HST.508/Biophysics 170: Quantitative genomics Module 1: Evolutionary and population genetics Lecture 4: natural selection – its modeling and detection

Professor Robert C. Berwick

Topics for this module

1. The basic forces of evolution; neutral evolution and drift
2. Computing 'gene geneaologies' forwards and backwards; the coalescent
3. Natural selection and its discontents
4. Detecting selection: Molecular evolution; from classical methods to modern statistical inference techniques

Agenda for today

1. Natural selection: from the basic dynamical system equation to the diffusion approximation: how can genes survive?
2. How can we detect selection in our data?

To think about from *Nature*

"Protein sequences evolve through random mutagenesis with selection for optimal fitness" – Russ, Lowery, Mishra, Yaffe, Ranganathan, sept. 2005, 437:22, p. 579. *Natural-like function in artificial WW domains.*

The new reality game show - "Survivor"
1 gene in 2 different forms (alleles)

genotype	AA	Aa	aa
frequency	p^2	$2pq$	q^2
Viability	w_{11}	w_{12}	w_{22}
after selection	$w_{11}\, p^2$	$w_{12}\, 2pq$	$w_{22}\, q^2$

survivors

Intuitively, w is a 'growth rate'

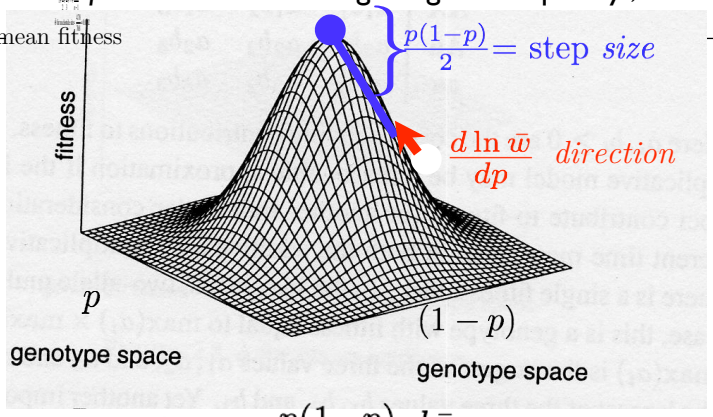
Note that if N_t = # before selection, the total # after selection is:

$N_{t+1} = \bar{w} N_t$ where

$\bar{w} = w_{11}p^2 + w_{12}2pq + w_{22}q^2$

mean fitness $= \bar{w}$

Sewall Wright's adaptive landscape:
Understanding the formula
Evolution *equated* to mean 'change in gene frequency', delta *p*



$$\Delta p = \frac{p(1-p)}{2\bar{w}} \frac{d\bar{w}}{dp}$$

$$\Delta p = \frac{p(1-p)}{2} \frac{d \ln(\bar{w})}{dp}$$

Some dissection...

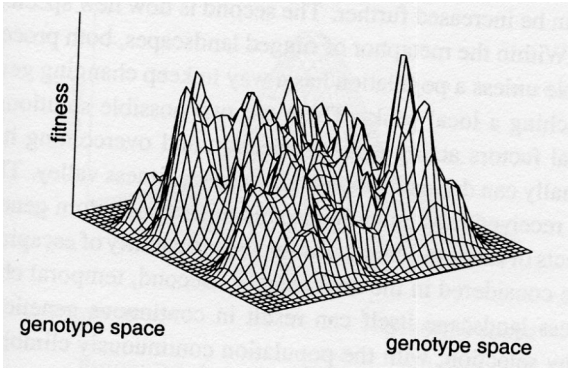
$$\triangle p = \frac{p(1-p)}{2} \frac{d(\bar{w})}{\bar{w} dp}$$

Variance component of allele A within genotype

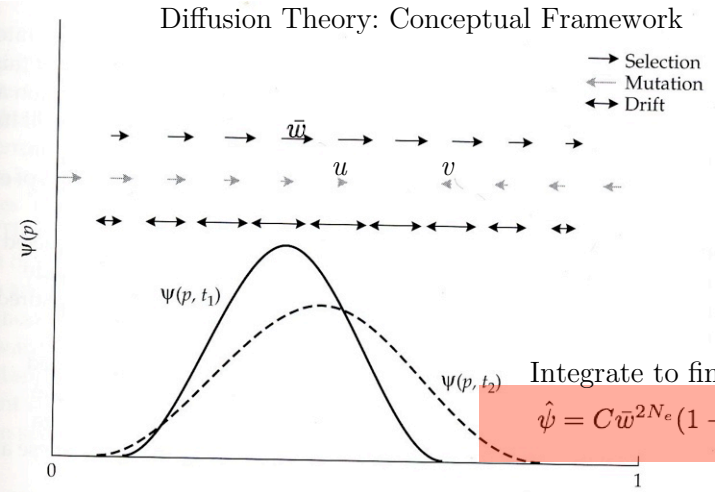
Slope of fitness function divided by mean population fitness – a *potential function*?

Why variance? Draw from pool of A, a gametes many many times: binomial sampling – frequency of A within a genotype is either 1, 1/2, or 0; variance is $p(1\text{-}p)/2$ ("heterozygosity")

But…



Diffusion Theory: Conceptual Framework



Integrate to find:

$$\hat{\psi} = C\bar{w}^{2N_e}(1-p)^{4N_e u-1}p^{4N_e v-1}$$

The prob distribution of the # populations – i.e., the prob density – with different allele frequencies *shifts* under the directional effects of selection & mutation (& migration) and *flattens out* under the effects of drift

