PLENARY LECTURE SPONSOR

European Molecular Biology Organization

Sponsoring ● Mentoring ● Networking ● Training

- → PhD & Post-doc students
- → Young Investigators
- → Group Leaders
- → Senior scientists

Promoting excellence in molecular life sciences since 1964

EMBO

www.embo.org

Looking after the neighbourhood: Noise abatement and genome evolution.

Laurence D. Hurst
University of Bath, UK

Martin Lercher
Araxi Urrutia
Adam Pavlicek
Csaba Pal
Elizabeth Williams
Juan Poyatos
Nizar Batada
Karoly Kovacs
Balazs Papp

Is the genome a precisely engineered swiss movement or a mickey mouse watch?

What "..if the different parts had been differently shaped from what they are, or placed after any other manner or in any other order than that in which they are placed…" Paley 1802

Gene order evolution as a test case

Is gene order random, and if not why not?

Prior to the genomic age it was believed that in eukaryotes gene order should be random:

"In eukaryotes... there will not be selection for gene clustering to control gene dosage in eukaryotes"

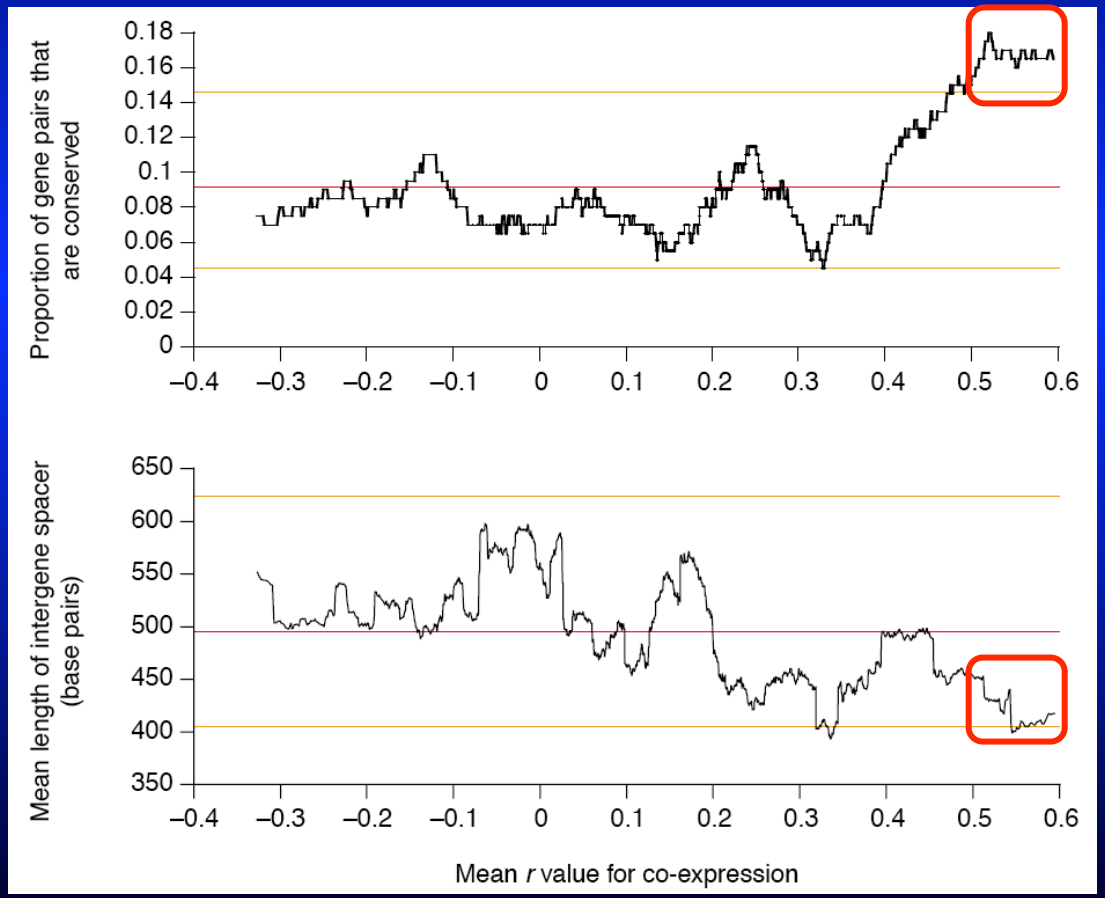
Cavalier-Smith, T.(1993)

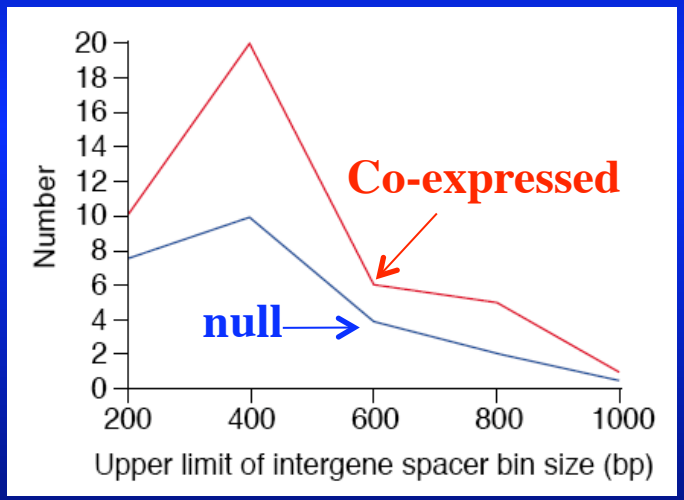
in P. Broda, S. Oliver, and P. Sims, eds. The Eukaryotic Genome: organization and regulation. Cambridge University Press, Cambridge

There is now abundant evidence that, in every well studied eukaryotic genome, gene order and gene expression are coupled e.g.:

- In worm there exist operons
  - Blumenthal et al. 2002 *Nature* 417, 851-854
- In yeast, highly co-expressed genes are linked more often than expected.
  - Cohen et al. 2000, *Nat Genet* 26 183-186
- In Arabidopsis bidirectional promoters probably explain many incidences of local co-expression
  - Williams, E.J.B. and Bowles, D.J. 2004 *Genome Res*. **14:** 1060-1067
- In Drosophila, chromosomal domains of similar expression have been described.
  - Spellman, P. T., and G. M. Rubin, 2002 *J. Biol*. **1:** 5.
- In humans, broadly expressed genes cluster (in GC rich regions)
  - Lercher, M. J., A. O. Urrutia, and L. D. Hurst, 2002. *Nat. Genet*. **31:** 180-183.
  - Lercher, M. J et al., 2003 *Hum. Mol. Genet*. **12:** 2411-2415.

At least some of this order is likely owing to selection favouring conservation of gene order to preserve co-expression





$\chi 2 = 17.0, P = 0.004$

Is selection for co-expression the sole reason that genomes are organized?

- Why are essential genes clustered but not co-expressed?
- Why is gene order in metabolic operons sometimes colinear with the order of reactions in metabolic pathways?

Model: selection on noise in protein levels drives both organizations

- Noise is variation in transcript/protein abundance between otherwise identical cells
- Noise = standard deviation in expression between cells/mean

The inevitability of noise: A. The ubiquity of transcription factor binding sites

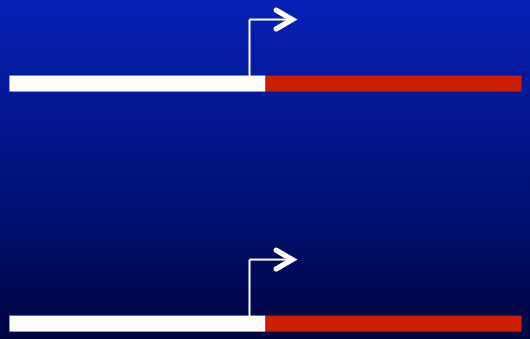
Question: In a sequence of 100 random nucleotides how many TF bindings sites and what proportion of sequence is covered by TF binding sites?

Answer: according to TRANSFAC specification of TF binding sites in 100 bp of human sequence there are on average 15 TF binding sites occupying about 60% of the sequence.

