**Nielson 2005:**

Definitions:

Balancing/diversifying/disruptive selection: selection that increases variability within a population

Positive/Directional selection: selection acting on new advantageous mutations

$d_N/d_S > 1$

Negative/Purifying selection: selection acting against new deleterious mutations

$d_N/d_S < 1$

Neutrality test: statistical test of a model which assumes all mutations are either neutral or strongly deleterious

Frequency spectrum:

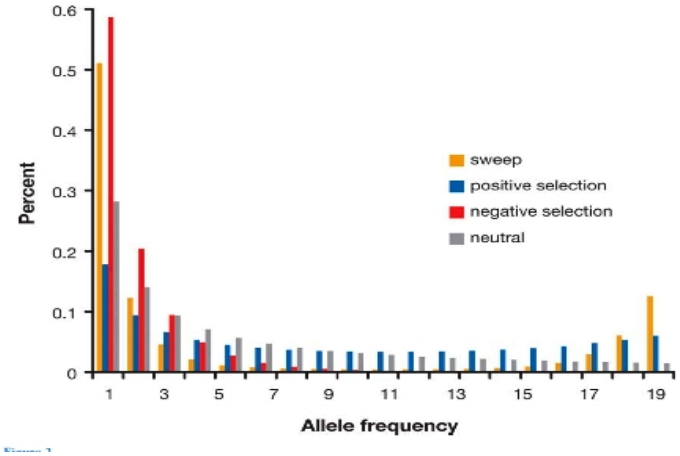
definition – count of the number of mutations that exist in a frequency of: $x_i = i/n$ for $i = 1,2...n-1$, in a sample of size n.

in standard neutral model, expected value of $x_i$ is proportional to $1/i$

Selection against deleterious mutations will increase the fraction of mutations segregating at low frequencies in the sample

A selective sweep has a similar effect

Positive selection will increase frequency in a sample of mutations segregating at high frequencies




Linkage disequilibrium(LD):

correlation among alleles from diff. loci

Reduce LD with:

old polymorphism under balancing selection

Increase LD with:

new polymorphism

selective sweep

Tests based on patterns of LD may be particularly sensitive to underlying model assumptions, because they contain strong assumptions of the underlying recombination rates

Signatures of Selection in Comparative Data

major tool: dN/dS – ratio of nonsyn mutations per nonsyn site to the number of syn mutations per nonsyn site

no selection: dN/dS =1

negative selection: dN/dS < 1

positive selection: dN/dS > 1

because negative selection will tend to dominate in evolution dN/dS is very conservative

Tests:

**Tajima's D:**

Average number of nucleotide differences between pairs of sequences is compared with the total number of segregating sites (SNPs). If the difference between these two measures of variability is larger than what is expectd on the standard neutral model, this model is rejected.

Tajima's D will frequently reject a neutral model in the presence of population growth (similar to a selective sweep)

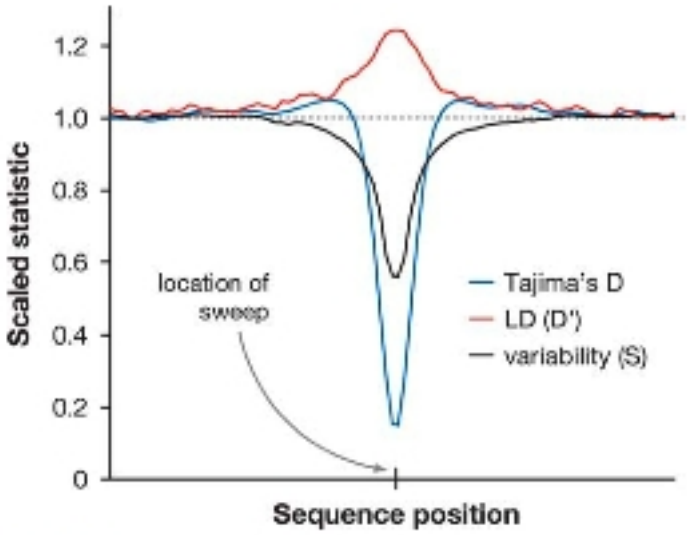



Figure 1

McDonald-Kreitman:

Nonsynonymous and synonymous mutations/polymorphisms within and between species

| | Within | Between |
|---|---|---|
| Syn | a | b |
| Non-syn | c | d |

If ratios of non-syn to syn muations differ significantly, this can provide evidence of selection

MK test is robust to any demographic assumption

Not very suitable for detecting recent selective sqeeps because both nonsyn and syn mutations linked to the beneficial mutation will be affected similarly

MK test cannot distinguish between past and present selection

Population Genetic Signatures of Selection:

(population genetic because it looks at variation within populations not species)

Population Differentiation:

Selection may increase differentiation between populations

When a locus shows extraordinary levels of differentiation compared with other loci, this may be interpreted as evidence for positive selection

Table 2 A very incomplete list of methods for detecting selection from DNA sequence and SNP data

| Test | Data | Pattern | Requires multiple loci | Robust to demographic factors? | References |
|---|---|---|---|---|---|
| Tajima's D and related | Population genetic data | Frequency spectrum | No | No | (28, 32–34, 112) |
| Modeling of selective sweep—spatial pattern | Population genetic data | Frequency spectrum/spatial pattern | No | No | (55, 56) |
| Tests based on LD | Population genetic data | LD and/or haplotype structure | No | No | (4, 23, 47, 54, 87) |
| $F_{ST}$ based and related tests | Population genetic data | Amount of population subdivision | Yes | No[a] | (1, 9, 10, 53, 92, 114) |
| HKA test | Population genetic and comparative data | Number of polymorphisms/substitutions | Yes | No | (48) |
| Macdonald-Kreitman-type tests | Population genetic and comparative data | Number of nonsynonymous and synonymous polymorphisms | No | Yes | (16, 69) |
| $d_N/d_S$ ratio tests | Comparative data or population genetic data without recombination (6) | Nonsynonymous and synonymous substitutions | No | Yes | (49, 78, 104, 123, 128, 129) |

**Perry 2007:**

**Individuals from populations with high-starch diets have more AMY1 copies than those with traditionally low-starch diets, a possible example of positive selection on a copy number-variable human gene.**

Estimated diploid AMY1 copy number using quantitative PCR (genes with higher copy number amplified faster)

Compared the extent of population differentiation of the AMY1 locus to other loci in the genomes of Yakut and Japanese populations.

-extent of Japanese-Yakut differentiation exceeded that for >97% of the microsat loci
-suggests natural selection has shaped AMY1 copy number variation in either the Japanese, Yakut, or both

Schlenke 2004:

**A chromosomal segment from California D. simulans shows extremely reduced heterozygosity but typical levels of divergence between species, suggesting strong directional selection(a "selective sweep"). The authors examined African simulans, which had high levels of polymorphism, and California D. melanogaster, which also had one locus in the segment with reduced heterozygosity, despite being highly diverged from simulans.**

**The locus contained a gene that encodes for a protein that is involved in DDT resistance, and the only mutation meeting the criteria for a selected site was a transposon upstream of the coding region in the California simulans.(the melanogaster had a different transposon in the same area).**

**One potential candidate is a gene associated with a transposon that is itself associated with a phenotype linked to insecticide resistance in D. melanogaster.**

