In What Direction is Evolution Going?

Stanley Sawyer, Washington University in St. Louis, USA*

Are most new fixed mutations deleterious or advantageous?

Clues: McDonald-Kreitman tables: For DNA sequences from two closely-related species:

	mono. at diff. bases	poly. in either spp.
Replacement	K_a	S_a
Silent	K_s	S_s

(Replacement means that it changes an amino acid.) Excess/Deficit of replacement fixed differences suggest possible positive/negative selection.

Drosophila species tend to show a significant excess of replacement fixed differences in these tables over many loci, suggesting overall positive selection on fixed differences. Other pairs of related species show a significant deficit, suggesting overall negative selection.

(*)-Joint work with Rob Kulathinal, Carlos Bustamante, and Dan Hartl

How to model? Can we estimate the amount of selection involved?

Many events tend to happen on a scale of N_e generations, where N_e is the effective population size.

It is useful to consider five different kinds of mutations, where s is the rate of selection per generation:

(i)	$s < 0$, $\lvert sN \rvert \gg 1$	Evolutionary lethal
(ii)	$s < 0$, $\lvert sN \rvert = O(1)$	Weakly deleterious
(iii)	$s = 0$	Neutral
(iv)	$s > 0$, $\lvert sN \rvert = O(1)$	Weakly advantageous
(v)	$s > 0$, $\lvert sN \rvert \gg 1$	Hopeful monsters(?)

Evolutionary lethal mutations can be ignored since they rapidly disappear in this time scale, and hopeful monsters are essentially never polymorphic.

This will be a theory of (ii,iii,iv). This ignores the most interesting mutations (v), but they may be rare.

Looking ahead, we will find that the expected proportions of beneficial mutations among nonlethal replacement mutations in 56 Drosophila loci are

New (nonlethal) mutations	19%
Polymorphic in samples	47%
Fixed differences	93%

The model: We assume

- All new mutations occur at a new site.

- Sites are unlinked; that is, are statistically independent. (Seems OK by forwards simulation for applications with two related species. Also, many loci show evidence of strong short-segment gene conversion, which could randomize sites.)

- Directional selection for each new mutant site, with no epistasis or dominance over sites.

- Silent sites are neutral. For replacement mutations (that change an amino acid), each new $\gamma = (N_e)s$ is drawn from a normal distribution with parameters

  $$N(\gamma_i, \sigma_w^2)$$

  where γ_i depends on the i^{th} locus. (σ_w^2 should also, but we would need more data. Bustamante etal 2002 is the same model with $\gamma \equiv \gamma_i$.)

- The γ_i for loci are drawn from another normal distribution

  $$N(\mu_\gamma, \sigma_b^2)$$

  so that the distribution of γs for new mutations is the same as a random-effects model in statistics.

PRF model: The probability of survival of a new mutant is approximately $p(\gamma) = (1/N)(2\gamma/(1-\exp(-2\gamma))$, so that most new mutants are lost. However, a proportion $p(\gamma)$ of these will eventually be fixed.

The sites in the general population that are polymorphic will vary from time to time, but there will always be a random set of sites that are polymorphic. These will have a random set of population site frequencies p for these random sites, all moving independently (since sites are unlinked).

In the limit as $N \to \infty$, with $O(1)$ new mutations per generation, the result is a Poisson random field of population site frequencies.

If γ is fixed, the polymorphic population frequencies form a Poisson random field on $0 < p < 1$ as $N \to \infty$ with densities

$$\theta_r \frac{1 - e^{-2\gamma(1-p)}}{1 - e^{-2\gamma}} \frac{dp}{p(1-p)} \qquad \text{(Replacement)}$$

$$\theta_s \frac{dp}{p} \qquad \text{(Silent)}$$

(Sawyer and Hartl 1992) Here θ_r and θ_s are the replacement and silent-site mutation rates per generation, and mutant replacement bases have a relative selective advantage of γ/N_e.

