School of Doctoral Studies in Biological Sciences

University of South Bohemia in České Budějovice

Faculty of Science

Structural and functional characterization of giant plant Ogre-like retrotransposons

Ph.D. Thesis

RNDr. Veronika Steinbauerová

Supervisor: RNDr. Jiří Macas, Ph.D.
Institute of Plant Molecular Biology, BC AS CR

České Budějovice 2012

This thesis should be cited as:

Steinbauerová V, 2012: Structural and functional characterization of giant plant Ogre-like retrotransposons. Ph.D. Thesis, University of South Bohemia, Faculty of Science, School of Doctoral Studies in Biological Sciences, České Budějovice, Czech Republic.

Annotation

Ogre elements represent a distinct group of Ty3/gypsy LTR retrotransposons occurring in a range of dicot plants. They are characterized by two specific features – presence of long extra open reading frame in 5´ untranslated region with unknown function and a non-coding sequence containing several stop codons separating protease and reverse transcriptase domains which was proposed to be removed by splicing. This thesis describes the functional analysis of intron splicing in Ogre retrotransposons. Further, it investigates additional coding information not only in Ogre retrotransposons but in the whole group of Ty3/gypsy retroelements.

Declaration [in Czech]

Prohlašuji, že svoji disertační práci jsem vypracovala samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své disertační práce, a to v úpravě vzniklé vypuštěním vyznačených částí archivovaných Přírodovědeckou fakultou elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

České Budějovice, 4. 4. 2012

....

Veronika Steinbauerová

This thesis originated from a partnership of Faculty of Science, University of South Bohemia, and Institute of Plant Molecular Biology, Biology Centre of the ASCR, supporting doctoral studies in the Molecular and Cell Biology study programme.

Financial support

Work was supported by grants IAA500960702 and AVOZ50510513 from the Academy of Sciences of the Czech Republic, LC06004 from the Ministry of Education, Youth and Sport of the Czech Republic and P521/12/G090 from the Czech Science Foundation.

Acknowledgements

I would like to thank my supervisor Dr. Jiří Macas and other members of our laboratory for their support during my PhD study. Special thanks go to Jana Látalová and Helena Štěpančíková for excellent technical assistance.

List of papers and author's contribution

The thesis is based on the following papers (listed chronologically):

Veronika Steinbauerová, Pavel Neumann, Jiří Macas (2008) Experimental evidence for splicing of intron-containing transcripts of plant LTR retrotransposon Ogre. Molecular Genetics and Genomics 280: 427-436

Veronika Steinbauerová performed all experiments. All authors contributed to the manuscript preparation and approved its final version.

Veronika Steinbauerová, Pavel Neumann, Petr Novák, Jiří Macas (2012) A widespread occurrence of extra open reading frames in plant Ty3/gypsy retrotransposons. Genetica: accepted

Veronika Steinbauerová performed most of the bioinformatic analyses using scripts and programs developed by co-authors. All authors contributed to the manuscript preparation and approved its final version.

Contents

INTRODUCTION 1

1. Eukaryotic transposable elements 1

1.1 Classification of TEs 1

1.2 Genomic impact of TEs 3

2. LTR retrotransposons 6

2.1 Structure of elements 6

2.2 Replication cycle 7

3. Ogre retrotransposons 9

3.1 Splicing of Ogre elements 11

3.2 Extra open reading frames in Ogre-like retrotransposons 13

AIMS OF THE WORK 16

RESULTS 17

Part I - Functional analysis of splicing of Ogre retrotransposons 17

Part II - Analysis of additional open reading frames in Ogre-like and other Ty3/gypsy elements 19

CONCLUSIONS 21

REFERENCES 22

INTRODUCTION

1. Eukaryotic transposable elements

A significant portion of eukaryotic genomes is composed of transposable elements (TEs) that in some cases account for 50 – 70% of the genome (Meyers et al. 2001; Biemont and Vieira 2006; Gentles et al. 2007). TEs were first discovered in maize by Barbara McClintock and her breeding experiments provided the first descriptions of these genetic agents (McClintock 1951). Because TEs exploit host resources for selfish purposes they have been considered as genomic parasites. However, recent observations showed that mobile DNA plays a major role in the structural and functional evolution of genes and genomes in various organisms (Kazazian 2004; Feschotte 2008).

1.1 Classification of TEs

Eukaryotic TEs are divided into two classes according to whether their replication is mediated by RNA (class I or retrotransposons) or DNA (class II or DNA transposons). According to the classification system for eukaryotic TEs by Wicker et al. (2007) these two classes are further subdivided into subclasses, orders, superfamilies, etc. based on structural relationships, details of the transposition mechanism, and sequence similarities.

Class I of TEs contains five orders that have been distinguished on the basis of their transposition mechanism, organization and phylogeny of reverse transcriptase (RT, the enzyme encoded by retroelements, see below). This group of mobile genetic elements is characterized by a replicative (copy-and-paste) mode of transposition that involves transcription of genomic copy of element, reverse transcription of the resulting RNA into DNA, and subsequent integration of a new element into another locus. This replicative process can rapidly increase the copy number of retrotransposons and thus these elements may be important players in the evolution of genome size (Kumar and Bennetzen 1999; Vitte and Panaud 2005; Sabot and Schulman 2006).

First order of class I, LTR retrotransposons, possess long terminal repeats (LTR) that flank the internal region. In all autonomous elements the inner part encodes both structural

proteins (Gag) and enzymes (protease, reverse transcriptase, RNaseH and integrase) that are essential for TEs life cycle. Second and third order, LINEs (long interspersed elements) and SINEs (short interspersed elements), were previously defined as non-LTR retrotransposons because of their lack of long terminal repeats. Autonomous LINEs encode at least reverse transcriptase and a nuclease. The best known family of this order, L1 retroelements, represents approximately 17% of the human genome (Cordaux and Batzer 2009). The SINE elements originate from random retrotransposition events of various polymerase III transcripts (Kramerov and Vassetzky 2005). The internal polymerase III promoter allows them to be expressed. Nevertheless, they have no protein coding capacity and their replication and integration is dependent on proteins encoded by LINE elements (Ohshima et al. 1996; Kajikawa et al. 2002; Dewannieux et al. 2003). Members of the fourth order of retrotransposons, DIRS-like elements encode tyrosine recombinase instead of integrase and probably use different mechanism of integration than other retrotransposons (Capello et al. 1985; Goodwin and Poulter 2004). The last order consists of Penelope-like elements that were first identified in *Drosophila virilis*. The transpositionally active *Penelope* retrotransposon is responsible for the hybrid dysgenesis syndrome in this species (Evgenev et al. 1997). These retrotransposons encode reverse transcriptase domain with similarity to telomerase and an endonuclease (Evgenev and Arkhipova 2005).

Class II of TEs consists of two subclasses that are distinguished by the number of DNA strands that are cut during transposition event (Wicker et al. 2007). Nevertheless, these elements never move via RNA and the genomic DNA itself is the mobile intermediate.

First subclass contains orders TIR and Crypton that both use the classical cut and paste mechanism of transposition. Presence of terminal inverted repeats (TIRs) is the characteristic feature of nine superfamilies (with well-described members such as *P* element from *Drosophila* or *Ac-Ds* from maize; Rio 1990; Courage et al. 1984) belonging to the TIR order. Transposition process is mediated by a transposase enzyme that recognizes the TIRs. The transposase excises the transposon and integrates it into the target site. The gaps at the integration site and double-stranded DNA breaks in the donor DNA are repaired by the host DNA repair machinery (Miskey et al. 2005). The poorly known Crypton elements comprise the second order of this subclass. They contain a long gene interrupted by several introns that encodes putative tyrosine recombinase (that was found also in DIRS-like retrotransposons) but lack reverse transcriptase domain which suggests that they transpose via DNA intermediate (Goodwin et al. 2003).

