MK test software v2.0

What the software does: The software implements various multi-locus versions of the McDonald-Kreitman test (McDonald & Kreitman 1991). This test compares divergence and polymorphism data from two classes of site, one of which is assumed to be selectively neutral; these are assumed in the notes below to be non-synonymous and synonymous changes from protein-coding genes. The tests estimate a statistic α, which can be viewed as the proportion of the divergence attributable to positive selection (Smith & Eyre-Walker 2002).
The software implements most previous estimators of α (Fay et al. 2001; Smith & Eyre-Walker 2002; Bierne & Eyre-Walker 2002; Welch 2006), and also allows formal likelihood-based tests for differences in α between different classes of loci (Betancourt et al. in press).

Compilation: The software requires the GNU Scientific library (Galassi et al. 2004), which is available free from http://www.gnu.org/software/gsl/. After this library is installed, the software can be compiled with the following lines.

```
$ tar -xzf MKtest-2.0.tar.gz
$ cd MKtest-2.0
$ ./configure
$ make
```

Running the software: The software is run from a terminal by typing:

$ ./Mktest [*optional arguments*] *inputfilename*

for example "./Mktest -a 1 -p 2 -o my.outputfile example_data.csv"

Optional command line arguments specify the estimators, models, output file, etc. They are explained in detail below. A full list of options can also be found by specifying "-?".

Credits: Version 1.0 of the software was described in Welch (2006). The "original" methods are closely based on Bierne and Eyre-Walker (2004), with elements from Sawyer and Hartl (1992). The code contains routines by Ted Phelps, and from the GNU Scientific library (Galassi et al. 2004).

Changes from Version 1.0: The software is less user unfriendly. Labelling of models now differs from Welch (2006), and the expected neutral divergence parameter is redefined so that it does not scale with N_e. Despite these changes, all previous models can still be run, as can many new parameterisations that make more biological sense.

Input file: Input data should be placed in a comma-separated file (.csv), with each line containing observations for a single locus. For each locus, the required data are:

1. D_N and D_S: Integer-valued counts of between-species divergence for non-synonymous and synonymous substitutions (or other neutral and non-neutral substitutions)
2. P_N and P_S: Integer-valued counts of within-species polymorphism for non-synonymous and synonymous mutants
3. L_N and L_S: The number of non-synonymous and synonymous sites in the gene. These need not be integers, and should ideally use the "mutational opportunity" definition of a site (Bierne and Eyre-Walker 2003). Separate counts are given for divergence and polymorphism data in case they differ (e.g., if some sites were unalignable between species, but alignable within species)
4. *alleles*: The number of alleles sampled to get the polymorphism counts

There are also optional data items:

5. *Chr*: This flag was used with the original *Drosophila* data sets to indicate *X*-linked versus autosomal loci. X-linked loci are indicated with a zero, and any other integer will specify an autosomal locus. Depending on the model specified, the effective population size of *X*-linked genes is set at 3/4 of the value for autosomal loci, or allowed to vary as a second free parameter.
6. *Class*: An arbitrary integer flag used to specify different classes of locus (e.g., immunity versus non-immunity genes). Different classes can be assigned different parameters, e.g., different α values or mutation rates.
7. *r*: recombination rate. This is used in models where the expected neutral polymorphism is a linear function of the recombination rate

The order of the 12 items in each row of the input file should be as follows:

1	2	3	4	5	6	7	8	9	[10	11	12]
D_N	L_N(D)	P_N	L_N(P)	D_S	L_S(D)	P_S	L_S(P)	*alleles*	*Chr*	*Class*	*r*

It is important to adhere to this order or the results will be nonsense. Any of the optional parameters can be left off the end, but if you want to include later parameters, e.g., recombination rate, then all earlier parameters must also be specified for that locus.

An example data file "example_data.csv" is included with the software.

