

Measuring & Modeling Cellular Metabolic & Regulatory Networks

Metabolic Engineering Working Group (MEWG)
May 31, 2000 sp

We thank for support:

Government and private grant agencies:

NSF, ONR, DOE, DARPA, HHMI, Lipper, NIST

Corporate collaborators & sponsors:

Affymetrix, GTC, Pangea, Mosaic, Aventis

Lipper Center for Computational Genetics

1930

Search the WWW-ome:

1916

Genome DNAs	Transcriptome RNAs	Proteome Proteins	Physiome Metabolites	Biome Environments
DNA Motifs for 17 genomes (Aug'99)	Gene clustering			ExpressDB (Aug'99)
Genome sequence comparisons	Yeast RNA quant. (Oct'98)	E.coli protein quant.	LC-ESI-MS	Conditions & genotypes: Biomolecule Interaction Growth, Expression Database (BIGED)
New DNA sequencing methods	DNA Motif software (Oct'98)	LC-ESI-MS		Natural antibiotics
E.coli in-frame genome engineering (Dec'98): pKO3	E.coli DNA motifs (Aug'98)	Subcellular compartments	Network models	E.coli multiplex competition selections (phenome)

arep.med.harvard.edu

Lab Members: Aach@rascal.med.harvard.edu, TChen@salt2.med.harvard.edu, Cohen@rascal.med.harvard.edu, Church@arep.med.harvard.edu, Adnan@genetics.med.harvard.edu, WRindone@arep.med.harvard.edu, Philip_Juels@harvard.edu, Lam@genetics.med.harvard.edu, Ralston@rascal.med.harvard.edu , Steffen@rascal.med.harvard.edu, Reyes@arep.med.harvard.edu, hxhua@genetics.med.harvard.edu, vasudeo@genetics.med.harvard.edu, dudley@rascal.med.harvard.edu, sudarsanam@rascal.med.harvard.edu, VKMootha@yahoo.com, JEdwards@bioeng.ucsd.edu, Cheng@cheng.med.harvard.edu, MLBulyk@rascal.med.harvard.edu , PEStep@rascal.med.harvard.edu, JHughes@fas.harvard.edu, JJJohnson@fas.harvard.edu, AMcguire@fas.harvard.edu, RMitra@rascal.med.harvard.edu , Phillips@rascal.med.harvard.edu, Selinger@rascal.med.harvard.edu , Tavazoie@rascal.med.harvard.edu, Rebecca_Spencer@student.hms.harvard.edu, DJanse@student.med.harvard.edu, JShendur@student.med.harvard.edu, Petti@fas.harvard.edu, NReppas@fas.harvard.edu, KCheung@fas.harvard.edu



