

# Unusual pattern of single nucleotide polymorphism at the *exuperantia2* locus of *Drosophila pseudoobscura*

SOOJIN YI<sup>1\*</sup> AND BRIAN CHARLESWORTH<sup>2</sup>

<sup>1</sup> Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637-1573, USA

<sup>2</sup> Institute of Cell, Animal and Population Biology, University of Edinburgh, King's Buildings, Edinburgh, EH9 3JT, UK

(Received 9 May 2003 and in revised form 28 July 2003)

## Summary

We have investigated the pattern of DNA sequence variation at the *exuperantia2* locus in *Drosophila pseudoobscura*. This adds to the increasing dataset of genetic variation in *D. pseudoobscura*, a useful model species for evolutionary genetic studies. The level of silent site nucleotide diversity and the divergence from an outgroup *Drosophila miranda* are comparable with those for other X-linked loci. One peculiar pattern at the *exu2* locus of *D. pseudoobscura* is a complete linkage disequilibrium between two SNPs, one of which is a replacement site. As a result, there are two distinct haplotype groups in our dataset. Based upon the comparisons with the outgroup sequences from *D. miranda* and *Drosophila persimilis*, we show that the newly derived haplotype group has lower diversity than the ancestral haplotype group. The pattern of protein evolution at *exu2* shows some deviation from the neutral model. Together, these and other characteristics of the *exu2* locus suggest the action of selection on the pattern of SNP variation, consistent with a partial selective sweep associated with the newly derived haplotype.

## 1. Introduction

The significance of *Drosophila pseudoobscura* as ‘the second model species’ in the genus *Drosophila* for genetic studies has been increasing. For example, the genome of *D. pseudoobscura* has been sequenced using a whole genome shotgun method, to approximately sevenfold coverage (<http://www.hgsc.bcm.tmc.edu/projects/drosophila>). In fact, this species has long been a subject of evolutionary studies. In particular, the studies of inversion polymorphisms in *D. pseudoobscura* in relation to its geographic distribution have made invaluable contributions to our understanding of genome evolution (see Powell, 1997, Chapter 3, for a review). A Northern American native, this species is estimated to have diverged from *D. melanogaster* around 25 million years ago (Mya) (Russo *et al.*, 1995).

Even though some aspects of evolutionary genetics of *D. pseudoobscura* are well known, the amount of data about DNA sequence variation investigated is small compared with that of *D. melanogaster* (see

Table 3 of Moriyama & Powell, 1996; also see Table 6 of Yi *et al.*, 2003). It is desirable to understand better the patterns of molecular variation in this species, which can elucidate crucial parameters such as the effective population size and the degree of population subdivision (Kovacevic & Schaeffer, 2000). From the available data, the underlying genetic variation in this species is comparable with that of *D. simulans* (Table 6 of Yi *et al.*, 2003).

The gene *exuperantia2* (*exu2*) first came to our attention in the course of an investigation of genome evolution in the sibling species *Drosophila miranda* (Yi & Charlesworth, 2000; Bachtrog & Charlesworth, 2003). One feature of the data from *D. pseudoobscura* that caught our attention was the existence of complete linkage disequilibrium between two single nucleotide polymorphisms (SNPs). Notably, one of them encodes an amino acid replacement. The two sites are 222 bp apart and the variants at these sites are associated with each other in all the lines surveyed so far. In this report, we present data on nucleotide sequence variation at the *exu2* locus from 31 *D. pseudoobscura* lines caught in the wild in recent years, and compare them with those from other X-linked

\* Corresponding author. 1101 East 57th St, Chicago, IL 60637, USA. Tel: +1 773 834 3964. Fax: +1 773 702 9740. e-mail: soojinyi@midway.uchicago.edu

