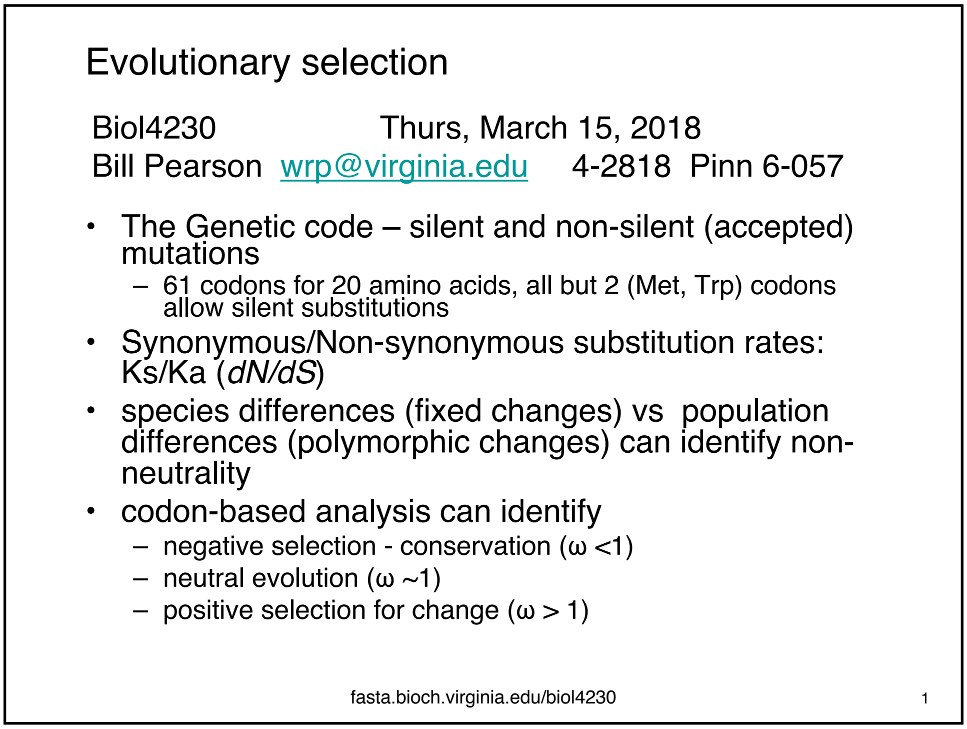
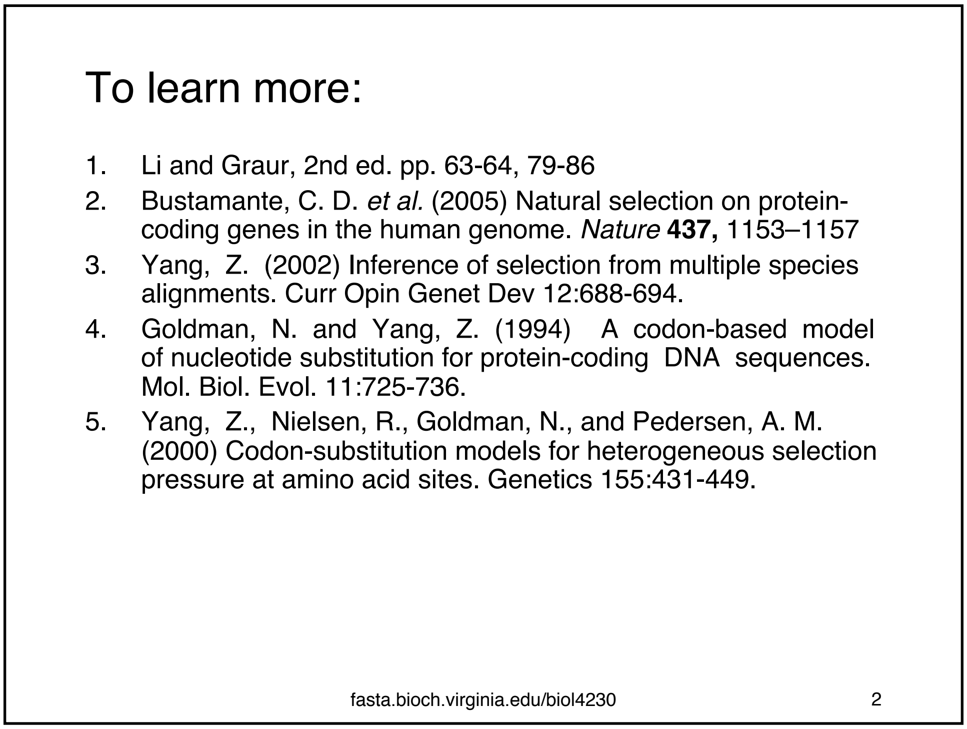


Evolutionary selection

Biol4230 Thurs, March 15, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

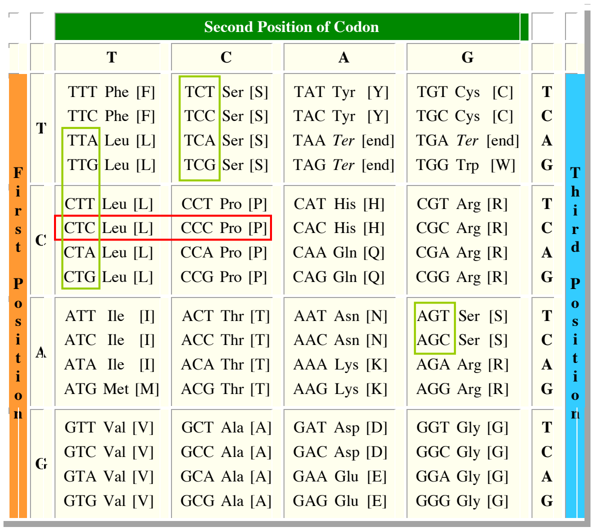
- The Genetic code – silent and non-silent (accepted) mutations
  - 61 codons for 20 amino acids, all but 2 (Met, Trp) codons allow silent substitutions
- Synonymous/Non-synonymous substitution rates: Ks/Ka (*dN/dS*)
- species differences (fixed changes) vs population differences (polymorphic changes) can identify non-neutrality
- codon-based analysis can identify
  - negative selection - conservation (ω <1)
  - neutral evolution (ω ~1)
  - positive selection for change (ω > 1)



To learn more:

1. Li and Graur, 2nd ed. pp. 63-64, 79-86
2. Bustamante, C. D. *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature* **437,** 1153–1157
3. Yang, Z. (2002) Inference of selection from multiple species alignments. Curr Opin Genet Dev 12:688-694.
4. Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725-736.
5. Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. M. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

The Genetic Code



		Second Position of Codon					
		T	C	A	G		
First Position	T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T	Third Position
		TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C	
		TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A	
		TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G	
	C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T	
		CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C	
		CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A	
		CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G	
	A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T	
		ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C	
		ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A	
		ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G	
	G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T	
		GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C	
		GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A	
		GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G	

Silent (dS, Ks)

Non-synomymous (dN, Ka) (accepted)

Positive selection for change

```
GTM1_HUMAN       R  F  L  P  R  P  V  F  S  K  M  A  V  W  G  N  K    217
gtm1_human      cgcttcctcccaagacctgtgttctcaaagatggctgtctggggcaacaag   651
GTM4_HUMAN       R  F  L  P  K  P  L  Y  T  R  V  A  V  W  G  N  K
gtm4_human      cgcttcctcccaaaacctctgtacacaagggtggctgtctggggcaacaag
GTM2_HUMAN       R  F  L  P  R  P  V  F  T  K  M  A  V  W  G  N  K
gtm2_human      cgcttcctcccaagacctgtgttcacaaagatggctgtctggggcaacaag
GTM5_HUMAN       Q  F  L  R  G  L  L  F  G  K  S  A  T  W  N  S  K
gtm5_human      caattcctccgaggtcttttgtttggaaagtcagctacatggaacagcaaa
GTM7_MOUSE       R  F  L  P  R  P  M  F  T  K  M  A  T  W  G  S  N
gtm7_mouse      cgcttcctcccaagacccatgttcacaaagatggcaacttggggcagcaat
GTM2_MOUSE       R  F  L  S  K  P  I  F  A  K  M  A  F  W  N  P  K
gtm2_mouse      cgcttcctctccaagccaatctttgcaaagatggcctttttggaacccaaag
GTM1_MOUSE       R  Y  I  A  T  P  I  F  S  K  M  A  H  W  S  N  K
gtm1_mouse      cgctacatcgcaacacctatatttcaaagatggcccactggagtaacaag
GTM3_MOUSE       R  F  L  P  R  P  V  F  T  K  I  A  Q  W  G  T  D
gtm3_mouse      cgcttcctcccaagacctgtgtttactaagatagcccagtggggcactgat
GTM6_MOUSE       R  F  L  P  S  P  V  Y  L  K  Q  A  T  W  G  N  E
gtm6_mouse      cgcttccttccaagtcctgtgtacttaaaacaggccacgtggggcaatgag
                 :  :  *  :        :  :     :     .     *  .     .
                ** *   *       *   * *     *     **    ***       *
model 2                   +  +           *           *
model 3                   *  *     +     *     +     *
```

