Mol Gen Genomics (2004) 272: 47–56
DOI 10.1007/s00438-004-1028-2

ORIGINAL PAPER

T. J. D. Goodwin · R. T. M. Poulter
M. D. Lorenzen · R. W. Beeman

DIRS retroelements in arthropods: identification of the recently active TcDirs1 element in the red flour beetle *Tribolium castaneum*

Received: 24 January 2004 / Accepted: 17 May 2004 / Published online: 19 June 2004

Abstract Members of the DIRS family of retrotransposons differ from most other known retrotransposons in that they encode a tyrosine recombinase (YR), a type of enzyme frequently involved in site-specific recombination. This enzyme is believed to insert the extrachromosomal DNA intermediate of DIRS element retrotransposition into the host genome. DIRS elements have been found in plants, a slime mold, fungi, and a variety of animals including vertebrates, echinoderms and nematodes. They have a somewhat patchy distribution, however, apparently being absent from a number of model organisms such as *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Drosophila melanogaster*. In this report we describe the first DIRS retroelement to be identified in an arthropod. This element, TcDirs1, was found in the red flour beetle *Tribolium castaneum* (Coleoptera). It is generally similar in sequence and structure to several previously described members of the DIRS group: it is bordered by inverted terminal repeats and it has a similar set of protein-coding domains (Gag, reverse transcriptase/ribonuclease H, and the YR), although these are arranged in a novel fashion. TcDirs1 elements exhibit several features indicative of recent activity, such as intact coding regions, a high level of sequence similarity between distinct elements and polymorphic insertion sites. Given their presence in an experimentally tractable host, these potentially active elements might serve as useful models for the study of DIRS element retrotransposition. An element closely related to TcDirs1 was also detected in sequences from a second arthropod, the honey bee *Apis mellifera* (Hymenoptera), suggesting that these retrotransposons are long-term residents of arthropod genomes.

Keywords Insect · Transposable element · Tyrosine recombinase

Communicated by G. Reuter

T. J. D. Goodwin (✉) · R. T. M. Poulter
Department of Biochemistry,
University of Otago, Cumberland Street,
Dunedin, New Zealand
E-mail: timg@sanger.otago.ac.nz
Fax: +64-3-4797866

M. D. Lorenzen
Division of Biology,
Kansas State University,
Manhattan, KS 66506, USA

R. W. Beeman
Grain Marketing and Production Research Center,
U.S. Department of Agriculture,
1515 College Ave., Manhattan,
KS 66502, USA

Introduction

The DIRS elements comprise an unusual group of retrotransposons, the mobile genetic elements that replicate via an RNA intermediate. Analyses of their reverse transcriptase (RT; Xiong and Eickbush 1990) and ribonuclease H (RH; Doolittle et al. 1989; Malik and Eickbush 2001) sequences suggest that they are related to the long-terminal-repeat (LTR) retrotransposons (Eickbush and Malik 2002). They differ in a number of important aspects from canonical LTR retrotransposons, however. Most importantly, DIRS elements do not encode a DDE integrase (Fayet et al. 1990), but instead, encode a tyrosine recombinase (Goodwin and Poulter 2001; Duncan et al. 2002). Tyrosine recombinases (YRs) are typically involved in site-specific recombinations between similar or identical DNA sequences. Representative examples include the Cre recombinase of bacteriophage P1, the FLP recombinase of yeast 2-micron circle plasmids, and the XerC and XerD recombinases of *Escherichia coli* (Nunes-Duby et al. 1998; see van Duyne 2002, for a recent review). During the replication cycle, the YRs encoded by DIRS elements are thought to mediate the insertion into the host genome of a circular double-stranded DNA copy of the DIRS element, produced by the action of the element's RT. In addition to

►

Fig. 1A–E Structures of DIRS elements. **A** Two previously described DIRS retrotransposons (DIRS1 from *Dictyostelium discoideum* and TnDirs1 from *Tetraodon nigroviridis*) with inverted terminal repeats (*hatched boxes*) and long overlapping ORFs (*shaded boxes*). Locations of the various protein domains are indicated (MT, a domain similar in sequence to phage methyltransferases; Goodwin and Poulter 2004). **B** TcDirs1.1 and flanking sequences. Exons in the flanking sequences are indicated by the *black boxes*. The orientation of these exons is indicated by the *arrows*. **C** A full-length TcDirs1 element assembled from two partial sequences (see text for details). The extent of the deletion in TcDirs1.1 is indicated. **D** Examples of TcDirs1-like elements from echinoderms: SpDirs1 from *Strongylocentrotus purpuratus* and ApDirs1 from *Arbacia punctulata*. The sequence of the extreme right end of ApDirs1 is not yet available (indicated by the *question mark*). The scale for panels **A–D** is indicated to the right of panel **D**. **E** ORF maps for the coding regions of TcDirs1 and ApDirs1. ATG codons and stop codons for each of the three forward reading frames are shown *above and below the lines*, respectively. Sources of sequences are listed in the Materials and methods section

encoding a YR, DIRS elements differ from typical LTR retrotransposons in that they encode a protein domain similar in sequence to phage methyltransferases, although the function of this domain is not known at present (Goodwin and Poulter 2004); they also lack genes for aspartic proteases. Furthermore, DIRS elements usually contain one of two distinct arrangements of terminal repeat sequences, each of which differs from that of typical LTRs. For instance, DIRS1, isolated from the slime mold *Dictyostelium discoideum* (Cappello et al. 1985), has inverted terminal repeats (ITRs), and the outer extremities of these ITRs are repeated internally in a sequence known as the internal complementary region (ICR; Fig. 1A). In contrast, the PAT element from the nematode *Panagrellus redivivus* (de Chastonay et al. 1992) has split direct repeats (SDRs), in which the terminal sequences are repeated adjacent to each other in the internal region in a nested fashion (A_1-B_1 A_2-B_2). The terminal repeats of DIRS elements may, nevertheless, play a similar role to that of the LTRs of canonical LTR retrotransposons, i.e., allowing the synthesis of a full-length DNA copy of the element from slightly-less-than-full-length transcripts.

Phylogenetic analyses based on alignments of RT-RH sequences divide the DIRS elements into two subgroups (Goodwin and Poulter 2004). Each subgroup is composed of elements with similar arrangements of repeat sequences, i.e., one contains all the elements with inverted terminal repeats, and the other contains all the elements with split direct repeats. We shall refer to these as the ITR and SDR subgroups, respectively. These two subgroups can also be distinguished by the arrangement of their coding regions: in the SDR elements the ORF encoding the YR either does not overlap, or has only a very short overlap, with the RT-RH ORF. In contrast, in the ITR elements the ORF encoding the YR overlaps the entire RT-RH ORF, and extends sufficiently far in the 5′ direction that it also overlaps the 3′ end of the first ORF (encoding a putative Gag protein). The actual YR-encoding sequence lies downstream of the RT-RH ORF (Fig. 1A). It has been suggested that this unusual arrangement of ORFs in the ITR elements enables the downstream YR coding region to be expressed in such a way that the YR protein does not end up being covalently attached to the RT-RH protein (Goodwin and Poulter 2001); i.e., if, as has been found in several LTR retroelements (e.g., Farabaugh et al. 1993), translation of the downstream ORFs is mediated by programmed ribosomal frameshifts, then a shift to the +1 reading frame at the end of ORF1 would result in translation of the YR ORF, whereas a shift to the −1 frame would permit independent translation of the RT-RH ORF.