Table 1. *Geographic origins of the species and lines used in this study*

Species	Geographic origin	Strains
<i>D. pseudoobscura</i>	American Fork Canyon, UT	afc3, afc7
	Flagstaff, AR	f17, f20
	Goldendale, WA	g98
	James Reserve, CA	jr10, jr274
	Mather, CA	m32, m48, m52
	Mt. St. Helena, CA	msh15, msh2, msh37, msh9
	Kaibab National Forest, AZ	ps1, ps2, ps3, ps4, ps5,
	Davis Mt. State Park, TX	ps7, ps8, ps9, ps10
	Zimalan-Hidalgo, Mexico	ps11
	Tempe, AZ	ps12, ps13
	Madera Canyon, AZ	ps14, ps15
	Pomona, CA	ps16, ps17, ps18
<i>Drosophila miranda</i>	British Columbia, Canada	0101-3
<i>Drosophila persimilis</i>	Mather, California	

genes. We investigate further to determine whether there is any evidence of selection to account for this unusual pattern of linkage disequilibrium.

## 2. Materials and methods

### (i) *Exu2 locus of D. pseudoobscura*

The gene *exu2* is homologous to the *exu* locus of *D. melanogaster*, which is known to function in mRNA localization as well as in spermatogenesis (Hazelrigg & Tu, 1994; Theurkauf & Hazelrigg, 1998). Luk *et al.* (1994) showed that in *D. pseudoobscura* there are two homologs of *exu*, *exu1* and *exu2*. The *exu1* locus seems to retain most of the homologous functions of *exu* of *D. melanogaster*, whereas the function of *exu2* is unclear (Luk *et al.*, 1994). The *exu2* locus is located proximally on the XL chromosome arm of *D. pseudoobscura* (Yi & Charlesworth, 2000).

### (ii) *Lines used in this study*

Recent collections of *D. pseudoobscura* flies were provided by several groups of people (see acknowledgements). The geographic origins of each line are shown in Table 1.

### (iii) *PCR and sequencing strategy*

The sequenced region encompasses a total of 1022 bps spanning the second exon of the *exu2* gene to near the end of the third exon, including 50 bps of the second intron. Figure 1 shows the structure of the gene and the sequenced region. Genomic DNA (gDNA) from single male flies from each line that had been preserved in ethanol was extracted after drying, following a modified protocol for the Puregene DNA extraction kit (Gentra). Less than 20 ng gDNA was used for each 25  $\mu$ l PCR reaction; the usual yield of genomic DNA from a single fly prepared in this way easily

exceeded 500 ng. Primers for the PCR reaction were designed based upon the published sequence (GenBank accession number L22553): forward, TTTCC-AGATTGTCAGTT; reverse, GAGTGCCATTGCCAGAGC. The sequence segments used for the analyses correspond to nucleotides 276–1297 of the GenBank accession. Additional pairs of primers were designed to provide sequences from both strands. The BIG dye-termination cycle sequencing kit (Perkin Elmer) was used for sequencing reactions. All sequences were run on an ABI 377 sequencer.

### (iv) *Sequence data analyses*

Sequences obtained from the sequencer were imported into the program Sequencher v3.0 (Gene Codes) and then proofread against the raw chromatograms. The sequences were then aligned and assembled into the whole sequenced region. A few reactions were redone using the original genomic DNA to provide complete coverage. Sequences were then edited according to the purpose of each analysis. The DNA SP program v3.97 (Rozas & Rozas, 1999) was used for most of the analyses. The population recombination parameter was estimated using Hudson's reduced likelihood method (Hudson, 2001) and Wall's full likelihood method (Wall, 2000). All the sequences obtained from this study are deposited in GenBank (accession numbers AY337547-AY337577).

