6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

Computational Biology 6.047

10/09/08 Guest Lecture:

Molecular evolution: traditional tests of neutrality

Dr. Daniel Neafsey

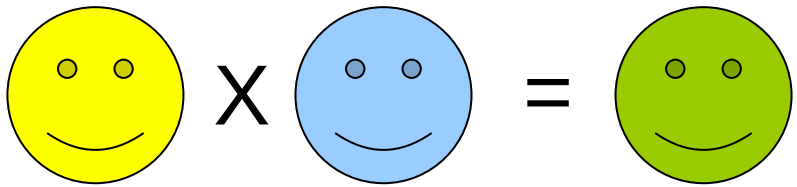
Research Scientist, Broad Institute

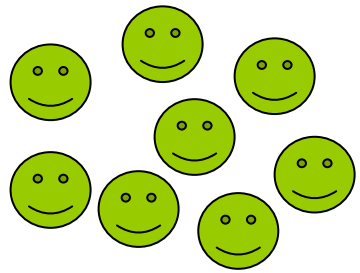
Mutation+Selection=Evolution

Relative importance of each for maintaining variation in population?

Early Criticism of Darwin

Blending inheritance, ‘gemmules’





Fleeming Jenkin (1867):

$Var[X(t+1)] = \frac{1}{2}\, Var[X(t)]$

Mendelian Inheritance

published 1865-66, rediscovered 1900

Law of Segregation:

- allelic variation
- offspring receive 1 allele from each parent
- dominance/recessivity
- parental alleles ‘segregate’ to form gametes

Law of Independent Assortment

Simple case: no selection

The Hardy-Weinberg Law (1908)

Requires:

- infinite population size
- random mating
- non-overlapping generations
- no selection, mutation, or migration

The Hardy-Weinberg Law

Genotype:	*AA*	*Aa*	*aa*
Frequency at time 0:	u_0	v_0	w_0

$$u_0 + v_0 + w_0 = 1$$

frequency of A (p_0) = $u_0 + v_0/2$

frequency of a (q_0) = $w_0 + v_0/2$

$$p_0 + q_0 = 1$$

The Hardy-Weinberg Law

Genotype:	*AA*	*Aa*	*aa*
Frequency at time 0:	u_0	v_0	w_0

Mating Pair	Frequency	Offspring		
		AA	Aa	aa
AA x AA	u_0^2	1	0	0
AA x Aa	u_0v_0	½	½	0
Aa x AA	u_0v_0	½	½	0
Aa x Aa	v_0^2	¼	½	¼

Frequency of AA in next generation: $u_1 = u_0^2 + u_0v_0 + 1/4\ v_0^2$

$= (u_0 + v_0/2)^2$

$= p_0^2$

The Hardy-Weinberg Law

If assumptions met:

- allele frequencies don't change
- after a single generation of random mating, genotype frequencies are:

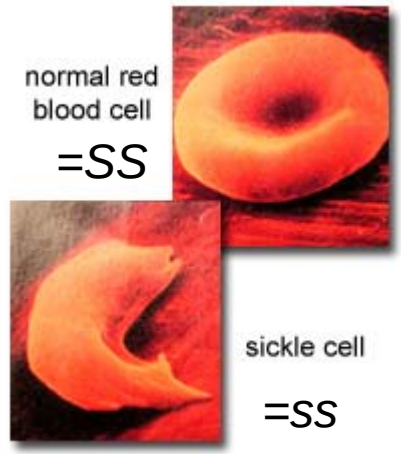
$u = p^2$ $v = 2pq$ $w = q^2$

- entire system characterized by one parameter (p)

Deviation from expectations indicates failure of 1 or more assumptions—selection?

HW application: Sickle cell anemia

	Observed Counts	Expected Counts
SS	834	
Ss	161	*2pq* *1000= 129
ss	5	



$p = \sqrt{0.834} = 0.91$

$q = \sqrt{0.005} = 0.071$

Approach: Detect selection through comparison to neutral expectation

Kimura: neutral theory

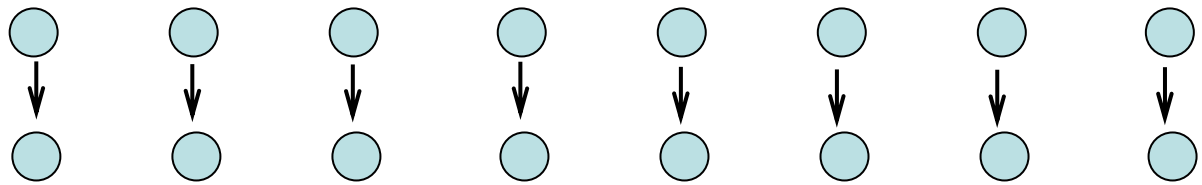
Ewens: sampling formula

Coalescence

Neutral Theory History

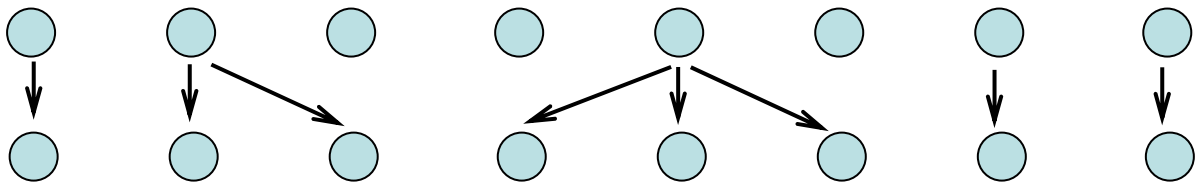
- Motoo Kimura (1924-1994)
- 1968: a large proportion of genetic change is not driven by selection
- Adapted diffusion approximations to genetics
- Dealt with finite pops