Mutation vs. drift: set $\bar{w} = 1 =$ constant, $u = v$,

$$\hat{\psi}(p) \propto [p(1-p)]^{4N_e u-1}$$

Drift wins when $4N_e u \ll 1$

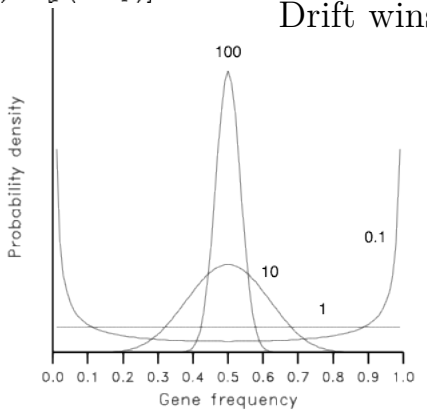
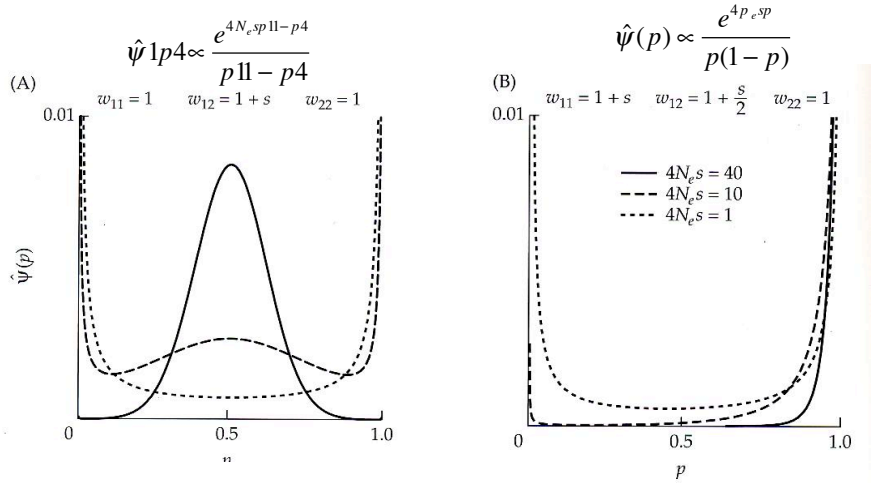


Figure 7.6: Equilibrium distribution of gene frequencies under mutation and genetic drift. In this case $4N_e u = 4N_e v$. The values of $4N_e u$ are shown next to the curves.

$$\hat{\psi}(p) = C\bar{w}^{2N_e}(1-p)^{4N_e u-1}p^{4N_e v-1}$$



Balancing selection Directional selection

Drift wins when $4N_e \ll 1$

Cannot say how effective selection is <u>without</u> knowing effective population size!!!

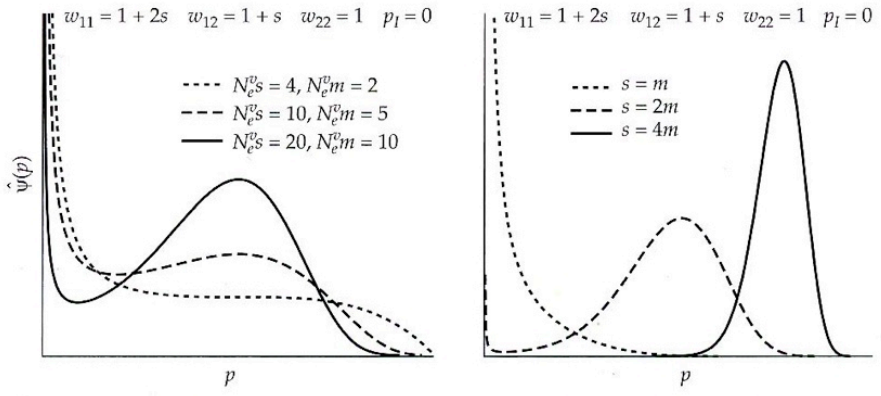


Figure 5.7 Equilibrium distributions of allele frequencies under selection, drift, and migration. The curves in (A) are for different effective population sizes with $s = 2m$ in each case. The curves in (B) correspond to different values of s for $N_e^v m = 5$. Note that increasing the effective population size makes the distribution more peaked, but has little effect on the position of the peak. Increasing the strength of selection relative to migration, though, shifts the peak to the right.

"Follow the variation": some famous data about individual variation in *Drosophila melanogaster* (Marty Kreitman)

Allele	39	226	387	393	441	513	519	531	540	578	606	615	645	684
Reference	T	C	C	C	C	C	T	C	C	A	C	T	A	G
Wa-S	.	T	T	.	A	A	C	.	.	.	.	.	.	.
Fl-1S	.	T	T	.	A	A	C	.	.	.	.	.	.	.
Af-S	.	.	.	.	.	.	.	.	.	.	.	.	.	A
Fr-S	.	.	.	.	.	.	.	.	.	.	.	.	.	A
Fl-2S	G	.	.	.	.	.	.	.	.	.	.	.	.	.
Ja-S	G	.	.	.	.	.	.	.	T	.	T	.	C	A
Fl-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Fr-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Wa-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Af-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Ja-F	G	.	.	A	.	.	.	G	T	C	T	C	C	.

Table 1.1: The 11 *ADH* alleles. A dot is placed when a nucleotide is the same as the nucleotide in the reference sequence. The numbers refer to the position in the coding sequence where the 14 variant nucleotides are found (see Figure 1.1). The first two letters of the allele name identify the place of origin. The S alleles have a lysine at position 192 of the protein; the F alleles have a threonine.

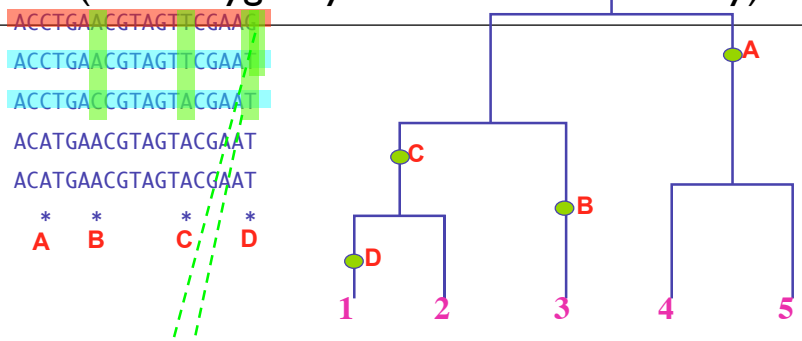
Kreitman 1983 original data set for melanogaster Adh sequences
Kreitman,M (1983): Nucleotide polymorphism at the alcohol #ehydrogenase locus of Drosophila melanogaster.
Nature 304, 412-417.

Different aspects of the data used to test neutrality

- Nucleotide diversity
- Allelic frequency spectrum
- Polymorphism / Divergence
- K_A / K_s (amino-acid changing vs. 'silent' substitutions in DNA

No 'one size fits all'!

