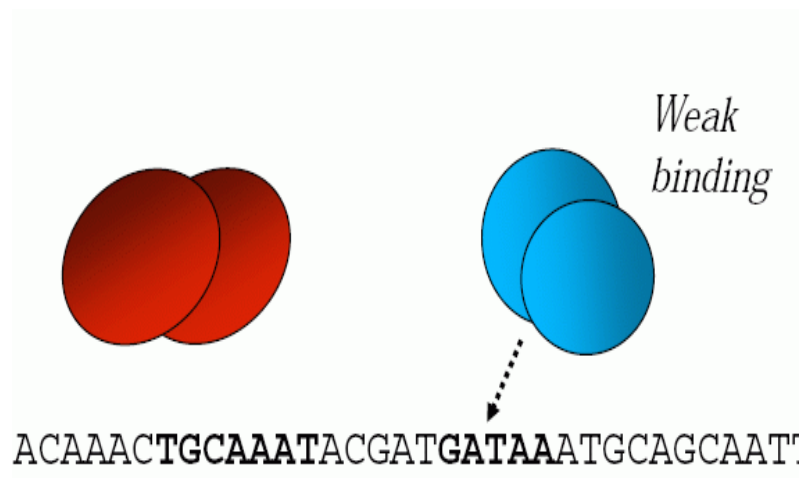
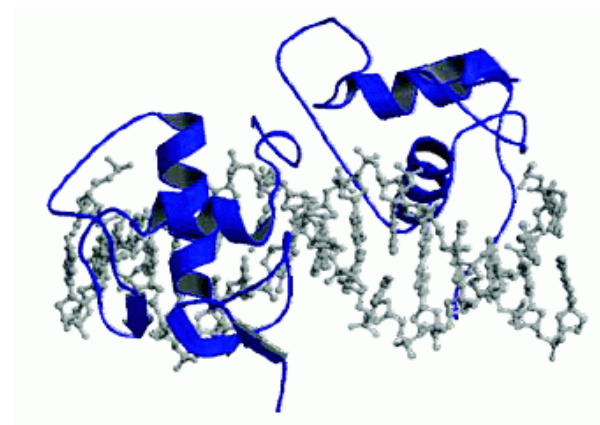
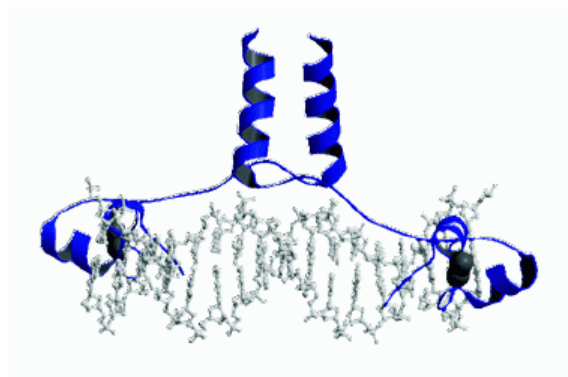


## 2.6. ChIP-on-Chip: Further tools

- Affymetrix:
  - TAS (Tiling Analysis Software)
  - Integrated Genome Browser: Visualization of ChIP-Chip data along the chromosome
  
- NimbleGen:
  - NimbleScan
  - SignalMap: Visualization
  
- Bioconductor:
  - Ringo
  - BAC (Bayesian Analysis of ChIP-Chip): Model based analysis of tiling arrays
  - ...
  
- Standalone:
  - CisGenome
  - JBD (joint binding deconvolution)
- ...



### 3. Sequence based analysis of TF binding sites (TFBS)





## 3. Sequence based analysis of TFBS (2)

### 1. Discover a motif within a set of sequences (*de novo* motif discovery)

Data sources:

- **Co-regulated** genes from gene expression analysis/ clustering
- **evolutionary conserved** regions
- **ChIP-on-Chip** data

Algorithms for Motif Discovery:

- find motifs that are **statistically overrepresented** in the dataset

### 2. Does a sequence **contain** a motif?

### 3. Sequence based analysis of TFBS (3)

Sequences bound by TF:

```

ATCGACTACAAATGCAAA GCTTACGATGTGATAAATGCAGCAAAT
ACTTACTAGCATGGCCATCATCAAATGATAAGCAGGTTGTGCC
GGATAAATGTAATGTATT CATACGATCAGCATCAGATATCGATTG
TACGATGTATATACAGGATTAGCCTGTCTCCACTACA AATGCAAT
ATAAATGCCCAATTGATTTGTCTCCACTACA AATGCAATTACGATG
ACGTGTCTGCTACA AATGCAAA TACGATGATAAATGCAGCAATTG
ACGT AATGTATT TACGATGATAAATGCAGCAACCGTTATCGACTTG
GTAACTCATCATAGCATGGCGATCAAATGATAA CAGGTTGTGCC
    
```

Position Weight Matrix:

	1	2	3	4	5	6	7	8
A	0.53	0.00	0.00	1.00	0.20	0.02	0.01	0.30
C	0.02	0.10	0.00	0.00	0.39	0.03	0.86	0.25
G	0.43	0.01	1.00	0.00	0.20	0.03	0.08	0.22
T	0.02	0.89	0.00	0.00	0.21	0.92	0.05	0.23
	a	T	G	A	c	T	C	.

*De novo* motif discovery

Sequence Logos:





## 3. Sequence based analysis of TFBS (4)

### De novo motif discovery tools

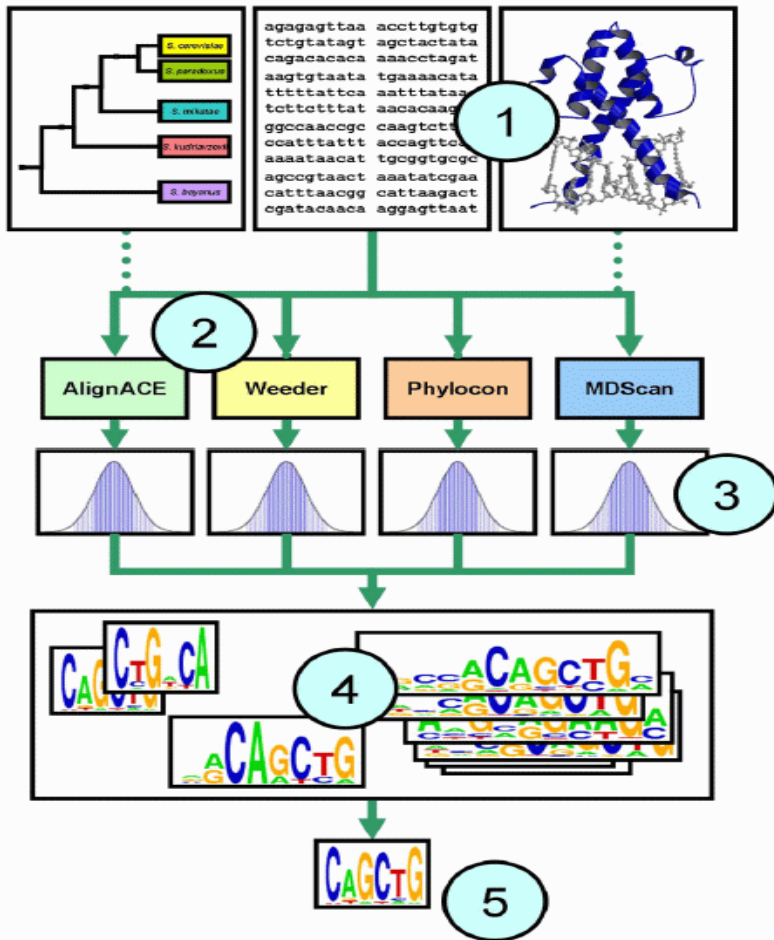
**Table 1** Details about the operation principles, basic technical data and URLs of 13 analyzed tools