Genetic Drift



no drift

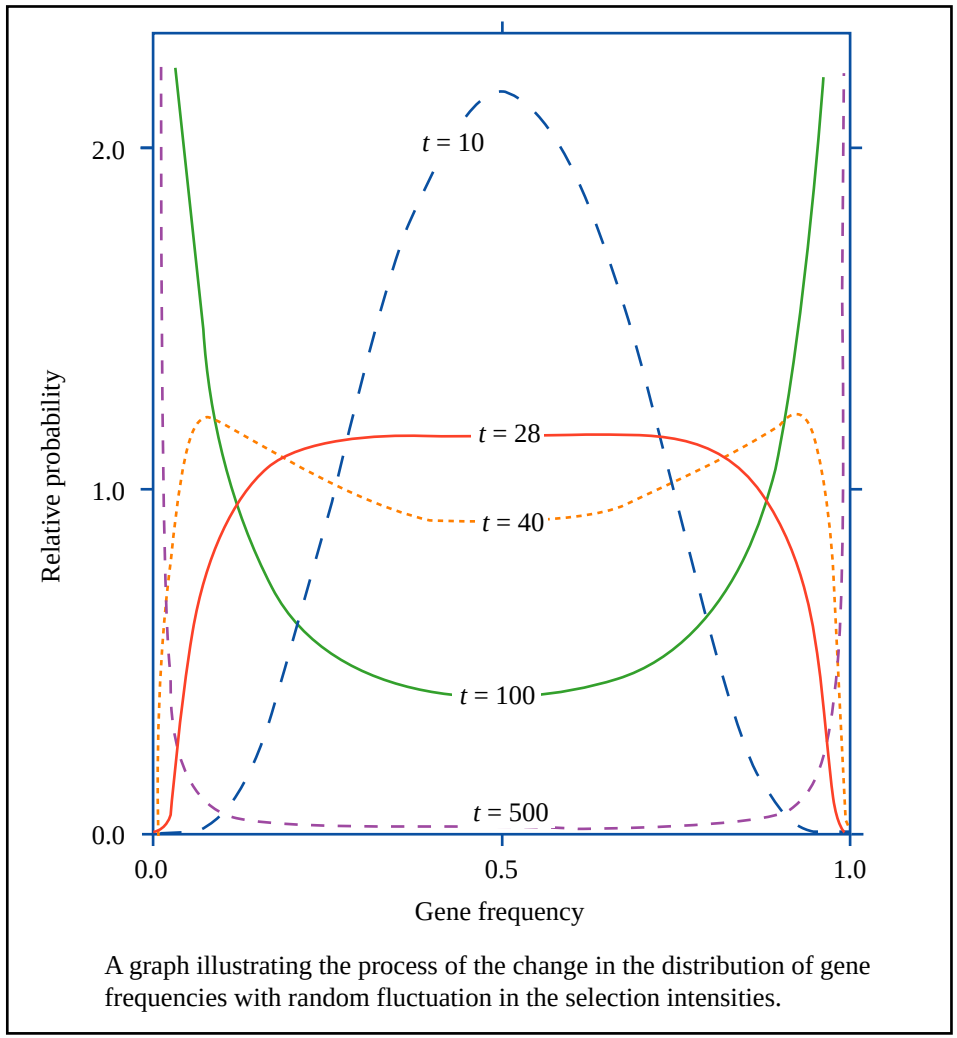
infinite pop



drift

finite pop

Neutral allele diffusion



A graph illustrating the process of the change in the distribution of gene frequencies with random fluctuation in the selection intensities.

Figure by MIT OpenCourseWare, based on:

Kimura, Motoo. "Process Leading to Quasi-Fixation of Genes in Natural Populations due to Random Fluctuation of Selection Intensities." *Genetics* 39, no. 3 (1954): 280-295.

Ewens sampling formula (1972)

- built on foundations of diffusion theory
- extended idea of ‘identity by descent’ (ibd)
- sample-based
- shifted focus to inferential methods
- introduced ‘infinite alleles’ model

Infinite alleles model

- infinite number of states into which an allele can mutate, therefore each mutation assumed unique (protein-centric)
- $2N\mu$ new alleles introduced each generation, derived from existing alleles
- initial allele frequency = $1/(2N)$
- every allele eventually lost

Infinite alleles model

Under diffusion, probability of an allele whose frequency is between x and x+δx is:

$$f(x)\partial x = \Theta x^{-1}(1-x)^{\Theta-1}\partial x$$

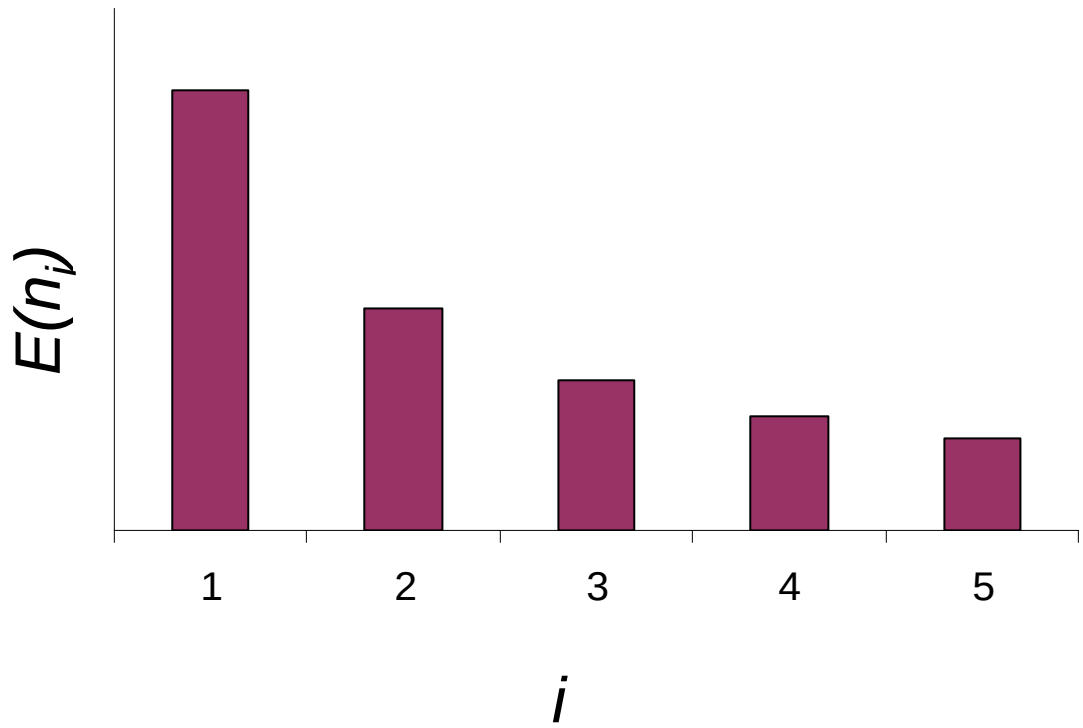
where

$\Theta = 4Nu$

N = population size

μ = mutation rate

Expected Site Frequencies



Ewens Sampling formula

Probability that a sample of *n* gene copies contains k alleles and that there are a_1, a_2, …, a_n alleles represented 1,2, …,*n* times in the sample:

$$P(a_1, a_2, ..., a_n) = \frac{n!\Theta^k}{\Theta_{(n)}} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!}$$