gcggatttaqctcagt

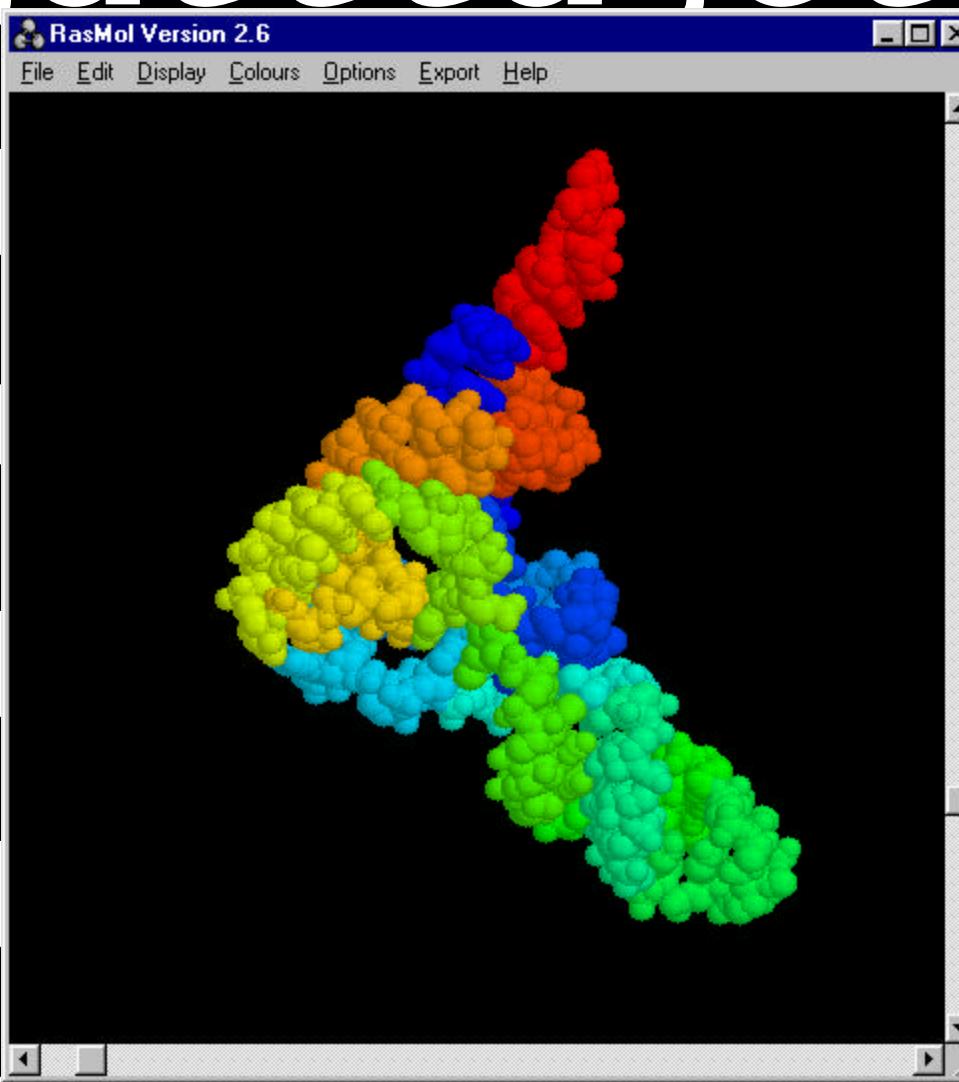
tggg

caga

tttg

tgtg

cacagaatttcgcacca



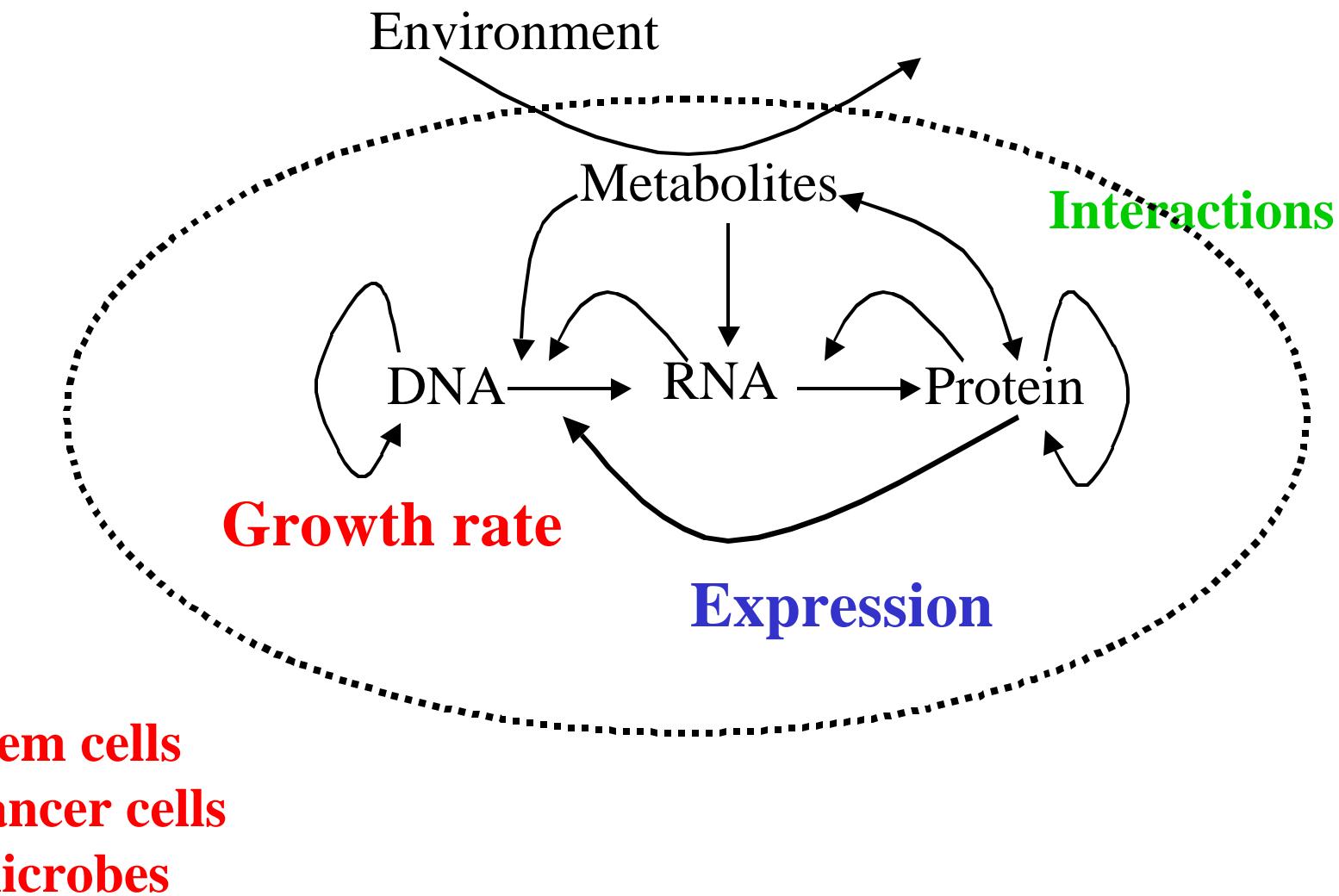
gcgc

aaga

gtcc

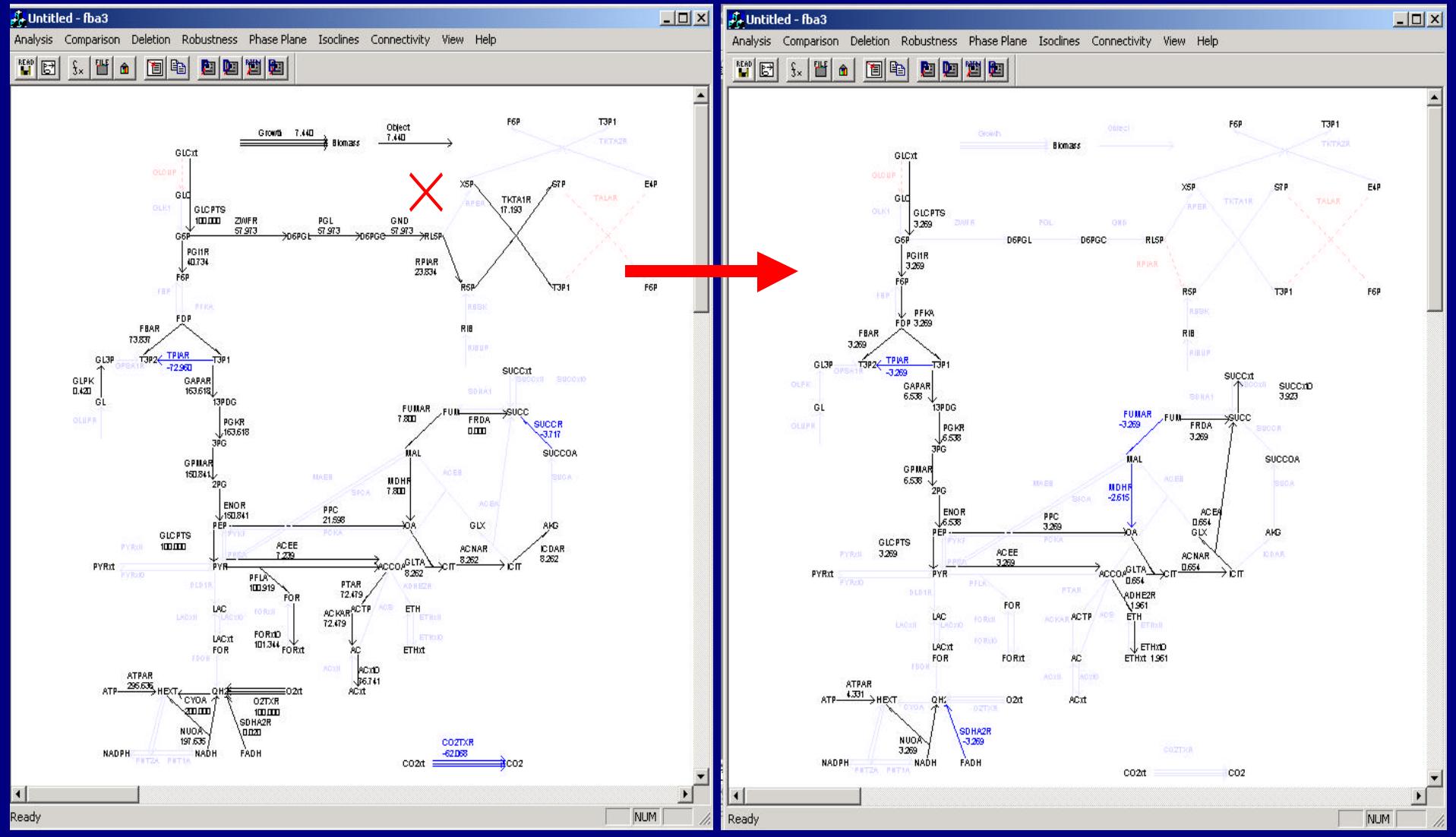
gatc

Network genomics



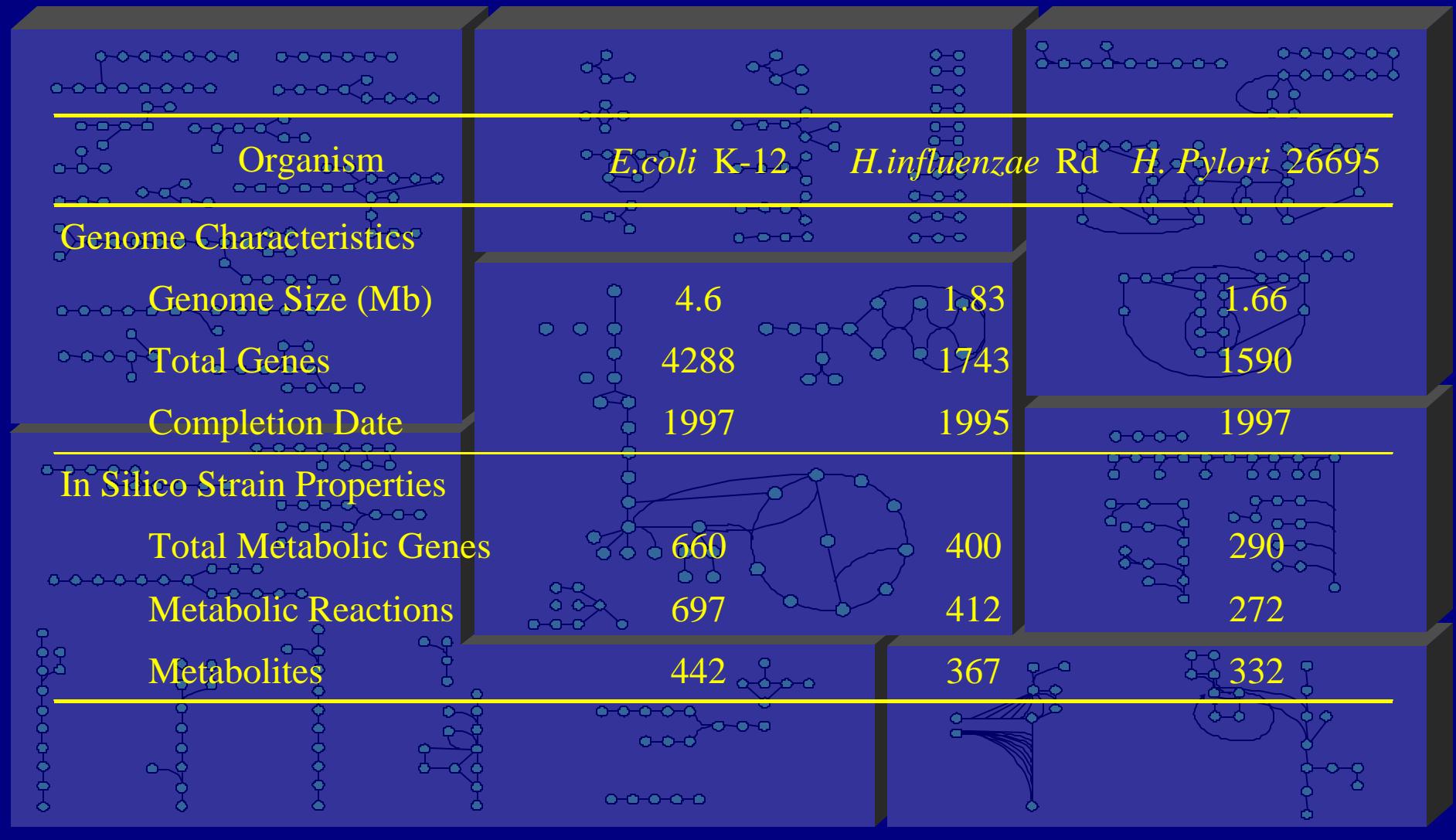
Goal #2: Enzyme gene deletions

Jeremy Edwards & Dereth Phillips
with Bernard Palsson group at UCSD

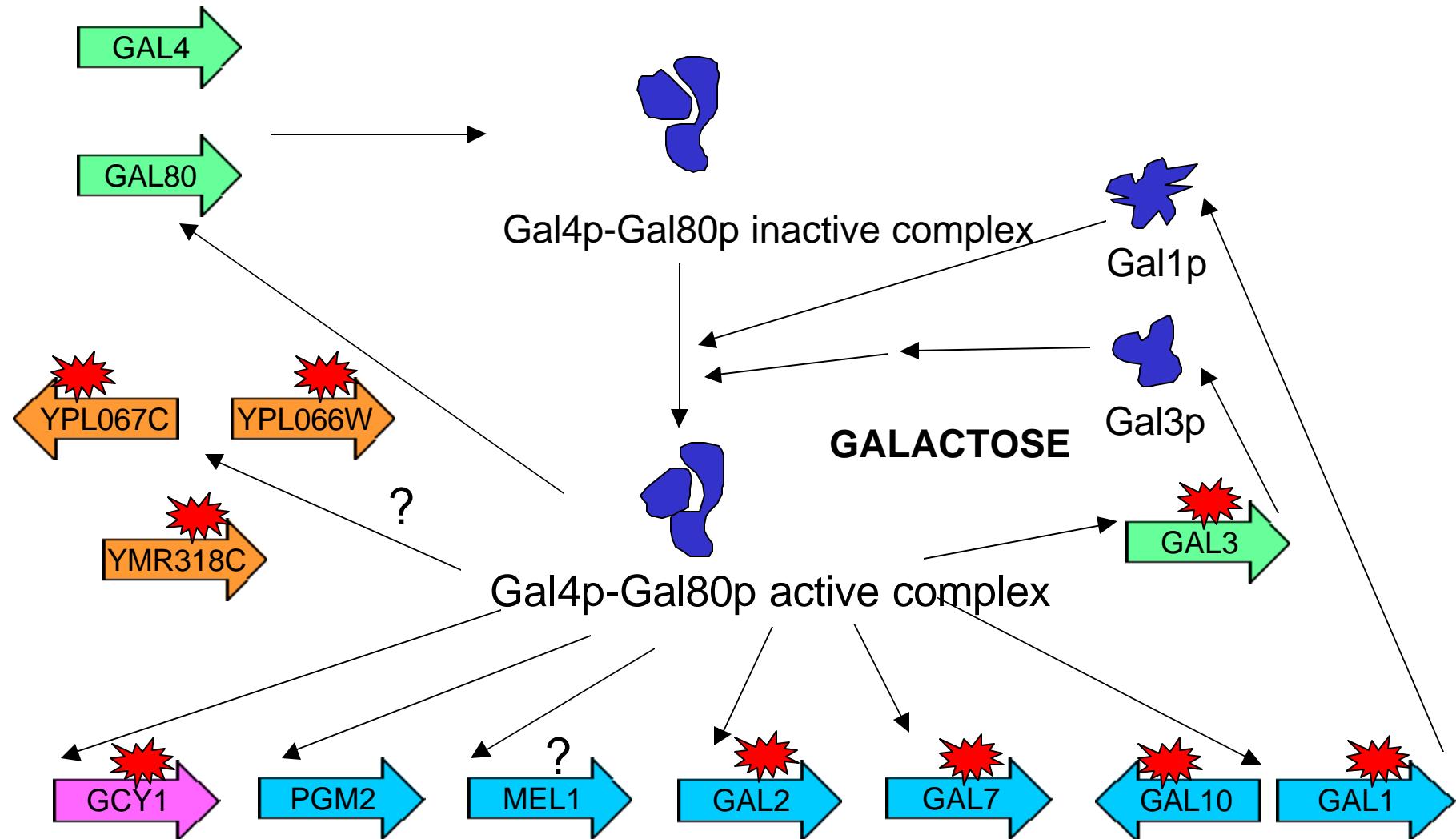


Goal #1:Reconstructing Metabolic Networks

*Tzachi Pilpel, Felix Lam, Saeed Tavazoie
with Bernard Palsson group at UCSD*



Galactose Regulatory Network



Roth, et al. Nat Biotech. 16, 939 (1998) Nat Genet 1999 Jul;22(3):281-5

RNA quantitation

(Frequently Asked Questions)

Is less than a 2-fold RNA-ratio ever important?

Yes; 1.5-fold in trisomies.

Why oligonucleotides rather than cDNAs?

Alternative splicing, 5' & 3' ends; gene families.

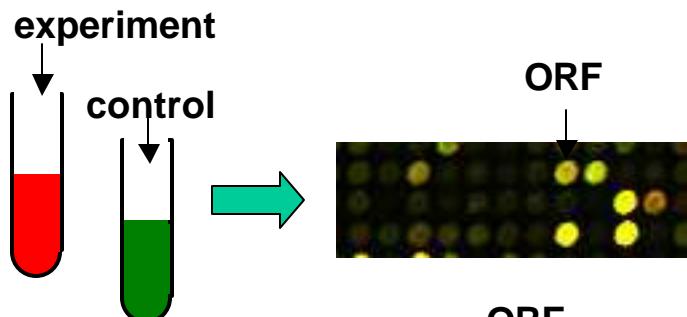
What about using a subset of the genome?

It makes trouble for later (meta) analyses.