Fixations occur at the relative rates

$$\theta_r \frac{2\gamma}{1 - \text{exp}(-2\gamma)} \qquad \text{and} \qquad \theta_s$$

These are *population* fixation rates and polymorphism frequencies. For *samples* of m and n sequences from two closely-related species, the counts K_a, S_a, K_s, S_s are independent Poisson with means

$$E(K_a) = \theta_r \left(\frac{2\gamma}{1 - e^{-2\gamma}}\right) \big(t + G(m) + G(n)\big)$$

$$E(S_a) = \theta_r \left(\frac{2\gamma}{1 - e^{-2\gamma}}\right) \big(F(m) + F(n)\big)$$

$$E(K_s) = \theta_s \left(t + \frac{1}{m} + \frac{1}{n}\right)$$

$$E(S_s) = \theta_s \big(L(m) + L(n)\big)$$

Here t is the scaled divergence time of the two species and

$$G(n) = \int_0^1 (1-p)^{n-1} \frac{1 - e^{-2\gamma p}}{2\gamma p}\, dp$$

$$F(n) = \int_0^1 \frac{1 - p^n - (1-p)^n}{1-p} \frac{1 - e^{-2\gamma p}}{2\gamma p}\, dp$$

$$L(n) = \sum_{i=1}^{n-1} \frac{1}{i}$$

If the γs are chosen independently from $N(\gamma_i, \sigma_w^2)$ within each locus, the formulas for $E(K_a)$ and $E(S_a)$ are replaced by double integrals, with a Gaussian integral on the outside. The sampling formulas are valid if e.g. $m = 1$, as long as the other sample size $n > 1$. In that case, all of the polymorphism information comes from the species with $n > 1$.

The model allows $\theta_{ri} \neq \theta_{si}$, so that $\theta_{ri}/(2\theta_{si}) = q_i$ gives an estimate of the average number of possible nonlethal amino-acid replacements at the i^{th} locus.

Data: We started with $T = 72$ loci with $n_i > 1$ sequences (at the i^{th} locus) from *D. simulans* and one from *D. melanogaster*. Unfortunately, to get our model to converge, we were forced to set $q_i = q$ across loci. We threw out locus outliers (estimated $q_i > 0.28$) and some other suspicious loci, reducing $T = 72$ loci to $T = 56$. The 16 dropped genes were mostly apparent pseudogenes and rapidly-evolving Acp loci.

This left us with two "local" parameters for each locus

$$\theta_{si},\ \gamma_i \qquad 1 \leq i \leq 56$$

and five "global" parameters (shared by all loci)

$$\mu_\gamma,\ \sigma_b^2,\ \sigma_w^2,\ q,\ t$$

with $2T + 5 = 117$ parameters for $4T = 224$ observations.

We used MCMC (Markov Chain Monte Carlo), which essentially estimates parameters by asking where most of the mass of the likelihood

$$L(\theta_{si}, \gamma_i, \mu_\gamma, \sigma_b, \sigma_w, q, t,\ K_{ai}, S_{ai}, K_{si}, S_{si}) \qquad (*)$$

is, viewed as a function of the parameters, with the observed data $K_{ai}, S_{ai}, K_{si}, S_{si}$ held constant.

MCMC works by defining a Markov chain with (*) as a stationary distribution and computing averages and quantiles over long runs of this Markov chain. With $q_i = \theta_{ri}/(2\theta_{si}) = q$ fixed at a global parameter, the Markov chain converged very nicely.

Technically speaking, we used $n =$ 10,000 "burnin" Markov chain steps to stabilize the parameters and then $n =$ 1,000,000 further steps, sampling only every 10$^{\mathrm{th}}$ step to lessen autocorrelation. The last 10 "subchains", each with 100,000 steps (10,000 samples), gave very similar results. The following discusses the parameter estimates that resulted from this run.

Results for global parameters were
(len=100,000, 10 subchains)

Var	Mean±1.96×SD	GR	R^2
$\mu_\gamma > 0$	0.01 ± 0.21	1.0001	0.0024
μ_γ	-7.3 ± 10.3	1.0027	0.0248
σ_b	5.69 ± 2.43	1.0013	0.0123
σ_w	6.79 ± 4.92	1.0025	0.0228
$\sigma_b/(\sigma_b + \sigma_w)$	0.47 ± 0.12	1.0015	0.0142
$\theta_r/2\theta_s = q$	0.16 ± 0.11	1.0025	0.0231
t	2.48 ± 0.31	1.0000	0.0007

