Estimating the distribution of mutational effects in Drosophila

Laurence Loewe[1], Brian Charlesworth,
Carolina Bartolomé[2] and Véronique Nöel
Institute of Evolutionary Biology,
University of Edinburgh, UK

Why is this important?

The distribution of mutational effects (DME) is of fundamental importance in evolutionary theory. For example, it is not possible to predict observable parameters such as diversity within species and divergence between species without making assumptions about the distribution of mutational effects. Thus we need to know the DME if we want to understand genome evolution. Several methods have been devised to estimate the DME (ref.1), but since all methods are indirect, it is very difficult to assess their reliability. We developed a new method that will hopefully use new aspects of observable data to increase overall reliability of DME estimates.

Ecological Setting

Our method is applicable to settings where large sets of diversity data are available for at least two closely related species with significantly different effective population sizes N_e (Fig. 1). Then we need an *a priori* criterion to divide diversity into that which is from effectively neutral and that which is from sites that are effectively under selection. Here we assume that partitioning into silent and non-synonymous sites will achieve this. Finally, the ecology of these species must be similar enough to assume that they are all subject to the same absolute DME, which we assume to be unimodal. In other words, some mutations that are removed as deleterious in the species with the large N_e can accumulate as effectively neutral in the species with the small N_e.

Principle

We exploit the effects of this difference in N_e on π_A, the diversity at the non-synonymous sites that we assume to be under selection. Assuming a constant mutation rate μ and genic selection coefficient *s* suggests the following simple principle: *Effectively neutral sites (N_es<<1):* diversity is mainly determined by N_e, since mutation and drift are most important and selection is negligible (π=4N_eμ). *Strongly selected sites (N_es>>1):* diversity is mainly determined by selection, since mutation and selection are most important and drift is negligible (π=2μ/s). *Intermediate sites:* use mutation-selection-drift-equilibrium to predict diversity (ref.2).

Algorithm

1. Use diversity at silent sites and assume a mutation rate μ to estimate N_e in both species.
2. Assume an arbitrary unimodal DME that is characterized by two parameters only (shape and location).
3. Compute π_A for each species by integrating over the full DME (i.e. sum up the contribution of all potential mutational effects between neutral and lethal, weighted by their frequency as determined from the DME).
4. Compare predictions and observations:
   a. Predictions are correct: Bingo! You found the right shape and location.
   b. Predictions outside of acceptable error-margins: Go back to (2) and try better parameters.

Much of the success of this algorithm depends on an intelligent way to guess increasingly better values for shape and location of the DME. We use the 'Amoeba'-algorithm as implemented in the statistical script language R(ref.3).

Extensions & Simplifications

We assume above that all amino-acid substitutions are effectively deleterious. If c_n% of them are completely neutral (see Fig.2), then π_A is easy to correct. If all mutations are either neutral or strongly deleterious (bimodal DME), then formulae simplify greatly. Key results like s_h, the harmonic mean of effectively deleterious *s* and the fraction of effectively neutrals c_{ne} are remarkably stable to changes in the assumptions.

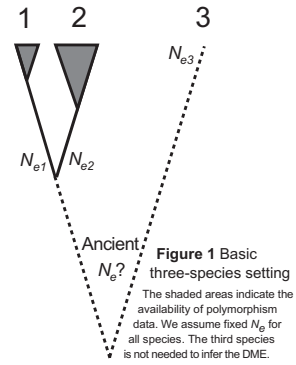


Figure 1 Basic three-species setting
The shaded areas indicate the availability of polymorphism data. We assume fixed N_e for all species. The third species is not needed to infer the DME.

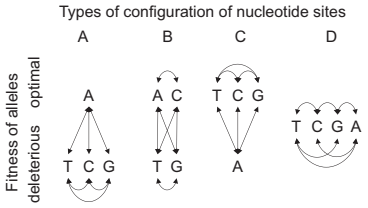


Figure 2 Fitness relevant configurations of sites.
The four alternatives A, T, C and G stand for any of the four nucleotides that can occupy a base pair. This work assumes that configurations B and C do not exist and captures the frequency of D in the parameter c_n.

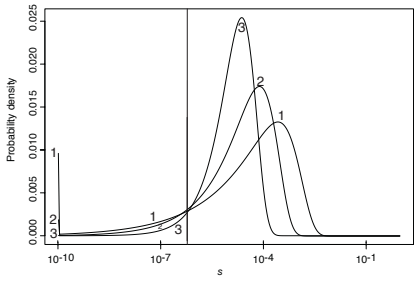


Figure 3 Realistic distributions of mutational effects
Estimated from the variance weighted data assuming c_n = 0%, 2.5% and 5% (curves 1,2 and 3, respectively) and a gamma DME. The spike below 10^{-10} corresponds to the integral down to 0. The vertical line correspnds to N_es=0.5 for *D. miranda*. See Table 1 for more details.

