Published online 30 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W157–W162***
doi:10.1093/nar/gkn337

Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites

Raquel Egea*, Sònia Casillas and Antonio Barbadilla

Genomics, Bioinformatics and Evolution Group, Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

Received February 18, 2008; Revised April 30, 2008; Accepted May 10, 2008

ABSTRACT

The McDonald and Kreitman test (MKT) is one of the most powerful and extensively used tests to detect the signature of natural selection at the molecular level. Here, we present the standard and generalized MKT website, a novel website that allows performing MKTs not only for synonymous and nonsynonymous changes, as the test was initially described, but also for other classes of regions and/or several loci. The website has three different interfaces: (i) the standard MKT, where users can analyze several types of sites in a coding region, (ii) the advanced MKT, where users can compare two closely linked regions in the genome that can be either coding or noncoding, and (iii) the multi-locus MKT, where users can analyze many separate loci in a single multi-locus test. The website has already been used to show that selection efficiency is positively correlated with effective population size in the *Drosophila* genus and it has been applied to include estimates of selection in DPDB. This website is a timely resource, which will presumably be widely used by researchers in the field and will contribute to enlarge the catalogue of cases of adaptive evolution. It is available at http://mkt.uab.es.

INTRODUCTION

The neutral theory of molecular evolution plays a fundamental role within the theoretical framework of population genetics, since it constitutes a null model on which statistical models can be developed to compare patterns of genetic variation and to infer the evolutionary forces governing molecular evolution (1). The relative amounts of neutral, deleterious and adaptive mutations and their selection coefficients can be estimated by various methods using population genetic data within and between species (2–4). The test proposed by McDonald and Kreitman (MKT) is among the most powerful and widely used tests and it has been applied to numerous model organisms (5–10).

The MKT compares the amount of variation within a species to the divergence between species at two types of sites, one of which is putatively neutral and used as the reference to detect selection at the other types of sites. As the test was initially described (2), these sites were synonymous (putatively neutral) and nonsynonymous sites in a coding region. Under the null hypothesis, all nonsynonymous mutations are expected to be neutral and then the ratio of nonsynonymous to synonymous variation within species (Pn/Ps) is expected to equal the ratio of nonsynonymous to synonymous variation between species (Dn/Ds). However, these ratios will not be equal if some nonsynonymous variation is under either positive or negative selection. Deleterious mutations rarely become fixed in the population and thus do not contribute to divergence, but they still make a significant contribution to polymorphism. As a result, Dn/Ds is lower, on average, than Pn/Ps. On the contrary, since mutations under positive selection spread through a population quickly, they do not contribute to polymorphism but have a cumulative effect on divergence and, as a result, Dn/Ds is greater, on average, than Pn/Ps. Some variants of the classical test have been developed to infer the proportion of substitutions that have been driven to fixation by the action of positive selection (11–13).

One main concern of the MKT refers to population dynamics. The test assumes that all nonsynonymous mutations are either strongly deleterious, neutral or strongly advantageous. However, when some of the nonsynonymous mutations are mildly deleterious, then Pn/Ps becomes sensitive to the population's demographic history, and Pn/Ps < Dn/Ds can hold if there has been an increase in population size (the reverse to what is expected under demographic stability) as a result of slightly deleterious mutations contributing disproportionately to divergence

*To whom correspondence should be addressed. Tel: +34 935 812730; Fax: +34 935 812387; Email: raquel.egea@uab.es

compared to polymorphism (14). The exclusion of low frequency polymorphisms has been used to detect adaptive selection as it increases the power of the MKT (15,16). However, it may also make the test more sensitive to increases in the effective population size (14) and thus in these cases the history of the population must be known in order to assure constancy of the effective population size.

Caution should be taken under other less likely conditions. For example, with increasing between-species divergence, the probability of multiple mutational hits at the same site increases. Because these hidden changes will be biased towards the faster evolving sequence category, they could skew the results. It is then advisable to incorporate sequence divergence estimates that take multiple mutational hits into account.

Even though the standard test compares synonymous and nonsynonymous sites in coding regions, the MKT can potentially be generalized to any two types of sites provided that one of them is assumed to evolve neutrally and that both types of sites are closely linked in the genome. These variations to the classical test may be used to infer selection at protein-binding sites (17), preferred codons (18), 5′ regions of genes (19), or intergenic noncoding regions (15,20). The test can even be used to compare sites sampled from multiple genomic regions if a proper multilocus test is applied (11).