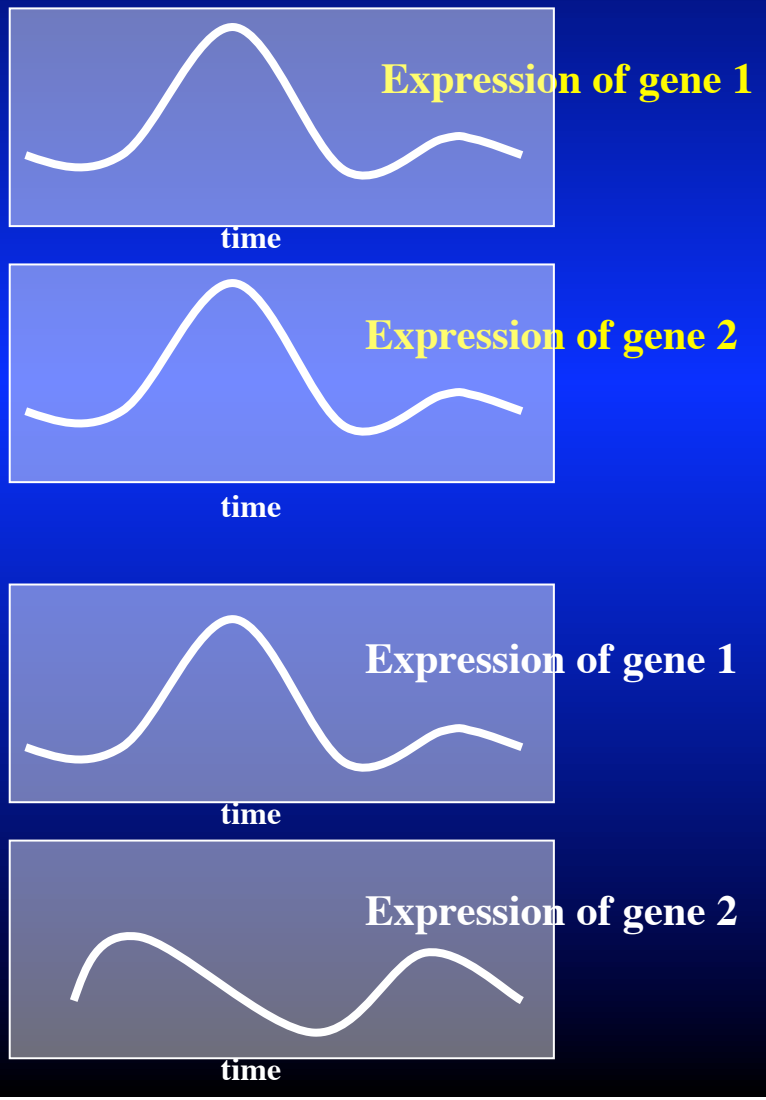
The inevitability of noise: B. Chromatin opening causes de facto co-expression of linked genes

a) Two tandem transgenes

b) The same two transgenes but unlinked



Raj, A. et al. 2006. Plos Biology **4**:e309.



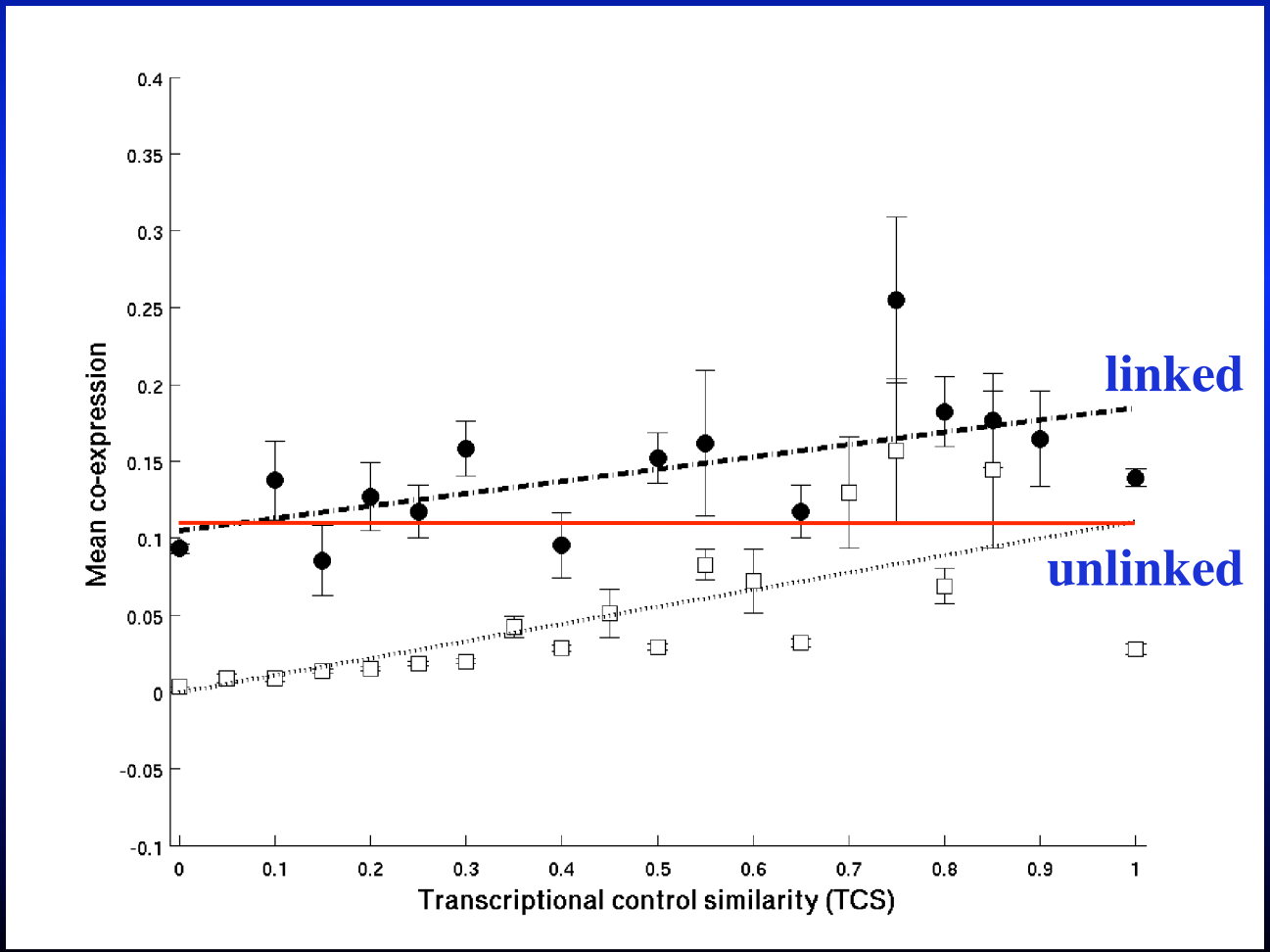
How important is this? Do linked genes have higher co-expression than unlinked genes with similar transcriptional control?

- Take complete yeast genome and microarray expression data (23 time course experiments).

