Lecture 2: More complex evolutionary scenarios

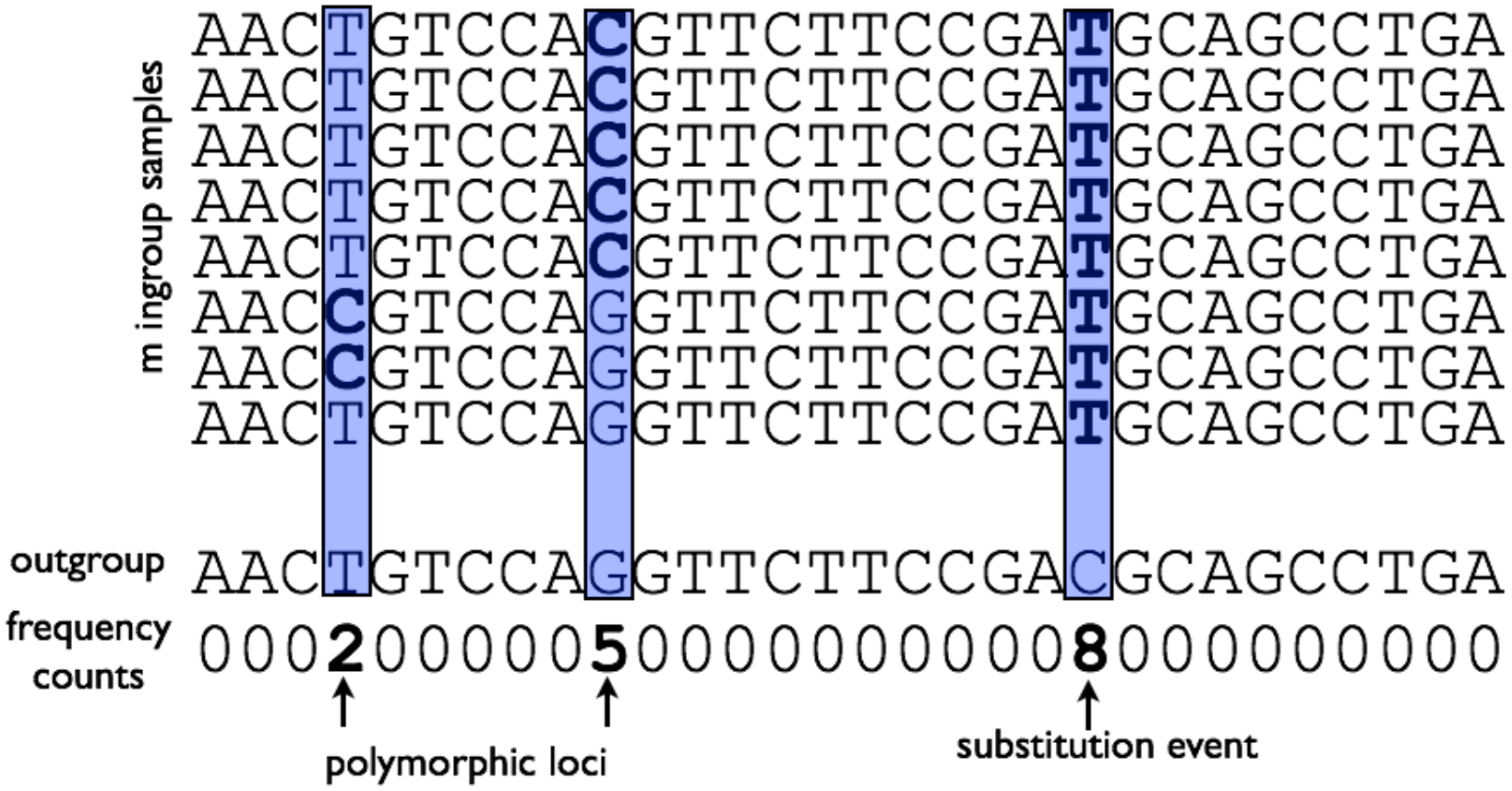
Phenotypic variation

Cichlid fish

Photo: Ryan Bloomquist

Genetic variation

Cross- and intra-species comparison



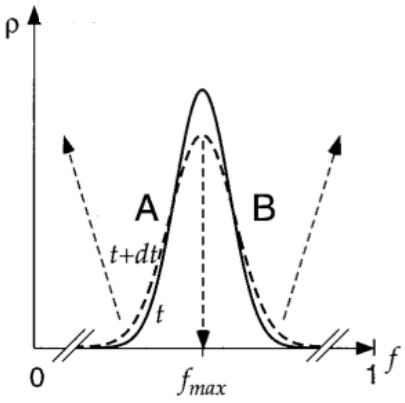
Drift, selection, and mutation

Kimura's diffusion equation

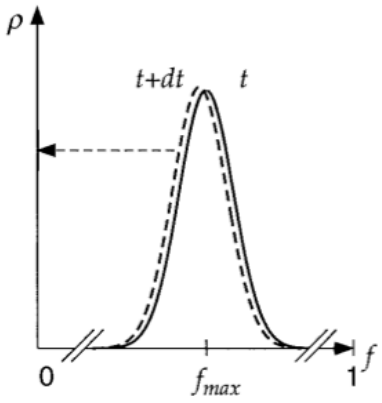
$p(x,t)$ Probability that allele frequency is x at time t

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\frac{x(1-x)}{N}p(x,t) - \frac{\partial}{\partial x}[\sigma x(1-x) + \mu(1-2x)]p(x,t)$$

Drift



Selection



Mutation

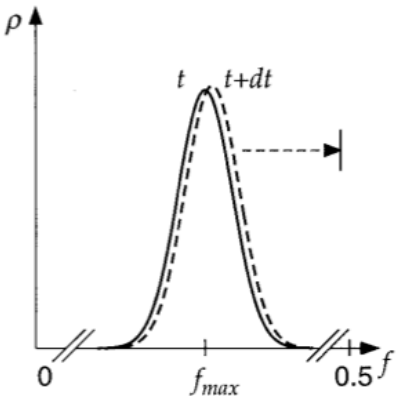


Image: Rouzine et al, 2001

Wright-Fisher process

Repeated binomial sampling

Allele frequency

$$q_a = N_a/N$$
$$q_A = N_A/N$$

Propagation

$$P(m, N, t+1) = \binom{N}{m} q_A(t)^m (1 - q_A(t))^{N-m}$$

Mean and variance

$$\langle q_A(t+1) \rangle = q_A(t)$$
$$\langle q_A(t+1)^2 \rangle - \langle q_A(t+1) \rangle^2 = \frac{q_A(t)(1 - q_A(t))}{N}$$

Activity

Build your own Wright-Fisher simulation

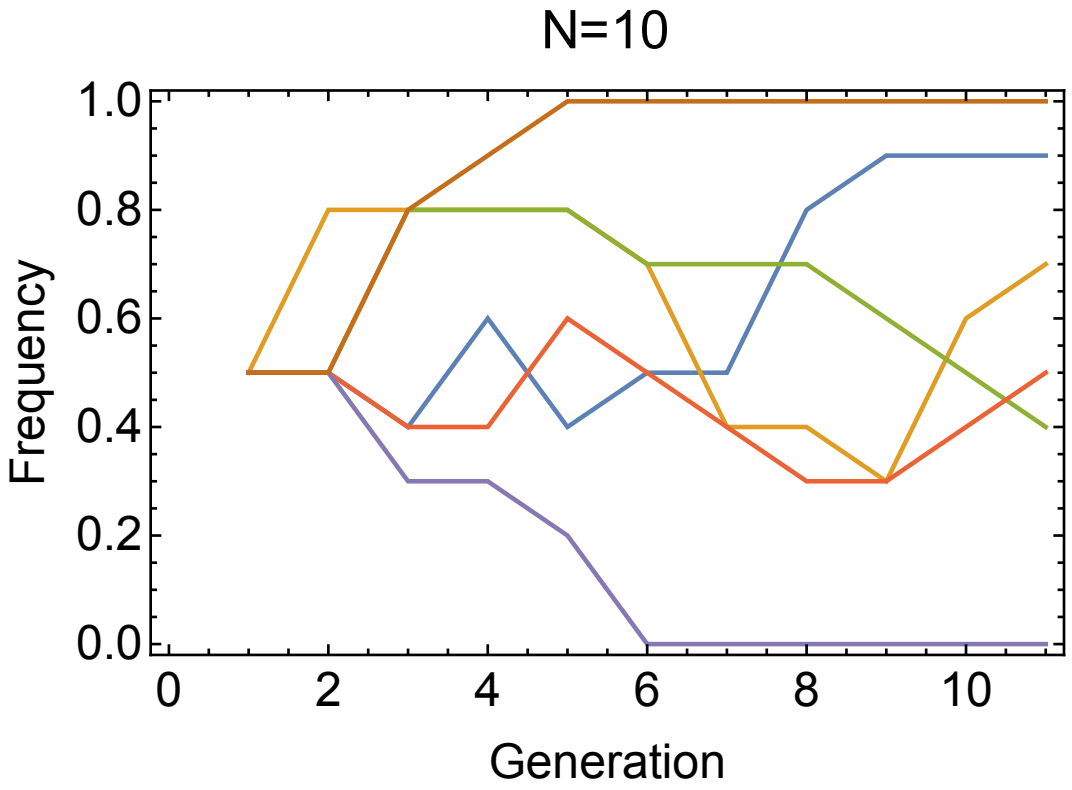
Consider a two-allele, one-locus, haploid model, with population size N

Examine how the population size affects the extent of genetic drift in the population

Hint: Each generation can be modelled as a binomial sample from the previous generation

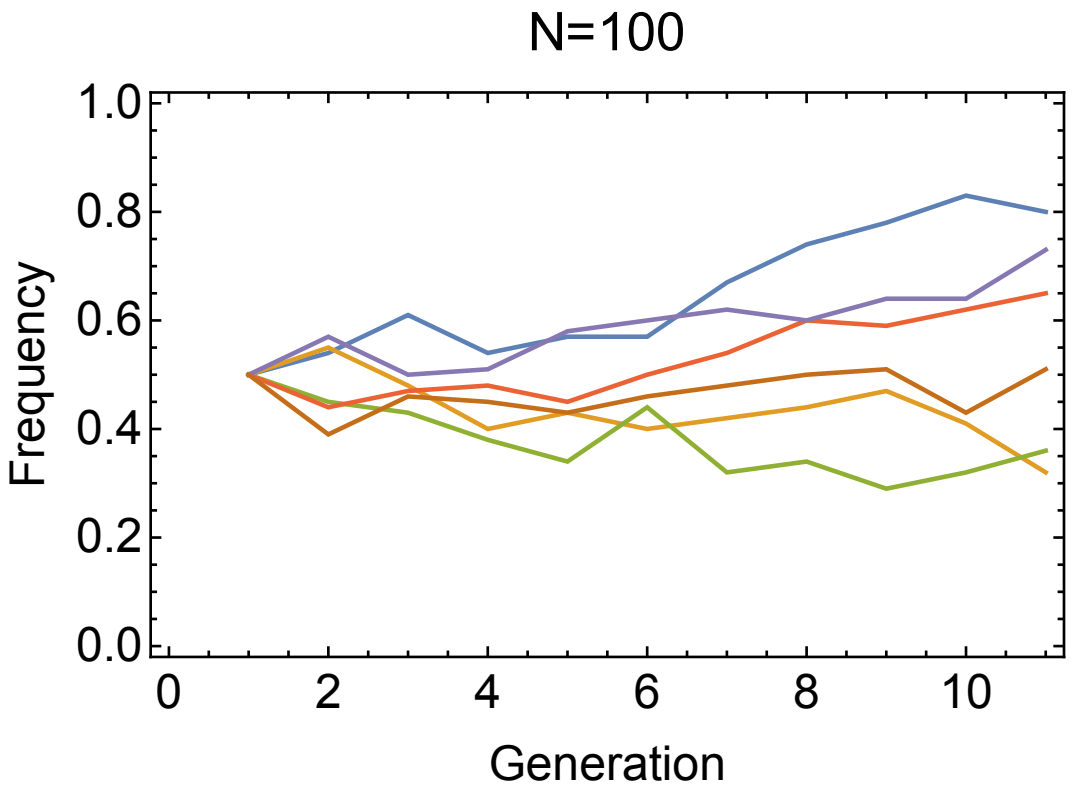
Wright-Fisher process

Rate of drift depends upon population size



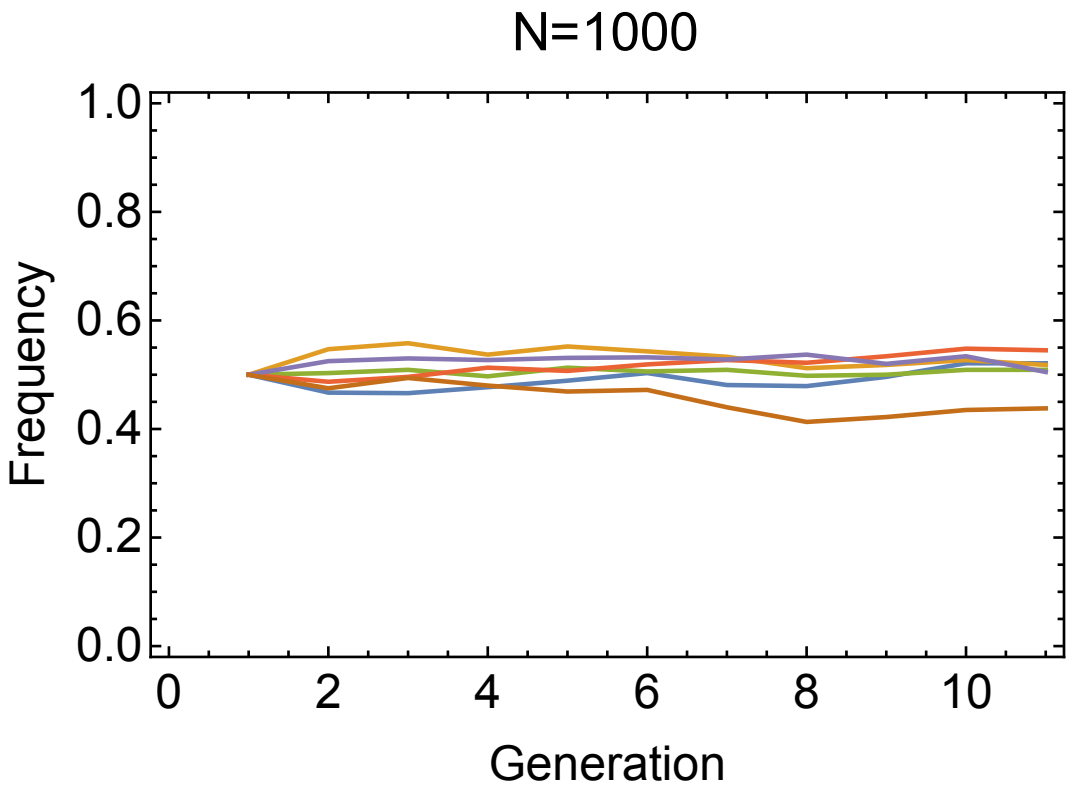
Wright-Fisher process

Rate of drift depends upon population size



Wright-Fisher process

Rate of drift depends upon population size



Rate of frequency change of the order 1/N.

Fitness and diploid genomes

Two copies of each allele: one from each parent

Genotypes are AA, Aa, and aa

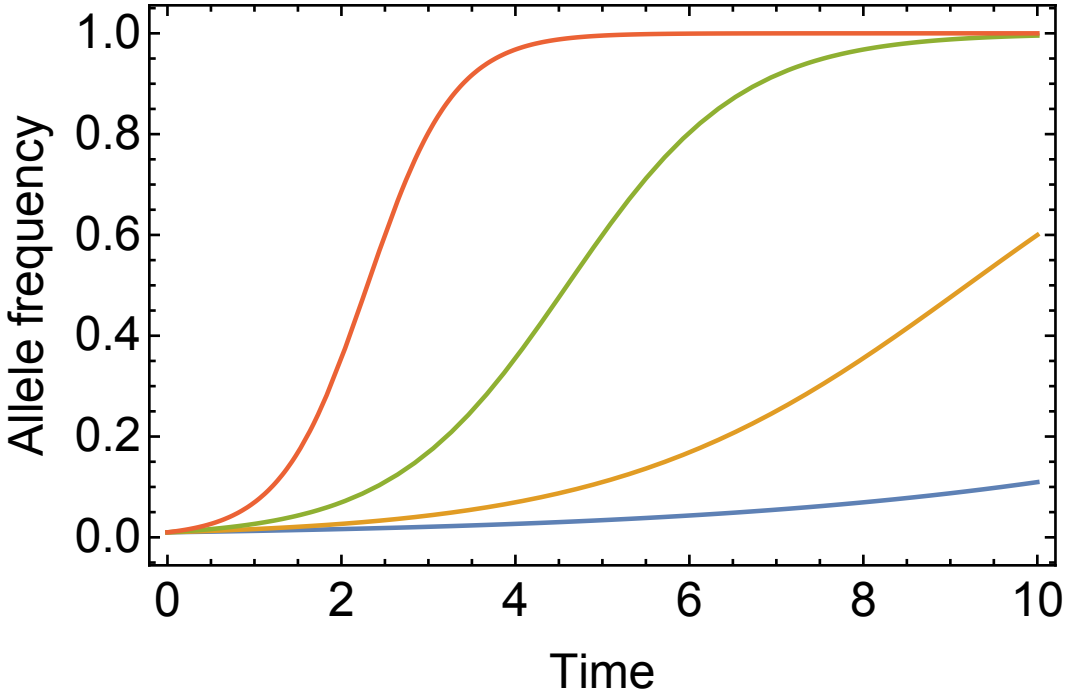
Dominance parameter h gives fitness of heterozygote

Type	Fitness
AA	$1+2\sigma$
Aa	$1+2h\sigma$
aa	1

Selection

Equation of allele frequency change under selection

$$\dot{q}_A = \sigma q_A(1 - q_A) \qquad q_A(t) = \frac{q_A(0)e^{\sigma t}}{1 - q_A(0) + q_A(0)e^{\sigma t}}$$



