THE ACTIVITY AND EVOLUTION OF THE *DAPHNIA* DNA TRANSPOSON *POKEY*

A Thesis

Presented to

The Faculty of Graduate Studies

of

The University of Guelph

by

TYLER ADAM ELLIOTT

In partial fulfilment of requirements

for the degree of

Master of Science

January, 2011

Your file Votre référence
ISBN: 978-0-494-80087-4
Our file Notre référence
ISBN: 978-0-494-80087-4

ABSTRACT

THE ACTIVITY AND EVOLUTION OF THE *DAPHNIA* DNA TRANSPOSON *POKEY*

Tyler Adam Elliott
University of Guelph, 2011

Co Advisors:
Professor T. J. Crease
Professor T.R. Gregory

The DNA transposon *Pokey* from the freshwater microcrustacean *Daphnia pulex* is unique for its ability to insert into a highly conserved region of the 28S rDNA gene. My thesis consists of the extraction and characterization of 136 *Pokey* elements from the *Daphnia pulex* genome and the measurement of the excision rate of an artificially constructed non-autonomous *Pokey* element. Elements grouped into four clusters of two full length and two non-autonomous MITE groups and showed evidence of co-evolution with rDNA loci. An excision assay performed in yeast showed that the rate of excision of a non-autonomous *Pokey* element from a reporter plasmid (3×10^{-10}) is several orders of magnitude lower than most DNA transposons. This, combined with evidence derived from the elements analyzed from the sequenced genome, suggests that intra-genomic selection pressure from frequent periods of apomixis has shaped the co-evolutionary relationship between *Pokey* and its host organism.

ACKNOWLEDGEMENTS

This thesis represents over two years of hard work that could not have been accomplished without the help and support of a great many people. First and foremost I would like to thank Teri Crease for her patience, sage-like wisdom and always having an open door policy to talk about just about anything. I'll always be grateful for the opportunity she gave me to get my hands dirty and discover and play with my new-found intellectual love, transposable elements. Thanks are also due to Ryan Gregory for introducing me to evolutionary biology, giving me the chance to do some real research and helping me learn that if everything goes right the first time you do it, it's probably not science. Dr. Terry Van Raay was always able to offer sound advice and a much needed different perspective on my research that only strengthened it.

I would especially like to thank Dr. Robin Floyd for teaching me the ropes in the lab when I started that first summer. I'll never forget any of the tricks he taught me. Dr. Guojun Yang and Dr. Nathan Hancock provided both resources and technical expertise in carrying out and troubleshooting the yeast excision assay and without their help it would have been impossible. Dr. Yves Bigot first suggested to us that *Pokey* might have an intron and Dr. Deborah Stage provided the sequence of a *Pokey* MITE that she found which was instrumental in the analysis of elements from the genome.

The staff of the AAC Genomics Facility, Angela, Jing and Jeff, were always cheery and helpful with all the sequencing that needed to be done and never seemed to mind my persistent need to spin down my yeast cultures, even if it was very late in the

day. Thank-you so much.

I would also like to thank past and present members of the labs in room 1403 for their help, friendship and support: Nick Jeffery, João Lima, Ola Pierossi, Brent Saylor, Alison Forde, Shannon Eagle, Syed Omar, Chandler Andrews, Jillian Smith, Alex Ardila Garcia, Martin Brummell and past members of the Hebert and Lynn Labs. You all made working in one lab or another these past three years some of the best times I've ever had. Thanks.

I would like to thank my friends Dan, Matt, Bill, Brandon, Ryan and Shane for their friendship and for keeping me sane throughout this degree and for years beforehand. They're the greatest friends a guy could ask for and if I never make another friend until they day I die, I'll be happy.

I'd also like to thank everyone in my family, both old and new: Adrian, Shannon, Kirstin, Marissa, Rhianna, Derek, Oliver and Kevin. I love you all, even if I don't say it that often. Lastly, I owe a great deal of thanks to my parents Sonia O'Brien and Douglas Elliott. Without them I wouldn't be here, literally, and because they always encouraged my interest and love of science from a very early age. I dedicate this thesis to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Mobile DNA was first discovered by geneticist Barbara McClintock in the 1940's while she was observing the behaviour of x-ray induced broken chromosomes in maize (McClintock, 1946; 1947). McClintock observed that some changes in corn kernel phenotype were due to the excision and reinsertion of particular DNA segments she named Dissociation (*Ds*), controlled by an independent locus named Activator (*Ac*) (McClintock, 1950). Because of the highly visible and ontogenetically-timed effects these segments could have on the phenotype, she postulated that they were the key to development through gene expression modulation (McClintock, 1961). Thus she termed *Ac*/*Ds* and other loci like them 'controlling elements' due to their supposed role in the orchestration of development. A contemporary of McClintock, the corn geneticist R.A. Brink, observed similar behaviour with his Modulator of Pericarp (*Mp*) element system but felt that the functional connotations associated with the term 'controlling elements' were unfounded, leading him to coin the more neutral term, 'transposable elements' (Wood and Brink, 1956).

Transposable elements (TE) are a unique category of DNA that possesses the ability to mobilize and replicate themselves, sometimes to high copy numbers. This action has littered eukaryotic genomes with large amounts of this repetitive DNA, which prompted researchers to propose functional roles for it (Britten and Davidson, 1971; Cohen, 1976; Nevers and Saedler, 1977). Although this was not the sole viewpoint (Östergren, 1945; Peterson, 1970), this adaptationist tendency was the dominant one until the publication of two seminal papers in *Nature* in 1980 (Doolittle and Sapienza, 1980;

Orgel and Crick, 1980). The Selfish DNA hypothesis melded evolutionary theory with the ubiquity and seeming redundancy of TE sequences to propose that their ability to replicate was an adequate default explanation for their presence within genomes. Although this hypothesis did not preclude the possibility of them acquiring a secondary function for the host, the authors suggested that TEs and some other repetitive sequences are no more than molecular parasites inhabiting the genome of a host organism. This view still represents the best explanation for the presence, persistence and profuseness of most TEs. However, a more sophisticated and complex theory of how TEs and host genomes interact is necessary for total understanding (Kidwell and Lisch, 2001; Venner et al., 2009). The molecular and genomic revolutions have provided the tools and raw information required for the characterization and classification of TEs, which has increased substantially in the last 20 years. Despite a debate over the classification and naming of TEs (Wicker et al., 2007; Kapitonov and Jurka, 2008; Seberg and Petersen, 2009; Petersen and Seberg, 2009) all classification systems include two classes based on the presence or absence of an RNA intermediate during transposition (Finnegan, 1989).

Class I TEs

Class I consists of retrotransposons and is further subdivided into smaller groups based on the mode of insertion and the inclusion or exclusion of inverted or direct repeats on the 5' and 3' ends. Class I elements featuring long-terminal repeats (LTR) and an integrase-mediated system of insertion make up one major subdivision of retrotransposons (Havecker et al., 2004). LTR elements are similar to retroviruses in their manner of transposition and in their possession of intervening sequences coding for Gag,

protease, reverse transcriptase (RT), integrase and RNase H proteins. The TE is transcribed in the nucleus with the aid of an RNA polymerase II promoter within the 5' LTR, whereupon the mRNA transcript passes through the nuclear envelope and the encoded proteins are translated from the transcript in the cytoplasm (Voytas and Boeke, 2002; Sandmeyer et al., 2002). The Gag proteins form a virus-like particle (VLP) around the RT, RNase H, two RNA transcripts of the element and a cellular tRNA, which primes the reverse transcription of the mRNAs into cDNA copies (Havecker et al, 2004). Through the action of the integrase, the transcript is returned to the nucleus where it is integrated into a new genomic site at which a target site duplication of 4 to 5 base pairs (bp) occurs (Havecker et al., 2004). LTR elements are notable for having proliferated greatly in the genomes of plants (Wessler et al., 1995). A related subdivision known as DIRS-like elements transpose through similar means except their repeats can be either direct or inverted and integration is accomplished using a tyrosine recombinase rather than an integrase (Goodwin and Poulter, 2001; 2004).

The second major subdivision within Class I elements consists of elements that lack both inverted repeats and LTRs. They are named by exclusion, as in non-LTR elements or long interspersed nuclear elements (LINEs), and exhibit a different duplication and insertion mechanism than LTR elements (Han, 2010). Autonomous non-LTR elements consist of one or several open reading frames (ORFs) encoding proteins or domains with endonuclease (EN), RT, and other motifs, often flanked on their 3' end with a poly-A tail reminiscent of those found on processed mRNAs (Eickbush, 2002; Moran and Gilbert, 2002). Non-LTR TEs retrotranspose through a process known as

target-primed reverse transcription. In this mechanism, the TE is transcribed and the transcript is translated, after which the protein product of the second (or only) ORF binds to the transcript. The protein-mRNA complex returns to the nucleus, the EN domain of the protein cleaves one strand of the insertion site and uses an exposed 3' end of the sequence to prime reverse transcription of the non-LTR mRNA into the insertion site (Luan et al., 1993, Christensen and Eickbush, 2005). This mechanism is imperfect and copies that are truncated at the 5' end are often observed ("dead-on-arrival" or "DOA" copies), which is thought to be due to premature termination by the RT before synthesis of the second strand is complete (Eickbush, 2002). Non-LTR elements have proliferated greatly in many vertebrates, and are the dominant TEs in both the chicken and human genomes (International Human Genome Sequencing Consortium, 2001; International Chicken Genome Sequencing Consortium, 2004). In particular the element *Alu*, a short interspersed nuclear element (SINE) which uses the proteins of non-LTR retrotransposons for mobility, has reached a copy number of over 1 million in the human genome (International Human Genome Sequencing Consortium, 2001). The related *Penelope*-like elements appear to use a similar means of insertion but do in fact possess both inverted and direct repeats (Evgen'ev et al., 1997).

Class II TEs

Class II elements, known as DNA transposons, move via an element-encoded transposase protein which recognizes and removes the element, opens a new insertion site, and reintegrates the element back into the genome (Feschotte and Pritham, 2007). Typical eukaryotic DNA transposons consist of a transposase ORF flanked on either end

by inverted terminal repeats (ITRs). Most transposons are between 1kb and ~7kb in size, with ITRs of between 10 to 1000 bp (Feschotte and Pritham, 2007). The ITRs serve as recognition sites for the DNA binding domain(s) of the transposase and serve to direct the catalytic residues of the protein to cut the DNA outside the ITRs (Curcio and Derbyshire, 2003). Upon cutting a target site and re-inserting the element back into the genome, part of the host DNA is duplicated, which is known as a target site duplication (TSD). This can be from 2 to 10 bp and may or may not be symmetrically duplicated on both ends of the element (Feschotte and Pritham, 2007). If a DNA transposon excises from a locus its removal is usually not perfect, causing what is known as a footprint, characterized by insertions or deletions of nucleotides surrounding the excision site. Not all DNA transposons cause footprints when they excise though, notable examples being elements from the *piggyBac* and PIF/*Harbinger* superfamilies (Fraser, 2000; Hancock et al., 2010). The sequence similarity of the ITRs, transposase ORF and TSD combined are diagnostic features used to assign elements to particular superfamilies

This "cut and paste" method of transposition does not result in a change in the copy number of the element. However, the copy number of DNA transposons can increase through sister chromatid recombination, or if the element moves from a location that has already been replicated into an area that has yet to be replicated during mitosis or meiosis (Engels et al., 1990; Chen et al., 1992). Most currently known DNA elements do not have a specific target site, with some exceptions like *Pokey,* which belongs to the *piggyBac*-superfamily (Penton et al., 2002). Three other superfamilies of DNA transposons have been characterized that utilize quite distinct mechanisms for

duplication. *Helitrons*, or rolling-circle elements, are thought to use a rolling circle form of replication similar to the one used by some plasmids, accompanied by reintegration back into the genome (Kapitonov and Jurka, 2001). *Mavericks* are 15-25 kb DNA elements that encode 5-11 proteins and are thought to be able to self-synthesize new copies via protein-primed replication and reintegration, as do some viruses (Kapitonov and Jurka, 2006; Pritham et al., 2007). A third superfamily, *Cryptons*, are a little-studied group of DNA elements that lack repeats and contain a tyrosine recombinase ORF rather than a transposase (Goodwin et al., 2003).They are suspected to replicate via the formation of extrachromosomal, circular intermediates that are recombined back into the genome via the tyrosine recombinase.

TE Biology and Evolution

In general, TEs exist in one of two types of transposition-competent states determined by the presence or absence of the ORF's encoding the proteins necessary for duplication and transposition. Autonomous elements possess the protein-coding sequences necessary for mobility while non-autonomous ones do not. Because TE insertions are frequently selectively neutral or deleterious at the host level, insertions with ORFs rendered inactive by mutations tend to accumulate (Kaplan et al., 1985). Provided these insertions still possess the repeats necessary to recruit transposase, and/or the promoters necessary to produce mRNA transcripts for retroelements, they can use the proteins produced by autonomous elements to replicate and move to a new insertion site. Sometimes the loss of an active transposase for a particular TE and subsequent internal deletions leads to a different, but often more successful evolutionary trajectory. Miniature

inverted-repeat TEs, or MITEs, comprise a structure-based phenetic grouping of non-autonomous DNA transposons, characterized by small sizes, often between 100-700 bp. They often have high sequence identity to autonomous DNA transposon representatives only in the TSD and ITR regions (Feschotte et al., 2002). The restriction of high sequence identity between certain MITEs and autonomous DNA transposons to very limited regions of the element has led some to propose that MITEs can sometimes arise *de novo* from non-TE genomic DNA that happens to be flanked by repeats similar or identical to the ITRs of an active DNA transposon (Feschotte et al., 2002). MITEs have proliferated to a substantial degree in the genomes of many species, reaching copy numbers as high as the tens of thousands (Feschotte et al., 2002). Non-autonomous elements for both Class I and Class II elements tend to outnumber autonomous ones over time, which could in turn lead to the extinction of that particular element in the genome when the last autonomous insertion becomes inactive through mutation (Le Rouzic and Capy, 2006).

How do TEs avoid this spiral towards inactivation? One common hypothesis is through the process of horizontal transfer (HT), whereby an active element is transferred from one organism to another where it can proliferate anew until the mutational inactivation process occurs once again (Silva et al., 2004). Cases of HT have also been reported for LTR and non-LTR retrotransposons, with evidence that non-LTR element transfers are less frequent than those of LTR elements, although it has been argued that this may be an artefact of biased investigation (Schaack et al., 2010a). It has sometimes been argued that due to the less prolific replication process of DNA transposons, they should be more dependent on HT for long term survival over evolutionary time, with the

drosophilid *P* and *mariner* elements cited as examples (Lohe et al., 1995; Clark and Kidwell, 1997). However, the universality of this claim has been challenged by several examples involving Tc1/*mariner* elements from drosophilids and flowering plants (Lampe et al., 2001; Feschotte and Wessler, 2002). Lampe et al. (2001) proposed that multiple lineages of DNA transposons could persist within the same genome if ITR and transposase sequences diverged to a degree where only certain transposases could recognize specific ITR structures. Feschotte and Wessler (2002) characterized multiple lineages and sub-lineages of Tc1/*mariner* elements in 31 species of flowering plants that appeared to have been vertically propagated for approximately 25-50 million years. How often vertical diversification of DNA transposons within a host lineage occurs and is a successful strategy for their long-term survival is unknown.

Another aspect of host biology that is of particular importance for the understanding of TE dynamics is mode of reproduction. The presence of active TEs within an asexual lineage is thought to result in one of several outcomes; extinction of the lineage through excessive TE-induced mutation, eventual elimination of active TEs through selection acting at the host level, or domestication of TEs for the host's benefit (Nuzhdin and Petrov, 2003). Conversely, sex is thought to not only increase the power of selection within the host to remove TEs, but to also furnish the elements with a pathway to colonize new genomes within the same species, and thus acts as a double-edged sword (Hickey, 1982; Wright and Finnegan, 2001). Sexual reproduction also allows for possibly deleterious ectopic recombination to occur between non-allelic TE insertions, resulting in damaging chromosomal rearrangements (Mieczkowski et al., 2006; Delprat et al., 2009).

TE Contributions to Eukaryotic Genome Evolution

One of the simplest and most significant contributions of TEs to eukaryotic genomes is their influence on the total quantity of DNA in the genome (i.e., genome size) (Gregory, 2001). In general, it has been shown that eukaryote genome size correlates positively with the quantity of TEs present in that species (Kidwell, 2002; Lynch and Conery, 2003; Gregory, 2005a). For example, humans have a genome size of 3.423 Gb and approximately 45% of that DNA is composed of recognizable TE sequences. In contrast, the fruit fly, *Drosophila melanogaster* has a genome size of 176 Mb and only ~5% of its DNA is composed of TE sequences (Gregory, 2005a, 2010; Quesneville et al., 2005). Genome size correlates positively with cell size and negatively with metabolic and cell division rate, which are important parameters at both the cellular and whole-organism level (Gregory, 2005b).

Element insertion is largely undirected, but TEs tend to be found in the more densely packed, gene poor heterochromatic regions of the genome where recombination rates are low. Whether this is a case of some target site selection or the elimination of insertions in high-recombination regions that affect host fitness, or a combination of both forces, is not well understood (Rizzon et al., 2002; Simons et al., 2007; Gao et al., 2008). However, some elements do show very specific target site preferences, which may be advantageous to the elements. These sites allow them to insert specifically in regions where they will be less disruptive to the transcription and regulation of host genes and give them a better chance at continued propagation (Craig, 1997; Bushman, 2003; Winckler et al., 2005). Insertions that do occur within or near genes tend to be disruptive

and there are many examples of element insertions changing protein coding sequences and regulatory regions, or causing splicing errors and a variety of other mutations that can lead to dysfunction and disease (Boisinnot et al., 2006; Belancio et al., 2008; Kim et al., 2007; Deragon and Capy, 2007). It is this potential for harmful mutations that may have provided the selection in favour of the evolution of molecular countermeasures against TEs and other selfish genetic elements, such as silencing mechanisms involving epigenetic modification and RNA interference (Yoder et al., 1997; Jordan and Miller, 2009).

A TE insertion might be harmless initially but the presence of multiple, homologous element sequences at different locations in the genome can lead to ectopic recombination. Homologous TEs from different insertion sites can pair and recombine, possibly causing deletions or duplications of flanking sequences (Hedges and Deininger, 2007). The insertion and excision mechanisms of transposable elements can create double-strand breaks in genomic DNA as well (Izsvák et al., 2009). If not successfully repaired, these can cause the loss of large segments of chromosome or the ligation of large fragments to other chromosomes, causing a potentially lethal translocation and/or inversion (Hedges and Deininger, 2007). The mutagenic potential of TEs also means that they can create variation upon which selection can act and TE-derived sequences have been found in protein coding exons (Lorenc and Makalowski, 2003; Piriyapongsa et al., 2007; Polavarapu et al., 2008) and regulatory sequences (van de Lagemaat et al., 2003; Jordan et al., 2003). Sometimes large portions or entire TE protein-encoding sequences have been co-opted by the host genome for its use via a process known as exaptation or

molecular domestication (Gould and Vrba, 1982; Miller et al., 1992). Numerous examples exist, including one of the proteins responsible for antibody production in vertebrates, RAG1, which is composed of sequences derived from the transposases of *Transib* and *Chapaev* elements (Kapitonov and Jurka, 2005; Panchin and Moroz, 2008).

The co-evolutionary interaction between TEs and the host genome has also created situations which are more complex. In several species lacking telomerase activity, the ends of chromosomes are maintained by site-specific non-LTR retrotransposons, such as *Het-A* in drosophilids and *Zepp* in the algae *Chlorella vulgaris* (Higashiyama et al , 1997; Pardue and DeBaryshe, 1999). In these situations, both the elements and the host benefit from transposition, resulting in more of a mutually exploitative relationship rather than a host-parasite one (Kidwell and Lisch, 2001). The position of TEs and TE-derived sequences in highly heterochromatized regions of the genome has led some to suggest that they play an essential functional role in X-chromosome inactivation in mammals (Avner and Heard, 2001; Abrusán et al., 2008), and centromere formation in a wide array of taxa (Wong and Choo, 2004).

Insertions in Ribosomal RNA Genes

Ribosomes consist of ribosomal RNAs (rRNAs) transcribed from rDNA loci that are present from one to thousands of copies per haploid genome (Long and Dawid, 1980). The rDNA units occur in tandem arrays and consist of a transcription unit composed of an external transcribed spacer (ETS), the 18S rRNA gene, an internal transcribed spacer (ITS), the 5.8S rRNA gene, a second ITS, and the 28S rRNA gene (Eickbush and Eickbush, 2007) separated by an intergenic spacer (IGS). Individual rRNAs are spliced

from the primary transcript and fold into the configuration necessary for their incorporation into the subunits of the ribosome (Eickbush and Eickbush, 2007). Since ribosome synthesis is essential for survival, rRNA genes display a high level of sequence conservation between different species – certain locations within the genes are nearly identical across very distantly related taxa (Hillis and Dixon, 1991; Ganley and Kobayashi, 2007). Other regions under less functional constraint are free to diverge and can be quite different when compared between species. Certain regions of rDNA units are maintained with such high conservation due to the homogenizing force of selection acting on the arrays, resulting in concerted evolution (Liao, 1999). Homogeneity can also be achieved by drift and selection acting at a level below the organism through a process known as molecular drive (Dover, 1982). Concerted evolution occurs through mechanisms such as unequal crossing over and gene conversion, between sister chromatids and homologous chromosomes, which can drive selectively favourable variants to spread through arrays.

Given these constraints, one might expect that insertions into rRNA genes by TEs of any kind would not be tolerated. However, the high copy number and sequence conservation of rDNA appears to have been exploited by multiple forms of selfish DNA across the tree of life (Figure 1). Organisms as diverse as Archaea, ciliates, and slime moulds have been found to harbour mobile introns inserting into a much conserved region of the 28S rRNA gene (Kruger et al., 1982; Muscarella and Vogt, 1989; Kjems and Garrett, 1991). Insertions into this same region were discovered in *Drosophila melanogaster* when unusually long 28S rRNA gene sequences were observed (Long and

Dawid, 1979). The extra sequences were identified as insertions by the TEs R1 and R2, which are 5.3 and 3.5 kb non-LTR retrotransposons that insert approximately 74 bp apart, and are part of a larger group of rDNA-specific retrotransposons (Xiong and Eickbush, 1988; Kojima and Fujiwara, 2003; 2004). R2 elements are found in arthropods, echinoderms, chordates and platyhelminthes and reconstruction of the phylogenetic relationships between elements from different species suggests that they evolved shortly after animals and fungi diverged from a common ancestor (Kojima and Fujiwara, 2005; Burke et al., 1998). Phylogenetic analysis of R2 RT domains from a variety of species showed that R2 elements in different arthropod species can be quite diverged from one another. However, phylogenetic trees of these same R2 sequences show that the branching order matches that of the host species, suggesting that R2 elements have been stably vertically inherited for hundreds of millions of years (Burke et al., 1998). It is speculated that R2 elements remain in their host cell by simply transposing to one of the other numerous, conserved 28S rRNA insertion sites that exist in the genome (Eickbush, 2002). This is a delicate balance: if the elements insert into too many 28S rRNA genes, the cell will not be able to produce functional ribosomes and will die. However, if they do not transpose often enough they will be eliminated from the genome by recombination and host-level selection acting on the rDNA. Most eukaryotes have more rDNA copies then are necessary for survival so there is an ample genomic niche in which R2 and other rDNA-inserting TEs can exist.

