Neutral Theory

- Kimura: accumulation of genetic differences between species can be explained by the combined effects of mutation and genetic drift **w/o selection**
- neutral theory recognizes the role of strong negative (or purifying selection) but suggests that positive selection plays a minor role in determining overall patterns of genetic variation
- neutral theory provides the null hypothesis against which tests for selection are made

Tests for departures from neutrality

- comparison of a pair of sequences between species is a dull tool for detecting selection at individual sites
- tests making use of **population** genetic data:
  - HKA test
  - MacDonald-Kreitman test
  - Site Frequency Spectrum
  - Tajima's D
  - Genome-wide patterns of divergence and linkage disequilibrium

fixed (divergence) vs. polymorphic synonymous vs. nonsynonymous

```
Sp1    GCG TGC GAC TCA CCA TTG
Sp2.1  GCG TTC GAC TCA CCG TTG
Sp2.2  GCA TTC GAC TCA CCG TTG
Sp2.3  GCA TTC GAC TCA CCG TTG
Sp2.4  GCG TTC GAC ACA CCG TTG
```

s_P n_F n_P s_F

neutral expectation: $n_P/s_P = n_F/s_F$

neutral expectation: $n_P/s_P = n_F/s_F$

$$S = \theta\sum_{i=1}^{n-1}\frac{1}{i} = 4N\mu\sum_{i=1}^{n-1}\frac{1}{i}$$

$$k = 2\mu t$$

$$k = 2\mu t + 4N_A\mu$$

- if there is one mutation rate for neutral synonymous mutations μ_S and another for neutral non-synonymous mutations μ_N, then:

$$\frac{n_P}{s_P} = \frac{4N\mu_N\sum_{i=1}^{n-1}\frac{1}{i}}{4N\mu_S\sum_{i=1}^{n-1}\frac{1}{i}} = \boxed{\frac{\mu_N}{\mu_S}}$$

and...

$$\frac{n_F}{s_F} = \frac{2\mu_N t}{2\mu_S t} = \boxed{\frac{\mu_N}{\mu_S}}$$

McDonald-Kreitman test

- contingency test for similar proportions of synonymous and non-synonymous changes within and between sister species
- if all changes are neutral, the ratio of nonsynonymous to synonymous polymorphism within species should equal the ratio of nonsynonymous to synonymous fixed differences between species

McDonald & Kreitman 1991 Adaptive protein evolution at the ADH locus in *Drosophila*. *Nature* 351: 652-654.

- compared ADH sequences (n = 30) from 3 species of *Drosophila* (*melanogaster*, *simulans*, *yakuba*)

	# of sites		G-test p = 0.006
Differences	Fixed	Polymorphic	Total
Synonymous	s_F (17)	s_P (42)	59
Nonsynonymous	n_F (7)	n_P (2)	9
Sum	24	44	68

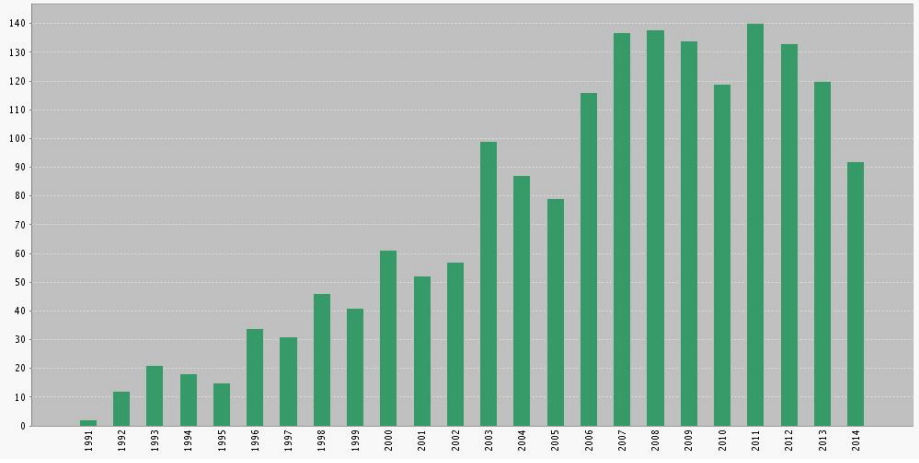
- relative excess of nonsynonymous fixed differences suggests positive diversifying selection on ADH locus
- nearly neutral evolution is unlikely explanation for results

Complications in interpreting the McDonald-Kreitman test

- no accounting for "multiple hits"
  - therefore best for recently diverged species
  - but not so recent that there are shared polymorphisms
- inferences about non-synonymous substitutions requires an assumption of neutrality for synonymous mutations
  - synonymous mutations also subject to selection
  - e.g., codon usage bias
    - 82% GC at 3rd codon positions in *Drosophila*
    - e.g., GCC >> GCT in Ala codons

$$n_P/s_P = n_F/s_F$$

McDonald & Kreitman 1991 Adaptive protein evolution at the ADH locus in *Drosophila*. *Nature* 351: 652-654.