Program	Operating principle	Technical data	URL	Reference
AlignACE	Gibbs sampling algorithm that returns a series of motifs as weight matrices that are over-represented in the input set	Judges alignments sampled during the course of the algorithm using a maximum <i>a priori</i> log likelihood score, which gauges the degree of overrepresentation. Provides an adjunct measure (group specificity score) that takes into account the sequence of the entire genome and highlights those motifs found preferentially in association with the genes under consideration.	<a href="http://atlas.med.harvard.edu/">http://atlas.med.harvard.edu/</a>	7
ANN-Spec	Models the DNA-binding specificity of a transcription factor using a weight matrix	Objective function based on log likelihood that transcription factor binds at least once in each sequence of the positive training data compared with the number of times it is estimated to bind in the background training data. Parameter fitting is accomplished with a gradient descent method, which includes Gibbs sampling of the positive training examples.	<a href="http://www.cbs.dtu.dk/~workman/ann-spec/">http://www.cbs.dtu.dk/~workman/ann-spec/</a>	8
Consensus	Models motifs using weight matrices, searching for the matrix with maximum information content	Uses a greedy method, first finding the pair of sequences that share the motif with greatest information content, then finding the third sequence that can be added to the motif resulting in greatest information content, and so on.	<a href="http://bifrost.wustl.edu/consensus/">http://bifrost.wustl.edu/consensus/</a>	9
GLAM	Gibbs sampling-based algorithm that automatically optimizes the alignment width and evaluates the statistical significance of its output	Since the basic algorithm cannot find multiple motif instances per sequence, long sequences were fragmented into shorter ones, and the alignment was transformed into a weight matrix and used to scan the sequences to obtain the final site predictions.	<a href="http://zlab.bu.edu/glam/">http://zlab.bu.edu/glam/</a>	10
The Improbizer	Uses expectation maximization to determine weight matrices of DNA motifs that occur improbably often in the input sequences	As a background (null) model it uses up to a second-order Markov model of background sequence. Optionally, Improbizer constructs a Gaussian model of motif placement, so that motifs that occur in similar positions in the input sequences are more likely to be found.	<a href="http://www.soe.ucsc.edu/~kent/improbizer">http://www.soe.ucsc.edu/~kent/improbizer</a>	11
MEME	Optimizes the E-value of a statistic related to the information content of the motif	Rather than sum of information content of each motif column, statistic used is the product of the <i>P</i> values of column information contents. The motif search consists of performing expectation maximization from starting points derived from each subsequence occurring in the input sequences. MEME differs from MEME3 mainly in using a correction factor to improve the accuracy of the objective function.	<a href="http://meme.sdsc.edu/">http://meme.sdsc.edu/</a>	12
MITRA	Uses an efficient data structure to traverse the space of IUPAC patterns.	For each pattern, MITRA computes the hypergeometric score of the occurrences in the target sequences relative to the background sequences and reports the highest scoring patterns.	<a href="http://www.calit2.net/compbio/mitra/">http://www.calit2.net/compbio/mitra/</a>	13
MotifSampler	Matrix-based, motif-finding algorithm that extends Gibbs sampling by modeling the background with a higher order Markov model	The probabilistic framework is further exploited to estimate the expected number of motif instances in the sequence.	<a href="http://www.esat.kuleuven.ac.be/~dna/Biol/Software.html">http://www.esat.kuleuven.ac.be/~dna/Biol/Software.html</a>	14

Table 1 continued on following page

# 3. Sequence based analysis of TFBS (5)

## TAMO package

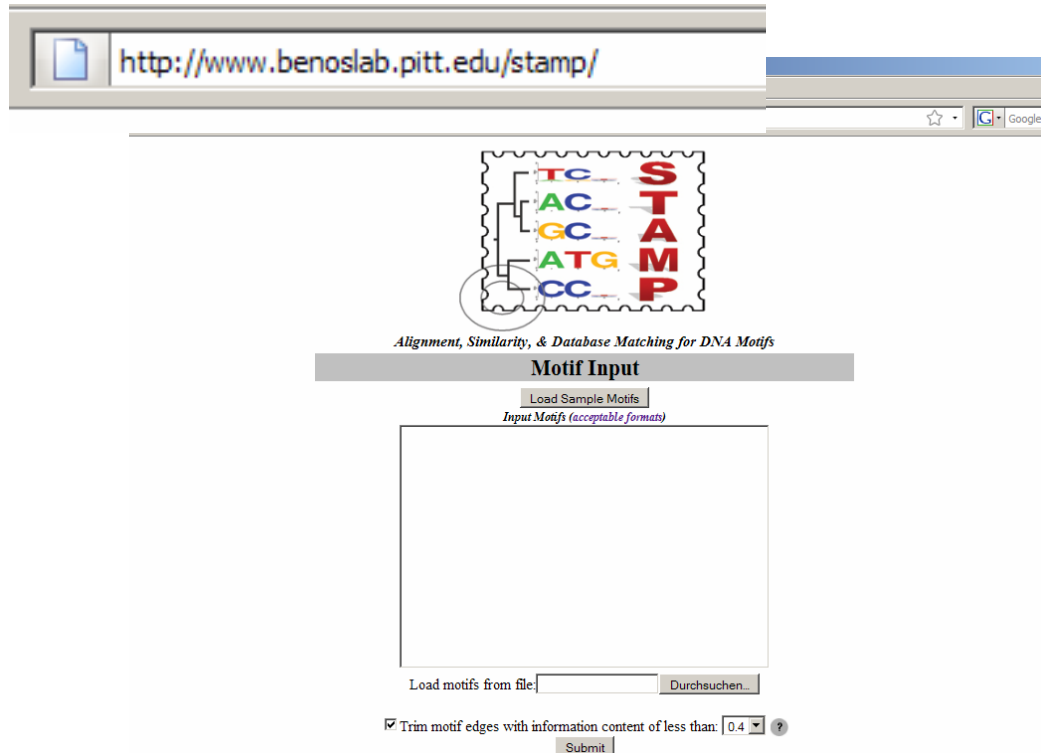


- Assemble input data.** Results may be improved by restricting the input to high-confidence sequences. Some algorithms achieve improved performance by using phylogenetic conservation information from orthologous sequences or information about protein DNA-binding domains.
- Choose several motif discovery programs for the analysis.** For recommended programs see Figure 3.
- Test the statistical significance of the resulting motifs.** Use control calculations to estimate the empirical distribution of scores produced by each program on random data.
- Clustering and post-processing the motifs.** Motif discovery analyses often produce many similar motifs, which may be combined using clustering. Phylogenetic conservation information may be used to filter out statistically significant, but non-conserved motifs that are more likely to correspond to spurious sequence patterns.
- Interpretation of motifs.** Algorithms exist for linking motifs to transcription factors and for combining motif discovery with expression data.

Kenzie D. MacIsaac, Ernest Fraenkel: “*Practical Strategies for Discovering Regulatory DNA Sequence Motifs*”; PLOS Computational Biology 2006

## 3. Sequence based analysis of TFBS (6)

### STAMP: Alignment, Similarity and Database Matching for DNA Motifs



The screenshot shows a web browser window with the address bar containing the URL <http://www.benoslab.pitt.edu/stamp/>. Below the browser window is the STAMP web interface. At the top, there is a logo for STAMP (Sequence, Similarity, and Database Matching for DNA Motifs) featuring a dendrogram and the letters S, T, A, M, P. Below the logo is the title "Alignment, Similarity, & Database Matching for DNA Motifs". The main section is titled "Motif Input" and contains a "Load Sample Motifs" button. Below this is a large empty text area for "Input Motifs (acceptable format)". At the bottom, there is a "Load motifs from file:" field with a "Durchsuchen..." button, a checkbox for "Trim motif edges with information content of less than: 0.4", and a "Submit" button.

