Gene regulation: From biophysics to evolutionary genetics

Michael Lässig

Institute for Theoretical Physics
University of Cologne

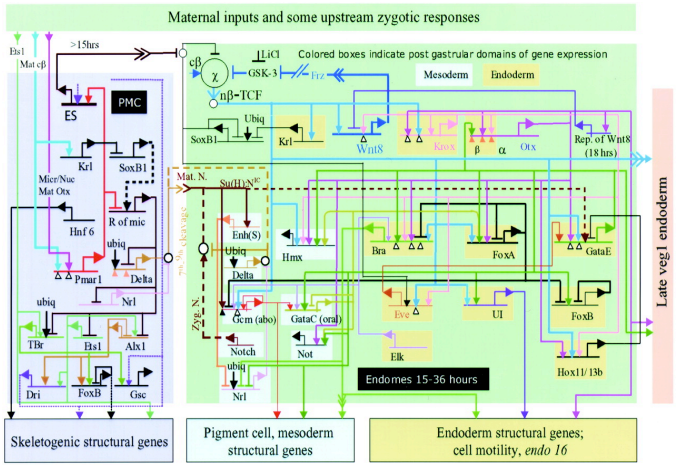
Thanks

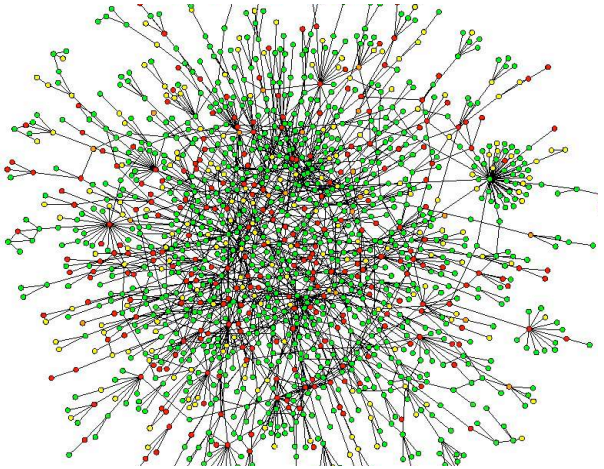
Ville Mustonen
Johannes Berg
Stana Willmann

Curt Callan (Princeton)
Justin Kinney (Princeton)
Luca Peliti (Naples)

SFB TR 12
SFB 680
European Research Training Network STIPCO

Structure and dynamics of molecular networks





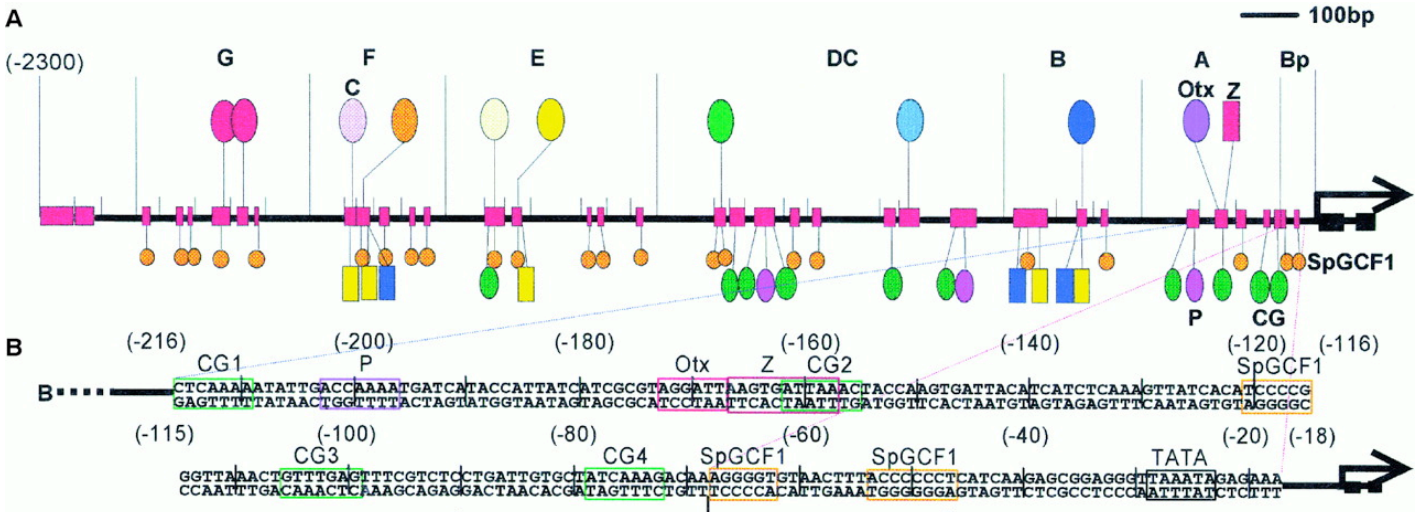
- Structure: Random parts?
  Functional design?

- Evolution: Pathways?
  Tempo?

1. Evolution of regulatory DNA

Genomic encoding of network interactions

- Multiple binding sites allow for complex regulation of individual genes in higher organisms:



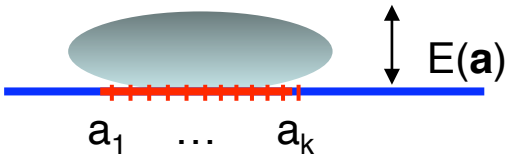
[Bolouri and Davidson, 2002]

- Input-output relation?

