GENOME ANNOTATION

OF FRITILLARIA AGRESTIS BAC CLONE

A University Thesis Presented to the Faculty

of

California State University, East Bay

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Biological Science

By

Rajhalutshimi Narayanaswamy

September 2015

GENOME ANNOTATION

OF FRITILLARIA AGRESTIS BAC CLONE

By

Rajhalutshimi Narayanaswamy

Approved:

Dr.Chris Baysdorfer — Date: July 26, 2015

Dr.Claudia Udhe-Stone — 2015 Aug 26

Dr.Kenneth Curr — August 26, 2015

Acknowledgments

I would like to thank my advisors Dr.Chris Baysdorfer, Dr.Claudia Udhe-Stone and Dr.Kenneth Curr for their support, guidance and encouragement that helped me to complete this project successfully.

I would also like to thank my family members who have been supportive during the course of my project.

Table of Contents

List of Figures

1 Introduction

Genome size is the amount of DNA in the unreplicated gametic nucleus (also known as 1cx value) (Leitch et al., 2009). Genome size is measured either by its mass (in picograms, pg) or by the number of base pairs (bp) (Gregory, 2005). The disparity between the genome size and organism complexity or the gene number called the *c-value* paradox was resolved with the discovery of non-coding DNA in 1970 (Gregory, 2005). The resolution of the paradox commenced the journey in answering several questions that arose with the answer, which was termed c-value enigma by Gregory (Gregory, 2005). The c-value enigma aimed at answering important questions: The nature of the non-coding DNA, its evolution, and the process involved in maintaining it, the non random distribution of genome size variation among various taxonomic groups, the relationship between c-value and cell sizes, and cell division rates. The most prominent among these is the nature of the non-coding DNA. One discipline to learn about the non-coding DNA, is complete genome sequencing, followed by annotation (Gregory, 2005).

1.1 Genome size variation

Genome size currently varies 66,000 fold across eukaryotes, among the 10,000 species studied. Angiosperms stand out as one of the most variable group with genome sizes, varying 2400 fold (Pellicer, Fay, & Leitch, 2010). Species with large genomes are largely restricted to monocots, most notably in *Alliaceae, Asparagaceae, Liliaceae, Melanthiaceae* and *Orchidaceae* (Leitch et al., 2009). The smallest angiosperm genome so far reported is in *Genlisea margaretae* with 0.0648pg. Until recently, *Fritillaria assyriaca* with a genome size of 127.4 pg and belonging to *Liliaceae* Family was considered to have the largest size genome. Recently, Pellicer and Leitch et al have reported *Paris japonica* from the Family *Melanthiaceae* to have the largest genome size of 152.23 pg (Pellicer et al., 2010).

1.2 Brief outline of the mechanisms of genome size variation

Several mechanisms are involved in the expansion and shrinkage of genome size. Expansions are caused by polyploidy, transposable elements, intron proliferation, segmental duplication and small scale duplications. Decrease in genome size is known to result from non-homologous recombination, DNA replication errors and chromosome loss (Hawkins, Grover, & Wendel, 2008; Whitney et al., 2010). Although polyploidy and transposable elements are the main source of genome size variation, polyploidy only contributes to increase in c-value and not the size of the genome (Bennett & Leitch, 2005). Moreover, molecular studies in plants have shown, mobile elements contribute to

the majority of the genome in some large genome plants (Jeffrey L. Bennetzen, 2000, 2002; R. B. Flavell, Bennett, Smith, & Smith, 1974). Plants are primarily composed of LTR retrotransposons (Jeffrey L. Bennetzen, Ma, & Devos, 2005; P. SanMiguel et al., 1996). More than 60% of plant genome is comprised of transposable elements (Jeffrey L. Bennetzen, 2002; Gregory, 2005). In maize more than 70% of the plant genome is comprised of LTR-retrotransposons (Jeffrey L Bennetzen, 2005; P. SanMiguel et al., 1996).

1.3 Repetitious DNA

There are two main types of repetitious DNA. They are simple-sequence DNA and interspersed repeats. The simple-sequence DNA is composed of short sequences repeated tandemly. When the repeating unit is composed of 1-4 units or at the maximum 6 bp, then it is called microsatellites. In microsatellites, dinucleotide repeats dominate followed by mono and tetranucleotide repeats, trinucleotide repeats are less dominant compared to the other types. The number of iterations of these units can vary. Repeats with longer repeated units are called minisatellites, in extreme cases, where the repeat unit is longer than 30bp, it is called satellite DNA. The majority of these repeats are embedded in non-coding DNA, either in introns or intergenic sequences (Ellegren, 2004).

There is a positive correlation between microsatellites and genome size, especially in mammals. But, in plants are negatively correlated. This can be explained by the fact that plant genomes are involved in genome expansion and thereby their repetitive DNA is highly represented by long terminal repeats of transposable elements (Ellegren, 2004).

Transposable elements

Transposable elements (TE) have two basic characteristics. First, is to move from place to place in the genome, hence termed mobile DNA. Second, is to amplify their copy number within the genome. By their activity they bring about changes in the structure of the genome and, or activity of the gene. Because of this nature Barbara McClintock who discovered TEs called them "controlling elements" (Jeffrey L. Bennetzen, 2000).

TEs are classified by their modes of transposition via an RNA or DNA intermediate. DNA transposons or class II elements although found in all organisms are a major class of transposable elements in prokaryotes. They range in size from a few hundred bases to 10 kb (Jeffrey L. Bennetzen, 2000). The genomes that harbor the highest density and diversity of DNA transposons are rice, nematodes and humans (Feschotte & Pritham, 2007). DNA transposons move by using single or double stranded DNA (Feschotte & Pritham, 2007). DNA transposons are divided into three main subclasses (i) the "cut and paste" transposons, (ii) Helitrons, that use rolling circle replication; (iii) and Mavericks which use use self-encoded DNA polymerase (Feschotte

& Pritham, 2007). A family of DNA transposons is characterized by the TIR sequences that they share. The TIRs range in size from 11bp (Ac/Ds) Activator/Dissociator to a few hundred bases (Mutator) (Jeffrey L. Bennetzen, 2000). All DNA transposons have terminal inverted repeat (TIR) except for Helitrons and a few "cut and paste" transposons. There are ten superfamilies of cut and paste transposons and they are characterized by the sequence similarity of the transposase encoded by the autonomous copies and the TIRs. Mavericks also known as Polintons are very large transposons that can code for multiple proteins (Feschotte & Pritham, 2007).

MITEs (Miniature inverted terminal repeats) are short transposons of 100-600 bp in length. They are characterized by the absence of autonomous element and are distinguished from other non-autonomous elements by their high copy number and length homogeneity. Again, rice, nematodes and human genomes are filled with the largest copies of MITEs (Feschotte & Pritham, 2007). It is likely that MITEs employ *trans*-acting transposition function from the host machinery (Jeffrey L. Bennetzen, 2000).

Class I elements are retroelements, which are comprised of retrotransposons and retroviruses. Retrotransposons are abundant in eukaryotes, particularly in plant genomes (Jeffrey L. Bennetzen, 2000; Kumar & Bennetzen, 1999; Staginnus, Desel, Schmidt, & Kahl, 2010). While, retroviruses are exclusively found in vertebrates (A. J. Flavell & Smith, 1992) . Retrotransposons are believed to be progenitors of retroviruses or on the contrary as descendants of retroviruses by the loss of their envelope (env) gene (Eickbush & Jamburuthugoda, 2008). They contribute to more than 70% of DNA in maize (Phillip SanMiguel & Bennetzen, 1998; Staginnus et al., 2010). Despite their abundance in the

plant kingdom only a few elements are active (Grandbastien, 1992). All class I elements transpose through reverse transcription of an RNA intermediate and integration of the resulting cDNA into another location. During transposition the sequence information flows from RNA to DNA to RNA. These elements do not excise, instead leave a copy that inserts elsewhere (Jeffrey L. Bennetzen, 2000). Retrotransposons are characterized by the presence of long terminal repeats (LTR). Retrotransposons are classified based on the presence or absence of LTR into LTR retrotransposons and non-LTR retrotransposons. The non-LTR retrotransposons are further divided into long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). SINEs are non-autonomous elements in the non-LTR retrotransposons. They range in size from 100-300bp. They are abundant in human with *Alu* sequences. They do not encode any known peptides (Jeffrey L. Bennetzen, 2000). Most SINEs possess promoters and motifs for RNA polymerase III. Most of them are derivatives of tRNA or snRNA (Jeffrey L. Bennetzen, 2000; Hitoshi Nakayashiki, 2011).

There are two sub groups of autonomous non-LTR retrotransposons based on the phylogenetic analysis of their reverse transcriptase sequences. They are LINEs and Penelope-like elements (PLEs). PLEs have short terminal repeats when compared to LINEs and sometimes retain introns. PLEs have been identified in protists, plants and fungi (Hitoshi Nakayashiki, 2011).

Autonomous non-LTR retrotransposons contain one or two open reading frame (ORFs) (Han, 2010). They encode for reverse transcriptase and endonuclease. ORF1 contains RNA binding activity and nucleic acid chaperone activity and are likely to be

similar to Gag proteins. A short internal promoter is present within the 5' untranslated region (UTR), but the sequences are not well conserved (Han, 2010; Hitoshi Nakayashiki, 2011). There are specific sequences within the 3' UTR that are recognized by ORF2 (Han, 2010). Most non-LTR retrotransposons are truncated at the 5' end, due to incomplete reverse transcription. Only a small percentage of them are full length and active. For example, out of 500,000 L1NEs in human genome only 80-100 are full length and active (Han, 2010).

LTR retrotransposons as the name implies are defined by the presence of long terminal repeats at the 3' and 5' ends of the elements. LTR retrotransposons evolved from a late-branching lineage of non-LTR retrotransposons as found from the phylogeny study of the RNase H domain of these elements (Malik & Eickbush, 2001). The LTR retrotransposons were first identified and characterized in *Saccharomyces cerevisiae* and *Drosophila melanogaster* (Boeke, Garfinkel, Styles, & Fink, 1985; Cameron, Loh, & Davis, 1979; Clare & Farabaugh, 1985; Mount & Rubin, 1985; Rubin, 1983). LTR retrotransposons are common in invertebrate eukaryotes, especially in higher plant kingdom (A. J. Flavell, 1992; Kumar & Bennetzen, 1999). LTR retrotransposons in plants range in size from 2kb to 18kb with LTRs varying in size from a few hundred bases to several kilobases (Kumar & Bennetzen, 1999; Vitte & Panaud, 2005). For example, Ty1 element of *Saccharomyces cerevsiae* is 5.6 kb long with non-inverted repeats of 0.25 kb on either ends (Cameron et al., 1979).

Retrotransposons are classified into vertebrate retroviruses (*Retroviridae*), Ty1/Copia (*Pseudoviridae*), Ty3/Gypsy (*Metaviridae*), *hepadenoviruses, caulimoviruses,*

BEL and DIRS1 groups (Malik & Eickbush, 2001) (Gorinsek, Gubensek, & Kordis, 2004). The two main superfamilies of LTR retrotransposons are Ty1/Copia (*Pseudoviridae*) and Ty3/Gypsy (*Metaviridae*). This classification is based on the phylogeny of the RT domains (Eickbush & Jamburuthugoda, 2008).

Ty are elements from *Saccharomyces cerevisiae*. The Ty family consists of approximately 30 transposable elements (Clare & Farabaugh, 1985); of which Ty1 consists of 5.6 kb length sequence inclusive of a 0.25kb direct repeats (Cameron et al., 1979). Whereas, Ty3 is composed of 5.2 kb which includes a 4.7 kb of central coding region flanked by 0.34kb of direct repeats (Clark, Bilanchone, Haywood, Dildine, & Sandmeyer, 1988), while Copia, Bel and Gypsy are elements of *Drosophila melanogaster*. There are a total of 11 Ty1-Copia group identified so far (A. J. Flavell, 1992).

The Ty3/gypsy and Ty1/Copia elements mainly differ in the order in which they package the genes in the central coding region. These two families differ in the placement of integrase gene in pol ORF. The domains of Copia elements are arranged as LTR-Gag-Ap-In-Rt-Rh-LTR (fig. 1), whereas Gypsy elements are organized as LTR-Gag-Ap-Rt-Rh-In-LTR (J. A. Tanskanen, F. Sabot, C. Vicient, & A. H. Schulman, 2007). Both these groups are present in high copy numbers in plants with large genomes (Kumar & Bennetzen, 1999). Copia and Gypsy elements are also present in animals and fungi, while the Bel clade is found only in animals (Eickbush & Jamburuthugoda, 2008; J. Tanskanen, F. Sabot, C. Vicient, & A. Schulman, 2007). Moreover, the components of

each are more similar in sequence to the same superfamily in other organisms than to elements of the other superfamily in the same organism.(J. A. Tanskanen et al., 2007).

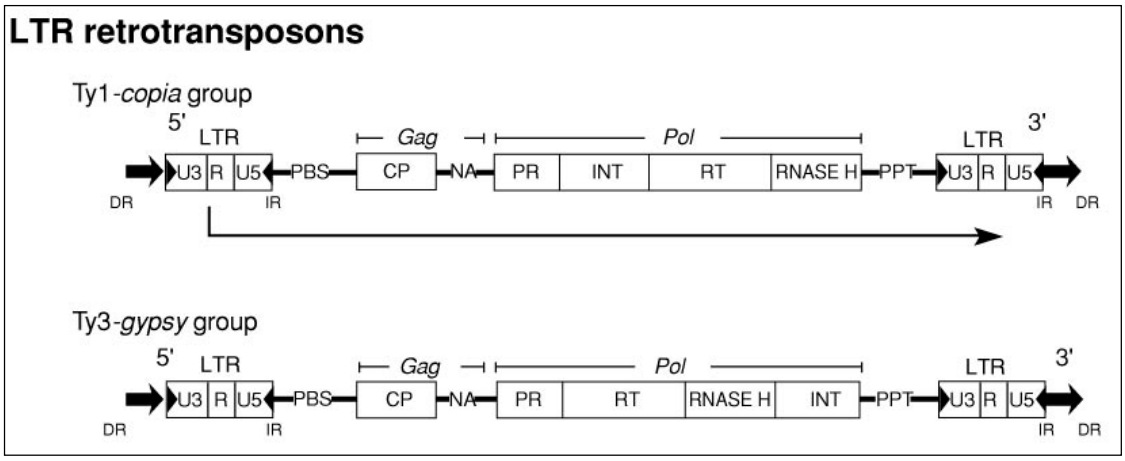

Fig. 1 Basic structural features of LTR retrotransposon.

Sequences within LTRs are unique 3' RNA (U3), repeated RNA (R), unique 5' RNA (U5), primer binding site (PBS), polypurine tract (PPT). The genes within the retrotransposons encode: capside-like proteins (CP), integrase (INT), protease (PR), reverse transcriptase (RT), and ribonuclease (RNase H). Other sequences include DR (flanking target site direct repeat), IR (inverted terminal repeats), NA (nucleic acid binding moiety) (Kumar & Bennetzen, 1999).

The LTR retrotransposons are usually terminated by the dinucleotides 5'-TG….CA-3' (Kumar & Bennetzen, 1999; Ma, Devos, & Bennetzen, 2004). The LTR is functionally divided into three regions U3, R and U5. Transcription proceeds from the

U3/R boundary in the left LTR to R/U5 in the right LTR and produces an mRNA transcript which has the R region repeated at both ends (Matthews, Goodwin, Butler, Berryman, & Poulter, 1997). The mRNA molecule of LTR retrotransposon has the structure 5'-R-U5- PBS-coding region - PPT- U3 – R-3', where R, U5, PBS, PPT and U3 stand for repeated RNA, unique 5' RNA, primer binding site, polypurine tract (PPT) and unique 3'RNA (Kumar & Bennetzen, 1999). The LTRs contain regulatory sequences that are recognized by the host cell transcription machinery. Although, both the LTRs have identical sequences, they vary in their functions. Transcription initiation takes place in the 5'-LTR, while cleavage and polyadenylation/ transcription termination signals is driven by sequences in the 3'-LTR (Klaver & Berkhout, 1994). Most of the eukaryotic genes, including HIV-1, carry the promoter sequence for RNA polymerase II with a TATA box consensus sequence TATAAAA. In HIV-1 the TATA box is located 28 bp upstream of the transcription start site in the U3 region (van Opijnen, Kamoschinski, Jeeninga, & Berkhout, 2004).