Subclass 2 of DNA transposons contains two orders, Helitron and Maverick. Elements of both orders most likely rely on distinct transposition mechanisms involving replication without double-stranded cleavage. These elements probably transpose through a replicative, copy-and-paste process (Feschotte and Pritham 2007). Helitrons probably replicate via a rolling-circle (RC) mechanism that is similar to the RC replication strategy of some plasmids, single-stranded bacteriophages, plant geminiviruses and some prokaryotic transposable elements (Kapitonov and Jurka 2001; Feschotte and Wessler 2001). Helitrons have no TIRs and autonomous copies encode helicase and nuclease/ligase. Remarkable feature of Helitrons is their ability to carry gene fragments that have been captured from the host genome (Morgante et al. 2005; Lai et al. 2005). Transposable elements of the order Maverick (also known as Polintons) are very large transposons with long TIRs and coding capacity for multiple proteins. Their mechanism of transposition is not yet well understood, but they probably replicate using a self-encoded DNA polymerase (Feschotte and Pritham 2005; Kapitonov and Jurka 2006; Pritham et al. 2007).

Both retrotransposons and DNA transposons can be either autonomous or nonautonomous. The latter group lacks some (or all) genes encoding the essential proteins for their transposition, so they must "borrow" these proteins from autonomous TEs. To the nonautonomous groups of TEs belong large retrotransposon derivatives (LARDs), terminal repeat retrotransposons in miniature (TRIMs) or miniature inverted-repeat transposable elements (MITEs). LARDs and TRIMs are LTR retrotransposon derivatives (Kalendar et al. 2004; Witte et al. 2001) while MITEs are deletion derivatives of DNA transposons with copy number reaching over thousands in genomes (Feschotte et al. 2002).

1.2 Genomic impact of TEs

Mobile elements affect the genomes in many different ways and they play a key role in the evolution of eukaryotic genomes and gene regulation.

TEs make up large portions of most eukaryotic genomes with 45 % of human genomic content, 37 % of mouse genome (Deininger et al. 2003) and in plants, where TEs can be extremely successful in their amplification, they constitute up to 70 % of the genome (Meyers et al. 2001). To explain such phenomenon, the skeletal role of this "junk DNA" has been

hypothesized (Cavalier-Smith and Beaton 1999). According to this theory, optimal cellular function requires a relatively constant cytonuclear ratio (ration between cell volume and nuclear volume, i.e. genome size). Therefore, amplification of TEs that increases the genome size may play a role in maintaining critical nuclear volume and the amount of non-coding DNA derived from TEs is maintained by positive selection.

Mobile elements can affect the regulation of nearby gene expression at several levels (for review see Feschotte et al. 2008). At the transcriptional level a transposable element can influence the expression of close gene by inserting the regulatory sequences such as promoters or transcription factor binding sites. For example, it was shown that almost 25% of the analyzed promoter regions of human genome contain TE-derived sequences (Jordan et al. 2003). At the post-transcriptional level, a TE can introduce an alternative polyadenylation site and it can cause various forms of alternative splicing by inserting within the intron sequence. Alternatively, transposon can be exonized which leads to either the formation of new protein isoform or the degradation of mRNA by nonsense-mediated decay pathway if the premature stop codon is introduced. Analysis of TE exonization events in human and mouse genome revealed that this process occurs preferentially in the beginning of protein coding sequences and can be population-specific. Therefore, the exonizations may enhance divergence and lead to speciation (Sela et al. 2010).

Proteins or protein domains encoded by transposable elements have been recruited by their hosts during evolution of diverse eukaryotic lineages. Host genomes have domesticated a various proteins encoded by TEs, such as gag proteins, integrases, transposases or envelope proteins of retroviruses (for review see Miller et al. 1999; Volff 2006). The well-studied domesticated transposase is Rag1 protein that plays an important role in formation of various antibodies produced by immune system of jawed vertebrates. Rag1 together with Rag2 catalyze the recombination of the V(D)J locus which leads to the high diversity of immunoglobulins and T-cell receptors. It was shown that Rag1 core region required for its catalytic activity as well as the recombination signal sequences are derived from an ancient DNA transposons *Transib* (Agraval et al. 1998; Hiom et al. 1998; Roth and Craig 1998; Kapitonov and Jurka 2005).

The key enzyme of retroelements, the reverse trancriptase (RT) is thought to be relative to telomerase that replicates telomeres of chromosomes. Telomerase is an RNA-dependent DNA polymerase that shares sequence similarities with reverse transcriptase of

mobile elements and thus seems to be created by domestication of RT of a retroelement (Nakamura et al. 1997). Transposable elements play a key role in maintaining telomeres of *Drosophila melanogaster* using a different strategy. *Drosophila* cannot regenerate the ends of chromosome using telomerase since this species does not encode this enzyme. Instead, two transposable elements, *Het-A* and *TART*, have been domesticated by the genome to perform this function. These elements repeatedly insert at the telomeres and thus they are able to extend the chromosome ends and slow terminal loss (Levis et al. 1993; Sheen and Levis 1994).

Besides their positive contribution to their host genome, transposable elements can promote mutations and chromosome rearrangements leading to various disorders. In humans, *de novo* insertions of transposable elements *L1*, *Alu* and *SVA* were shown to cause various diseases such as haemophilia, cystic fibrosis, muscular dystrophy, β-thalassemia or breast and colon cancers (Callinan and Batzer 2006; Chen et al. 2005; Cordaux and Batzer 2009).

Due to this harmful effect of TEs, organisms employ various strategies to protect their genome including DNA methylation, modification of histones (especially methylation of lysine 9 of histone H3, H3mK9), chromatin remodelling and RNA interference (Lippman et al. 2003; Zilberman and Henikoff 2004). In RNA interference (RNAi) pathway, dsRNA is cleaved into short RNAs which are subsequently associated with RNA-induced silencing complex (RISC) and they guide degradation or translation silencing of complementary transcripts (Aravin and Tuschl 2005). dsRNA from transposable elements can be generated from two overlapping antiparallel transcripts that base pair together as it was shown for human *L1* retroelement (Yang and Kazazian 2006). Short RNAs can function also in alternate pathway (RNAi-mediated chromatin modifications) where a different complex cleaves nascent transcripts that are still attached to RNA polymerase II and the DNA strand. This region of chromatin is subsequently modified by methylation of DNA and/or histone H3 (H3mK9) which leads to the formation of heterochromatin that is condensed and inaccessible to transcription (Slotkin and Martienssen 2007).

The precise control of retroelement transposition was shown to play an important role in generating neuronal sequence diversity (for review see Singer et al. 2010). Non-LTR retrotransposons *L1* are mobilized during the formation of central nervous system and later during adult neurogenesis. Transposition occurs independently in individual cells which leads to the neuron-to-neuron variation in genomic DNA. *L1* transposition is repressed in neural

stem cells by a complex comprised of the transcription factor SOX2 and histone deacetylase 1 (HDAC1). During the transition from neural stem cells to neural progenitor cells, derepression of the SOX2/HDAC1 complex occurs and Wnt signaling pathway induces expression of *L1* retrotransposons (Kuwabara et al. 2009). Ectopic activation of previously silenced retrotransposons leads to the somatic mosaicism and thus to the phenotypic diversity of the neurons since new *L1* insertions may influence the expression of nearby genes (Muotri et al. 2005). This mechanism may affect neural plasticity, cognition and ultimately behaviour in individuals (Singer et al. 2010).

2. LTR retrotransposons

Long terminal repeat (LTR) retrotransposons are less abundant in animals, but represent a major fraction of repetitive sequences in plant genomes. They are closely related to the retroviruses since they share the mechanisms of intracellular element transcription, replication and integration. One distinguishing feature between these two types of retroelements is the presence of *envelope* (*env*) gene in the retroviruses that enables them to be infectious. Envelope proteins associate with cell membrane and allow the budding of viral particles from infected cells (for review see Coffin et al. 1997; Kuman and Bennetzen 1999).

