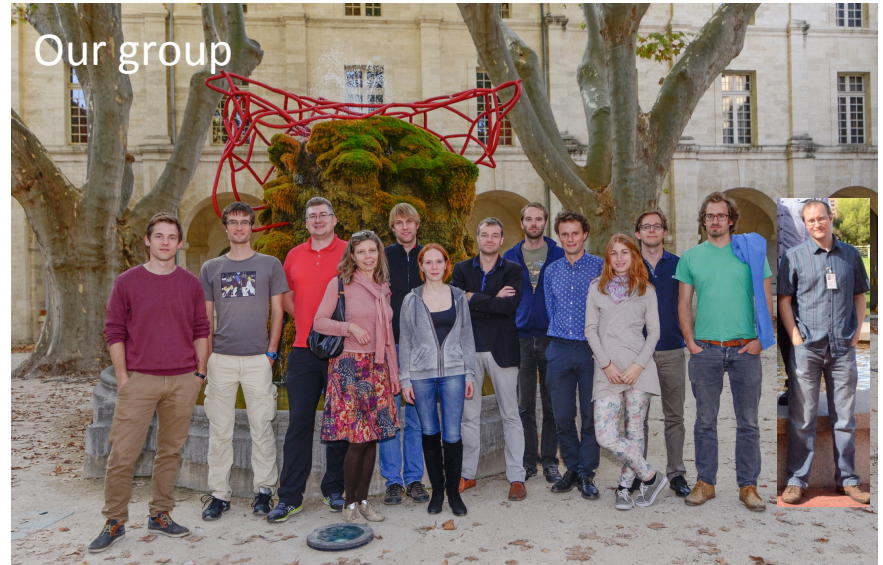


Structure, function and evolution of gene regulatory networks



Erik van Nimwegen
*Biozentrum, University of Basel,
and Swiss Institute of Bioinformatics*

Science is sort of done

The Lagrangian of the Standard Model

$$\mathcal{L} = -\frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{8}\text{tr}(\mathbf{W}_{\mu\nu}\mathbf{W}^{\mu\nu}) - \frac{1}{2}\text{tr}(\mathbf{G}_{\mu\nu}\mathbf{G}^{\mu\nu})$$

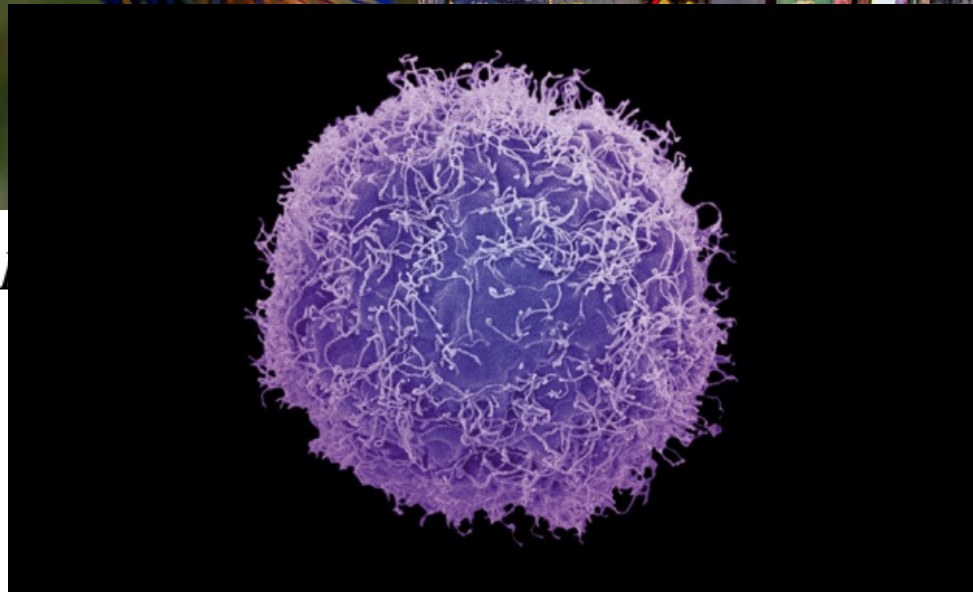
(U(1), SU(2) and SU(3) gauge terms)

$$D_{\mu}\nu_R + (\text{h.c.})$$

(lepton dynamical term)

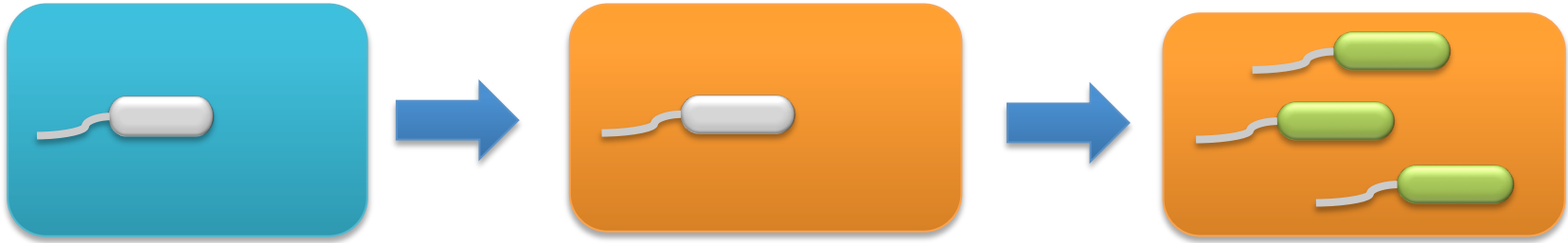


Gravity:

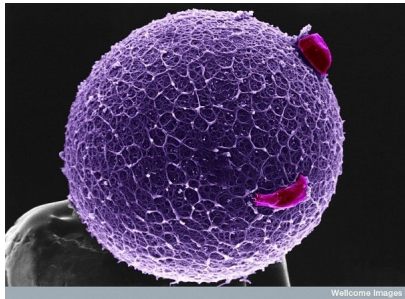


Gene regulatory networks

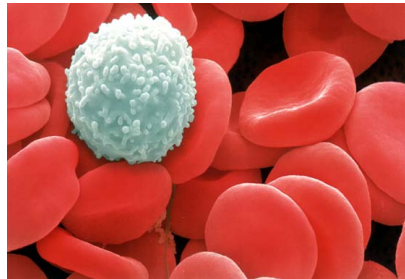
Allows cells to respond 'on their own behalf' to changing environments.



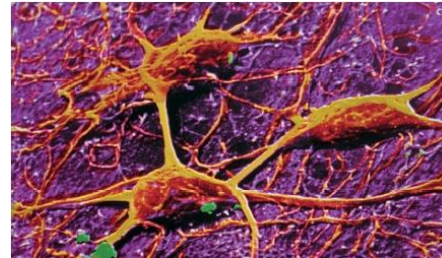
Allows multi-cellular organisms to express diverse phenotypes from one genome



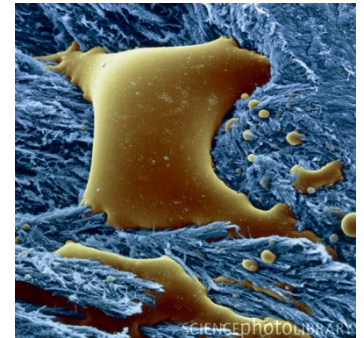
egg cell with 2 coronal cells



white and red blood cells

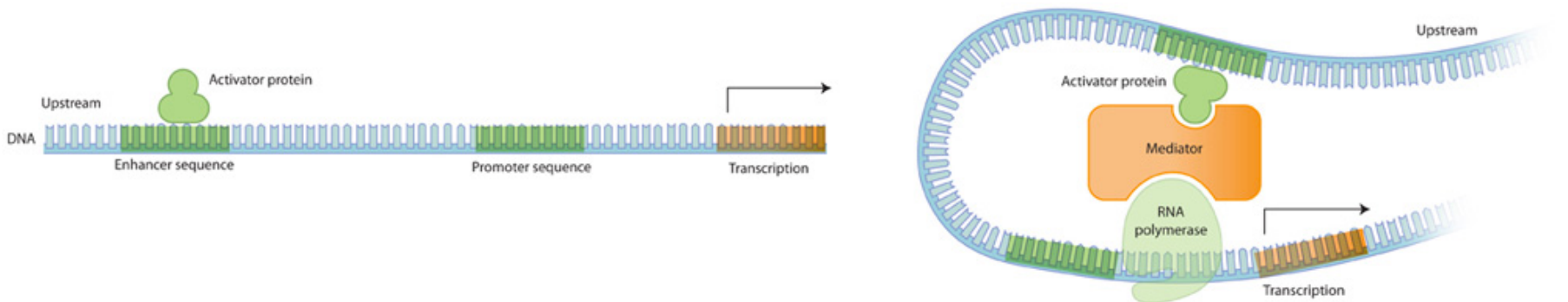


three neurons



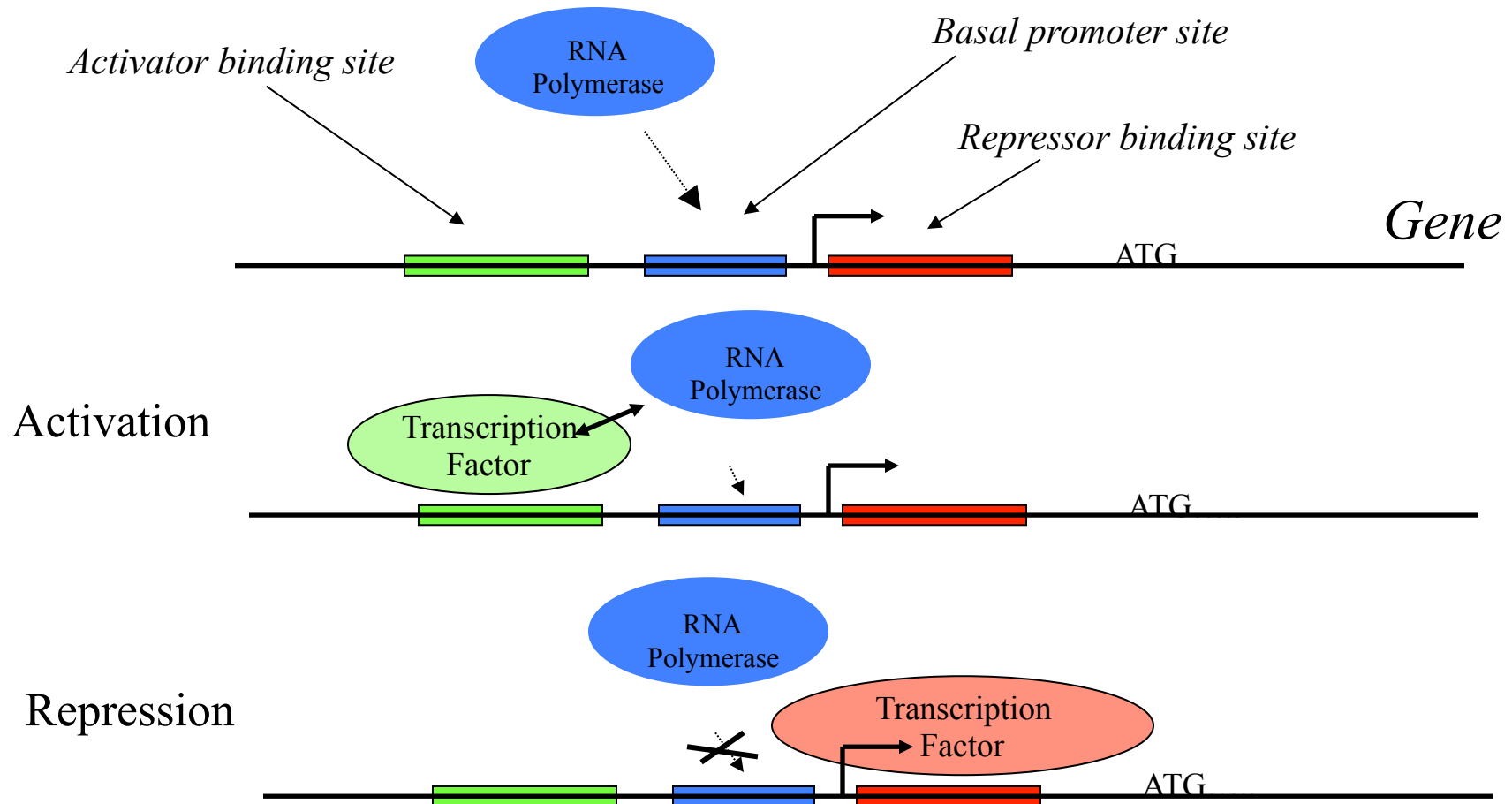
osteoclasts

The basic physical mechanisms are clear: Recognition of short DNA sequence fragments



Transcription regulation cartoon

- Transcription factors bind DNA in a *sequence-specific* manner (some sequences are much more strongly bound than others).
- Transcription factors often bind near *the transcription start site* of a gene.
- The binding of a transcription factor effects the rate of transcription initiation from the nearby start site.

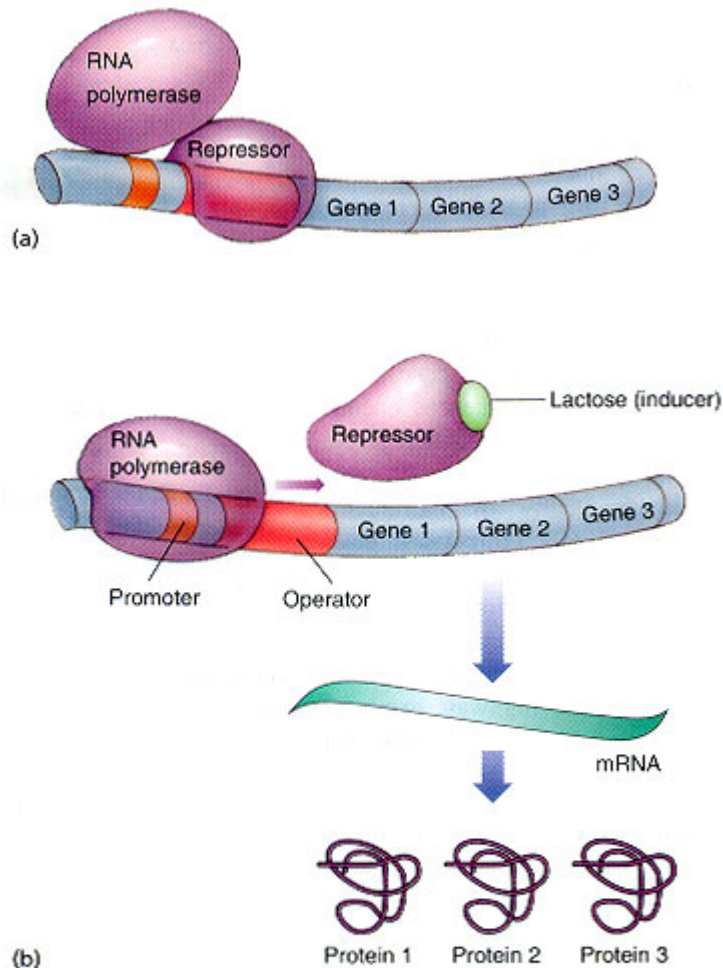


The hydrogen atom of gene regulation: the lac operon

J Mol Biol. 1961 Jun;3:318-56.

Genetic regulatory mechanisms in the synthesis of proteins.

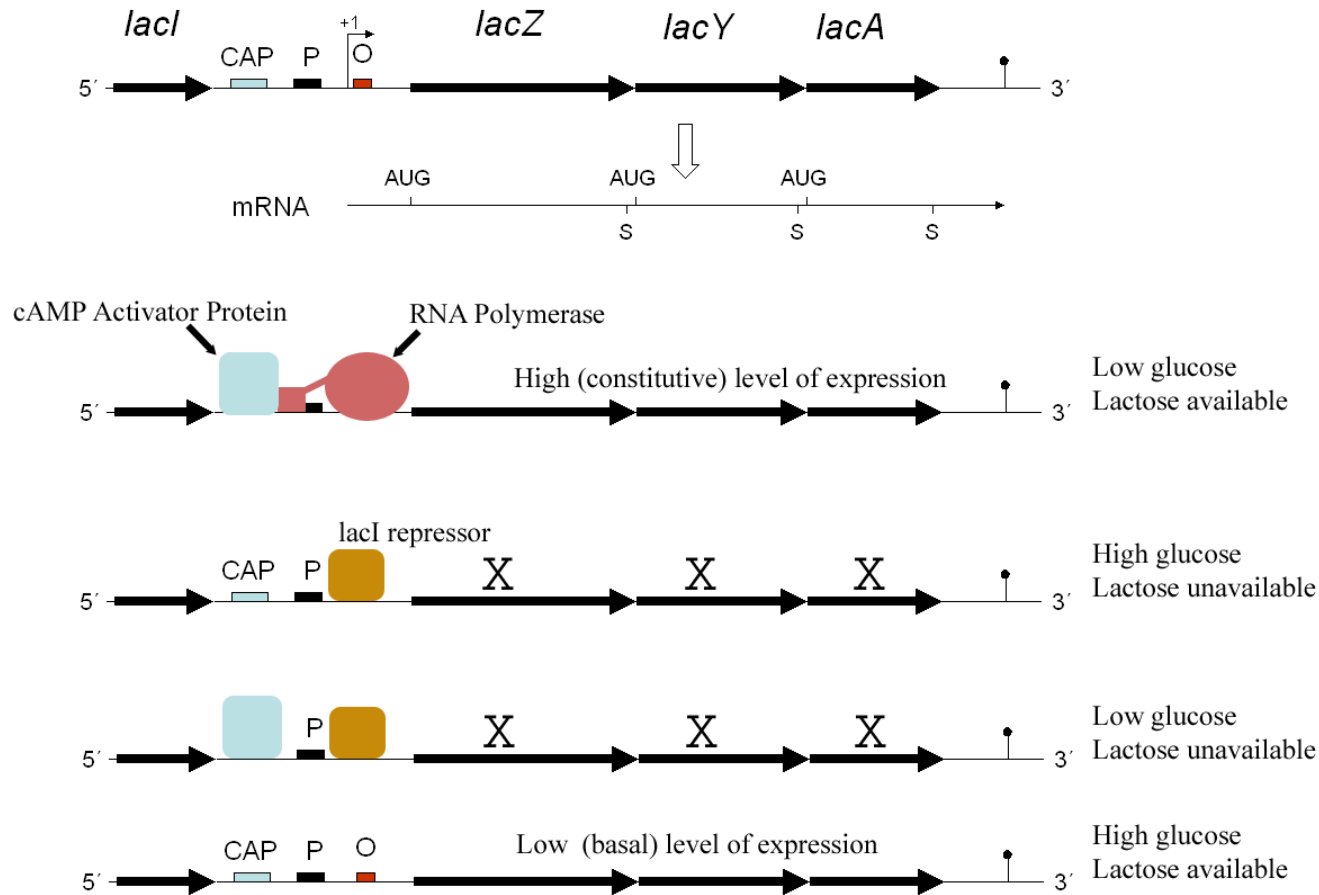
JACOB F, MONOD J.



- **Lac-operon:** *E. coli* has a set of three genes that code for proteins that are used to metabolize lactose.
- It additionally has a transcription factor, *lacI*, that *represses* transcription of the lac-operon.
- When lactose is present, it will *bind* to *lacI*.
- *LacI* bound by lactose *cannot bind DNA*.
- Given sufficiently high lactose concentration, the repressors are thus inactivated.
- As long as lactose remains present *E. coli* will keep transcribing the lac-operon and produce the enzymes to 'eat the lactose'.
- When the lactose disappears the repressor *lacI* will again bind to its site and shut transcription of the lac-operon back off.

Regulatory sites on the lac operon

The *lac* Operon and its Control Elements



Roughly speaking: the constellation of binding sites in the promoter determines how gene expression is regulated as a function of TF concentrations.

Outline of the lectures

Day 1

1. Computational methods for determining the constellations of regulatory sites.
2. From constellations of regulatory sites to genome-wide gene expression patterns.

Day 2

3. Large-scale patterns in genomes and gene regulatory networks.
4. Gene expression noise and its role in the evolution of *de novo* gene regulation.

Day 3

5. *How do bacteria genomes evolve in the wild?*

What is a binding site?

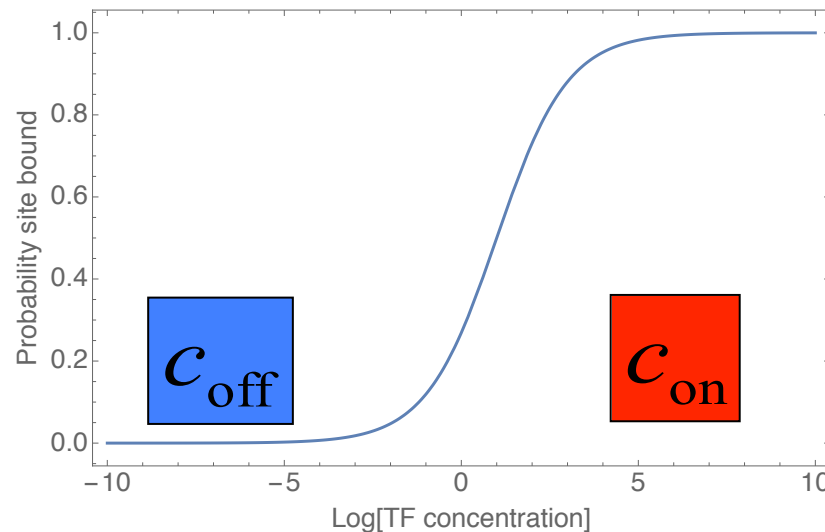
- **Question:** What does it mean to say that a *binding site* for a given transcription factor (TF) appears at some position in the DNA?



The interaction between the TF and the binding site is in essence characterized by two parameters:

1. The *binding energy* E of the interaction between TF and binding site.
2. The concentration c of the transcription factor.

As the TF concentration increases, the proportion of time the TF is bound to the site increases.



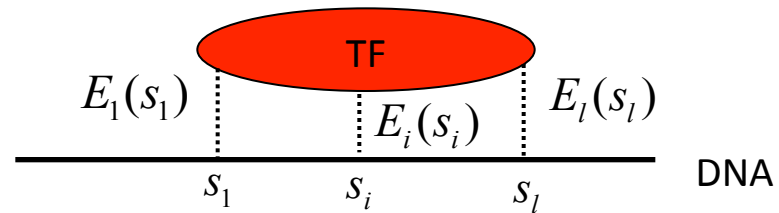
$$P_{\text{bound}} = \frac{ce^{\beta E}}{ce^{\beta E} + K}$$

If the TF's concentration goes between a low 'off' state and a high 'on' state, then binding sites are characterized by binding energies E such that they will be mostly unbound in the 'off' state and mostly bound in the 'on' state.

From binding energies to weight matrices

1. We *assume* the binding energy of a sequence s is an additive function of the individual bases:

$$E(s) = \sum_{i=1}^l E_i(s_i)$$



2. The probability for the site to be bound (ignoring other TFs) is a Fermi-function of energy $E(s)$ and concentration of the transcription factor c :

$$P_{\text{bound}}(s) = \frac{ce^{\beta E(s)}}{ce^{\beta E(s)} + K}$$

3. Assume that the only constraint on 'functional binding sites' is that the need to have some characteristic *average energy* E_* . Then we get using **maximum entropy**:

$$P(s) = \frac{e^{\lambda E(s)}}{\sum_{s'} e^{\lambda E(s')}} = \prod_{i=1}^l \left[\frac{e^{\lambda E_i(s_i)}}{\sum_{\alpha} e^{\lambda E_i(\alpha)}} \right]$$

where the Lagrange multiplier λ is chosen such that $\sum_s E(s)P(s) = E_*$

From binding energies to weight matrices

$$P(s) = \frac{e^{\lambda E(s)}}{\sum_{s'} e^{\lambda E(s')}} = \prod_{i=1}^l \left[\frac{e^{\lambda E_i(s_i)}}{\sum_{\alpha} e^{\lambda E_i(\alpha)}} \right]$$

This can be rewritten in terms of a *weight matrix* (WM) w containing probabilities:

$$P(s) = \prod_{i=1}^l P_i(s_i) \equiv \prod_{i=1}^l w_{s_i}^i \qquad w_{\alpha}^i = \frac{e^{\lambda E_i(\alpha)}}{\sum_{\alpha'} e^{\lambda E_i(\alpha')}}$$

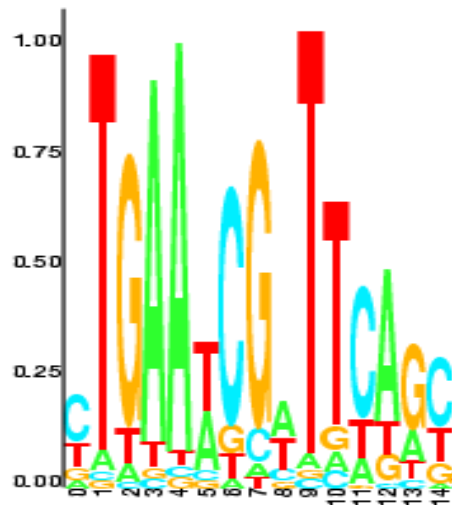
The probability that a binding site for the TF will have a sequence s is given by:

$$P(s|w) = \prod_{i=1}^l w_{s_i}^i$$

Note: This is **not** the probability that a segment with sequence s is a binding site!

Example weight matrix

Alignment of known binding sites for the *E. coli* TF **fruR**:



```
AAGCTGAATCGATTTTATGATTTGGT
AGGCTGAATCGTTTCAATTCAGCAAG
CTGCTGAATTGATTCAGGTCAGGCCA
GTGCTGAAACCATTCAAGAGTCAATT
GTGGTGAATCGATACTTTACCGGTTG
CGACTGAAACGCTTCAGCTAGGATAA
TGACTGAAACGTTTTTGCCCTATGAG
TTCTTGAAACGTTTCAGCGCATCTT
ACGGTGAATCGTTCAAGCAAATATAT
GCACTGAATCGGTTAACTGTCCAGTC
ATCGTTAAGCGATTTCAGCACCTTACC

**gcTGAAtCG*TTcAg**c*****
```

From this we estimate a weight matrix:

w_{α}^i = Probability of finding base α at position i .

For instance : $w_A^3 = 0.267$, $w_C^3 = 0.2$, $w_G^3 = 0.467$, $w_T^3 = 0.067$

Probability that a site for the TF represented by w will have sequence s :

$$P(s \mid w) = \prod_{i=1}^l w_{s_i}^i$$

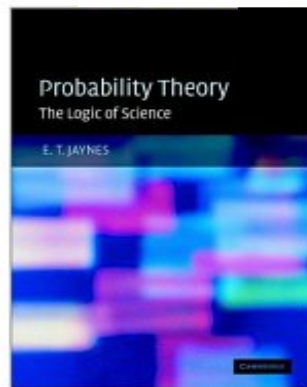
Finding sites for a known WM

Sidebar on the general Bayesian inference approach

1. Given a dataset D enumerate all hypotheses H that could have accounted for the data.
2. Assign *prior probabilities* $P(H)$ to each hypothesis H .
3. Define a likelihood model that gives the probability $P(D|H)$ of obtaining the data D under each of the hypotheses H .
4. The posterior probability $P(H|D)$ that hypothesis H produced the data is given by Bayes' theorem:

$$P(H|D) = \frac{P(D|H)P(H)}{\sum_{\tilde{H}} P(D|\tilde{H})P(\tilde{H})}$$

The most useful book I ever read: E.T. Jaynes: Probability Theory, the Logic of Science



Probability as the uniquely consistent extension of logic from true/false statements to statements with any degree of plausibility in the range $[0,1]$.

Finding a site for a known WM

Simplest example: We are given a sequence s of length L that is known to contain a *single site* for the TF represented by w . Task: find where the site is.

- **Hypotheses:** Positions i at which the site could start. $0 \leq i \leq L - l$ Prior: $P(i) = \frac{1}{L - l + 1}$
- **Likelihood:** Combination of the probability of the site and the other bases under a 'background' model.

$$\begin{array}{c}
 \underbrace{1 \quad \quad \quad i+1 \quad \quad i+l \quad \quad \quad L}_{\text{Sequence indices}} \\
 \text{GAGATCGCTTAGGTAGTTTAATAGTT} \text{AGTCACACCC} \text{ACCCACCCATTGCTTAGAATATAGAATATAGAGCAGTTGCAGT} \\
 \underbrace{\hspace{10em}}_{P(s_{[0,i]} | b)} \quad \underbrace{\hspace{10em}}_{P(s_{[i,l]} | w)} \quad \underbrace{\hspace{10em}}_{P(s_{[i+l,L-i-l]} | b)}
 \end{array}$$

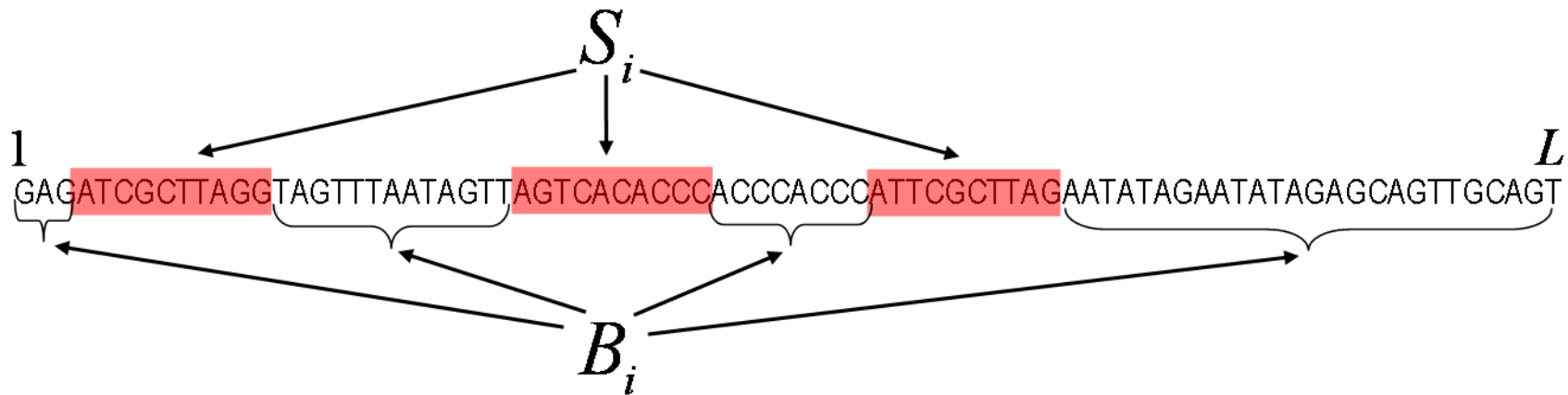
$$\begin{array}{lll}
 \text{Site:} & \text{Left background:} & \text{Right background:} \\
 P(s_{[i,l]} | w) = \prod_{k=1}^l w_{s_{i+k}}^k & P(s_{[0,i]} | b) = \prod_{k=1}^i b_{s_k} & P(s_{[i+l,L-i-l]} | b) = \prod_{k=i+l+1}^L b_{s_k}
 \end{array}$$

$$P(D | i) = P(s_{[0,i]} | b) P(s_{[i,l]} | w) P(s_{[i+l,L-i-l]} | b)$$

Posterior that site occurs at position i :
$$P(i | D) = \frac{P(D | i)}{\sum_{j=0}^{L-l} P(D | j)}$$

Finding multiple sites for a known WM

Arbitrary site configurations: $i = (i_1, i_2, \dots, i_n)$



Likelihood:

$$P(D|i) = \left[\prod_{\sigma \in B_i} b_{\sigma} \right] \prod_{s \in S_i} P(s|w)$$

Prior: Assume there is a constant probability π per position for a site to occur.

$$P(i) \propto \pi^{n(i)} (1 - \pi)^{L - \ln(i)}$$

Posterior:

$$P(i|D) = \frac{P(D|i) \pi^{n(i)} (1 - \pi)^{L - \ln(i)}}{\sum_j P(D|j) \pi^{n(j)} (1 - \pi)^{L - \ln(j)}}$$

Note: In the denominator we have to sum over all possible binding site configurations j .

Finding multiple sites: recursion relation for the partition sum

Sum over all configurations: $P(D) = \sum_j P(D | j) P(j)$

F_n = Forward sum up to position n in the sequence.

Recursion relation: $F_n = F_{n-1}(1 - \pi)b_{s_n} + F_{n-l}\pi P(s_{[n-l,l]} | w)$

$$\begin{array}{c}
 F_n \\
 \text{GAGATCGCTTAGGTAGTTTAATAGTTAGTCACACCCACCCACCCATT} \text{CGCTTAGAATATAGAATATAGAGCAGTTGCAGT} \\
 \\
 \parallel \\
 \begin{array}{cc}
 F_{n-1} & b_{s_n} \\
 \text{GAGATCGCTTAGGTAGTTTAATAGTTAGTCACACCCACCCACCCATT} \text{CGCTTAGAATATAGAATATAGAGCAGTTGCAGT}
 \end{array} \\
 + \\
 \begin{array}{cc}
 F_{n-l} & P(s_{[n-l,l]} | w) \\
 \text{GAGATCGCTTAGGTAGTTTAATAGTTAGTCACACCCA} \text{CCACCCATT} \text{CGCTTAGAATATAGAATATAGAGCAGTTGCAGT}
 \end{array}
 \end{array}$$

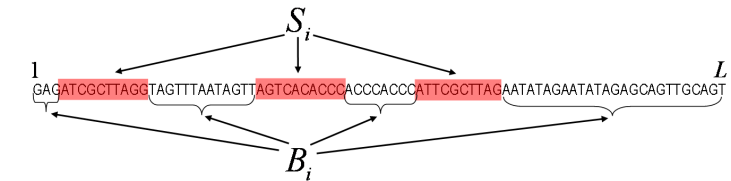
Similarly for the backward sum from position n to the end:

$$R_n = b_{s_n}(1 - \pi)R_{n+1} + P(s_{[n-1,l]} | w)\pi R_{n+l}$$

Finding multiple sites: expected number of sites

Partition sum: $P(D) = F_L = R_1$

Posterior:
$$P(i | D) = \frac{P(D | i) \pi^{n(i)} (1 - \pi)^{L - n(i)}}{P(D)}$$

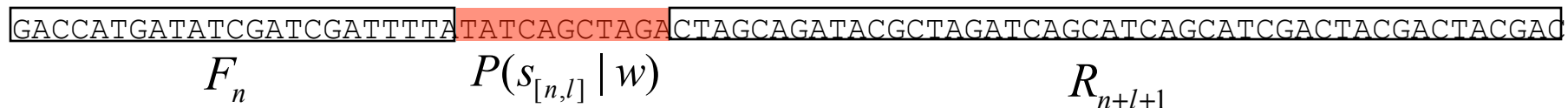


The posterior of any particular configuration i is not very useful: there are so many possible configurations that no individual configuration is going to have a high posterior probability.

Probability to find a site at position n independent of everything else.

$\{n\}$ = Set of all configurations i that have a site at position n .

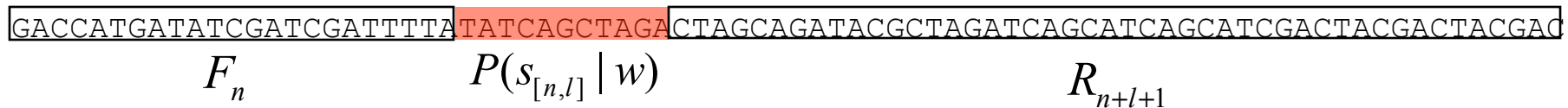
$$P(\{n\} | D) = \sum_{i \in \{n\}} P(i | D) = \frac{F_n P(s_{[n,l]} | w) \pi R_{n+l+1}}{F_L}$$



Finding multiple sites: summary of the results

Expected total number of sites:

$$\langle n \rangle = \sum_{n=0}^{L-l} P(\{n\} | D) = \sum_{n=0}^{L-l} \frac{F_n P(s_{[n,l]} | w) \pi R_{n+l+1}}{F_L}$$



Summary

When given as input:

- A sequence s .
- A weight matrix w .
- The *a priori* expected density of sites π ,

we can determine, in a time linear in the sequence length:

- The probability $P(\{n\} | D)$ for a site to occur at any position n .
- The probability $P(i | D)$ of any binding site configuration i .
- The probability F_n of the entire sequence up to n averaged over all possible configurations.
- The expected number of binding sites $\langle n \rangle$ in the sequence.

Optimizing the prior π

What if we do not know the prior π ?

Probability theory tells us that, to obtain the probabilities independent of π , we need to specify a prior $P(\pi)$ and integrate over all possible π , e.g.:

$$P(i | D) = \int_0^1 P(i, \pi | D) d\pi = \int_0^1 P(i | \pi, D) P(\pi) d\pi$$

Similarly, to estimate π we would like to calculate its posterior probability:

$$P(\pi | D) = \frac{P(D | \pi) P(\pi)}{\int_0^1 P(D | \pi) P(\pi) d\pi}$$

Unfortunately these integrals can not be easily calculated.

Alternative: estimate π by maximization of the probability. This is a reasonable approach especially when $P(D | \pi)$ is sharply peaked as a function of π .

Optimizing the prior π

Optimization with respect to π :

$$\begin{aligned} \frac{d \log(P(D))}{d\pi} &= \sum_i \frac{P(D|i)}{P(D)} \frac{dP(i)}{d\pi} = \sum_i \frac{P(D|i)}{P(D)} \frac{d[\pi^{n(i)}(1-\pi)^{L-n(i)l}]}{d\pi} = \\ &= \sum_i \frac{P(D|i)}{P(D)} P(i) \left[\frac{n(i)}{\pi} - \frac{L-n(i)l}{1-\pi} \right] = \sum_i P(i|D) \left[\frac{n(i)}{\pi} - \frac{L-n(i)l}{1-\pi} \right] = \frac{\langle n \rangle}{\pi} - \frac{L-\langle n \rangle l}{1-\pi} \end{aligned}$$

The optimal value of π obeys :
$$\pi = \frac{\langle n \rangle}{\langle n \rangle + (L - \langle n \rangle l)}$$

At the optimum, π matches the expected fraction of sequence that is sites.

Expectation-Maximization procedure:

1. Start with some value of π .
2. Calculate $\langle n \rangle$
3. Set π according to the equation on the left.
4. Go to step 2.

Partition sum with multiple WMs

All calculations described so far easily generalize to an arbitrary number of WMs.

Let π_w denote the prior for WM w (for simplicity we can consider the background one of the WMs which happens to have length 1).

Let l_w denote the length of WM w .

Recursion relation for the forward sum:
$$F_n = \sum_w F_{n-l_w} \pi_w P(s_{[n-l_w, l_w]} | w)$$

AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACATAAGTTGATATTCCTTTGATATCGACGACTA

AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACATAAGTTGATATTCCTTTGATATCGACGACTA

AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACATAAGTTGATATTCCTTTGATATCGACGACTA

AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACATAAGTTGATATTCCTTTGATATCGACGACTA

AAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACATAAGTTGATATTCCTTTGATATCGACGACTA

Equations determining the optimum prior:
$$\text{constant} = \frac{d \log(P(D))}{d \pi_w} = \frac{d \log(F_L)}{d \pi_w} = \frac{\langle n(w) \rangle}{\pi_w}$$

Update equation for the Expectation Maximization:
$$\pi_w = \frac{\langle n(w) \rangle}{\sum_{\tilde{w}} \langle n(\tilde{w}) \rangle}$$

Inferring an optimal WM

Recollect the derivative of the partition sum of the data $P(D)$ with respect to a WMs prior:

$$\frac{d \log(P(D))}{d\pi} = \frac{\langle n \rangle}{\pi}$$

Similarly, one can show that the derivative with respect to a *component of the WM* is given by:

$$\frac{d \log [P(D|w)]}{dw_{\alpha}^k} = \frac{\langle n_{\alpha}^k \rangle}{w_{\alpha}^k}$$

$\langle n_{\alpha}^k \rangle$ = The expected number of sites for w with letter α at position k .

$\langle n_{\alpha}^k \rangle$ = Sum over all putative sites with letter α at position k , each weighed with its posterior probability .

At the optimum we must have : $w_{\alpha}^k = \frac{\langle n_{\alpha}^k \rangle}{\langle n \rangle}$

Expectation maximization of the WM:

Bailey and Elkan
 Proc Int Conf ISMB 1994

1. Start with a random WM w .
2. Calculate the total number of expected sites $\langle n \rangle$ and the expected numbers $\langle n_{\alpha}^k \rangle$ with this WM w .
3. Set the new WM entries $w_{\alpha}^k = \frac{\langle n_{\alpha}^k \rangle}{\langle n \rangle}$
4. Iterate.

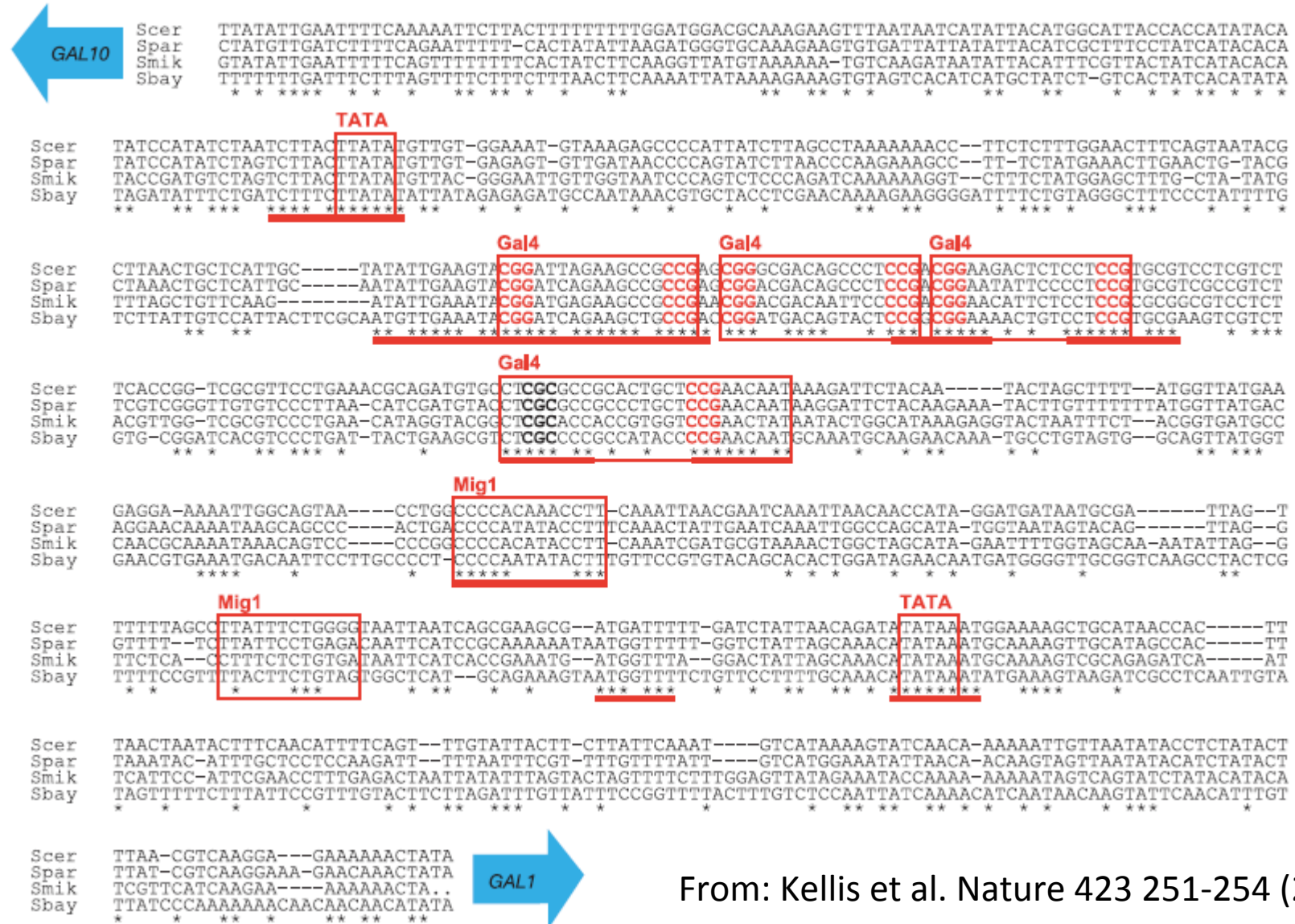
Phylogenetic Footprinting:

Finding regulatory sites through comparative genomics

Rationale

- Related genomes will share much of their gene regulation.
- Many regulatory sites were present already in the common ancestor of related species and will often have been conserved through evolution.
- One can try to identify regulatory sites by searching for short sequence segments in intergenic regions that are surprisingly conserved across related species.

Phylogenetic Footprinting: conceptual idea



From: Kellis et al. Nature 423 251-254 (2003)

Phylogenetic Footprinting: your mileage may vary

```
Scer ATGTTTTTTTAATGATATATGTAACGTACATTCTTTCTCTACCACTGCCAATTCGGTATTATTTAATTGTGTTTAGCGCTATTTAC
Spar -ATGTTTTTTTAATGATATATGTAACGTACATTCTTC--CTACTGCTACCAAGTCGGTATTATTTAATTGTGTTTAGCGCTATTTAC
Smik -----TCTTTTCTCTA--CCACTACTACCAATTCGGTATTATTTAATTGTGTTTAGCACTATTTAC
Sbay --ATGTTCTTAATGATATATATAACGTACATTTTTT--CCTCTACTAGCCAATCGGTGTTATTTAATTGTGTTTAGCTCTATTTAC
          *   * * *   *   * *   * *   * *   * *   * *   * *   * *   * *   * *   * *   * *   * *   * *   *
```

```
Scer TAATTAAGTAGAACTCAATTTTTAAAGGCAAAGCTCGCTGACCT--TTCCTGATTTTCGTGGATGTTATACTATCAGTTACTCTTC
Spar CCACTAAGTAGAACTCGATTTTTAAAGGCAAATTCAGTGTCT--TTCCTAGTTTTGCAGATGTCTTGCTATCAGCTACTTCCC
Smik TCCTAAC-AAAACTCAATTTTGAAGGGCTGA-TTAAATATCCTCCTTTAATAGTTTTGCGCTTAGCCTGTTATCA--TATAAGTA
Sbay TCCTTAACAAAAAACCAACTTCAAAAGTATAATAACAATAATTC-TCCGTTGATCTTGTGAAGTACATGCTATCACTTATTTGCC
      * * *   * * * *   * *   * * * *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
```

GCN4

ABF1

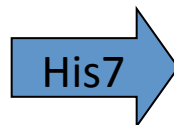
```
Scer TGCAAAAAAAAA-----TTGAGTCATATCGTAGCTTTGGGATTATTTTTCT-CTCTCTCCACGGCTAATTAGGTGATCATG
Spar TGCAGAAAAGAAAAATA-----TTGAGTCATATCATCGTCTAGGAAGTGTTTTCT-CTCTCTCCACGGATAGTTAAGTGATCATG
Smik TACAAAAGAGAAATAT-----TTGAGTCATATCATCGCCTAGGAAGTATTTTTTTCTCTCTCTCCACGGTTAATTAGGTGATTTCT
Sbay TGTA AAAAGAAAATCGTTTCGTTTGGAGTCATATCATGTTCTCATAA-TATTTTTTTT--TTCCTTAGCGATTAA-----
      *   * * * *   *   * * * * * * * *   *   *   * * * * * *   *   * *   * *   *
```

GCN4

```
Scer AAAAAATGAAAAATTCATGAGAAAGAGTCAGACATC-GAAACATACATAA--GTTGATATTC-CTTTGATATCG-----ACGACTA
Spar AAAAAATGAAAAATTCATGAGAAAGAGTCAGACATC-GAAACATACATAA--ATTGATATTC-CTTTAGCTTTT----AAAGACTA
Smik GAAAAACGAAAAATTCATG-GAAAGAGTCAGACGTC-GAAACATACATAA--ACCGATATTT-CTTTAGCTTTTCGACAAAAATCTG
Sbay GAAAAATAAAAAGTGATTG-GAAAGAGTCAGATCTCCAAACATACATAATAACAGGTTTTTACATTAGCTTTT---GAAAACTA
      * * * *   * * * *   * *   * * * * * * * *   * *   * * * *   *   *   *   *   *
```

```
Scer CTCAATCAGG-TTTTAAAAGAAAAGAGGCA-GCTATTGAAGTAGCAGT-ATCCAGTTTAGGTTTTTTAATTATTTACAAGTAAA-GA
Spar CTCAATCAA--GTTTAAATAGAAGAAAGAGG-AAGGTTGAGATAGGTAT-ATCCAGTTTAGGTTTC--AATTATTTAATAATAAA-GG
Smik CAATATTCATTATTTCAAACTCAAAAGAAG-AAGGTTGCAATTGGTGT-GTCCAGTTTAGGCTCT--AATTGTTGAATAATAAAAGG
Sbay TCCACCACAA-ATTGAAGGTGAGGAAGAAACAAAGTTAAAGCAAGAATCGGCTTGTGTCTTTTTT--GATTGCGTATT--TGAAAGG
          * *   * *   * *   * *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
```

```
Scer AAAAGAGA-----
Spar TAAAGAA-----
Smik CGAAGAAATAACGATCCAAAAA
Sbay TAAAGGAATACAACAAAAA---
      * * *
```

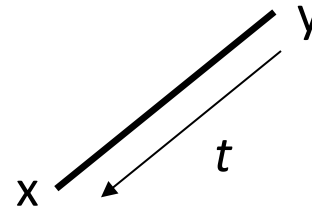


These approaches are *ad hoc*.
Use of an explicit evolutionary model is desirable.

Nucleotide evolution along a single branch of a phylogenetic tree

Probability to evolve from ancestor base y to descendant base x :

$$P(x | y, t, w)$$



- This probability depends both on mutation rates and on the effects of selection.
- In regulatory sites the preferred bases vary from position to position and we thus want to allow selection pressures to vary from position to position.

μ_{xy} = The rate of mutations from y to x (per unit time).

f_{xy} = The probability that a mutation from y to x will be fixed in the population.

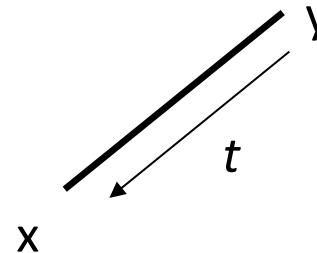
- Halpern and Bruno (*Mol. Biol. Evol.* 1998) derived that, if in the limit of large time there is a probability w_x to find nucleotide x then the probabilities of fixation are given by

$$f_{xy} = \mu_{xy} \log \left[\frac{\mu_{yx} w_x}{\mu_{xy} w_y} \right] \left(1 - \frac{\mu_{xy} w_y}{\mu_{yx} w_x} \right)^{-1}$$

Simpler model for evolution along a branch

Probability to end up with base x after time t , starting from base y , assuming WM column w .

$$P(x | y, t, w)$$



General time evolution:

$$\frac{dP(x | y, t, w)}{dt} = \sum_z \mu(x \leftarrow z | w) P(z | y, t, w) - \mu(z \leftarrow x | w) P(x | y, t, w)$$

Assumptions:

$$P(x | y, \infty, w) = w_x$$

$$\mu(x \leftarrow y | w) = \mu(x | w)$$

Solution:

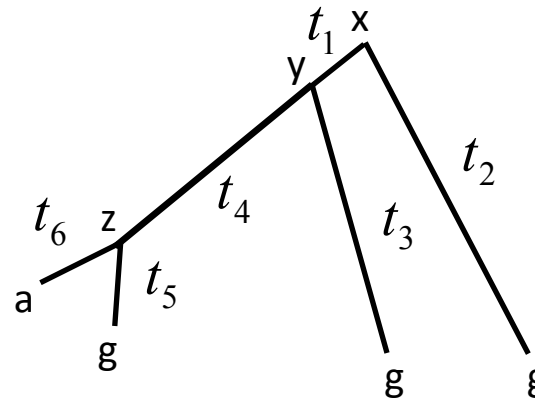
$$\mu(x | w) = w_x \mu$$

$$P(x | y, t, w) = \delta_{xy} e^{-\mu t} + w_x (1 - e^{-\mu t})$$

Probability of an alignment column under the evolutionary model

species A	acgtaactagtg
species B	acgttgctagatg
species C	tcgttgctataat
species D	aggtagcgagaag

S



Phylogenetic tree
relating the species.

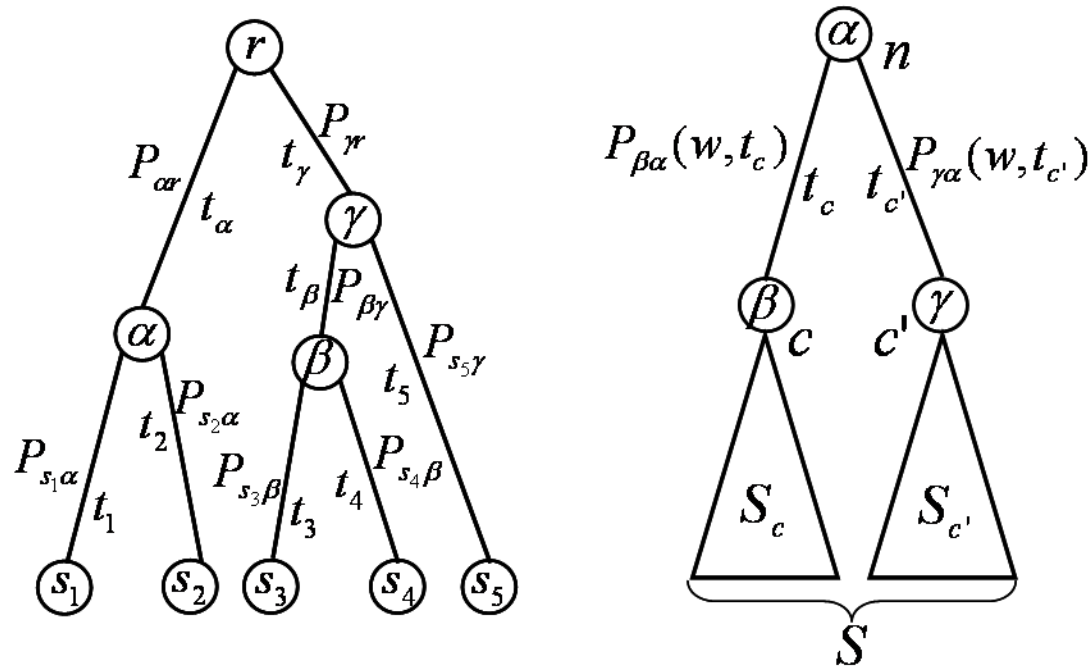
The probability $P(S | T, w)$ of the bases at the leafs given the tree T and the limit frequencies w is the product over the transition probabilities along each of the branches, summed over the possible bases at the internal nodes:

$$P(S | T, w) = \sum_{x,y,z} w_x P(y | x, t_1, w) P(g | x, t_2, w) P(g | y, t_3, w) P(z | y, t_4, w) P(g | z, t_5, w) P(a | z, t_6, w)$$

Note:

It seems that to evaluate this expression we have to sum over the states at all internal nodes, so that the number of terms would grow exponentially with the number of species. However, we can use a recursion relation on the tree (Felsenstein, 1981).

Probability of an alignment column: recursion relation on the tree



At every tree node, the probability of the data under that node, given the node is α is given by:

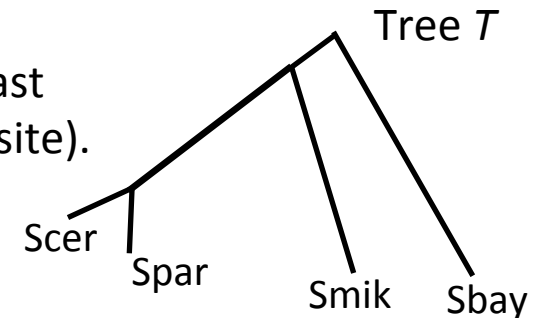
$$D_{\alpha}(n | w) = \prod_{m \in c(n)} \left[P(\alpha | \beta, t_m, w) D_{\beta}(m | w) \right]$$

The probability of the alignment column given the tree and the WM is given by summing over the base at the root:

$$P(S | T, w) = \sum_{\alpha} w_{\alpha} D_{\alpha}(r | w)$$

Finding binding sites for a known WM in multiple alignments

A multiple alignment of orthologous intergenic regions from yeast species, with a hypothesized binding site configuration (a single site).



```
Scer  AAAAAATGAAAAATTCATGAGAAAGAGTCAGACATC-GAAACATACATAA--GTTGATATTC-CTTTGATATCG-----ACGACTA
Spar  AAAAAATGAAAAATTCATGAGAAAGAGTCAGACATC-GAAACATACATAA--ATTGATATTC-CTTTAGCTTTT----AAAGACTA
Smik  GAAAAACGAAAAATTCATG-GAAAAGAGTCAACCGTC-GAAACATACATAA--ACCGATATTT-CTTTAGCTTTTCGACAAAAATCTG
Sbay  GAAAAATAAAAAGTGATTG-GAAAAGAGTCAGATCTCCAAACATACATAATAACAGGTTTTTACATTAGCTTTT----GAAAACTA
```

$$P(S | w, T)$$

A column that is part of a binding site is assigned a probability by taking the limit frequencies w equal to the WM entries at the corresponding position in the weight matrix.

$$P(S | b, T)$$

A column that is part of the background is assigned a probability by taking a limit frequencies w the average base frequencies b in intergenic regions.

MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model

Alan M Moses^{1,2} ✉, Derek Y Chiang³ ✉, Daniel A Pollard¹ ✉, Venky N Iyer⁴ ✉ and Michael B Eisen^{2,3,4} ✉

¹ Graduate Group in Biophysics, University of California, Berkeley, CA 94720, USA

² Center for Integrative Genomics, University of California, Berkeley, CA 94720

³ Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA

⁴ Department of Genome Sciences, Genomics Division, Ernest Orlando Lawrence Berkeley National Lab, 1 Cyclotron Road, CA 94270, USA

✉ author email ✉ corresponding author email

MotEvo: Finding sites for multiple WMs in multiple alignments

Arnold et al, *Bioinformatics*. 2012 Feb 15;28(4):487-94.

$$F_{n-1} \quad P(S_n | b, T)$$

	F_{n-1}	$P(S_n b, T)$
Scer	AAAAAATGAAAAATTCATGAGAAAAGAGTCA	GACATC-GAAACATACATAA--GTTGATATTC-CTTTGATATCG-----ACGACTA
Spar	AAAAAATGAAAAATTCATGAGAAAAGAGTCA	GACATC-GAAACATACATAA--ATTGATATTC-CTTTAGCTTTT----AAAGACTA
Smik	GAAAAACGAAAAATTCATG-GAAAAGAGTCA	ACCGTC-GAAACATACATAA--ACCGATATTT-CTTTAGCTTTTCGACAAAAATCTG
Sbay	GAAAAATAAAAAGTGATTG-GAAAAGAGTCA	GATCTCCAAAACATACATAATAACAGGTTTTTACATTAGCTTTT----GAAAACATA

$$F_{n-l} \quad P(S_{[n-l,l]} | w, T)$$

	F_{n-l}	$P(S_{[n-l,l]} w, T)$
Scer	AAAAAATGAAAAATTCATGAGAA	AAGAGTCAGACATC-GAAACATACATAA--GTTGATATTC-CTTTGATATCG-----ACGACTA
Spar	AAAAAATGAAAAATTCATGAGAA	AAGAGTCAGACATC-GAAACATACATAA--ATTGATATTC-CTTTAGCTTTT----AAAGACTA
Smik	GAAAAACGAAAAATTCATG-GAA	AAGAGTCAACCGTC-GAAACATACATAA--ACCGATATTT-CTTTAGCTTTTCGACAAAAATCTG
Sbay	GAAAAATAAAAAGTGATTG-GAA	AAGAGTCAGATCTCCAAAACATACATAATAACAGGTTTTTACATTAGCTTTT----GAAAACATA

WM probs single sequences

$$P(s | w) > P(s | b)$$



S. par TGCTATCAGCTACTTCCC

$$P(s | w) < P(s | b)$$



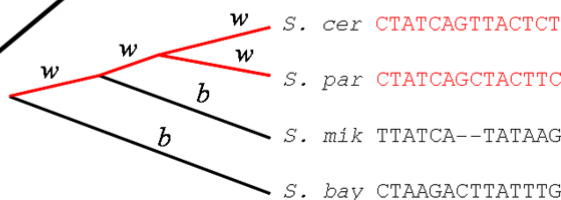
S. bay TGCTAAGACTTATTTGCC

$$S_{[i,l]}$$



<i>S. cer</i>	TA	CTATCAGTTACTCT	TC
<i>S. par</i>	TG	CTATCAGCTACTTC	CC
<i>S. mik</i>	TGTTATCA--	TATAAGTA	
<i>S. bay</i>	TG	CTAAGACTTATTTGCC	

Selection pattern



We sum over all binding site configurations for multiple motifs.

Instead of assuming that the site is maintained in all species, we allow sites to be under selection only in some species.

MotEvo also uses EM to optimize the WMs.

Collections of WMs

Using a combination:

- Known binding sites from the experimental literature.
- Large-scale binding experiments (ChIP-chip and ChIP-seq).
- Co-regulation of gene expression.
- Computational motif prediction.

a large number of WMs has been curated in for different organisms and these are available through a number of databases.

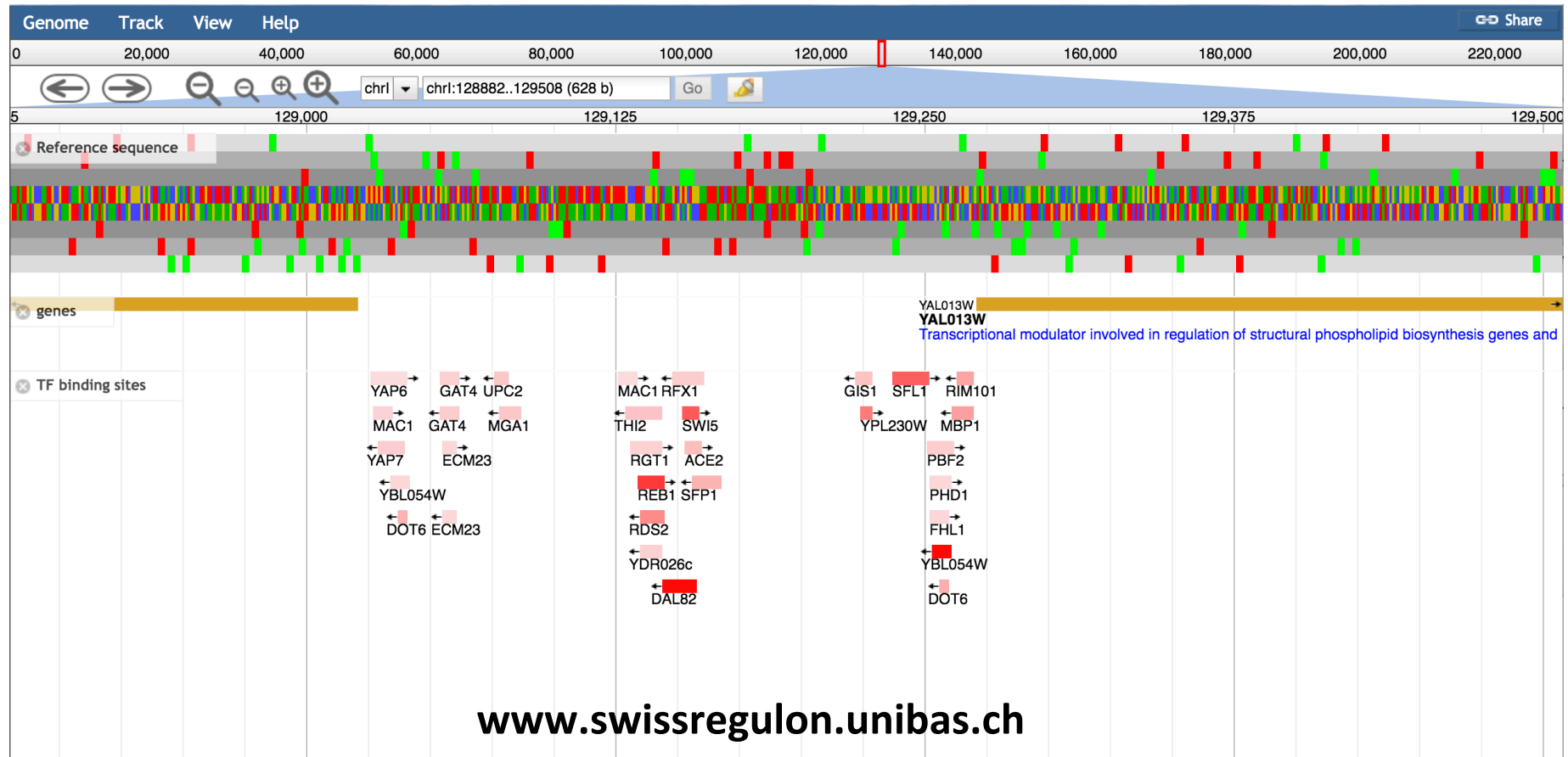
Examples:    *Regulatory Element Database for Drosophila*

- RegulonDB (E. coli) containing WMs for over 100 E. coli TFs (out of about 250).
- Jaspar (several organisms). 500 Vertebrate WMs (of about 2000 TFs).
- RedFly and Jaspar (Drosophila) > 140 WMs (of about 600).