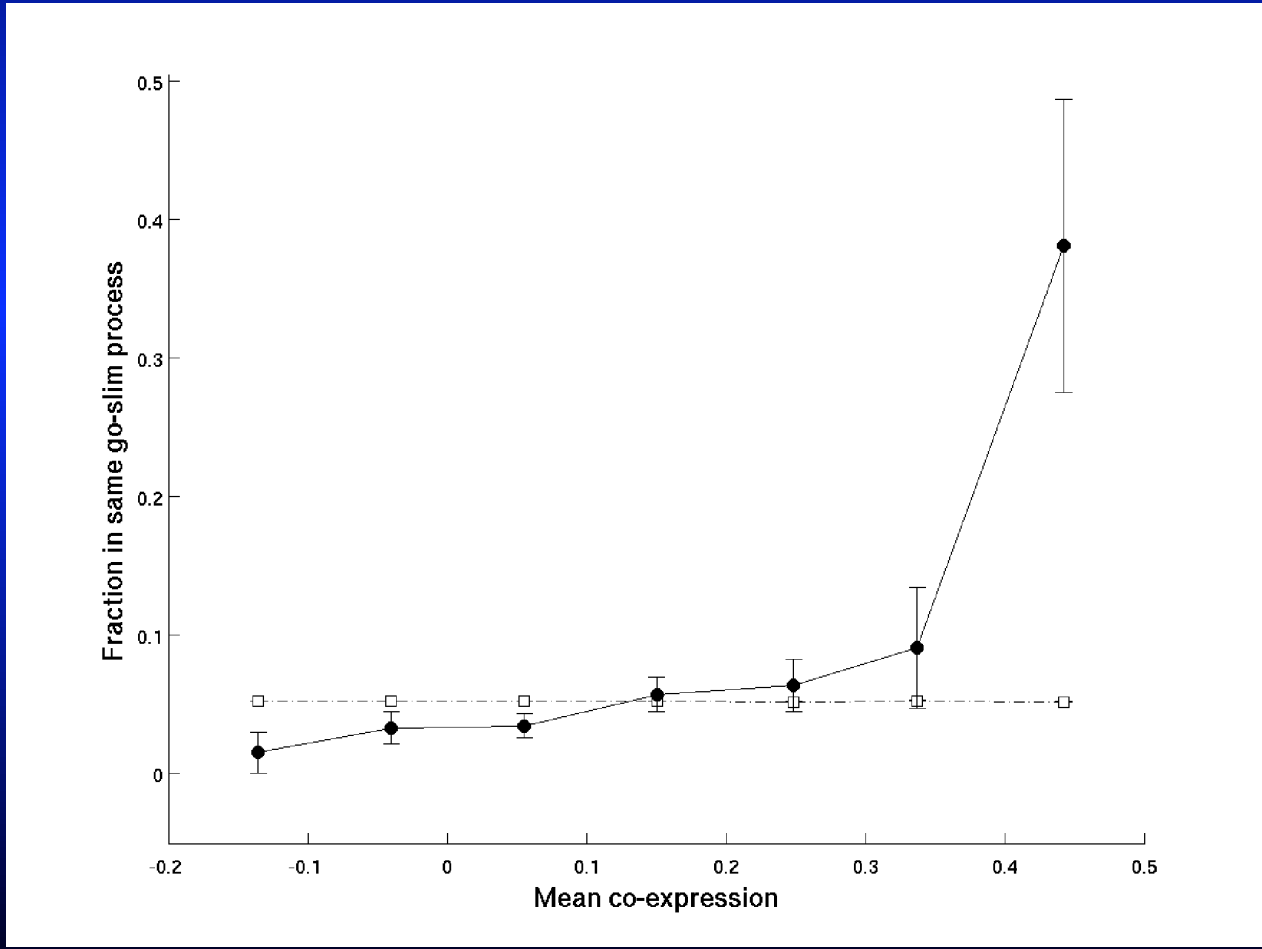
- For all gene pairs (nearest neighbours and non-neighbours), calculate *r*, the correlation coefficient, from expression data - this is the measure of co-expression.

- From data on 157 transcription factors, 4410 genes and 12,873 regulatory interactions, calculate transcriptional control similarity between all gene pairs

Controlling for transcriptional control similarity linked genes show much higher co-expression rates



Only for the very most highly co-expressed genes are genes more similar in their function than expected by chance.



Co-expression of linked genes supports the view that gene expression is intrinsically noisy

- Much low to moderate co-expression of linked genes is likely a side consequence of chromatin opening/closing.
- Consistent with this we find that chromosomal domains associated with fluctuating chromatin have slightly higher co-expression of linked genes (data not shown).
- Only for very highly co-expressed gene pairs need we suspect selection for functional co-ordination.
- Conversely, most co-expression of linked genes is better explained as functionally irrelevant noise.

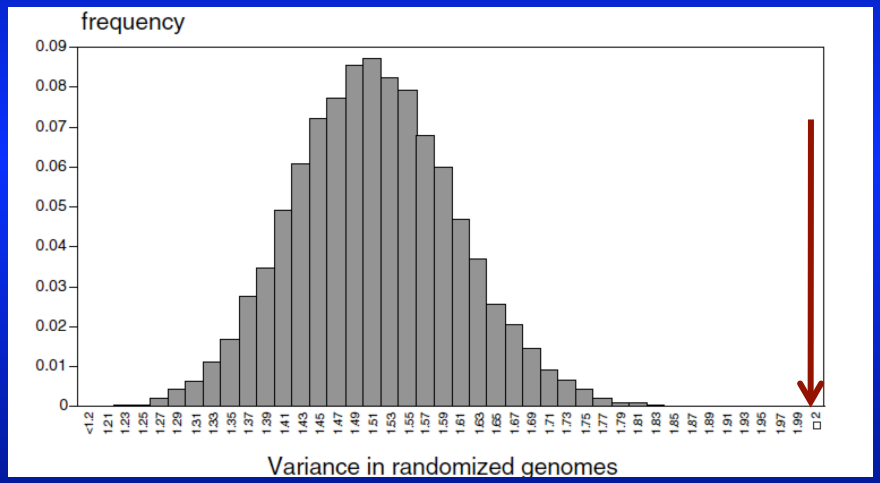
Case history I
Why might essential genes cluster?

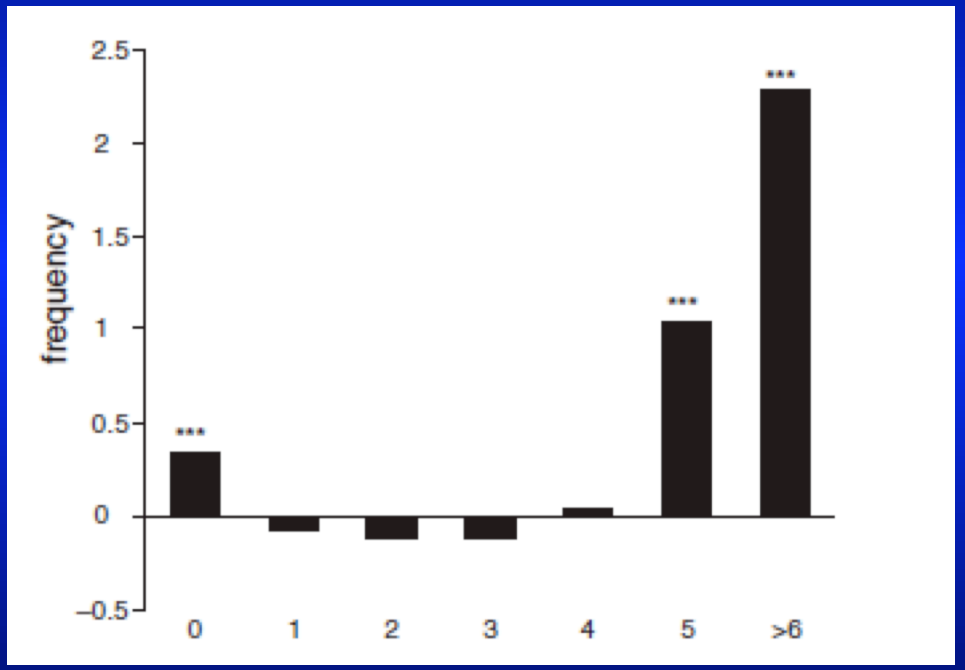
1.Evidence for clustering

2.Evidence that this is independent of selection for co-expression

3.Model: open chromatin favours "expression when needed" and hence clustering of essential genes.

