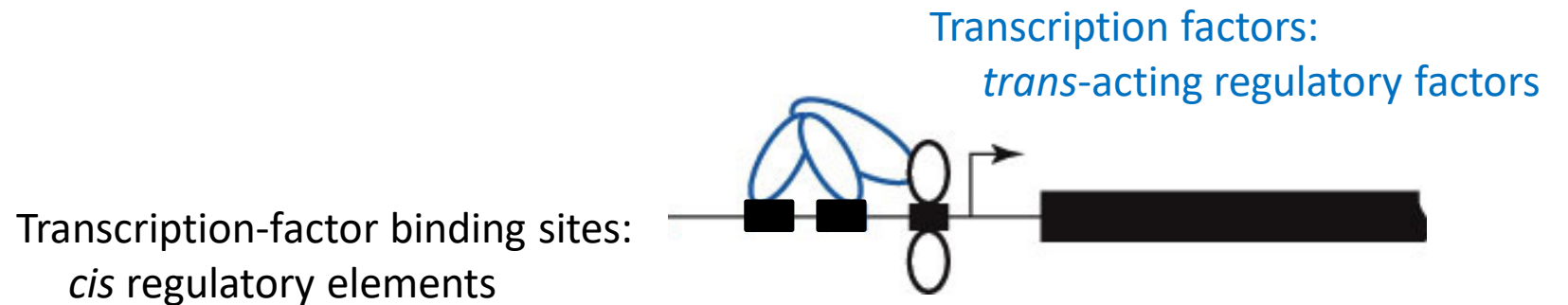
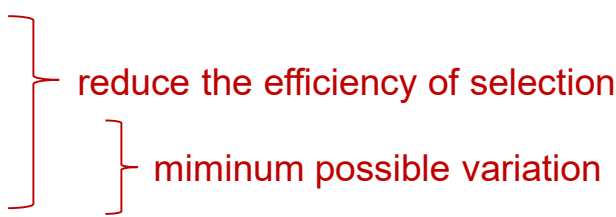


# Transcription and Gene Regulation

- Molecular stochasticity in single cells.
- Transcription factors and their regulatory motifs.
- Biophysics of recognition: facilitated diffusion and the search for regulatory motifs.
- Evolution of the regulatory vocabulary.
- Evolutionary rewiring of transcription networks.



# Variation in Gene Expression in Single Cells

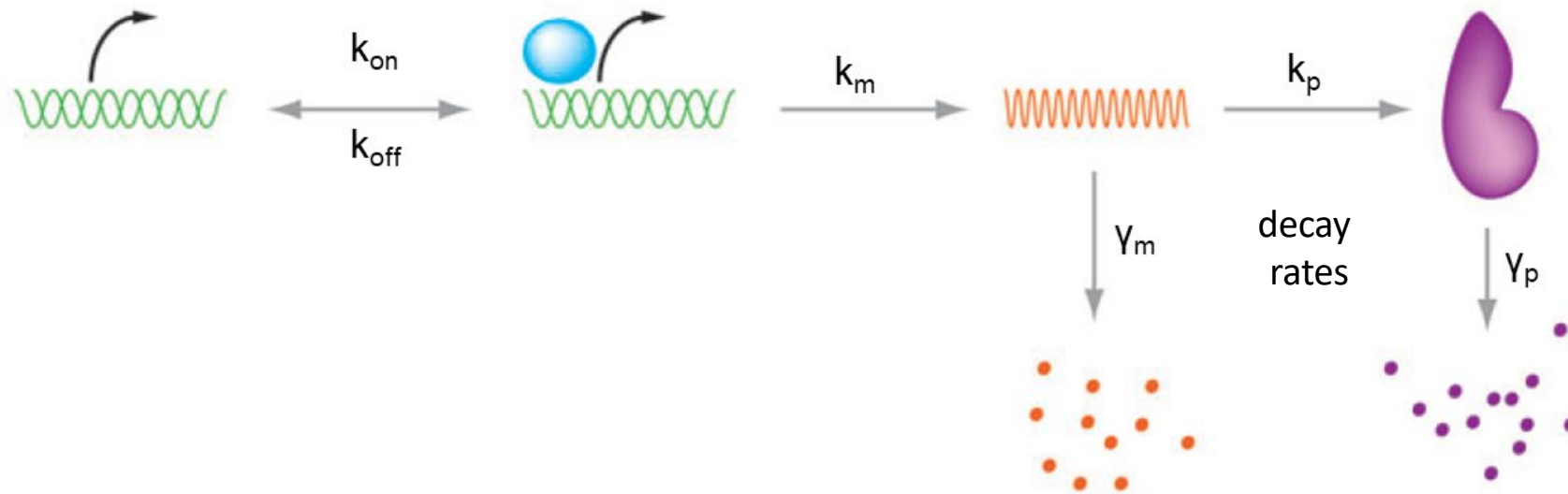
- Typically, gene expression is measured in populations of cells, obscuring the variation that exists at the individual level.
- Because the average numbers of proteins of a particular type is often well below 100 / cell, stochastic cell-to-cell variation can be large.
- Total among-cell variation in gene expression can be subdivided into three components:
  - 1) genetic variation among cells, ← essential for evolutionary change
  - 2) extrinsic environmental variance,
  - 3) intrinsic noise due to the vagaries of random molecular motion and production.

The diagram consists of a large red curly bracket on the right side of the list, spanning items 2 and 3. To its right is the text "reduce the efficiency of selection". A smaller red curly bracket is positioned to the right of item 3, with the text "mimumum possible variation" (sic) to its right.
- Heritability ( $h^2$ ) =  $V_G / (V_G + V_{Ee} + V_{Ei})$ .
- Response to Selection = Heritability x Selection Differential

# Central Kinetic Components of Protein Production

## Exponentially distributed times

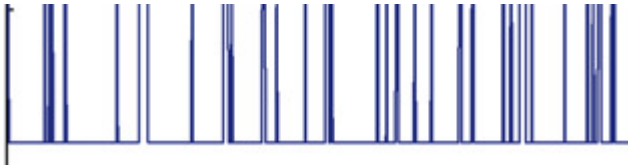
In ON and OFF promoter states  
Between mRNA, protein\* bursts



- Probability gene is on:  $P_{on} = k_{on} / (k_{on} + k_{off})$

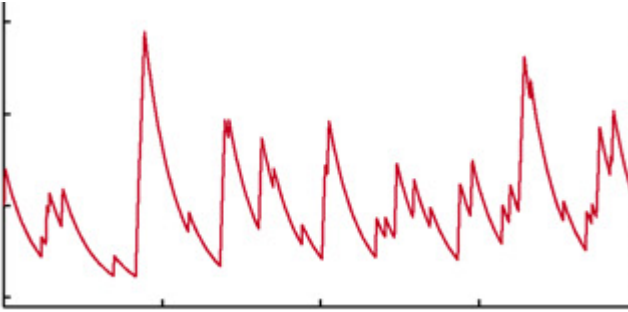
# Temporal Variation in Gene Expression Within a Cell

Promoter on



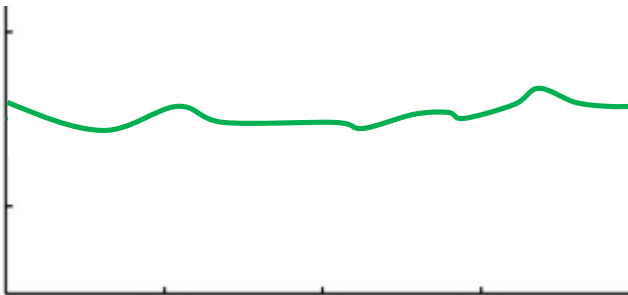
- Gene is stochastically on / off depending on the binding of cognate transcription factors.

mRNA copies



- Messenger RNAs are produced during on periods, but decay away at exponential rates during off periods.

Protein copies



- Protein numbers rise during periods of mRNA abundance, and decline slowly via during periods of mRNA rarity. Fluctuations in protein numbers are damped, owing to their greater longevities than mRNA molecules.