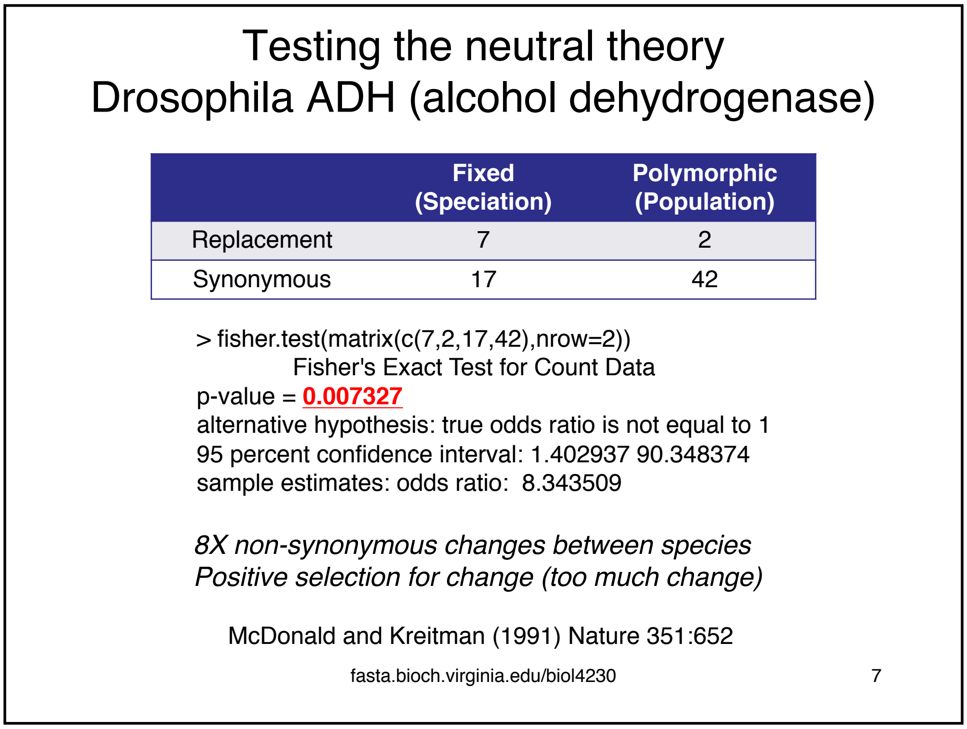
Observed Non-synonymous and Synonymous Mutation Rates

- Codon substitutions are either silent (redundancy of genetic code yields *synonymous* residue) or amino acid altering (*nonsynonymous, accepted*)
- Rate of observed *synonymous* (*dS*) mutations is similar to mutation rate of noncoding DNA
- *Nonsynonymous* mutation rate (*dN*) is lower at conserved positions, e.g. catalytic active site residues, structural determinants (purifying selection)

Testing the neutral theory

- Neutral theory of evolution (mutation)
  - most mutations are neutral, they have no effect on "fitness" (random drift)
  - deleterious mutations are rapidly lost; what is left has a very small effect
- McDonald-Kreitman test for neutrality
  - the ratio of silent/non-silent substitutions between species should match the ratio within a species
  - if not, positive or negative selection

McDonald and Kreitman (1991) Nature 351:652



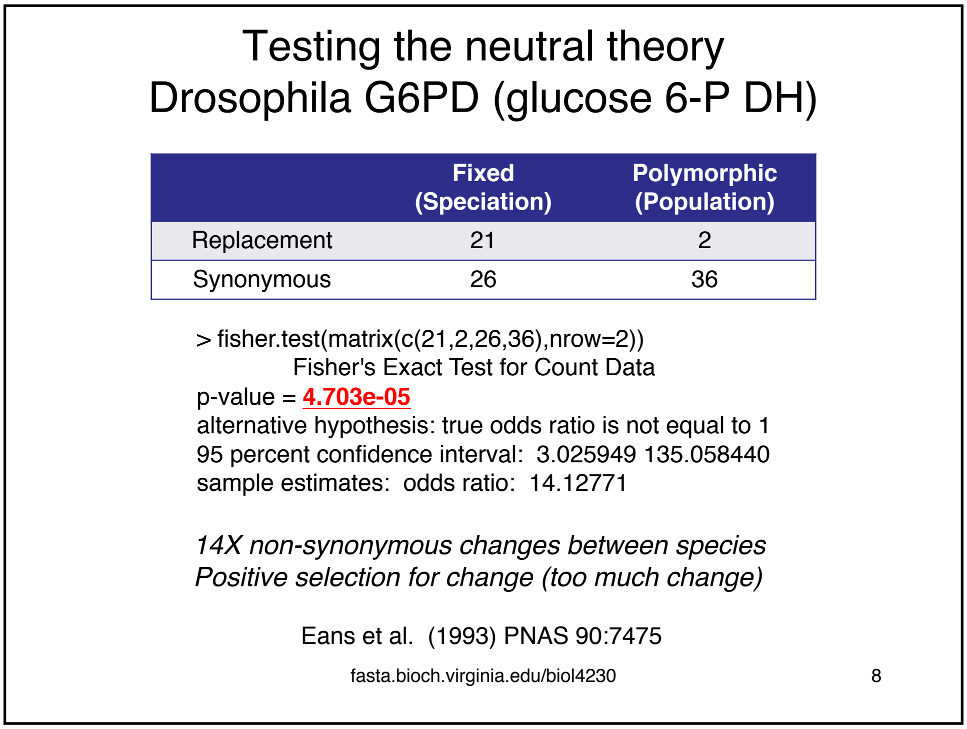
Testing the neutral theory Drosophila ADH (alcohol dehydrogenase)

	Fixed (Speciation)	Polymorphic (Population)
Replacement	7	2
Synonymous	17	42

```
> fisher.test(matrix(c(7,2,17,42),nrow=2))
         Fisher's Exact Test for Count Data
p-value = 0.007327
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval: 1.402937 90.348374
sample estimates: odds ratio:  8.343509
```

8X non-synonymous changes between species
Positive selection for change (too much change)

McDonald and Kreitman (1991) Nature 351:652



Testing the neutral theory Drosophila G6PD (glucose 6-P DH)

	Fixed (Speciation)	Polymorphic (Population)
Replacement	21	2
Synonymous	26	36

```
> fisher.test(matrix(c(21,2,26,36),nrow=2))
         Fisher's Exact Test for Count Data
p-value = 4.703e-05
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:  3.025949 135.058440
sample estimates:  odds ratio:  14.12771
```

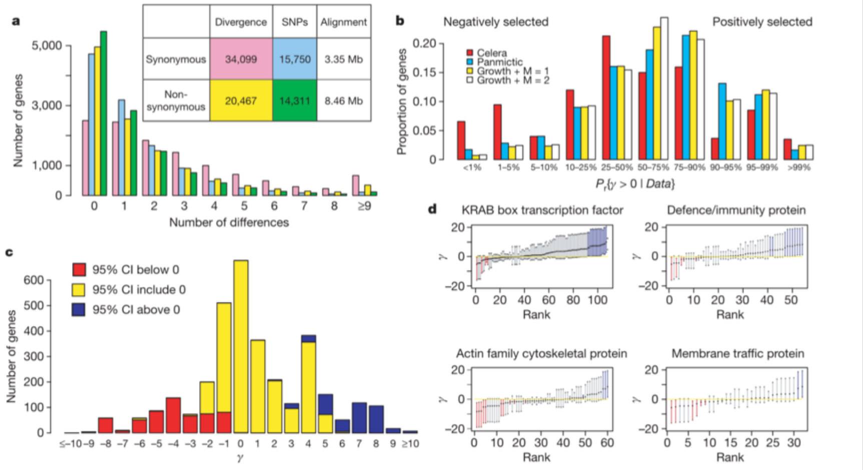
14X non-synonymous changes between species
Positive selection for change (too much change)

