Algorithms in Computational Genomics at Tau

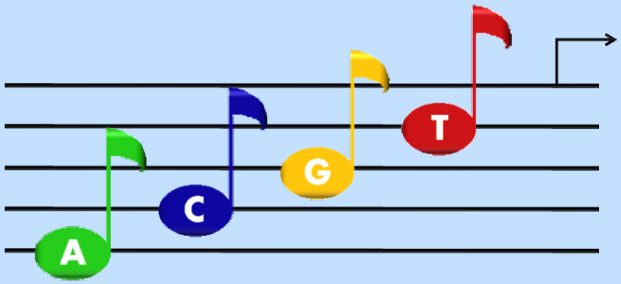
Gene regulation, protein networks and disease – a computational perspective

Ron Shamir
School of Computer Science
Tel Aviv University

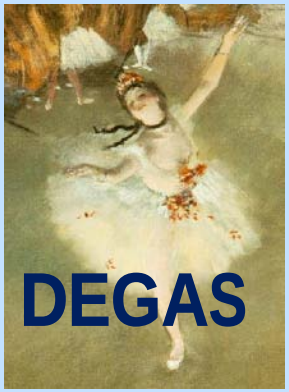
CPM Helsinki July 3 2012

Outline

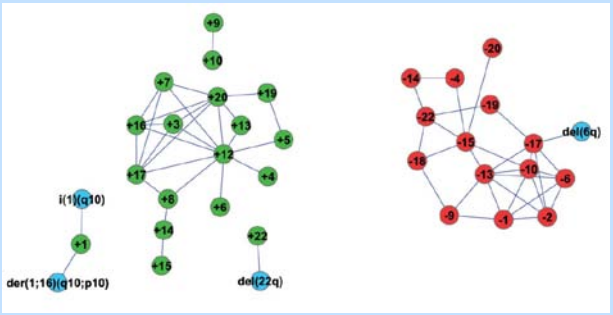
- Finding regulatory motifs I, II, III



- Utilizing case-control expression profiles and networks I, II

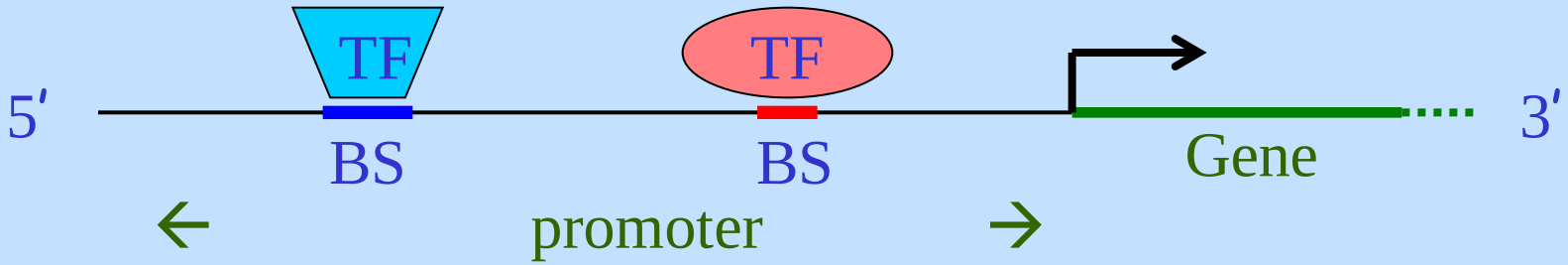


- Chromosomal aberrations in cancer



Regulation of Transcription

- A gene's ranscription regulation is mainly encoded in the DNA in a region called the promoter
- Each promoter contains several short DNA subsequences, called binding sites (BSs) that are bound by specific proteins called transcription factors (TFs)



Position Weight Matrix (PWM)

Score: product of base probabilities.

Need score threshold for hits.

A	0.1	0.8	0	0.7	0.2	0
C	0	0.1	0.5	0.1	0.4	0.6
G	0	0	0.5	0.1	0.4	0.1
T	0.9	0.1	0	0.1	0	0.3

ATGCAGGATACACCGATCGGTA 0.0605

GGAGTAGAGCAAGTCCCGTGA 0.0605

AAGACTCTACAATTATGGCGT 0.0151

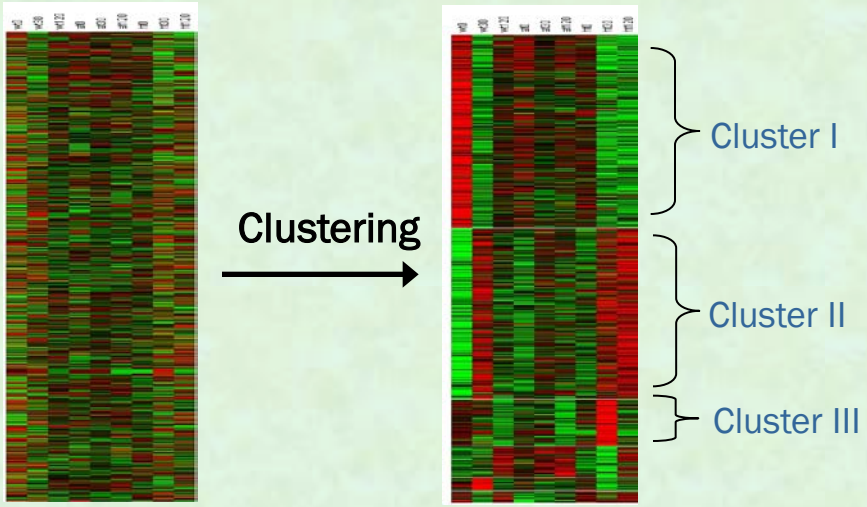
C. Linhart, Y. Halperin *Genome Research 08*

I. Finding Regulatory Motifs

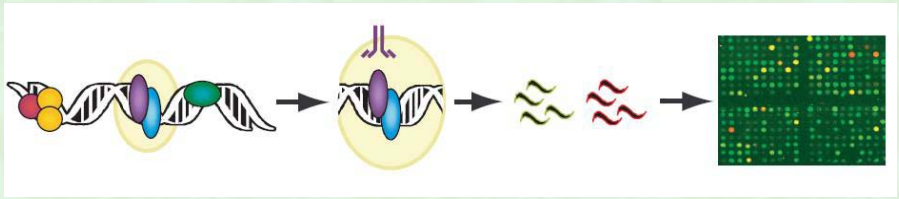
Motif discovery:
The two-step strategy

Co-regulated gene set

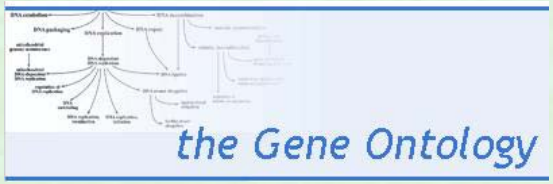
Gene expression microarrays



Location analysis (ChIP-chip, ...)

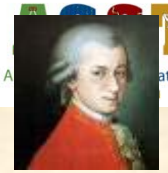


Functional group (e.g., GO term)



Promoter sequences

Motif discovery



Amadeus

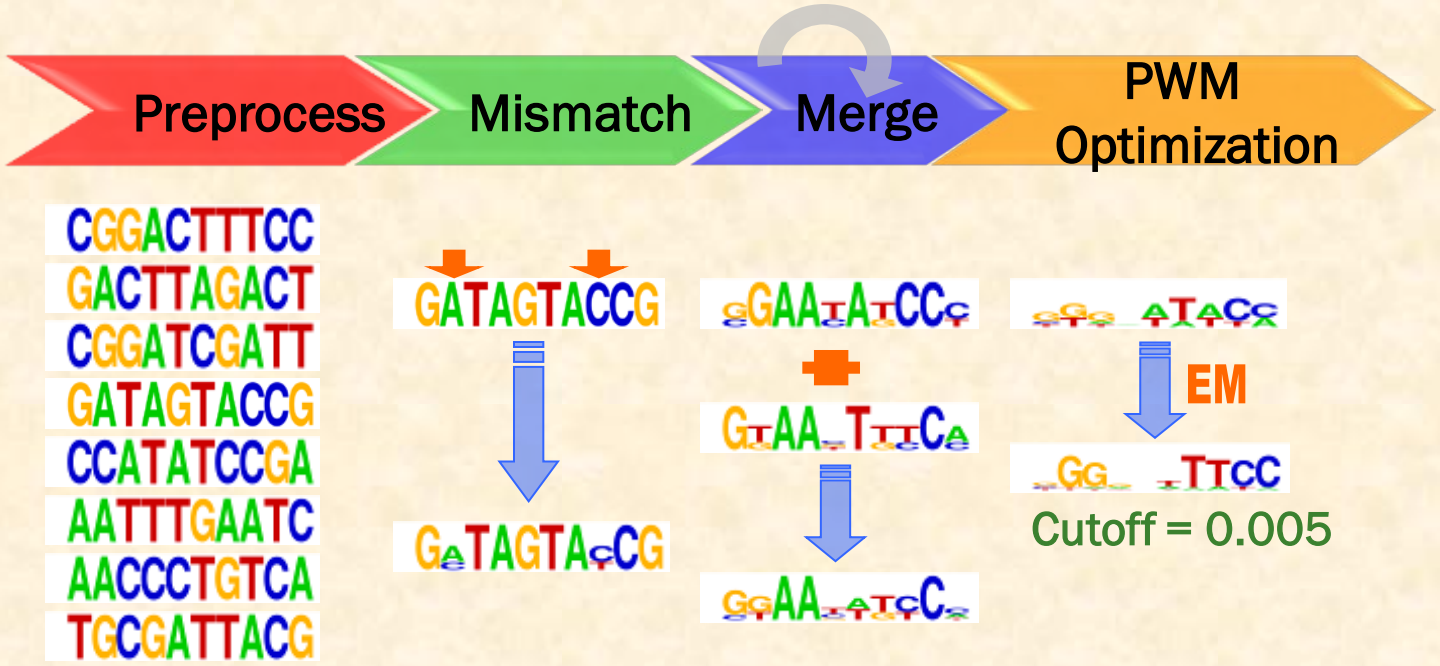
A Motif Algorithm for Detecting Enrichment in mUltiple Species

