THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

Estimating the ...ers of selective swe... patterns of genetic di... wild house mice

REDUX

Tom Booker

Peter Keightley
Brian Charlesworth

Labgroup 11th May 2018

Adaptation in Protein-Coding Versus Regulatory Regions

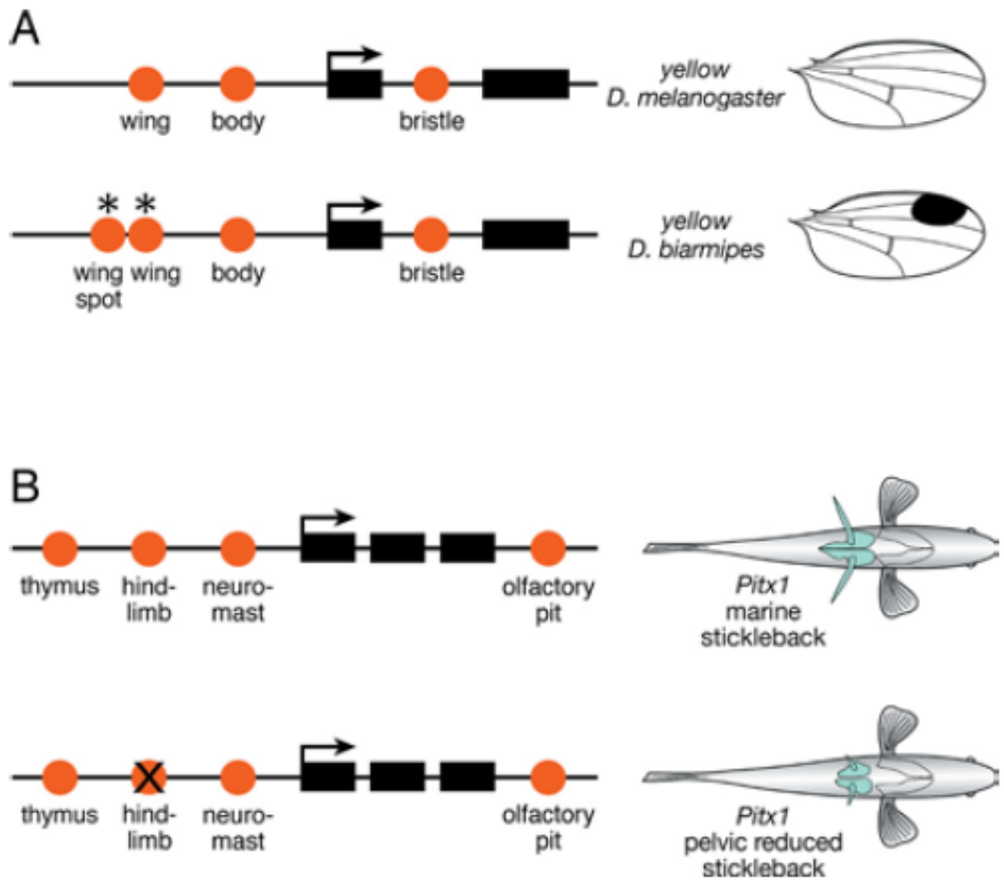
‘Their macromolecules are so alike that regulatory mutations may account for their biological differences.’

Evolution at Two Levels in Humans and Chimpanzees - King and Wilson 1975

Adaptation in Protein-Coding Versus Regulatory Regions

‘Their macromolecules are so alike that regulatory mutations may account for their biological differences.’

Evolution at Two Levels in Humans and Chimpanzees - King and Wilson 1975



Carroll 2005 PLoS Genetics

A (very) simplified model of change in population mean fitness due to adaptive evolution

Change in fitness per generation due to adaptive mutations (ΔW):

Adaptive mutations occur at a rate μ_a ($p_a\mu$)
The selection coefficients (s_a) of new adap. mutations are drawn from some distribution with probability $f(s_a)$
New mutations become fixed with probability $u(s_a)$

In the whole genome there are n_a sites at which adaptive mutations can arise, leading to the following:

$$\Delta W \propto n_a \int \mu_a u(s_a) s_a f(s_a) ds_a,$$

If we assume that selection on adaptive mutations is strong relative to drift…

$$\Delta W \propto n_a \int \mu_a u(s_a) s_a f(s_a) ds_a,$$

Leads to

$$\Delta W \propto n_a \mu_a E(s_a^2),$$

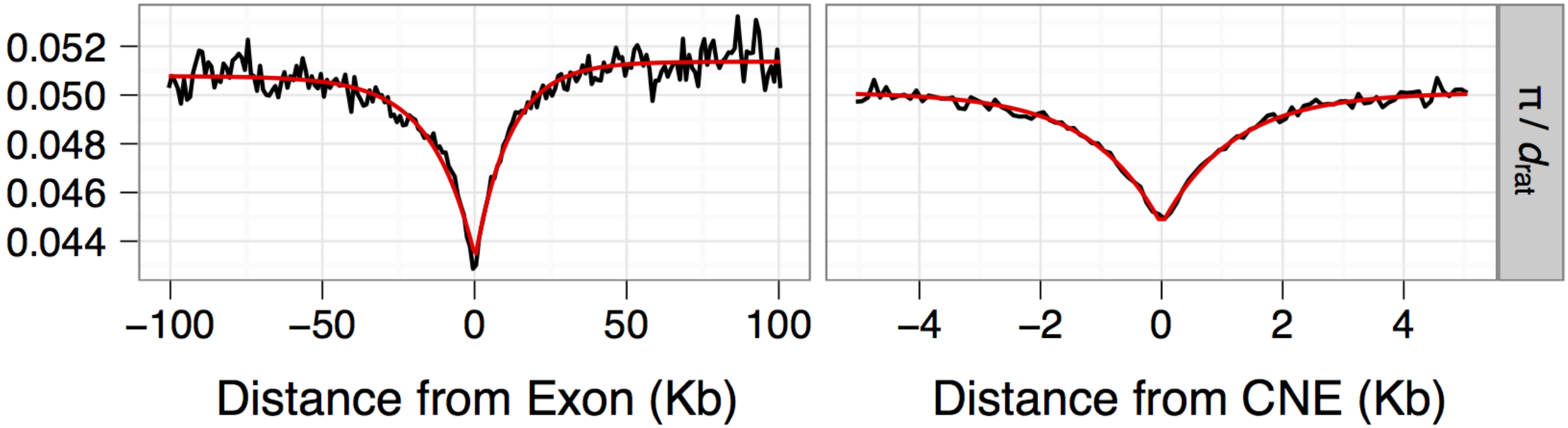
If the point mutation rate is the same for all sites in the genome, then ΔW is proportional to $n_a p_a s_a^2$