Recently, a second group of tyrosine recombinase-encoding retrotransposons, the Ngaro group, has been described (Goodwin and Poulter 2004). Ngaro elements are broadly similar in structure to DIRS elements of the SDR subgroup, and, like DIRS elements, they lack genes for a DDE integrase and an aspartic protease. They can be distinguished from DIRS elements, however, by comparisons of the RT, RH and YR domains, and by the absence from all known Ngaro elements of the methyltransferase-like domain characteristic of DIRS elements (Goodwin and Poulter 2004).

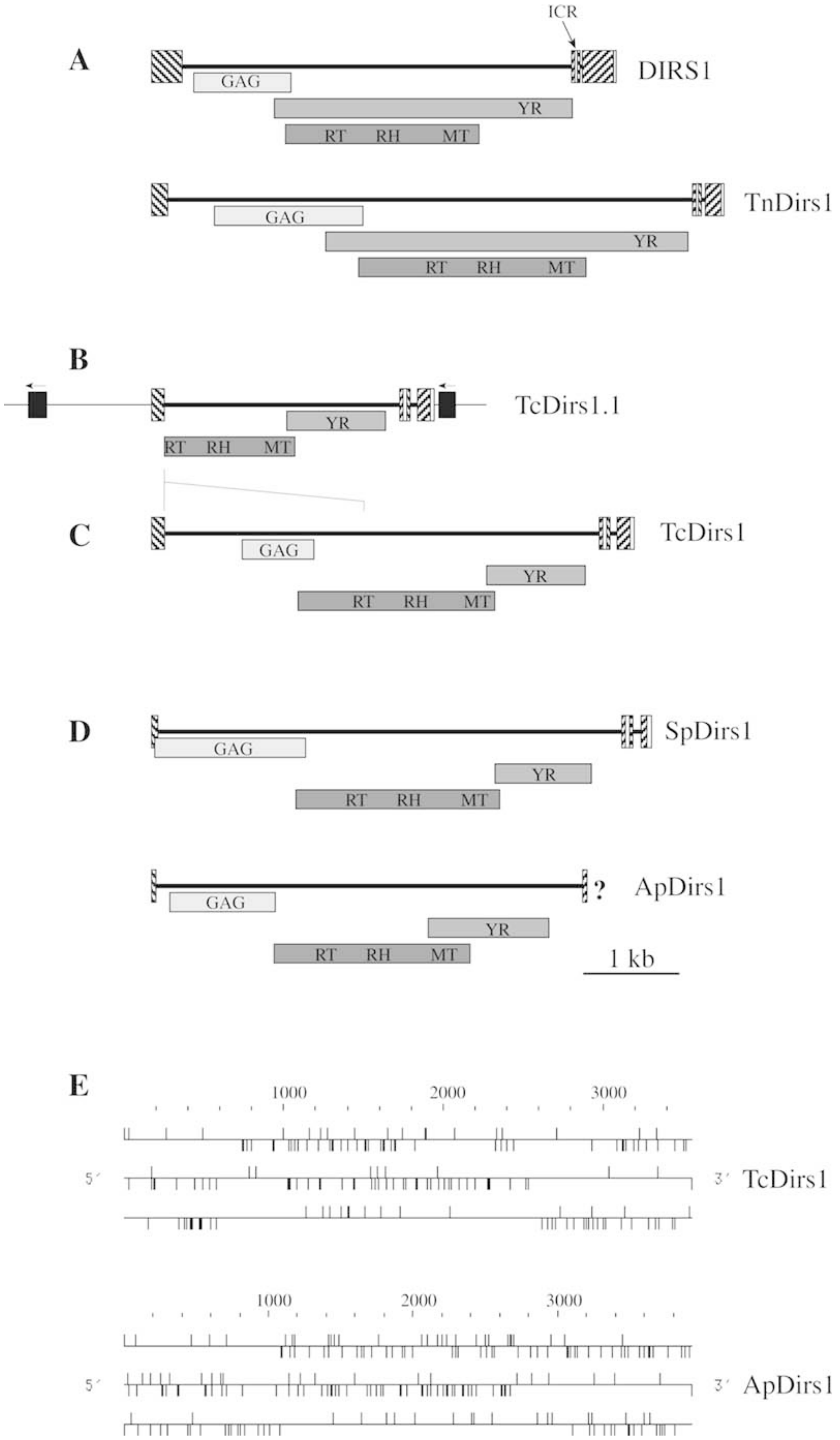
DIRS-like retrotransposons have been identified in a wide range of eukaryotes, including a slime mold (Cappello et al. 1985), fungi (Ruiz-Perez et al. 1996), plants (Duncan et al. 2002), and various animals including nematodes (de Chastonay et al. 1992), echinoderms and vertebrates (Goodwin and Poulter 2001). No DIRS elements have previously been identified in arthropods, however, despite the intense study of insect transposable elements and the availability of the genome sequences of *Drosophila melanogaster* (Diptera; Adams et al. 2000) and *Anopheles gambiae* (Diptera; Holt et al. 2002).

Here we report the first identification of a DIRS-like retrotransposon in an arthropod. This element, TcDirs1, was found in the red flour beetle *Tribolium castaneum* (Coleoptera). TcDirs1 falls clearly within the ITR subgroup of DIRS elements, both on the basis of its inverted terminal repeats and on phylogenetic analyses of predicted protein sequences. It can, however, be distinguished from previously described members of the ITR group by the arrangement of its coding regions. TcDirs1 exhibits several features which suggest that it was recently mobile, and some copies may still be active. Such elements may make useful models for analysing the retrotransposition of DIRS elements, given their presence in a relatively tractable host.

Materials and methods

Strains and culture conditions

The majority of *T. castaneum* field strains used in this work were collected between 1985 and 1990 from farms, mills, grain storage facilities and warehouses around the world. These strains have been maintained at the


A
ICR
DIRS1
GAG
YR
RT RH MT
TnDirs1
GAG
YR
RT RH MT
B
TcDirs1.1
YR
RT RH MT
C
TcDirs1
GAG
YR
RT RH MT
D
SpDirs1
GAG
YR
RT RH MT
ApDirs1
?
GAG
YR
RT RH MT
1 kb
E
1000
2000
3000
5´
3´
TcDirs1
1000
2000
3000
5´
3´
ApDirs1

Tribolium Stock Center (USDA-ARS-GMPRC, Manhattan, Kansas) under standard conditions (Beeman et al. 1986). The strains used, and their places of origin, are as follows: JPN-2 (Japan), NDG-2 (Canada), IND-1 (India), AUS-2 (Australia), PHL-1 (Philippines), THA-1 (Thailand), CRI-1 (Costa Rica), UGA-1 (Uganda). Strain M^1, derived from a Singaporean strain, is described in Beeman and Friesen (1999), and GA-2 is a near-homozygous inbred derivative of GA-1 (S. Thompson, unpublished). GA-1 is a standard laboratory strain that was originally collected from a farmer's corn bin in the USA (Haliscak and Beeman 1983).