Our collection SwissRegulon (www.swissregulon.unibas.ch) has:

- About 680 mammalian weight matrices (of about 2000 TFs).
- 160 yeast weight matrices (out of about 200 yeast TFs).

Using such WM collections we can predict TF binding site constellations genome-wide.



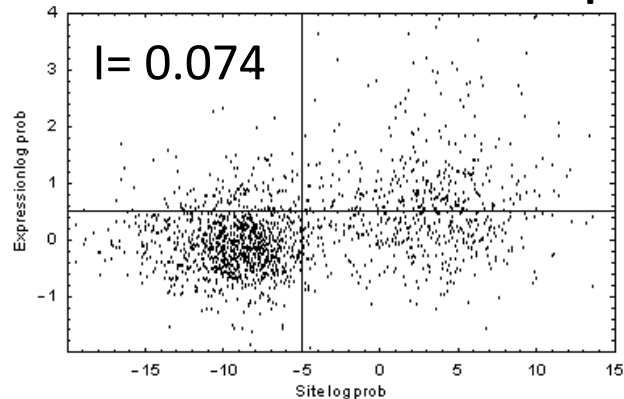
To what extent can we trust such binding site annotations?
Can we validate these experimentally?

Annotation reliability:

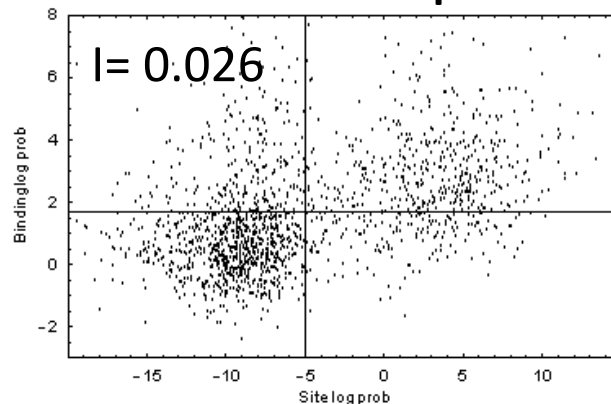
Comparison with high-throughput binding and expression data

- In depth experimental analysis of one yeast TF that targets many genes: ABF1.
 1. Genome-wide measurement of TF binding to using ChIP-on-chip.
 2. Genome-wide measurement of gene expression changes in response to turning of the TF.
- Compare the predictions of which genes are targeted by ABF1 using the mutual information I between each pair of predictions..

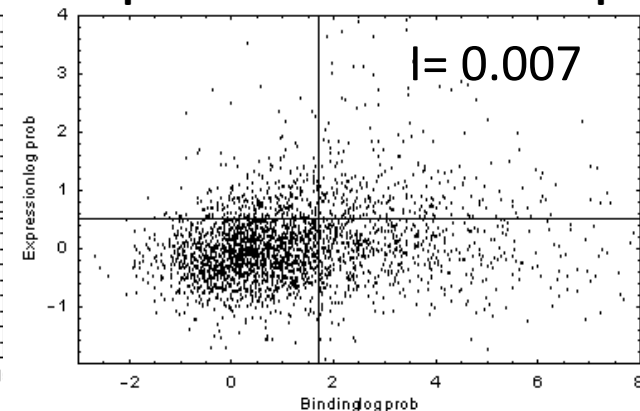
Annotation vs. ChIP-on-chip



Annotation vs. Expression



Expression vs. ChIP-on-chip



The computational annotations correlate much better with both experimental data-sets than the experimental datasets with each other!

Annotation reliability: Comparison with literature sites

SCPD



The Promoter Database of *Saccharomyces cerevisiae*

(Michael Zhang, Cold Spring Harbor)

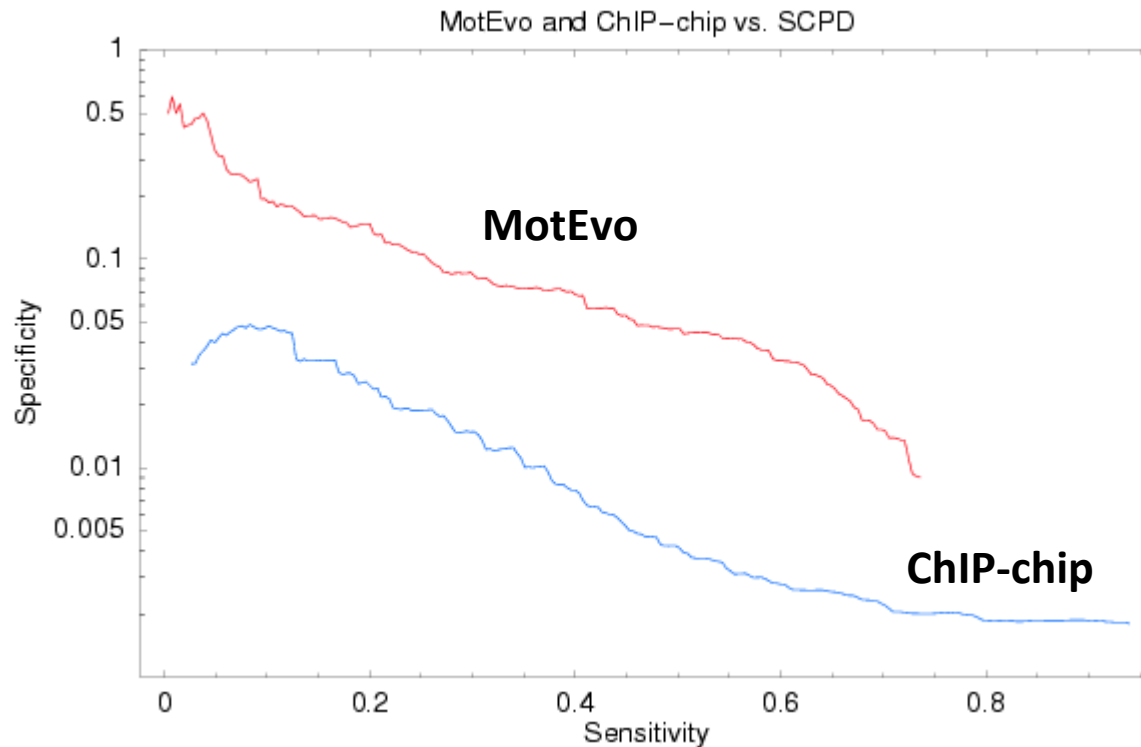
437 known binding sites upstream of 200 cerevisiae genes.

- For each TF with at least 1 known site: determine the probability for each intergenic region to have at least 1 site.
- Using results of large-scale Chip-chip data, determine for each intergenic region the probability that the region is bound (Harbison et al, Nature, 2004).
- Compare the predictions of MotEvo and Chip-chip taking literature sites as reference.

Sensitivity: Fraction of regions with known sites that are predicted to indeed have a site (for the right TF).

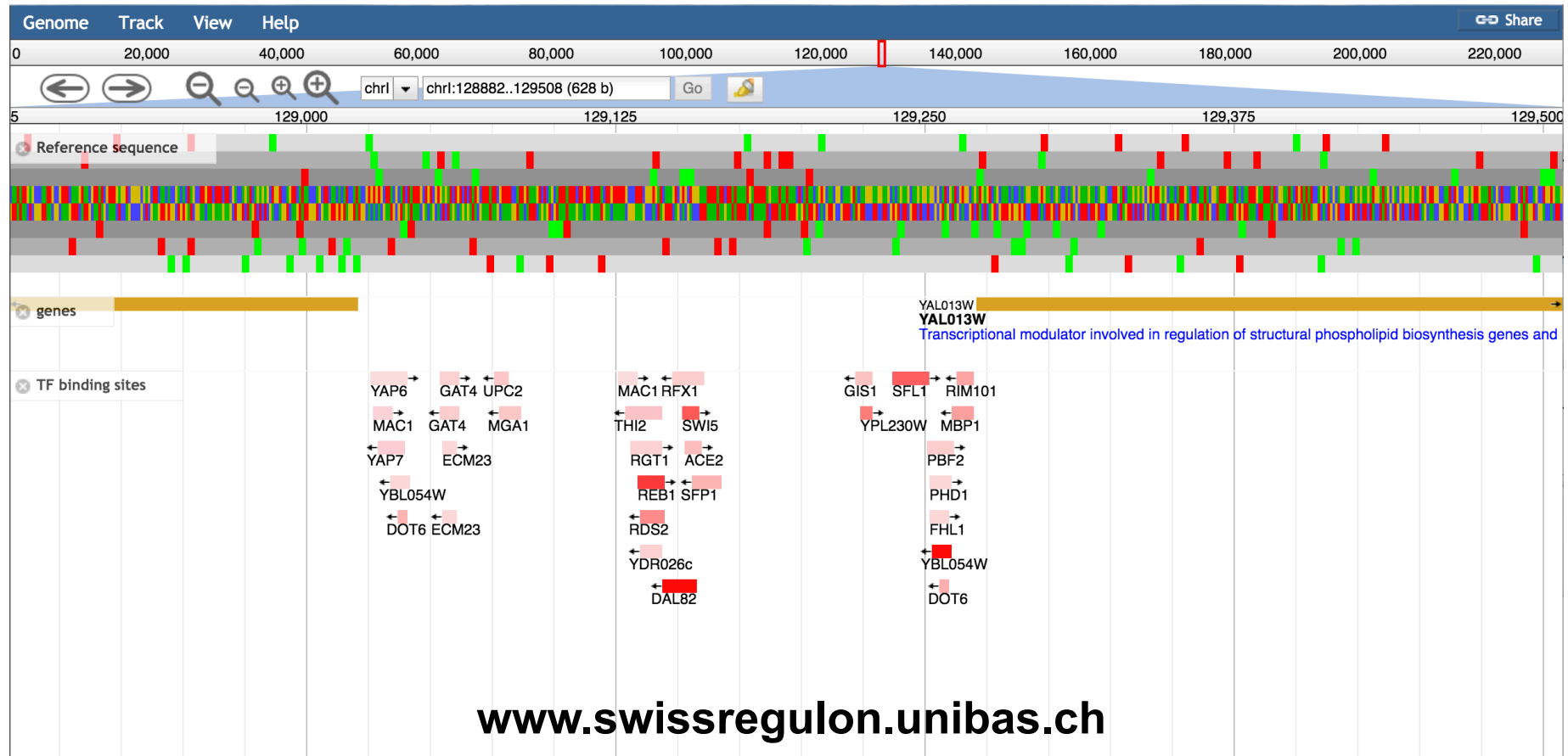
Specificity: Fraction of all regions predicted to contain a site that are found in the literature (for the right TF).

Annotation reliability: Comparison with literature sites



To hit the same number of literature regions ChIP-chip has to predict 8-10 times as many regions.

My interpretation: Computational predictions identify target genes with significantly higher accuracy than high-throughput TF binding data.



Can we use the regulatory site annotations
to understand how regulatory networks functions?

Outline of the lectures

Day 1

1. Computational methods for determining the constellations of regulatory sites.
2. From constellations of regulatory sites to genome-wide gene expression patterns.

Day 2

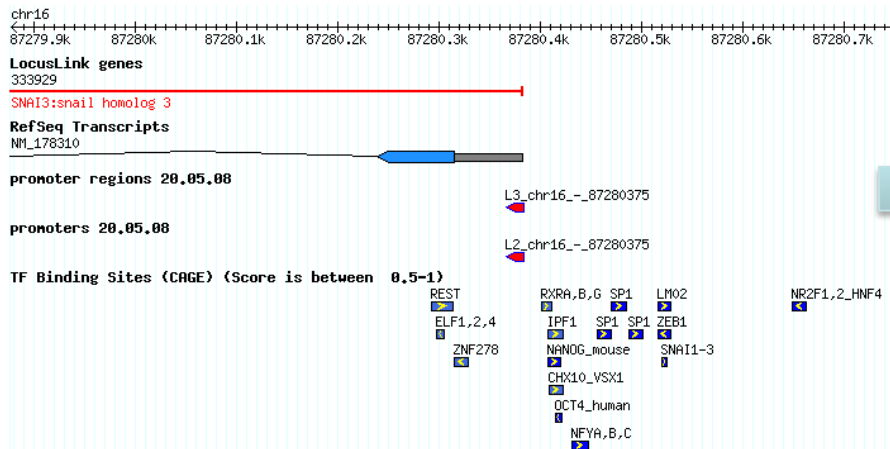
3. Large-scale patterns in genomes and gene regulatory networks.
4. Gene expression noise and its role in the evolution of *de novo* gene regulation.

Day 3

5. *How do bacteria genomes evolve in the wild?*

From constellations of regulatory sites to genome-wide gene expression patterns

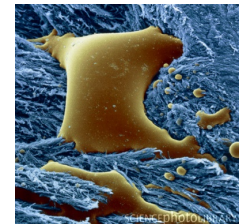
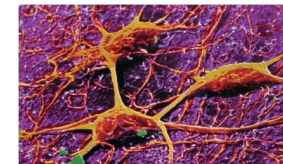
Genome-wide TF binding site predictions



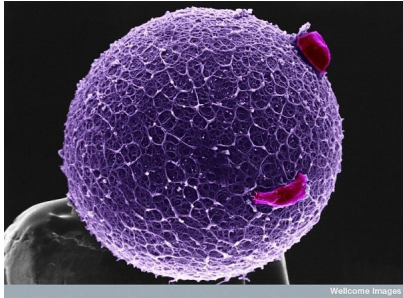
?



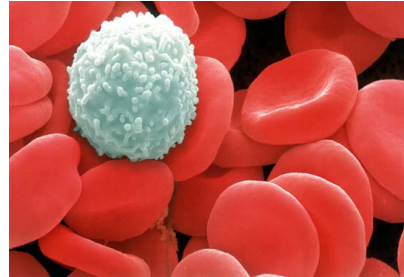
Stable gene expression 'states'.



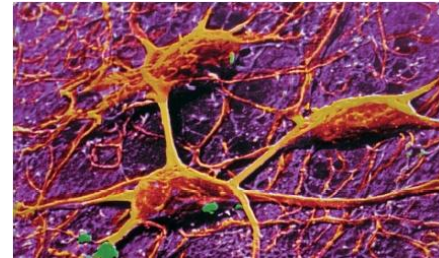
How is the regulatory code in the DNA 'read out' to control cell fate and identity?



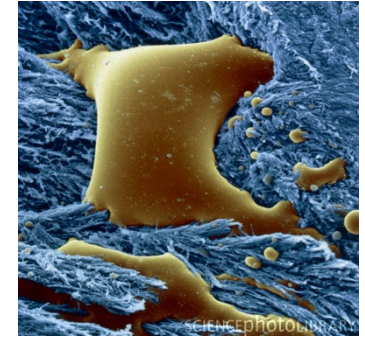
egg cell with 2 coronal cells



white and red blood cells



three neurons



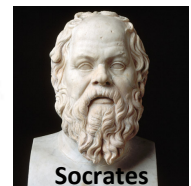
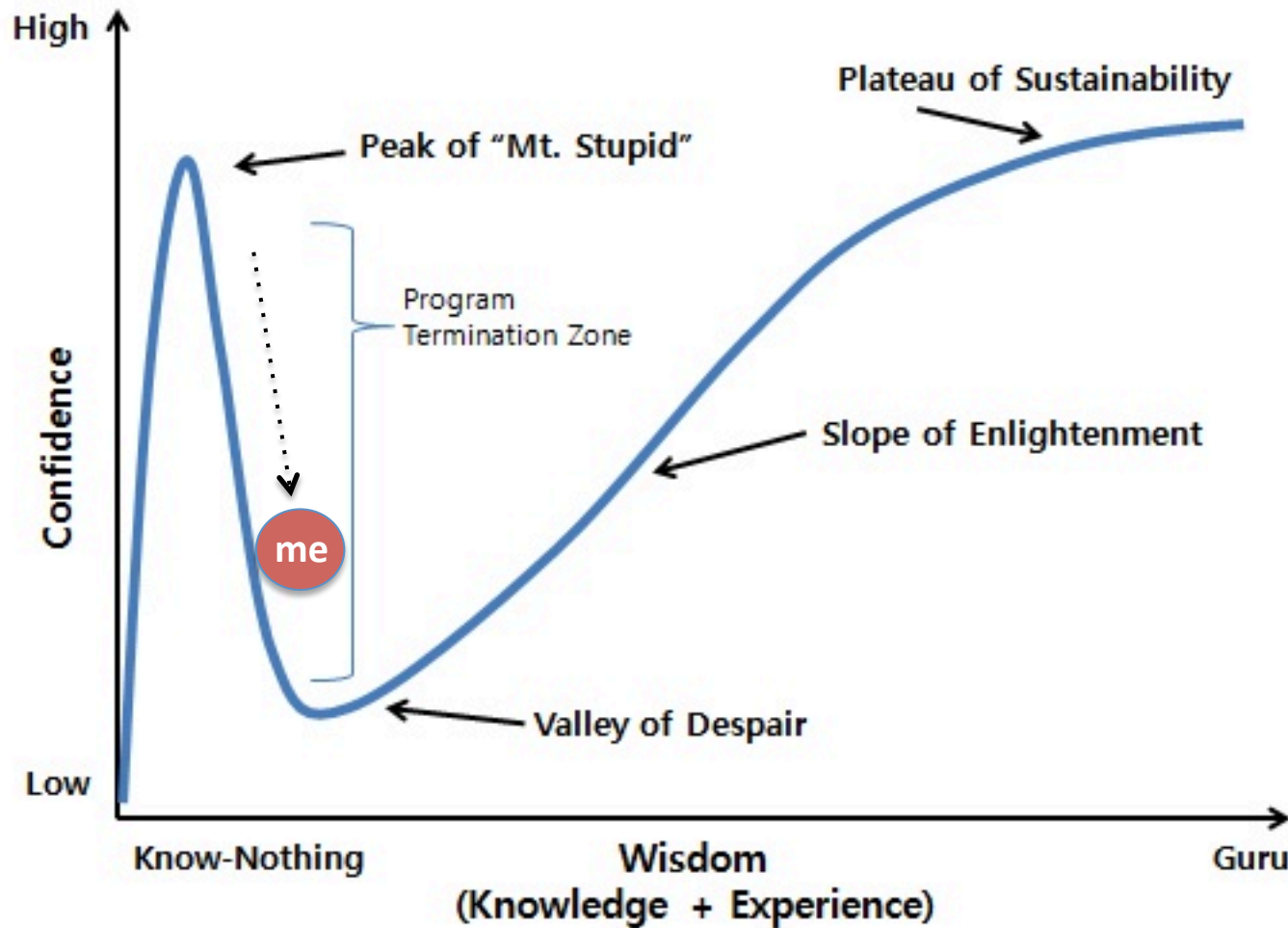
osteoclasts

How do gene regulatory networks function as *systems*?

- What is a cell type?
- How is cell identity stabilized?
- What variables contain the crucial information? What does not matter?



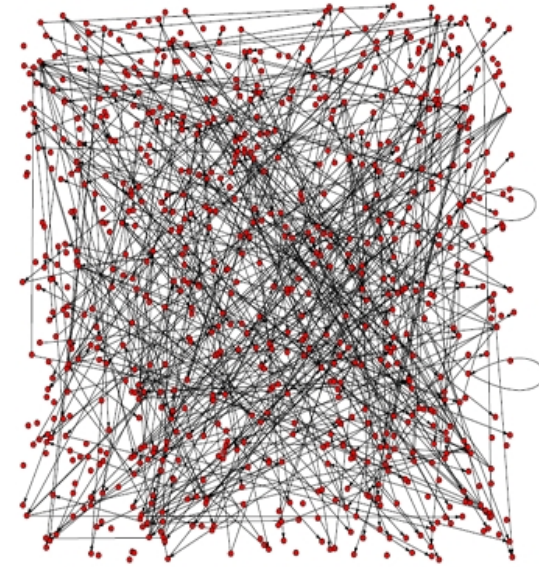
Dunning-Kruger Effect



Detailed models are out of the question

Current knowledge is still very incomplete:

- We know binding specificity for < 750 of ~1500 mammalian TFs.
- Chromatin state regulation mostly not understood.
- mRNA processing, transport, translation and stability are all also regulated.
- Regulator activity depends on post-translational modifications, on interactions with co-factors, localization.
- ‘Grammar’ of regulatory site constellations.
- and so on *and on....*



No use putting everything *we know* into a mathematical model without facing up to the fact that there is much more *we do not know*.

This does not help

How can computational/theoretical analysis make a constructive contribution?

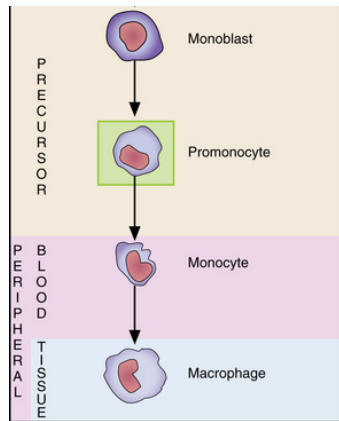
1. Develop methods to help guide experimental efforts.
2. Identify large-scale patterns/invariants to establish overall organizing principles.

What does my high-throughput data say about regulation in my system?

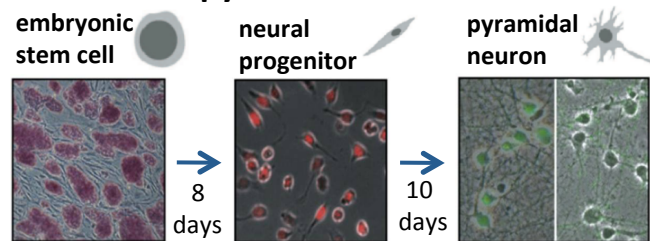
Typical questions:

What are the key regulators? What are their roles? Which pathways do they target?

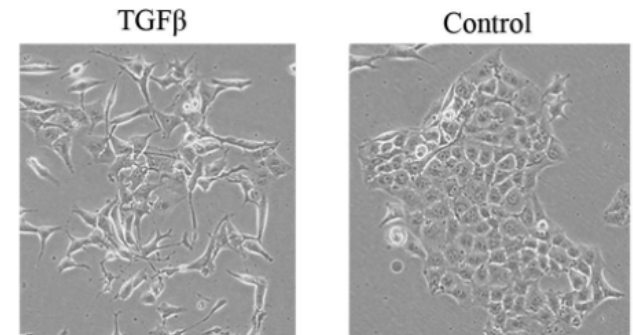
monoblast to macrophage differentiation



Mouse ES cells differentiating into pyramidal neurons



TGF- β induced EMT



Challenges

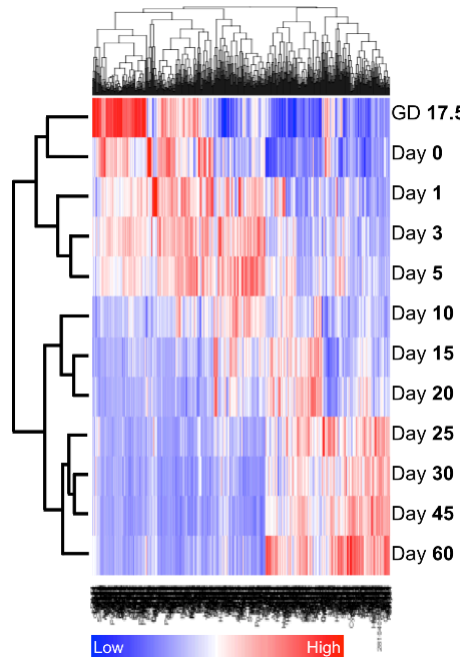
- Cannot do saturating genetic screens (too many candidate TF/miRNA regulators).
- Easy to do high-throughput measurements (microarray, RNA-seq, ChIP-seq).
- *Experimental labs do not have the expertise to analyze such data.*
- Collaborations with dedicated computational labs on a *per case* basis are big investment of time and effort.

Typical analysis of transcriptomic data

Basic processing

- Map raw reads to transcripts.
- Find all genes that are expressed.
- Find genes that are *differently expressed* across pairs of conditions, e.g. *DESeq*.

Clustering genes with similar expression



METHOD | OPEN ACCESS

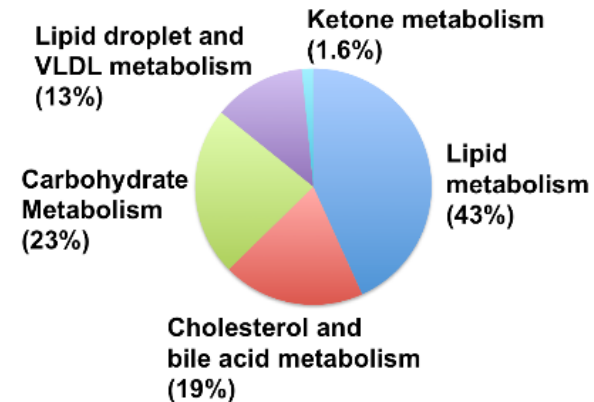
Differential expression analysis for sequence count data

Simon Anders and Wolfgang Huber

Genome Biology 2010 11:R106 | <https://doi.org/10.1186/gb-2010-11-10-r106> | © Anders et al 2010

Received: 20 April 2010 | Accepted: 27 October 2010 | Published: 27 October 2010

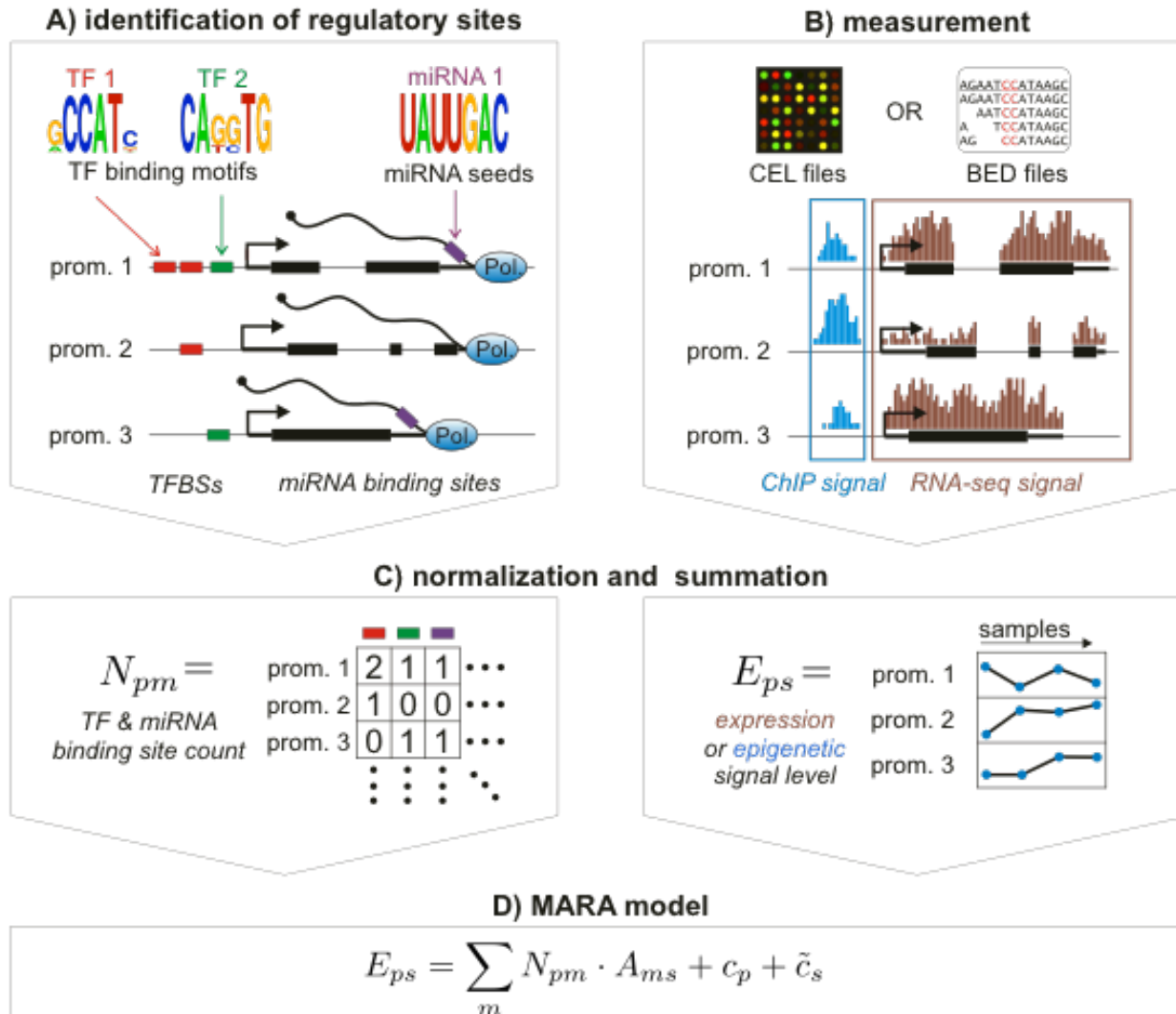
Enriched categories among gene sets



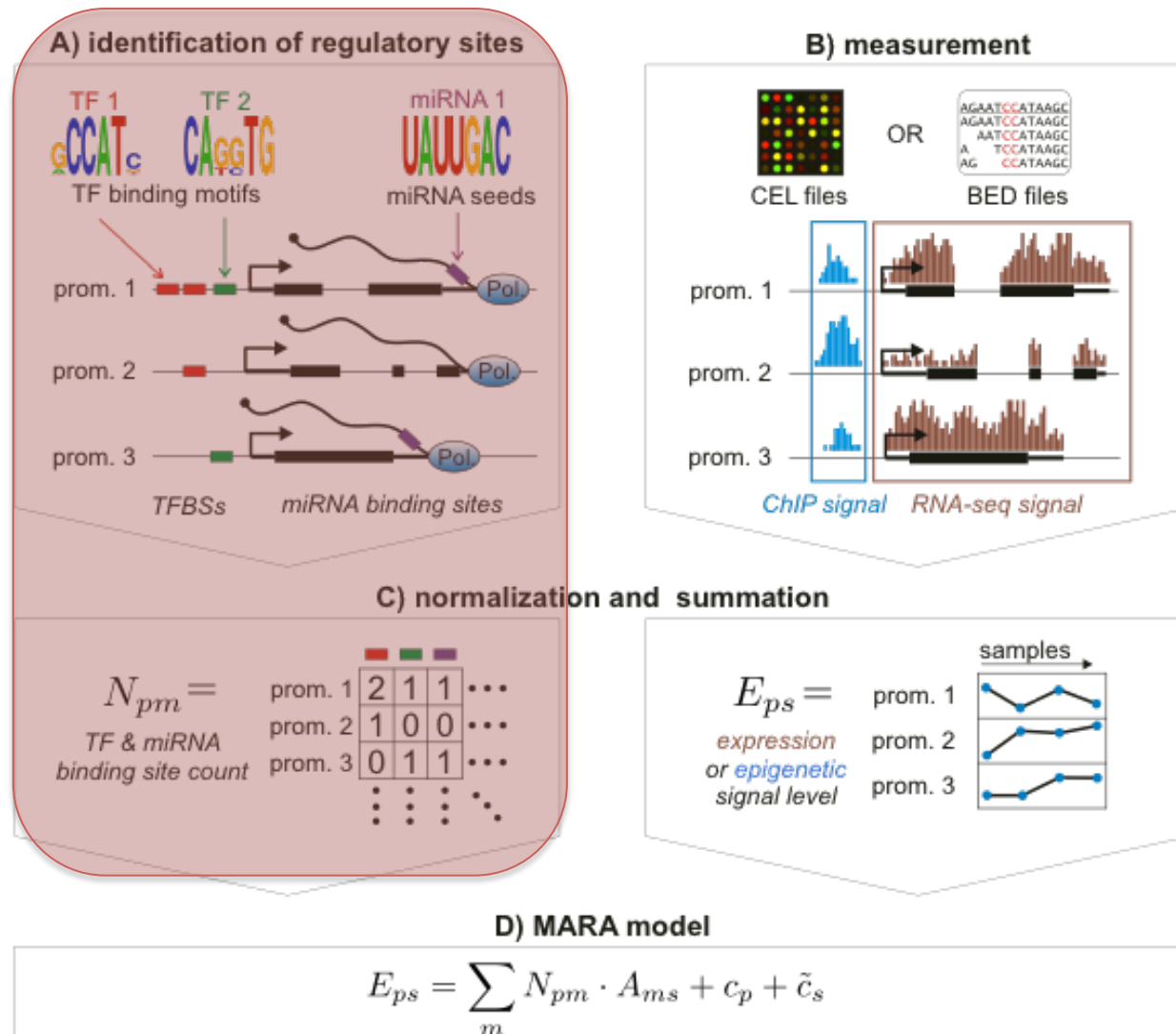
Limitations of these traditional approaches

- Do not make concrete predictions about gene *regulation*.
- Unclear how to experimentally follow-up.

Modelling gene expression and chromatin state in terms of TFBS using a linear model

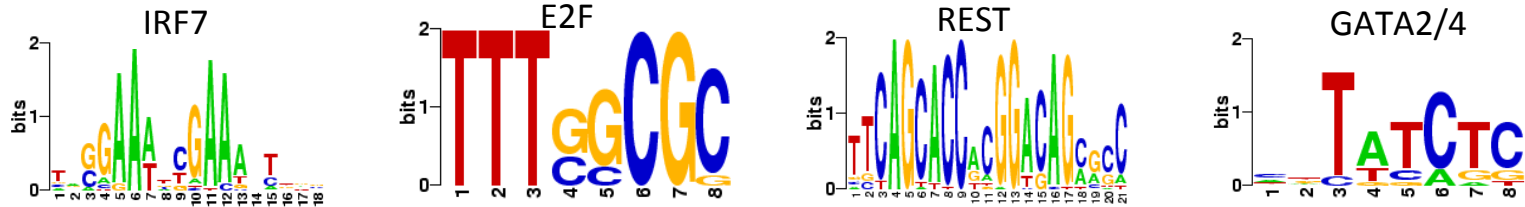


Modelling gene expression and chromatin state in terms of TFBS using a linear model

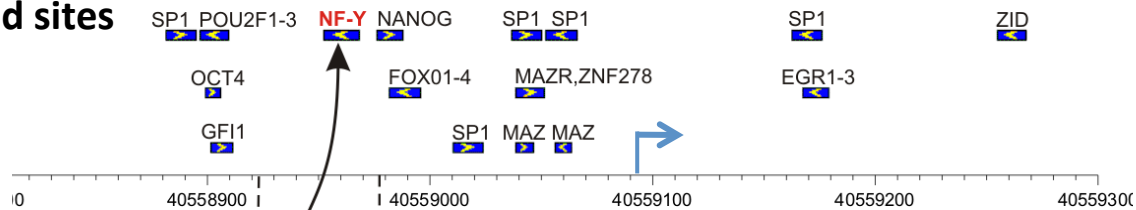


Regulatory site predictions using comparative genomics

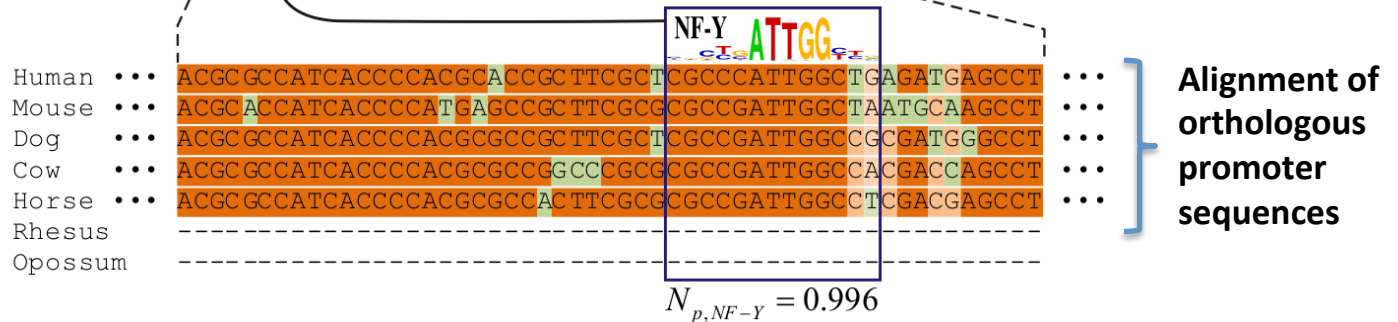
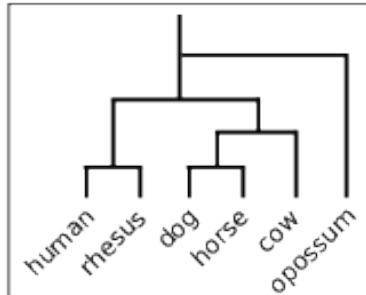
Sequence-specificities of TFs are mathematically represented as *weight matrices*.



Predicted sites



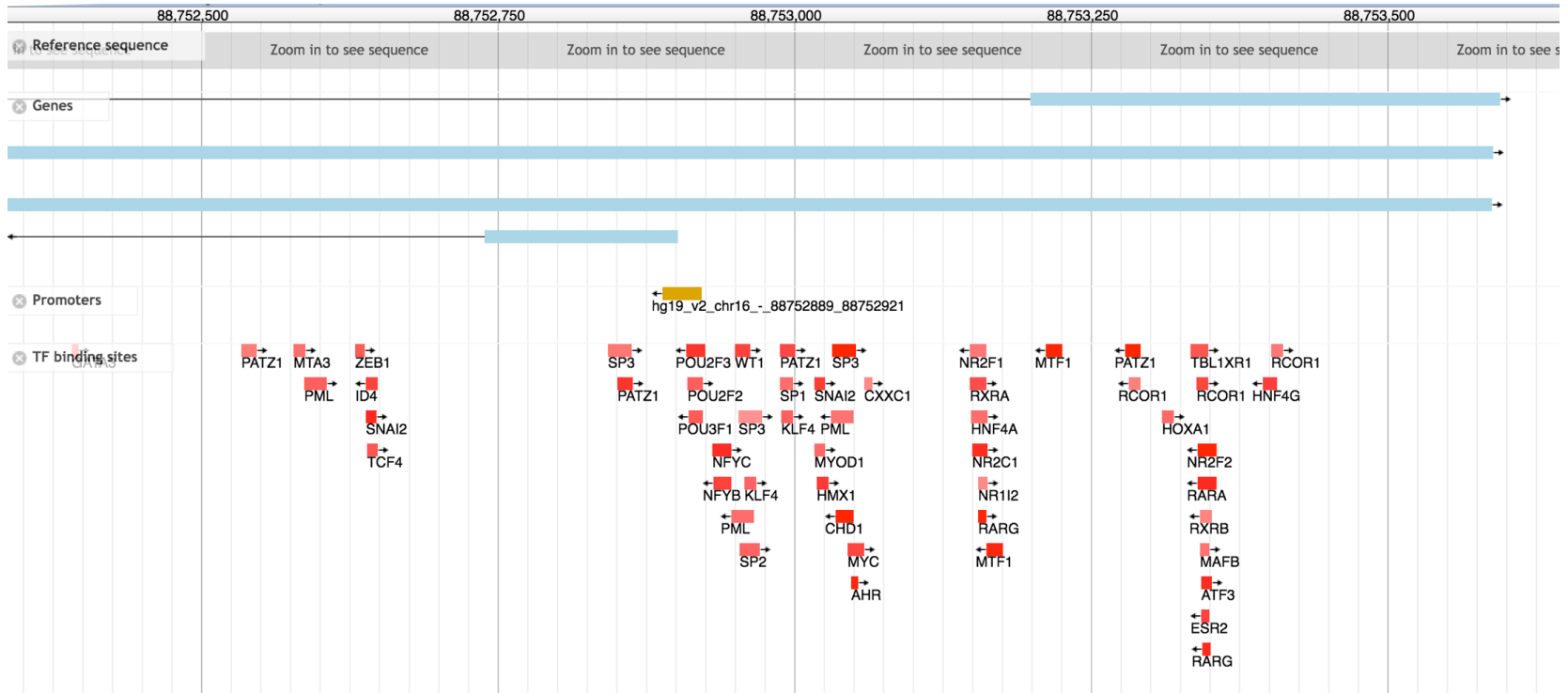
Phylogeny:



- **PROCSE:** van Nimwegen et al, *PNAS* 2002 99(11): 7323-7328
- **Stubb:** Sinha et al, *Bioinformatics* 2003 19 (suppl 1): i292-i301
- **PhyloGibbs:** Siddharthan et al, *PLoS Comp Biol*, 2005 (1)7: e67
- **MotEvo:** Arnold et al, *Bioinformatics*. 2012 Feb 15;28(4):487-94.

Genome-wide annotation of regulatory sites in proximal promoters

Example: Predicted TFBSs in the proximal promoter of the SNAI3 TF.

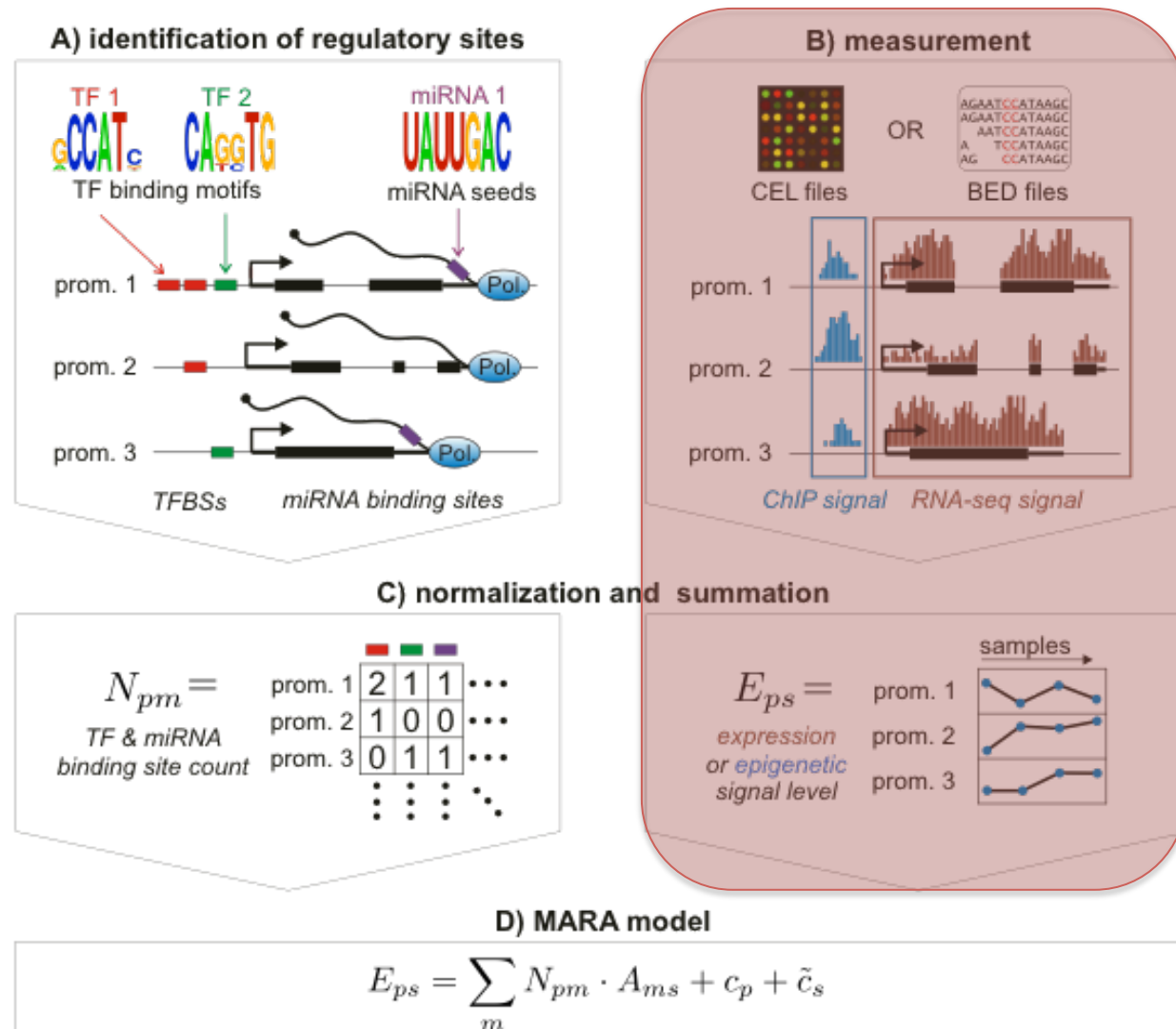


Summarizing the TFBS predictions

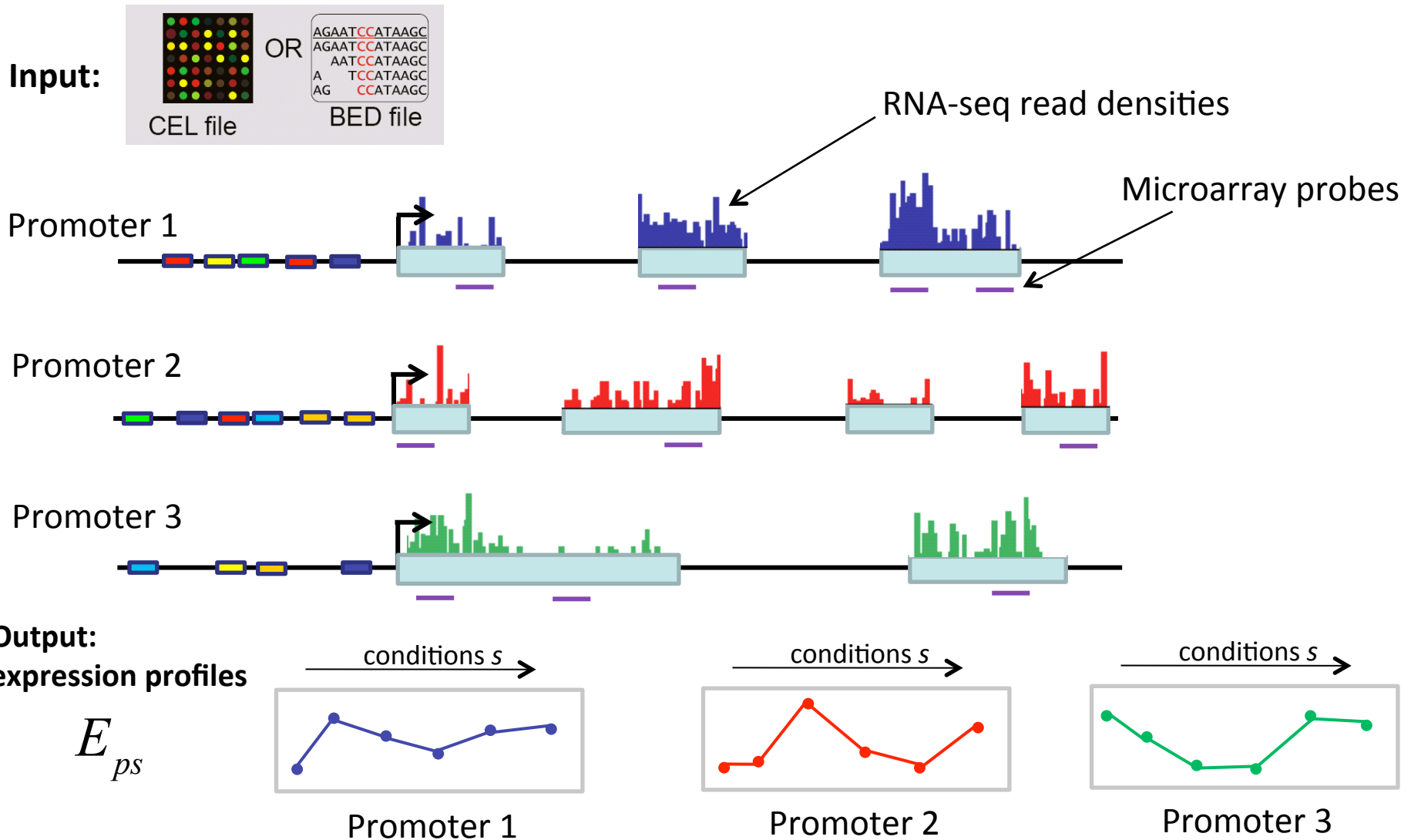
Sum the posteriors of the predicted sites for each motif to obtain a **matrix of site-counts**:

$$N_{pm} = \text{Total number of sites for motif } m \text{ in promoter } p.$$

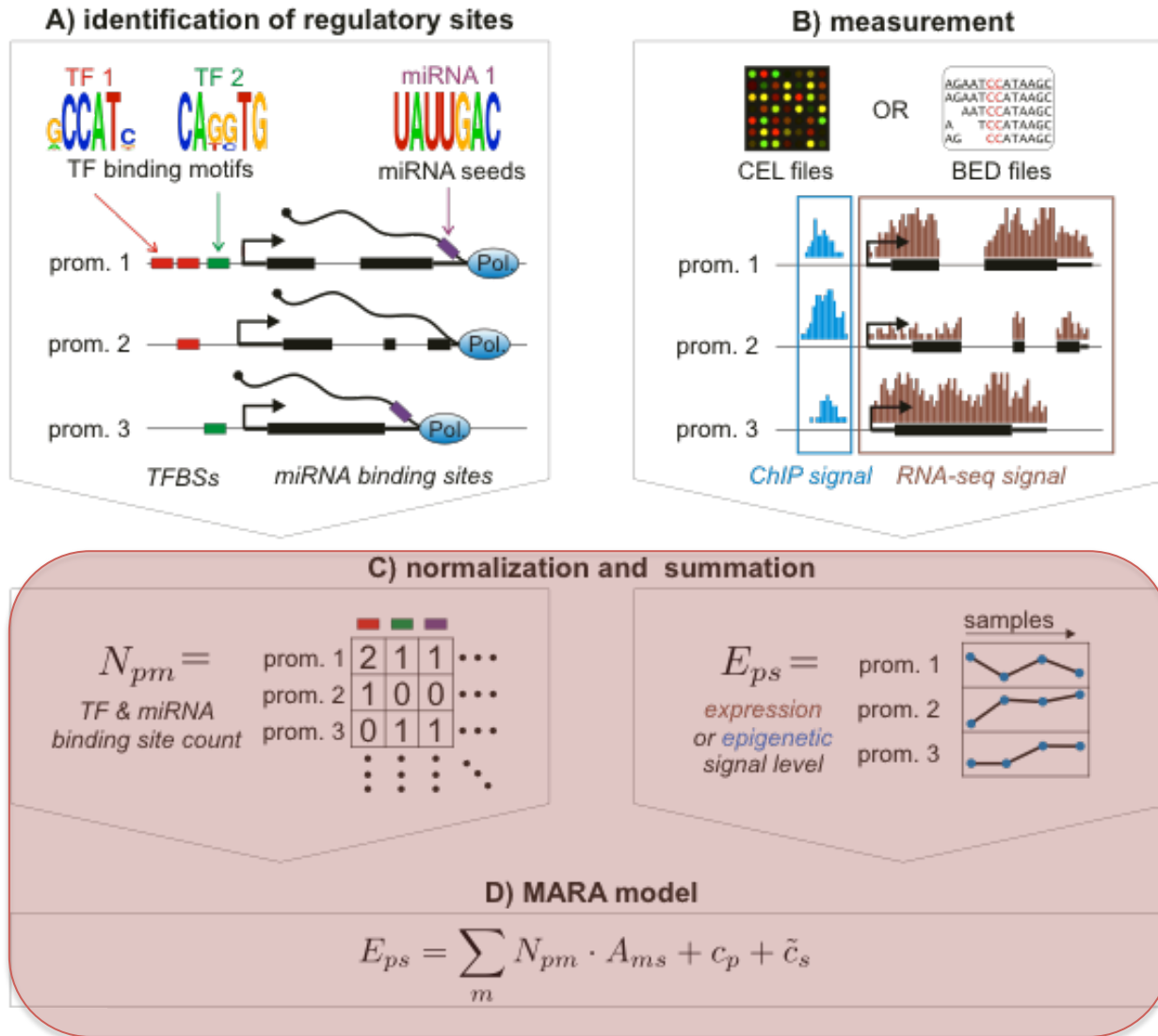
Modeling gene expression and chromatin state in terms of TFBS using a linear model



Quantifying genome-wide expression



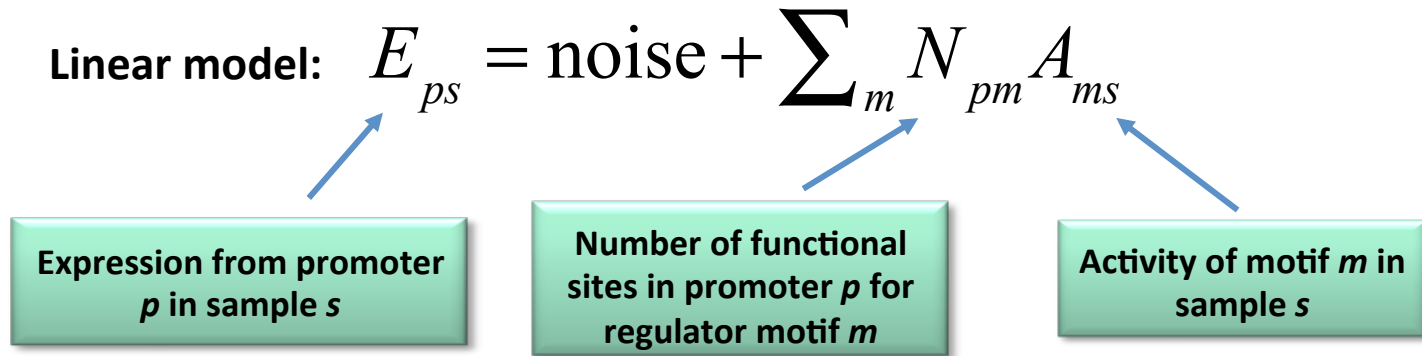
Modeling gene expression and chromatin state in terms of TFBS using a linear model



Forrest et al.
Nat Genet 2009

Balwierz et al.
Genome Res
2014

Fitting MARA's linear model (conceptual)



Bayesian inference of the motif activities

Obtain both best-fit activities and error-bars on the activities:

A_{ms}^* = Fitted activity of motif m in sample s .

δA_{ms} = Error-bar on the activity.

Significance of motif m :

$$z_m = \sqrt{\frac{1}{S} \sum_{s=1}^S \left(\frac{A_{ms}^*}{\delta A_{ms}} \right)^2}$$

Notes

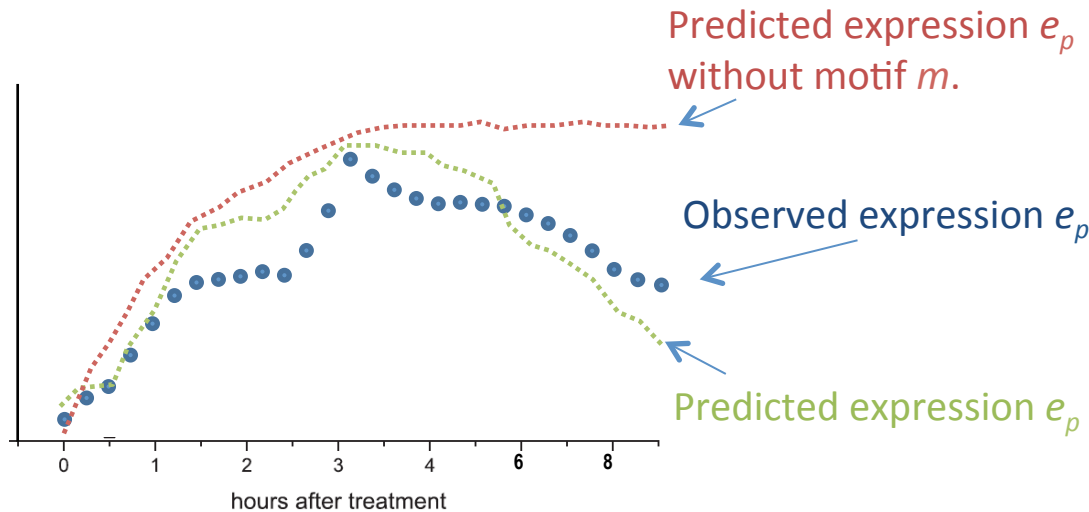
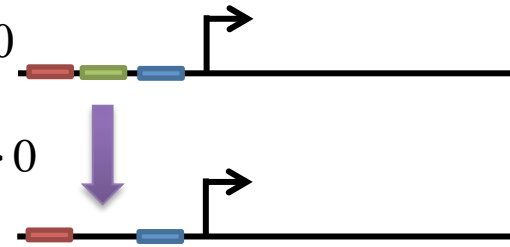
- Motif activities capture the expression *changes* across the input samples.
- Activity meaning: A_{ms}^* is the average amount by which log-expression would go up in sample s when a site for motif m is added to the promoter.
- Significance meaning: z_m is the average number of standard-deviations that the activity of motif m is away from its average of zero.

Predicting targets of each motif

- For each **motif**, select promoters with predicted sites, i.e with $N_{pm} > 0$
- *Mutate* promoter p to *remove* the binding site(s) for **motif** m : $N_{pm} \rightarrow 0$
- Updated site-count matrix $N \rightarrow \tilde{N}$
- Log-likelihood ratio of fitting *all data* with N versus the mutated \tilde{N} :

$$S_{pm} = \log \left[\frac{\int dAP(E | N, A)}{\int dAP(E | \tilde{N}, A)} \right]$$

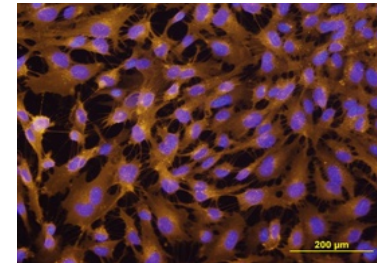
Quantifies the contribution of motif m to explaining the expression pattern of promoter p .



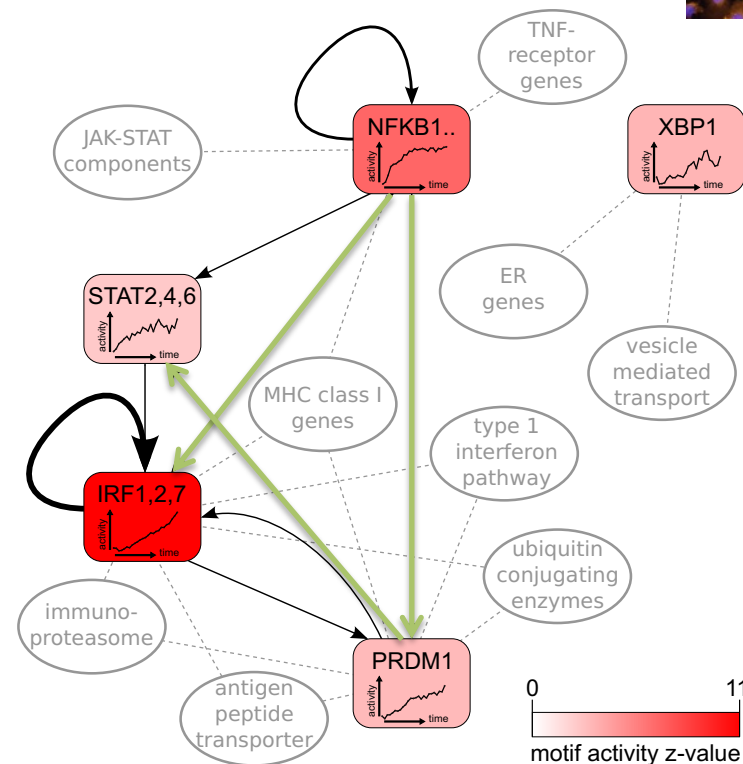
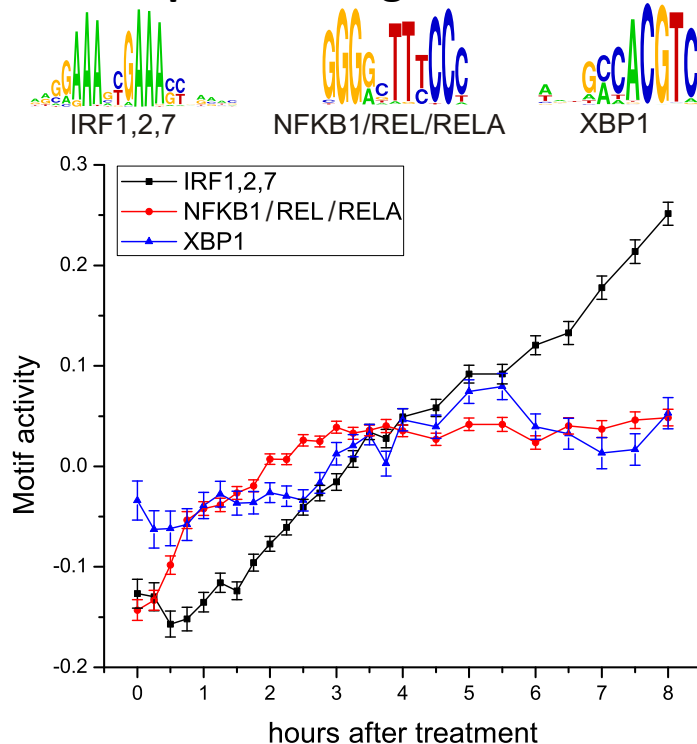
The log-likelihood ratio S_{pm} quantifies how much the quality of the fit is reduced when the sites for motif m in promoter p are removed.

Example: Response of Human umbilical vein endothelial cells to treatment with TNF α

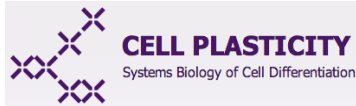
Time course measurements: Wada *et al.* A Wave of nascent transcription on activated human genes. *PNAS* 2009



Top 3 most significant motifs



This works fairly consistently



Polycomb recruitment during mouse ES cell differentiation

- Arnold *et al.* **Genome Research** 23(1):60-73 (2013)

Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting

Epithelial mesenchymal transition

- Tiwari *et al.* **Cancer cell** 23(6) 768-783 (2013)

Sox4 Is a Master Regulator of Epithelial-Mesenchymal Transition by Controlling Ezh2 Expression and Epigenetic Reprogramming.

- Tiwari *et al.* **PloS ONE** 8(2):e57329 (2013)

Klf4 is a transcriptional regulator of genes critical for EMT, including Jnk1 (Mapk8)

T-cell development

- Vigano MA *et al.* **Eur J Immunol** doi: 10.1002/eji.201344022 (2013)

An epigenetic profile of early T-cell development from multipotent progenitors to committed T-cell descendants.

Breast cancer

- Meier-Abt *et al.* **Breast Cancer Research** 15(2):R36 (2013)

Parity induces differentiation and reduces Wnt/Notch signaling ratio and proliferation potential of basal stem/progenitor cells isolated from mouse mammary epithelium.

- Aceto *et al.* **Nature medicine** 18(4), 529-37 (2012)

Tyrosine phosphatase SHP2 promotes breast cancer progression and maintains tumor-initiating cells via activation of key transcription factors and a positive feedback signaling loop.

Hepatitis C treatment

- Dill *et al.* **Journal of Clinical Investigation** in press (2014)
Pegylated interferon-alpha regulates hepatic gene expression by transient activation of the Jak-STAT pathway

Dirk Schubeler
FMI



Gerhard Christofori
DBM Uni Basel



Ton Rolink
DBM Uni Basel



Mohamed Bentires-Alj
FMI



Markus Heim
Uni Hospital Basel



Swiss Institute of
Bioinformatics

This works fairly consistently

Co-factor PGC-1 α

- Baresic *et al.* **Mol Cell Biol** revision (2014)

Transcriptional network analysis in muscle reveals AP-1 as a partner of PGC-1 α in the regulation of the hypoxic gene program.

