How to setup a motif discovery analysis

A quick survey of studies that successfully applied motif discovery to find and validate new binding sites

Maximilian Haeussler
Developpement du systeme nerveux des chordees
CNRS Gif-sur-Yvette

Motif Discovery in cis-regulation

Input 1:
A set of regulatory regions which we assume to contain a common “word” (6-16bp) as the regions are supposed to be bound by the same transcription factor

Input 2:
Background sequences, not bound by transcription factor

Motif Discovery Program

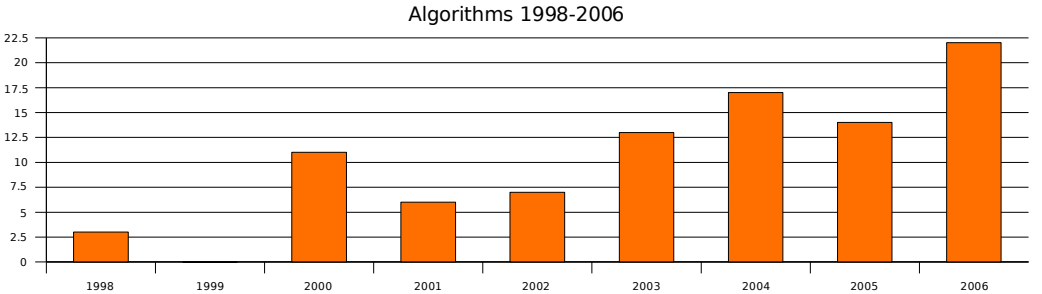
Output:
A ranked list of overrepresented motifs, ranked by some score

~100 Motif Discovery Algorithms

MEME	1994	RISA	2003	MOST	2005
MACAW	1994	DMotifs	2003	EBMF	2005
CoResearch	1996	Kamvysselis	2003	POCO	2005
R'MES	1997	Mermaid	2003	A-GLAM	2005
Oligo-Analys	1998	STARS	2003	Gemoda	2005
Teiresias	1998	SDDA	2003	SVD	2005
Yebis	1998	TreeGibbs	2003	BDTree	2005
Consensus	1999	Superposition	2003	PhyloGibbs	2005
Dyad-Analys	2000	CUBIC	2003	DME	2005
AlignACE	2000	ProfileBranch	2003	MEDUSA	2005
SP-Star	2000	PatternBranc	2003	Mamf	2006
Ann-Spec	2000	GMS-MP	2004	THEME	2006
Verbumculus	2000	Mermaid	2004	GAME	2006
by Anderson/	2000	Wu2004	2004	COMODE	2006
YMF	2000	PRUNER	2004	REFINEMENT	2006
ELPH	2000	GRAM	2004	RevJump	2006
Winnower	2000	GLAM	2004	Gertz et al	2006
MobyDick	2000	BioOptimizer	2004	PRISM	2006
SMILE	2000	Uniform Proje	2004	wordspy	2006
Bioprospecto	2001	DWE	2004	BEAM	2006
Co-Bind	2001	Combine	2004	EMD	2006
Tsukuba BB	2001	cWinnower	2004	SOMBRERO	2006
ITB	2001	MoDEL	2004	BoCaTfbs	2006
Weeder	2001	QuickScore	2004	GibbsILR	2006
Mitra	2002	EC	2004	MotifCut	2006
Spexs	2002	BiPad	2004	GibbsST	2006
Multiprofiler	2002	PhyME	2004	ALSE	2006
Projection	2002	Emnem	2004	Gertz et al	2006
MDScan	2002	COOP	2005	PRIORITY	2006
Li2002	2002	NestedMICA	2005	Reddy et al.	2006
ScanSeq	2002	EOMM	2005	MotifSeeker	2006
Mitra-PSSM	2003				

- Published in peer-reviewed journals
- Each one is proven to better than a couple of the others
- Most comprehensive benchmark by Tompa et al in 2005 with 12 participants.

from MUMDAB
(Max's Useless Motif Discovery Algorithm Database),
www.stud.uni-potsdam.de/~haussler/master/



Many open questions:

- Motif search method: Brute force? Statistical sampling? Dynamic Programming? Self-organizing Maps? ...
- Background model: HMM versus nucleotide distribution? Whole genome as background?
- Comparative Genomics: Which genomes and how align them (local/global) ?

