The Transcriptional Regulation of Retrotransposon *BARE*

Wei Chang

Institute of Biotechnology and
Faculty of Biological and Environmental Sciences and
Finnish Graduate School in Plant Biology
University of Helsinki
MTT Agrifood Research

Academic Dissertation

To be presented for public criticism, with the permission of the Faculty of Biological and Environmental Sciences , University of Helsinki, in the auditorium 1041, Viikki Biocenter, Viikinkaari 5, Helsinki, on 9 December 2011, at 12 o'clock noon.

Supervisor:	Prof. Alan H. Schulman Director, Genomics Research, MTT Group Leader, Institute of Biotechnology, Univ. Helsinki
Followup group members:	Kristiina Mäkinen, Ph.D Department of Food and Environmental Sciences University of Helsinki *and* Mikko Frilander, Ph.D. Institute of Biotechnology University of Helsinki, Finland
Reviewers:	Kristiina Mäkinen, Ph.D Department of Food and Environmental Sciences University of Helsinki *and* Carlos Vicient, Ph.D Center for Reserarch in Agrigenomics (CRAG-CSIC) Barcelona, Spain
Opponent:	Professor Andrew Flavell Plant Science Devision University of Dundee
Custos:	Acting Professor Pekka Heino Department of Biosciences University of Helsinki

ISBN 978-952-10-7300-7 (pbk.)
ISBN 978-952-10-7301-4 (PDF)

Yliopistopaino, Helsinki University printing house, Helsinki 2011

To my family

Contents

Original publications

Abbreviations

Abstract ... 1

1. Introduction ... 2

1.1 The discovery of transposable elements ... 2

1.2 The classification of transposable elements ... 3

1.2.1 Class I (retrotransposons) ... 5

1.2.2 Class II (DNA transposons) ... 5

1.3 Autonomous and non-autonomous TEs ... 6

1.4 LTR retrotransposons and retroviruses ... 7

1.5 Structure, replication, life cycle, and autonomy of the LTR retrotransposons ... 8

1.5.1 Structure ... 8

1.5.2 Replication ... 9

1.5.3 Life Cycle ... 11

1.5.4 Autonomy ... 12

1.6 Retrotransposon *BARE1* ... 13

1.6.1 Structural features of *BARE1* ... 14

1.6.2 *BARE1* insertion site preferences and evolutionary conservation of RNA and cDNA processing sites ... 17

1.6.3 The role of *BARE1* in *Hordeum* genome evolution ... 17

1.7 *BARE2* is a chimeric and defective retrotransposon ... 18

1.8 *Cassandra* elements ... 19

1.9 The transcription regulation of retrotransposon, capping, splicing and polyadenylation ... 20

1.9.1 Capped & Uncapped RNA ... 20

1.9.2 Splicing ... 21

1.9.3 Polyadenylated & Non-polyadenylated RNA ... 22

1.10 The impact of REs on genome structure, function and evolution ... 24

1.11 Retrotransposons as molecular markers ... 26

2. Aims of the study ... 27

3. Materials and methods ... 28

3.1 Materials ... 28

3.2 Methods ... 28

3.2.1 DNA extraction ... 29

3.2.2 RNA extraction ... 29

3.2.3 Primer Design 29
3.2.4 Particle bombardment 29
3.2.5 LUC and GUS assays 29
3.2.6 *In vitro* transcription 30
3.2.7 RACE-PCR 30
3.2.8 Nuclease protection assay 31
3.2.9 Virus-like particle isolation 31
3.2.10 Polyribosome isolation 32
3.2.11 5' RLM-RACE 32
3.2.12 3' RLM-RACE 33
4. Results and discussion 34
4.1 The *BARE1* LTR functions as a promoter and some *BARE1* elements are transcriptionally active 35
4.1.1 *BARE1* is active in all tested barley tissues 35
4.1.2 Determination of the start site for *BARE* transcripts 35
4.1.3 Control of *BARE* transcription in barley tissues 36
4.2 *BARE1* transcript termination 37
4.2.1 Polyadenylated transcripts 38
4.2.2 Non-polyadenylated transcripts 38
4.3 Capping of *BARE* transcripts 39
4.4 *BARE1* transcripts are partially spliced 40
4.5 The features of non-autonomous element *BARE2* 41
4.6 *BARE1* transcripts are sorted: one pool for translation and another pool for encapsidation 41
4.7 The 3' ends of *Cassandra* transcripts are polyadenylated 42
5. General conclusions and future prospects 44
6. Acknowledgements 47
7. References 48

Original publications

This thesis is based on the following publications, which will be referred to in the text with their Roman numerals.

I

Chang W, Schulman AH 2008. *BARE* retrotransposons produce multiple groups of rarely polyadenylated transcripts from two differentially regulated promoters. The Plant Journal, 56(1): 40 - 50.

II

Kalendar R, Tanskanen J, **Chang W**, Antonius K, Sela H, Peleg O, Schulman AH 2008. *Cassandra* retrotransposons carry independently transcribed 5S RNA. Proc Natl Acad Sci U S A, 105(15): 5833-5838.

III

Chang W, Jääskeläinen M, Li S-P, Schulman AH. Distinct RNA pools as a replication strategy for the *BARE* retrotransposon (submitted).

Statement of my contribution on published articles:

Article I:

Planning: I made the experimental plan together with my supervisor Alan Schulman. Experimental part: I did all the experimental part of this work. Writing: I wrote manuscript draft and carried out revisions based on interactions with my supervisor.

Article II:

Planning: I participated in the work plan. Experimental part: I designed the experiment on mapping the beginning of transcripts of the *Cassandra* element, designed and did the experiment on mapping the end of *Cassandra* element. These two experiments solved the manuscript's weakness found by the reviewer and made the article's publication possible. Writing: I wrote these two parts of manuscript draft, and took part in the whole article revision.

Article III:

Planning: I made the experimental plan together with my supervisor. Experimental part: I did 70% of experimental work. Writing: I wrote the manuscript draft and carried out revisions based on interactions with my supervisor.

Abbreviations

Alu	the most common SINE family in primates
BAGY2	a barley retrotransposon family, *Gypsy* superfamily
BARE	a barley retroelement family
BARE1a	the fully sequenced *BARE1* clone in accession Z17327
BARE2	a non autonomous *BARE* retroelement family found in barley
bp	base pair
BPB	bromophenacyl bromide
cDNA	complementary deoxyribonucleic acid
copia	a retrotransposon family found in *Drosophila*
CTAB	cetyltrimethyl ammonium bromide
DR	direct repeat
DTT	dithiothreitol
EGTA	ethylene glycol tetraacetic acid
env	envelope gene
EST	expressed sequence tag
gag	GAG gene
GFP	green fluorescent protein
gypsy	a retrotransposon family found in *Drosophila*
HIV	human immunodeficiency virus
HTLV-1,-2	human T-lymphotropic virus
in	integrase gene
IR	inverted repeat
IRES	internal ribosomal entry sites
LARD	large retrotransposon derivative
LINE	long interspersed nuclear element
LTR	long terminal repeat
luc	gene coding for firefly luciferase protein, LUC
MITE	miniature inverted repeat transposable element

Morgane	a LTR retrotransposon family containing *pol*-like sequence only
mRNA	messenger ribonucleic acid
nt	nucleotide
ORF	open reading frame
PBS	primer binding site
pol	open reading frame containing possibly several genes
PPT	polypurine tract
R domain	a repeat present at both ends of a LTR-retrotransposon transcript
RACE	rapid amplification of cDNA ends
REs	retroelements
RSV	respiratory syncytial virus
rt	reverse transcriptase gene
SDS	sodium dodecyl sulfate
SINE	short interspersed nuclear element
SVA	short interspersed element (SINE-R), variable number of tandem repeats (VNTR) and Alu
TATA	motif for transcription initiation
TE	transposable element
TRIM	terminal repeat in miniature
tRNA	transfer ribonucleic acid
TSD	target site duplication
udiA	gene coding for GUS protein
UTL	untranslated leader sequence
VLP	virus-like particle
Wis	wheat insertion sequence, a retrotransposon family

Abstract

The purpose of this research project was to understand the steps of the retrotransposon *BARE* (BArley REtrotransposon) life cycle, from regulation of transcription to Virus-Like Particle (VLP) formation and ultimate integration back into the genome. Our study concentrates mainly on *BARE1* transcriptional regulation because transcription is the crucial first step in the retrotransposon life cycle. The *BARE* element is a Class I LTR (Long Terminal Repeat) retrotransposon belonging to the *Copia* superfamily and was originally isolated in our research group. The LTR retrotransposons are transcribed from promoters in the LTRs and encode proteins for packaging of their transcripts, the reverse transcription of the transcripts into cDNA, and integration of the cDNA back into the genome. *BARE1* is translated as a single polyprotein and cleaved into the capsid protein (GAG), integrase (IN), and reverse transcriptase-RNaseH (RT-RH) by the integral aspartic proteinase (AP). The *BARE* retrotransposon family comprises more than 10^4 copies in the barley (*Hordeum vulgare*) genome. The element is bound by long terminal repeats (LTRs, 1829 bp) containing promoters required for replication, signals for RNA processing, and motifs necessary for the integration of the cDNA. Members of the *BARE1* subfamily are transcribed, translated, and form virus-like particles.

Several basic questions concerning transcription are explored in the thesis: *BARE1* transcription control, promoter choice in different barley tissues, start and termination sites for *BARE* transcripts, and *BARE1* transcript polyadenylation (I). Polyadenylation is an important step during mRNA maturation, and determines its stability and translatability among other characteristics. Our work has found a novel way used by *BARE1* to make extra GAG protein, which is critical for VLP formation. The discovery that *BARE1* uses one RNA population for protein synthesis and another RNA population for making cDNA has established the most important step of the *BARE1* life cycle (III). The relationship between *BARE1* and *BARE2* has been investigated. Besides *BARE*, we have examined the retrotransposon *Cassandra* (II), which uses a very different transcriptional mechanism and a fully parasitic life cycle. In general, this work is focused on *BARE1* promoter activity, transcriptional regulation including differential promoter usage and RNA pools, extra GAG protein production and VLP formation. The results of this study give new insights into transcription regulation of LTR retrotransposons.

1. Introduction

Transposable elements are the most abundant components of most eukaryotic genomes. In the Triticeae, they comprise up to 85% of the genomic DNA. Transposable elements consist of retrotransposons (also called Class I transposable elements) and DNA transposons (Class II transposable elements). LTR retrotransposons (retrotransposons containing long terminal repeat sequences at both ends) are the most abundant transposable element class in grass genomes, of which barley is a member. The *BARE1* retrotransposon in barley, which was discovered by our group in previous work (Manninen and Schulman, 1993), is an especially active system and has been demonstrated to be transcriptionally active in somatic tissues and translated, processed, and assembled into virus-like particles, known as the VLPs (Suoniemi, 1996b; Jääskeläinen *et al.*, 1999). Transcription is the major step in many retrotransposons' life cycles, for example, tobacco retrotransposon Tto 1 (Hirochika, 1993) and the rice retrotransposon Tos 17 (Hirochika *et al.*, 1996). *BARE1* is a major, dispersed component of the *Hordeum* genome and is highly conserved in its functional domains (Suoniemi *et al.*, 1996a). Our group has shown that *BARE1* is a major factor in genome size dynamics in barley and its genus *Hordeum*, and that intra-element recombination plays a major role in controlling genome expansion resulting from *BARE1* integration (Vicient and Schulman, 2005). Very similar retrotransposons are transcribed as RNA and expressed as proteins in other cereals and grasses (Vicient *et al.*, 2001a). A large proportion of the plant LTR retrotransposons are partly or completely unable to synthesize their own machinery for transposition and are therefore non-autonomous elements. However, it is likely that most of these inactive or non-autonomous elements are able to retrotranspose as shown by their insertional polymorphism (Witte *et al.*, 2001). For example, the non-autonomous element *BARE2* has the possibility to be a partial parasite of the *BARE1* element because the GAG protein synthesized by *BARE1* can complement the defective GAG of the *BARE2* (Tanskanen *et al.*, 2007).

1.1 The discovery of transposable elements

Barbara McClintock (1902-1992) discovered the concept of transposable elements, called by her 'controlling elements', in maize by studying chromosome breakage and its genetic consequences. By studying patterns of coloration, she identified a mutation system with two

elements: one element caused the mutation and a second element controlled the activity of the first. She named these elements *Ds* (dissociator) and *Ac* (activator) respectively (Mcclintock, 1953). Both elements were noted to have the ability to change their position on chromosome 9, and evidence was assembled that *Ac* controlled its own mobility. She called this movement 'transposition'. In 1956, she reported another system of transposition in maize, the suppressor-mutator system, involving two genes and a series of transposable elements (Mcclintock, 1956). Her pioneering work revolutionized our thinking about genome stability and genome organization; she was awarded the Nobel Prize in 1983. Transposable elements (TEs) are mobile; their movement from one location to another in the genome was experimentally verified in bacteria in 1968 (Jordan *et al.*, 1968). The molecular details of McClintock's controlling elements were finally clarified as a transposable element in 1983 (Shure M *et al.*, 1983). TEs were found in *Drosophila melanogaster* and the yeast *Saccharomyces cerevisiae* in late 1970s (Finnegan *et al.*, 1978; Cameron *et al.*, 1979) and in *Caenorhabditis elegans* and human during the 1980s (Rosenzweig *et al.*, 1983; Paulson *et al.*, 1987). Their existence in filamentous fungi was discovered in the 1990s (Daboussi *et al.*, 1991). By now, mobile elements have been found in genomes of almost all organisms. They constitute more than 75% of the maize genome (Baucom *et al.*, 2009), 15% of the fruit fly genome (Hoskins *et al.*, 2002), more than 35% of the mouse genome (Waterston *et al.*, 2002) and about 50% of human DNA (Lander *et al.*, 2001). Many genes have been assembled or amplified by the action of the transposable elements.

1.2 The classification of transposable elements

All functional transposable elements have the ability to move from place to place in the genome—hence their designation as transposable elements—and most of them have their ability to amplify their copy number within the genome via this transposition, thereby providing a selectable function for their selfish or parasitic DNA (Le Rouzic *et al.*, 2007).

In 1989, Finnegan proposed the first TE classification system, which distinguished two classes by their transposition intermediate: RNA (Class I or retrotransposons) or DNA (Class II or DNA transposons). The transposition mechanism of Class I is commonly called 'copy-and-paste', and that of Class II, 'cut-and-paste' (Finnegan, 1989). The discovery of bacterial (Duval-Valentin *et al.*, 2004) and eukaryotic (Morgante *et al.*, 2005) TEs that copy and paste but without RNA intermediates, and of highly reduced non-autonomous TEs called miniature

inverted-repeat transposable elements (MITEs), has challenged the two-class system. A new classification system was proposed in the year 2007 (Wicker *et al.*, 2007), and it is the first unified hierarchical classification system that maintains two classes while applying mechanistic and enzymatic criteria. The highest level (Class) divides TEs by the presence or absence of an RNA transposition intermediate as before (Finnegan, 1989). Subclass, previously used to separate LTR from non-LTR (long and short interspersed nuclear element, LINE and SINE) Class I TEs, is used here to distinguish elements that copy themselves for insertion from those that leave the donor site to reintegrate elsewhere (Fig.1).

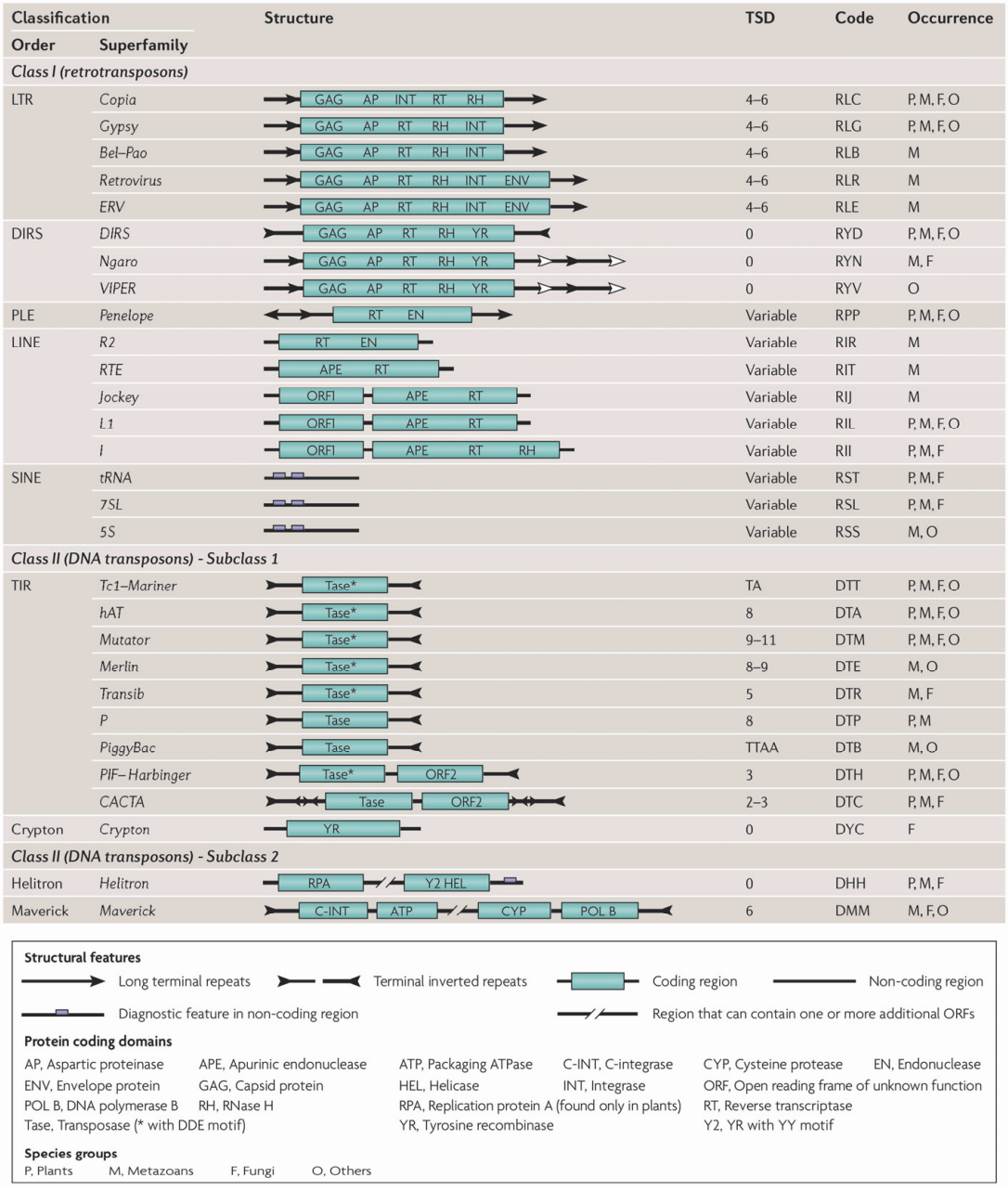


Figure 1. Proposed classification system for transposable elements (Wicker *et al.*, 2007).

