Lecture 1: Concepts in Population Genetics

Chris Illingworth

Course overview

How can genetic data be used to learn about evolutionary processes?

1-4: Introduction to population genetics: evolutionary forces and dynamics

Evolutionary experiments, selection in multi-locus systems, viral evolution

5-8: Aylwyn Scally, Hannes Svardal, Richard Durbin: Human evolution, Neanderthal admixture, demographic histories of natural populations

Course overview

Assessment

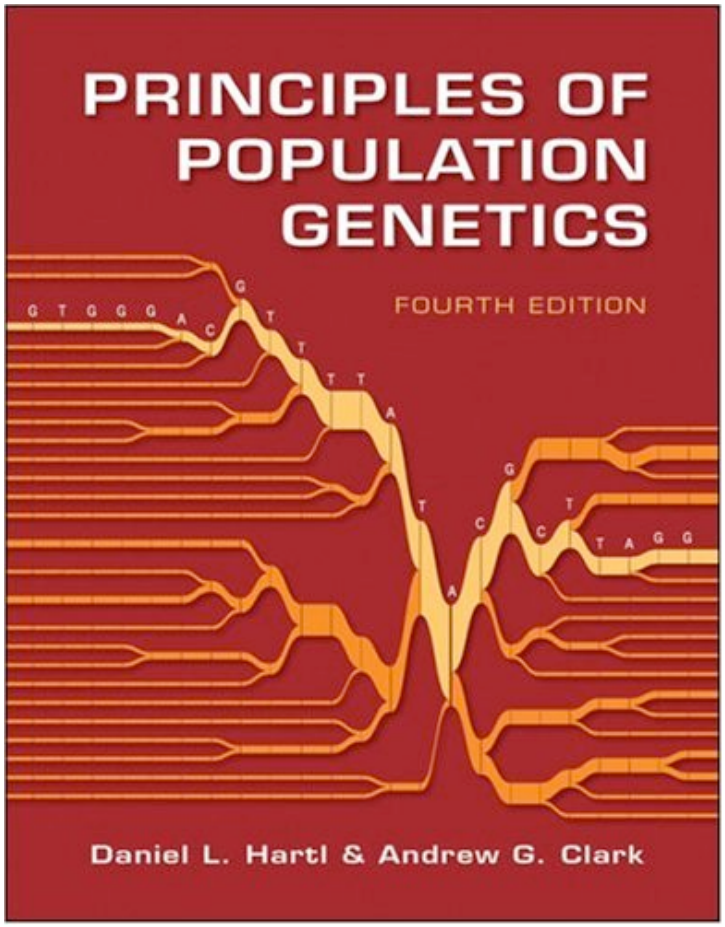
Two example sheets

Please ask if anything is unclear…

cjri2@cam.ac.uk

Book recommendation

Principles of Population Genetics, Hartl and Clark



What is population genetics?

"The study of the distribution of inherited variation among a group of organisms of the same species" – O.E.D.

