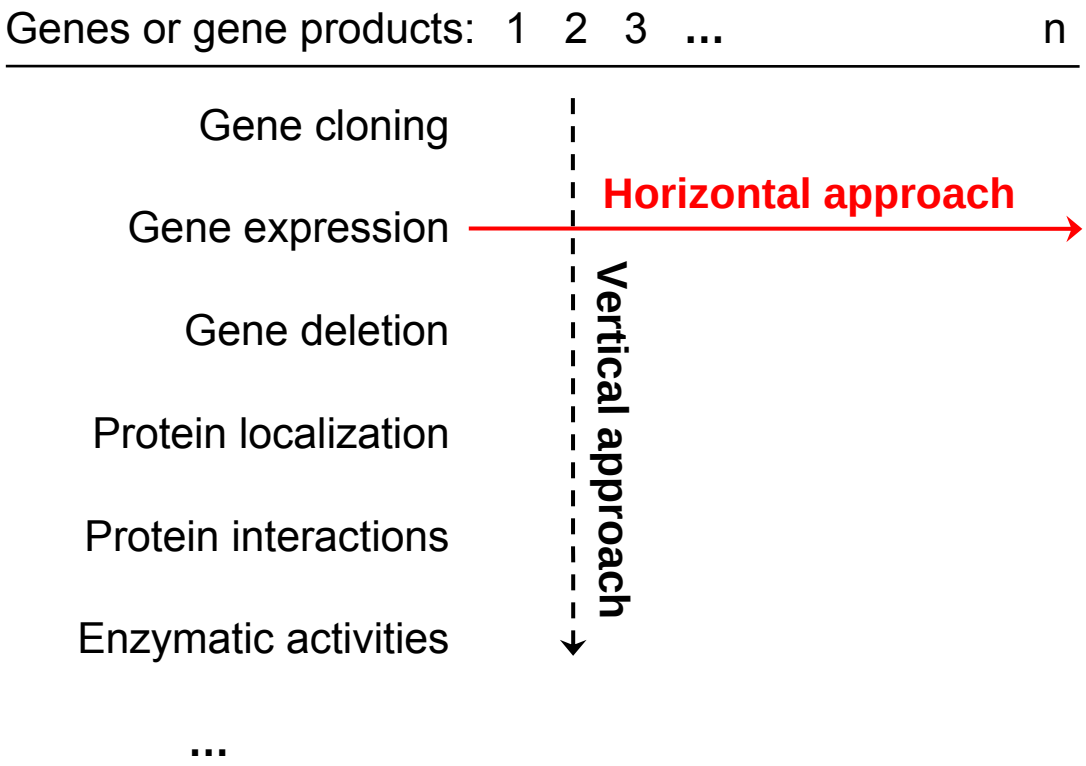
1. Analysis and mining of large-scale microarray data sets

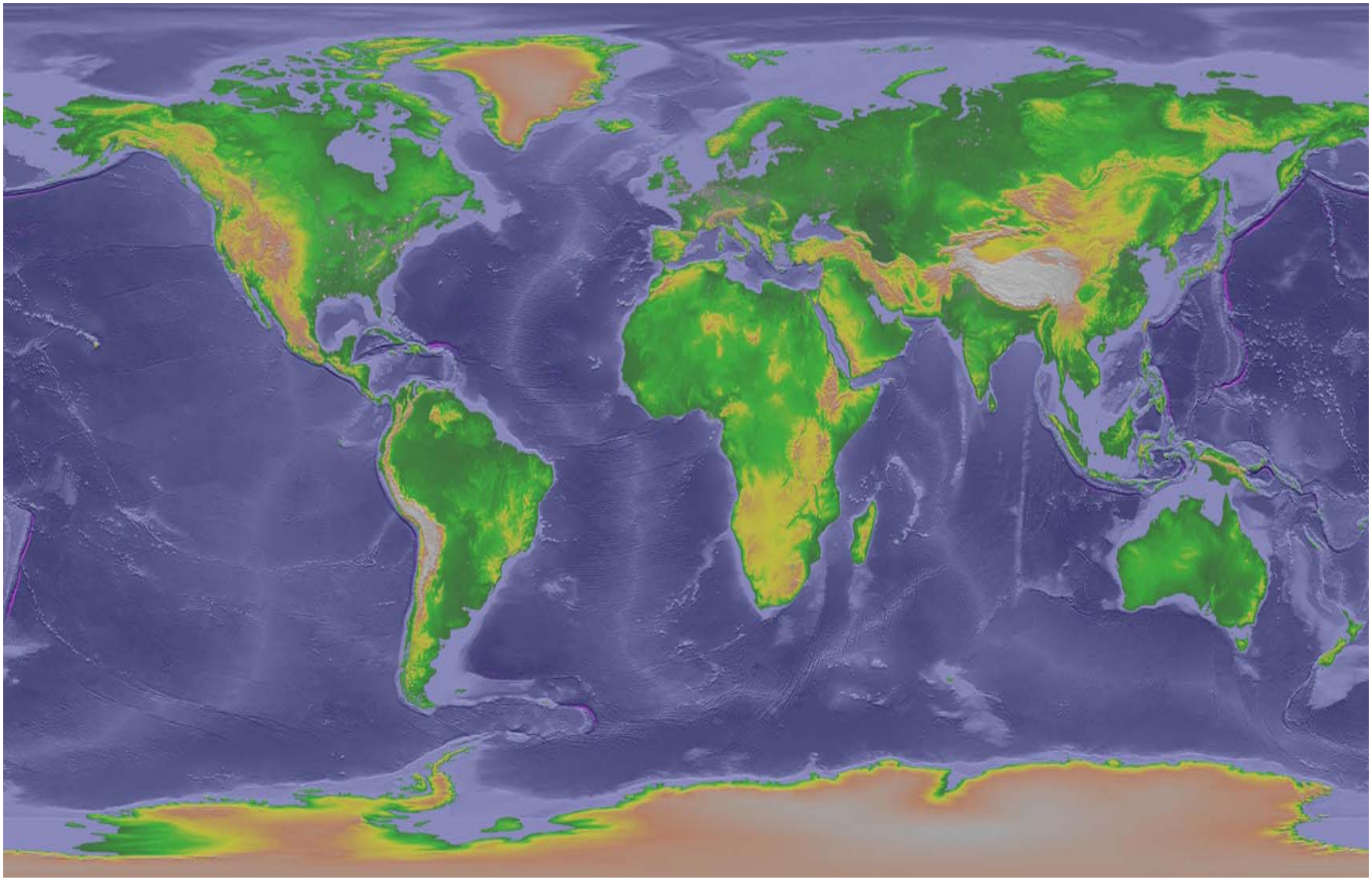
2. Global Gene Expression Programs in Fission Yeast

Slides available at:
http://www.sanger.ac.uk/PostGenomics/S_pombe/

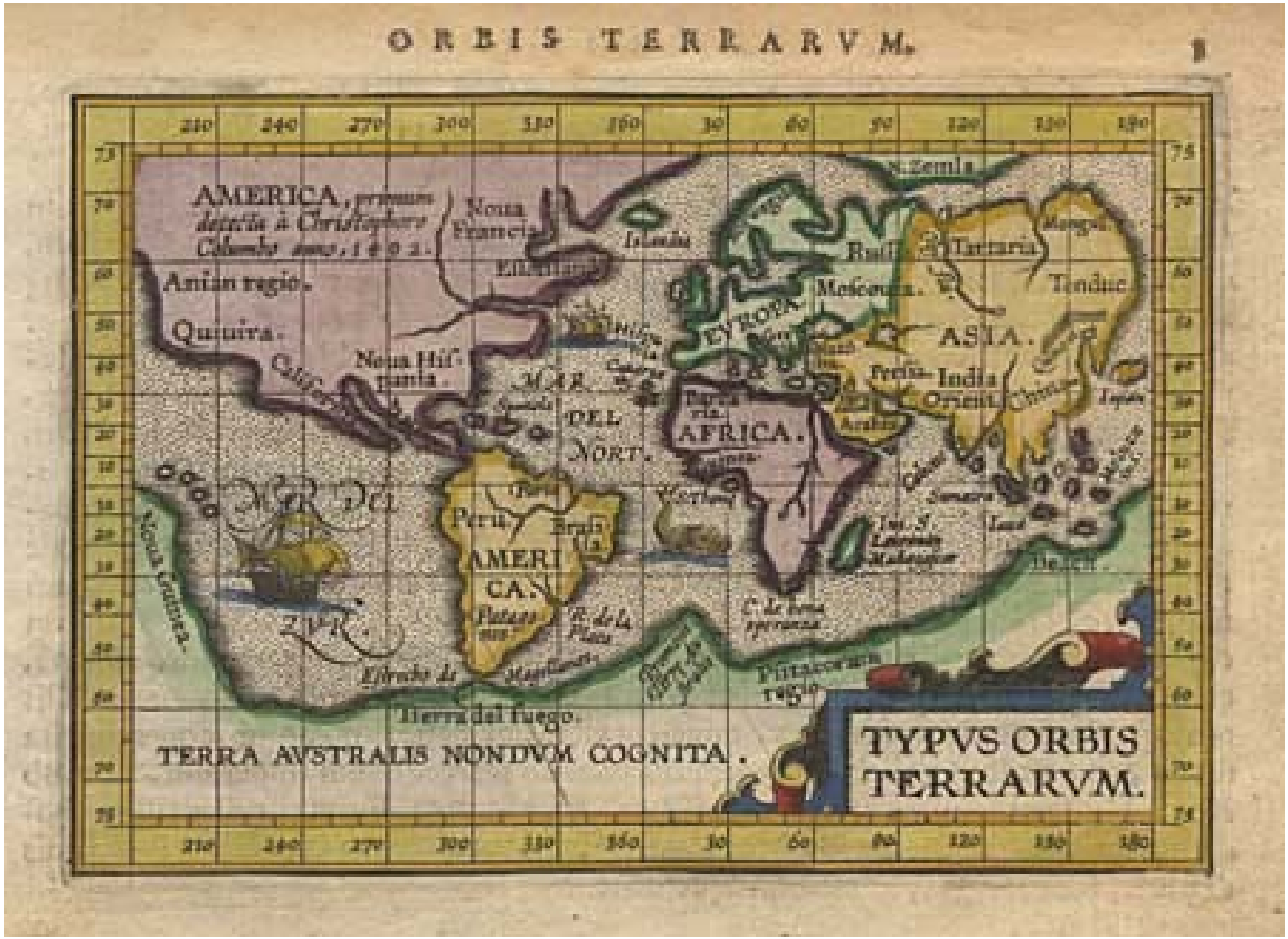
Post-genomic vs traditional experiments:



The global view…



Global map as an approximation of reality:



Gene expression data

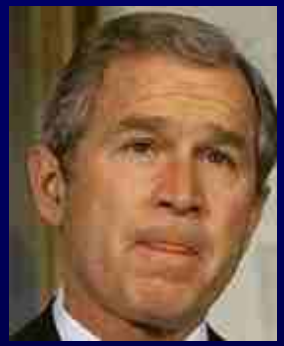
Evolution has carried out the perfect functional genomic experiment:

a multi-million year, genome-wide conditional knock-out mutagenesis, accompanied by exhaustive phenotypic screens.

The results are recorded in gene expression programs activated under various conditions.

Microarrays: Principle of Differential Hybridization

Reference



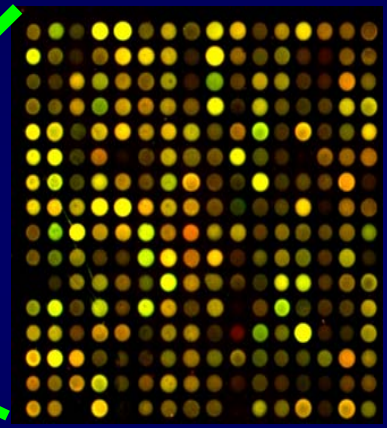
Test

RNA isolation

Fluorescently labeled probes

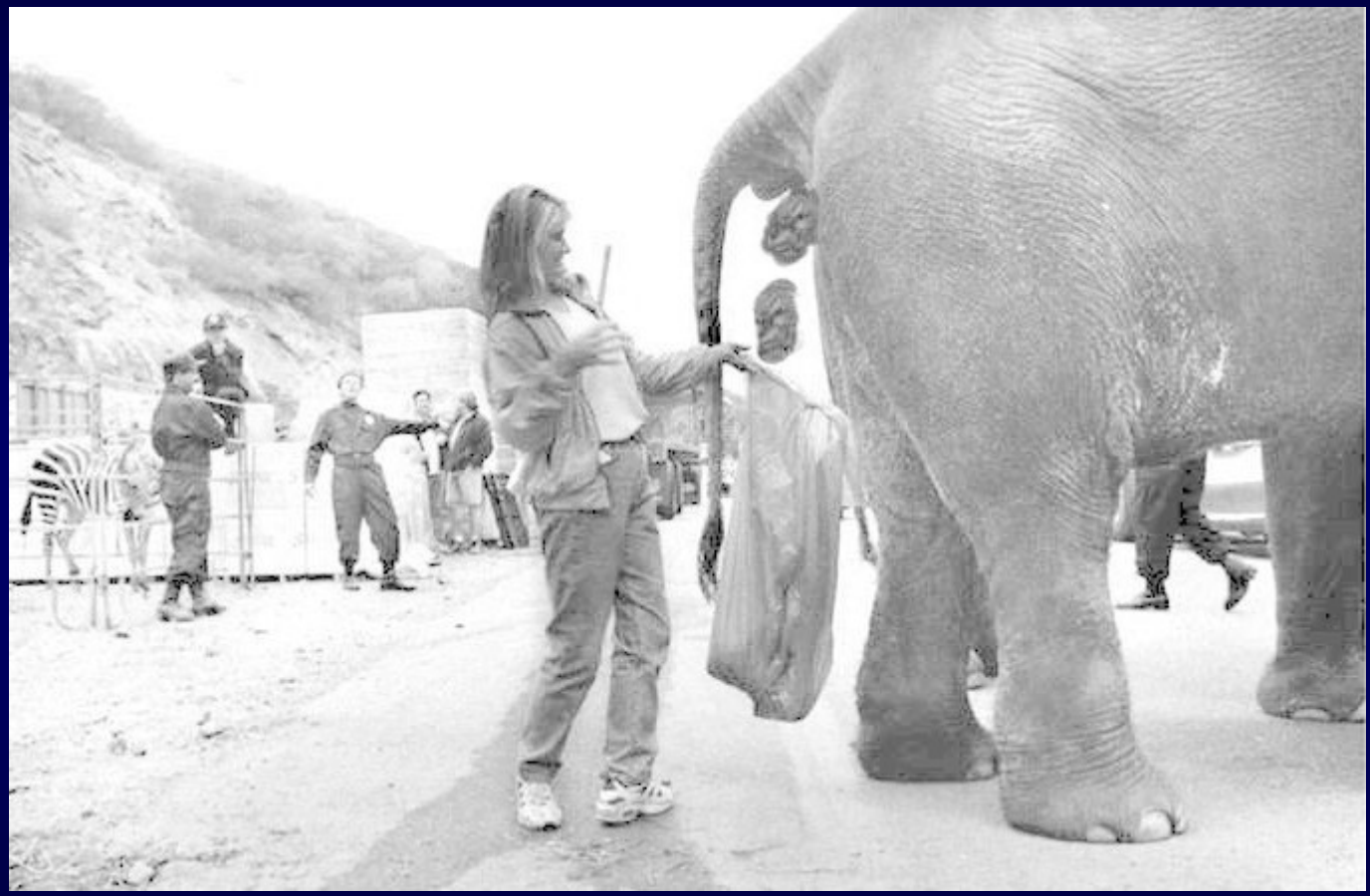
Hybridize
Wash

Measure fluorescence signal ratios red/green



Analyze data to identify patterns of gene expression

Data Collection



Data evaluation

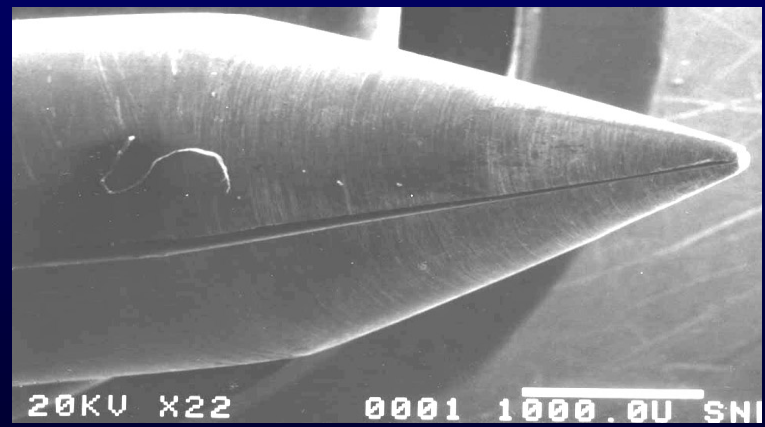
Programming and curation support

Arrayer

Biorobotics MicroGrid II

- 48 pins
- 200 μm spots, 250 μm spacing
- >20,000 spots/slide
- 100 slide capacity

Split pin used for printing



Scanner

Genepix 4000B (Axon)

- 5µm resolution
- simultaneous dual laser scanning
- GenePix analysis software

Data Processing Pipeline

1 Image Analysis Software: determine signal ratios (GenePix)

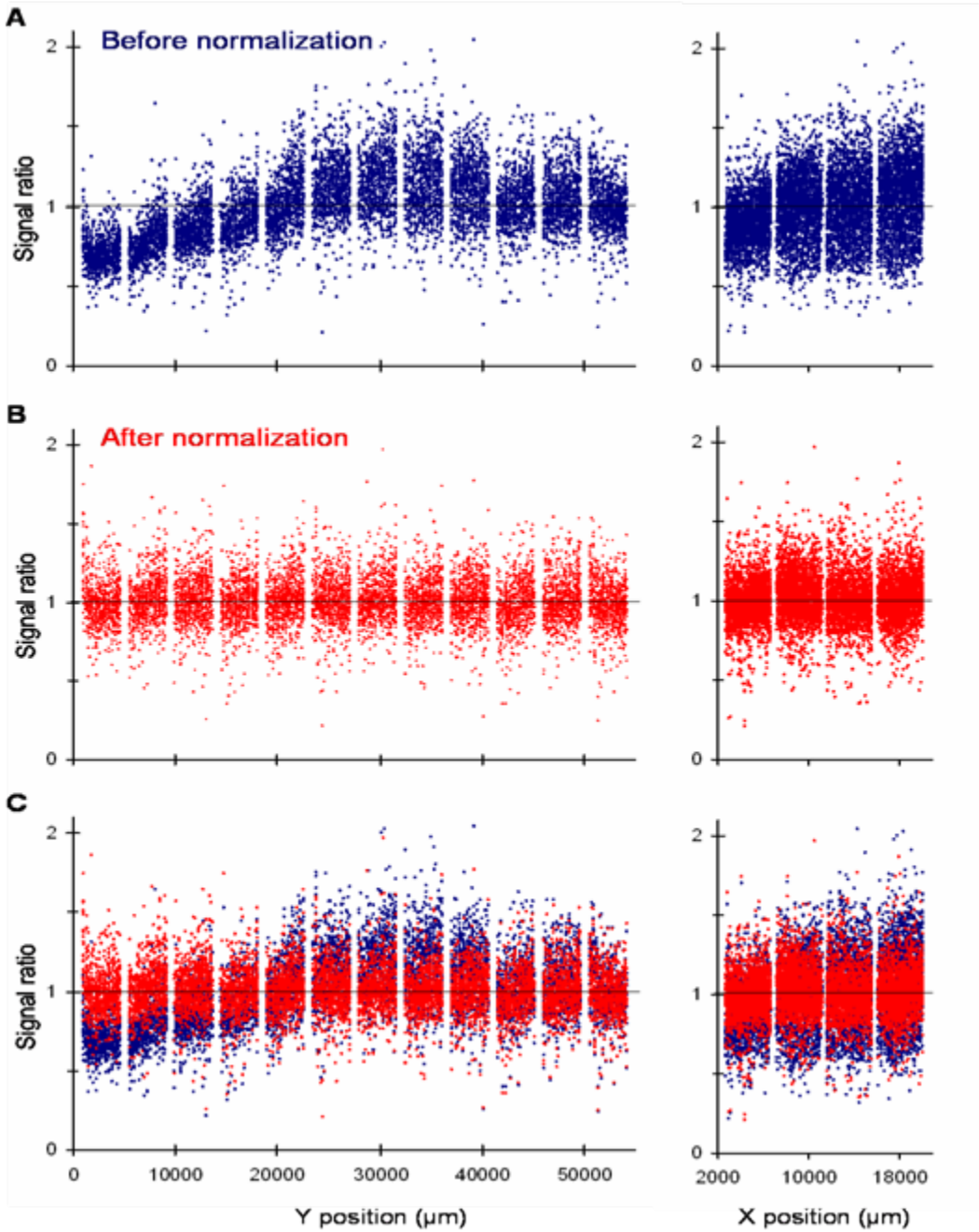
2 InHouse program for initial data processing: filter weak and irreproducible signals, local normalization, quality control

3 Data mining using various software (GeneSpring, R/Bioconductor, SAM, …)

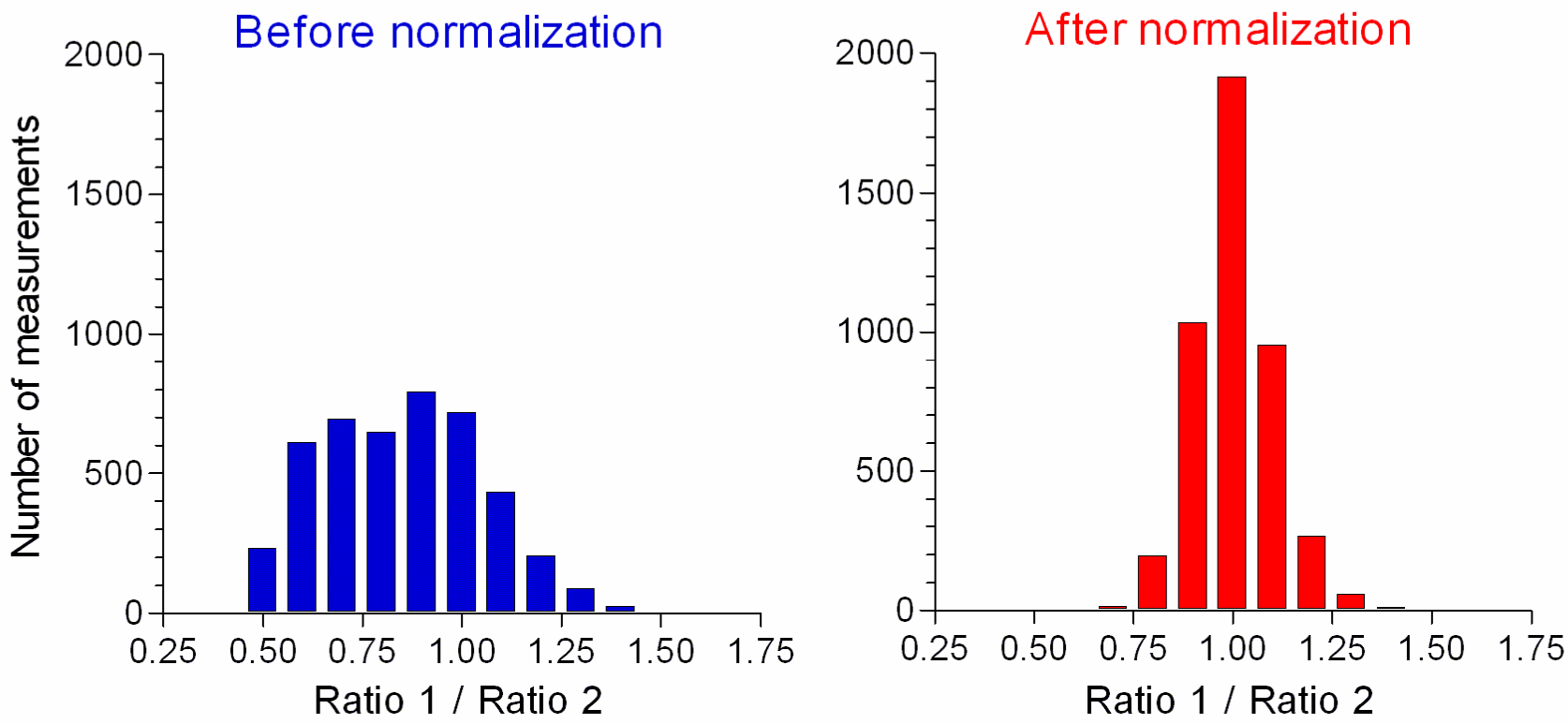
4 Public Database – ArrayExpress

Local Normalization:

Running window
1000 spots

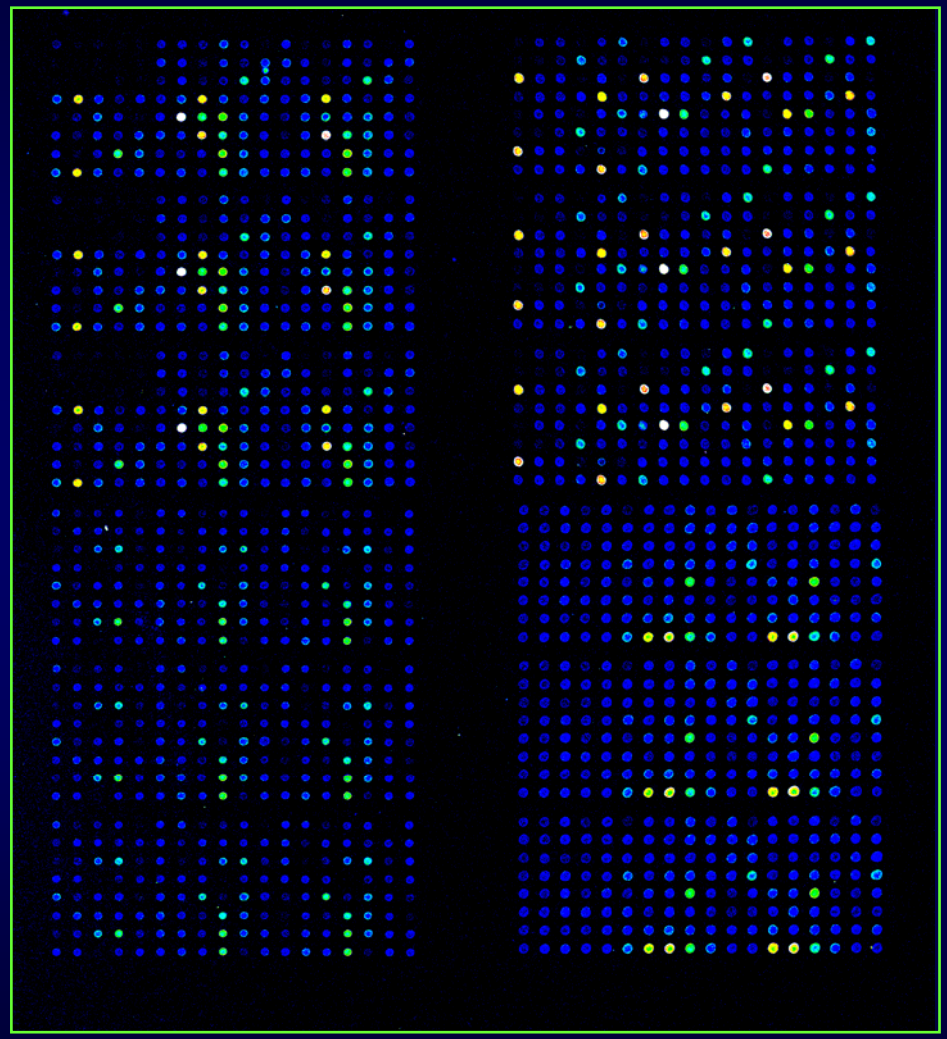


Replicate data are more similar to each other after normalization:

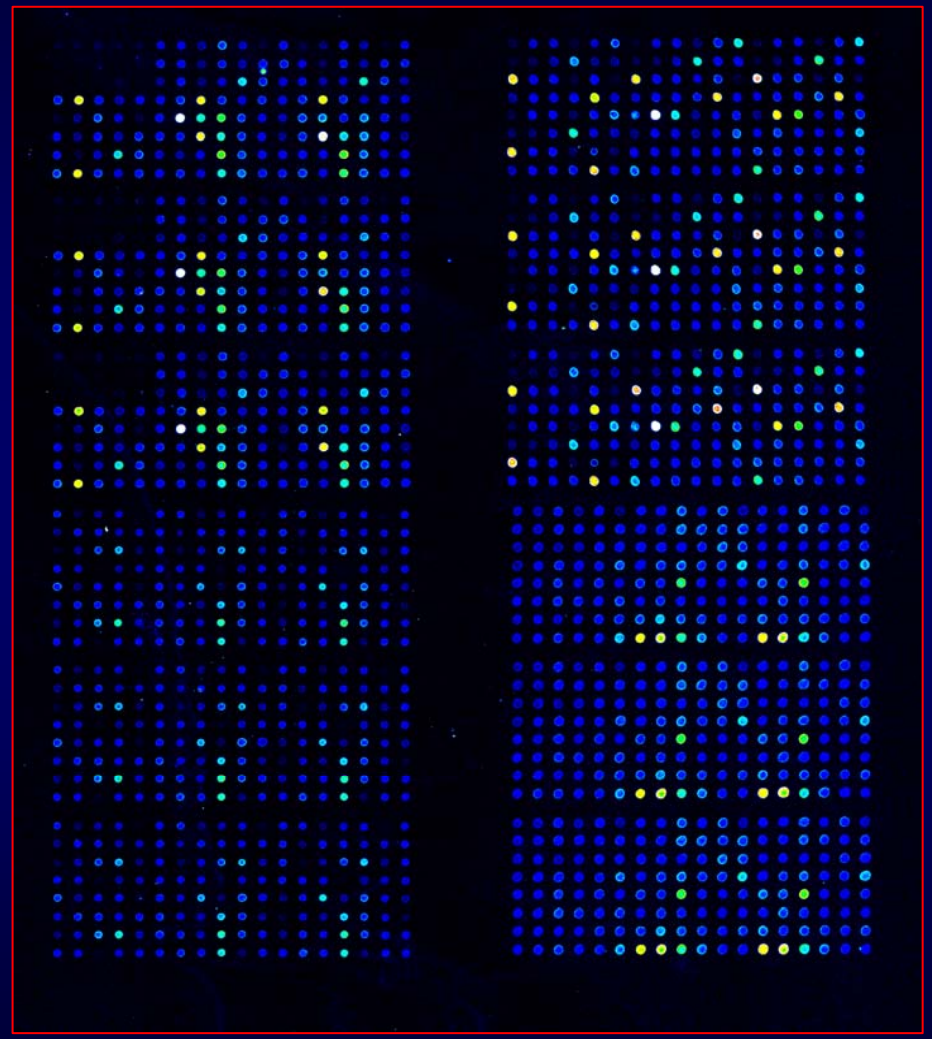


Lyne et al. (2003). *BMC Genomics* 4:27

Self vs Self: same sample

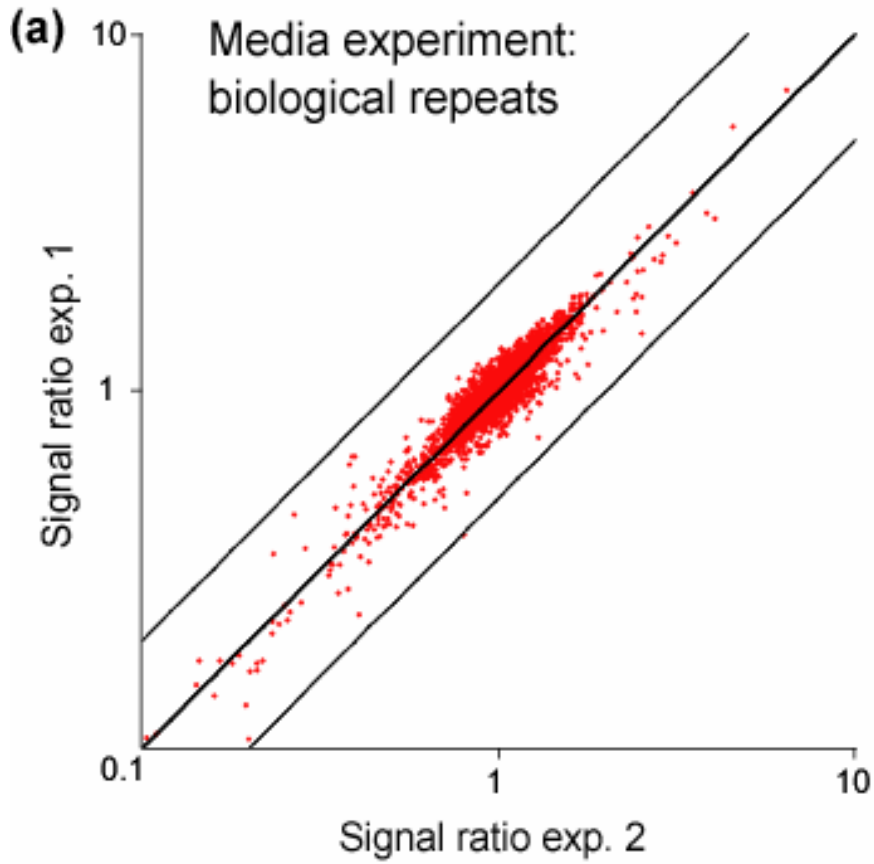


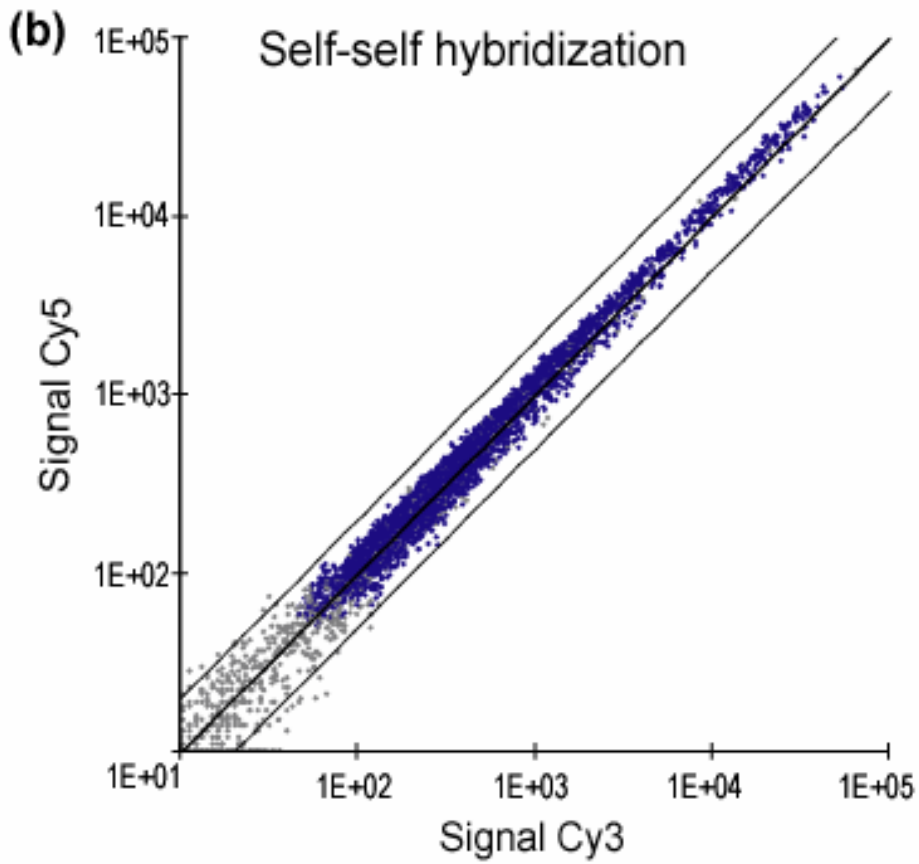
Cy3



Cy5

Reproducibility of signal ratios and intensities:



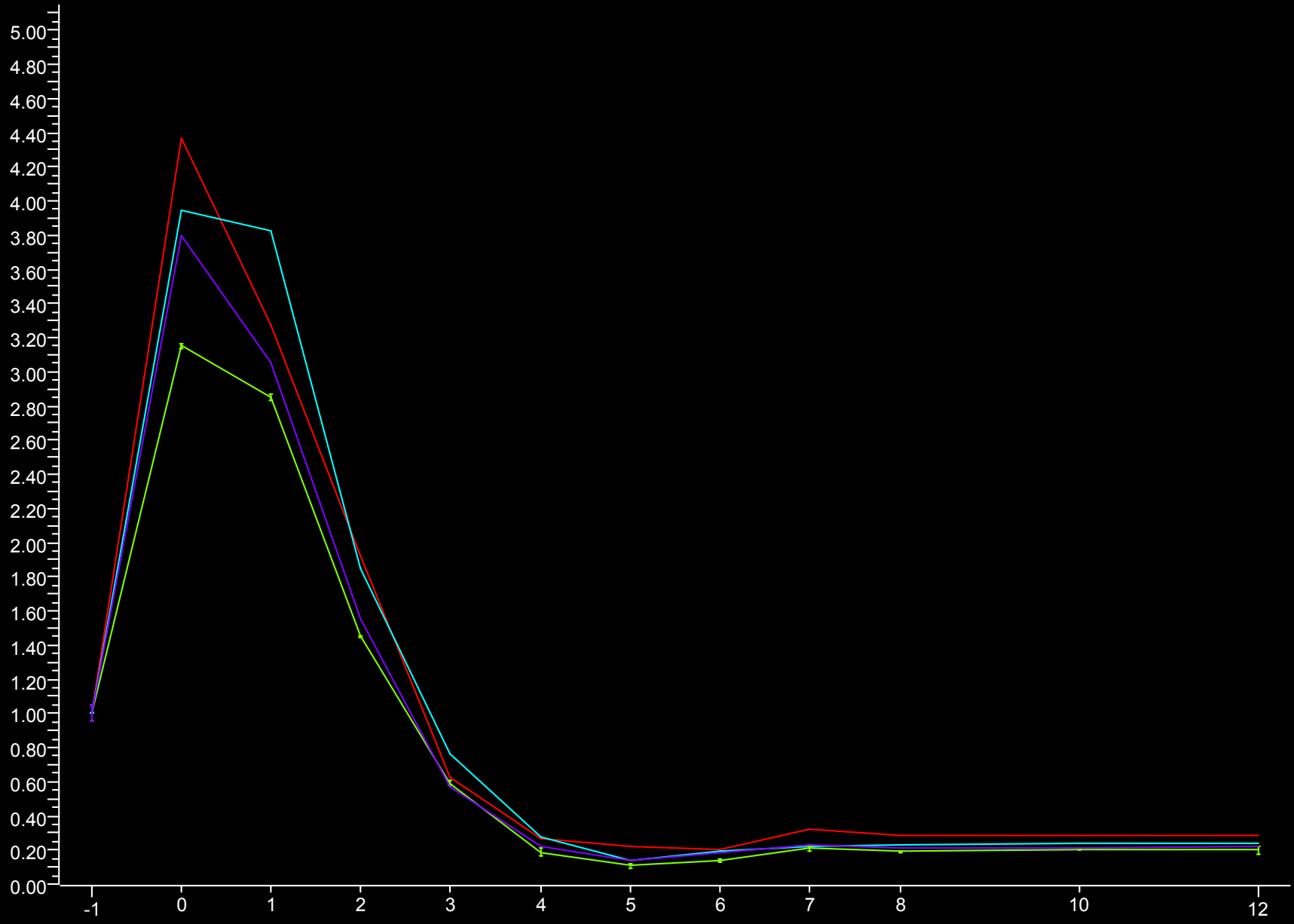


Reproducibility of array data:

Measurement	SD Mean (Range)	CV (Range)
Within array replicates	0.04 (0.03-0.06)	4.4% (3.1-6.2%)
Technical repeats	0.04 (0.02-0.06)	4.5% (2.5-6.3%)
Biological repeats	0.07 (0.05-0.10)	6.4% (4.9-8.1%)

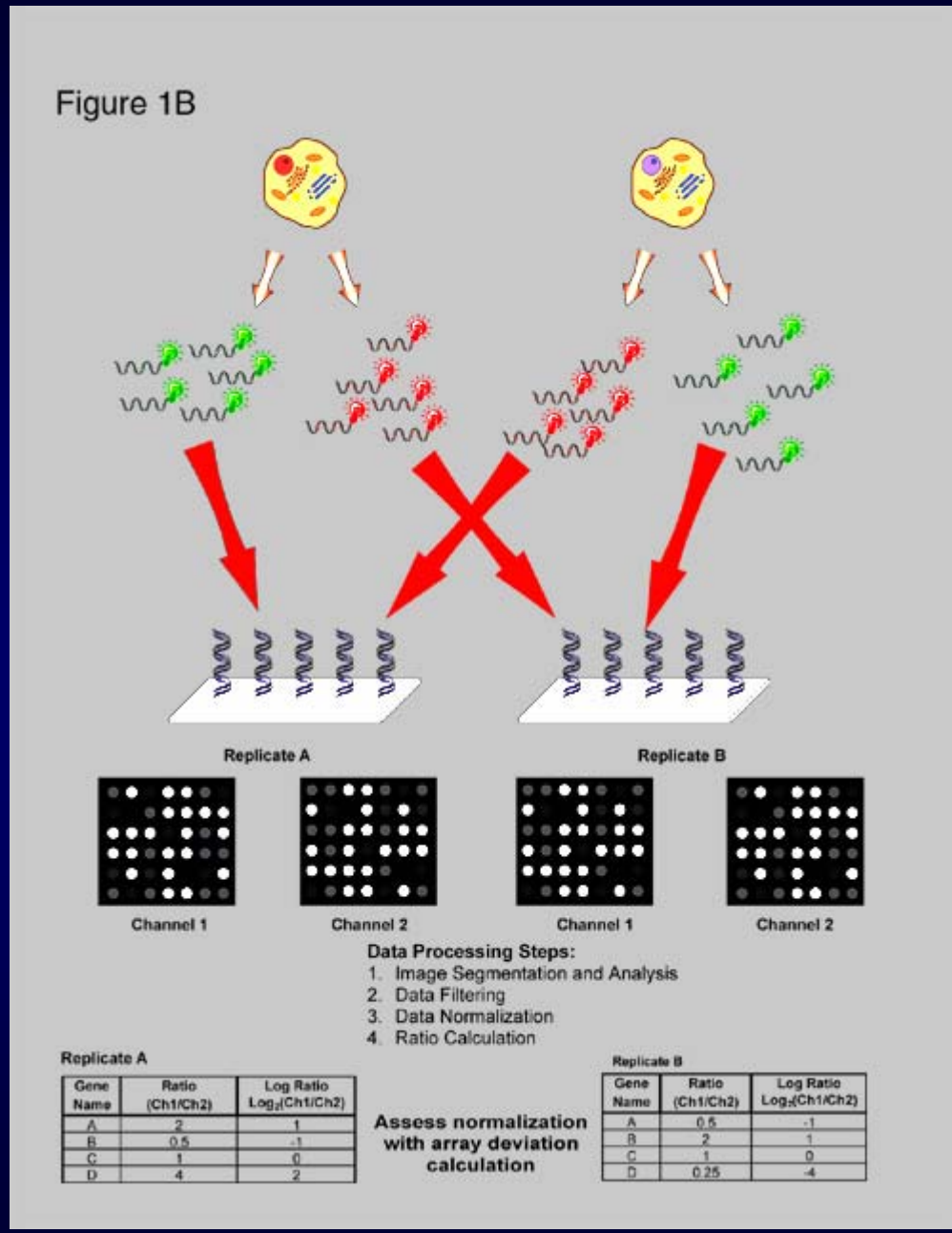
Lyne et al. (2003). *BMC Genomics* 4:27

Reproducibility in 4 experiments: ste11 gene



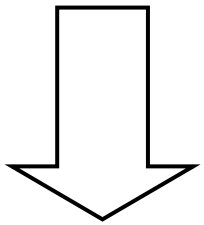
DESIGN OF DYE REVERSAL REPLICATE

- Replicate experiment in which we assess the same mRNA pools but invert the dyes used
- The replicates are independent biological experiments
- Balanced, deals with biases of dye incorporation



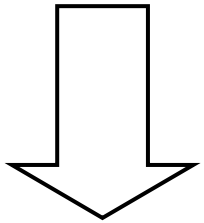
Clustering techniques

Aim: Discover structure/patterns in the data



- Classify/cluster genes according to their expression profiles
- Apply special visualisation techniques

Aim: Group together (cluster) genes that behave similarly across different conditions



Define/quantify similarity

Fundamental: Different definitions will lead to different classifications


PEPSI
diet


NEW
Coca-Cola
vanilla

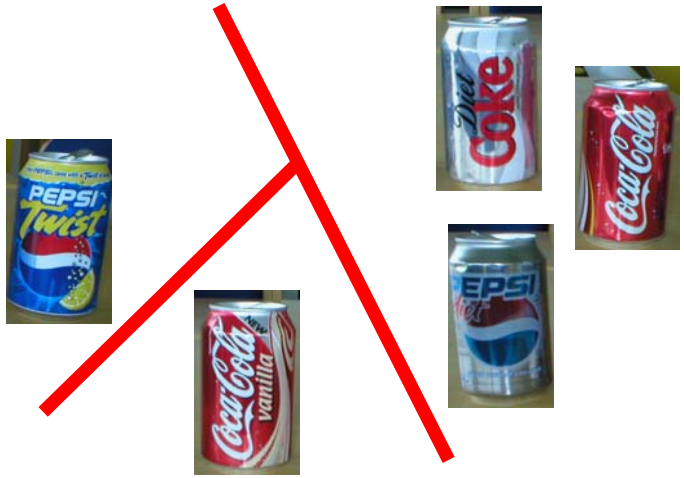

Diet
Coke


Coca-Cola


PEPSI
Twist

Sugar content

Flavour



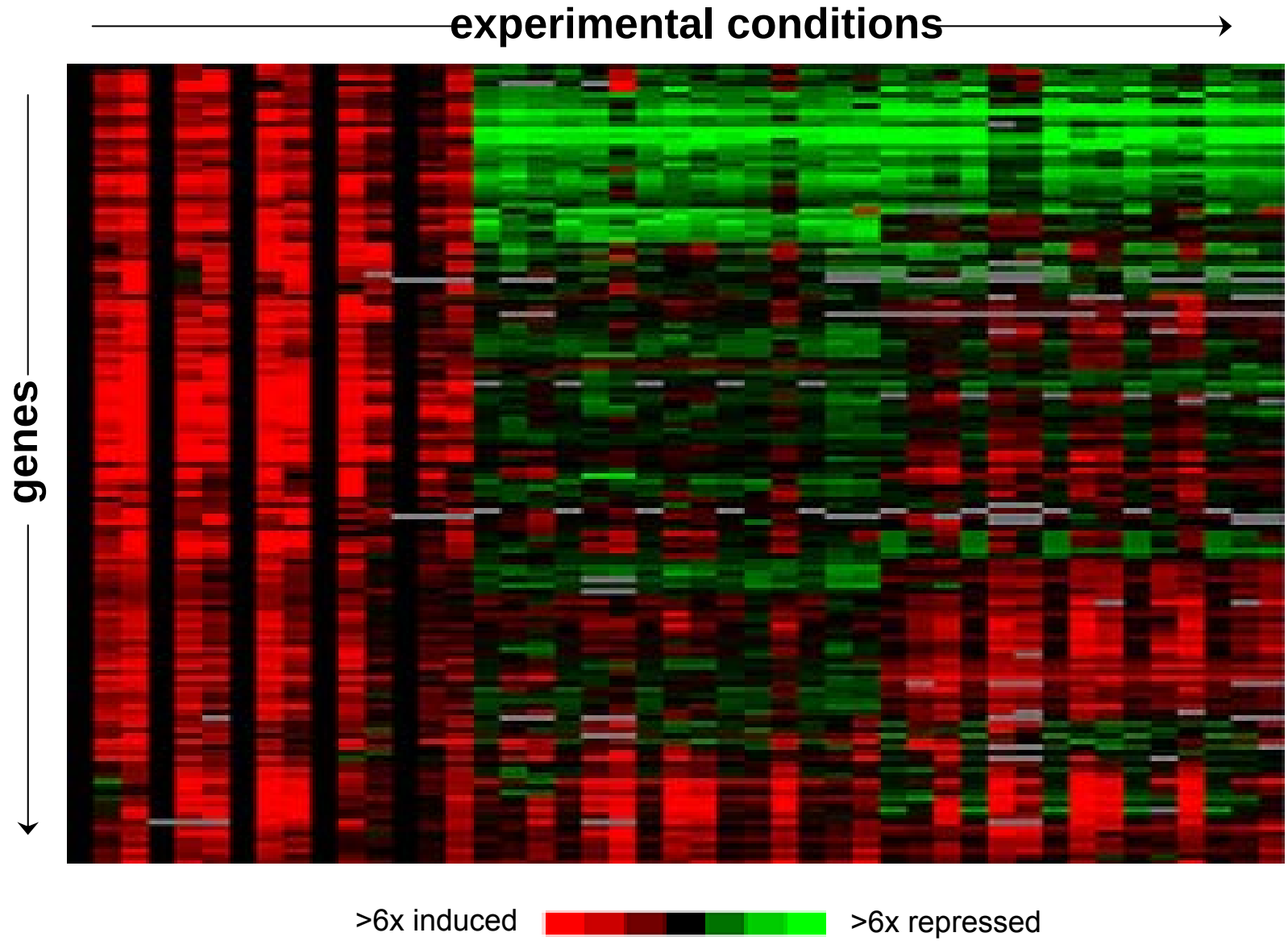
Colour

Brand

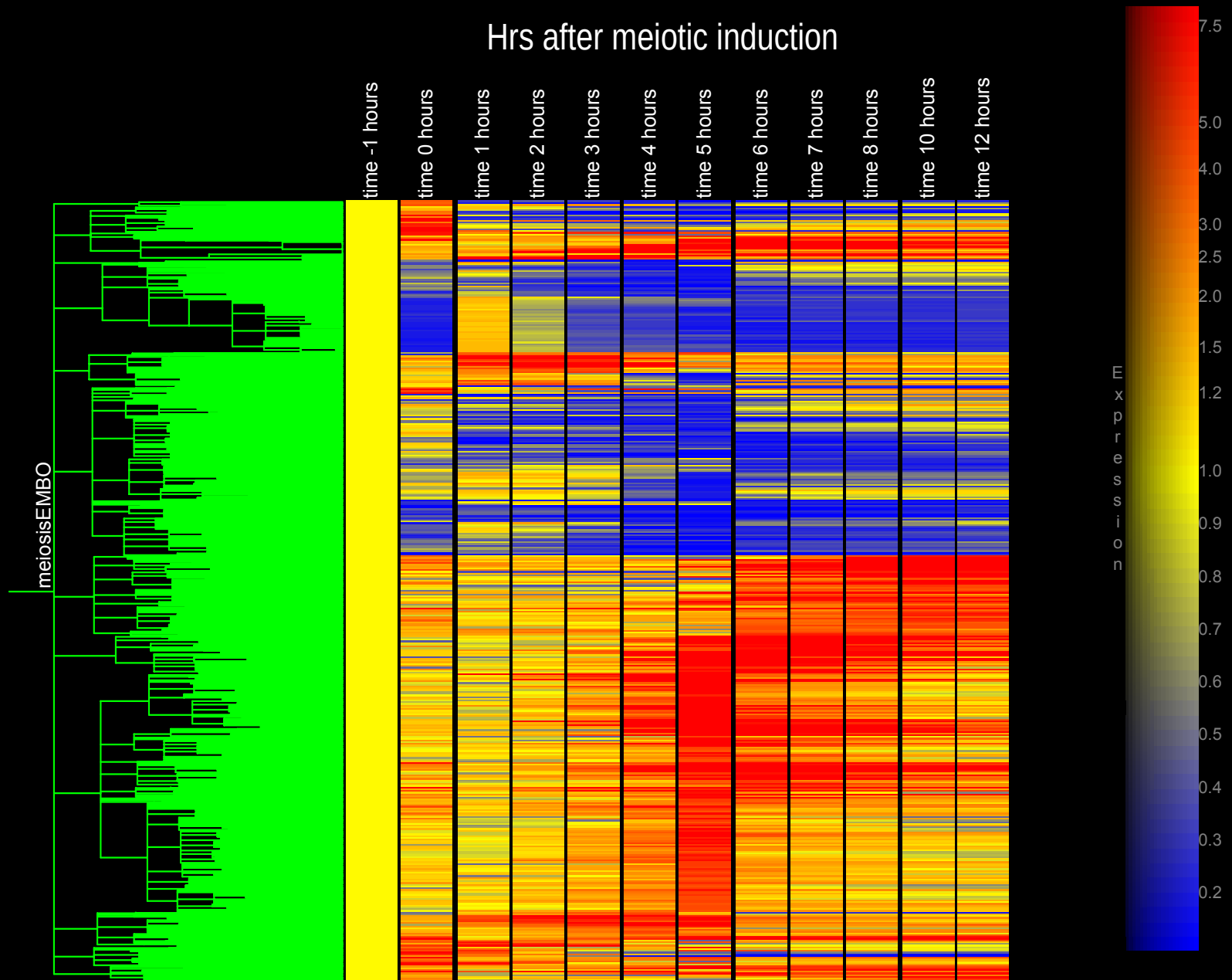
Measures of similarity/distance for gene expression