I. Evidence that essential genes cluster in the yeast genome





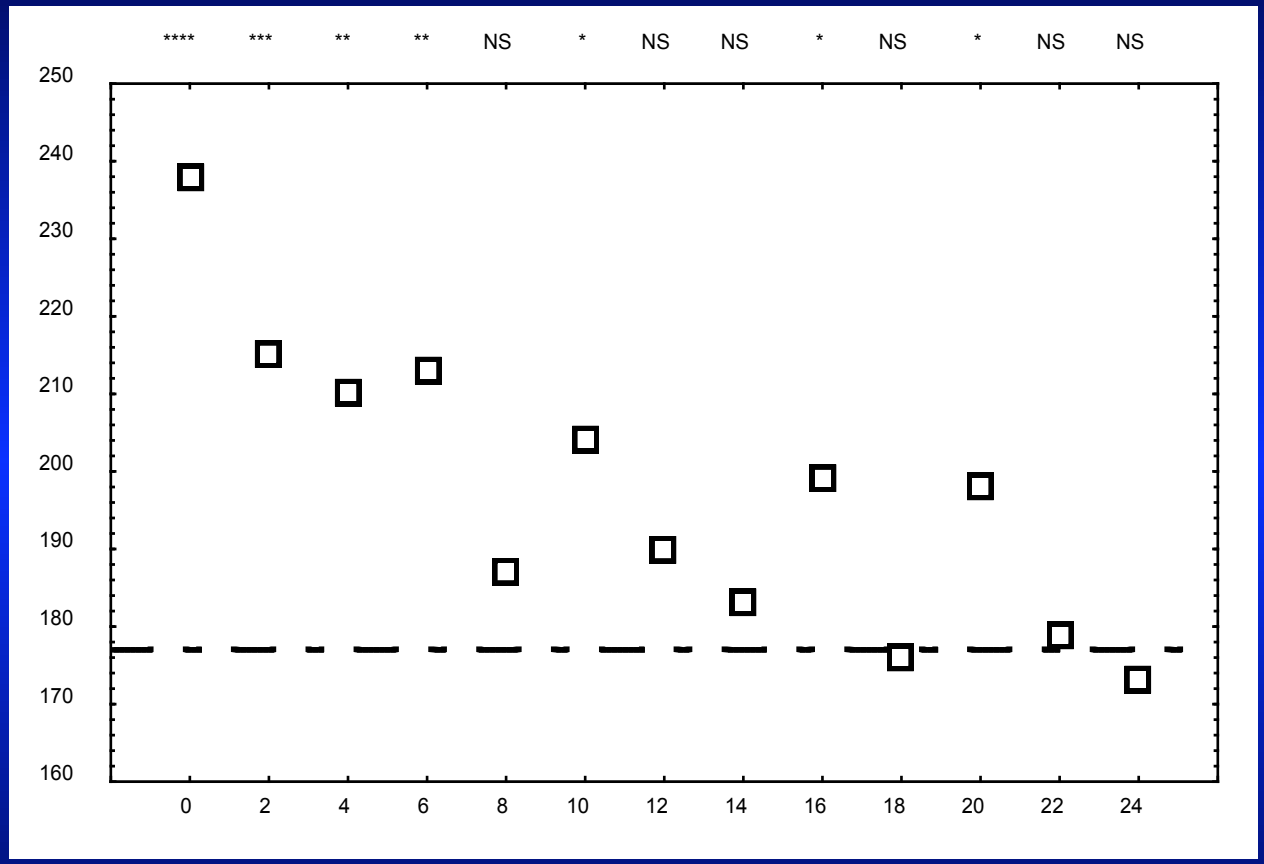
Number of essential genes in blocks of 10 genes

II. Evidence that this clustering is evolutionary constrained: essential gene pairs are preferentially conserved from *S. cerevisiae* to *Candida albicans*

	Observed	**Exp.**
Number of essential, conserved adjacent gene pairs	30	17
Number of non-essential, conserved adjacent gene pairs	163	176

Chi square test: $P<0.002$ NB. Intergene distance for essential genes is the same as that for non-essential genes

III Evidence that clustering of essential genes is not a result of selection for co-expression

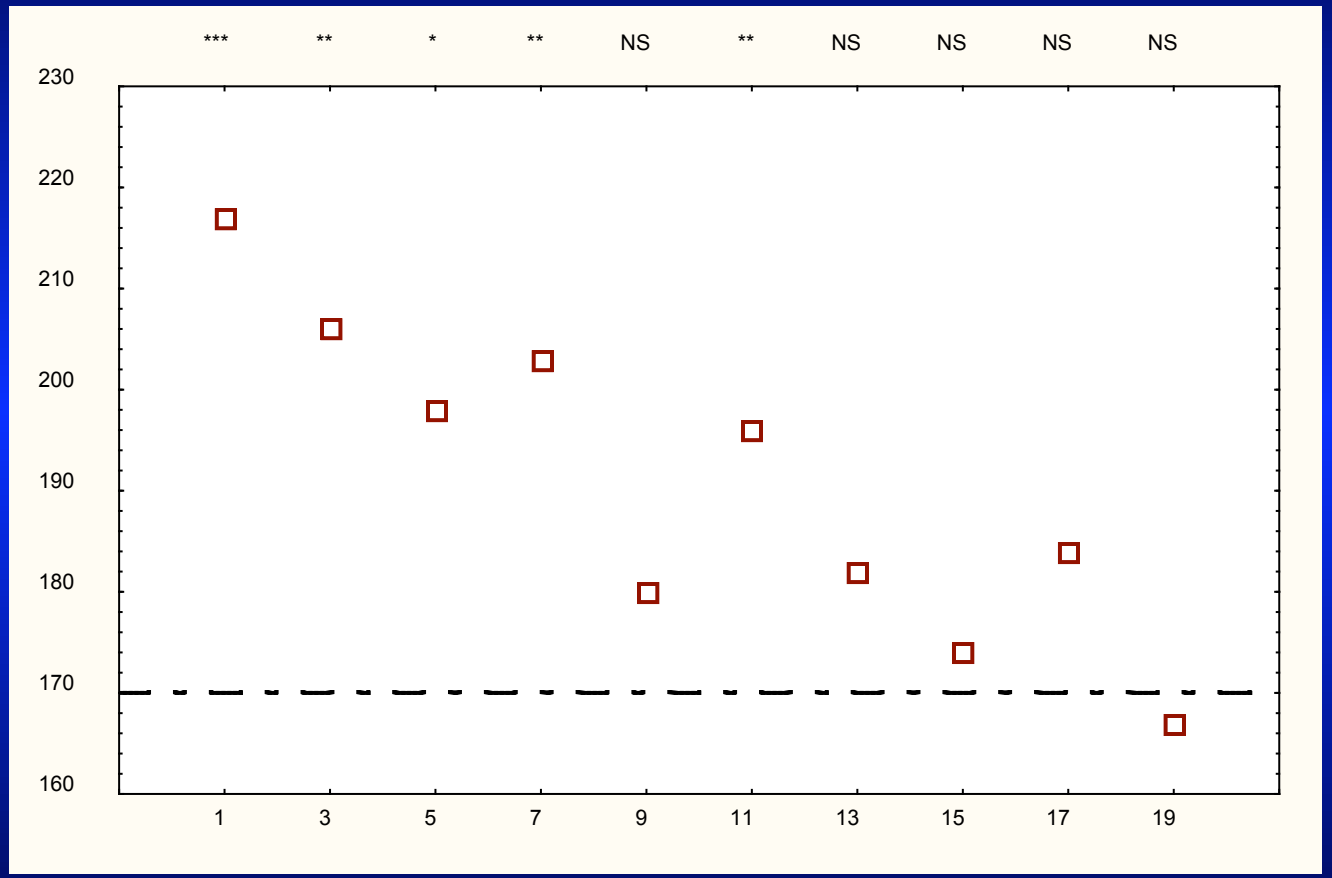


Number of incidences

Number of genes between a focal essential gene and downstream essential genes

---------- expected number

After controlling for co-expression and tandem duplicates clustering of essential genes remain



