1. Introduction
   - Gene regulation
   - Genomics and genome analyses

2. Gene regulation tools and methods
   - Regulatory sequences and motif discovery
   - Transcription factor binding sites

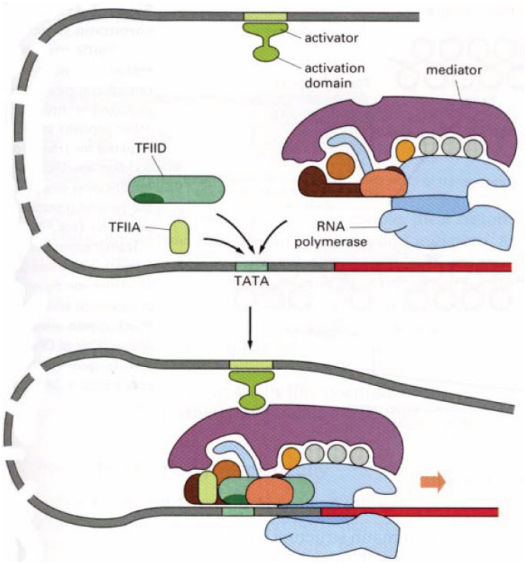
3. Technologies
   - Microarrays
   - Deep sequencing and applications
   - Single-cell RNAseq and spatial transcriptomics

4. Clustering
   - Unsupervized clustering (HCA,K-means, PCA, SOM)
   - Supervized clustering (classification)

5. Gene ontology, Pathways, Enrichment analysis
   - Databases and tools
   - Gene set enrichment analysis

6. Biomolecular networks
   - Small world networks
   - Topology and parameter
   - Network motifs

Regulation of transcription



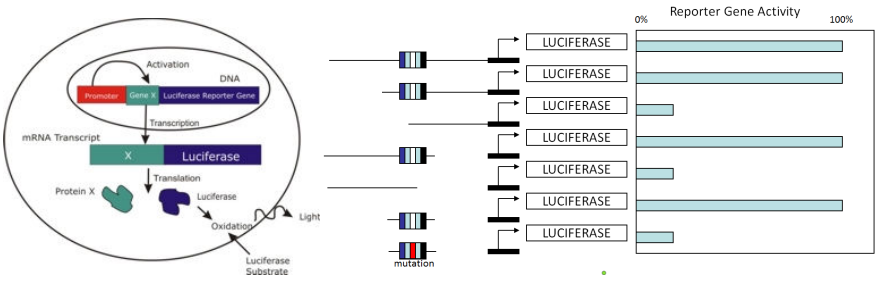
Identification of transcription factor binding sites

- Experimental methods
- Computational methods

Experimental methods

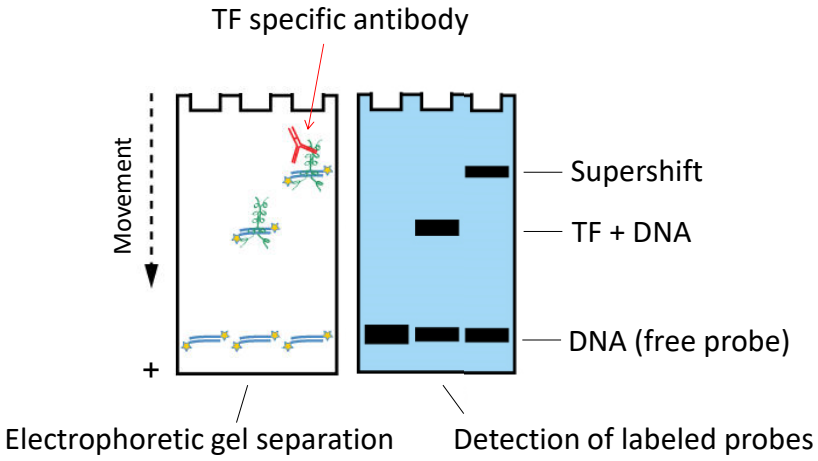
- Reporter gene assays (luciferase)
- Electro mobility shift assays (EMSA)
- DNase I and Exonulease Footprinting
- SELEX
- Protein binding microarrays (PBMs)
- Chromatin immuno precipitation (ChIP)

Luciferase reporter assays

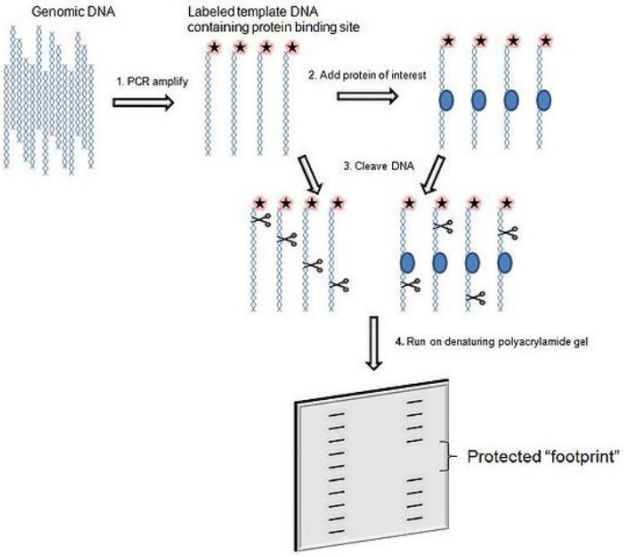


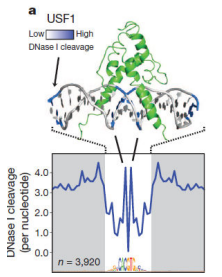
- Identify functional regulatory region within a sequence and delineate specific TFBS through mutagenesis
- Evidence that TF binding has an effect on transcription (not only binding to DNA)