Θ_T estimated from pairwise differences (heterozygosity or nucleotide diversity)



A mutation on an *interior* branch will have higher weight

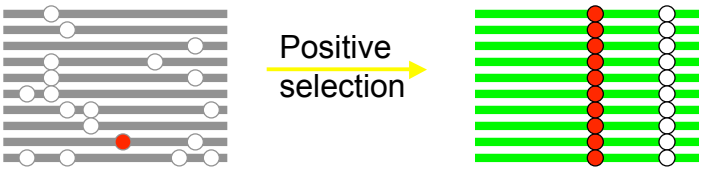
Θ_T = Average Pairwise Distance

$$\theta_T = \sum_{i<j} \frac{D(i,j)}{\binom{n}{2}}$$

$= (1+3+3+3+2+2+2+2+2+2)/10=2$

Deviations from the neutral model

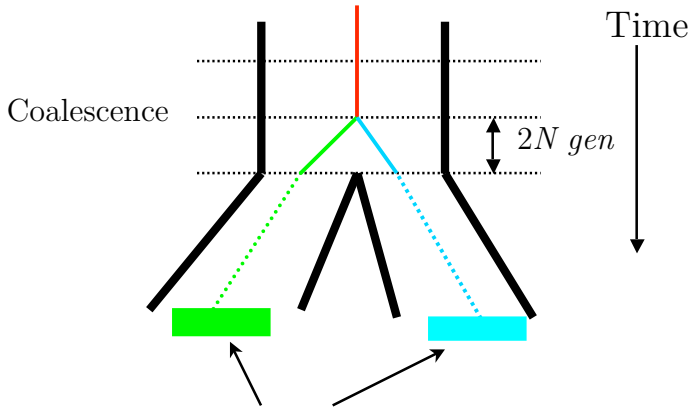
- Positive selection



● Advantageous mutation

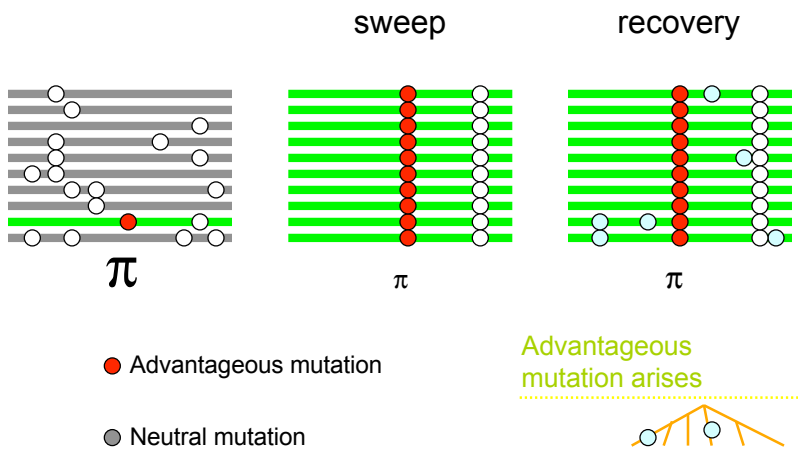
○ Neutral mutation

The Great Obsession: variation (polymorphism) entangled with descent

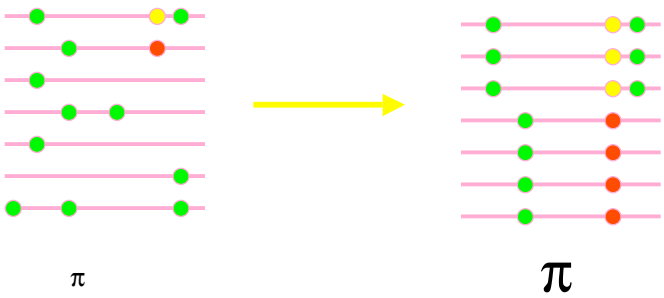


Observed sample variation: is it from descent (tree), or from biology?

Positive selection will decrease nucleotide diversity (π)



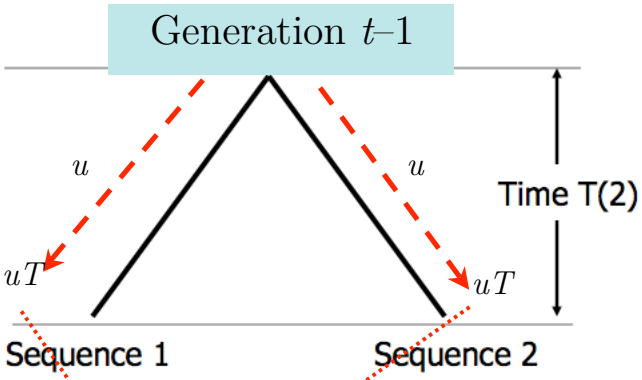
Balancing selection will increase nucleotide diversity (π)



balanced' mutation

Neutral mutation

Estimating nucleotide divergence as θ

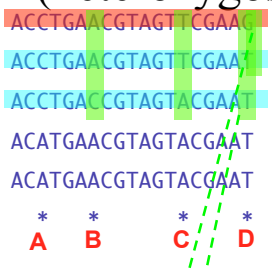


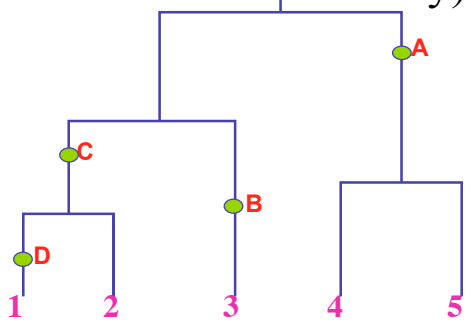
Notation: T_i= time to collapse of i genes, sequences,…

$2uT = 2u\ 2N = 4uN = \theta$

20

Θ_T estimated from pairwise differences (heterozygosity or nucleotide diversity)





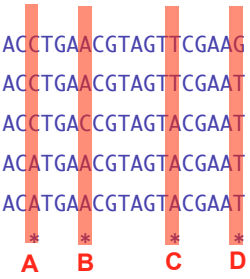
A mutation on an *interior* branch will have higher weight

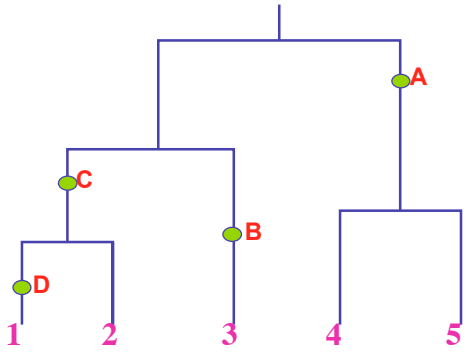
Θ_T = Average Pairwise Distance

= (1+3+3+3+2+2+2+2+2+2)/10=2

$$\theta_T = \sum_{i<j} \frac{D(i,j)}{\binom{n}{2}}$$

$\Theta_W=4N\mu$ estimated from # segregating sites





Expected number of segregating sites:

$$S_n = \Theta_W \sum_{i=1}^{k-1} \frac{1}{i}$$

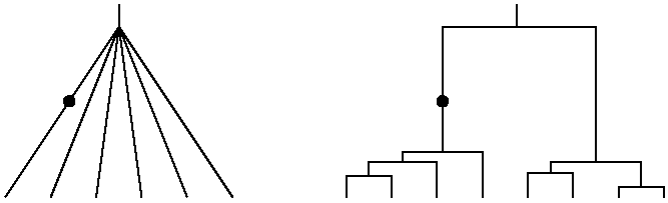
(coalescent theory)

$$E(S_n) = uE(T_c) = \sum_{i=1}^{n-1} \frac{1}{i}$$

Θ_W= 4/(1+1/2+1/3+1/4)=24/11=2.1818

Watterson, 1975

Different coalescent patterns (relative branch lengths) yield different estimates for theta even though total branch length is the same and # segregating sites remains the same

