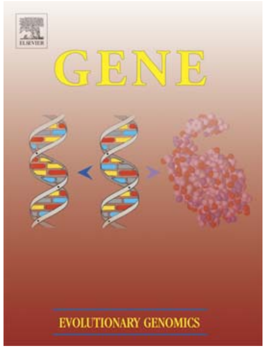


# *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses

Ellen J. Pritham, Tasneem Putliwala, Cédric Feschotte

PII: S0378-1119(06)00537-3
DOI: doi: 10.1016/j.gene.2006.08.008
Reference: GENE 35594

To appear in: *Gene*

Received date: 14 July 2006
Accepted date: 2 August 2006

Please cite this article as: Pritham, Ellen J., Putliwala, Tasneem, Feschotte, Cédric, *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses, *Gene* (2006), doi: 10.1016/j.gene.2006.08.008

# *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses

Key words: DNA polymerase B, c-integrase, transposon, adenovirus, Polinton, *Trichomonas vaginalis*

**Ellen J. Pritham*, Tasneem Putliwala and Cédric Feschotte**

**The University of Texas at Arlington, The Department of Biology, Arlington, TX, USA**

Abbreviations: aa amino acid(s), bp base pair(s), CMG conserved *Maverick* gene, ds double-strand(ed), *hAT hobo*/*Activator*/*Tam3*, kb kilobase(s) or 1000 bp, LTR long terminal repeat, *Mav Maverick*, Mb megabase(s) or 1 million bp, ORF open reading frame, ss single strand(ed), TE transposable element, TIR terminal inverted repeat, *Tlr Tetrahymena thermophila* long repeat, TSD target site duplication.

*Corresponding author
Ellen J. Pritham
Department of Biology
Box 19489
The University of Texas at Arlington
Arlington, TX  76019
phone 817-272-0981
fax      817-272-2855
pritham@uta.edu

## Abstract

We previously identified a group of atypical mobile elements designated *Mavericks* from the nematodes *Caenorhabditis elegans* and *C. briggsae* and the zebrafish *Danio rerio.* Here we present the results of comprehensive database searches of the genome sequences available, which reveal that *Mavericks* are widespread in invertebrates and non-mammalian vertebrates but show a patchy distribution in non-animal species, being present in the fungi *Glomus intraradices* and *Phakopsora pachyrhizi* and in several single-celled eukaryotes such as the ciliate *Tetrahymena thermophila*, the stramenopile *Phytophthora infestans* and the trichomonad *Trichomonas vaginalis*, but not detectable in plants. This distribution, together with comparative and phylogenetic analyses of *Maverick*-encoded proteins, is suggestive of an ancient origin of these elements in eukaryotes followed by lineage-specific losses and/or recurrent episodes of horizontal transmission. In addition, we report that *Maverick* elements have amplified recently to high copy numbers in *T. vaginalis* where they now occupy as much as 30% of the genome. Sequence analysis confirms that most *Mavericks* encode a retroviral-like integrase, but lack other open reading frames typically found in retroelements. Nevertheless, the length and conservation of the target site duplication created upon *Maverick* insertion (5 or 6-bp) is consistent with a role of the integrase-like protein in the integration of a double-stranded DNA transposition intermediate. *Mavericks* also display long terminal inverted repeats but do not contain ORFs similar to proteins encoded by DNA transposons. Instead, *Mavericks* encode a conserved set of 5 to 9 genes (in addition to the integrase) that are predicted to encode proteins with homology to replication and packaging proteins of some bacteriophages and diverse eukaryotic double-stranded DNA viruses, including a DNA polymerase B homolog and putative capsid proteins. Based on these and other structural similarities, we speculate that *Mavericks* represent an evolutionary missing link between seemingly disparate invasive DNA elements that include bacteriophages, adenoviruses and eukaryotic linear plasmids.

## 1. Introduction

Most eukaryotic genomes harbor a vast amount of interspersed repetitive DNA, which mostly consists of transposable elements (TEs) or their remnants. The rapidly increasing amount of DNA sequences accumulating in the public database and the completion of many genome sequencing projects have contributed to reveal an extraordinary diversity of TEs, in terms of their structure, survival strategies and dynamics of amplification (e.g. Lander et al., 2001; Feschotte et al., 2002; Kapitonov and Jurka, 2003; Brookfield, 2005; Hua-Van et al., 2005). It has become clear that deciphering the origin and biology of these elements is an essential facet of genomic research, the value of which is not just limited to the assistance provided to sequence assembly and annotation, but is also critical to understanding how genes and genomes evolve.

Traditionally, transposable elements have been divided into 2 classes based on their mechanism of transposition (Finnegan, 1989; Capy et al., 1998). Class 1 elements (or retroelements) transpose through reverse-transcription of an RNA intermediate, while class 2 elements (or DNA transposons) move directly through a DNA intermediate. Retrotransposition involves relatively complex enzymatic machinery and often requires the sequential action of multiple proteins such as reverse transcriptase, endo- or ribonucleases and integrase (Craig et al., 2002). In contrast the transposition of most eukaryotic DNA transposons requires a single element-encoded protein called transposase, which catalyzes all the steps of the so-called cut-and-paste reaction (Craig et al., 2002). Retrotransposons are further divided into 2 major types based on their structure and transposition mechanism. Non-long terminal repeat (non-LTR) retroelements (LINEs and SINEs) transpose through a target-primed mechanism where integration and reverse-transcription is coupled and initiation occurs via a single-strand nick at the insertion site (Luan et al., 1993). In contrast, reverse-transcription of LTR retroelements (or retroviral elements) occurs in retroviral particles and precedes integration (Voytas and Boeke, 2002; Curcio and Derbyshire, 2003).

The integration mechanism of LTR retroelements and retroviruses is very similar to those of DNA transposons (Haren et al., 1999; Curcio and Derbyshire, 2003). Indeed, the catalytic domains of many integrases and transposases display significant similarities at the primary sequence level and adopt a very similar tridimensional fold (Capy et al., 1996; Haren et al., 1999; Rice and Baker, 2001; Hickman et al., 2005). Therefore, it is likely that at least a subset of retroviral integrases and transposases share a common ancestor (Capy et al., 1996). This observation led several authors to hypothesize that an ancestral LTR retrotransposon evolved from a non-LTR retrotransposon by assimilation of a nested DNA transposon (Capy et al., 1998; Malik and Eickbush, 2001; Eickbush and Malik, 2002). A subsequent step in the evolution of LTR retroelements is their occasional transition into infectious retroviruses by acquisition of envelope genes from multiple diverse viral sources (Temin, 1980; Malik et al., 2000; Eickbush and Malik, 2002). On the other hand, there is some evidence that retroviruses may also subsequently lose their infectious capacities (i.e. envelope gene) and go back to an intracellular life style (Lerat and Capy, 1999; Gifford and Tristem, 2003; Herve et al., 2004; Yano et al., 2005).

The traditional dichotomy of transposable elements (class 1 vs. class 2) has been recently challenged by the discovery of atypical mobile elements, which share no sequence or structural similarities to previously described class 1 or class 2 elements, and thus may define entirely new classes of TEs. For example, a group of transposons called *Helitrons* with common structural and coding capacities were recently identified in the genomes of plants, animals and fungi (Kapitonov and Jurka, 2001; Poulter et al., 2003; Hood, 2005). *Helitrons* are apparently related to genetic elements that replicate through a rolling-circle mechanism, including the circo- and geminiviruses, a group of single-stranded DNA (ssDNA) viruses that infect plant and animals. Akin to the reversible switch from retrotransposons to retroviruses, it has been proposed that *Helitrons* originated from and/or gave rise to geminiviruses (Feschotte and Wessler, 2001; Kapitonov and Jurka, 2001; Murad et al., 2004).

Together these studies tend to blur the distinction between transposable elements and viruses and suggest that there are frequent exchanges of functional modules and domains between TEs and viruses. This could reflect similar evolutionary trajectories and, in some cases, the common ancestry of TE and viruses. Recently we identified a new assemblage of large transposable elements, called *Maverick*s (Feschotte and Pritham, 2005), that share structural and sequence similarity to *Tlr1*, an atypical group of mobile element described from the ciliate *Tetrahymena thermophila* (Wells et al., 1994; Wuitschick et al., 2002). *Maverick* elements were initially identified in the nematodes *Caenorhabditis elegans* and *C.*

*briggsae* and in the zebrafish *Danio rerio* and we proposed that, together with *Tlr1*, they define an entirely new class of TEs (Feschotte and Pritham, 2005). More recently, Kapitonov & Jurka (2006) described in more detail some of these *Mavericks* as well as additional related elements that they identified in several eukaryotic genomes. They refer to this group of elements as *Polintons* collectively and proposed a model for the origin and transposition mechanism of these unusual transposons (Kapitonov and Jurka, 2006).

In the present study, we further expand the distribution of *Mavericks* in eukaryotes and offer a detailed characterization of their structure and coding capacity. We also provide evidence that the parabasalid protozoa *Trichomonas vaginalis* has recently experienced an explosive amplification of *Mavericks*, which now appear to contribute up to one third of its genome. Finally, we highlight the unique characteristics of *Mavericks* among endogeneous mobile elements and their striking similarities to an anciently related group of invasive DNA that includes adenoviruses, bacteriophages and eukaryotic linear plasmids.

## 2. Materials and Methods

### 2.1 Computational detection of *Mavericks*

Candidate *Mavericks* were identified by homology-based searches at NCBI, beginning in April of 2005 until final preparation and submission of this manuscript. All sequence data that has been deposited in the NR, HTGS and WGS databases at NCBI during this period were subjected to multiple searches using the various BLAST tools (Altschul et al., 1990; McGinnis and Madden, 2004). The queries used included the DNA or putative protein sequences derived from previously identified *Maverick* elements (Feschotte and Pritham, 2005). Generally, a hit was considered significant when the *e*-value was lower than $10^{-4}$. *Maverick* sequences reported in this article are available upon request to the authors. Accession numbers from the various databases and the nucleotide coordinates of complete elements are reported in Table 1; Supplemental Table 1. Additional *Mavericks* copies were identified using BLAT (Kent, 2002) against genomes available through the UCSC genome browser (http://genome.ucsc.edu/). Due to the relatively large size of complete *Maverick* elements (~15-30 kb), many contigs could be identified that contained incomplete *Mavericks*, that is elements lacking one or both of their extremities. We considered a contig as containing an incomplete *Maverick* when at least two putative conserved *Maverick* protein sequences were identified through tblastn searches within 10 kb of each other.

### 2.2 *Maverick* annotation and phylogenetic analyses

Open reading frames (ORFs) were identified through conceptual translations with the program Translate (www.expasy.org/tools/dna.html) or with ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). Putative introns were identified using NetGene2 (www.cbs.dtu.dk/services/NetGene2/) and when possible confirmed through alignments with matching cDNA and EST sequences. Sequences were aligned with ClustalW or TCoffee using default parameters and alignments were refined manually using GeneDoc (Thompson et al., 1994; Nicholas et al., 1997; Notredame et al., 2000). When necessary, frameshifts were judiciously intriduced according to nucleotide alignments of closely related sequences. Phylogenetic trees were generated with MEGA v.3 using the neighbor-joining method with poisson correction or equal input allowing for multiple substitutions at sites and allowing for a variation in rates between branches (Kumar et al., 2004). For phylogenetic analyses, representative sequences were included that necessitated only minor corrections in the ORF, for example they did not have large insertions or deletions or were not incomplete as an artifact of cloning, sequencing or genome assembly. The function of *Maverick*-encoded proteins was predicted by homology to proteins of known function, by the presence of conserved domains identified through CDD search (Marchler-Bauer et al., 2005) and by protein threading (Kelley et al., 2000) through the analysis of conserved protein folds using Phyre (http://www.sbg.bio.ic.ac.uk/phyre/).