These elements share a basic structure of a few hundred base pairs of direct repeats flanking a central coding region. The central coding region is made up of open reading frames (ORF) which codes for group specific antigen (gag) and pol (Havecker, Gao, & Voytas, 2004; Staginnus et al., 2010). Most plant retrotransposons encode gag and pol in a single ORF (Gao, Havecker, Baranov, Atkins, & Voytas, 2003). It is classically found in Ty1/Copia and BEL clades (Gao et al., 2003). The gag and the pol in some LTRs are separated by a stop codon or a frame shift (Havecker et al., 2004). Although rare, gag and pol separated by stop codon are found in the Ty3/Gypsy and

RIRE2 elements in rice (Ohtsubo, Kumekawa, & Ohtsubo, 1999) and BEL element in kamikaze from *Bombyx mori* (Abe et al., 2001). Only a few retroelements have pol in the +1-frame relative to gag. In plants only two elements have been identified with a +1-frameshift. Both the elements are in Arabidopsis (AtChr2_44644 and AtChr2_4188838) (Gao et al., 2003). Furthermore, ORFs in some elements can be overlapping (Staginnus et al., 2010) and multiple ORFs can be found in some retroelements, mainly plant retrotransposons as a result of mutations. One such example can be found in maize (U68408) with three large ORFs (Gao et al., 2003).

The gag ORF in retrotransposons encodes the structural components, the capsid (CA) and the nucleocapsid (NC) proteins that form the virus like particle (VLP), inside where transcription takes place. Additionally, retroviral gag also encodes the matrix (MA). Most retroelements, which include both retrotransposons and retroviruses, have functionally similar capsid and nucleic acid binding regions. Additionally some retroelements contain short spacers (SP), which aids in particle assembly but not code for mature protein. For example, in alpharetroviruses (RSV) mutations in SP caused formation of budding tubules instead of spherical viruses, moreover mutants completely lacking SP in RSV are noninfectious (Clemens et al., 2011). The gag is least conserved due to the fast rate of evolution (Doolittle, Feng, Johnson, & McClure, 1989). Studies on gag sequences from the Gypsy database has revealed that gag showed sequence similarity to its own lineage counterparts than to other gag sequences. In other words, sequence homology of gag is not found in distantly related retroelements. This was demonstrated in Ty3/Gypsy (Llorens, Futami, Bezemer, & Moya, 2008). With a few exceptions,

retroviruses and retrotransposon gags are characterized by the presence of one or two zinc-finger motifs, which has the amino acid sequence C-X2-C-X4-H-X4-C in the NC protein. Moreover, retrovirus Gags carry one or more additional domains. A similar extra domain has also been reported in pyret, a Ty3/Gypsy element between NC and protease in *Magnaporthe grisea*, characterized by the WCCH motif (H. Nakayashiki et al., 2001). In addition, the capsid (CA) region carries a stretch of ~ 20 amino acids termed the major homology region, which is conserved in most retroviruses except spuma-retroviruses (Craven, Leure-duPree, Weldon, & Wills, 1995; H. Nakayashiki et al., 2001; Patarca & Haseltine, 1985). Sequence alignment of retrovirus CA proteins yielded the consensus sequence (H)XQGX2E(S)X3FX2RLX2(SH), where H indicates a hydrophobic residue and S indicates P,S or T. A major homology region (MHR) like motif has also been identified in *Saccharomyces cerevisiae* TY3 in its CA domain in the form of QGX2xex(S)X3FX3LX3(H) (Orlinsky, Gu, Hoyt, Sandmeyer, & Menees, 1996). Later, a similar domain has also been identified in chromoviruses (H. Nakayashiki et al., 2001). Chromoviruses are chromodomain containing retrotransposons which is the only *Metaviridae* (Ty3/Gypsy) that has wide distribution in Eukaryotes (Kordiš, 2005). However, this domain in chromoviruses are reported to be significantly different from the MHR domain of retroviruses by having some sequence variation and a gap in the chromoviruses (H. Nakayashiki et al., 2001). It has been noted that this domain is not universally conserved in chromodomains. For example, several members of chromoviruses like grasshopper and tf1, both of which are retroelements do not have this

motif (H. Nakayashiki et al., 2001). This motif is not identified in Ty1/Copia (Clare & Farabaugh, 1985; Orlinsky et al., 1996).

In Sireviruses, a plant specific lineage of Ty1/Copia, Gag lengths ranged from about 548 to 961 amino acids. The Gag proteins were found to be more conserved in their N-terminal halves, encoding a central CCHC zinc-knuckle. In addition, some Sireviruses have a second CCHC zinc knuckle at the C-terminus. With the exception of rice element, Osr7 and Osr8, all of the Sireviruses contained a gag extension downstream of the central zinc-knuckle, which was responsible for the variations in the length of the gag ORF (Havecker, Gao, & Voytas, 2005).

The pol gene encodes for several enzymatic functions. They are the protease, reverse transcriptase, integrase and RNase H. The catalytic active sites of these enzymes share a high degree of sequence conservation in both retrotransposons and retroviruses (Moore & Garfinkel, 2009).

The protease (PR) processes the Gag protein precursor to yield mature Gag proteins and cleave the Gag-Pol protein precursor by post-translational cleavage. Deletion of protease in *Saccharomyces cerevisiae* resulted in unprocessed gag-pol precursor polyprotein and also altered the morphology of VLP (Youngren, Boeke, Sanders, & Garfinkel, 1988). Proteases of retrotransposons are less extensively characterized compared to that of retroviruses. Retroviral protease family has sequence similarity to the active site of aspartic proteases. The sequence Asp-Thr/Ser-Gly, which is conserved in the active site of aspartic protease is also conserved in retroviruses, but the viral protease correspond to a single domain of the aspartic protease (Pearl & Taylor,

1987), (Toh, Ono, Saigo, & Miyata, 1985). Most retrotransposons such as Ty1/Copia, Ty2 and Ty3/Gypsy have Asp-Ser-Gly at their protease active site, except the gypsy-like Drosophila element 17.6 which carries the sequence Asp-Thr-gly at the predicted active site (Boeke & Sandmeyer, 1991). The similarity of substrate specificity between retrotransposons and retroviruses is not clear because retrotransposon protease processing sites have not been characterized yet (Kirchner & Sandmeyer, 1993) . In all the retrotransposon cases studied so far, the substrate has been element encoded Gag and Gag-Pol protein precursor. The products have been capsid (CA) and nucleocapsid (NC) from Gag and protease, reverse transcripase and integrase from Gag-Pol region (Barrett, Woessner, & Rawlings, 2004).

Reverse transcriptase is the key enzyme in retroelement replication. This enzyme exhibits both RNA-and DNA-dependent polymerase activity, as well as RNase H activity as this enzyme is found adjunct to the reverse transcriptase (Nanni, Ding, Jacobo-Molina, Hughes, & Arnold, 1993). This enzyme is highly conserved (Doolittle et al., 1989). For example in retroviruses due to the high level of sequence conservation of this region in Pol , the RT region was used to study phylogenetic relationship among retroviruses (Xiong & Eickbush, 1988) . Most studies on reverse transcriptase of LTR retrotransposons involve direct comparison of this enzyme with that of retroviral RTs, needless to say this domain has been extensively studied in retroviruses. Next to retroviruses, this domain has been studied well in TY1 and Ty3 of Saccharomyces (Eickbush & Jamburuthugoda, 2008). HIV-1 RT is composed of two subunits. The p66 and the p51 subunits. The p66 subunit is composed of a polymerase domain, a connection

and the RNase H domains. P51 is formed by proteolytic processing of p66 and thereby removing RNase H domain from p66 domain (Klumpp & Mirzadegan, 2006). RT domain of LTR retrotransposons are severely truncated when compared to other reverse transcriptases (Malik & Eickbush, 2001). Sequence comparison of all RTs revealed three conserved aspartic acid residues in the active site. Two of the three aspartic acid residues are present in the conserved YXDD motif. The third residue is found about 75-100 amino acids N- terminal to this motif (Eickbush & Jamburuthugoda, 2008). Mutational studies of these three residues in HIV-1 resulted in loss of RT activity both *in vitro* and *in vivo* studies (Kaushik et al., 1996). The overall sequence similarity between different RTs of different classes of retroelements (retroviruses vs. transposable elements) as well as among different elements of the same class is low. For example, the amino acid sequence similarity between mammalian retrovirus MuLV and Visna virus is only 25%. But, there is a high level of sequence similarity in the conserved regions of RT – like sequences in all elements except Copia and Ty. Xiong and Eickbush identified 7 distinct regions in all elements (Xiong & Eickbush, 1988). Each of this region carries a series of conserved amino acids. Copia and Ty have the least sequence similarity compared to other elements but have relatively higher similarity amongst each other. Among the seven regions identified Copia and Ty had unambiguous sequence similarity only in regions/boxes 3, 5 and 7 (Xiong & Eickbush, 1988). Each Ty1-copia RT family in mungbean was more closely related to representative elements in other plant species than to other families of mungbean (Dixit, Ma, Yu, Cho, & Park, 2006)

RNase H is an enzyme, which has the dual responsibility of degrading the template RNA strand, as well as releasing the PPT, as the PPT remains resistant to the enzymatic activity of RNase H. Also, removal of the primer RNA from the newly synthesized cDNA is a function of this enzyme. In non-LTR retrotransposons and most LTR retrotransposons the RNase H domain is found adjunct to the reverse transcriptase domain. Whereas in vertebrate retroviruses both these domains are connected by a linker domain (Malik, 2005). RNase H can be broadly classified into type 1 and type 11 enzymes, based on the difference in their amino acid sequences. Most organisms carry at least one type (Cerritelli & Crouch, 2009). Retroelements carry RNase H1. Based on composition the most conserved region of RNase H of all eukaryotes are the N and C termini, which is separated by a variable connecting domain in the middle. There are four key catalytic residues found within the RNase H active site. They are Asp443, Glu478, Asp498, Asp549 (DDED) (Sharon J Schultz & James J Champoux, 2008). These residues are conserved in all RNase H domains. The position of the residues reported here are the ones found in HIV-1 (S. J. Schultz & J. J. Champoux, 2008). LTR retrotransposons carry a weaker RNase H that lacks a histidine residue in the catalytic site. This reduced activity of the enzyme helps in retaining the PPT of LTR retrotransposons without digesting them. The necessity of this function has probably rendered this enzyme evolve in a conservative fashion in retrotransposons. Host cell RNase H activity is restricted to the nucleus and organelles. As most of the life cycle of LTR retrotransposons takes place in the cytoplasm or virus-like particle, the retrotransposons do not have access to the host RNase H and hence must possess their own RNase H to carry out the replication. This too

is another reason for the high conservation of the RNase H domain in these elements (Malik, 2005; Malik & Eickbush, 2001).

Integrase, integrates the double stranded cDNA into the host genome after reverse transcription. The integrase region varies in length from 280-480 residues in length. Integrase is divided into three main functionally distinct regions. They are the N-terminal domain, the core catalytic region and the C-terminal domain (Peterson-Burch & Voytas, 2002). The N-terminal domain is characterized by the sequence HHCC which is conserved in retrotransposons and retrovirus. This region binds a single zinc ion and provides stability to the N-terminal region of integrase (Moore & Garfinkel, 2009). This zinc binding domain (ZBD) is very similar to CCHH zinc finger motif found in dna binding proteins (Asante-Appiah & Skalka, 1997).

The central core catalytic domain is characterized by the presence of three amino acid residues Asp-64, Asp-116 and Glu-152 which form the D,D (35) E motif (ALAN Engelman & Craigie, 1992). This domain is also conserved in retrotransposons and retroviruses. Interestingly, these three residues are also conserved in certain bacterial transposases. Both integrase and transposase perform the same function of insertion of the genetic material into the host genome (Polard & Chandler, 1995) (A Engelman, Englund, Orenstein, Martin, & Craigie, 1995) . This domain binds metal cofactors such as Mn2+ or MG2+ (Asante-Appiah & Skalka, 1997). Core domain mutants in HIV exhibited abolished 3'processing, DNA strand transfer and disintegration activity (A Engelman et al., 1995).

The C- terminal domain is the least conserved of the three domains showing great variation in length and sequence (Asante-Appiah & Skalka, 1997; Malik & Eickbush, 1999). This domain is important for non-specific binding, and therefore is thought to interact with target DNA while strand transfer. This subdomain is well conserved among several members of TY3/GYPSY and certain vertebrate retroviruses. The conserved region of this module is loosely identified to be G-(D/E)-X $_{10\text{-}20}$ –K-L-X2-(R/K)-(F/Y/W)-X-G-P-(F/Y)-X-(I/V), where X refers to any amino acid and the other ltrs refer to the single letter amino acid code. This module is commonly referred to as GPY/F for brevity and to distinguish the best conserved residues (Malik & Eickbush, 1999). Surprisingly, this module is not universal in either TY3/GYPSY or the retrovirues. Phylogenetic studies of TY3/GYPSY and its related classes of LTR retrotransposons along with TY1/copia and vertebrate retroviruses revealed that this GPY/F module is differentially retained in both TY3/GYPSY and retroviruses (Malik & Eickbush, 1999). To add more to the differentiation this module is completely absent in TY1/copia, but carry a GKGY motif ~60 residues downstream of the catalytic core unit (Peterson-Burch & Voytas, 2002). It has been noted that the clades that lost this module acquired another ORF that is env (envelope) – like (Malik & Eickbush, 1999).

The envelope (env) ORF facilitates intercellular infectivity, it is believed to differentiate retrovirus from retrotransposons structurally (Havecker et al., 2004; Hitoshi Nakayashiki, 2011). It is usually found upstream of 3'LTR (Kumar 1999). This distinction is somewhat fluid because some gypsy elements encode a functional env-like protein which in some conditions also exhibit infectious properties, but did not cluster

along with retroviruses during phylogenetic studies, instead clustered with non-env elements (Hitoshi Nakayashiki, 2011).

There is one other domain called the chromo domain, the chromatin organization modifier domain. The Chromo domain mediates interaction between different proteins (Malik & Eickbush, 1999; Platero, Hartnett, & Eissenberg, 1995), RNA and DNA (Novikova, 2009). Chromo domain was originally identified in Drosophila melanogaster as a conserved sequence between two proteins, the polycomb and heterochromatin protein-1 (HP1). Later this domain was identified in various proteins involved in chromatin remodeling and regulation of gene expression. Further, this same domain is found adjunct to the C-terminus of integrase in some members of the Ty3/Gypsy clade (Novikova, 2009). These Gypsy elements were classified as chromoviruses. Chromoviruses are one of the three genera of Ty3/gypsy ltr retrotransposons. They are widely distributed in the plant kingdom with high copy number. It is usually 40-50 amino acids long (Kordis, 2005). The chromo domain is classified into 'classical' and 'shadow' chromo domains. The classical chromo domain carries three distinct conserved residues Y34, W45, and Y48, which together form the aromatic pocket. The shadow chromo domain is characterized by the absence of the first and usually the third residues that are conserved in the classical group. Plant retrotransposons carry a chromo domain, which resemble the shadow chromo domain (Novikova, 2009). Example of plant chromoviruses are *Arabidobsis* (Legolas), Oryza (RIRE3) and *Lilium henryi* (del1-46) (Kordis, 2005) .

