



Origin and evolution of a new gene expressed in the *Drosophila* sperm axoneme*

José María Ranz¹, Ana Rita Ponce¹, Daniel L. Hartl^{1,*} & Dmitry Nurminsky²

¹Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 33143, USA;

²Department of Anatomy and Cell Biology, Tufts University School of Medicine, Boston, MA 02111, USA; *Author for correspondence: (Phone: +1-617-496-3917; Fax: +1-617-496-5854; E-mail: dhartl@oeb.harvard.edu)

Key words: axoneme, dynein intermediate chain, exon shuffle, gene fusion, spermatogenesis

Abstract

Sdic is a new gene that evolved recently in the lineage of *Drosophila melanogaster*. It was formed from a duplication and fusion of the gene *AnnX*, which encodes annexin X, and *Cdic*, which encodes the intermediate polypeptide chain of the cytoplasmic dynein. The fusion joins *AnnX* exon 4 with *Cdic* intron 3, which brings together three putative promoter elements for testes-specific expression of *Sdic*: the distal conserved element (DCE) and testes-specific element (TSE) are derived from *AnnX*, and the proximal conserved element (PCE) from *Cdic* intron 3. *Sdic* transcription initiates within the PCE, and translation is initiated within the sequence derived from *Cdic* intron 3, continuing through a 10 base pair insertion that creates a new splice donor site that enables the new coding sequence derived from intron 3 to be joined with the coding sequence of *Cdic* exon 4. A novel protein is created lacking 100 residues at the amino end that contain sequence motifs essential for the function of cytoplasmic dynein intermediate chains. Instead, the amino end is a hydrophobic region of 16 residues that resembles the amino end of axonemal dynein intermediate chains from other organisms. The downstream portion of *Sdic* features large deletions eliminating *Cdic* exons v2 and v3, as well as multiple frameshift deletions or insertions. The new protein becomes incorporated into the tail of the mature sperm and may function as an axonemal dynein intermediate chain. The new *Sdic* gene is present in about 10 tandem repeats between the wildtype *Cdic* and *AnnX* genes located near the base of the X chromosome. The implications of these findings are discussed relative to the origin of new gene functions and the process of speciation.

Abbreviations: dynein IC – dynein intermediate polypeptide chain; DCE – distal conserved element; PCE – proximal conserved element; TSE – testes-specific element.

Introduction

The evolution of novel gene functions is thought to occur primarily by one of two mechanisms, either by duplication and divergence or by exon shuffling. Both mechanisms are known to occur. There are many examples of evolution by gene duplication in *Drosophila*. These are usually recognized by similarities between paralogous genes sometimes, but not always, found in small gene clusters, such as the maltase gene cluster (Snyder & Davidson, 1983), the chorion

protein gene cluster (Martinez-Cruzado et al., 1988), the larval cuticle protein gene cluster (Steinemann & Steinemann, 1990), the alcohol dehydrogenase and alcohol dehydrogenase-related genes (Jeffs, Holmes & Ashburner, 1994), and the alpha-esterase gene cluster (Robin et al., 1996). These examples are by no means exhaustive, and many more can be found in FlyBase (<http://flybase.bio.indiana.edu>).

The other primary mechanism for creating new gene functions is exon shuffling (Gilbert, 1978; Long, Rosenberg & Gilbert, 1995; Long, 2001). There are also examples of exon shuffling in *Drosophila*. Perhaps the most dramatic is that of *jingwei*, a newly

* The authors José María Ranz and Ana Rita Ponce contributed equally to this work.

evolved gene of unknown function in the lineage leading to *D. teissieri* and *D. yakuba* (Long & Langley, 1993; Wang et al., 2000). The chimeric *jingwei* gene was created by the insertion of part of a reverse transcript of the alcohol dehydrogenase (*Adh*) coding sequence into the third intron of a different gene, denoted *yande*. The *jingwei* coding sequence thereby includes the initial *yande* exon to which is appended the 'shuffled' and already spliced *Adh* exons.

This paper examines a novel gene that originated as a duplication and gene fusion accompanied by recruitment of new promoter elements and the formation of a new exon encoding the amino end of the polypeptide chain. These events fused exon 4 of the gene *AnnX*, which encodes an annexin protein, with intron 3 of the gene *Cdic*, which encodes the intermediate polypeptide chain for the cytoplasmic dyneins. The new gene, called *Sdic* (for sperm-specific dynein intermediate chain), is expressed primarily if not exclusively in testes, and it encodes a protein that features a re-fashioned amino end to which is appended much of the carboxyl end of what originally encoded a cytoplasmic dynein intermediate chain. The *Sdic* protein localizes to the sperm tail and may function as an axonemal dynein intermediate chain. The *Sdic* gene features some unprecedented 'fudging' of the genetic functions: exon 4 sequences in the wildtype *AnnX* gene have become a part of the *Sdic* promoter that is not transcribed, and intron 3 sequences in the wildtype *Cdic* gene that have no preexisting coding function now, through multiple mutations, including a 10-bp insertion, encode the amino end of the *Sdic* protein. Remarkably, the sperm-specific promoter element was formed by the gene fusion itself. In addition, the *Sdic* gene has become tandemly duplicated and is present in about 10 copies in the base of the X chromosome of *D. melanogaster*. The origin and evolution of this gene and its tandem duplicates is recent, since it is not found in closely related species that diverged within the last 1–3 million years.

Origin of the chimeric *Sdic* gene

The *Sdic* gene was discovered through an anomalous cDNA recovered in a study of alternative splicing of cytoplasmic intermediate-chain dynein transcripts (Nurminsky et al., 1998a). Cytoplasmic dynein is a multisubunit complex composed of two heavy chains, three intermediate chains (ICs), several light intermediate chains (LICs), and one light chain (LC) (Paschal

et al., 1992; King et al., 1996). It acts as a minus end-directed microtubule motor, participating in a number of events, including slow axonal transport (Dillman, Dabney & Pfister, 1996), anterograde organelle movement (Schroer, Steuer & Sheetz, 1989; Cortesy-Theulaz, Pauloin & Rfeffer, 1992; Aniento et al., 1993), mitosis (Vaisberg, Koonce & McIntosh, 1993), and nuclear migration (Xiang, Beckwith & Morris, 1994).