### 2.3 Identification of paralogous ‘empty’ sites

To illustrate the mobility of *Mavericks*, paralogous sites (empty sites) devoid of the insertion were queried. To identify empty sites, searches were completed using the sequences flanking each side of the insertion as a query using either blastn or BLAT. An empty site is reported when another region is identified in the same genome that lacks the insertion yet contains the unduplicated target site.

## 3. Results and Discussion

### 3.1 Identification and distribution of *Mavericks* across the tree of life

Recently, we showed that non-mammalian c-integrases from the genomes of the nematodes *C. elegans* and *C. briggsae* and in the zebrafish are carried by large transposable elements with terminal-inverted repeats, called *Mavericks* (Feschotte and Pritham, 2005). In order to conduct a more comprehensive analysis of the structure and distribution of these elements, we used the *Mavericks* from *C. elegans* and *D. rerio* as queries in BLAST searches of the nr, WGS and HTGS Genbank databases. These searches confirmed the presence of *Mavericks* in *C. briggsae* and *Takifugu rubripes* as indicated by the initial study of c-integrases (Gao and Voytas, 2005), but also suggested that their occurrence was more widely distributed than previously anticipated (Table 1).

In addition to the elements previously reported from *C. elegans*, *C. briggsae* and *D. rerio* (Feschotte and Pritham, 2005), we could identify complete *Mavericks* (as defined by elements with TIRs and flanked by a 5 or 6-bp putative TSD) in twelve additional species from diverse eukaryotic phyla. These include a wide spectrum of invertebrate and vertebrate animals, as well as the mychorrizal fungus *Glomus irradices*, the oomycete *Phytophthora infestans* and the parabasalid *Trichomonas vaginalis* (Table 1, Fig. 1). BLAT searches (Kent, 2002) using the TIRs of individual *Maverick* elements against their respective draft genome sequence (when available) allowed the identification of additional copies in all of these species.

With the notable exception of *Trichomonas vaginalis* (see Section 3.5), *Maverick* elements occur as relatively low copy number families (3-50 copies per genome) that share between 85-99% nucleotide similarity between copies of the same family. In most species, very distinct families apparently co-exist within the same genome. For example, *Mav_Pd1.1* and *Mav_Pd2.1* from the annelid *Platynereis dumerilii* (see Supplemental Table 1) cannot be confidently aligned at the nucleotide level, but both are clearly *Maverick* elements sharing conserved structural features and encoding distantly related genes (data not shown). Members of highly divergent *Maverick* families are also apparent in *C. elegans*, *Strongylocentrus purpuratus*, *Oikopleura dioica* and *Danio rerio* (see Fig. 1 and Table 1).

It should be noted that elements closely related to those reported here for *T. castaneum*, *X. tropicalis*, *S. punctata* and *G. irradices* were also independently identified by Kapitonov and Jurka and recently described as *Polintons* (Kapitonov and Jurka, 2006). However, the elements that we identified in *S. purpuratus* and *T. vaginalis* are only distantly related to those described in these species by Kapitonov and Jurka and presumably belong to divergent families co-existing within the same genome. *Maverick* elements from *C. remanei*, *P. dumerilii*, *Drosophila persimilis*, *D. ananassae*, *O. dioica* and *Phytopthora infestans* have not been reported elsewhere. Since we previously coined the name *Mavericks* for the first elements discovered in nematodes and zebrafish and predicted in this study that these *Mavericks* would delineate an entirely new group of TEs, we refer hereafter to all related elements as *Maverick*s (Feschotte and Pritham, 2005). One exception is *Tlr1* from *Tetrahymena thermophila*, whose composite sequence was originally assembled from several overlapping genomic clones, but yet appears to represent an incomplete copy of a *Maverick*-like transposon family (see below) (Wells et al., 1994; Wuitschick et al., 2002).

BLAST searches of the databases also led to the identification of protein-coding DNA segments that likely represent internal regions of *Maverick* elements in other eukaryotic species, but for which it was not possible to unambiguously identify the termini. This was either because of gaps in the genome assembly, because the assembled contigs themselves are smaller than a typical *Maverick* (~15-20 kb) or due to secondary TE insertions or other recombination events truncating and rearranging the elements. Nevertheless, when significant hits to multiple conserved *Maverick* genes (described below) were detected in the same genomic neighborhood (i.e. at least two genes less than 10 kb of each other), we considered the segment to be of likely *Maverick* origin. In this way, it can be established that *Mavericks* have also colonized the genomes of the hagfish *Eptatretus stoutii*, the three-spined stickleback *Gasterosteus aculeatus*, the pufferfishes *Takufugu rubripes* and *Tetraodon nigroviridis*, the fruit flies *D. grimshawi*, *D.* melanogaster, *D. pseudobscura, D. virilis*, *D. willistoni* and *D. yakuba*, the sea squirt *Ciona intestinalis*, the plant pathogenic fungi *Phakopsora pachyrhizi* and the ciliate *Tetrahymena thermophila* (Table 1).

To complete this overview of *Maverick* distribution, it should also be added that sequences clearly related to *Maverick*-encoded integrases (see Section 3.2) could be identified in numerous other animal species, including the aphid *Acyrthosiphon pisum,* ratite bird *Apteryx australis mantelli* and the cnidarians *Hydractinia echinata* and *Nematostella vectensis*, suggesting that *Maverick*s are also present in the genomes of species from the phyla Hemiptera, Aves and Cnidaria. However, these sequences were either from very short genomic clones (ratite) or from EST projects (aphids, cnidarians) and therefore could not be physically associated with other *Maverick* genes.

Together, these data provide evidence that *Maverick*s are present in a very broad range of eukaryotes, albeit with an intriguingly patchy distribution. For example, no *Maverick* elements or *Maverick*–related genes could be identified in any plant species despite the presence of large amounts of genomic data for several plant species.  Also we could not detect any *Maverick*-related sequences in the now complete or nearly complete genome sequences of the mosquito *Anopheles gambiae*, the bee *Apis melifera* or the silkworm *Bombyx mori*, although *Maverick*s can be readily detected in virtually all other insect and invertebrate genomes reasonably represented in the databases.  Similarly, there are no detectable *Maverick*s in the impressive amount of mammalian genomic sequences currently available in the databases, besides a distantly related integrase gene, possibly derived from an ancient *Maverick* element (Feschotte and Pritham, 2005; Gao and Voytas, 2005).  Finally, *Mavericks* are also poorly represented in fungi as they only could be detected in *Glomus intraradices*, a mychorrizal fungus and *Phakosora pachyrhizi*, a plant pathogenic fungus but not in any of the other 50 fungal genome sequences completed or nearing completion (mostly ascomycetes and basidiomycetes, see http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=fungi). We believe that this broad but erratic distribution reflects an ancient origin of *Maverick*s, and their differential success at adapting to different host species and/or a propensity for stochastic loss during evolution.  A less parsimonious yet not mutually exclusive alternative is that the current distribution of *Maverick*s was shaped by repeated horizontal invasions of only a subset of already differentiated eukaryotic lineages, followed by vertical diversification within these lineages.

### 3.2 Phylogenetic analysis of the *Maverick* c-integrase

The c-integrase gene family, was first reported by Gao and Voytas (2005), as encoding a monophyletic group of cellular integrases possibly co-opted for host function. They showed that the c-integrase proteins from nematodes and fish contain a ~150-aa region with significant similarity to the catalytic core domain of retroviral integrases (RVE) and harbor a chromodomain at their C-terminus. The chromodomain is also found at the C-terminus of integrases from the chromoviruses, a subgroup of *Ty3/gypsy* LTR retrotransposons, where it modulates integrase activity and might be involved in the targeting of these elements to particular chromatin environments (Malik and Eickbush, 1999; Hizi and Levin, 2005; Nakayashiki et al., 2005).  We subsequently showed that c-integrases from nematodes and zebrafish are in fact encoded by *Maverick* elements and likely involved in their chromosomal integration (Feschotte and Pritham, 2005). Annotation of the newly identified *Mavericks* confirm that almost all of them contain a gene highly similar to the previously described c-integrase (Fig. 1), which can encode a ~400--aa protein with both RVE and chromodomains.

To determine the relationship of the *Maverick* c-integrases both to each other, as well as to the integrases encoded by other elements, the RVE domain from representative *Mavericks* was aligned to those present in the ciliate *Tlr1*, slime mold *Tdd4*, budding yeast *Ty1* and Drosophila *copia* elements and the alignment was used for phylogenetic reconstruction with the neighbor-joining method (Fig. 2).  When the tree is rooted with the retroviral sequences *Ty1* and *copia*, the *Maverick* c-integrases group with the integrases of *Tlr1* and *Tdd-4*, forming a monophyletic clade reasonably well supported by bootstrap analysis (82%).  The most basal position in this clade is occupied by the c-integrase of *Mav_Tv1.1* from the parabasalid *T. vaginalis*. This topology indicates that both *Tlr1* and *Tdd-4* integrase proteins share a more recent common ancestor with *Maverick* integrases than with the *Ty1*/*Copia* group and probably should be considered part of the c-integrase family, as was suggested by the initial study of Gao & Voytas.

The most basal position of the *T. vaginalis* c-integrase in the RVE phylogeny (bootstrap value 53%) is consistent with the deep-branching position of this species in reconstructions of the eukaryotic tree of life (Baldauf et al., 2000; Horner and Embley, 2001) and the hypothetical status of parabasalids as one of the most primitive group of eukaryotes. The detached and similarly basal position of the c-integrases from the other protozoan eukaryotes *P. infestans* and *T. thermophila* relative to a (weakly supported) monophyletic group of c-integrases from animal, fungi and *D. discoideum* is also reminiscent of the position occupied by these species in phylogenetic analyses of large nuclear and mitochondrial gene sets (Bapteste et al., 2002; Lang et al., 2002; Steenkamp et al., 2006). This data provides support to the idea that *Mavericks* are very ancient components of the eukaryotic genome that have been vertically inherited and diversified for over a billion years of evolution.

A more surprising finding in our c-integrase RVE phylogeny is the highly supported grouping (100% bootstrap value) of the mycorrhizal fungus with the two cnidarian species as a sister clade to all other animal sequences. Cnidarians are thought to be the most primitive extant animals so it is intriguing that they would group with a fungus rather than with the other animals. To our knowledge, there is no obvious biological link between cnidarians and mycorrizhal fungi that could indicate a case of horizontal transfer between these organisms and explain this unexpected topology. In light of the well-established monophyly of the opisthokonts (animal and fungi), the simplest explanation for this grouping is that it directly descends from an ancient lineage of *Mavericks* present in the common ancestor of fungi and animals.

Phylogenetic resolution is limited among the other animal sequences, although there is high bootstrap support for groupings of the four different nematode species and of the teleost fishes and weak support for grouping of the teleosts with the clawed frog and the ratite bird (Fig. 2). In sum, the topology of the c-integrase tree appears largely congruent with the established relationships of its diverse eukaryotic host species, which we interpret as evidence that *Maverick* elements are very ancient components of these genomes and probably originated prior to the eukaryotes.