DNA isolation

DNA isolation from single beetles was performed using the Wizard Genomic DNA isolation kit (Promega, Madison, Wis.) according to the manufacturer's protocol.

PCR

PCRs were performed in an Eppendorf Mastercycler Gradient instrument, using either the Expand High-Fidelity PCR system or the Expand Long-Template PCR system (Roche). Oligonucleotides were obtained from Proligo (Singapore). Primers used to amplify the TcDirs1.1 insertion site were TribF1 (5′-CTTGGGATCCTAGTGTTTGTTGAGAAAACC-3′) and TribR1 (5′-CTTCGCATGCGTGTTGGACTTTTACGTC-3′). The primers used to amplify the region of TcDirs1 that is missing from TcDirs1.1 were TribFR4 (5′-CTTGGGATCCGTTAGATGGTAGCACAC-3′) and TribR2 (5′-CTTCGCATGCTAGCAAGAAATCGTCGAG-3′). The underlined bases correspond to sequences that are not complementary to the target DNA, but were added to the primers to facilitate cloning.

Cloning and sequencing

Recombinant DNA manipulations were carried out using standard procedures (Sambrook and Russell 2001). Bacterial plasmid DNA was prepared using an alkaline lysis/polyethylene glycol precipitation method from Applied BioSystems. Sequencing was performed at the University of Otago using an ABI377 DNA Sequencer.

Sequence analyses

General sequence analyses were performed using the programs in the Wisconsin GCG package (Genetics Computer Group 1994) and the Australian National Genomic Information Service node located at the University of Otago (http://angis.otago.ac.nz). Sequence similarity searches were performed using the BLAST servers at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/) and the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/E-mail/homology.html). Multiple sequence alignments were constructed using CLUSTAL_X (Thompson et al. 1997) and adjusted using SEAVIEW (Galtier et al. 1996). Phylogenetic trees were constructed using PAUP* 4b10 (Swofford 1998). Full-length consensus sequences of several elements were constructed from whole-genome shotgun sequence data. For this purpose, multiple overlapping fragments of each element were first identified using BLAST (http://www.ncbi.nlm.nih.gov/BLAST/mmtrace.shtml). The termini were then defined by comparisons of the sequences of copies of the element inserted at different loci, and by comparisons between occupied and related empty sites. A representative full-length sequence was then constructed, and this was converted to a consensus by identifying the most common base at each position using the 'query-anchored with identities' display option of the BLAST results. Construction of a consensus sequence usually required only a small number of changes, as the sequences involved were generally of high quality and the various different elements of the families usually appear to be homogeneous in sequence.

Accession numbers

Sequences described in this report have been submitted to the DDBJ/EMBL/GenBank databases under the following Accession Nos.: TcDirs1.1, AY531876; segment of TcDirs1 missing from TcDirs1.1, AY531877; TcDirs1.1 related empty site, AY531878. The AmDirs1 element is present on sequences AADG02016821 and AADG02016822. The sequences of SpDirs1 and ApDirs1 are available in the Third Party Annotation section of the DDBJ/EMBL/GenBank databases under Accession Nos. TPA: BK005158 and BK004821, respectively. The assembled sequences of the SpDirs elements, the assembled full-length sequence of TcDirs1, and the alignment used to generate the phylogenetic tree (see below) are available on the Poulter laboratory website (http://biocadmin.otago.ac.nz/retrobase/home.htm).

Sources of additional sequences mentioned in this work were as follows: bacteriophage lambda, J02459; bacteriophage P1 Cre, X03453; *Caenorhabditis briggsae* CbPat1, AC090521; *Chlamydomonas reinhardtii*TOC3, *C. reinhardtii* draft genome sequence (http://genome.jgi-psf.org/chlre1/chlre1.home.html) scaffold 2543 (443–5971); *Danio rerio*DrDirs1, AL590134; DrDirs2, BK001257; DrDirs3, BK001259; *Dictyostelium discoideum* DIRS1, M11339; *Drosophila melanogaster* gypsy, M12927; *Panagrellus redivivus* PAT, X60774; *Phycomyces blakesleeanus* Prt1, Z54337; *Rhizopus oryzae* RoDirs1, assembled consensus sequence available at http://biocadmin.otago.ac.nz/retrobase/home.htm; *Saccharomyces cerevisiae* Ty3, M23367; *S. purpuratus* SpPat1, assembled consensus sequence available at http://biocadmin.otago.ac.nz/retrobase/home.htm; *Tetraodon nigroviridis* TnDirs1, AF442732; *Volvox carteri* Kangaroo1, AY137241; *Xenopus tropicalis* XtDirs1, AC144974; XtDirs2, AC145807.