Although the heavy chain comprises the catalytic dynein subunit and by itself can bring about an ATP-dependent force on the microtubules (Mazumdar et al., 1996), the presence of other subunits is apparently required for dynein function *in vivo*. The roles of these so-called accessory subunits still remain unclear. At least two accessory subunits of cytoplasmic dynein, the intermediate-chain and light-chain subunits, share significant homology with the corresponding subunits of the axonemal dyneins, suggesting similarity of their functions in these two complexes. In the case of the ICs, this suggestion lead to the hypothesis that the IC subunits are responsible for linking the cytoplasmic dynein to the intracellular targets (Paschal et al., 1992), because in the axonemal dyneins the ICs have been localized in the base of complex, binding directly to the A-microtubule. The ICs of cytoplasmic dyneins possess a large carboxyl-terminal portion containing a series of WD-40 repeats, which is present in the axonemal ICs as well (Wilkerson et al., 1995). The amino-terminal part shows no homology with the axonemal ICs.

In *Drosophila*, the multiple forms of the dynein ICs are created by alternative splicing of the transcript of a single-copy gene, denoted *Cdic*, located in region 19E near the base of the X chromosome (Nurminsky et al., 1998a). The anomalous IC cDNA was unusual in that the apparent amino end of the coding sequence was missing both the coiled-coil domain and the serine-rich domain necessary for the interaction between dynein intermediate chains and the p150/Glued protein that participates in attaching the dynein complex to its cytoplasmic targets. Instead, the amino-terminal end of the protein had a novel sequence that was very hydrophobic (Nurminsky et al., 1998a).

The region of the genome encoding this anomalous cDNA was also in the base of the X chromosome near *Cdic*. Immediately upstream of the transcription start site was a sequence derived from the gene for Annexin X, denoted *AnnX*. The annexins are a large family of proteins that bind to phospholipids in a calcium-dependent manner. They appear to have a

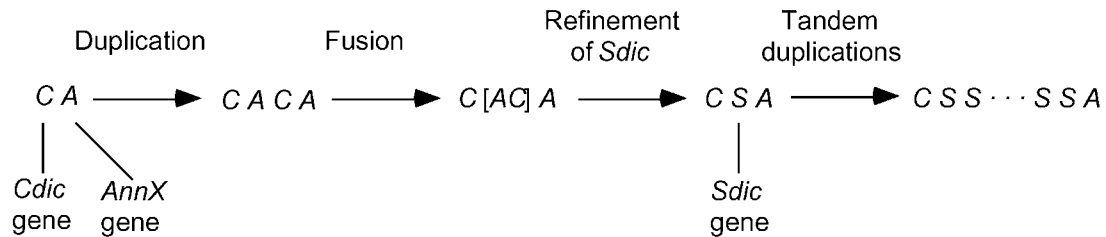
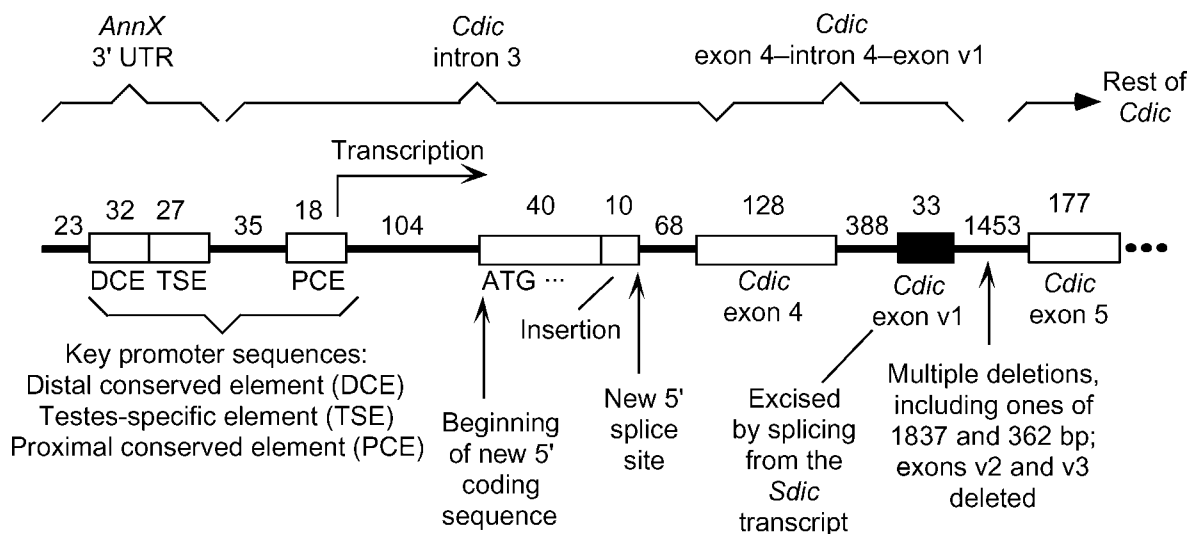
A. Origin of *Sdic*B. Structure of *Sdic*

Figure 1. Origin and structure of *Sdic*. (A) A duplication of the region containing *Cdic* (C) and *AnnX* (A) took place, followed (or accompanied) by several large deletions that created a gene fusion between *AnnX* and *Cdic* [AC], which was the progenitor of the gene *Sdic* (S). In present-day populations of *D. melanogaster*, *Sdic* is present in about 10 nonidentical tandem repeats. (B) The *Sdic* promoter elements were formed by the fusion of *AnnX* exon 4 (DCE and TSE) and *Cdic* intron 3 (PCE). The amino end of the new *Sdic* protein derives from sequences in *Cdic* intron 3 and a 10-bp insertion that creates a new splice donor site that is spliced to the normal 3' splice acceptor site of *Cdic* intron 3. The exons are designated as in Nurminsky et al. (1998a).

wide variety of functions and have been implicated in cytoskeletal interactions, phospholipase inhibition, intracellular signalling, anticoagulation, membrane fusion, and apoptosis (Barton et al., 1991; Geisow, 1991).

A curious observation was that, in the region transcribed to yield the anomalous cDNA, the sequence similar to *AnnX* was upstream of the 5' end of the *Cdic*-like cDNA, whereas in the genome the position of these genes is reversed. The inferred explanation for this orientation is shown in Figure 1(A). Prior to the origin of *Sdic*, the *Cdic* (C) and *AnnX* (A) genes were both single-copy genes oriented C A, as

shown at the left. A duplication of this region led to the configuration C A C A, and a series of at least three large deletions fused the 3' end of *AnnX* with the 5' end of *Cdic*, producing the configuration C [AC] A, where the square brackets denote the gene fusion. There is at present no way of knowing when these deletions occurred, nor the order in which they occurred. One possibility is that they all took place simultaneously with the formation of the duplication, another possibility is that they occurred sequentially after the duplication was already in place, and there are also other scenarios. Whatever the process, the [AC] fusion created the framework of *Sdic*



Figure 2. Sequence of the *Sdic* promoter and the wildtype sequences of *AnnX* exon 4 and *Cdic* intron 3. Single underline, distal conserved element (DCE); wavy underline, testes-specific element (TSE); double underline, proximal conserved element (PCE); dashed underline ATG, initial codon of *Sdic* protein.