Daphnia *and* Pokey

The genus *Daphnia* consists of three subgenera (*Daphnia, Hyalodaphnia,* and *Ctenodaphnia*) of obligately or cyclically parthenogenetic, filter-feeding microcrustaceans native to freshwater habitats across the globe (Hebert, 1978). During the spring and summer, and during favourable conditions, *Daphnia* reproduce via direct developing apomictic eggs that develop into females. When winter approaches, or environmental conditions become inhospitable, females of cyclically parthenogenetic populations produce diploid, male offspring as well as haploid eggs which must be fertilized by the sperm from males. These fertilized eggs are protected by a hard case called an ephippium and remain dormant over the winter, after which they hatch as females in the spring and continue the cycle (Hebert, 1978). Some populations in the same species may also be obligate parthenogens and will never revert to a sexual mode of reproduction. In this case, the ephippial eggs are also produced apomictially. This system of alternating sexual and apomictic reproduction has been of interest to researchers studying it from ecological and evolutionary standpoints, and much more recently for its effect on TE dynamics (Sullender and Crease, 2001; Valizadeh and Crease, 2008; Schaack et al., 2010b).

In 1993, Barry Sullender discovered an insertion in 10% of the 28S rRNA genes screened from a phage DNA library of *Daphnia pulex*, in the same conserved region where previous forms of mobile DNA were found (Sullender, 1993). Further analysis showed this was not a retrotransposon, but a 7.2 kb, putative DNA transposon with 16 bp ITRs and a four bp TTAA TSD on either end (Sullender, 1993). This element, which was

named *Pokey,* was also found in the closely related species, *Daphnia pulicaria* as 5 kb and 6.6 kb insertions in the 28S rRNA gene (Figure 2, Penton et al., 2002). These elements contain a ~1.5 kb ORF, which, when compared to other transposases, is most similar to those from the *piggyBac* superfamily (Penton et al., 2002). More extensive comparison between *piggyBac* transposases, including *Pokey*, revealed the presence of a possible DDD amino acid motif, a catalytic triad known to be essential for transposition, which was later confirmed empirically to be necessary for *piggyBac* function (Keith et al., 2008a). Y. Bigot (pers. comm.) has suggested that the published *Pokey* transposase sequence might be incomplete as he identified what appears to be a 68 bp intron in the transposase gene, which contains what was originally thought to be the stop codon for the transposase ORF (Penton et al. 2002). Whether the 68 bp sequence is actually an intron requires validation.

The *D. pulicaria* elements sequenced by Penton et al. (2002) also contain repeats in their 5' end that are homologous to part of the IGS region downstream of the 28S rRNA gene (Penton et al., 2002). *Pokey* elements have also been found outside the 28S rRNA gene in various genomic locations and some elements might even display preference for a target site other than TTAA (Sullender and Crease, 2001; Valizadeh and Crease, 2008). These genomic *Pokey* elements were found to have similar copy numbers in both cyclical and obligate strains of *D. pulex* but there were fewer insertion variants in obligates, consistent with theory concerning TEs and mode of reproduction (Sullender and Crease, 2001). Because obligate parthenogens lack recombination they cannot remove *Pokey* insertions as easily as cyclical parthenogens, and would be expected to

accumulate more insertions. However, if selection acts more strongly against obligate clones with more *Pokey* insertions than both obligate and cyclical clones with less insertions, only those obligate clones with a lower *Pokey* copy number would be expected to survive and reproduce (Sullender and Crease, 2001). Penton and Crease (2004) extracted partial *Pokey* elements from 14 species in the subgenus *Daphnia* and found multiple, distinct lineages in *D. obtusa* named *Pokey*A and *Pokey*B. A phylogeny constructed in the same study showed strong evidence for strict vertical inheritance of *Pokey* where the branching order of *Pokey* elements mirrored that of their host species. This is similar to the relationship found between R2 elements and their hosts (Burke et al., 1998). However, a phylogeny constructed of *piggyBac* elements from the silk moth *Bombyx mori* showed one element grouping closely with the *D. pulicaria* 6.6 kb *Pokey* element (Xu et al., 2006). The authors suggested a horizontal transfer event might be the cause, although the direction is unknown.

Sullender and Crease (2001) observed high frequencies of *Pokey* insertion site heterozygosity in some *D. pulex* populations. Because one would expect insertions to be either fixed or lost by drift, Sullender and Crease (2001) suggested that this high frequency may have been caused by recent transposition of *Pokey* elements. The presence of both a complete promoter upstream and a polyadenylation signal sequence downstream of the ORF, accompanied by the intact DDD motif, suggested that the elements characterized by Penton et al. (2002) could be active. More recently, data from *D. pulex* mutation accumulation lines, looking at the effect of breeding system on TE copy number, point to possible somatic transposition events occurring for several DNA

transposon families including *Pokey*, providing more indirect evidence for activity (Schaack et al., 2010c).

Goals of the Present Study

Despite what is known about *Pokey*, numerous questions remain unanswered. Research goals in this thesis were organized around two, overarching questions:

1) What is the evolutionary history, copy number and overall diversity of the *Pokey* elements in the *Daphnia pulex* genome? All current knowledge concerning *Pokey* stems from the two elements extracted from the *D. pulicaria* genome and partial sequences amplified from several other species in the *D. pulex* species complex (Penton et al., 2002; Penton and Crease, 2004). The great sequence and structural diversity found in these few elements suggests that the population of *Pokey* elements in the *D. pulex* genome could be quite heterogeneous but this has yet to be investigated. With this in mind, I extracted and characterized intact *Pokey* elements from the recently sequenced genome of *Daphnia pulex* to determine the copy number of autonomous and non-autonomous elements, and their structural and sequence diversity. In addition, I extracted the sequence flanking *Pokey* insertions sites to determine what, if any, preference *Pokey* has for insertion sites. I also determined whether or not the sequence identified as an intron within the *Pokey* transposase is functionally relevant.
2) Is *Pokey* an active TE? Indirect evidence indicates *Pokey* is active (Penton and Crease, 2002; Schaack et al., 2010c), however its relative transposition rate is not known. I investigated this using a yeast-based assay to characterize the excision

rate of the transposase encoded by the 6.6 kb element from *D. pulicaria*, and a non-autonomous derivative of that element.

METHODS

Characterization and analysis of *Pokey* elements within the *Daphnia pulex* genome

Recovery and annotation of Pokey *elements from the* D. pulex *genome*

I recovered *Pokey* elements from the *D. pulex* genome by querying the database maintained by the US Department of Energy Joint Genome Institute. The 6.6 kb element from *D. pulicaria* (AY115589.1) was used as query for BLAST-algorithm mediated searches under default settings. In addition, a small (~750 bp) non-autonomous *Pokey* element identified by Deborah Stage (pers. comm.) was also used as a query. Contigs, and occasionally whole scaffolds, containing highly significant hits were downloaded and elements were manually annotated, aided by cross-referencing with putative *Pokey*-location data taken from the supplementary material of Schaack et al. (2010b). When structurally novel or divergent elements were discovered, they were used as new queries in subsequent BLAST searches.

I identified ORFs using a combination of the ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/) and the ExPASy Proteomics Server Translate (http://www.expasy.ch/tools/dna.html) tools and sequence similarity to the 5 kb and 6.6 kb *D. pulicaria* transposase coding sequences. Promoter sequences were inferred using the Neural Network Promoter Prediction program (http://www.fruitfly.org/seq_tools/promoter.html). Polyadenylation signals were predicted using the Webgene portal (http://zeus2.itb.cnr.it/~webgene/wwwHC_polya.html\). The amino acid sequence of the

transposase from the 6.6 kb element was analyzed using the PSORTII suite of protein analysis programs (http://psort.hgc.jp/) to determine what, if any protein localization motifs are present. Each element was assigned an identifier as follows: "[s+scaffold #] – [bp position where element starts] –[length of insertion]". Concensus sequences for each cluster of elements were generated by BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html\) using alignments of all elements lacking uncalled bases, followed by manual adjustment.

Phylogenetic analysis of Pokey *sequences*

I aligned *Pokey* sequences using a combination of the CLUSTAL, MUSCLE and MAFFT multiple sequence alignment programs available from the EMBL-EBI website (http://www.ebi.ac.uk/Tools/sequence.html). Alignments were then manually adjusted in the program BioEdit. Only sequences with less than 5% un-called bases across the aligned region were used in phylogenetic analyses. Measurements of pairwise sequence divergence were calculated using the Kimura 2-parameter (Kimura, 1980) method in MEGA4 (Tamura et al., 2007). Neighbor-joining (NJ) trees (Saitou and Nei 1987) were also constructed in MEGA4. Bootstrap analysis was performed on 500 pseudo-replicates for each tree (Felsenstein, 1985). The dataset of full-length elements excluded the highly variable repeat region between the 5' ITR and the region upstream of the transposase coding region. A dataset including the last ~1600 bp of the 3' end of *Pokey* elements from various species within the subgenus *Daphnia* (Table 1) was aligned with the *Pokey* elements from the *D. pulex* genome sequence and used to generate NJ trees. A global dataset of all elements from the genome sequence was generated by aligning the 5' and 3'

terminal regions and removing all intervening un-alignable sequence in between. Full alignments can be found in Appendix 2.

Insertion site analysis of genomic elements

Pokey elements having intact ITRs as well as the same tetranucleotide TSD at both ends of the insertion were selected, both from the genome sequence and other sources, for consensus site analysis. These criteria were used to avoid selecting older, more mutated insertion events where any signal of a preferred concensus site surrounding the element may have been degraded. In total 118 insertion sites were selected. Sixteen bp on either side of the tetranucleotide TSD were extracted and the complete 36 bp sequence of each insertion site was analyzed using the WebLogo program (Crooks et al., 2004). Sixteen bp was chosen because a previous study that analyzed *piggyBac* insertion sites in human cell lines failed to find any consensus in 10 bp on either side of the element (Wilson et al., 2007). Five insertion sites came from transposon display data from a previous study (Valizadeh and Crease, 2008). Only the 3' flanking sequence was available so I used the *D. pulex* genome sequence to find the sequence upstream of the element insertion. One insertion was extracted from a *D. pulex* 37609 bp genomic DNA sequence from clone JGIAZSN-5P10 reported in GenBank (AC167692.2).

Cloning, PCR and sequencing

DNA Amplification

Polymerase chain reaction (PCR) was used to amplify all DNA sequences in this study. Primers are described in Table 2. To quickly amplify cloned fragments from *E. coli* colonies, colony PCR was employed. Individual colonies were re-suspended in 10 µL of water and heated to 95 °C for 2 minutes before standard PCR was performed on 1 µL of the lysate. Unless otherwise noted, all standard reactions were performed in 25 µL of PCR mix containing 1 unit of *Taq* polymerase (New England Biolabs), 1 mM dNTPs, 2.5 µL of 10X ThermoPol Buffer (100 mM KCl [pH 7.4], 0.1 mM EDTA, 1mM dithiothreitol, 0.5% Tween 20, 0.5% NP-40 and 50% glycerol), 0.1 µL of each 10 µM primer and 10-25 ng of template DNA.

When required, high-fidelity PCR was also performed using either Phusion Hot Start *Taq* (New England Biolabs) or Platinum *Taq* (Invitrogen) kits. Unless otherwise noted all Phusion reactions were performed in 25 µL of PCR mix containing 1 unit of Phusion *Taq*, 1mM dNTPs, 5 µL of 5X Phusion HF Buffer (7.4 mM $MgCl_2$), 0.1 µL of each 10 µM primer and 10-25 ng of template DNA. All Platinum reactions were performed in 25 µL of PCR mix containing 1 unit of Platinum *Taq*, 1mM dNTPs, 2.5 µL of 10X Platinum *Taq* Buffer, 0.1 µL each of 10 µM primer, 1.5-2.0 µL of 50 mM $MgCl_2$ and 10-25 ng of template DNA.

Both standard and high-fidelity PCR reactions were carried out on a PTC-100 thermal cycler (MJ Research Inc.) under the following conditions unless otherwise noted:

1) Two minutes and 30 seconds for dsDNA melting at 94 °C

2) 30 seconds for primer annealing at 55 °C

3) One minute for polymerase extension at 72 °C

4) Steps 1-3 were carried out a further 34 more times

5) Five minutes at 72 °C for a final extension step

Electrophoretic Gels

DNA fragments were run on 0.8%-1% agarose gels in TAE buffer and visualized using the Gel Red (Biotium Inc.) nucleotide specific stain and UV light.

DNA Concentration

The concentration of all DNA was determined using a Nanodrop 8000 Spectrophotometer (Thermo Scientific).

Sequencing

DNA was sequenced using the Big Dye Terminator v3.1 Kit (Applied Biosystems Inc.). Reactions were carried out in 12 μL volumes containing 0.3 μL of Dye terminator, 1 μL of 5X Sequencing Buffer, 1 μL of 10 μM primer and 10-50 ng of template DNA. Sequences were resolved on an ABI 3730 Automated Sequencer.

E. coli *Transformation*

Between 16 μL and 25 μL of chemically competent *E.coli* cells were thawed briefly on ice and transformed with 50-200 ng of the relevant plasmid. Cells were incubated on ice for 20 minutes and then heat shocked at 42 °C for 45 seconds to 2 minutes depending on the strain used. Cells were allowed to recover on ice for 5 minutes and then added to 250-400 μL of Terrific Broth (TB) liquid culture or SOC liquid medium (Invitrogen) and shaken at 200 rpm at a temperature of 37 °C for 1 hour. Whole volumes or aliquots were then pipetted onto selective Luria Broth (LB) agar plates and spread using sterile silicate beads. Plates where then placed at 37 °C for incubation.

E. coli *Plasmid DNA Extraction*

E. coli colonies containing desired plasmids were grown overnight, at 37 °C with shaking at 200 rpm, in 5 mL of LB or TB liquid media containing an appropriate selective drug. Plasmid DNA was extracted from *E. coli* using the Roche Diagnostics High Pure Plasmid Isolation Kit (Roche) and standard manufacturer's protocols. Plasmid DNA was re-suspended in 25-50 μL of water.

Isolation of the transposase coding region

Validation of putative intron

Reverse-transcriptase PCR was performed on RNA samples extracted from sexual *D. pulex* isolates to determine if *Pokey* transposase transcripts were present and to provide further evidence for the presence of an intron in the transposase ORF. RNA was extracted from the parthenogenetic offspring of sexual *D. pulex* isolates Disp 325, Can 2-

18 and Can 3-57 donated by Dr. M. Cristescu of the University of Windsor. Samples were stored in RNA Later (Qiagen) at -20 °C and RNA was extracted using the RNAqueous-4PCR kit (Ambion) and standard manufacturer's protocols. Absence of DNA contamination was verified using standard PCR and *Pokey* ORF primers (Table 2). Reverse transcription of the RNA used the SuperScript III One-Step RT-PCR with Platinum *Taq* Kit (Invitrogen) and standard manufacturer's protocols. Primers used were Pok4065F with Pok4488R and Pok5026F with Pok5985R (Table 2).

Cloning the transposase ORF

I used PCR to amplify the transposase of the *D. pulicaria* 6.6 kb element using the Pc17 plasmid DNA as template. Pc17 contains the nearly full-length 6.6 kb element extracted from a single *D. pulicaria* clonal line from Humboldt Lake, Saskatchewan (Penton et al., 2002). Primers PokattB1ORFF and PokORFDelR (Table 2) were used in a high-fidelity PCR reaction to amplify exon 1 of the transposase, with double the standard volume of dNTPs due to the large size of the fragment. Standard thermal cycler conditions were also used with a two minute extension time to produce a fragment 1449 bp long. In the reaction to amplify exon 2, primers PokORFDelF and PokattB2ORFR (Table 2) were used under the same conditions as those used for exon 1, except a one minute extension time was used. This produced a fragment 615 bp long.

Removal of the putative 68 bp intron found within the 6.6 kb *Pokey* transposase was done using overlap extension PCR (Figure 3, Lee et al., 2004). This method makes use of chimeric primers to amplify sections of a single template in a primary PCR reaction whose products are then used as template in a secondary reaction to produce a

final product composed of the two sections joined together. The PokORFDelF primer sequence contains 20 bp upstream of the intron and 19 bp downstream of the intron. The reverse primer contains the reverse complement of that sequence. Secondary PCR was then performed using 1 μL of a 10 fold dilution of each exon fragment as template, and the PokattB1ORFF and PokattB2ORFR primers under standard PCR conditions with double the standard volume of dNTPs and an extension time of two minutes. This yielded a 2064 bp fragment which was subsequently blunt-end cloned into pSC-B-amp (Stratagene) to produce pSC-Pok6.6 ORF. Ligation of the two exons and complete removal of the intron was confirmed via sequencing using several *Pokey* ORF-specific primers (Pok4065F, Pok4410F, Pok5026F, Pok4488R and Pok5985R).

Quantifying DNA Transposon Activity

Inter-plasmid excision or transposition assays were first developed to measure the activity of the Lepidopteran transposon *piggyBac* and to determine exactly what is required for its successful excision (Fraser et al., 1995; Elick et al., 1996). Measuring the excision rate of a DNA transposon is most efficient when the two components of transposition, an ITR-bearing sequence and the transposase, are cloned into separate plasmids (Elick et al., 1996). A transposase ORF under the control of a galactose-inducible promoter is cloned into an expression vector while a complementary non-autonomous DNA transposon is cloned into the phosphoribosylaminoimidazole carboxylase (*ade2*) gene, which is important for synthesis of the nucleobase adenine, in the donor plasmid pWL89A (Weil and Kunze, 2000; Yang et al., 2006). Yeast is

transformed with the two plasmids and grown on media selective for their presence. Transposition is then induced by growing the transformed cells in media lacking adenine and providing galactose as the only carbon source. The reversion rate on media lacking adenine is compared to the viable cell count to calculate an excision rate for the element in the system. The unique *Hpa*I restriction site in the *ade2* gene of pWL89A is ideal for cloning a non-autonomous *Pokey* element because it contains the TTAA target site. By adding an AA dinucleotide to the 5' end and a TT dinucleotides to the 3' end of the non-autonomous element, the TTAA TSD characteristic of *piggyBac* superfamily elements is created after ligation into the plasmid (Fraser, 2000; Sarkar et al., 2003).

Construction of pAG413-Pok6.6 ORF expression plasmid

I amplified the full-length transposase sequence from pSC-Pok6.6 ORF plasmid DNA using standard PCR, the PokattB1F and PokattB2R primers and double the standard volume of dNTPs under standard conditions with a two minute extension time. This ORF sequence with flanking att sequences was then used in a BP Gateway cloning reaction (Appendix 1.1) with 150 ng of pSC-Pok6.6 ORF and 150 ng of pDONR-221 to create the pDONR-Pok6.6 ORF entry clone. pDONR-221 was graciously provided by Dr. A. Walhout of the University of Massachusetts Medical School. The entry clone was transformed into *E. coli*, extracted and sequenced to ensure that the fidelity of the ORF was maintained. Because a non-proofreading *Taq* polymerase was used to amplify the full-length ORF for use in the BP reaction, many clones containing a variety of mutations in the ORF were obtained. The clone selected for use in the excision assay contained a synonymous mutation at amino acid position 389 in the ORF. An LR clonase reaction

was performed using 150 ng of pDONR-Pok6.6 ORF and 150 ng of the destination vector pAG413GAL-ccdB (Addgene ID# 14141), which was obtained from the Addgene plasmid repository (http://www.addgene.org/pgvec1). This produced the expression plasmid pAG413-Pok6.6 ORF which was transformed into *E. coli*, extracted and sequenced.

Amplification and construction of a non-autonomous Pokey *element*

An overlap extension PCR approach was also used to create a non-autonomous *Pokey* element from the 6.6 kb *D. pulicaria* element for use in the excision assay. Sequence from the 5' and 3' ends of the element was amplified and ligated together using this approach to create an artificial non-autonomous *Pokey* element containing both ITRs with intervening sequence derived from the first 356 bp and last 426 bp of the 6.6 kb element. The 5' half was amplified from *D. pulex* genomic DNA extracted from an individual from a local pond population (Wellington County, Ontario) using primers 28S2974F and PokMITE5'R in a standard PCR reaction, which yielded a 438 bp fragment. The 3' half was amplified from the same template and reaction conditions using primers PokMITE3'F and 28S3282R to yield a product 631 bp long. Secondary PCR was then performed using 1 μL of a 10 fold dilution of each fragment as template and the PokMITE5'F and Pok MITE3'BR primers in a high-fidelity PCR reaction under standard conditions. This produced a 787 bp fragment which was then blunt-end cloned into pSC-B-amp to produce pSC-Pok6.6NA. The clone was subsequently sequenced to confirm that successful ligation had occurred and the sequence was nearly identical to that of the 5' and 3' ends of the 6.6 kb element.

Construction of pWL89A-Pok6.6NA donor plasmid

To construct the donor plasmid for the excision assay, I digested plasmid pWL89A with *Hpa*I, to create a blunt-end cut in the middle of a single GTTAAC site within the *ade2* gene (Appendix 1.2). pSC-Pok6.6NA was digested with *Kpn*I and *Bam*HI to free a fragment containing the non-autonomous element sequence and to prevent the re-ligation of pSC-Pok6.6NA in downstream reactions. The digested plasmid was then used as template in a high-fidelity PCR reaction using primer Pok MITE5' F and Pok MITE3' B R and standard conditions. These primers added the AA and TT dinucleotides to the ends of the fragment to generate the TTAA TSD after ligation. The ligation reaction (Appendix 1.3) contained 500 ng of this PCR product and 250 ng of digested pWL89A to create pWL89A-Pok6.6NA, 50 ng of which were transformed into *E. coli*, extracted and sequenced using primers Ade2F and Ade2R, which flank the *Hpa*I site in pWL89A.

Yeast excision assay

The excision assay was performed in the haploid *Saccharomyces cerevisiae* strain DG2523 (*MATα, ura3-167, trp1-hisG, leu2-hisG, his3-del200, ade2-hisG*) provided by Dr. G. Yang of the University of Toronto. First, 50 μL of chemically competent DG2523 yeast cells (Appendix 1.4) were transformed with 100-150 ng each of pAG413-Pok6.6 ORF and pWL89A-Pok6.6NA and plated on selective yeast media lacking histidine and uracil (SD-U-H, Appendix 1.4, Figure 4). Cells transformed with pAG413-GAL (no insert) and pWL89A-Pok6.6NA were used as a negative control. Plates were incubated at 30 °C for 3-5 days until pink colonies appeared. Single colonies were then picked into

7 mL of SD-U-H liquid media and incubated at 30 °C for 66-72 hours with shaking at 200 rpm to grow cells to saturation. Cultures were then centrifuged for 1 minute at 7000G to pellet the cells which were then washed in 5-6 mL of sterile water to remove any glucose from the liquid media, which inhibits galactose induction. Cells were re-suspended in 400 µL of sterile water of which 390 µL was plated onto SD-Ade+2% galactose+1% raffinose media to select for excision of the non-autonomous *Pokey* element by adenine reversion (inducing plates). Both 10^4 and 10^5 fold dilutions were made from the remaining 10 µL of cells by adding them to 90 µL of water and repeating until the desired level of dilution was achieved. The 100 µL of diluted cell suspension was then plated on SD-U-H plates to obtain a viable cell count (counting plates). Both inducing and counting plates were incubated at 30 °C for 3-21 days until colonies appeared. The number of colonies on both the inducing and counting plates was used to calculate an excision frequency of the non-autonomous *Pokey*. For example, if 1500 colonies grew on the 10^4 fold dilution counting plate, then 5.85×10^8 viable cells (1500 x 10 000 x 39 = 585 000 000) were plated on the inducing plate. If 10 colonies grew on this plate then the rate of excision is $10/ 5.85 \times 10^8 = 1.17 \times 10^{-8}$ excision events/cell.

