

# Selective Sweep in the Evolution of a New Sperm-Specific Gene in *Drosophila*

Rob J. Kulathinal<sup>1</sup>, Stanley A. Sawyer<sup>2</sup>, Carlos D. Bustamante<sup>3</sup>,  
Dmitry I. Nurminsky<sup>4</sup>, Rita Ponce<sup>1</sup>, José M. Ranz<sup>1</sup> and Daniel L. Hartl<sup>1</sup>

<sup>1</sup> Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA  
02138

<sup>2</sup> Department of Mathematics, Washington University, St. Louis, MO 63130

<sup>3</sup> Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY  
14853

<sup>4</sup> Department of Anatomy and Cell Biology, Tufts University School of Medicine, Boston, MA  
02111

**Keywords:** axoneme/ dynein intermediate chain / novel gene / spermatogenesis / selective sweep

**Abbreviations:** dynein IC, dynein intermediate polypeptide chain; DCE, distal conserved  
element; PCE, proximal conserved element; TSE, testis-specific element

Corresponding author: Rob. J. Kulathinal

Department of Organismic and Evolutionary Biology

Harvard University, Cambridge, MA 02138

Phone: (617) 496-5540

Fax: (617) 496-5854

email: rkulathinal@oeb.harvard.edu

## Abstract

The *Sdic* gene cluster at the base of the X-chromosome is unique to the lineage of *Drosophila melanogaster*. The repeating unit in the cluster was formed from a duplication and fusion of the genes, *AnnX* and *Cdic*, which juxtaposed the 3' untranslated region of *AnnX* to the third intron of *Cdic*. *AnnX* encodes Annexin 10 and *Cdic* encodes a cytoplasmic dynein intermediate chain. The 3' untranslated region of *AnnX* contains two promoter elements, including a testis-specific element, and *Cdic* intron 3 contains a third promoter element; together these elements result in testis-specific transcription of *Sdic*. The *Sdic* protein features a novel amino terminus derived in part from *Cdic* intron 3 which contains motifs similar to those in axonemal dyneins. It has been demonstrated that the *Sdic* protein becomes incorporated into the tails of mature sperm. The evolution of the *Sdic* cluster required several deletions, at least one insertion, at least eleven nucleotide substitutions, and an estimated tenfold tandem duplication, all of which took place in the 1–3 million years since the divergence of *D. melanogaster* from *D. simulans*. Evidence for the ongoing evolution of *Sdic* including a recent selective sweep is found in the low levels of polymorphism across neighboring genes in the region, a large number of fixed amino acid replacements relative to fixed synonymous nucleotide substitutions, and a frequency spectrum of polymorphic nucleotides skewed toward rare variants. The analysis of polymorphism and divergence in the *Sdic* region, however, is complicated by the possible effects of background selection caused by deleterious new mutations, owing to the reduced amount of recombination in the region associated with its proximity to centromeric heterochromatin. We present the rapid evolution of this novel gene as a fascinating example of male-driven evolution incurred by recurrent selective sweeps.

## Introduction

Recent analyses of amino acid polymorphisms within species and differences between species of *Drosophila* have provided evidence that amino acid replacements are frequently driven by positive selection (Bustamante *et al.* 2002; Fay *et al.* 2002; Smith and Eyre-Walker 2002). In all three analyses, the principal conclusion rests primarily on the observation that the ratio of amino acid replacements to synonymous nucleotide substitutions between species is greater than the ratio of amino acid polymorphisms to synonymous nucleotide polymorphisms within species (McDonald and Kreitman 1991; Sawyer and Hartl 1992). From their analysis of polymorphism and divergence in *D. simulans* and *D. yakuba*, Smith and Eyre-Walker (2002) deduce that about 45% of the amino acid replacements between these species have been driven by positive selection. Their data suggest that these species have undergone one amino acid replacement every 20 years (~200 generations), or about 600,000 substitutions altogether, of which 270,000 were driven by selection. Fay *et al.* (2002) have carried out a similar analysis of data from 45 genes in *D. melanogaster* and *D. simulans* and have come to a somewhat different conclusion. Although they also noted strong evidence for positive selection in the data as a whole, they attributed most of the positive selection to 11 genes (*Acp26Aa*, *Acp29Ab*, *anon1A3*, *anon1E9*, *anon1G5*, *ci*, *est-6*, *Ref2P*, *Rel*, *tra* and *Zw*) and regarded the remaining 34 genes as evolving essentially neutrally with respect to amino acid replacements.

Bustamante *et al.* (2002) have carried out a hierarchical Bayesian analysis of polymorphism and divergence data, using a set of 34 genes in *D. melanogaster* and *D. simulans*, partly overlapping the set of genes analyzed by Fay and colleagues (Fay *et al.* 2002). We found the Bayesian approach appealing because the data are analyzed in the aggregate to estimate the average selection coefficient of each gene individually, and each estimate has an accompanying 95% credible interval, which is the Bayesian analog of the 95% confidence interval. The credible

intervals emerge naturally because the Bayesian analysis is implemented by a Markov chain Monte Carlo stochastic process whose stationary distribution coincides with the posterior distribution of the parameters conditional on the data (Gilks *et al.* 1996).