"distribution" : Statistical measures

"inherited variation" : Includes genetic properties

"Group of organisms" : Usually not the whole species

"of the same species" : Implies inter-breeding

Phenotypic variation

Rift Valley cichlids

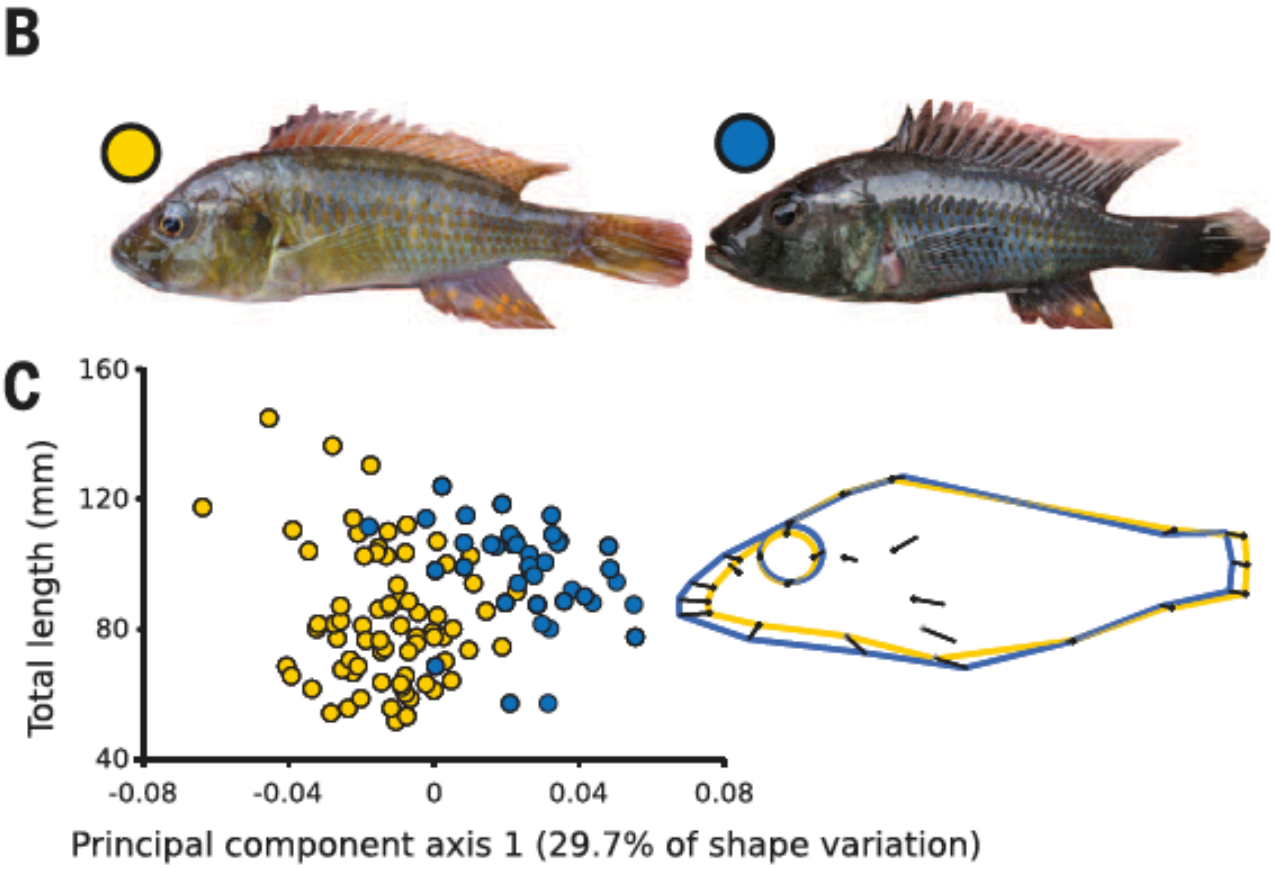
Around 500 species

Phenotypic variation

Cichlid fish

Photo: Ryan Bloomquist

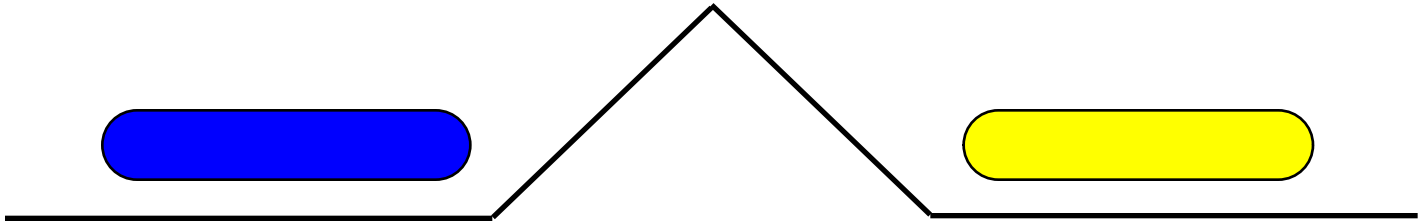
Phenotypic variation



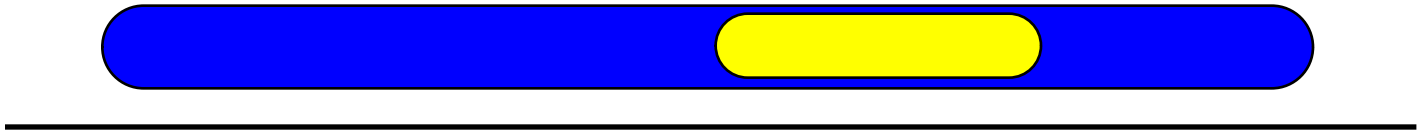
Malinsky et al., Science 2015

Speciation

Allopatric speciation



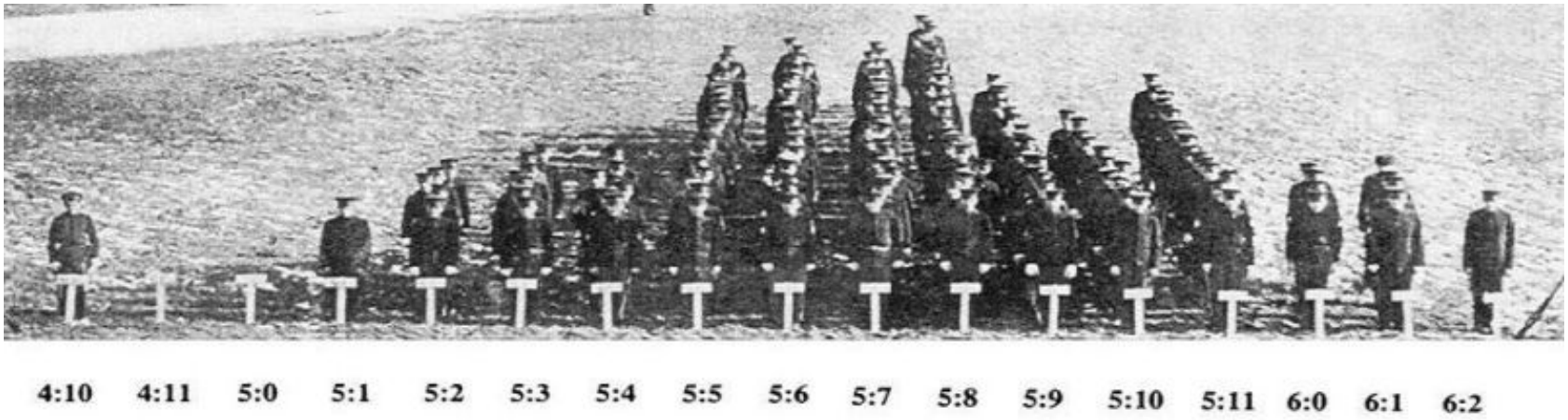
Sympatric speciation

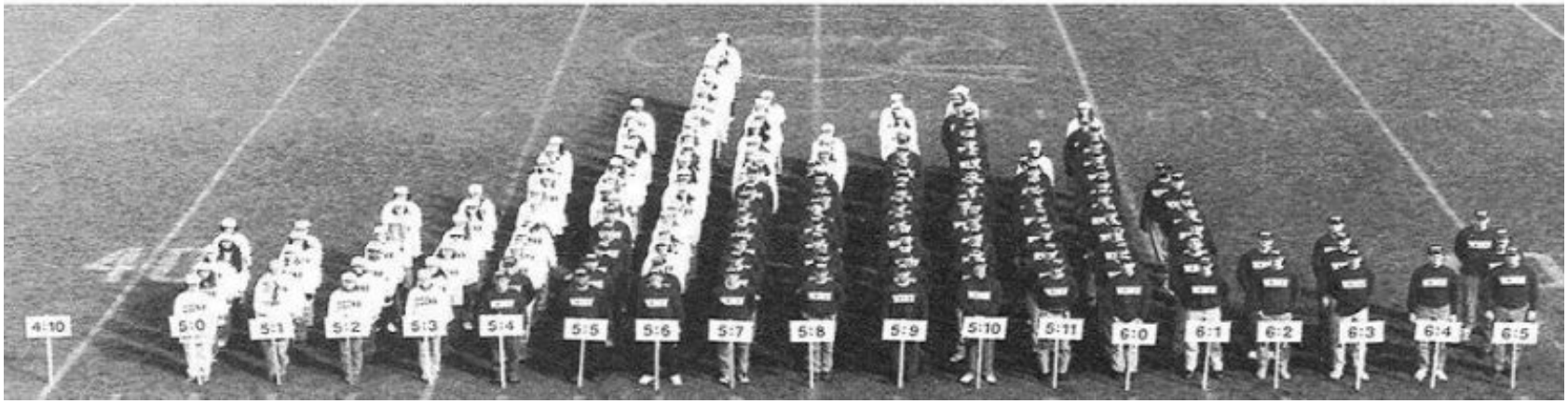


Malinsky et al., Science 2015

Phenotypic variation

Human height





Phenotypic variation

Human height

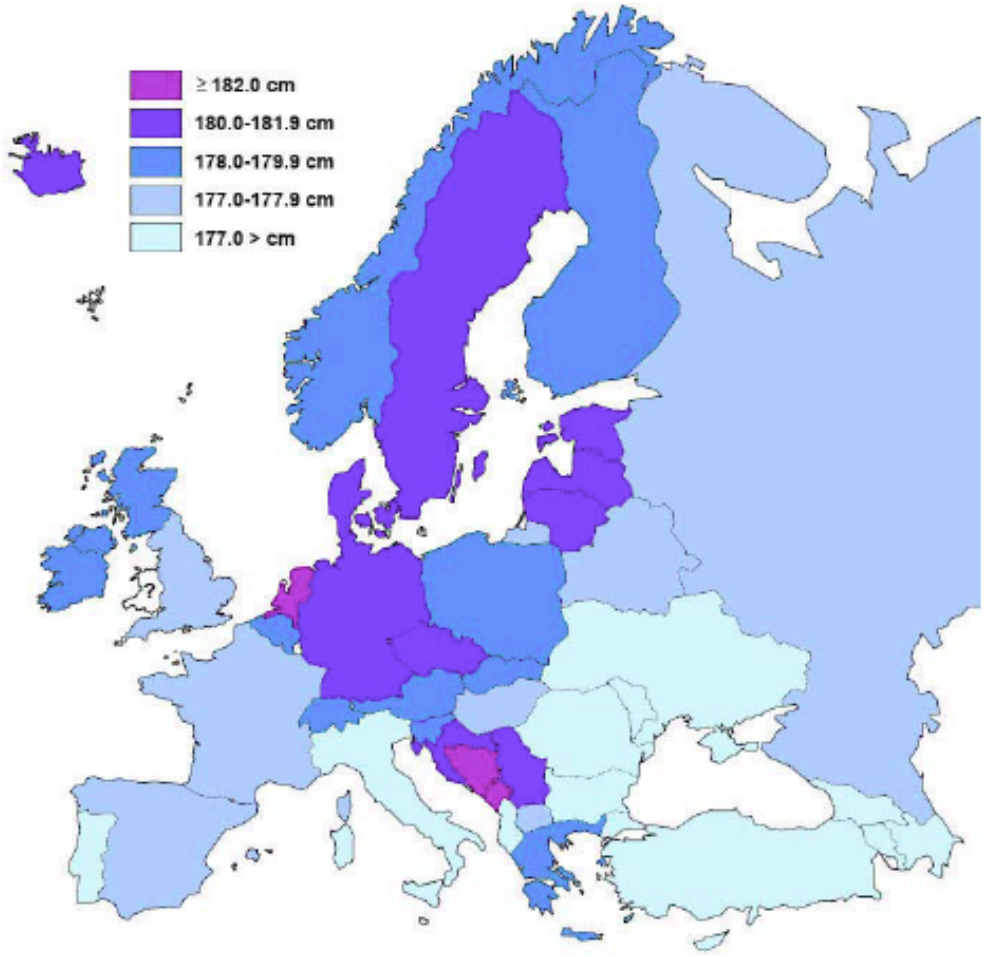
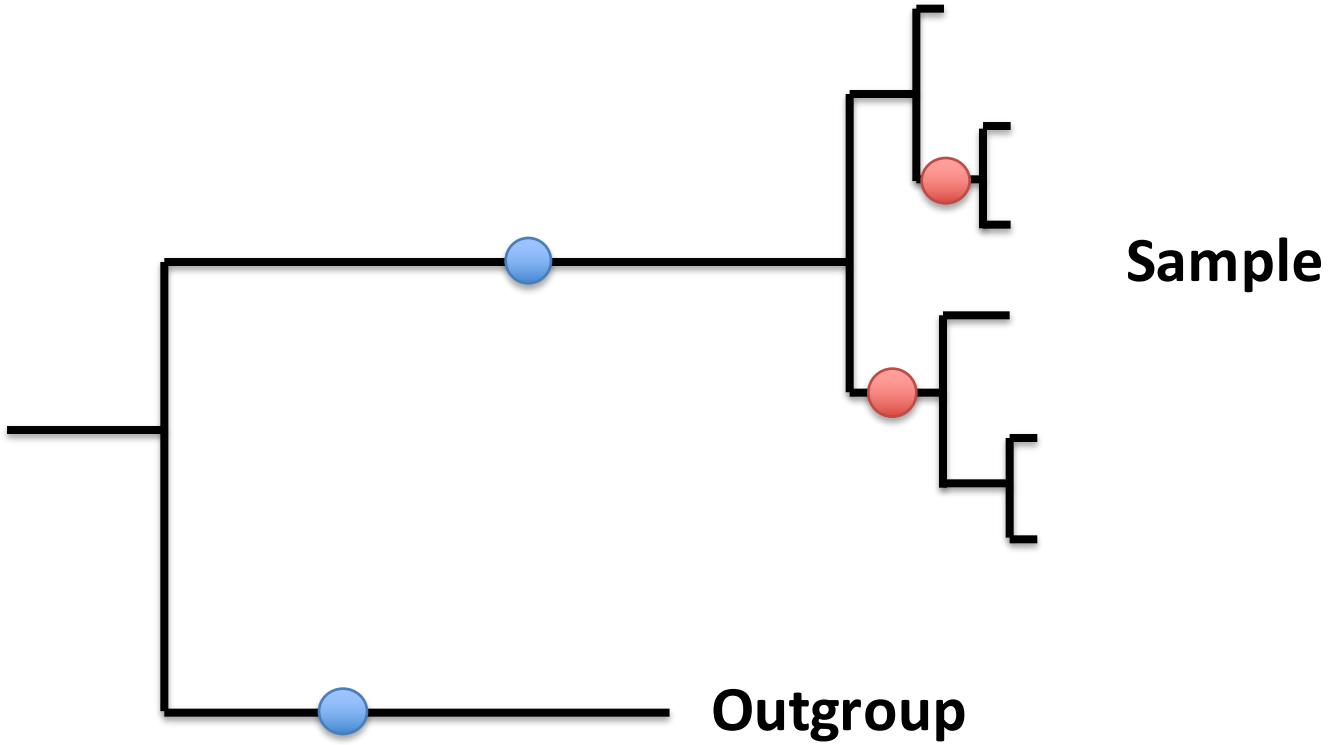


