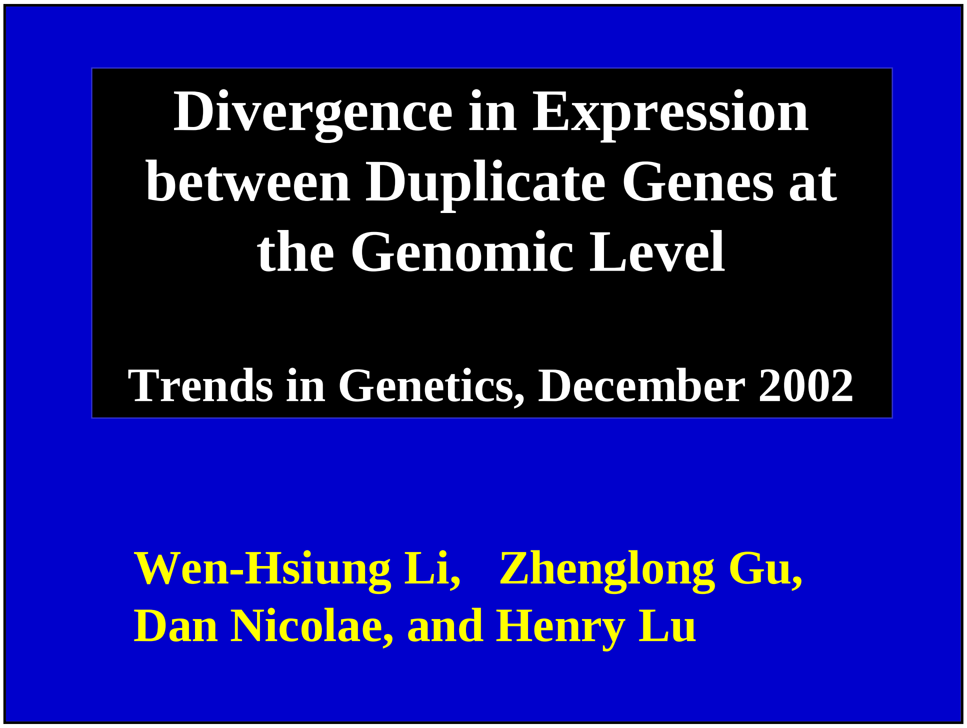


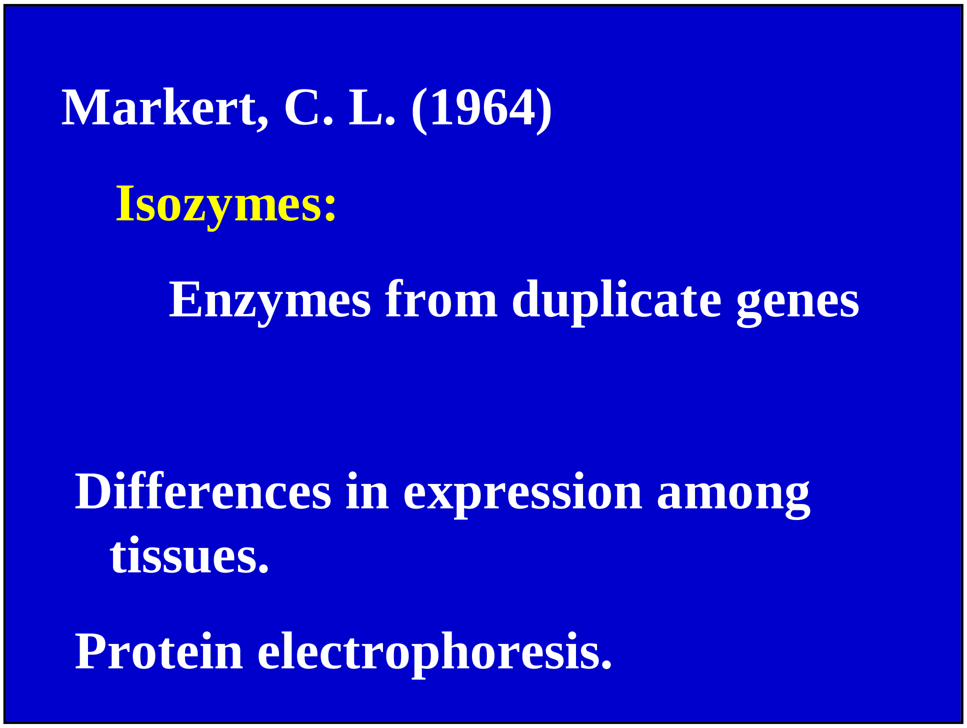
Divergence in Expression between Duplicate Genes at the Genomic Level

Trends in Genetics, December 2002

**Wen-Hsiung Li, Zhenglong Gu,
Dan Nicolae, and Henry Lu**

Expression Divergence between Duplicate Genes:

An old issue



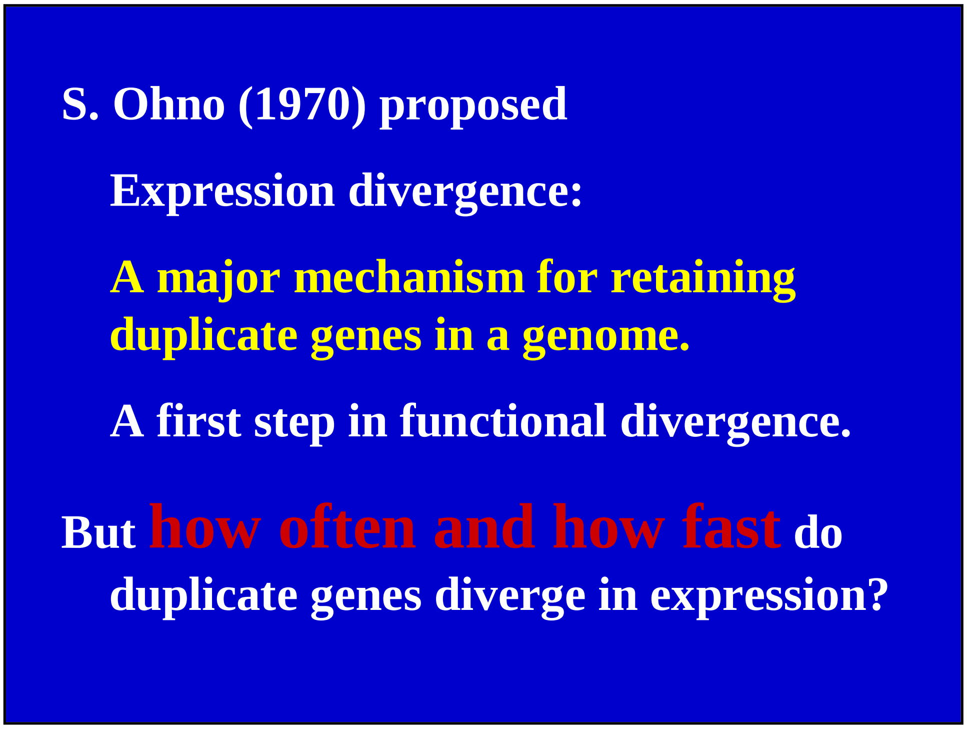
Markert, C. L. (1964)

Isozymes:

Enzymes from duplicate genes

Differences in expression among tissues.

Protein electrophoresis.



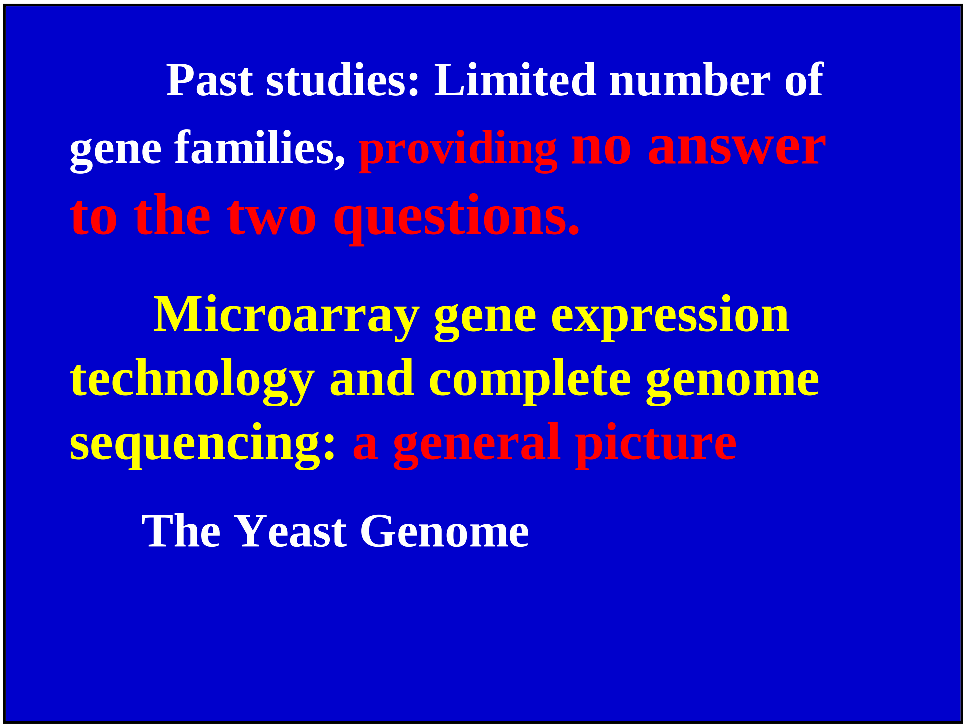
S. Ohno (1970) proposed

Expression divergence:

A major mechanism for retaining duplicate genes in a genome.

A first step in functional divergence.

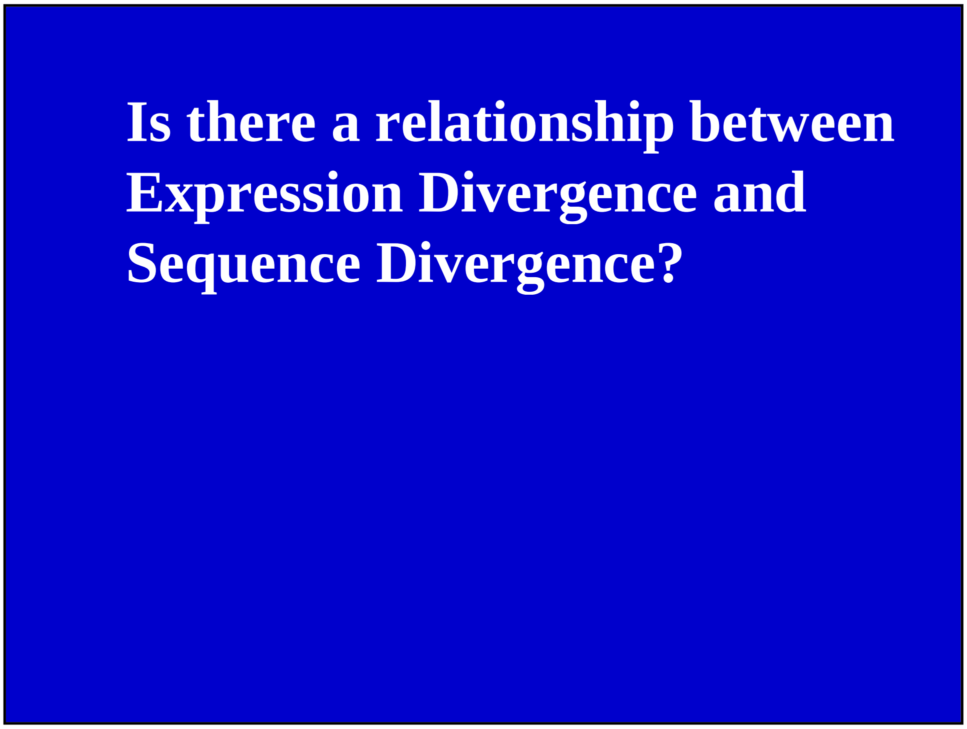
But how often and how fast do duplicate genes diverge in expression?



Past studies: Limited number of gene families, providing no answer to the two questions.

Microarray gene expression technology and complete genome sequencing: a general picture

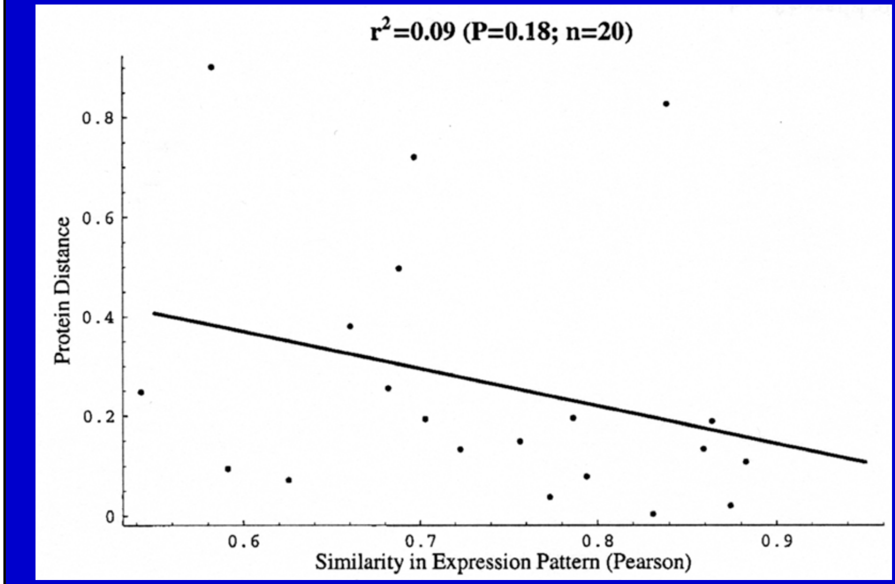
The Yeast Genome



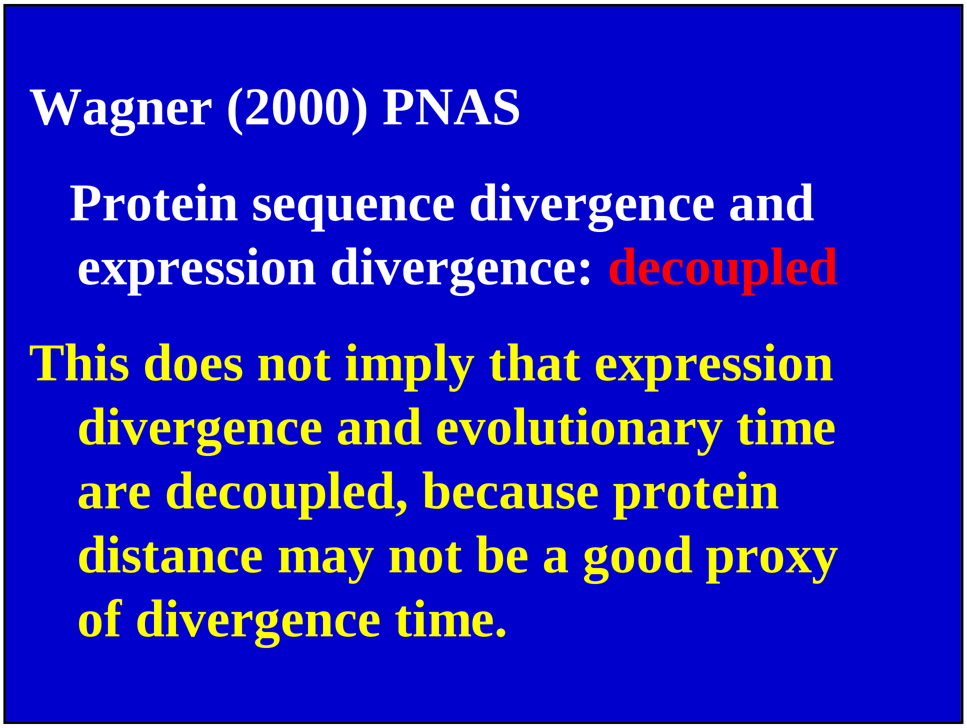
Is there a relationship between Expression Divergence and Sequence Divergence?