The Bayesian analysis on the *Drosophila* data yields average scaled selection coefficients,  $N_e s$ , ranging from  $-1.12$  to  $+4.12$ , where  $N_e$  is the haploid effective population size and  $s$  is the conventional selection coefficient. Among the 34 estimates, 32 are positive, again suggesting an important role for positive selection. Included among the most strongly positively selected genes, whose 95% credible interval does not overlap zero, are *Acp26Aa*, *Acp29Ab*, *anon1A3*, *anon1E9*, *ci* and *Zw*, which are found on the list of eleven rapidly evolving genes to which Fay *et al.* (2002) attribute most of the positive selection. Three genes in their list (*anon1G5*, *est-6* and *Ref2P*) are not among the most strongly positively selected genes in the Bayesian analysis, however, but are intermixed among the others. Hence, the Bayesian analysis supports that of Fay *et al.* (2002), but not completely.

The Bayesian analysis also supports that of Smith and Eyre-Walker (2002), but again not completely. Considering the 95% credible intervals across all genes, about 80% of the total span of the credible intervals is positive. This is much larger than the 45% positively selected amino acid replacements estimated in their study (Smith and Eyre-Walker 2002). However, 57% of the total span of the credible intervals has  $N_e s > 1$  and 49% has  $N_e s > 2$ ; likewise 65% of the mean values of  $N_e s$  are greater than one and 38% are greater than two. These proportions of positively selected amino acid replacements can be reconciled with those of Smith and Eyre-Walker (2002) if their method identifies amino acid replacements as positively selected provided that  $N_e s > \sim 2$ .

Details of the analyses aside, there seem to be a significant number of amino acid replacements that are driven by positive selection. As judged by the Bayesian analysis, however, the intensity of selection is relatively small. Across all genes, the average value of  $N_e s$  equals 1.5.

This intensity of selection is sufficiently weak that genetically linked neutral polymorphisms would hardly be affected unless the linkage is very tight (Wiehe and Stephan 1993).

Yet there is also considerable evidence for "selective sweeps" which describes positive selection of a certain magnitude affecting linked neutral variation (Maynard Smith and Haigh 1974). Its presence is revealed by nonneutral haplotype frequencies, typically as an excess of rare alleles across a region of the genome or as an excess in the frequency of a single haplotype. Although the interpretation of such observations is potentially complicated by demographic factors such as population subdivision, changes in population size, or founder effects, examples of apparent selective sweeps in *D. melanogaster* include regions containing the genes, *Sod* (Hudson *et al.* 1994), *white* (Kirby and Stephan 1995, 1996), *Suppressor of Hairless* (Depaulis *et al.* 1999) and *Fbp2* (Benassi *et al.* 1999). In *D. simulans*, they include regions containing the genes, *Pgd* (Begun and Aquadro 1994), *runt* (Labate *et al.* 1999), *Zw* and *vermilion* (Hamblin and Veuille 1999), and *ocnus* (Parsch *et al.* 2001).

In this paper, we summarize evidence for one or more selective sweeps in the region of a newly evolved gene found on the X-chromosome of *D. melanogaster*. The gene, denoted *Sdic*, encodes the intermediate chain for an axonemal dynein; it is expressed specifically in the testes and its novel protein is incorporated into the mature sperm tails (Nurminsky *et al.* 1998b). The novel gene is found only in *D. melanogaster* and not in any of its sibling species, including *D. simulans* (Nurminsky *et al.* 1998b). We first examine what is known about the origin and genetic structure of *Sdic*, examine the evidence for one or more selective sweeps, describe the results of a hierarchical Bayesian analysis of polymorphism and divergence in the *Sdic* region and briefly discuss *Sdic*'s rapid divergence in the more general context of the faster evolution of male-specific genes. The emphasis in this paper is on the evidence for selective sweeps. Further details about the origin and molecular structure of *Sdic* can be found in (Ranz *et al.* 2002).

## The origin of *Sdic*

The *Sdic* gene was discovered through an anomalous cDNA sequence recovered in a study of alternative splicing of cytoplasmic dynein intermediate-chain transcripts (Nurminsky *et al.* 1998a). Dynein intermediate chains are one component of the multisubunit dynein complex whose function in the cytoplasm is to act as a minus end-directed microtubule motor (Paschal *et al.* 1992; King *et al.* 1996). In *Drosophila*, the multiple forms of the dynein intermediate chains are created by alternative splicing of the transcript of a single-copy gene, denoted *Cdic*, located in polytene chromosome region 19A near the base of the X-chromosome (Nurminsky *et al.* 1998a).

The anomalous intermediate chain cDNA was unusual in that the apparent amino end of the coding sequence was missing two conserved amino-terminal domains necessary for interacting with proteins that help attach the dynein complex to its cytoplasmic targets. Instead, the amino-terminal end of the protein had a novel sequence resembling axonemal dynein intermediate chains (Nurminsky *et al.* 1998b). The intermediate chains of the axonemal dyneins are localized at the base of dynein complex and are thought to bind directly to the A-microtubule (Paschal *et al.* 1992). In a genomic clone containing the coding sequence for the anomalous cDNA, the region upstream from the transcription start site was a sequence closely resembling the single-copy gene, *Annexin X* (denoted *AnnX*) (Benevolenskaya *et al.* 1998), which encodes one of a large family of proteins that bind to phospholipids in a calcium-dependent manner and appears to have a wide variety of functions (Barton *et al.* 1991; Geisow 1991). It soon became apparent that, in *D. melanogaster*, both *Cdic* and *AnnX* had been duplicated, and that the anomalous cDNA resulted from a gene fusion that is expressed specifically in the testes and that encodes a putative axonemal dynein intermediate chain that becomes incorporated into the axoneme of the tail of the mature sperm (Nurminsky *et al.* 1998b).