Eans et al. (1993) PNAS 90:7475

Natural selection on protein-coding genes in the human genome (2005) Nature 437:1153

- Sequenced 39 humans (20 European, 19 African), 1 chimpanzee
  - 11,624 genes
  - 34,099 fixed synonymous human/chimp differences (d_S=1.02%); 20,247 fixed non-synonymous human/chimp differences (d_N =0.242%)
  - 15,750 syn, 14,311 non-syn SNPs among humans (p_S=0.470%, p_N=0.169%)
  - dN/dS=23.76%, pN/pS=38.42%, excess of variation/vs divergence=> weak selection
  - 304/3,277 (9%) showed positive selection (too much change)
  - 813/6,033 (13.5%) showed negative selection (too little change)

Selection in humans



	Divergence	SNPs	Alignment
Synonymous	34,099	15,750	3.35 Mb
Non-synonymous	20,467	14,311	8.46 Mb

Figure 1 | Summary distributions of McDonald-Kreitman cell entries and mkprf analyses. **a,** Distributions of synonymous and non-synonymous SNPs and fixed differences across 11,624 genes. **b,** Distribution of the posterior probability that a gene is positively selected for simulated data under three neutral demographic scenarios and for the Celera data conditioning on a gene having at least four variable amino acid positions (IPS data). **c,** Distribution of the estimated average selection coefficient for the 3,277 genes in the Celera IPS data set. **d,** Distribution of γ for four molecular functions showing an excess of non-neutral loci. Bars represent 95% CIs with blue and red denoting CIs completely above or below $\gamma = 0$.

Phylogenetic alignment predicts less selection

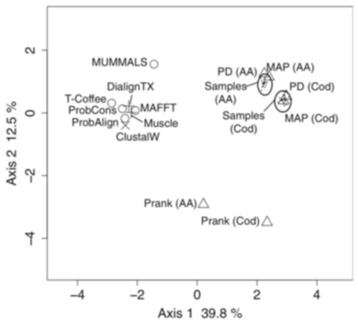


FIG. 1. PCoA plot of mean alignment distances (d_{evol}) for alignments made across 200 data sets from The Adaptive Evolution Database. "PD" and "MAP" refer to the BAli-Phy posterior decoding and maximum a posteriori summary alignments. "Samples" refers to the 20 samples taken from each BAli-Phy run.

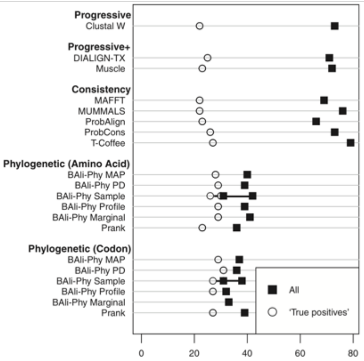


FIG. 3. Total number of families (out of 200) inferred to be under adaptive evolution (P 0:05) found, and the number of families that agree with the BAli-Phy Marginal Codon estimate (putative "true positives," see text).

Blackburne and Whelan (2012) Mol. Biol. Evol 30:642

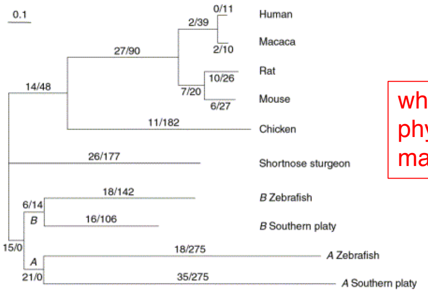
Selection in populations

- McDonald-Kreitman test compares "fixed" mutations (between species) with "variable" mutations (polymorphic, within a population)
  - dN/dS > pN/pS suggests selection *for* change (high dN/dS)
  - dN/dS < pN/pS suggests selection *against* change (low dN/dS)
- In Drosophila populations (very short generation time), many genes appear to be changing fast (dN/dS > pN/pS)
- In humans, see both positive and negative selection

Selection at codons (amino acid sites) Nonsynonymous / Synonymous Mutation Rate Ratio ω

- ω = 0: purifying selection (no aa change)
- 0 < ω < 1: biased selection
  - Varying preference for certain residues (structural residues, binding site, etc); some mutations deleterious, others tolerated
  - Most residues fall into this class
- ω = 1: neutral evolution (non-syn=syn)
- ω > 1: adaptive selection (positive selection for change)

Adaptive selection on branches



Current Opinion in Genetics & Development

where on this tree will phylogenetic alignment matter?

Fig. 1. The phylogeny of the TPI genes. Branch *A* represents gene duplication leading to the new A isozyme. The unrooted tree is used in the analysis, although the root is most likely to be along the branch ancestral to chicken and mammals [22]. The branch lengths are measured by the expected number of nucleotide substitutions per codon, estimated under the free-ratio model which estimates one for each branch. The numbers along each branch are the likelihood estimates of nonsynonymous and synonymous changes (*n**/*s**) under the same model. Estimates under other models are listed in Table 1 for branch *A*.

Yang (2002) Curr Opin Genet Dev. 12:688-94

Adaptive selection at sites

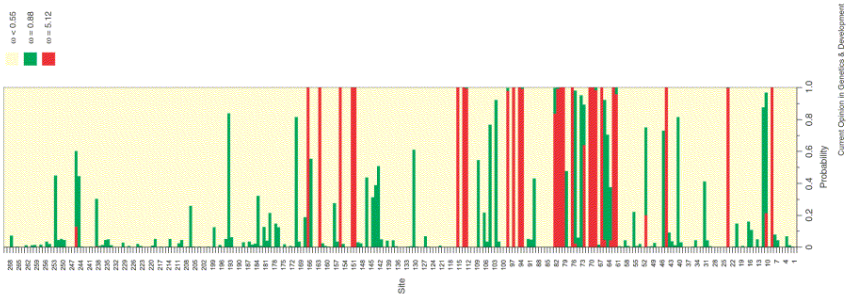


Fig. 2. Posterior probabilities of site classes for sites along the MHC class I gene. A dataset of 192 alleles from the human class I MHC alleles was analysed under the random-sites model M8 (beta&). Maximum likelihood parameter estimates suggest 90.0% of conserved sites with ratios from the distribution *B*(*p* = 0.168, *q* = 0.710) and 10.0% of positive selection sites with = 5.122. Ten equal-probability categories are used to approximate the distribution [31], with ratios of 0.000, 0.000, 0.000, 0.003, 0.015, 0.048, 0.128, 0.286, 0.548, 0.881, and 5.122. The first nine categories are collapsed into one category represented by < 0.55. Site numbering is according to the structure file 1AKJ in Protein Data Bank (chain A). From [27].

Yang (2002) Curr Opin Genet Dev. 12:688-94

Table 1. Basic statistics for data sets analyzed in this article

Data set	*s*	*n*			*S*	PS
D1: mitochondrial gene from hominoids	7	3331	14.25	0.041	2.79	Y
D2: ß-globin gene from vertebrates	17	144	2.07	0.237	7.12	Y
D3: Drosophila alcohol dehydrogenase (*adh*) gene	23	254	1.58	0.094	4.20	N
D4: flavivirus E-glycoprotein gene	22	490	3.94	0.052	12.36	N
D5: human influenza virus A hemagglutinin (HA) gene	28	329	4.62	0.391	0.85	Y
D6: HIV-1 *vif* gene	29	192	3.72	0.644	2.88	Y
D7: HIV-1 *pol* gene	23	947	4.89	0.196	1.18	Y
D8: Japanese encephalitis *env* gene	23	500	9.52	0.051	2.54	N
D9: tick-borne flavivirus NS-5 gene	18	342	2.25	0.025	26.13	N
D10: HIV-1 *env* gene V3 region	13	91	2.47	0.901	1.76	Y

s, number of sequences; *n*, number of codons in the sequence; , transition/transversion rate ratio (/ß in the notation of Kimura 1980); , nonsynonymous/synonymous rate ratio, averaged over sites (d_N/d_S); *S*, tree length, measured by the number of nucleotide substitutions along the tree per codon; PS, positive selection; Y, yes; N, no.