Similarity between expression patterns of two genes

R **= the correlation coefficient of the expression levels of the two genes over different time points of an experiment (a process)**



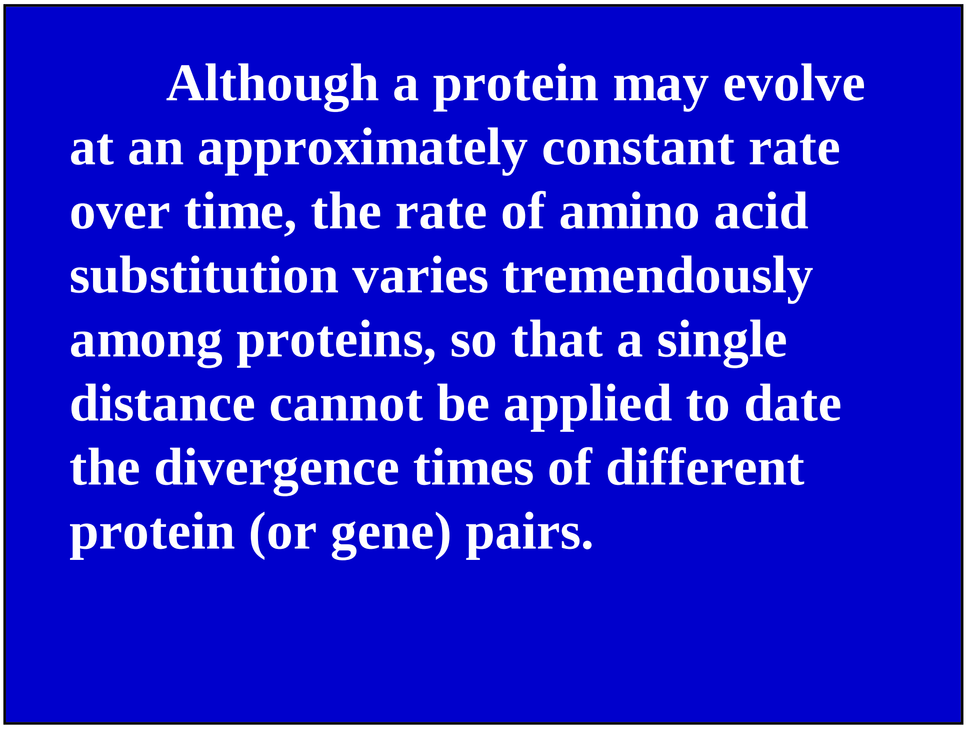
Wagner, 2000



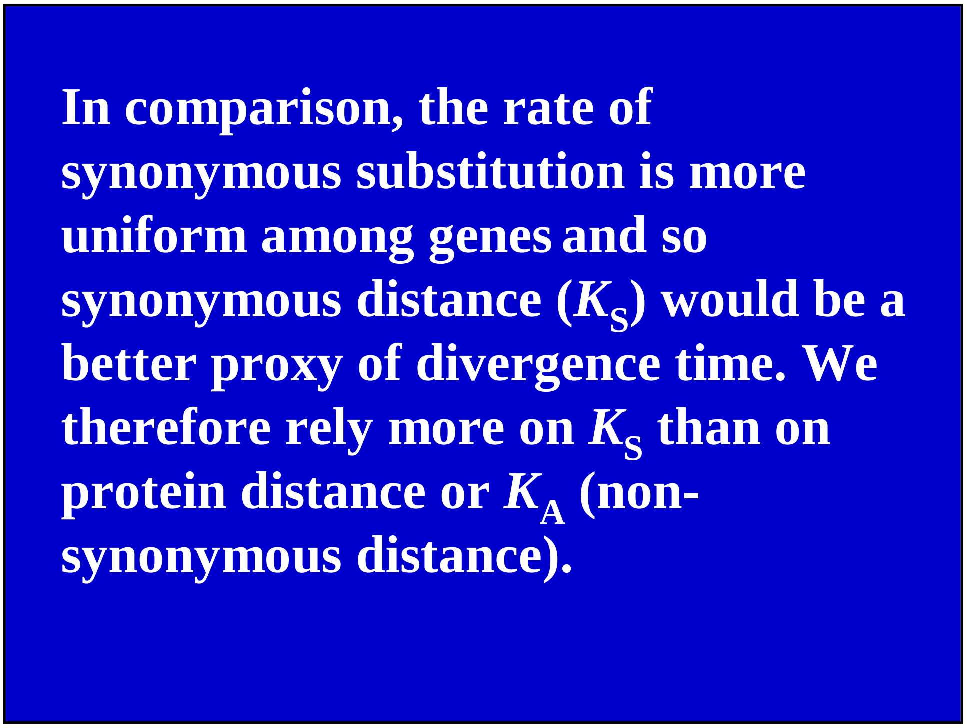
Wagner (2000) PNAS

Protein sequence divergence and expression divergence: decoupled

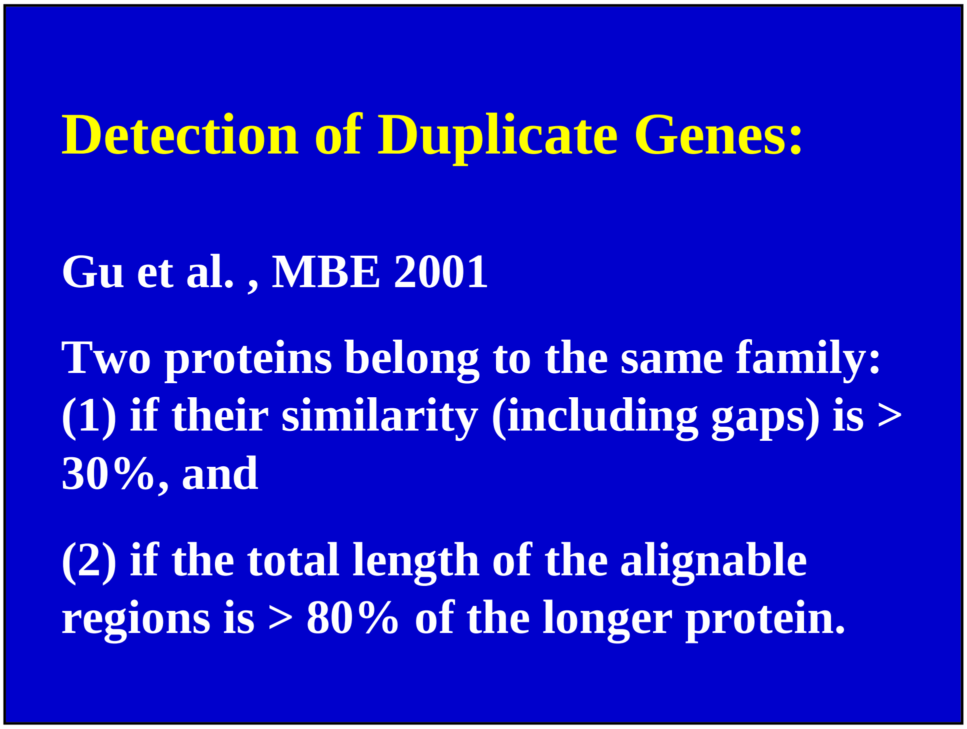
This does not imply that expression divergence and evolutionary time are decoupled, because protein distance may not be a good proxy of divergence time.



Although a protein may evolve at an approximately constant rate over time, the rate of amino acid substitution varies tremendously among proteins, so that a single distance cannot be applied to date the divergence times of different protein (or gene) pairs.



In comparison, the rate of synonymous substitution is more uniform among genes and so synonymous distance (K_S) would be a better proxy of divergence time. We therefore rely more on K_S than on protein distance or K_A (non-synonymous distance).

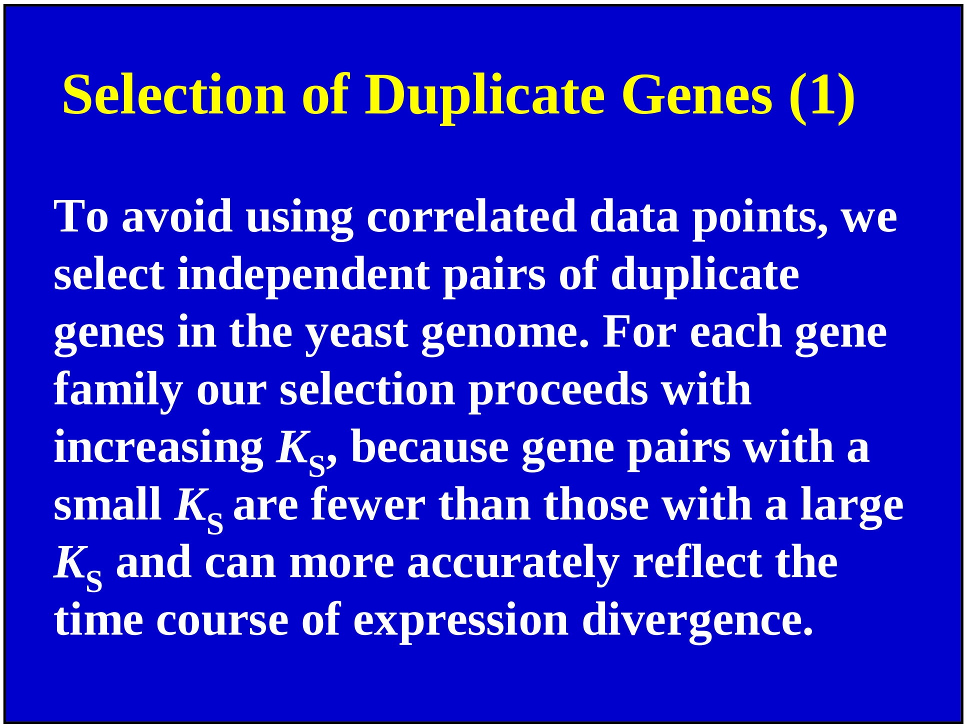


Detection of Duplicate Genes:

Gu et al. , MBE 2001

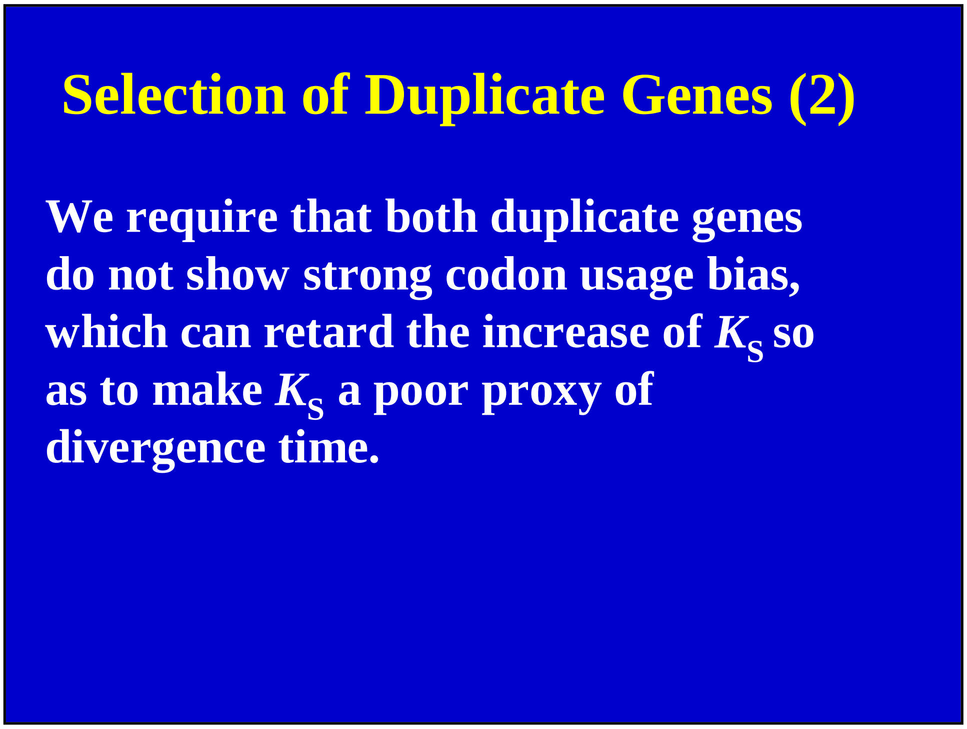
Two proteins belong to the same family:
(1) if their similarity (including gaps) is > 30%, and

(2) if the total length of the alignable regions is > 80% of the longer protein.



Selection of Duplicate Genes (1)

To avoid using correlated data points, we select independent pairs of duplicate genes in the yeast genome. For each gene family our selection proceeds with increasing K_S, because gene pairs with a small K_S are fewer than those with a large K_S and can more accurately reflect the time course of expression divergence.



Selection of Duplicate Genes (2)

We require that both duplicate genes do not show strong codon usage bias, which can retard the increase of K_S so as to make K_S a poor proxy of divergence time.

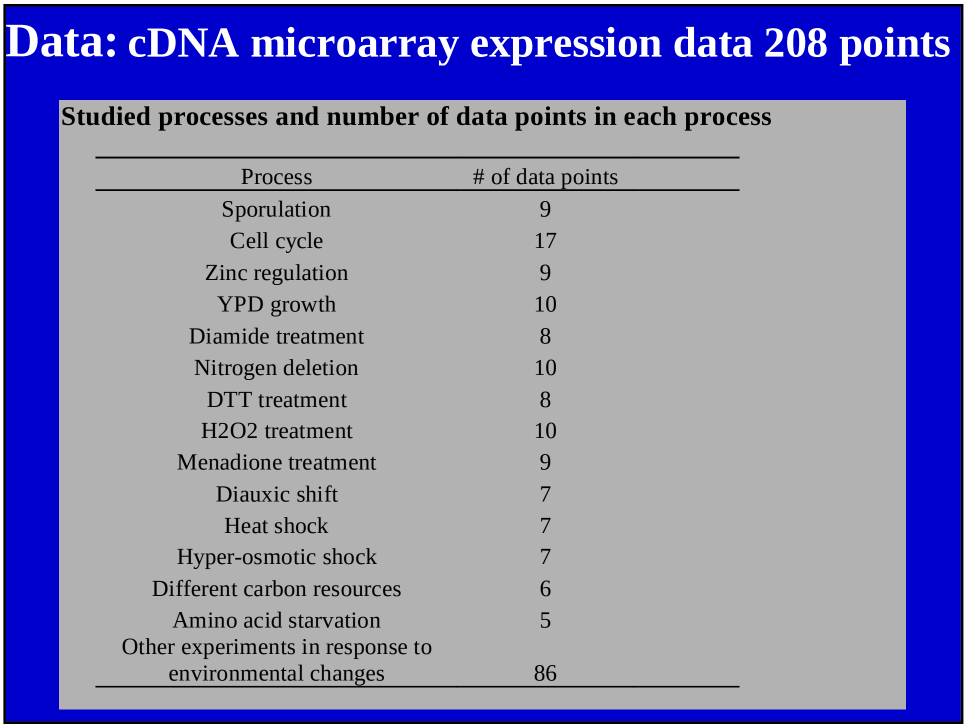
Data

A total of 400 pairs were selected.

Linear regression analysis

Since R is bounded by –1 and 1, the transformation $\ell n((1+R)/(1-R))$ was used.