In the genome of *D. simulans* and other sibling species of *D. melanogaster*, the orthologs of *Cdic* and *AnnX* are situated in the order, Telomere ···–*AnnX*–*Cdic*–··· Centromere, and transcription of each gene takes place from right to left. In the origin of *Sdic*, it is clear that there was a duplication of the region including *AnnX* and *Cdic*, leading to the structure, Telomere ···–*AnnX*–*Cdic*–*AnnX*–*Cdic*–··· Centromere. A series of deletions fused the middle two genes in such a way that intron 3 of the *Cdic* gene became juxtaposed with the 3' untranslated region of the *AnnX* gene, which may be represented as Telomere ···–*AnnX*–[*Cdic*–*AnnX*]–*Cdic*–··· Centromere (where again transcription takes place from right to left and the square brackets represent the gene fusion). This [*Cdic*–*AnnX*] fusion was the nascent novel *Sdic* gene, which after additional evolutionary refinement, became tandemly duplicated approximately tenfold (Benevolenskaya *et al.* 1998), yielding its present situation in the genome as Telomere ···–*AnnX*–[*Sdic*]<sup>~10</sup>–*Cdic*–··· Centromere (Nurminsky *et al.* 1998b).

### **The molecular structure of *Sdic***

The reconstituted portion of the *Sdic* repeating unit (in terms of novel promoter and 5' coding regions) is illustrated in Figure 1, in which the gene is oriented so that transcription takes place from left to right. This means that the centromere of the chromosome is far to the left and the telomere of the chromosome is much farther to the right. In each region of the gene, the numbers of nucleotides are indicated. This appears to be the structure of the *Sdic* gene nearest the 5' end of the cluster (nearest to *Cdic*) but there is some variation in sequence and structure from one repeating unit to the next (J. M. Ranz and R. Ponce, unpublished data).

The promoter region of *Sdic* is formed from a fusion between the exon for the 3' untranslated region of *AnnX* and intron 3 of *Cdic*. The new promoter shares two similar domains, the distal conserved element (DCE) and the proximal conserved element (PCE), as defined

within the wildtype promoter of *Cdic* (Nurminsky *et al.* 1998a). The similarity appears to be fortuitous, since neither the *Sdic* DCE nor the *Sdic* PCE are derived from the *Cdic* promoter. Indeed, the *Sdic* DCE derived from the *AnnX* 3' UTR matches the *Cdic* promoter DCE in 25 out of 34 base pairs (bp). The *Sdic* PCE is derived from *Cdic* intron 3 but matches the *Cdic* promoter PCE in 16/20 bp. Both PCE's include the conserved Initiator (*Inr*) sequence motifs (Purnell *et al.* 1994; Arkhipova 1995). Another important component of the *Sdic* promoter is the testis-specific element or TSE. This sequence matches the TSE of the testis-specific *betaTub85D* promoter in 21/27 bp. Yet the *Sdic* TSE appears to derive from the 3' UTR of *AnnX*, in which there is a sequence that matches in 22/27 bp. The *Sdic* promoter is sufficient to drive the testis-specific transcription of a construct encoding the *Sdic* protein fused to a green fluorescent protein reporter (Nurminsky *et al.* 1998b).

Although the *Sdic* protein includes the carboxyl end of *Cdic*, it is missing 84 amino acids from the amino-terminal end of *Cdic*. Instead, the *Sdic* amino-terminus consists of a novel exon encoding 93 amino acids that derives largely from *Cdic* intron 3. The *Sdic* amino end includes motifs that are similar to those at the amino end of axonemal dyneins (Nurminsky *et al.* 1998b).

As diagrammed in Figure 1, transcription of *Sdic* begins in the PCE. Translation begins 140 nucleotides downstream with an initiation codon that encodes the novel amino end of the *Sdic* protein. An insertion of 10 base pairs creates a novel splice site, which serves as a donor site for splicing with the wildtype 3' splice acceptor of *Cdic* exon 4. The variable exons (v1–v3) present in *Cdic* between exons 4 and 5 (Nurminsky *et al.* 1998a) are not present in *Sdic* mRNA; exon v1 is removed by RNA splicing, and exons v2 and v3 have been deleted from the *Sdic* genomic sequence. The alternatively spliced exon 5 (which includes exon v4) is spliced in *Sdic* in the longer mode, as found in *Cdic*. The structure and splicing patterns of *Cdic* and *Sdic* are similar for exons 5, 6, and 7, although there are some additional differences near the carboxyl end of the protein.

## Reduced polymorphism in the region of *Sdic*

The current molecular structure of *Sdic* suggests that in the course of the evolution of this multigene family there was an initial duplication of the region including *AnnX* and *Cdic*, at least three deletions resulting in the *AnnX–Cdic* gene fusion, two more insertions or deletions including one that created a novel splice junction, 11 nucleotide substitutions including reversal of a chain-terminating codon, and an estimated tenfold tandem reiteration of the newly fashioned *Sdic* gene (Nurminsky *et al.* 1998a). All of these mutations and gene fixations have occurred in a relatively short time after the divergence of *D. melanogaster* and *D. simulans*, and evolutionary refinement may still be taking place.