- **Supports diverse motif discovery tasks:**
  1. Find over-represented motifs in given sets of genes.
  2. Identify motifs with global spatial features given only the genomic sequences.
- **How?**
  - A general pipeline architecture for enumerating motifs.
  - Different statistical scoring schemes of motifs for different motif discovery tasks.

Motif search algorithm

- Pipeline of refinement phases of increased complexity

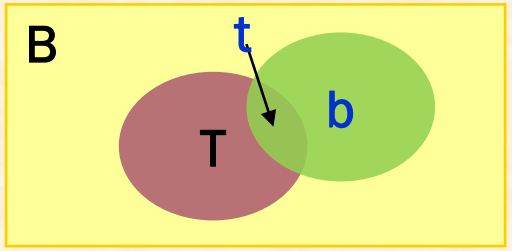
➢ Phases:



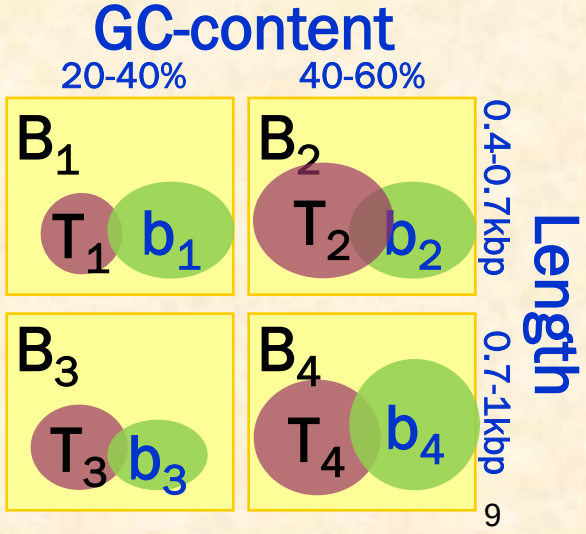
➢ Motif Model:

Scoring over-represented motifs

- **Input:** Target set (size *T*) = co-regulated genes
  Background (**BG**) set (size *B*) = entire genome
- **Motif enrichment scoring:**
- Hyper-geometric