Selection at linked sites in the house mouse *Mus musculus castaneus*

Mice are an excellent model organism for studying molecular evolution in mammals

Mouse from Phifer-Rixey and Nachman 2015
(It's actually *domesticus* though, sorry)

Evidence for natural selection is pervasive in the mouse genome



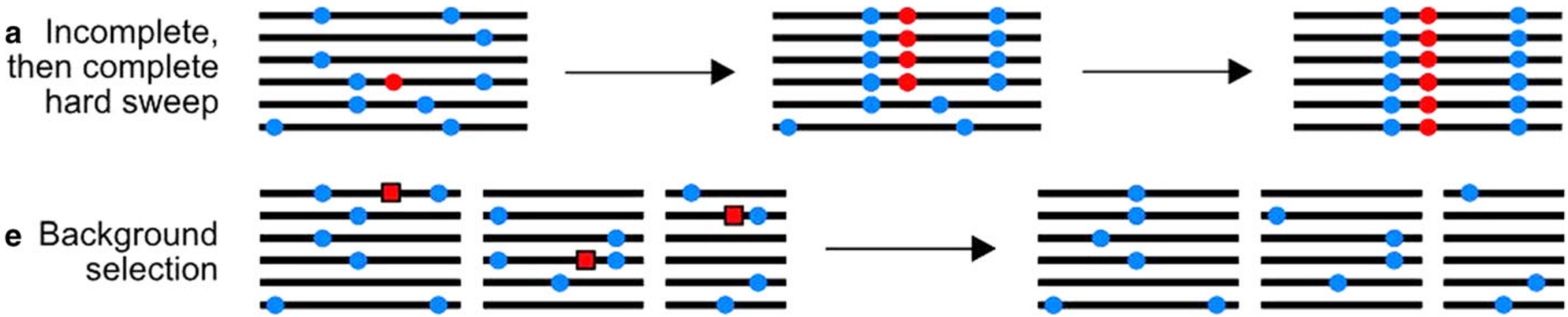
OPEN ACCESS Freely available online

PLOS | GENETICS

Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents

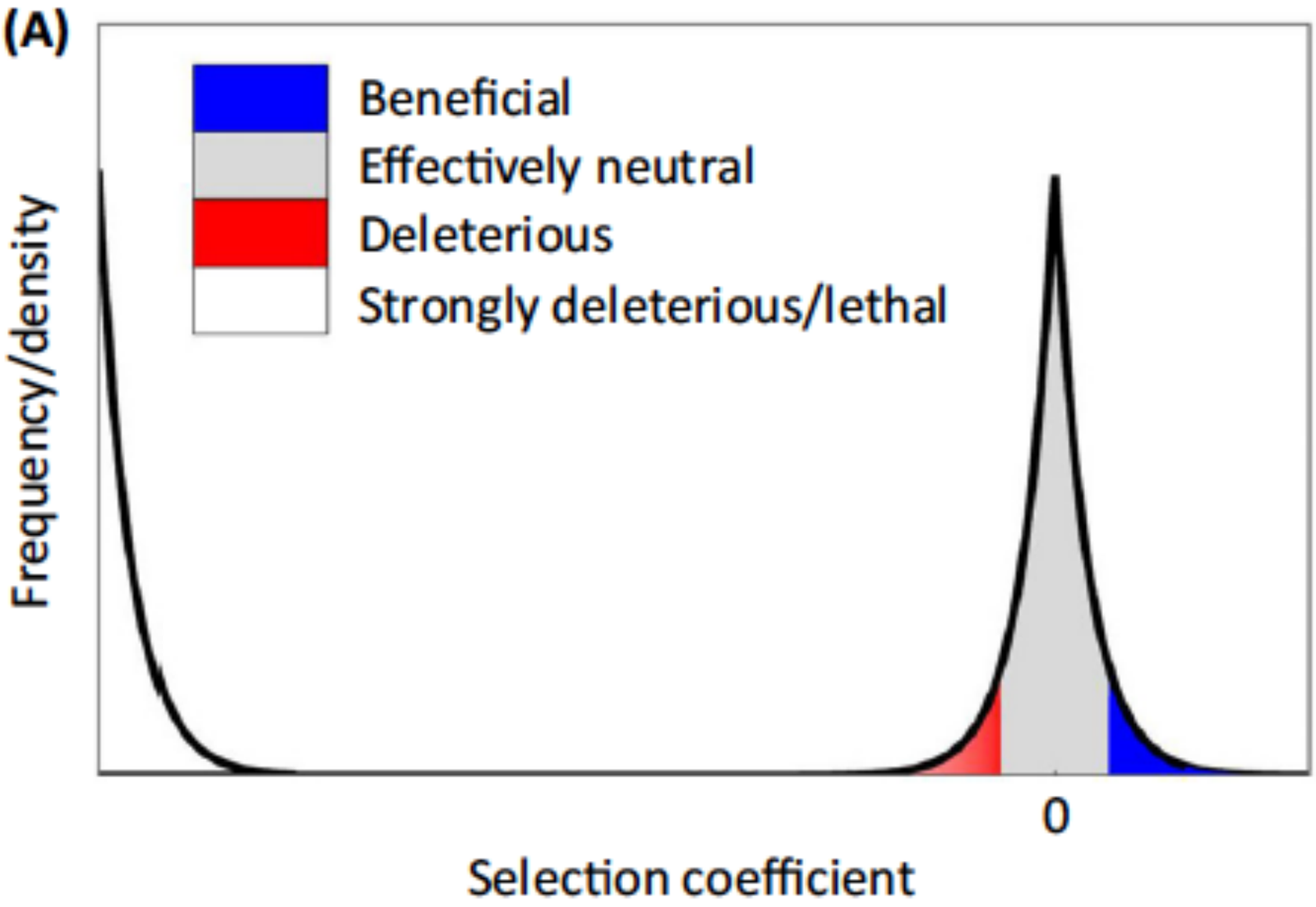
Daniel L. Halligan[1], Athanasios Kousathanas[1¤], Rob W. Ness[1], Bettina Harr[2], Lél Eöry[3], Thomas M. Keane[4], David J. Adams[4], Peter D. Keightley[1*]