- Eisele *et al.* **J. Biol. Chem.**, 288(9):6589 (2013)

The peroxisome proliferator-activated receptor coactivator 1/(PGC-1) coactivators repress the transcriptional activity of NF κ -B in skeletal muscle cells

- Perez-Schindler *et al.* **Mol Cell Biol** 32(24):4913-4924 (2012)

The corepressor NCoR1 antagonizes PGC-1 α and estrogen-related receptor α in the regulation of skeletal muscle function and oxidative metabolism.



Christoph Handschin
Biozentrum

Non-genotoxic tumorigenesis

- Luisier *et al.* **Nucleic Acids Res** gtk1415 (2014)

Computational modeling identifies key gene regulatory interactions underlying phenobarbital-mediated tumor promotion.



Raphaëlle Luisier
Novartis/UniBas



Remi Terranova
Novartis

Other

- Summers *et al.* **Eur. J. Hum. Genet.**, 18(11):1209-1215 (2010)

Co-expression of FBN1 with mesenchyme-specific genes in mouse cell lines: implications for phenotypic variability in Marfan syndrome

- Arner *et al.* **Diabetes** 61(8):1986-1993 (2012)

Adipose Tissue MicroRNAs as Regulators of CCL2 Production in Human Obesity.

- Suzuki *et al.* **PLoS ONE** 7(3):e33474 (2012)

Reconstruction of monocyte transcriptional regulatory network accompanies monocytic functions in human fibroblasts.

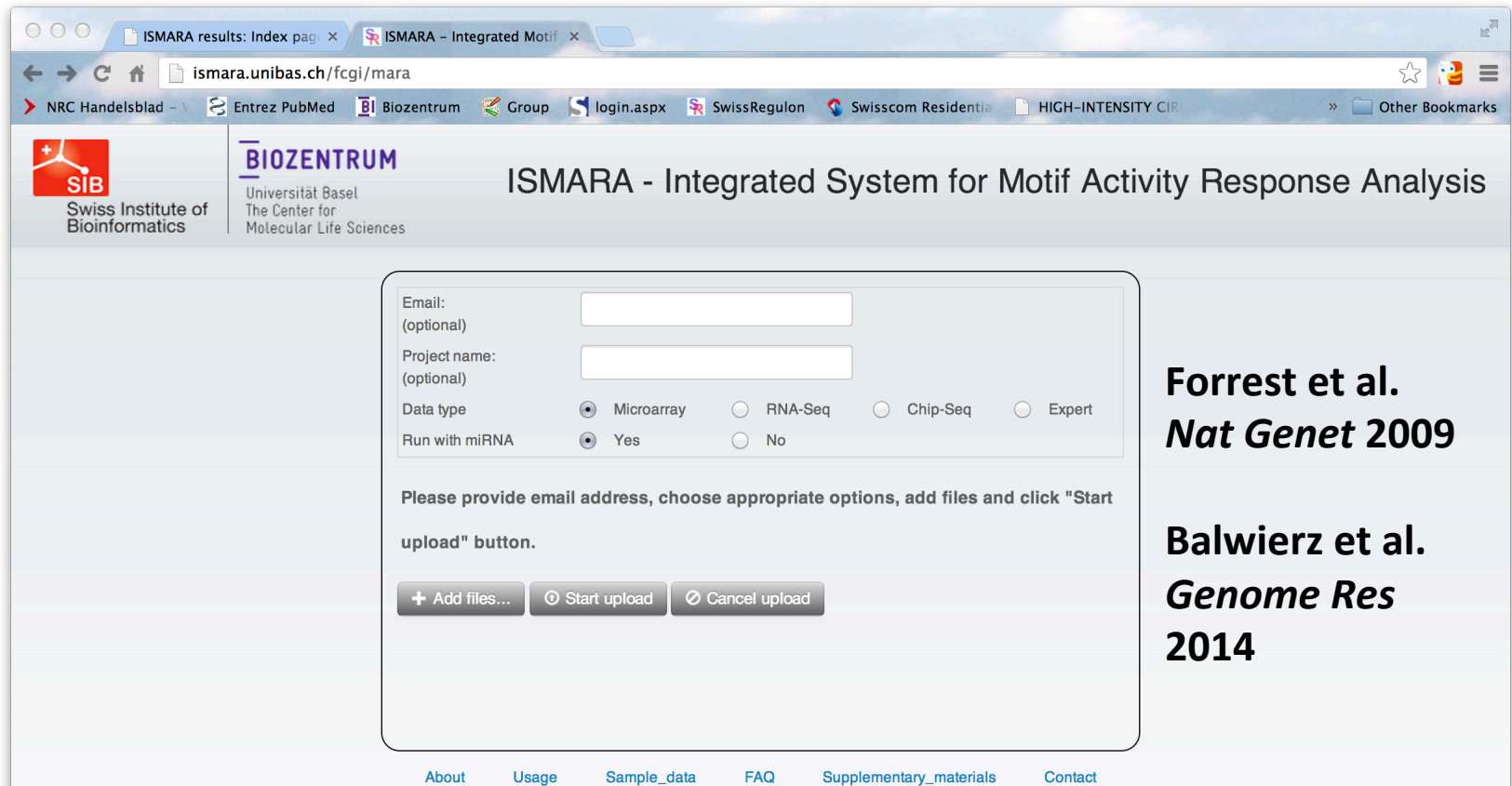
- Hasegawa *et al.* **Exp. Cell Res.** 319(3):68-76 (2013)

Identification of ZNF395 as a novel modulator of adipogenesis.

- Vervoort *et al.* **PLoS ONE** 8(1):e53238 (2013)

SOX4 mediates TGF-beta induced expression of mesenchymal markers during mammary cell epithelial to mesenchymal transition.

Completely automated prediction of regulatory interactions from high-throughput data



The screenshot shows the ISMARA web interface in a browser. The browser tabs include 'ISMARA results: Index page' and 'ISMARA - Integrated Motif'. The address bar shows 'ismara.unibas.ch/fcgi/mara'. The browser's bookmark bar contains links to 'NRC Handelsblad', 'Entrez PubMed', 'Biozentrum', 'Group', 'login.aspx', 'SwissRegulon', 'Swisscom Residential', and 'HIGH-INTENSITY CIR'. The page header features the SIB logo (Swiss Institute of Bioinformatics) and the BIOZENTRUM logo (Universität Basel, The Center for Molecular Life Sciences). The main title is 'ISMARA - Integrated System for Motif Activity Response Analysis'. The central form contains the following fields and options:

- Email: (optional) [text input]
- Project name: (optional) [text input]
- Data type: ☒ Microarray, ☐ RNA-Seq, ☐ Chip-Seq, ☐ Expert
- Run with miRNA: ☒ Yes, ☐ No

Below the form, a message states: 'Please provide email address, choose appropriate options, add files and click "Start upload" button.' At the bottom of the form are three buttons: '+ Add files...', 'Start upload', and 'Cancel upload'. The footer of the page contains links: 'About', 'Usage', 'Sample_data', 'FAQ', 'Supplementary_materials', and 'Contact'.

Forrest et al.
***Nat Genet* 2009**

Balwierz et al.
Genome Res
2014

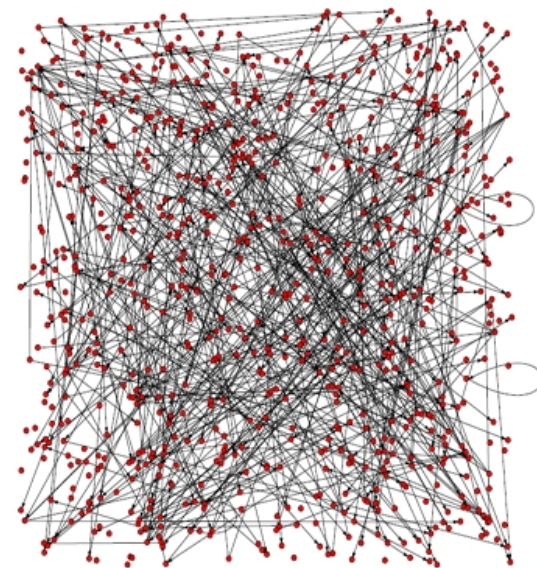
Upload micro-array, RNA-seq, or ChIP-seq data and predict:

- Key regulators (TFs/miRNAs) in the system.
- Regulator activities across the input samples.
- Sets of target genes and pathways for each regulator.
- The regulatory sites on the genome through which the regulators acts.
- Interactions between the regulators.

Detailed models are out of the question

Current knowledge is still very incomplete:

- We know binding specificity for < 750 of ~1500 mammalian TFs.
- Chromatin state regulation mostly not understood.
- mRNA processing, transport, translation and stability are all also regulated.
- Regulator activity depends on post-translational modifications, on interactions with co-factors, localization.
- `Grammar' of regulatory site constellations.
- and so on *and on....*



No use putting everything *we know* into a mathematical model without facing up to the fact that there is much more *we do not know*.

This does not help

How can computational/theoretical analysis make a constructive contribution?

1. Develop methods to help guide experimental efforts.
2. Identify large-scale patterns/invariants to establish overall organizing principles.

Outline of the lectures

Day 1

1. Computational methods for determining the constellations of regulatory sites.
2. From constellations of regulatory sites to genome-wide gene expression patterns.

Day 2

3. Large-scale patterns in genomes and gene regulatory networks.
4. Gene expression noise and its role in the *de novo* evolution of gene regulation.

Day 3

5. *How do bacterial genomes evolve in the wild?*

Why is there still almost no predictive quantitative evolutionary theory?

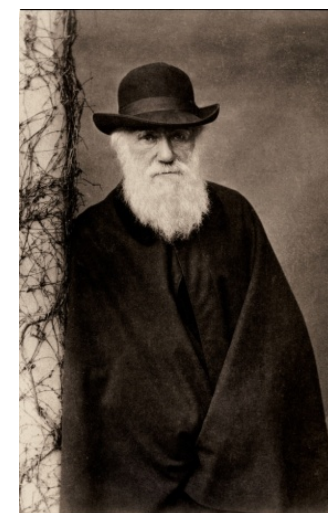
Feynman on how physics works (my paraphrase):

1. We observe phenomena that interest us.
2. We formalize the phenomena by rigorously defined measurable quantities.
3. We search for specific quantitative relationships, i.e. 'laws' that connect the measurable quantities.
4. The glory of physics is when we find a theory that, in one full sweep, explains many of these relationships and makes them self-evident.



The original sins of mathematical population genetics

- No collection of laws about measurable quantities in evolving populations in nature.
- Not even clear what measurable quantities should be analyzed to look for such 'laws'.
- Took a qualitative framework and just started making up toy models.
- **The key quantities in mathematical population genetics models are impossible to measure, e.g:**
 - Fitness
 - Effective population size



Why is there still almost no predictive quantitative evolutionary theory?

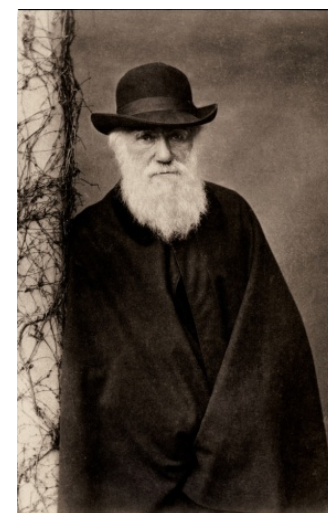
Feynman on how physics works (my paraphrase):

1. We observe phenomena that interest us.
2. We formalize the phenomena by rigorously defining measurable quantities.
3. We search for specific quantitative relationships, i.e. 'laws' that connect the measurable quantities.
4. The glory of physics is when we find a theory that, in one full sweep, explains many of these relationships and makes them self-evident.

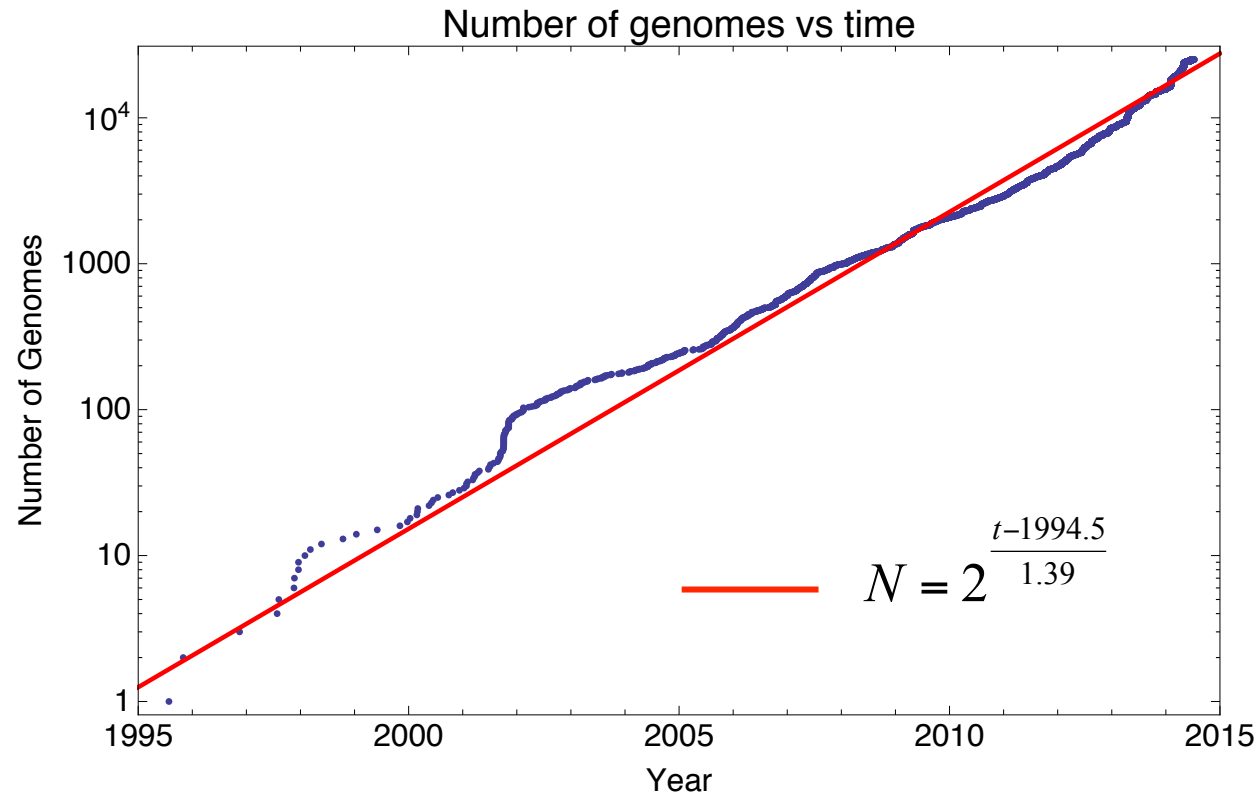


The original sins of mathematical population genetics

- No collection of laws about measurable quantities in evolving populations in nature.
- Not even clear what measurable quantities should be analyzed to look for such 'laws'.
- Took a qualitative framework and just started making up toy models.
- The key quantities in mathematical population genetics models are impossible to measure:
 - Fitness
 - Effective population size



The number of prokaryotic genomes has been doubling every 16-18 months



- First complete genome: *Haemophilus influenzae* in 1995.
- 100 genomes around the year 2002.
- Over 20'000 (virtually) complete genomes in 2015.

[illegible]

Accomplishments of Bioinformatics

Given only the DNA sequence of an organism computational tools are able to:

- Find all the genes in the genome.
- Find out the amino-acid sequence that each gene codes for.
- Find out what other genes in other genomes the genes are evolutionary related to.
- Find out what the *function* of the protein is (at least roughly).
- Find out which parts of the protein are responsible for different aspects of its function.
- Often know what 3-dimensional structure the protein has and know which amino-acids occur where in this structure.

All of this is ultimately based on pairwise and multiple alignment of sequences.

NCBI

Microbial Genomes

HOME

SEARCH

SITE MAP

Genome Project

Genome

Prokaryotic Projects

Collaborators

gMap

ProtMap

TaxPlot

BLAST

FTP

Contact us

Microbial Genomes Resources presents public data from prokaryotic genome sequencing projects. The sequence collection contains data from finished genomes as well as draft assemblies.

Microbial Genome Annotation Tools: We are pleased to announce the availability of [GeneMark](#) and [Glimmer](#), gene prediction tools for microbial genome annotation.

Genome Annotation Pipeline: NCBI has developed a pipeline for annotation of prokaryotic genomes. This service is available to all users by request. If interested, please send an email to [NCBI Genomes](#).

NEW Submission Check Tool: Microbial genome submission check is for the validation of genome submissions to Genbank.

The Concise BLAST database allows for faster calculation times and a broader taxonomic view by eliminating similar proteins within a genus.

Prokaryotes are the earliest forms of life, appearing on earth 4 billion years ago. During the course of their evolution they have extensively altered the biology and chemistry of our planet. More advanced organisms developed as once free-living bacteria took up symbiotic residence inside other cells. These organisms eventually became the organelles found in modern eukaryotes. Energy-producing mitochondria and chloroplasts are examples of organelles in eukaryotic cells.

The **Prokaryotes** include the **Archaea**, which include inhabitants of some of the most extreme environments on the planet, and the **Bacteria**, which include both important pathogens and producers of fermented food, antibiotics, and vitamins.

Genomes

Genome Projects

Prokaryotic Projects

Microbial Genomes

Home

Complete Genomes

Draft Assemblies

Registered

Entrez Genome

Submit a Genome

Sequin

Submission Guide

Register a Project

Submit a Genome

Submit Traces

Tools

Resources

Sequencing Centers

Collaborators

Statistics

Browse Genomes

Database: Genomes Organisms: Bacteria Completeness: All Collapsing level: Limit by class

class:Actinobacteria 732 sequence(s)

class:Bacilli 504 sequence(s)

class:Clostridia 390 sequence(s)

class:Negativicutes 45 sequence(s)

class:Erysipelotrichia 28 sequence(s)

species:Firmicutes bacterium JGI 0000112-J22 1 sequence(s)

species:Firmicutes bacterium JGI 0000112-M16 1 sequence(s)

species:Firmicutes bacterium JGI 0000112-H15 1 sequence(s)

species:Firmicutes bacterium JGI 0000112-P22 1 sequence(s)

species:Firmicutes bacterium JGI 0000112-L22 1 sequence(s)

species:Firmicutes bacterium ASF500 1 sequence(s)

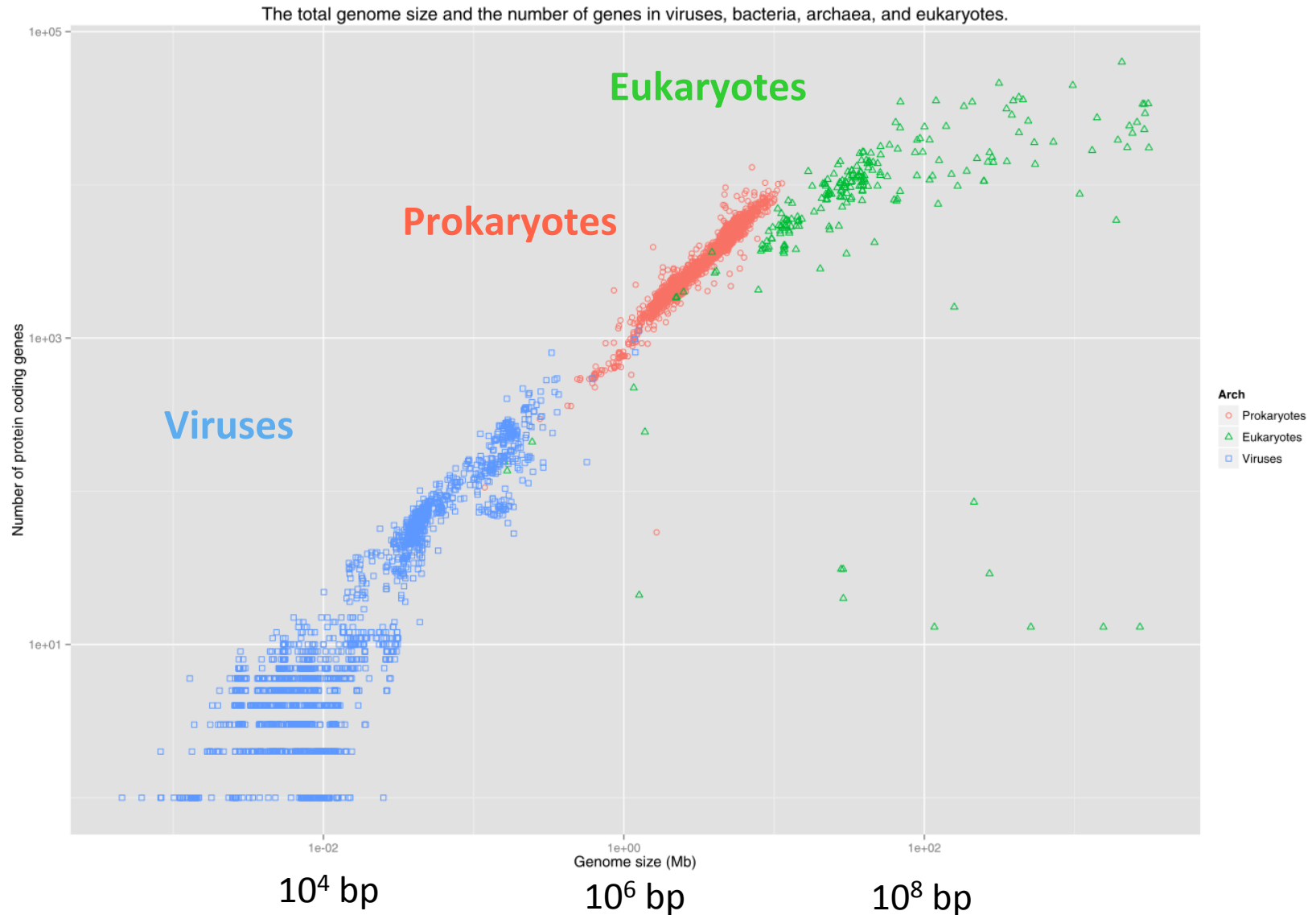
species:Firmicutes bacterium JGI 0000119-P10 1 sequence(s)

species:Firmicutes bacterium JGI 0000119-C08 1 sequence(s)

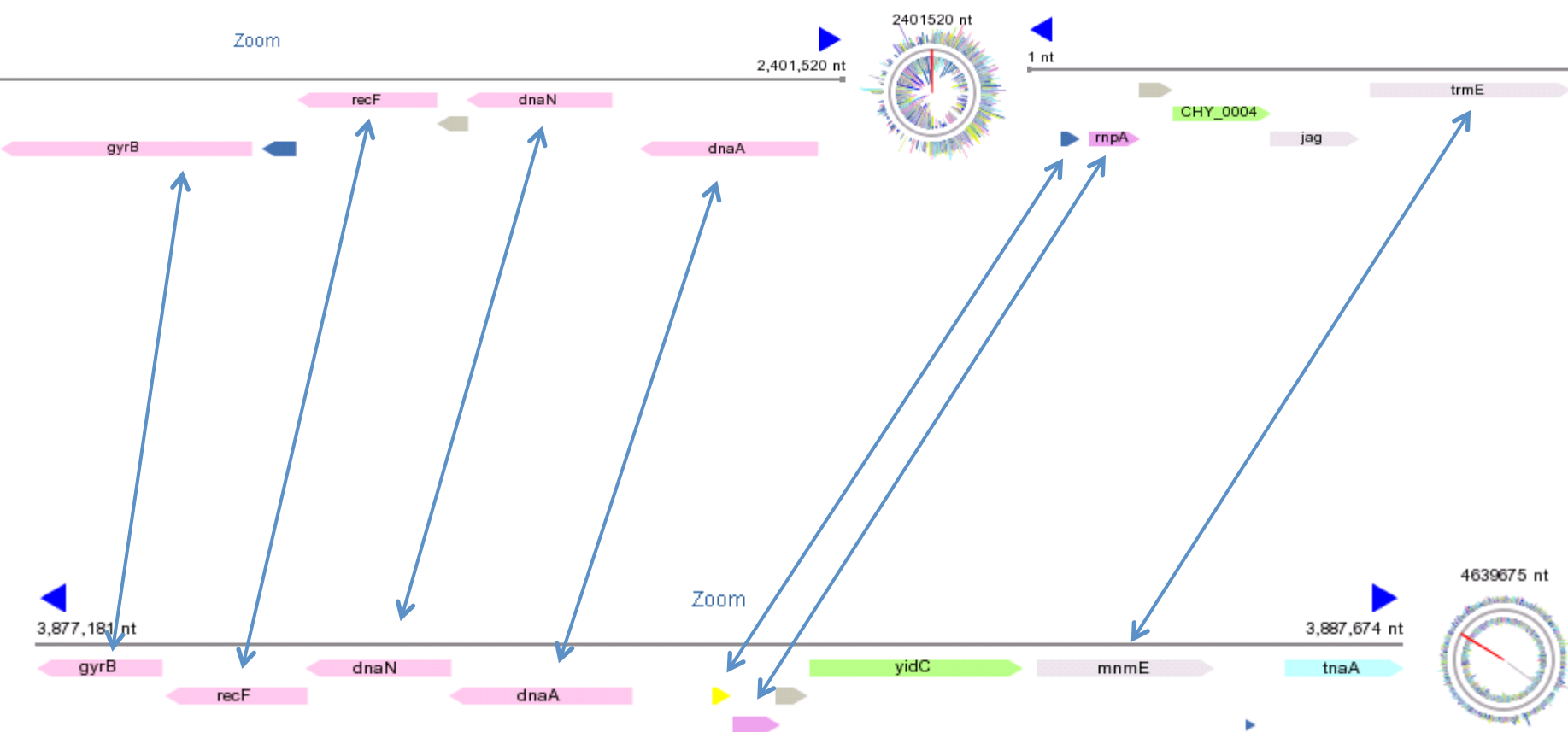
species:Firmicutes bacterium M10-2 1 sequence(s)

class:Alphaproteobacteria 541 sequence(s)

Genome size vs gene number in viruses, prokaryotes, and eukaryotes

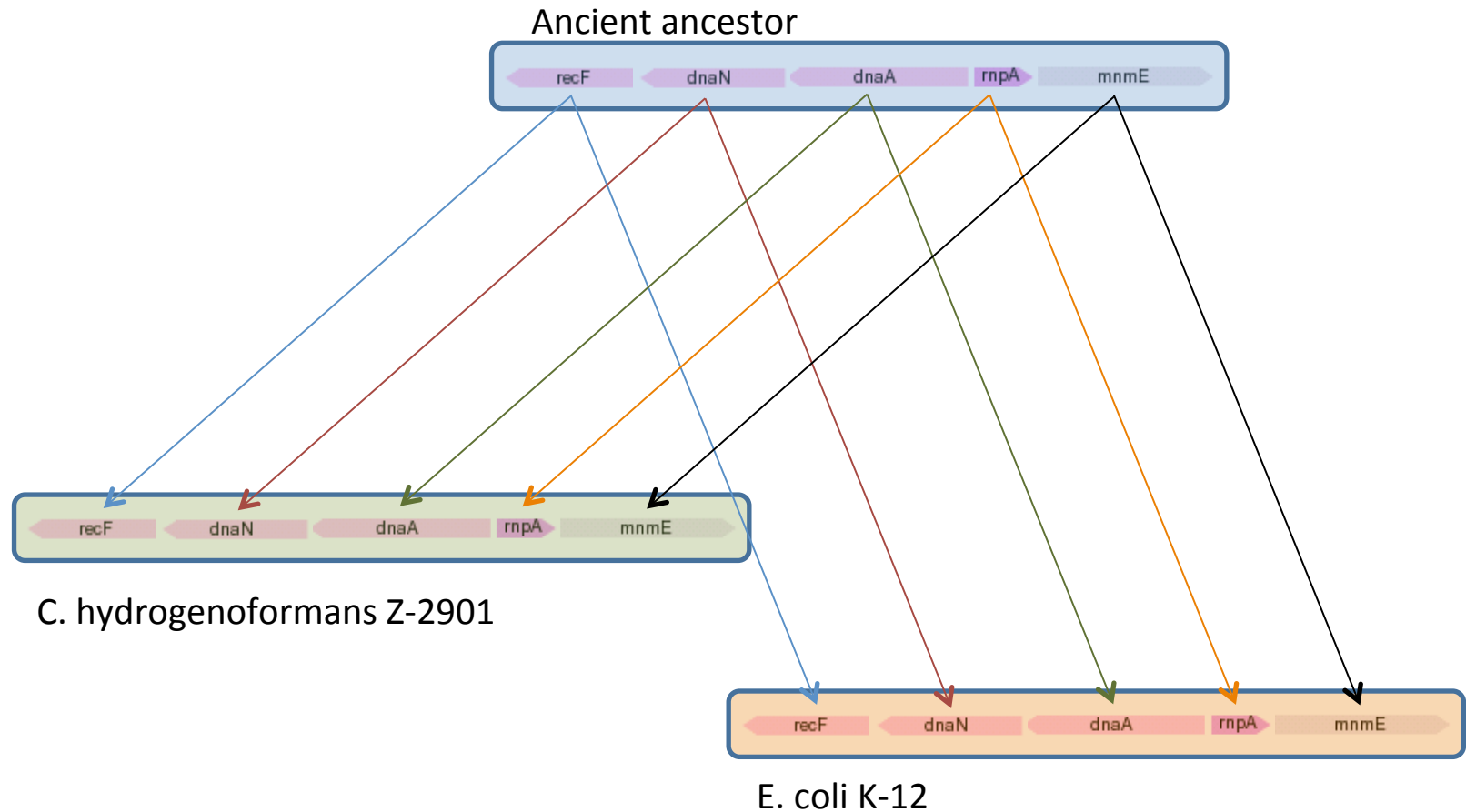


A small region of the Carboxydothemus hydrogenoformans Z-2901 genome near the replication origin.



A small region of the *E. coli* K-12 genome

These two genomes are examples of some of the most distantly related bacteria, i.e. they have likely evolved separately for > a billion years.



- The recF genes in the two species derive from a recF gene that was already present in the ancient ancestral bacterium.
- The same holds for the other 4 genes in this example.
- These are examples of *orthologous* pairs of genes.
- Orthologous genes typically perform roughly the same function today as they already performed in the ancestor.

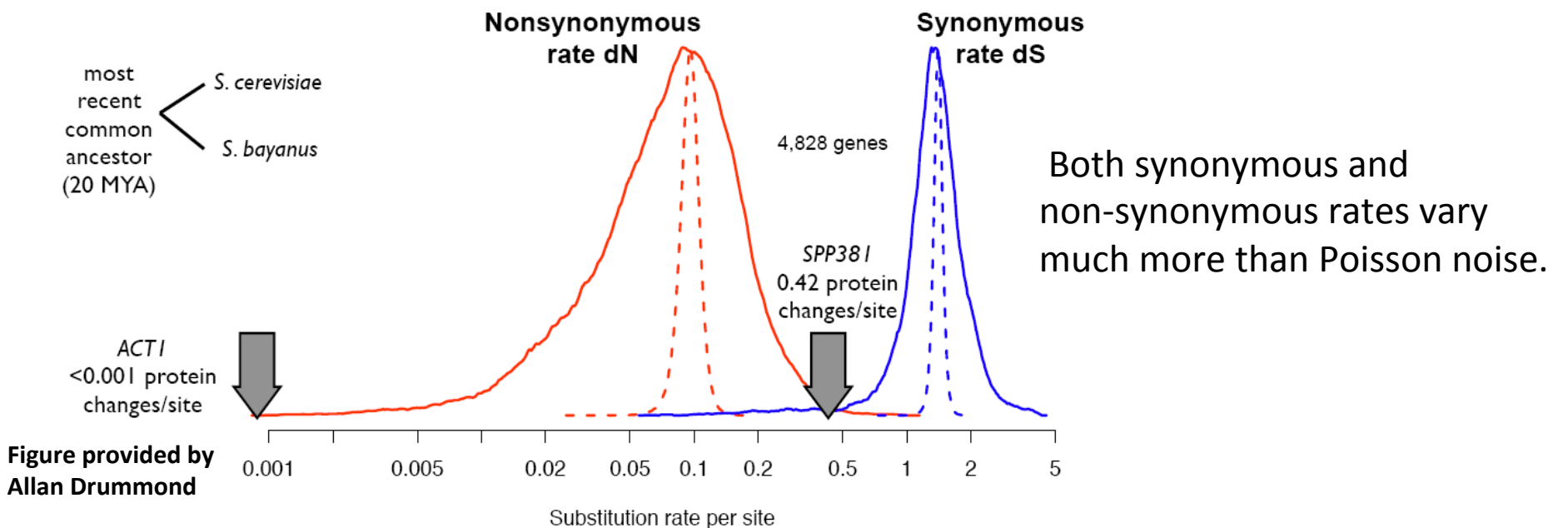
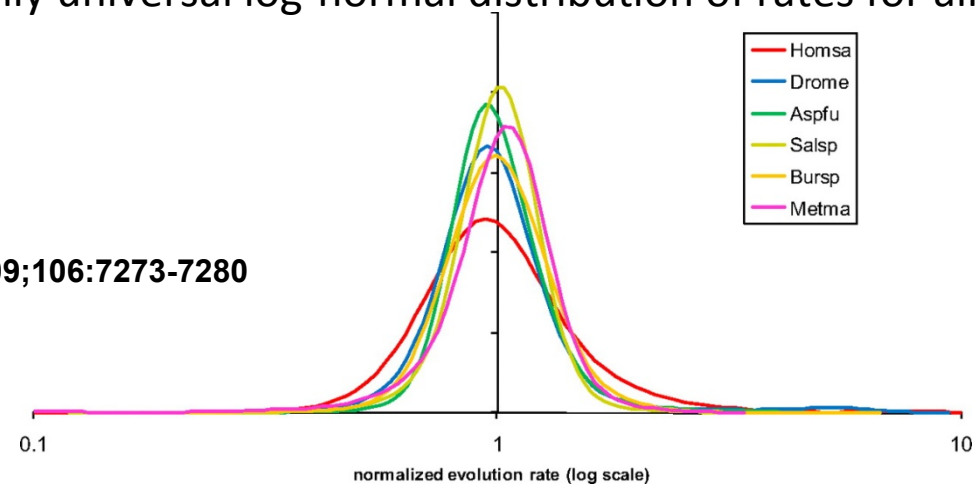


Figure provided by Allan Drummond

Roughly universal log-normal distribution of rates for all species.

Wolf Y I et al. PNAS 2009;106:7273-7280



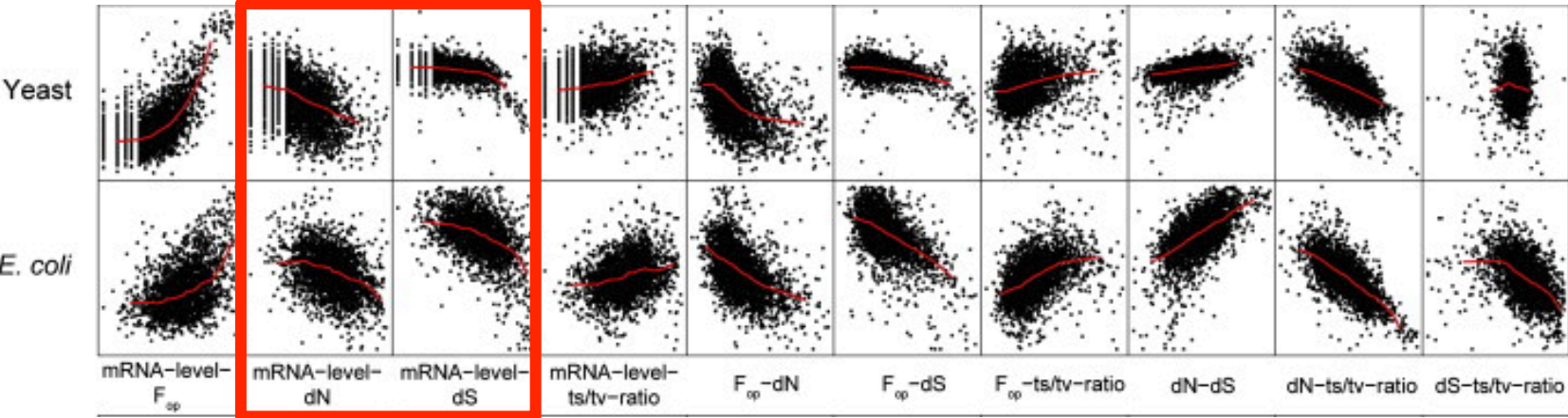
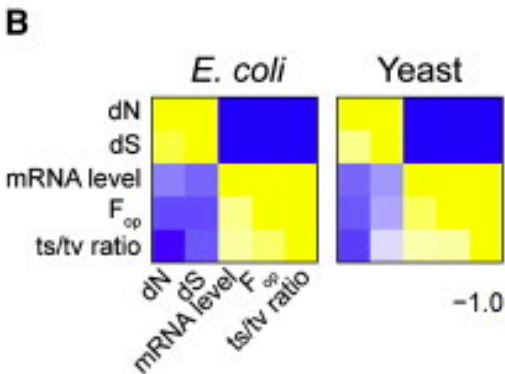
What determines if a gene evolves slow or fast?

Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution

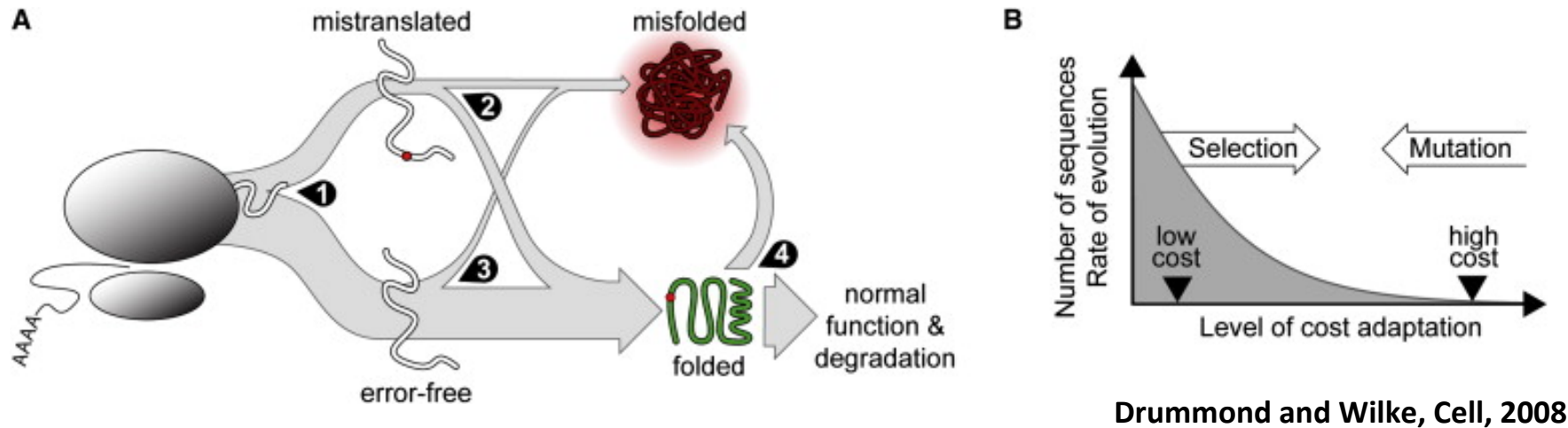
D. Allan Drummond¹, Claus O. Wilke²

Show more

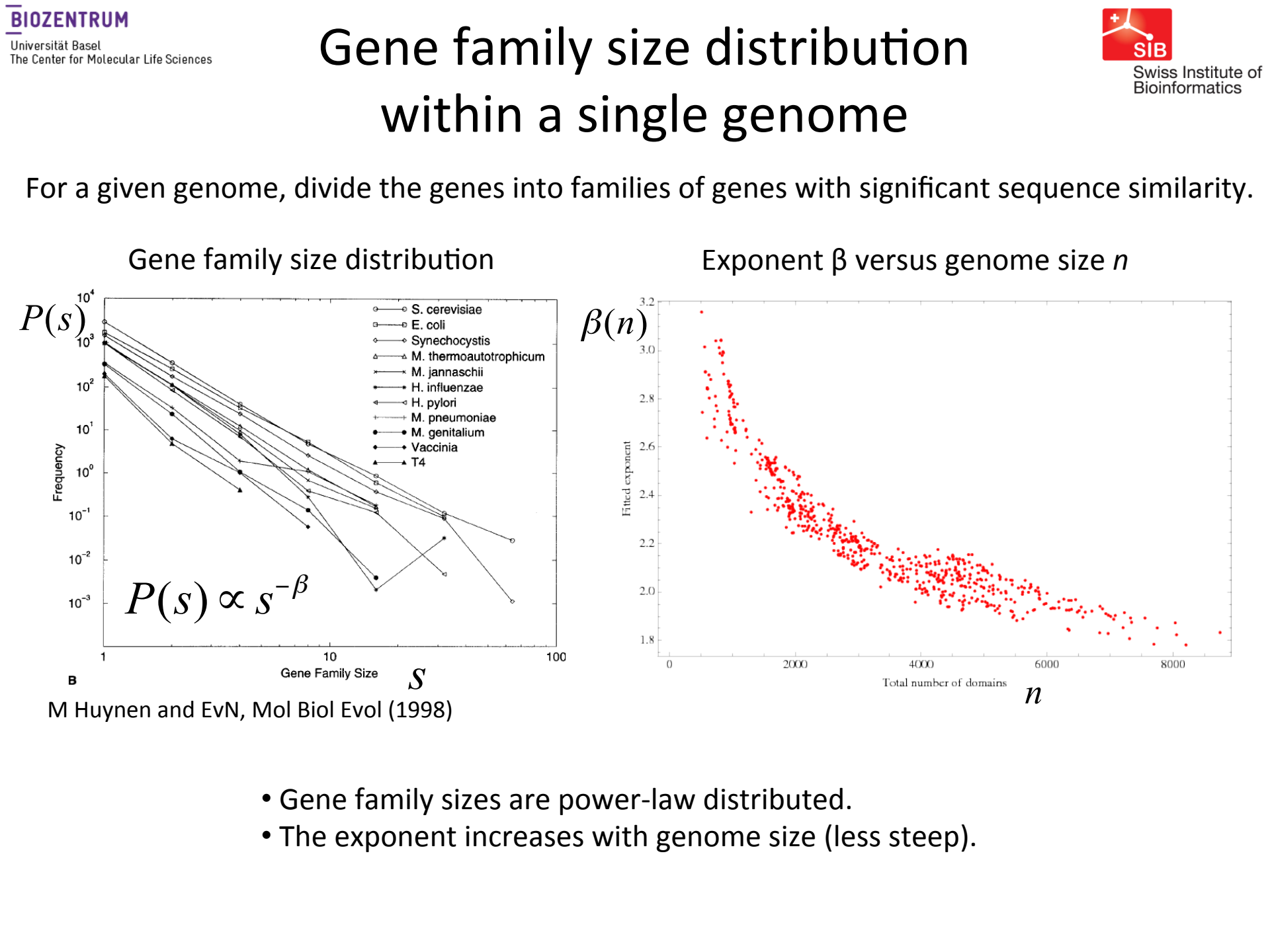
DOI: 10.1016/j.cell.2008.05.042



- Both dN and dS correlate negatively with mRNA expression level of the gene.



- Compared to RNA and DNA polymerases, *ribosomes are highly error-prone*.
- A non-negligible fraction of proteins have *wrong* amino acids in them when they are made.
- This may cause proteins to misfold: misfolded proteins are nonfunctional and may be toxic.
- **Adaptation:** The fraction of misfolded proteins for a given gene can be lowered by:
 - Using codons that are less likely to be mistranslated.
 - Use amino acids that stabilize folding.
- **Cost/Benefit:**
 - Cost/benefit of adaptation is proportional to the gene's expression level.
 - Highly adapted sequences are rare -> lower evolutionary rate.



Protein domains: Functional and evolutionary ‘atoms’

To some reasonable approximation protein domains are:

- Functional units.
- Structural units (can fold by themselves).
- ‘Atomic’: Cannot be split into smaller functional units.

All occurrences of a given protein domain are likely the result of duplication and mutation of a common ancestral domain.



[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#)



Pfam 23.0 (July 2008, 10340 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

Pfam: Large collection of Hidden Markov Models of (mostly) *non-redundant* protein domains.

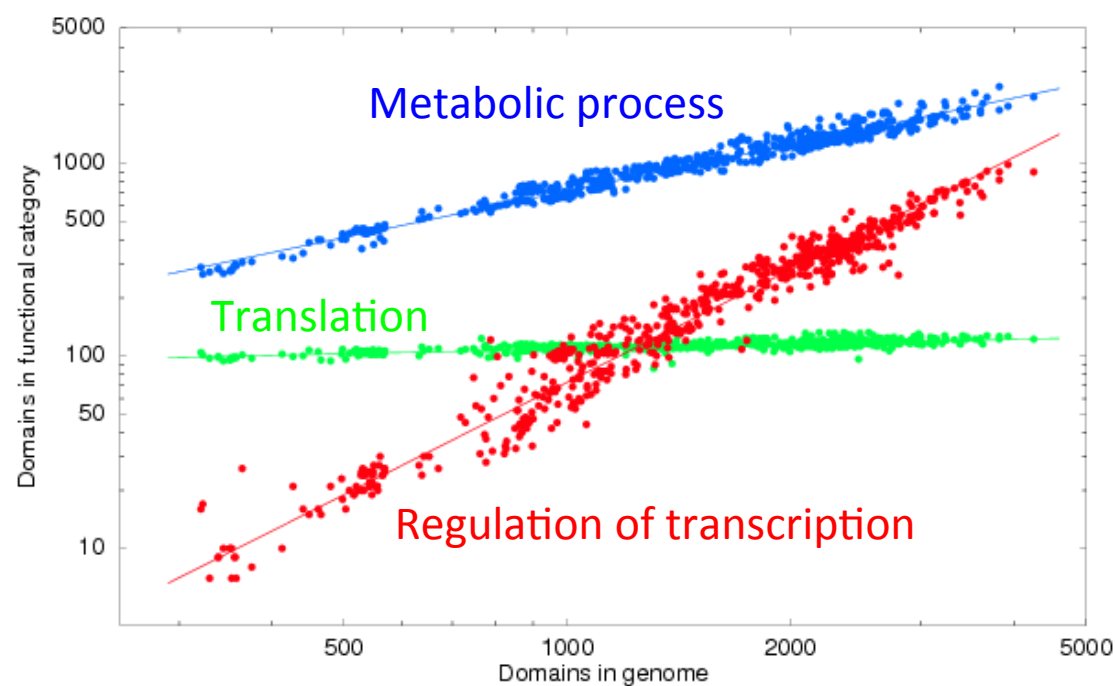


HMMER

biosequence analysis using profile hidden Markov models



Standard tools are available for predicting Pfam-family occurrences in whole genomes. Pfam domains can be associated with more or less specific functions.



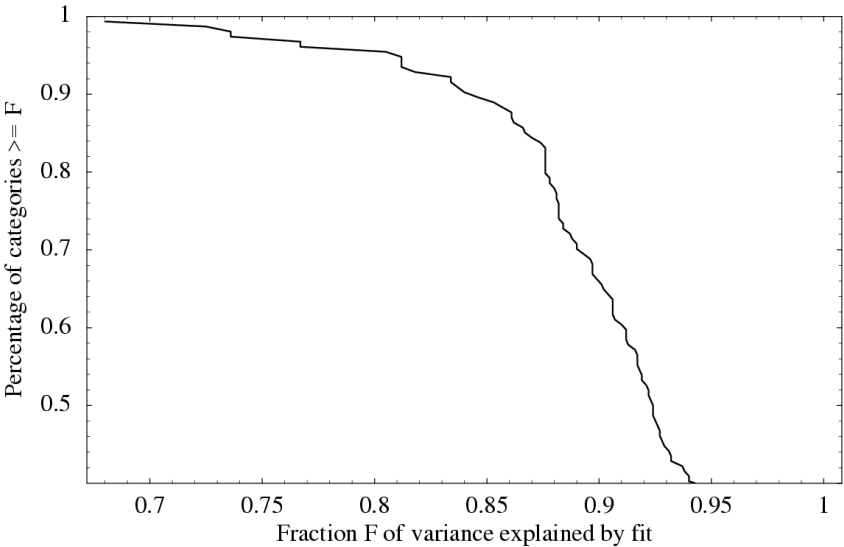
GO term	Exponent α_c
translation	0.08 ± 0.01
DNA replication	0.41 ± 0.04
DNA repair	0.61 ± 0.04
metabolic process	0.80 ± 0.01
carbohydrate metabolic process	1.10 ± 0.06
transport	1.30 ± 0.06
porphyrin metabolic process	1.66 ± 0.16
regulation of transcription	1.94 ± 0.04
secretion	1.98 ± 0.20
DNA recombination	2.23 ± 0.23

EvN Trends in Genetics (2003)

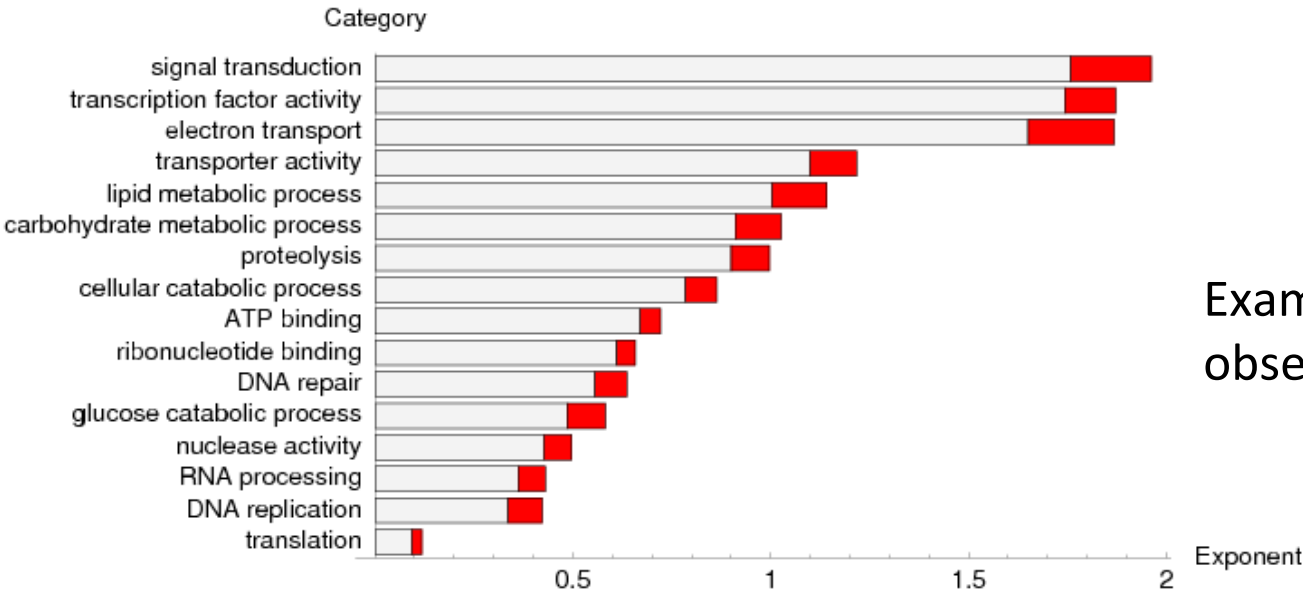
N Molina &EvN, Biol Direct (2008)

$$n_c = A_c n^{\alpha_c} \Leftrightarrow \log(n_c) = \beta_c + \alpha_c \log(n)$$

Scaling laws are observed for most high-level functional categories



Reverse cumulative distribution of the *fraction of variance* explained by the fit for all ubiquitous categories.

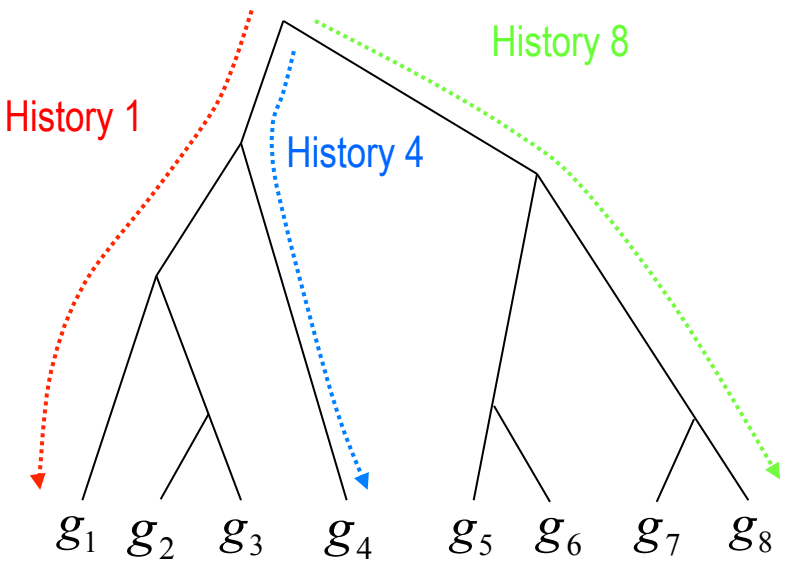


Examples of the observed exponents α_c

Constraints on domain-count dynamics

Define $x(g,t) = \log[n(g,t)]$ $x_c(g,t) = \log[n_c(g,t)]$

$\{x_c(0)\}, x(0)$ **Common ancestor**



$$x_c(g_8, t_{\text{today}}) = x_c(0) + \int_0^{t_{\text{today}}} \frac{dn_c(g_8, t)}{n_c(g_8, t)}$$

$$x(g_8, t_{\text{today}}) = x(0) + \int_0^{t_{\text{today}}} \frac{dn(g_8, t)}{n(g_8, t)}$$

The scaling laws say that $x_c(g, t_{\text{today}}) = \beta_c + \alpha_c x(g, t_{\text{today}})$ **for all genomes!**


Therefore all histories obey:

$$\int_0^{t_{\text{today}}} \frac{dn_c(g, t)}{n_c(g, t)} = \alpha_c \int_0^{t_{\text{today}}} \frac{dn(g, t)}{n(g, t)} \quad \forall g$$

Time invariance:


Assume that if we had collected genomes, 10, 50, 100, or 500 million years ago we would have found the **same** scaling laws.

$$\int_0^{t_{\text{today}}} \frac{dn_c(g,t)}{n_c(g,t)} = \alpha_c \int_0^{t_{\text{today}}} \frac{dn(g,t)}{n(g,t)} \quad \forall g$$




$$\frac{dn_c}{n_c} = \alpha_c \frac{dn}{n}$$

Equivalently:



$$\frac{dn_c}{dn} = \alpha_c \frac{n_c}{n}$$



Fraction of all domains that are category c.

Fraction of all domain gain/losses involving domains from category c.

Prediction:

- If over a short interval a total of dn domain gain/losses occur, then the fraction of those involving domains from category c is α_c as large as would be expected if one were to select domains at random from the genome.

Test using closely-related pairs of genomes

For a pair of closely-related genomes (g, g') we have roughly:

$$f_c \approx \frac{n_c(g) + n_c(g')}{n(g) + n(g')}$$

- We estimate the total number of domain-count gain/losses:

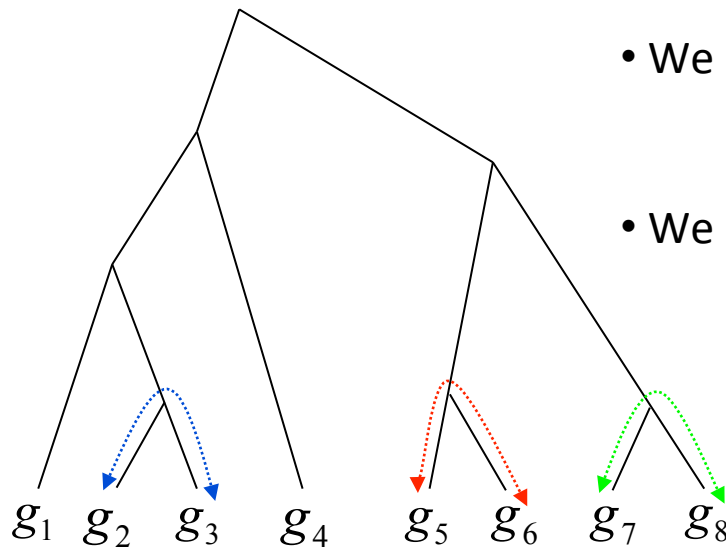
$$\Delta n$$

- We estimate the gains/losses in category c :

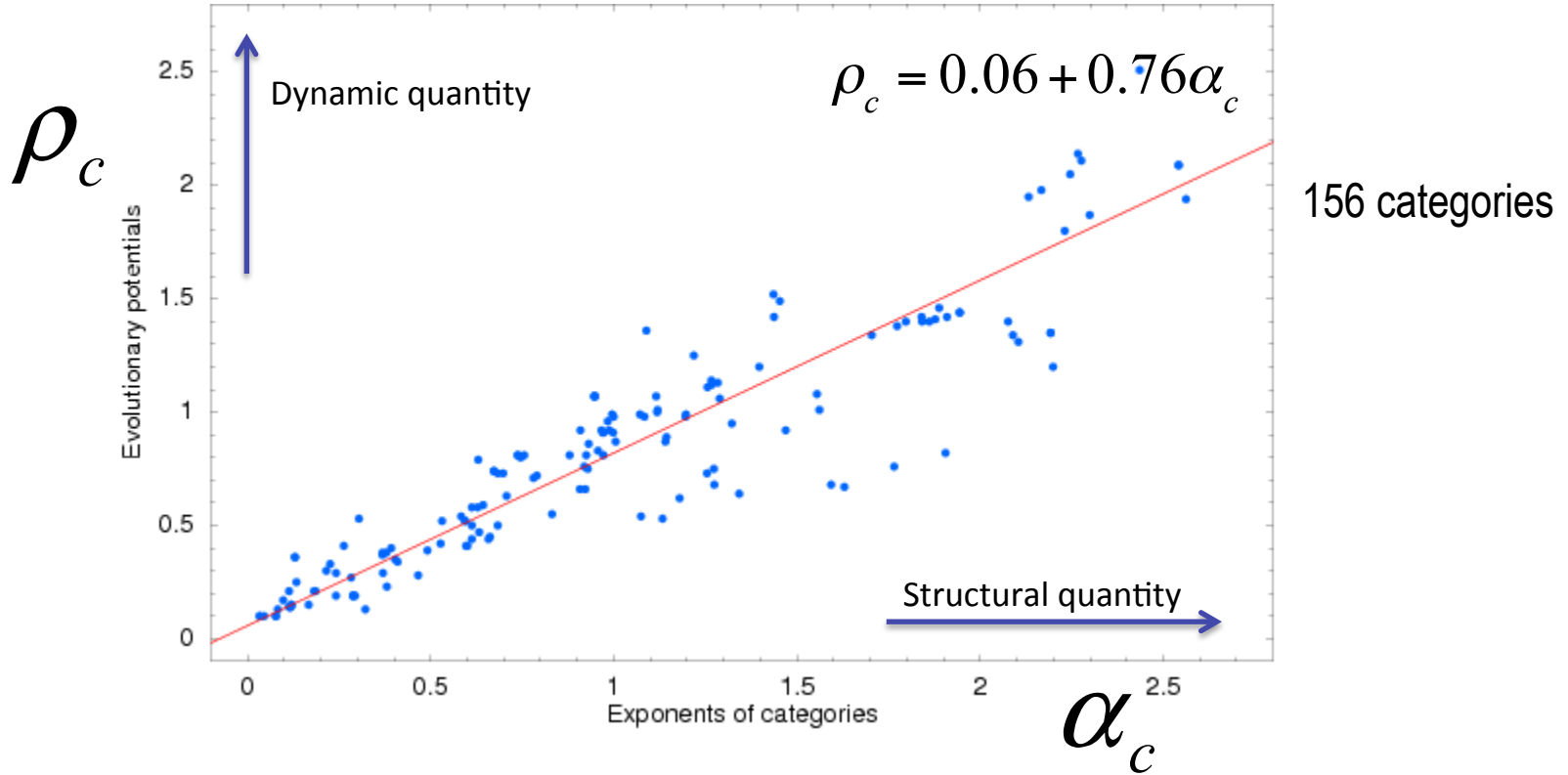
$$\Delta n_c$$

Define the evolution potential as:

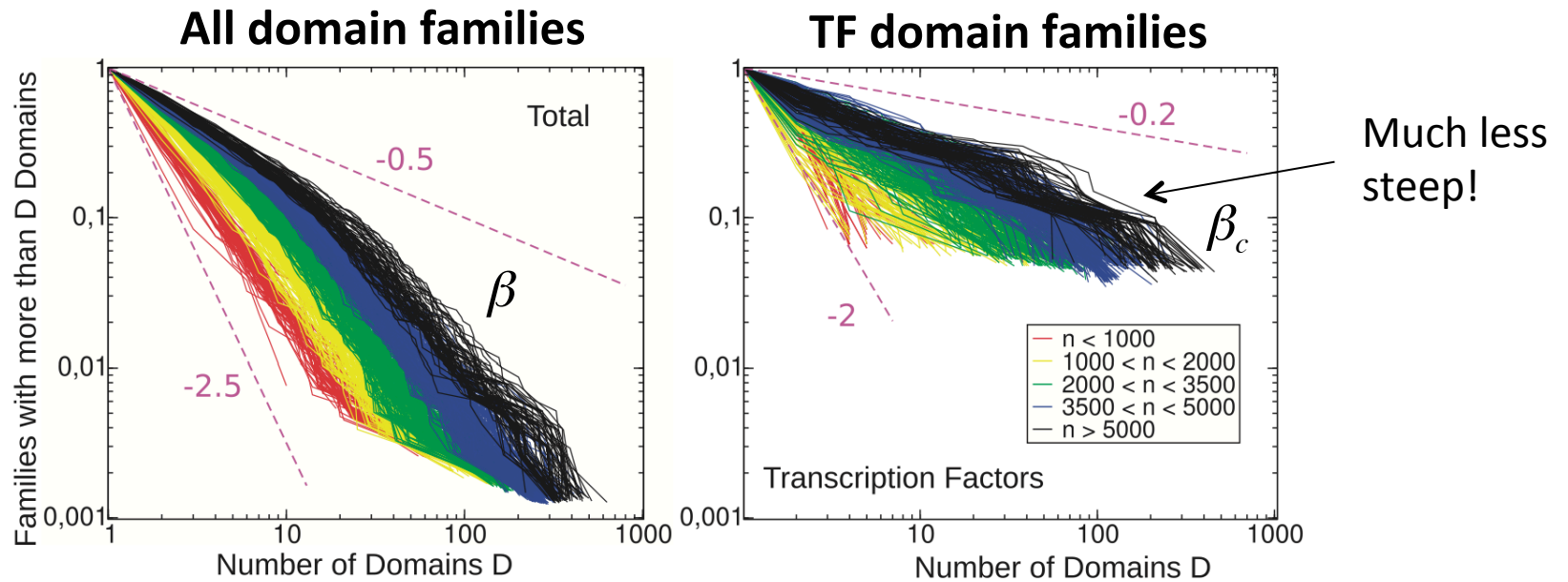
$$\rho_c = \frac{\Delta n_c(g, g')}{\Delta n(g, g')} \frac{1}{f_c}$$



Thus the model predicts: $\rho_c = \alpha_c$



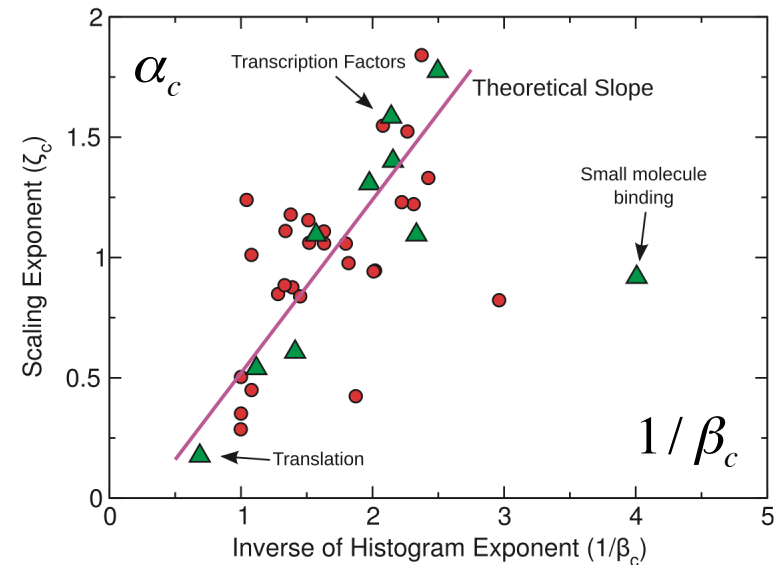
- Conclusions:**
- The evolutionary potentials (ρ_c, α_c) are fundamental constants of genome evolution.
 - They quantify the relative rates at which additions and deletions of domains in that class are fixed in evolution.
 - What determines the values of these exponents is still not clear.