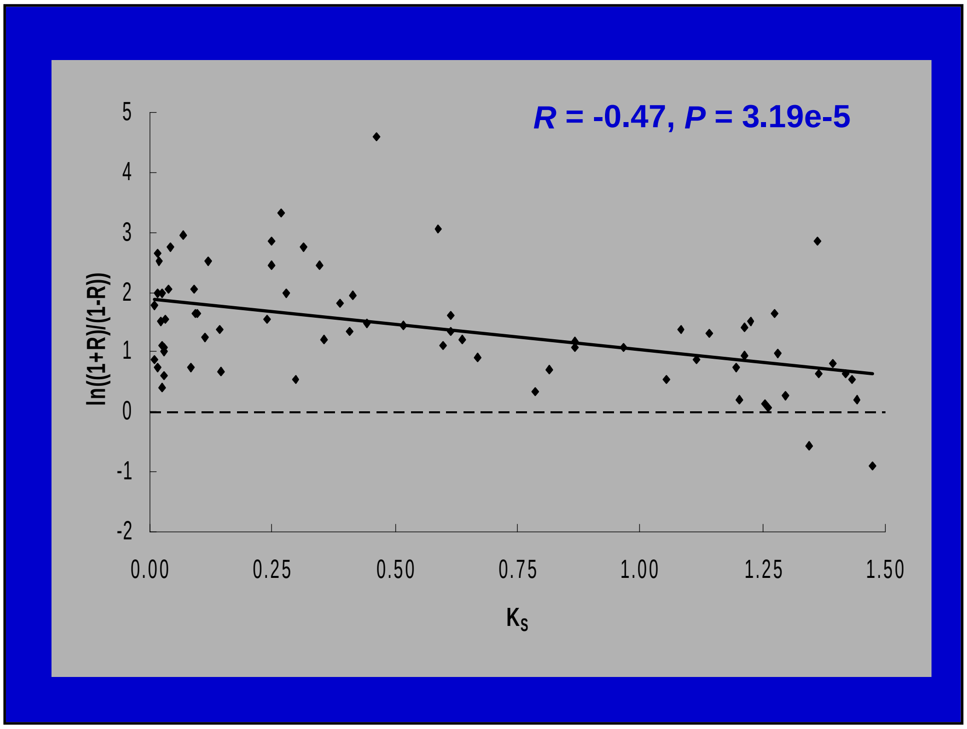
The normal linear regression was then carried out between K_S (K_A) and the transformed R .

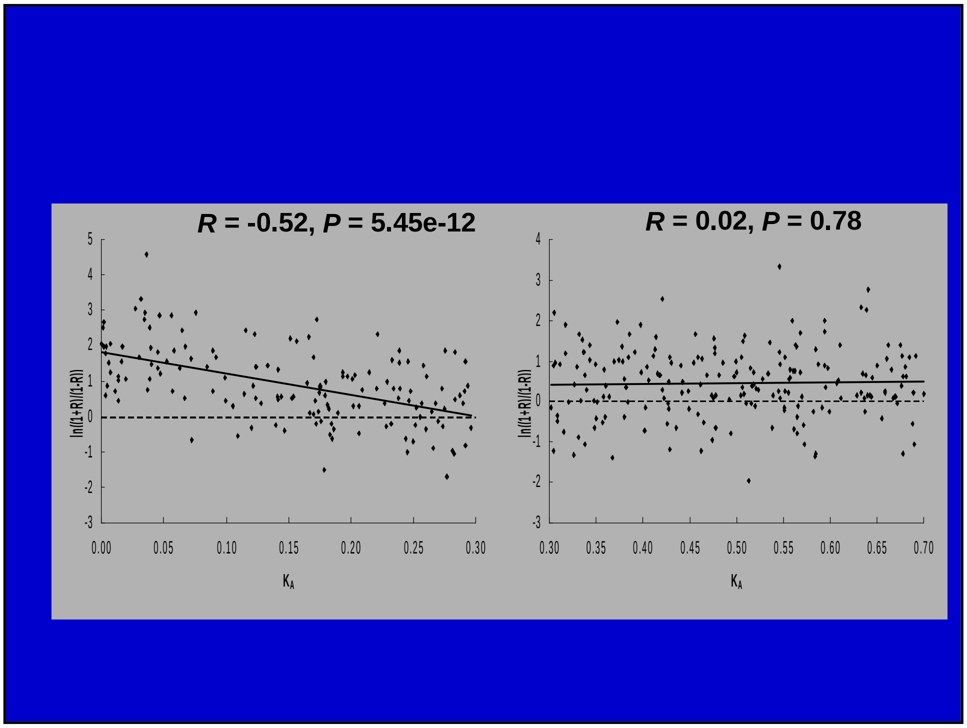


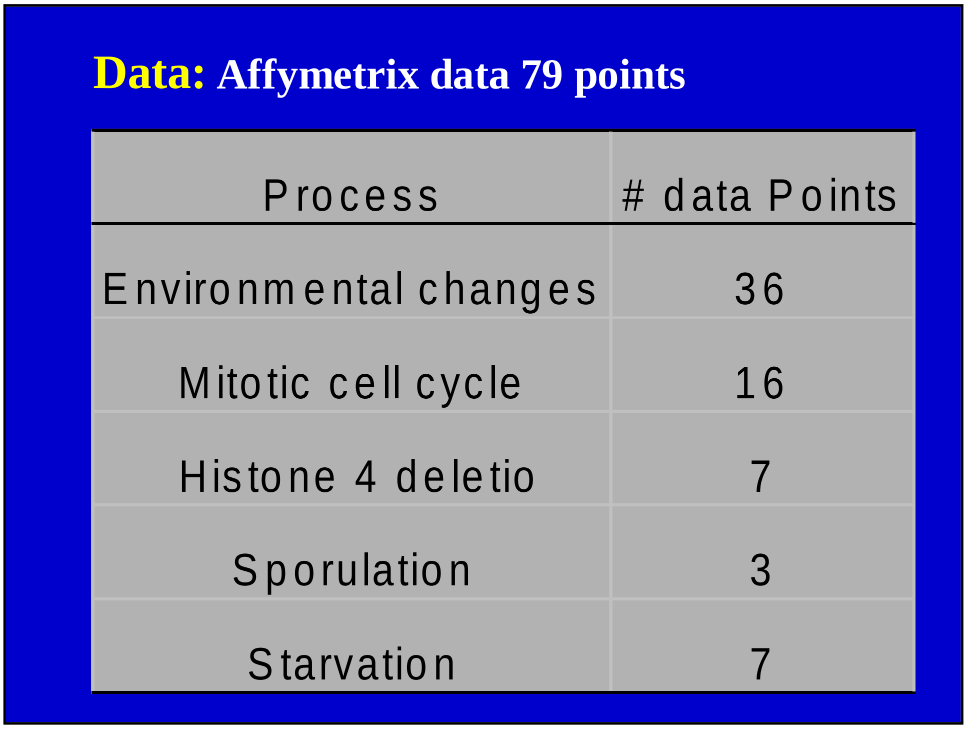
Data: cDNA microarray expression data 208 points

Studied processes and number of data points in each process

Process	# of data points
Sporulation	9
Cell cycle	17
Zinc regulation	9
YPD growth	10
Diamide treatment	8
Nitrogen deletion	10
DTT treatment	8
H2O2 treatment	10
Menadione treatment	9
Diauxic shift	7
Heat shock	7
Hyper-osmotic shock	7
Different carbon resources	6
Amino acid starvation	5
Other experiments in response to environmental changes	86

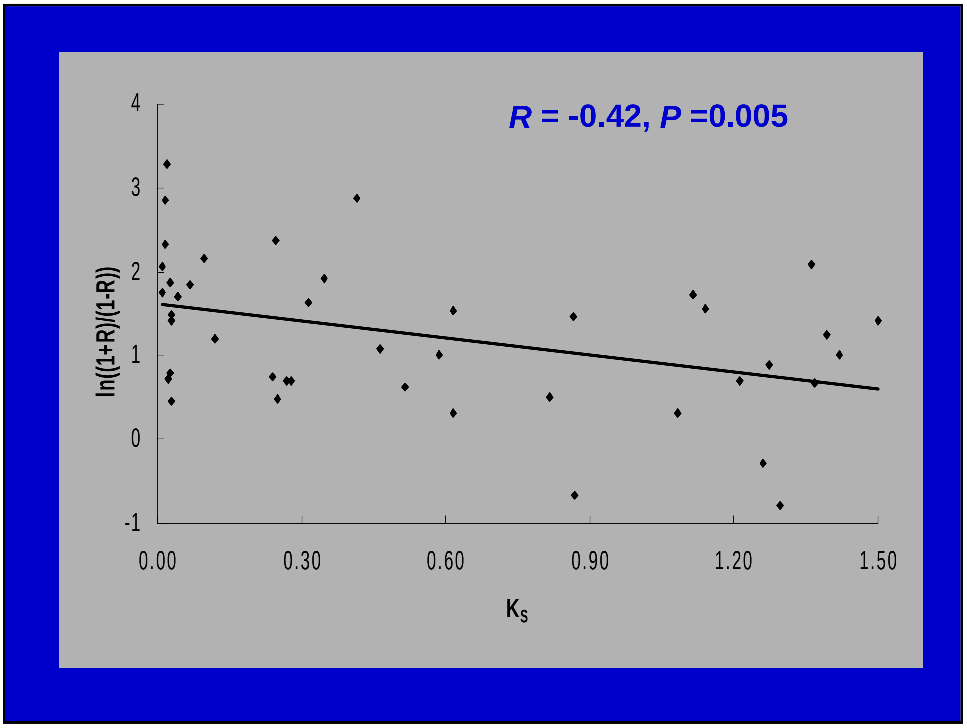


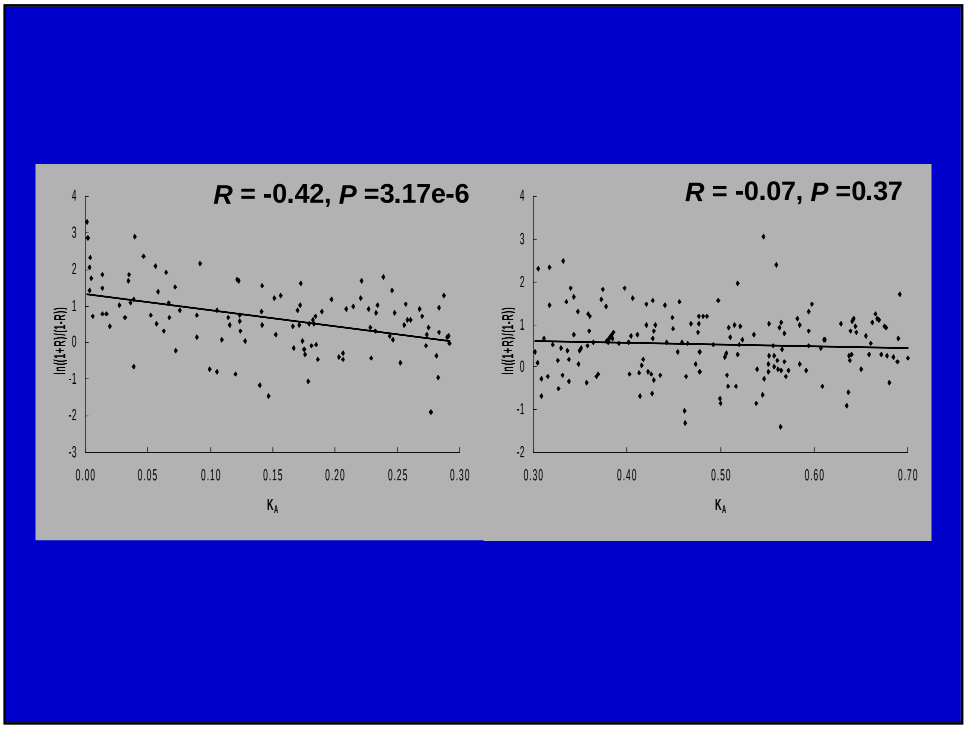


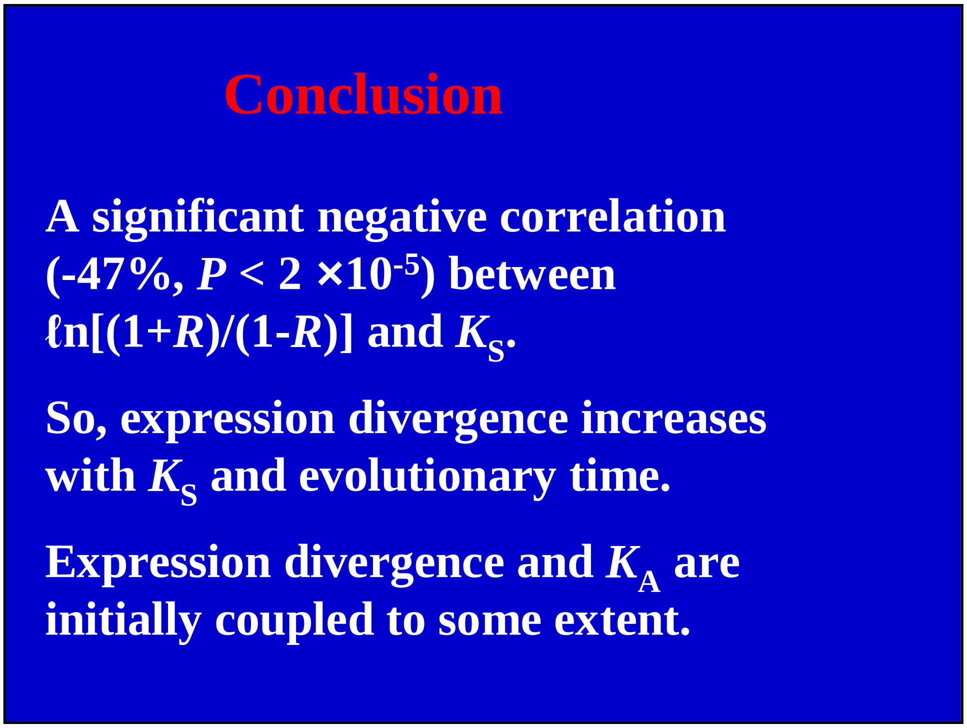


Data: Affymetrix data 79 points

Process	# data Points
Environmental changes	36
Mitotic cell cycle	16
Histone 4 deletio	7
Sporulation	3
Starvation	7


R = -0.42, P =0.005
4
3
2
1
0
-1
ln((1+R)/(1-R))
0.00
0.30
0.60
0.90
1.20
1.50
K_S


R = -0.42, P =3.17e-6
4
3
2
1
0
-1
-2
-3
ln((1+R)/(1-R))
0.00
0.05
0.10
0.15
0.20
0.25
0.30
K_A
R = -0.07, P =0.37
4
3
2
1
0
-1
-2
ln((1+R)/(1-R))
0.30
0.40
0.50
0.60
0.70
K_A