(*S* in Figure 1(A)), which has also undergone about a tenfold amplification yielding the present configuration *C S S . . . S S A* (Nurminsky et al., 1998b). Evidence that the gene is newly evolved is that it is present in all wildtype strains of *D. melanogaster* so far examined, but neither the novel gene nor any evidence of a tandem repeat is found in wildtype strains of *D. simulans* nor in any other member of the *D. melanogaster* species subgroup (Nurminsky et al., 1998b).

Molecular structure of the *Sdic* gene

The molecular structure of an *Sdic* repeating unit is shown in Figure 1(B). Within the *Sdic* cluster, however, there may be variation in sequence or structure from one unit to the next. The numbers above each box or line segment give the number of nucleotides present in the region.

As shown at the left, the promoter region of *Sdic* is formed from a fusion between the exon for the 3' untranslated region of *AnnX*, joined to intron 3 of *Cdic*. Upstream of *Cdic* intron 3, one noncoding exon and two coding exons with open reading frames of 155 and 147 nucleotides are deleted, which accounts for the missing amino end of *Cdic* in the *Sdic* protein encoded in the cDNA originally discovered.

As indicated in Figure 1(B), the promoter region consists of three discrete elements, called the distal conserved element (DCE), the proximal conserved element (PCE), and a testes-specific element (TSE). As we shall see in a moment, the DCE and the PCE are somewhat similar to corresponding promoter elements in the wildtype *Cdic* gene, although the *Sdic* elements have a completely different origin.

Transcription of *Sdic* begins in the PCE (Figure 1(B)), and 140 nucleotides downstream translation begins with an initiation codon that encodes the novel amino end of the *Sdic* protein. An insertion of 10 bp creates a novel splice site, which serves as a

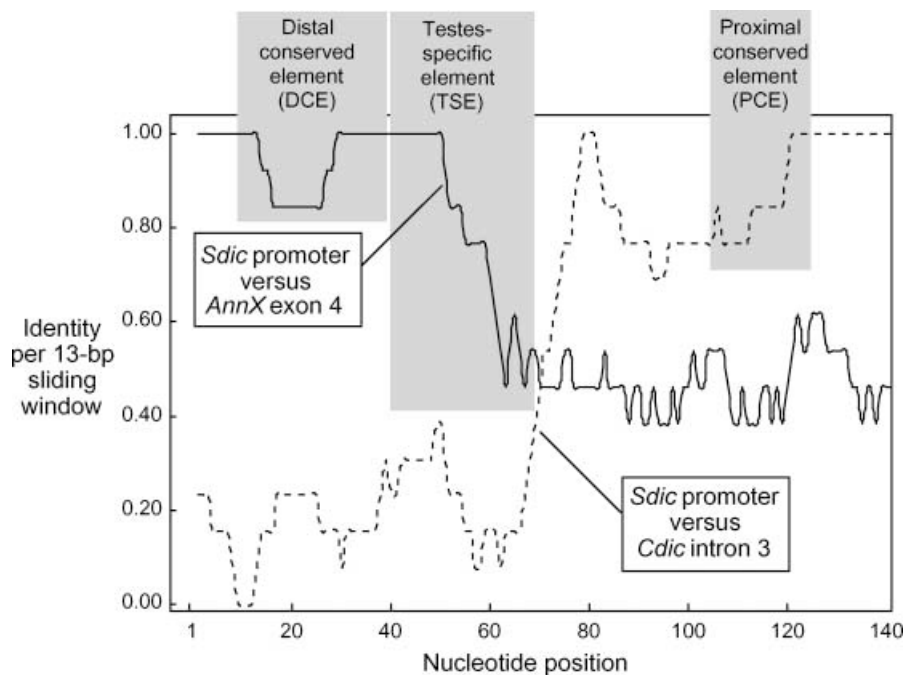


Figure 3. Percent identity of nucleotide sequences in a 13-bp sliding window across the *Sdic* promoter, compared with the sequence of wildtype *AnnX* exon 4 (solid line) and wildtype *Cdic* intron 3 (dashed line). The shaded rectangles indicate the positions of the *Sdic* promoter elements.

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 are deleted from the *Sdic* genomic DNA. The alternatively spliced exon 5 is spliced in *Sdic* in the longer mode. 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 terminus of the proteins.

The promoter region

Two points need to be emphasized in the present analysis of the *Sdic* sequence. The first is that the comparisons are based on *Sdic*, *Cdic* and *AnnX* as they exist in *D. melanogaster* today. Experiments to infer the ancestral sequences are in progress but at present incomplete. The second point is that the canonical *Sdic* sequence, illustrated in Figure 1(B), is based on a single cloned copy of the repeat, which we know to be functional because of cDNA sequencing and confirmation by germline transformation

experiments with an *Sdic*::GFP fusion protein (Nurminsky et al., 1998b).

Bearing in mind these caveats, the overall structure of the *Sdic* promoter is diagrammed in Figure 2. As noted, the promoter consists of three elements. The distal conserved element (DCE, single underline) and the proximal conserved element (PCE, double underline) are similar in size and sequence with promoter elements present in *Cdic* (Nurminsky et al., 1998b). Their spacing is somewhat different: they are separated by 62 bp in *Sdic* but by only 29 bp in *Cdic*. The third promoter element is a testes-specific element (TSE, wavy underline).

In Figure 2, the *AnnX* sequence is the 3' untranslated region of exon 4, and that of *Cdic* is a region of 229 bp from about the middle of the 375-bp intron 3. The ATG used for the translational start of *Sdic* is dash underlined at the lower right. This ATG is included in *Cdic* intron 3 and is spliced out of the *Cdic* transcript during RNA processing. The key point is that the wildtype *Cdic* intron 3 does not contain a full set of promoter elements, hence *Cdic* transcription begins far upstream of the region shown in Figure 2. The new *Sdic* promoter is unique, formed by the fusion between *AnnX* exon 4 and *Cdic* intron 3.