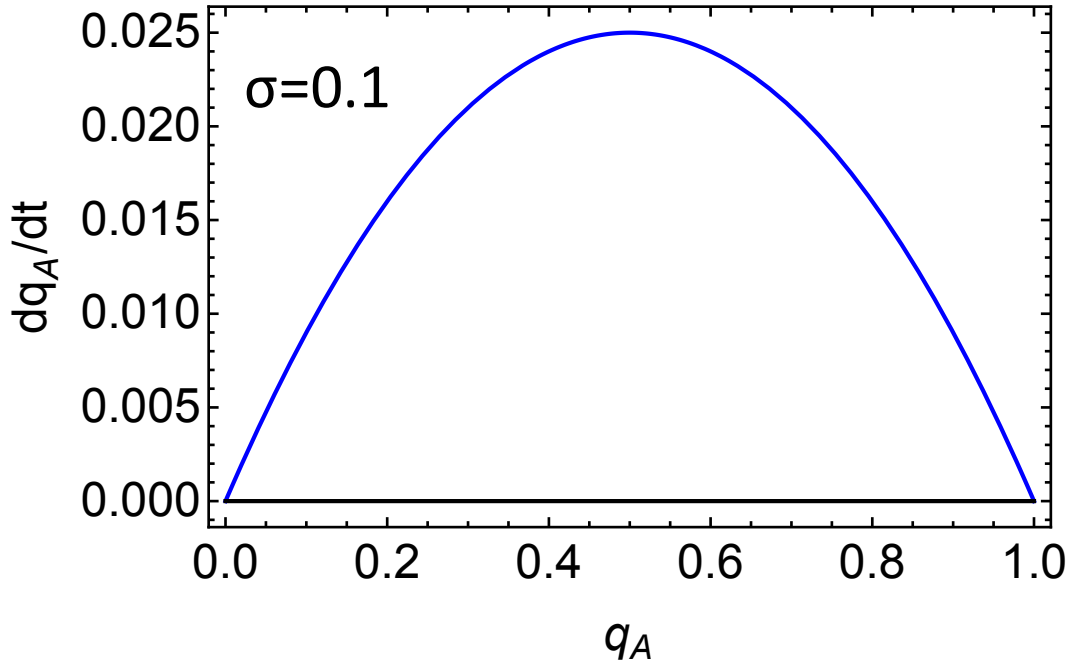
Rate of change proportional to σ

Diploid genomes

Change in allele frequency under selection

General equation $$\frac{dq_A}{dt} = 2\sigma q_A(1-q_A)(q_A + h(1-2q_A))$$

Case h=½: Additive selection



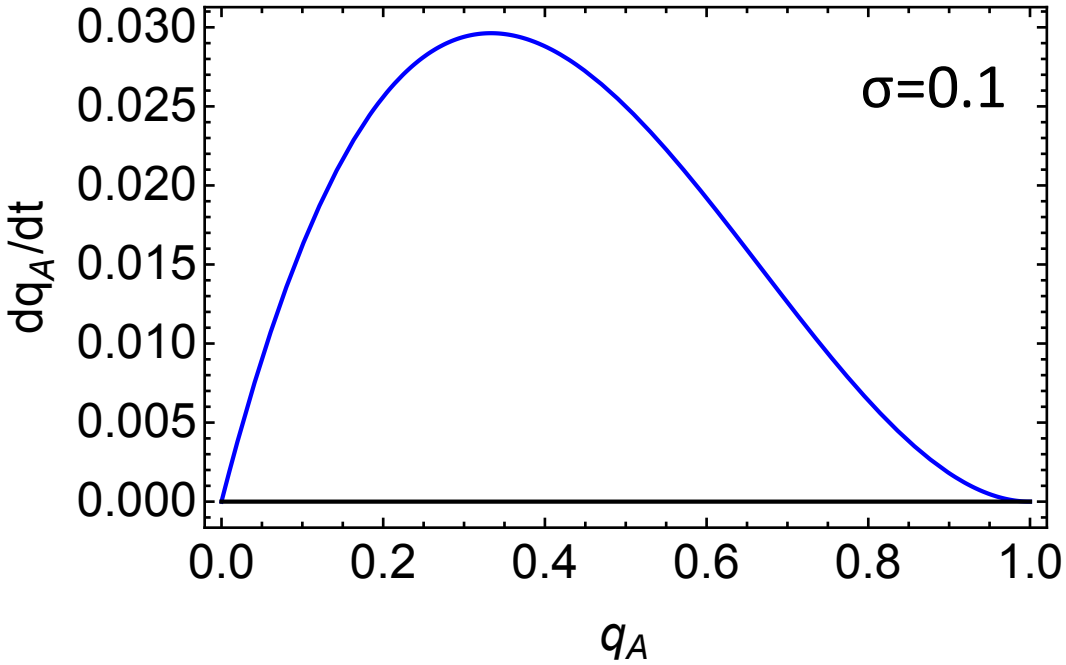
Type	Fitness
AA	1+2σ
Aa	1+2hσ
aa	1

Diploid genomes

Change in allele frequency under selection

General equation $$\frac{dq_A}{dt} = 2\sigma q_A(1-q_A)(q_A + h(1-2q_A))$$

Case h=1: A is dominant, a is recessive



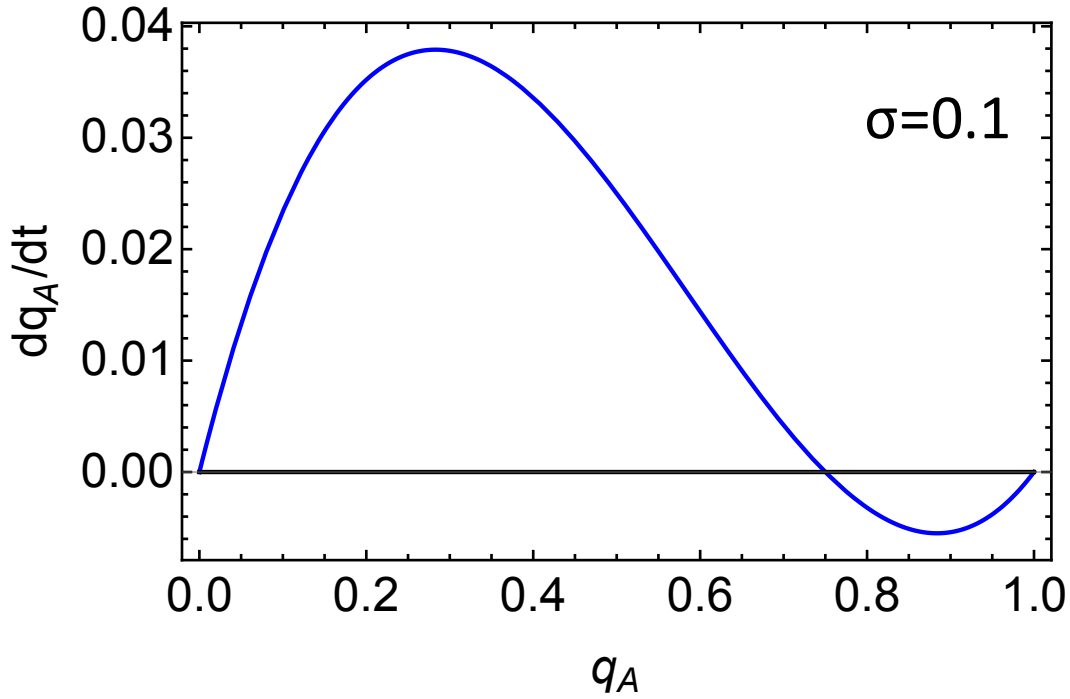
Type	Fitness
AA	1+2σ
Aa	1+2hσ
aa	1

Diploid genomes

Change in allele frequency under selection

General equation $$\frac{dq_A}{dt} = 2\sigma q_A(1-q_A)(q_A + h(1-2q_A))$$

Case h>1: Overdominance, heterozygote advantage



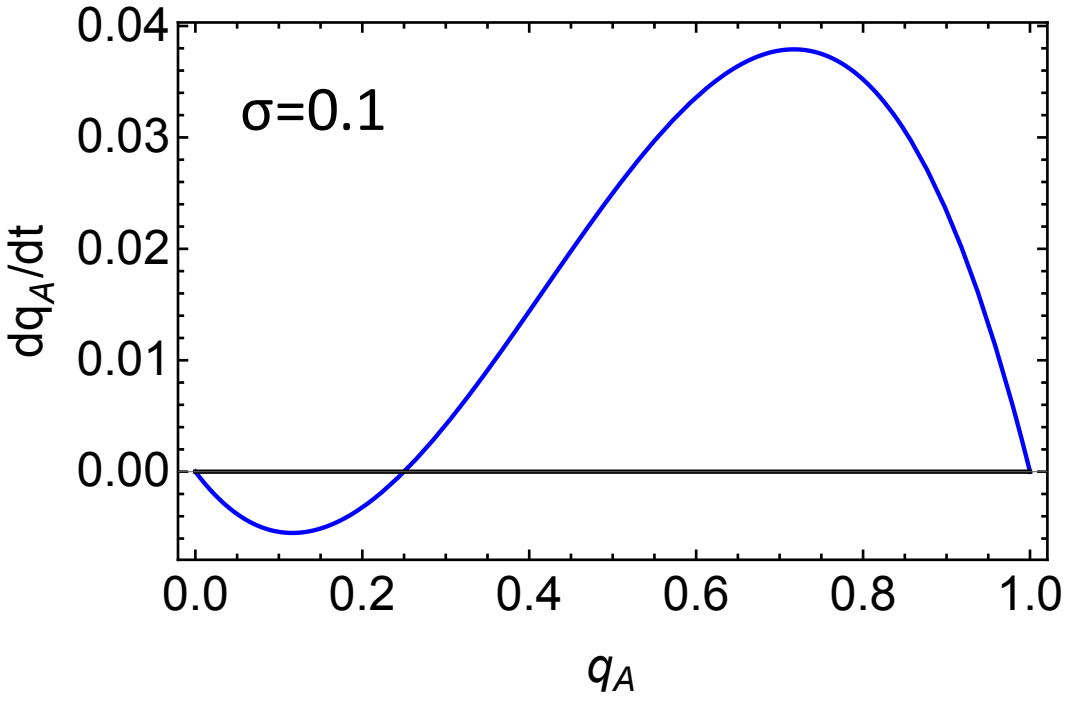
Type	Fitness
AA	1+2σ
Aa	1+2hσ
aa	1

Diploid genomes

Change in allele frequency under selection

General equation $$\frac{dq_A}{dt} = 2\sigma q_A(1-q_A)(q_A + h(1-2q_A))$$

Case h<0: Underdominance, homozygote advantage



Type	Fitness
AA	1+2σ
Aa	1+2hσ
aa	1

Frequency-dependent selection

Hawk-dove game

Resource of value V Cost of losing a fight C

Payoff matrix:

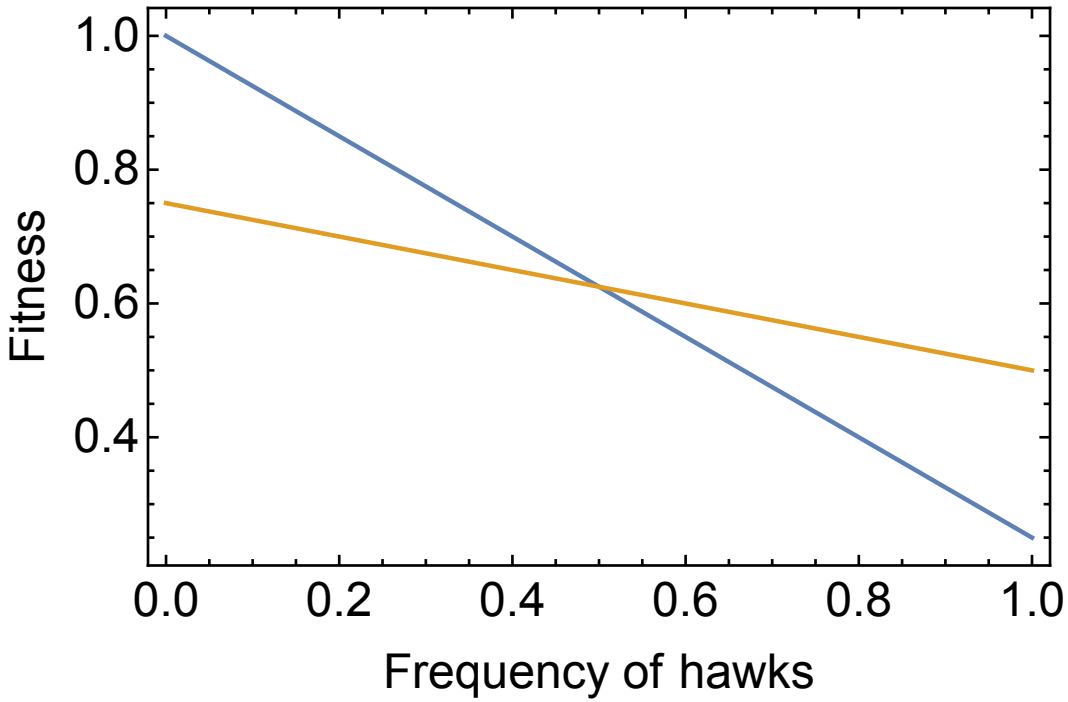
	Hawk	Dove
Hawk	V/2 – C/2	V
Dove	0	V/2

Fitness given by base fitness B plus mean outcome of game

Frequency dependent selection

Hawk-dove game

V=0.5; C=1; B=0.5: Fitness of Doves Hawks

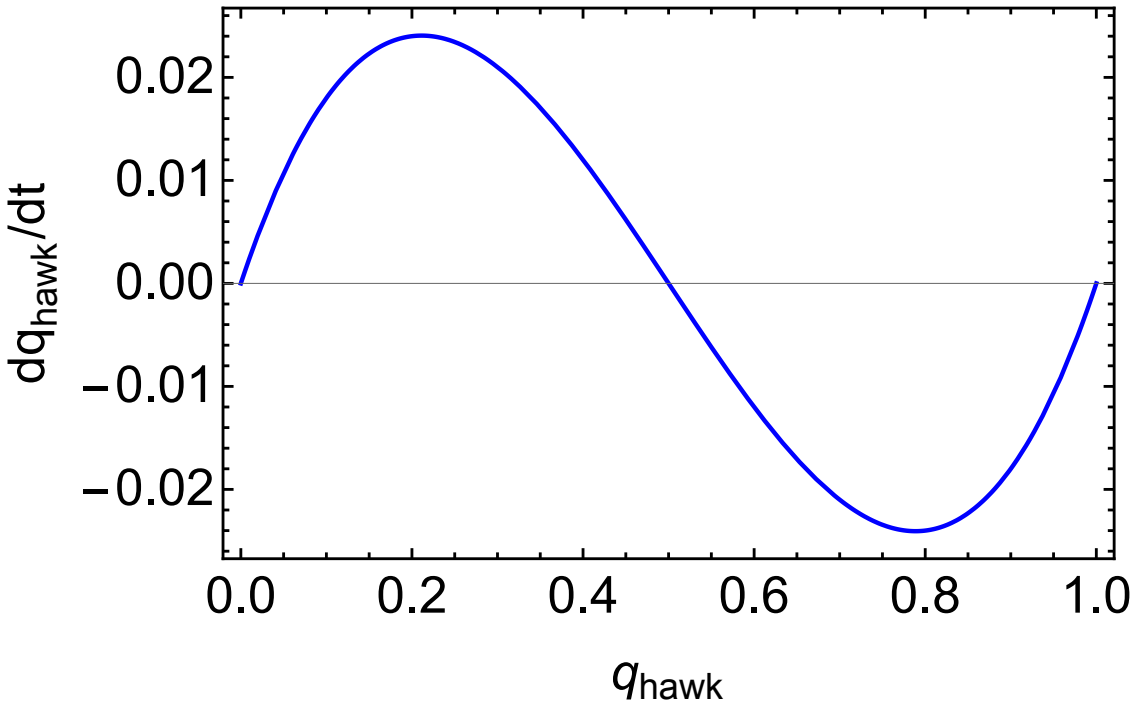


System vulnerable to invasion by each type

Frequency dependent selection

Hawk-dove game

V=0.5; C=1; B=0.5



Evolutionary game theory

Improved strategy: Assess the situation

Act as a hawk if you can win the fight, or as a dove otherwise

Desert spiders: Web competition

Frequency-dependent selection

Prisoner's dilemma

Cooperation versus selfish behaviour

	Cooperate	Defect
Cooperate	1	3
Defect	0	2

Optimal strategy in iterated game…

Effective population size

Effective population size defined by behaviour under drift

Variance of Wright-Fisher model:

$$\langle q_A(t+1)^2 \rangle - \langle q_A(t+1) \rangle^2 = \frac{q_A(t)(1-q_A(t))}{N}$$

Suppose a population evolves according to a different variance, v.

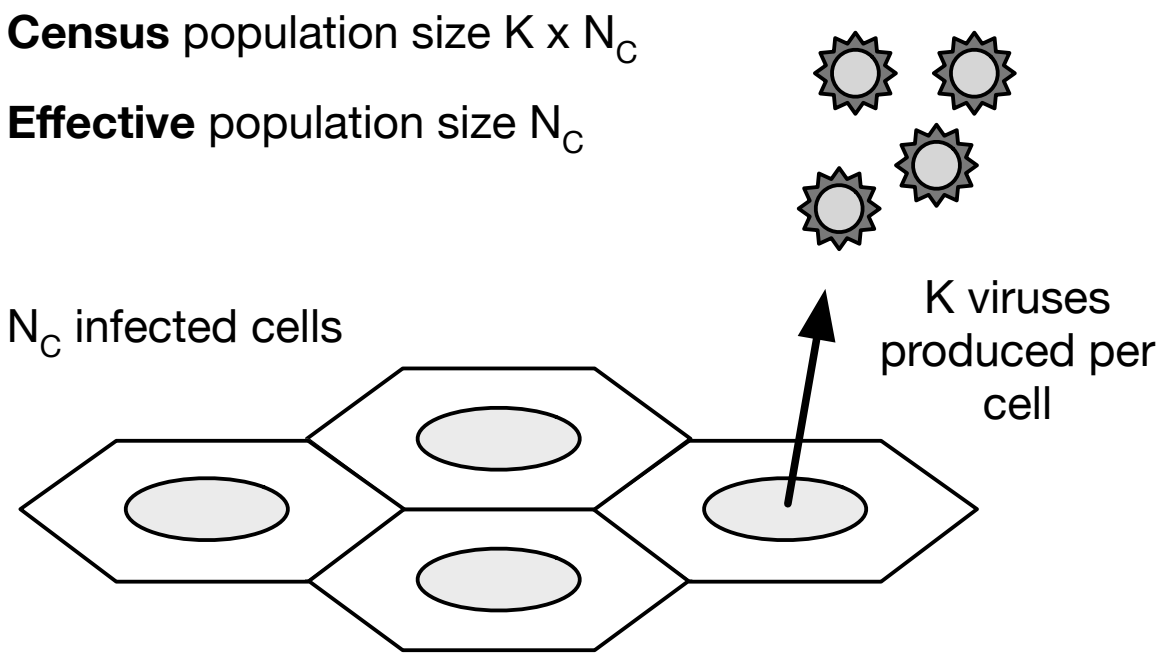
Can define the effective population size by

$$N_e = \frac{q_A(t)(1-q_A(t))}{v}$$

Effective population size

Effective population size defined by behaviour under drift

Retroviral population:



Effective population size

Effective population size defined by behaviour under drift

Variable population size:

If the real population size at time t is given by N_t

$$\frac{1}{N_e} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}$$

Different sex ratios:

$$N_e = \frac{4N_mN_f}{N_m + N_f}$$

Inference of evolutionary parameters from genomic data

Genomic Data: A very rough guide

Technology	Read Length	Accuracy
Sanger sequencing	c. 1kb	Very high
Illumina MiSeq / HiSeq etc.	150-300 bases (paired end)	High
Pacific Biosciences	c. 2kb	Moderate
Oxford Nanopore	>100kb	Poor but improving

Fitting a model to data

Probability

Probability of an outcome given an underlying model

e.g. Probability of flipping 5 heads and 2 tails given an unbiased coin

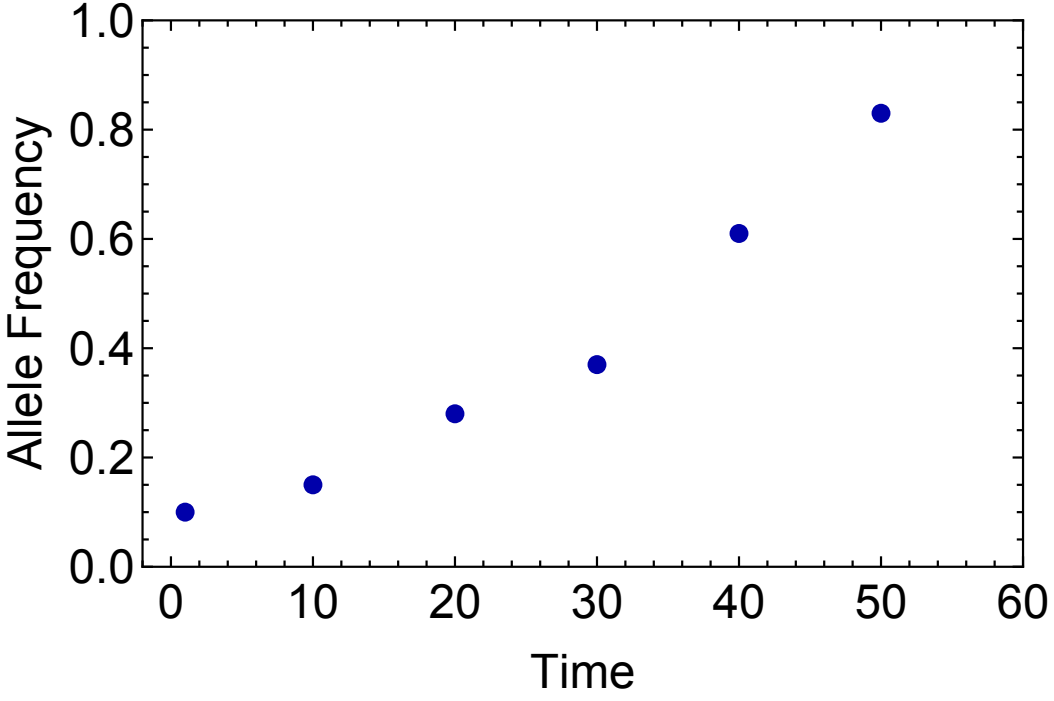
Likelihood

Likelihood of a model given an observed outcome

e.g. Likelihood that the coin is unbiased given I just flipped 5 heads and 2 tails

Example I: Inference of selection

Data from an evolutionary experiment



$n_A(t_k), N(t_k)$

Observations of allele A

Total observations

Assumptions: Large population size (neglect genetic drift)
Low rate of mutation (neglect mutation)
Constant strength of selection

Example I: Inference of selection

Data from an evolutionary experiment

Assumptions: Large population size (neglect genetic drift)
Low rate of mutation (neglect mutation)
Constant strength of selection

Dynamic model:

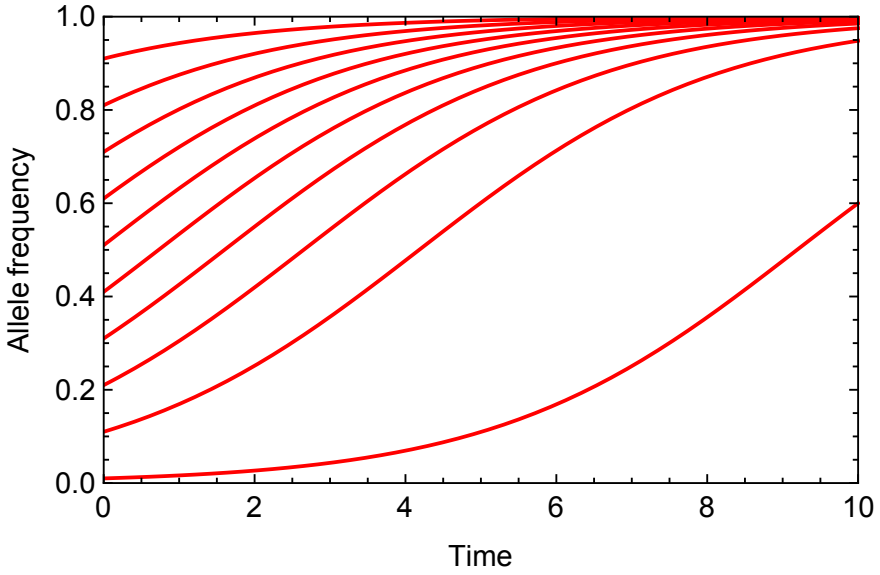
$$\mathcal{M} : q_A(t) = \frac{q_A(0)e^{\sigma t}}{1 - q_A(0) + q_A(0)e^{\sigma t}}$$