During the insertion of the retrotransposon into the host genome a staggered cut is created in the target site. In the case of TY1 there is a 5bp staggered cut (Kenna,

Brachmann, Devine, & Boeke, 1998). After the integration, a small gap remains that is filled by DNA polymerase and host gap repair machinery, which forms the characteristic target site duplication (Lin, Nymark-McMahon, Yieh, & Sandmeyer, 2001). The length of the direct repeat generated at the site of transposition varies form 4-6 bp in length within the LTRs (Wicker et al., 2007). The number of base pairs generated is specific to the transposable element. For example Ty1/Copia generate 5bp of direct repeat (Dunsmuir, Brorein Jr, Simon, & Rubin, 1980) .

Polypurine tract (PPT) is located immediately upstream of 3'LTR. PPT functions as RNA primers for plus-strand cDNA synthesis. PPT is 13-18 nucleotides long (Noad, Al-Kaff, Turner, & Covey, 1998) and is involved in positive strand synthesis. The TY3 retrotransposon of *Saccharomyces cerevisiae* has a PPT that is 12pb long. In the human immunodeficiency virus Type 1 (HIV-1) and most lentiviruses, there is a second PPT present within the integrase region near the center of the genome (Powell & Levin, 1996). A conserved T-rich sequence is found upstream of the PPT which aids retrotransposon replication (Ilyinskii & Desrosiers, 1998; Wilhelm, Uzun, Mules, Gabriel, & Wilhelm, 2001).

The primer binding site (PBS) is located just downstream of the U5 region of 5'LTR. The length of the PBS in retrotransposons varies from 8 to 18 nts (Marquet, Isel, Ehresmann, & Ehresmann, 1995). The PBS sequence is complementary to a part of their primer tRNA. Most PBS base pairs with 3'-end of tRNA. The tRNA acts as a primer for minus/first strand cDNA synthesis. Different retroelements use different species of tRNAs as primers for replication, that bind to the primer binding site and initiate reverse

transcription (Lund, Duch, & Pedersen, 2000). It has been noted that plants always use the 3' end of tRNA methionine as initiator, while animal retroelements use more diverse species of tRNA (Grandbastien, 1992; Hirochika, Fukuchi, & Kikuchi, 1992; Jin & Bennetzen, 1989). *Fritillaria imperialis* sequences with accession numbers GU188682, GU188679, GU188681, GU188678 and *Fritillaria affinis* sequences with accession numbers GU188677, GU188676, GU188680, GU188675 have pbs sequences similar to trna met.

1.4 Annotation

Genome annotation involves the prediction of various features on the DNA sequence, such as coding genes, regulatory regions, repeat elements etc. (Reeves, Talavera, & Thornton, 2009). Annotation tools and databases used together in the annotation process disseminate biological meaning to the raw sequence data (Reeves et al., 2009). Coding sequences receive most attention, primarily because peptides serve as drug targets for drug discovery (Ofran, Punta, Schneider, & Rost, 2005). Moreover, the repetitive nature of the TEs make the annotation processes a challenge.

The complexity in TE annotation arises from the fragmented nature of the TE as a result of events such as, nesting of TEs into each other, recombination among TEs and sequence divergence. Also, interspersion of TEs into other type of repeat classes makes it difficult to identify them (Bergman & Quesneville, 2007). The copies of elements belonging to the same family, although similar in sequence, are not identical due to

evolutionary mechanisms that create point mutations, indels and rearrangements. These biological realities pose recurrent difficulties in automated detection of these elements (Lerat, 2010).

There are three main methods of searches, the *de novo* or *ab initio* method, which is aimed at discovering new TEs without using prior information about structure or similarity to known TEs. This method is useful to detect repeats in high copy number (Bergman & Quesneville, 2007). We did not use this method, since we were analyzing only 55kb of sequences. The second method is a homology based method. This method is typically applied to assembled genome sequences and capitalizes on the cumulated knowledge present in the large number of previously reported TEs. For example, RepeatMasker uses homology based algorithms to search a known repository of TEs such as RepbaseUpdate (Steinbiss, Willhoeft, Gremme, & Kurtz, 2009). This method detects TEs based on protein homology to known TE protein-coding sequences. But, homology based methods are biased towards recently active elements that retains protein homology (Bergman & Quesneville, 2007). The third method is a structure-based method, this method searches common structural signals, such as the size range of the LTR sequences, the distance between the LTRs of an element, the percentage of identity between the two LTRs, the presence of PBS, PPT and TSDs (Lerat, 2010). LTR_STRUC, LTR_FINDER and LTRHarvest all employ this methodology. This method is less biased by similarity to a known set of elements like the homology based method. This method can also identify new elements falling within the same class of repeats. Both structure based method and homology based methods can detect TEs with low copy number. However,

the distinction between these methodologies is not absolute (Bergman & Bensasson, 2007).

Although there are several tools available, they are not cross tested by researchers. Secondly, most of these tools are developed by labs to answer a particular question pertaining to their organism of interest. Hence, cross functionality of these tools is questionable. For example, LTR_par works with yeast genome for which it was originally developed and not on *D. melanogaster*. These programs require prior knowledge on the approximate parameters to input into the program for example, the distance between the two LTRs or knowledge on the PBS. For a newly sequenced organism with repeats, of which there is no prior knowledge on the elements present, it becomes increasingly difficult to decide on the particular parameters to be used to detect the repeat elements (Lerat, 2010). Hence, according to Lerat, using several different programs and cross comparing the results obtained, is a reliable strategy than using a single program (Lerat, 2010). Thus, a careful approach of a combination of homology based methodology, structure based methodology and blast searches were used in *Fritillaria agrestis*.

1.5 Standards

An intact element is a one that has two relatively intact LTRs flanked by TSDs and also has PBS and PPT identified. Truncated elements are those that have deletions at the 5' or 3' ends. These include elements with one or both LTRs partially deleted,

elements with one or both LTRs completely deleted, elements with one partially deleted and one completely deleted LTR. Other elements showing homology with partial retrotransposon sequences, but not having any recognized structural features are called "remnants/fragments" (Ma et al., 2004).

1.6 Objective

Objective of the study is to annotate the genomic DNA sequences of *Fritillaria agrestis* in order to learn about the sequences responsible for genomic obesity in plants. Specific aims pertain to identifying LTR retrotransposons and annotating the structure of intact elements, to the extent of mining catalytic domains.

2 Materials And Methods

2.1 Material

Genomic DNA sequence of 1 BAC clone from *Fritillaria agrestis* was provided by Dr.Chris Baysdorfer. The sequence size is 54802 base pairs.

2.2 Initial similarity search

Blastx and tBlastx searches were done on non-redundant (nr) databases at Genbank with a cut off value of e-value of e-5 (Yuan, SanMiguel, & Bennetzen, 2003). Later a stricter e-value of e-10 was set. The searches were done in small fragments of 200bp and the sequences at the junction of these fragments were taken at 200bp blocks and the same search was repeated again. The entire search was automated with ActivePerl 5.10.1 and BioPerl 1.6.923 on windows platform. The results were captured in Excel VBA.

2.3 Strategies for Mining Repetitive Elements

A combination of manual procedures and tools were used to identify intact and fragmented transposable elements. The following tools were used to mine repetitive

elements. They are Censor, RepeatMasker, LTR_Finder, and LTRStruct. Also, the previously mentioned Blastx/tBlastx (McGinnis & Madden, 2004) search against non-redundant (nr) database with e-value of -10 was done. Together, *Fritillaria imperialis* and *Fritillaria affinis* sequences in the Genbank database were used for comparative analysis, especially for gag and LTR sequences. Blastx/tBlastx search was also done against DNA transposon sequences obtained in the lab from the following plant species; *Lillium pardalinum, Fritillaria agrestis, Clintonia uniflora, Prosartes hookerii, Scoliopus bigelovii* and *Smilax californica*.

Parameters used for various searches:

Censor :Search against *Panicoideae*, *Oryza, Triticum, Arabidopsis, Poaceae*, and *Viridiplantae* sequences in Repbase Update (Jurka et al., 2005)

Also, force translated search and report simple repeats parameters were enforced.

LTRFinder (Zhao & Wang, 2007) main parameter changed was output threshold score set to 4/5

Blastx : nr database

RepeatMasker ver 4.0 DNA source: *Tritiaceae, Panicoideae* and Oryza. Parameter: default abblast.

LTRStruct: Default parameters were used.

2.4 Identification of Long Terminal Repeats and Target Site Duplications.

Here, since the domain identification preceded the LTR identification, the approximate position of the LTRs became known. A Blast2Sequence of the region upstream of protease and downstream of the chromodomain was done. The matched regions and the other parameters like the presence of TSD, qualifying it as LTRs were used to identify LTRs. The polyprotein annotation was used to determine the orientation of the long terminal repeat regions.

Using the terminal repeat regions mined, the rest of the unannotated sequence was searched for repeated regions using the newly found LTRs as a reference sequence. Also, the *Fritillaria affinis*, sequences reported in NCBI were used as reference sequence to search for LTRs using NCBI BLAST.

The TSDs were identified by looking for 4-6 bp of matching direct repeats on either ends of the 3' and 5' LTRs.

The Polypurine tract (PPT) was identified based on purine frequency and the thymidine rich region upstream of the PPT (Chaparro, Guyot, Zuccolo, Piegu, & Panaud, 2006) .

The Primer Binding Site (PBS) was searched using methionine tRNA sequence since it was reported in other *Fritillaria* retroelements. Also, tried the tool tRNASCAN-SE (Lowe & Eddy, 1997). The tool did not work with *Fritillaria agrestis* and for the rice sequence (accession number: AB030283) for which poly-A signal was identified and reported in Genbank. So I just used the tRNA methionine sequence to look for PBS.

2.5 Identification of Regulatory Sequences in the LTRs

The Poly-A site and Poly-A signal were identified by using PASPA, a webserver for polyA site prediction in plants and algae (http://bmi.xmu.edu.cn/paspa/index.html) (Ji et al., 2015), and PLACE (A database of plant Cis-acting Regulatory DNA elements) (Higo, Ugawa, Iwamoto, & Korenaga, 1999).

2.6 Domain Identification

Proteins are made up of small building blocks called ‘domains’ or ‘modules’. They appear as distinct regions in the 3D structure. Knowledge of protein domain architecture and boundaries help in delineation of boundaries for homologous domains in multiple sequence alignment (Kong & Ranganathan, 2004).

All domain databases can be classified into either structure-based or sequence-based domains. Sequence-based databases are where our interests lie. The main sequence-based databases are ProDom, DOMO, Pfam, SMART, COGs, BLOCKS, SBASE and Interpro. Of these databases, I was interested in using databases that accepted nucleotide sequence as input.

Conserved Domain Database (CDD) is a protein annotation resource which mainly uses NCBI-curated domains as well as domain models from Pfam, SMART, COG, PRK and TIGRFAM. CDD (Marchler-Bauer & Bryant, 2004; Marchler-Bauer et al., 2013). Initially the domain region was identified through homology based search

with Blastx. The initial results obtained from Blastx searches, served as input to the CDD search. Presence of motifs and conserved residues were used to confirm the CDD results.

2.7 Identification of Microsatellites

RepeatMasker was used to identify microsatellites and the results were verified manually having a cutoff value of minimum five iterations. RepeatMasker scans from 2-5 bp units. It picks up repeats that have less than 10% divergence from perfect repeat. Mononucleotide repeats were identified manually.

2.8 Identification of Segmental Duplication

Blast2seq was done to find the sequence identity (Khaja, MacDonald, Zhang, & Scherer, 2006).

2.9 Annotation Workflow

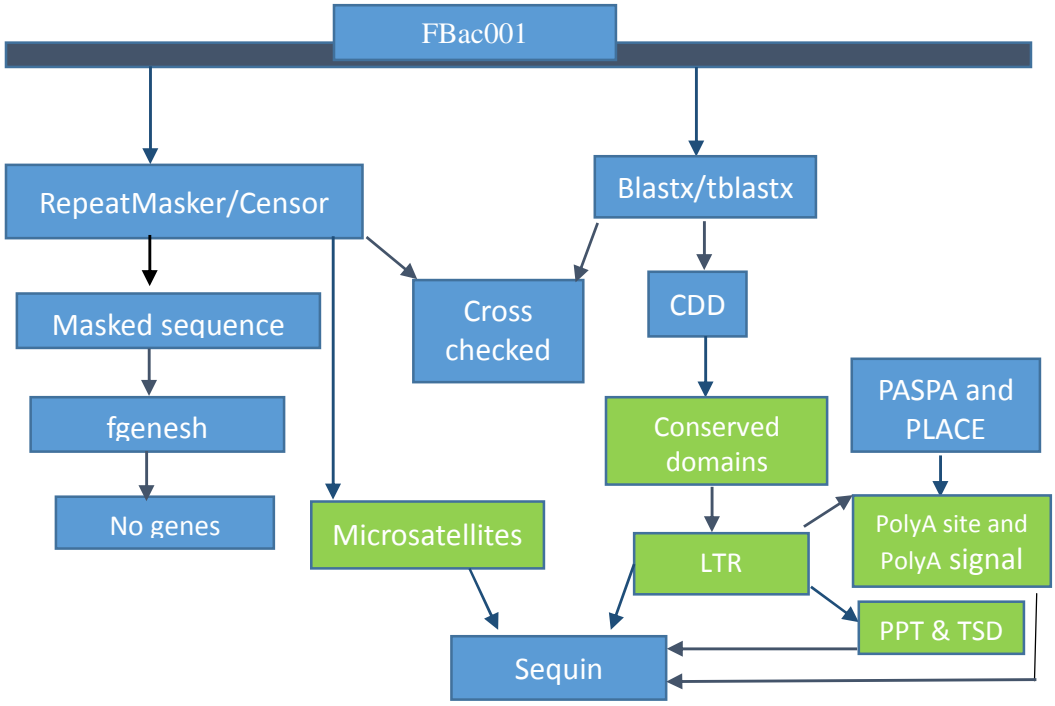

Fig. 2 Workflow of the annotation process

Fig. 2 is a workflow of the annotation process. To find if there were any genes, the masked file using RepeatMasker was used in the program fgenesh. Fgenesh is a gene prediction program for genomic DNA. No genes were identified with fgenesh. The results of RepeatMasker and Censor were compared to Blast search results, and the resulting sequences were fed to CDD. Once the domains were identified, like mentioned before they were used to identify LTRs. The LTR sequences were then fed into PLACE and PASPA to identify Poly-A signal and Poly-A site. Although, RepeatMasker identified simple repeats, they were manually checked and less perfect repeats were

removed. All the results were tabulated and fed into sequin, a Genbank annotation tool. We received the Genbank accession number KT290211, but the data is not published yet.

3 Results

The initial Blastx and tBlastx search with e-4 gave usable relevant results, then the automated search with e-10 was done to set a stricter filter. The results were tabulated in the excel sheet. The results obtained from RepeatMasker and Censor were compared with that of BLAST. The Censor and RepeatMasker analysis showed that the sequences mainly had retrotransposons, both Copia and Gypsy types. A few DNA transposons were reported by censor, but did not return any results against Genbank. Also, in a search against the DNA transposon sequences obtained from the lab, did not return any similarity. Hence, it can be said that the *Fritillaria agrestis* sequence does not contain any readily identifiable DNA transposons.

3.1 Censor

Censor is a tool used to find repeated elements. Censor uses Repbase Update, a database maintained by the Genetic information research institute (GIRI). Repbase Update is also used by RepeatMasker. Repbase is a repository of repetitive DNA from different eukaryotic species (Jurka et al., 2005). The masked sequences represent low complexity regions, tandem repeats, CpG islands and isochores which do not have significant biological meaning. This region if not masked, increases the number of non-homologous similarities that are stronger than homologous similarity (Frith, 2010). The search of repeat elements was done against the *Panicoideae, Oryza, Triticum*, *Arabidopsis, Poaceae,* and *Viridiplantae*. Collating all the results it can be seen that all

the hits are against transposable elements, with most of them being long terminal repeats. Among the LTRs majority of them are gypsy and the remaining are Copia.