## 3. Results and discussions

### (i) *Polymorphism and divergence at the exu2 locus*

We found 21 SNPs from 31 sequences of the 1022 bp segment (Fig. 1, Table 2). Among these, three were within introns and 11 were synonymous. The silent site diversity based on the number of segregating sites ( $\hat{\theta}_w$ ; Watterson, 1975) is 0.013. Nonsynonymous sites are much less polymorphic ( $\hat{\theta}_w = 0.0024$ ), as expected

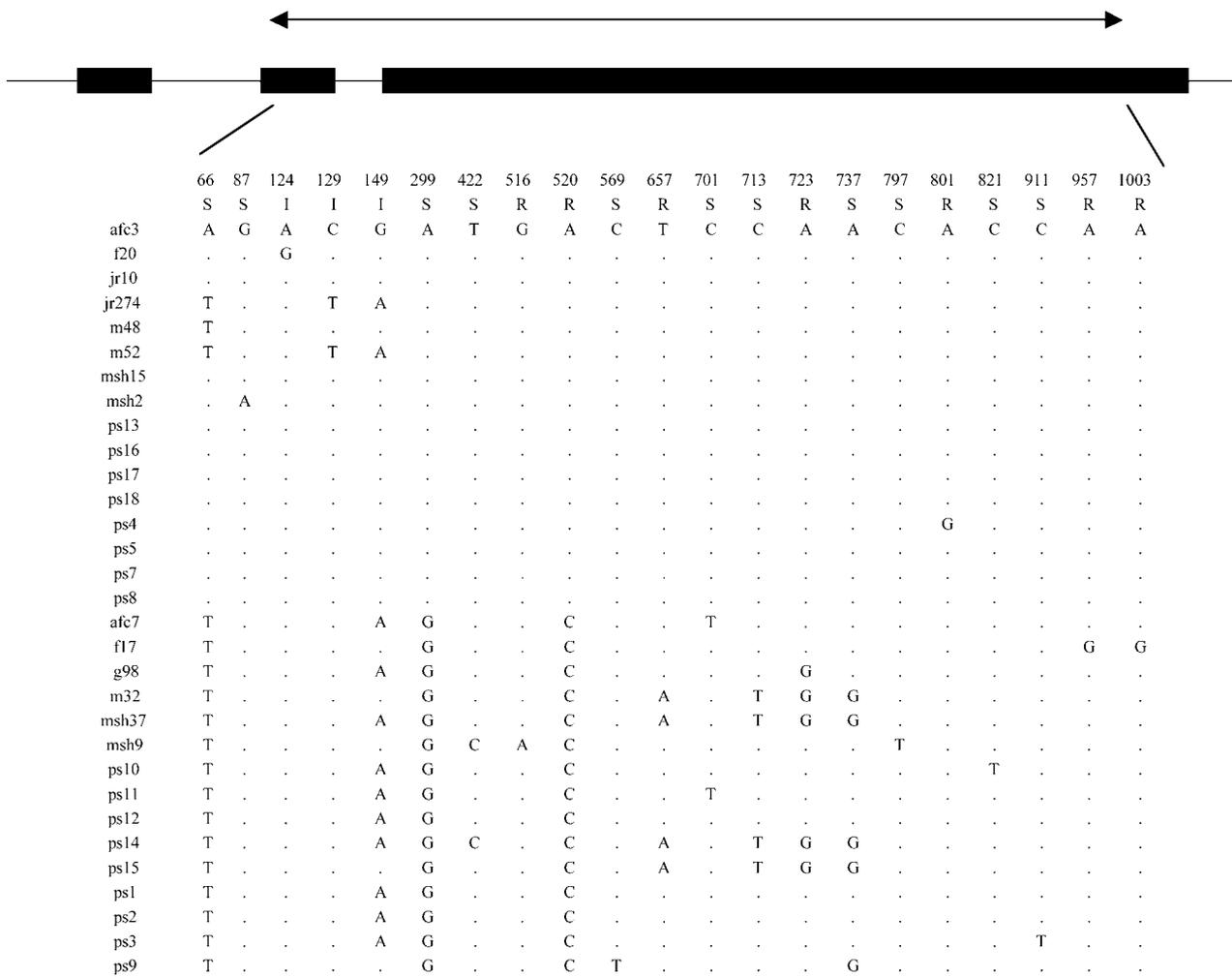


Fig. 1. SNPs at the *exu2* locus of *D. pseudoobscura*. The structure of the *exu2* locus and the sequenced region are shown at the top. Filled boxes indicate exons. The sequenced region is indicated by the double-headed arrow above the gene structure. The nature of each SNP is specified as follows: I, intron variant; R, replacement variant; S, synonymous variant.