  Evolutionary dynamics?

Biophysics of transcriptional regulation

- **Transcription factor proteins** bind to **specific DNA sites** catalyzing transcription.



- **Binding energy** E(**a**) can be obtained from
  - low-throughput measurements [Fields et al. 97]
  - position weight matrix of functional sites [Berg and v.Hippel 86]
  - ChIP-chip data [Float et al. 05, Kinney et al. 06]
  - *high-throughput measurements* [Maerkl and Quake 07].

- E(**a**) depends on the site sequence **a** = (a_1,…,a_k):

$$E = \sum_{i=1}^{k} \epsilon_i(a_i) \quad + \text{ nonlinear terms?}$$

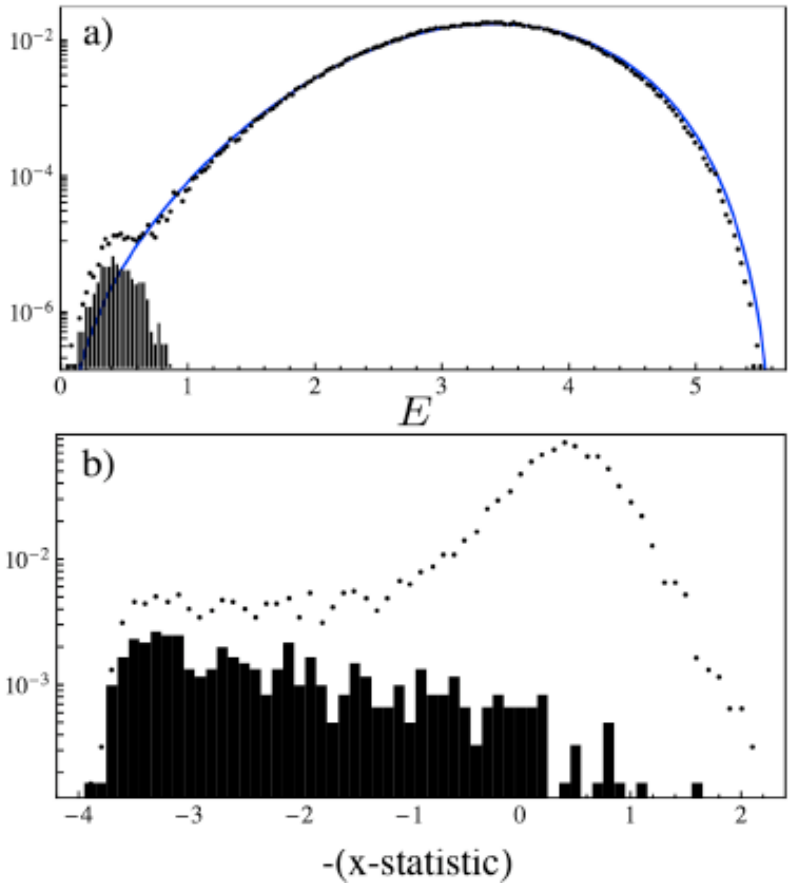
- E(**a**) is the **molecular phenotype** of a site, which **quantifies its functionality**.

Cis-regulatory elements: from sequence to phenotype

- The binding energy E(**a**) is the **molecular phenotype** of a site, which **quantifies its functionality**.

- This phenotype predicts **binding intensity** to promoters in yeast:

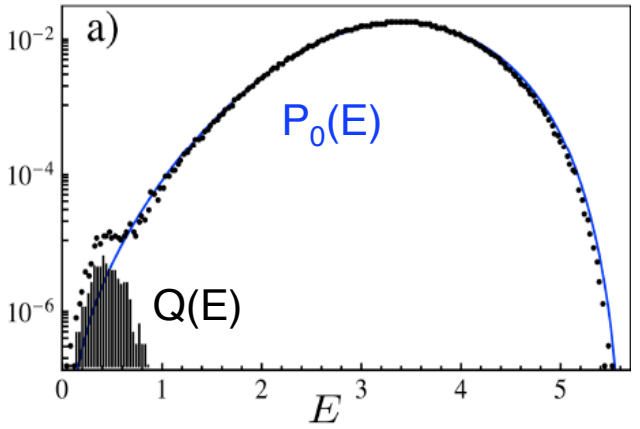
[Mustonen, Kinney, Callen, M.L., PNAS 2008]



ChIP-chip data for Abf1 binding in yeast
[Lee et al., Science 2002]

Cis-regulatory elements: from phenotype to fitness

- For broad-acting transcription factors, **high-affinity sites** ($E < E_b$) are statistically **overrepresented**.

