amplification of *Dhh* and *Ihh*, the *ScShh* fragment was obtained by nested PCR. The amino-acid sequences used for degenerate primers were: *ScShh* first PCR, KQFIPNVA and AHIHCSV *ScShh* second, nested PCR, PNYNPDI and GFDWVYYE. The longer *ScShh* (348 bp) fragment was obtained from cDNA pools of stage 27 *S. canicula* embryos by RT–PCR using a degenerate forward primer and a *ScShh*-specific reverse primer (GRYEGKIT and TTCGTAGTAGACCCAGTC). The nucleotide sequences of *ScEn1*, *ScShh*, *ScTbx4*, *ScTbx5*, *ScdHand* and *ScBmp4* cDNA are deposited in the GenBank database under the accession numbers: AF393834–AF393837, AY057890 and AY057891.

### *In situ* hybridization

*S. canicula* embryos were removed from their egg casings and dissected from the yolk mass. Wholemount *in situ* hybridization on younger *S. canicula* embryos was carried out as previously described[23] and this is based on methods used for other vertebrate embryos[24]. Older embryos were treated with dimethyl sulphoxide (DMSO) instead of proteinase K treatment by placing them in 2 ml of DMSO/methanol (1:1) on ice until they sank. Then 0.5 ml of 10% Triton X-100 (Sigma) in distilled water was added, and the embryos were incubated for an additional 20 minutes at room temperature[25,26]. After washing in PBT (1% Tween 20 (Sigma) in PBS), embryos were hybridized with probes as described previously for chick embryos[24]. Some whole-mount *in situ* samples were embedded in gelatin, and frozen sections were cut.

### RT–PCR

RT–PCR was performed as previously described[27]. The primers used for PCR amplification of *ScShh* (172 bp) were 5′-GAGCTGACAGGCTGATGACAC-3′ and 5′-TGGTGATGTCCACAGCTCGGC-3′. The PCR cycle was at 96 °C for 20 seconds, 55 °C for 40 seconds and 72 °C for 1 minute for 32 cycles. Relative levels of transcripts were compared to levels of internal control using 18S ribosomal RNA primers (Ambion). Both *ScShh* and 18S rRNA primers were added into the same reaction solution.

### Observation of cartilaginous pattern

*S. canicula* embryos to be stained for cartilage were fixed in 5% TCA (trichloroacetic acid), stained in 0.1% Alcian blue in 70% acid alcohol, dehydrated in ethanol and cleared in methyl salicylate.

### DiI labelling

DiI (1,1-dioctadecyl-3,3,3′,3′-tetramethylindo-carbocyanine perchloride; Molecular Probes; 3 mg $ml^{-1}$ in DMSO) was injected into chick wing-bud using a micropipette to label a small group of cells. Embryos were then incubated for 96 h and fixed in 4% paraformaldehyde in PBS. The average size of the initial DiI injected dot was 40–50 μm.

Received 2 July; accepted 17 December 2001.