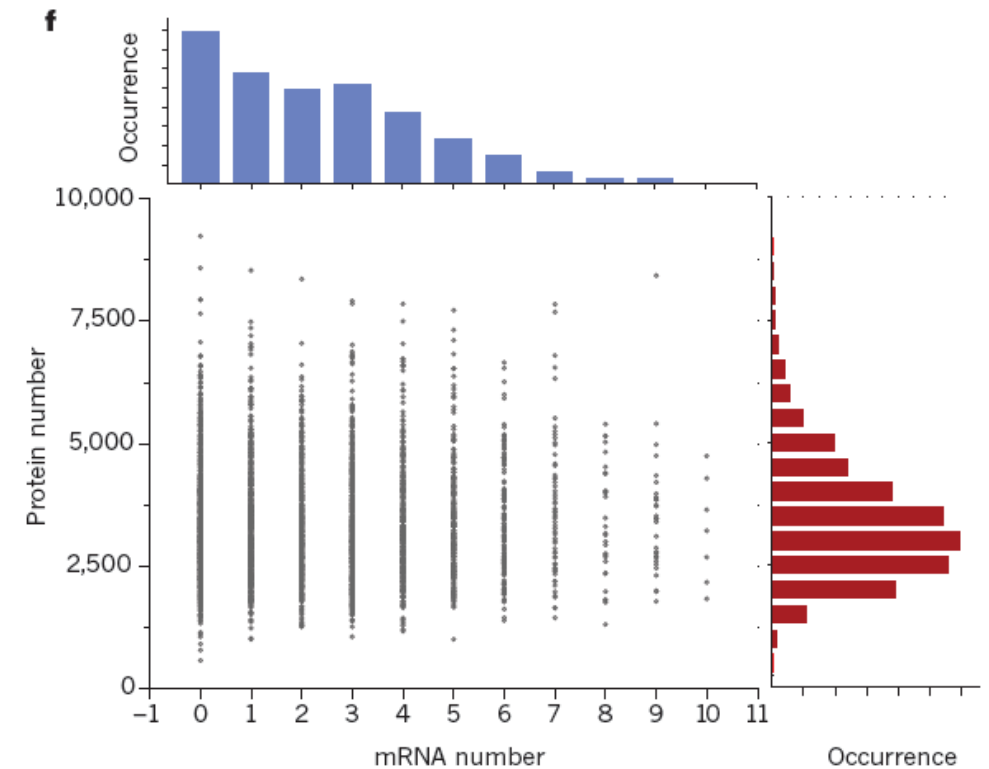
Time

## Because mRNA Degradation Rates Are High Relative to Transcription Rates, the Average Number of mRNAs / Gene is Small

### Median half lives of mRNAs:

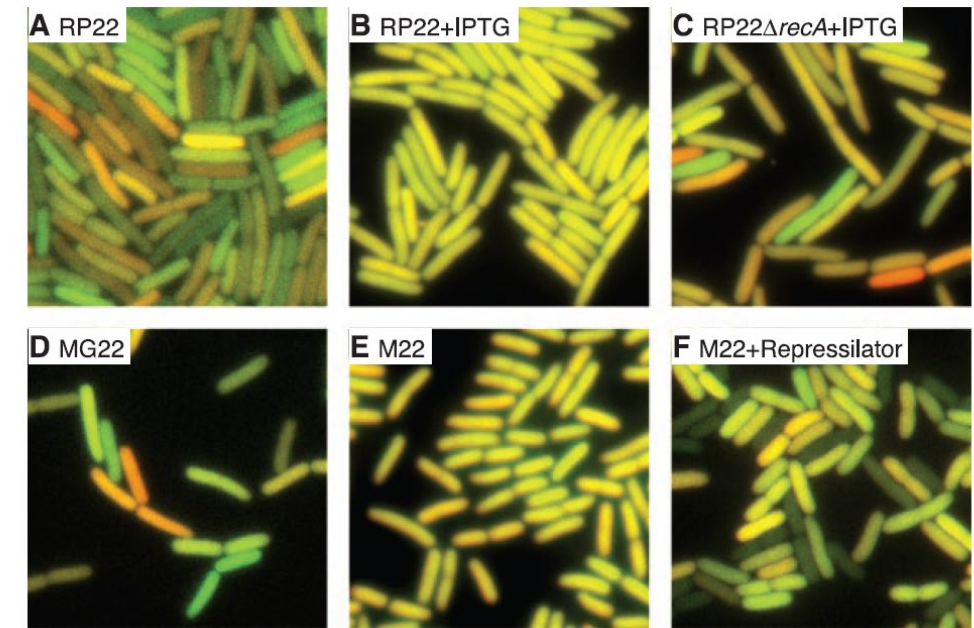
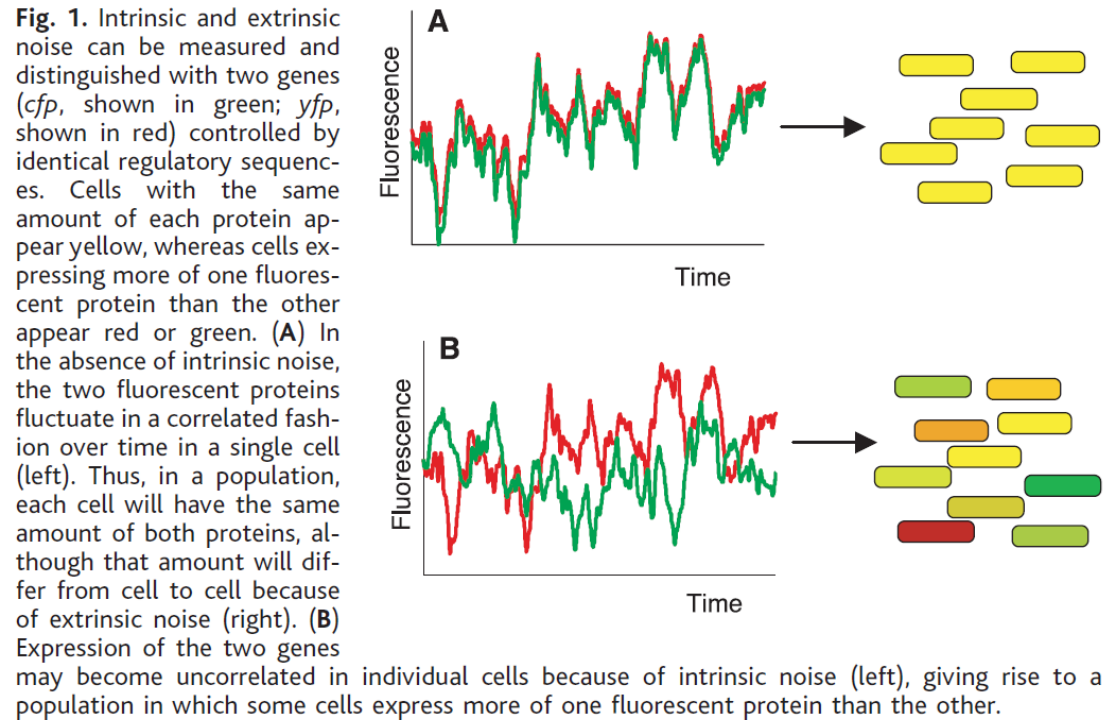
<i>E. coli</i>	5 minutes	(Taniguchi et al. 2010)
<i>S. cerevisiae</i>	22 minutes	(Wang et al. 2002)
Mouse fibroblast	9 hours	(Schwanhaussner et al. 2011)

- There is no correlation between the number of mRNAs and the number of protein molecules within individual cells.
- Proteins numbers are much higher than mRNA numbers.
- Raises issue about single-cell transcriptomics studies.



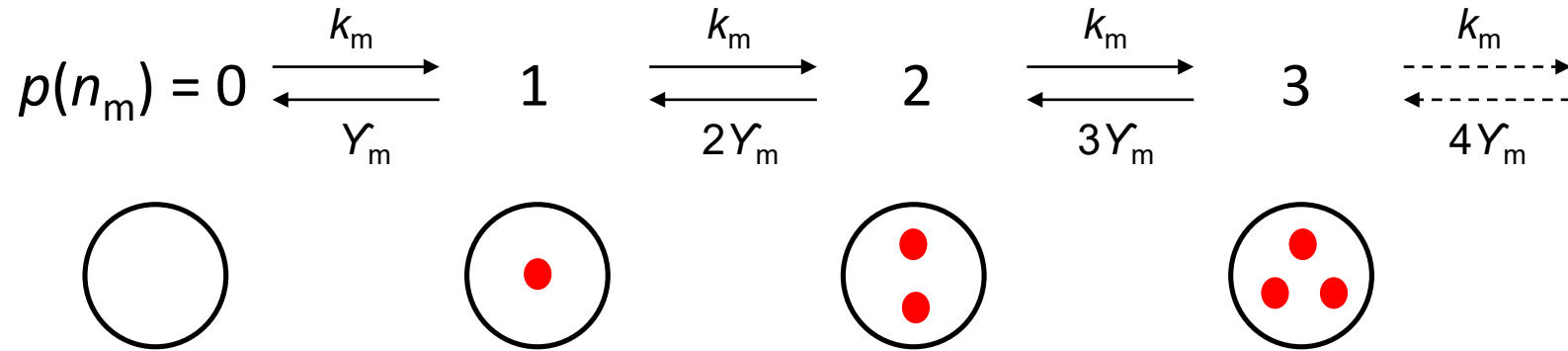
# Stochastic Gene Expression in Single Cells

- Two identical genes in the same bacterial cells, with the proteins labeled by different fluorescent markers, will have the same expression level only if there is no intrinsic noise within the cell.



**Fig. 2.** Noise in *E. coli*. CFP and YFP fluorescence images were combined in the green and red channels, respectively. (A) In strain RP22, with promoters repressed by the wild-type *lacI* gene, red and green indicate significant amounts of intrinsic noise. (B) RP22 grown in the presence of lac inducer, 2 mM IPTG. Both fluorescent proteins are expressed at higher levels and the cells exhibit less noise. (C) As in (B), except the *recA* gene has been deleted, increasing intrinsic noise. (D) Another wild-type strain, MG22, shows noise characteristics similar to those of RP22. (E) Expression levels and noise in unrepresed *lacI*<sup>-</sup> strain M22 are similar to those in *lacI*<sup>+</sup> strains induced with IPTG (B). (F) M22 cells regulated by the Repressator (16), an oscillatory network that amplifies intrinsic noise.

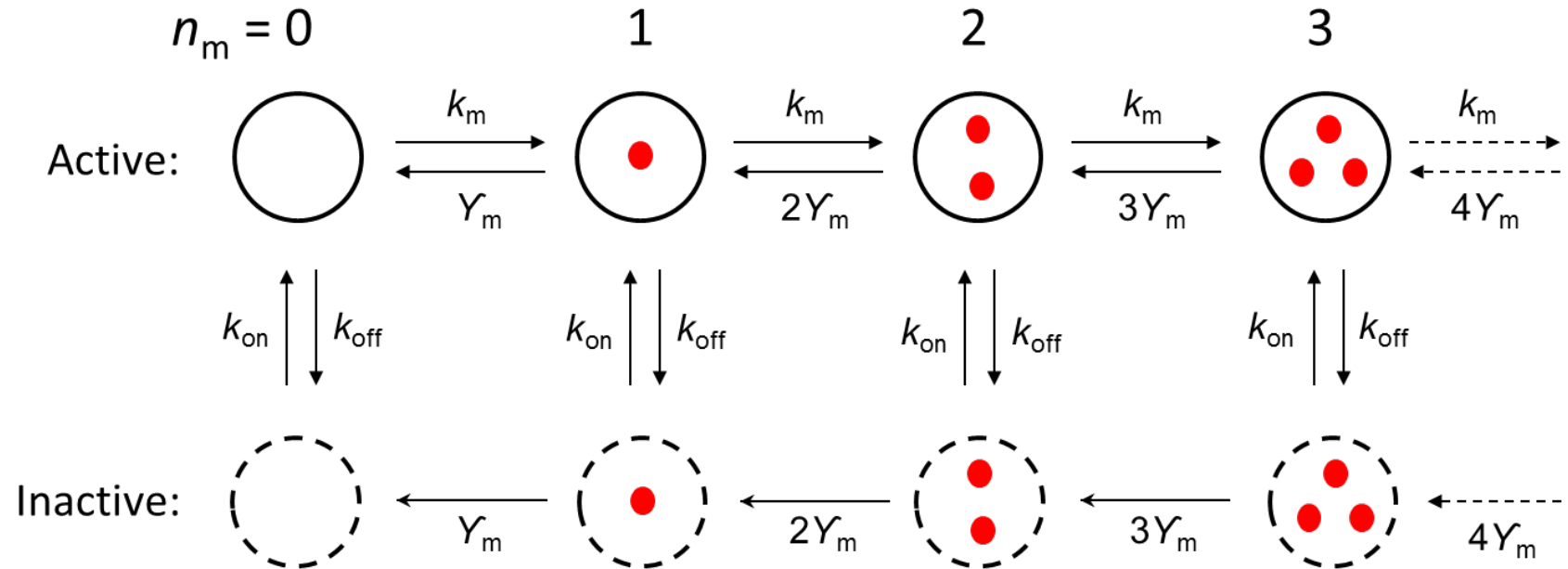
## The Number of mRNAs / Cell is Expected to be Poisson Distributed for a Constitutively Expressed Gene