Electromobility/Gel Shift Assays



DNase I and Exonuclease footprinting

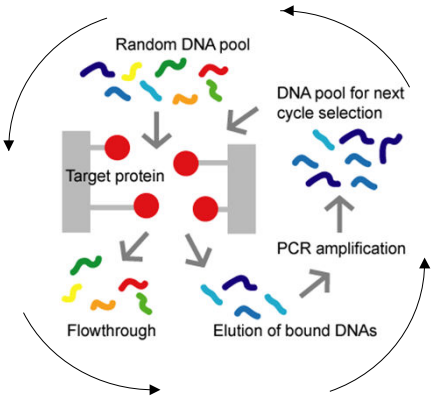




Neph et al., Nature, 2012

SELEX

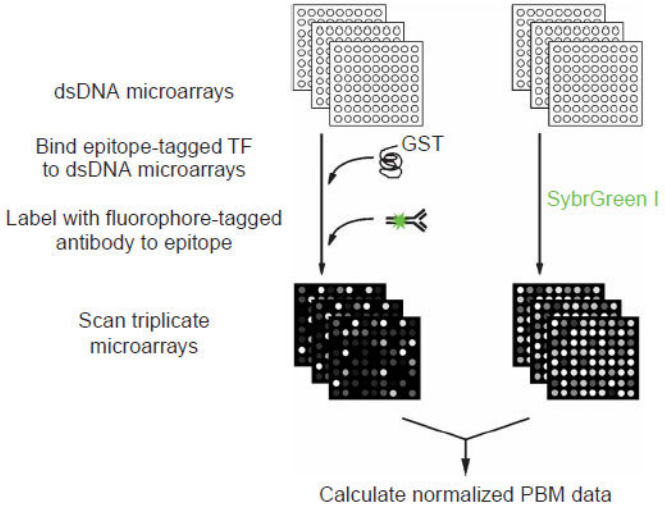
Systematic evolution of ligands by exponential enrichment



several cycles

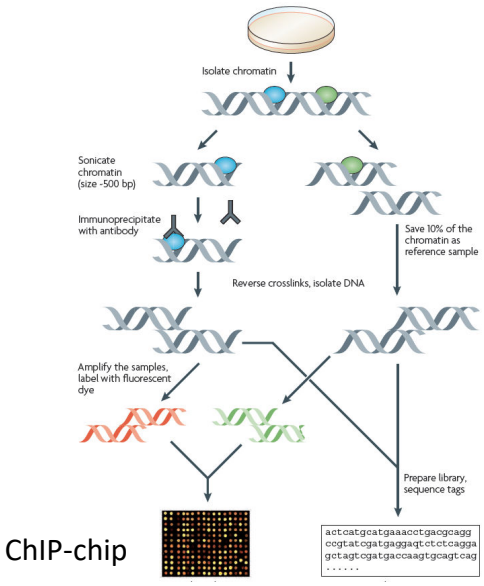
Most position weight matrices (PWMs) in the databases are derived by SELEX

Protein binding microarrays (PBMs)



Mukherjee et al., Nature Genetics, 2004

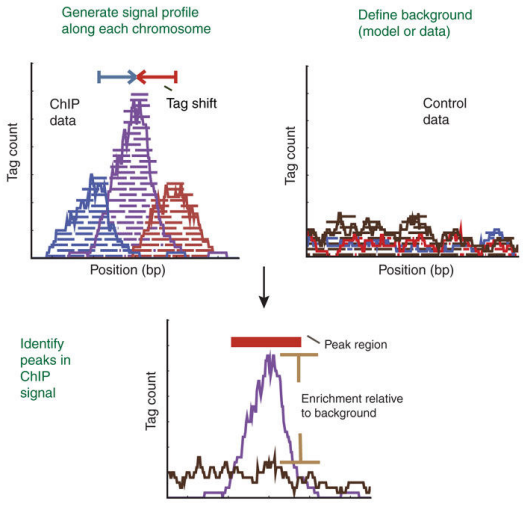
ChIP procedure



ChIP-chip ChIP-seq

Farnham, Nature Rev Genetics, 2009

ChIP-seq (Peak calling)