2.1 Structure of elements

LTR retrotransposons have direct LTRs that can range from a few hundreds nucleotides to several kb in size (Ogre retrotransposons possess LTRs over 6 kb long; Macas and Neumann 2007). The LTRs contain the sequences associated with the transcription such as promoter and polyadenylation signal. LTRs terminate in short inverted repeats, usually 5´–TG...CA–3´ and flank the internal *gag-pol* region that encodes both structural and enzymatic proteins. The *gag* gene encodes structural proteins that form virus-like particles where the reverse transcription occurs. The *pol* gene encodes the enzymatic functions required for replication, including protease (PRO), reverse transcriptase/RNaseH (RT/RH) and integrase (INT). The proteins encoded by *gag* and *pol* are synthesized as a polyprotein and individual

functional peptides are released by the action of PRO (Kumar and Bennetzen 1999; Havecker et al. 2004).

Gag and Pol proteins are encoded either within a single open reading frame (ORF) or the overlapping or adjacent ORFs are present. In latter cases the translation of proteins downstream of the stop codon or change in frame is facilitated by ribosomal frameshifting that was well studied in the retroviruses but occurs also in retrotransposons (Coffin et al. 1997; Gao et al. 2003). Ribosome occasionally slips at specific sites one nucleotide backward (-1 frameshifting) of forward (+1 frameshifting) and translation can continue in alternate downstream frame. Another translation-recoding mechanism is stop codon readthrough, where a stop codon is occasionally misread as a sense codon allowing translation to continue into a downstream ORF (Coffin et al. 1997). Finally, internal ribosome entry site (IRES) was proposed to function as a translation mechanism used by some retrotransposons (Meignin et al. 2003; Li et al. 2006). Since virus-like particle assembly requires many more copies of Gag than Pol the translation recoding mechanisms ensure the proper ratio of these proteins (Wilson et al. 1986; Kumar and Bennetzen 1999; Gao et al. 2003). For the elements with single ORF the required Gag/Pol ratio may be achieved by posttranslational Pol degradation as it was shown for yeast retrotransposons *Tf1* and *Ty5* (Atwood et al. 1996; Irwin and Voytas 2001).

Two major groups of LTR retrotransposons are Ty1/copia and Ty3/gypsy. The classification is based on phylogenetic relationship of the catalytic proteins and on the organization of the *pol* region (Hull 2001; Wicker et al. 2007). In the Ty1/copia group, INT precedes RT/RH, whereas in the Ty3/gypsy group INT resides at the 3' end of the *pol* domain. Ty3/gypsy group possess the same *pol* configuration as retroviruses and RT of both groups of retroelements show higher similarity while Ty1/copia group is more distantly related (Xiong and Eickbush 1990).

2.2 Replication cycle

The replication cycle of LTR retrotransposons comprises transcription, translation, reverse transcription and integration of element cDNA into genome (for review see Coffin et al. 1997; Kuman and Bennetzen 1999; Sabot and Schulman 2006).

First, synthesis of retrotransposon mRNA molecule is mediated by the cellularly encoded RNA polymerase II from a promoter located within the 5' LTR. This mRNA encodes the proteins needed for replication, and also serves as the template for reverse transcription. Synthesis of mRNA is a well-studied mechanism in plant retrotransposons. A lot of them are currently inactive due to a combination of mutation and epigenetic silencing (Hirochika et al. 1996; 2000). Some of the elements are transcriptionally activated by various biotic and abiotic factors (Grandbastien 1998). For example, the expression of tobacco *Tnt1* and *Tto1* retrotransposons is greatly increased by several abiotic stresses, including wounding and application of methyl jasmonate, $CuCl_2$ and salicylic acid (Mhiri et al. 1997; Takeda et al. 1998). Similarly, various biotic stress factors such as inoculation with bacterial or fungal pathogens have been shown to activate transcription of these retrotransposons (Pouteau et al. 1994; Takeda et al. 1999; Mhiri et al. 1999). Regulatory sequences involved in the activation of retrotransposons by stress are similar to those of various stress responsive plant genes (Grandbastien et al. 1997). The stress conditions may be benefical for retrotransposons since it may be difficult for the host to silence transcription of the retrotransposon and maintain the defensive response at the same time (Sabot and Schulman 2006). On the other hand, induction of transcription by stress stimuli and subsequent transposition of retrotransposons may play a role in adaptive mutagenesis in the host and might be a general way for genomes to evolve (Morillon et al. 2000).

Translation of retrotransposon's mRNA which specifies the proteins needed for replication occurs in the cytoplasm. Gag proteins form virus-like particles (VLPs) where the reverse transcription occurs. Typically, two RNA molecules are packaged in one virus-like particle. This process is generally selective for the RNA encoding Gag protein that forms the VLP. The selectivity is directed by packaging signal in the RNA sequence that is recognized by Gag proteins (Sabot and Schulman 2006).

Reverse transcriptase is a key enzyme encoded by *pol* region that is responsible for the complex process of transcribing viral RNA into double-stranded DNA. Reverse transcription reaction is primed from a cellular tRNA that pairs to a sequence near the 5' LTR (the primer-binding site, pbs). Resulting DNA intermediate of discrete length is transferred from the 5' LTR to the 3´ end of a retrotransposon mRNA molecule, where reverse transcription proceeds. This transfer is mediated by identical sequences which are present at both 5´ and 3´ ends of the mRNA. A second priming event initiates at a short polypurine rich fragment of RNA (polypurine tract, PPT) near the 3' LTR. The resulting cDNA undergoes an additional

strand transfer. Finally, reverse transcriptase completes double-stranded DNA synthesis (for review of this complex process see Wilhelm and Wilhelm 2001; Kuman and Bennetzen 1999).

The integration of cDNA into genome is a unique enzymatic process shared by retroviruses and retrotransposons that is divided into several stages including binding and processing of LTRs ends, recognition and cleavage of a target DNA and joining of LTRs to a host genome (Hindmarsh and Leis 1999; Lewinski and Bushman 2005). All phases are catalyzed by enzyme integrase that is encoded within *pol* region. As a result of integration a short direct repeat flanking the retroelement (target site duplication, TSD) is produced. Specificity of integration has been studied in both retroviruses and retrotransposons. Variable factors are supposed to potentially influence DNA target selection and it was shown that specific retroelements have unique integration site preferences. For example, the HIV-1 retrovirus favours integration into actively transcribed regions and MLV retrovirus prefers integration near transcription start sites whereas ASLV retrovirus shows no such preference (Mitchell et al. 2004). Integration site selection was also well-studied in yeast LTR retrotransposons. *Ty1* and *Ty3* retrotransposons integrate upstream of PolIII transcribed genes while *Ty5* preferentially inserts into heterochromatin at telomeres (Buschman 2003). On the other hand, most of *Drosophila* LTR retrotransposons exhibit no specific integration into the host genome with the exception of three subgroups (*gypsy*, *ZAM* and *Idefix*) where the integration specificity was influenced by structural features of target DNA (Nefedova et al. 2011). In plants, centromeric regions of chromosomes consist of arrays of highly repetitive satellite DNA and centromeric retrotransposons (CRs). These retroelements possess an integrase chromodomain and they represent a fraction of CRM clade of Chromoviruses belonging to Ty3/gypsy group of LTR retrotransposons (Gorinsek et al. 2004; Gorinsek et al. 2005; Kordis 2005; Neumann et al. 2011).