1 Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom, 2 Max-Planck Institute for Evolutionary Biology, Plön, Germany, 3 The



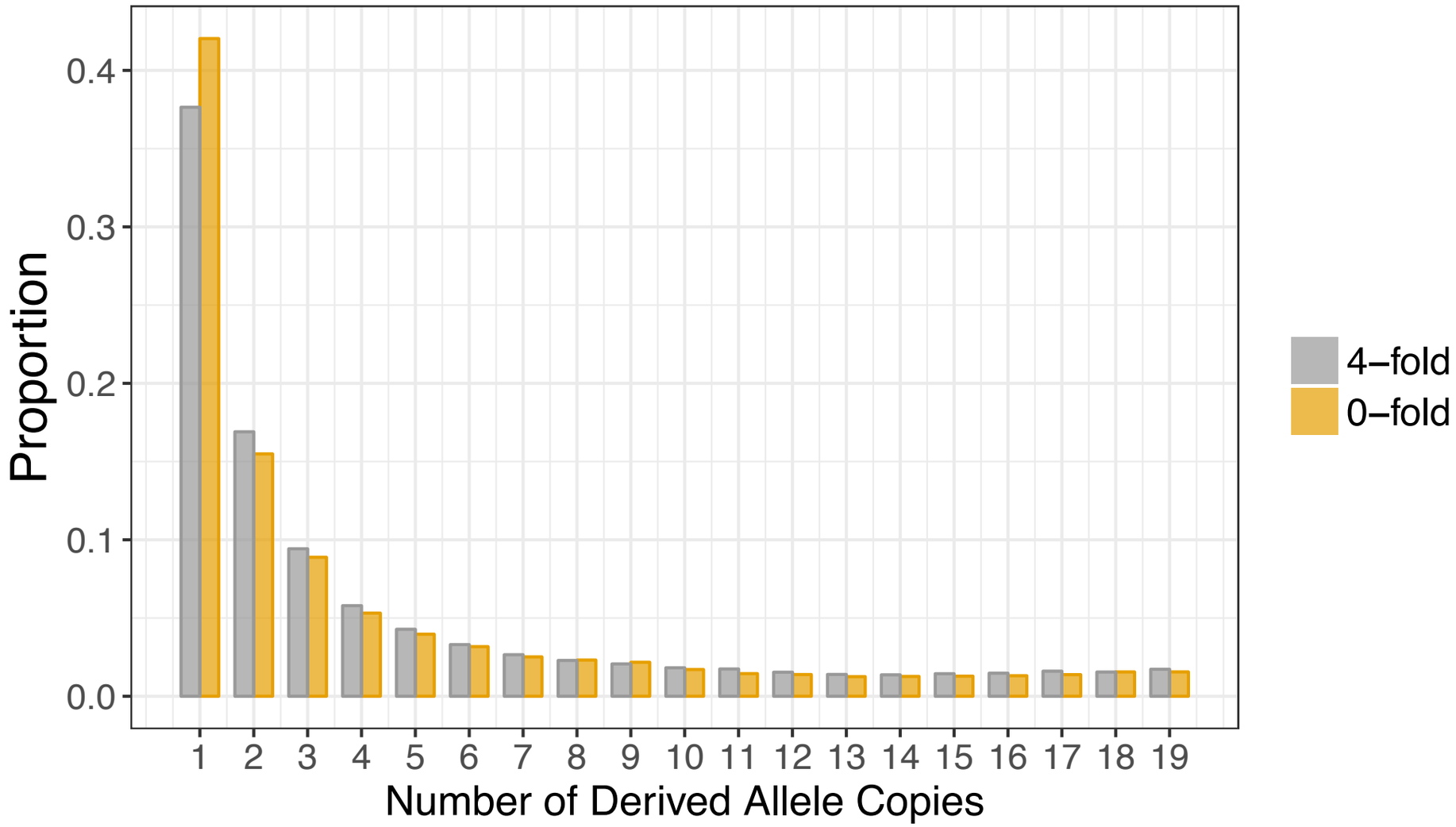
**This is Ben Jackson's figure*

Distribution of fitness effects (DFE)



Bank et al. 2014

Estimating selection parameters from the uSFS



Differences between the uSFS for selected and neutral sites carry information on the rate and strength of selected mutations

Assumes that selected mutations are segregating in the population

Can strongly selected, rare advantageous mutations be detected by analysis of the uSFS?