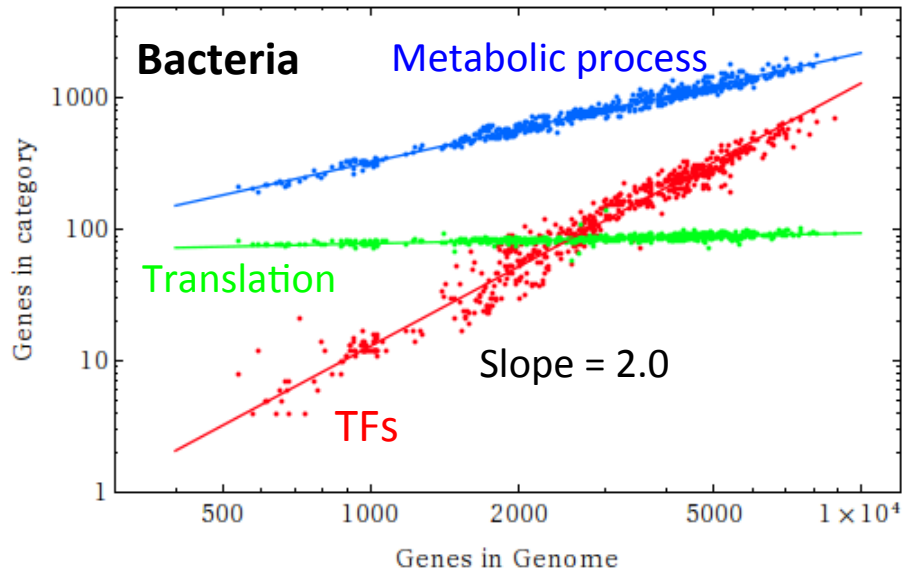
In general the *inverse* of the family-size distribution exponent is roughly proportional to the *scaling exponent* of the category:

$$\alpha_c = \frac{\beta}{\beta_c}$$

From: Grilli et al, Nucl. Acids Res, 2012



How many TFs are needed to regulate a genome with G genes?



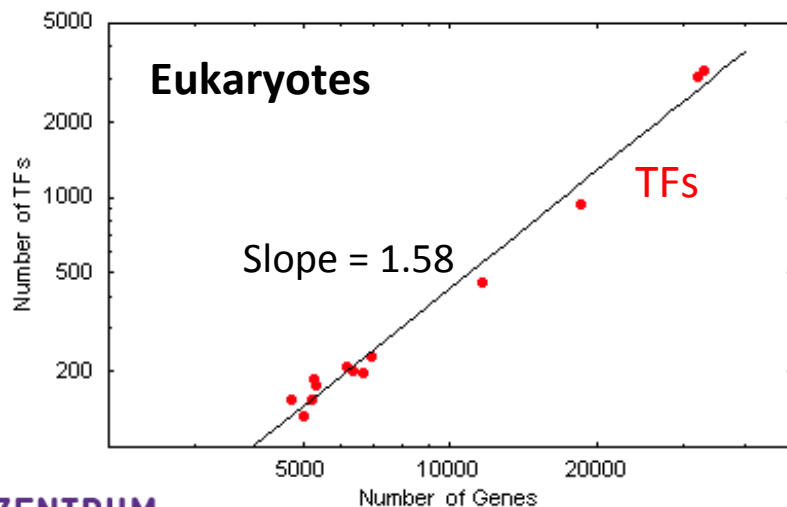
Quadratic in bacteria:

$$\text{Number of TFs} \propto G^{2.0}$$

Example other categories:

$$\text{Number of metabolic genes} \propto G^{0.83}$$

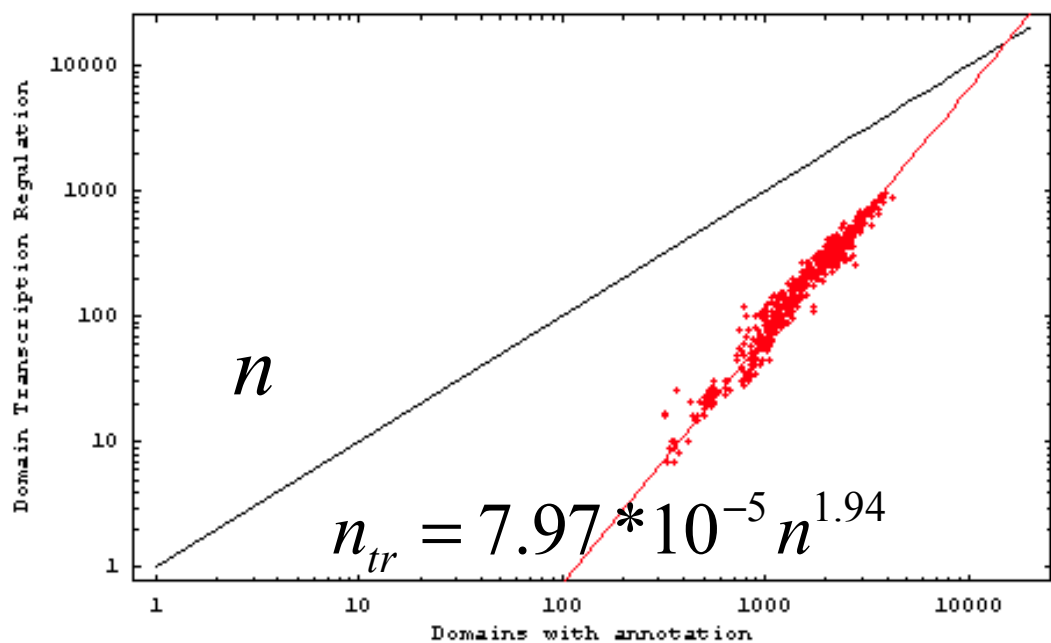
$$\text{Number of translation genes} \propto G^{0.08}$$



In eukaryotes still super-linear:

$$\text{Number of TFs} \propto G^{1.58}$$

If one naively extends the scaling law one would eventually have more transcription regulatory domains than there are domains.



Implies upper bound of $n = 15,500$

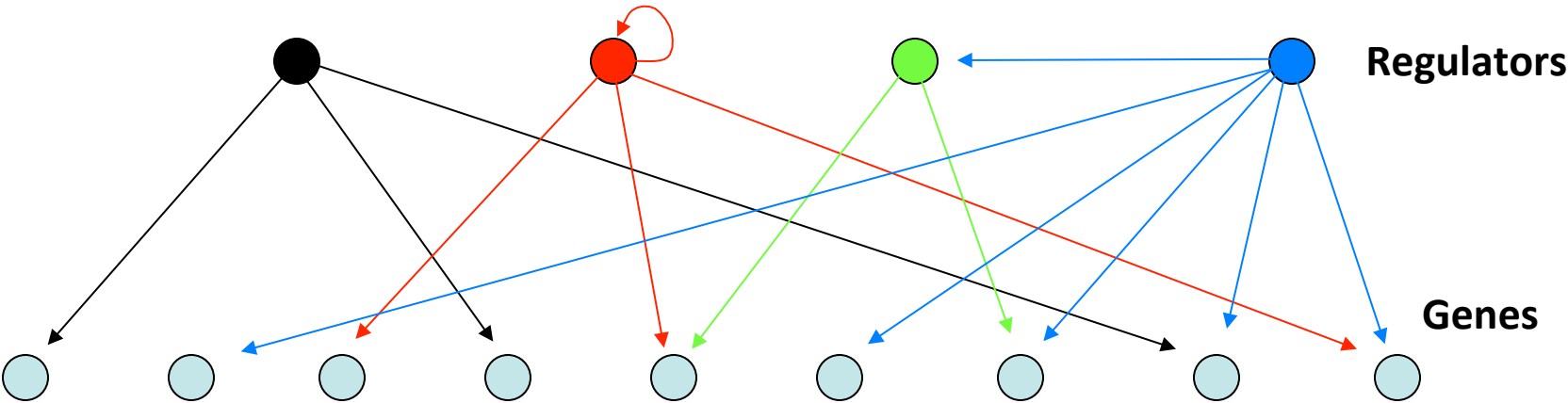
Observed maximum: $n=4254$

Another bound is obtained by assuming that the number of transcription regulatory domains cannot increase faster than the number of domains:

$$dn_{tr} \leq dn \Leftrightarrow A_{tr} (n+1)^{\alpha_{tr}} \leq A_{tr} n^{\alpha_{tr}} + 1$$

This implies an upper bound of $n = 7723$. Too large by a factor of 1.8.

Consequences for the topology of the transcription regulatory network



$\langle i(g) \rangle$ = average number of incoming arrows per gene in genome with g genes.

$\langle o(g) \rangle$ = average number of outgoing arrows per regulator in genome with g genes.

$r(g) \propto g^2$ = number of regulators in genome with g genes.

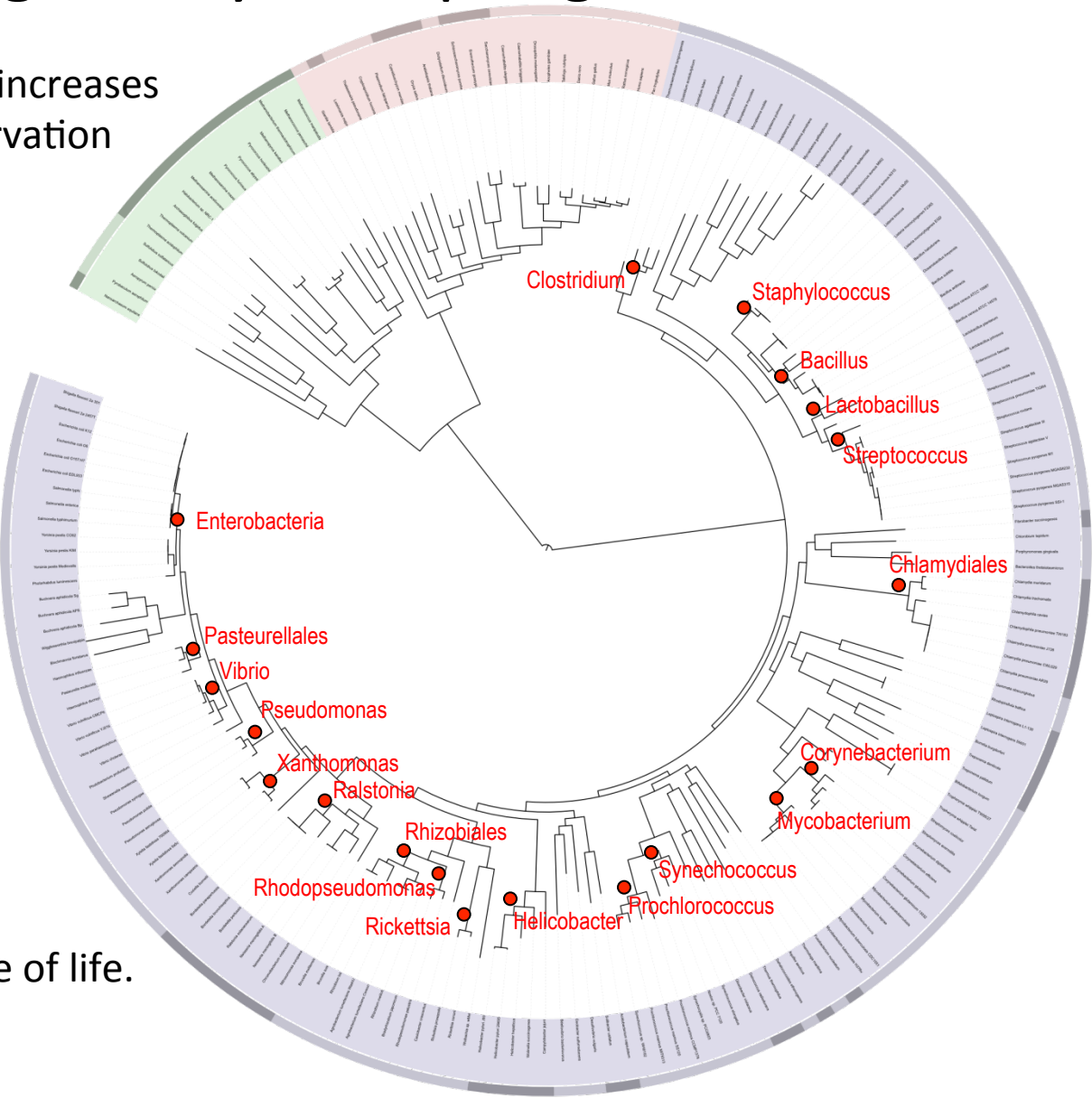
It follows that
$$r(g)\langle o(g) \rangle = \langle i(g) \rangle g \Leftrightarrow \frac{\langle i(g) \rangle}{\langle o(g) \rangle} \propto g$$

Two possibilities:

- Genes in larger genomes are regulated by more regulators.
- Regulators in larger genomes regulate smaller numbers of genes.

Estimating the number of regulatory sites per gene

- If the number of sites per gene increases linearly with genome size, conservation patterns should show this.
- Take sets of related bacteria.
- Of different genome sizes.
- Study conservation patterns in intergenic regions.



- Data: 22 Bacterial clades on tree of life.
105 genomes in total.

Quantifying selection at non-coding positions genome-wide

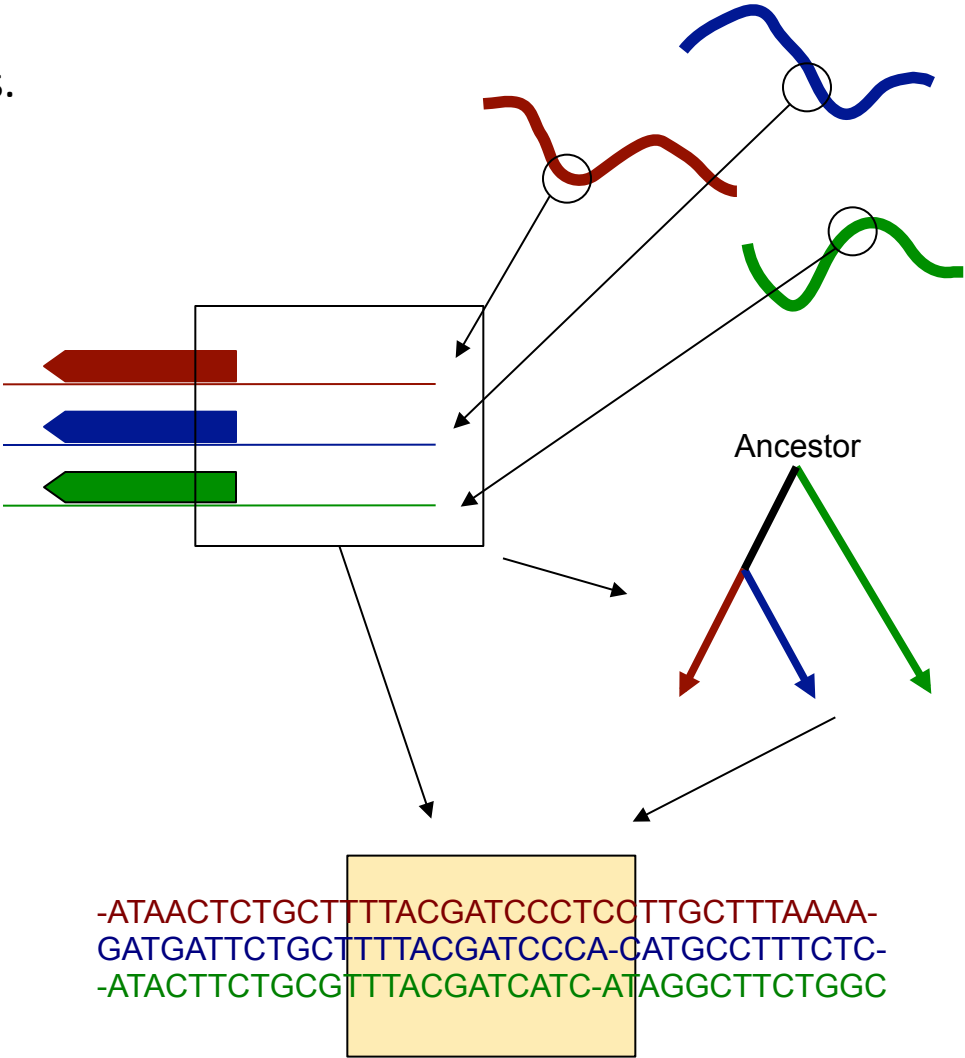
Genome sequences of a set of related species.

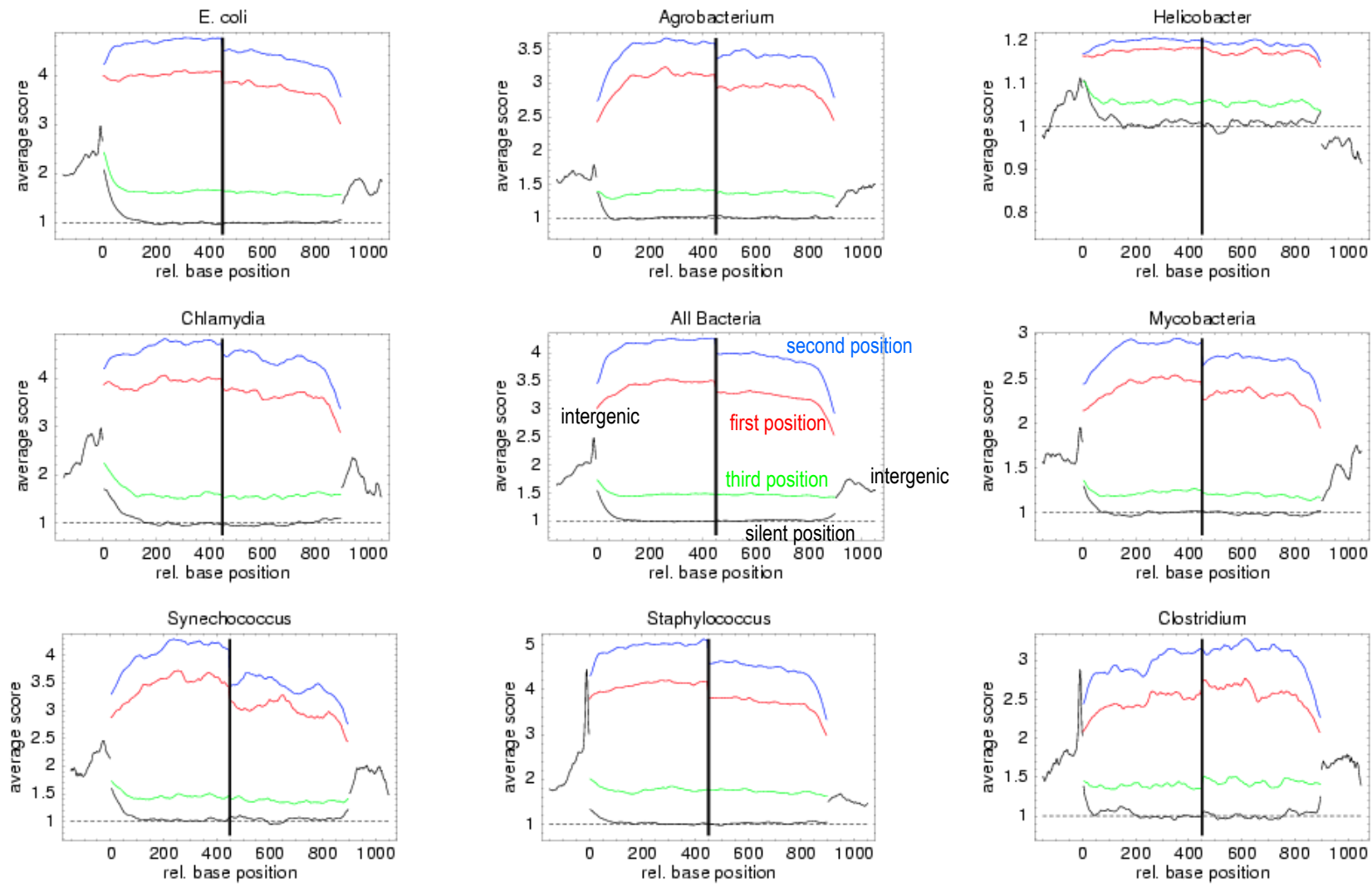
Identification of sets of orthologous genes.

Phylogenetic tree reconstruction.

Alignment of orthologous intergenic regions

Genome-wide evaluation of selection acting at non-coding positions





No correlation whatsoever between genome size and amount of conservation upstream.

Intergenic region sizes are independent of genome size



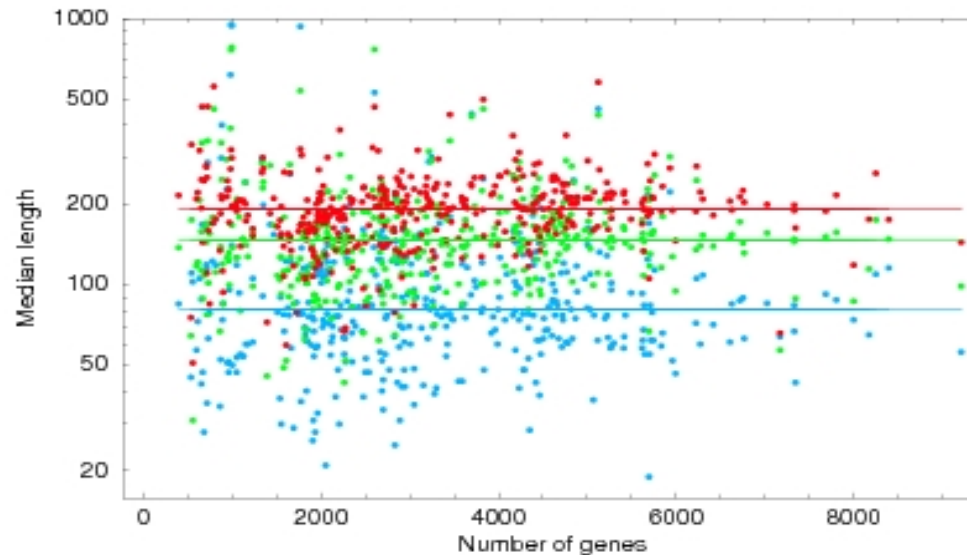
Double regulatory intergenic regions (**DR**)



Single regulatory intergenic regions (**SR**)



Non regulatory intergenic regions (**NR**)



$NR < SR < DR$

But no correlation with genome size

$\langle i(g) \rangle$ = average number of regulatory inputs per gene in genome with g genes.

$\langle o(g) \rangle$ = average number of regulatory outputs per regulator in genome with g genes.

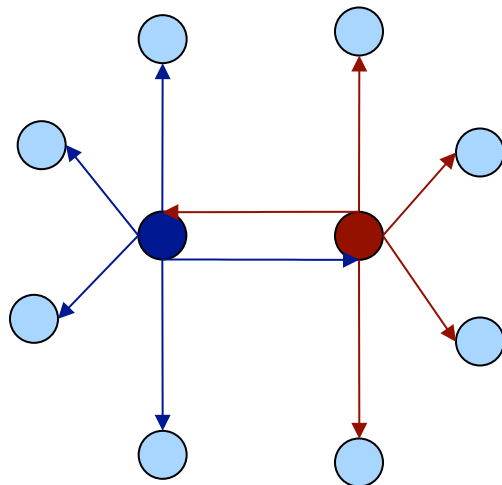
$r(g) \propto g^2$ = number of regulators in genome with g genes.

$$r(g) \langle o(g) \rangle = \langle i(g) \rangle g \Leftrightarrow \frac{\langle i(g) \rangle}{\langle o(g) \rangle} \propto g$$

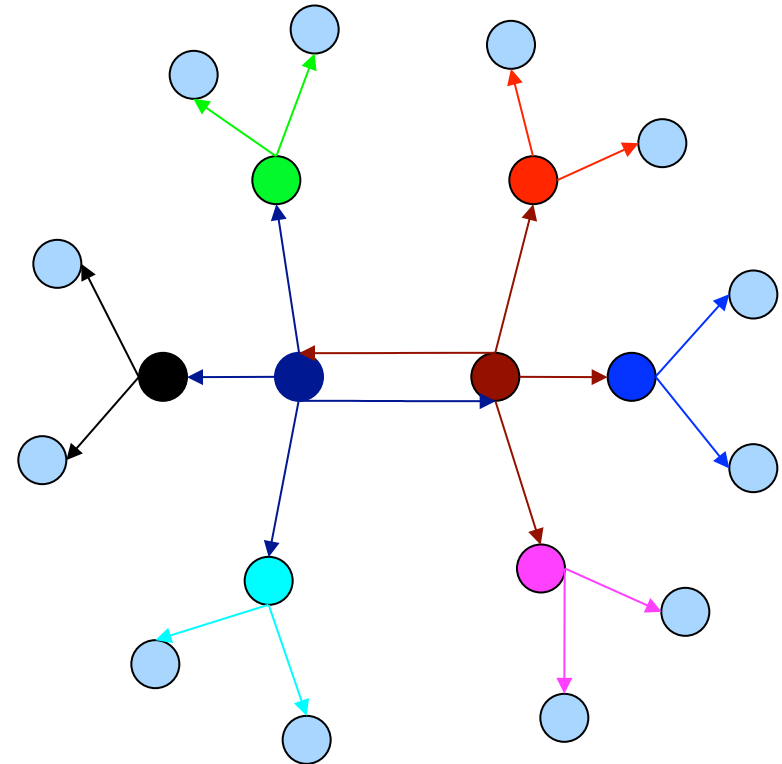
We find: $\langle i(g) \rangle = \text{constant}$

Consequence: $\langle o(g) \rangle \propto g^{-1}$

The number of genes that each regulator regulates decreases with genome size.



x2



- 10 genes
- 2 regulators
- 10 interactions
- Indegree 1
- Outdegree 5



- 20 genes
- 8 regulators
- 20 interactions
- Indegree 1
- Outdegree 2.5

Fundamental difference between regulation in eukaryotes and bacteria

Bacteria

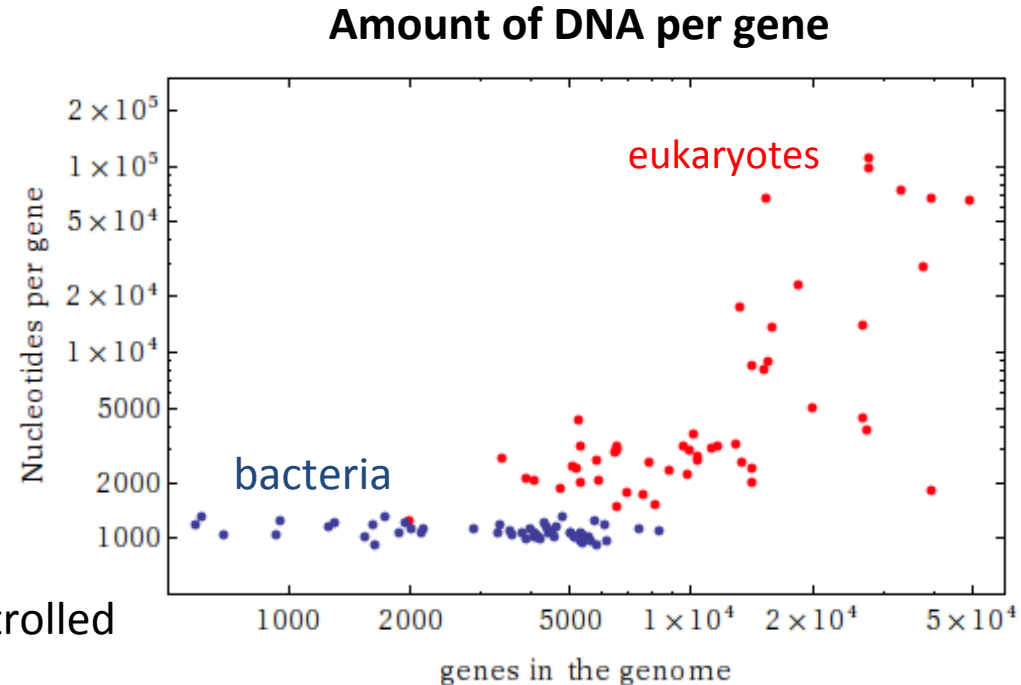
$$I(G) = \text{constant}$$

Bacteria have not evolved mechanisms to put a gene under the control of many regulators.

Multi-cellular eukaryotes

$$I(G) = \text{Increases with genome size}$$

In higher eukaryotes one gene can be controlled by many *distal* regulatory elements.

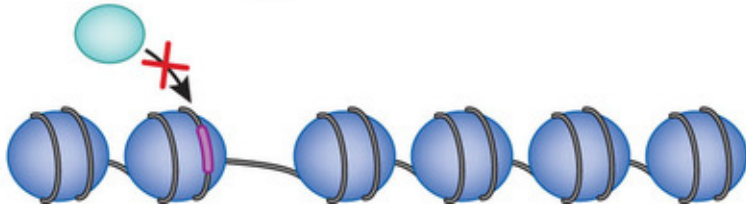


What do eukaryotes have that bacteria have not?

Hypothesis: Chromatin structure.

Which parts of the DNA are accessible for TFs is dynamically regulated.

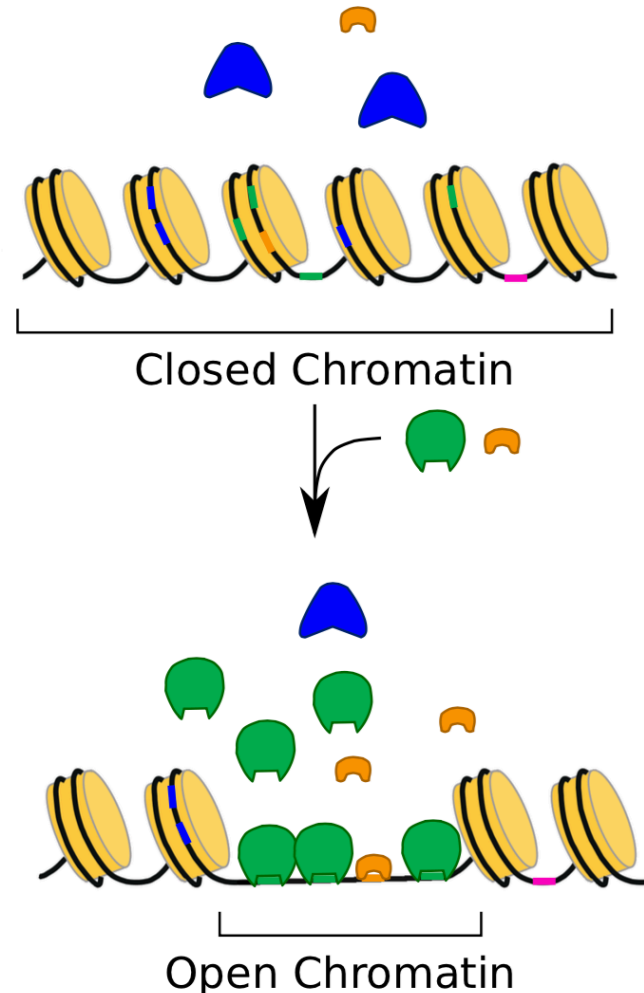
Competition between TFs and nucleosomes for binding to DNA gives combinatorial regulation 'for free'



Nucleosome coverage typically prohibits individual TFBSs to be accessed.

Resolution of the specificity problem:

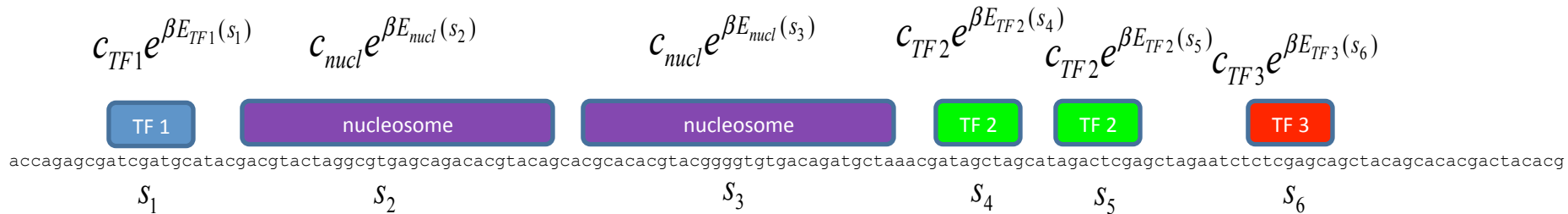
- TFs can *outcompete* the nucleosomes only at those places where *clusters* of TFBSs occurs for the TFs that are *co-expressed*.
- Cooperativity between nearby TFBSs is induced through competition with common nucleosome.
- Consistent with enhancers/promoters typically being 1-3 nucleosomes wide.
- Allows easy evolution of combinatorial regulation.
- Explains lack of 'binding site grammar' beyond clustering of sites within 150-450 bp regions.



Mirny L A PNAS 2010;107:22534-22539

Thermodynamic model for nucleosome/TF competition

Statistical weight of a configuration



Forward algorithm to calculate partition sum

$$F_n = F_{n-1} + F_{n-L} c_{nucl} e^{\beta E_{nucl}(s_{n-L+1} \dots s_n)} + F_{n-l} c_{TF1} e^{\beta E_{TF1}(s_{n-l+1} \dots s_n)} + \dots + F_{n-l} c_{TFN} e^{\beta E_{TFN}(s_{n-l+1} \dots s_n)}$$

accagagcgcgatcgatgcatacgcgctactaggcgtgagcagacacgtacagcacgcacacgtacggggtgtgacagatgctaaccgatagctagcatagactcgagctagaatctctcgagcagctacagcacacgactacacg

$$F_n = F_{n-1} +$$

accagagcgcgatcgatgcatacgcgctactaggcgtgagcagacacgtacagcacgcacacgtacggggtgtgacagatgctaaccgatagctagcatagactcgagctagaatctctcgagcagctacagcacacgactacacg

$$F_{n-L} c_{nucl} e^{\beta E_{nucl}(s_{n-L+1} \dots s_n)} +$$

nucleosome

accagagcgcgatcgatgcatacgcgctactaggcgtgagcagacacgtacagcacgcacacgtacggggtgtgacagatgctaaccgatagctagcatagactcgagctagaatctctcgagcagctacagcacacgactacacg

$$F_{n-l} c_{TF1} e^{\beta E_{TF1}(s_{n-l+1} \dots s_n)} +$$

TF 1

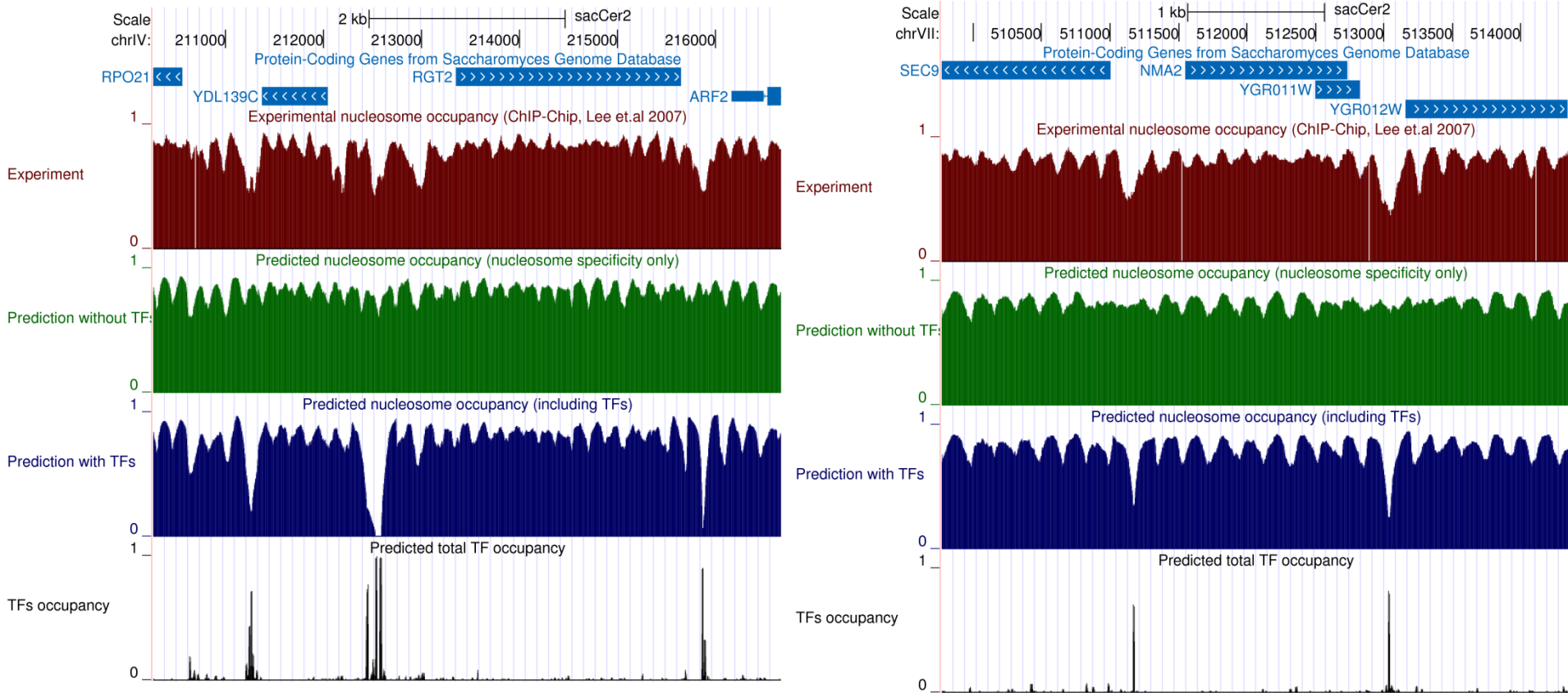
accagagcgcgatcgatgcatacgcgctactaggcgtgagcagacacgtacagcacgcacacgtacggggtgtgacagatgctaaccgatagctagcatagactcgagctagaatctctcgagcagctacagcacacgactacacg

$$\dots + F_{n-l} c_{TFN} e^{\beta E_{TFN}(s_{n-l+1} \dots s_n)}$$

TF N

accagagcgcgatcgatgcatacgcgctactaggcgtgagcagacacgtacagcacgcacacgtacggggtgtgacagatgctaaccgatagctagcatagactcgagctagaatctctcgagcagctacagcacacgactacacg

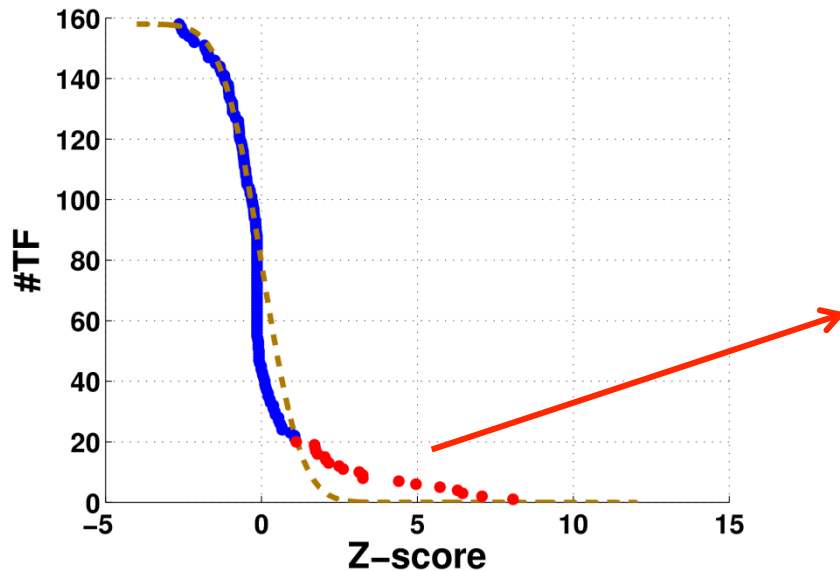
Competition with TFs helps explain nucleosome coverage patterns in yeast



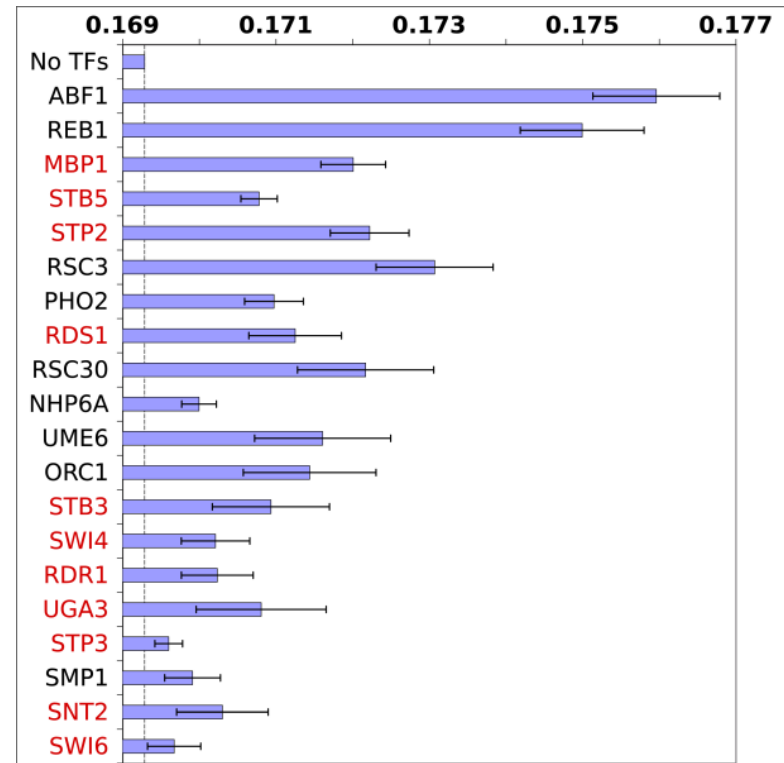
Ozonov & van Nimwegen (PLoS Comp Biol, 2013)

Only a small set of TFs known to interact with chromatin remodellers affect nucleosome coverage

Significance of the contribution of each TF



Yeast TFs that affect nucleosome coverage



Might indicate positive feedback mechanism: A specific class of TFs stabilizes TF binding by recruiting modifiers that cause local nucleosome eviction. **Pioneer factors?**

Outline of the lectures

Day 1

1. Computational methods for determining the constellations of regulatory sites.
2. From constellations of regulatory sites to genome-wide gene expression patterns.

Day 2

3. Large-scale patterns in genomes and gene regulatory networks.
4. Gene expression noise and its role in the evolution of *de novo* gene regulation.

Day 3

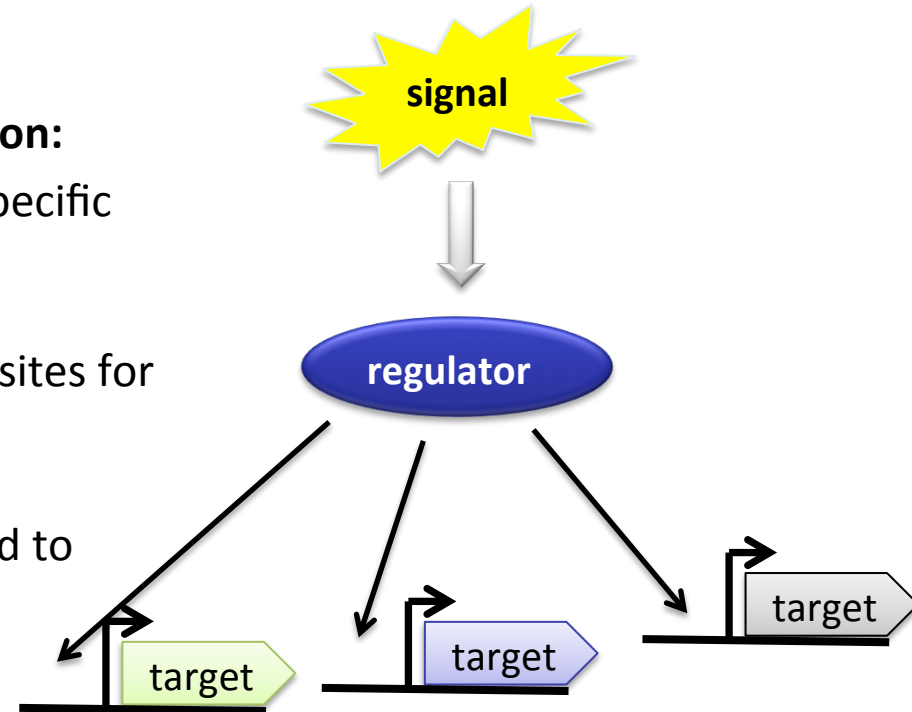
5. *How do bacterial genomes evolve in the wild?*

Where does gene regulation come from?

Gene regulatory interactions are assumed to be finely-tuned to ensure appropriate regulation. But how can this evolve from a situation without regulation?

Requirements for *de novo* evolution of regulation:

- Regulatory protein evolves to respond to a specific signal.
- The appropriate target genes evolve binding sites for the regulator.
- Strength of regulator/target interaction tuned to ensure appropriate induction levels.

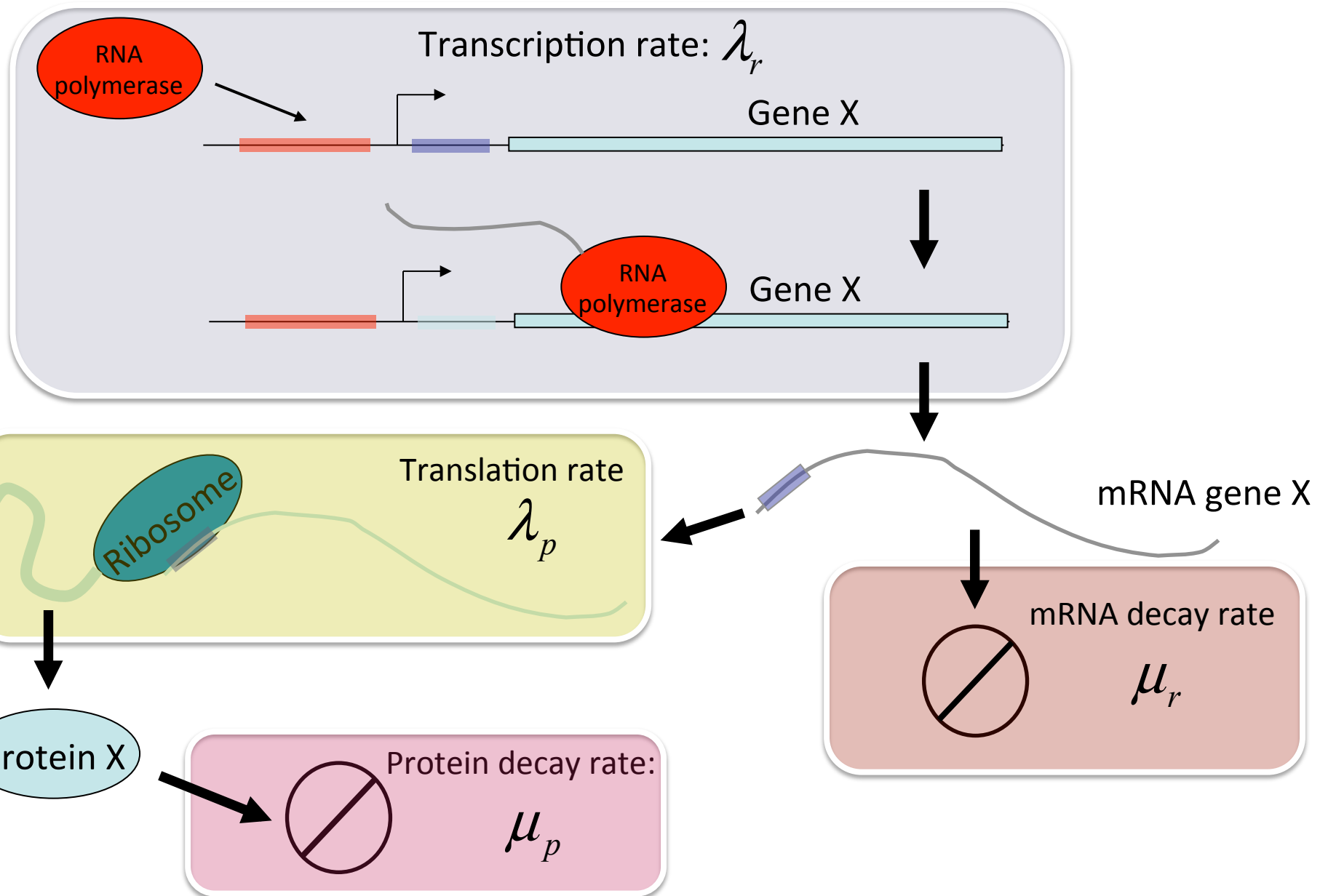


How does this come about?

Insights from a study on gene expression noise.

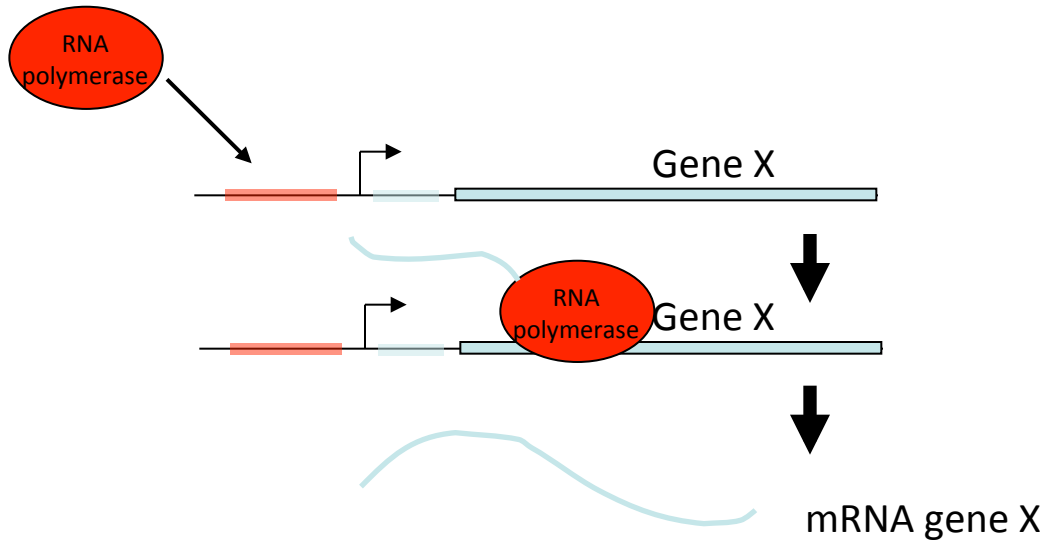
Expression noise facilitates the evolution of gene regulation.

Cartoon of the steps in gene expression

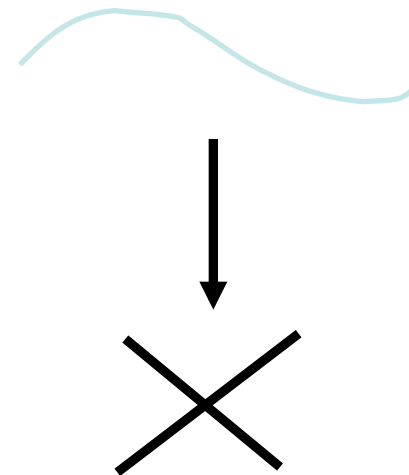


Stochastic transcription and decay

Probability λ_r per unit time to transcribe a new mRNA.



Probability μ_r per mRNA per unit time that it will decay.



Probability that there are n mRNAs at time t : $P_n(t)$

Differential equation for the distribution:

$$\frac{dP_n(t)}{dt} = \lambda_r P_{n-1}(t) + \mu_r (n+1) P_{n+1}(t) - (\lambda_r + \mu_r n) P_n(t)$$

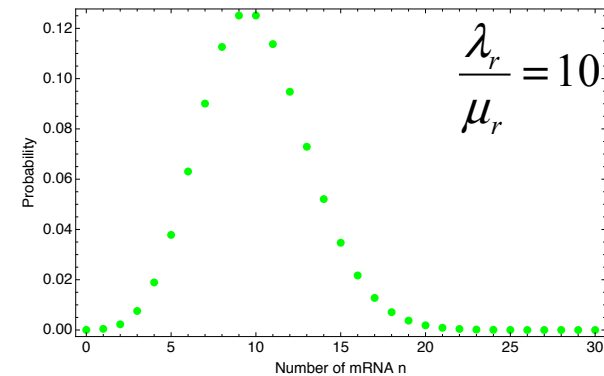
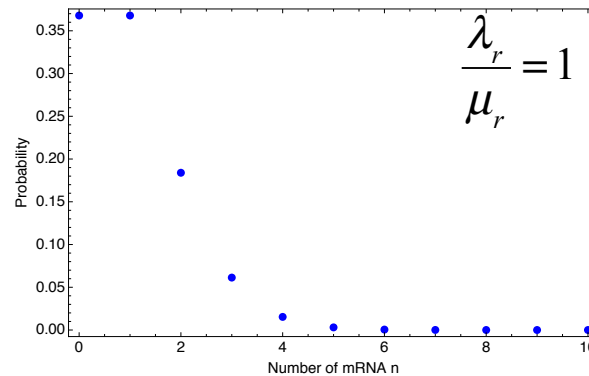
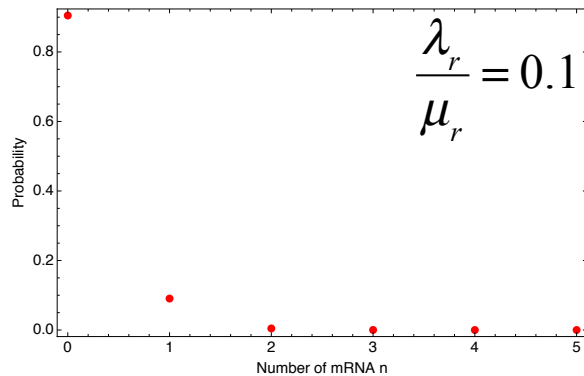
Steady-state is Poisson distribution

Probability to have n mRNAs: $P_n = \frac{1}{n!} \left(\frac{\lambda_r}{\mu_r} \right)^n e^{-\lambda_r/\mu_r}$

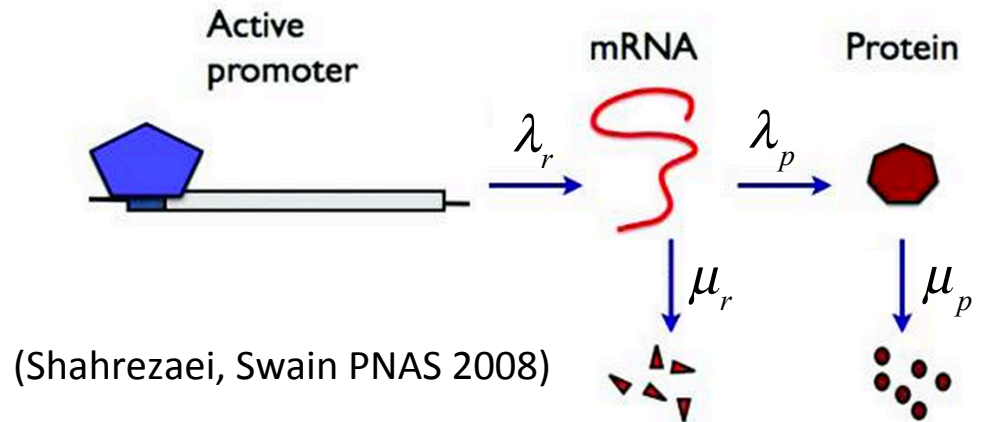
Mean: $\langle n \rangle = \frac{\lambda}{\mu}$

Variance: $\text{var}(n) = \langle n \rangle = \frac{\lambda}{\mu}$

Standard-deviation: $\sigma(n) = \sqrt{\langle n \rangle}$



- λ_r transcription
- μ_r mRNA decay
- λ_p translation
- μ_p protein decay



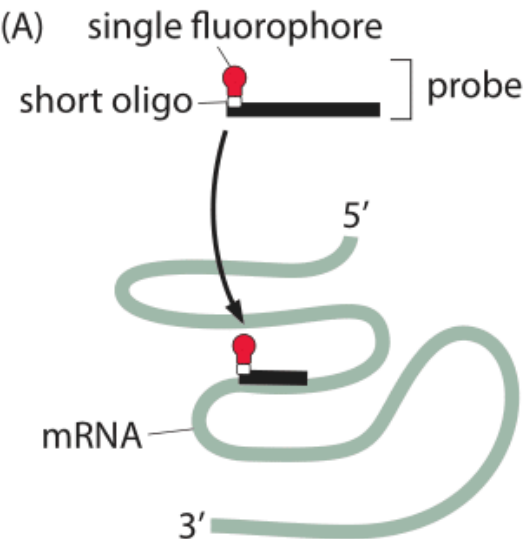
- Proteins are often long-lived: approximation protein-decay slow relative to mRNA decay.
- Solution in terms of two ratios:

$$a = \frac{\lambda_r}{\mu_p} \text{ Transcription events per protein lifetime. } b = \frac{\lambda_p}{\mu_p} \text{ "burst size": translations per mRNA lifetime.}$$

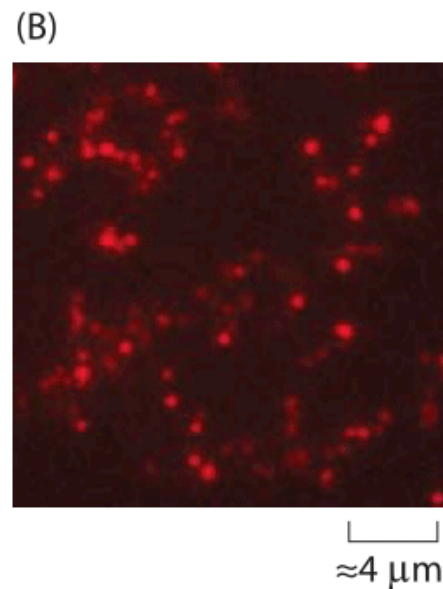
$$P_n = \frac{\Gamma(a+n)}{\Gamma(a)n!} \left(\frac{b}{b+1}\right)^n \left(1 - \frac{b}{b+1}\right)^a \quad \text{mean and variance: } \langle n \rangle = ab, \quad \text{var}(n) = (b+1)\langle n \rangle$$

$$\text{noise: } \eta(n) = \frac{\sigma(n)}{\langle n \rangle} = \sqrt{\frac{\text{var}(n)}{\langle n \rangle^2}} = \sqrt{\frac{b+1}{\langle n \rangle}}$$

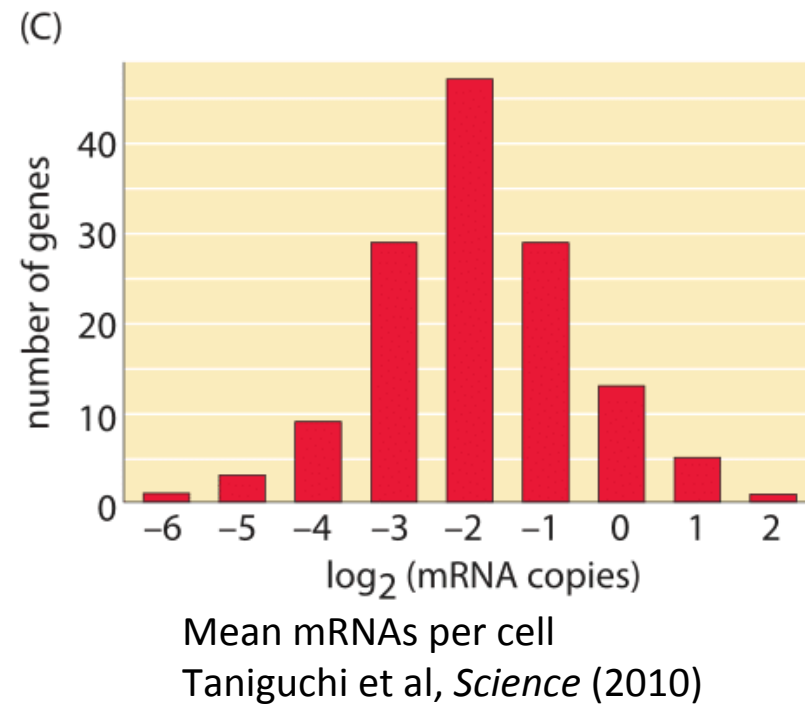
Typical genes have less than 1 mRNA per cell in E coli



Fluorescently labeling single mRNAs (Fluorescence In Situ Hybridization).



Counting mRNAs per cell under the microscope.

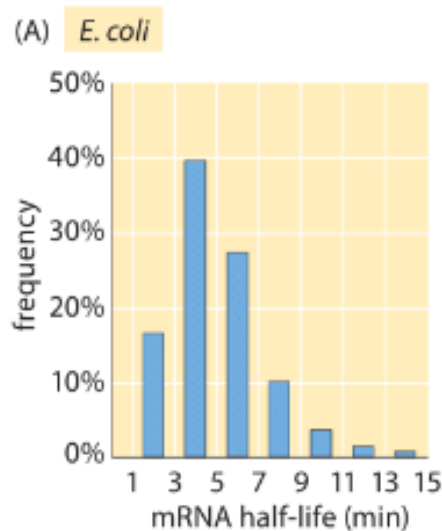


From: Milo and Phillips, Cell Biology by the numbers

Some additional numbers for *E. coli*

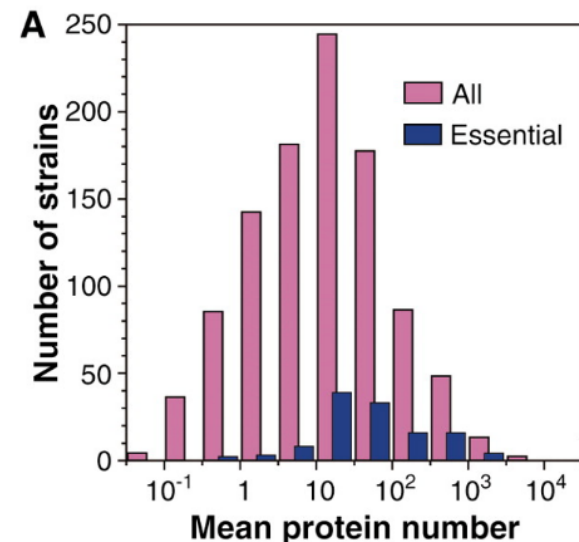
- RNA polymerases per cell: 1'500-10'000 (depending on growth rate).
- Ribosomes per cell: 14'000 (1 doubling per hour) – 45'000 (2 doublings per hour).
- mRNA decay rate: 1-15 minutes half-life.
- Protein decay rate: typically a few hours.
- Protein dilution rate: cell doubling time, i.e. 30 min to 2 hours.

