The retrotransposon landscape of the *Beta vulgaris* genome: Evolutionary conservation and diversity

Das genomische Profil von Retrotransposons in *Beta vulgaris*: Evolutionäre Konservierung und Diversität

DISSERTATION

zur
Erlangung des akademischen Grades Doktor der Naturwissenschaften (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Technischen Universität Dresden

vorgelegt von

Diplom-Biochemikerin Tony Heitkam

geboren am 30.09.1981 in Forst/Lausitz

1. Gutachter:	Prof. Dr. Thomas Schmidt, Technische Universität Dresden
2. Gutachter:	Prof. Dr. John Seymour Heslop-Harrison, University of Leicester

eingereicht am: 27. Mai 2011
verteidigt am: 17. Oktober 2011

“It would not surprise me if nature has evolved rather special and ingenious mechanisms so that evolution can proceed at an extremely rapid rate [...].”

Francis Crick about motors of evolution (1970)

In 1983, the Nobel Prize in Physiology or Medicine was awarded to Barbara McClintock "*for her discovery of mobile genetic elements*". Her nomination was supported, amongst others, by Francis Crick (Comfort, 2001). Since then, researchers try to understand how and to which degree transposable elements drive and accelerate genome evolution.

Acknowledgements

More than four years have passed since I started working on this dissertation. Looking back, I realize that I made enormous progress, both in scientific as well as in private aspects. That, however, would not have been possible on my own. I want to use this opportunity to thank those who supported me and those who contributed to the production of this thesis.

Above all, I am truly indebted and thankful to Prof. Dr. Thomas Schmidt for convincing me to dive into the topics of genome evolution and the role of transposable elements. When I started this thesis, I did not imagine that there was so much to tell. Thank you for the relative freedom of research, the lively discussions and your support.

Furthermore, I wish to express my gratitude to Prof. Dr. John Seymour (Pat) Heslop-Harrison and Dr. Trude Schwarzacher for the invitations to their laboratory and home that provided great opportunities to discuss research and future. I am especially honored that Prof. Dr. Heslop-Harrison agreed to review my dissertation. Not everybody has the luck to have two experts in the field as referees.

I am also obliged to the collegues and friends, who read and improved the dissertation. Dr. Gerhard Menzel and Falk Zakrzewski agreed to correct parts of this thesis in spite of a full schedule. Futhermore, Kathrin Seibt spent a weekend on my couch reading this dissertation, while Dr. Beatrice Weber examined parts of the manuscript despite a heavy tooth ache.

As for support in practical work, I want to acknowledge our FISH expert Ines Walter, our all-round coordinator Nadin Fliegner and our intern Johanna Sonntag for excellent technical assistance.

Of course, support does not only include scientific mentoring, practical advice or technical help: I wish to thank Kathrin Seibt, Dr. Beatrice Weber and Cora Wollrab for a great office atmosphere and for taking care of essential needs like tea and chocolate. Dr. Gerhard Menzel is acknowledged for his efforts in making my desk more beautiful – Thank you for the desiccation and darkness resistent potted plant! Even if not mentioned individually, all members of the TU Dresden research group “Plant cell and molecular biology” contributed to a pleasant and stimulating work environment and made sure that the lab was a place to find stability and friendship.

I also want to thank our cooperation partners Dr. Juliane Dohm, André Minoche and Dr. Heinz Himmelbauer from Barcelona and Prof. Dr. Bernd Weisshaar from Bielefeld for providing me with the first drafts of the sugar beet genome prior to publication. Analysis of these sequences opened a completely new perspective and enabled genome-wide views of transposon family origination and speciation.

Furthermore, there is a big need to acknowledge the bioinformatics teaching of Prof. Dr. Michael Schröder. It was his lectures that gave me the idea of delving into Hidden Markov Models.

Finally, I want to thank my family for their lifelong support. Mutti und Vati, danke dass ihr zu jeder Zeit für mich da wart und noch immer seid. Ihr hattet ein offenes Ohr für mich und die kleinen Hürden des Alltags und habt mich immer auf meinem Lebensweg unterstützt, sowohl moralisch, mit Rat und Tat, als auch finanziell, egal, wie die Umstände aussahen. Außerdem möchte ich den Wandlitzern – Tante Marika, Onkel Rainer, Yvonne, Yara und Yannis – dafür danken, dass Sie die Dinge ins rechte Licht rücken und mich daran erinnern was wirklich wichtig ist.

I wish to thank my husband Sascha for always believing in my abilities and simply for making my life more fulfilling. Especially in the "hot phase of writing", he ensured that I had enough time to focus on research and strictly controlled my progress by page and time statistics. He helped me in all aspects of life, including cooking, tax computation, and teaching me all there is to know about incompressible Navier-Stokes-Equations. And, last, but not least, thank you for allowing me to use your desk and chair for thesis writing.

Contributions to joint projects

Parts of this thesis have already been published in peer-reviewed scientific journals. Data and phrasings taken from these publications have not been cited individually. My co-authors Dr. Beatrice Weber, Dr. Torsten Wenke, Ulrike Frömmel and Prof. Dr. Thomas Schmidt also contributed to the projects as indicated in Chapter 11.

Acknowledgement of funding

I received financial support by the FAZIT foundation who offered me a pre-doctoral fellowship, as well as additional travel grants allowing me to visit the two conferences "*Genomic Impact of Eukaryotic Transposable Elements*" in Asilomar (USA 2009) and "*Annual Conference of the German Genetics Society*" in Cologne (Germany 2009). Similarly, a travel grant issued by the German Academic Exchange Office (DAAD) allowed me to participate in the "*Mobile DNA*" conference in Montréal (Canada 2010).

Graphical abstract

Word cloud showing the 150 most frequently used words in this dissertation. Font size directly relates to the word count. It illustrates the textual thematics of this thesis in a weighted form. Common words and words that do not convey scientific information (e.g. ‘or’, ‘and’, ‘been’, ‘Figure’, ‘Table’ and ‘*et al.*’) have been removed.

prepared with ‘wordle’ (Feinberg, 2009)

Abstract

Retrotransposons are major components of plant genomes influencing their genome size, organization and evolution. In the frame of this work, retrotransposons of the *Beta vulgaris* genome have been identified by molecular methods and whole genome bioinformatics approaches.

Neither belonging to the rosids nor asterids, *B. vulgaris* (cultivated beet including sugar beet, beet root and mangold) is taxonomically placed at a key position at the root of the core eudicots, and considerably different from traditional plant model species such as thale cress or rice. Its genome has been sequenced, and annotation is under way.

In order to compare different evolutionary lineages of *B. vulgaris* retrotransposons, long terminal repeat (LTR) and non-LTR retrotransposon family have been analyzed in detail. Full-length members have been isolated and characterized by bioinformatics, Southern and fluorescent *in situ* hybridization. Hallmarks of the LTR retrotransposon family Cotzilla are an additional *env*-like open reading frame (ORF), homogeneity of the members and the very high abundance. Most family members are evolutionarily young, and have most likely been created during recent bursts of amplification during species radiation.

In contrast, the non-LTR retrotransposon family BNR has fewer copies and is much more diverged. Although the BNR ORF2 resembles previously analyzed long interspersed nuclear elements (LINEs) of the L1 clade, its ORF1 sequence differs strongly. It lacks the zinc finger domain described for plant LINEs, but contains instead an RNA recognition motif (RRM) likely to have an RNA-binding function. Database searches revealed the presence of similar LINE families in higher plant genomes such as poplar, lotus and soybean. Comparing their reverse transcriptase regions with other retrotransposons, these BNR-like LINEs form a separate group of L1 LINEs designated as BNR subclade.

Availability of the *B. vulgaris* genome sequence allowed retrotransposon analyses on a genome-wide scale. A Hidden Markov Model-based detection algorithm has been developed in order to retrieve retrotransposon information directly from the database. Nearly 6000 *B. vulgaris* reverse transcriptase sequences have been isolated and classified into LTR retrotransposons of the Ty3-*gypsy* and Ty1-*copia* type, and non-LTR retrotransposons of the LINE type. As a result, a comprehensive overview of the retrotransposon spectrum of the *B. vulgaris* genome has been generated.

Since plant LINEs have been only rarely investigated, the *B. vulgaris* LINE composition was studied in detail. Out of 28 described LINE clades, only members of the L1 and RTE clades have been identified. Based on a minimal shared sequence identity of 60 %, they form at least 17 L1 families and one RTE family. Full-length members of all investigated L1 families have been analyzed regarding their sequence, structure and diversity.

In order to transfer the algorithm tested in *B. vulgaris* to other angiosperm genomes, twelve additional plant genomes have been queried for LINE reverse transcriptases. Key finding is the presence of only two LINE clades (L1 and RTE) in the analyzed genomes of higher plants. Whereas plant L1 LINEs are highly diverse and form at least seven subclades with members across species borders, RTE LINEs are extremely homogenized and constitute most likely only a single family per genome.

In summary, this work's results help to gain an understanding of the different strategies of retrotransposon evolution in plants, whereas the generated data directly contributes to the *B. vulgaris* genome annotation project.

Kurzfassung

Retrotransposons sind eine wesentliche Komponente von Pflanzengenomen, die sowohl die Größe und Organisation als auch die Evolution dieser Genome wesentlich beeinflussen können. Im Rahmen dieser Arbeit wurden verschiedene Gruppen von Retrotransposons des *Beta vulgaris* Genoms mittels molekularer und bioinformatischer Methoden identifiziert.

Innerhalb der dikotyledonen Blütenpflanzen gehört *B. vulgaris* (kultivierte Rübe einschließlich Zuckerrübe, Roter Beete und Mangold) weder zu den Rosiden noch zu den Asteriden, sondern nimmt eine Schlüsselposition innerhalb der Kerneudikotyledonen ein. Somit zeigt das Rübengenom wesentliche Unterschiede zu traditionellen Modellpflanzen wie *Arabidopsis thaliana* oder *Oryza sativa*. Das Genom ist bereits sequenziert, die Annotation jedoch noch nicht abgeschlossen.

Um verschiedene evolutionäre Linien von *B. vulgaris* Retrotransposons vergleichend zu untersuchen wurden insbesondere *Long Terminal Repeat* (LTR)- und Non-LTR-Retrotransposon-Familien detailliert analysiert. Vollständige Mitglieder wurden isoliert und mittels bioinformatischer Methoden, Southern- und Fluoreszenz-*in situ*-Hybridisierung untersucht. Die LTR-Retrotransposon-Familie Cotzilla ist durch einen zusätzlichen *env*-ähnlichen offenen Leserahmen (ORF), Homogenität ihrer Mitglieder und eine hohe Abundanz gekennzeichnet. Die meisten Cotzilla-Kopien sind evolutionär jung und wurden wahrscheinlich innerhalb eines kurzen Zeitraumes während der Artentstehung stark amplifiziert.

Im Gegensatz zur Cotzilla-Familie besitzt die Non-LTR-Retrotransposon-Familie BNR weniger Kopien und ist wesentlich divergierter. Während der BNR-spezifische ORF2 starke Ähnlichkeiten zu anderen pflanzlichen *Long Interspersed Nuclear Elements* (LINEs) der L1-Klade aufweist, unterscheidet sich der BNR ORF1 von diesen sehr stark. Im Gegensatz zu bereits beschrieben pflanzlichen LINEs kodiert er kein Zinkfingermotiv, sondern substituiert dieses durch ein RNA-Erkennungsmotiv (RRM). Durch Datenbanksuche konnten BNR-ähnliche LINEs in den Genomen höherer Pflanzen wie Soja, Lotus und Pappel identifiziert werden. Ein Vergleich der entsprechenden Reversen Transkriptasen (RT) mit den RTs anderer Retrotransposons zeigt, dass die BNR-ähnlichen LINEs eine separate Gruppe innerhalb der L1 LINEs bilden. Diese wurde daher als BNR-Subklade definiert.

Die Untersuchung von Retrotransposons auf Genomebene wurde durch die *B. vulgaris* Genomsequenz ermöglicht. Um Retrotransposon-Informationen direkt aus dem Genom zu extrahieren, wurde ein Hidden Markov Modell (HMM)-basierter Detektionsalgorithmus entwickelt. Annähernd 6000 *B. vulgaris* Reverse Transkriptase-Sequenzen konnten identifiziert und in LTR-Retrotransposons des Ty3-*gypsy*- beziehungsweise des Ty1-*copia*-Typs und in Non-LTR-Retrotransposons des LINE-Typs klassifiziert werden. Somit wurde ein umfassender Überblick über die Bandbreite der *B. vulgaris* Retrotransposons arhalten.

Da pflanzliche LINEs bisher nur wenig erforscht sind, wurde die *B. vulgaris* LINE Zusammensetzung genauer untersucht. Von 28 beschriebenen LINE-Kladen konnten nur Mitglieder der L1- und der RTE-Klade identifiziert werden. Basierend auf einer Identität von mindestens 60 % bilden die Sequenzen 17 L1 Familien und eine RTE Familie. Vollständige Mitglieder aller L1 Familien wurden hinsichtlich ihrer Sequenz, Struktur und Diversität analysiert.

Um den in *B. vulgaris* getesteten HMM-basierten Algorithmus auf andere Angiospermengenome zu übertragen, wurden zwölf weitere Pflanzengenome auf das Vorhandensein von LINE-spezifischen Reversen Transkriptasen untersucht. Wesentlichstes Ergebnis ist der Nachweis von nur zwei LINE-Kladen (L1 und RTE) in höheren Pflanzen. Während pflanzliche L1 LINEs hochgradig divers sind und über Artgrenzen hinaus mindestens sieben Subkladen mit Vertretern verschiedener Pflanzen bilden, sind RTE LINEs extrem homogenisiert und stellen höchstwahrscheinlich nur eine einzelne Familie pro Genom einer Art dar.

Zusammenfassend ermöglichen die Ergebnisse dieser Arbeit eine Erweiterung des Verständnisses der unterschiedlichen Evolutionsstrategien von Retrotransposons in Pflanzen. Zusätzlich tragen die gewonnen Daten zur Annotation des *B. vulgaris* Genoms bei.

Content

List of figures

List of tables

1 Introduction

In a special issue of *Science* celebrating the journal's 125th anniversary, editors and writers have chosen 125 key questions that address critical knowledge gaps in our scientific understanding of all research fields (Kennedy and Norman, 2005). The work which provided the foundation of this thesis aims to contribute to the unraveling of two of these crucial problems:

(1) "Why are some genomes really big and others quite compact?", and

(2) "What is all that 'junk' doing in our genomes?"

1.1 Plant genome sizes and the impact of repetitive DNA

For every genome there is a size. It was a puzzling observation in the beginning of genomic research that DNA content does not correlate with organismal complexity (Thomas, 1971). Especially when working on plant genomes, one of the most remarkable facts is the considerable variation in genome size. Current record holders having the smallest and the largest angiosperm genome, respectively, are the carnivorous species *Genlisea margaretae* with 63 Mb and the Japanese endemic herb *Paris japonica* with 149 Gb (Greilhuber *et al.*, 2006; Pellicer *et al.*, 2010). Such differences in DNA content are observable in all major eukaryotic groups; however, among higher organisms of increased complexity, the range of genome size in flowering plants is exceptionally large (Gregory, 2005). Variation in genome sizes of plant genomes amounts to nearly three orders of magnitude, contrasting with a much narrower mammalian genome size variation of only 5-fold, and points to less constraints on angiosperm genome size (Kejnovsky *et al.*, 2009).

The apparent lack of correlation between phenotypic complexity and genome size is generally referred to as C-value enigma, with C being the haploid DNA content given in Mb or pg. C-values of approximately 1.8 % of the global angiosperm flora have already been measured and more are still to come (Bennett and Leitch, 2011).

Despite the large differences in angiosperm DNA content, current estimations of gene numbers resemble each other and range between approximately 26,000 for thale cress and 40,000 for rice (Sterck *et al.*, 2007; Bennetzen *et al.*, 2004). If it is not the number of genes, what else causes the genome size differences?

Experiments applying DNA reannealing kinetics in the 1970s gave first clues that – apart from polyploidy and chromosome segment duplication – most DNA content variations are due to repeated DNA sequences (Flavell *et al.*, 1974). Later, comparative analysis of the gene and repeat composition of grass genomes led to the emergence of a simplified genomic model as illustrated in Figure 1.1. According to this hypothesis, differently sized genomes are made up of a similar set of genes, but also of differently amplified repetitive DNA (Moore, 1995; Vitte and Bennetzen, 2006; Flowers and Purugganan, 2009). The term "repetitive DNA" refers to homologous DNA fragments that are present in multiple copies in the genome (Jurka *et al.*, 2007). Probably the best analyzed tandemly repeated sequences are ribosomal DNA, telomeres, satellites and their corresponding arrays (reviewed in Hemleben *et al.*, 2007; Richard *et al.*, 2008). Regardless of the final genome size, repetitive DNA could be mapped to most chromosomal regions, while genes occur in clusters between blocks of repeats (Schmidt and Heslop-Harrison, 1998; Heslop-Harrison and Schwarzacher, 2011).

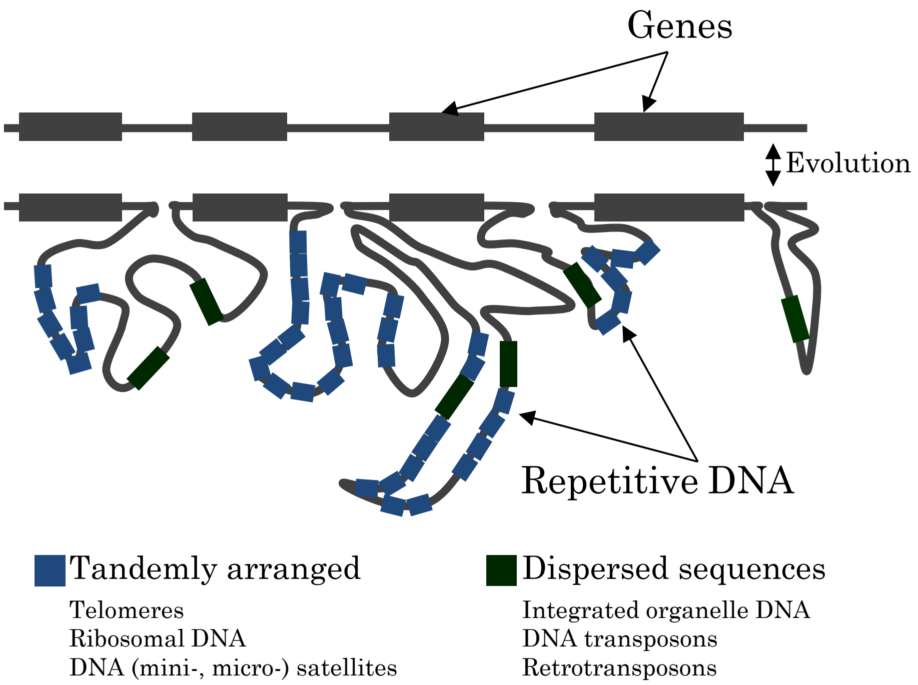


Figure 1.1: Schematic representation of two genomes with different sizes.
Genome size differences are mainly caused by a varying content of repetitive DNA. Based on their organization, repeats can be distinguished into tandemly arranged and dispersed sequences (modified after Moore, 1995).

1.2 Transposable elements

1.2.1 Transposable elements play an important role in genome evolution

One of the major fractions of the repeated genomic content is made up of transposable elements (TEs), often simply named “mobile DNA” or “jumping genes”. These are DNA sequences that have the capability to change position in the genome or to create duplicates of themselves in a process termed (retro)transposition. TEs are highly amplified and omnipresent in nearly all eukaryotes: They constitute 3 % of the yeast genome, 45 % of human DNA and nearly 85 % of some grass genomes (Boeke, 1989; Lander *et al.*, 2001; Schnable *et al.*, 2009). A nearly linear relationship between total TE DNA and genome size was proposed and proven for rice and relatives (Kidwell, 2002; Zuccolo *et al.*, 2007).

Currently, a change in perception of the relevance of TEs is experienced. After it was finally accepted that the genome was not “static, stable and immobile” (as termed later by Kazazian, 2011), transposons have been largely referred to as “junk” or “parasitic DNA” cluttering up the genome (Orgel and Crick, 1980). Now that their commonness among plants, animals, fungi and even some bacteria becomes clear, opinions grow that they are essential for life and evolution. This led to an increase of TE research in the last years. In 2002, Holmes distinguished several main fields of interest: Usage of TEs as experimental tool, the TE transposition mechanism in contrast to viral replication machinery, genome annotation and evolutionary biology. In the frame of this thesis, the latter two aspects have been considered.

Since the advent of next generation sequencing methods, there is the need to separate repeats from genic DNA. Genome annotators have to recognize TEs in order to mask them out prior to gene-related annotation. However, in most cases, the repetitive part is the last genomic fraction to be assembled correctly (Wicker *et al.*, 2006).

In the last decade, the impact of TEs on species evolution has become evident. As early as in 1984, McClintock postulated a mobilization of TEs, if the genomic regulatory system was disrupted, e.g. by repeated chromosomal breakages and rearrangements. TE activation events according to this “genomic stress hypothesis” have been confirmed after interspecific hybridization or allopolyploidization (Liu and Wendel, 2000; Comai, 2000). Furthermore, stress conditions like wounding, tissue culture, UV light and even spaceflight have been proven to lead to single amplification events or so-called “bursts of amplification” in plant genomes (Wessler, 1996; Grandbastien, 1998; Ramallo *et al.*, 2008; Long *et al.*, 2009). These environmental triggers are mediated by epigenetic modifications. The correlation of DNA demethylation and transposition, for example, was shown by TE activation in *Arabidopsis thaliana* methylation-deficient mutants (Tsukahara *et al.*, 2009). These processes, normally silenced in somatic cells, take place in the germ line of plants and animals and allow fixation of the changes in the offspring (Slotkin *et al.*, 2009; reviewed by Feng *et al.*, 2010). The release of transpositional repression enables genomic variation on a much larger scale than the accumulation of point mutations. In primate evolution, for example, all major divergence points correspond temporally with TE amplification bursts (Kim *et al.*, 2004). The increased mutagenic activity seems to facilitate a fast environmental adaptation that may even lead to species formation (Oliver and Greene, 2009; Zeh *et al.*, 2009).

In order to compensate genome size increases by retrotransposition, molecular mechanisms evolved that have the opposite effect (Bennetzen and Kellogg, 1997; Devos *et al.*, 2002). During TE amplification bursts, numerous new targets for legitimate or illegitimate recombination are generated. The rate of these DNA elimination processes can be sufficient to completely reverse the genome expansion process (Hawkins *et al.*, 2009).

1.2.2 Classification of transposable elements

The sheer amount of TEs using different transposition mechanisms and exhibiting various structural features led to several attempts for a TE classification system. Today, this is still cause of controversies.

Already in the early years of transposon research, a major split into two classes was proposed depending on the TE's mechanism of transposition (Finnegan, 1989): "Class I elements transpose by reverse transcription of an RNA intermediate, while class II elements transpose directly from DNA to DNA." In other words, class II elements, today better known as DNA transposons, are able to move in the genome in a *cut and paste* manner. Only small DNA duplications, "transposon footprints" are left at sites where they excise. Key enzyme facilitating this jump is a transposase. Alternatively, a replicative mode of transposition has been observed to occur in certain cell types or stages, e.g. in pregerminal and postmeiotic cells of maize (Raizada *et al.*, 2001). During this mechanism, exisions do not occur, and DNA transposons duplicate using a semi-conservative DNA replication (reviewed in Craig, 1995).

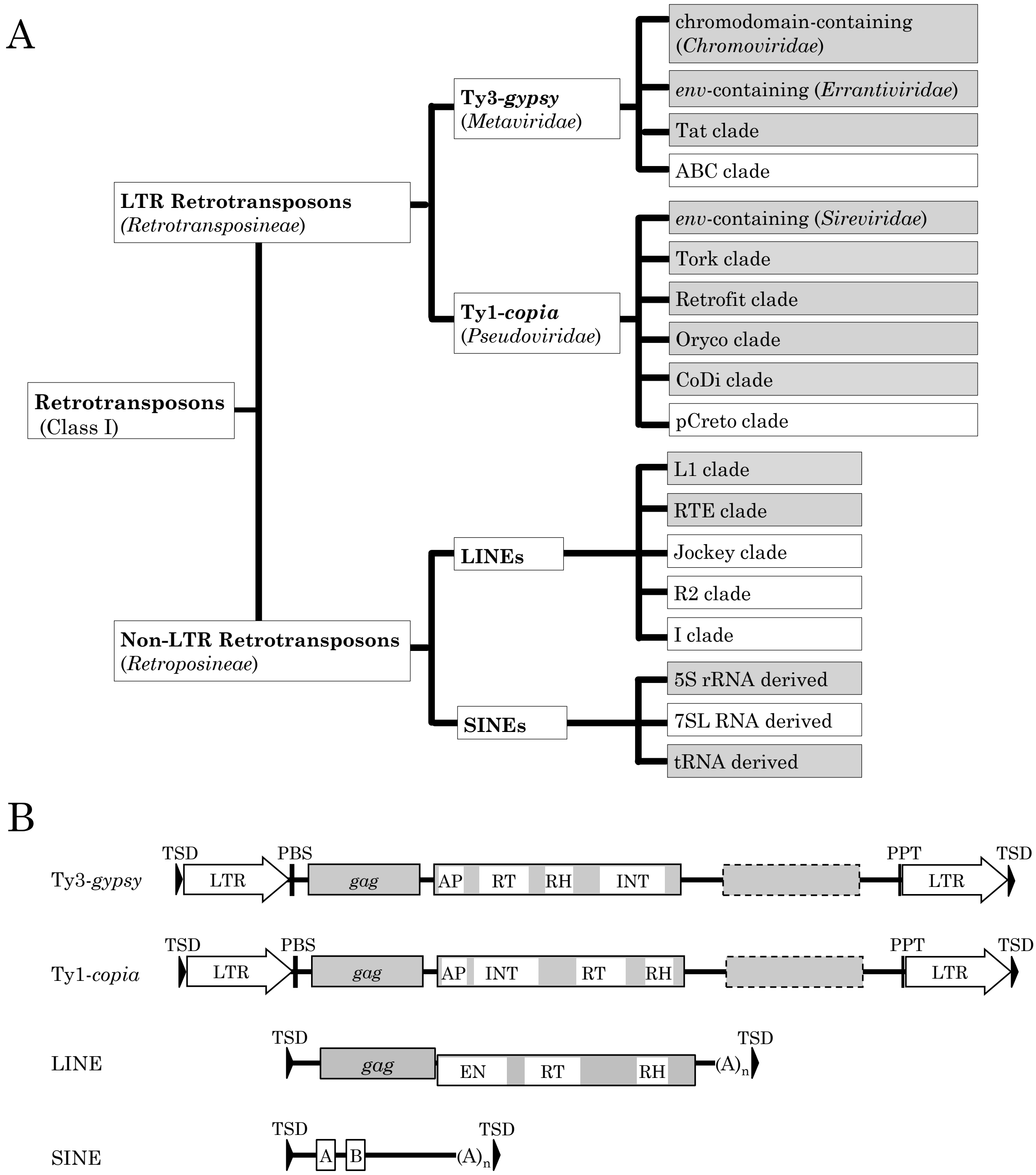


Figure 1.2: Main groups of class I transposable elements (retrotransposons).
(A) The dendrogram shows a simplified classification system including all orders and clades mentioned in this thesis (Finnegan, 1989; Wicker *et al.*, 2007; Llorens *et al.*, 2009; Kapitonov *et al.*, 2009). Retroelement clades containing members of plant genomes are shaded in grey.
(B) Structure of the four different retrotransposon orders. The grey rectangles with continuous borders represent ORFs, while rectangles with dashed borders mark optional ORFs. Conserved domains are: *gag*, aspartatic protease (AP), integrase (INT), endonuclease (EN), reverse transcriptase (RT) and RNaseH (RH). The open terminal arrows delimiting Ty3-*gypsy* or Ty1-*copia* retrotransposons represent long terminal repeats (LTRs). All retrotransposons are flanked by target site duplications (TSDs). The untranslated regions (UTRs), preceding and following the ORFs are shown as continuous black lines. Further conserved sequence features have been the primer binding site (PBS) and polypurine tract (PPT) of LTR retrotransposons, and the poly(A) tail of non-LTR retrotransposons. SINEs do not harbor protein-coding regions. Instead, two conserved regions derived from cellular RNA genes, named A- and B-box, are contained within. The drawing is not to scale.

The DNA-directed mechanisms of DNA transposons contrast with the reverse transcriptase-guided machinery of retrotransposons, the *copy and paste* elements of class I. They are able to produce copies of themselves that integrate at another position, while the original transposon remains unchanged. Here, an RNA serves as intermediate and transmits the genetic information to the target site. Based on sequence similarity and structural features, a subclassification of both classes is possible according to one of the proposed guidelines (Capy, 2005; Jurka *et al.*, 2007; Wicker *et al.*, 2007). Figure 1.2 A shows a simplified classification system of retrotransposons that was strictly applied to all TEs that have been analyzed in this thesis.

Depending on the presence of long terminal repeats (LTRs), retrotransposons are grouped into two categories: LTR and Non-LTR retrotransposons. Comparisons of animal and plant genomes show great differences in abundance of both retrotransposon groups (Figure 1.3). While non-LTR retrotransposons underwent enormous amplification in mammals, they populate plant genomes in much less frequently. A contrasting picture provides the analysis of LTR retrotransposon distribution: Though highly abundant in angiosperm genomes, they occur in mammals only in low copy numbers.

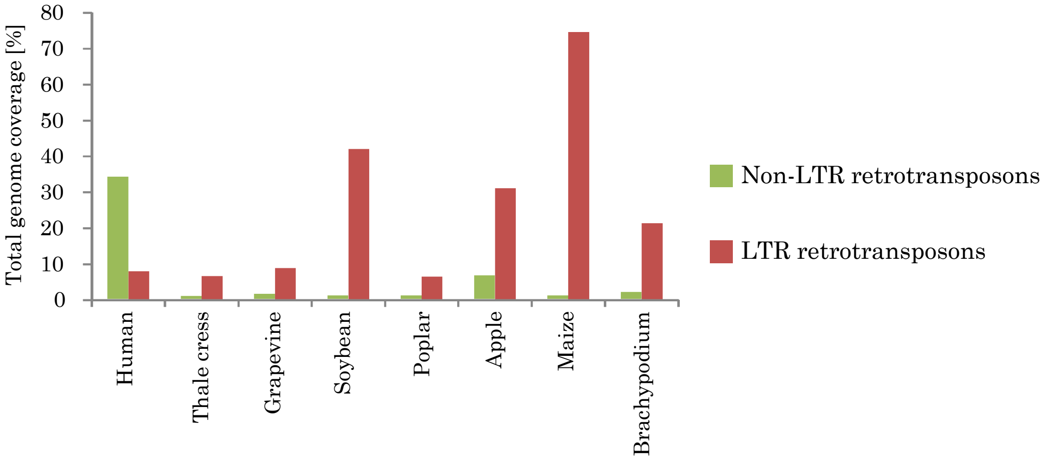


Figure 1.3: Differences of retrotransposon abundance in mammalian and plant genomes. While Non-LTR elements are much more frequent than LTR retrotransposons in mammalian genomes (e.g. human), their abundance in plant genomes is reversed. Data was taken from whole genome sequencing and annotation reports (Lander *et al.*, 2001; Swarbreck *et al.*, 2008; Jaillon *et al.*, 2007; Schmutz *et al.*, 2010; Tuskan *et al.*, 2006; Velasco *et al.*, 2010; Baucom *et al.*, 2009; Vogel *et al.*, 2010). Green = Non-LTR retrotransposons and red = LTR retrotransposons.

1.2.2.1 LTR retrotransposons

Hallmarks of LTR retrotransposons are two identical LTRs that have been generated during their transposition and flank the TE's coding regions. Transcriptional promotors are contained within them, and while the 5' LTR drives transcription, the 3' LTR works as a transcription terminator.

Depending on the order of encoded proteins and structure of their reverse transcriptases (RT), LTR retrotransposons are either classified as Ty3-*gypsy*-like or Ty1-*copia*-like elements (Kumar and Bennetzen, 1999; Hull, 2001). LTR retrotransposons, in particular Ty1-*copia* families, have been characterized in a wide range of plant taxa and constitute a heterogenous population of retrotransposon sequences. The extreme diversity and the vast amount of subfamilies are often a result of the error prone transposition mechanism (Casacuberta *et al.*, 1997; Vershinin and Ellis, 1999). Based on their RT sequence, Ty3-*gypsy* and Ty1-*copia* retrotransposons are subclassified into numerous clades (Figure 1.2 A). The classification system chosen here is based on Llorens *et al.* (2009).

LTR retrotransposons resemble retroviruses and encode the two open reading frames (ORFs) *gag* and *pol*, which are sometimes fused to form a *gag-pol* ORF: Their typical structure is presented in Figure 1.2B. The *gag* ORF encodes a nucleocapsid forming protein, while the *pol* polyprotein includes enzymatic domains for reverse transcription and integration of new copies. The *pol* gene encodes an aspartic proteinase (AP), responsible for the post-translational processing of the *pol* protein product, a reverse transcriptase (RT) and RNaseH (RH) carrying out reverse transcription, and an integrase (IN) facilitating genomic insertion of the new LTR retrotransposon copy.

Retroviruses and LTR retrotransposons are related and replicate through a cycle of successive transcription, reverse transcription, and integration into the genome (Boeke and Corces, 1989). The major structural difference between most retrotransposons and retroviruses is the presence of an envelope gene (*env*) in retroviruses, which is essential for infectivity. However, some LTR retroelements, mostly Ty3-*gypsy*-like, encode an additional ORF with similarities to a retroviral *env* ORF (Song *et al.*, 1994; Wright and Voytas, 2002). Because of the high similarity of the retroviral and Ty3-*gypsy* RT gene, a common origin of retroviruses and *env*-like Ty3-*gypsy* retrotransposons has been suggested (Malik *et al.*, 2000). With the identification of *SIRE*1 in soybean (Laten and Morris, 1993; Laten *et al.*, 1998), a Ty1-*copia* retrotransposon possessing a putative *env* gene has been discovered. Further members of the so-called Sireviruses have been identified in plants such as *Arabidopsis thaliana*, maize, maritime pine, tomato and *Lotus japonicus* (SanMiguel *et al.*, 1996; Kapitonov and Jurka, 1999; Peterson-Burch *et al.*, 2000; Holligan *et al.*, 2006; Miguel *et al.*, 2008). Since their *env*-like ORFs are highly variable (Havecker *et al.*, 2005), different scenerios regarding the origin of Sireviruses have been proposed (Kumar, 1998; Bousios *et al.*, 2010).

1.2.2.2 Non-LTR retrotransposons

Non-LTR retrotransposons are subdivided into long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). While LINEs are able to proliferate in the genome as an autonomous entity, SINEs do not encode any proteins and rely on the amplification machinery of LINEs.

Based on structural features and the phylogeny of RTs, LINEs are grouped into five main groups, called L1, RTE, R2, I and Jockey (Figure 1.2 A), which can be subdivided in currently 28 clades (Kapitonov *et al.*, 2009). In plant genomes, so far only LINEs of the L1 and RTE clade have been reported (Wenke *et al.*, 2009; Zupunski *et al.*, 2001).

Though structured simpler, many of the catalytical domains present in LTR retrotransposons are also encoded by LINEs. L1 and most RTE elements have two ORFs: Similar to the *pol* protein, LINE ORF2 encodes an endonuclease, a reverse transcriptase, and a putative RNaseH domain. The RT works as key enzyme to enable target-primed reverse transcription (TPRT), the transposition mechanism proposed for LINEs (Dhellin *et al.*, 1997; Ostertag and Kazazian, 2001a; Zingler *et al.*, 2005). For ORF1 proteins, a nucleic acid chaperone function is postulated, whereby a multimeric complex is formed that binds LINE mRNA and protects it from degradation (Martin, 2006). Therefore, the amino acid sequence of ORF1 generally has the capacity to bind and stabilize LINE mRNA. All plant LINEs characterized so far contain a zinc finger domain to exert RNA-binding function.

The first plant LINE, cin4, was discovered in *Zea mays* (Schwarz-Sommer *et al.*, 1987). Other LINEs have been isolated and characterized in the genomes of *Arabidopsis thaliana* (Ta*11-1* and *ATLN*), *Lilium speciosum* (*del2*), *Cannabis sativa* (LINE-CS), *Hordeum vulgare* (BLIN) and *Chlorella vulgaris* (Zepp) (Wright *et al.*, 1996; Noma *et al.*, 2000; Leeton and Smyth, 1993; Sakamoto *et al.*, 2000; Vershinin *et al.*, 2002 ; Higashiyama *et al.*, 1997). For two LINE families, *Karma* from *Oryza sativa* and *LIb* of the *Ipomoea batatas* genome, transpositional activation has been shown (Komatsu *et al.*, 2003; Yamashita and Tahara, 2006).

Publication of the human genome revealed that over 20 % of its DNA consists of LINE L1 sequences (Lander *et al.*, 2001). Since then, a large body of knowledge about mammalian LINEs has been accumulated (reviewed in Babushok and Kazazian, 2007; Belancio *et al.*, 2008). In plants, LINEs are – though ubiquitiously present (Turcotte *et al.*, 2001) – much less abundant and not regularly studied. A flood of data was expected from recent plant genome annotation projects. However, except for maize LINEs

(Baucom *et al.*, 2009) which were grouped into 31 families, no stringent analyses of element structure, family number or sequence diversity have been conducted, leaving plant LINE research without significant output for the past few years.

It still has to be verified, if the findings of mammalian LINE research can be applied to their plant homologues. Therefore, it is important to investigate in which detail mammalian and plant LINEs differ from or resemble each other.

1.3 Sugar beet genome analysis

1.3.1 Introduction to sugar beet breeding and genetics

Sugar beet is a cultivated variety of *Beta vulgaris*. Due to the high concentration of sucrose in its root, it serves as crop plant. Alongside sugar cane, it is one of only two crops that produce the sugar consumed world-wide. However, contrary to cane, sugar beet can also be grown in temperate climates (Draycott, 2006).

The wild ancestors of today's sugar beet have been domesticated for approximately 2500 years. Depending on the selection for leafy or swollen parts, different kinds of cultivars have been bred. Today, four groups are differentiated based on application in agriculture and industry: Sugar beet, fodder beet, garden beet and leaf beet (Lange *et al.*, 1999). Present-day varieties of sugar beet can reach a sugar content of up to 17 % and are not only grown for the food industry, but also for production of bioethanol and biogas (www.kws.de).

In order to achieve novel breeding aims like pest or herbicide resistance, drought or salt tolerance and yield increase, a thorough understanding of the crop's genetic base is essential. *B. vulgaris* has a diploid, relatively small genome (1C = 758 Mb, Arumuganathan and Earle, 1991) and a moderate number of chromosomes (2n = 18). Therefore, it is well-suited for the analysis of genomic organization through cytological and molecular means (Gindullis *et al.*, 2001b, Desel *et al.*, 2002). In the near future, beet genome research will reach a new stage, since the beet genome has been sequenced and annotation is under way (www.gabi.de).

1.3.2 Systematic position of sugar beet

Taxonomically, *B. vulgaris* is placed at a key position at the root of the core eudicots within the flowering plants, neither belonging to the rosids nor asterids (Figure 1.4 A; Angiosperm Phylogeny Group, 2009). Consequently, this plant is not closely related to any of the traditional plant models like thale cress, rice, maize, soy, poplar or grapevine, which also have been covered by whole genome sequencing efforts (Arabidopsis Genome Initiative, 2000; Goff *et al.*, 2002; Schnable *et al.*, 2009; Schmutz *et al.*, 2010; Tuskan *et al.*, 2006; Jaillon *et al.*, 2007). Sugar beet genome analysis does therefore not only contribute to crop breeding, but also provides essential new information on plant genome evolution.

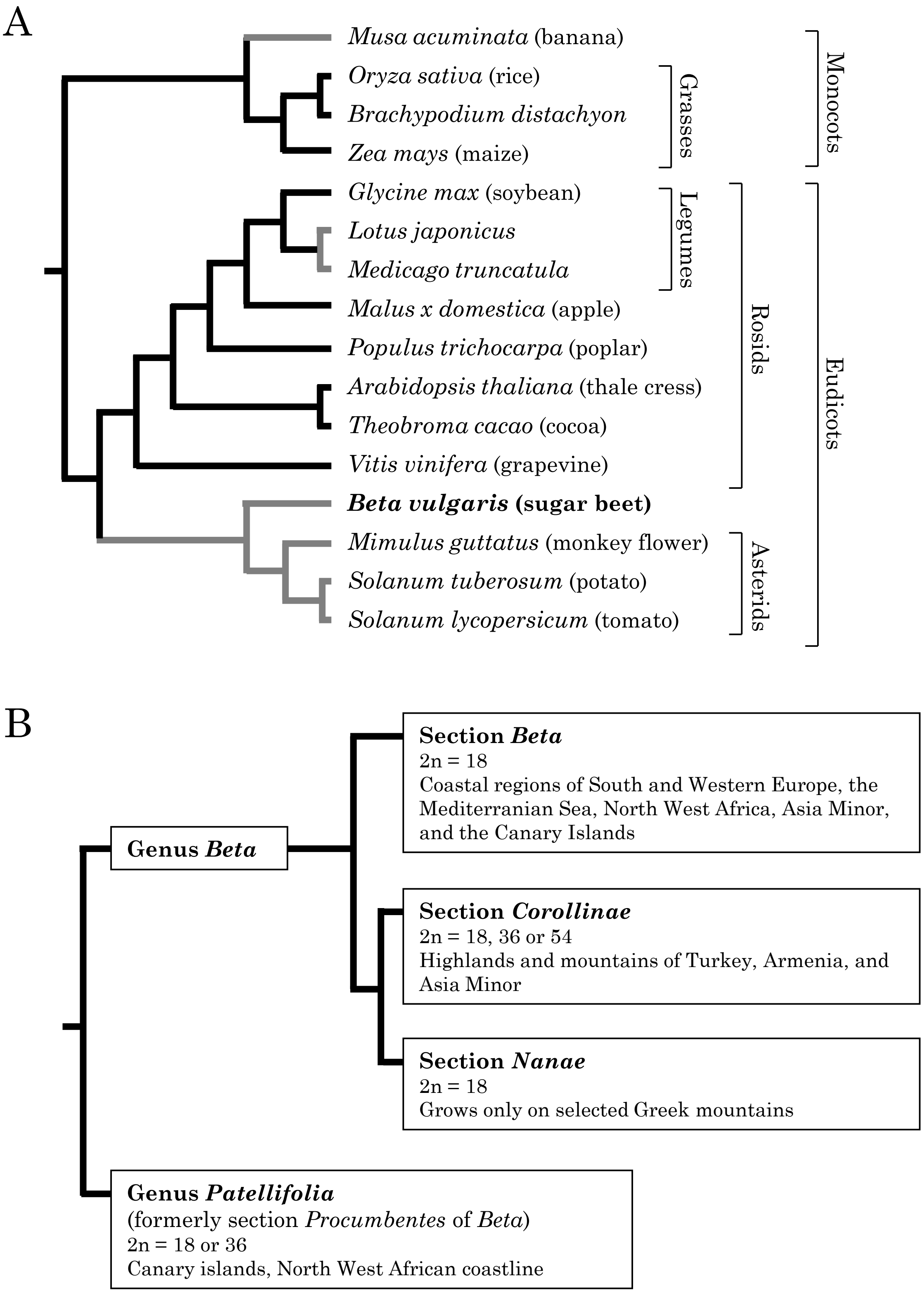

A
Musa acuminata (banana)
Oryza sativa (rice)
Brachypodium distachyon
Zea mays (maize)
Glycine max (soybean)
Lotus japonicus
Medicago truncatula
Malus x domestica (apple)
Populus trichocarpa (poplar)
Arabidopsis thaliana (thale cress)
Theobroma cacao (cocoa)
Vitis vinifera (grapevine)
Beta vulgaris (sugar beet)
Mimulus guttatus (monkey flower)
Solanum tuberosum (potato)
Solanum lycopersicum (tomato)
Grasses
Monocots
Legumes
Rosids
Eudicots
Asterids
B
Genus Beta
Section Beta
2n = 18
Coastal regions of South and Western Europe, the Mediterranian Sea, North West Africa, Asia Minor, and the Canary Islands
Section Corollinae
2n = 18, 36 or 54
Highlands and mountains of Turkey, Armenia, and Asia Minor
Section Nanae
2n = 18
Grows only on selected Greek mountains
Genus Patellifolia
(formerly section Procumbentes of Beta)
2n = 18 or 36
Canary islands, North West African coastline

Figure 1.4: Taxonomy of *B. vulgaris* and the genus *Beta*.
(A) Dendrogram showing the relationship of *B. vulgaris* to a selection of angiosperms according to APG III orders (Angiosperm Phylogeny Group, 2009). Black branches indicate that the corresponding plant genome is published, while a branch in grey denotes that sequencing or publishing is in progress. Branch lengths are not to scale.
(B) Overview of relatedness, distribution and chromosome numbers of plants belonging to the genera *Beta* and *Patellifolia*.

Apart from *B. vulgaris* cultivars, there is a number of wild beet species with a wide gene pool, potentially interesting for breeding purposes (Figure 1.4 B). They belong to the genus *Beta* (family *Amaranthaceae*, order *Caryophyllales*), which includes the sections *Beta* (with all its cultivars), *Corollinae*, *Nanae* and formerly *Procumbentes*, now named genus *Patellifolia* (Ulbrich, 1934; Scott *et al.*, 1977). Based on molecular marker assays, a merging of the section *Corollinae* with the unispecific section *Nanae* is proposed (Kadereit *et al.*, 2006).

Closely related to the genus *Beta* and considered here as biological outgroup and reference species are the leafy crop spinach (*Spinacia oleracia*) and the new world grain quinoa (*Chenopodium quinoa*).

1.3.3 Repeats dominate the *B. vulgaris* genome

Intense cytogenetic and molecular analysis already led to a basic understanding of *B. vulgaris* genome organization. The dimension of the repetitive fraction was measured to account for 63 % by Flavell *et al.* as early as 1973. Contributing to this number are major tandem repeat families that constitute centromeres, intercalary regions and subtelomeres in high copy numbers (Schmidt and Metzlaff, 1991; Schmidt *et al.*, 1991; Dechyeva and Schmidt, 2006). In the last years, a *c_0t-1* library containing only highly repetitive clones has been produced providing the basis for the identification of a number of dispersed minisatellites and other repeat families (Zakrzewski *et al.*, 2010). DNA transposon families of *mariner* TEs and their corresponding non-autonomous MITEs have also been characterized regarding their chromosomal distribution and integration pattern (Jacobs *et al.*, 2004; Menzel *et al.*, 2006).

Apart from tandemly organized repeats and class II transposons, first insights in *B. vulgaris* retrotransposon composition have been obtained by analyses of reverse transcriptase sequences amplified with degenerate primers. The presence of RT sequences similar to LINEs and Ty1-*copia* TEs, designated BNR and Tbv, respectively, has been proven (Schmidt *et al.*, 1995; Kubis *et al.*, 1998). In further approaches, full-length members of the LINE family BvL and the Ty1-*copia* family SALIRE have been isolated and sequenced (Wenke *et al.*, 2009; Weber *et al.*, 2010).

Additionally, nested LTR Ty3-*gypsy* elements, *Beetle*1 and *Beetle*2, containing a chromodomain responsible for targeted integration into centromeres have been detected in the sister genus *Patellifolia*. Hybridization of their RT to *B. vulgaris* chromosomes showed weak signals indicating the presence of related elements in sugar beet (Weber and Schmidt, 2009).

In summary, the *B. vulgaris* genome has accumulated members of each TE class, as well as a number of tandem repeats during its evolution, with many of them already being characterized. This research also provided the base for the construction of a linear chromosome model describing the organization of repeats and genes in higher plants (Schmidt and Heslop-Harrison, 1998). However, exact information regarding TE numbers, family structure, diversity and homogeneity is still missing.

1.4 Aim of this work

Retrotransposons have been reported to occur in vast numbers and high diversity, enabling them to colonize large fractions of plant genomes (Kumar and Bennetzen, 1999). In preliminary works, however, analysis of *B. vulgaris* full-length members has only be carried out for the Ty1-*copia* LTR retrotransposon family SALIRE and the LINE family BvL (Weber *et al.*, 2010; Wenke *et al.*, 2009). In order to illuminate retrotransposon diversity within the genome of a single species, two additional retrotransposon families of *B. vulgaris* shall be detected and investigated.

Based on the analyses of a *c₀t-1* DNA library, as well as on sequence comparisons with already analyzed TEs, the LINE family BNR and the Ty1-*copia* family Cotzilla have been chosen for in-depth examination. Full-length members of the selected families will be investigated by bioinformatics, Southern and fluorescent *in situ* hybridization. Their differences in sequence, structure, diversity and integration preference will be comparatively analyzed.

During this thesis, the first drafts of the *B. vulgaris* genome became available, allowing a TE analysis on a larger scale. The foundation of this work is provided by the information about the *B. vulgaris* retrotransposable fraction, which is deeply buried inside the sequence database. The development of a method to extract the retrotransposon data from the database, and the creation of an overview about the type of retrotransposons present in *B. vulgaris*, is one of the main aspects of this thesis. This retrotransposon extraction method can be also used to gain more information on plant LINE evolution. For this purpose, the whole spectrum of *B. vulgaris* LINE families will be characterized and annotated in order to expose sequence differences and similarities. Furthermore, comparison with the LINE content of selected higher plants might lead to an expansion of their classification scheme.

In conclusion, the retrotransposon variety in the *B. vulgaris* genome will be presented, allowing quick referencing. Whole genome detection approaches are to be complemented by an in-depth analysis of single family members in order to provide a general picture.

2 Material and Methods

2.1 Material

2.1.1 Plant Material

Plants were grown under long day conditions in the greenhouse of the Institute of Botany, TU Dresden. The species examined and their accessions are listed in Table 2.1. Fodder beet Brigadier seeds and leaf beet Vulkan seeds were aquired from the "Quedlinburger Saatgut GmbH", while spinach and quinoa were purchased at retail. Seeds of other accessions were provided by the IPK Gatersleben.

Table 2.1: Plants of the genera *Beta, Patellifolia, Chenopodium* and *Spinacia*

Genus	Section	Species/Subspecies	Common name	Accession
Beta	*Beta*	*Beta vulgaris* ssp. *vulgaris* var. conditiva "KWS 2320"	Sugar beet	BETA 1261
		Beta vulgaris ssp. *vulgaris* var. altissima "Brigadier"	Fodder beet	175V
		Beta vulgaris ssp. *vulgaris* var. crassa	Garden beet	41
		Beta vulgaris ssp. *vulgaris* var. vulgaris "Vulkan"	Leaf beet	806032
		Beta vulgaris ssp. *maritima*	Wild beet	BETA 999
		Beta vulgaris ssp. *adanensis*	Wild beet	BETA 1473
		Beta macrocarpa	Wild beet	BETA 574
		Beta patula	Wild beet	BETA 548
	Corollinae	*Beta corolliflora*	Wild beet	BETA 408
		Beta macrorhiza	Wild beet	BETA 545
	Nanae	*Beta nana*	Wild beet	BETA 541
Patellifolia		*Patellifolia procumbens*	Wild beet	BETA 951
		Patellifolia patellaris	Wild beet	BETA 534
		Patellifolia webbiana	Wild beet	BETA 927
Chenopodium		*Chenopodium quinoa*	Quinoa	
Spinacia		*Spinacia oleracea*	Spinach	

The double haploid *Beta vulgaris* genotype "KWS 2320" was used as reference. If not denoted otherwise, experiments were performed using the corresponding genomic DNA.

2.1.2 PRO1 BAC library

The fragment addition line PRO1, produced by a series of crosses and backcrosses of *B. vulgaris* and *P. procumbens*, carries a single, stably inherited chromosomal fragment of *P. procumbens* as well as the complete chromosomal set of *B. vulgaris* (Jung and Wricke, 1987). DNA of this chromosomal mutant line was used to create a PRO1 BAC library. This library contains 50,304 bacterial artificial chromosomes (BACs) with an average insert size of 125 kb. Based on the haploid genome size of 758 Mb, the library represents eight genome equivalents. Approximately 99.8 % of its genomic content correspond to *B. vulgaris*, while 0.2 % equate to *P. procumbens* DNA (Gindullis *et al.*, 2001a).

2.1.3 Culture media and antibiotics

Culture media were produced with desalinated water, followed by autoclaving at 121 °C and 2 bar for 20 min.

Luria-Bertani (LB) medium		
Tryptone/Pepton	1	%
Yeast extract	0.5	%
NaCl	1	%

LB agar plates
LB medium with 1.5 % agar

LB indicator plates		
LB agar plates with		
IPTG	0.5	mM
X-Gal	0.004	%

LB freezing medium		
LB medium with		
K_2HPO_4	36	mM
KH_2PO_4	13.2	mM
NaCitrate	1.7	mM
$MgSO_4$	0.4	mM
$(NH_4)_2SO_4$	6.8	mM
Glycerine	4.4	% (v/v)

SOC medium	*Storage: -20 °C*	
Tryptone/Pepton	2	%
Yeast extract	0.5	%
NaCl	10	%
KCl	2.5	mM
$MgCl_2$	10	mM
pH 7.0		
Addition after autoclaving:		
$MgSO_4$	10	mM
Glucose	20	mM
Sterile filtration		

Antibiotics		
Ampicillin	100	µg/ml Medium
Tetracycline	5	µg/ml Medium
Chloramphenicol	12.5	µg/ml Medium

2.1.4 Buffers and solutions

Buffers and solutions were prepared as listed below, using desalinated water.

CTAB (1x)

Tris/HCl (pH 8.0)	0.1	M
EDTA (pH 8.0)	10	mM
NaCl	0.7	M
CTAB	1	% (w/v)
Addition before usage:		
ß-Mercaptoethanol	0.2	% (v/v)

DAPI solution

Stock: DAPI in H_2O	100	µg/ml
Final: DAPI in McIlvaine buffer	2	µg/ml

Denhardt solution (100x) *Storage: -20 °C*

PVP	2	%
BSA (Fraction V)	2	%
Ficoll 400	2	%

Denhardt medium *Storage: -20 °C*

Denhardt solution	5	x
SSC	5	x
SDS	0.5	%

Enzyme buffer (10%)

Citric acid (pH 4.5)	40	mM
Sodium citrate	60	mM

Enzyme solution

Cellulase (*Aspergillus niger*, 0,45 U/mg)	2	%
Cellulase (Onozuka-R10, 1,3 U/mg)	4	%
Cytohelicase (*Helix pomatia*)	2	%
Pectolyase (*Aspergillus japonicus*, 3,9 U/mg)	0.5	%
Pectinase (*Aspergillus niger*, 445 U/ml)	5	%
in sterile 1x enzyme buffer		

Fixation Solution (freshly prepared)

Methanol	75	%
Glacial acetic acid	25	%

Loading Buffer (10x)

TAE	1	x
Glycerine	50	%
Bromophenol blue	0.1	%
Xylene cyanol	0.1	%

McIlvaine buffer

Na_2HPO_4 x 12 H_2O	164	mM
Citric acid	8	mM
pH 7.0		

Pre-hybridization medium

Denhardt medium	50	ml
Salmon sperm (denatured)	1	ml
EDTA (0.5 M)	1	ml
BSA (Fraction V)	5	mg

SSC (20x)

NaCl	3	M
Sodium citrate	0.3	M

SSC/Tween (4x)

SSC	4	x
Tween	0.2	%

TAE Buffer (50x)

TrisBase	242	g
EDTA (pH 8.0)	50	mM
Glacial acetic acid	57.1	ml
ad H_2O	1000	ml

TE buffer

EDTA	1	M
Tris/HCl	10	mM

2.1.5 Chemicals, consumables, kits and enzymes

For molecular biology experiments, a number of chemicals, consumables, kits and enzymes have been used. These are listed in Table 2.2, Table 2.3, Table 2.4, and Table 2.5.

Table 2.2: Chemicals and consumables

Name	**Company/Supplier**
α-[32P]-dATP, 3000 Ci/mmol	Amersham Pharmacia Ltd, UK
α-[32P]-dCTP, 3000 Ci/mmol	Amersham Pharmacia Ltd, UK
Acetic acid	Merck, Darmstadt
Acetone	Roth, Karlsruhe
Agarose, Seakem® LE	Biozym, Hess. Oldendorf
Ammonium persulfate	Merck, Darmstadt
Ammonium sulfate	Roth, Karlsruhe
Ampicillin	Roth, Karlsruhe
Anti-digoxigenin antibody FITC	Roche Diagnostics GmbH, Mannheim
ATP	MBI Fermentas GmbH, St. Leon-Rot
Bacto-agar	Roth, Karlsruhe
Agar-agar	Roth, Karlsruhe
Biotin-16-dUTP	Roche Diagnostics GmbH, Mannheim
Blocking solution	Roche Diagnostics GmbH, Mannheim
β-Mercaptoethanol	Roth, Karlsruhe
Boric acid	Appli Chem, Darmstadt
Bromophenol blue	Biomol Feinchemikalien, Hamburg
BSA (Fraction V)	Roth, Karlsruhe
Chloramphenicol	Serva Feinchemikalien, Heidelberg
Chloroform	AppliChem, Darmstadt
Chromic acid	Merck, Darmstadt
Citifluor AF1	Chem. Lab. Canterburry, UK
Citric acid	Serva Feinchemikalien, Heidelberg
DAPI	Fluka Chemie GmbH, Buchs, Schweiz
DEPC	Roth, Karlsruhe
Dextran sulfate	Roth, Karlsruhe
DIG-nick translation mix	Roche Diagnostics GmbH, Mannheim
Digoxigenin-11-dUTP	Roche Diagnostics GmbH, Mannheim
Dimethylformamide	Roth, Karlsruhe
dNTP mix	MBI Fermentas GmbH, St. Leon-Rot
DMSO	Roth, Karlsruhe
EDTA	Roth, Karlsruhe
Ethanol	Appli Chem, Darmstadt or Roth, Karlsruhe
Ethidium bromide	Roth, Karlsruhe
Ficoll® 400	Pharmacia BiotechAB, Uppsala, Schweden
Formamide	Sigma-Aldrich Chemie GmbH, Steinheim
GeneRuler™ 100bp DNA Ladder	MBI Fermentas GmbH, St. Leon-Rot
GeneRuler™ 1kb DNA Ladder	MBI Fermentas GmbH, St. Leon-Rot
Glucose	Roth, Karlsruhe
Glycerine	Roth, Karlsruhe
G/T mix	MBI Fermentas GmbH, St. Leon-Rot
Hybond N+, Nylon membrane	Amersham Biosciences, Freiburg
Hydroxyquinoline	Merck, Darmstadt
Hydrochloric acid	Roth, Karlsruhe
Hyperfilm™	Amersham Biosciences, Freiburg
Immersion oil 518C	Carl Zeiss, Oberkochen
IPTG	Roth, Karlsruhe
Isoamyl alcohol	Roth, Karlsruhe
Isopropanol	AppliChem, Darmstadt
Klenow buffer	MBI Fermentas GmbH, St. Leon-Rot
λ-DNA	MBI Fermentas GmbH, St. Leon-Rot
Lysozyme	Serva Feinchemikalien, Heidelberg
Magnesium chloride	Merck, Darmstadt
Magnesium sulfate	Merck, Darmstadt
Maleic acid	Roth, Karlsruhe
Methanol	Roth, Karlsruhe
Microscope slides	Menzel Gläser®, Walter, Kiel
Nucleotides	MBI Fermentas GmbH, St. Leon-Rot
Paraformaldehyde	Sigma-Aldrich Chemie GmbH, Taufkirchen
PCR buffer	Promega Corporation, Madison, USA

Name	Company/Supplier
PEG 6000	Roth, Karlsruhe
Phenol	Biomol Feinchemikalien GmbH, Hamburg
Polyvinylpyrrolidone	Fluka Chemie GmbH, Buchs, Schweiz
Potassium acetate	Merck, Darmstadt
Potassium dihydrogen phosphate	Merck, Darmstadt
Potassium hydrogen phosphate	Merck, Darmstadt
Proteinase K	Roth, Karlsruhe
Random hexamer primer	MBI Fermentas GmbH, St. Leon-Rot
Rotiphorese® NF-acrylamide/solution 0.4	Roth, Karlsruhe
Rotiphorese® NF-urea	Roth, Karlsruhe
Saccharose	Roth, Karlsruhe
Salmon sperm DNA	Roth, Karlsruhe
Sephadex G50	Amersham Biosciences, Freiburg
Sodium acetate	Roth, Karlsruhe
Sodium citrate	Roth, Karlsruhe
Sodium chloride	Roth, Karlsruhe
Sodium dodecyl sulfate	Roth, Karlsruhe
Sodium dihydrogenphosphate	Roth, Karlsruhe
Sodium hydroxide	Roth, Karlsruhe
Streptavidin-Cy3	Sigma-Aldrich Chemie GmbH, Steinheim
Sulfuric acid	Merck,Darmstadt
T4 DNA ligase10x buffer	Amersham Biosciences, Freiburg
TEMED	AppliChem,Darmstadt
Tris	Roth, Karlsruhe
Tris-acetate	Roth, Karlsruhe
Triton X100	Roth, Karlsruhe
Tryptone/Pepton	Roth, Karlsruhe
Tween 20	Sigma-Aldrich Chemie GmbH, Steinheim
Urea	Roth, Karlsruhe
X-Gal	MBI Fermentas GmbH, St. Leon-Rot
X-ray developer and fixer Adefo	MS Laborgeräte,Wiesloh
Xylene cyanole	Roth, Karlsruhe
Yeast extract	Roth, Karlsruhe

Table 2.3: Kits and their application

Name	Application	Company/Supplier
CEQ™ DTCS Quick Start Kit	Sequencing	Beckman Coulter, Fullerton USA
GeneJET™ Plasmid Miniprep Kit	Plasmid purification	MBI Fermentas GmbH, St. Leon-Roth
Nucleo Bond® Xtra Maxi Kit	BAC purification	Macherey-Nagel GmbH & Co. KG, Düren
pGEM®-T Vector System	Cloning of PCR fragments	Promega Corporation, Madison, USA
Invisorb Gel Extraction Kit	Purification of PCR-derived amplicons from agarose gels	InviTek GmbH, Berlin

Table 2.4: Enzymes and antibodies

Name	Company/Supplier
Cellulase (*Aspergillus niger*)	Sigma-Aldrich Chemie GmbH, Steinheim
Cellulase (Onozuka R-10)	Serva Feinchemikalien, Heidelberg
Cy3-α-streptavidin antibody	Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim
Cytohelicase (*Helix pomatia*)	Sigma-Aldrich Chemie GmbH, Steinheim
FITC-α- digoxigenin antibody	Sigma-Aldrich Chemie GmbH, Steinheim

Name	Company/Supplier
GoTaq **DNA polymerase**	Promega Corporation, Madison, USA
Klenow fragment	MBI Fermentas GmbH, St. Leon-Rot
Lysozym	Serva Feinchemikalien, Heidelberg
Pectolyase (*Aspergillus japonicus*)	Fluka Chemie GmbH, Buchs, Schweiz
Pectinase (*Aspergillus niger*)	Sigma-Aldrich Chemie GmbH, Steinheim
Pepsin	Sigma-Aldrich Chemie GmbH, Steinheim
Plasmid safe™ ATP dependent DNase	Epicentre Technologies, Madison, USA
Ribonuclease A	AppliChem, Darmstadt
Shrimp alkaline phosphatase **(SAP)**	Epicentre Technologies, Madison, USA
Taq **DNA polymerase**	MBI Fermentas GmbH, St. Leon-Rot
T4 DNA ligase	MBI Fermentas GmbH, St. Leon-Rot

Table 2.5: Plasmids and bacterial host strains

Name	Resistance	Application	Company/Supplier
pBeloBAC11	Chloramphicol	Cloning of large restriction fragments (BAC library)	Woo *et al.*, 1994
pGEM-T	Ampicillin	Cloning of PCR fragments	Promega Corporation, Madison, USA
pUC18	Ampicillin	Cloning of restriction fragments	Roche Diagnostics GmbH, Mannheim
E. coli **XL-1 Blue**	Tetracycline	Amplification of single copy plasmids	Stratagene, La Jolla, USA

2.1.6 Oligonucleotides

The oligonucleotides in Table 2.6 were used for PCR analysis, sequencing, and probe generation.