There is some software available to perform MKTs, among which DnaSP (21) is usually the preferred one. However, none of these programs can be used online. Given the extensive use of the MKT, it would be very convenient to have a web-based resource to perform this test. The standard and generalized MKT website is a novel resource where users can perform standard or generalized MKTs online. The website includes three separate interfaces to the program: the standard MKT, the advanced MKT and the multi-locus MKT. The standard MKT allows users to create 2×2 contingency tables by comparing two types of coding sites, such as synonymous and nonsynonymous changes, from the same coding region. The advanced MKT is an extension of the standard test, which allows comparing any two closely linked regions in the genome, including noncoding DNA. Finally, the multi-locus MKT performs single multi-locus tests on multiple genomic regions. At a time in which DNA polymorphism data from hundreds of species are being sequenced and poured into public databases, the standard and generalized MKT website appears to be a timely resource which will presumably be widely used by researchers in the field.

METHODS

Input sequences

The input for the MKT is a set of homologous sequences from two related species in FASTA format. Two or more sequences must be available for one of the species, from which polymorphism will be calculated, and at least one sequence in the other species (for divergence estimates). However, polymorphism data can also be included for both species, and in this case polymorphism will be added up together. The user can also limit the region of each sequence that will be analyzed in the annotation boxes. Input sequences can already be aligned as gapped FASTA-formatted sequences, or be alternatively aligned on our server previous to the MKT analysis. In the latter case, users can choose the alignment program, either Muscle (22) or ClustalW2.0 (23), and its parameters.

Types of sites that can be analyzed

Depending on the input genomic region different types of sites can be analyzed. The genomic region can be either coding or noncoding. Within coding regions, users can choose to analyze synonymous sites, nonsynonymous sites, 4-fold degenerate sites, 2-fold degenerate sites and nondegenerate sites. Noncoding sites are analyzed as a single class of sites. Only sites that are mutually exclusive can be selected for an analysis and one of them can be freely specified to be used as the neutral class. Results will show all the possible pairwise comparisons among the selected classes of sites and the neutral class. See the Help section in the website for details.

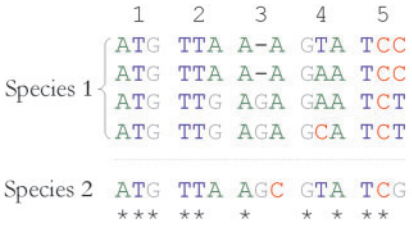
Treatment of sites or codons with multiple changes

When performing the test, counts in the contingency table can be either (i) sites (e.g. positions in the alignment which are either polymorphic within a species or divergent among species) or (ii) changes (i.e. the estimated number of mutations that have occurred in the position since the ancestor of both analyzed species). Sites and changes are computed differently (Figure 1). When *sites* are analyzed, each site can be counted only once. E.g. if a site has more than two variants in a species, it will be counted as one polymorphic site (Figure 1, codon 4). For the same reason, when a site is polymorphic and divergent at the same time, this site is not taken into account because it cannot be categorized into a single class (Figure 1, codon 5). Otherwise, when *changes* are analyzed, one site (position) might eventually involve several changes (mutations) if it has more than two variants, and thus that site will be counted more than once in the contingency table (Figure 1, codon 4). In this case, there is not any statistical problem by considering the same site as polymorphic and divergent at the same time (Figure 1, codon 5).

When sites cannot be classified into a unique class, only changes are computed. This is the case of synonymous and nonsynonymous changes in coding regions. In any other type of site, both the number of changes and the number of sites are computed. Then, when a comparison includes synonymous or nonsynonymous changes, only changes are compared; but if the comparison includes two analyses in which both changes and sites can be analyzed, then two comparisons are obtained: one for changes and the other for sites.

Estimation of the number of synonymous and nonsynonymous changes

We use a maximum parsimony criterion to estimate the number of synonymous and nonsynonymous changes after excluding those codons with gaps or undetermined nucleotides in any of the sequences (Figure 2).