1.2.1 Class I (retrotransposons)

Class I transposons are also known as retrotransposons because these elements use reverse transcriptase to move via an RNA intermediate. RNA polymerase II transcribes the original DNA into mRNA and this mRNA is then used as a template for reverse transcriptase to create a cDNA copy ready for insertion back into the genome (Havecker *et al.*, 2004). This is a replicative or 'copy-and-paste' method of transposition, and can generate high copy numbers, which in turn leads to an increase in genome size. In eukaryotes, LTR retrotransposons are the most widespread type of transposable elements. In plants especially, retrotransposons are the major constituents of the genome and are generally present in high copy numbers (Kumar and Bennetzen, 1999). Retrotransposons can be divided into five orders (Fig.1) on the basis of their mechanistic features, organization and reverse transcriptase phylogeny: LTR retrotransposons, *DIRS*-like elements, *Penelope*-like elements (PLEs), LINEs, and SINEs.

DIRS-like elements have several unusual structural features that distinguish them from typical LTR elements. For instance, they each encode a tyrosine recombinase (YR), but not a DDE-type integrase or an aspartic protease (Poulter and Goodwin, 2005). *Penelope*-like elements (PLEs) have been isolated from *Drosophila virilis*. The single ORF encoded by PLE consists of two principal domains, reverse transcriptase (RT) and endonuclease (EN), thereby forming a novel class of eukaryotic retroelements (Evgen'ev and Arkhipova, 2005). LINEs (Long Interspersed Elements) are widespread, autonomous non-LTR retrotransposons. They are 5–8 kb long elements with an internal polymerase II promoter, a poly(A) stretch and ORFs encoding the proteins necessary for their retrotransposition. The SINEs (Short Interspersed Elements) are short polymerase III transcribed elements, with an internal promoter and generally a poly(A) end. They are non-coding elements and thus depend on other genes for their mobility (Dewannieux and Heidmann, 2005).

1.2.2 Class II (DNA transposons)

Class II elements use DNA as the intermediate form in transposition and transpose directly mostly through a conservative 'cut-and-paste' mechanism (Finnegan, 1989). Class II elements have been divided into two subclasses. Subclass I consists of TIR (contains terminal inverted repeats in the sequence) and Crypton (contains tyrosine recombinase in the sequence), whereas Subclass II contains the helitron and Maverick groups (Fig. 1). A helitron is a transposon found in eukaryotes that is thought to replicate by a rolling-circle mechanism, whereas Maverick displays long terminal-inverted repeats but does not contain ORFs similar

to proteins encoded by other DNA transposons. Thus, not all DNA transposons transpose through a cut-and-paste mechanism; Subclass II of Class II uses a mechanism in which the transposon replicates itself to a new target site. In replicative transposition, the transposable element is duplicated during the reaction so that the transposing entity is a copy of the original element. Therefore replicative transposition is characteristic, not only for retrotransposons, but also in some Class II transposons (Duval-Valentin *et al.*, 2004).

1.3 Autonomous and non-autonomous TEs

Both classes of transposable elements may lose their ability to synthesize reverse transcriptase or transposase through spontaneous mutation, yet continue to move from one place to another in the genome because other elements are still producing the necessary enzymes. Moreover, transposable elements can be classified as either "autonomous" or "non-autonomous". An element is defined as autonomous simply if it appears to encode all the domains that are typically necessary for its transposition, without implying that the element is either functional or active. Non-autonomous TEs are defined as any group of elements that lacks some (or all) of the coding sequences found in autonomous elements (Tanskanen *et al.*, 2007). Usually, non-autonomous elements have a highly degenerate coding region, or even completely lack coding capacity. Occasionally, non-autonomous TEs lack some genes but still contain others; for example, members of the *Caspar* family (superfamily *CACTA*) often lack the transposase gene but still contain the second ORF (Wicker *et al.*, 2003), whereas the *BARE2* elements in the Triticeae have a conserved deletion that inactivates *gag* (Tanskanen *et al.*, 2007). Nevertheless, non-autonomous and autonomous elements usually still share strong sequence conservation and specific characteristics within their termini and in the 5' UTR (LTR retrotransposons), because these are required for packaging and transposition. Some non-autonomous elements might be cross-activated by autonomous partners from different families; for example, the Alu element in human mobilization depends on L1 protein of LINE-1 (Dewannieux *et al.*, 2003).

1.4 LTR retrotransposons and retroviruses

The LTR retrotransposons share similarities with retroviruses both in their genomic arrangement and in the mechanism of transposition (Fig. 2). The encoded proteins are organized 5′-LTR-*gag-ap-rt-rh-in*-LTR-3′ in superfamily *Gypsy* retrotransposons and in retroviruses. Retroviruses are more similar to *gypsy* of *Drosophila melanogaster* than they are to *copia* elements (Kumar and Bennetzen, 1999); retroviruses also have an *env* gene between *in* and the 3′ LTR. The strong internal sequence similarities within the *Copia* and *Gypsy* superfamilies suggest that they are lineages that have been separated since early in eukaryote evolution (Xiong and Eickbush, 1990). The IN structure of *BARE1* is extremely well conserved with HIV-1 and ASV over the enzymatic core domain defined by comparison to retroviral INs (Suoniemi *et al.*, 1998). Like retroviruses, LTR retrotransposons replicate through reverse transcription of their genomic RNA and they encode proteins with homology to the GAG and POL proteins of retroviruses. The main difference is that most LTR retrotransposons do not encode an envelope gene (*env*) and are not infectious, i.e. they carry on their replication cycle within a single cell (Wilhelm and Wilhelm, 2001). For retroviruses, three possible models can be invoked to explain the relationship between translation and encapsidation. In Model 1, any RNA can be translated or encapsidated. In Model 2, RNA is sorted into two non-equilibrating pools, one for translation and one for encapsidation. In Model 3, RNA can only be capsidated after it has been translated (Kaye and Lever, 1999). Among the retroviruses, Murine Leukemia Virus (MLV) uses distinct RNA pools for translation and reverse transcription (Messer, 1981), whereas for Human Endogenous Retrovirus 1 and 2 (HIV-1, -2), there is no such separation (Dorman and Lever, 2000). The question of RNA partitioning into pools for translation and reverse transcription does not seem to have been investigated for retrotransposons before. We attempted to show that *BARE1* forms two different RNA populations, one for translation and another for encapsidation.

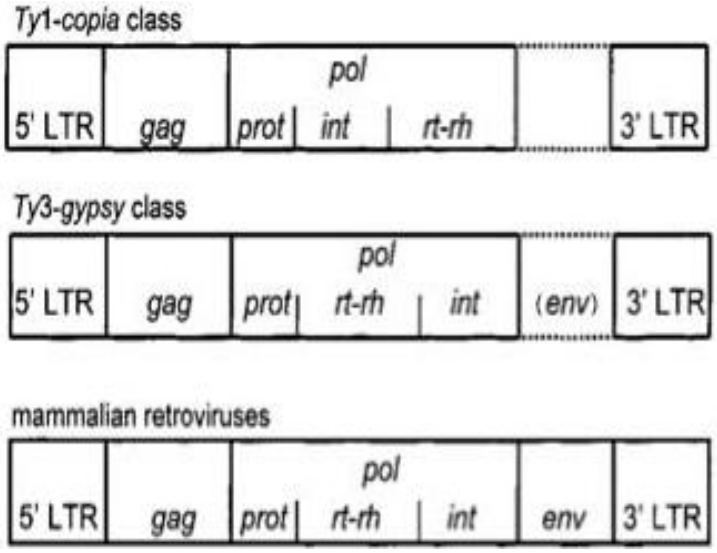


Figure 2. Organization of LTR REs

1.5 Structure, replication, life cycle, and autonomy of the LTR retrotransposons

1.5.1 Structure

All LTR retrotransposons are bounded by direct LTRs which in turn are flanked by short inverted repeats, usually containing 5'-TG…CA-3'. In autonomous elements, the 5' LTR provides the promoter function and the 3' LTR provides terminator and polyadenylation activities (Casacuberta and Santiago, 2003). The LTR sequence can range from a couple of hundred base pairs (*Bs1* of maize and *Tos17* of rice) to several thousand base pairs (*RIRE3* of rice) (Kumar and Bennetzen, 1999; Witte *et al.*, 2001). Between the LTRs is the internal coding region. In autonomous elements, this consists of the *gag* gene, which encodes proteins needed in the packaging of the retrotransposon RNA into the VLP inside which reverse transcription takes place, and the *pol* gene, which encodes protease, reverse transcriptase and RNaseH, and integrase. The protease cleaves the POL polyprotein, the reverse transcriptase and RNaseH are required for replication of the RNA strand back into DNA, and the integrase integrates the new cDNA copy of the retrotransposon into a new location in the genome (Havecker *et al.*, 2004). Superfamily *Gypsy* and *Copia* elements are autonomous and are defined according to their integrase (*in*) gene placement as well as their sequence similarities. *Gypsy* retrotransposons resemble retroviruses in gene order (LTR-*gag-ap-rt-rh-in*-LTR). *Copia* retrotransposons have a different gene order, where the integrase gene is placed

between the aspartic proteinase (*ap*) and reverse transcriptase (*rt*) genes (LTR-*gag-ap-in-rt-rh*-LTR) (Fig. 3). Morgane contains a partial Pol-like sequence in contrast to TRIM (terminal-repeat in miniature) and LARD (large retrotransposon derivatives) elements, which contain only non-coding sequences. All three groups are non-autonomous.

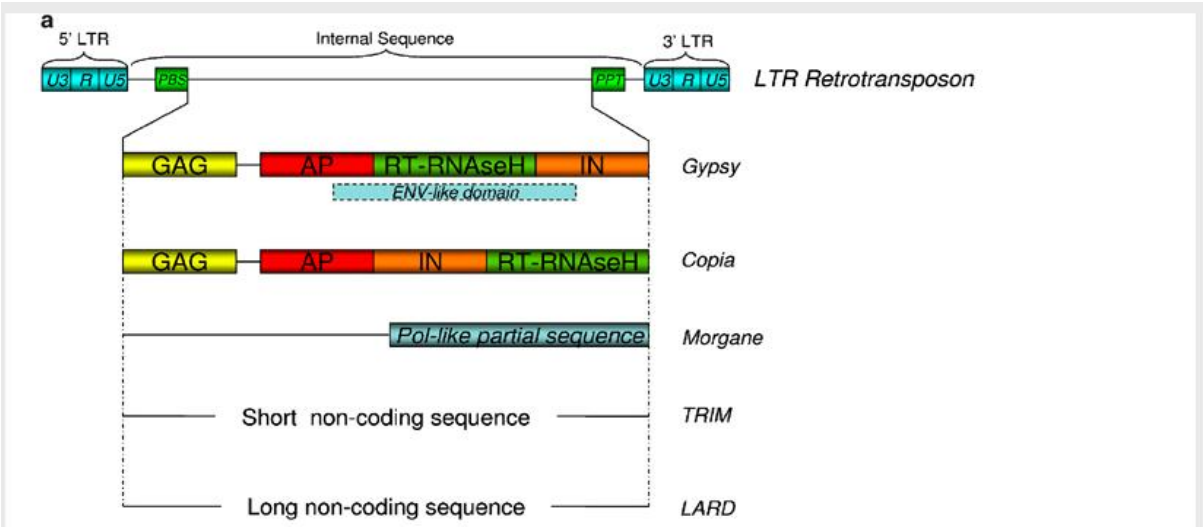


Figure 3. LTR retrotransposon structure and retrotransposon groups. The groups are separated according to the presence or absence of the *gag* and *pol* ORFs.

1.5.2 Replication

The LTR retrotransposon replicative cycle can be divided into several stages (Fig. 4), which are outlined below.

1. Retrotransposon mRNA molecules are synthesized by RNA polymerase II as are most cellular genes. The polyprotein mRNA molecule has the structure: 5'-R-U5-PBS-coding region-PPT-U3-R-3' (R= repeated RNA, U5= unique 5' RNA, PBS= primer binding site, PPT= polypurine tract, U3= unique 3' RNA). Transcription initiates at the 5' end of R in the 5' LTR and terminates at the 3' end of R in the 3' LTR.
2. The mRNA acts as a template for reverse transcriptase (RT) to synthesize a new DNA strand complementary to the element's internal sequences. The PBS on the mRNA molecule is complementary to a cellular RNA, usually the 3' end of a host tRNA. Thus, the tRNA can hybridize with retrotransposon RNA, and a free 3' hydroxyl group is provided from the tRNA which allows reverse transcriptase to synthesize a cDNA complement to the R and U5 regions of the 5'LTR. The reverse transcriptase then comes to the end of the template, the 5' end of the retrotransposon mRNA, and

cannot synthesize any more DNA. However, the RNaseH molecule encoded by the retrotransposon specifically digests the RNA in a DNA: RNA hybrid, thus freeing up a single-stranded DNA with homology to the R sequence that is also found at the 3' end of retrotransposon mRNA. Thus, the first template switch occurs. Hybridization between these sequences leads to a circular structure that allows a continuation of the reverse transcription until a single-stranded DNA complementary to all of the element-internal sequences is synthesized to generate a single-stranded DNA circle.

3. Second strand DNA synthesis is completed by the action of reverse transcriptase and RNaseH, primed from a PPT that lies just 5' to the 3' LTR, and involves a second template switch and breakage of base pairs in the U3-R region.
4. A double stranded linear DNA molecule is integrated back into the genome by integrase, which cuts both the donor and target molecules and leaves nicks staggered by 3-5bp. This causes the creation of the flanking direct repeats.

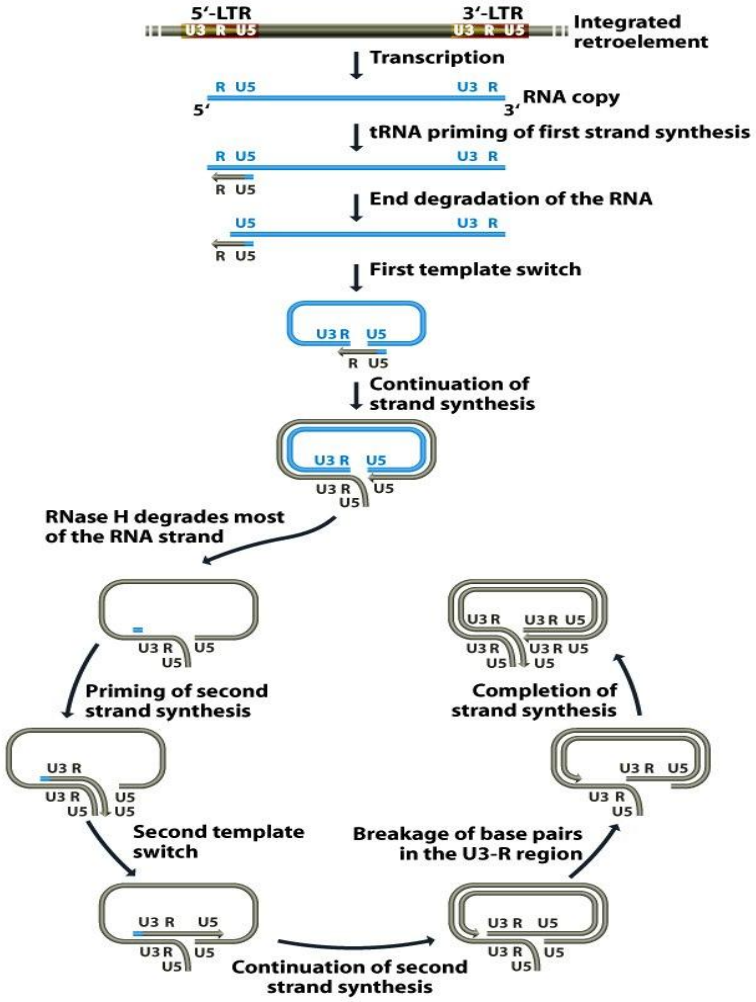


Figure 4. The replication of LTR retrotransposon (Source:'Genome III')

1.5.3 Life Cycle

Much of what we know about the mechanisms of LTR retrotransposition (Fig. 5) comes from work on yeast retrotransposons (Voytas and Boeke, 2002), but it is generally assumed that the mechanism is very similar among LTR retrotransposons from divergent hosts. First, a retrotransposon's RNA is transcribed by the general cellular RNA polymerase II from a promoter located within the 5' LTR. As in most of the cellular genes, the transcription starts after a TATA box. The RNA is then translated in the cytoplasm to synthesize the proteins that form the VLP, and carry out the reverse transcription and integration steps. Many retrotransposon RNAs contain a dimerization signal as do those of retroviruses. Typically, two RNA molecules are packaged into one VLP, and the RNA is subsequently made into a full-length cDNA copy through a reverse transcription reaction. The double-stranded cDNA molecule is formed by the two steps of strand transfer (Fig. 4). Another copy of the retrotransposon is added to the genome by integration back to the host DNA (Havecker *et al.*, 2004). Integration of DNA copies in a host genome is a necessary stage for the completion of the life cycle of retroviruses and LTR retrotransposons. All the LTR retrotransposons of the *Gypsy* superfamily demonstrate strict specificity in target DNA selection. Other LTR-retrotransposons do not show specificity of integration. The integration process can be divided into the following stages: (1) binding and processing of LTR ends; (2) recognition and cleavage of a target DNA in a host genome; and (3) joining of LTRs to the target DNA. The first stage occurs in the cytoplasm and the others in the nucleus. A target site duplication (TSD) is produced by repairing proteins at the last stage.