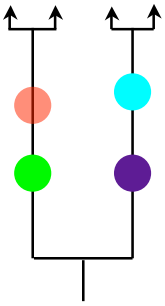


Second type of mutation counted more times when calculating the average pairwise distance – typical when there's a 'burst' after a population bottleneck

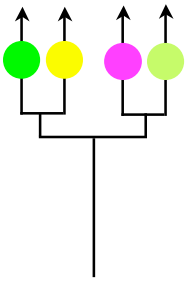
Use the *difference* between the two estimates to figure out a statistical measure that can pick out these two patterns

Consider these coalescent pattern differences & what they imply about possible *patterns* of variation (heterozygosity) if there are *neutral* mutations sprinkled on these patterns…

Note that S= # segregating sites remains the same…



Expect: *more* mutations on interior branches, sample heterozygosity *higher*



Expect: *fewer* mutations on interior branches, sample heterozygosity *lower*

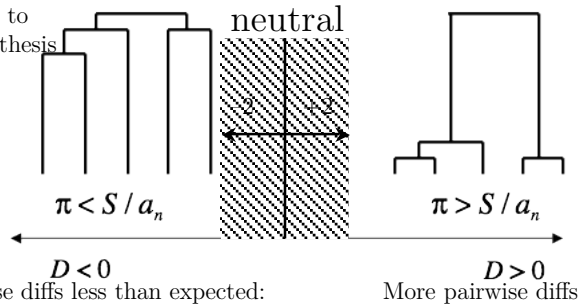
Tajima's D is the *difference* between these two estimates, normalized by a variance measure

If neutral model holds, this difference should be 0 - test whether difference from 0 could be due to chance or not

$$D = \frac{\pi - S/a_n}{\sqrt{\mathrm{Var}(\pi - S/a_n)}} \qquad a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Tajima (1989)

Complex estimate for variance - to get null hypothesis distribution



Pairwise diffs less than expected: Long external branches, mutations at low frequency

More pairwise diffs than expected from # of segregating sites: mutations at high freq

Human mitachondrial DNA

Ingman *et al.* (2000)

52 complete molecules from a worldwide sample (linguistic groups)

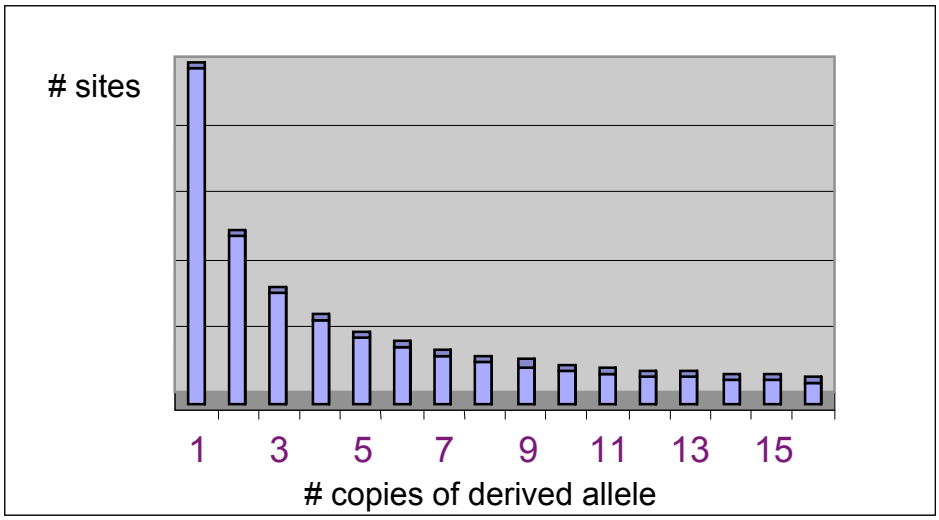
521 segregating sites excluding D-loop

$$\pi = 44.2$$
$$a_{52} = 4.52$$
$$S/a_{52} = 115.3$$
$$\sqrt{\hat{\mathrm{V}}(d)} = 31.8$$
$$D = \frac{44.2 - 115.3}{31.8} = -2.23$$

Probability of observing such an extreme value under neutrality = 0.01

Human mtDNA have an excess of low-frequency variants

Expected distribution under neutral model

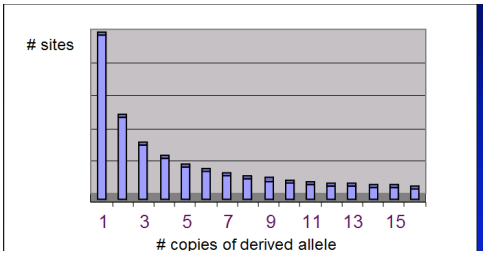


We can also use this to predict the allelic frequency spectrum – given the number of observed segregating sites, we can estimate what the nucleotide diversity will be – under the null model.