Recent adaptive evolution of *Sdic* might be detectable as a selective sweep, which in principle could be detected as a reduction in the level of genetic polymorphisms in the *Sdic* region and a frequency distribution of genetic variation skewed toward rare alleles. A reduced level of polymorphism in the *Sdic* region was noted in the original report (Nurminsky *et al.* 1998b). In particular, the nucleotide sequences of 1200 bp of *Sdic* and 985 bp of *Cdic* from each of nine strains of geographically diverse origin yielded estimates of nucleotide polymorphism ( $\theta$ ) of  $1.23\text{E-}3 \pm 0.83\text{E-}3$  and  $0.78\text{E-}3 \pm 0.66\text{E-}3$ , respectively, and estimates of nucleotide diversity ( $\pi$ ) of  $0.89\text{E-}3 \pm 0.73\text{E-}3$  and  $0.45\text{E-}3 \pm 0.50\text{E-}3$ , respectively. These are among the lowest estimates of nucleotide variation found in nuclear genes of diverse geographic isolates of *Drosophila* (Moriyama and Powell 1996) and are consistent with a relatively recent selective sweep in the *Sdic* region.

## The issue of background selection

Charlesworth and Charlesworth (1999) were quick to point out, correctly, that while a showing of reduced polymorphism is necessary to infer a selective sweep, it is not sufficient. They argued that a reduced level of polymorphism in a region of low recombination, such as at the base of the X-chromosome, is also consistent with background selection due to deleterious mutations. Background selection results from the fact that each new deleterious mutation that

occurs dooms some genetically linked region of chromosome to eventual extinction. The lower the rate of recombination, the larger the region of chromosome that is affected. The population effect of any new deleterious mutation is thus to reduce by one the number of chromosomes that the affected region of the genome can contribute to remote future generations. If there is absolute linkage, then the whole chromosome is affected; if there is recombination, then a smaller region flanking the mutation is affected. In either case, a sufficient density of harmful mutations will reduce the number of surviving lineages to such an extent that the degree of polymorphism will be smaller than expected, given the actual population size, and the tighter the linkage the greater the disparity.

Nurminsky and colleagues' (1999) rejoinder was based on the amount of codon usage bias in the region. In *Drosophila*, highly expressed genes tend to have a biased pattern of codon usage (Shields *et al.* 1988), which apparently results from weak selection that favors more rapid or more accurate translation (Akashi 1993, 1995). Background selection in a region of relatively tight linkage would, owing to the reduction in effective population size, be expected to result in a diminution in codon usage bias in genes across the region. Although the data available at the time showed an extremely sharp increase in codon usage bias as the gene locations proceeded outward from the centromeric heterochromatin of the X-chromosome, the complete genomic sequence of *D. melanogaster* (Adams *et al.* 2000) reveals a less dramatic pattern. Figure 2 shows the codon usage bias of 201 genes at the base of the X-chromosome, oriented with the centromere off to the right, taken from data compiled by Hey and Kliman (2002). Codon usage bias is scaled according to the effective number of codons, ENC (Wright 1990), a scale in which a smaller effective number of codons corresponds to a greater bias in codon usage. There is gradual, statistically significant ( $P < 0.01$ ) decrease in codon usage bias as the gene positions become closer to the centromeric heterochromatin (i.e. towards cytological band 20). This pattern is consistent with an

increase in background selection closer to the centromeric heterochromatin. However, the level of codon usage bias in the *Sdic* region (19A) is not markedly different from that of the *Zw* region (18D). These observations suggest that background selection does have some effect in the *Sdic* region, but not likely a sufficiently strong effect to reduce the level of polymorphism to that observed for *Sdic* and *Cdic*.

### **Further evidence for a selective sweep**

But of course, a general argument based on codon usage bias is indirect and uncertain. A more rigorous analysis was carried out by Nurminsky *et al.* (2001), who studied the level of polymorphism of ten genes at the base of the X-chromosome in a worldwide sample of 15 isofemale lines of *D. melanogaster* and 7 isofemale lines of *D. simulans*. The data from *D. simulans* served for comparison and showed a linear decrease in the level of polymorphism as a function of a gene's proximity to the centromeric heterochromatin. The data from *D. melanogaster* revealed a similar trend, but included a statistically significant "dip" in the level of polymorphism in the *Sdic* region. This pattern is entirely consistent with a selective sweep at or close to the *Sdic* locus.

A recent selective sweep was also implied by the frequency spectrum of polymorphisms (Nurminsky *et al.* 2001). In *D. melanogaster*, the frequency spectrum across the base of the X-chromosome was skewed toward rare variants, considering either synonymous polymorphisms only (Wilcoxon signed-rank test  $P = 0.04$ ) or for synonymous and nonsynonymous polymorphisms combined (Wilcoxon signed-rank test  $P = 0.01$ ). The corresponding  $P$ -values for the data from *D. simulans* were 0.44 and 0.28, respectively.

More evidence of a selective sweep can be gathered by comparing the *Sdic* locus to its progenitor sequence, *Cdic*. Between these two genes' aligned coding regions, Nurminsky *et al.*

(1998b) found six replacement changes but only two synonymous changes. This higher than average nonsynonymous to synonymous ratio of substitutions suggests that positive Darwinian selection has played a role in the evolution of *Sdic* although a decrease in selective constraints, particularly after a gene duplication event (Ohno 1970), can also explain this pattern. Further, a surprisingly complex pattern of deletions in the 3' exon has been recently found among *Sdic* copies and in relation to *Cdic* (J.M. Ranz and R. Ponce, unpublished data).

### **Bayesian analysis of polymorphism and divergence in the *Sdic* region**

Results of a hierarchical Bayesian analysis of polymorphism and divergence of genes across the *Sdic* region is shown in Figure 3, where an estimate of  $N_e s$  for each gene and its 95% credible interval is indicated (Bustamante *et al.* 2002). *Sdic* is not included, since the gene cluster does not exist in *D. simulans*.