### 3.3 Other conserved *Maverick* genes

To gain further insight into the origin and mode of transposition of *Maverick* elements, we analyzed the coding capacity of representative *Mavericks* using BLAST searches of the protein databases, conserved motif scans, and other gene discovery tools available through the NCBI server (Wheeler et al., 2005); as well as tri-dimensional fold prediction by protein threading (Kelley et al., 2000). These combined analyses revealed that most *Mavericks* harbor multiple ORFs, which constitute a set of four to nine predicted protein-coding genes (Fig. 1). In addition to c-integrases, 5 distinct groups of conserved coding sequences were detected in *Maverick* elements from at least three distant species (Fig. 1). Reiterated and reciprocal BLAST analyses (Tatusov et al., 2003) showed that within each group of conservation, the *Maverick* putative proteins are closer to each other than to other non-*Maverick* eukaryotic proteins. We therefore refer to these as clusters of 'conserved *Maverick* genes' (CMGs, Fig. 1, Supplemental Table 2) and provide below a description of their coding potential. Additionally, numerous other putative proteins were identified in only one *Maverick* element or only from *Mavericks* from a single species interestingly these proteins often displayed homology with known viral proteins (Fig. 1)

**CMG1 encodes a DNA polymerase B-like protein.** CMG1 is predicted to encode a large protein (~1,100-1,500-aa) detected in all *Maverick* families examined (Fig. 1). CMG1 proteins from different *Maverick* elements have variable degrees of conservation, which, as for the c-integrases, are roughly congruent with the taxonomic distance of their host species. For example, the CMG1 putative proteins from two different nematode elements *Mav_Ce1.1* and *Mav_Cb1.2* are 70% identical (80% similar) and each is ~45% identical (62% similar) to the CMG1 protein of the beetle *Mav_Tc1* and ~25% identical (37% similar) to the CMG1 protein from the zebrafish *Mav_Dr1.2*.

Despite broad variation in their level of conservation, CMG1 proteins contain a ~400-aa region that is consistently matching the complete DNA_pol_B_2 domain (PFAM accession **03175.10**) in searches of the conserved domain database (Fig. 3). This domain is characteristic of the type-B family of viral DNA polymerases. Furthermore, BLAST searches with CMG1 proteins return best hits (other than other CMG1 proteins) with protein-primed DNA polymerases encoded by phages from the tectiviridae (e.g. Bam35c) and podoviridae (e.g. phi29) families, adenoviruses and linear plasmids found in the cytoplasm and mitochondria of some plants, fungi and stramenopiles. For example, the *Mav_Gi1.1* CMG1 aligns with a protein from a *Zea mays* mitochondrial linear plasmid (GenBank accession **X02451.1**; *e*-value 2e=-04, 40% similarity over 452-aa). In line with the BLAST results, the DNA_pol_B_2 domain identified in CMG1 proteins could be confidently aligned to those of protein-primed polymerases (Fig. 3), but not to DNA- and RNA-primed DNA polymerases (data not shown). Type-B family viral

DNA polymerases typically display DNA binding, polymerase and 3'-5' exonuclease activities (Knopf, 1998). A partial alignment of the pol_B_2 domain of *Mav_Cr1.1* CMG1 with those of a maize linear plasmid, the PZA bacteriophage and human adenovirus 17 reveals that the *Maverick* CMG1 protein contains all of the critical residues involved in these functions (Fig. 3).

CMG1 are found in multiple copies within the genomes of the species that harbor *Maverick* elements and are almost always found in proximity of c-integrase genes (Fig.1). CMG1 and c-integrases generally occupy the most outer regions of *Maverick* internal sequences (except in *Mav_Tc1*), with a tail-to-head orientation in invertebrates and in *X. tropicalis* and a tail-to-tail orientation in all other chordate species (Fig. 1). Despite these variations, the data indicate that a DNA polymerase-like gene is an integral component of *Maverick* elements and that the resulting protein may be involved in their propagation. Interestingly, *Maverick* DNA polymerase-like genes were identified in close proximity to genes highly similar to those described in the *Tlr* elements from *T. thermophila*, suggesting that complete *Tlr* elements likely include a DNA polymerase B-like gene and therefore are likely to belong to the *Maverick* superfamily.

In addition to the pol_B_2 domain, several *Maverick* CMG1 proteins also contain regions with significant similarity to other domains including VSR, Smc and PolBc (Supplemental Table 2). In the case of the VSR (very short repair endonuclease), this domain is actually embedded within the DNA_pol_B_2 domain. The significance and function of a putative endonuclease activity for *Maverick* transposition is unclear, although it could be involved in initiating replication by single-strand nicking at the *Maverick*-host DNA junction.

**CMG2/3 encode coiled-coil domain proteins.** CMG2/3 are predicted to encode proteins that range in size from ~200 to ~900 amino acids. Most products have only weak or no significant similarities with eukaryotic proteins of known function and therefore they are currently annotated as unknown or hypothetical proteins in the respective sequenced genomes (if annotated at all). The proteins predicted from CMG2 and CMG3 genes typically harbor one to four coiled-coil domains and are only weakly related to each other (data not shown). In general the coiled coil domain functions in enabling protein-protein interactions, allowing oligomerization and/or association between disparate proteins. In light of the possible infectious capacity of *Maverick*s (see Section 3.6), it is notable that coiled coil proteins are frequently encoded by viruses and involved in many viral functions such as capsid formation (Ma et al., 2004), switch from latent to lytic infection (Zhang et al., 1994) and cell recognition (Plisson et al., 2005). For viruses and also for many parasitic unicellular eukaryotes, coiled coil proteins often have a critical role in binding and recognizing the host cell membrane (Werner et al., 1998; Kostyuchenko et al., 1999; Rayner et al., 2004). Like many of these known coiled coil proteins, the *Maverick* coiled coil domain proteins include regions rapidly diverging among or even within *Maverick* families (data not shown), suggesting that they may also have a role in host cell recognition and infectivity and/or a structural role such as viral proteins involved in capsid formation (see also Section 3.6). Further experiments will be necessary to determine the role of these proteins in *Maverick* transposition if any.

**CMG4 encodes a putative ATPase.** The CMG4 putative proteins display similarities to hypothetical or known proteins encoded by various dsDNA viruses, most of which are annotated as packaging ATPases. According to *e*-values in BLASTX searches, hits ranged from strong (*e*-value <$10^{-50}$) to weak significance (*e*-values ~$10^{-2}$-$10^{-3}$) depending on the queries, but they generally encompass large coding regions (~100-300-aa) and they are repeatedly obtained with different *Maverick* queries (details on these hits are given in Fig. 1; Supplemental table 2). A multiple alignment of some *Maverick* CMG4 proteins with representative viral ATPases confirm this relationship and the presence of the most conserved walker A and B motifs characteristic of this class of ATPases (data not shown). It should be noted that the *Tlr1* element from *T. thermophila* was also found to contain one ORF (Tlr6R) with significant similarity to ATPases involved in viral DNA packaging (Wuitschick et al., 2002) and it too can be aligned with CMG4 proteins.

**CMG5 encodes a protein homologous to adenoviral cysteine protease.** The CMG5 proteins are predicted to be between 170 and 180 amino acids in length and BLAST searches revealed that they are closely related to the cysteine proteases encoded by adenoviruses, the *Acanthamoeba polyphaga* mimivirus and the African swine fever virus. For example, the *Mav_Dr1.1* CMG5 (178-aa) is 29% identical and 50% similar over 128-aa to the protease encoded by the ovine adenovirus D (**NP_659524.1**; 201-aa). CMG5 proteins shares all of the residues that form the catalytic triad of cysteine proteases as well as the glutamine residue predicted to create the oxyanion hole in the active site (Fig. 4) indicating that CMG5 likely acts as a bonafide cysteine protease.

Most large viruses encode at least one protease, independent of whether their genomes are single- or double-stranded or composed of RNA or DNA (DiMaio and Coen, 2001). The proteases are utilized in a variety of functions. RNA viruses depend on a protease to cleave the polyproteins that are translated from in single transcription unit (Fitzgerald and Springer, 1991), while adenoviruses utilize their protease to cleave the preterminal protein (pTP) from the DNA polymerase B, which is necessary for the replication initiation of its linear chromosome as well as a trigger of other related events (Webster et al., 1994). Since the CMG1 proteins of *Mavericks* is most closely related to protein-primed DNA polymerase B, it is tempting to speculate that CMG5 functions in a similar manner to cleave a preterminal protein bound to the end of a *Maverick* replication intermediate.

### 3.4 Evidence for recent chromosomal integration of *Mavericks* likely mediated by their integrase

Most integrase and transposase enzymes are known to induce a staggered double-stranded cut in the host target DNA prior to TE insertion. Repair following TE integration results in a duplication of the target site (target site duplication, TSD). The size and sometimes the sequence of this TSD is a conserved property of the group of TE to which the enzyme belongs. For example, members of the *hAT* superfamily are flanked by an 8 bp TSD (Kunze and Weil, 2002) and most LTR elements are flanked by a 5 bp TSD. A comparison of the DNA flanking the insertion to an ‘empty’ paralogous insertion site reveals the presence of the TSD. In addition, the presence of an empty paralogous site provides direct evidence of the past mobility of the TE. Previously, we reported such empty paralogous sites for *Mavericks* identified in *C. briggsae* and *D. rerio* (Feschotte and Pritham, 2005). Here we present additional empty sites for *Mav_Cb1.5* from *C. briggsae*, for *Mav_Xt1.3* and *Mav_Xt1.8* from *X. tropicalis*, and for *Mav_Tv1.21* and *Mav_Tv1.23* from *T. vaginalis* (Supplemental Fig. 1). The empty sites are part of other repeats in the corresponding genome and the *Maverick* insertion occurred in just one of the repeats. In all five instances, comparison of the repeat copies with and without the

*Maverick* insertion showed that a 5 or 6-bp flanking sequence was present in only one copy in the pre-integration site and was duplicated upon insertion of the *Maverick* element (Supplemental Fig. 1). Together, these data demonstrate that integration of *Maverick* elements induces a TSD of conserved length, a hallmark of TE integration.

The size of the *Mavericks* TSD (5 or 6-bp) is comparable to the size of the TSD typical of other mobile elements that utilize an integrase for chromosomal insertion (most and LTR retroelements and retroviruses). For example, murine retroviral elements of the ET/MusD family induce a 6-bp TSD and insertion of most copia-like LTR retrotransposon provoke a 5-bp TSD. Together, this data strongly suggests that the integration of *Maverick* elements into the host chromosome is likely mediated by the *Maverick*-encoded integrase.

### 3.5 Recent explosive amplification of *Maverick* elements in the *T. vaginalis* genome

While most *Maverick* elements belong to low or moderate copy number families in their respective genome, several lines of evidence indicate that they have recently proliferated to high copy numbers in the genome of the parabasalid protozoa *T. vaginalis*. First, a blastn search using the first 225-bp of the TIR of *Mav_Tv1.1* as a query against a database predicted to cover the complete *T. vaginalis* genome (177 Mb) yielded 2298 hits with *e*-values < $10^{-4}$ (>84% identical over at least half of the query). Since this search utilized a region of the element that is expected to occur twice per element this result indicated the presence of 1149 closely related elements (2298/2). However, because this search was done at the DNA level using a non-coding region that is expected to evolve faster than *Maverick* genes, it is likely to represent an underestimate of total *Mavericks* in the genome of *T. vaginalis*.

To identify *Mavericks* that might be more distantly related we utilized tblastn searches with multiple queries representing different well-conserved regions of various *T. vaginalis* CMGs (see Fig. 5; see Supplemental table 3 for queries used). Two tblastn searches with ~ 150-aa queries representing two non overlapping well conserved regions of the DNA polymerase B protein domain yielded 1,312 and 1,304 hits respectively with *e*-

value < $10^{-4}$. Further intersection and inspection of these hits confirmed that they belong to multiple *Maverick* families co-existing within the *T. vaginalis* genome and not to other types of mobile elements (data not shown). This suggests that the genome of *T. vaginalis* is likely to include ~1,300 DNA polymerase encoding *Mavericks*. Since, only 5 of the 10 *T. vaginalis* complete *Mavericks* that we identified encode a DNA polymerase, this might also be an under representation of the total number of *Mavericks* in the genome (Fig. 5).

Additional searches were therefore carried out with similar sized queries representing the most conserved region of the c-integrase and the most conserved region of the hyper variable protein, which are present in more than half of the *Mavericks* that we identified in the *T. vaginalis* genome. These searches yielded 3,362 and 3,809 hits, respectively. Based on the size of the queries, on the genomic coverage of the database and on the location of the corresponding genes in *Maverick* elements (see Fig. 5) these results indicate that this species might harbor over 3,000 *Mavericks*. Considering an average size of *Mavericks* in *T. vaginalis* of 15 to 20 kb (see below), one can infer that *Maverick* elements occupy an enormous fraction of the haploid genome, possibly as much as 60 Mb or one third of the genome. To our knowledge, this makes *Mavericks* the most prominent component of the *T. vaginalis* genome and the most prevalent TE family ever described in a single-celled eukaryote.