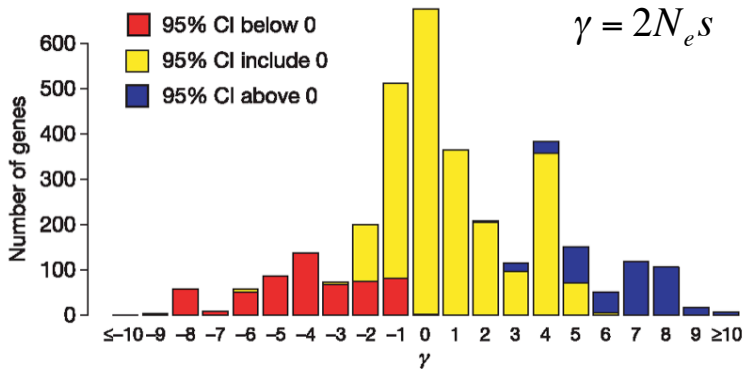
Bustamante *et al*. 2005 Natural selection on protein-coding genes in the human genome. *Nature* 437: 1153-1157

- Celera Genomics sequenced 20,362 protein-coding loci in 20 EAs 19 AAs and one male chimpanzee
- compared 11,624 genes using the MK test
- s_F = 34,099 fixed synonymous differences between 39 humans and chimpanzee across 11.81 Mb of sequence
- s_P = 15,750; n_F = 20,467; n_P = 14,311
- n_F / s_F = 23.76% < n_P / s_P = 38.42%
  - $p < 2 \times 10^{-16}$

Bustamante *et al*. 2005 Natural selection on protein-coding genes in the human genome. *Nature* 437: 1153-1157

- separate analyses of 3,277 genes with at least four non-synonymous sites (i.e., $n_F + n_P \geq 4$)
- estimated selection from patterns of polymorphism and divergence
  - as in the McDonald-Kreitman test

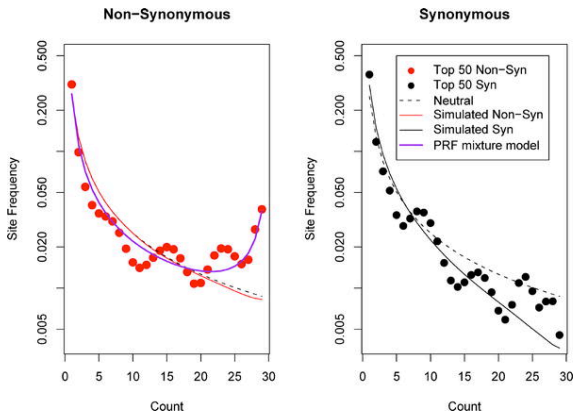
Bustamante *et al*. 2005 Natural selection on protein-coding genes in the human genome. *Nature* 437: 1153-1157



Nielsen *et al*. 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology* 3: e170.

- tested for dN/dS ratios > 1, comparing 20,000+ genes between human and chimpanzee
- 25 of the top 50 genes were then tested for higher dN/dS in primates in the context of the phylogeny
  - ((human,chimp),(mouse,rat))
  - 5/25 faster in humans, 5 in chimps, 8 in both

Nielsen *et al*. 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology* 3: e170.



Nielsen *et al*. 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology* 3: e170.