A. Distal conserved element (DCE)

```

AnnX      CCCGA GGATG TGGCA GCTGG TCCGC CCAAT ATTTT ATTCG TGTT- -AATA
          |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||
Sdic      CCCGA AGATG TGGCA GCTGG TCCGC CCAAT ATTTT ATTCG TGTT- -AATA
          |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||
Cdic DCE      A TCTAT GTAAA CCAGC CGCGC CCATT ATCTC CTTCG TGTC CAATA

```

```

AnnX      GCTTT GATCG TAGTG GCCTT TTAGG AAAAT CGCTT TTAAT GTCGT CTGCGC
          |||||  |||||  |||||  |||||  ||  ||  |||||  ||  ||  ||  ||  ||  ||
Sdic      GCTTT GATCG TAGTG GCCTT TGGGG GAAAT TCTGT TGGAT -TCCC CATCT-
          |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||
Cdic DCE      ACTGT GCGAA GTAAC GGCCG TGGGC

```

B. Testes-specific element (TSE)

```

AnnX      TTGGA TCGTA GTGTG CCTTT TAGGA AA-ATC GCTTT TAATG TCGTC TGCG
          |||||  |||||  |||||  |||||  ||  ||  ||  ||  ||  ||  ||  ||  ||
Sdic      TTTGA TCGTA GTGTG CCTTT GGGGG AA-ATC CTGTT GGAT- TCCCC ATCT
          |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||
βTub85D TSE  GGAAG TCGTA GTA-G CCTAT TTGTG AACATT CGGTG TAGTA ATCCA AGCC
          |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||

```

C. Proximal conserved element (PCE)

```

Cdic      ACA-A GCTTA ACAT- AAAAA AATAT C-AG- TAGTC AAAA- TTGGA ATCCT
          |  |||||  ||  ||  |||||  |||||  ||  ||  |||||  |||||  |||||  |||||
Sdic      ATG-A GTTTA ACGTC AAAAC AATAT C-AG- TAGTC AAAA- TTGGA ATCCT
          |  |||||  ||  ||  |||||  |||||  |||||  |||||  |||||  |||||  |||||
Cdic PCE      GCGCA TGTTT CATCT CTAGT TAGTC AAAAC ATTAA AAGTC

```

Figure 4. Promoter elements of *Sdic* compared with the wildtype sequences of present-day *AnnX* and *Cdic*, and with similar promoter elements from wildtype *Cdic* (DCE and PCE) and the TSE of the gene encoding a testes-specific β -tubulin (β Tub85D).

In *Sdic*, transcription begins within the PCE (double underline).

Examination of the similarity between the sequences in Figure 2, makes it clear that the *Sdic* DCE and most if not all of the TSE derive from *AnnX* exon 4. The PCE clearly derives from *Cdic* intron 3, as do regions further upstream. The exact breakpoint of the fusion is difficult to specify, but it appears to be somewhere between the AAATT near the end of the TSE and the GATTC 7 bp downstream. The situation is illustrated graphically in Figure 3, which shows the proportion of identical nucleotides in a 13-bp sliding window between *Sdic* and *AnnX* exon 4 (solid line) and between *Sdic* and *Cdic* intron 3. The positions of the promoter elements are indicated by the shaded rectangles. There is clearly a region immedi-

ately downstream of the TSE where the *Sdic* promoter becomes more similar to *Cdic* intron 3 than to *AnnX* exon 4.

Distal conserved element (DCE)

Detailed comparisons of the *Sdic* promoter elements with analogous elements from other genes are shown in Figure 4. The *Sdic* DCE (single underline in Figure 4(A)) matches the sequence of *AnnX* exon 4 in all 33 of 33 bp. It matches the DCE of the wildtype *Cdic* gene in 25/34 bp (74%). Under the binomial distribution, assuming equal proportions of the base pairs, the probability of 25 or more matches in a sequence of 34 bp is 3.9×10^{-9} ; requiring matches only of pyrimidines with pyrimidines and purines with

<i>Cdic</i> genomic	atg ggc tta gta ctg att aag ttt taa cga tca atg tat	
<i>Cdic</i> 'protein'	m g l v l i k f end r s m y	
<i>Sdic</i> genomic	ATG GGC TTA GTA CTG ATT AAG TTT <u>TTA</u> CGA TCA <u>ACG</u> TAT	
<i>Sdic</i> cDNA	ATG GGC TTA GTA CTG ATT AAG TTT TTA CGA TCA ACG TAT	
<i>Sdic</i> protein	M G L V L I K F L R S T Y	
<i>Cdic</i> genomic	t	gtactattgctattgccatttaaccgattcctaac
<i>Cdic</i> 'protein'		c t i a i a i end p i p n
<i>Sdic</i> genomic	<u>TCT ACT TTG AG</u>	gtactat <u>aa</u> ctattgccatttaacc <u>a</u> ttcctaac
<i>Sdic</i> cDNA	TCT ACT TTG AG	
<i>Sdic</i> protein	S T L S	
<i>Cdic</i> genomic	taatacc <u>g</u> atttcttatgtgcacccccaccag	C GGC GGA AAG AAA
<i>Cdic</i> 'protein'	end y q f l m c t p h q	G G K K
<i>Sdic</i> genomic	taataccgatttcttatgtgcacccccaccag	C GGC GGA AAG AAA
<i>Sdic</i> cDNA		C GGC GGA AAG AAA
<i>Sdic</i> protein		G G K K
<i>Cdic</i> genomic	CAG CCC CTC AAC CTA AGC GTC TAC AAT GTG CAG GCT ACG	
<i>Cdic</i> 'protein'	Q P L N L S V Y N V Q A T	
<i>Sdic</i> genomic	CAG <u>CCT</u> CTC AAC CTA AGC GTC TAC AAT GTG CAG GCT ACG	
<i>Sdic</i> cDNA	CAG CCT CTC AAC CTA AGC GTC TAC AAT GTG CAG GCT ACG	
<i>Sdic</i> protein	Q P L N L S V Y N V Q A T	

Figure 5. Sequences of wildtype genomic DNA for *Cdic* and that of genomic DNA and cDNA of *Sdic*. Lowercase letters indicate intronic sequences. Also shown are the amino acid sequences. That of *Cdic* intron 3 is a 'virtual protein' (lowercase letters) conceptually translated in the same reading frame in which *Sdic* is translated. Double underlines in *Sdic* genomic DNA denote differences between the sequences.

purines, the binomial probability is 0.004. Hence the likelihood of such a long sequence matching the *Cdic* DCE so well is quite remote.

Testis-specific element (TSE)

Similarly, the *Sdic* TSE matches a canonical TSE to a greater extent than expected by chance. Figure 4(B) shows a comparison between the *Sdic* TSE (wavy underline) and the TSE of the gene *betaTub85D* for the testes-specific beta-2 tubulin (Michiels et al., 1989). The *Sdic* TSE matches the sequence of *AnnX* exon 4 in 22/27 bp (81%) and that of *betaTub85D* in 21/27 bp. In this case the binomial probability of a random match of 21 or more is 1.3×10^{-8} , and considering only pyrimidines and purines it is 0.003. Interesting, of the 5 bp in which the *Sdic* TSE differs from the sequence of *AnnX* exon 4, three match the *betaTub85D* TSE and two do not. However, only one of these is in the 14-bp region required for testes-specific expression (Michiels et al., 1989).