Image: Grasgruber et al., 2013

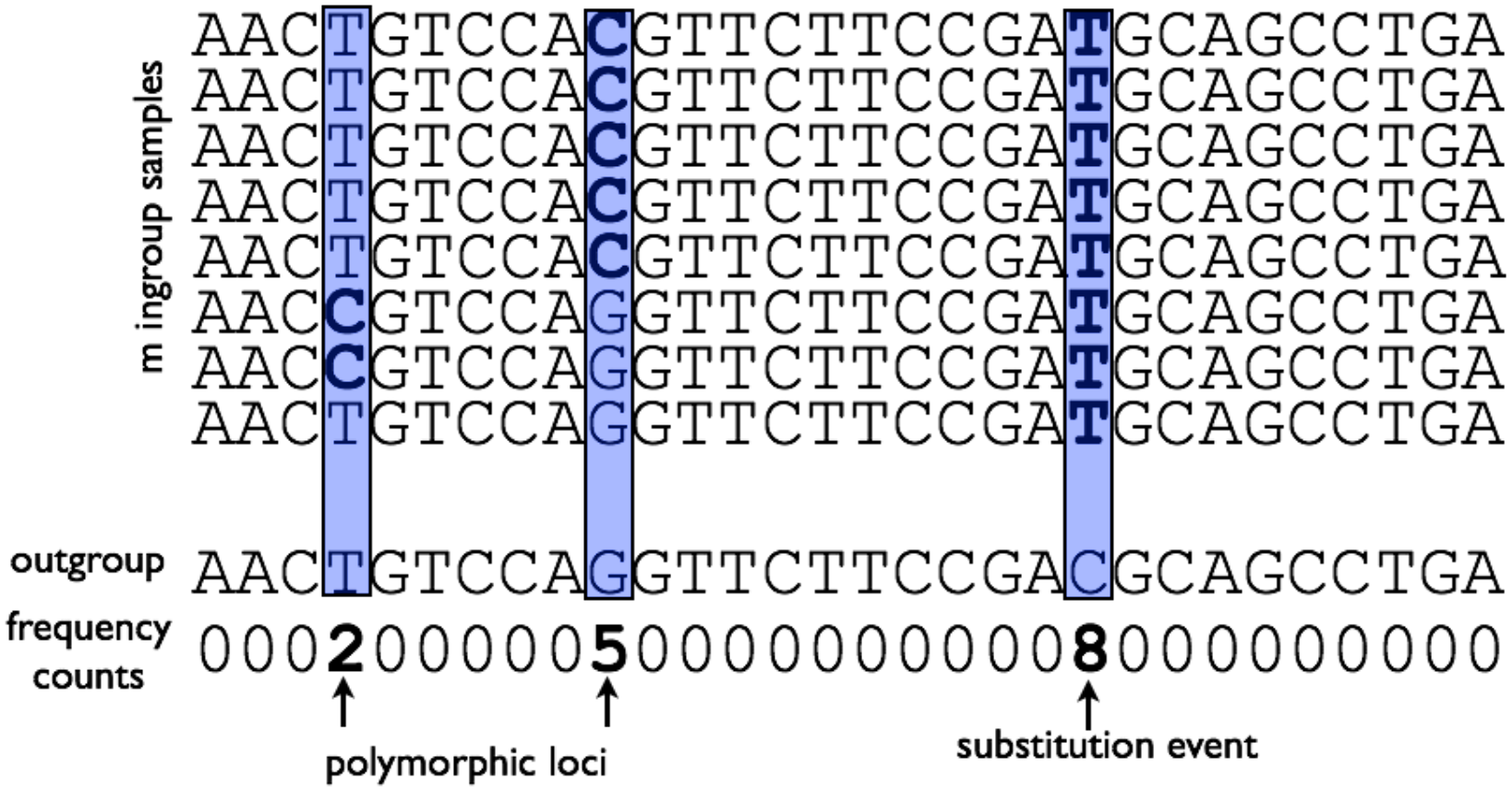
Genetic variation

Cross- and intra-species comparison



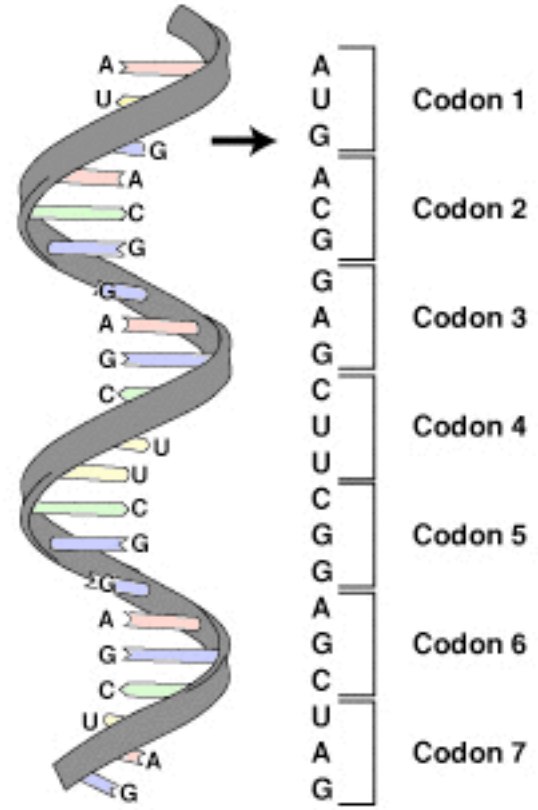
Genetic variation

Cross- and intra-species comparison



Genetic variation

Synonymous and non-synonymous substitutions



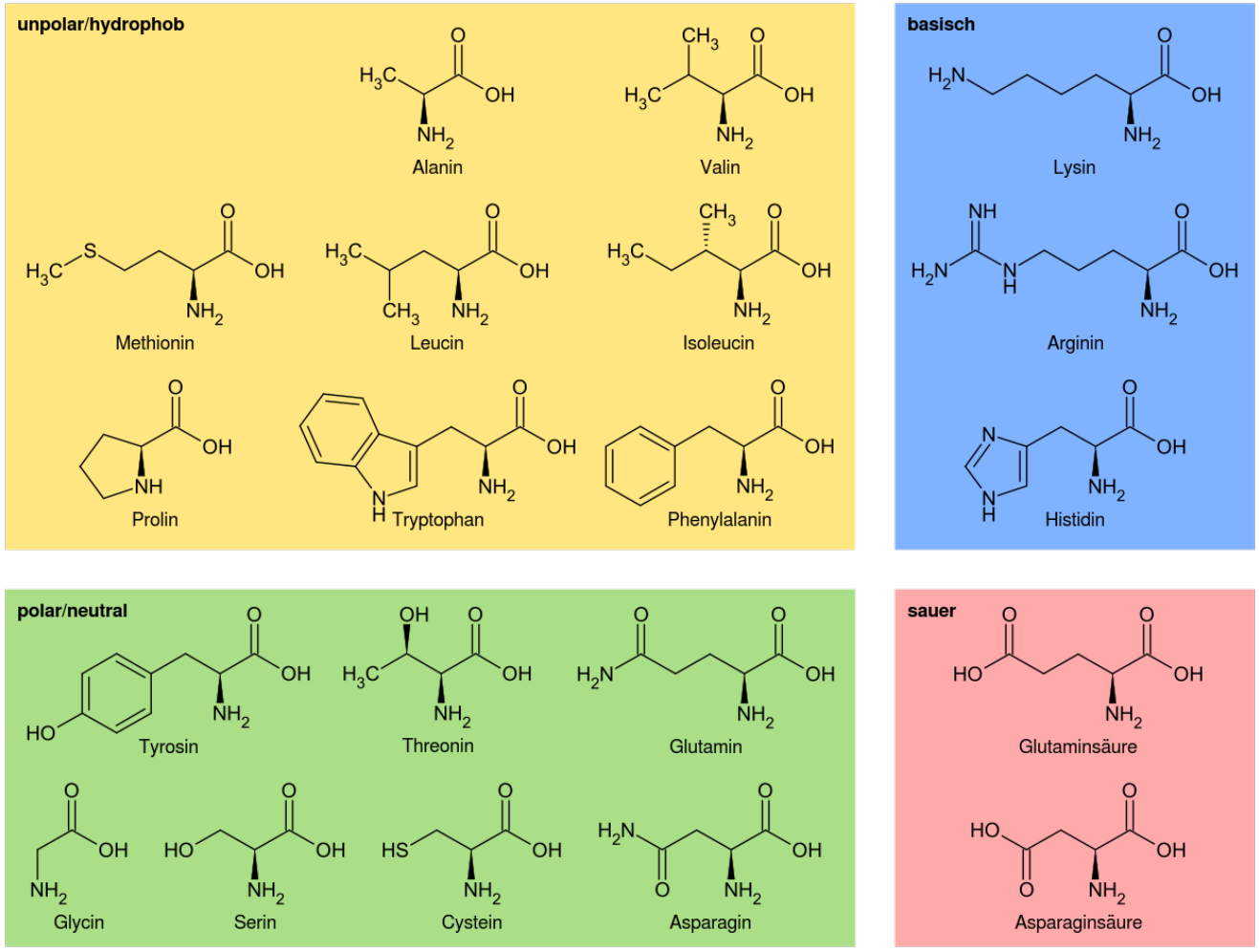
Standard genetic code

1st base	2nd base U			C			A			G			3rd base
U	UUU	(Phe/F) Phenylalanine		UCU	(Ser/S) Serine		UAU	(Tyr/Y) Tyrosine		UGU	(Cys/C) Cysteine		U
	UUC			UCC			UAC			UGC			C
	UUA	(Leu/L) Leucine		UCA			UAA	Stop (*Ochre*)		UGA	Stop (*Opal*)		A
	UUG			UCG			UAG	Stop (*Amber*)		UGG	(Trp/W) Tryptophan		G
C	CUU			CCU	(Pro/P) Proline		CAU	(His/H) Histidine		CGU	(Arg/R) Arginine		U
	CUC			CCC			CAC			CGC			C
	CUA			CCA			CAA	(Gln/Q) Glutamine		CGA			A
	CUG			CCG			CAG			CGG			G
A	AUU	(Ile/I) Isoleucine		ACU	(Thr/T) Threonine		AAU	(Asn/N) Asparagine		AGU	(Ser/S) Serine		U
	AUC			ACC			AAC			AGC			C
	AUA			ACA			AAA	(Lys/K) Lysine		AGA	(Arg/R) Arginine		A
	AUG[A]	(Met/M) Methionine		ACG			AAG			AGG			G
G	GUU	(Val/V) Valine		GCU	(Ala/A) Alanine		GAU	(Asp/D) Aspartic acid		GGU	(Gly/G) Glycine		U
	GUC			GCC			GAC			GGC			C
	GUA			GCA			GAA	(Glu/E) Glutamic acid		GGA			A
	GUG			GCG			GAG			GGG			G

Image: Graham Beards

Genetic variation

Synonymous and non-synonymous substitutions



Genetic variation

Synonymous and non-synonymous substitutions

Synonymous substitution: No change in amino acid

CU**C** to CU**A** : Both encode for Leucine

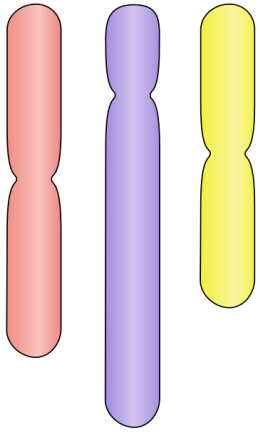
Non-synonymous substitution: Change in amino acid

CUC to **A**UC : Leucine to Isoleucine

Ploidy

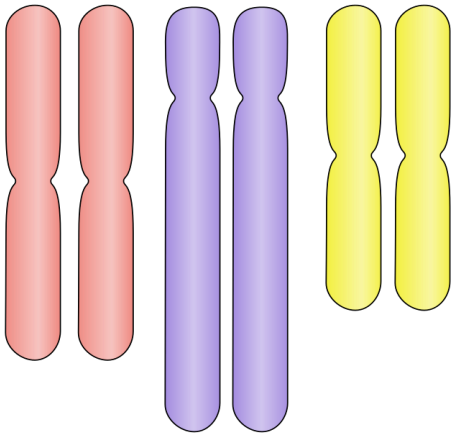
Ploidy: number of copies of each chromosome per cell

Haploid (N)



Bacteria
Plasmodium falciparum
in human stage
Yeast
Influenza virus

Diploid (2N)

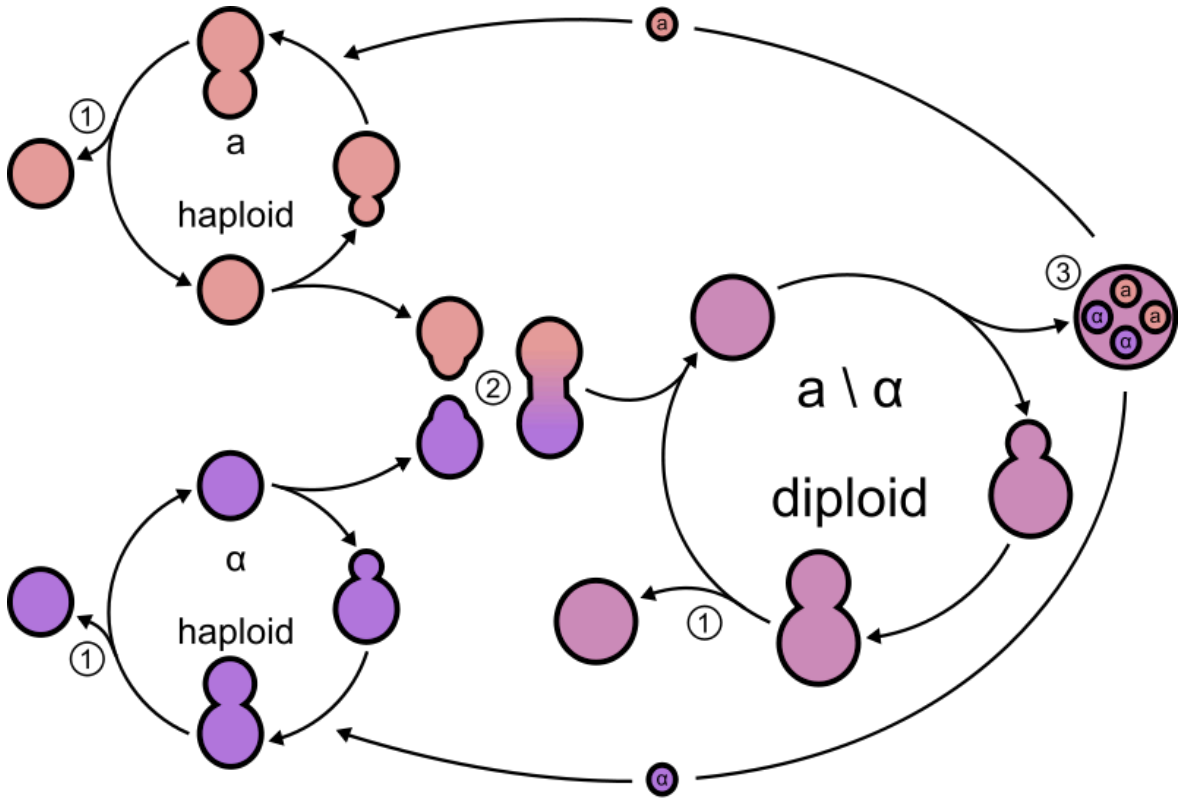


Humans
Yeast
Plasmodium falciparum
in mosquitoes

Image: Ehamburg

Ploidy

Yeast life-cycle



Ploidy

Cancer cells

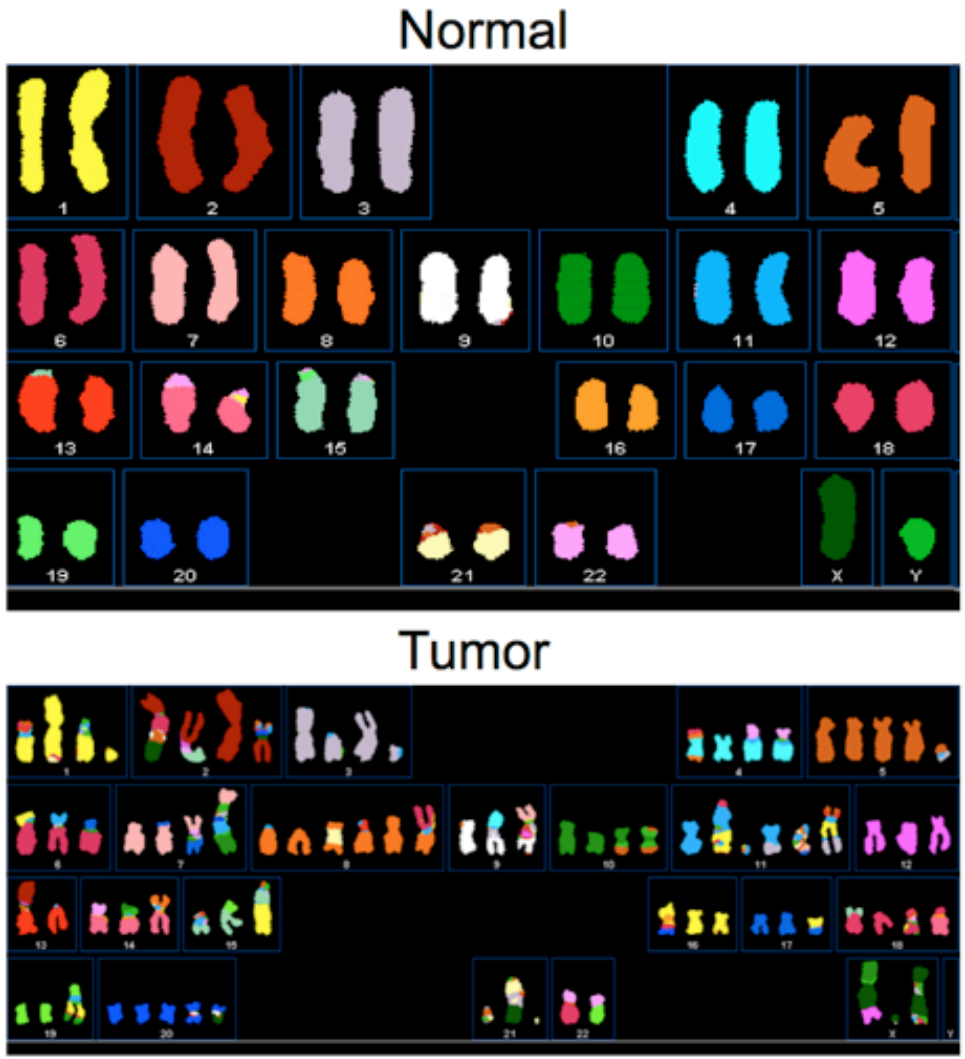
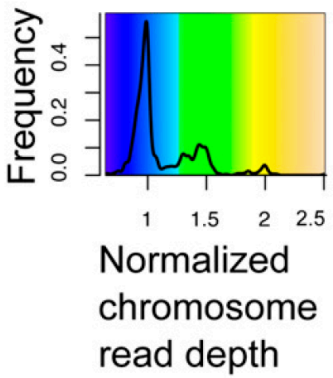
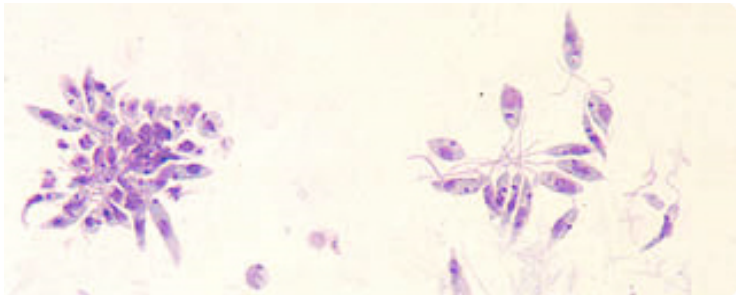