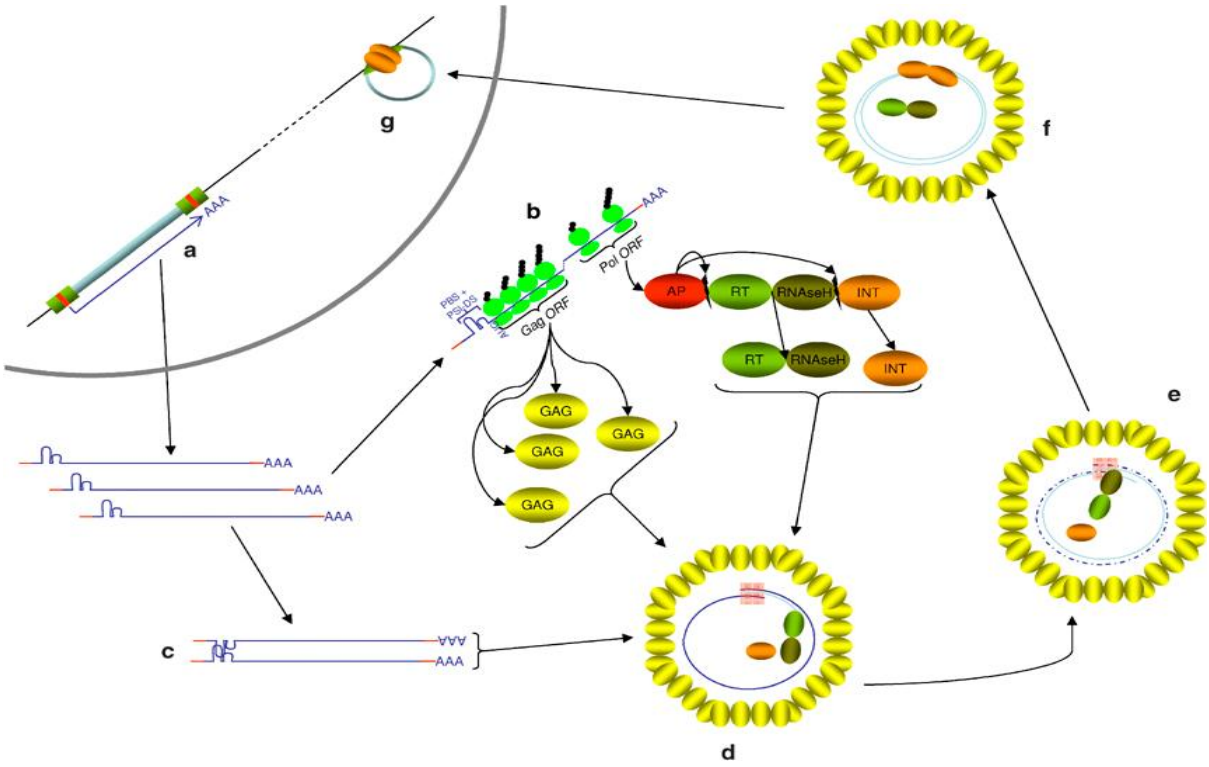


Figure 5. Theoretical life cycle of LTR retrotransposons. (a) Transcription of the mRNA, starting from the 5' R region to the 3' R region. (b) Translation and protein synthesis of active elements in GAG and POL, POL is further internally cleaved by AP into AP, RT-RNAseH and INT. (c) Dimerization of RNA before or during packaging. (d) Packaging of RNA and start of reverse transcription. (e) Degradation of the RNA matrix and initiation of synthesis of the second strand of the cDNA. (f) Completion of double-stranded cDNA synthesis. (g) Double-stranded break and integration of the newly synthesized copy in a new genomic location.

1.5.4 Autonomy

Families of retrotransposons containing individuals with an internal domain that is able to code for the requisite proteins are autonomous retrotransposon families. Individual copies may be, to varying degrees, transcriptionally or translationally competent (translation leading to a functional protein) or active. Transcriptionally and translationally active elements may complement the life cycle blocks of inactive or incompetent members of the same family in *cis* and that of other families or groups in *trans* to the extent that the complementation reduces the ability of the active element to propagate (Tanskanen *et al.*, 2007). Recent findings have identified large, structurally uniform retrotransposon groups in which no member contains the *gag*, *pol* or *env* internal domains. These groups are non-autonomous, yet individual elements may be active or inactive transcriptionally. Examples of non-autonomous groups are *LARD*, *TRIM* and *Morgane* (Fig. 6). Recent findings in soybean research have demonstrated that autonomous and non-autonomous retrotransposons appear to be both abundant and active in *Glycine* and *Phaseolus*. The impact of non-autonomous

retrotransposon replication on genome size appears to be much greater than previously appreciated (Wawrzynski *et al.*, 2008) and region-specific swapping of non-autonomous elements with autonomous elements generate various non-autonomous recombinants with LTR sequences from autonomous elements of different evolutionary lineages (Du *et al.*, 2010).

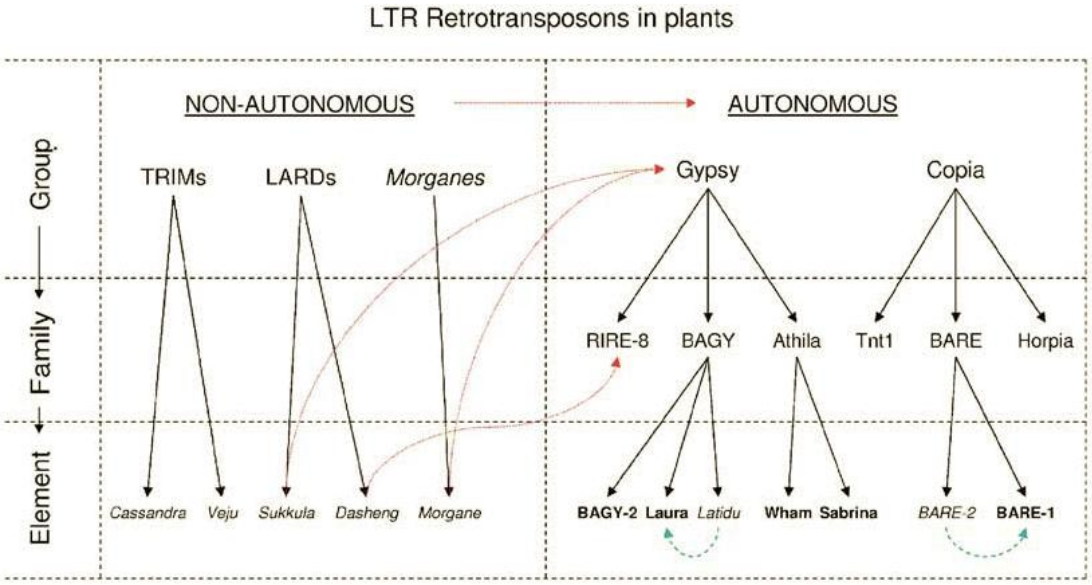


Figure 6. Autonomy and non-autonomy. Non-autonomous groups lack coding capacity for GAG and POL. Autonomous groups encode GAG and POL, which may be nevertheless inactive owing to mutations. The parasitic families (italics) are proposed to use the machinery of host elements (bold) in a *cis*- (dashed green arrows) or a *trans*- (dashed red arrows) mode (Sabot and Schulman, 2006).

1.6 Retrotransposon *BARE1*

The *BARE1* retrotransposon is a major, active component of the genome of barley (*Hordeum vulgare* L.) and other *Hordeum* species. *Copia*-like in its organization, it consists of 1.8 kb LTRs bounding an internal domain of 5275 bp that encodes a predicted polyprotein of 1301 residues. The polyprotein contains the key residues, structural motifs, and conserved regions associated with retroviral and retrotransposon GAG, AP, IN, RT, and RNaseH polypeptides. *BARE1* is actively transcribed and translated (Jääskeläinen *et al.*, 1999). Full-length members of the *BARE1* family constitute 2.8% of the barley genome. The *in situ* hybridization experiments for *BARE1* showed a uniform hybridization pattern over the whole of all chromosomes, excepting the centromeric, telomeric, and nucleolus organizer regions

(Suoniemi *et al.*, 1996a). Locally, *BARE1* occurs more commonly in repetitive DNA than in coding regions, forming clusters of nested insertions. Both barley and other *Hordeum* genomes contain a high proportion of *BARE1* solo LTRs (Vicient *et al.*, 1999a).

1.6.1 Structural features of *BARE1*



Figure 7. Retrotransposon *BARE1* organization. The regions of *BARE1* are abbreviated as follows: LTR, long terminal repeat (represented by grey boxes); 5' UTL, 5' untranslated leader; GAG, the capsid protein; AP, aspartic proteinase, IN, integrase; RT, reverse transcriptase; RH, RNaseH (each expressed protein represented by green boxes); 3' UTL, 3' untranslated leader. The red triangles represent inverted repeat and black errors represent host direct repeat. The TATA boxes contained in 5' LTR and start codon ATG contained in GAG are shown. The 5' UTL start from both TATA boxes of 5' LTR and the 3' UTL end in the 3' LTR are not shown in this figure.

LTR of *BARE1*

The *BARE1* LTRs are especially long, about 1.8 kb, and are conserved in *BARE1* populations (Suoniemi *et al.*, 1996a; Vicient *et al.*, 1999b). The organization of *BARE1* is represented in Fig. 7. Sequence examination revealed that *BARE1* LTRs contain two canonical TATA boxes (Manninen and Schulman, 1993), both of them being able to direct RNA transcription but under different conditions (Suoniemi, 1996b). Cellular RNA polymerase II is responsible for transcription of *BARE1*. The LTRs contain both the promoter necessary for transcription and the terminator and polyadenylation signals needed for RNA processing (Suoniemi, 1996b; I). The termini of the cDNA integration intermediate of *BARE1* LTRs are symmetrical and identical to that of HIV-1, but different from other plant retrotransposons (Suoniemi *et al.*, 1997). In addition, LTRs also contain the R region, lying between the transcription start and termination. Because the promoter functions in the 5′ LTR and the terminator in the 3′ LTR, the R region is found at both ends of the transcript. Deletion analysis of the promoter allows identification of regions important for expression in protoplasts (Suoniemi, 1996b). A region of 165 bp from the 3′ end of the LTR is composed of an array of tandemly repeated short sequences. Upstream from the tandem array, the region containing the two *BARE* promoters is the most variable region of the LTR. For the LTRs examined, more sequence divergence was found in the region surrounding the second TATA box than that surrounding the first

(Vicient and Schulman, 2005). The *BARE1* LTRs contain 6 bp imperfect inverted repeats at their ends with the canonical 5' TG…CA 3' terminal sequences present in most retroviruses and retrotransposons. The genome TSD is established to be 5 bp (Suoniemi *et al.*, 1997).

Internal domain of *BARE1*

The region between the LTRs in retrotransposons forms the internal domain. The PBS, a tRNA-complementary sequence at the end of 5' LTR, is used by reverse transcriptase to initiate the cDNA minus-strand synthesis (Marquet *et al.*, 1995). Priming of the cDNA plus-strand is initiated at the PPT located just 5' of the 3' LTR (Heyman *et al.*, 1995). The PBS of *BARE1*, as well as of several other plant retrotransposons is complementary to the tRNAi ^{met} (Suoniemi *et al.*, 1997). The PPT is also highly conserved in the *BARE1* family. Two important domains that exist in this part of the sequence before the translation start codon are the Packaging SIgnal (PSI) and the DImerization Signal (DIS) domains. The PSI is responsible for packaging of the retroviral mRNA into its specific viral particle (Clever *et al.*, 2002), whereas DIS directs kissing-loop interaction and is involved in the dimerization of the retroviral RNA during or just before mRNA packaging (Proudfoot, 2004). The untranslated leader (UTL) between the start of transcription and the start of translation is about 2 kb in *BARE1* and is conserved in length among the various copies of the element. Two other retrotransposon families, *stonor* of maize (Marillonneta and Wesslera, 1998) and *RIRE1* of rice (Noma *et al.*, 1997) also possess a 2 kb UTL sequence. The *BARE1* UTL region contains at least 51 putative ATG codons, similar to picornavirus. The RNA of picornavirus initiates translation internally, via an internal ribosome entry site (IRES) element present in their 5' untranslated region (Jackson and Kaminski, 1995). However, we do not have knowledge on the translation initiation of *BARE1*. The *BARE1* internal domain encodes a predicted polypeptide, the key residues of the peptides which, when aligned with their counterparts from retrovirus and other *copia*-like retrotransposons, are well conserved (Suoniemi *et al.*, 1998).

- *gag* gene product

Gag encodes proteins that form VLPs, which package retroelement mRNAs (Irwin and Voytas, 2001). *BARE1 gag* is 843 nt in length; the predicted protein contains a typical zinc finger domain (CCHC) sequence at the C terminal and nuclear localization signal at the N terminal. The zinc finger domain is also present in plant retrotransposons *del*1-46 of lily (Smyth *et al.*, 1989), *Tnt1* of tobacco (Grandbastien *et al.*, 1989) and *Zeon*-1 from maize (Hu

et al., 1995). The retroviral GAG precursor is formed from three essential domains, namely the matrix (MA), the capsid (CA) and the nucleocapsid (NC). The nucleocapsid protein consists of two CCHC zinc fingers flanked by highly basic regions (Morellet *et al.*, 1992). *BARE1* has been translated and the capsid protein and integrase components of the predicted polyprotein are processed into polypeptides of expected size (Jääskeläinen *et al.*, 1999).

- *pol* gene product

Products of the *pol* gene include the aspartic proteinase (PR) that processes the retroelement polyproteins, an integrase (IN) that inserts the cDNA into a new site in the host chromosome and a reverse transcriptase (RT) and its associated RNase H (RH), which synthesize a cDNA copy of the retroelement from the template mRNA (Irwin and Voytas, 2001).

- Proteinase

PR is encoded by *pol*, located at its N-terminus. It is required to release the other enzymes from the Pol precursor and is involved in processing of GAG (Gulnik *et al.*, 2000). The *BARE1* protease is 451 nucleotides long and its active catalytic site, which is located at its N terminus, contains a conserved motif DTG. Additionally, a tryptophan located three residues upstream of the catalytic aspartate is also conserved among *copia* elements (Peterson-Burch and Voytas, 2002).

- Integrase

The IN binds and inserts the retroelement cDNA into host chromosomal DNA. The IN features three domains: the N-terminal domain containing a 'zinc finger' – like motif (HHCC domain), the catalytic domain included in a central region of approximately 150 amino acids characterized by the DD(35)E motif, and the C-terminal domain, which is not highly conserved but contains the GKGY motif, unique to the *Copia* superfamily (Peterson-Burch and Voytas, 2002). The DD(35)E motif is a constellation of three invariant acidic amino acids, the last two separated by 35 amino acids. These acidic residues are required for all catalytic functions of IN and have been proposed to bind the essential metal cofactor(s), Mn^{2+} or Mg^{2+} (Andrake and Skalka, 1996). The integrase sequence of *BARE1* covers 1215 nt; its predicted translation and the secondary and tertiary structures are extremely well conserved when compared to HIV and ASV INs, although the sequence region used for modeling are only about 245 nt identical (Suoniemi *et al.*, 1998).

- Reverse transcriptase-RNaseH

The reverse transcriptase-RNaseH region is needed both for the synthesis of the polypurine, plus-strand primer and for the reverse transcription of the RNA transcript into a cDNA copy. The reverse transcriptase enzyme converts the RNA (5' R-U5---ORF---U3-R 3') molecule into a double-stranded DNA (5' U3-R-U5---ORF---U3-R-U5 3'). The *BARE1 rt* is 603 nt and *rnaseH* is 784 nt. Reverse transcriptase activity has been detected in VLPs showing that cDNA synthesis can happen inside them (Jääskeläinen *et al.*, 1999). The RT region between Tyr^{864} and Ala^{889} is fairly conserved among *BARE1* and other *copia* retrotransposons (Manninen and Schulman, 1993; Xiong and Eickbush, 1990).

1.6.2 *BARE1* insertion site preferences and evolutionary conservation of RNA and cDNA processing sites

In previous work carried out in our research group, inverse PCR was used to examine the sequences flanking the *BARE1* insertion sites. It was established the TSD as 5 bp, indicating that the *BARE1* IN generates a 5 bp staggered cut during the integration reaction. Of the thirteen identified integration sites, nine were other *BARE1* elements and three were other retrotransposons, one of them was a *Grande*-like element previously reported only for maize and its near relatives (Suoniemi *et al.*, 1997). The termini of the cDNA integration intermediate were found to be symmetrical and identical to that of HIV, but different from other plant retrotransposons. The dinucleotides at the end of cDNA were identified to be symmetrical 5' AC (coding strand) 3' CA (non-coding strand). The dinucleotides of tobacco Tnt1 element examined were in most cases AT at the 5' end of the linear cDNA (coding strand) and 3' TC at the 3'end (non-coding strand) (Feuerbach *et al.*, 1997).

1.6.3 The role of *BARE1* in *Hordeum* genome evolution

The *BARE1* family is a major, dispersed component of the barley genome (Suoniemi *et al.*, 1996a). Among the species in the *Hordeum* genus, *BARE1* is present on an average in 14,000 copies, with 16,000 copies in barley (Vicient *et al.*, 1999a); the number varies across the genus. Based on these copy numbers and on the genome sizes, full-length *BARE1* comprises 0.8% to 5.7% of the genome in genus, and this measure also varies across the genus (Kankanpää *et al.*, 1996). The *BARE1* copy number and genome size are positively correlated, indicating that *BARE1* is an important although not the sole contributor to the

differences in genome size among the species of genus *Hordeum*. Sequencing of *BARE1* flanking regions demonstrates that 69% of the flanking sequences are retrotransposons with 62% of these being *BARE1* elements and that *BARE1* elements are generally clustered near each other in the genome (Suoniemi *et al.*, 1997). *Hordeum* genomes contain a large excess of *BARE1* LTR sequences relative to the internal domain, and the excess LTRs appear to have their origin in homologous recombination between the LTRs of a single element, which removes the internal regions and leaves behind a single recombinant LTR. Recombination between LTRs would be expected to reduce the complement of functional retrotransposon in the genome, limiting but not eliminating the contribution of *BARE1* to the genome size. Consistent with these observations, the LTR excess is inversely correlated with the proportion of genome occupied by *BARE1* (Vicient *et al.*, 1999a; Kalendar *et al.*, 2000). The tandem copies of *BARE1* generated by recombination between the right LTR of one element and the left LTR of another element downstream from the first could eliminate non-*BARE1* DNA located between LTRs. *BARE1* displays nearly a three-fold intra-specific copy number variation in natural populations of the wild barley *Hordeum spontaneum* (Kalendar *et al.*, 2000). Correlations between *BARE1* copy number, genome size, and local environmental conditions suggest, for the first time, a testable molecular mechanism linking habitat with retrotransposon induction in natural populations (Wendel and Wessler, 2000).