Table 2.6: Oligonucleotides for PCR, sequencing and probe generation

Primer	Sequence (5' → 3')[a]	Melting temperature [°C]
Amplification of LINE fragments from genomic DNA		
BNR-5'[b]	AAR CNT TYG AYA G	36
BNR-3'[b]	GCR TCR TCN GCR TA	44
BNR RTcons for	GGC STT TGA YAG YRT CTC NTG G	55
BNR RTcons rev	CCA YCC AWA NWT CAT GCC AAA A	56
RTE02 for	CAY ACN ATG AAR MTN TGG GA	49-59
RTE02 rev	TCN GCR AAN ARC ATR CAC CAN GG	57-67
BNR1 sequencing primer		
BNR LINE walk01	TTT TGT TCA ATT TCA TTA CCC	50
BNR LINE walk02	GGA AAG TGA TTT TGC TGC TGC C	60
BNR LINE walk03	GGG ACC TGA TGA TGA CGA AA	57
BNR LINE walk04	CCA GTT CCT CCA TGT CAT CAT C	60
BNR LINE walk05	TAG GCG CTA GAA TTA AGC GC	58
BNR LINE walk06	TCC AAT GTG AAA GAT TTC ACT GAG	57
BNR LINE walk07	GGA TTA GAT TAC AGA TAC CTC GCC	61
BNR LINE walk08	GAA TAG CCT TCT TCA TGG TTG AAG	59
BNR LINE walk09	GAC CTT CAT GGA CAA GTT GAG ATT	59
BNR LINE walk10	ATA GCA GTT GAG ATG CCA TGT GAT	59

Primer	Sequence (5’ → 3’)[a]	Melting temperature [°C]
BNR LINE walk11	GCA TAT TCT CAC TGA TCC TTT CTC	59
BNR LINE walk12	TTG TGG ACC ATT TGG AAG GAA AG	59
BNR LINE walk13	ATC ATA TAA TCG CCG AAG TGC TA	57
BNR LINE walk14	TCT TCG GGT GTG AGT ATT CAC TA	59
Gap closure of BNR2		
BvBAC132gap for	CCG AAA GAG AAC ATG CCC	56
BvBAC132gap rev	GCT GGT TGA ACT TTT AGG AC	55
Generation of LINE probes for hybridization		
BNR 5’Ende for	TCC AAA GCC AAC GCC AAT TCT	58
BNR 5’Ende rev	TTC AGA TGA TTG CGA AGA CGG	58
BNR RT2 for	CGA TAA GTA TGA TAG GGT GTG	56
BNR RT2 rev	GAG AGT CCA TTC CAT AAA TCC	56
BNR RNaseH for	GTG GTG GTT GAG TAT TTG GGG A	60
BNR RNaseH rev	TCC ATA GGA GAC ATT GCG GGT T	60
Belle1_1 RTfor	CCG GTT GAA AGA TTT CCT CCC	60
Belle1_1 Rtrev	TCT AGC GGG AGT AAC AGA ACC	60
Belle2_8 RTfor	TTA ATC GCC TCA AAC CAA TTC	54
Belle2_8 RTrev	CGA GTT GGA CAA ATT TTT TCC	54
Belle3_2 Rtfor	CGG CTA AAA CCA ATA CTA AAG	54
Belle3_2 RTrev	ACG ATG GTT TGA ATC TTG TAG	54
Belle4_2 RTfor	ACA GAT TGA GCA CTA TTT TAC C	55
Belle4_2 RTrev	TTG TGA AGT GAA ATG CCC ATG	56
Belle5_2 RTfor	TAA CAA AGA GAT TGC AGG CAG	56
Belle5_2 RTrev	TCT GGG CAT CAA AAG GAA TAG	56
BvL19 RTfor	CAC CAA TCG CCT AAA AAT CAC	56
BvL19 RTrev	TAG AAG GGA AGA ACC TTT CAC	56
BNR18 RTfor	GAA TTC AAA GGG TAA TGA GTT CCC	59
BNR18 RTrev	CCA CGA TGA AGT TTG AAG GGG	60
Generation of LTR retrotransposon probes for hybridization		
Cotzilla LTR for	AGT CAT GCC TAG ATT ATA GG	53
Cotzilla LTR rev	CTA TTA TGA GAG AAT GAA GGC	54
Cotzilla RT for	CAA ATG GAT GTT AAG TGT GC	53
Cotzilla RT rev	TCA ACA TAT ATT TGA ACA AGC	50
Cotzilla env for	GAA CCG AAA CCT AAG AGG C	57
Cotzilla env rev	TAC CCT TGG AAC TAG AGG C	57
Standard and satellite primers		
Epi M13 for	CGC CAG GGT TTT CCC AGT CAC GAC	68
Epi M13 for	AGC GGA TAA CAA TTT CAC ACA GGA	59
pAv34 f1[c]	GAA TTG TTG AAA TCT TAA GAA AAA TGG	56
pAv34 r1[c]	CGG AGT TAG TGA ACC GGG	58

[a] N = A,C,G,T; W = A,T ; S = G,C; Y = T,C; R = A,G; M = A,C

[b] Schmidt *et al.*, 1995

[c] Dechyeva and Schmidt, 2006

2.1.7 DNA probes

DNA probes were generated by PCR to enable hybridization experiments for analysis of abundance and distribution of repeats. Probes used in this work are shown in Table 2.7.

Table 2.7: Repeat-specific DNA probes

Probe	Length [bp]	Target sequence
BNR1_5end	204	LINE BNR1, RRM region in ORF1
BNR1_RT	254	LINE BNR1, reverse transcriptase region in ORF2
BNR1_RNaseH	282	LINE BNR1, RNaseH region in ORF2
Belline1_19 (BNR19)	316	LINE Belline1_19, RT region
Belline2_4	324	LINE Belline2_4, RT region
Belline5_2	317	LINE Belline5_2, RT region
Belline7_18 (BvL18)	326	LINE Belline7_18, RT region
Belline9_5	325	LINE Belline9_5, RT region
Belline12_2	322	LINE Belline12_2, RT region
Belline17_1	325	LINE Belline17_1, RT region
Cotzilla_LTR	496	Ty1-*copia* retrotransposon Cotzilla, LTR region
Cotzilla_RT	269	Ty1-*copia* retrotransposon Cotzilla, RT region
Cotzilla_env	460	Ty1-*copia* retrotransposon Cotzilla, env region
pTa71[a]	4642	18S-5.8S-25S rRNA genes

[a] Dechyeva and Schmidt, 2009

2.1.8 Sequence databases

During the course of this work several *B. vulgaris* sequence databases became available (Table 2.8). Except for the *c₀t-1* library made up of repetitive DNA clones and the small RNA database containing short RNA, all sugar beet sequence databases contain genomic DNA. The *RefBeet* datasets are contig assemblies of Illumina- and 454-generated sequences of *B. vulgaris* ssp. *vulgaris* KWS 2320.

In order to compare *B. vulgaris* repeats with those of other organisms, the corresponding whole genome sequence databases have been analyzed (Table 2.9).

Table 2.8: *B. vulgaris* sequence databases

Database	Size [Mb]	No. of contigs/ sequences	Reference
BAC end library	121	117,496	McGrath *et al.*, 2004
***C₀t-1* library**	0.45	1763	Zakrzewski *et al.*, 2010
Fosmid end library	33	45,296	Lange *et al.*, 2008
***RefBeet* 0.1.1**	628	340,529	Weisshaar & Himmelbauer *et al.*, personal comm.
***RefBeet* 0.2**	895	906,894	Weisshaar & Himmelbauer *et al.*, personal comm.
***RefBeet* 0.4**	1006	853,494	Weisshaar & Himmelbauer *et al.*, personal comm.
Chromosome 9 partial BAC sequences	0.67	80	Schulte *et al.*, 2006
Small RNA	159	6,762,678	Himmelbauer *et al.*, personal comm.

Table 2.9: Plant and animal genome sequence databases used in this work

Species	Sequence database	Size [Mb]	No. of contigs	Reference	Downloaded from
Plant genomes					
Arabidopsis thaliana	TAIR8	119.7	7	Swarbreck *et al.*, 2008	www.arabidopsis.org/download/index.jsp
Brachypodium distachyon	v1.0	271.9	83	Vogel *et al.*, 2010	www.phytozome.org
Glycine max	v1.01	973.3	1168	Schmutz *et al.*, 2010	www.phytozome.org
Malus x domestica	v1.0	881.3	122,107	Velasco *et al.*, 2010	www.rosaceae.org/projects/apple_genome
Mimulus guttatus	v1.0	321.7	2216	Mimulus Genome Project, DoE Joint Genome Institute	www.phytozome.org
Oryza sativa	MSU Release 6.0	373.7	14	Ouyang *et al.*, 2007	www.phytozome.org
Populus trichocarpa	v1.0, June 2004	485.5	22,012	Tuskan *et al.*, 2006	genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html
Solanum lycopersicum	v2.40	781.7	13	The international tomato genome sequencing consortium	http://mips.helmholtz-muenchen.de/plant/tomato/download
Solanum tuberosum	*S. phureja* DM1-3 516R44 v3.0	717.5	9171	Potato genome sequencing consortium	http://potatogenomics.plantbiology.msu.edu
Theobroma cacao	v1.0	291.4	25,912	Argout *et al.*, 2011	cocoagendb.cirad.fr/gbrowse/download.html
Vitis vinifera	12X, March 2010	486.2	33	Jaillon *et al.*, 2007	www.phytozome.org
Zea mays	B73 v2.0 RefGen	2058.8	10	Schnable *et al.*, 2009	ftp.maizesequence.org/current
Human and animal genomes					
Bombyx mori	v2.0	480.8	43,622	Xia *et al.*, 2004	silkworm.genomics.org.cn
Danio rerio	v9.60	1412.5	1133	Danio rerio Sequencing Project	www.ensembl.org/info/data/ftp/index.html
Drosophila melanogaster	v5.32	159.4	9	Adams *et al.*, 2000	flybase.org
Homo sapiens	GRCh37	3095.7	24	Lander *et al.*, 2001	www.ncbi.nlm.nih.gov/projects/genome/assembly/grc

2.1.9 Software

For the analysis of genomic sequences a multitude of bioinformatics tools were used as summarized in Table 2.10.

Table 2.10: Computational biology software

Software	Function	Reference	Website
***Adobe Photoshop* 7.0**	Editing of autoradiographs and microscopy images	---	www.adobe.com
Bioedit	Sequence storage and management	---	www.mbio.ncsu.edu/BioEdit/bioedit.html
BOXSHADE	Printing of muliple sequence alignments	---	www.ch.embnet.org/software/BOX_form.html
BLAST	Homology search (*BLAST* via *EBI* or *NCBI* website, local *BLAST* via *Bioedit*)	Altschul *et al.*, 1990	www.ebi.ac.uk/Tools/sss/ http://blast.ncbi.nlm.nih.gov/Blast.cgi
***Case Data Manager Expo* 4.5.0.28**	Analysis of images obtained by the fluorescence microscope	---	www.spectral-imaging.com
Circoletto	Circular visulization of sequence similarity	Darzentas, 2010	http://tools.bat.ina.certh.gr/circoletto
DNA Block Aligner	Annotation of target site duplications	---	www.ebi.ac.uk/Tools/Wise2/dbaform.html
***DNAStar* (*Seqman*)**	Sequence assembly, Sequence editing	---	www.dnastar.com
FASTA	Homology search (*FASTA* via *EBI* website, local *FASTA* as standalone version)	Pearson, 1990	www.ebi.ac.uk/Tools/sss
***Format Converter* (*HCV Tools*)**	Sequence format conversion (e.g. *FASTA* to Stockholm)	---	hcv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html
Geneious	Sequence assembly, phylogenetic analyses	Drummond *et al.*, 2011	www.geneious.com
GeneWise	ORF annotation	Birney *et al.*, 2004	http://www.ebi.ac.uk/Tools/Wise2/index.html
HHPred	Prediction of amino acid motifs	Söding *et al.*, 2005	toolkit.tuebingen.mpg.de/hhpred
***HMMER*3**	Hidden Markov Model-based sequence search	Eddy, 1998	hmmer.janelia.org
Interpro	Prediction of amino acid motifs	Zdobnov and Apweiler, 2001	www.ebi.ac.uk/Tools/pfa/iprscan
***JPred*3**	Prediction of amino acid secondary structure	Cole *et al.*, 2008	www.compbio.dundee.ac.uk/www-jpred
LTR Finder	Prediction of complete LTR retrotransposons	Xu and Wang, 2007	tlife.fudan.edu.cn/ltr_finder
***MEGA*4**	Phylogenetic analyses, visualization of multiple sequence alignments	Tamura *et al.*, 2007	www.megasoftware.net
***Molecular Toolkit* (Translate)**	Visualization of ORFs	---	arbl.cvmbs.colostate.edu/molkit/translate/index.html
MUSCLE	Multiple sequence alignments (via EBI website, standalone version for more than 500 sequences)	Edgar, 2004	www.ebi.ac.uk/Tools/msa/muscle
MyHits Protein Motif Scan	Prediction of amino acid motifs	Pagni *et al.*, 2004	myhits.isb-sib.ch/cgi-bin/motif_scan
OligoAnalyzer	Analysis of primers (melting temperature, hairpins, dimers)	---	eu.idtdna.com/analyzer/Applications/OligoAnalyzer
PCOILS	Coiled-coil domain prediction	Lupas, 1996	http://toolkit.tuebingen.mpg.de/pcoils
PLACE	Promotor motif prediction	Higo *et al.*,	http://www.dna.affrc.go.jp/PLA

Software	Function	Reference	Website
		1999	CE/index.html
PlantCARE	Promotor motif prediction	Lescot *et al.*, 2002	http://bioinformatics.psb.ugent.be/webtools/plantcare/html/
***Python* 2.7.1**	General Programming Usage of several libraries: *Numpy* 1.5.1 (array objects) *Biopython* 1.56 (biological computation) *Matplotlib* 1.0.1 (2D plotting)	Cock *et al.*, 2009 ---	www.python.org numpy.scipy.org biopython.org matplotlib.sourceforge.net
RepeatMasker	Annotation of known repetitive elements	Smit *et al.*, 2008	www.repeatmasker.org
***RTclass*1**	Non-LTR retrotransposon assignment to known LINE clades	Kapitonov *et al.*, 2009	www.girinst.org/RTphylogeny/RTclass1
SERV	Identification of tandem repeats	Legendre *et al.*, 2007	www.igs.cnrs-mrs.fr/SERV

2.1.10 Script programs

In order to facilitate work with sequence databases, a number of script programs have been developed (Table 2.11). These were written in the script language *Python* using the modules *Numpy*, *Biopython* and *Matplotlib*. The program files and documentation are included on the supplemental CD-Rom.

Table 2.11: Python programs

Category	Script name	Function
BLAST parsing	*LocalBlastBioedit-tBLASTn.py*	Retrieves sequences with a defined e-value/identity that correspond to a *Bioedit* local *BLAST* output (exact matches)
	LocalBlastBioedit-tBLASTn-completeseq.py	Retrieves sequences with a defined e-value/identity that correspond to a *Bioedit* local *BLAST* output (complete sequence)
	LocalBlastBioedit-tBLASTn-cut.py	Retrieves sequences with a defined e-value/identity that correspond to a *Bioedit* local *BLAST* output (exact matches plus/minus an additional number of nucleotides)
	NCBI-tBLASTn-parser.py	Retrieves sequences with a defined e-value from *NCBI tBLASTn* Output (XML-Format)
HMMER parsing	*HMMER-Parse.py*	Retrieves sequences with a defined bit score from the translated database using the *HMMER* machine-readable tabular output (*domtblout* option)
	get-nt-seq-from-HMMER-parse.py	Retrieves sequences using a list of sequence names (derived of *HMMER-Parse.py* output; exact match)
	get-nt-seq-from-HMMER-parse_completeseq.py	Retrieves sequences using a list of sequence names (derived of *HMMER-Parse.py* output; complete sequence)
	get-nt-seq-from-HMMER-parse_crop.py	Retrieves sequences using a list of sequence names (derived of *HMMER-Parse.py* output; plus/minus an additional number of nucleotides)
Sequence retrieving	*Sequence-Retrieve-from-EBI.py*	Retrieves a list of sequences from EBI (using accession numbers)

Category	Script name	Function
	Sequence-Retrieve-from-fasta.py	Retrieves a list of sequences from a fasta file
Plotting	*Plot-small-RNA-Alignment_v1_allRNAs*	Visualizes of a small RNA assembly to a target sequence (input: Geneious assembly in fasta format)
Translating	*Translate-from-FASTA.py*	Translates a nucleotide sequence database in all six frames
Various	*Intersection_2Sets.py*	Detects duplicates in two lists of names
	PartitionSequence.py	Partitiones a sequence
	RandomSeq.py	Shuffles a sequence database

2.1.11 Retrotransposon references and accessions

Sequence data from this thesis can be found in the EMBL/GenBank as indicated in Table 2.12. Additionally, LINE sequences and their accessions are listed in Table 3.1 (BNR/Belline1 family in *B. vulgaris*), Table 3.2 (BNR-like LINEs in a number of angiosperm genomes) and Table 3.3 (reference members of 17 Belline families).

Table 2.12: Accession numbers of retrotransposon sequences

Order	Superfamily/ Clade	Name	Organism	Accession
LINE	I	I dm	*Drosophila melanogaster*	M14954
	Jockey	Jockey dm	*Drosophila melanogaster*	M22874
	L1	BLIN	*Hordeum vulgare*	AJ270056
		BNR1-1	*Beta vulgaris*	Z38073
		BNR1-2	*Beta vulgaris*	Z38074
		BNR1-6	*Beta vulgaris*	Z38075
		BvL1	*Beta vulgaris*	FM993986
		BvL2	*Beta vulgaris*	DQ374076 (6602-4165 nt) & DQ374077 (4234-1 nt)
		BvL3	*Beta vulgaris*	DQ374017 (14920-24242 nt)
		BvLi3	*Beta vulgaris*	Y13368
		cin4	*Zea mays*	Y00086
		del2	*Lilium speciosum*	Z17425
		Karma	*Oryza sativa*	AB081316
		L1 hs	*Homo sapiens*	U93574
		L1 rat	*Rattus norvegicus*	U83119
		L1 dog	*Canis familiaris*	AB012223
		LIb	*Ipomoea batatas*	AB231839
		LINE-CS	*Cannabis sativa*	AB013908
		Swimmermed	*Oryzias latipes*	AF055640
		Swimmerpup	*Cyprinodon macularius*	AF055643
		Ta11-1	*Arabidopsis thaliana*	L47193
		Zepp	*Chlorella vulgaris*	D89938
	R2	R2 dm	*Drosophila melanogaster*	X51967
	RTE	*Ghost*1	*Beta vulgaris*	FR852837
		RTE1	*Caenorhabditis elegans*	AF025462
		RTE1 zm	*Zea mays*	Kapitonov *et al.*, 2009

Order	Superfamily/ Clade	Name	Organism	Accession
Ty1-*copia* LTR retrotransposon	Ty1-*copia*	CIRE1	*Citrus sinensis*	AM040263
		Rire1	*Oryza australiensis*	D85597
		SALIRE1	*Beta vulgaris*	FN357199
		Tnt1	*Nicotiana tabacum*	X13777
		Tto1	*Nicotiana tabacum*	D83003
	env-containing Ty1-*copia* (*Sireviridae*)	Cotzilla1	*Beta vulgaris*	EF101866 (26271-37103 nt)
		Cotzilla3	*Beta vulgaris*	DQ374087 (71870-81547 nt)
		Hopie	*Zea mays*	AC116033
		Opie2	*Zea mays*	AF090446
		PREM-2	*Zea mays*	U41000
		*SIRE*1	*Glycine max*	U96295
Ty3-*gypsy* LTR retrotransposon	Ty3-*gypsy*	*Beetle*1	*Beta procumbens*	AJ539424
		*Beetle*2	*Beta procumbens*	FM242082
	env-containing *Ty3-gypsy (Errantiviridae)*	*Athila4-2*	*Arabidopsis thaliana*	AB026642
		Calypso	*Glycine max*	AF378070
		Cyclops-2	*Glycine max*	AF186182-86
Retrovirus		*Osvaldo*	*Drosophila buzzatii*	AJ133521

2.2 Molecular techniques

2.2.1 Isolation of DNA

2.2.1.1 Isolation of plant DNA

Plant genomic DNA was isolated from young leaves using the CTAB extraction method (Saghai-Maroof *et al.*, 1984, modified). The detergent CTAB serves to destabilize cell walls and to separate DNA from proteins and lipids. The chelating agent EDTA binds Mg^{2+} and thus inhibits nucleases and protects the DNA from degradation. Removal of proteins and RNA was achieved by phenol-chloroform extraction and RNase treatment, respectively. The method was completed with an isopropanol precipitation and washing steps in order to eliminate salt remnants and solvent carryovers.

- Overnight lyophilization of 3-5 g leaf material in a vaccum chamber at -60 °C and 0.2 mbar; storage at -20 °C in 50 ml tubes
- Pulverization of freeze-dried leafs with ceramic beads
- Addition of 12.5 ml CTAB buffer (including ß-mercaptoethanol) and incubation for 30 min at 65 °C
- Addition of 1 volume phenol-chloroform-isoamylalcohol (24:24:1), careful vortexing, and incubation for 10 min in an overhead mixer
- Centrifugation for 30 min at 3200 g
- Transfer of the upper phase to a new 50 ml tube; addition of 1 volume chloroform-isoamylalcohol (24:1), vortexing, and incubation for 10 min in an overhead mixer
- Centrifugation for 30 min at 3200 g
- Transfer of the upper phase to a new 50 ml tube, addition of 50 µl RNase A (10 mg/ml) and incubation for at least 1 h at 37 °C
- Incubation on ice for at least 5 min
- Mixing with 0.7 volumes cold isopropanol leads to precipitation of DNA
- Transfer of DNA to a new 2 ml tube containing 76 % ethanol
- 2 washing steps with 76 % ethanol
- After air-drying of DNA, addition of 100 to 500 µl TE buffer
- Overnight resuspension at 16 °C

2.2.1.2 Isolation of bacterial plasmid DNA

High copy number plasmids pGEM-T and pUC18 were isolated from bacterial culture using the GeneJET™ Plasmid Miniprep Kit (Fermentas) according to the manufacturer's instructions. The method is based on the alkaline lysis of bacterial cells followed by protein precipitation. Subsequently, plasmid DNA is bound to a nitrocellulose or glass fiber matrix and washed with ethanol-containing buffer. Plasmids were eluted using 50 µl water with a yield of approximately 5 to 15 µg.

2.2.1.3 Isolation of BAC DNA

For preparation of BAC DNA for sequencing purposes, the NucleoBond® Xtra Kit (Macherey-Nagel) was utilized as described by the manufacturer. 4 ml of starter culture (LB with chloramphenicol) were grown for 8 h at 37 °C and 220 rpm. It was diluted 1:1000 into the desired final volume of LB with chloramphenicol and grown overnight under the same conditions until an OD_{600} of approximately 400 was reached.

The BAC DNA pellet was resuspended using an appropriate volume of water.

2.2.2 Agarose gel electrophoresis

In order to separate DNA fragments according to their size, horizontal agarose gel electrophoresis was performed. DNA probes were mixed with 10x loading buffer and loaded on an agarose gel and an electric field was applied. Because of their negative charge, DNA molecules migrate to the positive pole with a size-dependent velocity: Small molecules are able to migrate faster than large ones. Separation took place in 1x TAE at 1-5 V/cm. The gel concentration was varied between 0.6 and 1.2 % agarose according to the size of the expected DNA fragments. In order to visualize the DNA bands, ethidium bromide was added to the agarose gel to a final concentration of 0.005 %. Images were captured with the GelDoc2000 system.

2.2.3 Polymerase chain reaction

A standard polymerase chain reaction (PCR) allows the amplification of short genomic regions with a size of up to 3 kb. Long range methods enable the generation of even longer products. A typical PCR reaction is described below.

PCR reaction

DNA		50	ng
GoTaq buffer (5x)		4	µl
dNTPs (2 mM)		2	µl
Forward primer (10 µM)		1	µl
Reverse primer (10 µM)		1	µl
GoTaq DNA polymerase (5 U/µl)		0.2	µl
H_2O	ad	20	µl

PCR program

Initial denaturation	94 °C	5 min	
Denaturation	94 °C	1 min	35 cycles
Annealing	dependent on primer	30 sec	
Elongation	72 °C	1 min/1000 bp	
Final elongation	72 °C	5 min	
Hold	04 °C	∞	

Annealing temperature and elongation duration have been chosen according to the primer's base composition and length of the expected product, respectively.

2.2.4 Molecular cloning

2.2.4.1 Restriction of plasmid and genomic DNA

Bacterial type II restriction enzymes recognize specific nucleotide motifs and are able to cut the DNA at these sites. 5 U enzyme/µg DNA, restriction buffer, DNA and water were mixed and incubated for 1 to 8 h at the temperature specified by the manufacturer.

2.2.4.2 Dephosphorylization of plasmid vectors

In order to avoid religation of linearized plasmid molecules, a dephosphorylization with alkaline phosphatase was performed according to the manufacturer's instructions. Subsequently, the DNA was purified by ethanol precipitation according to Sambrook *et al.* (1989), followed by a separation on an agarose gel (2.2.2). The corresponding plasmid DNA band was cut from the gel and purified as described in 2.2.4.3.

2.2.4.3 Elution of DNA fragments from agarose gels

After PCR amplification (2.2.3) or restriction (2.2.4.1), DNA fragments have been gel-purified before cloning. DNA products have been separated on an agarose gel (2.2.2) and cut out using a scalpel. In order to extract the fragments of the desired size from the agarose gel, the Invisorb Gel Extraction Kit (InviTek GmbH) has been utilized according to the manufacturer's instructions.

2.2.4.4 Ligation of DNA

After gel-purification (2.2.4.3), PCR-generated DNA fragments have been integrated into pGEM-T vector according to the instructions of the manufacturer. Restricted DNA fragments were ligated into pUC18 vector. In order to accommodate the DNA insert, pUC18 was pretreated with the corresponding restriction enzyme (2.2.4.1; 2.2.4.2). A 3:1

molar ratio of insert to vector was used. The reaction was incubated for 1 h at 37 °C, followed by an overnight exposure to 16 °C.

Ligation to pUC18:

DNA		x	µl
Vector (50 ng/µl)		1	µl
Ligation buffer (10x)		2	µl
ATP (5 mM)		4	µl
T4-Ligase		2	U
H_2O	ad	20	µl

2.2.4.5 Plasmid transformation

Recombinant plasmids have been amplified in *E. coli* cells. Competent *E. coli* cells were transformed with the insert-carrying vectors by electroporation.

- Thawing of 50 µl of frozen electrocompetent cells on ice
- Addition of 1 µl ligation reaction and transfer to a 0.2 cm elecroporation cuvette (Gene Pulser®, BioRad)
- Electroporation with the EasyjecT Prima (Equibio) at 2.5 kV
- Immediate incubation in 1 ml preheated SOC medium
- Recovering of the cells for 1 h at 37 °C and shaking at 300 rpm
- Plating of 50 to 500 µl of the cell suspension to indicator plates including the corresponding antibiotic
- Overnight incubation at 37 °C

2.2.5 DNA sequencing

DNA was sequenced with the automated capillary electrophoresis system CEQ 8000 (Beckman Coulter), which utilizes the dideoxy method invented by Sanger *et al.* (1977). The capillaries have been filled with polyacrylamide and the samples were processed automatically according to the applied program.

Cycle sequencing PCR reactions have been carried out using fluorescent dyes and other components from the CEQ™ DTCS Quick Starter Kit according to the manufacturer's instruction. Additives like betain have been used to enhance the quality of the sequencing result.

2.2.5.1 Plasmid sequencing

A typical sequencing reaction of a plasmid insert with Epi M13 primers is detailed below.

Cycle sequencing reaction (20 µl)

DNA	according to the instructions of the manufacturer
H_2O	ad 12 µl

Denaturation: 5 min at 94 °C

Premix	4 µl
Sequencing Reaction Buffer	2 µl
Betain (5 M)	1 µl
Primer EpiM13 (10 µM)	1 µl

Cycle sequencing program

Initial Denaturation	94 °C	2:00	min	
Denaturation	94 °C	0:20	min	30 cycles
Annealing	58 °C	0:20	min	
Elongation	60 °C	4:00	min	
Hold	4 °C	∞		

CEQ 8000 sequencing program

Capillary	50	°C	
Denaturation	90	°C	120 sec
Injection	2	kV	20 sec
Separation	4.2	kV	150 min

2.2.5.2 BAC sequencing

For sequencing of BAC DNA with internal primers, it is important that the oligonucleotides have a minimum length of 22 nt, and a melting temperature above 55 °C. Furthermore, they should end with a G or C. For BACs with unknown inserts, it is recommended to test the amount of primer, the number of cycles, and the sequencer settings. A typical reaction setup is described below.

Cycle sequencing reaction (20 µl)

DNA	according to the instructions of the manufacturer
H_2O	ad 12 µl

Denaturation: 5 min at 94 °C

Premix	4 µl
Sequencing Reaction Buffer	2 µl
Betain (5 M)	1 µl
Primer EpiM13 (10 µM)	2 µl

Cycle sequencing program

Initial Denaturation	94 °C	2:00	min	
Denaturation	94 °C	0:20	min	50 cycles
Annealing	dependent on primer	0:20	min	
Elongation	60 °C	2:00	min	
Hold	4 °C	∞		

CEQ 8000 sequencing program

Capillary	55	°C	
Denaturation	90	°C	120 sec
Injection	2	kV	15 sec
Separation	3	kV	180 min

2.2.6 Southern hybridization

Southern hybridization is based on the ability of a DNA probe to specifically bind to a DNA target. Target DNA was prepared by separation in an 1.2 % agarose gel (2.2.2), followed by transfer onto a positively charged nylon membrane (2.2.6.1). Using a radioactively labeled probe (2.2.6.3), it was possible to identify complimentary regions using a hybridization assay (2.2.6.4). This method allowed quantitative analysis of genomic regions similar to the probe utilized.

2.2.6.1 Southern transfer of DNA from agarose gels

After restriction and electrophoretic separation, the target DNA was conveyed to a positively charged nylon membrane using alkaline capillary transfer. This transfer was carried out overnight according to Sambrook *et al.* (1989), using a denaturing solution of 0.5 M NaOH/ 1.5 M NaCl. Then, the membrane was washed in 2x SSC for 5 min and fixed at 80 °C for 2 h.

2.2.6.2 Southern transfer of DNA from colony plates

Bacterial cultures harboring the desired plasmids have been transferred from agar plate to LB freezing medium in 96 or 384 well plates. After overnight growth at 37 °C, they were stamped onto a positively charged nylon membrane placed on an LB agar plate. Subsequent to an overnight incubation at 37 °C for bacterial growth on the membrane, the plasmid DNA was treated by the following procedure.

- Placement of the membrane on Whatman paper
- Treatment with 0.5 M NaOH for 5 min, 1 M Tris/HCl, pH 7,5 / 3 M NaCl for 10 min, and 1 M Tris Base, pH 6.5 for 10 min

- Air-drying of membrane, followed by fixation for 20 min at 80 °C
- Washing in 2x SSC prior to hybridization

2.2.6.3 Labeling of DNA probes

DNA probes were labeled using random priming according to Feinberg and Vogelstein (1983). Radioactively labeled α-[^{32}P]-dATP und α-[^{32}P]-dCTP isotopes have been incorporated into the backbone of the DNA probe. Non-incorporated radioactive nucleotides were removed by gel filtration.

- Dilution of 100 ng of DNA in water to a total amout of 76 µl
- Denaturation for 5 min at 99 °C
- Addition of nucleotides, primers, buffer and enzyme
- Incubation for 1 h at 37 °C
- Gel filtration of the sample using a Sephadex G50 column (equilibrated in 1x TE buffer)

<u>Labeling reaction</u>

DNA template		100	ng
dGTP/dTTP mix (0.5 mM)		5	µl
Random primer (0.2 µg/µl)		5	µl
Klenow buffer (10x)		10	µl
Klenow fragment (2U/µl)		1	µl
α-[^{32}P]-dATP (3000 Ci/mmol)		1.5	µl
α-[^{32}P]-dCTP (3000 Ci/mmol)		1.5	µl
H_2O	ad	100	µl

2.2.6.4 Hybridization of DNA probes

In order to ensure specific binding of the DNA probe, unspecific binding sites have been saturated using salmon sperm DNA. After hybridization, the membrane was washed with a washing stringency of approximately 75 %. Factors influencing the stringency are salt concentration, temperature and G/C content of the probe.

- Incubation of the membrane in pre-hybridization medium for at least 2 h
- Transfer of the membrane into a hybridization tube
- Denaturation of labeled DNA probe for 10 min at 99 °C
- Mixture of the denaturated probe with 15 ml Denhardt medium, and addition to the hybridization tube
- Overnight hybridization at 60 °C

- Washing of membrane with 2x SSC / 0.1 % SDS for at least 15 min at 60 °C (three times)
- Wrapping of membrane in plastic foil and exposure of an X-ray film for a few hours up to several days at -80 °C
- Development of X-ray film

2.2.6.5 Removal of DNA probe from the membrane

The following procedure enables the removal of the DNA probe from the membrane by alkaline treatment, and thus facilitates the rehybridization of the membrane, if the membrane was kept wet.

- Shaking of the membrane in 0.2 M NaOH / 0.1 % SDS for 15 min to remove the probe
- Washing of membrane under running water for 10 min
- Shaking of membrane in 3 M NaCl / 0.5 M Tris Base, pH 7.0 for 20 min in order to rebuffer the membrane
- Washing in 2x SSC
- Drying of membrane for 30 min at 80 °C

2.2.7 Molecular cytogenetics

Fluorescent *in situ* hybridization (FISH) is a technique that uses a labeled complimentary DNA probe to localize specific DNA target sequences in tissues, cells, cell nuclei or on chromosomes. DNA probes tagged with fluorophores can be detected directly using a fluorescence microscope, while hapten-labeled probes need binding and detection of a second fluorochrome-tagged antibody. For the production of a double color FISH, differently labeled DNA probes were used in order to image two DNA regions of interest.

2.2.7.1 Fixation of plant chromosomes

In order to obtain mitotic chromosomes, it is important to capture plant cells in the process of division. Therefore, in this thesis, meristematic tissues of young leaves have been used for chromosome preparation.

- Collection of young leaves 4-5 h after dawn or illumination
- Incubation in 2 mM 8-hydroxyquinoline for 2-3 h and transfer to fixation solution
- Change of fixation solution every 30 min, until the leaves lose their color
- Storage of fixed plant leaves for up to several months at -20 °C

2.2.7.2 Preparation of plant chromosomes

Mitotic plant chromosomes have been prepared using a slightly modified variant of the dropping method described by Schwarzacher and Heslop-Harrison (2000). However, before the cell nuclei were applied to the slides, fixed leaves were treated enzymatically to cause degradation of the cell wall and the cytoplasm.

Glass slides were pretreated by incubation in chromic acid for 2 h and rinsed with running water for 20 min. An overnight drying at 37 °C followed. Before the slides were used for chromosome spreading, they had to be rinsed with 70 % ethanol.

Fixed leaves were treated enzymatically according to the following procedure:

- Washing in water for 2x 5 min
- Washing in enzyme buffer for 2x 5 min
- Transfer of leaf to a reaction tube with enzyme solution and overnight incubation at room temperature
- Incubation at 37 °C for 20-30 min
- Careful maceration with forceps, preparative needle and by pipetting
- Incubation for another 10-15 min at 37 °C
- Removal of undegraded debris
- Careful exchange of enzyme buffer without disturbance of the nuclei suspension in two washing steps by centrifugation for 5 min at 1200 g at room temperature
- Replacement of buffer with fixation solution without disturbance of the nuclei suspension, followed by two additional washing steps for 5 min at 1200 g at room temperature
- Last washing step with fixation solution for 6 min at 1500 g at room temperature
- Careful removal of supernatant without disturbance of the nuclei suspension, leaving only 100 µl suspension in the tube
- Addition of 50-100 µl of fixation solution to rinse the tube walls
- Dropping of 13 µl of the suspension onto pretreated glass slides from a height of 50 cm
- Spreading of chromosomes by blowing on the slide sharply, followed by rapid slide shaking
- Overnight incubation at 37 °C

The slides were examined using a phase-contrast microscope (Zeiss Axioscope 40) at a magnification of 10x and 40x. Chromosome spreads clear of cytoplasm have been selected for FISH.

2.2.7.3 Labelling of DNA probes for fluorescent *in situ* hybridization

DNA probes have been PCR-labeled using modified nucleotides that compete with their unmodified counterparts for incorporation in the DNA probe. Haptens like biotin or digoxygenin fused to the C5 atom of uridine have been used, to allow an immunological recognition by fluorochrome-labeled antibodies. The probes have been amplified from plasmids using internal or Epi M13 primers.

PCR was carried out in a 50 µl approach using the concentration specifications as described in 2.2.3. Either biotin-16-dUTP or digoxigenin-11-dUTP was added.

PCR reaction

DNA			20	ng
GoTaq buffer (5x)			10	µl
dNTPs (2 mM)			5	µl
Forward primer (10 µM)			2.5	µl
Reverse primer (10 µM)			2.5	µl
either	Biotin-16-dUTP (1 mM)		3.5	µl
or	Digoxigenin-11-dUTP (1 mM)		1.75	µl
GoTaq DNA polymerase (5 U/µl)			0.5	µl
H_2O		ad	50	µl

The PCR program was set as in 2.2.3. The labeled probe was purified by ethanol precipitation as described by Sambrook *et al.* (1989).

2.2.7.4 Fluorescent *in situ* hybridization

Fluorescent *in situ* hybridization (FISH) was performed according to Heslop-Harrison *et al.* (1991), modified by Schmidt *et al.* (1994) for *B. vulgaris.*

All washing steps have been performed in shaking coplin jars, while all incubation steps have been carried out in a moist chamber at 37 °C. Small volumes have been applied directly to the slide. After application, the slides needed to be protected from dessication by coverage with a plastic cover slip. Throughout the procedure, it is especially important that the slides, once wet, do not dry out.

Pretreatment of slides

- Washing in 2x SSC for 1 min at room temperature
- Addition of 200 µl RNase A (0,1 µg/µl RNase A in 2x SSC) solution, coverage with plastic cover slip and incubation for 1 h at 37 °C
- Washing in 2x SSC for 3x 5 min at room temperature
- Equilibration in 0.01 N HCl for 1 min at room temperature
- Addition of 200 µl pepsin (10 µg/ml, in 0,01 M HCl), coverage with plastic cover slip and incubation for 15 min at 37 °C
- Washing in 2x SSC for 3x 5 min at room temperature
- Incubation in 4 % paraformaldehyde for 15 min at room temperature
- Consecutive washing in 2x SSC for 3x 10 min, in 70 % ethanol for 3 min, and at in 100 % ethanol for 3 min at room temperature
- Air-drying at room temperature

Hybridization

All LINE probes have been hybridized to interphase and metaphase spreads using a stringency of 76 %. That was achieved by hybridization in a 50 % formamide, 2x SSC environment at 37 °C.

- Preparation of hybridization mixture which is detailed below, and preheating to 70 °C for 10 min
- Addition of 30 µl hybridization mixture to slides, coverage by plastic cover slips
- Denaturing and stepwise cooling using the denaturation program with the *in situ* thermocycler Touchdown (ThermoHybaid)
- Overnight incubation at 37 °C for hybridization

Hybridization mixture

DNA probe		0.5-2	µg
Formamide (100 %)		15	µl
Dextran sulfate (50 %)		6	µl
SSC (20x)		3	µl
SDS (10 %)		0.5	µl
Salmon sperm DNA (1 µg/µl)		1	µl
H_2O	ad	30	µl

Denaturation program

70 °C	8 min
55 °C	5 min
50 °C	2 min
45 °C	3 min
37 °C	10 min

Washing steps

For washing of all LINE probes, a washing stringency of 79 % was chosen.

- Removal of plastic cover slip in 2x SSC
- Stringent washing in formamide solution (20 % formamide in 0.1x SSC), twice at 42 °C
- Three subsequent washing in 2x SSC for 5 min, twice at 42 °C and once at 37 °C

Detection

- Washing in 4x SSC / 0,2 % Tween20 for 5 min at 37 °C
- Addition of 200 µl preheated blocking solution, coverage with plastic cover slip and incubation for 30 min at 37 °C
- Careful removal of plastic cover slip, addition of 50 µl antibody solution (200 µg/ml FITC-α-digoxigenin or 1 mg/ml Cy3-α-streptavidin in 4x SSC / 0,2 % Tween20), recoverage with plastic cover slip
- Incubation for 1 h at 37 °C
- Washing in 4x SSC / 0,2 % Tween20 for 3x 10 min at 37 °C
- Addition of 15 µl DAPI in CitiFluor AF1
- Careful enclosure using a glass cover slip
- Drainage of excess liquids using filter paper
- Examination of slides (2.2.7.6)
- Dry and cool storage at 4 °C

2.2.7.5 Rehybridization of chromosomes

In order to reuse a glass slide for FISH, the DNA probe was removed according to Schwarzacher and Heslop-Harrison (2000). If not indicated otherwise, all steps were performed at room temperature.

- Heating of slides to 60 °C, followed by careful removal of coverslip
- Consecutive washing in 2x SSC for 2x 5 min at 42 °C, in 4x SSC / 0,2 % Tween20 for 30 min, in 2x SSC for 3x 10 min
- Incubation in 4 % paraformaldehyde for 10 min
- Consecutive washing in 2x SSC for 3x 10 min, in 70 % ethanol for 3 min, and at in 100 % ethanol for 3 min
- Air-drying of slides

2.2.7.6 UV microscopy and digital image processing

After excitation with UV-light of a certain wavelength, fluorochromes are able to emit light which is slightly shifted towards longer wavelengths. It is possible to visualize and photograph the emitted light signals, if it passes suitable filters. Depending on the kind of fluorochrome, different filter sets have to be used (Table 2.13). Slides were examined using a Zeiss Axioplan2 *imaging* UV-fluorescence microscope equipped with a filter set as described in Table 2.14.

Table 2.13: Properties of the fluorochromes utilized

Fluorochrome	Color of fluorescence	Excitation [nm]	Emission [nm]	Filter set used for detection
DAPI	Blue	358	461	01
FITC	Green	495	523	09
Cy3	Red	550	570	15

Table 2.14: Properties of filter sets

Filter set	Excitation filter	Emission filter	Beam splitter	Suitable fluorochromes
01	BP 365/12	LP 397	FT 395	DAPI
09	BP 450-490	LP 515	FT 510	FITC
15	BP 546/12	LP 590	580	Cy3
25 (triple filter)	TBP 400+495+570	TBP 460+530+610	TFT 410+505+585	DAPI/FITC/Cy3

Images were taken with a Zeiss MC 80 DX camera using a magnification of 1600x. Subsequently, the pictures were analyzed and edited using the software *Case Data Manager Expo* 4.5.0.28 and *Adobe Photoshop* 7.0.

2.3 Computational methods

References for the applied software and tools can be found in Table 2.10. Additionally, Table 2.11 includes information about script programs for automatization of routine tasks.

2.3.1 Homology searches

2.3.1.1 Web-based *BLAST* searches

DNA homology searches were performed using the web-based *BLAST* search from *NCBI*. For identification of BNR1 ORF1 homologues, *tBLASTn* searches were performed. This method enables to search a nucleotide database with a protein query. The resulting hits were saved as *extensible markup language* (*XML*) file. Subsequently, the computer script *NCBI-tBLASTn-parser.py* to retrieve the homologous sequences listed in the XML file directly from the NCBI database.

2.3.1.2 Local *BLAST* searches

Local databases were queried using the *BLAST* option in *Bioedit*. In addition to the human readable output, a tabular output file was generated, which was analyzed by one of the *LocalBlastBioedit-tBLASTn.py* scripts. These procedures enabled a fast sequence retrieval of either the exact *BLAST* matches or a controlled output of flanking sequences.

2.3.2 Multiple sequence alignments and assemblies

In order to compare a multitude of DNA or protein sequences, a multiple sequence alignment was created using the *MUSCLE* algorithm. For more than 500 sequences, the *MUSCLE* standalone software was used. In case of very large alignments exhausting the main storage of the computer, a cruder alignment was produced with only two iterations (instead of a flexibly allocated number of repetitions).

For comparison and alignment of one sequence to a database of sequences (e.g. small RNAs), the *Geneious* assembler was applied.

2.3.3 Visualization of multiple sequence alignments

Comparative retrotransposon sequence analysis was conducted using the software *MEGA*4. Neighbor-Joining consensus trees (Saitou and Nei, 1987) were constructed using 1000 bootstrap replicates. The evolutionary distances were computed using the

Poisson correction method and all positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons. Alternatively, *Geneious* was applied to build Neighbor-Joining consensus trees, if branch-specific access to the underlying sequence data was needed. Dendrograms were exchanged between both programs using the *Newick* tree format. In case of very large alignment files, in order to shorten computational time, a calculation of bootstrap support was disregarded.

2.3.4 Hidden Markov Model (HMM)-based motif search for the genome-wide identification of retrotransposons

For the genome-wide detection of retrotransposon RT sequences, a Hidden Markov Model (HMM)-based approach was applied. A Hidden Markov Model is a statistical model of a multiple sequence alignment that takes into account the conservation of amino acids at a certain position as well as the probability of their neighboring amino acids. The software *HMMER*3 was used to build HMMs and query local databases, while computer scripts enabled controlled sequence extraction. A typical HMM workflow for identification, annotation and presentation of reverse transcriptase sequences is presented in Figure 2.1.

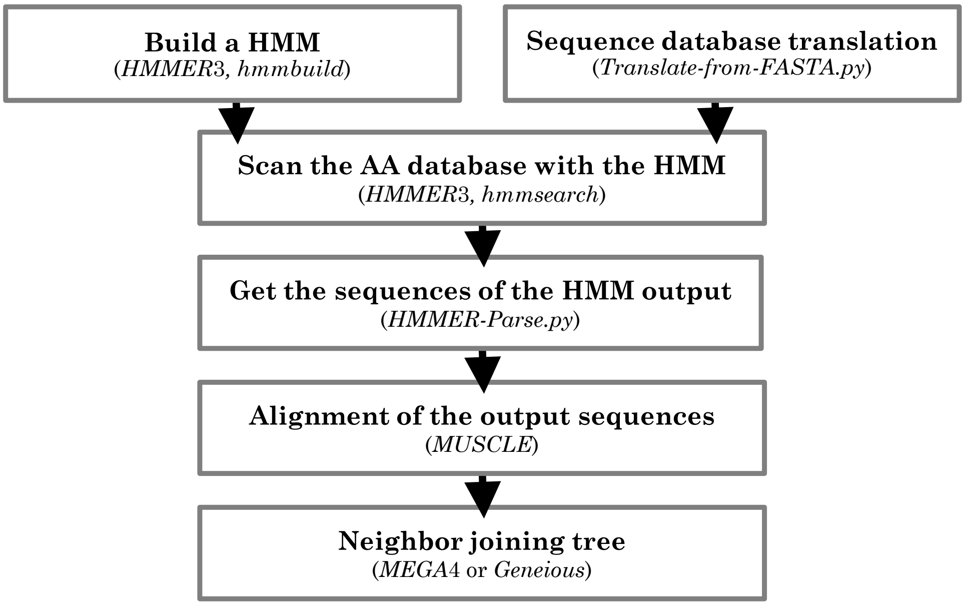

Build a HMM
(*HMMER*3, *hmmbuild*)
Sequence database translation
(*Translate-from-FASTA.py*)
Scan the AA database with the HMM
(*HMMER*3, *hmmsearch*)
Get the sequences of the HMM output
(*HMMER-Parse.py*)
Alignment of the output sequences
(*MUSCLE*)
Neighbor joining tree
(*MEGA*4 or *Geneious*)

Figure 2.1: Workflow for identification, annotation and presentation of *HMMER*-derived reverse transcriptase sequences.

2.3.4.1 Creation of a HMM

Hidden Markov Models were constructed with the *hmmbuild* function of *HMMER*3 using an alignment of transposon-typical amino acid reverse transcriptases. It is crucial that the underlying alignment is balanced in sequence as well as in organism diversity. For analysis of LINEs, the LINE RT alignment provided by Kapitonov *et al.* (2009) was shortened to contain only the eight characterized RT domains (Malik *et al.*, 1999; Wright *et al.*, 1996; Xiong and Eickbush, 1990). For analysis of Ty3-*gypsy*, Ty1-*copia* and BEL-Pao retrotransposons, RT alignments from the *Gypsy Database* have been applied without change (Llorens *et al.*, 2010; gydb.org). These alignments had to be converted to the *HMMER*3-compatible *Stockholm* format using the *Format Converter* software prior to HMM generation.

2.3.4.2 Searching a local database with a HMM

Plant genomes were translated in all six reading frames using the script *Translate-from-FASTA.py*. In case of large contig lengths, the sequences were partitioned into 100,000 bp fragments with 2000 bp overlaps prior to translation. Sequence fragmentation was performed using the program *PartitionSequence.py*.

The *hmmsearch* function of *HMMER*3 was applied to query the amino acid database with the HMM. A machine-readable tabular output (*domtblout* option) was saved and parsed by application of *HMMER-Parse.py* with consideration of the *HMMER* score and the alignment length. With this method, it is possible to retrieve the exact matches to the HMM query in fasta format. By application of one of the *get-nt-seq-from-HMMER-parse.py* scripts, it was also possible to extract the nucleotide sequences and, if desired, flanking regions.

2.3.4.3 Calibration of the alignment

For parsing of the HMM output, parameters were calibrated by a search against a set of previously identified retrotransposon ORFs containing reverse transcriptases. This set included LINEs, Ty1-*copia*, Ty3-*gypsy* and BEL-Pao retrotransposons, retroviruses and endogenous plant pararetrovirus sequences from Kapitonov *et al.* (2009) and from the *Gypsy Database* (Llorens *et al.*, 2010; gydb.org). The *HMMER*3 score threshold was defined 50, as hits with a higher score only included reverse transcriptases of the desired retrotransposon type.

2.3.5 Annotation of open reading frames, amino acid composition and secondary structure motifs

Specialized computational tools were employed to detect ORFs and define sequence features. These tools and their area of application are listed in Table 2.10.

3 Results

In order to study evolutionary routes and diversity of LTR and non-LTR retrotransposons in *B. vulgaris* genomes, exemplary families of each retrotransposon subclass have been analyzed in detail.

In a first approach, a *c₀t-1* DNA library, containing only highly repetitive sequences, has been screened for sequences homologous to retrotransposable elements (Zakrzewski *et al.*, 2010). Nearly half the *c₀t-1* clones contained TE sequences homologous to a single family of LTR Ty1-*copia* retrotransposons, the Cotzilla family (Chapter 3.1). In a second approach, sequences of the BvL LINEs (Wenke *et al.*, 2009). have been compared with short PCR-generated LINE regions. The BNR fragments (Schmidt *et al.*, 1995; Kubis *et al.*, 1998) have been found to harbour LINE-typical reverse transcriptases, but are considerable different from the BvL members (Chapter 3.2).

The BNR and Cotzilla retrotransposon families have been analyzed regarding their sequence, structure, abundance and localization on chromosomes to provide a base for the understanding of TE evolution in *B. vulgaris*.

3.1 Detection and characterization of the highly abundant retrotransposon family Cotzilla

Results from this chapter have been published in peer-reviewed journals (Weber et al., 2010).

3.1.1 Identification of a retrotransposon family in a *c₀t-1* library

During preliminary works, a *B. vulgaris* plasmid library of 1763 clones has been constructed and sequenced. These plasmids did contain only highly repetitive DNA, so-called *c₀t-1* DNA, which has been produced using renaturation kinetics of genomic DNA (Zakrzewski *et al.*, 2010).

A *BLASTx* analysis against the EBI protein database revealed the presence of 118 sequences with homology to retrotransposons. These 118 *c₀t-1* sequences have been assembled into contigs. The contig containing the highest number (20) of uncharacterized sequences (Figure 3.1) was marked by the presence of typical LTR features such as flanking inverted dinucleotide repeats TG/CA, and the primer binding site (PBS).

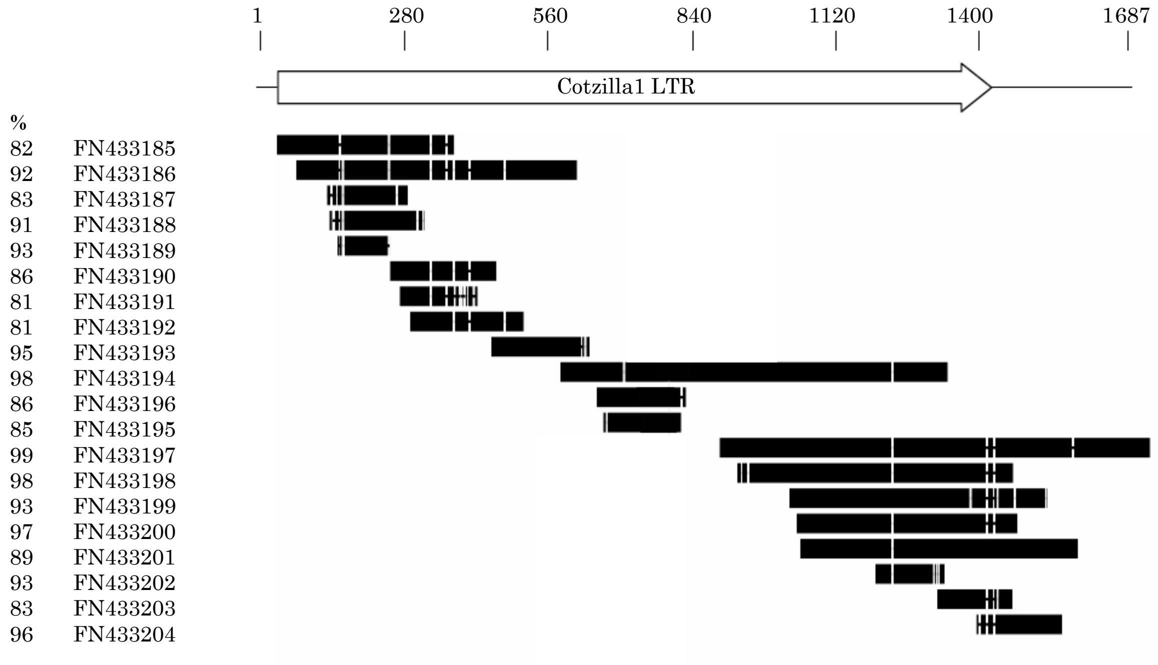


Figure 3.1: Schematic representation of a contig containing 20 *c₀t-1* sequences with similarities to a retrotransposon LTR.
Black bars represent sequences. The percentage identity relative to the Cotzilla1 5' LTR is indicated on the left side.

A *BLAST* search with the resulting LTR consensus sequence as query revealed a large number of *B. vulgaris* BAC sequences containing similar sequence motifs. A full-length member of this LTR retrotransposon family was detected on BAC EF101866 (26271-37103 nt) and designated Cotzilla1, a name referring to the *c₀t-1* DNA library where it was detected and its unusual size. This BAC has been sequenced by Kuykendall *et al.* (2009), who also detected this retrotransposon and provided initial analysis by comparing the sequence data to *BLASTp* hits. A highly diverged Cotzilla copy (Cotzilla3) was identified on BAC DQ374087 (71870-81547 nt). The molecular structures of Cotzilla1 and Cotzilla3 are shown in Figure 3.2 A, and the complete and annotated Cotzilla1 sequence is presented in Appendix 1.

In order to calculate the fraction of Cotzilla sequences in the *c₀t-1* DNA library, a homology search using the complete Cotzilla1 sequence revealed 40 *c₀t-1* sequences homologous to internal Cotzilla regions. Together with the 20 LTR-specific sequences, more than 50 % of all 118 retrotransposon-containing clones show homologies to Cotzilla.

Both LTR retrotransposons contain characteristic catalytic regions (RNA binding site/ protease/ integrase/ reverse transcriptase/ RNaseH) of the retroviral genes *gag* and *pol*. Based on the order of the enzymatic regions, these LTR retrotransposons can be classified as Ty1-*copia* retrotransposons. Cotzilla1 and Cotzilla3 differ in length and amino acid composition of the coding regions and most probably form two subfamilies. In the following, Cotzilla1 is considered as the reference element for the Cotzilla family.

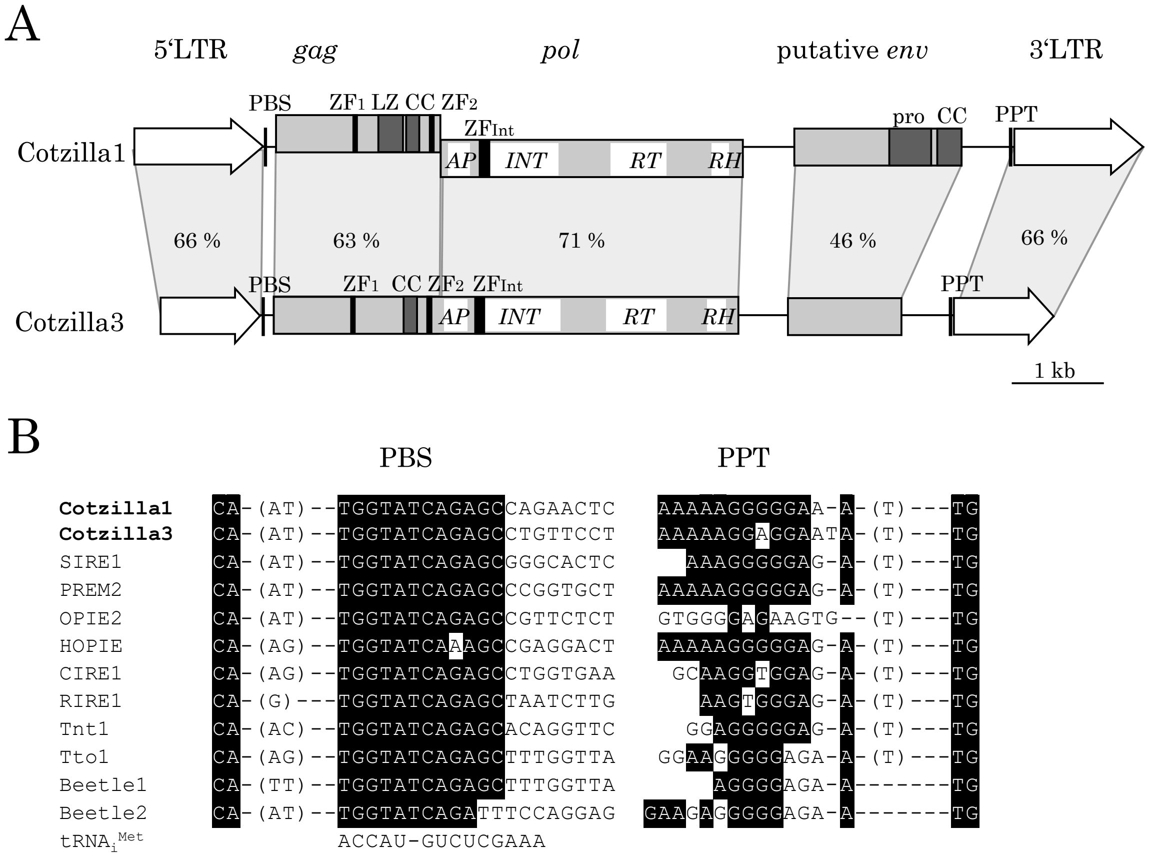

A
5'LTR
gag
pol
putative env
3'LTR
PBS
ZF1 LZ CC ZF2
ZFInt
pro CC
PPT
Cotzilla1
AP
INT
RT
RH
66 %
63 %
71 %
46 %
66 %
PBS
ZF1 CC ZF2 ZFInt
PPT
Cotzilla3
AP
INT
RT
RH
1 kb
B
PBS
PPT
Cotzilla1 CA-(AT)--TGGTATCAGAGCCAGAACTC AAAAAGGGGGAA-A-(T)---TG
Cotzilla3 CA-(AT)--TGGTATCAGAGCCTGTTCCT AAAAAGGAGGAATA-(T)---TG
SIRE1 CA-(AT)--TGGTATCAGAGCGGGCACTC AAAGGGGGAG-A-(T)---TG
PREM2 CA-(AT)--TGGTATCAGAGCCCGGTGCT AAAAAGGGGGAG-A-(T)---TG
OPIE2 CA-(AT)--TGGTATCAGAGCCGTTCTCT GTGGGGAGAAGTG--(T)---TG
HOPIE CA-(AG)--TGGTATCAAAGCCGAGGACT AAAAAGGGGGAG-A-(T)---TG
CIRE1 CA-(AG)--TGGTATCAGAGCCTGGTGAA GCAAGGTGGAG-A-(T)---TG
RIRE1 CA-(G)---TGGTATCAGAGCTAATCTTG AAGTGGGAG-A-(T)---TG
Tnt1 CA-(AC)--TGGTATCAGAGCACAGGTTC GGAGGGGGAG-A-(T)---TG
Tto1 CA-(AG)--TGGTATCAGAGCTTTGGTTA GGAAGGGGGAGA-A-(T)---TG
Beetle1 CA-(TT)--TGGTATCAGAGCTTTGGTTA AGGGGAGA-A-------TG
Beetle2 CA-(AT)--TGGTATCAGATTTCCAGGAG GAAGAGGGGGAGA-A-------TG
tRNAiMet ACCAU-GUCUCGAAA

Figure 3.2: Schematic representation of the sugar beet Ty1-*copia* retrotransposons Cotzilla1 and Cotzilla3.
(A) The open terminal arrows at each end of Cotzilla1 and Cotzilla3 represent the long terminal repeats (LTRs). Conserved domains are shown: *gag*, protease (AP), integrase (INT), reverse transcriptase (RT), RNaseH (RH), primer binding site (PBS) and polypurine tract (PPT). The grey rectangles represent ORFs, which contain protein or nucleotide binding motifs like zinc finger domains (ZF — black), coiled coils (CC — dark grey), putative leucine zippers (LZ — dark grey) and proline-rich domains (pro — dark grey). The percentages between Cotzilla1 and Cotzilla3 represent their nucleic acid sequence identity.
(B) Alignment of Ty1-*copia* retrotransposon primer binding sites (PBS) complementary to the initiator tRNA of methionine ($tRNA_i^{Met}$) and polypurine tracts (PPT). Shading indicates identity of 67 %.

The full-length Cotzilla1 has a length of 10,833 bp (including 1372 bp LTRs) and is flanked by a target site duplication of five base pairs (ATTTT), which is typically generated upon integration of LTR retrotransposons.

Using PlantCARE and PLACE databases, sequence stretches containing putative eukaryotic core promotor motifs (numerous TATA boxes) in the LTRs have been detected. Additional putative DNA motifs related to light, defense, stress and low temperature response have also been present (Figure 3.3).