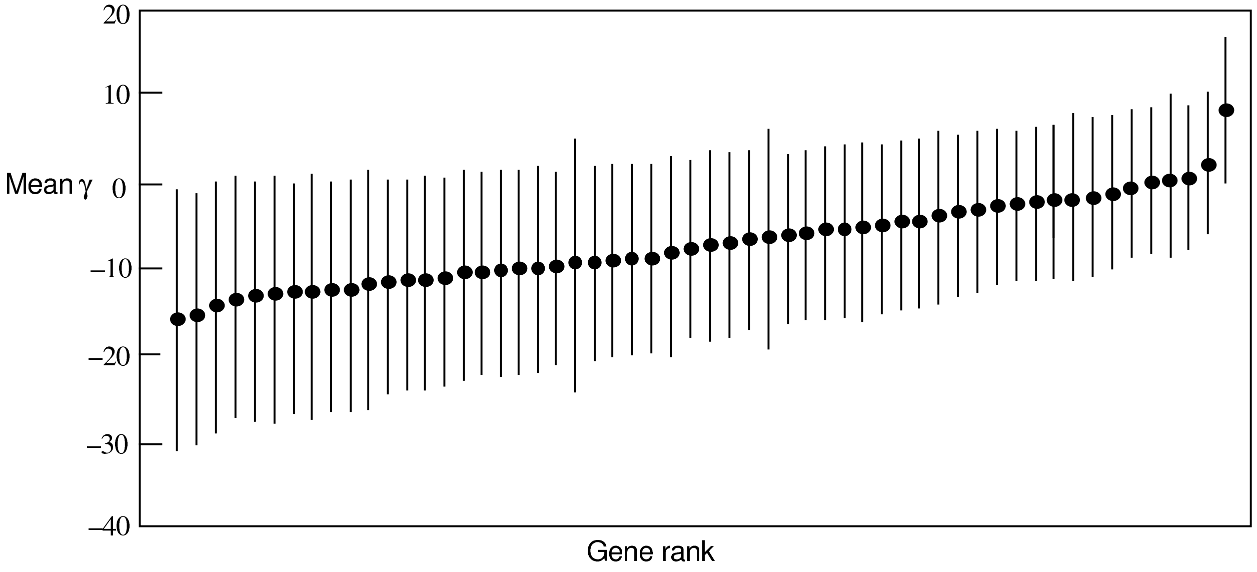
(last subchain, len=10,000)

Var	Median	95% credible interval
μ_γ	-5.74	(-20.7, -0.34)
σ_b	5.42	(3.70, 8.46)
σ_w	6.20	(2.87, 12.7)
$\sigma_b/(\sigma_b + \sigma_w)$	0.47	(0.37, 0.61)
$\theta_r/2\theta_s = q$	0.14	(0.08, 0.32)
t	2.47	(2.18, 2.80)

GR and R^2 are diagnostics for MCMC convergence. GR < 1.02 is considered very good.

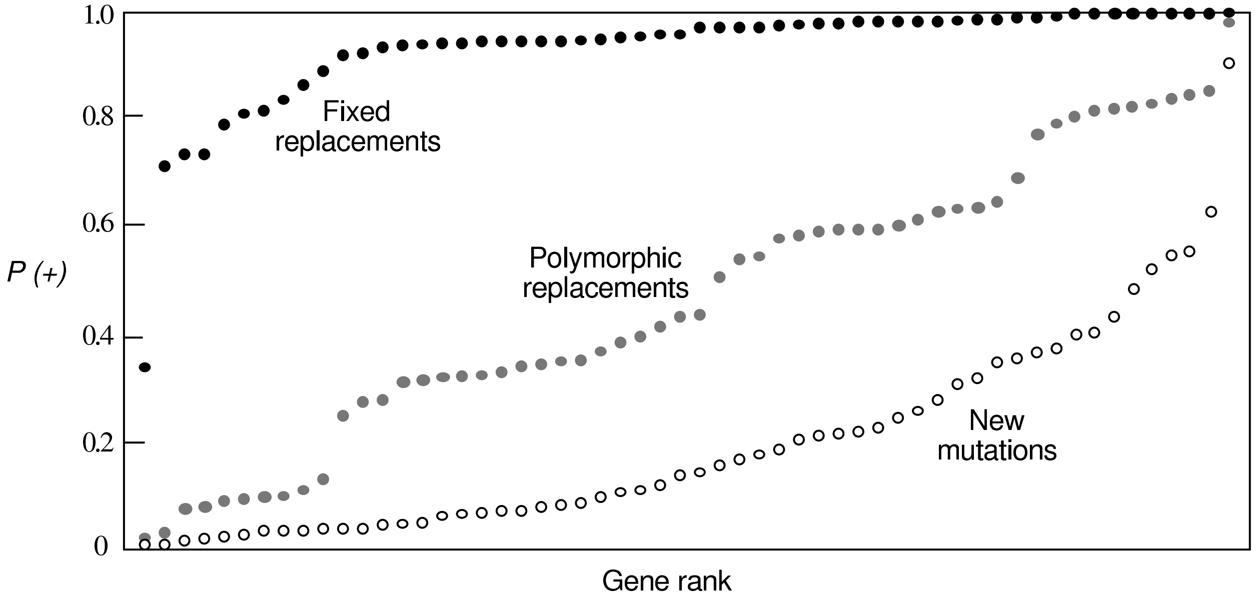
In particular, $\sigma_b/(\sigma_b + \sigma_w)$ had median 0.47 and varied in the range $(0.37, 0.61)$ (middle 95% quantiles), so that about half of the γ-variability was within locus and about half was between locus.