Results

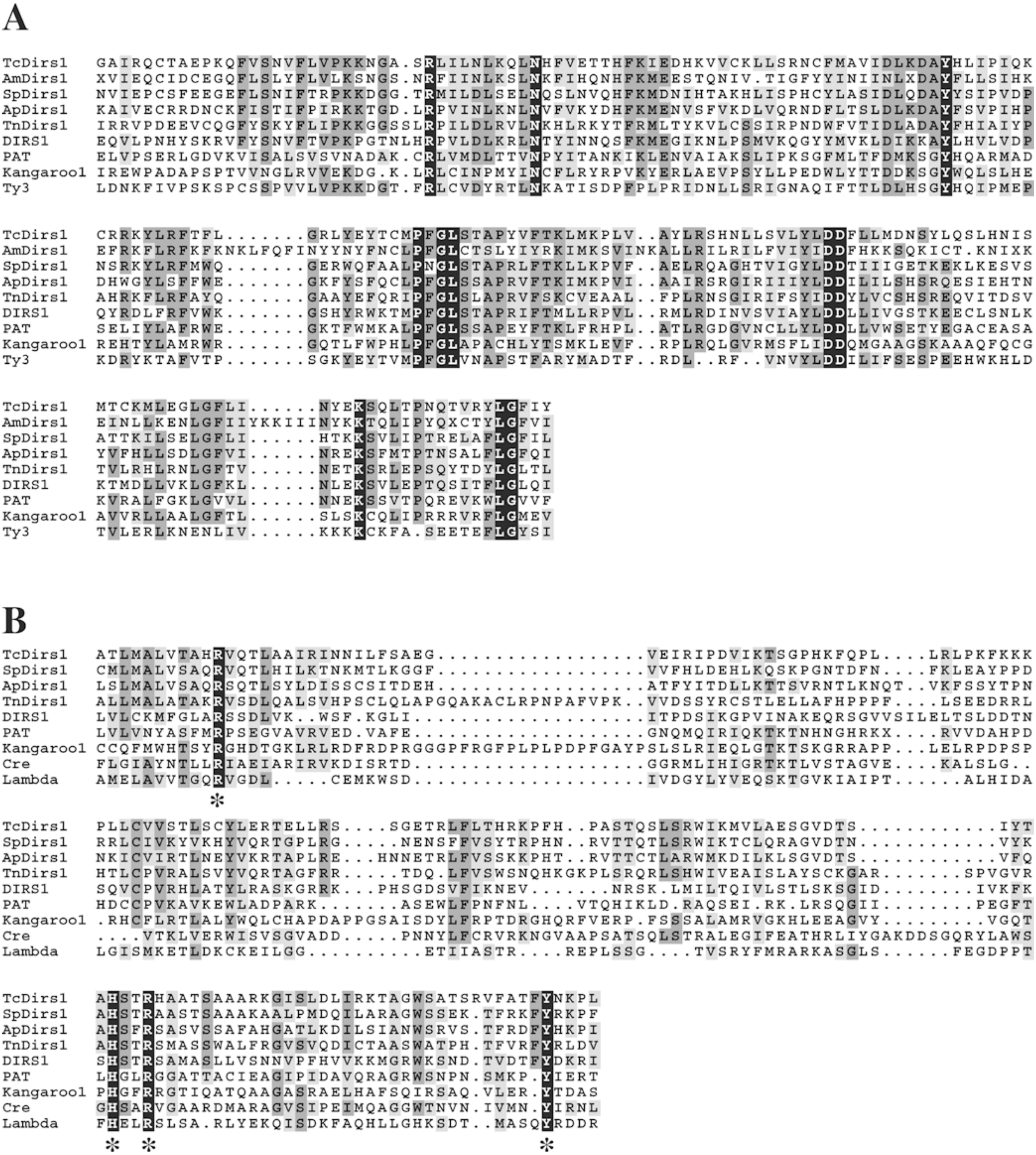
A DIRS element in *T. castaneum*

A retrotransposon with DIRS-like features was first identified in *T. castaneum* during the sequencing of a GA-2 BAC clone isolated during the positional cloning of the *Medea* ^{1}locus (R. W. Beeman, unpublished data). This element was found to contain two long ORFs (Fig. 1B), the first representing a RT-RH ORF, and the second encoding a tyrosine recombinase. The predicted products of each ORF were found to be most similar in sequence to proteins encoded by previously identified DIRS elements (Fig. 2, and data not shown). The ORFs are flanked by a set of repeat sequences similar to those characteristic of the ITR subgroup of DIRS elements, i.e., there are inverted terminal repeats, and downstream of the YR ORF there is an internal complementary region corresponding to the outer extremities of the terminal repeats (Fig. 1B). Because of its DIRS-like nature, this element was named TcDirs1.1 (GenBank Accession No. AY531876), while elements of this general family are referred to as TcDirs1. TcDirs1.1 appears to be an internally deleted element, as its RT-RH ORF lacks the 5′ end of the RT-encoding region, and unlike other DIRS elements it does not contain an additional long 5′ ORF. Apart from this deletion, however, the element appears to be intact—the ORFs are free of nonsense mutations and the predicted protein products contain all the highly conserved residues characteristic of DIRS-like RT, RH and YR proteins (Fig. 2, and not shown).

The identification of a DIRS-like retrotransposon in *T. castaneum* is of interest, as this is a relatively easy organism to work with and it is amenable to molecular genetic manipulation (Lorenzen et al. 2003). Few, if any, of the other species known to host DIRS elements are as experimentally tractable, with the result that no model system for studying transposition of these elements has been developed to date. We were therefore eager to learn whether or not any full-length and potentially active elements remain, and to discover whether there is any evidence for recent mobility of such elements.

As an initial test for recent mobility of the TcDirs1 element, and in order to define the termini precisely, we sought to identify ‘related empty sites’ in a variety of other *T. castaneum* strains. [Related empty sites are sequences related (and perhaps allelic) to the site of a mobile element’s insertion, but which lack the element]. TcDirs1.1 was found to lie within an intron of a gene homologous to the *A. gambiae* gene encoding protein EAA11927 (Fig. 1B, and not shown). Therefore, to maximise the chances of identifying related empty sites in other strains, PCR primers were designed to the flanking exon sequences. PCR products of an appropriate size to represent empty sites (~1.3 kb) were subsequently obtained from three of the four strains tested. (The strain that did not give a product was GA-2, the strain from which TcDirs1.1 was obtained; this may be because the element is homozygous at this locus in this highly inbred strain.) One related empty site was cloned and sequenced (Accession No. AY531878). It was found to be highly similar in sequence to the regions flanking TcDirs1.1 (97.8% identity over a 1251-bp overlap). The only major difference between the two sequences is the insertion of the TcDirs1 element. This finding suggests that TcDirs1 has transposed into this site fairly recently (sufficiently recently that these intronic sequences have not diverged substantially, and that the site is polymorphic in the species). Comparisons of the inserted and empty site sequences (Fig. 3) revealed that, as expected, the termini of the element correspond precisely to the ends of the repeats. Furthermore, this work showed that the element had inserted at a sequence consisting of an ‘AA’ dinucleotide. This is identical to the sequence that would form the putative circular junction of the termini of the extrachromosomal intermediate (underlined in Fig. 3). The element may thus have inserted into this locus by recombination between these two identical sequences, a strategy similar to the insertion mechanism postulated for other DIRS elements (Goodwin and Poulter 2001).