3. Ogre retrotransposons

Ogre elements were first discovered in legumes, *Pisum sativum* and *Vicia pannonica* (Neumann et al. 2003; Neumann et al. 2006) and subsequently in other plant taxa (Salicaceae and Solanaceae; Macas and Neumann, 2007). Schematic representation of Ogre element from *Pisum sativum* is shown in Figure 1. These Ty3/gypsy plant retrotransposons are extreme in

their size (reaching up to 25 kbp) and they are often constitutively transcribed in contrast to most plant retrotransposons described so far. Their primer binding site (pbs) is complementary to the 3´ end of tRNAarg and their 3´ untranslated region (UTR) often contains array of tandem repeats varying in total size, sequence, and length of the repeat monomers (Macas et al. 2009). In some species they can constitute a large portion of the genome. In *Vicia pannonica* they make up 38 % of nuclear DNA and Ogre elements from *Glycine max* (*GmOgre*) described subsequently by Laten et al. (2009) represent the largest copy number family in soybean, accounting for 15 % of all elements identified in soybean genome (Du et al. 2010).

RT domains of *Ogre*-like retrotransposons were found also in Caryophyllaceae family (in *Silene latifolia)* during the screen of genomic library with probes prepared by amplifying the conservative parts of transposable elements (Cermak et al. 2008). FISH analysis revealed that Ogre elements are localized on all *Silene* chromosomes with the exception of Y chromosome where they are almost absent. In *Vicia pannnonica* and *Pisum sativum* the Ogre elements are evenly dispersed on all chromosomes (Neumann et al. 2006; Macas et al. 2009).

This distinct group of Ty3/gypsy elements is characterized by two specific features – presence of long extra ORF with unknown function upstream to *gag* and the stop codon-rich region separating the *gag-pro* and *rt/rh-int* coding sequences. It was shown that this region is spliced out of part of *Ogre-PS* transcripts (Neumann et al. 2003).

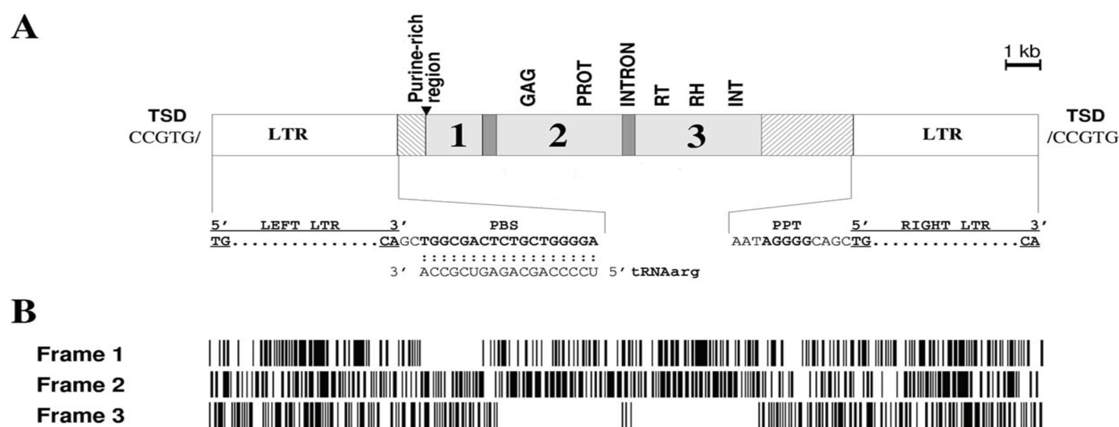

Figure 1. Schematic representation of Ogre element from pea. A. Structure of *Ogre-PS*. The coding sequences ORF1, ORF2 and ORF3 are represented by light grey boxes, dark grey boxes indicate regions separating individual ORFs, and cross-hatched boxes represent putative 5´ and 3´ untranslated region. LTR, long terminal repeat; PBS, primer binding site with complementarity to the 3´ end of tRNAarg; PPT, polypurine tract; TSD, target site duplication. B. Positions of stop codons in *Ogre-PS* sequence plotted in three reading frames. Each vertical line represents a single stop codon (Neumann et al. 2003, modified).

3.1 Splicing of Ogre elements

RNA splicing is a process that removes intervening sequences (introns) and joins coding regions (exons) in primary transcript. The chemistry of splicing involves two transesterification reactions. In the first one, a nucleophilic attack by the 2′-OH of an adenosine nucleoside in the intron (the branchpoint) cleaves the 5′ splice site. This results in the formation of a free 5′ exon and a lariat structure composed of the intron attached to the 3′ exon. In the second, the free 3′-OH of the 5′ exon attacks 3′ splice site breaking the phosphodiester bond while forming a new bond between the 5′ exon and 3′ exon, free intron is released in the form of lariat. The 5′ and 3′ termini of introns in most cases contain dinucleotides GT and AG, respectively (Breathnach et al. 1978; Mount 1982; Burset et al. 2000). Splice site usage can be modulated by auxiliary *cis*-elements known as exonic and intronic splicing silencers and enhancers which are recognized by *trans*-acting splicing factors (Ladd and Cooper 2002; Caceres and Kornblihtt 2002). The splicing process is catalyzed by the spliceosome, a complex RNA-protein aggregate, which consists of small nuclear ribonucleoprotein particles (snRNPs), each containing a small stable RNA bound by several proteins, and a large number of splicing factors (Jurica & Moore 2003).

Transcripts of retroelements can undergo the splicing events that generate subgenomic RNAs. Full-length retroviral RNA transcripts serve as a template for translation of *gag-pol* region and they also represent genomic RNAs which are packaged into progeny virion particles. Spliced transcripts of simple retroviruses encode the Env protein which is required for cell-to-cell transfer. Complex retroviruses generate both singly and multiply spliced transcripts that encode the *env* gene product and the sets of regulatory and accessory proteins (Coffin et al. 1997). *Drosophila* LTR retrotransposon *copia* uses alternative splicing of RNA to generate a subgenomic transcript in order to regulate the ratio between Gag and Pol proteins. Full length *copia* RNA containing *gag* and *pol* regions is expressed at a far lower level than subgenomic spliced RNA encoding *gag* products exclusively (Brierley and Flavell 1990). Alternative splicing as a mechanism of regulation of Gag:Pol ratio was also proposed for Ty3/gypsy centromeric retrotransposon of rice (CRR). The removing of RT-coding domain by splicing together with alternative usage of several different donor splice sites can suppress translation of *pol* region while *gag-pro* domains remain unaffected (Neumann et al. 2007). Splicing was also reported for *Bagy-2* (Ty3/gypsy retrotransposon of barley possessing an *envelope*-like ORF) where the spliced transcript is supposed to generate a subgenomic *env* product (Vicient et al. 2001).

In all the cases described above the splicing involves the coding part of the elements. Nevertheless, in some retrotransposons the non-coding sequences was shown to be spliced out and these events resemble the splicing of cellular introns. Functional introns were described in zebrafish non-LTR LINE element *ZfL2-1* (Tamura et al. 2007), and elements *Penelope* in *Drosophila virilis* and *Athena* in bdelloid rotifers (Arkhipova et al. 2003).

In Ogre retrotransposons the sequence between protease and reverse transcriptase domains contains several stop codons and it was predicted to be removed by splicing (Neumann et al. 2003). Splicing of this region would allow efficient translation of *rt/rh-int* region which can be in frame with *gag-pro* sequence. In *Pisum sativum* two forms of *Ogre-PS* transcripts were detected – the full-length and the spliced one. As splicing is only partial and transcripts with retained introns possess stop codons localized upstream of *rt/rh-int*, translation occurs for *gag-pro* sequences only and thus excess of Gag and PRO proteins can be expected. Thus, splicing as a mechanism for modulation of the proportion of element-encoded proteins might be used by Ogre retrotransposons. On the other hand, only full-length transcripts of Ogres were detected in *Vicia pannonica* (Neumann et al. 2006).