ExpressDB: meta-analyses

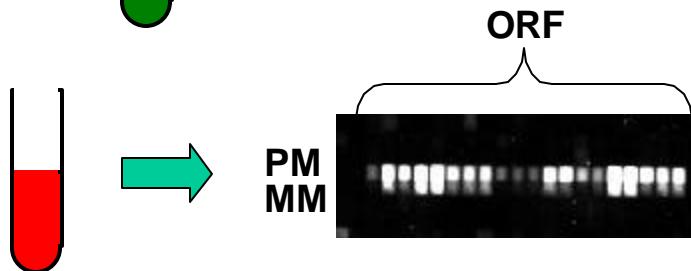
Aach, Rindone

- Microarrays¹



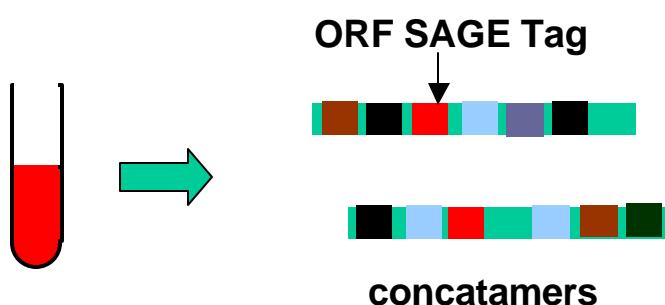
- R/G ratios
- R, G values
- quality indicators

- Affymetrix²



- Averaged PM-MM
- “presence”
- feature statistics
- 25-mers

- SAGE³



- Counts of SAGE 14-mers sequence tags for each ORF

¹ DeRisi, et.al., *Science* **278**:680-686 (1997)

² Lockhart, et.al., *Nat Biotech* **14**:1675-1680 (1996)

³ Velculescu, et.al, Serial Analysis of Gene Expression, *Science* **270**:484-487 (1995)

Genomic oligonucleotide microarrays

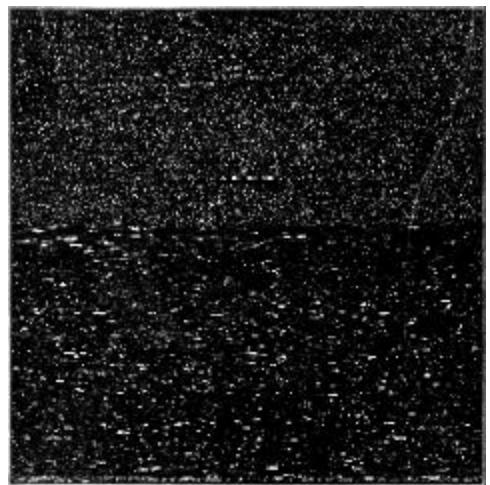
295,936 oligonucleotides (including controls)

Intergenic regions: ~6bp spacing Genes: ~70 bp spacing

Not polyA (or 3' end) biased

Strengths: Gene family paralogs,
RNA fine structure (allelic variation, adjacent promoters),
untranslated and antisense RNAs, DNA-protein interactions.

E. coli
25-mer array



Protein coding
25-mers

Non-coding sequences
(12% of genome)

tRNAs, rRNAs

Affymetrix: Mei, Gentalen,
Johansen, Lockhart(Novartis Inst)

HMS: Church, Bulyk, Cheung,
Tavazoie, Petti, Selinger

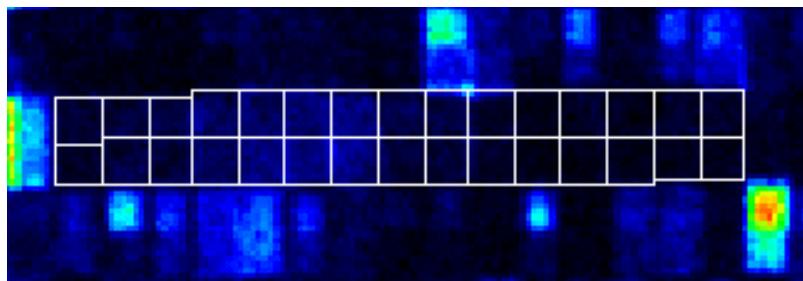
Sensitivity of Different Abundance Metrics

Metric	Scored Present	Negative Scores
Smax		
Log	85.6%	312
Stationary	95.4%	18
Median		
Log	20.0%	1,208
Stationary	90.5%	20
AD		
Log	18.2%	997
Stationary	87.3%	43

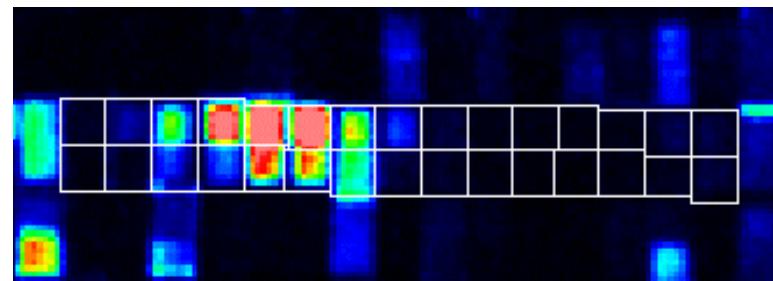
Presence for **Smax** and **Median** is $>3 \sigma$ above negative controls.
Presence for **AD** is by a proprietary Affymetrix formula.

Detection of Antisense and Untranslated RNAs

Expression Chip

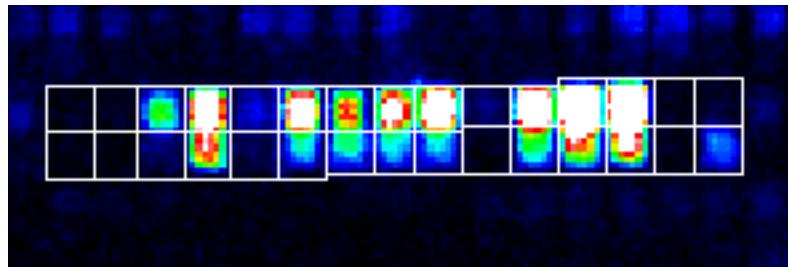


Reverse Complement Chip

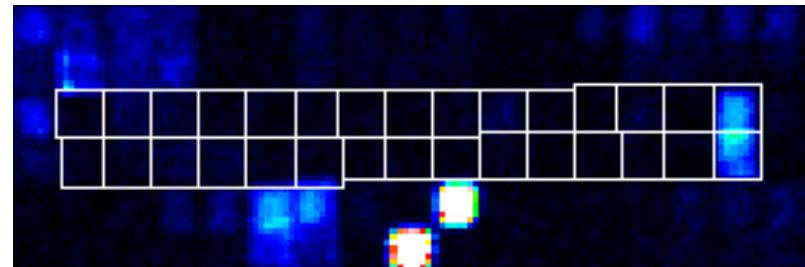


b0671 - predicted ORF of unknown function, tiled in the wrong orientation

Crick Strand

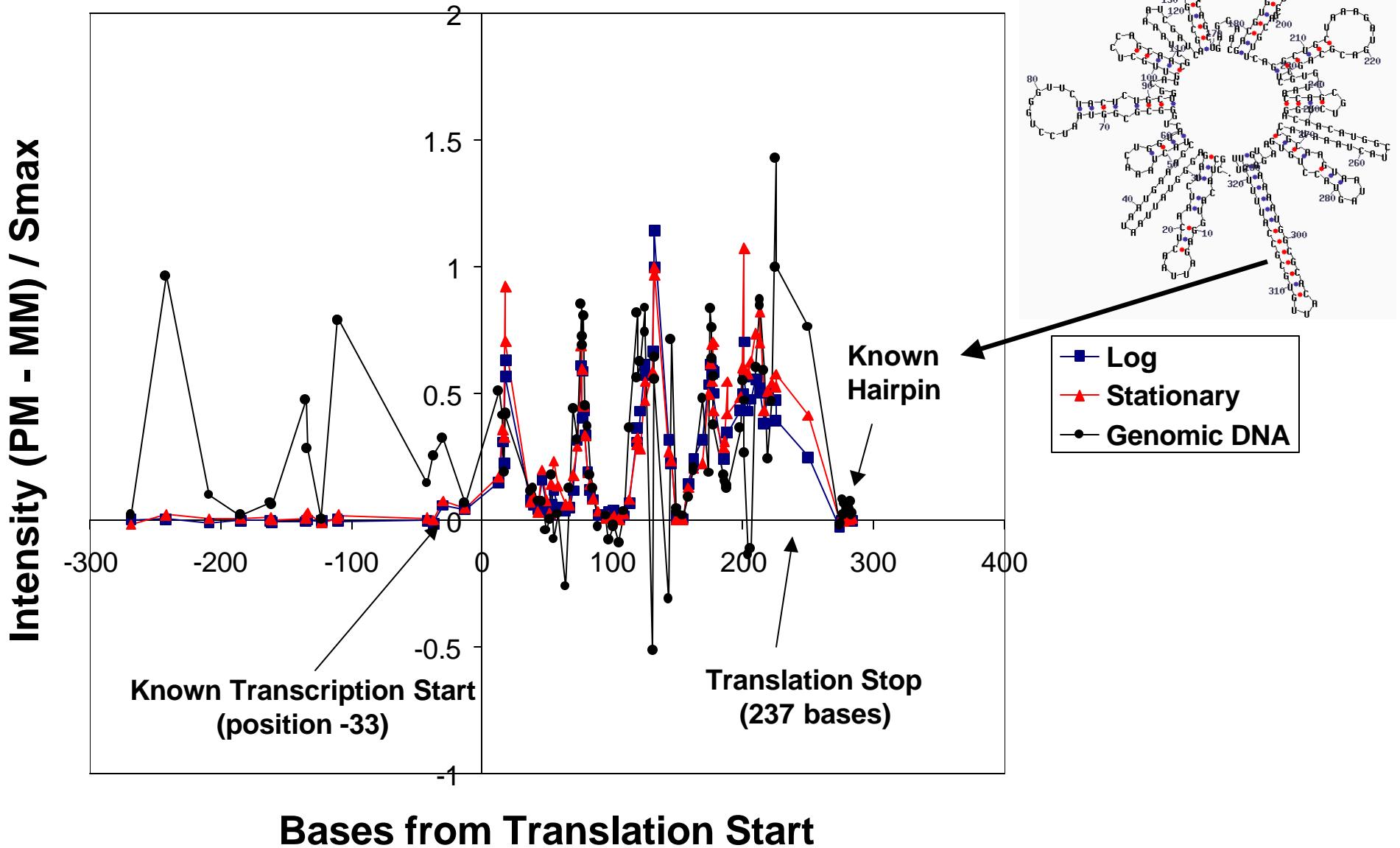


Watson Strand (same chip)



“intergenic region 1725” - actually *csrB*, a small untranslated RNA

Lpp mRNA start & structure



E.coli RNA/protein timecourses in progress

Cells grown at 37° C in rich LB:

- 1) Mid-**log** phase ($OD_{600} = 0.6$) in a fermentor
 - 2) Late **stationary** phase in an overnight shaken culture
-
- 3) Minimal Phos/NH3/SO4/Glucose:
Aerobic to anaerobic transition (and reverse)
 - 4) Minimal **glucose to acetate** (and reverse)

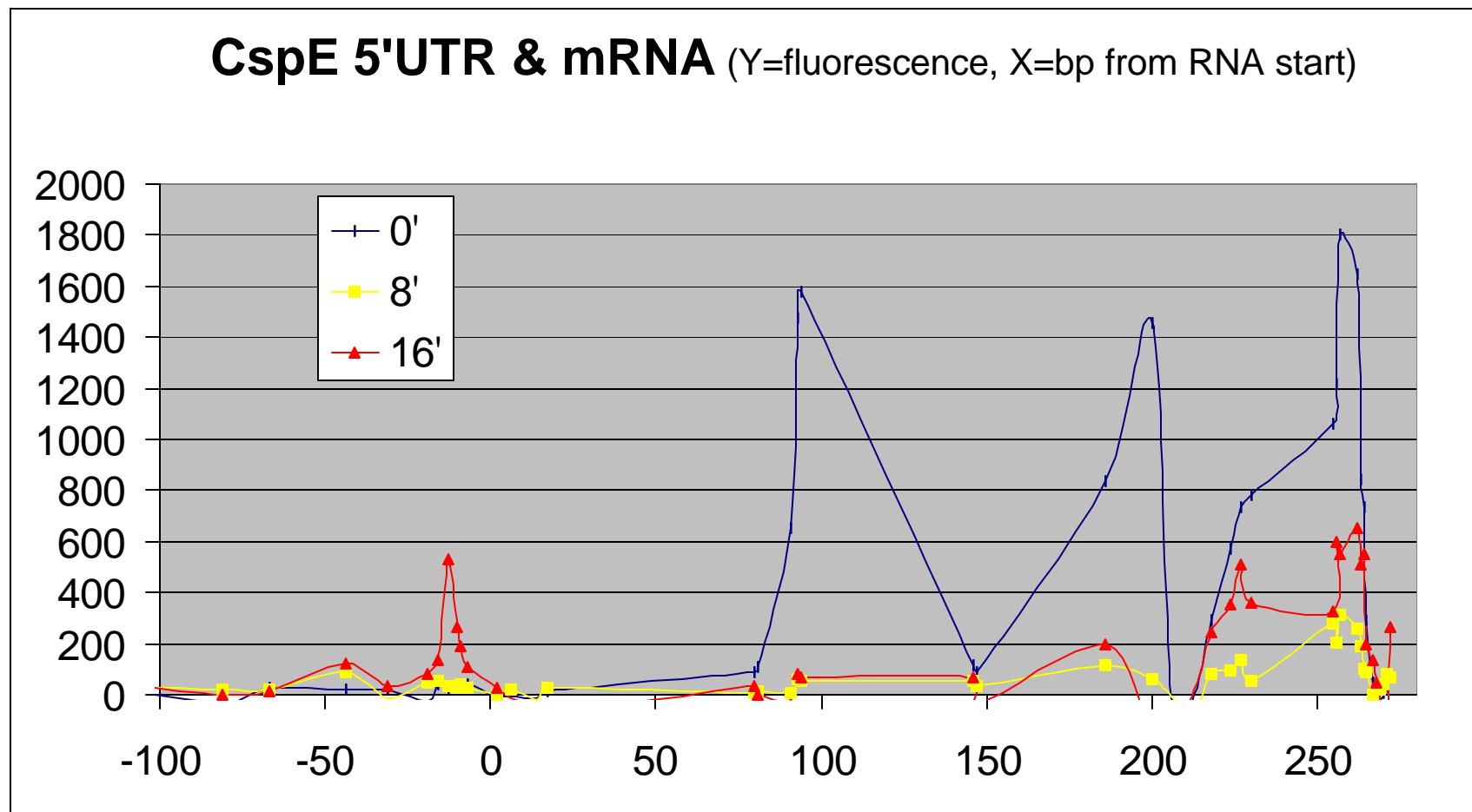
Cells taken **directly** into acid phenol:chloroform,
RNA purified, labeled in duplicate, & hybridized
to chips.