- Binned enrichment score
- Binomial

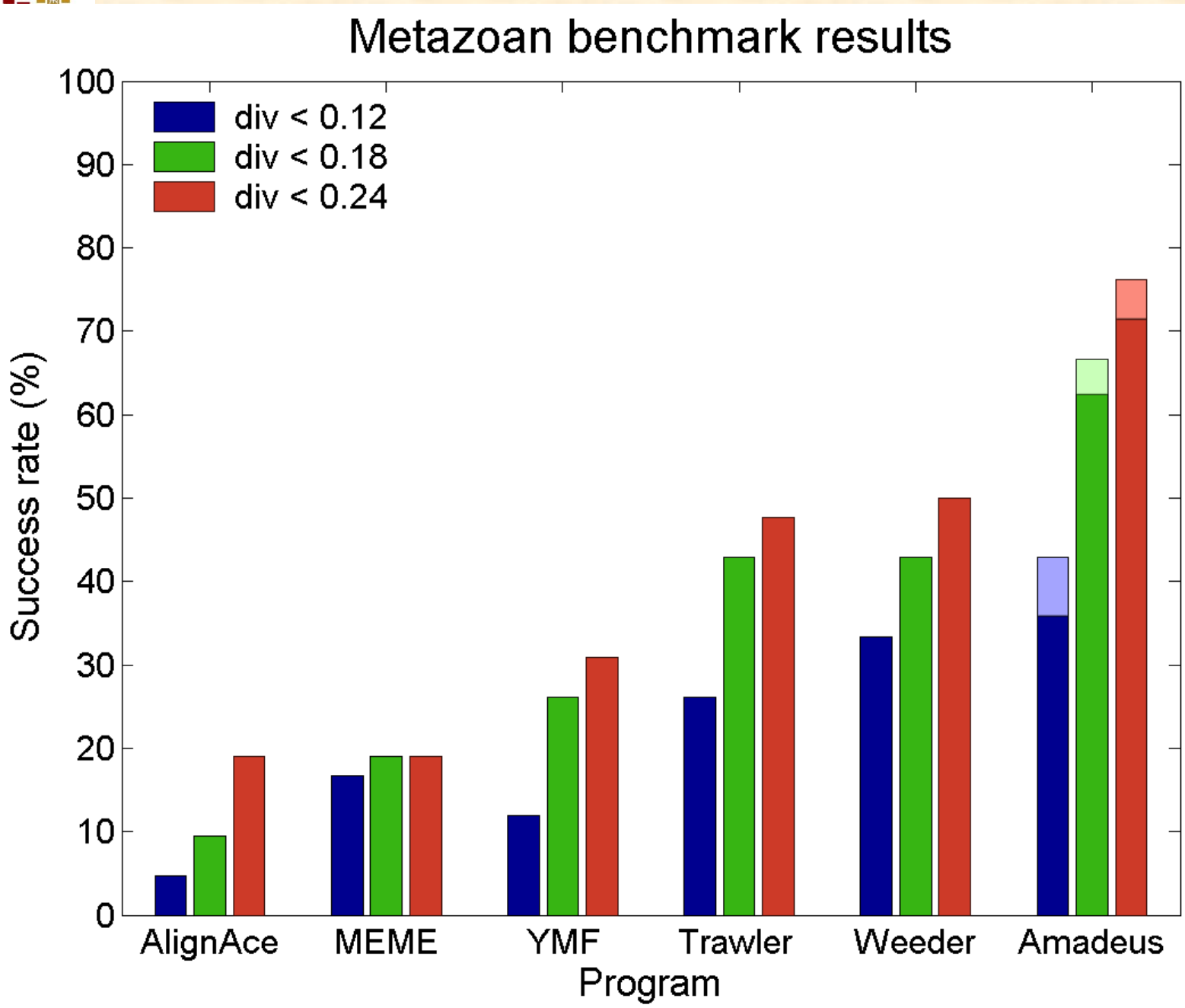


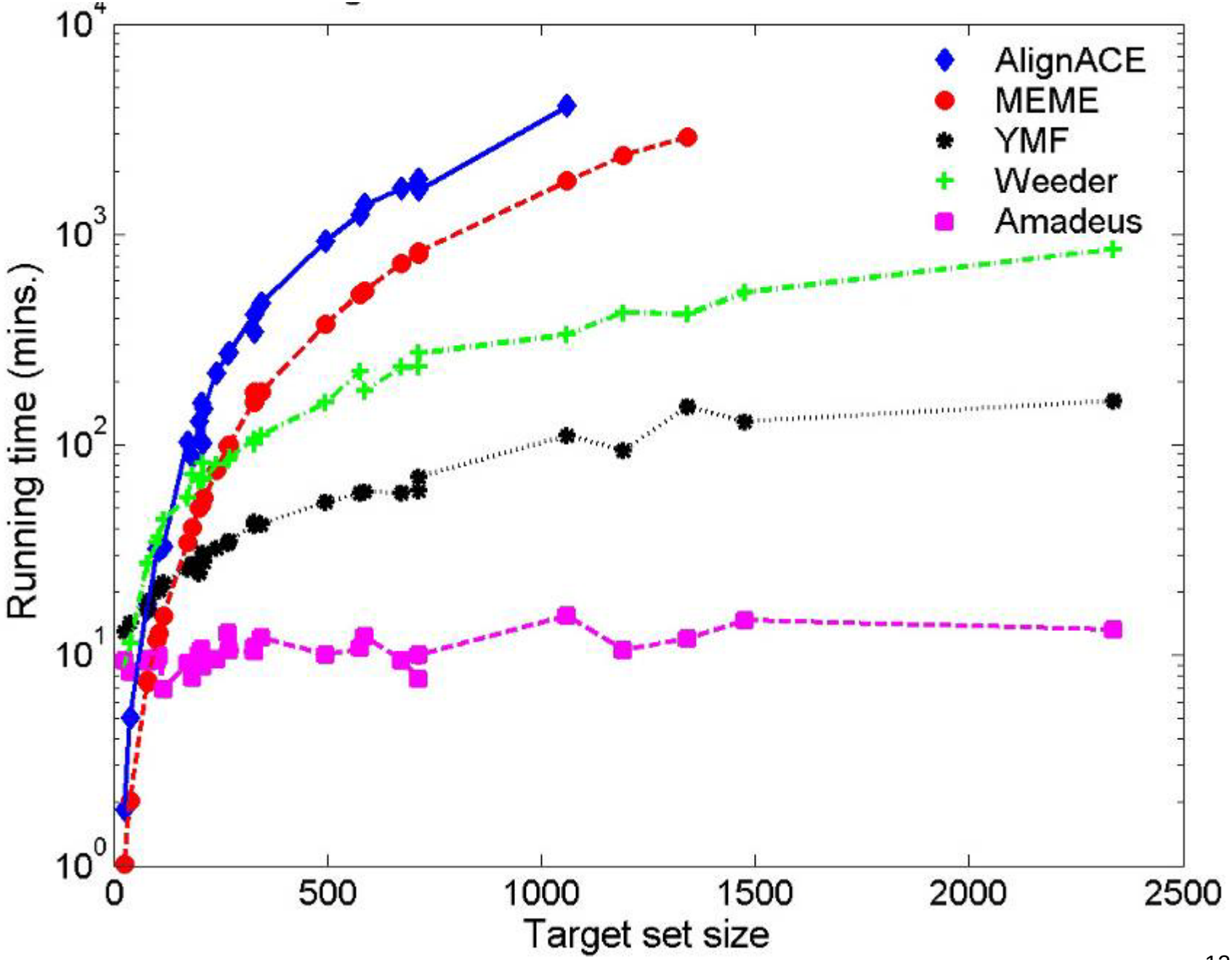
Metazoan motif discovery benchmark:

42 target sets of 26 TFs, 8 miRNAs from 29 studies (expression, Chip-ChIP,..) in human, mouse, fly, worm.

All motifs are experimentally verified

Average target set size: 400 genes (383 Kbp)

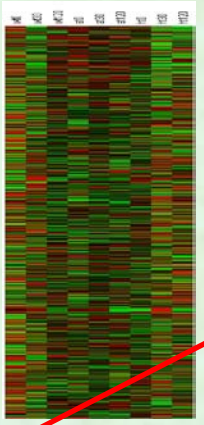

Metazoan benchmark results
div < 0.12
div < 0.18
div < 0.24
Success rate (%)
100
90
80
70
60
50
40
30
20
10
0
AlignAce
MEME
YMF
Trawler
Weeder
Amadeus
Program


AlignACE
MEME
YMF
Weeder
Amadeus
Running time (mins.)
Target set size
10^0
10^1
10^2
10^3
10^4
0
500
1000
1500
2000
2500

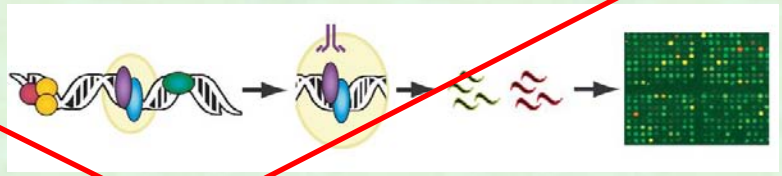
Amadeus – Global spatial analysis

Co-regulated gene set

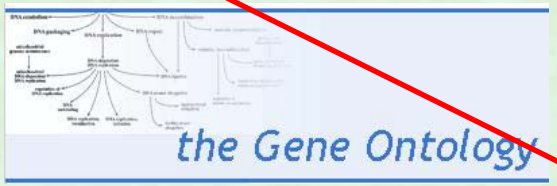
Gene expression microarrays



Location analysis (ChIP-chip, ...)



Functional group (e.g., GO term)



Promoter sequences

Output

Motif(s)

Task II: Global analyses

Scores for spatial features of motif occurrences

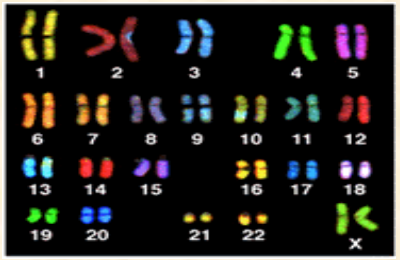
Input: Sequences (no target-set / expression data)

Motif scoring:

- Localization w.r.t the TSS

- Strand-bias

- Chromosomal preference



Global analysis:

Chromosomal preference in *C. elegans*

Input:

- All worm promoters (~18,000)
- Score: chromosomal preference

Results:

Novel motif on chrom IV

Motif 1 p-value = 8.1E-63

Scores

Organism	Enrichment		Strand bias	Chromosomal preference
	HG	Binned		
C. elegans (1)	1.0	0.5055	3.7E-7 (+)	8.1E-63 (IV)

Organism	Localization			
	BG	Target	Target vs. BG	
C. elegans (1)	0.0357	0.0357	1.0	

Cell
Volume 127, Issue 6, 15 December 2006, Pages 1193-1207

Large-Scale Sequencing Reveals 21U-RNAs and Additional MicroRNAs and Endogenous siRNAs in *C. elegans*

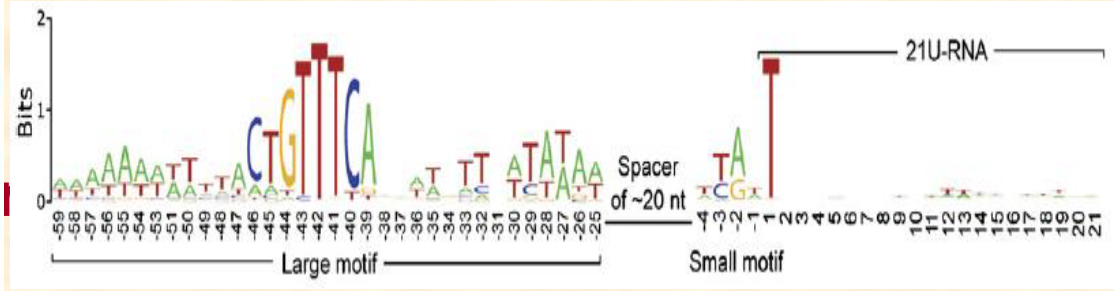
J. Graham Ruby,[1,2] Calvin Jan,[1,2] Christopher Player,[1] Michael J. Axtell,[1,4] William Lee,[3] Chad Nusbaum,[3] Hui Ge,[1] and David P. Bartel[1,2,*]

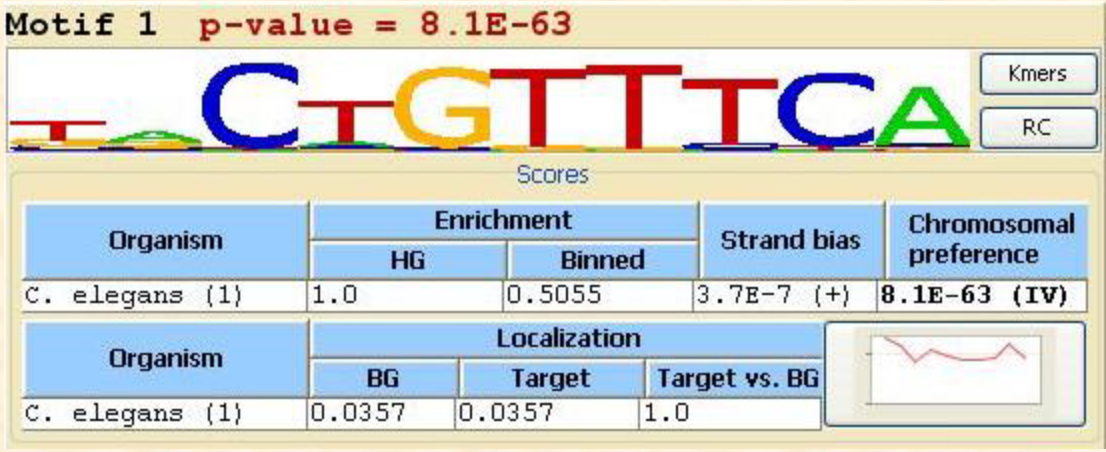
Global analysis:

Chromosomal preference in *C. elegans*

SUMMARY

We sequenced ~400,000 small RNAs from *Caenorhabditis elegans*. Another 18 microRNA (miRNA) genes were identified, thereby extending to 112 our tally of confidently identified miRNA genes in *C. elegans*. Also observed were thousands of endogenous siRNAs generated by RNA-directed RNA polymerases acting preferentially on transcripts associated with spermatogenesis and transposons. In addition, a third class of nematode small RNAs, called 21U-RNAs, was discovered. 21U-RNAs are precisely 21 nucleotides long, begin with a uridine 5′-monophosphate but are diverse in their remaining 20 nucleotides, and appear modified at their 3′-terminal ribose. 21U-RNAs originate from more than 5700 genomic loci dispersed in two broad regions of chromosome IV—primarily between protein-coding genes or within their introns. These loci share a large upstream motif that enables accurate prediction of additional 21U-RNAs. The motif is conserved in other nematodes, presumably because of its importance for producing these diverse, autonomously expressed, small RNAs (dasRNAs).