Selection of studies:

- Gene regulation in metazoans (animals and plants)
- Apply a motif discovery algorithm to discover a new binding site specific to a set of genes
- One prediction has to be tested by wet-lab assays (Mutation + Reporter-gene, gel shifts, etc) and shown to play some role

Approach of most studies

- Get sequences:
  - Either a set of known enhancers
  - or upstream sequences of genes assumed to be co-regulated
- Mine them for common motifs
- Rank these motifs
- Search genome for best matches to these motifs
- Test these enhancers experimentally and/or test the motifs by mutating them

Less papers to read!

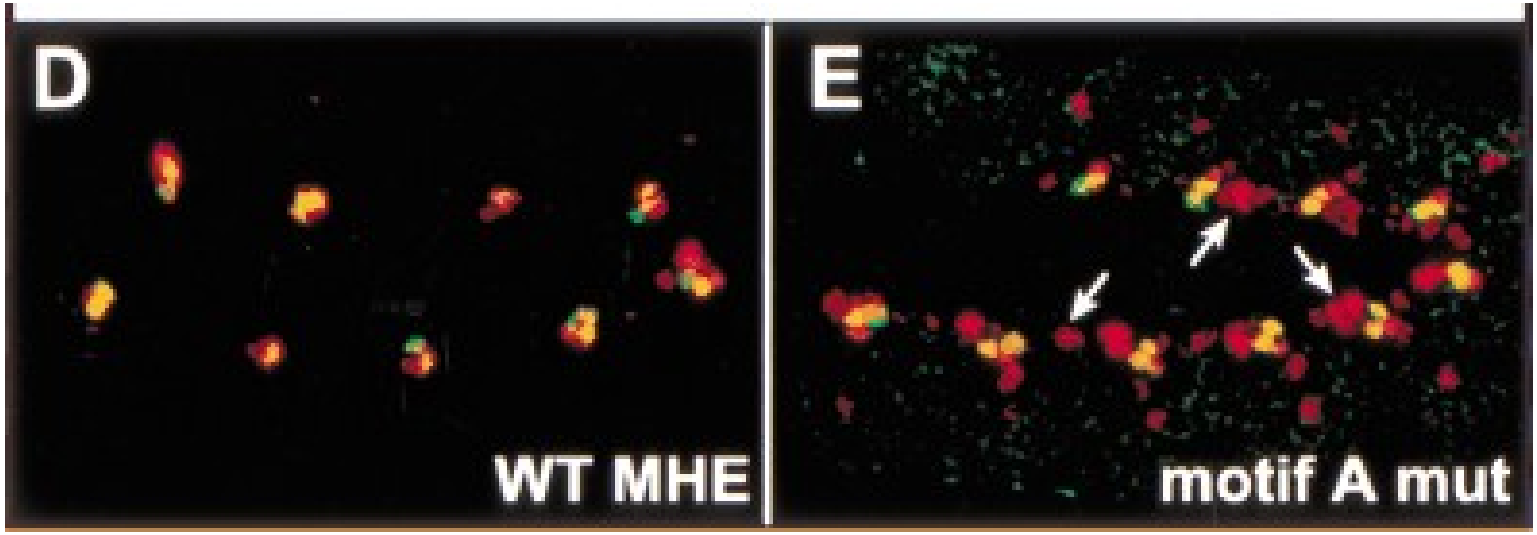
- Found around 10 studies
- Four selected for this talk: Drosophila, C. elegans, Ciona, Mouse
- **Focus:** Setup of the motif discovery analysis
  - Selection of genes
  - Selection of sequences
  - Searching for and scoring of candidate motifs