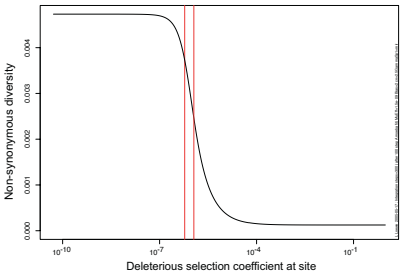
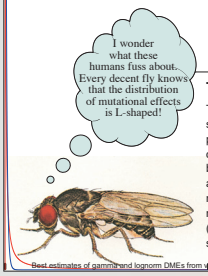


Figure 4 Contribution to diversity declines with increasing *s*
This graph shows the diversity at non-synonymous sites for a population of *D. miranda* where all mutations have the effect specified on the x axis and c_n=2.5%.

c_n(%)	α	*loc*	Species	N_es (am)	N_es (hm)	N_es (5%)	N_es (95%)	CV	c_{ne} (%)	P_{a1} ; P_{a2} (%)	P_{a3} (%)
0	0.299 (0.0782 / 0.741)	0.000266 (1.14x10^{-5} / 17000)	*mir*	255 (11.6 / 95700) 263 (12.4 / 97500)	8.35 (3.49 / 17.9) 13.3 (4.82 / 31.7)	1.31 (0.843 / 3.01) 2.22 (1.40 / 5.06)	1100 (36.9 / 558000) 1100 (37.8 / 560000)	1.67 (1.08 / 2.15) 1.64 (1.02 / 2.11)	12.7 (7.91 / 17.0) 15.5 (11.0 / 20.2)	-42.8 (-125 / 14.7)	0.902 (-83.5 / 60.3)
			pso	1370 (55.3 / 521000) 1400 (56.2 / 534000)	14.4 (8.69 / 25.0) 23.0 (11.8 / 43.8)	2.50 (1.55 / 5.47) 3.72 (2.39 / 8.47)	6220 (188 / 3110000) 6240 (188 / 3200000)	1.74 (1.14 / 2.22) 1.72 (1.12 / 2.20)	7.57 (2.35 / 13.4) 9.19 (3.65 / 14.7)	15.7 (-72.1 / 73.7)	
2.5	0.448 (0.0996 / 1.32)	7.67x10^{-5} (7.12x10^{-6} / 89.5)	*mir*	70.6 (6.96 / 76500) 73.1 (7.32 / 78800)	6.86 (3.19 / 17.2) 10.1 (3.99 / 29.9)	1.21 (0.877 / 2.96) 1.93 (1.37 / 4.80)	274 (18.4 / 463000) 275 (18.8 / 476000)	1.39 (0.832 / 2.08) 1.36 (0.789 / 2.05)	11.4 (5.73 / 16.6) 14.4 (9.04 / 19.3)	-31.1 (-117 / 27.9)	7.93 (-78.7 / 57.7)
			pso	382 (35.4 / 392000) 388 (35.4 / 401000)	14.7 (9.48 / 28.4) 21.5 (12.2 / 45.5)	2.73 (1.77 / 7.31) 3.96 (2.59 / 9.77)	1550 (97.2 / 2450000) 1550 (97.2 / 2490000)	1.45 (0.865 / 2.13) 1.43 (0.857 / 2.12)	6.58 (2.87 / 13.4) 7.95 (3.30 / 14.6)	16.4 (-70.0 / 59.1)	
5	0.831 (0.228 / 1.75)	2.33x10^{-5} (6.84x10^{-6} / 0.0179)	*mir*	20.4 (6.53 / 11400) 21.1 (6.72 / 11600)	5.42 (3.39 / 18.9) 7.07 (4.08 / 31.5)	1.21 (0.931 / 3.53) 1.76 (1.42 / 5.40)	63.6 (15.8 / 55700) 63.7 (16.2 / 55800)	1.05 (0.737 / 1.95) 1.02 (0.702 / 1.93)	9.27 (6.40 / 16.1) 12.2 (8.41 / 19.0)	-11.2 (-113 / 22.6)	15.2 (-78.3 / 38.3)
			pso	112 (31.7 / 80400) 113 (32.0 / 81200)	17.4 (10.5 / 35.2) 21.5 (12.5 / 53.0)	4.28 (2.13 / 10.9) 4.73 (2.93 / 13.1)	362 (80.1 / 413000) 362 (80.2 / 415000)	1.09 (0.753 / 1.98) 1.08 (0.746 / 1.97)	6.01 (5.08 / 12.1) 6.72 (5.22 / 13.7)	14.4 (-73.1 / 35.6)	