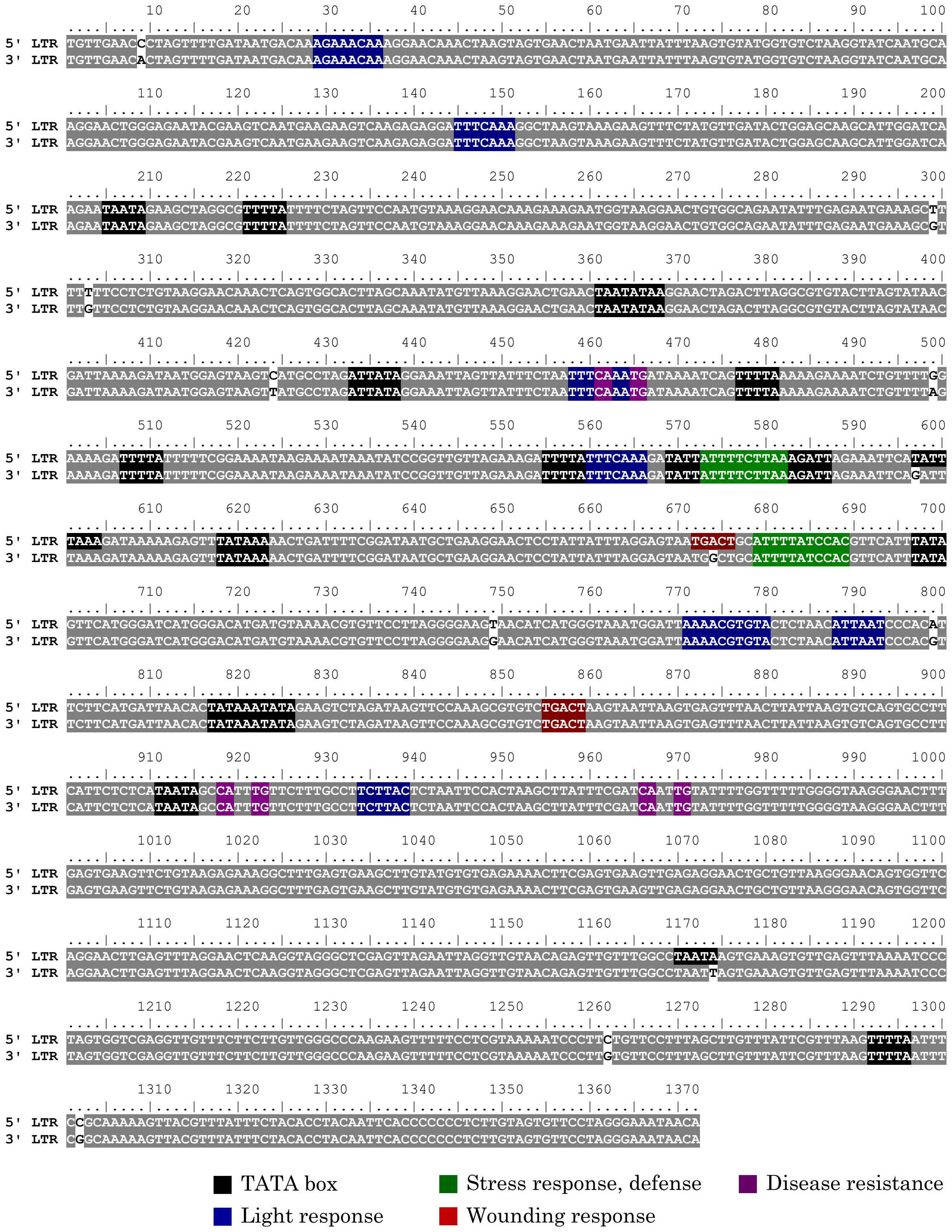


Figure 3.3: Sequence alignment of the Cotzilla1 5‘ and 3‘ LTRs showing putative eukaryotic promotor motifs.

The Cotzilla 5’ and 3’ LTR sequences have been aligned. Identical residues of both LTRs are shaded in grey. Additionally, putative eukaryotic core promotor motifs, as well as putative DNA motifs related to light (blue), stress (green), wounding response (red) and disease resistance (violet) occurring on the LTR sense strand.

Primer binding sites (PBS) are located downstream of the 5' LTR and show complementarity to the 3' end of the $tRNA_i^{Met}$. Adjacent to the 3' LTR, the polypurine tracts (PPT) were identified, consisting of 13 purines for Cotzilla1 (Figure 3.2 B).

While conventional Ty1-*copia* retrotransposons contain the *gag* and *pol* genes in a single continuous ORF, a +1 frameshift separates these genes in Cotzilla1. The *gag* ORF is terminated by a stop codon and a conserved motif upstream of the stop (Figure 3.4 A) as described for Sireviruses (Gao *et al.*, 2003). This palindromic motif might facilitate hairpin formation of the corresponding RNA (Figure 3.4 B). However, no start codon could be identified at the beginning of the *pol* ORF upstream of the protease motif D(S/T)G as defined by Peterson-Burch and Voytas (2002).

For verification of this frameshift, a homology search using the EBI database was carried out, resulting in 39 Cotzilla sequences with a similar *gag-pol* transition (Figure 3.4 A). Of these hits, the majority (31) have the frameshift between *gag* and *pol* and a stop codon terminating the *gag* ORF. While 20 of 39 sequences have the *gag* stop codon at the same position as Cotzilla1, 12 sequences terminate at a different, but conserved position indicating the existence of two subfamilies. The remaining seven Cotzilla sequences (including Cotzilla3) are more diverged from the consensus and do not contain any stops. Another difference to conventional Ty1-*copia* elements like Tnt1 and Tto1 (Grandbastien *et al.*, 1994; Hirochika *et al.*, 1996) is the existence of an extended *gag* region in Cotzilla1 with more than 300 additional amino acid residues. An accumulation of secondary structure motifs was detected in the expanded *gag* region of Cotzilla1, such as an additional zinc finger ($Cx_2Cx_4Hx_4C$), a putative leucine zipper and a predicted coiled coil domain, suggesting RNA- and protein-binding functions. A peculiar feature of the Cotzilla family is the presence of an additional putative ORF adjacent to the 3' LTR as described for the Sirevirus *env*-like ORFs. In Cotzilla1, this ORF was identified 561 nt downstream of the *pol* polyprotein (Figure 3.2). Using the first putative start codon, the *env* ORF of Cotzilla1 has a length of 1819 bp and contains one frameshift. Transmembrane domains were absent in the coding sequence. Instead, a putative coiled coil structure next to a proline-rich motif upstream of the coil is present at the C-terminus of this ORF.

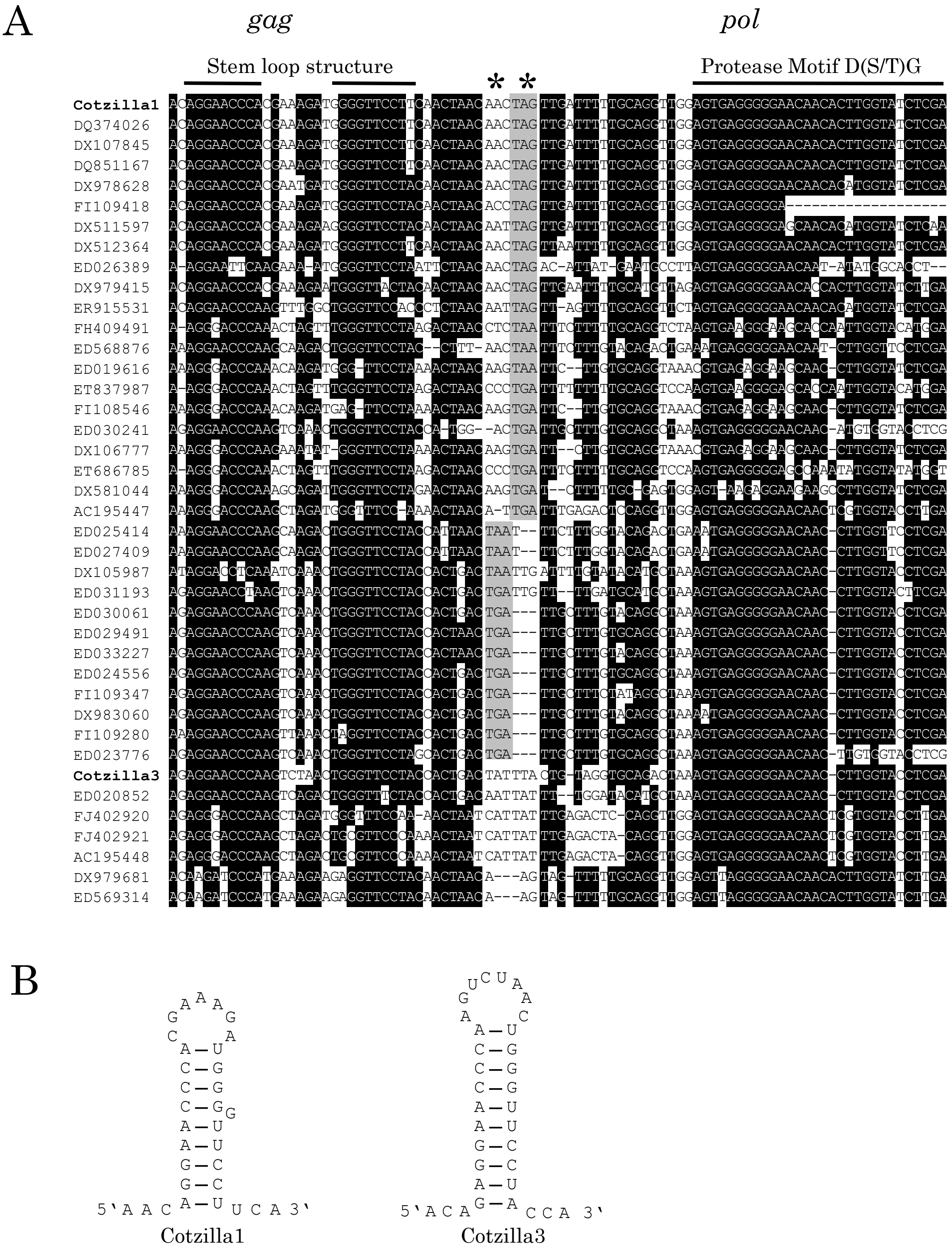


Figure 3.4: DNA sequence alignment of Cotzilla *gag-pol* frameshift regions.

(A) Forty Cotzilla sequences containing the *gag-pol* transition are aligned. Shading indicates sequence identity of at least 60 %. A conserved motif in the *gag* ORF enabling hairpin formation of the corresponding RNA and the protease motif D(S/T)G of the *pol* protein are indicated in the alignment. Furthermore, two subsets of Cotzilla elements with a stop codon terminating the *gag* ORF at different positions were identified. The stops are shaded in grey and marked with asterisks.

(B) The Cotzilla1 and Cotzilla3 hairpins correspond to their respective Cotzilla RNAs.

For analysis of the relationship of the Cotzilla elements to other plant LTR retrotransposons, an unrooted dendrogram was constructed using an amino acid sequence alignment of the reverse transcriptase (RT) domains defined by Xiong and Eickbush (1990) (Figure 3.5). SALIRE1, a conventional Ty1-*copia* retrotransposon of *B. vulgaris* has been included in the analysis (Weber *et al.*, 2010). However, whereas SALIRE1 is grouped together with typical Ty1-*copia* elements like Tnt1 and Tto1 (Grandbastien *et al.*, 1994; Hirochika *et al.*, 1996), Cotzilla1 and Cotzilla3 cluster with *env*-containing Sireviruses such as *SIRE*1 (Laten *et al.*, 1998). The high bootstrap values of the dendrogram as well as the identified structural motifs support this assignment.

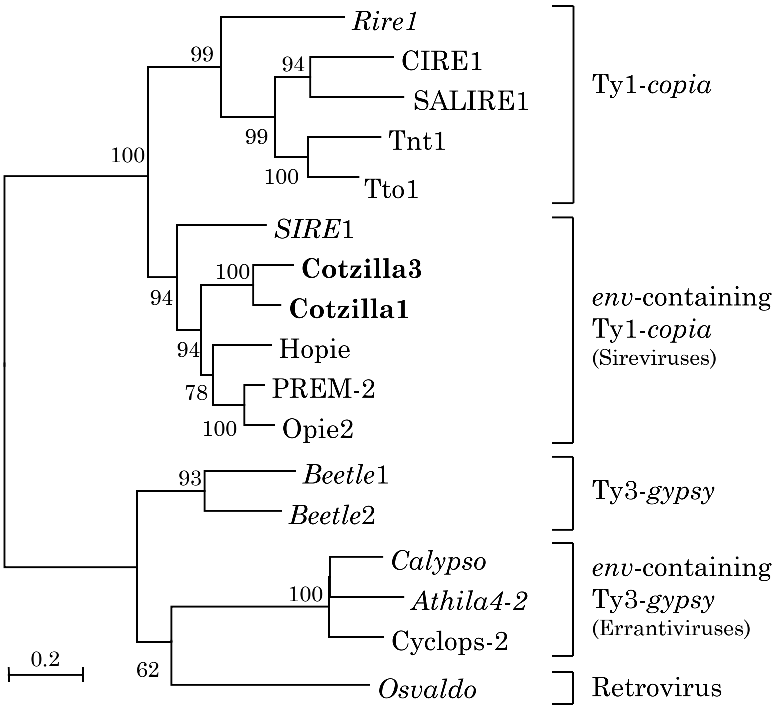


Figure 3.5: Dendrogram showing the relationship of Cotzilla RT amino acid sequences to other LTR retrotransposons.
The RT domains of the plant retrotransposons of the following species were analyzed: *C. sinensis* CIRE1; *N. tabacum* Tnt1 and Tto1; *O. australiensis Rire1*; *G. max* SIRE1; *Zea mays* Hopie, Opie2 and PREM-2; *B. procumbens Beetle*1 and *Beetle*2; *G. max Calypso* and *A. thaliana Athila4-2*. For comparison the retrovirus RT sequence of *Osvaldo* from *D. buzzatii* was included. Bootstrap values are indicated as a percentage of 1000 replicates. Branch lengths are proportional to genetic distance. The scale bar represents 0.2 substitutions per site.

3.1.2 Abundance, genomic organization, and distribution of Cotzilla-like sequences in the genera *Beta* and *Patellifolia*

The genomic organization of Cotzilla retrotransposons was investigated by Southern hybridization. Genomic DNA was restricted with five enzymes and hybridized with probes spanning the RT domains III to IV, a part of the LTR, or of the putative *env* ORF of Cotzilla1 (Figure 3.6 A). The presence of strong signals and conserved fragments visible after one day exposure indicates a high abundance and conserved restriction sites in Cotzilla elements or in their flanking genomic regions. After longer exposure, a strong signal smear was visible showing hybridization to DNA fragments over a wide range of molecular weight (not shown).

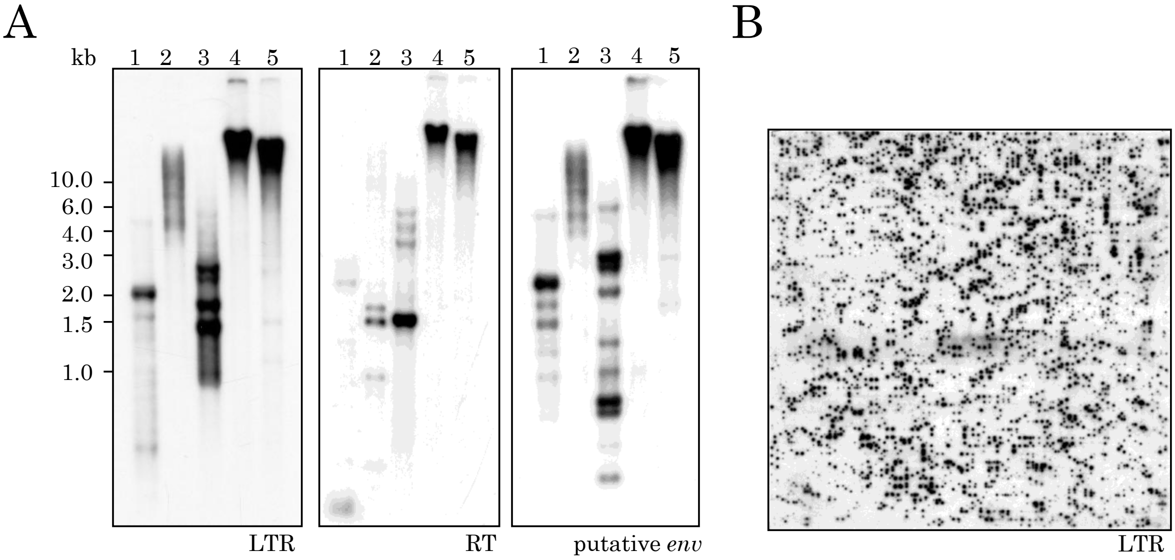


Figure 3.6: Southern hybridization showing the genomic organization and abundance of Cotzilla1.
(A) Genomic DNA of *B. vulgaris* was restricted with *Dra*I (1), *Eco*RI (2), *Hae*III (3), *Hpa*II (4) and *Msp*I (5). Autoradiograms are shown after hybridization of probes specific for the Cotzilla1 LTR, RT and *env*-like region after one day exposure.
(B) Hybridization of Cotzilla1 LTR to a high density BAC filter with of 9216 BAC clones. 1275 positive clones were detected after 1 day exposure.

Cytosine methylation was analyzed by comparative hybridization to genomic DNA restricted with *Hpa*II and *Msp*I (lanes 4 and 5). While *Hpa*II cuts only unmethylated CCGG sequences, *Msp*I is able to tolerate methylation of the internal cytosine. When using *Hpa*II restricted DNA, all Cotzilla1 probes hybridized predominantly to large fragments, which cannot be resolved by conventional gel electrophoresis. This shows strong methylation of one or both cytosines at CCGG sites of Cotzilla retrotransposons in the *B. vulgaris* genomes. *Msp*I restriction generates smaller fragments instead, resulting in a shift of the signal smear. The existence of weak bands at 1.5 kb and 2.5 kb indicate that some CCGG sites of Cotzilla members are not methylated at the outer cytosine of the restriction site.

Homology of Cotzilla1 to many *B. vulgaris c₀t-1* clones indicates the existence of a plethora of homologous copies. To test this assumption, a hybridization of *B. vulgaris* high-density filters containing 9216 BAC clones in duplicates (approximately 1.5 genome equivalents) with the Cotzilla1 LTR probe was performed (Figure 3.6 B). Massive signals corresponding to 1275 BACs have been observed. This verifies high abundance in the *B. vulgaris* genome.

In order to investigate the diversity and abundance of this Sirevirus family within the beet genera *Beta* (sections *Beta*, *Corollinae* and *Nanae*) and *Patellifolia*, Southern hybridization was carried out to *Hae*III restricted DNA from representative species. The same LTR probe as above has been used for hybridization (Figure 3.7).

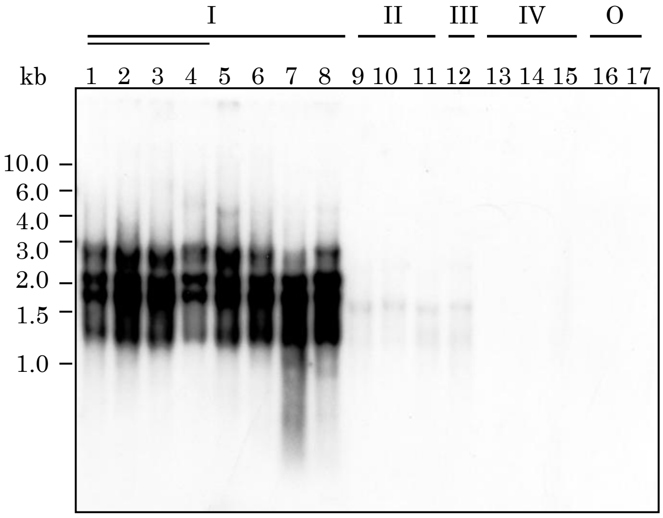


Figure 3.7: Distribution of Cotzilla1 sequences in the genera *Beta* and *Patellifolia*.
Genomic *Hae*III-restricted DNA was analyzed by comparative Southern hybridization using a probe from Cotzilla1 LTR (exposure time = 1 day). The following cultivars of *B. vulgaris* ssp. *vulgaris* in the section *Beta* (I) were tested: sugar beet (1), fodder beet (2), garden beet (3), chard (4). Wild beet species were used from the section *Beta* (I): *B. vulgaris* ssp. *adanensis* (5), *B. vulgaris* ssp. *maritima* (6), *B. patula* (7), *B. macrocarpa* (8); species of the section *Corollinae* (II): *B. corolliflora* (9), *B. macrorhiza* (10), *B. lomatogona* (11); species of the section *Nanae* (III): *B. nana* (12); species of the genus *Patellifolia* (IV): *P. procumbens* (13), *P. patellaris* (14), P. *webbiana* (15). As outgroup species (O) served *Chenopodium quinoa* (16) and *Spinacia oleracea* (17).

After one day exposure, strong hybridization to DNA of all species in the section *Beta* was detected, showing high abundance in this section. The observed pattern was similar for all species, indicating conserved restriction sites in Cotzilla elements or in their flanking regions. Furthermore, longer exposure revealed weak signals which were conserved throughout the sections *Corollinae* and *Nanae*. In the genus *Patellifolia* and in the outgroups *Chenopodium quinoa* and *Spinacia oleracea*, no hybridization of Cotzilla retrotransposons has been observed.

3.1.3 Chromosomal localization of the Cotzilla family

Localization of the Cotzilla families on *B. vulgaris* chromosomes was investigated by fluorescent *in situ* hybridization (FISH). The same probes as used for Southern analyses were labelled with biotin-11-dUTP, followed by hybridization to *B. vulgaris* metaphase and prometaphase chromosomes and interphase nuclei. The FISH images show signals of varying intensity with dispersed hybridization on all chromosomes (Figure 3.8 A, C and E).

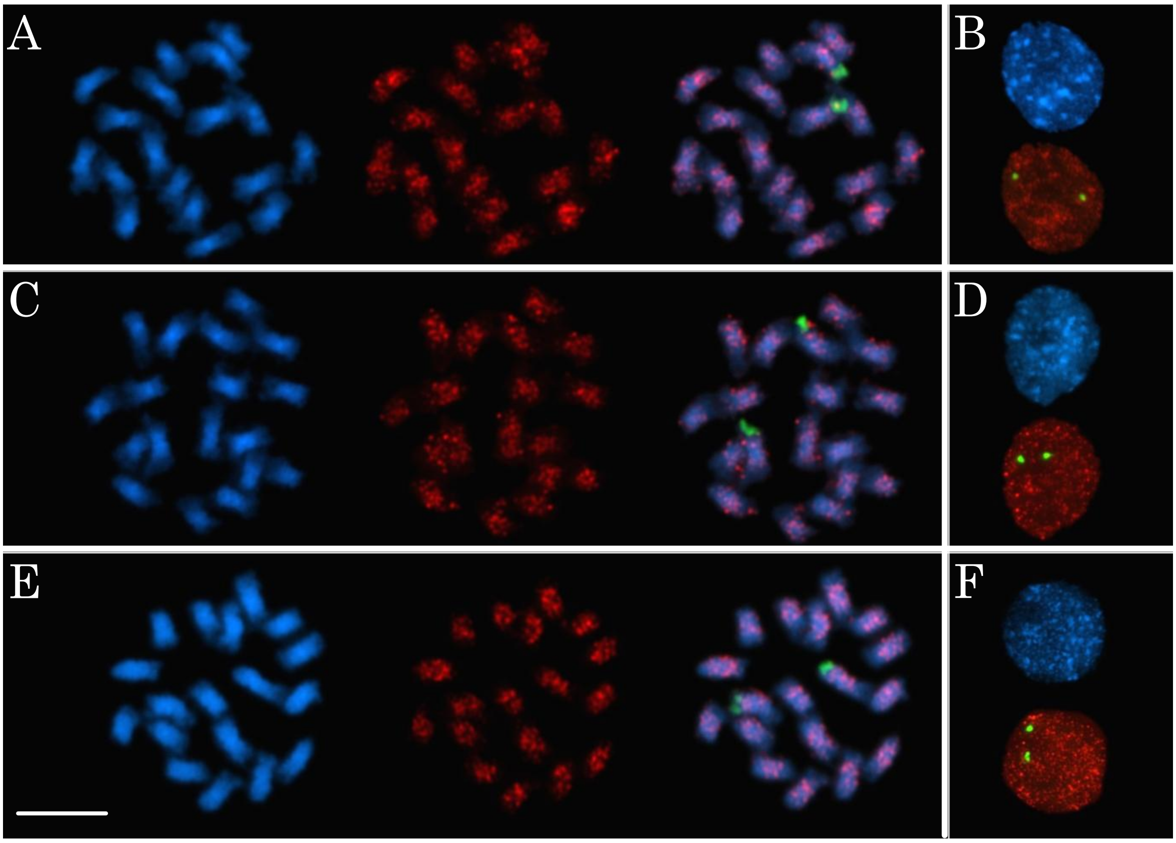


Figure 3.8: Chromosomal distribution of the Cotzilla LTR retrotransposon family along *B. vulgaris* chromosomes by fluorescent in situ hybridization (FISH).
In each panel, the DAPI-stained DNA (blue fluorescence) shows the morphology of the chromosomes. Retrotransposon hybridization signals are visible as red signals, while green fluorescence indicates hybridization with the 18S-5.8S-25S rDNA on chromosome 1. Metaphase and interphase nuclei were hybridized with an RT-specific **(A and B)**, an *env*-specific **(C and D)**, and an LTR-specific Cotzilla1 probe **(E and F)**. The scale bar in **(E)** corresponds to 10 µm.

Strong signal clusters forming blocks on both chromosome arms of most chromosomes were detected by hybridization of Cotzilla1 RT, *env* and LTR probes to metaphase chromosomes. These signals were preferentially located in the intercalary and pericentromeric heterochromatin, while hybridization at centromeres was strongly reduced. Furthermore, an exclusion of Cotzilla from the outermost distal euchromatin was detected with an LTR probe (Figure 3.8 E). Since the LTR is specific for a retrotransposon family, usage of an LTR probe allows unambiguous detection of Cotzilla retrotransposons only. However, a few Cotzilla copies have been detected at chromosome termini with RT and *env* probes. At higher resolution, depletion of Cotzilla retrotransposons from most DAPI-negative euchromatic regions and nucleoli could be observed in interphase nuclei (Figure 3.8 B, D and F).

Double-color-FISH with an 18S-5.8S-25S rDNA probe (green fluorescence) showed an exclusion from the rRNA gene arrays. However, weak but clear Cotzilla signals located in vicinity of rRNA genes have been visualized in Figure 3.8 A and Figure 3.8 C.

3.1.4 LTR variability and estimation of insertion time

In order to determine LTR diversity within the family, homology searches were carried out against the EBI database to identify similar LTR sequences present in 23,068 BAC end sequences corresponding to approximately 18 Mb (McGrath et al. 2004). Only sequences which were anchored to the PBS were considered for the estimation of LTR divergence, thus allowing an assignment to 5' LTRs of individual retrotransposon copies. In total, 50 Cotzilla LTR sequences upstream of the PBS with an average length of 298 bp, 22 % of total LTR length, have been analyzed. These LTRs had an average identity of 96 % ranging from two identical sequences to sequences harboring indels resulting in a minimum identity of 89 %. These very high numbers show the presence of many identical Cotzilla copies in the genome.

As LTR pairs are usually identical upon integration, the divergence of 5' and 3' LTR sequences can be used to estimate the time that passed after the transposition of the Cotzilla copies (SanMiguel *et al.*, 1998). Cotzilla1 has 12 nucleotide mismatches in the 1372 bp long LTRs and therefore integrated approximately 290,000 years ago, based on an average synonymous substitution rate of 1.5 x 10^{-8} mutations per site per year. Similarly, for Cotzilla3 an age of 850,000 years could be estimated, indicating that Cotzilla3 is older than Cotzilla1.

The Cotzilla1 retrotransposition event is relatively young on an evolutionary time scale and proof of recent Cotzilla amplification. The presence of young Cotzilla members also serves to explain the homogeneity of the detected LTRs. However, the detection of the older, more degenerated Cotzilla3 element shows that only a subset of the Cotzilla family is homogenous and young.

3.2 The BNR LINE family defines a novel subclade of L1 LINEs

Results from this chapter have been published in peer-reviewed journals (Heitkam et al., 2009).

In order to get a first impression of non-LTR retrotransposon diversity on the scale of full-length elements, at least two LINE families have to be compared. Prior to this thesis' work, already four fragments of the BNR family with a size ranging from 0.3 to 1.3 kb (Schmidt *et al.*, 1995; Kubis *et al.*, 1998) and three full-length members of the BvL family have been identified (Wenke *et al.*, 2009). Those family abbreviations correspond to Beet non-LTR retrotransposon (BNR) and *Beta vulgaris* LINE (BvL). All of these sequences contained a similar 300 bp region of their RT (Figure 3.8 A). Comparison of these regions showed large differences in amino acid composition with an average RT identity of only 47 % between BvL and BNR members. This finding was visualized by a Neighbor-Joining dendrogram showing that both families occupy different branches (Figure 3.8 B). Analysis of full-length BNR family members is supposed to reveal, whether the LINE diversity observed is limited to the RT part, or whether the BNR family characteristics differ from the ones observed for BvL.

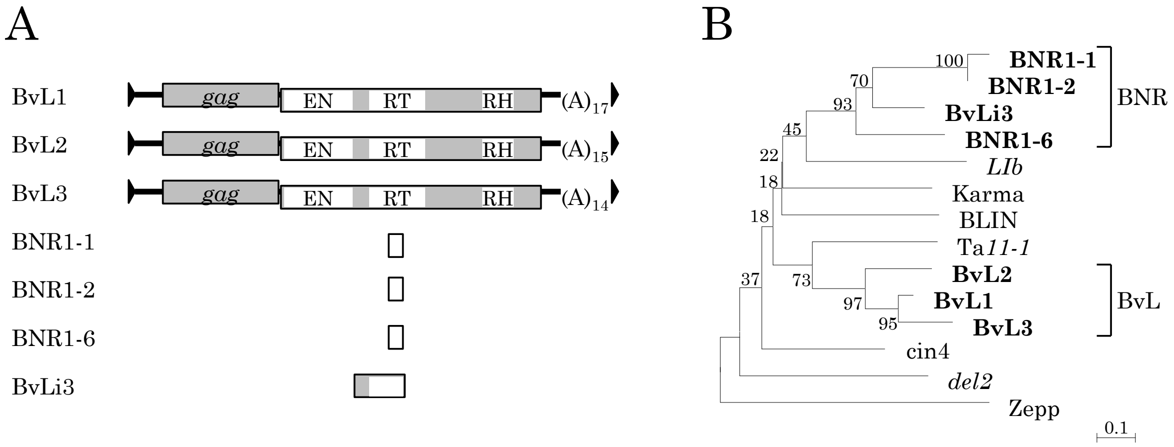


Figure 3.9: *B. vulgaris* LINEs and LINE fragments that have been described prior to this thesis.
(A) Sequences of three full-length BvL family members (Wenke *et al.*, 2009), three approximately 300 bp and one 1.3 kb BNR RT fragments (Schmidt *et al.*, 1995; Kubis *et al.*, 1998) provided the basis for further LINE isolation. They are represented here as schematic drawings.
(B) Visualization of a LINE RT alignment by application of the Neighbor-Joining algorithm. All known *B. vulgaris* LINE sequences have been compared with a set of characterized plant LINEs (Table 2.12). BvL and BNR LINEs do not group together, but occupy different branches of the tree. Bootstrap values are indicated as a percentage of 1000 replicates. The scale bar represents 0.1 substitutions per site.

3.2.1 Isolation of BNR sequences from *B. vulgaris*

A high density filter consisting of 9216 BAC clones (with approximately 1.5 genome equivalents) was probed with the LINE-fragment BNR1-2, to identify a BAC containing the sequence of a corresponding non-LTR retroelement from *B. vulgaris* (Figure 3.10 A). Three BACs with a strong hybridization signal were detected, isolated, restricted and probed again (Figure 3.10 B). All three BACs gave an identical hybridization pattern using the enzymes *Eco*RI (one fragment) and *Hin*dIII (two fragments). The presence of a unique *Hin*dIII restriction site in the BNR probe explains the two signals after *Hin*dIII hybridization. This and the single hybridization band after *Eco*RI restriction indicates that only one LINE copy is present on each BAC. The 6 kb *Eco*RI restriction fragment of BAC 47M6 was selected for subcloning into pUC18. Ninety-six sub clones were selected, grown on a 96 well plate and a nylon membrane, and hybridized again with BNR1-2 (Figure 3.10 C).

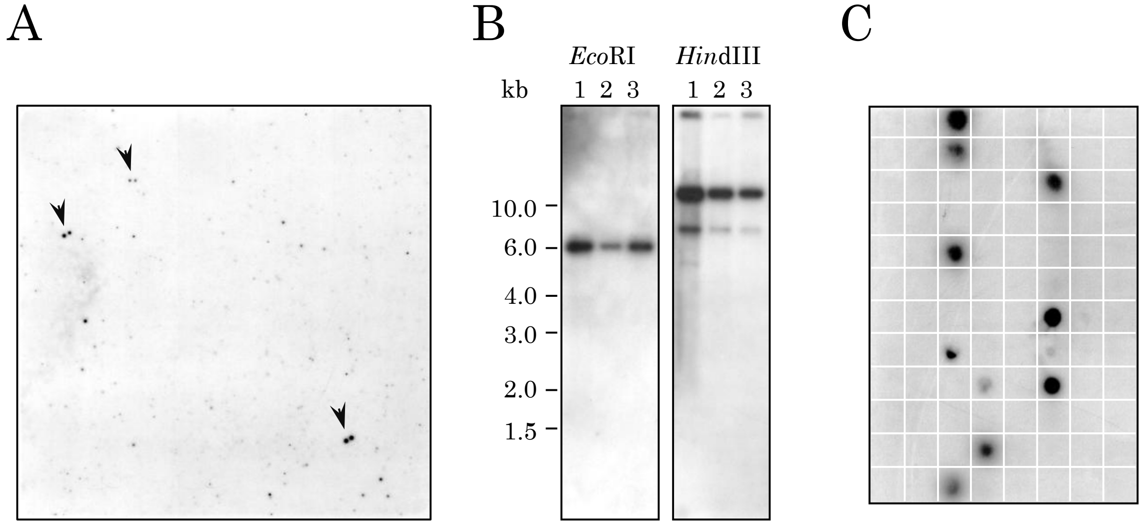


Figure 3.10: Steps of the isolation of a complete BNR retrotransposon.
The regions containing a BNR LINE fragment have been narrowed down by consecutive hybridizations with the probe BNR1-2.
(A) Hybridization of a high density filter containing 9216 clones in duplicates shows three strong signals (arrows).
(B) Three BACs (1: 33F21; 2: 47M6; 3: 27I20) have been selected for restriction, gel separation, and hybridization. The autoradiogram showed either a single or two bands for *Eco*RI and *Hin*dIII restriction, respectively.
(C) *Eco*RI restricted DNA of BAC 47M6 was selected for subcloning. Clones have been grown on a 96 well plate and a corresponding nylon membrane. BNR-positive subclones have been identified by hybridization of the membrane with the BNR probe. The corresponding autoradiogram is shown.

The positive clones were tested for insert size, and one of the plasmids with a 6 kb insert was chosen for sequencing by primer walking. By combination of subclone and BAC sequencing, a 6700 bp sequence was obtained that ended with a poly(A) tail and was flanked by a 16 bp target site duplication. A *BLAST* search confirmed similarity to non-LTR retrotransposons, more exactly to LINEs. The identified LINE copy has been designated BNR1 (EU564339) as has been reported (Heitkam and Schmidt, 2009).

Using BNR1 as query for a homology search in *B. vulgaris* BAC sequences submitted to EBI, two further BNR copies, BNR2 and BNR3 have been identified. BNR2 was detected on sequenced BAC fragments (DQ374060 and DQ374061), but was incomplete at the 5' end. Primers were constructed for gap closure between the DQ374061 and DQ374062 by PCR using *B. vulgaris* DNA as template and resulting in the complete sequence of BNR2. The complete sequence of BNR3 was identified on BAC DQ374017.

The structure of these full-length BNR LINEs along with their nucleic acid sequence similarity is shown in Figure 3.11. Furthermore, Appendix 2 presents the annotated sequence of the reference LINE BNR1.

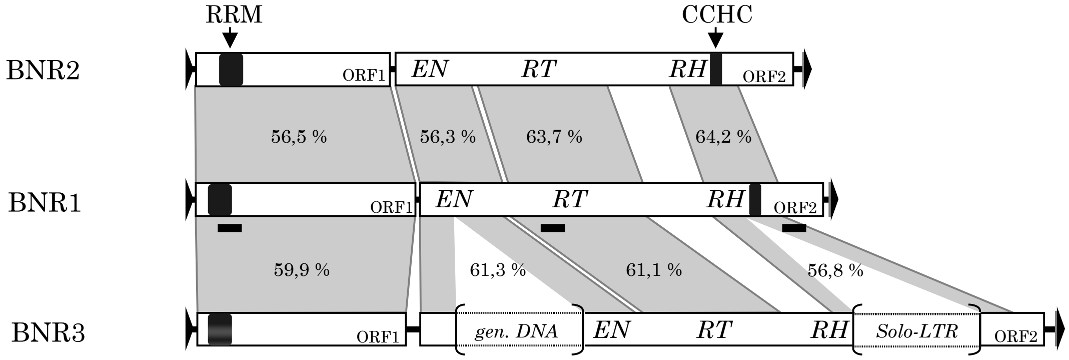


Figure 3.11: Schematic representation of the sugar beet LINEs BNR1, BNR2 and BNR3. The rectangles represent two ORFs with conserved motifs (ORF1: black - RNA recognition motiv (RRM); ORF2: black - zinc finger (CCHC) in the RNaseH domain). EN, RT and RH refer to the catalytic regions of the endonuclease, reverse transcriptase and RNaseH. Insertions in BNR3 are shown in brackets. The shaded regions show the nucleic acid sequence identity between the BNR copies. The bold lines below BNR1 represent the regions of the probes used for Southern hybridization.

3.2.2 BNR1 and related elements constitute a novel family of LINEs

The BNR1 amino acid sequence of ORF2 was deduced by comparison with the sequences of the active LINEs *Karma* and *LIb* (Komatsu *et al.*, 2003; Yamashita and Tahara, 2006). Analysis of ORF2 showed that this LINE belongs to the L1 clade which also includes human and mouse LINEs as well as all plant LINEs characterized so far. ORF2 encodes a polyprotein consisting of the conserved regions of an endonuclease, eight domains of the reverse transcriptase RT (Malik *et al.*, 1999; Wright *et al.*, 1996; Xiong and Eickbush, 1990) and a zinc finger $Cx_2Cx_8Hx_4C$ in the RNaseH domain.

ORF1 was identified as a continuous sequence upstream of ORF2 and contains a single stop codon. It is supposed to encode a *gag*-like nucleic acid chaperone activity (Martin, 2006). However, the ORF1 of BNR1 does not encode a zinc finger motif to facilitate RNA binding, but a different RNA-binding motif located in the N-terminus. The start codon of ORF1 delimits the short 5' UTR of 54 bp upstream of ORF1. The LINE is terminated by a 3' UTR of 160 bp and a poly(A) tail consisting of eight adenine residues.

BNR2 has a length of 6402 bp, is flanked by a target site duplication of 22 bp, and codes for two ORFs having features similar to BNR1. Instead, BNR3 has a rearranged sequence. It also encodes two ORFs, however many frameshifts had to be introduced to optimize the alignment with other BNR members. In ORF2, insertions of 1380 bp and 1360 bp were found, interrupting the endonuclease and RNaseH region. The insertion in the endonuclease region has a high A/T content and contains four different degenerated repeats unrelated to LINEs. The insertion in the RNaseH gene was identified as a Solo-LTR because of its 73 % sequence identity with the LTR of an internally deleted *env*-like LTR retrotransposon on the *B. vulgaris* BAC fragment DQ374067 from 8438 nt to 14,904 nt (Cora Wollrab, TU Dresden, personal communication). Including both insertions, BNR3 has a length of 9323 bp and a 7 bp target site duplication.

Excluding the insertions, computational translation of the ORF2 revealed a high similarity (72-81 %) or identity (55-63 %) in the RT regions between BNR1, BNR2 and BNR3. With an overall amino acid identity of 35 %, complete BNR RT regions do not show significant similarity to RT genes of the conventional LINE family BvL in *B. vulgaris* (Wenke *et al.*, 2009).

3.2.3 Comparative analysis of BNR members shows existence of subfamilies and conserved ORF1 domains

During progress of this thesis' work, first draft versions of the *B. vulgaris* genome sequence became available (Table 2.8) and allowed the isolation of a number of additional full-length BNR sequences and susequently, their comparative analysis.

By *tBLASTn* with the ORF1 of BNR1, LINEs homologous to BNR1 to BNR3 were detected in the *B. vulgaris* genome databases. Fifty-six of them were flanked by a TSD and were mostly of full length (Table 3.1). Interestingly, nine BNR copies do not contain mutations like internal stops and frameshifts. In theory, these LINEs might be retrotranspositionally competent, and are referred to as 'intact'.

Table 3.1: 56 BNR LINEs in *B. vulgaris* and their features.

LINE	Accession	Length [bp]	TSD	Remarks
BNR1	EU564339	6700	AACACTCACGCGTCTA	
BNR2	DQ374060/61	6402	AATGCAAGTGAGATAATATAA	
BNR3	DQ374017	9323	AAAAAGGG	rearranged; two insertions
BNR7	FR852838	6563	AAATCTGTATGTACAACA	no internal frameshifts or stops
BNR8	FR852839	6674	AAAGTTTAACTACATACAA	
BNR9	FR852840	6612	AAGAACAATATATCAGAT	
BNR10	FR852841	6612	AAAGTAGCACCCCCAAAA	
BNR12	FR852842	6615	AATAATTAATGACATGG	
BNR19	FR852795	6670	AAACAATACTTTAGTGAAA	no internal frameshifts or stops
BNR21	FR852843	6586	AATCATGTCAACAAGAT	
BNR22	FR852844	6618	AATATTGAAATA	no internal frameshifts or stops
BNR23	FR852845	6459	TGTCATTGTTAT	
BNR25	FR852846	6680	AATTTTGTAAAATTGAAA	
BNR29	FR852847	6403	AATCAACCAAACAACTA	
BNR30	FR852848	6423	AACATGATCAACTAACAGTAATAA	
BNR31	FR852849	6401	AACACTCCCCAAGACTC	
BNR33	FR852850	6526	CCCTGAAATGAA	
BNR34	FR852851	6424	AACTTTTTAAGATAAAAAT	
BNR37	FR852852	6482	AACTATCTCATATTTTCTA	
BNR38	FR852853	6501	AATAAAACATTTTTAA	
BNR39	FR852854	6572	AATTTTTTTACAAAA	
BNR41	FR852855	6592	AGTAATATAAGGTATAGT	
BNR45	FR852856	6652	AATAGAATTAGCATATAA	no internal frameshifts or stops
BNR51	FR852857	6609	AACAATTTATATAT	no internal frameshifts or stops
BNR52	FR852858	6450	AACATGTCTTTGTTTTA	
BNR55	FR852859	6583	AACATTTGTTTTTT	
BNR57	FR852860	6547	AAGATAATGTTGACAACA	
BNR59	FR852861	6549	AAGATCAAAGTGTAATAA	no internal frameshifts or stops
BNR60	FR852862	6463	AATAGGATGAAAAGATT	
BNR69	FR852863	6524	AAACAACAATAAAGAGCG	
BNR70	FR852864	6339	AACAAAAAAACGGAG	
BNR74	FR852865	6601	AAAAAGAATCAGAGGAA	
BNR76	FR852866	6593	AACTAAGAAAAGTTCCTTA	no internal frameshifts or stops
BNR79	FR852867	6355	AAAGAAGATGCTGGAAAAGA	
BNR81	FR852868	6376	AACATCGATCGTTCTTATA	
BNR83	FR852869	6493	ATGCATTACTTATGCTAA	
BNR84	FR852870	6411	AGGCTAATAGAGATTAT	5‘ truncated
BNR85	FR852871	6539	AACTCTATAATTGCATGCAACAC	
BNR86	FR852872	6472	ACCTGGTTTAGCGCAAA	
BNR87	FR852873	6506	AAATAAAGCATATGGGTAA	
BNR89	FR852874	6018	AAATCACTTATCTT	
BNR90	FR852875	6670	AAAGTTACTTTTCCTA	
BNR91	FR852876	6439	AAGGATGAAAGAGATG	
BNR95	FR852877	6521	AATTAATATGCACCGGG	
BNR96	FR852878	6393	AACATGGAATGAAA	no internal frameshifts or stops
BNR98	FR852879	6498	AATAATTAACCATATA	
BNR99	FR852880	6567	AAGCCATATAACGAGTTA	
BNR101	FR852881	6358	AAGTGTGCAATAGATT	
BNR103	FR852882	6340	AATTAGAATGCAAACAA	
BNR107	FR852883	6538	ATATGAAAGG	
BNR110	FR852884	5908	AGGTGGTTTTTTC	5‘ truncated
BNR114	FR852885	6578	AATTCTATTTTAATATGT	no internal frameshifts or stops
BNR115	FR852886	6855	AAGAGTATAAGCCACGT	
BNR116	FR852887	6319	ATAAATAATTTTTCCTT	
BNR119	FR852888	7091	AATGTGTAACTTTATGATA	~ 400 bp insertion in ORF2
BNR122	FR852889	6465	AAGAATACAATATCTTCTA	

In order to characterize sequence similarity of BNR family members, the nine intact LINEs have been compared to BNR1 using the *Circoletto* visualization software (Figure 3.11).

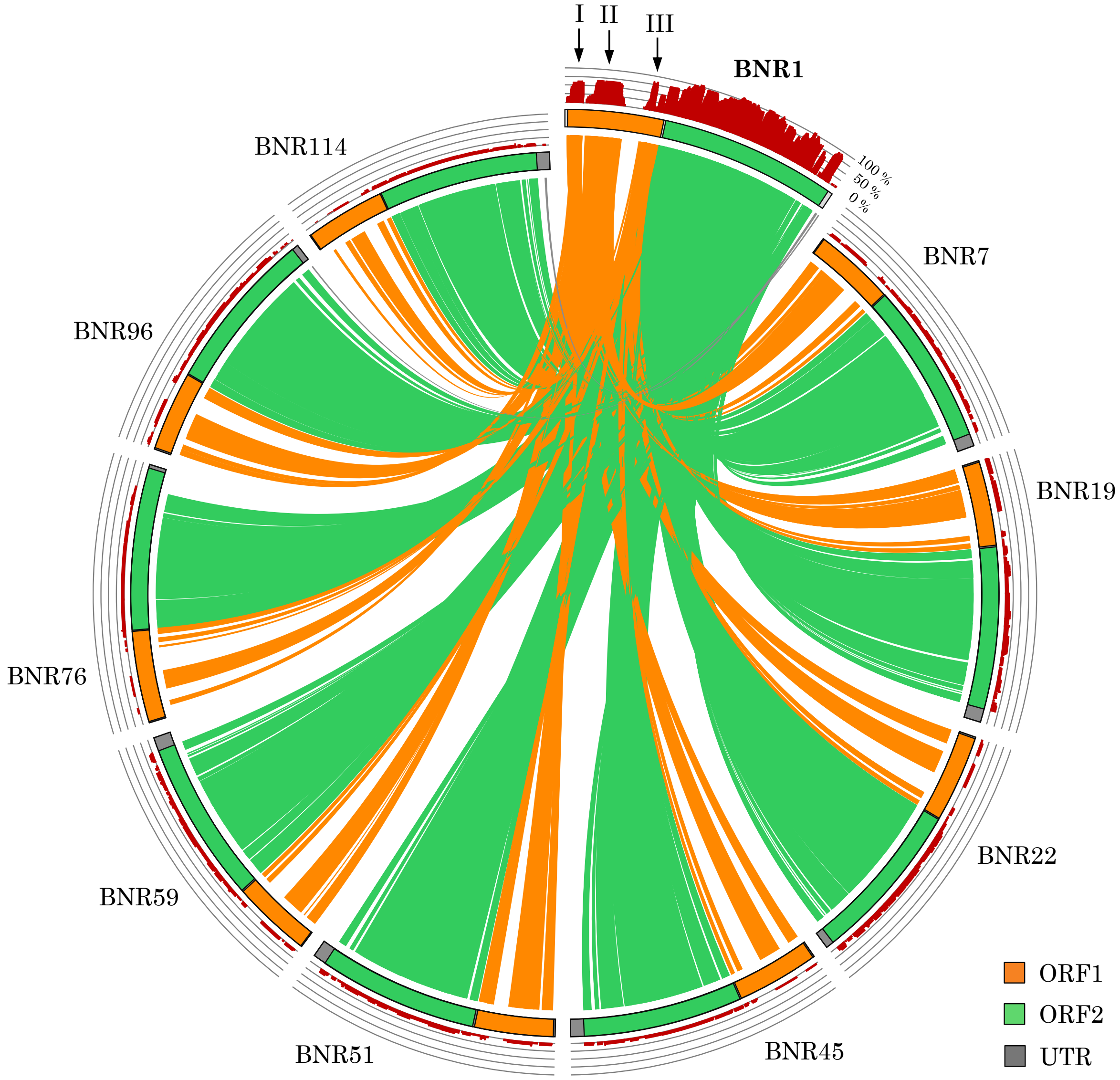


Figure 3.12: Sequence similarity of BNR LINEs.
Sequence similarity of the intact BNR LINEs has been visualized by comparison with the reference element BNR1. Graphical representations of BNR elements are arranged clockwise on a circle, whereby ORFs are represented by colors (orange = ORF1, green = ORF2). In order to define similar regions to BNR1, a *tBLASTx* search against the nine intact BNR elements has been performed. Conserved BNR regions have been detected as *BLAST* hits, and are shown by connecting lines in orange (ORF1 regions) and green (ORF2 regions). Additionally, similarity to BNR1 is summarized by a conservation histogram (red) outside of the cicle of arranged BNRs.
While ORF2 shows continuous conservation, homology in BNR ORF1 is limited to the three single regions I to III (arrows).

The highest degree of BNR sequence conservation has been observed in the coding regions. However, ORF1 and ORF2 similarity vary greatly. BNR ORF2 sequences are highly identical, and harbor enzymatic functions of an endonuclease, a reverse transcriptase and an RNaseH. However, different levels of conservation have been observed for ORF1, which consists of distinct domains (Figure 3.12, arrows), joined by variable linker sequences. The conserved region closest to the ORF1 N-terminus, region I, corresponds to an RNA-binding domain, the RNA recognition motif, present in all BNR LINEs (as will be described in Chapter 3.2.4).

Furthermore, two additional motifs have been identified. Region II is situated in the center of ORF1 and is composed of more than 100 aa, whereas region III marks the C-terminus and contains predominantly basic residues like arginine and lysine. For region II and III, no function could be deduced. However, their conservation in all analyzed sequences implies that they are important for ORF1 function.

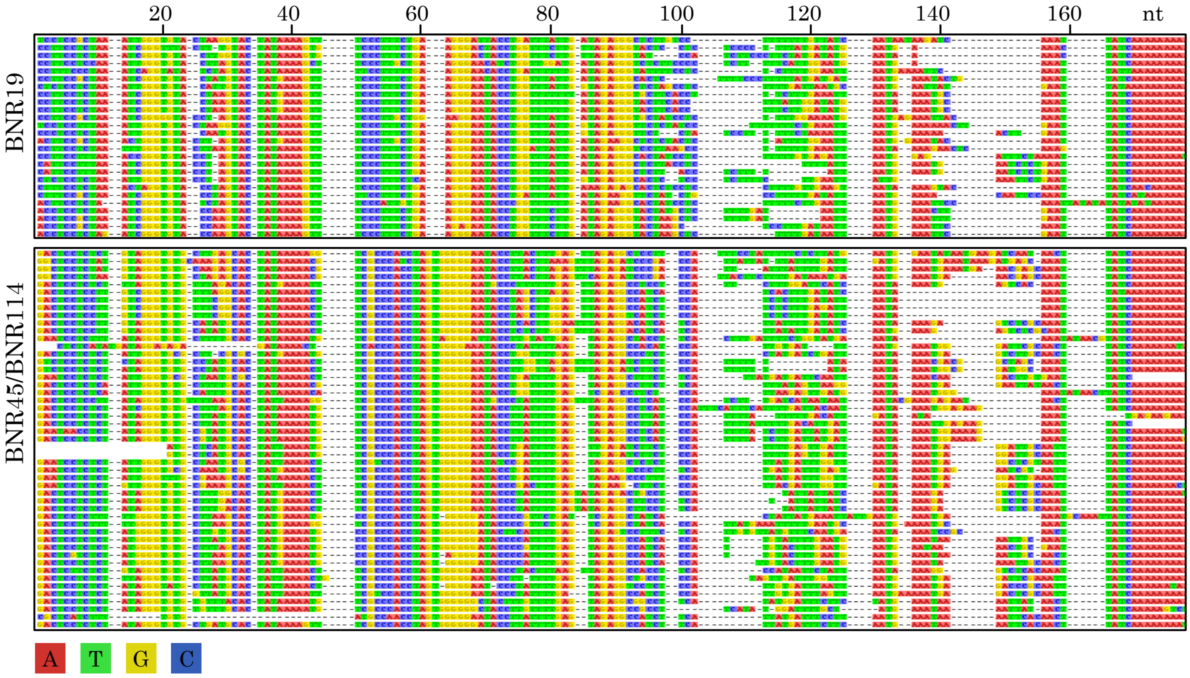


Figure 3.13: Multiple sequence alignment of 3' UTRs similar to BNR19, BNR45 and BNR114. Twenty-five and forty-seven 3' UTR regions similar to BNR19 and BNR45/BNR114, respectively, are shown. The LINE poly(A) tail is indicated at the right by a stretch of adenine residues. (red = A; green = T; yellow = G; blue = C)

Integrity of ORF sequences as has been documented for the intact BNR elements is a prerequisite for transposition. If retrotransposition takes place, most of the newly inserted LINE copies are 5' truncated (Szak *et al.*, 2002). In order to find out, if one of the nine intact BNR elements spawned copies by retrotransposition, a *BLASTn* search was conducted with their 3' UTR regions. Only three of them, BNR19, BNR45 and BNR114, gave multiple *BLAST* hits. Since BNR45 and BNR114 have nearly identical 3' sequences, they produced similar *BLAST* results and therefore, were analyzed in combination. Forty-seven sequences similar to BNR45/BNR114 and twenty-five similar to BNR19 have been identified and aligned (Figure 3.13). Their average identities have been 70.9 % and 74.9 %, respectively. This finding illustrates the formation of BNR subfamilies by single transpositionally active copies. Out of the nine still intact LINEs, BNR19, BNR45 and BNR114 are probably the ones that have been most successful in the generation of novel copies.

3.2.4 A novel subclade of plant LINEs is defined by the RNA recognition motif in ORF1 and present in a variety of plants

A striking feature of BNR retrotransposons is the absence of the typical CCHC-type zinc finger in the ORF1 which is necessary for LINE mRNA binding. Instead, a different secondary structure motif, the RNA recognition motif (RRM), also known as ribonucleoprotein domain (RNP) is located in the N-terminus of ORF1 (Figure 3.11; Figure 3.12).

Using BNR1 as query sequence, nine LINEs of *Populus trichocarpa, Glycine max, Glycine tomentella* and *Lotus japonicus* were identified in the EMBL database which contained a similar N-terminal RNA recognition motif in the ORF1 (for designation see Table 3.2). The alignment with members of the *B. vulgaris* BNR family revealed two regions of conservation in ORF1 (Figure 3.14 and Figure 3.15), that correspond to region I (RRM) and region II in Figure 3.12.

Table 3.2: Selected LINEs harboring an RNA recognition motif in their ORF1

LINE	Organism	Accession	Position[a] [nt]	Strand[b]	Length [bp]	TSD [bp]
BNR1	*B. vulgaris*	EU564339	89 - 6788	+	6700	16
BNR2	*B. vulgaris*	DQ374060 DQ374061	1482 - 6135 1 - 1451	- -	6402	22
BNR3	*B. vulgaris*	DQ374017	14920 - 24242	-	9323	8
Populus3	*P. trichocarpa*	AC209099	110966 - 117825	+	6860	46
Populus5	*P. trichocarpa*	AC210386	~76000 - 82308	+	>6000[c]	--
Populus7	*P. trichocarpa*	AC215899	41011 - 47997	+	6987	17
Lotus2	*L. japonicus*	AP006095	~17500 - 23844	+	>6100[c]	--
Lotus7	*L. japonicus*	AP004952	5334 - 11844	+	6511	16
Lotus8	*L. japonicus*	AP008086	18990 - ~27000	-	>7700[c]	--
Glycine1	*G. max*	AC196857	78680 - ~85100	-	>6100[c]	--
Glycine2	*G. max*	AC152885	149070 - ~155200	-	>6200[c]	--
Glycine3	*G. tomentella*	AC208298	28503 - ~35000	-	>6100[c]	--

[a] Position of the LINE in the sequence

[b] +, upper strand; -, lower strand

[c] LINE length has been given a minimum estimate, because the flanking TSD has not been identified

Close to the N-terminus, a stretch of approximately 80 amino acids forms a computationally predicted ßαßßαß–secondary structure which is typical for RRM (Cole *et al.*, 2008). The degenerate amino acid consensus is very similar to the RRM consensus ([U]- [x]- [U]- [x]$_{2}$- [L]- [x]$_{3\text{-}9}$- [Z]- [x]$_{3\text{-}4}$- [L]- [x]$_{3}$- [F]- [x]$_{3\text{-}4}$]- [G]- [x]- [U]- [x]$_{2}$- [Z]- [x]$_{6\text{-}12}$- [U]- [x]- [V]- [x]- [F]- [x]$_{6\text{-}7}$- [Z]- [x]$_{2}$- [A], with x being any residue, U indicates uncharged residues like L, I, V, A, G, F, W, Y, C, M, and Z = U + S, T) as reported (Birney *et al.*, 1993). More important, the two conserved key motifs RNP1 ([RK]- [G]- [FY]- [ILV]- [X]- [FY]) on the β3-strand and RNP2 ([ILV]- [FY]- [ILV]- [X]- [N]- [L]) on β1-strand (Maris *et al.*, 2005) were identified. The alignment of plant LINE RRMs with typical RRM domains from plant and animal genomes revealed considerable similarity, even outside the conserved amino acid signatures (Figure 3.14).

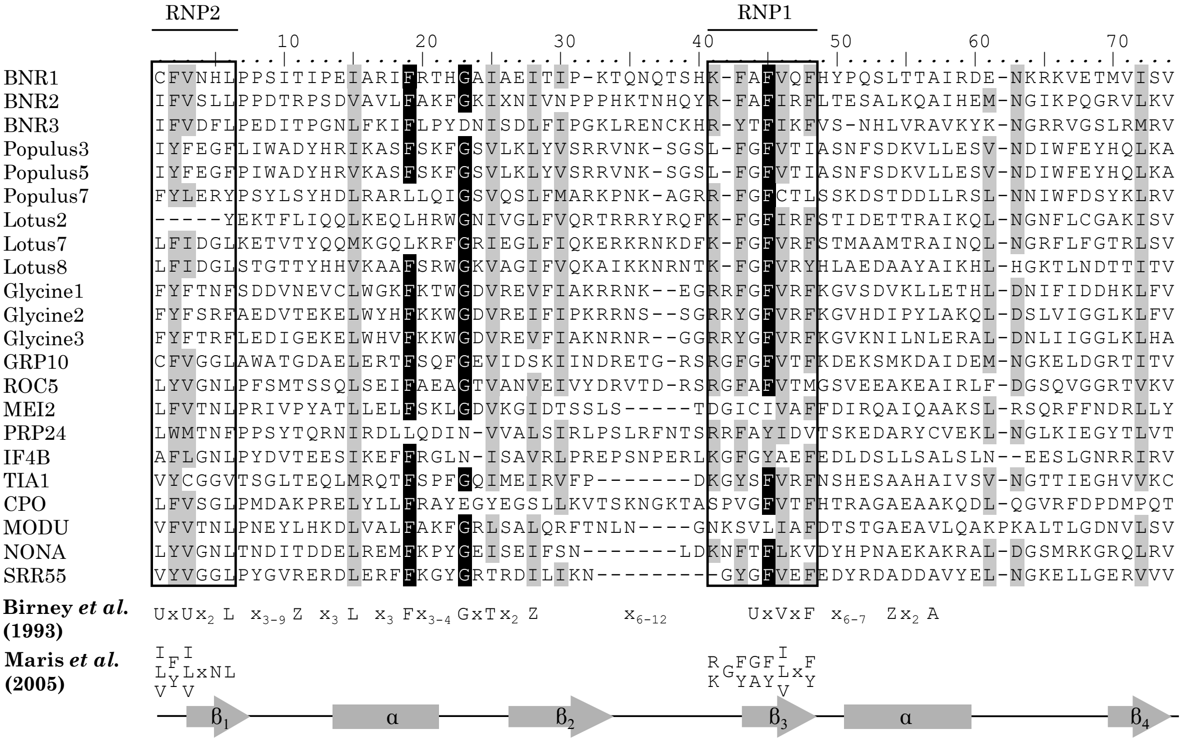

RNP2
RNP1
10
20
30
40
50
60
70
BNR1 CFVNHLPPSITIPEIARIFRTHGAIAEITIP-KTQNQTSHK-FAFVQFHYPQSLTTAIRDE-NKRKVETMVISV
BNR2 IFVSLLPPDTRPSDVAVLFAKFGKIXNIVNPPPHKTNHQYR-FAFIRFLTESALKQAIHEM-NGIKPQGRVLKV
BNR3 IFVDFLPEDITPGNLFKIFLPYDNISDLFIPGKLRENCKHR-YTFIKFVS-NHLVRAVKYK-NGRRVGSLRMRV
Populus3 IYFEGFLIWADYHRIKASFSKFGSVLKLYVSRRVNK-SGSL-FGFVTIASNFSDKVLLESV-NDIWFEYHQLKA
Populus5 IYFEGFPIWADYHRVKASFSKFGSVLKLYVSRRVNK-SGSL-FGFVTIASNFSDKVLLESV-NDIWFEYHQLKA
Populus7 FYLERYPSYLSYHDLRARLLQIGSVQSLFMARKPNK-AGRR-FGFCTLSSKDSTDDLLRSL-NNIWFDSYKLRV
Lotus2 -----YEKTFLIQQLKEQLHRWGNIVGLFVQRTRRRYRQFK-FGFIRFSTIDETTRAIKQL-NGNFLCGAKISV
Lotus7 LFIDGLKETVTYQQMKGQLKRFGRIEGLFIQKERKRNKDFK-FGFVRFSTMAAMTRAINQL-NGRFLFGTRLSV
Lotus8 LFIDGLSTGTTYHHVKAAFSRWGKVAGIFVQKAIKKNRNTK-FGFVRYHLAEDAAYAIKHL-HGKTLNDTTITV
Glycine1 FYFTNFSDDVNEVCLWGKFKTWGDVREVFIAKRRNK--EGRRFGFVRFKGVSDVKLLETHL-DNIFIDDHKLFV
Glycine2 FYFSRFAEDVTEKELWYHFKKWGDVREIFIPKRRNS--SGRRYGFVRFKGVHDIPYLAKQL-DSLVIGGLKLFV
Glycine3 FYFTRFLEDIGEKELWHVFKKWGDVREVFIAKNRNR--GGRRYGFVRFKGVKNILNLERAL-DNLIIGGLKLHA
GRP10 CFVGGLAWATGDAELERTFSQFGEVIDSKIINDRETG-RSRGFGFVTFKDEKSMKDAIDEM-NGKELDGRTITV
ROC5 LYVGNLPFSMTSSQLSEIFAEAGTVANVEIVYDRVTD-RSRGFAFVTMGSVEEAKEAIRLF-DGSQVGGRTVKV
MEI2 LFVTNLPRIVPYATLLELFSKLGDVKGIDTSSLS-----TDGICIVAFFDIRQAIQAAKSL-RSQRFFNDRLLY
PRP24 LWMTNFPPSYTQRNIRDLLQDIN-VVALSIRLPSLRFNTSRRFAYIDVTSKEDARYCVEKL-NGLKIEGYTLVT
IF4B AFLGNLPYDVTEESIKEFFRGLN-ISAVRLPREPSNPERLKGFGYAEFEDLDSLLSALSLN--EESLGNRRIRV
TIA1 VYCGGVTSGLTEQLMRQTFSPFGQIMEIRVFP-------DKGYSFVRFNSHESAAHAIVSV-NGTTIEGHVVKC
CPO LFVSGLPMDAKPRELYLLFRAYEGYEGSLLKVTSKNGKTASPVGFVTFHTRAGAEAAKQDL-QGVRFDPDMPQT
MODU VFVTNLPNEYLHKDLVALFAKFGRLSALQRFTNLN----GNKSVLIAFDTSTGAEAVLQAKPKALTLGDNVLSV
NONA LYVGNLTNDITDDELREMFKPYGEISEIFSN-------LDKNFTFLKVDYHPNAEKAKRAL-DGSMRKGRQLRV
SRR55 VYVGGLPYGVRERDLERFFKGYGRTRDILIKN---------GYGFVEFEDYRDADDAVYEL-NGKELLGERVVV
Birney et al. (1993) UxUx2 L x3-9 Z x3 L x3 Fx3-4 GxTx2 Z x6-12 UxVxF x6-7 Zx2 A
Maris et al. (2005) I/L/V F/Y I/L/V xNL R/K G F/Y G/A F/Y I/L/V x F/Y
ß1
α
ß2
ß3
α
ß4

Figure 3.14: The conserved RNA recognition motif in the ORF1 of BNR subclade members. LINE ORF1 amino acid sequences of BNR elements and similar elements from poplar, lotus and soybeans were aligned and the percentage of amino acid conservation was illustrated by 67 % conservation shading. Black boxes indicate identical and light grey similar residues. Below, results of Jpred3 secondary structure prediction are shown. Rectangles represent α-helices while arrows indicate ß-sheets. For comparison, RRM from different proteins of the following species were shown as well: *Brassica napus* (GRP10), *Nicotinia sylvestris* (ROC5), *Saccharomyces pombe* (MEI2), *Saccharomyces cerevisiae* (PRP24), *Homo sapiens* (IF4B and TIA1), *Drosophila melanogaster* (CPO). The RRM consensus sequences of Birney *et al.* (1993) and Maris *et al.* (2005) were included.

