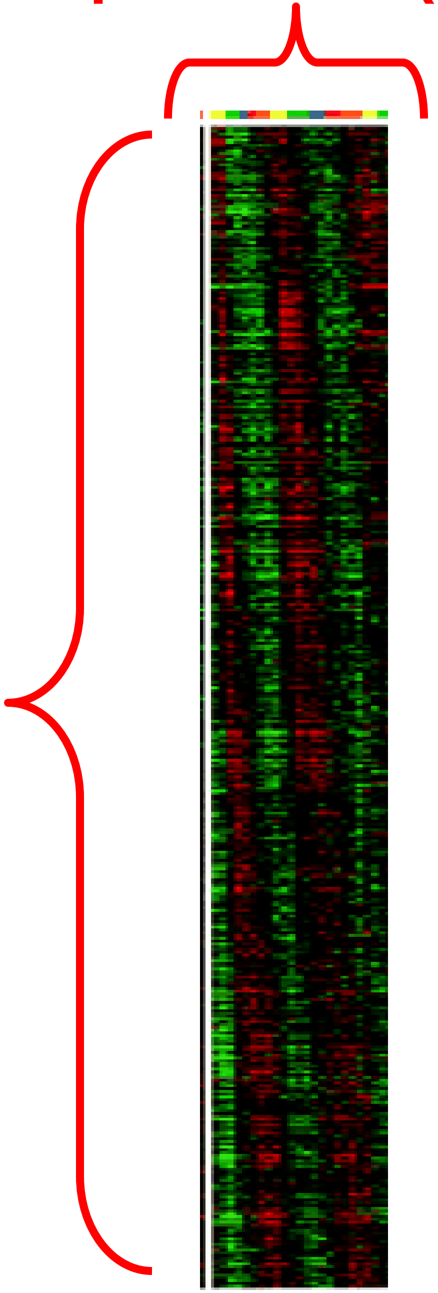
02-716
Cross species analysis of genomics data

Cross species analysis of expression data:
Studying the cell cycle in multiple species

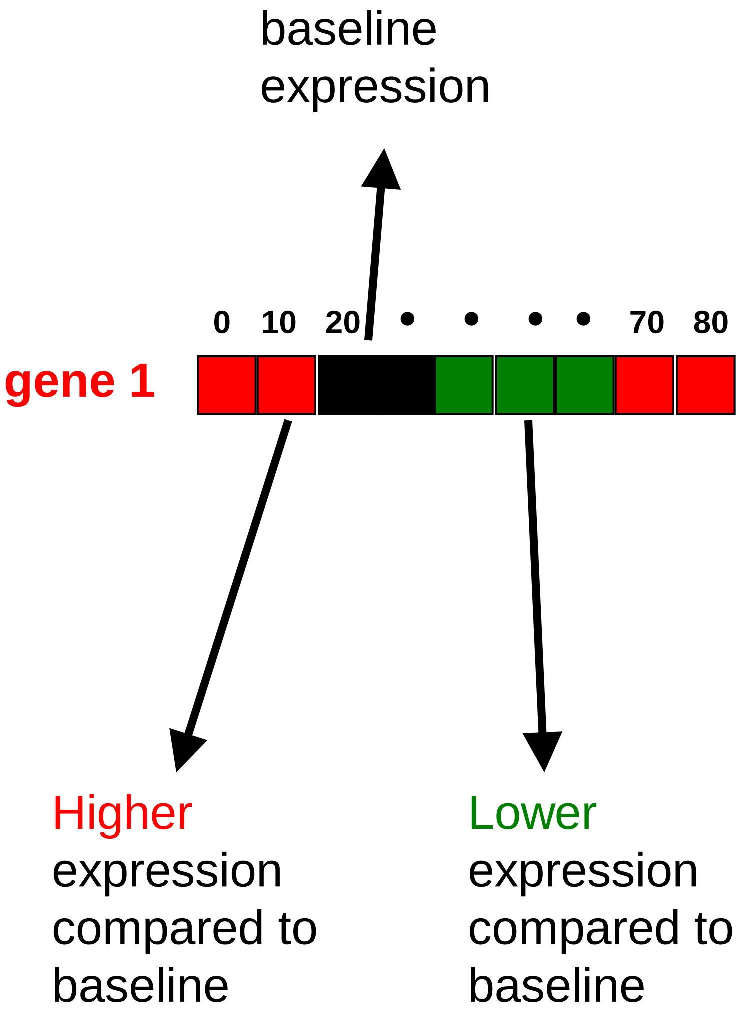
Time series expression data

Expression = level of gene (protein) in this experiment

Experiments (over time)