Frequency of allele A at time t

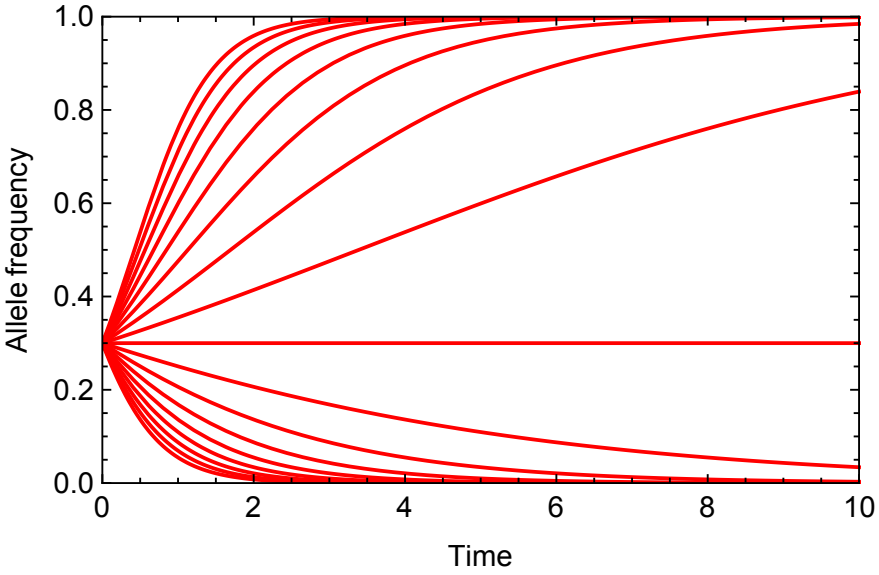
Parameters are: $q_A(0)$, σ

Example I: Inference of selection

Model trajectories



Constant σ
Varying $q_A(0)$



Constant $q_A(0)$
Varying σ

Example I: Inference of selection

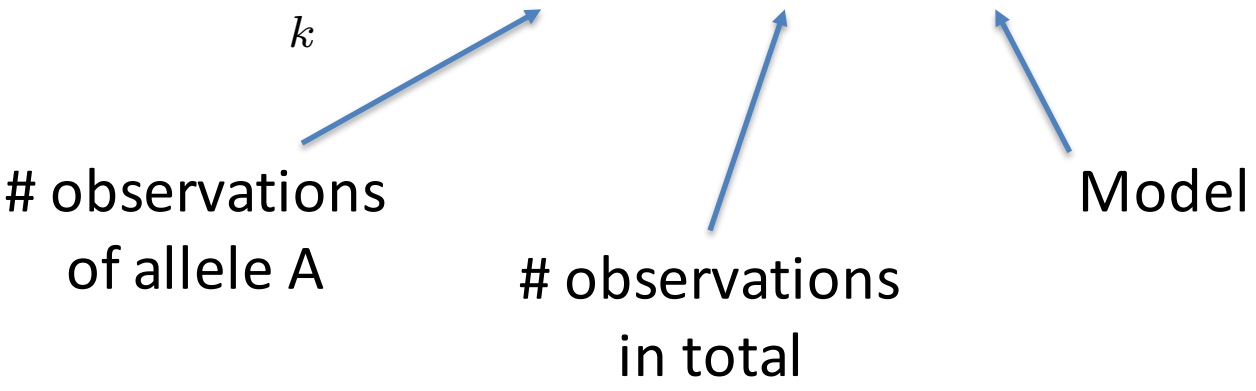
Fitting model to data

Dynamic model:

$$\mathcal{M} : q_A(t) = \frac{q_A(0)e^{\sigma t}}{1 - q_A(0) + q_A(0)e^{\sigma t}}$$

Log likelihood:

$$\mathcal{L} = \sum_k \log P(n_A(t_k)|N(t_k), \mathcal{M})$$



Example I: Inference of selection

Fitting model to data

Dynamic model:

$$\mathcal{M} : q_A(t) = \frac{q_A(0)e^{\sigma t}}{1 - q_A(0) + q_A(0)e^{\sigma t}}$$

Log likelihood:

$$\mathcal{L} = \sum_k \log P(n_A(t_k)|N(t_k), \mathcal{M})$$

Binomial model

$$n_a(t_k) = N(t_k) - n_A(t_k)$$

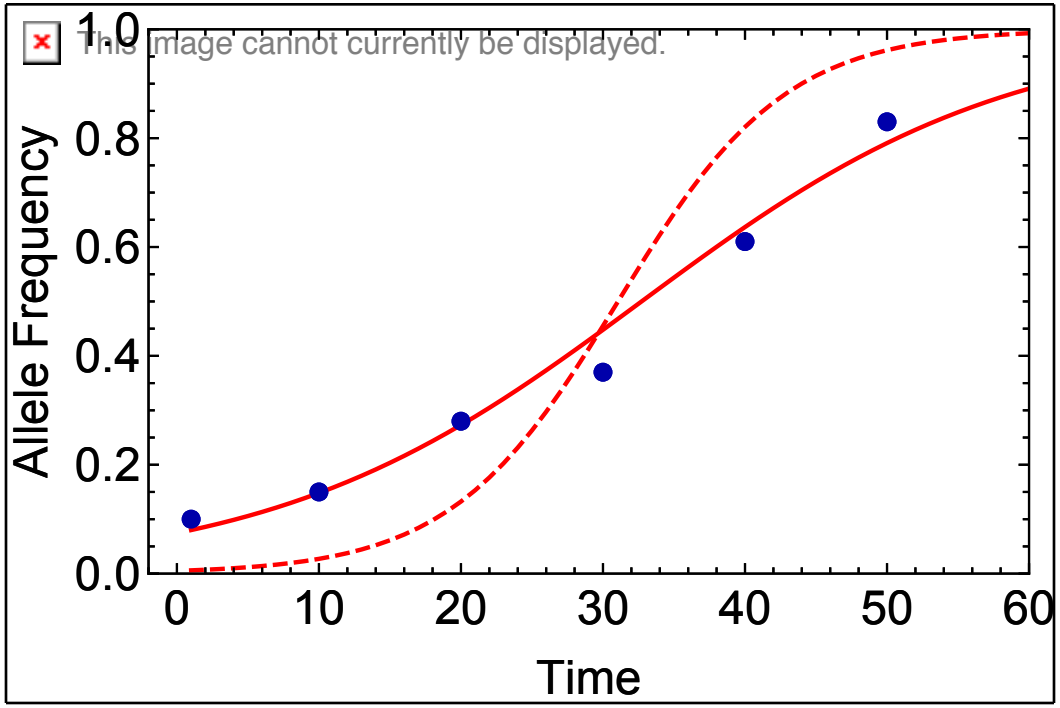
$$q_a(t) = 1 - q_A(t)$$

$$P(n_A(t_k)|N(t_k), \mathcal{M}) = \frac{N!}{n_A(t_k)!n_a(t_k)!} q_A(t_k)^{n_A(t_k)} q_a(t_k)^{n_a(t_k)}$$

Example I: Inference of selection

Learning a selection coefficient

Model of growth under constant selection



Parameters are:

$$q_A(0)$$

$$\sigma$$

Find the model parameters which best explain the observations (maximum likelihood model)

Exercise: Optimisation

Given a function, find the maximum (or minimum) value

Find the minimum value of:

$$y = x^4 - 5x^3 - x^2 + 15x + 1$$

Basic Strategy:

Guess an initial value of x

Calculate y

Change x. Keep the new x if this reduces y.

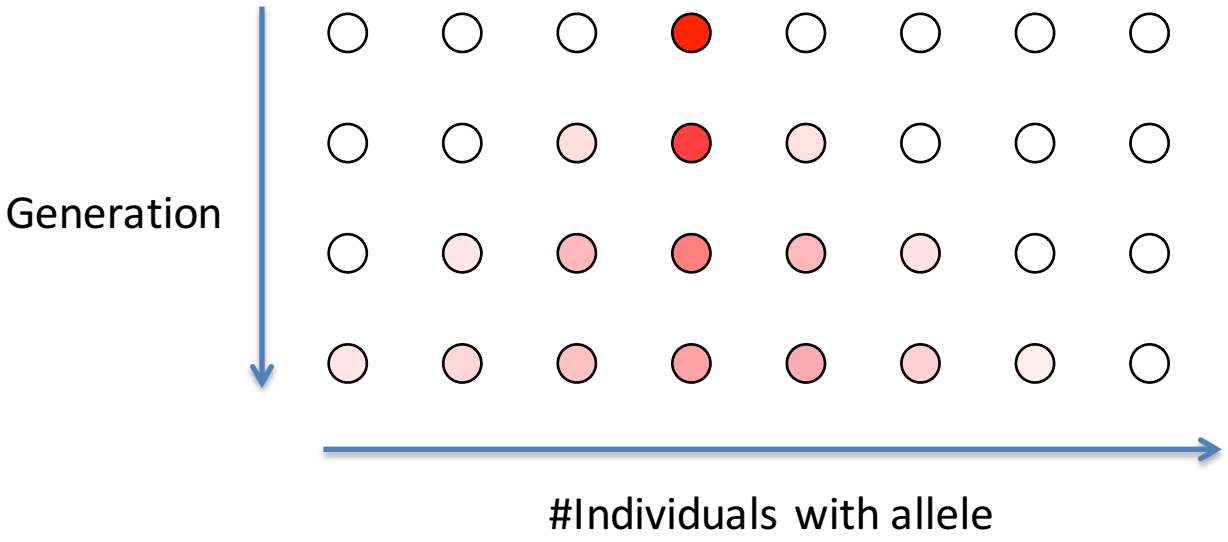
Inference of selection given genetic drift

Smaller population size: have to account for genetic drift

Case 1: Know the value of the population size N

Wright-Fisher model:

Repeated binomial sampling



Fitting a distribution to data

Model is probabilistic

Given the likelihood $L(n_A(t)|q_A(t))$

Calculate: $\int_{q=0}^{q=1} P(q_A = q)L(n_A(t)|q)dq$

Disadvantages of approach:

Need to know population size N and generation time

Have to calculate entire distribution

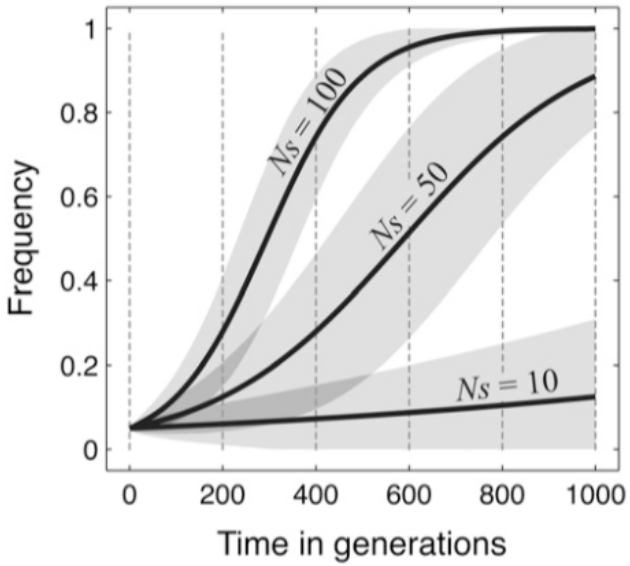
Inference of selection given genetic drift

Case 2: Don't know N or generation time

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\frac{x(1-x)}{N}p(x,t) - \frac{\partial}{\partial x}[\sigma x(1-x) + \mu(1-2x)]p(x,t)$$

Diffusion: Can be approximated by a Gaussian:

Propagate mean and variance

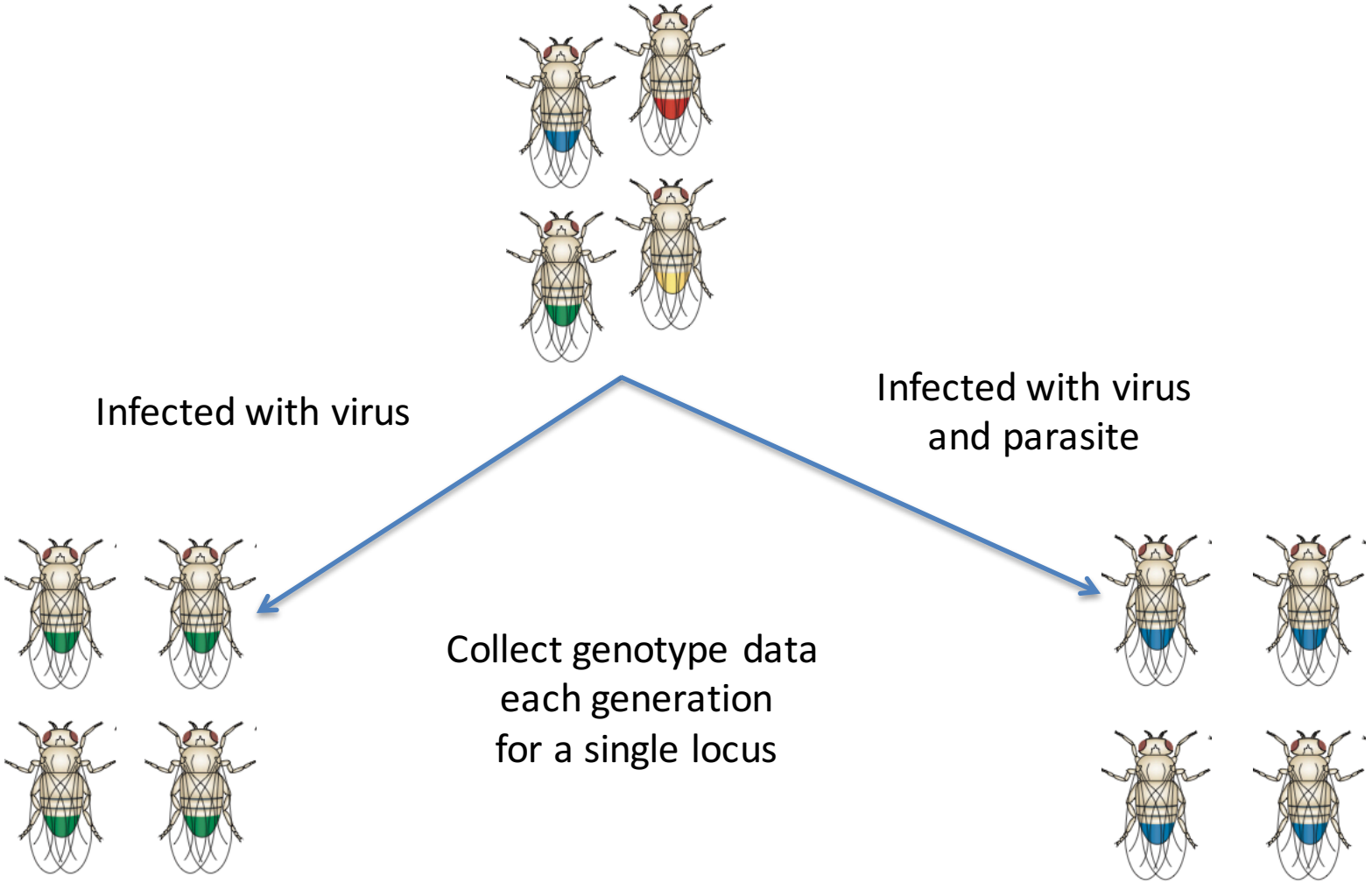


Good approximation at intermediate frequencies

Feder et al., Genetics 2014

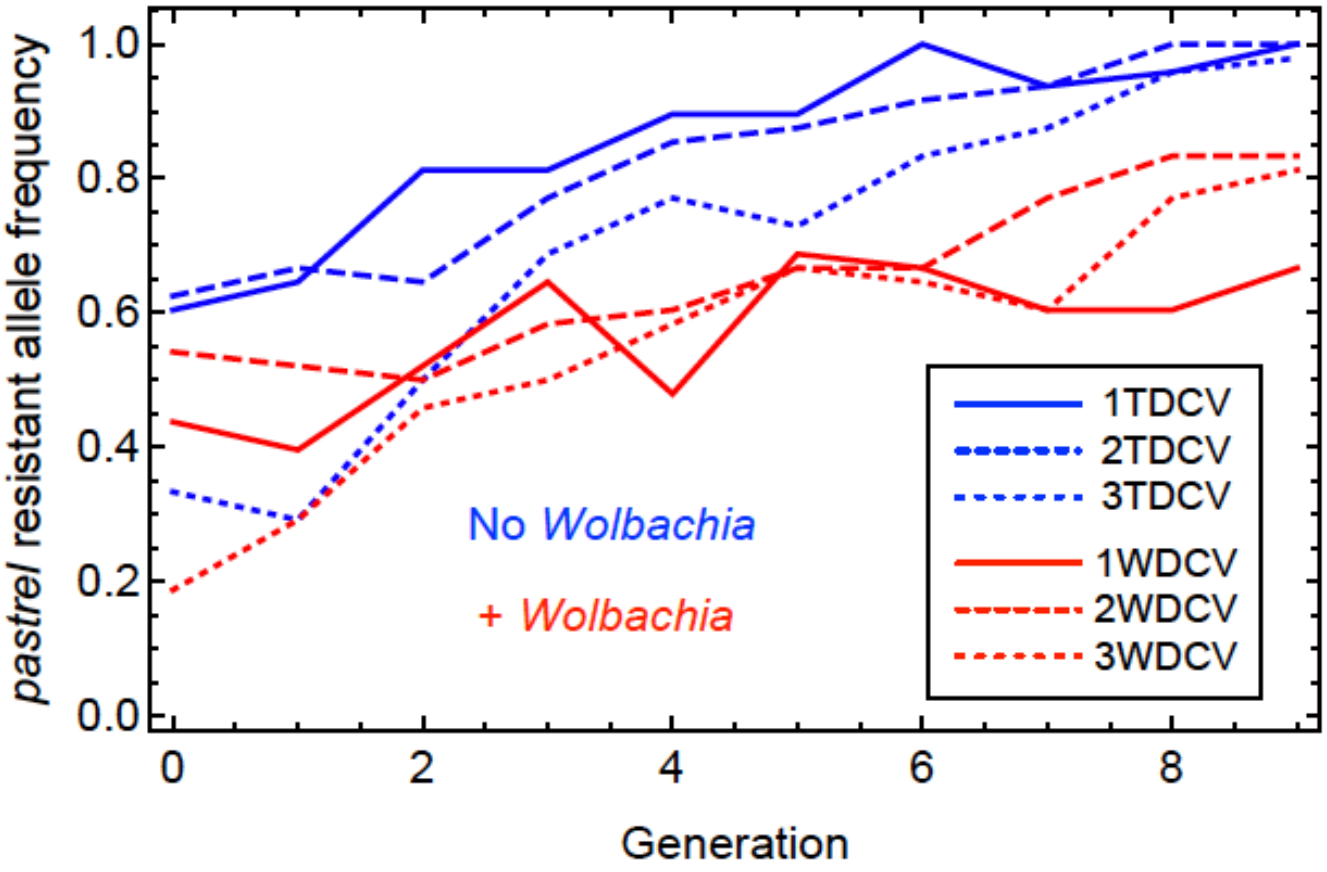
Example II: Selection and drift in a diploid system

Data from evolutionary experiment



Example II: Selection and drift in a diploid system

Allele frequency plots for three replica experiments (each case)



Example II: Selection and drift in a diploid system

Explicit counts of homozygotes and heterozygotes

Generation	CC	CT	TT
0	10	9	5
1	10	11	3
2	17	5	2
3	16	8	0
4	19	5	0
5	19	5	0
6	24	0	0
7	22	2	0
8	22	2	0
9	24	0	0

Example II: Selection and drift in a diploid system

Data from an evolutionary experiment

Assumptions: Large population size (neglect genetic drift)
Low rate of mutation (neglect mutation)
Constant strength of selection

Dynamic model: Must account for drift
N=300 (with parasite)
N=600 (without parasite)

Example II: Selection and drift in a diploid system

Gaussian model propagation

Mean: $\mu_{t+1} = \frac{(1+S)\mu_t}{1+S\mu_t}$

Variance:

$$\sigma^2_{t+1} = \frac{1}{N}\left[\frac{(1+S)\mu_t}{1+S\mu_t}\left(1-\frac{(1+S)\mu_t}{1+S\mu_t}\right)\right] + \left[\frac{1+S}{(1+S\mu_t)^2}\right]\sigma^2_t$$

S = advantage of T allele over C

Lacerda et al., Genetics 2014

Example II: Selection and drift in a diploid system

Express S in terms of diploid parameters

Fitness parameters:

$$w_{CC} = 1 + s$$
$$w_{CT} = 1 + hs$$
$$w_{TT} = 1$$

Frequency parameters:

$$\{q_{CC}, q_{CT}, q_{TT}\}$$

$$p = q_{CC} + \frac{1}{2} q_{CT}$$

$$S = \frac{s(h + p - 2hp)}{1 + hps}$$

Example II: Selection and drift in a diploid system

Replica experiments:

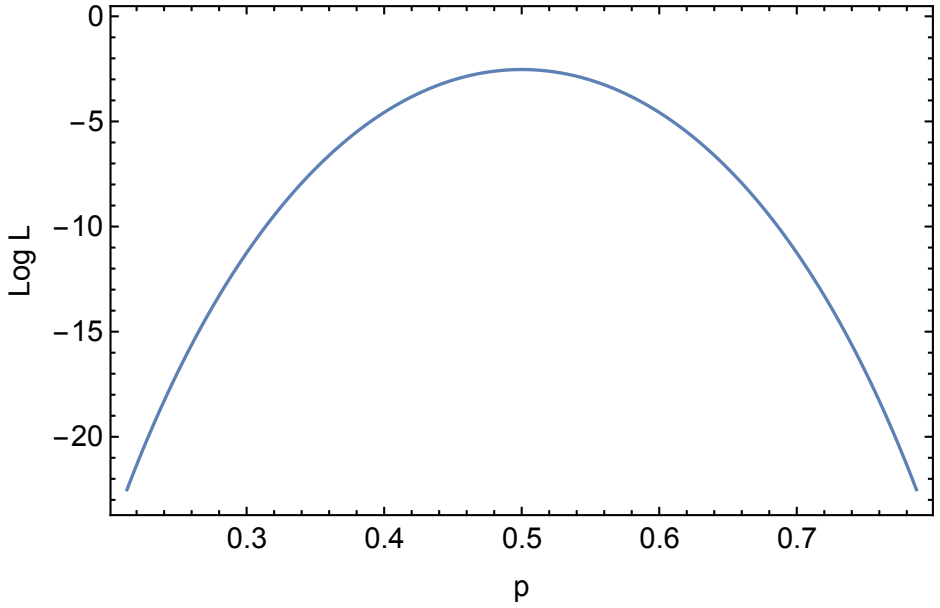
Fit parameters to all datasets combined – assume identical

