V6: Core Pluripotency Network

Mouse ES cells were isolated for the first time in 1981 from mouse blastocysts.

Maintenance of the self-renewing state of mouse ES cells requires the cytokine **leukemia inhibitory factor (LIF).**

The binding of LIF to its receptor activates STAT3 through phosphorylation. **STAT proteins** (Signal Transducer and Activator of Transcription) regulate many aspects of growth, survival and differentiation in cells.

Cytokines (Greek cyto-, cell; and -kinos, movement) are small signaling protein molecules that are used extensively in **intercellular communication**..

They are secreted by the glial cells of the nervous system and by numerous cells of the immune system.

Chen et al., Cell 133,
1106-1117 (2008)

Cytokines

Cytokines can be classified as proteins, peptides, or glycoproteins.

The term "cytokine" encompasses a large and diverse family of regulators produced throughout the body by cells of diverse embryological origin.

Sometimes, the term "cytokine" is restricted to immunomodulating agents, such as **interleukins** and **interferons**.

Biochemists disagree as to which molecules should be termed cytokines and which hormones.

www.wikipedia.org

Cytokines vs. hormones

Concentration:

Classic **protein hormones** circulate in nanomolar (10^{-9} M/l) concentrations that usually vary by less than one order of magnitude.

In contrast, some **cytokines** (such as IL-6) circulate in picomolar (10^{-12} M/l) concentrations that can increase up to 1,000-fold during trauma or infection.

Production:

Virtually all nucleated cells, but especially endo/epithelial cells and resident macrophages are potent producers of IL-1, IL-6, and TNF-α.

In contrast, classic hormones, such as insulin, are secreted from discrete glands (e.g., the pancreas).

www.wikipedia.org

Response to cytokines

Each cytokine has a matching **cell-surface receptor**.

Subsequent **cascades of intracellular signalling** then alter cell functions.

This may include
- the upregulation and/or downregulation of several genes and their TFs, resulting in the production of other cytokines,

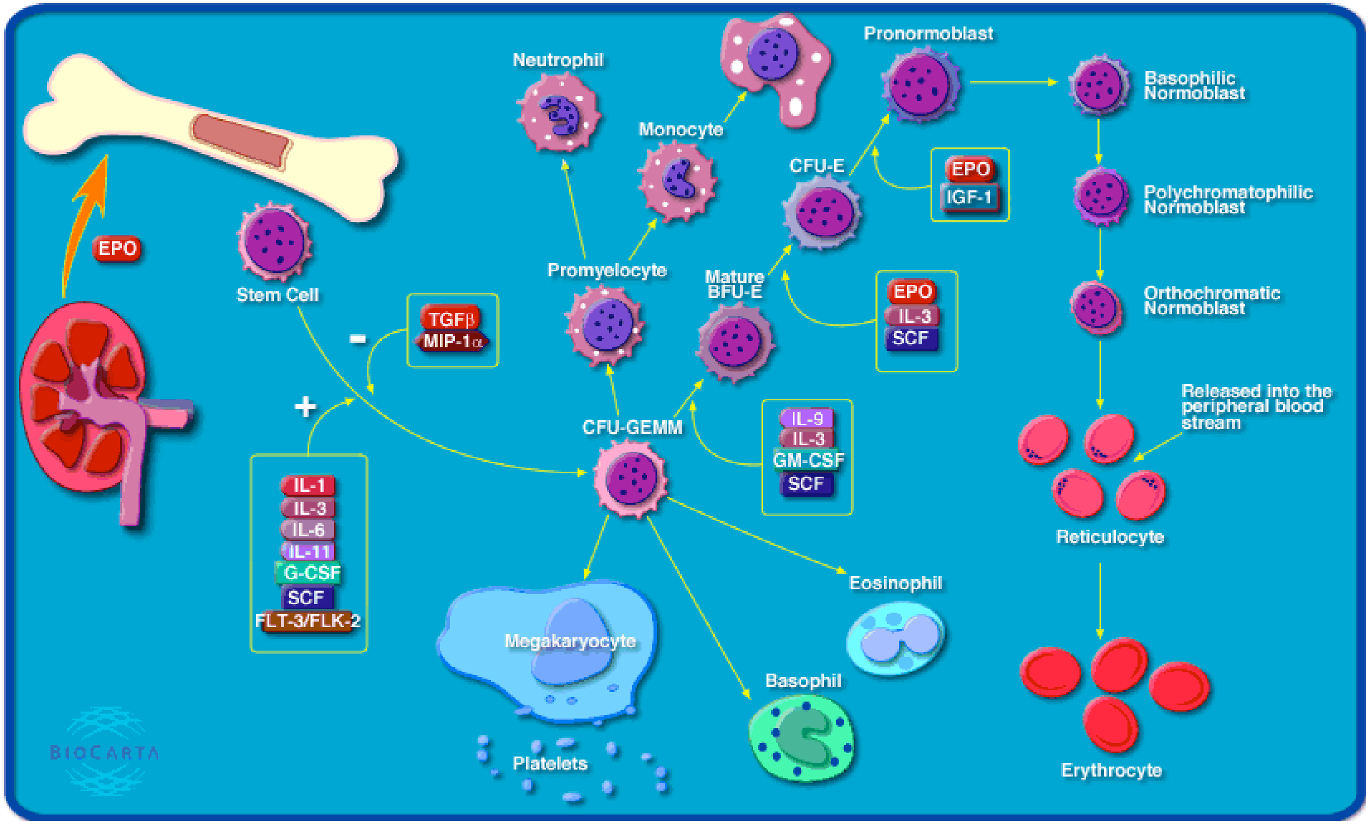
- an increase in the number of surface receptors for other molecules,
-
- or the suppression of their own effect by feedback inhibition.

www.wikipedia.org

Erythrocyte differentiation pathway

Stem cells in the bone marrow produce a variety of hematopoietic cell types from common progenitor cells under the influence of cytokines and growth factors.

CFU-GEMM cells are a key intermediate in the differentiation of granulo–cytes, erythrocytes, monocytes and megakaryocytes.



Erythropoietin (EPO) is a cytokine produced in the kidneys that, along with other cytokines, induces red blood cell (erythrocyte) differentiation in the bone marrow from CFU-GEMM cells.
As the erythrocyte lineage progresses, cells lose their nuclei, and move out of the bone marrow into circulation. The ability of EPO to selectively induce red blood cell differentiation has allowed extensive therapeutic use of the recombinant form of this cytokine to treat anemias. EPO is also used for doping!

www.biocarta.com

BMPs

LIF alone, however, is not sufficient to maintain ES cells, as their maintenance also requires the presence of **fetal calf serum**.

Fetal bovine serum (**FBS**) or **fetal calf serum** is the blood fraction remaining after the natural coagulation of blood (*dt. Blutgerinnung*), followed by centrifugation to remove any remaining red blood cells.

Fetal bovine serum is the most widely used serum-supplement for the ***in vitro*** **cell culture** of eukaryotic cells.
It contains very low levels of antibodies, many growth factors. One of its major components is the globular protein, bovine serum albumin (BSA).

Bone morphogenetic proteins (BMPs) appear to be key serum-derived factors that act in conjunction with LIF to enhance the self-renewal and pluripotency of mouse ES cells.

Chen et al., Cell 133, 1106-1117 (2008)

Integrating effect of signaling pathways

The binding of **BMP4** to its receptors triggers the phosphorylation of **Smad1** and activates the expression of members of the **Id** (inhibitor of differentiation) **gene family**.

ES cells overexpressing Ids can self-renew in the absence of BMP4.

Thus, induction of Id expression is likely a critical contribution of the BMP/Smad pathway.

Hence, the LIF and BMP signaling pathways play a central role in the maintenance of a pluripotential stem cell phenotype.

Besides these signaling pathways, which sense the presence of extrinsic growth factors in the environment, **intrinsic factors** such as **transcription factors (TFs)** are also essential for specifying the undifferentiated state of ES cells.

Chen et al., Cell 133, 1106-1117 (2008)

TFs in Core Pluripotency Network

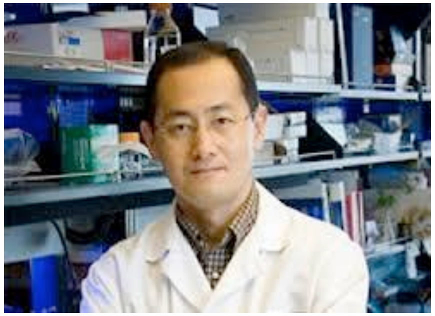
Oct4, encoded by ***Pou5f1***, is a POU domain-containing TF known to be essential to ES cells and early embryonic development.

Oct4 interacts with **Sox2** (an HMG-containing TF).
Genome-wide mapping of OCT4 and SOX2 sites in human ES cells shows that they cotarget multiple genes.

The cis-regulatory element to which the Sox2-Oct4 complex is bound consists of neighboring sox (50-CATTGTA-30) and oct (50-ATGCAAAT-30) elements.

Recent works indicate that Oct4 and Sox2, along with **c-Myc** and **Klf4**, are sufficient for reprogramming fibroblasts to **induced pluripotent stem cells (iPS),** which are functionally similar to ES cells.

(→ Yamanaka factors).



Shinya Yamanaka
noble price for medicine 2012

Chen et al., Cell 133, 1106-1117 (2008)

Other TFs in Core Pluripotency Network

Hence, these TFs can exert a dominant role in reconstructing the transcriptional regulatory network of ES cells.

A third well-studied TF in ES cells is **Nanog**. Nanog is a homeodomain-containing TF that can sustain pluripotency in ES cells even in the absence of LIF.

Other transcriptional regulators are required as well to maintain ES cells in the state of pluripotency.