genes



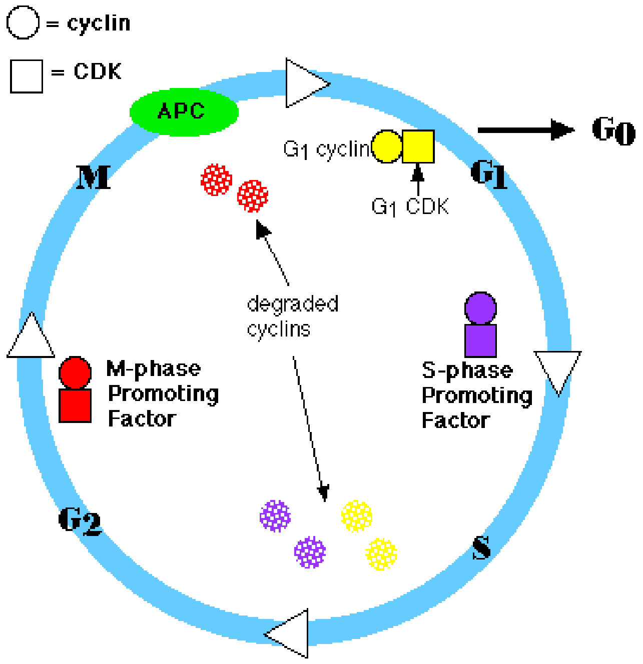
Higher expression compared to baseline

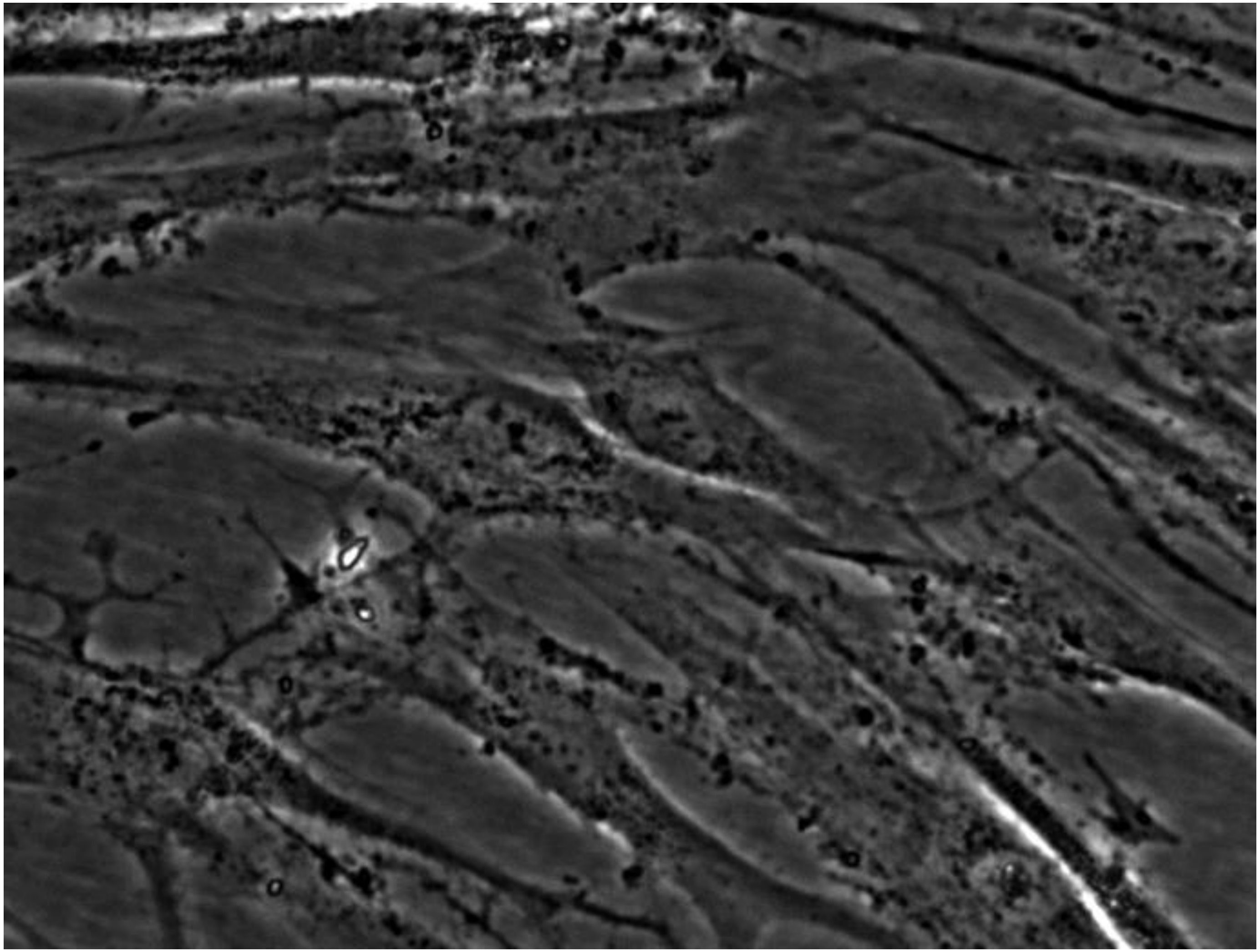
Lower expression compared to baseline

Spellman *et al Mol. Biol. Cell* 1998

The Cell Cycle

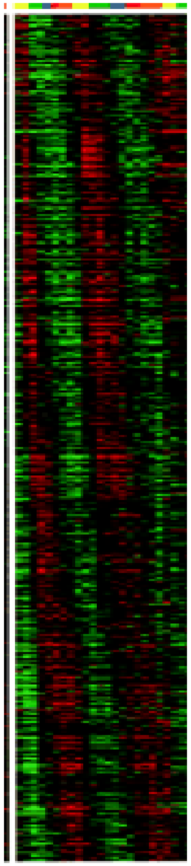
- The process in which cells divide.
- Plays key role in development and cancer.





Cell cycle expression: time line

- 1997, 1998 – budding yeast



Spellman *et al Mol. Biol. Cell* 1998

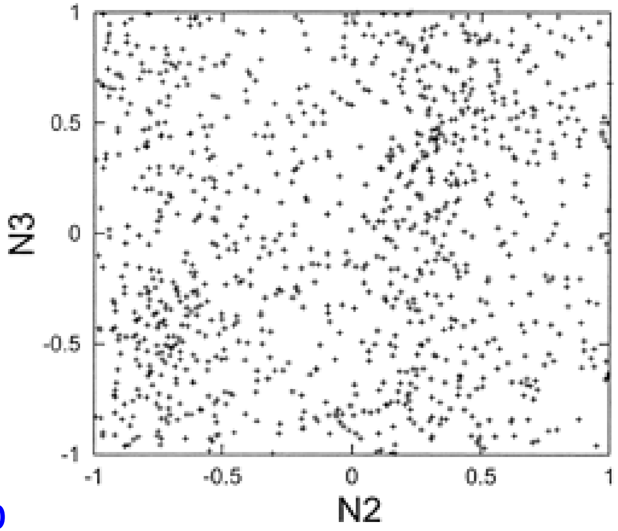
Cell cycle expression: time line

- 1997, 1998 – budding yeast
- 2000 - bacteria
- 2000 – plants
- 1999, 2000 - human
- 2001 – mouse

Cho *et al*, *Nature Genetics* 2000

Cell cycle expression: time line

- 1997, 1998 – budding yeast
- 2000 - bacteria
- 2000 – plants
- 1999, 2000 - human
- 2001 – mouse
- 2002 – human data is noise!



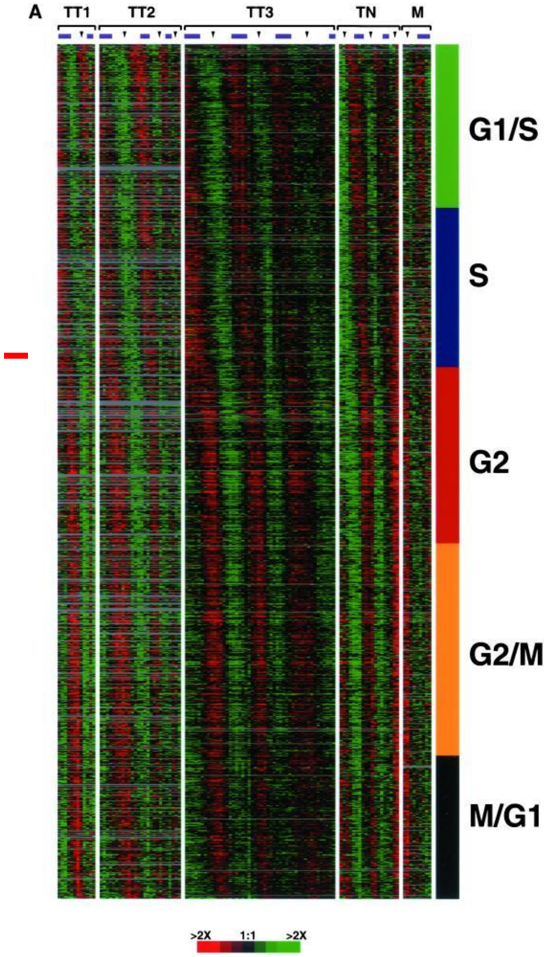
Reproducibility of peak between two repeats

Shedden & Cooper, PNAS, 2002

Cell cycle expression: time line

- 1997, 1998 – budding yeast
- 2000 - bacteria
- 2000 – plants
- 1999, 2000 - human
- 2001 – mouse
- 2002 – human data is noise!
- 2002 – Cancer cell cycle expression

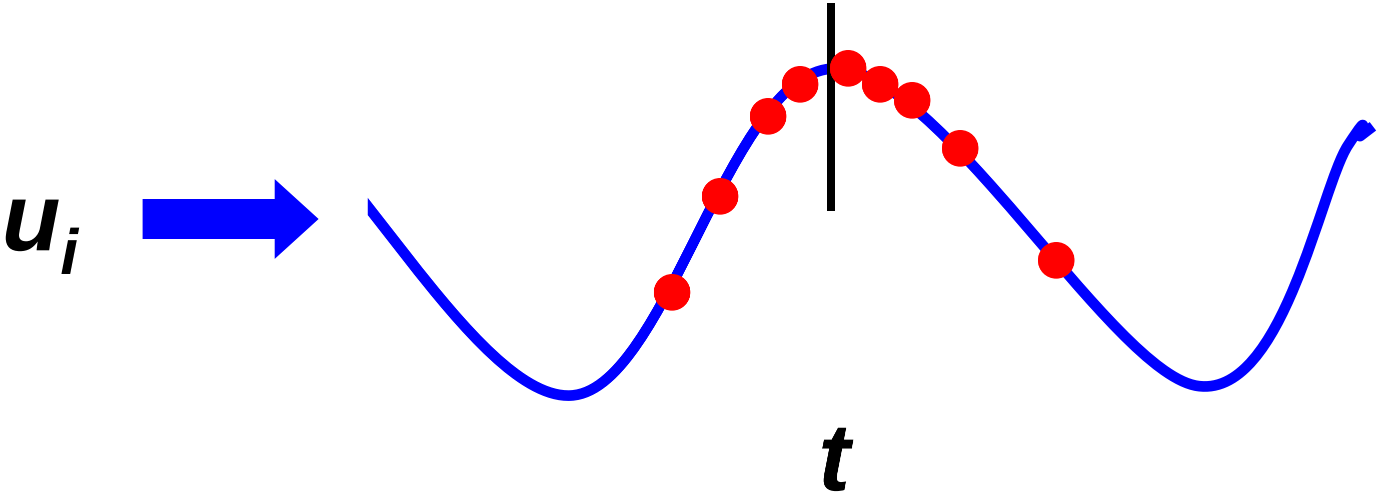
Can we compare cancer and normal expression programs?