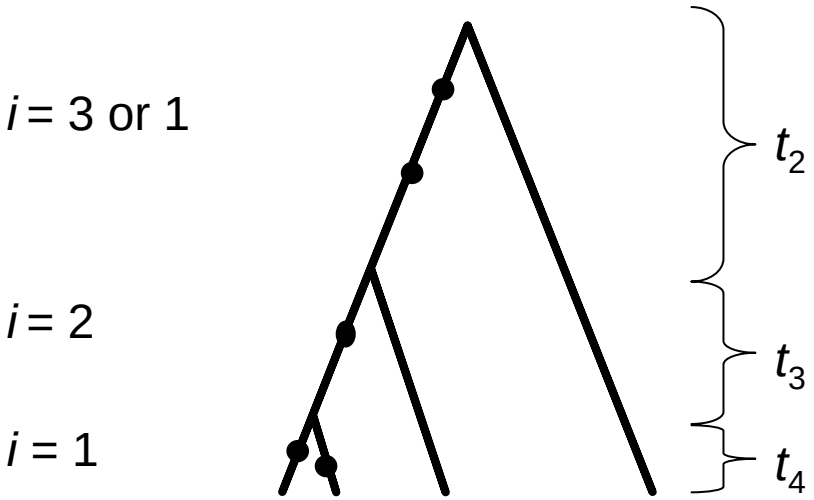
where

$$\Theta_{(n)} = \Theta(\Theta+1)...(\Theta+n-1)$$

and a_j is the number of alleles found in *j* copies

The Coalescent

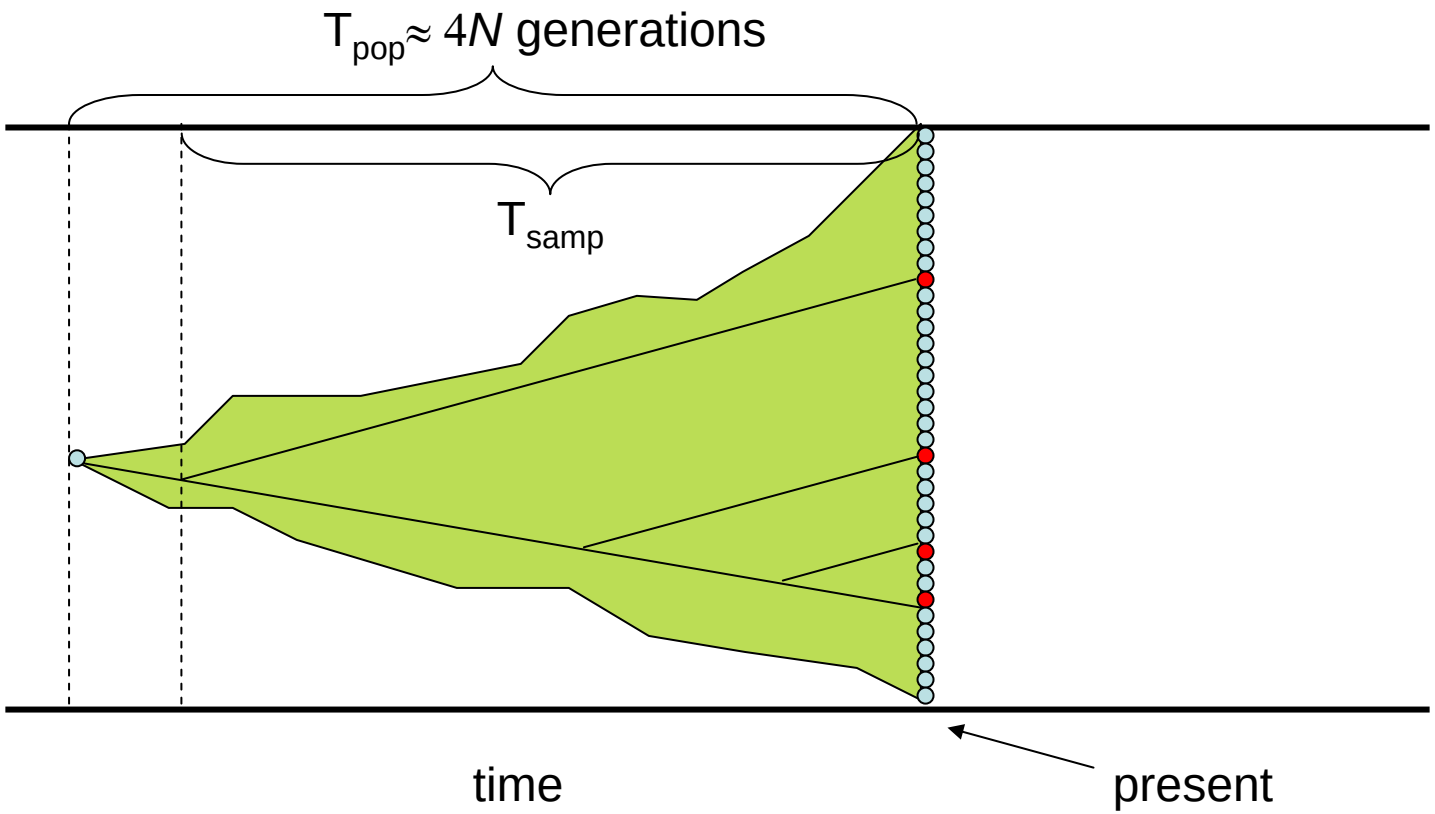
Alternate, ‘backwards’ approach to generating expected allele frequency distributions



infer tree structure (genealogy), because tree structure dictates pattern of polymorphism in data

The Coalescent

How far back in time did a sample share a common ancestor?