In order to examine the distribution of TcDirs1 elements among *T. castaneum* strains, and to test for the possible presence of full-length TcDirs1 elements, PCR primer pairs that correspond to various regions of TcDirs1.1 were made. These were then employed in reactions with genomic DNA from ten different strains from diverse geographical locations. Firstly, primers corresponding to regions flanking the site of the predicted deletion in TcDirs1.1 were used to test for the presence of potentially full-length elements (Fig. 4). Six of the ten strains examined in this way yielded a major PCR product of ~2.2 kb in size, together with a range of smaller bands. The finding that a majority of strains gave this ~2.2-kb product, but none gave a larger one, suggested that these ~2.2-kb fragments might be derived from full-length versions of TcDirs1. The range of smaller and less abundant PCR products found in many strains may represent other elements carrying various deletions. Only one strain, GA-2 (from which the BAC library was derived), gave a PCR product of a size corresponding to the particular deletion found in TcDirs1.1 (0.2 kb; Fig. 4, lane 1). Of the ten strains examined for the presence of TcDirs1 elements, eight gave a positive result with PCR (Fig. 4 and not shown). The remaining two strains (JPN-2 and THA-1) failed to give a positive result with any of three different primer pairs which were shown to detect TcDirs1 elements in other strains, but did give a positive result with primers



```
TcDirs1    GAIRQCTAEPKQFVSNVFLVPKKNGA.SRLILNLKQLNHFVETTHFKIEDHKVVCKLLSRNCFMAVIDLKDAYHLIPIQK
AmDirs1    XVIEQCIDCEGQFLSLYFLVLKSNGS.NRFIINLKSLNKFIHQNHFKMEESTQNIV.TIGFYYINIINLXDAYFLLSIHK
SpDirs1    NVIEPCSFEEGEFLSNIFTRPKKDGG.TRMILDLSELNQSLNVQHFKMDNIHTAKHLISPHCYLASIDLQDAYYSIPVDP
ApDirs1    KAIVECRRDNCKFISTIFPIRKKTGD.LRPVINLKNLNVFVKYDHFKMENVSFVKDLVQRNDFLTSLDLKDAYFSVPIHP
TnDirs1    IRRVPDEEVCQGFYSKYFLIPKKGGSSLRPILDLRVLNKHLRKYTFRMLTYKVLCSSIRPNDWFVTIDLADAYFHIAIYP
DIRS1      EQVLPNHYSKRVFYSNVFTVPKPGTNLHRPVLDLKRLNTYINNQSFKMEGIKNLPSMVKQGYYMVKLDIKKAYLHVLVDP
PAT        ELVPSERLGDVKVISALSVSVNADAK.CRLVMDLTTVNPYITANKIKLENVAIAKSLIPKSGFMLTFDMKSGYHQARMAD
Kangaroo1  IREWPADAPSPTVVNGLRVVEKDG.K.LRLCINPMYINCFLRYRPVKYERLAEVPSYLLPEDWLYTTDDKSGYWQLSLHE
Ty3        LDNKFIVPSKSPCSSPVVLVPKKDGT.FRLCVDYRTLNKATISDPFPLPRIDNLLSRIGNAQIFTTLDLHSGYHQIPMEP

TcDirs1    CRRKYLRFTFL.......GRLYEYTCMPFGLSTAPYVFTKLMKPLV..AYLRSHNLLSVLYLDDFLLMDNSYLQSLHNIS
AmDirs1    EFRKFLRFKFKNKLFQFINYYNYFNCLPFGLCTSLYIYRKIMKSVINKALLRILRILFVIYIDDFHKKSQKICT.KNIXK
SpDirs1    NSRKYLRFMWQ.......GERWQFAALPNGLSTAPRLFTKLLKPVF..AELRQAGHTVIGYLDDTIIIGETKEKLKESVS
ApDirs1    DHWGYLSFFWE.......GKFYSFQCLPFGLSSAPRVFTKIMKPVI..AAIRSRGIRIIIYLDDILILSHSRQESIEHTN
TnDirs1    AHRKFLRFAYQ.......GAAYEFQRIPFGLSLAPRVFSKCVEAAL..FPLRNSGIRIFSYIDDYLVCSHSREQVITDSV
DIRS1      QYRDLFRFVWK.......GSHYRWKTMPFGLSTAPRIFTMLLRPVL..RMLRDINVSVIAYLDDLLIVGSTKEECLSNLK
PAT        SELIYLAFRWE.......GKTFWMKALPFGLSSAPEYFTKLFRHPL..ATLRGDGVNCLLYLDDLLVWSETYEGACEASA
Kangaroo1  REHTYLAMRWR.......GQTLFWPHLPFGLAPACHLYTSMKLEVF..RPLRQLGVRMSFLIDDQMGAAGSKAAAQFQCG
Ty3        KDRYKTAFVTP.......SGKYEYTVMPFGLVNAPSTFARYMADTF..RDL..RF..VNVYLDDILIFSESPEEHWKHLD

TcDirs1    MTCKMLEGLGFLI......NYEKSQLTPNQTVRYLGFIY
AmDirs1    EINLLKENLGFIIYKKIIINYKKTQLIPYQXCTYLGFVI
SpDirs1    ATTKILSELGFLI......HTKKSVLIPTRELAFLGFIL
ApDirs1    YVFHLLSDLGFVI......NREKSFMTPTNSALFLGFQI
TnDirs1    TVLRHLRNLGFTV......NETKSRLEPSQYTDYLGLTL
DIRS1      KTMDLLVKLGFKL......NLEKSVLEPTQSITFLGLQI
PAT        KVRALFGKLGVVL......NNEKSSVTPQREVKWLGVVF
Kangaroo1  AVVRLLAALGFTL......SLSKCQLIPRRRVRFLGMEV
Ty3        TVLERLKNENLIV......KKKKCKFA.SEETEFLGYSI
```

```
TcDirs1    ATLMALVTAHRVQTLAAIRINNILFSAEG..................VEIRIPDVIKTSGPHKFQPL....LRLPKFKKK
SpDirs1    CMLMALVSAQRVQTLHILKTNKMTLKGGF..................VVFHLDEHLKQSKPGNTDFN....FKLEAYPPD
ApDirs1    LSLMALVSAQRSQTLSYLDISSCSITDEH..................ATFYITDLLKTTSVRNTLKNQT..VKFSSYTPN
TnDirs1    ALLMALATAKRVSDLQALSVHPSCLQLAPGQAKACLRPNPAFVPK..VVDSSYRCSTLELLAFHPPPF....LSEEDRRL
DIRS1      LVLCKMFGLARSSDLVK..WSF.KGLI....................ITPDSIKGPVINAKEQRSGVVSILELTSLDDTN
PAT        LVLVNYASFMRPSEGVAVRVED.VAFE....................GNQMQIRIQKTKTNHNGHRKX....RVVDAHPD
Kangaroo1  CCQFMWHTSYRGHDTGKLRLRDFRDPRGGGPFRGFPLPLPDPFGAYPSLSLRIEQLGTKTSKGRRAPP...LELRPDPSP
Cre        FLGIAYNTLLRIAEIARIRVKDISRTD....................GGRMLIHIGRTKTLVSTAGVE....KALSLG..
Lambda     AMELAVVTGQRVGDL.....CEMKWSD....................IVDGYLYVEQSKTGVKIAIPT......ALHIDA
                     *

TcDirs1    PLLCVVSTLSCYLERTELLRS....SGETRLFLTHRKPFH..PASTQSLSRWIKMVLAESGVDTS............IYT
SpDirs1    RRLCIVKYVKHYVQRTGPLRG.....NENSFFVSYTRPHN..RVTTQTLSRWIKTCLQRAGVDTN............VYK
ApDirs1    NKICVIRTLNEYVKRTAPLRE...HNNETRLFVSSKKPHT..RVTTCTLARWMKDILKLSGVDTS............VFQ
TnDirs1    HTLCPVRALSVYVQRTAGFRR.....TDQ.LFVSWSNQHKGKPLSRQRLSHWIVEAISLAYSCKGAR.......SPVGVR
DIRS1      SQVCPVRHLATYLRASKGRRK...PHSGDSVFIKNEV.......NRSK.LMILTQIVLSTLSKSGID........IVKFK
PAT        HDCCPVKAVKEWLADPARK.......ASEWLFPNFNL....VTQHIKLD.RAQSEI.RK.LRSQGII........PEGFT
Kangaroo1  .RHCFLRTLALYWQLCHAPDAPPGSAISDYLFRPTDRGHQRFVERP.FSSSALAMRVGKHLEEAGVY.........VGQT
Cre        ....VTKLVERWISVSGVADD.....PNNYLFCRVRKNGVAAPSATSQLSTRALEGIFEATHRLIYGAKDDSGQRYLAWS
Lambda     LGISMKETLDKCKEILGG..........ETIIASTR....REPLSSG....TVSRYFMRARKASGLS......FEGDPPT

TcDirs1    AHSTRHAATSAAARKGISLDLIRKTAGWSATSRVFATFYNKPL
SpDirs1    AHSTRAASTSAAAKAALPMDQILARAGWSSEK.TFRKFYRKPF
ApDirs1    AHSFRSASVSSAFAHGATLKDILSIANWSRVS.TFRDFYHKPI
TnDirs1    AHSTRSMASSWALFRGVSVQDICTAASWATPH.TFVRFYRLDV
DIRS1      SHSTRSAMASLLVSNNVPFHVVKKMGRWKSND.TVDTFYDKRI
PAT        LHGLRGGATTACIEAGIPIDAVQRAGRWSNPN.SMKP.YIERT
Kangaroo1  PHGFRRGTIQATQAAGASRAELHAFSQIRSAQ.VLER.YTDAS
Cre        GHSARVGAARDMARAGVSIPEIMQAGGWTNVN.IVMN.YIRNL
Lambda     FHELRSLSA.RLYEKQISDKFAQHLLGHKSDT..MASQYRDDR
            *  *                                 *
```