Codon		# Changes	# Sites
#	position		
2	3	1 polymorphic	1 polymorphic
3	3	0	0
4	2	2 polymorphic	1 polymorphic
5	3	1 polymorphic 1 divergent	0

Figure 1. Treatment of sites or codons with multiple changes. In a coding region, both sites and changes are counted on each codon. In the alignment above, codon number 2 is polymorphic. Because it contains two variants, the MKT website considers one polymorphic change and one polymorphic site. Since codon number 3 contains a gap, the divergent position is not taken into account, neither as change nor as site. In codon number 4, since there are three variants in the polymorphic position, the MKT website considers two polymorphic changes but one polymorphic site. The third position of codon number 5 is polymorphic and divergent at the same time. Therefore, the MKT website considers two changes: one polymorphic and one divergent. However, since the site cannot be categorized in a unique class, it is not taken into account when counting sites.
*Indicate nonvariable positions.

The number of estimated polymorphic changes is computed for each species independently and then added together. For each codon in the single-species alignment, all the different codons are connected by paths in which each step involves a single change. Then, the shortest path with the least number of required replacements is chosen as the most parsimonious path and the numbers of synonymous and nonsynonymous changes involved in this path are added up to the total. Some special cases apply. First, when one of the codons from the other species equals an intermediate codon in a path, this path are chosen as the most parsimonious one, and synonymous and nonsynonymous changes are added up accordingly (Figure 2, codon 8). Second, stop codons within the alignment are treated as another amino acid, and this is warned in the results (Figure 2, codon 10); they are only excluded if positioned at the last codon in the alignment. Furthermore, any paths involving intermediate stop codons are excluded except if both end codons are stop codons (Figure 2, codon 11). Finally, the number of divergent changes is estimated similarly, that is, by computing the different paths from each codon of one species to each codon of the other species and choosing the shortest path and with the least number of required replacements. See the Help section in the website for details.

Estimation of the number of degenerate sites/changes

The degree of degeneracy for each coding position in the alignment is assigned according to the genetic code (24), taking the first sequence as the template. Then the number of segregating sites/changes is computed position by position for all the sequences excluding those sites with gaps or undetermined nucleotides in any of the sequences.

Correction of divergence estimates for multiple hits

With increasing between-species divergence, the probability of multiple mutational hits at the same site increases. Jukes and Cantor's (25) correction to take multiple mutational hits into account is the simplest and most widely used estimate. All results can be visualized with and without this correction.

Exclusion of low-frequency variants

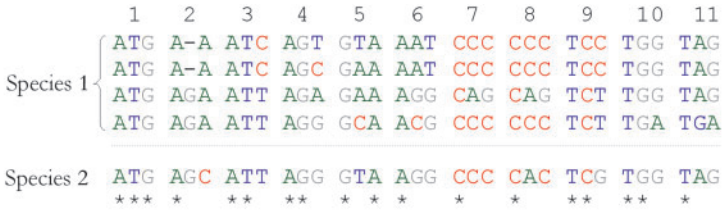
Polymorphisms segregating at low frequency typically represent slightly deleterious mutations that might bias the MKT results under certain conditions (see the Introduction section). These rare polymorphisms can be excluded from the analyses given a threshold frequency value.

Single-locus MKT

For each two classes of sites analyzed, the numbers of polymorphic and divergent sites/changes are shown in a 2×2 contingency table. For example, in the classical MKT the contingency table contains the number of synonymous and nonsynonymous polymorphic changes (Ps and Pn, respectively) and the number of synonymous and nonsynonymous replacements (Ds and Dn, respectively). From the contingency table, the following statistics are obtained: (i) its associated chi-square and *P*-value, (ii) the neutrality index [NI (26), computed as NI = (Pn/Ps)/(Dn/Ds)], and (iii) the proportion of adaptive substitutions [α (11), computed as $\alpha = 1 - \mathrm{NI}$ and ranging from $-\infty$ to 1]. NI indicates the extent to which the levels of polymorphic variation in the testing region depart from the expected in the neutral model. Under neutrality, Pn/Ps equals Dn/Ds and thus NI = 1. NI > 1 is interpreted as an excess of polymorphic variation compared to the neutral region due to negative selection, whereas NI < 1 is interpreted as an excess of variation between species due to positive selection.

Multi-locus MKT

This test allows the comparison of independent genomic regions in a single statistic test. A Mantel–Haenzel test (27) is performed to test the homogeneity of the multiple 2×2 contingency tables created for each locus independently, and to obtain a combined multi-locus statistic with the corresponding significance. Furthermore, the average proportion of adaptive substitutions is implemented here as $\bar{\alpha}$ (11), following a true multi-locus procedure.