Most top 25* Stationary-induced RNAs are unknowns.

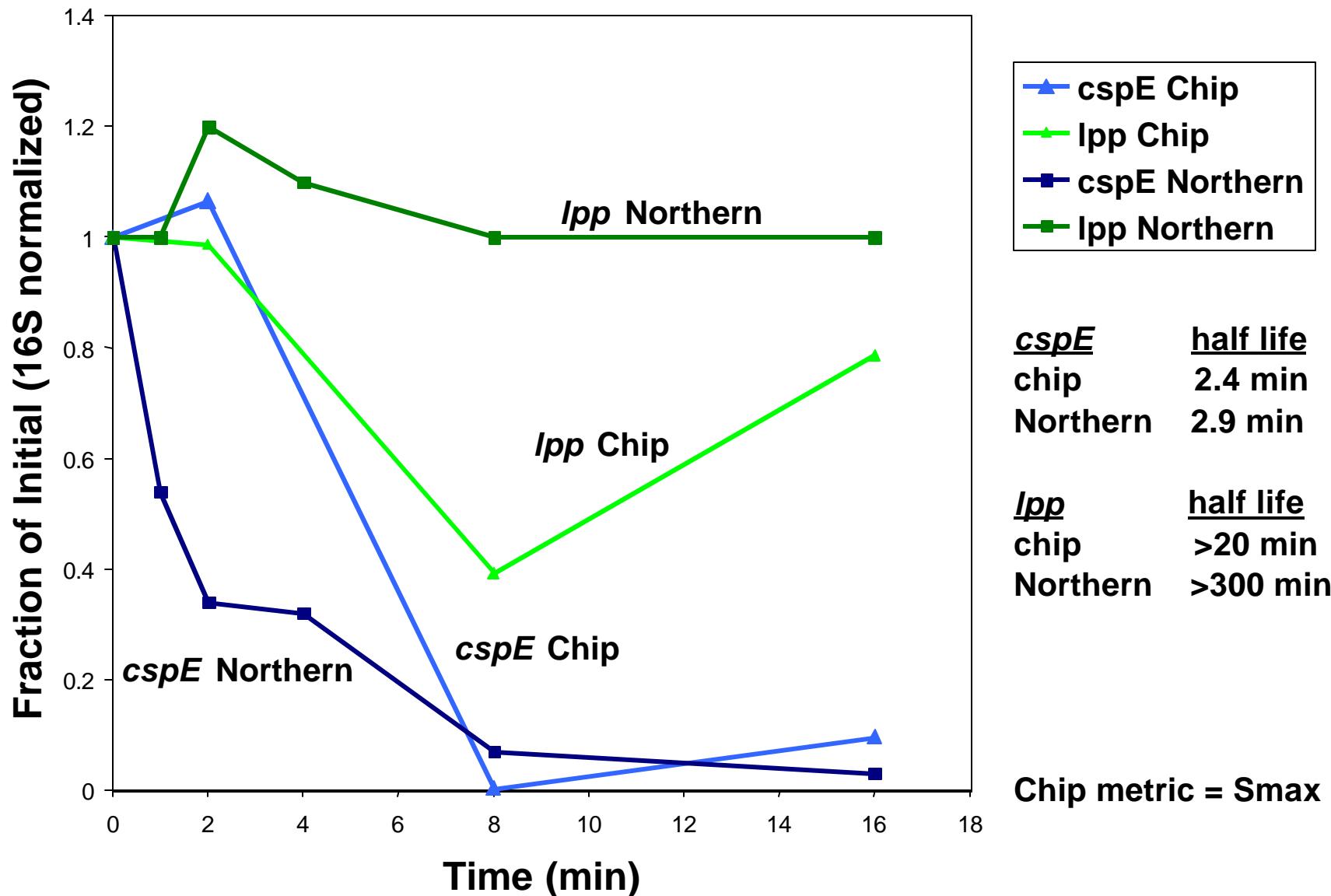
Bnumber	Gene	Fold Change	Annotation
b1005	<i>ycdF</i>	102.0	orf, hypothetical protein
b0836	-	>896.5	putative receptor
b0953	<i>rmf</i>	17.1	ribosome modulation factor
b3049	<i>glgS</i>	155.8	glycogen biosynthesis, <i>rpoS</i> dependent
b4045	<i>yjbJ</i>	9.0	orf, hypothetical protein
b3510	<i>hdeA</i>	41.4	orf, hypothetical protein
b0812	<i>dps</i>	55.2	global regulator, starvation conditions
b1480	<i>rpsV</i>	48.2	30S ribosomal subunit protein S22
b2665	<i>ygaU</i>	59.8	orf, hypothetical protein
b3555	<i>yiaG</i>	11.9	orf, hypothetical protein
b3239	<i>yhcO</i>	139.5	orf, hypothetical protein
b1240	-	3.8	orf, hypothetical protein
b1635	<i>gst</i>	80.7	glutathione S-transferase
b1051	<i>msyB</i>	10.8	acidic protein suppresses mutants lacking function of protein export
b0966	<i>yccV</i>	15.7	orf, hypothetical protein
b1318	<i>ycjV</i>	75.2	putative ATP-binding component of a transport system
b1154	<i>ycfK</i>	>123.6	orf, hypothetical protein
b1566	<i>flxA</i>	13.1	orf, hypothetical protein
b2212	<i>alkB</i>	5.6	DNA repair system specific for alkylated DNA
b1492	<i>xasA</i>	84.6	acid sensitivity protein, putative transporter
b2266	<i>elaB</i>	>92.1	orf, hypothetical protein
b1164	<i>ycgZ</i>	3.0	orf, hypothetical protein
b3183	<i>yhbZ</i>	6.7	putative GTP-binding factor
b1262	<i>trpC</i>	6.8	N-(5-phosphoribosyl)anthranilate isomerase and indole-3-glycerolphosphate synthetase
b1739	<i>osmE</i>	23.6	activator of <i>ntrL</i> gene

*ranked by absolute Smax change

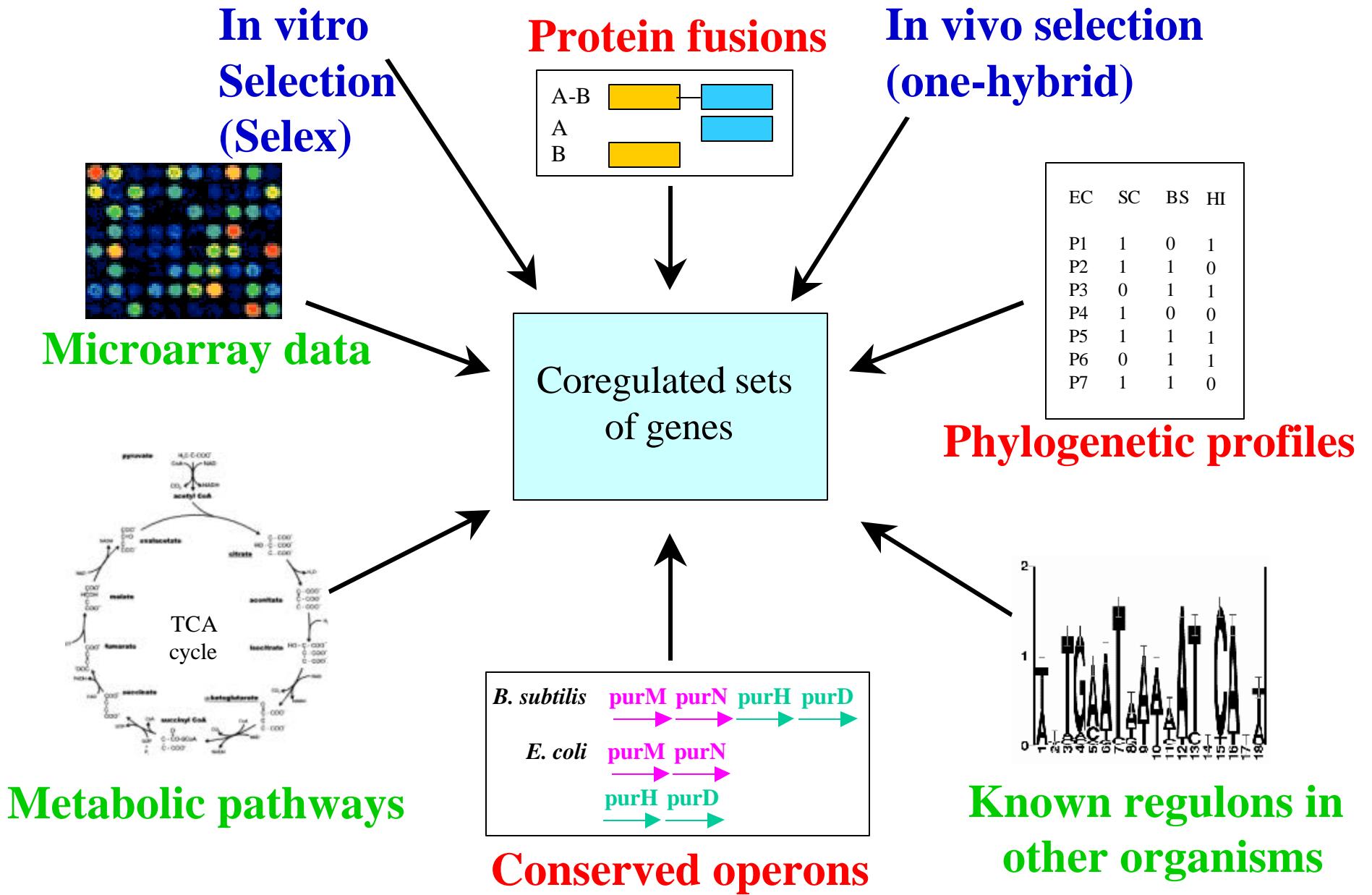
Rifampicin timecourse



mRNA turnover rates (rifampicin timecourse, array & gel assays)



8 methods for regulon discovery in 18 genomes



Why are RNAs co-clustered? (and how do we assess quality?)