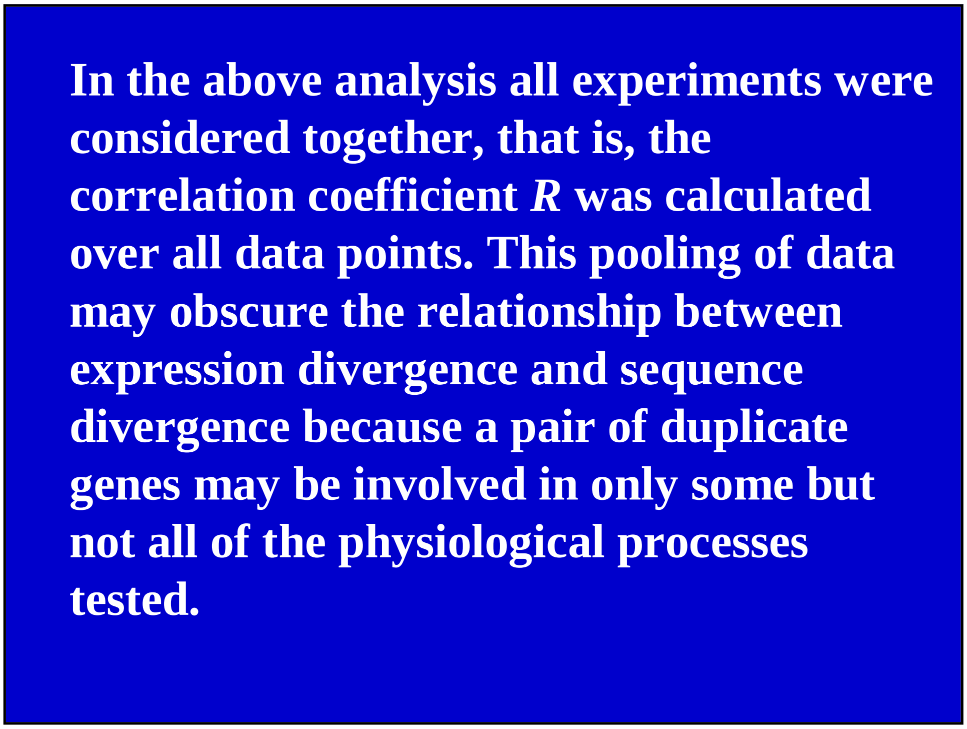


Conclusion

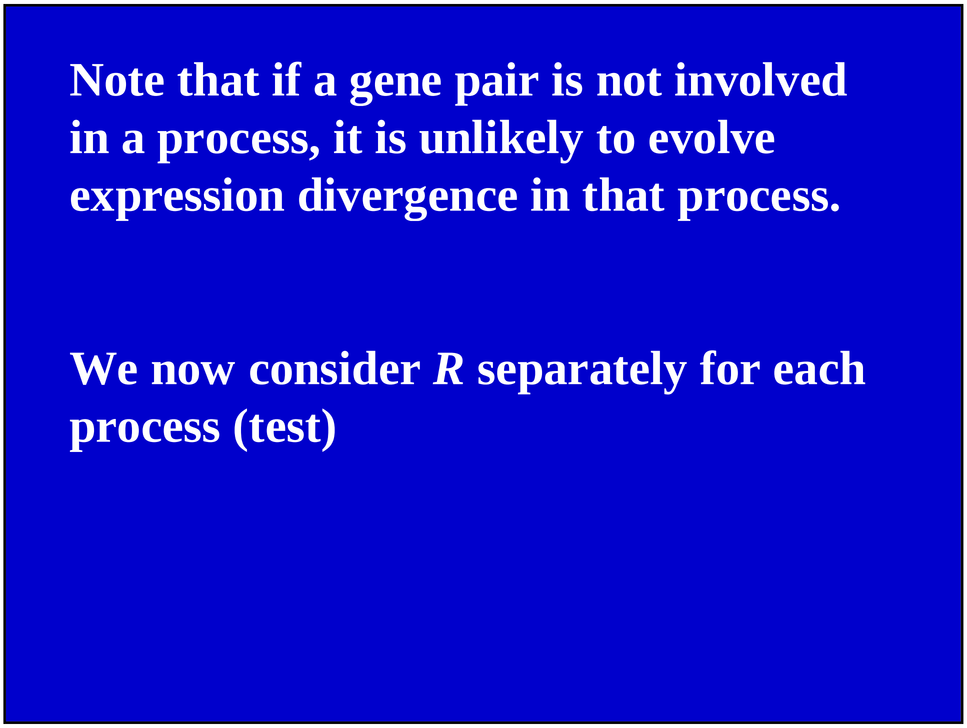
A significant negative correlation (-47%, $P < 2 \times 10^{-5}$) between $\ell n[(1+R)/(1-R)]$ and K_S.

So, expression divergence increases with K_S and evolutionary time.

Expression divergence and K_A are initially coupled to some extent.

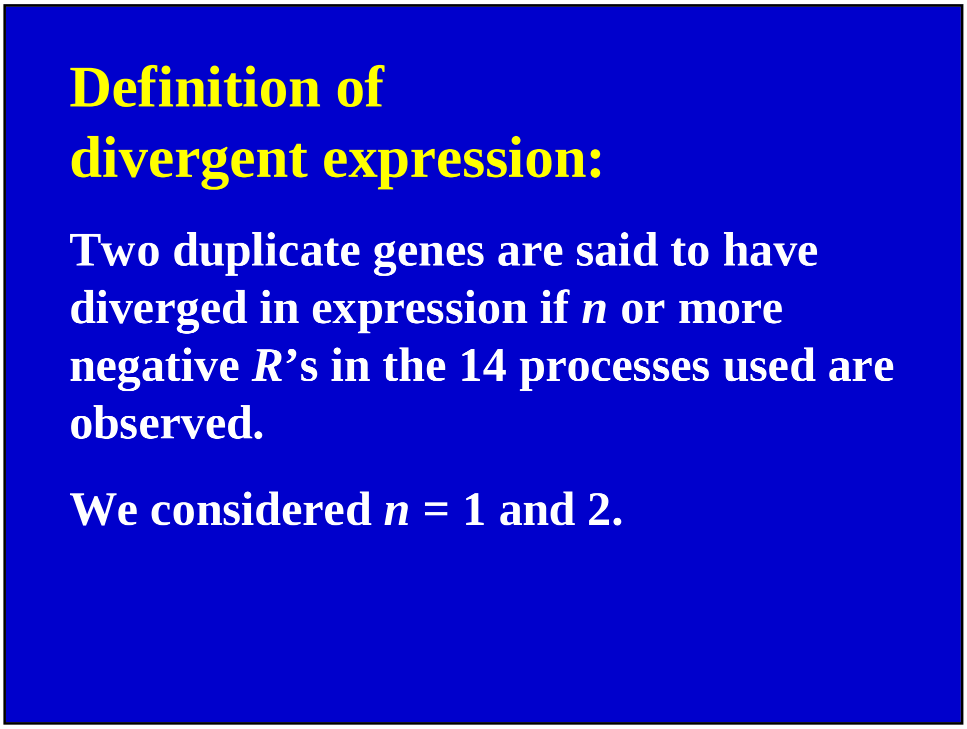


In the above analysis all experiments were considered together, that is, the correlation coefficient *R* was calculated over all data points. This pooling of data may obscure the relationship between expression divergence and sequence divergence because a pair of duplicate genes may be involved in only some but not all of the physiological processes tested.



Note that if a gene pair is not involved in a process, it is unlikely to evolve expression divergence in that process.

We now consider *R* separately for each process (test)



Definition of divergent expression:

Two duplicate genes are said to have diverged in expression if *n* or more negative *R*'s in the 14 processes used are observed.

We considered *n* = 1 and 2.

A sliding window analysis was used when the 14 processes used were treated separately.

For the gene pairs within the surrounding K_S (±0.25) or K_A (±0.05) window of each studied duplicate gene pair, the proportion of gene pairs with divergent expression is calculated.

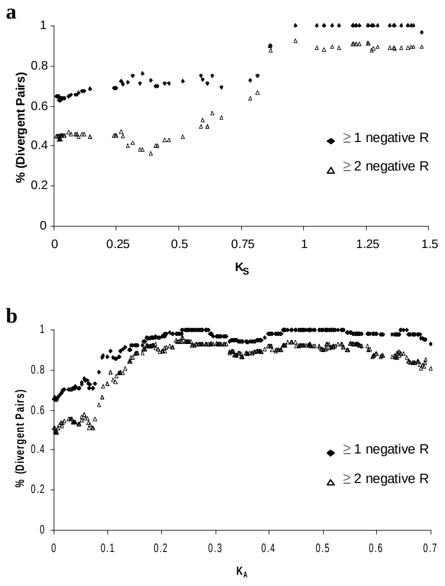


Figure 2

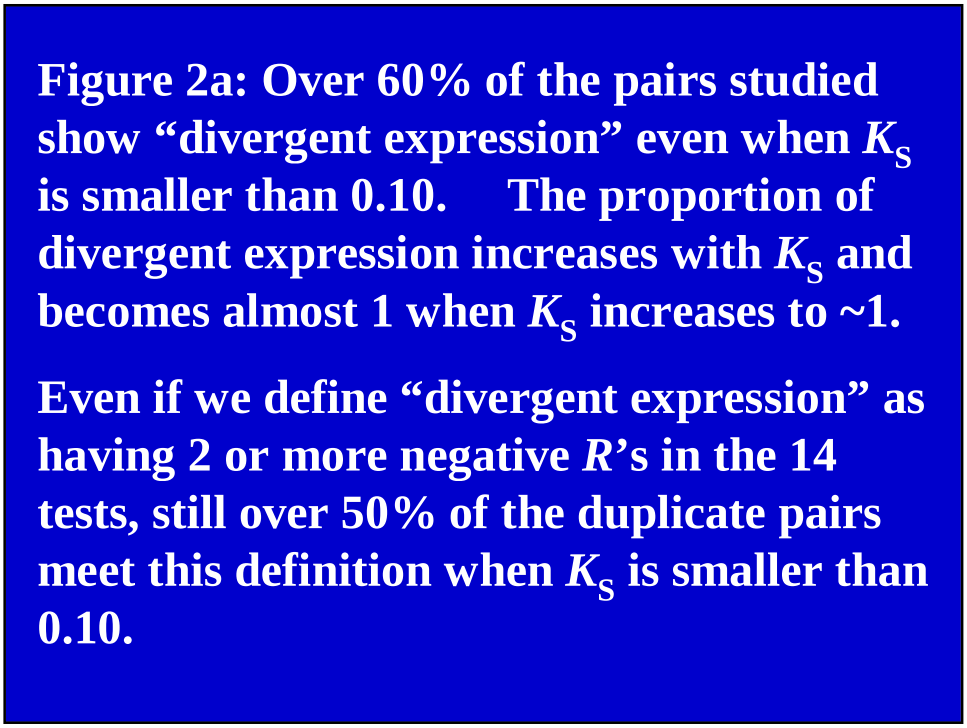
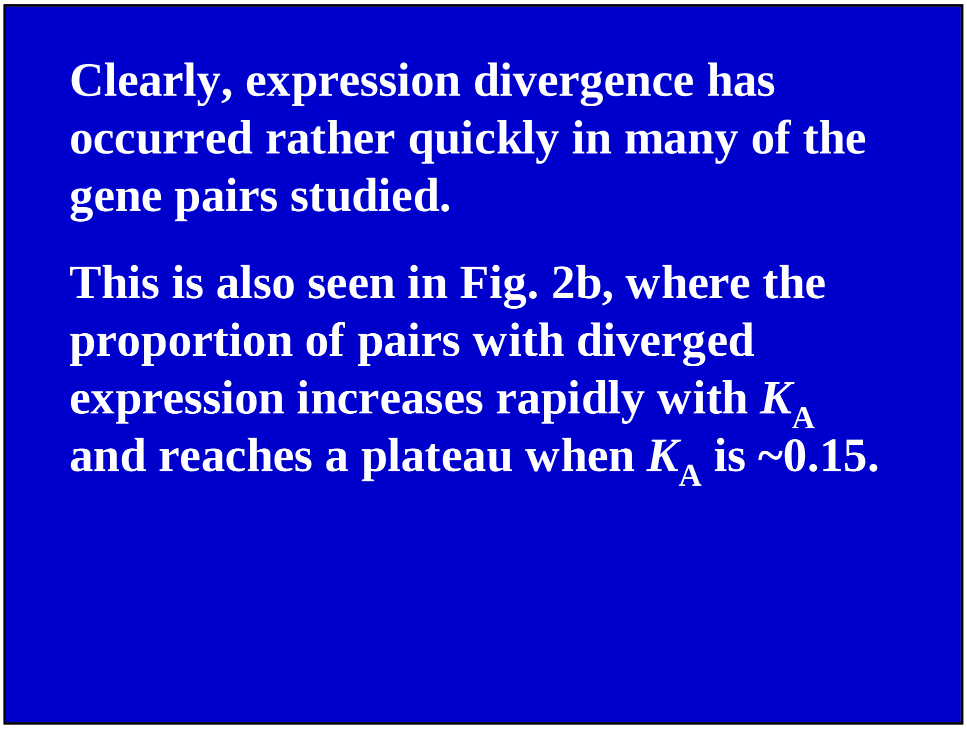


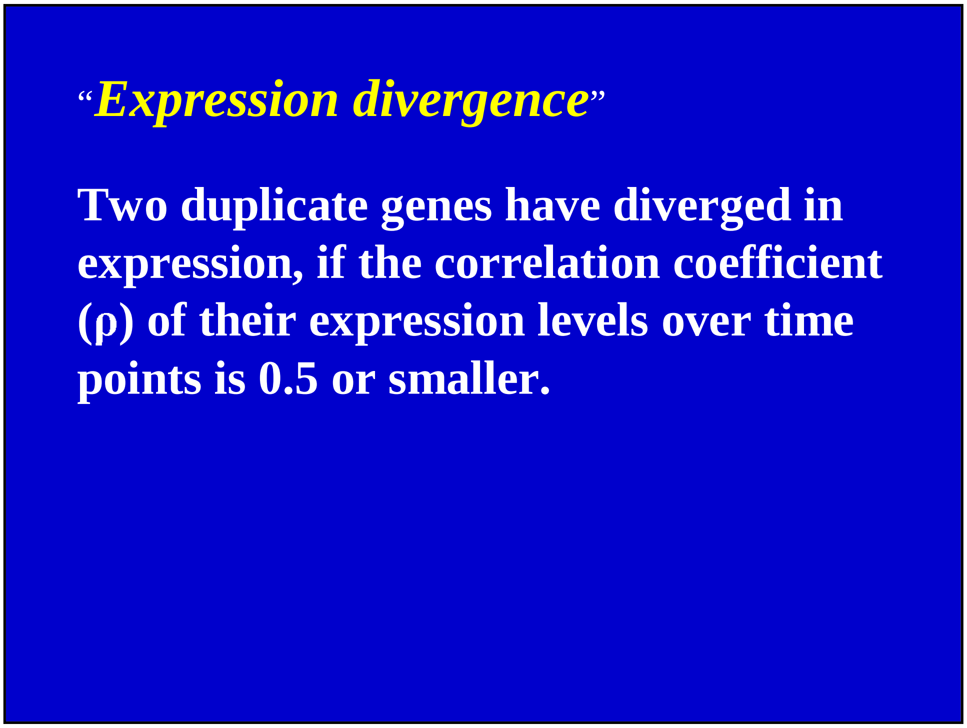
Figure 2a: Over 60% of the pairs studied show "divergent expression" even when *K*$_S$ is smaller than 0.10. The proportion of divergent expression increases with *K*$_S$ and becomes almost 1 when *K*$_S$ increases to ~1.

Even if we define "divergent expression" as having 2 or more negative *R*'s in the 14 tests, still over 50% of the duplicate pairs meet this definition when *K*$_S$ is smaller than 0.10.



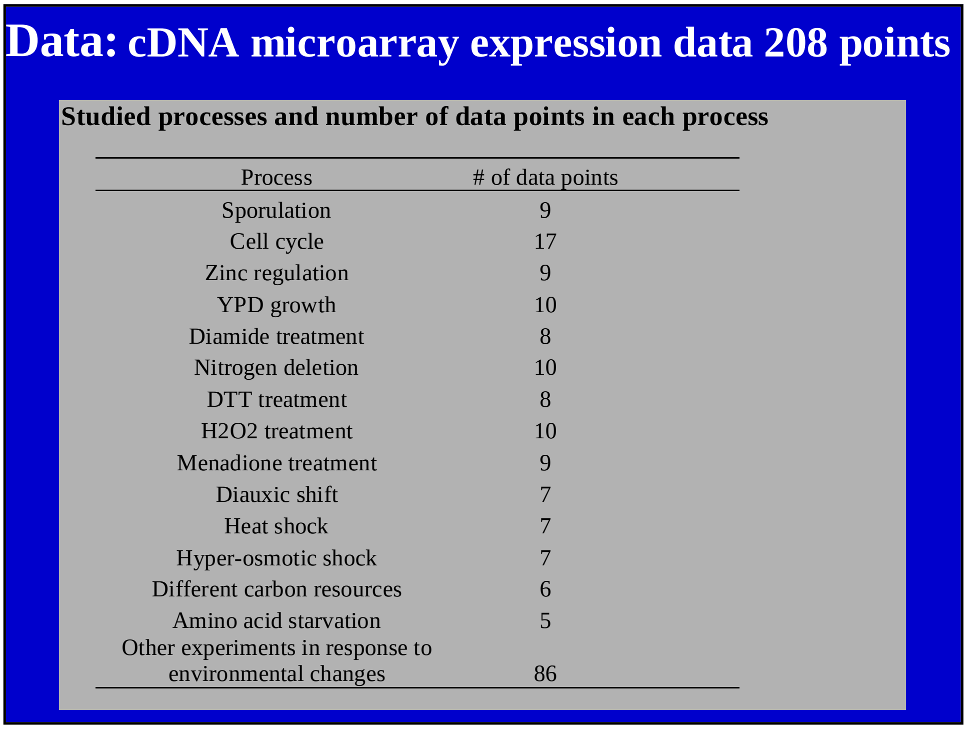
Clearly, expression divergence has occurred rather quickly in many of the gene pairs studied.