Sample size

$$\theta = s/\left(\sum_{i}^{n-1} 1/i\right)$$

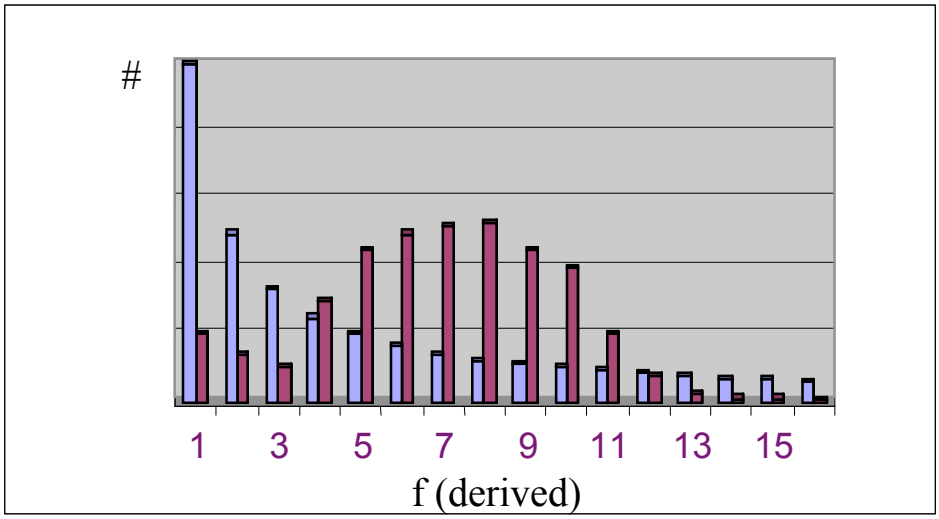
Number of SNPs



Under neutrality $\theta = \pi$

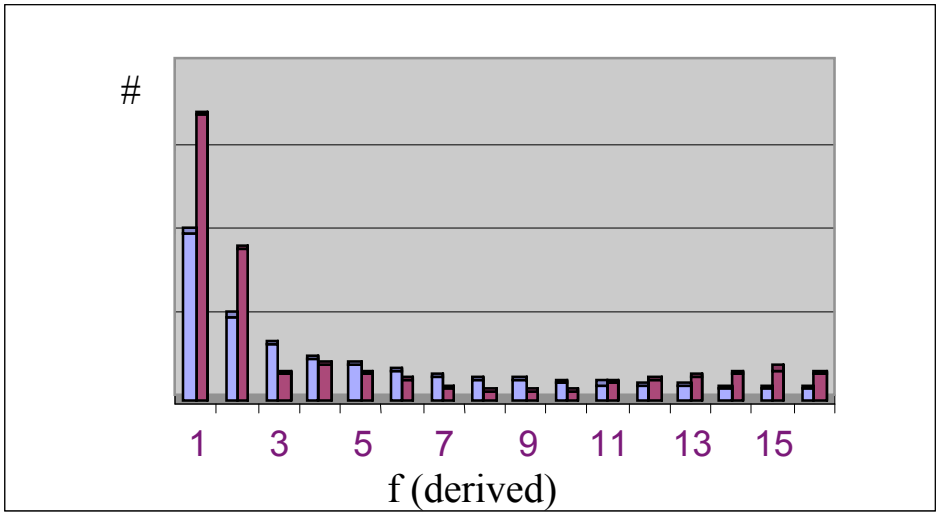
Tajima's *D* (Tajima 1989) → $D = \theta - \pi$ (normalized by the sd)

An excess of intermediate frequency alleles



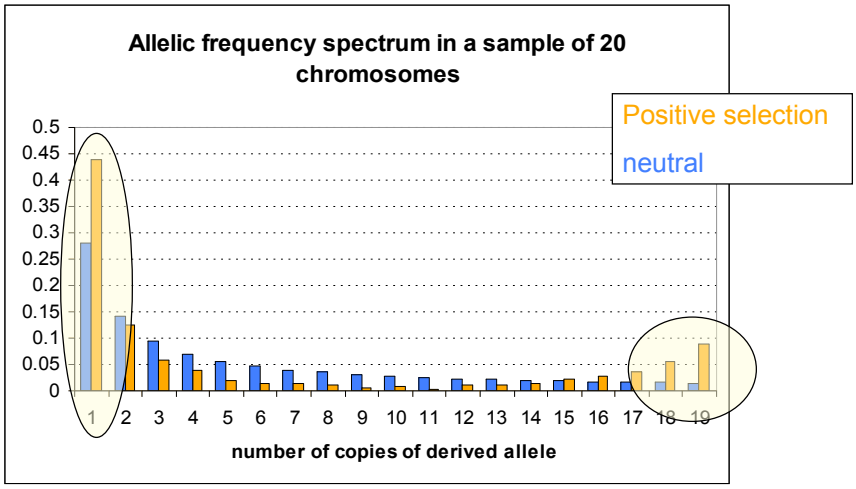
$D = \theta - \pi \rightarrow > 0$

An excess of rare alleles



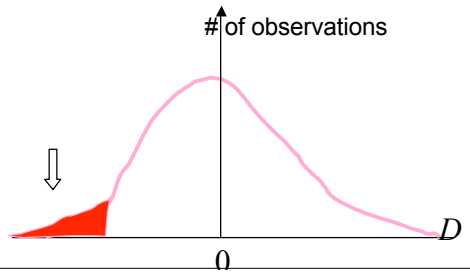
$D = \theta - \pi \rightarrow < 0$

Immediately after positive selection, the expectation is:



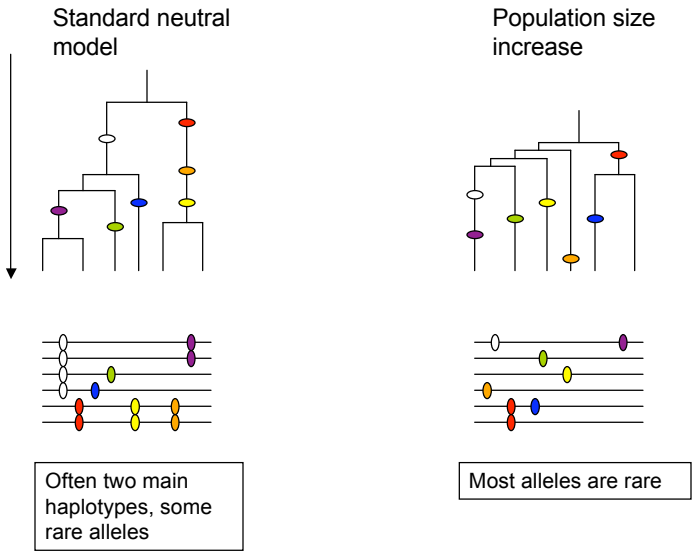
Goodness of fit tests of the allelic frequency spectrum

- Summarize your data, *e.g.*, by Tajima's *D*.
- Assume some simple null model, *e.g.* the standard neutral model, and build the distribution of the summary under that model (usually by simulation)
- Check how the actual value compares to the expected value, by looking at the probability of obtaining more extreme values
- If this probability is low (*e.g.*, $< 5\%$) reject the model



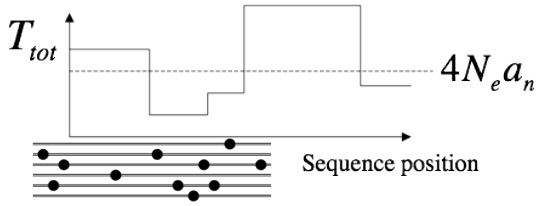
Purifying selection will also result in an excess of rare alleles