Coalescent inference

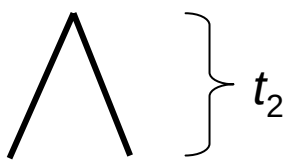
$$P(\text{pattern}) = \sum_{G} P(\text{appropriate mutations} \mid G) P(G)$$

summary statistics obviate need to actually sum over all genealogies

Sample of size 2:

P(coal) = 1/2*N*

$$f(t_2) = \frac{1}{2N} e^{-\frac{t_2}{2N}}$$



$$P(k) = \left(\frac{\Theta}{\Theta+1}\right)^{k} \left(\frac{1}{\Theta+1}\right)$$

P(mutation|event) *P*(coalescence|event)

Probability of *k* mutation events before two sequences coalesce

Turning neutral models into tests of neutrality

Three polymorphism summary statistics:

S no. of segregating sites in sample

π avg. no. of pairwise differences

η_i no. of sites that divide the sample into *i* and *n*-*i* sequences

Turning neutral models into tests of neutrality

$$S = \sum_{i=1}^{n/2} \eta_i$$

$$\Pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n/2} i(n-i)\eta_i$$

S	no. of segregating sites in sample
π	avg. no. of pairwise differences
n_i	no. of sites that divide the sample into *i* and *n-i* sequences

Turning neutral models into tests of neutrality

$$\Theta = 4N\mu$$

	Θ Estimator
$E(S) = \Theta \sum_{i=1}^{n-1} \frac{1}{i}$	$\frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$
$E(\pi) = \Theta$	π
$E(\eta_1) = \frac{n}{n-1}\Theta$	$\frac{n-1}{n}\eta_1$

Frequency-based neutrality tests

Tajima (1989) proposed:

$$D = \frac{\pi - S / a_1}{\sqrt{Var(\pi - S / a_1)}} \quad \text{where } a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

Fu and Li (1993) proposed:

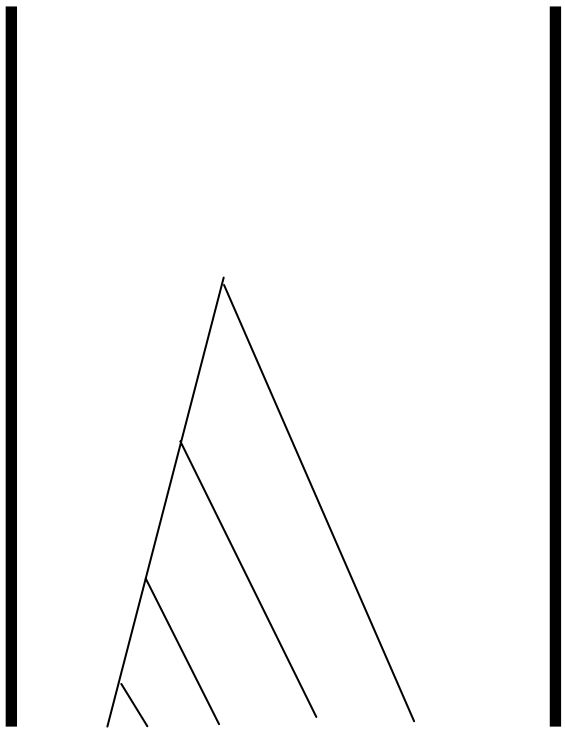
$$D^* = \frac{S / a_1 - \frac{n-1}{n}\eta_1}{\sqrt{S / a_1 - \frac{n-1}{n}\eta_1}}$$

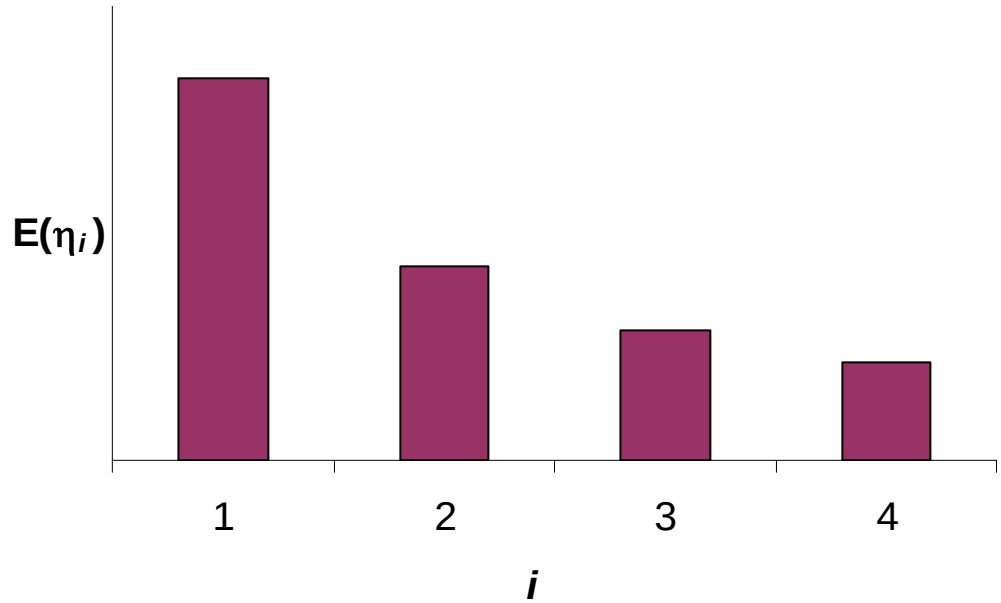
$$F^* = \frac{\pi - \frac{n-1}{n}\eta_1}{\sqrt{\pi - \frac{n-1}{n}\eta_1}}$$

Frequency-based neutrality tests

	$S = \eta_1$	$S = \eta_{[n/2]}$
	η_1 maximized	η_1 minimized
	π minimized	π maximized
$D \propto \pi - S / a_1$	Negative	Positive
$D^* \propto S / a \ -\frac{n-1}{n}\eta_1$	Negative	Positive
$F^* \propto \pi - \frac{n-1}{n}\eta_1$	Negative	Positive