Furthermore, downstream of the RRM, a second conserved amino acid motif encoded in the central region of ORF1 was identified (Figure 3.15), however, its function remains elusive. A different α/β–fold was predicted for this sequence, and especially the β-sheets show a high degree of conservation. The presence of both ORF1 domains is a characteristic feature of BNR-like LINEs present in the genomes of higher plants.

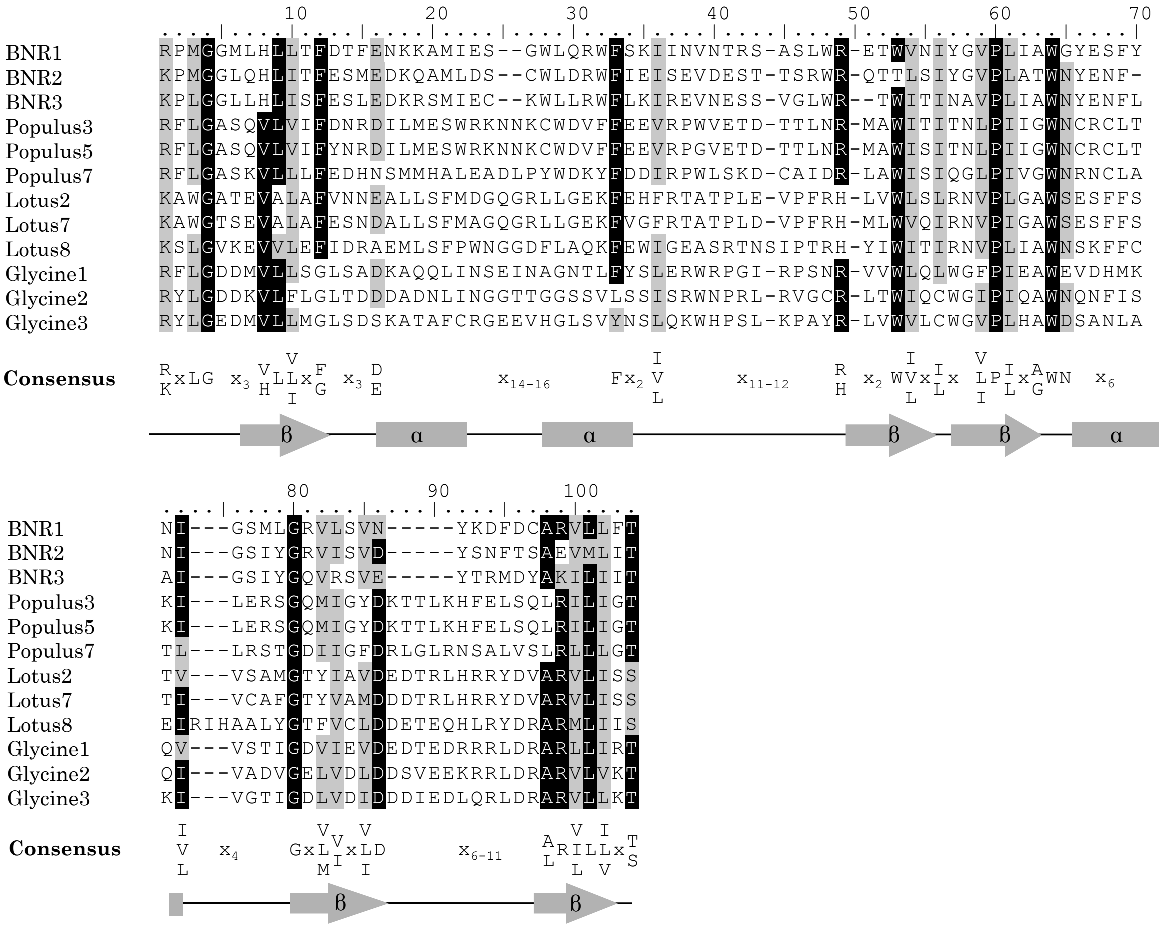

10 20 30 40 50 60 70
BNR1 RPMGGMLHLLTFDTFENKKAMIES--GWLQRWFSKIINVNTRS-ASLWR-ETWVNIYGVPLIAWGYESFY
BNR2 KPMGGLQHLITFESMEDKQAMLDS--CWLDRWFIEISEVDEST-TSRWR-QTTLSIYGVPLATWNYENF-
BNR3 KPLGGLLHLISFESLEDKRSMIEC--KWLLRWFLKIREVNESS-VGLWR--TWITINAVPLIAWNYENFL
Populus3 RFLGASQVLVIFDNRDILMESWRKNNKCWDVFFEEVRPWVETD-TTLNR-MAWITITNLPIIGWNCRCLT
Populus5 RFLGASQVLVIFYNRDILMESWRKNNKCWDVFFEEVRPGVETD-TTLNR-MAWISITNLPIIGWNCRCLT
Populus7 RFLGASKVLLLFEDHNSMMHALEADLPYWDKYFDDIRPWLSKD-CAIDR-LAWISIQGLPIVGWNRNCLA
Lotus2 KAWGATEVALAFVNNEALLSFMDGQGRLLGEKFEHFRTATPLE-VPFRH-LVWLSLRNVPLGAWSESFFS
Lotus7 KAWGTSEVALAFESNDALLSFMAGQGRLLGEKFVGFRTATPLD-VPFRH-MLWVQIRNVPIGAWSESFFS
Lotus8 KSLGVKEVVLEFIDRAEMLSFPWNGGDFLAQKFEWIGEASRTNSIPTRH-YIWITIRNVPLIAWNSKFFC
Glycine1 RFLGDDMVLLSGLSADKAQQLINSEINAGNTLFYSLERWRPGI-RPSNR-VVWLQLWGFPIEAWEVDHMK
Glycine2 RYLGDDKVLFLGLTDDDADNLINGGTTGGSSVLSSISRWNPRL-RVGCR-LTWIQCWGIPIQAWNQNFIS
Glycine3 RYLGEDMVLLMGLSDSKATAFCRGEEVHGLSVYNSLQKWHPSL-KPAYR-LVWVLCWGVPLHAWDSANLA
Consensus
β
α
α
β
β
α
80 90 100
BNR1 NI---GSMLGRVLSVN-----YKDFDCARVLLFT
BNR2 NI---GSIYGRVISVD-----YSNFTSAEVMLIT
BNR3 AI---GSIYGQVRSVE-----YTRMDYAKILIIT
Populus3 KI---LERSGQMIGYDKTTLKHFELSQLRILIGT
Populus5 KI---LERSGQMIGYDKTTLKHFELSQLRILIGT
Populus7 TL---LRSTGDIIGFDRLGLRNSALVSLRLLLGT
Lotus2 TV---VSAMGTYIAVDEDTRLHRRYDVARVLISS
Lotus7 TI---VCAFGTYVAMDDDTRLHRRYDVARVLISS
Lotus8 EIRIHAALYGTFVCLDDETEQHLRYDRARMLIIS
Glycine1 QV---VSTIGDVIEVDEDTEDRRRLDRARLLIRT
Glycine2 QI---VADVGELVDLDDSVEEKRRLDRARVLVKT
Glycine3 KI---VGTIGDLVDIDDDIEDLQRLDRARVLLKT
Consensus
β
β

Figure 3.15: A conserved domain identified in ORF1 of BNR subclade members.
LINE ORF1 amino acid sequences of BNR elements and similar elements from poplar, lotus and soybeans were aligned and the percentage of amino acid conservation was illustrated by 67 % conservation shading. Black boxes indicate identical and light grey similar residues. Below, results of Jpred3 secondary structure prediction are shown. Rectangles represent α-helices while arrows indicate β-sheets.

Despite their completely different ORF1 regions, the ORF2 sequences of BNR-like LINEs and all known plant LINEs share the typical domains for endonuclease, reverse transcriptase and RNaseH. An alignment of the reverse transcriptase regions according to Malik *et al.* (1999) and Permanyer *et al.* (2003) has been performed, followed by construction of a dendrogram showing their relationship using the Neighbor-Joining method (Saitou and Nei, 1987) (Figure 3.20). For comparison, retrotransposon sequences from four different LINE clades (R2, RTE, Jockey and I) were used. BNR-like LINEs cluster together, along with all plant non-LTR retrotransposons and a set of vertebrate L1 LINEs demonstrating their assignment to L1 LINEs. Plant LINEs are arranged separately from mammalian L1 LINEs, and BNR-like retrotransposons form a distinct subtree in the plant LINE group. This shows that BNR-like LINEs not only have a similar ORF1 structure, but significant similarities in their ORF2 sequences as well. They are distinguishable from other plant LINEs and form an L1 subclade, which was designated BNR subclade.

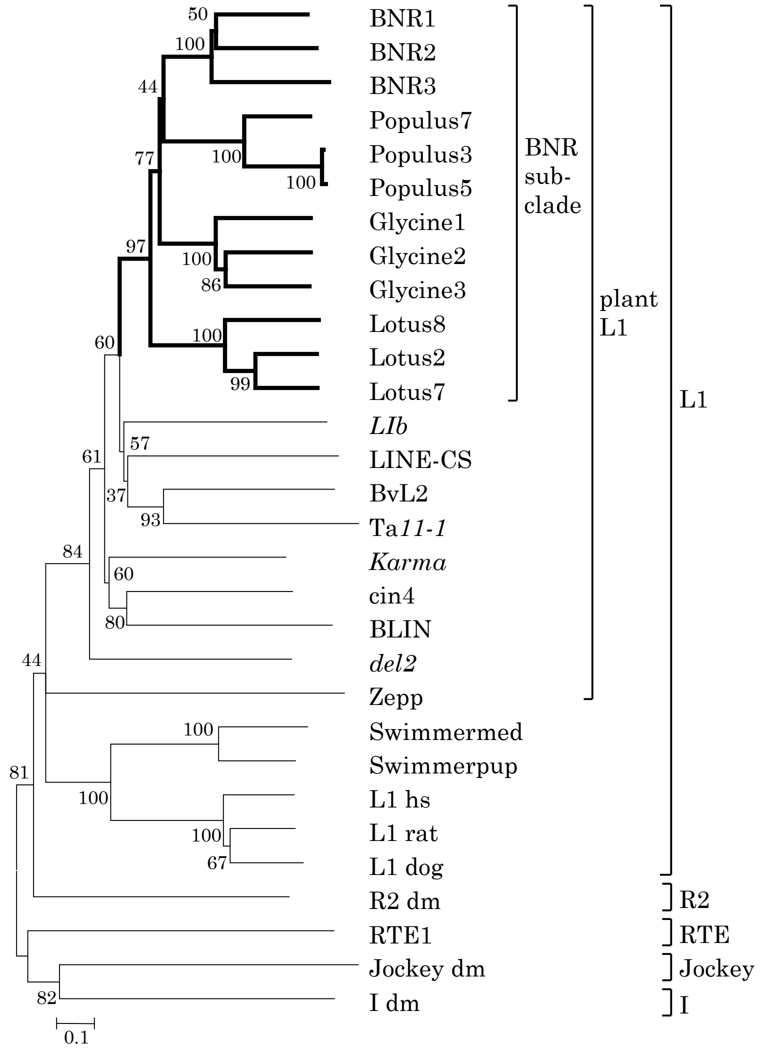


Figure 3.16: Dendrogram showing the relatedness of 30 ORF2 reverse transcriptase sequences. BNR-like retroelements from sugar beet, poplar, lotus and soybean form a separate group indicating the BNR subclade (bold). Included plant LINE sequences originate from the genomes of *Ipomoea batatas* (*LIb*), *Cannabis sativa* (LINE-CS), *Arabidopsis thaliana* (Ta*11-1*), *Beta vulgaris* (BvL), *Oryza sativa* (*Karma*), *Hordeum vulgare* (BLIN), *Zea mays* (cin4), *Lilium speciosum* (*del2*) and *Chlorella vulgaris* (Zepp). For comparison, the LINE ORF2 sequences of the L1, R2, RTE1, Jockey and I clade were analyzed from: *Oryzias latipes* (Swimmermed) and *Cyprinodon macularius* (Swimmerpub), *Homo sapiens* (L1 hs), *Rattus norvegicus* (L1 rat), *Canis lupus familiaris* (L1 dog), *Drosophila melanogaster* (R2, I, Jockey) and *Caenorhabditis elegans* (RTE1). The dendrogram was conducted by application of the Neighbor-Joining algorithm. Branch lengths are proportional to genetic distance. Bootstrap values are indicated as a percentage of 1000 replicates. The scale bar represents 0.1 substitutions per site.

3.2.5 Recent transposition of BNR1

Sequence analysis of the BNR1 flanking region shows integration of BNR1 into an array of subtelomeric pAv satellite repeats of *B. vulgaris* (Dechyeva and Schmidt, 2006). By PCR with an outward-facing primer of the BNR 5' end and a primer binding in the satellite pAv, a 1 kb PCR product specific for this transposition event was amplified (Figure 3.17 A). Sequencing of two clones from this amplicon revealed only integration of BNR1 in pAv, however, it cannot be excluded that BNR integration into pAv arrays on other chromosomes has occurred. This PCR product was identified also in the cultivar fodder beet, however, not in any other *Beta vulgaris* cultivar or in other species from the section *Beta* (Figure 3.17 B). Low molecular weight amplicons of approximately 400 bp originate from degenerated pAv repeats. Since the integration of BNR1 in pAv is unique for sugar and fodder beet, this indicates a relatively young transposition during the domestication, breeding and diversification of beet.

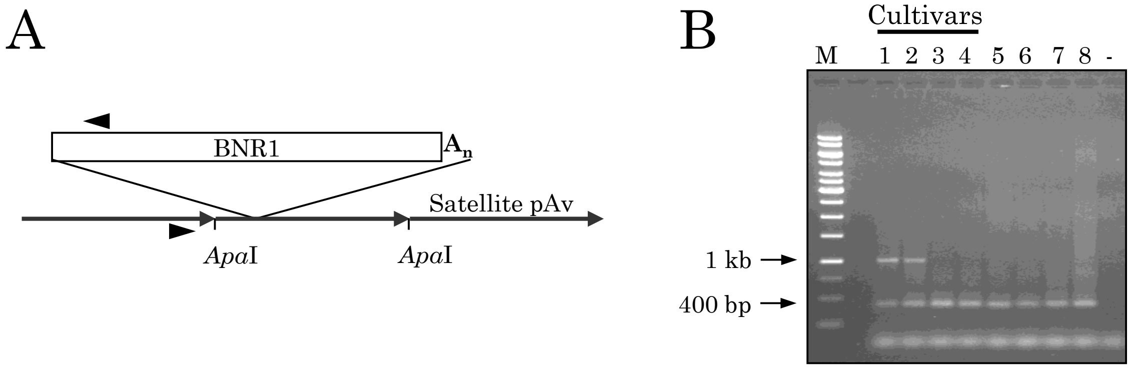


Figure 3.17: Integration site of BNR1 and similar LINEs.
(A) Schematic representation of the genomic integration of BNR1 in pAv. Arrow heads indicate primers used for the amplification of the integration site.
(B) The integration event of BNR1 into pAv (1 kb) was only detected in the cultivars sugar beet (1) and fodder beet (2), but not in cultivars such as garden beet (3) and chard (4) and also not in wild species of the section *Beta*: *Beta vulgaris* ssp. *maritima* (5), *Beta vulgaris* ssp. *adanensis* (6), *Beta macrocarpa* (7) and *Beta patula* (8). The second amplicon at 400 bp results from unspecific binding of the BNR primer to diverged pAv satellite repeats.

3.2.6 Truncation and methylation of LINEs of the BNR family

Target-primed reverse transcription often results in 5' truncation of LINEs during transposition. In order to investigate the extent of BNR truncation, comparative Southern hybridization to restricted genomic DNA was performed (Figure 3.18). Probes of the 5' end at position 628-831 bp, of the RT region at position 3868-4121 bp and of the 3' end at position 6209-6490 bp of BNR1 as indicated in Figure 3.10 were used. Only a few signals were detected by hybridization of the 5' end probe. However, an increase ranging from weak signals to a smear of strong hybridization signals was observed when using probes situated towards the 3' end. Fragments showing a strong distinct signal indicate conserved restriction sites in BNR or in the flanking regions. The increase in signals shows that the genome of *B. vulgaris* contains less full-length BNR1 copies than BNR1-derived 3' ends. This corresponds to the observation that LINEs can be heavily 5' truncated due to a premature abortion of reverse transcription during retrotransposition.

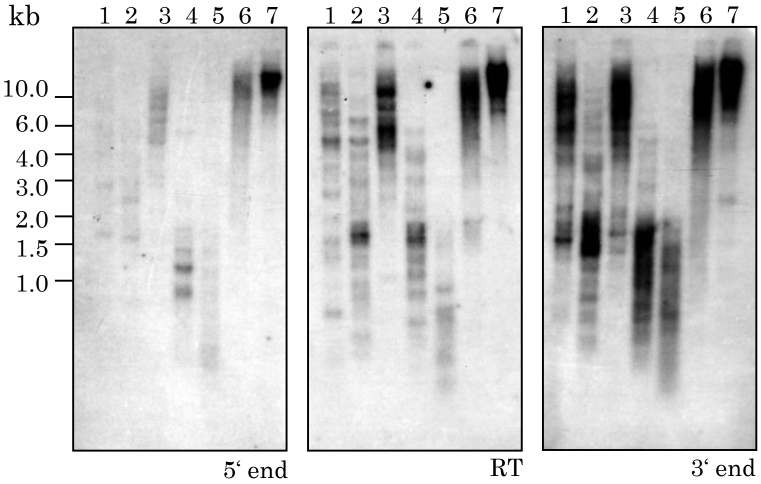


Figure 3.18: Southern hybridization of genomic *B. vulgaris* DNA with probes from different regions of BNR1.
Genomic DNA of *B. vulgaris* ssp. *vulgaris* was restricted with *Hin*dIII (1), *Dra*I (2), *Xba*I (3), *Rsa*I (4), *Alu*I (5), *Msp*I (6), *Hpa*II (7) and probed with sequences from three BNR1 regions as indicated in Figure 3.10.

In order to analyze cytosine methylation of BNR members, hybridization of genomic DNA restricted with *Hpa*II and *Msp*I was compared (Figure 3.18, lanes 6 and 7). While *Hpa*II cuts only unmethylated CCGG sequences, *Msp*I is able to tolerate methylation of the internal cytosine. When probing *Hpa*II digested DNA, hybridization of very large DNA fragments which are not resolved by conventional gel electrophoresis was detectable, indicating strong methylation of one or both cytosines of CCGG sites of BNR family members. In contrast, *Msp*I restricts genomic DNA into smaller fragments detectable as a smear of signals in Southern hybridization. The results show that most BNR copies and the adjacent genomic DNA are methylated at inner cytosines, while some outer cytosines in CCGG context are not methylated.

3.2.7 Chromosomal localization of BNR LINEs

The localization of BNR-like LINEs along *B. vulgaris* chromosomes was investigated by fluorescent *in situ* hybridization (FISH). A probe spanning the relatively conserved RT domains II and III of BNR1 was labelled with biotin-11-dUTP by PCR using BAC 47M6 as template and hybridized to metaphase chromosomes. It allowed the detection of full-length BNR-like LINEs and copies which are truncated upstream of the RT gene. The FISH images show signals of varying intensity with dispersed hybridization on all chromosomes (Figure 3.19). Strong signals originate most likely from BNR clusters preferentially located at the intercalary heterochromatin and in most, but not all centromeres. However, clustering has not been observed at sequence level. Consistent with the genomic association of BNR1 with the subtelomeric satellite family pAv, weak signals are also detected in the distal DAPI-negative euchromatic regions (arrows in Figure 3.19 A).

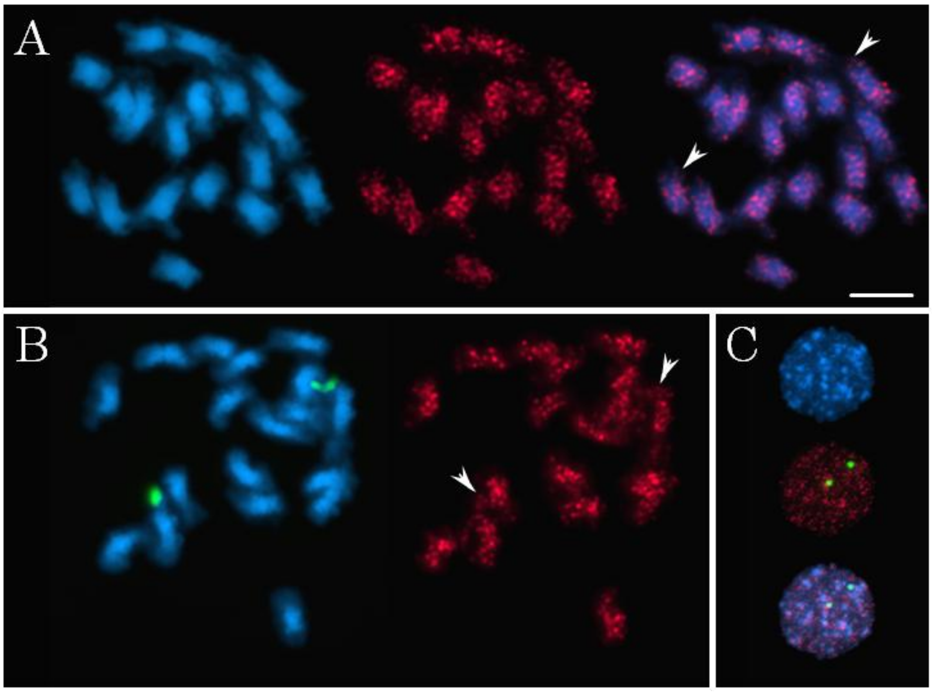


Figure 3.19: Physical mapping of BNR elements along *B. vulgaris* chromosomes.
Blue fluorescence shows DNA stained with DAPI, green fluorescence marks the 18S-5.8S-25S rRNA genes on chromosome 1, while red fluorescence shows hybridization with the same BNR1 reverse transcriptase probe as used for Southern hybridization. The scale bar corresponds to 10 µm.
(A) Hybridization of mitotic chromosomes visualizing the clustered organization of BNR LINEs. Most copies are located in the intercalary heterochromatin while some BNR LINEs are also integrated at subterminal positions (arrows).
(B) Double-color FISH shows single BNR copies in the 18S-5.8S-25S rRNA genes (arrows).
(C) Signals adjacent to brightly stained heterochromatic regions are visible at interphase nuclei.

Double-color FISH with an 18S-5.8S-25S rDNA probe (green fluorescence) shows that copies of BNR-like retrotransposons are integrated in the rRNA gene arrays located on chromosome 1 (arrowed in Figure 3.19 B). At higher resolution at interphase nucleus, a disperse localization, mostly in euchromatic DAPI-negative regions can be observed (Figure 3.19 C).

3.2.8 Diversity of the BNR family in the genera *Beta* and *Patellifolia*

In order to investigate the genomic distribution and organization of BNR family members in beet and relatives, Southern hybridization was carried out to *Hin*dIII restricted DNA from representative species of the genera *Beta* (sections *Beta*, *Corollinae* and *Nanae*) and *Patellifolia*. The region covering the RT domains II and III of BNR1 was used as a probe. Hybridization signals were visible in all species tested, including the outgroup species spinach, also belonging to the Amaranthaceae, indicating the widespread presence of BNR (Figure 3.16). However, the strongest signals and many shared fragments were observed for the *Beta* cultivars (lanes 1-5) indicating conserved restriction sites in the BNR RT sequence or the flanking region.

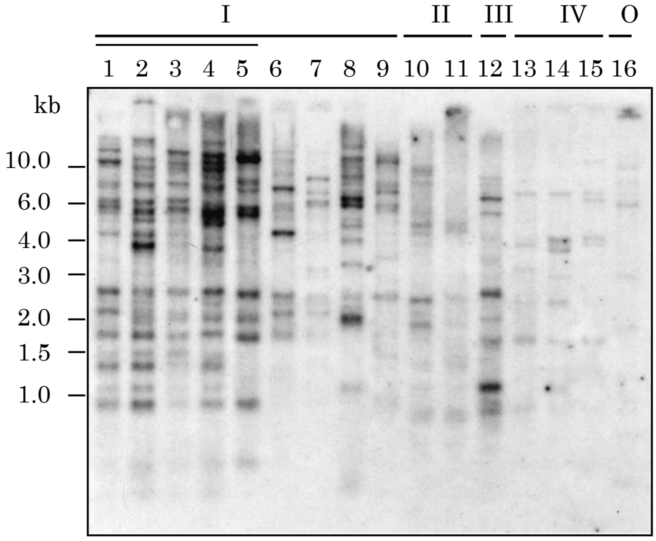


Figure 3.20: Distribution and organization of BNR copies in the genera *Beta* and *Patellifolia*. Genomic *Hin*dIII-restricted DNA was analyzed by comparative Southern hybridization using a probe from the BNR1 reverse transcriptase. BNR is present in all species of both genera showing a different genomic organization. Species tested were: Cultivars of *B. vulgaris* ssp. *vulgaris* in the section *Beta* (I): Rosamona (1), KWS 2320 (2), fodder beet Brigadier (3), fodder beet Eckdorot (4), chard (5); and wild beet species from the section *Beta* (I): *B. vulgaris* ssp. *maritima* (6), *B. vulgaris* ssp. *adanensis* (7), *B. macrocarpa* (8), *B. patula* (9); species of the section *Corollinae* (II): *B. corolliflora* (10), *B. macrorhiza* (11); species of the section *Nanae* (III): *B. nana* (12); species of the genus *Patellifolia* (IV): *P. procumbens* (13), *P. patellaris* (14), *P. webbiana* (15); outgroup species (O): *Spinacia oleracea* (16).

There are also distinct differences in the hybridization pattern for each cultivar. A reduced hybridization was detected for genomes of wild beet species of the sections *Beta*, *Corollinae* and *Nanae* indicating either a lower copy number or a higher divergence of

BNR-like LINEs. In the wild beet genus *Patellifolia*, formerly section *Procumbentes* of *Beta*, only faint signals were observed.

For a detailed insight into BNR RT diversity across species borders, sequences of the RT gene of BNR-like LINEs were isolated by PCR. Four amplicons of *B. vulgaris*, *B. adanensis*, *B. macrocarpa*, *B. patula*, *B. corolliflora*, *B. nana* and *P. patellifolia* have been cloned and sequenced. Deduced amino acid sequences, including the characterized LINEs *Karma*, *LIb*, BvL2, Ta*11-1*, BLIN, cin4, *del2* and Zepp were compared in a multiple sequence alignment (Appendix 4). RT sequences of BNR and the other LINEs share similarities only in their RT domains, however vary in all other amino acid residues. For quantification of BNR diversity in the genera *Beta* and *Patellifolia*, the sequences have been analyzed by pairwise sequence comparison, resulting in an identity matrix. The average, minimum and maximum sequence identities are presented in Figure 3.21.

Genus/ Section	**Species**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**O**
Beta	**I** *B. vulgaris*	97 **66** 58							
	II *B. adanensis*	62 **58** 50	99 **95** 90						
	III *B. macrocarpa*	68 **61** 53	65 **60** 48	99 **65** 54					
	IV *B. patula*	86 **67** 55	62 **58** 51	91 **64** 51	100 **67** 58				
Corollinae	**V** *B. corolliflora*	77 **67** 56	69 **62** 54	72 **66** 53	75 **68** 61	94 **75** 68			
Nanae	**VI** *B. nana*	83 **67** 59	64 **61** 55	99 **68** 54	90 **67** 59	75 **70** 63	99 **72** 65		
Patellifolia	**VII** *P. procumbens*	69 **65** 59	60 **58** 55	69 **63** 55	72 **64** 60	73 **69** 64	76 **70** 61	100 **81** 62	
Outgroups	**O** *Outgroups*	36 **31** 19	37 **30** 19	37 **31** 15	40 **32** 18	39 **32** 18	37 **32** 18	39 **32** 18	38 **28** 17

Average identity from [%]	0	10	20	30	40	50	60	70	80	90

Figure 3.21: Sequence identity of BNR-like reverse transcriptase sequences.
By PCR using genomic DNA of species from the beet genera *Beta* and *Patellifolia* 1 kb amplicons were amplified and four clones of each species were sequenced. The amino acid sequence was deduced and a fragment of RT domain IV to VII (containing 172 amino acids) was aligned and analyzed. The bold number indicates the average amino acid similarity, while the upper number marks the maximal and the lower number the minimal sequence similarity in percent. The characterized LINEs *Karma*, *LIb*, BvL2, Ta*11-1*, BLIN, cin4, *del2* and Zepp were used as outgroups.

RT sequences of the beet genera show homology to the full-length BNR LINEs which occur in all sections and are diverse with an identity ranging from 48 % to 100 %

(average 68 % similarity). However, their average intra-species identity is similar to their average inter-species identity ranging between approximately 60-70 %. Highest sequence identities were observed among *B. adanensis* and among *P. procumbens* BNR sequences.

As already shown by the multiple sequence alignment (Appendix 4), the BNR retrotransposons are significantly different from other characterized plant LINEs. An only low sequence identity (average 31 %) to BNR RTs has been observed. The novelty of BNR LINEs in plants has been further validated by those sequence differences.

3.3 Plant retrotransposon analysis on a genomic scale

Results from this chapter have been published in peer-reviewed journals (Wollrab et al., 2012; Weber et al., 2013; Heitkam et al. 2014), all of them after this thesis was defended.

With the advent of next generation sequencing methods (reviewed in Mardis, 2008), an increasing amount of genomic data is generated. During preparation of this thesis, three genome assemblies of *B. vulgaris* became available, from here on referred to as *RefBeet* 0.1.1, *RefBeet* 0.2 and *RefBeet* 0.4 (Table 2.8).

A large quantity of information on transposable elements is hidden in these sequence databases. The goal has been to extract this information and to generate an overview of the number and structure of retrotransposon families in beet. A conventional approach would have been the usage of *BLAST* or *FASTA* alignment algorithms in order to scan for low-matching homologues of known repeats like BNR, BvL or Cotzilla. This method is time-consuming, has a high rate of false-positives to be eliminated manually, and, more importantly, easily misses candidates of very low similarity.

Therefore, advanced methods are required to retrieve the desired information and to make them accessible for detailed analysis. An approach applying a Hidden Markov Model (HMM) was chosen for large scale analysis (see Chapter 2.3.4). With this method, conserved domains can be identified even if they are only remotely similar to an amino acid query. As basis for the HMM construction, not a single sequence is used, but a whole alignment. The query will have properties of a profile or consensus sequences, and additionally, will consider stretches of commonly occuring consecutive residues. Reverse transcriptase genes are especially well-suited for this approach, because they consist of short stretches of highly conserved domains typical for the retrotransposon order, but also of variable linkers which define the families.

3.3.1 Application of retrotransposon Hidden Markov Models

Four HHMs have been constructed which are sensitive to reverse transcriptase sequences of LINEs or of Ty3-*gypsy*, Ty1-*copia* or BEL-Pao LTR retrotransposon (Chapter 2.3.4.1).

3.3.1.1 Parameter calibration

After performance of a *HMMER* search, quality of the detected sequences can be measured by either the expectation value (e-value) or the bit score. An e-value is the number of hits that would be expected to have a score equal to or better than this by chance alone. A good e-value is much less than 1, for example, an e-value of 0.01 would mean that on average 1 false positive would be expected in every 100 searches with different query sequences. Bit scores are defined as logarithm of the ratio of the sequence's probability according to the profile (homology hypothesis) to the null model probability (non-homology hypothesis): The higher the bit score, the higher the propability of a true positive hit (Eddy, 1998; *HMMER*3 manual). While the database size influences the e-value, it does not have an effect on the bit score. Therefore, bit scores can be used to compare hits of different datasets.

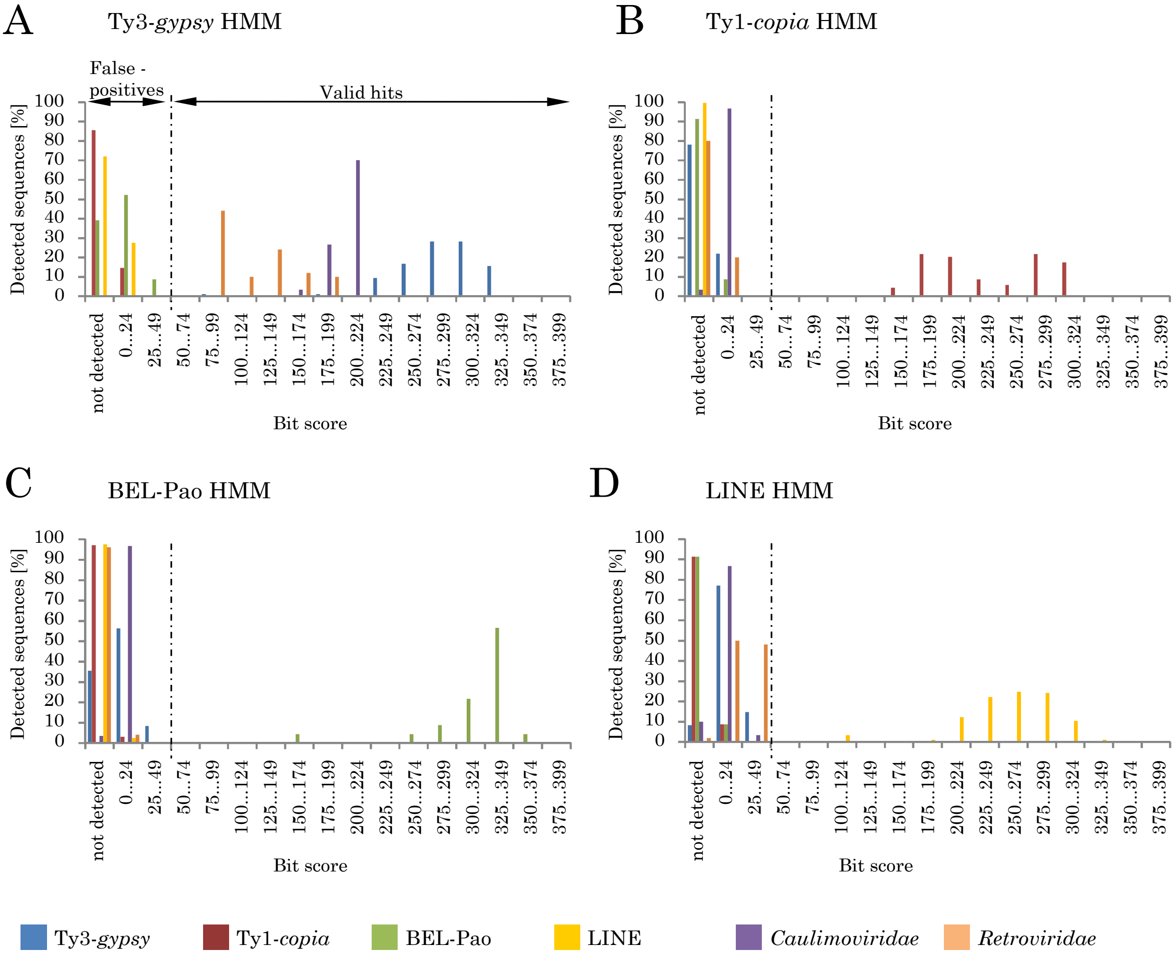


Figure 3.22: Calibration of four HMMs specific for reverse transcriptase sequences. Six databases containing RT sequences of different retrotransposon orders have been queried with each HMM. The number of hits having a specific bit score is presented in the diagrams A-D. The different datasets are color-coded (Ty3-*gypsy*, 96 sequences = blue; Ty1-*copia*, 69 sequences = red; *caulimoviridae*, 30 sequences = violet; LINE, 211 sequences = yellow; BEL-Pao, 23 sequences = light green; *retroviridae*, 50 sequences = orange). For further analysis a minimum bit score of 50 has been applied.

(A) The Ty3-*gypsy* HMM is not solely specific for Ty3-*gypsy* retrotransposons. Results with a bit score greater than 50 could also be of the related retrovirus or caulimovirus order.

(B) Only Ty-*copia* retrotransposons have had a bit sore greater than 50 by application of the Ty1-*copia* HMM.

(C) If a bit score minimum of 50 was considered, the BEL-Pao HMM gave only BEL-Pao RTs.

(D) After application of a bit score minimum of 50, the LINE HMM delivered only LINE RTs.

At first, the probability of random false-positive hits was tested by querying a shuffled and translated version of *RefBeet* 0.1.1 containing all residues, but in a random order. No sequences have been detected, indicating that random false-positives are not to be expected.

A second test was performed to measure the sensitivity of the retrotransposon models to different datasets containing RT amino acid sequences (referenced in Chapter 2.3.4.3). The detected sequences have been split up according to their bit scores in intervals of 25 (Figure 3.22). Using a bit score threshold of 50, searches with every HMM produced only sequences corresponding to the respective retrotransposons. Only the the Ty3-*gypsy* model also picked up related RTs of retro- and caulimoviruses. Therefore, no false-positives hits are to be expected when selecting only search outcomes with a minimum bit score of 50.

3.3.1.2 Suitability of *RefBeet* databases for retrotransposon analysis

In order to assess the suitability of the three *RefBeet* databases for retrotransposon identification, these have been queried with the models sensitive to LINEs or to Ty3-*gypsy*, Ty1-*copia* or BEL-Pao retrotransposons. No BEL-Pao-like sequences have been identified in the genome sequences, however, there have been many positive hits using the three remaining models.

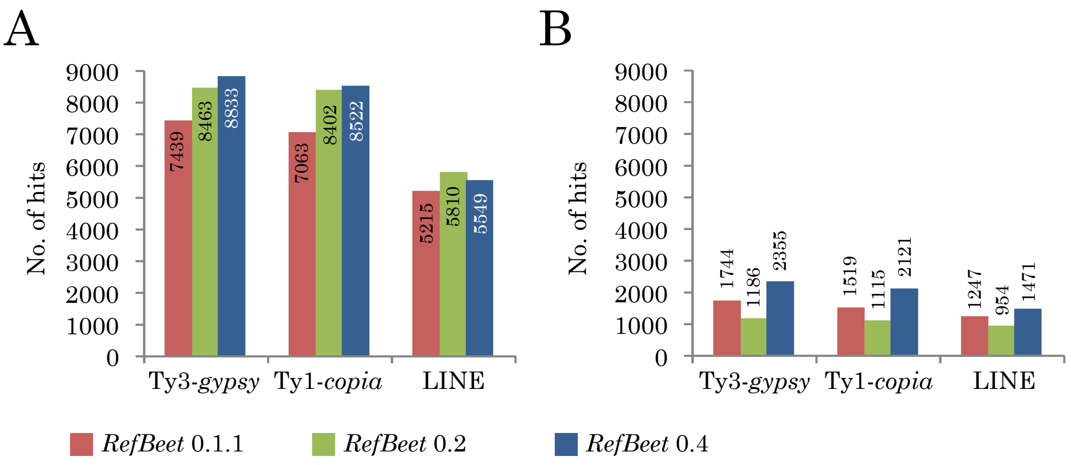


Figure 3.23: Comparative retrotransposon detection using all three *RefBeet* databases. The Ty3-*gypsy*, Ty1-*copia* and LINE HMMs have been used to query *RefBeet* 0.1.1 (blue), *RefBeet* 0.2 (red) and *RefBeet* 0.4 (light green). The number of hits is presented for every HMM and every database.
(A) Sequence hits with a bit score ≥ 50 are shown for each HMM and database.
(B) The fraction of sequence hits with a bit score ≥ 50 as well as a maximum length difference of 30 amino acids to the corresponding model is shown for each HMM and database.

In Figure 3.23 A, all hits with a bit score greater than 50 are presented for every model and database. The highest number of all retrotransposon hits occurs in *RefBeet* 0.2 and *RefBeet* 0.4. However, these numbers include also very short regions of homology that are impossible to use for sequence comparison in multiple alignments and generation of dendrograms. Therefore, these results have been filtered again to include only reverse transcriptase sequences that contain all domains and are fully homologous to the respective HMM. This was achieved by limiting the allowed length difference between the sequence hit and HMM to a maximum of 30 amino acids. This measure reduced the amount of data by elimination of low quality short reads (Figure 3.23 B). Database comparison showed that *RefBeet* 0.2 contains only few full-length RT sequences, probably a result of poor sequence assembly. This is in line with the fact that in comparison with the other databases, *RefBeet* 0.2 sequences have by far the shortest average contig lengths (compare with Table 2.8). This makes *RefBeet* 0.1.1 or the latest assembly, *RefBeet* 0.4, the choice for retrotransposon detection. Apart from this technical information, it is possible to deduce that Ty1-*copia* and Ty3-*gypsy* retrotransposon RTs are present in similar number in all *RefBeet* databases. The number of LINEs obtained by analysis of the same datasets has been lower. It is however not possible to draw conclusions concerning the total number of retrotransposon RTs in the *B. vulgaris* genome.

3.3.2 Overview of the retrotransposon landscape in the *B. vulgaris* genome

In order to get an impression of *B. vulgaris* retrotransposon diversity and family structure, graphical representations of all retrotransposon orders have been created. The *RefBeet* 0.4 database has been queried with the Ty3-*gypsy*, Ty1-*copia* and LINE HMMs, respectively. *HMMER* hits have been filtered according to Figure 3.22 B to yield 2355 Ty3-*gypsy*, 2121 Ty1-*copia* and 1471 LINE RT amino acid sequences, respectively.

Already characterized *B. vulgaris* retrotransposons (Chapters 1.3.3, 3.1 and 3.2) along with elements described by Kapitonov *et al.* (2009) and Llorens *et al.* (2009) have been used for classification of the detected reverse transcriptase sequences. *MUSCLE* alignment, followed by Neighbor-Joining analysis was used to generate an overview of the *B. vulgaris* retrotransposon landscape (Figure 3.24, Figure 3.25 and Figure 3.26). A summary of the detected retrotransposon lineages is presented in Table 3.3.

Table 3.3: Number of detected *B. vulgaris* reverse transcriptases by a HMM-based approach

HMM used for detection	**Lineage**	**No. of detected RTs**[a]	**Classification based on reference**
Ty3-*gypsy*	Chromoviruses: CRM	207	Llorens *et al.*, 2009
	Chromoviruses: Del/Tekay	743	Llorens *et al.*, 2009
	Chromoviruses: Galadriel	11	Llorens *et al.*, 2009
	Chromoviruses: Reina	119	Llorens *et al.*, 2009
	Tat	1080	Llorens *et al.*, 2009
	Errantiviruses	225	Llorens *et al.*, 2009
	ABC clade	5	Llorens *et al.*, 2009
	Caulimoviruses[b]	11	Llorens *et al.*, 2009
Ty1-*copia*	Retrofit	999	Llorens *et al.*, 2009
	Tork	326	Llorens *et al.*, 2009
	Sireviruses	636	Llorens *et al.*, 2009
	Oryco	54	Llorens *et al.*, 2009
	pCreto	8	Llorens *et al.*, 2009
	CoDi-D	2	Llorens *et al.*, 2009
LINE	L1	1468	Kapitonov *et al.*, 2009
	RTE	3	Kapitonov *et al.*, 2009

[a] in *RefBeet* 0.4

[b] Caulimoviruses do not belong to the Ty3-*gypsy* order, but have been cross-detected as indicated in Figure 3.22.

Main points of this analysis are:

(1) In the *B. vulgaris* genome LINEs, Ty3-*gypsy* and Ty1-*copia* retrotransposons occur in high copy numbers. All have more than 1000 members as indicated by the number of RT sequences.

(2) Elements of the BEL-Pao order have not been detected.

(3) Members of the occurring orders are highly diverse. They belong to several lineages, which in turn are subdivided into families.

(4) All previously identified retrotransposon families of the genera *Beta* and *Patellifolia* have been detected. Their addition to the dendrograms allows a comprehensive overview about the retrotransposon composition of beet genomes integrating both, data generated by targeted isolation and genome sequence analysis.

Analysis of *B. vulgaris* Ty3-*gypsy* retrotransposons (Figure 3.24) shows the presence of eight lineages. Most members belong to the chromoviruses, especially to the Del/Tekay clade, and to the Tat lineage with 1021 and 1080 RT sequences, respectively.

The Tat clade contains conventional Ty3-*gypsy* retrotransposons like RIRE2 from rice (Ohtsubo *et al.*, 1999) or Cinful-1 from maize (Sanz-Alferez *et al.*, 2003). In *B. vulgaris*, different Tat families exist, however no representative has been analyzed yet. Also belonging to this clade is the giant TE Ogre with a length of nearly 25 kb (Macas and Neumann, 2007). Few *B. vulgaris* RTs (8) homologous to Ogre have also been identified.

Currently, chromoviral clades (reviewed by Neumann *et al.*, 2011) in the *B. vulgaris* genome are intensively studied, with regard to their localization on chromosomes and conservation of their chromodomains (Beatrice Weber, TU Dresden, personal communication). This survey shows that RTs of all four chromoviral plant clades are present in varying numbers ranging from 11 (Galadriel) to 743 (Del/Tekay).

A third large group of Ty3-*gypsy* elements includes the Errantiviruses (first described in plants by Wright and Voytas, 2002), which are represented by 225 RTs.

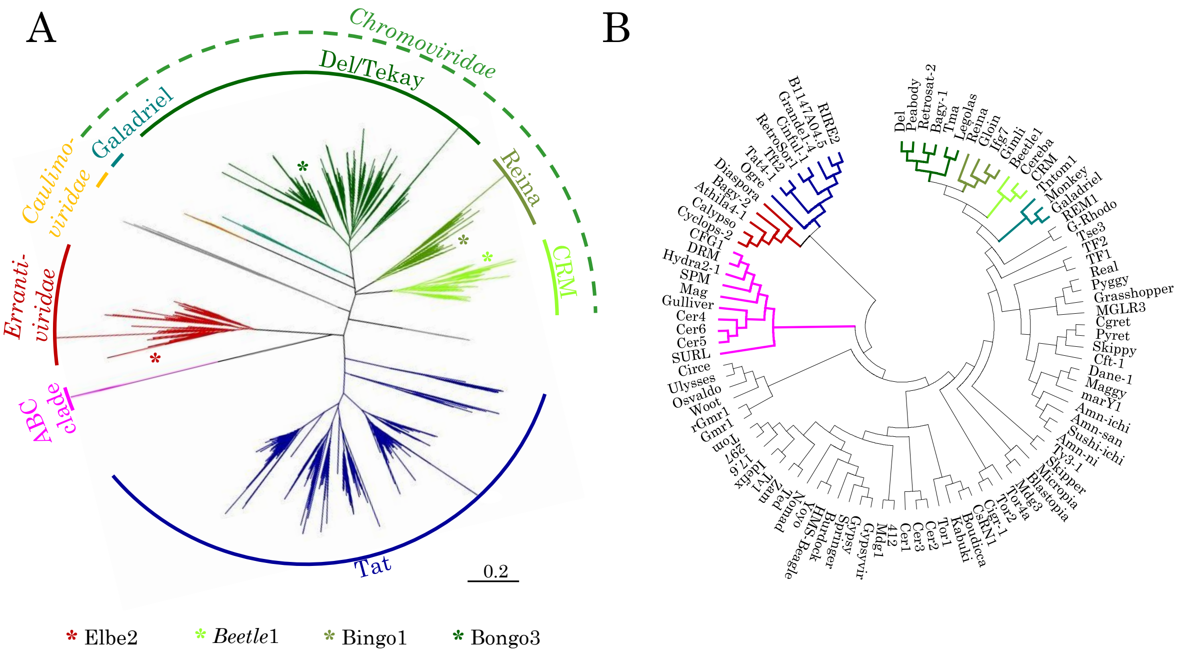


Figure 3.24: Graphical representation of Ty3-*gypsy* retrotransposon lineages.
The dendrograms are based on Ty3-*gypsy* RT (A) and *pol* (B) amino acid sequences. Different lineages have been marked by color (blue = Tat clade; pink = ABC clade; red = Errantiviruses; orange = Caulimoviruses; turquoise = Galadriel clade; dark green = Del/Tekay clade; olive = Reina clade and light green = CRM clade; grey = not assigned). The scale bar represents 0.2 substitutions per site.
(A) Overview of *B. vulgaris* Ty3-*gypsy* retrotransposons. 2355 *HMMER*-derived amino acid sequences with similarities to Ty3-*gypsy* RTs have been aligned using the *MUSCLE* algorithm. Subsequently, a dendrogram was constructed using the Neighbor-Joining method of *Geneious*. A second dendrogram was generated combining the *B. vulgaris* sequences and reference sequences from (B) to enable assignment of RT sequences to lineages (not shown). The positions of Elbe2 (Cora Wollrab, TU Dresden, personal communication), Bingo1 and Bongo3 (Beatrice Weber, TU Dresden, personal communication) from *B. vulgaris* and of *Beetle*1 (Weber and Schmidt, 2009) of *P. procumbens* are indicated by color-coded asterisks.
(B) A phylogeny of 96 Ty3-*gypsy* reference sequences has been recreated according to Llorens *et al.* (2009). Lineages whose presence was verified in *B. vulgaris* have been highlighted by their corresponding color.

A study focusing on these retrotransposons in *B. vulgaris*, designated Elbe, is in progress (Cora Wollrab, TU Dresden, personal communication).

Apart from these well-defined groups, two further lineages have been identified. 11 sequences have similarities to caulimovirus reverse transcriptases, and therefore do not belong to the Ty3-*gypsy* order, while 5 sequences show similarity to the so-called ABC-clade (Llorens *et al.*, 2009). The remaining sequences could not be assigned to any of these lineages.

Conventional retrotransposons of the Retrofit and Tork clade constitute the main portion of the analyzed *B. vulgaris* Ty1-*copia* sequences, numbering 999 and 326, respectively (Figure 3.25). Of the Tork clade, a representative member designated SALIRE1 has already been analyzed in detail (Weber *et al.*, 2010).

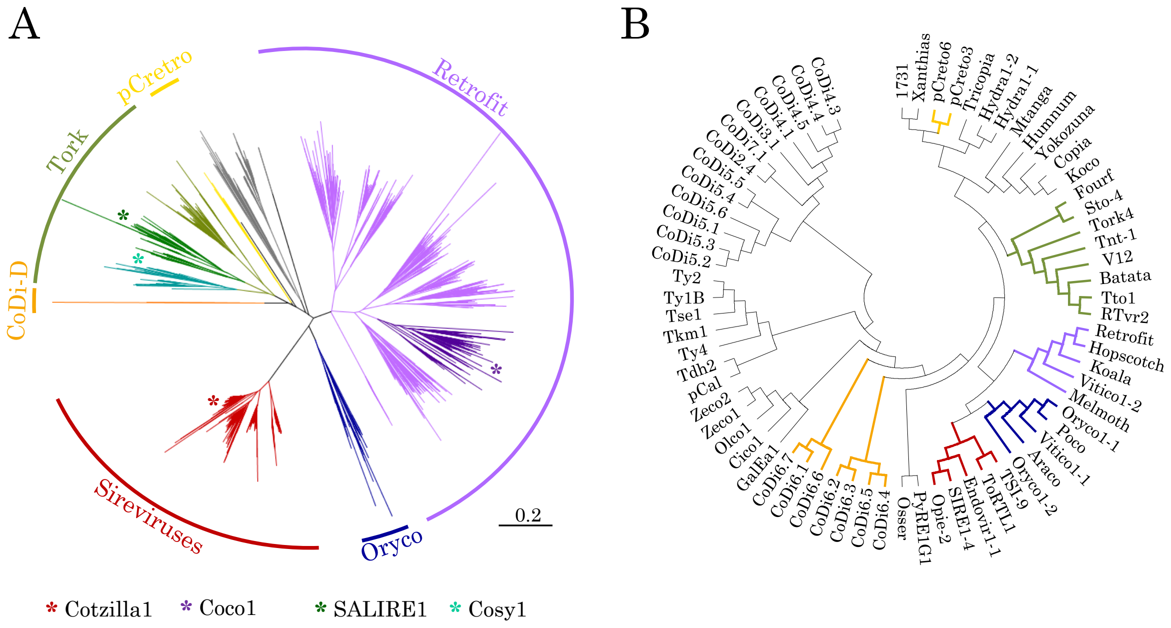


Figure 3.25: Graphical representation of Ty1-*copia* retrotransposon lineages.
The dendrograms are based on Ty1-*copia* RT (A) and *pol* (B) amino acid sequences. Different lineages have been marked by color (red = Sirevirus clade; blue = Oryco clade; violet = Retrofit clade; green = Tork clade; yellow = pCretro clade; orange = CoDi-D clade; grey = not assigned). The scale bar represents 0.2 substitutions per site.
(A) Overview of *B. vulgaris* Ty1-*copia* retrotransposons. The *HMMER*-derived amino acid sequences of 2121 Ty-*copia* RTs have been aligned using the *MUSCLE* algorithm. Subsequently, a dendrogram was constructed using the Neighbor-Joining method of *Geneious*. A second dendrogram was generated combining the *B. vulgaris* sequences and the reference sequences from (B) in order to assign the RT sequences to specific lineages (not shown). The positions of Cotzilla1 and SALIRE1 (this thesis and Weber *et al.*, 2010) from *B. vulgaris* and of Cosy1 and Coco1 (Conny Fiege, TU Dresden, personal communication) of *P. patellaris* are indicated by color-coded asterisks.
(B) A phylogeny of 69 Ty1-*copia* reference sequences has been recreated according to Llorens *et al.* (2009). Lineages whose presence was verified in *B. vulgaris* have been highlighted by their corresponding color.

Furthermore, a comparative study of Coco1 and Cosy1, representatives of both lineages in *P. patellaris*, is currently under way (Conny Fiege, TU Dresden, personal communication).

With 636 *B. vulgaris* RT sequences, the Sireviruses represent a third major group of Ty1-*copia* elements. All analyzed members are derivates of Cotzilla1, a highly abundant retrotransposon, analyzed in detail in Chapter 3.1. The remaining Ty1-*copia* members either belong to the Oryco clade, the pCreto or the CoDi-D group (54, 8 and 2 sequences) or could not been assigned to any lineage.

The vast majority (1468) of the analyzed LINE sequences (Figure 3.26) belong to the L1 clade, well-known for human LINE-1. The remaining sequences (3) group to the RTE-like LINEs. *B. vulgaris* L1 LINE sequences are extremely diverse and split into several families; two of them, BNR and BvL, have already been described (Chapter 3.2 and Wenke *et al.*, 2009).

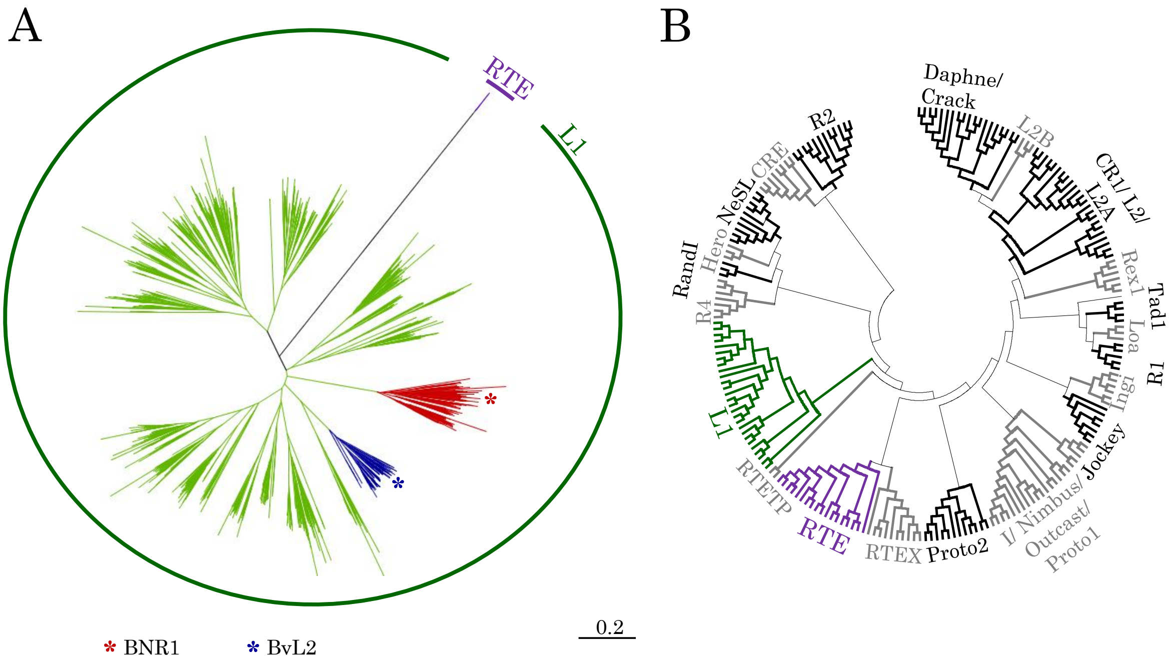


Figure 3.26: Graphical representation of LINE lineages.
The dendrograms are based on LINE RT amino acid sequences. Different lineages have been marked by color (green = L1; violet = RTE). The scale bar represents 0.2 substitutions per site.
(A) Overview of *B. vulgaris* LINEs. The *HMMER*-derived amino acid sequences of 1471 LINE RTs have been aligned using the *MUSCLE* algorithm. Subsequently, a dendrogram was constructed using the Neighbor-Joining method of *Geneious*. A second dendrogram was generated combining the *B. vulgaris* sequences and the reference sequences from (B) in order to assign the RT sequences to specific lineages (not shown). The positions of BNR1 and BvL2 (this thesis and Wenke *et al.*, 2009) are indicated by color-coded asterisks.
(B) A phylogeny of 211 LINE reference sequences has been recreated according to Kapitonov *et al.* (2009). Lineages whose presence was verified in *B. vulgaris* have been highlighted by their corresponding color.

3.3.3 A detailed analysis of *Beta vulgaris* LINEs

Apart from the information gained by a bird eye's view on the retrotransposable content of the *B. vulgaris* genome, these datasets can also be analyzed regarding the TE family composition and diversity. Furthermore, examination of full-length members of each family can gain valuable insights into organization of TE structure and evolution. In this Chapter, an in depth-analysis of retrotransposon diversity is exemplarily performed for *B. vulgaris* LINE sequences. For this study, the *RefBeet* 0.1.1 dataset has been used.

3.3.3.1 Classification of 17 L1 LINE families different in sequence and structure

As the overview of LINE reverse transcriptases in Chapter 3.3.2 shows, many diverse L1 sequences are present in *B. vulgaris* genomes, as opposed to a relatively low number of RTE sequences. Therefore, diversity and family structure of L1 LINEs was analyzed in detail. For this analysis, the first *B. vulgaris* genome assembly, *RefBeet* 0.1.1, has been used. Compared with *RefBeet* 0.4 with 1468 L1, it contains only 1238 L1 sequences.

Nucleotide sequences of the corresponding LINE RT regions have been extracted and aligned by *MUSCLE*. In order to define families, pairwise identity values have been obtained by *MEGA*4 (p-distances option). According to the classification system proposed by Wicker *et al.* (2007), two elements belong to the same family if they share 80 % (or more) sequence identity in at least 80 % of their coding or terminal repeat regions, or in both. Applying Wicker's rule, the 1238 L1 sequences would group into 611 families, many of them containing only one sequence. In order to define a more suitable threshold for L1 family classification, the number of LINE families was deduced as a function of the minimal shared identity (Figure 3.27).

With a decrease of identity restrictions, the family number drops nearly exponentially. A family threshold of at least 60 % identity in their RT sequence was chosen, leading to a number of 17 L1 families. Classified this way, the resulting LINE groups can also be distinguished in a dendrogram by a clear separation from another (Figure 3.28 A). Average pairwise identities between the 17 families are presented in Figure 3.28 B.

For easy reference, all LINE families have been termed Belline (*Be*ta L1 LINEs) with the appended numbers 1-17. Individual family members are named by a further apposition of their index number, e.g. Belline10_2. The previously identified LINE families, BNR and BvL (this thesis and Wenke *et al.*, 2009), will be in the following designated as Belline1 (BNR) and Belline7 (BvL), respectively.