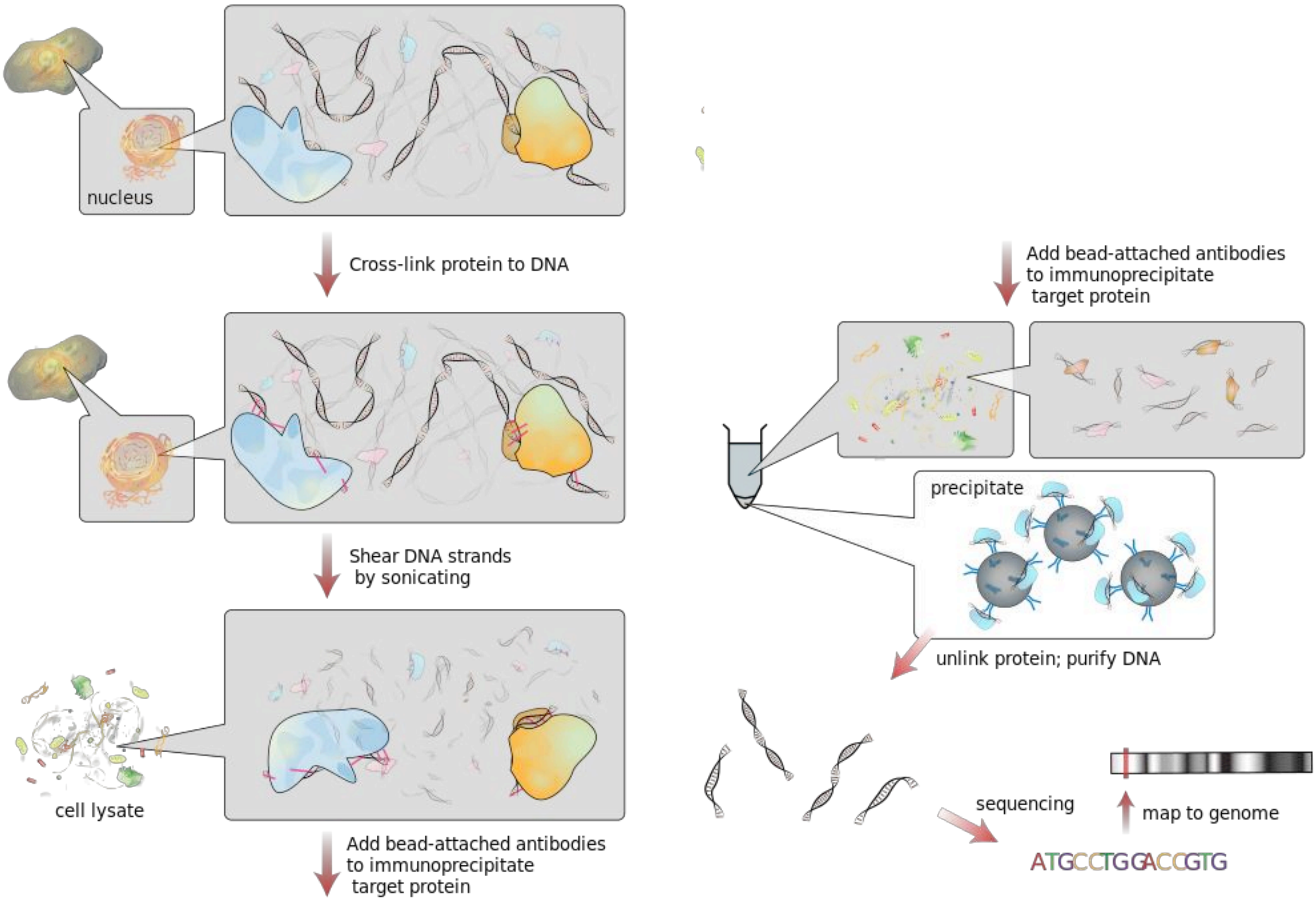
Through genetic studies, Esrrb and Zfx have been shown to regulate the self-renewal of ES cells.

Q: How do these TFs control cell fate?

Idea: map the *in vivo* binding loci for 13 sequence-specific TFs and 2 transcription coregulators in living mouse ES cells using ChIP-seq experiments.

Chen et al., Cell 133, 1106-1117 (2008)

Chromatin Immunoprecipitation



Cross-link protein to DNA

Shear DNA strands by sonicating

Add bead-attached antibodies to immunoprecipitate target protein

Add bead-attached antibodies to immunoprecipitate target protein

unlink protein; purify DNA

sequencing

map to genome

www.wikipedia.org

Other TFs in Core Pluripotency Network

Nanog, Oct4, Sox2, Esrrb, and **Zfx** are known regulators of pluripotency and/or self-renewal.

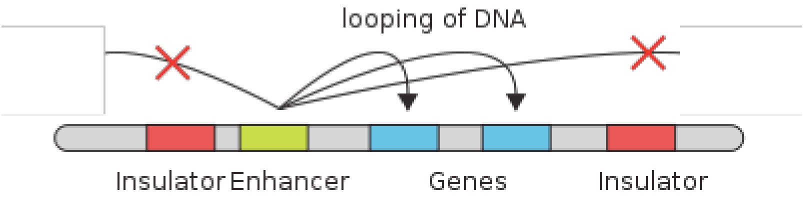
Smad1 and **STAT3** are key components of the signaling pathways mediated by BMP and LIF, respectively.

Tcfcp2l1 is preferentially upregulated in ES cells.

E2F1 regulates cell-cycle progression and associates extensively with promoter regions.

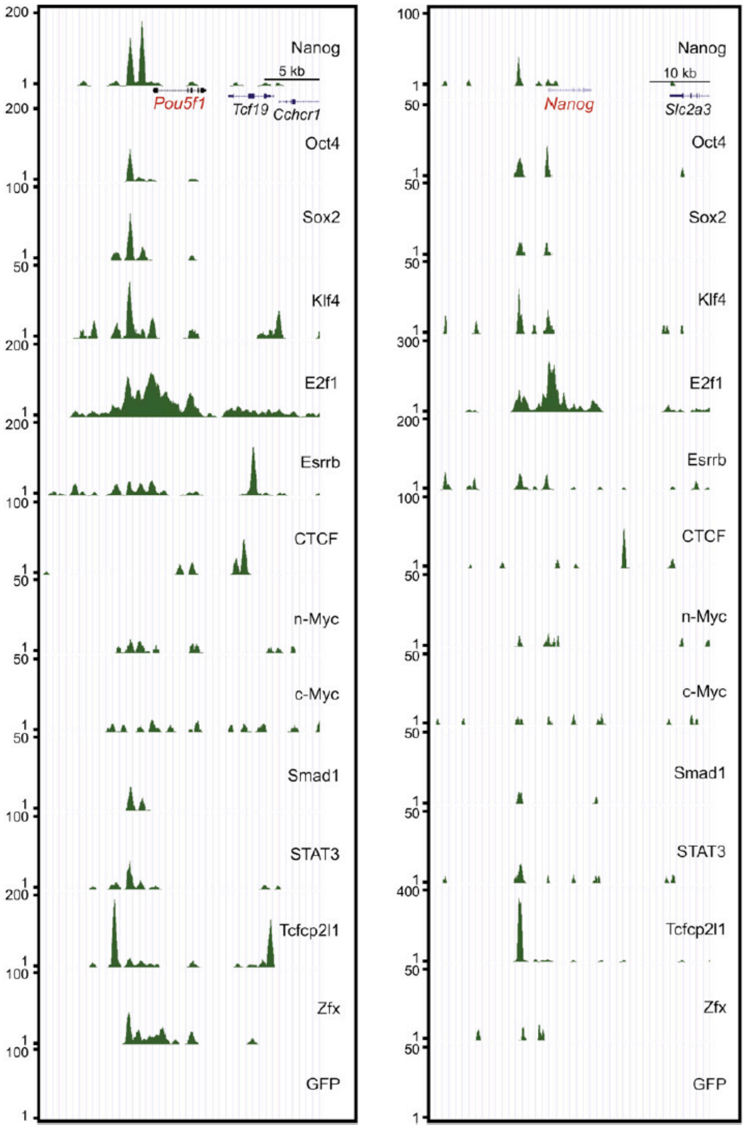
Klf4 and **Myc** TFs are reprogramming factors that are also implicated in the maintenance of the undifferentiated state of ES cells

CTCF is required for transcriptional insulation.



Chen et al., Cell 133, 1106-1117 (2008)