This is also seen in Fig. 2b, where the proportion of pairs with diverged expression increases rapidly with *K*$_A$ and reaches a plateau when *K*$_A$ is ~0.15.



"*Expression divergence*"

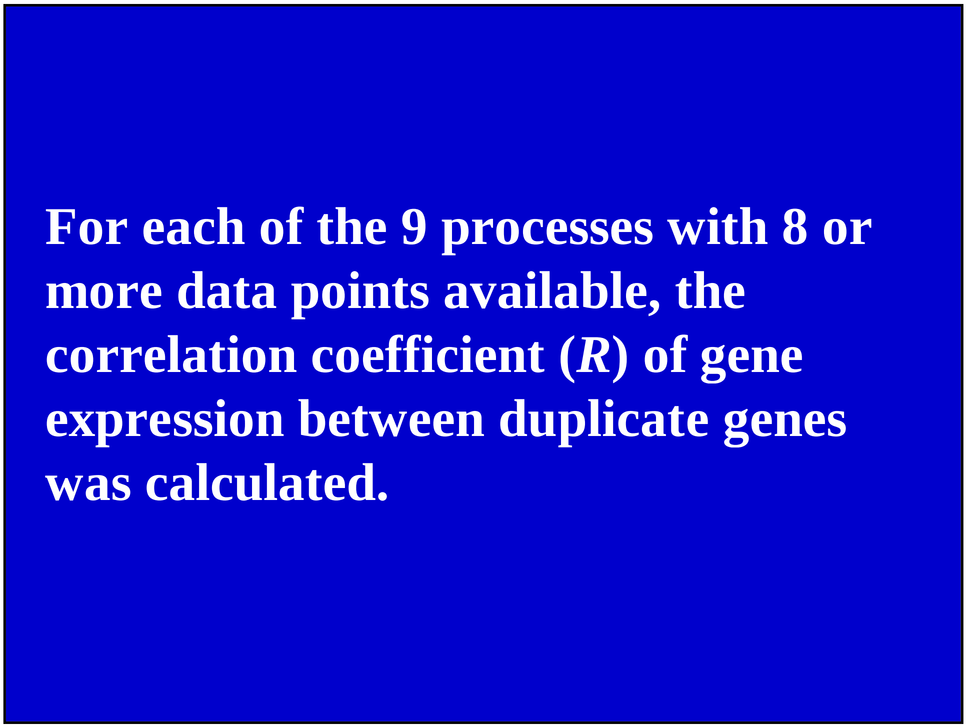
Two duplicate genes have diverged in expression, if the correlation coefficient (ρ) of their expression levels over time points is 0.5 or smaller.



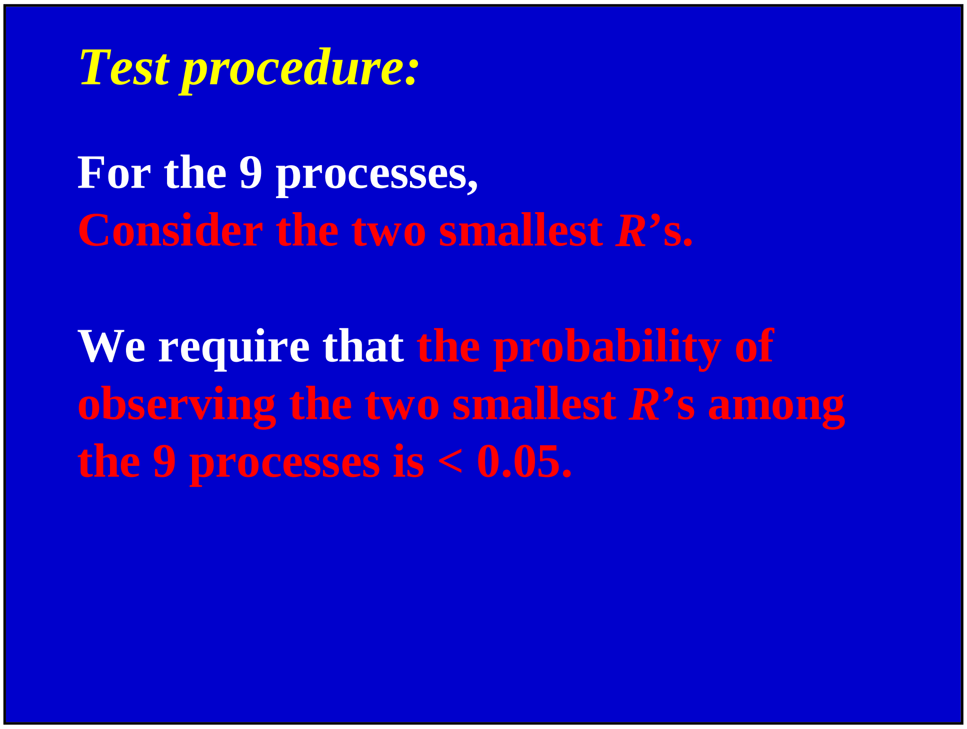
Data: cDNA microarray expression data 208 points

Studied processes and number of data points in each process

Process	# of data points
Sporulation	9
Cell cycle	17
Zinc regulation	9
YPD growth	10
Diamide treatment	8
Nitrogen deletion	10
DTT treatment	8
H2O2 treatment	10
Menadione treatment	9
Diauxic shift	7
Heat shock	7
Hyper-osmotic shock	7
Different carbon resources	6
Amino acid starvation	5
Other experiments in response to environmental changes	86



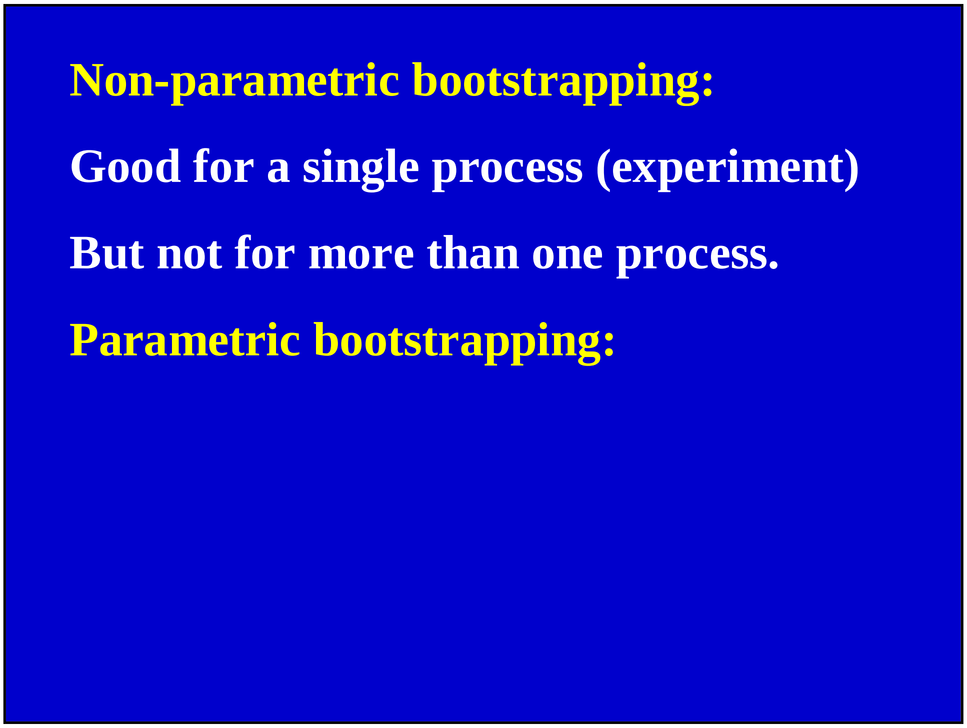
For each of the 9 processes with 8 or more data points available, the correlation coefficient (*R*) of gene expression between duplicate genes was calculated.



Test procedure:

For the 9 processes,
Consider the two smallest *R*'s.

We require that the probability of observing the two smallest *R*'s among the 9 processes is < 0.05.

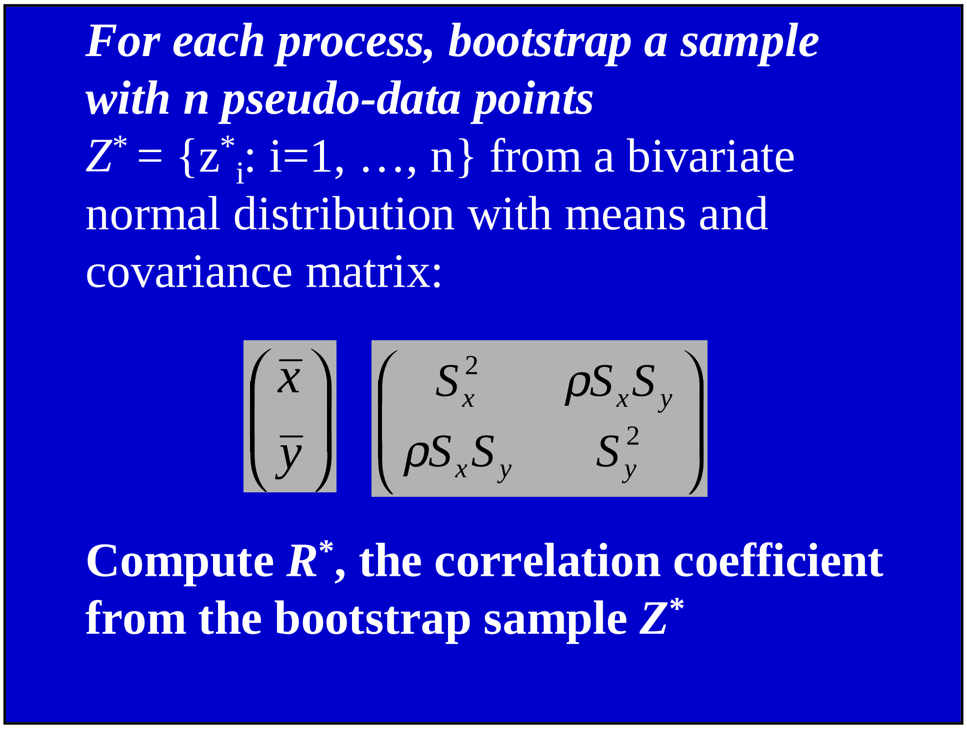


Non-parametric bootstrapping:

Good for a single process (experiment)

But not for more than one process.

Parametric bootstrapping:

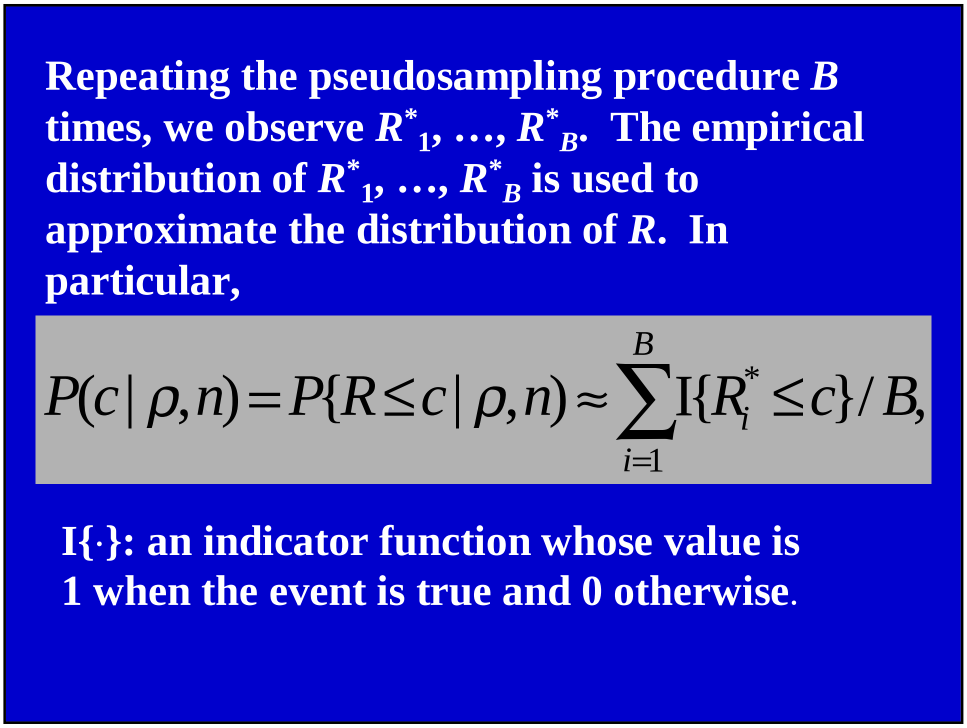


For each process, bootstrap a sample with n pseudo-data points

Z^* = {z^*_i: i=1, …, n} from a bivariate normal distribution with means and covariance matrix:

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \quad \begin{pmatrix} S_x^2 & \rho S_x S_y \\ \rho S_x S_y & S_y^2 \end{pmatrix}$$

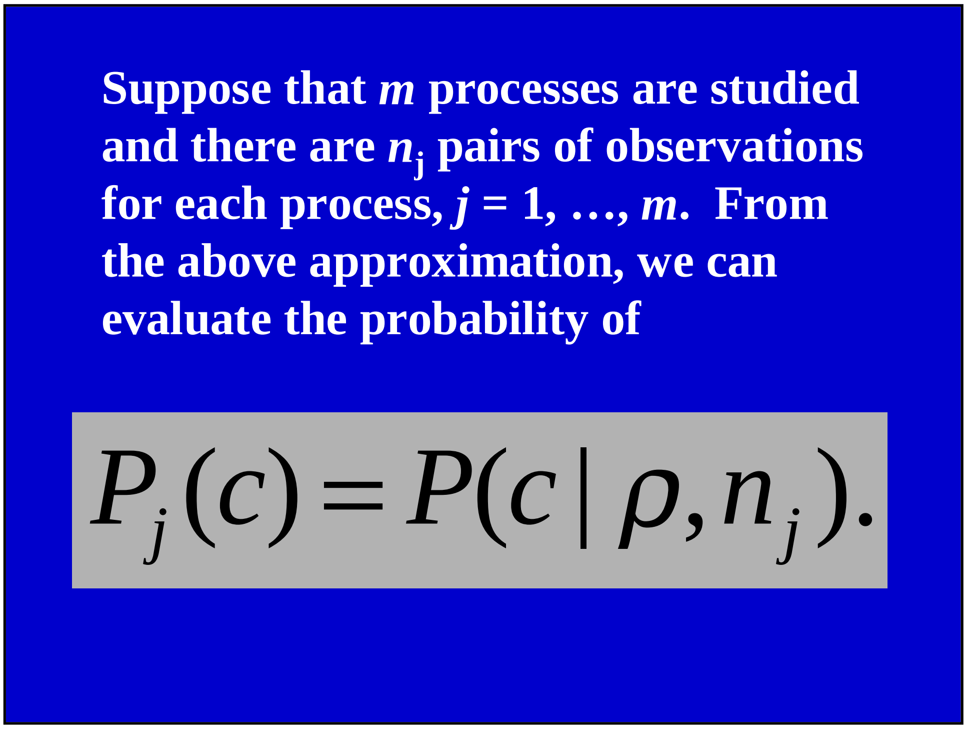
Compute R^*, the correlation coefficient from the bootstrap sample Z^*



Repeating the pseudosampling procedure *B* times, we observe R^*_1, …, R^*_B. The empirical distribution of R^*_1, …, R^*_B is used to approximate the distribution of *R*. In particular,

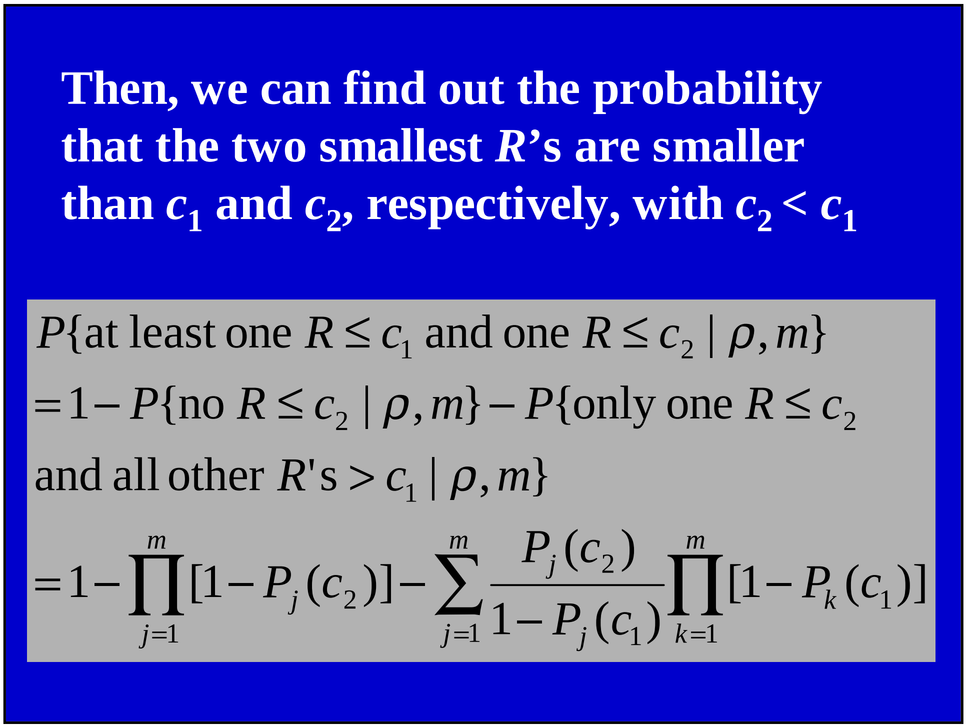
$$P(c\,|\,\rho,n)=P\{R\le c\,|\,\rho,n)\approx\sum_{i=1}^{B}\mathrm{I}\{R_i^*\le c\}/B,$$

I{·}: an indicator function whose value is 1 when the event is true and 0 otherwise.