1. Thacher, J. K. Median and paired fins, a contribution to the history of vertebrate limbs. *Trans. Conn. Acad.* **3,** 281–310 (1877).
2. Jarvik, E. in *Basic Structure and Evolution of Vertebrates* 109–131 (Academic, London, 1980).
3. Moy-Thomas, J. A. The evolution of the pectoral fins of fishes and the tetrapod forelimb. *School Sci. Rev.* **36,** 592–599 (1936).
4. Coates, M. I. The origin of vertebrate limbs. *Development* (1994 Suppl.), 169–180 (1994).
5. Shu, D.-G. *et al.* Lower Cambrian vertebrates from South China. *Nature* **402,** 42–46 (1999).
6. Neyt, C. *et al.* Evolutionary origins of vertebrate appendicular muscle. *Nature* **408,** 82–86 (2000).
7. Balfour, F. M. The development of elasmobranch fishes. *J. Anat. Physiol. Lond.* **11,** 128–172 (1876).
8. Ballard, W. W., Mellinger, J. & Lechenault, H. A series of normal stages for development of *Scyliorhinus canicula*, the lesser spotted dogfish (*Chondrichthyes*: *Scyliohinidae*). *J. Exp. Zool.* **267,** 318–336 (1993).
9. Altabef, M., Clarke, J. D. & Tickle, C. Dorso-ventral ectodermal compartments and origin of apical ectodermal ridge in developing chick limb. *Development* **124,** 4547–4556 (1997).
10. Tanaka, M. *et al.* Apical ectodermal ridge induction by the transplantation of En-1-overexpressing ectoderm in chick limb bud. *Dev. Growth Differ.* **40,** 423–429 (1998).
11. Tabin, C. & Laufer, E. *Hox* genes and serial homology. *Nature* **361,** 692–693 (1993).
12. Coates, M. I. *Hox* genes, fin folds and symmetry. *Nature* **364,** 195–196 (1993).
13. Tamura, K. *et al.* Evolutionary aspects of positioning and identification of vertebrate limbs. *J. Anat.* **199,** 195–204 (2001).
14. Ruvinsky, I., Silver, L. M. & Gibson-Brown, J. J. Phylogenetic analysis of T-box genes demonstrates the importance of amphioxus for understanding evolution of the vertebrate genome. *Genetics* **156,** 1249–1257 (2000).
15. Balinsky, B. I. Das extremitaetenseitenfeld, seine ausdehnung und beschaffenheit. *Roux Arch. Dev. Biol* **130,** 704–736 (1933).
16. Cohn, M. J., Izpisua-Belmonte, J. C., Abud, H., Heath, J. K. & Tickle, C. Fibroblast growth factors induce additional limb development from the flank of chick embryo. *Cell* **80,** 739–746 (1995).
17. Tanaka, M. *et al.* Distribution of polarizing activity and potential for limb formation in mouse and chick embryos and possible relationship to polydactyly. *Development* **127,** 4011–4021 (2000).
18. Riddle, R. D., Johnson, R. L., Laufer, E. & Tabin, C. *Sonic hedgehog* mediates the polarizing activity of the ZPA. *Cell* **75,** 1401–1416 (1993).
19. Ekker, S. C. *et al.* Patterning activities of vertebrate hedgehog proteins in the developing eye and brain. *Curr. Biol.* **5,** 44–55 (1995).
20. Cohn, M. J. Giving limbs a hand. *Nature* **406,** 953–954 (2000).
21. Tümpel, S. *et al.* Antero-posterior signaling in vertebrate limb development and stripes of *Tbx3* expression.*Dev. Biol* (submitted).
22. Kraus, P., Fraidenraich, D. & Loomis, C. A. Some distal limb structures develop in mice lacking Sonic hedgehog signaling. *Mech. Dev.* **100,** 45–58 (2001).
23. Mazan, S., Jaillard, D., Baratte, B. & Janvier, P. *Otx1* gene-controlled morphogenesis of the horizontal semicircular canal and the origin of the gnathostome characteristics. *Evol. Dev.* **2,** 186–193 (2000).
24. Wilkinson, D. G. *In Situ Hybridization: A Practical Approach* 75–83 (IRL Press/Oxford Univ. Press, Oxford, 1992).
25. Kuratani, S., Ueki, T., Aizawa, S. & Hirano, S. Peripheral development of cranial nerves in a cyclostome, *Lampetra japonica*: morphological distribution of nerve branches and the vertebrate body plan. *J. Comp. Neurol.* **384,** 482–500 (1997).
26. Schlosser, G. & Roth, G. Evolution of nerve development in frogs. I. The development of the peripheral nervous system in *Discoglossus pictus* (Discoglossidae). *Brain Behav. Evol.* **50,** 61–93 (1997).
27. Münsterberg, A. E., Kitajewski, J., Bumcrot, D. A., McMahon, A. P. & Lassar, A. B. Combinatorial signaling by Sonic hedgehog and Wnt family members induces myogenic bHLH gene expression in the somite. *Genes Dev.* **9,** 2911–2922 (1995).
28. Coates, M. I. Limb evolution. Fish fins or tetrapod limbs—a simple twist of fate?*Curr. Biol.* **5,** 844–848 (1995).


**Supplementary Information** accompanies the paper on *Nature's* website (http://www.nature.com).


### Acknowledgements

We are grateful to A. Wells for his assistance in maintenance of *S. canicula* embryos, S. Kuratani for information about *S. canicula* developmental studies before publication, S. Mazan for technical advice and *ScOtx1*and *ScOtx2* cDNA as positive control probes for establishing *in situ* hybridization methods and N. Helps for DNA sequencing. M.T. is supported by JSPS Postdoctoral Fellowships for Research Abroad, JSPS Research Fellowships for Young Scientists and the Inoue Research Award for Young Scientists. A.M. is supported by a Wellcome Trust research Career Development Award. C.T. is Foulerton Research Professor of The Royal Society.