Available motif databases via STAMP:

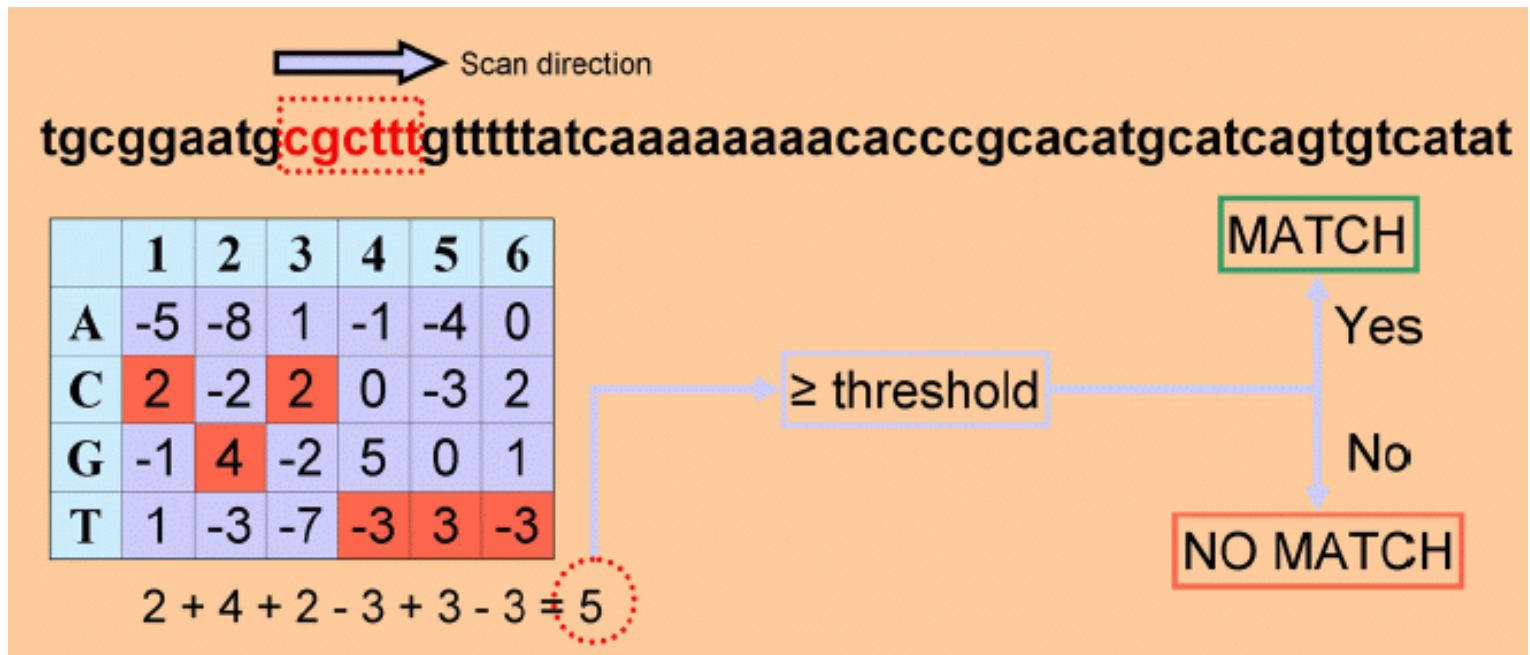
- TRANSFAC
- Jaspar

Mahony, S. & Benos, P. V.: **STAMP: a web tool for exploring DNA-binding motif similarities.** *Nucleic Acids Res* 2007, **35**:W253-W258.

### 3. Sequence based analysis of TFBS (7)

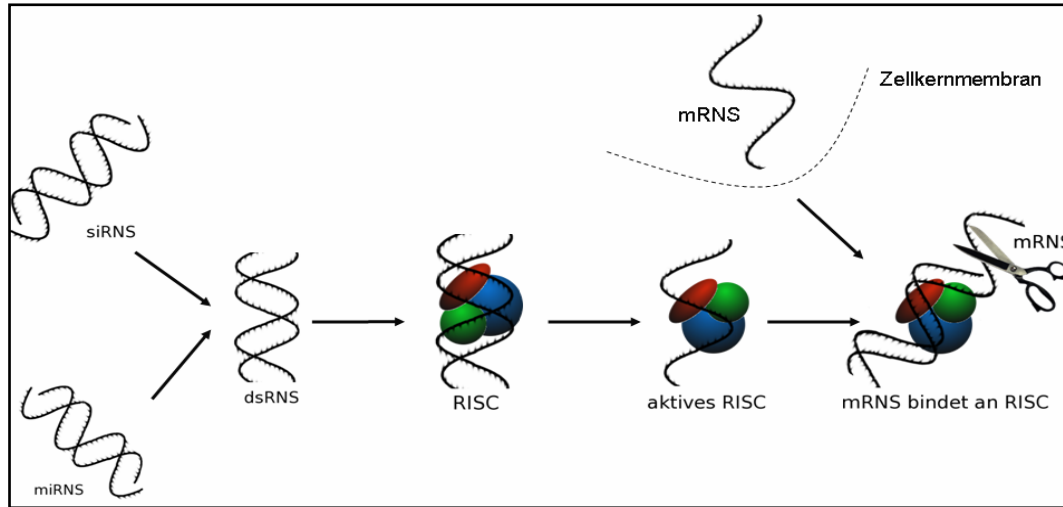
Does a sequence contain a given motif?

- Start with the PWM of a transcription factor
- Scan the promotor sequences of potential target genes





# 4. RNAi - Chip

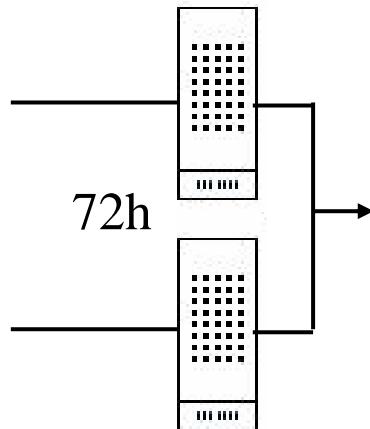


„posttranscriptional gene regulation“

specific  
knock down

72h

Unspecific  
knock down



Differential gene expression analysis identifies  
**direct** and **indirect** target genes

→ **causal** dependencies

→ Information about  $\uparrow$  and  $\downarrow$  regulation caused by the silenced transcription factor



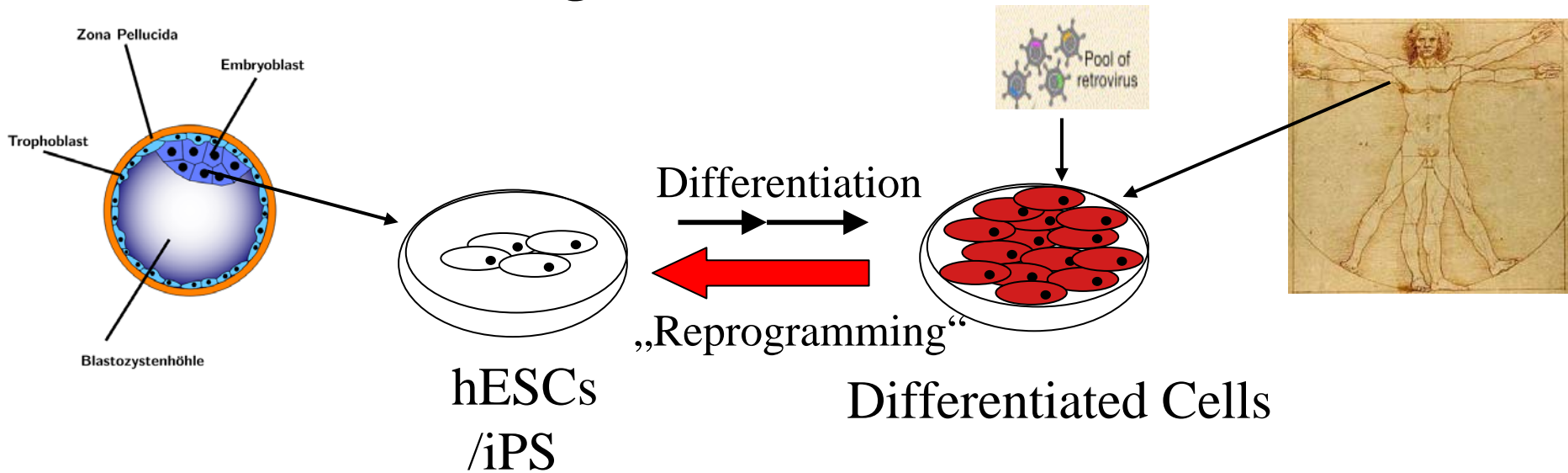
# 5. OCT4 as a key regulator of pluripotency