Analysis of Pok6.6NA element excision from donor plasmid

White revertant colonies were re-streaked onto SD-U-H media to grow cells to sufficient quantities for downstream manipulation. Plasmid DNA was extracted from revertants using the Zymoprep II Yeast Plasmid Miniprep Kit (Zymo Research) and standard manufacturer's protocols. The sequence spanning the *Hpa*I site in pPok6.6NA was amplified using Ade2F and R primers with a 65 °C annealing temperature under

standard PCR conditions. This produced ~263 bp fragments which were sequenced using the aforementioned Ade2 primers and compared to the homologous sequence in pWL89A to characterize the effects of excision.

RESULTS

Characterization and analysis of *Pokey* elements in the *Daphnia pulex* genome

Recovery and annotation of Pokey *elements from the* Daphnia pulex *genome*

A total of 75 full length (1402-9800 bp) and 61 MITE (612-1018 bp) *Pokey* elements were retrieved and annotated from the *D. pulex* genome sequence based on the high sequence similarity to the 6.6 kb element described by Penton et al. (2002) and the presence of intact ITR sequences. Elements were classified as either full length or MITE based upon the presence of absence of a transposase ORF, or the remains thereof. No sequence derived from the *Pokey* transposase coding region could be annotated with any confidence in the MITEs. Combined, the full length and MITE elements comprise 431,344 bp of the genome sequence, nearly two-fold larger than what was originally reported by Schaack et al. (2010b). An additional 84 'incomplete' sequences were found ranging from 400 - 4400 bp long. These were identified as highly significant hits in BLAST-mediated searches, but the entire element, from 5' ITR to 3'ITR, could not be recovered.

Full length elements

The full length elements, with an average size of 5009 bp, comprise those containing nearly complete or degenerated transposase ORFs. Many elements contained sizeable insertions and deletions, resulting in elements larger than 9500 bp. Although much of the size difference is due to variation in the 5' repetitive region, described in a later section, various unique insertions and deletions were found in other positions in

several elements. For example, the 1196 bp insertion in element s82-18392-6388 was found to be a non-autonomous DNA transposon belonging to an unknown family, and was not annotated in a recent paper on the Class II elements of *D. pulex* (Schaack et al., 2010b).

I constructed a NJ tree of the full-length elements that revealed two clusters, with the previously characterized 6.6 and 5kb elements falling into the same cluster (Figure 5). In addition, full length elements characterized from other *Daphnia* species also grouped with these elements. In accordance with the grouping of *Pokey* elements by Penton and Crease (2004), this cluster was designated *Pokey*A. Average sequence divergence between *Pokey*A elements from the genome sequence was calculated to be 6.8% while average divergence between elements in the second cluster is 5.0% (Table 3.). Average divergence between the two clusters is 40.8%.

A second group of elements, termed *Pokey*B, was previously identified by Penton and Crease (2004) in *Daphnia obtusa* based on analysis of the 3' terminal 1600 bp and was found to be highly divergent (>50%) from *Pokey*A elements. To ascertain whether the second cluster of elements from the genome sequence was the same as the *Pokey*B elements previously identified in *D. obtusa*, a second NJ tree was constructed from the 1600 bp dataset of Penton and Crease (2004) supplemented with data from the newly characterized elements from the *D. pulex* genome (Figure 6). The overall topology of the tree is similar to that of the tree in Figure 5 and the NJ tree generated by Penton and Crease (2004) with all *Pokey*A elements clustering together with high support. The *Pokey*B elements from *D. obtusa* group with the second cluster of elements from the *D.*

pulex genome sequence with high bootstrap support. Therefore this second *D. pulex* cluster was designated as *Pokey*B.

Intron analysis

Sequencing of RT-PCR products, performed on RNA extracted from sexual *D. pulex* isolates as template, showed that the intron sequence is spliced out of *Pokey* ORF mRNA transcripts (Figure 7). Introns were identified in *Pokey*A elements based on sequence similarity to the intron identified in the 6.6 kb element. Introns were identified in *PokeyB* elements via multiple sequence alignments between *Pokey*A and *Pokey*B. Evidence that the homologous sequence in *PokeyB* elements is actually an intron was provided by GENSCAN (http://genes.mit.edu/GENSCAN.html) analysis. *Pokey*A introns range from 68-74 bp long, with most differences due to the expansion of a poly-thymine region in the 3' half of the intron. *PokeyB* introns are larger on average, ranging from 79-84 bp long. The presence of this intron extends the coding region of the *D. pulicaria* 6.6 kb element by 582 bp which codes for an additional 194 amino acid residues.

Transposase coding regions

I identified transposase genes, or the easily recognizable remains thereof, in 23 of the 75 elements from the *Pokey*A and *Pokey*B clusters. These were aligned to each other, and to the ORFs of the *D. pulicaria* 5kb, 6.6kb elements as well as several elements cloned from other species in the subgenus *Daphnia* (Table 1). Fewer complete transposases were identified in *PokeyB* elements due to a greater preponderance of un-called bases in the members of this cluster. Based on comparisons with the amino acid

sequence of the transposase from the 6.6kb and 5kb elements during the recovery and annotation process, I found that four elements from the genome sequence, two *Pokey*A and two *Pokey*B, contain complete ORFs which could theoretically code for a functional transposase (Figure 8). Based on data from transcription start site predictions (Table 4.) and sequence comparison, it appears that *PokeyB* elements possess ORFs with start codons both up and downstream of the putative start site of *Pokey*A elements. The longest contiguous *Pokey*A and *Pokey*B ORFs possess the three conserved motifs making up the putative DDD catalytic amino acid triad of *piggyBac* elements (Figure 9).

PSORTII analysis revealed the presence of a putative nuclear localization signal (NLS) in the second exon, which contains a PGAKRRK amino acid motif and is well conserved across both *Pokey*A and *Pokey*B ORFs. A second putative NLS (PIRKIRP) was also identified in exon 1, upstream of the catalytic motifs, but this was only found in *Pokey*A ORFs. Both *Pokey*A and *PokeyB* proteins were classified as nuclear, with >60% probability, by the *k*-nearest neighbour classifier sub-program in PSORTII, which is consistent with their proposed function as transposases. Alignment between *Pokey*A and *Pokey*B transposases to other *piggyBac* transposases (Figure 9) shows the conservation of two CXXC motifs in the C-termini of each protein; a feature that was first seen in the *piggyBac* multiple sequence alignment of Keith et al. (2008a). This C-rich region is thought to be a RING-type zinc-finger, and is thought to be important in protein-protein interactions (Keith et al., 2008a). These motifs, along with the putative NLS in exon 2, are also well conserved in the partial *Pokey* ORFs recovered from other species in the subgenus *Daphnia* (Figure 10). One difference between the *Pokey* ORF and other

piggyBac elements is the lack of conservation of a fourth D residue two amino acids downstream of the third D in the proposed catalytic triad. The *Pokey* transposase has an N residue at this position, which is conserved not only across all *Pokey*A elements so far analyzed from the subgenus *Daphnia*, but also across *Pokey*B elements as well.

Repeats upstream of the transposase

The 5' end of both the 5 and 6.6 kb *Pokey* elements from *D. pulicaria* possess repeats derived from the IGS region of the rDNA repeat unit (Penton et al., 2002). These repeats, named A1, A2 and A3, are all derived from a unique sequence of ~220 bp with levels of sequence divergence ranging from 5.5-8.7%. The spacing of the repeats between the two elements differs, with the 5 kb element possessing a 1701 bp deletion between the A1 and A2 repeats relative to the 6.6 kb element, suggesting that the presence of multiple repeats in this region is very unstable (Penton et al., 2002). In addition, a second sequence derived from the IGS, B, was found between the A1 and A2 repeats of the 6.6 kb element in the 1701 bp insertion (Penton et al., 2002). To investigate variation in this 5' repetitive region in other *Pokey* elements, BLAST searches were performed using whole element sequences to determine the possible genomic origin of any sequence tracts found within elements. The three highly significant hits (E values: 0.0-9.0 x 10^{-10}) that appeared most often came from (1) a 199 bp region within the IGS of D. *tenebrosa* (EU595562.1), which is similar to the sequence identified by Penton et al. (2002); (2) a 49 bp sequence upstream from a *D. pulicaria* microsatellite marker (AY619422.1) and (3) a 47 bp sequence from the ITS of *D. galeata* (HM161704.1). The sequence matching the microsatellite marker was also found between the A2 and A3 repeats of the 5 and 6.6

kb elements. Divergence between the *Pokey* repeats and the unique IGS sequence varied from 1-10% and from 0-2% for the sequence similar to that near the microsatellite marker, henceforth known as C. C is 6% divergent from the BLAST hit from the *D. galeata* ITS2 sequence and was commonly found either between the A2 and A3 repeats or both immediately upstream and downstream of one of them. Order of appearance (5'to 3') and copy number of A1, A2, A3 and C repeats varied quite dramatically with many elements possessing unique configurations of A and C repeats. One element was found to have no A repeats whatsoever. A1 repeats were common in *Pokey*A elements but could only be identified in four of the *Pokey*B elements. One element (s257-24384-8894) has an insert of several thousand bases between an A3 and A2 repeat, some of which is highly similar to several sequence tracts from a clone of *D. pulex* genomic DNA deposited in GenBank (AC167683.2).

Pokey *MITEs*

As was previously stated, sequence similarity between full length elements and MITEs was restricted to the first and last several hundred bp of the elements. Sequence similarity between the MITE clusters was determined from the first ~160 bp and last ~360 bp of each element as well as two internal regions of 37 bp and 162 bp. The small size and the fact that sequence similarity between the small and full length elements was restricted to several hundred bp downstream and upstream of the 5' and 3' ITRs, respectively led to the labelling of these small elements as *Pokey* MITEs. These MITEs, with an average size of 760 bp, also appear to be composed of three well separated lineages (MITE1, MITE2 and MITE3) based on the NJ tree in Figure 11. Representative

MITE1 and MITE2 sequences were used as queries in BLAST-mediated searches to identify repeats similar to those found in the 5' ends of the large elements, but none were found. Both clusters exhibit low average intra-group pairwise divergence with values of 2.2% for MITE1 and 3.2% for MITE2 (Table 3). Four other elements also clustered with MITE1 albeit with low bootstrap support. When these are included in divergence estimates, the intra-group divergence of MITE1 increases by 1%, but the between group divergence remains unchanged. Divergence between groups is much higher than within at 24.9%, but smaller than divergence between the *Pokey*A and *PokeyB* elements. Average divergence between MITEs and full length elements is much higher than divergence between MITEs, ranging from 43.9-64.5%. MITE1 elements are most similar to *PokeyB* elements while MITE2 elements are most similar to *Pokey*A elements. Closer inspection of MITE3 sequences revealed they may not represent a distinct cluster. These elements do not share a common ITR structure, nor do they share one with any other *Pokey* elements. MITE3 elements also have numerous insertions not found in MITE1 and MITE2 elements, and intra-group divergence for MITE3 elements was found to be 20.4%, which is higher than any other intra-group divergence for any cluster. MITE3 elements are least divergent from MITE2 elements, 20.0%, and most likely represent older, more mutated copies of this group that may not be able to transpose.

Shared Pokey *features*

The global alignment of all elements recapitulated previous results with all elements clustering into the aforementioned groups, excluding the 11 previously mentioned elements, which technically belong to the MITE1 cluster (Figure 12).

Alignment of all the elements also revealed two distinct ITR structures (Figure 13). The original imperfect 16 bp ITR (ITR#1) characterized by Penton and Crease (2002) is found in the *Pokey*A and MITE2 elements. A second ITR structure (ITR#2) is also imperfect but only 12 bp long and is found in the *Pokey*B and MITE1 elements. Although the difference between the two structures is arguably minor (4 single nucleotide substitutions and reduction in size of the complementary region between the 5' and 3' ITR sequence), it is notable that the elements with ITR#2 are found in higher copy number than their ITR#1 relatives; 26 ITR#1 elements were annotated while the remaining 110 possess ITR#2. Despite these differences, there are features shared by all lineages of *Pokey* elements characterized (Figure 14). Alignment of consensus sequences that were generated for each cluster reveal that the only regions of extensive conservation shared by all four clusters are the first ~130 bp and the terminal ~430 bp of each element; regions where interaction between the element and transposase would most likely occur during transposition. Within these regions, several poly-adenine and poly-thymine tracts were of particular note.

Insertion site analysis of genomic elements

WebLogo analysis did not reveal a strong preference for any particular insertion sequence (Figure 15). The highest preference lies within the TSD tetranucleotide itself, although there is more flexibility in *Pokey* elements than has been seen in other *piggyBac* elements (Wilson et al., 2007; Wang et al., 2008; Wang et al., 2010). Most of the TSDs found conform to the typical *piggyBac* site, however over 10% of insertions possessed a limited range of non-typical TSDs (Table 5). This is consistent with previous data about

TSDs characterized from TE display data taken from cyclical and obligate isolates of *D. pulex* (Valizadeh and Crease, 2008). Bases occurring at high frequency at positions 13, 16, 21, 26 and 29-32 on the figure were the same as those found in the 28S rRNA target site where *Pokey* is known to insert.

Yeast excision assay

The measurement of excision rate was performed ten times using separate transformed colonies each time. An equal number of control colonies were also used. The rate of excision was calculated to be 3.0×10^{-10} excision events/ yeast cell (Standard error: 2.134×10^{-10}) with no *Ade*2 reversion in the control experiments. This provides the first direct evidence that the *Pokey* transposase from the 6.6 kb element is active and can in fact mobilize a non-autonomous derivative, albeit at a low rate. Plasmid DNA was extracted from the three revertants that were recovered and the regions flanking the *Hpa*I restriction site was amplified and sequenced. Of the three, only one revertant showed perfect excision, whereby the original insertion site was restored, while the remaining two left footprints (Figure 16). The sample size (n=3) is not large enough to comment on the relative frequency of perfect excision events.

DISCUSSION

Pokey *diversity in the* D. pulex *genome*

Extraction and analysis of more than 130 *Pokey* elements from the *D. pulex* genome revealed four well supported clusters of elements (Figure 12). Large elements, with an average size of ~5000 bp, form two clusters, termed *Pokey*A and *Pokey*B based on previous work (Penton and Crease, 2004). Both clusters include autonomous members, which appear to possess a gene encoding a functional transposase, as well as numerous non-autonomous elements. In addition, two clusters of MITEs were found, each bearing a unique ITR structure which corresponded to a particular cluster of full length *Pokey* elements. The fact that annotated elements bearing ITR#2 (*Pokey*B and MITE1) outnumber those with ITR#1 over 4:1 suggests that this ITR could be better at attracting and/or binding transposase proteins and thus give it a transposition advantage over ITR#1 elements. Whether this is true or if the ITR#2 elements possess another motif that is more selectively favourable for transposition would require empirical investigation. For example, excision and electrophoretic mobility-shift assays (EMSA) could be used to measure relative excision frequency and the strength of binding between ITR variants and transposases, respectively. The differences in sequence between the two ITR structures are arguably minor, but Casteret et al. (2009) demonstrated that a small number of single nucleotide changes to the ITR of the drosophilid DNA transposon, *Mos1* produce significant increases in transposition rate. Assays to measure the mobilization of *Stowaway* MITEs by *Osmar* transposases from rice showed that some elements had excision rates 30 fold higher than that of non-autonomous *Osmar* elements

(Yang et al., 2009). Whether or not the situation is as dramatic concerning *Pokey* MITEs could be investigated using excision assays.

Pokey *transposases*

Analysis of the *Pokey* transposase revealed the presence of two putative nuclear localization signals (NLS), only one of which is conserved between A and B elements, as well as two CXXC motifs in the C-terminus. The C-terminal NLS is in a homologous position to the one identified in *piggyBac* by Keith et al. (2008b), but whether the one identified in *Pokey* is functional would require testing. The CXXC motifs are features conserved across all *piggyBac*-superfamily elements and have been postulated to be a RING-type zinc finger important for protein-protein interactions (Keith et al., 2008a), or part of a chromatin-interacting PHD domain (Mitra et al., 2008). This domain was found to be unnecessary for transposition *in vitro* by Mitra et al. (2008) in excision assays with *piggyBac*, but the authors postulated that transposition *in vivo* might require transposase-chromatin interactions for success. This putative zinc-finger/PHD domain is conserved across *Pokey* elements suggesting its presence is just as critical for *Pokey* function as it is for *piggyBac*. If this domain is a RING-finger, it could be integral for transposase interaction with other *Daphnia* proteins, or with other *Pokey* transposases for the formation of multimers, which are known to be important in the transposition process of most DNA transposons (Butler et al., 2006; Hickman et al., 2010). Whatever its function, it could be investigated using site-directed mutagenesis or targeted deletions in the C-terminus to observe their effects on excision frequency or even their ability to bind to other *D. pulex* proteins from a cDNA library in a yeast 2-hybrid system.

Pokey is found in two distinct, divergent clusters in the *D. pulex* genome and limited evidence suggests that this is also the case in *D. obtusa* (Penton and Crease, 2004).This type of vertical diversification of TEs within the same genome could be driven by drift, selection or more likely, a combination of the two. For example, Lampe et al. (2001) observed a lack of interaction between the ITRs and transposases of Tc1/*mariner* elements from different subfamilies with sequence divergence greater than 16%. They postulated that silencing mechanisms based on sequence similarity might drive the transposase and ITR sequences of related DNA transposons to diverge to escape silencing. If multiple copies of the same element are nearly identical in sequence, then RNA molecules produced by one element could be processed by the RNAi machinery and used to silence all of the elements. Thus, intragenomic selection could favour the sequence divergence of elements over time to overcome this silencing, and could also result in lack of cross-mobilization between divergent insertions as a side-effect. A second possibility is the presence of numerous non-autonomous elements within a DNA element family driving the divergence of their transposase and ITR sequences (Feschotte and Pritham, 2007). The ability of numerous non-autonomous members to titrate the transposase away from autonomous members could have fitness consequences for the transposase producers, so much so that intra-genomic selection might favour divergent elements that can only recognize themselves. *Pokey*A and *PokeyB* clusters are 40.8% divergent, which means they may not be able to cross-mobilize if their relationship is similar to that between divergent Tc1/*mariners*. The fact that each *Pokey* cluster also has a distinct ITR structure suggests that cross-mobilization is unlikely. This could be investigated using the yeast excision assay, possibly supplemented with yeast 1-hybrid or

EMSA to determine the strength of interaction between the transposases and ITRs of each group. The specificity of the transposases for particular MITE clusters could also be tested this way to determine if *Pokey*A transposase can only mobilize MITE2 elements, and *PokeyB* transposases only MITE1 elements, and whether transposase binding specificity is stronger for MITEs than for autonomous elements.

The 5' region

The origin of the repeats in the 5' region of *Pokey* was first proposed by Penton et al. (2002) to have been mediated by recombination between *Pokey* and the rDNA IGS. The presence of the C repeat and its similarity to ITS2 sequence suggests that further recombination occurred between *Pokey* and other regions of the rDNA. The instability of these repeats was first suggested by their differential spacing in the 5 kb and 6.6 kb elements from *D. pulicaria*. Data from the *D. pulex* genome sequence further support this as certain combinations of A and C repeats are unique to single or small groups of elements. One possible mechanism for this instability might be the same one postulated to have created the varied repeat structures of the IGS in different lineages of *D. pulex*, namely unequal crossing over (Crease, 1995). During recombination, *Pokey* insertions pair with one another but this pairing could be disrupted in the 5' repeats, sometimes resulting in the unequal exchange of sequence creating elements with duplications of a particular repeat structure. Other configurations appear to be the consequence of internal recombination between repeats within the same element, resulting in the loss of the intervening sequence, as suggested for the 5 kb element (Penton et al., 2002). Sequence exchange in this 5' repetitive region does not appear to be limited to other *Pokey*

elements and rDNA, as indicated by the ~3600 bp of DNA acquired by element s257-24384-8894, ~1100 bp of which is clearly derived from a unique region on another scaffold in the *D. pulex* genome. TEs acquiring DNA derived from the host is not uncommon for both Class I and Class II TEs. Both *Helitron* elements and non-LTR retrotransposons have been shown to acquire host DNA downstream of insertion sites as a consequence of aberrant mobilization mechanisms (Moran and Gilbert, 2002; Lal et al., 2009). Using evidence from the excision process of *Ds* elements, Langer et al. (2007) proposed that DNA transposons could acquire host sequence if the transposase slides after binding but before cutting, or if cryptic ITR-like sequences exist downstream of an element in the genome. Because of the multiple origins and complexity of arrangements of the repeats carried by *Pokey,* neither of these two mechanisms seems very plausible, and recombination represents the most likely explanation.

What is the significance, if any, of these repetitive sequences? The unique region in the IGS from which the A repeats are derived is several hundred bp upstream of the putative transcription start site of each rDNA unit (Crease, 1993; Penton et al., 2002). In mammals, the IGS is known to bind the protein Upstream-Binding Factor (UBF), which not only recruits other proteins to initiate polymerase I transcription of the rRNA genes, but seems to bind non-specifically to the entire unit, including both genes and spacers, and prevents silencing of the unit by blocking epigenetic modifications (Sanij and Hannan, 2009). The sequence to which UBF binds in the IGS is not known, nor is it known whether or not UBF homologs exist in *D. pulex* or other non-vertebrate eukaryotic lineages. It is possible that the A repeats upstream of the *Pokey* transposase gene recruit

some protein or complex of proteins that aid in the transcription of *Pokey*, or act as a defence against silencing mechanisms. Conversely, the repeats could be used to attract silencing machinery to prevent the *Pokey* element from transcribing and transposing too promiscuously. This would need to be investigated empirically, perhaps by identifying *D. pulex* proteins with the ability to bind to these repeats and investigating their function. Conversely, the variability in configuration, and even presence/absence polymorphism of A repeats may be nothing more than the consequence of aberrant recombination, and has no fitness impact on *Pokey* elements.

Pokey *MITEs*

The copy number of MITE1 and MITE2 when compared to MITEs from other genomes is quite different, with some MITEs reaching copy numbers into the tens of thousands in several plant genomes (Feschotte et al., 2002). As well, *Pokey* MITEs appear to be atypical based on their relatively large size of ~750 bp as compared to other MITEs which can be as small as ~130 bp (Bureau and Wessler, 1992). One explanation for this disparity between *Pokey* MITEs and other MITEs might lie in how MITEs are thought to arise. One mechanism is through progressive internal deletion of autonomous DNA transposons and subsequent selection for better and better transposition among the resultant elements over time (Feschotte et al., 2002). If MITEs have only evolved recently in *Pokey,* this may explain their larger size relative to MITEs in other organisms. Recent evolution may also explain the low copy number of MITE1 and MITE2, although the added constraint of being in the much smaller genome of *D. pulex* as compared to some plant species might also be a contributing factor. The smaller, more gene-dense

genome of *D. pulex* might make it very difficult for MITEs to proliferate to high copy numbers without having very serious, deleterious mutagenic effects.