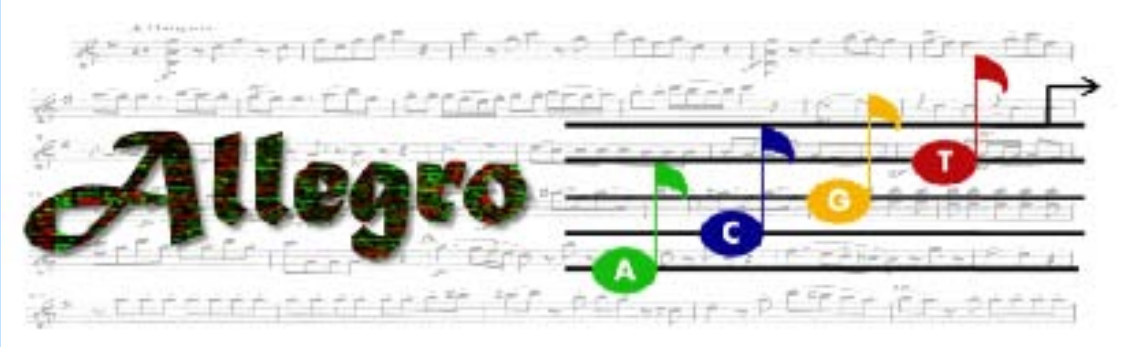


Organism	Enrichment		Strand bias	Chromosomal preference
	HG	Binned		
C. elegans (1)	1.0	0.5055	3.7E-7 (+)	**8.1E-63 (IV)**

Organism	Localization			
	BG	Target	Target vs. BG	
C. elegans (1)	0.0357	0.0357	1.0	

Y. Halperin, C. Linhart, I. Ulitsky *NAR 10*

II. Finding Transcriptional Programs



Goal

Given expression profiles, find the transcriptional programs active in them:

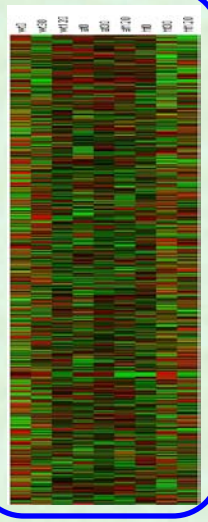
- the co-regulated genes,
- the motifs that govern their co-regulation

Our goal: bypass the two-step approach

Expression data

Co-regulated gene set

Gene expression microarrays

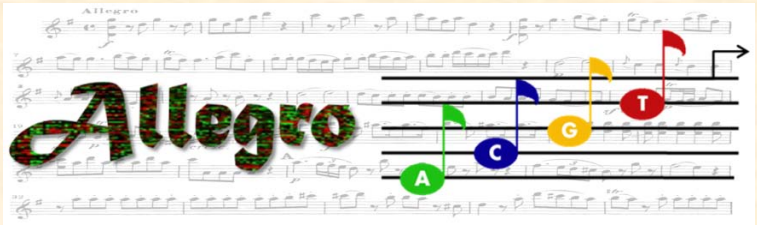


Clu

Simultaneous inference of the motifs and the exp profiles of their targets

Promoter sequences

Output



Motif(s)

Allegro: expression model

- Discretization of expression patterns

e_1=Up (U) ≥1.0
e_2=Same (S) (-1.0, 1.0)
e_3=Down (D) ≤-1.0

Expression pattern

	c_1	c_2	…	c_m
g	-2.3	-0.8		1.5

Discrete expression Pattern (DEP)

	c_1	c_2	…	c_m
g	D	S	…	U

- Condition frequency matrix (**CFM**)

$$F =$$

	c_1	c_2	…	c_m
U	0.05	0.1	…	0.78
S	0.9	0.2	…	0.14
D	0.05	0.7	…	0.08

- Condition weight matrix (**CWM**)

$$F^{(W)} = \left\{ \log\left(f_{ij} / r_{ij} \right) \right\}$$

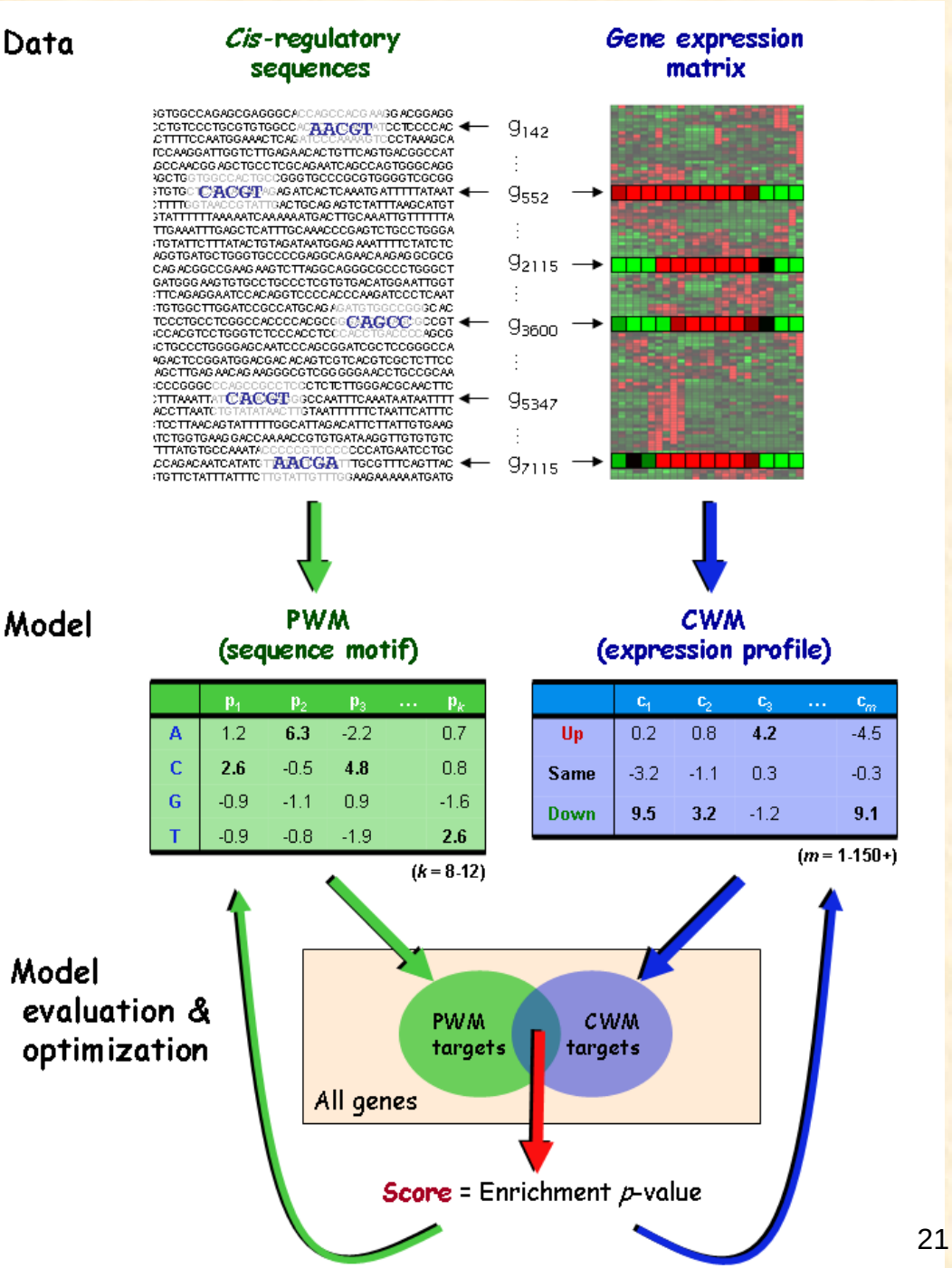
($R=\{r_{ij}\}$ is the BG CFM)

$\Rightarrow$ Log-likelihood ratio (**LLR**) score

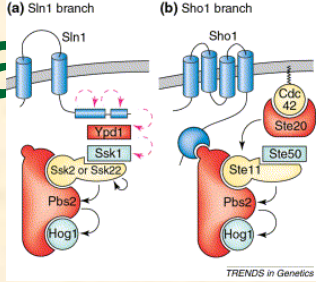
Algorithms in Computational Genomics at Tau