**Takahashi, Yamanaka (2007, Nov. 30th):** *Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors* , Cell 2007

→ **Oct4, Sox2, C-Myc and Klf4**

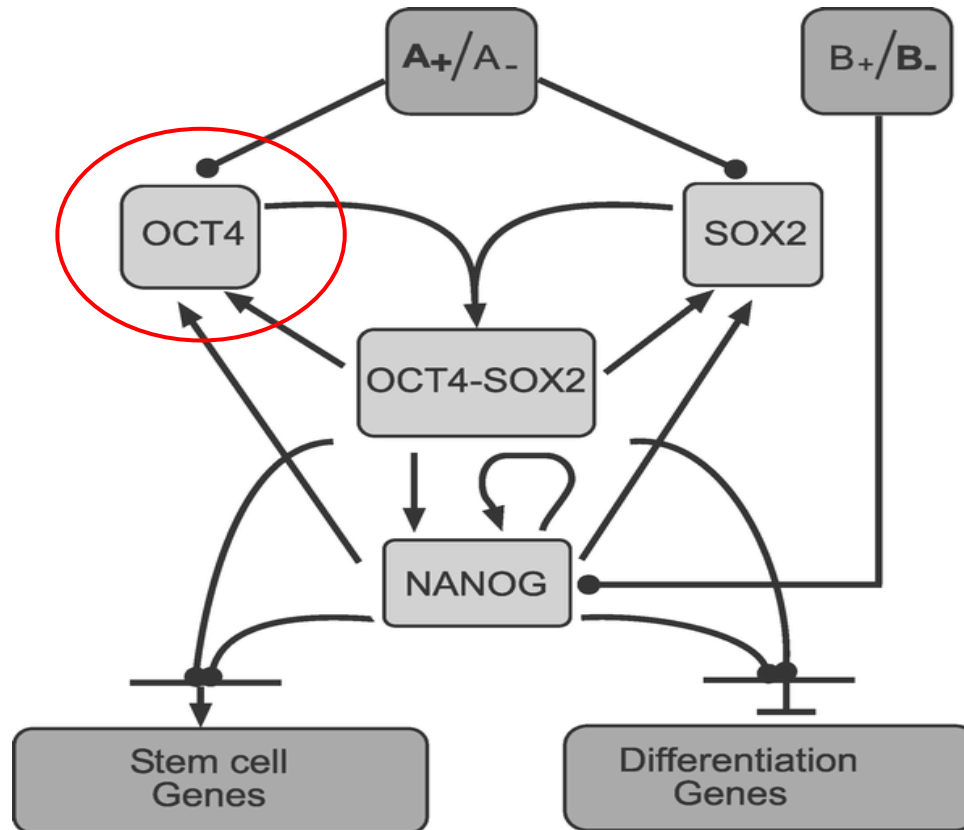
**Yu, Thomson (2007, Nov, 20th):** *Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells*, Scienceexpress 2007

→ **Oct4, Sox2, Nanog and Lin28**





## 5.1. OCT4 dependent gene regulatory network

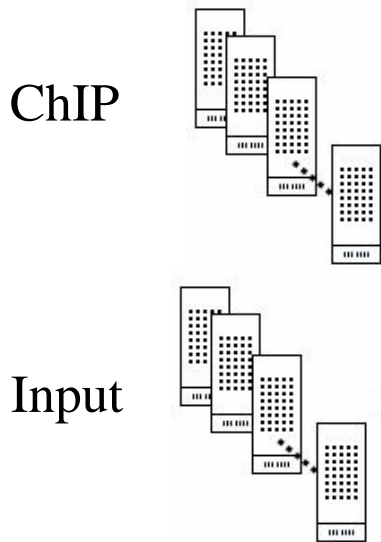


Chickarmane, Peterson et al.: *Transcriptional Dynamics of the Embryonic Stem Cell Switch*, PloS Computational Biology 2006



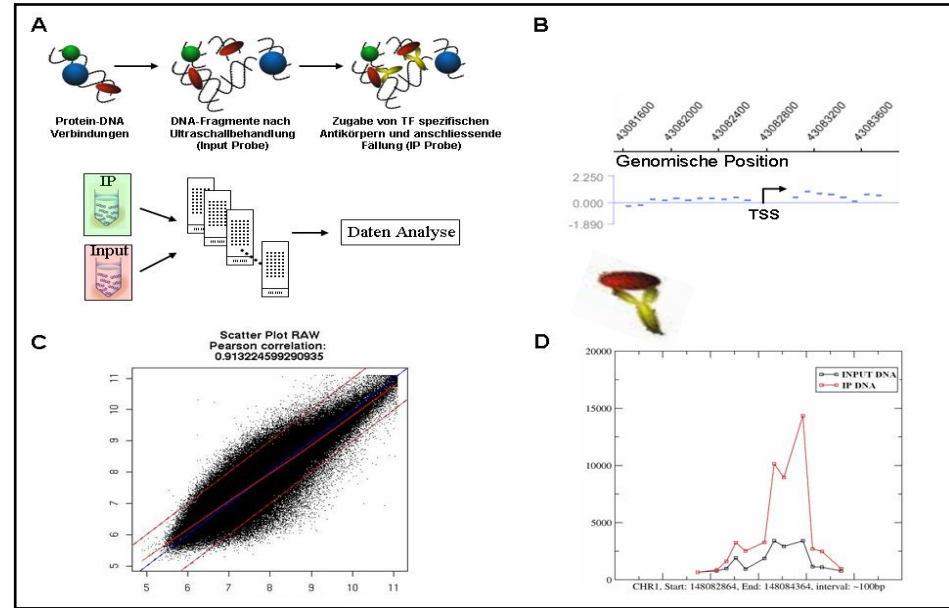
# 5.2. OCT4 ChIP-on-Chip in hESCs

- OCT4, SOX2 and NANOG
- Set of 10 Agilent arrays
- ~400.000 oligos
- scanning promotor regions around the TSS (-8kb - +2kb) of ~22.000 known genes