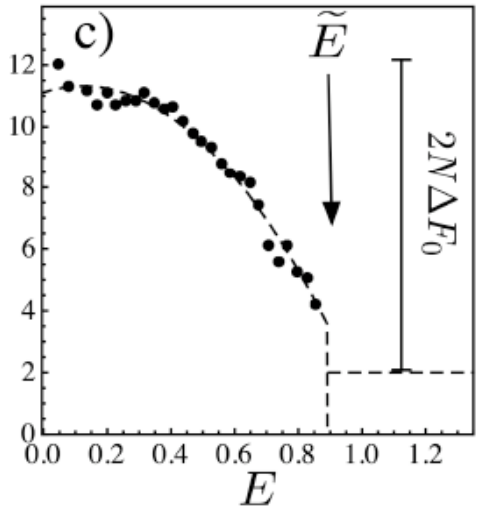


- At stationarity, the ensembles of functional and background sites determine the **average fitness landscape** F(E) of a site:

$$2NF(E) = \log \frac{Q(E)}{P_0(E)} + \text{const.}$$

- This predicts a **moderate fitness effect per functional site:**

$$2NF_0 \approx 10$$



Abf1 binding sites in S. cerevisiae

[Berg, Willmann, M.L., BMC Evol. Biol. 2004,
Mustonen, M.L., PNAS 2005, Mustonen, Kinney, Callen, M.L., PNAS 2008]

Population genetics

- **Selection**: sequence state **a** has fitness

$$F(\mathbf{a}) = \frac{d}{dt}\langle \log N(\mathbf{a})\rangle_{\mu=0} - \text{const.}$$

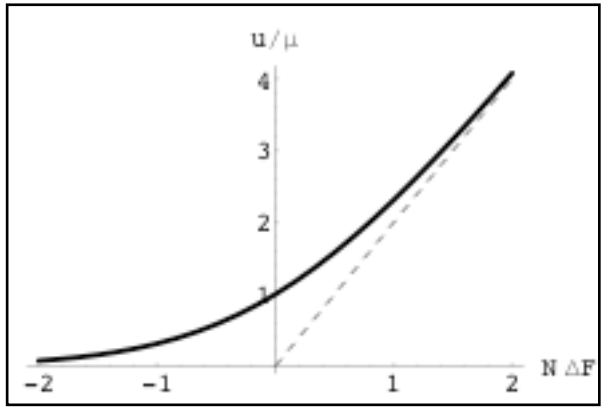
- **Point mutations:**

$$\mathbf{a} = (\ldots, a, \ldots) \rightarrow \mathbf{b} = (\ldots, b, \ldots)$$

Population genetics

- **Genetic drift:**

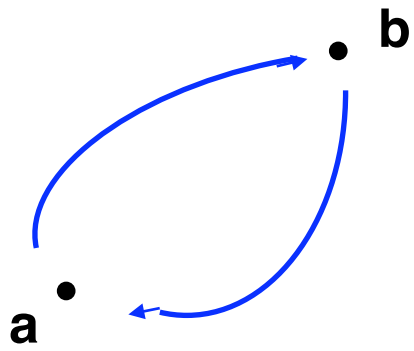
  Kimura-Ohta substitution rates



$$u_{\mathrm{a}\to\mathrm{b}} = \mu_{\mathrm{a}\to\mathrm{b}} \frac{1-\exp[-2(F(\mathrm{a})-F(\mathrm{b}))]}{1-\exp[-2N(F(\mathrm{a})-F(\mathrm{b}))]}$$

Ratio of forward and backward rates:

$$\frac{u_{\mathrm{a}\to\mathrm{b}}}{u_{\mathrm{b}\to\mathrm{a}}} = \frac{\mu_{\mathrm{a}\to\mathrm{b}}}{\mu_{\mathrm{b}\to\mathrm{a}}} \exp[2N(F(\mathrm{b}) - F(\mathrm{a}))]$$



Population genetics

- **Evolutionary equilibria in sequence space:**

  Given two families of loci,
  - **background loci** with stationary sequence distribution $P_0(\mathbf{a})$ under neutral evolution
  - **functional loci** with stationary sequence distribution $Q(\mathbf{a})$ under selection

    the fitness landscape $F(\mathbf{a})$ for the functional loci is given by

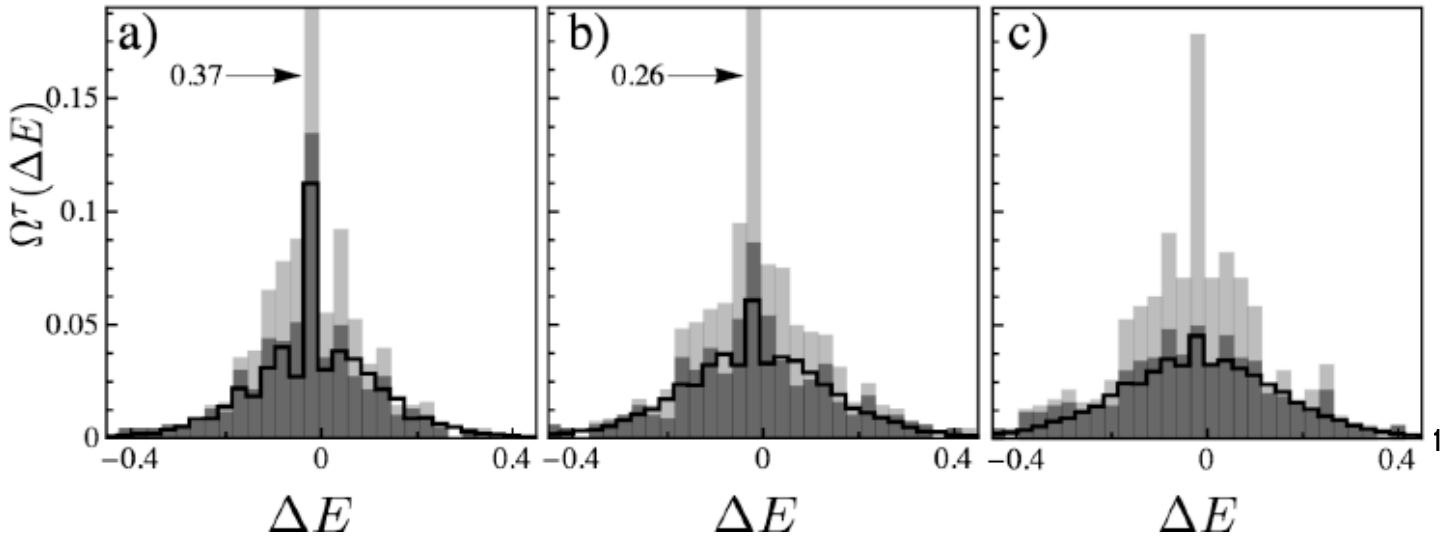
$$Q(\mathbf{a}) = P_0(\mathbf{a}) \exp[2NF(\mathbf{a}) + \text{const.}]$$