Correspondence and requests for materials should be addressed to M.T. (e-mail: m.tanaka@dundee.ac.uk).

# The cost of inbreeding in *Arabidopsis*


**Carlos D. Bustamante*†, Rasmus Nielsen‡, Stanley A. Sawyer§, Kenneth M. Olsen‖, Michael D. Purugganan‖ & Daniel L. Hartl***

* *Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA*
‡ *Department of Biometrics, Cornell University, Ithaca, New York 14853-2801, USA*
§ *Department of Mathematics, Washington University, St Louis, Missouri 63130, USA*
‖ *Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695-7614, USA*



**Population geneticists have long sought to estimate the distribution of selection intensities among genes of diverse function across the genome. Only recently have DNA sequencing and analytical techniques converged to make this possible. Important advances have come from comparing genetic variation within species (polymorphism) with fixed differences between species (divergence)[1,2]. These approaches have been used to examine individual genes for evidence of selection. Here we use the fact that the time since species divergence allows combination of data across genes. In a comparison of amino-acid replacements among species of the mustard weed *Arabidopsis* with those among species of the fruitfly *Drosophila*, we find evidence for predominantly beneficial gene substitutions in *Drosophila* but predominantly detrimental substitutions in *Arabidopsis*. We attribute this difference to the *Arabidopsis* mating system of partial self-fertilization, which corroborates a prediction of population genetics theory[3–6] that species with a high frequency of inbreeding are less efficient in eliminating deleterious mutations owing to their reduced effective population size.**


We analysed *Arabidopsis* data for 12 genes of diverse function for

† Present address: Department of Statistics, University of Oxford, Oxford OX1 3TG, UK.

which alleles were sequenced from two species. In each case, multiple alleles were sequenced from the partly self-fertilizing species *A. thaliana* and one allele from the closely related outcrossing species *A. lyrata*. For comparison, we analysed 34 genes from *D. melanogaster* and its sibling species *D. simulans*. Presentation of the data for each individual gene conventionally takes the form of a two-by-two table (Table 1), in which the number of nucleotide differences in each gene that do not change the encoded amino acid (synonymous) are categorized either as fixed within species but different between species ($K_s$) or as polymorphisms within one or both species ($S_s$). Similarly, the number of nucleotide differences in each gene that do change the encoded amino-acid (replacement) are categorized as either fixed ($K_a$) or polymorphic ($S_a$).

We use the term DPRS to refer to the layouts in Table 1 because, in clockwise order, the headings are divergence (D), polymorphism (P), replacement (R), and synonymous (S). Table 1 shows the means and s.d. (and range) of the counts in each cell for the genes from *Arabidopsis* and *Drosophila*. By itself, none of the individual DPRS tables is very illuminating as to whether amino-acid replacements ($K_a$) are selectively neutral relative to synonymous substitutions ($K_s$). In the absence of selection, the ratio of the expectations of $K_s$ to $S_s$ should equal the ratio of the expectations of $K_a$ to $S_a$, so that selection is detected as a significant $P$ value in a conventional test for homogeneity[1,2]. Among the individual DPRS tables, when the $P$ values are corrected for multiple comparisons, there is only one gene in the *Arabidopsis* data that is significant at the 5% level, and only two genes in the *Drosophila* data that are significant at the 5% level (Fisher exact tests, data not shown).