- Aim: quantify the degree of similarity between two gene expression profiles
- There are dozens of similarity metrics
- Examples:
  - Euclidean distance
  - Standard correlation
  - Pearson correlation

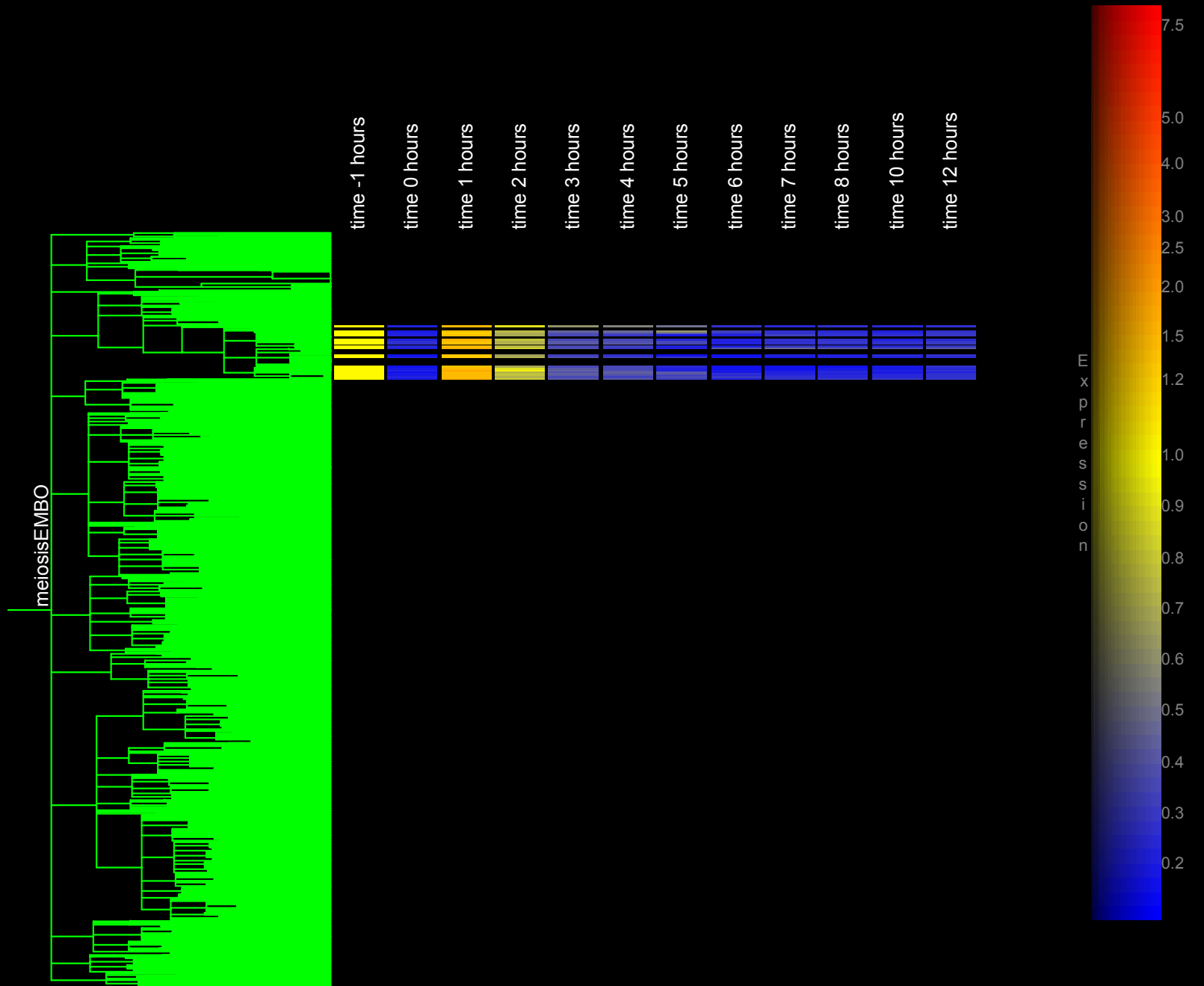
Hierarchical gene clustering:

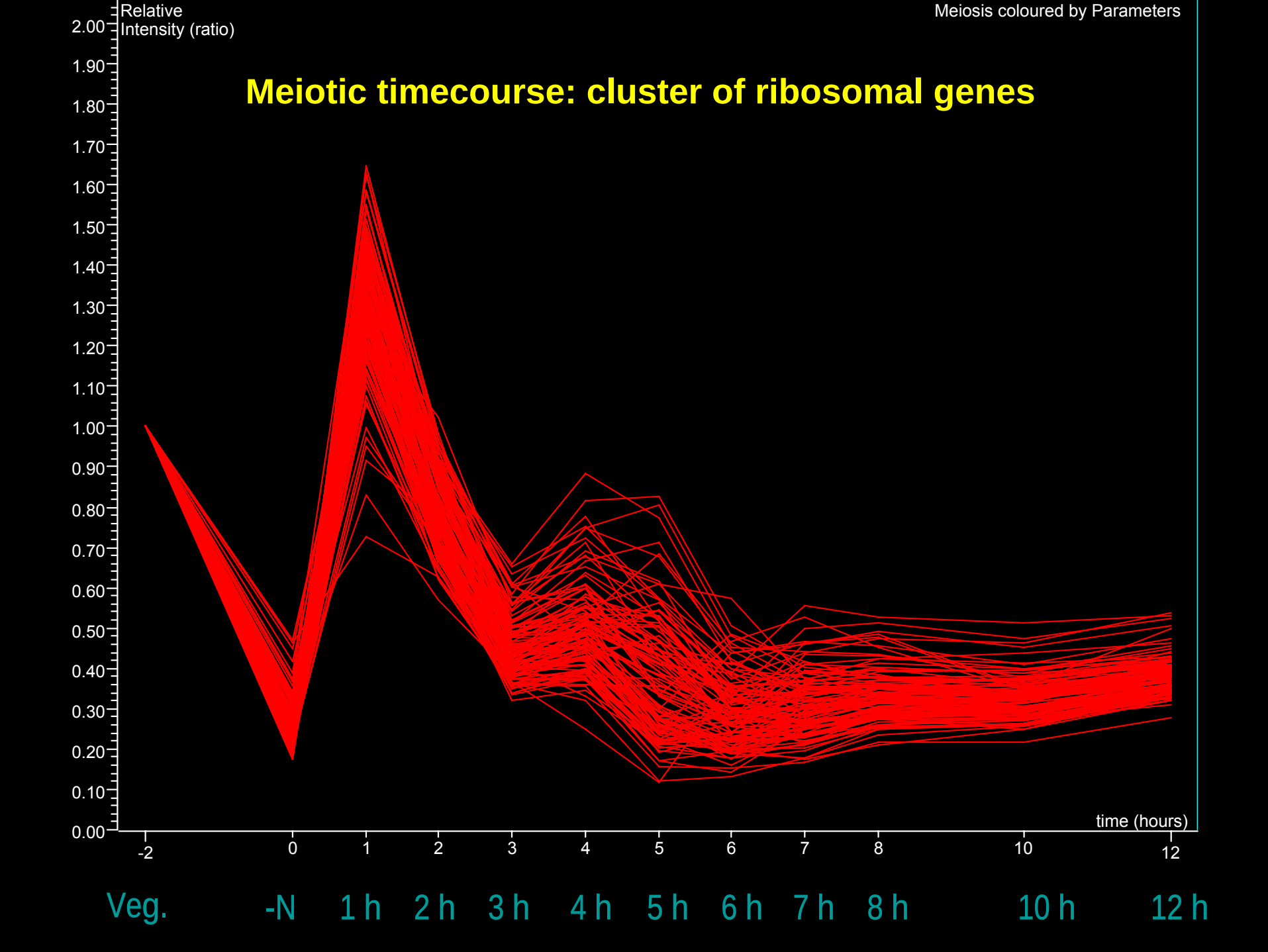


Hierarchical Gene Clustering:

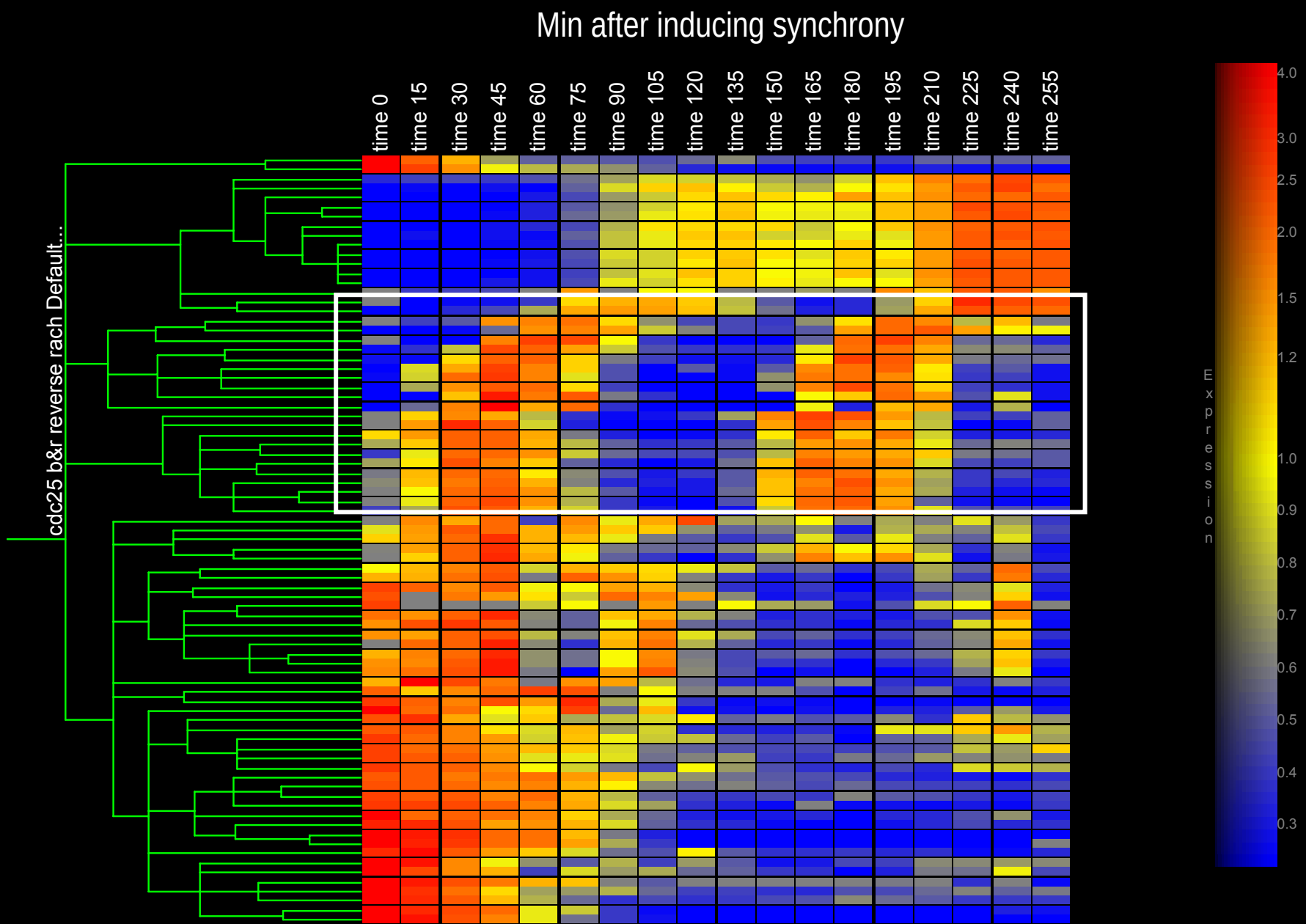


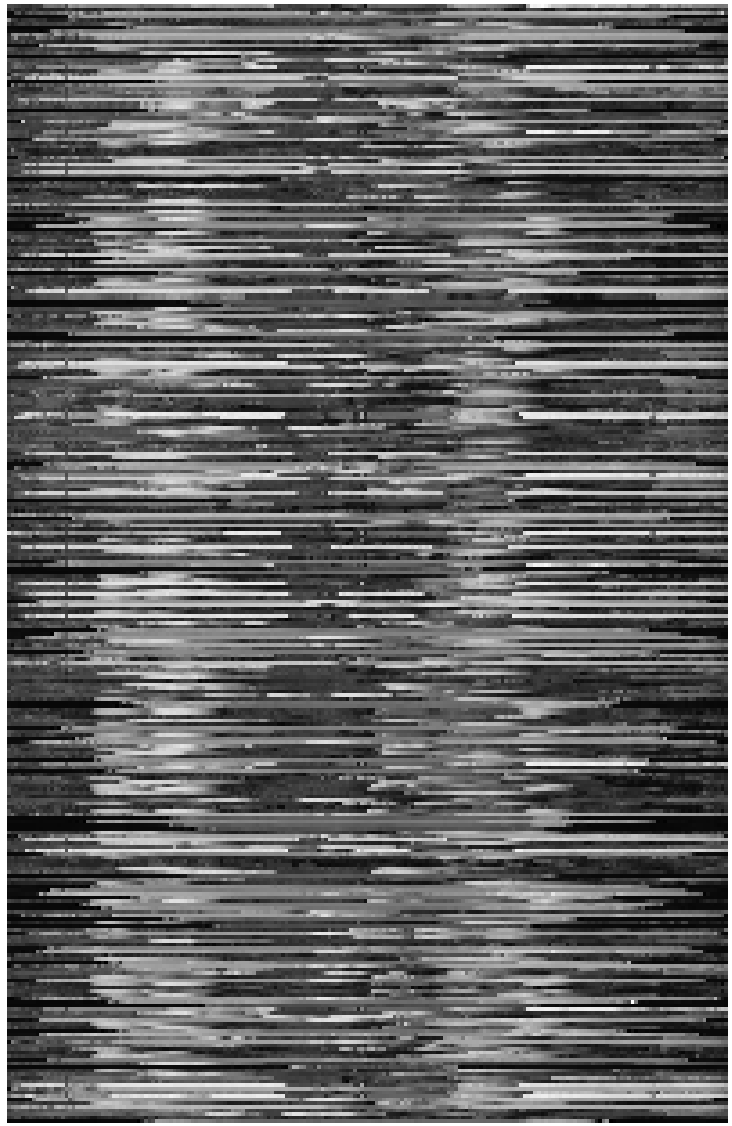
Ribosomal Gene Cluster

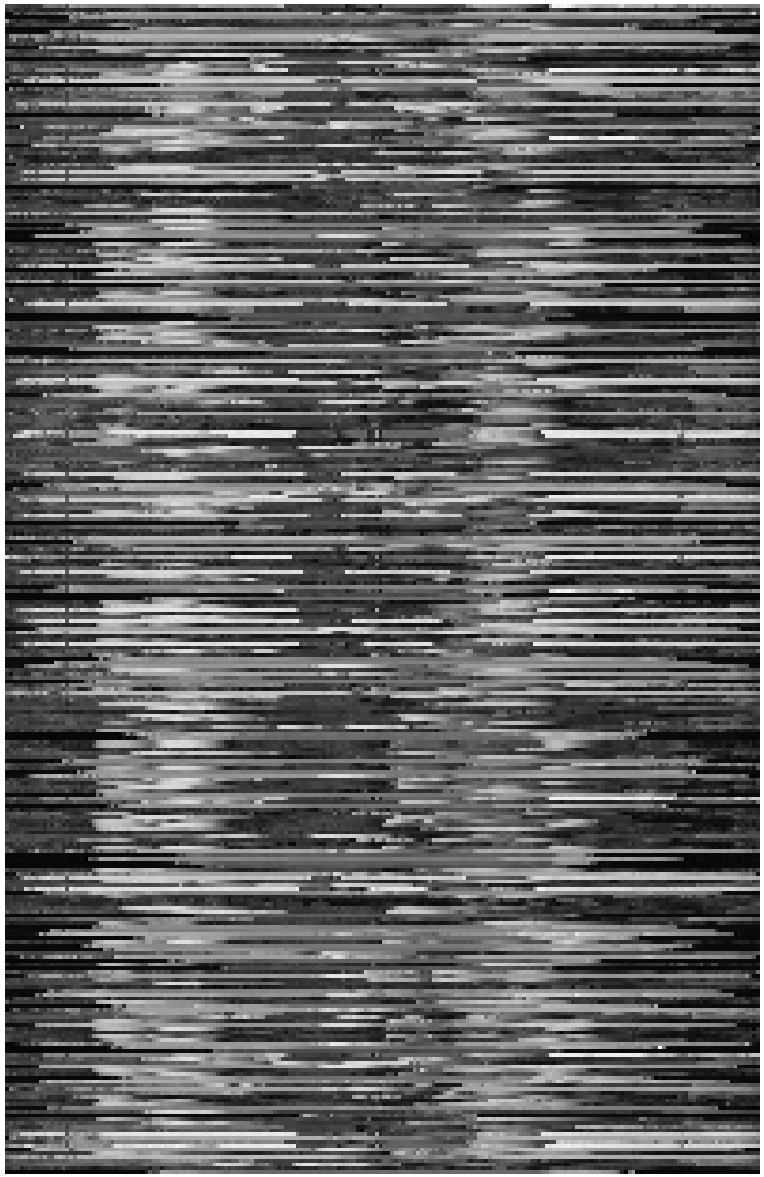


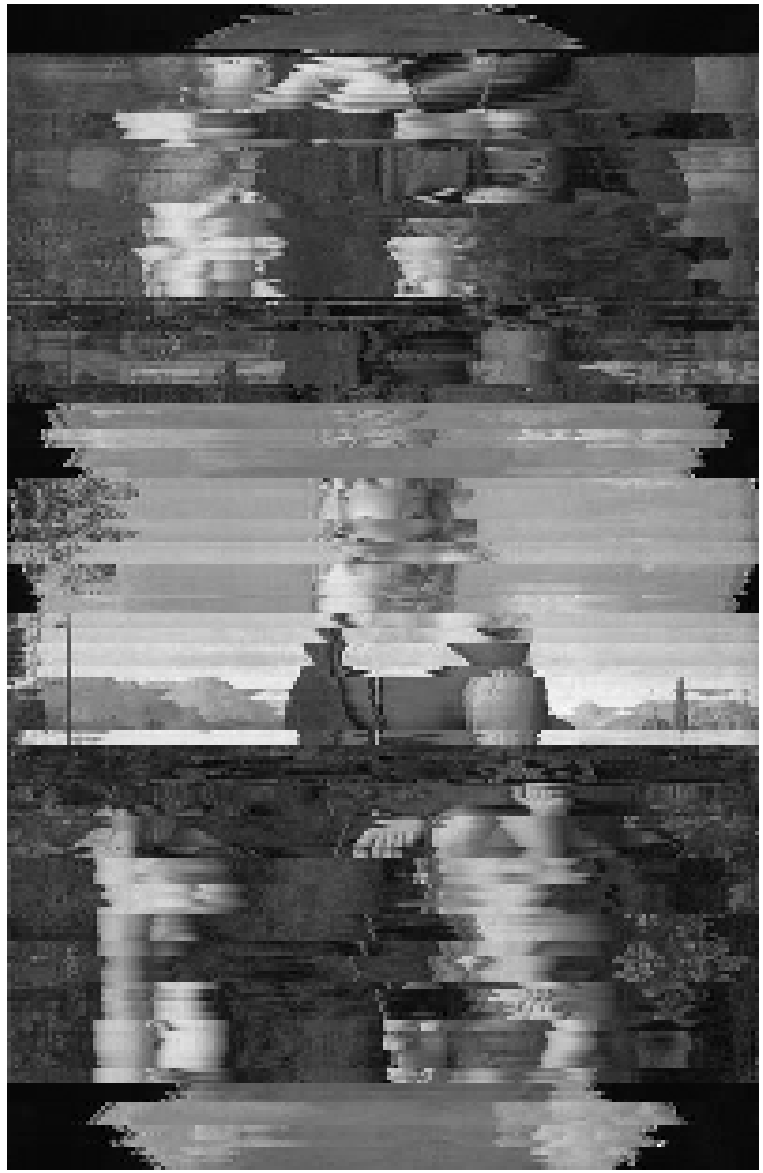

Relative
Intensity (ratio)
Meiosis coloured by Parameters
Meiotic timecourse: cluster of ribosomal genes
2.00
1.90
1.80
1.70
1.60
1.50
1.40
1.30
1.20
1.10
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
time (hours)
-2
0
1
2
3
4
5
6
7
8
10
12
Veg.
-N
1 h
2 h
3 h
4 h
5 h
6 h
7 h
8 h
10 h
12 h

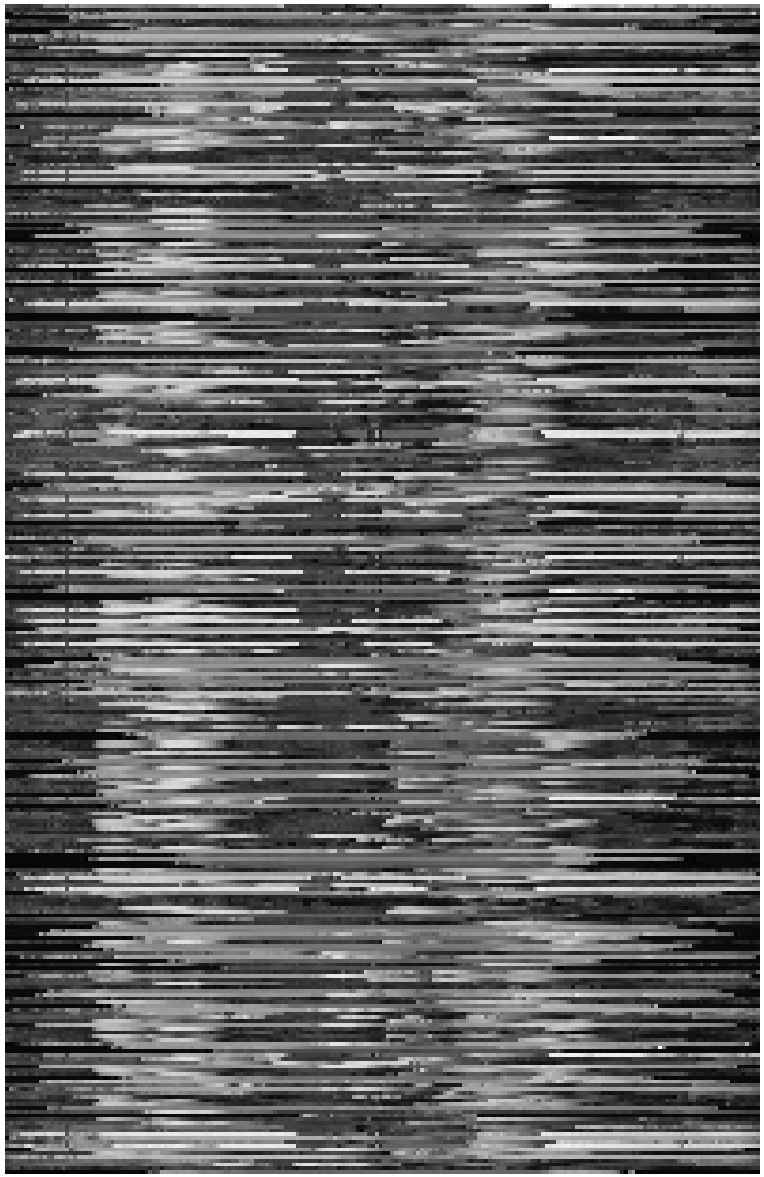
Clustering of genes regulated during the cell cycle:









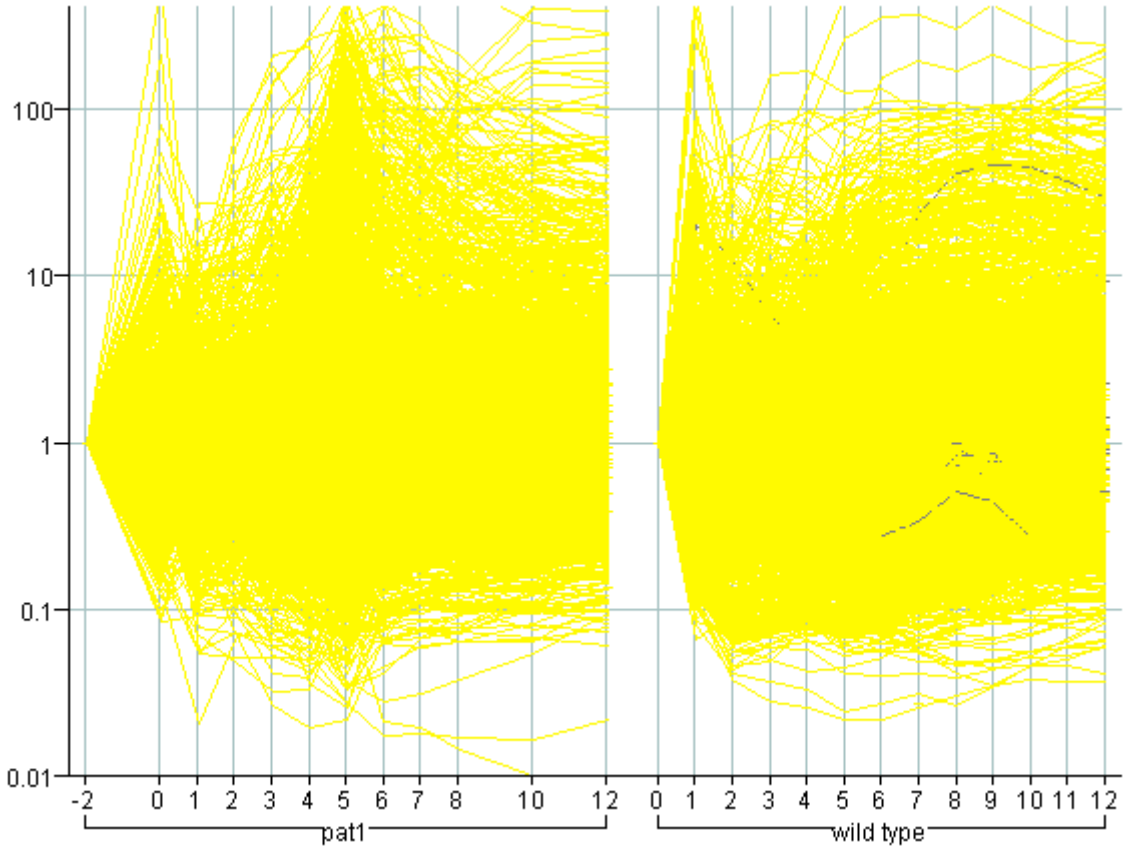


Mike Eisen

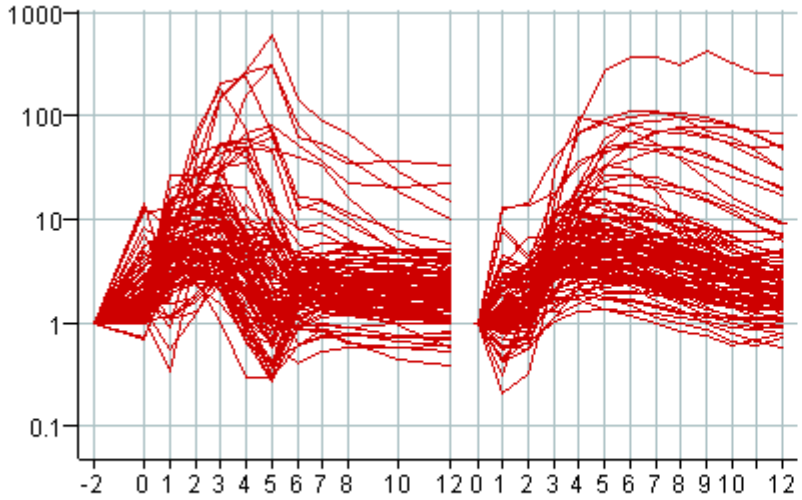
K-means clustering

- Classifies genes into non-overlapping groups
- The number of clusters (k) is specified by the user
- Unsupervised method

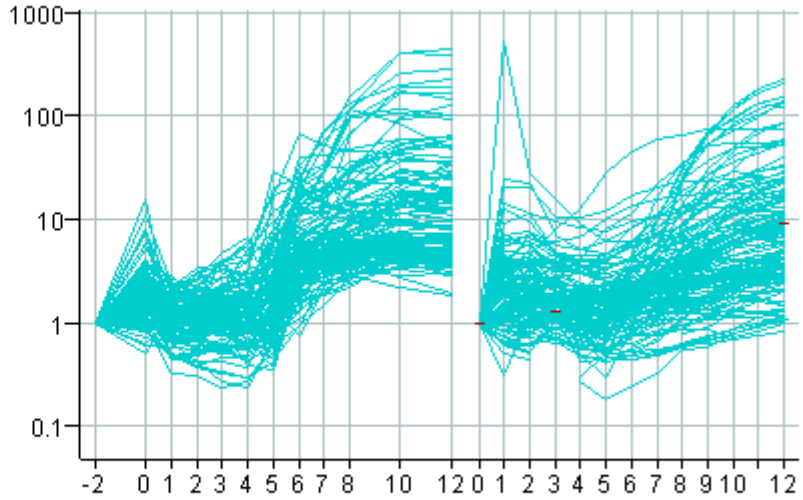
k-Means Clustering



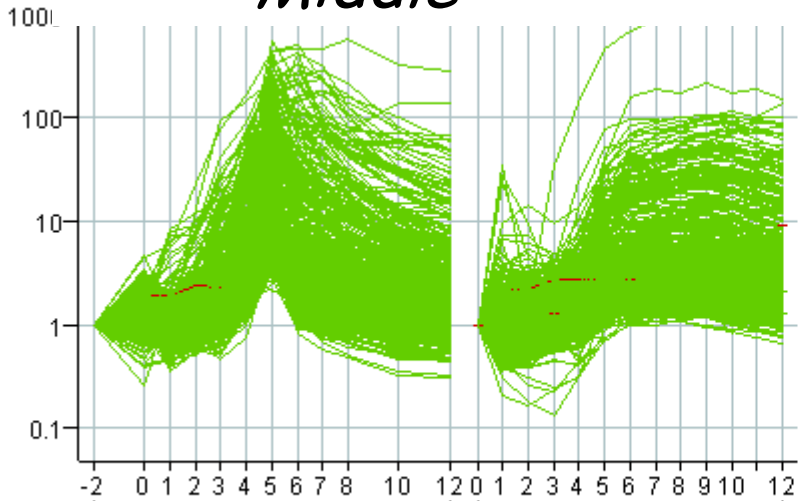
Early



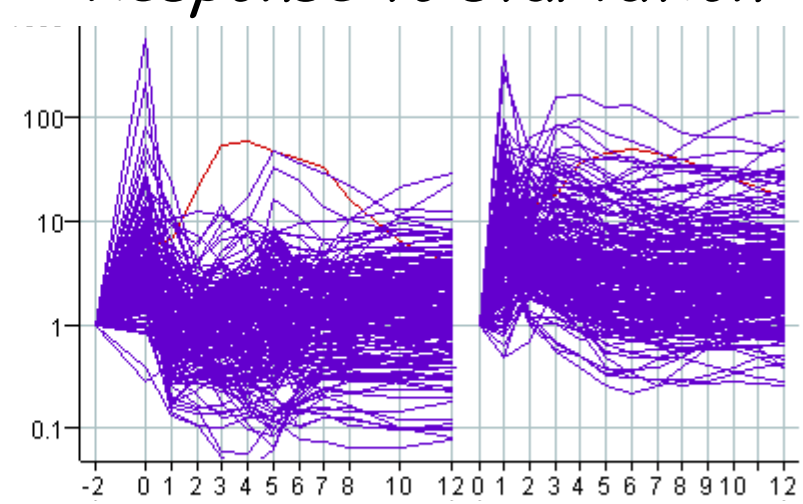
Late



Middle

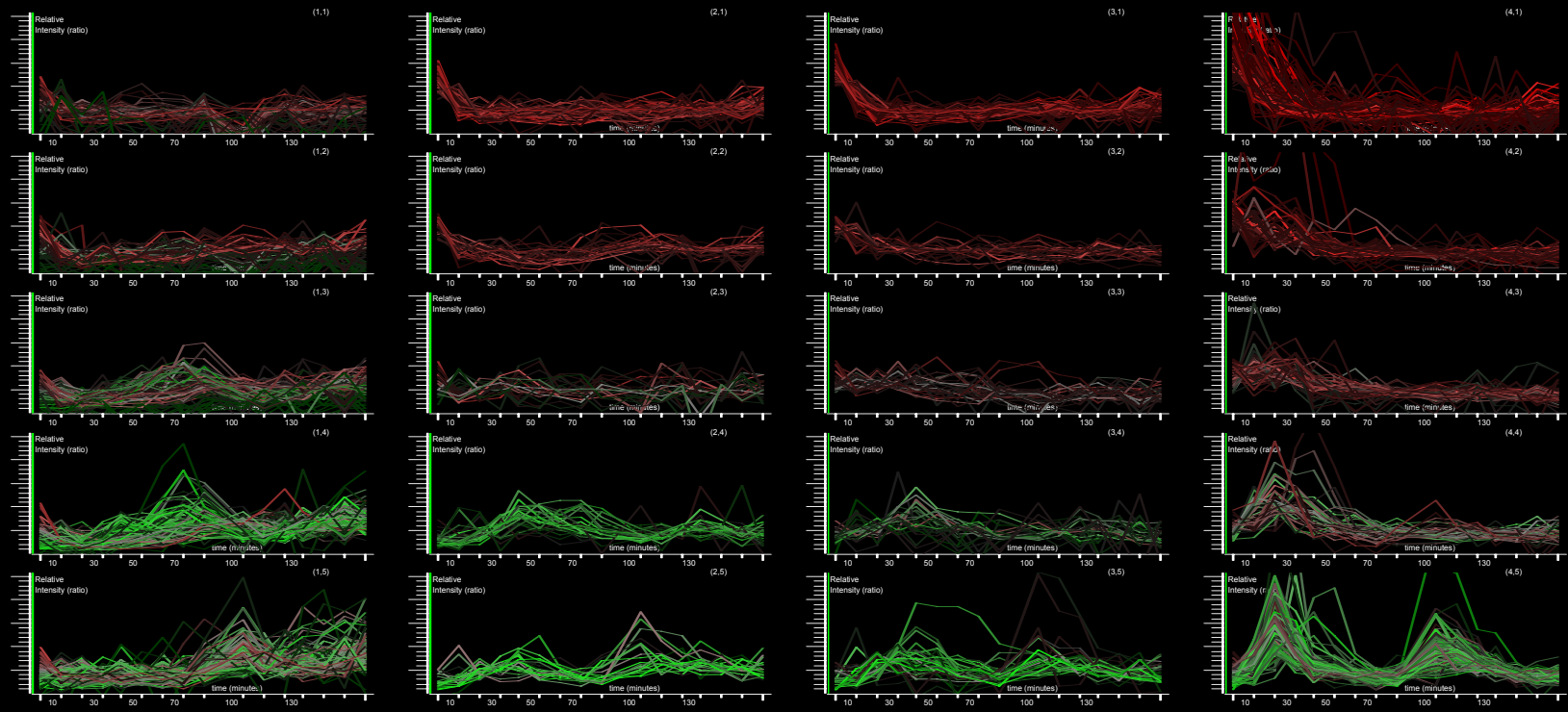


Response to starvation



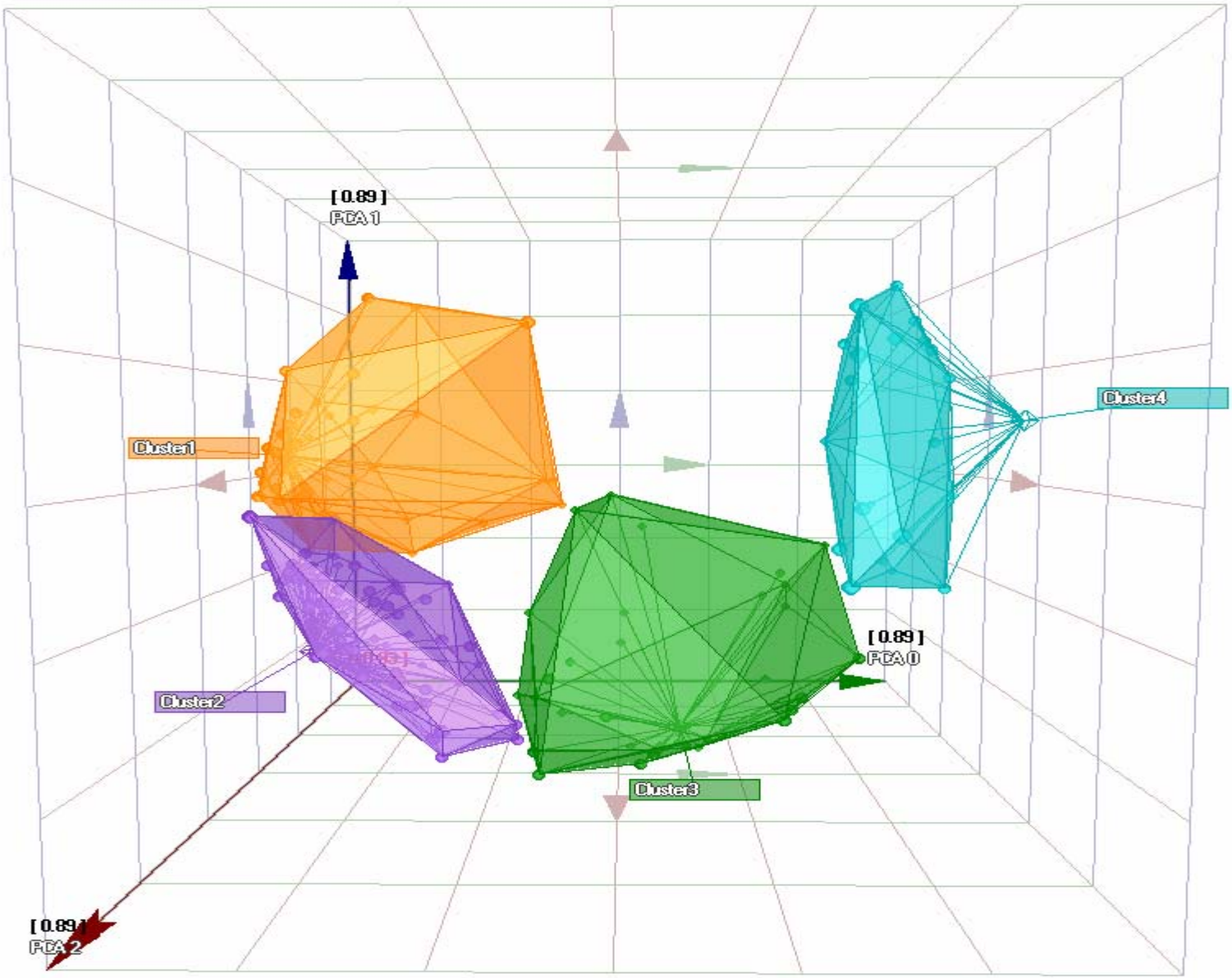
Self-Organizing Maps (SOMs)

- User specifies number of sets by indicating how many rows and columns their data should be divided into
- There is relation within a set AND relation of one set to another
- Particularly suitable for large datasets



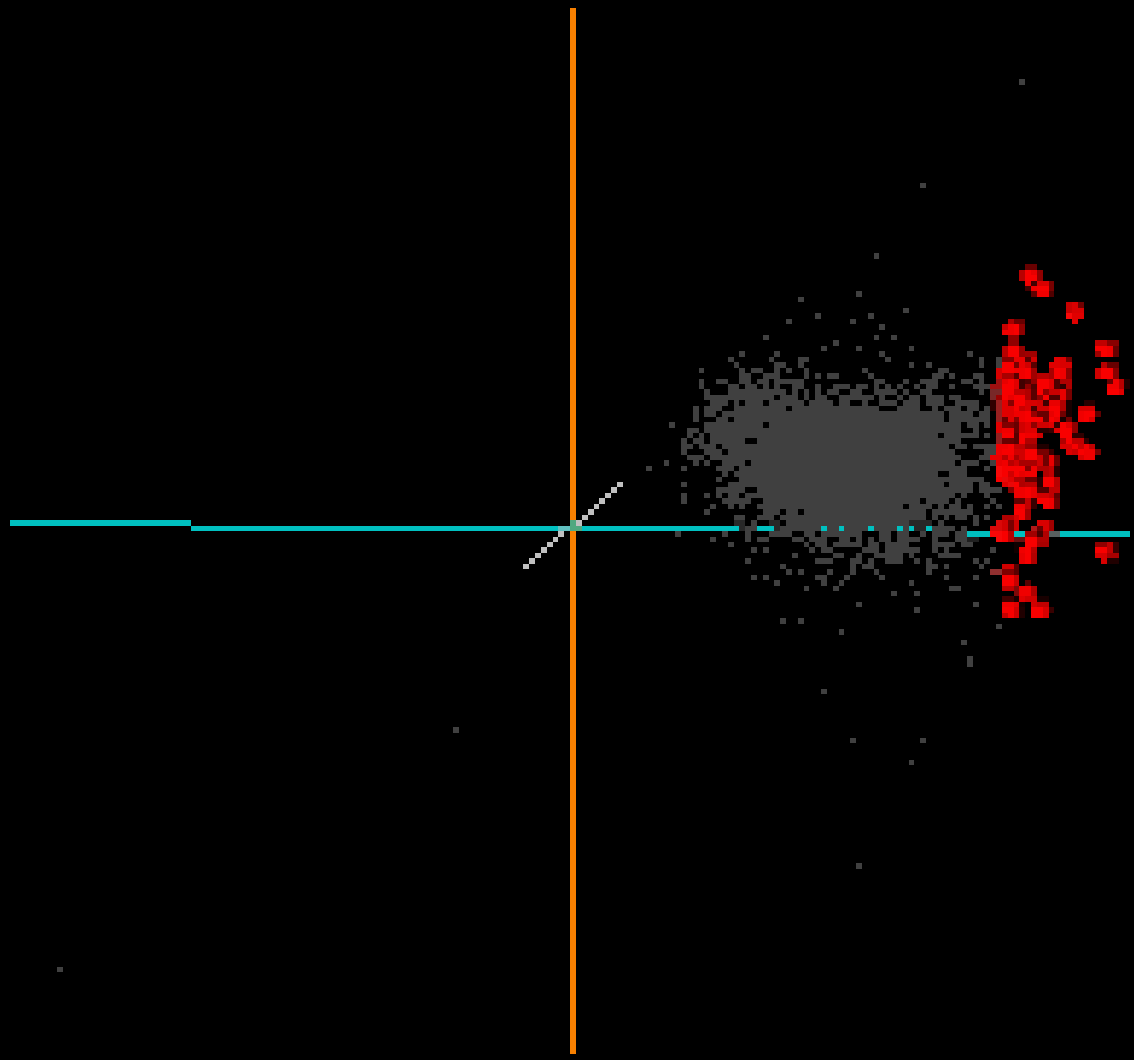
Principal component analysis:

Project multi-dimensional data into 2 or 3 dimensions

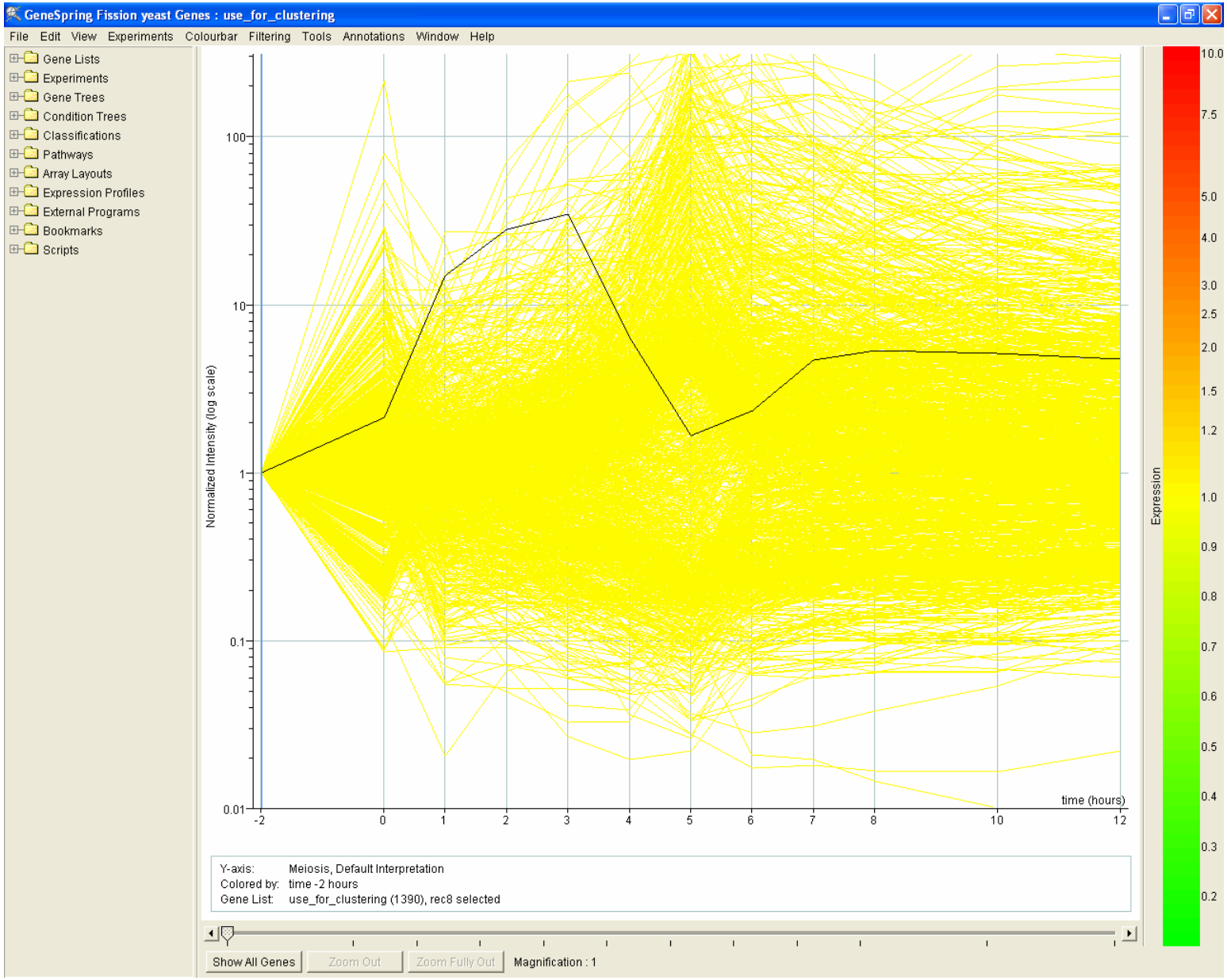


Principal Components Analysis

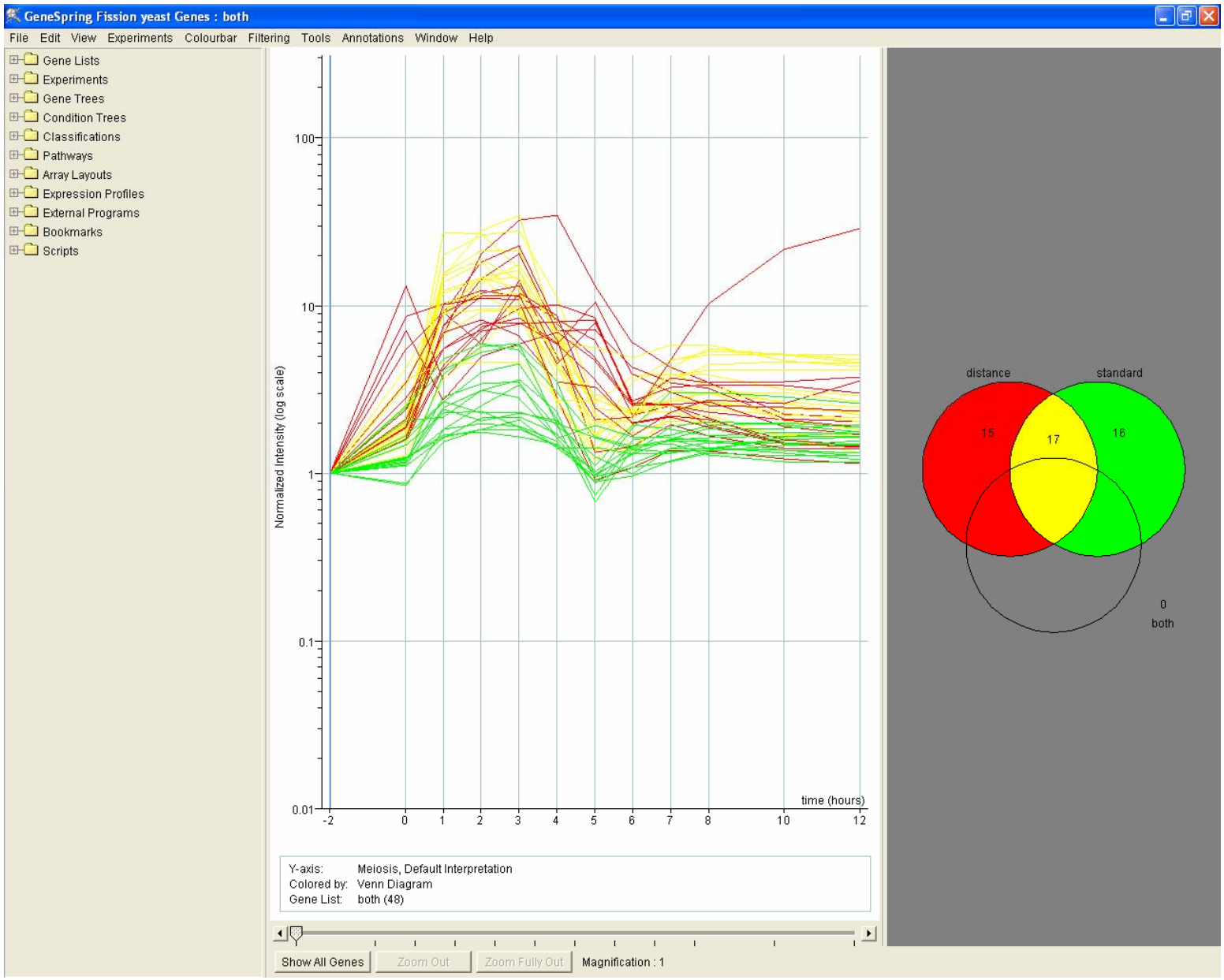
Reduce multi-dimensional datasets to 3 dimensions:



Finding genes with similar expression profiles: significant correlations

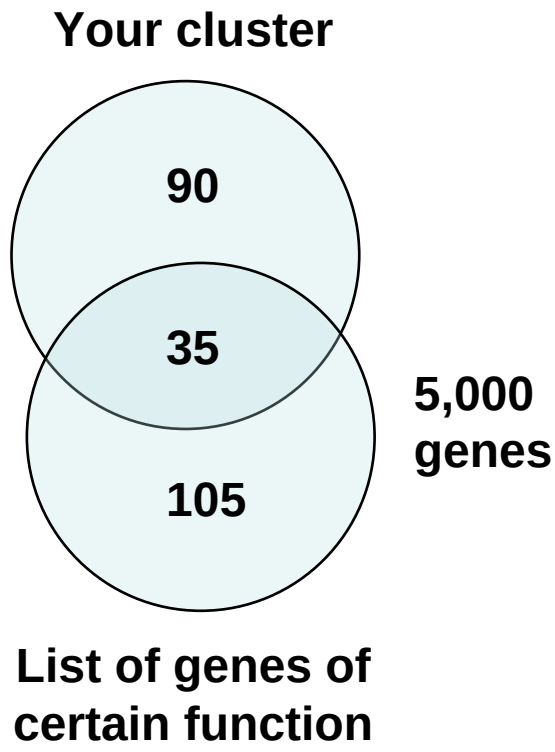


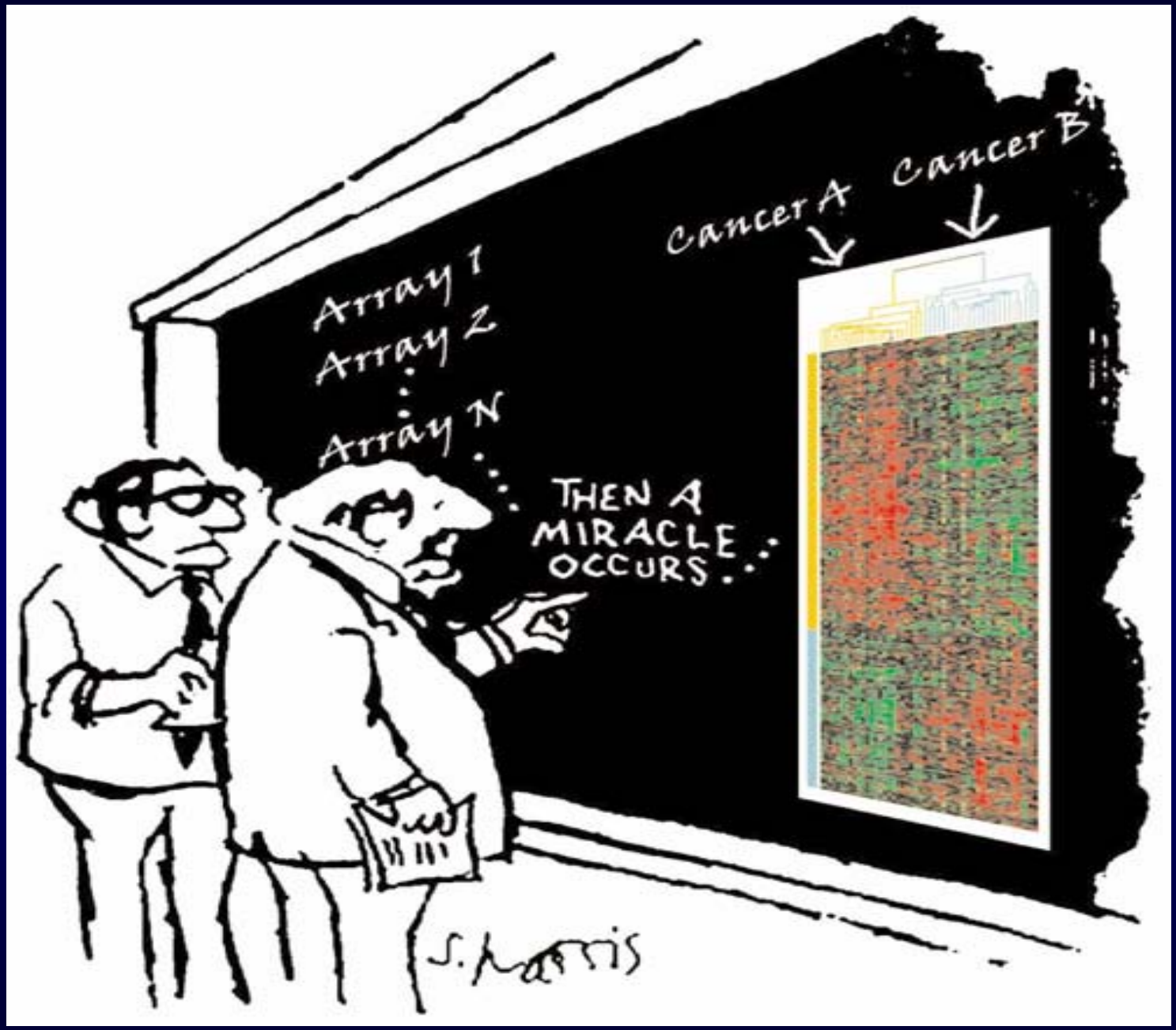
Comparing lists using Venn diagrams



Which types of genes are enriched in a cluster?

- Idea: Compare your cluster of genes with lists of genes with common properties (function, expression, location).
- Find how many genes overlap between your cluster and a gene list.
- Calculate the probability of obtaining the overlap by change (using the hypergeometric distribution). This measures if the enrichment is significant.
- This analysis provides an unbiased way of detecting connections between expression and function.





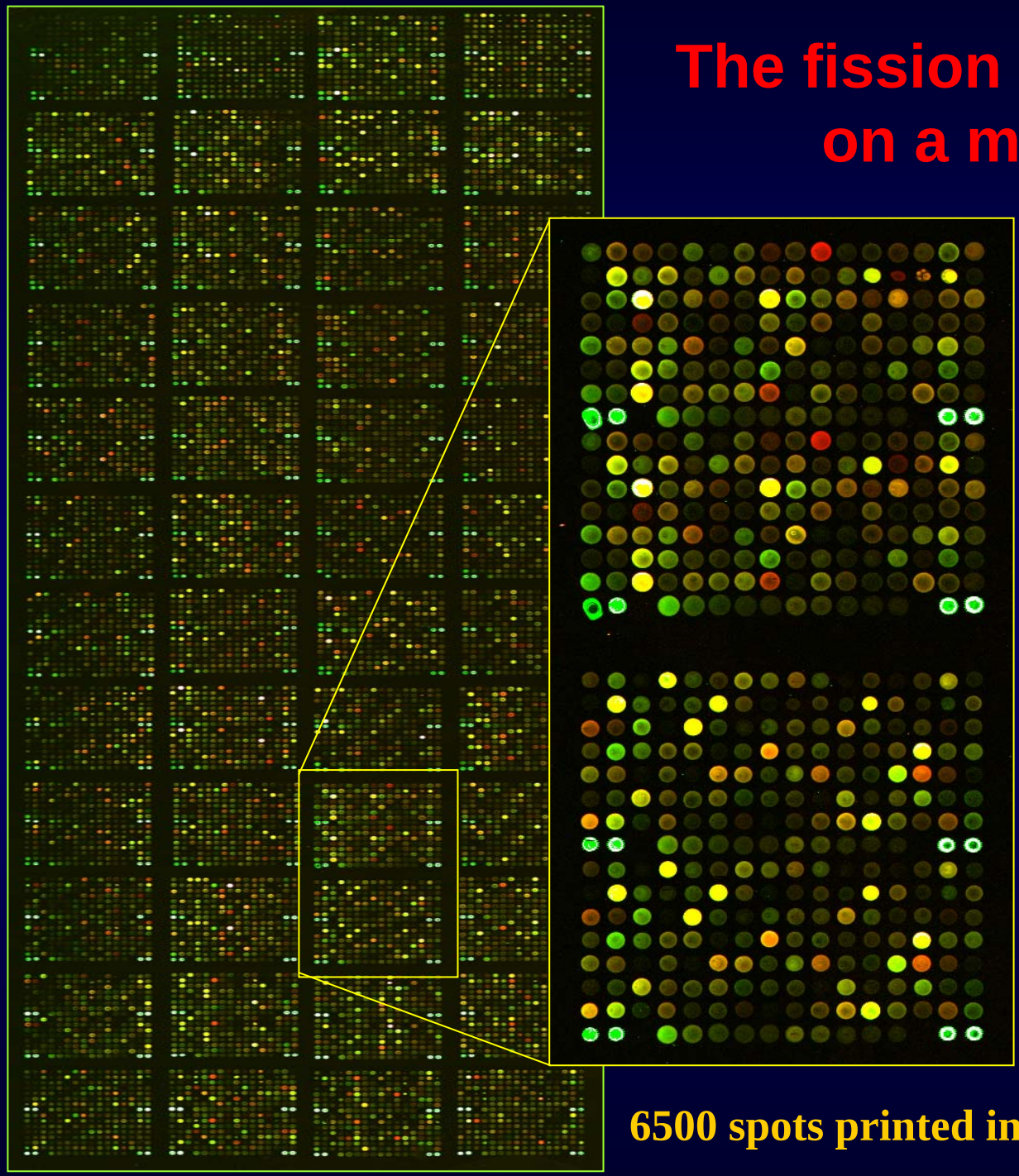
"I think you should be more explicit here in step 2."

Schizosaccharomyces pombe (Fission Yeast)

- unicellular eukaryote (fungus)
- genome: ~5,000 genes
- easy to handle / genetics
- evolutionary distant to *S. cerevisiae*
- simple model system
- no beauty – but what a beast!

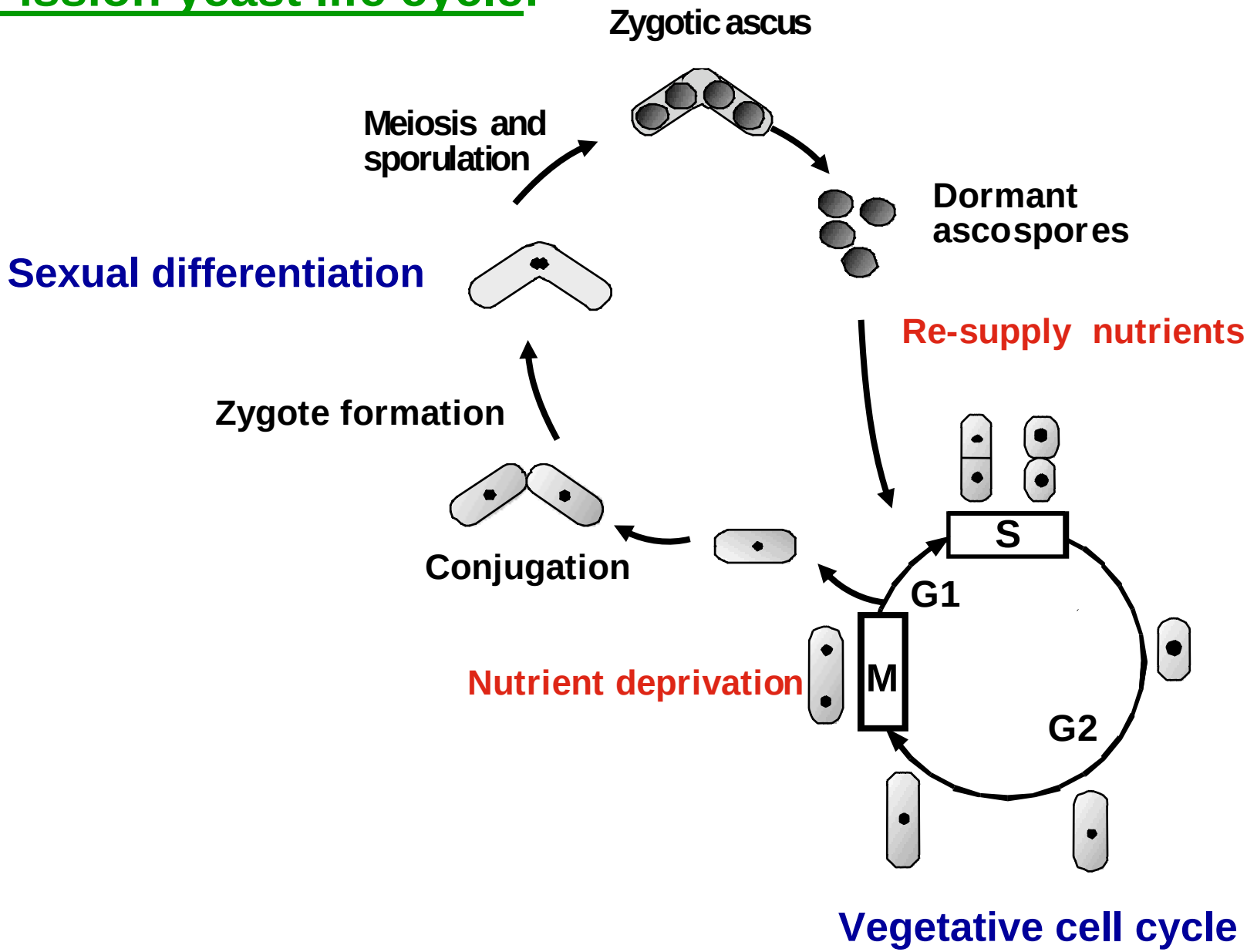
Now –which one is the ‘higher’ eukaryote?!

The fission yeast genome on a microarray

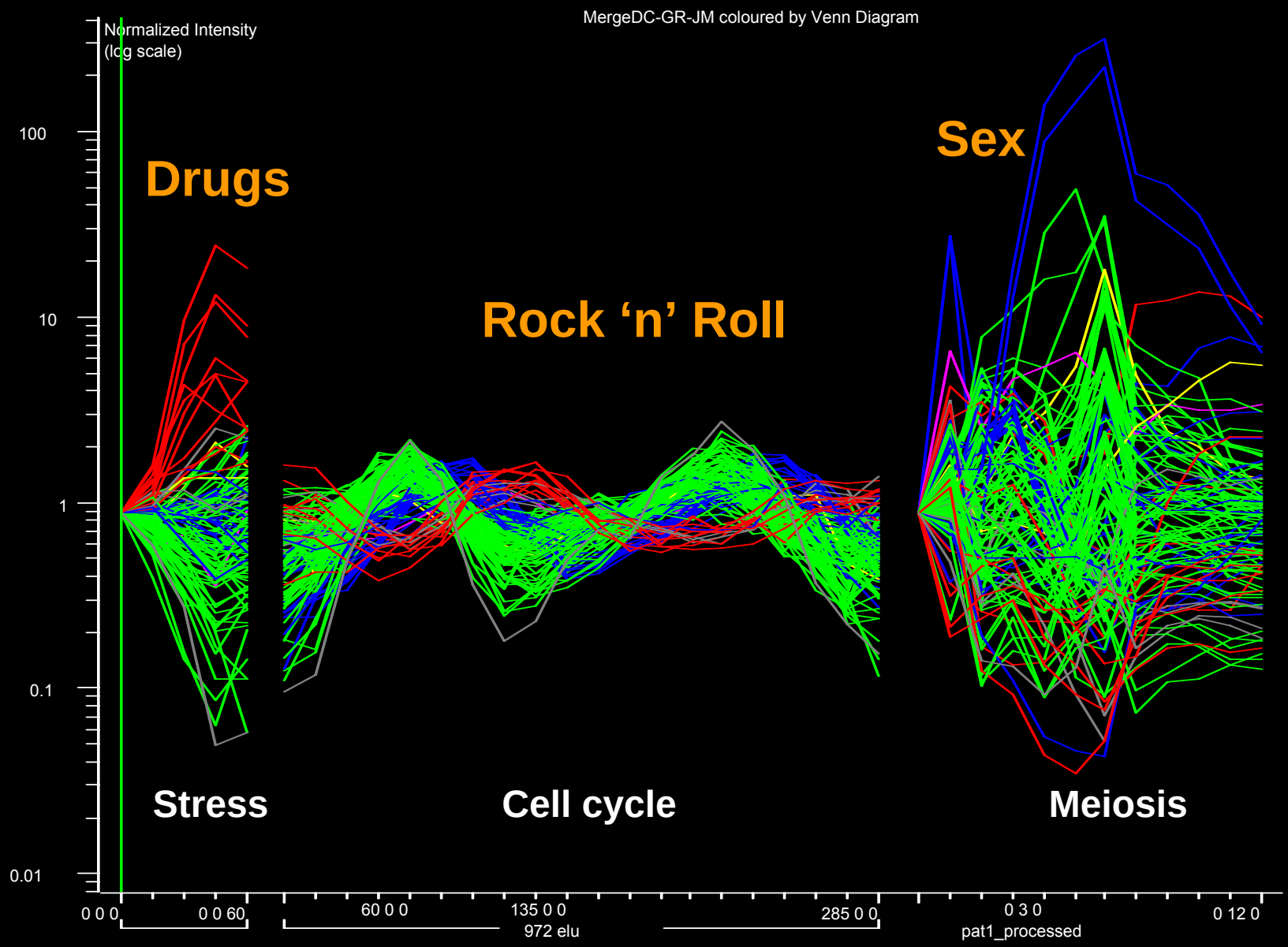


6500 spots printed in duplicate: 13,000 spots

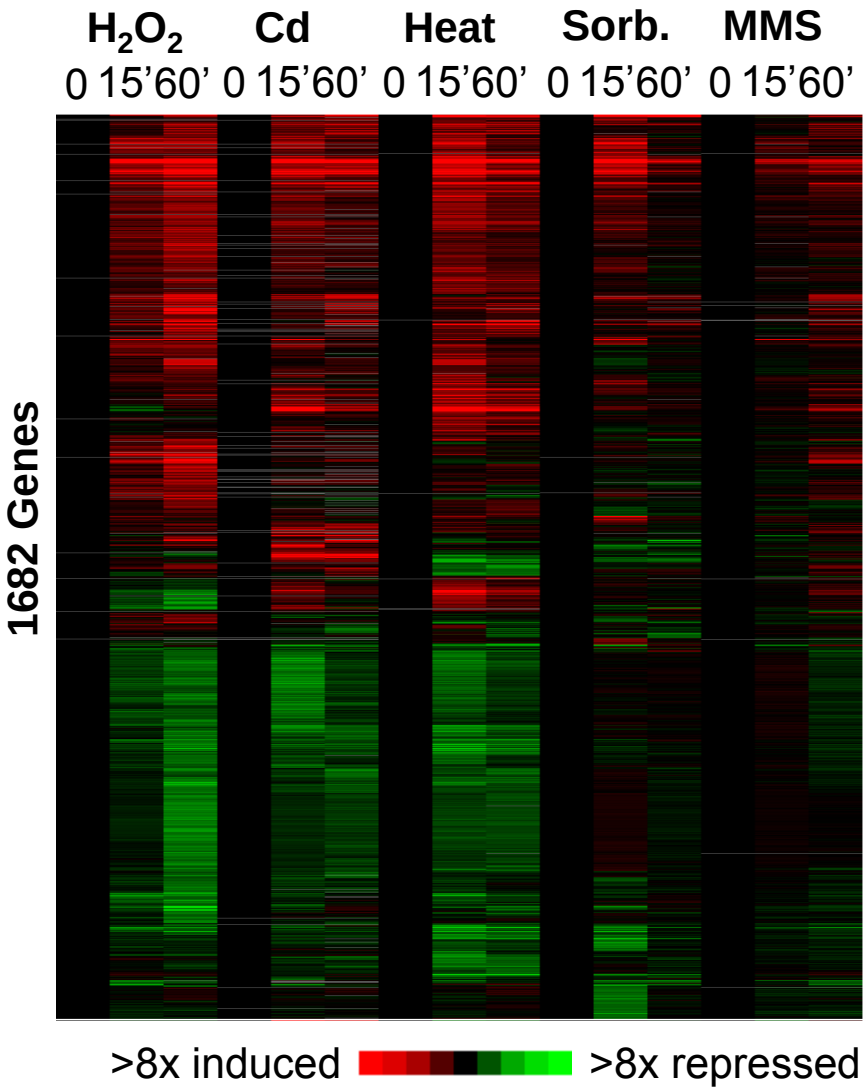
Fission yeast life cycle:

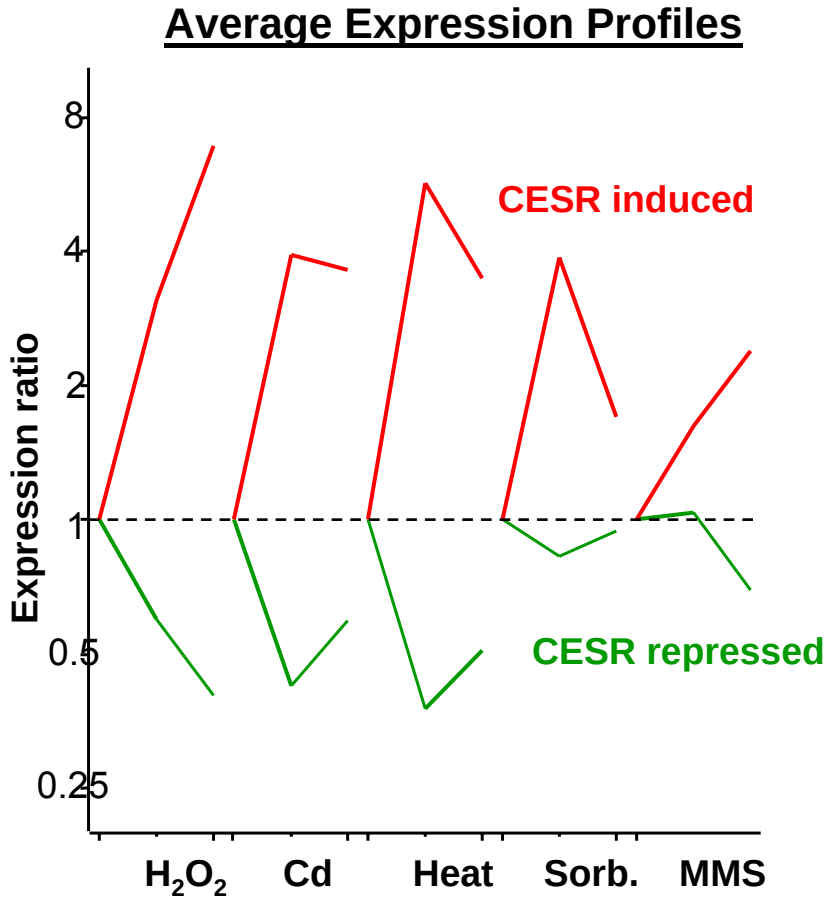


Three main projects:

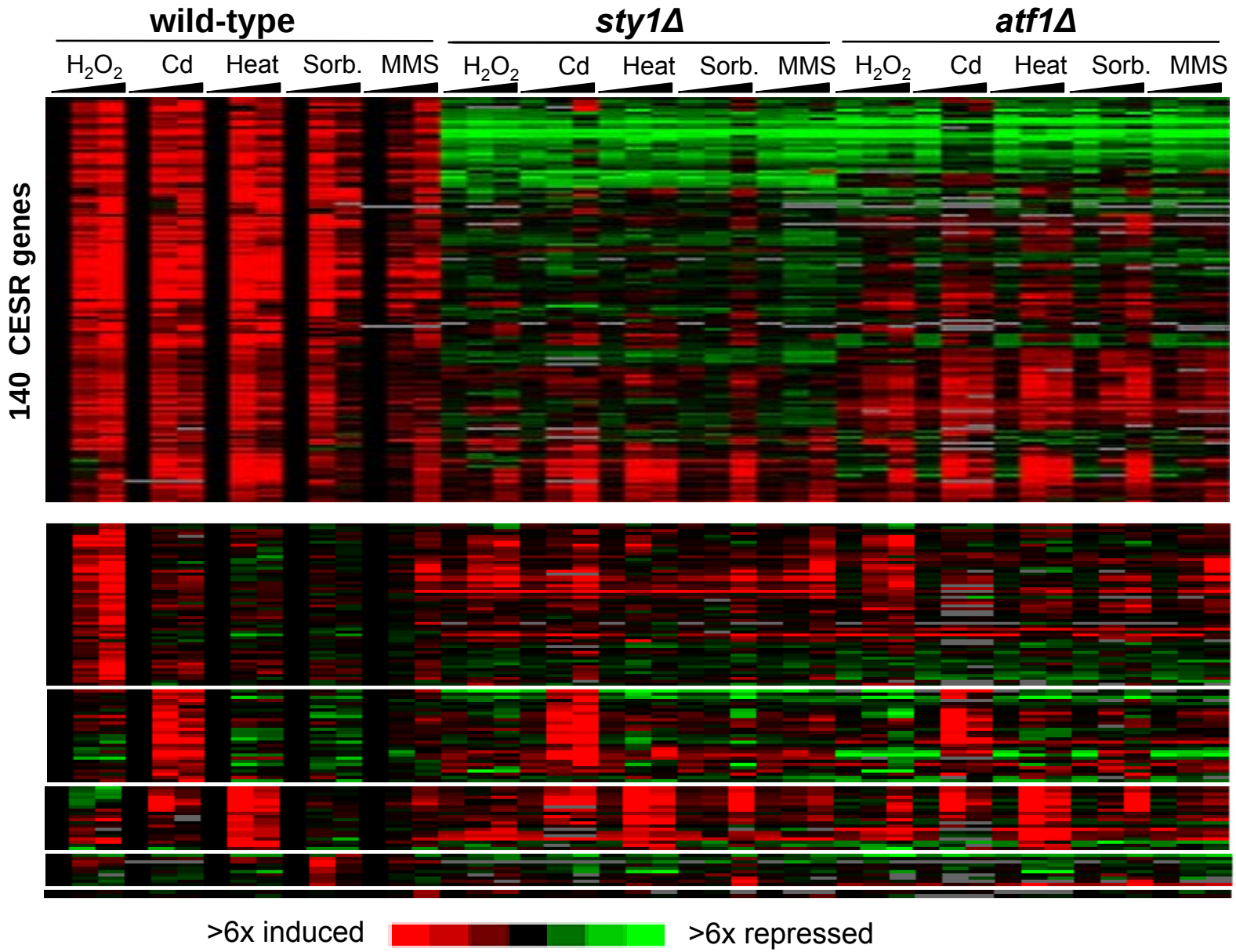


CESR: Core Environmental Stress Response

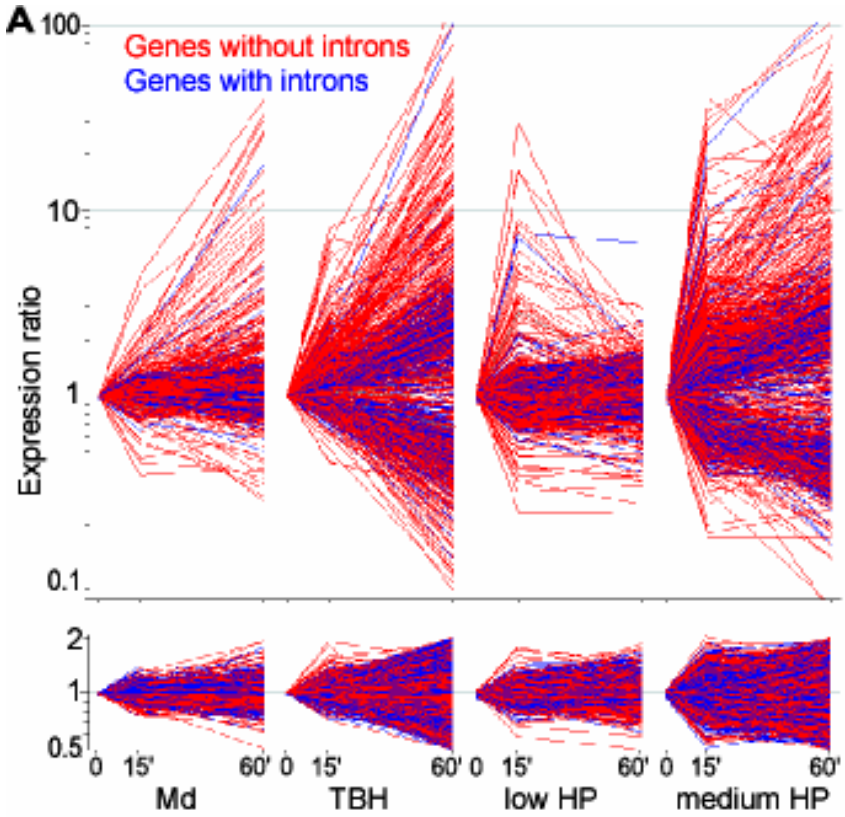


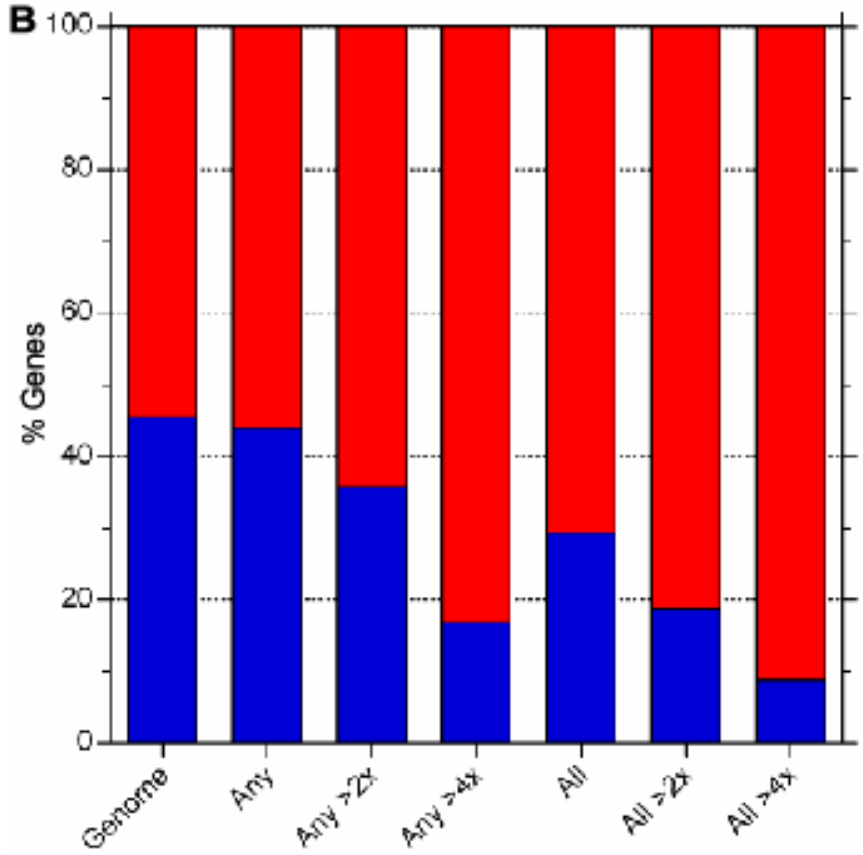


Regulation of stress response genes:



Stress-regulated genes tend to have no introns:





Bias against introns among oxidative stress genes:

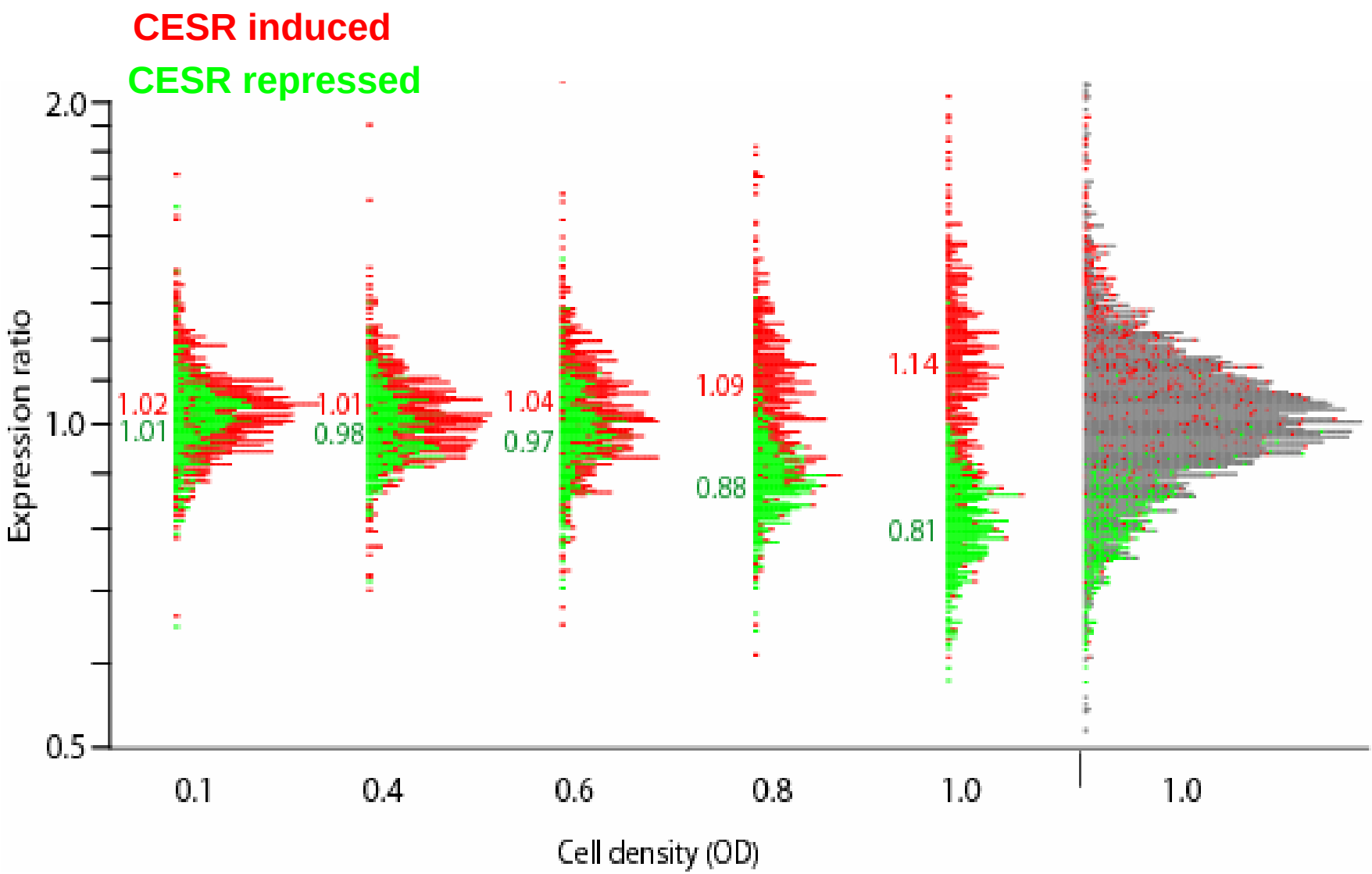
Gene list	Genes with introns	Genes without introns	Proportion of genes with introns	Probability
Genome	2289	2741	45.5%	-
Any	1349	1721	43.9%	0.003
Any >2x	467	840	35.7%	6.7E-17
Any >4x	55	272	16.8%	1.2E-29
Any $\leq$2x	882	881	50.0%	1.0
Any $\leq$4x	1294	1449	47.2%	0.996
All	37	89	29.4%	1.3E-04
All >2	12	52	18.8%	6.1E-06
All >4	3	31	8.8%	4.0E-06

- Oxidative stress response enriched for genes without introns
- Bias most significant for genes that are most highly regulated

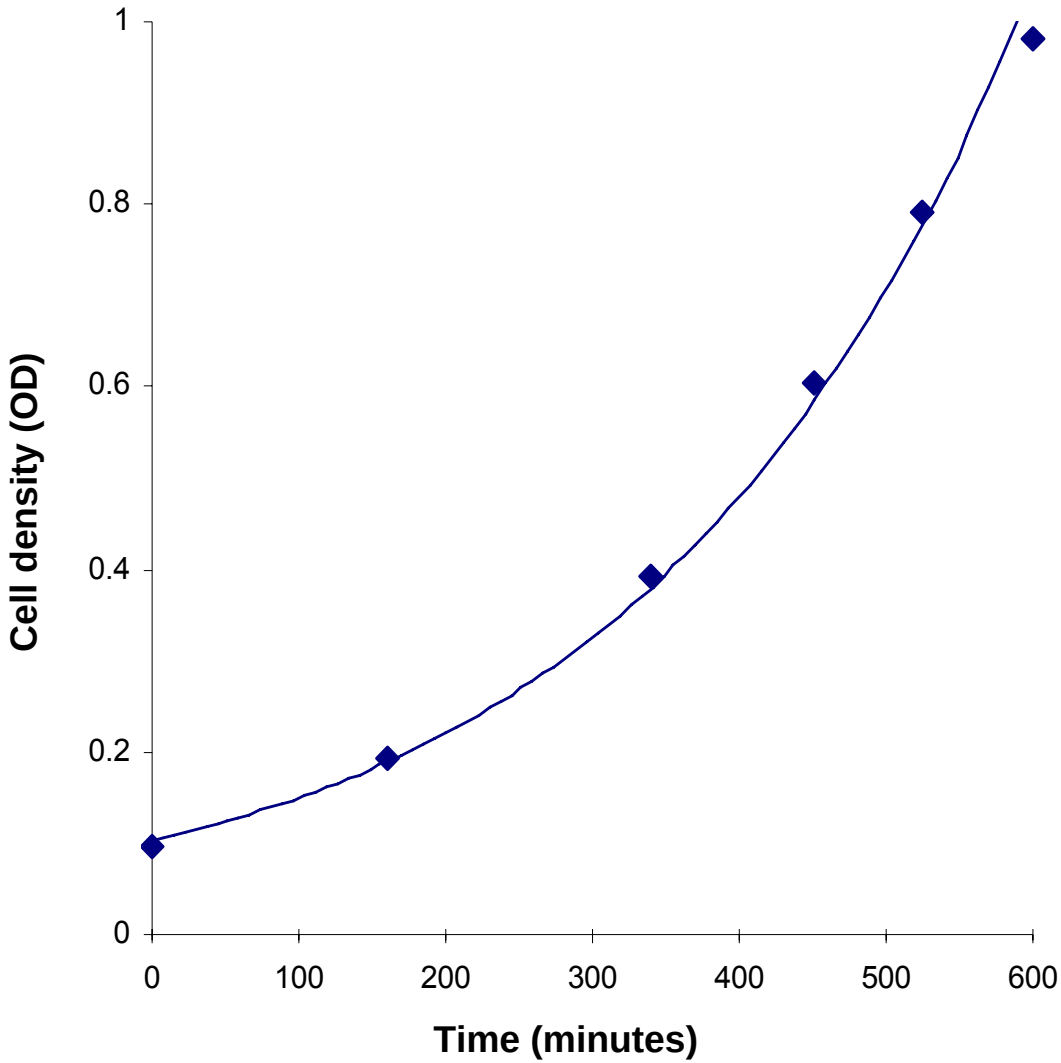
➢ Splicing takes longer than transcription:
Adaptation for fast response?

➢ Splicing sensitive to oxidative stress?

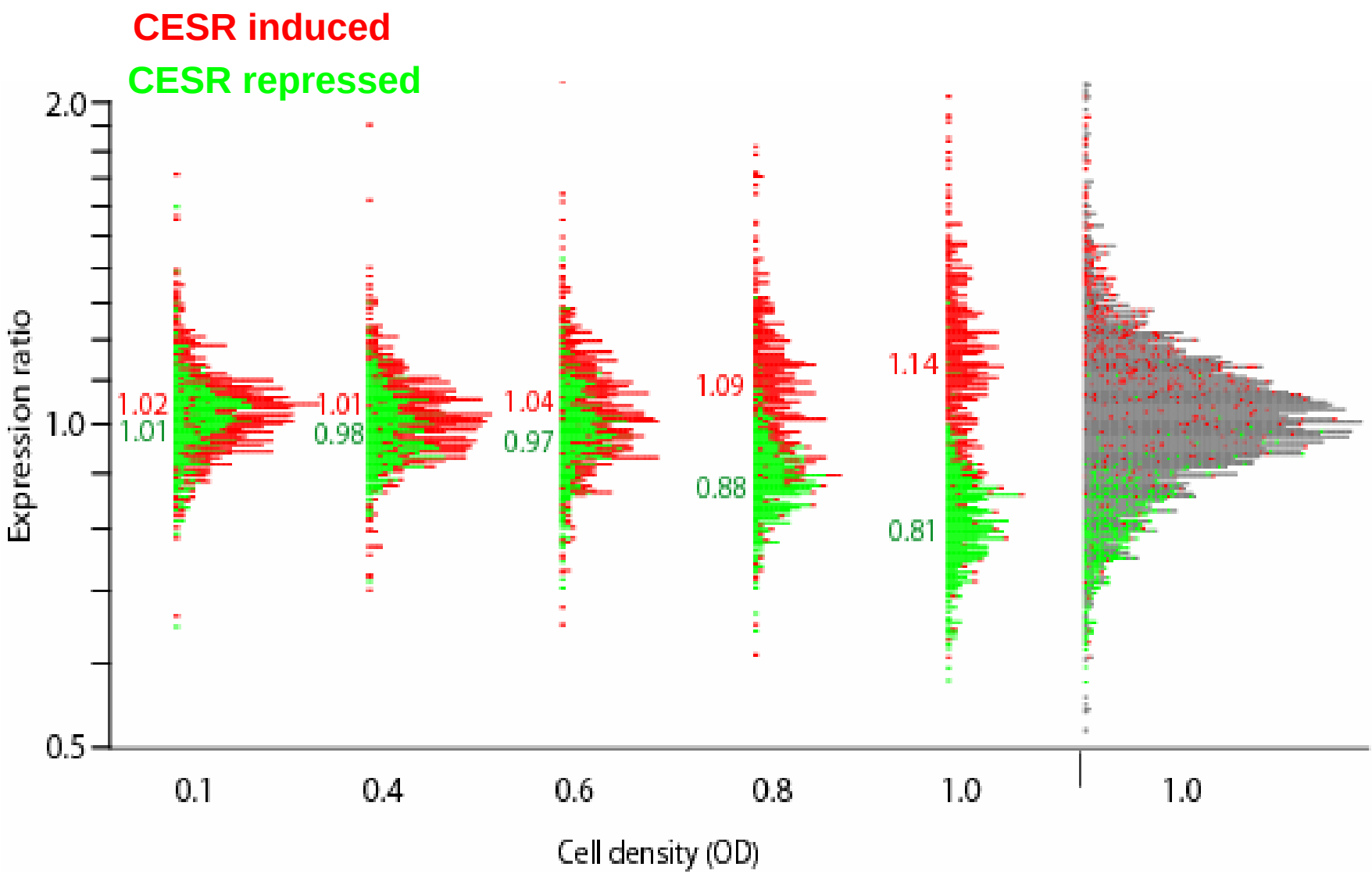
Increasing cell density and stress gene expression:



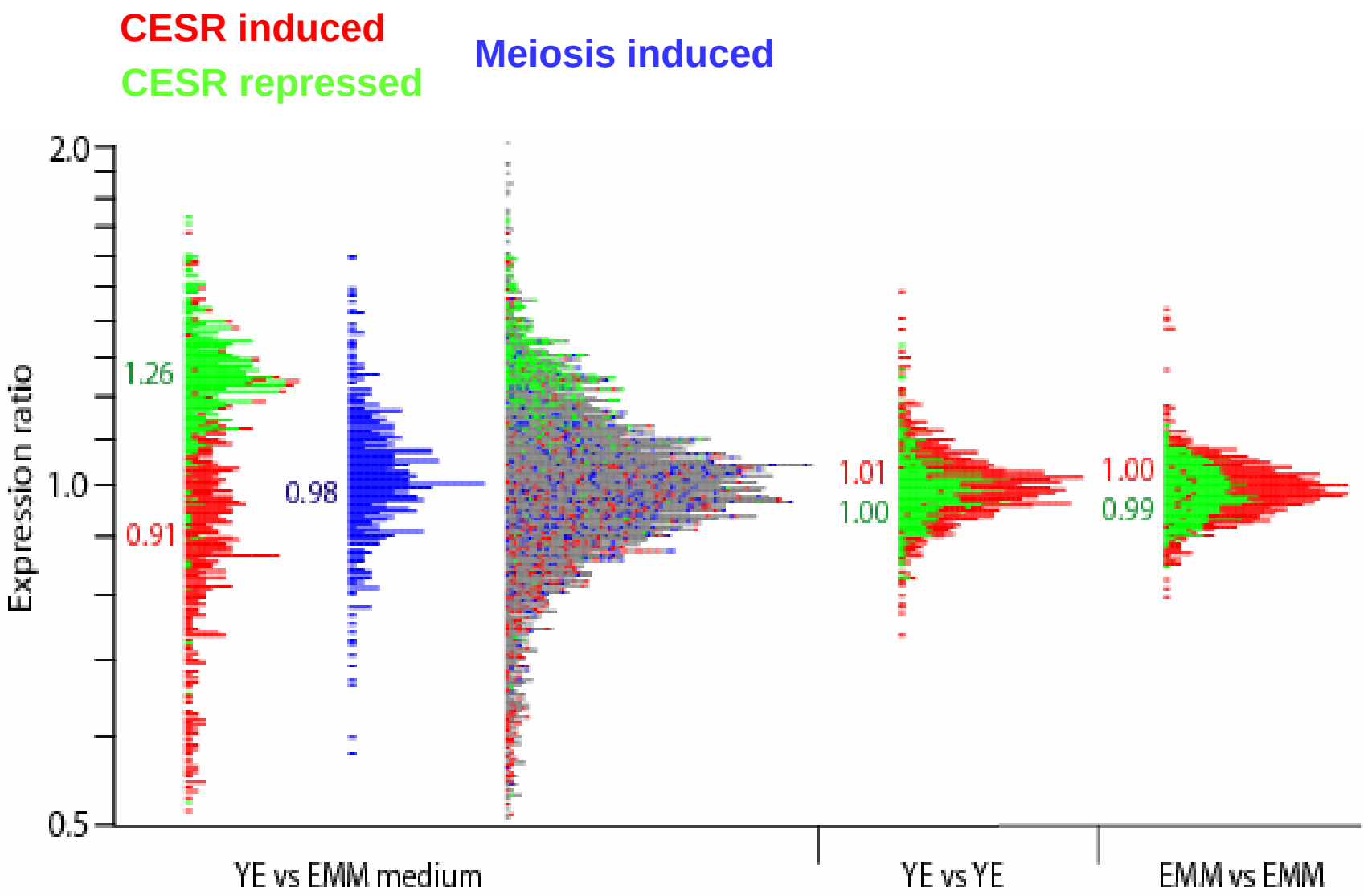
Growth curve:



Increasing cell density and stress gene expression:



Growth media influence stress gene expression:



- **Stress response is not an all or nothing effect**

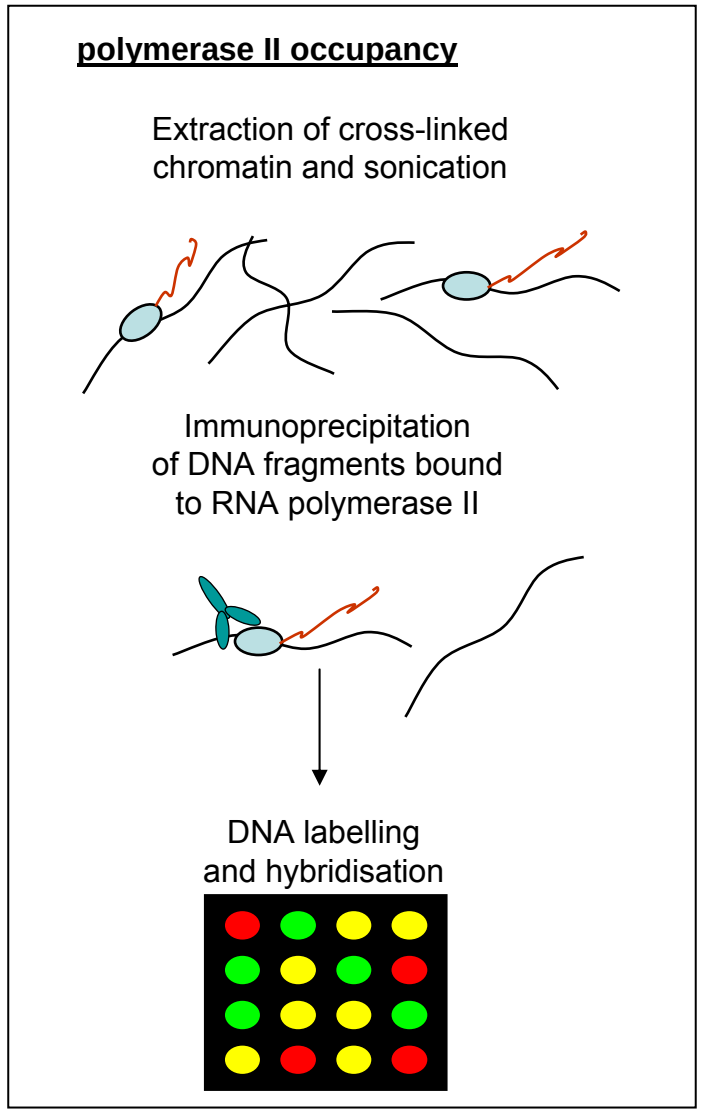
- **Cells fine-tune and adjust regulation in response to subtle changes in environment**

- **Stress response genes show highly variable expression levels reflecting sophisticated regulation to various external factors**

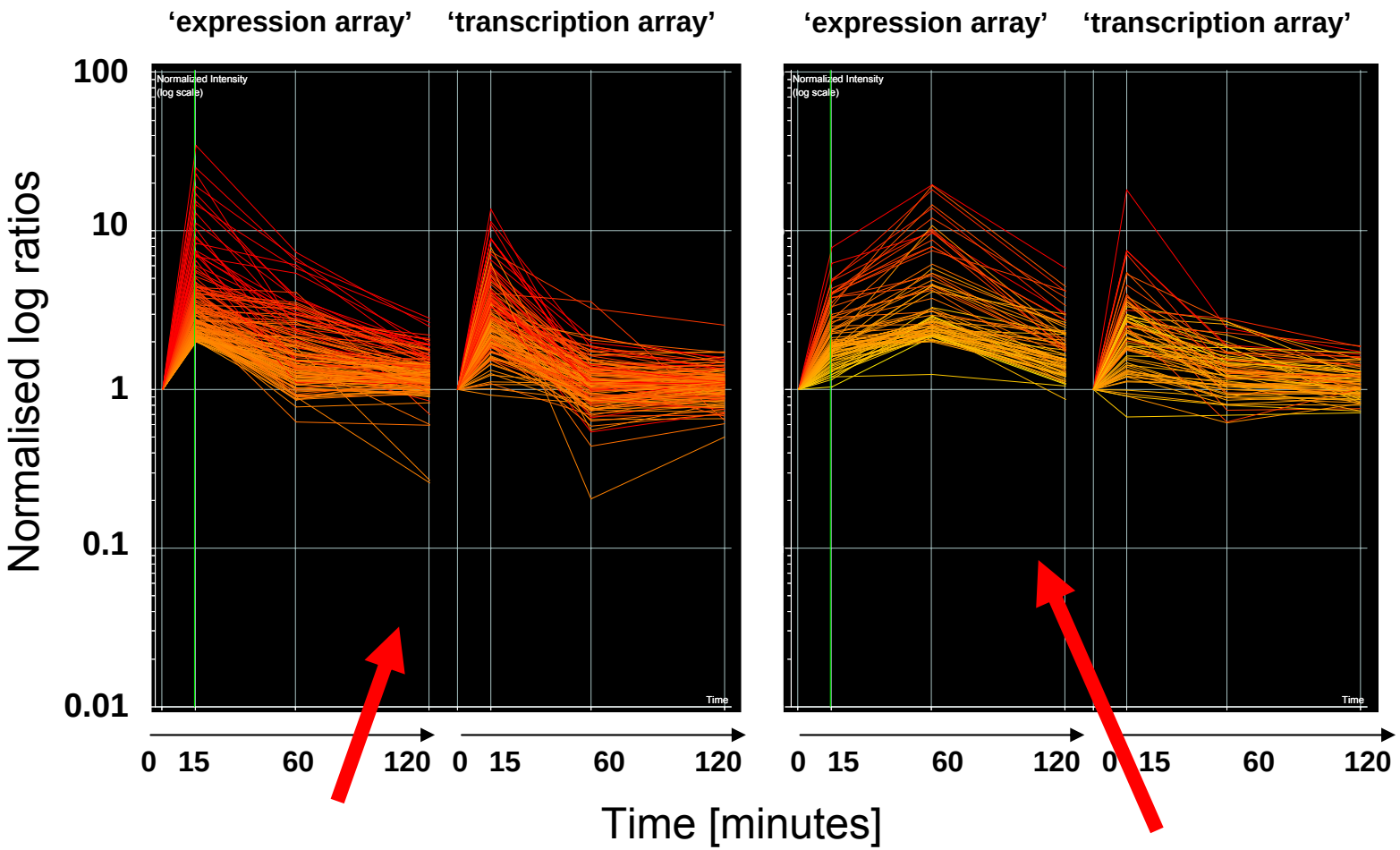
-> tightly controlled conditions
essential for microarray experiments

Microarrays measure mRNA steady-state levels, not transcription

Global estimates of transcriptional efficiency: ChIP-chip with RNA polymerase II



Cause of differences in gene expression profiles?