Allegro

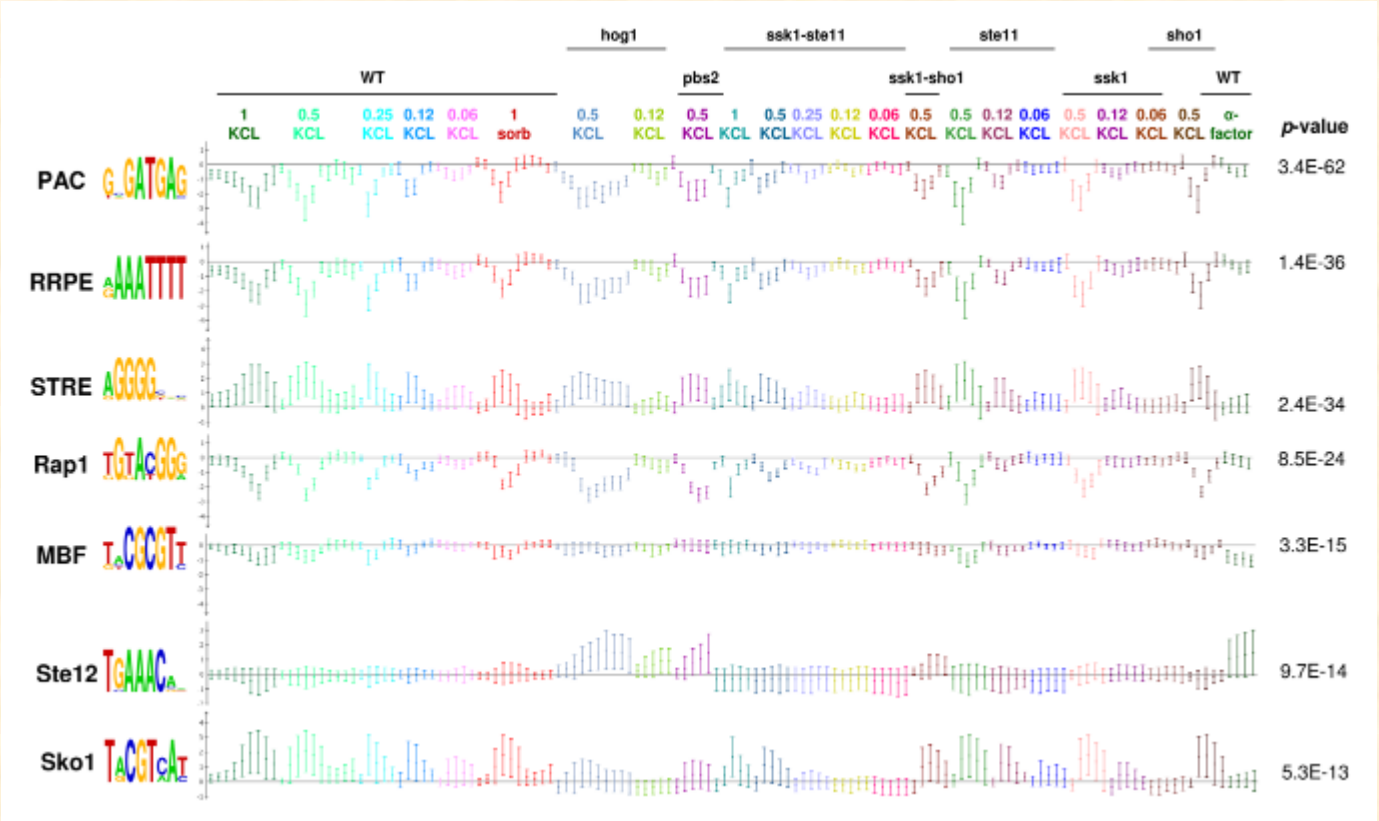
overview



Yeast osmotic shock pathwa



- ~6,000 genes, 133 conditions [O'Rourke et al. '04]



- Allegro can discover multiple motifs with diverse expression patterns, even if the response is in a small fraction of the conditions
- Extant two-step techniques recovered only 4 of the above motifs:
  - **K-means/CLICK + Amadeus/Weeder:** RRPE, PAC, MBF, STRE
  - **Iclust + FIRE:** RRPE, PAC, Rap1, STRE

3' UTR analysis: Human stem cells

- ~14,000 genes, 124 conditions (various types of proliferating cells) [Mueller et. al, Nature'08]
- Biases in length / GC-content of 3' UTRs, e.g.:

100 highly-expressed genes in…	3' UTR: length	GC
Embryoid bodies	584	47%
Undifferentiated ESCs	774	44%
ESC-derived fibroblasts	1240	39%
Fetal NSCs	1422	43%

(ESCs = embryonic stem cells, NSCs = neural stem cells)

- Extant methods / Allegro with HG score: report only false positives

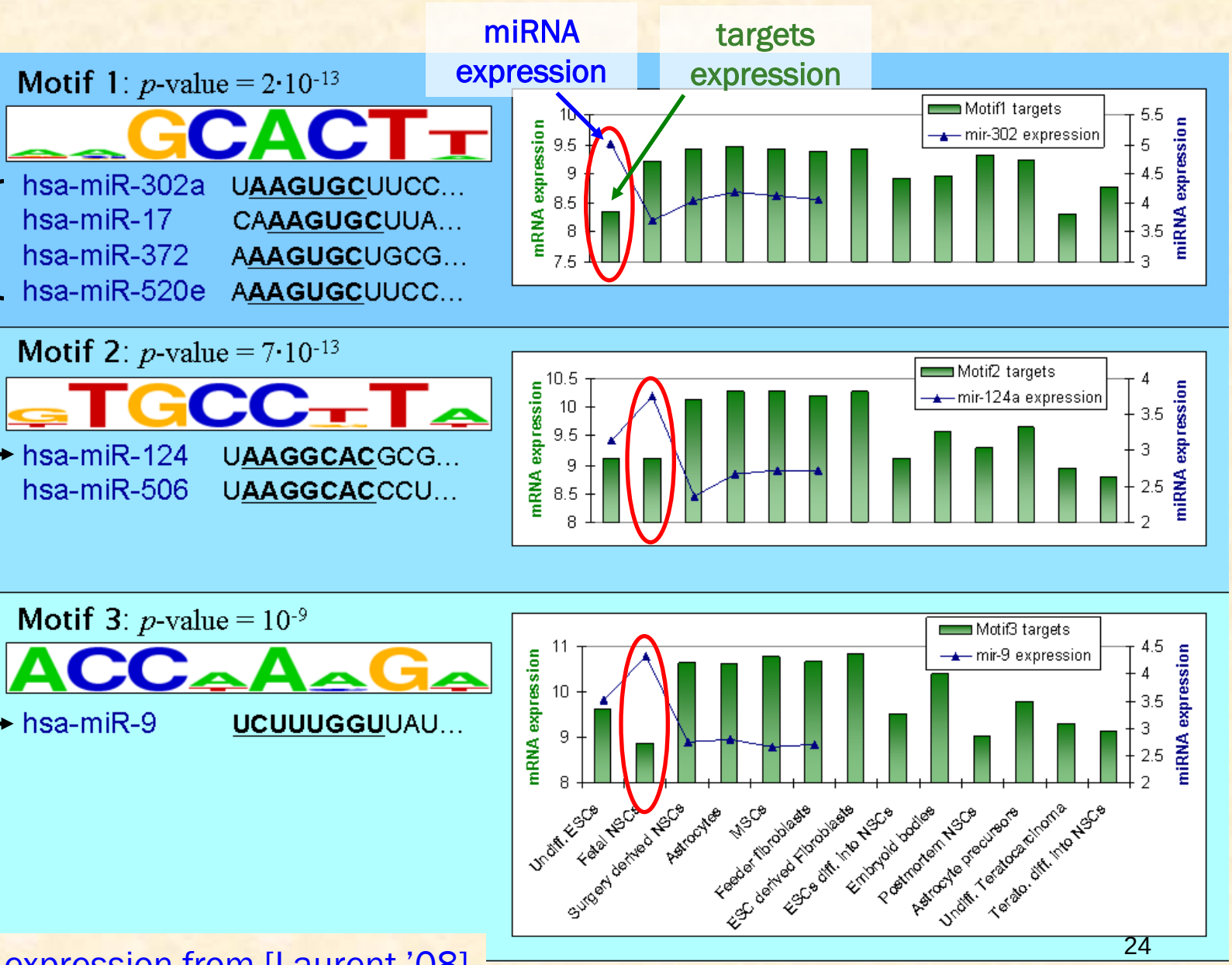
Human stem cells: results using binned score

Current knowledge

Most highly expressed miRNAs in human/mouse ESCs

Abundant & functional in neural cell lineage

Expressed specifically in neural lineage; active role in neurogenesis



miRNA expression from [Laurent '08]