Number of incidences

Number of genes between a focal essential gene and downstream essential genes

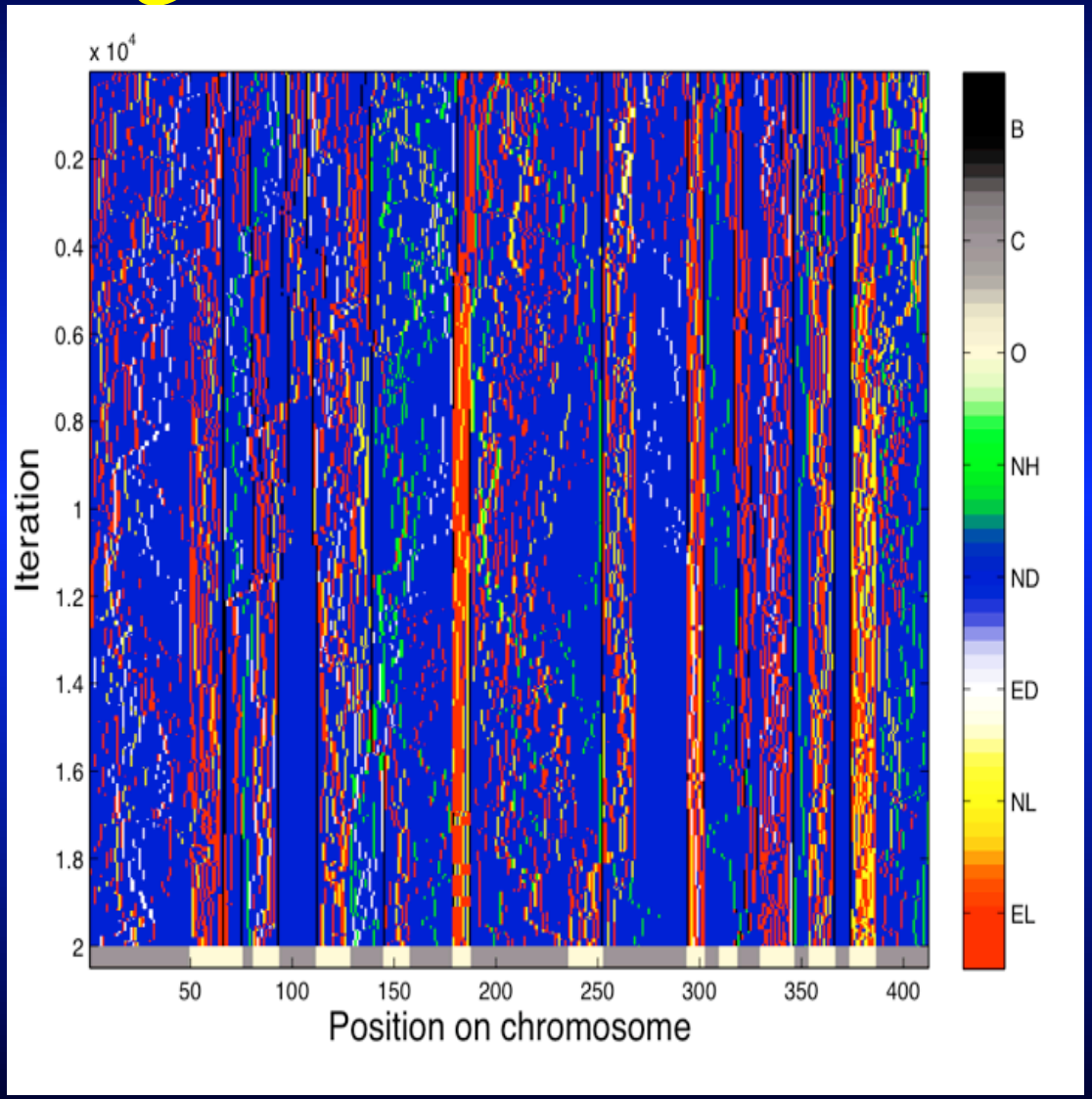
Hypothesis: reduction in gene expression noise as a driver of clustering of essential genes

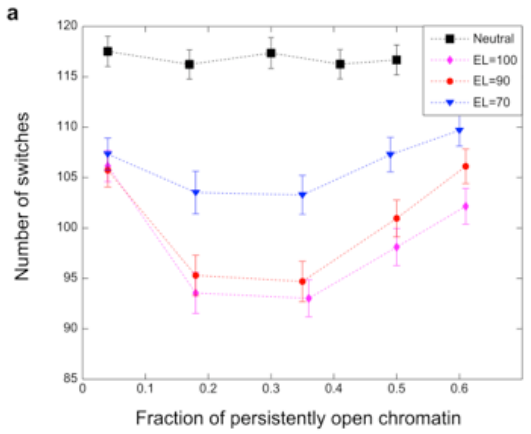
1. Noisy expression can mean a protein is present when not wanted or, potentially worse, absent when needed.
2. By definition, for essential genes the latter can be lethal (if dosage drops to zero).
3. Essential genes should be under selection to have low noise.
4. Among non-essential genes those with lower knockout fitness should have lower noise.
   - Abundance corrected noise versus KO fitness $r = 0.17$, $P < 10^{-10}$, Spearman
5. A major source of noise is random opening and closing of chromatin associated with transcriptional bursting.
6. Constantly open chromatin should a) be a low noise domain and b) be a sink for essential genes.

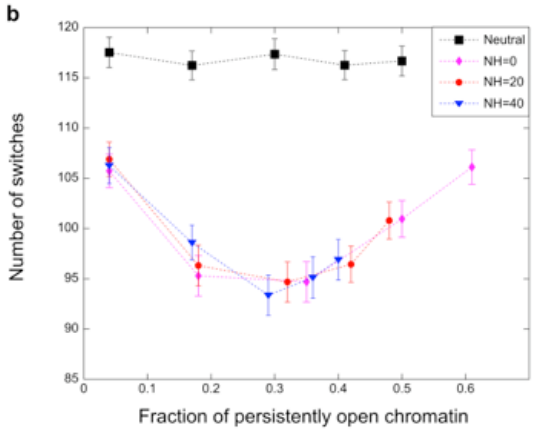
Simulating the model

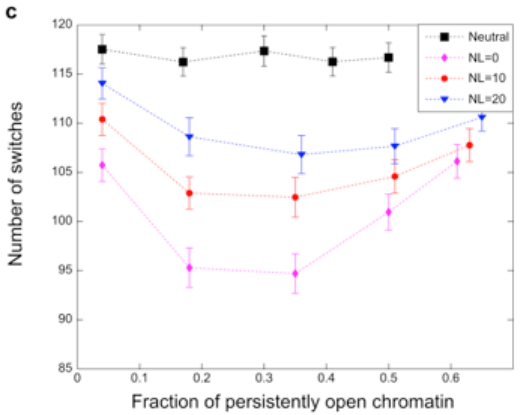
Assume:

- Boundary elements define domains of persistently open chromatin associated with low noise
- 18% of genes are essential
- Some proportion of essential genes like low noise, others are noise neutral
- Some proportion of non-essential gene like low noise, some like high noise, remainder are noise neutral.
- Re-arrangements occur via inversions which are not permitted to cut within genes
- Sub-telomeres do not have persistently open chromatin
- Weak selection









Clustering occurs and is maximal for intermediary frequency of the genome in persistently open chromatin

Allowing non-essentials preferring high noise doesn't aid clustering

Allowing non-essentials preferring low noise diminishes clustering of essentials, but the low noise non-essentials cluster with the essentials

Predictions 1. All genes of similar dispensability should mutually aggregate in the genome

1. **Non-essential genes favoured to have low noise (i.e. low KO fitness non-essentials) should cluster with the essential genes.** Hence there should a correlation between number of neighbouring essential genes around a given non-essential gene and the fitness of the non-essential gene on KO.
   - From simulation $r = -0.15$, $P = 6 \times 10^{-3}$.
   - From real data: $r \sim -0.1$, $P < 0.0001$.
2. **More generally, this model predicts a correlation between neighbouring genes in their fitness effects:**
   - simulation data: correlation between fitness of adjacent genes: $r = 0.18$, $P < 3 \times 10^{-4}$;
   - real data:
     - all genes: $r = 0.2$, $P < 10^{-47}$,
     - ignoring essential genes: $r = 0.19$, $P < 10^{-29}$.

Predictions 2. Sub-telomeres should be high noise domains with few essentials:

Essential genes are rare subtelomerically:

2% of genes in subtelomeric domains are essential versus 18% on average, (P<<0.0001).