Conveying uncertainty in an inference:

Deviation from maximum likelihood

Likelihood surface for
Binomial distribution

N=100
n=50



Example II: Selection and drift in a diploid system

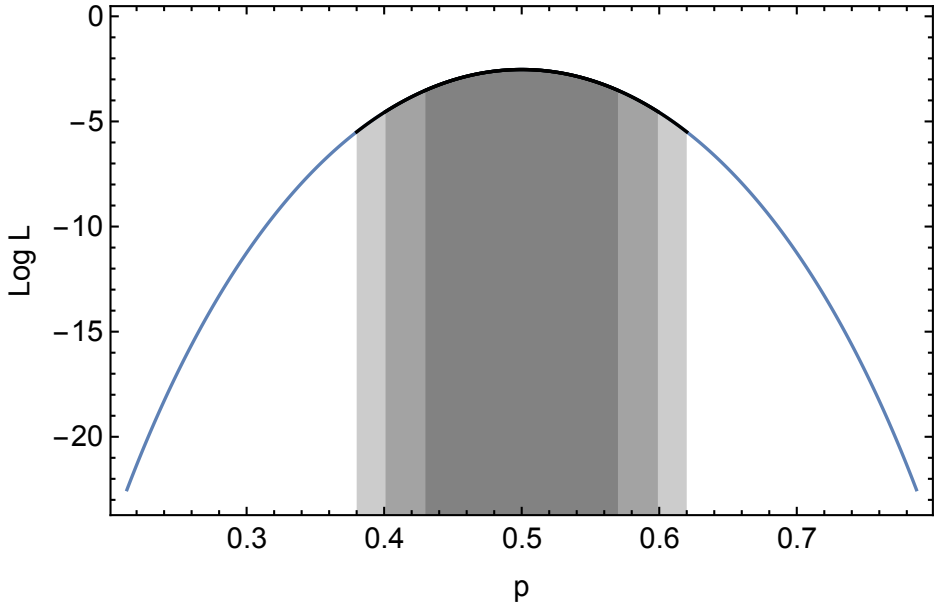
Replica experiments:

Fit parameters to all datasets combined

Conveying uncertainty in an inference:

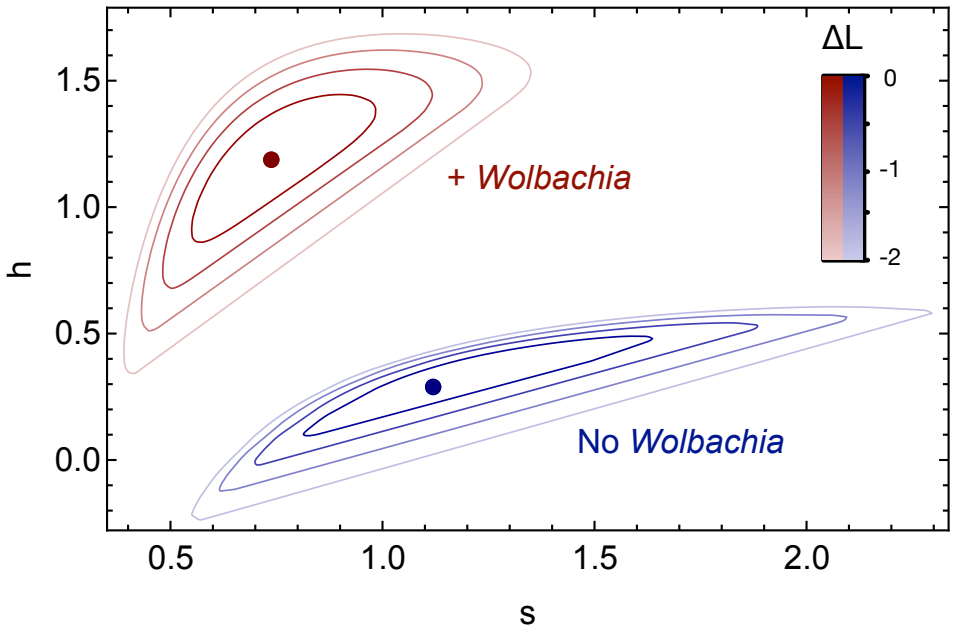
Deviation from maximum likelihood

Intervals of 1, 2, and 3 likelihood units around the maximum



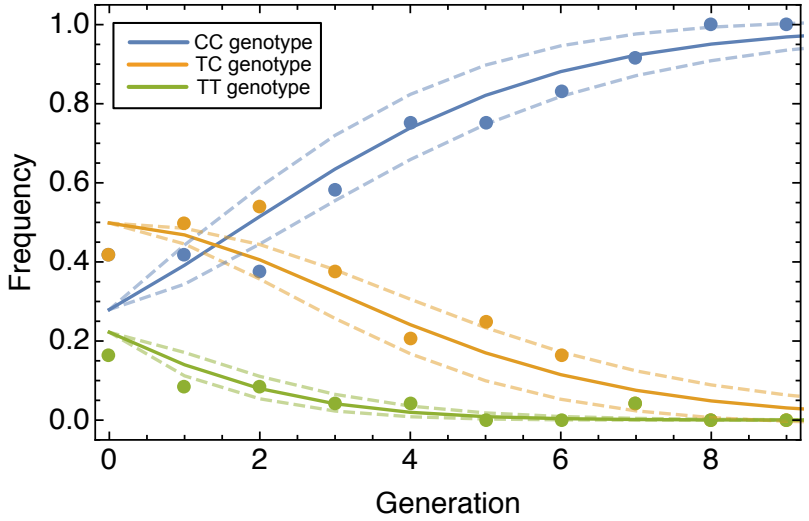
Example II: Selection and drift in a diploid system

Fit to data

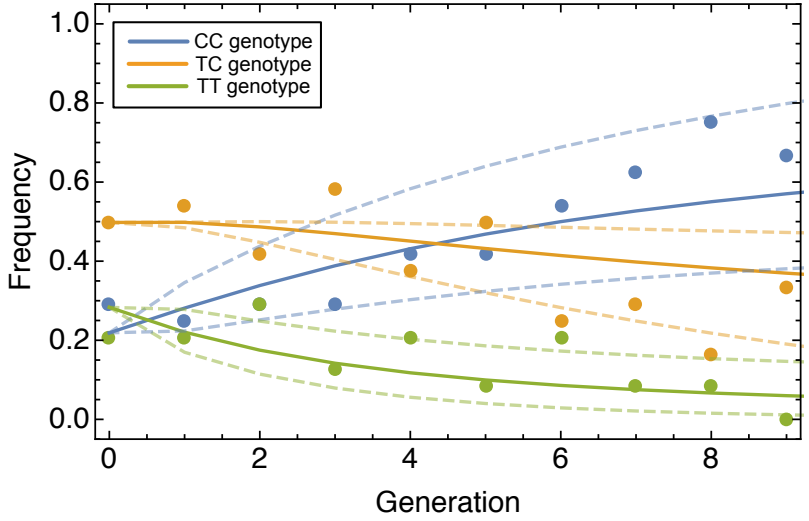


Observe likely overdominance with Wolbachia parasite

No *Wolbachia*

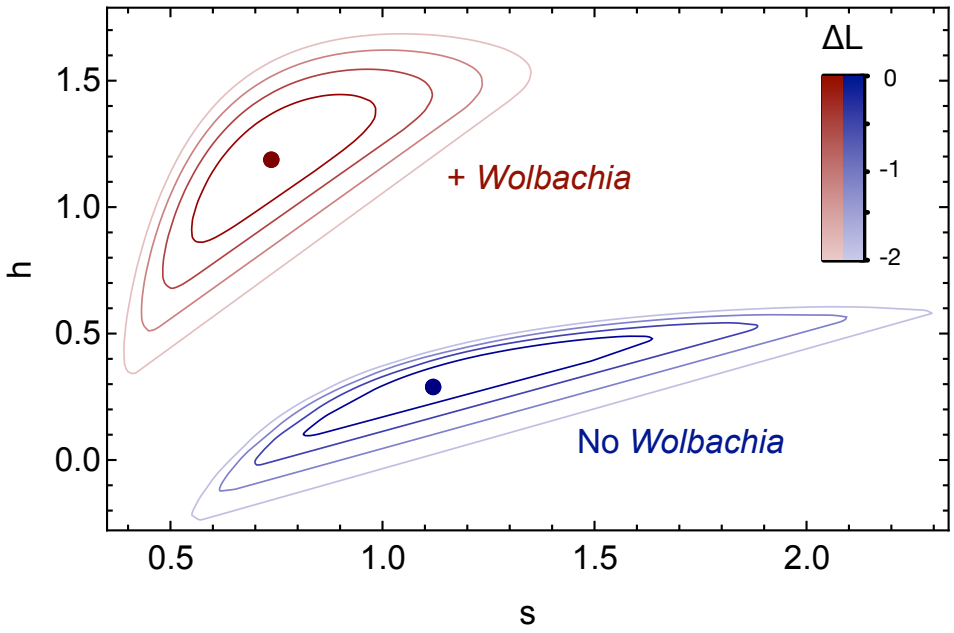


+ *Wolbachia*



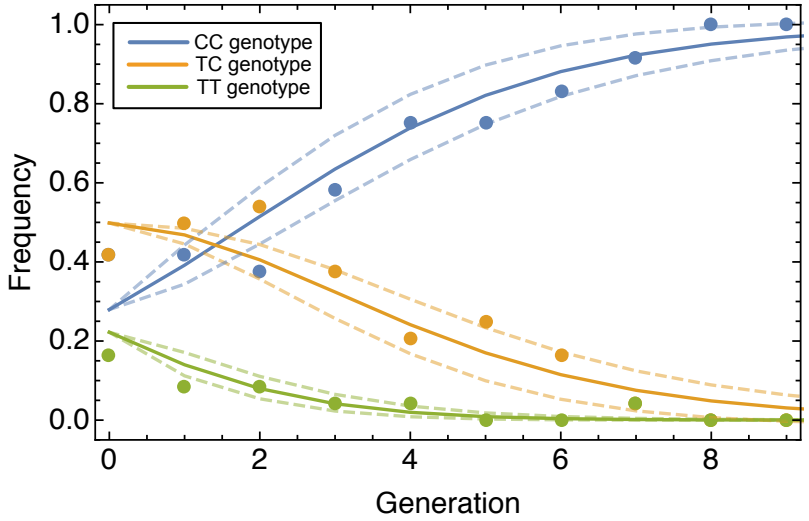
Implications: Parasite addiction

Fit to data

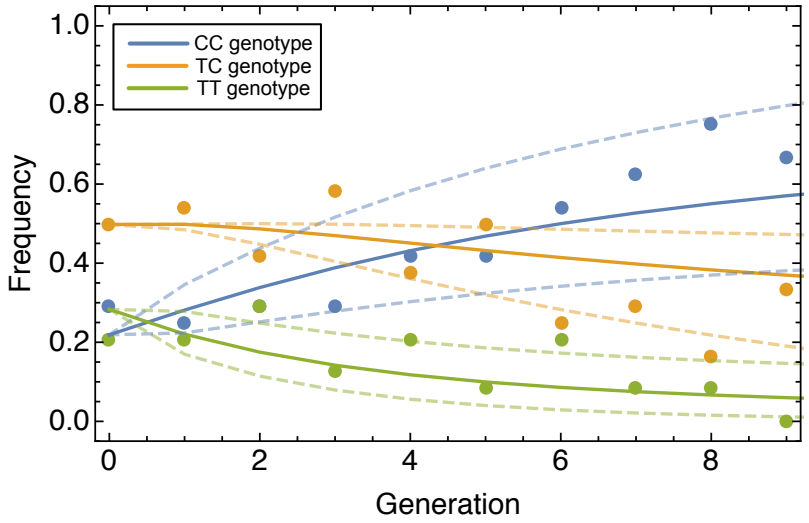


Martinez, et al, 2016
Proc Roy Soc B

No *Wolbachia*



+ Wolbachia



Multi-locus models

Two-locus, two-allele model

Minimal model for linkage, recombination, etc.

Denote loci i,j. Alleles 0,1 at each.

Sequences
00
01
10
11

Frequencies $q^{00}, q^{01}, q^{10}, q^{11}$

Fitnesses $f^{00}, f^{01}, f^{10}, f^{11}$

Frequency change under selection:

$$\frac{dq_{ij}^{00}}{dt} = f^{00}q^{00} - q^{00} \sum_{a,b \in \{0,1\}} f^{ab}q^{ab}$$

$$\vdots \qquad\qquad \vdots$$

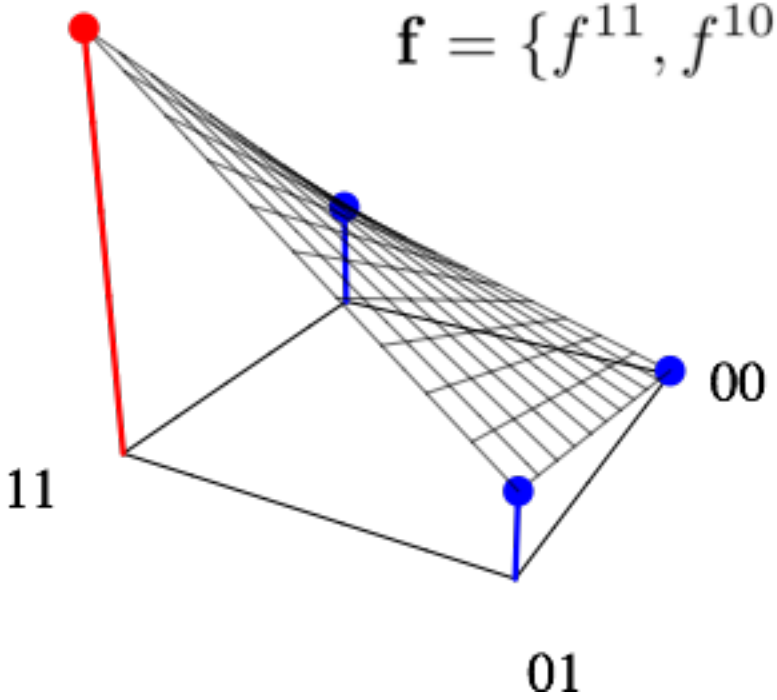
$$\frac{dq_{ij}^{11}}{dt} = f^{11}q^{11} - q^{11} \sum_{a,b \in \{0,1\}} f^{ab}q^{ab}$$