The NCBI BLASTX search result was compared against the Censor results of *Panicoideae* and *Viridiplantae* for confirmation. All the results were the same, expect for the DNA transposons and the *Viridiplantae* result of Gypsy-10_PAb-l. Moreover, the name of this particular repeat was not listed in Repbase. As mentioned earlier, the DNA transposon results did not give any results against Blastx and tBlastx search.

3.2 RepeatMasker

RepeatMasker search was done against *Triticum, Oryza and Zea*. Comparing all three results, *Oryza* results are closer to what is actually obtained with *Fritillaria*, containing three Gypsy elements and two Copia elements. All three have not reported any DNA transposons. All three have only reported LTR elements of Copia and Gypsy type, along with some simple repeats. This tool more precisely helped in identifying the simple repeats and helped in confirmation of the results obtained for the Blast search.

3.3 LTRs, TSDs and Regulatory Sequences

The LTRs annotated had 83% identity and lacked the terminal inverted repeat sequence TG… CA. A 4bp direct repeat (AGAC) Target site duplication (TSDs) was found in one full length (we call it full length although the PBS was not identified) gypsy

element. At the 5' end, the TSD was found just upstream of the LTR, but at the 3' end, the TSD was found 10 to 20 bp downstream of the LTR. This is probably due to the degeneration of the LTRs, that have accumulated mutations over time. Also, the TSDs were found only in one gypsy element, of the two full length gypsy elements.

LTR_Finder a web based tool to search LTRs was used to find LTRs. The search was done against *oryza sativa, Zea mays and Glycine max*. The default settings of LTR_Finder did not return any results. Lowering the output score threshold from default 6 to 5 and not having any domain restrictions returned some results. Most of the LTRs obtained were in minus strand. For one, the domain region was identified as LTRs.

```
5'-LTR   : 8226 - 9197 Len: 972
3'-LTR   : 23773 - 24744 Len: 972
```

For example the above mentioned region is identified as LTR, in all the plant searches. But, when checked, this region actually corresponds to RNase H and Reverse transcriptase. Likewise, every one of the output had to be checked for the validity of the results. While, using these software to predict genomes with little prior knowledge, post processing is highly recommended to weed out false positives (Ellinghaus, Kurtz, & Willhoeft, 2008). In this case it is done by doing a manual check of the results.

Repeat models that have a threshold score higher than 5 were returned. When the domain restrictions were applied, along with the lowered output threshold score, no results were obtained. All the results obtained from LTR_Finder were false positives, except one true positive with the opposite orientation of what was identified manually. But, LTR_Finder gave results with the rice sequence (accession number: AB030283).

Hence, LTR_Finder results could not be used. It is probably because the PBS is mutated that LTR_Finder did not work for *Fritillaria agrestis*.

The potential polypurine tract (PPT) was found adjacent to 3'ltr, but about 90 bps upstream of the 3' ltr. This too is attributed to the degeneration of the ltrs.

The Primer binding site (PBS) at which reverse transcription initiates could not be identified. There was similarity to 11bp, 21bp downstream of 5'LTR, but 4 bp were mutated. So could not take that for PBS. The distance between 5'LTR and PBS varies in LTR retrotransposons. But, in HIV there are 2bp between them. The reason the sequences 21bp downstream of 5'LTR was considered was because the element has accumulated mutations which is evident from the 83% identity of the LTRs.

The polyA signal and polyA site were identified and the exact location of these can be found in the annotation table.

3.4 Conserved Domains Database Search

Refer to Fig. A1. Ty3/Gypsy Gag in Appendix 2 for the following 3 paragraphs. The sequence identified is a Ty3/Gypsy Gag which presents similarity to the major homology region (MHR) motif by having the sequence QX2GESX2EX2EXFX5QXPXH, when aligned against pfam 03732 consensus sequence. The central motif identified by pfam 03732 (Retrotrans_gag) is

QGX2EX5FX2LX2H. *Fritillaria* sequence varies from the other sequences by a variation in the number of bases in between the highly conserved six residues. We can

see here that the sequence has been identified with an introduction of a gap next to Glutamine (Q) in the MHR motif. Although, CCHC zinc finger motif has not been identified in this sequence, it still stands confirmed as a gag of Ty3/gypsy due to the presence of the MHR motif.

Also, this same gag sequence when searched against probable gag sequence clusters, generated in Nextgen sequencer, did not present any sequence homology. This is probably due to the high variation in the gag sequences.

Refer to Fig. A2. Protease in Gypsy element in the region 14805bp to 15068 present in Appendix 2 for the following two paragraphs.

A hit was obtained against retropepsins. Pepsin (A1) and retropepsin (A2) are two distantly related sub-families of peptidases. Retropepsins have been identified in retrotransposons and retroviruses. Pepsin corresponds to the cellular proteases which are bilobed with active site cleft located between the lobes. Retropepsins on the other hand have a single lobe and have only a single aspartic acid residue compared to pepsins which have a pair of aspartic acid residue contributed by one lobe each. Retropepsins and pepsins share low homology except in the active site region, where the residues in the active site are conserved (Barrett et al., 2004; Rawlings & Barrett, 1995).

In the alignment obtained, the highlighted residues 13, 14 and 15 of *Fritillaria*, which are V T I correspond to DSG of the aspartic acid active site residues, in the aligned consensus sequence of retropepsins. In *Fritillaria*, we have Thr in place of Ser. The other two residues differ.

Refer to Fig. A3. Protease in Gypsy element in the region 33132bp to 33395bp in Appendix 2 for the following two lines.

The aspartic acid residues DSG found between the residues 10 and 20 in the consensus sequence aligns with DMV of Fritillaria.

Refer to Fig. A4. Protease in Gypsy element in the region 52344bp to 52529bp in Appendix 2 for the following two lines.

The residues DTG between residue position 20 and 30 in the consensus sequence, aligns with DLR in Fritillaria.

All the three proteases identified are Gypsy elements. Proteases were not found in the two Copia elements identified. Among the three gypsy elements, the conserved residues we obtained in Fritillaria are VTI, DMV and DLR. The aspartate (D) remains fairly conserved. For the next residue, *Fritillaria* has used Threonine instead of Serine.

Refer to Fig. A5. Integrase core domain in gypsy element in the region 17118bp to 17453bp in Appendix 2 for the following paragraph.

The central core region is comprised of approximately 120 amino acids. The highly conserved residues in the core domain D64, D116 and E152 are also conserved in *Fritillaria*, except for glu (E), which is replaced by lysine (k). This substitution is found even in barley (*Hordeum vulgare*) in its Copia-like BARE-1 retrotransposon family. In barley both variations were found in the amino acid position 152, although the lysine variation was rare (Suoniemi, Tanskanen, Pentikainen, Johnson, & Schulman, 1998). Similarly in *Fritillaria*, of the two gypsy sequences, one has lysine and the other one has glutamic acid.

Refer to Fig. A6. Integrase core domain in gypsy element in the region 35499bp to 35792bp in the Appendix 2 for the following paragraph.

In this sequence, glu is conserved. Also, the 5' end of the core domain sequence is truncated with the first aspartate missing. Both the sequences presented here belong to the Gypsy type. Although, two copia type integrase region were identified in *Fritillaria* against Blast search, they were not identified with the conserved domain search because the conserved core domain region is missing in the integrase.

Refer to Fig. A7. Copia reverse transcriptase in the region 8452 to 9043 present in Appendix 2 for the following paragraph.

This Copia reverse transcriptase element is oriented in the 3' to 5' direction. The alignment with the consensus sequence here is in the minus frame. The alignment is achieved with a frame shift in between, leaving three residues from 81-83 in the query sequence. The omitted nine bases can be attributed to mutation since this is a very old element. The conserved residues have not been identified in pfam 07727. Some of the conserved residues that were identified are highlighted in yellow. The conserved motif YXDD is found in the consensus sequence as YVDD. Whereas, in Fritillaria we have YVIN.

Refer to Fig A. 8 Gypsy reverse transcriptase in the region 15424 to 15945 present in Appendix 2 for the following paragraph.

The alignment is obtained in three different frames. Also, the results show the alignment of consensus sequences from both NCBI database and pfam. There are seven conserved regions that are common to all retro elements in reverse transcriptase. All the

gaps in the alignment are localized between the seven commonly found regions (Xiong & Eickbush, 1988). The first frame shows the region 1 conserved residue K, as initially identified by Eickbusch. Following the first frame there is a frame shift to third frame from residue 29 as shown in the alignment with cd01647 consensus sequence. The third frame subject sequence is aligned against both pfam and cdd consensus sequences. Of which the cdd consensus sequence has a better alignment with the region 5 Y+DD motif of the consensus sequence, aligned against the FIYD of the subject sequence. Following this, there is another frame shift, from third frame to the second frame for the rest of the sequence. The conserved residues LG is identified in the subject as well as the query sequence.

Refer to Fig. A9. Copia reverse transcriptase in the region 23999 to 24601 in Appendix 2 for the following two lines.

The Copia reverse transcriptase gene runs in the 3' to 5' direction in a single frame. YVIN is found in place of YVDD.

Refer to Fig. A10. Gypsy reverse transcriptase in the region 33752bp to 34279bp and Fig A11 in the Appendix 2 for the following few lines.

The Gypsy reverse transcriptase gene runs from 5' to 3' direction. The sequence is obtained in a single frame and is highly conserved.

Refer to Fig. A12. Copia RNase H in the region 7735bp to 8105bp present in Appendix 2 for the following few lines.

The Copia RNase H is in 3’ to5’ direction. The alignment here is achieved in two different frames. The sequence is truncated hence the last conserved residue in DED [D/E] is missing. Moreover, there is a gap in place of the third conserved residue D.

Refer to Fig A13. Gypsy RNase H in the region 16224bp to 16565bp in Appendix 2 for the following lines.

In the Gypsy RNase H sequence, all four catalytic residues DEDD are found in a single frame.

Refer to Fig. A14 in Appendix 2 Copia RNase H in the region 23353bp to 23723bp for the following lines.

The Copia RNase H sequence is aligned with a frame shift in between. Also, only the first and second conserved residues D and E are found, similar to the previous Copia RNase H.

Refer to Fig. A15 in Appendix 2 Gypsy RNase H in the region 34621bp to 34889bp for the following line.

In the above Gypsy RNase H sequence, all four catalytic residues DEDD are found.

Refer to Fig. A16 in Appendix 2 Gypsy RNase H in the region 54160bp to 54513bp for the following line.

This Gypsy RNase H sequence is in a single frame with all four catalytic residues (DEDD) found.

Chromo domain was not obtained from conserved domain search. Chromo domain gave results with Blast search, but conserved domains could not be identified. So did not annotate chromo domain.

3.6 Microsatellites

The location of the microsatellites can be found in the sequin table (Refer to Appendix 3). The majority of them are dinucleotides along with a small number of mononucleotide repeats. Among the dinucleotide there is a high representation of (AT) n.

3.7 Segmental Duplication

Segmental duplication also called low copy repeats are near identical regions of DNA at two or more sites in the genome (Khaja et al., 2006). They can vary in size from few base pairs to several megabases. They can be tandem duplication or interspersed within the same chromosomes or distinct chromosomes (Mendivil Ramos & Ferrier, 2012). They arise as a result of unequal recombination.

Dot plot in Fig. 3 shows several parallel diagonal lines on either side of the middle diagonal line indicating the presence of repeats. The single long diagonal repeat region along with the adjacent diagonal repeat region were compared in the blast2seq. The blast2seq gave a sequence identity of 90% with 99% coverage for 18,358bp clearly indicating segmental duplication. The regions are 5847 to 21391 and 21393 to 39751 in the Fbac001.gb, DNA sequence file. Each segment comprises of a Ty1/copia fragment

and a full length Ty3/gypsy element. The Ty3/gypsy element has inserted into the Ty1/copia element truncating the Ty1/copia element before the duplication event occurred.

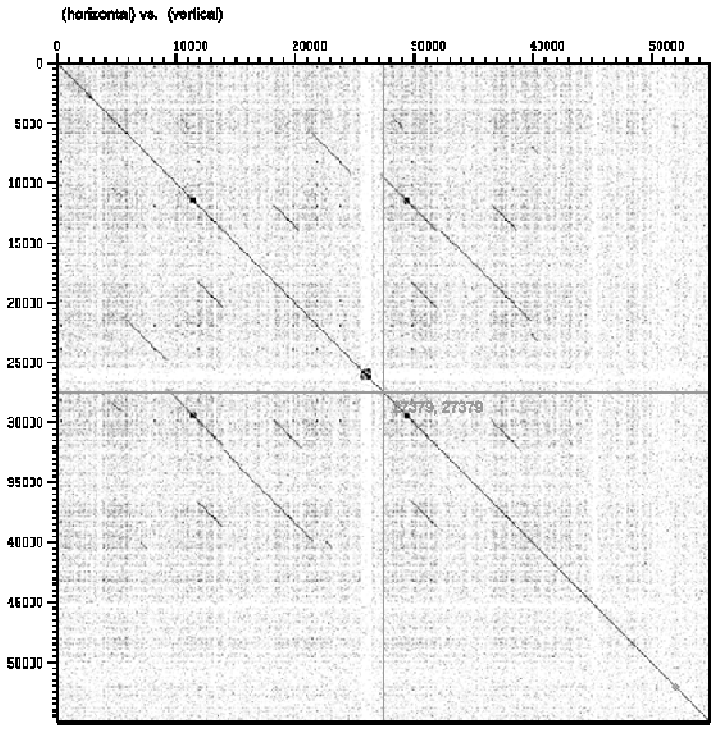

Fig. 3 Dotplot of Fritillaria agrestis intra-sequence comparison

All the information obtained with the processing of the DNA sequence file were tabulated and submitted to genbank through sequin version 13.70 and obtained the Accession number KT290211.

4 Discussion

The sequences of 1 BAC clone from *Fritillaria agrestis* were analyzed were analyzed for the presence of repetitive DNA and genes. The results showed that it was predominantly composed of LTR retrotransposons of the Gypsy and Copia type elements. In the 55kb sequence, we identified 3 Gypsy elements and 2 Copia fragments.

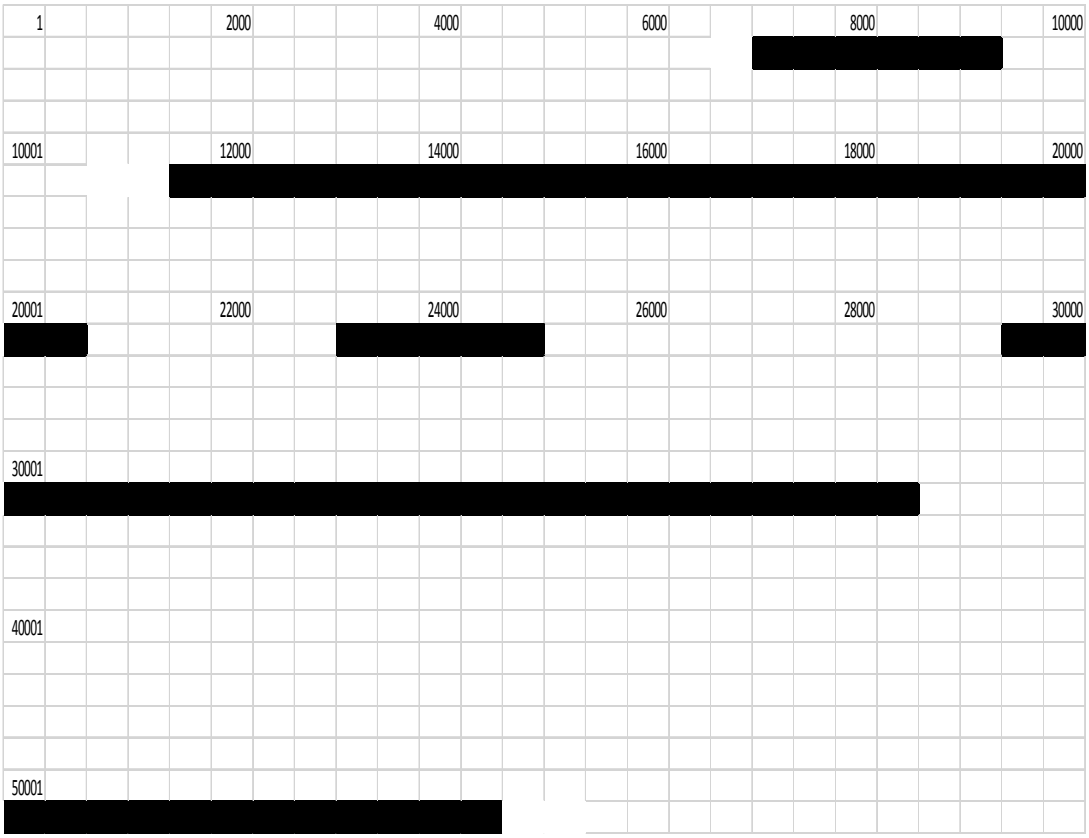

Fig. 4 Approximate length and location of the retrotransposon elements

The two small fragments around 8,000bp and 24,000bp are LTR retrotransposon Copia fragments and the 3 long elements i) running between 10,000bp and 21,000bp, ii)

element between 28,000bp and 40,000bp, iii) element between 50,000bp and 55,000bp are all LTR retrotransposon Gypsy elements.