Mean number of mRNAs ( $n_m$ ) / cell =  $k_m / \gamma_m$

Mean = Variance

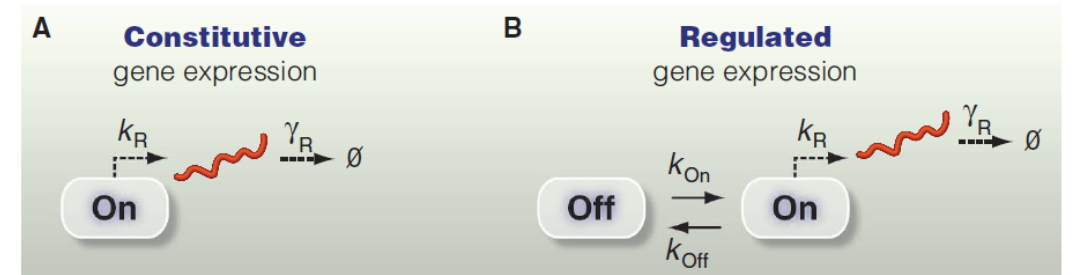
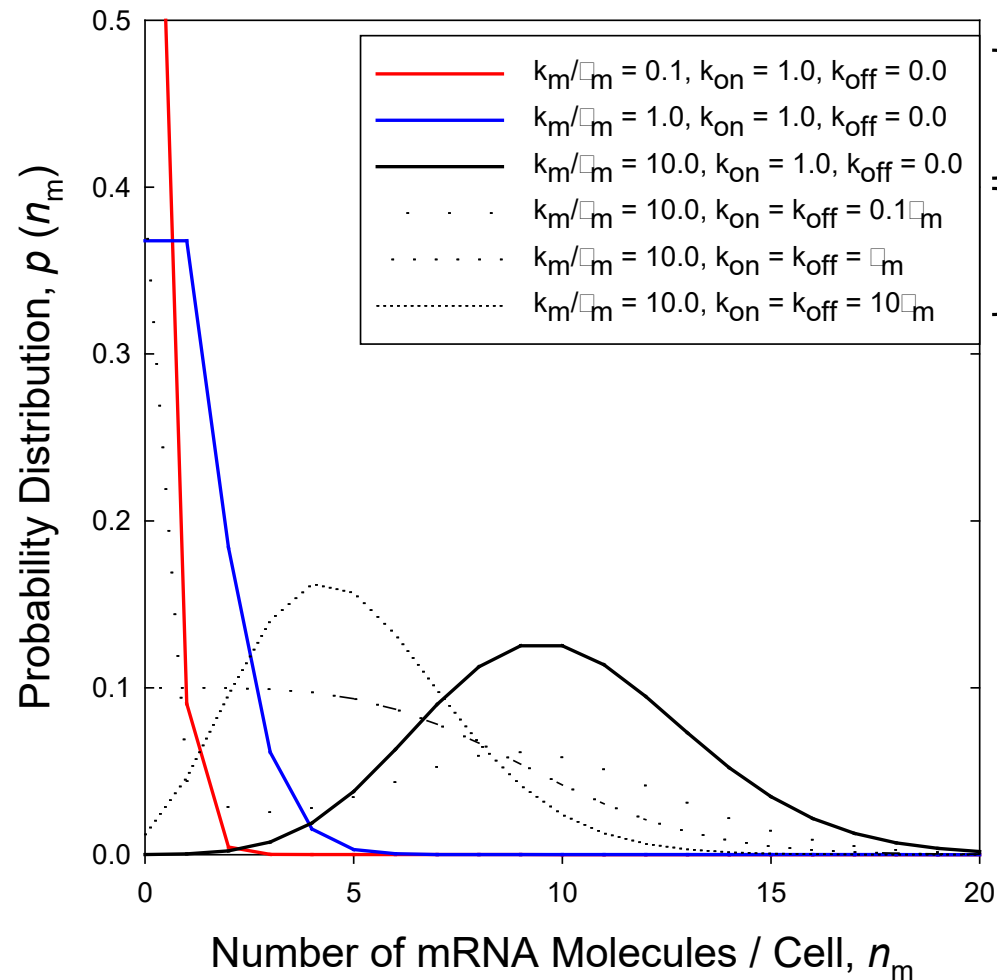
If the Gene is Regulated and Only Active for a Fraction of Time  $P_{\text{on}}$ , Number of mRNA Molecules / Cell Follows a Mixture of Distributions



$$\text{Mean number of mRNAs } (n_m) / \text{cell} = P_{\text{on}} k_m / Y_m$$



Distributions of Transcript Numbers per Cell: regulated genes exhibit elevated levels of among-cell variance in expression, and can even be bimodal.

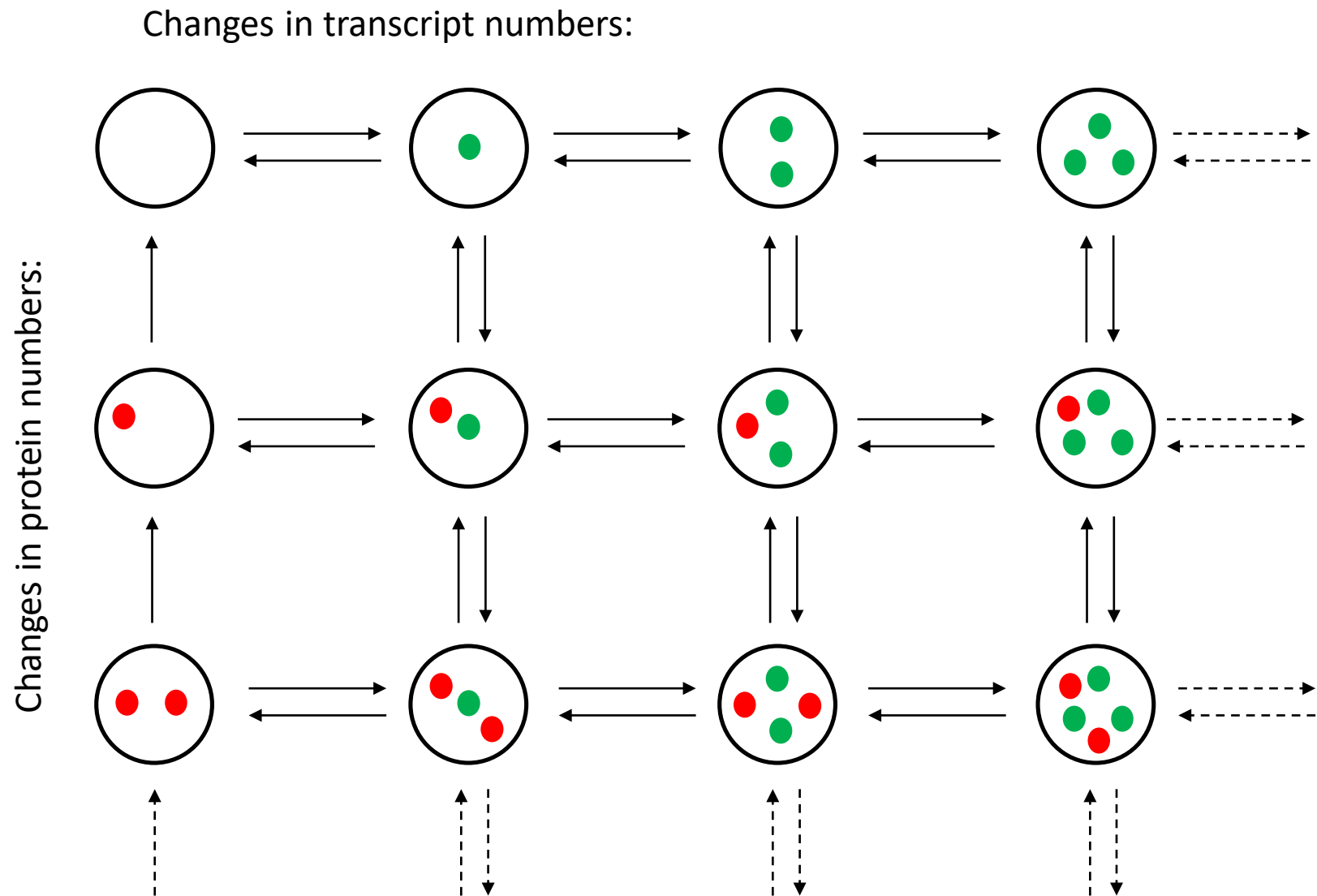


**Fig. 1.** Constitutive versus regulated gene expression. **(A)** Schematic of a constitutive gene expression model with transcription rate  $k_R$  and mRNA degradation rate constant  $\gamma_R$ . **(B)** Schematic of a two-state (On, Off) model with transition rates  $k_{on}$  and  $k_{off}$ .

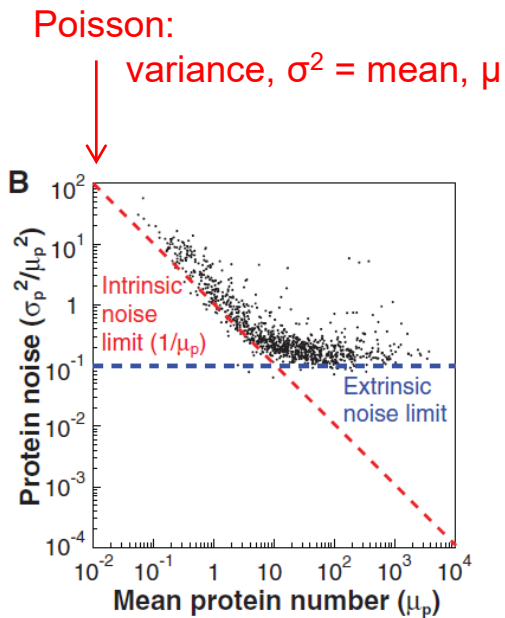
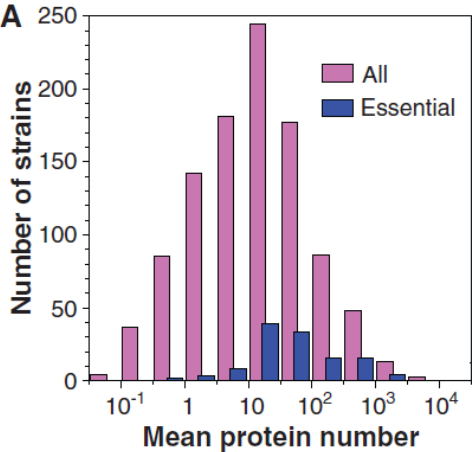
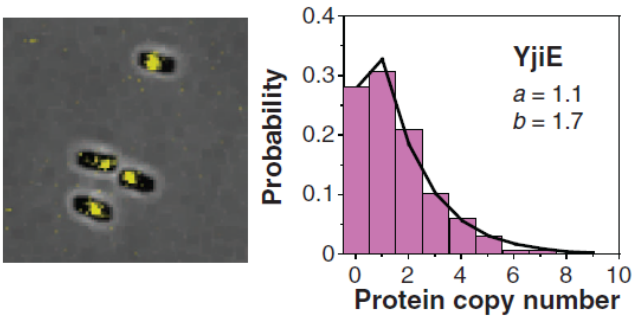
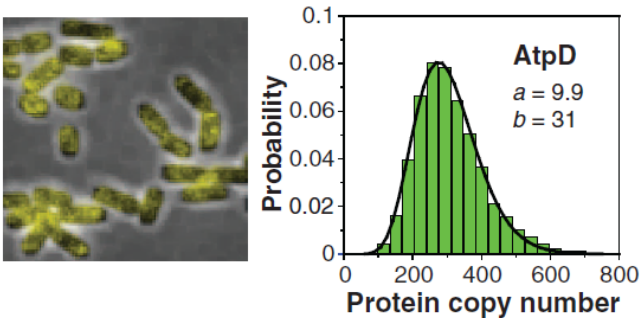
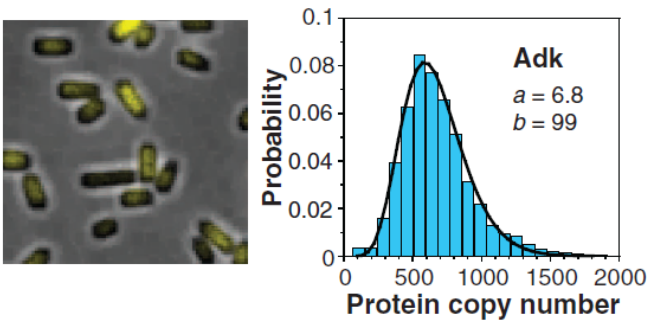
Munsky et al. (Science, 2012)

Extension to Protein Numbers / Cell

● transcript  
● protein



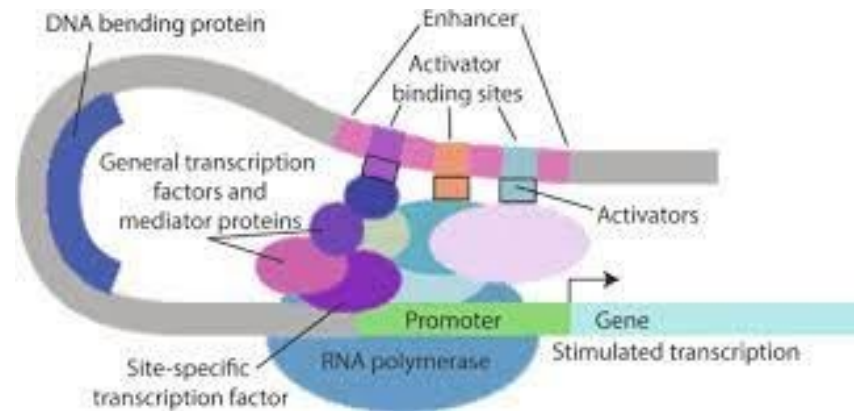
# Average Numbers of Protein Molecules in Individual *E. coli* Cells Are Small