To relate the data in Figure 3 to the full analysis of 43 genes in Bustamante *et al.* (2002), note that the value of  $N_e s$  for *Zw* ranks second highest among the full set of 43 genes, and the values of  $N_e s$  for eight of the nine genes in Figure 3 rank in the top 60% of the genes in the full set. Hence, although only two of the genes in Figure 3 (*Zw* and *runt*) have significant values of  $N_e s$  by the criterion that their 95% credible intervals do not overlap zero, the generally large values of  $N_e s$ , averaging 1.73, seem to reflect the apparent action of positive selection across the region. What is not so clear is the extent to which the apparent level of selection indicated is selection at each locus individually as opposed to the effects of genetic linkage with one or two strongly selected genes in the region. Nevertheless, the analysis of polymorphism and divergence reinforces the conclusion reached from the frequency spectrum of synonymous polymorphisms that there has been at least one positively selected sweep in this region. The genetic linkage across the region complicates the interpretation, because the Bayesian analysis assumes that the

genes are independent, but on the other hand, any reduction in  $N_e$  in the region that results from background selection implies that the values of  $s$  are actually greater than the estimated values of  $N_e s$  would imply. In any case, the results in Figure 3 suggest to us that there may well have been more than one selective sweep in the region, perhaps in more than one gene, since a selective sweep can impel to fixation only those amino acid replacements with which the favorable mutation happens to be linked.

One interesting sidelight of the data has to do with the effective population size of *D. simulans* relative to *D. melanogaster*. Analysis of synonymous substitutions suggests that  $N_e$  for *D. simulans* is larger than that for *D. melanogaster* (Akashi 1996). Maximum likelihood estimates of the ratio of the effective population sizes in the *Sdic* region yield an estimated ratio of 1.486 (95% confidence interval 0.723–2.249) for all *D. melanogaster* populations taken together. However, when the analysis is restricted to *D. melanogaster* lines from Zimbabwe, the estimated ratio of effective sizes is 0.994 (95% confidence interval 0.581–1.407). These are obviously not significantly different, but they do serve to support the inference that worldwide *D. simulans* has an effective population size about 50% greater than that of *D. melanogaster* and additionally, that there is more genetic variation in African, particularly Zimbabwe, populations of *D. melanogaster* than there is in North American populations (Begun and Aquadro 1993).

The higher effective population size found among Zimbabwe lines compared to other global *D. melanogaster* lines, as suggested by the Bayesian analysis, supports this population's distinct, isolated, and presumably stable nature (Wu *et al.* 1995; Hollocher *et al.* 1997; Andolfatto and Przeworski 2001). More importantly, it presents us with another opportunity to test the selective sweep hypothesis in the *Sdic* region. Once a selective sweep occurs, it takes approximately  $N_e$  generations (depending on the strength of selection) for the population to return back to equilibrium (Perlitz and Stephan 1997). Since the Zimbabwe population has a higher  $N_e$

relative to other more recently diverged *D. melanogaster* populations, deviations from neutrality would be easier to detect. Table 1 shows that although values of the Tajima's *D* statistic are not significantly different from zero, all ten loci (located in the *Sdic* region) with samples solely from Zimbabwe populations of *D. melanogaster*, produce negative Tajima's *D* values. This observed skew in frequency towards rare variants was not found in *D. simulans* nor with other *D. melanogaster* populations and together, with the previously reported pattern of low polymorphism, suggests that a recent sweep(s) has taken place in African *D. melanogaster* populations in or around the *Sdic* locus.

### **Rapid evolution of male-specific genes**

The accumulated set of observations which include the rapid formation of the *Sdic* gene cluster, the low level of *Sdic* nucleotide diversity and the frequency distribution of rare *Sdic* variants, as well as the observed patterns of variation in genes neighboring the *Sdic* locus – the suppressed levels of genetic variation, the lower than expected decrease in codon bias, the consistently negative Tajima's *D* values in African populations, and the slightly positive selection intensities estimated from the data – together provide strong evidence that a selective sweep, or a series of recurrent sweeps, has taken place at the *Sdic* locus. This inference also fits into the wider context of the faster evolution of male-specific traits, particularly those involved in fertility (Wu 1993; Civetta and Singh 1995). As a protein expressed specifically in the sperm tail, *Sdic* may be positively selected under a variety of sexual selection mechanisms. For example, sperm competition (Clark *et al.* 1995; Civetta and Clark 2000), sexual conflict (Rice 1996) and sexual coevolution (Swanson and Vacquier 2002) have been demonstrated in *Drosophila* and may be a potent force in the molecular evolution of sperm-specific genes.

Recently, a number of male-specific genes have been identified that, like *Sdic*, possess a high ratio of replacement to silent fixed substitutions (Singh and Kulathinal 2000). This pattern of high amino acid divergence in male-specific proteins appears to be a general phenomenon among a wide variety of taxa but is especially evident in *Drosophila* (Civetta and Singh 1999; Singh and Kulathinal 2000). For example, many of the most rapidly evolving genes, as revealed by two-dimensional electrophoresis of *Drosophila* proteins, are male-specific (Coulthart and Singh 1988; Thomas and Singh 1992; Civetta and Singh 1995). Other rapidly evolving male-specific genes or genetic systems in *Drosophila* include segregation distortion (Wu *et al.* 1988; McClean *et al.* 1994), sex ratio in *D. simulans* (Atlan *et al.* 1997), *Mst40* (Russell and Kaiser 1994) and *Stellate* (Palumbo *et al.* 1994; Bozzetti *et al.* 1995; Mckee and Satter 1996).