In the pea genome the spliced copies of *Ogre-PS* were detected (Neumann et al 2003). Spliced copy of LINE *ZfL2-1* retroelement was also revealed in zebrafish genomic DNA (Tamura et al. 2007). Nevertheless these events are very rare since the portion of spliced copies for either *Ogre-PS* or *ZfL2-1* retroelements in corresponding genomes is very low (in *Pisum sativum* the ratio of full-length to spliced *Ogre-PS* sequences is approximately 65:1). Integration of spliced copies into the genome must be suppressed because replication of the spliced transcripts would lead to gradual replacement of intron-containing elements with their spliced variants during genome evolution. Mechanism of recognition of full-length genomic RNA for encapsidation was well studied in retroviruses. In general the RNA sequences necessary and sufficient for RNA encapsidation (referred to as encapsidation signal or packaging signal) which is usually present only in the unspliced genomic RNA is recognized by the viral Gag protein. Nevertheless spliced mRNA can be also encapsidated even thought with very low frequency (Coffin et al. 1997; Jewell and Mansky 2000). This occasional packaging may allow subsequent reverse transcription and reintegration of spliced copy into the genome. Spliced and reintegrated transcripts in the genome have been observed also for HERV-H elements, a large family of endogenous retrovirus-like sequences found in the genomes of humans and other primates, which are supposed to be remnants of exogenous retrovirus infection of the germ line which became fixed in the population (Goodchild et al.

1995). Thus, such packaging signal might occur within the Ogre intron sequence that would lead to the preferable integration of unspliced copies into the genome.

Besides splicing of *Ogre-PS* transcripts, donor and acceptor splice sites were predicted for most of Ogre retrotransposons (Macas and Neumann 2007). The putative introns from different Ogre families are variable in both length and sequence. Hypothetical splicing of predicted introns results in a fusion of *gag-pro* and *pol* ORFs into one reading frame, and therefore a full polyprotein sequence could be translated.

3.2 Extra open reading frames in Ogre-like retrotransposons

Even though the *gag* and *pol* genes are believed to be necessary and sufficient for LTR retrotransposon life cycle, some of the elements possess additional open reading frames. The extra ORF found in the same position as *env* gene in retroviruses, downstream to *pol* region in the 3' untranslated region (3' UTR), is often referred to as an *env*-like gene. The *env*-like gene was found in a number of invertebrate Ty3/gypsy elements (Lerat and Capy, 1999) but *gypsy* of *Drosophila* is the only retroelement outside of the retroviruses for which *env*-like gene encoded protein is known to mediate infection (Kim et al, 1994; Song et al. 1994). The *env*-like genes were also reported for plants although their function remains unclear since the cell wall is thought to present a barrier to retroviral infection. These so-called endogenous plant retroviruses include Ty1/copia elements *SIRE-1* from *Glycine max, Endovir* from *Arabidopsis thaliana, copia_Endovir*-like elements from *Lotus japonicus* and *ToRTL1* from *Solanum lycopersicum* (Laten et al. 1998; Kapitonov and Jurka 1999; Laten 1999; Peterson- Burch et al. 2000; Holligan et al. 2006). Ty3/gypsy-like endogenous retroviruses comprise *Athila* from *A. thaliana*, *Cyclops* and *Pigy* from *Pisum sativum*, *Calypso* from *Glycine max, Bagy2* from *Hordeum vulgare* and *Rigy2* from *Oryza sativa* (Chavanne et al. 1998; Vicient et al. 2001; Wright and Voytas 2002; Neumann et al. 2005). The hypothetical Env-like proteins show little primary sequence similarity. Nevertheless, *env* genes of retroviruses also display low degree of conservation despite their functional role in cellular infection (Coffin et al. 1997; Laten et al. 1998). One of the most conserved features of the divergent retroviral Env proteins is the presence of transmembrane domain. This domain was predicted in the majority of Env-like proteins of plant endogenous retroviruses (Peterson-Burch et al. 2000; Wright and Voytas 2002).

In other plant retrotransposons the extra open reading frames (ORFs) in antisense orientation downstream to *pol* were found, namely in *Retand-2* from *Silene latifolia*, in *RIRE2* from rice and in *Grande1* from maize (Kejnovsky et al. 2006; Ohtsubo et al. 1999; Martínez-Izquierdo et al. 1997). The function of these hypothetical proteins remains unclear.

The presence of additional ORFs in plant retrotransposons is not restricted to the 3' UTR only. The additional ORF upstream to *gag* (called orf0) were previously described in *RIRE3, RIRE8A* and *RIRE8B* from *Oryza sativa* (Kumekawa et al. 1999) and in related element *FRetro3* in *Oryza brachyantha* (Gao et al. 2009).

Another LTR retrotransposons possessing extra ORF in 5' UTR are Ogre elements. It was shown for *Ogre-PS* from pea that its extra ORF contains predicted transmembrane domain (Neumann et al. 2003) and therefore might resemble *env*-like genes although it is located at different position. On the other hand, transmembrane domain was reported to be present in a broad spectrum of proteins (Krogh et al. 2001) and its prediction does not necessarily indicate the Env protein function. In Ogre retrotransposon from soybean (*GmOgre*) another extra ORF downstream to *pol* region that resembles *env*-like genes was identified. Env-like protein shares approximately 31% identity with that of *Calypso* (Laten et al. 2009, Du et al. 2010). Occurrence of extra ORFs in both 3' and 5' UTR makes *GmOgres* exceptional among the plant retrotransposons. Whether the extra ORFs found in Ogres have the same function as *env* genes of retroviruses remains unclear.

In a few cases, plant retrotransposons have acquired sequences that do not seem to play any role in their life cycle. LTR retrotransposon *Bs1* from maize has transduced sequences from three different cellular genes, namely proton-dependent membrane ATPase, 1,4-β-xylan endohydrolase and 1,3-β-glucanase. The transduction events generated a hybrid extra ORF containing *Bs1 gag* domain fused to the transduced sequences (Bureau et al., 1994; Jin and Bennetzen 1994; Palmgren 1994; Elrouby and Bureau, 2001). Moreover, the BS1 protein encoded by a novel chimeric gene may be involved in reproductive development of maize (Elrouby and Bureau, 2010). 3' noncoding region of another element, *Tat1-3* retrotransposon from *Arabidopsis thaliana*, contains the sequence similar to the nontranslated leader sequence of a gene for pyrroline-5-carboxylate reductase (AT-P5C1) and a remnant of related retrotransposon (Wright and Voytas 1998).

The capture of cellular genes is the unique mechanism of pathogenicity of close relatives of Ty3/gypsy retrotransposons, the retroviruses. Simplistic model for retroviral

transduction of cellular genes involves a retroviral integration upstream of the gene to be transduced and readthrough transcription from the retrovirus that generates a large RNA containing downstream cellular sequence. Alternatively, the viral and cellular sequences are fused by the deletion in DNA and the transcription produces a hybrid viral/cellular RNA. A chimeric RNA molecule and a normal viral RNA molecule are packaged into one virion. Non-homologous template switches between these two RNA molecules during reverse transcription may lead to the incorporation of the cellular gene sequence into the retroviral genome (reviewed in Coffin et al. 1997). This mechanism was proposed to give rise to the ORF1 of *Bs1* retrotransposon since this chimeric ORF is reminiscent of many transduced cellular genes (lack of native promoter and introns, fusion with GAG sequence; Elrouby and Bureau, 2001).

The extra ORFs of plant retrotransposons may represent transduced sequences as it was shown for *Bs1* element or they can play some role in elements life cycle. The question regarding the function of the extra ORFs still remains unanswered.

AIMS OF THE WORK

My PhD research was focused on two specific features of Ogre elements which make them unique among other retrotransposons: an extra open reading frame (ORF1) upstream of the *gag-pol* sequences and a stop codon-rich region separating protease and reverse transcriptase coding domains.