Proximal conserved element (PCE)

The PCE in *Sdic* (double underline in Figure 4(C)) is derived from *Cdic* intron 3 and matches it in 17/18 bp (94%). The match with the wildtype *Cdic* PCE is 16/20 bp. The binomial probability of an equal or greater number of exact matches is $3.9 \times$

10^{-7} , and for pyrimidine and purine matches is 0.006. Once again, these values seem unexpectedly small.

Fashioning a protein-coding region from an intron

The *Sdic* cDNA encodes a ~60 kDa protein derived largely from the C-terminal portion of the dynein intermediate chain molecule (Nurminsky et al., 1998a). This region is responsible for the interaction of the intermediate chain with other components of the dynein complex (Ma et al., 1999), and its sequence is strongly conserved between cytoplasmic and axonemal dynein intermediate chains. It seemed unlikely that the *Sdic* product could function as a subunit of cytoplasmic dynein, since it is missing the first two protein-coding exons of *Cdic*. These exons code for an N-terminal coiled-coil domain as well as a serine-rich, dynactin-binding domain, both of which are essential for the function of cytoplasmic dynein intermediate chains (Steffen, 1997).

Analysis of the *Sdic* cDNA revealed a novel 5' exon coding for the N-terminal domain of 16 amino acids (Figure 5). This new exon is derived from sequences present in intron 3 of the *Cdic* gene (a rare example of noncoding sequences evolving into coding sequences, with implications for the origin of

```

Cdic  MDRKAELERK KAKLAALREE KDRRRREKEI KDMEEAAGRI GGGAGIDKDO

Sdic                                     MGLVLI KFLRSTYSTL
Cdic  RKDLDEMLSS LGVAPVSEVL SSLSSVNSMT SDNSNTQTPD ASLQATVNGQ

Sdic  SGGKKQPLNL SVYNVQATNI PPKETLVYTK QTQTTSTGGG  NGD
Cdic  SGGKKQPLNL SVYNVQATNI PPKETLVYTK QTQTTSTGGG  NGDGYMEDWW

Sdic                                     VLA FDAQGDDEES SLQNLGNGFT
Cdic  RPRKAHATDY YDEYNLNPLG EWEDEFVLA FDAQGDDEES SLQNLGNGFT

Sdic  SKLPPGYLTH GLPTVKDVAP AITPLEIKKE TEVKKEVNEL  SEEQKQMIL
Cdic  SKLPPGYLTH GLPTVKDVAP AITPLEIKKE TEVKKEVNEL  SEEQKQMIL

Sdic  SENFQRFVVR AGRVIERALS ENVDIYTDYI GGDSEEDAND  ERSCHARLSLN
Cdic  SENFQRFVVR AGRVIERALS ENVDIYTDYI GGDSEEDAND  ERSCHARLSLN

Sdic  RVFYDERWSK NRCITSMDWS THFPELVVGS YHNNEESPNE  PDGVVMVWNT
Cdic  RVFYDERWSK NRCITSMDWS THFPELVVGS YHNNEESPNE  PDGVVMVWNT

Sdic  KFKKSTPEDV FHCQSAVMST CFAKFNPNLI LGGTYSQIV  LWDNRVQKRT
Cdic  KFKKSTPEDV FHCQSAVMST CFAKFNPNLI LGGTYSQIV  LWDNRVQKRT

Sdic  PIQRTPLSAA AHTHPVYCLQ MVGTQNAHNV ISISSDGKLC  SWSLDMLSQP
Cdic  PIQRTPLSAA AHTHPVYCLQ MVGTQNAHNV ISISSDGKLC  SWSLDMLSQP

Sdic  QDTLELQQRQ SKAIAITSMA FPANEINSLV MGSEEDGYVYS  ASRHGLRSGV
Cdic  QDTLELQQRQ SKAIAITSMA FPANEINSLV MGSEEDGYVYS  ASRHGLRSGV

Sdic  NEVYERHLGP ITGISTHYNQ LSPDFGHLFL TSSIDWTIKL  WSLKDTKPLY
Cdic  NEVYERHLGP ITGISTHYNQ LSPDFGHLFL TSSIDWTIKL  WSLKDTKPLY

Sdic  SFE                                     QYIAWSPVRRQPPGPKTQPRH
Cdic  SFEDNSDYVM DVAWSPVHPA LFAAVDGSGR LDLWNLNQDT  EVPTASIVVA

Sdic  GAPALRRDSW TPSGL--CIG DEAGKLYVYD VAENLAQPSR  DEWSRFNTHL
Cdic  GAPALNRVSW TPSGLHVCIG DEAGKLYVYD VAENLAQPSR  DEWSRFNTHL

Sdic  SEIKMIQGDEI
Cdic  SEIKMNSDEV

```

Figure 6. Comparison between *Sdic* and *Cdic* proteins. Single underline, genomic sequence present in *Cdic* but not in *Sdic* [note that internal exons (exons v2 and v3) totaling 23 codons are deleted from *Sdic*]; double underline, new 5' exon in *Sdic* derived from intron 3 of *Cdic*; dotted underline, alternatively spliced exons present in *Cdic* but not in *Sdic*; wavy underline, multiple frameshift deletions allow only partial and inexact alignment in this region.

exons). This hydrophobic N-terminal sequence shows some similarity with the N-terminal amino acid sequences of axonemal dynein intermediate chains from other organisms, with 44% amino acid identity and 62% amino acid similarity across the first 16 residues (Nurminsky et al., 1998b).

In Figure 5, protein-coding nucleotide sequences are shown in uppercase letters and intronic sequences in lowercase. Differences between the genomic sequences of *Sdic* and *Cdic* are denoted by double underlines in the *Sdic* sequence. The lowercase 'protein' sequences are 'virtual proteins' that would be derived by translation across an intron. As indicated

by the double underlines, the sequence encoding the N-terminal region of *Sdic* differs from that of wildtype *Cdic* intron 3 at two positions, one of which eliminates a putative termination codon. The 5' coding sequence of *Sdic* also includes a 10-bp insertion (wavy underline), which creates a new splice donor site at its 3' end that attacks the normal acceptor splice site at the downstream end of *Cdic* intron 3, splicing the exons in the correct reading frame to allow translation to proceed. There are also four nucleotide differences in the part of *Cdic* intron 3 that remains an intron in *Sdic*, and one nucleotide difference in the initial part of the fused *Cdic* exon 4.