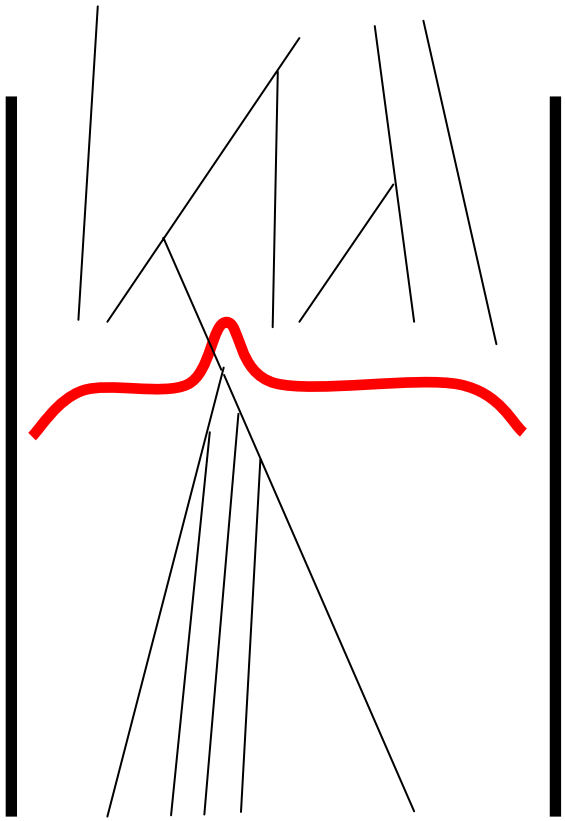
Neutral Expectation (no selection, no structure, constant population size)

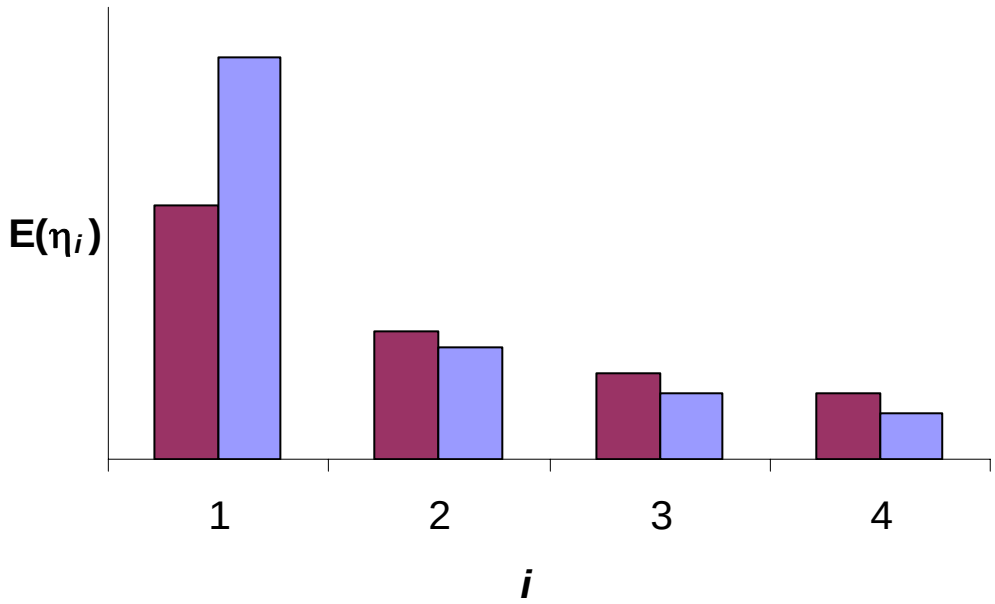




D, D*, F* ≈ 0

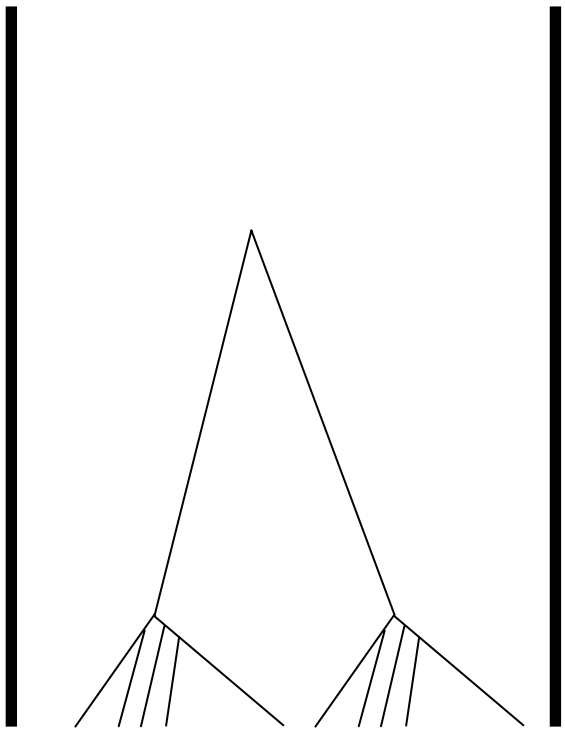
Positive Selection (Sweep)