Whitfield et al, *MBC*, 2002

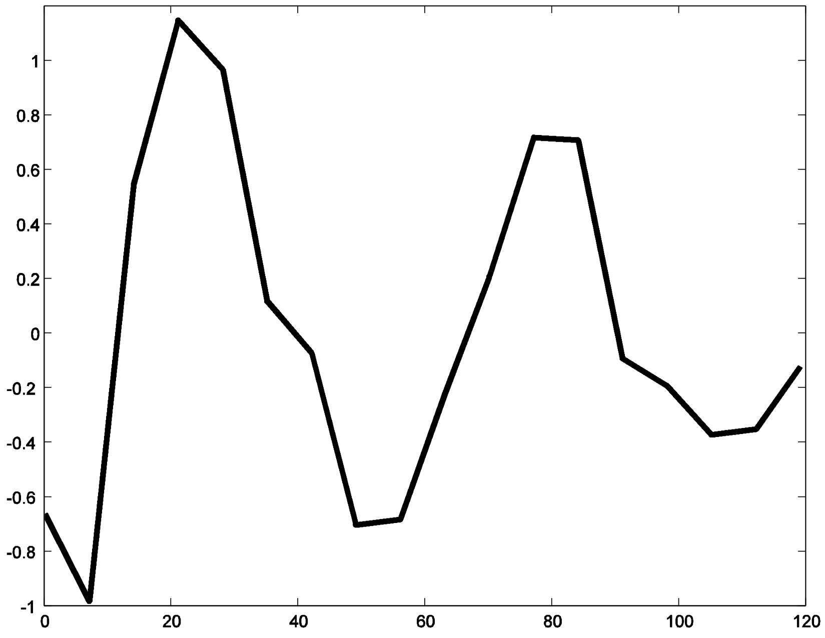
Main problem: Population effects

- Microarray experiments profile population of cells.
- Cells are artificially synchronized, not all cells are arrested.
- Even for those that are, synchronization is lost over time.

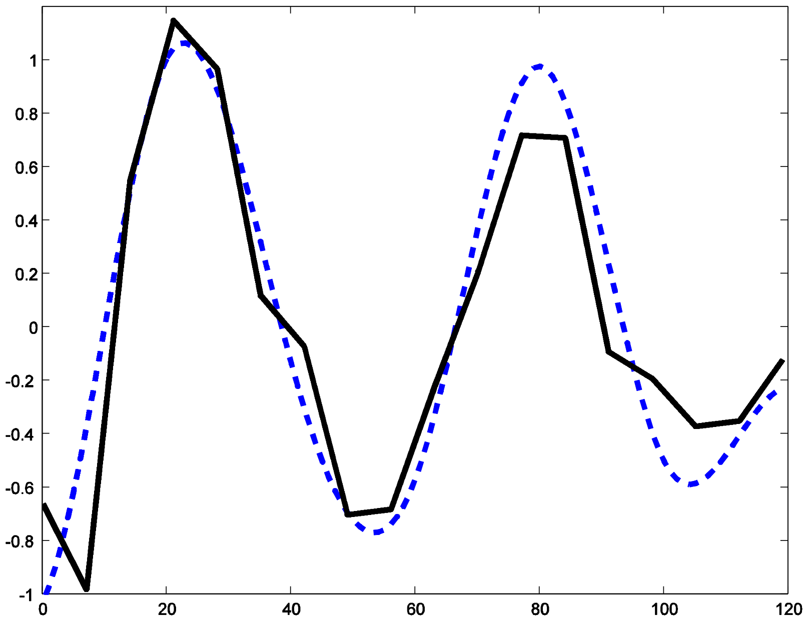


Synchronization

Smc3: observed values



Smc3: reconstructed values

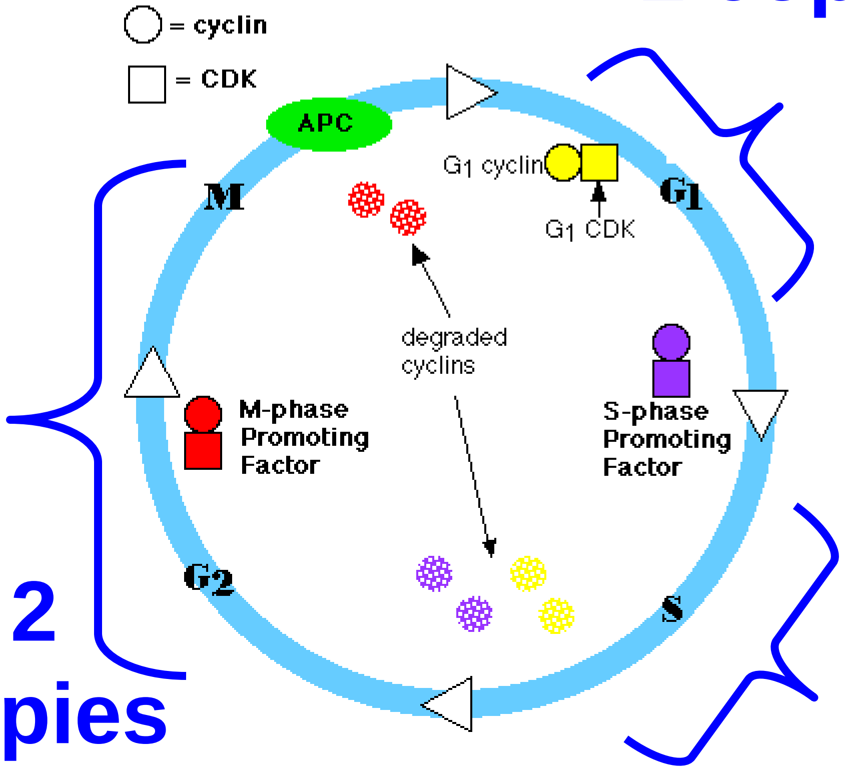


A major problem with human data (less than one cycle is synchronized)

Complementing expression data

- The main problem we face are population effects.

- While mRNAs cannot be measured on a single cell basis, bigger molecules, like DNA, can.