Although few of the individual DPRS tables are significant, each contains information about the selective forces impinging on the amino-acid replacements. A theoretical framework for extracting this information derives from considerations of the equilibrium flux of fixations and limiting probability densities of nucleotide substitutions affected by mutation, selection and random genetic drift taking place simultaneously and independently at multiple sites in a DNA sequence[2]. In this framework, known as a Poisson random field (PRF), the magnitude of each cell observed in a DPRS table is an independent Poisson random variable whose expected value[2] is given by the corresponding equation in Table 1 (bottom). The symbols $n$ and $m$ are the number of alleles sequenced from each of the species being compared. The quantity $\frac{1}{2}\theta_s$ is the expected number of synonymous mutations that occur in the gene in the entire population in any generation; $\frac{1}{2}\theta_a$ is the corresponding quantity for non-synonymous mutations (amino-acid replacements), excluding those mutations that are so deleterious that they have a negligible chance of becoming polymorphic or fixed. The parameter of greatest present interest is $\gamma$, which is the selection intensity in favour of (if $\gamma > 0$) or against (if $\gamma < 0$) amino acid replacements; $\gamma$ is scaled according to the haploid effective population number $N_e$; hence, $\gamma$ equals the selection coefficient multiplied by $N_e$. (The haploid effective size is twice the diploid effective size.) Although synonymous sites in some genes are known to be under selection for optimal codon usage, this effect is usually small[7,8] and is not explicitly taken into account in the present formulation. Consequently, $\gamma$ can be thought of as the magnitude of selection affecting amino-acid replacements relative to that affecting synonymous substitutions. Because our main interest is in variation in the intensity of selection among genes, we assume that all amino-acid polymorphisms and fixed differences in the same gene have a common $\gamma$, but that $\gamma$ can differ from one gene to the next. The fourth parameter, $t$, is the number of generations since the species diverged from a common ancestor, again expressed as a multiple of $N_e$. The three functions in Table 1 (bottom) are defined as

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

$$F(n) = \int_0^1 \frac{1 - x^n - (1-x)^n}{1-x} \frac{1 - e^{-2\gamma x}}{2\gamma x} dx$$

$$G(n) = \int_0^1 (1-x)^{n-1} \frac{1 - e^{-2\gamma x}}{2\gamma x} dx$$

Although each DPRS table has four parameters ($\theta_s$, $\theta_a$, $\gamma$ and $t$) and four observations ($K_a$, $K_s$, $S_a$ and $S_s$), the divergence time $t$ is a common parameter. This implies that each gene contributes valuable information about the distribution of $\gamma$ among genes. We therefore chose an analytical method that borrows information from all the genes to make inferences about the magnitude of selection for any individual gene. This approach greatly increases the power and accuracy of the inferences regarding selection.

A suitable framework is provided by the hierarchical Bayesian model described in Box 1 (refs 9, 10). For each species pair, we assume that the magnitude of $\gamma$ for each gene is drawn randomly and independently from a normal distribution with mean $\mu$ and standard deviation $\sigma$. The hierarchical structure is achieved by assuming that $\mu$ and $\sigma$ are themselves random variables. On the basis of this model we estimate the probability distribution of $\mu$ given the observed data and show that this distribution for genes in *Arabidopsis* is significantly different from that for genes in *Drosophila*.

The equation for the posterior distribution $\pi(\gamma, t, \theta, \mu, \sigma \mid D)$ given in the box is analytically intractable. Nevertheless, parameter estimates based on the posterior distribution are possible. The approach is first to define a Markov chain that has $\pi(\gamma, t, \theta, \mu, \sigma \mid D)$ as its stationary distribution. The means, modes, variances and other quantities for individual parameters are approximated by sampling from one or more long trajectories of this Markov chain with different starting positions. The Markov chain is defined by a sampling method, called Markov Chain Monte Carlo (MCMC)[11], whose properties result in convergence to $\pi(\gamma, t, \theta, \mu, \sigma \mid D)$. By

**Table 1 Observed and expected numbers of amino-acid changes that are fixed between species or polymorphic within species**

| Amino-acid change | Divergence | Polymorphism |
|---|---|---|
| *Arabidopsis* | | |
| Synonymous | $K_s = 26.7 \pm 14.0$ (range 1–55) | $S_s = 9.5 \pm 7.1$ (range 1–25) |
| Replacement | $K_a = 11.2 \pm 6.6$ (range 3–26) | $S_a = 9.3 \pm 6.5$ (range 2–19) |
| *Drosophila* | | |
| Synonymous | $K_s = 18.5 \pm 12.5$ (range 1–49) | $S_s = 17.4 \pm 17.2$ (range 0–69) |
| Replacement | $K_a = 10.9 \pm 16.3$ (range 0–75) | $S_a = 5.7 \pm 8.9$ (range 0–37) |
| Poisson random-field expected values | | |
| Synonymous | $E(K_s) = \theta_s\left(t + \frac{1}{m} + \frac{1}{n}\right)$ | $E(S_s) = \theta_s[L(m) + L(n)]$ |
| Replacement | $E(K_a) = \theta_a\left(\frac{2\gamma}{1-e^{-2\gamma}}\right)[t + G(m) + G(n)]$ | $E(S_a) = \theta_a\left(\frac{2\gamma}{1-e^{-2\gamma}}\right)[F(m) + F(n)]$ |