Table 2. Measures of genetic diversity at the *exu2* locus in *D. pseudoobscura*

Genetic diversity	Replacement	Silent	Synonymous
$S^*$	7	14	11
$\hat{\theta}_W^\dagger$	0.0024	0.013	0.0096
$\hat{\pi}^{\dagger\dagger}$	0.0018	0.010	0.013
$K^\S$	0.047	0.088	0.087

\* Number of segregating sites.

† Measure of nucleotide site variability based on the number of segregating sites (Watterson, 1975).

†† Pairwise nucleotide diversity,  $\pi$  (Nei, 1987).

§ Per-site divergence from *D. miranda* with Jukes and Cantor correction for multiple hits (Jukes & Cantor, 1969).

if there are selective constraints on the protein product of the gene. Pairwise differences between sequences per site ( $\pi$ ; Nei, 1987) were slightly lower than the above values and, as a result, Tajima's  $D$  for the whole gene is negative ( $-0.79$ ), although not significantly so (Table 3). Other measures of the frequency

spectrum were also negative, suggesting an excess of rare variants compared with equilibrium neutral expectation (Table 3).

These statistics showed marginal significance when tested against the equilibrium neutral model by coalescent simulations, incorporating the inferred recombination parameter (Table 3; see below for a discussion of the population recombination parameter). However, because most of genes from *D. pseudoobscura* exhibit a negatively skewed frequency spectrum, this is likely to reflect simply the history of a recent population expansion in this species (Kovacevic & Schaeffer, 2000; Machado *et al.*, 2002).

Compared with nine other X-linked genes from *D. pseudoobscura* whose diversity levels have been investigated (Kovacevic & Schaeffer, 2000; Machado *et al.*, 2002), the silent site diversity of the *exu2* locus ranks fourth. Extrapolating from the physical and recombinational map of the XL chromosome arm, the *exu2* locus is located between the *sisA* and *per loci*, near the base of the XL (recombination map from

Table 3. Measures of the frequency spectrum and the probability of the observed value under the neutral model incorporating the population recombination parameter of  $C = 20$

Measures of the frequency spectrum		Probability	Refs
Tajima's $D$	-0.79	0.09	Tajima, 1989
Fu and Li's $D$	-1.65	0.01	Fu & Li, 1993
Fu and Li's $F$	-1.63	0.04	Fu & Li, 1993
Fu and Li's $D^*$	-1.38	0.07	Fu & Li, 1993
Fu and Li's $F^*$	-1.40	0.05	Fu & Li, 1993

Kovacevic & Schaeffer, 2000; physical map from Yi & Charlesworth, 2000). The genetic diversity of the *exu2* locus is intermediate between those of the other two loci. However, no simple global relationship between chromosomal location and genetic diversity could be inferred from these data.

We estimated the average divergence per nucleotide site using a *D. miranda* sequence as the outgroup, correcting for multiple hits (Jukes & Cantor, 1969). We arbitrarily chose the strain 0101.3 of *D. miranda* for this purpose. As the level of genetic diversity at *exu2* in *D. miranda* is very low (Yi & Charlesworth, 2000), this choice will not greatly affect the estimate. The average silent site divergence ( $K_S$ ) is 8.8%. This is rather high, twice that of the average silent site divergence between these two species based upon 12 nuclear loci from *D. miranda* (Yi *et al.*, 2003). Interestingly, the average nonsynonymous site divergence ( $K_A$ ) is also very high (4.6%), as shown in a previous analysis using a single sequence from *D. pseudoobscura* (Yi & Charlesworth, 2000). This can be explained either by a relaxed selective constraint, coupled with a high mutation rate (as evidenced by the high silent site divergence), or by faster protein evolution owing to positive selection in either lineage. When we applied the test of McDonald & Kreitman (1991) to the data, the pattern of polymorphism and divergence of nonsynonymous and silent sites was found to be incompatible with the neutral model ( $p < 0.04$  by Fisher's exact test; Table 4). This is mainly due to the reduced number of segregating non-synonymous sites within the derived haplotype (see below); if this haplotype is removed from the dataset, the test becomes non-significant. In other words, the properties of this haplotype provide some evidence for selection on the replacement sites of the *exu2* locus compared with the pattern of molecular evolution of the silent sites.

