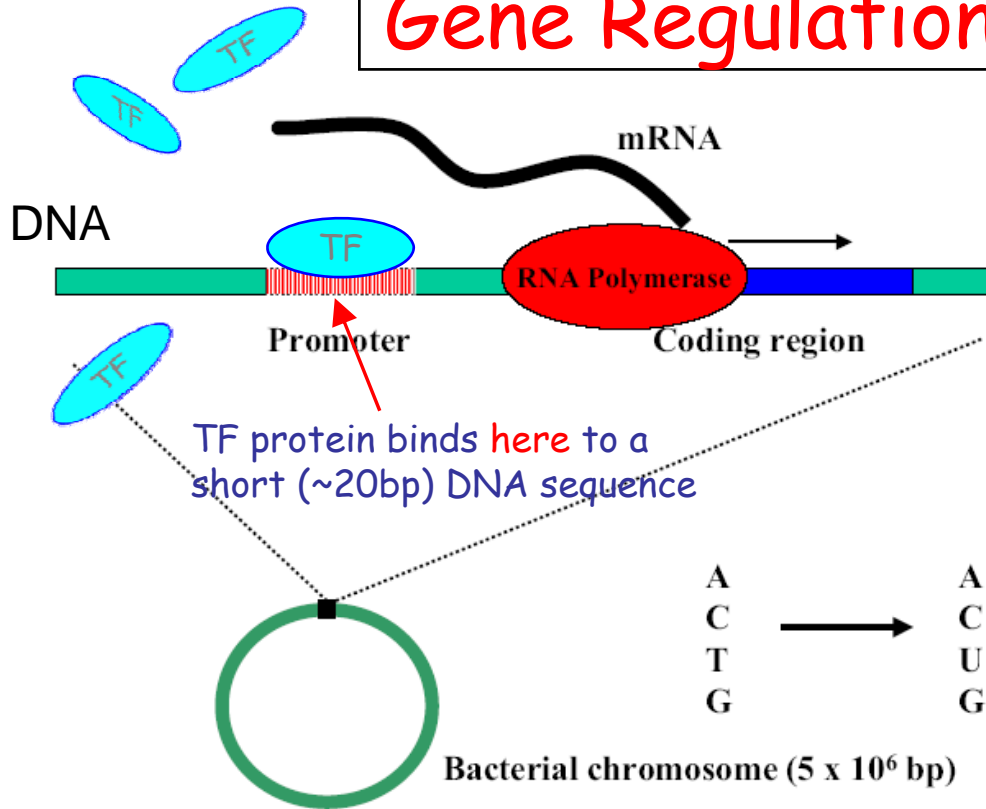


# Evolution of gene regulation: a simple stochastic model meets the real world (and proves fit)

Curtis Callan  
Physics Department  
Princeton University

Based on work with M. Laessig and V. Mustonen (Koeln) and J. Kinney (Princeton). An exploration of physical and statistical ideas about how evolution works in a very simple context where theory can be confronted with data.

# Gene Regulation: Overview



Transcription factor proteins (TFs) bind to promoter to help (hinder) RNAP copy gene to mRNA.

Sequence-dependent TF-DNA binding thermodynamics controls which TF binds to which gene.  
**Specificity is governed by energy.**

RNAP protein complex makes a mRNA copy of the gene. Ribosome translates triplets of bases into amino acids via the "genetic code".

**Coding Problem:** Same TF binds to many different sequences. No analog of 3bp codons. Binding loci statistically defined at best.

**Binding Energy:** Generic non-specific binding is weak; sequence-dependent binding provides a random energy landscape; strongest binding sites are the relevant ones (since they become occupied at lowest [TF]).

# Transcription Factor Binding Energies

First need energy as a function of sequence being “read” by the TF:

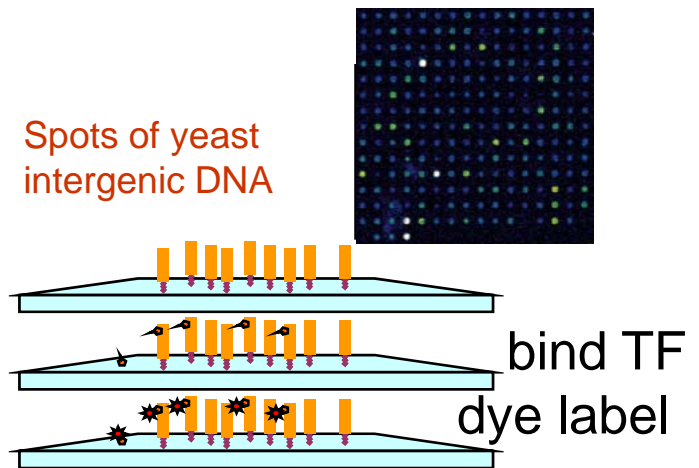


Simple “energy matrix” does a good job of capturing sequence-dependent TF binding energy:  
 $E(\text{TGTGAC})=0.7$  in the example