Two-locus, two-allele model

Fitness landscape

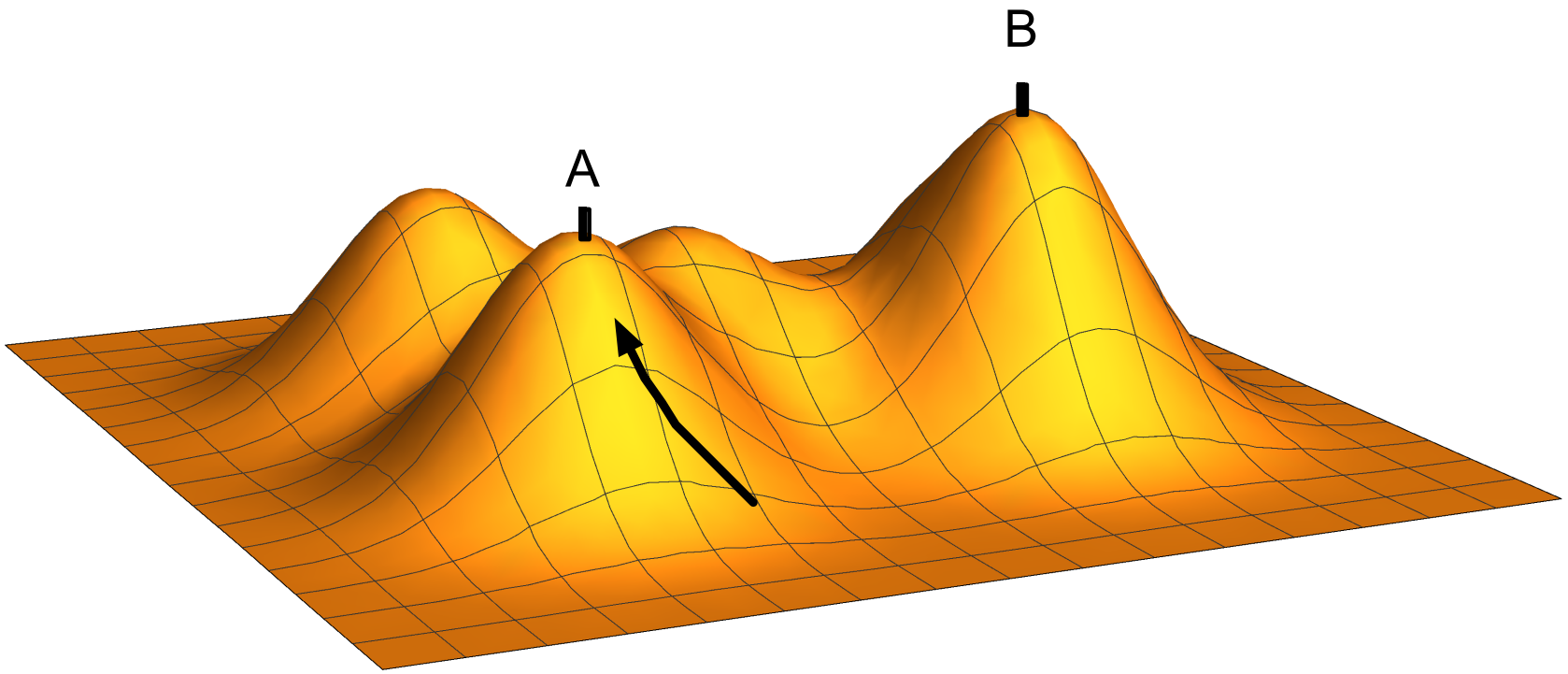
Representation of haplotype frequencies

$$\mathbf{f} = \{f^{11}, f^{10}, f^{01}, f^{00}\}.$$



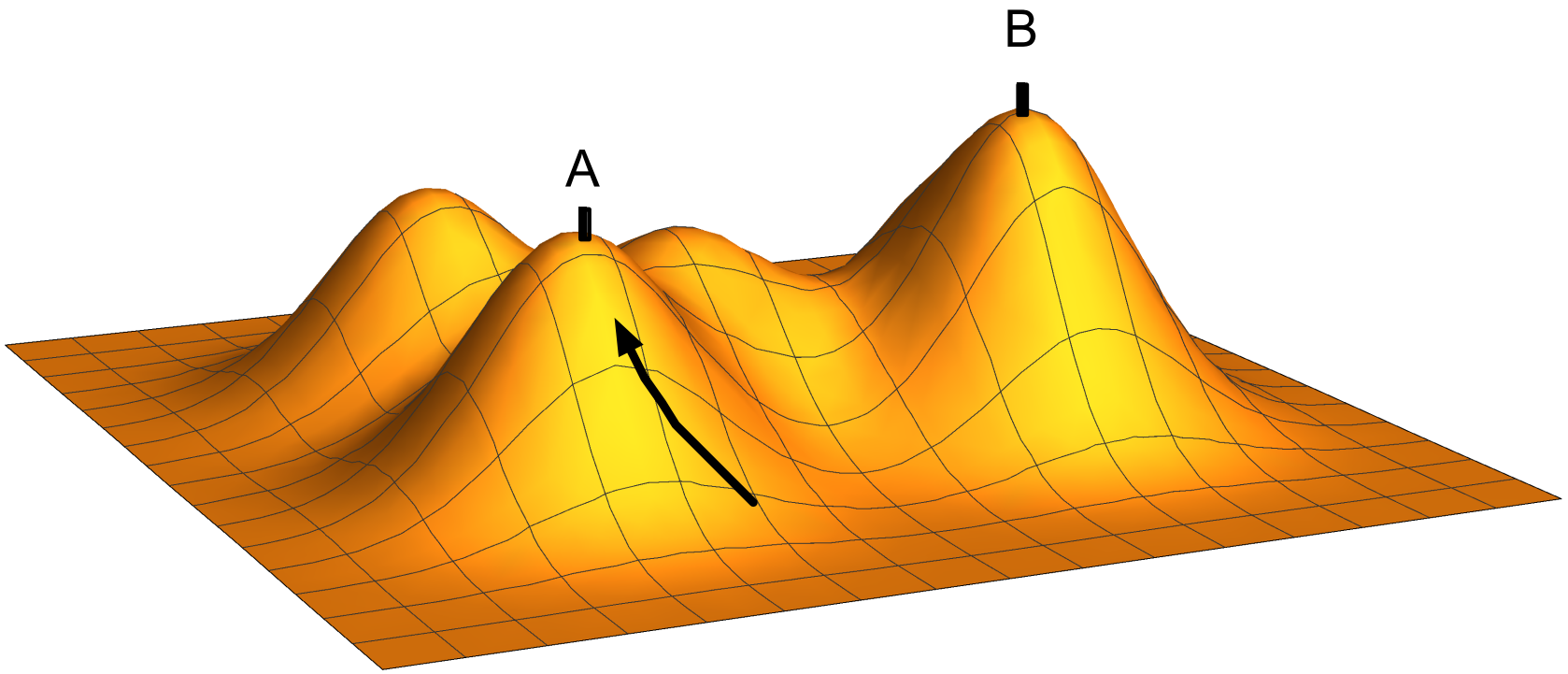
Fitness landscape

General case



Fitness landscape

More general case: time-dependence. Fitness seascape



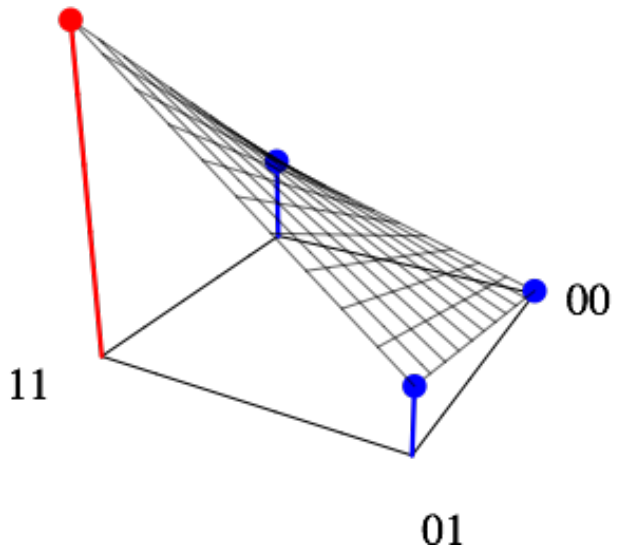
Two-locus, two-allele model

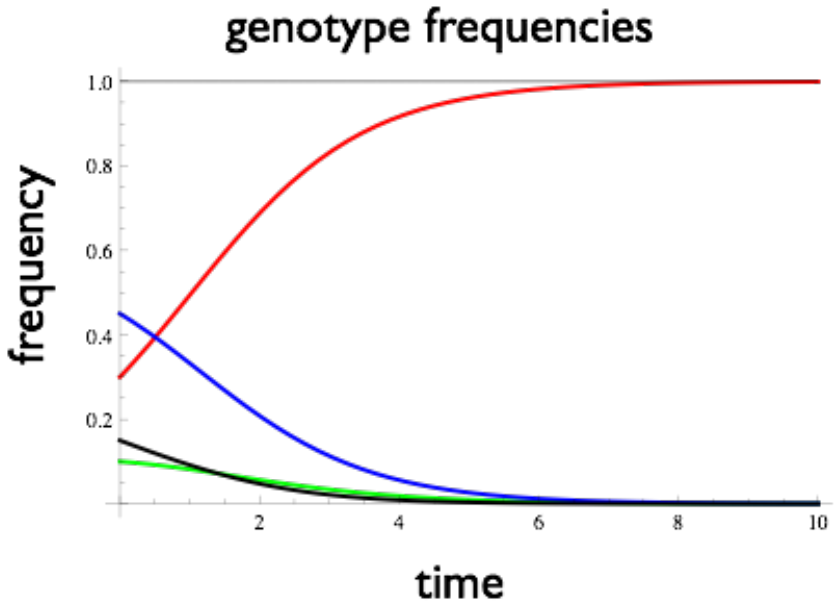
Categories of fitness landscapes

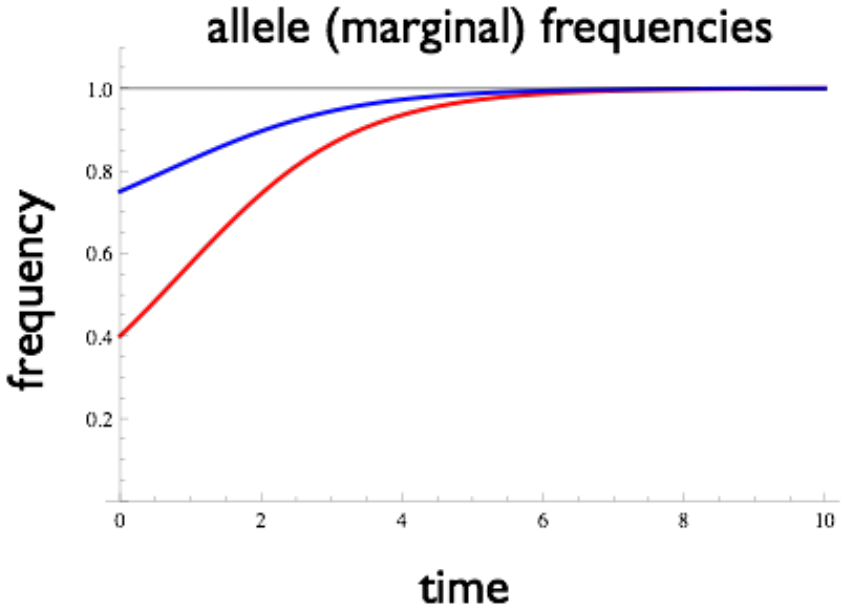
Unique maximum

Genotypes: 11, 10, 01, 00

Marginals: 11+10, 11+01







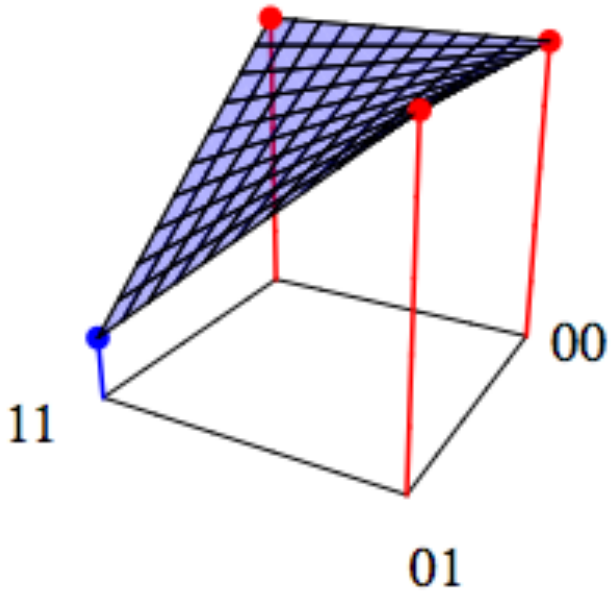
Two-locus, two-allele model

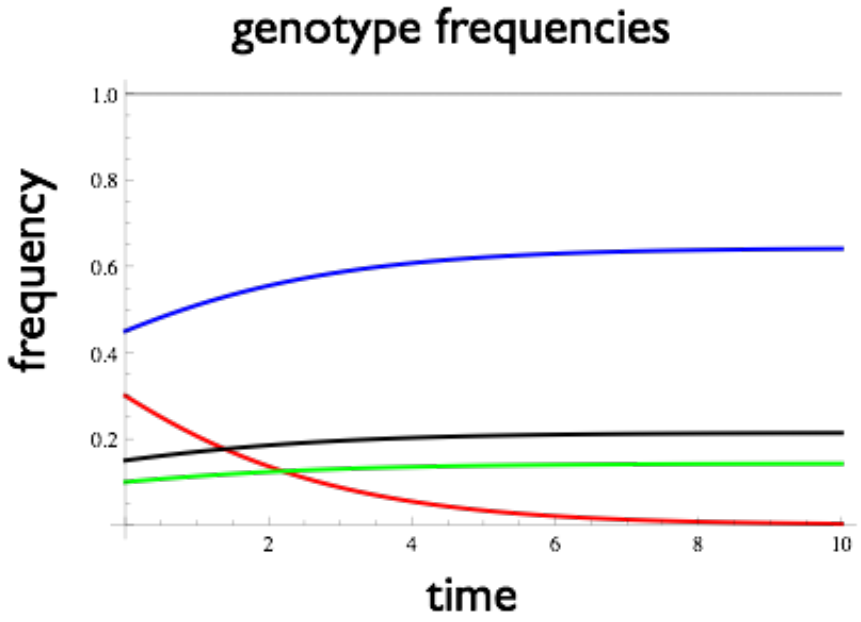
Categories of fitness landscapes

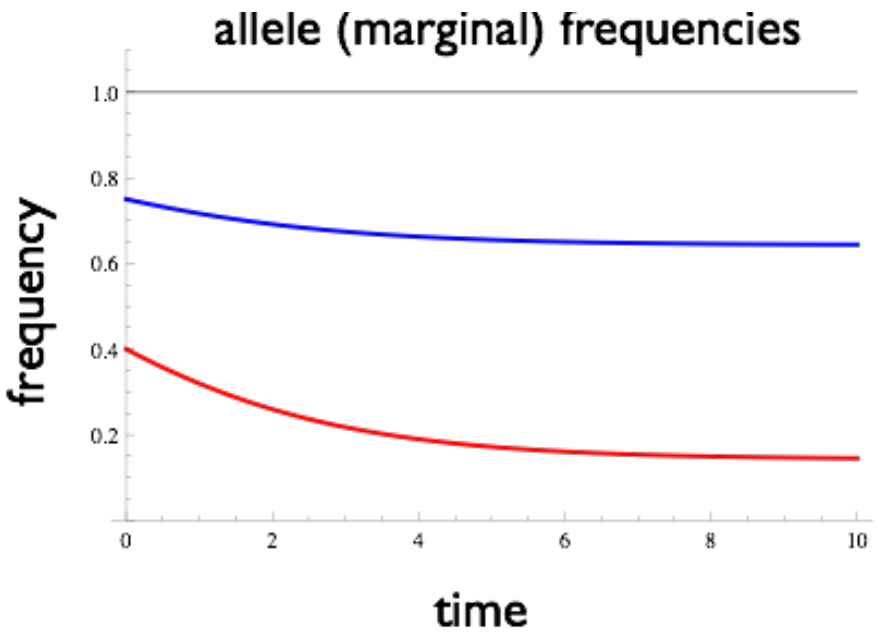
Three-fold degeneracy

Genotypes: 11, 10, 01, 00

Marginals: 11+10, 11+01



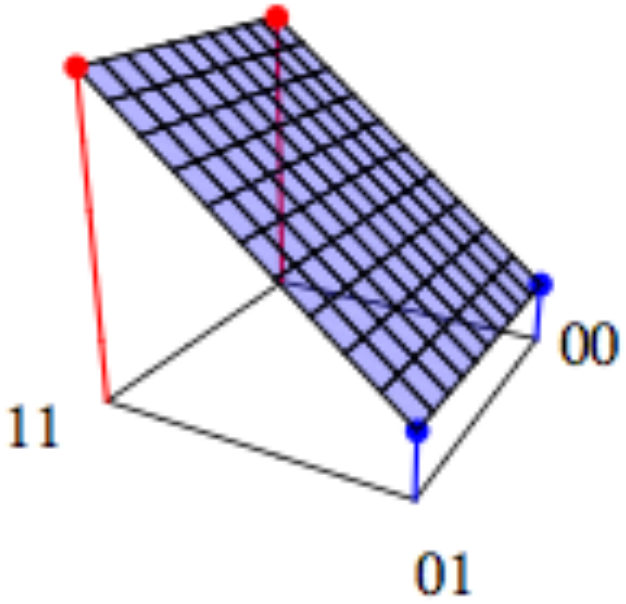


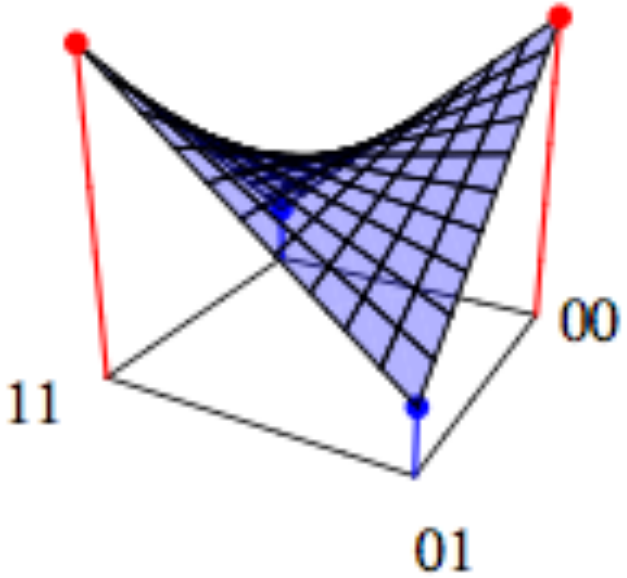


Two-locus, two-allele model

Categories of fitness landscapes

Two-fold degeneracy

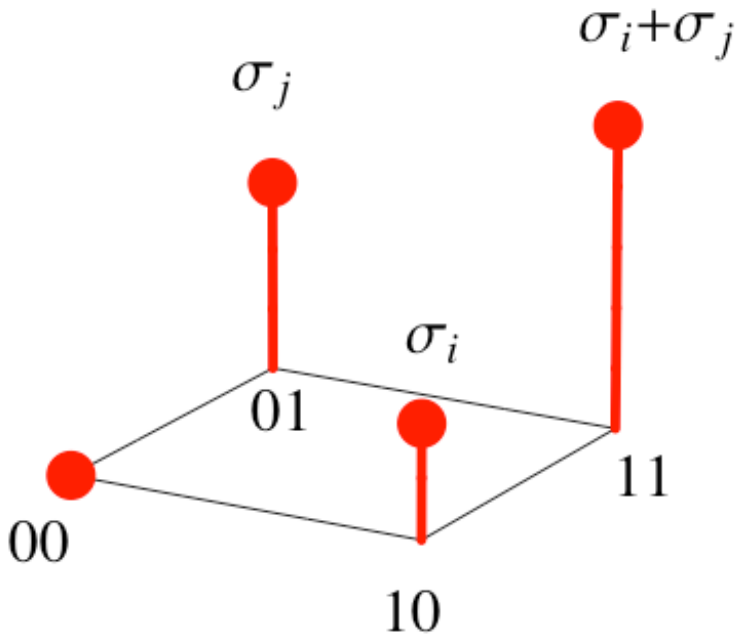




Epistasis

Deviation from additive selection: Epistasis

Additive selection



Epistasis

Epistasis is likely to be prevalent among mutations to proteins

Proteins must:

- Fold into the correct shape
- Perform their proper function

Free energy change associated with protein folding

$$\Delta G_{\text{folding}}$$

Too much destabilisation: protein will no longer fold

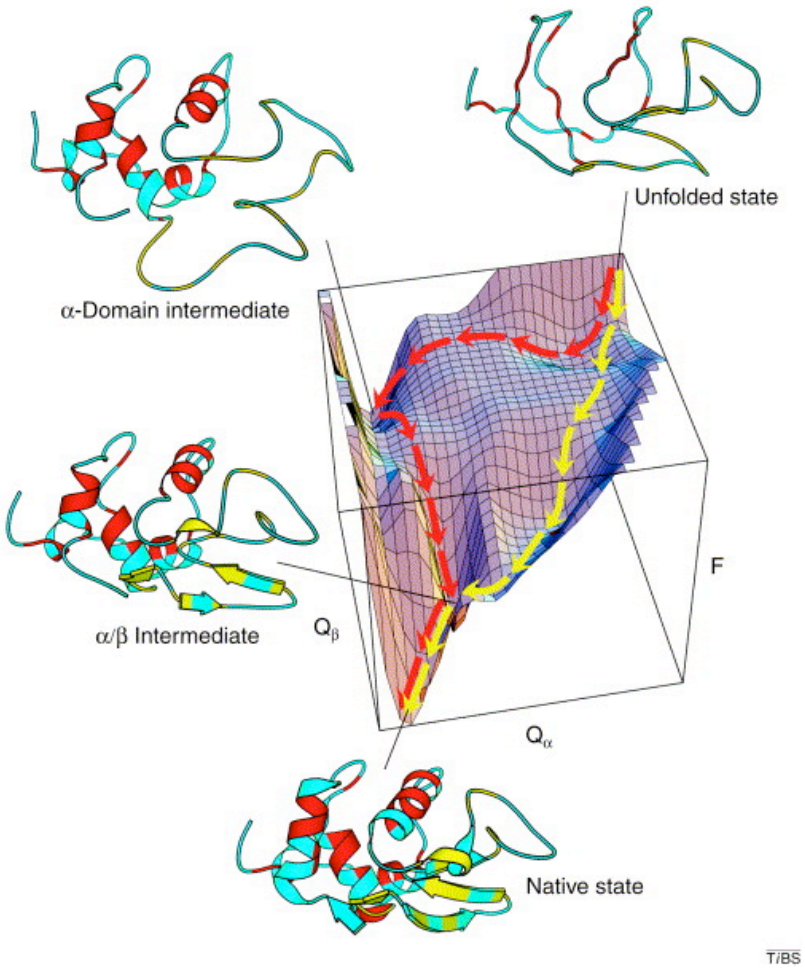
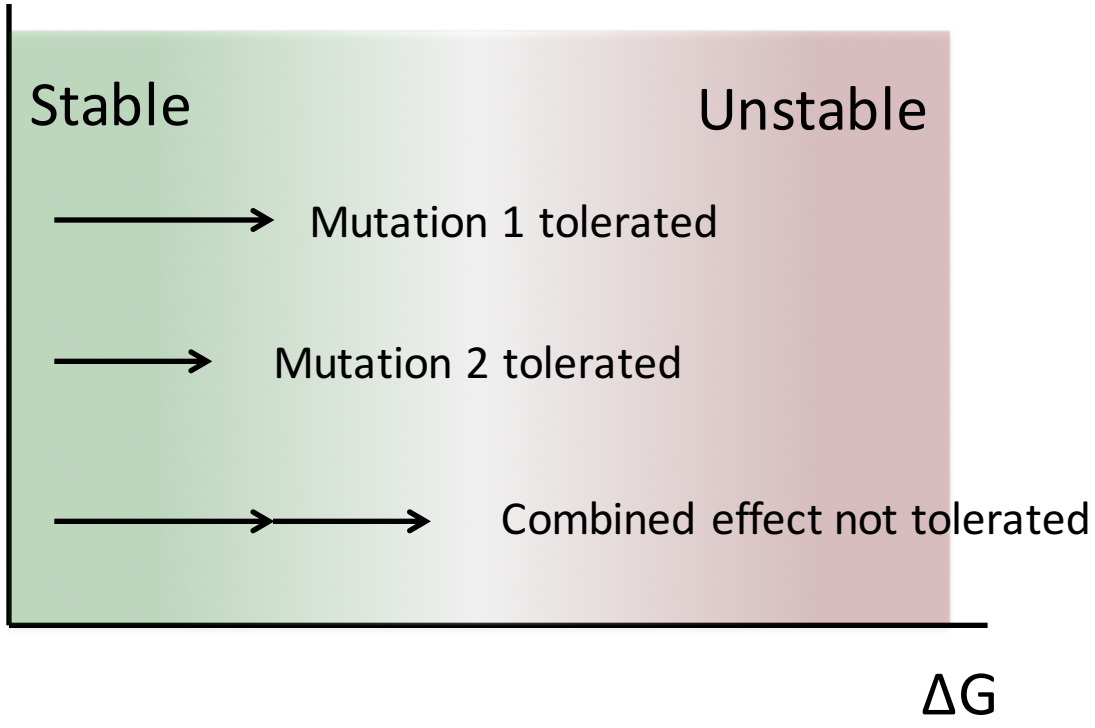


Figure: Dinner et al, Trends in Biochemical Sciences, 2001

Epistasis

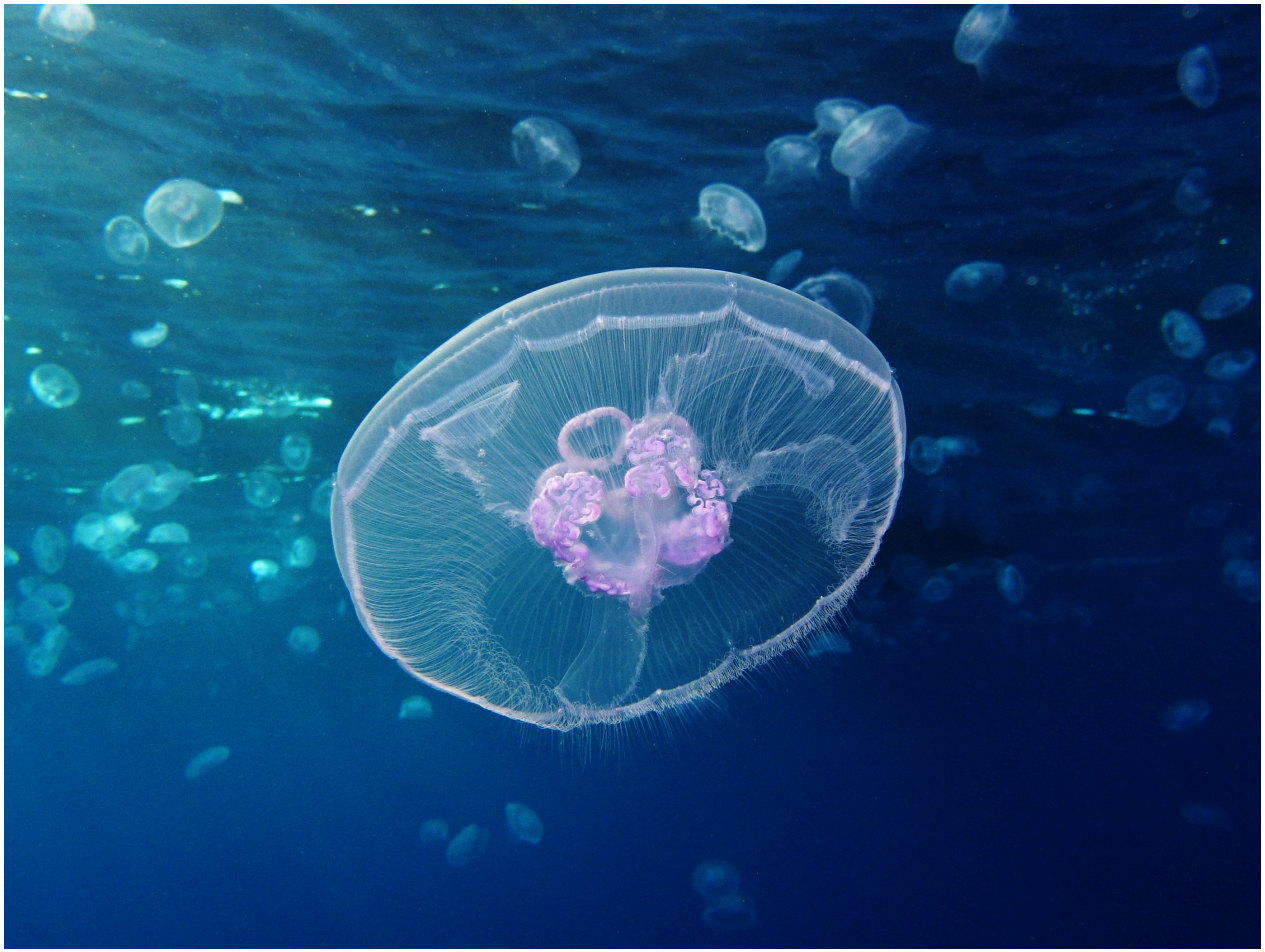
Epistasis is likely to be prevalent among mutations to proteins