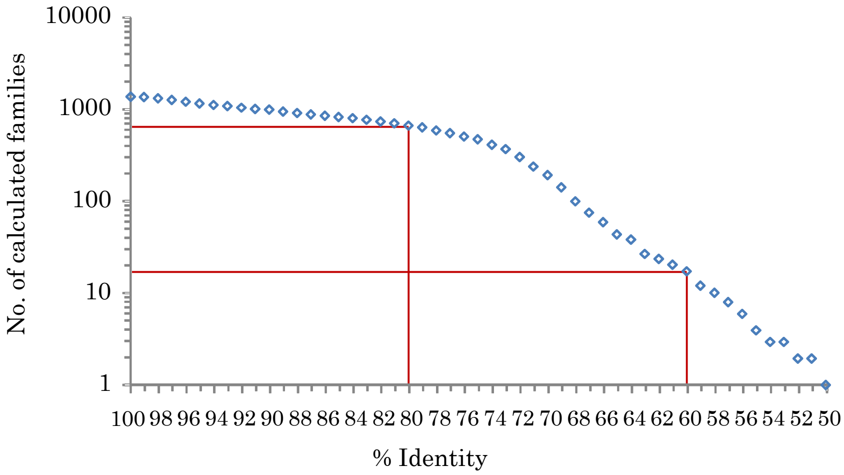


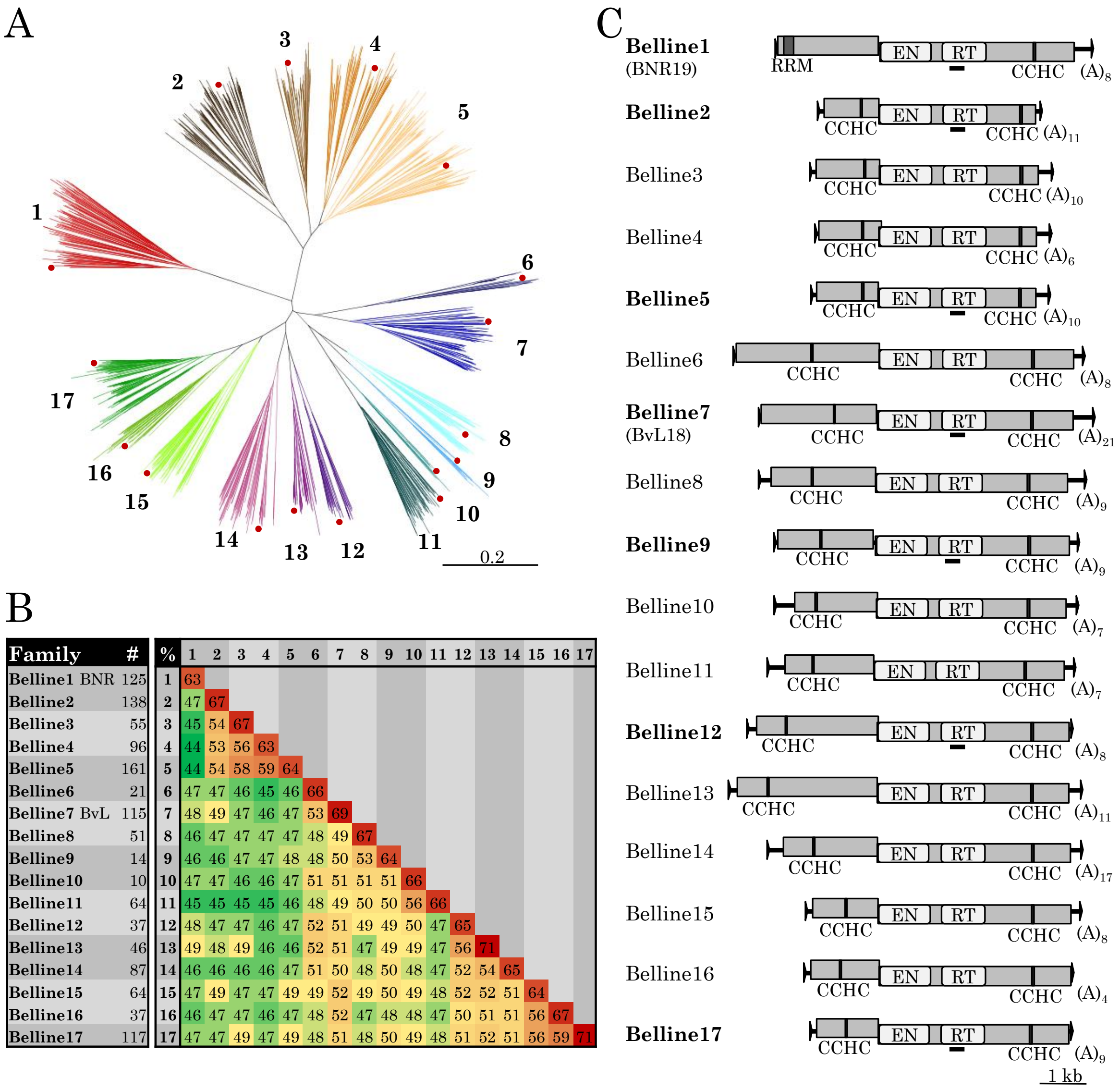
Figure 3.27: *B. vulgaris* L1 LINE family number as a function of the minimal shared identity. A decrease of identity restrictions, leads to a nearly exponential drop of the family number. The minimal shared identity proposed by Wicker *et al.* (2007) and the finally chosen percentage – 80 and 60 % – are marked by red lines.

In order to better understand the family structure of LINEs not only their RT sequences, but also full-length members of Belline1-17 have been analyzed. These have been identified by a *tBLASTn* search in *RefBeet* 0.1.1 using the corresponding consensus RT sequences as query. For each family one representative LINE was selected based on its integrity, structural completeness and absence of internal frameshifts. Comparative analysis of their sequence structures shows several main points (Figure 3.28 C; Table 3.4):

(1) All analyzed full-length LINEs have two ORFs and encode a reverse transcriptase, an endonuclease and a zinc finger of an RNaseH-like domain in ORF2. They all terminate by a poly(A) tail and are flanked by target site duplications of varying length.

(2) Except for one family (Belline1 (BNR), see Chapter 3.2), which encodes an RNA recognition motif in ORF1, all other LINE families code for the zinc finger typical for plant LINEs.

(3) Full-length members of a LINE family are characterized by family-typical element and ORF lengths. Families Belline2, Belline3, Belline4 and Belline5, which originate from one branch in the dendrogram, have sequence lengths of approximately 5000 ± 200 nt. ORF1 and ORF2 of these Belline families are reduced in the number of amino acids, however still contain all structural motifs essential for transposition. Full-length retrotransposons belonging to the other Belline families have a minimum length of 5600 nt.



Family	#	%	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Belline1 BNR	125	1	63																
Belline2	138	2	47	67															
Belline3	55	3	45	54	67														
Belline4	96	4	44	53	56	63													
Belline5	161	5	44	54	58	59	64												
Belline6	21	6	47	47	46	45	46	66											
Belline7 BvL	115	7	48	49	47	46	47	53	69										
Belline8	51	8	46	47	47	47	47	48	49	67									
Belline9	14	9	46	46	47	47	48	48	50	53	64								
Belline10	10	10	47	47	46	46	47	51	51	51	51	66							
Belline11	64	11	45	45	45	45	46	48	49	50	50	56	66						
Belline12	37	12	48	47	47	46	47	52	51	49	49	50	47	65					
Belline13	46	13	49	48	49	46	46	52	51	47	49	49	47	56	71				
Belline14	87	14	46	46	46	46	47	51	50	48	50	48	47	52	54	65			
Belline15	64	15	47	49	47	47	49	49	52	49	50	49	48	52	52	51	64		
Belline16	37	16	46	47	47	46	47	48	52	47	48	48	47	50	51	51	56	67	
Belline17	117	17	47	47	49	47	49	48	51	48	50	49	48	51	52	51	56	59	71

Figure 3.28: Overview of the Belline families, i.e. L1 LINEs in *B. vulgaris*.
(A) Dendrogram based on an alignment of nucleic acid sequences of 1238 Belline RTs. Based on a mimimum shared identity of 60 %, 17 families have been determined. These are represented by individual colors. Positions of representative full-length members have been marked by red dots. Their structure is shown in (C). The scale bar represents 0.2 substitutions per site.
(B) Average pairwise identities of Belline family members in percent. Based on their value, the percentages are shaded by a gradient from green (low similarity) to red (high similarity).
(C) Graphical representation of the structure of representative Belline LINEs from each family. The rectangles represent two ORFs with conserved motives (CCHC ... zinc finger; RRM ... RNA recognition motif). EN and RT refer to the catalytic regions of the endonuclease and reverse transcriptase. The bold lines below individual Belline members represent the regions of the probes used for Southern hybridization or FISH.

(4) Belline15, Belline16 and Belline17 families also have a shortened ORF1 sequence. However, length, sequence and structure of their ORF2 is more similar to the ORF2 of the families Belline6 to Belline14.

(5) Based on Figure 3.28 A and C, three main lineages of Belline LINEs can be distinguished with unique structural features: The Belline1 (BNR) family

constitutes the first lineage and is marked by an RRM in ORF1. Members of the families Belline2 to Belline5 belong to the second lineage. Compared with the majority of *B. vulgaris* LINEs, they have shorter ORFs and are generally more compact in sequence. LINEs of the families Belline6 to Belline17, represent conventional LINEs like *LIb* from sweet potato (Yamashita and Tahara, 2006).

The 17 Belline families have been defined based on a minimum sequence identity of 60 %. This is an average value taking into account conserved domains as well as variable spacer sequences. In order to illustrate how the sequences of the *B. vulgaris* LINE families differ, an amino acid alignment of the RT region of one representive member of each Belline family was generated (Figure 3.28). Domains playing an important role for retrotransposition are highly conserved throughout all families. The sequence variation that allows classification of families is solely found in regions of low conservation linking the conserved domains.

Table 3.4: Structural features of Belline reference LINEs.

LINE[a]	Acces-sion	Size [bp]	ORF1 [aa]	ORF2 [aa]	TSD	Remarks
Belline1_1 (BNR1)	EU564339	6700	742	1398	AACACTCACGCGTCTA	complete, 3 internal frameshifts and 2 stops in ORF2
Belline1_2 (BNR2)	DQ374060	6402	665	1379	AATGCAAGTGAGATAATATAA	complete, 2 internal frameshifts and 3 stops in ORF2
Belline1_19 (BNR19)	FR852795	6670	714	1379	AAACAATACTTTAGTGAAA	complete, no internal frameshift/stop
Belline2_1	FR852796	5162	488	1114	CGCAACCCTTT	complete, 2 internal frameshifts in ORF2
Belline2_2	FR852797	2559	---	790	AATCTTTGTTGTGTA	5' truncated, no internal stop at ORF2
Belline2_3	FR852798	4807	428	1110	GAAAGGATTACAAAGAA	complete, no internal frameshift/stop
Belline2_4	FR852799	4762	390	1114	TAAGTC	complete, no internal frameshift/stop
Belline3_2	FR852800	4967	452	1129	GATATTTAAAAAAA	complete, 1 frameshift in ORF2, many internal stops
Belline4_1	FR852801	4809	429	1122	AAAAGTCATTTCCA	complete, no internal frameshifts, but many stops, 1 mutation in TSD
Belline4_2	FR852802	4765		1122	AAGGAACTG	complete, 1 frameshift in ORF1
Belline5_1	FR852803	4906	425	1131	AAGTATACAAACTC	complete, 1 frameshift in ORF2, a few stops
Belline5_2	FR852804	4873	439	1123	AACCTAAGAGTTTT	complete, 1frameshift in ORF1
Belline6_1	FR852805	7298	1016	1382	ATATAGCTTAAGTACAAA	complete, 3 internal frameshifts
Belline7_1 (BvL1)	FM993986	6807	829	1385	CAATGATGT	complete, 5 internal frameshifts and 1 stop
Belline7_2 (BvL2)	FM993987	6718	795	1387	AAAGAACACAAGGATTTA	complete, 1 internal frameshifts and 1 stop in ORF1
Belline7_18 (BvL18)	FR852806	6730	803	1389	GAAAATTTGAAATAGAAT	complete, no internal frameshift/stop
Belline8_1	FR852807	6755	744	1363	AGGACATGAAGAGATA	complete, no internal frameshifts
Belline8_2	FR852808	6582	633	1366	GATCTA	5‘ truncated, 1 frameshift in ORF2
Belline9_1	FR852809	6262	633	1364	TGGGGATGAACAAT	complete, 1 internal frameshift in ORF2
Belline9_2	FR852810	6333	648	1371	AAtTTGTaAATATCAA	complete, 1 internal frameshift in ORF2, Stop in ORF1
Belline9_3	FR852811	6350	658	1375	ACATATGGTGAA	complete, 4 internal frameshifts
Belline9_4	FR852812	6290	661	1354	ATGATAGAAGGAGTTCA	5' truncated, 1 internal frameshift in ORF2
Belline9_5	FR852813	6471	678	1378	ATATATTGAAACCTGATCT	complete, no internal frameshift/stop
Belline10_1	FR852814	4230	---	1333	AGGTAGGTAATAAAC	5‘ truncated

LINE[a]	Acces-sion	Size [bp]	ORF1 [aa]	ORF2 [aa]	TSD	Remarks
Belline10_2	FR852815	6282	606	1349	AAAACATTCAGATTT	complete, 5 internal frameshifts, ORF2 zinc finger not intact
Belline11_1	FR852816	6647	742	1374	AATTGCTTGGTATGGT	complete, readthrough between ORF1 and ORF2, internal stop in ORF1
Belline11_2	FR852817	6342	631	1353	AACTTATGCGACGAGC	complete, 1 internal frameshift in ORF2, ORF1 zinc finger not intact
Belline12_1	FR852818	6686	848	1364	AAATAATGGGAACAT	5' truncated, no internal frameshift/stop
Belline12_2	FR852819	6903	859	1365	AACATGCTCCTC	complete, no internal frameshift/stop
Belline13_1	FR852820	7360	965	1356	ACATATACTCCCTCC	complete, 5 internal frameshifts
Belline13_2	FR852821	7293	994	1361	AAGAGAT	complete, 2 internal frameshifts and 1 stop in ORF2
Belline14_2	FR85282	5783	542	1372	AAGTTCA	5' truncated, 1 internal frameshift in ORF2
Belline14_4	FR852823	6497	669	1372	GAAGAGATAATAGTT	complete, no internal frameshifts, 1 stop in ORF2
Belline14_5	FR852824	6487	727	1246	GAAAAAGTAGAGTGGTGA	complete, 1 internal frameshift and 2 stops in ORF2
Belline15_1	FR852825	5661	424	1372	AGGGGGGAA	complete, 1 internal frameshift in ORF2
Belline15_2	FR852826	5638	466	1371	AAACAAGTTTAATTA	complete, 1 internal frameshift in ORF1
Belline15_3	FR852827	5667	465	1369	GACTCGCATTTT	complete, no internal frameshift/stop
Belline16_1	FR852828	5811	509	1359	AAGTATAGCTTAATCCG	complete, 1 internal frameshift in ORF1
Belline16_2	FR852829	5698	488	1362	AAACTGCTTATCACTTGAGC	complete, no internal frameshift/stop
Belline16_3	FR852830	5631	475	1361	AGCTTCAACCATTAATA	complete, 2 internal frameshifts in ORF2
Belline17_1	FR852831	5533	438	1355	AACTACTA	complete, no internal frameshift/stop
Belline17_2	FR852832	5631	455	1355	AAACTTTGTTCTCTAGAA	complete, 1 internal frameshift in ORF2
Belline17_3	FR852833	5644	479	1358	AAGACTTAGTTAAATGC	complete, no internal frameshift/1 stop in ORF2
Belline17_4	FR852834	5614	465	1357	AATCATTCTAGATGATTAA	complete, no internal frameshift/stop
Belline17_5	FR852835	5537	433	1356	AAGTGAAACAGTTAGT	complete, 1 internal frameshift in ORF2
Belline17_6	FR852836	5555	452	1355	ACTAATATACTCCA	complete, no internal frameshift/stop

[a] Underlined names denote reference elements whose structure is presented in Figure 3.27.

3.3.3.2 Belline retrotransposons are differentially amplified and organized in genomes of the beet genera *Beta* and *Patellifolia*

In order to verify the results shown in the dendrogram in Figure 3.27 A, and to determine the occurence of LINEs of the Belline families in related species, comparative Southern hybridization to *Hin*dIII restricted genomic DNA has been performed. As DNA source, plants of the sections *Beta*, *Corollinae* and *Nanae* from the genus *Beta*, as well as one species from the genus *Patellifolia* have been chosen. *Spinacia oleracea* and *Chenopodium quinoa* DNA has been used as outgroup. RT fragments of seven Belline representatives (Table 2.7) from main branches of the dendrogram (Figure 3.28 A) have been selected to probe seven identically loaded membranes (Figure 3.30). Washing stringency has been approximately 75 %.

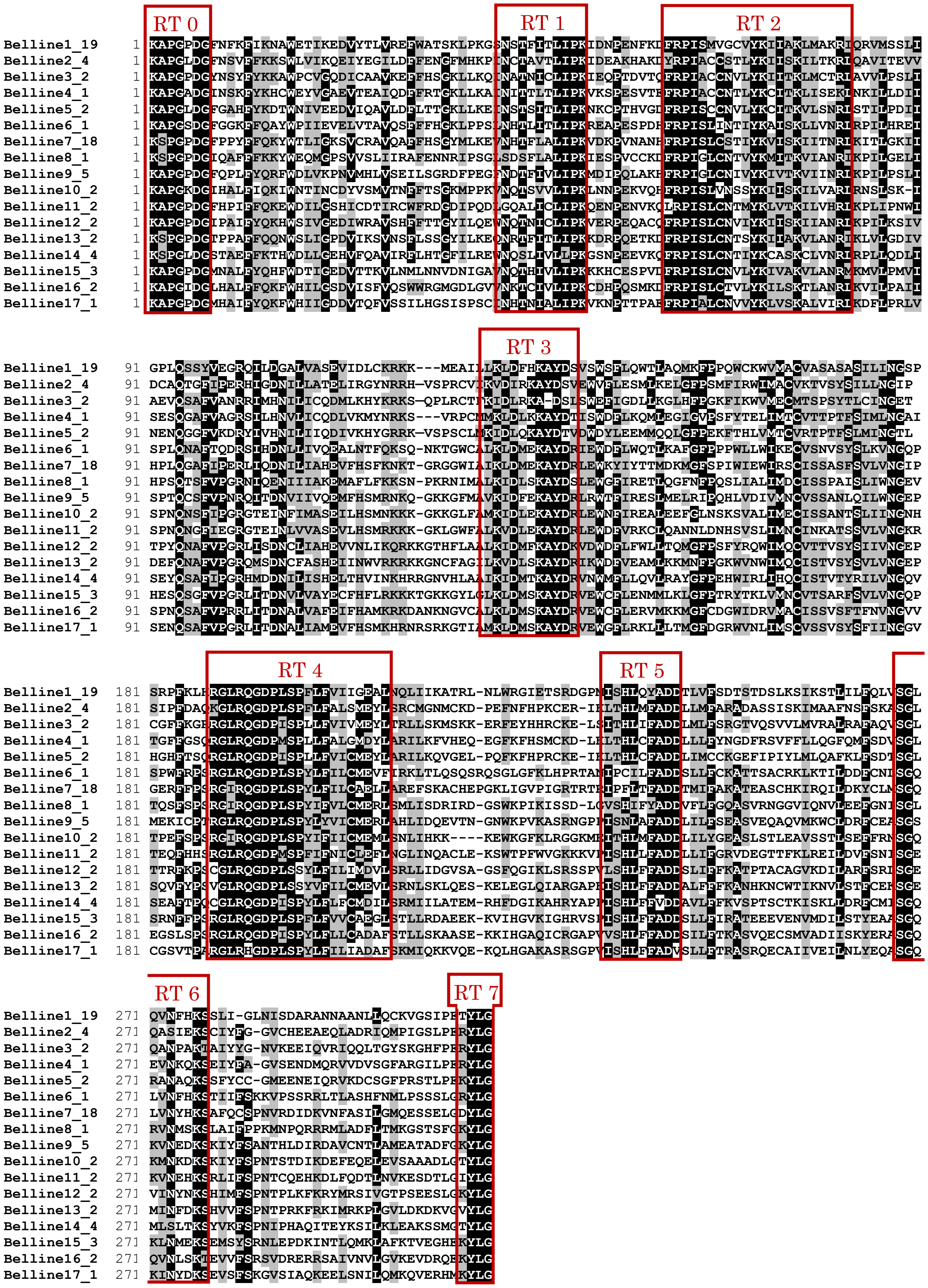


Figure 3.29: Alignment of the RT domains of Belline references.
Red boxes indicate reverse transcriptase domains 0 to 7 (Malik *et al.*, 1999; Wright *et al.*, 1996; Xiong and Eickbush, 1990). Residues with an identity or similarity greater than 60 % have been shaded in black or grey, respectively. The shading shows that LINE-typical domains are conserved, while variation occurs in the spacer sequences between the RT domains.

A varying number of copies has been detected for the Belline families analyzed by computation, ranging from 14 for Belline9 to 138 for Belline2. At first, it was investigated, if this number corresponds to the signal intensity of *B. vulgaris* DNA hybridization (Figure 3.29, lane 1 of each panel).

Autoradiograms obtained by probing with Belline2_4, Belline7_18/BvL_18, Belline12_2 and Belline17_1 show a great number of strong signals. Except for Belline12, these families also comprise a high number of copies (>100). Furthermore, the representative elements used are all situated in a dense region with short branches in the dendrogram. This is an indication for the presence of many sequences with an identity higher than the 60 % chosen for family classification.

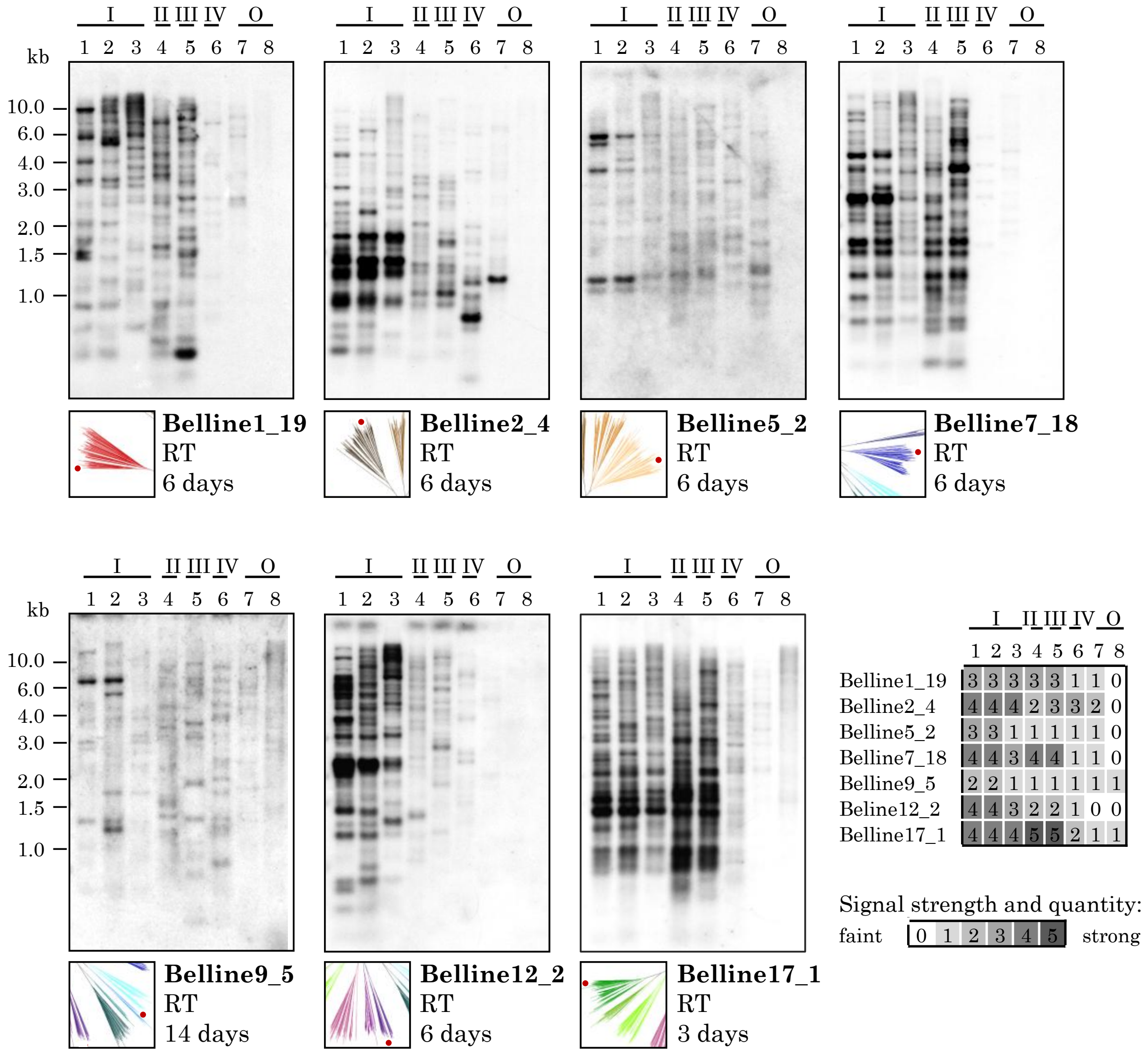
Not all members of the family branch hybridize as can be observed for Belline1_19 and Belline5_2 hybridization. Their families also have more than 100 members, however, hybridization shows only few or faint signals. It can be seen, especially well for Belline5_2, that the selected reference sequences derivate slightly from the consensus and are therefore situated in a less dense branch region.

Even after extended exposure time, Belline9_5 shows only faint hybridization, which corresponds well with the low number of family members. These differences in hybridization indicate (1) an absence of cross-hybridization with members of other families and (2) hybridization to the most homologous sequences of a family, only.

While keeping in mind these limitations of Southern hybridization, the abundance of Belline transposons in related species will be described in the following (Figure 3.30, all lanes).

Belline1_19 and Belline 7-18/BvL reverse transcriptases hybridized in nearly equal signal strengths and numbers to the tested species of the genus *Beta*. Hybridization with a Belline7 (BvL) probe revealed the existence of many common bands, however, for Belline1_19-like LINEs, all species, even among the cultivars, have only very few shared signals. This indicates a different genomic organization of Belline1 retrotransposons with few conserved restriction sites, and suggests retrotransposition, while the opposite is true for Belline7 (BvL). For both probes, only faint signals have been observed for genus *Patellifolia* and *S. oleracea*, while no hybridization occurred to *C. quinoa* DNA.

Belline2-4 hybridization resulted in equal hybridization patterns in the section *Beta*. Very strong and similar bands show conserved restriction sites within the LINE or the flanking region, indicating Belline2 proliferation after the sections *Beta* and *Corollinae* diverged. A reduced number and strength of signals are visible in the sections *Corollinae* and *Nanae*, however the major bands are still visible. For distantly related species of the genus *Patellifolia* and *S. oleracia*, a few strong signals are still visible. Only in *C. quinoa* no hybridization was detected.



	I	I	I	II	III	IV	O	O
	1	2	3	4	5	6	7	8
Belline1_19	3	3	3	3	3	1	1	0
Belline2_4	4	4	4	2	3	3	2	0
Belline5_2	3	3	1	1	1	1	1	0
Belline7_18	4	4	3	4	4	1	1	0
Belline9_5	2	2	1	1	1	1	1	1
Beline12_2	4	4	3	2	2	1	0	0
Belline17_1	4	4	4	5	5	2	1	1

Signal strength and quantity:

faint 0 1 2 3 4 5 strong

Figure 3.30: Distribution of Belline copies in the genera *Beta* and *Patellifolia*.
Genomic *Hin*dIII-restricted DNA was analyzed by comparative Southern hybridization using a probe from LINE reverse transcriptases. Species tested were: Cultivars of *B. vulgaris* ssp. *vulgaris* in the section *Beta* (I): sugar beet KWS 2320 (1), chard (2); and wild beet species from the section *Beta* (I): *B. patula* (3); species of the section *Corollinae* (II): *B. corolliflora* (4); species of the section *Nanae* (III): *B. nana* (5); species of the genus *Patellifolia* (IV): *P. procumbens* (6); outgroup species (O): *Spinacia oleracea* (7) and *Chenopodium quinoa* (8).
Below the autoradiograms, a clipping of Figure 3.27 A shows the position of the respective Belline reverse transcriptase in the dendrogram of all detected Belline RTs. Exposure is indicated in days.
In the lower right corner, abundance of Belline RTs is comparatively presented by assignment of a value between 0 and 5 to each lane based on relative signal intensity.

Hybridization with a Belline12_2 probe generated many strong and similar signals in the section *Beta*, too. However, the signals are faint and rare in the other sections or in the genus *Patellifolia*. No hybridization signals have been detected in the outgroup species, pointing to a relatively young family that underwent only recent amplification in the section *Beta*.

Contrasting with other LINE families, Belline17_1 shows the strongest signals after hybridization to species of the sections *Corollinae* and *Nanae* indicating amplification in the wild beet sections or some loss from the cultivated species. However, for the section *Beta* numerous signals are still present that show size conservation in the whole genus. For *Patellifolia*, the signals are less frequent, and for the outgroup species they are barely visible.

Hybridization of Belline5_2 and Belline9_5 probes resulted in very few and faint signals. Distinct signals are only visible using DNA of *Beta* cultivars. All other species show only very faint signals, probably by hybridization to diverged sequences of the same family.

In summary, LINEs have been differently amplified in the evolutionary history of the genera *Beta* and *Patellifolia*. Most of the analyzed families are abundant in the section *Beta* and much less common in the other species. However, there are exceptions: Of the analyzed families, only LINEs of the Belline2 family show a strong signal in the *Patellifolia* genus and even in *S. oleracea*, and Belline17 LINEs are more abundant in *Corollinae* and *Nanae* than in *Beta*.

3.3.3.3 Chromosomal localization of two exemplary LINE families

Two *B. vulgaris* LINE families have already been analyzed in regard to their distribution along chromosomes: The Belline1 (BNR) and the Belline7 (BvL) family whose members occur on all chromosomes in a dispersed pattern (this thesis, Chapter 3.2.6, and Wenke *et al.*, 2009). In order to generalize this statement for *B. vulgaris* LINEs, two additional families have been selected for FISH (Figure 3.31): Belline2 belongs to the lineage of short LINEs, while Belline17 is a conventional LINE family like BvL, however, situated on a different branch in the dendrogram (see Chapter 3.3.3.1). Similar to Southern hybridization, probes spanning approximately 325 bp of the RT have been used.

For both LINE families, signals have been detected on all chromosomes (Figure 3.31 A and C). They occur mainly in intercalary and distal chromosomal regions. Signals of varying intensity point to the formation of LINE clusters. Hybridization to interphase spreads shows a preference for weakly DAPI-stained euchromatic regions (Figure 3.31 B and D). However, heterochromatic signals have also been observed.

It can be assumed that most or even all analyzed *B. vulgaris* LINE families are distributed on all chromosomes in a dispersed organization. They often occur in clusters as indicated by strong signals.

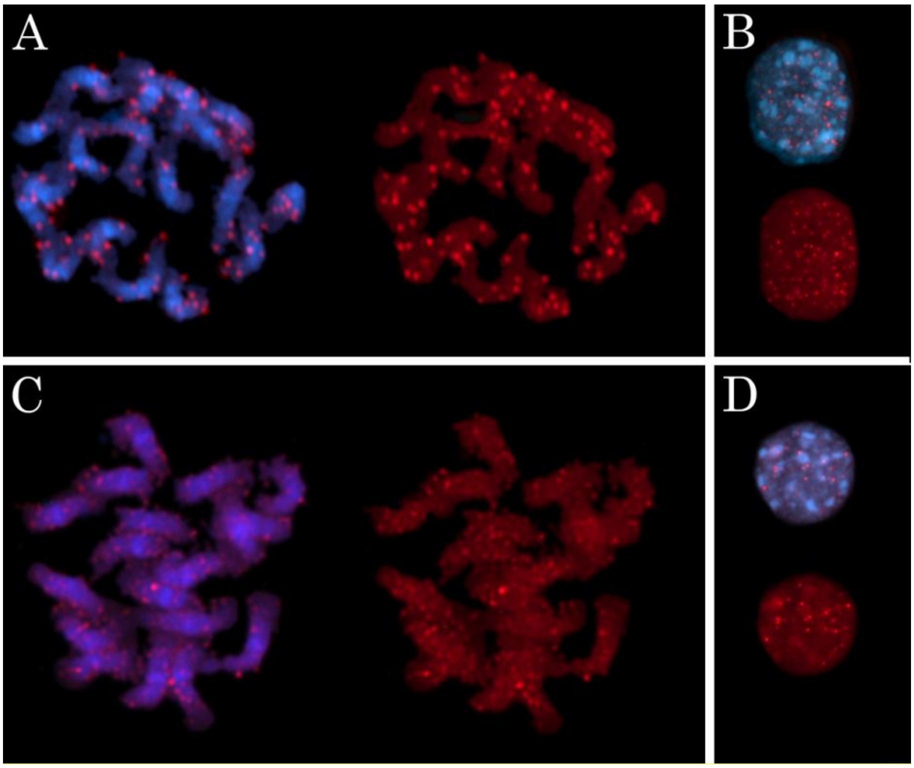


Figure 3.31: Localization of Belline2 and Belline17 elements along *B. vulgaris* chromosomes. Blue fluorescence shows DNA stained with DAPI, while red fluorescence shows hybridization with LINE reverse transcriptase sequences of Belline2 **(A-B)** or Belline17 **(C-D)**. The scale bar in panel D corresponds to 10 µm.
Metaphase **(A)** and interphase **(B)** nuclei were hybridized with a 324 bp probe specific for the RT of Belline2. A Belline17 RT region of 325 bp was hybridized to metaphase **(C)** and interphase **(D)** nuclei.

3.3.3.4 To which extent do *B. vulgaris* RTE LINEs exist?

PCR fragments of RTE LINEs have been described for a few monocotyledonous as well as dicotyledonous plant genomes (Zupunski *et al.*, 2001). However, in the *B. vulgaris RefBeet* databases only very few sequences containing the corresponding RT domains have been identified. Detailed analysis of these contigs revealed only a single RTE representative flanked by target site duplications. This 5' truncated RTE LINE has a length of 2775 bp, an 11 bp TSD and a poly(TTG) tail, a typical feature of RTE elements (Figure 3.32 A, annotated sequence in Appendix 3). Because *B. vulgaris* RTE LINEs escaped their detection, retrotransposons of this family have been designated *Ghost*. Additional *Ghost* elements have not been found by computational approaches like *BLAST* or *HMMER* search. Comparison of the *Ghost* RT amino acid sequence with ORFs of other LINEs confirms its membership to the RTE clade (Figure 3.32 B).

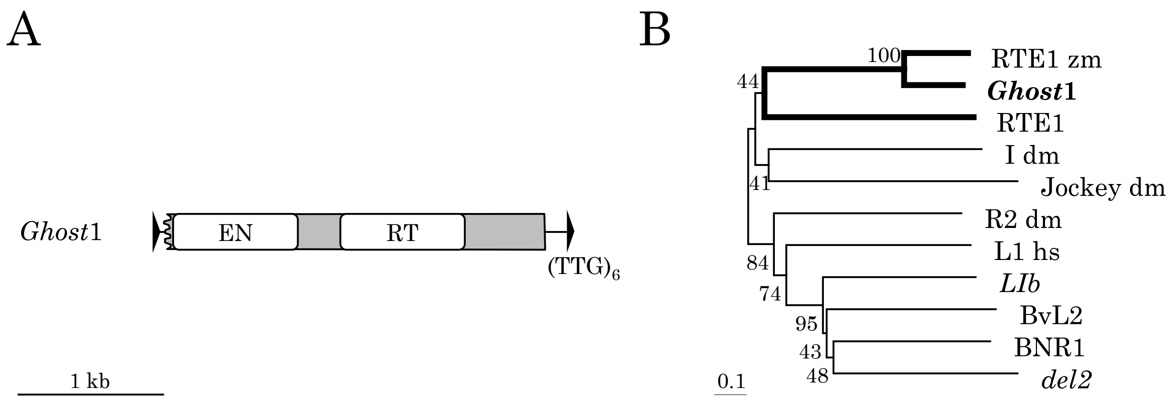
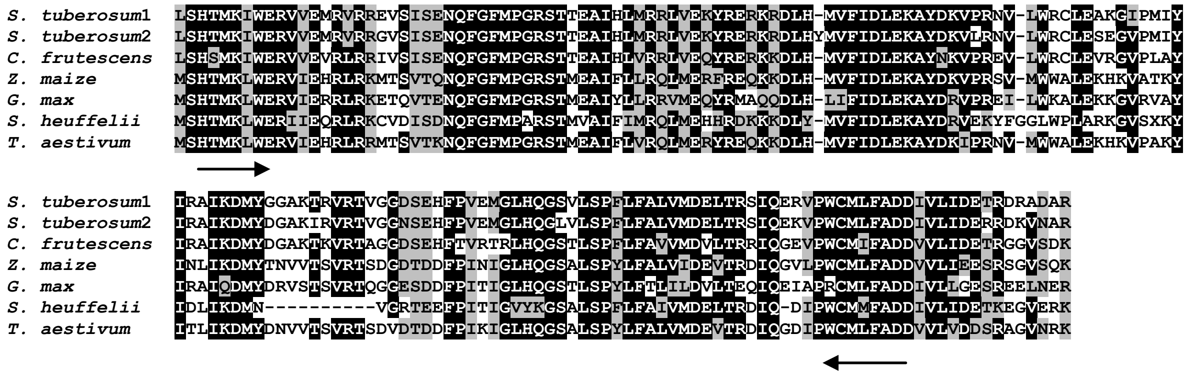


Figure 3.32: Schematic representation of the 5' truncated RTE LINE *Ghost*1 and its position in a phylogenetic tree.
(A) Schematic drawing of *Ghost*1. The rectangle indicates an open reading frame including endonuclease (EN) and reverse transcriptase (RT) domains. Triangles mark target site duplications. The LINE ends with a (TTG)$_6$ tail. A wiggly line symbolizes 5' truncation.
(B) Dendrogram showing the relatedness of *Ghost*1 to characterized reverse transcriptases from the RTE, L1, I, Jockey and R2 clade. The *B. vulgaris Ghost*1 element groups with LINEs of the RTE clade. For comparison, LINE ORF2 sequences were analyzed from: *Caenorhabditis elegans* (RTE1), *Zea mays* (RTE1 zm), *Homo sapiens* (L1 hs), *Ipomoea batatas* (*LIb*), *Beta vulgaris* (BNR1 and BvL2), *Lilium speciosum* (*del2*) and *Drosophila melanogaster* (R2, I, Jockey). The scale bar represents 0.1 substitutions per site.

Based on the *Ghost* results, degenerate primers specific for plant RTE LINEs were designed. A *Solanum tuberosum* RTE sequence (Döbel, 2008) has been used as *BLAST* query to detect full-length elements in *S. tuberosum* as well as in other plant genomes. Their RT amino acid sequences have been deduced, aligned and used for primer generation (Figure 3.33).



```
S. tuberosum1  LSHTMKIWERVVEMRVRREVSISENQFGFMPGRSTTEAIHLMRRLVEKYRERKRDLH-MVFIDLEKAYDKVPRNV-LWRCLEAKGIPMIY
S. tuberosum2  LSHTMKIWERVVEMRVRRGVSISENQFGFMPGRSTTEAIHLMRRLVEKYRERKRDLHYMVFIDLEKAYDKVLRNV-LWRCLESEGVPMIY
C. frutescens  LSHSMKIWERVVEVRLRRIVSISENQFGFMPGRSTTEAIHLVRRLVEQYRERKKDLH-MVFIDLEKAYNKVPREV-LWRCLEVRGVPLAY
Z. maize       MSHTMKLWERVIEHRLRKMTSVTQNQFGFMPGRSTMEAIFLLRQLMERFREQKKDLH-MVFIDLEKAYDKVPRSV-MWWALEKHKVATKY
G. max         MSHTMKLWERVIERRLRKETQVTENQFGFMPGRSTMEAIYLLRRVMEQYRMAQQDLH-LIFIDLEKAYDRVPREI-LWKALEKKGVRVAY
S. heuffelii   MSHTMKLWERIIEQRLRKCVDISDNQFGFMPARSTMVAIFIMRQLMEHHRDKKKDLY-MVFIDLEKAYDRVEKYFGGLWPLARKGVSXKY
T. aestivum    MSHTMKLWERVIEHRLRRMTSVTKNQFGFMPGRSTMEAIFLVRQLMERYREQKKDLH-MVFIDLEKAYDKIPRNV-MWWALEKHKVPAKY

S. tuberosum1  IRAIKDMYGGAKTRVRTVGGDSEHFPVEMGLHQGSVLSPFLFALVMDELTRSIQERVPWCMLFADDIVLIDETRDRADAR
S. tuberosum2  IRAIKDMYDGAKIRVRTVGGNSEHFPVEMGLHQGLVLSPFLFALVMDELTRSIQEKVPWCMLFADDIVLIDERRDKVNAR
C. frutescens  IRAIKDMYDGAKTKVRTAGGDSEHFTVRTRLHQGSTLSPFLFAVVMDVLTRRIQGEVPWCMIFADDVVLIDETRGGVSDK
Z. maize       INLIKDMYTNVVTSVRTSDGDTDDFPINIGLHQGSALSPYLFALVIDEVTRDIQGVLPWCMLFADDVVLIEESRSGVSQK
G. max         IRAIQDMYDRVSTSVRTQGGESDDFPITIGLHQGSTLSPYLFTLILDVLTEQIQEIAPRCMLFADDIVLLGESREELNER
S. heuffelii   IDLIKDMN----------VGRTEEFPITIGVYKGSALSPFLFAIVMDELTRDIQ-DIPWCMMFADDIVLIDETKEGVERK
T. aestivum    ITLIKDMYDNVVTSVRTSDVDTDDFPIKIGLHQGSALSPYLFALVMDEVTRDIQGDIPWCMLFADDVVLVDDSRAGVNRK
```

Figure 3.33: Alignment of RTE LINE reverse transcriptase fragments of various plants.
The RTE sequences have been identified in the following genomes: *Solanum tuberosum* (AC232019 and AC233635), *Capsicum frutescens* (DQ913814), *Zea mays* (AJ850302), *Glycine max* (AC235178), *Silene heuffeli* (*in silico* fusion of AY720863 and AY720885), *Triticum aestivum* (CT009585). Amino acid regions that have been used for design of degenerated primers (RTE02 for and rev) are marked by arrows.

Using the primer pair RTE02 for and rev (Table 2.6) for PCR with *B. vulgaris* DNA, the expected product with a size of 450 bp has been amplified (Figure 3.34 A), cloned and five clones have been sequenced. One of the clones, with a 96 % identity to *Ghost*1, has been used for high density filter hybridization (Figure 3.34 B). Out of a BAC library (Chapter 2.1.2), 9216 BACs have been spotted onto the filter, equating to 1.5 *B. vulgaris* genome equivalents. Therefore, a number of 140 signals relates to approximately 90-100 *Ghost* copies in the *B. vulgaris* genome. This indicates that *Ghost* RTE LINEs are at least middle repetitive in the genome of *B. vulgaris*, but underrepresented in the *RefBeet* databases.

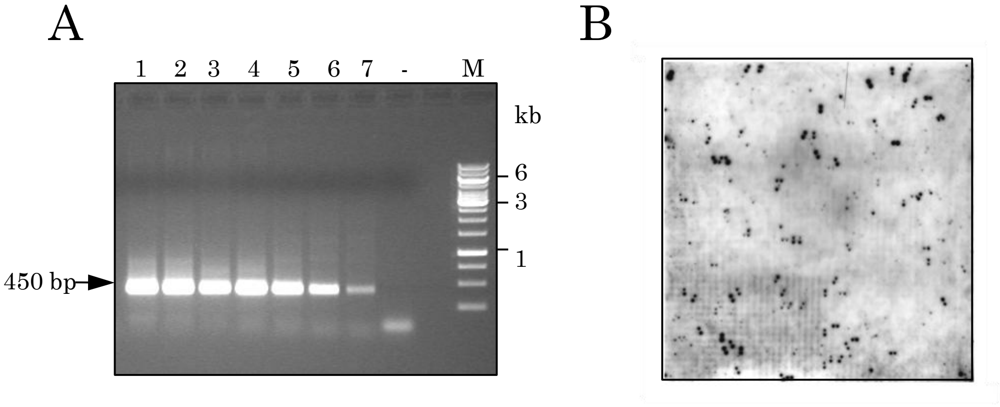


Figure 3.34: Experimental strategy for the detection of RTE LINE RTs in *B. vulgaris*.
(A) A PCR was performed using *B. vulgaris* DNA and the primer pair RTE02 for and rev. The amplicon had the expected size of 450 bp. Lane 1-7 show different annealing temperatures: 48 °C (1), 50 °C (2), 51,4 °C (3), 52,9 °C (4), 54,5 °C (5), 56,1 °C (6), 57,5 °C (7). In addition, a negative probe (-) was tested.
(B) A high density filter was probed with a sequenced clone resulting from the 450 bp amplicon. The corresponding autoradiogram is shown after exposure for three days.

3.3.3.5 LINE transcription is tightly controlled by small RNAs

A shown above, *B. vulgaris* LINEs are present in many families with varying abundance. These findings point to a tight control of transcription and translation. TE amplification control can be mediated by small RNAs (sRNAs) complementary to the respective elements. While 24 nt sRNAs induce transcriptional gene silencing and heterochromatin formation, 21 nt sRNAs inhibit translation by post-transcriptional gene silencing (reviewed by Simon and Meyers, 2010).

In order to detect if these processes regulate LINE amplification in *B. vulgaris*, a database containing sRNAs (Himmelbauer *et al.*, personal communication) has been assembled to LINEs of each family using the *Geneious* assembler. The assembled RNA sequences have a minimum identity of 90 % to the respective LINE. Introduction of sequence gaps has not been permitted. Furthermore, a random nucleotide sequence of

100,000 nt has been analyzed to define the number of unspecific assemblies with these settings: Every 1000 nt, approximately four sRNAs match the random sequence, indicating a low background in the investigation. Only few of the analyzed LINEs show a high number of matching sRNAs that belong mostly to the group of 24 nt sRNAs (Table 3.5). This illustrates that if sRNA-mediated silencing occurs, it is induced largely at the transcriptional level.

Table 3.5: Number of small RNAs that match to representatives of each LINE family

LINE	No. of 21 nt RNAs	No. of 24 nt RNAs	Total no. of small RNAs
Belline1_1	5	31	98
Belline1_19	4	21	58
Belline2_4	1	15	36
Belline3_2	5	0	61
Belline4_1	0	5	37
Belline5_2	2	0	22
Belline6_1	**16**	**247**	**386**
Belline7_18	**12**	**143**	**238**
Belline8_1	13	32	139
Belline9_1	0	3	41
Belline9_5	1	11	37
Belline10_2	6	10	53
Belline11_2	6	24	90
Belline12_2	7	19	70
Belline13_2	7	47	117
Belline14_4	4	24	60
Belline15_3	1	14	66
Belline16_2	8	10	80
Belline16_3	4	36	95
Belline17_1	8	99	155
Belline17_4	**19**	**350**	**565**
***Ghost*1**	**25**	**498**	**631**

Assemblies with the highest number of 24 nt sRNAs have been analyzed in greater detail. Thereby, all sRNAs of 21 and 24 nt have been mapped directly to the sequence (Figure 3.35). Interestingly, most of the 24 nt sRNAs bind to a region situated near the 3' end of the LINE sequence either corresponding to the 3' UTR or the RNaseH-like region of ORF2. For Belline6_1 and Belline7_18, the sRNA-matching regions contain palindromic stretches that might lead to the formation of double-stranded RNA (dsRNA). This dsRNA could be the source of the high number of sRNAs. However, for *Ghost*1 and Belline17_4, a similar correlation was not identified.

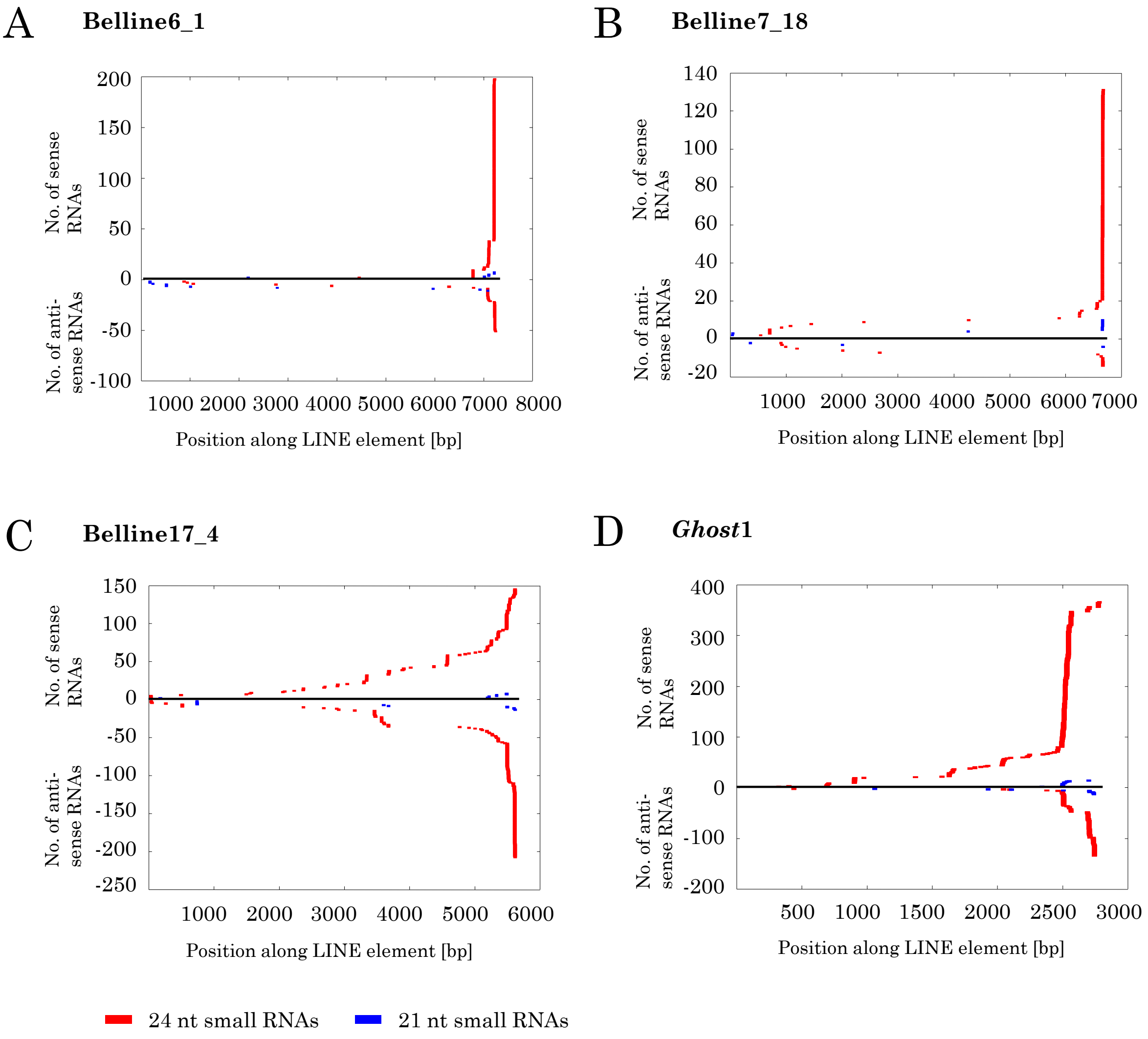


Figure 3.35: Mapping of small RNAs to selected representatives of the L1 families Belline6, Belline7 and Belline17 as well as of the RTE family *Ghost*.
Small RNAs with a length of 24 nt have been marked by red color, while 21 nt RNAs have been colored blue. The matching sRNAs have a minimum identity of 90 % to the reference LINE.

3.3.4 Is it possibile to generalize the findings of the *B. vulgaris* LINE landscape for higher plants?

LINEs have been well-analyzed in mammalian genomes like those of human and mouse. During evolution, LINEs of these genomes underwent several bursts of amplification and they are present now with nearly identical members in high copy numbers (Ostertag and Kazazian, 2001b; Konkel and Batzer, 2010). Recent estimations of retrotransposon content in plant genomes give only low copy numbers for LINEs in plant genomes (see Figure 1.3). Therefore, extreme sequence variation might not be expected. However, detailed analysis of *B. vulgaris* LINEs shows that they occur in high diversity, with at least one RTE and 17 L1 LINE families (Chapter 3.3.3). In the following, it shall be investigated, if this LINE diversity is typical for genomes of higher plants.

3.3.4.1 LINEs of higher plants are either divergent L1 or homogenous RTE elements

For LINE diversity analysis, genome sequences of twelve additional plants and four animals have been retrieved from open databases (Table 2.9). LINE RT sequences were extracted by a database search with the LINE Hidden Markov Model as described for *B. vulgaris* (Chapter 2.3.4). The resulting hits have been filtered by application of the same parameters as explained above (Chapter 3.3.1), followed by generation of a multiple sequence alignment and construction of a dendrogram using a Neighbor-Joining algorithm. The dendrograms giving an overview of LINE diversity of their respective genomes are shown in Figure 3.36, while Table 3.6 summarizes sequence hit and family numbers.

Two previous studies have already addressed the LINE content of plant genomes. These have been taken into consideration to validate and compare the *HMMER*-based results:

In 2000, Noma *et al.* identified *Arabidopsis thaliana* LINE sequences by *BLAST* homology search. The corresponding sequences have been identified by the HMM-based approach (Figure 3.36 A, red dots) as well as an additional main group of LINE RTs (Figure 3.36 A, arrow). This indicates that HMM searches are much more sensitive to distantly related sequences than common *BLAST* searches. All detected *A. thaliana* RT sequences belong to the L1 LINE clade.

In the scope of the B73 *Zea mays* genome annotation, 31 LINE families were identified by search of terminal sequence duplications flanking a block of appropriate sequence length (Baucom *et al.*, 2009). Sequences from each family have been detected (Figure 3.36 B, red dots), as well as additional RTs, which did not correspond to one of the previously described *Z. mays* LINEs. The LINE RT sequences group into two clades: RTE (105 sequences) and L1 (1008 sequences).

Indeed, LINEs of all analyzed plant genomes group into several L1 families. By application of a minimum identity of 60 %, calculated family numbers range from 5 (*Vitis vinifera*) to even 42 families (*Brachypodium distachyon*). RTE LINEs have not been detected in all analyzed genome assembles. Contrasting to L1 LINEs, RTE reverse transcriptases, if detected, are highly similar within the analyzed genomes. The 1129 RTE reverse transcriptases of *Malus x domestica* have, for example, an average identity of more than 98 %.

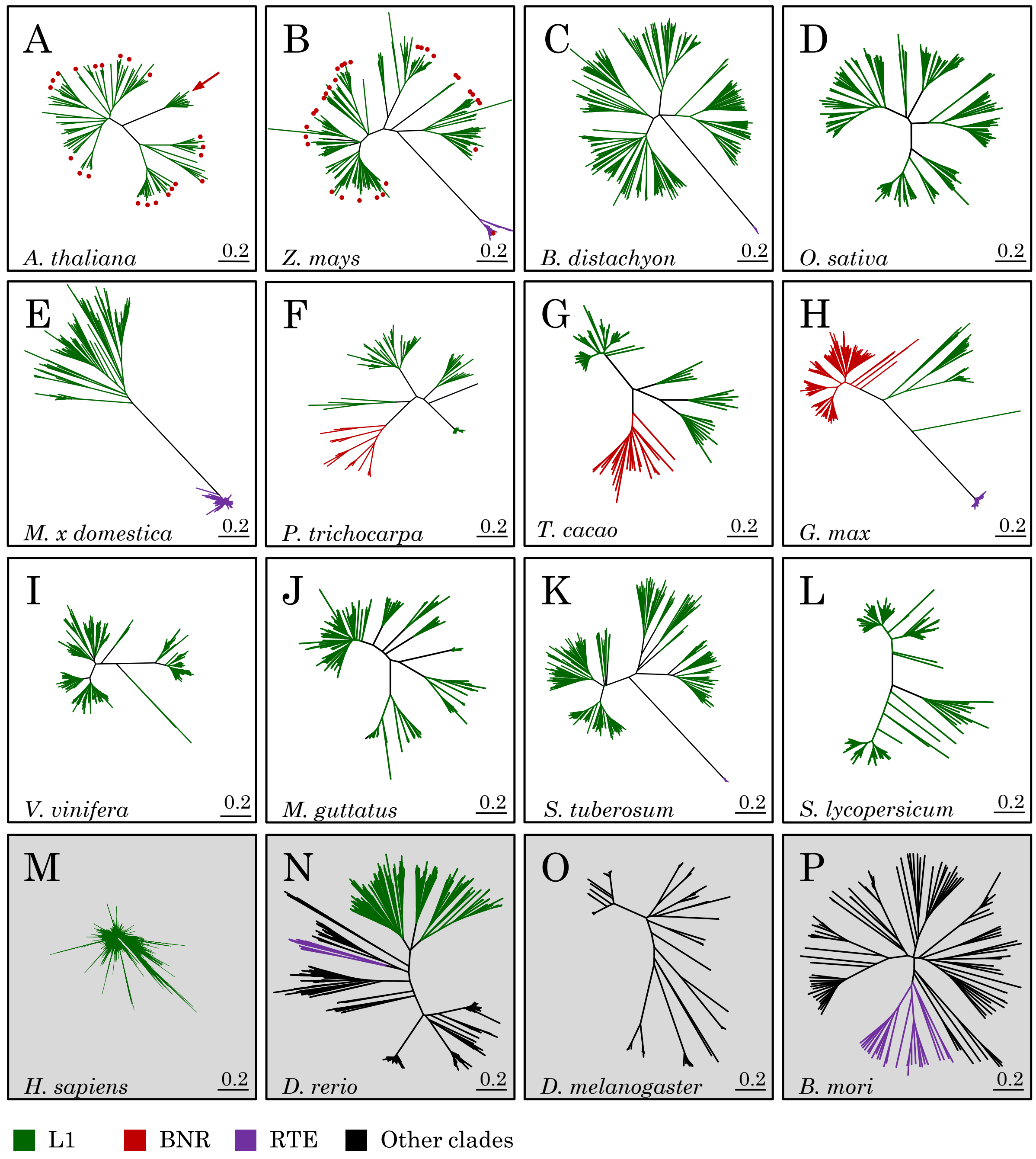


Figure 3.36: LINE clades in higher plants and in some animal reference genomes. LINE reverse transcriptases have been extracted from a number of plant **(A-L)** and animal **(M-P)** genomes using the *HMMER* strategy. Their amino acid sequences have been assembled, followed by the construction of a Neighbor-Joining dendrogram. Based on sequence similarities to the 211 RTs characterized by Kapitonov *et al.* (2009), LINE clades have been assigned to dendrogram branches. These have been colored according to the clade (Green = L1; red = L1 subclade BNR; violet = RTE; black = all others). The scale bars represent 0.2 substitutions per site.

To visualize the difference in LINE diversity of plant and well-characterized mammalian genomes, a dendrogram has also been generated for human LINE RTs (Figure 3.36 M). In contrast to plant LINE RTs, mammalian sequences do not form major branches. Instead, most sequences launch from one origin and form a star-like pattern that visualizes high sequence identities. Assuming a minimum sequence identity of 60 %, human L1 LINEs would constitute a single family. This analysis illustrates the contrast between highly identical and numerous human L1 LINEs, and the diverse variety with far less family members of plant L1 LINEs.

Table 3.6: Summary of the *HMMER*-based results of genome-wide LINE reverse transcriptase detection.

Species	**No. of LINE RTs**[a]		**LINE clades (hits)**[b]	**No. of families**[c]			
				60 % identity		**80 % identity**	
	Score ≥ 50	**Score ≥ 50 + length restriction**		**L1**	**RTE**	**L1**	**RTE**
Plant genomes							
Beta vulgaris (*RefBeet* 0.1.1)	5215	1238	L1 (1238), including BNR (125)	17	---	611	---
Arabidopsis thaliana	298	125	L1 (125)	15	---	103	---
Zea mays	5287	1113	L1 (1008); RTE (105)	16	1	138	3
Oryza sativa	839	261	L1 (261)	38	---	161	---
Brachypodium distachyon	1689	358	L1 (348); RTE (10)	42	1	256	1
Glycine max	2586	1026	L1 (708), including BNR (584); RTE (318)	17	1	99	1
Malus x domestica	5151	1443	L1 (314); RTE (1129)	23	1	148	1
Populus trichocarpa	608	146	L1 (146), including BNR (30)	9	---	53	---
Theobroma cacao	612	111	L1 (111)	13	---	59	---
Vitis vinifera	3021	929	L1 (929)	5	---	44	---
Mimulus guttatus	703	274	L1 (274)	11	---	117	---
Solanum lycopersicum	1237	174	L1 (174)	8	---	38	---
Solanum tuberosum	2611	453	L1 (434); RTE (19)	12	1	209	1
Animal genomes							
Homo sapiens	47715	5916	L1 (5916)	1	---	21	---
Danio rerio	7907	1795	L1 (637) including Tx1 (232); RTE(35); CR1 (890); Crack (35); Rex (120); NeSL (7); Hero (6)	31	4	135	7
Drosophila melanogaster	3690	254	Jockey (189); I (17); R1 (46); R2 (1); LOA (1)	---	---	---	---
Bombyx mori	6108	140	RTE (29); CR1 (36); I (20); Jockey (14); Ingi (2); R1 (31); R2 (1); R4 (4); Proto2 (3)	---	9	---	26

[a] *HMMER* hits have been filtered either only by bit score or by bit score and a minimum length of Length[minimum] = Length[HMM] - 30aa.

[b] Splitting into LINE clades was carried out only to RT hits filtered by bit score and hit length.

[c] Family numbers have been calculated with an 80 % threshold as defined by Wicker *et al.* (2007) and with a 60 % threshold as was used for *B. vulgaris* L1 LINE family classification (Figure 3.26).

Additionally, several genomes of lower animals (Figure 3.36 N-P) have been tested as well, in order to prove, if known LINE clades – apart from RTE and L1 – have been detectable. In summary, LINEs of 15 different clades have been identified directly from genomic data. That none of the analyzed plant genomes show any hits apart from L1 and RTE is an indication that these LINE clades dominate the genomes of higher plants.

3.3.4.2 Organization of plant LINEs into clades and subclades

In essence, Chapter 3.3.4.1 shows the extreme diversity of L1 LINEs in plants that contrasts with highly uniform RTE sequences. Subsequently, it will be investigated, if RTE and L1 reverse transcriptases group together in plant-specific clusters, or if they form subclades transcending species borders. One of the latter, the BNR subclade of L1 LINEs, has already been described in this thesis (Chapter 3.2.8).

Along with animal representatives, all plant L1 or RTE reverse transcriptases presented in the Chapters 3.3.3.1 and 3.3.4.1 have been aligned. Then, dendrograms have been constructed using the Neighbor-Joining method. Their branches have been colored according to the LINE source to allow easy species recognition (Figure 3.37).

The L1 LINE dendrogram has eight major branches, showing the presence of eight different L1 subclades, the BNR subclade being one of them (Figure 3.37 A). Four of these branches contain mainly members of higher dicotyledonous plants and have been named based on a prominent member: The LINE-CS subclade with 1372 sequences, the *LIb* subclade with 1393 sequences, the StL subclade with 1084 sequences, and the BNR subclade with 764 sequences (Sakamoto *et al.*, 2000; Yamashita and Tahara, 2006; Vogt, 2010; this thesis, Chapter 3.2)., Another branch, completely separated from the LINEs of higher plants, includes the selected animal LINEs.

The remaining three branches are constituted solely of 1465 grass LINEs of *Z. mays*, *O. sativa* and *B. distachyon* and can be summarized to three subclades named grasses I, II and III. Grass subclade I contains the previously described LINEs cin4 and BLIN and is consequently also called cin4 subclade, whereas the rice LINE *Karma* belongs to the Grass III subclade (Schwarz-Sommer *et al.*, 1987; Vershinin *et al.*, 2002; Komatsu *et al.*, 2003). Grass LINEs do not solely classify into these three subclades, but can also have representatives similar to *LIb*. Table 3.7 gives an overview of the identified L1 subclades and lists characterized LINE members, whereas Table 3.8 shows the number of plant members in each subclade.