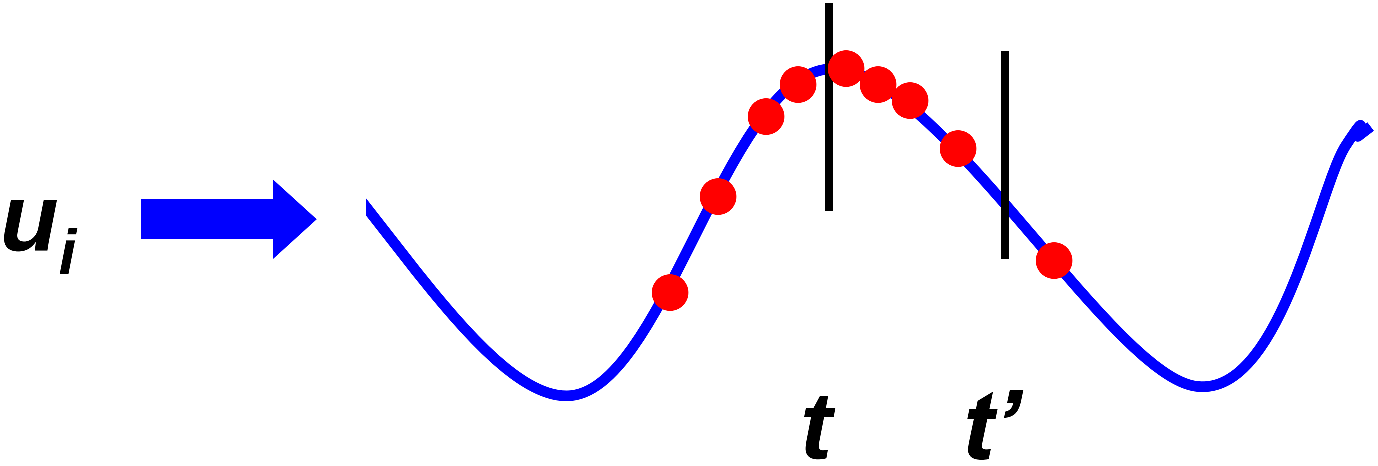
1 copy

2 copies

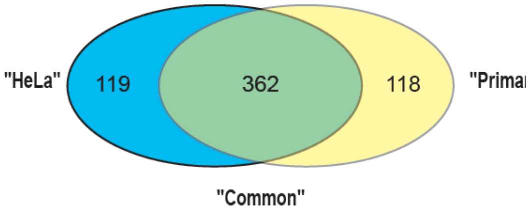
1-2 copies

Data integration to overcome synchronization loss

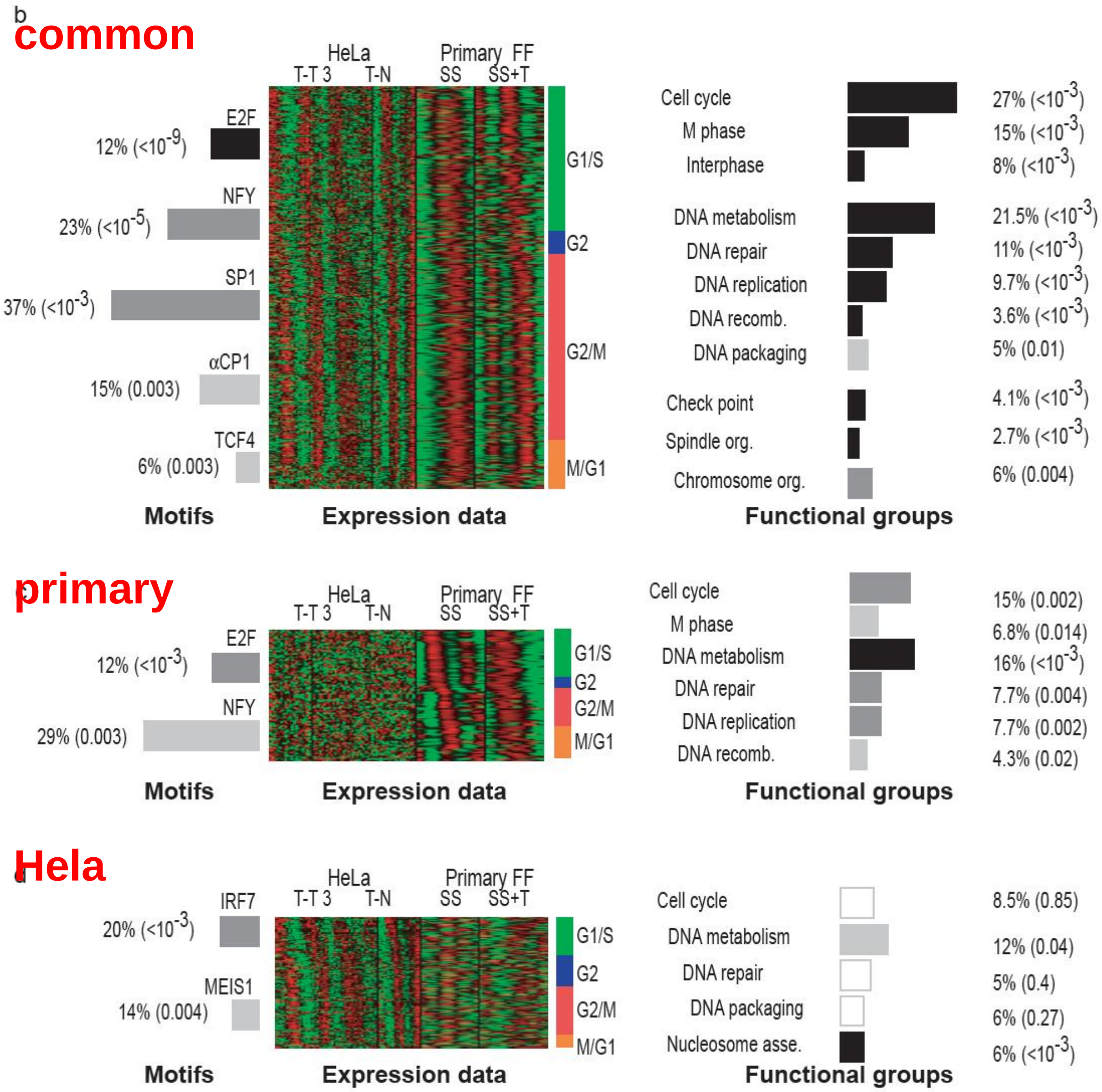
- We learn a synchronization loss model from independent measurements
- Using this model we estimate the proportion of cells at time ***t'*** when the real time is **t**
- We re-distribute the values measured for each gene according to the number of cells at this time



Results for human expression data



Validation by PCR


common

primary

Hela

Time line

- 1997, 1998 – budding yeast cell cycle expression
- 2000 – plants
- 1999, 2000 - human
- 2001 – mouse

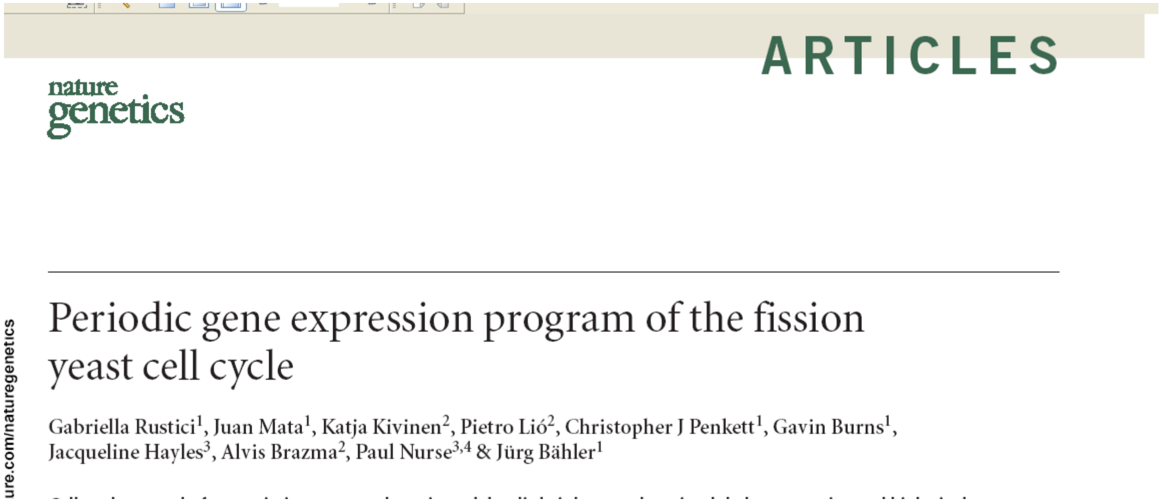
- 2002 – human data is noise !
- 2002 – cancer cell cycle expression (approximation)
- 2004, 2005 – deconvolution and Checksum
- 2008 – human cell cycle data

Time line

- 1997, 1998 – budding yeast cell cycle expression
- 2000 – plants
- 1999, 2000 - human
- 2001 – mouse

- 2002 – human data is noise !
- 2002 – cancer cell cycle expression (approximation)
- 2004, 2005 – deconvolution and Checksum
- 2008 human cell cycle data

- 2004 – fission yeast cell cycle data



ARTICLES

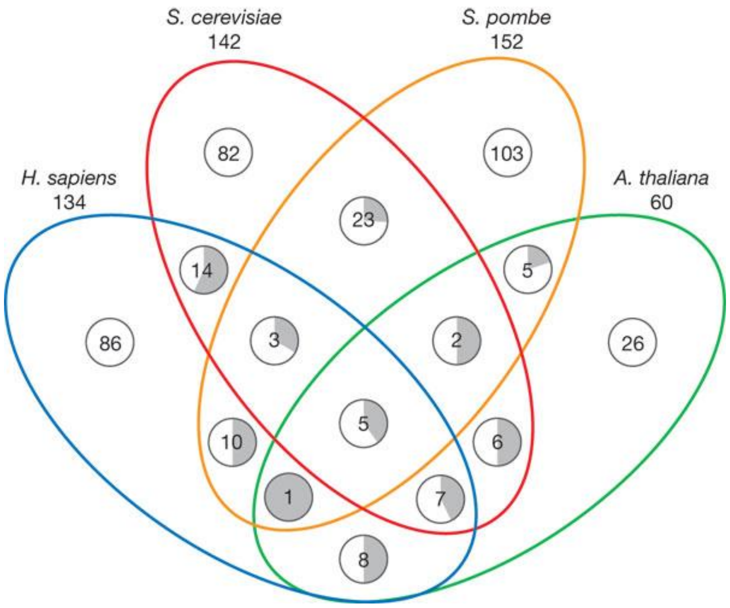
nature genetics

ure.com/naturegenetics

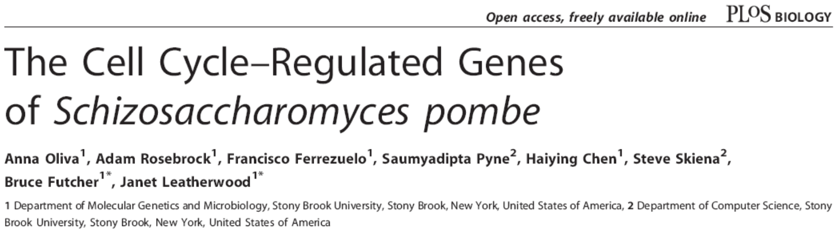
Periodic gene expression program of the fission yeast cell cycle

Gabriella Rustici[1], Juan Mata[1], Katja Kivinen[2], Pietro Lió[2], Christopher J Penkett[1], Gavin Burns[1], Jacqueline Hayles[3], Alvis Brazma[2], Paul Nurse[3,4] & Jürg Bähler[1]

“Our comparisons with budding yeast data revealed a surprisingly small core set of genes that are periodically expressed in both yeasts.”