*Arabidopsis* data are means ± s.d. among 12 genes from *A. thaliana* and *A. lyrata*; the number of alleles ranges from 14 to 21 (mean 17.6), and the length of coding region sequenced from 531 to 1,674 base pairs (bp) (mean 935 bp). *Drosophila* data are means ± s.d. among 34 genes from *D. melanogaster* and *D. simulans*; the number of alleles ranges from 5 to 62 (mean 10.8) and the length of coding region sequenced from 270 to 7,683 bp (mean 1172 bp).

custom, a relatively large number of iterations at the beginning of each realization of the Markov chain is disregarded as a 'burn-in' period. This is a protection against possible bias caused by the starting conditions.

For each trajectory in the MCMC simulations, we chose arbitrary initial values for $\mu$ and $\sigma$, the selection parameters, and the divergence time. Then, for each gene in turn, given its value of $\gamma$, a value for $\theta_a$ was drawn from a gamma distribution

$$\Gamma\left\{\alpha + K_a + S_a,\ \beta + \left(\frac{2\gamma}{1 - e^{-2\gamma}}\right)[t + G(m) + G(n) + F(m) + F(n)]\right\}$$

which is the posterior distribution of $\theta_a$ given the data and all other relevant parameters. Similarly, a value for $\theta_s$ was chosen from its posterior distribution

$$\Gamma[\alpha + K_s + S_s, \beta + t + 1/m + 1/n + L(m) + L(n)]$$

In both cases we choose $\alpha$ and $\beta$ to be close to 0 and hence uninformative.

Iterative updating of the selection parameters proceeds by Metropolis sampling[12] as follows. For each gene in turn, choose a proposed new value of $\gamma$–call it $\gamma'$–from a narrow uniform distribution centred on $\gamma$. Evaluate the Poisson means in Table 1 (bottom) and calculate the likelihood of the data for the value $\gamma'$ (using the current value of $\theta_a$) and multiply by the prior probability of $\gamma'$ calculated from the normal distribution with mean $\mu$ and variance $\sigma^2$. This is essentially the posterior probability for $\gamma'$, and if it is greater than the posterior probability for $\gamma$, set $\gamma = \gamma'$; otherwise set $\gamma = \gamma'$ only if a uniform random number in $[0, 1]$ is less than the ratio of the posterior probabilities. The mutation parameters for each gene in turn are updated by Gibbs sampling[13], which is implemented by choosing new values from the updated gamma distributions. Once all the selection and mutation parameters have been updated, the divergence time is updated by Metropolis sampling in a manner analogous to the updating of the $\gamma$ values. To update $\mu$ and $\sigma$, first calculate the sample mean ($\bar{\gamma}$) and variance ($s_\gamma^2$) of the updated $\gamma$ values. Then sample $\sigma^2$ from its posterior distribution, which is inverse-gamma-distributed with parameters that depend on $s_\gamma^2$ and the number of genes $k$. Finally, update $\mu$, which is normally distributed with mean and variance that depend on $\bar{\gamma}$, $\sigma^2$ and $k$. This procedure of updating the $\gamma$, $\theta$, $t$, $\mu$ and $\sigma^2$ is then repeated to form the Markov chain of parameter values whose stationary distribution is given by $\pi(\gamma, t, \theta, \mu, \sigma|D)$. The range of the uniform distributions used to update the selection parameters and the divergence time were chosen by preliminary trials to yield an acceptance rate of about 50%.