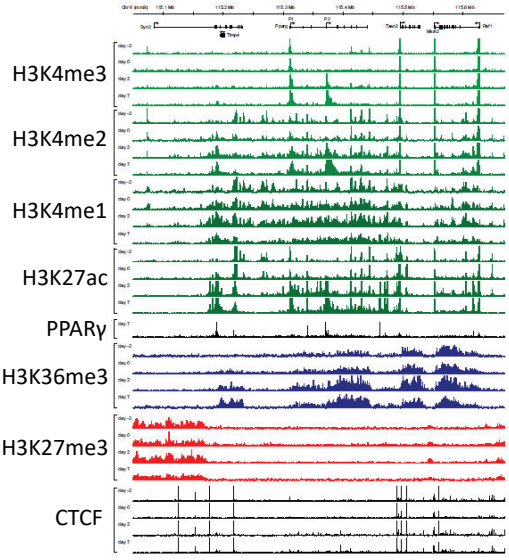


Pepke, Nature Methods, 2009

Tools:

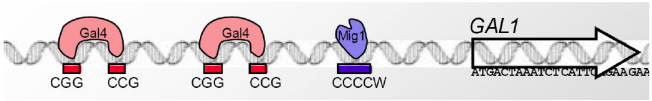
- CisGenome
- ERANGE
- FindPeaks
- F-Seq
- GLITR
- MACS
- PeakSeq
- QuEST
- SICER
- SiSSRs
- Spp
- Useq

Chromatin state and TF localization



Mikkelsen et al., Cell, 2010

Computational methods



Challenges:

- Sequences are short (e.g. 6-8bp)
- Sometimes degenerated (not always the same)
- Different distances from transcription start site
- TF binding might have no effect on gene expression
- Combination of TF and TF modules
- Many false positive predictions

Computational methods

- Matrix based methods (knowledge about TF)
  Position weight matrix (PWM), HMM
- Motif discovery
  Word counting, EM
- MicroRNA target prediction

Experimental verified binding sites

Gene	Organism	5'-3' Sequence	Ref
CYP4A6/P450 IV	rabbit	AACT AGGGCA A AGTTGA	[1]
CYP4A1/P450 IV	rat	AACT AGGGTA A AGTTCA	[2]
L-fatty acid binding protein	rat	ATAT AGGCCA T AGGTCA*	[3]
3-hydroxy-3-methyl-glutaryl-CoA-synthase	rat	AACT GGGCCA A AGGTCT*	[4]
Enoyl-CoA-hydratase	rat	ATGT AGGTAA T AGTTCA*	[1]
Malic enzyme	rat	TTCT GGGTCA A AGTTGA	[5]
Phosphoenolpyruvate carboxikinase	rat	AACT GGGATA A AGGTCT	[6]
Phosphoenolpyruvate carboxikinase)	rat	CCCA CGGCCA A AGGTCA*	[6]
▪ ▪ ▪ ▪			
Uncoupling protein 1	mouse	AGTG TGGTCA A GGGTGA*	[12]
Apolopoprotein C-III	human	GCGC TGGGCA A AGGTCA*	[1]
Acyl-CoA oxidase	human	TAGA AGGTCA G CTGTCA	[13]
Lipoprotein lipase	human	GTCT GCCCTT T CCCCCT*	[14]
Muscle type carnitine palmitoyltransferase I	human	CCTT TTCCCT A CATTTG	[15]
Consensus		AWCT AGGNCA A AGGTCA	[16]

Position frequency matrix

- Position frequency matrix (PFM, PWM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	10	8	4	3	11	0	1	1	2	19	15	17	2	0	0	0	16
C	3	4	11	5	1	1	2	6	15	0	1	4	1	1	2	17	2
G	3	2	4	2	7	20	19	6	1	1	2	1	17	15	1	4	1
T	6	8	3	12	3	1	0	7	4	2	4	0	2	6	19	1	3

- Position weight matrix (PWM),
  Position specific scoring matrix (PSSM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	0.86	0.54	-0.46	-0.87	1.00	-1.32	-2.46	-2.32	-1.46	1.79	1.45	1.63	-1.46	-1.32	-1.32	-1.32	1.54
C	-0.87	-0.46	1.00	-0.14	-2.46	-2.46	-1.46	0.26	1.45	-1.32	-2.46	-0.46	-2.46	-2.46	-1.46	1.63	-1.46
G	-0.87	-1.46	-0.46	-1.46	0.35	1.86	1.79	0.26	-2.46	-2.46	-1.46	-2.46	1.63	1.45	-2.46	-0.46	-2.46
T	0.13	0.54	-0.87	1.13	-0.87	-2.46	-1.32	0.49	-0.46	-1.46	-0.46	-1.32	-1.46	0.13	1.79	-2.46	-0.87

Position weight matrix (PWM)

Probability of base b at position i

$$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in \{A,C;G,T\}} s(b')}$$

N ... number of sites
s(b) ... pseudo counts
$F_{b,i}$ ... frequency of base b in position i