Command line arguments: The possible command line arguments are as follows:

-a *option_no* [0-5 or 999] (also -*c alpha_val*, -*A,* -*m min_Ps*)

If -a 999 is specified, heuristic estimators of the average α across loci are calculated. These estimators were introduced by Fay et al. ("FWW") and Smith and Eyre-Walker ("SEW"). Confidence intervals in parentheses are generated from 10,000 bootstraps of the data. In square brackets is the proportion of bootstraps for which $\alpha<0$. Also calculated is the Mantel-Haenszel *z* statistic ("MHz") for the comparison of multiple contingency tables. The SEW estimator deals with heterogeneity in N_e across loci, and works best when small P_S values are excluded (Smith and Eyre-Walker 2002; Welch 2006); the argument -*m min_Ps* can be used for this purpose.

If -a is set at 0-5 (or unspecified) maximum likelihood estimates of α are calculated. The six options allow for different assumptions about the way α varies between loci.

option	Description
0	α is fixed at zero at all loci
1	[default] α takes a single value at all loci, and this is estimated from the data
2	α at each locus is a random variable drawn from a Beta distribution. The two parameters of this Beta distribution are estimated from the data (Welch 2006)
3	α at each locus is a random variable drawn from a two-spiked distribution. The three parameters of this distribution are estimated from the data (Welch 2006)
4	α can take a unique value at each class of locus as specified in the input file
5	α can take a unique value at each locus

In addition, the argument -c *alpha_value* is used to fix α to a single specified value at all loci. This is useful for obtaining likelihood-based confidence intervals on α estimates (i.e., it can be used to find the values of α that reduce the log likelihood by any given amount).

Finally, instead of α, the option -*A* provides estimates of the number of adaptive substitutions per site (= $\alpha D_N/L_N$). This seems to work best when the parameter is allowed to take a unique value at each locus (achieved by setting arguments: -*a* 5 -*A*). Note that the screen output still refers to "alpha" even if this option is specified.

-p *option_no* [1-5]

This argument determines the way that the expected neutral diversity per site is parameterised. This quantity is often written as $\theta=4N_e\mu$. The five options are as follows.

option	Description
1	[default] θ takes a single value at all autosomal loci, and is set at ¾ of this value at X-linked loci. The best-fit value is estimated from the data. (X-linked and autosomal loci are specified by the *Chr* flag in the input file).
2	θ can take a unique value at each locus
3	θ can take a unique value at each class of locus (as specified by the *Class* flag in the input file)
4	θ takes one value at autosomal loci, and another value at X-linked loci (as specified by the *Chr* flag in the input file).
5	θ is assumed to be a linear function of the recombination rate as obtained from the input file. Parameters of the equation $\theta = a + br$ are then estimated from the data.

-f option_no [0-4]

The model assumes that a fraction 1-*f* of mutants are under strong purifying selection. This argument determines how *f* varies across loci. The five options are as follows.

option	Description
0	$f = 1$ at all loci, so there is no strong purifying selection.
1	[default] *f* takes a single value at all loci, and this is estimated from the data.
2	*f* can take a unique value at each locus
3	*f* can take a unique value at each class of locus as specified in the input file
4	*f* takes one value at autosomal loci, and another value at X-linked loci

-g option_no [0-2] (also -*R*)

An alternative way of modelling selective constraint is with an exponential distribution of selection coefficients with mean $\gamma = 4N_es$. This model is explained in Welch (2006) but is probably not useful, given the lack of information in the data to fit a distribution.

option	Description
0	[default] $\gamma = 0$: mutations are either effectively neutral or strongly selected.
1	γ takes a single value at all autosomal loci; γ reduced by 25% at X-linked loci.
2	γ can take a unique value at each locus

By default, the exponential distribution of selection coefficients is partially reflected (allowing for back-mutation), -*R* makes the distribution a one-sided exponential.