For each species pair, we ran five independent chains from overdispersed starting points for $10^6$ iterations each. The first $10^5$ iterations were treated as burn-in and disregarded. The chains converged and mixed very fast, yielding estimated scale reduction factors[14] close to 1 (1.00001 for *Arabidopsis* and 0.99994 for

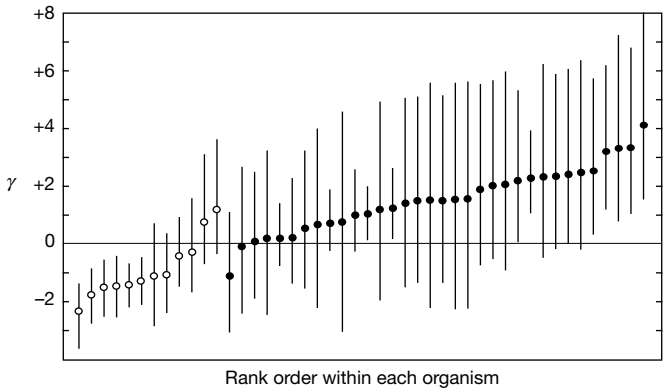


**Figure 1** Means of the posterior distributions of the selection parameters $\gamma$ (dots) and the 95% credible intervals (vertical lines), in rank order according to the mean, for the genes in *Arabidopsis* (open circles) and *Drosophila* (filled circles). From left to right the *Arabidopsis* genes are *AP3*, *PgiC*, *PI*, *AP1*, *ChiA*, *CAL*, *TFL1*, *FAH1*, *Adh1*, *F3H*, *LFY* and *CHI*, and the *Drosophila* genes are *per*, *Pgi*, *Acp26Ab*, *Adh*, *Est-6*, *pn*, *boss*, *CecA2*, *Anon1G5*, *tty*, *ref(2)p*, *Anon1E9*, *Tpi*, *Anon1A3*, *CecA1*, *sIgA*, *Pp4-19C*, *AnnX*, *Cdic*, *Bap*, *shakB*, *gld*, *Dhc-Yh3*, *runt*, *Acp26Aa*, *z*, *6-Pgd*, *Yp2*, *Rh3*, *ase*, *Acp29AB*, *cta*, *G6pd* and *ci*. Sources of the data are referenced in the Supplementary Information.

Box 1
**Hierarchical bayesian analysis of polymorphism and divergence**

In bayesian statistics, assumptions about the forms of the underlying distributions (prior distributions) of parameters are combined with current data by the use of a likelihood function. The result is the posterior distribution of the parameters, given the data. In our case, the parameters of interest include a vector $\gamma$ of continuous-time selection coefficients (Malthusian fitnesses). We assume that the individual $\gamma$ values, one for each gene, are independent samples from a normal distribution of selection coefficients with some unknown mean $\mu$ and unknown variance $\sigma^2$. The values of $\mu$ and $\sigma^2$ are of interest, as are the species divergence time ($t$) and the vector of mutation parameters ($\theta$). For $k$ DPRS tables, there are $k$ selection coefficients and $2k$ mutation parameters. We adopt the bayesian approach because it enables estimates with specified degrees of confidence even for correlated parameters, it allows the sharing of data across DPRS tables to estimate the divergence time, and it provides a powerful computational tool. The goal is to use the observed data ($D$) to make inferences about the posterior distribution of the parameters ($\gamma$, $t$, $\theta$, $\mu$, $\sigma$). This posterior distribution is symbolized as $\pi(\gamma, t, \theta, \mu\, \sigma|\, D)$. The bayesian formulation of our model is given formally by

$$\pi(\gamma, t, \theta, \mu, \sigma|D) = \frac{P(D|\gamma, t, \theta)f(\gamma|\mu, \sigma)g(\mu|\sigma)h(\sigma)p(\theta)q(t)}{\int\int\int\int\int[P(D|\gamma, t, \theta)f(\gamma|\mu, \sigma)g(\mu|\sigma)h(\sigma)p(\theta)q(t)]\mathrm{d}\gamma\mathrm{d}t\mathrm{d}\theta\mathrm{d}\mu\mathrm{d}\sigma}$$

Properties of individual parameters in the posterior distribution are used for estimates and credible intervals (bayesian confidence intervals). The value of the expression $P(D|\, \gamma, t, \theta)$ is calculated from the independent Poisson distributions whose means are given in Table 1 (bottom), and the other functions are the assumed prior probability distributions. For example, $f(\gamma \mid \mu, \sigma)$ is a normal distribution with mean $\mu$ and variance $\sigma^2$, whereas $g(\mu \mid \sigma)$ and $h(\sigma)$ embody the hierarchical feature that $\mu$ and $\sigma$ are themselves treated as random variables, with $g(\mu \mid \sigma)$ a normal distribution and $h(\sigma)$ such that $1/\sigma^2$ is a gamma distribution. These choices are motivated by the facts that the sample mean of independent observations from a normal distribution, given $\sigma$, is itself normal, and that the distribution of the sample variance is gamma. We also assume that the prior $p(\theta)$ for mutation is gamma, because with this choice the distribution of these parameters in the posterior distribution has the same form as in the prior distribution, and that the prior $q(t)$ for the divergence time is uniform. These priors are all chosen to be 'uninformative' in the sense that the parameters in the posterior distribution are determined primarily by the current data and not by the characteristics of the prior distribution.