Chaim Linhart

Yonit Halperin

Igor Ulitsky

Yaron Orenstein

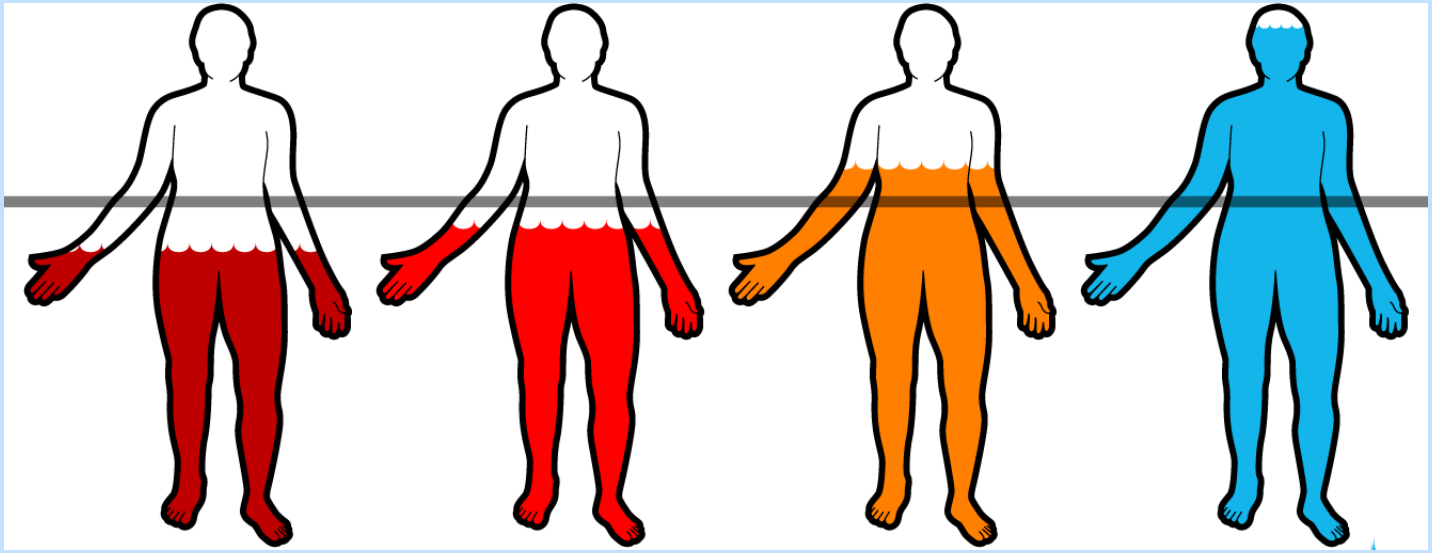
Open questions

- Better PWM inference: new scores, algs
- Richer models for in vivo / in vitro data – really helpful or diminishing return?
- How to evaluate model quality: match to literature? Ranking based? In vivo? In vitro?
- Integration of motif finding & expression
- Principled means to find motif pairs

I. Ulitsky, R. M. Karp *RECOMB 09*

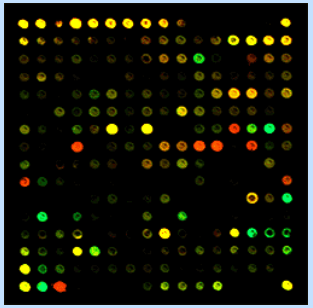
I. Ulitsky, A. Krishnamurthy, R. M. Karp *PLoS One 10*

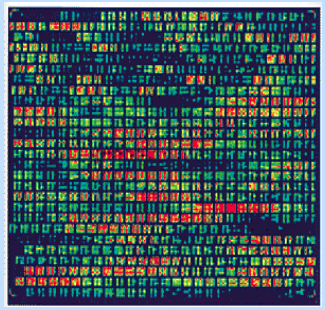
Using expression profiles and protein networks to understand cancer I



DNA chips / Microarrays

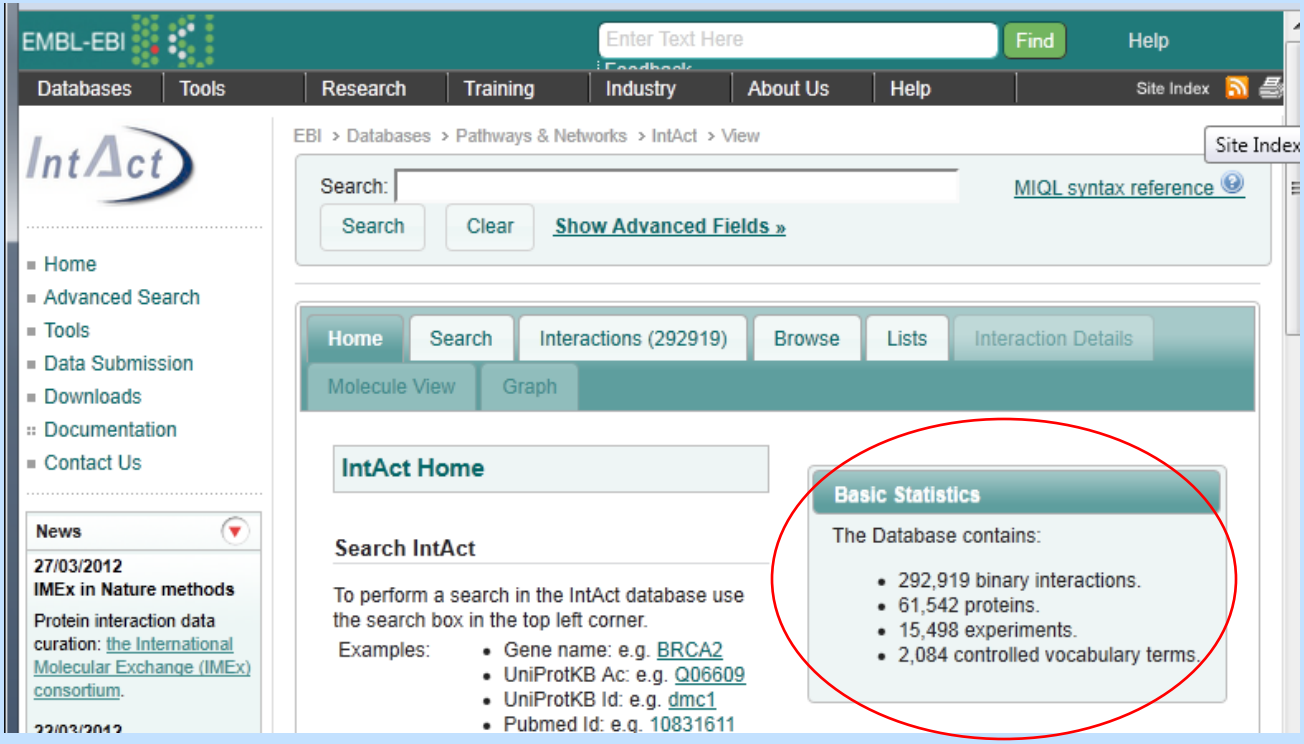
- Simultaneous measurement of expression levels of all genes.
- Global view of cellular processes.
- > 800,000 profiles available in ArrayExpress





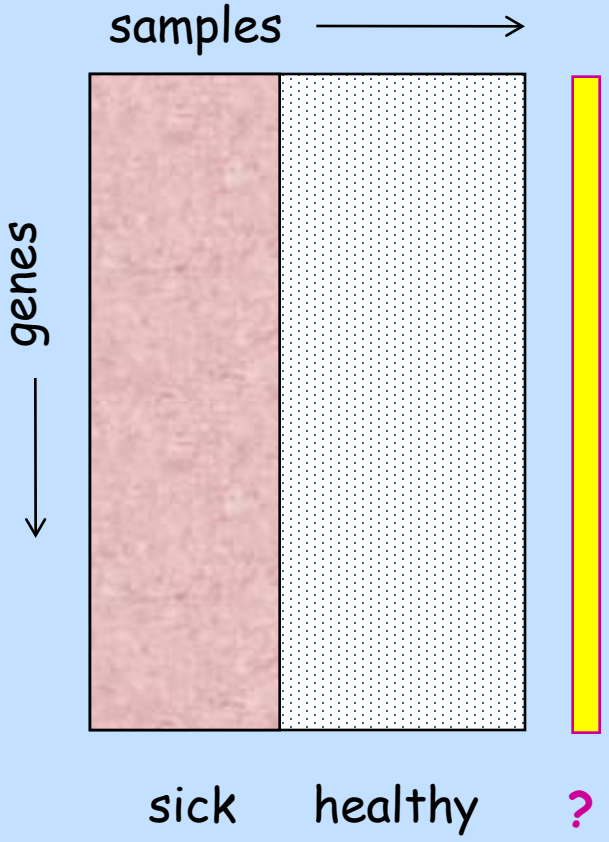
Protein-protein interactions (PPIs)

- A regulates/binds to B
- High throughput: abundant, noisy
- Large, readily available resource

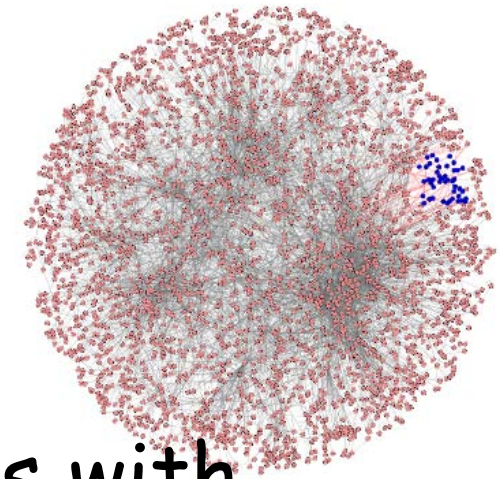


Case/control studies

- A typical study: 100s expression profiles of sick (case) & healthy (control) individuals
- Classification: Given a partition of the samples into types, classify the types of new samples
- Can the network help?