Fig. 5 Domain region highlighted approx. retrotransposons depicted

The two Copia elements have RNase H followed by reverse transcriptase in the 3' to 5'-direction. In the two full-length gypsy elements gag could not be identified. Whereas, in the last truncated gypsy element, gag was identified. In the first gypsy

element running between 10,000bp and 21,000bp, TSDs were also identified. All the gypsy elements are in 5' to 3' direction.

Two of the three Gypsy elements are full-length elements which got inserted into the Copia elements truncating the Copia elements. Both the full length Gypsy elements contain LTRs found with PPT, PolyA-site and PolyA-signal and all conserved domains with the exception of gag. The third Gypsy element is truncated; on one end the 55kb sequence terminates with RNase H. At the 5'-end there is no identifiable long terminal repeat upstream of gag. But, the gag was clearly identified with the conserved sequences, with some variation in the MHR region. Gag is known to be the least conserved among all domains. It is known to be diverse even within the same species. It is because of the diverse nature that the gags in the other two gypsy elements could not be identified. The PPT was identified not exactly where the 3'-LTR begins, but a few bases upstream of the 3'-LTR, in both the gypsy elements. This is probably due to the degeneration of the LTRs. As with Copia, both the elements are remnants, which only have reverse transcriptase and RNase H conserved domains of the pol protein. The copia integrase domain could not be annotated with conserved domains database as it did not generate any hit against the CDD database, because the conserved region was missing altogether due to truncation. But, it generated a hit against Genbank with tBlastx search. Both the Copia elements are found in reverse direction.

Mutations are the reason for the low identity match of 83% between the two LTR terminal repeats. At the time of insertion the sequences of 3' and 5'-LTRs are identical, as time elapses, they begin to accumulate mutations and thereby diverge from one

another in their nucleotide sequence. In other words, a low identity score means that the elements were not inserted recently.

This *Fritillaria agrestis* sequence also contains with the di nucleotide (AT)n microsatellites. Similar repeats have been reported earlier in other species of *Fritillaria* (Ambrozova et al., 2011). Among the dinucleotides, the (CA)n repeats are reportedly the most frequently occurring repeat, followed by (AT)n, (GA)n and (GC)n. However, in plant genomes (AT)n is the most frequently occurring motif (Ellegren, 2004).

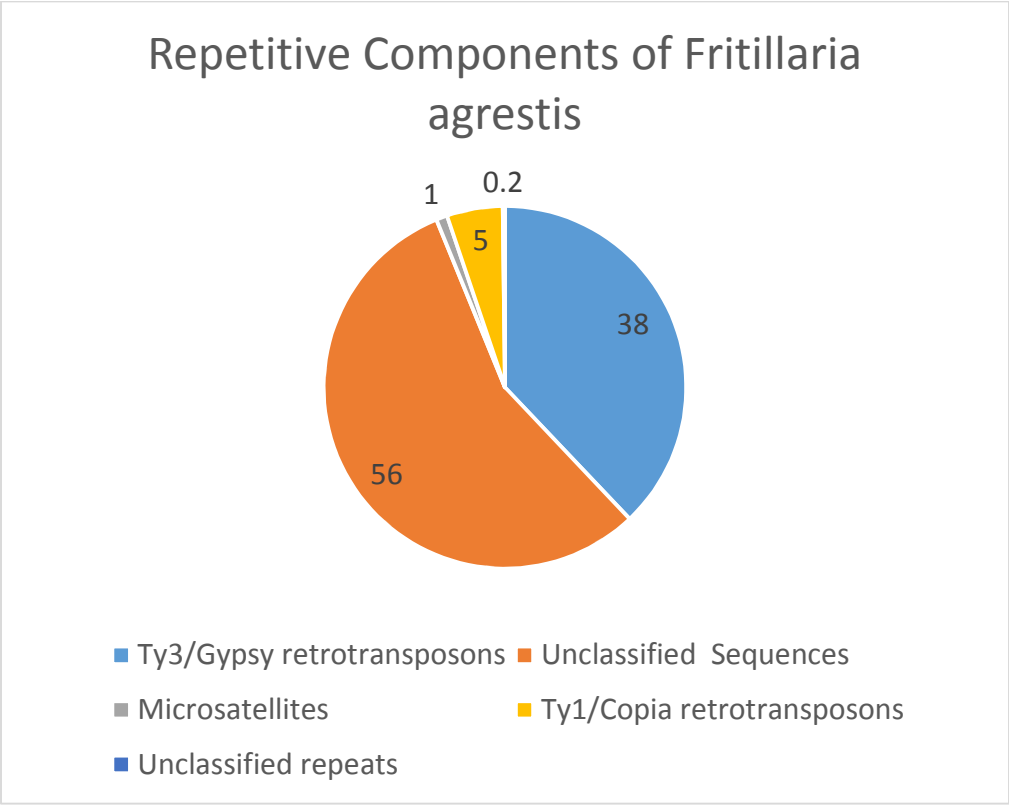

Fig. 6 Repetitive Components of Fritillaria agrestis

The major component of *Fritillaria agrestis* in this 55kb sequence is Ty3/Gypsy, which makes up approximately 40% of the sequence. It is followed by Ty1/Copia elements which makes up 5% of the sequence. Microsatellites only constitute 1% of the sequence. The rest of the unclassified sequences did not show any sequence similarity with sequences in Genbank and Repbase update. This implies, these unclassified sequences are 'unique DNA', which have diverged too far from ancient transposable elements over the years to be recognized with similarity searches as transposable elements (Lander et al., 2001).

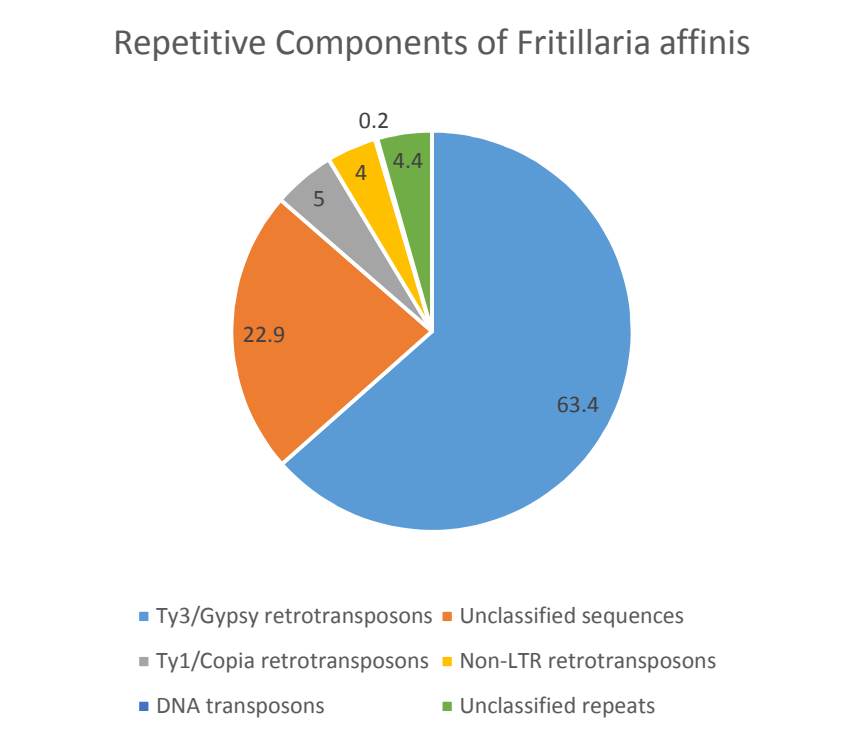

Fig. 7 Repetitive Components of Fritillaria affinis (Ambrozova et al., 2011)

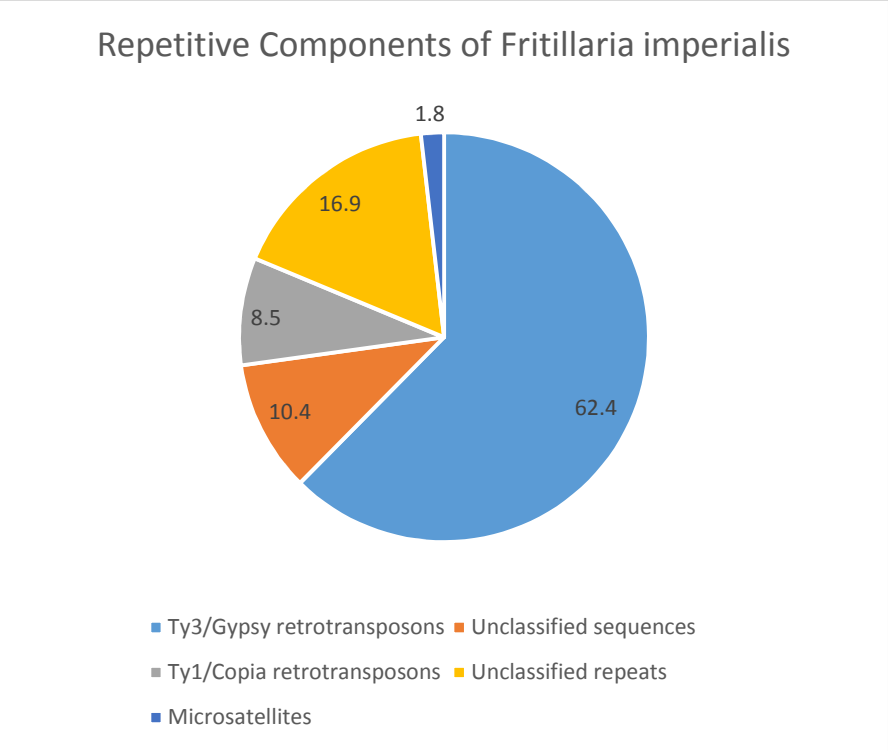

Fig. 8 Repetitive components of Fritillaria imperialis (Ambrozova et al., 2011)

The data from Fig 7 and Fig 8 were obtained from Ambrozova et al paper (Ambrozova et al., 2011) to compare the repetitive components of all three genomes. When comparing all three genomes of *Fritillaria agrestis, Fritillaria affinis* and *Fritillaria* imperialis, certainly retrotransposons represent at half or more of the sequences. And of the repeats, LTR retrotransposons form the major component. Among the LTR retrotransposons Ty3/Gypsy elements dominate the sequences.

5 Conclusion

Annotating the 55kb *Fritillaria agrestis* BAC sequences, revealed the composition of the sequence, which is primarily made up of LTR retrotransposons. Microsatellites form a small component of the sequence. Among the LTR retrotransposons in this clone, a Gypsy element has inserted into the Copia element thereby truncating the Copia element and a subsequent segmental duplication event then occurred, doubling both the events. When comparing *Frtillaria agrestis* sequence components with *Fritillaria affinis* and *Fritillaria imperialis*, it is evident that these genomes are mainly composed of LTR retrotransposons.

6 References

Abe, H., Ohbayashi, F., Sugasaki, T., Kanehara, M., Terada, T., Shimada, T., . . . Oshiki, T. (2001). Two novel Pao-like retrotransposons (Kamikaze and Yamato) from the silkworm species Bombyx mori and B. mandarina: common structural features of Pao-like elements. *Molecular genetics and genomics : MGG, 265*(2), 375-385.

Ambrozova, K., Mandakova, T., Bures, P., Neumann, P., Leitch, I. J., Koblízkova, A., . . . Lysak, M. A. (2011). Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of Fritillaria lilies. *Annals of Botany, 107*(2), 255-268. doi: 10.1093/aob/mcq235

Asante-Appiah, E., & Skalka, A. M. (1997). Molecular mechanisms in retrovirus DNA integration. *Antiviral Research, 36*(3), 139-156. doi: http://dx.doi.org/10.1016/S0166-3542(97)00046-6

Barrett, A. J., Woessner, J. F., & Rawlings, N. D. (2004). *Handbook of proteolytic enzymes* (Vol. 1): Elsevier.

Bennett, M. D., & Leitch, I. (2005). Genome size evolution in plants. *The evolution of the genome*, 89-162.

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology, 42*(1), 251-269.

Bennetzen, J. L. (2002). Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica, 115*, 29-36.

Bennetzen, J. L. (2005). Transposable elements, gene creation and genome rearrangement in flowering plants. *Current Opinion in Genetics & Development, 15*(6), 621-627.

Bennetzen, J. L., Ma, J., & Devos, K. M. (2005). Mechanisms of Recent Genome Size Variation in Flowering Plants. *Annals of Botany, 95*(1), 127-132.

Bergman, C. M., & Bensasson, D. (2007). Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster. *Proc Natl Acad Sci U S A, 104*(27), 11340-11345. doi: 10.1073/pnas.0702552104

Bergman, C. M., & Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics, 8*(6), 382-392. doi: 10.1093/bib/bbm048

Boeke, J. D., Garfinkel, D. J., Styles, C. A., & Fink, G. R. (1985). Ty elements transpose through an RNA intermediate. *Cell, 40*(3), 491-500.

Boeke, J. D., & Sandmeyer, S. B. (1991). 4 Yeast Transposable Elements. *Cold Spring Harbor Monograph Archive, 21*, 193-261.

Cameron, J. R., Loh, E. Y., & Davis, R. W. (1979). Evidence for transposition of dispersed repetitive DNA families in yeast. *Cell, 16*(4), 739-751.

Cerritelli, S. M., & Crouch, R. J. (2009). Ribonuclease H: the enzymes in eukaryotes. *FEBS Journal, 276*(6), 1494-1505.

Chaparro, C., Guyot, R., Zuccolo, A., Piegu, B., & Panaud, O. (2006). RetrOryza: a database of the rice LTR-retrotransposons. *Nucleic Acids Research, 35*(suppl 1), D66.

Clare, J., & Farabaugh, P. (1985). Nucleotide sequence of a yeast Ty element: evidence for an unusual mechanism of gene expression. *Proceedings of the National Academy of Sciences, 82*(9), 2829-2833.

Clark, D. J., Bilanchone, V. W., Haywood, L. J., Dildine, S. L., & Sandmeyer, S. B. (1988). A yeast sigma composite element, TY3, has properties of a retrotransposon. *Journal of Biological Chemistry, 263*(3), 1413-1423.

Clemens, K., Larsen, L., Zhang, M., Kuznetsov, Y., Bilanchone, V., Randall, A., . . . Sandmeyer, S. (2011). The TY3 Gag3 Spacer Controls Intracellular Condensation and Uncoating. *Journal of Virology, 85*(7), 3055-3066. doi: 10.1128/jvi.01055-10

Craven, R. C., Leure-duPree, A. E., Weldon, R. A., & Wills, J. W. (1995). Genetic analysis of the major homology region of the Rous sarcoma virus Gag protein. *Journal of Virology, 69*(7), 4213-4227.