623 direct target genes of OCT4

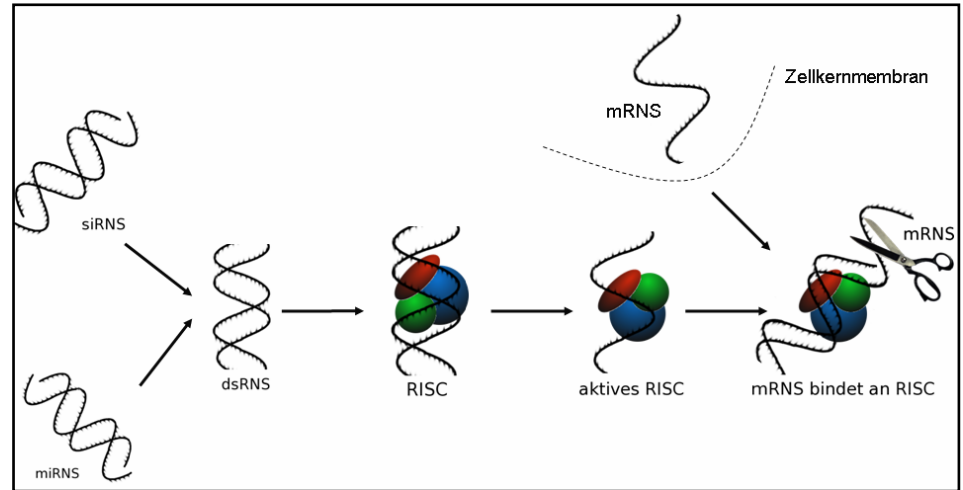
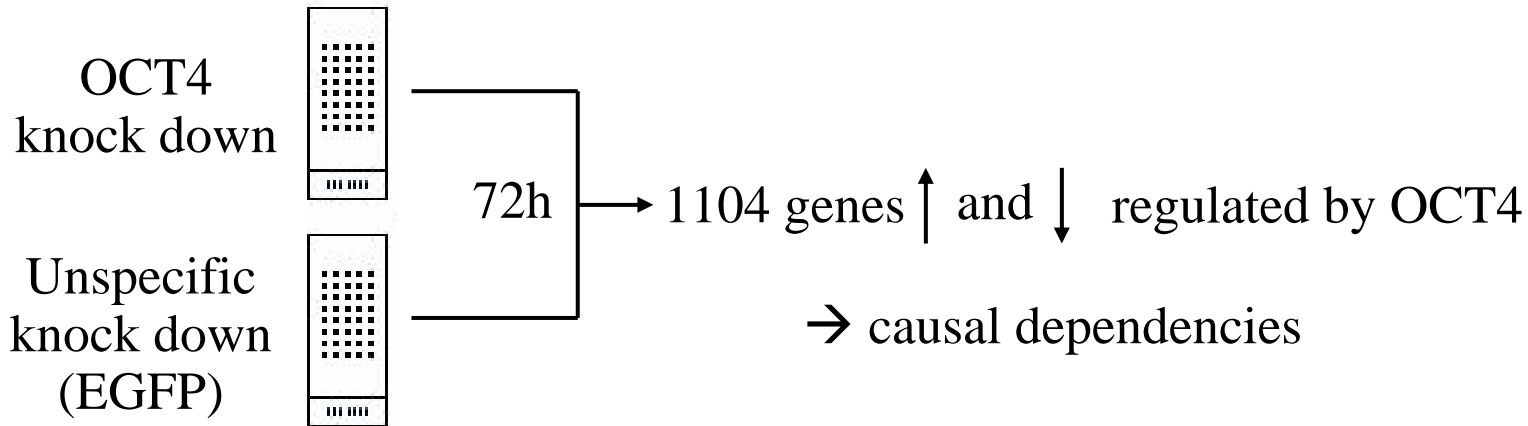
**Boyer et al.: Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells, Cell 2005**





## 5.3. OCT4 RNAi silencing in hESCs

In house cDNA array  
~16.000 genes



**Babaie, Greber, Adjaye et al.: Analysis of OCT4 dependent transcriptional networks regulating pluripotency in human embryonic stem cells, Stem Cells 2006**

## 5.4. OCT4 motifs

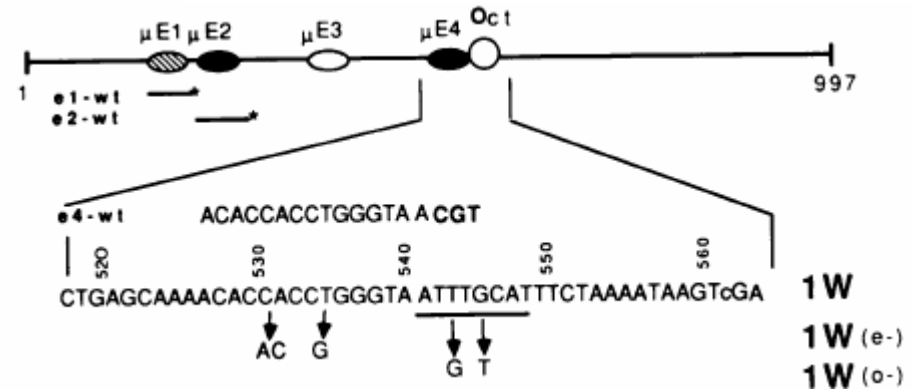
The EMBO Journal vol.8 no.9 pp.2551 – 2557, 1989

### Octamer binding proteins confer transcriptional activity in early mouse embryogenesis

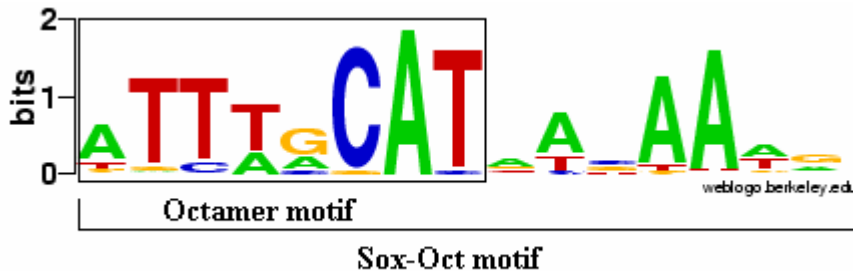
Hans R.Schöler, Rudi Balling,  
Antonis K.Hatzopoulos, Noriaki Suzuki and  
Peter Gruss

Max-Planck Institute for Biophysical Chemistry, Department of  
Molecular Cell Biology, 3400 Göttingen, FRG

Communicated by P.Gruss



→ DNA octamer motif interacting with POU domain factors like the homeodomain containing transcription factor Oct4



Motif recognizing a Sox2/Oct4 heterodimer was identified by

- sequence conservation analysis
- ChIP-PET and ChIP-Chip analysis in mouse and human



## 5.5 ChIP-on-Chip data analysis- update

**Boyer et al.:** *Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells*, Cell 2005

- 400.000 oligos designed for NCBI 35 by Boyer et al.

### 5.5.1 Probe assignment update:

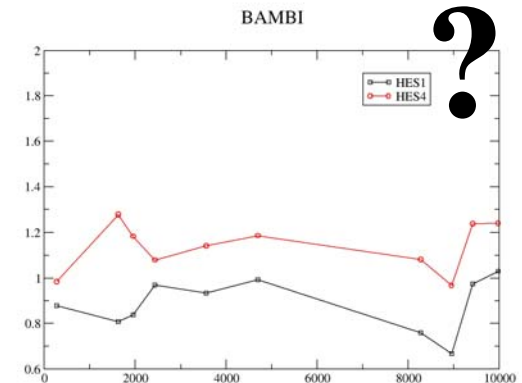
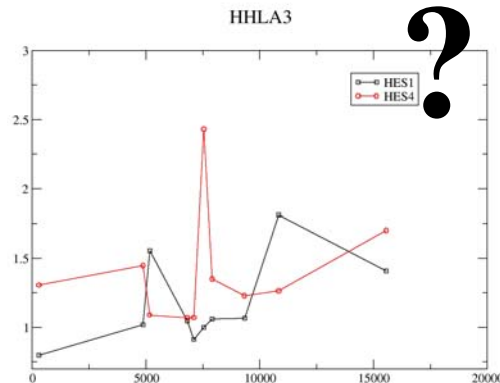
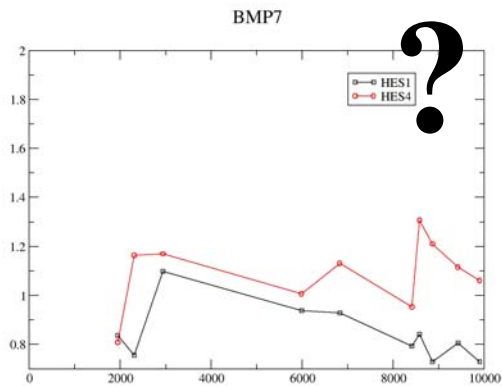
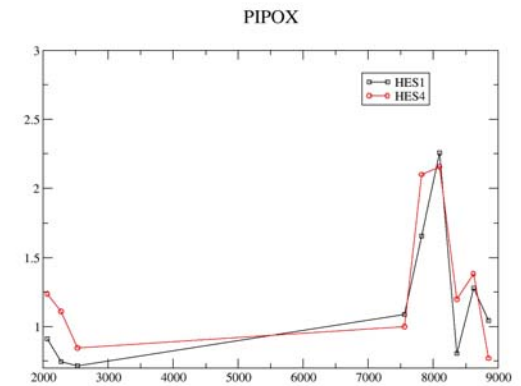
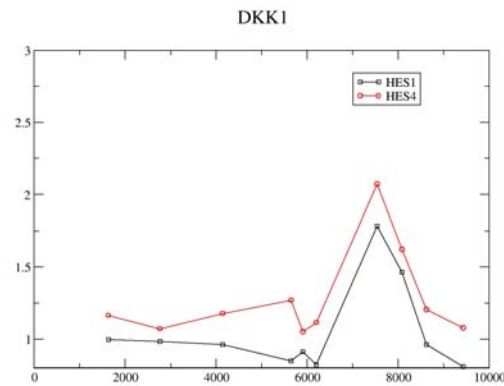
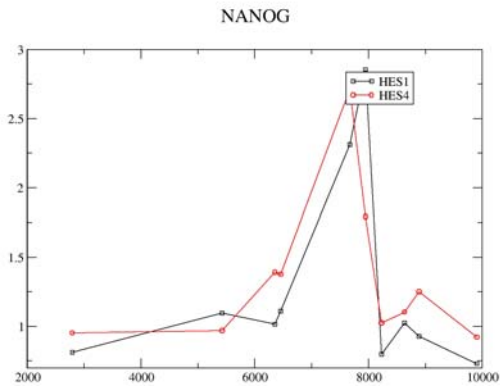
- Sequence alignment (BLAT) of all oligos against the human genome (NCBI 36)
  - Updated assignment: *oligo*  $\leftrightarrow$  *genomic position*
- RAW data selection (ChIP, IP and BG) for all uniquely mapped oligonucleotides