If advantage mutations are strongly selected, their sojourn times can be quite short

The rate of substitutions at selected sites:

$$V_a = 2\mu\ p_a\gamma_a$$

The per base-pair, per-generation mutation rate

Proportion of new mutations that are advantageous

Scaled strength of selection, $2N_es$

Table 1: Positive selection parameter estimates obtained by analysis of the uSFS for simulated poplations.

Divergence [a]	γ_a		p_a		$\gamma_a p_a$	Prop.
	Simulated	*Estimated*	*Simulated*	*Estimated*		Significant [b]
+	10	11.2 [5.60 - 20.0]	0.010000	0.00856 [0.00440 - 0.0199]	0.0954 [0.0838 - 0.115]	1.00
-		3.97 [1.13 - 27.2]		0.0201 [0.00472 - 0.0706]	0.0828 [0.0616 - 0.155]	1.00
+	20	16.6 [9.20 - 37.4]	0.005000	0.00568 [0.00241 - 0.0107]	0.0949 [0.0822 - 0.108]	1.00
-		19.9 [2.90 - 37.4]		0.00532 [0.00289 - 0.0207]	0.106 [0.0454 - 0.193]	0.97
+	50	37.4 [21.6 - 41.8]	0.002000	0.00257 [0.00202 - 0.00467]	0.0951 [0.0809 - 0.106]	1.00
-		37.3[1.87 - 65.5]		0.00266 [0.00125 - 0.0146]	0.0717 [0.0112 - 0.145]	0.86
+	100	37.43 [37.4 - 1530]	0.001000	0.00249 [0.0000738 - 0.00283]	0.0938 [0.0795 - 0.107]	1.00
-		0.323 [0.0371 - 1.25]		0.00259 [0.000525 - 0.0941]	0.00102 [0.0000620 - 0.0137]	0.00
+	200	37.4 [37.4 - 1,700]	0.000500	0.00251 [0.000220 - 0.00283]	0.0947 [0.0738 - 0.106]	1.00
-		0.272 [0.00546 - 1.911]		0.0122 [0.000690 - 0.138]	0.00310 [0.000104 - 0.0294]	0.07
+	400	37.4 [32.7 - 37.4]	0.000250	0.00245 [0.00199 - 0.00283]	0.0919 [0.0776 - 0.102]	1.00
-		12.3 [0.287 - 66.6]		0.00212 [0.000783 - 0.0104]	0.0338 [0.000250 - 0.0984]	0.22
+	800	37.4 [32.9 - 37.4]	0.000125	0.00222 [0.00186 - 0.00264]	0.0831 [0.0701 - 0.0936]	1.00
-		1.75 [0.111 - 43.0]		0.00240 [0.000343 - 0.0293]	0.0134 [0.0000515 - 0.0649]	0.12

[a] +/- indicates whether or not divergence was included when analysing the uSFS

[b] The proportion of bootstrap replicates where a full DFE gave a significantly better fit than a model containing just deleterious mutations

When advantageous mutations are strongly selected, between-species divergence is effectively the only information on the positive selection parameters

Table S1: Parameters of the distribution of fitness effects for harmful mutations obtained by analysis of the uSFS

Divergence[a]	Full DFE[b]	β[c]	$\hat{\gamma}_d$ [c]
+	+	0.203 [0.190 - 0.231]	-865 [-1120 - -561]
+	-	0.135 [0.127 - 0.140]	-6860 [-10100 - -4850]
-	+	0.217 [0.190 - 0.270]	-755 [-110000 - -483]
-	-	0.175 [0.166 - 0.184]	-1550 [-2100 - -1180]
+	+	0.199 [0.184 - 0.212]	-974 [-1390 - -744]
+	-	0.132 [0.125 - 0.142]	-8480 [-13200 - -5030]
-	+	0.199 [0.187 - 0.226]	-9831 [-1330 - -676]
-	-	0.176 [0.168 - 0.183]	-1620 [-2040 - -1230]
+	+	0.199 [0.179 - 0.210]	-979 [-1680 - -740]
+	-	0.136 [0.130 - 0.144]	-7260 [-11100 - -4930]
-	+	0.199 [0.187 - 0.215]	-944 [-1350 - -739]
-	-	0.186 [0.177 - 0.195]	-1220 [-1640 - -986]
+	+	0.195 [0.175 - 0.210]	-952 [-1780 - -661]
+	-	0.137 [0.129 - 0.144]	-5980 [-9350 - -4140]
-	+	0.193 [0.184 - 0.271]	-953 [-1270 - -637]
-	-	0.189 [0.182 - 0.199]	-1040 [-1310 - -790]
+	+	0.197 [0.174 - 0.210]	-1040 [-2060 - -748]
+	-	0.136 [0.130 - 0.144]	-7470 [-10700 - -5100]
-	+	0.207 [0.187 - 0.353]	-927 [-1320 - -498]
-	-	0.190 [0.183 - 0.199]	-1160 [-1470 - -917]
+	+	0.209 [0.192 - 0.224]	-745 [-1180 - -558]
+	-	0.148 [0.141 - 0.156]	-4010 [-5910 - -2810]
-	+	0.210 [0.199 - 0.229]	-727 [-939 - -541]
-	-	0.202 [0.193 - 0.212]	-840 [-1040 - -660]
+	+	0.210 [0.181 - 0.218]	-798 [-1500 - -592]
+	-	0.148 [0.139 - 0.157]	-3890 [-6000 - -2720]
-	+	0.205 [0.193 - 0.236]	-804 [-1020 - -543]
-	-	0.198 [0.189 - 0.209]	-889 [-1130 - -693]