Yang et al. (2000) Genetics 155:431-49

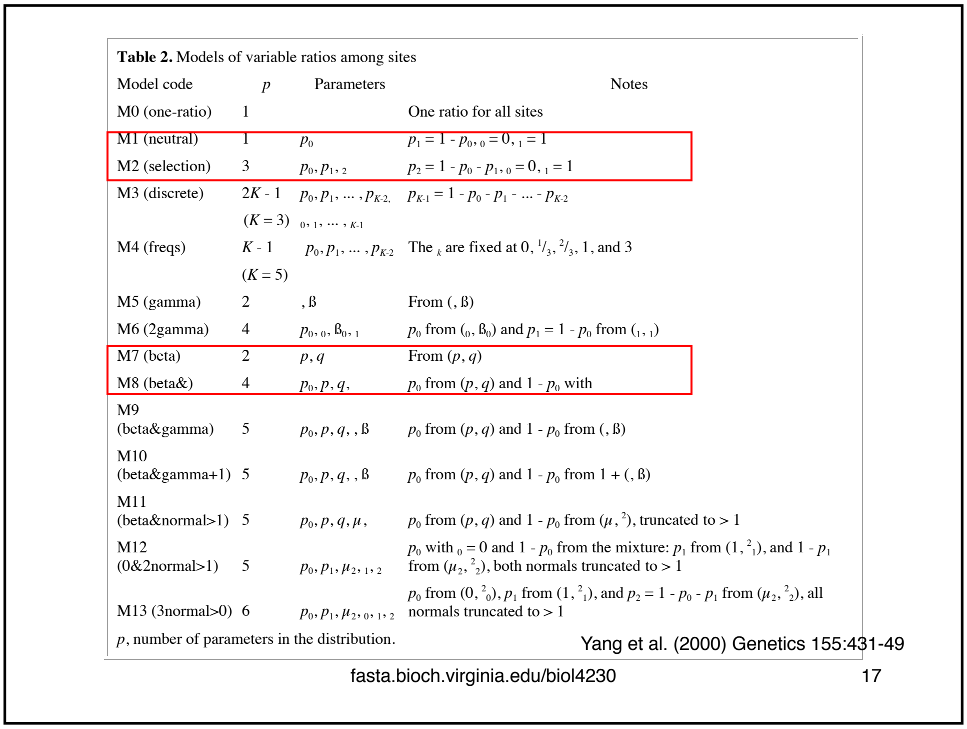


Table 2. Models of variable ratios among sites

Model code	p	Parameters	Notes
M0 (one-ratio)	1		One ratio for all sites
M1 (neutral)	1	p_0	$p_1 = 1 - p_0$, $_0 = 0$, $_1 = 1$
M2 (selection)	3	$p_0, p_1, _2$	$p_2 = 1 - p_0 - p_1$, $_0 = 0$, $_1 = 1$
M3 (discrete)	$2K - 1$ $(K = 3)$	$p_0, p_1, \dots, p_{K\text{-}2,}$ $_0, _1, \dots, _{K\text{-}1}$	$p_{K\text{-}1} = 1 - p_0 - p_1 - \dots - p_{K\text{-}2}$
M4 (freqs)	$K - 1$ $(K = 5)$	$p_0, p_1, \dots, p_{K\text{-}2}$	The $_k$ are fixed at 0, $^1/_3$, $^2/_3$, 1, and 3
M5 (gamma)	2	, ß	From (, ß)
M6 (2gamma)	4	p_0, $_0$, ß$_0$, $_1$	p_0 from ($_0$, ß$_0$) and $p_1 = 1 - p_0$ from ($_1$, $_1$)
M7 (beta)	2	p, q	From (p, q)
M8 (beta&)	4	$p_0, p, q,$	p_0 from (p, q) and $1 - p_0$ with
M9 (beta&gamma)	5	$p_0, p, q,$, ß	p_0 from (p, q) and $1 - p_0$ from (, ß)
M10 (beta&gamma+1)	5	$p_0, p, q,$, ß	p_0 from (p, q) and $1 - p_0$ from 1 + (, ß)
M11 (beta&normal>1)	5	$p_0, p, q, \mu,$	p_0 from (p, q) and $1 - p_0$ from (μ, ^{2}), truncated to > 1
M12 (0&2normal>1)	5	$p_0, p_1, \mu_2, _1, _2$	p_0 with $_0 = 0$ and $1 - p_0$ from the mixture: p_1 from (1, $^2{}_1$), and $1 - p_1$ from (μ_2, $^2{}_2$), both normals truncated to > 1
M13 (3normal>0)	6	$p_0, p_1, \mu_2, _0, _1, _2$	p_0 from (0, $^2{}_0$), p_1 from (1, $^2{}_1$), and $p_2 = 1 - p_0 - p_1$ from (μ_2, $^2{}_2$), all normals truncated to > 1

p, number of parameters in the distribution.

Yang et al. (2000) Genetics 155:431-49

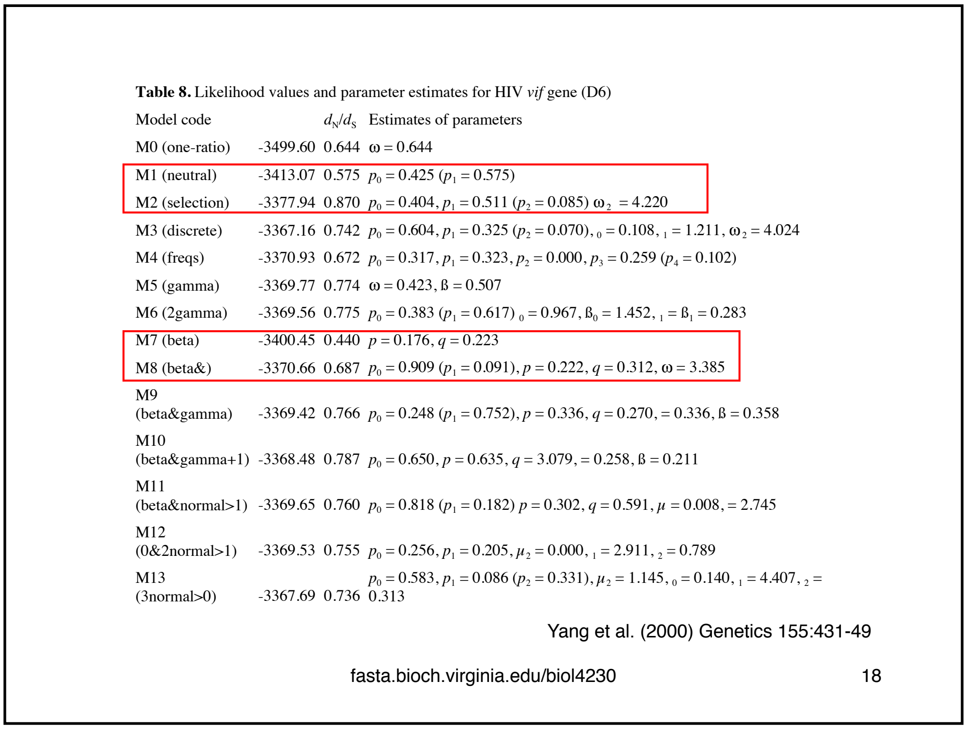
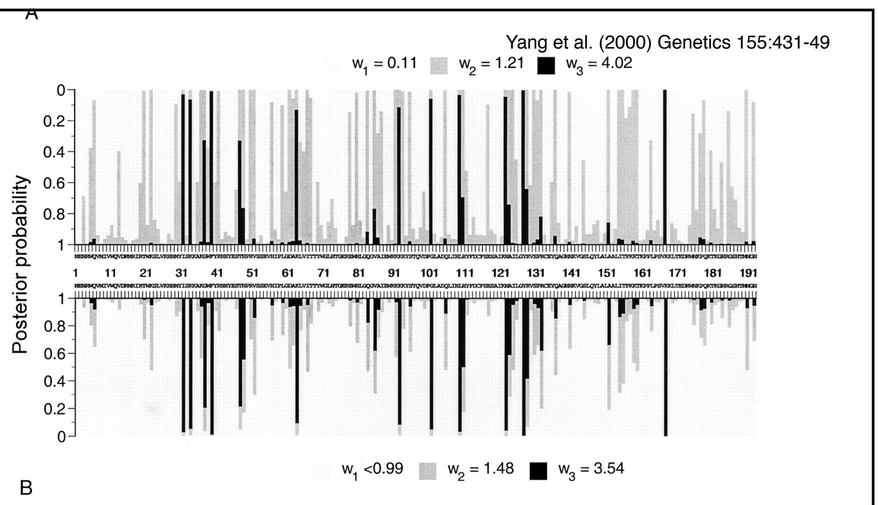


Table 8. Likelihood values and parameter estimates for HIV *vif* gene (D6)