The rapid evolution of the *Sdic* gene cluster also represents a remarkable example of gene evolution *in statu nascendi*. Interestingly, of the few known examples of incipient gene/domain formation among closely related species, many appear to be associated with male reproductive traits, particularly spermatogenesis. For example, the *jingwei* gene in the *D. teissieri* /*D. yakuba* lineage (Wang *et al.* 2000) has recently evolved and is expressed specifically in the testis. Similarly, *Odysseus* – although not a newly evolved gene - contains rapidly evolving homeodomains involved in sperm function that have been recently fixed solely in *D. mauritiana*, a sibling species in the *D. melanogaster* complex (Ting *et al.* 1998, 2000). Hence, it appears that while other genetic systems may possess a higher level of selective constraints, spermatogenesis may be more prone to allow for the co-opting of novel genes and function. Consequently, the greater potential for selective sweeps may be an intrinsic property of genes expressed in the male reproductive system. Therefore, the observed presence of selective sweep(s) in the *Sdic* region may be the result of the combination of *Sdic*'s location in a tightly linked region of the genome together with its potential fitness consequences on male fertility.

## **Acknowledgements**

This work was supported by NIH grants GM60035 (DH) and GM61549 (DN), NSF grant DMS-0107420 (SAS) and by fellowships from the Natural Sciences and Engineering Council of Canada (RJK), the Marshall-Sherfield fund (CDB), the National Research Council of Spain (JMR), the Foundation for Science and Technology of Portugal (ARP).

## References

- Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, *et al.* The genome sequence of *Drosophila melanogaster*. *Science* 2000; 287: 2185-95.
- Akashi H. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 1993; 136: 927-35.
- Akashi H. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* 1995; 139: 1067-76.
- Akashi H. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: Reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* 1996; 144: 1297-307.
- Andolfatto P, Przeworski M. Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. *Genetics* 2001; 158: 657-65.
- Arkhipova I. Promoter elements in *Drosophila melanogaster* revealed by sequence analysis. *Genetics* 1995; 139: 1359-69.
- Atlan A, Mercot H, Landre C, Montchampmoreau C. The sex-ratio trait in *Drosophila simulans*: geographical distribution of distortion and resistance. *Evolution* 1997; 51: 1886-95.
- Barton GJ, Newman RH, Freemont PS, Crumpton MJ. Amino acid sequence analysis of the annexin super-gene family of proteins. *Eur J Biochem* 1991; 198: 749-60.
- Begun DJ, Aquadro CF. African and North American populations of *Drosophila melanogaster* are very different. *Nature* 1993; 365: 548-50.
- Begun DJ, Aquadro CF. Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of *Drosophila*: selection and geographic differentiation. *Genetics* 1994; 136: 155-71.
- Benassi V, Depaulis F, Meghlaoui GK, Veuille M. Partial sweeping of variation at the *Fbp2* locus in a West African population of *Drosophila melanogaster*. *Mol Biol Evol* 1999; 16: 347-53.

Benevolenskaya E, Nurminsky D, Gvozdev V. Structure of the *Drosophila melanogaster* *annexin X* gene. DNA Cell Biol 1998; 14: 349-57.

Bozzetti MP, Massari S, Finelli P, Meggio F, Pinna LA, Boldyreff B, *et al.* The *Ste* locus, a component of the parasitic cry-ste system of *Drosophila melanogaster*, encodes a protein that forms crystals in primary spermatocytes and mimics properties of the beta subunit of casein kinase. Proc Natl Acad Sci USA 1995; 92: 6067-71.

Bustamante CR, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. The cost of inbreeding in *Arabidopsis*. Nature 2002; 46: 531-4.

Charlesworth B, Charlesworth D. How was the *Sdic* gene fixed? Nature 1999; 400: 519-20.

Civetta A, Singh RS. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. J Mol Evol 1995; 41: 1085-95.

Civetta A, Singh RS. Broad-sense sexual selection, sex gene pool evolution, and speciation. Genome 1999; 42: 1033-41.

Civetta A, Clark A. Correlated effects of sperm competition and postmating female mortality. Proc Natl Acad Sci U S A 2000; 97: 13162-5.

Clark AG, Aguade M, Prout T, Harshman LG, Langley CH. Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. Genetics 1995; 139: 189-201.

Coulthart MB, Singh RS. High level of divergence of male-reproductive-tract proteins between *Drosophila melanogaster* and its sibling species, *D. simulans*. Mol Biol Evol 1988; 5: 182-91.

Depaulis F, Brazier L, Veuille M. Selective sweep at the *Drosophila melanogaster* *Suppressor of Hairless* locus and its association with the *In(2L)t* inversion polymorphism. Genetics 1999; 152: 1017-24.

Fay JC, Wyckoff GJ, Wu CI. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* 2002; 415: 1024-6.

Geisow MJ. Annexins: forms without function but not without fun. *Trends Biotechnol* 1991; 9: 180-1.

Hamblin MT, Veuille M. Population structure among African and derived populations of *Drosophila simulans*: evidence for ancient subdivision and recent admixture. *Genetics* 1999; 153: 305-17.

Hey J, Kliman RM. Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics* 2002; 160: 595-608.

Hollocher H, Ting C-T, Wu M-L, Wu C-I. Incipient speciation by sexual isolation in *Drosophila melanogaster*: extensive genetic divergence without reinforcement. *Genetics* 1997; 147: 1191-201.

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ. Evidence for positive selection in the superoxide-dismutase (Sod) region of *Drosophila melanogaster*. *Genetics* 1994; 136: 1329-40.

King SM, Barbarese E, Dillman JF, Patel-King RS, Carson JH, Pfister KK. Brain cytoplasmic and flagellar outer arm dyneins share a highly conserved Mr 8,000 light chain. *J Biol Chem* 1996; 271: 19358-66.