The low intra-group sequence divergence between MITEs within the two major clusters suggests a history of recent transposition (Figure 11). Whether or not MITE elements possessing a particular type of ITR can only be mobilized by transposases produced by elements with the same ITR is unknown. Each MITE cluster is more similar to full length elements with which they share an ITR structure than they are to the other full length cluster (Table 3), suggesting that MITEs may have evolved independently twice. In contrast, the tree generated from the global alignment shows both MITE groups clustering with *Pokey*B, to the exclusion of *Pokey*A elements (Figure 12). However, the clustering of MITEs and *Pokey*B has no bootstrap support and when a tree was constructed using the complete deletion option, *Pokey*A and *Pokey*B clustered with no support to the exclusion of the MITEs (data not shown). If the evolution of MITEs was independent, both clusters have converged on internal sequence features shared between them, which might be crucial for their ability to transpose or attract transposases. Yang et al. (2009) demonstrated that rice *Stowaway* MITEs are mobilized by *Osmar* transposases, and that the MITEs themselves possess sequence motifs that appear to make excision more efficient than that of the *Osmar* elements themselves. Establishing whether or not these shared internal *Pokey* MITE sequences are important for transposition could be investigated using excision assays.

Asexuality, intra-genomic selection and Pokey *evolution*

The target site specificity observed for *Pokey* is a departure from the situation in *piggyBac,* which appears to have little to no specificity whatsoever (Wilson et al., 2007; Wang et al., 2008; Wang et al., 2010). The similarity of *Pokey*'s consensus genomic insertion site to the 28S rRNA target site suggests that it may have evolved a preference for this site or sites similar to it. Insertion into rDNA is likely to be favourable for the element and possibly less deleterious for the host, as it provides many identical insertion sites for the elements to exploit and because the host organism can still produce functional ribosomes provided that there is a sufficient number of un-inserted rDNA copies. This insertion site propensity, combined with the presence of rDNA-related repeats in the 5' end of large elements suggest that *Pokey* has diverged in genomic lifestyle quite noticeably from that of its presumably non-site-specific *piggyBac* ancestors. The unique breeding system of *Daphnia*, involving extended periods of apomictic reproduction, and the complete loss of sexuality in some lineages, in theory could have put strong selection pressure on ancestral *Pokey* elements to avoid causing deleterious mutations in their host, while still maintaining a transposition rate high enough to survive. The theory describing the interaction of TEs with asexual or partially asexual hosts predicts three possible outcomes: i) active elements are lost, ii) the host goes extinct due to TE-induced mutation, or iii) the elements become domesticated and the threat is neutralized (Nuzhdin and Petrov, 2003). Valizadeh and Crease (2008) and Schaack et al. (2010b) found evidence from natural isolates and mutation accumulation lines, respectively that obligately asexual *D. pulex* have lower TE loads than cyclical

parthenogens, consistent with the first prediction. However, Lockton and Gaut (2010) observed no significant differences in copy number between outcrossing *Arabidopsis lyrata* and selfing *A. thaliana*, although selfing and outcrossing both involve meiosis and recombination while apomixis does not. The proposed outcomes mentioned above tend to ignore a commonly overlooked fact; that variation exists within a population of elements within a genome, possibly imbuing it with the ability to respond to selection pressure caused by a host with an asexual lifestyle. Intra-genomic selection could favour element variants that are less deleterious to the host, perhaps those which have a slight target site preference for a highly conserved, multi-copy gene family such as the 28S rRNA gene. The fact that one of two suspected active TEs in the genome of the ancient asexual rotifer *Adineta vaga* is an rDNA-specific retrotransposon, R9, lends credence to this (Gladyshev and Arkhipova, 2009). These variants might have spread to other 28S rRNA genes via transposition or recombination between rDNA loci. If this recombination was sometimes aberrant, *Pokey* elements could have acquired DNA from other loci in the genome. The frequent recombination between rDNA loci may have facilitated a particular *Pokey* insertion acquiring sequence from the IGS, which may then in turn have given that element an advantage over copies without IGS repeats. It also cannot be ruled out that a proto-*Pokey* element acquired sequence from rDNA which then facilitated a higher probability of rDNA insertion. Overall, this scenario suggests a fourth option, that of a move towards stable coexistence between element and host like that seen between R1 and R2 elements and their respective hosts. Despite being from separate classes, *Pokey* and R elements share several common features aside from a shared target site, namely the presence of multiple lineages within the same genome and strong evidence for vertical

transmission (Burke et al., 1998; Burke et al., 2003; Gentile et al., 200l; Penton and Crease, 2004). These shared traits between a DNA transposon and non-LTR retrotransposons suggest that once specificity for rDNA evolves it drives the evolution of other features which might be inevitable consequences of long-term interaction with rDNA.

Another feature of the *Pokey* transposase is the fixation of an N residue rather than a D downstream of the third and final D of the catalytic triad. Keith et al. (2008a) determined that this fourth D residue is important in transposition by replacing it with an uncharged N and observing a reduction in excision frequency of three orders of magnitude from 10^{-2} to 10^{-5}. Similarly, the excision frequency measured for *Pokey* using the yeast-based assay was quite low in comparison to other DNA transposons, on the order of 10^{-10}. Using the same assay, Weil and Kunze (2000) measured an excision rate of 6.3-37.0 x 10^{-6} for the maize *hAT*-superfamily element *Ac/Ds*. Yang et al. (2006) recovered a rate of 1.51 x10^{-6} for the rice Tc1/*mariner* element *Osmar5* and Hancock et al. (2010) measured the excision rate of the PIF/*Harbinger* mobilized MITE *mPing* to be 6.0-17.0 x 10^{-6}. It would be interesting to mutate this conserved N back to a D in the *Pokey* transposase and observe whether or not this increases the rate of excision, and to what degree. The fixation in *Pokey* elements of an amino acid that is known to decrease the excision rate compared to *piggyBac* elements suggests that this change was favoured by intra-genomic selection, either before or after the target site specificity of *Pokey* evolved. In comparison, the transposition rates of R1 and R2 elements in *Drosophila melanogaster* have been calculated to be six orders of magnitude higher than the one

observed for *Pokey* (Pérez-González and Eickbush, 2002). That the rate is lower for *Pokey* seems odd but there are several key differences between the two systems which might reconcile this. For one, the creation of a new insertion during retrotransposition is in theory higher than that of DNA transposition due to the inherent replicative nature of retrotransposons. Secondly, non-LTR elements like R1 and R2 cannot excise from a location once inserted and new insertions always destroy the ability of that particular rDNA locus to produce a functional rRNA transcript (Eickbush et al., 2000). On the other hand, *Pokey* is a DNA transposon, and can excise itself from one site and move to another without causing a change in copy number. Unlike R1 or R2, *Pokey* exists at sites outside of rDNA in *D. pulex* which can act as either a source or sink population for element movement between rDNA and other sites in the genome. As well, Glass et al. (2008) observed that *Pokey*-inserted rDNA repeats can sometimes increase in frequency due to unequal crossing over, a means of copy number increase independent of transposition. Combined, these factors suggest that the lower rate observed for *Pokey* might be sufficient, or even necessary, for its continued existence within the genome of *D. pulex.*

CONCLUDING REMARKS

My thesis work used two approaches aimed at achieving a better understanding of the evolution and persistence of the rDNA-specific DNA transposon *Pokey* in the genome of *Daphnia pulex.* First, I extracted and characterized the diversity of *Pokey* elements in the genome sequence of *D. pulex.* A total of 136 complete elements were annotated, only four of which appear able to encode a functional transposase. I found that full length elements, with complete or mutated transposase coding regions, belonged to two divergent groups, *Pokey*A and *Pokey*B. The 5' sub-termini of both groups are highly variable due to the presence of rDNA-derived repeats, with few elements sharing the same configuration. I found that the genomic insertion sites of *Pokey* elements are weakly similar to the conserved region of the 28S rRNA gene into which *Pokey* inserts. Two clusters of non-autonomous *Pokey* MITEs were also identified that appear to have arisen independently from each group of full length elements. The low intra-group sequence divergence among MITEs suggests a history of recent transposition, which requires the presence of a working transposase. The second approach I used was to measure the excision rate of a non-autonomous derivative of the *D. pulicaria* 6.6 kb element by its transposase. Although the excision rate was found to be quite low at 3.0×10^{-10}/cell, this assay confirms that the transposase of at least one *Pokey* element is functional. Moreover, the protein was found to have an amino acid substitution previously found to substantially decrease the rate of activity in the related element, *piggyBac* (Keith et al., 2008a). Data generated from both the excision assay and the genome sequence analysis suggest a pathway for the evolution of *Pokey* whereby

selection pressure on the elements from a host with a cyclically apomictic lifestyle combined with a slight preference for rDNA opened up a new co-evolutionary relationship between element and host.

The results of my work suggest several possibilities for future research. For example, the putatively functional transposase in *Pokey*B elements could be used in excision assays to determine if they are more or less effective at mobilizing a non-autonomous *Pokey* element than the one used in this study. The excision rates of both MITE1 and MITE2 elements could also be measured to determine whether cross-mobilization is possible between elements bearing ITR1 and ITR2 sequences. This may provide insight into the existence of two lineages of *Pokey* and what, if any, selective forces may have contributed to their divergence. As well, the conserved N residue of the *Pokey* transposase could be mutated back to a D to determine its impact on excision frequency. If this change substantially increases the excision rate, as expected, it suggests that the low rate of *Pokey* excision is adaptive, possibly to a host which undergoes frequent periods of apomixis. It will also be useful to search for TEs in the rDNA of other genome sequences to see if the traits possessed by rDNA-specific elements are shared by all elements which have evolved to exploit this niche. I suspect that the current paucity of rDNA-specific DNA transposons is due more to a lack of investigation rather than their lack of existence. In conclusion, *Pokey* may represent an excellent case of co-evolution between host and element where element-level characteristics are not only important but vital in understanding the evolutionary history of a transposable element. All too often variation and traits at the level of the elements themselves is ignored or neglected in the

study of how TEs evolve, which is a denial of their status as units of evolution as important to understand as organisms, populations and species.

REFERENCES

Abrusán, G., J. Giordano, and P.E. Warburton. 2008. Analysis of transposon interruptions suggests selection for L1 elements on the X chromosome. *PLoS Genetics* 4: e1000172

Avner, P. and E. Heard. 2001. X-chromosome inactivation: counting, choice and initiation. *Nature Reviews Genetics* 2: 59-67

Belancio, V.P., D.J. Hedges, and P. Deininger. 2008. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Research* 18: 343-358

Boissinot, S., J. Davis, A. Entezam, D. Petrov, and A.V. Furano. 2006. Fitness cost of LINE-1 (L1) activity in humans. *Proceedings of the National Academy of Sciences of the United States of America* 103: 9590-9594

Britten, R.J. and E.H. Davidson. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *The Quarterly Review of Biology* 46: 111-138

Bureau, T.E. and S.R. Wessler. 1992. *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. *The Plant Cell* 4: 1283-1294

Burke, W.D., H.S. Malik, W.C. Lathe III, and T.H. Eickbush. 1998. Are retrotransposons long-term hitchhikers? *Nature* 392: 141-142

Burke, W.D., D. Singh, and T.H. Eickbush. 2003. R5 retrotransposons insert into a family of infrequently transcribed 28S rRNA genes of planaria. *Molecular Biology and Evolution* 20: 1260-1270

Bushman, F.D. 2003. Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell* 115: 135-138

Butler, M.G., S.A. Chakraborty, and D.J. Lampe. 2006. The N-terminus of *Himar1 mariner* transposase mediates multiple activities during transposition. *Genetica* 127: 351-366

Casteret, S., N. Chbab, J. Cambefort, C. Auge-Gouillou, Y. Bigot, and F. Rouleux-Bonnin. 2009. Physical properties of DNA components affecting the transposition efficiency of the *mariner Mos1* element. *Molecular Genetics and Genomics* 282: 531-546

Chen, J., I.M. Greenblatt, and S.L. Dellaporta. 1992. Molecular analysis of *Ac* transposition and DNA replication. *Genetics* 130: 665-676

Christensen, S.M. and T.H. Eickbush. 2005. R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Molecular and Cellular Biology* 25: 6617-6628

Clark, J.B. and M.G. Kidwell. 1997. A phylogenetic perspective on *P* transposable element evolution in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 94: 11428-11433

Cohen, S.N. 1976. Transposable genetic elements and plasmid evolution. *Nature* 263: 731-738

Craig, N.L. 1997. Target site selection in transposition. *Annual Review of Biochemistry* 66: 437-474

Crease, T.J. 1993. Sequence of the intergenic spacer between the 28S and 18S rRNA-encoding genes of the crustacean, *Daphnia pulex*. *Gene* 134: 245-249

Crease, T.J. 1995. Ribosomal DNA evolution at the population level: nucleotide variation in intergenic spacer arrays of *Daphnia pulex*. *Genetics* 141: 1327-1337

Crooks, G.E., G. Hon, J.-M. Chandonia, and S.E. Brenner. 2004. WebLogo: A sequence logo generator. *Genome Research* 14: 1188-1190

Curcio, M.J. and K.M. Derbyshire. 2003. The outs and ins of transposition: from *Mu* to *Kangaroo*. *Nature Reviews Molecular Cell Biology* 4: 1-13

Delprat, A., B. Negre, M. Puig, and A. Ruiz. 2009. The transposon *Galileo* generates natural chromosomal inversions in Drosophila by ectopic recombination. *PLoS One* 4: e7883

Deragon, J.M. and P. Capy. 2000. Impact of transposable elements on the human genome. *Annals of Medicine* 32: 264-273

Doolittle, W.F. and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284: 601-603

Dover, G. 1982. Molecular drive: a cohesive mode of species evolution. *Nature* 299: 111-117

Eickbush, D.G., D.D. Luan, and T.H. Eickbush. 2000. Integration of *Bombyx mori* R2 sequence into the ribosomal RNA genes of *Drosophila melanogaster*. *Molecular and Cellular Biology* 21: 213-223

Eickbush, T.H. 2002. R2 and related site-specific non-long terminal repeat retrotransposons. In *Mobile DNA II* (eds. N.L. Craig R. Craigie M. Gellert, and A.M. Lambowitz), pp. 813-835. American Society of Microbiology Press, Washington, DC, USA

Eickbush, T.H. and D.G. Eickbush. 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* 175: 477-485

Elick, T.A., C.A. Bauser, and M.J. Fraser. 1996. Excision of the *piggyBac* transposable element *in vitro* is a precise event that is enhanced by the expression of its encoded transposase. *Genetica* 98: 33-41

Engels, W.R., D.M. Johnson-Schlitz, W.S. Eggleston, and J. Svedt. 1990. High-frequency *P* element loss in *Drosophila* is homolog dependent. *Cell* 62: 515-525

Evgen'ev, M.B., H. Zelentsova, N. Shostak, M. Kozitsina, V. Barskiy, D.-H. Lankenau, and V.G. Corces. 1997. *Penelope*, a new family of transposable elements and its possible role in hybrid dysgenesis in *Drsophila virlis*. *Proceedings of the National Academy of Sciences of the United States of America* 94: 196-201

Feschotte, C. and S.R. Wessler. 2002. *Mariner*-like transposases are widespread and diverse in flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* 99: 280-285

Feschotte, C., X. Zhang, and S.R. Wessler. 2002. Miniature inverted-repeat transposable elements and their relationship to established DNA transposons. In *Mobile DNA II* (eds. N.L. Craig R. Craigie M. Gellert, and A.M. Lambowitz). American Society for Microbiology Press, Washington, D.C., USA

Feschotte, C. and E.J. Pritham. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics* 41: 331-368

Felsenstein, J. 1985.Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791

Finnegan, D.J. 1989. Eukaryotic transposable elements and genome evolution. *Trends in Genetics* 5: 103-107

Fraser, M.J., L. Cary, K. Boonvisudhi, and H.-G.H. Wang. 1995. Assay for movement of lepidopteran transposon IFP2 in insect cells using a baculovirus genome as a target DNA. *Virology* 211: 397-407

Fraser, M.J. 2000. The TTAA-specific family of transposable elements: identification, functional characterization, and utility for transformation of insects. In *Insect Transgenesis: Methods and Applications* (eds. A.M. Handler and A.A. James). CRC Press, Boca Raton, FA, USA

Ganley, A.R.D. and T. Kobayashi. 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Research* 17: 184-191

Gao, X., Y. Hou, H. Ebina, H.L. Levin, and D.F. Voytas. 2008. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Research* 18: 359-369

Gentile, K.L., W.D. Burke, and T.H. Eickbush. 2001. Multiple lineages of R1 retrotransposable elements can coexist in the rDNA loci of *Drosophila*. *Molecular Biology and Evolution* 18: 235-245

Gladyshev, E.A. and I.R. Arkhipova. 2009. Rotifer rDNA-specific R9 retrotransposable elements generate an exceptionally long target site duplication upon insertion. *Gene* 448: 145-150

Glass, S.K., A. Moszczynska, and T.J. Crease. 2008. The effect of transposon *Pokey* insertions on sequence variation in the 28S rRNA gene of *Daphnia pulex*. *Genome* 51: 988-1000

Goodwin, T.J.D. and R.T.M. Poulter. 2001. The DIRS1 group of retrotransposons. *Molecular Biology and Evolution* 18: 2067-2082

Goodwin, T.J.D., M.I. Butler, and R.T.M. Poulter. 2003. Cryptons: a group of tyrosine-recombinase encoding DNA transposons from pathogenic fungi. *Microbiology* 149: 3099-3109

Goodwin, T.J.D. and R.T.M. Poulter. 2004. A new group of tyrosine recombinase-encoding retrotransposons. *Molecular Biology and Evolution* 21: 746-759

Gould, S.J. and E.S. Vrba. 1982. Exaptation-a missing term in the science of form. *Paleobiology* 8: 4-15

Gregory, T.R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews* 76: 65-101

Gregory, T.R. 2005a. Synergy between sequence and size in large-scale genomics. *Nature Reviews Genetics* 6: 699-708

Gregory, T.R. 2005b. *The Evolution of the Genome*. Elsevier Academic Press, London, UK

Han, J.S. 2010. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mobile DNA* 1: 15

Hancock, C.N., F. Zhang, and S.R. Wessler. 2010. Transposition of the *Tourist*-MITE *mPing* in yeast: an assay that retains key features of catalysis by the class 2 *PIF/Harbinger* superfamily. *Mobile DNA* 1: 5

Havecker, E.R., X. Xiang Gao, and D.F. Voytas. 2004. The diversity of LTR retrotransposons. *Genome Biology* 5: 225

Hebert, P.D.N. 1978. The population biology of *Daphnia* (Crustacea,Daphnidae). *Biological Reviews* 53: 387-426

Hedges, D.J. and P.L. Deininger. 2007. Inviting instability: transposable elements, double-strand breaks, and the maintenance of genomic integrity. *Mutation Research* 616: 46-59

Hickey, D.A. 1982. Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* 101: 519-531

Hickman, A.B., M. Chandler, and F. Dyda. 2010. Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. *Critical Reviews in Biochemistry and Molecular Biology* 45: 50-69

Higashiyama, T., Y. Noutoshi, M. Fujie, and T. Yamada. 1997. Zepp, a LINE-like retrotransposon accumulated in the *Chlorella* telomeric region. *The European Molecular Biology Organization Journal* 16: 3715-3723

Hillis, D.M. and M.T. Dixon. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. *The Quarterly Review of Biology* 66: 411-453

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432: 695-716

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921

Izsvák, Z., Y. Wang, and Z. Ivics. 2009. Interactions of transposons with the cellular DNA repair machinery. In *Transposons and the Dynamic Genome* (eds. D.-H.

Lankenau and J.-N. Volff), pp. 133-176. Spring-Verlag, Berlin/Heidelberg, Germany

Jordan, I.K., I.B. Rogozin, G.V. Glazko, and E.V. Koonin. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics* 19: 68-72

Jordan, I.K. and W.J. Miller. 2009. Genome defence against transposable elements and the origins of regulatory RNA. In *Transposons and the Dynamic Genome* (eds. D.-H. Lankenau and J.-N. Volff), pp. 77-94. Springer-Verlag, Berlin/Heidelberg, Germany

Kapitonov, V.V. and J. Jurka. 2001. Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* 98: 8714-8719

Kapitonov, V.V. and J. Jurka. 2005. RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. *PLoS Biology* 3: e181

Kapitonov, V.V. and J. Jurka. 2006. Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* 103: 4540-4545

Kapitonov, V.V. and J. Jurka. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics* 9: 411-412

Kaplan, N.L., T. Darden, and C.H. Langley. 1985. Evolution and extinction of transposable elements in Mendelian populations. *Genetics* 109: 459-480

Keith, J.H., C.A. Schaeper, T.S. Fraser, and M.J. Fraser. 2008a. Mutational analysis of highly conserved aspartate residues essential to the catalytic core of the *piggyBac* transposase. *BMC Molecular Biology* 9: 73

Keith, J.H., T.S. Fraser, and M.J. Fraser. 2008b. Analysis of the *piggyBac* transposase reveals a functional nuclear targeting signal in the 94 C-terminal residues. *BMC Molecular Biology* 9: 72

Kidwell, M.G. and D.R. Lisch. 2000. Transposable elements and host genome evolution. *Trends in Ecology and Evolution* 15: 95-99

Kidwell, M.G. and D.R. Lisch. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* 55: 1-24

Kim, D., J. Huh, and H. Kim. 2007. Transposable elements in human cancers by genome-wide EST alignment. *Genes and Genetic Systems* 82: 145-156

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111-120

Kjems, J. and R.A. Garrett. 1991. Ribosomal RNA introns in archaea and evidence for RNA conformational changes associated with splicing. *Proceedings of the National Academy of Sciences of the United States of America* 88: 439-443

Kojima, K.K. and H. Fujiwara. 2003. Evolution of target specificity in R1 clade non-LTR retrotransposons. *Molecular Biology and Evolution* 20: 351-361

Kojima, K.K. and H. Fujiwara. 2004. Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets. *Molecular Biology and Evolution* 21: 207-217

Kojima, K.K. and H. Fujiwara. 2005. Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Molecular Biology and Evolution* 22: 2157-2165

Kruger, K., P.J. Grabowski, A.J. Zaug, J. Sands, D.E. Gottschling, and T.R. Cech. 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell* 31: 147-157

Lal, S., M. Oetjens, and L.C. Hannah. 2009. *Helitrons*: enigmatic abductors and mobilizers of host genome sequences. *Plant Science* 176: 181-186

Lampe, D.J., K.K.O. Walden, and H.M. Robertson. 2001. Loss of transposase-DNA interaction may underlie the divergence of *mariner* family transposable elements and the ability of more than one *mariner* to occupy the same genome. *Molecular Biology and Evolution* 18: 954-961

Langer, M., L.F. Sniderhan, U. Grossniklaus, and A. Ray. 2007. Transposon excision from an atypical site: a mechanism of evolution of novel transposable elements. *PLoS One* 10: e965

Lee, J., H. Lee, M. Shin, and W. Ryu. 2004. Versatile PCR-mediated insertion or deletion mutagenesis. *BioTechniques* 36: 398-400

Le Rouzic, A. and P. Capy. 2006. Population genetics models of competition between transposable element subfamilies. *Genetics* 174: 785-793

Liao, D. 1999. Concerted evolution: molecular mechanism and biological implications. *American Journal of Human Genetics* 64: 24-30

Lockton, S. and B.S. Gaut. 2010. The evolution of transposable elements in natural population of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. *BMC Evolutionary Biology* 10: 10

Lohe, A.R., E.N. Moriyama, D.-A. Lindholm, and D.L. Hartl. 1995. Horizontal transmission, vertical inactivation, and stochastic loss of *mariner*-like transposable elements. *Molecular Biology and Evolution* 12: 62-72