Model code		d_N/d_S	Estimates of parameters
M0 (one-ratio)	-3499.60	0.644	$\omega = 0.644$
M1 (neutral)	-3413.07	0.575	$p_0 = 0.425$ ($p_1 = 0.575$)
M2 (selection)	-3377.94	0.870	$p_0 = 0.404$, $p_1 = 0.511$ ($p_2 = 0.085$) $\omega_2 = 4.220$
M3 (discrete)	-3367.16	0.742	$p_0 = 0.604$, $p_1 = 0.325$ ($p_2 = 0.070$), $_0 = 0.108$, $_1 = 1.211$, $\omega_2 = 4.024$
M4 (freqs)	-3370.93	0.672	$p_0 = 0.317$, $p_1 = 0.323$, $p_2 = 0.000$, $p_3 = 0.259$ ($p_4 = 0.102$)
M5 (gamma)	-3369.77	0.774	$\omega = 0.423$, ß = 0.507
M6 (2gamma)	-3369.56	0.775	$p_0 = 0.383$ ($p_1 = 0.617$) $_0 = 0.967$, ß$_0$ = 1.452, $_1$ = ß$_1$ = 0.283
M7 (beta)	-3400.45	0.440	$p = 0.176$, $q = 0.223$
M8 (beta&)	-3370.66	0.687	$p_0 = 0.909$ ($p_1 = 0.091$), $p = 0.222$, $q = 0.312$, $\omega = 3.385$
M9 (beta&gamma)	-3369.42	0.766	$p_0 = 0.248$ ($p_1 = 0.752$), $p = 0.336$, $q = 0.270$, = 0.336, ß = 0.358
M10 (beta&gamma+1)	-3368.48	0.787	$p_0 = 0.650$, $p = 0.635$, $q = 3.079$, = 0.258, ß = 0.211
M11 (beta&normal>1)	-3369.65	0.760	$p_0 = 0.818$ ($p_1 = 0.182$) $p = 0.302$, $q = 0.591$, $\mu = 0.008$, = 2.745
M12 (0&2normal>1)	-3369.53	0.755	$p_0 = 0.256$, $p_1 = 0.205$, $\mu_2 = 0.000$, $_1 = 2.911$, $_2 = 0.789$
M13 (3normal>0)	-3367.69	0.736	$p_0 = 0.583$, $p_1 = 0.086$ ($p_2 = 0.331$), $\mu_2 = 1.145$, $_0 = 0.140$, $_1 = 4.407$, $_2 =$ 0.313

Yang et al. (2000) Genetics 155:431-49

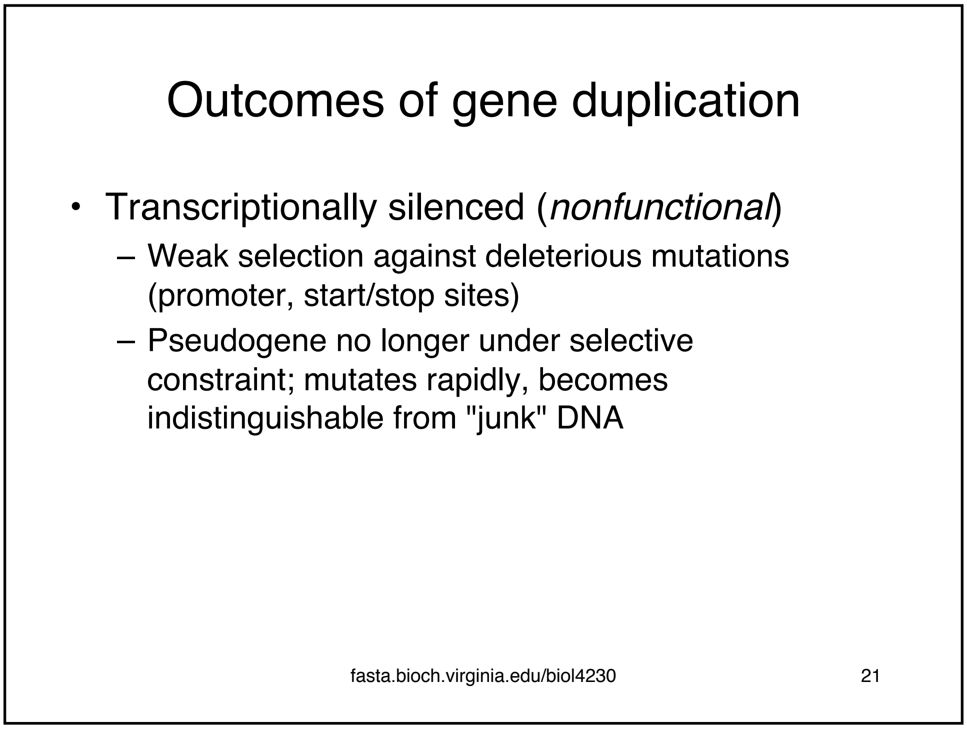


Yang et al. (2000) Genetics 155:431-49

Figure 2. Posterior probabilities of site classes along the HIV-1 *vif* gene (data set D6). (A) Model M3 (discrete) is used, which assumes three classes of sites in the gene. The estimated frequencies and ratios for the three classes are *p*0 = 0.604, *p*1 = 0.325, and *p*2 = 0.070 and 0 = 0.108, 1 = 1.211, and 2 = 4.024 (see Table 8). Those *pk* are the prior distribution for each codon (amino acid) site. The data (codon configurations in different sequences) at the site change that distribution into the posterior distribution, which may be very different from the prior. For example, the posterior probabilities for the three classes at site 1 are 0.9928, 0.0072, and 0.0000, and the site is under strong purifying selection. In contrast, the posterior probabilities at site 31 are 0.0000, 0.0301, and 0.9699, and the site is almost certainly under diversifying selection. (B) Model M9 (beta&gamma) is fitted to the data with 10 categories used to approximate the continuous mixture distribution. Estimates of parameters under the model are shown in Table 8. According to those estimates, the ratios for the 10 categories, each of probability 10%, are 0.00036, 0.00945, 0.04338, 0.11910, 0.25502, 0.46947, 0.76134, 0.98969, 1.47748, and 3.53812. The first 8 categories are combined in the plot. The amino acid sequence of one of the variants (SF2) is superimposed on the graph

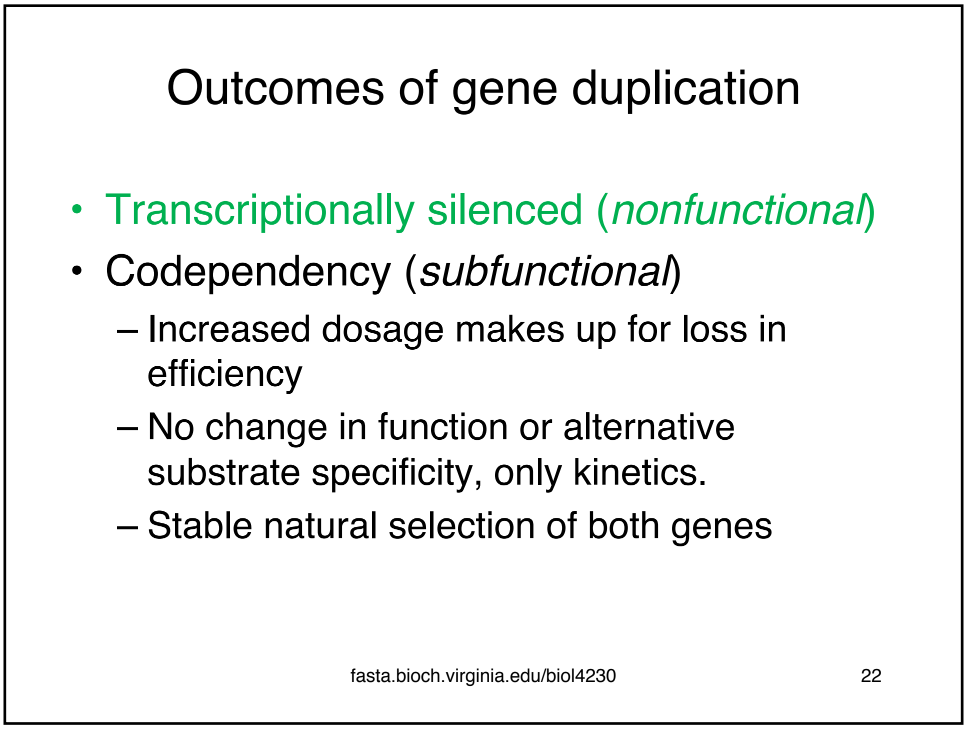
Functional inferences from paralogous glutathione transferase sequences