1.7 *BARE2* is a chimeric and defective retrotransposon

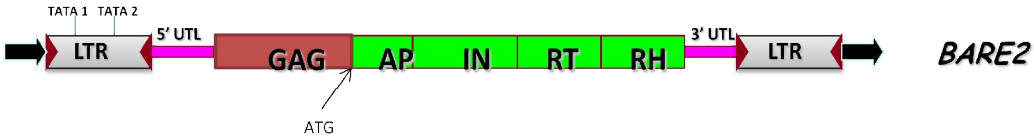


Figure 8. The structure of retrotransposon *BARE2*. The regions of *BARE2* are abbreviated as follows: LTR (5'), 5' UTL, GAG (the GAG of *BARE2* is not expressed, and therefore represented by a red box), AP, IN, RT, RH, RNaseH, 3' UTL and LTR (3'). The red triangles represent inverted repeats and black arrows represent host direct repeats. The TATA boxes contain in 5' LTR and start codon ATG contain at the end of GAG are shown. The 5' UTL start from both TATA boxes of 5' LTR and 3' UTR end in 3' LTR are not shown in this figure.

BARE2 LTRs are at least 96% similar to *BARE1* LTRs, but their internal sequences are very different. The *BARE2* group contains more genome copies than *BARE1* (Tanskanen *et al.*, 2007). It also creates a 5 bp target site during an insertion event as *BARE1* does. The full-length *BARE2* retrotransposon displays abrupt switches in sequence similarity between two

related families of elements, *BARE1* and *Wis*-2. Hence, it appears that the *BARE2* is a mosaic or chimeric element that was generated by stand switching during replication (Vicient and Schulman, 2005). These two elements are present in the Triticeae and related species, and are together polymorphic among closely related accessions. *BARE2* elements are unable to synthesize their own GAG protein because when the first ATG is lost from *gag* ORF, the succeeding ATG is downstream of *gag*. However, *BARE2* sequences are conserved with *BARE1* in several domains which are critical for its life cycle. The structure of retrotransposon *BARE2* is presented in Fig.8.

1.8 *Cassandra* elements

Cassandra elements belong to the TRIM group of retrotransposons (Fig.3). TRIM elements have the following features: terminal direct repeat sequences between 100 and 250 bp in length; an internal domain of 100–300 bp. The internal domain contains a PBS and PPT but lacks the coding domains required for mobility. Thus, TRIM elements are not capable of autonomous transposition and probably require the help of mobility-related proteins encoded by other retrotransposons (Witte *et al.*, 2001). *Cassandra* elements universally carry conserved 5S RNA sequences and associated RNA polymerase III promoters and terminators in their LTRs (II). They are found in all vascular plants that have been investigated. Uniquely for LTR retrotransposons, the full length *Cassandra* element is 565-860 bp in length in which the length of LTR is 240-350 bp and the core domain is 65-260 bp long. The core domain contains 5S RNA sequences with conserved A, IE and C domains. The *Cassandra* polymerase III promoter contains a termination signal although it is different from the canonical termination signal for cellular 5S. Polyadenylation of pol III transcripts is rare. However, many *Cassandra* 5S, but not cellular 5S genes, possess a putative polyadenylation signal, CAA(T/C)AA, located 17 nt before the pol III terminator at the beginning of the 5S domain; its distance from the terminator is quite typical. Although most of polymerase III transcripts are non-polyadenylated, polyadenylated cellular 5S RNA has been found (Fulnecek and Kovarik, 2007). The Structure of a *Cassandra* element is represented in Fig.9.

Cassandra elements otherwise possess the organization typical for a Class I retrotransposon, in which polymerase II is used for RNA synthesis. The polymerase III promoter located in the LTR therefore raises the question of which polymerase is used for *Cassandra* transcription. To answer this question, we set up experiments to find out the capping status of

the transcripts, because polymerase III transcripts are uncapped. We have also investigated the organization of the *Cassandra* element in the plant kingdom including the transcription motifs, organizations and insertional polymorphisms.



Figure 9. Structure of a *Cassandra* element. Flanking genomic DNA is indicated as a wavy line with the TSDs as arrowheads. The element components, including the PBS and PPT sites, are shown as boxes. The terminal inverted repeats (TIRs) of the LTRs are shown as black triangles, and the 5S domain is hatched. The internal sequence contains PBS and PPT sites and a non-coding sequence.

1.9 The transcription regulation of retrotransposon, capping, splicing and polyadenylation

1.9.1 Capped & Uncapped RNA

Nuclear capping occurs co-transcriptionally on all RNA polymerase II-synthesized RNAs in three steps (Lewis and Izaurralde, 1997; Furuichi and Shatkin, 2000; Gu and Lima, 2005). The N-terminal triphosphatase domain of capping enzyme hydrolyzes the 5′ triphosphate of nascent pre-mRNA to a diphosphate. Guanosine monophosphate (GMP) is subsequently transferred onto this diphosphate to create a G(5′)ppp(5′)N terminus, where N denotes the first transcribed nucleotide. This terminus is then methylated by RNA (guanine-7)-methyltransferase to generate the m^7G(5′)ppp(5′)N cap. The cap participates in many aspects of pre-mRNA and mRNA metabolism, including splicing, polyadenylation, nucleocytoplasmic transport, translation, quality control and stability (Maquat, 2004; Isken, 2007). Routine translation of the majority of mRNAs in eukaryotic cells is initiated by a cap-dependent mechanism. This involves recognition and binding of the cap structure (m^7GpppN) on the 5′ ends of mRNAs by the eukaryotic translation initiation factor, eIF4F. Upon binding an mRNA, eIF4F recruits the small ribosomal subunit via eIF3 interaction and additional initiation factors, and then this 43S complex scans 5′–3′ until the first AUG initiation codon is encountered. The 60S subunit is then recruited and elongation begins (Kozak, 1989).

There has been little evidence for uncapped mRNAs in eukaryotic cells. Compelling evidence for uncapped mRNAs appeared only recently through experiments in *Arabidopsis thaliana*.

The populations were identified by their 5′-monophosphate ends, which enable primer ligation (Gregory *et al.*, 2008). Another group also used the same method to tag uncapped mRNAs but they started with poly(A) RNA and the uncapped mRNA was identified by a microarray technique. The finding that the levels of specific uncapped transcripts varied depending on the stage of floral development independently of full-length mRNA abundance provided the evidence that decapping was regulated and physiologically significant (Jiao *et al.*, 2008). Sequence-related transcripts usually shared similar levels of uncapping, offering additional evidence that decapping is to some extent an active, rather than a passive, process (Schoenberg and Maquat, 2009). Many viruses have uncapped RNAs as their genomic RNA, for example, poliovirus, mengovirus and satellite tobacco necrosis virus (STNV) RNA. Uncapped poliovirus mRNAs harbor internal ribosome entry sites (IRES) in their long and highly structured 5' non-coding regions, and such IRES sequences are required for viral protein synthesis (Haller *et al.*, 1993). The IRES of hepatitis C virus (HCV) RNA contains >300 bases of a highly conserved 5′-terminal sequence, most of it located in the uncapped 5′-untranslated region (5′-UTR) upstream from the single AUG initiator triplet at which the translation of the HCV polyprotein begins (Lyons, 2001).

1.9.2 Splicing

In most eukaryotic genes, the coding information (exons) is interrupted by introns that are removed from pre-mRNA to produce mature mRNA. This process of intron removal (and subsequent exon ligation), termed splicing, is carried out by the host spliceosome, which is composed of over 200 different proteins and five small nuclear RNAs (U1, U2, U4, U5 and U6) (Jurica and Moore, 2003). The 5′ and 3′ termini of introns contain the highly conserved dinucleotides GU and AG respectively (Mount, 1982; Burset *et al.*, 2000). Other sequences around the 5′ and 3′ splice sites (the splice donor and acceptor sites, respectively) are, however, poorly conserved, suggesting that the exact recognition of a genuine splice site among many cryptic splice sites requires additional *cis* elements. Most of these *cis* elements are considered to be recognized by *trans*-acting splicing factors and generally are referred to as exonic or intronic splicing enhancers and exonic or intronic splicing silencers. Splicing, especially alternative splicing (in which the order of exon ligation varies), is a major contributor to protein diversity in metazoans (Black, 2003).

Although the splicing has been well documented for some groups of retroelements (REs) like retroviruses and LINEs (Rabson and Graves, 1997; Belancio *et al.*, 2006; Tamura *et al.*, 2007), it has so far been reported only for a few LTR retrotransposons. It occurs in the transcripts of the envelope-class retrotransposon *Bagy-2*, where it generates a subgenomic RNA lacking almost the entire *gag-pol* sequence, thereby enabling expression of the downstream *env* gene (Vicient *et al.*, 2001b). It has been demonstrated that alternative splicing of RNA from *Drosophila* retrotransposon *copia* is involved in the regulation of the ratio between GAG and the *pol* proteins, as the full-length *copia* RNA containing both *gag* and *pol* regions is translated to protein at a far lower level than spliced subgenomic RNA encoding GAG only (Brierley and Flavell, 1990). Differential expression is also observed in retroviruses and Ty, but for these the mechanism is a frameshift (Voytas and Boeke, 1993). In the case of VLPs of Tf1 retrotransposon from *Schizosaccharomyces pombe*, an excess of GAG protein is produced relative to integrase, because of a regulated degradation process (Atwood *et al.*, 1996).

Experimental evidence for the splicing of intron-containing transcripts of plant LTR retrotransposon *Ogre* has been provided recently (Steinbauerová *et al.*, 2008). This article describes a unique arrangement of the *gag-pol* region for *Ogre* elements where the *gag-pro* domains (ORF2) are separated from *rt/rh-int* (ORF3) by a region of about 150–350 bp, which includes several stop codons and is surrounded by GT/AG dinucleotides typical of the 5′ and 3′ termini of most introns (Breathnach *et al.*, 1978; Mount, 1982; Burset *et al.*, 2000). It has been proposed that this region represents an intron that is removed by splicing to reconstitute the full-length *gag-pol* coding region (Neumann *et al.*, 2003).

1.9.3 Polyadenylated & Non-polyadenylated RNA

Almost all eukaryotic mRNA precursors undergo a co-transcriptional cleavage followed by polyadenylation at the 3' end. The life of an mRNA is directed by the protein components of ribonucleoprotein particles (RNPs) whose roles include polyadenylation, transport, translation and degradation. Polyadenylation plays a key role in the life of an mRNA, regulating its transport, translation and turnover. A typical poly(A) signal required for transcriptional termination by RNA Pol II consists of three sequence elements that determine the exact site of the cleavage and polyadenylation. These elements are: hexanucleotide AAUAAA, cleavage/polyadenylation site and a GU- or U-rich region. The actual

cleavage/polyadenylation site is typically located 11–23 nt downstream of the hexamer and 10–30 nt upstream of the GU- or U-rich region (Proudfoot, 1991). For several retroviruses (*e.g.*, HTLV-1, HTLV-2, bovine leukemia virus, RSV, murine leukemia virus), the choice of polyadenylation site is straightforward since the major signal for the reaction (AAUAAA) occurs only once in the transcript. For other retroviruses (*e.g.*, HIV-1, equine infectious anemia, moloney murine leukemia virus), the situation is rendered more complex by the duplication of the polyadenylation signals (AAUAAA and the 3' G/U-rich sequence) at the 5' and 3' ends of the transcript (Cochrane *et al.*, 2006). Thus, controlling where polyadenylation occurs in the retroviral genome is critical for replication.

A number of functional transcripts are known to lack poly(A) tails. These non-polyadenylated transcripts include ribosomal RNAs generated by RNA polymerase I and III, other small RNAs generated by RNA polymerase III, replication-dependent histone mRNAs (Mullen and Marzluff, 2008) and a few recently described long non-coding RNAs (lncRNAs) (Wilusz *et al.*, 2008; Sunwoo *et al.*, 2009) synthesized by RNA polymerase II. Earlier evidence suggested the existence of non-histone polysomal-associated non-polyadenylated RNAs (Milcarek *et al.*, 1974; Salditt-Georgieff, 1981), but these were not characterized in detail. A group of scientists who work on the eukaryotic transcriptome have identified many uncharacterized transcripts and a group of mRNAs lacking a poly(A) tail in H9 and Hela cells (Yang *et al.*, 2011).

Processing of eukaryotic mRNA starts during transcription and is influenced by the RNA polymerase II elongation complex (Zorio and Bentley, 2004; Proudfoot, 2004). Capping, polyadenylation, and splicing have been seen to occur on nascent transcripts *in vitro*, and a variety of *in vivo* and *in vitro* approaches have strongly implicated the carboxyl-terminal domain (CTD) of the large subunit of RNA polymerase II in connecting transcription with these events (Proudfoot *et al.*, 2002; Cramer, 2001; Hirose and Manley, 2000; Shatkin and Manley, 2000). Like retroviruses, *BARE1* LTRs contain typical polymerase II promoters, which include a TATA box and regulatory elements; their gene expression requires transcription by the host RNA polymerase II. For retroviruses, the integrated viral DNA (the provirus) is transcribed by the host RNA polymerase II (pol II) to generate genome-length viral RNA that has a 5′ cap and a 3′ poly(A) tail (McNally, 2008).

1.10 The impact of REs on genome structure, function and evolution

Many experiments have demonstrated that REs may be an important creative force in genome evolution and in the adaptation of an organism to altered environments (Gogvadze and Buzdin, 2009). Table I summarizes the current state of knowledge concerning the ways in which retrotransposon may affect the structural and functional evolution of genes and genomes.

Table 1. The impact of REs on genome structure, function and evolution (modified from (Gogvadze and Buzdin, 2009))

RE function	Examples
Formation of new retrotransposons	Formation of SVA (Shen *et al.*, 1994) LTR-containing retrotransposons (Malik and Eickbush, 2001), and tRNA-derived SINEs (Ohshima *et al.*, 1996)
Recombination events	Recombination between REs may cause various diseases (Burwinkel and Kilimann, 1998; Kamp *et al.*, 2000; Goodier and Kazazian Jr, 2008)
Transduction of 3'-flanking sequences	SVA-mediated transduction duplicated the entire AMAC gene three times in the human genome (Xing *et al.*, 2006)
Formation of processed pseudogenes	Mouse PMSE2b (Zaiss and Kloetzel, 1999) and PHGP pseudogenes (Boschan *et al.*, 2002), TRIMCyp gene of owl monkey (Babushok *et al.*, 2007)
Template switch during reverse transcription	Formation of bipartite and tripartite chimeric elements in eukaryotic genomes (Fudal *et al.*, 2005; Buzdin *et al.*, 2007; Gogvadze *et al.*, 2007)
As promoters	LTRs cause placental-specific expression of CYP19 (van de Lagemaat *et al.*, 2003) and regulate transcription of the NAIP gene (Romanish *et al.*, 2007); LTRs represent the only known promoter for the liver-specific BAAT gene (Carlton *et al.*, 2003)
As transcriptional enhancers	Expression of salivary amylase in humans is a result of HERV-E integration (Meisler and Ting, 1993); ERV9 LTR are enhancer elements in the beta-globin locus control region (Long *et al.*, 1998); Alu sequence is part of enhancer element of human CD8 alpha gene (Hambor *et al.*, 1993)
Providers of novel splice sites	Muscle-specific inclusion of an Alu-derived exon in SEPN1 mRNA in humans (Lev-Maor *et al.*, 2008); generation of alternative VEGFR-3 transcript due to the use of a non-canonical acceptor splice site within LTR sequence (Hughes, 2001)
Sources of new polyadenylation signals	HERV-F LTR may function as an alternative polyadenylation site for gene ZNF195 (Kjellman *et al.*, 1999) HERV-H LTRs are major polyadenylation signals for human HHLA2 and HHLA3 genes (Mager *et al.*, 1999)
Regulate mRNA production	RNAs transcribed from mouse B2 and human Alu SINEs have been found to control mRNA production at multiple levels (Ponicsan *et al.*2010)
Transcriptional silencers	A part of Alu element is a transcriptional silencer of the human BRCA gene (Sharan *et al.*, 1999); endogenous retroviral sequence RTVL-la may serve as silencer of the human Hpr gene (Maeda and Kim, 1990)
Antisense regulators of host gene transcription	Human-specific HERV-K LTRs generate antisense transcripts to SLC4A8 and IFT172 mRNAs (Gogvadze *et al.*, 2009)
Insulator elements	B2 SINE element located in the murine growth hormone locus serves as a boundary to block the influence of repressive chromatin modification (Lunyak *et al.*, 2007) drosophila LTR retrotranspson *gypsy* in the 5' region of the gene yellow blocks the action of the upstream located enhancers and is responsible for the pigmentation of cuticula (Dorsett, 1993)
Regulators of translation	Alu and L1 segments in the 5'UTR of human ZNF177 gene modify gene expression on the protein level by decreasing translation efficiency (Landry *et al.*, 2001)
Play a role in cancer predisposition, development and progression	All three currently actively mobilizing non-LTR retrotransposon families-L1, SVA and Alu-have been identified as the causative agent of several genetic disorders (Konkel and Batzer, 2010)

1.11 Retrotransposons as molecular markers

Retrotransposons are ubiquitous, active, and abundant in plant genomes. Many retrotransposons' features make them appealing as the basis of molecular marker systems. They are usually dispersed throughout the genome and produce large genetic changes at the point of insertion that can be detected by PCR with specific primers (Schulman, 2007). Several molecular marker systems based on retrotransposons have been developed. SSAP (sequence-specific amplified polymorphism) relies on amplification of DNA between a retrotransposon integration site and a restriction site with a ligated adapter (Waugh and Thomas, 1997). IRAP (inter-retrotransposon amplified polymorphism) relies on amplification of DNA between two nearby retrotransposons or LTRs (Kalendar and Schulman, 2006). REMAP (retrotransposon-microsatellite amplified polymorphism) involves amplification of fragments which lie between a retrotransposon insertion site and a microsatellite site and RBIP (retrotransposon-based amplified polymorphism) detects loci either occupied by or empty of a retrotransposon (Agarwal *et al.*, 2008). All methods rely on amplification using a primer corresponding to the retrotransposon and a primer matching a section of the neighboring genome.