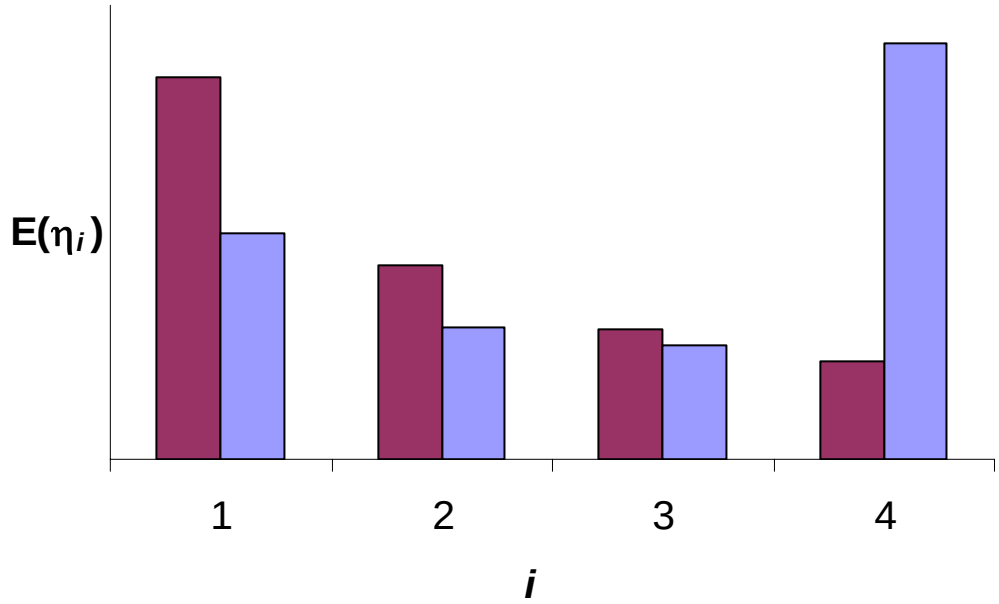




Negative D, D*, F*

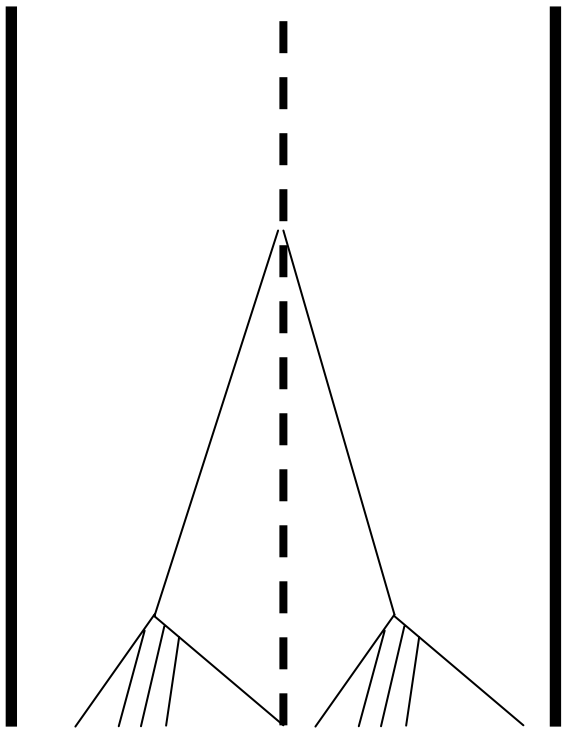
Balancing Selection

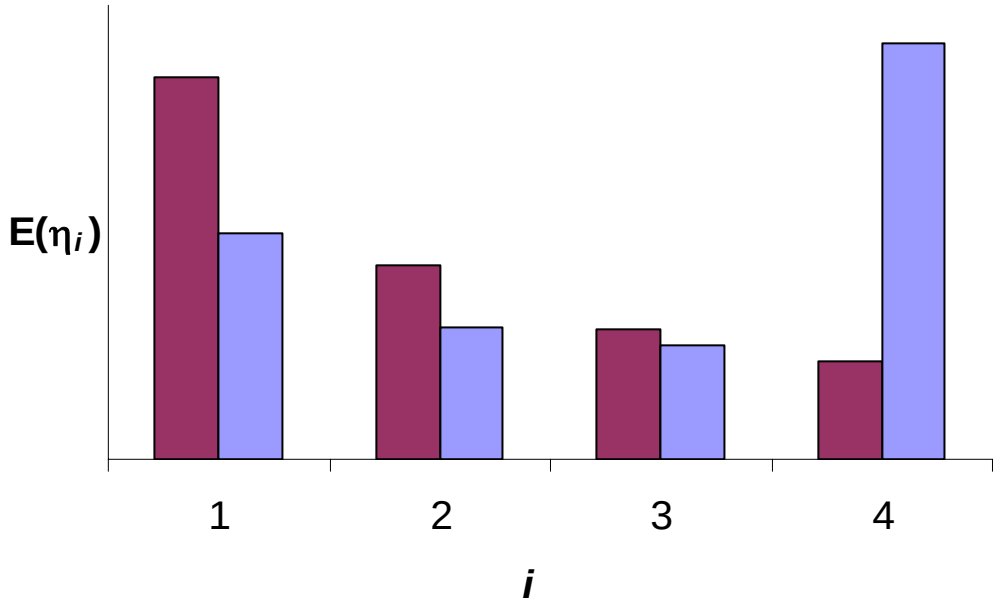




Positive D, D*, F*

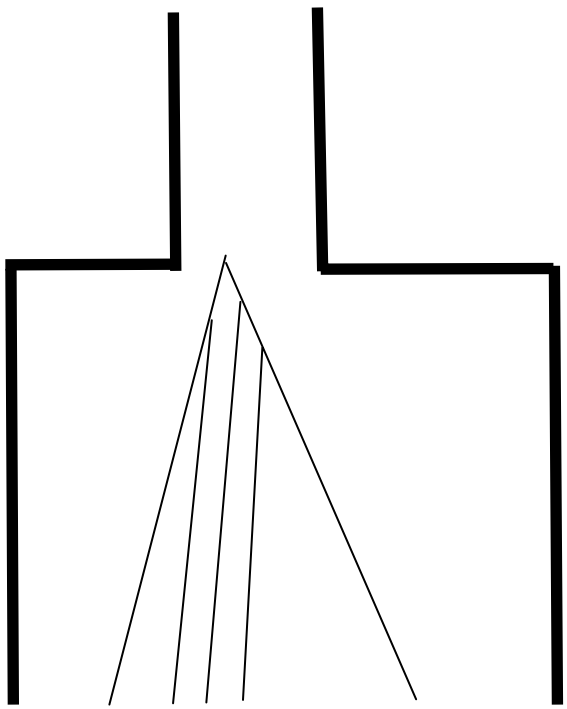
Population Structure/Subdivision

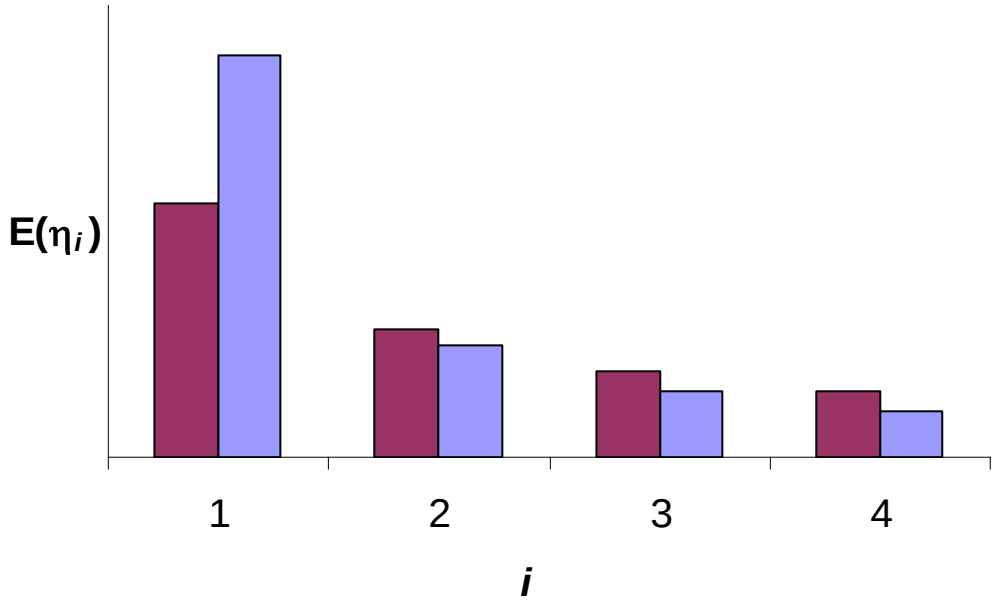




Positive D, D*, F*

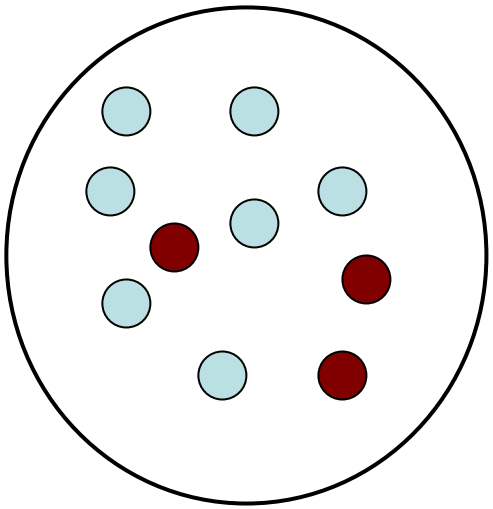
Population Expansion



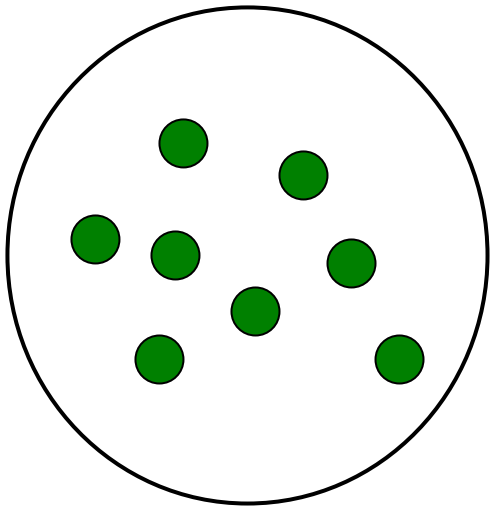


Negative D, D*, F*

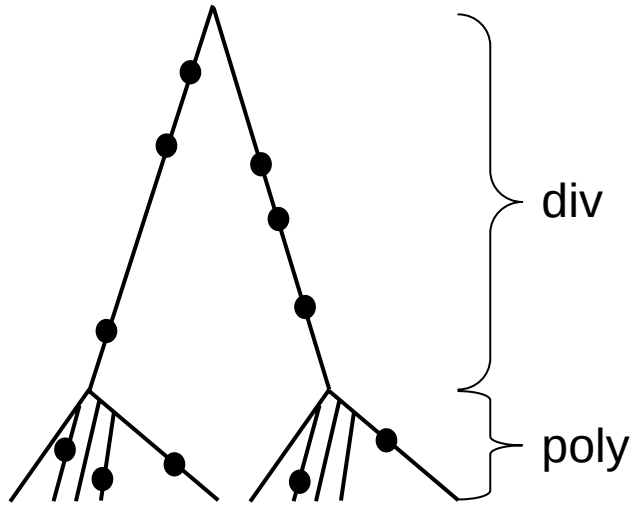
Polymorphism vs. Divergence



Species A



Species B



Divergence between species should reflect variation within species

HKA Test

Hudson, Richard, Martin Kreitman, and Montserrat Aguade. "A Test of Neutral Molecular Evolution Based on Nucleotide Data." *Genetics* 116, no. 1 (1987): 153-159.

	5' Flanking			*Adh* Locus		
	Length	No. sites compared	No. sites variable	Length	No. sites compared	No. sites variable