Kirby DA, Stephan W. Haplotype test reveals departure from neutrality in a segment of the *white* gene of *Drosophila melanogaster*. *Genetics* 1995; 141: 1483-90.

Kirby DA, Stephan W. Multi-locus selection and the structure of variation at the *white* gene of *Drosophila melanogaster*. *Genetics* 1996; 144: 635-45.

Labate JA, Biermann CH, Eanes WF. Nucleotide variation at the *runt* locus in *Drosophila melanogaster* and *Drosophila simulans*. *Mol Biol Evol* 1999; 16: 724-31.

Maynard Smith J, Haigh J. The hitch-hiking effect of a favorable gene. *Genet Res* 1974; 23: 23-5.

McClellan JR, Merrill CJ, Powers PA, Ganetzky B. Functional identification of the *segregation distorter* locus of *Drosophila melanogaster* by germline transformation. *Genetics* 1994; 137: 201-9.

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

McKee BD, Satter MT. Structure of the Y chromosomal *Su(Ste)* locus in *Drosophila melanogaster* and evidence for localized recombination among repeats. *Genetics* 1996; 142: 149-61.

Moriyama EN, Powell JR. Intraspecific nuclear DNA variation in *Drosophila*. *Mol Biol Evol* 1996; 13: 261-77.

Nurminsky DI, Benevolenskaya EV, Nurminskaya MV, Shevelyov YY, Hartl DL, Gvozdev VA. Cytoplasmic dynein intermediate chain isoforms with different targeting properties created by tissue-specific alternative splicing. *Mol Cell Biol* 1998a; 18: 6816-25.

Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* 1998b; 396: 572-5.

Nurminsky DI, Hartl DL. How was the *Sdic* gene fixed? *Nature* 1999; 400: 520.

Nurminsky DI, Aguiar DD, Bustamante CD, Hartl DL. Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*. *Science* 2001; 291: 128-30.

Ohno S. Evolution by Gene Duplication. Berlin: Springer-Verlag, 1970.

Palumbo G, Bonaccorsi S, Robbins LG, Pimpinelli S. Genetic analysis of stellate elements of *Drosophila melanogaster*. *Genetics* 1994; 138: 1181-97.

Parsch J, Meiklejohn CD, Hartl DL. Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. *Genetics* 2001; 159: 647-57.

Paschal BM, Mikami A, Pfister KK, Vallee RB. Homology of the 74-kD cytoplasmic dynein subunit with a flagellar dynein polypeptide suggests an intracellular targeting function. *J Cell Biol* 1992; 118: 1133-43.

Perlitz M, Stephan W. The mean and variance of the number of segregating sites since the last hitchhiking event. *J Math Biol* 1997; 36: 1-23.

Purnell B, Emanuel P, Gilmour D. TFIID sequence recognition of the initiator and sequences farther downstream in *Drosophila* class II genes. *Genes Dev* 1994; 8: 830-42.

Ranz JM, Ponce AR, Hartl DL, Nurminsky DI. *Genetica* (In press) 2002.

Rice WR. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature* 1996; 381: 232-4.

Russell SRH, Kaiser K. A *Drosophila melanogaster* chromosome-2L repeat is expressed in the male germ line. *Chromosoma* 1994; 103: 63-72.

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

Shields DC, Sharp PM, Higgins DG, Wright F. "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. *Mol Biol Evol* 1988; 5: 704-16.

Singh RS, Kulathinal RJ. Sex gene pool evolution and speciation: a new paradigm. *Genes Genet Syst* 2000; 75: 119-30.

Smith NGC, Eyre-Walker A. Adaptive protein evolution in *Drosophila*. *Nature* 2002; 415: 1022-4.

Swanson W, Vacquier V. The rapid evolution of reproductive proteins. *Nature Reviews Genetics* 2002; 3: 137-44.

Thomas S, Singh RS. A comprehensive study of genetic variation in natural population of *Drosophila melanogaster*. VII. Varying rates of genic divergence as revealed by two-dimensional electrophoresis. *Mol Biol Evol* 1992; 9: 507-25.

Ting C-T, Tsaur S-C, Wu M-L, Wu C-I, Davis A. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* 1998; 282: 1501-4.

Ting C-T, Tsaur S-C, Wu C-I. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. *Proc Natl Acad Sci U S A* 2000; 97: 5313-6.

Wang W, Zhang JM, Alvarez C, Llopart A, Long M. The origin of the *Jingwei* gene and the complex modular structure of its parental gene, *yellow emperor*, in *Drosophila melanogaster*. *Mol Biol Evol* 2000; 17: 1294-301.

Wiehe THE, Stephan S. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol Biol Evol* 1993; 10: 842-54.

Wright F. The 'effective number of codons' used in a gene. *Gene* 1990; 87: 23-9.

Wu C-I, Lyttle TW, Wu M-L, Lin G-F. Association between a satellite DNA sequence and the *Responder of Segregation Distorter* in *D. melanogaster*. *Cell* 1988; 54: 179-89.

Wu C-I, Hollocher H, Begun D, Aquadro C, Xu Y, Wu M-L. Sexual isolation in *Drosophila melanogaster*: a possible case of incipient speciation. *Proc Natl Acad Sci U S A* 1995; 92: 2519-23.

Wu C-I, Davis, AW. Evolution of postmating reproductive isolation: the composite nature of Haldane's rule and its genetic bases. *Am Nat* 1993; 142: 187-212.