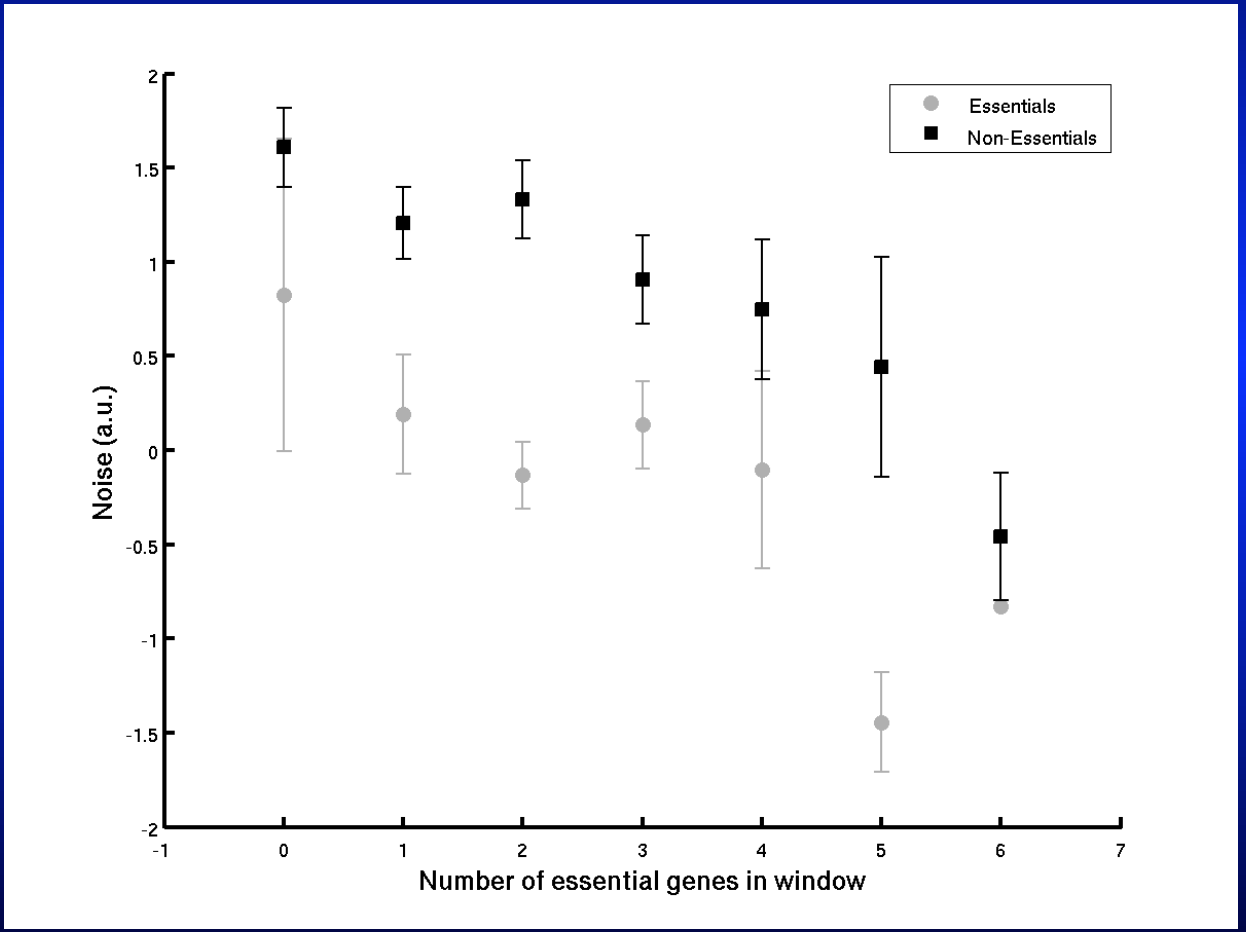
Noise is 30 fold higher subtelomerically:

Median noise of non-essentials in non-telomeric region = 0.14;

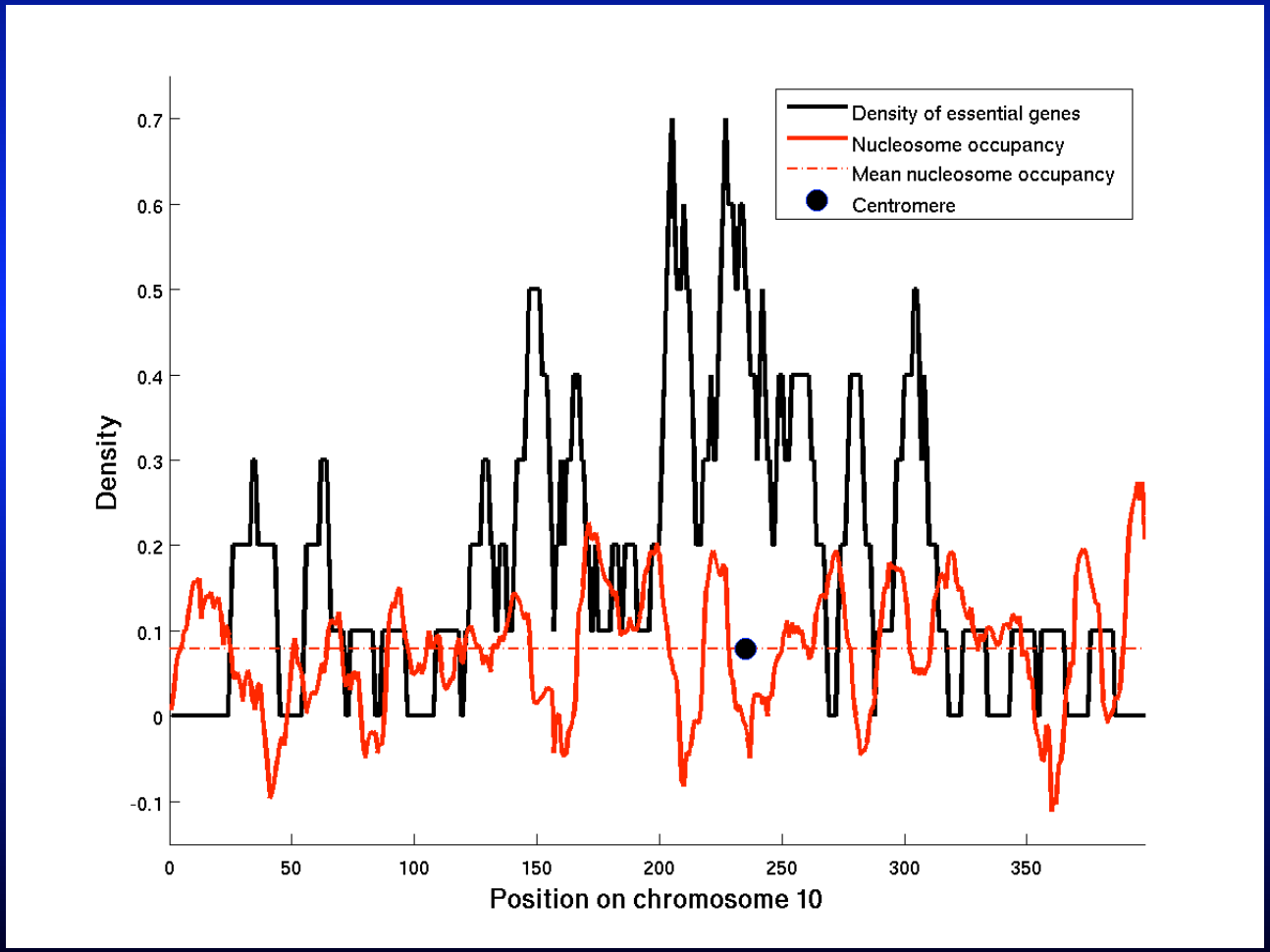
median noise of non-essentials in subtelomeric = 4.2;

P = 3.6 x 10-4, Mann-Whitney U test.

Predictions 3. Essential gene clusters should be low noise domains:



Predictions 4: Essential gene clusters should be in open chromatin (low nucleosome occupancy*)