- broad overlap with previous study in the specific genes identified as being under positive selection in the human lineage
  - genes involved in immune function, spermatogenesis, sensory perception (e.g., olfactory), and transcription factors
- cancer-related genes in the top 50
  - four putative tumor suppressors (*HYAL3, DFFA, PEPP-2, C16orf3)*, one associated with tumor progression *(MMP26)*, a gene with high similarity to melanoma-associated antigens *(FLJ32965)* and several genes involved in apoptosis *(PPP1R15A, HSJ001348, TSARG1,* and *GZMH)*

The HKA test (Hudson, Kreitman & Aguade)

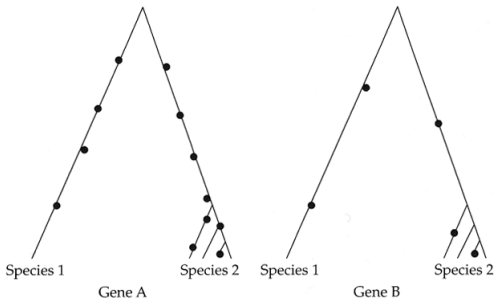
- similar logic to MK test but greater statistical power by comparing patterns of divergence and polymoprhism across genes
- for each gene i, determine (S = "segregating"):

S_i^A the number of polymorphic sites in species A

S_i^B the number of polymorphic sites in species B

D_i divergence : the average number of differences between alleles from A and B

The HKA test (Hudson, Kreitman & Aguade)



under neutrality, the ratio of polymorphisms to divergence should be equal across genes

FIGURE 7.17 Under the neutral theory, levels of divergence and polymorphism are expected to be correlated. For gene *A*, there is a high neutral mutation rate, driving many sites of nucleotide divergence and high levels of polymorphism within species 2. For gene *B*, the neutral mutation rate is low, so species 1 and 2 have few sites of divergence, but also species 2 has fewer polymorphic sites. The HKA test is a goodness-of-fit test to the observed levels of intraspecific diversity and interspecific divergence under a model whose parameters are population sizes, neutral mutation rates, and times of divergence.

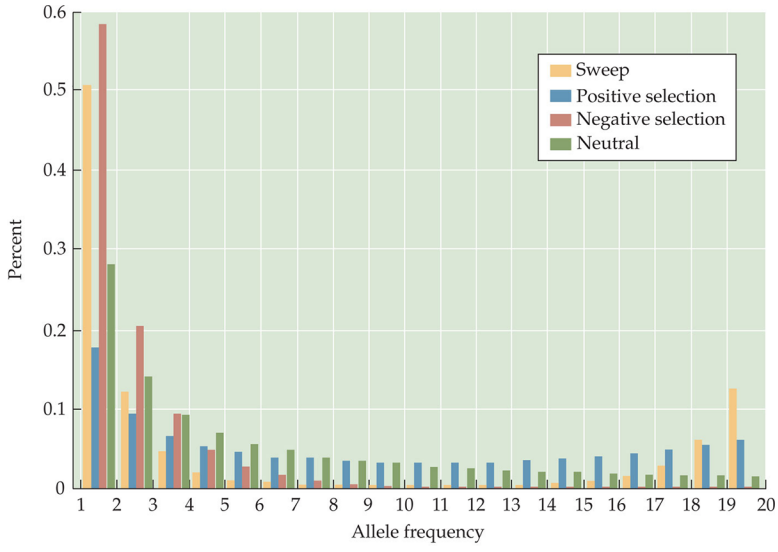
The HKA test (Hudson, Kreitman & Aguade)

- test statistic:

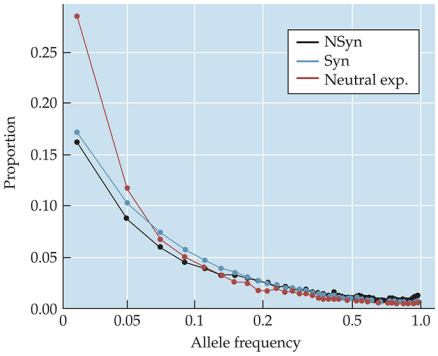
$$X^2 = \sum_{i=1}^{L}\left(S_i^A - \hat{E}\left(S_i^A\right)^2\right)\Big/\hat{V}ar\left(S_i^A\right) + \sum_{i=1}^{L}\left(S_i^B - \hat{E}\left(S_i^B\right)^2\right)\Big/\hat{V}ar\left(S_i^B\right) + \sum_{i=1}^{L}\left(D_i - \hat{E}\left(D_i\right)^2\right)\Big/\hat{V}ar\left(D_i\right)$$

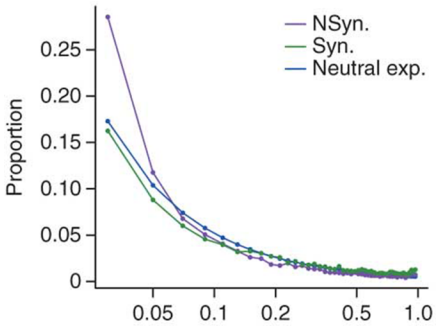
- is not based on dN/dS and thus allows comparison of both coding and non-coding regions

Site Frequency Spectrum



Site Frequency Spectrum





Li *et al*. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature Genetics* **42**, 969–972.