The *Sdic* protein may function as an axonemal dynein intermediate chain

The protein comparisons suggested that, if *Sdic* is functional, its product might well be an axonemal dynein IC. In *Drosophila*, the axoneme constitutes a major part of the sperm tail. Amplification of reverse-transcribed cDNA indicated that *Sdic* transcription is not only abundant in testes, but is also testes-specific (Nurminsky et al., 1998b). To follow the fate of the *Sdic* protein in more detail, we created an *Sdic::GFP* reporter cassette coding for the *Sdic* polypeptide fused to GFP (green fluorescent protein) at the carboxyl end, under the conserved of the *Sdic* promoter. Transgenic flies carrying this cassette exhibited green fluorescent *Sdic::GFP* fusion protein only in the testes and seminal vesicles. *Sdic::GFP* fusion protein is not present in the stem cells or proliferating spermatocytes, but first appears in the growing spermatocytes. The fluorescent label is especially abundant in bundles of maturing spermatocytes, and it is very strong all along the full length of the tails of mature sperm. The cytological preparations supporting these patterns of fluorescence are not shown because they require color, but they are dramatic, completely unambiguous, and highly reproducible (Nurminsky et al., 1998b).

Comparison of the *Sdic* and *Cdic* proteins

Earlier we mentioned that the ICs of axonemal dyneins possess a large carboxyl-terminal portion that is similar to those of cytoplasmic dyneins (Wilkerson et al., 1995). Figure 6, which compares the complete sequences of the *Cdic* and *Sdic* proteins, shows the extensive similarity across a large part of the carboxyl region. In Figure 6, the double underline denoted the novel amino end of *Sdic*, the single underlines denote residues in *Cdic* derived from exons that are present in *Cdic* cDNA but not in *Sdic* cDNA, and the dotted underline denotes residues present in some alternatively spliced versions of *Cdic* but not in *Sdic*. The wavy line toward the carboxyl end denotes a region in which multiple frameshift deletions prevent only partial an inexact alignment of the genomic sequences.

In spite of regions in which there are major differences between *Sdic* and *Cdic* due to the novel amino end of *Sdic* and to insertions or deletions, a total of 474 residues can be aligned without ambiguity. Among these only five are different, and all are concentrated near the extreme carboxyl end of the protein (in the

last two rows in Figure 6, which encompass *Sdic* codons 459–517).

Silent and replacement substitutions

As there appear to be few nonsynonymous (replacement) nucleotide substitutions that distinguish *Sdic* from *Cdic* (Figure 6), so too there are few synonymous substitutions. Among the 474 codons in *Cdic* and *Sdic* that can be unambiguously aligned, there are a total of five nonsynonymous substitutions and seven synonymous substitutions. The nonsynonymous substitutions are in *Sdic* codons 464, 466, 512, 514, and 517 (Figure 6). The synonymous substitutions are in *Sdic* codons 23, 343, 428, 467, 469, 472, and 515. It is odd that the majority of the synonymous substitutions (4/7) occur in the same carboxyl 10% of the coding sequence as all of the nonsynonymous substitutions.

Discussion

Sdic is a novel gene that has only recently been created and is apparently still in the process of evolving, ‘caught in the act’ as it were (Nurminsky et al., 2001). The reason why newly evolved genes warrant detailed analysis is that they give us rare, first-hand examples of the early stages of gene creation and evolution, from which we may hope to generalize the findings to other genes whose origin and functional elaboration cannot be observed directly. Most genetic functions, such as those involved in basic cellular and metabolic processes, are ancient. They came into existence so long ago that it is difficult to imagine the mechanisms of their origin and functional divergence. Yet we may hope that insights into such early evolutionary processes may be gained by studying the handful of recently evolved gene functions that happen to have been identified. At the very least we shall learn how new genes are created and evolve in contemporary organisms.

The fact that *Sdic* is male-specific in its function fits into a wider picture. One hypothesis to account for Haldane’s rule (‘when hybrid sterility occurs in only one sex, it is likely to be the heterogametic sex’) is that genes governing reproductive functions evolve faster in the heterogametic sex (Wu & Davis, 1993; Wu, Johnson & Palopoli, 1996; Laurie, 1997). In support of this hypothesis, among sequences for homologous

genes in closely related species present in GenBank, classified as to function of their protein product, there is a significantly high ratio of nonsynonymous to synonymous substitutions in the coding regions for 'sex-related' genes (defined as those affecting mating behavior, fertility, spermatogenesis, or sex determination), which is more pronounced between closely related species than more distantly related species, and which is due to an elevated proportion of nonsynonymous changes (Civetta & Singh, 1995). Consistent with this finding, two-dimensional electrophoresis of *Drosophila* proteins has revealed a surprising number of examples, anonymous as to function, that differ between closely related species, many of which are male specific (Thomas & Singh, 1992; Civetta & Singh, 1995; Coulthart & Singh, 1988). There are also some known *Drosophila* genes that fall into this category:

- The *segregation distorter* system, a spermatogenesis-specific meiotic drive found in *D. melanogaster* but not in *D. simulans*, which involves the interaction of two distinct genetic elements (Wu et al., 1988; McClean et al., 1994).
- A system of male X-chromosomal meiotic drive found in *D. simulans* but not in *D. melanogaster*, which also involves multiple genetic elements (Atlan et al., 1997).
- *Mst40*, a repeated locus coding for a male-specific transcript of unknown function found in *D. melanogaster* but not in *D. simulans* (Russell & Kaiser, 1994).
- In *D. melanogaster* only, a genetic interaction between the Y-linked *Suppressor of Stellate* [*Su(Ste)*] locus (Balakireva et al., 1992; Mckee & Satter, 1996) and the X-linked *Stellate* elements (Livak, 1990), mediated by short, double-stranded RNA (Aravin et al., 2001), in which the combination of *Stellate* elements with a deletion of *Su(Ste)* causes meiotic abnormalities in spermatogenesis, gamete-genotype dependent failure of sperm development, and deposition of protein crystals in spermatocytes (Palumbo et al., 1994; Bozzetti et al., 1995).
- The *jingwei* gene in the *teissieri/yakuba* lineage, which is expressed specifically in the testes (Wang et al., 2000).
- The homeobox gene *Odysseus*, a putative male-sterility gene in the *melanogaster/simulans* lineage, a homolog of the *C. elegans* neurogenesis gene *unc-4*, which was recruited for testes expression in *D. simulans* but not *D. melanogaster* (Ting et al., 1998; Ting, Tsauro & Wu, 2000).
- The *Sdic* repeats, the subject of this paper, which encode a novel sperm-specific dynein found only in *D. melanogaster* (Nurminsky et al., 1998b).