Drosophila 1: *eve*

- The five transcription factors that regulate *eve* are known
- **Genes/Sequences:** Genome-wide search identifies 34 sequences of <500bp where these 5 matrices match + simulation to see if this is significant
- **Motif Discovery:** AlignACE, gives 755 motifs
- **Ranking:**
  - filter out motifs that are not conserved in *D. virilis (=>25)*
  - cluster by similarity (=>14)
  - filter out all motifs that are not similar to Transfac (=>1)

Validation

- Motif that gave a match to Transfac was mutated and cloned, changes expression

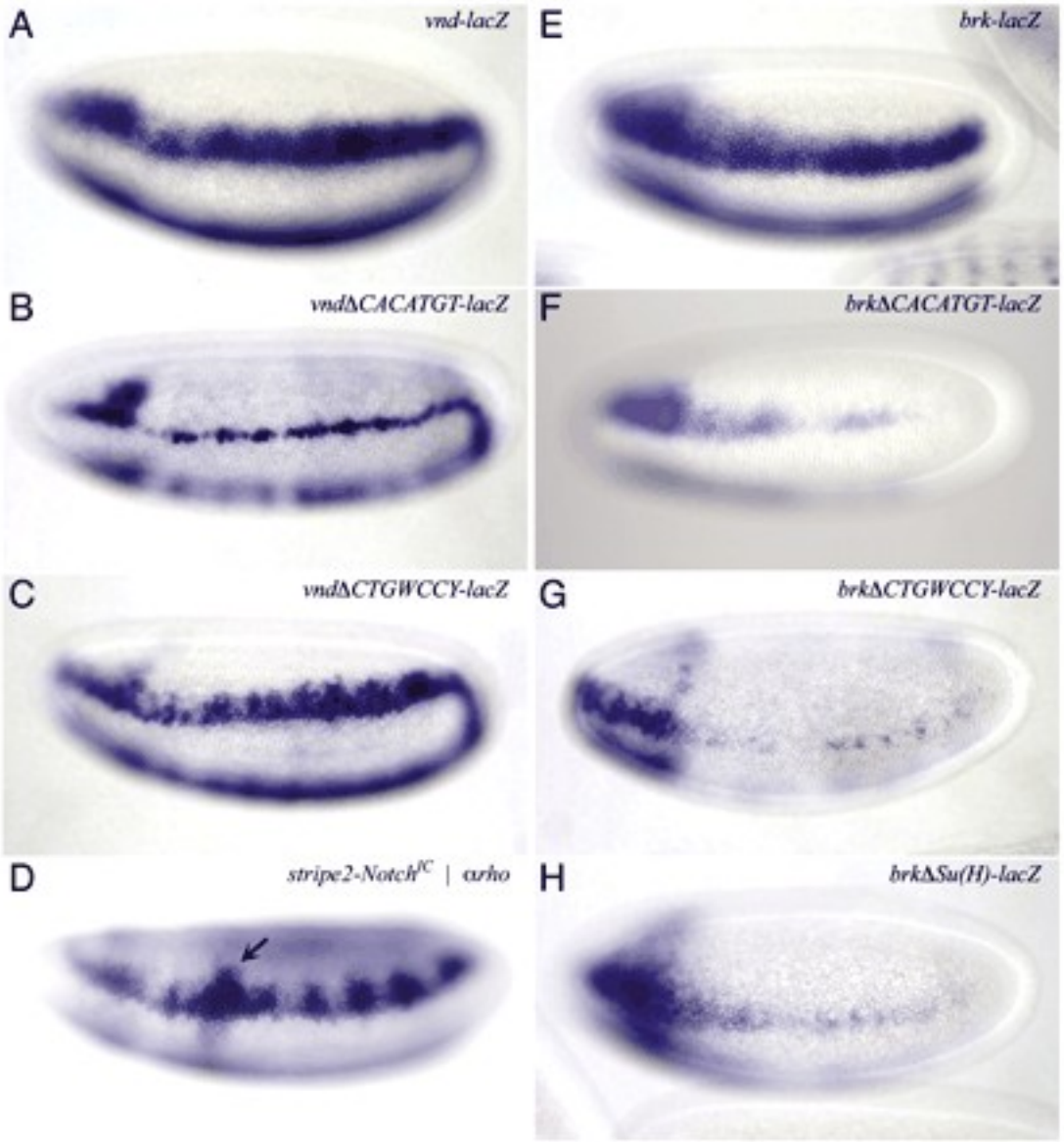


Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. Marc Halfon et al. , Genome research Jul 2002

Drosophila 2