Distribution mRNA half-lives

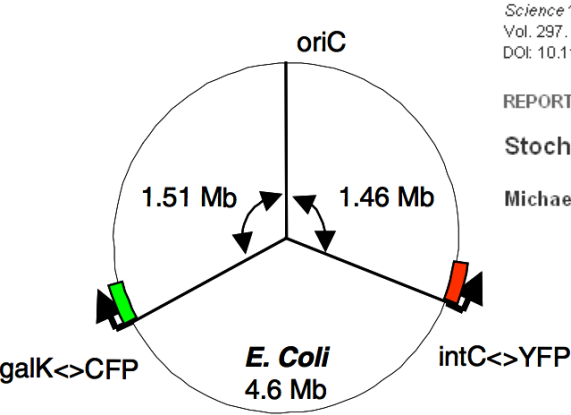


Bernstein et al, PNAS (2002)

Distribution mean proteins per cell



Taniguchi et al, Science (2002)



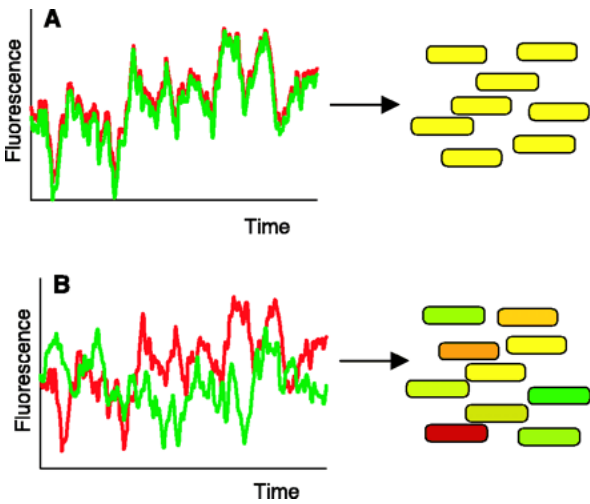
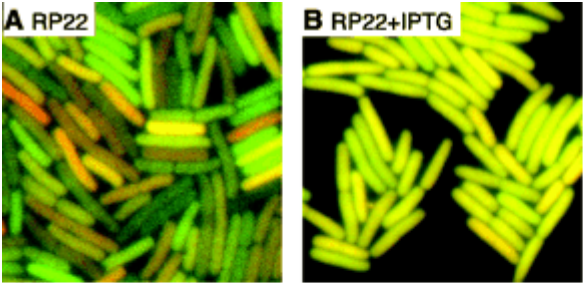
Two times the same promoter

Science 16 August 2002:
 Vol. 297, no. 5584, pp. 1183 - 1186
 DOI: 10.1126/science.1070919

REPORTS

Stochastic Gene Expression in a Single Cell

Michael B. Elowitz,^{1,2*} Arnold J. Levine,¹ Eric D. Siggia,² Peter S. Swain²



Intrinsic and extrinsic noise

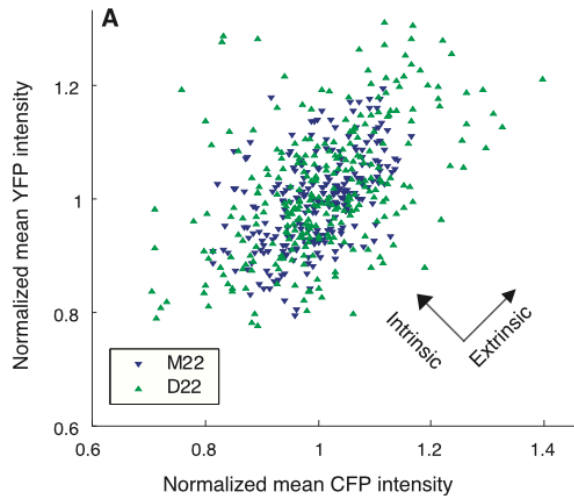
- Total variance in fluorescence per cell can be decomposed into two parts:

$$v_{tot} = \text{var}(g) + \text{var}(r) = v_i + v_e$$

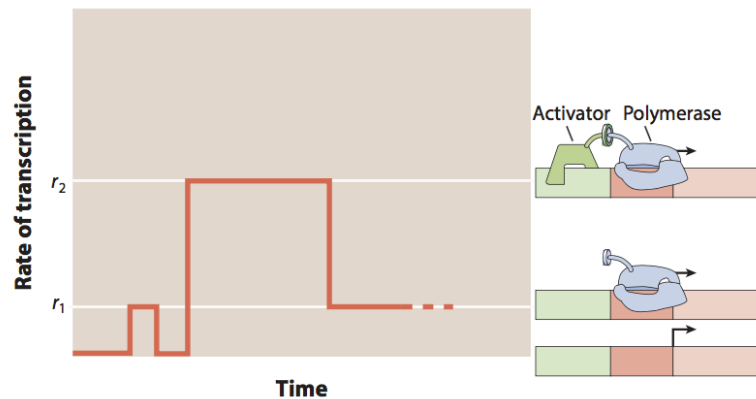
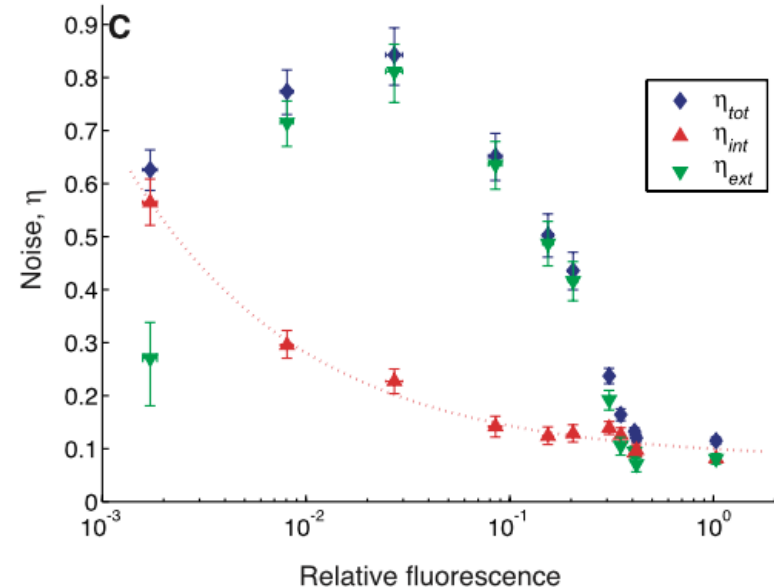
- Intrinsic = variance within cell: $v_i = \frac{1}{2} \langle (g - r)^2 \rangle$
- Extrinsic variance = the rest, i.e. variability across cells: $v_e = \langle gr \rangle - \langle g \rangle \langle r \rangle$

Extrinsic noise implies transcription/translation/decay rates fluctuate

Extrinsic noise in Elowitz et al:



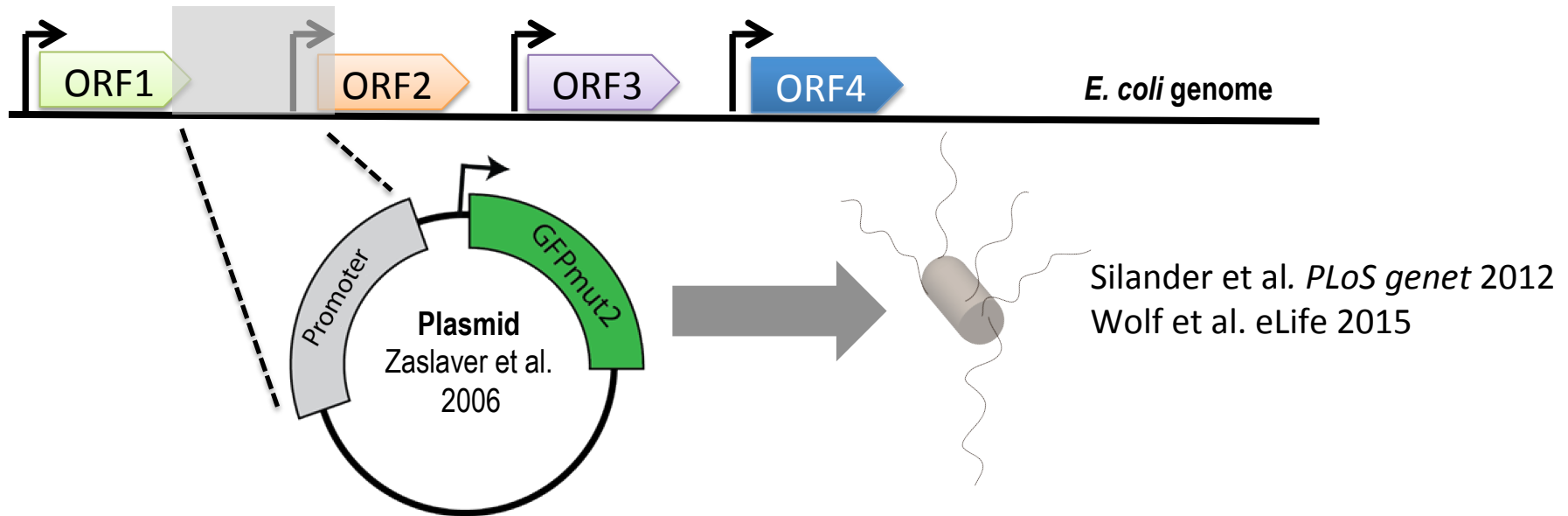
Intrinsic noise falls as the promoter is induced.
 Extrinsic noise peaks at intermediate induction.



R Phillips (Annu Rev Con Mat Phys, 2015)

- Transcription rate can vary when the promoter switches between different states.
- Switching rates depend on concentrations of DNA binding proteins (polymerases, TFs).
- These concentrations will fluctuate from cell to cell.

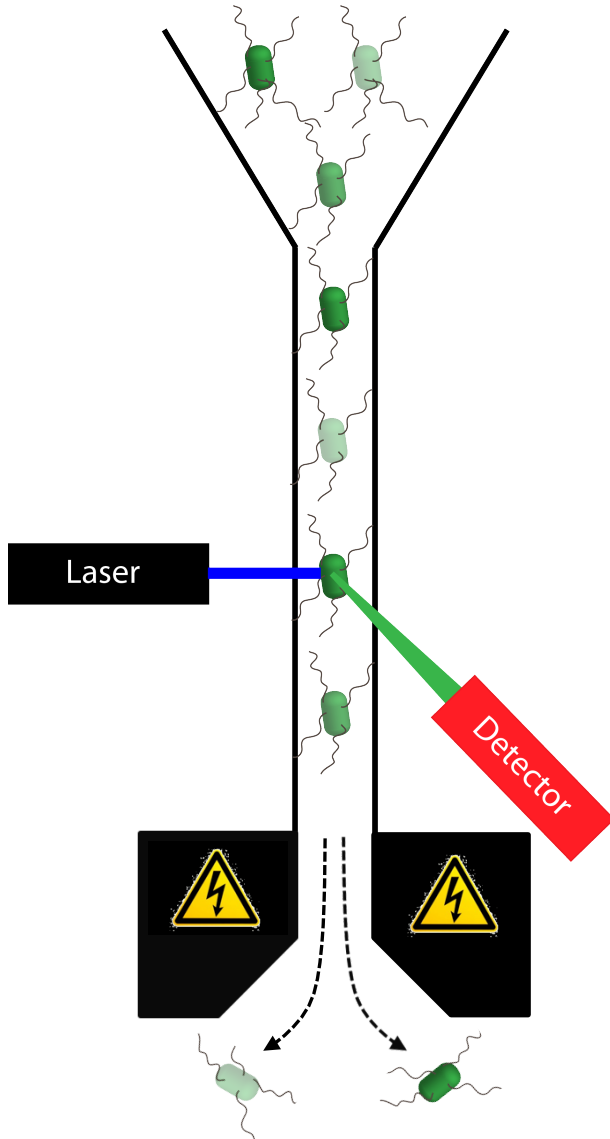
Measuring transcription from all *E. coli* promoters in single cells



- GFP fluorescence per cell proportional to protein number.
- GFP levels of single cells can be **measured in high-throughput using FACS**.
- Quantitatively characterize the distribution of expression levels across single cells, for all *E. coli* promoters.

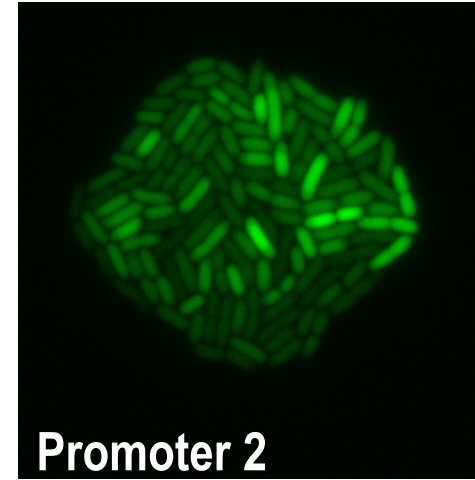
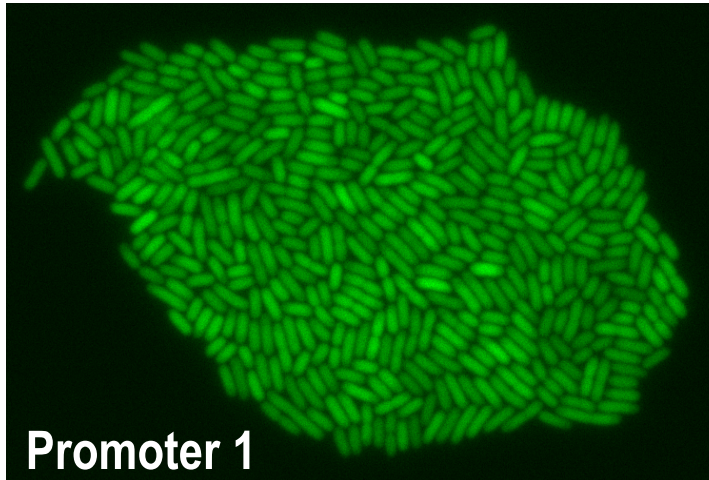
FACS:

Measuring and selecting single cells

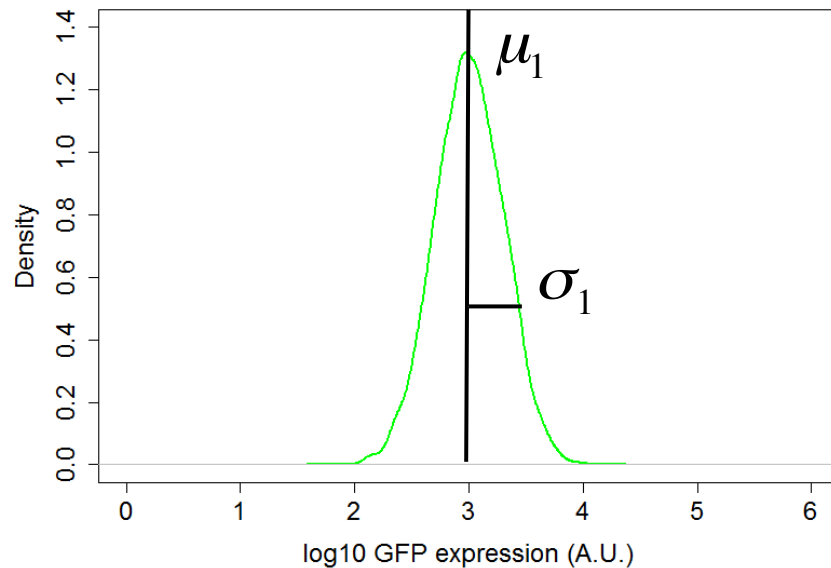


- Cells move one-by-one in a flow channel.
- Each cell passes in front of a laser and its fluorescence is measured.
- By selectively charging particles based on their measured fluorescence, one can select cells whose fluorescence lies in a certain range.

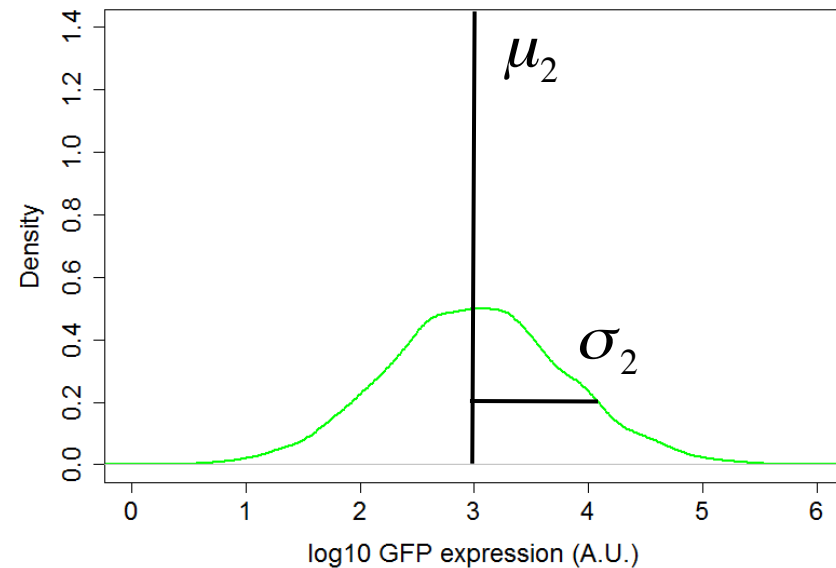
Gene expression distributions for two example promoters



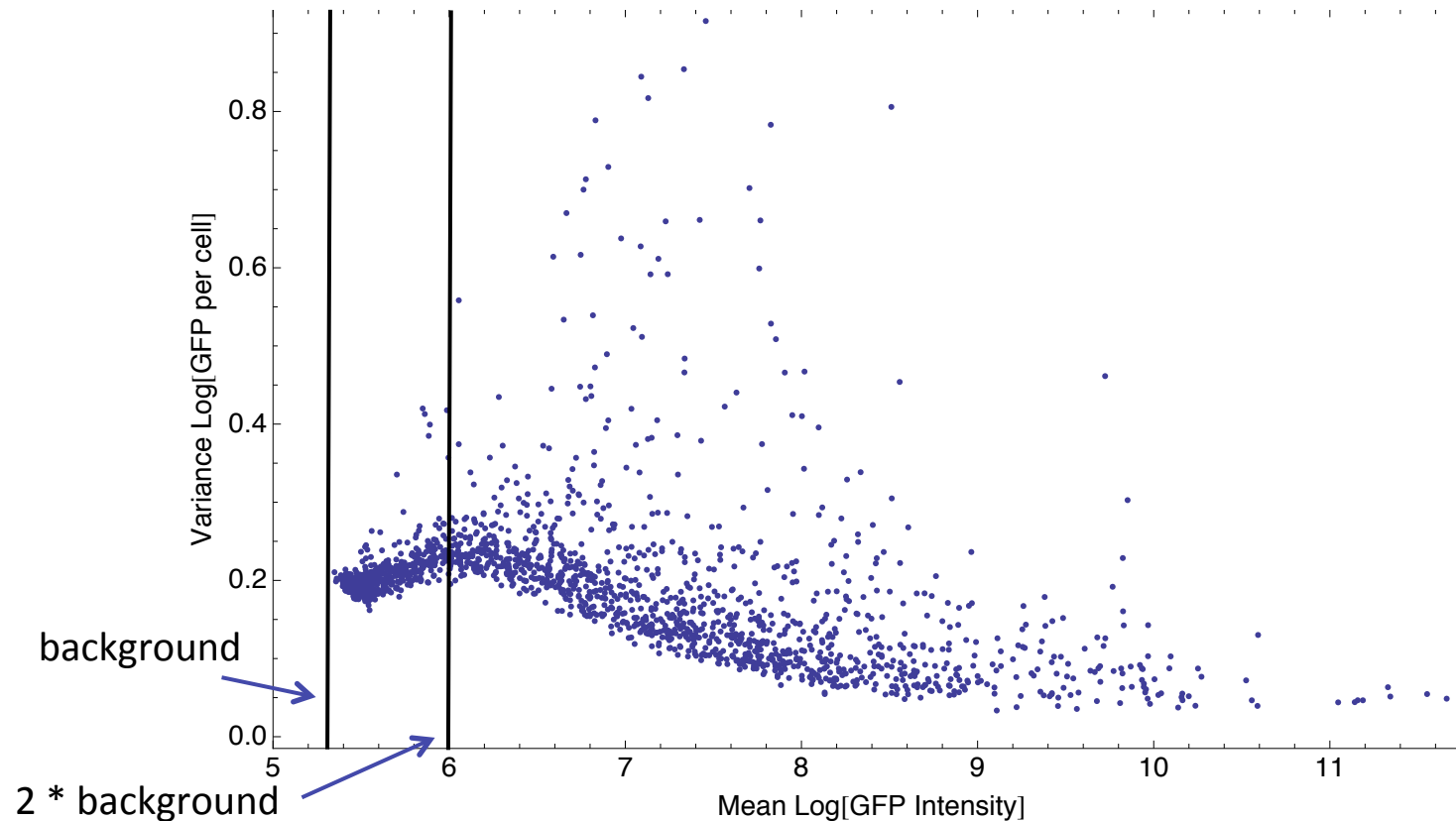
Distribution of GFP expression levels



Distribution of GFP expression levels

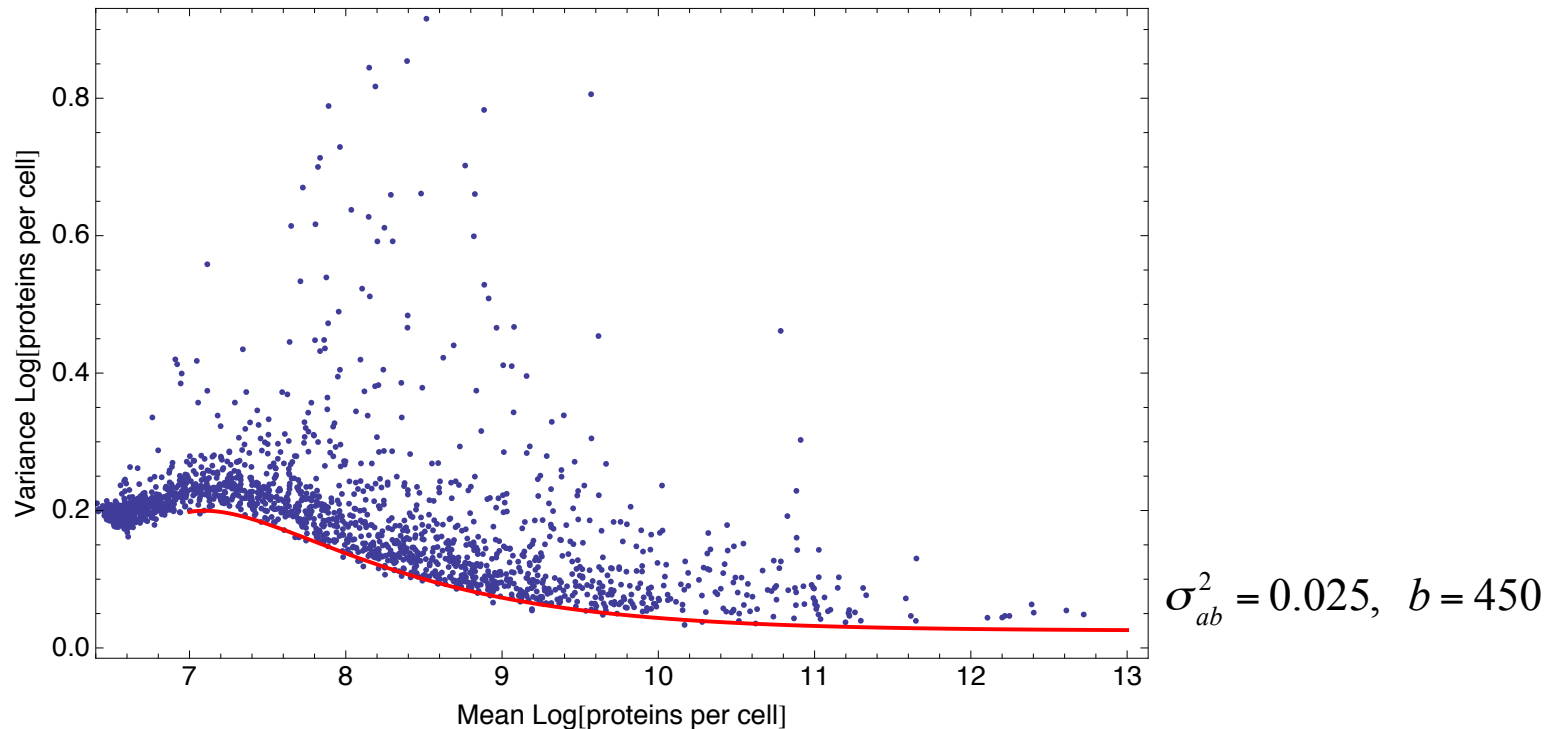


Means and variances of native *E. coli* promoters



- Variance in log-expression shows a trend of decreasing with mean expression.
- Different promoters with same mean can show significantly different variance.
- There seems to be a clear lower bound on variance as a function of mean.

Means and variances of native E. coli promoters



At constant transcription/translation/decay rates: $\langle n \rangle = ab$, $\text{var}(n) = (b+1)\langle n \rangle$

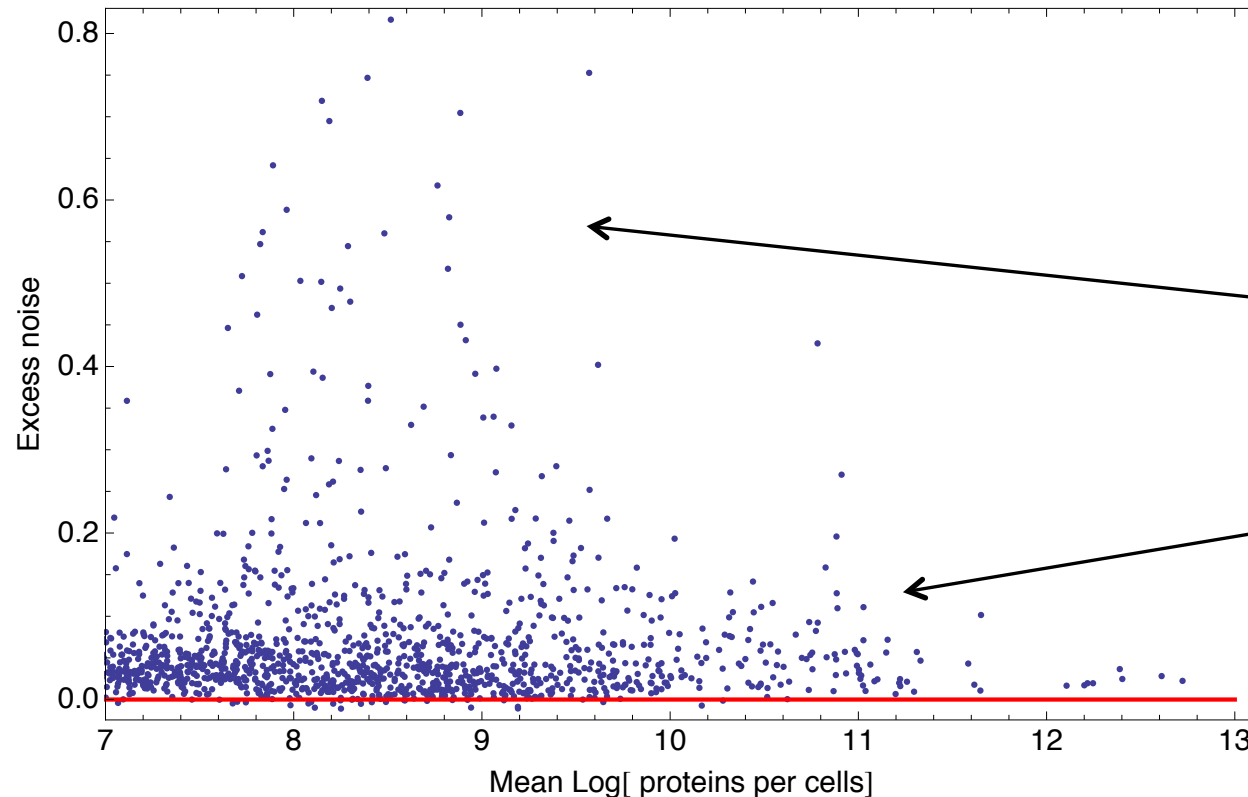
Assume a and b both fluctuate: $\text{var}(n) = (b+1)\langle n \rangle + \sigma_{ab}^2 \langle n \rangle^2$

Red curve:

$$n_{meas} = n_{bg} + \langle n \rangle + \varepsilon \sqrt{\text{var}(n)} \quad \Longrightarrow \quad \text{var}[\log(n_{meas})] = \sigma_{ab}^2 \left(1 - \frac{n_{bg}}{\langle n_{meas} \rangle} \right)^2 + \frac{(b+1)}{\langle n_{meas} \rangle} \left(1 - \frac{n_{bg}}{\langle n_{meas} \rangle} \right)$$

Noise levels vary across native *E. coli* promoters

Excess noise (variance – lower bound as func. mean)



Selection on noise levels

High noise

Drift? Selected for noise?

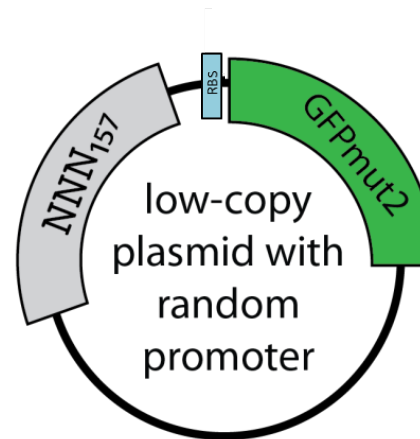
Low noise.

Selection to minimize noise?

What noise would one get without selection?

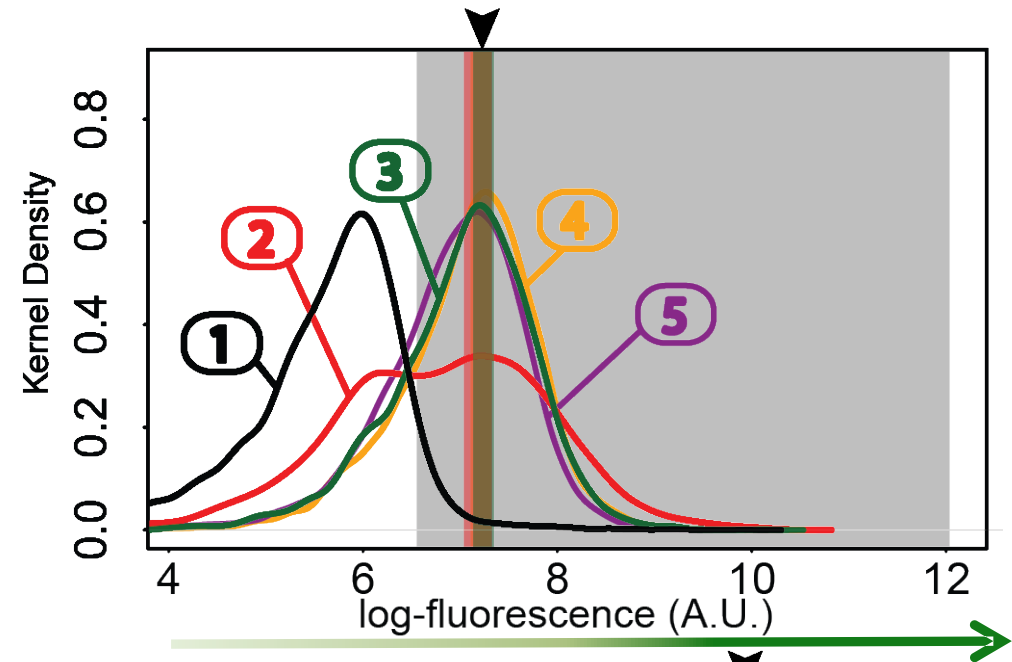
Evolve *synthetic promoters* in a precisely controlled selective environment.

Directed evolution of promoters that express at a desired level

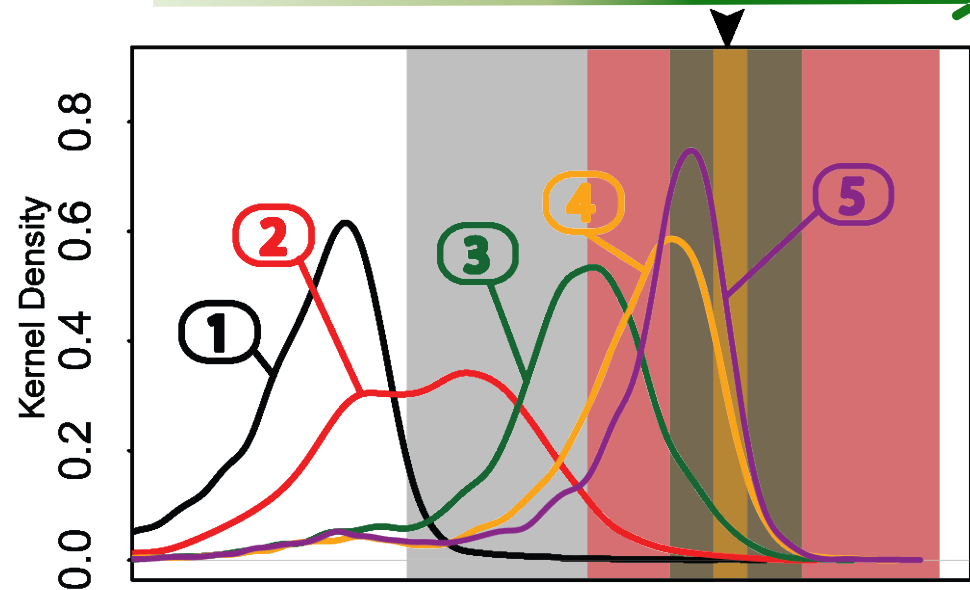


Evolution of population expression levels

Selecting for
Medium expression

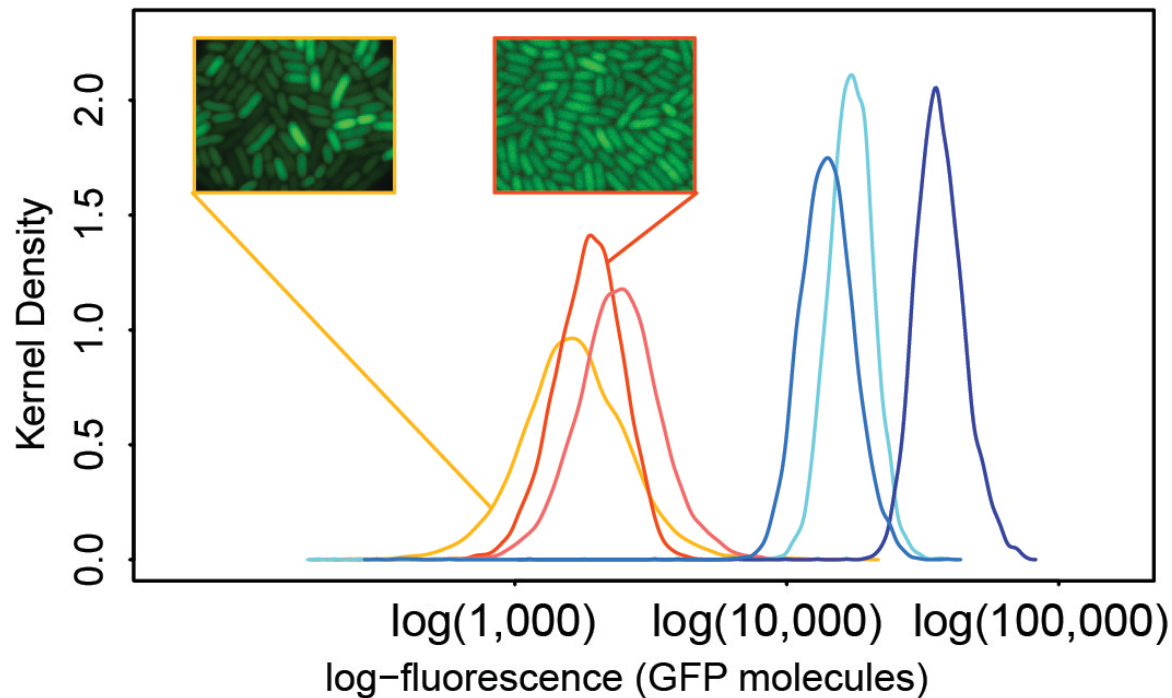


Selecting for
High expression



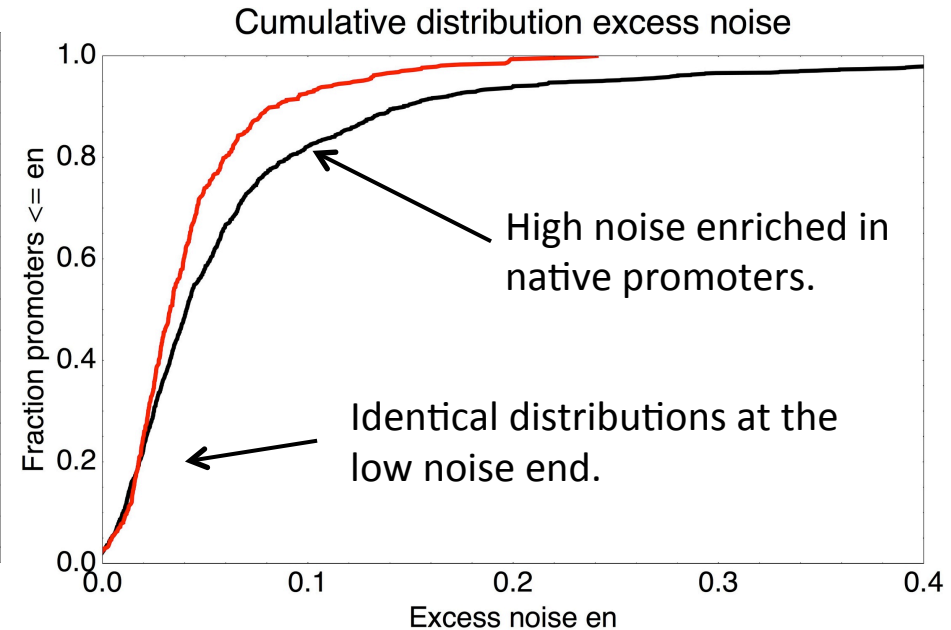
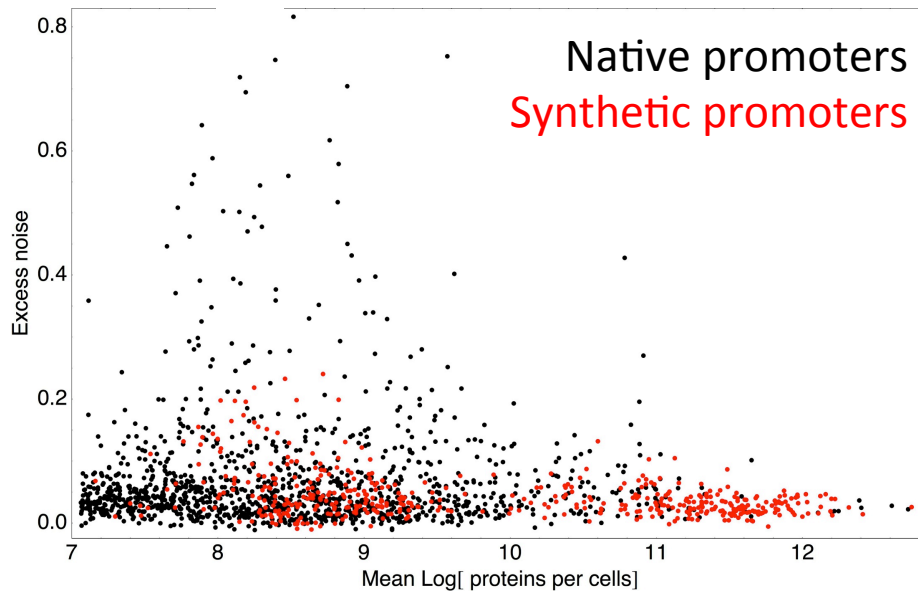
Expression distributions of individual synthetic promoters

- We isolated ~400 clones from evolutionary runs for both **medium** and **high** expression.
- Measured each clone's expression distribution.



How do noise levels of synthetic promoters compare with those of native promoters?

Selection caused increased noise in a substantial fraction native promoters



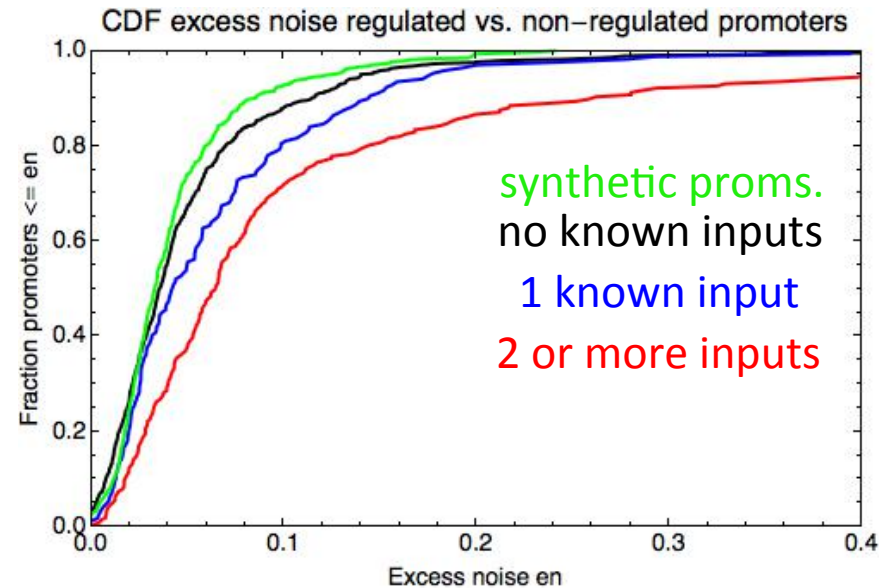
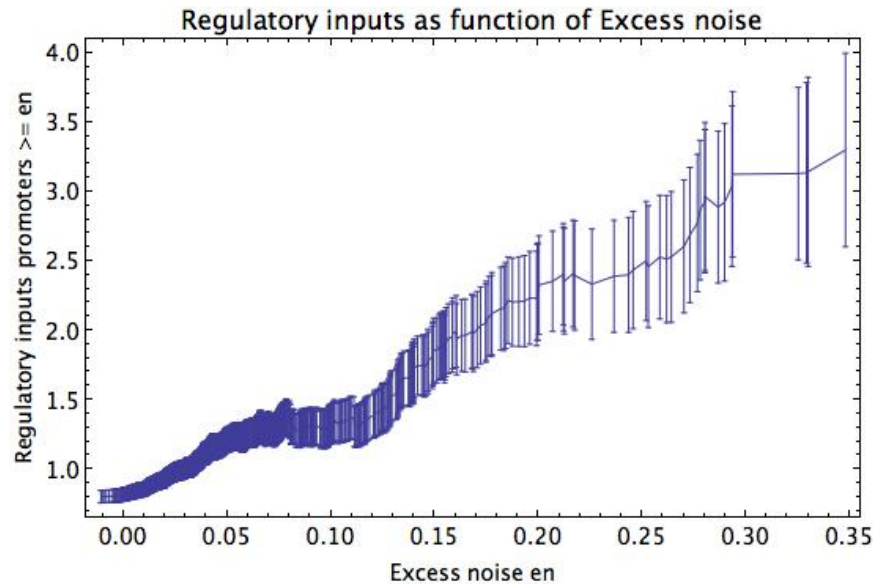
- Synthetic promoters were **not** selected on their noise properties.
- Low noise is the **default behavior** of E. coli promoters.
- Selection must have acted so as to increase the noise levels of some native promoters.

What is 'special' about native promoters that show high noise?

Noisy genes have more regulatory inputs



- 185 *E. coli* transcription factors (TFs).
- 4123 known regulatory interactions TF → promoter.



Genes with higher noise have (on average) higher numbers of known regulatory inputs.

Why is there a general association between noise and regulation?
Why did selection cause noise to increase?

Noise-propagation: nuisance or opportunity?

Noise as an unavoidable side-effect of regulation

- Explains the general association of noise and regulation.
- ‘Fluctuation-dissipation relation’: Genes that need complex regulation unavoidably couple to the noise in their regulators.
- Generally assumed to be *detrimental*: reduces the accuracy of regulation.

[Cell Mol Life Sci.](#) 2011 Mar;68(6):1005-10. doi: 10.1007/s00018-010-0589-y. Epub 2010 Nov 30.

Fluctuation and response in biology.

[Lehner B¹](#), [Kaneko K.](#)

Stochasticity as a bet-hedging strategy

- Phenotypic diversity can generally be selected for in fluctuating environments.

Evolution of Phenotypic Variance

J. J. Bull

Evolution, Vol. 41, No. 2 (Mar., 1987), pp. 303-315

[Science](#). 2005 Sep 23;309(5743):2075-8. Epub 2005 Aug 25.

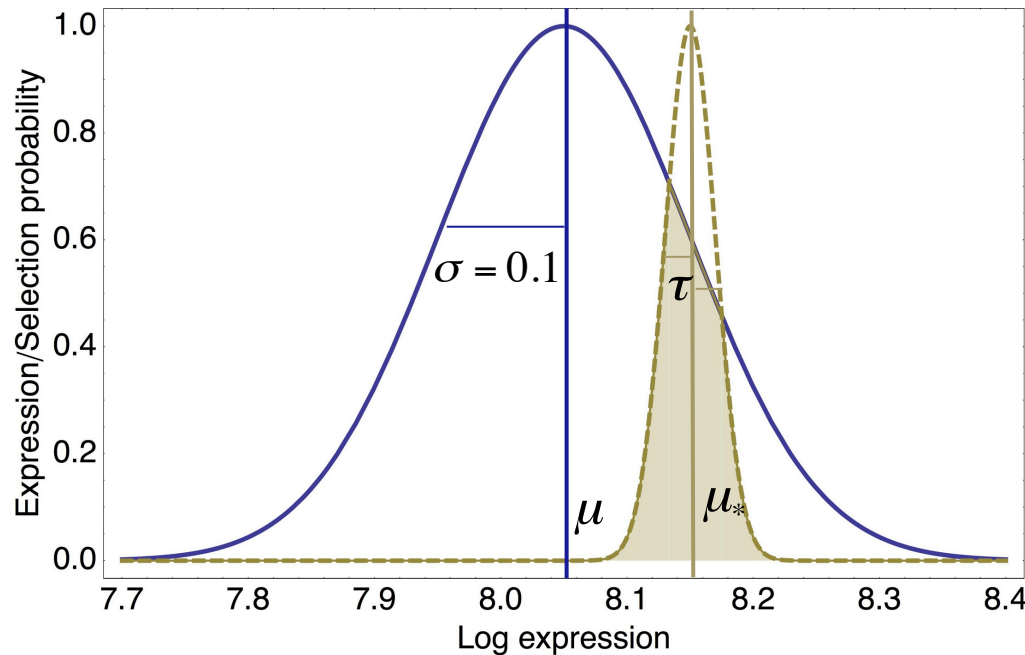
Phenotypic diversity, population growth, and information in fluctuating environments.

[Kussell E¹](#), [Leibler S.](#)

- Maybe noise-propagation can be *beneficial* in some circumstances?

Let's do some theory on how gene expression noise affects fitness

Fitness function in a single environment



Promoter expression distribution:

$$p(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Fitness (probability to be selected):

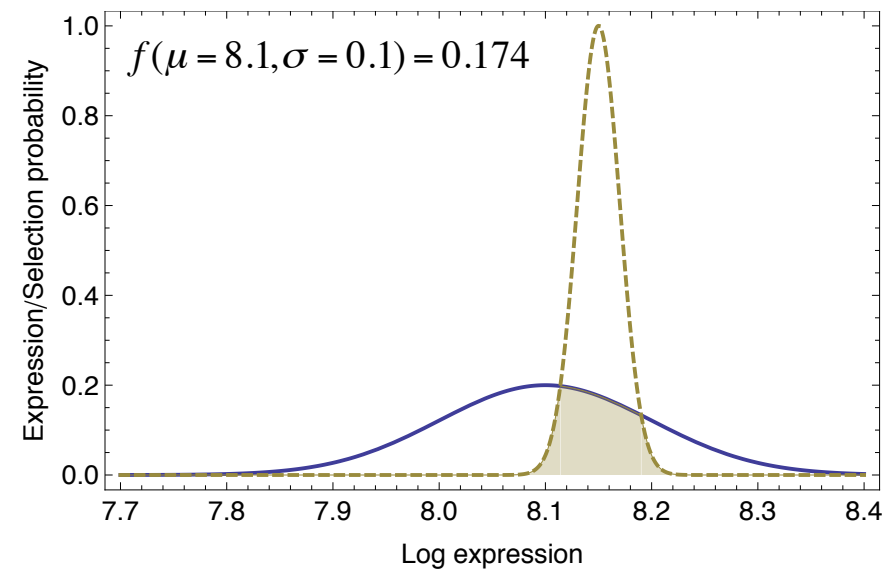
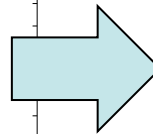
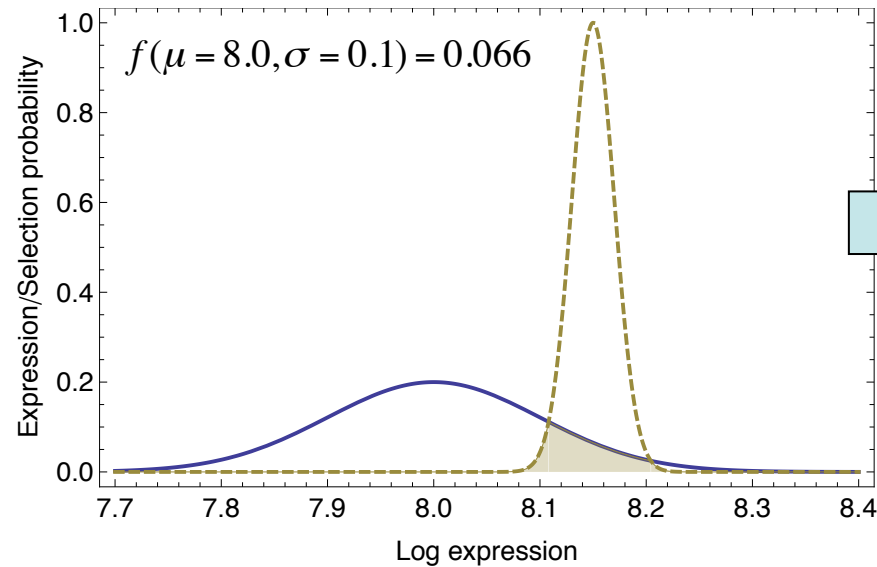
$$f(x | \mu_*, \tau) = \exp\left(-\frac{(x - \mu_*)^2}{2\tau^2}\right)$$

The fitness of a promoter 'genotype' (fraction of its cells selected) is a convolution of these two functions (approx. area on the intersection):

$$f(\mu, \sigma | \mu_*, \tau) = \int dx p(x | \mu, \sigma) f(x | \mu_*, \tau) = \sqrt{\frac{\tau^2}{\tau^2 + \sigma^2}} \exp\left(-\frac{(\mu - \mu_*)^2}{2(\tau^2 + \sigma^2)}\right)$$

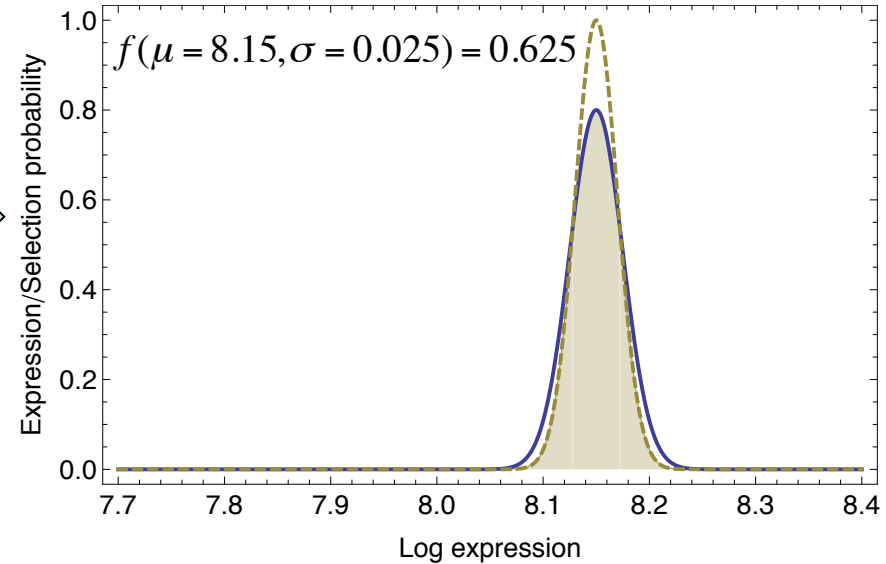
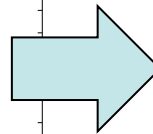
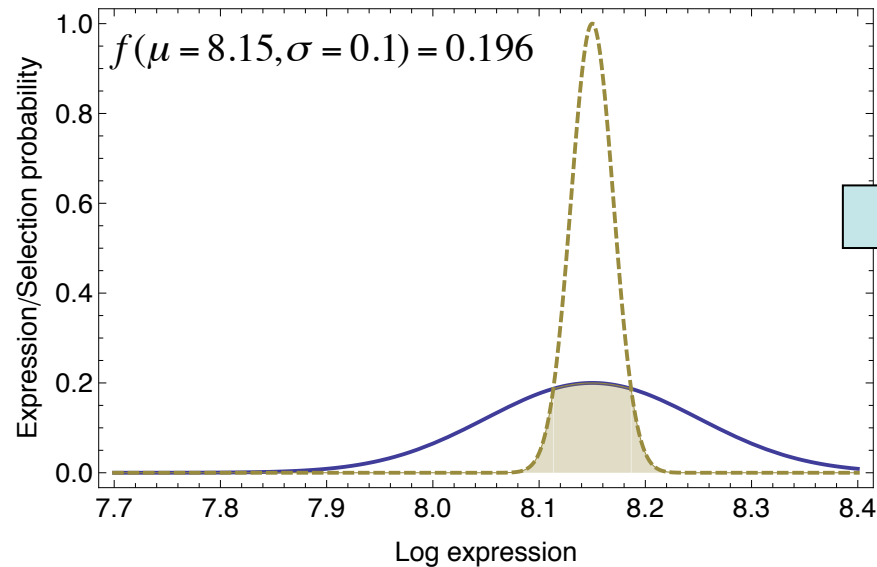
Moving the mean toward the desired level always increases fitness

$$f(\mu, \sigma | \mu_*, \tau) = \sqrt{\frac{\tau^2}{\tau^2 + \sigma^2}} \exp\left(-\frac{(\mu - \mu_*)^2}{2(\tau^2 + \sigma^2)}\right)$$

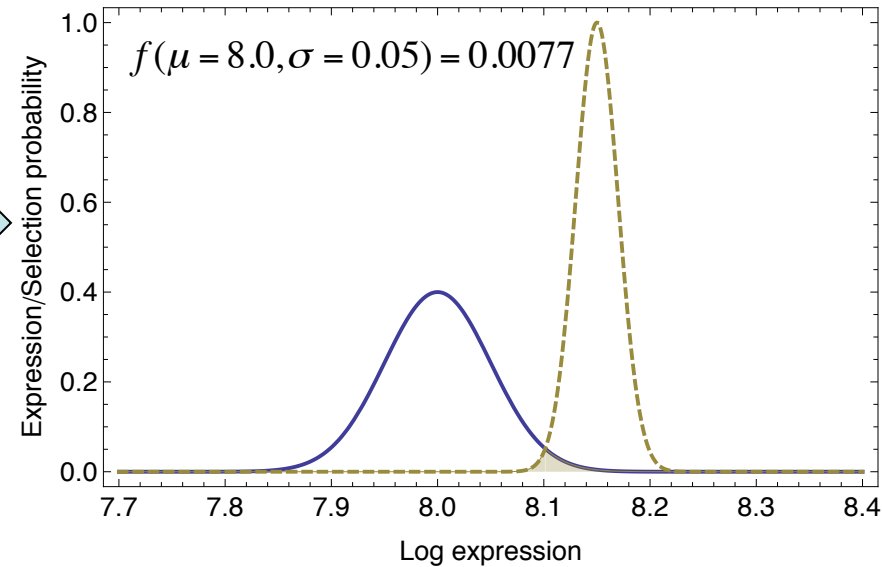
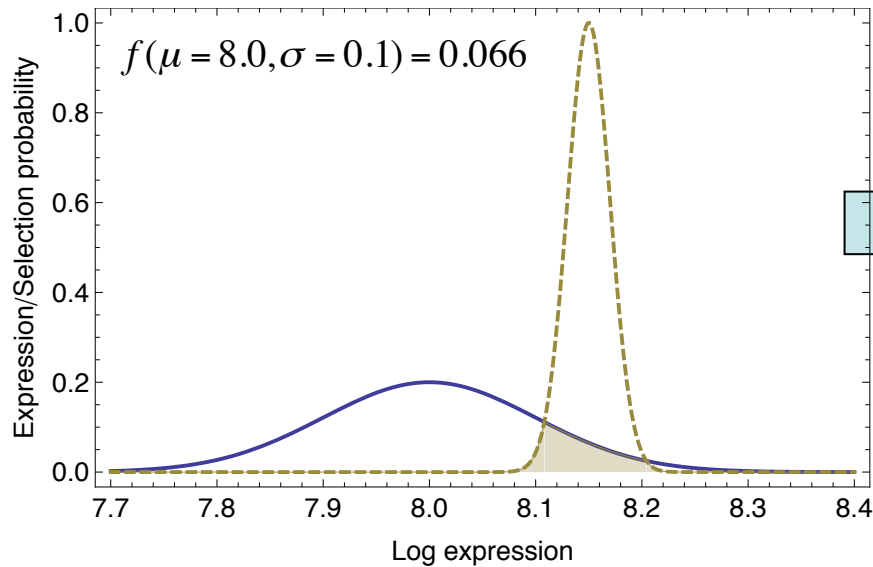


At optimal mean minimal noise is preferred

$$f(\mu, \sigma | \mu_*, \tau) = \sqrt{\frac{\tau^2}{\tau^2 + \sigma^2}} \exp\left(-\frac{(\mu - \mu_*)^2}{2(\tau^2 + \sigma^2)}\right)$$



As mean moves away from the optimum there is a bifurcation to nonzero optimal noise

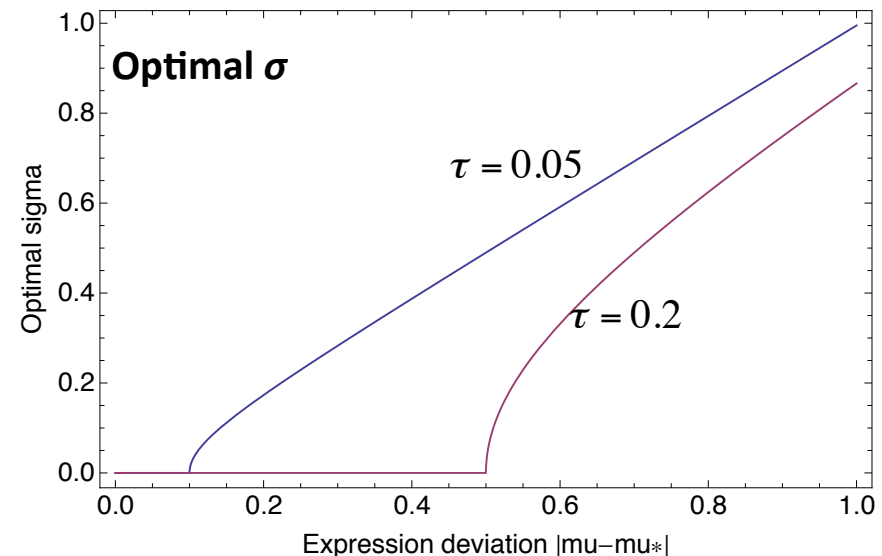


$$f(\mu, \sigma | \mu_*, \tau) = \sqrt{\frac{\tau^2}{\tau^2 + \sigma^2}} \exp\left(-\frac{(\mu - \mu_*)^2}{2(\tau^2 + \sigma^2)}\right)$$

'Bifurcation' in optimal σ

When $|\mu - \mu_*| \geq \tau$, the optimal noise level is non-zero:

$$\sigma_* = \sqrt{(\mu - \mu_*)^2 - \tau^2}$$



Variable environment: Fitness of an unregulated gene

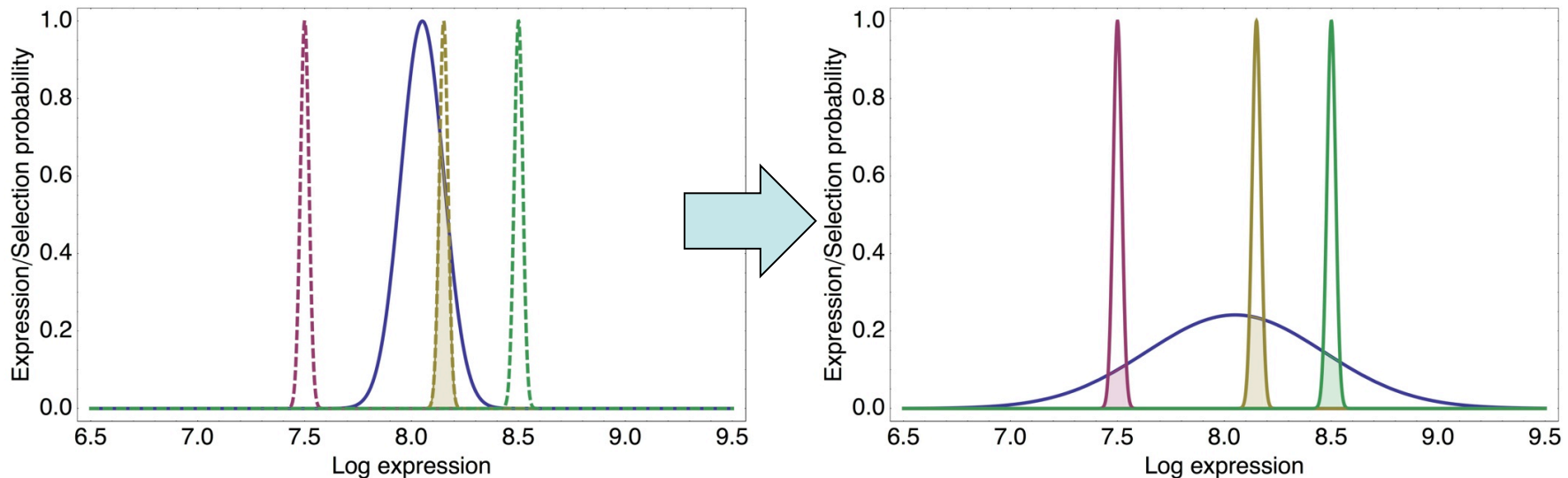
Log-fitness in a variable environment: $\log[f(\mu, \sigma)] = -\frac{\langle (\mu - \mu_e)^2 \rangle}{2(\tau^2 + \sigma^2)} + \frac{1}{2} \log \left[\frac{\tau^2}{\tau^2 + \sigma^2} \right]$

Assuming no regulation, optimal mean equals $\mu = \langle \mu_e \rangle$

Log-fitness becomes:

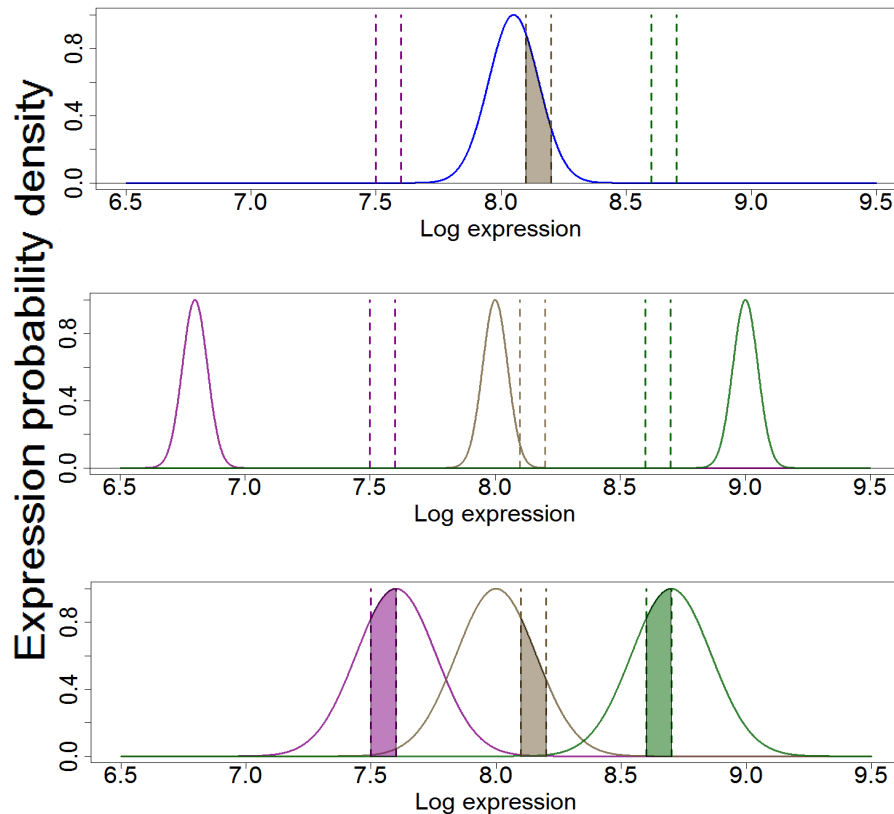
$$\log[f(\mu, \sigma)] = -\frac{\text{var}(\mu_e)}{2(\tau^2 + \sigma^2)} + \frac{1}{2} \log \left[\frac{\tau^2}{\tau^2 + \sigma^2} \right]$$

Optimal noise matches the variation in desired expression levels: $\sigma_{\text{opt}}^2 = \text{var}(\mu_e) - \tau^2$



This is the **bet hedging** scenario. But:
 Wouldn't it be better to evolve gene regulation?