- Glutathione transferases: large multifunctional gene family, important in the metabolism of oxidative toxins
- All classes (alpha, mu, theta, etc.) are multigenic; divergence of classes very ancient, duplications are more recent
- Paralogs within each class have distinct substrate specificity profiles



Outcomes of gene duplication

- Transcriptionally silenced (*nonfunctional*)
  - Weak selection against deleterious mutations (promoter, start/stop sites)
  - Pseudogene no longer under selective constraint; mutates rapidly, becomes indistinguishable from "junk" DNA



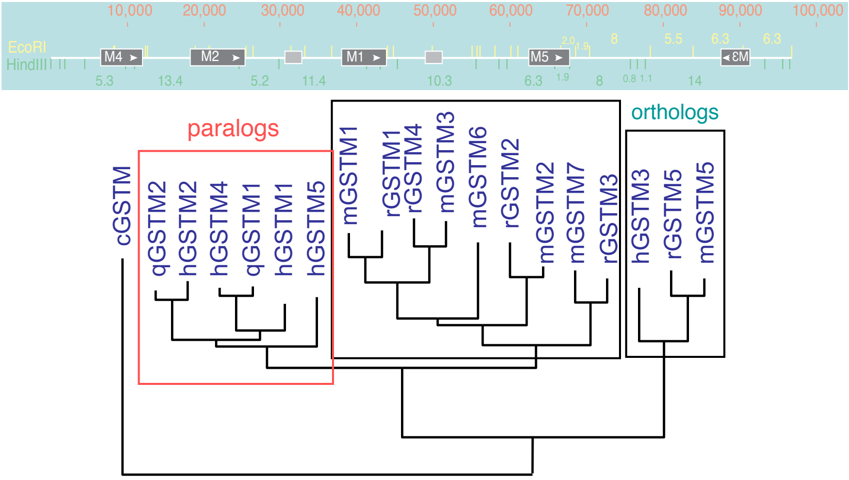
Outcomes of gene duplication

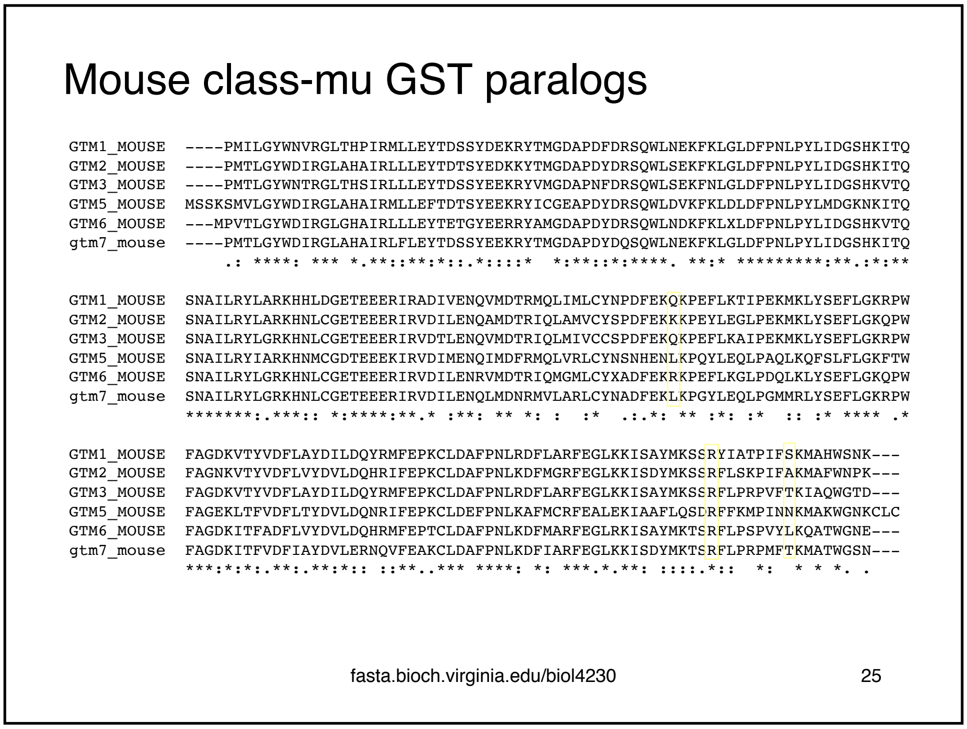
- Transcriptionally silenced (*nonfunctional*)
- Codependency (*subfunctional*)
  - Increased dosage makes up for loss in efficiency
  - No change in function or alternative substrate specificity, only kinetics.
  - Stable natural selection of both genes

Outcomes of gene duplication

- Transcriptionally silenced (*nonfunctional*)
- Codependency (*subfunctional*)
- Functional divergence (*neofunctional*)
  - Relaxed selection on redundant genes allows "exploration" of alternative function or substrate specificity.
  - Mutations that introduce novel advantageous function more likely to become fixed: *adaptive (positive) selection*