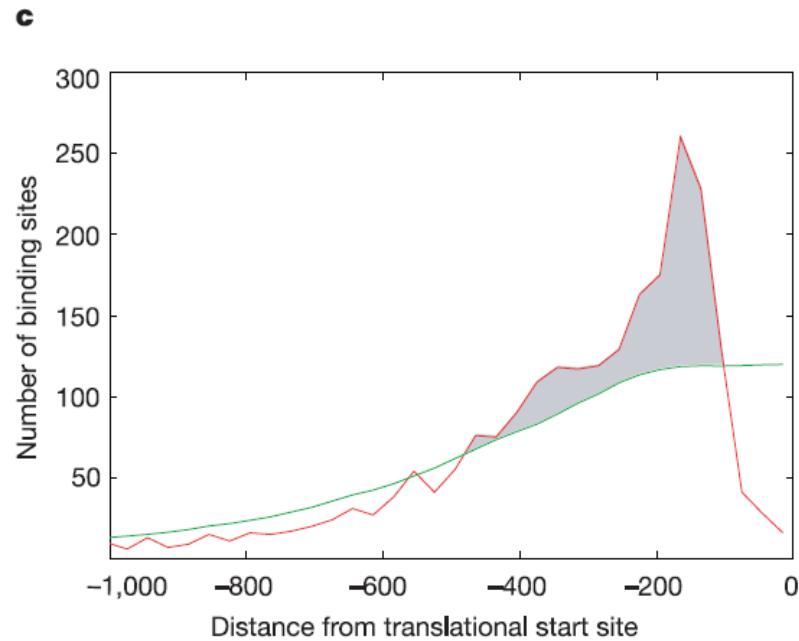
Individual distributions are roughly gamma in form, with a being a measure of burst frequency, and b a measure of burst size.

# Biological Features of Transcription Factors

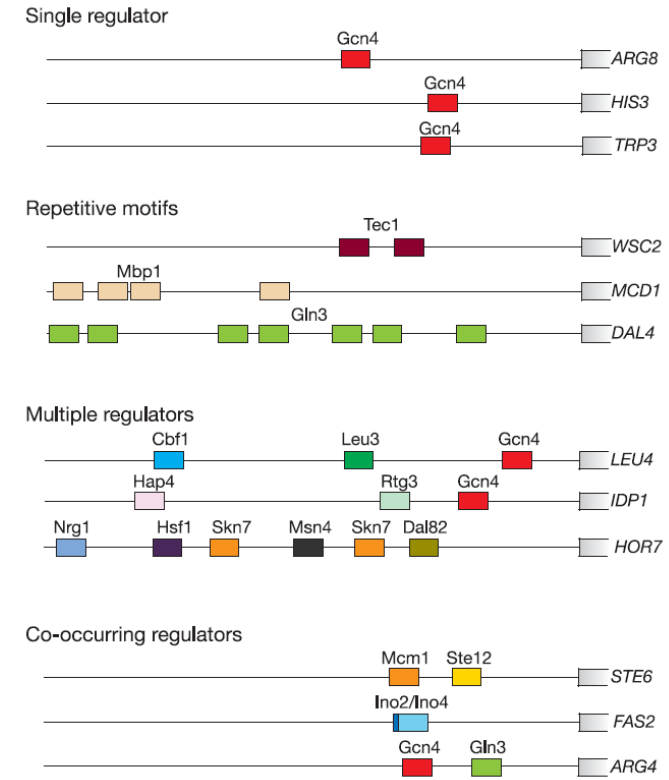
- Transcription is generally nonautonomous, as one to several accessory TFs must be present simultaneously for transcriptional activation.
- In eukaryotes, individual TFs often service multiple genes, which facilitates coregulation of gene expression
- *E. coli* has ~300 transcription factors: 7 control the expression of ~50% of regulated genes, whereas ~60 service single genes.



# Locations of Transcription-factor Binding Sites in Yeast



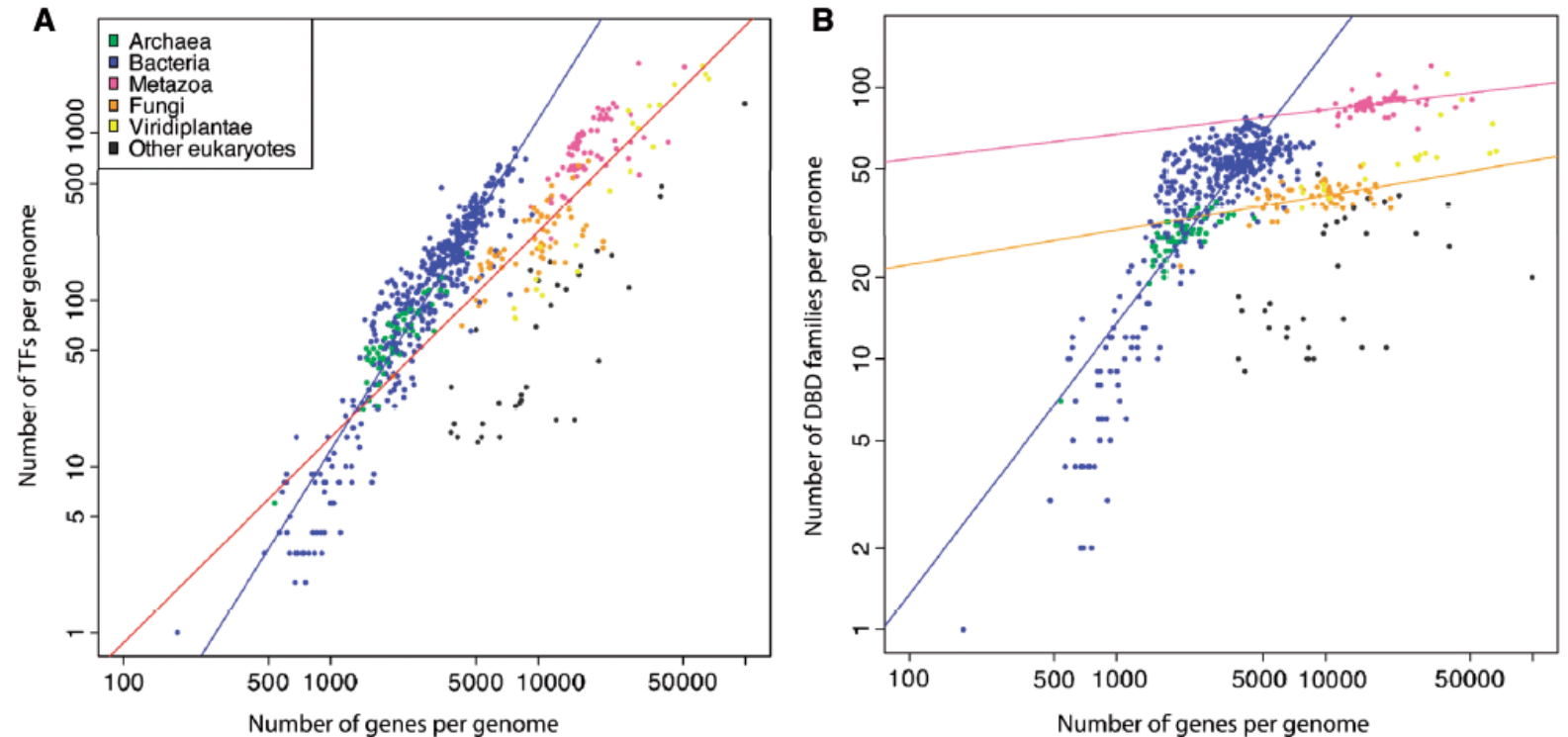
**Figure 2** Drafting the yeast transcriptional regulatory map. **a**, Portions of chromosomes illustrating locations of genes (grey rectangles) and conserved DNA sequences (coloured boxes) bound *in vivo* by transcriptional regulators. **b**, Combining binding data and sequence conservation data. The diagram depicts all sequences matching a motif from our compendium (top), all such conserved sequences (middle) and all such conserved sequences bound by a regulator (bottom). **c**, Regulator binding site distribution. The red line shows the distribution of distances from the start codon of open reading frames to binding sites in the adjacent upstream region. The green line represents a randomized distribution.



**Figure 3** Yeast promoter architectures: single regulator architecture, promoter regions that contain one or more copies of the binding site sequence for a single regulator; repetitive motif architecture, promoter regions that contain multiple copies of a binding site sequence of a regulator; multiple regulator architecture, promoter regions that contain one or more copies of the binding site sequences for more than one regulator; co-occurring regulator architecture, promoters that contain binding site sequences for recurrent pairs of regulators. For the purposes of illustration, not all sites are shown and the scale is approximate. Additional information can be found in Supplementary Tables 4–6.

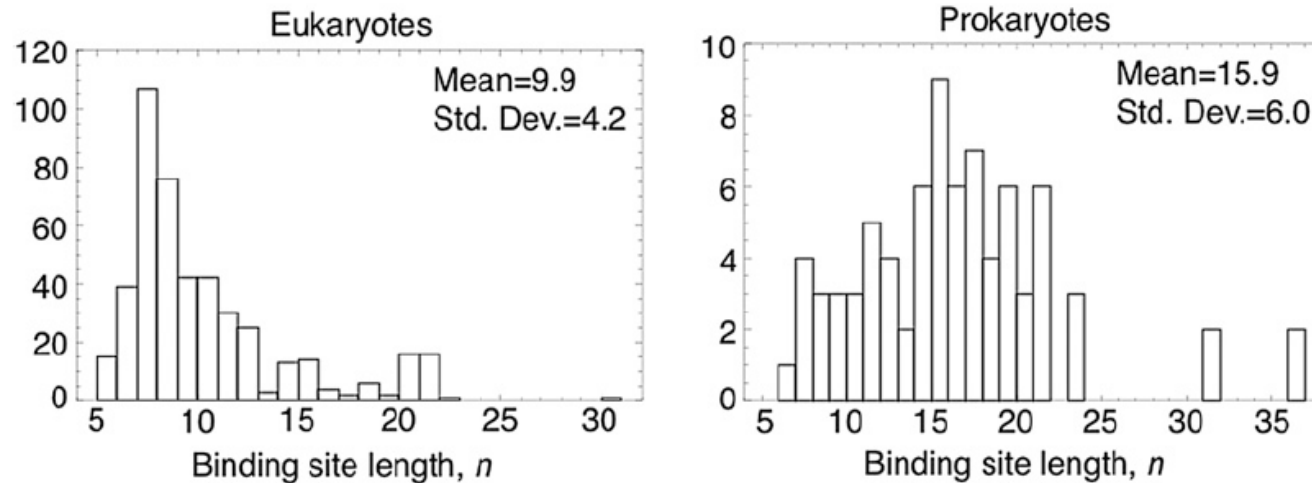
## Scaling of the Number of TFs with the Number of Protein-coding Genes – typically, 1 to 5% of the protein-coding genes within a genome are TFs

- Prokaryotes: number of TFs scales with  $\sim 2$  power of gene number.
- Eukaryotes: number of TFs scales with  $\sim 1$  power of gene number.