Growth will also result in an excess of rare alleles

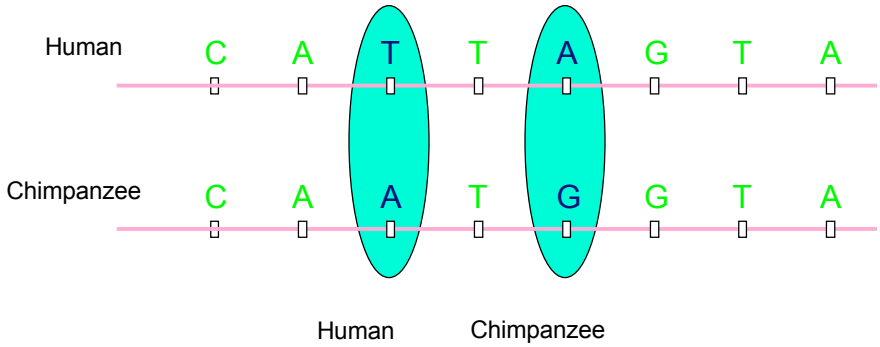


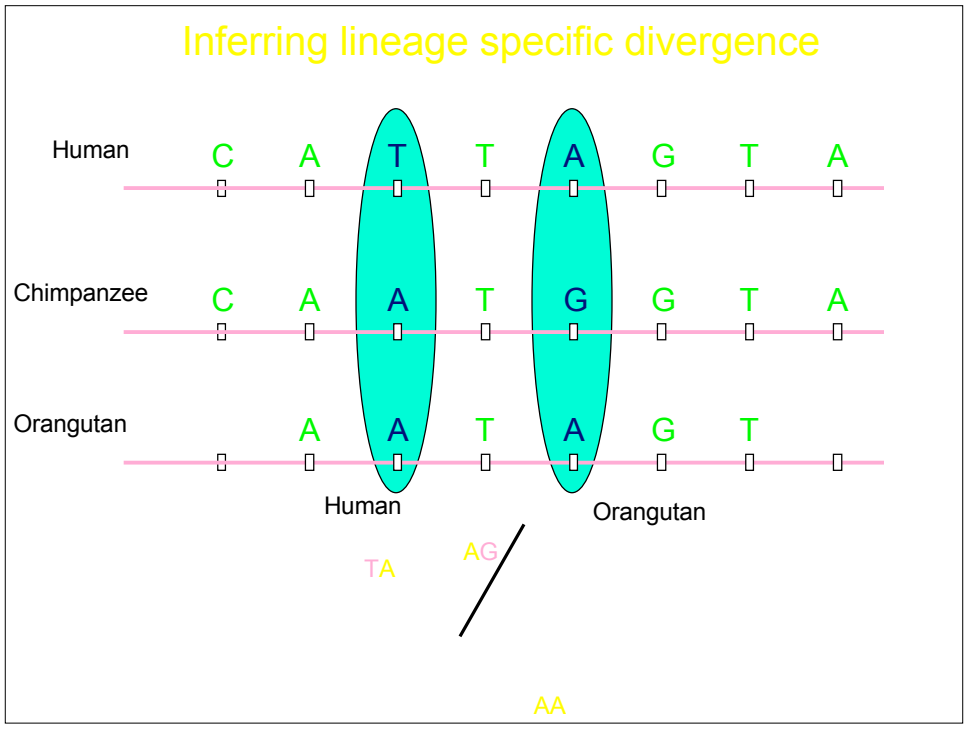
Factors affecting test power

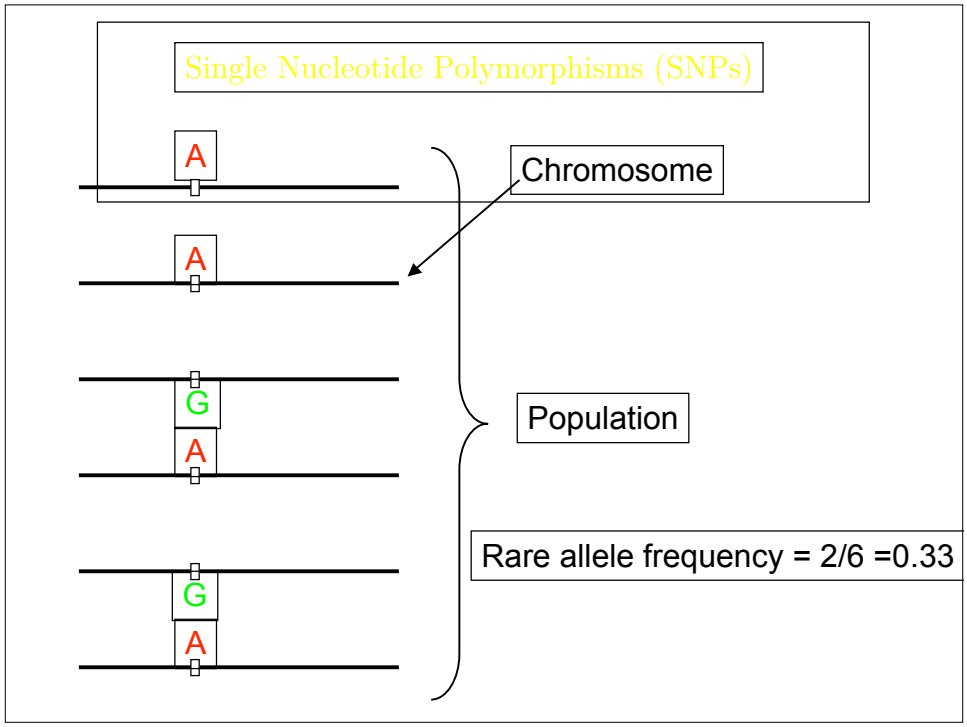
- The number of mutations in the sample is of critical importance
  - In general, sequencing a large region is more important than sequencing many individuals
- Recombination reduces the possibility of drawing trees from sequences, but evens out evolutionary stochasticity



Divergence - the number of fixed sites



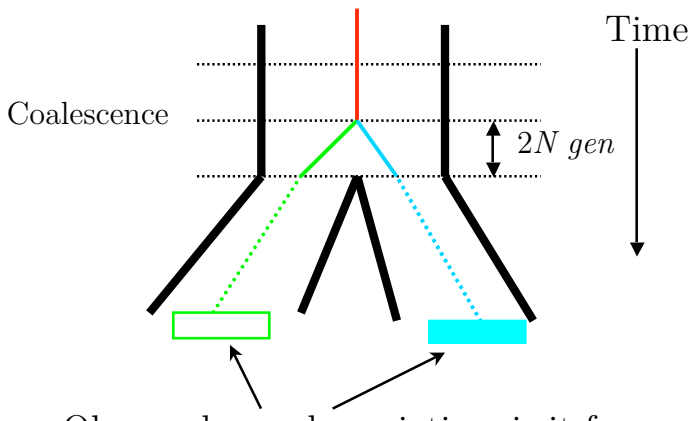

Inferring lineage specific divergence
Human
Chimpanzee
Orangutan
C A T T A G T A
C A A T G G T A
A A T A G T
Human
Orangutan
TA
AG
AA