Class-mu glutathione transferase genes



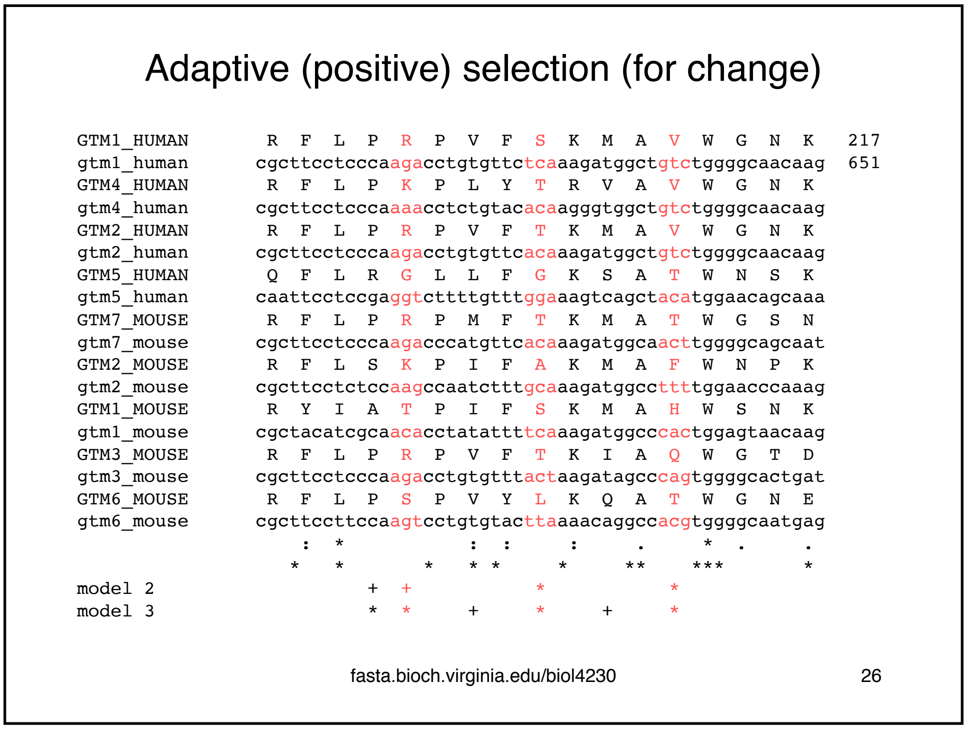


Mouse class-mu GST paralogs

```
GTM1_MOUSE  ----PMILGYWNVRGLTHPIRMLLEYTDSSYDEKRYTMGDAPDFDRSQWLNEKFKLGLDFPNLPYLIDGSHKITQ
GTM2_MOUSE  ----PMTLGYWDIRGLAHAIRLLLEYTDTSYEDKKYTMGDAPDYDRSQWLSEKFKLGLDFPNLPYLIDGSHKITQ
GTM3_MOUSE  ----PMTLGYWNTRGLTHSIRLLLEYTDSSYEEKRYVMGDAPNFDRSQWLSEKFNLGLDFPNLPYLIDGSHKVTQ
GTM5_MOUSE  MSSKSMVLGYWDIRGLAHAIRMLLEFTDTSYEEKRYICGEAPDYDRSQWLDVKFKLDLDFPNLPYLMDGKNKITQ
GTM6_MOUSE  ---MPVTLGYWDIRGLGHAIRLLLEYTETGYEERRYAMGDAPDYDRSQWLNDKFKLXLDFPNLPYLIDGSHKVTQ
gtm7_mouse  ----PMTLGYWDIRGLAHAIRLFLEYTDSSYEEKRYTMGDAPDYDQSQWLNEKFKLGLDFPNLPYLIDGSHKITQ
                .: ****: *** *.**::**:*::.*::::*  *:**::*:****. **:* *********:**.:*:**

GTM1_MOUSE  SNAILRYLARKHHLDGETEEERIRADIVENQVMDTRMQLIMLCYNPDFEKQKPEFLKTIPEKMKLYSEFLGKRPW
GTM2_MOUSE  SNAILRYLARKHNLCGETEEERIRVDILENQAMDTRIQLAMVCYSPDFEKKKPEYLEGLPEKMKLYSEFLGKQPW
GTM3_MOUSE  SNAILRYLGRKHNLCGETEEERIRVDTLENQVMDTRIQLMIVCCSPDFEKQKPEFLKAIPEKMKLYSEFLGKRPW
GTM5_MOUSE  SNAILRYIARKHNMCGDTEEEKIRVDIMENQIMDFRMQLVRLCYNSNHENLKPQYLEQLPAQLKQFSLFLGKFTW
GTM6_MOUSE  SNAILRYLGRKHNLCGETEEERIRVDILENRVMDTRIQMGMLCYXADFEKRKPEFLKGLPDQLKLYSEFLGKQPW
gtm7_mouse  SNAILRYLGRKHNLCGETEEERIRVDILENQLMDNRMVLARLCYNADFEKLKPGYLEQLPGMMRLYSEFLGKRPW
            *******:.***:: *:****:**.* :**: ** *: :  :*  .:.*: ** :*: :*  :: :* **** .*

GTM1_MOUSE  FAGDKVTYVDFLAYDILDQYRMFEPKCLDAFPNLRDFLARFEGLKKISAYMKSSRYIATPIFSKMAHWSNK---
GTM2_MOUSE  FAGNKVTYVDFLVYDVLDQHRIFEPKCLDAFPNLKDFMGRFEGLKKISDYMKSSRFLSKPIFAKMAFWNPK---
GTM3_MOUSE  FAGDKVTYVDFLAYDILDQYRMFEPKCLDAFPNLRDFLARFEGLKKISAYMKSSRFLPRPVFTKIAQWGTD---
GTM5_MOUSE  FAGEKLTFVDFLTYDVLDQNRIFEPKCLDEFPNLKAFMCRFEALEKIAAFLQSDRFFKMPINNKMAKWGNKCLC
GTM6_MOUSE  FAGDKITFADFLVYDVLDQHRMFEPTCLDAFPNLKDFMARFEGLRKISAYMKTSRFLPSPVYLKQATWGNE---
gtm7_mouse  FAGDKITFVDFIAYDVLERNQVFEAKCLDAFPNLKDFIARFEGLKKISDYMKTSRFLPRPMFTKMATWGSN---
            ***:*:*:.**:.**:*:: ::**..*** ****: *: ***.*.**: ::::.*::  *:  * * *. .
```



Adaptive (positive) selection (for change)

```
GTM1_HUMAN       R  F  L  P  R  P  V  F  S  K  M  A  V  W  G  N  K   217
gtm1_human      cgcttcctcccaagacctgtgttctcaaagatggctgtctggggcaacaag  651
GTM4_HUMAN       R  F  L  P  K  P  L  Y  T  R  V  A  V  W  G  N  K
gtm4_human      cgcttcctcccaaaacctctgtacacaagggtggctgtctggggcaacaag
GTM2_HUMAN       R  F  L  P  R  P  V  F  T  K  M  A  V  W  G  N  K
gtm2_human      cgcttcctcccaagacctgtgttcacaaagatggctgtctggggcaacaag
GTM5_HUMAN       Q  F  L  R  G  L  L  F  G  K  S  A  T  W  N  S  K
gtm5_human      caattcctccgaggtcttttgtttggaaagtcagctacatggaacagcaaa
GTM7_MOUSE       R  F  L  P  R  P  M  F  T  K  M  A  T  W  G  S  N
gtm7_mouse      cgcttcctcccaagacccatgttcacaaagatggcaacttggggcagcaat
GTM2_MOUSE       R  F  L  S  K  P  I  F  A  K  M  A  F  W  N  P  K
gtm2_mouse      cgcttcctctccaagccaatctttgcaaagatggcctttttggaacccaaag
GTM1_MOUSE       R  Y  I  A  T  P  I  F  S  K  M  A  H  W  S  N  K
gtm1_mouse      cgctacatcgcaacacctatatttttcaaagatggcccactggagtaacaag
GTM3_MOUSE       R  F  L  P  R  P  V  F  T  K  I  A  Q  W  G  T  D
gtm3_mouse      cgcttcctcccaagacctgtgtttactaagatagcccagtggggcactgat
GTM6_MOUSE       R  F  L  P  S  P  V  Y  L  K  Q  A  T  W  G  N  E
gtm6_mouse      cgcttccttccaagtcctgtgtacttaaaacaggccacgtggggcaatgag
                    :  *           :  :     :     .     *  .     .
                   *   *       *   * *     *     **    ***       *
model 2                   +  +           *           *
model 3                   *  *     +     *     +     *
```

PAML analysis of class-mu GSTs

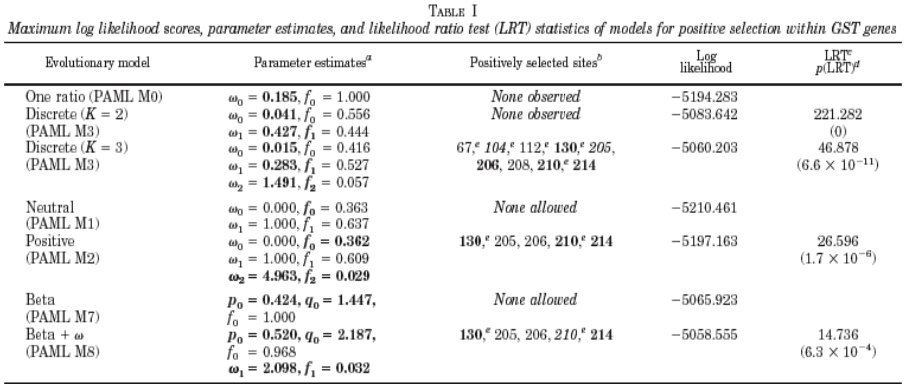
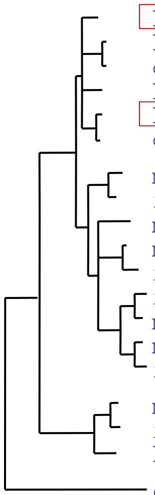


TABLE I

Maximum log likelihood scores, parameter estimates, and likelihood ratio test (LRT) statistics of models for positive selection within GST genes

Evolutionary model	Parameter estimates[a]	Positively selected sites[b]	Log likelihood	LRT[c] p(LRT)[d]
One ratio (PAML M0)	ω_0 = **0.185**, f_0 = 1.000	*None observed*	−5194.283	
Discrete (K = 2) (PAML M3)	ω_0 = **0.041**, f_0 = 0.556 ω_1 = **0.427**, f_1 = 0.444	*None observed*	−5083.642	221.282 (0)
Discrete (K = 3) (PAML M3)	ω_0 = **0.015**, f_0 = 0.416 ω_1 = **0.283**, f_1 = 0.527 ω_2 = **1.491**, f_2 = 0.057	67,[e] *104*,[e] 112,[e] **130**,[e] *205*, **206**, 208, **210**,[e] **214**	−5060.203	46.878 (6.6×10^{-11})
Neutral (PAML M1)	ω_0 = 0.000, f_0 = 0.363 ω_1 = 1.000, f_1 = 0.637	*None allowed*	−5210.461	
Positive (PAML M2)	ω_0 = 0.000, f_0 **= 0.362** ω_1 = 1.000, f_1 = 0.609 **ω_2 = 4.963, f_2 = 0.029**	**130**,[e] 205, 206, **210**,[e] **214**	−5197.163	26.596 (1.7×10^{-6})
Beta (PAML M7)	**p_0 = 0.424, q_0 = 1.447,** f_0 = 1.000	*None allowed*	−5065.923	
Beta + ω (PAML M8)	**p_0 = 0.520, q_0 = 2.187,** f_0 = 0.968 **ω_1 = 2.098, f_1 = 0.032**	**130**,[e] 205, 206, *210*,[e] **214**	−5058.555	14.736 (6.3×10^{-4})