Within species (n = 81)	4000	414	9	900	79	8
Between species	4052	4052	210	900	324	18

Distribution of polymorphism around the *Adh* locus in *D. melanogaster* and between *D. melanogaster* and *D. sechellia*

Figure by MIT OpenCourseWare, based on paper cited above.

apply chi-squared test to summary statistics of polymorphism, divergence

Conclusion: *Adh* exhibits excessive polymorphism

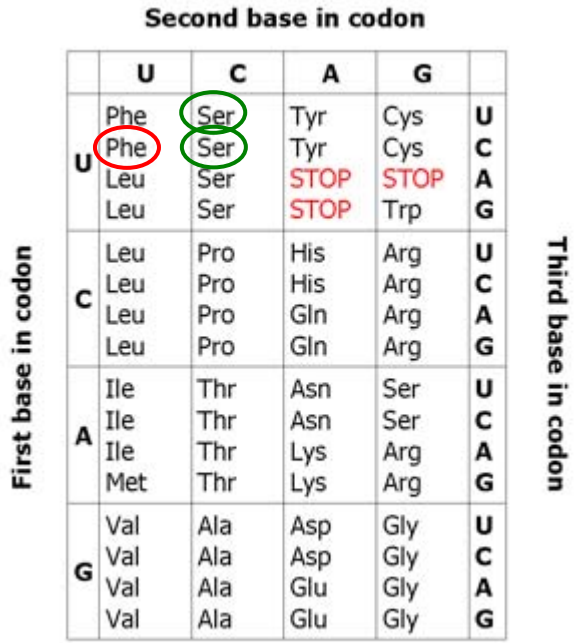
Polymorphism/divergence with a twist: site classes

Synonymous changes: don't affect amino acid

UCU $\Rightarrow$ UCC=Serine

Nonsynonymous (replacement) changes: new amino acid

UCU $\Rightarrow$ UUC= Phenylalanine



First base in codon	Second base in codon: U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

MK Test

McDonald, John, and Martin Kreitman. "Adaptive Protein Evolution at the *Adh* locus in *Drosophila*." *Nature* 351 (1991): 652-654.