Binding data at Oct4 and Nanog gene loci

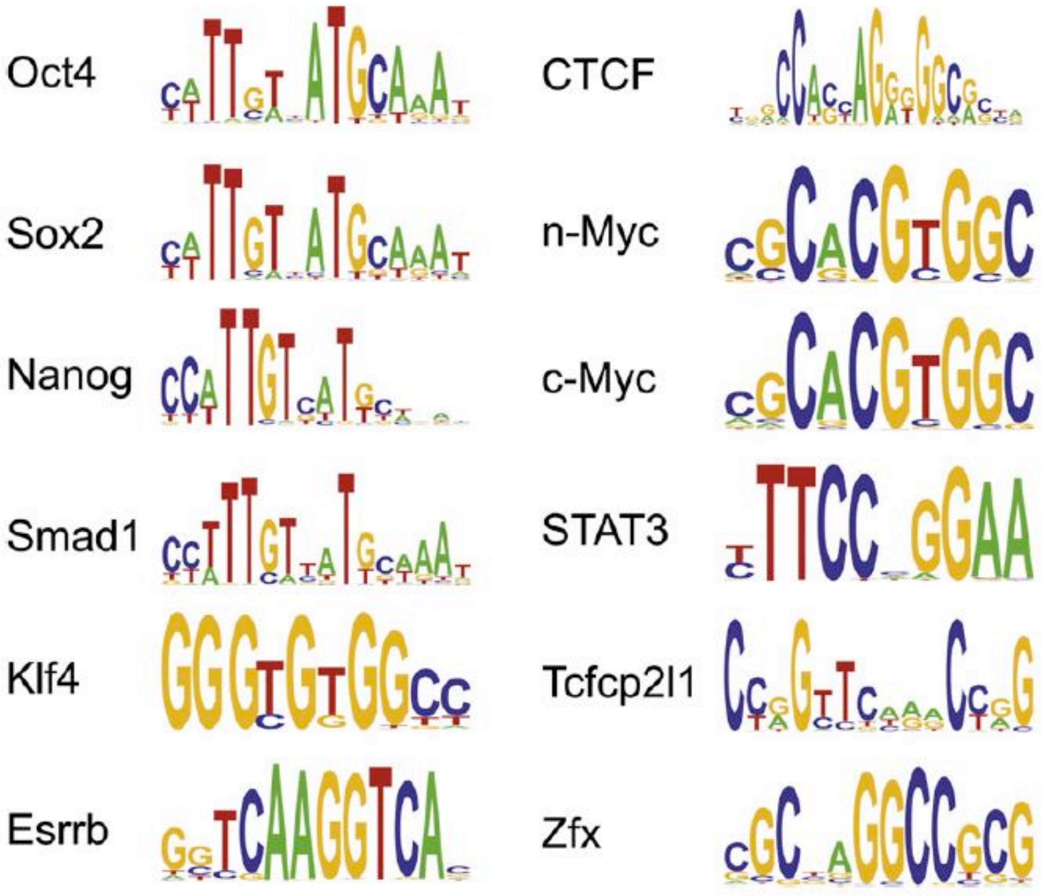


Shown as examples are the binding profiles for all 13 factors at the Pou5f1 (Oct4) and Nanog gene loci.

Between 1,126 and 39,609 TF-binding sites (TFBSs) were found for the 13 factors.

Chen et al., Cell 133, 1106-1117 (2008)

Derivation of binding motifs



To determine the in vivo sequence specificity of the TFs, consensus sequence motifs were derived by a de novo motif-discovery algorithm.

- sequences (±100 bp) from the top 500 binding peaks were selected from each factor,
- repeats were masked, and
- the program `Weeder` was used to find sequences that are overrepresented against what is expected to occur randomly.

Chen et al., Cell 133,
1106-1117 (2008)

Motif Discovery

Computational methods for the discovery of novel motifs in a set of sequences of co-regulated genes are typically based on two steps.

(1) One or more groups of oligonucleotides similar enough to each other (i.e. differing in some nucleotide substitutions) are detected in the sequences.

(2) Their presence is evaluated from a statistical point of view.

For this, algorithms estimate how likely each group would be to appear in a set of sequences that are

- either picked at random from the same organism (thus they are very unlikely to be coregulated) or
- built randomly with the same nucleotide composition as the input sequences (thus very likely to present a different oligo composition).

The best groups of oligos found are likely to be **binding sites** for some **TFs**.

Pavesi et al., Nucl Ac Res 32, W199-W203 (2004)

Sox2-Oct4 pair

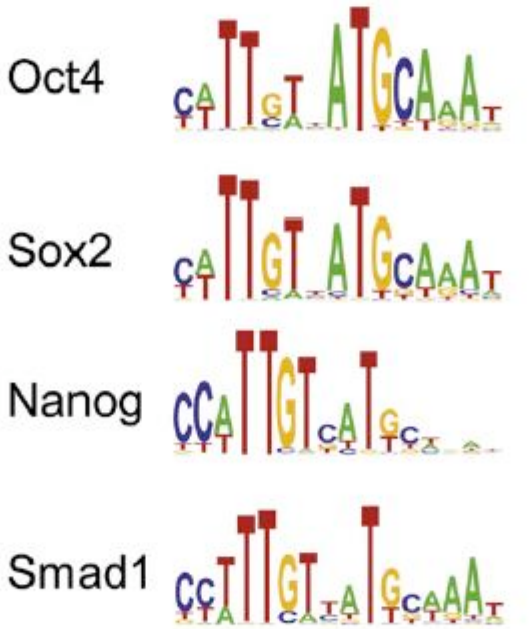
From both the Oct4 and Sox2 data sets, one obtains a sox-oct composite element consisting of

- a **Sox2-binding site consensus** (5’-CATTGTT-3’) and
- a **Oct4-binding sequence** (5’-ATGCAAAT-3’) adjacent to one another.

The presence of a common motif suggests that the Sox2 and Oct4 heterodimer is the functional binding unit.

Interestingly, the de novo-predicted matrices for Nanog- and Smad1-bound sequences resemble the sox-oct joint motif.

This reflects the frequent co-binding of Nanog and Smad1 with Sox2 and Oct4.



Chen et al., Cell 133, 1106-1117 (2008)

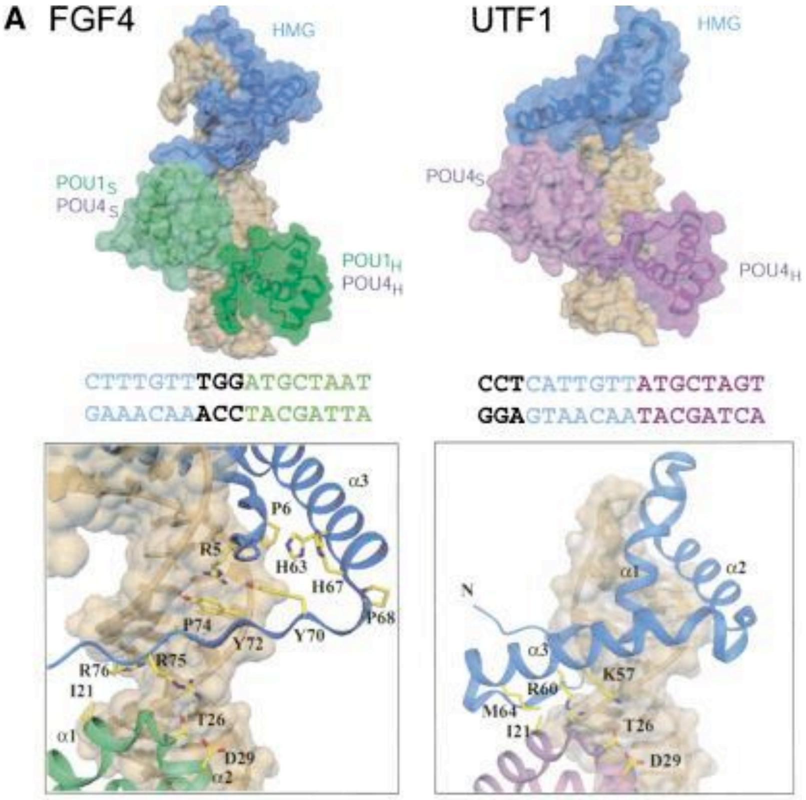
Complexes of the central regulator Oct4

The POU domain is a part of Oct1 and Oct4.

(Left) Atomic X-ray structure of the complexes of the transcription factors POU and Fibroblast growth factor 4 (FGF4) with DNA.

(Right) Atomic X-ray structure of the complexes of POU and undifferentiated embryonic cell transcription factor 1 (UTF1) with DNA.

Note the slightly different positions of the TFs which lead to recognition of slightly different DNA motifs.



Remenyi et al. Genes Dev 15, 2048 (2003)

MTLs: multiple TF-binding loci

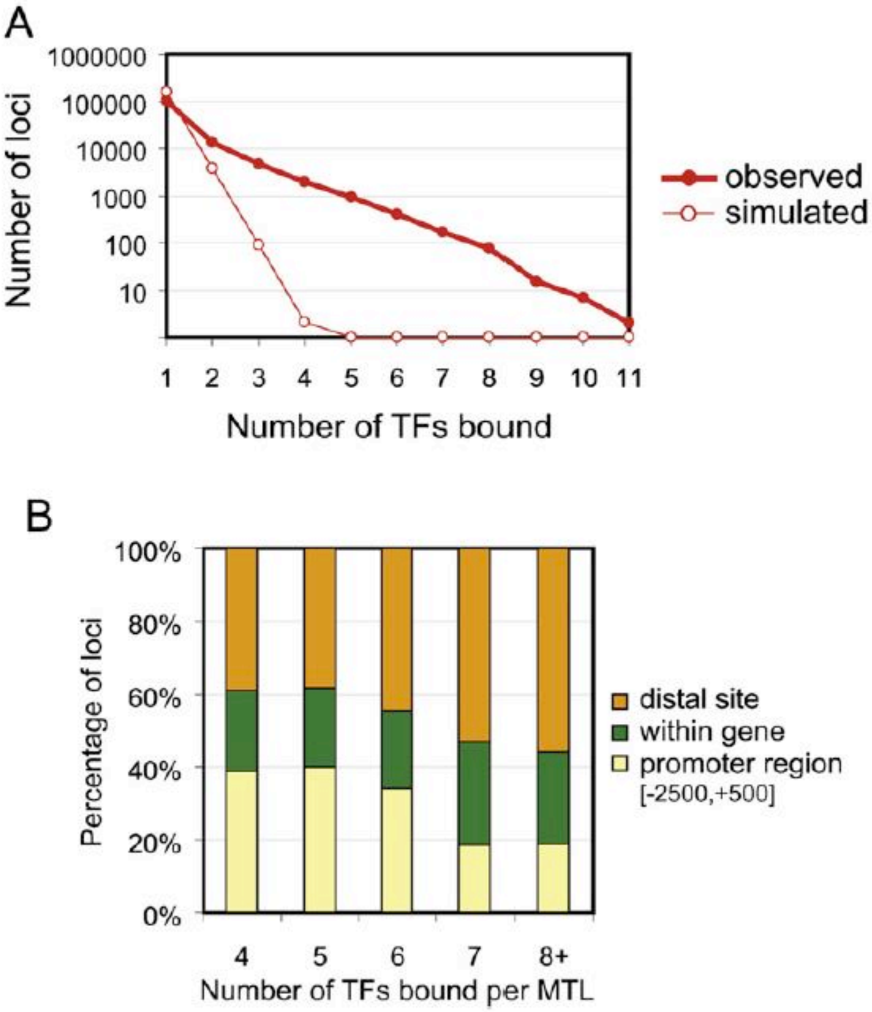
Upon close examination of the binding profiles from these 13 TFs, we found that a subset of binding sites was bound by many of these TFs.