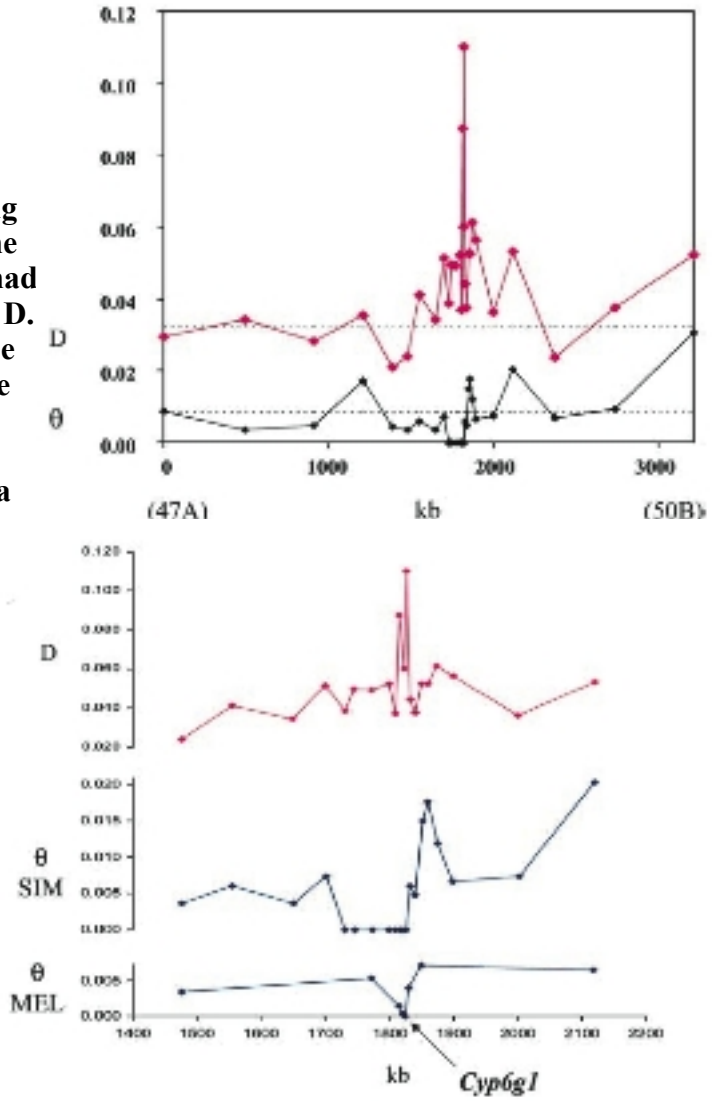



Fig. 2. Heterozygosity ($\theta$) and divergence ($D$) within and between California populations of *D. simulans* (SIM) and *D. melanogaster* (MEL) near the *Cyp6g1* locus.

**Kreitman 1994**

The neutral theory asserts that the great majority of evolutionary changes at the molecular level, as revealed by comparative studies of protein and DNA sequences, are caused not by Darwinian selection but by random drift of selectively neutral mutants (ie, mutants with a selection coefficient much smaller in absolute value than $1/2N_e$

The neutral theory is dead:

1.closest we can come to Ne is theta = 4Neu (under strict neutrality), which gives us an estimate of 10^6 for Ne, which means that s must be very close to zero

the problem: the high level of polymorphism in Drosophila(reason for the neutral theory) also tells us we can't get experimentally close to the threshold b/t drift and selection (because it's so tiny)

2.weak selection(only slightly greater than 1/Ne) is sufficient to have driven the evolution of highly biased codon usage in Drosophila

3. Protein evolution is likely not neutral

-unexpectedly high variance in rate of protein evolution (overdispersion of

the molecular clock)
-protein mutation rate of silent sites only weakly dependent on generation length
-positive selection drives protein evolution (McD-K test)

Long live the neutral theory:
-useful conceptual framework for thinking about molecular variation and evolution
-viable null alternative to selection
-(quasi-neutral) noncoding variation provides crucial barometer for detecting selection acting at linked sites; critical for success of the molecular evolutionary paradigm

**Andolfatto 2005:**

Several classes of noncoding DNA in Drosophila are evolving considerably slower than synonymous sites(hallmark of selective constraint), and yet show an excess of between-species divergence when compared with synonymous sites(signature of adaptive evolution), thus indicating that adaptive changes in noncoding DNA might have been more common in the evolution of D. melanogaster.