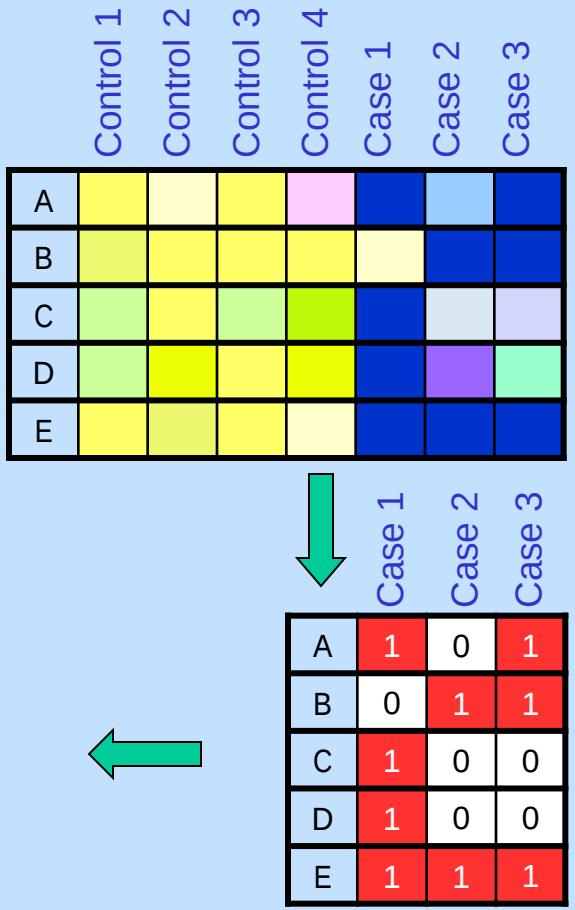
The network angle



- Integrate case-control profiles with network information
- Extract dysregulated pathways specific to the cases
- Account for heterogeneity among cases
- Meaningful pathway: connected

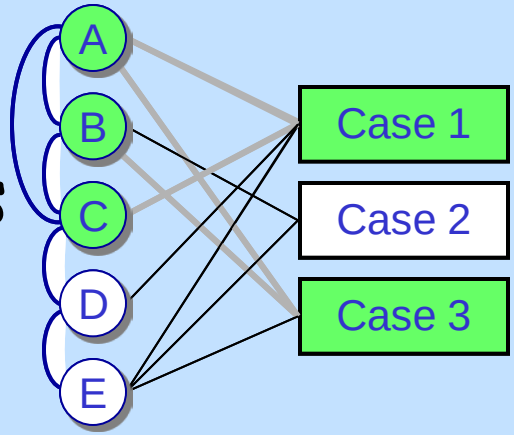
Preprocessing

- For each gene, use the distribution of values among the controls to decide if the gene is dysregulated in each of the cases



Dysregulated pathway

- Input:
  - Bipartite graph: genes, cases
  - Edge (gene g, case c) if g is dysregulated in c
  - A network over the genes
- Dysregulated pathway (DP): smallest connected subnetwork s.t. sufficiently many $\geq k$ genes are dysregulated in all but few $\leq l$ cases
- Small pathway → focused disease explanation
- Min connected set cover problem



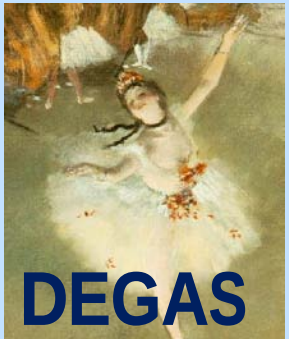
k=2,l=1

Complexity

- **Set cover** problem: Given sets of elements, find fewest sets that cover all elements

k	l	G	Problem
1	0	Clique	Set cover
k	0	Clique	Set k-cover
1	>0	Clique	Partial set cover
1	0	Any	Connected set cover (Shuai & Hu 06)

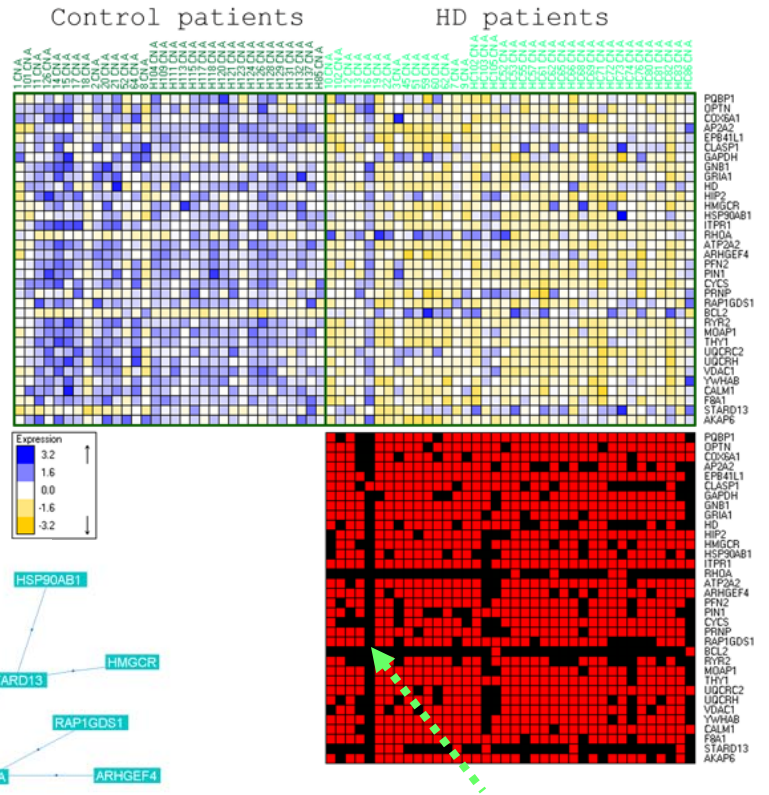
- All are NP-Hard
- Devised approximation and heuristic algs

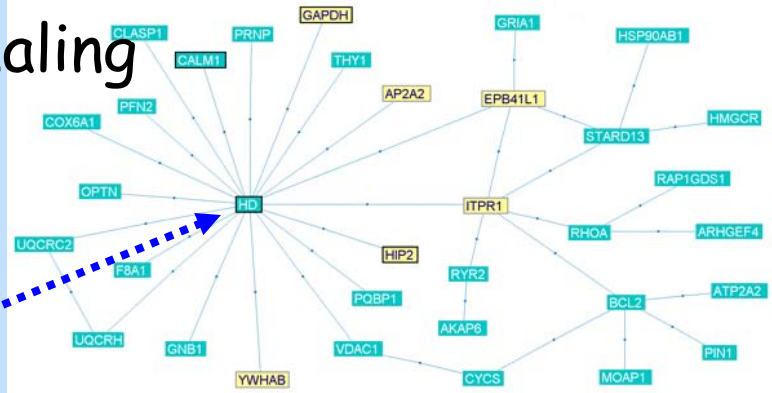


DysrEgulated Gene set Analysis via Subnetworks

Huntington Disease down-regulated pathway

- Brain exp profiles of 38 patients, 32 controls (Hodges et al 06)
- The most significant pathway found for k=25 ($p < 0.005$)
- Enriched with:
  - HD modifiers
  - HD relevant genes
  - Calcium signaling