Additional evidence for the recent activity of *Mavericks* in *T. vaginalis* includes the discovery of a *Maverick* insertion embedded within a *Mar1 Mariner* element (Supplemental Fig. 1). This family of *Mariner* elements was reported to display all the hallmarks of an active transposable element and to have recently amplified to hundreds of copies in the genome of *T. vaginalis (Silva et al., 2004)*. Discovery of a *Maverick* element within a *Mar1* suggests that the *Maverick* element was active more recently or at least around the sametime as the *Mar1* family.

In order to characterize the *Mav_Tv1.1 family* in more detail all the complete closely related were identified using the *Mav_Tv1.1* copy as a query. To our surprise, we could only identified 11 elements closely related to *Mav_Tv1.1* with two clear termini (e.g. contained within the same contig and flanked by a 5-bp TSD), despite the fact that the equivalent in sequence to a 7.2 times coverage of the genome is in the database. The isolated *Mav_Tv1* elements range in size from 8.2 to 22.2 kb (average 16.2 kb) and share over 94% nucleotide similarity in pairwise comparisons (Fig. 5). All of the elements except *Mav_Tv1.1* contain multiple sequence gaps of unknown length, so their sizes may be underestimated. In addition, *Mav_Tv1.1*, which contained no gaps and is the largest element, did not contain all the proteins predicted to be important for transposition (Section 3.6). Thus, full-length and autonomous *Maverick_Tv1* copies would likely be over 22 kb long. This may in part explain why elements with two ends are difficult to identify in the current assembly of the *T. vaginalis* genome. Indeed 70% of the contigs are less than 10 kb in length (data not shown), and are therefore probably much shorter than the elements themselves. Due to their highly repetitive nature, high level of sequence similarity and large size, it is likely that *Maverick* elements represent a major burden for the correct assembly of the *T. vaginalis* genome sequence. Indeed, we observed that a very large number of contig ends and gaps coincide within *Maverick*-related sequences.

### 3.6 *Mavericks* from *Trichomonas vaginalis* potentially encode many additional proteins

To determine the complete set of possible proteins encoded by this family of *Trichomonas Maverick* elements, every ORF larger than 200 residues (starting with the first methionine) was identified for each element (see Fig. 5). For simplicity, we consider each unique ORF as potentially encoding a different protein. However, it remains possible that some ORFs could be transcribed and translated as part of the same polypeptide. Each putative protein was assigned a functional classification based on comparisons to predicted proteins from other *Maverick*s, to searches of the protein databases and based on structural homology of predicted protein folds (see Section 2). In total, we were able to assign 9 of the 11 putative *Trichomonas Maverick* proteins to two broad functional classes: DNA metabolism and structural proteins. The function of the remaining 2 putative proteins could not be predicted with confidence using these approaches.

**DNA metabolism.** Several putative proteins were identified that are predicted to be involved in DNA metabolism that share a common function with the CMGs from the other *Mavericks* (described above). Included in this group are the DNA polymerase B, the c-integrase and two ATPases (Fig. 5). In addition to these proteins, *T. vaginalis Mavericks* also encode two other putative proteins involved in DNA-binding and/or modification (see Fig. 5). A putative protein (280-aa) with homology to the DNA binding Kil-A domain (Iyer et al., 2002) is present in 5 of the 12 *T. vaginalis Mavericks*. The most closely related proteins containing the Kil-A domain annotated in the database are from the canary and fowl poxviruses (25% identical; 46% similar over 194-aa with canary poxvirus **AY318871**) and the *Acanthamoeba polyphaga* mimivirus. The function of these proteins is unknown. In addition, to the Kil-A domain, the corresponding protein from *Mav_Tv1.1* was predicted with an estimated precision of 55% to share a fold with the transcription repressor chain A of the regulatory protein sir4. Taken together, this data suggests a possible role of this putative protein in the transcriptional regulation of *Maverick* genes. An additional 285-aa ORF encoded by 6 of the 12 *Trichomonas Mavericks* was predicted to share a common fold with an UDP-Glycosyltransferase from T4 phage (with an estimated precision of 75%). UDP-Glycosyltransferases are DNA modifying enzymes that catalyzes the transfer of a glucose from uridine diphosphoglucose (UDP-Glc) to 5-hydroxymethylcytosine (5-HMC) in double-stranded DNA. They are thought to function in shielding the T4 phage DNA from attack by cellular nucleases.

**Structural proteins.** Three putative proteins (S1-S3) ranging in size from 285-aa to 1169-aa were each predicted to share a common fold with bacteriophage capsid or core proteins (Fig. 5). The largest of these predicted proteins, structural protein S1, ranges in size between 994-1169-aa and is encoded by 7 of the 12 complete elements (see Fig. 5). An alignment of these putative proteins reveals several well-conserved regions located near the N and C-termini flanking a central hypervariable region (Supplemental Fig. 2). Given the variability of the central region, structural fold prediction tools were more informative about the possible function of these proteins and relationships to other proteins than those based on primary sequence homology (e.g. PSI-BLAST). Nevertheless, both approaches converged in revealing an intriguing recurrent relationship to proteins involved in the assembly of the phage core and capsid as well as proteins involved in cell adhesion. For example, one region of the *Mav_Tv1.1* protein spanning the entire hypervariable region (from ~100 to 730-aa) is predicted with confidence levels between 75%-90% to share a common fold with the Reovirus core proteins p3 and p7 (Nakagawa et al., 2003; Fang et al., 2005) and the major capsid protein of bacteriophage PRD1 (Bamford et al., 2002). This fold extends completely through the hypervariable region of structural protein S1 (figure 5; Supplemental Fig. 2). The reovirus core p3, p7 proteins and the PRD1 major capsid proteins are the building blocks of the viral core and capsid structures of the respective viruses. In addition, PSI-BLAST searches revealed significant similarity in the C-terminal region with the gp10 baseplate wedge subunit and tail pin proteins of the bacteriophage RB49 (**AAQ15390**; *e*-value= 2e-04, 25% identity; 40% similar over 252-aa). The baseplate wedge and tail pin proteins are involved in host cell recognition and adsorption of the phage to host cells. Additionally, two other putative proteins were identified that are also predicted to share a common fold with the PRD1 major capsid protein, suggesting either that these proteins are encoded by members of a multi gene family or that they converged to the same structure (Fig. 5). It has been previously suggested that viral genes encoding proteins involved in capsid formation have arisen via gene duplication events (Boyko et al., 1992; Merckel et al., 2005), thus supporting the idea that the putative *Maverick* structural genes are related by descent.

## 4. Conclusions and relationship of *Mavericks* to other forms of invasive DNA

The results presented in this study confirm and expand our initial prediction that *Mavericks* represent an entirely new class of mobile elements widespread in eukaryotic genomes. Below we summarize several of the unique features of these elements in an attempt to shed some light on their evolutionary origin and their replication mechanism.

**Structural features.** One of the most striking characteristics of *Maverick*s in all species examined is their large size. In the absence of any genetic evidence of the sequence and biochemical activities required for *Maverick* mobility, we can only speculate based on the genetic organization and coding capacity of a canonical full-length, autonomous *Maverick* element. Nonetheless, most elements that are devoid of obvious truncation or internal deletion reach an average size of ~15-20 kb. Very few transposable elements have been previously shown to reach such a size. To our knowledge, elements larger than 15 kb have been seldom described and they are often limited to one subtype or family of elements within a superfamily of generally smaller TEs. For example, *Ogre* elements are a group of plant gypsy-like LTR retrotransposons that are ~25 kb

(Neumann et al., 2006). The *Candystripe1* element from sorghum is a 23-kb DNA transposon of the *CACTA* superfamily that can still be excised at high frequency (Chopra et al., 1999). In contrast, gigantism appears to be the norm for *Maverick* elements and it does not hinder their evolutionary success or their proliferation (e.g. ~3,000 copies in the relatively compact genome of *T. vaginalis*).

Another distinctive structural feature of *Maverick*s is the presence of large (typically 400-700 bp) and often perfect TIRs. DNA transposons are also characterized by TIRs, but these are usually much shorter (typically 10-200 bp). Given their large size and long TIRs, the genome of *Maverick*s is more reminiscent of the linear genome of other types of mobile and invasive DNA such as linear plasmids, bacteriophages and many eukaryotic DNA viruses. In these elements, the TIRs are directly involved in their replication mechanism (de Jong et al., 2003) and/or are thought to play a role in the protection of the ends from degradation by cellular enzymes (DeLange et al., 1986). In addition, all identified *Maverick* TIRs began with a short simple repeat motif (AGTAGT, TGTGTG, AGAGAGAG), which is seldom observed in DNA transposons, but resemble the short terminal repeat motifs at or near the end of the TIRs of adenoviruses (e.g. CATCAT in FADV-9 and CADV-1) and linear plasmids (AAAAATACTAC, **NC_004946**; CCCGTGTGTGT, **NC_004935**) (Supplemental Fig. 3). These short repeats are generated through a so-called sliding-back mechanism typical of the protein-primed replication of their genome (McNeel and Tamanoi, 1991; de Jong et al., 2003). We also observed that the *Maverick* TIRs are generally more similar to each other than to any other TIRs detectable within the same genome (data not shown). This suggests that either one TIR is replicated using the other TIR as a template (akin to the LTRs of retroelements) during *Maverick* replication or that some other mechanisms are responsible for homogenization of the TIRs after *Maverick* integration (e.g. gene conversion). The structural similarities between *Maverick*s and some DNA viruses suggest an important role of the TIRs in the replication of *Maverick*s.

**Genetic organization.** We have shown that *Maverick*s typically contain a set of 4-9 genes and occasionally many more (e.g. Section 3.6). We identified a set of 6 genes (c-integrase and CMG1-5) that are most common and relatively conserved among various *Maverick*s identified across the eukaryotic tree, but it is apparent that there is a wide variation in the coding capacity and genetic organization of the elements. Even the relative arrangement of the CMGs in relatively closely related species can greatly vary (see Fig. 1). This high gene content and plasticity of *Maverick*s is unusual among TEs and here again it is more reminiscent of viral genomes (Dolja and Koonin, 2006). Furthermore, the flexibility of gene arrangements in *Maverick*s also suggests that they contain multiple independent transcription units, and possibly that each gene possess its own core promoter.

At present, nothing is known about the transcription of *Maverick* elements and it would be an interesting area of future research. Blast searches of the EST database using *Maverick* CMGs as queries reveals that these putative proteins are frequently transcribed in many different taxa (data not shown). It is beyond the scope of this study to describe the origin and structure of these transcripts. It is worth noting, however, that several of the genes are predicted with a high level of confidence to be interrupted by spliceosomal introns and on some cases, the predicted gene structure is further supported by spliced EST/cDNA sequences (e.g. **K09F6.6** in *Mav_Ce2.3*). The presence of introns in CMGs suggests that transposition is unlikely to occur through reverse-transcription of an RNA intermediate, but rather via a DNA intermediate. Consistent with this view, no coding sequences with homology to reverse transcriptase or other retroelement-encoded proteins were detected in *Maverick* elements (besides the c-integrases and those occasionally deposited by secondary retrotransposon insertions, data not shown).