Combination of mutations changing ΔG



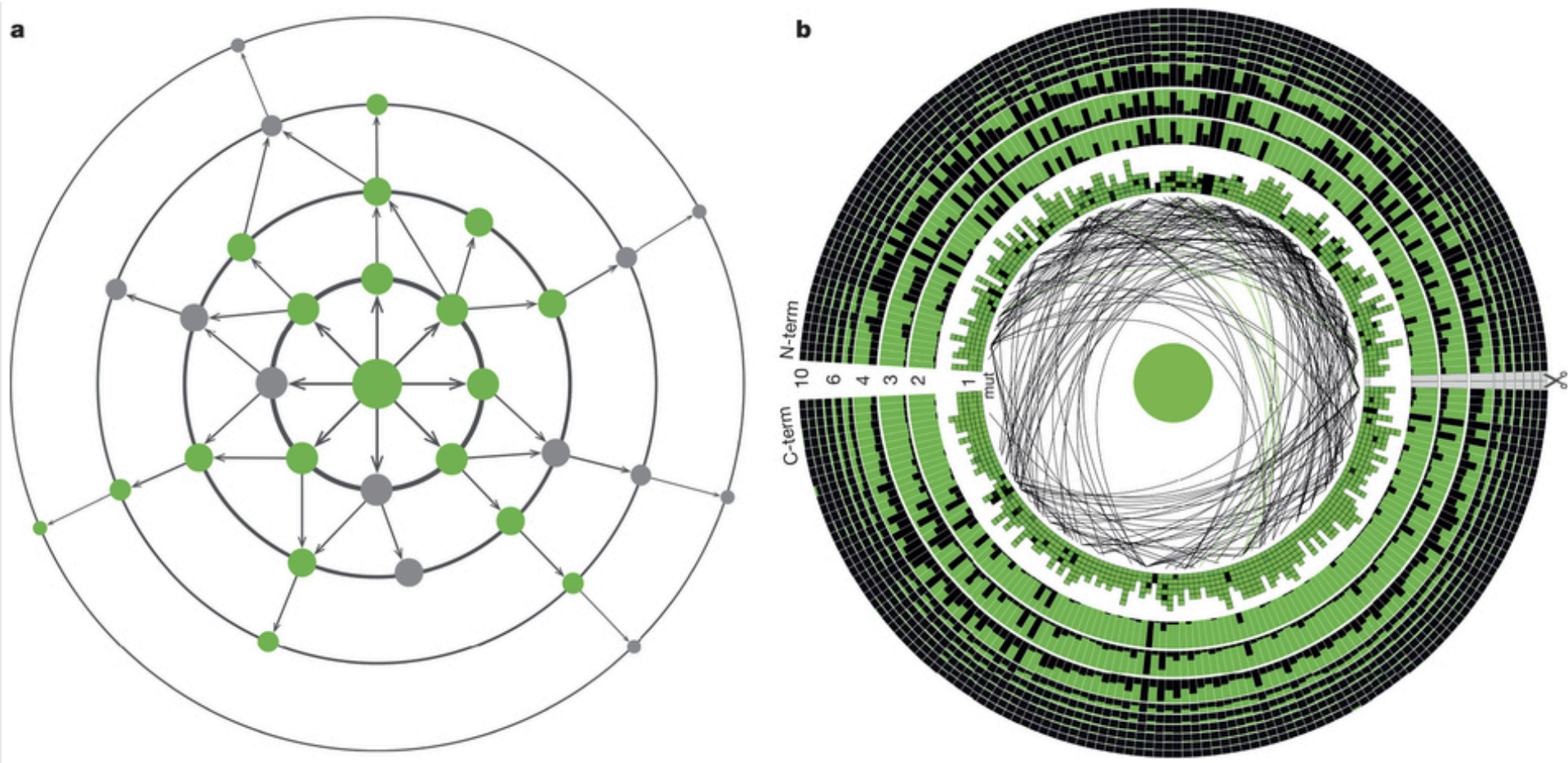
Green fluorescent protein

Identified in jellyfish



Green fluorescent protein

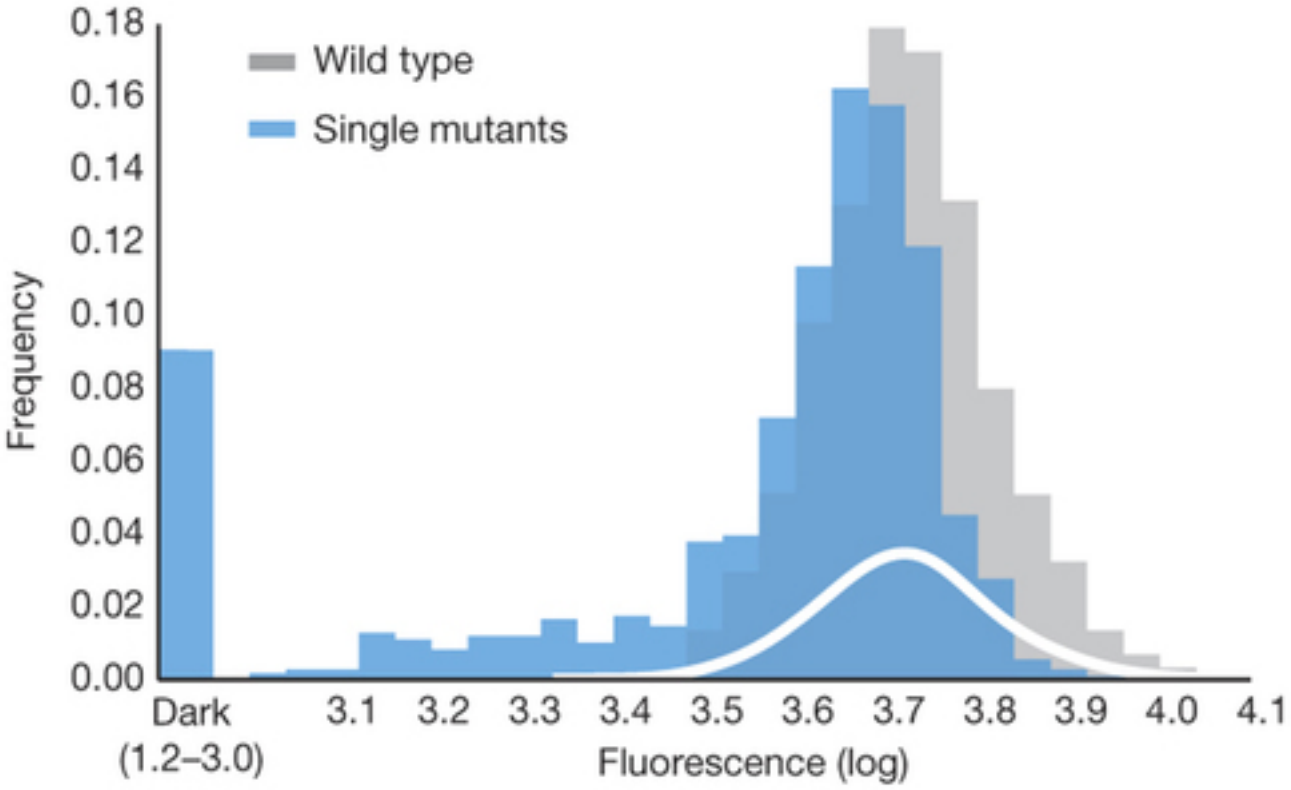
Mutational study: Effects on fluorescence



Sarkysian et al, Nature 2016

Green fluorescent protein

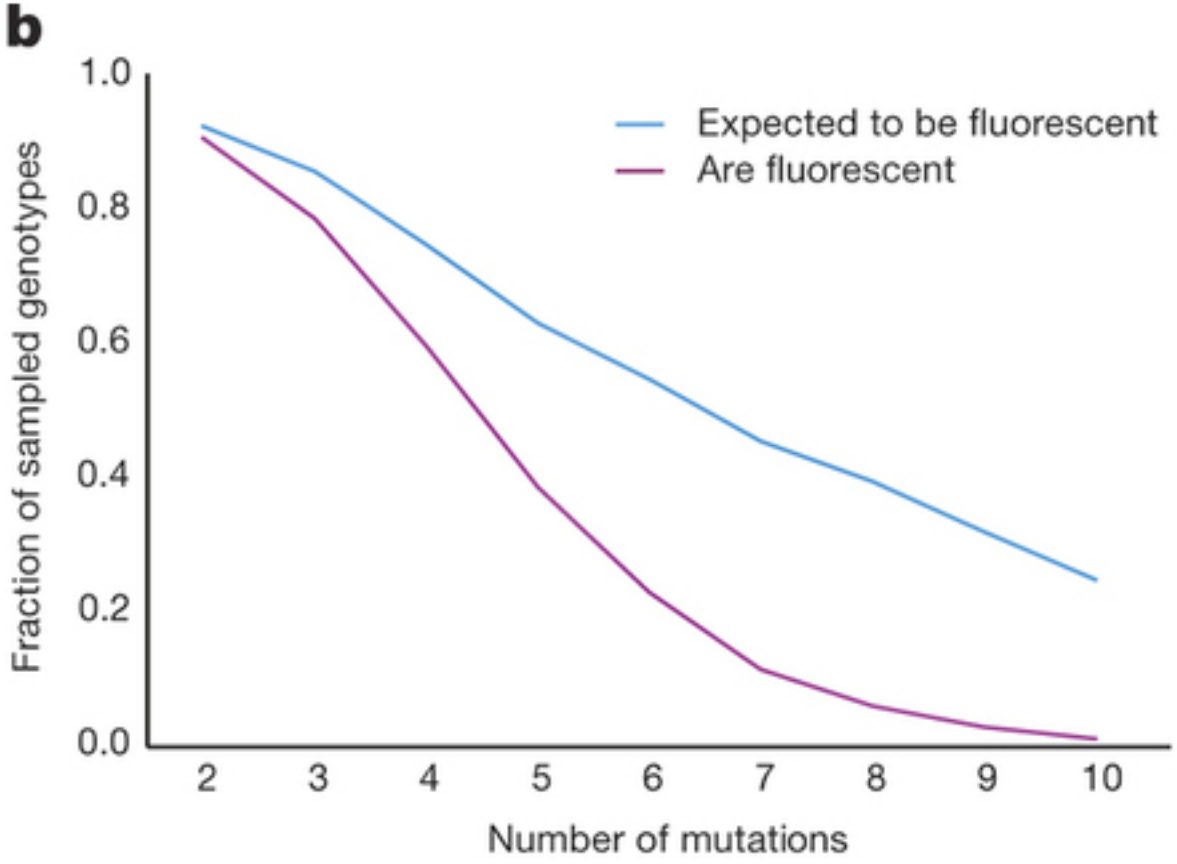
Effect of single mutations



Sarkysian et al, Nature 2016

Green fluorescent protein

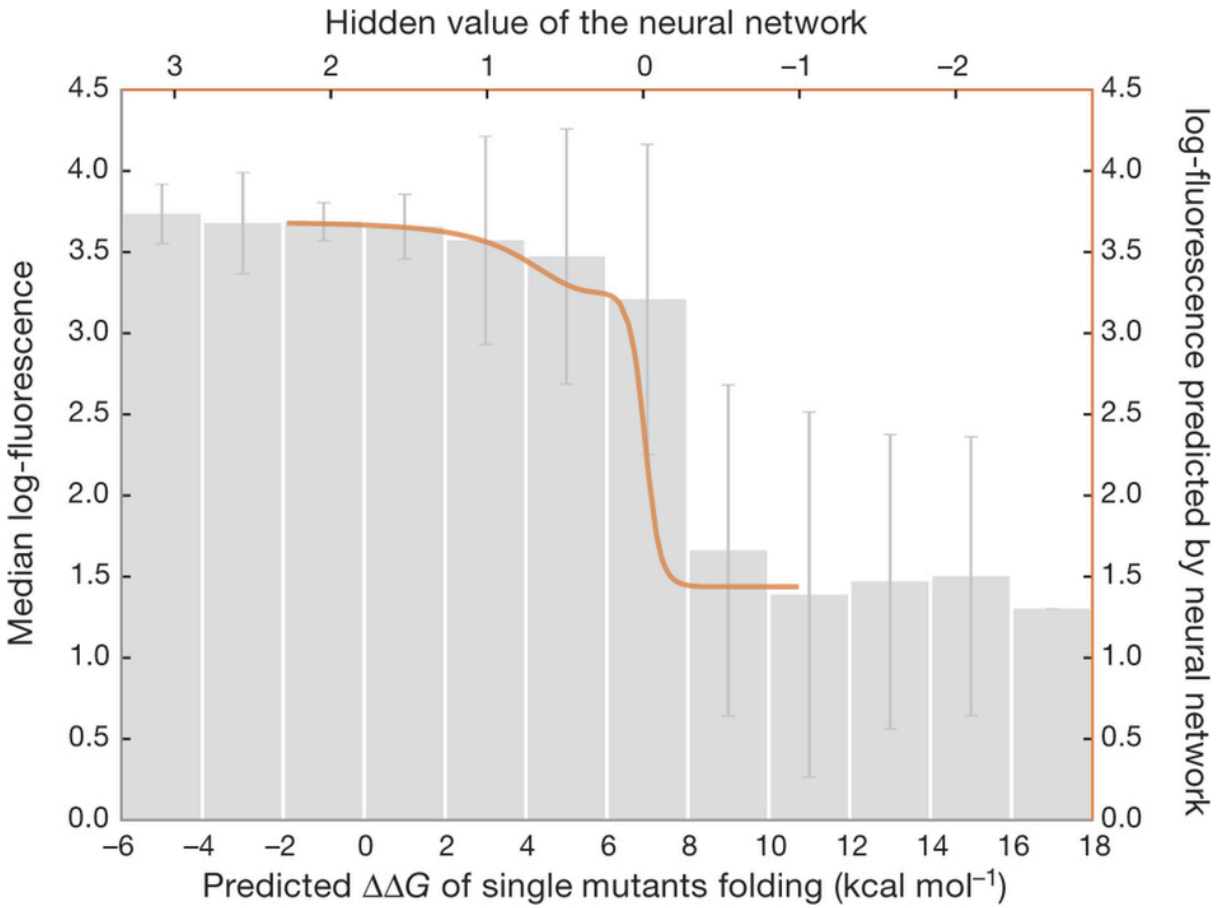
Expected and observed fluorescence of multiple mutations



Sarkysian et al, Nature 2016

Green fluorescent protein

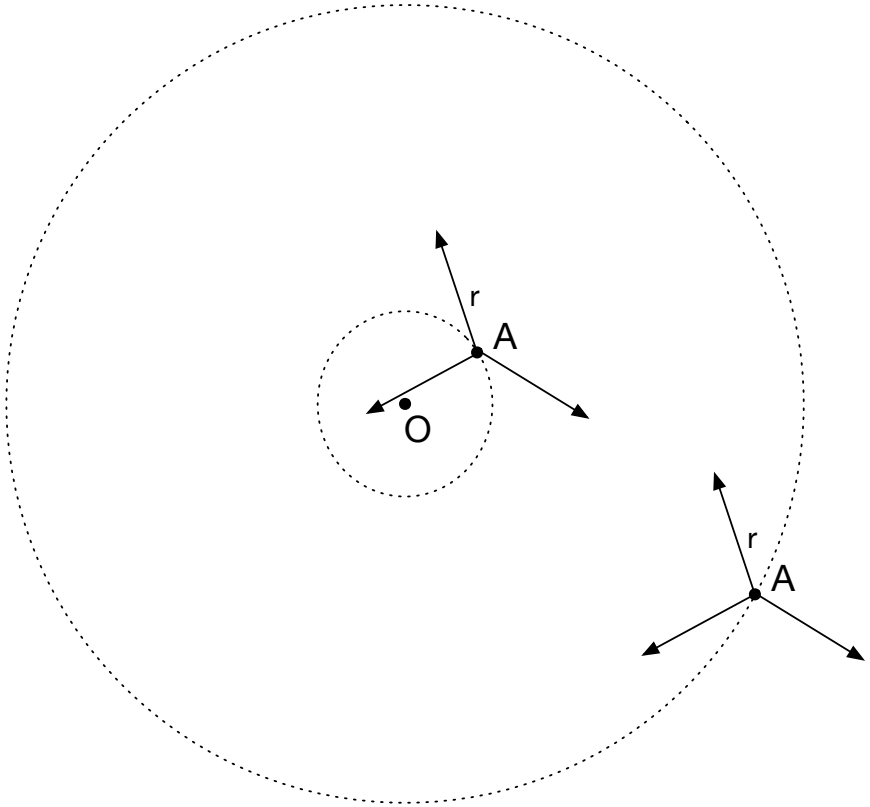
Mutations versus predicted change in protein folding energy



Sarkysian et al, Nature 2016

Another approach: Fisher's geometric model

Generic model of fitness and adaptation



Point O represents fitness maximum

Summary

More complex evolutionary scenarios:

Diploid genomes

Frequency-dependent selection: Evolutionary games

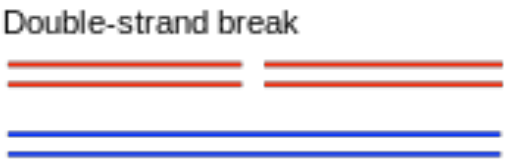
Two locus, two allele model

Fitness landscapes and epistasis

Recombination

Molecular process

Recombination occurs through a process of DNA breakage and repair

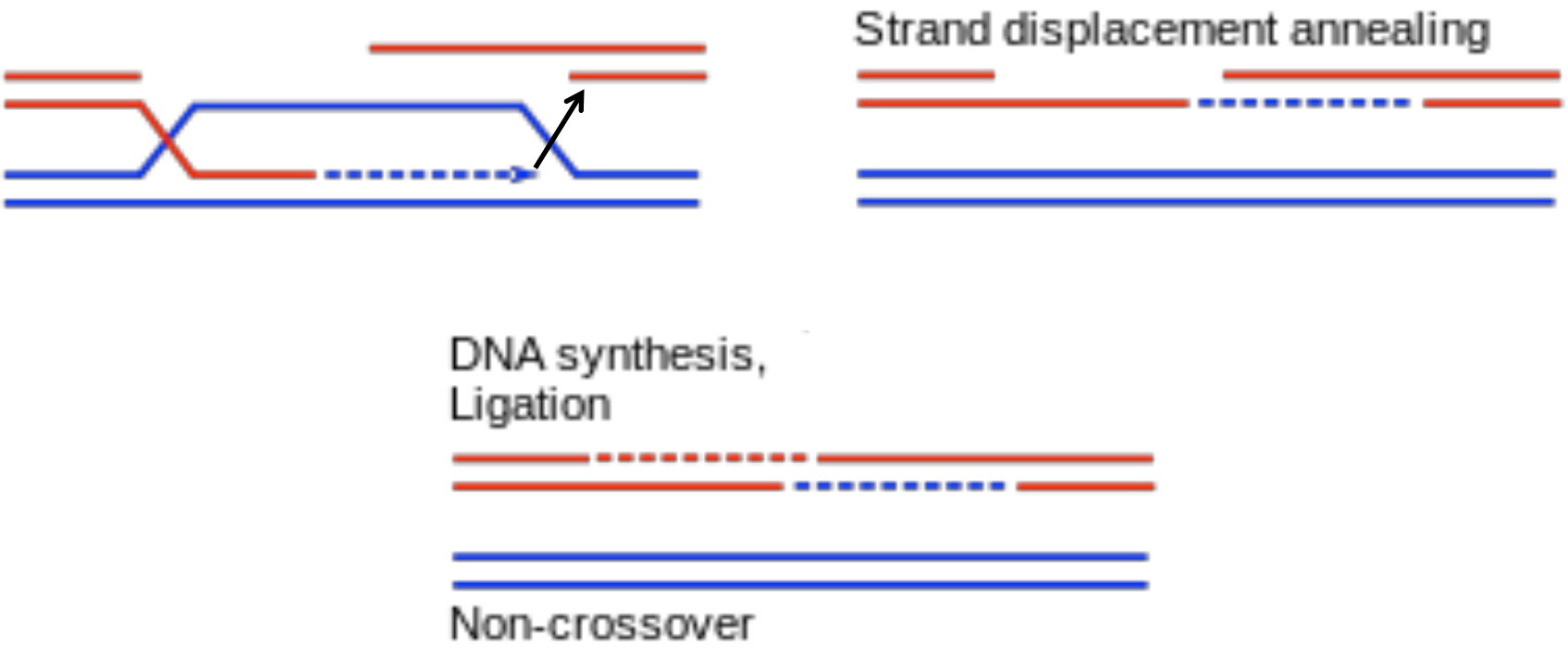


Sung and Klein, Nat Rev Mol Cell Biol, 2006

Genetic recombination

Molecular process

Recombination occurs through a process of DNA breakage and repair

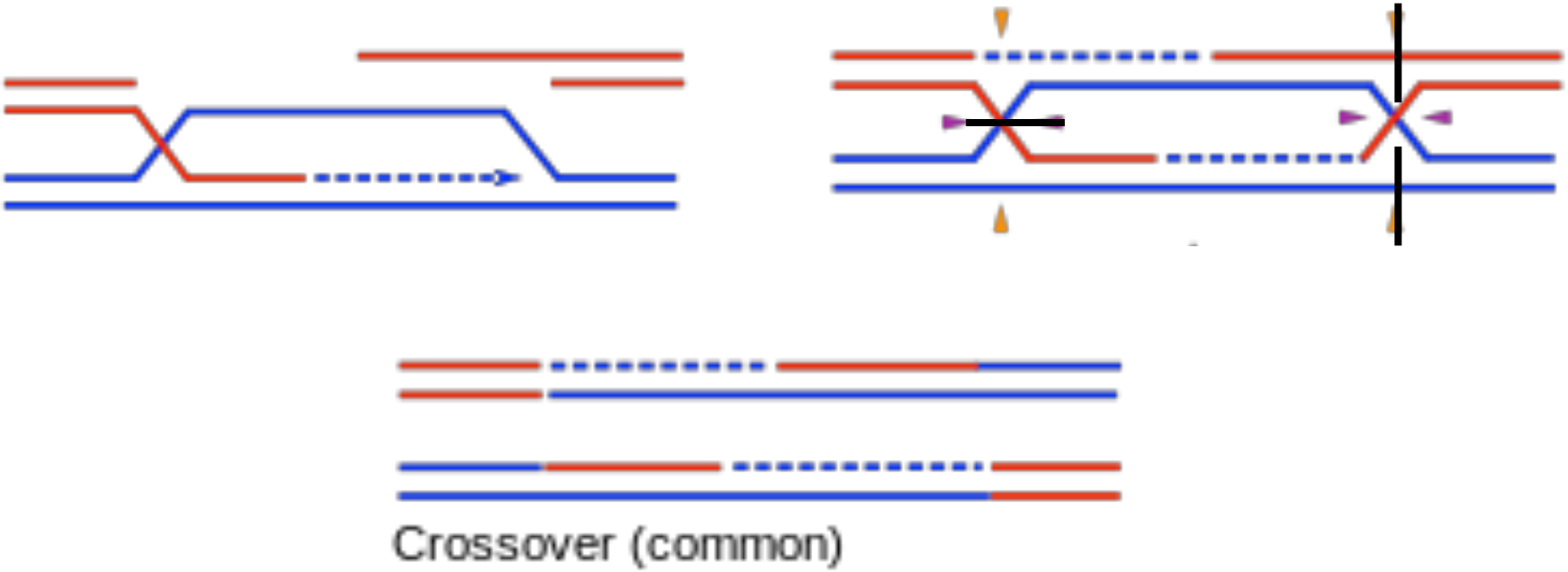


Sung and Klein, Nat Rev Mol Cell Biol, 2006

Genetic recombination

Molecular process

Recombination occurs through a process of DNA breakage and repair

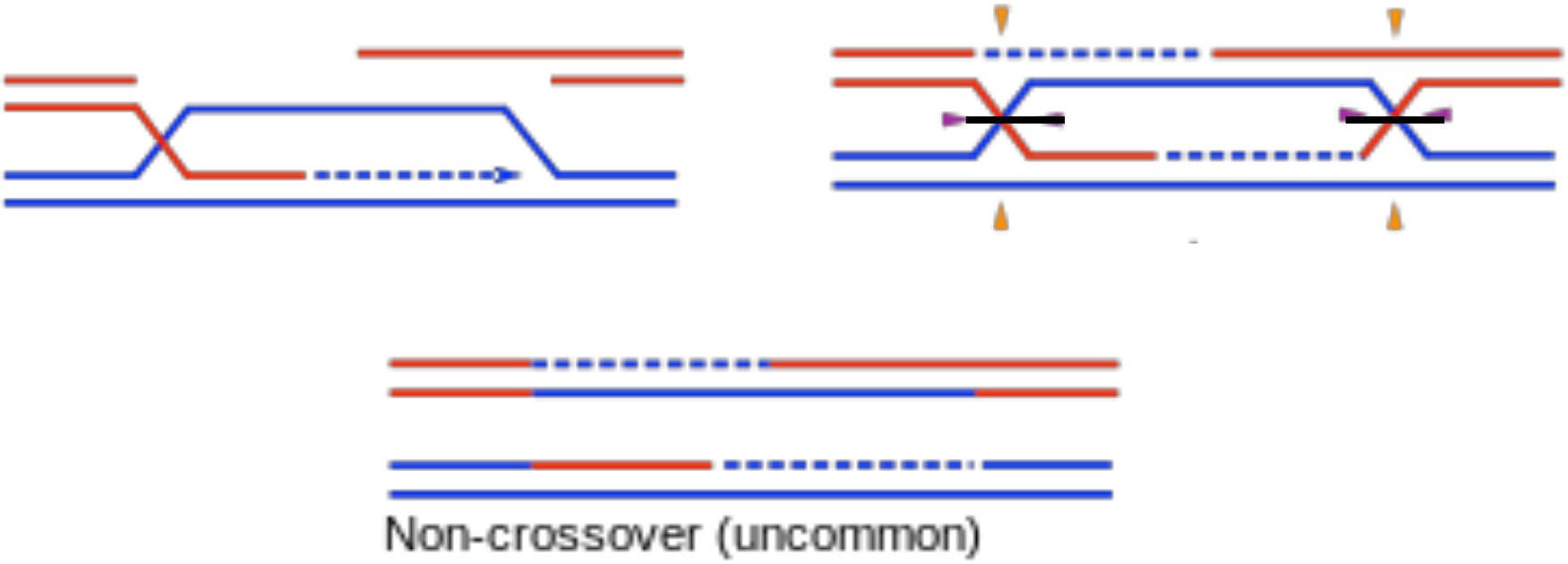


Sung and Klein, Nat Rev Mol Cell Biol, 2006

Genetic recombination

Molecular process

Recombination occurs through a process of DNA breakage and repair

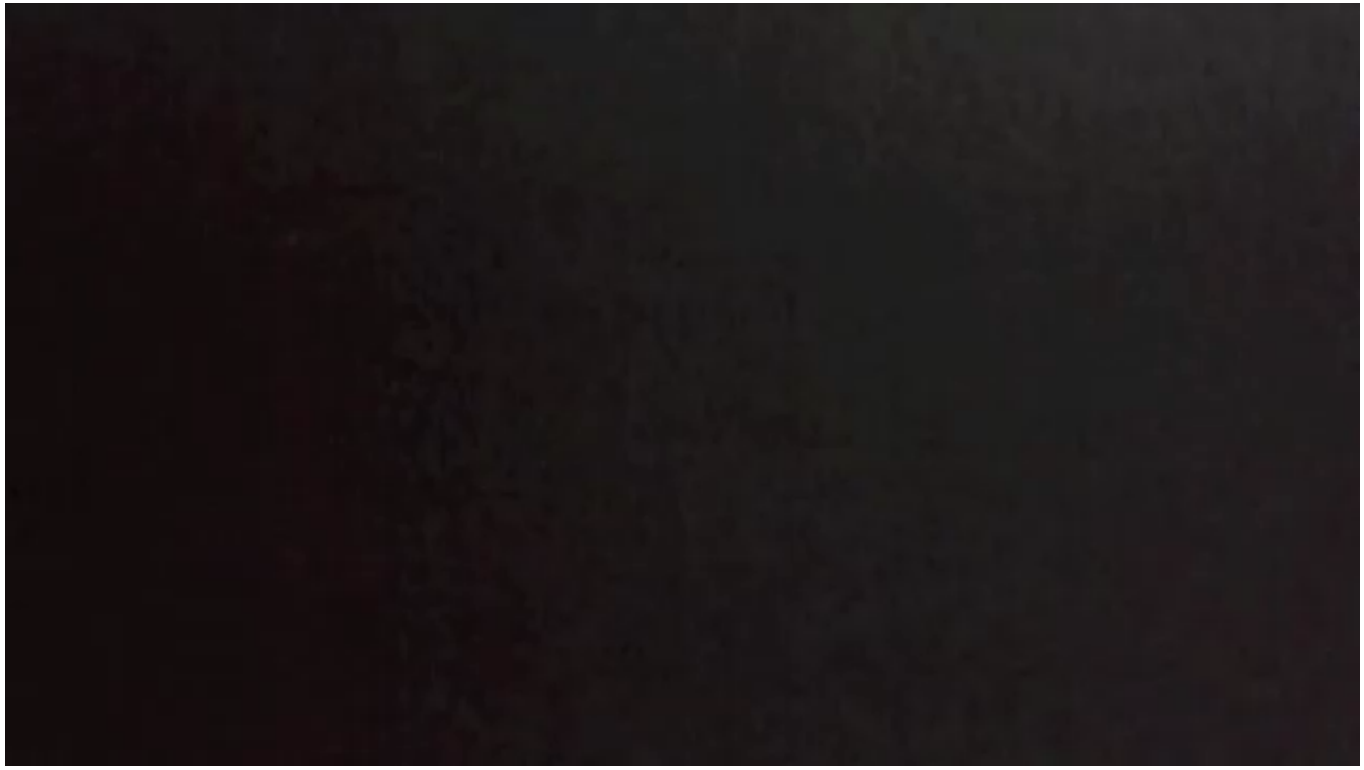


Sung and Klein, Nat Rev Mol Cell Biol, 2006

Genetic recombination

Molecular process

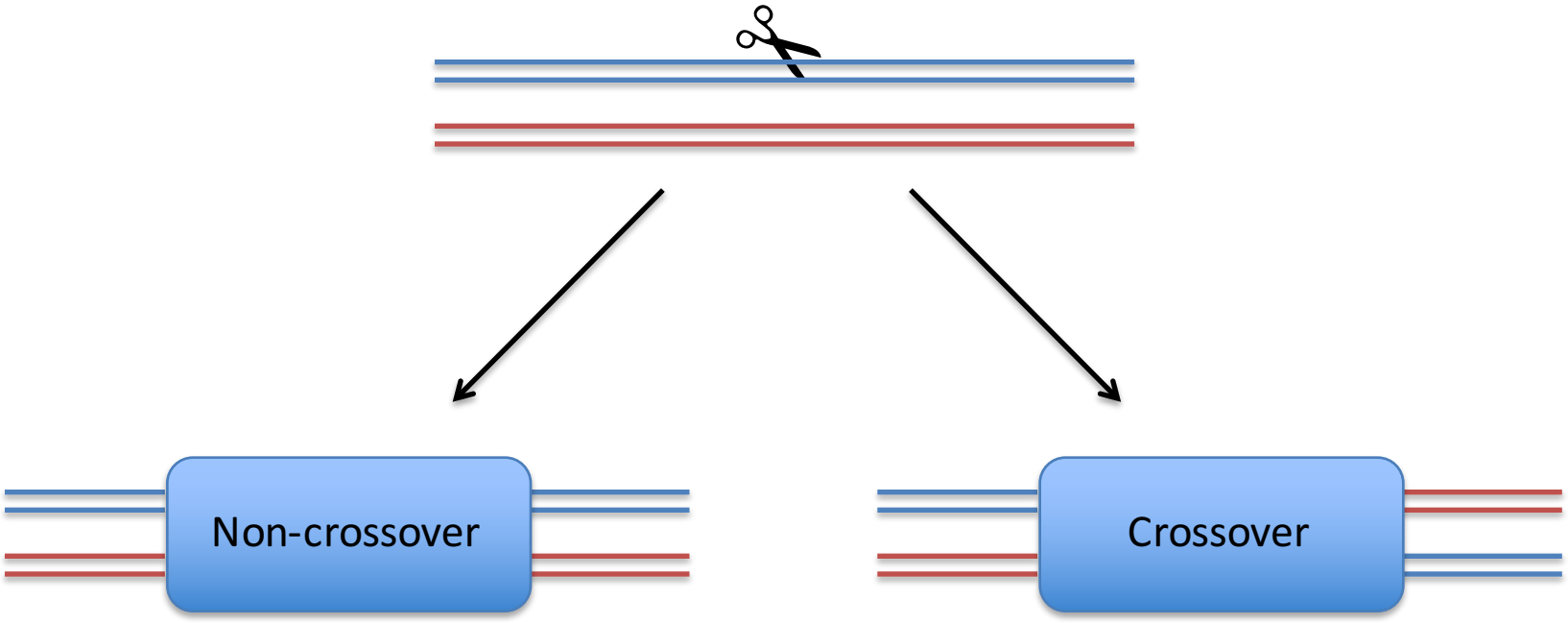
Recombination occurs through a process of DNA breakage and repair



Genetic recombination

Molecular process

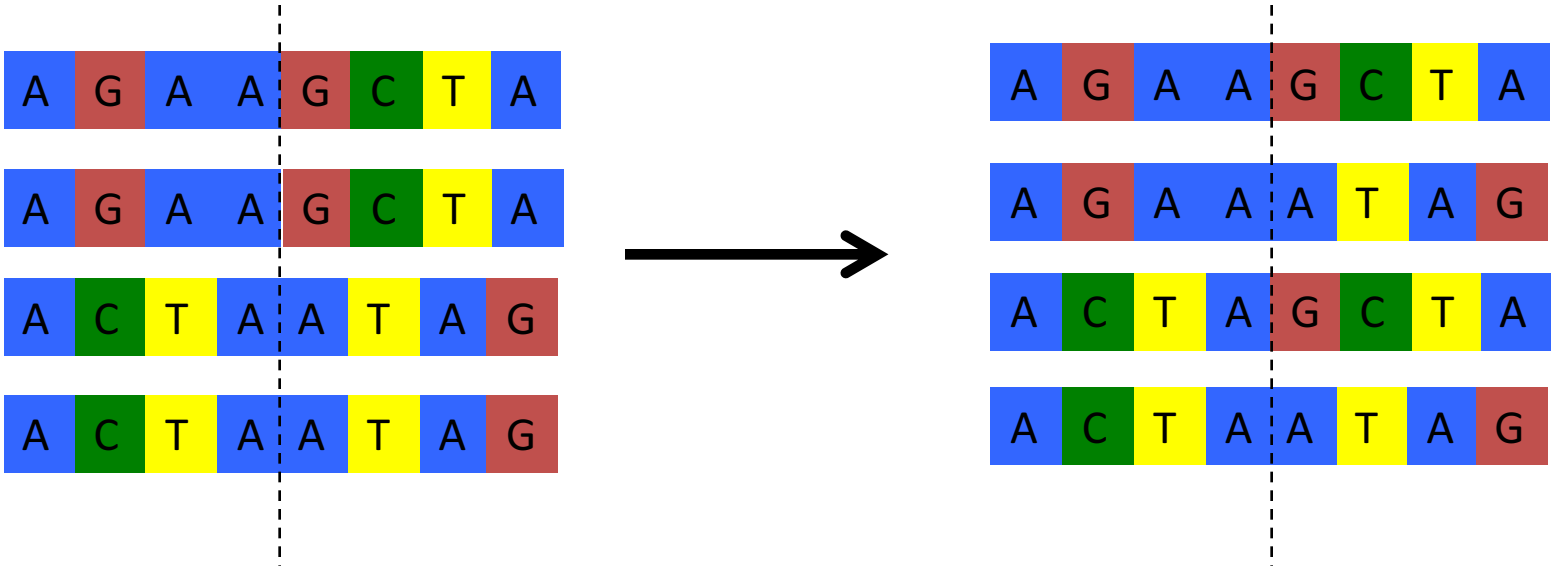
Recombination occurs through a process of DNA breakage and repair



Genetic recombination

Effect of recombination

Produce new combinations of alleles

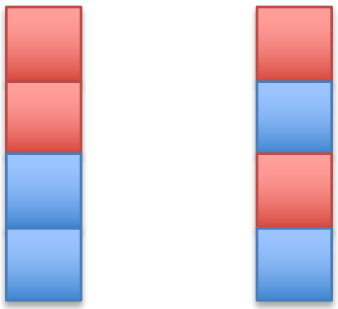


Linkage disequilibrium (LD)

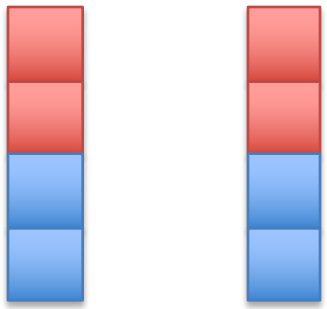
Definition

Non-random association of alleles at two or more loci

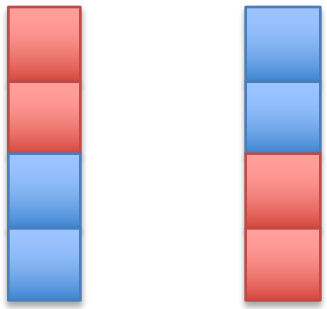
$$D_{ij} = q_{ij}^{11} - q_i^1 q_j^1$$



$D_{ij} = 0$



$D_{ij} = 0.25$

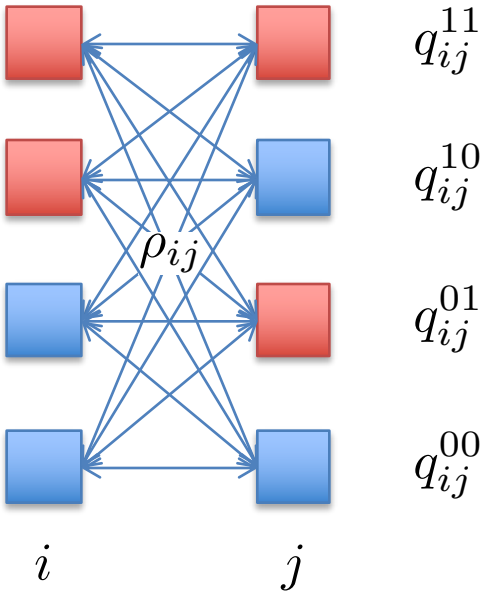


$D_{ij} = -0.25$

Linkage disequilibrium

LD and Recombination

Recombination breaks linkage disequilibrium



Allele frequencies unchanged by recombination

Among non-recombinant sequences:

$$q_{ij}^{11}(t+1) = q_{ij}^{11}(t)$$

Among recombinant sequences:

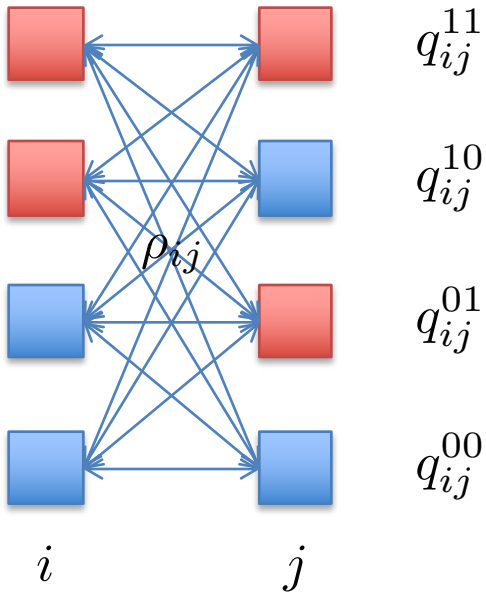
$$q_{ij}^{11}(t+1) = q_i^1(t) q_j^1(t)$$

$$D_{ij} = q_{ij}^{11} - q_i^1 q_j^1$$

Linkage disequilibrium

LD and Recombination

Recombination breaks linkage disequilibrium



$$q_{ij}^{11}(t+1) = (1-\rho_{ij})q_{ij}^{11}(t) + \rho_{ij}q_i^1(t)q_j^1(t)$$

$$D_{ij}(t+1) = (1-\rho_{ij})D_{ij}(t)$$

Linkage disequilibrium decays to zero exponentially over time

$$D_{ij} = q_{ij}^{11} - q_i^1 q_j^1$$

Summary

More complex evolutionary scenarios:

Diploid genomes

Frequency-dependent selection: Evolutionary games

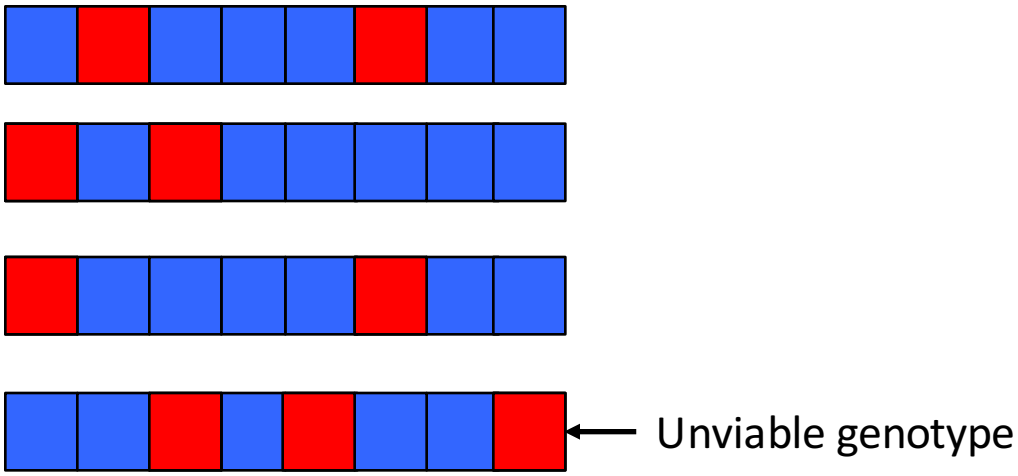
Two locus, two allele model

Fitness landscapes and epistasis

Recombination and linkage disequilibrium

Importance of recombination: Muller's Ratchet

Accumulation of deleterious mutations

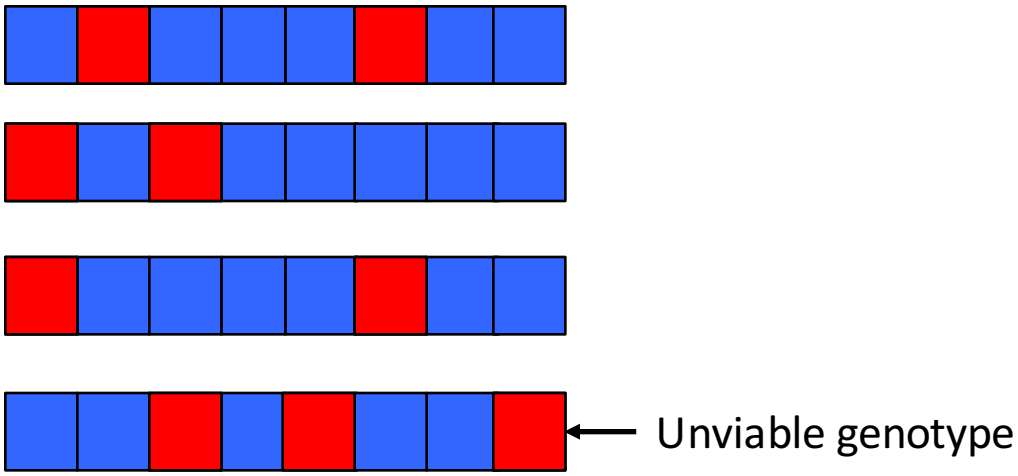


In each generation, each individual copies itself

Some rate of mutation: most mutations are deleterious

Deleterious mutations: Muller's Ratchet

Accumulation of deleterious mutations

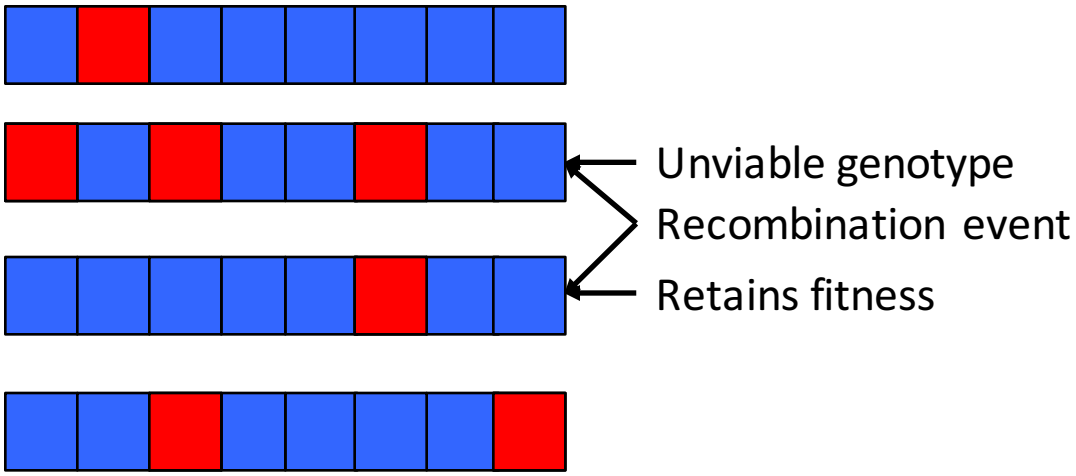


Population fitness decreases over time

Fitness can only be recovered by reverse mutation

Deleterious mutations: Muller's Ratchet

Sexual reproduction allows mutations to be recombined out



Recombination swaps combinations of alleles

Allows for fitter genotypes to be maintained; deleterious mutants recombined out of the population.