Effects of coupling a gene to a regulator



Gene without regulation



Regulator's activity



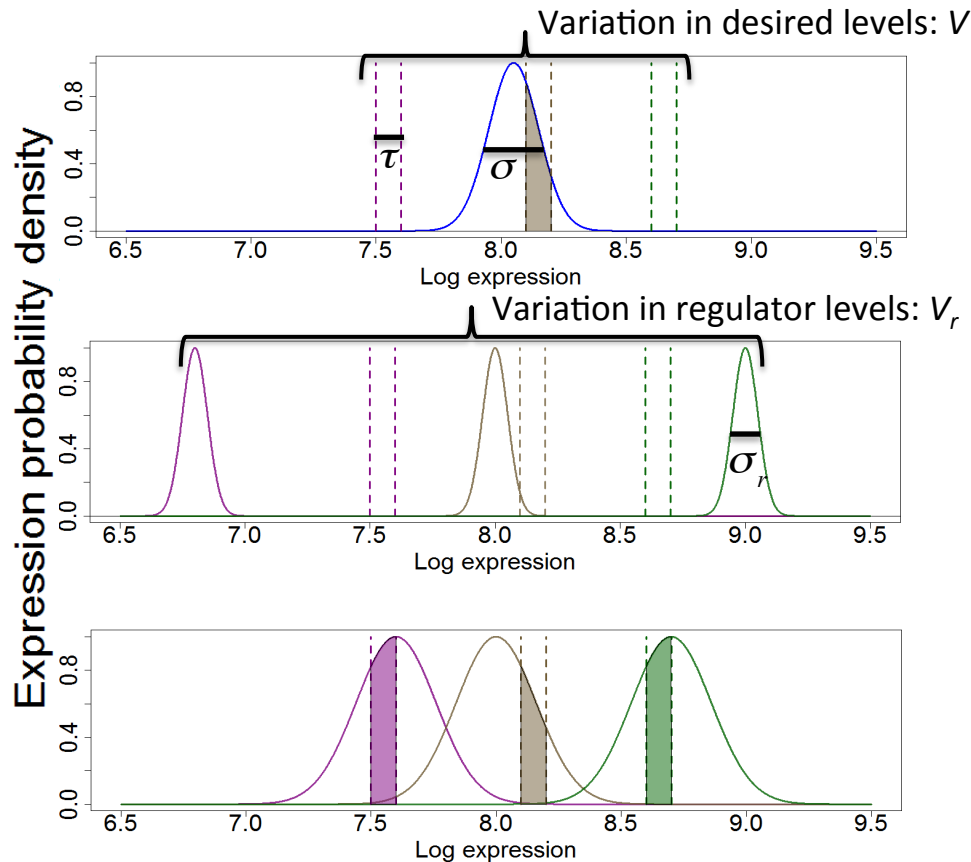
Gene coupled to the regulator.

Two main effects on the gene's expression:

1. **Condition-response:** Mean depends on regulator's (condition-dependent) activity.
2. **Noise-propagation:** Noise increases due to propagation of the regulator's noise.

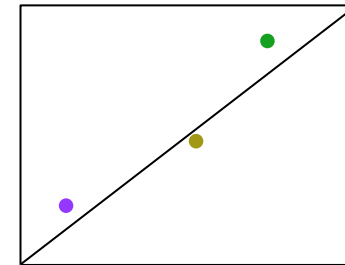
We developed a general theory to calculate how these effects conspire to affect fitness.

Fitness depends on only 4 effective parameters



1. Expression mismatch: $Y^2 = \frac{V}{\sigma^2 + \tau^2}$

2. Signal-to-noise of the regulator: $S^2 = \frac{V_r}{\sigma_r^2}$



3. Correlation regulator/desired levels: R

4. Coupling strength: X

Fitness effect of the regulatory interaction:

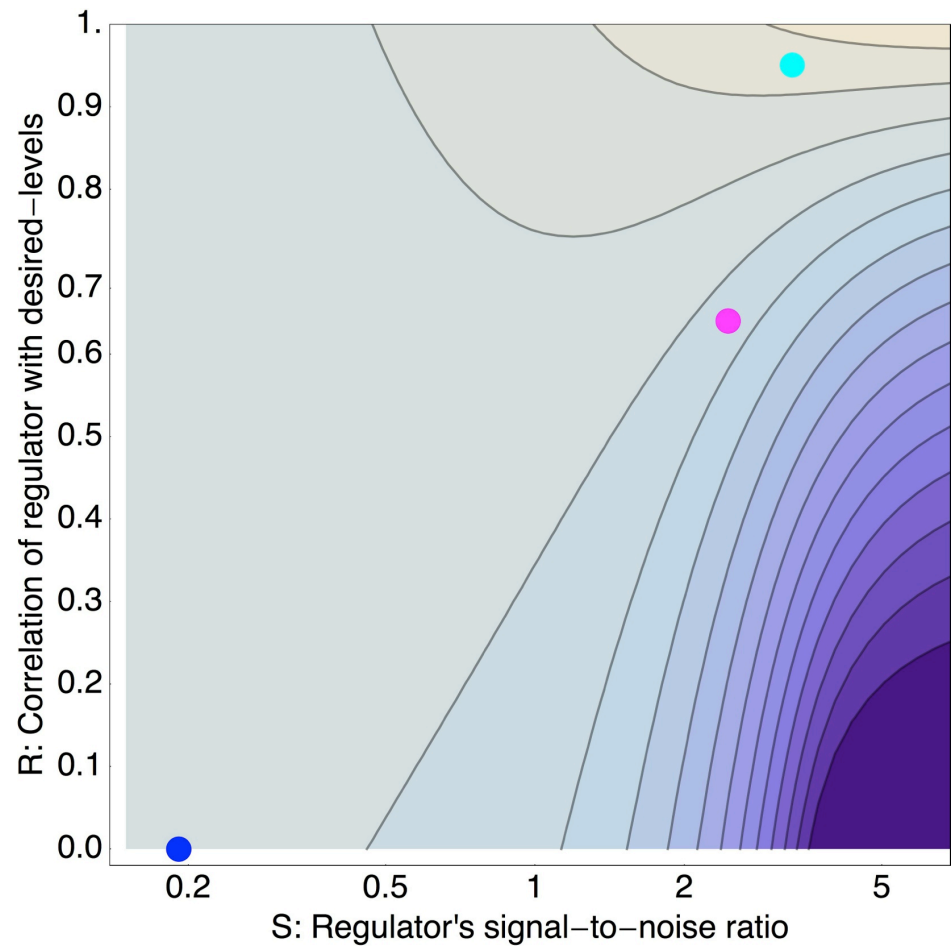
$$\log[f] = -\frac{1}{2} \frac{Y^2(1-R^2) + (SX - RY)^2}{(1+X^2)} - \frac{1}{2} \log[1+X^2]$$

Scenario: Start with unregulated promoter. What fitness can be obtained by coupling to regulator with signal-to-noise S and correlation R ?

Fitness with optimal coupling to a regulator of given correlation R and signal-to-noise S

Perfect correlation

No correlation



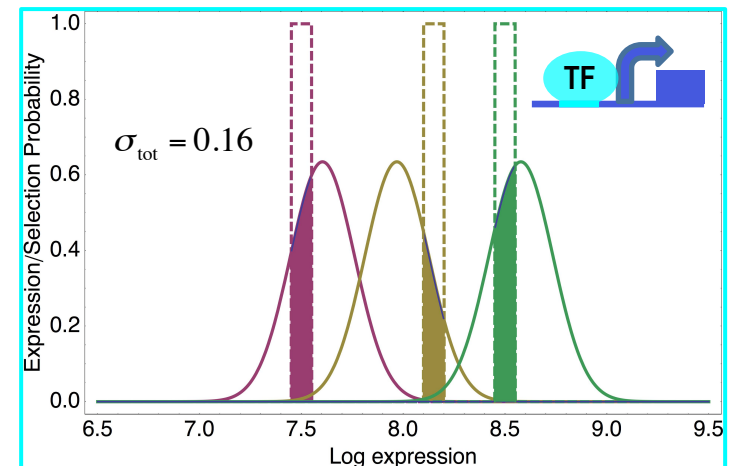
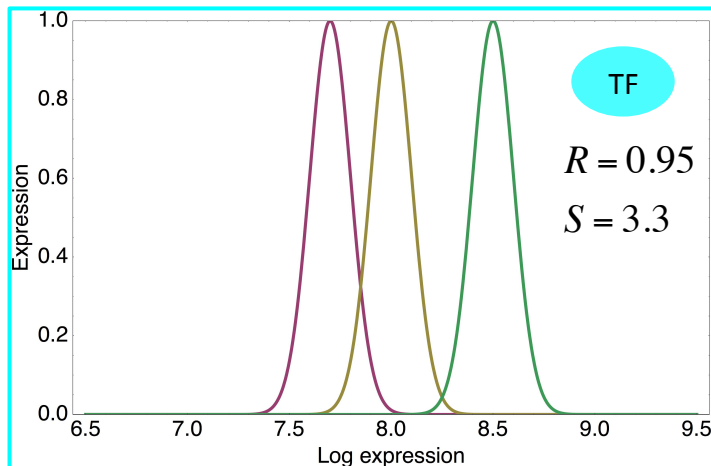
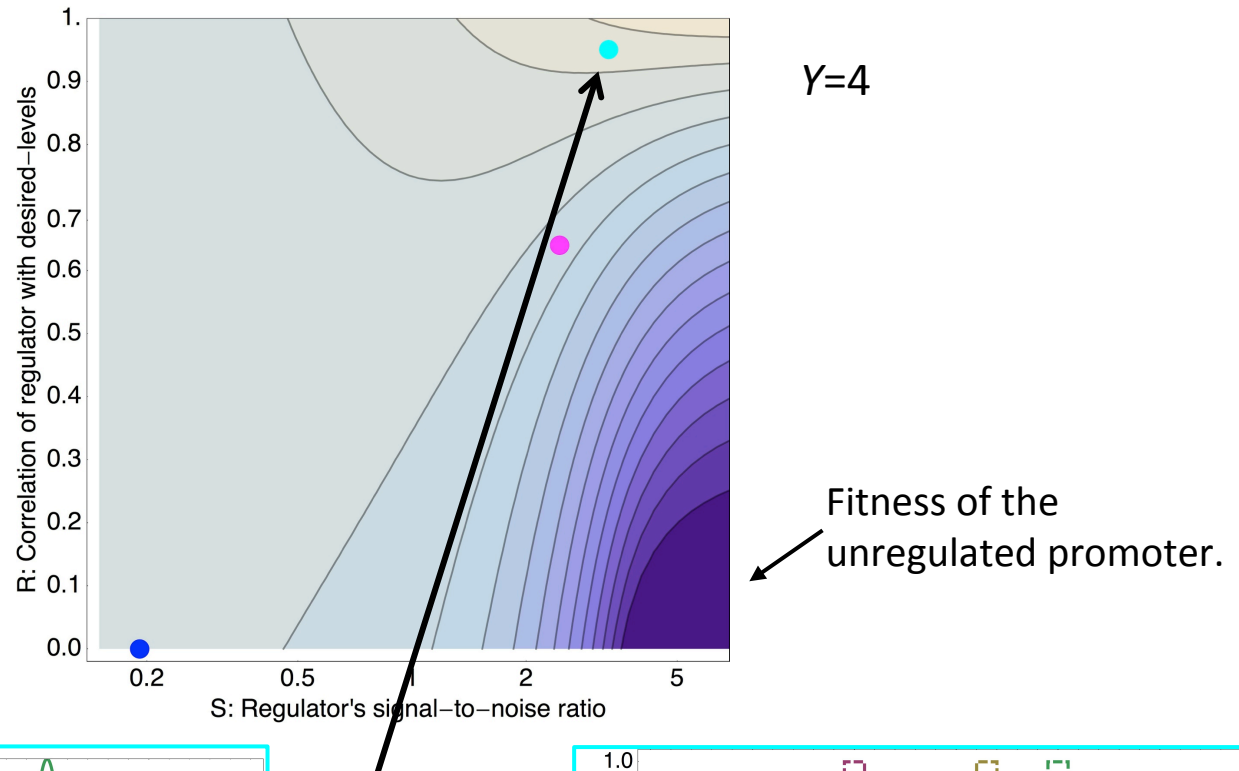
$\gamma=4$

Fitness of the unregulated promoter.

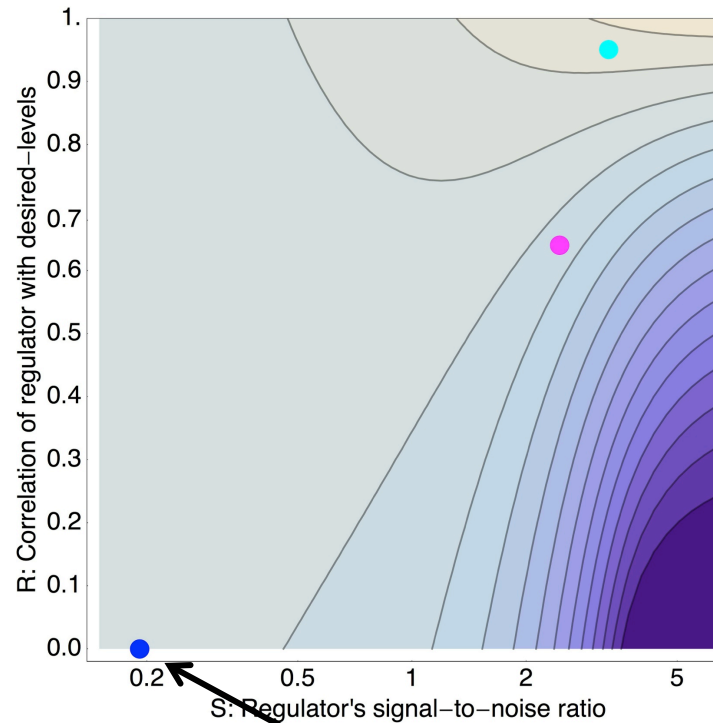
Noisy regulator

Precise regulator

Coupling to a near optimal regulator: condition-response effect

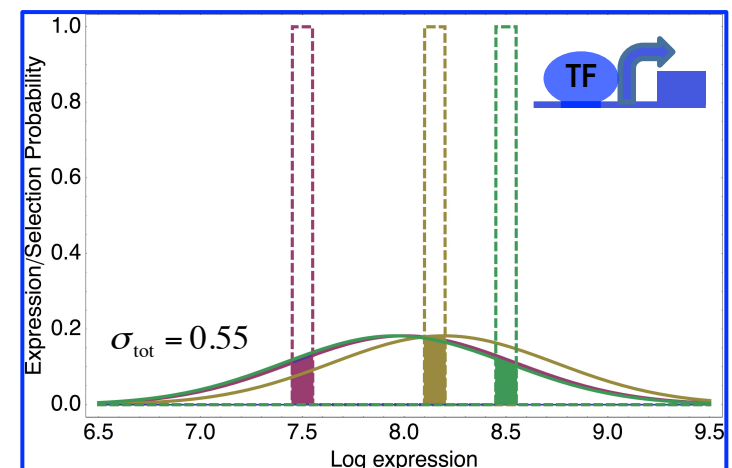
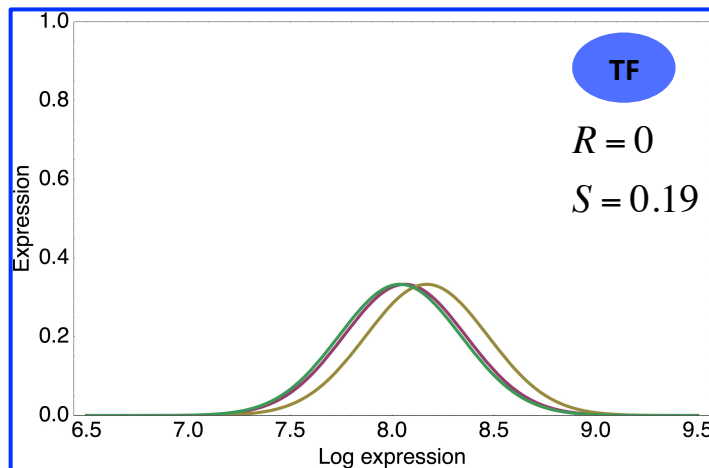


Coupling to a noisy uncorrelated regulator: noise-propagation implements bet hedging strategy

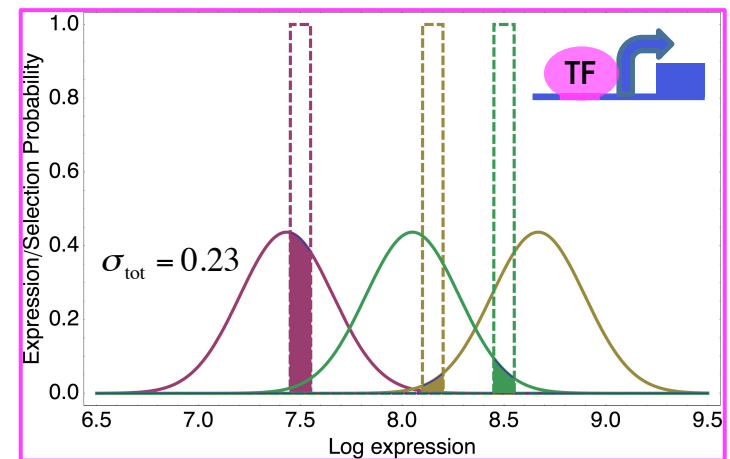
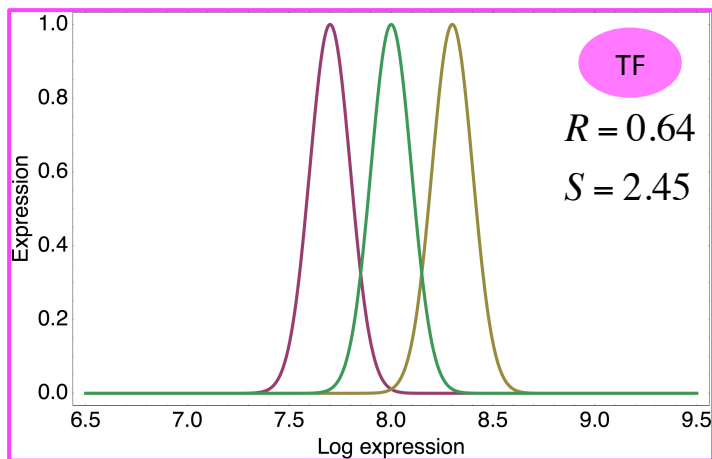
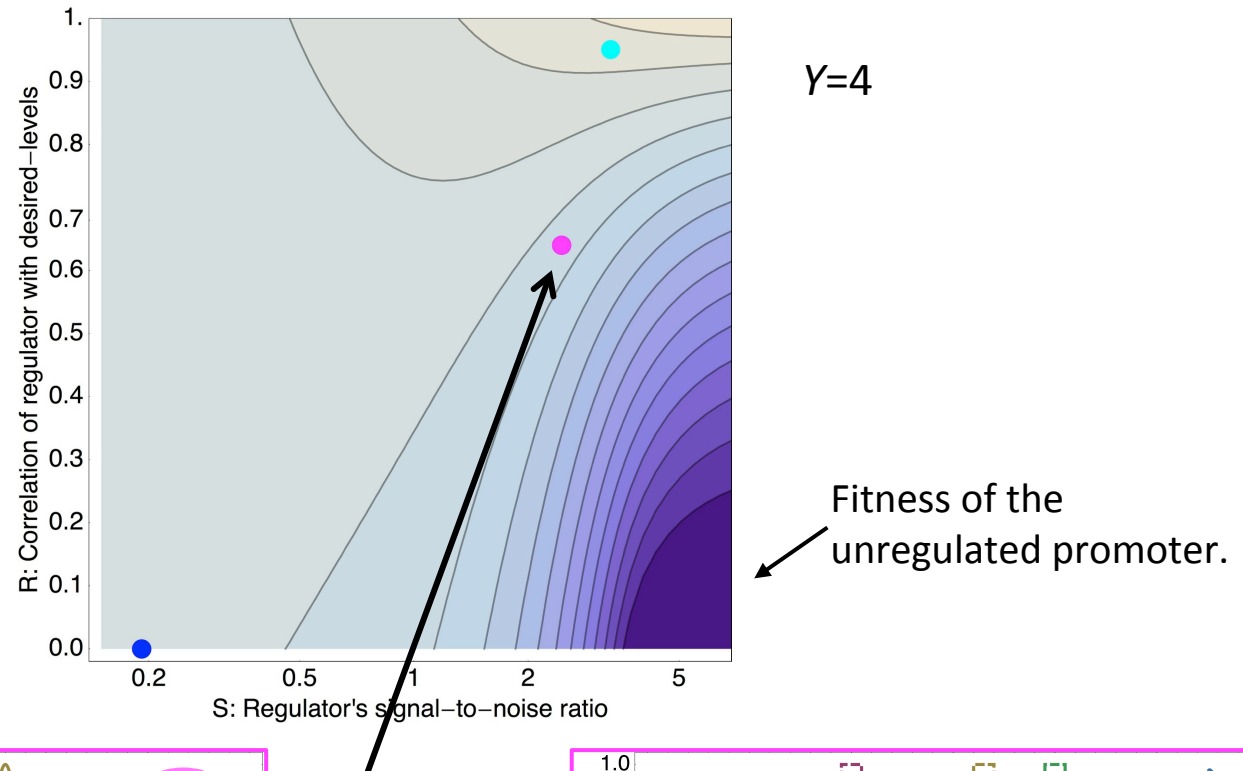


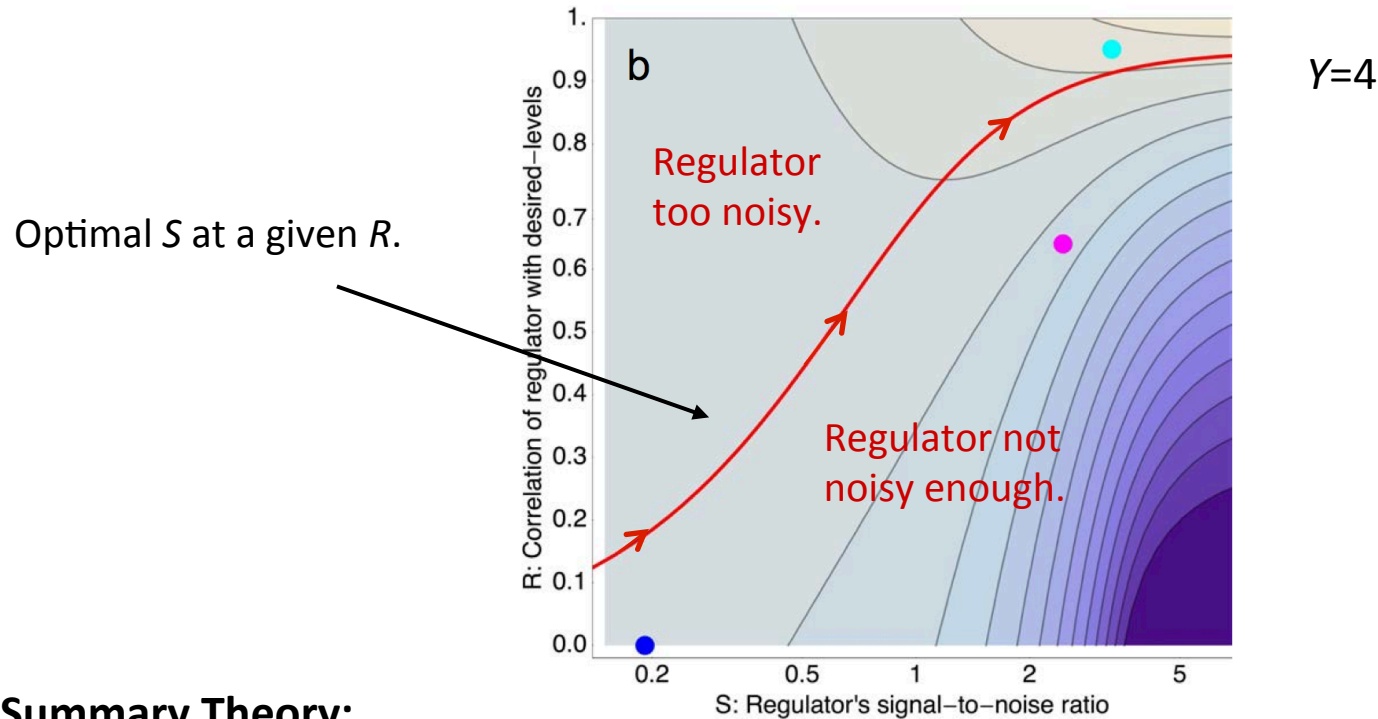
$\gamma=4$

Fitness of the
 unregulated promoter.



Intermediate case: a moderately correlated regulator

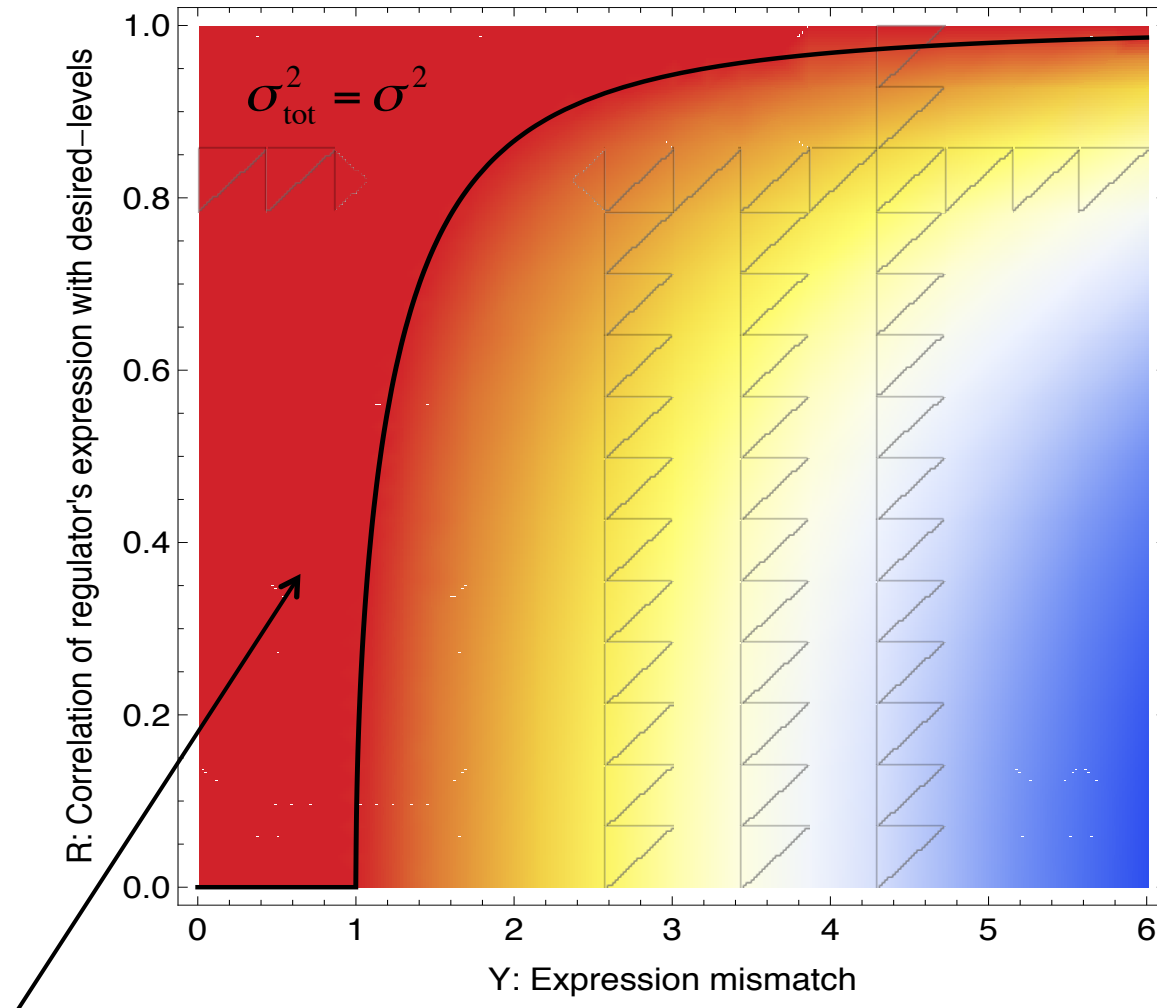




Summary Theory:

- Noise-propagation is often *functional*, acting as a rudimentary form of regulation.
- *De novo* evolution of regulation: Starting from pure noise-propagation ($R=0, S=0$) there is a continuum of solutions of increasing accuracy along which condition-response and noise-propagation optimally complement each other.
- Regulated genes are noisy because, whenever the condition-response is imperfect, maximal fitness requires noisy regulators.

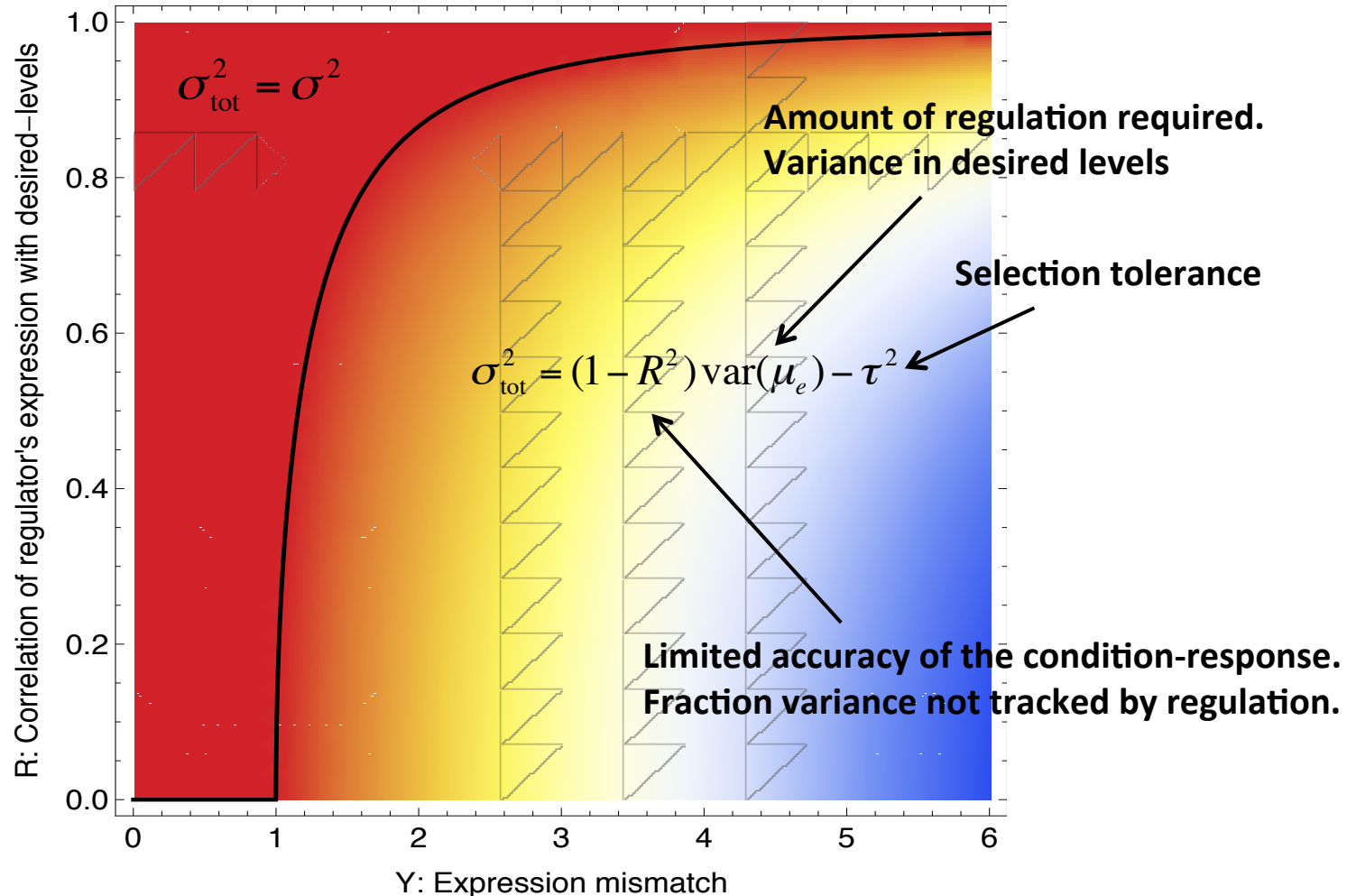
Phase diagram of final noise after coupling to regulators with optimal noise levels.



Low noise regime:

Promoters with low expression mismatch $Y < 1$ 'do not bother' to be regulated.
For extremely correlated regulators, zero noise-propagation is the optimum.

Phase diagram of final noise after coupling to regulators with optimal noise levels.



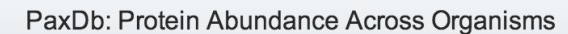
Noise-propagation regime:

The final noise level matches the fraction of variance in desired levels *not tracked* by the condition-response.

1. dN: Amino acid substitution rate.
2. dS: Synonymous site substitution rate.
3. Promoter conservation.

Biozentrum, the University of Basel, and Swiss Institute of Bioinformatics, 4056-CH, Basel, Switzerland.

4. Average protein level.
5. Average mRNA level.
6. Promoter mean expression (our data).



Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA

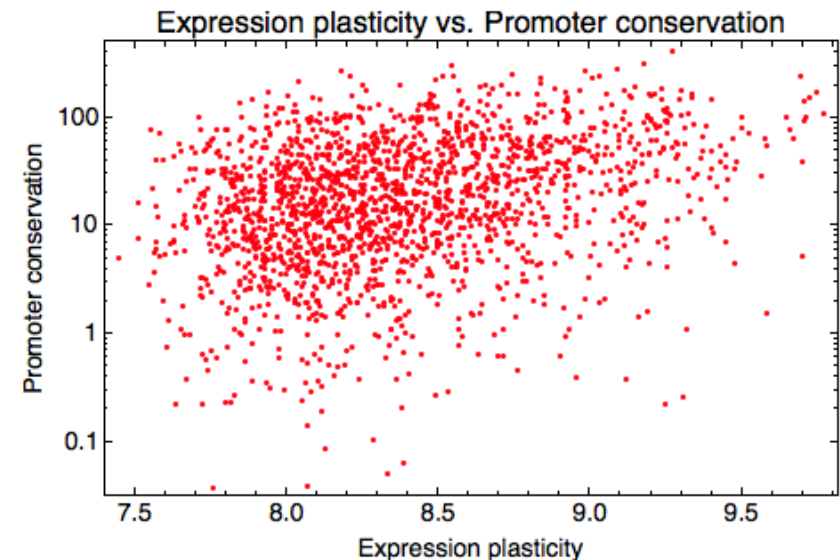
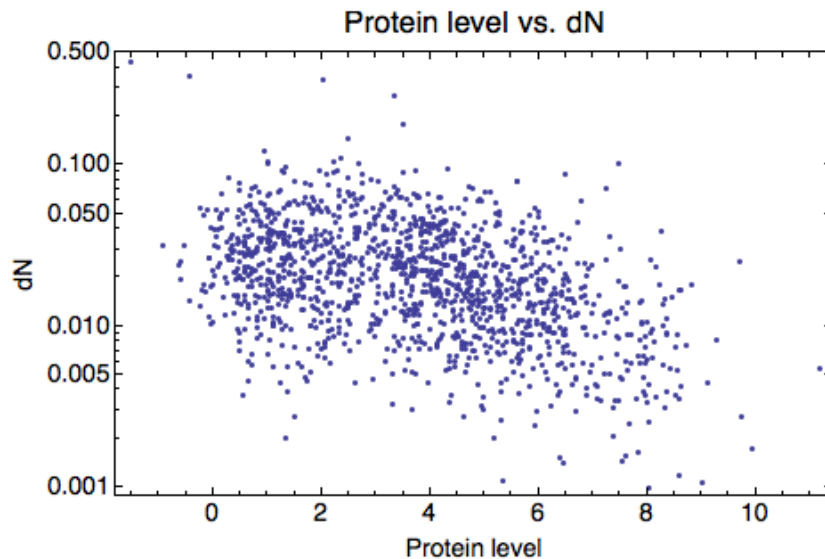
7. Number of regulatory inputs.
8. Expression plasticity.



9. Promoter excess noise (our data).

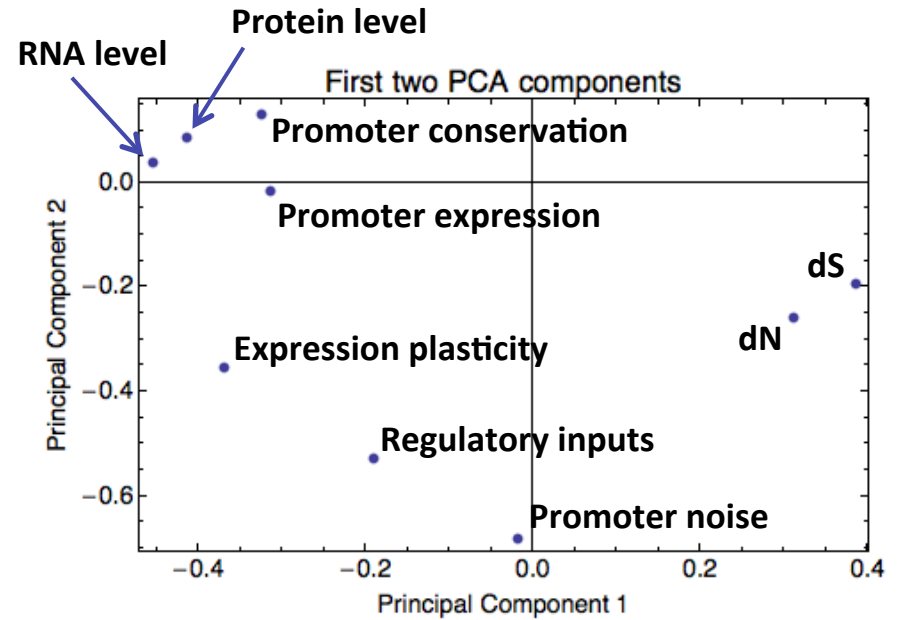
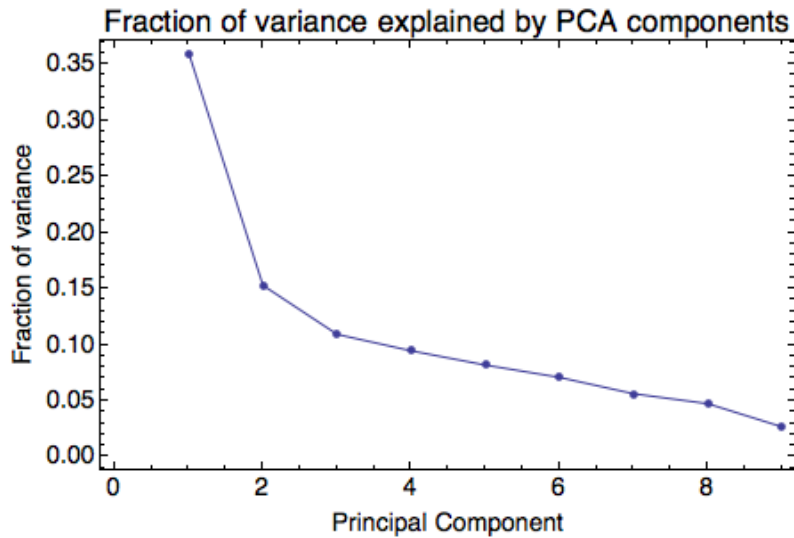
Calculate matrix of correlations of all pairs of characteristics

Two examples of correlated characteristics



We then perform *Principal Component Analysis* on the matrix of correlations.
(identifying linear combinations of characteristics that are mutually uncorrelated)

First two principal components



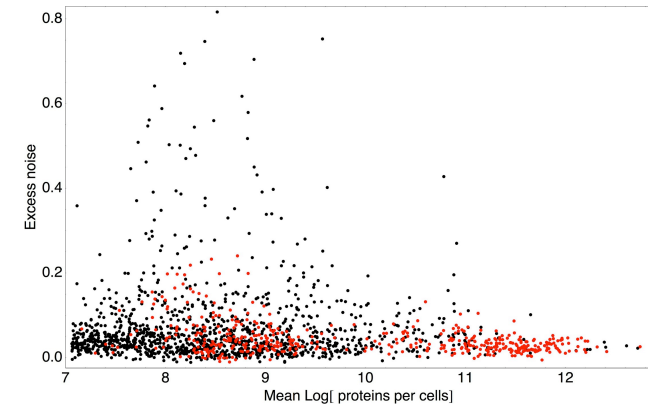
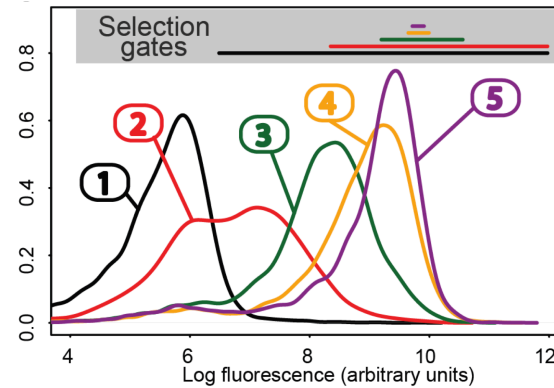
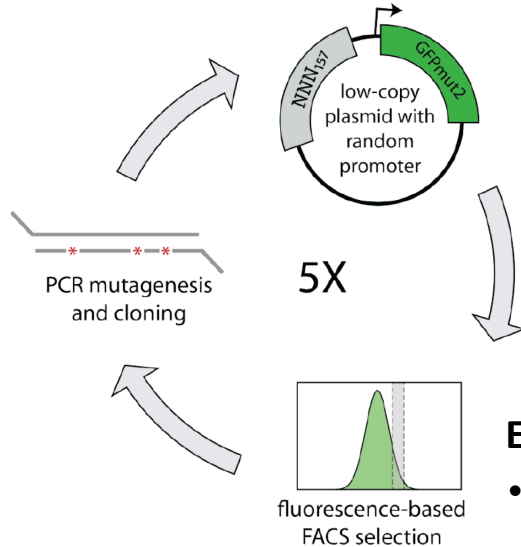
First principal Component

RNA level	dS
Protein level	dN
Expression plasticity	Promoter conservation
Promoter conservation	Expression plasticity
dN	Protein level
dS	RNA level

Second principal Component

Promoter noise	Promoter conservation
Regulatory inputs	Protein level
Expression plasticity	RNA level
RNA level	Expression plasticity
Protein level	Regulatory inputs
Promoter conservation	Promoter noise

Conclusions

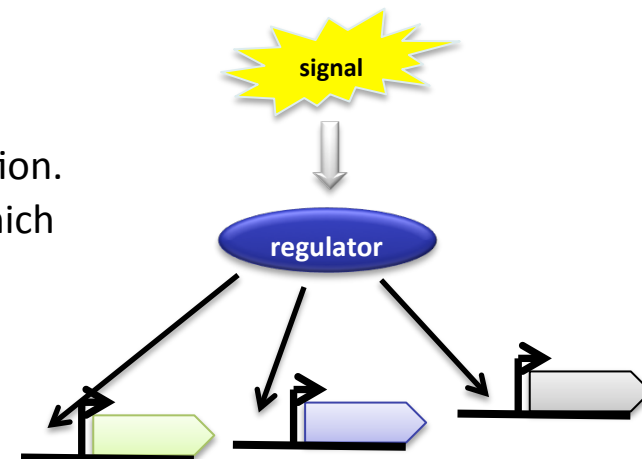


Experimental observations

- We evolved synthetic promoters *de novo* in *E. coli* under carefully-controlled selective conditions.
- No evidence *E. coli* promoters have been selected to lower noise.
- Regulated genes have been selected to increase noise.

Theory

- Coupling a regulator to a target promoter has two effects:
 1. Condition-response.
 2. Noise-propagation.
- Noise-propagation alone can act as a rudimentary form of regulation.
- Accurate regulation can evolve smoothly along a continuum in which noise-propagation and condition-response act in concert.
- Whenever the condition-response has limited accuracy, noisy regulation is preferred.
- Explains the general association between noise and regulation.



Outline of the lectures

Day 1

1. Computational methods for determining the constellations of regulatory sites.
2. From constellations of regulatory sites to genome-wide gene expression patterns.

Day 2

3. Large-scale patterns in genomes and gene regulatory networks.
4. Gene expression noise and its role in the evolution of *de novo* gene regulation.

Day 3

5. *How do bacterial genomes evolve in the wild?*

Why is there still almost no predictive quantitative evolutionary theory?

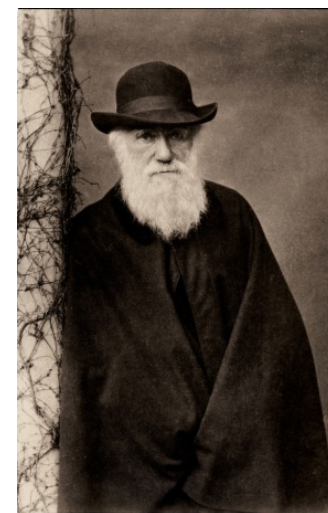
Feynman on how physics works (my paraphrase):

1. We observe phenomena that interests us.
2. We formalize the phenomena by rigorously defining measurable quantities.
3. We search for specific quantitative relationships, i.e. 'laws' that connect the measurable quantities.
4. The glory of physics is when we find a theory that, in one full sweep, explains many of these relationships and makes them self-evident.



The original sins of mathematical population genetics

- No collection of laws about observable quantities in evolving populations in nature.
- Not even clear what measurable quantities should be analyzed to look for such 'laws'.
- Took a qualitative framework and just started making up toy models.
- The key quantities in mathematical population genetics models are impossible to measure:
 - Fitness
 - Effective population size

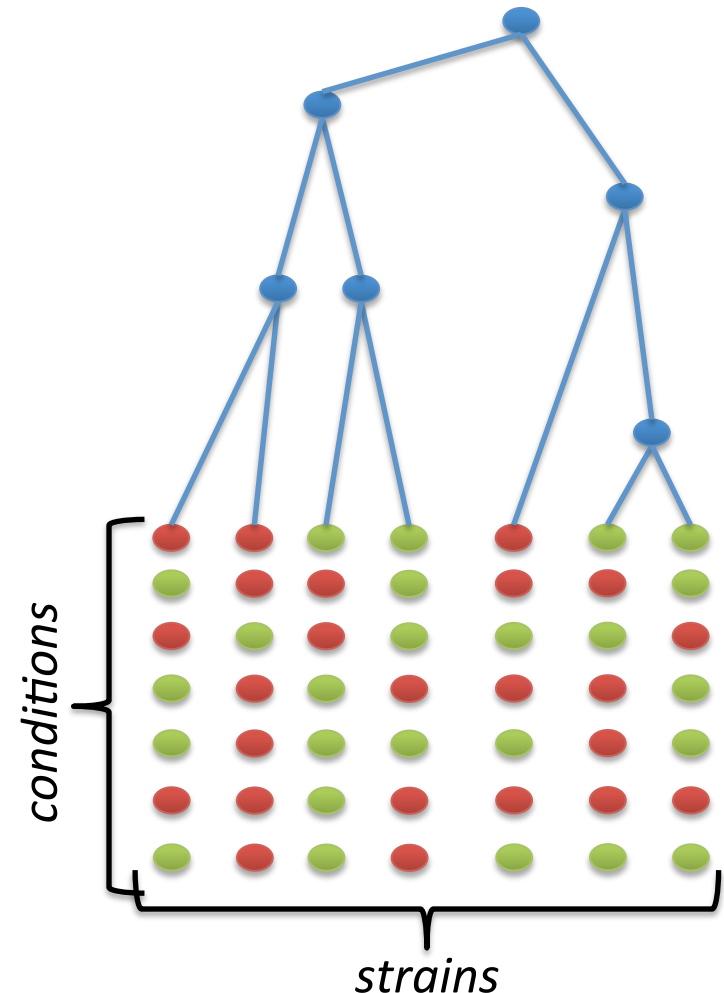


Original research plan:

- Isolate wild strains of closely-related bacteria from a single habitat.
- Sequence their genomes.
- Reconstruct their phylogeny.
- Phenotype them.
(growth/non-growth across many conditions)



- Analyze statistics of how phenotypes evolve along the phylogeny.
- Do we see any recurring quantitative patterns?



Evolution of *E. Coli* in the wild



46.701111,....

Report a problem

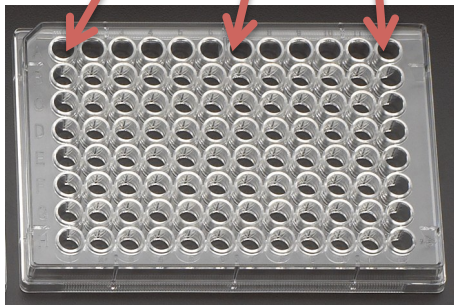
The SC Strain Collection



Water

Watershed Soil

Inland Soil



Filtered by phenotype and MLST for *E. coli*.

≈500 Strains, focused on one plate

→ Phenotype and **Genome** data

Ishii et al., Appl. Environ. Microb., 2006

Core Phylogeny

Core Genome:

Sequence that exists
unambiguously in all strains

E. coli

SC1 Collection

K12 Strains

- *S. dysenteriae*

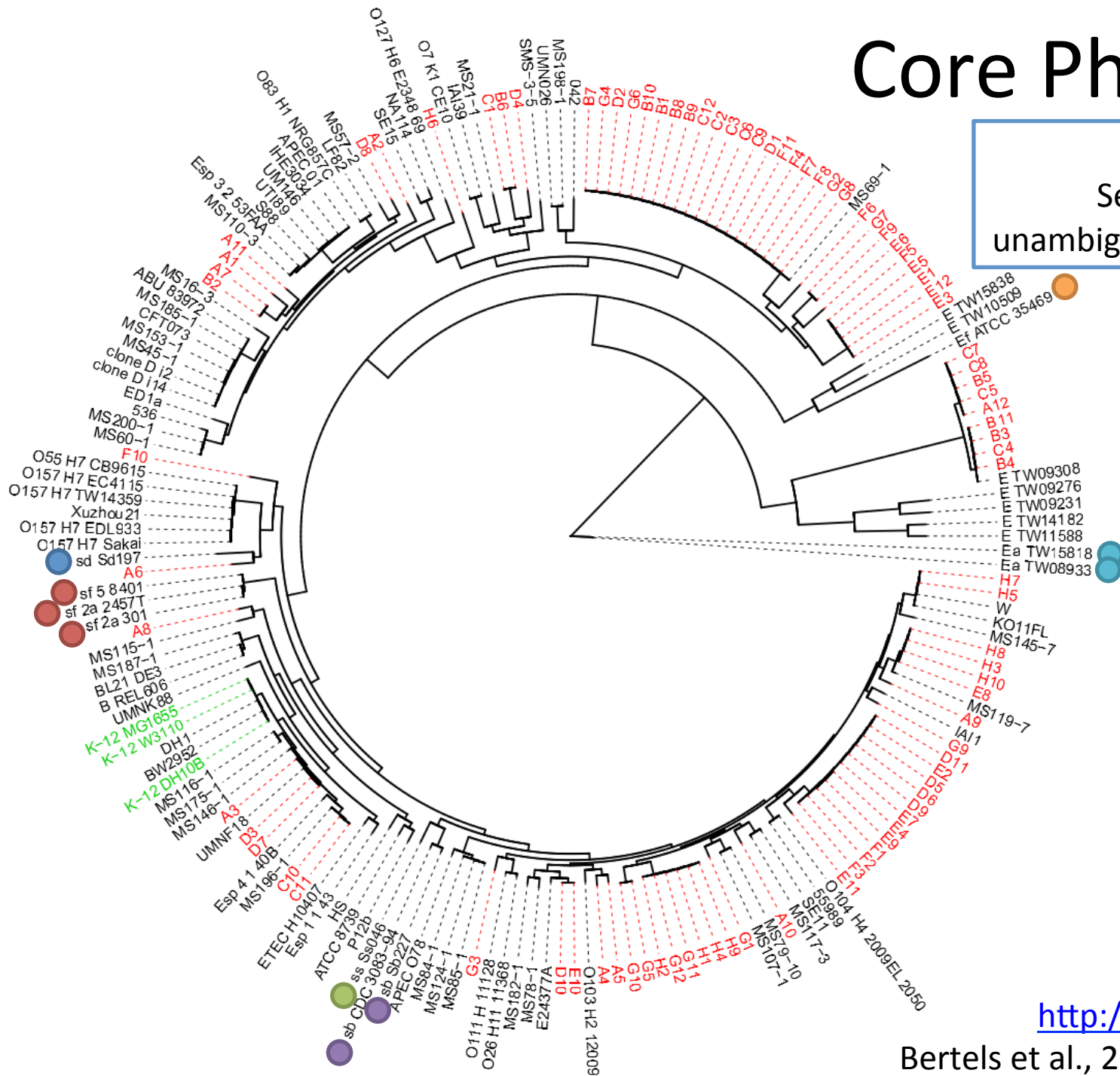
- *S. flexneri*

● S. sonnei

● S. Boydii

- *E. albertii*

● *E. fergusonii*

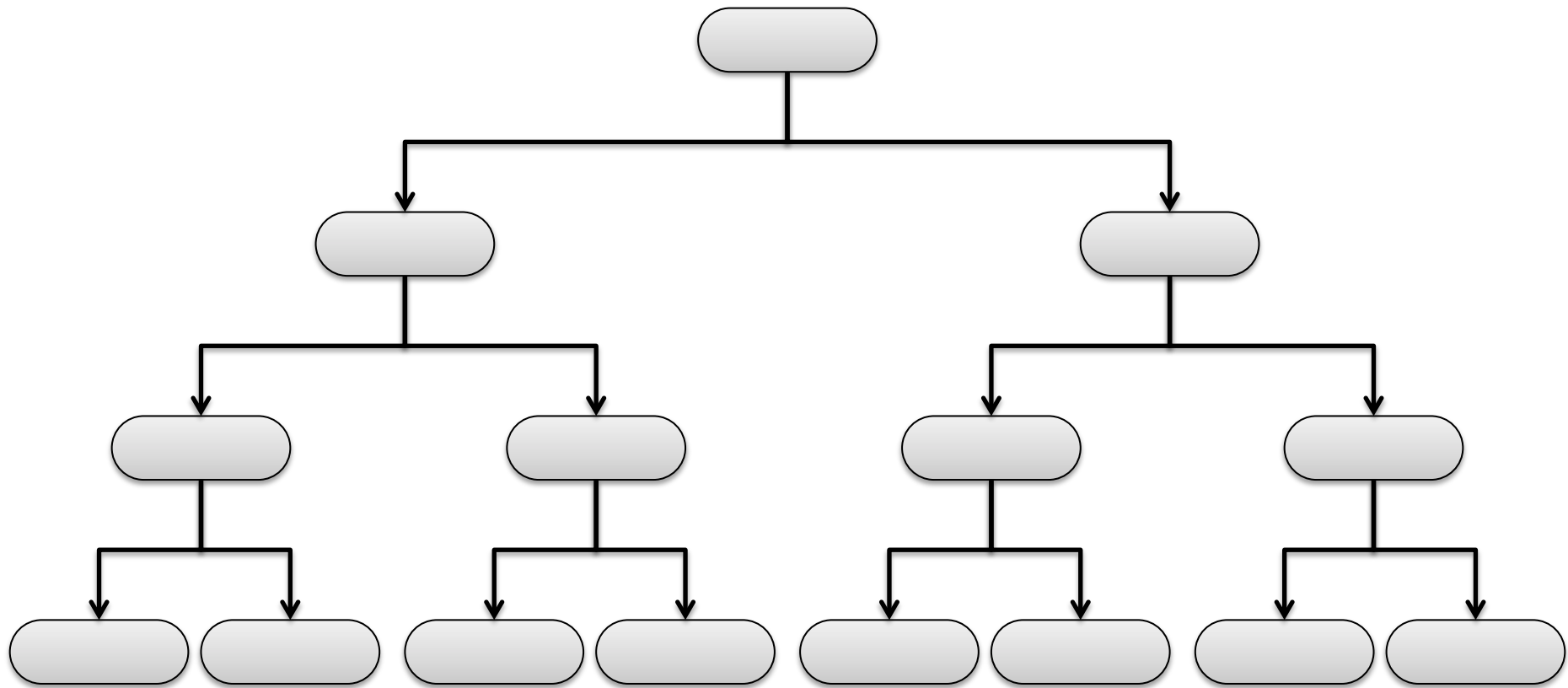


<http://realphy.unibas.ch/>

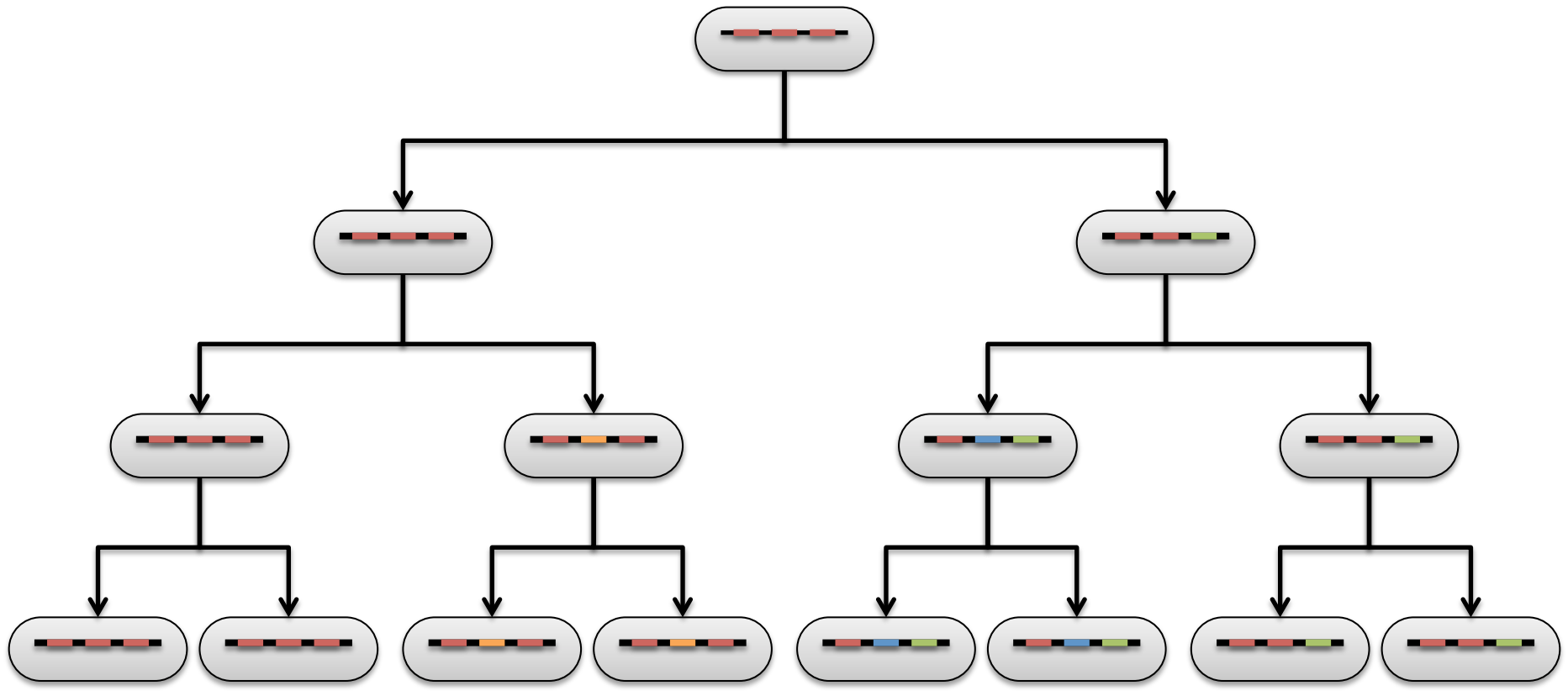
Bertels et al., 2014, Mol. Biol. Evol.

Cell Division

The clonal phylogeny

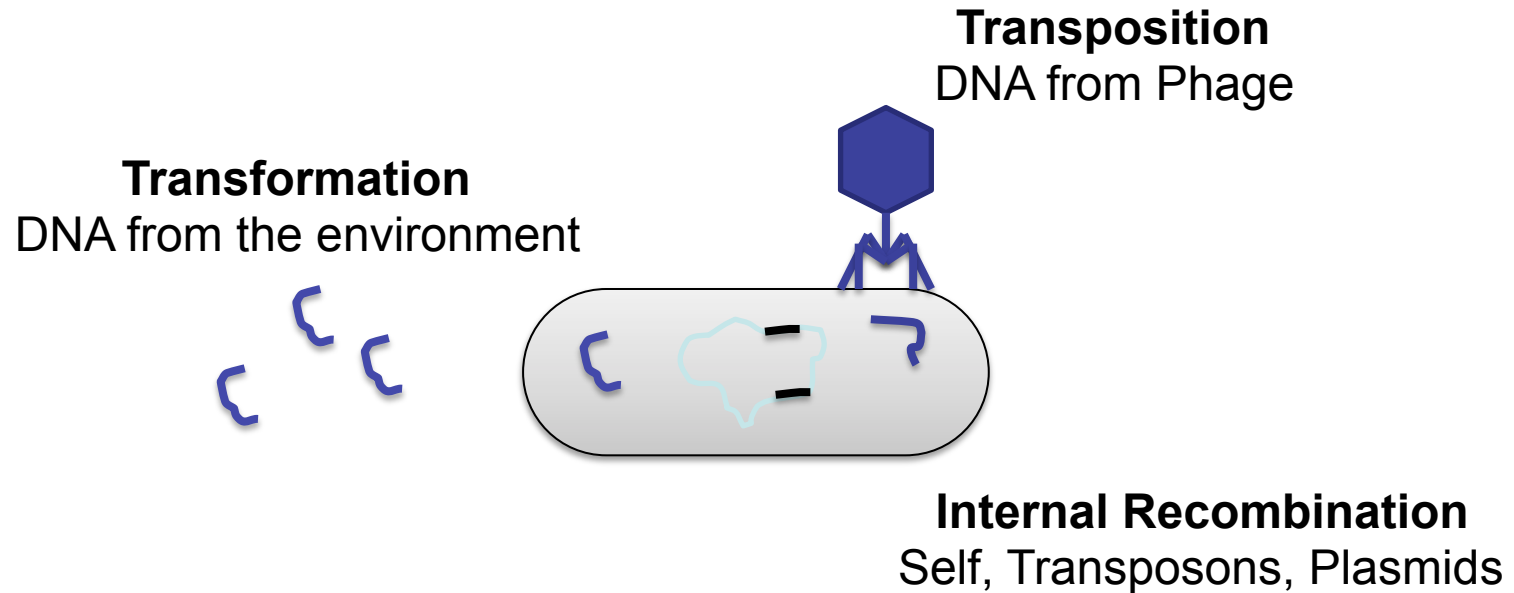


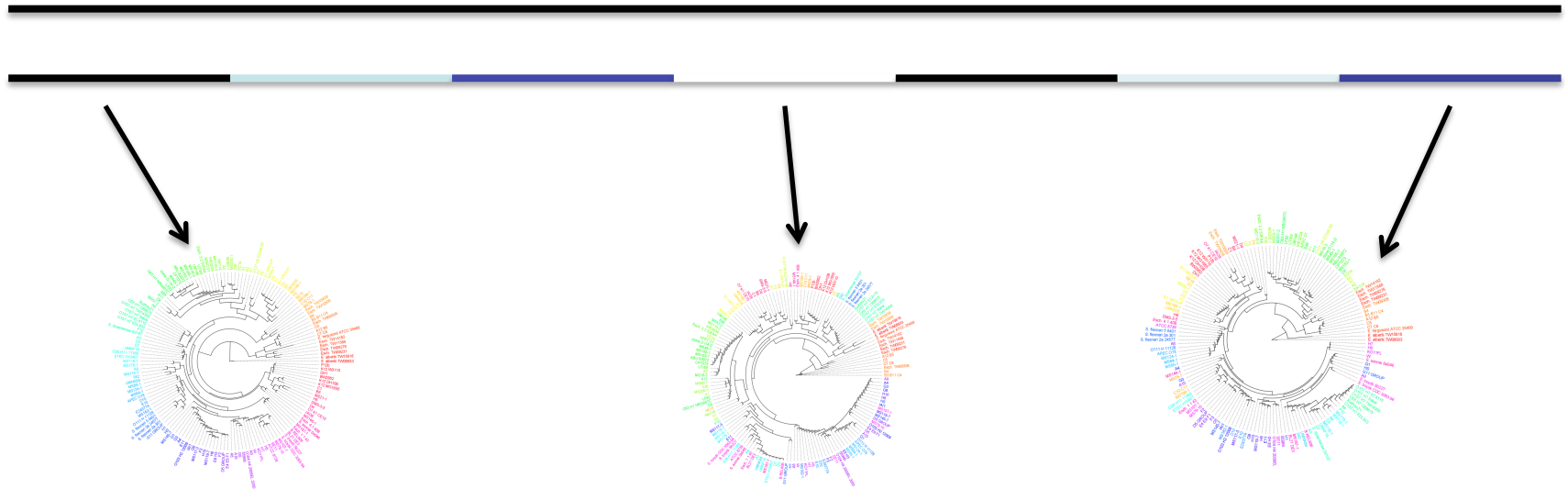
Clonal inheritance: sequence relations reflect phylogeny



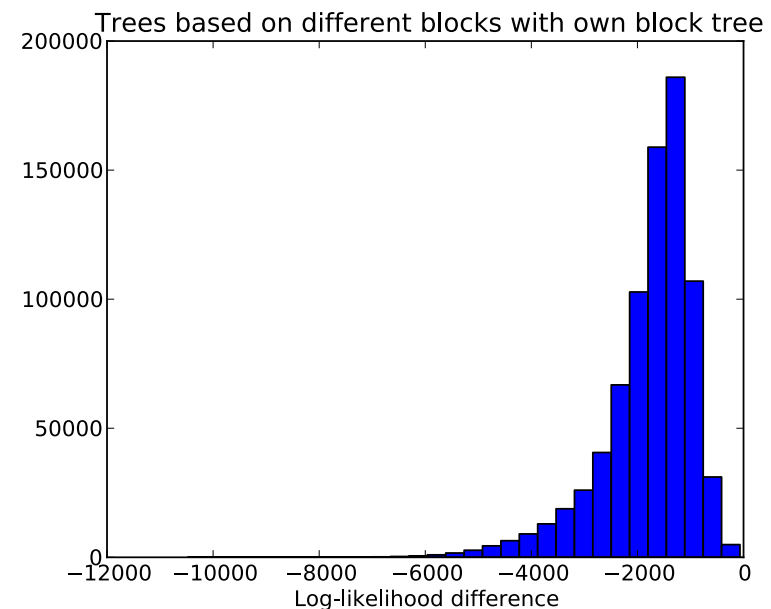
Not all DNA is vertically inherited

Homologous Recombination





- Constructed a phylogeny from each 2Kb block in the alignment.
- **Each block significantly rejects the phylogenies from all other blocks.**
- **But:** Phylogeny created from any large set of blocks is essentially *the same*.
- Is that the clonal phylogeny?