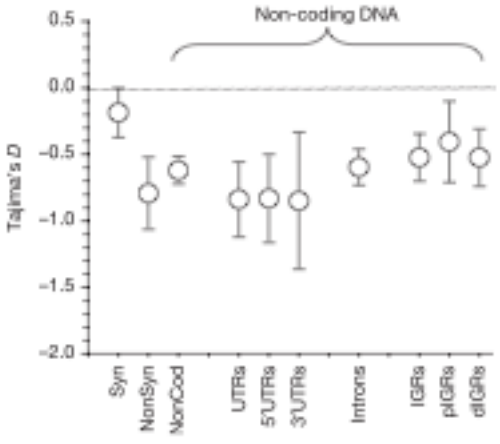


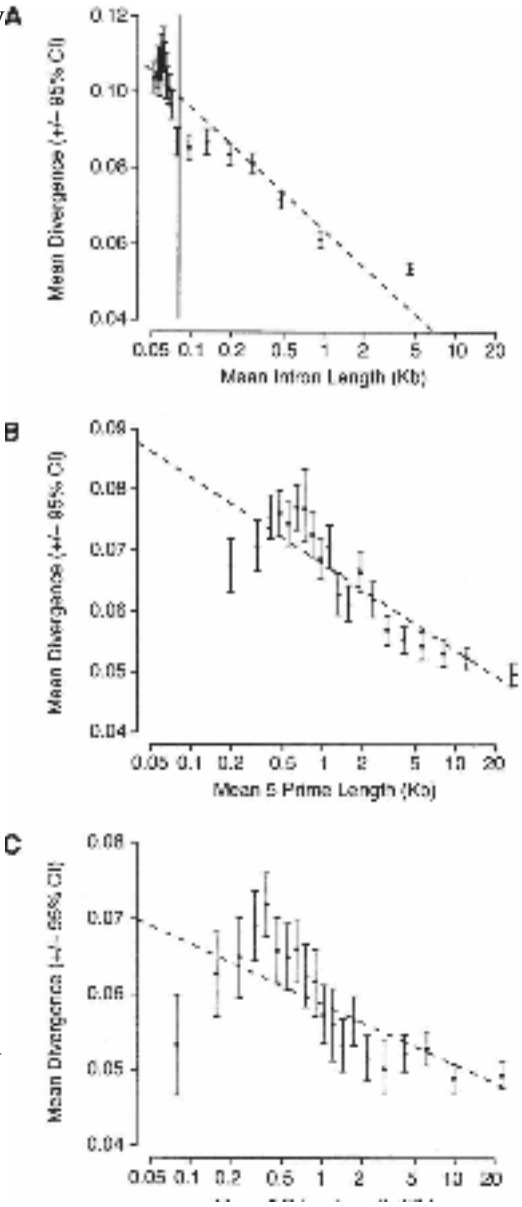

**Figure 1 | Mean Tajima's *D* values for coding and non-coding DNA.** Means across loci are given with bars indicating two standard errors. The expectation of *D* under the neutral model is shown as a dotted line. Syn, synonymous sites; NonSyn, non-synonymous sites; NonCod, pooled non-coding DNA.

Used combination of population variability and comparative genomic(ie. species) data
Noncoding regions have reduced polymorphism: functional constraint or lower mutation rate?
look at polymorphism frequencies: negative selection would keep them at lower than expected values if they were neutral
result: Tajima's D shows non-syn values negatively skewed relative to syn sites, suggesting that noncoding regions subject to purifying selection
Reduced polymorphism + purifying selection suggests that noncoding mutations sujbect o stronger negative selection than synonymous sites
Use extension of McK-D test to show that a substantial fraction of nucelotide divergence due to positive selection, especially for UTRs
These results suggests that large fraction of the non-translated genome is both important and subject to adaptive evolution

Halligan 2006

Selective constraint: fraction of mutations removed by selection

Constraint is positively correlated with intronic and intergenic sequence length and is strong in non-coding DNA, so more than half of all point mutations in the Drosophila genome are deleterious. The results also suggest blocks of constrained nucleotides concentrated in long non-coding sequences, probably related to expression control. There appears to be 3x the amount of functional noncoding DNA as coding DNA, and most deleterious mutations occur in noncoding DNA, and may make an important contribution to a wide array of evolutionary processes.