Jensen et al *Nature* 2006



Open access, freely available online PLoS BIOLOGY

The Cell Cycle–Regulated Genes of *Schizosaccharomyces pombe*

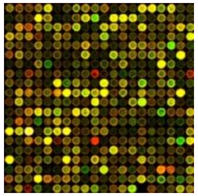
Anna Oliva[1], Adam Rosebrock[1], Francisco Ferrezuelo[1], Saumyadipta Pyne[2], Haiying Chen[1], Steve Skiena[2], Bruce Futcher[1*], Janet Leatherwood[1*]

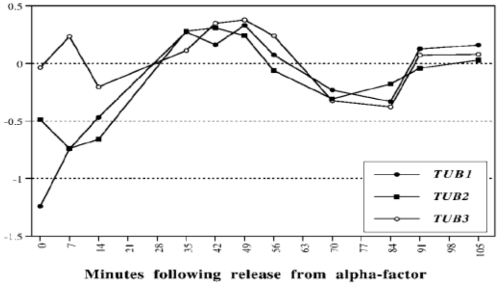
1 Department of Molecular Genetics and Microbiology, Stony Brook University, Stony Brook, New York, United States of America, 2 Department of Computer Science, Stony Brook University, Stony Brook, New York, United States of America

“Of our top 200 ranked cell cycle regulated genes, 72 (36%) had S. cerevisiae homologs that cycled”

From expression values to score

- Cells are *synchronized* to the same phase
- Microarray experiments at *multiple time points* after release from synchronization
- Scores derived from multiple expression time series
- Rank genes based on their scores, and use a *cutoff score* to identify cycling genes

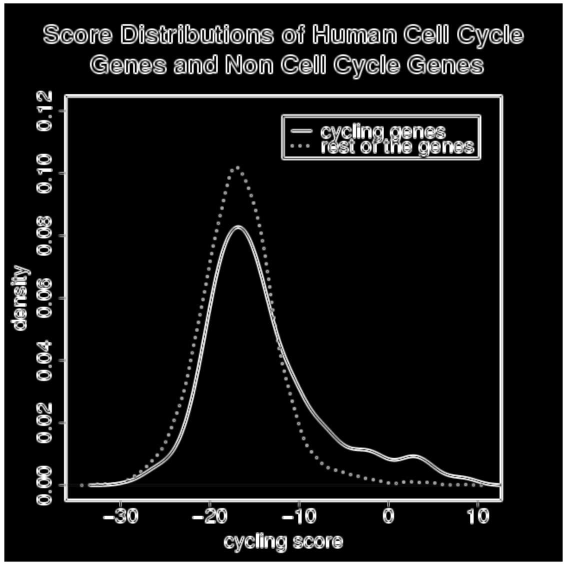




Spellman et al. (1998)

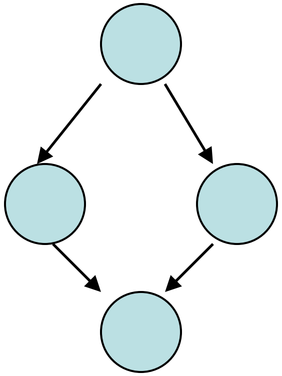
Problems

- Different scoring methods result in different lists
- Microarray data are noisy
- Hard to separate scores for cycling and non-cycling genes
    - Score distribution of cell cycle genes (derived from GO) versus the rest
        - solid curve: cycling genes
        - dotted curve: the rest



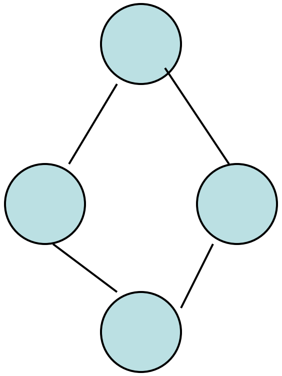
Graphical models

- Efficient way to represent and reason about *joint distributions*
- Graphs in which nodes represent random variables and edges correspond to dependency assumptions
- Two major types: Directed and undirected



$$\prod_i p[x_i \mid Pa(x_i)]$$

- Bayesian networks
- Hidden Markov models



$$\prod_{i,j} \psi_{i,j}(x_i, x_j)$$

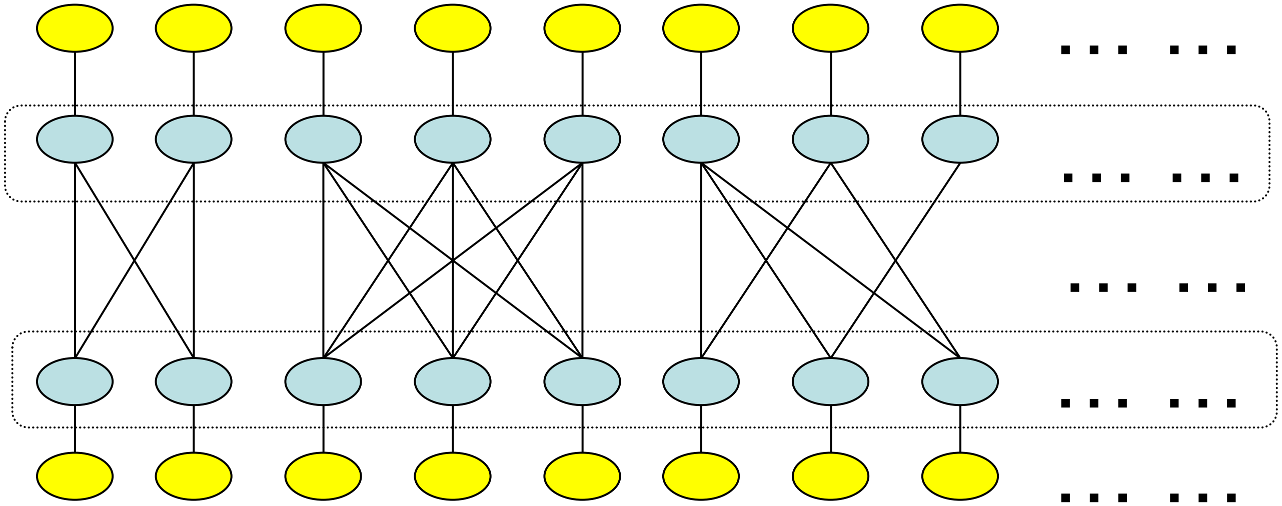
- Markov random fields

Graphical models (cont)

- Parameters are used to specify the conditional probability distribution (directed graphs) or the potential functions (undirected graphs)
- Computational questions:

  - Determining the structure of the model (sometimes)
  - Estimating the parameters of the model
  - Inference

Probabilistic graphical model for combining expression and sequence homology

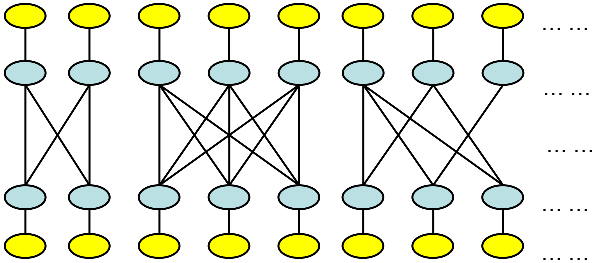


🔵 : C_i : Cycling Status Nodes (unobserved)

🟡 : S_i : Score Nodes (observed) —— : Encodes Dependency Relations

Numeric summary of expression time series

Likelihood of the model



- Node Potential:
- Edge Potential:

$$\psi_i(C_i) = Pr(C_i|S_i)$$

$$\psi_{ij}(C_i, C_j) = 2^{-\lambda w_{ij}(C_i - C_j)^2}$$

need to be learned from data

controls contribution from each source

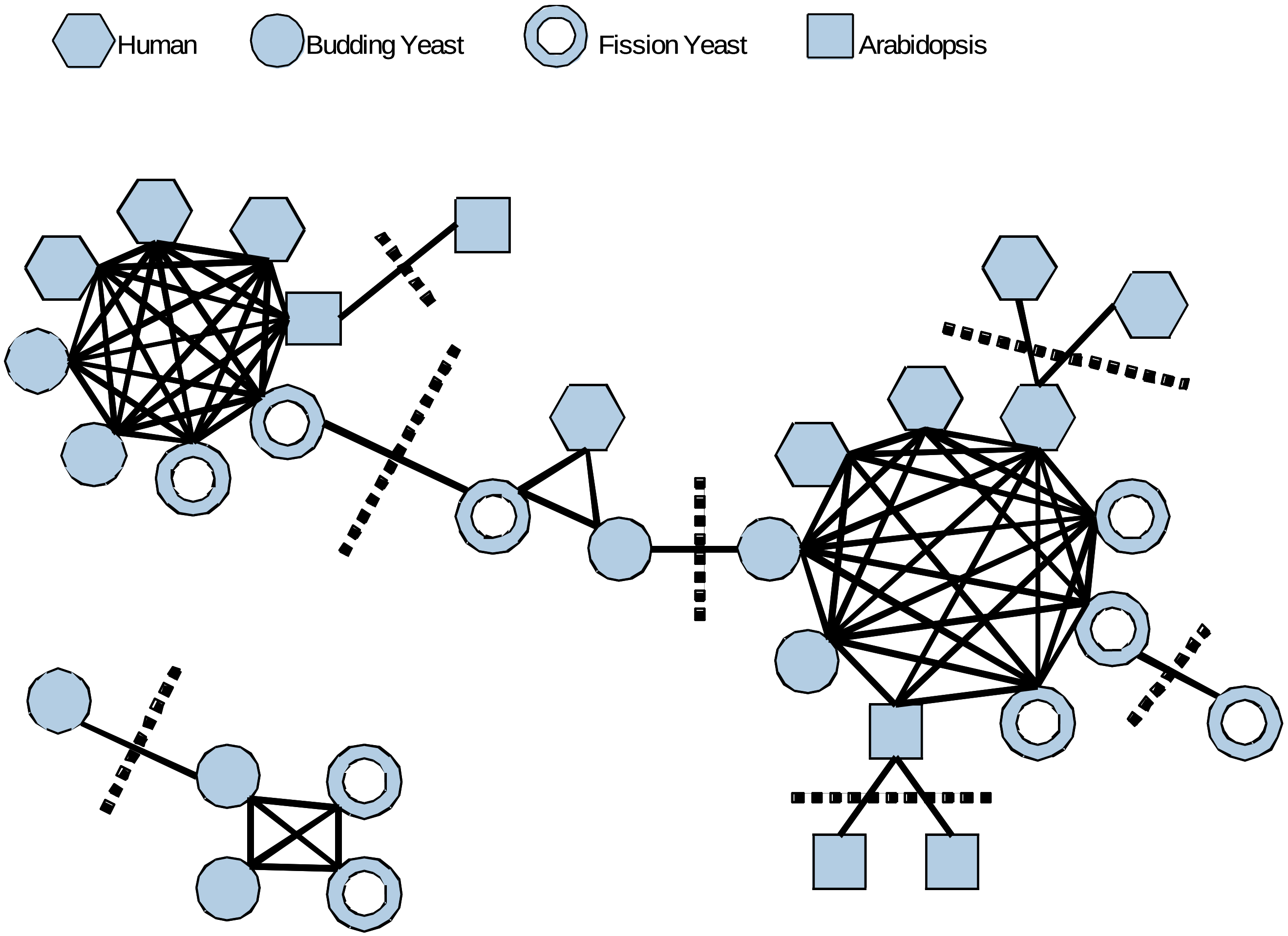
weight (from homology)

- Joint probability distribution

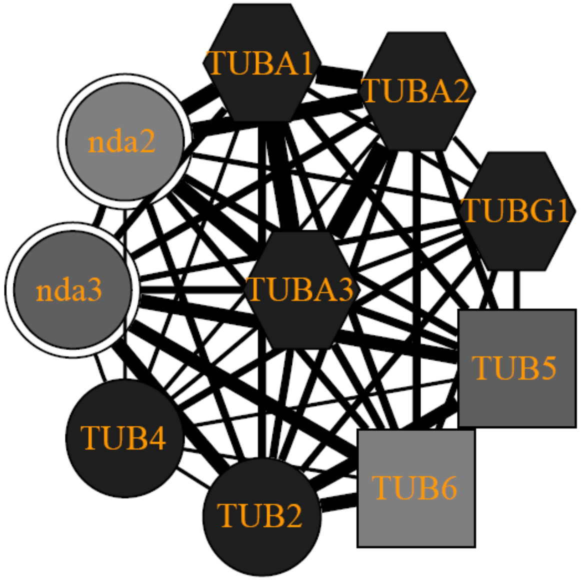
$$L = \frac{1}{Z} \prod_i \psi_i(C_i) \prod_{i,j} \psi_{ij}(C_i, C_j)$$

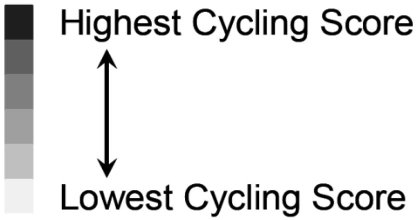
Once scores are assigned: Identify conserved genes

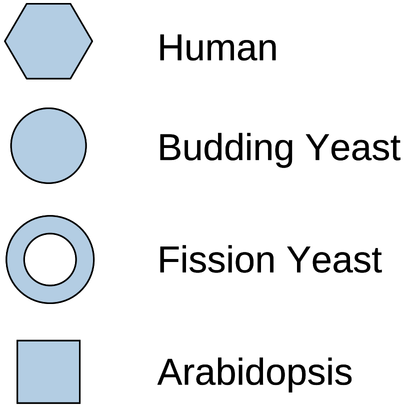
- A cutoff is selected
- All genes with scores below cutoff are removed
- The rest of the genes are grouped into homologues cliques



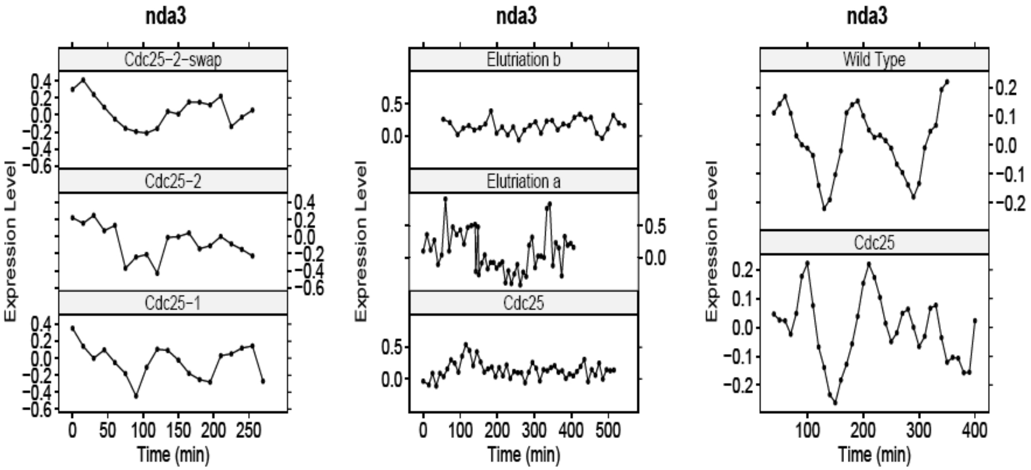
Analysis One Result: Clique of cycling genes from multiple species





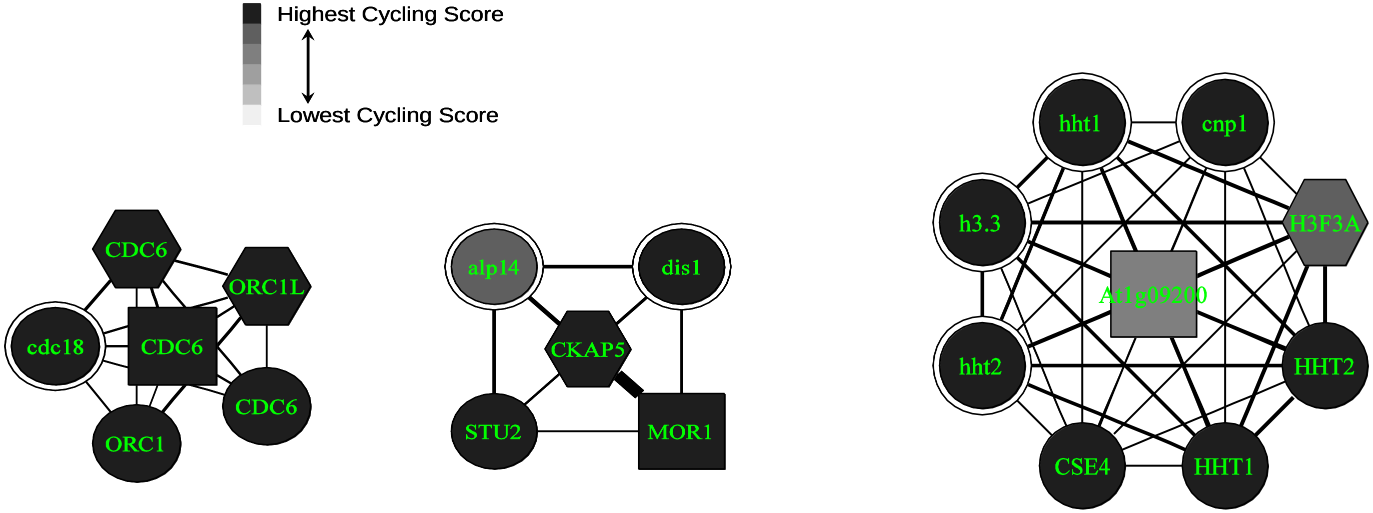


- Nda3 is a fission yeast gene required for chromosome separation
- It would have been missed due to noise in some of the expression datasets