Dixit, A., Ma, K.-H., Yu, J.-W., Cho, E.-G., & Park, Y.-J. (2006). Reverse transcriptase domain sequences from Mungbean (Vigna radiata) LTR retrotransposons: Sequence characterization and phylogenetic analysis. *Plant Cell Reports, 25*(2), 100-111.

Doolittle, R. F., Feng, D. F., Johnson, M. S., & McClure, M. A. (1989). Origins and evolutionary relationships of retroviruses. *Quarterly Review of Biology, 64*(1), 1-30.

Dunsmuir, P., Brorein Jr, W. J., Simon, M. A., & Rubin, G. M. (1980). Insertion of the drosophila transposable element< i> copia</i> generates a 5 base pair duplication. *Cell, 21*(2), 575-579.

Eickbush, T. H., & Jamburuthugoda, V. K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Research, 134*(1-2), 221-234. doi: S0168-1702(07)00463-7 [pii]
10.1016/j.virusres.2007.12.010

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics, 5*(6), 435-445.

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics, 9*, 18.

Engelman, A., & Craigie, R. (1992). Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function in vitro. *Journal of Virology, 66*(11), 6361-6369.

Engelman, A., Englund, G., Orenstein, J. M., Martin, M. A., & Craigie, R. (1995). Multiple effects of mutations in human immunodeficiency virus type 1 integrase on viral replication. *Journal of Virology, 69*(5), 2729-2736.

Feschotte, C., & Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review of Genetics, 41*(1), 331-368. doi: doi:10.1146/annurev.genet.40.110405.090448

Flavell, A. J. (1992). Ty1-copia group retrotransposons and the evolution of retroelements in the eukaryotes. *Genetica, 86*(1-3), 203-214.

Flavell, A. J., & Smith, D. B. (1992). A Ty1-copia group retrotransposon sequence in a vertebrate. *Molecular and General Genetics, 233*(1-2), 322-326.

Flavell, R. B., Bennett, M. D., Smith, J. B., & Smith, D. B. (1974). Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochemical Genetics, 12*(4), 257-269.

Frith, M. C. (2010). A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Research*. doi: 10.1093/nar/gkq1212

Gao, X., Havecker, E. R., Baranov, P. V., Atkins, J. F., & Voytas, D. F. (2003). Translational recoding signals between gag and pol in diverse LTR retrotransposons. *Rna, 9*(12), 1422-1430.

Gorinsek, B., Gubensek, F., & Kordis, D. (2004). Evolutionary Genomics of Chromoviruses in Eukaryotes. *Molecular Biology and Evolution, 21*(5), 781-798.

Grandbastien, M.-A. (1992). Retroelements in higher plants. *Trends in Genetics, 8*(3), 103-108.

Gregory, T. R. (2005). The C-value Enigma in Plants and Animals: A Review of Parallels and an Appeal for Partnership. *Annals of Botany, 95*(1), 133-146.

Han, J. S. (2010). Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mobile DNA, 1*(1), 15. doi: 10.1186/1759-8753-1-15

Havecker, E. R., Gao, X., & Voytas, D. F. (2004). The diversity of LTR retrotransposons. *Genome Biology, 5*(6), 225.

Havecker, E. R., Gao, X., & Voytas, D. F. (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 Physiology, 139*(2), 857-868.

Hawkins, J. S., Grover, C. E., & Wendel, J. F. (2008). Repeated big bangs and the expanding universe: Directionality in plant genome size evolution. *Plant Science, 174*(6), 557-562.

Higo, K., Ugawa, Y., Iwamoto, M., & Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Research, 27*(1), 297-300.

Hirochika, H., Fukuchi, A., & Kikuchi, F. (1992). Retrotransposon families in rice. *Molecular and General Genetics, 233*(1-2), 209-216.

Ilyinskii, P., & Desrosiers, R. (1998). Identification of a sequence element immediately upstream of the polypurine tract that is essential for replication of simian immunodeficiency virus. *The EMBO journal, 17*(13), 3766-3774.

Ji, G., Li, L., Li, Q. Q., Wu, X., Fu, J., Chen, G., & Wu, X. (2015). PASPA: a web server for mRNA poly(A) site predictions in plants and algae. *Bioinformatics, 31*(10), 1671-1673. doi: 10.1093/bioinformatics/btv004

Jin, Y.-K., & Bennetzen, J. L. (1989). Structure and coding properties of Bs1, a maize retrovirus-like transposon. *Proc Natl Acad Sci U S A, 86*(16), 6235-6239.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research, 110*(1-4), 462-467.

Kaushik, N., Rege, N., Yadav, P. N., Sarafianos, S. G., Modak, M. J., & Pandey, V. N. (1996). Biochemical analysis of catalytically crucial aspartate mutants of human immunodeficiency virus type 1 reverse transcriptase. *Biochemistry, 35*(36), 11536-11546.

Kenna, M. A., Brachmann, C. B., Devine, S. E., & Boeke, J. D. (1998). Invading the yeast nucleus: a nuclear localization signal at the C terminus of Ty1 integrase is required for transposition in vivo. *Molecular and Cellular Biology, 18*(2), 1115-1124.

Khaja, R., MacDonald, J. R., Zhang, J., & Scherer, S. W. (2006). Methods for identifying and mapping recent segmental and gene duplications in eukaryotic genomes *Gene Mapping, Discovery, and Expression* (pp. 9-20): Springer.

Kirchner, J., & Sandmeyer, S. (1993). Proteolytic processing of Ty3 proteins is required for transposition. *Journal of Virology, 67*(1), 19-28.

Klaver, B., & Berkhout, B. (1994). Comparison of 5' and 3' long terminal repeat promoter function in human immunodeficiency virus. *Journal of Virology, 68*(6), 3830-3840.

Klumpp, K., & Mirzadegan, T. (2006). Recent progress in the design of small molecule inhibitors of HIV RNase H. *Current pharmaceutical design, 12*(15), 1909-1922.

Kong, L., & Ranganathan, S. (2004). Delineation of modular proteins: Domain boundary prediction from sequence information. *Briefings in Bioinformatics, 5*(2), 179-192. doi: 10.1093/bib/5.2.179

Kordis, D. (2005). A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. *Gene, 347*(2), 161-173.

Kordiš, D. (2005). A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. *Gene, 347*(2), 161-173.

Kumar, A., & Bennetzen, J. L. (1999). Plant retrotransposons. *Annual Review of Genetics, 33*, 479-532.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921.

Leitch, I. J., Kahandawala, I., Suda, J., Hanson, L., Ingrouille, M. J., Chase, M. W., & Fay, M. F. (2009). Genome size diversity in orchids: consequences and evolution. *Annals of Botany, 104*(3), 469-481. doi: 10.1093/aob/mcp003

Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity, 104*(6), 520-533.

Lin, S. S., Nymark-McMahon, M. H., Yieh, L., & Sandmeyer, S. B. (2001). Integrase Mediates Nuclear Localization of Ty3. *Molecular and Cellular Biology, 21*(22), 7826-7838. doi: 10.1128/mcb.21.22.7826-7838.2001

Llorens, C., Futami, R., Bezemer, D., & Moya, A. (2008). The Gypsy Database (GyDB) of mobile genetic elements. *Nucleic Acids Research, 36*(suppl_1), D38-46. doi: 10.1093/nar/gkm697

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research, 25*(5), 0955-0964.

Lund, A. H., Duch, M., & Pedersen, F. S. (2000). Selection of functional tRNA primers and primer binding site sequences from a retroviral combinatorial library: identification of new functional tRNA primers in murine leukemia virus replication. *Nucleic Acids Research, 28*(3), 791-799. doi: 10.1093/nar/28.3.791

Ma, J., Devos, K. M., & Bennetzen, J. L. (2004). Analyses of LTR-Retrotransposon Structures Reveal Recent and Rapid Genomic DNA Loss in Rice. *Genome Research, 14*(5), 860-869.

Malik, H. S. (2005). Ribonuclease H evolution in retrotransposable elements. *Cytogenetic and Genome Research, 110*(1-4), 392-401.

Malik, H. S., & Eickbush, T. H. (1999). Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. *Journal of Virology, 73*(6), 5186-5190.

Malik, H. S., & Eickbush, T. H. (2001). Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Research, 11*(7), 1187-1197.

Marchler-Bauer, A., & Bryant, S. H. (2004). CD-Search: protein domain annotations on the fly. *Nucleic Acids Res, 32*(Web Server issue), W327-331. doi: 10.1093/nar/gkh454

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., . . . Bryant, S. H. (2013). CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res, 41*(Database issue), D348-352. doi: 10.1093/nar/gks1243

Marquet, R., Isel, C., Ehresmann, C., & Ehresmann, B. (1995). tRNAs as primer of reverse transcriptases. *Biochimie, 77*(1-2), 113-124.

Matthews, G. D., Goodwin, T. J., Butler, M. I., Berryman, T. A., & Poulter, R. T. (1997). pCal, a highly unusual Ty1/copia retrotransposon from the pathogenic yeast Candida albicans. *Journal of Bacteriology, 179*(22), 7118-7128.

McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research, 32*(suppl 2), W20-W25. doi: 10.1093/nar/gkh435

Mendivil Ramos, O., & Ferrier, D. E. (2012). Mechanisms of gene duplication and translocation and progress towards understanding their relative contributions to animal genome evolution. *International journal of evolutionary biology, 2012*.

Moore, S. P., & Garfinkel, D. J. (2009). Functional Analysis of N-terminal Residues of Ty1 Integrase. *Journal of Virology*, JVI.00159-00109. doi: 10.1128/jvi.00159-09

Mount, S. M., & Rubin, G. M. (1985). Complete nucleotide sequence of the Drosophila transposable element copia: homology between copia and retroviral proteins. *Molecular and Cellular Biology, 5*(7), 1630-1638. doi: 10.1128/mcb.5.7.1630

Nakayashiki, H. (2011). The Trickster in the genome: contribution and control of transposable elements. *Genes to Cells, 16*(8), 827-841. doi: 10.1111/j.1365-2443.2011.01533.x

Nakayashiki, H., Matsuo, H., Chuma, I., Ikeda, K., Betsuyaku, S., Kusaba, M., . . . Mayama, S. (2001). Pyret, a Ty3/Gypsy retrotransposon in Magnaporthe grisea contains an extra domain between the nucleocapsid and protease domains. *Nucleic Acids Res, 29*(20), 4106-4113.

Nanni, R., Ding, J., Jacobo-Molina, A., Hughes, S., & Arnold, E. (1993). Review of HIV-1 reverse transcriptase three-dimensional structure: Implications for drug design. *Perspectives in Drug Discovery and Design, 1*(1), 129-150. doi: 10.1007/BF02171659

Noad, R. J., Al-Kaff, N. S., Turner, D. S., & Covey, S. N. (1998). Analysis of Polypurine Tract-associated DNA Plus-strand Priming in Vivo Utilizing a Plant Pararetroviral Vector Carrying Redundant Ectopic Priming Elements. *Journal of Biological Chemistry, 273*(49), 32568-32575. doi: 10.1074/jbc.273.49.32568

Novikova, O. (2009). Chromodomains and LTR retrotransposons in plants. *Communicative and Integrative Biology, 2*(2), 158-162.

Ofran, Y., Punta, M., Schneider, R., & Rost, B. (2005). Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug discovery today, 10*(21), 1475-1482.

Ohtsubo, H., Kumekawa, N., & Ohtsubo, E. (1999). RIRE2, a novel gypsy-type retrotransposon from rice. *Genes and Genetic Systems, 74*(3), 83-91.

Orlinsky, K. J., Gu, J., Hoyt, M., Sandmeyer, S., & Menees, T. M. (1996). Mutations in the Ty3 major homology region affect multiple steps in Ty3 retrotransposition. *Journal of Virology, 70*(6), 3440-3448.

Patarca, R., & Haseltine, W. A. (1985). A major retroviral core protein related to EPA and TIMP. *Nature, 318*, 390.

Pearl, L. H., & Taylor, W. R. (1987). A structural model for the retroviral proteases.

Pellicer, J., Fay, M. F., & Leitch, I. J. (2010). The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society, 164*(1), 10-15. doi: 10.1111/j.1095-8339.2010.01072.x

Peterson-Burch, B. D., & Voytas, D. F. (2002). Genes of the Pseudoviridae (Ty1/copia Retrotransposons). *Molecular Biology and Evolution, 19*(11), 1832-1845.

Platero, J. S., Hartnett, T., & Eissenberg, J. (1995). Functional analysis of the chromo domain of HP1. *The EMBO journal, 14*(16), 3977.

Polard, P., & Chandler, M. (1995). Bacterial transposases and retroviral integrases. *Molecular Microbiology, 15*(1), 13-23. doi: 10.1111/j.1365-2958.1995.tb02217.x

Powell, M. D., & Levin, J. G. (1996). Sequence and structural determinants required for priming of plus-strand DNA synthesis by the human immunodeficiency virus type 1 polypurine tract. *Journal of Virology, 70*(8), 5288-5296.

Rawlings, N. D., & Barrett, A. J. (1995). [7] Families of aspartic peptidases, and those of unknown catalytic mechanism. *Methods in Enzymology, 248*, 105-120.

Reeves, G. A., Talavera, D., & Thornton, J. M. (2009). Genome and proteome annotation: organization, interpretation and integration. *Journal of the Royal Society Interface, 6*(31), 129-147.

Rubin, G. M. (1983). Dispersed repetitive DNAs in drosophila. *Mobile Genetic Elements, 1*, 329-362.

SanMiguel, P., & Bennetzen, J. L. (1998). Evidence that a Recent Increase in Maize Genome Size was Caused by the Massive Amplification of Intergene Retrotransposons. *Annals of Botany, 82*(Supplement A), 37-44.

SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., . . . Bennetzen, J. L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science, 274*(5288), 765-768.

Schultz, S. J., & Champoux, J. J. (2008). RNase H activity: structure, specificity, and function in reverse transcription. *Virus Research, 134*(1), 86-103.

Schultz, S. J., & Champoux, J. J. (2008). RNase H activity: structure, specificity, and function in reverse transcription. *Virus Research, 134*(1-2), 86-103. doi: S0168-1702(07)00460-1 [pii]
10.1016/j.virusres.2007.12.007

Staginnus, C., Desel, C., Schmidt, T., & Kahl, G. (2010). Assembling a puzzle of dispersed retrotransposable sequences in the genome of chickpea (Cicer arietinum L.). *Genome, 53*(12), 1090-1102. doi: g10-093 [pii]
10.1139/g10-093

Steinbiss, S., Willhoeft, U., Gremme, G., & Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Research, 37*(21), 7002-7013. doi: 10.1093/nar/gkp759

Suoniemi, A., Tanskanen, J., Pentikainen, O., Johnson, M. S., & Schulman, A. H. (1998). The core domain of retrotransposon integrase in Hordeum: predicted structure and evolution. *Molecular Biology and Evolution, 15*(9), 1135-1144.

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

Tanskanen, J. A., Sabot, F., Vicient, C., & Schulman, A. H. (2007). Life without GAG: The BARE-2 retrotransposon as a parasite's parasite. *Gene, 390*(1-2), 166-174.

Toh, H., Ono, M., Saigo, K., & Miyata, T. (1985). Retroviral protease-like sequence in the yeast transposon Ty 1. *Nature, 315*, 691.

van Opijnen, T., Kamoschinski, J., Jeeninga, R. E., & Berkhout, B. (2004). The Human Immunodeficiency Virus Type 1 Promoter Contains a CATA Box Instead of a TATA Box for Optimal Transcription and Replication. *Journal of Virology, 78*(13), 6883-6890. doi: 10.1128/jvi.78.13.6883-6890.2004

Vitte, C., & Panaud, O. (2005). LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenetic and Genome Research, 110*(1-4), 91-107.

Whitney, K., Baack, E., Hamrick, J., Godt, M., Barringer, B., Bennett, M., . . . Leitch, I. (2010). A role for nonadaptive processes in plant genome size evolution? *Evolution, 64*, 2097-2109.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., . . . Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet, 8*(12), 973-982.