Codon#	Ps	Pn	Ds	Dn	Explanation
1	0	0	0	0	All the sequences have the same codon
2	0	0	0	0	Codons with gaps are not taken into account
3	1	0	0	0	**ATC** (Ile) → **ATT** (Ile)
4	2	1	0	0	**AGT** (Ser) → **AGC** (Ser) → **AGA** (Arg) → **AGG** (Arg)
5	0	2	0	0	**GTA** (Val) → **GAA** (Glu) → **GCA** (Ala)
6	1	2	0	0	**AAT** (Asn) → ACT (Thr) → **ACG** (Thr) → **AGG** (Arg)
7	1	1	0	0	**CCC** (Pro) → CCG (Pro) → **CAG** (Gln)
8	0	2	0	0	**CCC** (Pro) → CAC (His) → **CAG** (Gln) There are the same codons as in Codon #7 but the outgroup codon is different. We assume that the most parsimonious path is the one that includes the outgroup codon
9	1	0	1	0	**TCC** (Ser) → **TCT** (Ser) → **TCG** (Ser)
10	0	1	0	0	**TGG** (Trp) → **TGA** (Stop) A warning will be displayed but this Stop codon will be counted as another amino acid
11	2	0	0	0	**TAG** (Stop) → TAA (Stop) → **TGA** (Stop) The intermediate Stop codon is allowed only because both end codons are Stop codons

Figure 2. Estimation of the number of synonymous and nonsynonymous changes. This alignment shows several possible situations and how they are resolved for counting the numbers of synonymous and nonsynonymous changes. In the table, counts for the numbers of polymorphic synonymous and nonsynonymous changes (Ps and Pn, respectively) and the numbers of divergent synonymous and nonsynonymous changes (Ds and Dn, respectively) are given for each codon. The table also shows the most parsimonious path (or one of them in case of tie) that connects all the different codons in each codon position.
*Indicate nonvariable positions.

PRACTICAL GUIDE TO THE MKT WEBSITE

Standard MKT

In the standard MKT interface, users can analyze different types of sites in a single coding region (see the Methods section). This interface is recommended for users who want to perform the standard MKT including also other types of sites (4-, 2-fold and nondegenerate sites) from the same coding region.

Advanced MKT

The advanced MKT interface allows comparing two different genomic regions that can be either coding or noncoding. However, note that for the analysis to make sense the two regions must be tightly linked in the genome. This interface is recommended for more complex analyses such as the comparison of two related coding regions (e.g. two exons from the same gene), two related noncoding regions (e.g. two introns from the same gene or the UTR of a gene with its introns), or a coding with a noncoding region (e.g. a gene with a related pseudogene or the exons with the introns of a gene). In any case, the user must select at least two classes of sites to be analyzed, one of which must be selected as neutral.

Multi-locus MKT

Using the multi-locus MKT interface, users can analyze multiple regions in a single multi-locus test. Note that in this case regions do not need to be linked in the genome, since the heterogeneity among regions is taken into account in the statistical tests (see the Methods section). Therefore, this interface can be used to discover whether a set of genes are evolving homogeneously and how.

Main parameters

In this section, users can choose different parameter values that apply to any of the interfaces. Specifically, they can choose to exclude low-frequency variants under a given threshold (i.e. rare polymorphisms), they can choose the genetic code to be used for coding sequences, and different parameters related to the alignment of the sequences: whether they want to align the sequences previous to

the analyses or not, and if so they can choose the alignment program, either Muscle (22) or ClustalW2.0 (23), the alignment order (either align all the sequences at the same time, or first align the sequences of each species independently and then join both alignments), and some parameters specific to the chosen alignment algorithm.

For huge alignments, the standard and generalized MKT website performance is better when input sequences have already been aligned. Therefore, in these cases we recommend to align the sequences first and enter manually curated alignments to the interface.

Output

The output of each analysis is shown in an HTML page. At the top of the page there is a summary table with all the performed comparisons, indicating, for each comparison, which is the neutral region and which is the testing region, and whether the analysis has been performed on sites or on changes (see the Methods section). There is also a link to the estimates for the corresponding comparison.

Below the summary table, basic information is provided on the main parameters chosen and the input sequences for each region: the number of input sequences for each species, the length of the alignment, the percentage of gaps within the alignment and the alignment itself in any selected formats. When both species have been aligned separately, the alignment for each species is also shown.

Finally, detailed results for each performed comparison are shown. These results include a 2×2 contingency table for each analyzed locus and a set of statistical estimates that depend on whether the test is a one-locus test or a multi-locus test (see the Methods section). Different warnings alert when stop codons interrupt a coding sequence or when the percentage of gaps within the alignment is $>30\%$.