Long, E.O. and I.B. Dawid. 1979. Expression of ribosomal DNA insertions in *Drosophila melanogaster*. *Cell* 18: 1185-1196

Long, E.O. and I.B. Dawid. 1980. Repeated genes in eukaryotes. *Annual Review of Biochemistry* 49: 727-764

Lorenc, A. and W. Makalowski. 2003. Transposable elements and vertebrate protein diversity. *Genetica* 118: 183-191

Luan, D.D., M.H. Korman, J.L. Jakubczak, and T.H. Eickbush. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* 72: 595-605

Lynch, M. and J.S. Conery. 2003. The origins of genome complexity. *Science* 302: 1401-1404

McClintock, B. 1946. Maize genetics. *Carnegie Institute of Washington Yearbook* 45: 176-186

McClintock, B. 1947. Cytogenetic studies of maize and *Neurospora*. *Carnegie Institute of Washington Yearbook* 46: 146-152

McClintock, B. 1950. The origin and behaviour of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America* 36: 344-355

McClintock, B. 1961. Some parallels between gene control systems in maize and in bacteria. *The American Naturalist* 95: 265-277

Mieczkowski, P.A., F.J. Lemoine, and P. T.D. 2006. Recombination between retrotransposons as a source of chromosome rearrangements in the yeast Saccharomyces cerevisiae. DNA Repair 5: 1010-1020

Miller, W.J., S. Hagemann, E. Reiter, and W. Pinsker. 1992. *P*-element homologous sequences are tandemly repeated in the genome of *Drosophila guanche*. *Proceedings of the National Academy of Sciences of the United States of America* 89: 4018-4022

Mitra, R., J. Fain-Thornton, and N.L. Craig. 2008. *piggyBac* can bypass DNA synthesis during cut and paste transposition. *The European Molecular Biology Organization Journal* 27: 1097-1109

Moran, J.V. and N. Gilbert. 2002. Mammalian LINE-1 retrotransposons and related elements. In *Mobile DNA II* (eds. N.L. Craig R. Craigie M. Gellert, and A.M. Lambowitz), pp. 836-869. American Society of Microbiology Press, Washington, DC, USA

Muscarella, D.E. and V.M. Vogt. 1989. A mobile group I intron in the nuclear rDNA of *Physarum polycephalum*. *Cell* 56: 443-454

Nevers, P. and H. Saedler. 1977. Transposable genetic elements as agents of gene instability and chromosomal rearrangements. *Nature* 268: 109-115

Nuzhdin, S.V. and D.A. Petrov. 2003. Transposable elements in clonal lineages: lethal hangover from sex. *Biological Journal of the Linnean Society* 79: 33-41

Orgel, L.E. and F.H.C. Crick. 1980. Selfish DNA: the ultimate parasite. *Nature* 284: 604-607

Östergren, G. 1945. Parasitic nature of extra fragment chromosomes. *Botaniska Notiser* 2: 157-163

Panchin, Y. and L.L. Moroz. 2008. Molluscan mobile elements similar to the vertebrate Recombination-Activating Genes. *Biochemical and Biophysical Research Communications* 369: 818-823

Pardue, M.-L. and P.G. DeBaryshe. 1999. *Drosophila* telomeres: two transposable elements with important roles in chromosomes. *Genetica* 107: 189-196

Penton, E.H., B.W. Sullender, and T.J. Crease. 2002. *Pokey*, a new DNA transposon in *Daphnia* (Cladocera:Crustacea). *Journal of Molecular Evolution* 55: 664-673

Penton, E.H. and T.J. Crease. 2004. Evolution of the transposable element *Pokey* in the ribosomal DNA of species in the subgenus *Daphnia* (Crustacea: Cladocera). *Molecular Biology and Evolution* 21: 1727-1739

Pérez-González, C.E. and T.H. Eickbush. 2002. Rates of R1 and R2 retrotransposition and elimination from the rDNA locus of *Drosophila melanogaster*. *Genetics* 162: 799-811

Petersen, G. and O. Seberg. 2009. *Stowaway* MITEs in *Hordeum* (Poaceae): evolutionary history, ancestral elements and classification. *Cladistics* 25: 1-11

Peterson, P.A. 1970. Controlling elements and mutable loci in maize: their relationship to bacterial episomes. *Genetica* 41: 33-56

Piriyapongsa, J., N. Polavarapu, M. Borodovsky, and J. McDonald. 2007. Exonization of the LTR transposable elements in human genome. *BMC Genomics* 8: 291

Pritham, E.J., T. Putliwala, and C. Feschotte. 2007. *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* 390: 3-17

Quesneville, H., C.M. Bergman, O. Andrieu, D. Autard, D. Nouaud, M. Ashburner, and D. Anxolabéhère. 2005. Combined evidence annotation of transposable elements in genome sequences. *PloS Computational Biology* 1: e22

Rizzon, C., G. Marais, M. Gouy, and C. Biémont. 2002. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Research* 12: 400-407

Saitou, N. and Nei, M. 1987.The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425

Sandmeyer, S.B., M. Aye, and T. Menees. 2002. Ty3, a position-specific *gypsy*-like element in *Saccharomyces cerevisiae*. In *Mobile DNA II* (eds. N.L. Craig R. Craigie M. Gellert, and A.M. Lambowitz), pp. 663-683. American Society of Microbiology Press, Washington, DC, USA

Sanij, E. and R.D. Hannan. 2009. The role of UBF in regulating the structure and dynamics of transcriptionally active rDNA chromatin. *Epigenetics* 4: 374-382

Sarkar, A., C. Sim, Y.S. Hong, J.R. Hogan, M.J. Fraser, H.M. Robertson, and F.H. Collins. 2003. Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related ''domesticated'' sequences. *Molecular Genetics and Genomics* 270: 173-180

Schaack, S., C. Gilbert, and C. Feschotte. 2010a. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology and Evolution* 25: 537-546

Schaack, S., E. Choi, M. Lynch, and E.J. Pritham. 2010b. DNA transposons and the role of recombination in mutation accumulation in *Daphnia pulex*. *Genome Biology* 11: R46

Schaack, S., E.J. Pritham, A. Wolf, and M. Lynch. 2010c. DNA transposon dynamics in populations of *Daphnia* pulex with and without sex. *Proceedings of the Royal Society of London B* 277: 2381-2387

Seberg, O. and G. Petersen. 2009. A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nature Reviews Genetics* 10: 276

Silva, J.C., E.L. Loreto, and J.B. Clark. 2004. Factors that affect the horizontal transfer of transposable elements. *Current Issues in Molecular Biology* 6: 57-72

Simons, C., I.V. Makunin, M. Pheasant, and J.S. Mattick. 2007. Maintenance of transposon-free regions throughout vertebrate evolution. *BMC Genomics* 8: 470

Sullender, B.W. 1993. Preliminary characterization and population genetic survey of the *Daphnia* rDNA transposable element, *Pokey*. University of Oregon, Eugene

Sullender, B.W. and T.J. Crease. 2001. The behavior of a *Daphnia pulex* transposable element in cyclically and obligately parthenogenetic populations. *Journal of Molecular Evolution* 53: 63-69

Tamura, K., Dudley J., Nei, M. & Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24:1596-1599

Valizadeh, P. and T.J. Crease. 2008. The association between breeding system and transposable element dynamics in *Daphnia pulex*. *Journal of Molecular Evolution* 66: 643-654

van de Lagemaat, L.N., J.-R. Landry, D.L. Mager, and P. Medstrand. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends in Genetics* 19: 530-536

Venner, S., C. Feschotte, and C. Biémont. 2009. Dynamics of transposable elements: towards a community ecology of the genome. *Trends in Genetics* 25: 317-323

Voytas, D.F. and J.D. Boeke. 2002. Ty1 and Ty5 of *Saccharomyces cerevisiae*. In *Mobile DNA II* (eds. N.L. Craig R. Craigie M. Gellert, and A.M. Lambowitz), pp. 631-662. American Society of Microbiology Press, Washington, DC, USA

Wang, J., Y. Dua, S. Wang, S.J. Brown, and Y. Park. 2008. Large diversity of the *piggyBac*-like elements in the genome of *Tribolium castaneum*. *Insect Biochemistry and Molecular Biology* 38: 490-498

Wang, J., E.D. Miller, G.S. Simmons, T.A. Miller, B.E. Tabashnik, and Y. Park. 2010. *piggyBac*-like elements in the pink bollworm, *Pectinophora gossypiella*. *Insect Molecular Biology* 19: 177-184

Weil, C.F. and R. Kunze. 2000. Transposition of maize *Ac/Ds* transposable elements in the yeast *Saccharomyces cerevisiae*. *Nature Genetics* 26: 187-190

Wessler, S.R., T.E. Bureau, and S.E. White. 1995. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Current Opinion in Genetics & Development* 5: 814-821

Wicker, T., F. Sabot, A. Hua-Van, J.L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A.H. Schulman. 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* 8: 973-982

Winckler, T., K. Szafranski, and G. Glöckner. 2005. Transfer RNA gene-targeted integration: an adaptation of retrotransposable elements to survive in the compact *Dictyostelium discoideum* genome. *Cytogenetic and Genome Research* 110: 288-298

Wilson, M.H., C.J. Coates, and A.L. George. 2007. *PiggyBac* transposon-mediated gene transfer in human cells. *Molecular Therapy* 15: 139-145

Wong, L.H. and K.H.A. Choo. 2004. Evolutionary dynamics of transposable elements at the centromere. *Trends in Genetics* 20: 611-616

Wood, D. and R.A. Brink. 1956. Frequency of somatic mutation to self color in maize plants homozygous and heterozygous for variegated pericarp. *Proceedings of the National Academy of Sciences of the United States of America* 42: 514-519

Wright, S. and D. Finnegan. 2001. Genome evolution: sex and the transposable element. *Current Biology* 11: 296-299

Xiong, Y. and T.H. Eickbush. 1988. The site-specific ribosomal DNA Insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. *Molecular and Cellular Biology* 8: 114-123

Xu, H.-F., Q.-Y. Xia, C. Liu, T.-C. Cheng, P. Zhao, J. Duan, X.-F. Zha, and S.-P. Liu. 2006. Identification and characterization of *piggyBac*-like elements in the genome of domesticated silkworm, *Bombyx mori*. *Molecular Genetics and Genomics* 276: 31-40

Yang, G., C.F. Weil, and S.R. Wessler. 2006. A rice Tc1/*mariner*-like element transposes in yeast. *The Plant Cell* 18: 2469-2478

Yang, G., D.H. Nagel, C. Feschotte, C.N. Hancock, and S.R. Wessler. 2009. Tuned for transposition: molecular determinants underlying the hyperactivity of a *Stowaway* MITE. *Science* 325: 1391-1394

Yoder, J.A., C.P. Walsh, and T.H. Bestor. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics* 13: 335-340

Table 1. *Pokey* sequences that were not derived from the *D. pulex* genome sequence. PCU= Penton and Crease, unpublished, PC= Penton and Crease, 2004, NA = not available

Element	**GenBank #**	**Source**
PCU-D.pulex1	NA	Penton and Crease, Unpublished
PCU-D.obtusa9	NA	Penton and Crease, Unpublished
PCU-D.obtusa11	NA	Penton and Crease, Unpublished
PCU-D.obtusa17	NA	Penton and Crease, Unpublished
PCU-E.D.pulex9	NA	Penton and Crease, Unpublished
PCU-D.tenebrosa4	NA	Penton and Crease, Unpublished
PCU-D.tenebrosa	NA	Penton and Crease, Unpublished
PCU-D.pulex4	NA	Penton and Crease, Unpublished
PCU-D.pulicaria10	NA	Penton and Crease, Unpublished
PCU-D.retrocurva7	NA	Penton and Crease, Unpublished
PCU-D.retrocurva3	NA	Penton and Crease, Unpublished
PC-D.cheraphila	AY630592	Penton and Crease, 2004
PC-D.catawba1	AY630597	Penton and Crease, 2004
PC-D.minnehaha10	AY630596	Penton and Crease, 2004
PC-D.pulicaria2	AY115590	Penton and Crease, 2004
PC-D.arenata	AY630585	Penton and Crease, 2004
PC-D.pulicaria1	AY115589	Penton and Crease, 2004
PCU-D.middendorffiana3	NA	Penton and Crease, Unpublished
PC-D.pulex1	AY630580	Penton and Crease, 2004

PC-D.pulex2	AY630581	Penton and Crease, 2004
PC-D.pulex3	AY630579	Penton and Crease, 2004
PC-E.D.pulex1	AY630583	Penton and Crease, 2004
PC-E.D.pulex2	AY630584	Penton and Crease, 2004
PC-D.retrocurva1	AY630595	Penton and Crease, 2004
PC-D.parvula1	AY630593	Penton and Crease, 2004
PC-D.parvula2	AY630594	Penton and Crease, 2004
PC-D.pileata1	AY630586	Penton and Crease, 2004
PC-D.obtusa4	AY630591	Penton and Crease, 2004
PC-D.obtusa11	AY630587	Penton and Crease, 2004
PC-D.obtusa1	AY630589	Penton and Crease, 2004
PC-D.obtusa2	AY630588	Penton and Crease, 2004
PC-D.obtusa7	AY630590	Penton and Crease, 2004
PCU-D.obtusaB10	NA	Penton and Crease, Unpublished
PCU-D.obtusaB6	NA	Penton and Crease, Unpublished
PC-D.ambigua1	AY630598	Penton and Crease, 2004
PC-E.D.pulicaria5	AY630582	Penton and Crease, 2004
PC-D.pulicaria3	AY115590	Penton and Crease, 2004
PC-D.obtusa2-7	AY630590	Penton and Crease, 2004

Table 2. Polymerase chain reaction primers.

Primer Name	Description	Sequence
PokattB1ORFF	attB1 recombination tail + 5' end of Pokey 6.6 kb ORF	5'-gggg aca agt ttg tac aaa aaa gca ggc ttg - caa aag aag gcc gat gcc aaa aaa gtc g-3'
PokORFDelR	Reverse overlap extension primer to remove the intron and ligate the 2 transposase exons together	5'-c cag ata att ttc ctc gac - aat atc att gaa gca tat cc-3'
PokORFDelF	Forward overlap extension primer to remove the intron and ligate the 2 transposase exons together	5'-gg ata tgc ttc aat gat att - gtc gag gaa aat tat ctg g-3'
PokattB2ORFR	attB2 recombination tail + 3' end of Pokey 6.6 kb ORF	5'-gggg ac cac ttt gta caa gaa agc tgg gtc - ttg ttg gaa atc atc ata atc atc aat cat ata gcc ttc-3'
Pok4065F	ORF fidelity confirmation primer	5'-tga ttc acc gag gcc tca gtt c-3'
Pok441 F	ORF fidelity confirmation primer	5'-gtc gat gtt ctg gga gca gc-3'
Pok5026F	ORF fidelity confirmation primer	5'-tcg aac ctg cag ccg gac gaa ttt gca g-3'
Pok4488R	ORF fidelity confirmation primer	5'-gaa tcg ctc gcg agt cat gg-3'
Pok5985R	ORF fidelity confirmation primer	5'-cac gtc ggt tag aat att ctg gct cgt cgg-3'
M13F	Standard primer for sequencing across the cloning sites of multiple plasmids	5'-gtt gta aaa cga cgg cca gtg-3'
M13R	Standard primer for sequencing across the cloning sites of multiple plasmids	5'-cag gaa aca gct atg acc atg-3'
PokMITE5F	AA + 5' ITR to insert into *Hpa*I site (GTT AAC) of pWL89A	5'-aa ccc ttt ttc gac tga cgg gac gtt ttt ttt gc-3'
PokLigPCRR	Deletion/ligation reverse primer for 5' half of Pok6.6NA	5'-tga ctc tca ttc ac - gga tcc tta tca cta t - ggc aat tca att ctg tag g-3'
PokLigPCRF	Deletion/ligation forward primer for 3' half of Pok6.6NA	5'-aca gaa ttg aat tgc c - ata gtg ata agg atc c -gtg aat gag agt caa gc-3'
PokMITE3'BR	TT + 3' ITR to insert into *Hpa*I site (GTT AAC) of pWL89A	5'-aa ccc ttt atc gac cgc cac aaa agcg-3'
Ade2F	Amplifies across the *Hpa*I insertion site of pWL89A and derivatives	5'-tcg tct tga agt cga gga ctt tgg ca-3'
Ade2R	Amplifies across the *Hpa*I insertion site of pWL89A and derivatives	5'-aac gga gtc cgg aac tct agc agg cgc a -3'

Table 3. Sequence divergence estimates within and between clusters of *Pokey* elements. Estimates were calculated using the Kimura 2-parameter and pairwise deletion options in MEGA4. Estimates above the diagonal are based on a global alignment of both full length elements and MITEs (77 sequences, 567 positions in the dataset).Estimates on and below the diagonal are based on separate alignments of MITEs (49 sequences, 1327 positions in the dataset) and of *Pokey*A and *Pokey*B elements (29 sequences, 17 401 positions in the dataset).

Lineage	***Pokey*A**	***Pokey*B**	**MITE1**	**MITE2**
***Pokey*A**	0.068	0.488	0.645	0.439
***Pokey*B**	0.408	0.050	0.469	0.553
MITE1			0.022	0.315
MITE2			0.249	0.032

Table 4. Features of complete or nearly complete *Pokey* transposase ORFs. Protein lengths were estimated using the stop codon of the 5kb and 6.6 kb elements (premature stop codons were ignored). Promoter strength was predicted by the Neural Network Promoter Prediction program (http://www.fruitfly.org/seq_tools/promoter.html). Optimum scores are those closest to 1.0. Assignment of putative function was based upon comparison to the 6.6kb element from *D. pulicaria*. Element s92-184466-7920 has a premature stop codon whose inclusion may or may not affect transposase function.

Element	Cluster	Protein Length	Premature Stop Codons	Promoter Strength	Putatively Functional	Poly A Signal
s112-301057-4390	B	735	1	0.97	No	Yes
s14-88789-6277	B	642	0	1.0	Yes	Yes
s201-111870-4656	B	734	2	1.0	No	Yes
s212-104907-5537	B	594	0	0.99	Yes	Yes
s38-755471-6930	A	667	2	1.0	Yes	Yes
s92-184466-7920	A	600	1	1.0	?	Yes
s69-38572-4729	A	668	0	1.0	Yes	Yes
D.pulicaria-6.6kb	A	666	0	1.0	Yes	Yes
D.pulicaria-5kb	A	648	1	1.0	Yes	Yes

Table 5. Target site duplications (TSD) found in *Pokey* insertion sites from the *Daphnia pulex* genome.

TSD	% of Insertions
TTAA	88.13
TTAT	5.93
ATAA	5.09
CTAA	0.85

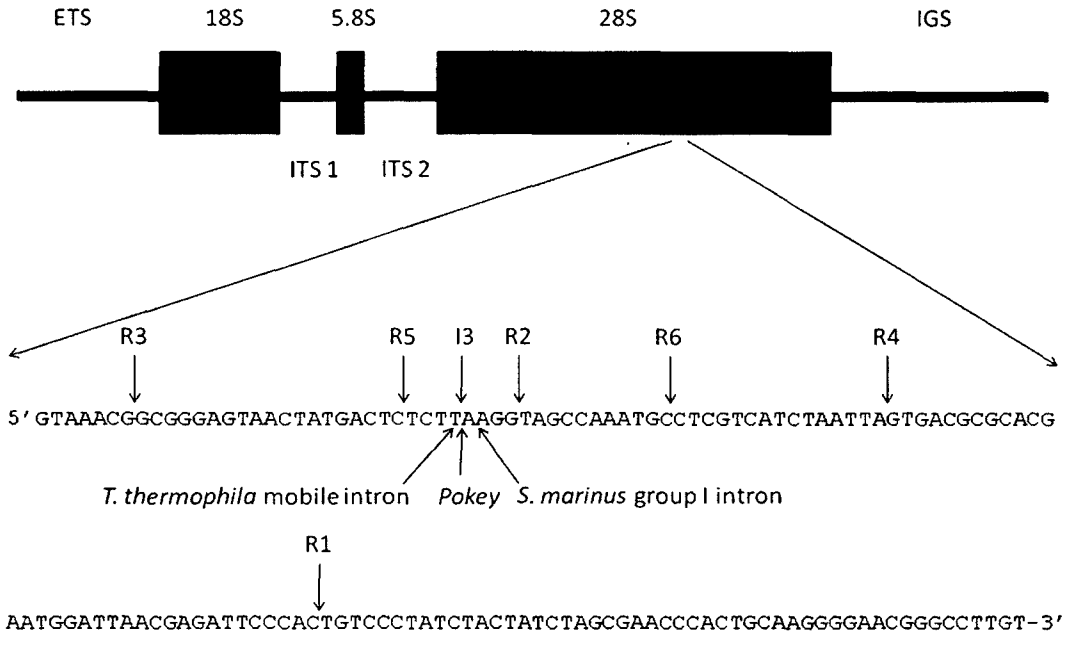

Figure 1. Location of the insertion site of *Pokey* and other mobile DNA into a conserved region of the 28S rRNA gene (Kruger et al., 1982; Muscarella and Vogt, 1989; Kjems and Garrett, 1991; Burke et al., 1993; Burke et al., 1995; Eickbush, 2002; Penton et al., 2002; Kojima and Fujiwara, 2003; Burke et al., 2003). The sequence shown is from *Daphnia pulicaria*. Arrows indicate insertion sites of each element. ETS= external transcribed spacer, ITS= internal transcribed spacer, IGS= intergenic spacer.

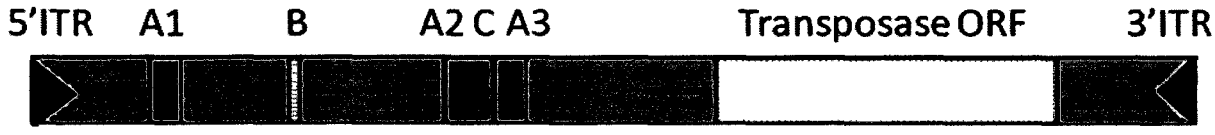

Figure 2. Organization of the 6.6 kb *Pokey* DNA transposon from *Daphnia pulicaria* (Penton et al., 2002). ITR= inverted terminal repeat, ORF= open reading frame. A, B and C refer to sequences with similarity to other regions of the *D. pulex* genome.