**Enzymatic capacity of *Maverick*-encoded proteins and evolutionary implications.**
Among the six CMGs, all but the c-integrases have a direct (sequence similarity) or indirect link (common domains and fold prediction) with proteins encoded by and involved in the replication cycle of DNA viruses. In addition numerous putative proteins encoded by a single *Maverick* are display homology to viral proteins (Fig.1). This raises several testable models considering the transposition mechanism of *Mavericks* some of which have been proposed elsewhere (Kapitonov and Jurka, 2006). Yet, at this point we can only be confident about the chromosomal integration step of the replication cycle of *Maverick*s; it most likely involves the encoded c-integrase and is accompanied by the duplication of 5-6 bp of host DNA. The detection of a homolog of DNA polymerase B and of a likely homolog of cysteine proteases found in adenoviruses point to a close

evolutionary link to genetic elements that use protein-primed DNA replication. Bamford and colleagues have proposed the existence of a common viral ancestor for a seemingly disparate group of double stranded DNA viruses that infect bacteria and animals including, the Tectiviridae bacteriophage PRD1 and the Adenoviridae (Benson et al., 1999; Benson et al., 2004). This hypothesis is based on a remarkably analogous architecture of the virion echoed by highly similar tri-dimensional folding of their major capsid proteins, despite the lack of detectable primary sequence homology. This relationship is further supported by the relationship of the viral encoded DNA polymerase B that is used to replicate both the PRD1 and adenovirus linear chromosome, and is also encoded by most *Mavericks* (Fig. 1 and Fig. 3). Given the structural and sequence similarities of *Mavericks* with some of these viruses and their ancient eukaryotic origin, it is possible that they all descend from the same common ancestral virus. *Mavericks* may have adapted to an intragenomic lifestyle by acquistion of an integrase-like protein in an early eukaryote ancestor. On the other hand, the prediction that a protein fold in one of the coil coiled domain protein of *Mav_Tv1* is shared with p3 and p7 core proteins of reovirus and the PRD1 P3 major capsid protein (see Section 3.6) and the presence of numerous other viral related putative proteins raises the intriguing and testable hypothesis that at least some *Maverick*s have retained an extracellular component in their replication cycle and an infectious capacity.

## Acknowledgements

We would like to thank John Pace (UTA) for technical assistance, Dr. David Edgell (University of Western Ontario) and Dr. Harri Savilahti (University of Helsinki) for stimulating discussions, during the preparation of this manuscript. This work was supported by start up funds from the Department of Biology at the University of Texas at Arlington.

## References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J.: Basic local alignment search tool. J Mol Biol 215 (1990) 403-10.
Baldauf, S.L., Roger, A.J., Wenk-Siefert, I. and Doolittle, W.F.: A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290 (2000) 972-7.
Bamford, J.K., Cockburn, J.J., Diprose, J., Grimes, J.M., Sutton, G., Stuart, D.I. and Bamford, D.H.: Diffraction quality crystals of PRD1, a 66-MDa dsDNA virus with an internal membrane. J Struct Biol 139 (2002) 103-12.
Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M. and Philippe, H.: The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci U S A 99 (2002) 1414-9.
Benson, S.D., Bamford, J.K., Bamford, D.H. and Burnett, R.M.: Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. Cell 98 (1999) 825-33.
Benson, S.D., Bamford, J.K., Bamford, D.H. and Burnett, R.M.: Does common architecture reveal a viral lineage spanning all three domains of life? Mol Cell 16 (2004) 673-85.
Boyko, V.P., Karasev, A.V., Agranovsky, A.A., Koonin, E.V. and Dolja, V.V.: Coat protein gene duplication in a filamentous RNA virus of plants. Proc Natl Acad Sci U S A 89 (1992) 9156-60.
Brookfield, J.F.: The ecology of the genome - mobile DNA elements and their hosts. Nat Rev Genet 6 (2005) 128-36.
Capy, P., Bazin, C., Higuet, D. and Langin, T.: Dynamics and evolution of transposable elements. Springer-Verlag, Austin, Texas, 1998.
Capy, P., Vitalis, R., Langin, T., Higuet, D. and Bazin, C.: Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? J Mol Evol 42 (1996) 359-68.
Chopra, S., Brendel, V., Zhang, J., Axtell, J.D. and Peterson, T.: Molecular characterization of a mutable pigmentation phenotype and isolation of the first active transposable element from Sorghum bicolor. Proc Natl Acad Sci U S A 96 (1999) 15330-5.
Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M.: Mobile DNA II. American Society for Microbiology Press, Washington, D.C., 2002.
Curcio, M.J. and Derbyshire, K.M.: The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol 4 (2003) 865-77.
de Jong, R.N., van der Vliet, P.C. and Brenkman, A.B.: Adenovirus DNA replication: protein priming, jumping back and the role of the DNA binding protein DBP. Curr Top Microbiol Immunol 272 (2003) 187-211.
DeLange, A.M., Reddy, M., Scraba, D., Upton, C. and McFadden, G.: Replication and resolution of cloned poxvirus telomeres in vivo generates linear minichromosomes with intact viral hairpin termini. J Virol 59 (1986) 249-59.
DiMaio, D. and Coen, D.M.: Replication Strategies of DNA Viruses. In: Knipe, D.M. and Howley, P.M. (Eds.), Fields Virology. Lippencott Williams & Wilkens, Philidelphia, 2001, pp. 119-132.
Dolja, V.D. and Koonin, E.V.: Virus Research. Elsevier B.V., 2006.
Eickbush, T.H. and Malik, H.S.: Origins and evolution of retrotransposons. In: Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (Eds.), Mobile DNA 2. ASM Press, Washington, DC, 2002, pp. 1111-1144.
Fang, Q., Shah, S., Liang, Y. and Zhou, Z.H.: 3D reconstruction and capsid protein characterization of grass carp reovirus. Sci China C Life Sci 48 (2005) 593-600.
Feschotte, C., Jiang, N. and Wessler, S.R.: Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3 (2002) 329-41.
Feschotte, C. and Pritham, E.J.: Non-mammalian c-integrases are encoded by giant transposable elements. Trends Genet 21 (2005) 551-2.
Feschotte, C. and Wessler, S.R.: Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. Proc Natl Acad Sci U S A 98 (2001) 8923-4.
Finnegan, D.J.: Eukaryotic transposable elements and genome evolution. Trends Genet. 5 (1989) 103-107.
Fitzgerald, P.M.D. and Springer, J.P.: Structure and function of retroviral proteases. Annual Review of Biophysics and Biophysical Chemistry 20 (1991) 299-320.
Gao, X. and Voytas, D.F.: A eukaryotic gene family related to retroelement integrases. Trends Genet 21 (2005) 133-7.
Gifford, R. and Tristem, M.: The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26 (2003) 291-315.
Haren, L., Ton-Hoang, B. and Chandler, M.: Integrating DNA: transposases and retroviral integrases. Annu Rev Microbiol 53 (1999) 245-81.

Herve, C.A., Forrest, G., Lower, R., Griffiths, D.J. and Venables, P.J.: Conservation and loss of the ERV3 open reading frame in primates. Genomics 83 (2004) 940-3.

Hickman, A.B., Perez, Z.N., Zhou, L., Musingarimi, P., Ghirlando, R., Hinshaw, J.E., Craig, N.L. and Dyda, F.: Molecular architecture of a eukaryotic DNA transposase. Nat Struct Mol Biol 12 (2005) 715-21.

Hizi, A. and Levin, H.L.: The integrase of the long terminal repeat-retrotransposon tf1 has a chromodomain that modulates integrase activities. J Biol Chem 280 (2005) 39086-94.

Hood, M.E.: Repeat-induced point mutation and the population structure of transposable elements in Microbotryum violaceum. Genetics (2005).

Horner, D.S. and Embley, T.M.: Chaperonin 60 phylogeny provides further evidence for secondary loss of mitochondria among putative early-branching eukaryotes. Mol Biol Evol 18 (2001) 1970-5.

Hua-Van, A., Le Rouzic, A., Maisonhaute, C. and Capy, P.: Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. Cytogenet Genome Res 110 (2005) 426-40.

Iyer, L.M., Koonin, E.V. and Aravind, L.: Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. Genome Biol 3 (2002) RESEARCH0012.

Kapitonov, V.V. and Jurka, J.: Rolling-circle transposons in eukaryotes. Proc. Natl. Acad. Sci. USA 98 (2001) 8714-8719.

Kapitonov, V.V. and Jurka, J.: Molecular paleontology of transposable elements in the Drosophila melanogaster genome. Proc Natl Acad Sci U S A 100 (2003) 6569-74.

Kapitonov, V.V. and Jurka, J.: Self-synthesizing DNA transposons in eukaryotes. Proc Natl Acad Sci U S A 103 (2006) 4540-5.

Kelley, L.A., MacCallum, R.M. and Sternberg, M.J.: Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299 (2000) 499-520.

Kent, W.J.: BLAT--the BLAST-like alignment tool. Genome Res 12 (2002) 656-64.

Knopf, C.W.: Evolution of viral DNA-dependent DNA polymerases. Virus Genes 16 (1998) 47-58.

Kostyuchenko, V.A., Navruzbekov, G.A., Kurochkina, L.P., Strelkov, S.V., Mesyanzhinov, V.V. and Rossmann, M.G.: The structure of bacteriophage T4 gene product 9: the trigger for tail contraction. Structure 7 (1999) 1213-22.

Kumar, S., Tamura, K. and Nei, M.: MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5 (2004) 150-63.

Kunze, R. and Weil, C.F.: The hAT and CACTA superfamilies of plant transposons. In: Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (Eds.), Mobile DNA II. American Society for Microbiology Press, Washington DC, 2002, pp. 565-610.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al.: Initial sequencing and analysis of the human genome. Nature 409 (2001) 860-921.

Lang, B.F., O'Kelly, C., Nerad, T., Gray, M.W. and Burger, G.: The closest unicellular relatives of animals. Curr Biol 12 (2002) 1773-8.

Lerat, E. and Capy, P.: Retrotransposons and retroviruses: analysis of the envelope gene. Mol Biol Evol 16 (1999) 1198-207.

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

Ma, L., Jones, C.T., Groesch, T.D., Kuhn, R.J. and Post, C.B.: Solution structure of dengue virus capsid protein reveals another fold. Proc Natl Acad Sci U S A 101 (2004) 3414-9.

Malik, H.S. and Eickbush, T.H.: Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. J Virol 73 (1999) 5186-90.

Malik, H.S. and Eickbush, T.H.: Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res 11 (2001) 1187-97.

Malik, H.S., Henikoff, S. and Eickbush, T.H.: Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10 (2000) 1307-18.

Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L.Y., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Marchler, G.H., Mullokandov, M., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Yamashita, R.A., Yin, J.J., Zhang, D. and Bryant, S.H.: CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res 33 Database Issue (2005) D192-6.

McGinnis, S. and Madden, T.L.: BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32 (2004) W20-5.

McNeel, D.G. and Tamanoi, F.: Terminal region recognition factor 1, a DNA-binding protein recognizing the inverted terminal repeats of the pGKl linear DNA plasmids. Proc Natl Acad Sci U S A 88 (1991) 11398-402.

Merckel, M.C., Huiskonen, J.T., Bamford, D.H., Goldman, A. and Tuma, R.: The structure of the bacteriophage PRD1 spike sheds light on the evolution of viral capsid architecture. Mol Cell 18 (2005) 161-70.

Murad, L., Bielawski, J.P., Matyasek, R., Kovarik, A., Nichols, R.A., Leitch, A.R. and Lichtenstein, C.P.: The origin and evolution of geminivirus-related DNA sequences in Nicotiana. Heredity 92 (2004) 352-8.

Nakagawa, A., Miyazaki, N., Taka, J., Naitow, H., Ogawa, A., Fujimoto, Z., Mizuno, H., Higashi, T., Watanabe, Y., Omura, T., Cheng, R.H. and Tsukihara, T.: The atomic structure of rice dwarf virus reveals the self-assembly mechanism of component proteins. Structure 11 (2003) 1227-38.

Nakayashiki, H., Awa, T., Tosa, Y. and Mayama, S.: The C-terminal chromodomain-like module in the integrase domain is crucial for high transposition efficiency of the retrotransposon MAGGY. FEBS Lett 579 (2005) 488-92.