Ivarsson (2003) J Biol Chem. 278:8733-8



gene	104	130	205	206	210	214	CDNB	cDMNG	TSO
hGSTM1	T	E	P	R	S	V	+	+	+++++
hGSTM4	A	E	P	K	T	V			
qGSTM1	A	G	P	K	T	V			
hGSTM5	V	E	R	G	G	T			
hGSTM2	F	A	P	R	T	V	+	++	+
qGSTM2	L	G	P	R	T	V			
mGSTM7	L	Q	P	R	T	T			
rGSTM3	L	Q	P	R	T	I			
mGSTM6	V	G	P	S	L	T			
mGSTM2	A	G	S	K	A	F			
rGSTM2	A	G	S	K	A	F			
rGSTM4	V	S	P	R	T	Q			
mGSTM3	V	A	P	R	T	Q			
mGSTM1	V	T	A	T	S	H			
rGSTM1	V	T	S	T	S	Q			
mGSTM5	I	Q	K	M	N	K			
rGSTM5	I	Q	K	M	N	K			
hGSTM3	V	E	K	M	N	Q			
cGSTM	L	L	K	A	W	L			

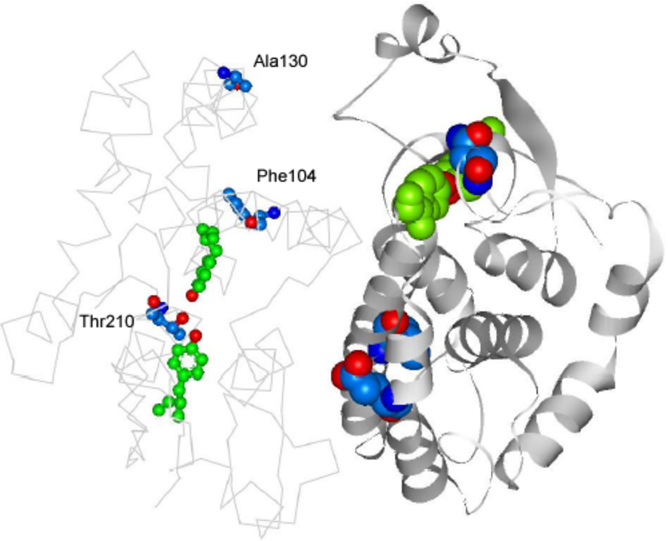
Residues selected by codon-substitution models

	210	104	130
M1-1	Ser	Thr	Ala
M2-2	Thr	Phe	Glu

Specific activities (μmol min^{-1} mg^{-1}) of wild-type and mutant human Mu class GSTs with various substrates

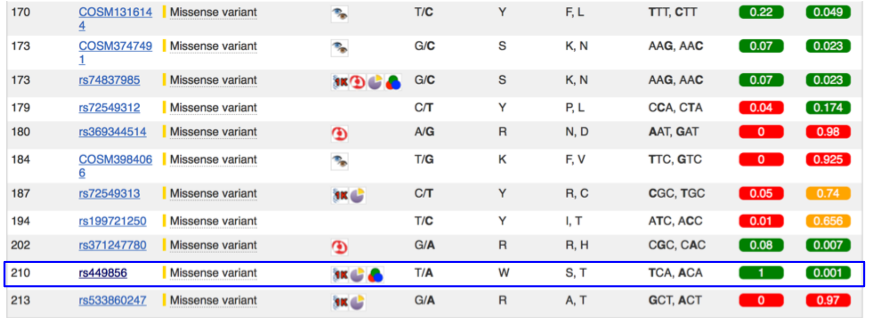
Enzyme	trans-stilbene oxide	aminochrome	CDNB
GST M2-2	0.0002 ± 0.00003	120 ± 7	426 ± 5
GST M2-2 T210S	0.17 ± 0.03	108 ± 6	482 ± 14
GST M2-2 T210S/F104T	0.19 ± 0.02	82 ± 7	547 ± 12
GST M2-2 T210S/F104T/A130E	0.28 ± 0.01	132 ± 8	600 ± 16
GST M1-1	3.00 ± 0.02	0.73 ± 0.02	136 ± 6
GST M1-1 S210T	0.026 ± 0.001	0.94 ± 0.05	112 ± 3

Ivarsson (2003) J Biol Chem. 278:8733-8



Ivarsson (2003) J Biol Chem. 278:8733-8

ENSEMBL – protein variation (missense)



170	COSM1316144	Missense variant		T/C	Y	F, L	TTT, CTT	0.22	0.049
173	COSM3747491	Missense variant		G/C	S	K, N	AAG, AAC	0.07	0.023
173	rs74837985	Missense variant		G/C	S	K, N	AAG, AAC	0.07	0.023
179	rs72549312	Missense variant		C/T	Y	P, L	CCA, CTA	0.04	0.174
180	rs369344514	Missense variant		A/G	R	N, D	AAT, GAT	0	0.98
184	COSM3984066	Missense variant		T/G	K	F, V	TTC, GTC	0	0.925
187	rs72549313	Missense variant		C/T	Y	R, C	CGC, TGC	0.05	0.74
194	rs199721250	Missense variant		T/C	Y	I, T	ATC, ACC	0.01	0.656
202	rs371247780	Missense variant		G/A	R	R, H	CGC, CAC	0.08	0.007
210	rs449856	Missense variant		T/A	W	S, T	TCA, ACA	1	0.001
213	rs533860247	Missense variant		G/A	R	A, T	GCT, ACT	0	0.97

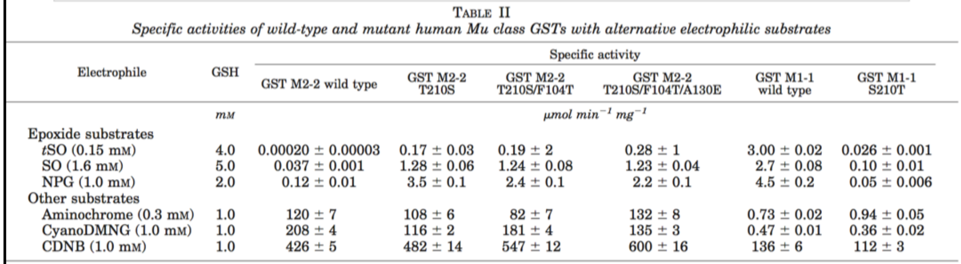


TABLE II

Specific activities of wild-type and mutant human Mu class GSTs with alternative electrophilic substrates

Electrophile	GSH	Specific activity					
		GST M2-2 wild type	GST M2-2 T210S	GST M2-2 T210S/F104T	GST M2-2 T210S/F104T/A130E	GST M1-1 wild type	GST M1-1 S210T
	mM	*μmol min⁻¹ mg⁻¹*					
Epoxide substrates							
*t*SO (0.15 mM)	4.0	0.00020 ± 0.00003	0.17 ± 0.03	0.19 ± 2	0.28 ± 1	3.00 ± 0.02	0.026 ± 0.001
SO (1.6 mM)	5.0	0.037 ± 0.001	1.28 ± 0.06	1.24 ± 0.08	1.23 ± 0.04	2.7 ± 0.08	0.10 ± 0.01
NPG (1.0 mM)	2.0	0.12 ± 0.01	3.5 ± 0.1	2.4 ± 0.1	2.2 ± 0.1	4.5 ± 0.2	0.05 ± 0.006
Other substrates							
Aminochrome (0.3 mM)	1.0	120 ± 7	108 ± 6	82 ± 7	132 ± 8	0.73 ± 0.02	0.94 ± 0.05
CyanoDMNG (1.0 mM)	1.0	208 ± 4	116 ± 2	181 ± 4	135 ± 3	0.47 ± 0.01	0.36 ± 0.02
CDNB (1.0 mM)	1.0	426 ± 5	482 ± 14	547 ± 12	600 ± 16	136 ± 6	112 ± 3

Ivarsson, Y. et al. (2003) *J Biol Chem* **278,** 8733

Evolutionary selection: dN/dS

- The Genetic code – silent and non-silent (accepted) mutations
  - 61 codons encode amino acids, 20 amino acids, all but 2 (Met, Trp) codons allow silent substitutions
- Synonymous/Non-synonymous substitution rates: Ks/Ka (*dN/dS*)
- species differences (fixed changes) vs population differences (polymorphic changes) can identify non-neutrality (McDonald-Kreitman)
- codon-based analysis can identify
  - negative selection - conservation (ω <1)
  - neutral evolution (ω ~1)
  - positive selection for change (ω > 1)