Molecular markers play an essential role today in all aspects of plant breeding, ranging from the identification of genes responsible for desired traits to the management of backcrossing programs. They are useful to determine pedigrees and phylogenies and serve as biodiversity indicators (Schulman, 2007). Furthermore, they are used to analyze genome structure and elucidate gene function. For example, *BARE* retrotransposon markers have been used to map a major gene in barley that conditions resistance to the important phytopathogen *Pyrenophora teres* (net blotch; Manninen, 2000). This class of retrotransposon markers has also been used to add knowledge on evolution of *Hordeum spontaneum* in response to climate (Kalendar *et al.*, 2000).

2. Aims of the study

The main objective of the present study was to investigate transcription of LTR retrotransposons. For the LTR retrotransposon family *BARE*, which includes *BARE1* and its parasitic non-autonomous partner *BARE2*, I studied transcript processing and transcriptional regulation. For the non-autonomous *Cassandra* retrotransposon, I investigated the transcript structure and transcriptional features. The specific goals were as follows:

1. To elucidate the promoter activity of the *BARE1* LTR in different barley tissues.

2. To investigate the capping and polyadenylation of *BARE1* transcripts.

3. To study the RNA pools of *BARE1* for translation and /or encapsidation.

4. To study the potential for extra GAG protein formation for *BARE1*.

5. To study the parasitism of the non-autonomous *BARE2* element on *BARE1*.

6. To characterize the features of the transcripts of *Cassandra* elements

3. Materials and methods

3.1 Materials

Hordeum vulgare cv. Himalaya was used for bombardment.

Hordeum vulgare cv. Kymppi line K19 (gift of VTT Biotechnology and Food Research, Espoo, Finland) was used for making callus cell culture.

Hordeum vulgare cv. Bomi was used for DNA and RNA isolation for all other experiments.

3.2 Methods

Table 2. Methods used in this dissertation

METHOD	**I**	**II**	**III**
Plant DNA isolation	x	x	x
Plant RNA isolation	x	x	x
In-vitro transcription	x		
Nuclease protection assay	x		
RACE–PCR	x		
5'RLM–PCR		x	x
3' RLM–PCR	x		x
Particle bombardment	x		
Virus-like particle isolation			x
Polyribosome isolation			x
RT-PCR	x	x	x
LUC & GUS assay	x		
Sequencing & analysis	x	x	x

3.2.1 DNA extraction

DNA was extracted according to the CTAB method for DNA isolation from barley tissue. The detailed protocol is at http://primerdigital.com/dna.html

3.2.2 RNA extraction

For small amounts of RNA (several micrograms), the Qiagen Plant RNeasy kit was used. For large amounts of RNA, we used phenol and chloroform. For the RNA isolation from sucrose gradients, we have used the phenol and chloroform method.

3.2.3 Primer Design

All the primers were designed with the FastPCR program:
http://www.biocenter.helsinki.fi/bi/Programs/download.htm

3.2.4 Particle bombardment

Seeds from cv. Himalaya were germinated aseptically at room temperature; embryos were excised from grains as materials for bombardment. Seeds were sown in 15 cm pots filled with vermiculite and grown in a controlled environment room for collection of roots and shoots. Callus lines were harvested 10-14 days after subculture. Transient transformation was carried out by bombardment. The various LTR-*luc* constructs were mixed with a GUS control plasmid (pBI221) at a ratio of 1:1, and precipitated on to gold particles (~1 μm diameter) prior to bombardment. After bombardment, the tissues were incubated on the sealed agar plate for 24 h at room temperature in darkness.

3.2.5 LUC and GUS assays

Expressed LUC and GUS proteins were extracted for enzymatic measurement with Promega LUC kit and Tropix GUS-light kit respectively according to the manufacturer's instructions. The final results were calculated as: EXO-LUC/LTR-GUS : LTR-LUC/LTR-GUS, where EXO-LUC is the LTR deletion fused to *luc*, LTR-LUC , the full-length LTR fused to *luc*, and LTR-GUS , the full-length LTR fused to *uidA*. The LTR-GUS construct was co-transformed with each LUC construct and worked as transformation control. Visualization of GUS

activity was performed by submerging bombarded tissues in a stain buffer from the kit. The blue stain was then visualized by light microscopy and the photos were taken.

3.2.6 *In vitro* transcription

Probes for ribonuclease protection assays for identifying *BARE1* transcripts were synthesized by *in vitro* transcription using T7 RNA polymerase on templates of synthetic DNA containing the T7 promoter. Several partially single-stranded templates were prepared. The *in vitro* transcription reaction was carried out with the T7 MEGAshortscript™ kit (Ambion) according to the manufacturer's protocol. Fluorescent probes were made by incorporating labeled Fluorescein-12-UTP and unlabeled UTP in 1:1 ratio into the transcription reaction. The internal control RNA probe P18S was generated by *in-vitro* transcription with T3 polymerase from the pTRI RNA 18S (Ambion). The probes produced were gel purified and quantified.

3.2.7 RACE-PCR

The reverse transcription reaction for production of cDNA was carried out using the SMART RACE cDNA Amplification Kit according to the manufacturer's instructions. The PCR product was first purified with the Qiagen PCR purification kit and cloned into the pGEM-T vector (Promega). Colonies containing inserts were screened by PCR, using the gene-specific primer which binds to both TATA1 and TATA2 transcripts together with the universal primer mix from the kit (smart™ RACE cDNA amplification kit). Positive colonies were sequenced by an in-house service (Fig. 10).

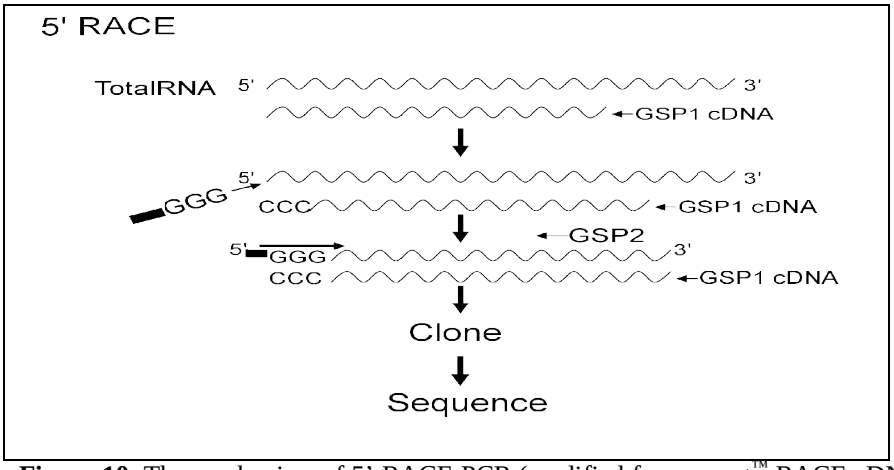


Figure 10. The mechanism of 5' RACE PCR (modified from smart™ RACE cDNA amplification kit protocol). GSP represents gene specific primer.

3.2.8 Nuclease protection assay

Ribonuclease protection assays (RPAs) were made with the RPA III kit (Ambion) according to the manufacturer's instructions. The precipitated, protected fragment generated by the assay was dissolved in 10 µl of loading buffer before loading onto a 5% acrylamide sequencing gel in a Pharmacia LKB ALF DNA sequencer.

3.2.9 Virus-like particle isolation

Barley cv. Kymppi callus (frozen by liquid nitrogen) was homogenized into powder, and 4.8 g powder was added into VLP extraction buffer (150mM KCl, 10mM HEPES-KOH pH 7.2, 10mM EDTA, 5mM, $MgCl_2$ 3mM DTT, 0.5% Triton X-100, 1% protease inhibitor cocktail (Promega). The extracts were transferred into 1.5 ml Eppendorf tubes and centrifuged at 1000x*g* for 10 min at 4°C in a tabletop centrifuge, followed by 12,000x*g* and 18,000x*g* centrifugation under the same conditions. The supernatant from the extraction was filtered through a 0.2 µm membrane, the filtered supernatant was placed into ultracentrifuge tubes that contained 20% sucrose and then the VLPs were purified from the supernatant by ultracentrifugation at 35,000 rpm (Sorvall TH 641 rotor) for three hours at 4°C. Nine fractions were taken and pellet was saved. Protein from each fraction was precipitated by TCA followed by Western blotting. The fraction which contained VLP was identified by the existence of mature GAG protein (32 kDa).

3.2.10 Polyribosome isolation

Barley cv. Kymppi callus was cultured in Medium 108 as described by Salmenkallio-Marttila et al. (1995). The cells were collected, frozen under liquid N_2, and then pulverized. Approximately 3.2 g powder was thawed in polysome extraction buffer (0.2 M sucrose, 0.2 M Tris-HCl pH 8.5, 0.4 M KCl, 35 mM $MgCl_2$, 25 mM EGTA, 10 mM DTT) and the mixture was gently homogenized. The mixture was centrifuged at 2,000 g for 5 min at 4 °C. The supernatant was adjusted to 1 % (v/v) for Triton X-100 and centrifuged at 20,000 g for 20 min at 4 °C. The supernatant was collected and supplemented with 400 mM KCl, then incubated for 10 min at room temperature, after that, the supernatant was layered onto 10%-50% sucrose gradients prepared in polysome buffer and centrifuged at 4°C for 4 h at 36,000 rpm in a Sorvall TH-641 rotor. Then 1ml fractions were collected and the absorbance of the gradient at 260 nm was monitored. The total RNA was isolated from each fraction and an RNA gel was run to identify those fractions which contained polyribosome.

3.2.11 5’ RLM-RACE

5’ RLM-RACE was used to investigate the capping state of *BARE1* transcripts. The total RNA or polyadenylated RNA was directly ligated to an RNA linker to detect the uncapped RNA population. The ligation of Calf Intestine Alkaline Phosphatase (CIP) and Tobacco Acid Pyrophosphatase (TAP) -treated RNA enables detection of the capped population by nested PCR. The idea of 5’ RLM-RACE is presented in Fig.11.

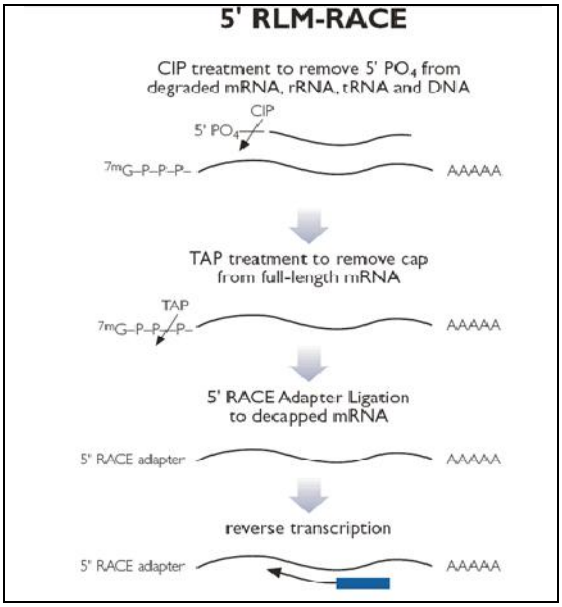


Figure 11. The mechanism of 5’ RLM-RACE (RLM-RACE Procedure from Ambion)

3.2.12 3’ RLM-RACE

3’ RLM RACE was used to investigate the 3’ end of *BARE1* transcripts and also to examine the 3’ ends of *Cassandra* elements. The method is summarized in Fig. 12. The RNA was ligated with a phosphorylated oligo first, then the cDNA was synthesized with a primer which binds to the oligo. Nested PCR was then carried out with a gene-specific primer together with the cDNA primer, and the PCR product cloned and sequenced.

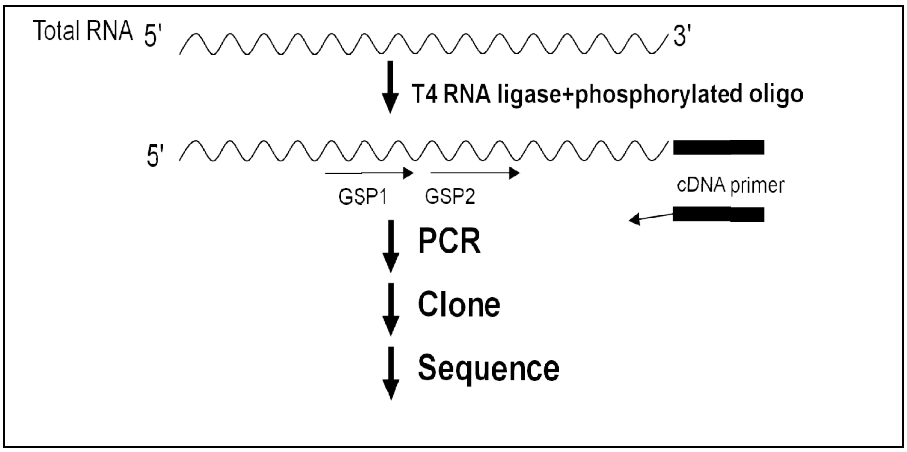


Figure 12. The mechanism of 3’ RLM-RACE. GSP represents a gene specific primer.

4. Results and discussion

We have carried out a study of the transcription of *BARE* element within different barley tissues and analyzed the promoter activity of the *BARE* LTR. The two promoters of the LTR vary independently in activity by tissue. Regarding translation of a reporter gene, promoter element TATA1 is almost inactive in embryos, whereas transcription in callus appears to be less tightly regulated than in other tissues. Deletion analyses of the LTR identified strong positive and negative regulatory elements. The promoters produce multiple groups of transcripts, and these transcripts are distinct by their start and stop points, by their sequences, and by whether a poly(A) tail is added to the end of transcript. Some of these groups do not share the common end structures needed for template switching, which is critical for their replication and life cycle. Only about 15% of *BARE* transcripts are polyadenylated.

Many viruses have uncapped genomic RNA, for example, tobacco necrosis virus RNA (Shen and Miller, 2004) and barley yellow dwarf virus (Allen *et al.*, 1999). The close relationship between retrotransposons and retroviruses attracted our interest to investigate the capping situation of *BARE1* transcripts. Surprisingly, we found that only TATA2 transcripts are capped, and transcripts starting from TATA1 are uncapped. Furthermore, the experimental results showed that capped transcripts are polyadenylated, and also that capping and polyadenylation are important steps for mRNA maturation and translation. The bombardment experiment with LTR constructs gave strong evidence that this population of transcripts is used for making polyprotein, and that, in contrast, constructs containing TATA1 do not give any LUC activity, which means that this population is not available for translation. The evidence that capped and polyadenylated mRNA cannot provide an R domain, which is a critically needed in cDNA synthesis, and that the non-polyadenylated transcripts can provide different lengths of R regions demonstrate that *BARE* transcripts can be sorted into two pools, capped ones for translation and uncapped transcripts for encapsidation. The existence of two nonequilibrating pools has been demonstrated for moloney murine leukemia virus before (Levin *et al.*, 1974; Levin and Rosenak, 1976).

BARE1 has only one open reading frame like many other superfamily *Copia* elements. Unlike superfamily *Gypsy* elements and retroviruses which create extra GAG by frameshifting, we found *BARE1* has a novel way to produce extra GAG: by splicing. Partial transcripts of *BARE1* are spliced to produce a subgenomic RNA encoding only GAG, the

capsid protein that forms the VLPs. The *BARE2* retrotransposon, which lacks the capacity to produce its own GAG, is not spliced. *BARE2* is packaged into the VLPs formed by the GAG protein of *BARE1*. Furthermore we found capping, splicing, and polyadenylation are connected, and that capped, spliced, polyadenylated mRNA is polyribosome related. Taken together, we believe that we have uncovered an important feature of *BARE* transcription, and we will investigate RNA and cDNA intermediates in virus-like particles to uncover the details of *BARE* replication.

4.1 The *BARE1* LTR functions as a promoter and some *BARE1* elements are transcriptionally active

The LTRs of retrotransposon *BARE1* have promoter activity. This has been established by histochemical assays of GUS, which is translated using an LTR-*gus* construct in various tissues (I), and protoplast transformation (Suoniemi *et al.*, 1997). Most retrotransposons are thought to be transcriptionally inactive (Kumar and Bennetzen, 1999) or transcriptionally silent in somatic tissues, but active during certain stages of plant development or stress conditions (Grandbastien, 1998). *BARE1* was shown to be transcriptionally active in leaves (Suoniemi *et al.*, 1997). In our work, we wanted to find out if *BARE1* is active in other barley tissues as well, and about promoter choice in various tissues.

4.1.1 *BARE1* is active in all tested barley tissues

The *BARE* retrotransposon is unusual in containing two promoters, TATA1 and TATA2. The question of how both promoters may be under selection for maintenance prompted us to investigate the relative activity of TATA1 and TATA2 in leaves, roots, shoots, embryos, and callus by the RNase protection assay (Melton, 1984) to evaluate expression level. Our results showed *BARE1* is active in all tested tissues, and that the transcription level is almost the same for the transcripts that start from TATA1. For TATA2 transcripts, however, embryos give the strongest signal, and leaves give a weaker signal (I).

4.1.2 Determination of the start site for *BARE* transcripts

Whole-genome transcriptional analyses indicate that putative alternative transcriptional start sites are not uncommon in Arabidopsis (Alexandrov *et al.*, 2006). Many retrotransposons also have this feature, for example, yeast retrotransposon Ty4 (Hug and Feldmann, 1996), the *D.*

melanogaster TART family of telomeric retrotransposons (Maxwell *et al.*, 2006), and the *D. melanogaster mdg1* element, which displays multiple start sites for the antisense direction (Arkhipova and Ilyin, 1991). We have examined the 5′ ends of *BARE* transcripts from callus and shoots in more detail by 5′ rapid amplification of cDNA ends (RACE). For transcripts starting after TATA1, we were also able to separate the results for *BARE1* and *BARE2*. Shoots yielded clones for transcripts from both *BARE* subfamilies and TATA boxes. Callus cells contained both TATA1-driven *BARE2* transcripts and TATA2-driven *BARE* transcripts. Multiple, closely spaced 5′ ends were obtained, with four different 5′ ends for TATA1 products in 13 sequenced RACE clones and five different 5′ ends for TATA2 products among 24 RACE clones. Our earlier results (Suoniemi, 1996b) indicated a start for TATA1 at nt 1323 and for TATA2 at nt 1670, both for callus. Our current results (I) differ from these by 25–60 nt for TATA1 and from 5 to 19 nt for TATA2, which is not surprising: The earlier positions were estimated from the relative mobility of RNA fragments in sequencing ladders, whereas the current positions have been determined directly by sequencing.