Many simplifications result from this bayesian formulation as applied to DPRS tables. For example, because the DPRS tables are independent, the multivariate distribution $f(D \mid \gamma, t, \theta)$ can be written as a product of univariate distributions $f(D \mid \gamma_i, t, \theta_i)$, where the subscript $i$ indexes the gene ($i$ = 1, 2, ..., $k$). Likewise, the multivariate normal distribution $g(\gamma \mid \mu, \sigma)$ can be written as a product of univariate normal distributions of the form $g(\gamma_i|\mu,\sigma)$. Because mutation enters the equations in the DPRS tables multiplicatively, updating the mutation parameters conditional on $\gamma$ and $t$ is straightforward. These simplifications also imply that many of the parameters become uncorrelated in their conditional distributions.

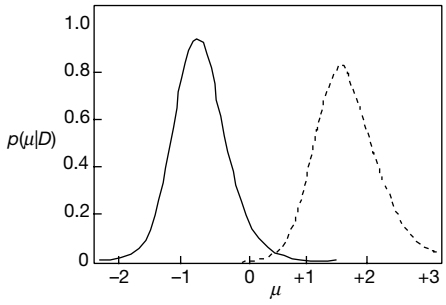


**Figure 2** Estimated distribution of the mean value of the selection parameter ($\gamma$) across genes in *Arabidopsis* (solid line) and *Drosophila* (dashed line).

*Drosophila*). After the burn-in, each chain was sampled every 100 generations, yielding 45,000 sample points for each species pair.

Using the results of MCMC, we estimated the mean of the posterior distribution of the selection intensity $\gamma$ for each gene. These are shown in rank order in Fig. 1, along with the 95% credible intervals. For the 12 *Arabidopsis* genes, 10 have their mean $\gamma < 0$, whereas for the 34 *Drosophila* genes, 32 have their mean $\gamma > 0$. Among the 12 *Arabidopsis* genes, 6 credible intervals are entirely negative (do not overlap 0), and among the 34 *Drosophila* genes, 9 are entirely positive. The detection of non-zero selection parameters for so many genes contrasts with conventional significance tests and demonstrates the power of the bayesian analysis.

The posterior distribution of $\mu$ also shows a significant difference between the organisms (Fig. 2). For *Arabidopsis* genes, most of the posterior distribution for the average selective effect of polymorphic or fixed amino-acid replacements has $\mu < 0$, and in fact the probability that $\mu > 0$ is about 0.03. In contrast, for *Drosophila* genes, most of the posterior distribution has $\mu > 0$, and indeed the probability that $\mu < 0$ is about 0.0004. The overall probability that $\mu$ for *Arabidopsis* is actually greater than $\mu$ for *Drosophila* is $P = 0.001$, which we estimated by comparing the 45,000 sample values for each data set in all possible pairs.

We conclude from Figs 1 and 2 that the average amino-acid replacement that is polymorphic or fixed in *Drosophila* is beneficial, whereas the average amino-acid replacement that is polymorphic or fixed in *Arabidopsis* is slightly deleterious. This result is consistent with a recent decrease in effective population size in *A. thaliana*, which is predicted[3–6] to result from its mode of reproduction largely by self-fertilization[15]. *A. lyrata* reproduces by outcrossing[16], and phylogenetic evidence indicates that partial selfing is a derived trait[17].

If *A. thaliana* and *A. lyrata* were to share many ancestral polymorphisms, some of these would be scored as fixed differences. However, these species show substantial sequence divergence (Table 1, top) and are estimated to have diverged 5–6 Myr ago[17]. Our analysis indicates that the 95% credible interval for the divergence time $t$ between the species, in multiples of the haploid effective population size, is 6.9–11.2. Considering that a newly arisen neutral mutation that is destined to become fixed has an average fixation time of $t = 2$ (as a multiple of the haploid effective size), extensive shared polymorphism seems unlikely. For the *Drosophila* species, which diverged 1–3 Myr ago[18], the 95% credible interval for $t$ is 3.7–4.8.