## 5.5. ChIP-on-Chip quality control and data analysis (2)

### 5.5.1. Quality Control (2):

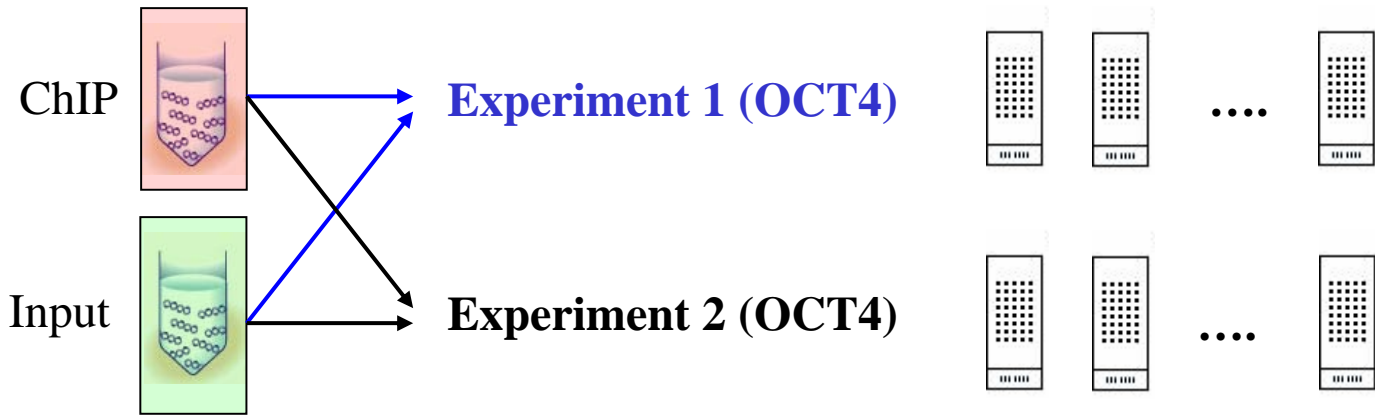
OCT4 RAW data plotted (ChIP/IP ratio) for both replicas:



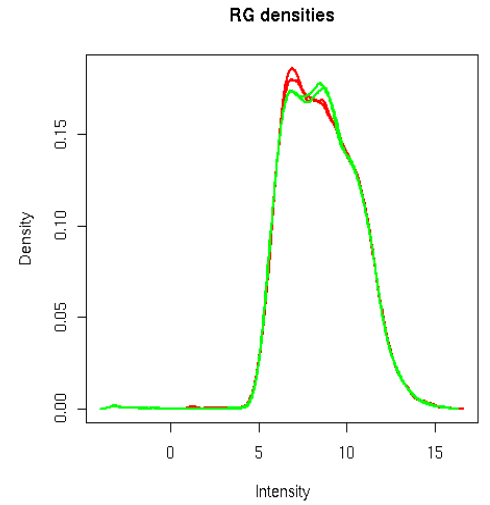
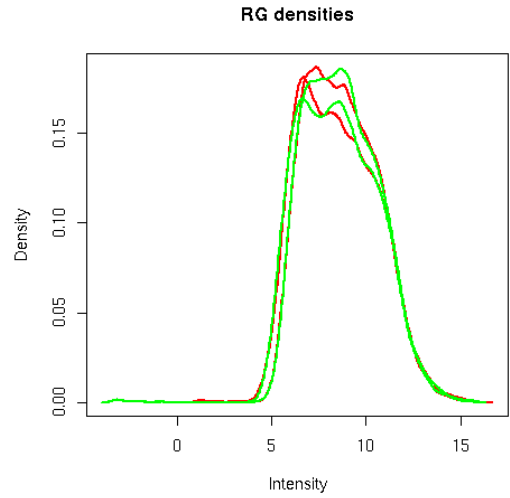
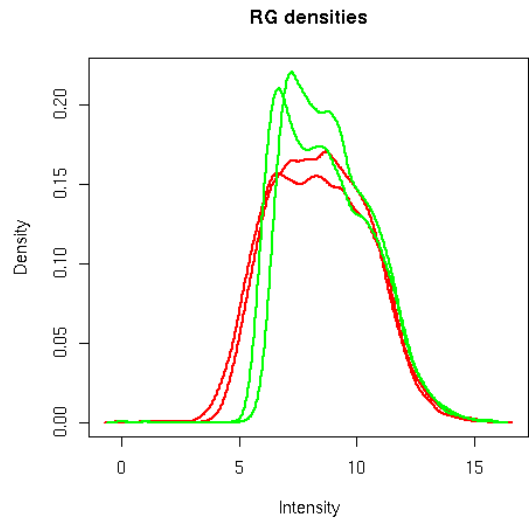


# 5.5. ChIP-on-Chip quality control and data analysis (3)

## 5.5.2. Data Normalization (biconductor/ limma package)



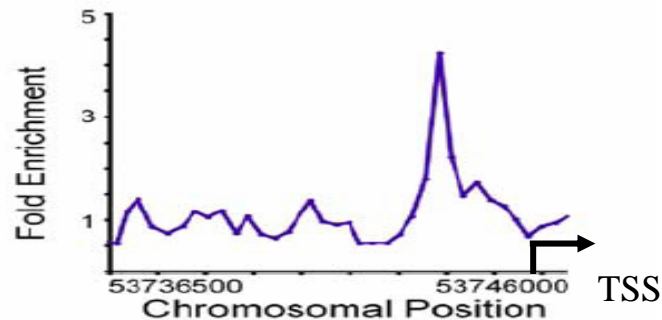
For each pair of arrays:



## 5.5. ChIP-on-Chip quality control and data analysis (4)

### 5.5.3 Enrichment analysis and peak annotation

- **ratio** calculation
- **threshold** determination  $t_0$  based on the ratio distribution
- **interval** analysis
- gene **annotation** of enriched regions