Importance of recombination: Adaptation

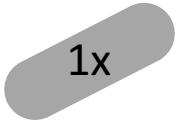
***E. coli* populations with differing mutation supply rates**

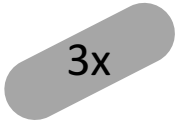
Two strains :

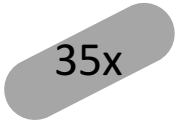
Adapted (post-10k generations)

Non-adapted

Varying mutation rates:

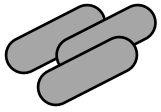




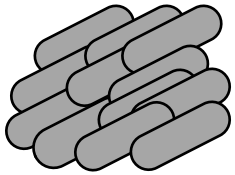


Disable DNA-repair genes

Varying population sizes:



Standard size

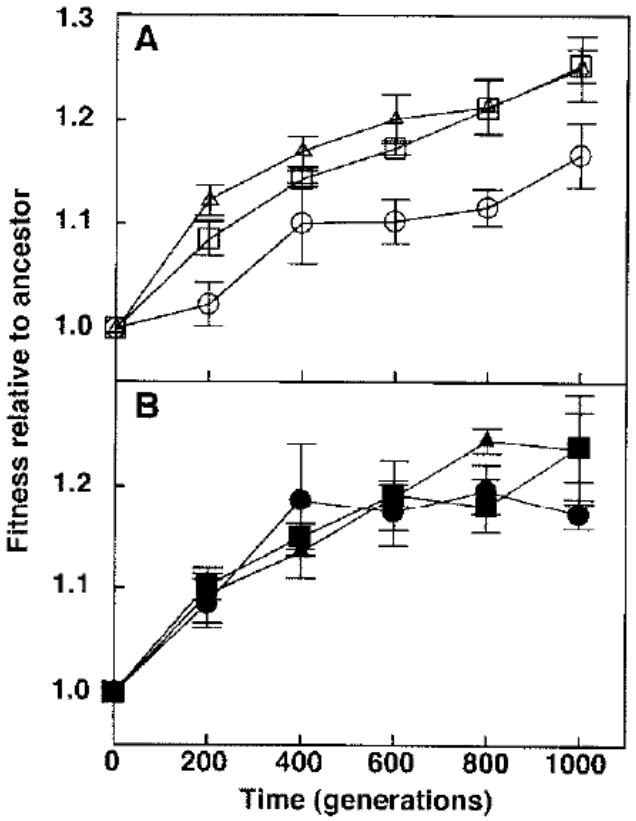


Extra-large size (50x individuals)

“Speed limit” on adaptive asexual evolution

Non-adapted population fitness

Increase seen in every case



Small population

Large population

Relative mutation rate

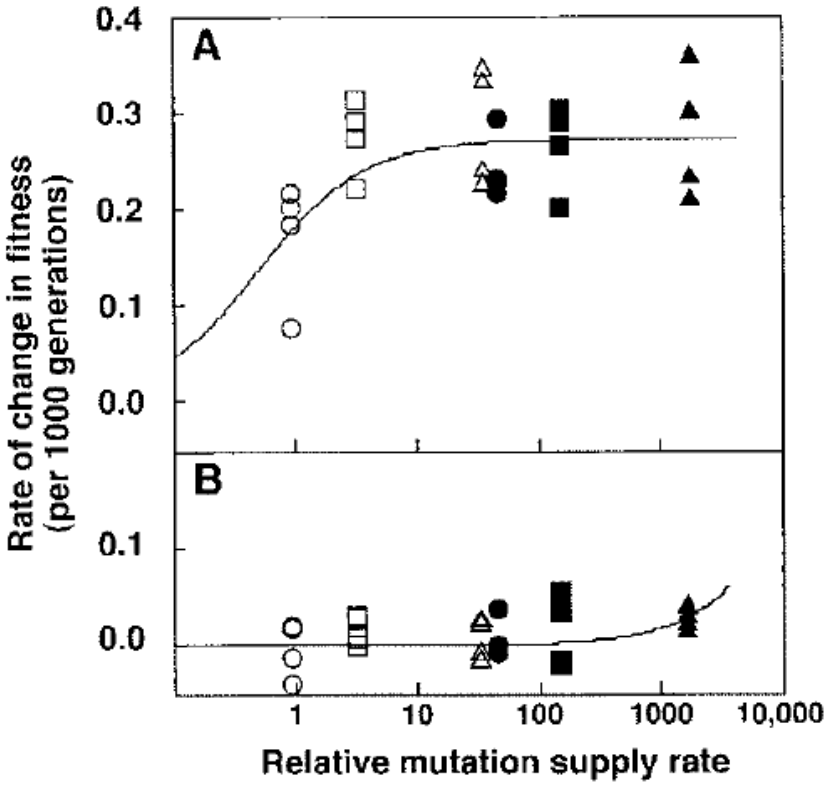
O	1
□	3.3
Δ	34.5

de Visser et al, Science, 1999

“Speed limit” on adaptive asexual evolution

Adaptation rate vs mutation supply Nμ

Plateau in the rate of adaptation



Non-adapted population

Adapted population

Relative mutation rate

O	1
□	3.3
Δ	34.5

Small population

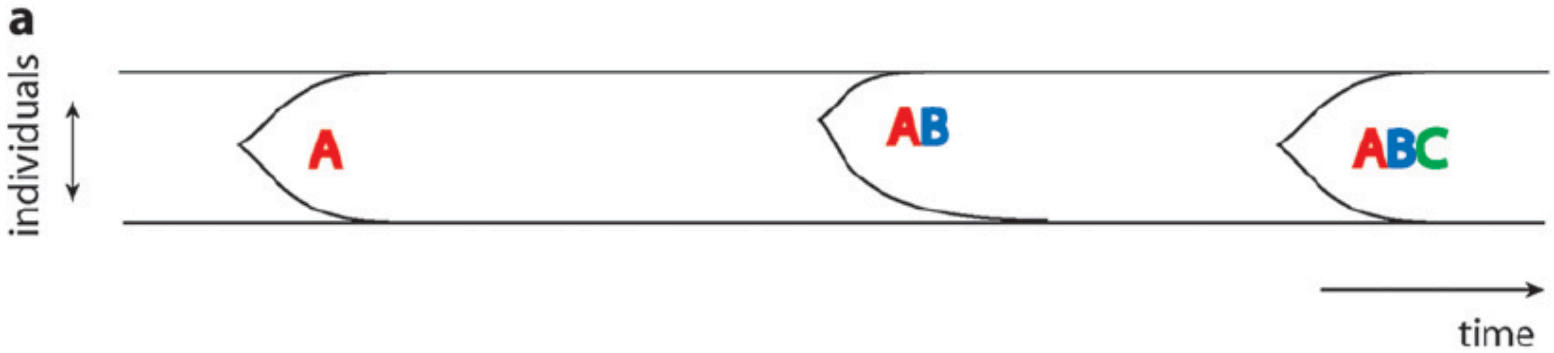
Large population

de Visser et al, Science, 1999

“Speed limit” on adaptive asexual evolution

Clonal interference slows adaptation

Low mutation rates: beneficial mutations appear slowly, but are likely to fix

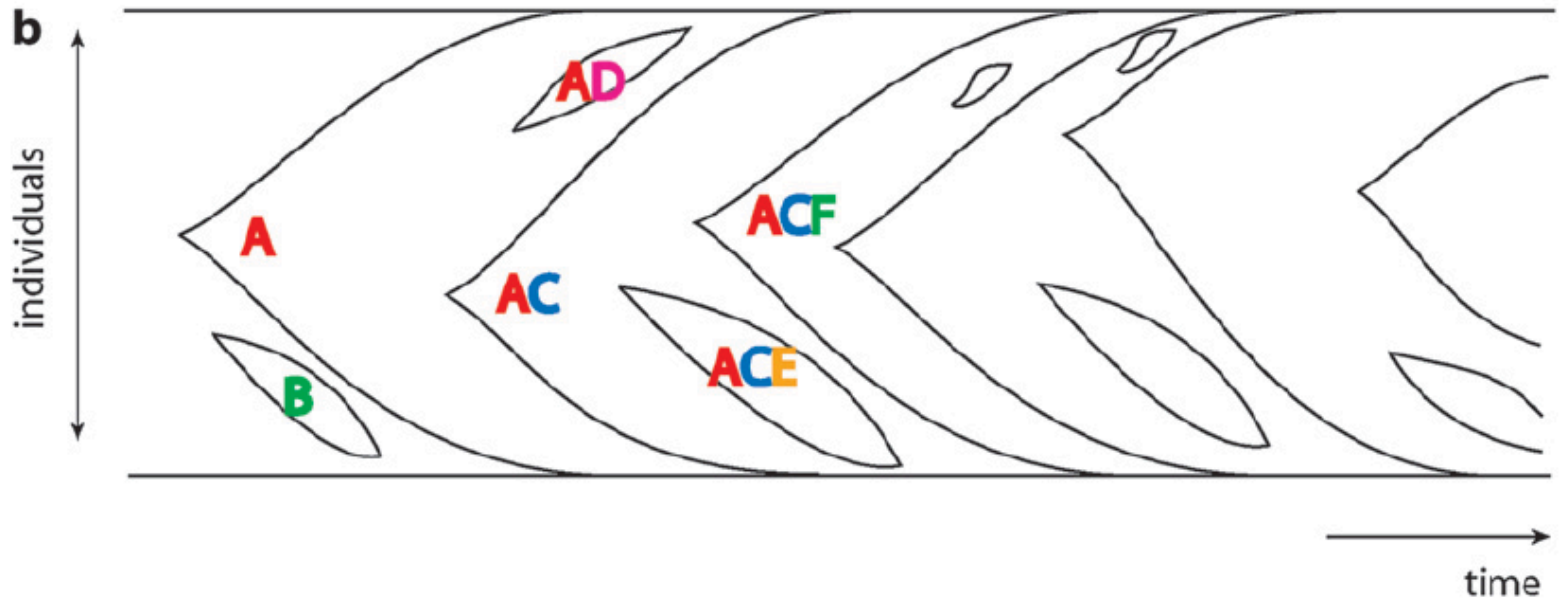


Desai and Fisher, Genetics, 2007

“Speed limit” on adaptive asexual evolution

Clonal interference slows adaptation

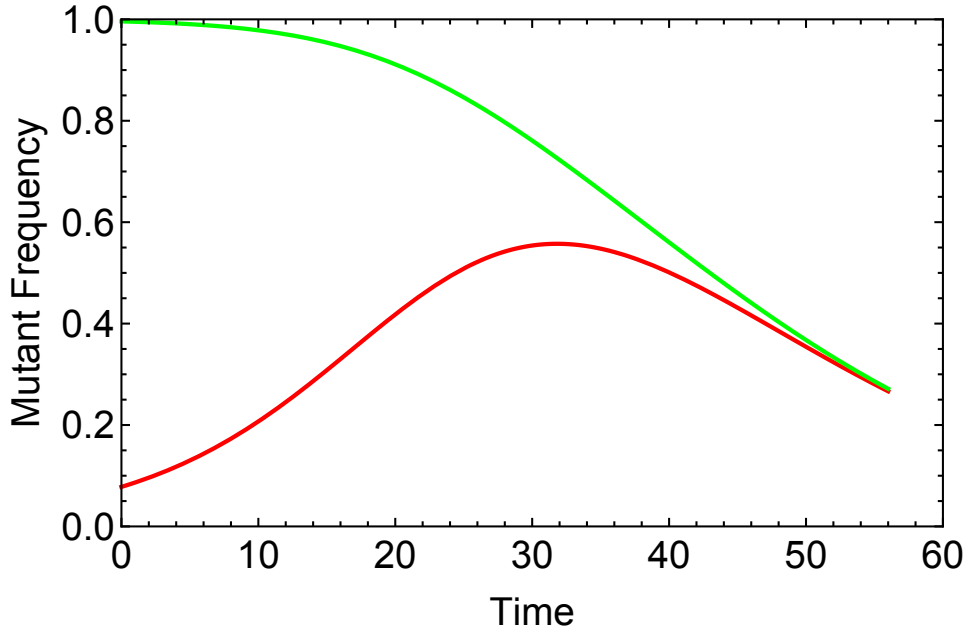
High mutation rates: beneficial mutations appear rapidly, but outcompete one another: only the strongest survive.



Desai and Fisher, Genetics, 2007

“Speed limit” on adaptive asexual evolution

Clonal interference slows adaptation



Sexual vs asexual reproduction

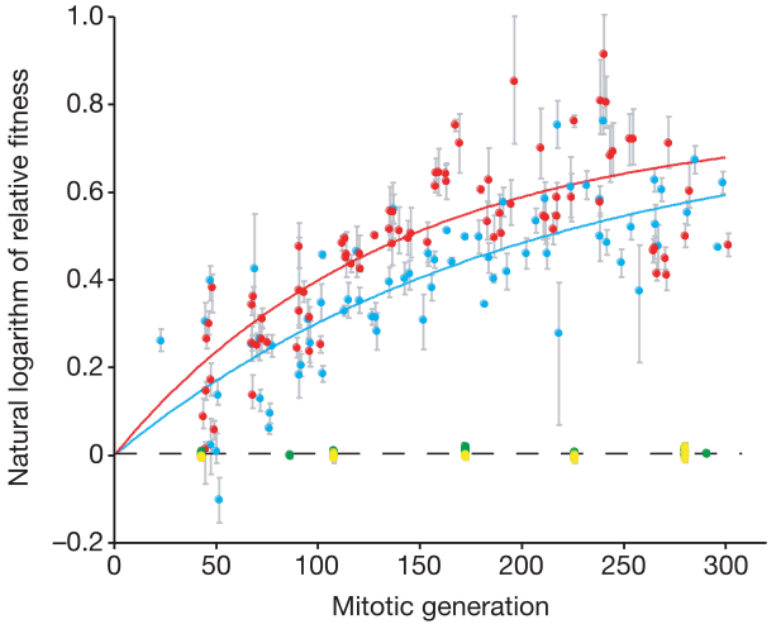
Reproduction in yeast

Sexual reproduction allows for more rapid adaptation to harsh environmental conditions

Yeast grown in chemostat at 37°C, and at 0.2M concentration NaCl

Wild-type yeast

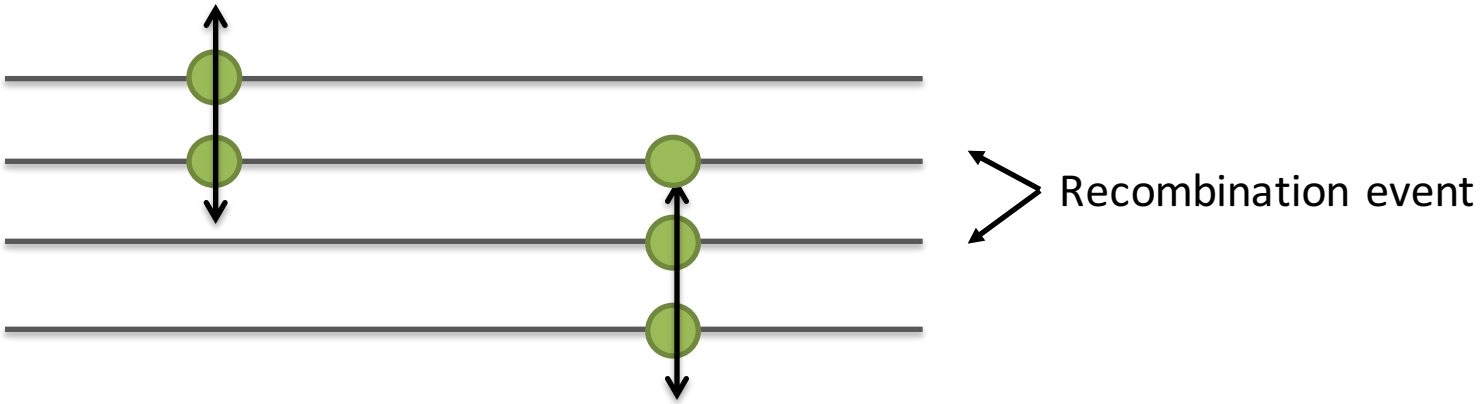
Mutant yeast, deletion of gene required for sexual reproduction



Goddard et al., Nature, 2005

“Speed limit” on adaptive asexual evolution

Recombination breaks the speed limit



Beneficial mutations on opposing haplotypes oppose each other

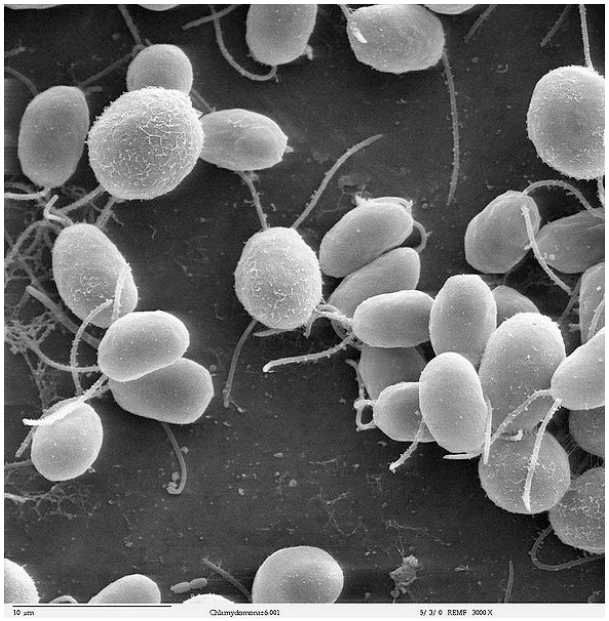
Recombination allows haplotypes to share beneficial mutations

Under asexual evolution, the double mutant could only be created by two mutations. Sex allows for more rapid adaptation

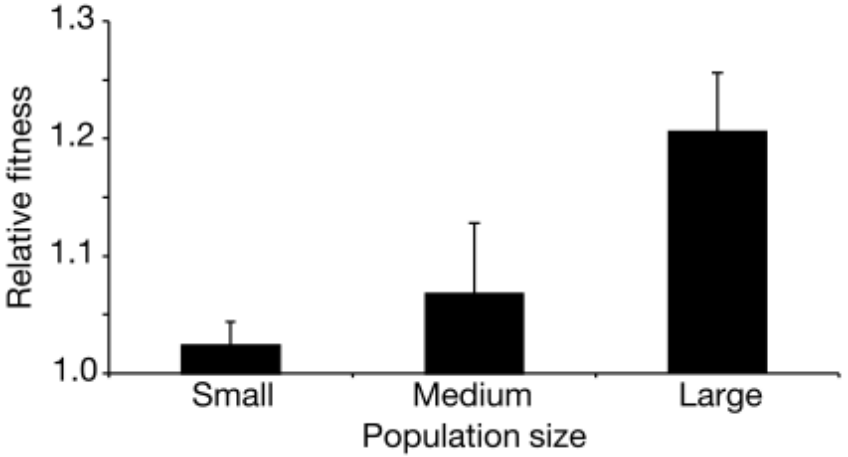
Importance of Recombination

Recombination and adaptation

Experimental evolution of algae *Chlamydomonas reinhardtii*



Difference in fitness between sexual and asexual lines



In larger populations, with greater supply of mutations, sexual reproduction is of greater benefit to the population

Colegrave, Nature, 2002

Pathogenic origin to sexual reproduction?

Red Queen hypothesis

Evolutionary race between pathogens and hosts:

Increased variance of offspring confers an evolutionary advantage

Sexual reproduction increases variance