The *Sdic* system also suggests a new model for the long-term fate of some gene duplications which, to our knowledge, has not been observed previously, but may be quite important. It relates to the fact that *Sdic* itself is duplicated about tenfold in tandem repeats, but DNA sequencing as well as recovery of cDNAs suggests that at least some of the copies may be defective or transcriptionally silent (unpublished observations). One scenario to explain this unexpected contrast is that, in the early stages of gene evolution, when the rate of transcription may be limiting to fitness, perhaps the 'easiest' kind of favorable mutation to arise is a duplication leading to a tandem repeat or to multiple copies. Duplications are quite common, for example, of the *Adh* region (Jeffs, Holmes & Ashburner, 1994; Nurminsky et al., 1995; Begun, 1997; Luque, Marfany & González-Duarte, 1997). As time goes on, one of the duplicated copies may undergo point mutations (or other rearrangement) that increases the promoter efficiency, making the other duplicated copy (or copies) superfluous. Over additional time the superfluous duplicated copy (or copies) would be expected to undergo mutational degeneration and, given the high rate of DNA loss in *Drosophila* (Petrov, Lozovskaya & Hartl, 1996; Petrov & Hartl, 1997; Petrov & Hartl, 1998), eventually complete elimination. This suggests a more general principle that, except in special cases governed by different constraints (e.g., rDNA repeats or histone repeats), duplications may persist over long stretches of evolutionary time only if they diverge in function enough to be wholly or partially noncomplementing. Rapid acquisition and loss of duplications may help to explain why, for example, *D. virilis* and *D. melanogaster* both have a maltase gene cluster, but the origin of each cluster from the ancestral maltase is completely different (Vieira, Vieira & Hartl, 1997), and why both major phylads of the *D. virilis* species group have *Adh* duplications, but the duplications are of completely independent origin (Nurminsky et al., 1995). We recognize that it is impossible to generalize from any single example. But on the other hand, so few genes are caught in the act of evolving that every example contributes potentially important insights.

Acknowledgements

This work was supported by NIH grants 60035 (DH) and GM61549 (DN), and by fellowships from the National Research Council of Spain to JMR and the Foundation for Science and Technology of Portugal to ARP.

References

- Aniento, F., N. Emans, G. Griffiths & J. Gruenberg, 1993. Cytoplasmic dynein-dependent vesicular transport from early to late endosomes. *J. Cell Biol.* 123: 1373–1387.
- Aravin, A.A., N.M. Naumova, A.V. Tulin, V.V. Vagin, Y.M. Rozovsky & V.A. Gvozdev, 2001. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr. Biol.* 11: 1017–1027.
- Atlan, A., H. Mercot, C. Landre & C. Montchampmoreau, 1997. The sex-ratio trait in *Drosophila simulans*: geographical distribution of distortion and resistance. *Evolution* 51: 1886–1895.
- Balakireva, M.D., Y.Y. Shevelyov, D.I. Nurminsky, K.J. Livak & V.A. Gvozdev, 1992. Structural organization and diversification of Y-linked sequences comprising *Su(Ste)* genes in *Drosophila melanogaster*. *Nucl. Acids Res.* 20: 3731–3736.
- Barton, G.J., R.H. Newman, P.S. Freemont & M.J. Crumpton, 1991. Amino acid sequence analysis of the annexin super-gene family of proteins. *Eur. J. Biochem.* 198: 749–760.
- Begun, D.J., 1997. Origin and evolution of a new gene descended from *alcohol dehydrogenase* in *Drosophila*. *Genetics* 145: 375–382.
- Bozzetti, M.P., S. Massari, P. Finelli, F. Meggio, L.A. Pinna, B. Boldyreff, O.G. Issinger, G. Palumbo, C. Ciriaco, S. Bonaccorsi & S. Pimpinelli, 1995. 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* 92: 6067–6071.
- Civetta, A. & R.S. Singh, 1995. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *J. Mol. Evol.* 41: 1085–1095.
- Corthesy-Theulaz, I., A. Pauloin & S.R. Rfeffer, 1992. Cytoplasmic dynein participates in the centrosomal localization of the Golgi complex. *J. Cell Biol.* 118: 1333–1345.
- Coulthart, M.B. & R.S. Singh, 1988. High level of divergence of male-reproductive-tract proteins between *Drosophila melanogaster* and its sibling species, *D. simulans*. *Mol. Biol. Evol.* 5: 182–191.
- Dillman, J.F., L.P. Dabney & K.K. Pfister, 1996. Cytoplasmic dynein is associated with slow axonal transport. *Proc. Natl. Acad. Sci. USA* 93: 141–144.
- Geisow, M.J., 1991. Annexins: forms without function but not without fun. *Trends Biotechnol.* 9: 180–181.
- Gilbert, W., 1978. Why genes in pieces? *Nature* 271: 501.
- Jeffs, P.S., E.C. Holmes & M. Ashburner, 1994. The molecular evolution of the *alcohol dehydrogenase* and *alcohol dehydrogenase-related* genes in the *Drosophila melanogaster* species subgroup. *Mol. Biol. Evol.* 11: 287–304.
- King, S.M., E. Barbarese, J.F. Dillman, R.S. Patel-King, J.H. Carson & K.K. Pfister, 1996. Brain cytoplasmic and flagellar outer arm dyneins share a highly conserved Mr 8,000 light chain. *J. Biol. Chem.* 271: 19358–19366.
- Laurie, C.C., 1997. The weaker sex is heterogametic: 75 years of Haldane's rule. *Genetics* 147: 937–951.
- Livak, K.J., 1990. Detailed structure of the *Drosophila melanogaster* *Stellate* genes and their transcripts. *Genetics* 124: 303–316.
- Long, M., 2001. Evolution of novel genes. *Curr. Opin. Genet. Dev.* 11: 673–680.
- Long, M. & C.H. Langley, 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* 260: 91–95.
- Long, M., C. Rosenberg & W. Gilbert, 1995. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl. Acad. Sci. USA* 92.
- Luque, T., G. Marfany & R. González-Duarte, 1997. Characterization and molecular analysis of *Adh* retrosequences in species of the *Drosophila obscura* group. *Mol. Biol. Evol.* 14: 1316–1325.
- Ma, S., L. Trivinos-Lagos, R. Graf & R.L. Chisholm, 1999. Dynein intermediate chain mediated dynein–dynactin interaction is required for interphase microtubule organization and centrosome replication and separation in Dictyostelium. *J. Cell Biol.* 147: 1261–1273.
- Martinez-Cruzado, J.C., C. Swimmer, M.G. Fenerjian & F.C. Kafatos, 1988. Evolution of the autosomal chorion locus in *Drosophila*. I. General organization of the locus and sequence comparisons of genes *s15* and *s19* in evolutionary distant species. *Genetics* 199: 663–677.
- Mazumdar, M., A. Mikami, M.A. Gee & R.B. Vallee, 1996. *In vitro* motility from recombinant dynein heavy chain. *Proc. Natl. Acad. Sci. USA* 93: 6552–6556.
- McClellan, J.R., C.J. Merrill, P.A. Powers & B. Ganetzky, 1994. Functional identification of the *segregation distorter* locus of *Drosophila melanogaster* by germline transformation. *Genetics* 137: 201–209.
- Mckee, B.D. & M.T. Satter, 1996. Structure of the Y chromosomal *Su(Ste)* locus in *Drosophila melanogaster* and evidence for localized recombination among repeats. *Genetics* 142: 149–161.
- Michiels, F., A. Gasch, B. Kaltschmidt & R. Renkawitz-Pohl, 1989. A 14 bp promoter element directs the testis specificity of the *Drosophila* beta 2 tubulin gene. *EMBO J.* 8: 1559–1565.
- Nurminsky, D.I., E.N. Moriyama, E.R. Lozovskaya & D.L. Hartl, 1995. Molecular phylogeny and genome evolution in the *Drosophila virilis* group: duplications of the *alcohol dehydrogenase* gene. *Mol. Biol. Evol.* 13: 132–149.
- Nurminsky, D.I., E.V. Benevolenskaya, M.V. Nurminskaya, Y.Y. Shevelyov, D.L. Hartl & V.A. Gvozdev, 1998a. Cytoplasmic dynein intermediate chain isoforms with different targeting properties created by tissue-specific alternative splicing. *Mol. Cell Biol.* 18: 6816–6825.
- Nurminsky, D.I., M.V. Nurminskaya, D. De Aguiar & D.L. Hartl, 1998b. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* 396: 572–575.
- Nurminsky, D., D. De Aguiar, C.D. Bustamante & D.L. Hartl, 2001. Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*. *Science* 291: 128–130.
- Palumbo, G., S. Bonaccorsi, L.G. Robbins & S. Pimpinelli, 1994. Genetic analysis of *stellate* elements of *Drosophila melanogaster*. *Genetics* 138: 1181–1197.