**Figure 2.** (A) TF abundance against number of genes per genome in different lineages across the tree of life. Each colored dot represents a genome. Different colors are used to highlight genomes from different phylogenetic groups. According to the linear model fit on a log-log scale, TF expansion in bacteria strictly follows a power law increase, with an exponent close to quadratic ( $\log T = 1.98 \log G - 4.84$  with  $R^2 = 0.87$  where  $T$  is number of predicted TFs,  $G$  is number of genes and  $R^2$  is coefficient of determination). The TF increase in eukaryotes has a lower exponent as well as degree of correlation ( $\log T = 1.23 \log G - 2.53$  with  $R^2 = 0.61$ ). (B) The number of unique DBD families increases linearly with the total number of proteins in bacteria (power law exponent = 1.00,  $R^2 = 0.71$ ). In contrast, the number of families is independent of the number of genes in metazoans (pink, exponent = 0.09,  $R^2 = 0.11$ ) and fungi (orange, exponent = 0.13,  $R^2 = 0.23$ ). Grey dots in the figures represent other eukaryotic species that do not belong to the main kingdoms such as apicomplexan and euglenozoa.

# Transcription-factor Binding Motifs Are Small, and Longer in Prokaryotes Than Eukaryotes



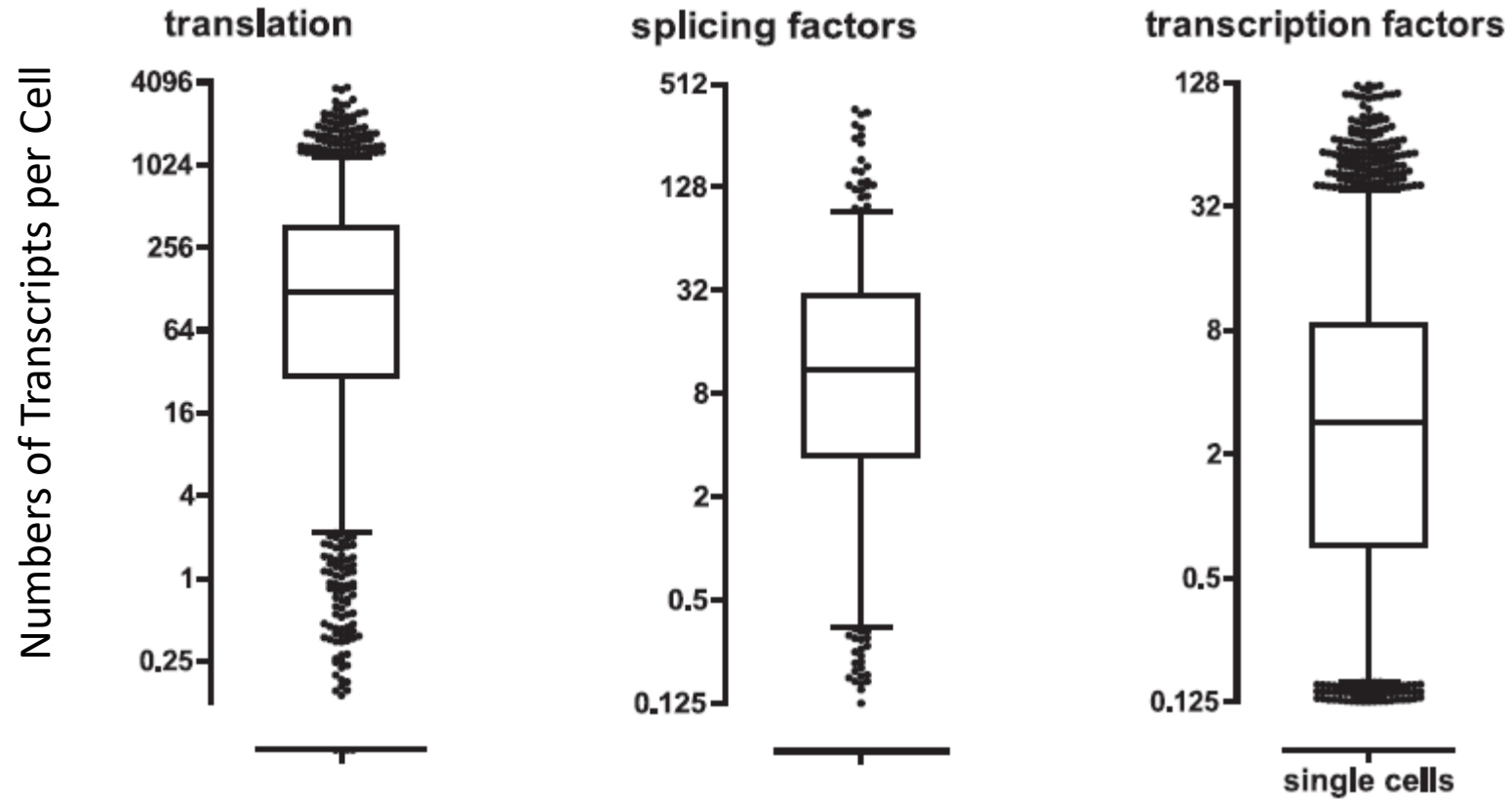
**Figure 1** The lengths of binding sites range from 5 nt to ~30 nt, in both eukaryotes (left, 454 curated transcription factor motifs) and prokaryotes (right, 79 motifs). The information content per nucleotide ranges from ~0.25 bits to 2 bits (see Figure S1).

From: Stewart and Plotkin (2012, Genetics)

- Typically, 10 to 50 amino-acid residues in the TF are involved in contacts with the DNA.

- Although each TF has maximum affinity for a specific DNA motif, there is no general regulatory code in TFs, i.e., no specific language involving one-to-one matching between the amino-acid sequence of a TF and the nucleotide sequence of its binding site.

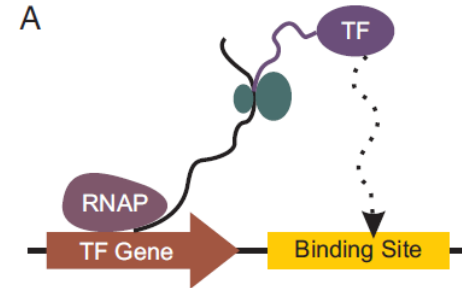
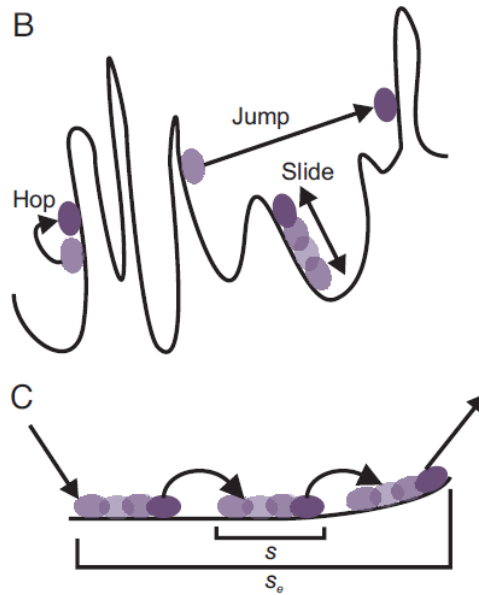
# Transcription-Factor Molecules Are Especially Rare in Cells





# Transcription-factor Molecules Locate Their Binding Sites by Facilitated Diffusion

- All TFs engage in promiscuous interactions with off-target sites as a consequence of the negatively charged phosphate backbones of the DNA and positively charged residues on the protein.



Pete von Hippel

In bacteria, transcription and translation are colocalized, and genes often appear in operons, increasing the chance of rapid localization.

A combination of one-dimensional sliding and three-dimensional jumping dramatically reduces the search time, relative to random diffusion.

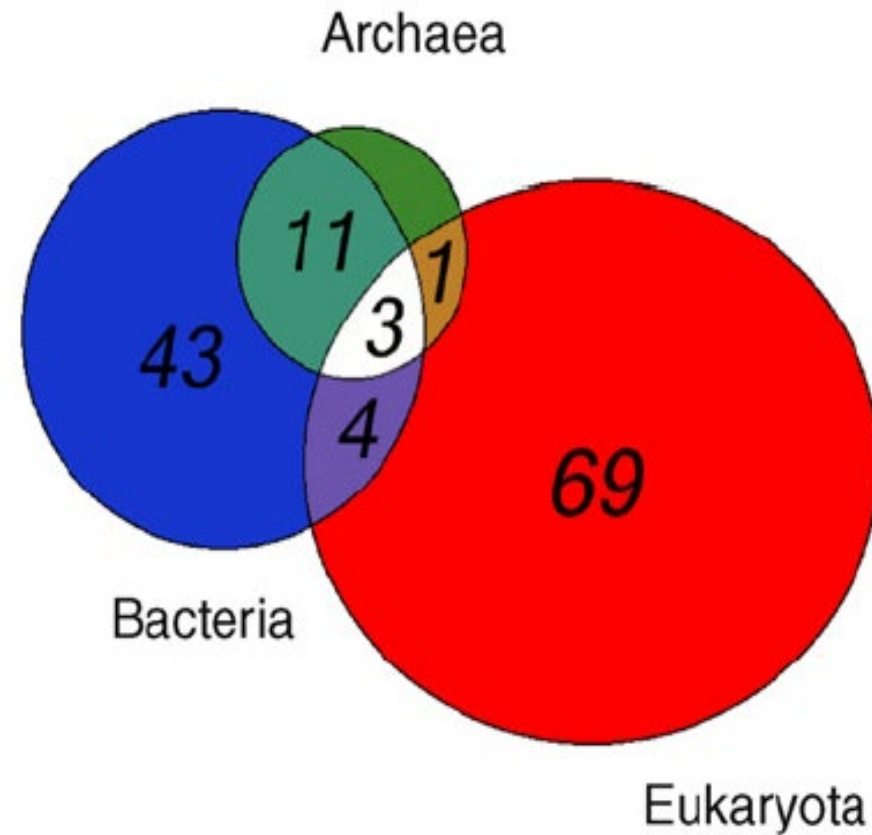
## How Rapidly Do TFs Find Their Cognate TFBs by Facilitated Diffusion?

---

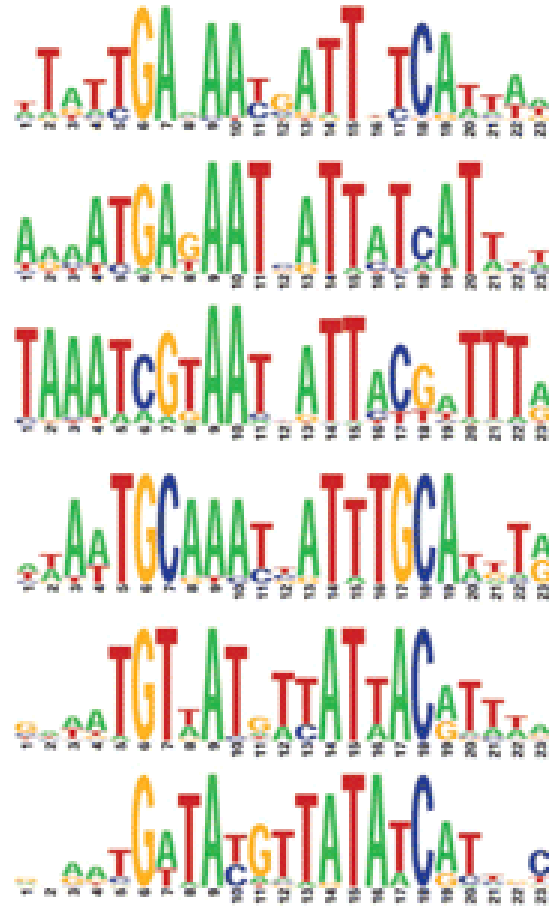
- With one-dimensional diffusion in *E. coli*, once on the DNA, it would take ~29 days for a TF to find a specific binding site.
- With three-dimensional diffusion, the encounter rate between jumps =  $4\pi(D_{3n} + D_{3p})(r_n + r_p)$ .
- Accounting for the size ( $r$ ) of nucleotides and TFs, and the diffusion rate of a TF ( $D_{3p}$ ), the time to jump from one location to another is  $\sim 2.5 \times 10^{-7}$  seconds.
- Once on the DNA, a TF spends  $\sim 0.0026$  sec diffusing over  $\sim 100$  bp, before falling off, so essentially all of the search time is spent directly interrogating the DNA, rather than jumping from spot to spot.
  - Because  $\sim 10^5$  100-bp scans are required to cover the entire genome, the estimated time to locate a site is  $10^5 \times 0.0026 = 260$  seconds.
  - With  $N_{tf}$  molecules in the cell, the search time would be reduced to  $260 / N_{tf}$  seconds.