Upstream: DNA motifs

(AlignACE, mutant RNA profiles)

Further up: DNA-binding proteins

(One-hybrid, *in vivo* crosslinking)

Downstream: enzymes and structural proteins

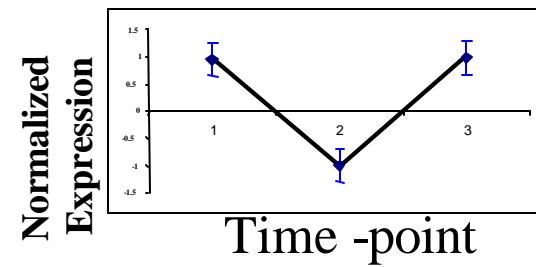
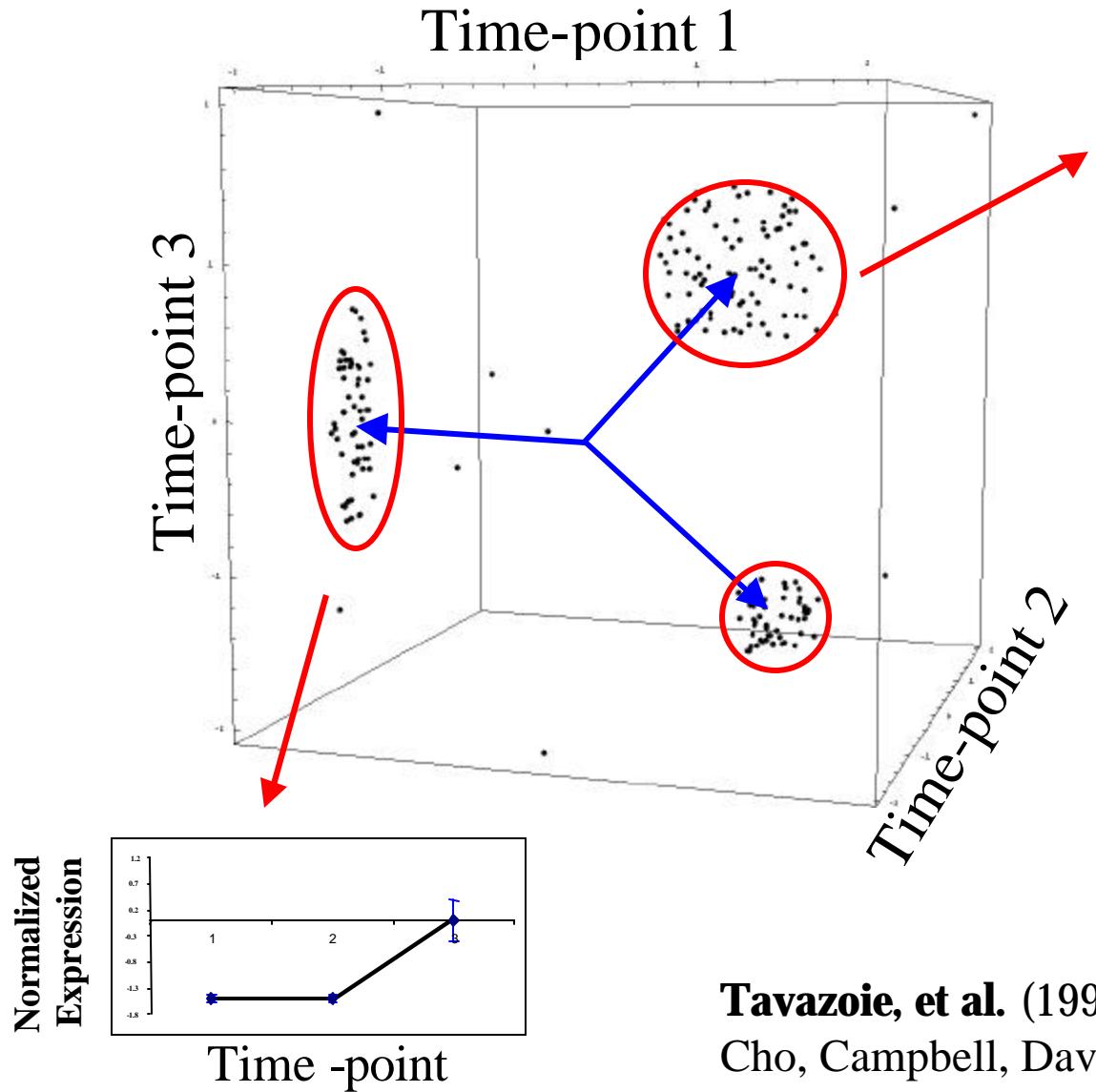
(ScanACE, co-classification odds)

Further down: Survival of the species

(multiplex mutant growth competition)

Yeast Expression Clustering & Motifs

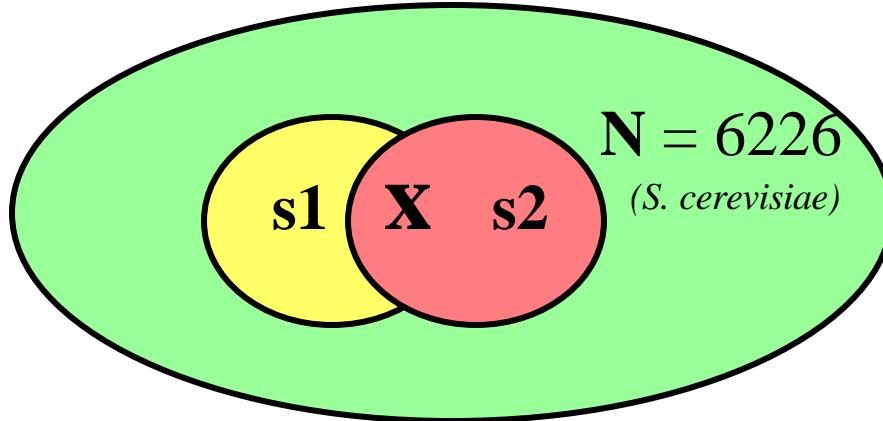
Saeed Tavazoie, Jason Hughes, Fritz Roth



17 time-points
(two cell cycles)
30 clusters

Tavazoie, et al. (1999) Nature Genetics **22**: 281-5
Cho, Campbell, Davis, Lockhart cell cycle data

Group Specificity Score (S_{group})



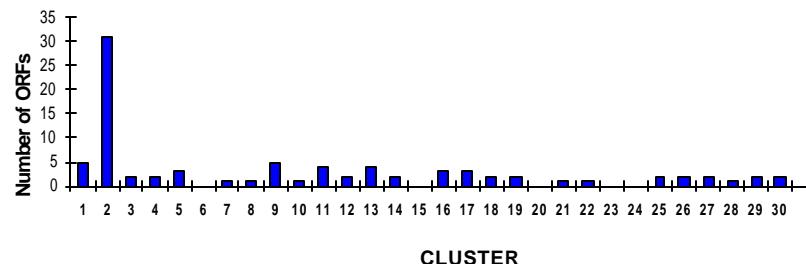
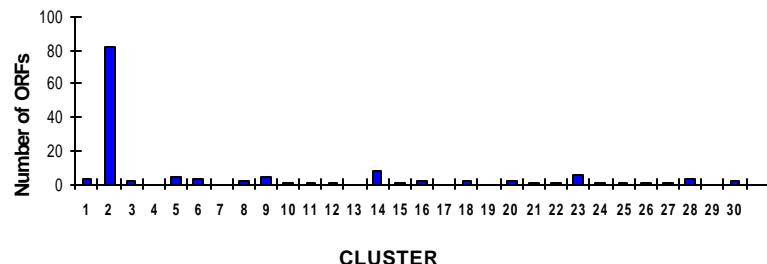
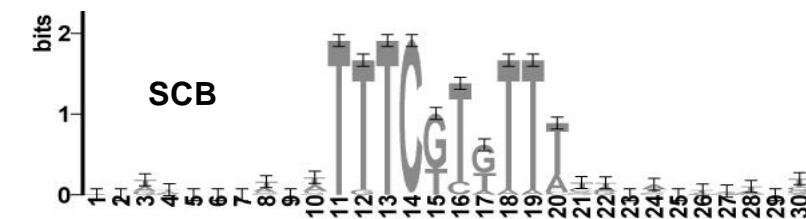
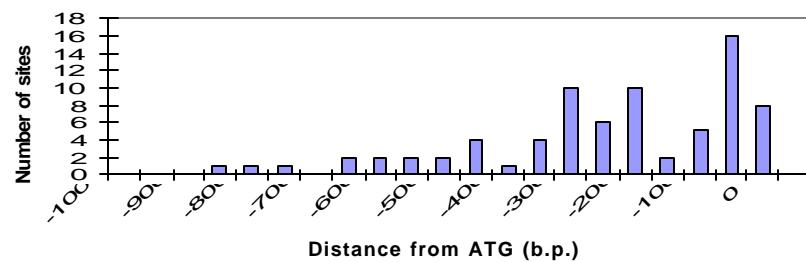
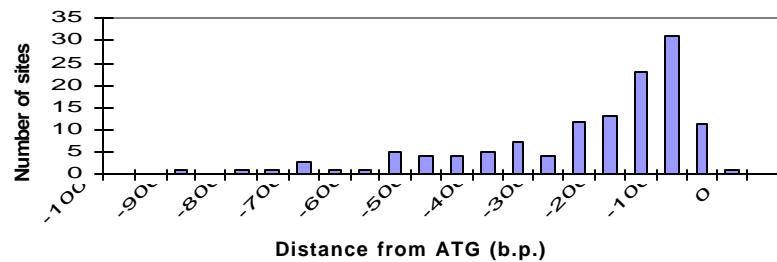
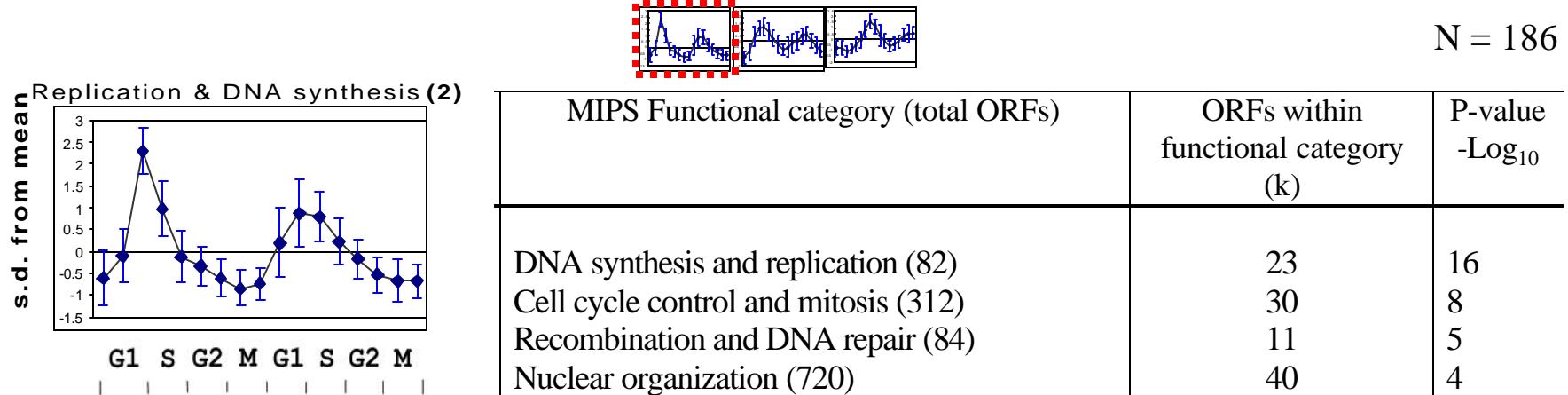
N = Total # of ORFs in the genome

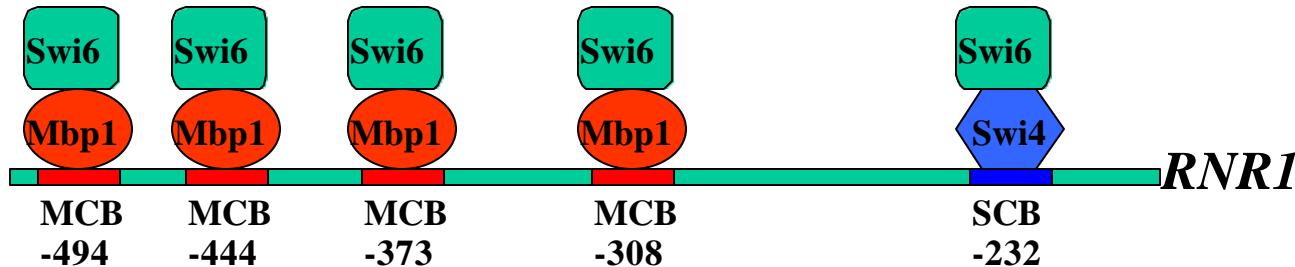
s_1 = # ORFs whose upstream sequences were used to align the motif

s_2 = # ORFs in the target list (~ 100 genes in the genome with the best sites for the motif near their translational starts)