(ii) *Recombination, gene conversion and neutrality tests based on the frequency spectrum*

There is evidence of substantial recombination from the SNP data. First, we inferred a minimum of three

Table 4. McDonald–Kreitman test on the *exu2* locus, showing the fixed differences from *D. miranda* and the number of polymorphic sites within *D. pseudoobscura*

	Fixed differences	Polymorphic sites	$p$
Silent variant	20	14	
Replacement variant	32	7	
Fisher's exact test			0.038
$G$ test			0.028
$\chi^2$ test			0.030

recombination events using the method of Hudson and Kaplan (1985), between sites 149 and 299, 422 and 657, and 723 and 737. The estimated value,  $\rho$ , of the population recombination parameter  $4N_e c$  (where  $c$  is two-thirds of the female recombination frequency, and  $N_e$  is the effective population size for X-linked loci) was 22.9 for the whole sequence, using a reduced maximum likelihood method (Hudson, 2001). We also used the method of Hudson (1987), and the full-likelihood estimation method as implemented by Wall (2000). Interestingly, all these methods provided very similar values, around 20–25 per gene. If we take 25 as the value, this implies that the ratio of the recombination parameter for adjacent nucleotide sites to the neutral mutation parameter, estimated as  $\rho/\hat{\theta}_w$  (using silent sites), is 1.88. This is intermediate in value for the genes whose polymorphism patterns have been studied in *D. pseudoobscura* (Kovacevic & Schaeffer, 2000; Machado *et al.*, 2002). However, different researchers have used different methods to estimate the recombination parameter and the confidence intervals are rather large, so this result should be treated with caution.

We were concerned that gene conversion rather than reciprocal exchange might be occurring, which might inflate the recombination rate estimate. To investigate this possibility, we used a full maximum likelihood method (Wall, 2000) to jointly estimate the recombination and gene conversion parameters. The maximum likelihood estimate of the population recombination parameter (for crossover only) was 20, whereas the ratio of gene conversion to recombination was 0.5. Therefore, we used the value of 20 as the estimated recombination parameter for the whole sequence for the rest of this work.

(iii) *Pattern of linkage disequilibrium and unusual haplotype structure at the *exu2* locus*

The SNPs at the *exu2* locus of *D. pseudoobscura* present an interesting pattern of linkage disequilibrium (LD). Figure 2 shows the pattern of pairwise LD between 11 informative SNPs, where the significance of each was assessed by Fisher's exact test. There are two

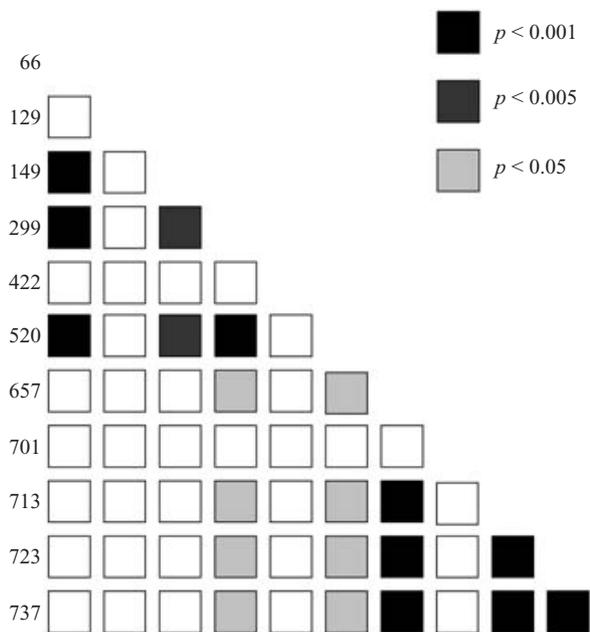


Fig. 2. Linkage disequilibrium among the 11 informative sites in our sample of 31 *exu2* sequences. Significance tested by Fisher's exact test is represented by different shades of gray and black. Ten cases ( $p < 0.001$ ) remained significant after Bonferroni correction for multiple tests.