[a] +/- indicates whether or not divergence was included when analysing the uSFS

[b] +/- indicates whether or not advantageous mutation parameters were inferred

[c] The shape parameter of the gamma distribution of deleterious fitness effects

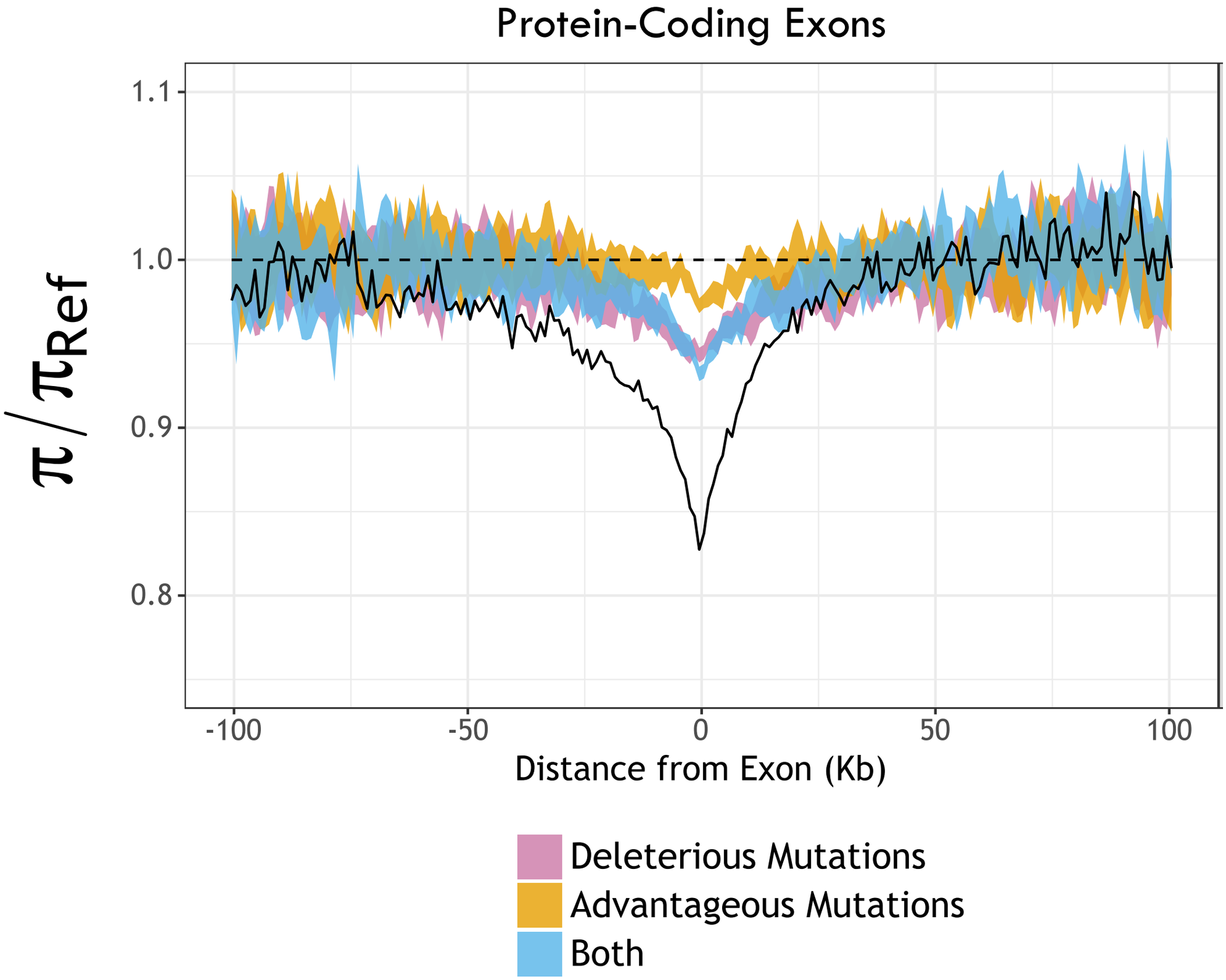
[d] Mean strength of selection of a new harmful mutation