PWM

$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)}$$

p(b) ... background probability of base b

Evaluation of sequences

$$S = \sum_{i=1}^{w} W_{b,i}$$

w … width of PWM
b … nucleotide in position i
S … PWM score of a sequence

	1	2	3	4	5	6
A	1.00	-1.32	-2.46	-2.32	-1.46	1.79
C	-2.46	-2.46	-1.46	0.26	1.45	-1.32
G	0.35	1.86	1.79	0.26	-2.46	-2.46
T	-0.87	-2.46	-1.32	0.49	-0.46	-1.46

…ACGTAGGTCATAGAGTA.. S=1+1.86+1.79+0.49+1.45+1.79=8.38

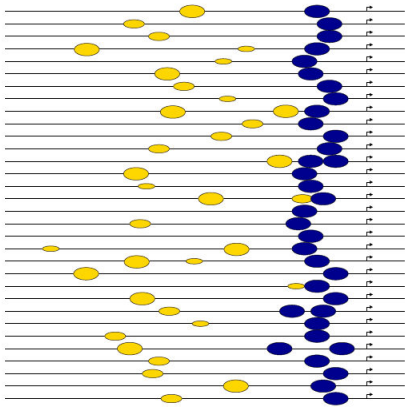
…ACGTAGGTCATAGAGTA.. S=-0.87-2.46-2.46+0.49-1.46-2.46=-9.22

Threshold for similarities

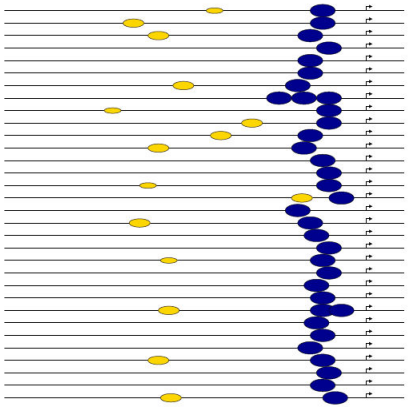
- Find optimized similarity score to minimize false predictions
- Difficult to determine because of lack of experimental data
- One approach could be to define the optimized threshold of a weight matrix as the matrix similarity that allows e.g 3 matches in 10000bp of none regulatory test sequences
- Using background sequences with similar nucleotide composition and using upper percentile to set cutoff (e.g. upper 5%).

Using a set of background sequences

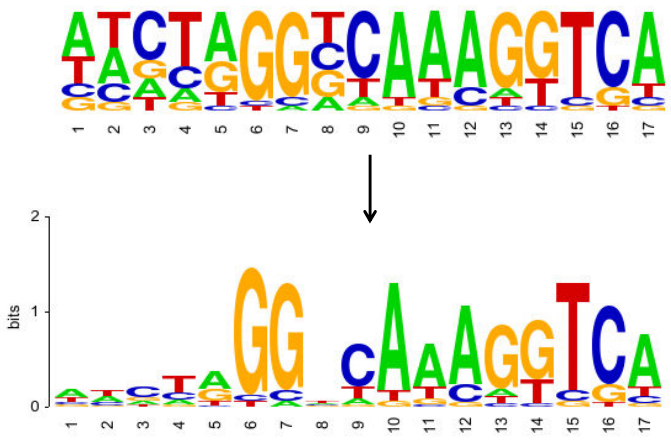
Foreground sequences



Background sequences



From Frequency to Sequence Logo



Information content in position i

$$D_i = 2 + \sum_b p(b,i)\log_2 p(b,i) - e(n)$$

e(n) … correction factor if only few samples n
D_i … information content at position i
b … base A,C,G, or, T

All bases with equal probabilities at position i

$D_i = 2+4*0.25*\log_2 0.25 = 0$ bits

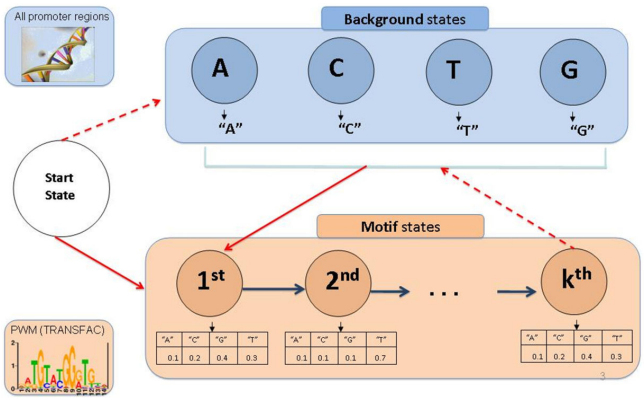
Only one base is present at position i

$D_i = 2+1*\log_2 1+3*0.001*\log_2 0.001 = 1.97$ bits

↑

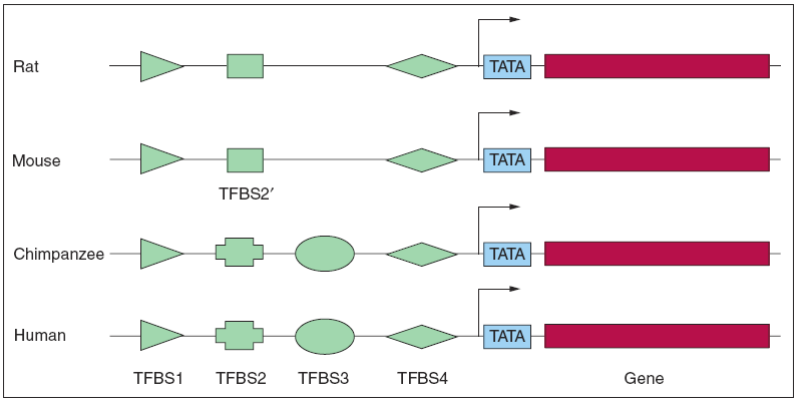
from pseudocounts ($\log_2 0$ is not defined!!)