Expression peaks at 15 minutes
Transcription peaks at 15 minutes

Expression peaks at 60 minutes
Transcription peaks at 15 minutes

General and simple recommendations:

- Repeat biological experiments to get statistically sound data
- Plan and design experiments carefully / controlled and standardized conditions
- Compare data from different experiments / Explore data with various tools

Global Gene Expression Programs in Fission Yeast

http://www.sanger.ac.uk/PostGenomics/S_pombe

Jürg Bähler

Wellcome Trust Sanger Institute / Cancer Research UK

Cell cycle control of gene expression

- Universal level of regulation during cell cycle
- Budding yeast: 400-800 periodically expressed genes
- Fission yeast: ~35 periodic genes reported

➢ Genome-wide overview of cell cycle-regulated gene expression in fission yeast

➢ Conservation of periodic gene expression programs?

Periodic Gene Expression during Cell Cycle:

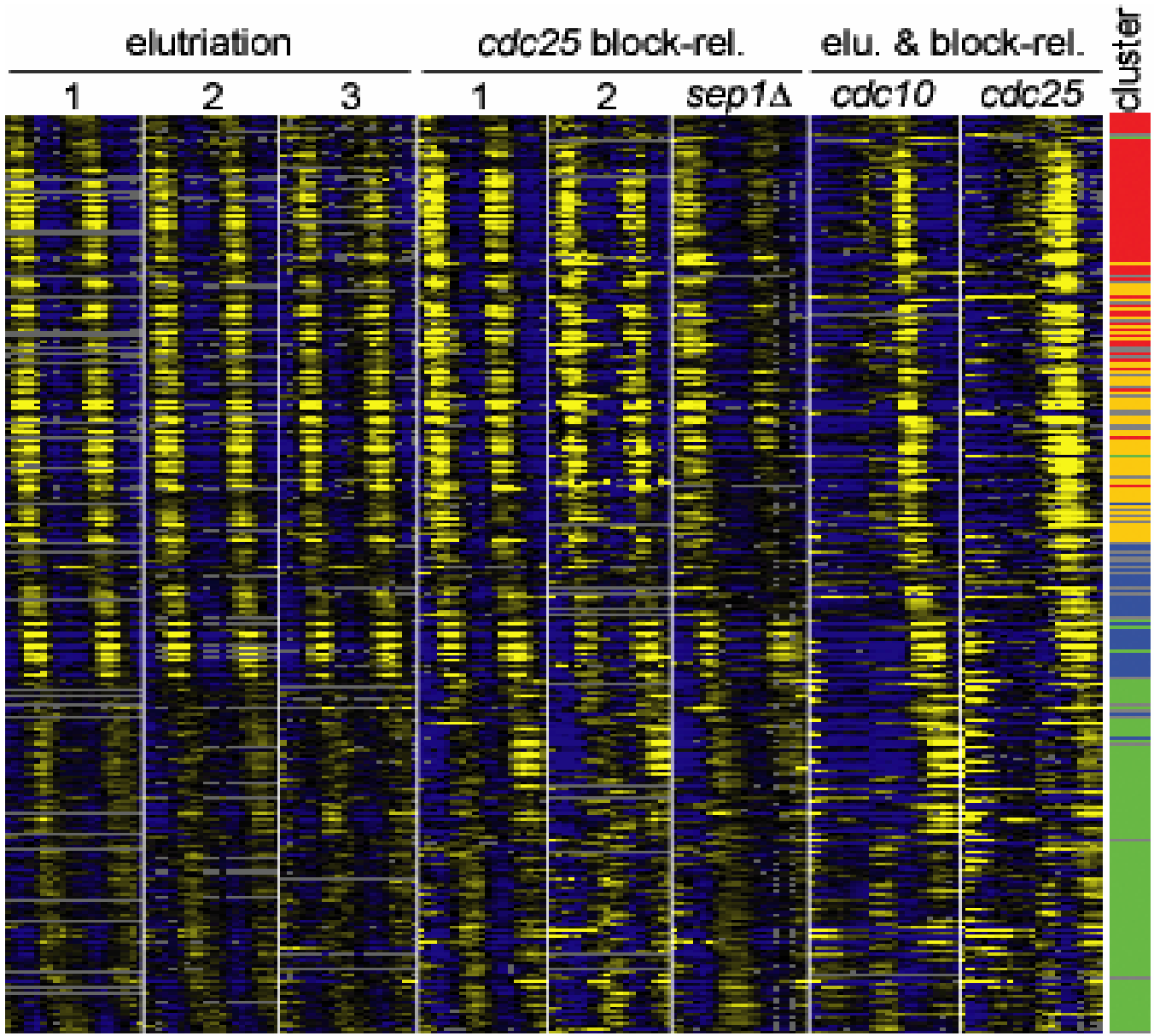
DNA microarrays to study stage-specific gene expression during cell cycle

Synchronization of cells in different ways:

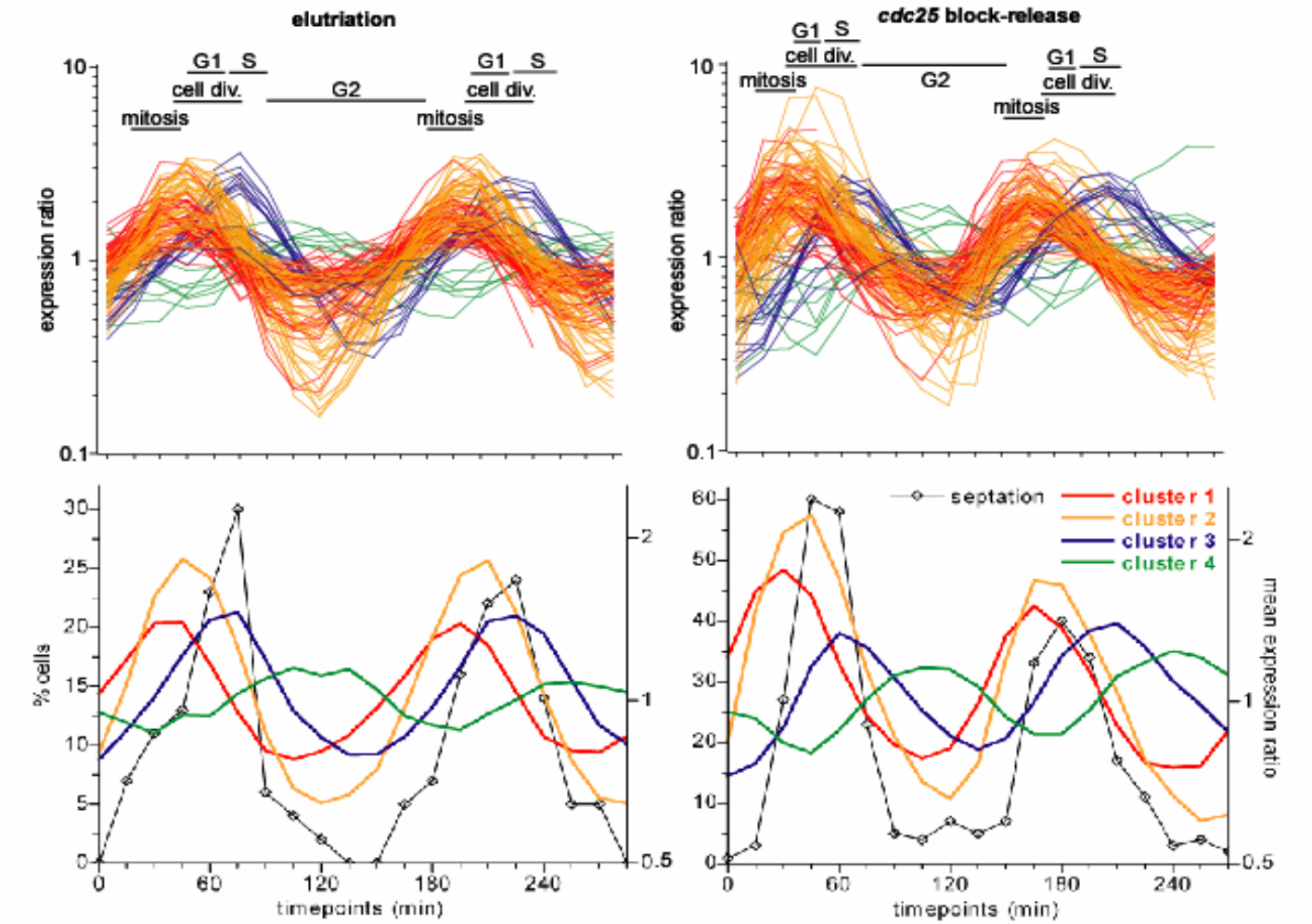
- centrifugal elutriation
- conditional cell cycle mutants

Rustici et al. (2004) *Nat Genet* 36:809

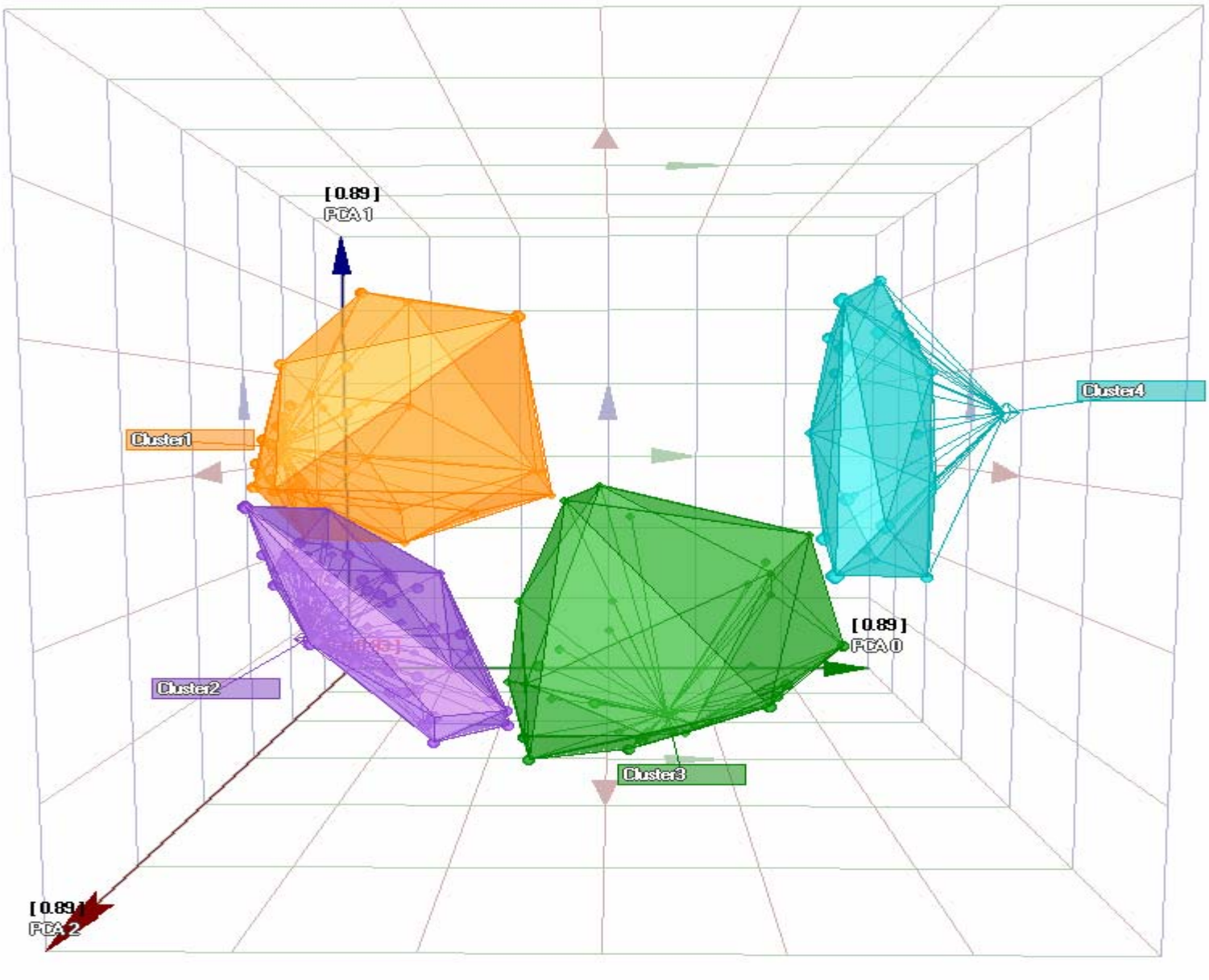
Phaseogram: ~400 periodically expressed genes,



Clustering: 4 major waves of gene expression



Principal component analysis:

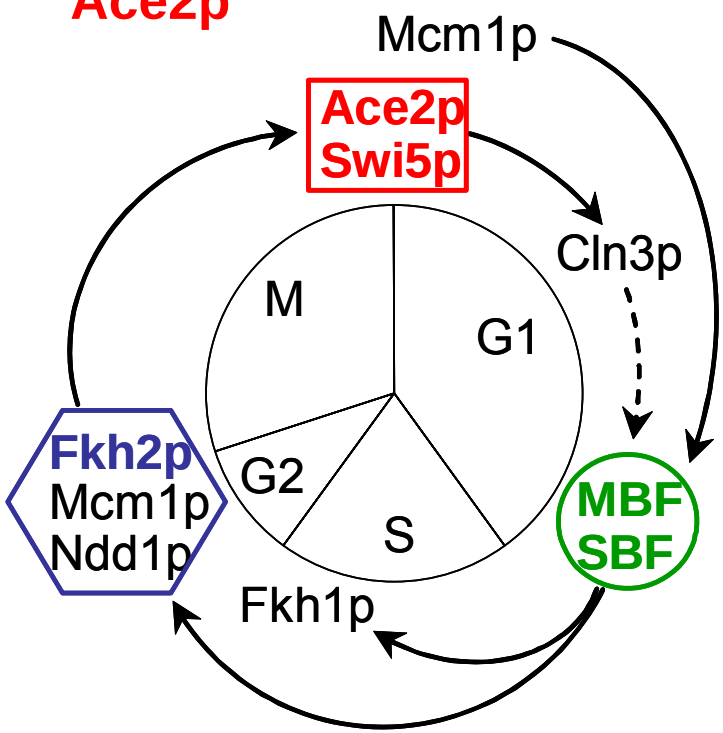


Major cell cycle transcription factors

Budding yeast:

?

Ace2p



Forkhead

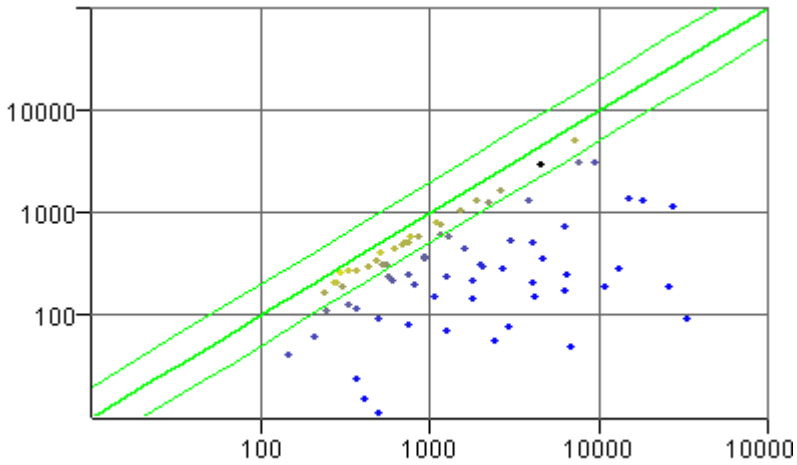
Sep1p
Fkh2p

MBF

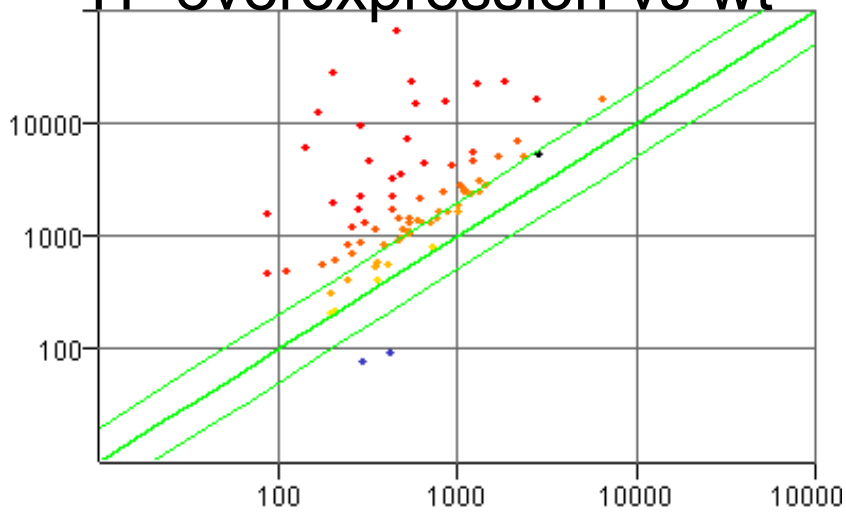
E2F

Regulation of periodically transcribed genes

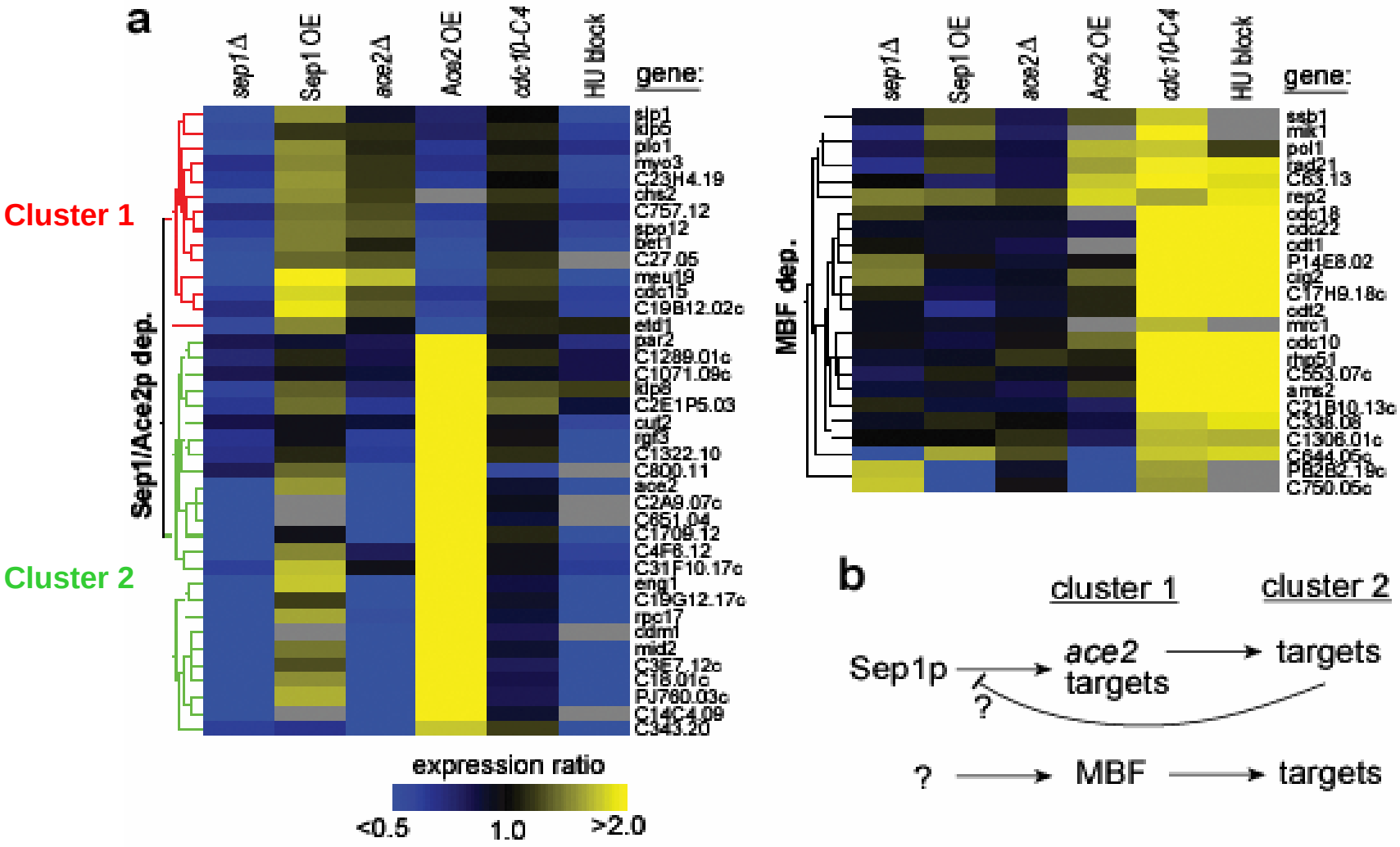
TF deletion vs wt



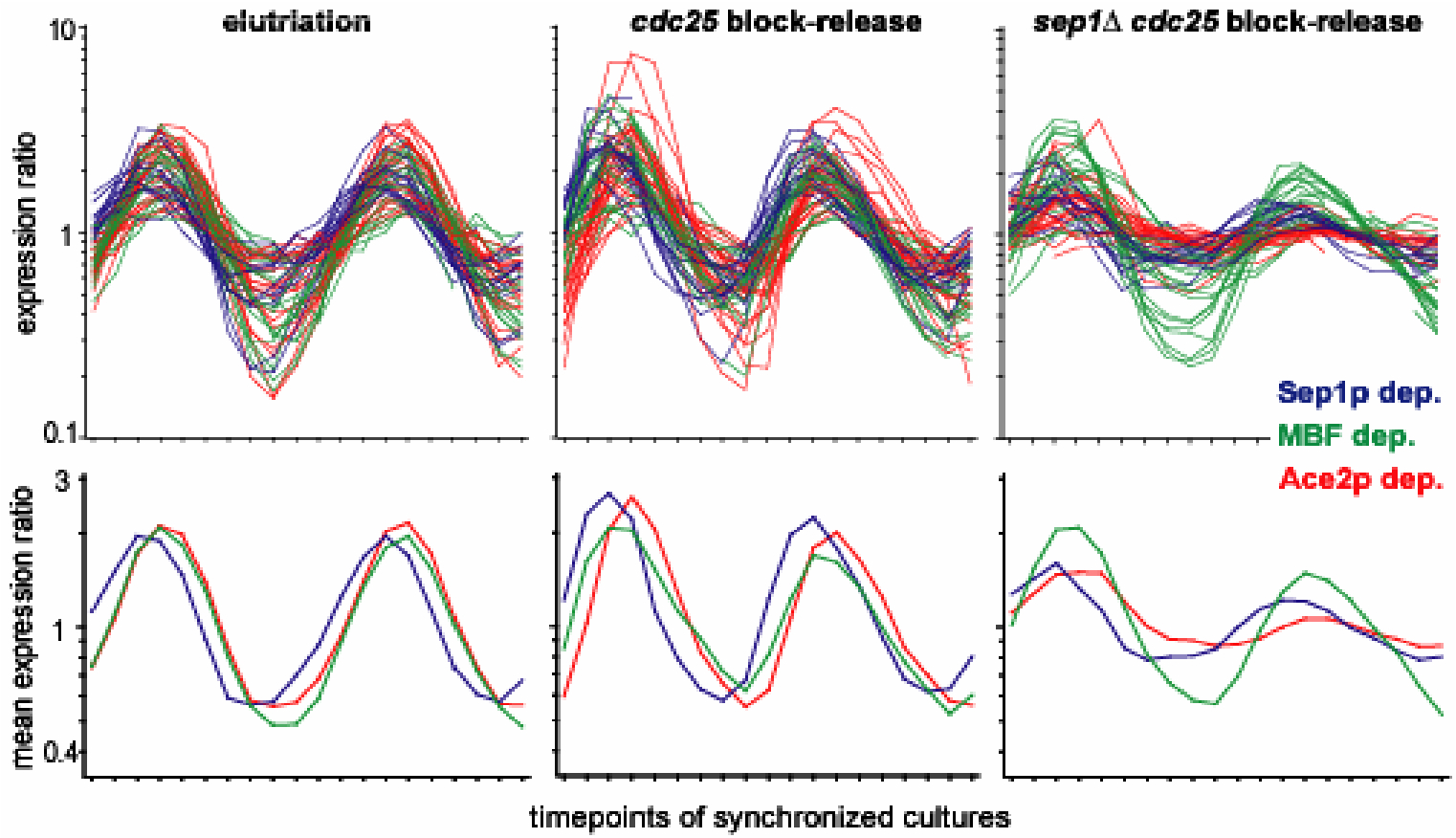
TF overexpression vs wt



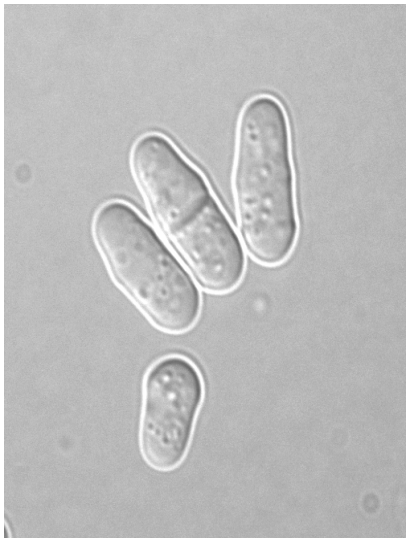
Transcriptional regulation of clusters 1 and 2