True values

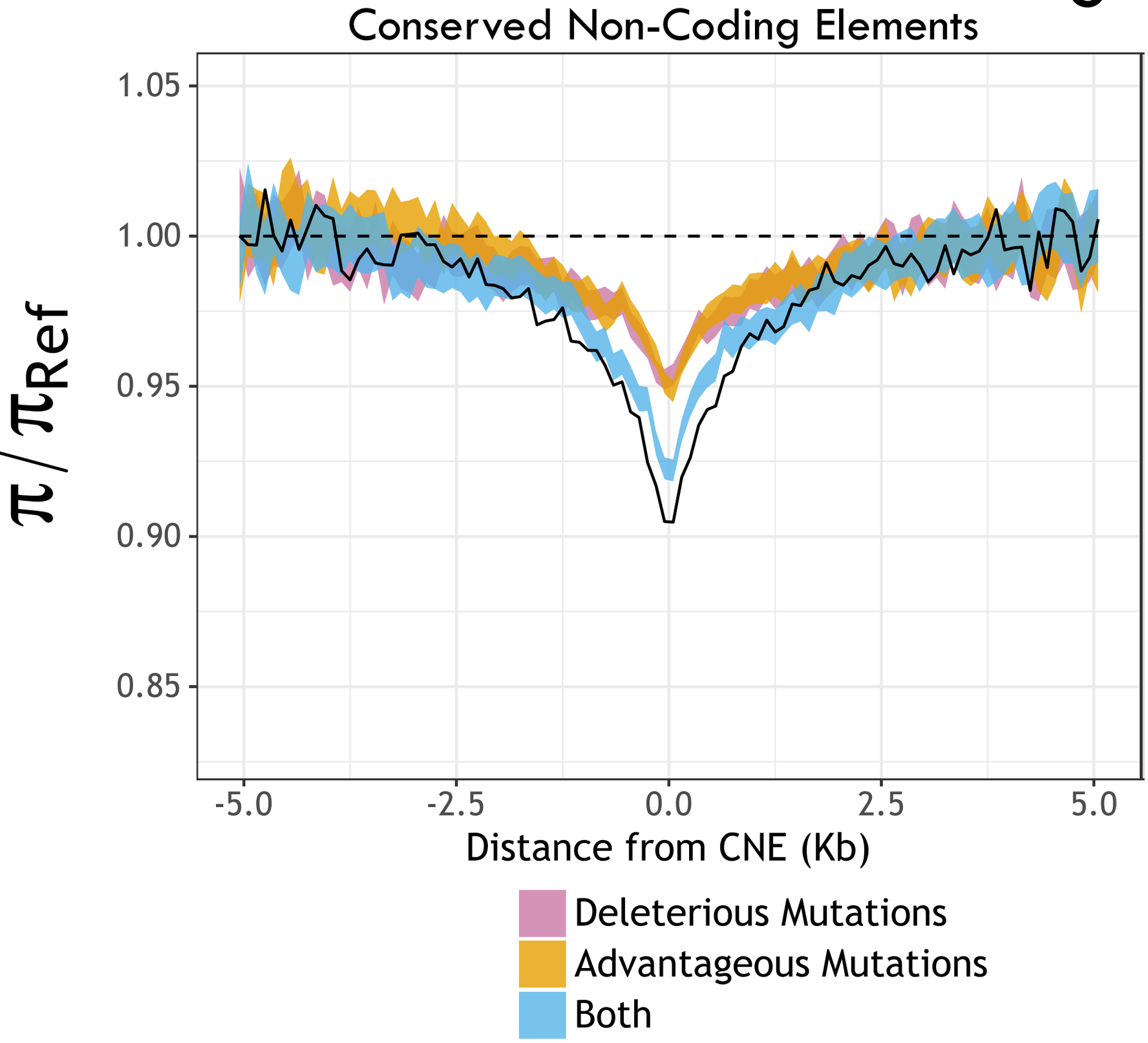
$\beta = 0.2$

$\gamma_d = -1000$

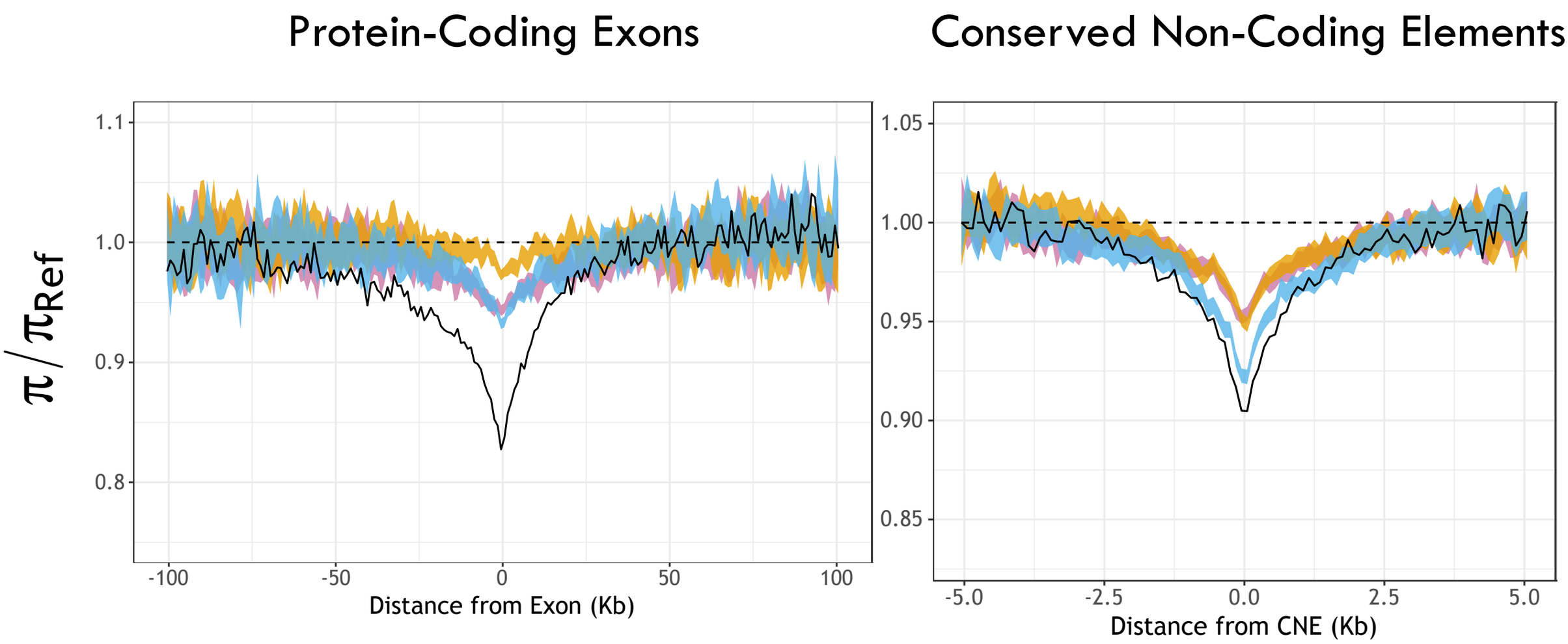
Natural selection in the mouse genome



Natural selection in the mouse genome



Natural selection in the mouse genome



Selection parameters obtained from the uSFS explain patterns of diversity around CNEs, but not exons