Fig 1: Newly arising nonlethal replacement mutations:



Mean ± 1.96 × StdDev for $\gamma_i = (N_e)s_i$, ranked by mean. The fraction with $\gamma > 0$ varies from 1% (*Pgm*) to 62% (*Rel*) and 90% (*mei-218*). (Average = 19%.)

Fig 2: Proportions of mutations that are beneficial (based on averages of functions of (γ_i, σ_w) over the MCMC run):

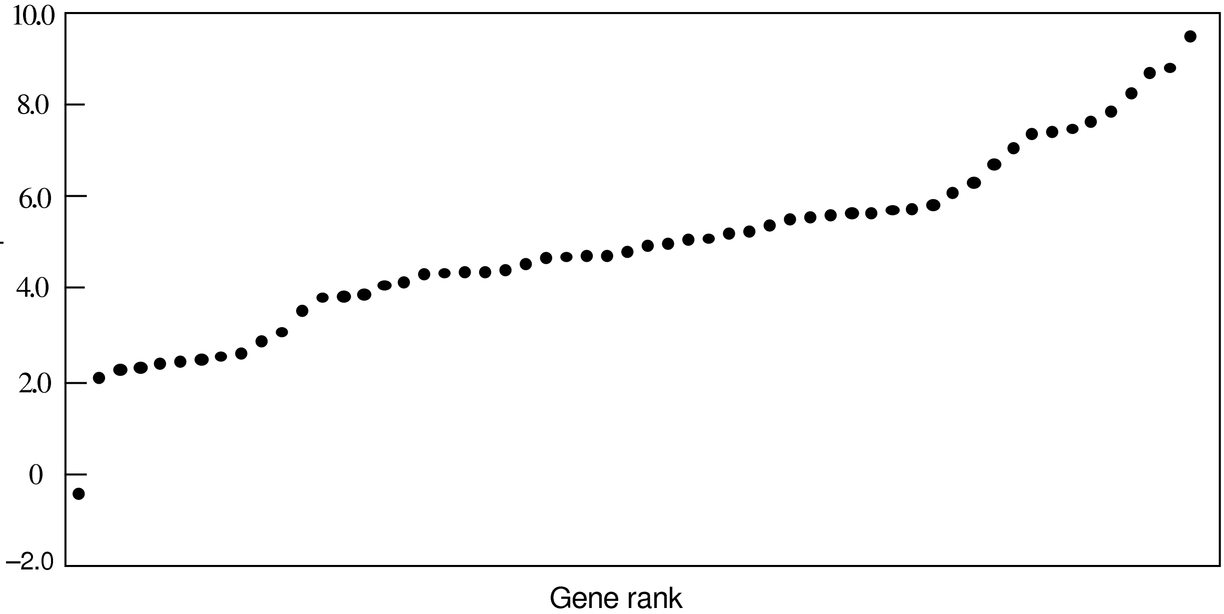


In Figure 2, the lower outlier for fixed replacements is *kl-5*, which is a Drosophila fertility factor gene on the Y chromosome.

Overall, averaging over all 56 loci, the expected proportions of beneficial mutations among replacement mutations are

New (nonlethal) mutations	19%
Polymorphic in samples	47%
Fixed differences	93%

Fig 3: Expected mean scaled selection coefficients (γ_f) among fixed replacement mutations:



With one exception, the average $\gamma_f = (N_e)s_f$ for fixed replacement mutations in the two populations varies in the range 2.0–10.0.

Thank you for coming.

(Joint work with Rob Kulathinal, Carlos Bustamante, and Dan Hartl.)