→ 308 validated direct OCT4 targets



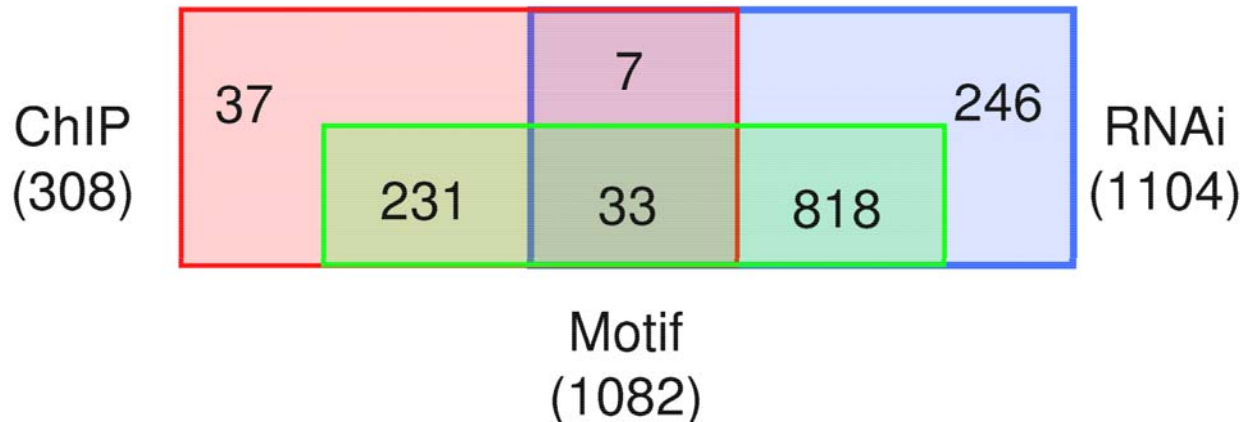
## 6. Integrative analysis

### 6.1. Motivation

- reduce noisiness of the different approaches
- combine information of direct targets and gene regulation
- get more secure targets

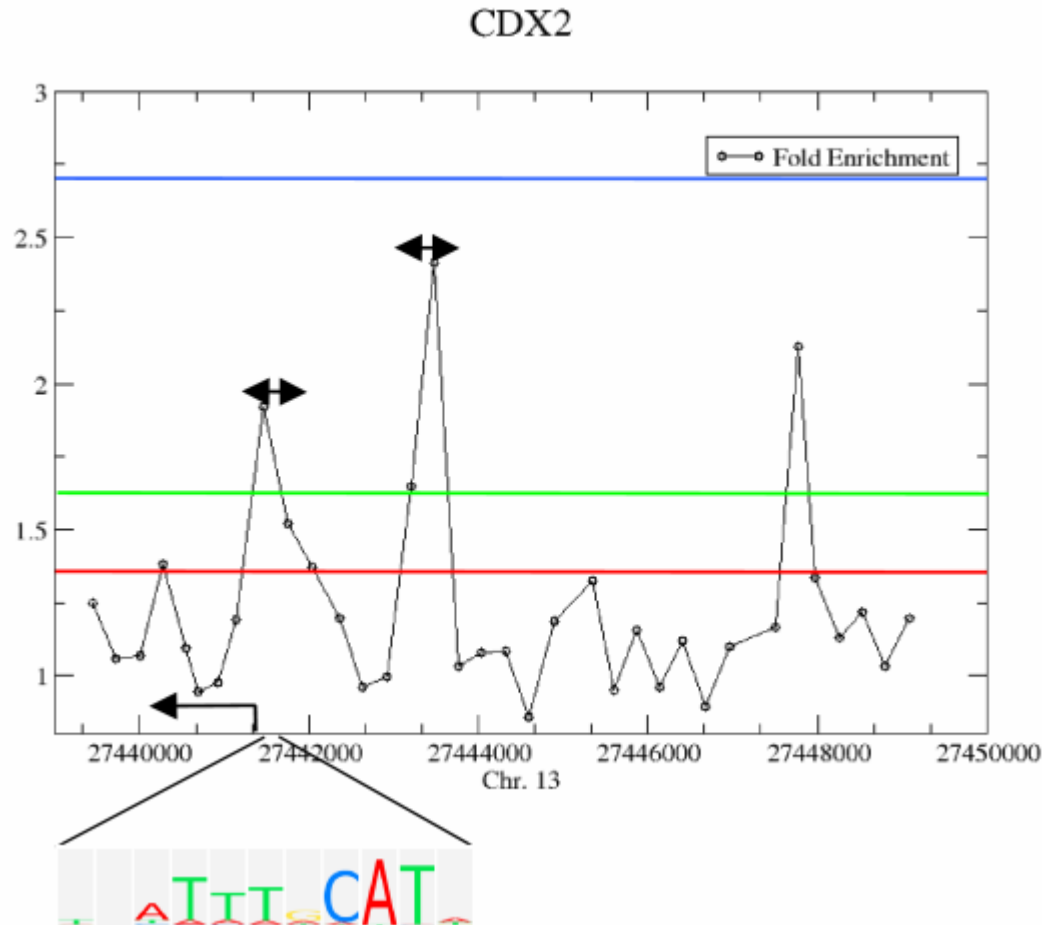
### 6.2. Results

Overlap of processed ChIP-on-Chip, RNAi  
and motif mapping results



## 6. Integrative analysis (2)

### 6.3. Example of core OCT4 target genes

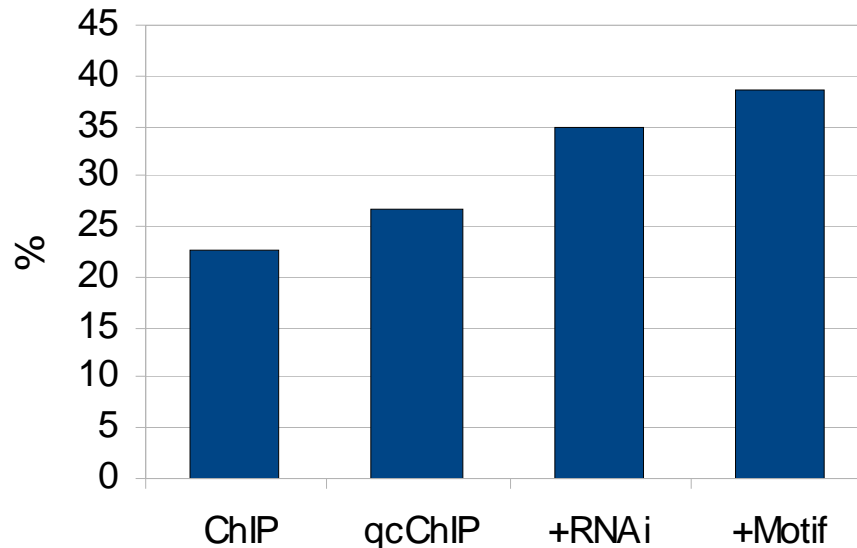


## 6. Integrative analysis (3)

### 6.4. Enrichment analysis

- Gene enrichment analysis: DAVIDS
- Most enriched GO term within the original Boyer data set:
  - **Transcription Factor Activity**

#### Transcription Factor Activity

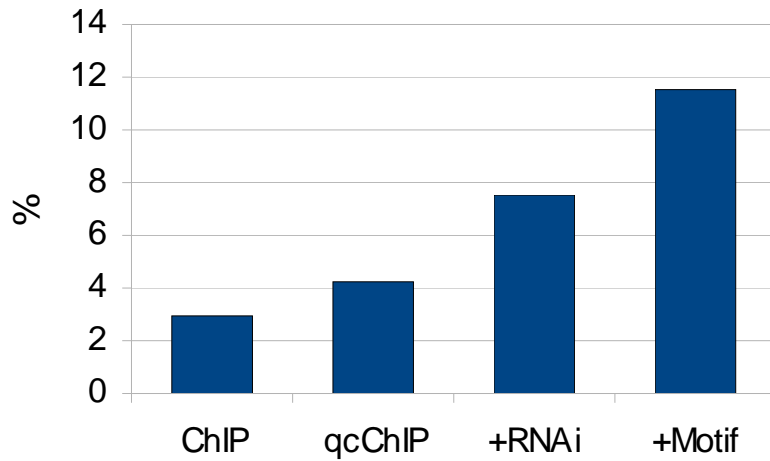




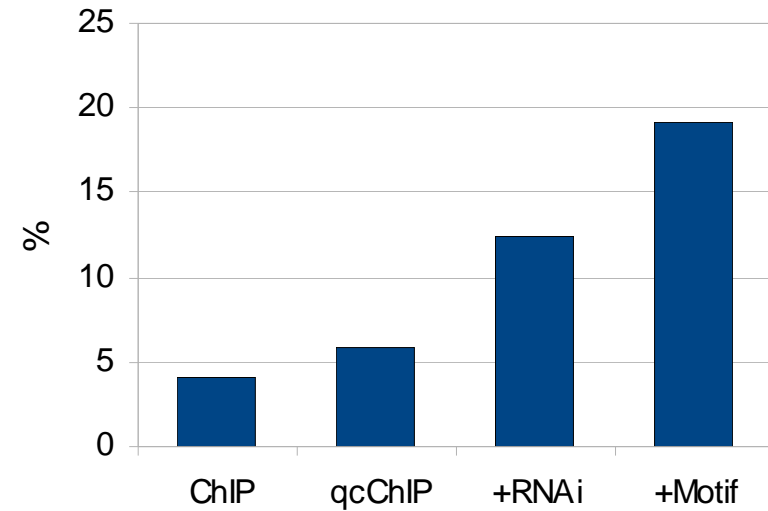
## 6. Integrative analysis (4)