  N: effective population size.

[J.Berg, S. Willmann, M.L., **BMC Evol. Biol.** (2004)]
[V. Mustonen, M.L., **Proc. Natl. Acad. Sci.** (2005)]

Phenotypic evolution of binding sites

- The inferred fitness landscape quantitatively **predicts the evolution of the phenotype E**:

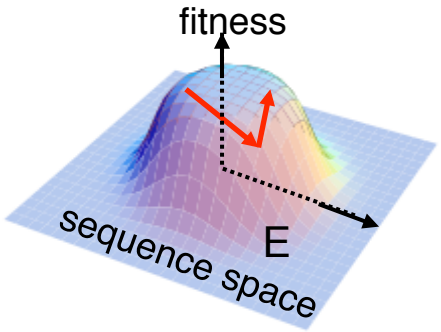


Abf1 binding energy differences of sites in S. cerevisiae, S.paradoxus, S. mikatae, S. bayanus

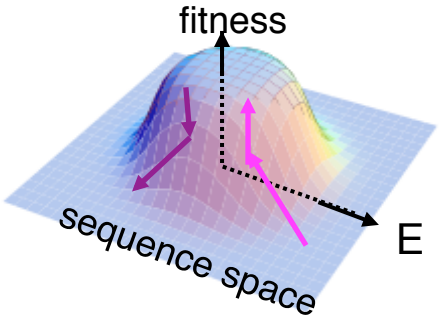
[Mustonen, Kinney, Callen, M.L., PNAS 2008]

Pathways of promoter evolution

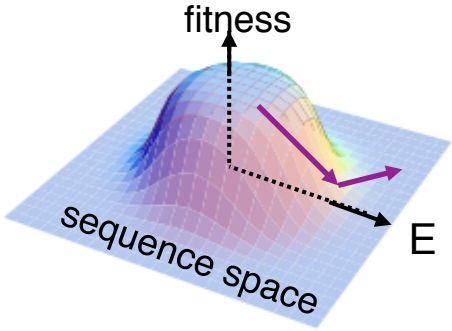
- **Conservation of site** and **function**



- **Site turnover**



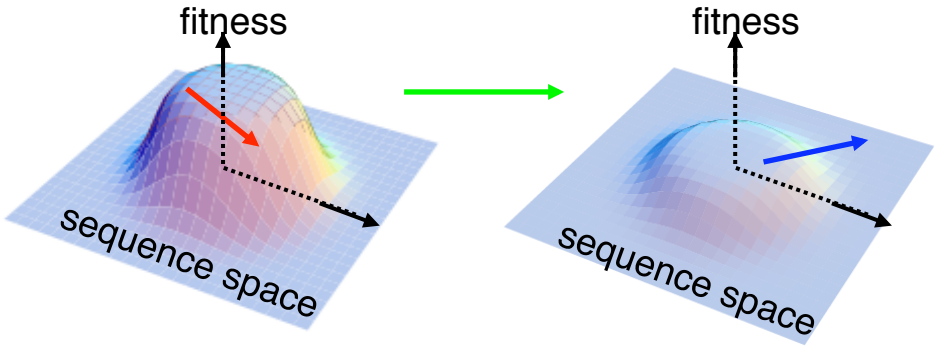
+ conservation of function



+ turnover of function

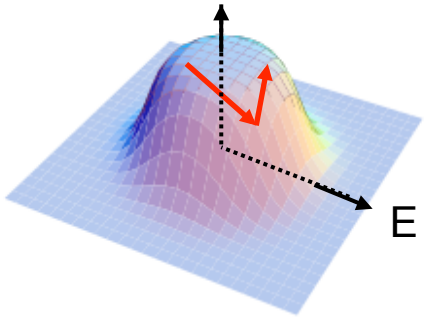
- **Time-dependent selection**

+ adaptation



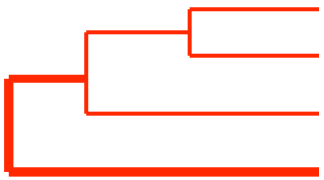
Conservation of binding sites

- Sequences of conserved sites evolve by **compensatory mutations**:



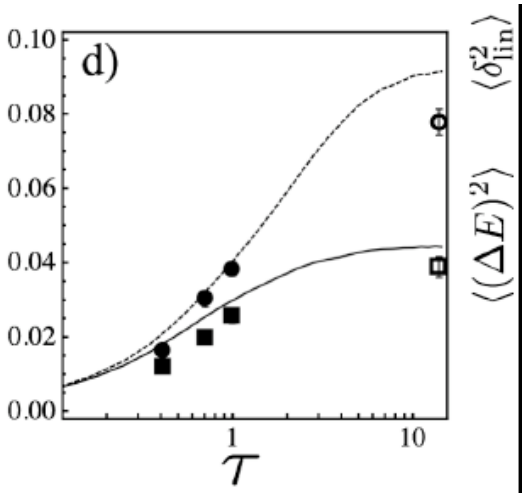
$$\Delta E = \sum_i \Delta\epsilon_i \qquad \text{but} \qquad \mathrm{var}(\Delta E) < \sum_i \mathrm{var}(\Delta\epsilon_i)$$

- Hence, the **energy phenotype is more constrained than the site sequence**:



[cf. Kellis et al., Nature 2003]

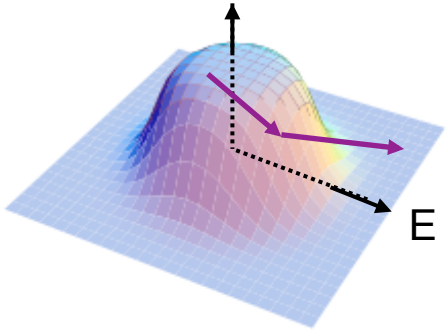
[Mustonen, Kinney, Callen, M.L., PNAS 2008]



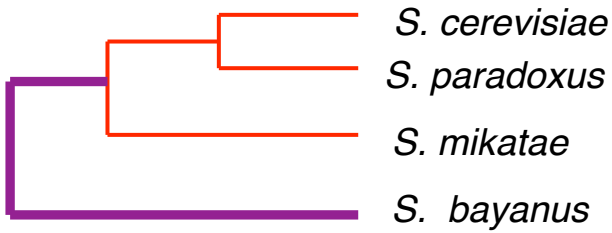
divergence time from *cer*

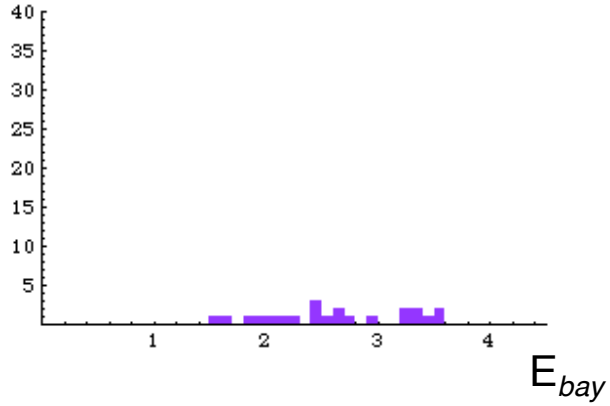
Loss and gain of function

- **Turnover** of **promoter function** determines loss and gain of regulatory interactions:



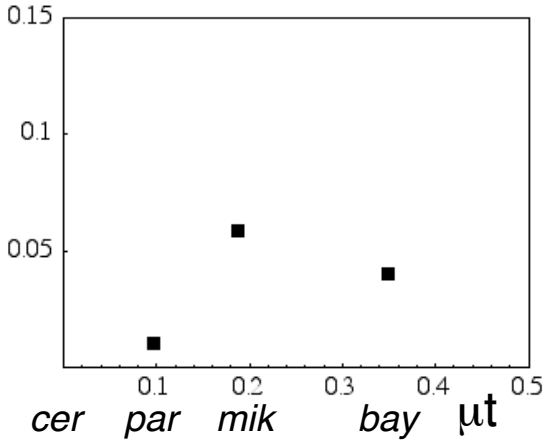
- **Species-specific loss of sites**:





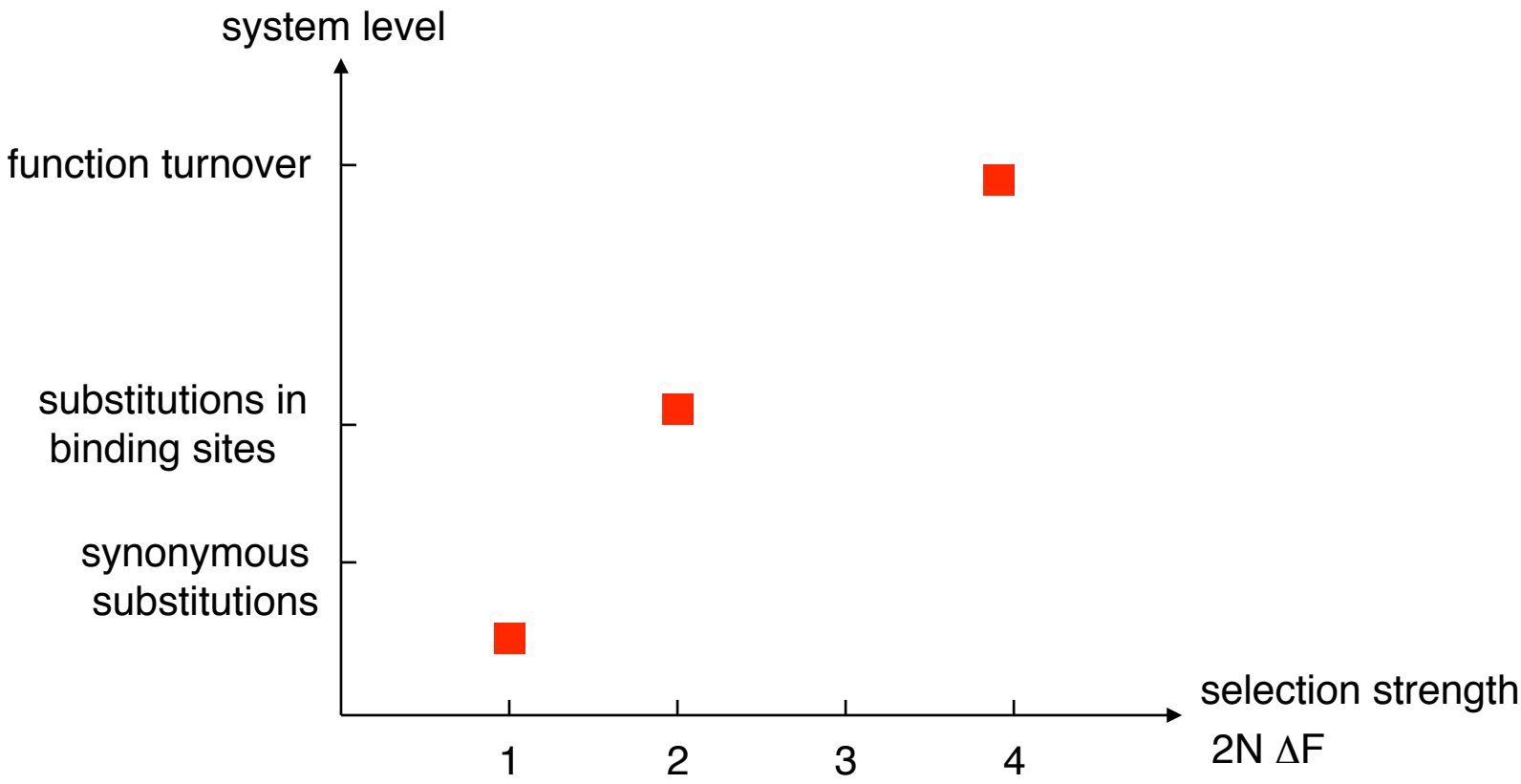
- **Functional turnover rate**

$$\gamma_f \sim 0.1\ \mu$$



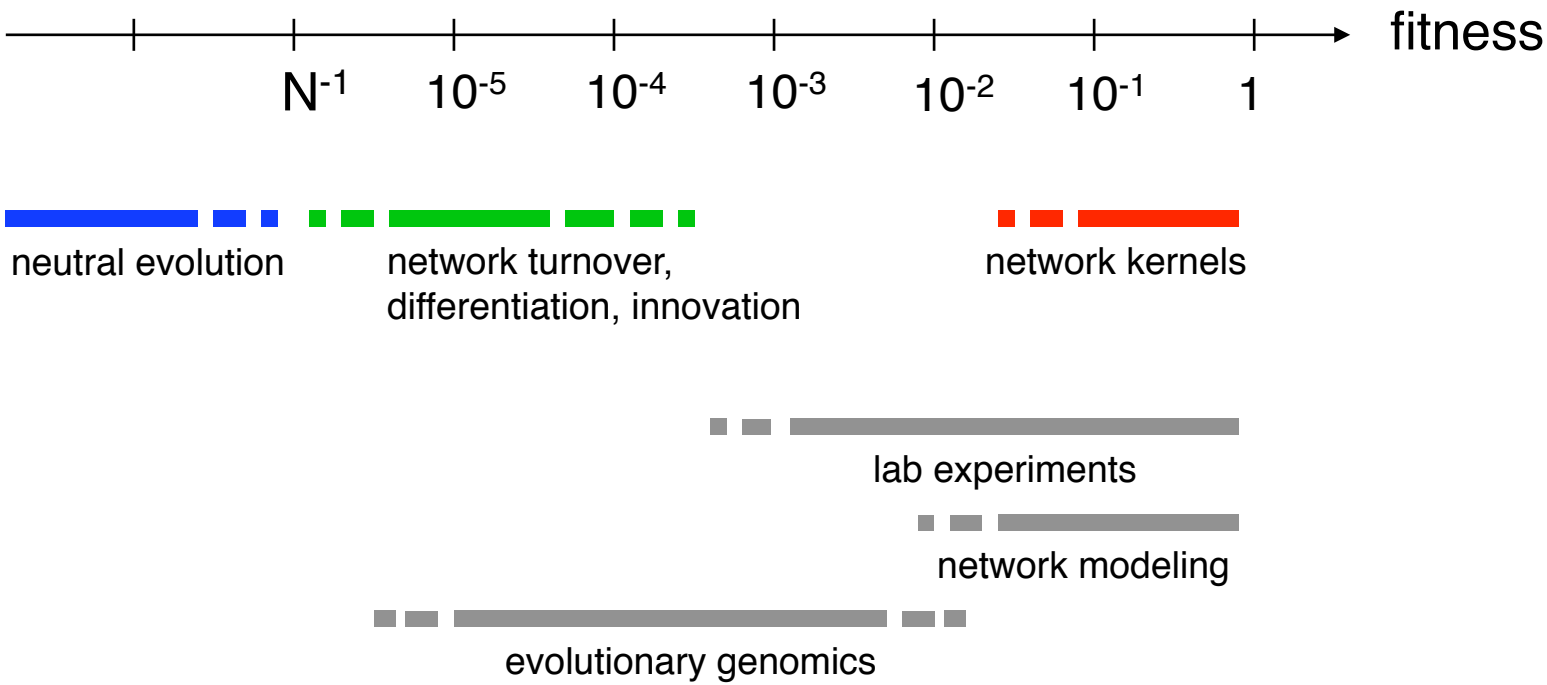
[J. Kinney, V. Mustonen, C. Callan, M.L., PNAS 2008]