Fig. 2A, B Reverse transcriptase and tyrosine recombinase alignments. **A** An alignment of the region encompassing the seven conserved domains of RT described by Xiong and Eickbush (1990). Note that it was necessary to make several changes (not shown) to the degenerate AmDirs1 sequence to reconstruct the RT coding sequence. **B** An alignment of the region encompassing the conserved RHRY residues in tyrosine recombinase (indicated by the *asterisks*). Sources of sequences are listed in the Materials and methods section

designed to amplify the genomic DNA adjacent to the TcDirs1.1 insertion. It is possible that these two strains lack the TcDirs1 element, or contain elements that are divergent in sequence.

Three copies of the ~2.2-kb PCR product, from two different strains, were cloned and sequenced. (One of these sequences is available in the GenBank/EMBL/

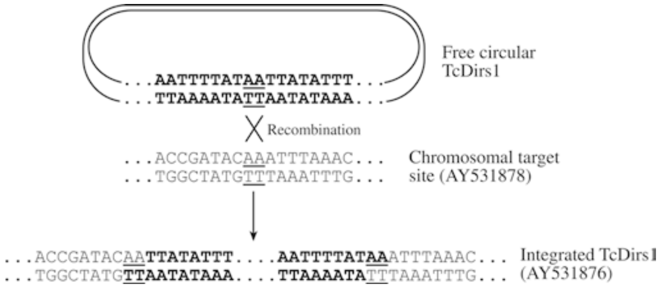

...AATTTTATAATTATATTT...
...TTAAAATATTAATATAAA...
Free circular TcDirs1
Recombination
...ACCGATACAAATTTAAAC...
...TGGCTATGTTTAAATTTG...
Chromosomal target site (AY531878)
...ACCGATACAATTATATTT....AATTTTATAAATTTAAAC...
...TGGCTATGTTAATATAAA....TTAAAATATTTAAATTTG...
Integrated TcDirs1 (AY531876)

Fig. 3 Flanking sequences, related empty site and possible mechanism of insertion of TcDirs1.1. TcDirs1.1 (*bold face*) may have inserted into the host genome (*standard type*) by a recombination between an AA dinucleotide (*underlined*) at the junction of the left and right ITRs in a circular molecule produced by reverse transcription, and an AA dinucleotide (*underlined*) in the host genome (present in the related empty site described in the text)

DDBJ databases under Accession No. AY531877.) The three sequences are highly similar to each other (≥99.5% identity over 2.2 kb) and each contains a 2071-bp insertion relative to TcDirs1.1. In the overlapping regions (143 bp) two of the sequences are identical to TcDirs1.1, while the third has a single base substitution. The new sequences each contain the section of the RT-RH ORF that is missing from TcDirs1.1 and also contain an additional upstream ORF, probably corresponding to the first ORFs of other DIRS elements. The finding that these three sequences are highly similar, with no insertions or deletions, and contain all the expected sequences suggests that they indeed represent the missing section of TcDirs1, and that full-length TcDirs1 elements are likely to be present in *T. castaneum*. The high level of sequence similarity found between different elements, and the paucity of inactivating mutations, suggest that elements of this family have transposed recently and raise the possibility that some may still be active.

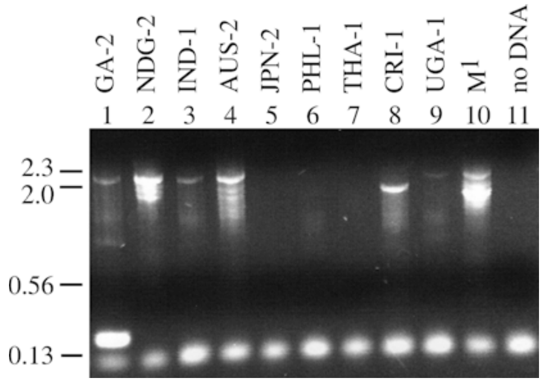


Fig. 4 Results of a PCR screen of ten *T. castaneum* strains for potentially full-length TcDirs1 elements. Primers designed to flank the site of the deletion in TcDirs1.1 were used in reactions with genomic DNA from each of the indicated strains. Fragment sizes (kb) are indicated on the *left*. The intense band running at ~200 bp obtained from strain GA-2 is of the expected size to be derived from TcDirs1.1 itself. The bands that have higher mobility than the 0.13-kb marker probably represent primer dimers

The sequence of TcDirs1.1 and one of the sequences of the section lost from this element were combined to generate a representative sequence of a full-length TcDirs1 element. The overall structure of this element is similar to that of other DIRS elements of the ITR group (Fig. 1C). For instance, it has a similar overall size and a similar complement of conserved protein coding domains. The nature of the repeat sequences is also very similar to that of previously described elements. For instance, as is often found with other ITR elements, the two ITRs of TcDirs1 are not identical. The right ITR contains an extension at its 3′ end relative to the left ITR and there are several base substitutions and an indel. These differences are confined to the outer regions of the ITRs; the inner portions are perfectly complementary. The outer regions of each ITR are, however, perfectly complementary to the corresponding regions of the ICR. These findings are consistent with the model for DIRS1 replication proposed by Cappello et al. (1985) in which much of the inner region of the left ITR is generated using the right ITR as a template, while the outer extremities of each ITR are copied off the ICR.