Can the clonal phylogeny be recovered from DNA?

Typical reasoning in the community:

- Horizontal gene transfer is common..... **But:**
 - ‘Core genes’ that all strains share may be relatively unaffected by HGT.
 - Genome-wide, effects HGT will ‘average out’ and the clonal phylogeny will emerge.
 - It is often assumed that HGT can be treated as a perturbation on top of the ‘clonal frame’ reconstructed from the core:

ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes.
X Didelot, DJ Wilson *PLoS computational biology* **11** (2), e1004041-e1004041 (2015)

Or is recombination so common that the clonal phylogeny cannot be recovered?

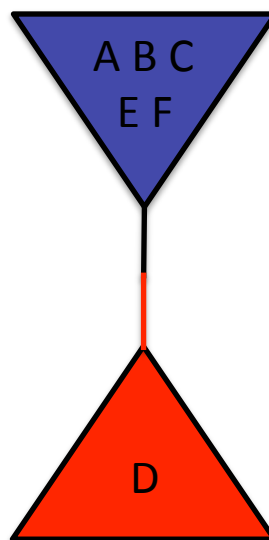
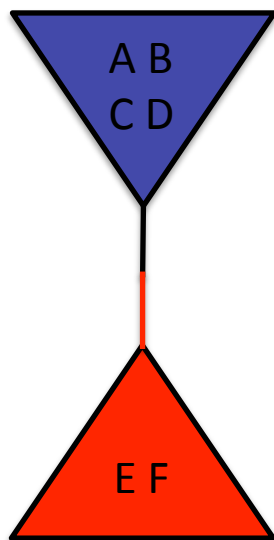
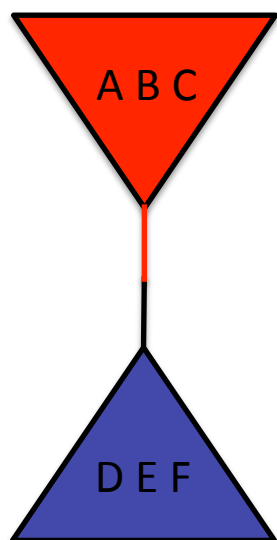
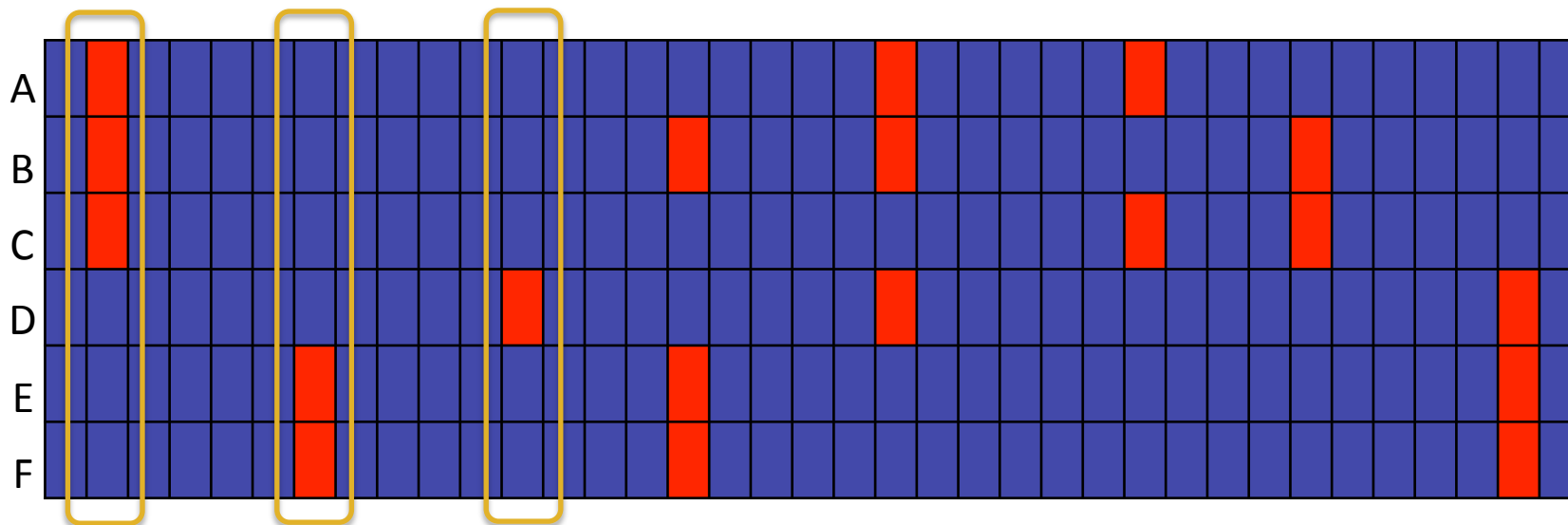
Recombinant transfer in the basic genome of *Escherichia coli*
PD Dixit, TY Pang, FW Studier, S Maslov *PNAS* **112** (29), 9070-9075 (2015)

Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche.

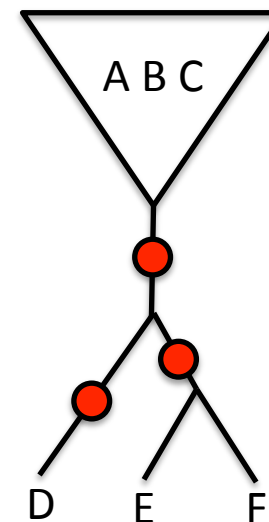
MJ Rosen, M Davison, D Bhaya, DS Fisher *Science* **348** (6238), 1019-1023 (2015)

Which is it?

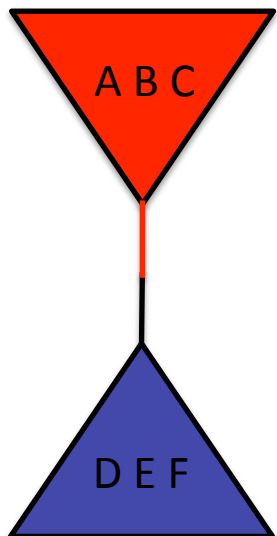
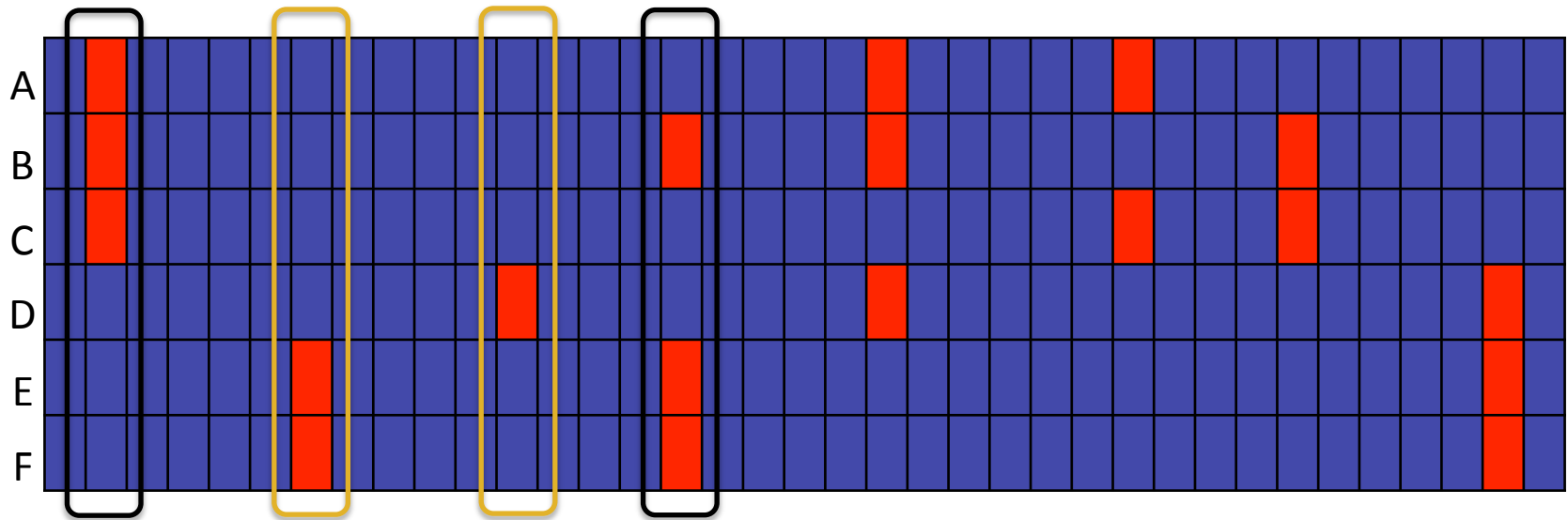
SNPs bi-partitions the strain set



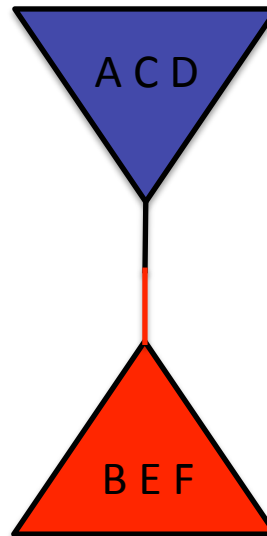
Phylogeny consistent with the 3 SNPs



SNPs can be incompatible with a common phylogeny

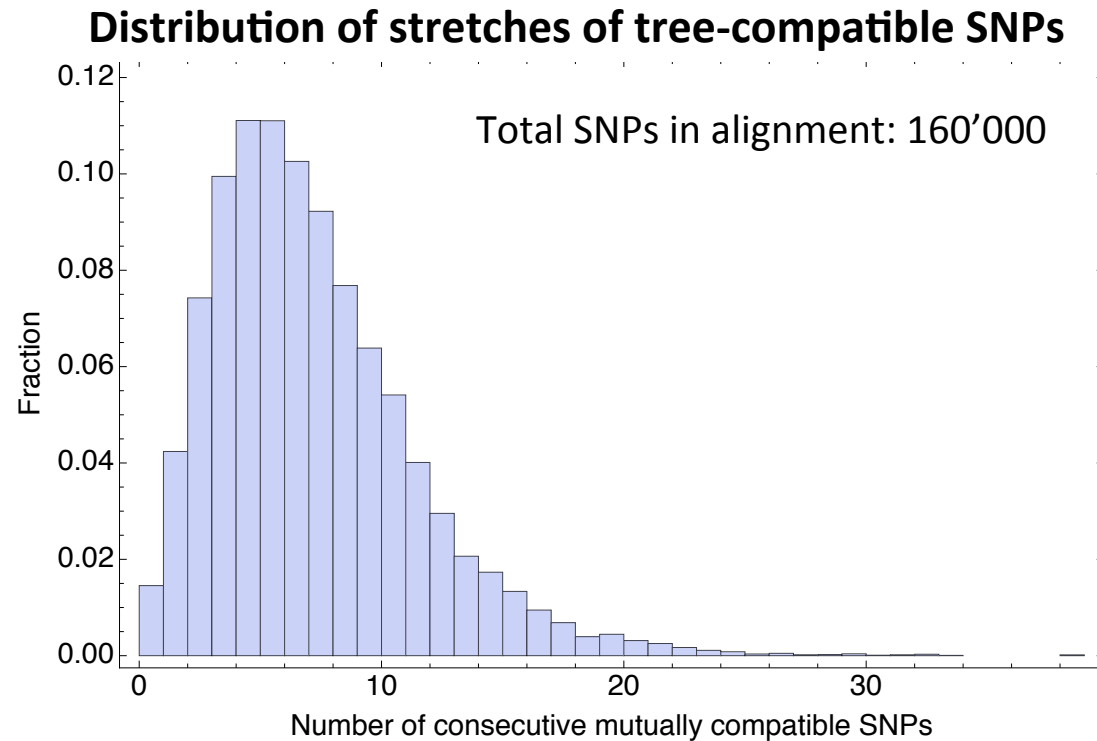


Incompatible!



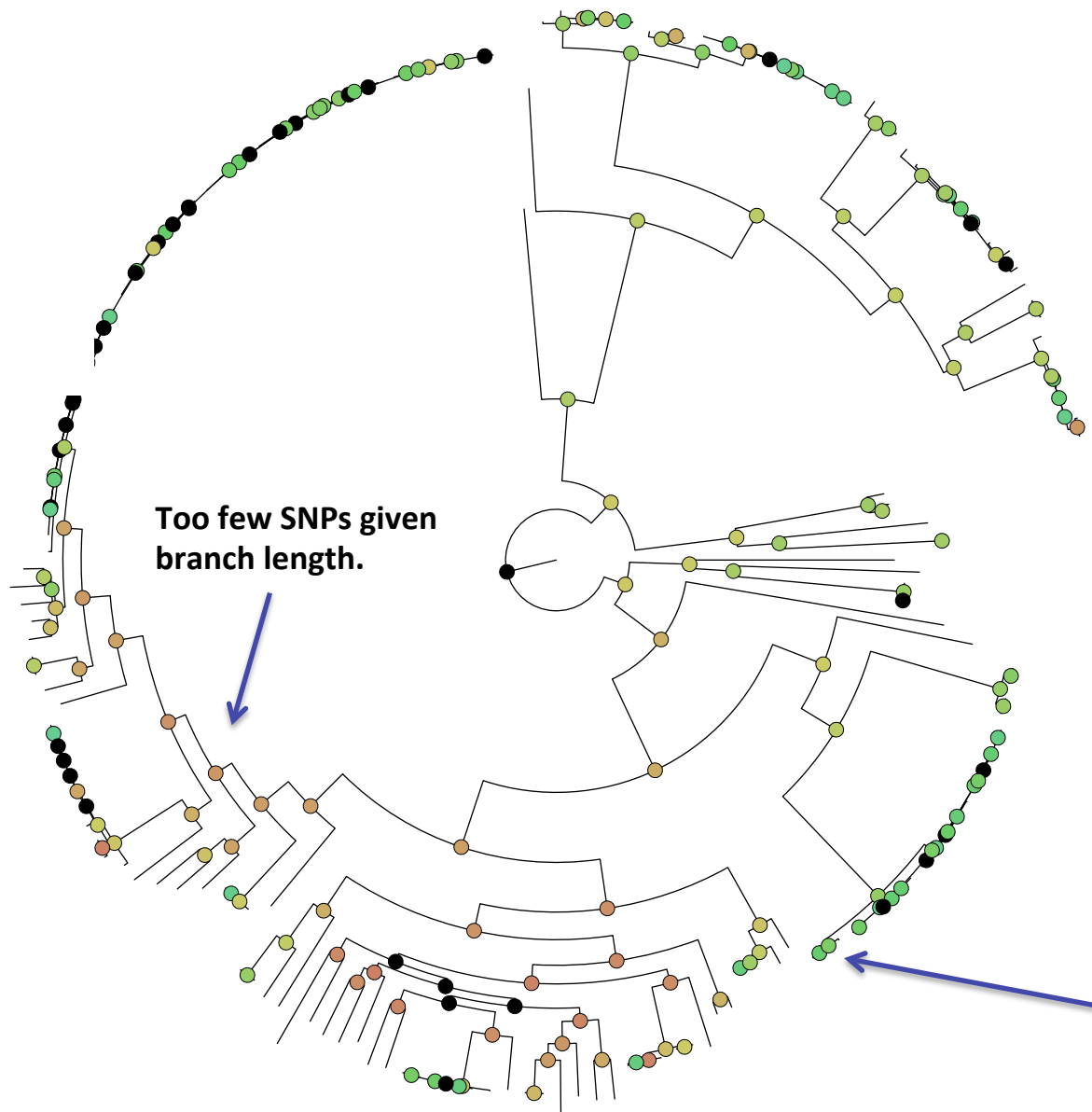
Three questions:

- How long are the stretches of mutually compatible SNPs along the alignment?
- How many of the SNPs are compatible with the core phylogeny?
- How many SNPs are compatible with *any* single phylogeny?



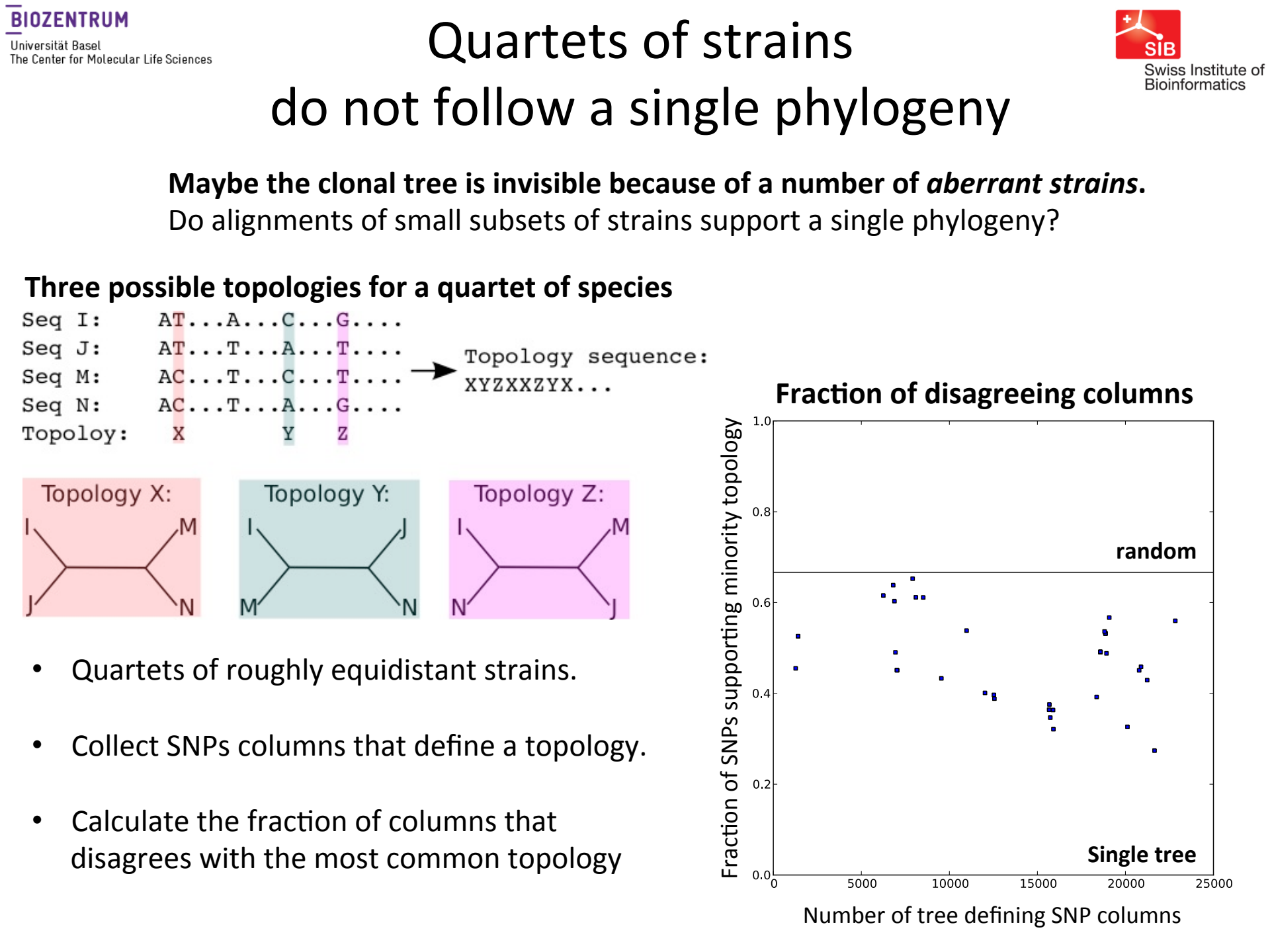
- Very short length-scale of SNP compatibility suggests lots of recombination.
- Phylogeny must change more than 15'000 times in the core alignment.

- But maybe one common `background' phylogeny is frequently interspersed with short recombined stretches?
- What fraction of all SNPs are consistent with the core phylogeny?



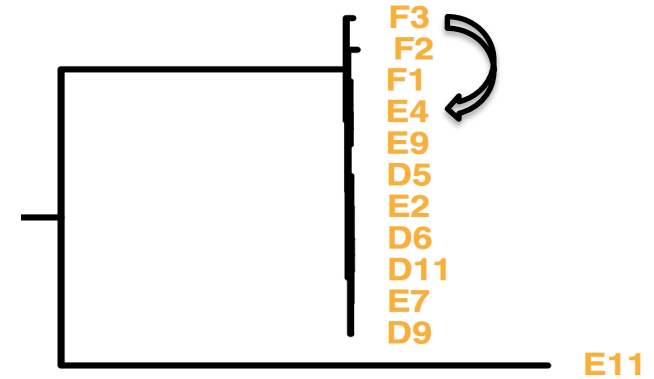
- 71.7% of SNPs are *inconsistent* with the core phylogeny.
- SNP frequency often orders of magnitude different from inferred branch length.
- **Core tree does not capture SNP stats at all.**
- Constructing a tree to maximize consistent SNPs does not help: 71.4% inconsistent with tree.

Too many SNPs given
branch length.



Pairwise alignment of very close pairs shows evidence of recombination

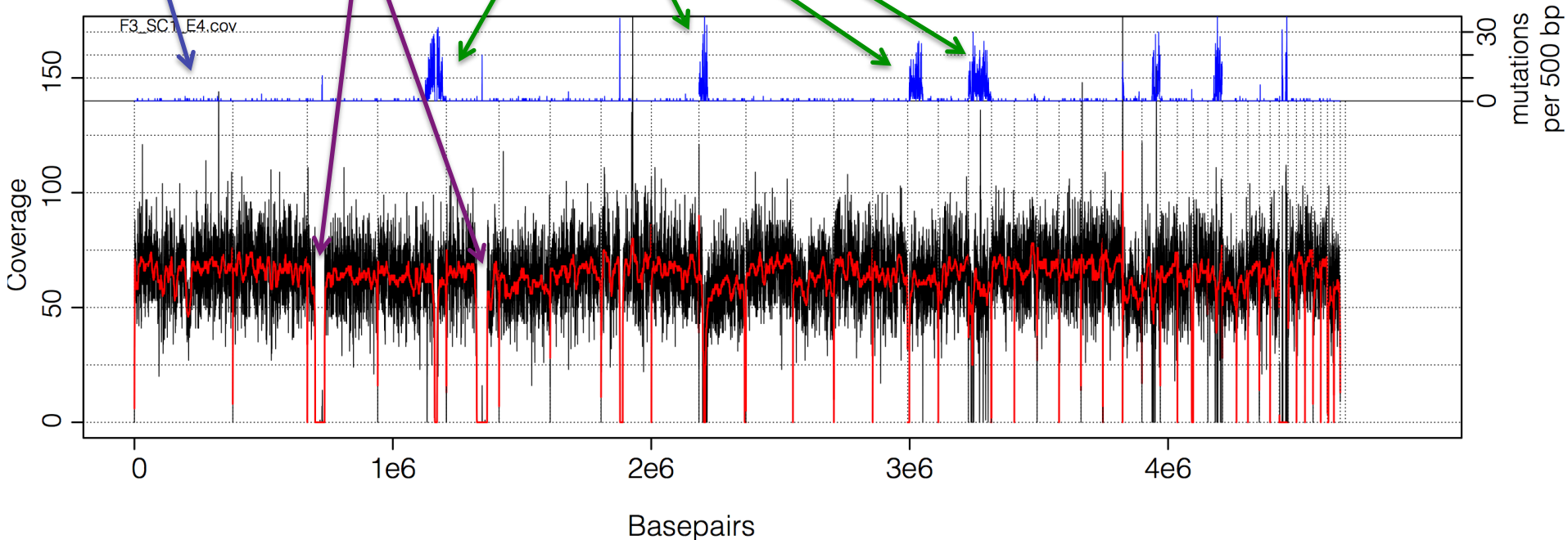
- Map F3 sequences to E4 genome.
- Calculate in sliding windows:
 - SNP density.
 - Coverage.



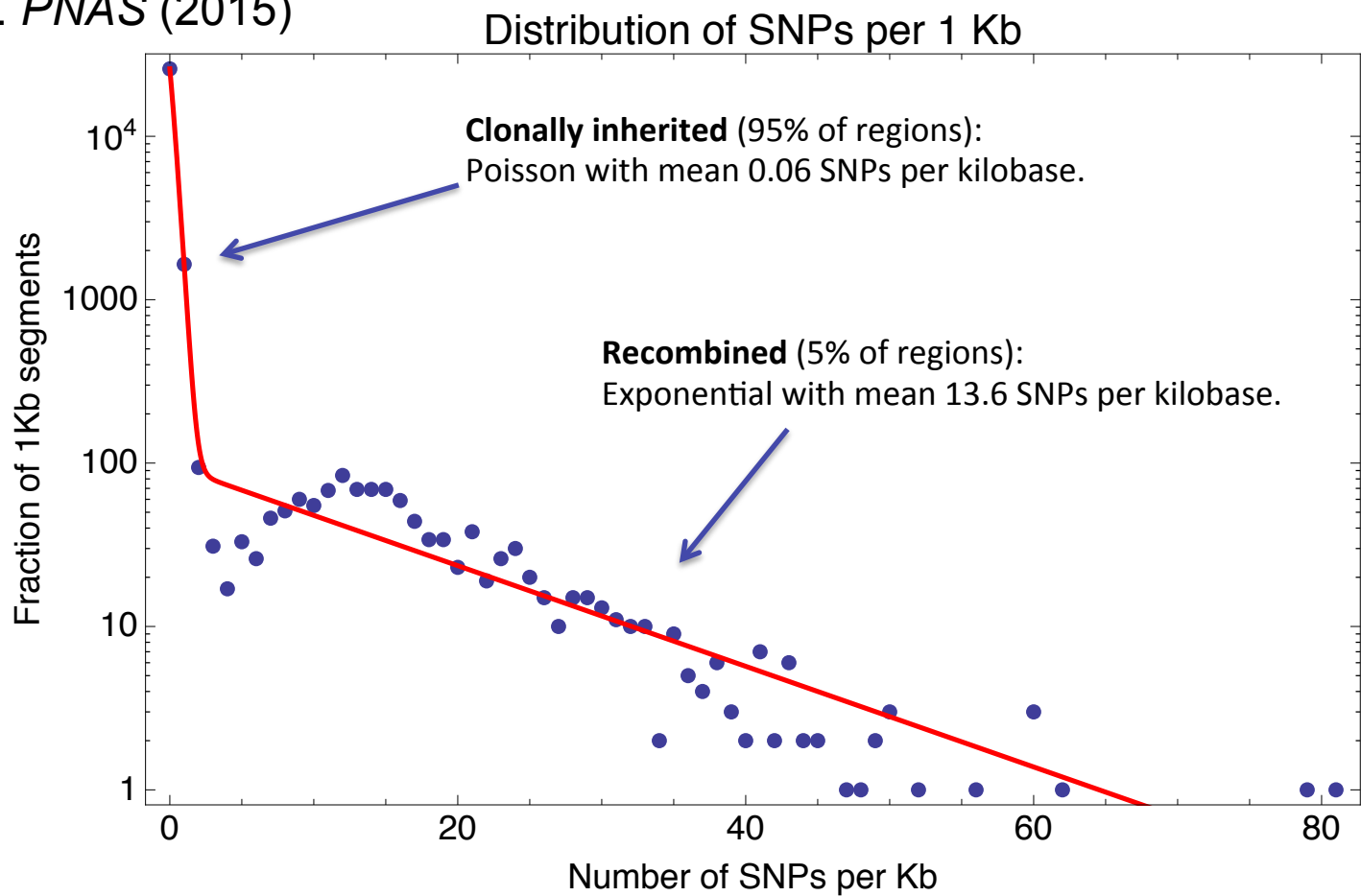
Low SNP density regions.

High SNP density regions

Missing in F3



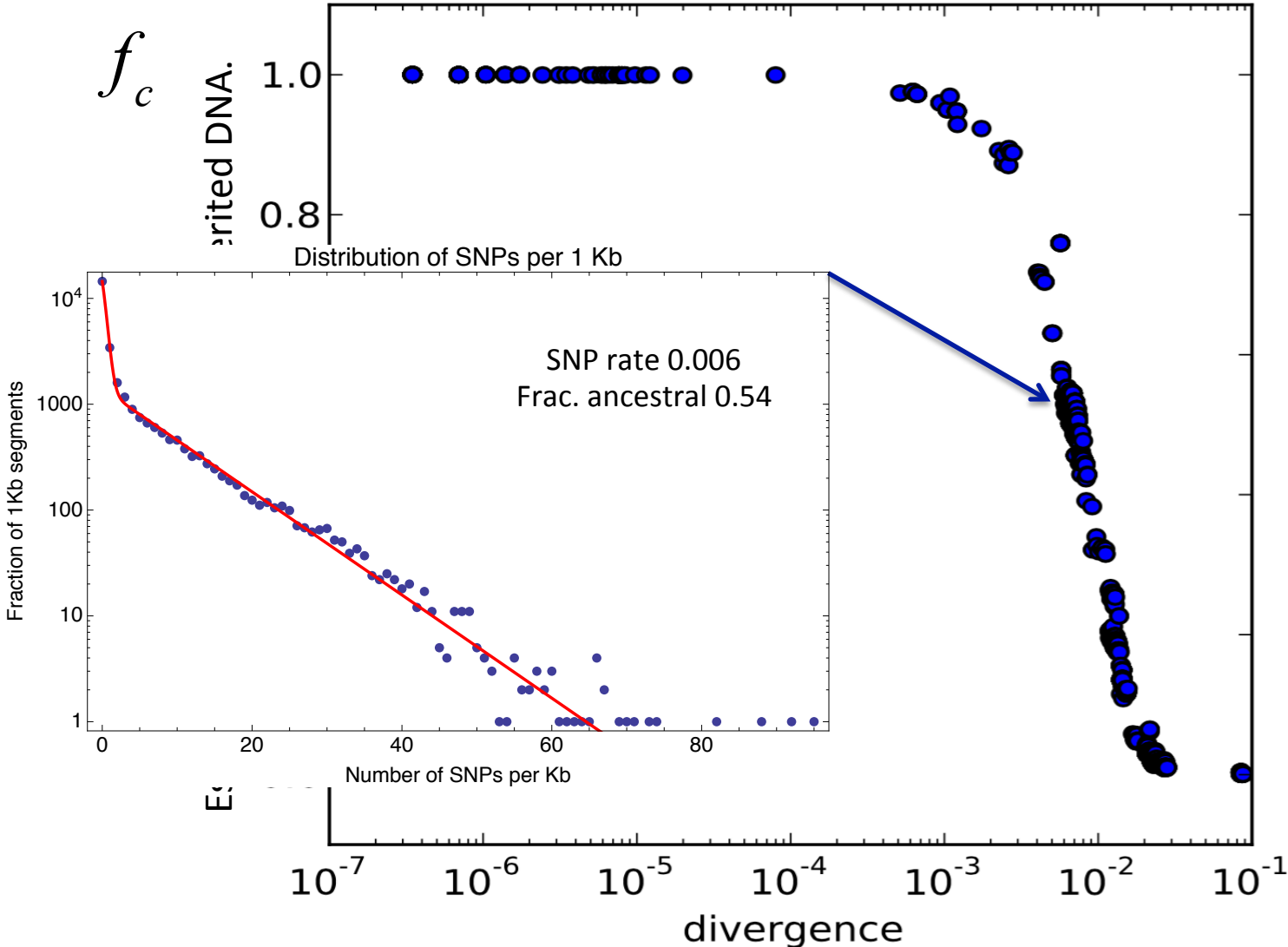
Dixit et al. *PNAS* (2015)

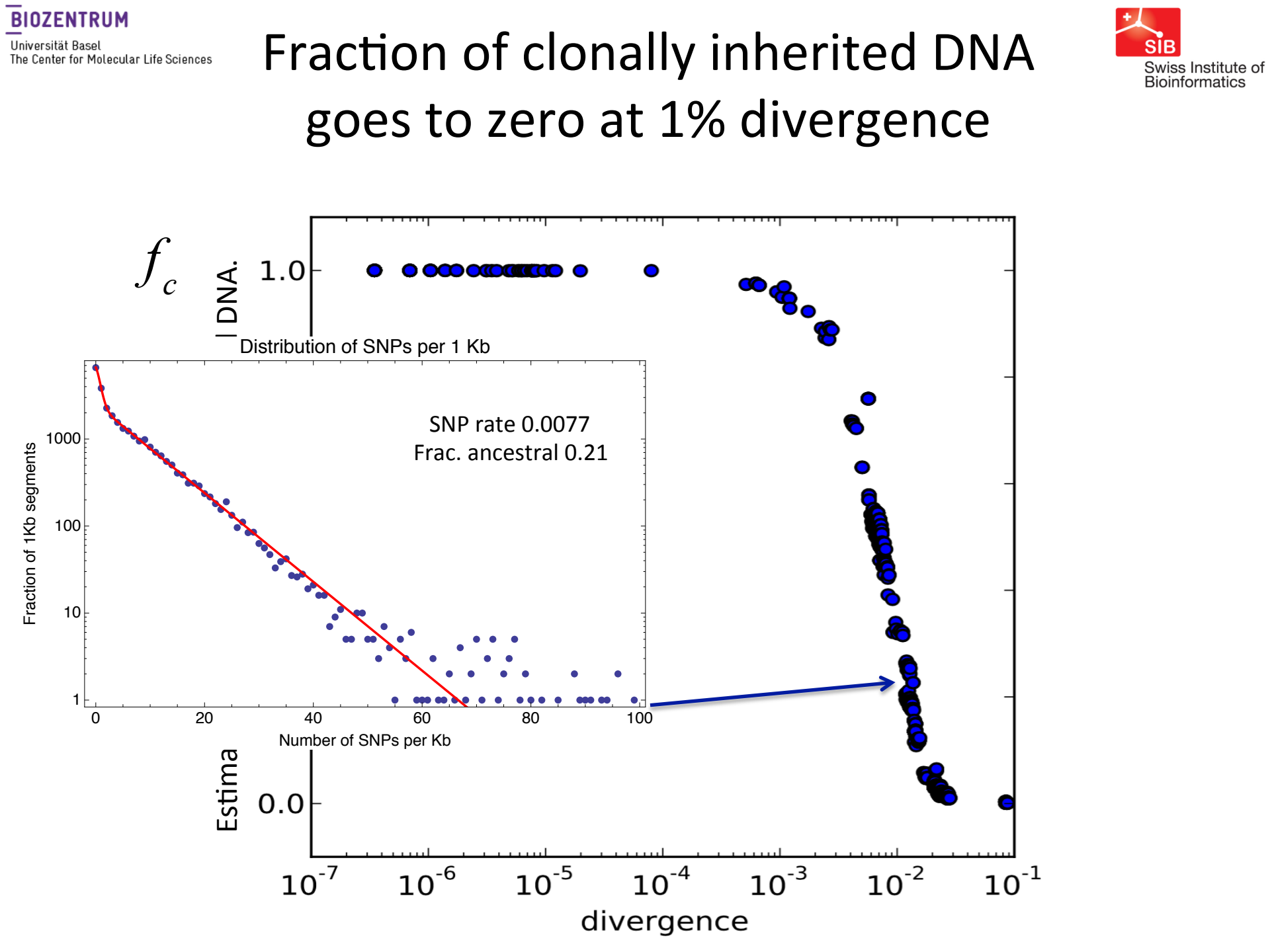


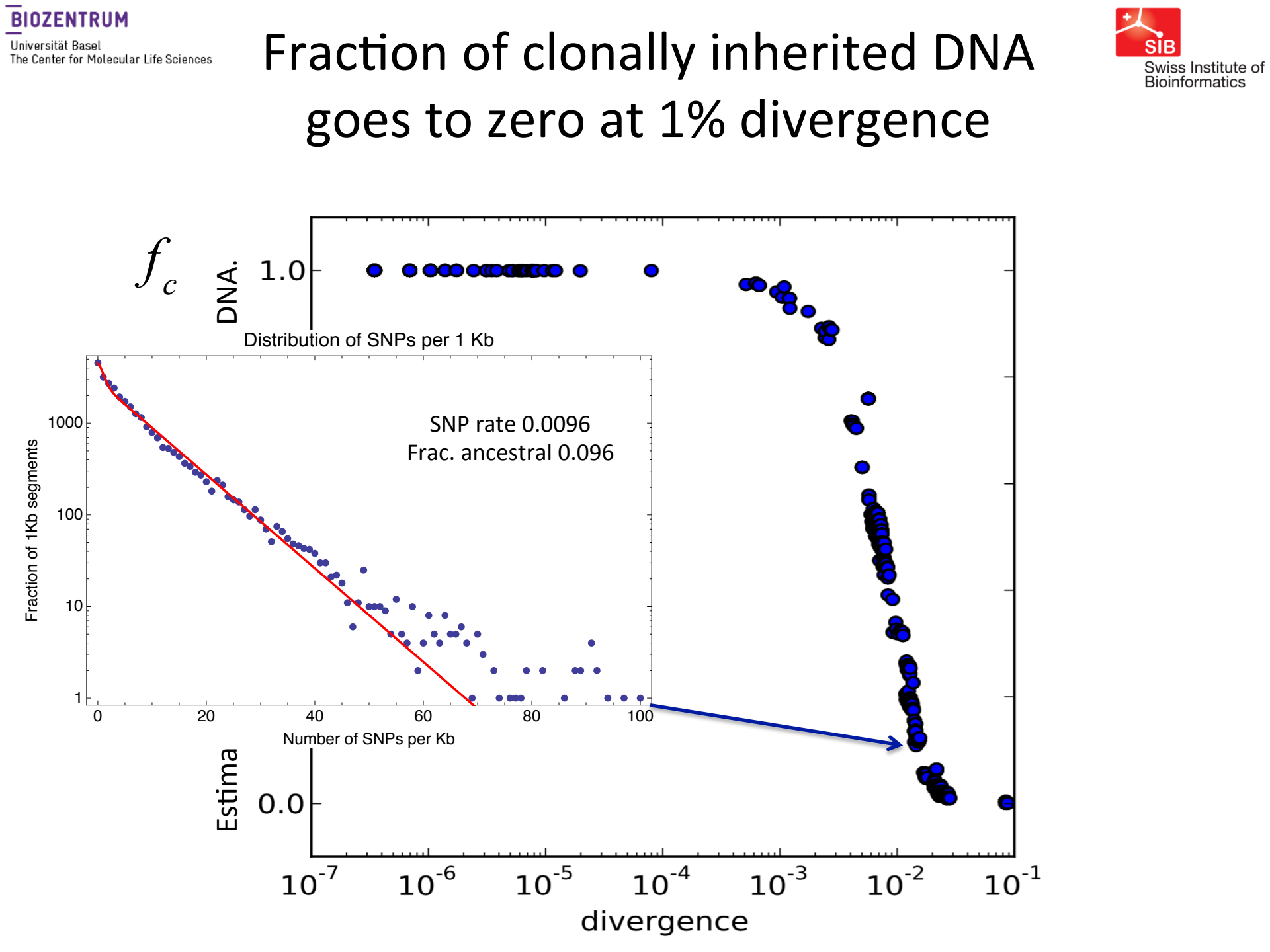
For each pair, fit a mixture of a Poisson (clonal) and a negative binomial (recombined):

$$P(n) = f_c \frac{r^n}{n!} e^{-r} + (1 - f_c)(1 - \lambda)^a \lambda^n \frac{\Gamma(n + a)}{n! \Gamma(a)} \quad a \geq 1$$

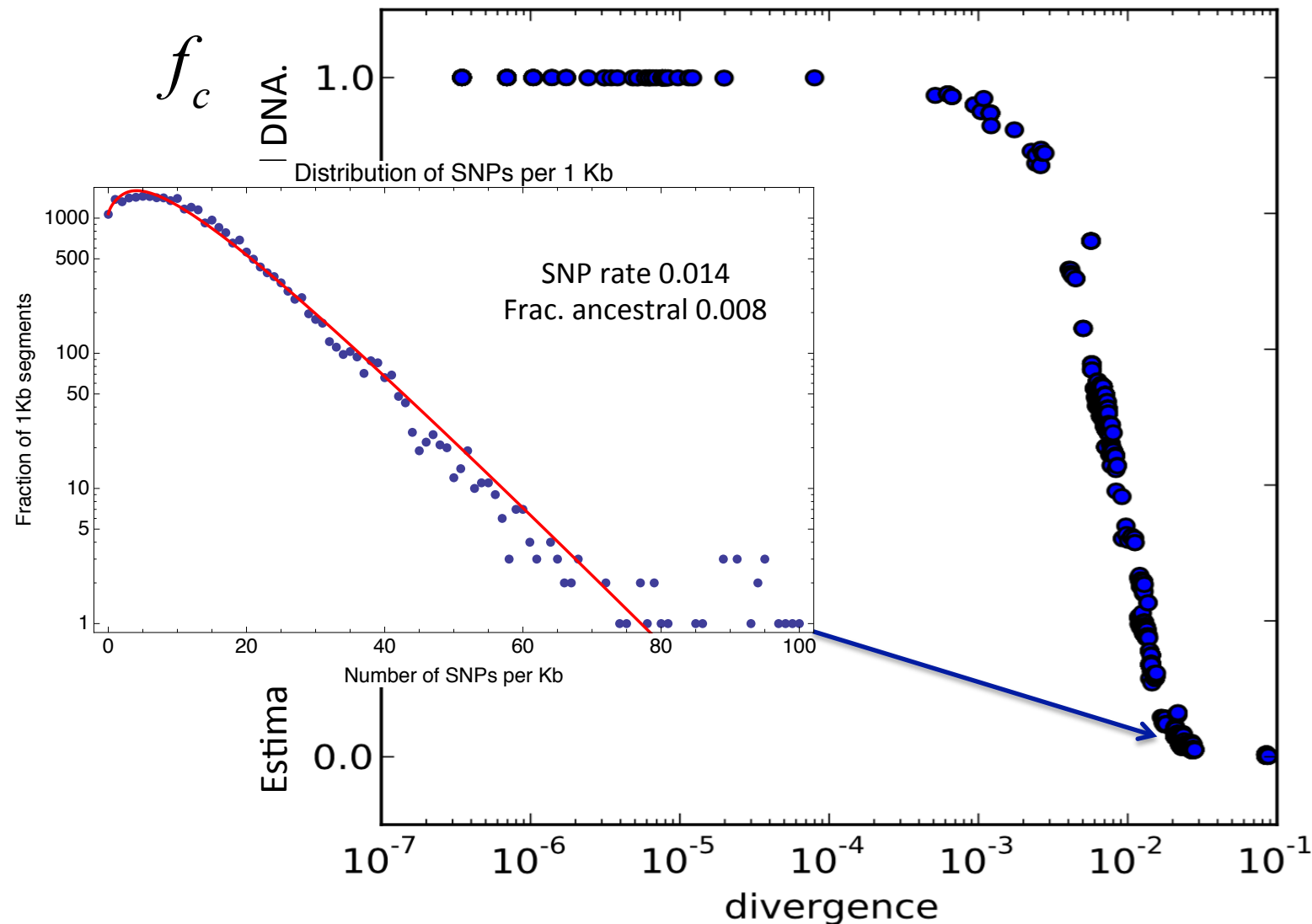
Fraction of clonally inherited DNA goes to zero at 1% divergence

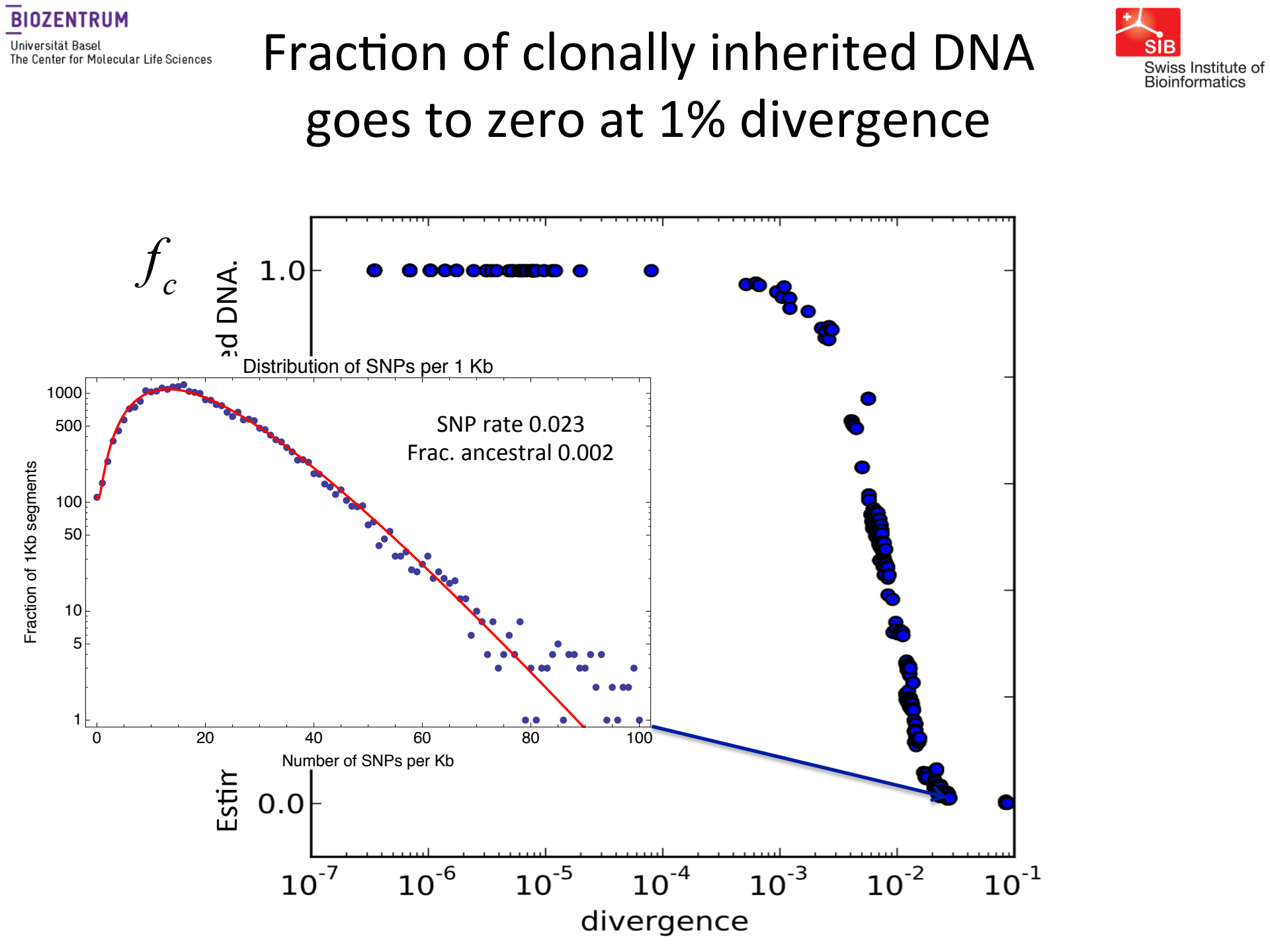






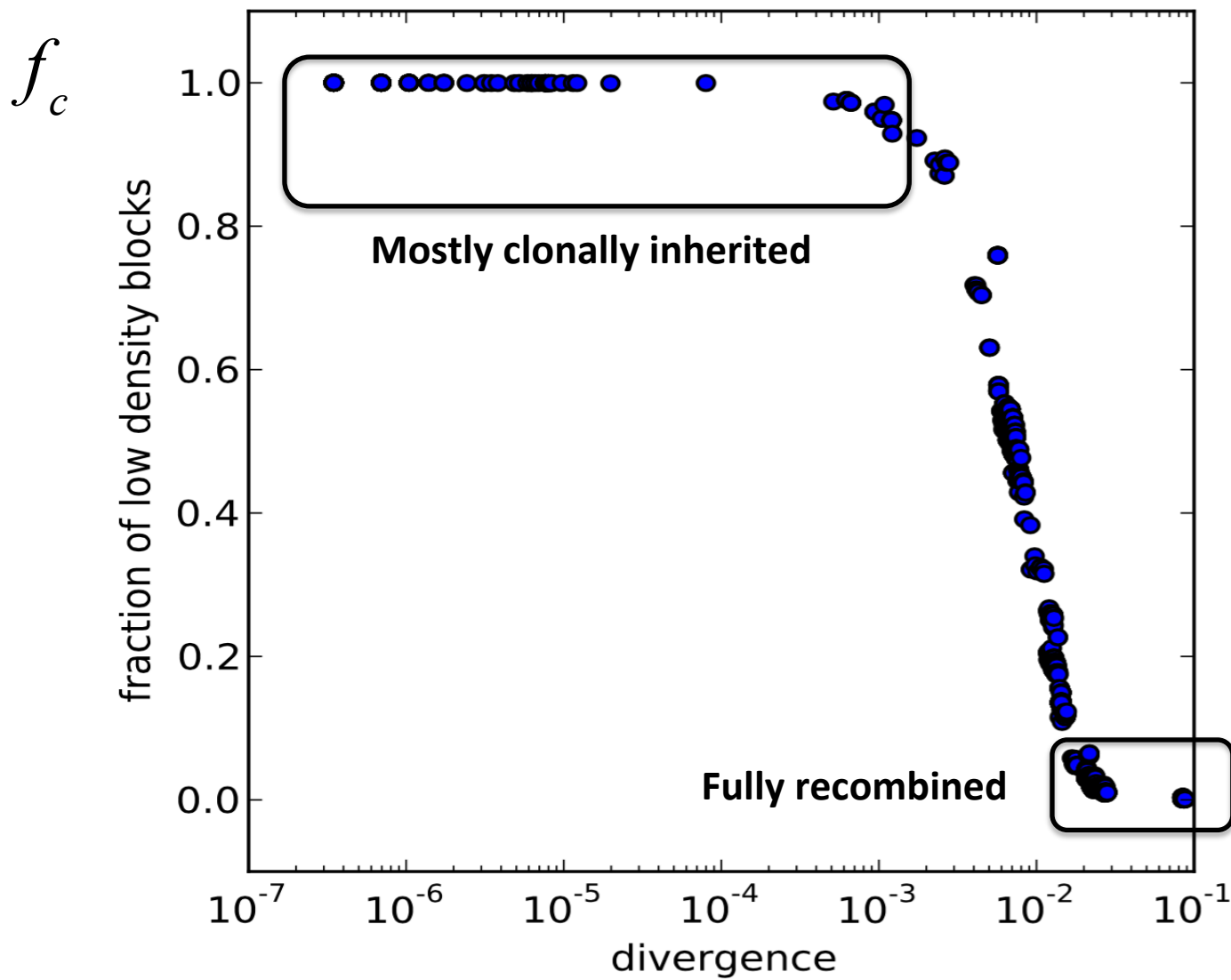
Fraction of clonally inherited DNA goes to zero at 1% divergence





Fraction of clonally inherited DNA goes to zero at 1% divergence

Similar to observations in: Dixit et al. *PNAS* (2015)



Simple model

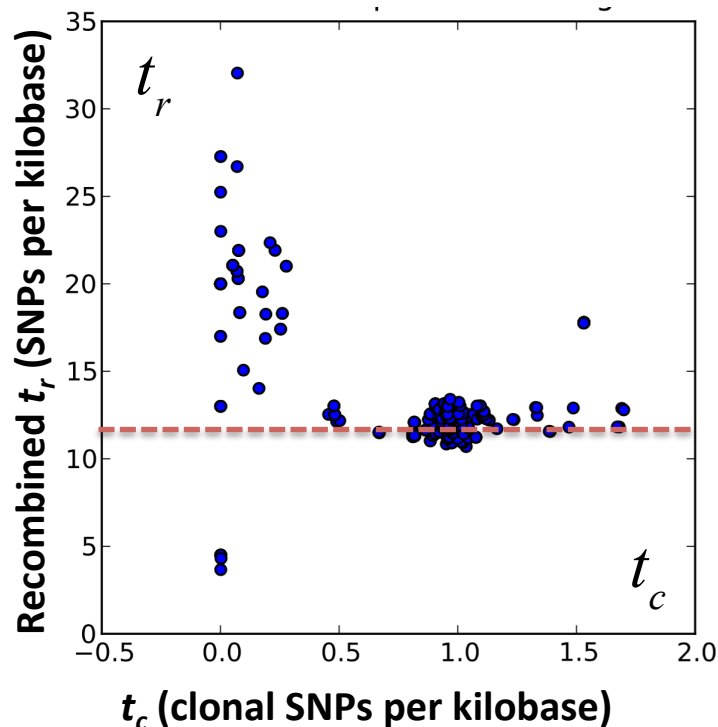
Three simple stats for each pair of strains:

f_c = Fraction of genome that is clonally inherited.

t_c = SNP rate in clonally inherited regions. Measure of time since clonal ancestor.

t_r = SNP rate in recombined regions. Average coalescence time to recombined ancestor.

$$t = f_c t_c + (1 - f_c) t_r \quad \text{Total divergence of the pair.}$$



Constant average age of recombined blocks.

- t_r is independent of t_c .
- $t_r = 0.013$ (13 SNPs per Kb)

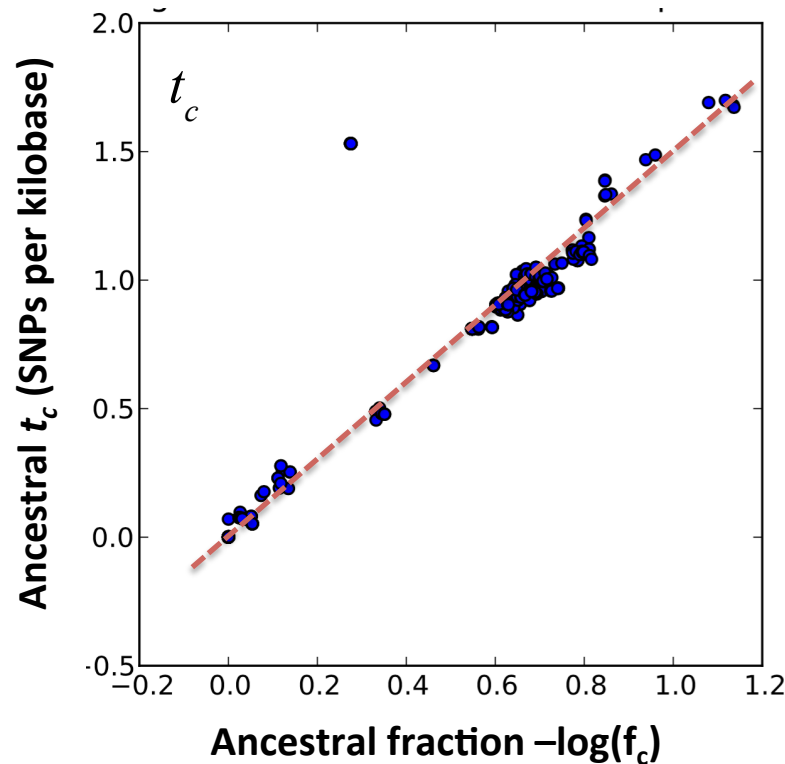
Three simple stats for each pair of strains:

f_c = Fraction of genome that is clonally inherited.

t_c = SNP rate in clonally inherited regions. Measure of time since clonal ancestor.

t_r = SNP rate in recombined regions. Average coalescence time to recombined ancestor.

$t = f_c t_c + (1 - f_c) t_r$ Total divergence of the pair.



f_c decreases exponentially with t_c .

$f_c = e^{-\rho t_c} \Rightarrow t_c = -\frac{\log(f_c)}{\rho}$

- effective recombination rate: $\rho \approx 670$

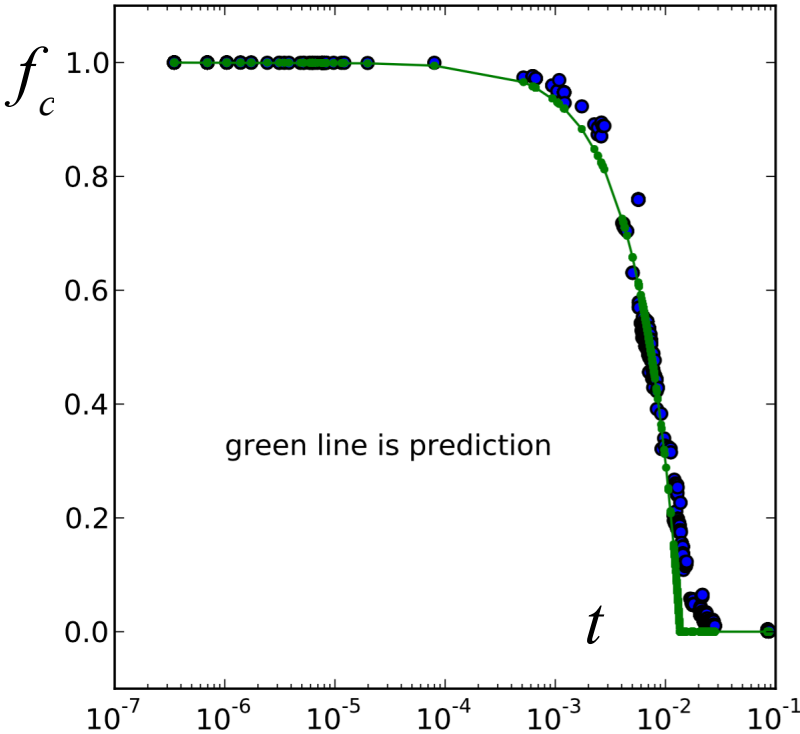
Three simple stats for each pair of strains:

f_c = Fraction of genome that is clonally inherited.

t_c = SNP rate in clonally inherited regions. Measure of time since clonal ancestor.

t_r = SNP rate in recombined regions. Average coalescence time to recombined ancestor.

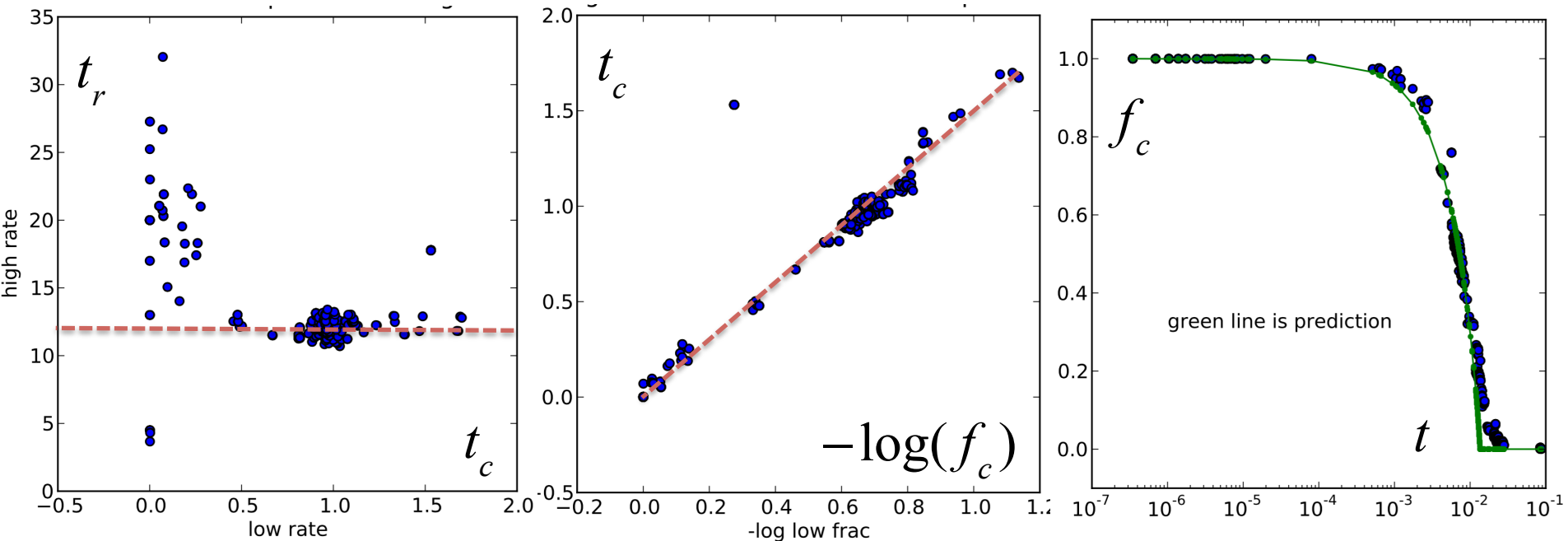
$t = f_c t_c + (1 - f_c) t_r$ Total divergence of the pair.