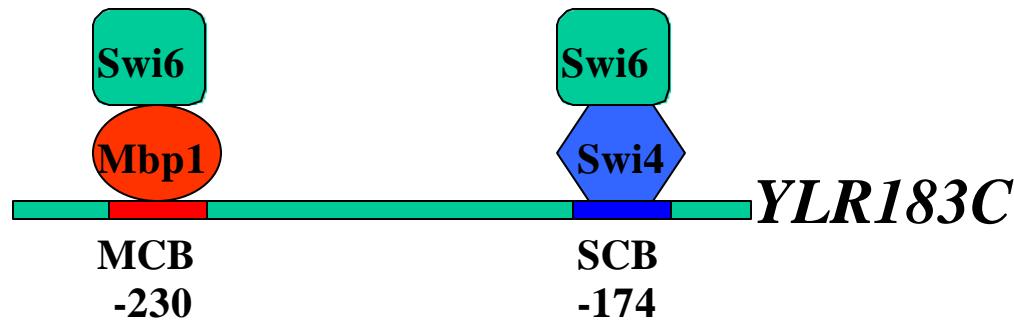
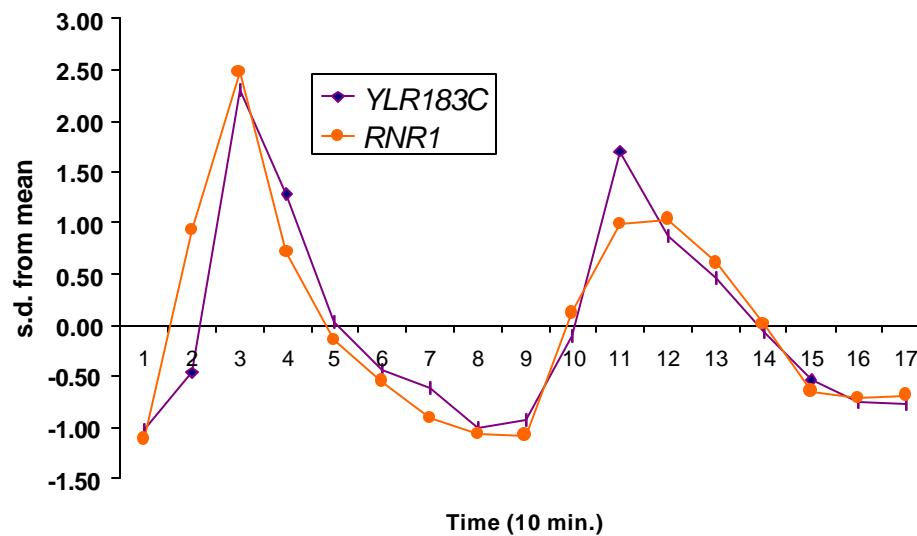
x = # ORFs in the intersection of these groups

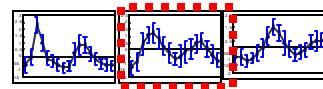
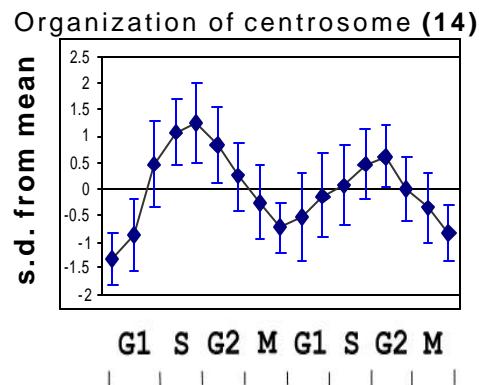
$$S_{group} = \sum_{i=x}^{\min(s_1, s_2)} \frac{\binom{s_1}{i} \binom{N-s_1}{s_2-i}}{\binom{N}{s_2}}$$





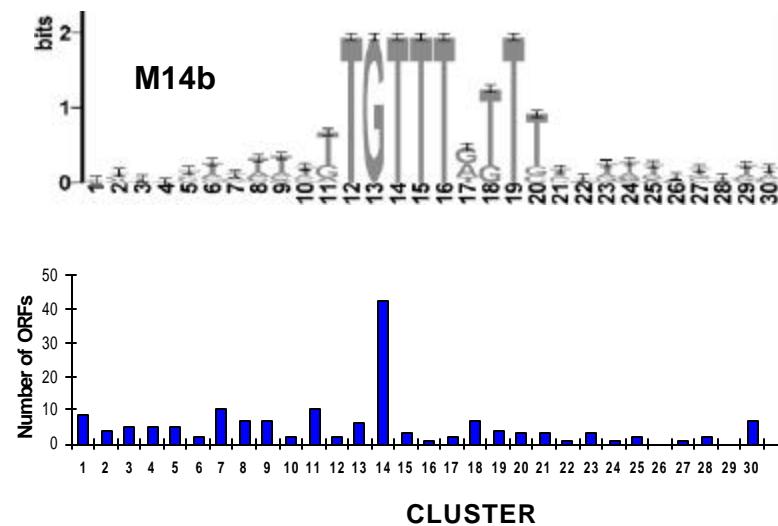
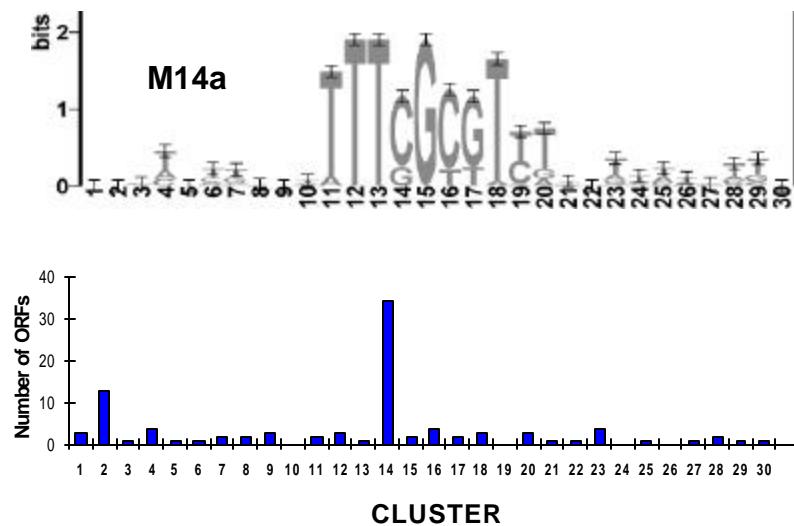
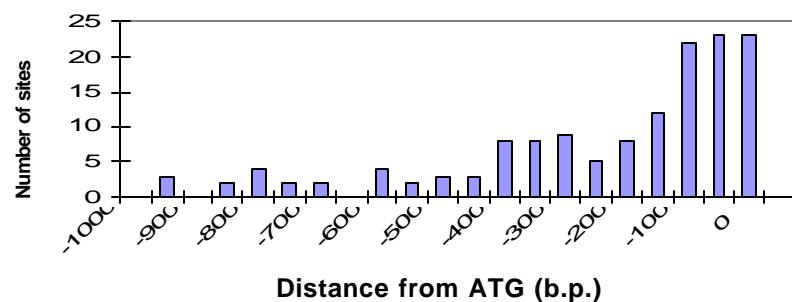
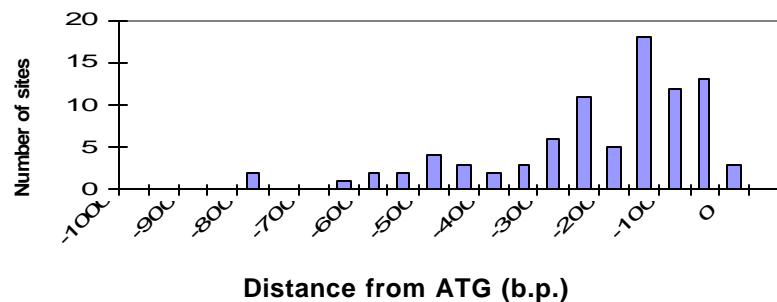
(G1-S)
Replication &
DNA synthesis



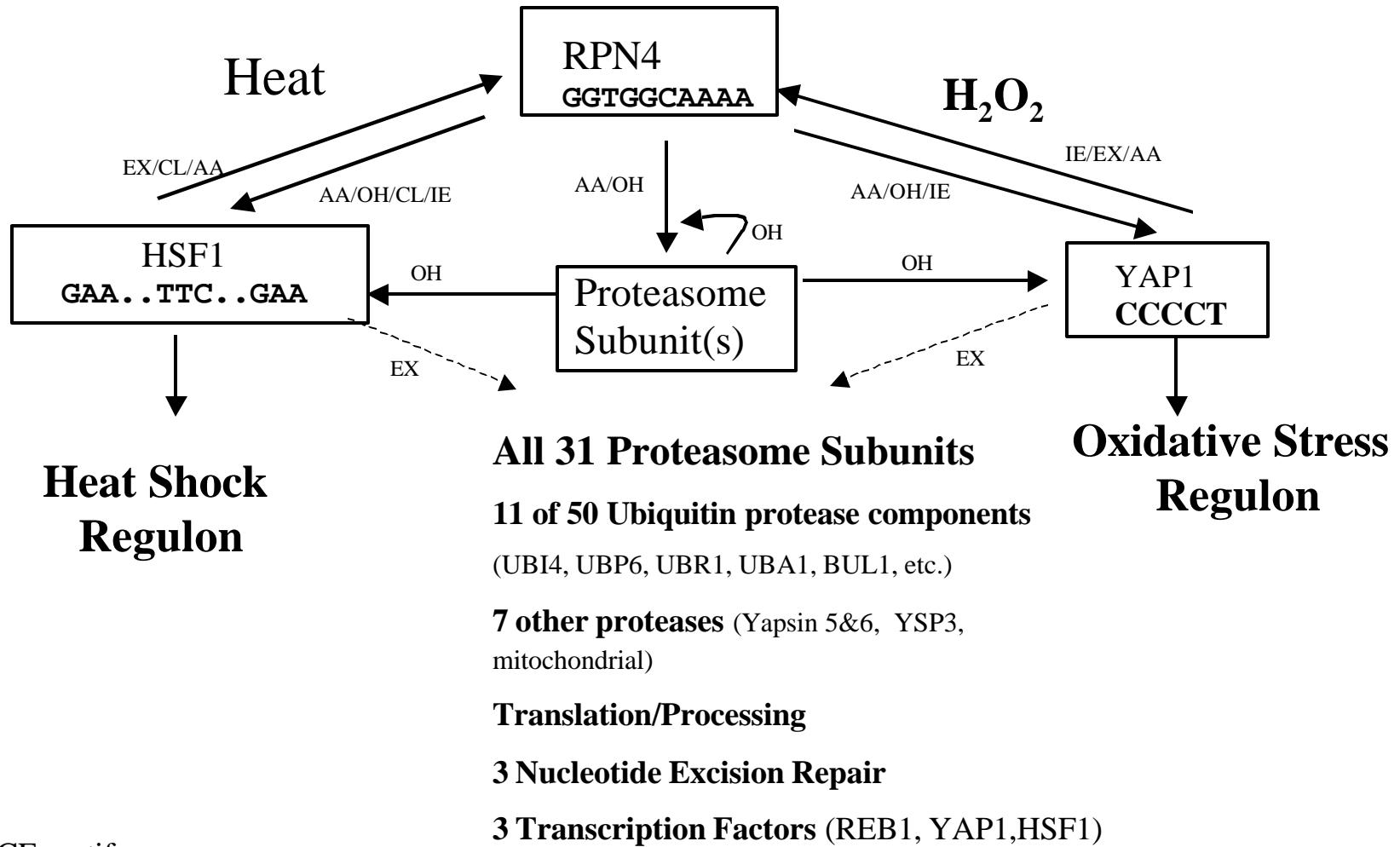


N = 74

MIPS Functional category (total ORFs)	ORFs within functional category (k)	P-value -Log ₁₀
Organization of centrosome (28)	6	6
Nuclear biogenesis (5)	3	5
Organization of cytoskeleton (93)	7	4*