- **Genes and Sequences:**
  - Previous genome-wide search for clustered *Dorsal* binding sites lead to 6 active enhancers, size 300-500bp
  - Background: 20 kb sequence
  - total Size: 3kb!
- **Motif Discovery**:
  - Exhaustive search for n-mers with their own algorithm "Mermaid"
  - Mermaid: consensus-based, allows up to 2 wildcards, parameters and scoring not specified
- Several known (Literature/Transfac) and one new motif found

Validation



Motifs tested by mutation

Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo.
Angelike Stathopoulos et al.
Cell 2002

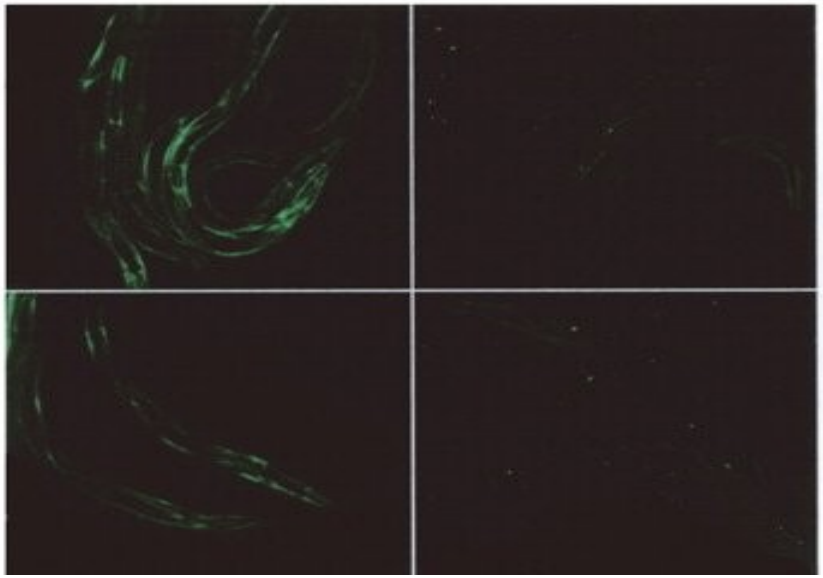
A regulatory code for neurogenic gene expression in the Drosophila embryo.
Michele Markstein et al.
Development 2004