4.1.3 Control of *BARE* transcription in barley tissues

In order to identify functional regulatory regions of *BARE1* LTR in plant cells, transient gene expression experiments were undertaken in leaf, embryo, and callus by particle bombardment (I). We concentrated on leaves as they are better bombardment targets for this purpose. Full-length and deleted LTR constructs, which contain a putative promoter, *luc* gene and *nos* terminator, were bombarded together with the construct LTR-*gus*-*nos* as a transformation control. Some constructs contain both TATA1 and TATA2, whereas some others contain only one of them. The expression of the *luc* gene driven by these constructs, and of the *gus* gene driven by the LTR was analyzed by histochemical and fluorimetric methods. Construct H and T0, which contain neither TATA1 nor TATA2, did not give any LUC activity, which shows that there is no other promoter. Previous assays of the promoter activity of the *BARE1* LTR were carried out only in leaf protoplasts (Suoniemi, 1996b), which may not accurately reflect expression in organized tissues. When we compared the LUC activity driven by deleted LTR constructs and by the full length LTR, we identified both positive and negative regulatory elements, inferred to be so because of their corresponding effect on reporter gene expression. (Fig. 13). The positive elements locate in 308-963 nt, the negative regulatory elements locate at 963-1444 nt and strong positive regulatory elements locate at 1444-1611 nt respectively. Callus tissue is a useful experimental system for studying the *BARE1* life cycle

due to the abundance of *BARE1* translation products in it (as well as in embryos) in comparison with their dearth in leaves (Jääskeläinen *et al.*, 1999; Vicient *et al.*, 2001b). For transcriptional comparisons between embryos and callus, we chose constructs to test the relative strengths of TATA1 and TATA2 and the result showed that both TATA1 and TATA2 were active in callus. TATA2 showed more activity in embryos than in callus. Furthermore, we found an interesting phenomenon: the construct that contained only TATA2 gave relatively strong *luc* expression, especially in callus tissue, which is comparable to full length LTR activity. On the other hand, constructs that contain only TATA1 did not give LUC activity. As we now know, TATA1 functions in transcription, but the transcripts that start from TATA1 may not be used as templates for translation.

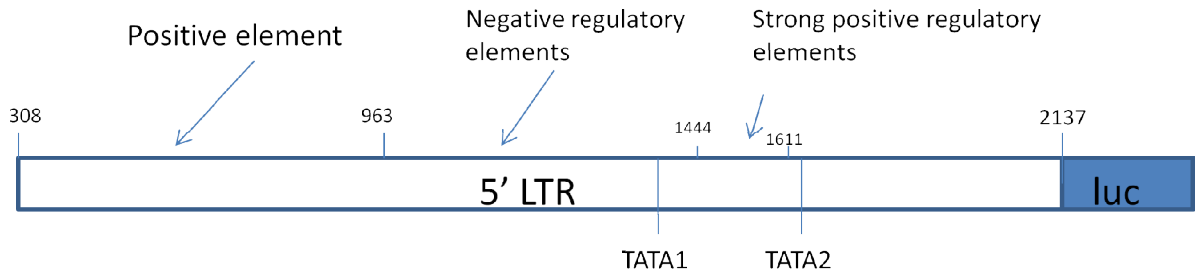


Figure 13. The location of regulatory element in *BARE1* 5' LTR.

4.2 *BARE1* transcript termination

BARE1 mRNA falls into two populations, one containing polyadenylated tails, the other lacking them, and the RNA also shows sequence heterogeneity within each group. Based on our experimental results (I), we observed that if polyadenylated RNA starts after TATA2, no R domain could be identified. If it starts after TATA1, a few different sizes for the R region can be identified. The R region is very important for the start of *BARE1* transcription. The experiment to map the 5' mRNA ends of polyadenylated *BARE1* transcripts by the direct ligation method did not give PCR products, which means that polyadenylated *BARE1* RNA has no free 5' phosphate. Hence, it appeared unlikely (I) that an R region is created from this population.

4.2.1 Polyadenylated transcripts

Polyadenylation is an important step during messenger RNA maturation that determines the RNA stability and translatability, among other characteristics. It consists of the addition of a poly(A) tract to the 3' end of the cleaved pre-mRNA and requires a number of *cis*-elements that are recognized by the cleavage and polyadenylation machinery. Suboptimal *cis*-elements have been shown to explain the inefficient polyadenylation and frequent transcriptional read-through of the Rous sarcoma virus, which is a retrovirus structurally related to LTR retrotransposons (Maciolek and McNally, 2008). We analyzed the mRNA 3' ends of *BARE1* by RT-PCR, using a modified oligo(dT) as the cDNA primer and carrying out PCR using two nested *BARE*-specific primers, positioned upstream of TATA1, paired with a cDNA primer lacking the dT segment. The *BARE*- specific primer guarantees that the amplification is from the U3 region of 3' LTR. We detected cDNA products, which comprised at least two populations (I): the major population stops 23nt after the TATA1 in the 3' LTR and has the contiguous structure (TATATA)(TATAA)(18nt)Poly(A), with the first motif being TATA1 and the second the likely polyadenylation signal; the other is a minor population, which stops at a few nucleotides downstream of a predicted polyadenylation signal in *BARE1*a (Z17327), with the sequence ATAA located at 1484 of Z17327 being proposed as the polyadenylation signal. Neither of these two populations will define an R domain from transcripts starting from TATA2, and this puzzle led us to investigate the existence of non-polyadenylated transcripts. The *D. melanogaster* TART family also has multiple polyadenylation ends (Maxwell *et al.*, 2006; Hernández-Pinzón *et al.*, 2009).

4.2.2 Non-polyadenylated transcripts

We performed 3' RLM-RACE PCR by first adding a linker to the end of the total transcript preparation and then, after purification, synthesizing cDNA by reverse transcription. The *BARE1* transcripts were screened by nested PCR using nested *BARE1*-specific forward primers and a primer which binds to the ligated linker. The PCR product showed several bands on gels, which represent transcripts ending after each TATA box when mapped on the *BARE1a* (Z17327). We also purified these PCR products and cloned them, then checked the clones using *BARE*-specific primer together with a primer specific for products with a poly(A) end. This approach showed that 15% of the *BARE* transcripts are polyadenylated. This result suggests that *BARE1* presents a weak transcriptional terminator that give rise to a population of transcripts polyadenylated at different positions, with a majority of transcripts

being non-polyadenylated. In the case of retroviruses, a weak polyadenylation signal allows some transcripts to escape premature polyadenylation at the 5' LTR (Furger *et al.*, 2001). Nevertheless, the presence of a weak polyadenylation site could help retrotransposons to minimize their deleterious effect in genomes and contribute to their maintenance in evolution (Hernández-Pinzón *et al.*, 2009).

4.3 Capping of *BARE* transcripts

Many viruses contain uncapped genomic RNA and use an internal ribosomal entry mechanism that promotes translation; for example, both hepatitis A virus (Brown *et al.*, 1991) and yeast retrotransposon Ty1 produce both capped and uncapped RNAs encapsidated into particles (Cheng and Menees, 2004). We set up an experiment to investigate the capping situation of *BARE1* transcripts to uncover the mechanism of translation and VLP formation. We made two parallel experiments: in one experiment, the DNase-treated total RNA was dephosphorylated, followed by decapping. The dephosphorylation treatment eliminated the phosphate on possible contaminating trace amounts of genomic DNA and on the uncapped RNA population. The decapping treatment exposed the free phosphate behind the G cap for ligation, allowing an RNA linker to be ligated to the 5' end. Nested PCR was then carried out with nested *BARE1*-specific primer pairs, each pair having a primer which binds to the ligated linker. Both TATA boxes were known to be in use, as demonstrated by the RACE – PCR experiment. Surprisingly, only transcripts that started after TATA2 gave bands in the various tissues. From the RACE experiment, we know both TATA boxes are used, so it was likely that transcripts starting after TATA1 were uncapped; in order to check this, another experiment was done in parallel in which the same amount of total RNA was directly ligated with the RNA linker, so the uncapped *BARE* transcripts should have a linker sequence at the 5' end, suitable for the same nested RT-PCR. The RT-PCR result from the RNA directly ligated to a linker identified transcripts starting after TATA1. No transcripts starting from TATA2 were identified; this corresponds to the result from decapped RNA. Hence, TATA2 transcripts are capped. We did the same decapping experiment for polyadenylated RNA, and found that polyadenylated *BARE1* transcripts are also capped and start after TATA2 (III).

4.4 *BARE1* transcripts are partially spliced

Alternative splicing (AS) creates multiple mRNA transcripts from a single gene and AS is known to contribute to gene regulation and proteome diversity in animals. AS in plants is not rare, as evidence suggests that AS participates in important plant functions, including stress response, and may impact domestication and trait selection (Barbazuk *et al.*, 2008). Retroelements (including retrotransposons and retroviruses) employ a variety of translational recoding mechanisms to express GAG and POL. In contrast to retroviruses, nearly half of the retrotransposons identified encode GAG and POL in a single ORF. *BARE1*, like many other *copia* elements, contains only one reading frame. For these elements, the required ratio of GAG to POL may be achieved post-translationally through preferential POL degradation, as has been observed for the Tf1 and Ty5 yeast retrotransposons (Levin, 1993; Atwood *et al.*, 1996; Irwin and Voytas, 2001). It is also possible that a post-transcriptional mechanism, such as alternative splicing, is utilized to express an excess of GAG, which is a strategy employed by the *Drosophila copia* element (Brierley and Flavell, 1990). In view of this, we carried out RT-PCR using a forward primer located before *gag* and a reverse primer located in the middle of *ap*. We found a faint band besides the major amplification band; the faint band was not amplified in genomic DNA controls. We cloned and sequenced both bands, and found that the faint short band represents sequences that have 104 nt deletions after the *gag* coding sequence. We searched the sequences that came from the cloned major band for splicing sites using the splicing site prediction database (http://www.cbs.dtu.dk/services/NetGene2/) and found both a splicing donor and an acceptor around the splicing border sequence. Thus, we found that *BARE1* transcripts are partially spliced and the spliced sequence contains typical GT/AG at both ends (III). The T nucleotides occupy 37.5% of this region. A high T proportion in intron sequences has been reported particularly for plants. In plants, moreover, branch site selection is relaxed and a polypyrimidine tract is not necessary (Goodall and Filipowicz, 1989). After splicing, the reading frame of the *BARE pol* gene is changed, and many stop codons are created after the 3' splicing junction resulting from the 104 nt sequence being spliced out. The 5' splicing junction is located just before the last two nucleotides of the predicted full-length *gag* sequence. Hence splicing may be a way for *BARE1* to produce more GAG protein for VLP formation given that *BARE1* is a retrotransposon that has only single reading frame. The position of the first stop codon after splicing is just a few amino acids away from the predicted GAG carboxyl terminus. If splicing is to make more GAG, then spliced transcripts

should be polyribosome-associated. The RT-PCR using total RNA isolated from polyribosomal fractions, gave the same spliced transcript band.

4.5 The features of non-autonomous element *BARE2*

BARE2 sequences are conserved with *BARE1* in the PBS, Psi and DIS domains. The Psi and DIS have been shown to be the motifs that are important for RNA dimerization and packaging in HIV-1 and HIV-2 (Ooms *et al.*, 2004; L'Hernault *et al.*, 2007). Our group's previous work has concluded that *BARE2* is probably a partial parasite of the *BARE1* element, because the machinery of the latter can complement the defective *gag* of the former (Tanskanen *et al.*, 2007). Our results showed that *BARE2* is indeed packaged into VLPs formed by GAG that is a product of *BARE1* translation. In addition, we found that *BARE2* is also capped though this is not a surprise, because in the conserved sequence between the *BARE1* and *BARE2* LTRs, the same protein complexes and the same capping mechanism could be used. *BARE2* transcripts are also polyribosome- associated, which means that *ap, in,* and *rt-rh* contained in the *pol* gene of *BARE2* are also expressed. This is consistent with the computer prediction: an ATG codon upstream of the AP domain may be used as the start codon for *BARE2* protein synthesis. *BARE2* and *BARE1* were expressed equally in callus tissue although *BARE2* has a higher genome copy number compared to *BARE1*; no spliced sequences of *BARE2* have been found (III). When *BARE2* RNA sequences were aligned with *BARE1* RNA sequences over the splicing region of *BARE1*, nucleotides around the splicing donor and acceptor of *BARE1* are very much different. No splicing signals were found when searching for a splicing signal in *BARE2* using the same splicing site prediction database as for *BARE1*. This is logical because the proposed purpose of splicing is to make more GAG protein for VLP formation. *BARE2* does not express GAG, and our experiment gave evidence that *BARE2* is packaged into VLPs. Hence, *BARE1* and *BARE2* are life cycle partners.

4.6 *BARE1* transcripts are sorted: one pool for translation and another pool for encapsidation

LTR retrotransposons and vertebrate retroviruses all share a transposition mechanism, which involves transcription of the integrated genomic DNA copies into RNA that contains all of the genetic information. Two functions have been attributed to this RNA. One is to be copied

by reverse transcription into extrachromosomal DNA, which becomes inserted into new chromosomal locations. The second is to be the template for the protein components encoded by both the *gag* gene and the *pol* gene. Thus, the full-length RNA serves as both mRNA and genomic RNA (Meignin *et al.*, 2003). For *BARE1*, we set the goal to find out if there is a single pool of full-length RNA within the cell that is translated and then encapsidated, or alternatively if there are two independent pools of these RNAs, with one of them being the template for translation and the second being the template for encapsidation. Our results show that *BARE1* transcripts are fall in two populations: one for translation and another pool for encapsidation (III). Here is the evidence which supports this view: i) only TATA1 transcripts contain the R domain, which is necessary for reverse transcription (I); ii) the promoter analyses of the *BARE* LTR, using a reporter gene linked to LTR deletion constructs, demonstrated that TATA2 is sufficient to give full promoter activity and that TATA1 alone gave no reporter expression in callus or protoplasts, with only negligible expression in embryos (I; Suoniemi, 1996b); iii) although the existence of internal ribosomal entry sites in some genes has been observed, the majority of eukaryotic cells 'scan' the capped transcripts for the start of translation; our results show that polyadenylated transcripts are also capped, spliced, and polyribosome associated. The polyadenylation is important for RNA stability and translation. Both capped and uncapped RNA were found in yeast retrotransposon Ty1 (Cheng and Menees, 2004); P-body components are required for Ty1 retrotransposition during assembly of VLP (Checkley *et al.*, 2010). Although the TATA1 transcripts were not detected with caps, they could be initially capped and then very efficiently decapped as is Ty1 (Dutko *et al.*2010). The TATA2 products, in contrast, are never seen decapped. Thus, TATA1 and TATA2 RNAs appear to follow very different pathways regarding RNA processing.

4.7 The 3' ends of *Cassandra* transcripts are polyadenylated

Uncapped, polyadenylated *Cassandra* transcripts starting from the polymerase III promoter have been detected, which means *Cassandra* 5S domains are transcriptionally functional. *Cassandra* specifically produces the LTR-to-LTR transcripts typical of REs at least in barley. Transcripts initiating from the internal RNA polymerase III promoter begin in the 5S domain of the 5' LTR and terminate in the 3' LTR at a canonical pol III terminator that is present in *Cassandra* but absent from within cellular 5S genes. An 18 nt R region, which is needed for

reverse transcription, is formed from the 5’ end of the 5S region. *Cassandra* transcripts terminate at the beginning of the 5S RNA sequence of 3’ end LTR, where the termination signal (TTTT) is located, with poly(A) tail; they are polymerase III transcripts (II). The full length of the transcripts is about 480bp. Capped, read-through transcripts containing *Cassandra* sequences can also be found in RNA and in EST databases. The predicted *Cassandra* RNA 5S secondary structures resemble those for cellular 5S rRNA, with a high information content especially in the pol III promoter region. *Cassandra* retrotransposons are also abundant and insertionally polymorphic (II).

5. General conclusions and future prospects

Retrotransposons represent a large fraction of the repetitive DNA of most eukaryotes. The LTR retrotransposons encode the protein components required for their movement in the genome. In the Triticeae, *BARE* elements or their relatives currently appear to be highly active, composing up to 10% of the genome of barley, in which transcription and line-specific insertion site polymorphisms are readily detected. The autonomous elements have mechanisms to copy or move themselves and thus alter the genome. The mechanisms may be under tight control by host factors, environmental factors or the element itself. Many retroelement copies appear to be transcriptionally silent, partially deleted, or contain stop-codons or frameshifts. Many of these elements can still be active in the genome following homologous recombination or be transposed using essential proteins supplied by other elements. Hence, the concept of autonomy, non-autonomy, and parasitism are the key to understand retrotransposon dynamics in the plants. Intra-element recombination plays a major role in controlling genome expansion resulting from *BARE1* integration through removal of all but a single LTR (Vicient *et al.*, 1999b; Shirasu *et al.*, 2000).

As evolutionary opportunists, retrotransposons' success depends on their interface with basic cellular processes including stress response and signal transduction, as well as the cell cycle. Hence, our basic goal is to understand the steps of retrotransposon replication, their regulation by the element itself and by the cell, and the impact of retrotransposons on the genome. To realize this goal, the particle bombardment method has been used to investigate LTR activity, and furthermore to define the enhancer region. Transcription start sites were first investigated by RACE–PCR and further established by RLM–RACE PCR. The sequencing result shows that only the TATA2 transcripts are capped. Surprisingly, only 15% of total transcripts are polyadenylated. It is likely that there are two *BARE* transcript pools, one that is uncapped and serves as the template for cDNA synthesis and another that is capped and polyadenylated, which is used for translation. *BARE1* transcripts are partially spliced, and spliced form may be the source of extra GAG production because they are translated (polyribosome associated).