Gene regulation in LUCA must have relied on transcription factors, but only a small fraction of known DNA-binding domains are shared across the three super-kingdoms.

---



## Transcription Factors Bind to Specific Motifs With Different Binding Affinities



- The motifs generally vary among individual client genes, and seldom match the consensus sequence.

# Towards a Physical Model for Understanding Gene-Expression Evolution: Characterizing Binding Sites by the Strength of Their Motif Sequences

- Affinity between TFs and their cognate TFBSs on the DNA are governed by hydrogen bonds.

Sum of the strength of binding over all nucleotide positions:

$$E(\mathbf{a}) = \sum_{i=1}^{\ell} \epsilon_i(a_i)$$

**Table 1.** Features of the motifs of well-studied transcription factors (TFs). Motif size is based on consensus sequences. The estimated costs of mismatches are obtained from binding-strength experiments in which single-base changes were made in motifs. Costs of single-base mismatches are in units of kcal/mol; these average to 1.4 across the full set of studies, or in terms of Boltzmann units ( $K_B T \simeq 0.6$  kcal/mol), to 2.3.

TF	Species	Motif (bp)	Cost of Mismatch		References
			Mean	Range	
CI	Lambda phage	17	1.4	0.5 – 3.5	Sarai and Takeda (1989)
Cro	Lambda phage	9	1.4	0.5 – 2.5	Takeda et al. 1989
Mnt	<i>Salmonella</i> phage P22	21	1.0	0.3 – 1.6	Fields et al. (1997); Berggrun and Sauer (2001)
CRP	<i>Escherichia coli</i>	22	1.7	0.9 – 2.5	Gunasekera et al. (1992); Kinney et al. (2010)
CRP	<i>Synechocystis</i> sp.	22	1.8	0.7 – 3.0	Omagari et al. (2004)
ArcA	<i>Shewanella oneidensis</i>	15	1.3	0.1 – 3.4	Schildbach et al. (1999); Wang et al. (2008)
Gcn4	<i>Saccharomyces cerevisiae</i>	11	1.0	0.5 – 1.7	Nutiu et al. (2011)
c-Myb	<i>Homo sapiens</i>	6	1.6	0.6 – 2.8	Oda et al. (1998)

## Probability a Particular Target Site with $m$ Matches is Bound

---

$N_{\text{tf}}$  = number of transcription-factor molecules in the cell,

$N_{\text{ot}}$  = number of competing functional binding sites for the transcription factor,

$G$  = nucleotide sites per genome (each of which can initiate non-specific binding),

$2$  = gain in binding in Boltzmann units.

$$P_{\text{on}} \simeq \frac{1}{1 + Be^{-2m}},$$

where  $B = G/(N_{\text{tf}} - N_{\text{ot}})$  is a measure of the concentration of background (non-specific) binding sites relative to the number of TF molecules available for the target site.

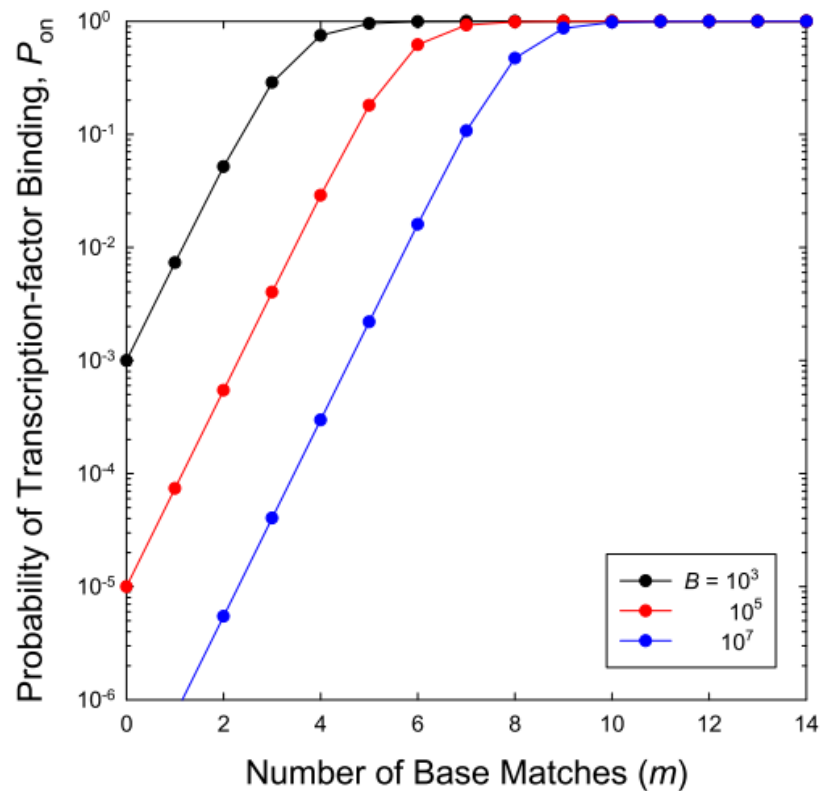
# What is the Magnitude of Background Off-Target Promiscuity?

---

- The number of off-target sites,  $G$ , is generally in the range of  $10^6$  to  $10^{10}$  bp, with prokaryotes falling at the lower end and multicellular eukaryotes at the higher end of the range.
- In bacteria, the numbers of molecules per cell for particular TFs,  $N_{tf}$ , are often in the range of 100 to 1000, with just a few cases ranging as high as 50,000. The number of genes serviced by a particular TF,  $N_{ot}$ , is generally  $<100$ .
- Thus,  $B$  is on the order of  $10^3$  to  $10^6$  for prokaryotes, and estimates for eukaryotes are in the same range.
- If other sources of interference exist (such as promiscuous binding to other proteins),  $B$  would be higher.

# The Probability of Binding-site Occupancy Typically Saturates at a Small Number of Matches

$$P_{\text{on}} \simeq \frac{1}{1 + Be^{-2m}}$$



← Implies a drift barrier

- Unless the level of promiscuous binding is enormous, there is little advantage of a binding-site lengths  $> 10$  bp.
- Two costs / limitations of using TFs to regulate genes:
  - 1) Changing the number of matches is a coarse-grained tuning mechanism.
  - 2) Owing to promiscuous binding to nonspecific sites, 100s of TF molecules need to be present in a cell to ensure that a host gene is turned on.



$$W(m) = 1 + \alpha P_{\text{on}} = 1 + \frac{\alpha}{1 + e^{-2(\ell - n) + \ln(B)}}$$

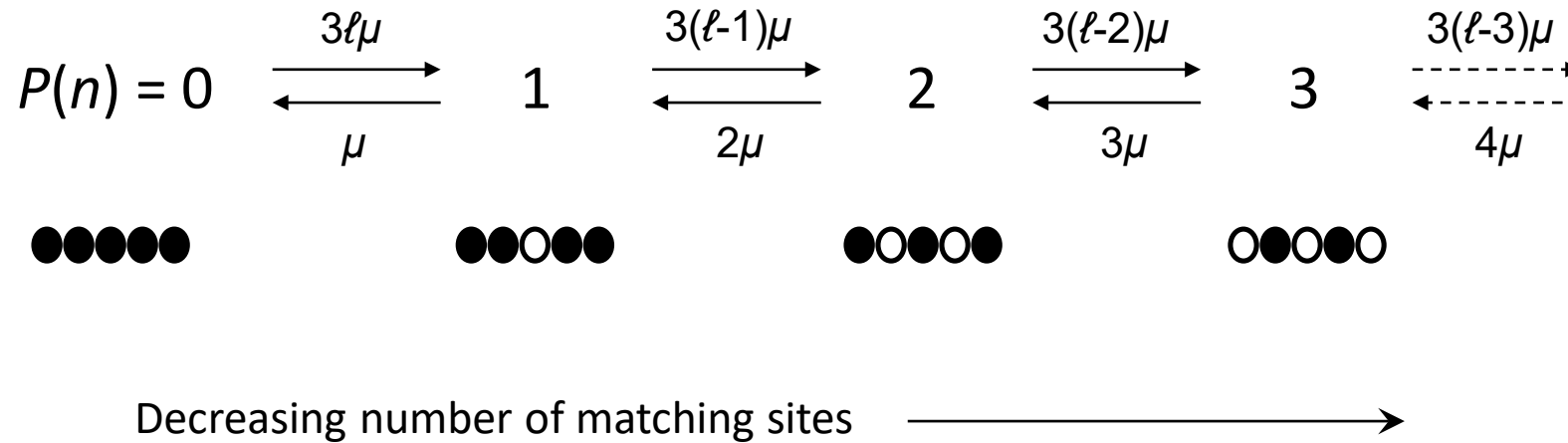
$n$  = number of mismatches

$\ell$  = length of the optimal motif sequence

$\alpha$  = scaling factor for the strength of selection

$B$  = measure of background interference

# Quasi-Equilibrium Evolutionary Distribution of Binding-Site Affinities: the Neutral Case



$\mu$  = the rate of mutation from nucleotide x to nucleotide y (reversion rate)

$3\mu$  = rate of loss of a correct site

The fixation probabilities are obtained from Kimura's (1962) diffusion equation for newly arisen mutations,

$$\phi_{x,y} = \frac{1 - e^{-2N_e s_{x,y}/N}}{1 - e^{-2N_e s_{x,y}}}$$

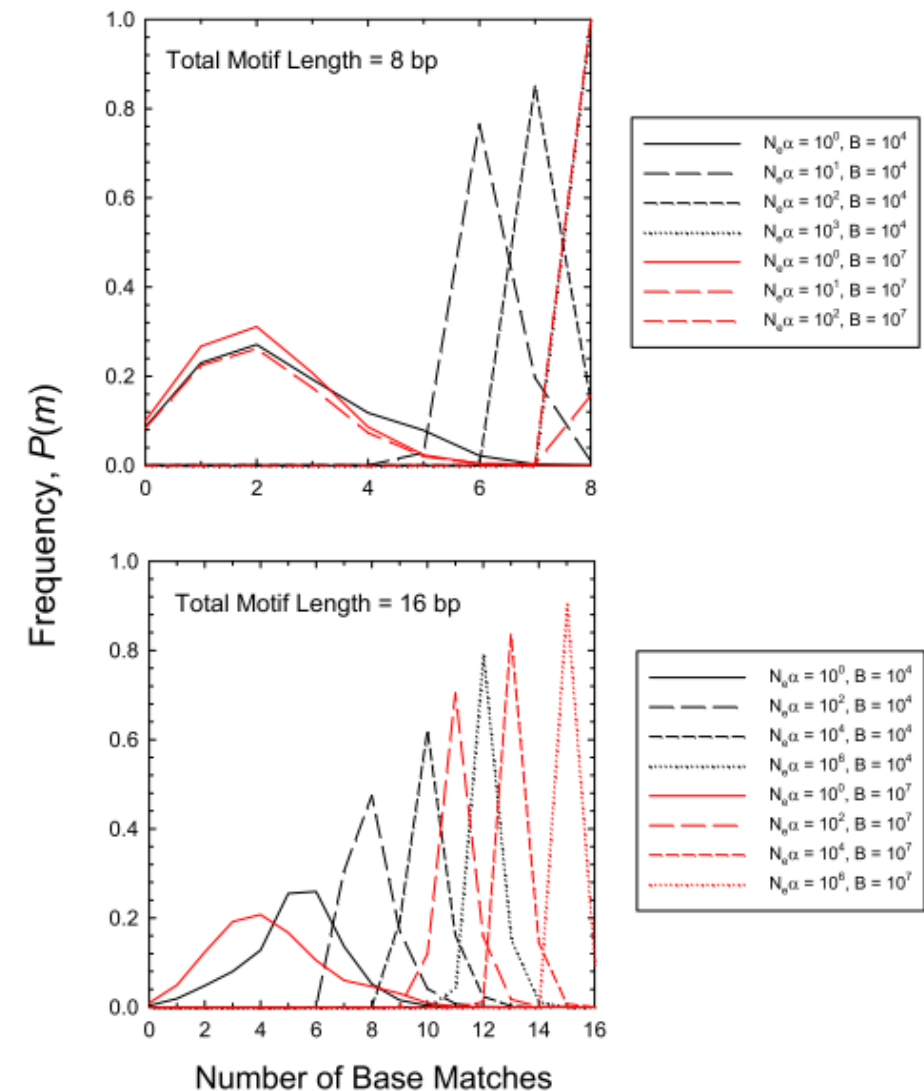
where  $N_e$  is the effective population size,  $1/N$  is the initial frequency of a mutation, and  $s_{x,y}$  is the fractional selective advantage of allelic class  $y$  over  $x$ .