There is one striking structural difference between TcDirs1 and the previously described ITR elements: the RT-RH ORF of TcDirs1 is only overlapped by a short 5′ extension of the YR ORF (Fig. 1C, 1E). This suggests that TcDirs1 uses a different mechanism to express the tyrosine recombinase from that employed by the other members of the ITR sub-group. Given that the 5′ end of the YR ORF overlaps the 3′ end of the RT-RH ORF, it is possible that translation of the YR occurs via a programmed ribosomal frameshift in the region of overlap (see below for further discussion of this point).

A TcDirs1-like element in the honey bee

A DIRS retroelement was also detected in the recently released genome sequence assembly of the honey bee *Apis mellifera*. This element (AmDirs1) is the closest known relative of TcDirs1 (see below) but, unlike the *T. castaneum* element, it is somewhat degenerate (suffering from a number of frameshift and nonsense mutations; not shown). It appears in just a single copy in the *A. mellifera* genome assembly [the remnants of the RT-RH ORF being present on sequence AADG02016821 and the remnants of the YR ORF lying on the adjacent (and partially overlapping) sequence AADG02016822]. The presence of this TcDirs1-like element in a hymenopteran, a distant relative of *T. castaneum* (Coleoptera) suggests that DIRS retroelements are long-term residents of insect genomes, rather than being recent acquisitions by horizontal transfer.

TcDirs1-like elements in echinoderms

In a previous report we described some partial DIRS elements from the sea urchin *Strongylocentrotus* *purpuratus* that have RT-RH ORFs which lack long overlapping ORFs in a different reading frame, and therefore might have structures that differ from those of other DIRS elements (Goodwin and Poulter 2001). Given that a large amount of genomic sequence data for *S. purpuratus*has recently become available, we have now been able to investigate the structure of these *S. purpuratus* elements in detail. Full-length consensus sequences of three elements (SpDirs1, 3 and 4) were constructed from the available *S. purpuratus* whole-genome shotgun sequence data (as described in Materials and methods). These elements were all found to have structures very similar to that of TcDirs1 (for example, see Fig. 1D). In particular, in each of these elements the YR ORF does not entirely overlap the RT-RH ORF; instead, the 5′ end of the YR ORF shows just a short overlap with the 3′ end of the RT-RH ORF. This arrangement is consistent with the possibility that translation of the YR is achieved via a programmed ribosomal frameshift near the end of the RT-RH ORF, as suggested above for TcDirs1.

An almost full-length TcDirs1-like element was also identified in a draft BAC sequence from the sea urchin *Arbacia punctulata* (AC146998). This element, ApDirs1, again has a very similar structure to TcDirs1 and SpDirs1, 3 and 4 (Fig. 1D), although the extreme 3′ end of this element is missing from this unfinished sequence.

Relationships among DIRS-like retrotransposons

Phylogenetic trees based on alignments of RT-RH sequences were constructed to examine the relationships among members of the DIRS group (a typical example, obtained by the neighbor-joining method, is shown in Fig. 5). As mentioned above, such trees separate the DIRS elements into two major groups, one containing the elements with split direct repeats, and the other those with inverted terminal repeats. As expected, the TcDirs1 element was found to group with the ITR elements. Within the ITR subgroup, TcDirs1 and AmDirs1 (from the honey bee) appear as each other's closest known relative, and these elements group with the echinoderm elements of similar structure, forming a monophyletic cluster (albeit without high levels of bootstrap support). A close relationship among these elements was also indicated by trees based on alignments of YR sequences (not shown), and by sequence similarity between the putative ORF1 proteins (not shown). The emergence of the TcDirs1 cluster from within the group of ITR elements that contain the long overlapping ORFs (rather than appearing as a sister taxon), suggests that the TcDirs1-like elements are descended from such an element, but have since evolved an alternative method for expressing the YR ORF. The presence of TcDirs1-like elements in both insects (Arthropoda) and sea urchins (Echinodermata), and their apparent monophyletic origin, suggests that these retrotransposons evolved relatively early in metazoan evolution.

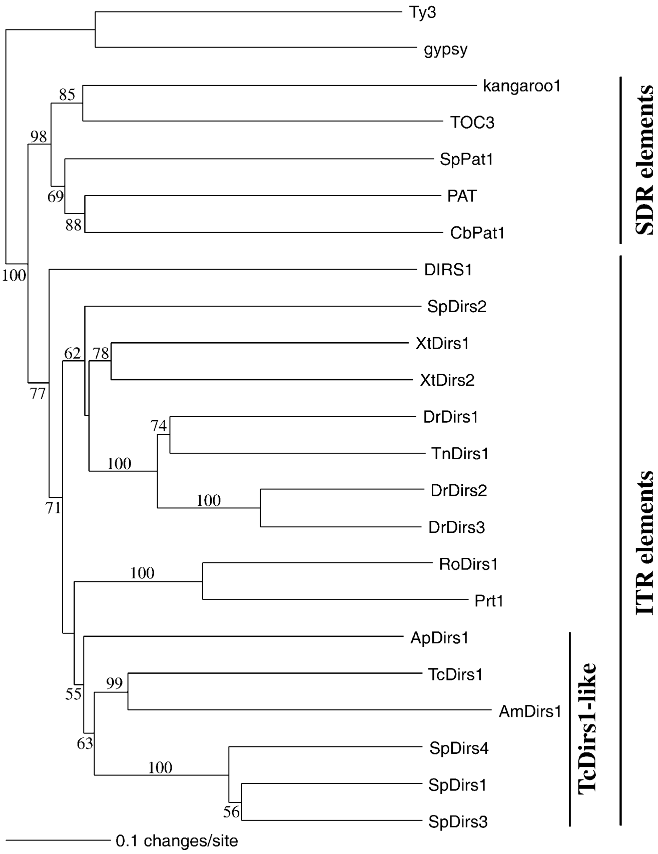


Fig. 5 Relationships among DIRS retrotransposons. This tree is based on an alignment of RT and RH sequences. It is a consensus of 100 bootstrap replicates and was obtained by the neighbor-joining method using the heuristic search option (PAUP*4b10; Swofford 1998). The levels of bootstrap support for nodes receiving > 50% support are indicated. Sources of sequences are listed in the Materials and methods section

Discussion

TcDirs1 is the first DIRS retrotransposon to be found in an arthropod. It has several features which suggest that it has transposed recently, such as a high level of sequence similarity between different elements, intact coding regions, and related empty sites highly similar to the sequences flanking a known insertion. These features are indicative of recent activity and transposition of these elements, and suggest that there is no particular barrier to the replication of DIRS elements in arthropods. The apparent absence of these elements from well-characterised insect species such *D. melanogaster* and *A. gambiae* is probably simply due to their random loss. Given that *T. castaneum* is an experimentally tractable organism (see, for example, Lorenzen et al. 2003), the likelihood that full-length and potentially active TcDirs1 retrotransposons exist in some strains is of particular interest, as these could form an attractive model system for analysis of the transposition of DIRS-like elements. It might also be possible to introduce such elements into *D. melanogaster* and take advantage of the genetic tools available for this species.