Profile hidden markov models (HMM)



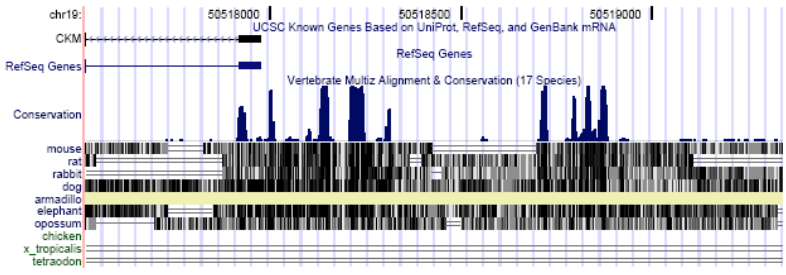
Levkovitz et al. PLoS One. 2010

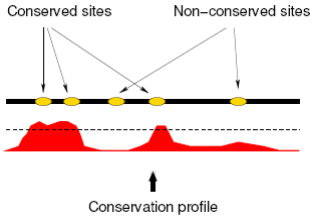
Phylogenetic footprinting



- Functional regulatory sites are conserved between species
- E.g. using multiz alignment of UCSC genome browser

Phylogenetic footprinting



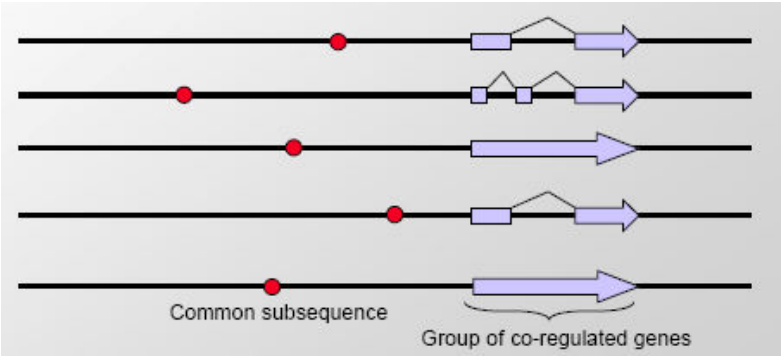


Motif discovery

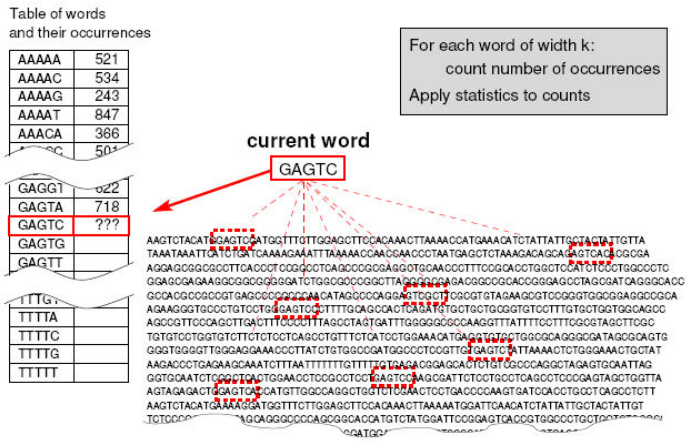
- Start with just sequences
- Identify strongly enriched motifs de novo
- Algorithmically on of the most challenging analysis tasks
- Use it when you suspect important unknown motifs in your data

- Word counting and enrichment
- Statistical algorithms
  - Expectation-maximum algorithm (EM)
  - Gibbs Sampling

Find common subsequence



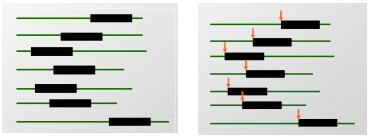
Motif discovery by word counting



Compared with background sequences (Fischer exact test => p-value)

Motif discovery

- Problem: Don't know what the motif looks like or where the starting positions are



- Expectation maximum (EM) (e.g. used by MEME, ChIP-MEME)
- Gibbs sampling (e.g. used by AlignACE)

- EM is a family of algorithms for learning probabilistic models in problems that involve *hidden state*
- In our problem, the hidden state is where the motif starts in each training sequence

Basic EM-approach

given: length parameter *W,* training set of sequences
set initial values for *p*
do
re-estimate *Z* from *p* (E –step)
re-estimate *p* from *Z* (M-step)
until change in $p < \varepsilon$
return: *p, Z*