The *BARE1* has very high number of genome copies but fairly low expression. One reason for the low expression is that many LTRs of *BARE1* are methylated. Retrotransposon LTRs generally are prime targets for DRD1/pol IVb-mediated cytosine methylation (Matzke *et al.*, 2006). Furthermore, retrotransposon transcripts may be subject to post-transcriptional

silencing. In our group's previous work, the expression of *BARE1* from the methylated LTR constructs was much lower than that from unmethylated LTR constructs (Suoniemi, 1997). For methylation research, we plan to demethylate the target genome first by using chemical 5-dAZAC and then use RT-PCR to compare if the transcription level was raised after treatment. Another way to investigate the relationship between the methylation pattern change and expression level is to use restriction enzymes, which specifically cut methylated cytosine followed by bisulfite sequencing.

Recently, many TE-related micro RNAs has been reported (Yuan *et al*., 2011). The siRNAs derived from transgenes and endogenous REs in plants have also been found (Hamilton *et al.*, 2002). RNA interference has an important role in defending cells against parasitic genes, such as retrotransposons. It would be interesting to find endogenous microRNAs from *BARE1* and furthermore to investigate if RNAi is one of the regulatory pathways in the transcription of the *BARE1* element. *BARE1* transcription may be regulated by RNAi because *BARE1* exists in the genome in both orientations; double stranded RNA which is a target of enzyme Dicer will be created if transcripts are produced from both orientations. On the another hand, because synthetic dsRNA introduced into cells can induce post-transcriptional gene silencing, we can make double–stranded RNA *in vitro* and then inject it into barley tissue by particle bombardment to investigate if double–stranded RNA induce the RNA interference pathway. The transformed barley (barley transformed with LTR-GFP or LTR deletion-GFP constructs by using Agrobacterium), which has been produced by us in another project, can be used as target system.

The *BARE1* elements and near relatives are not restricted to barley and other members of genus *Hordeum,* but appear to be seen in several other species of the tribe Triticeae and in oats (*Avena sativa*) and rice (*Oryza sativa*) as well (Vicient *et al.*, 2001a). It remains to be established at which point in the evolution of the Gramineae this element emerged as an active class. And another issue is where *BARE* elements play a role in meiotic chromosomal pairing and in speciation.

Cassandra elements contain a polymerase III promoter, and the usage of polymerase III has been established (II), but there appears to be the possibility that polymerase II is also used. The C Box motif is responsible for the transcription starts of pol III. We are going to make C box deletion by site-specific mutagenesis for the investigation of promoter usage by the *Cassandra* element, then transfect the mutated LTR-*luc* construct to mammalian cells to

verify the usage of polymerase III by comparing the signal to the signal produced from the non-mutated LTR-*luc* construct. And in the same time, we will investigate if the *Cassandra* 5S RNA is a component of large ribosomal unit as cellular 5S RNAs are. There remain many unanswered questions critical for understanding the role of retrotransposons in genome dynamics and cellular gene expression.

6. Acknowledgements

This work has been carried out at the MTT/BI Plant Genomics Group, a joint laboratory of MTT and the Institute of Biotechnology, University of Helsinki. Research funding was provided by the Academy of Finland and the travel grant to scientific meeting has been provided by the Finnish Graduate School in Plant Biology and Chancellor's Travel Grant.

I want to thank my supervisor Professor Alan H. Schulman for giving me the opportunity to work in his lab. His scientific guidance throughout the period of my Ph. D study is greatly acknowledged. I am especially grateful for the freedom he has allowed and supported me to develop my own ideas. He treats each group member as a co-worker, I am very grateful for that.

I wish to express my sincerest gratitude to my co-workers, Dr. Ruslan Kalendar , Dr. Cédric Moisy, Marko Jääskeläinen, Jaakko Tanskanen, Ursula Lönnqvist, Anne-Mari Narvanto, Elitsur Yaniv and my previous co-workers Dr. Carlos Vicient and Dr. François Sabot; their great knowledge in molecular biology, as well as their excellent comments have been invaluable.

Special thanks to Dr. Ruslan Kalender, his sense of humor often makes me laugh, it helped me forgot about failed experiments. Special thanks to Marko Jääskeläinen, he always has time when I need help on slides, figures, and also any questions on protein work. Special thanks to Anne-Mari Narvanto, for her help with my work at the end of my Ph.D study. Special thanks to Dr Lei Wang and Wei Wang for their support on organizing my dissertation defense.

I would like to thank the reviewers Dr. Kristiina Mäkinen and Dr. Carlos Vicient for their encouragement, valuable advice, and great patience in reading and revising my thesis manuscript and special thanks to Dr. Krishnan Narayanan who did the proofreading of my thesis. I also would like to thank Marko Jääskeläinen and Pirkko-Liisa Schulman for translating and reviewing the popular abstract of my dissertation.

I would like to thank my followup group members: Dr. Kristiina Mäkinen and Dr. Mikko Frilander for their valuable advice on my work yearly, especially Dr. Mikko Frilander for his support on *BARE1* splicing and *Cassandra* mutations.

Finally, I want to give my distinctive thanks to my husband Li Song Ping and our children Li Kaiyu , Li Kaiwen. The support I got from my husband was not only in everyday life but also on the scientific level; he takes my questions on work as his own and always try to find answers for me. Special thanks to my lovely children, their love and my love to them has made me feel every day to be full of sunshine; with the energy they gave to me, I finished my thesis happily.

Helsinki, December. 2011

7. References

Agarwal, M., Shrivastava, N. and Padh, H. (2008) Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Rep.* 27, 617-631.

Alexandrov, N.N., Troukhan, M.E., Brover, V., Tatarinova, T., Flavell, R.B. and Feldmann, K.A. (2006) Features of Arabidopsis genes and genome discovered using full-length cDNAs. *Plant Mol Biol* 60, 69-85.

Allen, E., Wang, S. and Miller, W.A. (1999) Barley yellow dwarf virus RNA requires a cap-independent translation sequence because it lacks a 5' cap. *Virology* 253, 139-144.

Andrake, M.D. and Skalka, A.M. (1996) Retroviral integrase, putting the pieces together. *J Biol Chem.* 271, 19633-19636.

Arkhipova, R. and Ilyin, Y.V. (1991) Properties of promoter regions of mdg1 *Drosophila* retrotransposons indicate that it belongs to a specific class of promoters. *EMBO J.* 10, 1169-1177.

Atwood, A., Lin, J.H. and Levin, H.L. (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.

Babushok, D.V., Ostertag, E.M. and Kazazian, H.H.J. (2007) Current topics in genome evolution: molecular mechanisms of new gene formation. *Cell Mol Life Sci.* 64, 542–554

Barbazuk, W.B., Fu, Y. and McGinnis, K.M. (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res.* 18, 1381-1392.

Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.M., Westerman, R.P., Sanmiguel, P.J. and Bennetzen, J.L. (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet.* 5: e1000732.

Belancio, V.P., Hedges, D.J. and Deininger, P. (2006) LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res.* 34, 1512–1521.

Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem.* 72, 291-336.

Boschan, C., Borchert, A., Ufer, C., Thiele, B.J. and Kuhn, H. (2002) Discovery of a functional retrotransposon of the murine phospholipid hydroperoxide glutathione peroxidase: chromosomal localization and tissue-specific expression pattern. *Genomics* 79, 387–394

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978) Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proc Natl Acad Sci U S A.* 75, 4853-7

Brierley, C. and Flavell, A.J. (1990) The retrotransposon copia controls the relative levels of its gene products post-transcriptionally by differential expression from its two major mRNAs. *Nucleic Acids Res.* 18, 2947–2951.

Brown, E.A., Day, S.P., Jansen, R.W. and Lemon, S.M. (1991) The 5′ nontranslated region of hepatitis A virus: secondary structure and elements required for in vitro translation. *J Virol*. 65, 5828–5838.

Burset, M., Seledtsov, I.A. and Solovyev, V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* 28, 4364-4375.

Burwinkel, B. and Kilimann, M.W. (1998) Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J Mol Biol.* 277, 513-517.

Buzdin, A., Gogvadze, E. and Lebrun, M.H. (2007) Chimeric retrogenes suggest a role for the nucleolus in LINE amplification. *FEBS Lett.* 581, 2877-2882.

Cameron, J.R., Loh, E.Y. and Davis, R.W. (1979) Evidence for transposition of dispersed repetitive DNA families in yeast. *Cell* 16, 739-751.

Carlton, V.E., Harris, B.Z., Puffenberger, E.G., Batta, A.K., Knisely, A.S., Robinson, D.L., Strauss, K.A., Shneider, B.L., Lim, W.A., Salen, G., Morton, D.H. and Bull, L.N. (2003) Complex inheritance of familial hypercholanemia with associated mutations in TJP2 and BAAT. *Nat Genet.* 34, 91-96.

Casacuberta, J.M. and Santiago, N. (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene* 311, 1-11.

Checkley, M.A., Nagashima, K., Lockett, S.J., Nyswaner, K.M. and Garfinkel, D.J. (2010) P-Body Components Are Required for Ty1 Retrotransposition during Assembly of Retrotransposition-Competent Virus-Like Particles. *Mol Cell Biol.* 30, 382-398.

Cheng, Z. and Menees, T.M. (2004) RNA Branching and Debranching in the Yeast Retrovirus-like Element Ty1. *Science* 303, 240-243

Clever, J.L., Miranda, D.J. and Parslow, T.G. (2002) RNA structure and packaging signals in the 5' leader region of the human immunodeficiency virus type 1 genome. *J Virol.* 76, 12381-12387.

Cochrane, A.W., McNally, M.T. and Mouland, A.J. (2006) The retrovirus RNA trafficking granule: from birth to maturity. *Retrovirology* 3: 18

Cramer, P., Bushnell, D. A., Kornberg, R. D. (2001) Structural Basis of Transcription:
RNA Polymerase II at 2.8 Angstrom Resolution. *Science* 292, 1863-1876.

Daboussi, M.-J., Langina, T., Deschampsa, F., Brygooa, Y., Scazzocchioa, C. and Burger, G. (1991) Heterologous expression of the Aspergillus nidulans regulatory gene nirA in Fusarium oxysporum. *Gene* 109, 155-160.

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

Dewannieux, M. and Heidmann, T. (2005) LINEs, SINEs and processed pseudogenes: parasitic strategies for genome modeling. *Cytogenet Genome Res.* 110, 35-48.

Dorman, N. and Lever, A. (2000) Comparison of Viral Genomic RNA Sorting Mechanisms in Human Immunodeficiency Virus Type 1 (HIV-1), HIV-2, and Moloney Murine Leukemia Virus. *J Virol,* 74, 11413-11417.

Dorsett, D. (1993) Distance-independent inactivation of an enhancer by the suppressor of Hairy-wing DNA-binding protein of Drosophila. *Genetics* 134, 1135-1144.

Dutko, J.A., Kenny, A.E., Gamache, E.R. and Curcio, M.J. (2010) 5' to 3' mRNA decay factors colocalize with Ty1 gag and human APOBEC3G and promote Ty1 retrotransposition. *J Virol.* 84, 5052-5066.

Duval-Valentin, G., Marty-Cointin, B. and Chandler, M. (2004) Requirement of IS911 replication before integration defines a new bacterial transposition pathway. *EMBO J.* 23, 3897-3906.

Evgen'ev, M.B. and Arkhipova, I.R. (2005) Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenet Genome Res.* 110, 510-521.

Feuerbach, F., Drouaud, J. and Lucas, H. (1997) Retrovirus-like end processing of the tobacco Tnt1 retrotransposon linear intermediates of replication. *J Virol*. 71, 4005-4015.

Finnegan, D.J. (1989) Eukaryotic transposable elements and genome evolution. *Trends Genet.* 5, 103-107.

Finnegan, D.J., Rubin, G.M., Young, M.W. and Hogness, D.S. (1978) Repeated gene families in Drosophila melanogaster. *Cold Spring Harb Symp Quant Biol* 42 Pt 2, 1053-1063.

Fudal, I., Bohnert, H.U., Tharreau, D. and Lebrun, M.H. (2005) Transposition of MINE, a composite retrotransposon, in the avirulence gene ACE1 of the rice blast fungus Magnaporthe grisea. *Fungal Genet Biol.* 42, 761–772

Fulnecek, J. and Kovarik, A. (2007) Low abundant spacer 5S rRNA transcripts are frequently polyadenylated in Nicotiana. *Mol Genet and Genomics* 278, 565-573

Furger, A., Monks, J. and Proudfoot, N.J. (2001) The retroviruses human immunodeficiency virus type 1 and Moloney murine leukemia virus adopt radically different strategies to regulate promoter-proximal polyadenylation. *J Virol.* 75, 11735–11746.

Furuichi, Y. and Shatkin, A.J. (2000) Viral and cellular mRNA capping: past and prospects. *Adv Virus Res.* 55, 135-184.

Gogvadze, E., Barbisan, C., Lebrun, M.-H. and Buzdin, A. (2007) Tripartite chimeric pseudogene from the genome of rice blast fungus Magnaporthe grisea suggests double template jumps during long interspersed nuclear element (LINE) reverse transcription. *BMC Genomics* 8: 360

Gogvadze, E. and Buzdin, A. (2009) Retroelements and their impact on genome evolution and functioning. *Cell Mol Life Sci.* 66, 3727-3742.

Gogvadze, E., Stukacheva, E., Buzdin, A. and Sverdlov, E. (2009) Human-specific modulation of transcriptional activity provided by endogenous retroviral insertions. *J Virol.* 83, 6098–6105

Goodall, G.J. and Filipowicz, W. (1989) The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell Mol Life Sci*. 58, 473-483.

Goodier, J.L. and Kazazian Jr, H.H. (2008) Retrotransposons revisited:the restraint and rehabilitation of parasites. *Cell Mol Life Sci.* 135, 23-35.

Grandbastien, M.-A. (1998) Activation of plant retrotransposons under stress conditions. *Trends Plant Sci.* 3, 181-187.

Grandbastien, M.A., Spielmann, A. and Caboche, M. (1989) Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. *Nature* 337, 376-380.

Gregory, B.D., O'Malley, R.C., Lister, R., Urich, M.A., Tonti-Filippini, J., Chen, H., Millar, A.H. and Ecker, J.R. (2008) A link between RNA metabolism and silencing affecting Arabidopsis development. *Dev Cell.* 14, 854-866.

Gu, M. and Lima, C.D. (2005) Processing the message: structural insights into capping and decapping mRNA. *Curr Opin Struct Biol.* 15, 99-106

Gulnik, S., Erickson, J.W. and Xie, D. (2000) HIV protease: enzyme function and drug resistance. *Vitam Horm.* 58, 213-256.

Haller, A.A., Nguyen, J.H. and Semler, B.L. (1993) Minimum internal ribosome entry site required for poliovirus infectivity. *J Virol.* 67, 7461-7471.

Hambor, J.E., Mennone, J., Coon, M.E., Hanke, J.H. and Kavathas, P. (1993) Identification and characterization of an Alu-containing, T-cell-specific enhancer located in the last intron of the human CD8 alpha gene. *Mol Cell Biol.* 13, 7056–7070

Hamilton, A., Voinnet, O., Chappell, L. and Baulcombe, D. (2002) Two classes of short interfering RNA in RNA silencing. *EMBO J.* 21, 4671-4679.

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

Hernández-Pinzón, I., de Jesús, E., Santiago, N. and Casacuberta, J.M. (2009) The frequent transcriptional readthrough of the tobacco Tnt1 retrotransposon and its possible implications for the control of resistance genes. *J Mol Evol.* 68, 269-278.

Heyman, T., Agoutin, B., Friant, S., Wilhelm, F.X. and Wilhelm, M.L. (1995) Plus-strand DNA Synthesis of the Yeast Retrotransposon Ty1 is Initiated at Two Sites, PPT1 Next to the 3′ LTR and PPT2 Within thepolGene. PPT1 is Sufficient for Ty1 Transposition. *J Mol Biol.* 253, 291-303.

Hirochika, H. (1993) Activation of tobacco retrotransposons during tissue culture. *Embo J.* 12, 2521-2528.

Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H. and Kanda, M. (1996) Retrotransposons of rice involved in mutations induced by tissue culture. *Proc Natl Acad Sci U.S.A.* 93, 7783-7788.

Hirose, Y and Manley, J (2000) RNA polymerase II and the integration of nuclear events. *Genes Dev.* 14, 1415-29

Hoskins, R.A., Smith, C.D., Carlson, J.W., Carvalho, A.B., Halpern, A., Kaminker, J.S., Kennedy, C., Mungall, C.J., Sullivan, B.A., Sutton, G.G., Yasuhara, J.C., Wakimoto, B.T., Myers, E.W., Celniker, S.E., Rubin, G.M. and Karpen, G.H. (2002) Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. *Genome Biol* 3: RESEARCH0085.

Hu, W., Das, O.P. and Messing, J. (1995) Zeon-1, a member of a new maize retrotransposon family. *Mol Gen Genet.* 248, 471-480.

Hug, A.M. and Feldmann, H. (1996) Yeast retrotransposon Ty4: the majority of the rare transcripts lack a U3-R sequence. *Nucleic Acids Res.* 24, 2338–2346.

Hughes, D.C. (2001) Alternative splicing of the human VEGFGR-3/FLT4 gene as a consequence of an integrated human endogenous retrovirus. *J Mol Evol.* 53, 77-79.

Irwin, P.A. and Voytas, D.F. (2001) Expression and processing of proteins encoded by the Saccharomyces retrotransposon Ty5. *J Virol.* 75, 1790-1797.

Isken, O., and Maquat, L.E. (2007) Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev.* 21, 1833-1856.

Jackson R and Kaminski A. (1995) Internal initiation of translation in eukaryotes: the picornavirus paradigm and beyond. *RNA* 1, 985-1000.

Jiao, Y., Riechmann, J.L. and Meyerowitz, E.M. (2008) Transcriptome-wide analysis of uncapped mRNAs in Arabidopsis reveals regulation of mRNA degradation. *Plant Cell Rep.* 20, 2571-2585.

Jordan, E., Saedler, H. and Starlinger, P. (1968) 0^o and strong polar mutations in the *gal* operons are insertions. *Mol Gen Genet.* 102, 353-360.