Other sections of the website

The website includes an example for each type of test. In the standard MKT, a classical MKT for the *tre1* gene is performed on five sequences of *Drosophila melangaster* using *D. yakuba* as the outgroup. In the advanced MKT, the introns of the *Adh* gene are compared to 4-fold degenerate sites of the coding sequence of same gene using 10 sequences of *D. melanogaster* and four sequences of *D. simulans*. In the multi-locus MKT, synonymous and nonsynonymous coding sites are compared across two genes in *Mus musculus* and *M. spicilegus*, using four sequences for the *h2-eb* gene in the first species and two sequences for the same gene in the second species, and two sequences for the *irbp* gene in both species. The website also includes a detailed Help page and links to related resources: two databases with MKT estimates obtained using the MKT website [the *Drosophila* Polymorphism DataBase DPDB (28,29) and the Mammalia Polymorphism Database MamPol (30)], and PDA, a pipeline to estimate nucleotide diversity in different functional regions (31).

APPLICATIONS OF THE MKT WEBSITE

The standard and generalized MKT website facilitates the study of selection on population genetic data. It has been used, even before publication, to test the hypothesis that selection efficiency is positively correlated with effective population size in the *Drosophila* genus (Petit and Barbadilla, manuscript submitted for publication). They accepted the previous hypothesis and also found that adaptive selection on coding sequences and on preferred codons are positively correlated with population size. To perform the analyses they used a subset of data from DPDB (28,29) and the analyses included MKTs performed in the standard and generalized MKT website.

In addition, the source code of the standard and generalized MKT website has been incorporated into DPDB. This database contains polymorphism estimates in the *Drosophila* genus. After the incorporation of outgroup sequences for each analyzed coding region, we have included MKTs for these regions as a pilot test of the large-scale application of this website. Outgroup sequences are obtained through the mapping of the alignments to the genomic sequence of *D. melanogaster*. This genomic sequence is the outgroup sequence used for the MKT when the polymorphic alignment is not from *D. melanogaster*. Otherwise, the outgroup sequence is the homologous region in *D. simulans* obtained from the UCSC table browser (32). At the time of writing, the database contained 2889 analyzed coding regions. We could obtain reliable outgroup sequences for 2630 of these regions, for which we performed MKTs. All these analyses are available through the DPDB website at http://dpdb.uab.es.

A similar procedure is being applied to MamPol (30). The MamPol database contains polymorphism estimates in the *Mammalia* class. This implementation will greatly extend the current estimates of nucleotide selection in these species and will provide a large amount of data on which to test a wide range of hypotheses.

The standard and generalized MKT website is a user-friendly web server on which even nonpopulation geneticists can easily search for the evidence of selection at one or more loci. The many options and interfaces that the website includes turn it into a powerful resource within the today's landscape of applications for molecular evolutionary biologists.

AVAILABILITY

The standard and generalized MKT is a free resource online, open to all users without login requirement at http://mkt.uab.es.

ACKNOWLEDGEMENTS

The authors would like to thank Natalia Petit for helping in developing the software, Emilio Centeno for helping in developing the website, Miquel Ràmia for the design of the banner, and two anonymous reviewers for worthy improvements on the website and valuable discussions on the article. R.E. was supported by the Departament

d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya i el Fons Social Europeu (Grant 2005FI 00328). This work was funded by the Ministerio de Educación y Ciencia, Spain (Grant BFU-2006-08640). Funding to pay the Open Access publication charges for this article was provided by the Ministerio de Educación y Ciencia, Spain (Grant BFU-2006-08640).

Conflict of interest statement. None declared.