# Equilibrium Distributions of the Number of Matches

- Distribution is independent of the mutation rate.
- Depends only on the multiplicity under neutrality.
- The distribution under selection is simply a weighting of the neutral expectation.
- Unless the power of selection relative to drift ( $N_e\alpha$ ) is extremely strong, the majority of motifs will contain mismatches.

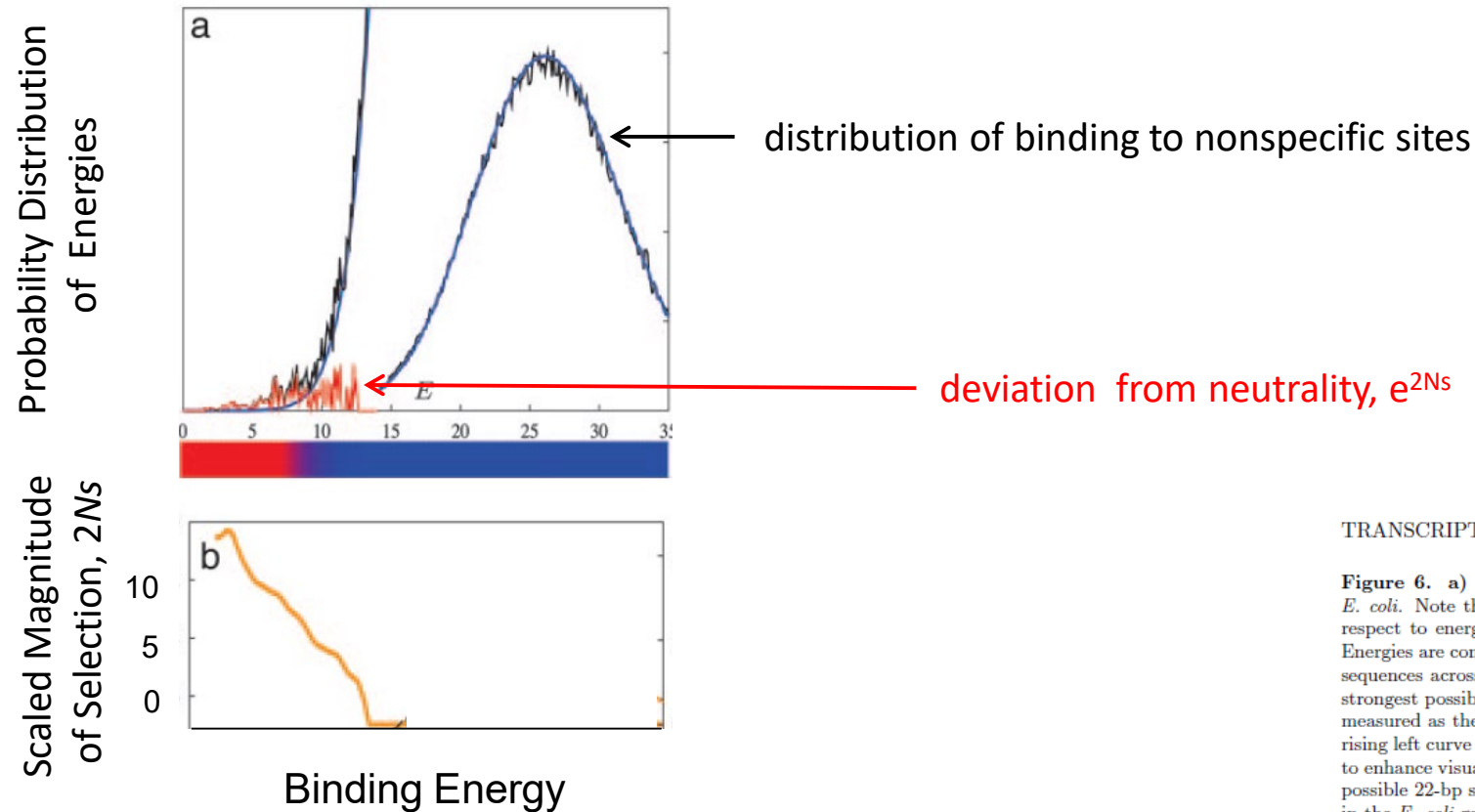
$$\tilde{P}(m) = C \left[ 3^n \binom{\ell}{n} \right]$$

the neutral expectation



# Using the Theory to Estimate the Strength of Selection on Binding Sites

- CRP motifs in *E. coli*.

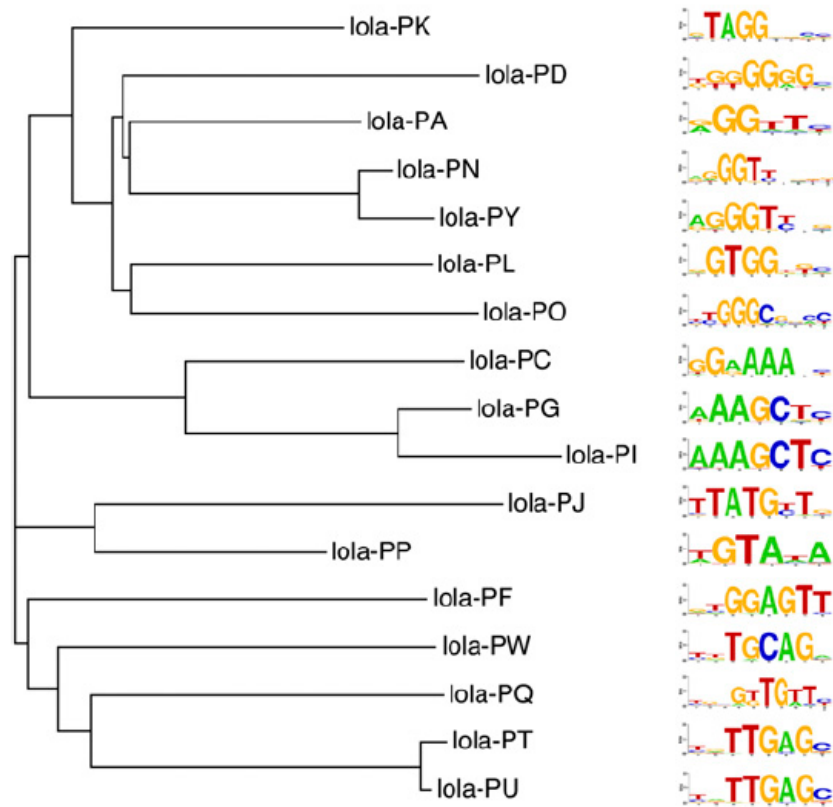


## TRANSCRIPTION

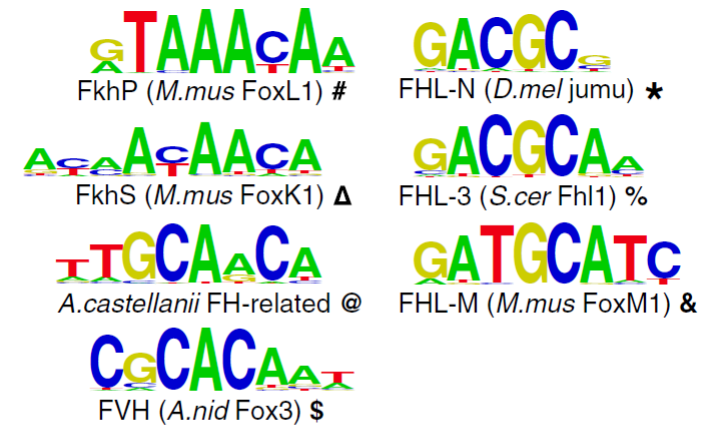
39

**Figure 6.** a) Distribution of binding energies associated with the transcription factor CRP in *E. coli*. Note that contrary to the approach in the text, the binding sites are characterized with respect to energy rather than mismatches, although the two scales are entirely interchangeable. Energies are computed using sliding windows of 22-bp (the length of the consensus TFBS for CRP) sequences across the entire *E. coli* genome. The energy scale is set such that  $E = 0$  denotes the strongest possible binding site, with all other (more weakly binding) motif sequences simply being measured as the deviation from this value (and appearing further towards the right). The rapidly rising left curve is the tail of the remainder of the energy distribution (to the right) multiplied by 30 to enhance visualization. The solid lines illustrate the expected distribution based on the full set of possible 22-bp sequences under a random (model using the known distribution of nucleotide types in the *E. coli* genome; these fit very well in the right portion of the distribution, which represents non-specific binding sites. The red line is the excess of motifs in the left tail from this neutral expectation. Motifs in the red region are viewed as true binding sites, whereas all others denote the background resulting from nonspecific binding. b) As discussed in the text, for TFBS motifs deemed to be functional, the logarithm of the ratio of observed abundance relative to that expected under neutrality (the red line),  $\tilde{P}/\tilde{P}_n$ , provides an estimate of  $2N_e s$ , which is equivalent to the selective advantage of each site relative to the power of drift. From Mustonen and Lässig (2005).

# Transcription-factor Binding Motifs Appear to Wander Over Evolutionary Time

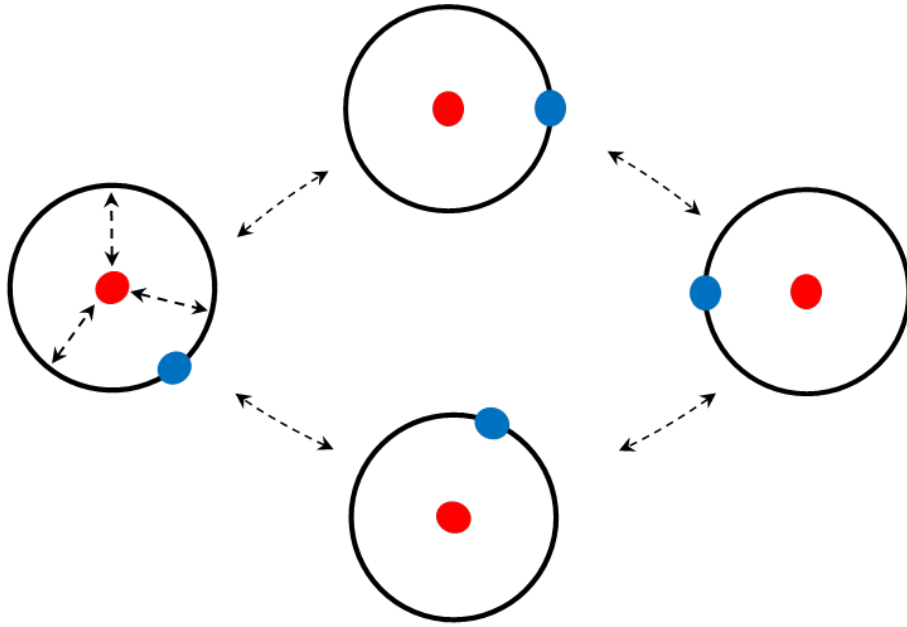


**Figure 2.** Comparison of isoform specificities. DNA-binding specificities of 17 Lola isoforms generated through alternate splicing. MatAlign cluster-gram emphasizing the diversity within the recognition motifs of the various Lola isoforms. All of the characterized ZFPs utilize a pair of zinc fingers to recognize DNA. Identical fingers are present in the lola-PN and -PY isoforms and the lola-PT and -PU isoforms, and both pairs have identical specificity.