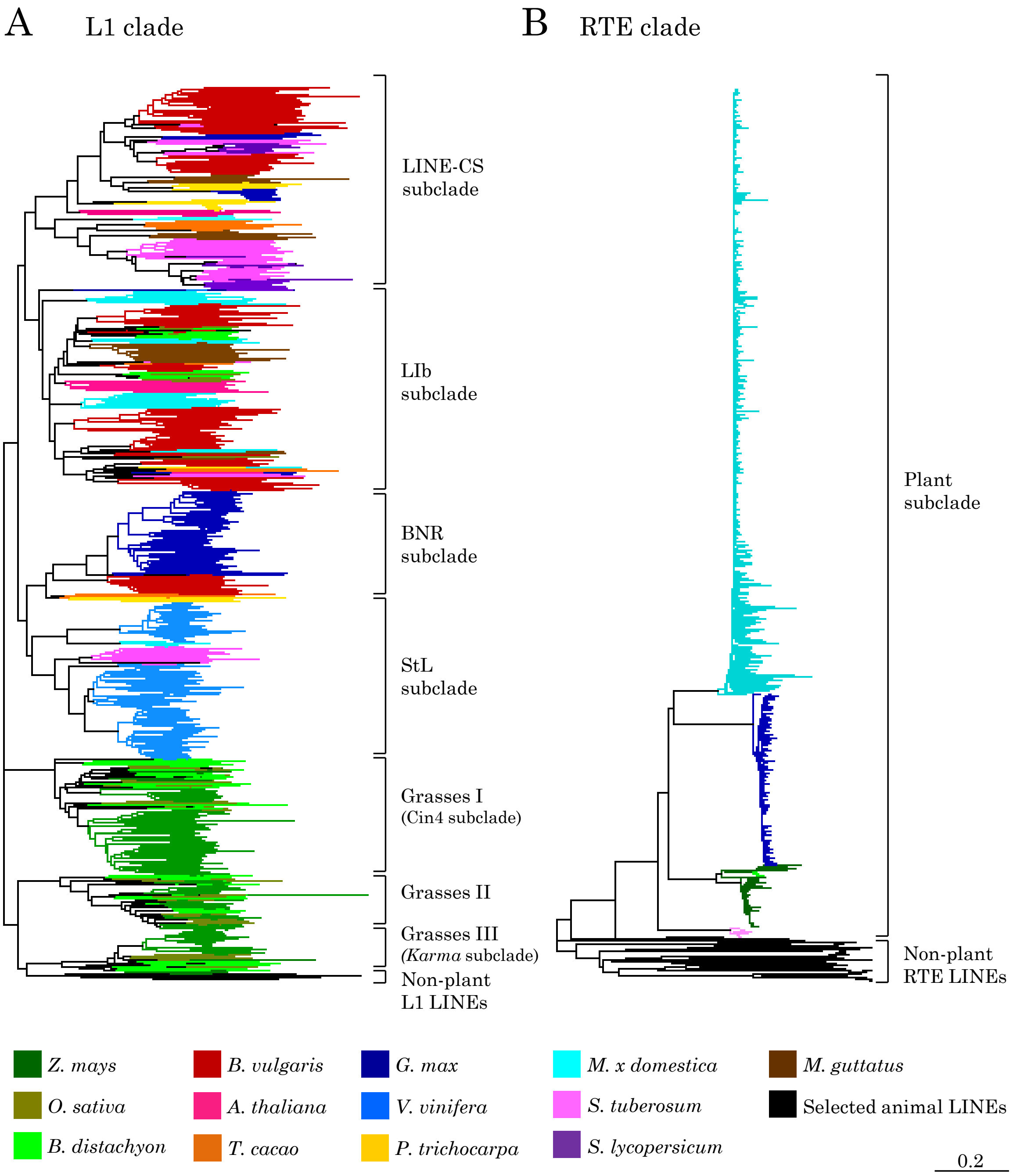


Figure 3.37: Relationship of plant L1 and RTE LINE reverse transcriptases.
A dendrogram has been constructed including all plant L1 **(A)** and RTE **(B)** LINEs presented in Chapter 3.3.3.1 and 3.3.4.1, as well as reference sequences of the corresponding clade described by Kapitonov *et al.* (2009). Amino acid sequences have been used. Branches have been colored according to the LINE source. Table 3.7 gives an overview of the plants represented in each L1 subclade. The scale bar represents 0.2 substitutions per site.
(A) 6081 plant L1 RT sequences from thirteen species have been included in the analysis. Additionally, 33 representative L1 RT sequences from animals have been included. Based on their reverse transcriptase domains, five subclades of plant L1 LINEs can be distinguished. They have been named based on the previously described LINEs *LIb*, BNR, LINE-CS, cin4 and StL.
(B) The RTE dendrogram is based on 1582 plant RT sequences from five species. Furthermore, 18 animal RTE references, along with the corresponding sequences from *D. rerio* and *B. mori* (Chapter 3.3.4.1) have been included.

Table 3.7: LINE representatives of L1 plant subclades and their hallmarks.

L1 subclade	No. of sequences	Average identity[a]	Reference members[b]	Length [nt]
LINE-CS subclade	1372	42 %	LINE-CS Shaline14 Belline families 2 to 5	4396 4346 ≈ 5000
LIb subclade	1393	44 %	*LIb* ATLINE1 Shaline16 Belline families 6 to 17	6454 5851 5758 > 5500
BNR subclade	764	59 %	Belline1 (BNR) family all LINEs of Table 3.2	> 6000
StL subclade	1084	57 %	StL1 family	
Grasses I (Cin4 subclade)	794	54 %	Cin4 BLIN	6822 6294
Grasses II	367	58 %	None	
Grasses III (*Karma* subclade)	304	59 %	*Karma*	7091

[a] Identity values based on RT amino acid sequences
[b] Accession numbers and host information can be found in Table 2.12

Members of the *LIb* subclade have been detected in all analyzed plant genomes. It is the only subclade that contains eudicot as well as monocot LINEs and is therefore considered the most common L1 subclade in angiosperms. LINEs of this subclade are also referred to as conventional plant LINEs.

Relatively many plant genomes contain LINEs of the LINE-CS type. They are marked by short element and ORF lengths as has been described for Belline2, Belline3, Belline4 and Belline5 (this thesis), LINE-CS and Shaline14 (Repbase Update).

Only few of the analyzed plant genomes have StL subclade members. Nevertheless, an exceptional high number of StL subclade LINEs has been detected in the *V. vinifera* genome. LINEs of this subclade have not yet been described in the literature. However, an *S. tuberosum* LINE family, StL1, and a related element from *V. vinifera* have been described in the frame of a recent diploma thesis (Vogt, 2010). Both LINEs belong to the StL subclade. Interestingly, none of the described ORF1 motifs, neither zinc finger nor RRM has been detectable in these LINE sequences.

LINEs of the BNR subclade have already been described in Chapter 3.2. In addition to *B. vulgaris*, *G. max*, *L. japonicus* and *P. trichocarpa*, whose RRM has been characterized before, the plant species *T. cacao* has been also found to contain BNR-like LINEs.

Table 3.8: L1 plant subclades and number of the detected members.

Species	LINE-CS subclade	*LIb* subclade	BNR subclade	StL subclade	Cin4 subclade (Grasses I)	Grasses II	*Karma* subclade (Grasses III)
***B. vulgaris* (*RefBeet* 0.1.1)**	450	663	125	-	-	-	-
A. thaliana	39	86	-	-	-	-	-
Z. mays	-	7	-	-	591	210	200
O. sativa	-	52	-	-	82	83	44
B. distachyon	-	92	-	-	121	74	60
G. max	114	9	584	-	-	-	-
M. x domestica	32	244	-	38	-	-	-
P. trichocarpa	110	6	30	-	-	-	-
T. cacao	63	24	24	-	-	-	-
V. vinifera	-	3	-	926	-	-	-
M. guttatus	120	154	-	-	-	-	-
S. lycopersicum	157	1	-	16	-	-	-
S. tuberosum	306	24	-	104	-	-	-

L1 LINE diversity contrasts strongly with RTE homogeneity in plants as was visualized by the dendrogram in Figure 3.37 B. RTE LINE RTs of five plant species have been compared. They all form plant-specific clusters. However, their short branch lengths, and average inter-species identity of more than 60 % (on basis of the RT amino acid sequence) indicate that they can be summarized in a single plant subclade of RTE LINEs.

Summarizing, L1 LINEs of higher plants – though highly divergent – group into several subclades encompassing members of many species, while RTE LINEs form a single subclade with families of highly identical plant elements.

4 Discussion

This thesis gives a comprehensive overview of the retrotransposon diversity in *B. vulgaris*. For a detailed depiction, exemplary sequence families of the LTR and Non-LTR retrotransposon orders have been analyzed. The focus has been on TE structure, family diversity and conservation, abundance, and chromosomal localization.

Furthermore, a large amount of genomic sequences allowed the genome-wide characterization of *B. vulgaris* retrotransposons based on their key enzyme, the reverse transcriptase. This enabled not only the integration of the identified families into a broader context, but also showed TE diversity of Ty1-*copia*, Ty3-*gypsy* and LINE elements. Since not extensively investigated in plants, the LINE population of *B. vulgaris* has been investigated in greater detail.

4.1 The LTR retrotransposon Cotzilla is a major component of *Beta* genomes

4.1.1 Structural characteristics, diversity and chromosomal localization of Cotzilla retrotransposons

Analysis of a *B. vulgaris* c_0t-*1* library, containing mostly highly repetitive sequences, led to the identification of Cotzilla retrotransposons. Based on the order of coding regions and RT domains, this TE family was grouped to a specific Ty1-*copia* lineage called Sireviruses. Hallmarks of this family in *Beta* are their abundance and extreme homogeneity.

Recently, a retrotransposon identical to Cotzilla1, named *SCHULTE,* has been detected by annotation of a *B. vulgaris* BAC containing a disease resistance-activating factor (Kuykendall *et al.*, 2009). Kuykendall *et al.* described a continuous *gag-pol* ORF and the presence of an additional hypothetical ORF. Detailled analysis described here, showed a separation of the *gag* and *pol* gene by a frameshift and supports an assignment of the hypothetical protein gene to the *env*-like ORFs. In the frame of this thesis, a comparative structural analysis of Cotzilla1 and a further member of the family, Cotzilla3 has been performed. In particular, the genomic organization of the Cotzilla family, their LTR variability and the distribution in the genus *Beta* and the chromosomal localization has been investigated.

Compared to conventional Ty1-*copia* retrotransposons, the *gag* reading frame of Cotzilla1 is extended and encodes several protein or nucleic acid binding domains typical for Sireviruses (Peterson-Burch and Voytas, 2002). Moreover, a putative *env* ORF upstream of the 3' LTR was detected which might enable infectivity of LTR retrotransposons mediated by transmembrane domains similar to the mechanism of retroviruses (Peterson-Burch *et al.*, 2000; Gallo *et al.*, 2003). If infective retrotransposons do exist in plants, infectivity might be generated by other yet unknown factors, since transmembrane domains cannot protrude plant cell walls. The putative Cotzilla1 *env* ORF does not contain a transmembrane domains and is instead characterized by a coiled coil domain and a proline-rich region. Similar coiled coil regions were also observed for the Sireviruses *SIRE*1 of *G. max* and ToRTL of *Solanum lycopersicum* (Havecker *et al.*, 2005). Because of its rigid structure, proline acts as a disruptor of regular secondary structure folds and is most commonly found in 'turn' motifs. The KRG and LTPL domains postulated for *env* ORFs of retroviruses and *env*-containing Ty3-*gypsy* retrotransposons (Lerat and Capy, 1999) were not detected. This is similar to the absence of these domains in the *env*-like ORF of the Ty1-*copia* element *SIRE*1 and supports the assumption that Ty3-*gypsy* and Ty1-*copia* retrotransposons acquired the *env*-like ORF independently (Kumar, 1998). Indeed, the existence of *env*-lacking Ty1-*copia* elements highly similar to *SIRE*1 has been reported in *G. max* genomes (Pearce, 2007), suggesting either acquisition or, more likely, loss of an *env*-like ORF.

As observed for Sireviruses (Gao *et al.*, 2003), the Cotzilla1 *pol* ORF is separated from the *gag* gene by a frameshift. The *gag* ORF of Cotzilla is terminated by a stop codon and a conserved sequence upstream of the stop allowing the corresponding RNA to form a hairpin structure. Interestingly, the DNA sequence of the loop itself is not conserved, indicating a strong selectional pressure on the hairpin-mediating sequence stretch. This sequence is not present in continuous *gag-pol* genes of typical Ty1-*copia* elements and could be responsible for a disassembly of the RNA translation machinery or a recruitment of proteins binding this hairpin, which might support the termination of *gag* translation or contribute to initiation of *pol* translation.

LTR sequences, which are specific for each retrotransposon family, give insights into retroelement diversity. Conventional Ty1-*copia* retrotransposons of the *B. vulgaris* SALIRE family – identified in context of this thesis – have LTRs highly diverse in sequence, indicating accumulation of mutations in ancient copies (Weber *et al.*, 2010). The structural reorganization or elimination of essential regulatory promotor motifs within the LTR might be responsible for limited SALIRE retrotransposition, as has been similarly described for Tnt1 (Le *et al.*, 2007). In contrast, Cotzilla LTRs are highly homologous and contain a plethora of regulatory motifs. This is a prerequisite for ongoing Cotzilla proliferation and a sign of recent amplificational bursts in the *Beta* genome.

Whereas the localization of Ty1-*copia* retrotransposons on chromosomes has been previously reported (Brandes *et al.*, 1997; Heslop-Harrison *et al.*, 1997; Francki, 2001; De Felice *et al.*, 2008; Jia *et al.*, 2009), only one study describes the chromosomal distribution of Sireviruses (Mroczek and Dawe, 2003). Members of the Cotzilla family and the maize elements Opie and PREM-2 analyzed by Mroczek and Dawe (2003) show a similar localization with a dispersed distribution along chromosomes.

LTR retrotransposons tend to integrate into other copies, thus generating nested insertions (SanMiguel *et al.*, 1996; Weber and Schmidt, 2009) which are often detected as clustered signals. As resolution of FISH to mitotic metaphase spreads is limited to approximately 3 Mb (De Jong *et al.*, 1999), it is assumed that each strong signal represents several clustered Cotzilla copies. Reduced numbers of both retrotransposon elements in centromeres, telomeres and 18S-5.8S-25S rDNA might be due to a rapid homogenization of these regions. Indeed, two satellite repeats are major components of *B. vulgaris* centromeres (Gindullis *et al.*, 2001b; Menzel *et al.*, 2008) and their amplification might lead to a depletion of newly integrated retrotransposon copies.

4.1.2 Cotzilla abundance might be caused by integration into heterochromatin

Cotzilla abundance has been proven using three different strategies. Firstly, Cotzilla elements form the largest group of retrotransposon sequences within a *c₀t-1* DNA library (Zakrzewski *et al.*, 2010). Secondly, 50 Cotzilla 5' LTR sequences have been detected in 18 Mb of *B. vulgaris* BAC end sequences. Transferring this ratio to 758 Mb, the genome size of *B. vulgaris*, a number of 2100 Cotzilla copies for a diploid genome can be estimated. Thirdly, hybridization of high-density filters revealed a 16-fold higher abundance of the Cotzilla family compared to the SALIRE family. This is in accordance with Southern hybridizations yielding strong Cotzilla1 signals after short exposure, and contrasts with weak SALIRE1 LTR signals detectable after very long exposure of 18 days. Summarizing, Cotzilla retrotransposons might represent one of the most abundant retroelement families in the *B. vulgaris* genome. It has a copy number similar to other highly repetitive Ty1-*copia* families such as CIRE1 from *C. sinensis* with 2200 copies (Kimura *et al.*, 2001) and a much higher copy number than *Reme1* from melon with 120 copies (Ramallo *et al.*, 2008) or the most abundant retrotransposon from grapevine *Gentil* with 212 copies (Moisy *et al.*, 2008).

Cotzilla retrotransposons make up to 3 % of the *B. vulgaris* genome, based on the following calculation: length of Cotzilla1 x copy number of Cotzilla / genome size of *B. vulgaris* (10.833 kb x 2100 / 758000 kb = 3 %). Thus, they are a prominent target for unequal homologous recombination or illegitimate recombination causing removal of Cotzilla retrotransposons and adjacent regions (Ma *et al.*, 2004). In contrast to retrotransposons with short LTR sequences, retrotransposons with long LTRs like Cotzilla elements, have been reported to be more frequently subjected to recombination mechanisms (Ramallo *et al.*, 2008). Thus, they might contribute to the antagonistic processes of genome expansion and contraction, which have a considerable impact on the restructuring of genomes.

FISH analysis of Cotzilla elements shows an exclusion from euchromatic regions and a preferential integration into heterochromatic regions. In *B. vulgaris*, the intercalary heterochromatin is well-characterized and mainly constituted of the satellite repeats pEV and pAp11 (Schmidt *et al.*, 1991; Schmidt and Heslop-Harrison, 1998; Dechyeva *et al.*, 2003). Since these regions have a lower gene density, they might tolerate accumulation of Cotzilla retrotransposons, and thus high copy numbers can be a consequence. Heterochromatic structures are typical for silenced genome regions and accompanied by DNA methylation and histone modifications, such as monomethylation of lysine 9 and 27 of histone H3 in *B. vulgaris* (Zakrzewski *et al.*, 2011). However, actively transcribed genes are not completely depleted in heterochromatic regions. There are numbers of reports describing active genes exclusively located in these regions, for example in the rice centromere 8 (Nagaki *et al.*, 2004). Furthermore, recent experiments demonstrated that the transcription of repetitive sequences is essential for the RNA interference-mediated heterochromatin assembly (Slotkin and Martienssen, 2007; Kloc and Martienssen, 2008). Detection of several EST sequences homologous to Cotzilla1 (Kuykendall *et al.*, 2009) as well as reduced methylation of few Cotzilla elements detected in genomic blot hybridization in this work indicates potential transcription and suggests putative activity of some Cotzilla members. Further studies are needed to confirm transcriptional and transpositional activity of Cotzilla retrotransposons. Presence of Cotzilla transcripts after 5-azacytidine treatment have been already verified by RT-PCR (Schmidt, 2010). 5-Azacytidine is a chemical that is incorporated into the DNA and replaces cytosine, but cannot be methylated. Thereby, the transcriptional repression by DNA methylation is lifted. Cotzilla transcription shows that some of the family members have functional LTR promotors or are in vicinity of external promotors that have the ability to interact with RNA polymerases, if the genetic environment is not repressive. Bisulfite sequencing of the Cotzilla body, LTR or the flanking region will show Cotzilla DNA methylation with a single nucleotide resolution. Moreover, future works might address the presence or absence of heterochromatic marks by immunostaining with antibodies against specific histone modifications. These experiments will show to which degree Cotzilla members are embedded into heterochromatic regions and thus further the understanding of TE repression.

4.1.3 Cotzilla retrotransposons are evolutionarily young

Fifty Cotzilla 5' LTR sequences have been detected in BAC end sequences which are characterized by an average identity of 96 % to Cotzilla1. This indicates the presence of many highly homologous and probably young Cotzilla copies in the *B. vulgaris* genome. Furthermore, these results indicate that Cotzilla3 is a diverged and older copy, not representative for the Cotzilla family. Hybridization of DNA from several *Beta* species with Cotzilla LTR shows its conserved distribution in all species of the phylogenetically youngest section *Beta* containing all cultivars. In the *Corollinae* and *Nanae* sections less copies with a lower similarity have been detectable, and in the wild beet genus *Patellifolia* or in species such as *S. oleracea* and *C. quinoa*, Cotzilla elements have not been detected. This gives a strong indication for massive bursts of amplification of Cotzilla retrotransposons after the separation of the section *Beta*. It is assumed that the majority of Cotzilla retrotransposons are evolutionarily younger than the SALIRE elements reported in Weber *et al.* (2010). This is in line with the calculated insertion times of 290 000 and 850 000 years for the individual copies Cotzilla1 and Cotzilla3, respectively, and 1.4 million years for the SALIRE1 copy. However, the calculated insertion times of single elements do not fully reflect the age of a whole family, which could be much older. For comparison, in *Medicago truncatula* most LTR retrotransposon members which have been analyzed, inserted within 400 000 years and only some have an age of 2 million years (Wang and Liu, 2008). Supposedly, the age and hence the divergence of SALIRE retrotransposons limited the possibility to identify full-length elements. Wicker and Keller (2007) estimated that after 790 000 years half of the rice retrotransposons have been truncated or otherwise rearranged, while Ma *et al.* (2004) suggest that after 6 million years the accumulation of mutations and recombinations make an identification of 50 % of the rice retrotransposons impossible. Summarizing, SALIRE elements are examples for ancient retrotransposons, which are in the progress of becoming eliminated, whereas Cotzilla retrotransposons show a recent burst in a short evolutionary time.

Together with SALIRE1, two Ty1-*copia* retrotransposon families in sugar beet have been isolated. With contrasting abundance, age, sequence diversity and chromosomal localization, they give insight in the heterogeneity of LTR retrotransposons populating a plant genome. Especially members of the highly repetitive Cotzilla family are likely to have a major impact on *B. vulgaris* genome evolution.

4.2 The BNR family represents a novel L1 subclade

4.2.1 Diversity, recent transposition and chromosomal organization of BNR retrotransposons

Full-length LINEs of the sugar beet BNR family have been identified by screening of a large-insert library, *in silico* analyses of database entries and targeted gap-closure of partial BAC sequences.

Remarkable differences in structure have been observed between members of the BNR family, in particular between BNR1 and BNR3. While BNR1 has only a single stop codon, BNR3 is structurally highly rearranged, containing two insertions, many frameshifts and internal stop codons, indicating an evolutionarily diverged family member. One of these insertions is a solo-LTR of an internally deleted *env*-like retrotransposon of the Elbe family, as concluded from the strong similarity of the LTRs to other *B. vulgaris* Elbe members (Cora Wollrab, TU Dresden, personal communication). The rearrangement of BNR3 has most likely taken place in several steps. First, integration of an *env*-like retrotransposon has disrupted this element. Later, the integrated *env*-like retroelement has probably been excised by recombination processes leaving behind the solo-LTR in BNR3. Unequal recombination is a common mechanism to reduce the number and transposition competence of retroelements in plant genomes (Devos *et al.*, 2002; Ma *et al.*, 2004). The origin and date of the upstream insertion in ORF2 remains elusive.

In contrast to BNR3, BNR1 is evolutionarily younger. The integration of BNR1 in a monomer of the subterminal satellite family pAv appears to be unique and is only detectable in two beet cultivars. This suggests that the BNR1 integration into pAv took place during domestication of beets dating back 2500 years ago (Lange *et al.*, 1999), indicating a more recent transposition compared to other LINEs: Southern hybridization supports the assumption of recent transposition and revealed a higher abundance of BNR in cultivars than in wild *Beta* species. Although many hybridizing fragments are conserved, higher copy number and polymorphic patterns in cultivars point to multiple transposition events during breeding, indicating expansion of the BNR numbers over the last 2500 years. The identification of nine structurally intact BNR members in the *B. vulgaris* genome further also favors the assumption of active, but silenced TEs. However, sequencing of domain IV to VII of the BNR reverse transcriptase gene of seven *Beta* species demonstrated the high diversity even within cultivars which is a result of accumulation of mutations over a long time period. Taken together, these results signify a recent activity of a relatively old LINE family.

FISH analyses of plant LINEs have only rarely been performed. So far, only the barley BLIN and the LINE-CS from cannabis have been physically mapped by FISH (Sakamoto *et al.*, 2000). BLIN occurs in a dispersed distribution with weak signals on the majority of the chromosomes, while LINE-CS is exclusively located at the end of the long arm of the Y chromosome. The chromosomal distribution of BNR is different and stronger, consistent with the results from gel blot experiments. The presence of strong signals, mostly detectable as doublettes, indicates a clustered organization with multiple BNR copies per chromosomal region. This has also been observed by FISH analysis of BNR fragments generated by PCR from genomic DNA (Schmidt *et al.*, 1995). Resolution of FISH to mitotic metaphase spreads is limited and therefore, it is thought that each strong signal represents BNR copies in a 3 Mb chromosomal interval (De Jong *et al.*, 1999). The reduced number of BNR copies in some centromeres is in contrast to *Beta* Ty3-*gypsy* retrotransposons which exclusivly insert into centromeric satellite arrays mediated by a their encoded chromodomain (Weber and Schmidt, 2009).

The significant increase of Southern hybridization signals using probes located at the 3' end suggests the 5' truncation during the target-primed reverse transcription of LINEs. Full-length LINE integration occurs in only 15 % of all transposition events in human (Szak *et al.*, 2002) and plant LINEs suffer also from a high rate of 5' truncation as has been observed (Sakamoto *et al.*, 2000; Wenke *et al.*, 2009).

4.2.2 A conserved RRM domain in ORF1 characterizes a novel subclade of L1 plant LINEs

The most striking feature of the BNR family is the presence of a novel mRNA binding motif in ORF1 including two conserved domains (Heitkam and Schmidt, 2009). Unlike all plant non-LTR retrotransposons reported to date, the ORF1 of BNR contains a conserved RNA recognition motif (RRM) located at the N-terminus. Very recently, the presence of remote homologues to RRM domains deviating significantly from the RRM consensus was discovered in eukaryotic LINEs (Khazina and Weichenrieder, 2009). The ORF1 RRM domain reported in this thesis is different and has a considerably higher similarity to the functional consensus RRM motif (Maris *et al.*, 2005; Lunde *et al.*, 2007) and is lacking a typical zinc finger motif found in all plant LINEs identified so far. An overview of ORF1 structure is presented in Figure 4.1.

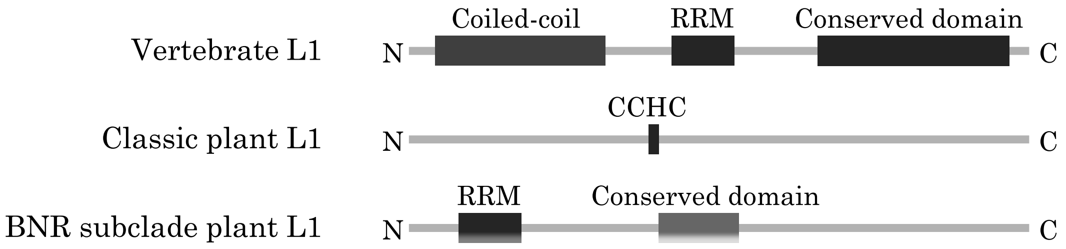


Figure 4.1: Structure of ORF1 proteins in selected L1 LINEs.
The thin bar represents the entire length of the protein encoded by ORF1. Thicker bars show the position of defined domains in vertebrate and plant LINEs (according to Martin, 2006; Khazina and Weichenrieder, 2009). Structure motifs are as follows: RRM – RNA recognition motif; CCHC – zinc finger; Coiled coil – Coiled coil secondary structure motif). The conserved domains of vertebrate and plant LINEs are not related to another and are of unknown function.

RRMs are the most common and evolutionarily oldest eukaryotic RNA-binding domains (Maris *et al.*, 2005; Lunde *et al.*, 2007). RRMs can bind nucleic acids by a combination of salt bridge formation, stacking and hydrophobic interactions and are ubiquitious in proteins involved in RNA processing, editing and degradation. Significantly, the RRM-containing ORF1 is not only present in the sugar beet BNR family, but also has been identified in LINEs of soybean, lotus and poplar (Heitkam and Schmidt, 2009). Comparison of the reverse transcriptase genes of LINEs containing the typical eight domains described by Xiong and Eickbush (1990) and Malik *et al.* (1999) revealed that the RRM-containing LINEs of sugar beet, soybean, lotus and poplar form a closely related group in the L1 clade of eukaryotic LINEs. Furthermore, a second conserved domain was identified, however its function is still unclear. The function of this domain downstream of the RRM remains elusive although computational structure predictions revealed several conserved α- helices and ß-sheets. Its conservation across diverse plants species indicates its functional importance which might include dimerization or contribution to the mRNA binding of the RRM.

The BNR-type LINEs reported here have been found to form a separate L1 subclade designated as BNR subclade. This result is also supported by the analysis of ORF2 sequences of retroelements in the BNR subclade which showed a higher similarity to each other than to other plant LINEs. It is likely that this BNR subclade will grow with the increasing number of annotated plant genome sequences produced by next generation sequencing technologies (reviewed in Mardis, 2008).

4.2.3 Evolution of LINEs carrying an RRM ORF

The N-terminal RRM of the BNR family could provide the nucleic acid chaperone activity that is encoded by the ORF1 and has been reported for mammalian ORF1 proteins (Martin and Bushman, 2001). The functions of a nucleic acid chaperone include the facilitation of rearrangements of nucleic acids into their thermodynamically most stable form and the promotion of melting and annealing of nucleic acids (Martin, 2006). An example for the recognition of L1 transcript by an RNA recognition motif has been documented in human: 3' ends of L1 transcripts bind to the RNA-binding protein NXF which stimulates mRNA transport from nucleus to cytoplasm (Lindtner *et al.*, 2002). In contrast to the BNR-encoded RRM, NXF is a cellular human protein and not encoded by LINEs. However, this binding ability shows, that RRMs can principally bind and transport LINE-mRNA.

For the evolution of LINEs containing a RRM in ORF1 several scenarios are possible. First, the discovery of degenerated RRM homologues in eukaryotic LINEs (Khazina and Weichenrieder, 2009) might suggest the following evolutionary route: RRM domains could have been present in all LINEs and an aquisition of additional RNA-binding domains such as zinc finger motifs during evolution resulted in the loss of the primary RRM. However, BNR-like LINEs did not aquire any additional RNA-binding motifs and the RRMs reported by Khazina and Weichenrieder (2009) differ strongly from the RRMs reported by Maris *et al.* (2005) and as described in this thesis.

Alternatively, it is more likely that an integration of a LINE, shortly truncated at the 5' end, occurred directly in or downstream of a protein gene having an RRM (Figure 4.2). LINEs propagate by target site primed reverse transcription which is an error prone process resulting in truncation of the majority of copies (Szak *et al.*, 2002). The truncation of the LINE had to be upstream of ORF2, so that the ORF2 remains intact. A read-through transcription of the gene and the 5' truncated LINE generates chimeric LINE transcripts which are amplified by subsequent retrotransposition. If this model is correct, the rearrangement is a very ancient event which must have taken place before the separation of *B. vulgaris*, *P. trichocarpa*, *L. japonicus*, *G. max* and *G. tomentella*. Furthermore, the similar structure and organization of domains in ORF1 as well as the close relationship of ORF2 suggests that LINEs of the BNR subclade originate from a common ancestor.

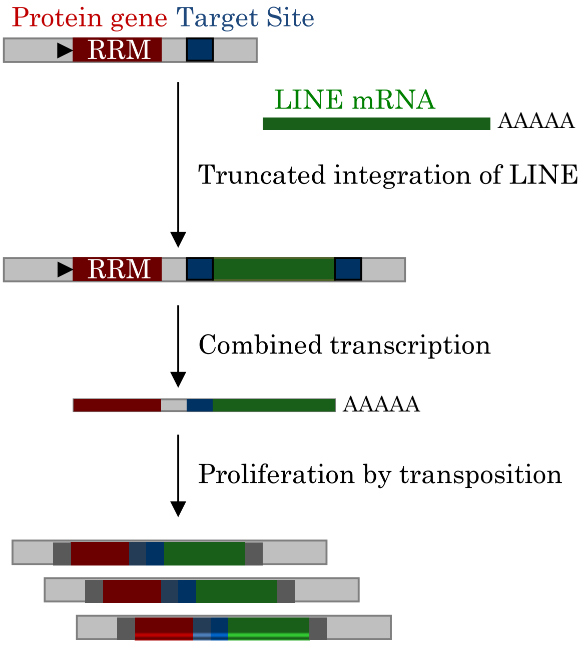


Figure 4.2: Possible mechanism of BNR generation by modular evolution.
This scenario begins with the retrotransposition of a 5' truncated LINE copy directly into a protein-coding ORF, next to an RRM. Read-through transcription by the gene's own promotor produces combined transcripts coding for an original ORF2 and a chimeric ORF1. If both are functional, the fused LINE can proliferate by retrotransposition.

LINEs of the BNR subclade are an example for rearrangements and modular evolution of retroelements and their success in colonizing plant genomes.

4.3 Retrotransposon evolutionary dynamics in *B. vulgaris* and other higher plants

4.3.1 A HMM-based algorithm allows to asses retrotransposon diversity in genomes of *B. vulgaris* and other angiosperms

The availability of complete or nearly complete genome sequences of higher plants provide the basis for the detection of TEs on a genome-wide scale and permit cross-species comparisons. However, poor conservation of TEs presents a problem for homology-based identification methods. Here, an algorithm based on a Hidden Markov Model (HMM) has been applied to detect *B. vulgaris* reverse transcriptases of the Ty1-*copia*, Ty3-*gypsy* and LINE orders. Sequence comparison allowed classification into the corresponding lineages according to Kapitonov *et al.* (2009) and Llorens *et al.* (2009) and thus gave insights into plant genome diversity.

HMMs have been widely used for the modeling of protein sequence families, but only rarely for DNA sequence family modeling. Therefore, common HMM tools are specialized for work with protein data. HMM algorithms to identify TEs on a DNA level are in development, but not yet applicable (Edlefsen and Liu, 2010). By translation of the genomic sequence into all frames, this obstacle has been overcome.

At first, it has to be considered that the numbers of reverse transcriptases shown in this thesis are probably underestimating their real numbers. In order to select for full-length *B. vulgaris* reverse transcriptases, incomplete sequences detected by the search algorithm have been eliminated. These short sequences were mostly characterized by degenerated RTs, RTs that were only partially covered by a contig, or RTs that harbored a frameshift in the translated sequence. Furthermore, the *RefBeet* assemblies do not yet contain the full *B. vulgaris* genome. Direct comparison of genomic and computational data shows this difference: Southern hybridization of Cotzilla (Figure 3.5) indicates that, most likely, more *B. vulgaris* Sirevirus sequences exist than have been detected by bioinformatics (Figure 3.25). The same has been observed for RTE LINEs of the *Ghost* family (Figure 3.26 and Figure 3.34 B). These findings indicate an underrepresentation of highly similar *B. vulgaris* retrotransposons in the *RefBeet* assemblies. This data loss occurred during the assembly process of short sequence reads generated by next generation methods. The presence of many highly repetitive short sequences often leads to multiple assembly possibilities. To avoid bias or incorrect assembly, many of these repetitive sequences have been removed (Heinz Himmelbauer, Centre for Genomic Regulation Barcelona, personal communication).

With help of the HMM-based detection method, *B. vulgaris* RTs of the Ty1-*copia* and Ty3-*gypsy* retrotransposon order have been identified (Figure 3.24; Figure 3.25; Table 3.3). They are highly diverse in sequence, and form several lineages and families. Annotation of complete or partial genomic sequences of *A. thaliana*, *L. japonicus*, *M. truncatula*, *O. sativa*, *Z. mays* and *G. max* also revealed a multitude of different retrotransposon sequences (Pereira, 2004; Holligan *et al.*, 2006; Wang and Liu, 2008; Tian *et al.*, 2009; Baucom *et al.*, 2009; Du *et al.*, 2010).

Ty3-*gypsy* retrotransposons of legumes, rice and thale cress have been especially well characterized (Wang and Liu, 2008; Du *et al.*, 2010), allowing a comparison with the related *B. vulgaris* elements. Based on an RT alignment, they also form lineages with multiple families, as has been shown for *B. vulgaris* Ty3-*gypsy* retrotransposons (Figure 3.24; Table 3.3). They also belong to the errantiviral, chromoviral and Tat clade, indicating that the composition of *B. vulgaris* Ty3-*gypsy* retrotransposons is typical for higher plants.

A comparison of the *B. vulgaris* Ty1-*copia* clades with the retrotransposable content of other plant genomes is more difficult: Having at least four different designations, the naming of plant Ty1-*copia* lineages has not been consistent throughout literature. For example, the Sirevirus lineage (Llorens *et al.*, 2009) has also been named 'Maximus' (Wicker and Keller, 2007), 'Endovir-like' (Holligan *et al.*, 2006) or 'Copia4' (Wang and Liu, 2008). Similar to the classification scheme that was introduced in Figure 1.2, and applied to *B. vulgaris* Ty1-*copia* retrotransposons (Figure 3.25; Table 3.3), the classification by Llorens *et al.* (2009) has always been used. Ty1-*copia* elements of *A. thaliana*, *O. sativa*, *L. japonicus* and *M. truncatula*, belong to the Retrofit, Sirevirus, Oryco, Tork and CoDi-D clade (Holligan *et al.*, 2006; Wicker and Keller, 2007; Holligan *et al.*, 2006). In legumes, nearly 70 % of the Ty1-*copia* have been found to belong to the Sireviruses (Du *et al.*, 2010). In this thesis, all of these clades have also been identified in the *B. vulgaris* genome. However, using the HMM-based method, only 30 % of the detected *B. vulgaris* Ty1-*copia* sequences belong to the Sirevirus clade. A much higher abundance of Cotzilla-like Sireviruses has been discussed though (Chapter 4.1.2). Again, this contradiction can be explained by the elimination of highly repetitive sequences from the *RefBeet* assemblies.

The ABC clade of Ty3-*gypsy* and the pCreto clade of Ty1-*copia* retrotransposons (Llorens *et al.*, 2009), detected in low numbers in the *B. vulgaris RefBeet* assembly, have not been previously described in higher plant genomes. Instead, they are more common in protostomes and deuterostomes (e.g. nematodes), or fungi, respectively (gydb.org; Llorens *et al.*, 2010). Their presence in the database can be explained either biologically by horizontal transfer, or, more likely, methodically by contamination of the sequenced *B. vulgaris* samples. A related inquiry gave the information that 17 % of the sequence data used for the *RefBeet* assemblies correspond to contaminants (Heinz Himmelbauer, Centre for Genomic Regulation Barcelona, personal communication). A PCR could be performed to verify the presence of those retrotransposon families in *B. vulgaris*.

Summarizing, the HMM-based approach presents an opportunity to obtain an overview of the retrotransposable content of the *B. vulgaris* genome. The types of LTR retrotransposons detected are comparable with those isolated from other angiosperm genomes, indicating that *B. vulgaris* has a plant-typical retrotransposon composition.

4.3.2 The LINE landscape of *B. vulgaris*

4.3.2.1 The *B. vulgaris* LINE population as an example to specify typical LINE characteristics

Apart from RTs of LTR retrotransposons, the HMM-based algorithm also enabled the isolation of non-LTR retrotransposon reverse transcriptases of the LINE order. These *B. vulgaris* sequences show very high sequence diversity (Figure 3.26; Figure 3.28 B). Nevertheless, these LINEs have been classified as members of only two out of 28 clades (Kapitonov *et al.*, 2009): The vast majority of the detected *B. vulgaris* RTs belongs to the L1 clade, but is characterized by very different sequences forming a number of subclades and families. In contrast, only three of the detected *B. vulgaris* sequences belong to the RTE clade of LINEs.

The existence of at least seven plant L1 subclades (Figure 3.37; Table 3.8; discussed in Chapter 4.3.3), of which three have been detected in *B. vulgaris*, shows that BNR and BvL-like L1 LINEs represent only a minority of the total *B. vulgaris* L1 content (summarized in Figure 4.3). Therefore, *B. vulgaris* LINEs have been examined in greater detail to obtain a comprehensive overview of the LINE population, and to further illuminate the similarities and differences between LINEs of the same species, but of different clades, subclades and families.

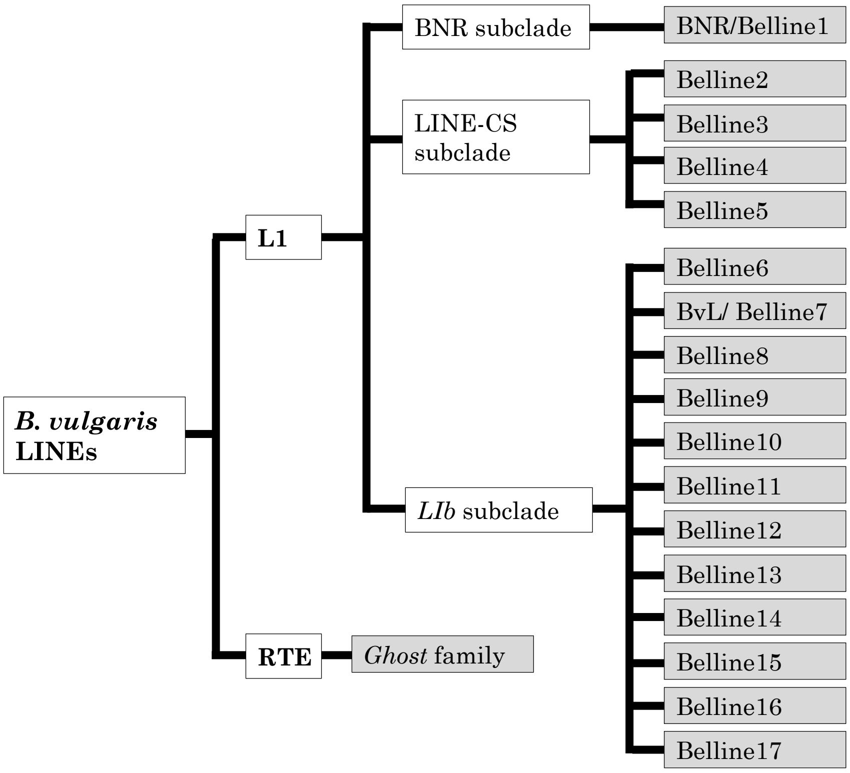


Figure 4.3: Classification of *B. vulgaris* LINEs based on their RT.
All detected *B. vulgaris* LINE families and their position in the classification scheme have been summarized. Families have been marked in grey.

According to Wicker *et al.* (2007), two elements belong to the same family, if they share 80 % (or more) sequence identity in at least 80 % of their coding or terminal repeat regions, or in both. This rule has been applied to LTR retrotransposons, whose families are distinguished based on the similarity of LTR sequences. However, for classification of very diverse plant L1 families, the rule of Wicker *et al.* is too strict (as shown in Chapter 3.3.3.1; Figure 3.27). In agreement with the phylogeny of LINE RTs (Figure 3.28 A), a minimal shared identity of 60 % has been applied to assign *B. vulgaris* L1 families (Figure 4.3). Based on these premises, 1238 *B. vulgaris* L1 LINE reverse transcriptases have been grouped into 17 Belline families. An RT alignment of representative family members illustrates distinct differences in amino acid sequences outside of conserved domains. These variable residues can be used to define family-specific sequences.

Based on the amino acid composition of the RT region, twelve Belline families belong to the *LIb* subclade (including Belline7/BvL), four to LINE-CS and one to the BNR subclade. In order to test, whether the other LINEs also have subclade-specific features, complete family members have been extracted and analyzed. All Belline families except BNR/Belline1 have a zinc finger in ORF1, a structural motif reported for canonical plant LINEs. Belline2, Belline3, Belline4 and Belline5 members are structurally very similar to each other. As discussed for LINE-CS retrotransposons, they are unusually short in sequence, especially when compared with ORF1 and ORF2 of other Belline families. The ORF2 structure of *LIb* subclade LINEs (Belline6-Belline17) is identical, however, there are differences in ORF1 length as well as in the position of the zinc finger. Hence, flexibility in length and structure of *LIb*-typical LINE can be deduced, which has not been found for Belline members of the LINE-CS and BNR subclades.

For the first time, different LINE families of one species have been comparatively analyzed regarding their abundance and amplification by Southern hybridization (Figure 3.30). Since a relatively low threshold of 60 % identity has been used for classification of L1 families, not all Belline family members predicted computationally have been detectable (Figure 3.30, Belline5 family) due to higher hybridization stringency of 75 %. However, it has been possible to describe the abundance of family members, of which the probes were derived from, in plants of the genera *Beta* and *Patellifolia*. Comparing the analyzed Belline families, different hybridization patterns have been observed: Members of three families (BNR/Belline1, BvL/Belline7, Belline17) occur in all species of the genus *Beta*, but are very reduced in *Patellifolia*, indicating amplification after the separation of both genera. Members of the Belline17 family are the only LINEs observed that proliferated predominantly in the closely related sections *Corollinae* and *Nanae*. High abundance exclusively in the section *Beta* has been shown for two LINE families (Belline2, Belline12). This points to increased levels of retrotransposition after the species separated. However, single strong Belline2 signals have been observed also in the genera *Patellifolia* and even *Spinacia*, indicating that the Belline2 family is an ancient LINE family. Two of the families tested (Belline5, Belline9) showed only faint hybridization in all species. This can be explained by two different reasons: Since the Belline5 family is very divergent, the probe did probably not detect all family members. The Belline9 family however, contains only few members as has been also indicated by bioinformatic analysis.

In conclusion, LINE evolution in the genera *Beta* and *Patellifolia* did not follow a general pattern. Even families belonging to the same L1 subclade vary in abundance and rate of proliferation. This suggests that some Belline families have been more active than others, however, it has not been possible to calculate valid copy numbers as reported for LTR retrotransposons (Chapter 4.1). Increases in retrotransposition have been observed after differentiation of related sections or genera. However, amplification of Belline families has not been concerted, but occurred most likely at different points in the evolutionary time scale.

Summarizing the FISH analyses of this thesis and Wenke *et al.* (2009), the chromosomal localization of four *B. vulgaris* LINE families have been investigated: BNR/Belline1 LINEs are representatives of the BNR subclade, and the Belline2 family belongs to the LINE-CS subclade. Since the *LIb* subclade contains the most Belline families, LINEs of the Belline17 family have been selected for FISH additional to the BvL/Belline7 family (Wenke *et al.*, 2009). LINEs of all analyzed families have been found to be evenly distributed along all chromosomes, with a slightly higher abundance in distal chromosomal regions. This has also been stated for maize LINEs by computation of an *in silico* FISH (Baucom *et al.*, 2009). In *C. sativa* however, the vast majority of the LINE-CS retrotransposons is situated at the distal ends of the Y chromosome (Sakamoto *et al.*, 2000). Those or other pecularities have not been found in this thesis. Instead, a general trend towards random insertion into gene-rich euchromatic regions has been observed as seen by hybridization to interphase nuclei. Baucom *et al.* (2009) generalized this integration pattern for low-copy TE families (most LINE families), whereas high-copy families have been reported to most often accumulate inside other retrotransposons.

Apart from L1-like retrotransposons, RTE LINEs also exist in the *B. vulgaris* genome as shown by bioinformatics and PCR. Though largely underrepresented in the genomic databases, RTE LINEs of the *Ghost* family are at least middle repetitive in the sugar beet genome as proven by high density filter hybridization. During the assembly process of the genomic draft, highly repetitive sequences have been removed from the database as has been discussed in Chapter 4.3.1. Furthermore, RTE LINEs detected in other plant genomes (Chapter 3.3.4.1) have been extremely homogenous. Therefore, it is assumed that RTE LINEs of the *Ghost* family are highly similar as well.

In conclusion, LINEs of *B. vulgaris* are very different in sequence, structure and abundance. For L1 LINEs general tendencies towards the formation of many families with relatively few members and high diversity can be observed.

4.3.2.2 Control of LINE transcription by small RNAs

Active TEs are highly mutagenic and can target gene-rich regions, cause chromosome breakage, or recombination. In addition to TE expansion control by intra- or interelement recombination, the expression of most TEs are suppressed by several epigenetic pathways (Slotkin and Martienssen, 2007; Lisch, 2009). Three pillars support the epigenetic network and can be examined in order to gain a full view on the processes taking place: DNA methylation, histone marks and small RNAs. These mechanisms interact in a complex fashion, with 21 nt sRNAs mediating post-transcriptional silencing and 24 nt sRNAs guiding RNA-dependent DNA methylation and heterochromatin maintenance (reviewed by Simon and Meyers, 2010). Here, a first insight into the epigenetic regulation of plant LINEs has been gained by investigation of small RNAs homologous to TEs.

Most analyzed LINE family members exhibit only a low number of matching small RNAs. They are probably not transcribed, because they are not in vicinity of a functional promotor motif. Or, they might be deeply buried in heterochromatized areas, as a lack of siRNAs has been described for TEs in these regions (Matthias Zytnicki, INRA Versailles, personal communication). Furthermore, methylated cytosines as present in heterochromatic regions mutate faster, leading to sequence erosion and permanent inactivation. Therefore, smaller numbers of sRNAs are derived from older TEs (Cantu *et al.*, 2010).

However, the L1 families Belline6, BvL/Belline7, and Belline17, as well as the RTE LINE family *Ghost* contain at least one member homologous to a relatively high number of sRNAs. Compared to the total amount of sRNAs, especially the number of 24 nt sRNAs is increased. This shows that these LINEs are probably transcriptionally silenced by RNA-directed LINE methylation and heterochromatization. Ony few 21 nt sRNAs match the investigated LINEs, indicating no necessity for posttranscriptional silencing.

The vast majority of complementary sRNAs match the LINE 3' UTR or RNaseH region, in both, sense and antisense orientation. This has been the case for all LINE members analyzed – L1 and RTE LINEs. These sRNAs might not only suppress transcription of retrotranspositionally competent TEs, but also the read-through transcription of similar 5' truncated LINEs. Rearranged and 5' truncated LINEs often occur in eukaryotic genomes, as has been shown for BNR, BvL and LINE-CS by comparative Southern hybridization (Chapter 3.2.5; Wenke *et al.*, 2009; Sakamoto *et al.*, 2000). L1 5' truncation in human has been even measured to account for 85 % of all LINE insertions (Szak *et al.*, 2002).

Similar to the work performed in this thesis, Cantu *et al.* (2010) mapped sRNAs to TEs in wheat. They reported that sRNAs match specific areas of repetitive elements exhibiting a relatively high diversity (e.g. LTRs of LTR retrotransposons, or TIRs of MITEs). The preferred LINE regions reported in this thesis are also characterized by high sequence variability. For human L1, the presence of many L1 elements in the transcriptome has been detected (Rangwala *et al.*, 2009). Though the mechanisms of epigenetics are slightly different in animals, sRNAs are also thought to play a role in TE silencing. Artificially generated sRNAs, for example, effectively suppressed L1 transposition in cell culture (Soifer *et al.*, 2005). However, efforts to detect sRNAs produced from human LINE-1 still failed (as reviewed by Soifer, 2006; Fedorov, 2009).

Summarizing, this thesis also presents new data describing the LINE regions where correponding sRNAs originate from, and provides initial insight into the epigenetic regulation of plant LINEs.

4.3.3 Plant L1 LINEs can be classified into at least seven subclades, whereas plant RTE LINEs are characterized by high homogeneity

The *B. vulgaris* LINE population is characterized by a high degree of variability. Diversity of L1 LINEs in plants has also been reported in the genomes of *A. thaliana*, *Z. maize* and in those of the cotton genus *Gossypium* by bioinformatics or PCR with degenerate RT primers (Noma *et al.*, 2000; Schnable *et al.*, 2009; Baucom *et al.*, 2009; Hawkins *et al.*, 2008; Hu *et al.*, 2010). These studies identified a separation of a species' LINE content into numerous L1 families with low sequence identities to each other. However, the work described in this thesis is the first to compare several LINE families, originating from one species, on the level of sequence, structure and chromosomes.

Because of the insufficiency of available data on plant LINEs, a comparative investigation to analyze LINE diversity has been performed in this thesis. Twelve additional plant genomes have been queried for LINE reverse transcriptases. According to the classification summarized by Kapitonov *et al.* (2009), the detected RTs have been organized into clades and families (Figure 3.36; Table 3.6). Out of 28 currently known LINE clades, only L1 and RTE elements have been detected to be present in higher plants. Whereas members of the L1 clade have been found in all analyzed plant genomes, RTE homologues have only been detectable in five plant genomes. Since highly repetitive sequences are difficult to assemble, their sequences are often excluded from the published genomic data. As discussed above, this has also been the case for the current assemblies of the *B. vulgaris* genome.

In order to investigate the organization of plant LINEs into clades and subclades, all identified L1 and RTE LINE reverse transcriptases have been compared to each other in two comprehensive dendrograms (Figure 3.37). Based on their RT, all identified L1 plant LINEs group into seven L1 subclades, named referring to an already characterized member of this branch. L1 LINEs of eudicots group into the *LIb*, LINE-CS, BNR or StL subclade, whereas L1 LINEs of monocots can be assigned to the *LIb*, Grasses I (cin4), Grasses II or Grasses III (*Karma*) subclade (Table 3.8).

LINEs of the *LIb* subclade occur in all of the analyzed angiosperm genomes and is the only subclade present in genomes of monocots and eudicots. Therefore, this subclade is probably the evolutionarily most ancient, and contains canonical plant LINEs often described (Yamashita and Tahara, 2006; Noma *et al.*, 2000; Wenke et al., 2009). LINEs of the LINE-CS subclade are not present in the monocots, however in nearly all eudicots analyzed. All characterized LINE-CS-like retrotransposons are unusually short in sequence (< 5000 bp), encoding shortened ORF1 as well as shortened ORF2 proteins. LINEs of the BNR and StL subclade are much less common in plant genomes. Interestingly, LINEs of these two subclades do not contain a CCHC-type zinc finger in ORF1. Members of the StL subclade have not been found to substitute this feature, whereas BNR-like LINEs contain an RNA recognition motif instead (this thesis; Heitkam and Schmidt, 2009). Interestingly, BNR elements have also been identified in *T. cacao* additional to the previously analyzed members in *B. vulgaris*, *L. japonicus*, *G. max* and *P. trichocarpa*. The last three subclades detected, Grasses I-III, have only been found in the three grass genomes analyzed and have not been investigated further. Future works might focus on the structural differences of members of the L1 subclades. Isolation and annotation of full-length LINEs of each subclade and species is a laborious, but important work, helping to define structural properties of each subclade.

Additional to the analysis of plant L1 LINEs, the detected RTE LINEs have been compared as well (Figure 3.37 B). RTE LINEs have been observed to be highly homogenous in the plants, where they have been detected, with high average identitity values of more than 90 % in *G. max*, *M. x domestica* and *S. tuberosum*. Even cross-species comparisons show high identity values of more than 60 %, indicating very high inter-species similarity, even between RTE elements of monocot and eudicot genomes.

In summary, the LINE population of higher plants has been found to be highly variable, though predominantly composed of L1 and RTE LINEs. Based on their RT, L1 LINEs can be classified into at least seven subclades containing many families. In contrast, RTE LINEs are highly homogenous and constitute most likely only a single family per plant genome. A comparison with the LINE population of *B. vulgaris* (discussed in Chapter 4.3.2) shows that the *B. vulgaris* LINE landscape is typical for higher plants containing both, homogenous RTE and diverse L1 LINEs.

4.3.4 Evolutionary strategies leading to the formation of different LINE populations in higher plants and mammals

Mammalian LINE-1 interspersed elements are not only the namesake of plant L1 LINEs, they also show a high degree of conservation in their RT domains. In mammals, a single lineage of successive L1 families has been generated in approximately 93 million years of replication and evolution (Khan *et al.*, 2006; Lee *et al.*, 2007). These L1 retrotransposons have nearly identical sequences and constitute 20 % of the total genomic content of *H. sapiens*.

Plant L1 LINEs do neither show the high identity values nor copy numbers of their human counterparts. In all analyzed plant genomes, they are marked by high diversity and structural variation. It was demonstrated that angiosperm LINEs can be divided based on their RT sequence into at least seven subclades. Each plant genome analyzed contained LINEs of at least two L1 subclades, and each subclade occured in at least three of the investigated species. This shows a general trend towards L1 diversification in plants. Each subclade can be subdivided again into numerous families with low copy numbers.

Comparable to plants, high L1 diversity has been observed in fish genomes containing many L1 families, with less than 100 copies each (see *D. rerio* in Figure 3.36 and Table 3.6; Volff *et al.*, 2003; Furano *et al.*, 2004), indicating that the mammalian evolutionary history of L1 LINEs might be unique. Different evolutionary retrotransposon dynamics have probably caused the different situations in mammalian, fish and plant genomes. Three of the parameters influencing retrotransposon accumulation and deletion have been suggested by Eickbush and Furano (2002): (1) Activation of LINEs and retrotransposition, (2) effect of a retrotransposed copy on gene activity, and (3) the rate of recombination. Interplay of these factors ensures a balance between accumulation and removal of retrotransposons. Even though the number of LINE copies in plant, fish and mammalian genomes differ by orders of magnitude, the bulk of human LINE retrotransposition is accounted for by only a few very active L1 members (Brouha *et al.*, 2003). These might be existent in plants as well. *B. vulgaris*, for example, has been shown to contain many L1 elements that are structurally intact and theoretically capable of retrotransposition. Under these circumstances, the relatively low number of plant L1 LINE family members implies that the intact LINEs are either epigenetically silenced or removed shortly after retrotransposition.

Removal of full-length LINEs by homologous recombination has been shown to take place in human (Boissinot *et al.*, 2001; Han *et al.*, 2008). Although deletion of large DNA regions has been reported, L1 recombinations occur relatively rarely, if set in relation to the elevated L1 copy number. Likewise, recombination has been described in plants as mechanism for TE elimination and genome shrinkage (Bennetzen *et al.*, 2005). Compared to mammals, much higher recombination rates have been calculated in angiosperms (Kejnovsky *et al.*, 2009). Therefore, plant retrotransposons probably need higher activity levels than the mammalian retroelements in order to populate plant genomes in high numbers. Summarizing, L1 activity does not seem to balance recombination rates in plants or fish, leading to a continual turnover of LINEs limiting L1 copy number. L1 diversification might be an alternative evolutionary route to increase the LINE population without increasing the rate of recombination.

The population of RTE LINEs is also different in animal and plant genomes. Animal RTE LINEs detected in *B. mori* and *D. rerio* have been observed to be far more divergent than those detected in plant genomes (Figure 3.36; Table 3.6). Members of the RTE clade have not been detected in human or mice, but in cattle, supposedly introduced by horizontal transfer. They are also much less identical than those in plants (Adelson *et al.*, 2009). Moreover, all plant RTE LINEs discovered are very similar to each other, even across species borders or between monocots and dicots. This indicates that they belong to a single lineage of RTE LINEs (Chapter 4.3.3).

Homogenous RTE and highly diverse L1 LINEs seem to be a general feature of the genomes of higher plants. Nevertheless, future genome sequencing projects might still offer surprises. For example, it was believed for a long time that mammalian DNA transposons were in the process of extinction, occurred in very low numbers and showed little to no activity. Recent studies of TEs in the bat genus *Myotis*, however, provided evidence for extensive amplification of helitron and hAT DNA transposons contributing to more than 3 % and 3.5 % of the genome (Pritham and Feschotte, 2007; Ray *et al.*, 2008). Likewise, it is possible that during the 140-180 million years of angiosperm evolution, differing scenarios have developed.

5 Conclusion

As has been demonstrated in this thesis, retrotransposons are major components of the *B. vulgaris* genome. In order to investigate their actual impact on *B. vulgaris* genomes, two representative retrotransposon families, of the LTR and of the non-LTR order, have been analyzed in detail. Cotzilla LTR retrotransposons belong to the Sirevirus lineage, with typical hallmarks such as the presence of a putative *env* ORF, an extended *gag* region, and a frameshift separating the *gag* and *pol* genes. In contrast, LINEs of the BNR family differ from canonical plant LINE references in sequence and structure. Instead of a zinc finger motif in ORF1, they harbor an RNA recognition motif, likely to have an RNA-binding function. LINEs similar to BNR occur in a wide range of higher plants. Based on their RT domains, they were assigned to LINEs of the L1 clade, but form a distinct group, referred to as BNR subclade. Whereas the LTR retrotransposon family Cotzilla belongs to the most abundant and homogenous retrotransposon families in the beet genome, BNR LINEs occur in much less copy numbers and are far more diverse. These discrepancies in amplification can be in part explained by different integration preferences: While Cotzilla retrotransposons seem to target heterochromatic regions, BNR LINEs are mostly excluded from those.

The availability of the *B. vulgaris* genome sequence drafts enabled TE detection on a whole genome level. By bioinformatic approaches, it was possible to extract the reverse transcriptase sequences of Ty3-*gypsy*, Ty1-*copia* and LINE-like retrotransposons. Thus, a detailed overview of the retrotransposable fraction in the *B. vulgaris* genome has been generated by classification into clades and subclades, providing the base for in-depth TE characterization. As an example, *B. vulgaris* LINEs have been selected for a detailed analysis. Seventeen L1 LINE families (including BNR) have been identified, which are characterized by high diversity in sequence and structure. Though highly underrepresented in the genome sequence, an additional RTE family has been detected by a combination of bioinformatic and molecular methods.

In contrast to mammalian LINEs, which have been excessively investigated, plant LINE data are only poorly analyzed. For a comparison with other plant LINE populations, thirteen plant genomes including for example *A. thaliana*, *S. tuberosum* and *Z. mays* have been queried for LINE RTs. All of them show accumulation of many divergent L1 families, and some also harbor nearly uniform RTE sequences. A comparative analysis, including 6081 plant L1 reverse transcriptase sequences, provided evidence for a major separation of plant L1 LINEs into at least seven subclades, the BNR subclade being one of them. This enormous L1 diversification distinguishes plant L1s from their highly homogeneous human counterparts and illustrates the differences of mammalian and plant LINE evolution.

In conclusion, the data generated in this work contributes to the unraveling of retrotransposon evolutionary strategies in higher plant genomes. Furthermore, it provides a base for future retrotransposon detection in beet, as well as a guideline for LINE classification in plants.

6 Application and outlook

Cotzilla-like Sireviruses have been shown to make up one of the largest transposon families in the *B. vulgaris* genome. Targeted integration into heterochromatic regions is probably the key to the colonization of large fractions of the genome as has been discussed in Chapter 4.1.2. The epigenetic regulation of Cotzilla transcription and transposition, as well as the genome reponse to TE proliferation are still unknown. Genome sequences of related wild beets of the sections *Corollinae* and *Nanae*, and the genus *Patellifolia* are currently generated and assembled. In those genomes, only few or no Cotzilla members have been detected by Southern hybridization (Figure 3.7). It will be a task of cross-species comparisons to investigate, whether a related Sirevirus is equally abundant in those genomes, or if another retrotransposon clade has been more successful.

Moreover, 17 highly divergent *B. vulgaris* L1 LINE families have been detected and analyzed in detail. Only a single LINE family, belonging to the RTE clade, was characterized by a higher identity, however a complete RTE LINE of this family has still to be identified and characterized.

Additional to the characterization of beet TEs, plant LINEs have been comparatively analyzed in thirteen species. LINEs of the RTE clade are highly similar and at least middle repetitive in these plant genomes, where they have been identified, indicating the possibility of ongoing proliferation by retrotransposition. Interestingly, these retrotransposons seem to be underrepresented in many published plant genome sequences. A PCR-based approach might show the presence or absence of RTEs in these genomes. Instead, L1 LINEs have been found to be highly diverse, and have been therefore classified into at least seven subclades based on their RT. A detailed investigation of the BNR subclade has been performed, revealing typical structural characteristics for members of this group. Six L1 subclades (Figure 3.37) still remain to be characterized by analysis of full-length members in a wide range of plant genomes. It is still to verify, if more subclade-specific hallmarks evolved in genomes of higher plants.

Furthermore, the sequence information collected during this work is valuable for annotation of the repetitive part of the *B. vulgaris* genome. For easy detection of characterized repeats in unknown sequences, a number of *B. vulgaris* retrotransposon sequences has been added to the *EBI* database. Furthermore, the datasets containing nearly 6000 beet reverse transcriptases (Table 3.3; supplemental DVD-Rom) can not only be used to generate an overview of the *B. vulgaris* retrotransposable genome content, but also as a *BLAST* query to isolate and characterize a wide range of additional retrotransposons.

In summary, this the results of this thesis help to gain an understanding of the different strategies of retrotransposon evolution in plants, whereas the generated data directly contributes to the *B. vulgaris* genome annotation project.