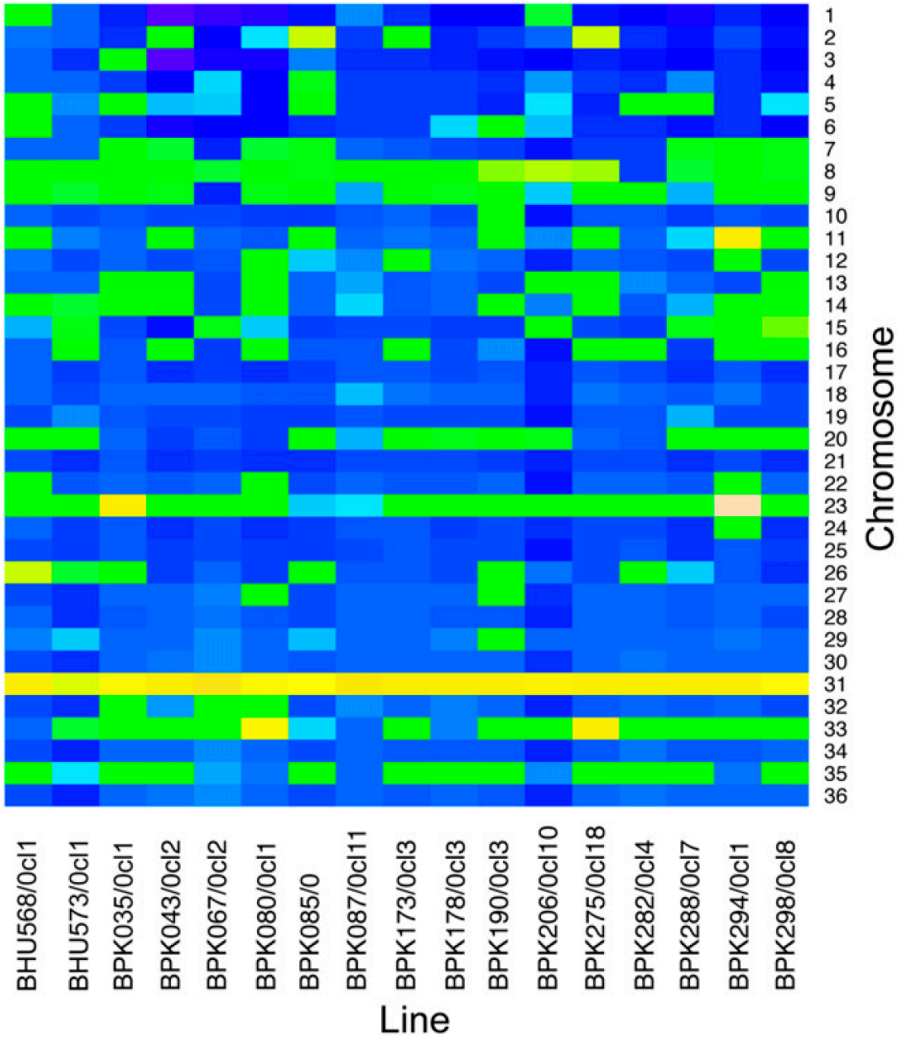
Image: Mira Grigorova and Paul Edwards

Ploidy

Leishmania donovani



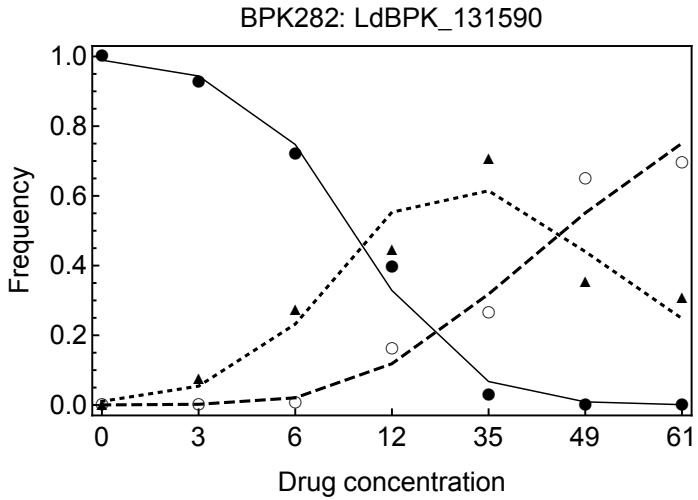


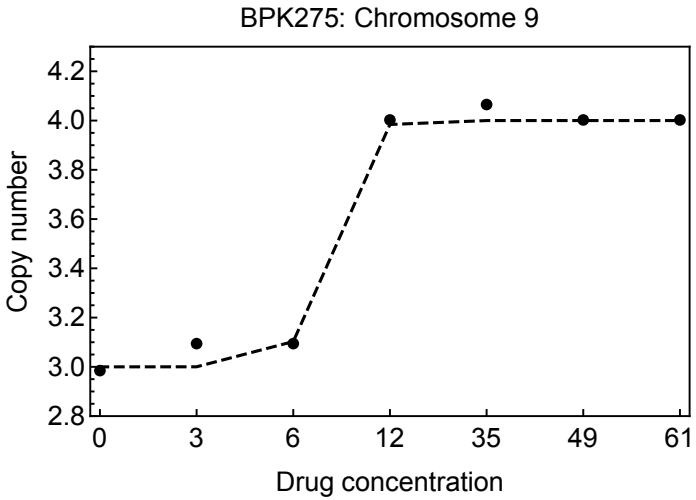


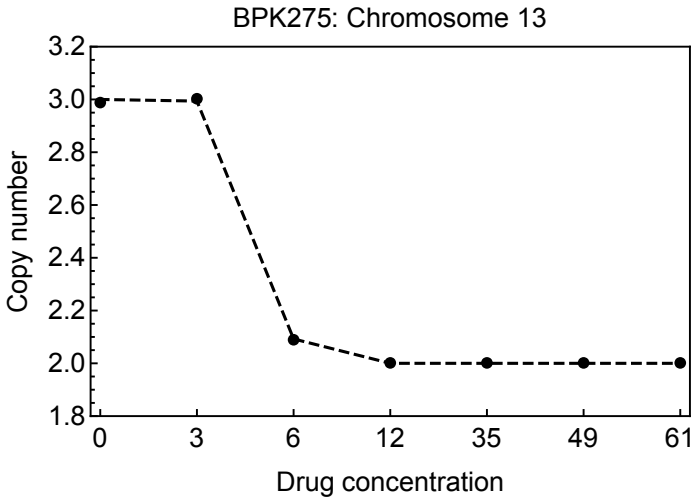
Images: Downing et al., Genome Research 2012; CDC

Ploidy

Adaptation to drug pressure







Absence of evolution: Hardy-Weinberg model

Hardy-Weinberg model

Assume random mating in a population

Diploid locus has alleles A and a, in proportions q_A and q_a. (Let $p = q_A$)
(q=1-p)

Results of mating

			Female parent	
			p	q
			A	a
Male parent	p	A	AA	Aa
	q	a	aA	aa

Diploid allele frequencies

Allele	Frequency
AA	p^2
Aa = aA	2pq
aa	q^2

Hardy-Weinberg model

Hardy-Weinberg model

Frequencies are maintained over time under random mating

			Female parent		
			p^2	2pq	q^2
			AA	Aa	aa
Male parent	q^2	AA			
	2pq	Aa			
	q^2	aa			

Frequency of AA in next generation is

$$p^4+2p^3q+p^2q^2$$

$$= p^2(p+q)^2$$

$$= p^2$$

Hardy-Weinberg model and F_{ST}

Hardy-Weinberg model

Other circumstances give different equilibrium frequencies

			Female parent		
			p^2	2pq	q^2
			AA	Aa	aa
Male parent	p^2	AA	AA	½AA ½Aa	Aa
	2pq	Aa	½AA ½Aa	¼AA ½Aa ¼aa	½Aa ½aa
	q^2	aa	Aa	½Aa ½aa	aa

Selection

Haplotypes may have different fitnesses

f_{AA}, f_{Aa}, f_{aa}

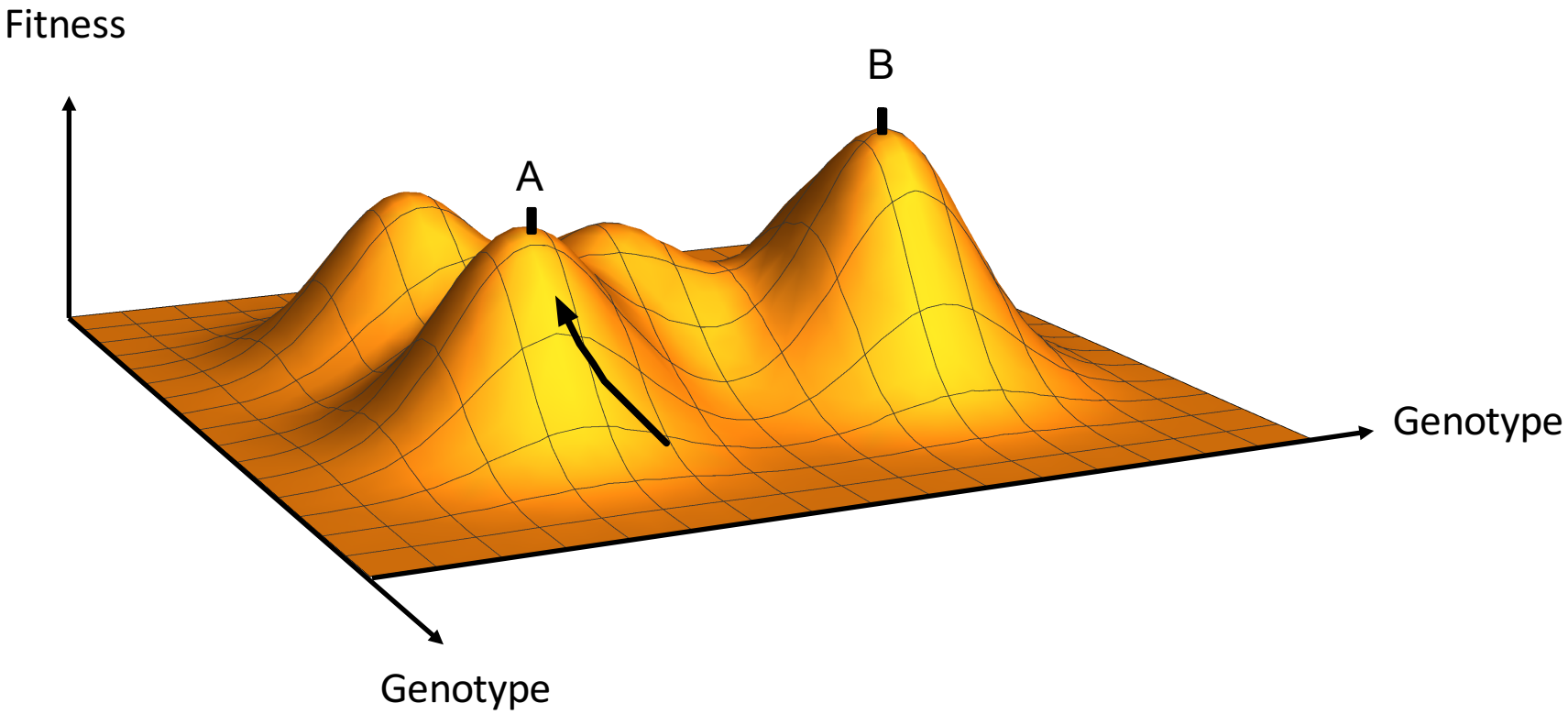
Stable equilibria occur at population fitness maxima

Mean fitness of population:

$$q_{AA}f_{AA} + q_{Aa}f_{Aa} + q_{aa}f_{aa}$$

Hardy-Weinberg model and F_{ST}

Adaptation increases the mean population fitness



Hardy-Weinberg model

Hardy-Weinberg model

Other circumstances give different equilibrium frequencies

			Female parent		
			p^2	2pq	q^2
			AA	Aa	aa
Male parent	p^2	AA	AA	½AA ½Aa	Aa
	2pq	Aa	½AA ½Aa	¼AA ½Aa ¼aa	½Aa ½aa
	q^2	aa	Aa	½Aa ½aa	aa

Inbreeding

Greater probability of mating between identical genotypes

Decrease in frequency of Aa allele

Heterozygosity

$$H = \frac{\#Aa}{\#AA + \#Aa + \#aa}$$

is reduced

Deviation from Hardy-Weinberg: Inbreeding

Related parents share more identical alleles

Expected fraction of heterozygous loci with no inbreeding 2pq

Let F be the probability that two parents are identical by descent at a given locus.

Then the probability that their offspring are heterozygous at that locus is (1-F) * 2pq

Measuring the level of inbreeding

Observed fraction of heterozygous loci q_{Aa}.

Estimate of F: $$F = 1 - \frac{q_{Aa}}{2pq}$$

			Female parent	
			p	q
			A	a
Male parent	p	A	AA	Aa
	q	a	aA	aa

Measurement of the extent of inbreeding

Level of inbreeding depends on the reference population

Comparison with regional, national, global populations

F-statistics

F-statistics describe the relationships between populations

F_{XY}: The correlation between random gametes, drawn from the same X, relative to Y

I: In an individual

S: In a subpopulation

T: In the total population

Often calculate F_{IS}, F_{IT}, and F_{ST}

F-statistics

F-statistics describe the relationships between populations

F_{XY}: The correlation between random gametes, drawn from the same X, relative to Y

I: In an individual

S: In a subpopulation

T: In the total population

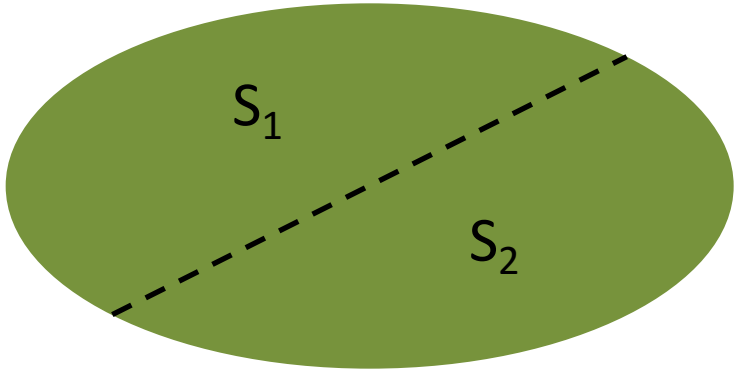
$$F_{IS} = 1 - \frac{q_{Aa}}{2p_S q_S} \qquad F_{IT} = 1 - \frac{q_{Aa}}{2p_T q_T}$$

F-statistics

F_{ST} compares sub-populations to the whole

$$F_{ST} = 1 - \frac{2p_S q_S}{2p_T q_T}$$

Total population T



Sub-populations S_1, S_2

F_{ST} = 0 : no difference in heterozygosity between populations

F_{ST} = 1 : complete segregation between subpopulations: alleles in each subpopulation are fixed

$$p_S q_S = \frac{N_{S1} p_{S1} q_{S1} + N_{S2} p_{S2} q_{S2}}{N_{S1} + N_{S2}}$$

F-statistics

Example: Measuring selection in the wild

What effect does parasite infection have upon *Daphnia magna*?