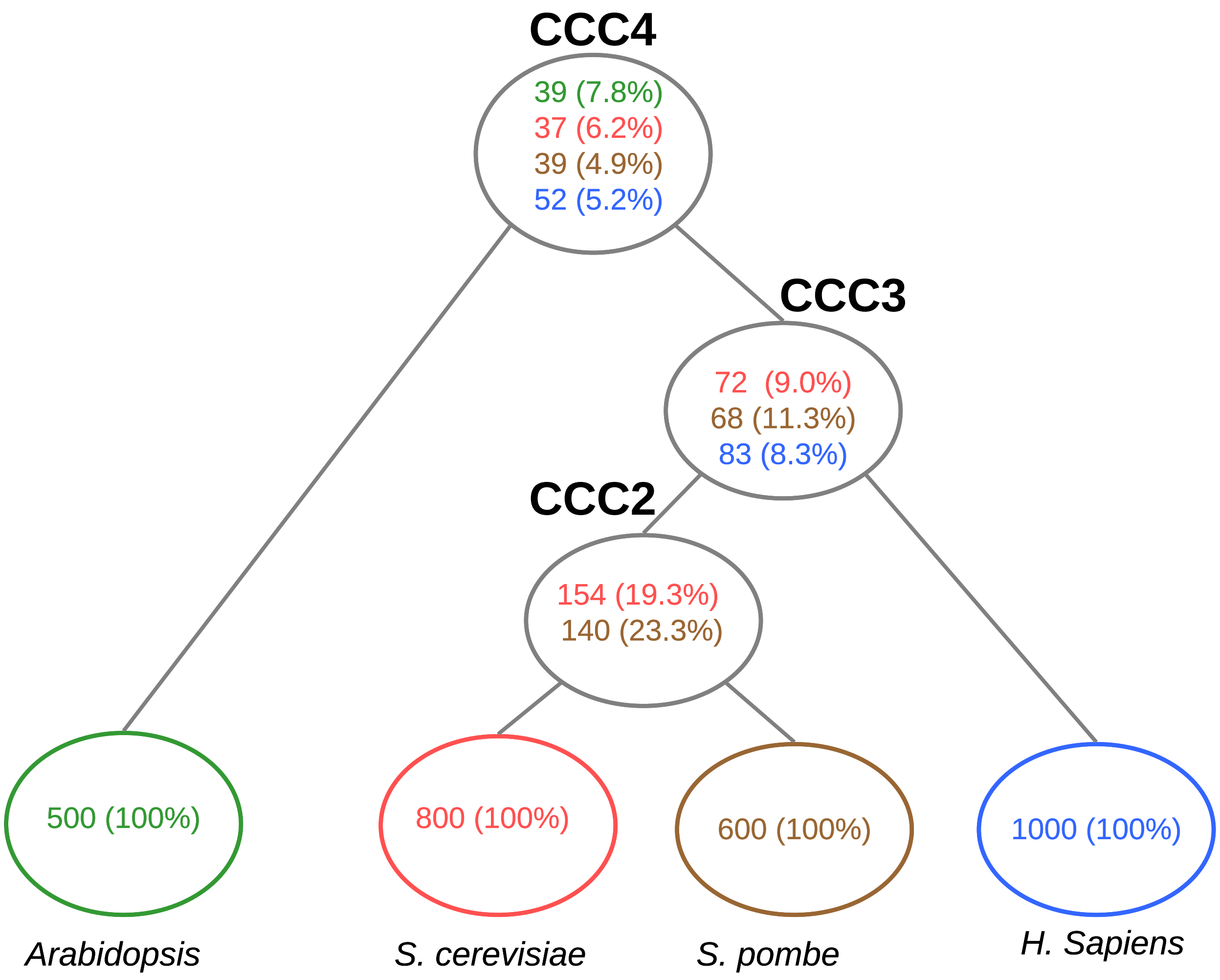


Resulting conserved cliques

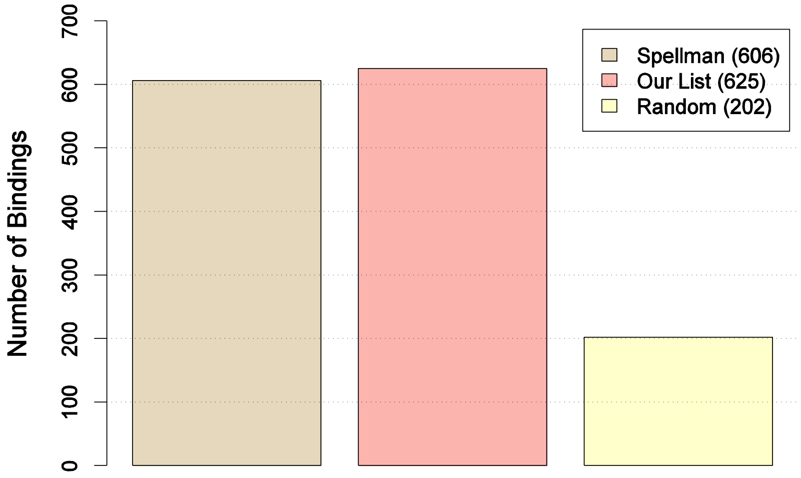
⬡ Human ○ Budding Yeast ◎ Fission Yeast □ Arabidopsis