To investigate their biological relevance, we first determined the significance of such enrichments of TFBSs.

Peak sites within 100 bp were iteratively clustered to define **multiple transcription factor-binding loci (MTL).**

Chen et al., Cell 133, 1106-1117 (2008)

Multiple TF-Binding Loci

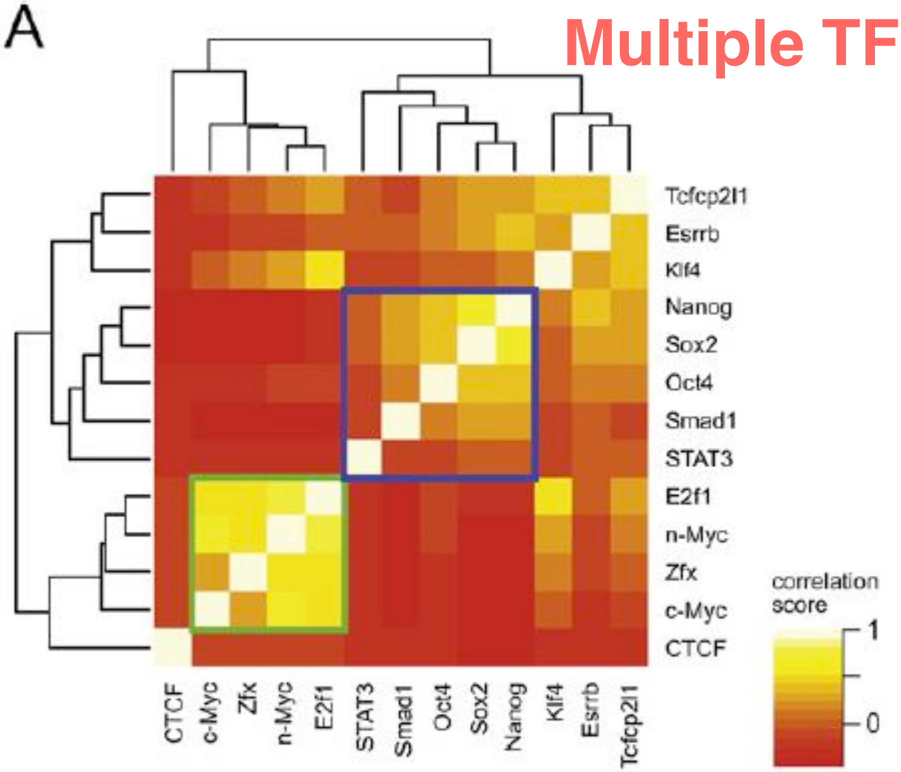


(A) Number of TFs bound per co-bound locus. The distribution of randomly occurring co-bound loci is obtained by simulation.
Loci bound by 4 or more TFs are highly significant ($p < 0.001$).
Of these, 40.2% were found in the intergenic regions, and the remaining loci were spread between promoter regions (37.2%) and within gene regions (22.6%).

Less than 20% of the clusters with 7 or more TFs are found at promoter regions (yellow columns), compared with 40% of the clusters that have fewer than 5 TFs.
Hence, the co-occurrence of TFBSs within the MTL is not mainly due to their occurrence at promoters.

Chen et al., Cell 133, 1106-1117 (2008)

Multiple TF-Binding Loci



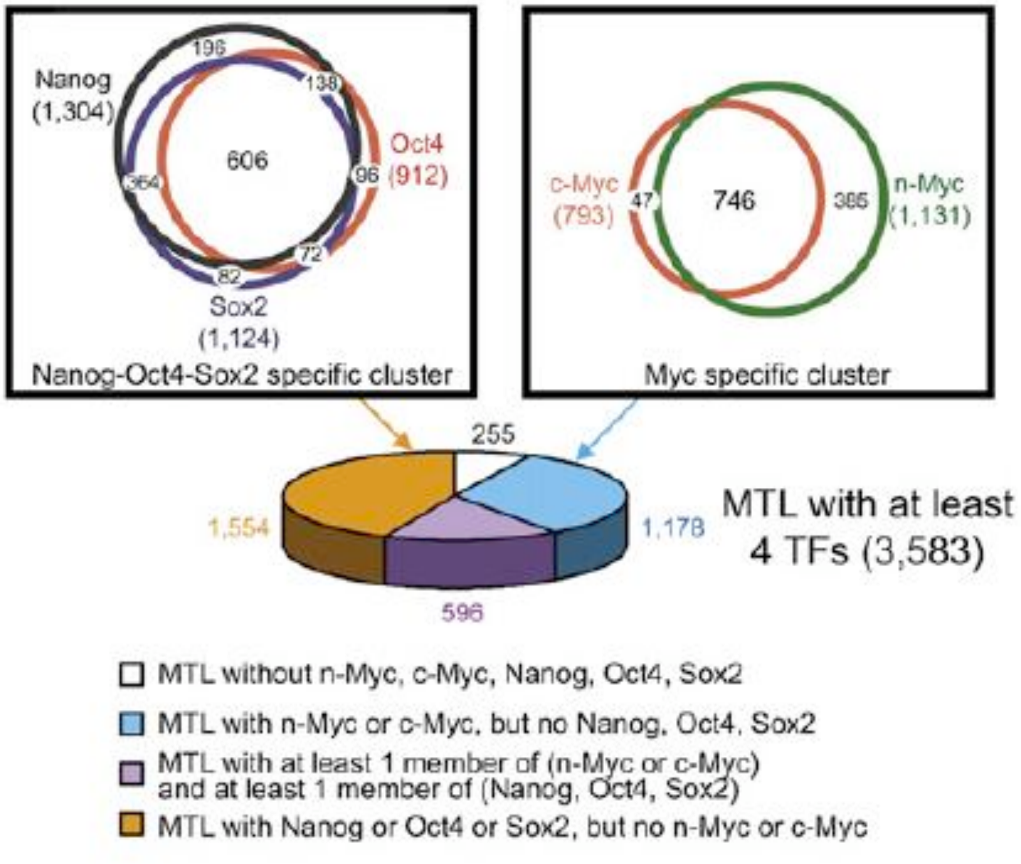
Among the 13 TFs, Nanog, Sox2, Oct4, Smad1, and STAT3 (**blue box**) tend to co-occur quite often.

The same is found for members of a second, distinct group comprised of n-Myc, c-Myc, E2f1, and Zfx (**green box**).

Co-occurrence of transcription factor (TF) groups within MTL. Colors in the heat map reflect the colocalization frequency (occurrence) of each pair of TFs in MTL.
Yellow means more frequently colocalized, red means less frequent.

Chen et al., Cell 133, 1106-1117 (2008)

TF-composition of MTLs



Two major clusters exist within the 3583 MTL.

The first group (**orange** sector) consists of Oct4, Nanog, or Sox2, but not n-Myc and c-Myc.

The second group (**light-blue** sector) consists of n-Myc or c-Myc, but not Oct4, Nanog, and Sox2.

The purple sector is a mixture of the first two groups (orange and light-blue sectors).

Chen et al., Cell 133, 1106-1117 (2008)

Determine association of genes with TFs (I)

Let ***g*** represent a **non-redundant gene** and ***tf*** represent a **TF** in our dataset,

We assume that the probability of gene *g* being the target of *tf* is dependent on the location of the nearest *tf* binding site, denoted by ***l*(*g*,*tf*)**, relative to the TSS of *g*.

To quantitatively estimate the location-dependent association between genes and TFs, we first aligned the TSSs of all 17,442 genes and divided the genomic locations into 16 bins, separated by

{-100k bps, -50k bps, -20k bps, -10k bps, -5k bps, -2k bps, -1k bps, 0 bps, 1k bps, 2k bps, 5k bps, 10k bps, 20k bps, 50k bps, 100k bps},

and the histogram of *l*(*g*,*tf*) was calculated based on these bins.

Chen et al., Cell 133, 1106-1117 (2008)

Determine association of genes with TFs (II)

Next, we randomized the *tf* binding sites by uniformly placing them in the whole genome and calculated the histogram of *l*(*g*,*tf*) based on the random model.

Comparing these two histograms, we are able to approximate the proportion of non-random *tf* -gene association for each bin.