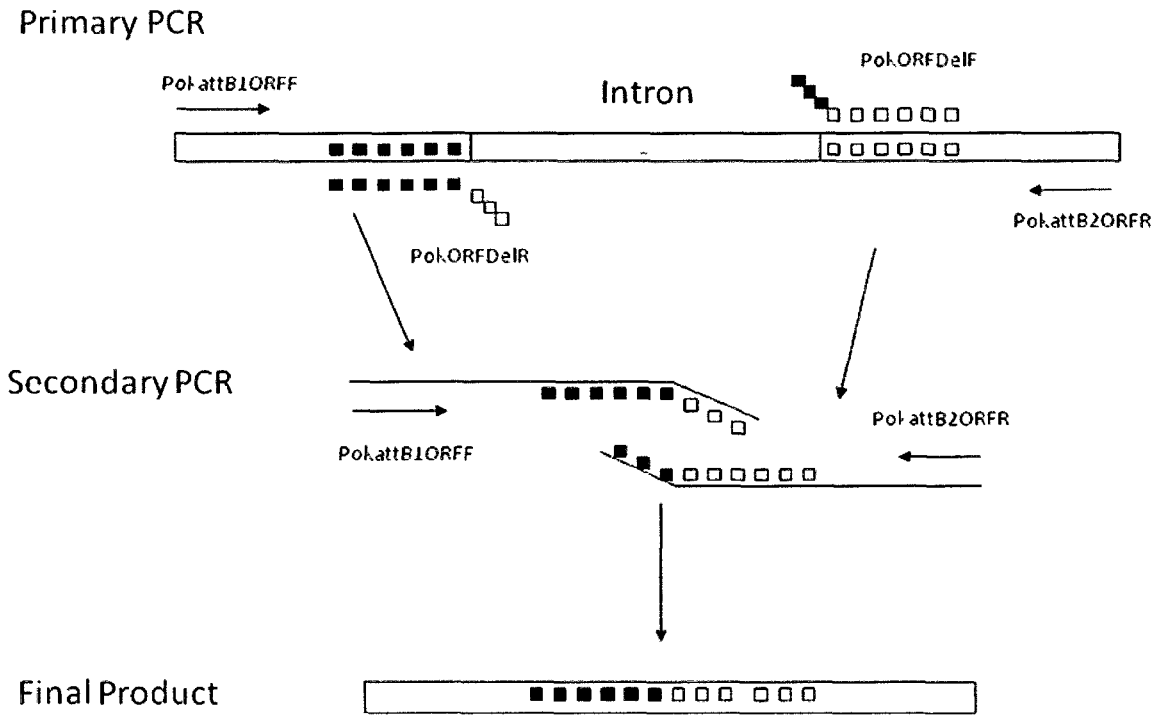

Figure 3. Overlap extension PCR method of Lee et al. (2004). Removal of the 68 bp intron from the 6.6 kb *Pokey* element transposase gene is shown.

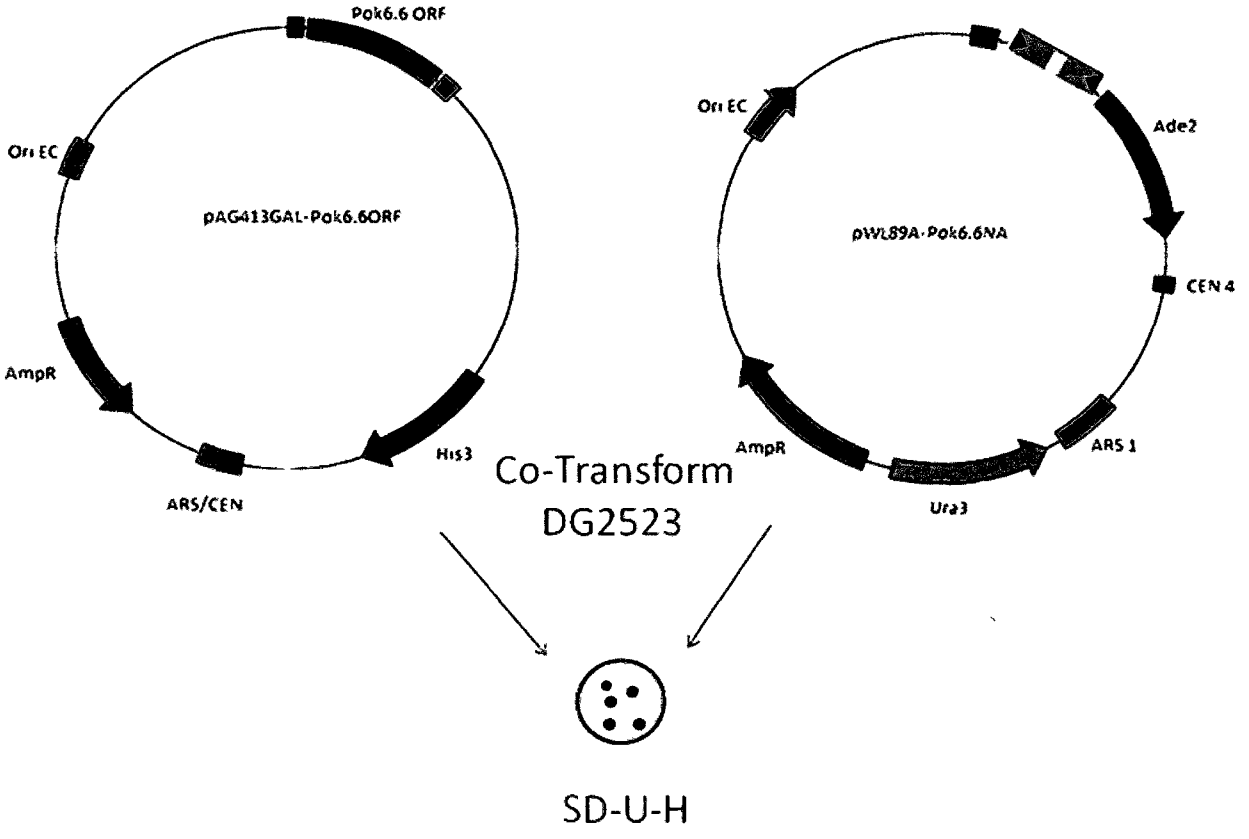

Figure 4. Plasmids used to measure *Pokey* excision rate in yeast strain, DG2523. Ori EC= *E.coli* origin of replication, AmpR= ampicillin resistance gene, ARS/CEN= yeast autonomous replicating sequence/centromere, His3= Imidazoleglycerol-phosphate dehydratase ORF, Ura3= Orotidine-5'-phosphate decarboxylase ORF, Ade2= Phosphoribosylaminoimidazole carboxylase ORF

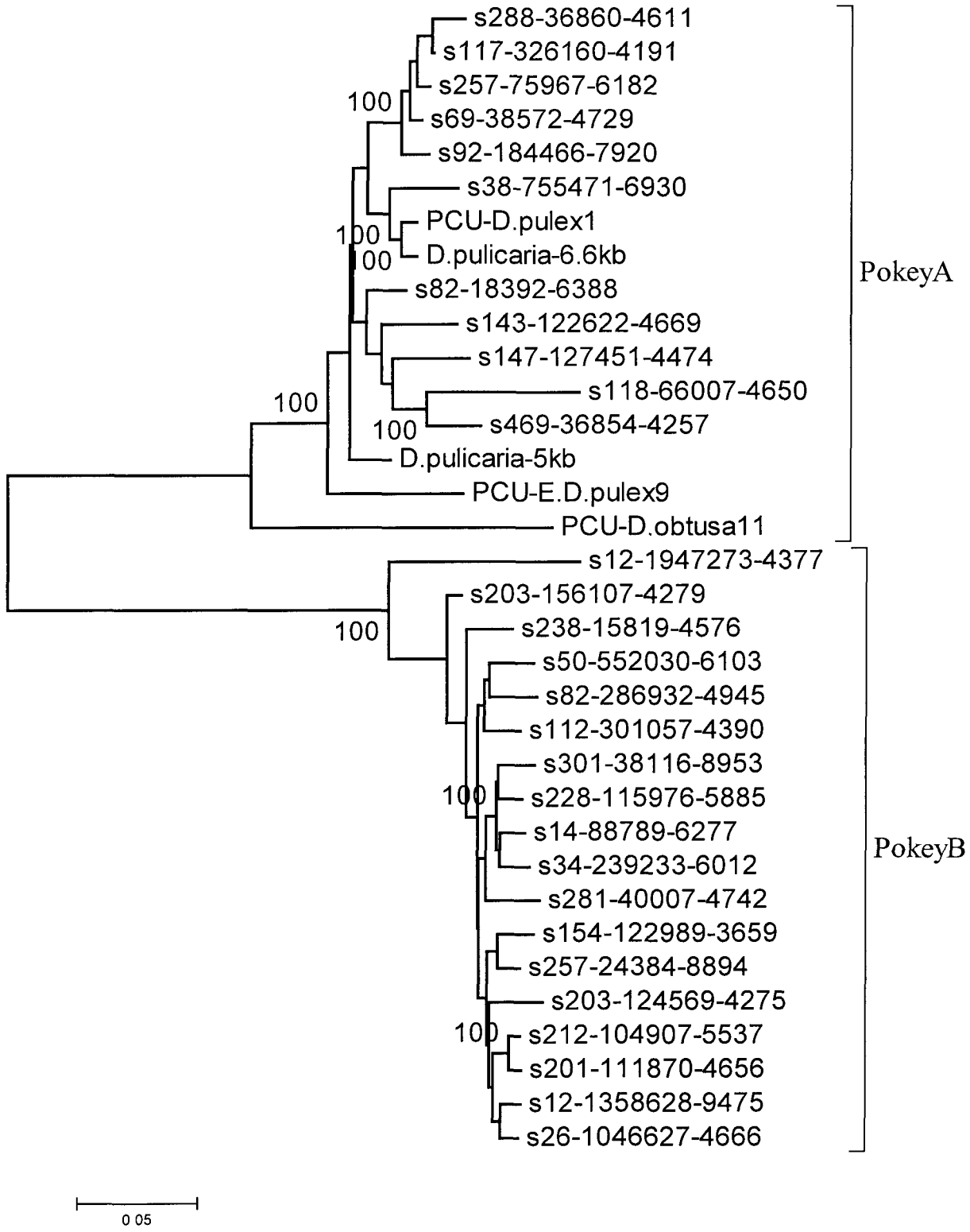
s288-36860-4611
s117-326160-4191
s257-75967-6182
100
s69-38572-4729
s92-184466-7920
s38-755471-6930
PCU-D.pulex1
100
D.pulicaria-6.6kb
00
s82-18392-6388
s143-122622-4669
s147-127451-4474
s118-66007-4650
100
s469-36854-4257
100
D.pulicaria-5kb
PCU-E.D.pulex9
PCU-D.obtusa11
PokeyA
s12-1947273-4377
s203-156107-4279
100
s238-15819-4576
s50-552030-6103
s82-286932-4945
s112-301057-4390
s301-38116-8953
100
s228-115976-5885
s14-88789-6277
s34-239233-6012
s281-40007-4742
s154-122989-3659
s257-24384-8894
s203-124569-4275
100
s212-104907-5537
s201-111870-4656
s12-1358628-9475
s26-1046627-4666
PokeyB
0 05

Figure 5. Unrooted NJ tree of 34 full length *Pokey* elements from the *Daphnia pulex* genome sequence. The elements form two clusters denoted *Pokey*A and *PokeyB*. All positions containing alignment gaps and missing data were eliminated only in pairwise comparisons (pairwise deletion option). There were a total of 17401 positions in the final dataset. Bootstrap values greater than 70 are indicated at the nodes of the tree.

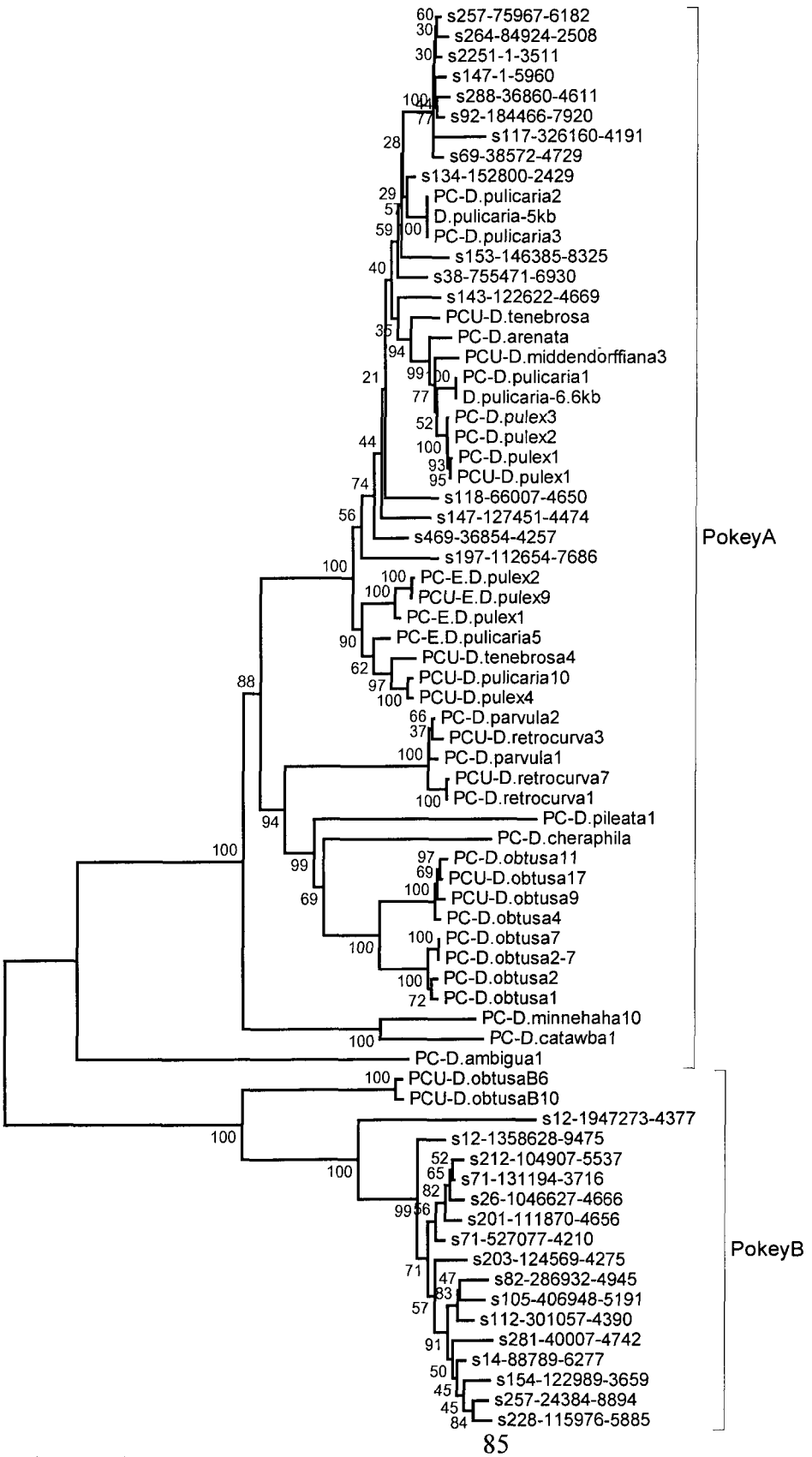
s257-75967-6182
s264-84924-2508
s2251-1-3511
s147-1-5960
s288-36860-4611
s92-184466-7920
s117-326160-4191
s69-38572-4729
s134-152800-2429
PC-D.pulicaria2
D.pulicaria-5kb
PC-D.pulicaria3
s153-146385-8325
s38-755471-6930
s143-122622-4669
PCU-D.tenebrosa
PC-D.arenata
PCU-D.middendorffiana3
PC-D.pulicaria1
D.pulicaria-6.6kb
PC-D.pulex3
PC-D.pulex2
PC-D.pulex1
PCU-D.pulex1
s118-66007-4650
s147-127451-4474
s469-36854-4257
s197-112654-7686
PC-E.D.pulex2
PCU-E.D.pulex9
PC-E.D.pulex1
PC-E.D.pulicaria5
PCU-D.tenebrosa4
PCU-D.pulicaria10
PCU-D.pulex4
PC-D.parvula2
PCU-D.retrocurva3
PC-D.parvula1
PCU-D.retrocurva7
PC-D.retrocurva1
PC-D.pileata1
PC-D.cheraphila
PC-D.obtusa11
PCU-D.obtusa17
PCU-D.obtusa9
PC-D.obtusa4
PC-D.obtusa7
PC-D.obtusa2-7
PC-D.obtusa2
PC-D.obtusa1
PC-D.minnehaha10
PC-D.catawba1
PC-D.ambigua1
PokeyA
PCU-D.obtusaB6
PCU-D.obtusaB10
s12-1947273-4377
s12-1358628-9475
s212-104907-5537
s71-131194-3716
s26-1046627-4666
s201-111870-4656
s71-527077-4210
s203-124569-4275
s82-286932-4945
s105-406948-5191
s112-301057-4390
s281-40007-4742
s14-88789-6277
s154-122989-3659
s257-24384-8894
s228-115976-5885
PokeyB
0.05

Figure 6. Unrooted NJ tree of 71 1600-bp sequences from the 3' end of *Pokey* elements. Elements from the *Daphnia pulex* genome sequence and others cloned from species in the sub-genus *Daphnia* (Penton and Crease, 2004) are included. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 2811 positions in the final dataset. Bootstrap values greater than 70 are indicated at the nodes of the tree.

Transposase DNA

5′-CTTCAATGATATT**GTGAGATCACAGAT~~~TTTTTCTTGACTTAG**GTCGAGGAAAATT-3`

Transposase mRNA

5′-TGAACAGCTGGATATGCTTCAATGATATTGTCGAGGAAAATTATCTGGAGGCTTACGA-3`

Figure 7. Comparison of the DNA and spliced mRNA of the *Pokey* transposase gene. The intron sequence is marked in bold and the codons flanking it are underlined.

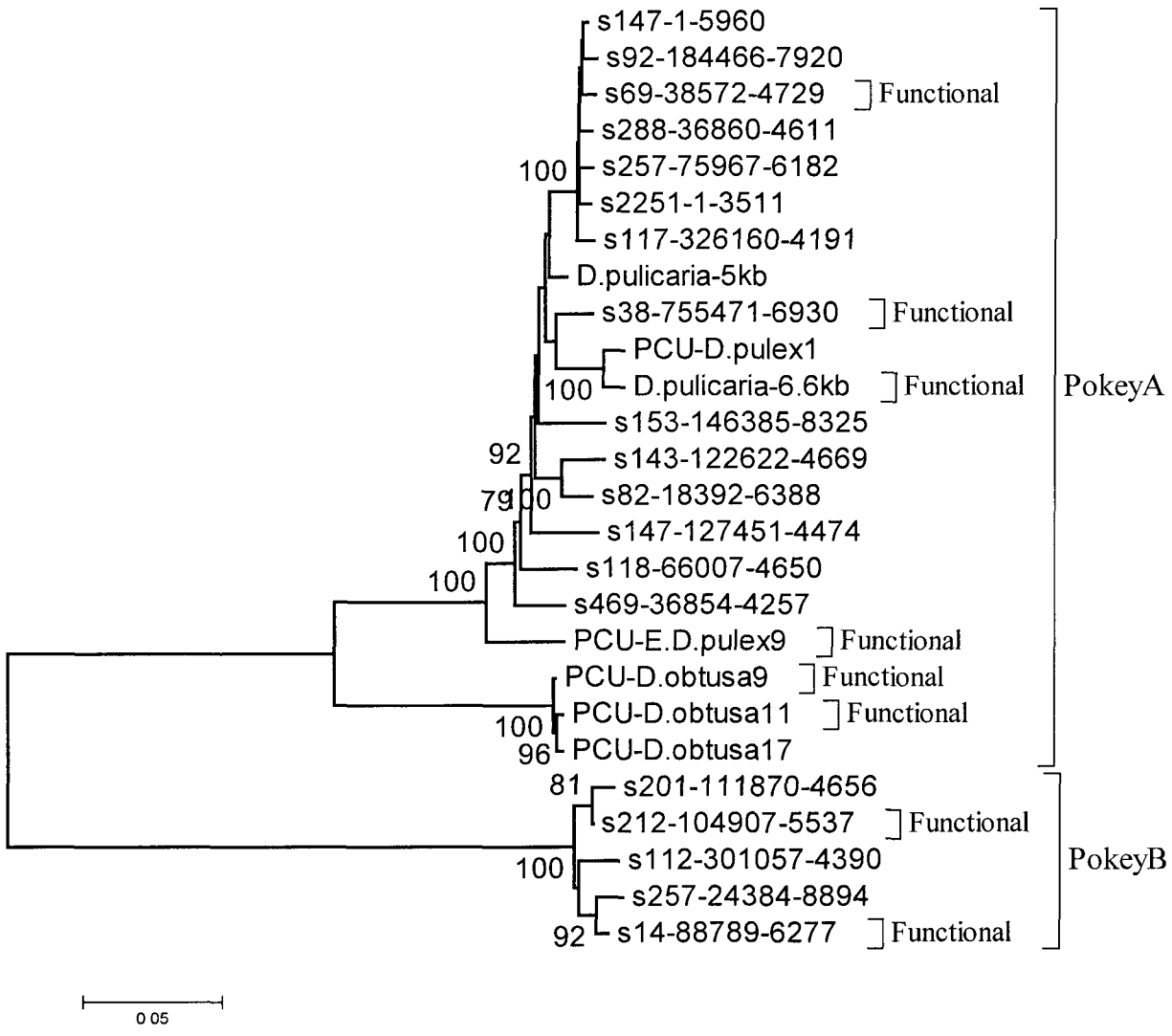

Figure 8.Unrooted NJ tree of 26 *Pokey* transposase coding regions. Sequences were obtained from elements in the *Daphnia pulex* genome sequence and those cloned from several other species from the sub-genus *Daphnia* (Table 1). "Functional" indicates elements whose transposase could be functional based on comparison to the 6.6 kb *D. pulicaria* element (Penton et al., 2002). All three codon positions were included in the analysis. All positions containing alignment gaps and missing data were eliminated in pairwise sequence comparisons (pairwise deletion option). There were a total of 3797 positions in the final dataset. Bootstrap values greater than 70 are indicated at the nodes of the tree.