Wilhelm, M., Uzun, O., Mules, E. H., Gabriel, A., & Wilhelm, F. X. (2001). Polypurine tract formation by Ty1 RNase H. *Journal of Biological Chemistry, 276*(50), 47695-47701.

Xiong, Y., & Eickbush, T. H. (1988). Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. *Molecular Biology and Evolution, 5*(6), 675-690.

Youngren, S. D., Boeke, J., Sanders, N., & Garfinkel, D. (1988). Functional organization of the retrotransposon Ty from Saccharomyces cerevisiae: Ty protease is required for transposition. *Molecular and Cellular Biology, 8*(4), 1421-1431.

Yuan, Y., SanMiguel, P. J., & Bennetzen, J. L. (2003). High-Cot sequence analysis of the maize genome. *The Plant Journal, 34*(2), 249-255.

Zhao, X., & Wang, H. (2007). LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons (2007). *Nucleic Acids Res, 35*, W265-W268.

Appendix 1

Pseudocode for Blast Automation

The following modules will be included

strict; Bio::SeqIO; Bio::PrimarySeqI; Bio::Root::Root; Bio::Tools::Run::RemoteBlast; Bio::SearchIO; Data::Dumper;

Initialize variables

Get filename and blast option such as blastx or tblastx through command line variable argv[1] and argv[2] respectively.

Set file to argv[1]

Set limit to 200

Set i to 101

Set j to 300

Set string to null

Set b to 0

Set k to 1

Set prog to argv[2]

Set db to nr

Set eval to 1e-10

Initialize the array params by passing the values in prog, db, eval, and searchIO

Instantiate the object RemoteBlast by passing the values in params and set it to remoteBlast

Instantiate the object SeqIO by passing the value in the variable file and 'GenBank' for format and set it to the variable $in

Call the method nextseq() in SeqIO which gets sequences, and then assign the result to the variable seqObj

Call the method length () in SeqIO to get the length of the sequence in the file and store the result in a variable len

Store the file length -100 in a variable in order to blast junction sequences where previously 200 bases were blast searched at a time

Get the modulus of 200 for the file length and check it there is a remainder, if there is a remainder then, one more iteration of the loop will be needed to complete the blast search for the entire file

Store the value got as a result of dividing the length of the file by 200 to a variable named times in order to iterate the loop that many times, also if modulus is greater than zero times variable has to be incremented by 1

The preceding block of code will determine the number of 200bp present in the sequence. Thereby setting the number of times the tblastx will be called since the search is made for 200bp at a time

Iterate through a while loop as long as the variable times is greater than zero

Then, in a for loop starting from 101 bp stored in a variable i, iterate the sequence file in 200 bp at a time, incrementing i with 200bp every time.

The 200bp string that is obtained has to be initialized as a primary sequence object for further handling of the sequence in bioperl modules.

Next call the submit_blast method passing the primary sequence object as a parameter in the remoteBlast module

In a while loop store the result id of the blast result in an array and within the while loop in a foreach loop iterate the result id array

Within the foreach loop call the method retrieve blast in remoteBlast module and store the result in a variable to check if the method fetched a result id for the call, if the result id is less than zero remove the result id by calling the method remove id in remoteBlast module else call the get the result by using the method next_result () and store it in an .out file

Call the save_output () method in remoteBlast to write the results to the .out file

Keep decrementing times within the while loop and exit the loop if times equal zero

For parsing the sequence file from the beginning instead of from the 100th base the same algorithm will be used expect the file will parsed from the first base pair.

Pseudocode for Sorting Blast Hits

Include the module File::Copy which copies directories and files

Accept command line argument and store it in the variable $d

Print the variable on the screen to check value stored in the variable $d

In the variable $dir store the path of the directory

Open the directory; if the directory cannot be opened print “cannot open file”

If the directory can be opened print “hello”

Initialize the array named @filesarray to store all the filenames from the directory

Iterate through the filesarray in a foreach loop and assign the filename to variable $file

If u do not get hit push the array if u do get a hit store the filename in a new array named final array and push the filesarray

Now when you have the finalarray iterate the final array in a foreach loop till the array is not empty and store the filename in the new directory

Opendir DH and pass the path as the second parameter

Else exit the loop and print “cannot open directory”

If directory is opened copy the file into the directory

Do clean up

Close dir BIN

Print the file name to ensure the file has been fetched

Within the loop perform the following function

Open a filehandle FH and pass the filename as a second parameter to the Open operator.

If the file cannot be opened, print “file cannot be opened”

Iterate through the file in a while loop and store each line in a variable $line

If $line has the value “No significant similarity found” print “no hit”

Algorithm for Blast Automation1

```
use strict;
use Bio::SeqIO;
use Bio::PrimarySeqI;
use Bio::Root::Root;
use Bio::Tools::Run::RemoteBlast;
use Bio::SearchIO;
use Data::Dumper;


my $file        = $ARGV[1];
my $limit       = 200;

my $i           = 101;
my $j           = 300;
my $string="";
my $b=0;
my $k=1;

# Set the parameters for blast and get blastx/tblastx option through command line argv[2]
my $prog = $ARGV[2];
my $db = "nr";
my $e_val = "1e-10";

my @params = ( '-prog' => $prog,
       '-data' => $db,
       '-expect' => $e_val,
       '-readmethod' => 'SearchIO' );

my $remoteBlast = Bio::Tools::Run::RemoteBlast->new(@params);



### OBJECT INSTANTIATION
my $in = Bio::SeqIO->new(
                        -file     => $file,
                        -format=> 'GenBank',
                        );
```

```
my $seqObj=$in->next_seq();
my $len=$seqObj->length();
my $evenlen=$len-100;
print $len,"\n";
my $remainder=$evenlen%200;#modulus
print $remainder,"\n";
my $times=($evenlen-$remainder)/200;
if($remainder!=0 ){
                    $times++;
}
print $times,"\n";



while($times>0){

          for ($i=101;$string=$seqObj->subseq($i,$j);$i=$i+200){

          my $seq = Bio::PrimarySeq->new( -seq =>$string );
          my $r = $remoteBlast->submit_blast($seq);
 my $v = 1;

   print STDERR "waiting..." if( $v > 0 );  ########  WAIT FOR THE RESULTS TO
RETURN!!!!!

 while ( my @rids = $remoteBlast->each_rid ) {
     foreach my $rid ( @rids ) {
       my $rc = $remoteBlast->retrieve_blast($rid);
       if( !ref($rc) ) {
         if( $rc < 0 ) {
           $remoteBlast->remove_rid($rid);
         }
         print STDERR "." if ( $v > 0 );
         sleep 5;
       } else {
         my $result = $rc->next_result();
         #save the output
        # my $filename = $result->query_name()."\.out";
            my $filename = $file."_tbx_".$e_val."_even_".$i."\.out";
                    $remoteBlast->save_output($filename);
         $remoteBlast->remove_rid($rid);
         #print "\nQuery Name: ", $result->query_name(), "\n";
                    print "\nQuery Name: ", $k, "\n";
```

```
        while ( my $hit = $result->next_hit ) {
          next unless ( $v > 0);
          print "\thit name is ", $hit->name, "\n";
          while( my $hsp = $hit->next_hsp ) {
            print "\t\tscore is ", $hsp->score, "\n";
          }
        }
      }
    }
  }



$k++;
print $i,"\n";
print $j,"\n";
$j=$j+200;
print $string,"\n";
$times--;
print $times,"times\n";
          if (($times==1)&& ($remainder!=0) ){
                    print "i before changes", $i,"\n";
                    $j=$len;
                    print "has remainder";
                    print"i is ",$i ,"\n";
                    print "j is ",$j ,"\n";

          }
          if ($times==0){
          exit;
          }


}

}



#undef is eof next_seq() returns nextseq or undef

=pod
for(my $i=$limit+1; my $seq=$in->next_seq(); $i++){
```

```
  if($i>=$limit){ $i=0;$seqO = Bio::SeqIO->new(-file=>">$file.chomp".(++$j).".fasta", -
format=>'Fasta')}
  $seqO->write_seq($seq);
print "hi";
}
```

if modulus is zero loop num-- else loop num+1--

when modulus is not zero
when num==1 set j to remainder;

=cut

Algorithm for Blast Automation2

```
use strict;
use Bio::SeqIO;
use Bio::PrimarySeqI;
use Bio::Root::Root;
use Bio::Tools::Run::RemoteBlast;
use Bio::SearchIO;
use Data::Dumper;


my $file        = $ARGV[1]; #FBac001-BACforwardcontig.gbank
my $limit       = 200;

my $i           = 1;
my $j           = 200;
my $string="";
my $b=0;
my $k=1;

#Here set the parameters for blast and get blastx/tblastx option through command line
my $prog = $ARGV[2]; #tblastx
my $db = "nr";
my $e_val = "1e-10";
```

```
my @params = ( '-prog' => $prog,
        '-data' => $db,
        '-expect' => $e_val,
        '-readmethod' => 'SearchIO' );

my $remoteBlast = Bio::Tools::Run::RemoteBlast->new(@params);

### OBJECT INSTANTIATION
my $in = Bio::SeqIO->new(
                            -file     => $file,
                            -format=> 'GenBank',
                            );
my $seqObj=$in->next_seq();
my $len=$seqObj->length();
print $len,"\n";
my $remainder=$len%200;
print $remainder,"\n";
my $times=($len-$remainder)/200;
if($remainder!=0 ){
                    $times++;
}
print $times,"\n";



while($times>0){

          for ($i=1;$string=$seqObj->subseq($i,$j);$i=$i+200){

          my $seq = Bio::PrimarySeq->new( -seq =>$string );
          my $r = $remoteBlast->submit_blast($seq);
 my $v = 1;

   print STDERR "waiting..." if( $v > 0 );  ########  WAIT FOR THE RESULTS TO
RETURN!!!!!

 while ( my @rids = $remoteBlast->each_rid ) {
     foreach my $rid ( @rids ) {
       my $rc = $remoteBlast->retrieve_blast($rid);
       if( !ref($rc) ) {
         if( $rc < 0 ) {
           $remoteBlast->remove_rid($rid);
         }
```

```
        print STDERR "." if ( $v > 0 );
        sleep 5;
       } else {
        my $result = $rc->next_result();
        #save the output
       # my $filename = $result->query_name()."\.out";
          my $filename = $file."_tbx_".$e_val."_odd_".$i."\.out";
                $remoteBlast->save_output($filename);
        $remoteBlast->remove_rid($rid);
        #print "\nQuery Name: ", $result->query_name(), "\n";
                print "\nQuery Name: ", $k, "\n";

        while ( my $hit = $result->next_hit ) {
          next unless ( $v > 0);
          print "\thit name is ", $hit->name, "\n";
          while( my $hsp = $hit->next_hsp ) {
            print "\t\tscore is ", $hsp->score, "\n";
          }
        }
       }
      }
    }



$k++;
print $i,"\n";
print $j,"\n";
$j=$j+200;
print $string,"\n";
$times--;
print $times,"times\n";
          if (($times==1)&& ($remainder!=0) ){
                print "i before changes", $i,"\n";
                $j=$len;
                print "has remainder";
                print"i is ",$i ,"\n";
                print "j is ",$j ,"\n";

          }
          if ($times==0){
          exit;
          }
```

```
}

}


#undef is eof next_seq() returns nextseq or undef

=pod
for(my $i=$limit+1; my $seq=$in->next_seq(); $i++){
  if($i>=$limit){ $i=0;$seqO = Bio::SeqIO->new(-file=>">$file.chomp".(++$j).".fasta", -
format=>'Fasta')}
  $seqO->write_seq($seq);
print "hi";
}

if modulus is zero loop num-- else loop num+1--

        when modulus is not zero
        when num==1 set j to remainder;

=cut
```

Program to Sort Blast Hits.

```
use File::Copy;

$dir="c:/perl/tbx/F_imp_rpl16";
opendir(BIN,$dir)||die ("cannot open dir");
print BIN."hello";
my @filesarray = grep { -T "$dir/$_" } readdir BIN;


        foreach $file (@filesarray){

                print $file,"\n";
                open(FH,$dir."/".$file) || die "cannot open file";
```

```
                        while($line=<FH>){

                                if ($line =~m/^<b>No significant similarity found/){
                                        print "no hit\n";
                                        unshift(@nohitsarray,$file);
                                        my $arraysize= @nohitsarray;
                                        print "no hits array size ".$arraysize."end\n";

                                }


                        }
        }

        foreach $nohit(@nohitsarray){
                print $nohit."nohitfile\n";

        }

my $filesarraysize= @filesarray;
print "allfilesarraysize is".$filesarraysize."\n";


# compare two arrays

my @finalarray;
OUTER:
for ( @filesarray ) {

         for my $nohit ( @nohitsarray ) {
                        if ( / \Q$nohit/ ) {
                        next OUTER;
                                }
                        }
 push @finalarray, $_;
}

foreach $final (@finalarray){

        print $final."finallist\n";
        opendir(DH,"c:/perl/tbx/hits") || die "directory cannot be opened";
        copy($dir."/".$final,"c:/perl/tbx/hits/".$final) || die "$file
                cannot be copied";
```

```
}


closedir(BIN);
closedir(DH);
```