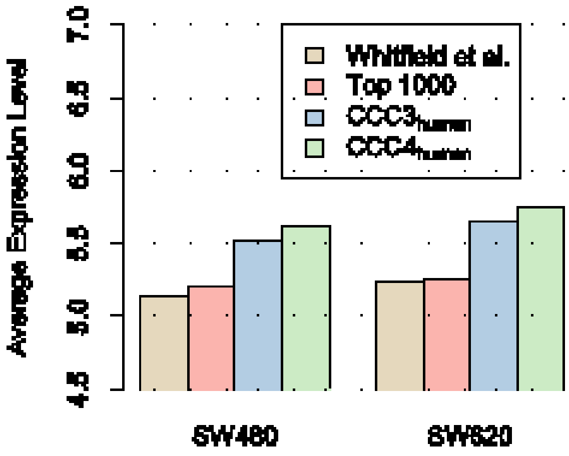


Overall conservation



Analysis using complementary high throughput datasets

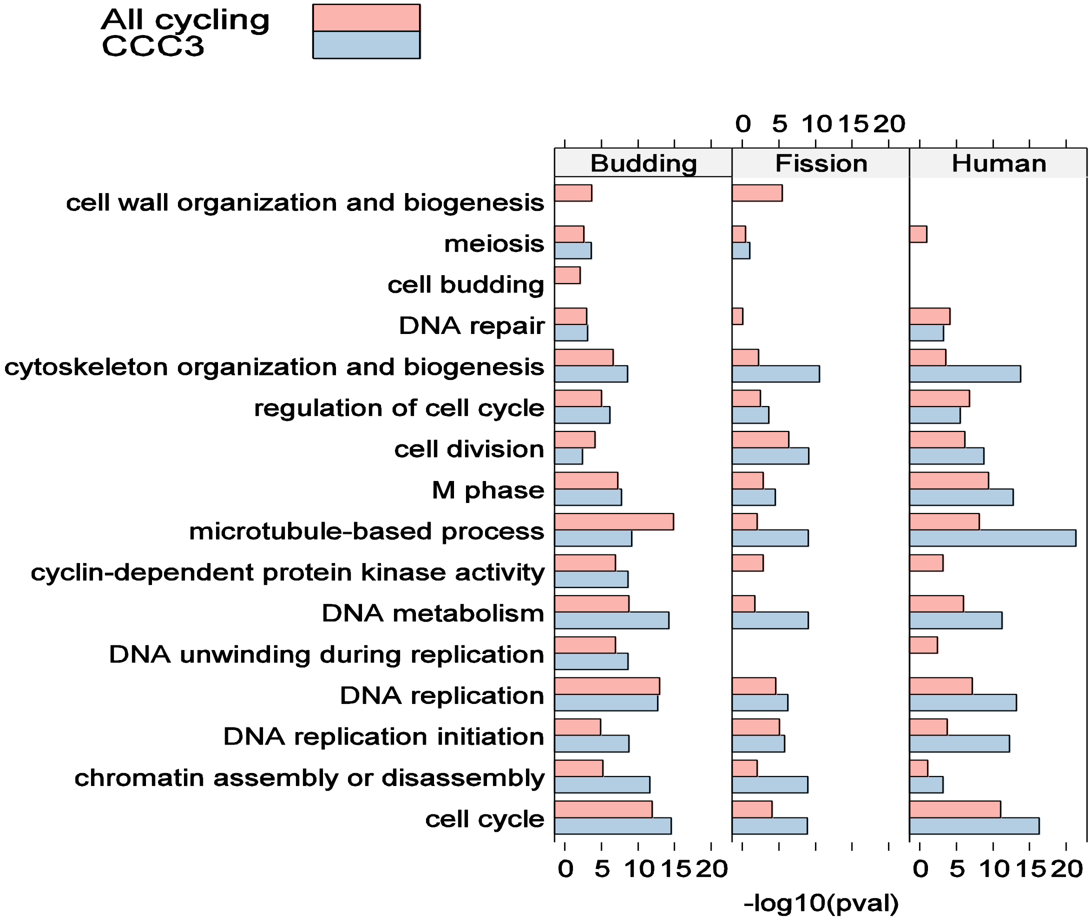

Budding Yeast Genes Bound by Cell Cycle Transcription Factors

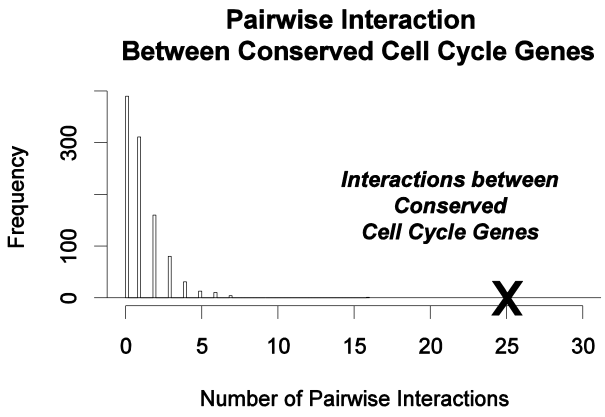

Human Gene Expression In Colon Cancer Cells

Motif analysis of conserved cycling genes

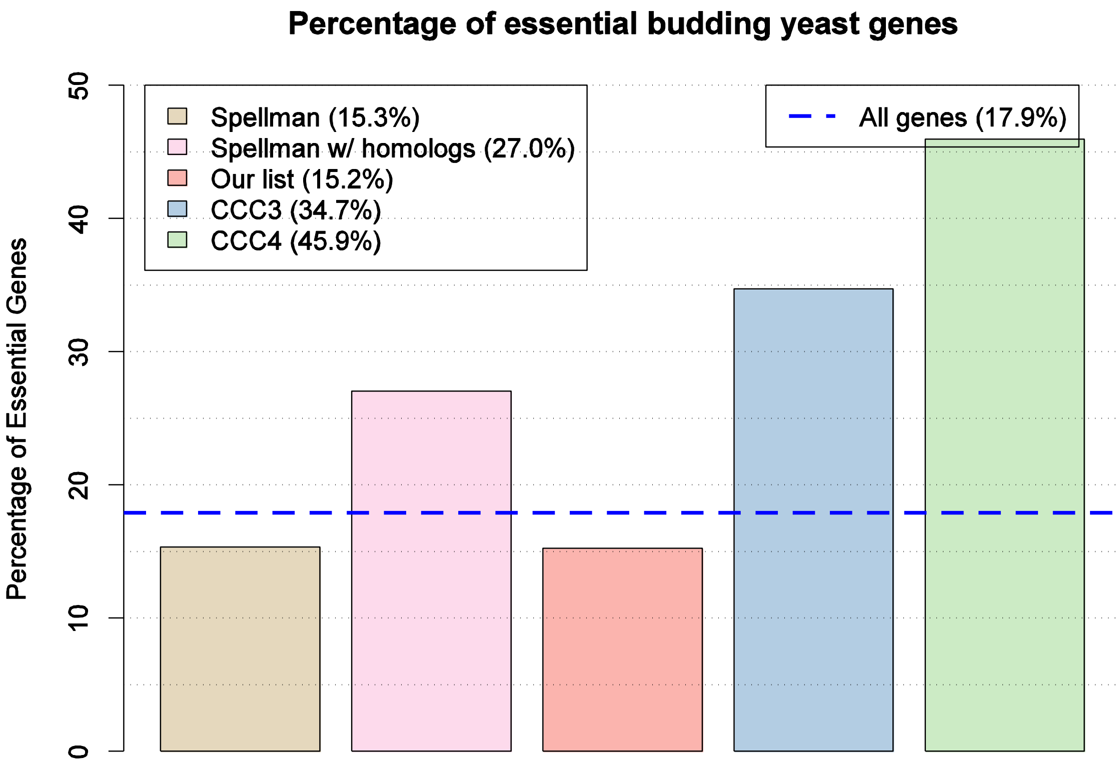
Budding yeast phase	**Transcription factor**	***fission yeast* cell-cycle genes**		**Negative control (*fission yeast* non cell-cycle genes)**		**Positive control (conserved budding yeast cell-cycle genes)**		**Extended positive control (all budding yeast CC genes)**	
G1/S	**SWI4**	4	43%	0	0%	4	96%	4	98%
	SWI6	4	97%	0	0%	4	100%	4	83%
	MBP1	4	59%	0	0%	4	93%	4	91%
G2/M	**FKH1**	0	0%	2	22%	1	62%	3	67%
	FKH2	2	45%	2	24%	1	74%	2	67%
	NDD1	0	0%	0	0%	4	100%	4	100%
M/G1	**MCM1^**	0	0%	0	0%	3	87%	4	88%
	ACE2	4*	86%	0	0%	0*	0%	4	88%
	SWI5	~2*	100%	0	0%	~2*	75%	1	0%

Functional analysis

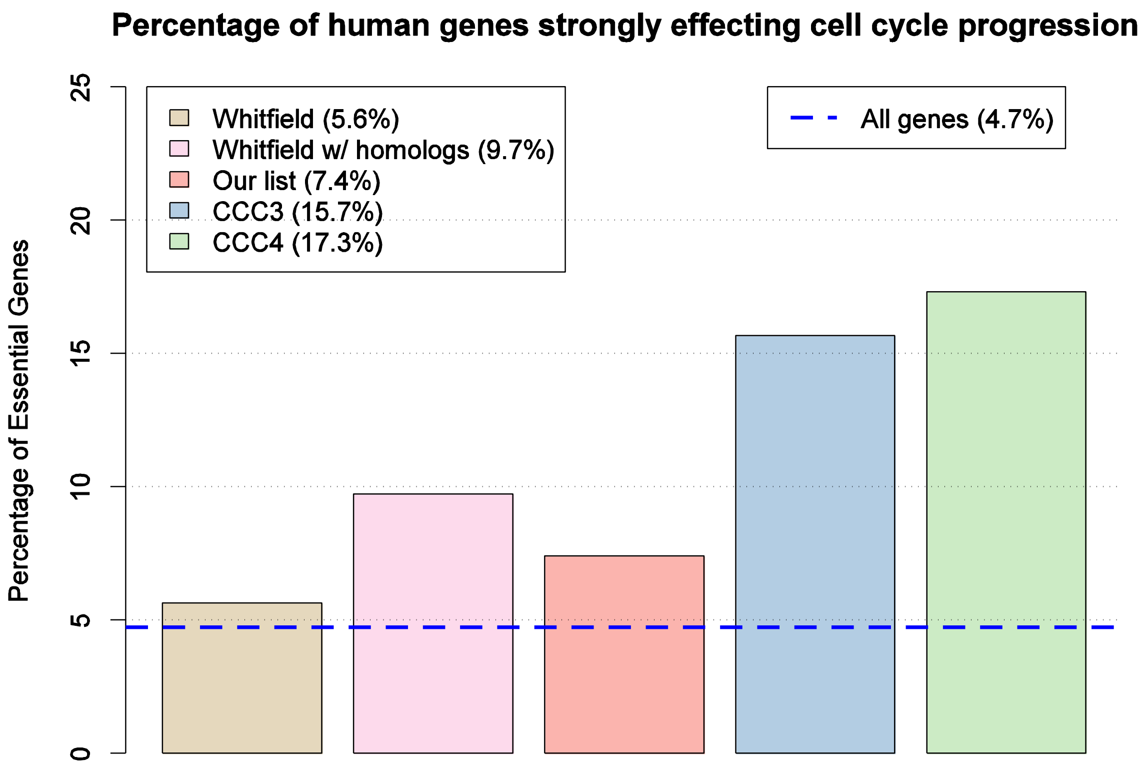




Importance of conserved sets



Similar analysis for human cells using RNAi data



What you should know

- Comparing expression experiments across species is usually harder than comparing sequence data:

  - Different time scales
  - Different conditions
  - Not necessarily one to one orthology matches

- Need methods that can overcome noise and support ‘soft’ cutoff for such comparisons
- When applied, such methods can idnetify conserved patterns which are missed by list comparison methods which are based on a species specific cutoff.