In DIRS1 and other previously characterised members of the ITR group, the 5′ end of the tyrosine recombinase ORF completely overlaps the RT-RH ORF. Indeed, it extends sufficiently far in the 5′ direction that it overlaps the 3′ end of ORF1 (the putatively *gag* -like ORF; Goodwin and Poulter 2001), as does the 5′ end of the RT-RH ORF. This unusual arrangement means that translation of both these ORFs could be achieved as a result of programmed ribosomal frameshifts near the end of ORF1: a shift into the +1 reading frame would result in translation of the YR ORF, whereas a shift to the −1 frame would result in translation of the RT-RH ORF. We previously suggested that this might represent a means by which DIRS elements could express the downstream YR ORF in such a way that the protein did not end up covalently attached to the RT-RH (Goodwin and Poulter 2001). The importance of separate RT-RH and YR proteins is suggested by the way in which the mature integrase and RT-RH proteins of canonical LTR retrotransposons are produced by cleavage of a polyprotein precursor. DIRS elements might not be able to cleave polyproteins, as none of these elements is known to encode a protease.

The relationship between the RT-RH and YR ORFs in TcDirs1 and related elements appears to be more similar to that in some members of the SDR subgroup, such as TOC3 of *Chlamydomonas reinhardtii* (Goodwin and Poulter 2004), than to other ITR elements. The finding that the YR ORF in the former elements does not overlap the first ORF indicates that translation of the YR cannot be achieved by frameshifting at the end of ORF1. The fact that the YR ORF shows a short overlap with the 3′ end of the RT-RH ORF instead suggests that translation might occur via a programmed ribosomal frameshift at the end of the RT-RH ORF. (The alternative possibilities—for example, that translation of the YR ORF is achieved either by splicing out of the RT-RH region or by internal initiation—seem less likely because with such mechanisms there is no apparent reason why the short overlap between the two ORFs would have to be conserved.) If this is the case, then the YR proteins in TcDirs1-like elements (and certain members of the SDR subgroup) would presumably be produced as translational fusions to the RT-RH proteins. This might not be a problem in these cases, however, because it is likely that the YR itself can tolerate a large N-terminal extension, as suggested by the long 5′ extensions of the YR ORFs found in other ITR elements. Similarly, because most programmed ribosomal frameshifts are relatively rare events, only a small proportion of the translation products of the RT-RH ORF would have the potentially inactivating YR domain attached to their C-termini.

Acknowledgements This work was carried out in compliance with the laws governing genetic experimentation in New Zealand and the United States of America. TJDG is supported by a NZ Science and Technology Post-Doctoral Fellowship (Contract No. UOOX0222). We thank the Exelixis Pharmaceutical Co. (S. San Francisco, Calif.) for BAC library preparation and screening, and for BAC sequencing, and the Human Genome Sequencing Center at the Baylor College of Medicine (Houston, Tex.) for provision of *S. purpuratus* and *A. mellifera* genome sequence data.

References

Adams MD, et al (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195

Beeman RW, Friesen KS (1999) Properties and natural occurrence of maternal-effect selfish genes (‘*Medea* ’ factors) in the red flour beetle,*Tribolium castaneum*. Heredity 82:529–534

Beeman RW, Johnson TR, Nanis SM (1986) Chromosome rearrangements in *Tribolium castaneum*. J Hered 77:451–456

Cappello J, Handelsman K, Lodish HF (1985) Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell 43:105–115

De Chastonay Y, Felder H, Link C, Aeby P, Tobler H, Muller F (1992) Unusual features of the retroid element PAT from the nematode *Panagrellus redivivus*. Nucleic Acids Res 20:1623–1628

Doolittle RF, Feng DF, Johnson MS, McClure MA (1989) Origins and evolutionary relationships of retroviruses. Quart Rev Biol 64:1–30

Duncan L, Bouckaert K, Yeh F, Kirk DL (2002) *kangaroo*, a mobile element from *Volvox carteri*, is a member of a newly recognised third class of retrotransposons. Genetics 162:1617–1630

Eickbush TH, Malik HS (2002) Origins and evolution of retrotransposons. In: Craig NL, Craigie R, Gellert M, Lambowitz AM (eds) Mobile DNA II. ASM Press, Washington, D.C., pp 1111–1144

Farabaugh PJ, Zhao H, Vimaladithan A (1993) A novel programed frameshift expresses the *POL3* gene of retrotransposon Ty3 of yeast: frameshifting without tRNA slippage. Cell 74:93–103

Fayet O, Ramond P, Polard P, Prere MF, Chandler M (1990) Functional similarities between retroviruses and the IS*3* family of bacterial insertion sequences? Mol Microbiol 4:1771–1777

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_-WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12:543–548

Genetics Computer Group (1994) Program manual for the Wisconsin package. Version 8. Genetics Computer Group, Madison, Wis.

Goodwin TJD, Poulter RTM (2001) The DIRS1 group of retrotransposons. Mol Biol Evol 18:2067–2082

Goodwin TJD, Poulter RTM (2004) A new group of tyrosine recombinase-encoding retrotransposons. Mol Biol Evol 21:746–759

Haliscak JP, Beeman RW (1983) Status of malathion resistance in five genera of beetles infesting farm-stored corn, wheat and oats in the United States. J Econ Entomol 76:717–722

Holt RA, et al (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298:129–149

Lorenzen MD, Berghammer AJ, Brown SJ, Denell RE, Klingler M, Beeman RW (2003) *piggyBac*-mediated germline transformation in the beetle *Tribolium castaneum*. Insect Mol Biol 12:433–440

Malik HS, Eickbush TH (2001) Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res 11:1187–1197

Nunes-Duby SE, Kwan HJ, Tirumalai RS, Ellenberger T, Landy A (1998) Similarities and differences among 105 members of the Int family of site-specific recombinases. Nucleic Acids Res 26:391–406

Ruiz-Perez VL, Murillo FJ, Torres-Martinez S (1996) *Prt1*, an unusual retrotransposon-like sequence in the fungus *Phycomyces blakesleeanus*. Mol Gen Genet 253:324–333

Sambrook J, Russell DW (2001) Molecular cloning: a laboratory manual (3rd edn). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Swofford DL (1998) PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer, Sunderland, Mass.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Van Duyne GD (2002) A structural review of tyrosine recombinase site-specific recombination. In: Craig NL, Craigie R, Gellert M, Lambowitz AM (eds) Mobile DNA II. ASM Press, Herndon, Virginia, pp 93–117

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353–3362