Figure 9. Partial alignment of six *Pokey* transposase amino acid sequences and four other *piggyBac*-superfamily transposases. The three conserved catalytic aspartic acid (D) residues, the four cysteine (C) residues thought to compose the zinc-finger/PHD motif and the putative NLS are highlighted. The asparagine (N) residue conserved in *Pokey* transposases is highlighted in grey. Other *piggyBac* elements have D at his position. yabusame-W = putative *Bombyx mori piggyBac* transposase, ABS18391.1 = *piggyBac* transposase from *Helicoverpa armigera*, Uribo2 = *piggyBac* transposase from *Xenopus tropicalis*, NP_689808.2 = *piggyBac* transposase-derived protein from *Homo sapiens*

```
s212-104907-5537        IRSLVDVLNKQFNECRRPPRWQSIDESMVKFKGRSMLRKTMKGKPIKSGFKIWSRC-CSR
s14-88789-6277          IRSLVDVLNKQFSECRRPPRWQSIDESMVKFKGRSMLRQTMKGKPIKSGFKIWSQC-CSR
D.pulicaria-5kb         IRPLVKRLNERYHACRKPPRGQSIDESMVKFKGRSVLRQTMKNKPIKSGFKIWSRC-CHR
D.pulicaria-6.6kb       IRPLVKRLNESYHVCRKPPRGQSIDESMVKYKGRSMLRQTMKNTPIKSGFKIWSRC-CLR
s69-38572-4729          IRPLVKRLNERYHACRKPPSGQSIDESMVKFKGRSVLRQTMKNKPIKSGFKIWSRC-CNR
s38-755471-6930         IRPLVKRLNERYHECRKPLRGQSIDESMVKYKGRSMLRQTIKNKPIKSGFKIWSRC-CHR
yabusame-W              FRSIFDQFVQCCQNAYSPSEFLTIDEMLLSFRGRCLFRVYIPNKPAKYGIKILALVDAKN
ABS18391.1              VRKIWEIFINQCRQNHVPGSNLTVDEQLLGFRGRCPFRMYIPNKPDKYGIKFPMMCAAAT
Uribo2                  LRPLIDSLSERFAAVYTPCQNICIDESLLLFKGRLQFRQYIPSKRARYGIKFYKLCESSS
NP_689808.2             IKPVFDFLVNKFSTVYTPNRNIAVDESLMLFKGPLAMKQYLPTKRVRFGLKLYVLCESQS

s212-104907-5537        GYTYKFEIYHGTR-IGETPKDS-NFTMVEGVVLDLCEPLAKIGHVVAFDRFFTSIALLDE
s14-88789-6277          GYTYKFEIYHGTR-IGETPKDS-NFTMVEGVVLDLCEPLAKIGHVVAFDRFFTSIALLDE
D.pulicaria-5kb         GYTYKFEIYQGAR-FGEKQGRSRNNEAVERVVVDLCQPLTDQGFVVAFDRFFTSIALLDK
D.pulicaria-6.6kb       GYTYKFEIYQGAR-FGEKQKRSRNNEAVERVVVDLCQPLTDQGFVVAFDRFFTSIALLDK
s69-38572-4729          GYTYKFEIYQGAR-FGEKQGRSRNNEAVERVVVDLCQPLTDQGFVVAFDRFFTSIALLDK
s38-755471-6930         GYTYKFEIYQGAR-FGEKQKRSRNNEAVERVVVDLCQPLTDQGFVVAFDRFFTSIALLDK
yabusame-W              FDVVNLEVYAGKQPSGPYAVSNRPFEVVERL----IQPVARSHRNVTFDNWFTGYELMLH
ABS18391.1              KYMIDAIPYLGKS-TKTN-----GLPLGEFYVKDLTKTVHGTNRNITCDNWFTSIPLAKN
Uribo2                  GYTSYFLIYEGKDSKLDPPGCPPDLTVSGKIVWELISPLLGQGFHLYVDNFYSSIPLFTA
NP_689808.2             GYVWNALVHTGP---GMNLKDSADGLKSSRIVLTLVNDLLGQGYCVFLDNFNISPMLFRE
```

```
s212-104907-5537        LDKR--GINAVGTILKTRVGQPIFTVNESNLPQDEFVAKFGGEPGTGRKGVFIW--KDTK
s14-88789-6277          LDKR--GINVVGTILKTRVGQPIFTVNESNLPQDEFVAKFGGEPGTGRKGLFIW--KDTK
D.pulicaria-5kb         LREN--GVNAVGTILPSRVNQPIMTKNESNLRPDEFAAKFGGEPGTCRKGIFVW--RDTK
D.pulicaria-6.6kb       LREN--GVNAVGTILPSRVNQPIMTKNESNLRPDEFAAKFGGEPGTCRKGIFVW--RDTK
s69-38572-4729          LREN--GVNAVGTILPSRVNQPIMTKNESNLRPDEFAAKFGGEPGTCRKGIFVW--RDTK
s38-755471-6930         LREN--GVNAVGTILPSRVNQPIMTKNESNLRPDEFAAKFGGEPGTCRKGIFVW--RDTK
yabusame-W              LLNE-YRLTSVGTVRKNKRQIPESFIR-TDRQPNSSVFGFQKDITLVS-----YAPKKNK
ABS18391.1              MLQAPYNLTIVGTIRSNKREMPEEIKNSRSRPVGSSMFCFDGPLTLVS-----YKPKPSK
Uribo2                  LYCL--DTPACGTINRNRKGLPRALLDK-KLNRGETYALRKNELLAIK-----F--FDKK
NP_689808.2             LHQN--RTDAVGTARLNRKQIPNDLKK--RIAKGTTVARFCGELMALK-----W--CDGK

s212-104907-5537        PFRVASNFNGSEVVKVQRKQRDGSFRTKSCPKAIGDYVDNMGGVDTANQLRSYYERDRKS
s14-88789-6277          PFRVASNFHGSEVVKVQKKQRDGSFRTKSCPKAIGDYVDNMGGVDTANQLRSYYERDRKS
D.pulicaria-5kb         AF--ASNYHGSNLVKVRRKQRDGSFKTKSCPKAIDDYVNNMGGVDTANQLRSYYERDRKA
D.pulicaria-6.6kb       AFRVASNYHGSDIVKVRRRQRDGSFRSKSCPKAIDDYVNNMGGVDTANQLRSYYERDRKA
s69-38572-4729          AFRVASNYHGSDIVKVRRKQRDGSFKTKSCPKAIDDYVNNMGGVDTANQLRSYYERDRKS
s38-755471-6930         AFRVASNYHGSDIVKVRRKQRDGSFKTKSCPKAIDDYVNNMGGVDTANQLRSYYERDRKA
yabusame-W              VVVVMSTMHHDNSIDESTGEKQK-------PEMITFYNSTKAGVDVVDELSANYNVSRNS
ABS18391.1              MVFLLSSCDENAVINESNGK----------PDMILFYNQTKGGVDSFDQMCKSMSANRKT
Uribo2                  NVFMLTSIHDESVIREQRVGRPPKNK----PLCSKEYSKYMGGVDRTDQLQHYYNATRKT
NP_689808.2             EVTMLSTFHNDTVIEVNNRNGKKTKR----PRVIVDYNENMGAVDSADQMLTSYPSERKR

s212-104907-5537        KK-WWHRLFYSLMETCMVNSWITYCDLVKGKTFLKDKRYLSLLEFKRSVTTSILFYWLNA
s14-88789-6277          KK-WWHRLFYSLMETCMVNSWITYCDLVKGKTFLKDKRYLSLLEFKRSVTTSLLLYGLNA
D.pulicaria-5kb         KK-WWHRLLYSLLETCLVNSWICFNDMVEENYLENFEAPMTFLEFKRNVTMGLLSHALNE
D.pulicaria-6.6kb       KK-WWHRLLYSLLETCLVNSWICFNDIVEENYLEAYEVQMPFLEFKRNVTMGLLSHALNE
s69-38572-4729          KK-FWHRLLYSLLETCLVNSWVCFNDMVENNYLEYFEVQITFLEFKRNVTMGLLSNALNE
s38-755471-6930         KK-WWHRLLYSLLETCLVNSWICFNDMVEENYLEHFEVPMTFLEFKRNVTMGLLSHALNE
yabusame-W              KR-WPMTLFYGVLNMAAINACIIYRA--------NKNVTIKRTEFIRSLGLSMIYEHLHS
ABS18391.1              NR-WPMAVFYGMLNMAFVNSYIIYCH----NKINKQEKPISRKEFMKKLSIQLTTPWMQE
Uribo2                  RA-WYKKVGIYLIQMALRNSYIVYKAAVPGP-------KLSYYKYQLQILPALLFGGVEE
NP_689808.2             HKVWYKKFFHHLLHITVLNSYILFKK-------DNPEHTMSHINFRLALIERML------
```

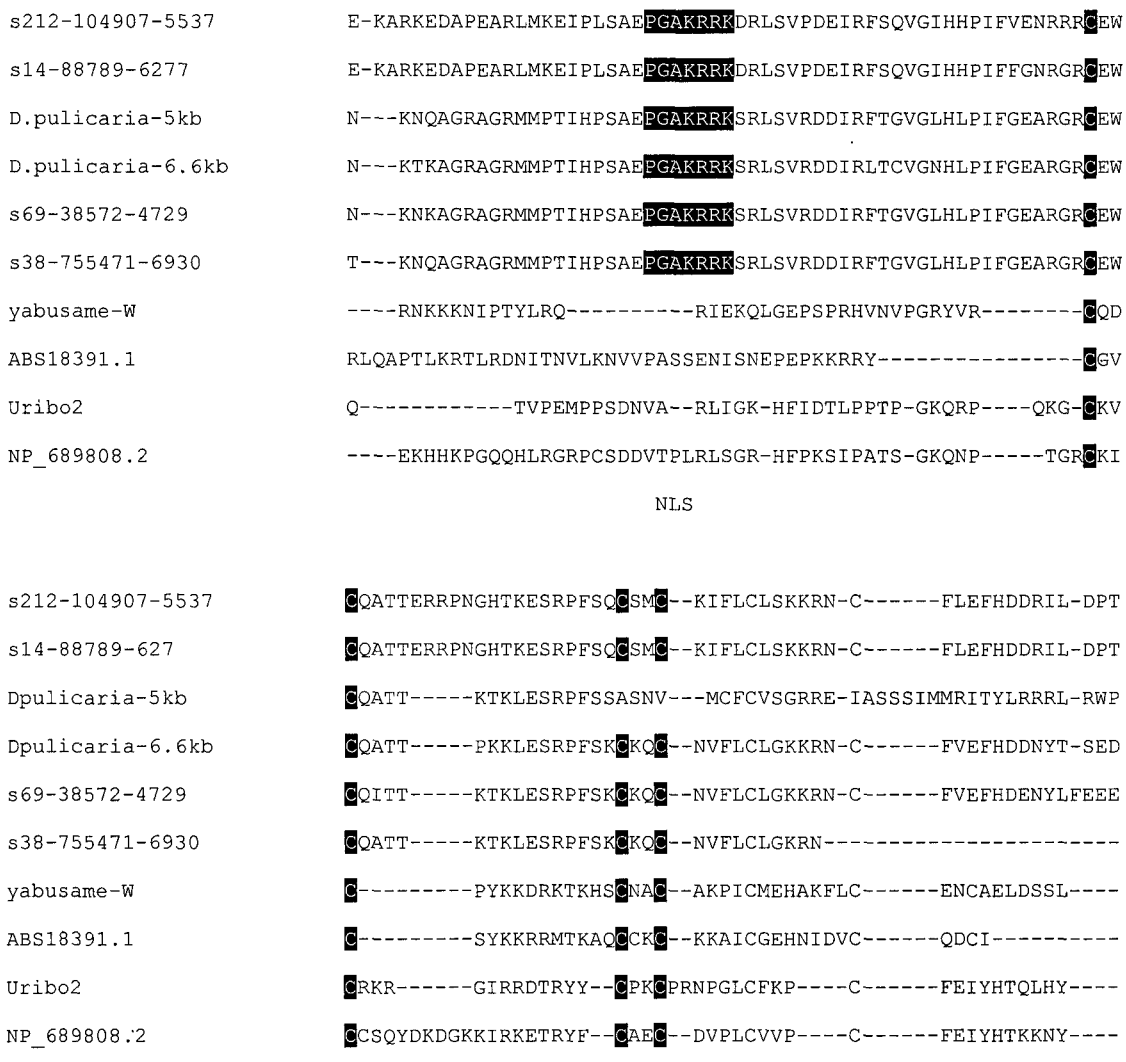
s212-104907-5537 E-KARKEDAPEARLMKEIPLSAEPGAKRRKDRLSVPDEIRFSQVGIHHPIFVENRRRCEW
s14-88789-6277 E-KARKEDAPEARLMKEIPLSAEPGAKRRKDRLSVPDEIRFSQVGIHHPIFFGNRGRCEW
D.pulicaria-5kb N---KNQAGRAGRMMPTIHPSAEPGAKRRKSRLSVRDDIRFTGVGLHLPIFGEARGRCEW
D.pulicaria-6.6kb N---KTKAGRAGRMMPTIHPSAEPGAKRRKSRLSVRDDIRLTCVGNHLPIFGEARGRCEW
s69-38572-4729 N---KNKAGRAGRMMPTIHPSAEPGAKRRKSRLSVRDDIRFTGVGLHLPIFGEARGRCEW
s38-755471-6930 T---KNQAGRAGRMMPTIHPSAEPGAKRRKSRLSVRDDIRFTGVGLHLPIFGEARGRCEW
yabusame-W ----RNKKKNIPTYLRQ----------RIEKQLGEPSPRHVNVPGRYVR--------CQD
ABS18391.1 RLQAPTLKRTLRDNITNVLKNVVPASSENISNEPEPKKRRY----------CGV
Uribo2 Q-----------TVPEMPPSDNVA--RLIGK-HFIDTLPPTP-GKQRP----QKG-CKV
NP_689808.2 ----EKHHKPGQQHLRGRPCSDDVTPLRLSGR-HFPKSIPATS-GKQNP-----TGRCKI
NLS
s212-104907-5537 CQATTERRPNGHTKESRPFSQCSMC--KIFLCLSKKRN-C------FLEFHDDRIL-DPT
s14-88789-627 CQATTERRPNGHTKESRPFSQCSMC--KIFLCLSKKRN-C------FLEFHDDRIL-DPT
Dpulicaria-5kb CQATT-----KTKLESRPFSSASNV---MCFCVSGRRE-IASSSIMMRITYLRRRL-RWP
Dpulicaria-6.6kb CQATT-----PKKLESRPFSKCKQC--NVFLCLGKKRN-C------FVEFHDDNYT-SED
s69-38572-4729 CQITT-----KTKLESRPFSKCKQC--NVFLCLGKKRN-C------FVEFHDENYLFEEE
s38-755471-6930 CQATT-----KTKLESRPFSKCKQC--NVFLCLGKRN----------
yabusame-W C---------PYKKDRKTKHSCNAC--AKPICMEHAKFLC------ENCAELDSSL----
ABS18391.1 C---------SYKKRRMTKAQCCKC--KKAICGEHNIDVC------QDCI----------
Uribo2 CRKR------GIRRDTRYY--CPKCPRNPGLCFKP----C------FEIYHTQLHY----
NP_689808.2 CCSQYDKDGKKIRKETRYF--CAEC--DVPLCVVP----C------FEIYHTKKNY----

Figure 10. Partial alignment of the translated transposase ORF sequences from 1600-bp *Pokey* fragments. Sequences were obtained from the *Daphnia pulex* genome and from Penton and Crease (2004). The conserved aspartic acid (D) and asparagine (N) residues are highlighted.

```
PCU-D.obtusa17              KVLRKQRDGSY--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
s153-146385-8325            KVRXKQRDGSF--KTKSCPKAIDDYVNNM--GGVDTLNQLRSYYERDRK
s143-122622-4669            KVRRKQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
s264-84924-2508             KVRRKQRDGSF--KTKSCPTAIDDYVNNM--GGVDTVNQLRSYYERDRK
s82-18392-6388              KVRRKQRDGSF--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.cheraphila             KVLRKQRDGSF--KQKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
s288-36860-4611             YCQGQEKAAXWFVQTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.catawba1               KVSRKQRDGSF--REKTCPKAIADYVDNM--GGVDTANQLSSYYERDRK
PC-D.minnehaha10            KAKRKQRDGSF--REKTCPKAIADYVDNM--GGVDTANQLSSYYERDRK
PCU-D.tenebrosa4            KVRRKQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
s2251-1-3511                KVRRKQRDGSF--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
D.pulicaria-5kb             KVRRKQRDGSF--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.pulicaria2             KVRRKQRDGSF--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-E.D.pulicaria5           KVRRKQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANRLRSYYERDRK
PC-D.arenata                KVRRKQRDGSF--XXKSCPKAIDDXVNNM--GGVDTANQLRSYYERDRK
D.pulicaria-6.6kb           KVRRRQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.pulicaria1             KVRRRQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PCU-D.middendorffiana3      KVRRKQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.pulex1                 KVRRKQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.pulex2                 KVRRKQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.pulex3                 KVRRKQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PCU-D.tenebrosa             KVRRKQRDGSF--RSKSCPXAIDDYXNNMETGGVDTANQLRSYYXRDRK
s118-66007-4650             KVRRKQRDGSF--RTKSCPKAIDDYVNNM--GGVDTVNQLRSYYERDRK
s69-38572-4729              KVRRKQRDGSF--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
s92-184466-7920             KVRRKQRDGSF--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
s147-1-5960                 KVRRKQRDGSF--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
s38-755471-6930             KVRRKQRDGSF--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
s134-152800-2429            KVRRKQRDGSF--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
```

```
s197-112654-7686          KVRRKQRDGTF--RSKSCPKAIDDYVNNM--GGVDTANQLRSXYERDRK
PCU-D.pulex4              KVRRKQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PCU-D.pulicaria10         KVRRKQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-E.D.pulex1             KVRRKQRDGSF--RSKSCPKAIDDXVNNM--GGVDTANQLRSYYERDRK
PC-E.D.pulex2             KVRRKQRDGSF--RSKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PCU-D.retrocurva7         KVLRKQRDGSF--RSKSCPKAIDDYVDNM--GGVDTANQLRSYYERDRK
PC-D.retrocurva1          KVLRKQRDGSF--RSKSCPKAIDDYVDNM--GGVDTANQLRSYYERDRK
PC-D.parvula1             KVLRKQRDGSF--RSKSCPKAIDDYVDNM--GGVDTANQLRSYYERDRK
PCU-D.retrocurva3         KVLRKQRDGSF--RSKSCPKAIDDYVDNM--GGVDTANQLRSYYERDRK
PC-D.parvula2             KVLRKQRDGSF--RSKSCPKAIDDYVDNM--GGVDTANQLRSYYERDRK
PC-D.pileata1             KVDRKQRDGSF--KKKSCPKAIADYVDNM--GGVDTSNQLRSYYERDRK
PCU-D.obtusa9             KVLRKQRDGSY--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.obtusa4              KVLRKQRDGSY--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.obtusa11             KVLRKQRDGSY--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.obtusa1              KVLRKQRDGSY--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.obtusa2              KVLRKQRDGSY--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.obtusa7              KVLRKQRDGSY--KTKSCPKAIDDYVNNM--GGVDTANQLRSYYERDRK
PC-D.ambigua1             QVERKQRDGSF--RKKPCPKAIADYVDNM--GGVDTANQLRSYYERDRK
s154-122989-3659          KVQRKQRDGSF--RTKSCPKAIGDYVDNM--CGVDTANQLRSXYERDRK
s281-40007-4742           KVQRKQRDGSF--RTKSCPKAIGDFVDNM--GGVDTANQLRSYYERDRK
s26-1046627-4666          KVQRKQRDGSF--RTKSCPKAIGDYVDNM--GGVDTANQLRSYYERNRK
s71-527077-4210           KVQRKQRDGSF--RTKSCPKAIGDYVDNM--GGVDTANQLRSYYERDRK
s257-24384-8894           KVQRKQRDGSF--RTKSCPKAIGDYVDNM--GGVDTANQLRSFYERDRK
PCU-D.obtusaB10           KVRRKQRDGSF--RTKSCPKAIADYVNNM--GGVDTANQLHSYYERDRK
PCU-D.obtusaB6            KVRRKQRDGSF--RTKSCPKAIADYVNNM--GGVDTANQLHSYYERDRK
s203-124569-4275          KVQRKQRDGSF--KTKSCPKAIGDYVDNM--GGVDTANQLHSYYERDRK
s14-88789-6277            KVQKKQRDGSF--RTKSCPKAIGDYVDNM--GGVDTANQLRSYYERDRK
s212-104907-5537          KVQRKQRDGSF--RTKSCPKAIGDYVDNM--GGVDTANQLRSYYERDRK
s71-131194-3716           KVQRKQRDGSF--RTKSCPKAIGDYVDNM--GGVDTANQLRSYYERDRK
s201-111870-4656          KVQRKQRDGSF--RTKSCPKAIGDYVDNM--GGVDTANQLRSYYERDRK
```

```
s112-301057-4390          KVQRKQRDGSF--RTKSCPKAIGDYVDNM--GGVDTANQLRSYYERDRK
s105-406948-5191          KVQRKQRDGSF--RTKSCPKAIGDYVDNM--GGVDTANQLRSYYERDRK
```

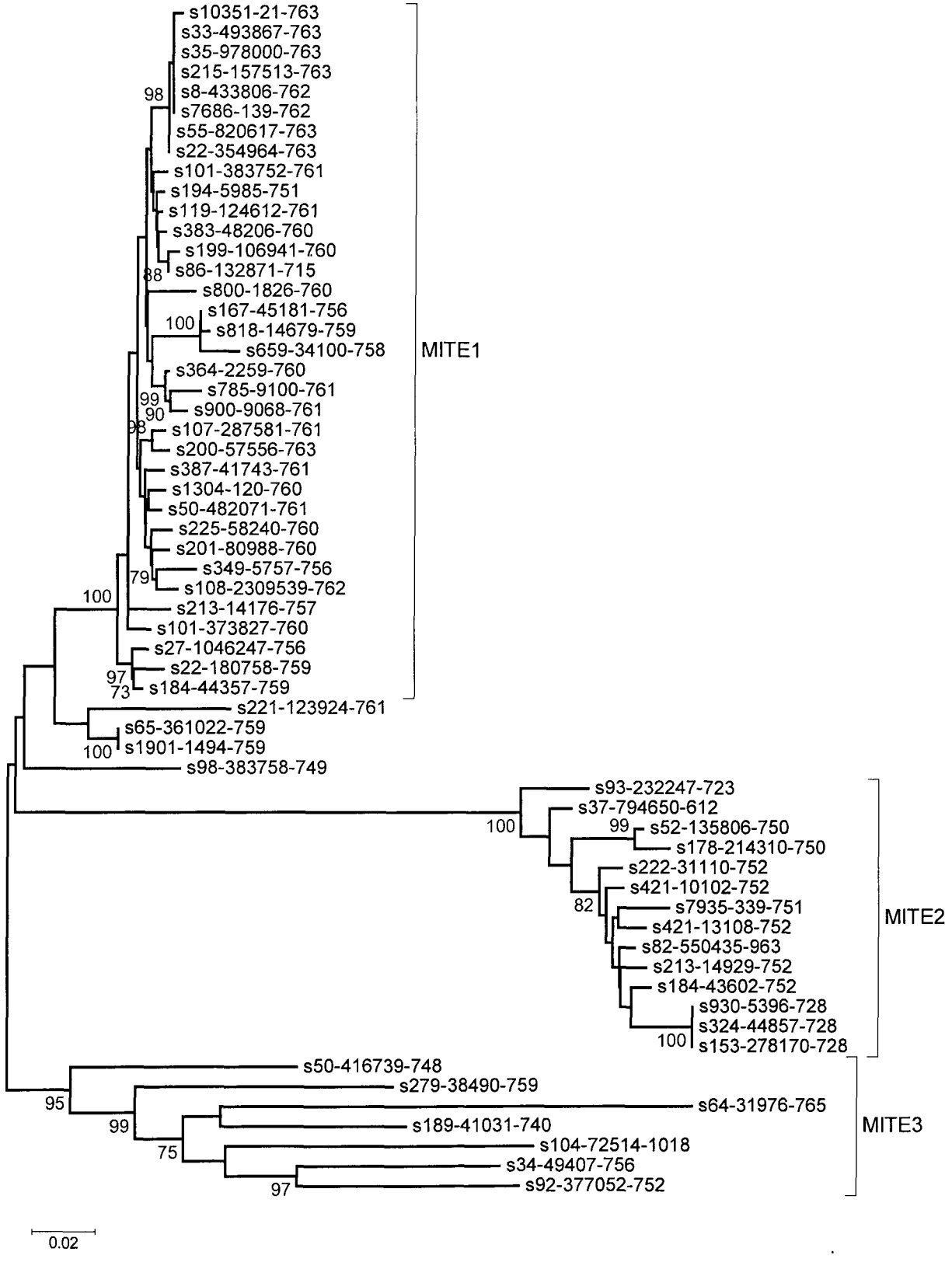
s10351-21-763
s33-493867-763
s35-978000-763
s215-157513-763
98
s8-433806-762
s7686-139-762
s55-820617-763
s22-354964-763
s101-383752-761
s194-5985-751
s119-124612-761
s383-48206-760
s199-106941-760
88
s86-132871-715
s800-1826-760
100
s167-45181-756
s818-14679-759
s659-34100-758
MITE1
s364-2259-760
99
s785-9100-761
90
s900-9068-761
98
s107-287581-761
s200-57556-763
s387-41743-761
s1304-120-760
s50-482071-761
s225-58240-760
s201-80988-760
79
s349-5757-756
s108-2309539-762
100
s213-14176-757
s101-373827-760
s27-1046247-756
97
s22-180758-759
73
s184-44357-759
s221-123924-761
s65-361022-759
100
s1901-1494-759
s98-383758-749
s93-232247-723
s37-794650-612
100
99
s52-135806-750
s178-214310-750
s222-31110-752
s421-10102-752
82
s7935-339-751
s421-13108-752
s82-550435-963
s213-14929-752
s184-43602-752
s930-5396-728
s324-44857-728
100
s153-278170-728
MITE2
s50-416739-748
s279-38490-759
95
s64-31976-765
99
s189-41031-740
75
s104-72514-1018
s34-49407-756
97
s92-377052-752
MITE3
0.02

Figure 11.Unrooted NJ tree of 60 *Pokey* MITE elements. The three major clusters are designated as MITE1, MITE2 and MITE3. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (pairwise deletion option). There were a total of 1327 positions in the final dataset. Bootstrap values greater than 70 are indicated at the nodes of the tree.