- **Natural selection acts on complex systems in a scale-dependent way:**



Evolutionary systems biology

- **Laboratory experiments, modeling**, and **evolutionary genomics** address **complementary aspects** of biological systems:



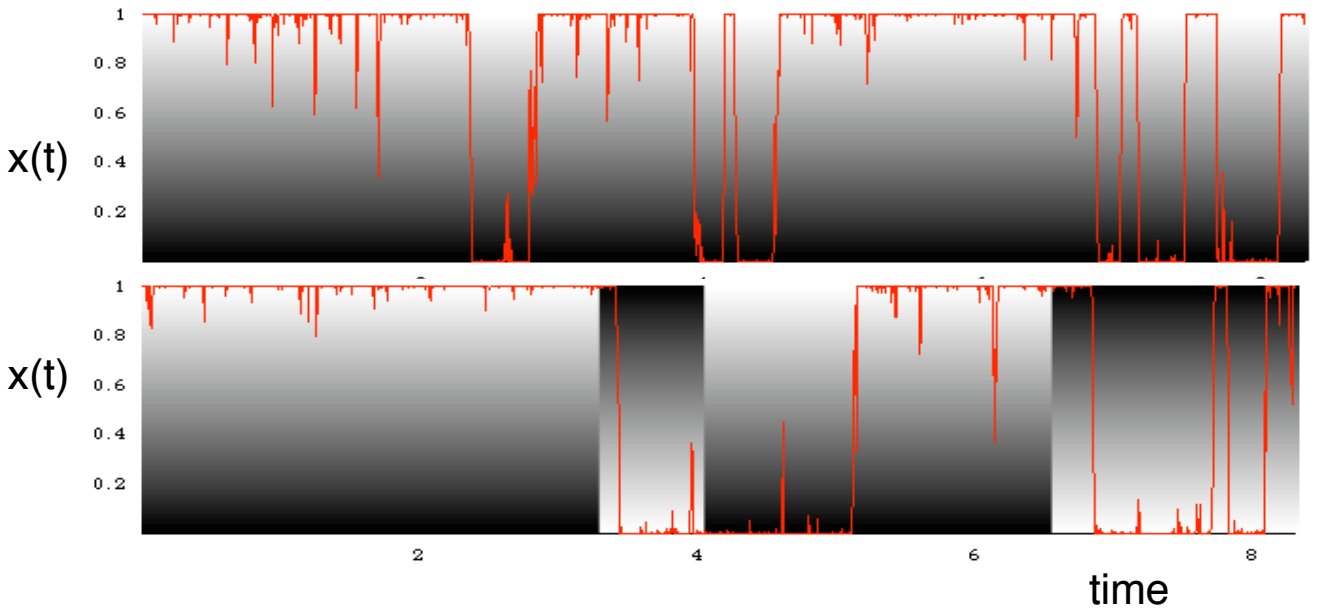
2. Evolution of the Drosophila genome

From fitness landscapes to seascapes

- Phenotypic concept of Darwinian selection:
  newly arising selection and response by **adaptation**.

- Can we trace the **time-dependence** of selection
  in genomic data?

Genome evolution under constant and fluctuating selection



- Allele frequency x(t) evolves under
  selection, mutations, stochastic fluctuations (genetic drift).

- **Constant selection** leads to **evolutionary equilibrium,** p_{eq} (x).

- **Fluctuating selection**

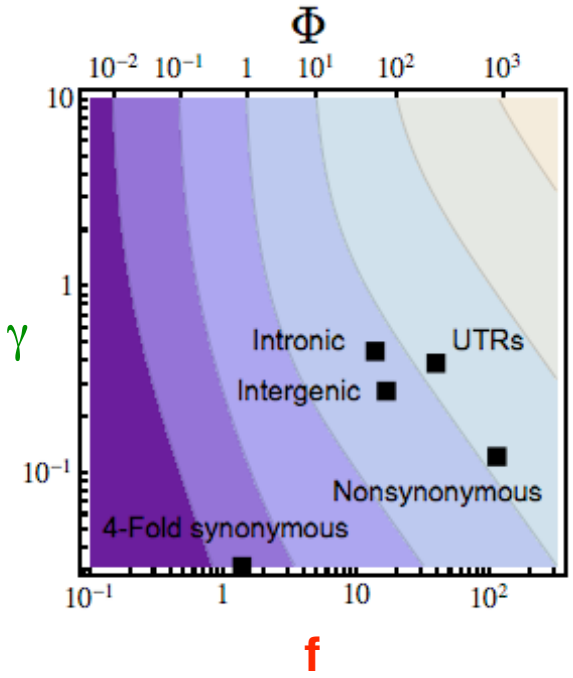
  $\Delta F(t) = f\,\chi(t)$ with **switching rate** γ,

  leads to **adaptation:** excess number of uphill mutations w/r to equilibrium.