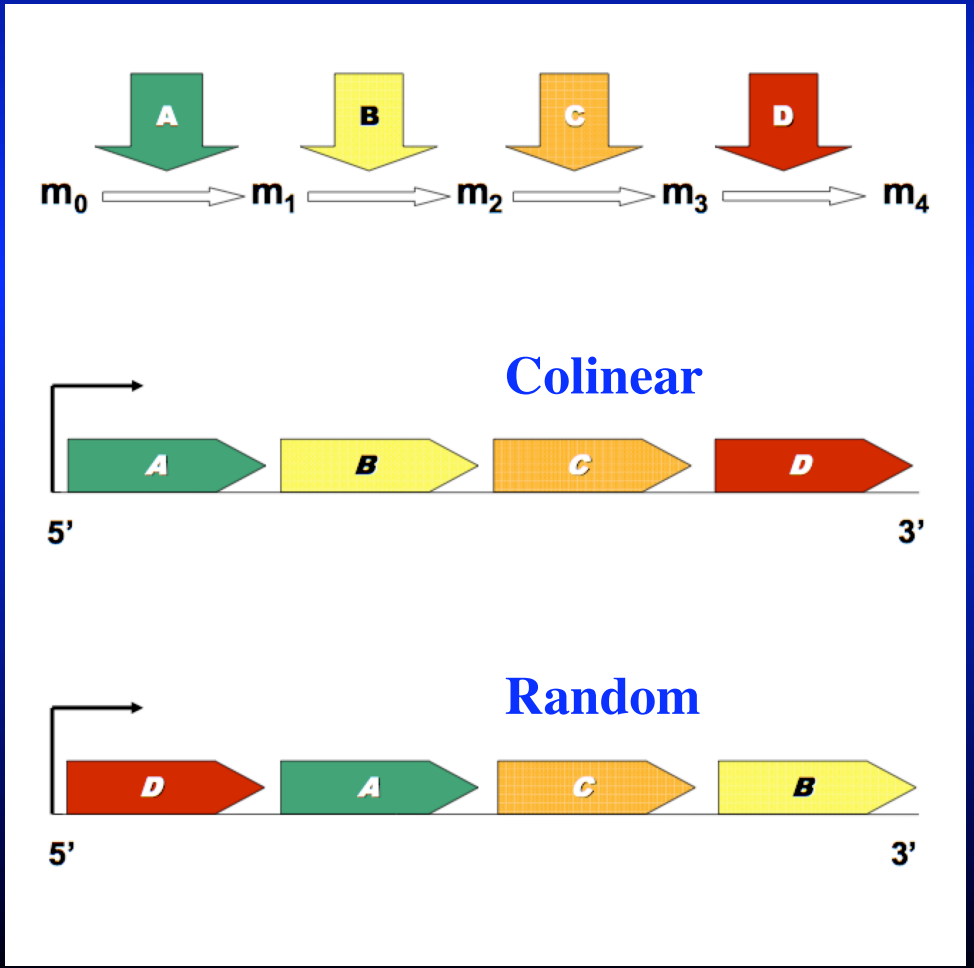
- Domains in which more than half of the genes are essential have occupancy rates half that of domains lacking essential genes ($P = 0.00025$).

- This is reflected in an anticorrelation between nucleosome occupancy and localized density of essential genes across yeast chromosomes ($< P < 0.0002$).

*Myc-tagged histone H4

Case history II: why might some metabolic operons in E. coli be colinear?

Colinearity in metabolic operons is a correspondence between order in a metabolic pathway and order within the operon.



As all genes in the same operon are expressed during the same transcriptional event, selection for co-expression is unlikely to explain any colinearity.

Q1. Are operons co-linear more than expected?

Q2. If so, why?

Colinearity is more common than expected:

- 70 operons
- 321 gene pairs

- 60% of gene pairs are colinear, more than expected by chance (50%, P=0.001).

This is paradoxical as at steady-state all enzymes derived from the same operon should be equally abundant

To show this we employ a standard Michaelis-Menten rate law applied to a pathway with 4 enzymes:

The metabolite concentrations are expressed as follows for the first three products (i = 1...3):

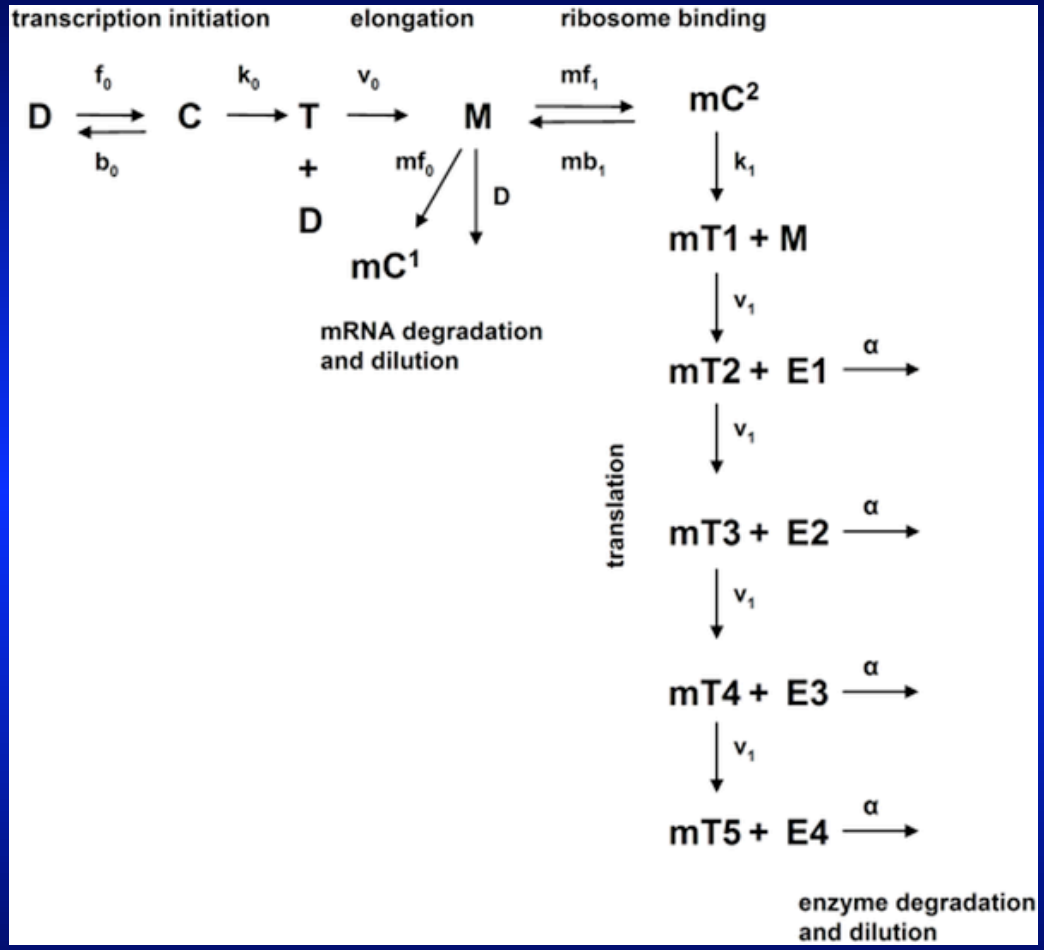
$$\frac{dS_i}{dt} = k_{cat} \cdot E_i \cdot \frac{S_{i-1}}{S_{i-1} + K_m} - k_{cat} \cdot E_{i+1} \cdot \frac{S_i}{S_i + K_m} - D \cdot S_i$$

Metabolic pathway productivity was defined as the amount of end product synthesized during a given time period after operon induction:

$$\frac{dS_4}{dt} = k_{cat} \cdot E_4 \cdot \frac{S_3}{S_3 + K_m}$$

This in turn was coupled to a model of operon operation:

Operon expression was modelled following the "read-through" operon model of Swain, in which ribosomes move directly from one gene to the next, hence translation events are completely correlated across intraoperonic genes. The rate of translation was fine tuned to achieve a delay between the appearances of consecutive gene products (Ei) that reflects empirically observed values, i.e., 60 s



As expected gene order makes no difference to rate of formation of the end-product

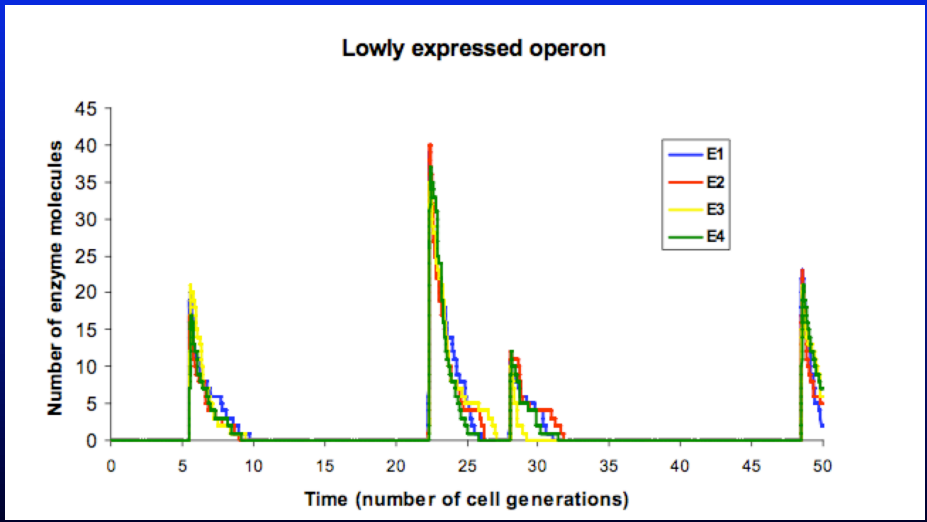
Gene order	**Flux through E_4 (mmol * s^{-1})**
ABCD	$1.64033624*10^{-16}$
DCBA	$1.64033624*10^{-16}$

Why then do we see colinerity?