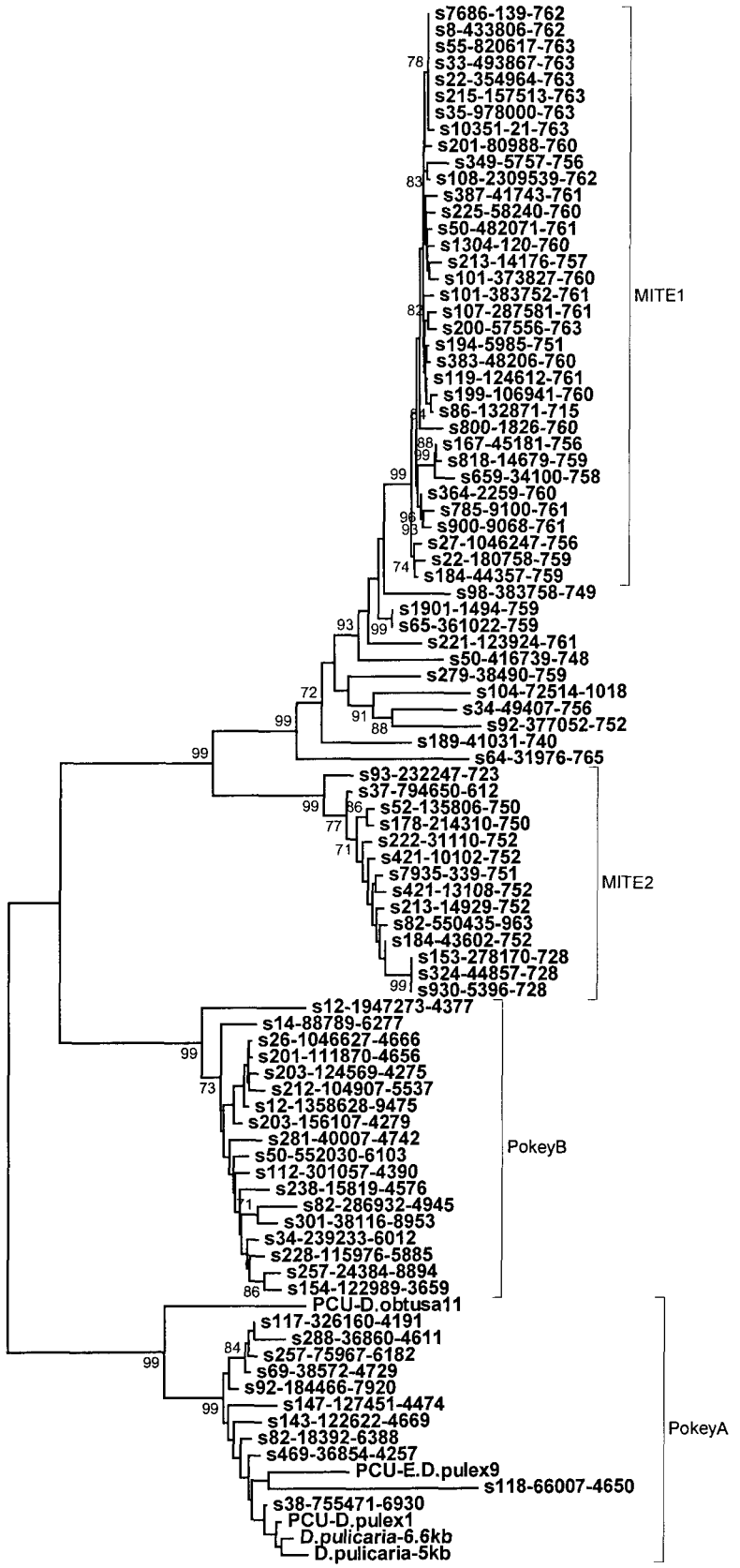
s7686-139-762
s8-433806-762
s55-820617-763
s33-493867-763
s22-354964-763
s215-157513-763
s35-978000-763
s10351-21-763
s201-80988-760
s349-5757-756
s108-2309539-762
s387-41743-761
s225-58240-760
s50-482071-761
s1304-120-760
s213-14176-757
s101-373827-760
s101-383752-761
s107-287581-761
s200-57556-763
s194-5985-751
s383-48206-760
s119-124612-761
s199-106941-760
s86-132871-715
s800-1826-760
s167-45181-756
s818-14679-759
s659-34100-758
s364-2259-760
s785-9100-761
s900-9068-761
s27-1046247-756
s22-180758-759
s184-44357-759
s98-383758-749
s1901-1494-759
s65-361022-759
s221-123924-761
s50-416739-748
s279-38490-759
s104-72514-1018
s34-49407-756
s92-377052-752
s189-41031-740
s64-31976-765
s93-232247-723
s37-794650-612
s52-135806-750
s178-214310-750
s222-31110-752
s421-10102-752
s7935-339-751
s421-13108-752
s213-14929-752
s82-550435-963
s184-43602-752
s153-278170-728
s324-44857-728
s930-5396-728
s12-1947273-4377
s14-88789-6277
s26-1046627-4666
s201-111870-4656
s203-124569-4275
s212-104907-5537
s12-1358628-9475
s203-156107-4279
s281-40007-4742
s50-552030-6103
s112-301057-4390
s238-15819-4576
s82-286932-4945
s301-38116-8953
s34-239233-6012
s228-115976-5885
s257-24384-8894
s154-122989-3659
PCU-D.obtusa11
s117-326160-4191
s288-36860-4611
s257-75967-6182
s69-38572-4729
s92-184466-7920
s147-127451-4474
s143-122622-4669
s82-18392-6388
s469-36854-4257
PCU-E.D.pulex9
s118-66007-4650
s38-755471-6930
PCU-D.pulex1
D.pulicaria-6.6kb
D.pulicaria-5kb
MITE1
MITE2
PokeyB
PokeyA
0 05

Figure 12. Unrooted NJ tree of 94 full length and MITE *Pokey* elements. The four major clusters are designated *Pokey*A, *Pokey*B, MITE1 and MITE2. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (pairwise deletion option). There were a total of 567 positions in the final dataset. Bootstrap values greater than 70 are indicated at the nodes of the tree.

ITR#1

PokeyA

5′-CCCTTTTTCGACTGAC----------GGCGGTCGATAAAGGG-3′

***Pokey* MITE2**

5′-CCCTTTTTCGACTGAC----------GGCGGTCGATAAAGGG-3′

ITR#2

PokeyB

5′-CCCTTTATCTAC----------GTCGAAAAAGGG-3′

***Pokey* MITE1**

5′-CCCTTTATCTAC----------GTCGATAAAGGG-3′

Figure 13. Two groups of Inverted Terminal Repeat (ITR) sequences found in *Pokey* elements from the *Daphnia pulex* genome. *Pokey*A and MITE2 share the ITR#1 sequence first identified by Penton et al. (2002). *Pokey*B and MITE1 share ITR#2. Complementary regions between the imperfect 5' and 3' ITRs are underlined.

5' End

```
PokeyAConsensus   TTAACCCTTTTTCGACTGACGGGACATTTTTTTTTCGACGGCTTT-CAGTAGGGTCTGGC
MITE2Consensus    TTAACCCTTTTTCGACTGACGGGACAAAAAAAATT-GACGGCTTTTCAGTAGGGTG-GTC
PokeyBConsensus   TTAACCCTTTATCTACCGTCGGGACAAAAAAAATTTGCSGGCTTTYCAGTAGGGTGGGGC
MITE1Consensus    TTAACCCTTTATCTACCGTCGGGACAAAAAAAATT-GACGGCTTTTCAGTAGGGTG-GCC
                  ********** ** ** * *******       ** *  ****** ********** * *

PokeyAConsensus   GGGAGCTCACCCACAGACAC-GAGAGATTTCA-AATCAGCTAAATTCATTCCTATCAAAG
MITE2Consensus    GGCGCAAATGTGCCCCCACACGAGAG-ATTTCAAATCAGCTAAATTCATTCTTTTCAAAG
PokeyBConsensus   GGGCGCAAGCTCGCAGAAAC-GAGAAAATTAA-AAYCAGACAAAGTGCCTCACACTTAMG
MITE1Consensus    GGCAGTAATGTGCCTCCACACGAGAGAAAAAAAAATCGGCAAAAGCGATCGCTCCTTTAG
                  **           *       ****         * *    ***

PokeyAConsensus   CTG
MITE2Consensus    CTG
PokeyBConsensus   CTG
MITE1Consensus    CTG
                  ***
```

3' End

```
PokeyAConsensus      ATATCTCGAAATCCATCCTACAGAATTGAATGRCCTTGBAAAAACTC----------
MITE2Consensus       ATATCTCGAAATCTATCCTACAGAATTGAATCGCCTCTTCAAATATTTGTTGTTGAC
PokeyBConsensus      TTTTCTCAATGATCTACCTACAAAATTGAATCGCCTTTTCAAGACTTTTATTAGTGC
MITE1Consensus       TTTTCTTGATGACCTACCTACAGAATTGAATCGCCATTATAGATATTTGTTGGTGAC
                      * ***  *       ****** ********  **          *

PokeyAConsensus   GAGCCGTCTGTCGGGACAAAAATTTGAACAAGTGGTAGTTGCATGCAA-CTATCGGGACA
MITE2Consensus    CAACCGACTGACGGGACAAATATTCACAAGAGTGGCTGCTGTGCGTAA-CTATCGGGACA
PokeyBConsensus   GAGCCGTCTGACGGGACAAATTTTTTTW-AAGTGGCTGCTGTGTGTAAACCGTCGGGACA
MITE1Consensus    CAACCGTCTGACGGGACAAATATTCRCAAGAGTGGCTGCTGTGTGTAA-CTGACGGGACA
                   * *** *** *********  **      *****  * **     ** *   *******

PokeyAConsensus   -AATTTTTTGTCCCGCTAGACGKCTCGAGTCGAAAAAAGTGTCTCTYTTTAGAGACCATG
MITE2Consensus    CCTATGGCGKGATTTTTCGCCATRAAAGT-------CCCGAC--AGACCTCAG--CGGTG
PokeyBConsensus   TTTTTTTTTGTCCCGGCGGACGGCAGTCGCTTGGAAATTTGGCCGTTTTT-----CGGTG
MITE1Consensus    CTT--------ATTTCGGGACATAGGTGT-------CCCGGC--AGACCTCAG--CGGTG
                             *                                      *  **

PokeyAConsensus   TCCGMCTGAAG---------GGACATTCCTGTCGGGACAACGGTGGCCRAAACGCGGTWA
MITE2Consensus    ----------GTGATATTT-GGACTTTCCTGTCGGGACAACGGTGGCCAAAACGCGCGGT
PokeyBConsensus   ----ACTGAAGTGCCGGTCGGGACTTTCCTGACGGGACAAAAAACAAAAAAGCAAAAAAA
MITE1Consensus    ----------GTGATATTT-GGACTTTCCTGTCGGGACAAAA--CAACAAAACGACCGAA
                            *         **** ****** ********         **

PokeyAConsensus   ---GGCCGGAAAAAAATCGGATATTCCGAATTTTTTTTAAATGAGTGGTCTTAGGACCAC
MITE2Consensus    AAGGCCCGGAAAAAATCGGMTATTCCG-AATTTTTTTTAAATGAGTGGTCTTAGGACCAC
PokeyBConsensus   ---GTACAGTTAAAAAAAGGAATTTTTTATTTCTTTTTAATTGAGTGGTTTGAGACGTCC
MITE1Consensus    AATGTACATTAAAAACTTGCCGAAAGG-TTTTTTTTTTTTTTGGGTGAGGTAATTDCGTC
                     *       ****   *           ** *****   ** ***   *        *

PokeyAConsensus   TCAATGATA--------TTTCGATCGAGCATTSAATCGAATCCGACCATCGGTTTTGTGG
MITE2Consensus    TCAATGATATTTCGATC-------GAGCATTGAATCGAAATCCGACCATCGGTTTTGTGG
PokeyBConsensus   TCAGCAATCT---GTACTTTTAAT------TTCAGTCAAGCACGGCCAAAAAAACTTTGG
MITE1Consensus    ACCTGGAATATGTGCAC-------TAACGGTCAGTTG-GCTCCGAC-AAAATTTTTCTTC
                  *                          *           ** * *       * *

PokeyAConsensus   CGGTCGATTAAAGGGTTAA
MITE2Consensus    CGGTCGA-TAAAGGGTTAA
PokeyBConsensus   C-GTCGAAAAAAGGGTTAA
MITE1Consensus    GCGTCGA-TAAAGGGTTAA
                    *****  **********
```

Figure 14. Partial alignment of the consensus sequences of the *Pokey*A, *Pokey*B, MITE1 and MITE2 clusters. Identical bases across all four consensus sequences are marked with stars.

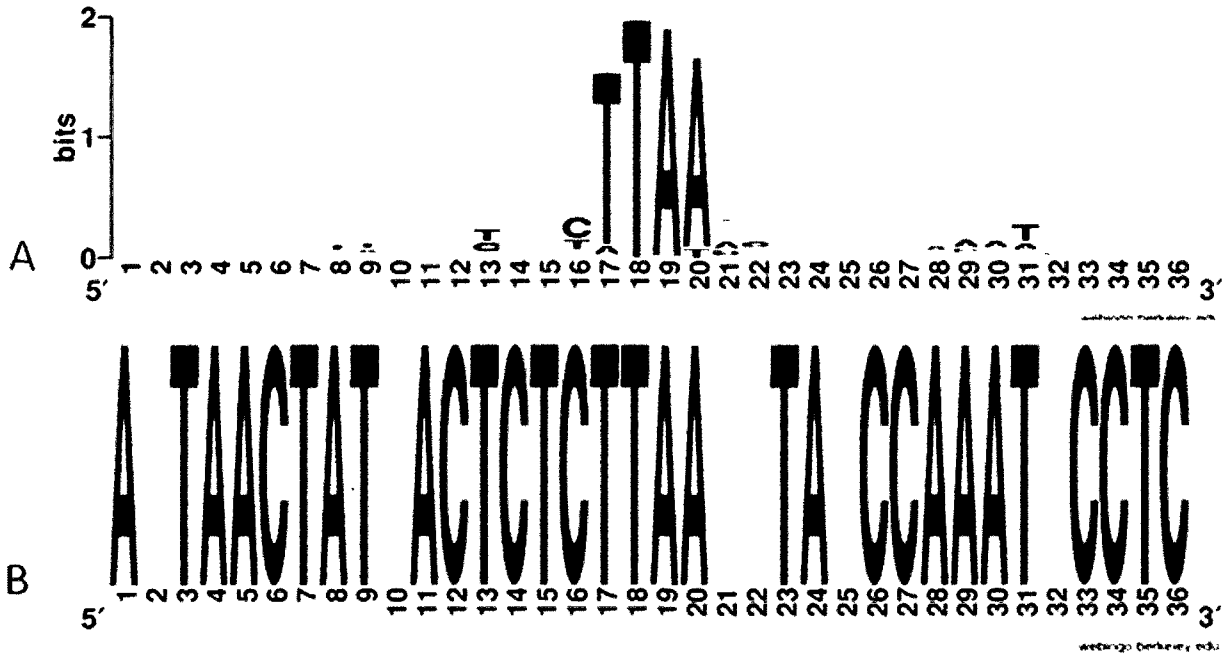

Figure 15. WebLogo output showing the consensus sequence of (a) 118 selected *Pokey* insertion sites from the *Daphnia pulex* genome and (b) the 28S rRNA gene target site. There is a weak but significant preference at positions 13, 16, 21, 29, 30, 31 and 32 for the nucleotides found in the 28S rRNA target site. Significant bases are those with letter heights higher than their corresponding error bars, which were calculated by the sample size correction of the WebLogo program (Crooks et al., 2004).

5′-GTCATGATTGTGAGGTCT**G**----------**TTAAC**GGTTTAGTGTTTTCTTAC-3′

5′-GTCATGATTGTGAGGTCT**G**----------G̲GGTTTAGTGTTTTCTTAC-3′

5′-GTCATGAT----------TTGTGGGACGGAGGGAGTA**AAC**GGTTTAGTGTTTTCTTAC-3′

Figure 16. Analysis of *Pokey* excision from pWL89A-Pok6.6NA. The *Hpa*I site, marked in bold, and flanking sequences from each revertant plasmid are shown. Excision footprints are underlined.

APPENDIX I

Molecular protocols

1.1 Gateway ® Cloning

The Gateway ® System (Invitrogen) of plasmids and enzymes was used to generate the transposase expression plasmids for use in the yeast excision assay. Protocols were modified slightly from the manufacturers:

BP Clonase II Reaction

-150 ng of att-tailed transposase secondary PCR product from overlap-extension PCR

-150 ng of pDONR-221 plasmid

- 1 μL of BP Clonase II enzyme mix

-add enough ultrapure water to adjust volume to 5 μL

-incubate for 1 hour at room temperature

The entire reaction was then transformed into DH5α (Invitrogen) or XL-1 Blue (Stratagene) chemically competent *E. coli* cells. When less than 50 μL of competent cells were used chilled (4°C) 100 mM $CaCl_2$ solution was added to adjust the volume to 50 μL to ensure proper efficiency of transformation. Cells were plated on selective media and colonies were screened via colony PCR using standard M13F and R primers. Fragments which were 1kb or larger were sequenced using a battery of *Pokey* ORF specific primers (Pok4065F, Pok4410F, Pok5026F, Pok4488R, Pok5985R) to ensure the fidelity of the

sequence. Colonies containing desired plasmids were grown up in selective liquid culture overnight and plasmid DNA was extracted.

Entry clone plasmid DNA was then subjected to the LR Clonase II reaction:

-150 ng of pDONR-Pok6.6ORF

-150 ng of pAG413GAL-ccdB

-1 μL of LR Clonase II enzyme mix

-added enough ultrapure water to adjust volume to 5 μL

-incubated for 1 hour at room temperature

The entire reaction was then transformed, plated and screen as before.

1.2 Plasmid DNA Digestion

Plasmid pWL89A was digested in a 50 μL volume of water as follows:

-5 μL of 10X #4 New England Biolabs Digestion Buffer

-5 μL of 10X Bovine serum albumen

-0.2 μL of *Hpa*I restriction endonuclease (New England Biolabs)

-2500 ng of pWL89A plasmid DNA

Digestion was carried out in a 37 °C water bath for 4 hours. *Hpa*I enzyme was removed from the reaction using the Micropure-EZ Enzyme Remover Kit (Millipore). Products were run on a gel to confirm digestion had taken place.

1.3 Ligation

Ligation of the Pok6.6NA fragment to digested pWL89A was carried out in the following 20 µL reaction:

-500-900 ng of Pok6.6NA amplified using a high fidelity PCR reaction and phosphate-labelled primers

-100-250 ng of pWL89A cut with *HpaI*

-1 µL of T4 DNA Ligase (New England Biolabs)

-2 µL of 10X T4 Buffer (50 mM Tris-HCl, 10 mM $MgCl_2$, 10mM DTT, 1mM ATP, pH 7.5 at 25 °C)

-adjust volume to 20 µL using water

The reaction was incubated overnight at 4 °C and the ligation reaction was terminated by heating at 65 °C for 10 minutes. 50-100 ng of ligated plasmid was transformed into chemically competent *E. coli.*

1.4 Yeast Transformation

Transformation of yeast strain DG2523 was carried out using a protocol modified from one provided by the Walhout Lab of the University of Massachusetts Medical School:

1) Cells derived from a single colony were scrapped with a toothpick and re-suspended directly in 75-100 mL of standard YEPD liquid media.

2) Liquid culture was incubated at 30 °C and shaken at 200 rpm for approximately 24-30 hours.

3) After 24 hours, 5 mL samples of culture were taken and the absorbance was measured at 600 nm on a Spectronic 20 Spectrophotometer (Bausch & Lomb). When an absorbance of between 0.4 and 0.6 was reached, relative to an un-inoculated YEPD standard sample, incubation was ceased.

4) 50mL of culture was aliquoted and centrifuged at 700g and room temperature for 5 minutes.

5) Supernatant was decanted and pelleted cells were re-suspended in 5mL of sterile water.

6) Cells were centrifuged for another 5 minutes at 700G and room temperature.

7) Supernatant was decanted and cells were re-suspended in 5mL of TE/LiAc solution (1.25 mL of 10X TE Buffer, 1.25 mL of 1M lithium acetate solution, 10 mL of sterile water) and centrifuged for a further 10 minutes at 700G and room temperature.

8) Supernatant was decanted a final time and cells were re-suspended in 25 μL of boiled salmon sperm DNA (10 mg/mL, boiled for 10 minutes) and TE/LiAc solution to adjust volume to 250 μL.

9) 50 μL of competent cell solution was used for each transformation reaction.

10) 50-100 ng of both donor and expression plasmids were aliquoted into 5 μL of sterile water and pipetted into the 50 μL competent cell solution.

11) 275 μL of TE/LiAc/PEG solution (150 μL of 10X TE Buffer, 150 μL of 1M lithium acetate solution, 1.2 mL of 50% polyethylene glycol solution) was added to each transformation reaction and cells were re-suspended gently.

12) Transformation reactions were incubated at 30 °C for 30 minutes to 1 hour.

13) Reactions were then heat shocked for 20 minutes in a 42 °C water bath.

14) Reactions were centrifuged for 1 minute at 7000G to pellet cells which were then re-suspended in 500 μL of sterile water.

15) Cells were then plated on SD-U-H media (Sunrise Scientific) to select for cells containing both expression and donor plasmids.

APPENDIX II

Sequence alignments of *Pokey* elements

The contents of this Appendix have been provided electronically on an accompanying CD

2.1 Figure 5 Data

Alignments of full length elements used to create the NJ tree in Figure 5 provided as both .txt files and .meg files. The intervening, un-alignable sequence between the conserved 5' and 3' ends of the elements has been removed. The original alignments of both the 5' and 3' ends separately have been provided as well.

2.2 Figure 6 Data

Alignments of the terminal 1600 bp from the 3' end of various elements from species in the subgenus *Daphnia* used to create the NJ tree in Figure 6, provided as both .txt and .meg files.

2.3 Figure 8 Data

Alignments of the nucleotide sequences of *Pokey* ORFs used to create the NJ tree in Figure 8, provided as both .txt and .meg files.

2.4 Figure 11 Data

Alignments of *Pokey* MITEs used to create the NJ tree in Figure 11, provided as both .txt and .meg files.

2.5 Figure 12 Data

Alignments of both full length and MITE elements used to create the NJ tree in Figure 12, provided as both .txt and .meg files. The intervening un-alignable sequence between the conserved 5' and 3' ends of all elements has been removed.