**Table 1**  
**Tajima's  $D$  on loci near the centromeric region of the X-chromosome**

Locus	Band	Length (bp)	<i>D. melanogaster</i>			<i>D. simulans</i>
			non-Zimbabwe	Zimbabwe	All populations	Global populations
<i>Zw</i>	18D13	537	0.38 (6)	<b>-0.22 (6)</b>	-0.33 (12)	-0.33 (7)
<i>Bap</i>	18E4-18E5	471	0.18 (8)	<b>-0.53 (5)</b>	-0.75 (13)	-1.21 (7)
<i>AnnX</i>	19C1	606	-1.29 (9)	<b>-0.93 (6)</b>	-0.83 (15)	0.85 (7)
<i>Sdic</i>	19C1	1146	-0.58 (9)	<b>-0.47 (5)</b>	0.43 (14)	N.A.
<i>Cdic</i>	19C1	777	-0.84 (9)	<b>-0.19 (6)</b>	0.08 (15)	0.16 (7)
<i>Pp4</i>	19C	339	0.33 (8)	<b>-0.93 (6)</b>	-0.34 (14)	0.69 (7)
<i>run</i>	19E2	585	0.00 (7)	<b>-0.79 (6)</b>	-0.72 (13)	-0.73 (7)
<i>shakB</i>	19E3	705	-1.24 (7)	<b>-0.93 (6)</b>	-1.65 (13)	-1.01 (7)
<i>tty</i>	19F4	594	-0.36 (9)	<b>-0.79 (6)</b>	-0.70 (15)	-0.35 (7)
<i>slgA</i>	19F6-20A1	495	-0.58 (9)	<b>-0.83 (6)</b>	-0.33 (15)	0.71 (6)
<i>su(f)</i>	20E	966	N.A.	<b>N.A.</b>	N.A.	0.21 (7)
<i>Variance</i>			0.38	<b>0.08</b>	0.32	0.54
<i>Sign Test</i>			P<0.75	<b>P&lt;0.01**</b>	P<0.11	P=1

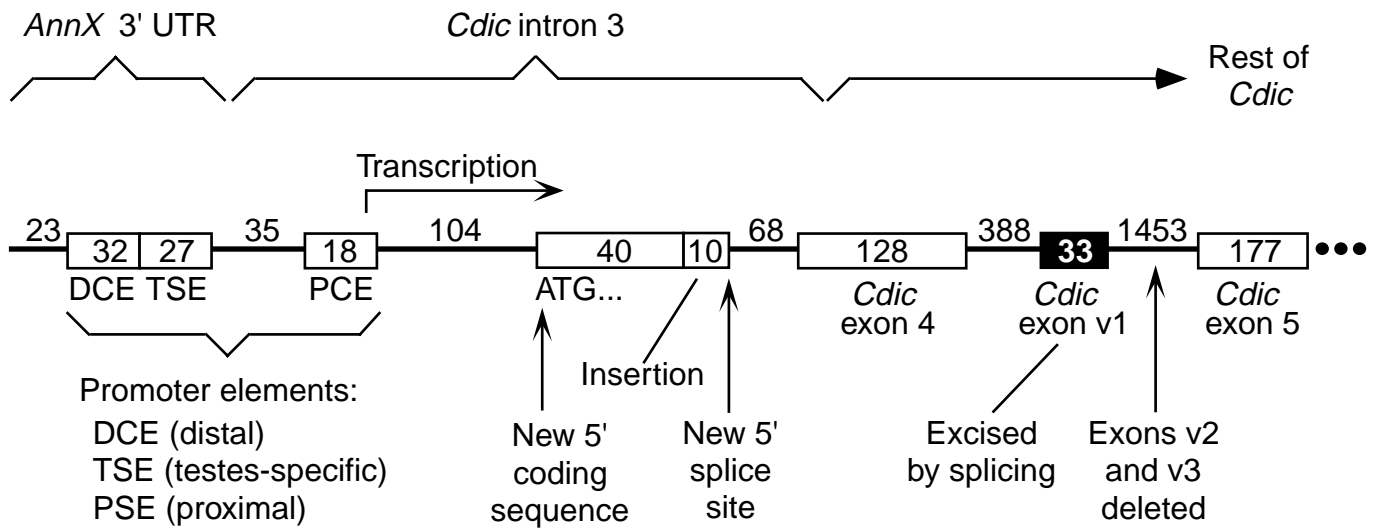
N.A., Data not available. Number of sequences in parentheses. Tajima's  $D$  is not significantly different from zero in all cases.

## Figure Legends

**Figure 1.** Molecular structure of a portion of one of the analyzed *Sdic* repeats showing the three key promoter elements created by the fusion of the 3' UTR of *AnnX* and intron 3 of *Cdic*. Part of the novel amino end of the *Sdic* protein derives from *Cdic* intron 3 sequences. Sequence length is indicated in base pairs.

**Figure 2.** Codon usage bias of genes in the base of the euchromatin of the X-chromosome, oriented with the centromeric heterochromatin off toward the right. The measure of codon bias is the effective number of codons (Wright 1990), which scales inversely with codon usage bias. Hence larger values of the ENC are associated with less biased codon usage. Based on data from Hey and Kliman (2002).

**Figure 3.** Estimates of the scaled average selection coefficient ( $N_{es}$ ) of amino acid replacements, and the 95% credible intervals, for a sample of genes across the base of the X-chromosome in *D. melanogaster* and *D. simulans*, based on the hierarchical Bayesian analysis outlined in Bustamante *et al.* (2002).



Kulathinal et al. Figure 1