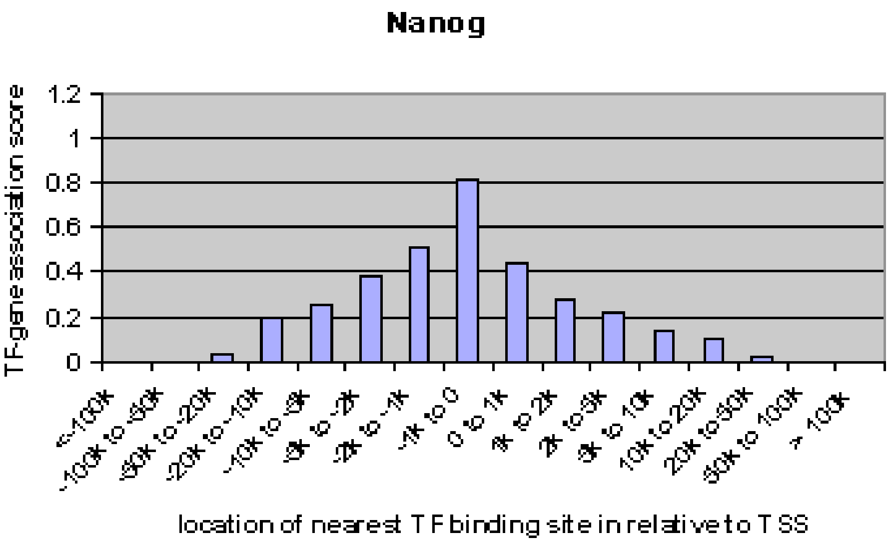
Given a pair, *g* and *tf* , let *k* represent the index of bin corresponding to *l*(*g*,*tf*), the **association score** was calculated by:

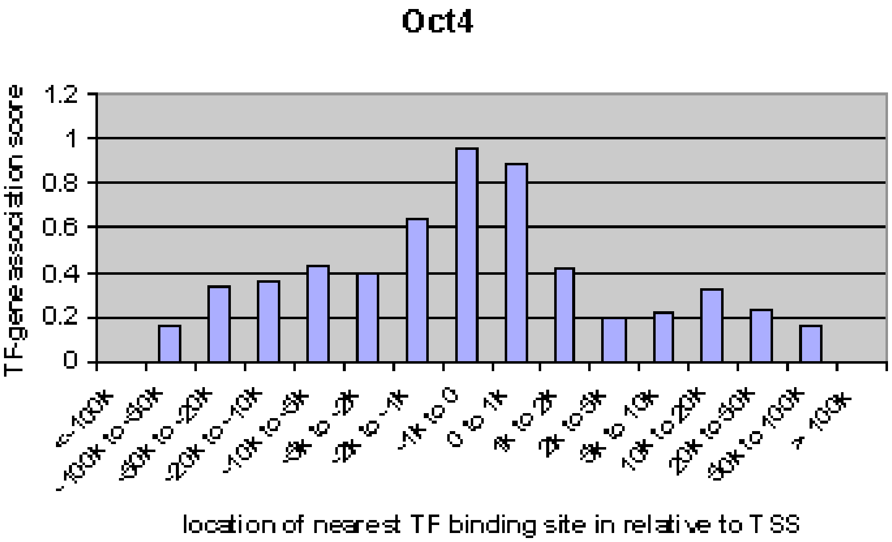
$$Score(g,tf) = \begin{cases} 0, & \text{if } Hist_{real}(k) < Hist_{rand}(k) \\ \dfrac{Hist_{real}(k) - Hist_{rand}(k)}{Hist_{real}(k)}, & \text{if } Hist_{real}(k) \geq Hist_{rand}(k) \end{cases}$$

where $Hist_{real}(k)$ represents the histogram calculated from the real data and $Hist_{rand}(k)$ represents the histogram calculated from the random model.

Chen et al., Cell 133, 1106-1117 (2008)

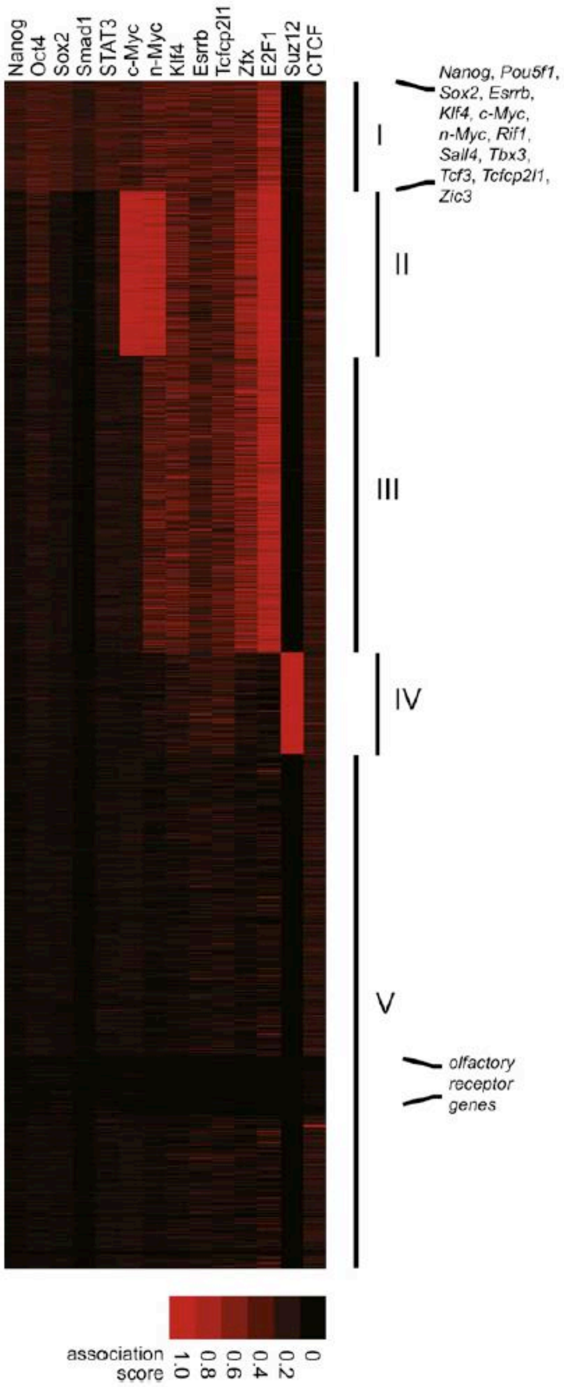
Association scores





Chen et al., Cell 133, 1106-1117 (2008)

Gene-TF association



Heatmap showing 5 classes of genes that are associated with a similar set of TFs obtained from k-means clustering based on the TF-gene association score.

Class I genes are enriched in binding sites for **Nanog, Oct4, Sox2, Smad1**, and **STAT3**.

Class II genes are bound heavily by **c-Myc** and **n-Myc**.

Class III genes show enrichment (more than 1-fold) in binding by **n-Myc, Klf4, Esrrb, Tcfcp2l1, Zfx,** and **E2f1**.

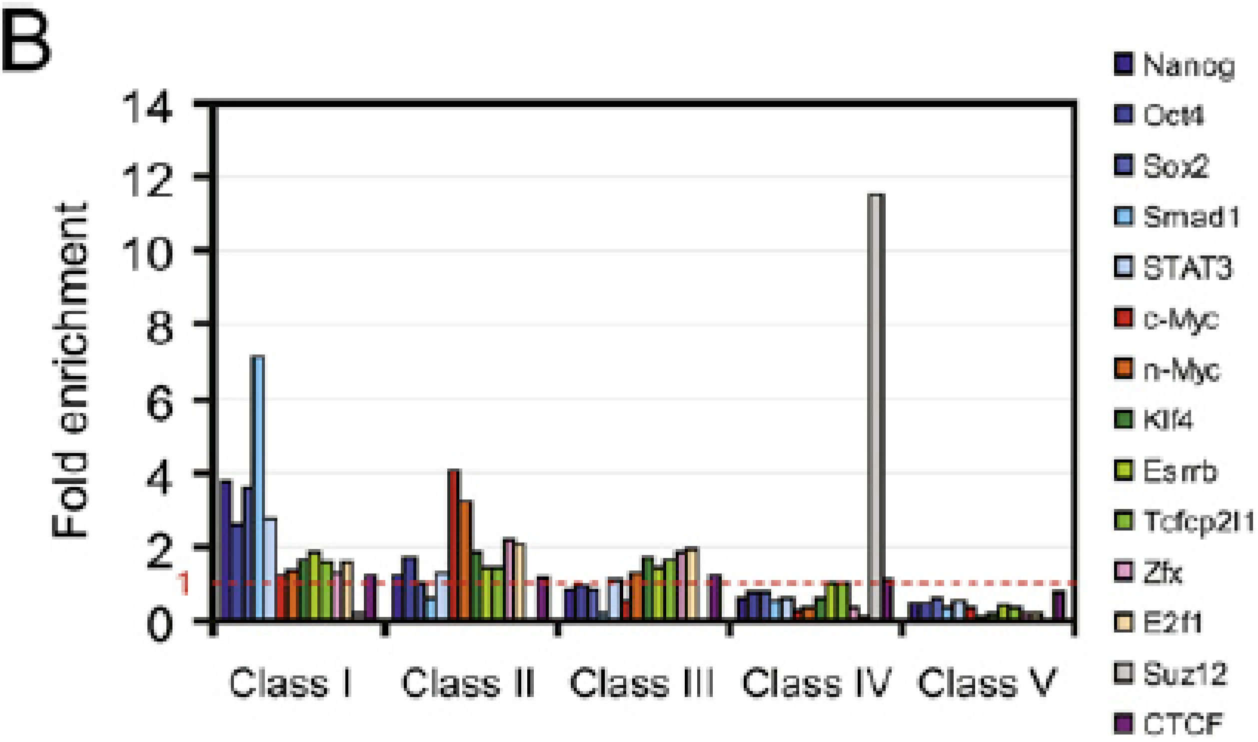
Class IV genes are highly enriched in **Suz12**-bound genes.

Class V genes are deficient in all TFs.

About half of all genes (48%) are deficient in TF binding by the 13 TFs (and thus belong to classes IV and V).

Chen et al., Cell 133, 1106-1117 (2008)

Enrichment of TFs in the 5 classes

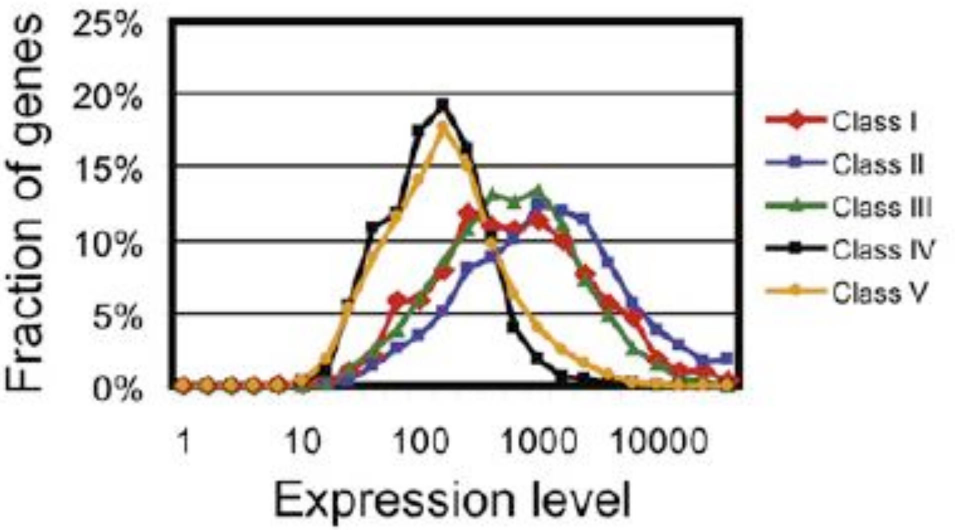


The y axis represents the ratio of the average TF-gene association score for the group to the average association score for all genes.