Tajima's D

- compares two estimators of θ
  - Waterson's

$$\hat{\theta}_W = S \Big/ \left(\sum_{i=1}^{n-1} 1/i \right)$$

  - Tajima's

$$\hat{\theta}_T = \frac{\sum_{(i,j),i<j} d_{ij}}{n(n-1)/2}$$

- Tajima's D

$$\frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{\hat{V}\left(\hat{\theta}_T - \hat{\theta}_W\right)}}$$

Tajima's D

nuc_div — sites

$$\frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{\hat{V}\left(\hat{\theta}_T - \hat{\theta}_W\right)}}$$

- Interpretation
  - selective sweep leads to negative values
  - balancing selection results in positive values
  - rapid population expansion leads to negative values
  - population bottlenecks can temporarily generate positive values

Neutrality index (Rand & Kann 1996)

- $NI = (n_P/n_F)/(s_P/s_F)$
- NI = 1 under strict neutrality
- NI > 1 implies "excess" amino acid variation within species
  - purifying selection
- NI < 1 implies "excess" amino acid variation between species
  - adaptive diversifying selection
- NI = (2/7)/(42/17) = 0.12 for ADH

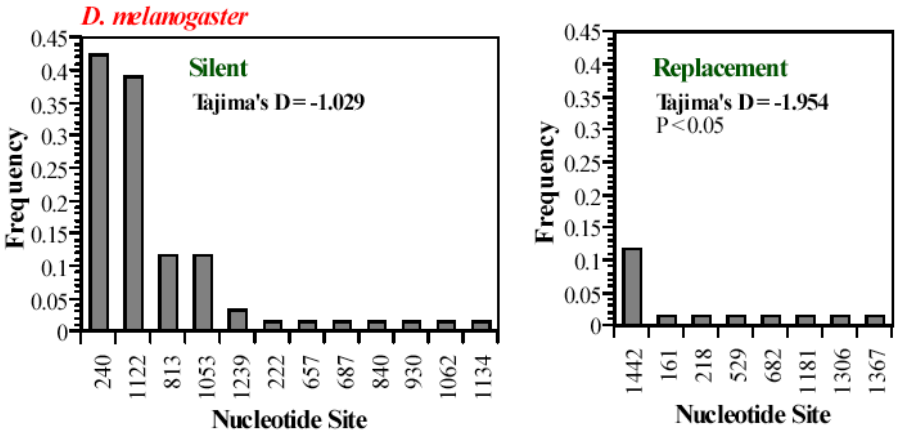
NI >> 1 in comparisons of mitochondrial genes

"mildly deleterious amino acid haplotypes are observed as ephemeral variants within species but do not contribute to divergence"

	Drosophila ND5		Hominid ND5[c]			
	59 mel. vs. 29 *si*II sim.		3 Humans vs. Chimp			
Replacement	15	11	51	5		
Silent	52	17	143	7		
Neutrality Index	2.24		2.00			
G_{adj}	2.659, N.S.		1.179, N.S.			
	Drosophila Cyt b[d]		Hominid Cyt[c]			
	17 mel. vs. 16 *si*II sim.		3 Humans vs. Chimp			
Replacement	1	4	27	5		
Silent	47	7	102	5		
Neutrality Index	26.86		3.78			
G_{adj}	8.505, $P < 0.005$		3.494, N.S.			
	Drosophila ND3 + ND5 + Cyt b		Hominid ND5 + ND3 + Cyt b		All Mitochondrial Proteins[c]	
	59 (17) mel. vs. 29 (16) *si*II sim.		3 Humans vs. Chimp		3 Humans vs. Chimp	
Replacement	18	16	80	21	179	31
Silent	112	29	268	25	915	55
Neutrality Index	3.43		2.81		2.88	
G_{adj}	9.446, $P < 0.002$		9.716, $P < 0.01$		17.519, $P < 0.0005$	

Rand & Kann 1996 *MBE* 13:735

polymorphisms in mitochondrial ND5 gene: lower frequencies and lower variance in frequencies for nonsynonymous polymorphisms suggest purifying selection against most replacement substitutions



$$\frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{\hat{V}\left(\hat{\theta}_T - \hat{\theta}_W\right)}}$$

Rand & Kann 1996 MBE 13:735