	Fixed	Polymorphic
Replacement	7	2
Synonymous	17	42

Number of replacement and synonymous substitutions for fixed differences between species and polymorphisms within species

Figure by MIT OpenCourseWare, based on paper cited above.

MK test requires only 1 locus, but polymorphism data from 2 species.

Adh exhibits an excessive proportion of replacement fixed differences.

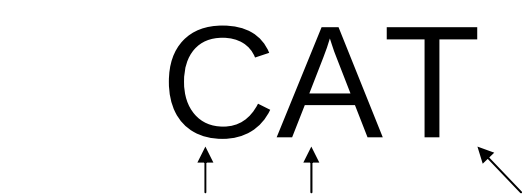
Rate-based selection metric: d_N/d_S

d_N = no. nonsynonymous changes/ no. nonsynonymous sites

d_S = no. synonymous changes/ no. synonymous sites

Counting codon ‘sites’ example: CAT

Histidine is encoded by only one other codon: CAC



P(T⇒C) = fractional syn sites

P(T⇒G or A) = fractional nonsyn sites

Rate-based selection metric: d_N/d_S

$d_N/d_S < 1$ purifying selection

$d_N/d_S = 1$ neutral expectation

$d_N/d_S > 1$ positive selection

Rate-based selection metric: d_N/d_S

- Can be calculated using various methods
- Goldman & Yang implementation (PAML):

nucleotide changes modelled as continuous-time Markov chain with state space = 61 codons

$$q_{ij} = \begin{cases} 0: \text{ if the two codons differ at } > 1 \text{ position} \\ \pi_j: \text{ synonymous transversion} \\ \kappa\pi_j: \text{ synonymous transition} \\ \omega\pi_j: \text{ nonsynoymous transversion} \\ \omega\kappa\pi_j: \text{ nonsynonymous transition} \end{cases}$$

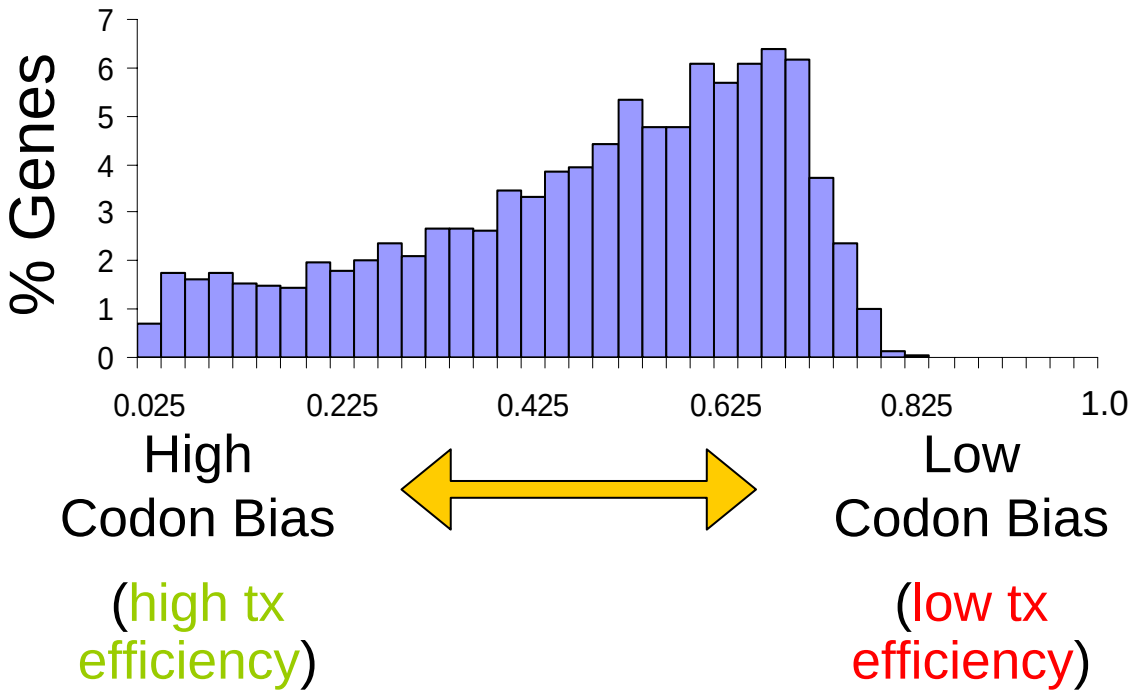
Rate-based selection metric: d_N/d_S

Are syn sites really neutral?

Codon Bias and Translation

Codon bias: the unequal usage of synonymous codons

-Thought to reflect selection for optimal translational efficiency and/or translational accuracy.



Distribution of Codon Bias Estimates for 6,453 *Cryptococcus* Genes

Correlates with d_N/d_S (or just d_N)

- expression level (-)
- dispensability (+)
- protein abundance (-)
- codon bias (-)
- gene length (+)
- number of protein-protein interactions (-)
- centrality in interaction network (-)

Neutrality Tests Summary

- Allelic frequency spectrum tests (Tajima's D)
- Polymorphism/divergence tests (HKA, MK)
- Rate-based metric: d_N/d_S

The future:

empirical tests based on genomic data that are not dependent on demographic assumptions (Pardis Sabeti)

tests that incorporate biophysical properties of amino acids into calculation of syn, nonsyn changes?