Divergence and intron size:
- Negative correlation b/t intron length and divergence
- Divergence decreases substantially at around 80bp, indicating constrained blocks within the large intron class

Divergence and intergenic size:
- Negative correlation between intergenic sequence length and divergence
- Divergence peaks at 500bp, then decreases as sequences get shorter
- Potential reason: constrained UTRs in the intergenic regions

Faster evolving sites: 2 types
- FEI (fastest evolving intronic sites) – base pairs 8-30 in introns
- FEF (fourfold degenerate sites)
- Neither type significantly different from the other; appears that FE sites are subject to little selective constraint

Constraints in introns:
- significant positive constraint in long introns at all non-splice sites
- significant positive constraint at sites involved in splicing
- Positive correlation between intron length and constraint due to even distrubition of constraint across long introns

Constraints in intergenic sequences:
- Larger intergenic sequences had more constraint
- Higher constraint in UTRs than non-UTR flanking regions
- Constrained elements within intergenic DNA are relatively evenly dispered

Genome-wide estimates of constraint:

mean constraint suggests that >50% of new mutants are removed by selection in long introns and intergenic sequences
constraint only marginally higher in coding than noncoding

Clustering of substitutions:
Highly conserved noncoding sequences are nonrandomly distributed in both Drosophila

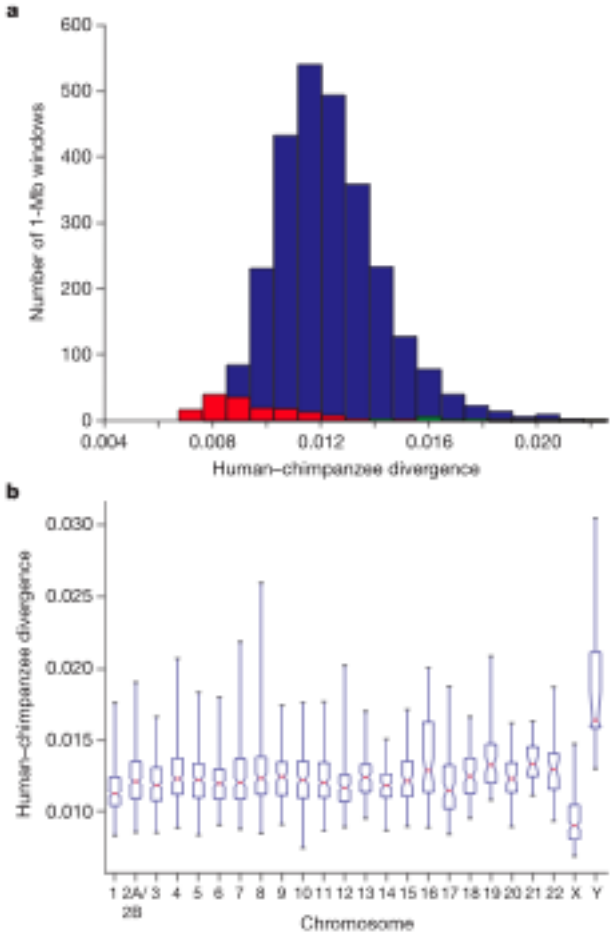
Estimate that the genomic deleterious mutation rate per diploid is U=1

Chimp Genome 2005:

Definitions:
Divergence time for orthologues:
t1(time since speciation, constant across loci) + t2 (coalescence time for orthologues within the common ancestral population, random variable that fluctuates across loci)