A motif is represented by a matrix of probabilities: P_{ck}

X_i = G C **T G T** A G

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

$p =$

The element Z_{ij} of the matrix Z represents the probability that the motif starts in position *j* in sequence *i.*

$Z =$

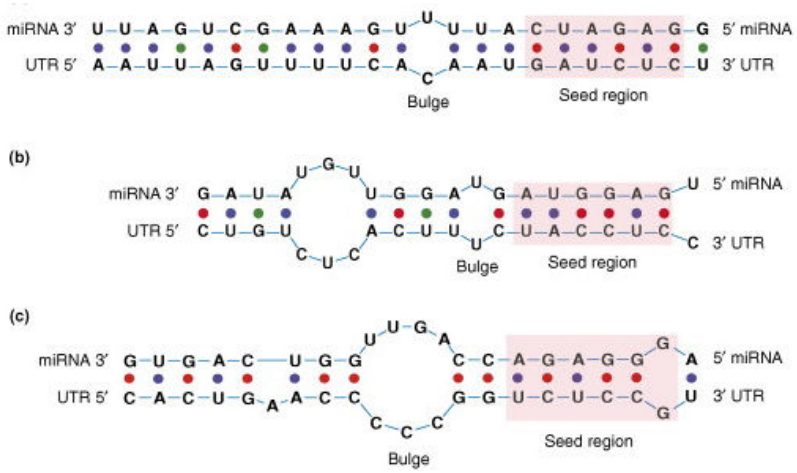
	1	2	3	4
seq1	0.1	0.1	0.2	0.6
seq2	0.4	0.2	0.1	0.3
seq3	0.3	0.1	0.5	0.1
seq4	0.1	0.5	0.1	0.3

Resources

- JASPAR (free)
- TRANSFAC (BIOBASE), Match
- Genomatix (commercial), MatInspector
- UniProbe
- DeepBind
- WebLogo
- SeqLogo
- ORegAnno, PAZAR
- Regulatory Sequence Analysis Tools (RSAT)
- MEME suite, FIMO
- ConTra
- LASAGNA, MAPPER2
- TOUCAN

microRNA target prediction

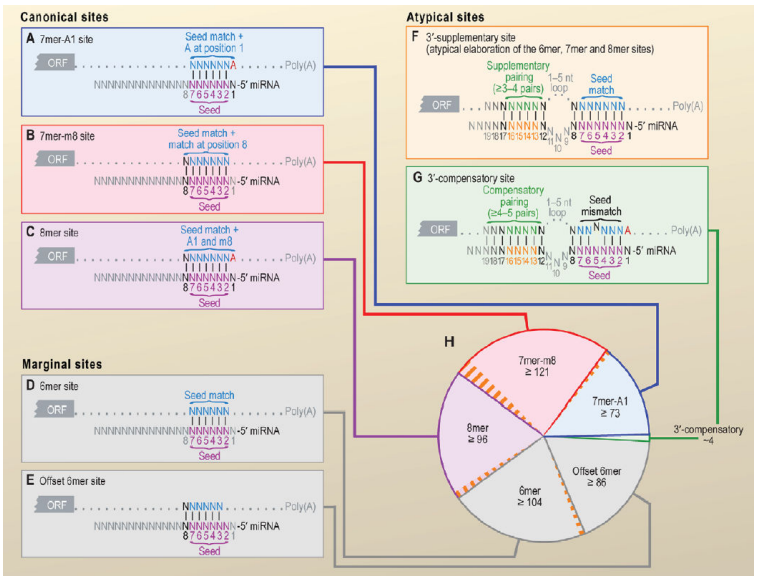
microRNA/mRNA pairing



Principles of microRNA target prediction

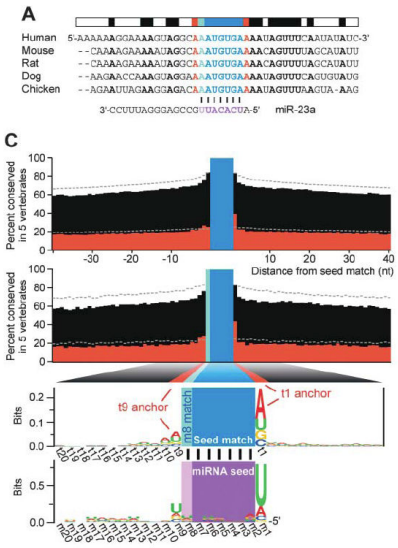
1. Sequence complementarity
2. Conservation
3. Thermodynamics
4. Site accessibility
5. UTR Context
6. Correlation of expression profiles

Sequence complementarity (TargetScan)



Bartel, Cell,2009

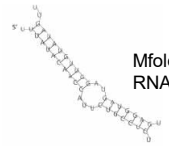
Conservation



Lewis BP et al., Cell, 2003

Thermodynamics

1. Minimum free energy



⇨ e

Mfold (Zuker et al.)
RNAfold (Hofacker et al.)