Estimating the parameters of positive selection from diversity troughs

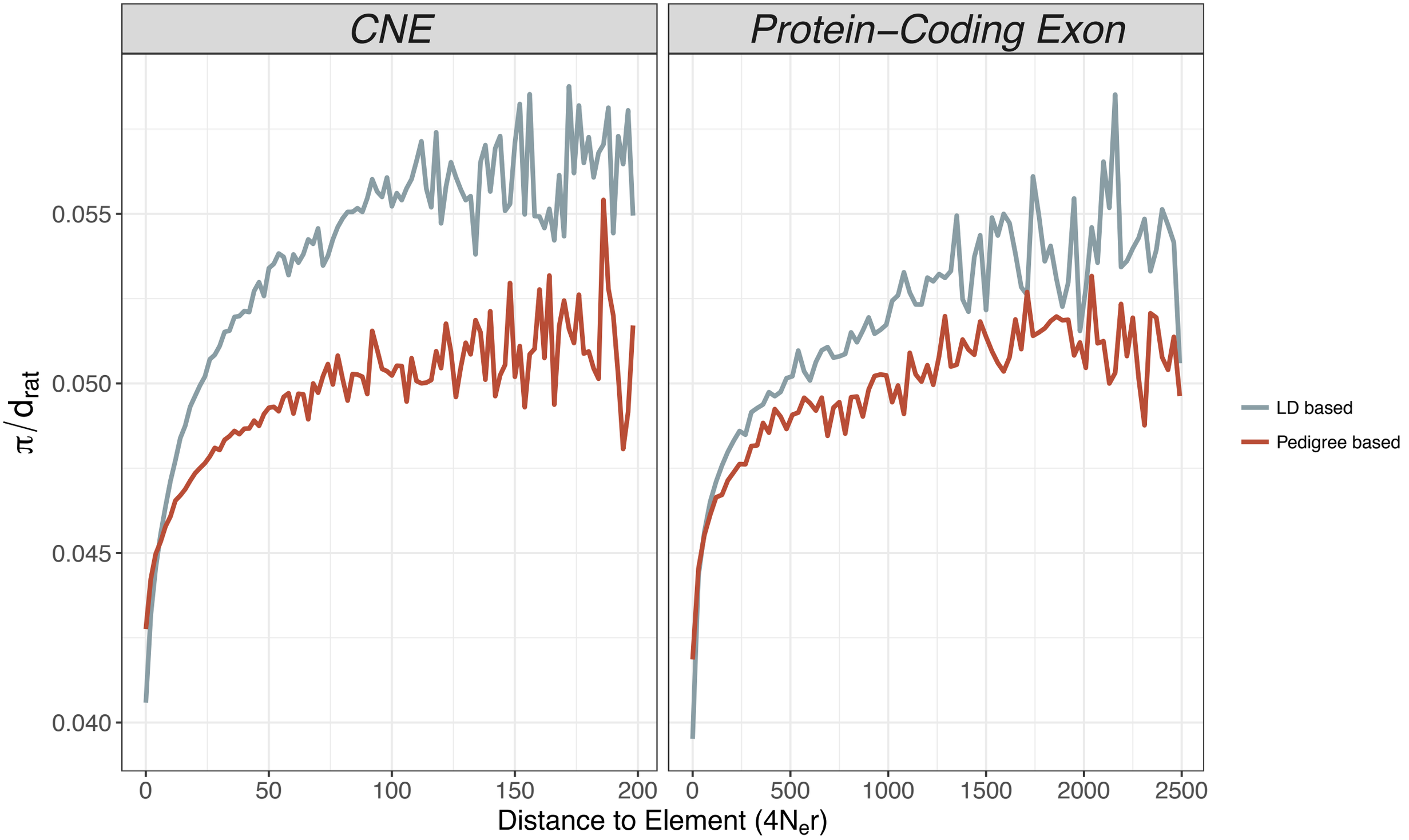
An approximation for the combined effects of background selection and selective sweeps