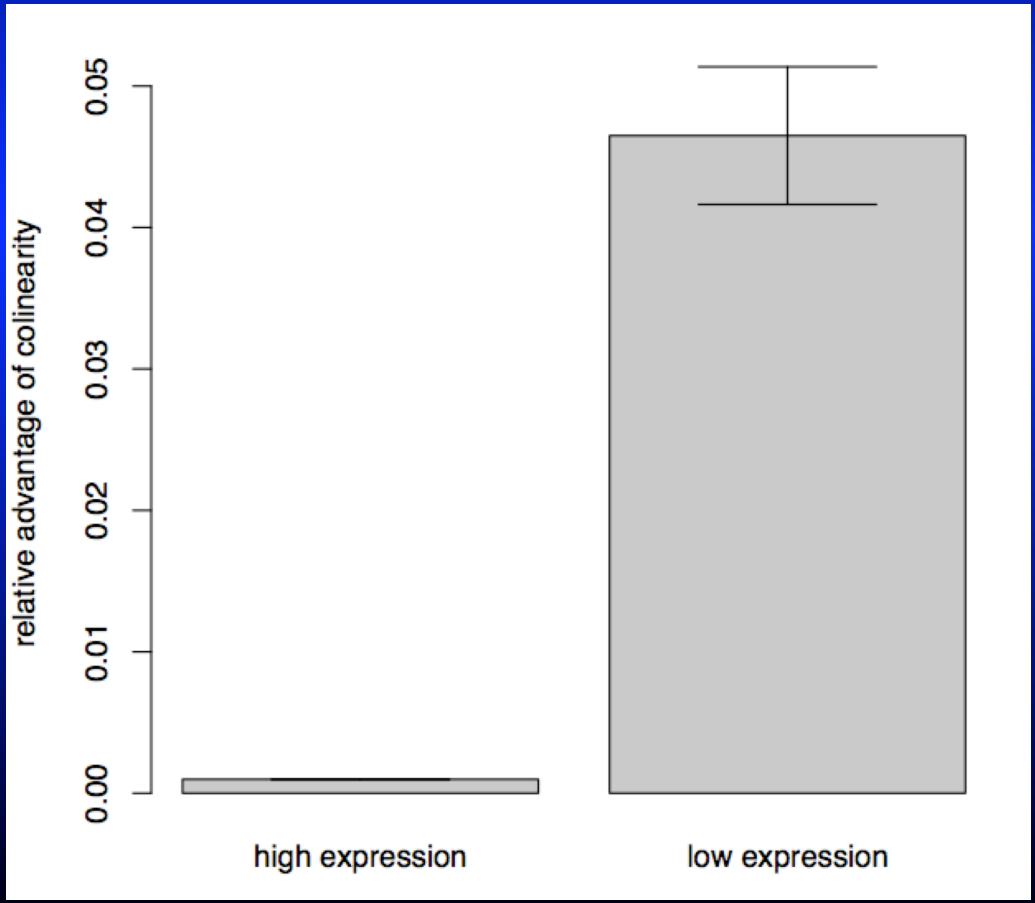
We looked at three hypotheses. To cut a long story short, only one had any predictive power…

This was that hypothesis that stochastic protein loss in lowly expressed operons selects for colinearity as a means to efficiently "reboot"stalled metabolism.

- To examine this we modified our deterministic model of expression from operons to produce a stochastic simulation.
- This suggests that if one protein from a rarely expressed operon is lost all are lost.

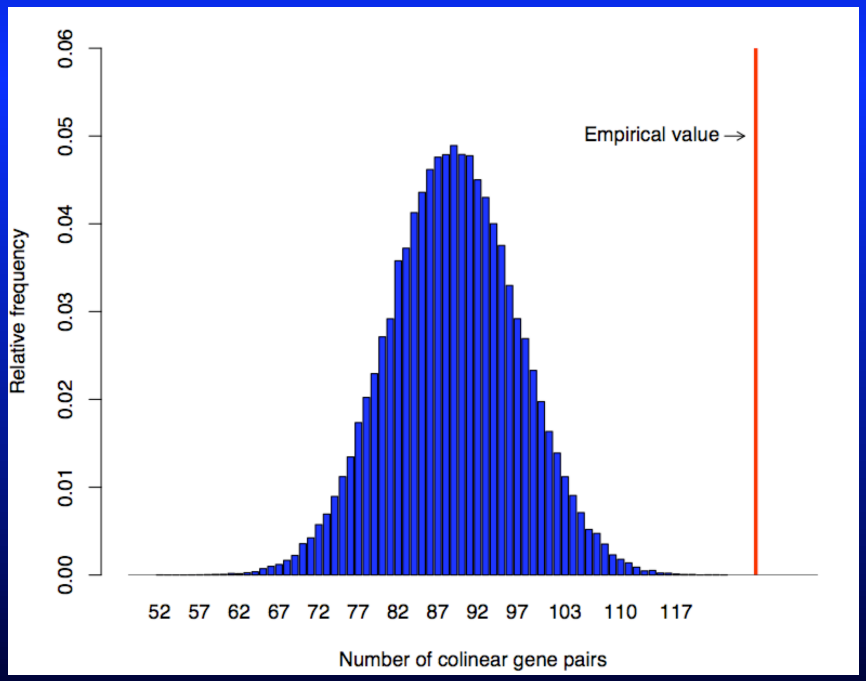


To restart metabolism it is fastest to translate A before B, rather than vice versa

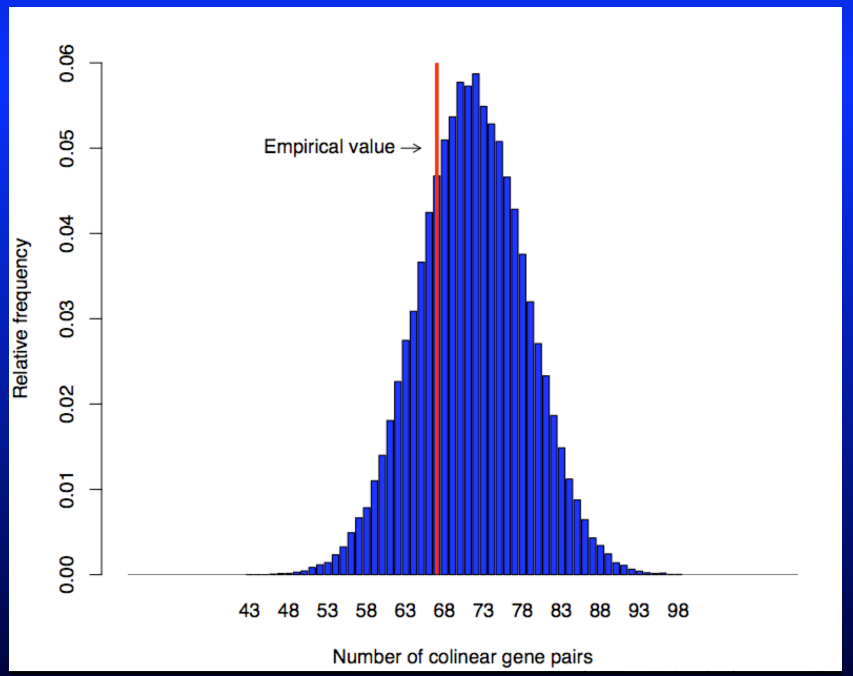


This model makes an unusual prediction: as no advantage to colinearity is seen for highly expressed operons this model proposes colinearity to be unique to lowly expressed operons. This is observed.

Lowly expressed (P<0.0006)
72% colinear



Highly expressed (ns)
45% colinear



The need for proteins to be present when needed can explain:

1. Clustering of essential genes into domains of open chromatin

2. Colinearity of metabolic genes in lowly expressed operons

Conclusions

The dichotomy between a noisy genome versus an ordered genome is spurious: genomes are ordered to minimize the impact of inevitable noise

PLENARY LECTURE SPONSOR

European Molecular Biology Organization

Sponsoring ● Mentoring ● Networking ● Training

- → PhD & Post-doc students
- → Young Investigators
- → Group Leaders
- → Senior scientists

Promoting excellence in molecular life sciences since 1964

EMBO

www.embo.org