A	1.2	1.8	5.6	2.5	0.0	6.0
C	3.7	0.0	0.5	1.2	5.2	0.0
G	2.9	0.1	0.8	0.6	1.3	1.4
T	0.0	3.0	0.0	0.0	3.1	3.2

C A T G T G A C C T

Use binding assay data from chip technology to infer parameters of the energy matrix (could even do massive direct E measurements). From now on, we’ll study TFs where  $E(\text{seq})$  is “known”.



# Evolution Through the Lens of Energy

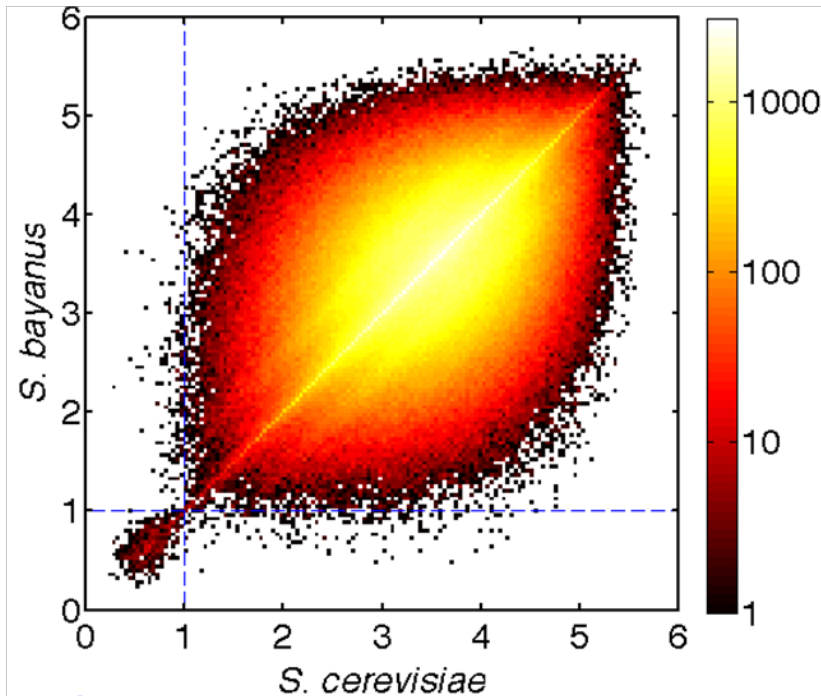
Binding site function is governed largely by energy and it, not sequence, should be conserved in evolution. Study by comparing “orthologous” binding sites between species (see example of aligned bacterial intergenic sequence)

*ecoli*  
*salmonella*

```

XXXXXXXXXXXXXXXXXXXXXXXXX
----TAAAGAGTGACGTAAATCACACTTTACAGCTAACTGTTTGTTTTTGTTTCATTGTA
AGTAAAAAATGTGATGTTCTGCAAACCTTACTGCTAATTGGCTGTTTTTGAACACTACTGTA
***  *****  **      **  *****  *****  **  *****  *  *****
    
```

← Crp site sequence varies (a lot) between *ecoli* and *salmonella*

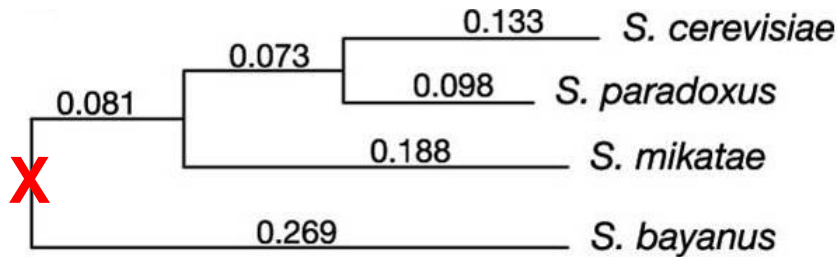


Need evidence that **energy is** conserved. **Yeast ABF1** is a great test case: 100s of binding sites, well-diverged family of sister species, accurate energy matrix is known.

Scatter plot all orthologous energy pairs for all intergenic regions of *S.cerevisiae* vs *S.bayanus* (40% sequence divergence)

For  $E < 1$  energy is conserved between two genomes (for 100s of sites). Evidence that selection acts on *energy* phenotype.

