In What Direction is Evolution Going?

Stanley Sawyer, Department of Mathematics

Are most new fixed mutations deleterious
or advantageous?

"Naïve pan-selectionism" would say that
most are advantageous.

Competing theories suggest that most are
selectively neutral or mildly deleterious.

Collaborators:

Dan Hartl, Harvard University

Carlos Bustamante, Cornell University

Rob Kulathinal, Harvard University

John Parsch, University of Munich

Haley Abel, Washington University

Amei Amei, University of Nevada Los Vegas

and many others

McDonald-Kreitman tables: For DNA sequences from two closely-related species at one locus:

	mono. at diff. bases	poly. in either spp.
Replacement	K_r	S_r
Silent	K_s	S_s

(Replacement means that it changes an amino acid.)

Excess/Deficit of replacement fixed differences suggest possible positive/negative selection.

The polymorphisms at most loci are too sparse to make any conclusion, but we can use the Mantel-Haenszel test to aggregate over loci:

	mono. at diff. bases	poly. in either spp.	
Replacement	K_r	S_r	T_r
Silent	K_s	S_s	T_s
(Totals)	K_t	S_t	N

Over many independent loci ($1 \leq i \leq L$):

$$Z = \sum_{i=1}^{L} \Big(K_{ri} - EK_{ri} \Big) \Big/ \sqrt{\sum_i \mathsf{Var}(K_{ri})}$$

has an $N(0,1)$ distribution if there are no deficits or excesses, where $EK_r = T_r K_t / N$ at the i^{th} locus. (Z is necessary to avoid Simpson's Paradox.)

Drosophila: $L = 91$ $Z = 3.30$ ($P < 0.001$)
D. melanogaster: $7 \leq m_i \leq 12$ sequences
D. simulans: $n_i = 1$ sequence ($i =$ locus)
(Baines et al 2007, Sawyer et al 2007)

Arabidopsis: $L = 12$ $Z = -4.47$ ($P < 10^{-5}$)
A. thaliana: $14 \leq m_i \leq 21$ sequences
A. lyrata: $n_i = 1$ sequence
(Bustamante et al 2002)

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

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 (i,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 a 91-locus Drosophila dataset are

New (nonlethal) mutations	7%
Polymorphic in samples	30%
Fixed differences	94%

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, $\gamma = (N_e)s$ for each new mutation is drawn from a normal distribution with parameters

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

  where γ_i depends on the i^{th} locus. (Bustamante etal 2002 has the same model with $\gamma \equiv \gamma_i$.)

- The γ_i for loci are themselves drawn from a normal distribution

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

  This means 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 immediately lost. However, a proportion 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 mean density

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

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

(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

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

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

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

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

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

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

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 $E(K_r)$ and $E(S_r)$ are replaced by double integrals, with a Gaussian integral on the outside. The sampling formulas are valid if e.g. $n = 1$, as long as the other sample size $m > 1$. In that case, all of the polymorphism information comes from the species with $m > 1$.

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.

A Drosophila dataset (Pröschel etal 2006) with $T = 91$ autosomal loci has $m_i > 1$ sequences (at the i^{th} locus) from *D. melanogaster* in Zimbabwe and one sequence from *D. simulans* (from North Carolina). The model has three "local" parameters for each locus

$$\theta_{ri},\ \theta_{si},\ \gamma_i \qquad 1 \leq i \leq 91$$

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

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

This makes $3T + 4 = 277$ parameters for $4T = 364$ observations.

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

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

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

Technically speaking, we used $n = 1{,}000{,}000$ "burnin" Markov chain steps to stabilize the parameters and then $n = 20{,}000{,}000$ further steps, sampling only every 10^{th} step to save memory. The last 10 "subchains", each with 2,000,000 steps (200,000 samples), gave very similar results.

Results for global parameters were
(2,000,000 samples, 10 subchains)

Var	Mean$\pm$1.96$\times$SD	GR
$\mu_\gamma > 0$	0.16 ± 0.72	1.009
μ_γ	-11.3 ± 30.9	1.020
σ_w	6.91 ± 11.5	1.021
σ_b	4.28 ± 4.38	1.018
$\sigma_w/(\sigma_b + \sigma_w)$	0.55 ± 0.28	1.015
t	4.48 ± 0.44	1.000

GR is a diagnostic for MCMC convergence. GR < 1.03 is considered good.

In the last subchain, $\sigma_w/(\sigma_b+\sigma_w)$ had median 0.59 and varied in the range $(0.24, 0.80)$ (middle 95% quantiles), so that about half of the γ-variability was within locus and about half was between locus.

Fig 1: For a 56-locus Drosophila dataset (Sawyer etal 2003), the proportions of mutations that are beneficial (based on averages of functions of (γ_i, σ_w) over the MCMC run):