Solve f_c as function of total divergence t .

$$t = -f_c \frac{\log(f_c)}{\rho} + (1 - f_c) t_r$$

$$f_c = \rho(t_r - t) W \left[\rho(t_r - t) e^{\rho(t_r - t)} \right]$$



Ancestral divergence when half of the genome is recombined:

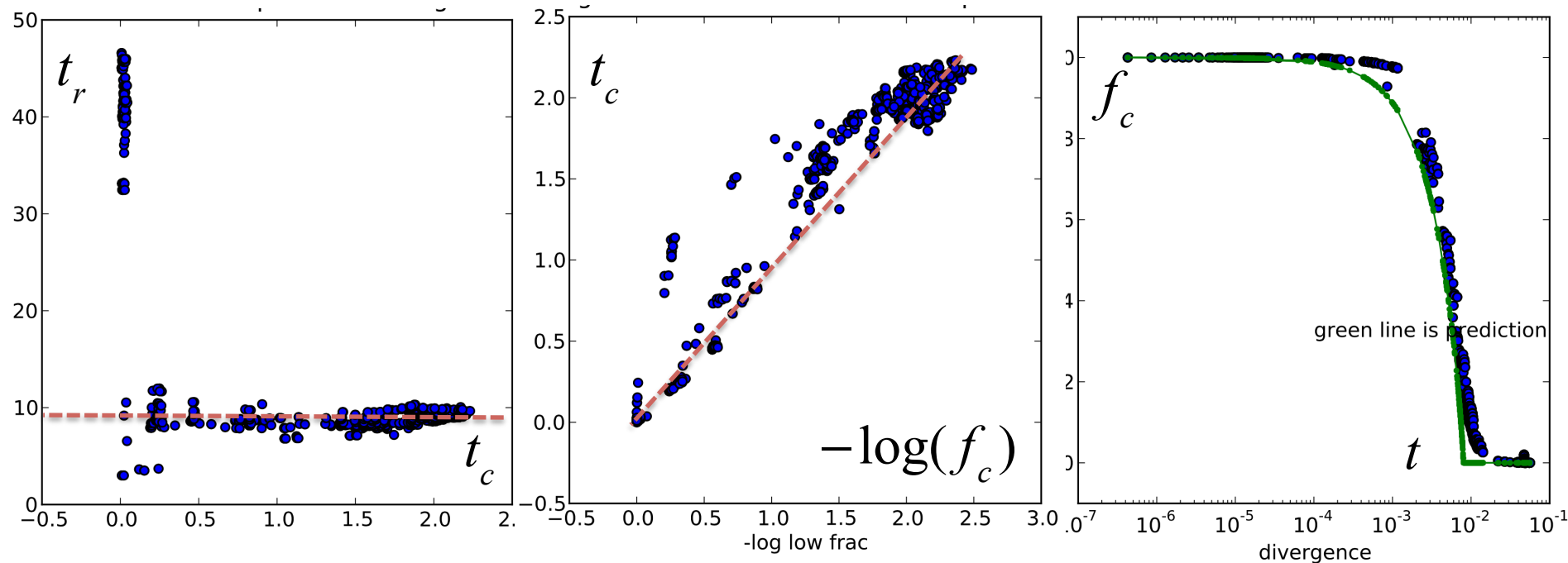
$$f_c = e^{-\rho t_c} = \frac{1}{2} \Rightarrow$$
$$t_c = -\frac{\log(0.5)}{\rho} \approx 0.001$$

Ratio of recombined to clonal mutations as divergence goes to zero:

$$\lim_{t_c \rightarrow 0} \frac{t_{recomb}}{t_{clonal}} = \frac{t_r (1 - e^{-\rho t_c})}{t_c e^{-\rho t_c}} = \rho t_r \approx 8.7$$

Clonally inherited DNA is lost at divergence an order of magnitude below typical divergence between strains.

Divergence of close strains is driven by recombination events.



$$\rho \approx 1000$$

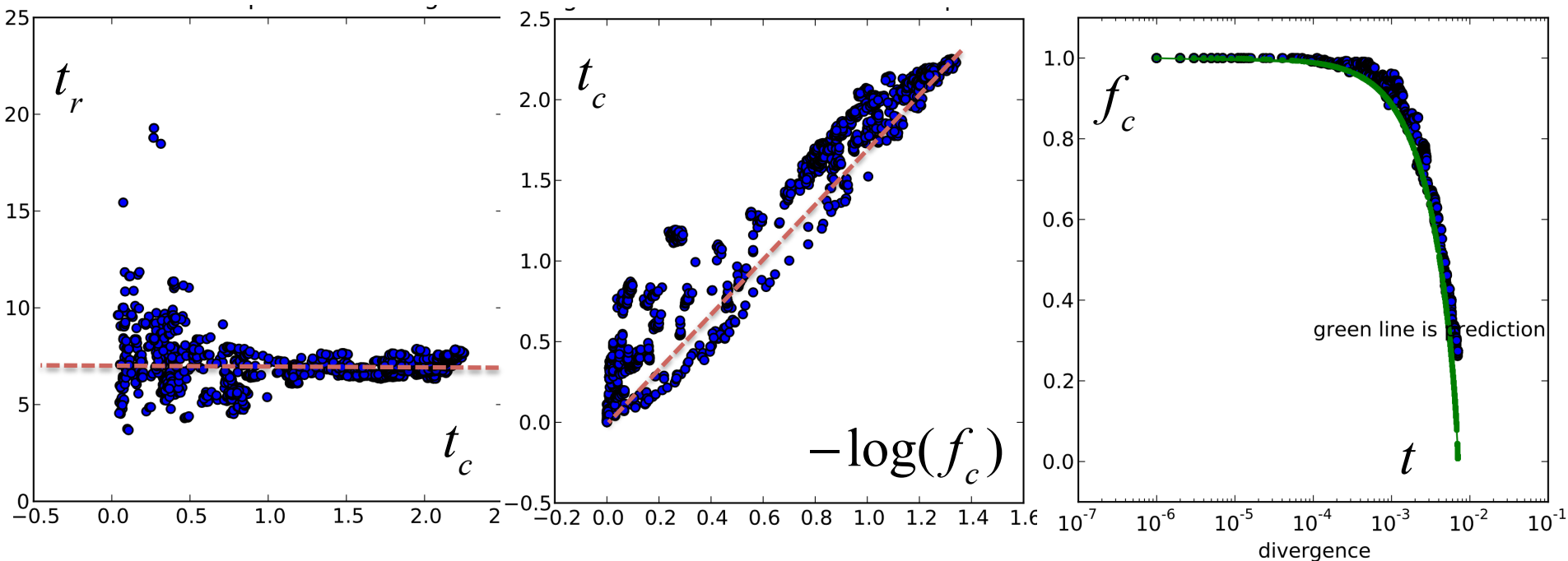
$$t_r \approx 0.008$$

Clonal SNP rate when half of the genome is recombined:

$$t_c = -\frac{\log(0.5)}{\rho} \approx 0.0007$$

Ratio of recombined to clonal mutations as divergence goes to zero:

$$\rho t_r \approx 8.0$$



$$\rho \approx 550$$

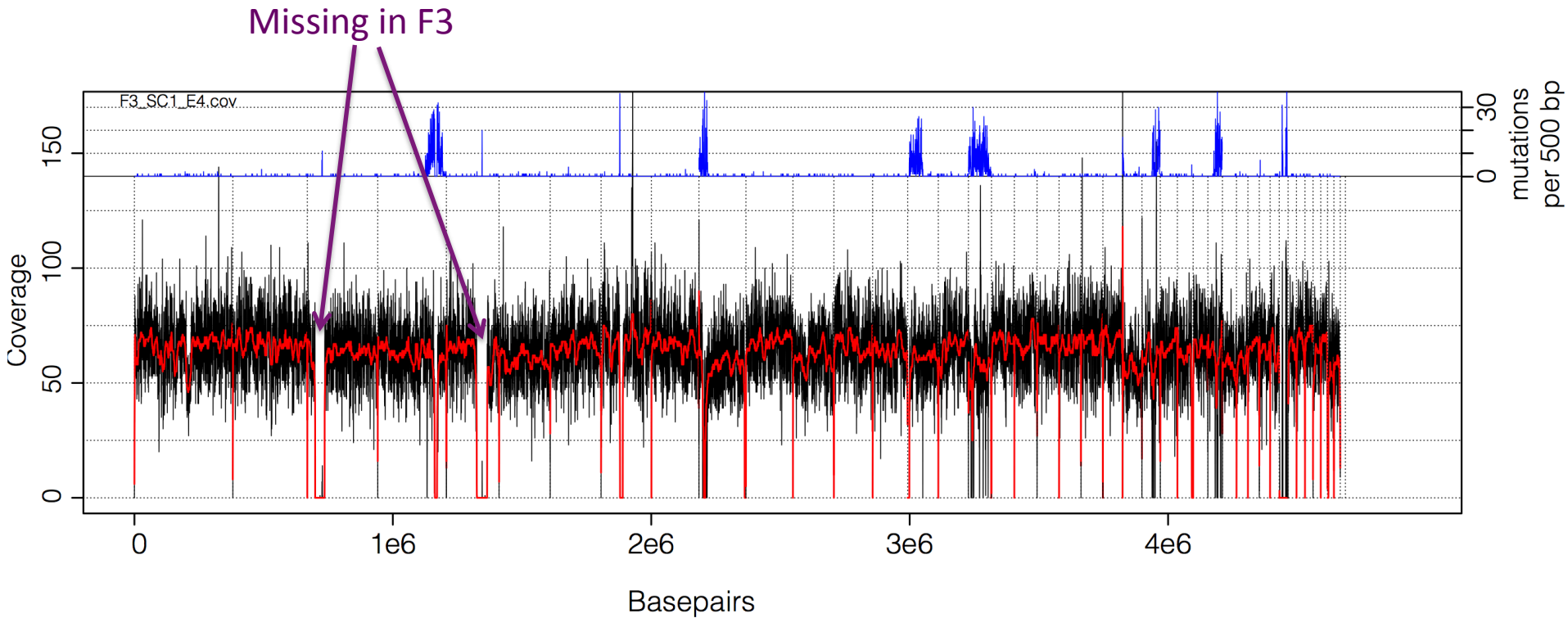
$$t_r \approx 0.007$$

Clonal SNP rate when half of the genome is recombined:

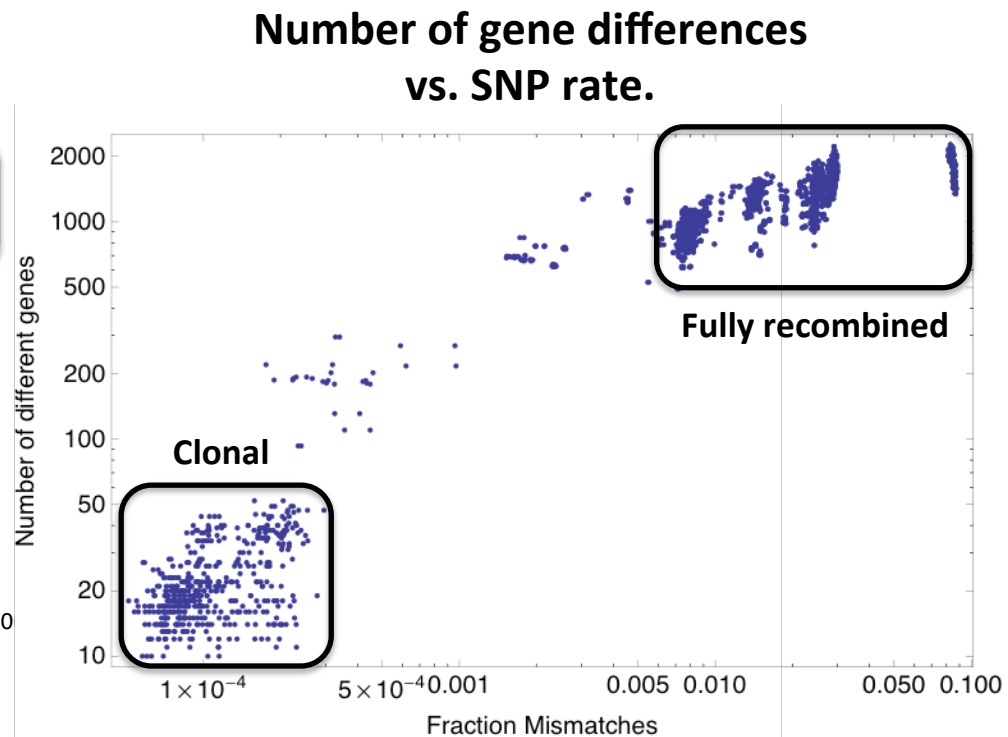
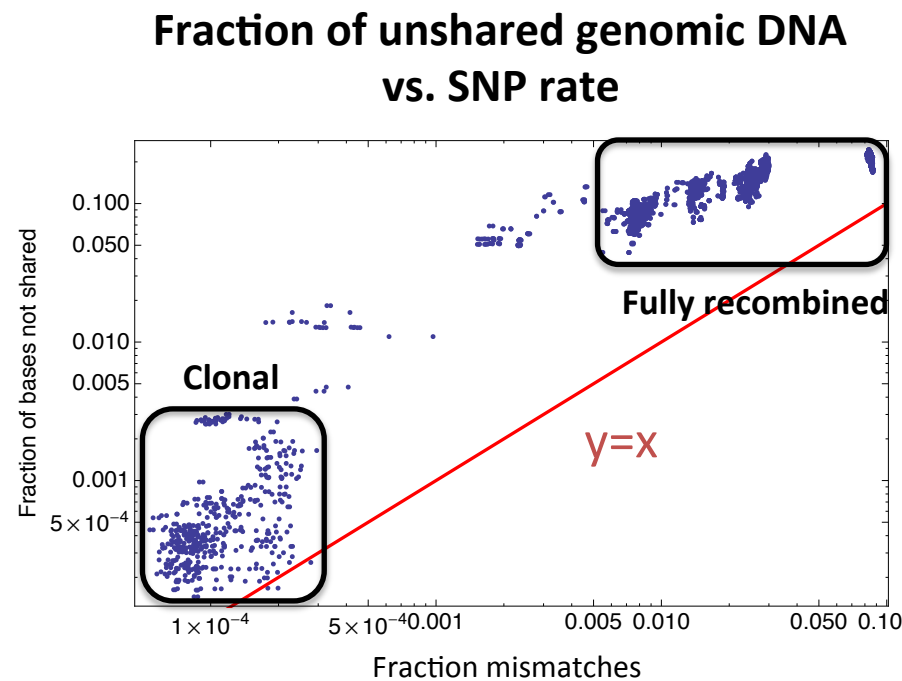
Ratio of recombined to clonal mutations as divergence goes to zero:

$$t_c = -\frac{\log(0.5)}{\rho} \approx 0.001$$

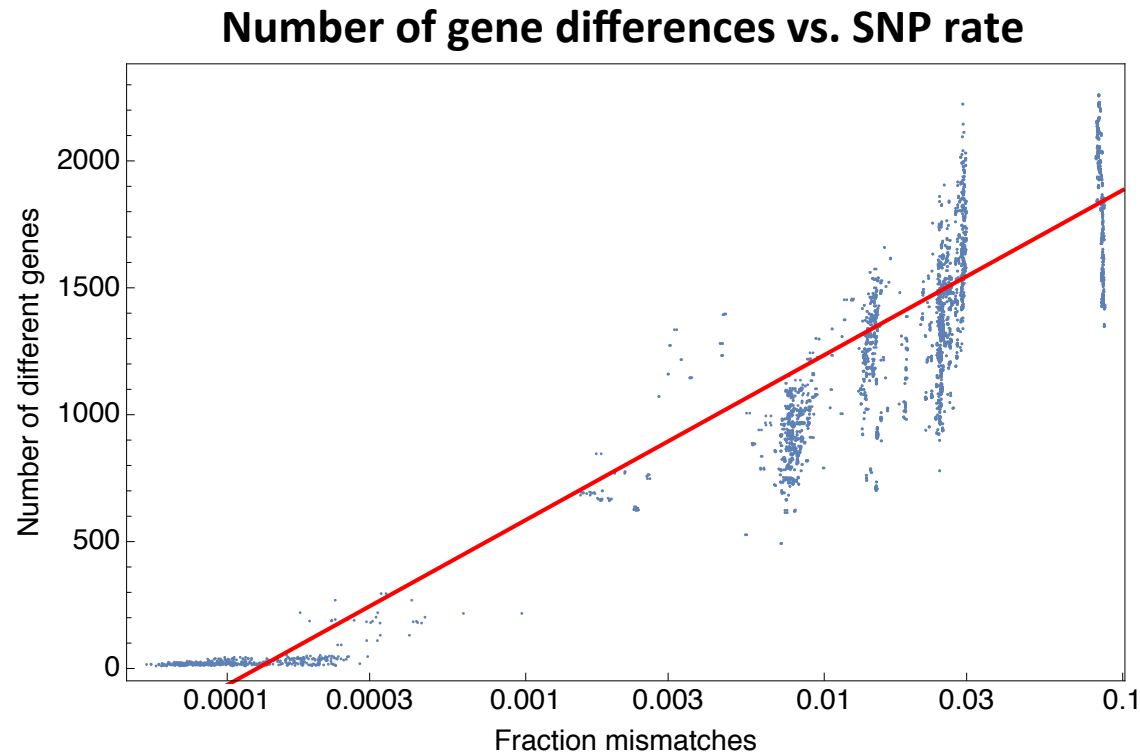
$$\rho t_r \approx 3.9$$



How does amount of genomic material that is unique to each strain scale with the divergence of a pair of strains?

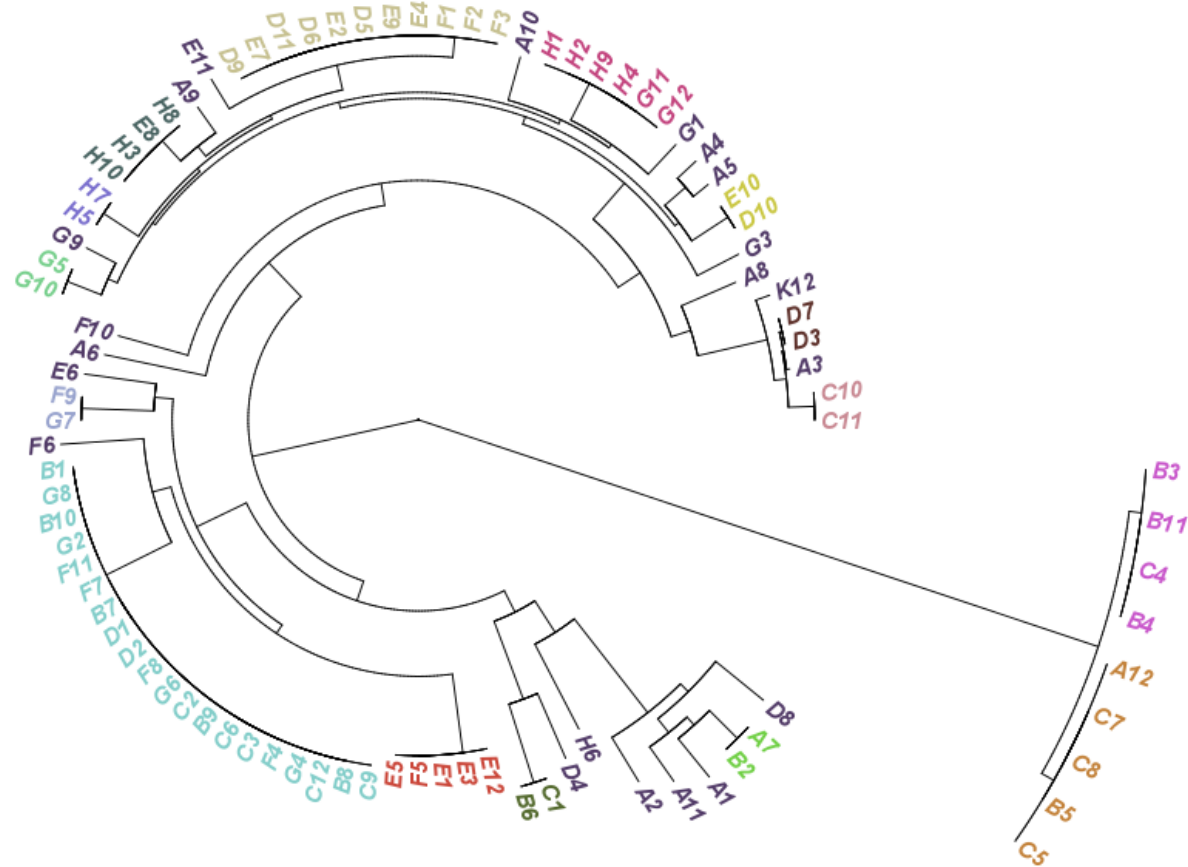


- Amount of differences due to unshared DNA is 10-fold higher than differences due to SNPs.
- Even clonal strains differ by 10-50 genes.
- Fully recombined strains differ in 500-2000 of 4500 genes.



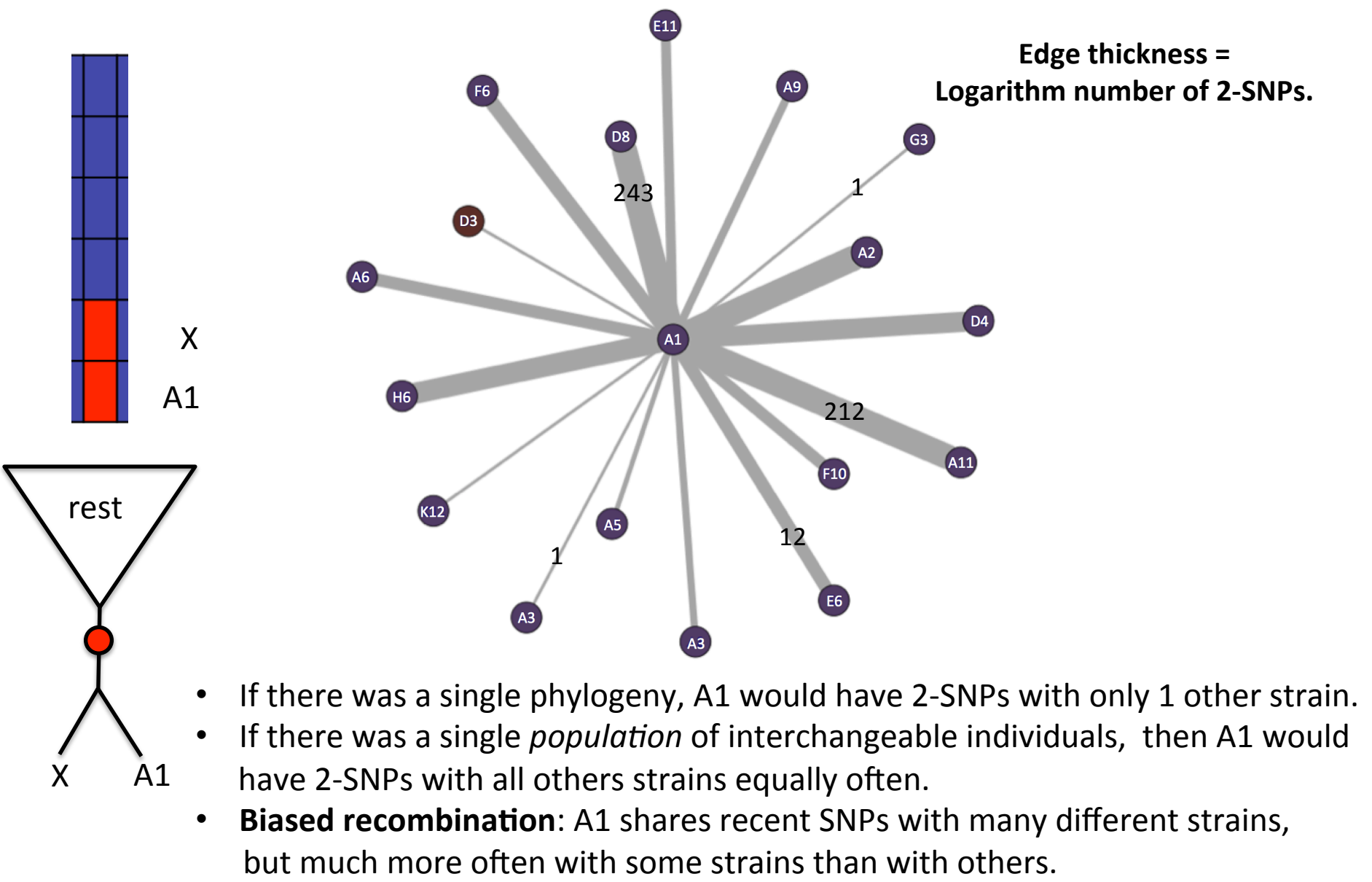
- For completely recombined pairs gene differences keep increasing with SNP rate.
- Increase roughly logarithmically.

Why is there an apparent phylogeny?

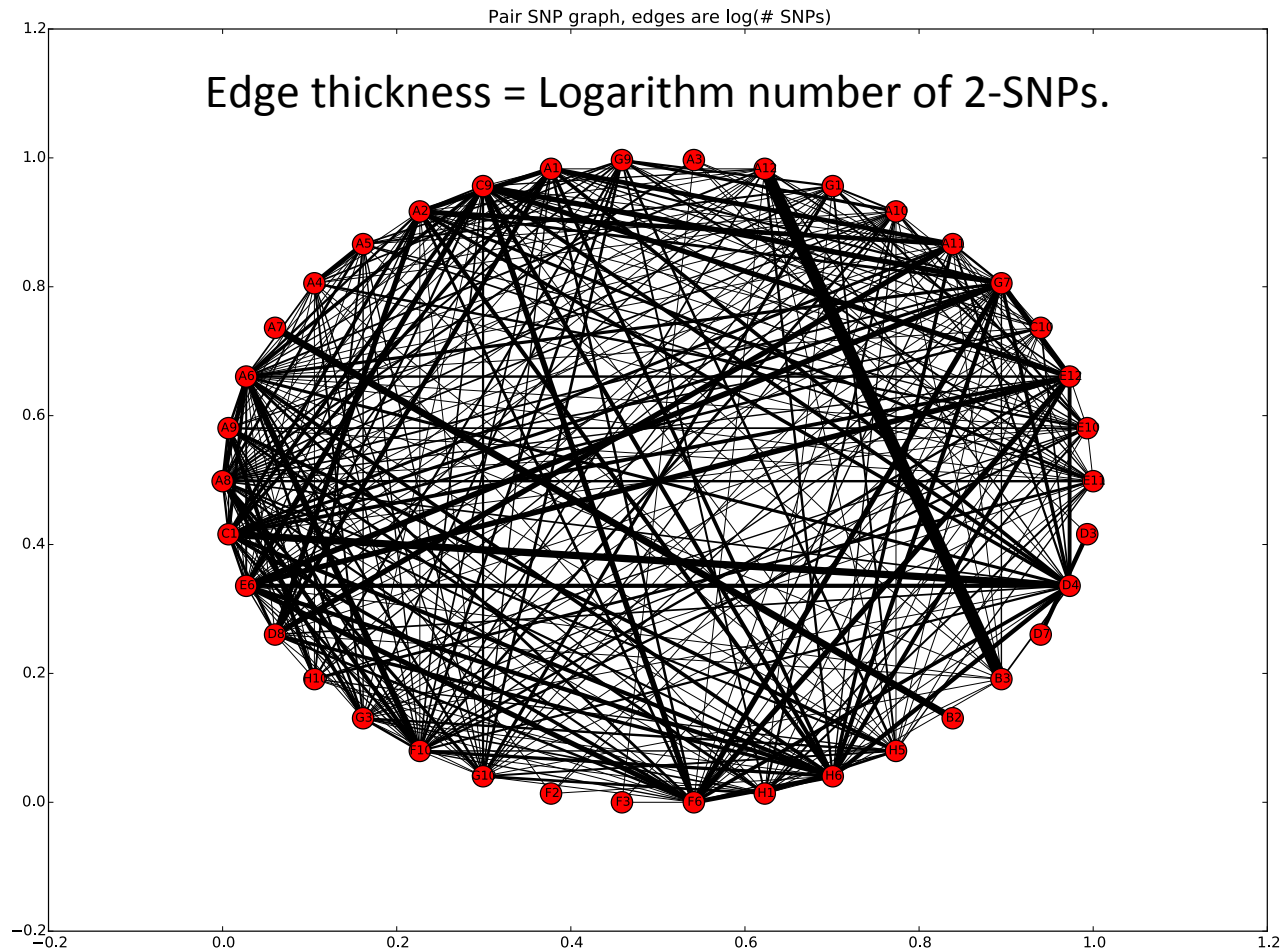


Why doesn't the tree look like a star-topology?

Why does one get a consistent phylogeny when using genome-wide alignments?

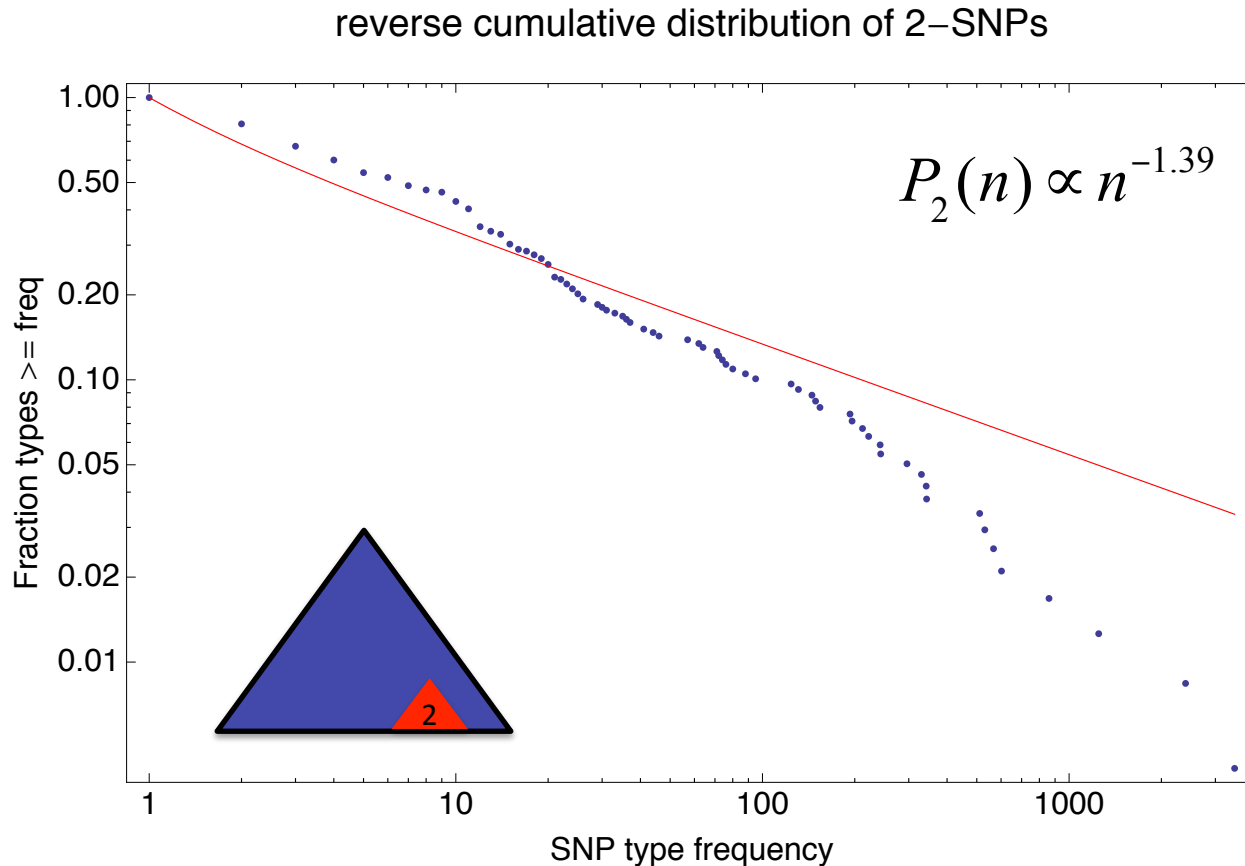


Graph of all 2-SNPs



Recombination: Each strain shares 2-SNPs with several other strains.

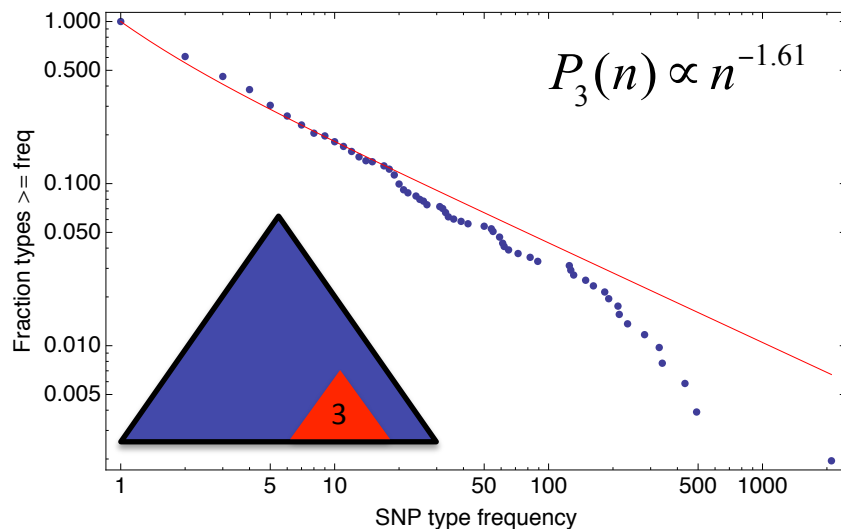
Unequal recombination rates: some pairs share 2-SNPs much more often than others.



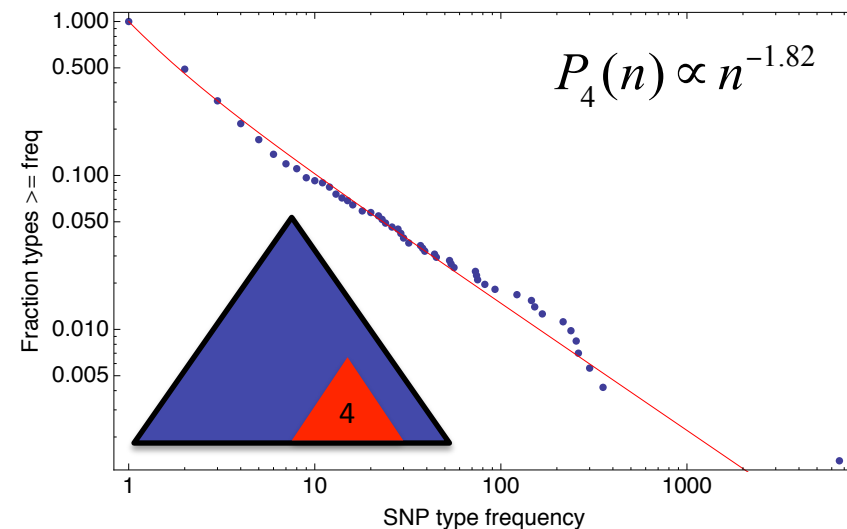
- Most pairs only share a SNP never or once, but some pairs share thousands of SNPs.
- Distribution is approximately power-law: no typical scale.
- **The picture of single recombining population, or separate subpopulations, is wrong.**
- Recombination rates of different pairs vary smoothly over > 3 orders of magnitude.

N-SNP frequency distributions are all long-tailed

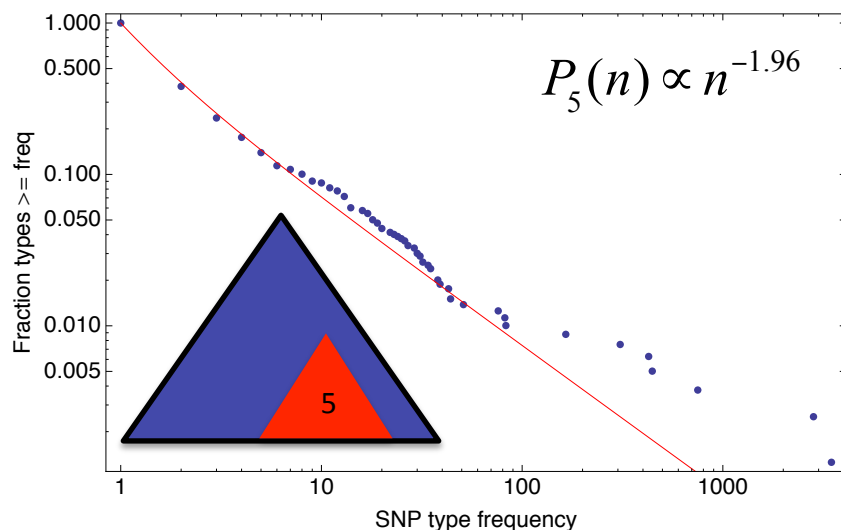
Reverse cumulative 3-SNPs



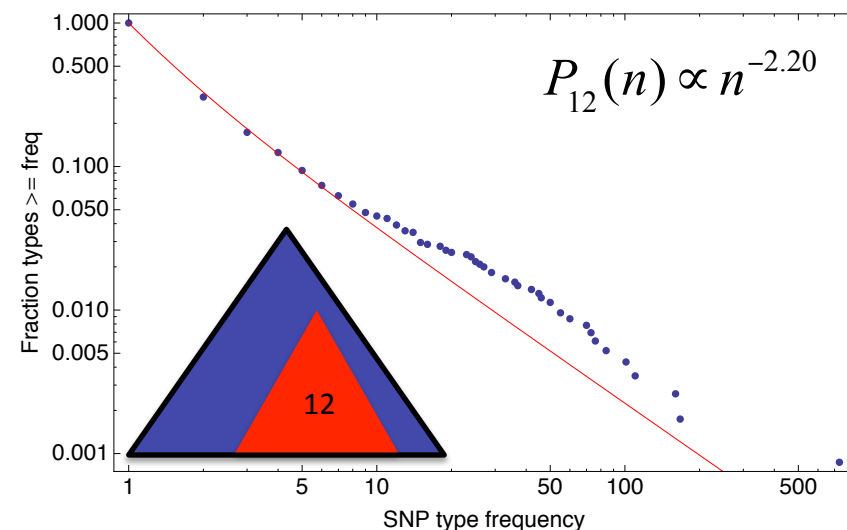
Reverse cumulative 4-SNPs



Reverse cumulative 5-SNPs



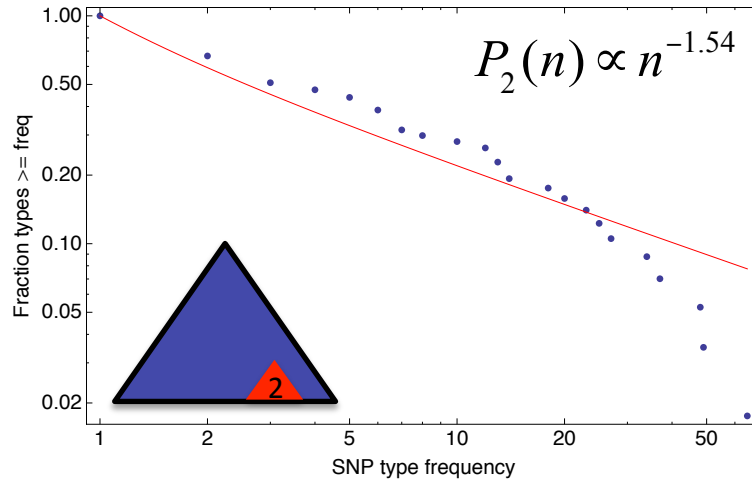
Reverse cumulative 12-SNPs



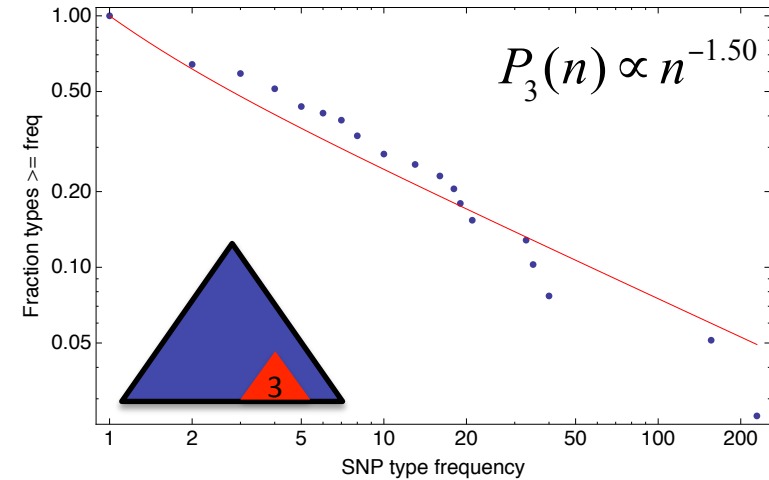
Other bacteria also show power-law N-SNP distributions

82 Chlamydia strains. 1 Mb core alignment. SNP rate: 0.015

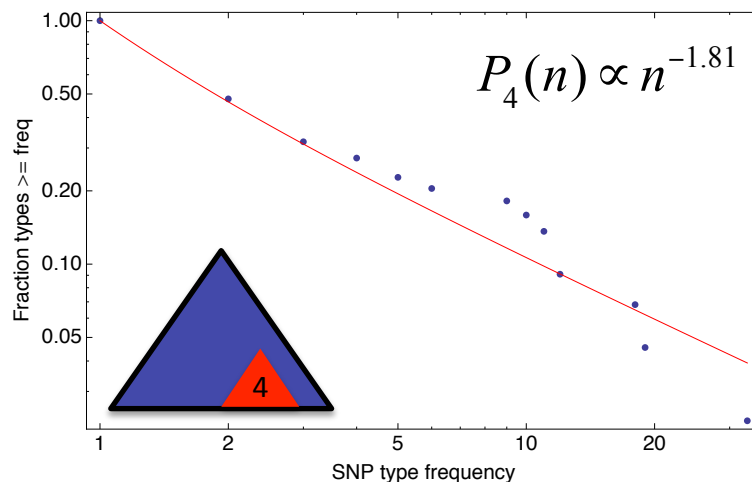
Reverse cumulative 2-SNPs



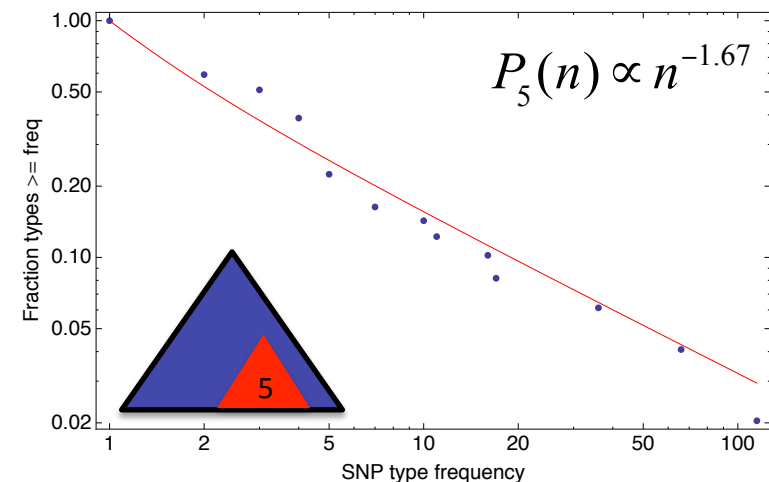
Reverse cumulative 3-SNPs



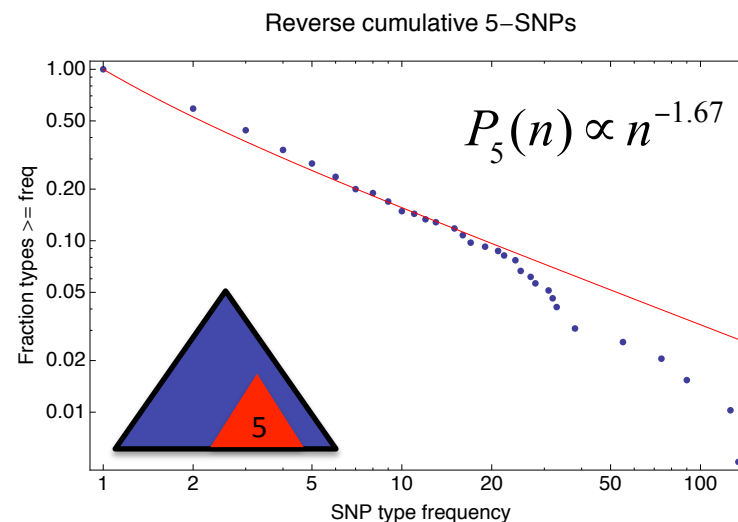
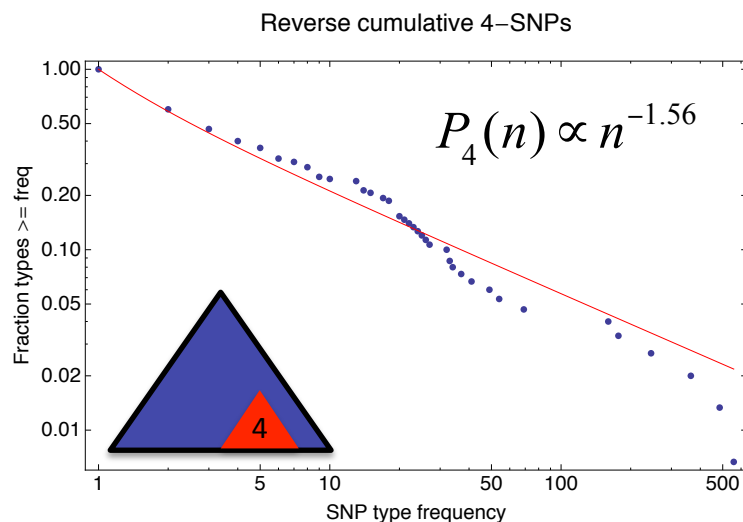
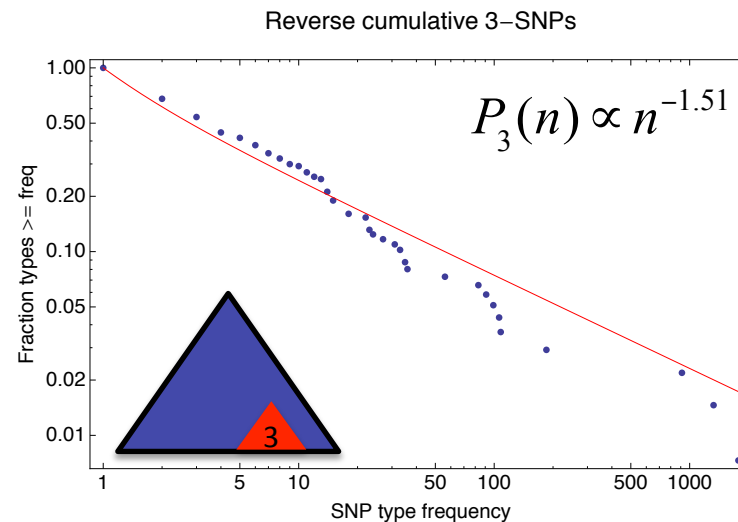
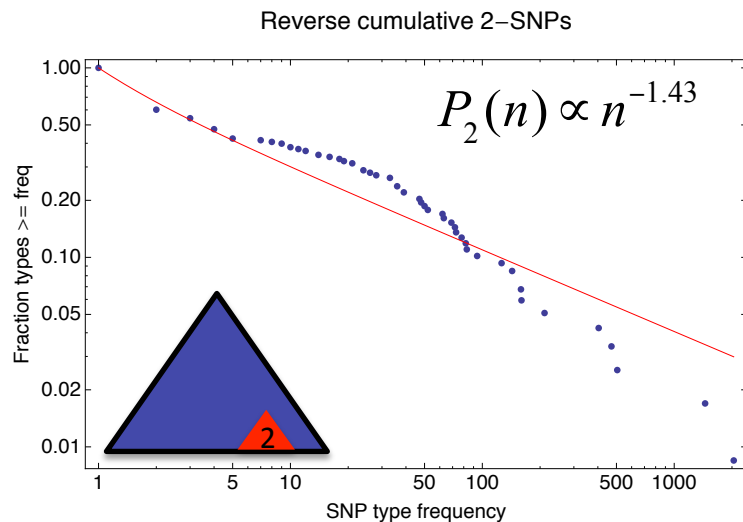
Reverse cumulative 4-SNPs



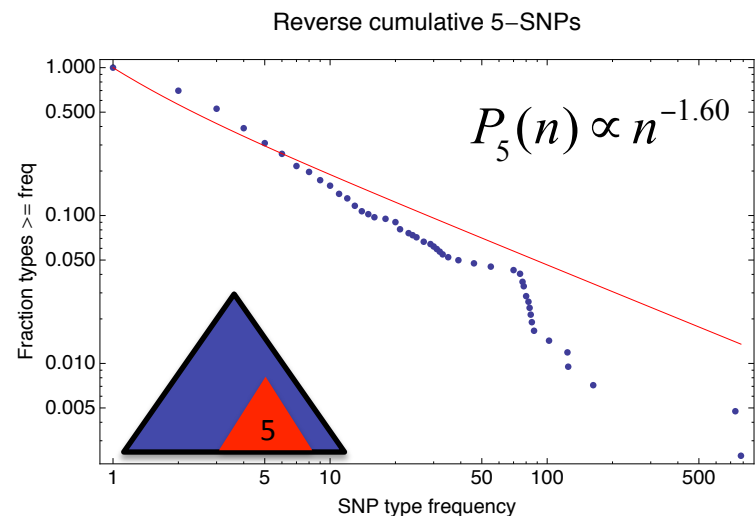
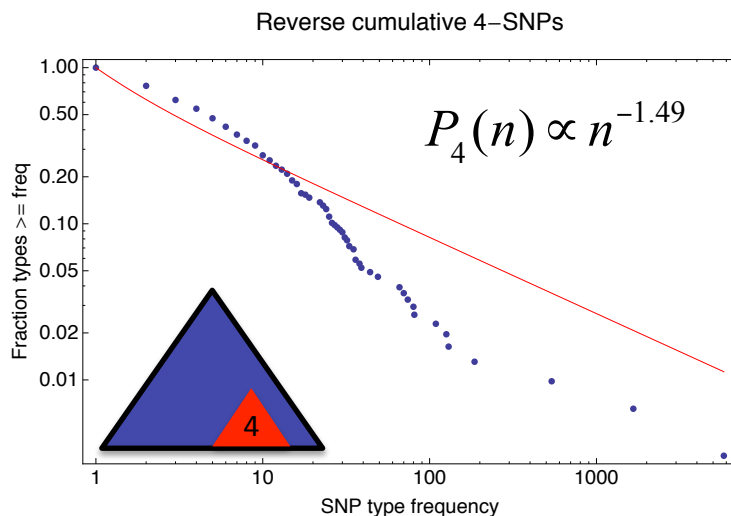
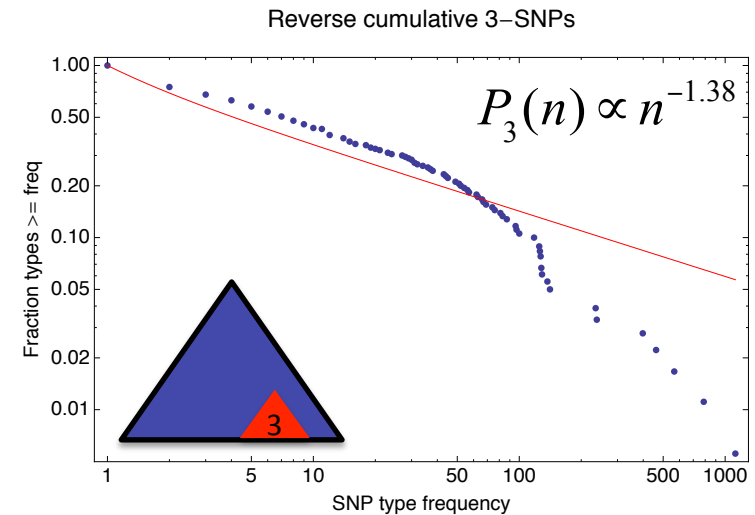
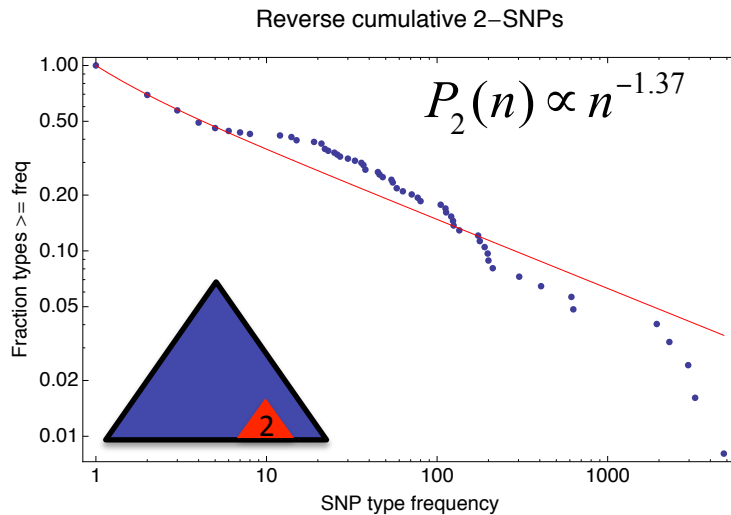
Reverse cumulative 5-SNPs

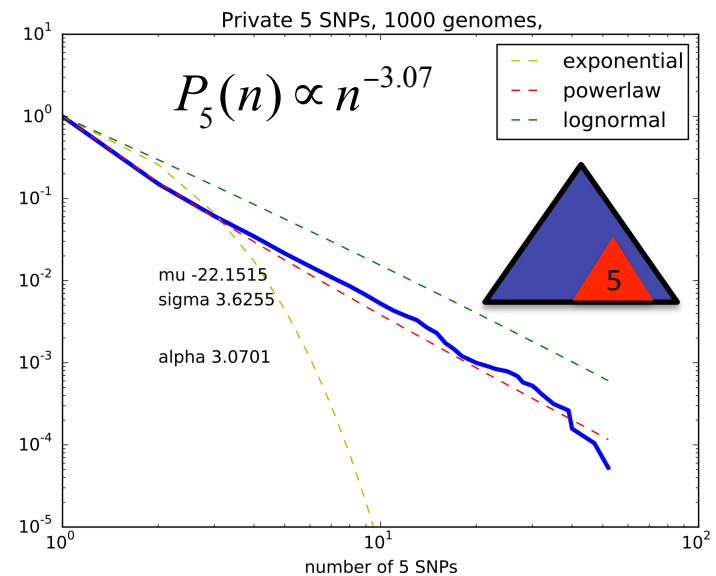
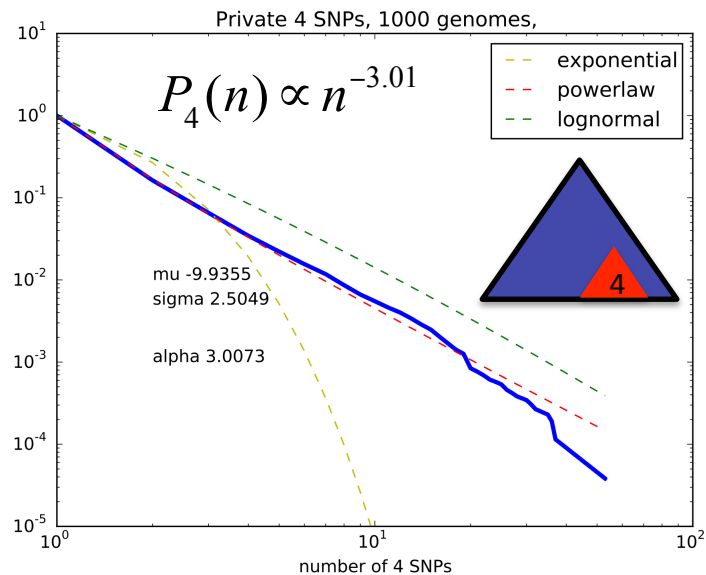
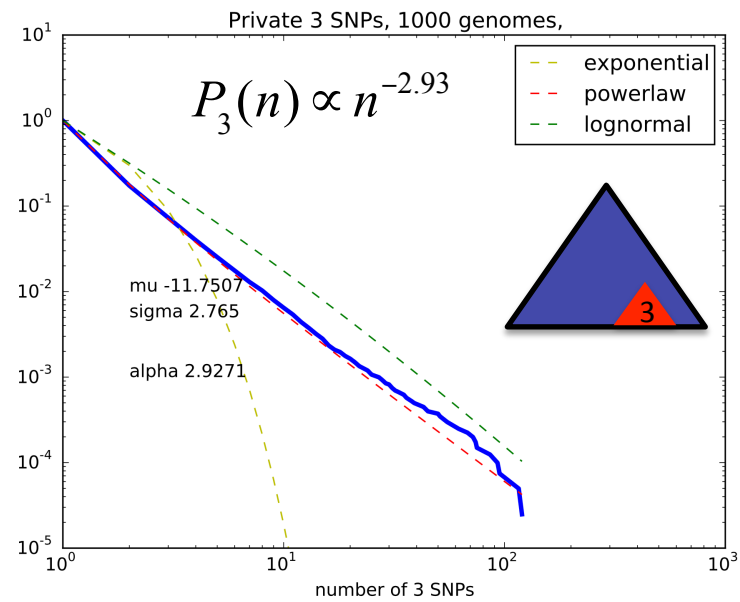
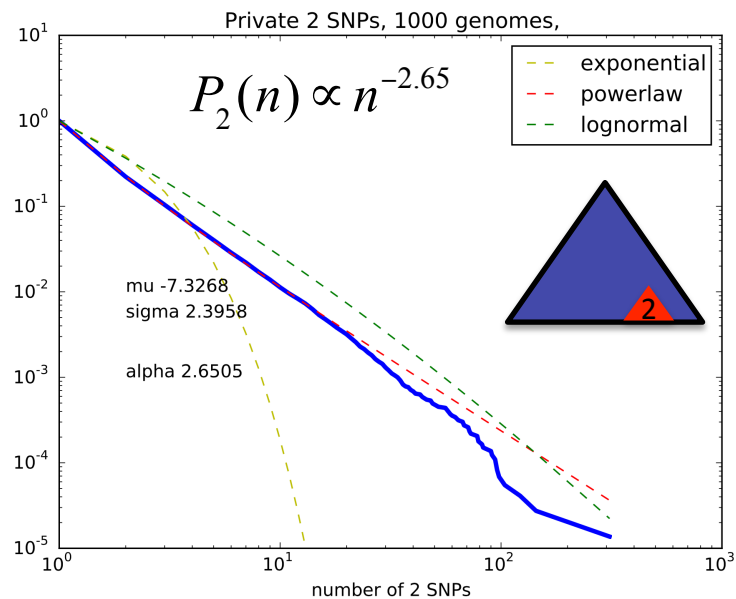


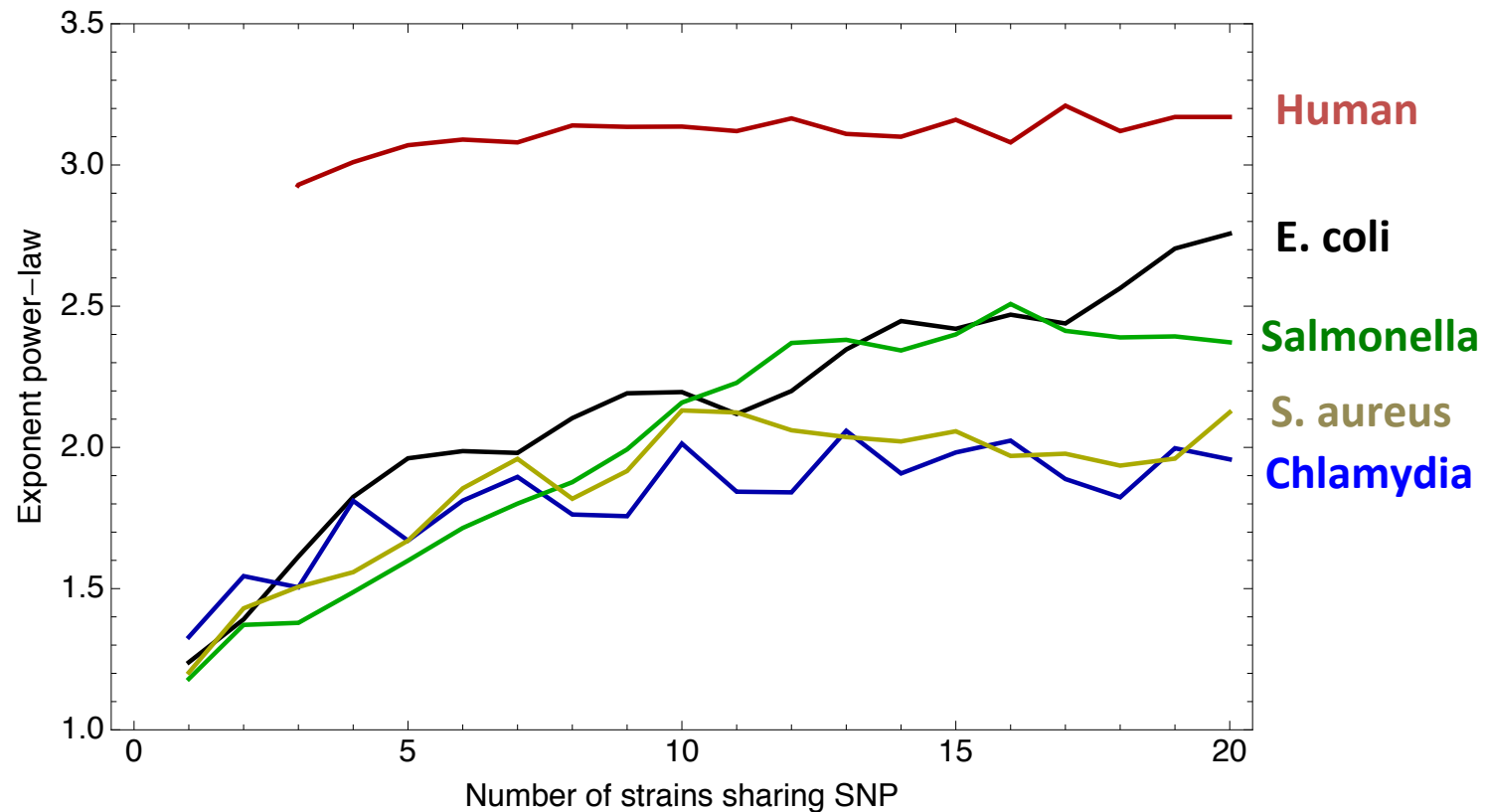
48 Staphylococcus aureus strains. 1.25 Mb core alignment. SNP rate: 0.09



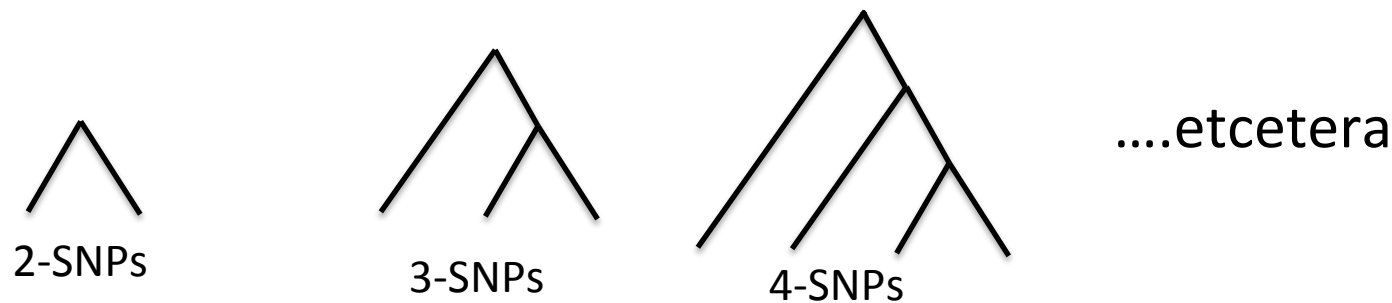
48 Salmonella strains. 1.25 Mb core alignment. SNP rate: 0.09







The N-SNP exponents encode population structure statistics as one goes further back in time.



Summary

- For most pairs of strains, none of the DNA in their alignment stems from their clonal ancestor, *all* has been recombined in at some point in their history.
- Recombination drives genome evolution: It introduces substitutions at almost 10-fold higher rate than point mutations.
- The clonal phylogeny of a species of bacteria *cannot* be reconstructed from DNA sequences.
- The apparent phylogenetic structure reflects *population structure*.
- There is not a single recombining population or separate subpopulations, but a long-tailed continuum of relative recombination rates.
- This population structure can be characterized by the exponents of the N-SNP frequency distributions.



Chris Field

Phenotyping, genome sequencing, assembly, gene annotation, orthology, and alignment.



Thomas Sakoparnig

Phylogeny and SNP statistics modeling