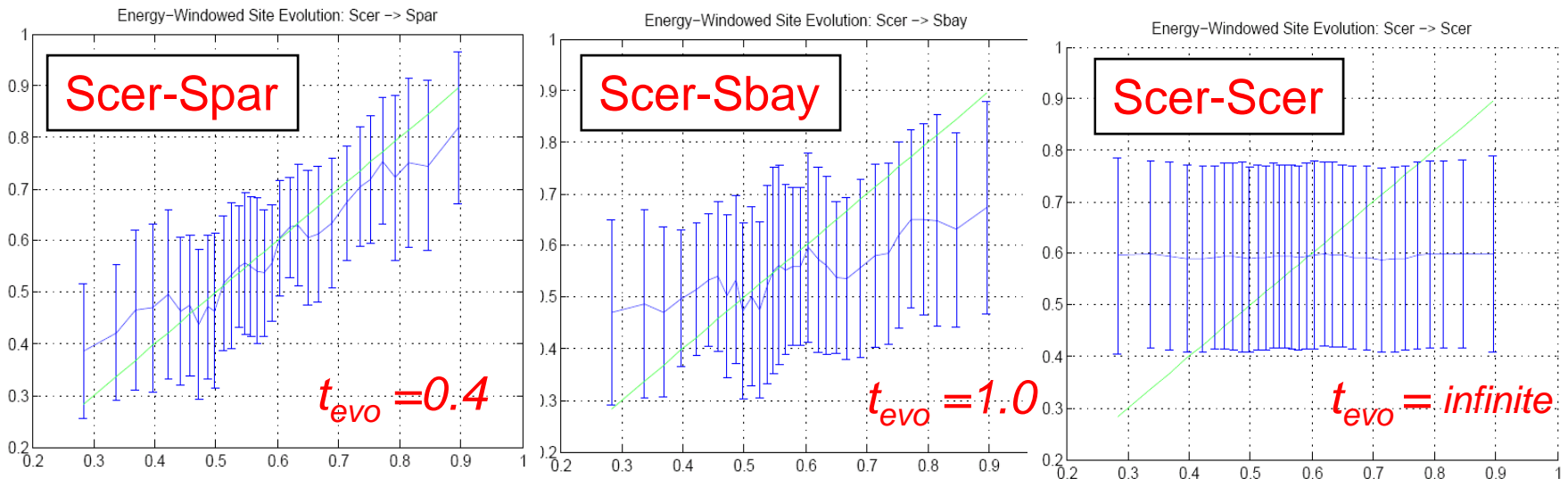
# Binding Energy Evolves in a Systematic Way



Phylogenetic tree of well-sequenced yeasts enables detailed study of site energy evolution statistics.

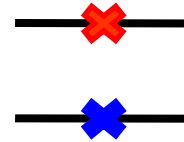
Identify functional sites by demanding orthologous site E conservation across four species: clean sample of ~600 sites

Order Scer sites by E, show how sites of similar E evolve to E-clouds in other species. Do for increasing divergence times (Scer to Spar, Sbay, ...). Simulate *infinite* divergence time by randomly pairing Scer sites with each other. Actual data seem to relax toward this equilibrium. Binding sites do a random walk in E like particles a confining potential: why?



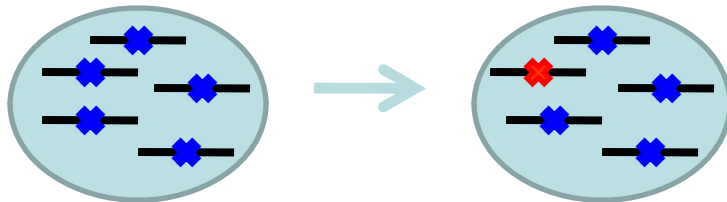
# Population Genetics of Binding Site Evolution (I)

Consider two different “alleles” of a specific TF binding site (red, blue). Base changes arise by mutation:

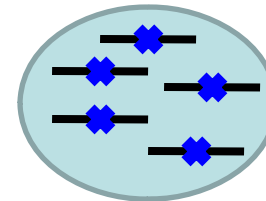


Symbol represents a single yeast; color stands for allele of a particular Abf1 site.

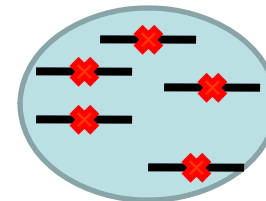
Clonal population of yeast organisms, of stable size  $N$ , living, reproducing, dying:



Background mutations occasionally cause a mutant allele to appear in **one** organism in the population of  $N$ .



Mutant allele usually dies out quite quickly.