Table 1 Estimates of the (gamma-) distribution of mutational effects for variance-weighted data

The N_es values reported here describe the distribution of deleterious mutational effects (DME) for all effectively non-neutral and non-lethal sites, where (am) denotes the arithmetic mean and (hm) the harmonic mean of such sites; (5%) and (95%) represent the lower and upper fifth percentiles of this truncated distribution; *mir* and *pso* refer to *D. miranda* and *D. pseudoobscura* respectively. For a given species and value of c_n, the upper and lower values in all N_es columns correspond to assumed borders to neutrality of N_es = 0.5 and 1, respectively. Values in brackets indicate the lower and upper approximate fifth percentiles from bootstrapping across genes, after omitting all values that did not give a perfect fit to the data. c_n is the fraction of completely neutral mutations assumed in the calculations; c_{ne} is the fraction of effectively neutral mutations (reported for N_es ≥ 0.5). All estimates assume a mutational bias of κ=2, a mutation rate of μ=1.5x10^{-9} and a gamma distribution of mutational effects (DME) with the given shape parameter α and location parameter β, which is computed from the given arithmetic mean of *s* (β=α/s_{am}). All predictions based on DMEs reported here fitted observations at 6 digits or more. P_{a1}, P_{a2} and P_{a3} gives the proportion of adaptive substitutions assuming an ancient N_e of *D. miranda*, *D. pseudoobscura* or the mean of both, respectively.



Best estimates of gamma- and lognorm DMEs from variance weighted data, assuming c_n=2.5%, μ=1.5e-9, κ=2. Ticks mark 2% heterozygous effects; red=dlnorm(x,meanlog=-9.4761,sdlog=3.7), blue=dgamma(x,shape=0.462,scale=0.0001504); plotted in R

References: * ** *** 1. Too many to list. Examples are Sawyer SA, Hartl DL (1992) Genetics 132:1161-1176. *** Bustamante CD, et al. (2003) Theor Popul Biol 63:91-103. *** 2. McVean GAT, Charlesworth B (1999) Genetical Research 74:145-158. *** Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge. R (2005) http://www.r-project.org/ *** 4. Wakeley J, Hey J (1997) Genetics 145:847-855.

Results

We obtained variance-weighted mean diversities for a set of 17 and 14 loci from *D. miranda (mir,* π_A=0.086%; θ_S=0.502%) and *D. pseudoobscura (pso,*π_A=0.294%; θ_s=2.86 %) and estimated N_e (*mir*=0.84x10^6; *pso*=4.77x10^6) from the number of segregating sites to minimize influence from recent population expansions in *pso*. Assuming a gamma DME and mutation-selection-drift equillibrium with backmutations, we estimate the DMEs shown in Fig. 3 with parameters reported in Tab. 1. Results show that most mutations are strongly deleterious, DME shapes are very leptokurtic, harmonic means of *s* are much more meaningful than arithmetic means and results do not depend critically on assumptions about the parameters c_n, μ, κ.

Adaptive mutations

In the presence of an outgroup (*D. affinis* in our case) one can attempt to quantify the proportion of adaptive amino-acid substitutions P_a by comparing observed and predicted substitution rates. However, results are very sensitive to assumptions about ancestral population sizes and confidence intervals are huge. While it is not uncommon to predict some fraction of advantageous mutations with this method, we cannot place any confidence in such a result, since it usually includes 0 (i.e. no adaptive mutations). Perhaps independent estimates of ancestral population sizes (ref.4) and inclusion of more genes in the diversity surveys will one day lead to better estimates.

[1]Email: Laurence.Loewe@evolutionary-research.net Address: Ashworth Laboratories, King's Buildings, West Mains Road, Edinburgh EH9 3JT, Scotland, UK
[2]Present address: Unidad de Xenética Evolutiva, Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain.
This poster was first presented at "ESEB X: 10th Congress of the European Society for Evolutionary Biology", 15th-20th August 2005, Jagiellonian University, Krakow, Poland; this version: 29 Nov 2005.
See the full report in Loewe L, Charlesworth B, Bartolomé C, Nöel V (2006) "Estimating selection on non-synonymous mutations", Genetics (online 19 Nov 2005, doi 10.1534/genetics.105.047217)