Chen et al., Cell 133, 1106-1117 (2008)

Expression levels in ES cells

Q: What are the implications of combinatorial patterns of TFBS for gene expression?



Shown are expression levels of the genes belonging to each class in a published microarray data set for undifferentiated ES cells.

The expression level is highest for genes in class II, followed by genes in class I and class III.

Genes in class IV and class V are either not expressed or are expressed at a very low level in ES cells.

Chen et al., Cell 133, 1106-1117 (2008)

Upregulated genes in ES cells

60% of the genes upregulated in ES cells are from class I and class II.

→ Gene clustering based on TF occupancies has the potential to predict ES-cell specific gene expression.

→ This suggests that the TF-binding patterns of these two groups are relevant in specifying ES-cell-specific expression.

In summary, combinatorial binding patterns of TFs have great predictive power for ES cell-specific expression.

Chen et al., Cell 133,
1106-1117 (2008)

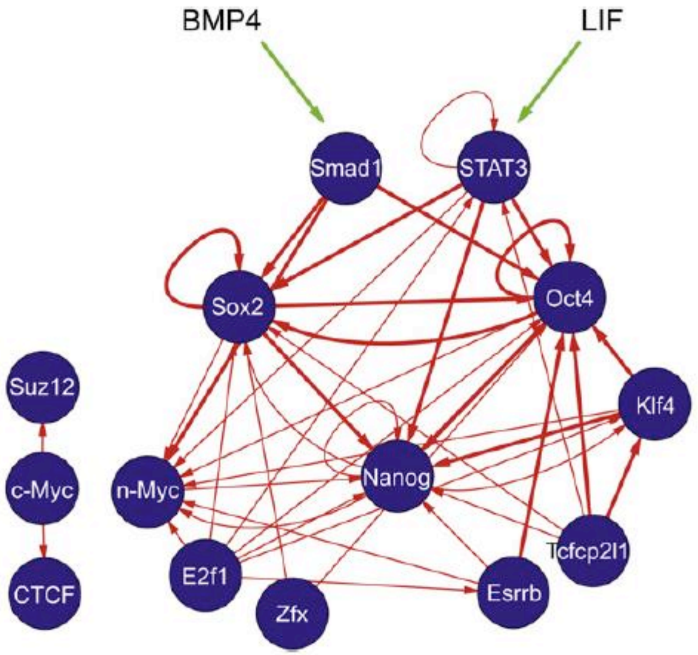
Identify Core Pluripotency Network

Next aim: construct regulatory network that specifies ES-cell-specific expression by using binding sites of transcriptional regulators under the undifferentiated state.

Define for each individual TF_i, the regulatory interactions of its target genes by intersecting the rank-ordered list of genes bound to TF_i and the rank-ordered list of differentially expressed genes.

Chen et al., Cell 133, 1106-1117 (2008)

Transcriptional Regulatory Network in ES cells



Nodes: ChIP-seq-assayed TFs.
Arrows point from the TF to the target gene.
2 sets of published experiments were used to define genes that are differentially expressed during differentiation.
Thick arrows : interactions inferred from binding data and both expression experiments
Thin arrows: interactions inferred from binding data and only one of the expression experiments.

Network of regulatory interactions inferred from ChIP-seq binding assays and from gene-expression changes during differentiation.

All regulatory interactions in this network involve higher-level expression in ES cells and lower-level expression during differentiation.

Consistent with previous studies, we identified regulatory feedback loops for Oct4, Sox2, and Nanog. Interesting is the high interconnectivity among 11 of the 13 TFs.

3 key signaling pathways are integrated to the Oct4, Sox2, and Nanog circuitries through Smad1 and STAT3.

Chen et al., Cell 133, 1106-1117 (2008)

Pluripotency Networks of Mouse vs. Human

the core of the regulatory network that maintains the pluripotent state is a set of **TFs**. Among these, **OCT4** seems to play a key role.

OCT4 co-occupies many regulatory sites together with SOX2 and NANOG.

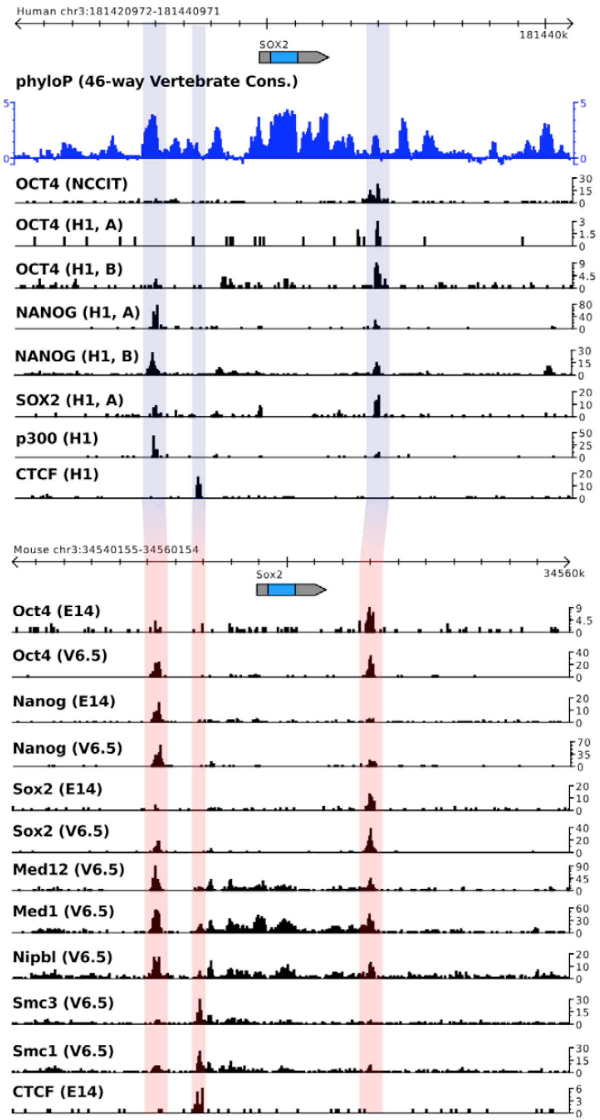
Many genes which are important for early embryogenesis have a conserved function in mouse and human.

BUT: only about 5% of binding events of the key pluripotency factors OCT4 and NANOG are conserved at orthologous genomic locations in mouse and human ES cells.

This is also true for other TFs. E.g. the liver TFs CEBP and HNF4 only showed 7% conserved binding events between human and mouse.

Göke et al., PLoS Comput Biol 7, e1002304 (2011)

Pluripotency Network in mouse and human ES cells



Overview of genome-wide CHiP-seq binding data in human H1 ES cells and various mouse embryonic stem cells and embryonal carcinoma cells.

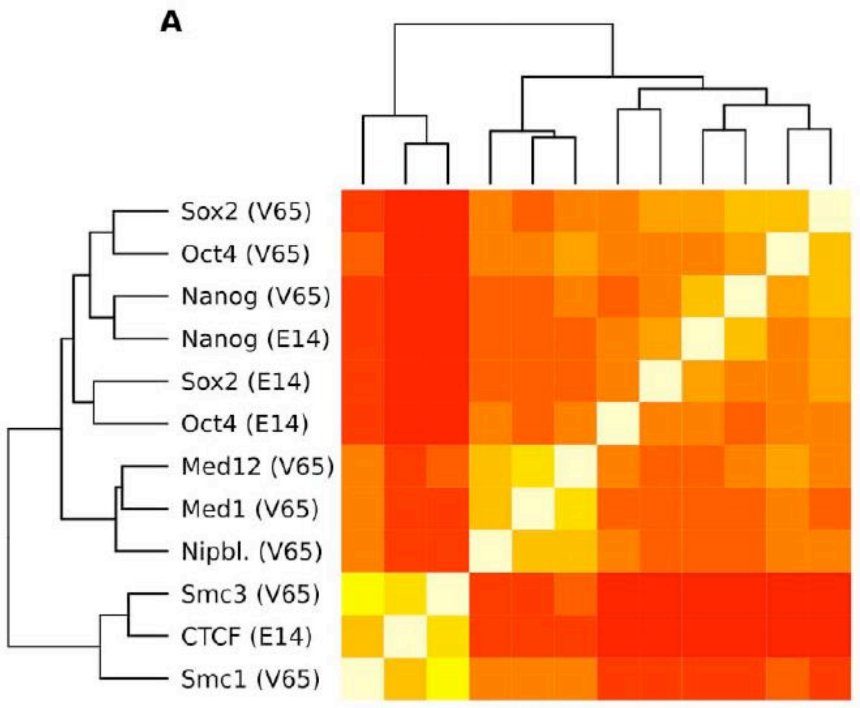
Note: human genes use CAPITAL letters (OCT4), mouse genes small letters (Oct4).

Shown is the locus of the SOX2 gene in the human genome (top), along with mapped reads for OCT4, SOX2, NANOG and p300. Individual experiments are shown separately.

The orthologous locus in the mouse genome is aligned at the bottom along with mapped reads from individual experiments.