Previous data suggested that the non-coding region between protease and reverse transcriptase represents an intron which can be spliced out of Ogre transcripts in *Pisum sativum* (Neumann et al. 2003). Nevertheless, pea genome contains a small fraction of Ogre copies that also lack this region. Therefore, the aim of my work was to find out whether the spliced transcripts are produced from the intron-less copies in the genome or the splicing of Ogre full-length transcripts occurs in the cell. We studied transcription patterns of Ogre subfamilies in *Medicago truncatula* as well as the splicing of intron sequence of pea Ogre *in vivo* using transgenic hairy root culture system.

The second aim of my work was to investigate the origin and potential function of the additional ORF located upstream of *gag* region in Ogre elements. This study was performed using bioinformatic approaches and included all Ty3/gypsy retrotransposons in order to put extra ORFs from Ogres into phylogenetic context of the whole group or related elements.

RESULTS

Part I

Functional analysis of splicing of Ogre retrotransposons

Steinbauerová V., Neumann P., Macas J. (2008) Experimental evidence for splicing of intron-containing transcripts of plant LTR retrotransposon Ogre. Mol Genet Genomics 280: 427-436

Abstract

Ogre elements are a distinct group of plant Ty3/gypsy-like retrotransposons characterized by several specific features, one of which is a separation of the *gag-pol* region into two non-overlapping open reading frames: ORF2 coding for Gag-Pro, and ORF3 coding for RT/RH-INT proteins. Previous characterization of Ogre elements from several plant species revealed that part of their transcripts lacks the region between ORF2 and ORF3, carrying one uninterrupted ORF instead. In this work, we investigated a hypothesis that this region represents an intron that is spliced out from part of the Ogre transcripts as a means for preferential production of ORF2-encoded proteins over those encoded by the complete ORF2–ORF3 region. The experiments involved analysis of transcription patterns of well-defined Ogre populations in a model plant *Medicago truncatula* and examination of transcripts carrying dissected pea Ogre intron expressed within a coding sequence of chimeric reporter gene. Both experimental approaches proved that the region between ORF2 and ORF3 is spliced from Ogre transcripts and showed that this process is only partial, probably due to weak splice signals. This is one of very few known cases of spliced LTR retrotransposons and the only one where splicing does not involve parts of the element's coding sequences, thus resembling intron splicing found in most cellular genes.

RESULTS

Part II

Analysis of additional open reading frames in Ogre-like and other Ty3/gypsy elements

Steinbauerová V., Neumann P., Novák P., Macas J. (2012) A widespread occurrence of extra open reading frames in plant Ty3/gypsy retrotransposons. Genetica: accepted

Abstract

LTR retrotransposons make up substantial parts of most higher plant genomes where they accumulate due to their replicative mode of transposition. Although the transposition is facilitated by proteins encoded within the *gag-pol* region which is common to all autonomous elements, some LTR retrotransposons were found to potentially carry an additional protein coding capacity represented by extra open reading frames located upstream or downstream of *gag-pol*. In this study, we performed a comprehensive *in silico* survey and comparative analysis of these extra ORFs in the group of Ty3/gypsy LTR retrotransposons as the first step towards our understanding of their origin and function. We found that extra ORFs occur in all three major lineages of plant Ty3/gypsy elements, being the most frequent in the Tat lineage where most (77%) of identified elements contained extra ORFs. This lineage was also characterized by the highest diversity of extra ORF arrangement (position and orientation) within the elements. On the other hand, all of these ORFs could be classified into only two broad groups based on their mutual similarities or the presence of short conserved motifs in their inferred protein sequences. In the Athila lineage, the extra ORFs were confined to the element 3' regions but they displayed much higher sequence diversity compared to those found in Tat. In the lineage of Chromoviruses the extra ORFs were relatively rare, occurring only in 5' regions of a group of elements present in a single plant family (Poaceae). In all three lineages, most extra ORFs lacked sequence similarities to characterized gene sequences or functional protein domains, except for two Athila-like elements with similarities to *LOGL4* gene and part of the Chromoviruses extra ORFs that displayed partial similarity to histone H3 gene. Thus, in these cases the extra ORFs most likely originated by transduction or recombination of cellular gene sequences. In addition, the protein domain which is otherwise associated with DNA transposons have been detected in part of the Tat-like extra ORFs, pointing to their origin from an insertion event of a mobile element.

CONCLUSIONS

This thesis has contributed to the functional and structural characterization of Ogre retrotransposons that represent a distinct group of plant Ty3/gypsy retroelements.

It was shown that Ogre LTR retrotransposons possess functional introns which are spliced out from their mRNA. At least some portion of spliced transcripts of Ogre elements in *Medicago truncatula* arise from splicing events in the cells and not from the transcription of a small fraction of intron-less copies present in the genome. Expression of chimeric intron-containing GUS constructs *in vivo* allowed us to demonstrate that splicing of the intron from pea Ogre element is functional but revealed its low efficiency which left a part of the transcripts unspliced. Thus, it can be speculated that these unspliced transcripts serve as preferred templates for Ogre replication, facilitating the persistence of the intron-containing elements in the Ogre population.

In order to reveal the origin and function of additional ORFs of plant retrotransposons we performed a comprehensive *in silico* survey and comparative study of these extra ORFs in the group of Ty3/gypsy LTR retrotransposons. We found that extra ORFs occur in all three major lineages of plant Ty3/gypsy elements (Tat, Athila and Chromoviruses), being the most frequent in the Tat lineage including Ogre retrotransposons. Sequence similarities detected between 5' extra ORFs from Ogres and 3' extra ORFs from related *ATLANTYS-LC*-like elements strongly suggest that additional ORFs in these two groups are of common origin. They probably originated from an insertion of a transposable element because of the presence of plant mobile domain that is often associated with the transposase of MULE transposons. Transduction of cellular gene sequences represent another potential mechanism of extra ORF origin. Extra ORFs with similarity to cellular genes pointing to such event were found in two elements belonging to Athila lineage and in some Chromoviruses. The function of additional ORFs still remains unclear.

REFERENCES

Agrawal A, Eastman QM, Schatz DG (1998) Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 394: 744-751

Aravin A, Tuschl T (2005) Identification and characterization of small RNAs involved in RNA silencing. FEBS Lett 579: 5830-5840

Arkhipova IR, Pyatkov KI, Meselson M, Evgenev MB (2003) Retroelements containing introns in diverse invertebrate taxa. Nat Genet 33: 123-124

Atwood A, Lin JH, Levin HL (1996) The retrotransposon Tf1 assembles virus-like particles that contain excess Gag relative to integrase because of a regulated degradation process. Mol Cell Biol 16: 338-346

Biemont C, Vieira C (2006) Genetics: junk DNA as an evolutionary force. Nature 443: 521-524

Breathnach R, Benoist C, O'Hare K, Gannon F, Chambon P (1978) Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc Natl Acad Sci USA 75: 4853-4857

Brierley C, Flavell AJ (1990) The retrotransposon copia controls the relative levels of its gene products post-ranscriptionally by differential expression from its two major mRNAs. Nucleic Acids Res 18: 2947-2951

Bureau TE, White SE, Wessler SR (1994) Transduction of a cellular gene by a plant retroelement. Cell 77: 479-480

Burset M, Seledtsov IA, Solovyev VV (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res 28: 4364-4375

Bushman FD (2003) Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. Cell 115: 135-138

Caceres JF, Kornblihtt AR (2002) Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet 18: 186-193

Callinan PA, Batzer MA (2006) Retrotransposable elements and human disease. Genome Dyn 1:104-115

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

Cavalier-Smith T, Beaton MJ (1999) The skeletal function of non-genic nuclear DNA: new evidence from ancient cell chimaeras. Genetica 106: 3-13

Cermak T, Kubat Z, Hobza R, Koblizkova A, Widmer A, Macas J, Vyskot B, Kejnovsky E. (2008) Survey of repetitive sequences in *Silene latifolia* with respect to their distribution on sex chromosomes. Chromosome Res 16: 961-976