Wild populations of *Daphnia* were infected or not infected with the *Octosporea bayeri* parasite.

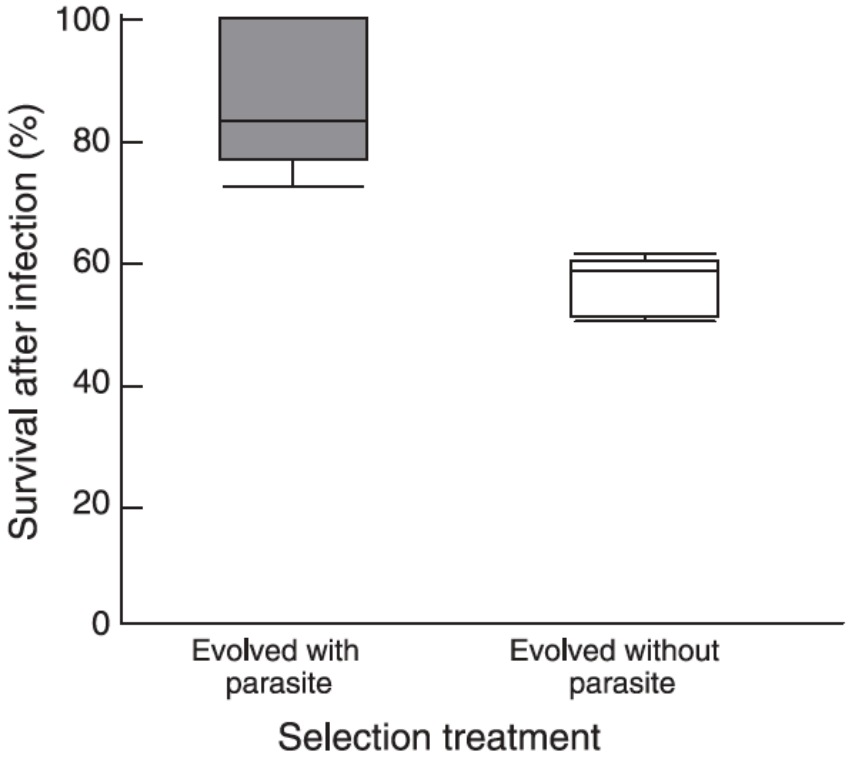
After 15 generations, the probability of individuals in each population surviving infection was measured.

Photo from Brewer et al, PLoS Genetics, 2011

F-statistics

Example: Measuring selection in the wild

Parasite infection triggers an evolutionary response



Daphnia which evolved in the presence of the parasite were more resistant to infection

Zbinden et al, J Evol Biol, 2008

Evolutionary experiments

Genetic differences identified between populations

Measure F_{IS}: not significantly different between populations

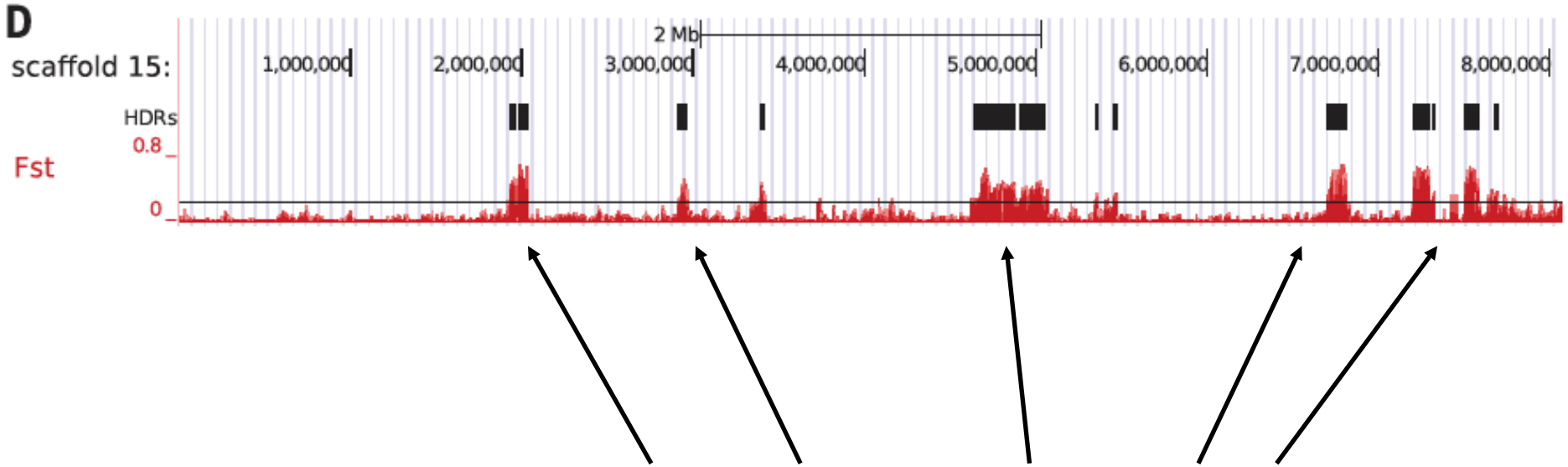
Locus	Frequency of allele 1 in control populations	Frequency of allele 1 in infected populations	F_{ST}	*P*-value
Aat	0.166	0.292	0.045	< 0.0001
Fum	0.538	0.466	0.009	0.0009
Gpi	0.335	0.415	0.013	0.0003

Aat : Aspartate amino transferase
Fum: Fumarase
Gpi : Glucose phosphate isomerase