Jurica, M.S. and Moore, M.J. (2003) Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell.* 12, 5-14.

Jääskeläinen, M., Mykkänen, A.-H., Arna, T., Vicient, C., Suoniemi, A., Kalendar, R., Savilahti, H. and Schulman, A.H. (1999) Retrotransposon *BARE*-1: Expression of encoded proteins and formation of virus-like particles in barley cells. *Plant J.* 20, 413-422.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., Schulman, A. (2000) Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proc Natl Acad Sci U S A.* 97, 6603-6607.

Kalendar, R. and Schulman, A.H. (2006) IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. *Nat Protoc.* 1, 2478-2484.

Kamp, C., Hirschmann, P., Voss, H., Huellen, K. and Vogt, P.H. (2000) Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events. *Hum Mol Genet.* 9, 2563–2572.

Kankanpää, J., Schulman, A.H. and Mannonen, L. (1996) The genome sizes of Hordeum species show considerable variation. *Genome Biol.* 39, 730-735.

Kaye, J.F. and Lever, A.M.L. (1999) Human Immunodeficiency Virus Types 1 and 2 Differ in the Predominant Mechanism Used for Selection of Genomic RNA for Encapsidation. *J Virol.* 3023-3031.

Kjellman, C., Sjogren, H.O., Salford, L.G. and Widegren, B. (1999) HERV-F (XA34) is a full-length human endogenous retrovirus expressed in placental and fetal tissues. *Gene* 239, 99–107

Konkel, M.K. and Batzer, M.A. (2010) A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol.* 20, 211-221.

Kozak, M. (1989) The scanning model for translation: an update. *J Cell Biol.* 108, 229-241.

Kumar, A. and Bennetzen, J.L. (1999) Plant retrotransposons. *Annu Rev Genet.* 33, 479-532.

L'Hernault, A., Greatorex, J.S., Crowther, R.A. and Lever, A.M. (2007) Dimerisation of HIV-2 genomic RNA is linked to efficient RNA packaging, normal particle maturation and viral infectivity. *Retrovirology* 4: 90.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Landry, J.R., Medstrand, P. and Mager, D.L. (2001) Repetitive elements in the 5′ untranslated region of a human zinc-finger gene modulate transcription and translation efficiency. *Genomics* 76, 110-116.

Le Rouzic, A., Boutin, T.S. and Capy, P. (2007) Long-term evolution of transposable elements. *Proc Natl Acad Sci U S A.* 104, 19375-19380.

Lev-Maor, G., Ram, O., Kim, E., Sela, N., Goren, A., Levanon, E.Y. and Ast, G. (2008) Intronic Alus influence alternative splicing. *PLoS Genet.* 4: e1000204.

Levin, H.L., Weaver, D. C., and Boeke, J. D. (1993) Novel gene expression mechanism in a fission yeast retroelement: Tf1 proteins are derived from a single primary translation product. *EMBO J.* 12, 4885-4895.

Levin, J.G., Grimley, P.M., Ramseur, J.M. and Berezesky, I.K. (1974) Deficiency of 60 to 70S RNA in murine leukemia virus particles assembled in cells treated with actinomycin D. *J Virol.* 14, 152-161.

Levin, J.G. and Rosenak, M.J. (1976) Synthesis of murine leukemia virus proteins associated with virions assembled in actinomycin D-treated cells: evidence for persistence of viral messenger RNA. *Proc Natl Acad Sci U S A.* 73, 1154-1158.

Lewis, J.D. and Izaurralde, E. (1997) The role of the cap structure in RNA processing and nuclear export. *Eur J Biochem.* 247, 461-469.

Long, Q., Bengra, C., Li, C., Kutlar, F. and Tuan, D. (1998) A long terminal repeat of the human endogenous retrovirus ERV-9 is located in the 5′ boundary area of the human beta-globin locus control region. *Genomics* 54, 542-555.

Lunyak, V., Prefontaine, G.G., Nunez, E., Cramer, T., Ju, B.G., Ohgi, K.A., Hutt, K., Roy, R., Garcia-Diaz, A., Zhu, X., Yung, Y., Montoliu, L., Glass, C.K. and Rosenfeld, M.G. (2007) Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* 317, 248-251.

Lyons, A.J., Lytle, J. R., Gomez, J. and Robertson, H. D. (2001) Hepatitis C virus internal ribosome entry site RNA contains a tertiary structural element in a functional domain of stem-loop II. *Nucleic Acids Res.* 29, 2535-2541.

Maciolek, N.L. and McNally, M.T. (2008) Characterization of Rous sarcoma virus polyadenylation site use in vitro. *Virology* 374, 468-476.

Maeda, N. and Kim, H.S. (1990) Three independent insertions of retrovirus-like sequences in the haptoglobin gene cluster of primates. *Genomics* 8, 671-683.

Mager, D.L., Hunter, D.G., Schertzer, M. and Freeman, J.D. (1999) Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3). *Genomics* 59, 255–263

Malik, H.S. and Eickbush, T.H. (2001) Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res.* 11, 1187-1197.

Manninen, I. and Schulman, A.H. (1993) BARE-1, a copia-like retroelement in barley (Hordeum vulgare L.). *Plant Mol Biol.* 22, 829-846.

Manninen, O., Kalendar, R., Robinson, J., and Schulman, A.H. (2000) Application of BARE-1 retrotransposon markers to the mapping of a major resistance gene for net blotch in barley. *Mol Gen Genet.* 264, 325–334.

Maquat, L.E. (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.* 5, 89-99.

Marillonneta, S. and Wesslera, S.R. (1998) Extreme Structural Heterogeneity Among the Members of a Maize Retrotransposon Family. *Genetics* 150, 1245-1256.

Marquet, R., Isel, C., Ehresmann, C. and Ehresmann, B. (1995) tRNAs as primer of reverse transcriptases. *Biochimie.* 77, 113-124.

Matzke, M., Kanno, T., Huettel, B., Daxinger, L. and Matzke, A.J.M. (2006) RNA-directed DNA Methylation and Pol IVb in Arabidopsis. *Cold Spring Harb Symp Quant Biol.* 71, 449-459.

Maxwell, P.H., Belote, J.M. and Levis, R.W. (2006) Identification of multiple transcription initiation, polyadenylation, and splice sites in the Drosophila melanogaster TART family of telomeric retrotransposons. *Nucleic Acids Res.* 34, 5498-5507.

McClintock, B. (1953) Induction of instability at selected loci in maize. *Genetics* 38, 379-599.

McClintock, B. (1956) Controlling elements and gene. *Cold Spring Harb Symp Quan Biol* 21, 197-216.

McNally, M.T. (2008) RNA processing control in avian retroviruses. *Front Biosci.* 13, 3869-3883.

Meignin, C., Bailly, J.-L., Arnaud, F., Dastugue, B. and Vaury, C. (2003) The 5' 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.

Meisler, M.H. and Ting, C.N. (1993) The remarkable evolutionary history of the human amylase genes. *Crit Rev Oral Biol Med.* 4, 503-509.

Melton, D.A., Krieg, P.A., Rebagliati, M.R., Maniatis, T., Zinn, K., and Green, M.R (1984) Efficient In Vitro Synthesis of Biologically Active RNA and RNA Hybridization Probes From Plasmids Containing a Bacteriophage SP6 Promoter. *Nucleic Acids Res.* 12, 7035-7056.

Messer, L.I., Levin, J.G., Chattopadhyay, S.K., (1981) Metabolism of viral RNA in murine leukemia virus-infected cells; evidence for differential stability of viral message and virion precursor RNA. *J Virol.* 40, 683-690.

Milcarek, C., Price, R. and Penman, S. (1974) The metabolism of a poly(A) minus mRNA fraction in HeLa cells. *Cell* 3, 1-10.

Morellet, N., Jullian, N., De Rocquigny, H., Maigret, B., Darlix, J.L. and Roques, B.P. (1992) Determination of the structure of the nucleocapsid protein NCp7 from the human immunodeficiency virus type 1 by 1H NMR. *Embo J.* 11, 3059-3065.

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

Mount, S.M. (1982) A catalogue of splice junction sequences. *Nucleic Acids Res.* 10, 459-472.

Mullen, T.E. and Marzluff, W.F. (2008) Degradation of histone mRNA requires oligouridylation followed by decapping and simultaneous degradation of the mRNA both 5′ to 3′ and 3′ to 5′. *Genes Dev.* 22, 50-65.

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

Noma, K., Nakajima, R., Ohtsubo, H. and Ohtsubo, E. (1997) RIRE1, a retrotransposon from wild rice Oryza australiensis. *Genes Genet Syst.* 72, 131-140.

Ohshima, K., Hamada, M., Terai, Y. and 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.

Ooms, M., Huthoff, H., Russell, R., Liang, C. and Berkhout, B. (2004) A riboswitch regulates RNA dimerization and packaging in human immunodeficiency virus type 1 virions. *J Virol*. 78, 10814-10819.

Paulson, K.E., Matera, A.G., Deka, N. and Schmid, C.W. (1987) Transcription of a human transposon-like sequence is usually directed by other promoters. *Nucleic Acids Res.* 15, 5199-5215.

Peterson-Burch, B.D. and Voytas, D.F. (2002) Genes of the Pseudoviridae (Ty1/copia retrotransposons). *Mol Biol Evol.* 19, 1832-1845.

Ponicsan, S.L., Kugel, J.F. and Goodrich, J.A. (2010) Genomic gems: SINE RNAs regulate mRNA production. *Curr Opin Genet Dev.* 20, 149-155.

Poulter, R.T. and Goodwin, T.J. (2005) DIRS-1 and the other tyrosine recombinase retrotransposons. *Cytogenetc Genome Res.* 110, 575-588.

Proudfoot, N.J. (1991) Poly(A) signals. *Cell* 64, 671-674

Proudfoot, N.J. (2004) New perspectives on connecting messenger RNA 3′ end formation to transcription. *Curr Opin Cell Biol*. 16, 272-278.

Proudfoot, N.J., Furger, A. and Dye, M.J. (2002) Integrating mRNA processing with transcription. *Cell* 108, 502-512.

Rabson, A.B. and Graves, B.J. (1997) Synthesis and processing of viral RNA. *Cold Spring Harbor Laboratory Press*, 205–262.

Romanish, M.T., Lock, W.M., van de Lagemaat, L.N., Dunn, C.A. and Mager, D.L. (2007) Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *PLoS Genet.* 3, 51-62.

Rosenzweig, B., Liao, L.W. and Hirsh, D. (1983) Sequence of the C. elegans transposable element Tcl. *Nucleic Acids Res.* 11, 4201-4209.

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

Salditt-Georgieff, M., Harpold, M.M., Wilson, M.C., and Darnell, J.E (1981) Large heterogeneous nuclear ribonucleic acid has three times as many 5' caps as polyadenylic acid segments, and most caps do not enter polyribosomes. *Mol Cell Biol.* 1, 179-187.

Schoenberg, D.R. and Maquat, L.E. (2009) Re-capping the message. *Trends in Biochem Sci.* 34, 435-442.

Schulman, A. (2007) Molecular markers to assess genetic diversity. *Euphytica* 158, 313-321.

Sharan, C., Hamilton, N.M., Parl, A.K., Singh, P.K. and Chaudhuri, G. (1999) Identification and characterization of a transcriptional silencer upstream of the human BRCA2 gene. *Biochem Biophys Res Commun.* 265, 285-290.

Shatkin, A.J. and Manley, J.L. (2000) The ends of the affair: capping and polyadenylation. *Nat Struct Biol.* 7, 838-842.

Shen, L., Wu, L.C., Sanlioglu, S., Chen, R., Mendoza, A.R., Dangel, A.W., Carroll, M.C., Zipf, W.B. and Yu, C.Y. (1994) Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *J Biol Chem.* 269, 8466-8476.

Shen, R. and Miller, W.A. (2004) The 3' untranslated region of tobacco necrosis virus RNA contains a barley yellow dwarf virus-like cap-independent translation element. *J Virol.* 78, 4655-4664.

Shirasu, K., Schulman, A.H., Lahaye, T. and Schulze-Lefert, P. (2000) Acontiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* 10, 908-915.

Shure M, Wessler S and N, F. (1983) Molecular identification and isolation of the Waxy locus in maize. *Cell* 35, 225-233.

Smyth, D.R., Kalitsis, P., Joseph, J.L. and Sentry, J.W. (1989) Plant retrotransposon from Lilium henryi is related to Ty3 of yeast and the gypsy group of Drosophila. *Proc Natl Acad Sci U S A.* 86, 5015-5019.

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

Sunwoo, H., Dinger, M.E., Wilusz, J.E., Amaral, P.P., Mattick, J.S. and Spector, D.L. (2009) MEN ε/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.* 19, 347-359.

Suoniemi, A. (1997) Retrotransposons as active and major components of the barley genome. *dissertation*. Helsinki 1997 ISBN. 951-45-7861-9.

Suoniemi, A., Anamthawat-Jónsson, K., Arna, T. and Schulman, A.H. (1996a) Retrotransposon *BARE*-1 is a major, dispersed component of the barley (*Hordeum vulgare* L.) genome. *Plant Mol Biol.* 30, 1321-1329.

Suoniemi A, N.A., Schulman AH. (1996b) The BARE-1 retrotransposon is transcribed in barley from an LTR promoter active in transient assays. *Plant Mol Biol*. 31, 295-306.

Suoniemi, A., Schmidt, D. and Schulman, A.H. (1997) BARE-1 insertion site preferences and evolutionary conservation of RNA and cDNA processing sites. *Genetica* 100, 219-230.

Suoniemi, A., Tanskanen, J., Pentikäinen, O., Johnson, M.S. and Schulman, A.H. (1998) The core domain of retrotransposon integrase in Hordeum: predicted structure and evolution. *Mol Biol Evol.* 15, 1135-1144.

Tamura, M., Kajikawa, M. and Okada, N. (2007) Functional splice sites in a zebrafish LINE and their influence on zebrafish gene expression. *Gene* 390. 221-31.

Tanskanen, J.A., Sabot, F., Vicient, C. and Schulman, A.H. (2007) Life without GAG: The *BARE*-2 retrotransposon as a parasite´s parasite. *Gene* 390, 166–174.

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

Vicient, C.M., Jääskelainen, M.J., Kalendar, R. and Schulman, A.H. (2001a) Active retrotransposons are a common feature of grass genomes. *Plant physiol.* 125, 1283-1292.

Vicient, C.M., Kalendar, R., Anamthawat-Jonsson, K. and Schulman, A.H. (1999a) Structure, functionality, and evolution of the BARE-1 retrotransposon of barley. *Genetica* 107, 53-63.

Vicient, C.M., Kalendar, R. and Schulman, A.H. (2001b) Envelope-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. *Genome Res.* 11, 2041-2049.

Vicient, C.M. and Schulman, A.H. (2005) Variability, recombination, and mosaic evolution of the barley *BARE*-1 retrotransposon. *J Mol Evol*. 61, 275-291.

Vicient, C.M., Suoniemi, A., Anamthawat-Jonsson, K., Tanskanen, J., Beharav, A., Nevo, E. and Schulman, A.H. (1999b) Retrotransposon BARE-1 and Its Role in Genome Evolution in the Genus Hordeum. *Plant Cell.* 11, 1769-1784.

Voytas, D.F. and Boeke, J.D. (1993) Yeast retrotransposons and tRNAs. *Trends Genet.* 9, 421-427.

Voytas D.F. and Boeke,J.D. (2002) Ty1 and Ty5 of Saccharomyces cerevisiae. *In Mobile DNA II. Edited by Craig NL, Craigie R, Gellert M, Lambowitz AL. Washington, DC: ASM Press;* 631-662.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-562.

Waugh, R., McLean, K., Flavell, A.J., Pearce, S.R., Kumar, A., Thomas, B.B., and Powell, W (1997) Genetic distribution of BARE-1 retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). *Mol Gen Genet.* 253, 687-694.

Wawrzynski, A., Ashfield, T., Chen, N.W.G., Mammadov, J., Nguyen, A., Podicheti, R., Cannon, S.B., Thareau, V., Ameline-Torregrosa, C., Cannon, E., Chacko, B., Couloux, A., Dalwani, A., Denny, R., Deshpande, S., Egan, A.N., Glover, N., Howell, S., Ilut, D., Lai, H., Martin del Campo, S., Metcalf, M., O'Bleness, M., Pfeil, B.E., Ratnaparkhe, M.B., Samain, S., Sanders, I., Ségurens, B., Sévignac, M., Sherman-Broyles, S., Tucker, D.M., Yi, J., Doyle, J.J., Geffroy, V., Roe, B.A., Saghai Maroof, M.A., D., Y.N. and Innes, R.W. (2008) Replication of nonautonomous retroelements in soybean appears to Be both recent and common. *Plant physiol.* 148, 1760-1771.

Wendel, J.F. and Wessler, S.R. (2000) Retrotransposon-mediated genome evolution on a local ecological scale. *Proc Natl Acad Sci U S A.* 97, 6250-6252.

Wicker, T., Guyot, R., Yahiaoui, N. and Keller, B. (2003) CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. *Plant physiol.* 132, 52-63.

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

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

Wilusz, J.E., Freier, S.M. and Spector, D.L. (2008) 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* 135, 919-932.

Witte, C.P., Le, Q.H., Bureau, T. and Kumar, A. (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc Natl Acad Sci U S A.* 98, 13778-13783.

Xing, J., Wang, H., Belancio, V.P., Cordaux, R., Deininger, P.L. and Batzer, M.A. (2006) Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc Natl Acad Sci U S A.* 103, 17608-17613.

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

Yang, L., Duff, M.O., Graveley, B.R., Carmichael, G.G. and Chen, L.-L. (2011) Genomewide characterization of non-polyadenylated RNAs. *Genome Biol*. 12: R16.

Yuan Z, Sun X, Liu H, J. X (2011) MicroRNA genes derived from repetitive elements and expanded by segmental duplication events in mammalian genomes. PLoS One 6: e17666

Zaiss, D.M. and Kloetzel, P.M. (1999) A second gene encoding the mouse proteasome activator PA28beta subunit is part of a LINE1 element and is driven by a LINE1 promoter. *J Mol Biol.* 287, 829-835.

Zorio, D.A. and Bentley, D.L. (2004) The link between mRNA processing and transcription: communication works both ways. *Exp Cell Res.* 296, 91-97.