Chavanne F, Zhang DX, Liaud MF, Cerff R (1998) Structure and evolution of Cyclops: a novel giant retrotransposon of the Ty3/Gypsy family highly amplified in pea and other legume species. Plant Mol Biol 37: 363-375

Chen JM, Stenson PD, Cooper DN, Ferec C (2005) A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum Genet 117:411-427

Coffin JM, Hughes SH, Varmus HE (1997). Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbour, NY

Cordaux R, Batzer MA (2009) The impact of retrotransposons on human genome evolution. Nat Rev Genet 10: 691-703

Courage U, Döring HP, Frommer WB, Kunze R, Laird A, Merckelbach A, Müller-Neumann M, Riegel J, Starlinger P, Tillmann E, et al (1984) Transposable elements Ac and Ds at the shrunken, waxy, and alcohol dehydrogenase 1 loci in *Zea mays* L. Cold Spring Harb Symp Quant Biol 49: 329-338

Deininger PL, Moran JV, Batzer MA, Kazazian HH (2003) Mobile elements and mammalian genome evolution. Curr Opin Genet Dev 13: 651-658

Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 35: 41-48

Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemakes RC, Ma J (2010) Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J 63: 584-598

Elrouby N, Bureau TE (2001) A novel hybrid open reading frame formed by multiple cellular gene transductions by a plant long terminal repeat retroelement. J Biol Chem 276: 41963-41968

Elrouby N, Bureau TE (2010) Bs1, a new chimeric gene formed by retrotransposon-mediated exon shuffling in maize. Plant Physiol 153: 1413-1424

Evgenev MB, Arkhipova IR (2005) Penelope-like elements - a new class of retroelements: distribution, function and possible evolutionary significance. Cytogenet Genome Res 110: 510-521

Evgenev MB, Zelentsova H, Shostak N, Kozitsina M, Barskyi V, Lankenau D-H, Corces VG (1997) *Penelope,* a new family of transposable elements and its possible role in hybrid dysgenesis in *Drosophila virilis.* Proc Natl Acad Sci USA 94: 196-201

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9: 397-405

Feschotte C, Ning Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nature Rev Genet 3: 329-341

Feschotte C, Pritham EJ (2005) Non-mammalian c-integrases are encoded by giant transposable elements. Trends Genet 21: 551-552

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41: 331-368

Feschotte C, Wessler SR (2001) Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. Proc Natl Acad Sci USA 98: 8923-8924

Gao D, Gill N, Kim H-R, et al. (2009) A lineage-specific centromere retrotransposon in *Oryza brachyantha*. Plant J 60: 820-831

Gao X, Havecker ER, Baranov PV, Atkins JF, Voytas DF (2003) Translational recoding signals between gag and pol in diverse LTR retrotransposons. RNA 9: 1422-1430

Gentles AJ, Wakefield MJ, Kohany O, Gu W, Batzer MA, Pollock DD, Jurka J (2007) Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. Genome Res 17: 992-1004

Goodchild NL, Freeman JD, Mager DL (1995) Spliced HERV-H endogenous retroviral sequences in human genomic DNA: evidence for amplification via retrotransposition. Virology 206: 164-173

Goodwin TJ, Butler MI, Poulter RT (2003) Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. Microbiology 149: 3099-3109

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

Gorinsek B, Gubensek F, Kordis D. (2004) Evolutionary genomics of chromoviruses in eukaryotes. Mol Biol Evol 21: 781-798

Gorinsek B, Gubensek F, Kordis D. (2005) Phylogenomic analysis of chromoviruses. Cytogenet Genome Res 110: 543-552

Grandbastien MA (1998) Activation of plant retrotransposons under stress conditions. Trends Plant Sci 3: 181-187

Grandbastien MA, Lucas H, Morel JB, Mhiri C, Vernhettes S, Casacuberta JM (1997) The expression of the tobacco Tnt1 retrotransposon is linked to plant defense responses. Genetica 100: 241-252

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biol 5: 225

Hindmarsh P, Leis J (1999) Retroviral DNA integration. Microbiol Mol Biol Rev 63: 836-843

Hiom K, Melek M, Gellert M (1998) DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell 94: 463-470

Hirochika H, Okamoto H, Kakutani T (2000) Silencing of retrotransposons in *Arabidopsis* and reactivation by ddm1 mutation. Plant Cell 12: 357-369

Hirochika H, Sugimoto K, Otusuki Y, Tsugawa H, Kanda M (1996) Retrotransposons of rice involved in mutations induced by tissue culture. Proc Natl Acad Sci USA 93: 7783-7788

Holligan D, Zhang X, Jiang N, Pritham EJ, Wessler SR (2006) The transposable element landscape of the model legume *Lotus japonicus*. Genetics 174: 2215-2228

Hull R (2001) Classifying reverse transcribing elements: a proposal and a challenge to the ICTV. International Committee on Taxonomy of Viruses. Arch Virol 146: 2255-2261

Irwin PA, Voytas DF (2001) Expression and processing of proteins encoded by the *Saccharomyces* retrotransposon Ty5. J Virol 75: 1790-1797

Jewell NA, Mansky LM (2000) In the beginning: genome recognition, RNA encapsidation and the initiation of complex retrovirus assembly. J Gen Virol 81: 1889-1899

Jin YK, Bennetzen JL (1994) Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. Plant Cell 6: 1177-1186

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19: 68–72

Jurica MS, Moore MJ (2003) Pre-mRNA splicing: awash in a sea of protein. Mol Cell 12: 5-14

Kajikawa M, Okada N (2002) LINEs mobilize SINEs in the eel through a shared 3' sequence. Cell 111: 433-444

Kalendar R, Vicient CM, Anamthawat-Jonsson K, Bolshoy A, Schulman AH (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics 166: 1437-1450

Kapitonov VV, Jurka J (1999) Molecular paleontology of transposable elements from *Arabidopsis thaliana*. Genetica 107: 27-37

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

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

Kapitonov VV, Jurka J (2006) Self-synthesizing DNA transposons in eukaryotes. Proc Natl Acad Sci USA 103: 4540-4545

Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303: 1626-1632

Kejnovsky E, Kubat Z, Macas J, et al. (2006) Retand: a novel family of gypsy-like retrotransposons harboring an amplified tandem repeat. Mol Genetics Genomics 276: 254-263

Kim A, Terzian C, Santamaria P, Pélisson A, Purd'homme N, Bucheton A (1994) Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proc Natl Acad Sci USA 91: 1285-1289

Kordis (2005) A genomic perspective on the chromodomain-containing retrotransposons: chromoviruses. Gene 347: 161-173

Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. Int Rev Cytol 247: 165-221

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33: 479-532

Kumekawa N, Ohtsubo H, Horiuchi T, Ohtsubo E (1999) Identification and characterization of novel retrotransposons of the gypsy type in rice. Mol Gen Genet 260:593-602

Kuwabara T, Hsieh J, Muotri A, Yeo G, Warashina M, Lie DC, Moore L, Nakashima K, Asashima M, Gage FH (2009) Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. Nat Neurosci 12: 1097-1105

Ladd AN, Cooper TA (2002) Finding signals that regulate alternative splicing in the post-genomic era. Genome Biol 3: reviews0008

Lai, J, Li Y, Messing J, Dooner HK (2005) Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci USA 102: 9068-9073

Laten HM (1999) Phylogenetic evidence for Ty1-copia-like endogenous retroviruses in plant genomes. Genetica 107: 87-93

Laten HM, Majumdar A, Gaucher EA (1998) SIRE-1, a copia/Ty1-like retroelement from soybean, encodes a retroviral envelope-like protein. Proc Natl Acad Sci USA: 95: 6897-6902

Laten HM, Mogil LS, Wright LN (2009) A shotgun approach to discovering and reconstructing consensus retrotransposons ex novo from dense contigs of short sequences derived from Genbank Genome Survey Sequence database records. Gene 448:168-73

Lerat E, Capy P (1999) Retrotransposons and retroviruses: analysis of the envelope gene. Mol Biol Evol 16: 1198-1207

Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen FM (1993) Transposons in place of telomeric repeats at a *Drosophila* telomere. Cell 75: 1083-1093

Lewinski MK, Bushman FD (2005) Retroviral DNA integration – mechanism and consequences. Adv Genet 55: 147-181

Li PW, Li J, Timmerman SL, Krushel LA, Martin SL (2006) The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: implications for retrotransposition. Nucleic Acids Res 34: 853-864

Lippman Z, May B, Yordan C, Singer T, Martienssen R (2003) Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification. PLoS Biol 1:E67

Macas J, Koblížková A, Navrátilová A, Neumann P (2009) - Hypervariable 3' UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene 448: 198-206

Macas J, Neumann P (2007) Ogre elements – a distinct group of plant Ty3/gypsy-like retrotransposons. Gene 390: 108-116

Martínez-Izquierdo JA, García-Martínez J, Vicient CM (1997) What makes Grande1 retrotransposon different ? Genetica 100: 15-28

McClintock B (1951) Chromosome organization and genic expression. Cold Spring Harbor Symp Quant Biol 16: 13-47

Meignin C, Bailly JL, Arnaud F, Dastugue B, Vaury C (2003). The 50 untranslated region and Gag product of Idefix, a long terminal repeat-retrotransposon from *Drosophila melanogaster*, act together to initiate a switch between translated and untranslated states of the genomic mRNA. Mol Cell Biol 23: 8246-8254

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11: 1660-1676

Mhiri C, De Witt PJGM, Grandbastien MA (1999) Activation of the promoter of the Tnt1 retrotransposon in tomato after inoculation with the fungal pathogen *Cladosporium fulvum*. Mol Plant-Microbe Interact 12: 592-603

Mhiri C, Morel JB, Vernhettes S, Casacuberta JM, Lucas H, Grandbastien MA (1997) The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. Plant Mol Biol 33: 257-266

Miller WJ, McDonald JF, Nouaud D, Anxolabehere D (1999) Molecular domestication – more than a sporadic episode in evolution. Genetica 107: 197-207

Miskey C, Izsvák Z, Kawakami K, Ivics Z (2005) DNA transposons in vertebrate functional genomics. Cell Mol Life Sci 62: 629-641

Mitchell RS, Beitzel BF, Schroder AR, Shinn P, Chen H, Berry CC, Ecker JR, Bushman FD (2004) Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol 2: E234

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37: 997-1002

Morillon A, Springer M, Lesage P (2000) Activation of the Kss1 invasive-filamentous growth pathway induces Ty1 transcription and retrotransposition in *Saccharomyces cerevisiae*. Mol Cell Biol 20: 5766-5776

Mount SM (1982) A catalogue of splice junction sequences. Nucleic Acids Res 10: 459-472

Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435: 903-910

Nakamura TM, Morin GB, Chapman KB, Weinrich SL, Andrews WH, Lingner J, Harley CB, Cech TR (1997) Telomerase catalytic subunit homologs from fission yeast and human. Science 277: 955-959

Nefedova LN, Mannanova MM, Kim AI (2011) Integration specificity of LTR-retrotransposons and retroviruses in the *Drosophila melanogaster* genome. Virus Genes 42: 297-306

Neumann P, Koblížková A, Navrátilová A, Macas J (2006) Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. Genetics 173: 1047-1056

Neumann P, Navrátilová A, Koblížková A, Kejnovský E, Hřibová E, Hobza R, Widmer A, Doležel J, Macas J (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mobile DNA 2: 4

Neumann P, Požárková D, Koblížková A, Macas J (2005) PIGY, a new plant envelope-class LTR retrotransposon. Mol Genet Genomics 273: 43-53

Neumann P, Požárková D, Macas J (2003) Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced. Plant Mol Biol 53: 399-410

Neumann P, Yan H, Jiang J (2007) The centromeric retrotransposons of rice are transcribed and differentially processed by RNA interference. Genetics 176: 749-761

Ohshima K, Hamada M, Terai Y, Okada N (1996) The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. Mol Cell Biol 16: 3756-3764

Ohtsubo H, Kumekawa N, Ohtsubo E (1999) RIRE2, a novel gypsy-type retrotransposon from rice. Genes Genet Syst 74:83-91

Palmgren MG (1994) Capturing of host DNA by a plant retroelement: Bs1 encodes plasma membrane H(+)-ATPase domains. Plant Mol Biol 25: 137-140

Peterson-Burch BD, Wright DA, Laten HM, Voytas DF (2000) Retroviruses in plants? Trends Genet 16: 151-152

Pouteau S, Grandbastien MA, Boccara M (1994) Microbial elicitors of plant defense response activate transcription of a retrotransposon. Plant J 5: 535-554

Pritham EJ, Putliwala T, Feschotte C (2007) Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390: 3–17

Rio DC (1990) Molecular mechanisms regulating *Drosophila* P element transposition. Annu Rev Genet 24: 543-578

Roth DB, Craig NL (1998) VDJ recombination: a transposase goes to work. Cell 94:411-414

Sabot F, Schulman AH (2006) Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity 97: 381-3888

Sabot F, Schulman AH (2006) Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity 97: 381-388

Sela N, Mersch B, Hotz-Wagenblatt A, Ast G (2010) Characteristics of transposable element exonization within human and mouse. PLoS ONE 5: e10907

Sheen FM, Levis RW (1994) Transposition of the LINE-like retrotransposon TART to *Drosophila* chromosome termini. Proc Natl Acad Sci USA 91: 12510-12514

Singer T, McConnell MJ, Marchetto MC, Coufal NG, Gage FH (2010) LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends Neurosci 33: 345-354

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8: 272-285

Song SU, Gerasimova T, Kurkulos M, Boeke JD, Corces VG (1994) An env-like protein encoded by a *Drosophila* retroelement: evidence that gypsy is an infectious retrovirus. Genes Dev 8: 2046-2057

Takeda S, Sugimoto K, Otsuki H, Hirochika H (1998) Transcriptional activation of the tobacco retrotransposon Tto1 by wounding and methyl jasmonate. Plant Mol Biol 36: 365-376

Takeda S, Sugimoto K, Otsuki H, Hirochika H (1999) A 13-bp cis-regulatory element in the LTR promoter of the tobacco retrotransposon Tto1 is involved in responsiveness to tissue culture, wounding, methyl jasmonate and fungal elicitors. Plant J 18: 383-393

Tamura M, Kajikawa M, Okada N (2007) Functional splice sites in a zebrafish LINE and their influence on zebrafish gene expression. Gene 390: 221-231

Vicient CM, Kalendar R, Schulman AH (2001) Envelope-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. Genome Res 11: 2041-2049

Vitte C, Panaud O (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res 110: 91-107

Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28: 913-922

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973-982

Wilhelm M, Wilhelm FX (2001) Reverse transcription of retroviruses and LTR retrotransposons. Cell Mol Life Sci 58: 1246-1262

Wilson W, Malim MH, Mellor J, Kingsman AJ, Kingsman SM (1986) Expression strategies of the yeast retrotransposon Ty: a short sequence directs ribosomal frameshifting. Nucleic Acids Res 14: 7001-7016

Witte C-P, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci USA 98: 13778-13783

Wright DA, Voytas DF (1998) Potential retroviruses in plants: Tat1 is related to a group of *Arabidopsis thaliana* Ty3/gypsy retrotransposons that encode envelope-like proteins. Genetics 149: 703-715

Wright DA, Voytas DF (2002) Athila4 of *Arabidopsis* and Calypso of soybean define a lineage of endogenous plant retroviruses. Genome Res 12: 122-131

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

Yang N, Kazazian HH (2006) L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nat Struct Mol Biol 13: 763-771

Zilberman D, Henikoff S (2004) Silencing of transposons in plant genomes: kick them when they're down. Genome Biol 5: 249