Göke et al., PLoS Comput Biol 7, e1002304 (2011)

Co-localization identifies know protein interaction

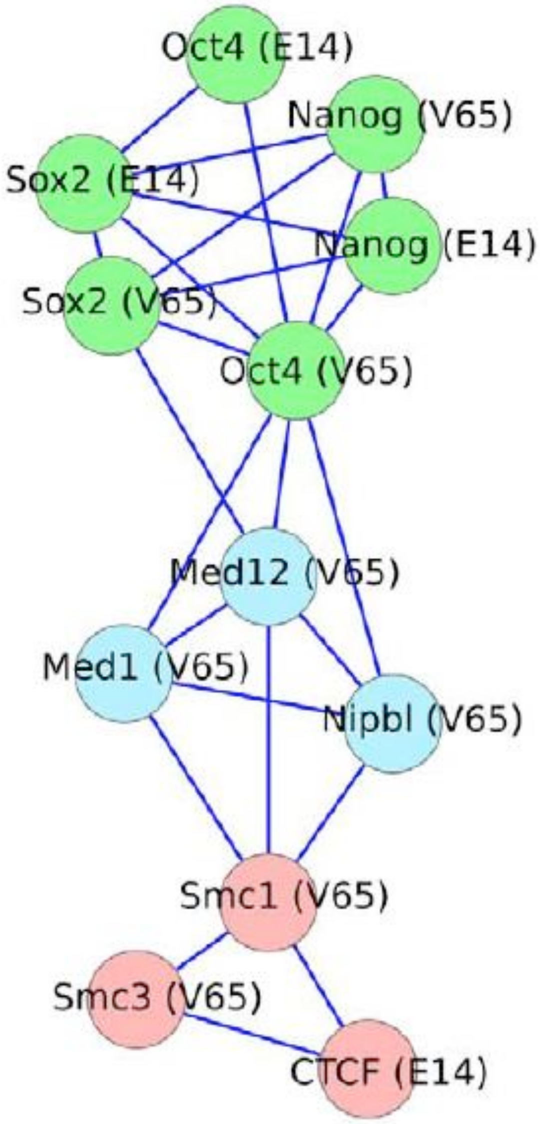


Clustering of genome-wide binding profiles from mES cells based on the number of shared binding events identifies 3 main classes:

- **Enhancer binding** (Oct4, Sox2, Nanog),
- **Insulator binding** (CTCF, Smc1, Smc3) and
- **Mediator associated binding** (Med1, Med12, Nipbl).

Göke et al., PLoS Comput Biol 7, e1002304 (2011)

Core Pluripotency Interaction Network



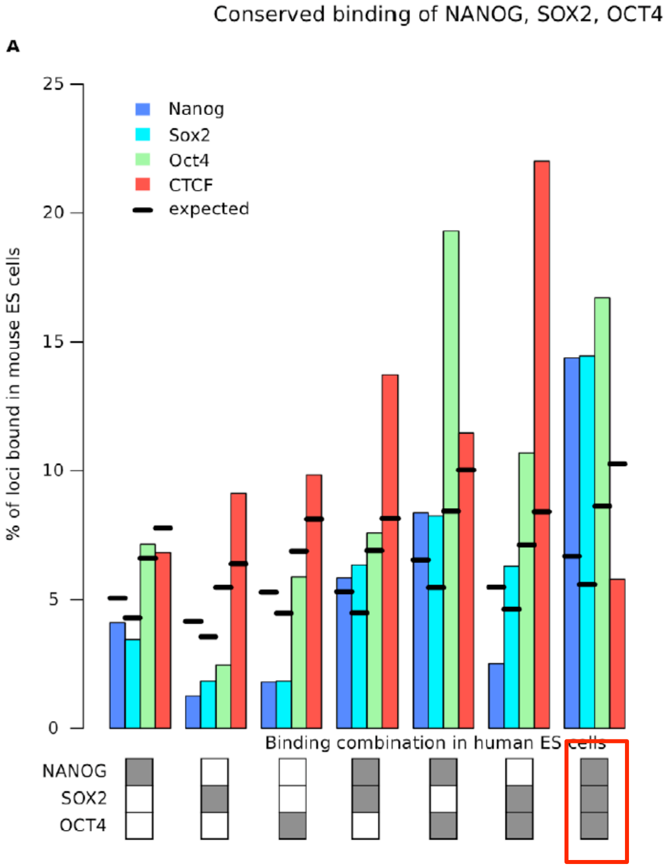
Protein network inferred on the similarity of genome-wide binding data. Edges represent the pairwise similarities with a Z-score above a threshold.

The same 3 distinct clusters are apparent:
- enhancer binding (green),
- insulator binding (red) and
- transcriptional co-activation (blue).

Interestingly, pairwise distances from genome-wide data on DNA-protein interactions reproduce known protein-protein interactions:
- CTCF interacts with Cohesin at insulator elements,
- Oct4, Sox2 and Nanog interact at enhancers, and
- Mediator plays a central role by integrating signals from distant regulatory elements and Cohesin.

Göke et al., PLoS Comput Biol 7, e1002304 (2011)

Combined binding of Oct4, Sox2 and Nanog



The combination of OCT4, SOX2 and NANOG influences conservation of binding events.

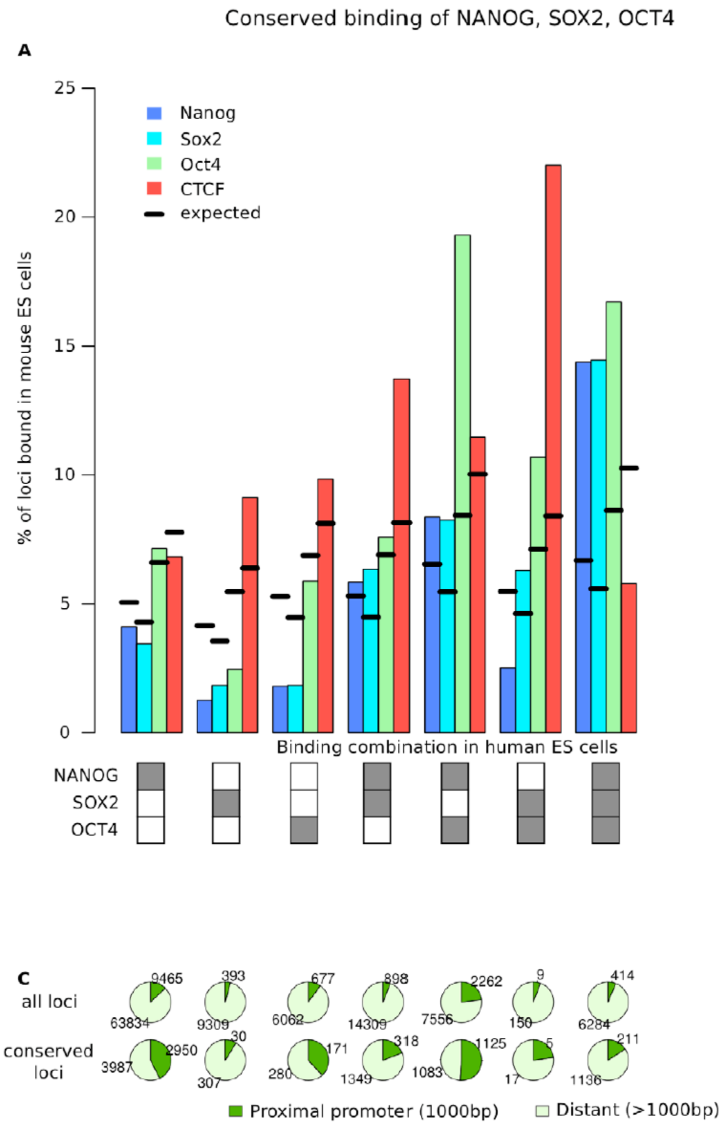
(A) Bars indicate the fraction of loci where binding of Nanog, Sox2, Oct4 or CTCF can be observed at the orthologous locus in mouse ES cells for all combinations of OCT4, SOX2 and NANOG in human ES cells as indicated by the boxes below.

Dark boxes indicate binding, white boxes indicate no binding (“AND” relation).

Combinatorial binding of OCT4, SOX2 and NANOG shows the largest fraction of conserved binding for Oct4, Sox2 and Nanog in mouse.

Göke et al., PLoS Comput Biol 7, e1002304 (2011)

Binding combinations



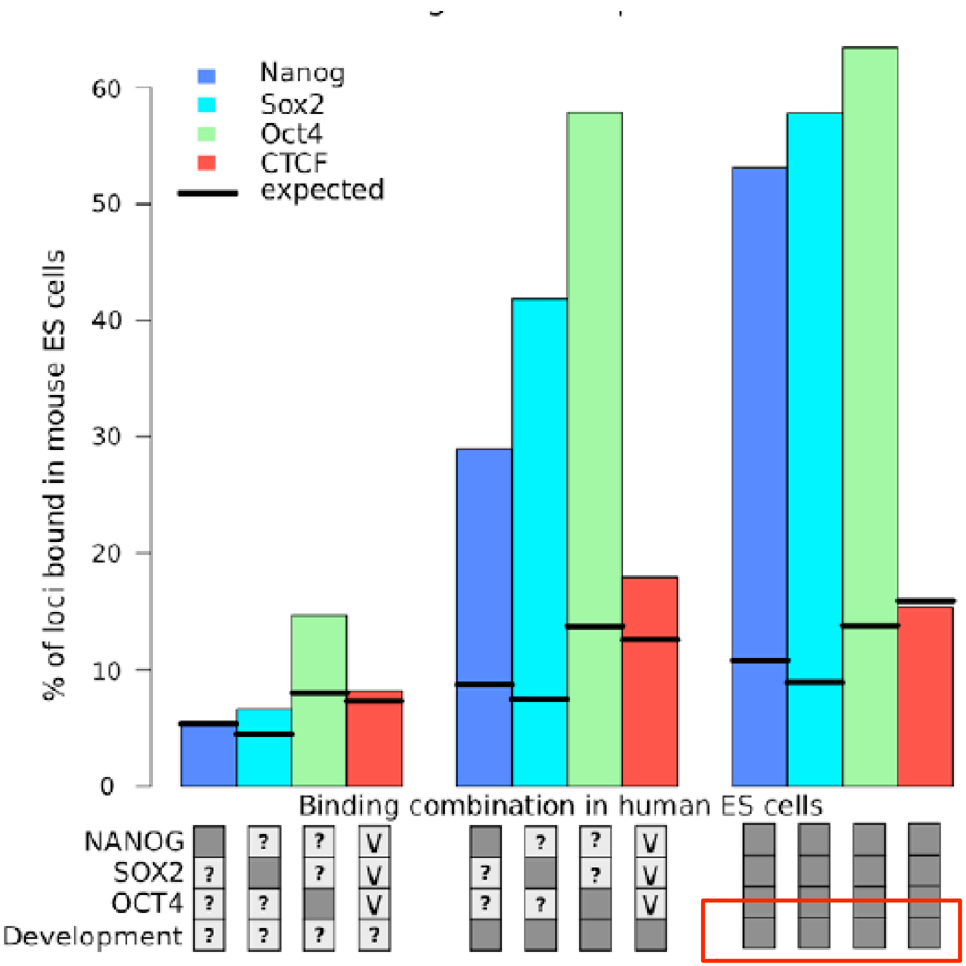
Göke et al., PLoS Comput Biol 7, e1002304 (2011)