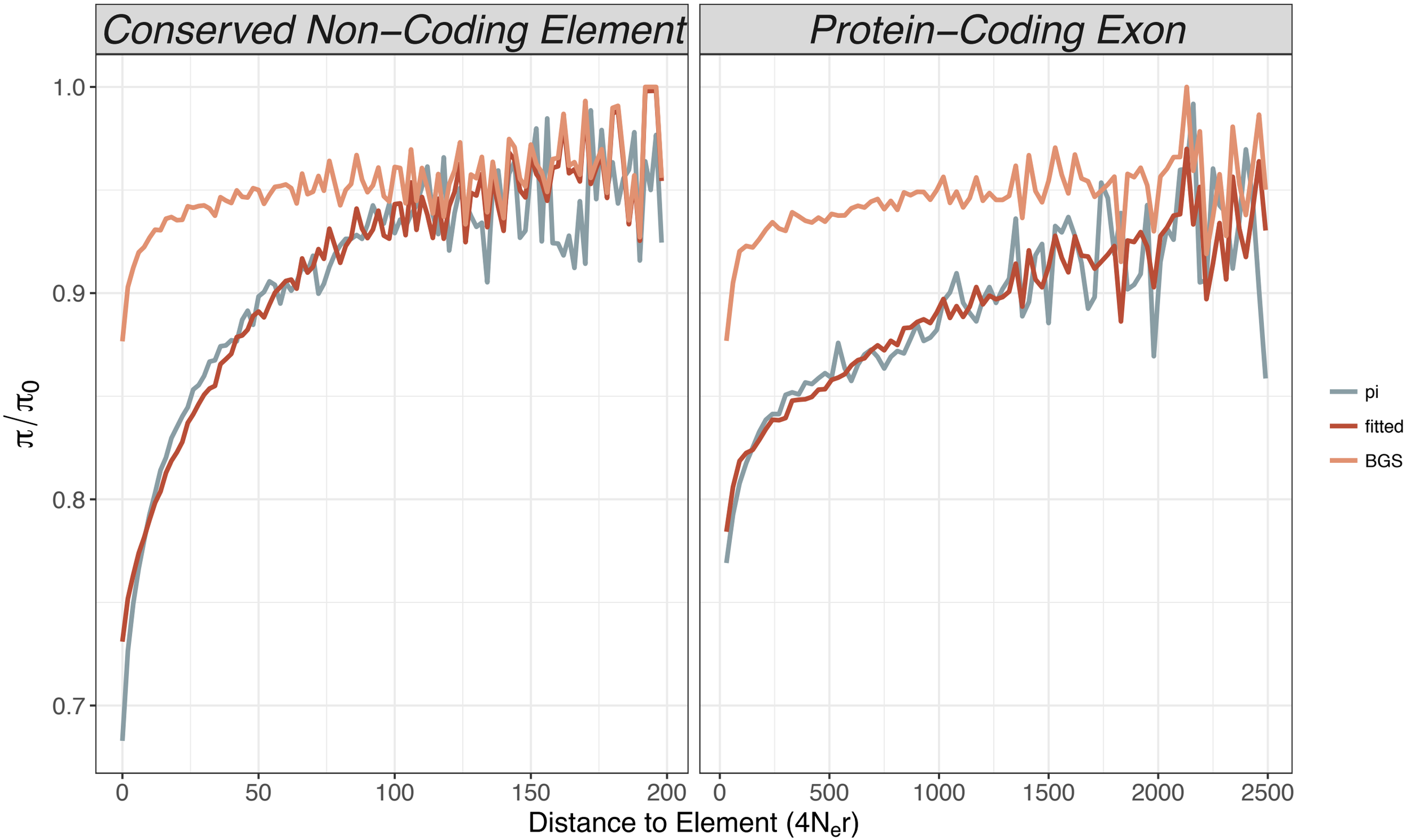
$$\frac{\pi_j}{\pi_0} \approx \frac{1}{B_j^{-1} + B2N_eP_{sc,j}}$$

$$P_{sc,j} \approx V_a \tau \gamma_a^{\frac{-4r_{i,j}}{s}}$$

Minor tweaks to the model used by Campos *et al.* 2017 PNAS

Choice of genetic map makes a difference to data analysis





Assuming the LD-based map

Element	Recombination Map	$\gamma_a = 2N_e s_a$	p_a	π_0
Protein-Coding Exon	LD-based	13,350 [2,035]	9.720×10^{-6} [2.391×10^{-6}]	0.00950
Protein-Coding Exon	Pedigree-based	5,971 [1,945]	1.139×10^{-5} [6.003×10^{-6}]	0.00895
Conserved Non-Coding Elements	LD-based	490.2 [41.60]	0.00111 [1.496×10^{-4}]	0.0103
Conserved Non-Coding Elements	Pedigree-based	336.3 [88.77]	6.55×10^{-4}] [2.65×10^{-4}]	0.00916

If we assume that selection on adaptive mutations is strong relative to drift…

$$\Delta W \propto n_a \int \mu_a u(s_a) s_a f(s_a) ds_a,$$

Leads to

$$\Delta W \propto n_a \mu_a E(s_a^2),$$

If the point mutation rate is the same for all sites in the genome, then ΔW is proportional to $n_a p_a s_a^2$

Adaptation in Protein-Coding Versus Regulatory Regions

LD-Based Recombination Map

	Nonsynonymous sites	Conserved Non-Coding (Regulatory)
n_a *(Mbp)*	25.8	82.1
p_a	0.00000972	0.00111
$2N_e s_a$	13,350	490.2
s_a^2	0.00025	0.00000034
$\Delta W \propto n_a p_a s_a^2$	0.0627	0.0310

Assuming an N_e for M. m. castaneus *of 420,000*

Adaptation in Protein-Coding Versus Regulatory Regions

Pedigree-Based Recombination Map

	Nonsynonymous sites	Conserved Non-Coding (Regulatory)
n_a *(Mbp)*	25.8	82.1
p_a	0.0000114	0.000655
$2N_e s_a$	5971	336.3
s_a^2	0.000051	0.00000016
$\Delta W \propto n_a p_a s_a^2$	0.0149	0.0086

Assuming an N_e *for* M. m. castaneus *of 420,000*

- Adaptation in proteins *seems* to contribute more to adaptive fitness change
- But we have chopped the genome up into pretty crude categorisations. A more sophisticated analysis might use ENCODE categorisations for CNEs and look at different classes of protein-coding genes