Suppose that *m* processes are studied and there are n_j pairs of observations for each process, *j* = 1, …, *m*. From the above approximation, we can evaluate the probability of

$$P_j(c)=P(c\,|\,\rho,n_j).$$



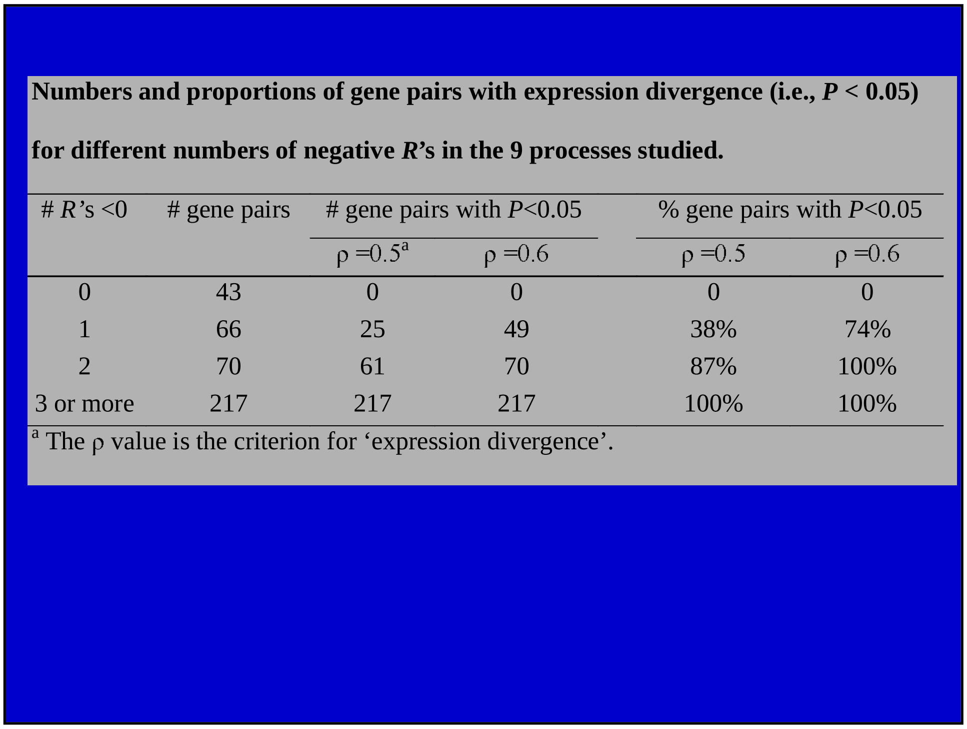
Then, we can find out the probability that the two smallest *R*'s are smaller than c_1 and c_2, respectively, with $c_2 < c_1$

$$P\{\text{at least one } R \le c_1 \text{ and one } R \le c_2 \mid \rho, m\}$$

$$= 1 - P\{\text{no } R \le c_2 \mid \rho, m\} - P\{\text{only one } R \le c_2$$

$$\text{and all other } R\text{'s} > c_1 \mid \rho, m\}$$

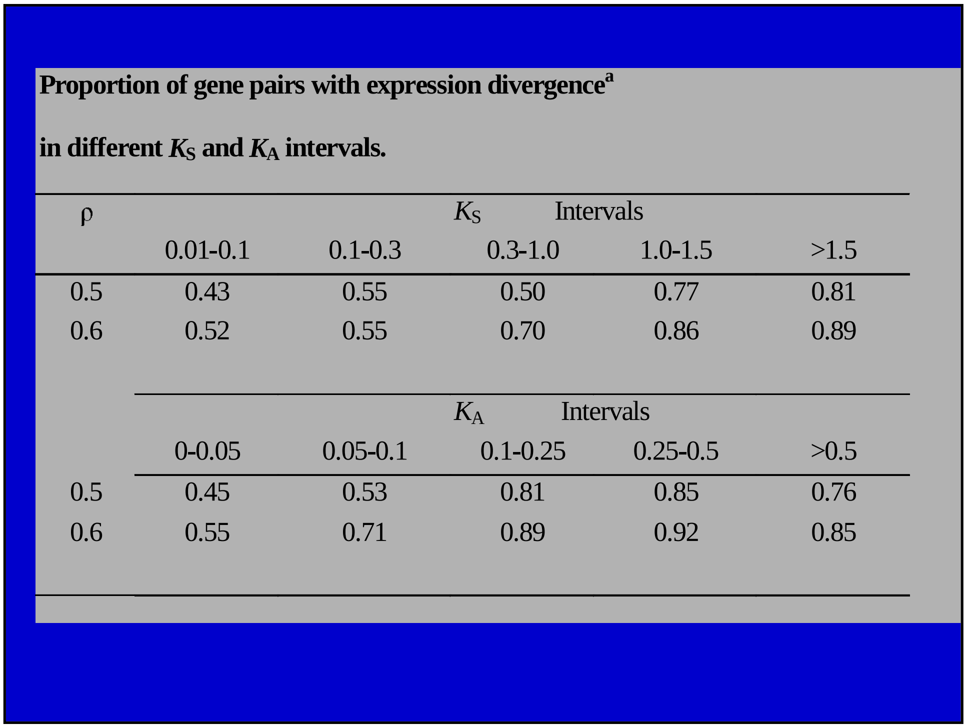
$$= 1 - \prod_{j=1}^{m} [1 - P_j(c_2)] - \sum_{j=1}^{m} \frac{P_j(c_2)}{1 - P_j(c_1)} \prod_{k=1}^{m} [1 - P_k(c_1)]$$



Numbers and proportions of gene pairs with expression divergence (i.e., *P* < 0.05) for different numbers of negative *R*'s in the 9 processes studied.

# *R*'s <0	# gene pairs	# gene pairs with *P*<0.05		% gene pairs with *P*<0.05	
		ρ =0.5[a]	ρ =0.6	ρ =0.5	ρ =0.6
0	43	0	0	0	0
1	66	25	49	38%	74%
2	70	61	70	87%	100%
3 or more	217	217	217	100%	100%

[a] The ρ value is the criterion for 'expression divergence'.

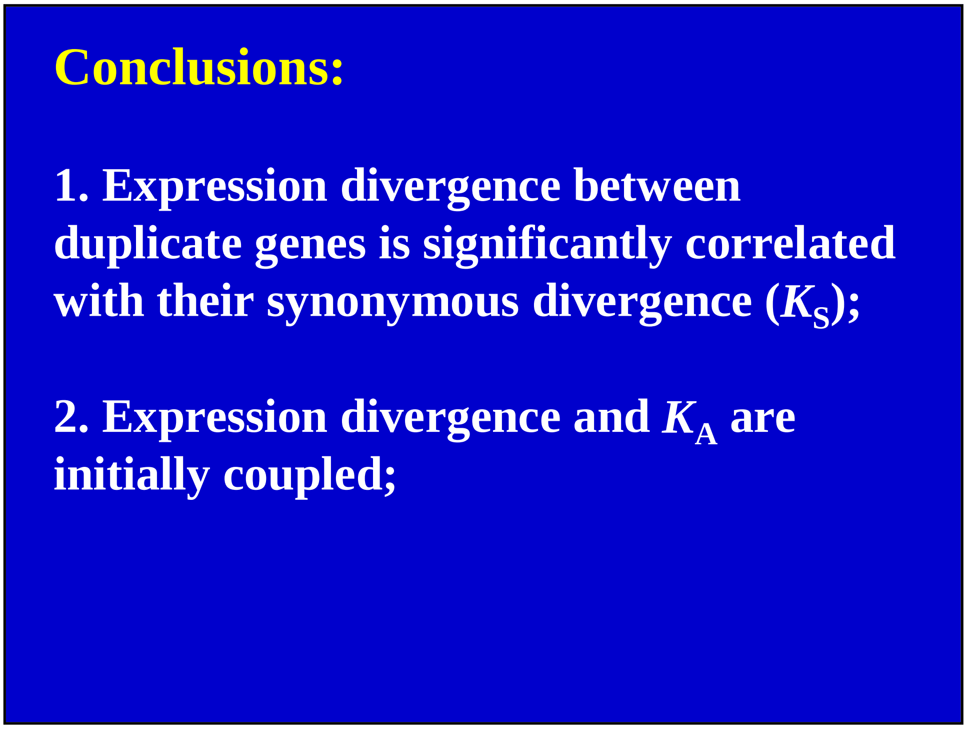


Proportion of gene pairs with expression divergence[a]

in different K_S and K_A intervals.

ρ			K_S	Intervals	
	0.01-0.1	0.1-0.3	0.3-1.0	1.0-1.5	>1.5
0.5	0.43	0.55	0.50	0.77	0.81
0.6	0.52	0.55	0.70	0.86	0.89

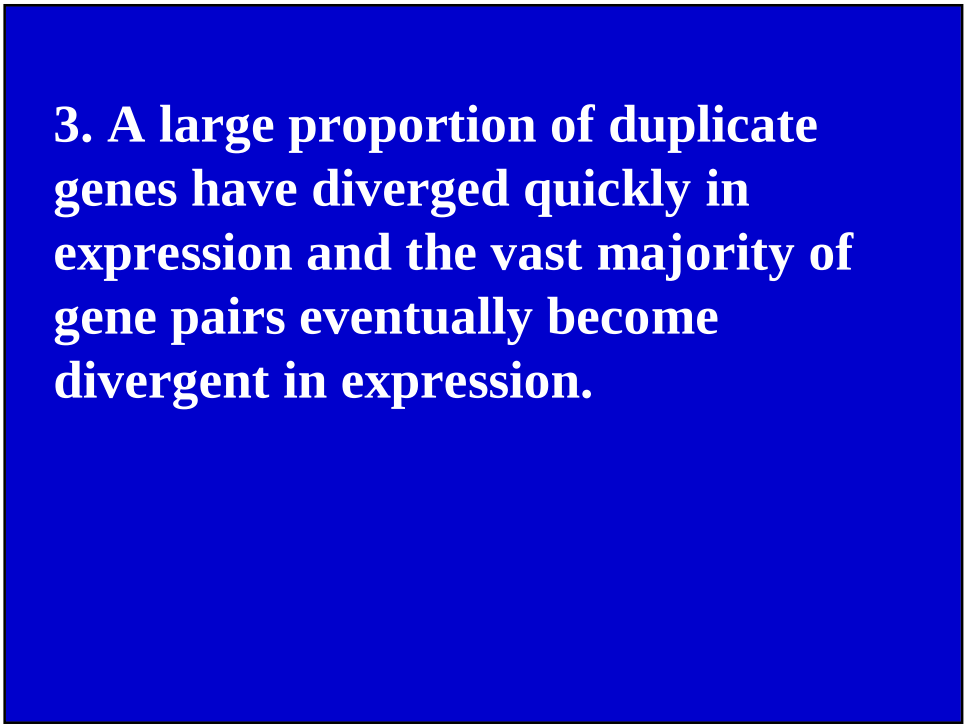
			K_A	Intervals	
	0-0.05	0.05-0.1	0.1-0.25	0.25-0.5	>0.5
0.5	0.45	0.53	0.81	0.85	0.76
0.6	0.55	0.71	0.89	0.92	0.85



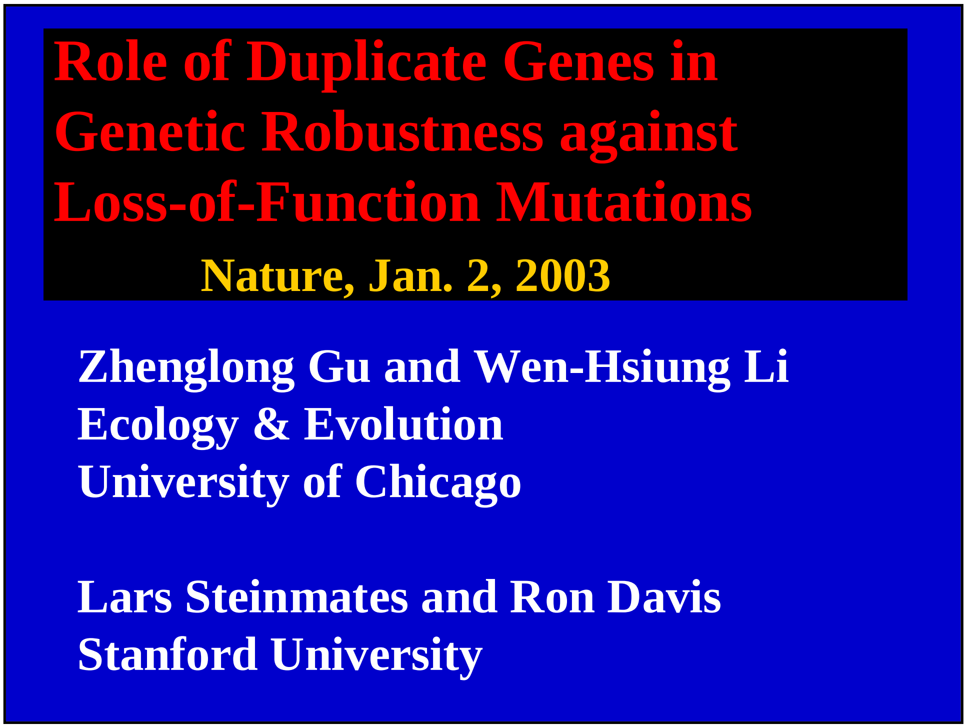
Conclusions:

1. Expression divergence between duplicate genes is significantly correlated with their synonymous divergence (K_S);

2. Expression divergence and K_A are initially coupled;



3. A large proportion of duplicate genes have diverged quickly in expression and the vast majority of gene pairs eventually become divergent in expression.