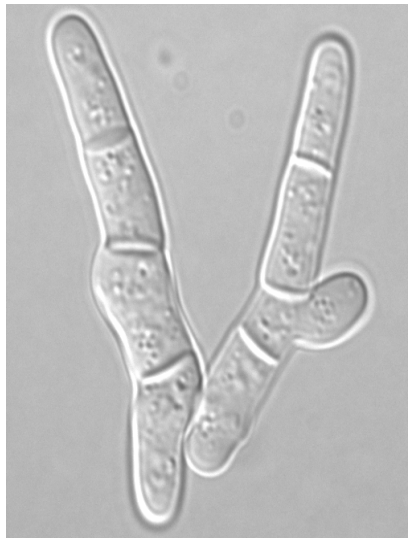
Transcriptional regulation by Sep1p and Ace2p:



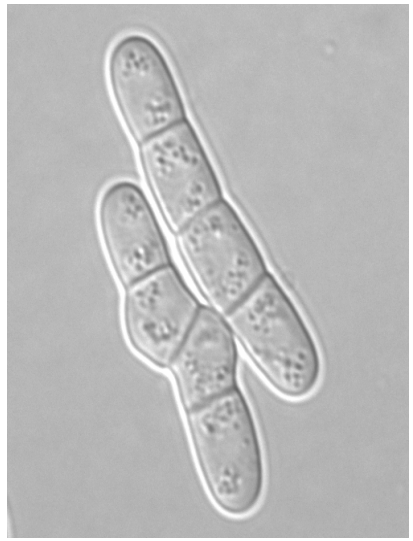
Mutant phenotypes



wild type



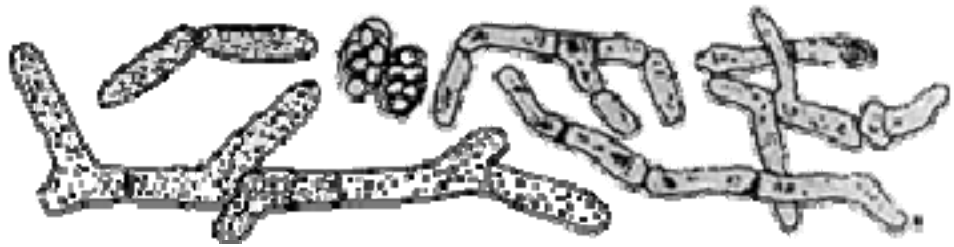
sep1Δ



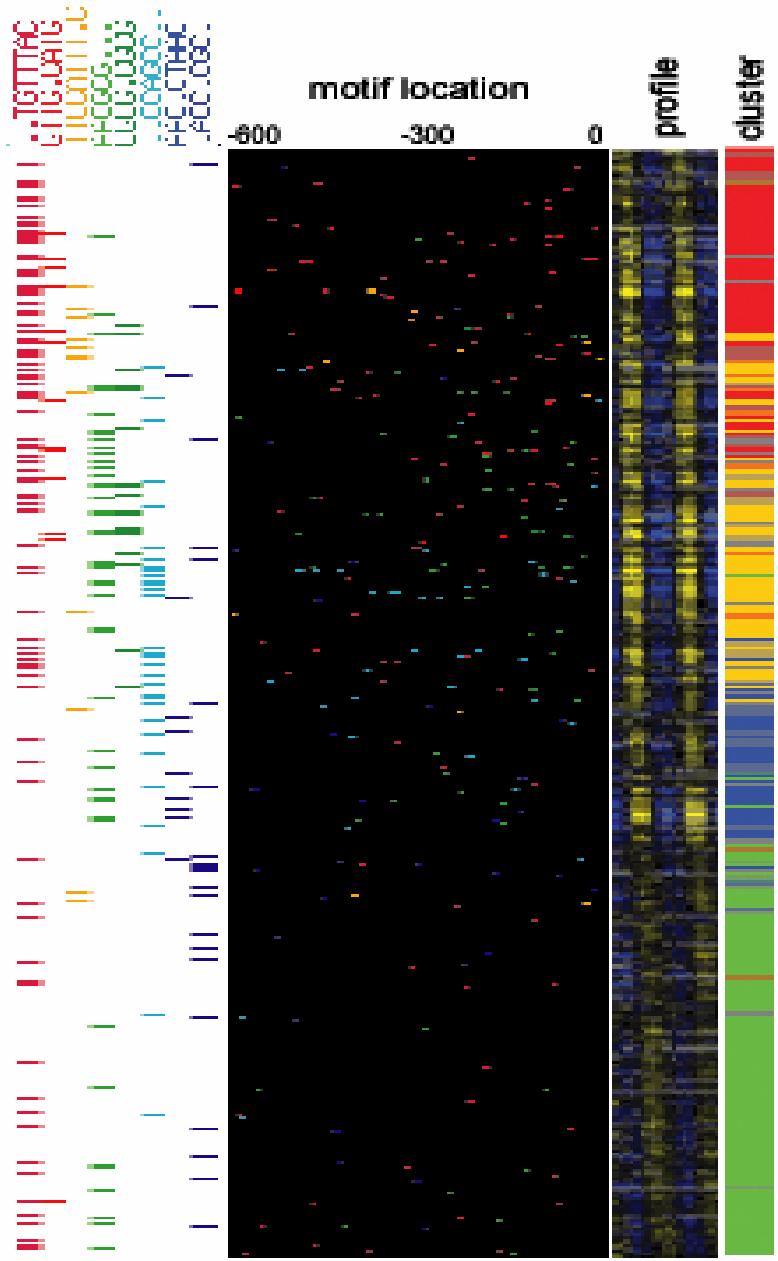
ace2Δ

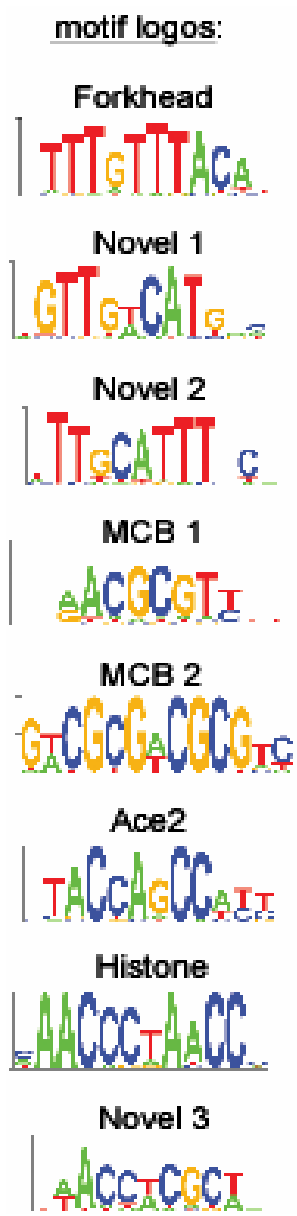
ace2Δ sep1Δ

P. Lindner, 1893



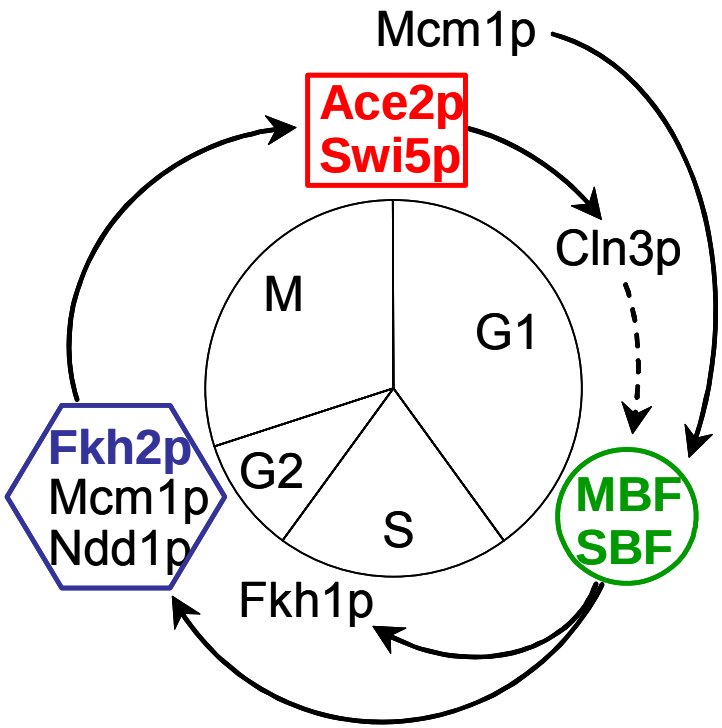
Identification of regulatory promoter motifs:



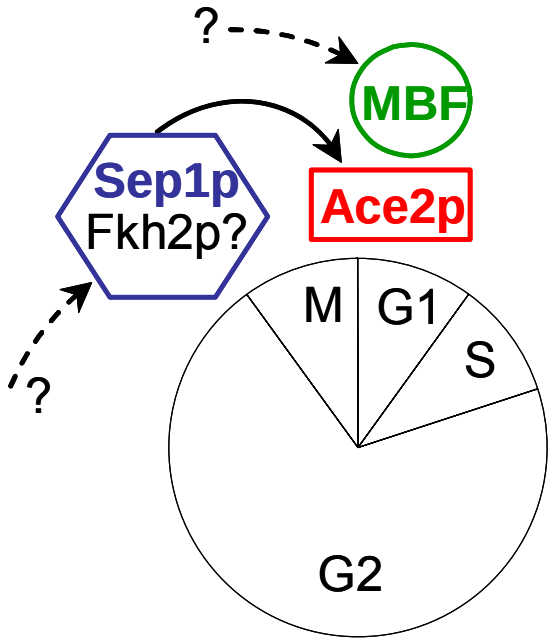


Regulatory gene expression networks:

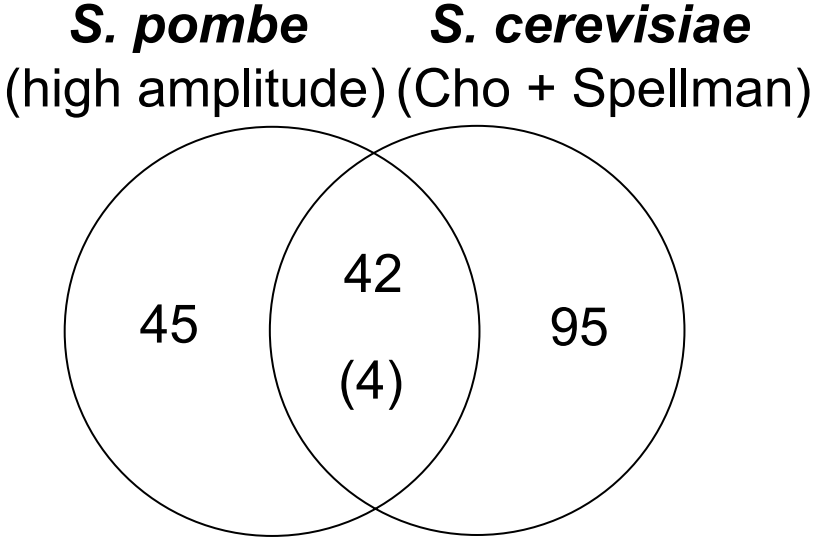
Budding yeast:



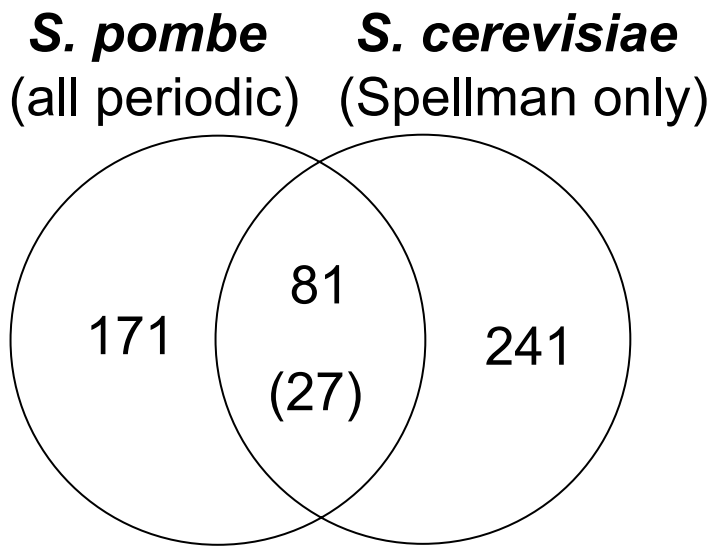
Fission yeast:



Core cell cycle-regulated genes:



2981 orthologs in total



Core cell cycle-regulated genes:

Mitosis and cell division:

CDC5, IPL1, KIN3	*plo1, ark1, fin1*	Polo, Aurora, and NimA kinases
CDC20	*slp1*	Activator of APC
SPO12	*wis3*	Putative cell-cycle regulator
KAR3, KIP1	*klp5, klp6, klp8*	Kinesin microtubule motor
MOB1, DBF2	*mob1, sid2*	Proteins involved in MEN/SIN
MYO1	*myo3*	Myosin II heavy chain
BUD4	*mid2*	Protein involved in cytokinesis
ACE2	*ace2*	Transcription factor
HOF1	*imp2*	Protein involved in cell division
DSE4	*eng1*	Glucanase for cell separation
CHS2	*chs2*	Protein involved in septum formation
TOS7	*mac1*	Putative role in cell separation

Core cell cycle-regulated genes:

DNA replication:

POL1, POL2	*pol1 and cdc20*	DNA polymerases α and ε
RFA1	*ssb1*	Single-stranded DNA-binding protein
CDC6	*cdc18*	Regulator of DNA replication initiation
MRC1	*mrc1*	DNA replication checkpoint protein
RNR1	*cdc22*	Ribonucleotide reductase
SMC3, MCD1	*psm3 and rad21*	Cohesins
HTZ1	*pht1*	Histone variant
8 histone genes	9 histone genes	Histones H2A, H2B, H3, and H4

Others:

mik1	*SWE1*	Kinase inhibiting cyclin-dependent kinase
cig2	*CLB1-CLB6*	B-type cyclins
msh6	*MSH6*	Mismatch-repair protein
rhp51	*RAD51*	DNA repair protein

Human and fission yeast only:

cdc2 kinase and *cdc25* phosphatase

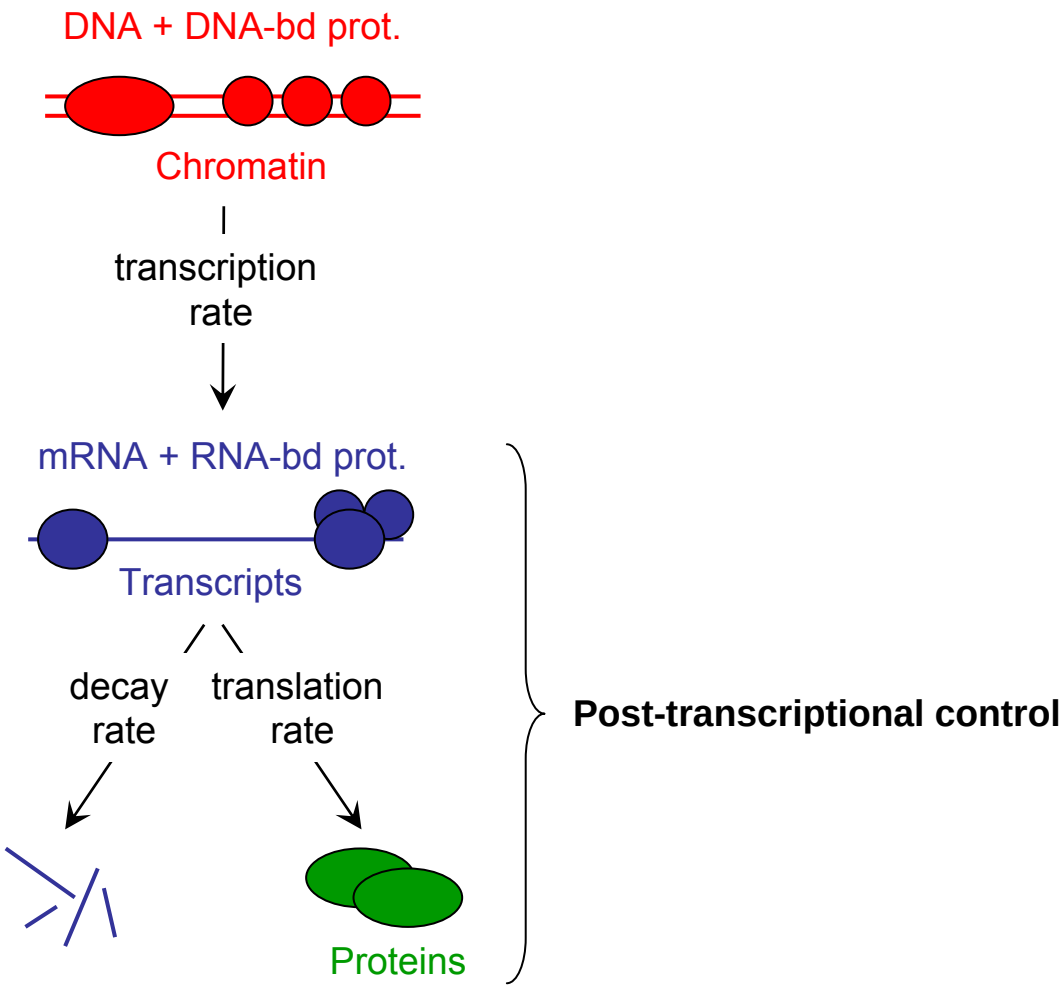
Human and budding yeast only:

MCM complex DNA replication genes

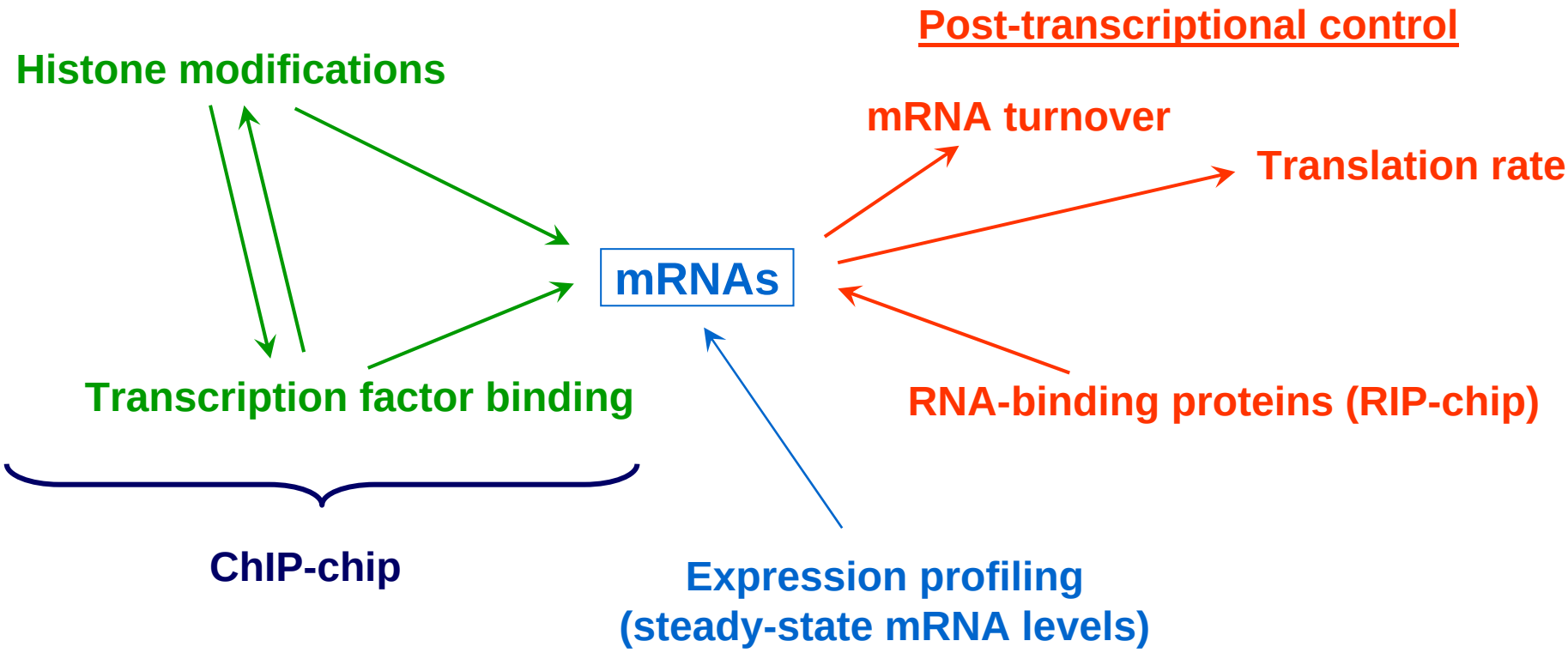
Major Conclusions

- 4 major waves of transcription,
  ~400 periodic genes (8% of genome)
- Conserved transcription factors but differences in regulatory circuits between fission and budding yeasts: rewiring during evolution to accommodate different cell cycle phases
- Periodic transcription not necessarily conserved, but core set of universally regulated genes with basic functions in cell cycle progression

Regulation of Gene Expression at Multiple Levels



Global data on different layers of gene expression control

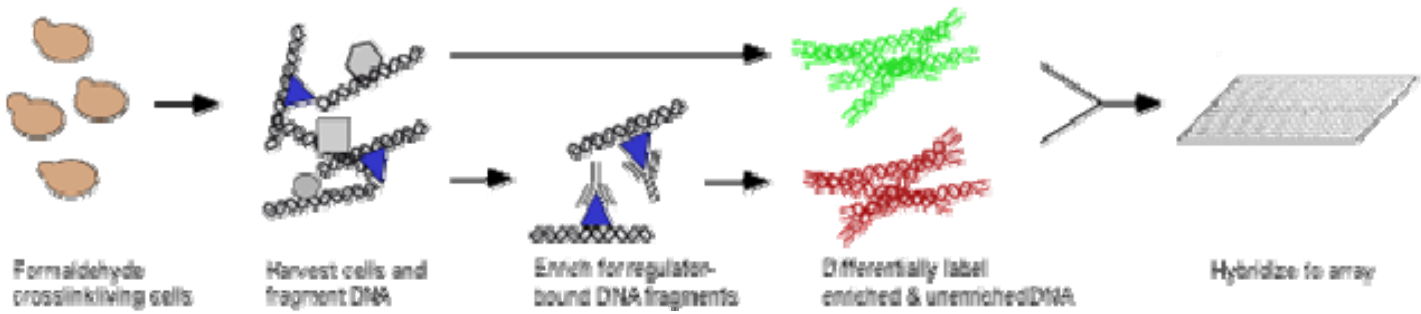


Microarray-based approaches to measure control at multiple levels

- **Integrate data from different levels of regulation**
- **Dynamic changes in regulation, genetic and environmental perturbations**

ChIP-on-chip:
Direct interaction of TFs with promoters

1. Tag TF (HA, TAP, by homologous recombination in the yeast)
2. 'Cross-link' the TF to chromatin
3. Immunoprecipitation of TF-chromatin complex
4. Label DNA and hybridise to array of intergenic regions



R. Young

ChIP-on-chip
(global mapping of transcription factor binding sites)

Vs

Expression profiling
(transcription factor mutant/overexpressor vs wild type?

Expression profiling: direct vs indirect effects?
functional role

ChIP-chip: direct
functional?
divergent genes?

- **Integrate** expression and ChIP-chip data

wellcome trust sanger institute

Thank you!

Gabriella Rustici
Daniel Lackner
Samuel Marguerat
Juan Mata
Val Wood
Brian Wilhelm
Chris Penkett
Stephen Watt
Luis López-Maury
Falk Schubert
Sofia Aligianni
Tannia Gracia

Victor Chang Cardiac Research Institute, Sydney, Australia
Thomas Preiss
Traude Beilharz

CANCER RESEARCH UK