Variation in divergence in regions of the genome could reflect:
mutation rate differences (likely)
genetic drift
positive selection
negative selection

Sex chromosomes are outliers:
Y high divergence, X low
likely explanation: higher germ line mutations in males
ratio of male/female mutation rates(alpha) = ~3-6
this would affect mutations from DNA replication, but not DNA damage; calculated alpha for CpG, got ~2




Contribution of CpG dinucleotides:
CpG dinucleotides divergence 10x higher than other bases but regional CpG and non-CpG divergence is highly correlated, suggesting that higher-order effects modulate these processes

Increased divergence in distal regions, possibly because large-scale chromosomal structure influences regional divergence patterns.
X chromosome has a skewed distribution of high and low values:

- low values: greater purifying selection b/c of hemizygous males
- high values: increased adaptive selection also from hemizygosity, if a considerable proportion of advantageous alleles are recessive

Why is the slope of P(ancestral)/allele freq less than 1 in humans?

- likely because of population bottlenecks

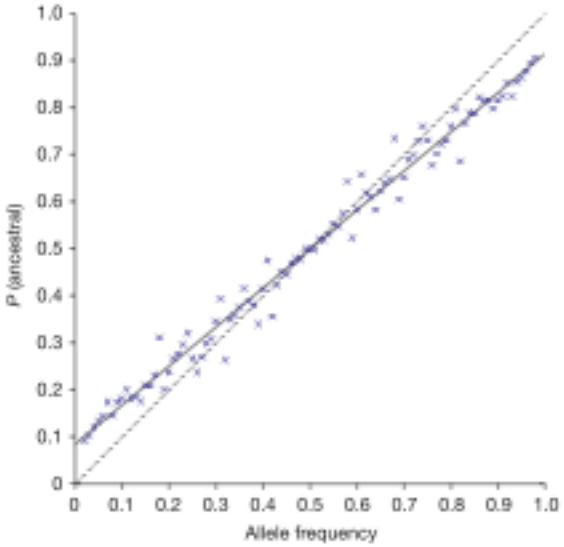



**Figure 13 | The observed fraction of ancestral alleles in 1% bins of observed frequency.** The solid line shows the regression ($b = 0.83$). The dotted line shows the theoretical relationship $p_A(x) = x$. Note that because each variant yields a derived and an ancestral allele, the data are necessarily symmetrical about 0.5.

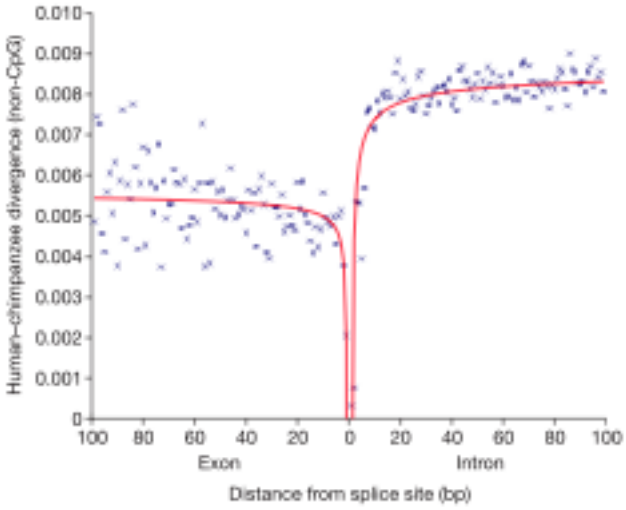



**Figure 10 | Purifying selection on synonymous sites.** Mean divergence around exon boundaries at non-CpG, exonic, fourfold degenerate sites and intronic sites, relative to the closest mRNA splice junction. The divergence rate at exonic, fourfold degenerate sites is significantly lower than at nearby intronic sites (Mann–Whitney $U$-test; $P < 10^{-27}$), suggesting that purifying selection limits the rate of synonymous codon substitutions.