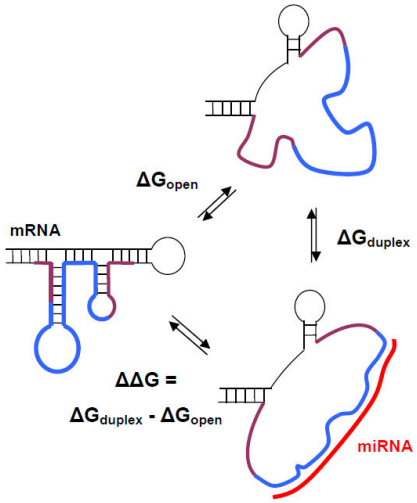
```
          mfe: -25.3 kcal/mol
           p-value: 0.010068

Target  5' A          UC         A 3'
             CACAG  UUG  UCUGCAGGG
             GUGUU  AGC  AGAUGUCCC
miRNA   3'       UA   CA          5'
```

2. Account for different sequence length

3. Extreme value distribution of MFE

Rehmsmeier M et al. RNA (2004)

Site accessibility

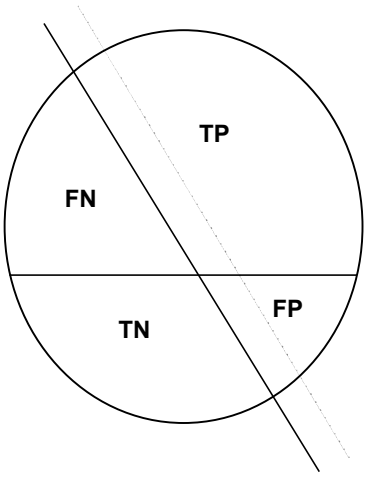


Leitner A, 2009

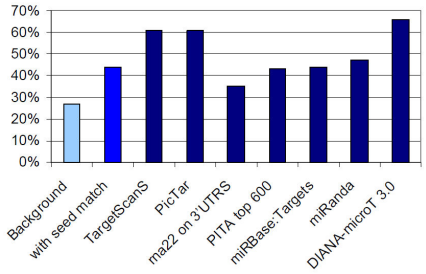
Validation

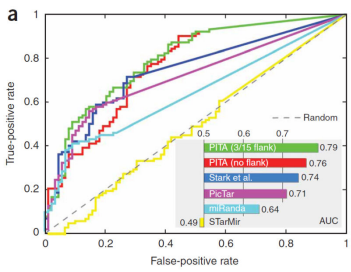
Sensitivity = TP / (TP+FN)

Specificity = TN / (FP+TN)