clusters of such significant LD, one linked to SNP 66 and the other among four SNPs between sites 657 and 737. The LD between SNP 299 and SNP 520 also stands out (see below). The extent of average linkage disequilibrium over all pairwise comparisons of a number of polymorphic sites, the  $Z_{nS}$  statistic (Kelly, 1997), is equal to 0.205. This is significantly more than expected amount of LD under the neutral model when the inferred recombination parameter is incorporated into the neutral coalescent process ( $p < 0.02$ ).

In particular, two sites (SNPs 299 and 520) appeared to be in complete LD with each other in all the alleles sequenced in this study. In 15 of the 31 lines surveyed, these two sites are encoded by G and C, whereas, in the other 16 lines, they were encoded by A and A. Compared with the *D. miranda* outgroup sequence, the GC haplotype appears to be ancestral, whereas the AA haplotype is newly derived. The *exu2* locus in *Drosophila persimilis* (Yi & Charlesworth, 2000), a closer relative to *D. pseudoobscura*, has the same G and C nucleotides at the sites of interest. Therefore, it appears that the AA haplotype is specific to *D. pseudoobscura*. There is no association between the geographic origins of the individuals (Table 1). Interestingly, SNP 520 is at a replacement site, coding for either the ancestral amino acid proline (CCC) or derived histidine (CAC). Proline is a nonpolar amino acid often used in reversing the direction of a peptide backbone, whereas histidine can be positively charged and is commonly external (Li, 1997).

Although there are no other fixed differences among haplotypes, if we divide the total sample into two groups according to the haplotypes of these two sites, the two groups show rather different patterns of SNP variation. The ancestral GC haplotype is more polymorphic than the derived AA haplotype. The silent site diversities differ about twofold in their pairwise nucleotide diversities, whereas the replacement sites show a tenfold difference.

The numbers of haplotypes are 11 and 6 for the GC and AA groups, respectively. We performed the haplotype diversity test (Depaulis & Veuille, 1998), which takes into account the haplotype frequency distribution. The probability that the haplotype diversity is equal to or less than observed in the sample is around 8% by coalescent simulation when we incorporated the population recombination parameter  $C=20$ . When we use  $C=25$ , the probability is less than 6%. In other words, the major haplotype (in this case, most of the AA groups) appears marginally more frequent than expected given the estimated amount of recombination.

#### (iv) Selection at the *exu2* locus

The excessive amount of LD, particularly a complete LD involving a replacement site, and the relative lack of variability within the derived haplotype suggest a history of selection at the *exu2* locus. A possibility is the presence of balancing selection near or at the GC/AA sites, possibly to maintain the proline/histidine polymorphism. Strong linkage disequilibria and polymorphism peaks are observed near these two sites, as expected from such selection (Charlesworth *et al.*, 1997). This does not, however, account for the lower diversity in the haplotype with the derived amino acid variant.

A more likely possibility is that this AA haplotype has originated relatively recently and reached an intermediate frequency, as expected from a partial selective sweep (Hudson *et al.*, 1994; Sabeti *et al.*, 2002). Because the sibling species *D. persimilis* has the ancient haplotype and is estimated to have diverged from *D. pseudoobscura* only 500,000 years ago (Aquadro *et al.*, 1991; Wang *et al.*, 1997), this haplotype must have emerged in the recent past to have reached the current substantial frequency. This is in accordance with several characteristics of the dataset, including lower diversity in the derived haplotype and the overall deficit of haplotypes compared with the equilibrium neutral model. In particular, such a deficit is opposite to what is expected with the pattern of population growth inferred for *D. pseudoobscura* (Fu, 1997; Machado *et al.*, 2002).