Adaptation and fitness flux

- Substitutions and polymorphism spectra [Glinka et al 2003, Ometto et al 2005] are used to infer a **surplus of beneficial over deleterious substitutions.**

- Adaptation is quantified by a positive
  fitness flux = (substititution rate) x
  (average selection coefficient of substitutions).



[Mustonen and M.L, PNAS 2007]

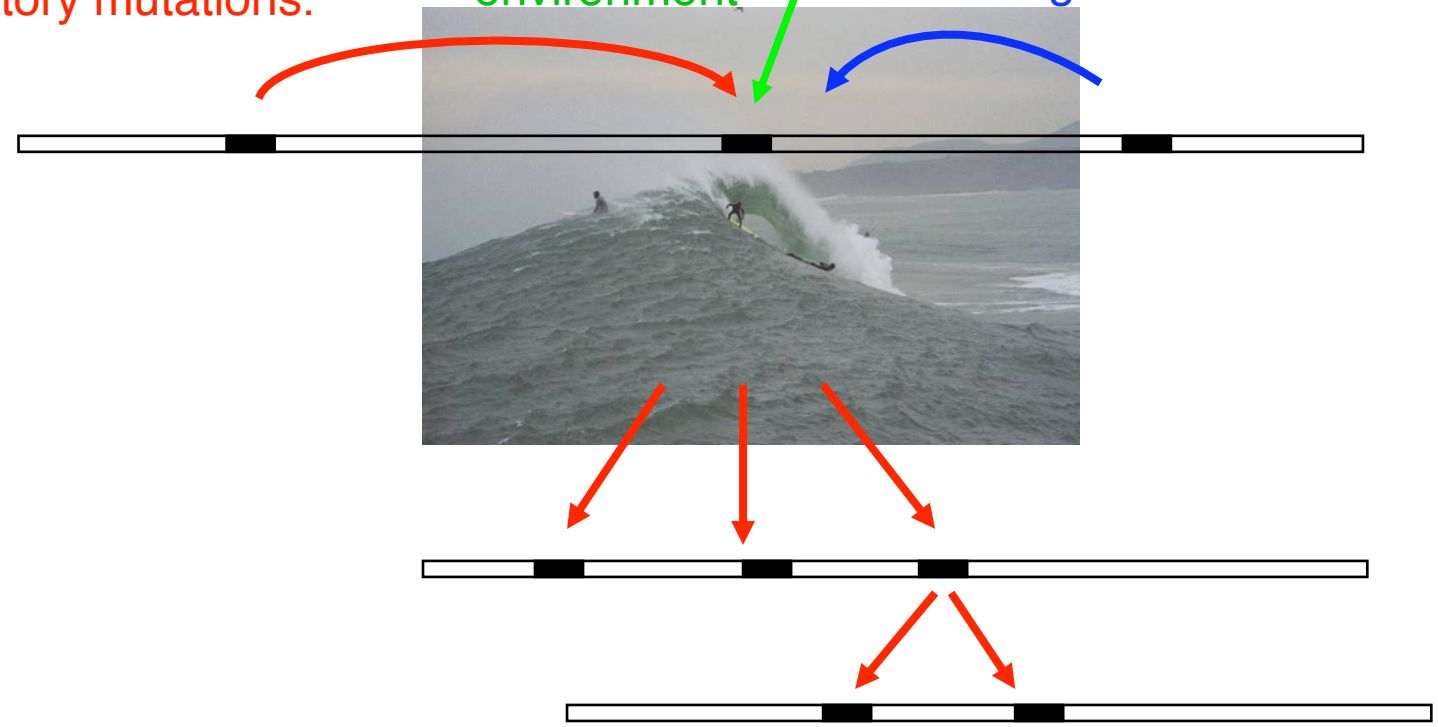
Fitness seascapes

- **What drives the waves**?

systems component:
correlations (epistasis) cause compensatory mutations.

external component:
time-dependent environment

genomic component:
linkage to other loci



- **Nonequilibrium + correlations:**
  one external change can trigger an avalanche of responses.

Conclusions

- Adaptive evolution should be viewed as a **nonequilibrium phenomenon.**
- Adaptation can be quantified by the **fitness flux** in a population over a given time interval.
- The biophysical **binding energy** is a **quantitative molecular phenotype** for regulatory sequences in yeast.
- Genomic **sequence analysis** can be used to infer **fitness landscapes** for this phenotype.
- In *Drosophila*, **fitness seascapes** drive **adaptive evolution**.
- Review articles:
  From Biophysics to evolutionary genetics, M.L., BMC Bioinformatics 2007
  From fitness landscapes to seascapes: The dynamics of selection and adaptation, V. Mustonen and M.L., Trends in Genetics 2009