7 Bibliography

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD*, et al.* (2000). The genome sequence of *Drosophila melanogaster*. Science 287:2185-2195

Adelson DL, Raison JM, Edgar RC (2009). Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. Proc Natl Acad Sci USA 106:12855-12860

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J Mol Biol 215:403-410

Angiosperm Phylogeny Group (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc 161:105-121

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796-815

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G*, et al.* (2011). The genome of *Theobroma cacao*. Nat Genet 43:101-108

Arumuganathan K, Earle E (1991). Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208-218

Babushok DV, Kazazian HH, Jr. (2007). Progress in understanding the biology of the human mutagen LINE-1. Hum Mutat 28:527-39

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A*, et al.* (2009). Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet 5:e1000732

Belancio VP, Hedges DJ, Deininger P (2008). Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. Genome Res 18:343-358

Bennett MD, Leitch IJ (2011). Nuclear DNA amounts in angiosperms: Targets, trends and tomorrow. Ann Bot 107:467-590

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W (2004). Consistent over-estimation of gene number in complex plant genomes. Curr Opin Plant Biol 7:732-736

Bennetzen JL, Kellogg EA (1997). Do plants have a one-way ticket to genomic obesity? Plant Cell 9:1509-1514

Bennetzen JL, Ma J, Devos KM (2005). Mechanisms of recent genome size variation in flowering plants. Ann Bot 95:127-32

Birney E, Clamp M, Durbin R (2004). GeneWise and Genomewise. Genome Res 14:988-995

Birney E, Kumar S, Krainer AR (1993). Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. Nucleic Acids Res 21:5803-5816

Boeke JD (1989). Transposable elements in *Saccharomyces cerevisiae*. In: Berg D, Howe M (eds) Mob DNA. American Society of Microbiology, Washington, DC

Boeke JD, Corces VG (1989). Transcription and Reverse Transcription of Retrotransposons. Annu Rev Microbiol 43:403

Boissinot S, Entezam A, Furano AV (2001). Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol 18:926-35

Bousios A, Darzentas N, Tsaftaris A, Pearce S (2010). Highly conserved motifs in non-coding regions of Sirevirus retrotransposons: the key for their pattern of distribution within and across plants? BMC Genomics 11:89

Brandes A, Heslop-Harrison JS, Kamm A, Kubis S, Doudrick RL*, et al.* (1997). Comparative analysis of the chromosomal and genomic organization of Ty1-*copia*-like retrotransposons in pteridophytes, gymnosperms and angiosperms. Plant Mol Biol 33:11-21

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH*, et al.* (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci USA 100:5280-5

Cantu D, Vanzetti L, Sumner A, Dubcovsky M, Matvienko M*, et al.* (2010). Small RNAs, DNA methylation and transposable elements in wheat. BMC Genomics 11:408

Capy P (2005). Classification and nomenclature of retrotransposable elements. Cytogenet Genome Res 110:457-61

Casacuberta J, Vernhettes S, Audeon C, Grandbastien M-A (1997). Quasispecies in retrotransposons: a role for sequence variability in Tnt1 evolution. Genetica 100:109-117

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ*, et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422-1423

Cole C, Barber JD, Barton GJ (2008). The *Jpred* 3 secondary structure prediction server. Nucleic Acids Res 36:W197-201

Comai L (2000). Genetic and epigenetic interactions in allopolyploid plants. Plant Mol Biol 43:387-99

Comfort NC (2001). From controlling elements to transposons: Barbara McClintock and the Nobel Prize. Trends Biochem Sci 26:454-457

Craig NL (1995). Unity in transposition reactions. Science 270:253-4

Crick F (1970). Molecular biology in the year 2000. Nature 228:613-615

Danio rerio Sequencing Project (http://www.sanger.ac.uk/Projects/D_rerio/). Wellcome Trust Sanger Institute

Darzentas N (2010). *Circoletto*: visualizing sequence similarity with *Circos*. Bioinformatics 26:2620-1

De Felice B, Wilson R, Argenziano C, Kafantaris I, Conicella C (2008). A transcriptionally active *copia*-like retroelement in *Citrus limon*. Cell Mol Biol Lett 14:289-304

De Jong HJ, Fransz P, Zabel P (1999). High resolution FISH in plants – techniques and applications. Trends Plant Sci 4:258-263

Dechyeva D, Gindullis F, Schmidt T (2003). Divergence of satellite DNA and interspersion of dispersed repeats in the genome of the wild beet *Beta procumbens*. Chromosome Res 11:3-21

Dechyeva D, Schmidt T (2006). Molecular organization of terminal repetitive DNA in *Beta* species. Chromosome Res 14:881-97

Dechyeva D, Schmidt T (2009). Molecular cytogenetic mapping of chromosomal fragments and immunostaining of kinetochore proteins in *Beta*. Int J Plant Genomics doi:10.1155/2009/ 721091

Desel C, Jansen R, Dedong GUE, Schmidt T (2002). Painting of parental chromatin in *Beta* hybrids by multi-colour fluorescent *in situ* hybridization. Ann Bot 89:171-181

Devos KM, Brown JKM, Bennetzen JL (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res 12:1075-1079

Dhellin O, Maestre J, Heidmann T (1997). Functional differences between the human LINE retrotransposon and retroviral reverse transcriptases for in vivo mRNA reverse transcription. EMBO J 16:6590-602

Döbel T (2008). Identifikation von *Short interspersed nuclear elements* (SINE)-Familien aus *Solanum tuberosum* und Anwendung für die Untersuchung der genetischen Diversität von Kartoffelgenotypen. Dresden University of Technology, Germany, Diploma Thesis

Draycott AP (2006). Sugar beet. Blackwell Publishing Ltd, Oxford

Drummond A, Ashton B, Buxton S, Cheung M, Cooper A*, et al.* (2011). Geneious v5.4, available from http://www.geneious.com/

Du J, Tian Z, Hans CS, Laten HM, Cannon SB*, et al.* (2010). Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J 63:584-98

Eddy SR (1998). Profile hidden Markov models. Bioinformatics 14:755-763

Edgar R (2004). *MUSCLE*: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113

Edlefsen P, Liu J (2010). Transposon identification using profile HMMs. BMC Genomics 11:S10

Eickbush TH, Furano AV (2002). Fruit flies and humans respond differently to retrotransposons. Curr Opin Genet Dev 12:669-74

Fedorov A (2009). Regulation of mammalian LINE1 retrotransposon transcription. Cell Tissue Biol 3:1-13

Feinberg AP, Vogelstein B (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 132:6-13

Feinberg J (2009). Wordle. www.wordle.net

Feng S, Jacobsen SE, Reik W (2010). Epigenetic reprogramming in plant and animal development. Science 330:622-627

Finnegan DJ (1989). Eukaryotic transposable elements and genome evolution. Trends Genet 5:103-107

Flavell RB, Bennett MD, Smith JB, Smith DB (1974). Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet 12:257-69

Flowers JM, Purugganan MD (2009). The evolution of plant genomes – scaling up from a population perspective. Curr Opin Genet Dev 18:565-570

Francki MG (2001). Identification of Bilby, a diverged centromeric Ty1-*copia* retrotransposon family from cereal rye (Secale cereale L.). Genome 44:266-274

Furano AV, Duvernell DD, Boissinot S (2004). L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. Trends Genet 20:9-14

Gallo SA, Finnegan CM, Viard M, Raviv Y, Dimitrov A*, et al.* (2003). The HIV *env*-mediated fusion reaction. Biochimica et Biophysica Acta (BBA) - Biomembranes 1614:36-50

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

Gindullis F, Dechyeva D, Schmidt T (2001a). Construction and characterization of a BAC library for the molecular dissection of a single wild beet centromere and sugar beet (*Beta vulgaris*) genome analysis. Genome 44:846-55

Gindullis F, Desel C, Galasso I, Schmidt T (2001b). The large-scale organization of the centromeric region in *Beta* species. Genome Res 11:253-265

Goff SA, Ricke D, Lan T-H, Presting G, Wang R*, et al.* (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92-100

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

Grandbastien M-A, Audeon C, Casacuberta JM, Grappin P, Lucas H*, et al.* (1994). Functional analysis of the tobacco Tnt1 retrotransposon. Genetica 93:181-189

Greilhuber J, Borsch T, Müller K, Worberg A, Porembski S*, et al.* (2006). Smallest Angiosperm genomes found in *Lentibulariaceae*, with chromosomes of bacterial size. Plant Biol 8:748-757

Han K, Lee J, Meyer TJ, Remedios P, Goodwin L*, et al.* (2008). L1 recombination-associated deletions generate human genomic variation. Proc Natl Acad Sci USA 105:19365-19370

Haren L, Ton-Hoang B, Chandler M (1999). Integrating DNA: transposases and retroviral integrases. Annu Rev Microbiol 53:245-81

Havecker ER, Gao X, Voytas DF (2005). The Sireviruses, a plant-specific lineage of the Ty1/*copia* retrotransposons, interact with a family of proteins related to dynein light chain 8. Plant Physiol 139:857-868

Hawkins JS, Hu G, Rapp RA, Grafenberg JL, Wendel JF (2008). Phylogenetic determination of the pace of transposable element proliferation in plants: *copia* and LINE-like elements in *Gossypium*. Genome 51:11-8

Hawkins JS, Proulx SR, Rapp RA, Wendel JF (2009). Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proc Natl Acad Sci USA doi: 10.1073/pnas.0904339106

Heitkam T, Schmidt T (2009). BNR - a LINE family from *Beta vulgaris* contains an RRM domain in open reading frame 1 and defines a L1 subclade present in diverse plant genomes. Plant J 59:872-882

Heitkam T, Holtgräwe D, Dohm JC, Minoche AE, Himmelbauer H, Weisshaar B, Schmidt T. 2014. Profiling of extensively diversified plant LINEs reveals distinct plant-specific subclades. Plant J 79:385-397

Hemleben V, Kovarik A, Torres-Ruiz RA, Volkov RA, Beridze T (2007). Plant highly repeated satellite DNA: molecular evolution, distribution and use for identification of hybrids. Syst Biodivers 5:277-289

Heslop-Harrison JS, Brandes A, Taketa S, Schmidt T, Vershinin A*, et al.* (1997). The chromosomal distributions of Ty1-*copia* group retrotransposable elements in higher plants and their implications for genome evolution. Genetica 100:197-204

Heslop-Harrison JS, Schwarzacher T (2011). Organization of the plant genome in chromosomes. Plant J 66:18-33

Heslop-Harrison JS, Schwarzacher T, Anamthawat-Jónsson K, Leitch AR, Shi M*, et al.* (1991). *In-situ* hybridization with automated chromosome denaturation. Technique 3:109-116

Higashiyama T, Noutoshi Y, Fujie M, Yamada T (1997). Zepp, a LINE-like retrotransposon accumulated in the *Chlorella* telomeric region. EMBO J 16:3715-23

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999). Plant *cis*-acting regulatory DNA elements (PLACE) database. Nucleic Acids Res 27:297-300

Hirochika H, Otsuki H, Yoshikawa M, Otsuki Y, Sugimoto K*, et al.* (1996). Autonomous Transposition of the Tobacco Retrotransposon Tto1 in Rice. Plant Cell 8:725-734

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

Holmes I (2002). Transcendent elements: Whole-genome transposon screens and open evolutionary questions. Genome Res 12:1152-1155

Hu G, Hawkins JS, Grover CE, Wendel JF (2010). The history and disposition of transposable elements in polyploid *Gossypium*. Genome 53:599-607

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

Jacobs G, Dechyeva D, Menzel G, Dombrowski C, Schmidt T (2004). Molecular characterization of Vulmar1, a complete mariner transposon of sugar beet and diversity of *mariner*- and En/Spm-like sequences in the genus *Beta*. Genome 47:1192-201

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C*, et al.* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463-467

Jia J, Yang Z, Li G, Liu C, Lei M*, et al.* (2009). Isolation and chromosomal distribution of a novel Ty1-*copia*-like sequence from *Secale*, which enables identification of wheat–*Secale africanum* introgression lines. J Appl Genet 50:25-28

Jung C, Wricke G (1987). Selection of diploid nematode-resistant sugar beet from monosomic addition lines. Plant Breed 98:205-214

Jurka J, Kapitonov VV, Kohany O, Jurka MV (2007). Repetitive sequences in complex genomes: Structure and evolution. Annu Rev Genom Hum Genet 8:241-259

Kadereit G, Hohmann S, Kadereit JW (2006). A synopsis of *Chenopodiaceae* subfam. *Betoideae* and notes on the taxonomy of *Beta*. Willdenowia - Annals of the Botanic Garden and Botanical Museum Berlin-Dahlem 36:9-19

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

Kapitonov VV, Tempel S, Jurka J (2009). Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene 448:207-213

Kazazian H (2011). Mobile DNA: Finding Treasure in Junk. FT Press, New Jersey

Kejnovsky E, Leitch IJ, Leitch AR (2009). Contrasting evolutionary dynamics between angiosperm and mammalian genomes. Trends Ecol Evol 24:572-582

Kennedy D, Norman C (2005). What don't we know? - So much more to know. Science 309:78-102

Khan E, Mack JP, Katz RA, Kulkosky J, Skalka AM (1991). Retroviral integrase domains: DNA binding and the recognition of LTR sequences. Nucleic Acids Res 19:851-60

Khan H, Smit A, Boissinot S (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res 16:78-87

Khazina E, Weichenrieder O (2009). Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. Proc Natl Acad Sci USA 106:725-730

Kidwell MG (2002). Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49-63

Kim T-M, Hong S-J, Rhyu M-G (2004). Periodic explosive expansion of human retroelements associated with the evolution of the hominoid primate. J Korean Med Sci 19:177-185

Kimura Y, Tosa Y, Shimada S, Sogo R, Kusaba M*, et al.* (2001). OARE-1, a Ty1-*copia* retrotransposon in oat activated by abiotic and biotic stresses. Plant Cell Physiol 42:1345-1354

Kloc A, Martienssen R (2008). RNAi, heterochromatin and the cell cycle. Trends Genet 24:511-517

Komatsu M, Shimamoto K, Kyozuka J (2003). Two-step regulation and continuous retrotransposition of the rice LINE-type retrotransposon Karma. Plant Cell 15:1934-44

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

Kubis SE, Heslop-Harrison JS, Desel C, Schmidt T (1998). The genomic organization of non-LTR retrotransposons (LINEs) from three *Beta* species and five other angiosperms. Plant Mol Biol 36:821-831

Kumar A (1998). The evolution of plant retroviruses: moving to green pastures. Trends Plant Sci 3:371-374

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

Kuykendall D, Shao J, Trimmer K (2009). A nest of LTR retrotransposons adjacent the disease resistance-priming gene NPR1 in *Beta vulgaris* L. U.S. Hybrid H20. Int J Plant Genom doi:10.1155/2009/576742

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature 409:860-921

Lange C, Holtgräwe D, Schulz B, Weisshaar B, Himmelbauer H (2008). Construction and characterization of a sugar beet (*Beta vulgaris*) fosmid library. Genome 51:948-951

Lange W, Brandenburg WA, De Bock TSM (1999). Taxonomy and cultonomy of beet (*Beta vulgaris* L.). Bot J Linn Soc 130:81-96

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

Laten HM, Morris RO (1993). SIRE-1, a long interspersed repetitive DNA element from soybean with weak sequence similarity to retrotransposons: initial characterization and partial sequence. Gene 134:153-159

Le Q, Melayah D, Bonnivard E, Petit M, Grandbastien M-A (2007). Distribution dynamics of the Tnt1 retrotransposon in tobacco. Mol Genet Genomics 278:639-651

Lee J, Cordaux R, Han K, Wang J, Hedges DJ*, et al.* (2007). Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. Gene 390:18-27

Leeton PR, Smyth DR (1993). An abundant LINE-like element amplified in the genome of *Lilium speciosum*. Mol Gen Genet 237:97-104

Legendre M, Pochet N, Pak T, Verstrepen KJ (2007). Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res 17:1787-1796

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

Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y*, et al.* (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. Nucleic Acids Res 30:325-327

Lindtner S, Felber BK, Kjems J (2002). An element in the 3' untranslated region of human LINE-1 retrotransposon mRNA binds NXF1(TAP) and can function as a nuclear export element. RNA 8:345-356

Lisch D (2009). Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol 60:43-66

Liu B, Wendel JF (2000). Retrotransposon activation followed by rapid repression in introgressed rice plants. Genome 43:874-80

Llorens C, Futami R, Covelli L, Dominguez-Escribá L, Viu JM*, et al.* (2010). The *Gypsy* Database (GyDB) of mobile genetic elements: release 2.0 Nucleic Acids Res doi: 10.1093/nar/gkq1061

Llorens C, Munoz-Pomer A, Bernad L, Botella H, Moya A (2009). Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct 4:41

Long L, Ou X, Liu J, Lin X, Sheng L*, et al.* (2009). The spaceflight environment can induce transpositional activation of multiple endogenous transposable elements in a genotype-dependent manner in rice. J Plant Physiol 166:2035-2045

Lunde BM, Moore C, Varani G (2007). RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol 8:479-490

Lupas A (1996). Prediction and analysis of coiled-coil structures. Methods in enzymology 266:513-525

Ma J, Devos KM, Bennetzen JL (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14:860-869

Macas J, Neumann P (2007). Ogre elements – a distinct group of plant Ty3/*gypsy*-like retrotransposons. Gene 390:108-16

Malik HS, Burke WD, Eickbush TH (1999). The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol 16:793-805

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

Malik HS, Henikoff S, Eickbush TH (2000). Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10:1307-1318

Mardis ER (2008). Next-generation DNA sequencing methods. Annu Rev Genom Hum Genet 9:387

Maris C, Dominguez C, Allain FH (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J 272:2118-31

Martin SL (2006). The ORF1 protein encoded by LINE-1: Structure and function during L1 retrotransposition. J Biomed Biotechnol 2006:45621

Martin SL, Bushman FD (2001). Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Mol Cell Biol 21:467-475

McClintock B (1984). The significance of responses of the genome to challenge. Science 226:792-801

McGrath JM, Shaw RS, de los Reyes BG, Weiland JJ (2004). Construction of a sugar beet BAC library from a hybrid with diverse traits. Plant Mol Biol Rep 22:23-28

Menzel G, Dechyeva D, Keller H, Lange C, Himmelbauer H*, et al.* (2006). Mobilization and evolutionary history of miniature inverted-repeat transposable elements (MITEs) in *Beta vulgaris* L. Chromosome Res 14:831-44

Menzel G, Dechyeva D, Wenke T, Holtgrawe D, Weisshaar B*, et al.* (2008). Diversity of a complex centromeric satellite and molecular characterization of dispersed sequence families in sugar beet (*Beta vulgaris*). Ann Bot 102:521-530

Miguel C, Simões M, Oliveira M, Rocheta M (2008). *Envelope*-like retrotransposons in the plant kingdom: Evidence of their presence in gymnosperms (*Pinus pinaster*). J Mol Evol 67:517-525

Moisy C, Garrison K, Meredith C, Pelsy F (2008). Characterization of ten novel Ty1/*copia*-like retrotransposon families of the grapevine genome. BMC Genomics 9:469

Moore G (1995). Cereal genome evolution: pastoral pursuits with 'Lego' genomes. Curr Opin Genet Dev 5:717-724

Mroczek RJ, Dawe RK (2003). Distribution of retroelements in centromeres and neocentromeres of maize. Genetics 165:809-819

Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M*, et al.* (2004). Sequencing of a rice centromere uncovers active genes. Nat Genet 36:138-145

Neumann P, Navratilova A, Koblizkova A, Kejnovsky E, Hribova E*, et al.* (2011). Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob DNA 2:4

Noma K, Ohtsubo H, Ohtsubo E (2000). ATLN elements, LINEs from *Arabidopsis thaliana*: identification and characterization. DNA Res 7:291-303

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

Oliver KR, Greene WK (2009). Transposable elements: powerful facilitators of evolution. Bioessays 31:703-714

Orgel LE, Crick FHC (1980). Selfish DNA: the ultimate parasite. Nature 284:604-607

Ostertag EM, Kazazian HH (2001a). Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res 11:2059-2065

Ostertag EM, Kazazian HH, Jr. (2001b). Biology of mammalian L1 retrotransposons. Annu Rev Genet 35:501-38

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M*, et al.* (2007). The TIGR rice genome annotation resource: improvements and new features. Nucleic Acids Res 35:D883-7

Pagni M, Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV*, et al.* (2004). MyHits: a new interactive resource for protein annotation and domain identification. Nucleic Acids Res 32:W332

Pearce S (2007). SIRE-1, A putative plant retrovirus is closely related to a legume Ty1-*copia* retrotransposon family. Cell Mol Biol Lett 12:120-126

Pearson WR (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Method Enzymol 183:63-98

Pellicer J, Fay MF, Leitch IJ (2010). The largest eukaryotic genome of them all? Bot J Linn Soc 164:10-15

Pereira V (2004). Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. Genome Biol 5:R79

Permanyer J, Gonzalez-Duarte R, Albalat R (2003). The non-LTR retrotransposons in *Ciona intestinalis*: new insights into the evolution of chordate genomes. Genome Biol 4:R73

Peterson-Burch BD, Voytas DF (2002). Genes of the *Pseudoviridae* (Ty1/*copia* retrotransposons). Mol Biol Evol 19:1832-1845

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

Pritham EJ, Feschotte C (2007). Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. Proc Natl Acad Sci USA 104:1895-900

Raizada MN, Nan GL, Walbot V (2001). Somatic and germinal mobility of the RescueMu transposon in transgenic maize. Plant Cell 13:1587-608

Ramallo E, Kalendar R, Schulman A, Martínez-Izquierdo J (2008). Reme1, a *copia* retrotransposon in melon, is transcriptionally induced by UV light. Plant Mol Biol 66:137-150

Rangwala SH, Zhang L, Kazazian HH, Jr. (2009). Many LINE1 elements contribute to the transcriptome of human somatic cells. Genome Biol 10:R100

Ray DA, Feschotte C, Pagan HJ, Smith JD, Pritham EJ*, et al.* (2008). Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. Genome Res 18:717-28

Richard GF, Kerrest A, Dujon B (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72:686-727

Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984). Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci USA 81:8014-8018

Saitou N, Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425

Sakamoto K, Ohmido N, Fukui K, Kamada H, Satoh S (2000). Site-specific accumulation of a LINE-like retrotransposon in a sex chromosome of the dioecious plant *Cannabis sativa*. Plant Mol Biol 44:723-32

Sambrook J, Fritsch EF, Maniatis T (1989). Molecular cloning: A laboratory manual. Cold Spring Harbor Laboratory Press

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463-5467

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998). The paleontology of intergene retrotransposons of maize. Nat Genet 20:43-45

SanMiguel P, Tikhonov A, Jin Y-K, Motchoulskaia N, Zakharov D*, et al.* (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765-768

Sanz-Alferez S, SanMiguel P, Jin YK, Springer PS, Bennetzen JL (2003). Structure and evolution of the Cinful retrotransposon family of maize. Genome 46:745-52

Schmidt M (2010). Untersuchung der Transkription und Transposition der Retrotransposon-Familie Cotzilla im Genom der Zuckerrübe (*Beta vulgaris*) nach 5-Azacytidin-Behandlung. Dresden University of Technology, Germany, Bachelor Thesis

Schmidt T, Heslop-Harrison JS (1998). Genomes, genes and junk: the large-scale organization of plant chromosomes. Trends Plant Sci 3:195-199

Schmidt T, Jung C, Metzlaff M (1991). Distribution and evolution of two satellite DNAs in the genus *Beta*. Theoretical and Applied Genetics 82:793-799

Schmidt T, Kubis SE, Heslop-Harrison JS (1995). Analysis and chromosomal localization of retrotransposons in sugar beet (Beta vulgaris L.): LINEs and Ty1-*copia*-like elements as major components of the genome. Chromosome Res 3:335-45

Schmidt T, Metzlaff M (1991). Cloning and characterization of a *Beta vulgaris* satellite DNA family. Gene 101:247-50

Schmidt T, Schwarzacher T, Heslop-Harrison JS (1994). Physical mapping of rRNA genes by fluorescent in-situ hybridization and structural analysis of 5S rRNA genes and intergenic spacer sequences in sugar beet (*Beta vulgaris*). Theor Appl Genet 88:629-636

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T*, et al.* (2010). Genome sequence of the palaeopolyploid soybean. Nature 463:178-183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F*, et al.* (2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112-1115

Schulte D, Cai D, Kleine M, Fan L, Wang S*, et al.* (2006). A complete physical map of a wild beet (*Beta procumbens*) translocation in sugar beet. Mol Genet Genomics 275:504-11

Schwarz-Sommer Z, Leclercq L, Gobel E, Saedler H (1987). Cin4, an insert altering the structure of the A1 gene in *Zea mays*, exhibits properties of nonviral retrotransposons. EMBO J 6:3873-3880

Schwarzacher T, Heslop-Harrison JS (2000). Practical *in situ* hybridization. BIOS Scientific Publishers, Oxford

Scott AJ, Ford Lloyd BV, Williams JT (1977). *Patellifolia*, nomen novum (Chenopodiaceae). Taxon 26:284

Simon SA, Meyers BC (2010). Small RNA-mediated epigenetic modifications in plants. Curr Opin Plant Biol 14:148-155

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

Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD*, et al.* (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell 136:461-472

Smit AFA, Hubley R, Green P (2008). RepeatMasker Open-3.0.

Söding J, Biegert A, Lupas AN (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244-W248

Soifer HS (2006). Do small RNAs interfere with LINE-1? J Biomed Biotechnol 2006:29049

Soifer HS, Zaragoza A, Peyvan M, Behlke MA, Rossi JJ (2005). A potential role for RNA interference in controlling the activity of the human LINE-1 retrotransposon. Nucleic Acids Res 33:846-856

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

Sterck L, Rombauts S, Vandepoele K, Rouzé P, Van de Peer Y (2007). How many genes are there in plants (... and why are they there)? Curr Opin Plant Biol 10:199-203

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M*, et al.* (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 36:D1009-D1014

Szak S, Pickeral O, Makalowski W, Boguski M, Landsman D*, et al.* (2002). Molecular archeology of L1 insertions in the human genome. Genome Biol 3:research0052.1-research0052.18

Tamura K, Dudley J, Nei M, Kumar S (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596-1599

Thomas CA (1971). The genetic organization of chromosomes. Annu Rev Genet 5:237-256

Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL*, et al.* (2009). Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res 19:2221-30

Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A*, et al.* (2009). Bursts of retrotransposition reproduced in Arabidopsis. Nature 461:423-426

Turcotte K, Srinivasan S, Bureau T (2001). Survey of transposable elements from rice genomic sequences. Plant J 25:169-179

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I*, et al.* (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596-1604

Ulbrich E (1934). Chenopodiaceae. In: Engler A, Prantl K (eds) Die natürlichen Pflanzenfamilien, ed. 2, Leipzig

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A*, et al.* (2010). The genome of the domesticated apple (*Malus x domestica* Borkh.). Nat Genet 42:833-839

Vershinin AV, Druka A, Alkhimova AG, Kleinhofs A, Heslop-Harrison JS (2002). LINEs and *gypsy*-like retrotransposons in *Hordeum* species. Plant Mol Biol 49:1-14

Vershinin AV, Ellis THN (1999). Heterogeneity of the internal structure of PDR1 , a family of Ty1/ *copia* -like retrotransposons in pea. Mol Gen Genet 262:703-713

Vitte C, Bennetzen JL (2006). Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103:17638-17643

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D*, et al.* (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763-8

Vogt J (2010). Identifizierung und Charakterisierung von Long Interspersed Nuclear Elements (LINEs) aus dem Genom der Kartoffel und Verbreitungsanalyse innerhalb der *Solanaceae*. Dresden University of Technology, Germany, Diploma Thesis

Volff JN, Bouneau L, Ozouf-Costaz C, Fischer C (2003). Diversity of retrotransposable elements in compact pufferfish genomes. Trends Genet 19:674-8

Wang H, Liu J-S (2008). LTR retrotransposon landscape in *Medicago truncatula*: more rapid removal than in rice. BMC Genomics 9:382

Weber B, Schmidt T (2009). Nested Ty3-*gypsy* retrotransposons of a single *Beta procumbens* centromere contain a putative chromodomain. Chromosome Res 17:379–396

Weber B, Wenke T, Frömmel U, Schmidt T, Heitkam T (2010). The Ty1-*copia* families SALIRE and Cotzilla populating the *Beta vulgaris* genome show remarkable differences in abundance, chromosomal distribution, and age. Chromosome Res 18:247-263

Weber B, Heitkam T, Holtgräwe D, Weisshaar B, Minoche AE, Dohm JC, Himmelbauer H, Schmidt T. 2013. Highly diverse chromoviruses of *Beta vulgaris* are classified by chromodomains and chromosomal integration. Mobile DNA 4:8

Wenke T, Holtgräwe D, Horn A, Weisshaar B, Schmidt T (2009). An abundant and heavily truncated non-LTR retrotransposon (LINE) family in *Beta vulgaris*. Plant Mol Biol 71:585-597

Wessler SR (1996). Turned on by stress. Plant retrotransposons. Curr Biol 6:959-61

Wicker T, Keller B (2007). Genome-wide comparative analysis of *copia* retrotransposons in *Triticeae*, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Res 17:1072-1081

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, *et al.* (2007). A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973-982

Wicker T, Schlagenhauf E, Graner A, Close T, Keller B, *et al.* (2006). 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7:275

Wollrab C, Heitkam T, Holtgräwe D, Weisshaar B, Minoche AE, Dohm JC, Himmelbauer H, Schmidt T. 2012. Evolutionary reshuffling in the Errantivirus lineage Elbe within the *Beta vulgaris* genome. Plant J 72:636-651

Woo S-S, Jiang J, Gill BS, Paterson AH, Wing RA (1994). Construction and characterization of bacterial artificial chromosome library of *Sorghum bicolor*. Nucleic Acids Res 22:4922-4931

Wright DA, Ke N, Smalle J, Hauge BM, Goodman HM, *et al.* (1996). Multiple non-LTR retrotransposons in the genome of *Arabidopsis thaliana*. Genetics 142:569-78

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

Xia Q, Zhou Z, Lu C, Cheng D, Dai F, *et al.* (2004). A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science 306:1937-1940

Xiong Y, Eickbush TH (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353-62

Xu Z, Wang H (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265-268

Yamashita H, Tahara M (2006). A LINE-type retrotransposon active in meristem stem cells causes heritable transpositions in the sweet potato genome. Plant Mol Biol 61:79-94

Zakrzewski F, Weisshaar B, Fuchs J, Bannack E, Minoche AE*, et al.* (2011). Epigenetic profiling of heterochromatic satellite DNA. Chromosoma 120:409-422

Zakrzewski F, Wenke T, Holtgräwe D, Weisshaar B, Schmidt T (2010). Analysis of a *c_0t-1* library enables the targeted identification of minisatellite and satellite families in *Beta vulgaris*. BMC Plant Biol 10

Zdobnov EM, Apweiler R (2001). *InterProScan* - an integration platform for the signature-recognition methods in *InterPro*. Bioinformatics 17:847-848

Zeh DW, Zeh JA, Ishida Y (2009). Transposable elements and an epigenetic basis for punctuated equilibria. Bioessays 31:715-726

Zingler N, Willhoeft U, Brose H-P, Schoder V, Jahns T*, et al.* (2005). Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. Genome Res 15:780-789

Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H*, et al.* (2007). Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. BMC Evol Biol 7:152

Zupunski V, Gubensek F, Kordis D (2001). Evolutionary dynamics and evolutionary history in the RTE clade of non-LTR retrotransposons. Mol Biol Evol 18:1849-1863

8 Abbreviations

Standard abbreviations, chemicals

	(v/v)	Volume per volume
	(w/v)	Weight per volume
	2n	Diploid chromosome set
	5S rRNA	Ribosomal RNA of the 5 Svedberg subunit
	7SL RNA	Signal recognition particle RNA
	λ-DNA	Lambda DNA
A	A	Adenine
	Amp	Ampicillin
	AP	Aspartic protease
	APG	Angiosperm Phylogeny Group
	ATP	Adenosine triphosphate
B	BAC	Bacterial artificial chromosome
	Belline	Beet L1 LINE
	BLAST	Basic local alignment search tool
	BNR	Beet non-LTR retrotransposon
	BSA	Bovine serum albumine
	BvL	*Beta vulgaris* LINE
C	C	Cytosine
	C-value	Carbon-value
	cDNA	DNA copy of an RNA
	CM	Chloramphenicol
	CTAB	Cetyl trimethylammonium bromide
	Cy3	Cyanine3 fluorochrome
D	DAPI	4',6-Diamidino-2-phenylindole
	dATP	Deoxyadenosine triphosphate
	dCTP	Deoxycytidine triphosphate
	ddNTP	Dideoxynucleotide
	DEPC	Diethylpyrocarbonate
	dGTP	Deoxyguanosine triphosphate
	DMSO	Dimethyl sulfoxide
	DNA	Deoxyribonucleic acid
	Dnase	Deoxyribonulease
	dNTP	Deoxyribonucleotide
	dsRNA	Double-stranded RNA
D	DTT	Dithiothreitol
	dTTP	Deoxythimidine triphosphate
	dUTP	Deoxyuridine triphosphate
E	e.g.	*Exempli gratia* (for example)
	EBI	European Bioinformatics Institute
	EDTA	Ethylenediaminetetraacetic acid
	EN	Endonuclease
	env	*envelope*, ORF with similarities to retroviral *envelope* proteins
	EST	Expressed sequence tag
	e-value	Expectation value
F	FISH	Fluorescent *in situ* hybridization
	FITC	Fluorescein isothiocyanate
G	G	Guanine
	g	Gravitational acceleration (≈ 9.81 m/s^2)
	gag	Group-specific antigen
H	H_2O	Water
	HCl	Hydrochloric acid
	HMM	Hidden Markov Model
I	I	Inosine
	i.e.	*Id est* (that is, in other words)
	int	Integrase
	IPTG	Isopropyl ß-D-1-thiogalactopyranoside
	KCl	Potassium chloride
	KH_2PO_4	Potassium dihydrogen phosphate
	K_2HPO_4	Potassium hydrogen phosphate
L	L1	LINE clade with a reverse transcriptase similar to human LINE-1
	LB	Luria-Bertani
	LINE	Long interspersed nuclear element
	LTR	Long terminal repeat
M	M	Marker of DNA sizes
	$MgCl_2$	Magnesium chloride
	$MgSO_4$	Magnesium sulfate
	MITE	Miniature inverted repeat transposable element
	mRNA	Messenger RNA
	MUSCLE	Multiple sequence comparison by log-expectation
N	N	Any nucleotide
	Na	Sodium
	NaCl	Sodium chloride
	Na_2HPO_4	Sodium dihydrogenphosphate
	NCBI	National Center for Biotechnology Information
	$(NH_4)_2SO_4$	Ammonium sulfate
	No.	Number
	NOR	Nucleolus organizer region

O	ORF	Open reading frame
P	^{32}P	Radioactive isotope of phosphorus
	PBS	Primer binding site
	PCR	Polymerase chain reaction
	pH	Negative decimal logarithm of hydrogen ion activity in solution
	pol	Polyprotein reading frame in retrotransposons
	PPT	Polypurine tract
	PVP	Polyvinylpyrrolidone
R	R	Nucleotides A and G (purines)
	rDNA	Ribosomal DNA
	RNA	Ribonucleic acid
	RH	RNaseH; Ribonuclease H
	RRM	RNA Recognition Motif
	rRNA	Ribosomal RNA
	RT	Reverse transcriptase
	RTE	LINE clade with a reverse transcriptase similar to retrotransposon-like element (RTE-1) in *Caenorhabditis elegans*
S	S	Nucleotides G and C
	SDS	Sodium dodecyl sulfate
	SINE	Short interspersed nuclear element
	sRNA	Small RNA
	SSC	Saline sodium citrate buffer
	subsp	Subspecies
T	T	Thymine
	TAE	Tris-Acetate-EDTA
	tBLASTn	*BLAST* using a protein query on a nucleotide database
	TE	Transposable element
	TE buffer	Tris-EDTA buffer
	TIR	Terminal inverted repeat
	Tris	Tris (hydroxymethyl) aminomethane
	TPRT	Target-primed reverse transcription
	tRNA	Transfer RNA
	TSD	Target site duplication
	Ty	Transposon in yeast
U	U	Uracil
	UTR	Untranslated region
	UV	Ultraviolet light
W	W	Nucleotides A and T
X	X	Any amino acid
	X-Gal	5-Bromo-4-chloro-3-indolyl-ß-galactopyranoside
	XML	Extensible markup language
Y	Y	Nucleotides T and C (pyrimidines)

Quantity abbreviations

°C	Degree(s) celsius
aa	Amino acid(s)
bp	Base pair(s)
Ci	Curie(s)
g	Gram(s)
h	Hour(s)
kb	Kilobase(s)
l	Liter(s)
M	Molar
m	Meter(s)
mA	Milliampere(s)
Mb	Megabase(s)
min	Minute(s)
mM	Millimolar
nm	Nanometer(s)
nt	Nucleotide(s)
pg	Picogram(s)
rpm	Round(s) per minute
s	Second(s)
U	Unit(s)
V	Volt(s)
vol	Volume(s)

Species abbreviations

A. thaliana	*Arabidopsis thaliana*
B. distachyon	*Brachypodium distachyon*
B. mori	*Bombyx mori*
B. vulgaris	*Beta vulgaris*
C. sativa	*Cannabis sativa*
D. melanogaster	*Drosophila melanogaster*
D. rerio	*Danio rerio*
E. coli	*Escherichia coli*
G. max	*Glycine max*
H. sapiens	*Homo sapiens*
L. japonicus	*Lotus japonicus*
M. x domestica	*Malus x domestica*
M. guttatus	*Mimulus guttatus*
O. sativa	*Oryza sativa*
P. procumbens	*Patellifolia procumbens*
P. trichocarpa	*Populus trichocarpa*
S. lycopersicum	*Solanum lycopersicum*
S. tuberosum	*Solanum tuberosum*
T. cacao	*Theobroma cacao*
V. vinifera	*Vitis vinifera*
Z. mays	*Zea mays*

9 Appendix

Appendix 1: Complete sequence and annotation of the Sirevirus Cotzilla1.

Appendix 2: Complete sequence and annotation of the integration event of LINE BNR1 into satellite pAv.

Appendix 3: Sequence annotation of the 5’ truncated *B. vulgaris* RTE LINE *Ghost*1.

Appendix 4: Alignment used for sequence identity caculation of BNR-like reverse transcriptase sequences.

Appendix 5: Content of the supplemental CD-Rom.

Appendix 1: Complete sequence and annotation of the Sirevirus Cotzilla1.
The complete sequence of Cotzilla1 with a length of 10833 bp has been presented. Hallmarks of this retrotransposon have been colorcoded. The coding sequences, deduced with the software *Genewise*, have been shown in grey shading, while their start and stop codons are colored in violet. Frameshifts are marked by a slash ('/'). RNA-binding zinc finger motifs have been highlighted in yellow, coiled coil domains in olive, and conserved protease, integrase, reverse transcriptase (RT) and RNaseH domains have been shown in blue, green red and magenta, respectively (Peterson-Burch and Voytas, 2002; Khan *et al.*, 1991; Haren *et al.*, 1999; Xiong and Eickbush, 1990; Malik and Eickbush, 2001). The target site duplication (TSD) flanking the element has been colored in red, the LTRs in blue and the PBS and PPT in green.

```
atttttgttgaaccctagttttgataatgacaaagaaacaaaggaacaaactaagtagtgaactaatgaattatttaagtgtatggtgtctaaggtatca 100
 TSD 5' LTR →

atgcaaggaactgggagaatacgaagtcaatgaagaagtcaagagaggatttcaaaggctaagtaaagaagtttctatgttgatactggagcaagcattg 200
gatcaagaataatagaagctaggcgttttattttctagttccaatgtaaaggaacaaagaaagaatggtaaggaactgtggcagaatatttgagaatgaa 300
agctttttttcctctgtaaggaacaaactcagtggcacttagcaaatatgttaaaggaactgaactaatataaggaactagacttaggcgtgtacttagt 400
ataacgattaaaagataatggagtaagtcatgcctagattataggaaattagttatttctaatttcaaatgataaaatcagttttaaaaagaaaatctgt 500
tttggaaaagattttattttcggaaaataagaaaataaatatccggttgttagaaagattttatttcaaagatattattttcttaaagattagaaattc 600
atatttaaagataaaaagagtttataaaaactgattttcggataatgctgaaggaactcctattatttaggagtaatgactgcattttatccacgttcat 700
ttatagttcatgggatcatgggacatgatgtaaaacgtgttccttaggggaagtaacatcatgggtaaatggattaaaacgtgtactctaacattaatcc 800
cacattcttcatgattaacactataaatatagaagtctagataagttccaaagcgtgtctgactaagtaattaagtgagtttaacttattaagtgtcagt 900
gccttcattctctcataatagccatttgttctttgccttcttactctaattccactaagcttatttcgatcaattgtattttggtttttggggtaaggga 1000
actttgagtgaagttctgtaagagaaaggctttgagtgaagcttgtatgtgtgagaaaacttcgagtgaagttgagaggaactgctgttaagggaacagt 1100
ggttcaggaacttgagtttaggaactcaaggtagggctcgagttagaattaggttgtaacagagttgtttggcctaataagtgaaagtgttgagtttaaa 1200
atccctagtggtcgaggttgtttcttcttgttgggcccaagaagtttttcctcgtaaaaatcccttctgttcctttagcttgtttattcgtttaagtttt 1300
aatttccgcaaaaagttacgtttatttctacacctacaattcaccccccctcttgtagtgttcctagggaaataacaattggtatcagagccagaactca 1400
                                                                                     PBS

cgtaagataggaaaccctattgaggaaaaaatccaaggaacaatgaataccaccgagaaacttgaagaaggttattccactcaaaggcctcctatgttca 1500
                                           M  N  T  T  E  K  L  E  E  G  Y  S  T  Q  R  P  P  M  F

atgggaagtactacaactattggaagaaccgcatggagatattcatcaaggctgaaaactatcaggtatggagagtcattgaagttggcgatttcgaagt 1600
N  G  K  Y  Y  N  Y  W  K  N  R  M  E  I  F  I  K  A  E  N  Y  Q  V  W  R  V  I  E  V  G  D  F  E  V

cacaactacaaatgacaaaaatgaggtaactcttaaacccctctctgattatgacaaatctgatttgaaaaaatggaagtaaatgccatggctattaaa 1700
  T  T  T  N  D  K  N  E  V  T  L  K  P  L  S  D  Y  D  K  S  D  F  E  K  M  E  V  N  A  M  A  I  K

ttattgcattgtggacttggaccccatgaacataataggatcatgggatgcaaatccgcaaagcaaatatgggatttactagaagtaacccacgaaggaa 1800
 L  L  H  C  G  L  G  P  H  E  H  N  R  I  M  G  C  K  S  A  K  Q  I  W  D  L  L  E  V  T  H  E  G

ctaatgaagttaaaagatcgaaaattgatttacttatgaatcaatatgaactgttctgcatgaaatccaaagaaagcatccgagacatgtttactcgttt 1900
T  N  E  V  K  R  S  K  I  D  L  L  M  N  Q  Y  E  L  F  C  M  K  S  K  E  S  I  R  D  M  F  T  R  F

tactaacattataaatgagttggcttctcttggaaagtttatttcatctgaggaacaggttcgaaaggttcttaggagtcttcctaaggacagatggatg 2000
  T  N  I  I  N  E  L  A  S  L  G  K  F  I  S  S  E  E  Q  V  R  K  V  L  R  S  L  P  K  D  R  W  M

actaaagtcacggctcttcaagaaacaaaagacttcactaagttcaacttagaacagctggcagggtcacttatgactcatgagcttcaccttgatactg 2100
 T  K  V  T  A  L  Q  E  T  K  D  F  T  K  F  N  L  E  Q  L  A  G  S  L  M  T  H  E  L  H  L  D  T

aatatggtgaaagctccaaatcaaaatcaattgctctcaaagcagatgatgaagatgattcggactctgaagaagaagaagcagccctcatggttcgaaa 2200
E  Y  G  E  S  S  K  S  K  S  I  A  L  K  A  D  D  E  D  D  S  D  S  E  E  E  E  A  A  L  M  V  R  K

atttcgaaagatgtacaggaacatgaagaatggaaacttcaaaggtaaaactaagaaattttctaacaaagctgcttgtcataaatgtggaagtacagat 2300
  F  R  K  M  Y  R  N  M  K  N  G  N  F  K  G  K  T  K  K  F  S  N  K  A  A  C  H  K  C  G  S  T  D
                                                                                  zinc finger motif

cacttcattaaagaatgtcctctttgggaaaacgacaaaaccaaagaaaggaacaaggaacgctttgctgaaaggaacaaagaatcaaaagcacctttct 2400
 H  F  I  K  E  C  P  L  W  E  N  D  K  T  K  E  R  N  K  E  R  F  A  E  R  N  K  E  S  K  A  P  F      gag

ctaaggcaaatgtgcgcaaagctatgatagctgcttggggagattcagagatggaagaagaagaggaacaaccaactgaagaaacagcaaacctttgcct 2500
S  K  A  N  V  R  K  A  M  I  A  A  W  G  D  S  E  M  E  E  E  E  E  Q  P  T  E  E  T  A  N  L  C  L

catggcaaacactgacgaaaaagaagatgaagaaataattgaggtaagtcaacatggacttgagtcagaactagatggtaaaactaggactgaaatatat 2600
  M  A  N  T  D  E  K  E  D  E  E  I  I  E  V  S  Q  H  G  L  E  S  E  L  D  G  K  T  R  T  E  I  Y

gatcttctttatgatgcaattcttgaatgtagagatgaacgtgaaaaacgtgaacaaacggaacaagccttaaaggtttgcaaggaacacatagaatggc 2700
 D  L  L  Y  D  A  I  L  E  C  R  D  E  R  E  K  R  E  Q  T  E  Q  A  L  K  V  C  K  E  H  I  E  W

ttaagaaaatgagaactgatgtcgaaactagatttttttgatttgtttgataaaaacttaaaaatcaaagaatgttatgagagtttgaaaaatgaaaatta 2800
L  K  K  M  R  T  D  V  E  T  R  F  F  D  L  F  D  K  N  L  K  I  K  E  C  Y  E  S  L  K  N  E  N  Y

tctaattaatttagagatttcacacttgaaggaatttgtccctgttctgagttcctttgatataagatcaacccctttgaaaggatacagatctgagatt 2900
  L  I  N  L  E  I  S  H  L  K  E  F  V  P  V  L  S  S  F  D  I  R  S  T  P  L  K  G  Y  R  S  E  I
                                                                         coiled coil

gaaaagataaataaagagaaaattattttaaaggaacagatctgtagtcttaaaaaggaactggttgtcgcaaggaacagaggcaaaatcatgaaagttc 3000
 E  K  I  N  K  E  K  I  I  L  K  E  Q  I  C  S  L  K  K  E  L  V  V  A  R  N  R  G  K  I  M  K  V

ctaagtggattgagaaaccacagaccaaaaggaaggaaggtcttggttttgaaaactacaagaaaaggaacaagacaaagaaatatgttgatctacccag 3100
P  K  W  I  E  K  P  Q  T  K  R  K  E  G  L  G  F  E  N  Y  K  K  R  N  K  T  K  K  Y  V  D  L  P  S
```

```
tgataagttctgtatgtattgtggcaaggtaggtcattacataaatcaatgtcaactaaaatctgagatgattggggaaaatgtgacaataagtaagaaa 3200
 D  K  F  C  M  Y  C  G  K  V  G  H  Y  I  N  Q  C  Q  L  K  S  E  M  I  G  E  N  V  T  I  S  K  K
           zinc finger motif

aaatgggttgtcaaaactaatcctagtcaagaacaggaacccacgaaagatggggttccttcaactaacaactagttgatttttgcaggttggagtgagg 3300
K  W  V  V  K  T  N  P  S  Q  E  Q  E  P  T  K  D  G  V  P  S  T  N  N/          F  L  Q  V  G  V  R

gggaacaacacttggtatctcgatagtggttgttccaagcacatgacaggagacaaaacgaagtttctctcactcacaacctatgaaggtggcagtgtaa 3400
G  N  N  T  W  Y  L  D  S  G  C  S  K  H  M  T  G  D  K  T  K  F  L  S  L  T  T  Y  E  G  G  S  V
          protease I

catttggtgataataagaaaggtaacattgttgctatgggtaaggtcggtaagtcaccattacactccattgaaaatgtattttttagttgaaggtctaaa 3500
T  F  G  D  N  K  K  G  N  I  V  A  M  G  K  V  G  K  S  P  L  H  S  I  E  N  V  F  L  V  E  G  L  K
  protease II

acataatctattaagcatatctcaattctgtgacaaaggaaatactgtaaaatttgataaagaaaaatgcttaatcatcaacattaaaactaagaaggtt 3600
 H  N  L  L  S  I  S  Q  F  C  D  K  G  N  T  V  K  F  D  K  E  K  C  L  I  I  N  I  K  T  K  K  V
             protease III

attttggaaggaacaaggaaagggaacacatacattgtggatctagaccttgttcctcaaatcaatttaacttgtcttagtgttattgaagatgattcac 3700
 I  L  E  G  T  R  K  G  N  T  Y  I  V  D  L  D  L  V  P  Q  I  N  L  T  C  L  S  V  I  E  D  D  S

ttctatggcacaagcgtctaggacatgctagcttttcgttgcttgagaaacttagatcaaaagaccttgttttaggattaccatctataaaatttcatat 3800
L  L  W  H  K  R  L  G  H  A  S  F  S  L  L  E  K  L  R  S  K  D  L  V  L  G  L  P  S  I  K  F  H  I
                                zinc finger motif of integrase

tgatcaagtttgtgatgcatgtgcacgaggtaaacaagtaagatcctctttcaaatctaaaacgattgtaagtaccactaaaccattagagttaattcat 3900
 D  Q  V  C  D  A  C  A  R  G  K  Q  V  R  S  S  F  K  S  K  T  I  V  S  T  T  K  P  L  E  L  I  H

atcgacttatgtggacccatgagaattcaaagcagaagtggaaagaggtatgtacttgtgattgtagatgactactctaggtacacttgggttattttc 4000
 I  D  L  C  G  P  M  R  I  Q  S  R  S  G  K  R  Y  V  L  V  I  V  D  D  Y  S  R  Y  T  W  V  I  F
                                integrase DDE motif

tctctagtaaagatgaaacttatgatgagttccttgtctttgctaaaagagttcaaaacctaagtggacataaaataatgcacattagatcagatcatgg 4100
L  S  S  K  D  E  T  Y  D  E  F  L  V  F  A  K  R  V  Q  N  L  S  G  H  K  I  M  H  I  R  S  D  H  G

caaggaattcgaaaattataaatttgatgacctaagcagagataatggcctagatcataatttttctgcccctagaacaccacaacaaaatggtgttgta 4200
  K  E  F  E  N  Y  K  F  D  D  L  S  R  D  N  G  L  D  H  N  F  S  A  P  R  T  P  Q  Q  N  G  V  V

gaaaggaaaaatagaactttggaggaaatgtctggaaccatgttaattgctagtgaacttccaaggaacttttgggctgaagctgttaatactgcttgtc 4300
 E  R  K  N  R  T  L  E  E  M  S  G  T  M  L  I  A  S  E  L  P  R  N  F  W  A  E  A  V  N  T  A  C

acataattaatcgtgccatgatgagacctatcattaataaaactccttatgaactttactttggaaagaaaccaaacatcacctatttagaacatttgg 4400
H  I  I  N  R  A  M  M  R  P  I  I  N  K  T  P  Y  E  L  Y  F  G  K  K  P  N  I  T  Y  F  R  T  F  G

atgcaaatgttatgtgcataataatggaaaagataatcttggaaaatttgatgcaaggagtgatgaagcaacattttaggatactcctcacatagtaaa 4500
  C  K  C  Y  V  H  N  N  G  K  D  N  L  G  K  F  D  A  R  S  D  E  A  T  F  L  G  Y  S  S  H  S  K
                                integrase GKGY motif

acttatagagtatttaacaaaagaactatgtgtgttgaagaaagcattcatgtaatctttgatgaatctgacaaacataatccaagcatacaggttgatg 4600
 T  Y  R  V  F  N  K  R  T  M  C  V  E  E  S  I  H  V  I  F  D  E  S  D  K  H  N  P  S  I  Q  V  D

actatgagatagggttggcgcaacctagtccaaggaacacagattcacaagaagaagaagaagtgaaagaggaaaggaacgaggaattcgaaaataatga 4700
D  Y  E  I  G  L  A  Q  P  S  P  R  N  T  D  S  Q  E  E  E  E  V  K  E  E  R  N  E  E  F  E  N  N  E

gaacaatgatattcctgttcctcaaggaggagaagatgctcctgttcctctaaatgaaggaactgagtccaccgaggctaacagaggaacaacagagcag 4800
  N  N  D  I  P  V  P  Q  G  G  E  D  A  P  V  P  L  N  E  G  T  E  S  T  E  A  N  R  G  T  T  E  Q

aactcctcatccactcaaggaactcatgaggaacaacatgttcctataagagagtttcaaccaaaaccttggagattacaaaagtcacatcctgtggaac 4900
 N  S  S  S  T  Q  G  T  H  E  E  Q  H  V  P  I  R  E  F  Q  P  K  P  W  R  L  Q  K  S  H  P  V  E

tcataatcagcgacatatccaaaggtacgcaaactagatctcagttaaggaacttttgtgcatttcatgcgttcctatcaatgatggaaccaagaaatca 5000
L  I  I  S  D  I  S  K  G  T  Q  T  R  S  Q  L  R  N  F  C  A  F  H  A  F  L  S  M  M  E  P  R  N  H

tgaagaagccttaattgattctaattggattattgcaatgcaagctgaattaaatgagtttgaaagaaacaaggtatggcatctagttccttcatctaaa 5100
  E  E  A  L  I  D  S  N  W  I  I  A  M  Q  A  E  L  N  E  F  E  R  N  K  V  W  H  L  V  P  S  S  K

caacaaaaagtaattggcttaaaatgagtgtttagaaacaaactagatgaatatggaaccattgtaagaaacaaagctagacttgtagtaaaaggttata 5200
 Q  Q  K  V  I  G  L  K  *  V  F  R  N  K  L  D  E  Y  G  T  I  V  R  N  K  A  R  L  V  V  K  G  Y
                    RT I                                            RT II

accaacaagaaggaattgattatgaagaaacctttgctcctgtagctagattagaagccattagaatcttaattgcatttgctgcttacatgggattcaa 5300
N  Q  Q  E  G  I  D  Y  E  E  T  F  A  P  V  A  R  L  E  A  I  R  I  L  I  A  F  A  A  Y  M  G  F  K

attatatcaaatggatgttaagtgtgcattttaaatggctatctaaacgaagatgtatatgttgaacaacccccctggtttcgaaaataacaatctccca 5400
 L  Y  Q  M  D  V  K  C  A  F  L  N  G  Y  L  N  E  D  V  Y  V  E  Q  P  P  G  F  E  N  N  N  L  P
              RT III

aatcatgtctacaagcttgataaagctctttatgggttaaaacaagcacctagatcatggtatgagagattatctaagttccttttagaaacaacttta 5500
 N  H  V  Y  K  L  D  K  A  L  Y  G  L  K  Q  A  P  R  S  W  Y  E  R  L  S  K  F  L  L  E  N  N  F
                          RT IV
```

gag

pol

```
aaagaggaaaggttgataaaaccttgttcctaaaatctaaaggaactgatattttgcttgttcaaatatatgttgatgacatcatattcggagctactaa 5600
K  R  G  K  V  D  K  T  L  F  L  K  S  K  G  T  D  I  L  L  V  Q  I  Y  V  D  D  I  I  F  G  A  T  N
                                                                       RT V

tgaaacattgtgcaaagaattctcaagactcgtgagcaatgaatttgaaatgagtatgatgggtgagttaaatttctttcttggactacaaatcaaacaa 5700
  E  T  L  C  K  E  F  S  R  L  V  S  N  E  F  E  M  S  M  M  G  E  L  N  F  F  L  G  L  Q  I  K  Q
                                           RT VI                           RT VII

actgaaaaaggtataattgttcaccaacaaaaatacataaaggaactattaaagaaatatggcctcgaaaattcaaagataaatcacactcctatgggaa 5800
 T  E  K  G  I  I  V  H  Q  Q  K  Y  I  K  E  L  L  K  K  Y  G  L  E  N  S  K  I  N  H  T  P  M  G

cttcaactagattagatgaagactccataggaacaagtgttgatcaaacaaaatatagaggaatgattggtttactttgtatctaactgctagtcgtcc 5900
T  S  T  R  L  D  E  D  S  I  G  T  S  V  D  Q  T  K  Y  R  G  M  I  G  L  L  L  Y  L  T  A  S  R  P

agacatagctttcagtgttggtttgtgtgctagatttcaagcaaatccaaaggagtctcacctcactgctgtaaagagaattcttagataccttaaagga 6000
  D  I  A  F  S  V  G  L  C  A  R  F  Q  A  N  P  K  E  S  H  L  T  A  V  K  R  I  L  R  Y  L  K  G

acagatgatcttggactctactatccaagaagtgatacattcgaactaaaaggatatgcagatgcagattatgctggagatcttgtaaataggaaaagca 6100
 T  D  D  L  G  L  Y  Y  P  R  S  D  T  F  E  L  K  G  Y  A  D  A  D  Y  A  G  D  L  V  N  R  K  S

cctcaggtatggcacaattccttggtcatagtttagtttcttggagtaccaagaaacaaaacacagttgctttatccactgctgaagcagaatatgtagc 6200
T  S  G  M  A  Q  F  L  G  H  S  L  V  S  W  S  T  K  K  Q  N  T  V  A  L  S  T  A  E  A  E  Y  V  A
                                                                          RNase motif I

tgctgctgcatgttgctctcaaatgctttggataaaacaacaacttagtgactatggaataaactttgaatgtgttcctatttattgtgacaatacaagt 6300
  A  A  A  C  C  S  Q  M  L  W  I  K  Q  Q  L  S  D  Y  G  I  N  F  E  C  V  P  I  Y  C  D  N  T  S

gctatcagtatatctaaagatccagtgcatcactctagagtaaagcatatacatattagacatcacttttaagagaaaatgtcgaaaaaggtttgatta 6400
 A  I  S  I  S  K  D  P  V  H  H  S  R  V  K  H  I  H  I  R  H  H  F  L  R  E  N  V  E  K  G  L  I
                                         RNase motif II

aacttgagttttgtcaaactgattatcagattgctgatatcctaactaagcctttgcaaagagatagatatgaaaaattaaggcttgagttatgcttaat 6500
K  L  E  F  C  Q  T  D  Y  Q  I  A  D  I  L  T  K  P  L  Q  R  D  R  Y  E  K  L  R  L  E  L  C  L  I

aaaaattaagtgactgttcctatgtcaattccccccatgtttgaatcaatttgagtgactttgtgtgctaaaaataattgtgtttaattgcctaagacca 6600
  K  I  K  *

ttgcaaattggaatgttgcaaaaacataaccatgaagtcaaagaaactaaacaactcacaaatgtcacataatactccctttactgttcctaaggtcaag 6700
aaacatgcatagtatgaaggatattgatatgtagcattagttgttcctatagcacaagttccttagcatttattttctttcaaaaattaaaattcataaa 6800
aatccgaaaaatccgataaaatcccaaaaaaataaggggaaccagtttaattaaatcttggttctttttatcttttcgtttttttttattattttttatttt 6900
tatttcttttgctttatttctttcccgcgtgtcataattactattacatgcaaaacttccttcttcaatccatctaaagaaaccgtgttcctctcaagcg 7000
cttcatcttcaccaaattcgaaacctcatcacaccctttgatactccataaaaacctccaaaccttcataaatcatggcaaaaacttcaagaaactcacc 7100
                                                                           M  A  K  T  S  R  N  S  P

ttttccaacaccaaaaacaaccgacacttcaaaaatggaaatttcttcatctccaactaaaacccttgaaactgacataattcaagaaaccccatagag 7200
  F  P  T  P  K  T  T  D  T  S  K  M  E  I  S  S  S  P  T  K  T  L  E  T  D  I  I  Q  E  N  P  I  E

gaacaacaattagaatctcctagaaggagcatgagaagaagaaaattgattggggaagacgatgaagaagaaaccccggaccaatctcaggggggaagaag 7300
 E  Q  Q  L  E  S  P  R  R  S  M  R  R  R  K  L  I  G  E  D  D  E  E  E  T  P  D  Q  S  Q  G  E  E

aagaaactgcacctgttccttccaaaggaactgtttcttcgtcttccaagaaaagggttgaacgaagtgaatctaggtatgtcttgggcttttgcgtcga 7400
E  E  T  A  P  V  P  S  K  G  T  V  S  S  S  S  K  K  R  V  E  R  S  E  S  R  Y  V  L  G  F  C  V  D

tttagaatgggcaaataaaattgaatttctagaattaacctccattctaaaggaacaaggcatggaattcttgtttgaattcgcctttaataatgctctt 7500
  L  E  W  A  N  K  I  E  F  L  E  L  T  S  I  L  K  E  Q  G  M  E  F  L  F  E  F  A  F  N  N  A  L

gtttctctagctatggatgagttttgtgcaaattttgttaatactaatggaacttgcacaactgaagttagaggcaagaaaataaggttcaatcaaagga 7600
 V  S  L  A  M  D  E  F  C  A  N  F  V  N  T  N  G  T  C  T  T  E  V  R  G  K  K  I  R  F  N  Q/ G

actgtttcctttccgaattttcctcaagataaaatcttagaatattttggaggaacagctggccaaaacaaacttaaccataatctactctctccccttc 7700
T  V  S  F  P  N  F  P  Q  D  K  I  L  E  Y  F  G  G  T  A  G  Q  N  K  L  N  H  N  L  L  S  P  L

acaaagtactgtttaacttggtttggagagcactcattcctcgaactgaaaagaggaatgaagtgggtctcttggatatgtgttacatgttttgcttgga 7800
H  K  V  L  F  N  L  V  W  R  A  L  I  P  R  T  E  K  R  N  E  V  G  L  L  D  M  C  Y  M  F  C  L  D

ccaacacatacaaattaatttcccttccctcttcattcaacacttaacccattgcattgaaaatcattgtgttattggctacggggccttgataacttcc 7900
  Q  H  I  Q  I  N  F  P  S  L  F  I  Q  H  L  T  H  C  I  E  N  H  C  V  I  G  Y  G  A  L  I  T  S

ttacttcatcattttgaagtaagacttactggttggaccaccataggggtcaagcaagggaacatattgaatgagaagaccctaaatggacttggattgt 8000
 L  L  H  H  F  E  V  R  L  T  G  W  T  T  I  G  V  K  Q  G  N  I  L  N  E  K  T  L  N  G  L  G  L

ccgtccaagaaggtgttctcattcaaggaagaatggcttccagttcctcaggaaaaagaaaaatgcccaaactggacttagaagaagatgatgatgtaat 8100
S  V  Q  E  G  V  L  I  Q  G  R  M  A  S  S  S  S  G  K  R  K  M  P  K  L  D  L  E  E  D  D  D  V  I

cttggataaccctgttcctaagccggaggaaccgaaacctaagaggcgcaaacacattgccactaagcctaaaccctctgttcctttgagagtgagaaag 8200
  L  D  N  P  V  P  K  P  E  E  P  K  P  K  R  R  K  H  I  A  T  K  P  K  P  S  V  P  L  R  V  R  K
                                proline-rich region

agcactcgtcgaaaaattgctcctgttcctgttccttccaacgaggactctcctttagtcctttcggataatgaggaacaggtaccctatattctcaaag 8300
S  T  R  R  K  I  A  P  V  P  V  P  S  N  E  D  S  P  L  V  L  S  D  N  E  E  Q  V  P  Y  I  L  K

aaccaatacctccaccaactcccaaactctctgttcctttcacaaaaatacctccatttcaccttcgtccttctttgttccttcaagttttgagcagtt 8400
E  P  I  P  P  P  T  P  K  L  S  V  P  F  T  K  I  P  S  I  S  P  S  S  F  F  V  P  S  S  F  E  Q  L

gaaggatgaaataaaaggaccaaatccttttgcttccacctccacaactcttccacccacaccaccatttcttcctcccatttatcacccttcaccttct 8500
  K  D  E  I  K  G  P  N  P  F  A  S  T  S  T  T  L  P  P  T  P  P  F  L  P  P  I  Y  H  P  S  P  S
```

pol

env

```
gtttccattcctggaacagatgaaccttccactcaaacctccccagttccttcacaacatcccccatctgcctctagttccaagggtaaggaacaggaaa 8600
 V  S  I  P  G  T  D  E  P  S  T  Q  T  S  P  V  P  S  Q  H  P  P  S  A  S  S  S  K  G  K  E  Q  E

atccggatgacactaactgggttaatgaagtctgccttcctgcacaacatatgataaaattctgcaatcatcttcattggactttagatgccaaaactga 8700
N  P  D  D  T  N  W  V  N  E  V  C  L  P  A  Q  H  M  I  K  F  C  N  H  L  H  W  T  L  D  A  K  T  D

                                                                                coiled coil

tactatgctaagtcttatgaagaaattgagtgatactgtggttgcgcagcaacagaagatggaaactatggctcgtaccttggaggagctgaaggaacag 8800
  T  M  L  S  L  M  K  K  L  S  D  T  V  V  A  Q  Q  Q  K  M  E  T  M  A  R  T  L  E  E  L  K  E  Q

gttgaactacatcatggggtggtgatgaaaaagttggaggatgtggttgaagaagaagttgtcgaggaacaccattccccttccgtttcttaatttcttt 8900
 V  E  L  H  H  G  V  V  M  K  K  L  E  D  V  V  E  E  E  V  V  E  E  H  H  S  P  S  V  S  *

aattgtcgttttgttgttcctgtctctttcatcaaggaacacaacttttgtttcttctttttcttttctttctttgttatggaacaagtgctatataatc 9000
aatgacttttgctatattgtctatgtttctaaacgctttgcataatttcttccttggaatggctagtattatgtgattgatgtttgagcaagttttctct 9100
ttgattttctaacttaaccactaatggtgcttgcatattggtcttaaggtgttaaatcaattcatcaattcatgtagagtctattctttttgaggatgac 9200
aaaggggaaaaggatttaaggaacaaactgtacataaacttcacataaacttaaaaaaatgaaaattgaaaaaaaatcaaaaattggataaggtaaatg 9300
aaaaatcaaaaataggttaaggtaaatgaaaaatcaaacttcagttactcacataattgtgtttttattttttcatttttcttctttatgtttttattttcat 9400
tcacttatattatttgttattttccttagtttgtacctttgttgtcatcatcaaaaaggggaaattgttgaacactagttttgataatgacaaagaaac 9500
                                                         PPT        3’ LTR →

aaaggaacaaactaagtagtgaactaatgaattatttaagtgtatggtgtctaaggtatcaatgcaaggaactgggagaatacgaagtcaatgaagaagt 9600
caagagaggatttcaaaggctaagtaaagaagtttctatgttgatactggagcaagcattggatcaagaataatagaagctaggcgttttattttctagt 9700
tccaatgtaaaggaacaaagaaagaatggtaaggaactgtggcagaatatttgagaatgaaagcgtttgttcctctgtaaggaacaaactcagtggcact 9800
tagcaaatatgttaaaggaactgaactaatataaggaactagacttaggcgtgtacttagtataacgattaaaagataatggagtaagttatgcctagat 9900
tataggaaattagttatttctaatttcaaatgataaaatcagttttaaaaagaaaatctgttttagaaaagattttatttttcggaaaataagaaaataa 10000
atatccggttgttagaaagatttatttcaaagatattattttcttaaagattagaaattcagatttaaagataaaaagagtttataaaaactgattttc 10100
ggataatgctgaaggaactcctattatttaggagtaatggctgcattttatccacgttcatttatagttcatgggatcatgggacatgatgtaaaacgtg 10200
ttccttaggggaaggaacatcatgggtaaatggattaaaacgtgtactctaacattaatcccacgttcttcatgattaacactataaatatagaagtcta 10300
gataagttccaaagcgtgtctgactaagtaattaagtgagtttaacttattaagtgtcagtgccttcattctctcataatagccatttgttctttgcctt 10400
cttactctaattccactaagcttatttcgatcaattgtattttggtttttggggtaagggaactttgagtgaagttctgtaagagaaaggctttgagtga 10500
agcttgtatgtgtgagaaaacttcgagtgaagttgagaggaactgctgttaagggaacagtggttcaggaacttgagtttaggaactcaaggtagggctc 10600
gagttagaattaggttgtaacagagttgtttggcctaattagtgaaagtgttgagtttaaaatccctagtggtcgaggttgtttcttcttgttgggccca 10700
agaagtttttcctcgtaaaatcccttgtgttcctttagcttgtttattcgtttaagttttaatttcggcaaaaagttacgtttatttctacacctacaa 10800
ttcacccccccctcttgtagtgttcctagggaaataacaatttt                                                         10843
                                         TSD
```

env

Appendix 2: Complete sequence and annotation of the integration event of LINE BNR1 into satellite pAv.