### 6.4. Enrichment analysis

Regulation of Cell  
Differentiation



Cell Fate Commitment

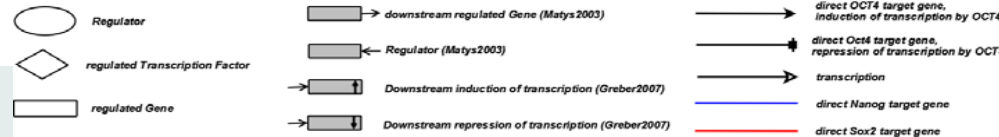
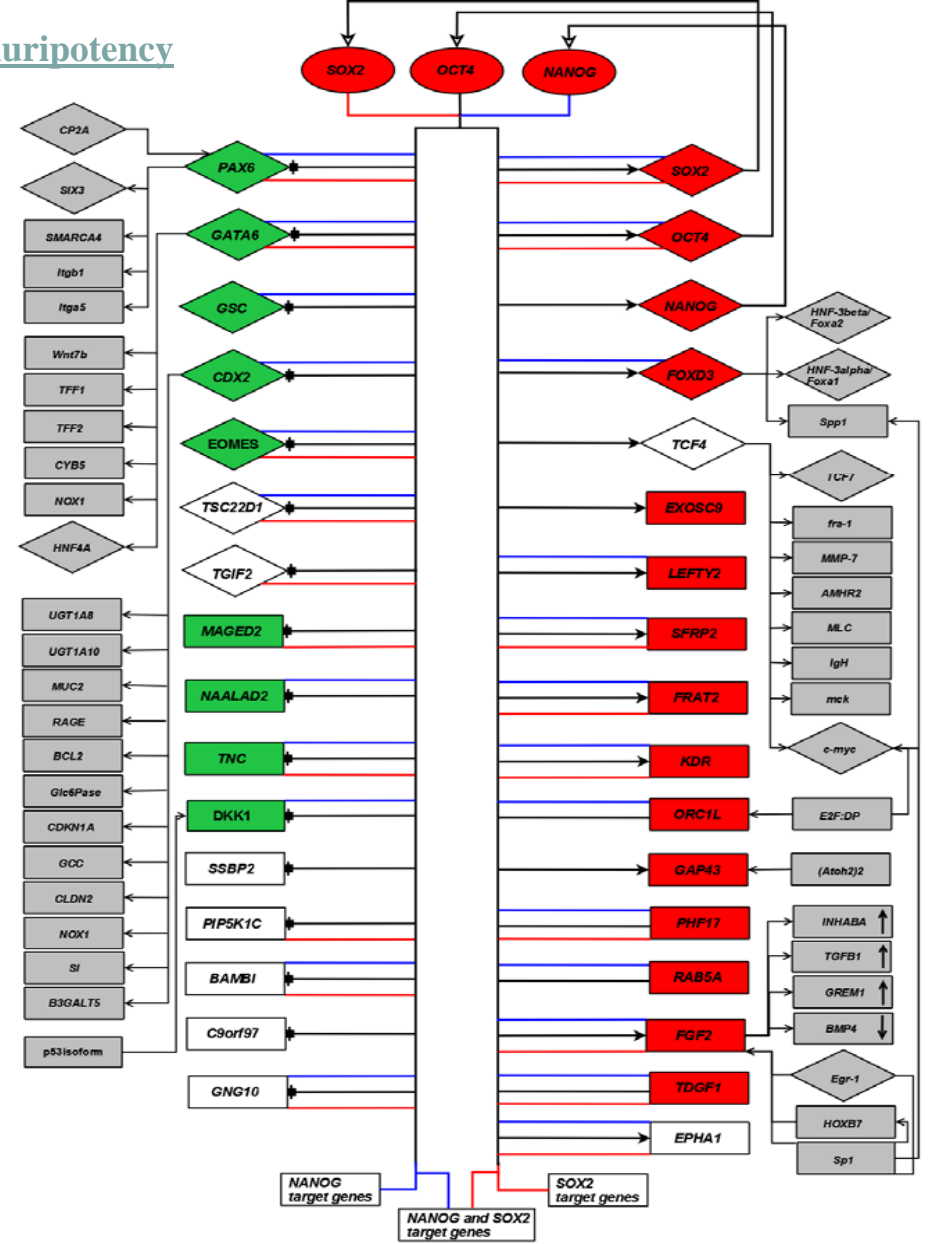




# 6. Integrative analysis (5)







## 6.5. Core Network

- 33 core OCT4 target genes
- shows positive and negative OCT4 influence
- points out core downstream transcription factors
- includes QC SOX2 and NANOG target genes
- functional cross-annotation of target genes: stemness/ differentiation genes
- extended by downstream targets from TRANSFAC



## 7. De novo motif discovery

- Based on promotor subsequences (200bp length) of highly confirmed functional OCT4 target genes

Motif	Complexity score	Genes containing motif in promoter sequence (-8kb to TSS)	Similar to known motifs in TRANSFAC and Jaspar (top 5 each)
	1.4654	PIP5K1C SFRP2 TDGF1 TSC22D1	<i>Nrf-1, Oct-1, TFE, DBP, AFP1, Arnt-Ahr, ZNF354C, Broad-complex 4, FOXD1, sna</i>
	1.3883	PIP5K1C	<i>TFE, bHLH66, TWI, AFP1, c-Myc, Mycn, Arnt, USF1, MYC-MAX, ARR10</i>
	1.1945	BAMBI CDX2 DKK1 EOMES EPHA1 EXOSC9 FGF2 FOXD3 GATA6 GNG10 LEFTY2 MAGED2 NAALAD2 NANOG PAX6 RAB5A SSBP2 TGIF2 TNC TSC22D1	<i>Oct-1, Octamer, HMG1Y, HMG, OCT-x, Agamous, NR2F1, RELA, Pdx1, Pax4, (HNF4A)</i>
	1.1945	BAMBI C9orf97 CDX2 DKK1 EOMES EPHA1 EXOSC9 FOXD3 FRAT2 GAP43 GATA6 GSC LEFTY2 NAALAD2 NANOG ORC1L PIP5K1C POU5F1 RAB5A SFRP2 SSBP2 TCF4 TGIF2 TSC22D1	<i>LUN-1, Alfin1, TAL1, HEB, ROM, Myf, NHLH1, ZEB1, SP1, sna</i>
	0.9730	C9orf97 CDX2 DKK1 EPHA1 EXOSC9 FOXD3 GAP43 GNG10 SFRP2 SOX2 TDGF1 TSC22D1	<i>Oct-4, Oct-1, Octamer, OCT-x, POU3F2, HMG-1Y, Pdx1, SQUA, Prrx2, STAT1</i>
	0.8941	C9orf97 CDX2 DKK1 EOMES EXOSC9 FOXD3 GAP43 GNG10 KDR NAALAD2 NANOG POU5F1 SFRP2 SOX2 TCF4 TDGF1 TGIF2 TNC TSC22D1	<i>PU.1, KROX, ZNF219, Retroviral, Nrf-2, GABPA, Klf4, SPI1, SPIB, id1</i>

→ Set of **18 motifs** including binding sites similar to **known motifs** recognized by TFs which are **closely interconnected** to the network.

# Acknowledgment

- Ralf Herwig
- James Adjaye
- Bioinformatics Group
- Hans Lehrach