Huntingtin

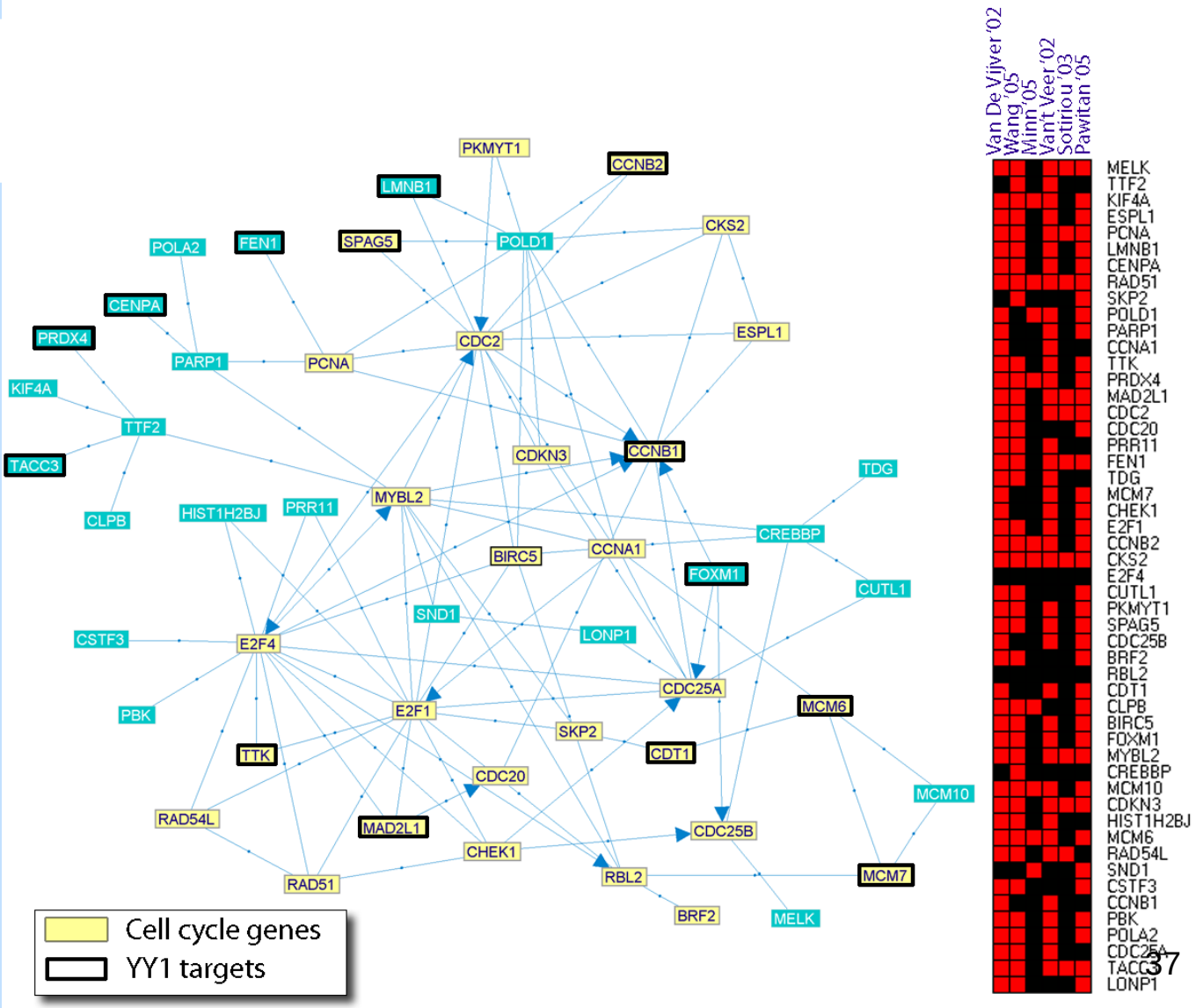
outlier

Breast cancer meta-analysis

- 6 breast cancer studies comparing poor and good prognosis
  - Van't Veer *et al. Nature* 2002
  - Van de Vijver *et al. NEJM* 2002
  - Wang *et al. Lancet* 2005
  - Minn *et al. Nature* 2005
  - Sotiriou *et al. PNAS* 2003
  - Pawitan *et al. Breast Cancer Research* 2005
- Poor prognosis = metastases within 5 years
- 1,004 patients in total
- Elements = studies
- Discovered 2 significant DPs associated with poor prognosis and one associated with good prognosis

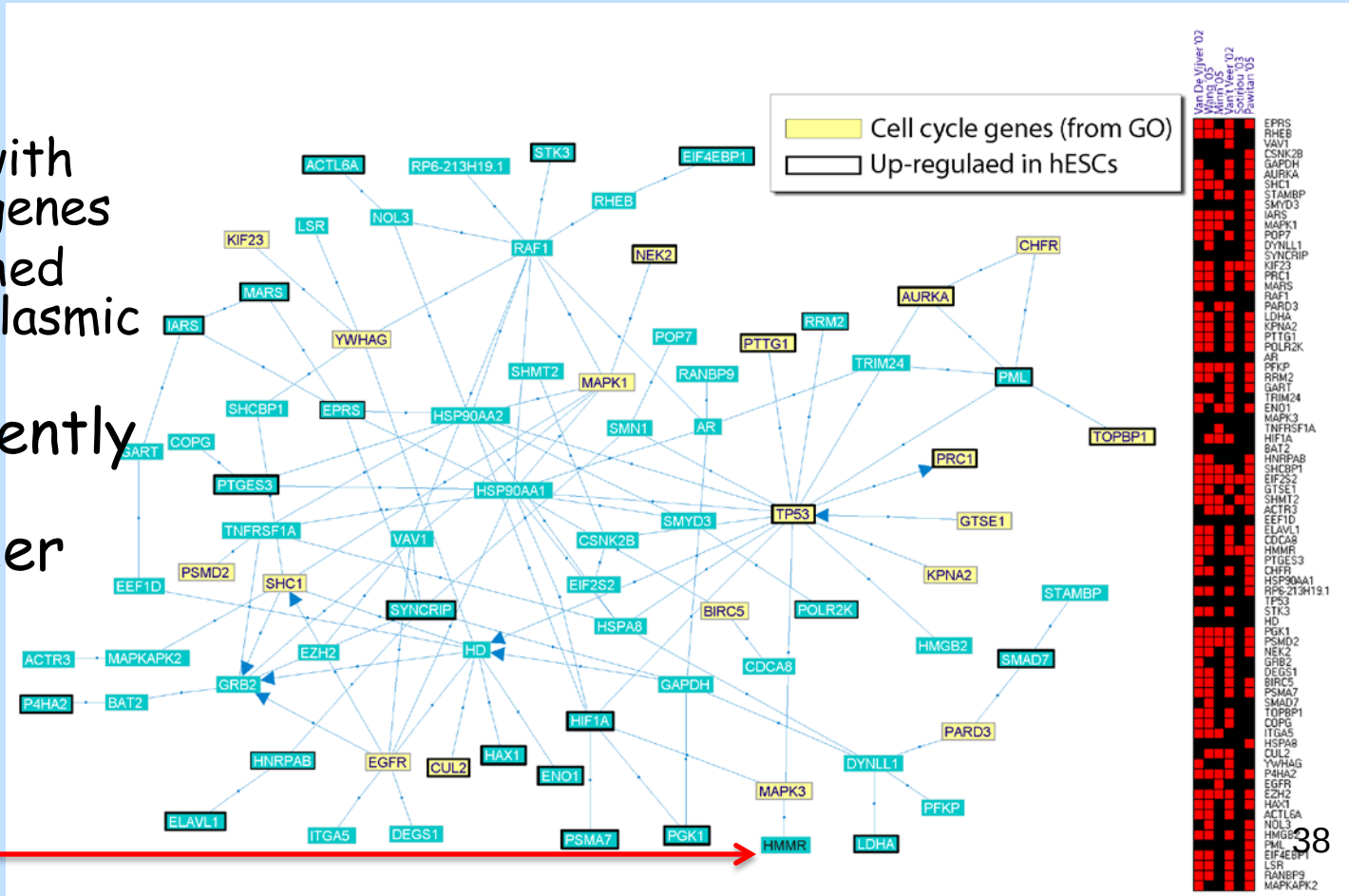
Poor prognosis network 1

- k = 40, l = 2, p<0.005
- Enriched with cell-cycle associated genes ($p=2\cdot10^{-26}$) & YY1 targets ($p=2.42\cdot10^{-16}$)
- Enriched with genes localized to the nucleus



Poor prognosis network 2

- Found by removing network 1 and repeating the search ($k=50$; $p < 0.005$)
- Also significantly enriched with cell cycle genes
- Not merely a segmentation of a single network:
  - Only DP2 network is strongly enriched with stem cell genes
  - DP2 enriched with cytoplasmic genes
- HMMR: recently discovered breast cancer risk factor



Summary

- A method for finding subnetworks of dysregulated genes
- Specific to cases, but allows outliers and exception
- Connected set cover paradigm
- Better approximations??

Dick Karp,
Berkeley

Igor Ulitsky,

Whitehead Inst

Akshay
Krishnamurthy
CMU

Support: Israel Academy of Sciences, Wolfson Foundation, Edmond J. Safra Foundation, US-Israel BSF, German-Israeli Fund, EU 6th and 7th Frameworks, Intel, IBM, I-CORE Gene regulation & disease.