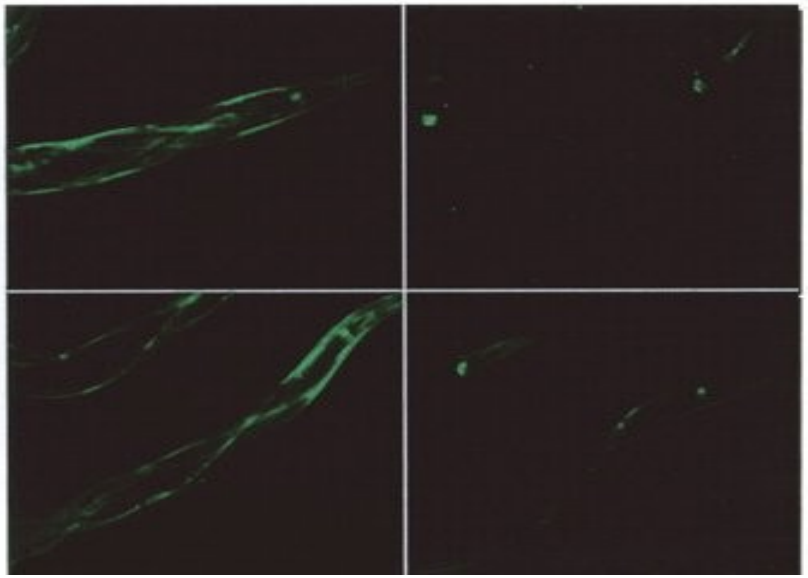
C. elegans

- **Genes:** 41 muscle-specific genes from the literature + 33 *C. briggsae* orthologs, background: 500 sets of 2000 random genes
- **Sequences:** 2kb upstream of selected genes
- **Motif Discovery:**
  - Search: Consensus and Ann-Spec, best three motifs
- **Ranking:**
    - Motifs: respective scores of the discovery programs
    - sequence sets: rank summed match-score over all sequences + test if ranks differ significantly (muscle versus random sequences) + Additional tests (some transfac motifs versus our motifs, conserved versus non-conserved sequences, etc)
    - Alignment with BLASTZ (local) and GLASS (global) (70% identity over 50bp) finds over-representation of hits in conserved regions but still misses more than 50% of the matches

Validation

- Ranking of all genes with the top motifs places some known muscle genes at the top
- Mutation of predicted motifs tested in two selected known muscle genes - strong reduction of expression:





Guhathakurta et al, Novel transcription regulatory elements in Caenorhabditis elegans muscle genes, Genome Res 2004

Ciona intestinalis

- Similar to the last study:
  - 20 known muscle enhancers of 300 bp
  - CisModule is configured to find 4 motifs
  - Searched conserved (*C. savigny:* BLAST+MLAGAN) parts of genome for conserved 3 motifs within 150 bp
  - kept 23 matches close to first exons
  - 7 muscle-enhancers found

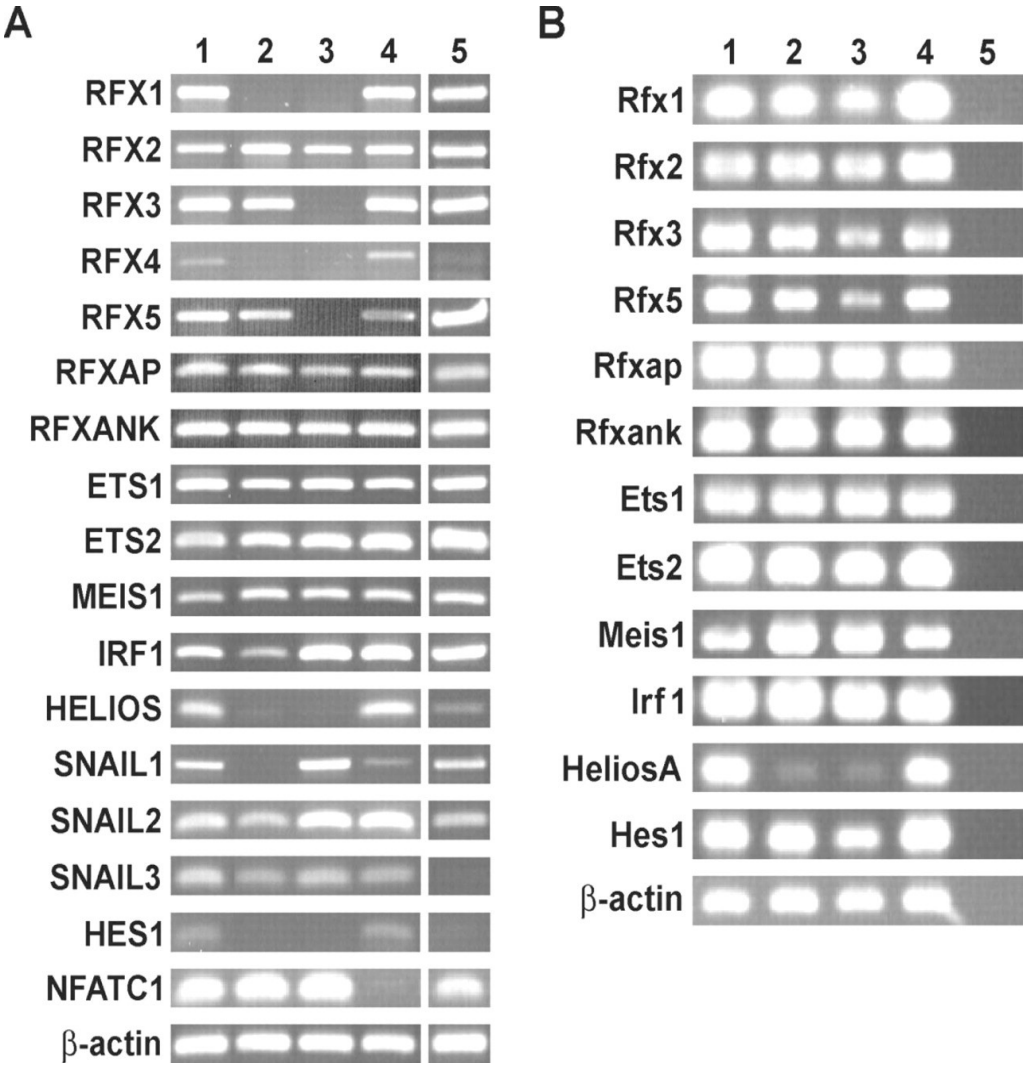
Johnson et al, De novo discovery of a tissue-specific gene regulatory module in a chordate, Gen Res 2005