It can, rarely, “sweep” the population (for finite  $N$ ). Even if two alleles have the same fitness!

As time marches on, different sequence states of this allele sweep the population. Neutral background rate for site **a** to go to site **b** is  $\mu_{ab}$ . What about site “fitness”?

# Population Genetics of Binding Site Evolution (II)

Some mutations lead to sites that bind very weakly: they will be unlikely to “fix”. Assume that “fitness” of a site depends on its sequence  $\sigma$  thru its energy:  $F(E(\sigma))$

Kimura-Ohta population genetics result: if mutation  $\mathbf{a}$  to  $\mathbf{b}$  leads to a fitness change  $\Delta F_{ab} = F(\mathbf{b}) - F(\mathbf{a})$ , fixation rate becomes

$$r_{ab} = \mu_{ab} \frac{2N\Delta F_{ab}}{1 - e^{-2N\Delta F_{ab}}}$$

Null site distribution  $P_0(\sigma)$  satisfies detailed balance under bkgd rate  $\mu_{ab}$ . Functional site distribution,  $Q(\sigma)$  satisfies detailed balance under K-O rate:

$$Q(\sigma) = \exp(NF(\sigma))P_0(\sigma)$$

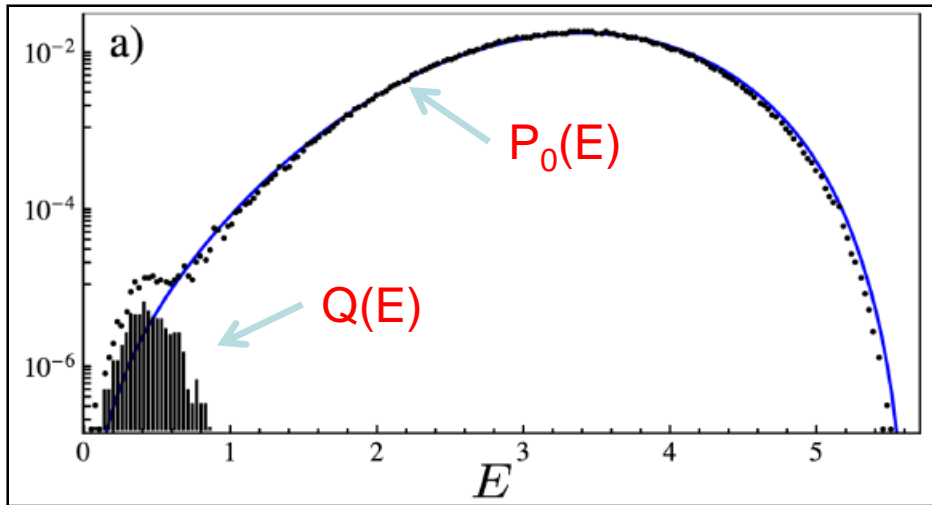
or, since  $F$  depends only on  $E$ ,

$$Q(E) = \exp(NF(E))P_0(E)$$

To exploit this, we assume all sites  $[\sigma_n]$  of a TF have same  $F(E)$ . Then phenotype distribution of one site over time equals its distribution over space (genomic sites of the TF).

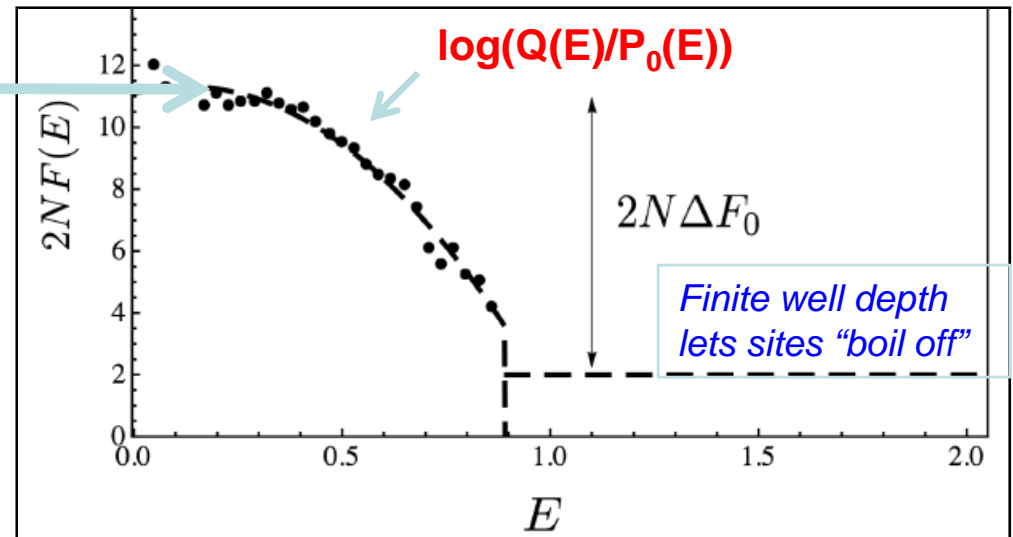
Read off fitness  $NF(E)$  from  $E$  distribution of TF binding sites in one species! Neat idea of Mustonen & Lassig (2005)