Single Nucleotide Polymorphisms (SNPs)
Chromosome
Population
Rare allele frequency = 2/6 =0.33

Different aspects of the data used to test neutrality

- Nucleotide diversity
- Allelic frequency spectrum
- Polymorphism / Divergence
- K_A / K_S

The Great Obsession: variation (polymorphism) entangled with descent



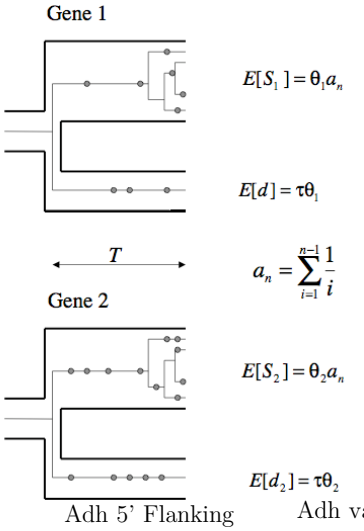
Observed sample variation: is it from descent (tree), or from biology?

HKA test: untangling *divergence* from polymorphism

- Divergence= $2ut + 4N_e u$
- Uses ratio of polymorphism to divergence and tests whether there is either more or less polymorphism compared to divergence at one locus compared to the other, using a test statistic based on a Chi-square goodness of fit; the ratio should be the same at both genes even if their mutation rates differ:

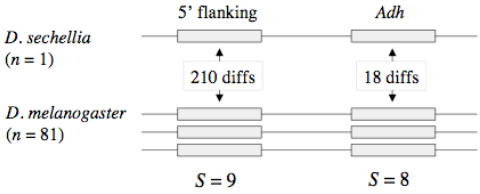
$$\frac{4N_e u}{(2ut + 4N_e u)} = \frac{4N_e u}{(2t + 4N_e)}$$

- Think of u as compound parameter uf, where f is the fraction of neutral sites. While f and u may be different at different genes, the ratio of polymorphism to divergence is independent of uf – so compare to a known neutral target…



	Adh 5' Flanking variable sites	Adh variable sites
Polymorphism	9	8
Divergence	210	18
P/D	0.04	0.44

Adh in Drosophila



Solving the simultaneous equations

$\hat{\tau} = 13.4 N_e$ generations

$\hat{\theta}_1 = 2.7$

$\hat{\theta}_2 = 0.7$

$\chi^2 = 6.09 \quad P = 0.016$

Fast / slow polymorphism in exon 4 leads to a two-fold difference in enzyme activity

Cline in polymorphism: Fast more common in northern America and at higher altitudes

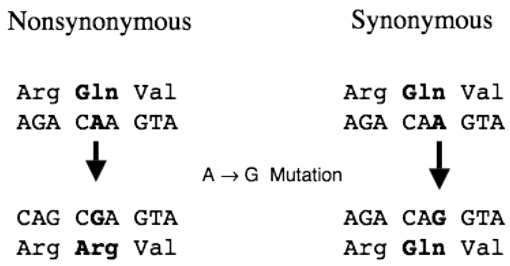
Different aspects of the data used to test neutrality

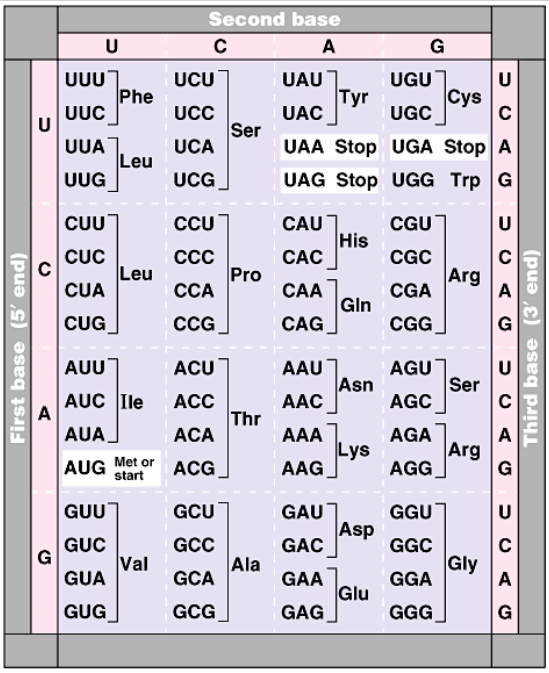
- Nucleotide diversity
- Allelic frequency spectrum
- Polymorphism / Divergence
- K_A / K_S

And last, but not least…the ratio of the rate of amino acid replacement substitutions to the rate of silent (aka 'synonymous' or 'semantaphoretic') substitutions (K_A / K_s)

Ka / Ks = 1 indicates equal rates of the two classes of substitutions, hence, 'neutral' evolution

Problem: very conservative





First base (5′ end)	Second base: U	Second base: C	Second base: A	Second base: G	Third base (3′ end)
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met or start	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

K_A = # nonsynonymous substitutions/# nonsynonymous sites

K_S = # synonymous substitutions/# synonymous sites

Test for selection by comparing d_N and d_S

$K_A / K_S = 1$: Neutral evolution

$K_A / K_S < 1$: Purifying selection

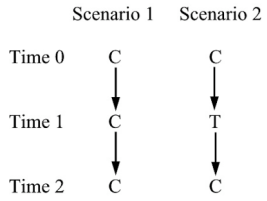
$K_A / K_S > 1$: Positive selection

The K_A / K_S ratio (ω) measures the selective pressure

Assumptions can affect calculation of K_A/K_S

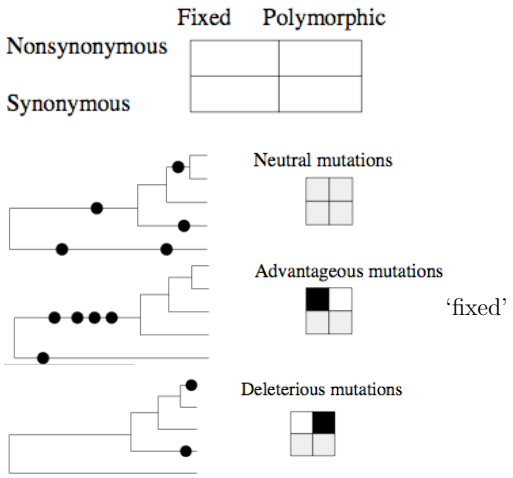
- All nucleotide sites change independently
- The substitution rate is constant over time and in different lineages
- The base composition is at equilibrium
- The conditional probabilities of nucleotide substitutions are the same for all sites, and do not change over time
- Most of these are not true in many cases...