Tight linkage of the nucleotide sites within a gene is also an issue. However, computer simulations indicate that the effect of linkage is to bias the estimates of $\gamma$ towards 0 (data not shown). Hence, linkage would tend to make the differences in the selection parameters between *Arabidopsis* and *Drosophila* seem somewhat smaller than they really are.

Beyond the present application, hierarchical bayesian analysis affords a theoretical framework for a science of evolutionary genomics. In principle it could discriminate between the evolutionary forces that affect proteins of diverse function, identifying effects specific to sex, tissue and developmental stage of expression, cellular location, three-dimensional structure and mechanism of action. The analysis could also sort out the evolutionary forces affecting upstream, downstream, intronic and intergenic regions of genes. Even in this era of high-throughput genomic sequencing, the acquisition of genome-wide polymorphism-divergence data presents a formidable challenge.

Received 6 August 2001; accepted 3 January 2002.

1. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351,** 652–654 (1991).
2. Sawyer, S. A. & Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics* **132,** 1161–1176 (1992).
3. Charlesworth, D., Morgan, M. T. & Charlesworth, B. The effect of deleterious mutations on neutral molecular variation. *J. Hered.* **84,** 321–325 (1993).
4. Kondrashov, A. S. Muller's ratchet under epistatic selection. *Genetics* **136,** 1469–1473 (1994).
5. Caballero, A. & Santiago, E. Response to selection from new mutation and effective size of partially inbred populations. I. Theoretical results. *Genet. Res.* **66,** 213–225 (1995).
6. Wang, J. L., Hill, W. G., Charlesworth, D. & Charlesworth, B. Dynamics of inbreeding depression due to deleterious mutations in small populations: Mutation parameters and inbreeding rate. *Genet. Res.* **74,** 165–178 (1999).
7. Hartl, D. L., Moriyama, E. N. & Sawyer, S. A. Selection intensity for codon bias. *Genetics* **138,** 227–234 (1994).
8. Akashi, H. Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA. *Genetics* **139,** 1067–1076 (1995).
9. Gelman, A., Carlin, J. S., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* (Chapman & Hall, London, 1997).
10. Carlin, B. P. & Louis, T. A. *Bayes and Empirical Bayes Methods for Data Analysis* (Chapman & Hall, London, 2000).
11. Gilks, R., Richardson, S. & Spiegelhalter, D. J. *Markov Chain Monte Carlo in Practice* (Chapman & Hall, London, 1996).
12. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21,** 1087–1091 (1953).
13. Geman, S. & Geman, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* **6,** 721–741 (1984).
14. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7,** 457–511 (1992).
15. Abbott, R. J. & Gomes, M. F. Population genetic structure and the outcrossing rate of *Arabidopsis thaliana*. *Heredity* **62,** 411–418 (1989).
16. Savolainen, O., Langley, C. H., Lazzaro, B. P. & Freville, H. Contrasting patterns of nucleotide polymorphism at the *Adh* locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. *Mol. Biol. Evol.* **17,** 645–655 (2000).
17. Kusaba, M. *et al.* Self-incompatibility in the genus *Arabidopsis*: Characterization of the *S* locus in the outcrossing *A. lyrata* and its autogamous relative *A. thaliana*. *Plant Cell* **13,** 627–643 (2001).
18. Caccone, A., Amato, G. D. & Powell, J. R. Rates and patterns of scnDNA and mtDNA divergence within the *Drosophila melanogaster* subgroup. *Genetics* **118,** 671–683 (1988).

**Supplementary Information** accompanies the paper on *Nature's* website (http://www.nature.com) and on the authors' website (http://www.oeb.harvard.edu/hartl/lab).

**Acknowledgements**

We thank D. Weinreich and D. Rand for providing the *Drosophila* data, and A. Kondrashov for numerous suggestions for improving the presentation. This work was supported by grants from the US Public Health Service, the US National Science Foundation, Howard Hughes and Marshall Sherfield Fellowships to C.D.B., and an Alfred P. Sloan Foundation Young Investigator Award to M.D.P.

**Competing interests statement**

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.L.H. (e-mail: dhartl@oeb.harvard.edu).