F_{ST} identifies the presence of selection

Zbinden et al, J Evol Biol, 2008

F_{st} in cichlid populations



Differentiation between species at a few regions in the genome

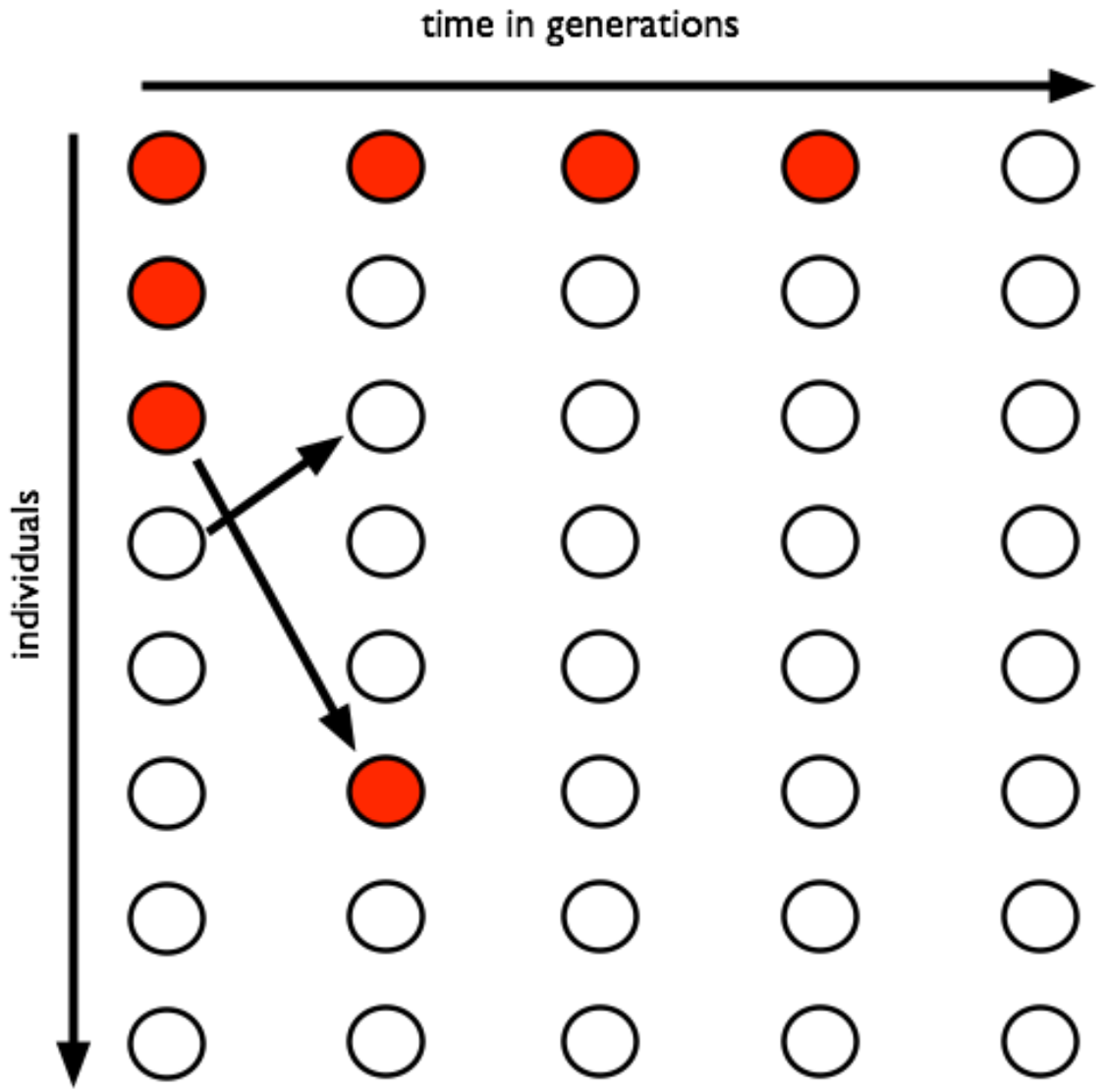
Population dynamics

How a population changes over time

Three key forces: Genetic drift, Selection, Mutation

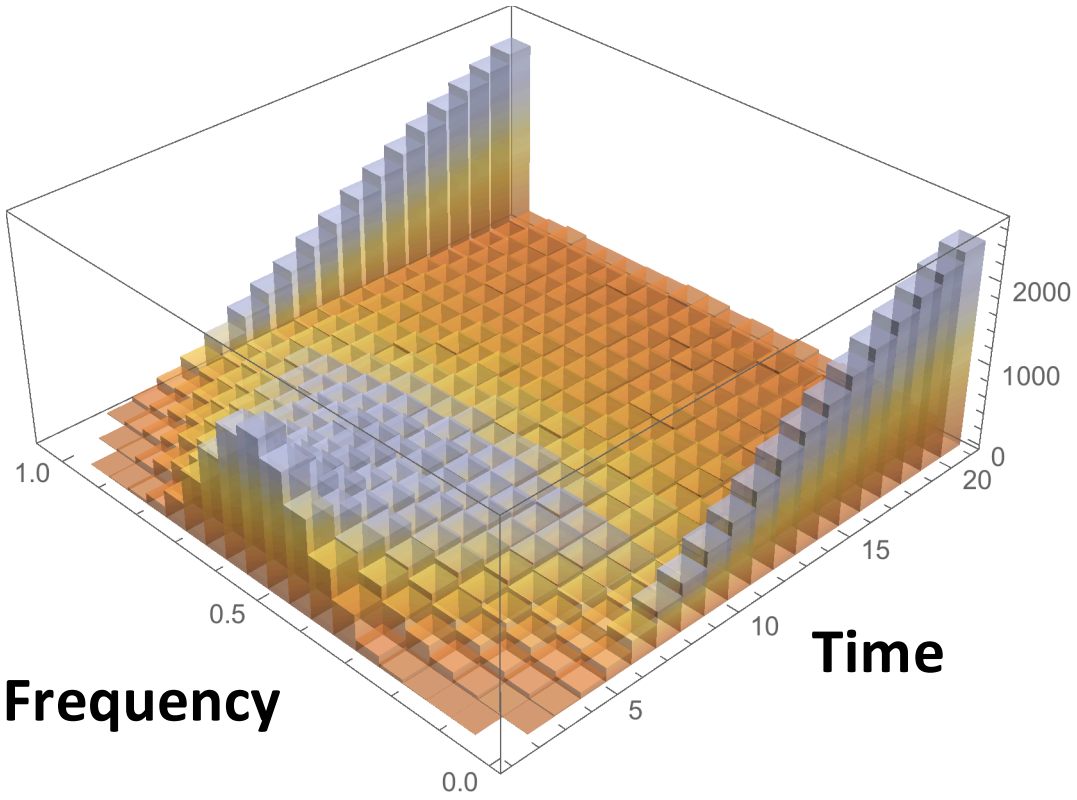
Here will assume a haploid population

Genetic drift: noise in reproduction



Genetic drift: noise in reproduction

Distribution of allele frequency over time



Wright-Fisher process

Repeated binomial sampling

Allele frequency

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

Propagation

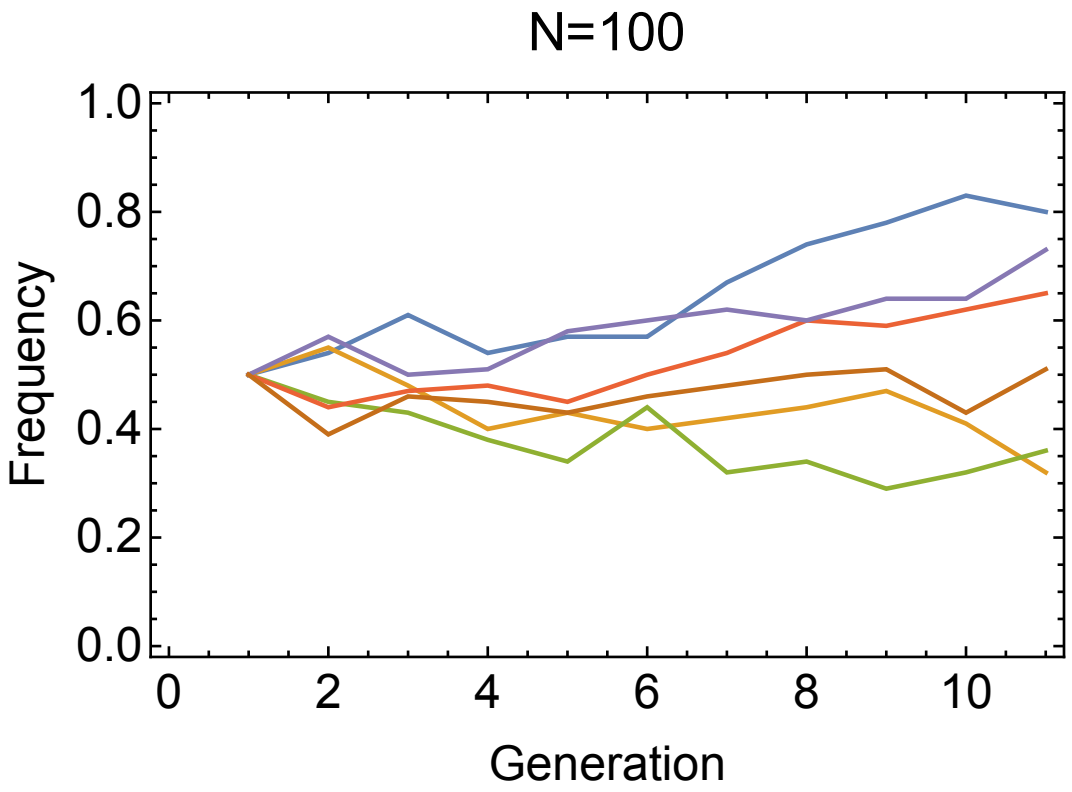
$$P(m, N, t+1) = \binom{N}{m} q_A(t)^m (1 - q_A(t))^{N-m}$$

Mean and variance

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

Wright-Fisher process

Rate of drift depends upon population size



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

Selection

Different measures for fitness

Malthusian fitness: c. f. Growth rate

Darwinian fitness: c.f. Relative probability of reproductive success

Selection

Differential reproductive success of individuals

Fitnesses of a and A alleles: w_a, w_A

Frequencies of a and A alleles: q_a, q_A

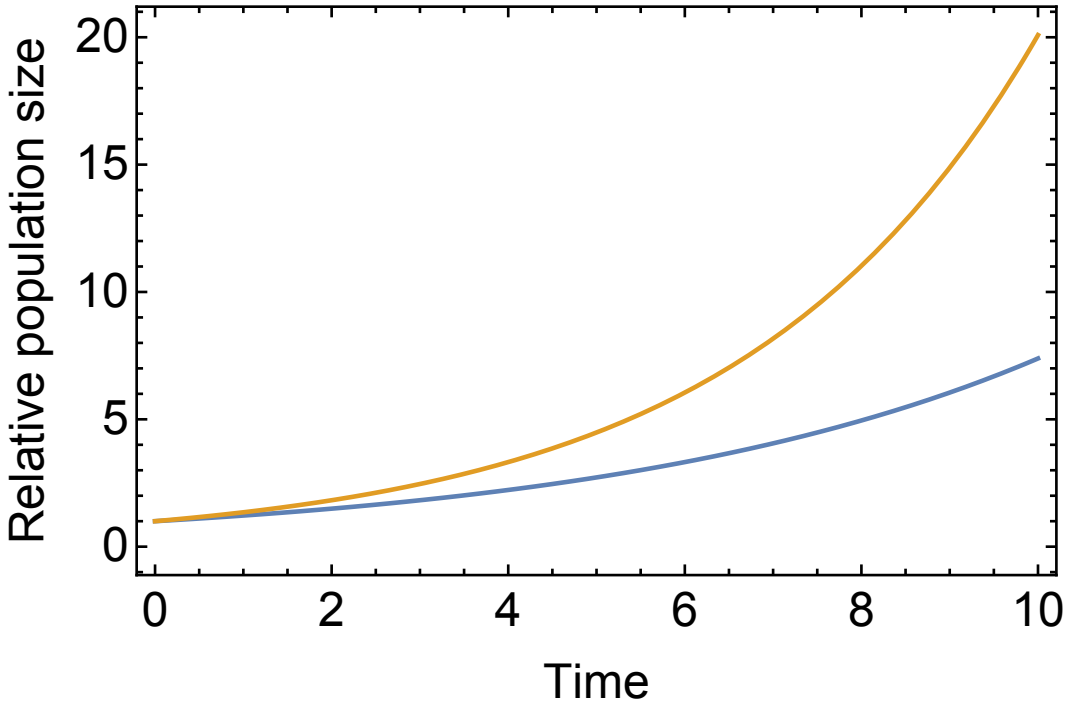
Mean fitness of the population $\bar{w}$ = $w_a q_a + w_A q_A$

Probability of allele a in the next generation $\frac{w_a q_a}{\bar{w}}$

Common notation: $w_a = 1$, $w_A = 1+s$

Selection

Differential reproductive success of individuals



$$\dot{N}_a = F_a N_a$$

$$\dot{N}_A = F_A N_A$$

Model of exponential growth in numbers

Selection

Assume competition in a constant population size

$$\dot{N}_a = F_a N_a \qquad \dot{N}_A = F_A N_A$$

$$q_a = N_a/N$$

$$q_A = N_A/N$$

$$\dot{q}_A = \frac{d}{dt}\frac{N_A(t)}{N_a(t)+N_A(t)}$$

$$= \frac{\dot{N}_A(t)}{N_a(t)+N_A(t)} - \frac{N_A(t)}{N_a(t)+N_A(t)}\frac{\dot{N}_a(t)+\dot{N}_A(t)}{N_a(t)+N_A(t)}$$

$$= \frac{F_A N_A(t)}{N} - \frac{N_A(t)}{N}\left(\frac{F_a N_a(t)+F_A N_A(t)}{N}\right)$$

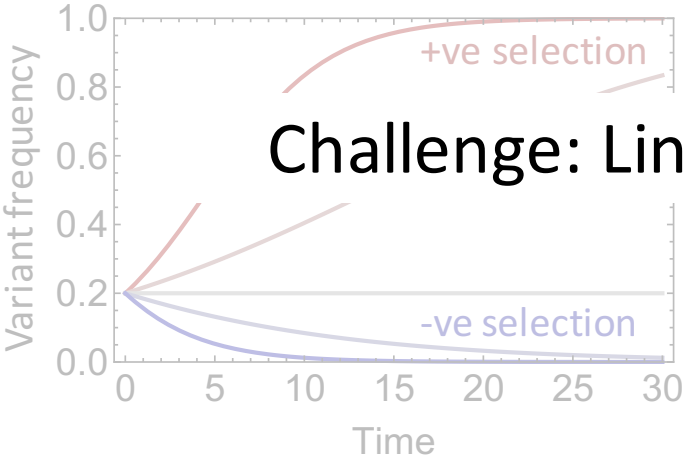
$$= (F_A - F_a)q_A(t)(1-q_A(t))$$

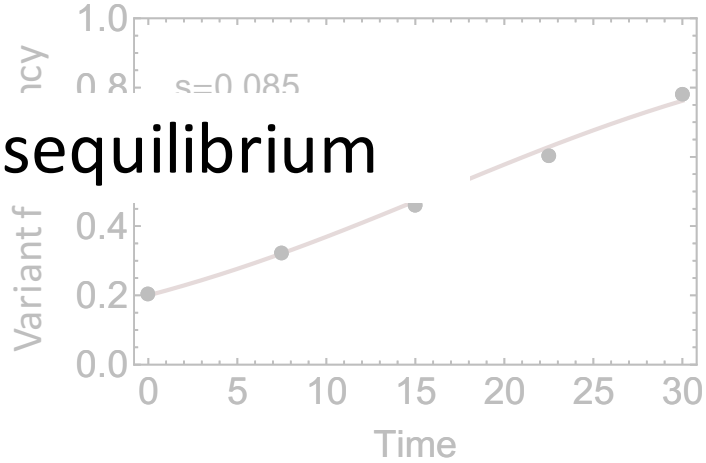
Can write $\sigma = (F_A - F_a)$

Population genetic theory

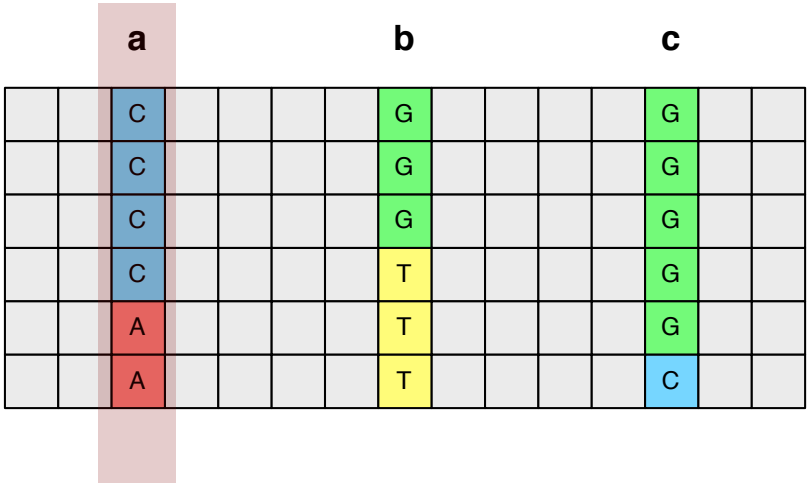
Selection changes the frequency of alleles over time

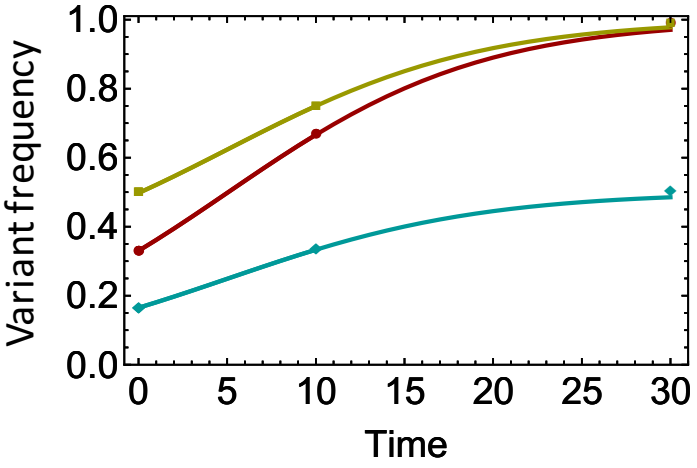
Genetic drift changes the frequency of alleles over time





Challenge: Linkage disequilibrium





Selection

Different measures for fitness

Malthusian fitness: In the above, denoted by F_A

Used with *continuous-time* models:

Exponential growth of individuals according to fitness

Darwinian fitness: Earlier, denoted by f_{AA}

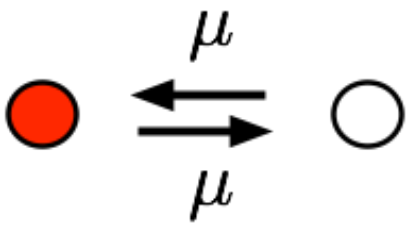
Used with *discrete-time* models:

Mean fitness is $\bar{f} = f_{AA}q_{AA} + f_{Aa}q_{Aa} + f_{aa}q_{aa}$

Mean frequency of AA next generation is $\frac{f_{AA}q_{AA}}{\bar{f}}$

Mutation : Source of variation

Changes the allele at a specific locus



$$\dot{N}_A = \mu N_a - \mu N_A$$

$$\dot{N}_a = \mu N_A - \mu N_a$$

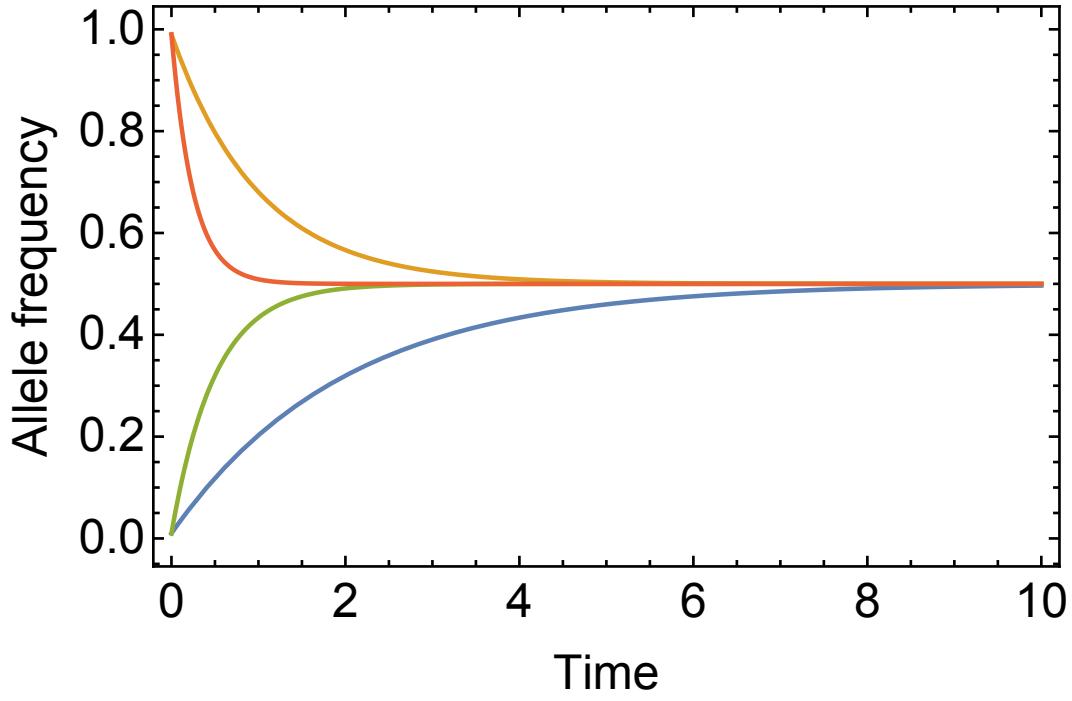
$$\dot{q}_A = \frac{\dot{N}_A(t)}{N_a(t) + N_A(t)} - \frac{N_A(t)}{N_a(t) + N_A(t)} \frac{\dot{N}_a(t) + \dot{N}_A(t)}{N_a(t) + N_A(t)}$$

$$= \mu(1 - 2q_A(t))$$

Mutation

Equation of allele frequency change under mutation

$$\dot{q}_A = \mu(1 - 2q_A(t)) \qquad q_A(t) = \frac{1}{2}(1 + e^{-2\mu t}(2q_A(0) - 1))$$

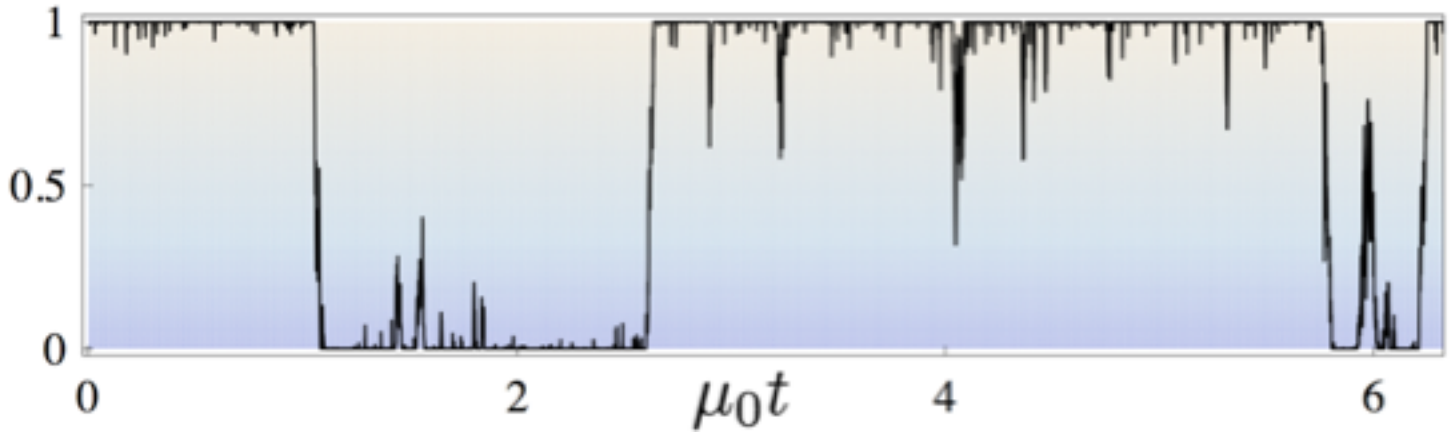


Rate of change proportional to μ

Population dynamics

How a population changes over time

Slow mutation: Fixation dynamics



Drift, selection, and mutation

Equation of allele frequency change : Langevin equation

$$\dot{q}_A(t) = \sigma[q_A(t)(1 - q_A(t))] + \mu[1 - 2q_A(t)] + \chi_q(t)$$

$$\langle \chi_q(t) \rangle = 0$$

$$\langle \chi_q(t)\chi_q(t') \rangle = \frac{q_A(1 - q_A)}{N}\delta(t - t')$$

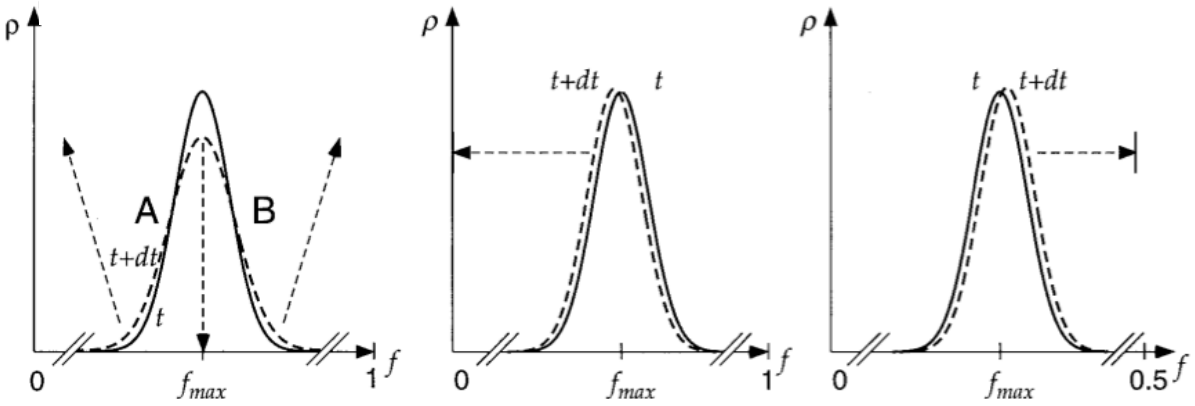
Drift, selection, and mutation

Kimura's diffusion equation

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

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

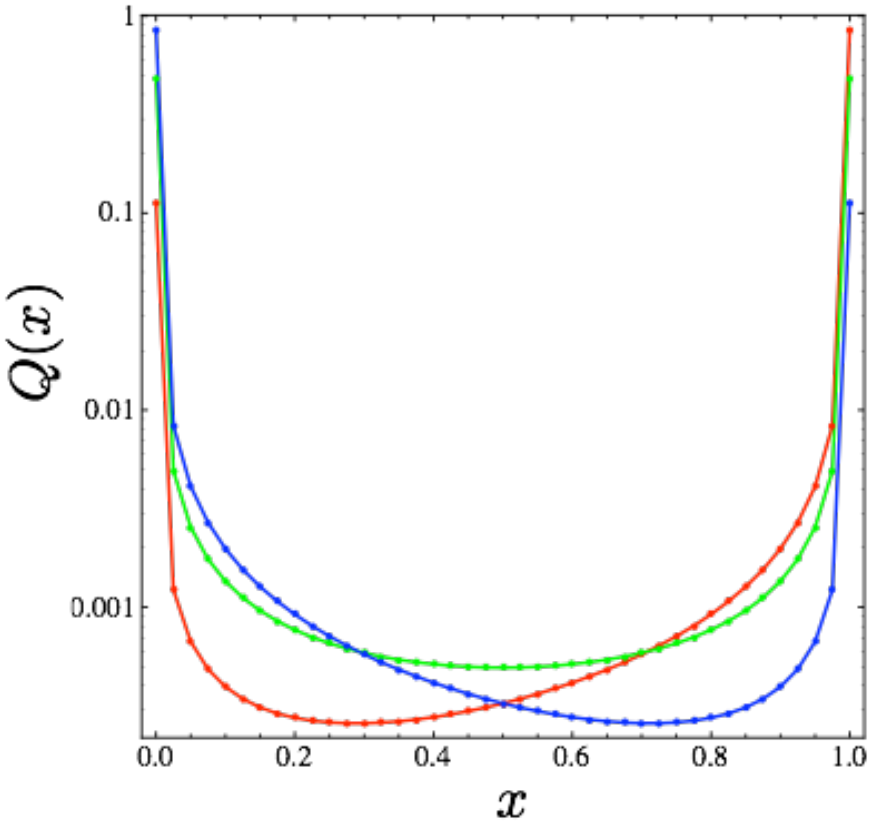
Used to calculate statistical properties of the system



Drift, selection, and mutation

Kimura's diffusion equation

Equilibrium allele frequency distribution



σ = 0

σ = 3

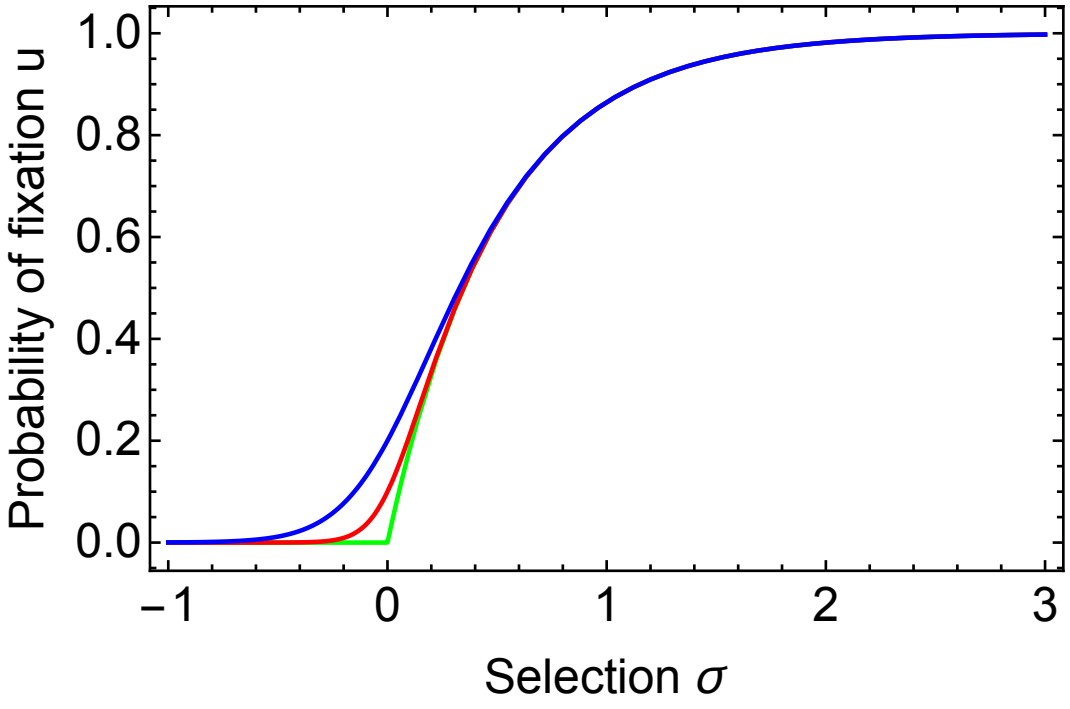
σ = -3

Drift, selection, and mutation

Probability of a new mutant fixing in a population

Suppose there are N **diploid** individuals in the population

$$u(N,\sigma) = \frac{1 - e^{-2\sigma}}{1 - e^{-4N\sigma}}$$



N = 1000

N = 10

N = 5

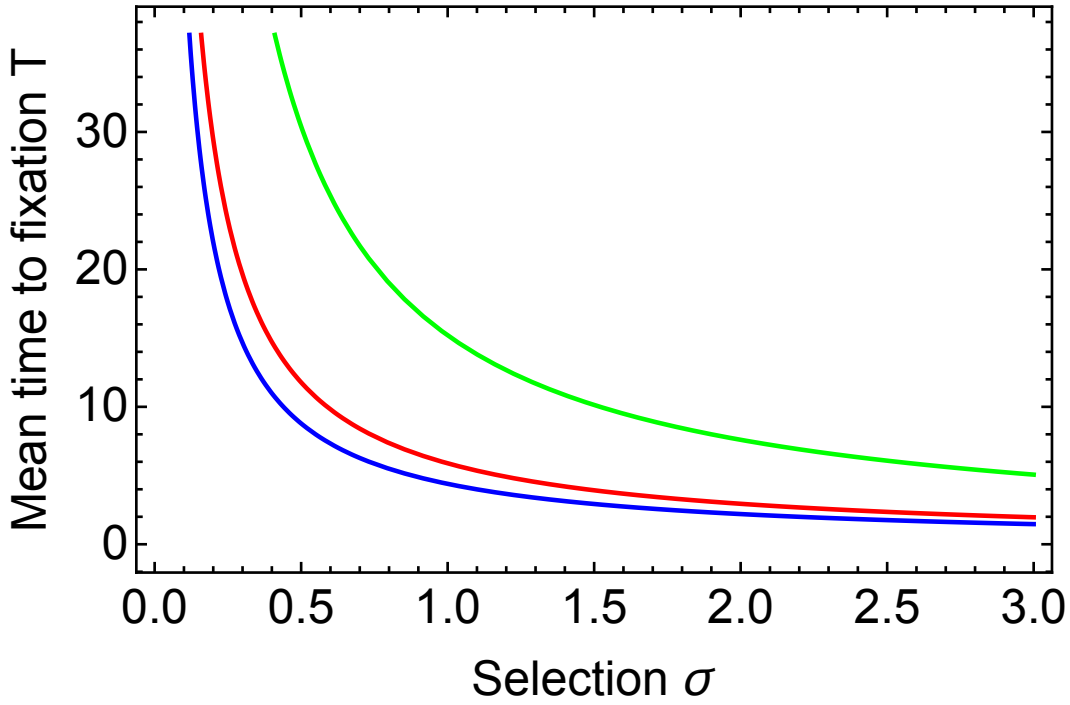
Drift, selection, and mutation

Mean time for a new mutant to fix in a population

(conditional upon fixation)