Most importantly: multiple hits, parsimony



There were no guarantees that a particular site had not undergone multiple changes. Two possible scenarios where multiple substitutions at a single site would lead to *underestimation* of the number of substitutions that had occurred if a simple count were performed.

MK test



McDonald-Kreitman test (MK)

Two-way table: compare *within group* nonsynonymous/synonymous substitutions (polymorphisms) vs. *between group* nonsynonymous/synonymous substitutions ('fixed') – use, e.g., Fisher's exact test to compute whether due to chance

MK Example

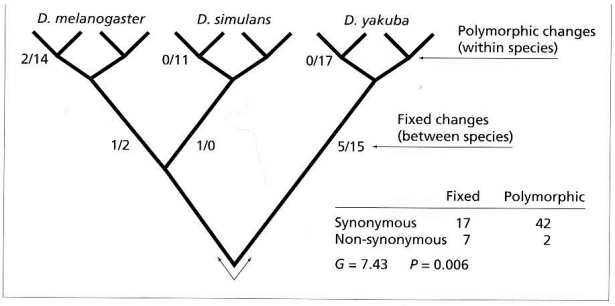


Fig. 7.22 Numbers of non-synonymous and synonymous substitutions in the *Adh* locus within (polymorphic) and between (fixed) three species of *Drosophila*. The results of a *G*-test show that there are significantly fewer non-synonymous polymorphisms than expected by the neutral theory given the number of fixed non-synonymous changes observed between species. The double-headed arrow at the bottom of the tree signifies that the fixed substitutions leading to *D. yakuba* could have occurred at any time since this species separated from *D. melanogaster* and *D. simulans*. Adapted from McDonald and Kreitman (1991).

Detecting selection… the landscape…

- *Levels of variation* (HKA test):
  1. Low levels of variation compared to a reference (HKA test)
  2. High levels compared to a reference
- *Frequency of variation* (gene or allele spectrum) (TD test):
  1. Excess of rare compared to common frequency variation (TD <0)
  2. Excess of common compared to rare frequency variation (TD>0)
  3. Excess of high compared to common frequency genes
- *Polymorphism vs. divergence (MK test):*

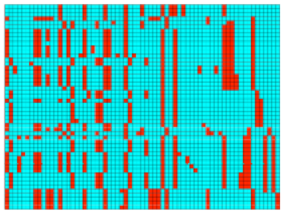
Kreitman's review - Ann. Rev. Genetics

TABLE 1 Statistical tests of selection[a]

Test	Type	Designed to detect	Best use	Caveats	Reference(s)
HKA	Within vs between spp. (two loci)	Differences in variation levels not accountable by constraints	Balancing selection; recent selective sweeps or other variation-reducing forces	High recombination rates may reduce effectiveness of test	49
McDonald (run test)	Within- vs between-spp. (contiguous region)	Regions with non-neutral patterns of poly. and div.	Equilibrium balancing selection	Has some advantages over the HKA test	71, 72
McDonald Kreitman G	Within- vs between-spp. (syn. vs nonsynon.)	Adaptive evolution	Adaptive protein evolution; mutation/selection	Selection on codon usage can seriously jeopardize test	73
Tajima's D	Within sp.	Skew in frequency spectrum	General purpose test of frequency spectrum skew	See reference 27 for situations in which the test performs poorly	96
Fu & Li's D	Within sp.	Recent vs ancient mutations	General purpose test of frequency spectrum skew	Fu's more recent tests may be more powerful	29
Fu W	Within sp.	Departures in frequency spectrum	Population subdivision	Hudson's Gst test is more powerful for detecting subdivision	27
Fu $G\eta$	Within sp.	Departures in frequency spectrum	Population subdivision, shrinkage, and overdominance selection	Little power against excess number of rare alleles 28	27
Fu $G\xi$	Within sp.	Departures in frequency spectrum	Population subdivision, shrinkage, and overdominance selection	Little power against excess number of rare alleles	27
Fu F_s	Within sp.	Excess or rare alleles (one sided)	Population growth, genetic hitchhiking, and background selection	May be best overall test for detecting genetic hitchhiking and population growth	28
Hudson	Within sp. and allele	Unexpectedly low variation within an allele class	Directional selection	A good test for young alleles driven to high frequency	45
Wall B ans Q	Within sp.	Linkage disequil. between adjacent segregating sites	Population subdivision, balancing selection	Q is more powerful when there is substantial recombination	100
Andolfatto's S_k	Within sp. (sliding window)	Non-neutral haplotype structure	Balancing and directional selection; pop. subdivision	Interpretation may be difficult	2

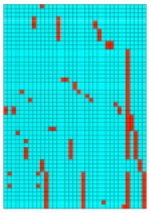
[a]Abbreviations: HKA, Hudson-Kreitman-Aguadé; syn., synonymous; nonsynon., nonsynonymous; disequil., disequilibrium; poly., polymorphism; div., divergence; pop., population.

Try different simulations…which matches data best?



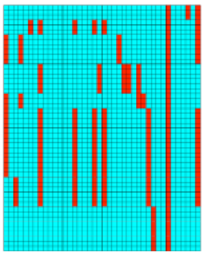
$\hat{\theta}_W = 15.0$
$\hat{\theta}_\pi = 16.3$
$\hat{\theta}_e = 17.0$
$\hat{\theta}_H = 12.7$
$K/S = 0.37$

Null model n=50, θ=10, ρ=10



$\hat{\theta}_W = 7.8$
$\hat{\theta}_\pi = 3.9$
$\hat{\theta}_e = 13.0$
$\hat{\theta}_H = 1.5$
$K/S = 0.63$

Growth n=50, θ=10, ρ=10, λ=5



$\hat{\theta}_W = 4.2$
$\hat{\theta}_\pi = 5.8$
$\hat{\theta}_e = 0.0$
$\hat{\theta}_H = 6.0$
$K/S = 0.42$

Recent bottleneck: n=50, θ=10, ρ=10,
10 ancestral lineages