Partners & validation of new motifs



AA: AlignACE motif

CL: Clustered with genes regulated by this factor

OH: One-hybrid

IE: Increase on over-expression of upstream factor

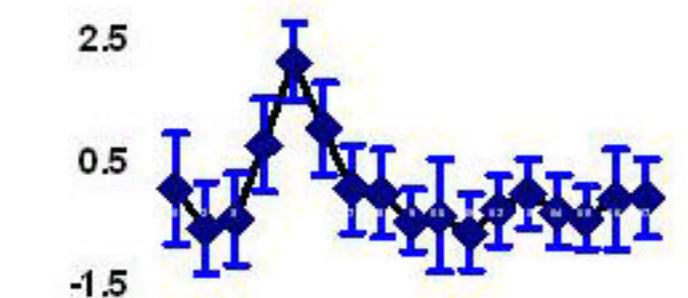
EX: Expressed conditionally (heat & peroxide)

Pete Estep

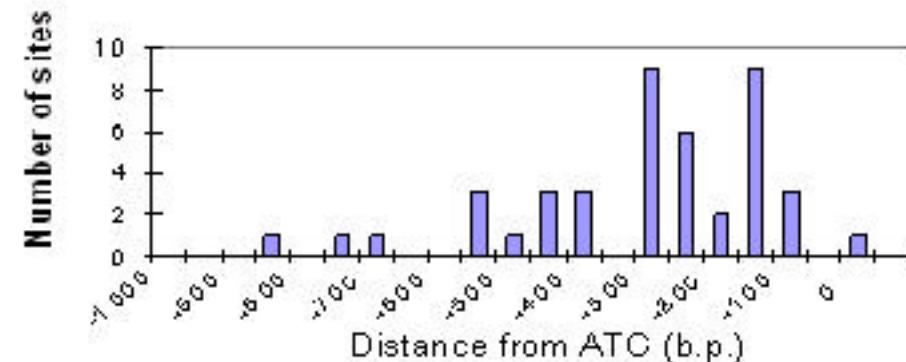
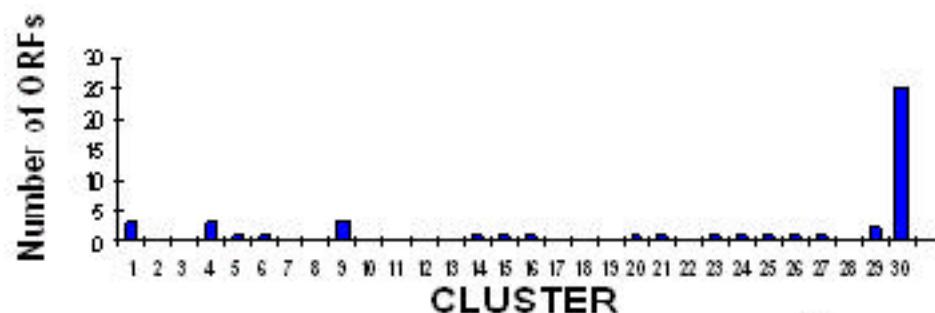
Cell-cycle non-periodic cluster #30