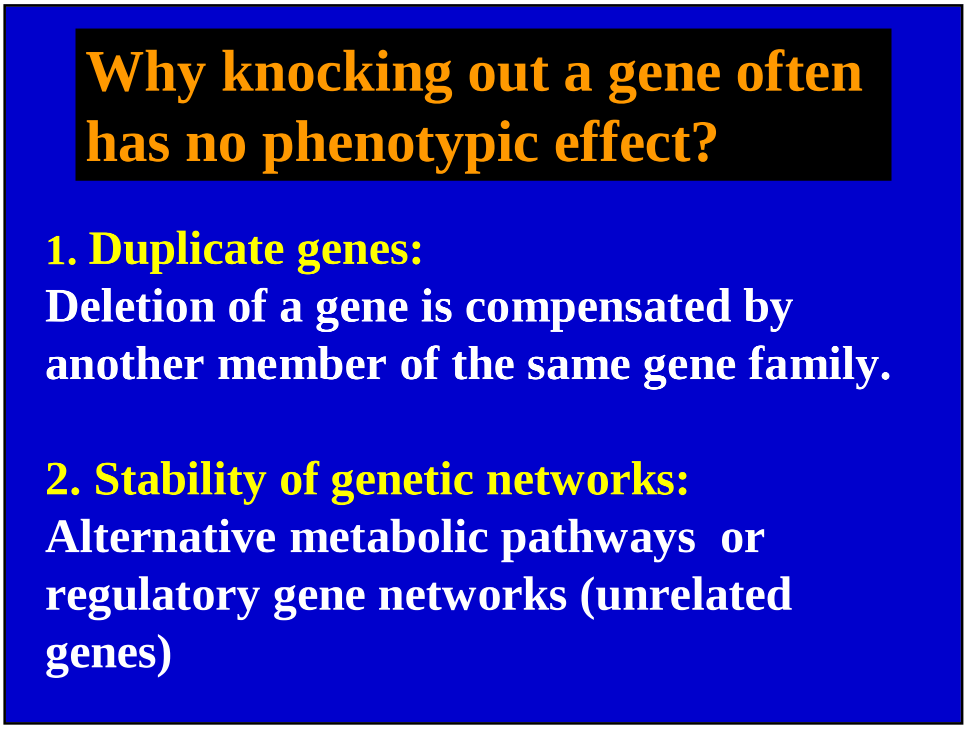


Role of Duplicate Genes in Genetic Robustness against Loss-of-Function Mutations

Nature, Jan. 2, 2003

Zhenglong Gu and Wen-Hsiung Li
Ecology & Evolution
University of Chicago

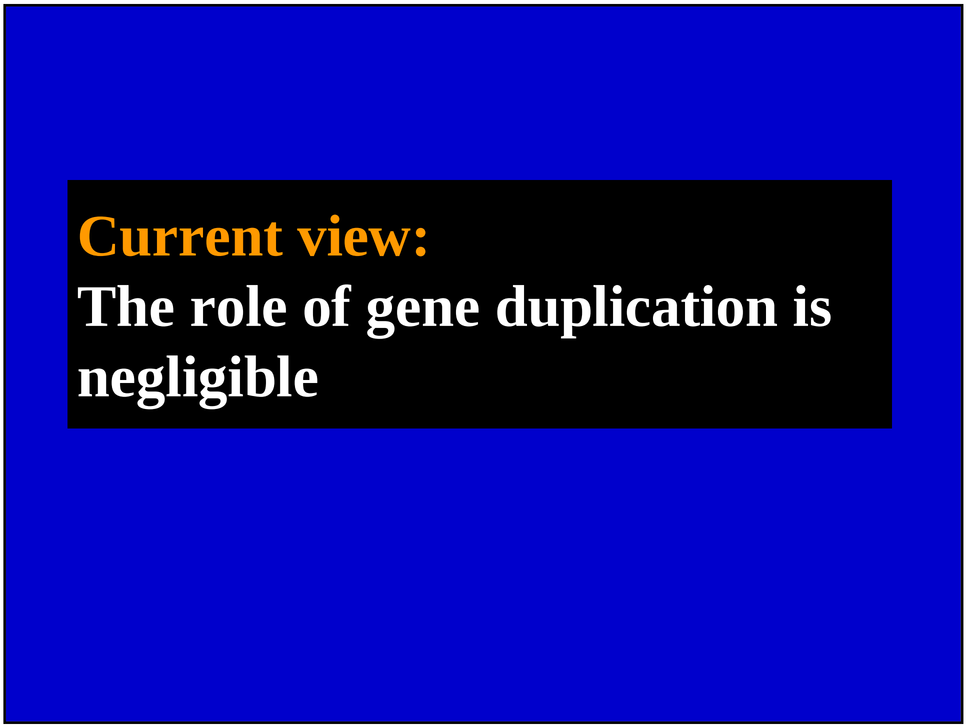
Lars Steinmates and Ron Davis
Stanford University



Why knocking out a gene often has no phenotypic effect?

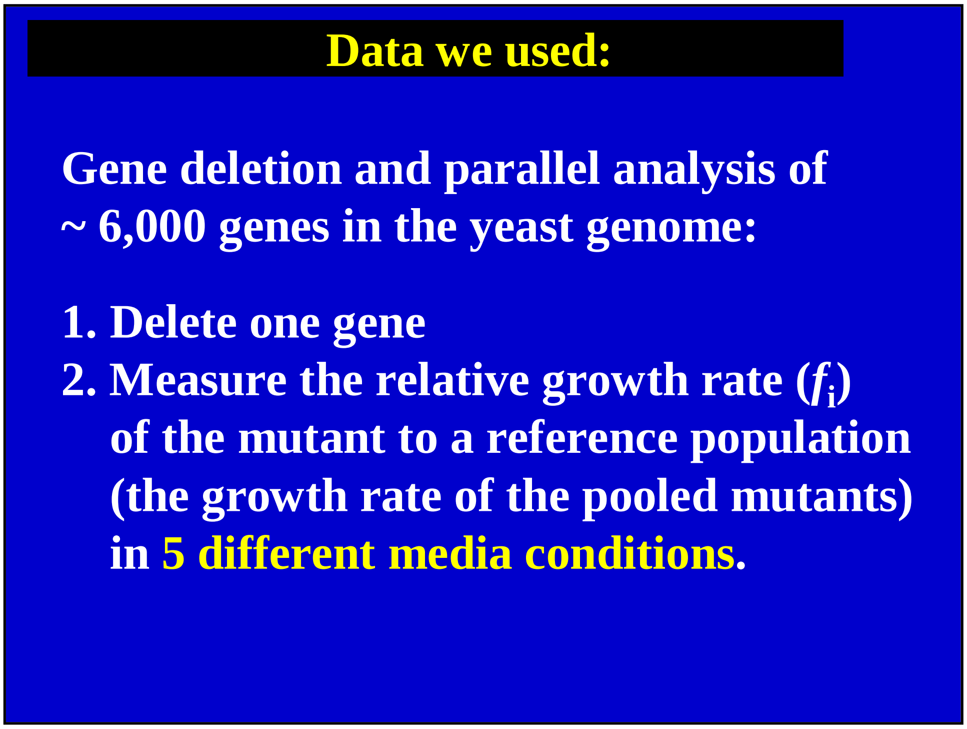
1. Duplicate genes:
Deletion of a gene is compensated by another member of the same gene family.

2. Stability of genetic networks:
Alternative metabolic pathways or regulatory gene networks (unrelated genes)



Current view:

The role of gene duplication is negligible

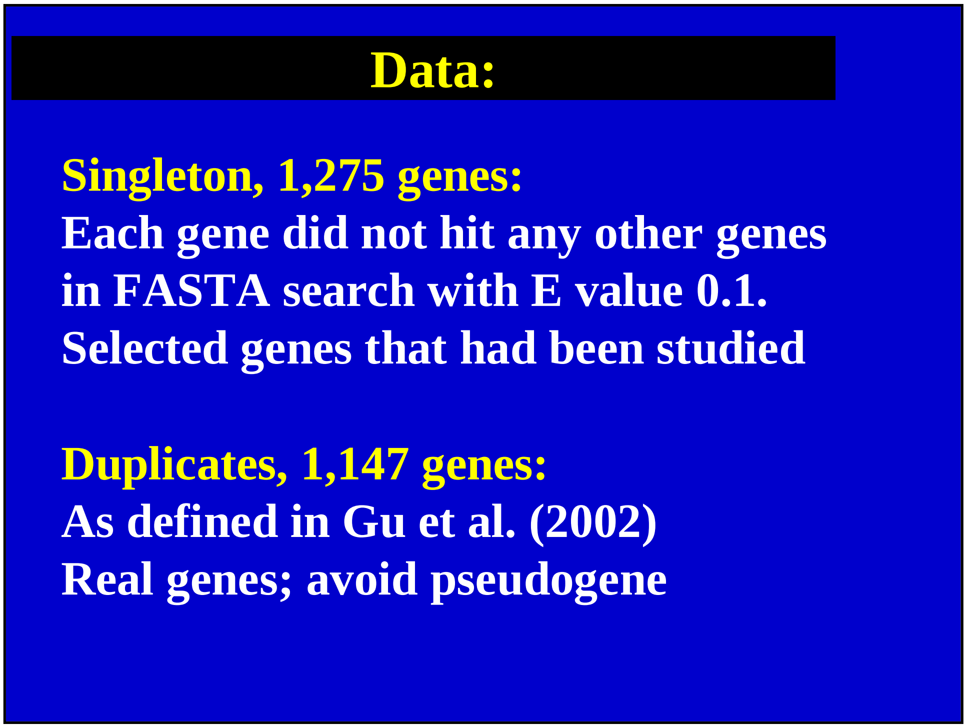


Data we used:

Gene deletion and parallel analysis of ~ 6,000 genes in the yeast genome:

1. Delete one gene

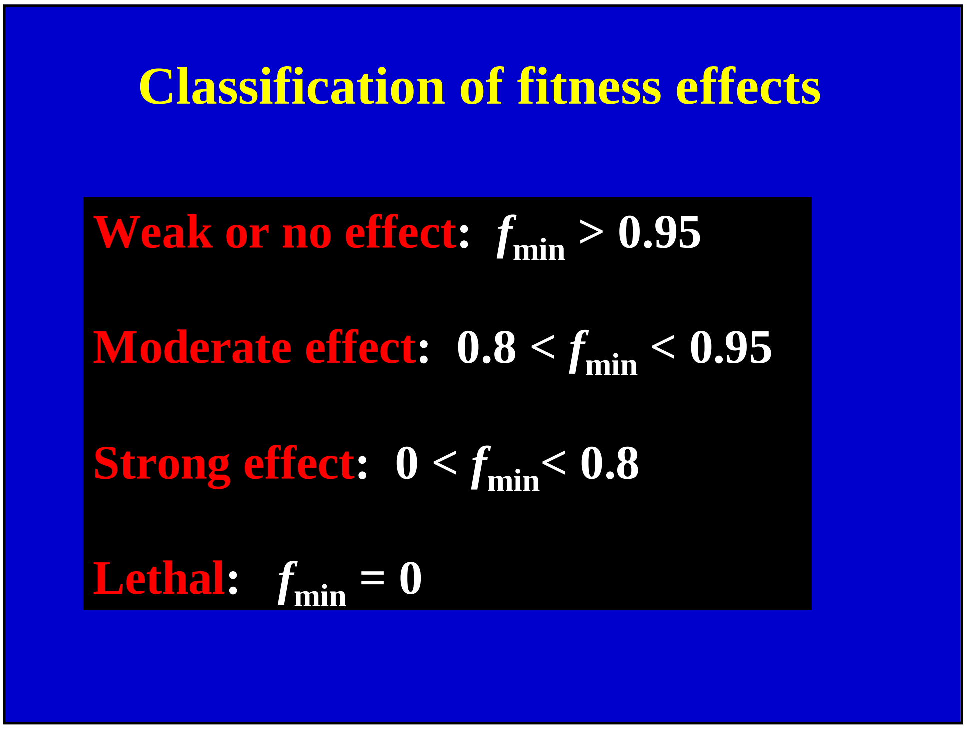
2. Measure the relative growth rate (f_i) of the mutant to a reference population (the growth rate of the pooled mutants) in 5 different media conditions.



Data:

Singleton, 1,275 genes:
Each gene did not hit any other genes in FASTA search with E value 0.1.
Selected genes that had been studied

Duplicates, 1,147 genes:
As defined in Gu et al. (2002)
Real genes; avoid pseudogene



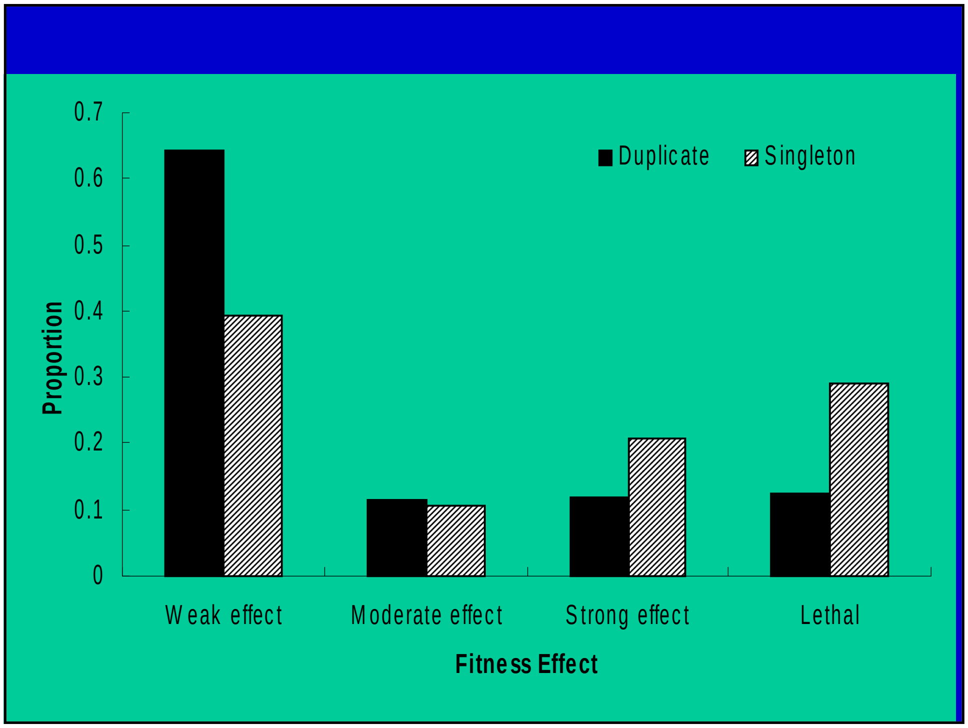
Classification of fitness effects

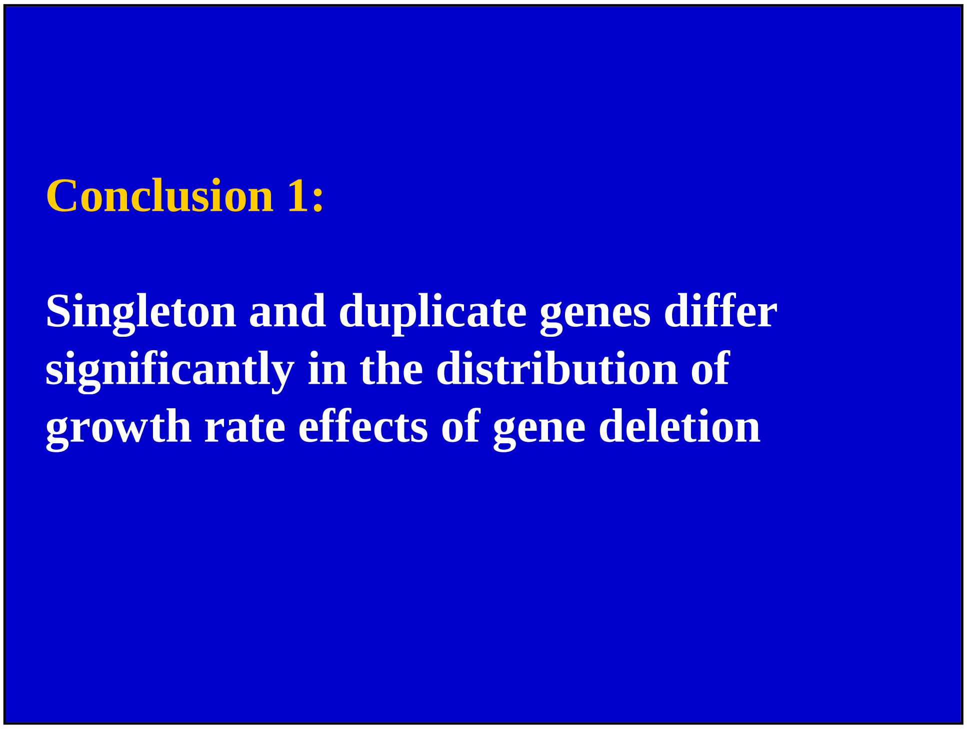
Weak or no effect: $f_{\mathrm{min}} > 0.95$