# Specific application to yeast ABF1



- Run energy matrix over *all* inter-genic sites to get  $E$  histogram (dots)
- Run over random genome to generate null model  $P_0(E)$  (blue line)
- Approx 600 “excess” low- $E$  sites give funct’l distribution  $Q(E)$  (bars)
- Functional sites confirmed by multi-genome  $E$  conservation

- Compute  $E$ -dependent fitness via Mustonen-Lässig relation:
- Identical results found using any genome on the yeast tree
- Sites evolves stochastically in a “potential well”  $U = -2NF(E)$
- “Temperature” is set by background point mutation rate.

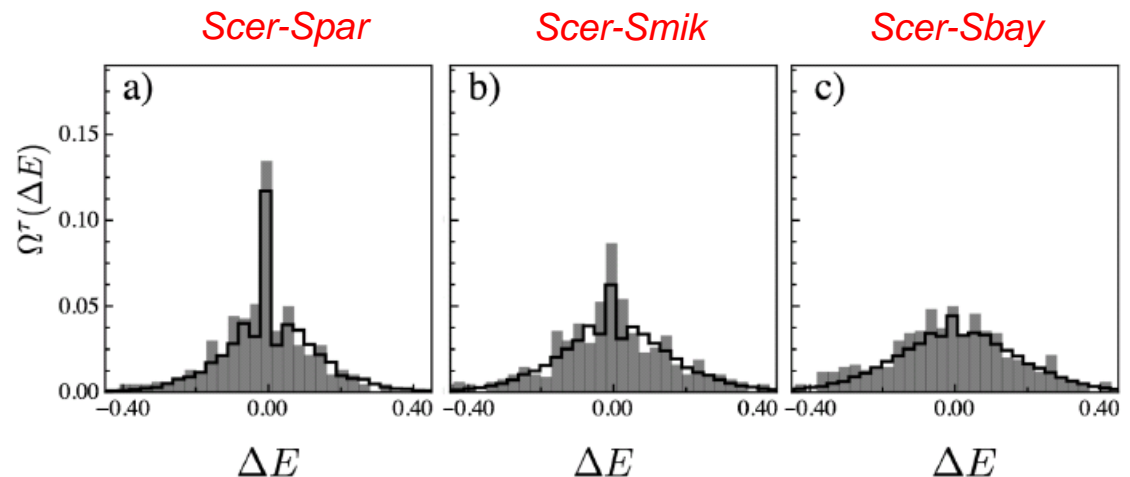


# Simulated Abf1 binding site evolution

- Start with a list of functional Abf1 sites in the initial genome  $\{\sigma\}_{Q_1}$ .
- Derive point mutation rate matrix  $\mu_{ba}$  from intergenic region average *Scer/Sbay* substitution rate (or synonymous codon usage).
- Use Gillespie algorithm to evolve sites in  $\{\sigma\}_{Q_1}$  over time  $t$  using **K-O** rates based on  $\mu_{ba}$  and fitness  $F(E)$  to assess single base substitution tries.
- Generate simulated sample  $\{\sigma\}_{Q_2}$  of Abf1 sites in evolved genome. Simulate phylogenetic tree by cloning sites at tree nodes and using branch times.
- Individual simulated site pairs carry no useful information, but statistics of site ensembles can be usefully compared with data

Diagnostic histogram of  $\Delta E$  values for evolved site pairs. Solid line: simulated data. Dark bars: real data.

Parameter-free account of site evolution as a process dominated by energy.



# What have we learned about evolution?

- Binding sites of broad-acting TFs like yeast Abf1 are a favorable arena for studying evolution.
  - A quantitative phenotype (energy) and its relation to phenotype (site sequence) are available
  - Existence of hundreds of independent binding sites makes statistical comparisons meaningful
  - Quantitative fitness function accessible (by inference) and simplistic assumptions (universality/time indep'ce) not crazy
- Stochastic picture suggests many informative experiments: direct energy measurement on a mass scale, directed genetic modification, ...
- Site turnover (loss/gain of function) lets us quantify max selective advantage of functional sites (quite small).