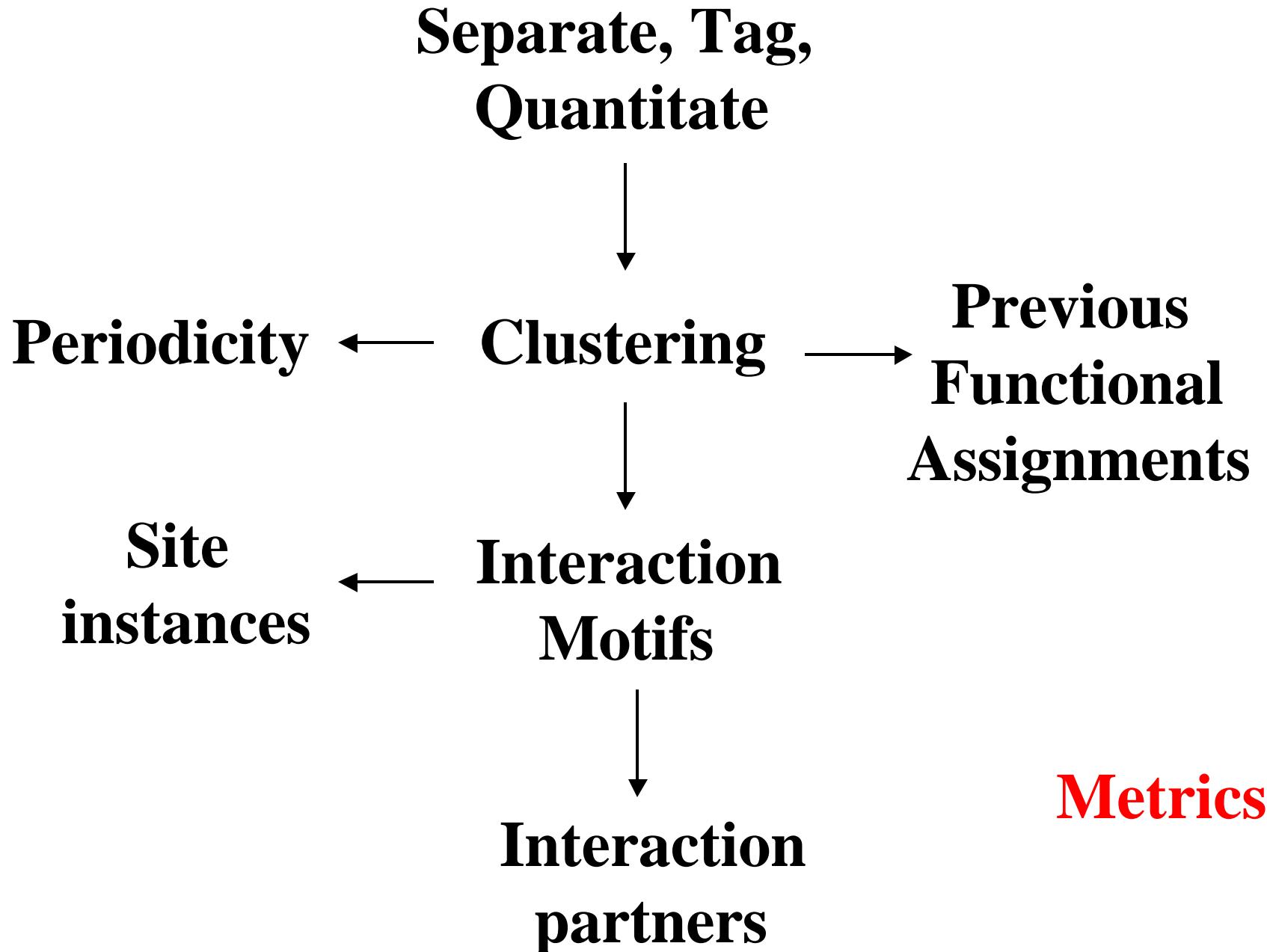
N,S metabolism $p < 10^{-8}$

AA metabolism $p < 10^{-7}$



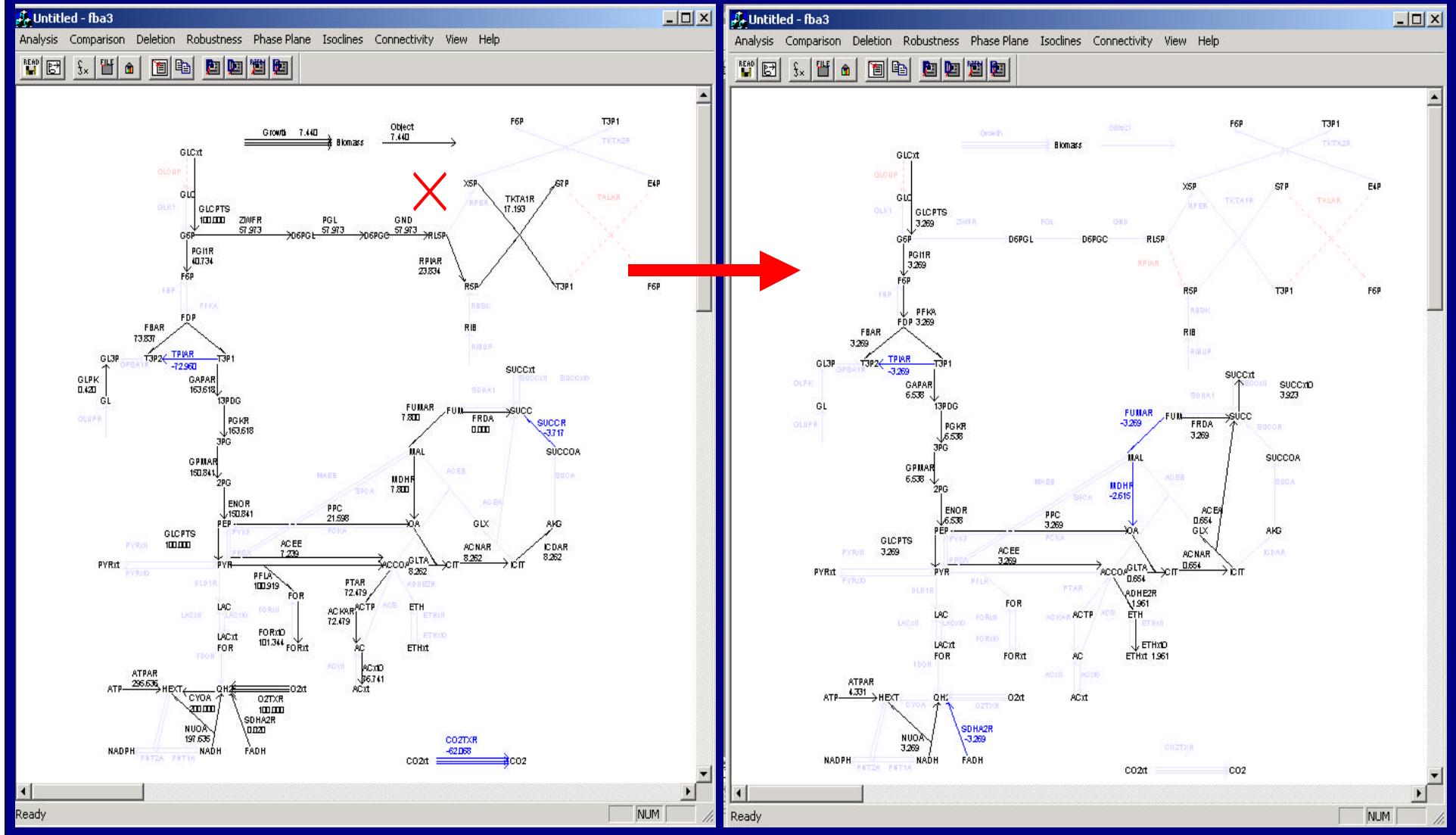
**Met31/32p
(30)**



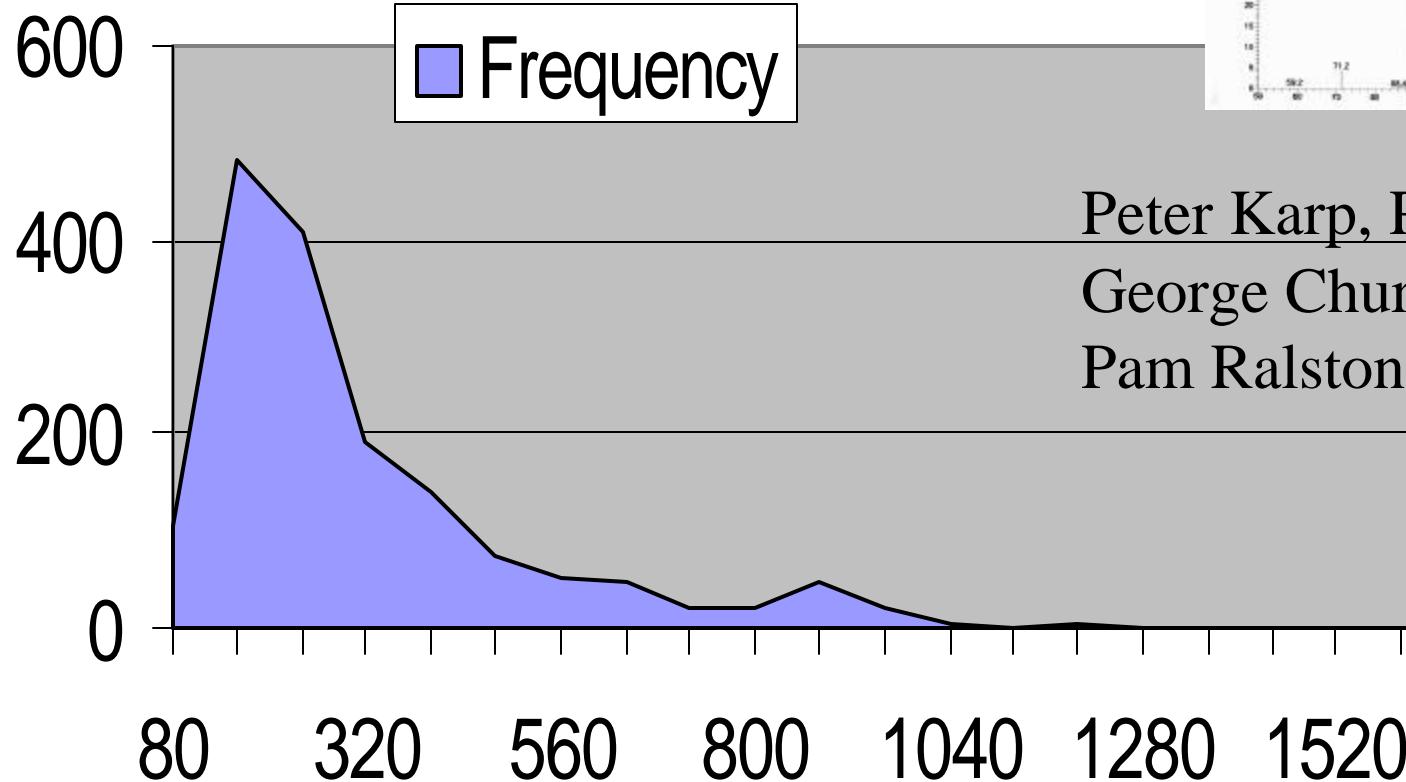


Goal #2: Enzyme gene deletions

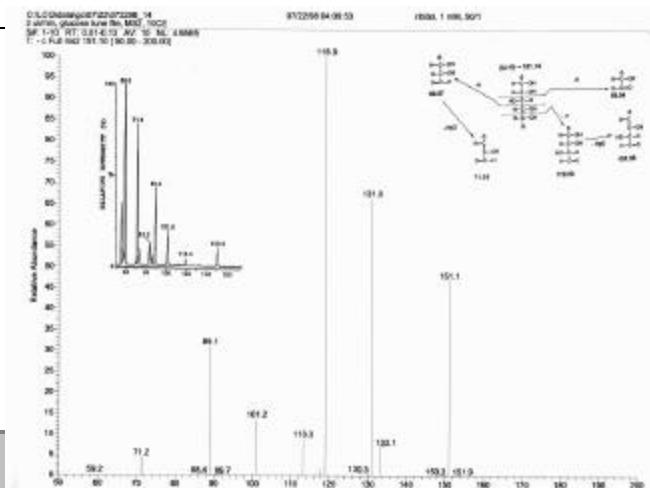
Jeremy Edwards & Dereth Phillips in collaboration
with Bernard Palsson group at UCSD



1634 Metabolite Masses 256 amino acids



Peter Karp, Pangea
George Church
Pam Ralston



Karp et al. (1998) *NAR* 26:50. EcoCyc; Selkov, et al. (1997) *NAR* 25:37. WIT
Ogata et al. (1998) *Biosystems* 47:119-128 KEGG

Genome Engineering

Challenges: Construct any mutant in any background, multiple mutants, minimizing hitchhiking mutants.

Avoid undesired residual activities and neomorphic effects on adjacent genes in most deletion, insertion nonsense, or antisense alleles.

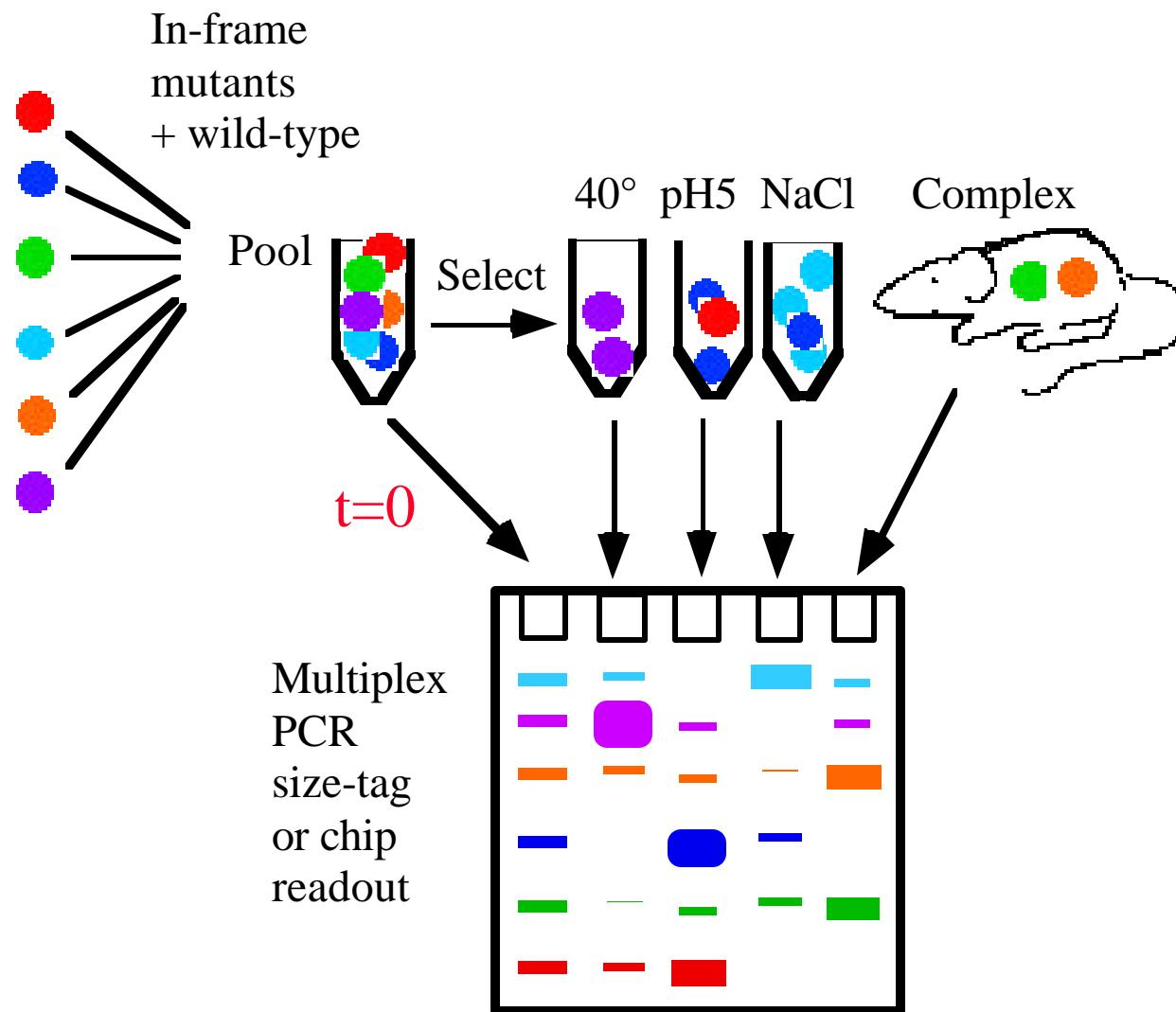
Full in-frame replacements, computationally track gene overlaps, primer & genomic repeats.

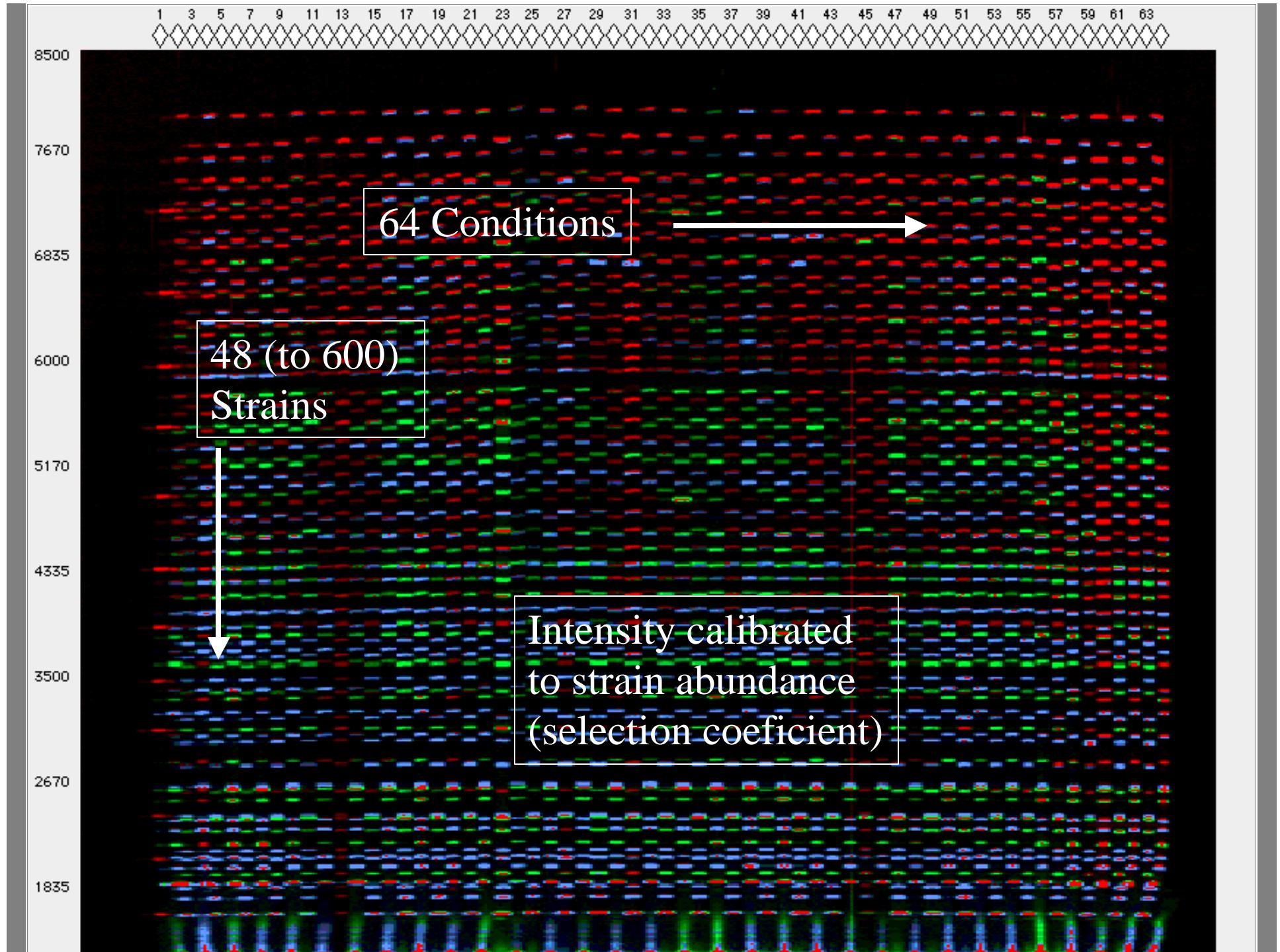


Link, Phillips, Church (1997) J. Bacteriol. 179: 6228-6237.

(pKO3) <http://arep.med.harvard.edu>

E.coli & Yeast Mutant Competitive Growth Experiments





Summary

- #1 Reconstructing RNA & protein networks
 - High-resolution arrays (one 25mer per >6 bp):
 - RNA steady-state levels & decay rates
 - RNA 5' and 3' ends & structure
 - Small untranslated RNAs & antisense
 - Beyond clustering to mechanisms
- #2 Flux-Balance & Gene Deletions
 - Electrophoretic and array based assays
 - in-frame & Tn based mutations