Suppose there are N **diploid** individuals in the population

$$T(N,\sigma) \approx \frac{2\ln(2N-1)}{\sigma}$$



N = 1000

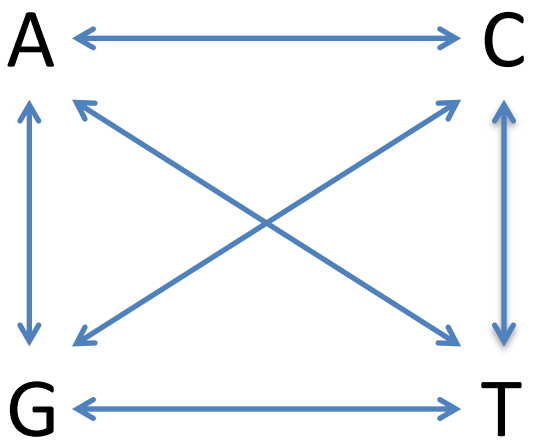
N = 10

N = 5

Processes of mutation

In a DNA sequence mutations occur with different frequencies

Different mutations caused by different chemical processes



e.g. UVB light induces C to T mutations

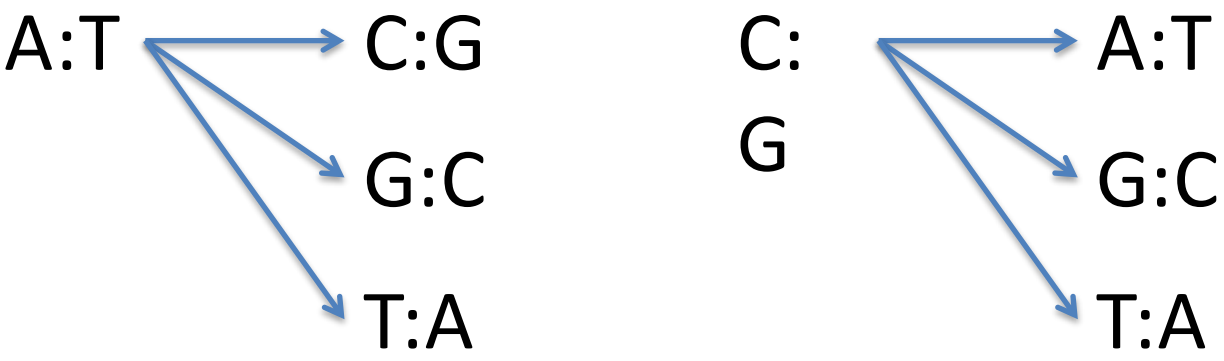
Question: What processes can be observed in cancer cells?

A. Fischer et al., Genome Biology, 2013 14:R39

Processes of mutation

Framework for mutational process inference

Potential mutational processes: Paired nucleotides



Six mutations

Allow for context: 96 mutations

aA:T**c** → **a**C:G**c**

Processes of mutation

Mutational processes

Each mutational process produces different mutations at different rates

μ_{aj} Probability that process *a* produces a mutation of type *j*

$$\sum_j \mu_{aj} = 1 \qquad 1 \leq j \leq 96$$

Activity

In any example, each process acts with some level of activity x_a

Processes of mutation

Opportunity

Each process acts upon a string of DNA with a non-uniform distribution of nucleotides

e.g. ATCGCGATATGGATGCATGTAGTCGATGTACGGATG
has a G in ⅓ of positions

ω_j^m Proportion of sites allowing for a mutation of type *j* in sample *m*

Total mutational output of process *a* in sample *m*

$$x_a \mu_{aj} \omega_j^m$$

Processes of mutation

Observed data

Tumour-normal matched genome samples: identify mutations

X_j^m Number of mutations of type *j* in sample *m*

Assumption: The same process act in different proportions upon each of the samples

Learn the processes, and the extent to which they acted in each sample

Processes of mutation

Expectation-maximisation algorithm

0. Guess some initial mutational processes μ_j

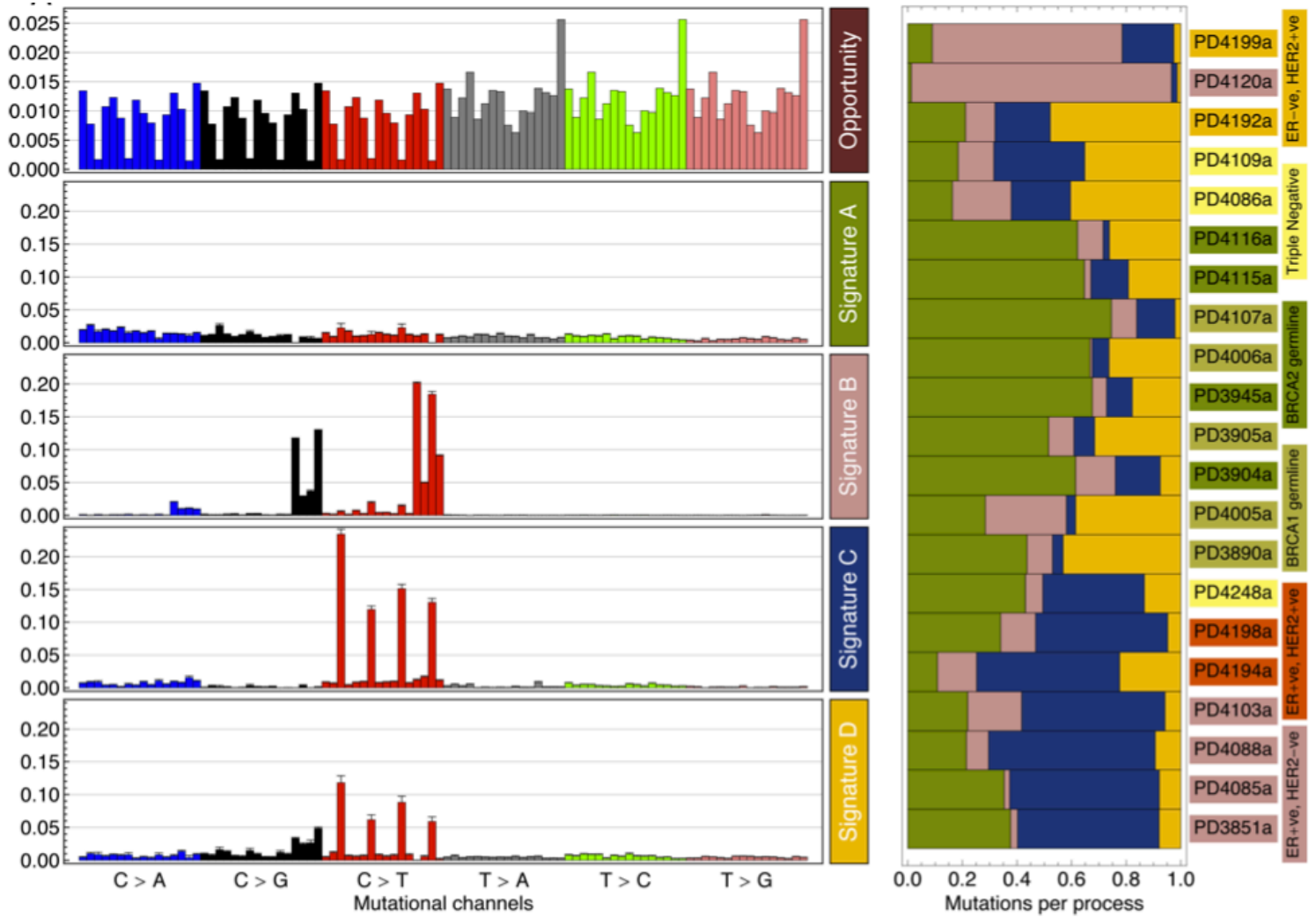
1. E-step: Given the estimated mutational processes, find the maximum likelihood activities x_a

2. M-step: Given the estimated activities, find the maximum likelihood mutational processes μ_j

3. Repeat until convergence

Processes of mutation

Results for breast cancer sample data



Processes of mutation

Localise processes in the genome

Given the processes identified, look for local patterns of mutation

For a specific tumour m, have global activities $x_a^{m,g}$

Use the global activities as a prior for learning local activities

Aside: Bayesian inference

Derive a posterior probability given a prior and some evidence

Hypothesis *H*

Prior probability of *H* is *P(H)*

New evidence *E*

Posterior probability

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Processes of mutation

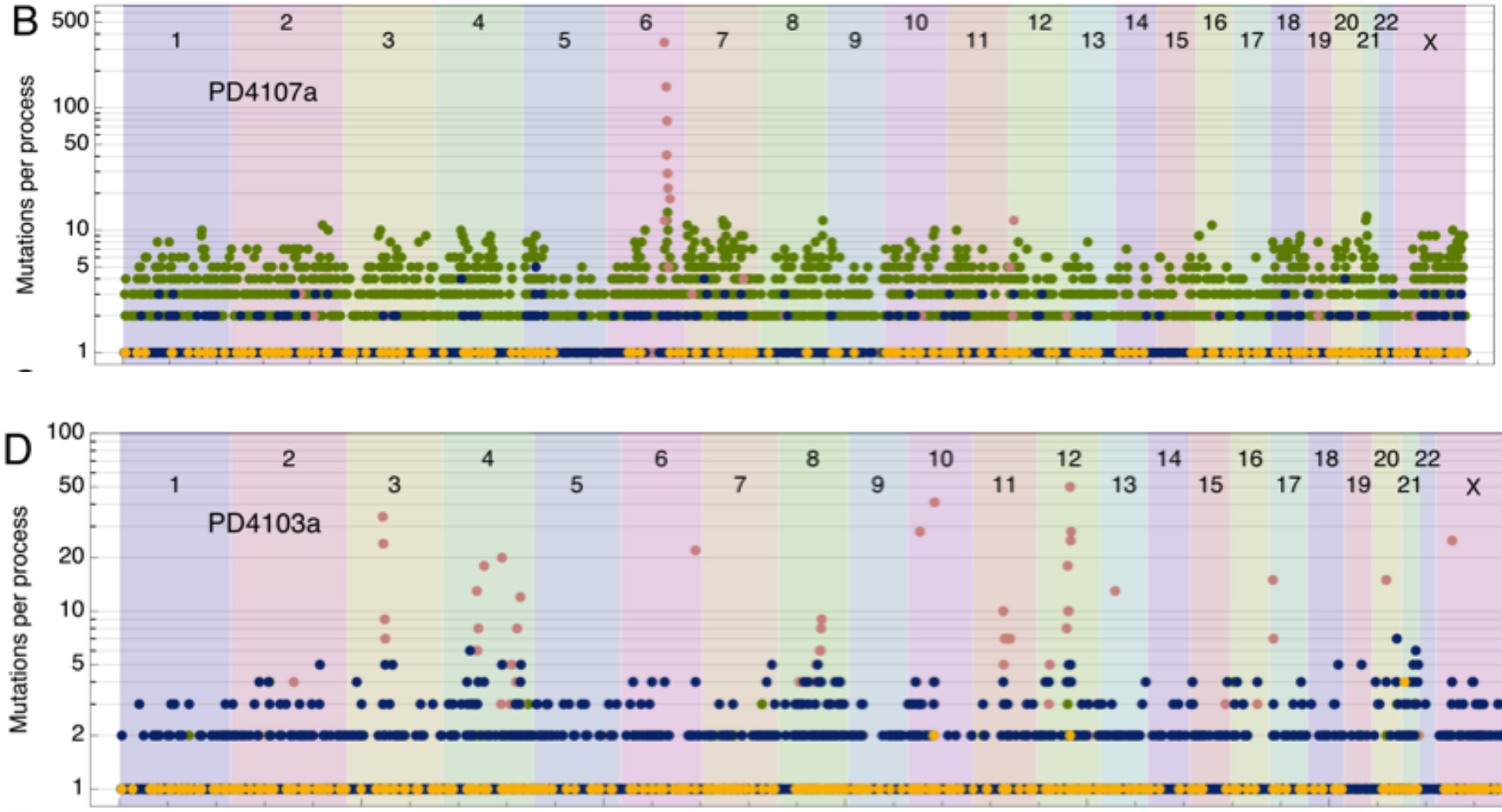
Localise processes in the genome

Given the processes identified, look for local patterns of mutation

For a specific tumour m, have global activities $x_a^{m,g}$

Use the global activities as a prior for learning local activities

Processes of mutation



A. Fischer et al., Genome Biology, 2013 14:R39

Summary

Genetic variation

Hardy-Weinberg equilibrium

Use of F-statistics

Basic forces of evolution: Genetic drift, mutation, selection
Kimura's diffusion equation

Inference of mutational spectra in cancer