-t option_no [1-4] (also *-L alleles* and *-B*)

This argument determines the way that the expected neutral divergence per site (μt) is parameterised. [Note, the definition of this parameter has changed from Welch 2006, and is no longer scaled by N_e]. The options for μt are as follows:

option	Description
1	[default] μt takes a single value at all loci, and this is estimated from the data.
2	μt can take a unique value at each locus
3	μt can take a unique value at each class of locus as specified in the input file
4	μt takes one value at autosomal loci, and another value at X-linked loci

Note that the expected *measured* divergence is assumed to have been inflated by unrecognised polymorphic variants in the outgroup sequence(s) (Sawyer & Hartl 1992; Welch 2006). The correction that is applied by default assumes that divergence has been inferred from a single sequence, i.e., not "polarized" along a single lineage.

If divergence *was* polarized and measured from an inferred ancestral sequence, then the option *-L alleles* should be set, with *alleles* set to the number of sequences used to infer the ancestral sequence. (e.g., if the common ancestor of human-chimp was inferred using one chimp and one gorilla sequence, then *-L* 2 should be set). To remove all of these corrections, use the option *-B*. In this case, the uncorrected equation $E[D_S] = \mu t\ L_S$, is applied. This is equivalent to the estimator of Bierne & Eyre-Walker (2004). These options should make little difference if the between-species divergence is much greater than within-species diversity; if this is not so, then estimates are sensitive to small changes in assumptions, and so should be treated with caution.

-o outputfilename

This specifies a filename for saving the estimated parameter values and final likelihood scores. This information is displayed on screen if no filename is entered. If the file already exists, it will be overwritten.

-P num_permutations

This is used to generate bootstrap or permutation tests for differences in α values.

A positive argument, e.g., *-P* 1000, will repeat the likelihood estimation with the *Class* flags permuted at random among the loci. A negative argument, e.g., *-P* -1000, will carry out a bootstrap, sampling loci with replacement within classes, before reestimating the parameters. For either a bootstrap or a permutation, each estimation is repeated as a check for convergence. Both runs are printed to the output file.

-X excluded_*class*, -*x excluded_locus*, -*K only_class*, -*k only_locus*

These arguments are used to obtain estimates for a single locus (-*k*) or class of loci (-*K*) or to exclude a single locus (-*x*) or class of loci (-*X*) without having to change the input file.

-*r cooling_rate*, -*i number_of_rounds*

These options might be useful if the simulated annealing algorithm isn't converging very well. They specify the cooling rate, and the maximum number of iterations.

The output: The software uses a simulated annealing algorithm, and while running, prints the log-likelihood value, the "temperature" and a sample parameter estimate every few rounds. These should all settle down to fixed values, although for parameter-rich models, this may never happen exactly. If the algorithm doesn't settle down adequately, then either the model is too big (too many parameters) or the algorithm is working poorly, in which case the iterations or cooling schedule can be adjusted using command line arguments (above). When estimation has finished, parameter estimates (plus any permutations or bootstraps) are printed to the output file (or screen if this has not been specified).

Also printed to the screen are the values of some model selection statistics (Welch 2006), and summary statistics about α.

References:

Betancourt, AJ, JJ Welch & B Charlesworth 2009 Reduced effectiveness of selection caused by a lack of recombination on the dot chromosome of *Drosophila americana. Curr. Biol.*

Bierne, N & A Eyre-Walker 2003 The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics* 165:1587-97

Bierne, N & A Eyre-Walker 2004 The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol. Biol. Evol.* 21: 1350-1360

Fay, JC, GJ Wycoff, & C-I Wu 2001 Positive and negative selection on the human genome. *Genetics* 158: 1227-1234

Galassi, M, J Davies, J Theiler, B Gough, G Jungman *et al*. 2004 GNU Scientific Library Reference Manual: Edition 1.6, for GSL Version 1.6. 0-541617-3-4.

McDonald, JH & M Kreitman 1991 Adaptive evolution at the Adh locus in *Drosophila*. *Nature* 351: 652-4

Sawyer, SA & DL Hartl 1992 Population genetics of polymorphism and divergence. *Genetics* 132: 1161-76

Smith, NGC and A Eyre-Walker 2002 Adaptive protein evolution in *Drosophila*. Nature 415: 1022-1024

Welch, JJ 2006 Estimating the genome-wide rate of adaptive protein evolution in *Drosophila*. *Genetics* 173: 821-37