Neumann, P., Koblizkova, A., Navratilova, A. and Macas, J.: Significant Expansion of Vicia pannonica Genome Size Mediated by Amplification of a Single Type of Giant Retroelement. Genetics 173 (2006) 1047-56.

Nicholas, K.B., Nicholas, H.B.J. and Deerfield, D.W.I.: GeneDoc:  Analysis and Visualization of Genetic Variation. EMBNEW.NEWs, 1997.

Notredame, C., Higgins, D.G. and Heringa, J.: T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302 (2000) 205-17.

Plisson, C., Uzest, M., Drucker, M., Froissart, R., Dumas, C., Conway, J., Thomas, D., Blanc, S. and Bron, P.: Structure of the mature P3-virus particle complex of cauliflower mosaic virus revealed by cryo-electron microscopy. J Mol Biol 346 (2005) 267-77.

Poulter, R.T., Goodwin, T.J. and Butler, M.I.: Vertebrate helentrons and other novel Helitrons. Gene 313 (2003) 201-12.

Rayner, J.C., Huber, C.S., Feldman, D., Ingravallo, P., Galinski, M.R. and Barnwell, J.W.: Plasmodium vivax merozoite surface protein PvMSP-3 beta is radically polymorphic through mutation and large insertions and deletions. Infect Genet Evol 4 (2004) 309-19.

Rice, P.A. and Baker, T.A.: Comparative architecture of transposase and integrase complexes. Nat Struct Biol 8 (2001) 302-7.

Silva, J.C., Bastida, F., Bidwell, S.L., Johnson, P.J. and Carlton, J.M.: A Potentially Functional Mariner Transposable Element in the Protist Trichomonas vaginalis. Mol Biol Evol (2004).

Steenkamp, E.T., Wright, J. and Baldauf, S.L.: The protistan origins of animals and fungi. Mol Biol Evol 23 (2006) 93-106.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A.: The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4 (2003) 41.

Temin, H.M.: Origin of retroviruses from cellular moveable genetic elements. Cell 21 (1980) 599-600.

Thompson, J.D., Desmond, D., Higgins, D.G. and Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (1994) 4673-4680.

Voytas, D.F. and Boeke, J.D.: Ty1 and Ty5 of Saccharomyces cerevisiae. In: Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (Eds.), Mobile DNA 2. American Society for Microbiology Press, Washington, DC, 2002, pp. 631-662.

Webster, A., Leith, I.R. and Hay, R.T.: Activation of adenovirus-coded protease and processing of preterminal protein. J Virol 68 (1994) 7292-300.

Wells, J.M., Ellingson, J.L., Catt, D.M., Berger, P.J. and Karrer, K.M.: A small family of elements with long inverted repeats is located near sites of developmentally regulated DNA rearrangement in Tetrahymena thermophila. Mol Cell Biol 14 (1994) 5939-49.

Werner, E.B., Taylor, W.R. and Holder, A.A.: A Plasmodium chabaudi protein contains a repetitive region with a predicted spectrin-like structure. Mol Biochem Parasitol 94 (1998) 185-96.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pontius, J.U., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L. and Yaschenko, E.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33 Database Issue (2005) D39-45.

Wuitschick, J.D., Gershan, J.A., Lochowicz, A.J., Li, S. and Karrer, K.M.: A novel family of mobile genetic elements is limited to the germline genome in Tetrahymena thermophila. Nucleic Acids Res 30 (2002) 2524-37.

Yano, S.T., Panbehi, B., Das, A. and Laten, H.M.: Diaspora, a large family of Ty3-gypsy retrotransposons in Glycine max, is an envelope-less member of an endogenous plant retrovirus lineage. BMC Evol Biol 5 (2005) 30.

Zhang, Q., Gutsch, D. and Kenney, S.: Functional and physical interaction between p53 and BZLF1: implications for Epstein-Barr virus latency. Mol Cell Biol 14 (1994) 1929-38.

## Figure legends

**Fig. 1. Coding capacity and genetic organization of *Maverick*s.**
The individual elements are labeled with the initials of the genus and species, except the previously described element *Tlr1* from *Tetrahymena thermophila*. Ce= *Caenorhabditis elegans*, Cb= *C. briggsae*, Cr= *C. remanei*, Pd= *Platynereis dumerilii*, Tc= *Tribolium castaneum*, Da= *Drosophila ananassae*, Dpe= *D. persimilis*, Stp= *Strongylocentrotus purpuratus*, Od= *Oikopleura dioica*, Dr= *Danio rerio*, Xt= *Xenopus tropicalis*, Sp= *Sphenodon punctatus,* Gi= *Glomus intraradices*, Pi= *Phytophthora infestans*, Tv=*Trichomonas vaginalis*. The black triangles represent the terminal-inverted repeats and the small open triangles the 5- or 6-bp TSD, when identified. Related ORFs are color-coded and a description of each ORF is given in the key. Solid colored arrows represent CMGs, which are defined as ORFs present in *Mavericks* elements from multiple species. Arrows not shaded represent ORFs present in *Mavericks* from a single species that are homologous to viral proteins. When secondary insertions of TEs were detected, their position is indicated and the type of nested TEs is given.

**Fig. 2. Comparative phylogenetic analysis of *Maverick* Integrase proteins.**
Bootstrapped neighbor-joining tree constructed with MEGA v. 3.1 from an alignment of a portion of the putative integrase protein corresponding to pfam00665 from species harboring *Maverick* elements. The phylogeny is rooted with the integrase of the *Ty1* polyprotein from yeast. Bootstrap values (1000 replicates) are given when > 50. If the sequence belongs to a defined TE it is labeled with the corresponding transposon name, or an abbreviated form, Mav=*Maverick* followed by the first initials of the genus and species names. If the sequence does not belong to a described element, than the first initials of the genus and species names are given, followed by the accession number. Ccv=*Cotesia congregata* Bracovirus, Nv= *Nasonia vitripennis*, Dp= *Drosophila persimilis*, Da= *D. ananassae*, Dv= *D. virilis*, Dg= *D. grimshawi*, Pp=*Pristionchus pacificus*, Cb= *Caenorhabditis briggsae*, Cr= *C. remanei*, Ce= *C. elegans*, Ap= *Acyrthosiphon pisum*, Tc= *Tribolium castaneum*, Od= *Oikopleura dioica*, Xm= *Xiphophorus maculates*, Tn= *Tetraodon nigroviridis*, Ga= *Gasterosteus aculeatus*, Dr= *Danio rerio*, Xt= *Xenopus tropicalis*, Aa= *Apteryx australis mantelli*, Stp= *Strongylocentrotus purpuratus*, Ci=*Ciona intestinalis*, Es= *Eptatretus stoutii*, Pd= *Platynereis dumerilii*, Gi= *Glomus intraradices*, Nev= *Nematostella vectensis*, He=*Hydractinia echinata*, Pi= *Phytophthora infestans*, Tt= *Tetrahymena thermophila*, Dd= *Dictyostelium discoideum*, Tv=*Trichomonas vaginalis*, Sc=*Saccharomyces cerevisiae*, Dm= *Drosophila melanogaster*. A chromodomain symbol is shown next to any putative protein where a chromodomain is detectable. The horizontal transmission event between the wasp and the bracovirus is indicated with a red arrow and the species are enclosed with a red box.

**Fig. 3. Domain structure and alignment of DNA polymerase B domain from linear plasmid, *Maverick*, bacteriophage and adenovirus.**
The first sequence in the alignment above is from a linear plasmid (mitochondria, *Zea mays*), the second sequence is from a *C. remanei Maverick,* the third sequence is from *Bacillus* phage PZA and the last is from human adenovirus 17. DNA polymerase B proteins are known to occur in some eukaryotic double-stranded DNA viruses, some bacteriophages and in cytoplasmic and mitochondrial linear plasmids from fungi and plants. DNA polymerase B proteins utilize a self-encoded protein to primer DNA replication. The residues known to be necessary for exonuclease activity in cysteine proteases are marked by black arrows. The residues involved in DNA binding and

stabilizing the unwound DNA are underlined and the polymerase catalytic site is boxed.

**Fig. 4. Alignment of CMG5 proteins with cysteine proteases from select adenoviruses, mimiviruses, poxviruses and from budding yeast.**
The black arrows indicate the residues required for catalytic activity of the protease. The first 12 sequences are from *Mavericks*, ASV=african swine virus, APV=acanthamoeba polyphaga virus, AdV=adenovirus, Sc=*Saccharomyces cerevisiae*

**Fig. 5. Coding capacity and organization of *Trichomonas vaginalis Maverick* elements.**
The approximate position and the strand orientation of each ORF encoding a putative protein of more than 200 residues (starting with a methionine codon) is illustrated by an arrow. The related ORFs are indicated with the same or similar colors and the predicted annotation is listed in the figure key.

Table 1: Taxonomic distribution of *Maverick* elements

| **Taxa** | Abbrev.[a] | Representative Accession [b] | Coordinates | DB [c] |
|---|---|---|---|---|
| **Nematodes** (round worms) | | | | |
| ***Caenorhabditis briggsae*** | Cb | CAAC01000115.1 | 402373 - 417795 | WGS |
| *C. elegans* | Ce | AL110478.1 | 26107 - 43380 | NR |
| *C. remanei* | Cr | AAGD01005187.1 | 27336 - 42574 | WGS |
| Annelids **(clam worm)** | | | | |
| ***Platynereis dumerilii*** | Pd | CT030680.1 | 112617 - 123393 | HTGS |
| **Insects** (fruit flies) | | | | |
| ***Drosophila ananassae*** | Da | AAPP01019595.1 | 518220 - 536360 | WGS |
| ***D. grimshawi*** | Dg | AAPT01000720.1 | Incomplete | WGS |
| ***D. melanogaster*** | Dm | AC016917 | Incomplete | HTGS |
| ***D. persimilis*** | Dpe | AAIZ01001788.1 | 2573 - 18340 | WGS |
| *D. pseudoobscura* | Dp | AAFS01001928.1 | Incomplete | WGS |
| *D. virilis* | Dv | AANI01015985.1 | Incomplete | WGS |
| *D. willistoni* | Dw | AAQB01006825.1 | Incomplete | WGS |
| *D. yakuba* | Dy | AAEU01001251.1 | Incomplete | WGS |
| **Insects** (flour beetle) | | | | |
| ***Tribolium castaneum*** | Tc | AC154128.3 | 19331 - 37011 | NR |
| **Echinoderms** (sea urchin) | | | | |
| ***Strongylocentrotus purpuratus*** | Stp | AAGJ01185250.1 | 1979 - 18488 | WGS |
| **Ascidians** (tunicates) | | | | |
| ***Ciona intestinalis*** | Ci | AABS01000479.1 | Incomplete | WGS |
| *Oikopleura dioica* | **Od** | AY449460.1 | 35615    58314 | NR |
| **Vertebrates** – **Teleost Fishes** | | | | |
| ***Danio rerio*** | Dr | BX294656.8 | 141050 - 156509 | NR |
| ***Eptatretus stoutii*** | Es | AY965680.1 | Incomplete | NR |
| ***Gasterosteus aculeatus*** | Ga | AANH01006348.1 | Incomplete | WGS |
| ***Takufugu rubripes*** | Tr | CAAB01011074.1 | Incomplete | WGS |
| ***Tetraodon nigroviridis*** | Tn | BX908814.1 | Incomplete | NR |
| **Vertebrates** – **Amphibians** (clawed frog) | | | | |
| *Xenopus tropicalis* | Xt | scaffold_959 | 38543 - 48326 | UCSC |
| **Vertebrates** – **Reptiles** (tuatara) | | | | |
| ***Sphenodon punctatus*** | Sp | AC153757 | 161293 - 173332 | NR |
| Fungi | | | | |
| ***Glomus intraradices*** | Gi | AC163889.2 | 5506 - 17459 | NR |
| ***Phakopsora pachyrhizi*** | Pp | AC149362.2 | Incomplete | NR |
| Stramenopiles **– oomycetes** | | | | |
| ***Phytophthora infestans*** | Pi | AC147544.2 | 81339 - 118719 | NR |
| **Alveolates** – ciliates | | | | |
| ***Tetrahymena thermophila*** | Tt | AAGF01001309.1 | Incomplete | WGS |
| **Parabasalids** - trichomonads | | | | |
| ***Trichomonas vaginalis*** | Tv | AAHC01000099.1 | 23568 - 46292 | WGS |

*Footnotes:*

a: species name abbreviation used throughout the figures.

b: accession number for sequence of a representative *Maverick* element.

c: database where the representative element was identified. WGS: GenBank whole genome shotgun sequencing project, NR: GenBank non redundant nucleotide database, HTGS, GenBank high-throughput genome sequencing, UCSC University of California Santa Cruz Genome Browser xenTro2 assembly.