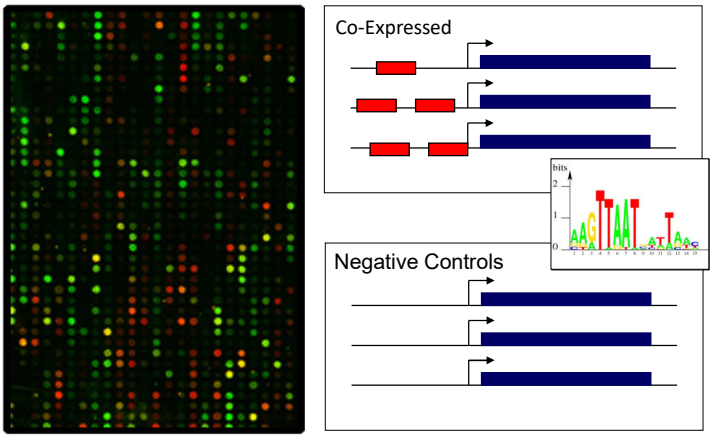
Reduced protein levels



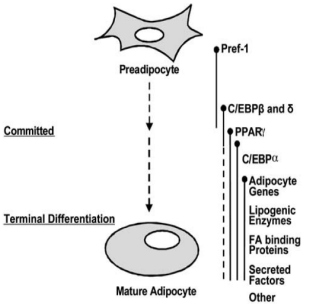


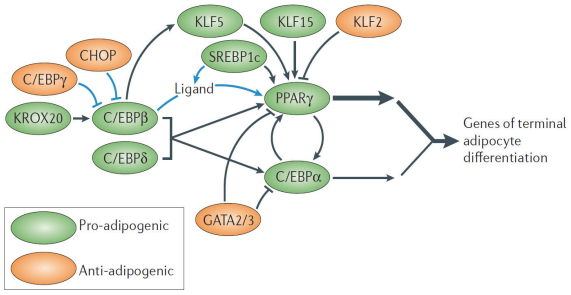
Case study

Deciphering Regulation of Co-Expressed Genes



Adipocyte differentiation

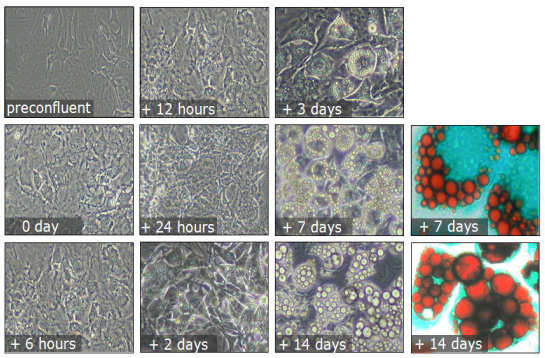




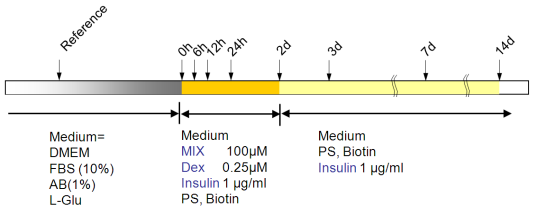
Gregoire et al., Physiol Rev, 1998

Rosen, MacDougald , Nat Rev Mol Cell Biol, 2006

3T3-L1 adipocyte differentiation timeseries

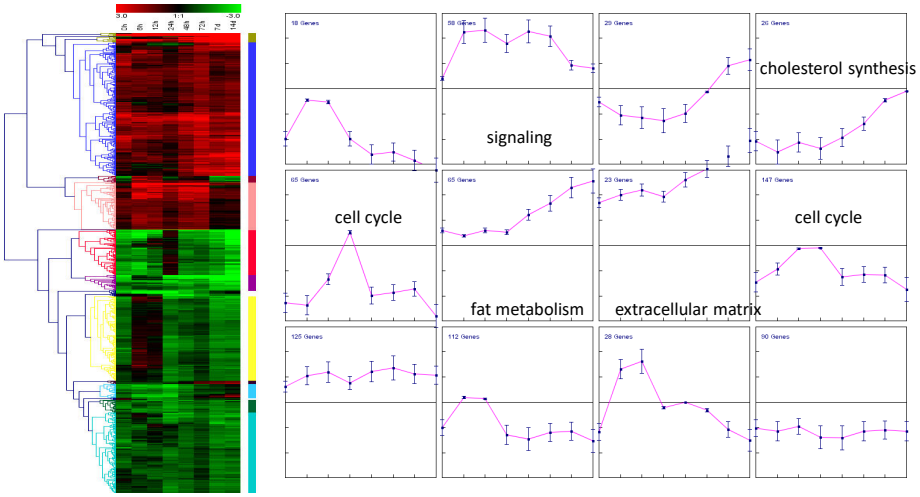


Microarray analysis



Hackl, Burkard et al. , Genome Biol, 2005

Correspondence between co-expression and function



Hierarchichal clustering and k-means clustering of 780 selected ESTs during fat cell differentiation

Correspondence between co-expression and co-regulation

- Functional transcription factor binding site: SREBP
- Binding sites for key regulators (PPARg, C/EBP) was not significantly over represented

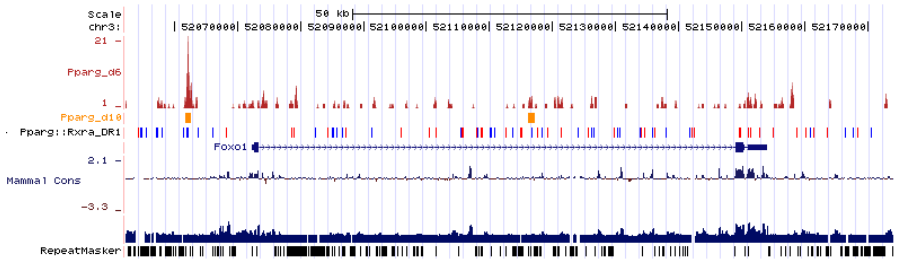
- In a follow up study Pparg binding sites were enriched

Matrix ID	Sequence logo	Transcription factor	Sim	# targets (cluster F)	# genes (cluster F)	# targets (RefSeq)	# genes (RefSeq)	p-value	FDR
M00183		c-Myb	0.99	165	206	15707	21408	1.5E-02	1.8E-01
M00191		ER (estrogen receptor)	0.94	140	206	12979	21408	1.7E-02	1.9E-01
M00474		FOXO1 (fork head box O1)	0.86	193	206	18983	21408	1.1E-02	1.5E-01
MG0001		PPARgamma/RXRalpha	0.84	123	206	10675	21408	2.7E-03	7.1E-02

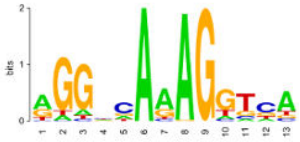
Hakim-Weber et al. BMC Research Notes, 2011

Pparg binding sites

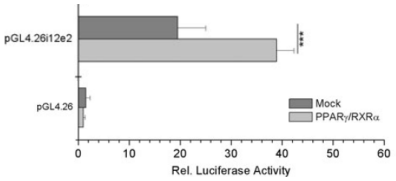
ChIPseq, ChIP-chip, and genomic organization (UCSC browser)



Pparg::Rxra DR1 motif search



Luciferase assays (Apmap promoter)



Bogner-Strauss et al. , Cell Mol Life Sci, 2010