The complete sequence of the LINE BNR1 (EU564339) with a length of 6700 bp, integrated into the satellite pAv (in italics; Dechyeva and Schmidt, 2006) is presented. The coding sequences, deduced with the software *Genewise*, have been shown in grey shading, while their start and stop codons are colored in violet. FRameshifts are marked by a slash (‘/’). RNA-binding motifs like the RRM in ORF1 and the zinc finger in ORF2 have been highlighted in yellow, and conserved reverse transcriptase (RT) and endonuclease (EN) domains have been shown in red and green, respectively (Malik *et al.*, 1999; Wright *et al.*, 1996; Xiong and Eickbush, 1990; Wenke *et al.*, 2009). The target site duplication (TSD) flanking the element has been colored in red.

```
ATGCACAGGAACCCTAAGTCTACTCGGGGACCAGAAGTGGCATCCTTAGTTTGATTTATAAAATCTCCGAAATAACACTCACGCGTCTAAAAAATATAAG 100
pAV →                                                               TSD       LINE BNR1 →

TCTTATTTCTAAGATAGAGAAACTGTATAAGAATCGAGGAGAGATGGAGGAGAATACCCCTAGAGAGAGAAACCCCACCAGAGACCATGAAAACCCAATC 200
                                            M  E  E  N  T  P  R  E  R  N  P  T  R  D  H  E  N  P  I

AAAATCCTTGAAAACCCATGGATAACCATCAAAAGAAGAAAATCCAAGCCTCCAAACCACCAAGCTTCTCGAACCTGTTTTGTAAATCATCTCCCCCCCT 300
 K  I  L  E  N  P  W  I  T  I  K  R  R  K  S  K  P  P  N  H  Q  A  S  R  T  C  F  V  N  H  L  P  P

CCATAACTATACCAGAAATAGCTAGAATCTTCAGAACCCATGGGGCAATCGCAGAAATAACAATCCCAAAAACCCAGAACCAAACCAGCCACAAATTTGC 400
S  I  T  I  P  E  I  A  R  I  F  R  T  H  G  A  I  A  E  I  T  I  P  K  T  Q  N  Q  T  S  H  K  F  A

         RNA Recognition Motif

TTTTGTTCAATTTCATTACCCACAATCCCTTACTACAGCCATACGAGATGAAAACAAAAGAAAAGTTGAAACCATGGTAATTTCTGTGCACCCAGCAAAA 500
  F  V  Q  F  H  Y  P  Q  S  L  T  T  A  I  R  D  E  N  K  R  K  V  E  T  M  V  I  S  V  H  P  A  K

TATGATAAAAACCTCTATGCTCGAGCCAACAACCCTCCAACCACCCTCACATACTCATCTCGCGTAGCAAATACAAAACAATTAGCCAAAACCCGAAAA 600
 Y  D  K  N  L  Y  A  R  A  N  N  P  P  T  T  L  T  Y  S  S  R  V  A  N  T  K  Q  I  S  Q  N  P  K

AAGCATATACCAGAGACCTTCGCACATACAAAGAAGCTGCAAATCCCACAAAAACCACTGAAAACAAAAAAACAAATCACCCGCAAACCCGACCTCCAAA 700
K  A  Y  T  R  D  L  R  T  Y  K  E  A  A  N  P  T  K  T  T  E  N  K  K  T  N  H  P  Q  T  R  P  P  K

GCCAACGCCAATTCTGCAGTTCCATAATCCCAATAACCCAGTTCTTCCTTTCGAGGAGTATGTCCCCACAACAACCTCCTGTAAACCTGAACCATCGGCA 800
  P  T  P  I  L  Q  F  H  N  P  N  N  P  V  L  P  F  E  E  Y  V  P  T  T  T  S  C  K  P  E  P  S  A

CATCGAAAAATGAGCTCACGAGCTCTTGGAGAAGACACCGAAAAAATCAGAAGCACACTGGGAGCCATTGATATGGAAAGTGATTTTGCTGCTGCCTTAA 900
 H  R  K  M  S  S  R  A  L  G  E  D  T  E  K  I  R  S  T  L  G  A  I  D  M  E  S  D  F  A  A  A  L

AAGGCAAGAGATGCAAAGAAAATGAAGACATGCTCCAAAGAAGTTCAATAGCCTTCTCCCCGTCTTCGCAATCATCTGAAATTATTATGGATCACATCTT 1000
K  G  K  R  C  K  E  N  E  D  M  L  Q  R  S  S  I  A  F  S  P  S  S  Q  S  S  E  I  I  M  D  H  I  L

GGCCGAGGGAGTTAACTGCCTCACAATTAGACCCATGGGAGGAATGTTACACCTTCTAACTTTTGACACATTTGAAAATAAGAAAGCAATGATTGAGAGT 1100
  A  E  G  V  N  C  L  T  I  R  P  M  G  G  M  L  H  L  L  T  F  D  T  F  E  N  K  K  A  M  I  E  S

GGATGGCTCCAAAGATGGTTTTCAAAAATAATAAACGTCAATACTAGAAGTGCATCCCTGTGGAGGGAAACATGGGTAAATATATATGGAGTGCCTCTCA 1200
 G  W  L  Q  R  W  F  S  K  I  I  N  V  N  T  R  S  A  S  L  W  R  E  T  W  V  N  I  Y  G  V  P  L

TTGCTTGGGGATATGAGAGCTTTTACAACATTGGAAGTATGCTGGGGAGAGTTTTATCAGTCAATTACAAGGACTTTGATTGTGCAAGAGTTCTCCTTTT 1300
I  A  W  G  Y  E  S  F  Y  N  I  G  S  M  L  G  R  V  L  S  V  N  Y  K  D  F  D  C  A  R  V  L  L  F

CACAGATTGCTTCTTTGACATAAGTTGCAAGATATCTTTTGAGATTGAAGATGAGAAATACCCAGTCTTTATTTCAGAGAAGCAACAACTCTGGCAGAAC 1400
  T  D  C  F  F  D  I  S  C  K  I  S  F  E  I  E  D  E  K  Y  P  V  F  I  S  E  K  Q  Q  L  W  Q  N

AAAACGAGCCCAGAGTACAAAATTAGCAAGATCAATGATGCAAATGACGGAAATCCACTAGGTACAATCAAAGATGCAAAGGGACCTGATGATGACGAAA 1500
 K  T  S  P  E  Y  K  I  S  K  I  N  D  A  N  D  G  N  P  L  G  T  I  K  D  A  K  G  P  D  D  D  E

GTCAGTAATCCTCATGCAGCCAACTTCAGGTAAGCCTACTCCtCATGGTCGTGAATGTTTCTGATAATGAGAAGACAGAAAGCCACTTTTTGAAAAATGA 1600
S  Q  *  S  S  C  S  Q  L  Q  V  S  L  L  L  M  V  V  N  V  S  D  N  E  K  T  E  S  H  F  L  K  N  D

TGAGTTAATTATTAATGTGAAGATTCCCGAAAATACCTCATTATTGGGAAATGATGATGCCCCTGCTGGCAATAAAAAAAGTGACAAGACCACTGATGAG 1700
  E  L  I  I  N  V  K  I  P  E  N  T  S  L  L  G  N  D  D  A  P  A  G  N  K  K  S  D  K  T  T  D  E

GTCCCATGTACCGAAAAAATTGTCGATAATGATGATGTGCAATACCCTAGCAATGTCAAACTGACACCCTCAAAAAACGGCGCACGTGACCTAGAAAACA 1800
 V  P  C  T  E  K  I  V  D  N  D  D  V  Q  Y  P  S  N  V  K  L  T  P  S  K  N  G  A  R  D  L  E  N

AAAAAAAATCCCAACCTTCTAATGATTTTAACATCATTGAAACCAACAAAACCTCCATCCAAAATCCCCCTGGGCCCTCGCCAGAACTGGGCCTAATCTC 1900
K  K  K  S  Q  P  S  N  D  F  N  I  I  E  T  N  K  T  S  I  Q  N  P  P  G  P  S  P  E  L  G  L  I  S

CTGCAACGCTTCCTCCCCACCTATACCAATCCAAAGTATCCATCAAACAAGCCCAACACAACCTCACTCTTCACCTCCTACCTCCCCTTTGTCCCCAATT 2000
  C  N  A  S  S  P  P  I  P  I  Q  S  I  H  Q  T  S  P  T  Q  P  H  S  S  P  P  T  S  P  L  S  P  I

CACCTAACTAATAGATTCAAAGCCCTAGTTAGGCCCAGTTCCTCCATGTCATCATCCTCATCACTATCTGGGCCTCTGTTCCCGCCGGGCTTCGAAAATG 2100
 H  L  T  N  R  F  K  A  L  V  R  P  S  S  S  M  S  S  S  S  S  L  S  G  P  L  F  P  P  G  F  E  N

ACATTCCCTTAACCATCAAAGCCATTCATAAAAAAAAAGAGAGAAAAGAAAATCCAAAAAAAGAAAAAGCCTCCCTTTCTCCCTCTTCCAAATCAATCCTC 2200
D  I  P  L  T  I  K  A  I  H  K  K  K  R  E  K  K  I  Q  K  K  K  K  P  P  F  L  P  L  P  N  Q  S  S

CCCATCCCTTACAAGAGCTCTTATCCCTTCATCTTCGGAAAACTCAGTCTCATCCATATTAGAAGTGGGCAAAAAGCTTGGGATGCGCTTTAATGGGCCT 2300
  P  S  L  T  R  A  L  I  P  S  S  S  E  N  S  V  S  S  I  L  E  V  G  K  K  L  G  M  R  F  N  G  P
```

ORF1

```
GAAACTATCCTTGAGGAACGAATTGAGTCGATTCTTCAGCAGCATAAAGCCAGCTGGAAGGCTAATCAGAATTAGCCTCTCTCCACTCCAATCCATAAAA 2400  ORF1
 E  T  I  L  E  E  R  I  E  S  I  L  Q  Q  H  K  A  S  W  K  A  N  Q  N  *

CAAAAAAATCTTTAATGATTCTTCTATCCTGGAATTGTCGTGGACTAGGCGCTAGAATTAAGCGCAATGCAGTTAGGAAATTGATACAAAAAAATGATCC 2500
               M  I  L  L  S  W  N  C  R  G  L  G  A  R  I  K  R  N  A  V  R  K  L  I  Q  K  N  D  P
                                EN I

CCACTTGATATTCATCCAAGAGTCGAAGCTAGAATCAATCAGCCCGAAAGTCATGAAATCCATCTGTGATGTCAATGACATGAACTCCGCAATAAGTCCC 2600
  H  L  I  F  I  Q  E  S  K  L  E  S  I  S  P  K  V  M  K  S  I  C  D  V  N  D  M  N  S  A  I  S  P
                  EN II

TCAAATGGTCCTTCAGGGGGCCTAATATCCATATGGAAGAATAGCCTTCTTCATGGTTGAAGAATCaAGATGTGAGTGTAATGGATCATACTCACACGGC 2700
 S  N  G  P  S  G  G  L  I  S  I  W  K  N/   A  F  F  M  V  E  E  S  R  C  E  C  N  G  S  Y  S  H  G
                  EN III

TCTATACTATCAATCAACATGAAATGCAGATTTATTAATTTGTACAACCCATGTGATGTTCAGAGGCGTACAGAAGTATGGCGCGAATTAATCCAAATTT 2800
 S  I  L  S  I  N  M  K  C  R  F  I  N  L  Y  N  P  C  D  V  Q  R  R  T  E  V  W  R  E  L  I  Q  I
                                                    EN IV

GCGAATCATCACCACTCCCGTGCCTTTTCATAGGCGATTTTAATGAAGTCCTTGAGGCTTCTGAGAGAGGTAGCCAACATATTTTAACTAACAGTTCGAC 2900
C  E  S  S  P  L  P  C  L  F  I  G  D  F  N  E  V  L  E  A  S  E  R  G  S  Q  H  I  L  T  N  S  S  T
                                          EN V

AGAATTTAAAGATTTTTTGCAGGTCCTTCATCTCATAGAAATACCATCCTCAAATTCCAAATTCACTTGGTTTCGTGGCCAATCTAAAAGCAAGCTTGAC 3000
  E  F  K  D  F  L  Q  V  L  H  L  I  E  I  P  S  S  N  S  K  F  T  W  F  R  G  Q  S  K  S  K  L  D
                                                                            EN VI

CGAGTCTTTGTTCAGGCTCAGTGGATATCTGCATACCCTCTTATTCGAGTCTCTCATCTACAAAGAGGATTGTCCGATCACTGCCCTATTCTAGTTCAAT 3100
 R  V  F  V  Q  A  Q  W  I  S  A  Y  P  L  I  R  V  S  H  L  Q  R  G  L  S  D  H  C  P  I  L  V  Q
                                                                                         EN VII

CAAAAGAAAAGAATTGGGGGCCGAGGCCCTTCCGATTCCTTAATTGCTGGCTCTCTCACCCAGGATGCATGAAAACTATCTCAGACACCTTGGGCTAAAT 3200
S  K  E  K  N  W  G  P  R  P  F  R  F  L  N/       A  L  S  P  R  M  H  E  N  Y  L  R  H  L  G  L  N

CCCCAAAaTATGACCTTCATGGACAAGTTGAGATTATTGAAGACCaACCTTGAAGAATGGAATGCTGCAGAATTTGGTATCATTGAGGAAAAAATTTCTT 3300
 P  Q  N  M  T  F  M  D  K  L  R  L  L  K  T  N  L  E  E  W  N  A  A  E  F  G  I  I  E  E  K  I  S

TCTTCGAGAACAAGATTCATGAATATGACCTTATTGCAAaCAATAGAATGCTTGATAAAGCTGAACTGGAGGAACGCAAAGCGGCTCAAATAGAACTTTG 3400
F  F  E  N  K  I  H  E  Y  D  L  I  A  N  N  R  M  L  D  K  A  E  L  E  E  R  K  A  A  Q  I  E  L  W

GCAATGGTCAAAGCGCAATGAATCCTTTTGGGCCCAACACTCAAGGGCCAAATGGATCCGAGAGGGGGACCGCAATACCCGCTACTTTCATGTGATGGCC 3500  ORF2
  Q  W  S  K  R  N  E  S  F  W  A  Q  H  S  R  A  K  W  I  R  E  G  D  R  N  T  R  Y  F  H  V  M  A
                                                                                   EN VIII

TCAATAAGAAGGAGGAAAAATAACATTGAGTATCTCAAGGAAGGTGAACATATGATTGAAGATCCCACTGAAATCAAAATGGCCGCGACAAATTTTTTTA 3600
 S  I  R  R  R  K  N  N  I  E  Y  L  K  E  G  E  H  M  I  E  D  P  T  E  I  K  M  A  A  T  N  F  F

AAAATCTATTCACGGAAGAGCATGAAATCAGACCTGTTTTTGAAGGTCTCGACTTCAAAAGGCTCGGTGAACAACACGAGCATATTCTCACTGATCCTTT 3700
K  N  L  F  T  E  E  H  E  I  R  P  V  F  E  G  L  D  F  K  R  L  G  E  Q  H  E  H  I  L  T  D  P  F

CTCTACTGCGGAAATAGATGCAGCAGTAGCCGCTTGTGACAGCTCGAAATCTCCCGGACCAGATGGtTTTAATTTCATGTTTATCAAGAACTCTTGGGAT 3800
  S  T  A  E  I  D  A  A  V  A  A  C  D  S  S  K  S  P  G  P  D  G  F  N  F  M  F  I  K  N  S  W  D
                                                      RT 0

CTTATCAAGGAAGACATCTATGCTATTGTGTTCGAATTTTGGCAAACTTCTCACCTTCCGAAAGGATGCAACACTGCTCTTGTTGCCTTAATCCCTAAAA 3900
 L  I  K  E  D  I  Y  A  I  V  F  E  F  W  Q  T  S  H  L  P  K  G  C  N  T  A  L  V  A  L  I  P  K
                                                                               RT I

CCAATCCTCCAAATGGTTTTAAGGACTTCAGGCCGATAAGTATGATAGGGTGTGTCTATAAAATAATATCGAAAATCCTCGCTAGGAGATTGCAGCAAGT 4000
T  N  P  P  N  G  F  K  D  F  R  P  I  S  M  I  G  C  V  Y  K  I  I  S  K  I  L  A  R  R  L  Q  Q  V
                                                  RT II

TATGGCGCATCTAGTAGGTTCTCACCAATCGGCCTTCATAAAAGGAAGGCAAATCCTGGATGGTGCCTTGATTGCGGGGGAAGTTATCGAATCTTGCAAA 4100
  M  A  H  L  V  G  S  H  Q  S  A  F  I  K  G  R  Q  I  L  D  G  A  L  I  A  G  E  V  I  E  S  C  K

CGGAACAAAGTGGATGCCACAATATTCAAAATTGACTTTCATAAGGCGTTTGATAGCGTCTCCTGGGGATTTATGGAATGGACTCTCACTCAAATGAATT 4200
 R  N  K  V  D  A  T  I  F  K  I  D  F  H  K  A  F  D  S  V  S  W  G  F  M  E  W  T  L  T  Q  M  N
                                      RT III

TTCCAAAGCAGTGGCGTGAATGGATTAGAGCTTGTGTCTTATCTGCCTCTGCAAGCATTCTCATTAATGGTTCTCCTACTTCTCCCATTAAACTTCGTCG 4300
F  P  K  Q  W  R  E  W  I  R  A  C  V  L  S  A  S  A  S  I  L  I  N  G  S  P  T  S  P  I  K  L  R  R

TGGCCTGAGACAGGGAGATCCCCTCTCCCCGTTCCTCTTTACTCTGATTGCGGAGCCACTAAATCTCCTTATTAAGAAGGCAGTTTCTCTAAGCCTATGG 4400
  G  L  R  Q  G  D  P  L  S  P  F  L  F  T  L  I  A  E  P  L  N  L  L  I  K  K  A  V  S  L  S  L  W
                        RT IV

GAGGGGGTCGAAATTTGCAGAGGTGGCCTCAAAATTACCCATCTTCAGTACGCAGACGACACTGTGCTCTTCTGCCCCCCGAAATTGGAGTTCCTGGAAA 4500
 E  G  V  E  I  C  R  G  G  L  K  I  T  H  L  Q  Y  A  D  D  T  V  L  F  C  P  P  K  L  E  F  L  E
                                           RT V

ATATAAAAAGAGTTCTAATACTCTTTCACCTTGCTTCTGGACTACAAATAAATTTCCACAAGAGTTCCCTCATGGGGATCAACATTGAACCGaAACTTCT 4600
N  I  K  R  V  L  I  L  F  H  L  A  S  G  L  Q  I  N  F  H  K  S  S  L  M  G  I  N  I  E  P  K  L  L
                                              RT VI
```

```
TGATCACATGGCATCTCAACTGCTATGCAAAGTGGGTTCACTTCCCTTTGTCTATTTAGGGCTTCCAATTGGTGGCAGCGCATCTCGTATTAATCTATGG 4700
  D  H  M  A  S  Q  L  L  C  K  V  G  S  L  P  F  V  Y  L  G  L  P  I  G  G  S  A  S  R  I  N  L  W
                                                   RT VII

GAGCCGGTCATAGCAAAAATCGAGAAAAAATTGGCCTCATGGAAAGGCAATCTTCTTTCCATTGGAGGAAGAGTCACACTCATAAAATCATGCCTTGCAA 4800
 E  P  V  I  A  K  I  E  K  K  L  A  S  W  K  G  N  L  L  S  I  G  G  R  V  T  L  I  K  S  C  L  A

GCCTCCCATTATACTACATGTCTCTTCTCCCGATGCCTAAAGGGGTGATTGAAAAAATCATTCAACTGCAAAGAAACTTTCTTTGGCGTGGTAGTTTGGA 4900
S  L  P  L  Y  Y  M  S  L  L  P  M  P  K  G  V  I  E  K  I  I  Q  L  Q  R  N  F  L  W  R  G  S  L  E

GAAGAAAGCTCTCCCCCTTGTGTCGTGGAATGTGCTTGAACTTCCCAAGCAATATGGAGGACTGAGTATTGGTAATCTTCATAATAAAAACACTGCTCTC 5000
  K  K  A  L  P  L  V  S  W  N  V  L  E  L  P  K  Q  Y  G  G  L  S  I  G  N  L  H  N  K  N  T  A  L

CTATTCAAATGGCTTTGGAGATTCATTCATGAGCCGAACTCATTATGGCGTCAAATAGTTCAAGCAAAATATGATATTGGACCGACCTTCACTATTCGGG 5100
 L  F  K  W  L  W  R  F  I  H  E  P  N  S  L  W  R  Q  I  V  Q  A  K  Y  D  I  G  P  T  F  T  I  R

ACTTGACAACCCCACCACATGGTGGCCCATGGCGAGGTATCTGTAATCTAATCCAATCTTCTTCTCATGCGTATCAAATTGCAACTCACATGATAAGAAA 5200
D  L  T  T  P  P  H  G  G  P  W  R  G  I  C  N  L  I  Q  S  S  S  H  A  Y  Q  I  A  T  H  M  I  R  K

GAATATAGGTGATGGTTCAAGCACTATGTTTTGGCATGATGTTTGGGTTGGTGAACATCCTCTCAAAGAGGTATGCCCCCGCCTCTTCCTCCTCTCTTTA 5300

  N  I  G  D  G  S  S  T  M  F  W  H  D  V  W  V  G  E  H  P  L  K  E  V  C  P  R  L  F  L  L  S  L

TCTCCAAATGCTCTAGTCTCCTCCTGTGGTTTTTGGGATGGACAAATCTGGCATTGGAGTCTCCATTGGAAGCGAAACCTTCGTCCGCAAGATCGATGCG 5400
 S  P  N  A  L  V  S  S/  V  V  F  G  M  D  K  S  G  I  G  V  S  I  G  S  E  T  F  V  R  K  I  D  A

AGtCGTGAATCCTTGCAAGTCCTTCTAGATAGAGCAGTGCTCTACCAGGATGGTCATGATCAAACCATCTGGACCCCAGCTAAATCCGGGAAATTCTCAG 5500
 S  R  E  S  L  Q  V  L  L  D  R  A  V  L  Y  Q  D  G  H  D  Q  T  I  W  T  P  A  K  S  G  K  F  S

TGAAATCTTTCACATTGGAACTAGCAAAAAAAGATGTACCGCAAAACTTTGATGCTTCAAAAGGATTCTGGAAGGGCCTTGTCCCTTTTAGAATTGAAAT 5600
V  K  S  F  T  L  E  L  A  K  K  D  V  P  Q  N  F  D  A  S  K  G  F  W  K  G  L  V  P  F  R  I  E  I

ATTTGTGTGGTTTGTACTTCTTGGTAGATTGAATACAAAAGAAAAACTATGGAGATTGGGAATCGTTCCAGAATCGGAAAAAAATTGCGTGCTCTGTAAT 5700
  F  V  W  F  V  L  L  G  R  L  N  T  K  E  K  L  W  R  L  G  I  V  P  E  S  E  K  N  C  V  L  C  N

ATTCACCCGGAGTCAGTTAATCATCTATTCATGGGTTGTACAGTGGCGTCGGAGCTTTGGCTGTGGTGGTTGAGTATTTGGGGAGTTTCGTGGGTTTTTC 5800
 I  H  P  E  S  V  N  H  L  F  M  G  C  T  V  A  S  E  L  W  L  W  W  L  S  I  W  G  V  S  W  V  F
    zinc finger motif

CCTCGACTTTAAAAAGCCTCCATAATCAGTGGCACGCTCCGTTCAGAGGATCTATCATTAAAAAATCATGGCAAGCAATATTTTTCATCATCTTGTGGAC 5900
P  S  T  L  K  S  L  H  N  Q  W  H  A  P  F  R  G  S  I  I  K  K  S  W  Q  A  I  F  F  I  I  L  W  T

CATTTGGAAGGAAAGAAACGGGCGAATCTTTGAAAATAAAGAATGCTCCATGTCTCAGTTGAAAGATCTAATTCTTCTAAGGCTAAGTTGGTGGTTGAAA 6000
  I  W  K  E  R  N  G  R  I  F  E  N  K  E  C  S  M  S  Q  L  K  D  L  I  L  L  R  L  S  W  W  L  K

GGGTGGGGAGATTTCTTCCCCTATAGTTCTACTGATACTCCTTCGAAACCCGCAATGTCTCCTATGGAATTCCAAGCCGTGCCCAAGTAATCACCTGCCT 6100
 G  W  G  D  F  F  P  Y  S  S  T  D  T  P  S  K  P  A  M  S  P  M  E  F  Q  A  V  P  K  *  S  P  A

AGCCGCCTACTGCTGTCGAGTCGTGGTCTCCTCCTCCATTTGGGTCGTTGAAGTGGAATGTAGATGCATCATGCAGCTCAATATTTGAATCCTCATCAAT 6200
*  P  P  T  A  V  E  S  W  S  P  P  P  F  G  S  L  K  W  N  V  D  A  S  C  S  S  I  F  E  S  S  S  I

TGGGGGTGTGCTCCGGGATCACAATGGTAATTTTAATCTGCATGTTTTCcGAGGCCTATCCCCGgTTTATGGAAATCAyATAATCGCCGAAGTGCTAGCT 6300
  G  G  V  L  R  D  H  N  G  N  F  N  L  H  V  F  R  G  L  S  P  V  Y  G  N  H  I  I  A  E  V  L  A

ATCCATCGTGCcACTTAAAAATATCAGCATCTTGTGAAAGATCTCATGAACCTGCCCATAATGATTGAGTCGGACTCCGTTAATGCAGTGAAATGGTGCA 6400
 I  H  R  A  T  *  K  Y  Q  H  L  V  K  D  L  M  N  L  P  I  M  I  E  S  D  S  V  N  A  V  K  W  C

AGCAAGATAAGGGTGGCCCATGGAACATGAATTTTATACTCAACTTCATTCGAGGGGAGTCGAGAAAGAGCCCGGGCATGTCAATTACCTACAAAGGTCG 6500
K  Q  D  K  G  G  P  W  N  M  N  F  I  L  N  F  I  R  G  E  S  R  K  S  P  G  M  S  I  T  Y  K  G  R

AGCATCAAATATGGTGGCAGACGCATTAGCTAAGCAAGGGTTAAGTAGAAGGGATGAATTTATTGCcTTGGCTCTAAAAATCACATGTATTTCATTTTAT 6600
  A  S  N  M  V  A  D  A  L  A  K  Q  G  L  S  R  R  D  E  F  I  A  L  A  L  K  I  T  C  I  S  F  Y

CTAGTTTGCTTCATAGTATAAGAGAGTTGTAATATGGTGTATGAACATTTTAAGGATGGAAATCCTCCTCCTCTTCGGGTGTGAGTATTCACTATGAAAG 6700
 L  V  C  F  I  V  *

TTTCCCTCCTTCTGAGGAGGAATACCTGGGTTGAGTCGAGGTTTTCTACCCTTAACTATAAATGAATAAAACCAAATTATCAAAAAAAAAACACTCACGC 6800
                                                                                  Poly(A)     TSD

GTCTACCCCGGTTCACTAACTCGGAGGATTTTAAAAATATATTTGATTTCTCATAAAATGACACTGTAAAGCACATTTGACCAAAAGCGCCCAAAATAGT 6900
      pAV →

GGAAGCCAAAGTCTTCCAAAAGTTTCGTGTCGCCTTTTCACGATAGTTTTAGGTAATCACGAAAATTGTGTCACAGGCGTGTACCGGGGGTCACCGAACA 7000
GAGAGAGTTTAAAGAATTGTTGAAATCTTTAGAAAAATGGTATCGGAAGGAACTTCGGCTCtGTATAGAATTCAATGCACAGGAACCCTAAGTCTACTCG 7100
```

ORF2

Appendix 3: Sequence annotation of the 5' truncated *B. vulgaris* RTE LINE *Ghost*1.
The sequence of the 5' truncated RTE LINE with a length of length of 2775 bp and an 11 bp TSD is presented. The point of truncation is located near the 5' site of the endonuclease gene. Therefore, ORF1 is missing and ORF2 is incomplete. Similarity to the endonuclease/ DNase superfamily has been detected by Interpro, however, the domains as suggested by Wenke *et al.* (2009) have been only vaguely related. Only EN IV contained at least two amino acids as described in this study (shown in green). Reverse transcriptase domains have been marked by red color (Malik *et al.*, 1999; Wright *et al.*, 1996; Xiong and Eickbush, 1990). The poly(GTT) tail typical for RTE LINEs has been highlighted as well.

```
TTTGCATTTTGCCGTCAGATTGACCATTGGTATTGGTATATGCAACAATAGGATTGGAAGCATTTATTAAACAACAATTTTGGGAGGATTTAGAGGAGGT 100
    TSD                   ... V  L  V  Y  A  T  I  G  L  E  A  F  I  K  Q  Q  F  W  E  D  L  E  E  V

TGTGCAACGAGTGCCTATTAATGAAAAATTGATCATTGGTGGGGATCTTAACGGACATGTGGGCACTAGTCGAGTTGGCTTTGAGAGCATTCATGGAGGT 200
  V  Q  R  V  P  I  N  E  K  L  I  I  G  G  D  L  N  G  H  V  G  T  S  R  V  G  F  E  S  I  H  G  G

TTTGGGTATGGGGAGAGAAATGAAGCTGGAAGTGGTATGTTGGATTTCACATTAGCCTATGGTTTTAGTATTATGAACACTTTGTTTGAGAAAAGAGAAT 300
 F  G  Y  G  E  R  N  E  A  G  S  G  M  L  D  F  T  L  A  Y  G  F  S  I  M  N  T  L  F  E  K  R  E
                                             Similarity to EN IV

CTCACCTTGTATAGGAGTGGAAGTAACGCGAGTCAGATTGACCTCTTTTTAGTAAGGATTGCTTGGCGGAAGAGTTATACAAACTGCAAGGTGATACCGG 400
S  H  L  V/  R  S  G  S  N  A  S  Q  I  D  L  F  L  V  R  I  A  W  R  K  S  Y  T  N  C  K  V  I  P

GTGAGAGCGCAACCACCCAACATAGACTAGTCGTGCTTGATTTCCGGGCTAGGGAGTCTTTAAGAAAGCGAAAACTCAATGTGGAGCCGCGAATCAAGTG 500
G  E  S  A  T  T  Q  H  R  L  V  V  L  D  F  R  A  R  E  S  L  R  K  R  K  L  N  V  E  P  R  I  K  W

GTGGAAGCTCCAAGGGGAGCACAAAGAGAACTTCTTAGCTAAGATGGCTAGTAAAGATATCTGGTCTTGTAACACGGAAATGGATGTAGATTCGACATGG 600
  W  K  L  Q  G  E  H  K  E  N  F  L  A  K  M  A  S  K  D  I  W  S  C  N  T  E  M  D  V  D  S  T  W

ACAAAGATAGAACTTAGTATAAAGGAAGTGGCGAAAGAGGTCCTTGGAGAATCTAAAGGAACTGTACCACCAAGTAAGGACACATCTGGGTGGAACGAGG 700
 T  K  I  E  L  S  I  K  E  V  A  K  E  V  L  G  E  S  K  G  T  V  P  P  S  K  D  T  S  G  W  N  E

CGGTGAGACAAGCGATAAAAAACAAACGAGAAAGTTATAAACTTTTGGGAAAAGGTATCAGTGATGTGAATTATGAGAAGTATACAGAGGCTAGAAAAGA 800
A  V  R  Q  A  I  K  N  K  R  E  S  Y  K  L  L  G  K  G  I  S  D  V  N  Y  E  K  Y  T  E  A  R  K  E

AGCCAAGAAGGCAGTACGAGAAGCTAAAGCACAGGTAAATAAAGAGCTTTATGCAAGATTGGATACGAAAGAAGGTGAAAAAGACATTTATAGAATTGCA 900
  A  K  K  A  V  R  E  A  K  A  Q  V  N  K  E  L  Y  A  R  L  D  T  K  E  G  E  K  D  I  Y  R  I  A

CGTATGAGGGACAGGAAGATAAGAGACATAAGAAAGGTAAAATGCGTCAAGGGTGTGGACCAAAGAGTGCTTGTGGAGGATGAGGAGATAAAGGTTAGAT 1000
 R  M  R  D  R  K  I  R  D  I  R  K  V  K  C  V  K  G  V  D  Q  R  V  L  V  E  D  E  E  I  K  V  R

GGAGATCGTATTTTGATACCTTGTTCAACGGTCACAAGCCATCGGGGATGTACACATCCCGCCAAGTATGTTAAATCGTGAGTTTATGAGGAGGATTCAA 1100
W  R  S  Y  F  D/      V  Q  R  S  Q  A  I  G  D  V  H  I  P  P  S  M  L  N  R  E  F  M  R  R  I  Q

AAGGTAGAAGTCGTTTTAGCATTAAAAAGATGGGATTGAAGAAAGCTACGGGGCTAGATGGCATACCTATAGAAGTTTGGAGGTGTTTAgGGGGGAGAGG 1200
K  V  E  V  V/   S  I  K  K  M  G  L  K  K  A  T  G  L  D  G  I  P  I  E  V  W  R  C  L  G  G  R  G
                                               RT 0

AATCGAATGGCTAACAACATTCTTCAACATGATTTGGAGACATAACAAAATGCCATTAGAGTGGAGGAAGAGTACTTTGATTCCTTTGTATAAGACCAAA 1300
  I  E  W  L  T  T  F  F  N  M  I  W  R  H  N  K  M  P  L  E  W  R  K  S  T  L  I  P  L  Y  K  T  K
                                                                      RT I

AACGATGTACAAGATTGTACCAACTACCGAGGAATCAAACTAATGAGTCATACTATGAAACTCTGGGAACGGGTGATTGAGGAAAGACTTAGGAAACATG 1400
 N  D  V  Q  D  C  T  N  Y  R  G  I  K  L  M  S  H  T  M  K  L  W  E  R  V  I  E  E  R  L  R  K  H
                                                RT II

TAAAGATATCGGAGAACCAATTTGGATTTATGCCTGGAAGATCGACTATAGAGGCCATCCACCTCATAAGACAAACAATGGAACATTACCGTGATAGGAA 1500
V  K  I  S  E  N  Q  F  G  F  M  P  G  R  S  T  I  E  A  I  H  L  I  R  Q  T  M  E  H  Y  R  D  R  K

GAAGGATCTACATATGGTTTTCATTGATTTGGAGAAGGCATATGATAGAGTACCAAGAGAAATACTTTGGTGGGCGTTGGCAAGGAAAGGTGTCCCACTG 1600
  K  D  L  H  M  V  F  I  D  L  E  K  A  Y  D  R  V  P  R  E  I  L  W  W  A  L  A  R  K  G  V  P  L
                            RT III

AAATATATAGATATCATCAAGGATATGTATGAGGGGGCAACTACGAGTGTGCAGACTACGGTGGGGAGGACAGAAGAGTTTCCTATTACGATCGGAGTGC 1700
 K  Y  I  D  I  I  K  D  M  Y  E  G  A  T  T  S  V  Q  T  T  V  G  R  T  E  E  F  P  I  T  I  G  V

ATCAAGGTTCTGTGCTTAGTCCCTTTCTCTTTGCTCTAGTCATGGACGAACTAACGAGGTCAGTTCAAGATGATATACCATGGTGTATGATGTTTGCGGA 1800
H  Q  G  S  V  L  S  P  F  L  F  A  L  V  M  D  E  L  T  R  S  V  Q  D  D  I  P  W  C  M  M  F  A  D
           RT IV                                                                              RT V

TGATATTGTGTTGATTGATGAGACAAAGGAAGGTGTTGAGAGTAAGTTGGAATTGTGGAGGCATACATTAGAAGCCCGTGGTTTTAGATTGAGCAGAAGT 1900
  D  I  V  L  I  D  E  T  K  E  G  V  E  S  K  L  E  L  W  R  H  T  L  E  A  R  G  F  R  L  S  R  S
                                                                                     RT VI

AAGACAGAGTATATGGAGTGCAAGTTCAGTGGGGTTACAAGCAGCGAGGCGGGGGCTGTCACTTTGGATAGCAAAGTGGTTCAAGGATCGAACGTCTTCC 2000
 K  T  E  Y  M  E  C  K  F  S  G  V  T  S  S  E  A  G  A  V  T  L  D  S  K  V  V  Q  G  S  N  V  F
```

ORF2

```
GTTACCTAGGATCCATTCTCCAAAAGGATGGAGAACTGGATGGAGACGTGGCGCATAGAATTAACGCCGGTTGGTTAAAGTGAAAAAGCGCCACCGAATT 2100
R  Y  L  G  S  I  L  Q  K  D  G  E  L  D  G  D  V  A  H  R  I  N  A  G  W  L  K  *  K  S  A  T  E  F
  RT VII

CCTTTGTGATTCCGGCATTCCCCAGAGATTGAAGGGAAAGTTCTACCACACTGCAATTAGGCCTGCTTTGTTATATGGCACCGAATGTTGGGCGGTGAAA 2200
  L  C  D  S  G  I  P  Q  R  L  K  G  K  F  Y  H  T  A  I  R  P  A  L  L  Y  G  T  E  C  W  A  V  K

CAATGCCACGTTCATAAGATGACCGTGGCGGAGATGCGTATGTTGCGGTGGATGTGTGGCCACACAAGGAAGGATCGATTAAGAAATGAACAAATCCTGG 2300
 Q  C  H  V  H  K  M  T  V  A  E  M  R  M  L  R  W  M  C  G  H  T  R  K  D  R  L  R  N  E  Q  I  L

AAAAAGTTGGAGTTGCATCCATTGAAGAAAAGATGAGGGAGAATCGATTAAGGTGGTTTGGTCATGTGAAAAGGAGATCAGGCGATGCACCAATGAGAAG 2400
E  K  V  G  V  A  S  I  E  E  K  M  R  E  N  R  L  R  W  F  G  H  V  K  R  R  S  G  D  A  P  M  R  R

AATTGAAGAGTGGAGTAATCAAATTGTAAAGGGTAGGGGAAGACCTAAGATGACTTGACTGAGGGTAATTGAAAGTGATATGAGGTTACTTGGGATTGAG 2500
  I  E  E  W  S  N  Q  I  V  K  G  R  G  R  P  K  M  T  *  L  R  V  I  E  S  D  M  R  L  L  G  I  E

GAGAGCATGACGTTGGATAGAAGTGGGTGGAGGGGCAGTATCTATGTGGAGGAAGGGGTCTGATTACTGCATTTAtTTGTTTTTCTCAATTGTGTTGGTA 2600
 E  S  M  T  L  D  R  S  G  W  R  G  S  I  Y  V  E  E  G  V  *

TTAAAACCTTTTGCTGAAAACCTCTTCTAAACACTTTTTGGTGTAAACCTTGTTTTCAGAAAAAAAAAAAAAACTTTTATCACTCTTGACGATGCCATCTC 2700
CTCTGGTCACGATTCCAGTTGACGATTCATGTTAGCCGACCCCAAATCACCTTGGGAATAAGGCTTAGTTGTTGTTGTTGTTGTTGTTTGCATTTTG      2797
                                                                       Poly(GTT)         TSD
```

ORF2

Appendix 4:. Alignment used for sequence identity caculation of BNR-like reverse transcriptase sequences.
By PCR using genomic DNA from the four sections of *Beta* 1 kb amplicons were amplified, cloned and four clones were sequenced of each species. A 172 amino acid sequence of RT domains IV to VII was deduced and aligned. Similar regions of the characterized LINEs *Karma* (*O. sativa*), BvL2 (*B. vulgaris*), *LIb* (*I. batatas*), Ta*11-1* (*A. thaliana*), BLIN (*H. vulgare*), cin4 (*Z. mays*), *del2* (*L. speciosum*) and Zepp (*C. vulgaris*) were used as outgroups. The percentage of amino acid conservation was illustrated by 67 % conservation shading. Black boxes indicate identical and light grey similar residues.

```
                                    10        20        30        40        50        60        70        80        90
                           ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
B. vulgaris BNR-RT1        FMEWTLTQMNFPKQWREWIRACVSSASASILINGSPTSPIKLRRGLRQGDPLSPFLFTLIAEPLNLLIKKA-VSLSLWEGVEICRGGLKI
B. vulgaris BNR-RT2        GMEWTLTQMNFP-QWREWIR-CVLSASASILINGSPTSPIKLRRGLRQGDPLSPFLFTLIAEPLNLLIKKA-VSLSLWEGVEICRGGLKI
B. vulgaris BNR-RT3        MLDWVLDQMKFPQQWKEWIYSCVMSASASILINGSPTTTFKLQRGLRXGDPLSPFLFNLAVEPLNLPFHKA-ISPRLWEGIELCKNGTTI
B. vulgaris BNR-RT5        LLDGVHEQMGFPLLWRAWIKSCVT-ASASILINGSPTHPVKLHRGLRQGDPLSPFLFNLAVETLNLILKKG-LDLGLWDGIATRPNGVVI
B. adanensis BNR-RT1       FLDWILQQMQFPPRWRQWINSCVMTTSASILINGSPTVPFKLQRRIQ-GDPLSPFLFDMVVESLSLLIQQA-ISTNLWEGIKIGNSGISI
B. adanensis BNR-RT2       FLDWILQQMQFPPRWRQWINSCVMTTSASILINGSPTVPFKLQRGLRQGDPLSPFLFDMVVESLSLLIQQA-ISTNLWEGIKIGNSGISI
B. adanensis BNR-RT3       FLDWILQQMQFP-RWRQWINSCVMTTSASILINGSRTVPFKLQRGLRQGDPLSPFLFDMVVESLSLLIQQA-ISTNLWEGIKIGNSGISI
B. adanensis BNR-RT4       FLDWILQQMQFPPRWRQWINSCVMTTSASXLINGSPTVPFKLQRGLRQGDPLSPFLFDMVVKSLSLLIQQA-ISTNLWEGIKIGNSGISI
B. macrocarpa BNR-RT1      FLDWTLGKMNFPPLWRKWIRSCLMTASASILINGCPTRPFKMQRGLRQGDPLSPFLFDIIVEPLNLIIQKA-SRLALWEGIEIGHKGLKI
B. macrocarpa BNR-RT4      FMDRVMTQMTFPLKWREXIKSCVMSASASILINGSPNN-FNLQQGLRQSDPLLPFLFDLVVESINLLFKKA-TSLNLWKGVGICWNGIKI
B. macrocarpa BNR-RT3      -----------PPLWRKWIRSCLMTASASILINGCPT-PFKMQRGLRQGDPLSPFLFDIIVEPLNLIIQKA-SRLALWEGIEIGHKGLKI
B. macrocarpa BNR-RT5      ---WTLTQMGFPPQWCNWIHSCITSAAASILVNGSPTTPFKLHRGLRQGDPLSPFLFDLVVETLSLVIQKA-VDLNLWDGVDMGNGELKL
B. patula BNR-RT1          FLDWTLAQMNFPMKWRDWIRTCVSSASASILINGSPTPPIKLSRGLRQGDPLSPFLFTLIAEPLNLLIKKA-VSLSLWDGIEVCRGGPKI
B. patula BNR-RT2          F--WTTGKMNFPPTMEEVDKKCVVTASTSILINGCPTRPFKMQRGLRQGDPLSPFLFDIIVEPLNLIIQKA-SRLALWEGIEIGHKGLKI
B. patula BNR-RT4          FLDWTLAQMNFPMKWRDWIR-CVSSASASILINGSPTPPIKLSRGLRQGDPLSPFLFTLIAEPLNLLIKKA-VSLSLWDGIEVCRGGPKI
B. patula BNR-RT5          YLGWVMKQMGFPPCWQRWMKACVMTASASILIXXSPTPPIKLHRGLRQGDPLSPFLFNLAVEPLNXLMKKG-XNXHLWEGIAPRPNGCAV
B. corolliflora BNR-RT6    FLDWVLEQMCFPSQWRGWIRSCVMSASASILVNGSPSFPFKLQHGLRQGDPLSPFLFDLVVEALSQLIQKA-TSLGLWEGIEVCKNGFKL
B. corolliflora BNR-RT1    FLDWVVDQMCFPPQWRDWIRSCVMSASASILINGSPTPPFKLQRGLRQGDPLSPFLFDLVVESLSLIIQKA-TSLNLWEGIEICRNGVKL
B. corolliflora BNR-RT4    FIEWTLVQMNFPEQWCKWMRACVMSATASILINGSPTAPIKLHRGLRQGDPLSPFLFDIIAESLNLLINKA-VSLGLWEGVEICRGGFKL
B. corolliflora BNR-RT5    FIEWTLVQMNFLEQWCKWMRACVMSATASVLINGSPTAPIKLHRGLRQGDPLSPFLFDIIAEPLNLLINNA-VSLGLWEGIEICRGGFKL
B. nana BNR-RT1            YLDWVLEQMKFPPQWRNWIQSCVMSASASILINGSPSSIFKLQRGLRQGDPLSPFLFNIAVEPLSLLFHKA-ISMNLWEGVEIRRNGLKI
B. nana BNR-RT3            FLDWTLGKMNFPPLWRKWIRSCLMTASASILINGCPTRPFKMQRGLRQGDPLSPFLFDIIVEPLNPIIQKA-SRLALWEGIEIGHKGLKI
B. nana BNR-RT4            FLDWTLIQMNFPKQWRDWIRSCVTSASASILINGSPTSPIKLRRGLRQGDPLSPFLFALIAEPLNILIKKA-VSLNLWDGVEVCRGGFKI
B. nana BNR-RT5            YLDWVLEQMKFPPQWRNWIQSCVMSASASILINGSPSSIFKLQRGLRQGDPLSPFLFNIAVEPLSLLFHKA-ISMNLWEGVXIRRNGLKI
B. procumbens BNR-RT2      YL-WVLEKMEFPPQWRQWISSCITSASASILINGSPSPTFKLQRGFRQGDPLSPFLFNIAVEPLNLLLQKS-ISLGLWEGVEICKNGSKV
B. procumbens BNR-RT3      VPDWVLEKMEFPPQWRQWISSCITSASASILINGSPSPTFKLQRGLRQGDPLSPFLFNIAVEPLNLLLQKS-ISLGLWEGVEICKNGSKV
B. procumbens BNR-RT5      FLEWTLEQMNFPKQWCTWIRTCVMTTSVSILINGSPTIPIKLRRGLRQGDPLSPFQFDLIAEPLNLLIKKA-VSLKLRDGISVCRNGLKI
B. procumbens BNR-RT6      --DWVLEKMEFPPQWRQWISSCITSASASILINGSPSPTFKLQRGLRQGDPLSPFLFNIAVEPLNLLLQKS-ISLGLWEGVEICKNGSKV
Karma                      SLFQILAVRGFPNNWISWIKSLLQTSKSAVLLNGIPGKWINCKKGLRQGDPLSPYLFILVADVLQRLLEKN-FQIR-----HPIYQDRPC
BvL2                       YIFMTLEKLGFSPIWIGWIKECISSSSFSVLVNGVPGEKFFPSRGIRQGDPISPYLFILCAELLARLLSSAANSPTKSVGMPVGKTGIRV
LIb                        YLEETLEDFGFPRRLIDLILFSLQESSLAILWNGGRLPPFKPGRGLRQGDPLAPYLFNLVMERLAHDIQTRVNARTWKP-VHITRGGTGI
Ta11-1                     YVRALLDALGFHQKWVGWIMFMISSVSYSVLINDKAFGNIVPSRGLRQGDPLSPFLFVLCSEGLT-HLMNRAERQGLLSGIRFSENGPAI
BLIN                       FLFEVLRCHGFGSRFLGWIALLLSSASTKVLLNGNPGPAIWHRRGLHQGDPVSPQLLVLAVDTLGRLFHRA---VELRVMQQLHPRRAIP
cin4                       YLLDVLKALGFTQKWRDWIATILGSSSSKIIINGQQTKEIKHMRGVRQGDPLSPFLFILAMDPLQRMIERA---AHEGLLGQVLPNGAKF
del2                       FLIECLRVKGFNDRWIGWVRSCLGSSRATVLVTGQPRHYFRVRRGVKQRCPLSPLLFILAAGVLGALIRGA-SARGSFPWSSGGGFVWLV
Zepp                       WIERCMAAVGFGPGVQRWVHILHSGTTSRVAFNGWHTDAFPVAAGVFQGSPLSPLLFVLARAPMAAHARML-AGQLAFQPIRLPSGEPAP

                                                                    Domain IV
```

Appendix 4 (ongoing):

```
                                   100       110       120       130       140       150       160       170
                           ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|..
B. vulgaris BNR-RT1        THLQYADDTVLFCPPKLEFPENIKRVLILFHLASGLQINFHKSSLMGIN-IEPKLLDHMASQLLCKVGSLPFVYLGLPIGGS
B. vulgaris BNR-RT2        THLQYADDTVLFCPPKLEFLENIKRVLILFHLASGLQINFHKSSLMGIN-IEPKLLDHMASQLLCKVGSLPFVYLG-AFGGS
B. vulgaris BNR-RT3        SHLQYVDDKIMFFPPNSQFLLNIKFVLILFQLTSGLQVNFHKISIIGLN-VTD-WMHQAANSLLCRTGSLPLSYLGLPIGGN
B. vulgaris BNR-RT5        SHLQYADDTIIFCPPKVEFLCNIKKTLIYHX--ASGLGVNHKSALYGIN-VDELWLS-HADLLLCRTGSLPYTYLGLPMGGN
B. adanensis BNR-RT1       SHLQYADDTILFCPPNTQQLMNVKRMLLLFHQASRLRVNFHKSSLLGIH-IQER-LDIMARFLPLQNCNIPIPYLRLLIGGN
B. adanensis BNR-RT2       SHLQYADDTILFCPPNTQQLMNVKRMLLLFHQASRLRVNFHKSSLLGIH-IQERQLDIMADSLHCKIAIFPFTYLGLLIGGN
B. adanensis BNR-RT3       SHLQYADDTILFCPPNTQQLMNVKRMLLLFHQASRLRVNFHKSSLLGIH-IQERQLDIMADSLHCKIAIFPFTYLGLPIGGN
B. adanensis BNR-RT4       SHLQYADDTILFCPPNTQQLMNVKRMLLLFHQASRLRVNFHKSSLLGIH-IQERQLDIMADSLHCKIAIFPFTYLGLPIGGN
B. macrocarpa BNR-RT1      SHLQYADDTIIFCPPKLEYLQNIKKALVAFQLASGLQTNFHKSSLMGIN-VLESWVKEAATVLHCKIGTLPFSYLGLPIGGS
B. macrocarpa BNR-RT4      --LSYSMRMIRFCSSLHDSISNIKK-LILFELAS-LQVNL--SQLV-LN-VDESYLFELAESLSCKIGSLPFTYIGLPI---
B. macrocarpa BNR-RT3      SHLQYADDTIIFCPPKLEYLQNIKKALVAFQLASGLQTNFHKSSLMGIN-VLESWVKEAATVLHCKICTLPFSYLGLPIGGS
B. macrocarpa BNR-RT5      THLQYADDTILFCPSNIESLLNVKKILILFHLSSGLKVNFHKSSMIGIH-TSEEWVKRAAEAFQCKIGSLPFSYLGLPIGGN
B. patula BNR-RT1          THLQFADDTVIFCPPKMEFLENIKRVLILFHMASGLQINFYKSSLMGIN-IDPSLLDHLASQLLCKVGSLPFTYLGLPIGGN
B. patula BNR-RT2          SHLQYADDTIIFCPPKLEYLQNIKNALVAFQLASGLQTNFHKSSLMGIN-VPESWVKEAATVLHCKIG--------------
B. patula BNR-RT4          THLQFADDTVIFCPPKMEFLENIKRVLILFHMASGLQINFYKSSLMGIN-IDPSLLDHLASQLLCKVGSLPFTYLGLPIGGN
B. patula BNR-RT5          THLQYADXXIIFSTPSLVSLCNIKKTLIAFQVASGLXVNFHKSAIYGIN-VEDSWLQQAAEALLCRXGDLPLKYLGLPIGGN
B. corolliflora BNR-RT6    SHLQYADDTILFCPPDLDSLMNIKKVLILFQLASGLQVNFHKSSIMGIN-VDEAWLQHSSNSLLCKIGSLPFTYLGLPIGGN
B. corolliflora BNR-RT1    SHLQYADDTILFCPPNLGSLLNIKKALILFQLASGLQVNFHKSSIVGIN-VDDSWLHDTSKALLCKIGSFPLTYLGLPIGGN
B. corolliflora BNR-RT4    THLQYADDTVIFCPPKLEFLANIKKTLILFQLASGLQVNFHKSSLXGVN-VNEXQMTAFASHLLCKVGE--FFITWLPIGGN
B. corolliflora BNR-RT5    THLQYADDTVIFCPPKLEFLANIKKTLILFQLASGLQVNFHKSSLIGVN-VNDSQXTAXAX-LLCKVGXFPFXYLGLPXGGN
B. nana BNR-RT1            SHLQYADDTLVFCPPDFHSLLNIKKVLILLQLSSGLQVNFHKSSIIGLN-VDNPWLQQVANILWCKIGSLPFSCLGLPIGGN
B. nana BNR-RT3            SHLQYADDTIIFCPPKLEYLQNIKKALVAFQLASGLQTNFHKSSLMGIN-VLESWVKEAATVLHCKIGTLPFSYLGLPIGGS
B. nana BNR-RT4            SHLQYADDTVIFCPPKLEFLQNIKKTLILFNLASGLQINFHKSSLMGIN-IDS-LLDHLASQLLCKG-KLPFTYPGAPDGGV
B. nana BNR-RT5            SHLQYADDTLVFCPPDFHSLLNIKKVLILFQLSSGLQVNFHKSSIIGLN-VDNPWLQQVAXILWCKIGSLPFSCLGLPIGGN
B. procumbens BNR-RT2      SHLQYADDTLIFCPPNIRSLANIKMTLILFQLSSGLQINFHKSSIYGIN-VDSSWVQDAANGLLCKMDTLPFTYLGLPIGGN
B. procumbens BNR-RT3      SHLQYADDTLIFCPPNIRSLANIKMTLILFQLSSGLQINFHKSSIYGIN-VDSSWVQDAANGLLCKMDTLPFTYLGLPIGGN
B. procumbens BNR-RT5      THLQYAEDAIIFCPPNFESLANIKKILILFQLASRLQANFHKSSLLGIN-VDDSQLGSLATHLLCRTDTLPITYFGLPIGGN
B. procumbens BNR-RT6      SHLQYADDTLIFCPPNIRSLANIKMTLILFQLSSGLQINFHKSSIYGIN-VDSSWVQDAANGLLCKMDTLPFTYLGLPIGGN
Karma                      ATIQYADDTLVICRAEEDDVLALRSTLLQFSKATGLQINFAKSTMISLH-IDRSKESSLSELLQCKLESLPMSYLGLPLSLH
BvL2                       PFLTFADDTMIIAKANNYSCLVIRQILDKYFSMSG-MVNYHKSAFQCTGNVSANEKQDFANILGMTESNSLGDYLGCPIITS
LIb                        SHLFFADDLMLFGEASEHQAQIMFDCLDSFSNASGLKVNFSKSLLFCSSNVNAGLKRAIGSILQVPVAESLGTYLGIPMLKE
Tal1-1                     HHLLFADDSLFMCKAVKEEVTVIKSIFKVYGDVTGQRINYDKSSITLGALVDEDCKVWIQAELGITNEGGASTYLGLPECFS
BLIN                       AISLYADDVILLCHPSPDDIAAVKEILQLFGRASGLHVNFQKSAAALIR-CDTEDAAHIVAHLGCPIVDFPLTYLGIPLKLR
cin4                       RCSLYADDAGVFVRADKLDLKVLKRILEAFEWCSGLKINFEKTEIFPIR-YPESLWSNLMEVFPGKYSNFPGKYLGLPLHFR
del2                       SGFQFADDTLIFSDHPQEAENLLSSCLVLAASVLELNCAKTQTLLLGTI----EERAADLASLRCSRGTLPMTYLGISQ---
Zepp                       VMHQHADDTSVHARTPGMLRSCWGPSVGLHCAATGARLQRSKSQALGLAASAISPGPIQSRGVTFAASSDGVKHLGIPLSTQ

                           Domain V               Domain VI                          Domain VII
```

Appendix 5: Content of the supplemental CD-Rom.

Folder	**Content**
Dissertation	PDF and DOCX file
***HMMER*3-related files**	HMMER workshop HMMs used in this thesis Sequence alignments used to produce the HMMs
Publications	PDF files of related publications
Script programs	Script programs listed in Table 2.11
Sequence data	Sequence data that provided the base of Figure 3.24, Figure 3.25, Figure 3.26 and Figure 3.37 DOC files with annotations of the LINEs presented in Table 3.1 and Table 3.7

10 Curriculum Vitae

The CV has been removed for the online version of this thesis.

11 Peer-reviewed publications that cover data presented in this thesis

For comprehensiveness, the online version of this thesis also lists publications derived from this work which have been published after thesis submission and defense.

Chapter 3.1: Detection and characterization of the highly abundant retrotransposon family Cotzilla

B. Weber, T. Wenke, U. Frömmel, T. Schmidt and **T. Heitkam** (2010): "The Ty1-*copia* families SALIRE and Cotzilla populating the *Beta vulgaris* genome show remarkable differences in abundance, chromosomal distribution and age." *Chromosome Research 18 (2)*, 247-263

> Contribution: I performed a major part of the bioinformatic analysis, designed and coordinated the Cotzilla-related experiments, analyzed and interpretated the results, created the figures, and wrote the manuscript for publication.

Chapter 3.2: The BNR LINE family defines a novel subclade of L1 LINEs

T. Heitkam and T. Schmidt (2009): "BNR – a LINE family from *Beta vulgaris* – contains a RRM domain in open reading frame 1 and defines a L1 subclade present in diverse plant genomes." *The Plant Journal* 59 (6), 872-882

Contribution: I performed the experiments and bioinformatic analysis, analyzed and interpreted the results, created the figures, and wrote the manuscript for publication.

Chapter 3.3: Plant retrotransposon analysis on a genomic scale

T. Heitkam, D. Holtgräwe, J. C. Dohm, A. E. Minoche, H. Himmelbauer, B. Weisshaar and T. Schmidt (2014): "Profiling of extensively diversified plant LINEs reveals distinct plant-specific subclades", *The Plant Journal* 79 (3), 385-397

> Contribution: I contributed to research design, performed the experiments and bioinformatic analysis, analyzed and interpreted the results, created the figures, and wrote the manuscript for publication.

B. Weber, **T. Heitkam**, D. Holtgräwe, B. Weisshaar, A. E. Minoche, J. C. Dohm, H. Himmelbauer and T. Schmidt (2013): "Highly diverse chromoviruses of *Beta vulgaris* are classified by chromodomains and chromosomal integration patterns", *Mobile DNA* 4, 8

> Contribution: HMM-based identification of Chromovirus-type Ty3-*gypsy* retrotransposons from the sugar beet genome and analysis of their reverse transcriptases. Generation of Figure 2. Contributed to manuscript writing and editing.

C. Wollrab, **T. Heitkam**, D. Holtgräwe, B. Weisshaar, A. E. Minoche, J. C. Dohm, H. Himmelbauer and T. Schmidt (2012): "Evolutionary reshuffling in the Errantivirus lineage Elbe within the *Beta vulgaris* genome." *The Plant Journal* 72 (4), 636-251

> Contribution: HMM-based identification of Errantivirus-type Ty3-*gypsy* retrotransposons from the sugar beet genome and analysis of their reverse transcriptases. Contributed to manuscript writing and editing.

Die vorliegende Arbeit wurde am Institut für Botanik am Lehrstuhl für Zell- und Molekularbiologie der Pflanzen der Technischen Universität Dresden unter der Betreuung von Prof. Dr. Thomas Schmidt angefertigt.

Versicherung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Dresden, den 27. Mai 2011 Tony Heitkam