Moderate effect: $0.8 < f_{\mathrm{min}} < 0.95$

Strong effect: $0 < f_{\mathrm{min}} < 0.8$

Lethal: $f_{\mathrm{min}} = 0$





Conclusion 1:

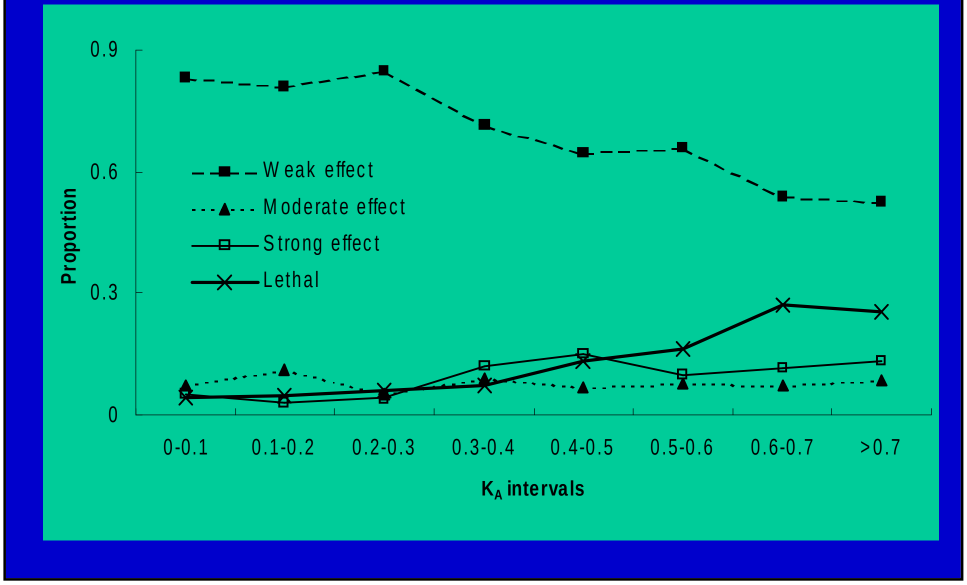
Singleton and duplicate genes differ significantly in the distribution of growth rate effects of gene deletion

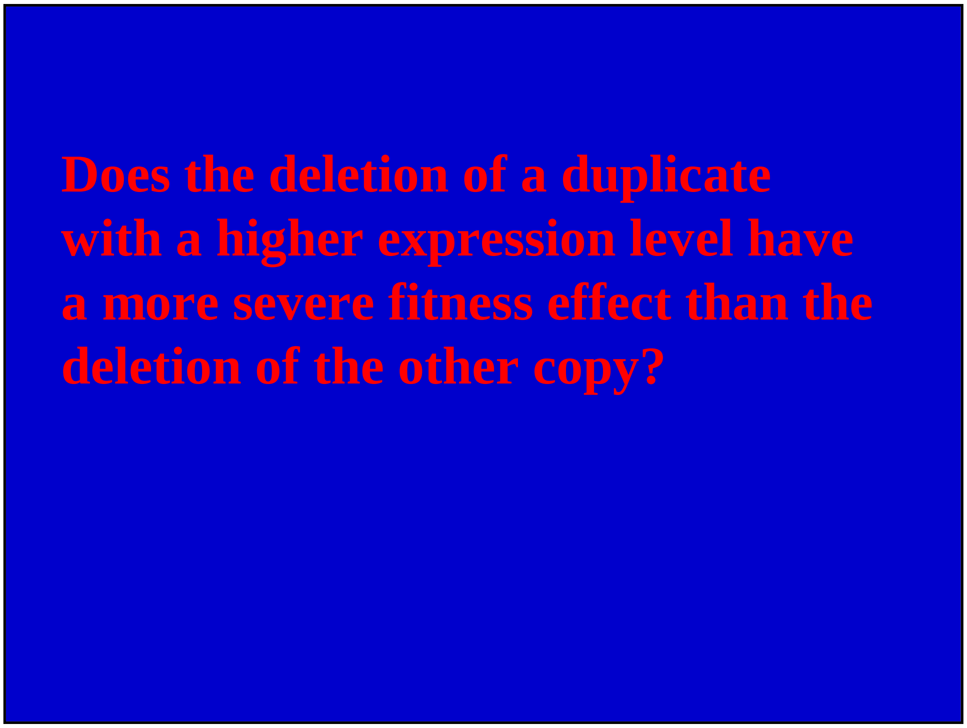
Hypothesis: Genes with closer homologs are compensated more often

1. Divide duplicate genes into different groups using the K_A value of each duplicate gene to its most similar homolog in the genome.

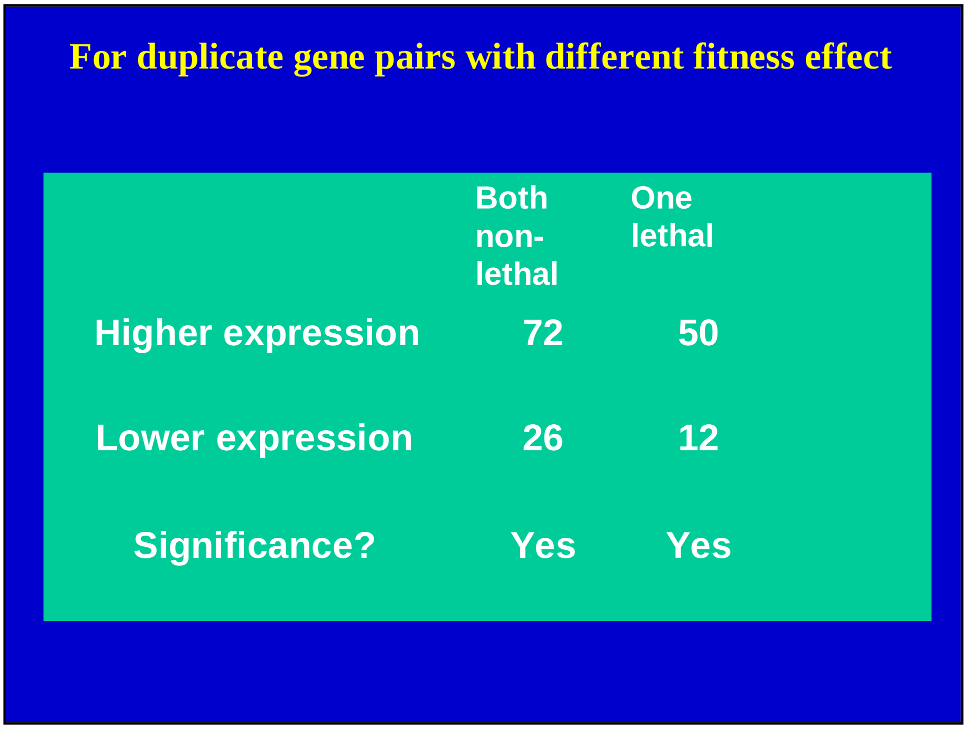
2. Calculate the distribution of fitness effect in each K_A interval.

Relationship between protein distance and fitness effect of deletion





Does the deletion of a duplicate with a higher expression level have a more severe fitness effect than the deletion of the other copy?



For duplicate gene pairs with different fitness effect

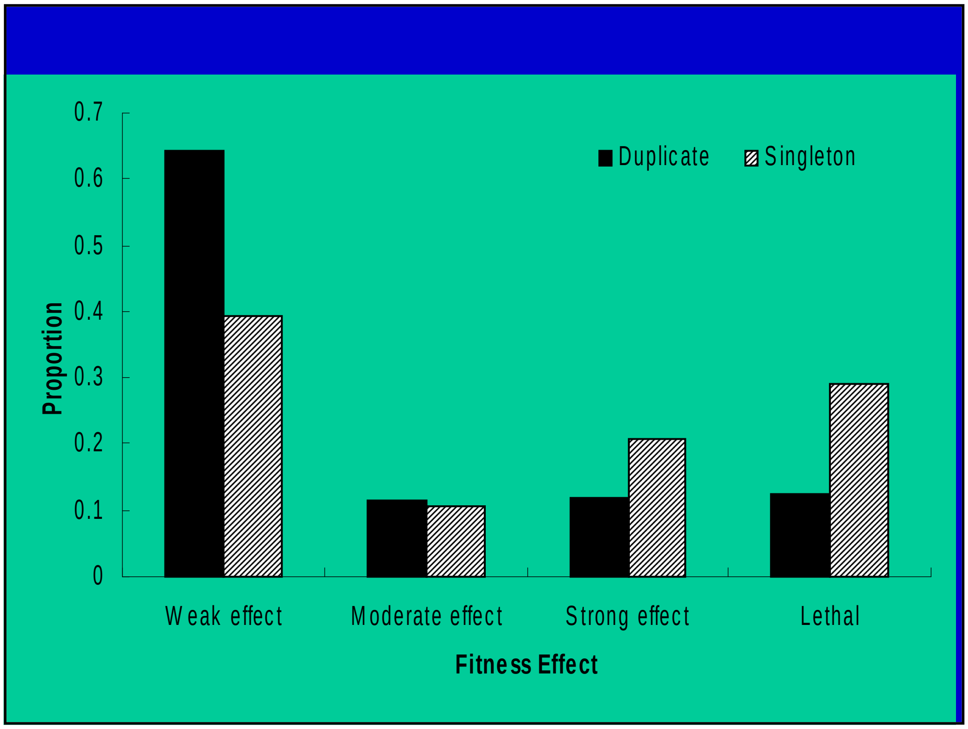
	Both non-lethal	One lethal
Higher expression	72	50
Lower expression	26	12
Significance?	Yes	Yes

Relative contribution of duplicate genes to genetic robustness:

Lower bound (23%):
The extra proportion of duplicate genes with weak or no effects compared to that for singletons is due to genetic redundancy.

284 genes are compensated due to gene duplication:
1,147 duplicates × (64.3% for duplicates – 39.5% for singletons)

Altogether 1,241 genes are compensated:
1,147 duplicates × 64.3% + 1,275 singletons × 39.5%



Upper bound (59%):

All the duplicate genes in the class of weak or no effect are due to genetic redundancy.

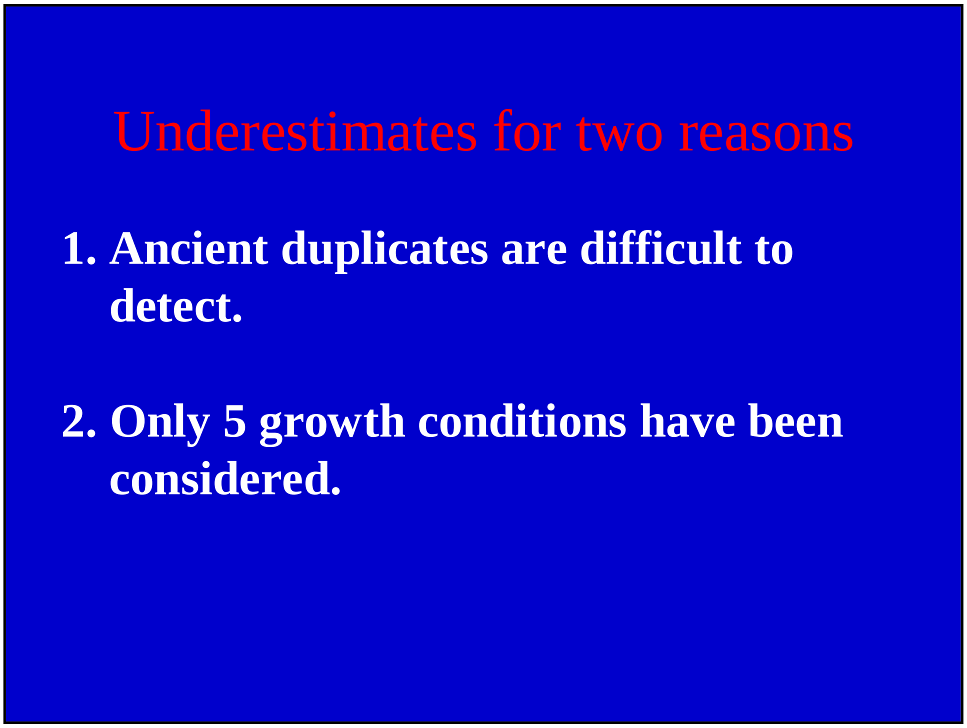
738 duplicate genes (1,147 duplicates × 64.3%) and 503 singleton genes (1,275 singletons × 39.5%) show weak or no effect after deletion

738/(738 + 503) = 59%

Conclusions:

1. **Duplicate genes contribute at least 25% to the genetic robustness against null mutations in the yeast genome**

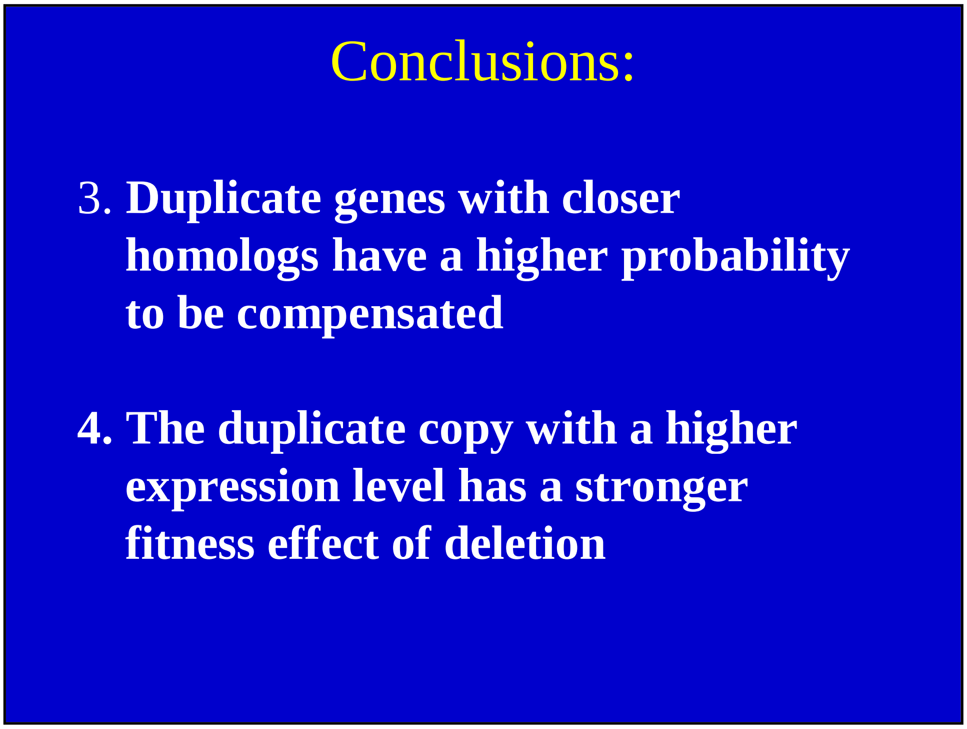
2. **Duplicate genes have more similar fitness effects of gene deletion than singletons**



Underestimates for two reasons

1. **Ancient duplicates are difficult to detect.**

2. **Only 5 growth conditions have been considered.**



Conclusions:

3. **Duplicate genes with closer homologs have a higher probability to be compensated**

4. **The duplicate copy with a higher expression level has a stronger fitness effect of deletion**

Thanks!