Figure(s)

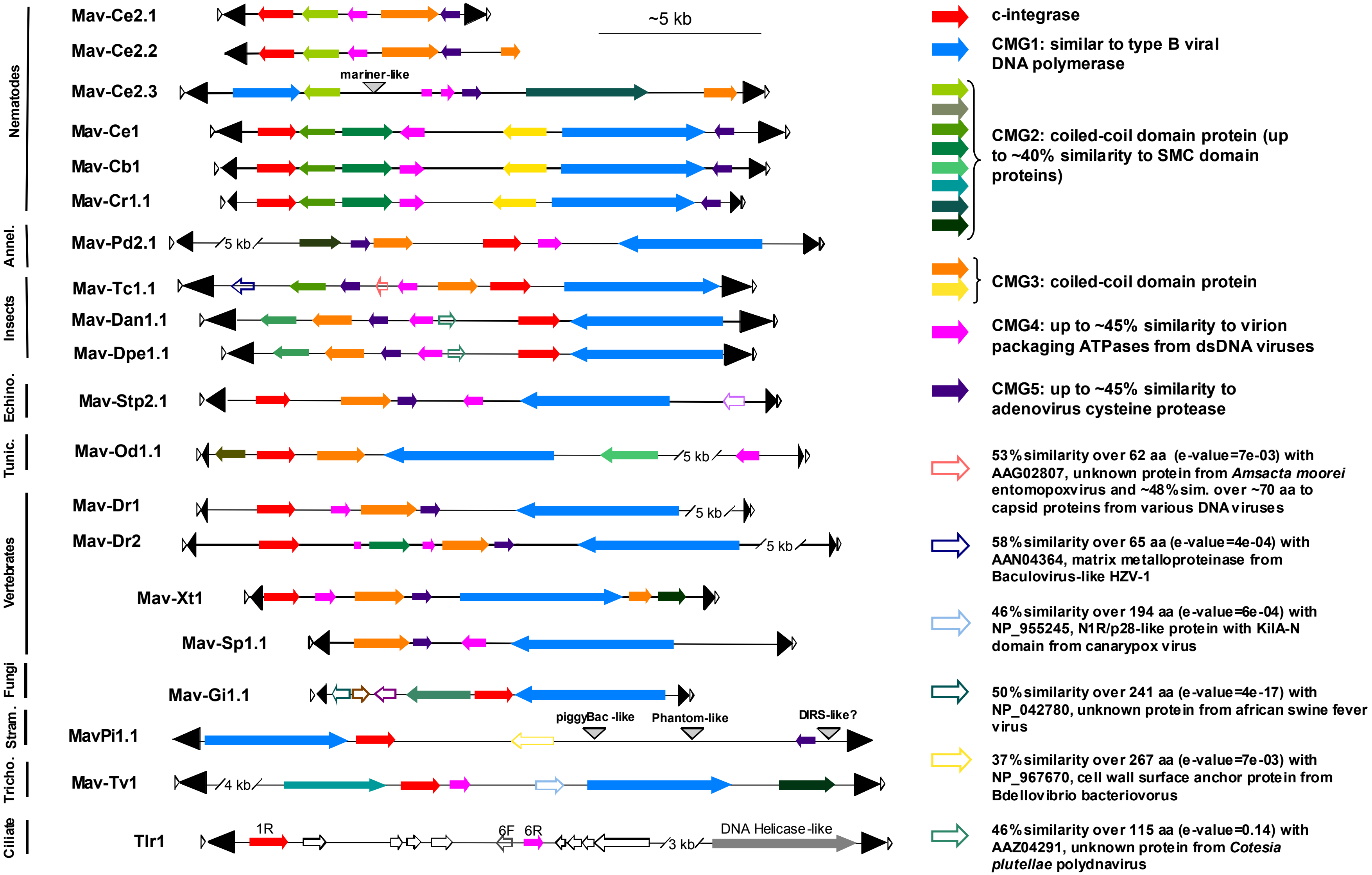

Nematodes
Annel.
Insects
Echino.
Tunic.
Vertebrates
Fungi
Stram.
Tricho.
Ciliate
Mav-Ce2.1
Mav-Ce2.2
Mav-Ce2.3
Mav-Ce1
Mav-Cb1
Mav-Cr1.1
Mav-Pd2.1
Mav-Tc1.1
Mav-Dan1.1
Mav-Dpe1.1
Mav-Stp2.1
Mav-Od1.1
Mav-Dr1
Mav-Dr2
Mav-Xt1
Mav-Sp1.1
Mav-Gi1.1
MavPi1.1
Mav-Tv1
Tlr1
~5 kb
mariner-like
5 kb
5 kb
5 kb
5 kb
piggyBac -like
Phantom-like
DIRS-like?
4 kb
1R
6F
6R
3 kb
DNA Helicase-like
c-integrase
CMG1: similar to type B viral DNA polymerase
CMG2: coiled-coil domain protein (up to ~40% similarity to SMC domain proteins)
CMG3: coiled-coil domain protein
CMG4: up to ~45% similarity to virion packaging ATPases from dsDNA viruses
CMG5: up to ~45% similarity to adenovirus cysteine protease
53% similarity over 62 aa (e-value=7e-03) with AAG02807, unknown protein from *Amsacta moorei* entomopoxvirus and ~48% sim. over ~70 aa to capsid proteins from various DNA viruses
58% similarity over 65 aa (e-value=4e-04) with AAN04364, matrix metalloproteinase from Baculovirus-like HZV-1
46% similarity over 194 aa (e-value=6e-04) with NP_955245, N1R/p28-like protein with KilA-N domain from canarypox virus
50% similarity over 241 aa (e-value=4e-17) with NP_042780, unknown protein from african swine fever virus
37% similarity over 267 aa (e-value=7e-03) with NP_967670, cell wall surface anchor protein from Bdellovibrio bacteriovorus
46% similarity over 115 aa (e-value=0.14) with AAZ04291, unknown protein from *Cotesia plutellae* polydnavirus