Mouse

- **Genes:** 41 genes from literature known to be expressed in the lung
- **Sequences:** -1000 to +200 bp relative to standard gene models; as a separate set: orthologous regions from human, background: 1000 random genes
- **Algorithm:** DME/removal of duplicates + Transfac-motifs
- **Ranking:**
  - “Classification error” (sequences that contain a motif are rather in foreground than in background)
  - Same done for mouse and human, only common motifs retained
  - Similarity to Transfac-matrices calculated

Validation

- Top motifs with match to Transfac selected, if their expression in lung has not been described yet: ETS, RFX, SNAIL
- RT-PCR on lung tissues for all paralogs to all members of ETS, RFX and SNAIL
- Factors really are expressed in lung tissues



DNA motifs in human and mouse proximal promoters predict tissue-specific expression.
Andrew Smith et al., PNAS Apr 2006
Computational prediction of novel components of lung transcriptional networks.
M Martinez et al. Bioinformatics Jan 2007

Conclusions

- Setup of your motif discovery analysis:
  - have as accurate sequence data as possible, rather not raw microarray results, not one single enhancer, at least 6-7 in the region of 500bp
  - Use any discovery algorithm yo like
  - Don't use the score of the discovery algorithm, devise your own score based on specificity of matches to foreground vs. background
  - use real promoter sequences as background, not markov models
  - filter with (globally) aligned sequences from one or several closely related species
  - annotate motifs with a database like Transfac

Remarks:

- Guhathakurta: “It is worth noting here that many of the known muscle genes are frequently observed to express also in neuronal tissues.“
- Martinez/Smith: “Many motifs that were shown to be conserved using the known motif analysis were not represented by DME motifs (e.g. SNAIL and HES-1) and vice versa (e.g. RFX and MEIS-1). In most cases, this may be due to DME motifs being shorter in length as compared with the binding sites contained in the TRANSFAC database, consequently making proper alignments difficult. A second possibility is that the DME input parameters excluded certain cis-regulatory elements. Additionally, some factors predicted by the novel motif analysis were not predicted by the known motifs. (...) **These findings suggest reliance on a single computational tool may be limiting”**
- Halfon: To see if a given combination of motifs is significant, take the matches, put them to random locations and re-scan