- Paschal, B.M., A. Mikami, K.K. Pfister & R.B. Vallee, 1992. Homology of the 74-kD cytoplasmic dynein subunit with a flagellar dynein polypeptide suggests an intracellular targeting function. *J. Cell Biol.* 118: 1133–1143.
- Petrov, D.A. & D.L. Hartl, 1997. Trash DNA is what gets thrown away: high rate of DNA loss in *Drosophila*. *Gene* 205: 279–289.
- Petrov, D.A. & D.L. Hartl, 1998. High rate of DNA loss in the *D. melanogaster* and *D. virilis* species groups. *Mol. Biol. Evol.* 15: 293–302.
- Petrov, D.A., E.R. Lozovskaya & D.L. Hartl, 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* 384: 346–349.
- Robin, C., R.J. Russell, K.M. Medveczky & J.G. Oakeshott, 1996. Duplication and divergence of the genes of the α -esterase cluster of *D. melanogaster*. *J. Mol. Evol.* 43: 241–252.
- Russell, S.R.H. & K. Kaiser, 1994. A *Drosophila melanogaster* chromosome-2L repeat is expressed in the male germ line. *Chromosoma* 103: 63–72.
- Schroer, T.A., E.R. Steuer & M.P. Sheetz, 1989. Cytoplasmic dynein is a minus end-directed motor for membranous organelles. *Cell* 7: 331–343.
- Snyder, M. & N. Davidson, 1983. Two gene families clustered in a small region of the *Drosophila* genome. *J. Mol. Biol.* 166: 101–118.
- Steffen, W., S. Karki, K.T. Vaughan, R.B. Vallee, E.L.F. Holzbaur, D.G. Weiss & S.A. Kuznetsov, 1997. The involvement of the intermediate chain of cytoplasmic dynein in binding the motor complex to membranous organelles of *Xenopus* oocytes. *Mol. Biol. Cell* 8: 2077–2088.
- Steinemann, M. & S. Steinemann, 1990. Evolutionary changes in the organization of the major *Lcp* gene cluster during sex chromosomal differentiation in the sibling species *Drosophila persimilis*, *D. pseudoobscura* and *D. miranda*. *Chromosoma* 99: 424–431.
- Thomas, S. & R.S. Singh, 1992. 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.* 9: 507–525.
- Ting, C.T., S.C. Tsaur & C.I. Wu, 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. *Proc. Natl. Acad. Sci. USA* 97: 5313–5316.
- Ting, C.T., S.C. Tsaur, M.L. Wu & C.I. Wu, 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* 282: 1501–1504.
- Vaisberg, E.A., M.P. Koonce & J.R. McIntosh, 1993. Cytoplasmic dynein plays a role in mammalian mitotic spindle formation. *J. Cell Biol.* 123: 849–858.
- Vieira, C.P., J. Vieira & D.L. Hartl, 1997. The evolution of small gene clusters: evidence for an independent origin of the maltase gene cluster in *D. virilis* and *D. melanogaster*. *Mol. Biol. Evol.* 14: 985–993.
- Wang, W., J.M. Zhang, C. Alvarez, A. Llopart & M. Long, 2000. The origin of the *Jingwei* gene and the complex modular structure of its parental gene, *yellow emperor*, in *Drosophila melanogaster*. *Mol. Biol. Evol.* 17: 1294–1301.
- Wilkerson, C.G., S.M. King, A. Koutoulis, G.J. Pazour & G.B. Witman, 1995. The 78,000 M(r) intermediate chain of *Chlamydomonas* outer arm dynein is a WD-repeat protein required for arm assembly. *J. Cell Biol.* 129: 169–178.
- Wu, C.-I. & A.W. Davis, 1993. Evolution of postmating reproductive isolation: the composite nature of Haldane's rule and its genetic bases. *Am. Nat.* 142: 187–212.
- Wu, C.-I., N.A. Johnson & M.F. Palopoli, 1996. Haldane's rule and its legacy: why are there so many sterile males? *Trends Ecol. Evol.* 11: 281–284.
- Wu, C.-I., T.W. Lyttle, M.-L. Wu & G.-F. Lin, 1988. Association between a satellite DNA sequence and the *Responder of Segregation Distorter* in *D. melanogaster*. *Cell* 54: 179–189.
- Xiang, X., S.M. Beckwith & N.R. Morris, 1994. Cytoplasmic dynein is involved in nuclear migration in *Aspergillus nidulans*. *Proc. Natl. Acad. Sci. USA* 91: 2100–2104.