REFERENCES

1. Akashi,H. and Schaeffer,S.W. (1997) Natural selection and the frequency distributions of 'silent' DNA polymorphism in Drosophila. *Genetics*, **146**, 295–307.
2. McDonald,J.H. and Kreitman,M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, **351**, 652–654.
3. Sawyer,S.A. and Hartl,D.L. (1992) Population genetics of polymorphism and divergence. *Genetics*, **132**, 1161–1176.
4. Templeton,A.R. (1996) Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. *Genetics*, **144**, 1263–1270.
5. Moriyama,E.N. and Powell,J.R. (1996) Intraspecific nuclear DNA variation in Drosophila. *Mol. Biol. Evol.*, **13**, 261–277.
6. Fay,J.C., Wyckoff,G.J. and Wu,C.I. (2001) Positive and negative selection on the human genome. *Genetics*, **158**, 1227–1234.
7. Fay,J.C., Wyckoff,G.J. and Wu,C.I. (2002) Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature*, **415**, 1024–1026.
8. Bustamante,C.D., Nielsen,R., Sawyer,S.A., Olsen,K.M., Purugganan,M.D. and Hartl,D.L. (2002) The cost of inbreeding in Arabidopsis. *Nature*, **416**, 531–534.
9. Bustamante,C.D., Fledel-Alon,A., Williamson,S., Nielsen,R., Hubisz,M.T., Glanowski,S., Tanenbaum,D.M., White,T.J., Sninsky,J.J., Hernandez,R.D. *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature*, **437**, 1153–1157.
10. Shapiro,J.A., Huang,W., Zhang,C., Hubisz,M.J., Lu,J., Turissini,D.A., Fang,S., Wang,H.Y., Hudson,R.R., Nielsen,R. *et al.* (2007) Adaptive genic evolution in the Drosophila genomes. *Proc. Natl Acad. Sci. USA*, **104**, 2271–2276.
11. Smith,N.G. and Eyre-Walker,A. (2002) Adaptive protein evolution in Drosophila. *Nature*, **415**, 1022–1024.
12. Bierne,N. and Eyre-Walker,A. (2004) The genomic rate of adaptive amino acid substitution in Drosophila. *Mol. Biol. Evol.*, **21**, 1350–1360.
13. Welch,J.J. (2006) Estimating the genomewide rate of adaptive protein evolution in Drosophila. *Genetics*, **173**, 821–837.
14. Eyre-Walker,A. (2002) Changing effective population size and the McDonald-Kreitman test. *Genetics*, **162**, 2017–2024.
15. Andolfatto,P. (2005) Adaptive evolution of non-coding DNA in Drosophila. *Nature*, **437**, 1149–1152.
16. Charlesworth,J. and Eyre-Walker,A. (2008) The McDonald-Kreitman test and slightly deleterious mutations. *Mol. Biol. Evol.*, **25**, 1007–1015.
17. Jenkins,D.L., Ortori,C.A. and Brookfield,J.F. (1995) A test for adaptive change in DNA sequences controlling transcription. *Proc. Biol. Sci.*, **261**, 203–207.
18. Akashi,H. (1995) Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in Drosophila DNA. *Genetics*, **139**, 1067–1076.
19. Kohn,M.H., Fang,S. and Wu,C.I. (2004) Inference of positive and negative selection on the 5' regulatory regions of Drosophila genes. *Mol. Biol. Evol.*, **21**, 374–383.
20. Casillas,S., Barbadilla,A. and Bergman,C.M. (2007) Purifying selection maintains highly conserved noncoding sequences in Drosophila. *Mol. Biol. Evol.*, **24**, 2222–2234.
21. Rozas,J., Sanchez-DelBarrio,J.C., Messeguer,X. and Rozas,R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.
22. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
23. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
24. Hartl,D.L. (2004) *A primer of population genetics*. Sinauer Associates Inc., SunderLand, Massachusetts.
25. Jukes,T.H. and Cantor,C.R. (1969) *Evolution of protein molecules*. Academic Press, New York, pp. 21–132.
26. Rand,D.M. and Kann,L.M. (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Mol. Biol. Evol.*, **13**, 735–748.
27. Sokal,R.R. and Rohlf,F.J. (1995) edn. *Biometry: the principles and practice of statistics in biological research*, 3rd edn. W.H. Freeman and Co, New York.
28. Casillas,S., Petit,N. and Barbadilla,A. (2005) DPDB: a database for the storage, representation and analysis of polymorphism in the Drosophila genus. *Bioinformatics*, **21 (Suppl 2)**, ii26–ii30.
29. Casillas,S., Egea,R., Petit,N., Bergman,C.M. and Barbadilla,A. (2007) Drosophila polymorphism database (DPDB): a portal for nucleotide polymorphism in Drosophila. *Fly*, **1**, 205–211.
30. Egea,R., Casillas,S., Fernandez,E., Senar,M.A. and Barbadilla,A. (2007) MamPol: a database of nucleotide polymorphism in the Mammalia class. *Nucleic Acids Res.*, **35**, D624–D629.
31. Casillas,S. and Barbadilla,A. (2006) PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA. *Nucleic Acids Res.*, **34**, W632–W634.
32. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.