(C) Fraction of proximal and distant binding sites for conserved and non-conserved binding events.

Conserved binding events are mostly found in distant regulatory elements.

But conserved binding events are more frequently in the proximal promoter than non-conserved binding events (share of dark green increases).

Increased Binding conservation in ES cells at developmental enhancers



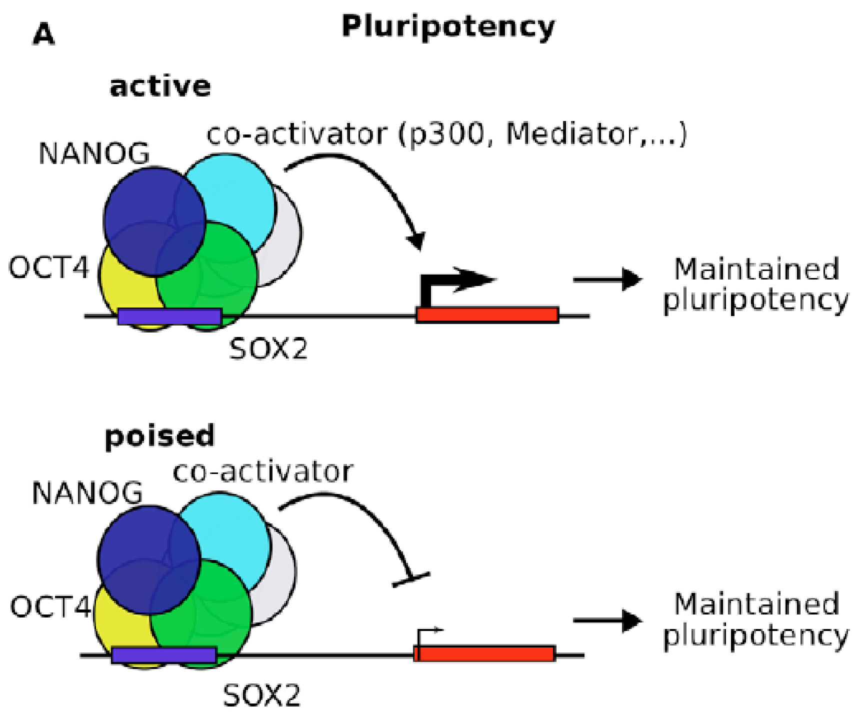
Göke et al., PLoS Comput Biol 7, e1002304 (2011)

Fraction of loci where binding of Nanog, Sox2, Oct4 and CTCF can be observed at the orthologous locus in mouse ESC.

Combinations of OCT4, SOX2 and NANOG in human ES cells are discriminated by developmental activity as indicated by the boxes below.
Dark boxes : "AND" relation,
light grey boxes with "v" : "OR" relation, '
'?" : no restriction.

Combinatorial binding events at developmentally active enhancers show the highest levels of binding conservation between mouse and human ES cells.

Model for "gene regulatory hotspots"

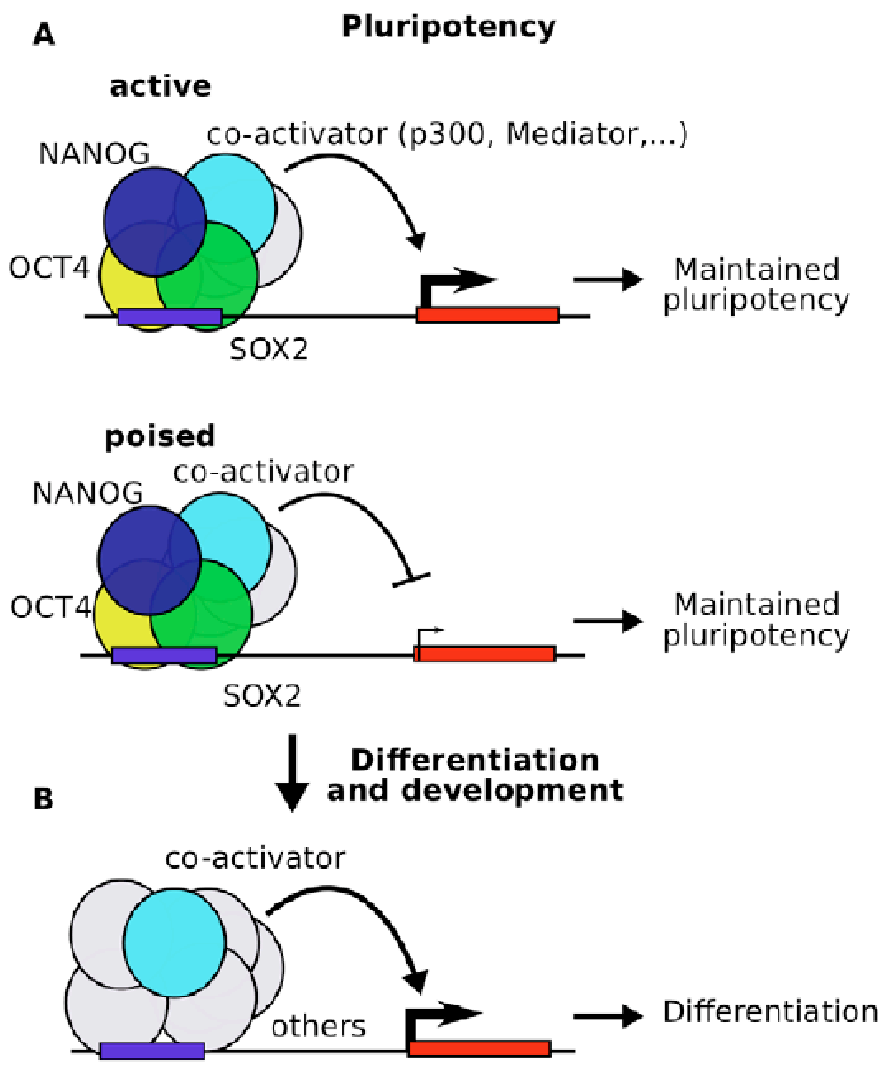


Enhancers are bound by OCT4, SOX2 and NANOG together with p300 in ES cells.

These enhancers maintain pluripotency by activating gene expression in ES cells (top) or poisoning expression for activation after differentiation (bottom)

Göke et al., PLoS Comput Biol 7, e1002304 (2011)

Model for "gene regulatory hotspots"



(B) After differentiation of the cell, the same enhancer DNA segments are bound by p300 in developmental tissues together with other TFs so that the target gene is expressed.

Enhancers which recruit multiple TFs in different stages of development can be considered gene regulatory hotspots which are crucial to connect the regulatory networks of pluripotency and development.

These enhancers potentially show higher sequence conservation compared to individual, isolated binding events which are active in single cell types.

Göke et al., PLoS Comput Biol 7, e1002304 (2011)

Conserved binding events

720 loci are conserved in human and mouse ES cells that are bound combinatorially by OCT4/Oct4, SOX2/Sox2, and NANOG/Nanog.

These loci are associated with 608 genes nearby.

Testing genes for enrichment of GO terms yields:

Pattern specification process ($p = 4.7\ e^{-13}$)
Regionalization ($p = 2.5\ e^{-12}$)
Development induction ($p = 8.4\ e^{-8}$)

as significantly enriched terms. All are related to **early development**.

Göke et al., PLoS Comput Biol 7, e1002304 (2011)

Summary

26% of combinatorially bound loci which are conserved between mouse and human ES cells are **developmental enhancers** in the mouse.
This suggests that many enhancers bound by OCT4, SOX2 and NANOG are also developmental enhancers in human.

The very same regulatory elements bound by key pluripotency factors in ES cells frequently act as enhancers during early development.

This finding provides an unknown link between the gene regulatory networks of ES cells and early development at the level of transcriptional regulation.

The finding that binding at these developmental enhancers are highly conserved in mouse and human ES cells suggests that these elements are crucial for the maintenance of the pluripotent state.

Göke et al., PLoS Comput Biol 7, e1002304 (2011)

Summary (II)

The fast evolutionary rewiring of regulatory networks mainly affects individual binding events.

In contrast to these events, there is a group of conserved enhancers in the genome which recruit multiple interacting factors and are active in multiple tissues of the developing embryo.

Many of these "gene regulatory hotspots" are under strong evolutionary constraints and seem to play a major role by linking the regulatory networks of cellular differentiation during early mammalian development.

Göke et al., PLoS Comput Biol 7, e1002304 (2011)