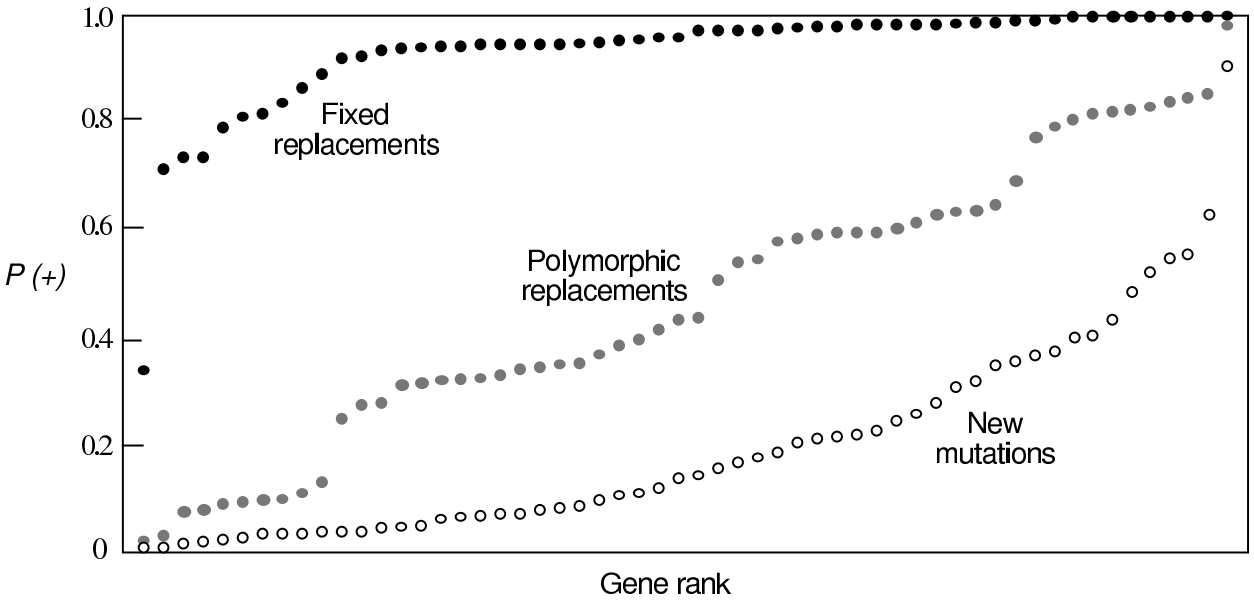
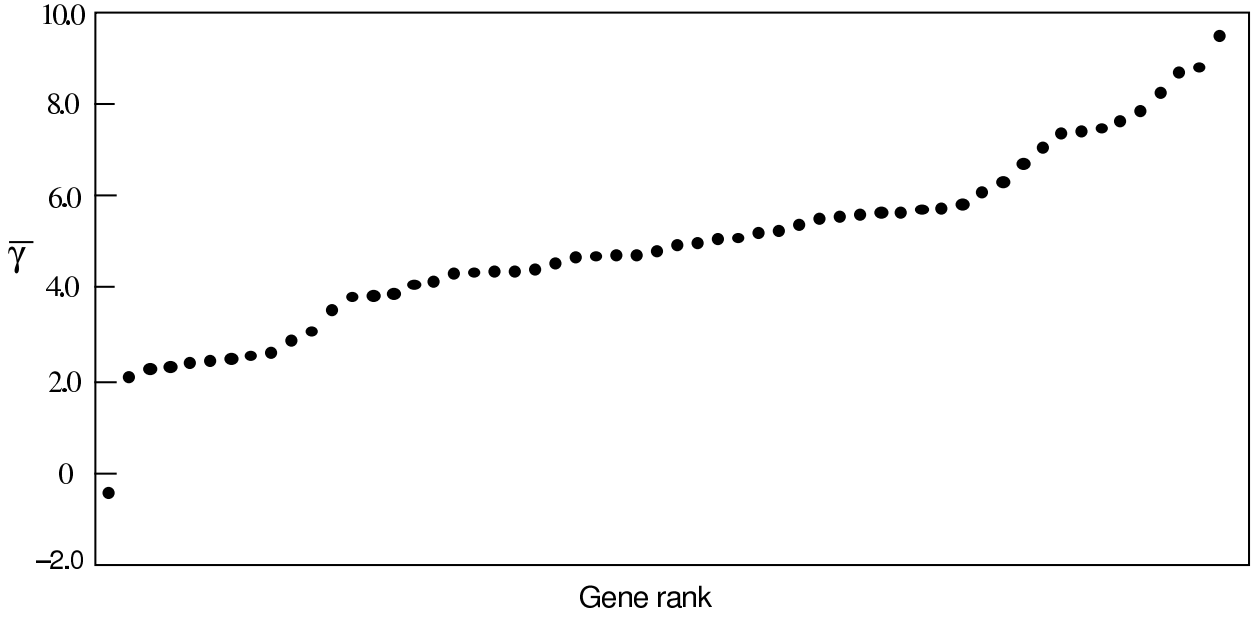


Fig 2: Among fixed replacement mutations, the average scaled selection coefficients:



1
0.8
0.6
0.4
0.2
P
$N_e s \leq 4$
$N_e s \leq 3$
$N_e s \leq 2$
$N_e s \leq 1$
$N_e s \leq 0$
20
40
60
80
Rank

Robustness of model to various assumptions:
(Checked by forwards simulations and other methods)

If linkage is tight, most parameter estimates are accurate but estimated confidence intervals are too small. (This means that conclusions that estimated parameters are nonzero can be wrong.)

If scaled recombination $R = N_e\rho$ is of moderate size ($R \geq 50$) and the breakpoints are uniformly distributed, conclusions are robust.

If the within-locus distribution of scaled selection coefficients of new replacement mutations is replaced by a double exponential (and so presumably also a gamma distribution) or by an even more heavy-tailed Student-t distribution, then conclusions are qualitatively the same.

Thank you for coming.