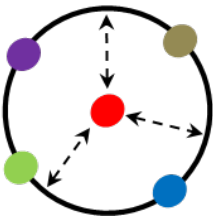
**Fig. 1.** DNA binding-site motifs bound by forkhead domain proteins. A representative member of each class of binding site discussed in the text is shown. Bold symbols are used to represent binding specificities in subsequent figures.

# Coevolution of the Regulatory Vocabulary: TFs and Their Binding Sites



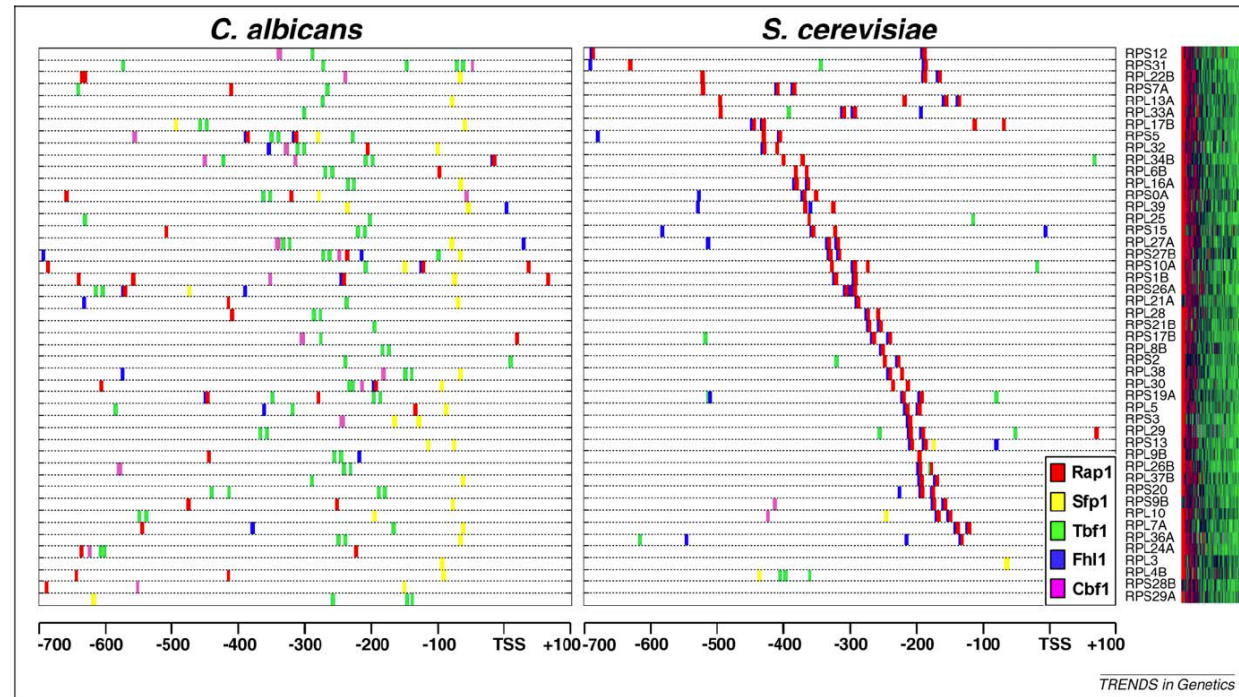
- TFs with larger numbers of target genes are more evolutionarily conserved at the amino-acid sequence level, including at the level of the recognition sequence.

- Decline in binding-site specificity with increasing numbers of genes serviced by a TF in both *E. coli* and yeast (Sengupta et al. 2002).



- Does such a condition evolve by selection so as to minimize the mutational burden on an organism?
- Or are TFs with low specificity recruited more frequently into various regulatory pathways over evolutionary time?

# Dramatic Rewiring of Regulatory Mechanisms Is Commonly Observed in Yeasts

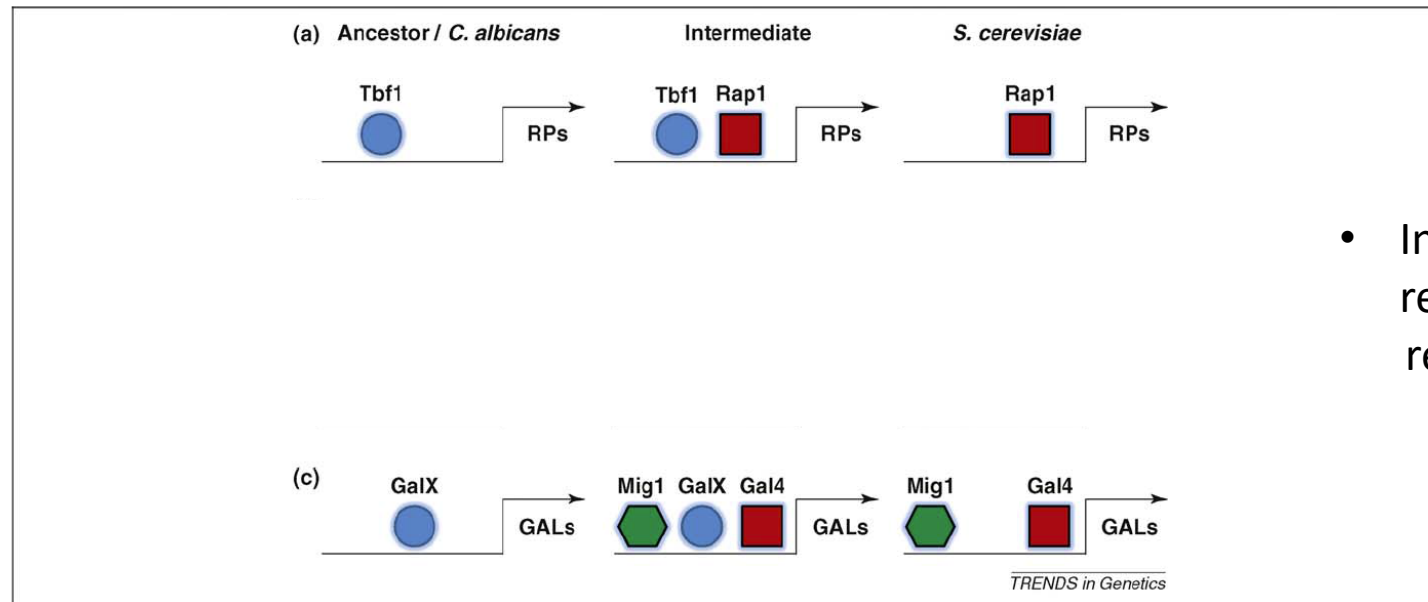


**Figure 1.** An overview of yeast Ribosomal Protein (RP) promoters. Promoter schematics and expression profiles of RPs in two yeast species. Each line displays information for one RP gene (left to right: promoter regions in *C. albicans* and *S. cerevisiae* (from -700 to +100 bp, relative to transcription start site, TSS), gene name and expression pattern). RPs were chosen based on annotations in *S. cerevisiae*, and restricted to those that have a 1:1 ortholog mapping to *C. albicans*, using InParanoid [109]. Colored boxes indicate locations of predicted TF binding sites (see key in bottom right corner). TFs were selected that have at least threefold binding site enrichment in promoter regions of RP genes in either species (relative to randomly selected promoter sets). Genes are sorted (top to bottom) in order of most distal appearance of Rap1 binding sites in *S. cerevisiae* (relative to TSS). Expression values range from twofold downregulated (green) to twofold upregulated (red). 111 experiments from [110] are shown that meet the criterion that at least 10 of the 47 genes have an absolute  $\log_{10}$  ratio of at least 0.1, sorted from highest to lowest average ratio among the 47 genes.

- Massive differences in the regulatory machinery associated with the ribosomal protein genes in the two yeasts *Saccharomyces cerevisiae* and *Candida albicans*. Nearly every TF used in Sc is utilized in a different way in Ca, and shifts in the consensus motifs for orthologous TFs occur as well.



# Potential Paths of Rewiring of Regulatory Modules Involving Intermediate Stages of Redundancy

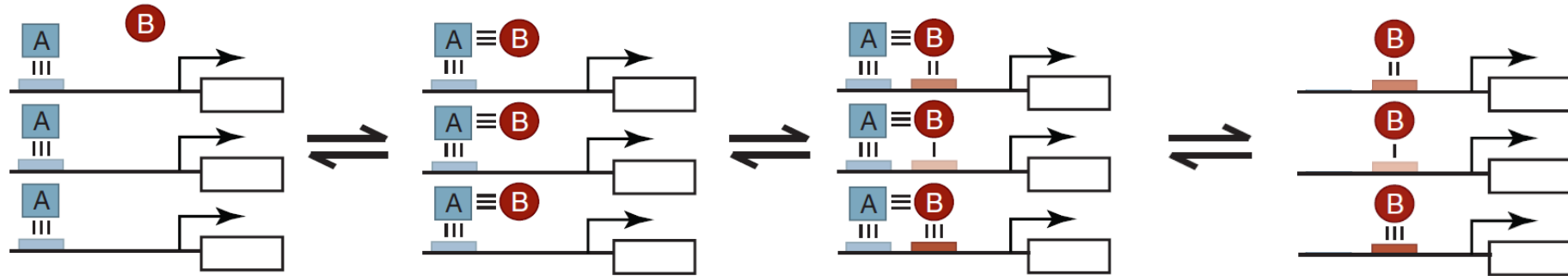


- Intermediate states of shared (redundant) regulatory motifs; subsequently experiencing reciprocal loss.

**Figure 2.** Mechanisms of TF switching in yeast. Three documented cases of TF switching between *C. albicans* and *S. cerevisiae*. TFs are depicted as colored shapes, with names depicted above. Gene regulatory regions are depicted as straight lines, with regulated genes indicated to the right. Arrows indicate the activation of the corresponding genes; thick lines ending in a bar indicate repression. (a) Mechanism for switching of RP gene control from Tbf1 to Rap1. (b) Mechanism for altering the transcriptional control of the mating type (MAT) locus while maintaining the same regulatory output of only expressing *asg1* in a-type cells. Promoter schematics for the MAT $\alpha$  and MAT $\alpha$  loci are depicted on the top and bottom, respectively. (c) Mechanism for switching the control of galactose metabolism (GAL) genes from an unknown ancestral regulator (GalX) to Gal4.

# Mechanisms of Regulatory Rewiring:

How can a large set of co-regulated genes experience a coordinated switch of their regulatory pathways?



**Fig. 2.** A plausible pathway to the concurrent rewiring of a large set of genes. In this scenario an interaction is acquired between TRs A and B, after which interactions between B and DNA are optimized gene-by-gene. Rewiring in this manner could avoid fitness barriers imposed by initially changing regulation one gene at a time.