Figure 2: Phylogeny of *Maverick* Integrase proteins

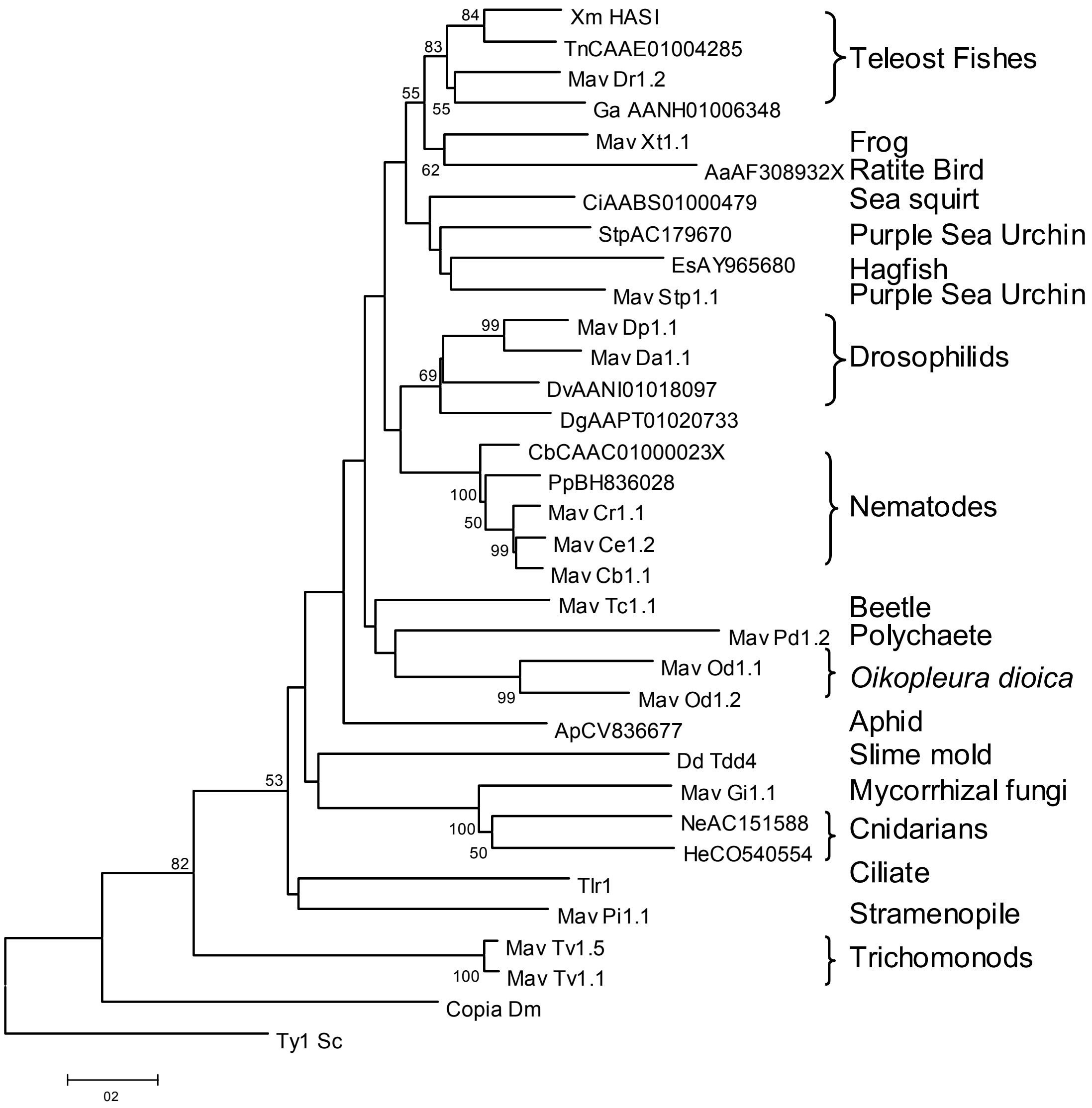


Figure 3

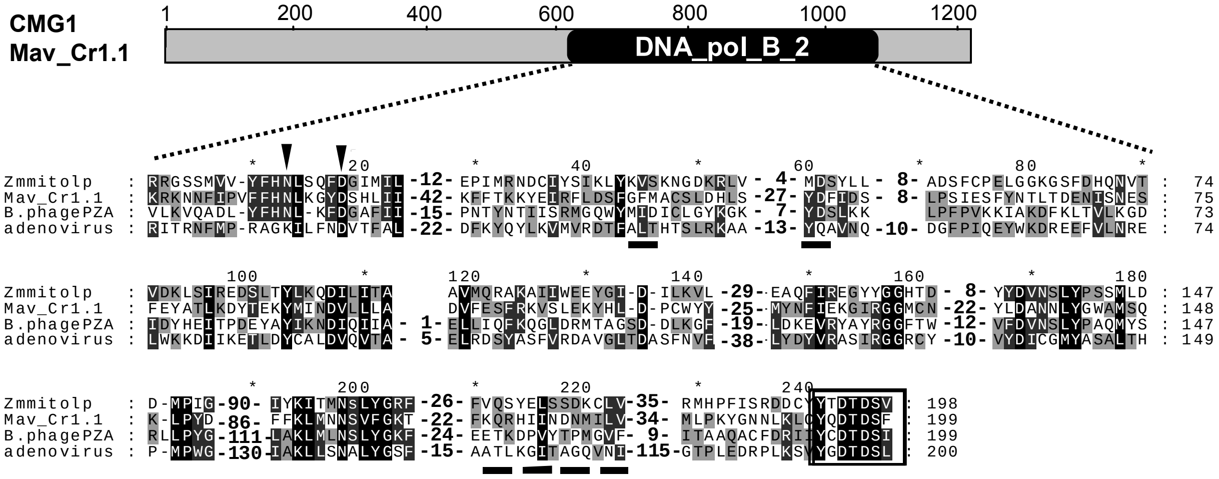
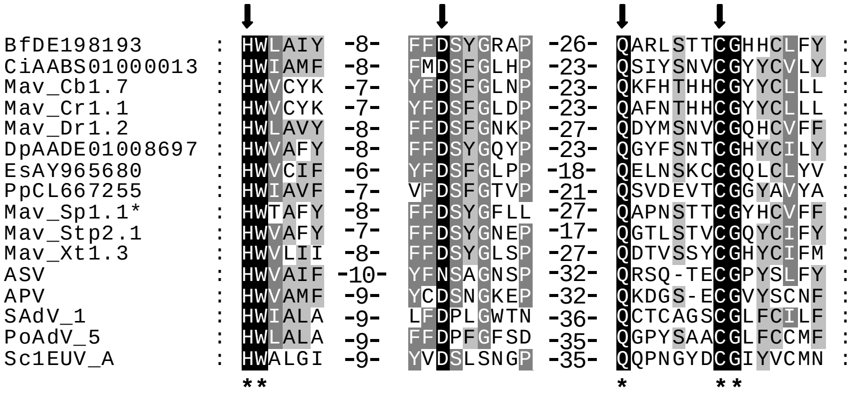

CMG1
Mav_Cr1.1
1
200
400
600
800
1000
1200
DNA_pol_B_2
* 20 * 40 * 60 * 80 *
Zmmitolp : RRGSSMVV-YFHNLSQFDGIMIL -12- EPIMRNDCIYSIKLYKVSKNGDKRLV - 4- MDSYLL - 8- ADSFCPELGGKGSFDHQNVT : 74
Mav_Cr1.1 : KRKNNFIPVFFHNLKGYDSHLII -42- KFFTKKYEIRFLDSFGFMACSLDHLS -27- YDFIDS - 8- LPSIESFYNTLTDENISNES : 75
B.phagePZA : VLKVQADL-YFHNL-KFDGAFII -15- PNTYNTIISRMGQWYMIDICLGYKGK - 7- YDSLKK LPFPVKKIAKDFKLTVLKGD : 73
adenovirus : RITRNFMP-RAGKILFNDVTFAL -22- DFKYQYLKVMVRDTFALTHTSLRKAA -13- YQAVNQ -10- DGFPIQEYWKDREEFVLNRE : 74
100 * 120 * 140 * 160 * 180
Zmmitolp : VDKLSIREDSLTYLKQDILITA AVMQRAKAIIWEEYGI-D-ILKVL -29- EAQFIREGYYGGHTD - 8- YYDVNSLYPSSMLD : 147
Mav_Cr1.1 : FEYATLKDYTEKYMINDVLLLA DVFESFRKVSLEKYHL-D-PCWYY -25- MYNFIEKGIRGGMCN -22- YLDANNLYGWAMSQ : 148
B.phagePZA : IDYHEITPDEYAYIKNDIQIIA - 1- ELLIQFKQGLDRMTAGSD-DLKGF -19- LDKEVRYAYRGGFTW -12- VFDVNSLYPAQMYS : 147
adenovirus : LWKKDIIKETLDYCALDVQVTA - 5- ELRDSYASFVRDAVGLTDASFNVF -38- LYDYVRASIRGGRCY -10- VYDICGMYASALTH : 149
* 200 * 220 * 240
Zmmitolp : D-MPIG -90- IYKITMNSLYGRF -26- FVQSYELSSDKCLV -35- RMHPFISRDDCYYTDTDSV : 198
Mav_Cr1.1 : K-LPYD -86- FFKLMNNSVFGKT -22- FKQRHIINDNMILV -34- MLPKYGNNLKLCYQDTDSF : 199
B.phagePZA : RLLPYG -111- LAKLMLNSLYGKF -24- EETKDPVYTPMGVF- 9- ITAAQACFDRIIYCDTDSI : 199
adenovirus : P-MPWG -130- IAKLLSNALYGSF -15- AATLKGITAGQVNI -115- GTPLEDRPLKSVYGDTDSL : 200

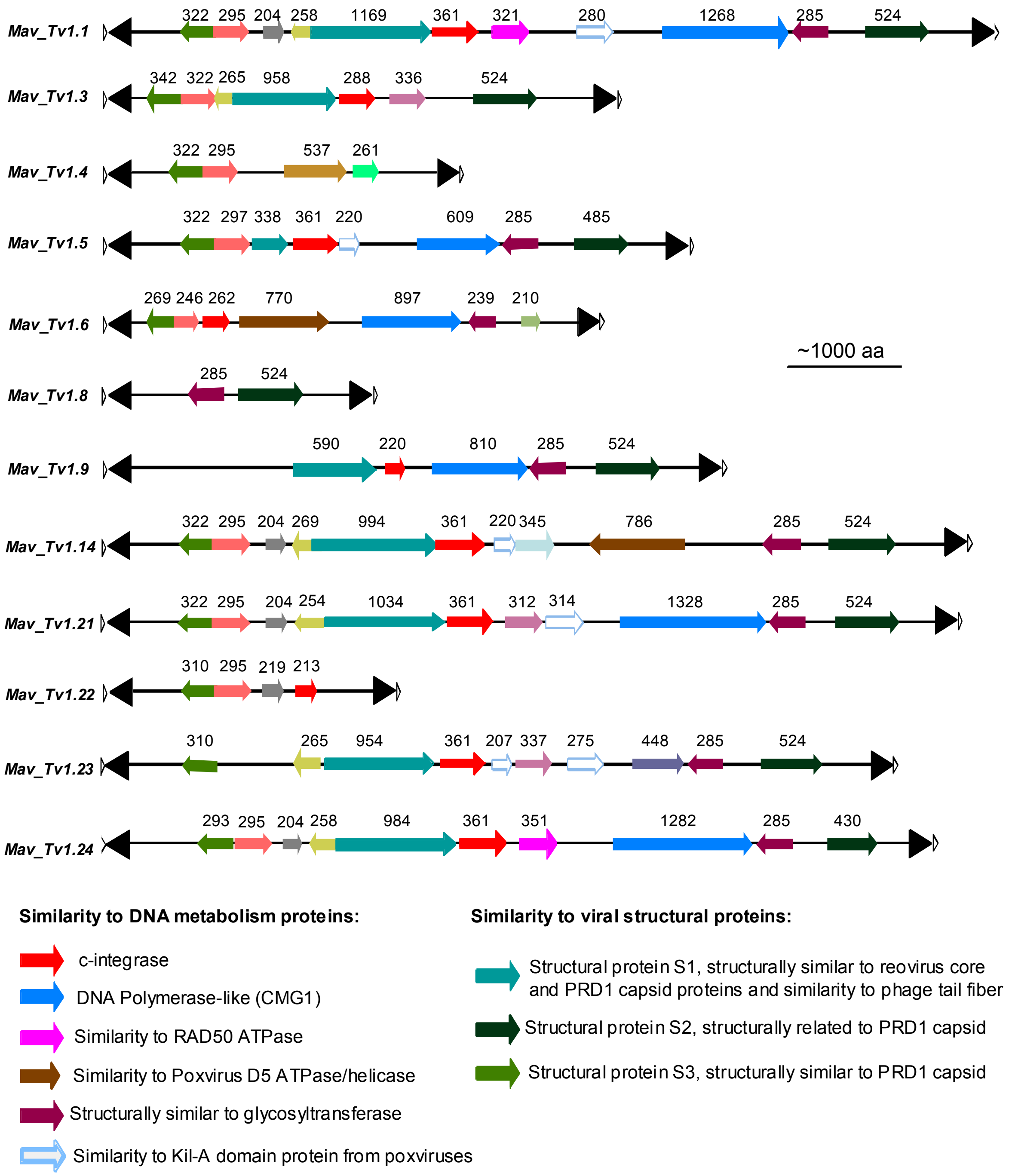
Figure 4



```
BfDE198193     : HWLAIY  -8-  FFDSYGRAP -26- QARLSTTCGHHCLFY :
CiAABS01000013 : HWIAMF  -8-  FMDSFGLHP -23- QSIYSNVCGYYCVLY :
Mav_Cb1.7      : HWVCYK  -7-  YFDSFGLNP -23- QKFHTHHCGYYCLLL :
Mav_Cr1.1      : HWVCYK  -7-  YFDSFGLDP -23- QAFNTHHCGYYCLLL :
Mav_Dr1.2      : HWLAVY  -8-  FFDSFGNKP -27- QDYMSNVCGQHCVFF :
DpAADE01008697 : HWVAFY  -8-  FFDSYGQYP -23- QGYFSNTCGHYCILY :
EsAY965680     : HWVCIF  -6-  YFDSFGLPP -18- QELNSKCCGQLCLYV :
PpCL667255     : HWIAVF  -7-  VFDSFGTVP -21- QSVDEVTCGGYAVYA :
Mav_Sp1.1*     : HWTAFY  -8-  FFDSYGFLL -27- QAPNSTTCGYHCVFF :
Mav_Stp2.1     : HWVAFY  -7-  FFDSYGNEP -17- QGTLSTVCGQYCIFY :
Mav_Xt1.3      : HWVLII  -8-  FFDSYGLSP -27- QDTVSSYCGHYCIFM :
ASV            : HWVAIF -10-  YFNSAGNSP -32- QRSQ-TECGPYSLFY :
APV            : HWVAMF  -9-  YCDSNGKEP -32- QKDGS-ECGVYSCNF :
SAdV_1         : HWIALA  -9-  LFDPLGWTN -36- QCTCAGSCGLFCILF :
PoAdV_5        : HWLALA  -9-  FFDPFGFSD -35- QGPYSAACGLFCCMF :
Sc1EUV_A       : HWALGI  -9-  YVDSLSNGP -35- QQPNGYDCGIYVCMN :
                 **                          *      **
```

Figure(s)

Figure 5


Mav_Tv1.1
322 295 204 258 1169 361 321 280 1268 285 524
Mav_Tv1.3
342 322 265 958 288 336 524
Mav_Tv1.4
322 295 537 261
Mav_Tv1.5
322 297 338 361 220 609 285 485
Mav_Tv1.6
269 246 262 770 897 239 210
~1000 aa
Mav_Tv1.8
285 524
Mav_Tv1.9
590 220 810 285 524
Mav_Tv1.14
322 295 204 269 994 361 220 345 786 285 524
Mav_Tv1.21
322 295 204 254 1034 361 312 314 1328 285 524
Mav_Tv1.22
310 295 219 213
Mav_Tv1.23
310 265 954 361 207 337 275 448 285 524
Mav_Tv1.24
293 295 204 258 984 361 351 1282 285 430
Similarity to DNA metabolism proteins:
c-integrase
DNA Polymerase-like (CMG1)
Similarity to RAD50 ATPase
Similarity to Poxvirus D5 ATPase/helicase
Structurally similar to glycosyltransferase
Similarity to Kil-A domain protein from poxviruses
Similarity to viral structural proteins:
Structural protein S1, structurally similar to reovirus core and PRD1 capsid proteins and similarity to phage tail fiber
Structural protein S2, structurally related to PRD1 capsid
Structural protein S3, structurally similar to PRD1 capsid