We thank M. Noor, M. Kovacevic, S. Schaeffer, T. Markow, J. Coyne and S. Bryant, who eagerly collected and sent flies.

K. Thornton generously helped with implementing and running the joint estimation of recombination and gene conversion rates using the full likelihood method.

## References

- Aquadro C. F., Weaver, A. L., Schaeffer, S. W. & Anderson, W. W. (1991). Molecular evolution of inversions in *Drosophila pseudoobscura*: the amylase gene region. *Proceedings of the National Academy of Sciences of the USA* **99**, 305–309.
- Bachtrog, D. & Charlesworth, B. On the genomic location of the *exuperantia* gene in *Drosophila miranda*—the limits of *in situ* hybridization experiments. *Genetics* (in press).
- Charlesworth, B., Nordborg, N. & Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research* **70**, 155–174.
- Depaulis, F. & Veuille, M. (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular Biology and Evolution* **15**, 1788–1790.
- Fu, Y.-X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925.
- Fu, Y.-X & Li, W.-H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.
- Hazelrigg, T. I. & Tu, C. (1994). Sex-specific processing of the *Drosophila exuperantia* transcript is regulated in male germ cells by the *tra-2* gene. *Proceedings of the National Academy of Sciences of the USA* **91**, 10752–10756.
- Hudson, R. R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetical Research* **50**, 245–250.
- Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817.
- Hudson, R. R. & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.
- Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J. & Ayala, F. J. (1994). Evidence for positive selection in the superoxide dismutase (*Sod*) region in *Drosophila melanogaster*. *Genetics* **136**, 1329–1340.
- Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism* (ed. H. Munro), pp. 21–132. New York: Academic Press.
- Kelley, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics* **146**, 1197–1206.
- Kovacevic, M. & Schaeffer, S. W. (2000). Molecular population genetics of X-linked genes in *Drosophila pseudoobscura*. *Genetics* **156**, 155–172.
- Li, W.-H. (1997). *Molecular Evolution*. Sunderland, MA, USA: Sinauer Associates.
- Luk, S., Kilpatrick, K. M., Kerr, K. & Macdonald, P. M. (1994). Components acting in localization of *bicoid* mRNA are conserved among *Drosophila* species. *Genetics* **137**, 521–530.
- Machado, C. A., Kliman, R. M., Markert, J. A. & Hey, J. (2002). Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Molecular Biology and Evolution* **19**, 472–488.
- McDonald, J. H. & Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654.
- Moriyama, E. N. & Powell, J. R. (1996). Intraspecific nuclear DNA variation in *Drosophila*. *Molecular Biology and Evolution* **13**, 261–277.
- Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.
- Powell, J. R. (1997). *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. New York: Oxford University Press.
- Rozas, J. & Rozas, R. (1999). DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**, 174–175.
- Russo, C. A. M., Takezaki, N. & Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. *Molecular Biology and Evolution* **12**, 391–404.
- Sabeti, P. C., Reich, D. R., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837.
- Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.
- Theurkauf, W. E. & Hazelrigg, T. (1998). *In vivo* analyses of cytoplasmic transport and cytoskeletal organization during *Drosophila* oogenesis: characterization of a multi-step anterior localization pathway. *Development* **125**, 3655–3666.
- Wang, R. L., Wakeley, J. & Hey, J. (1997). Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics* **147**, 1091–1106.
- Wall, J. D. (2000). A comparison of estimates of the population recombination rate. *Molecular Biology and Evolution* **17**, 156–163.
- Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276.
- Yi, S. & Charlesworth, B. (2000). A selective sweep associated with a recent gene transposition in *Drosophila miranda*. *Genetics* **156**, 1753–1763.
- Yi, S., Bachtrog, D. & Charlesworth, B. A survey of chromosomal and nucleotide sequence variation in *Drosophila miranda*. *Genetics* **164**, 1369–1381.