Appendix 2

Conserved Domains Database results

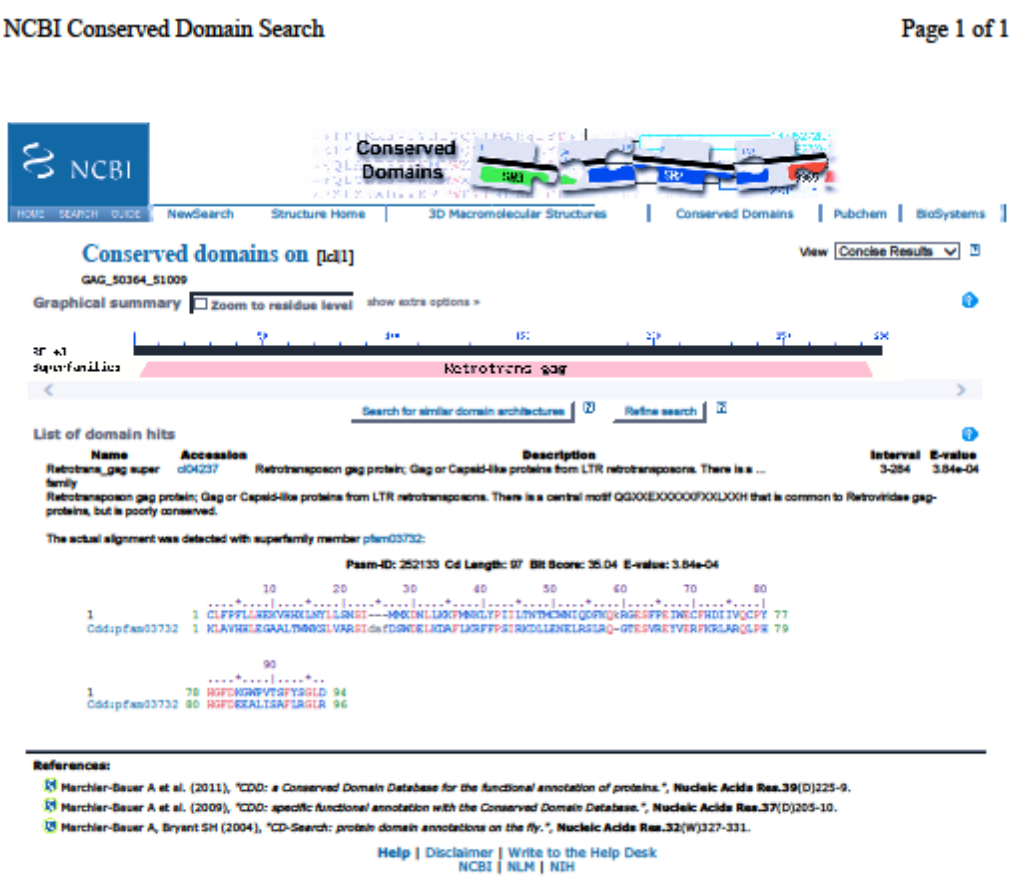

Fig. A1. Ty3/Gypsy gag in the region 50364bp to 51009bp

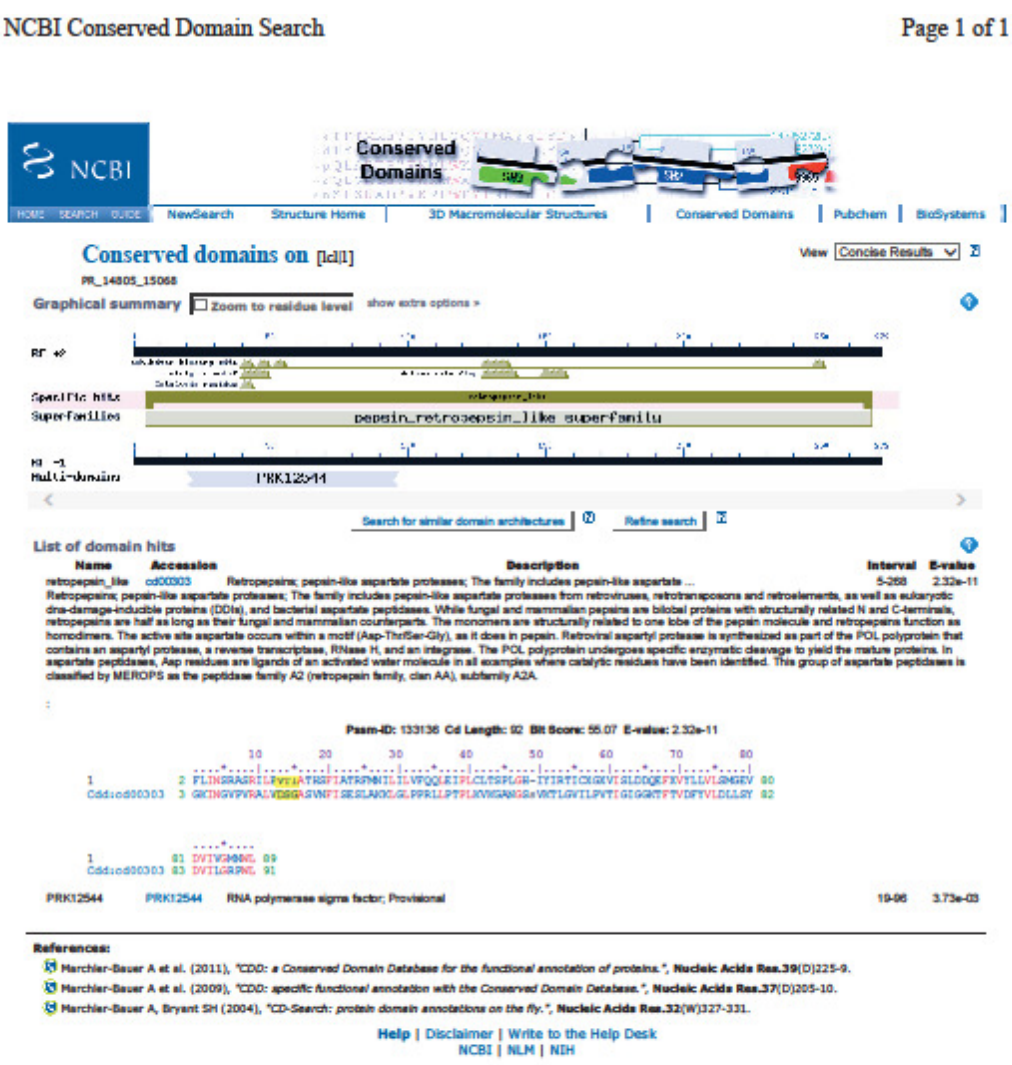

Fig. A2. Protease in Gypsy element in the region 14805bp to 15068bp

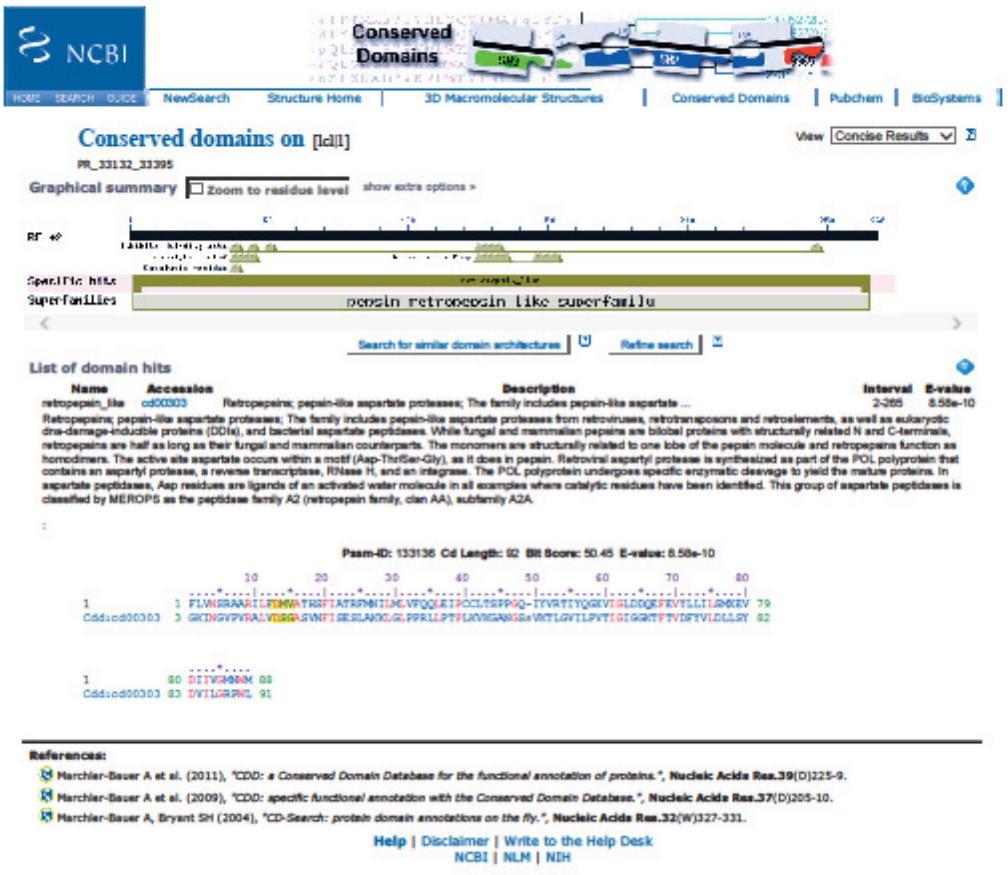

Fig. A3. Protease in Gypsy element in the region 33132bp to 33395bp

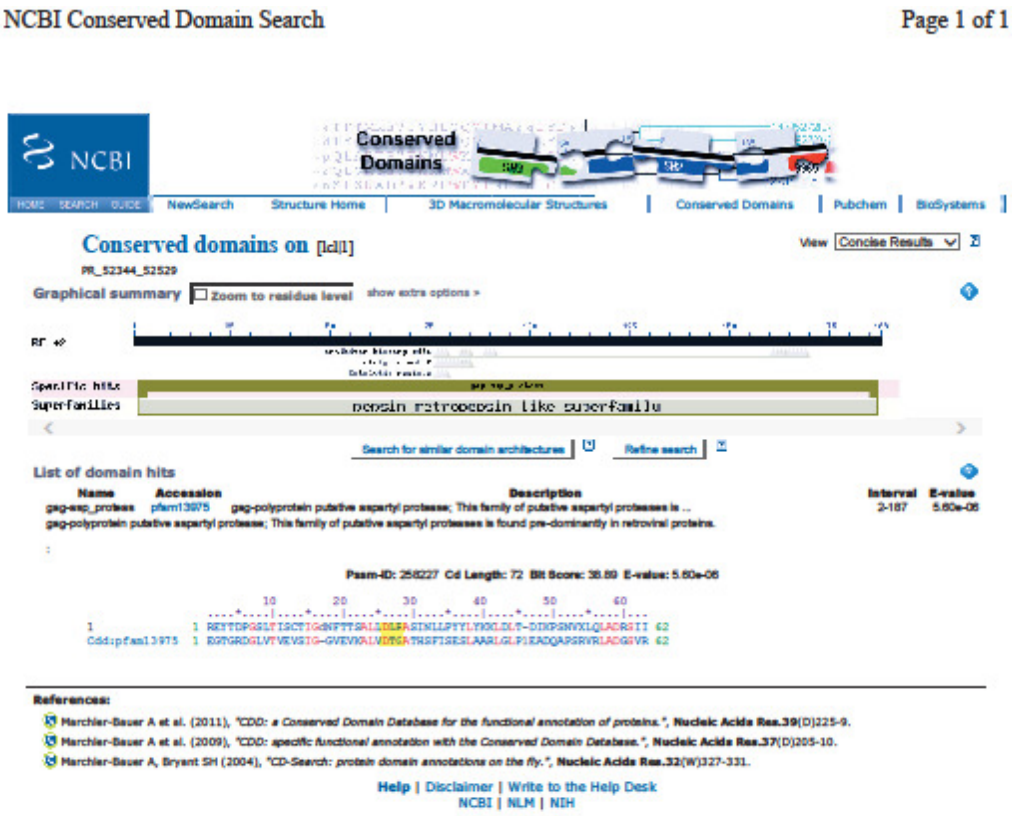

Fig. A4. Protease in Gypsy element in the region 52344bp to 52529bp

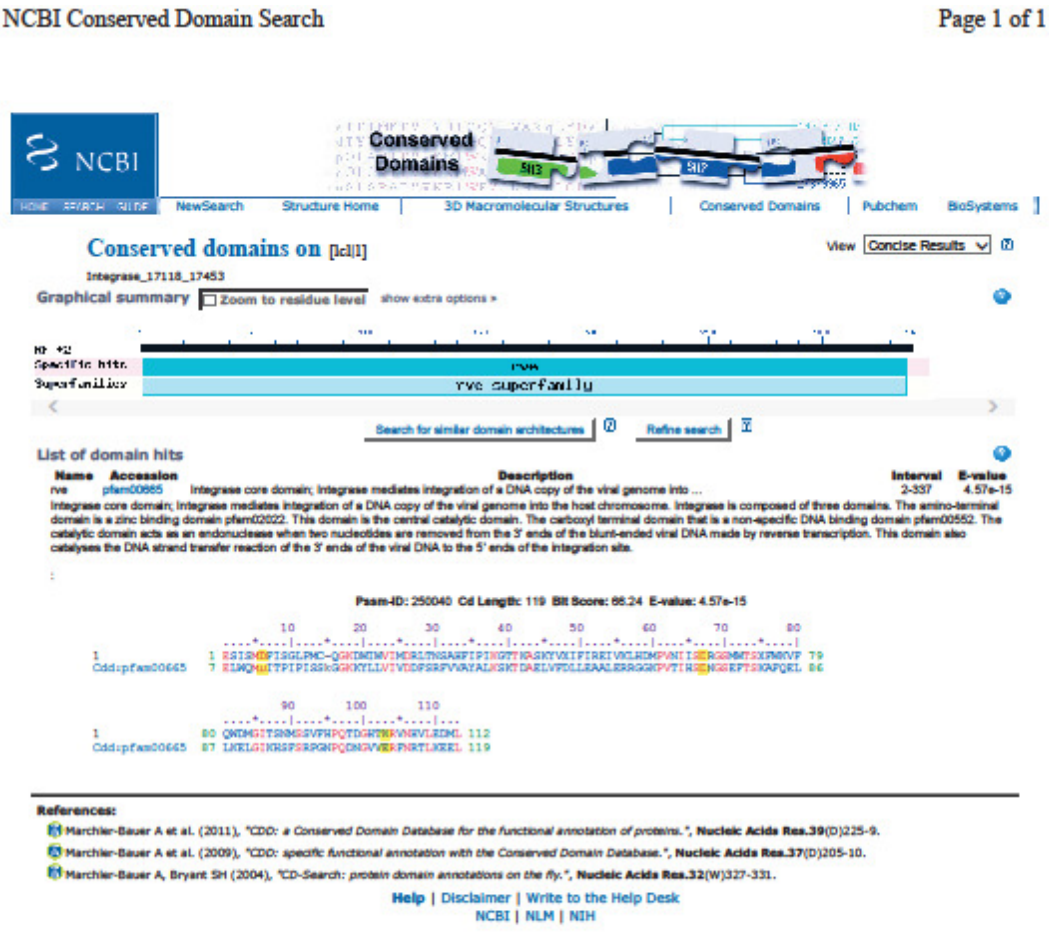

Fig. A5. Integrase core domain in gypsy element in the region 17118bp to 17453bp

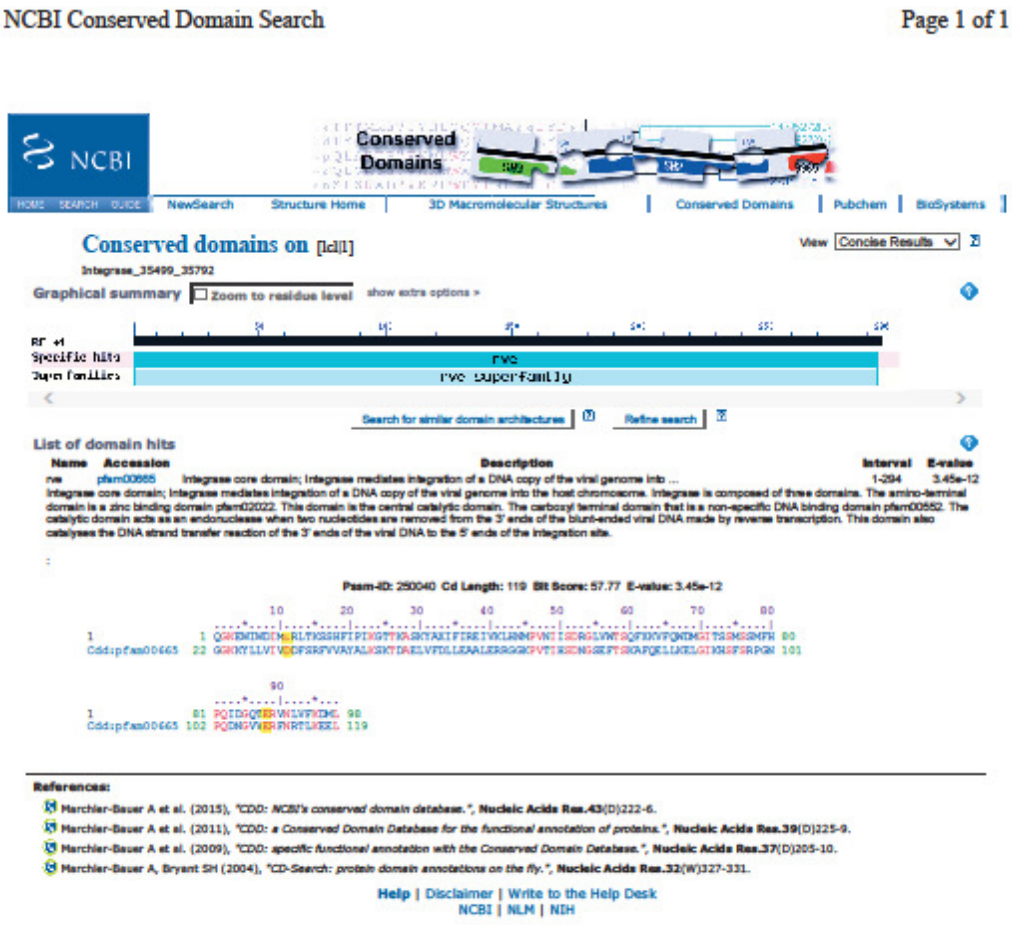

Fig. A6. Integrase core domain in gypsy element in the region 35499bp to 35792bp

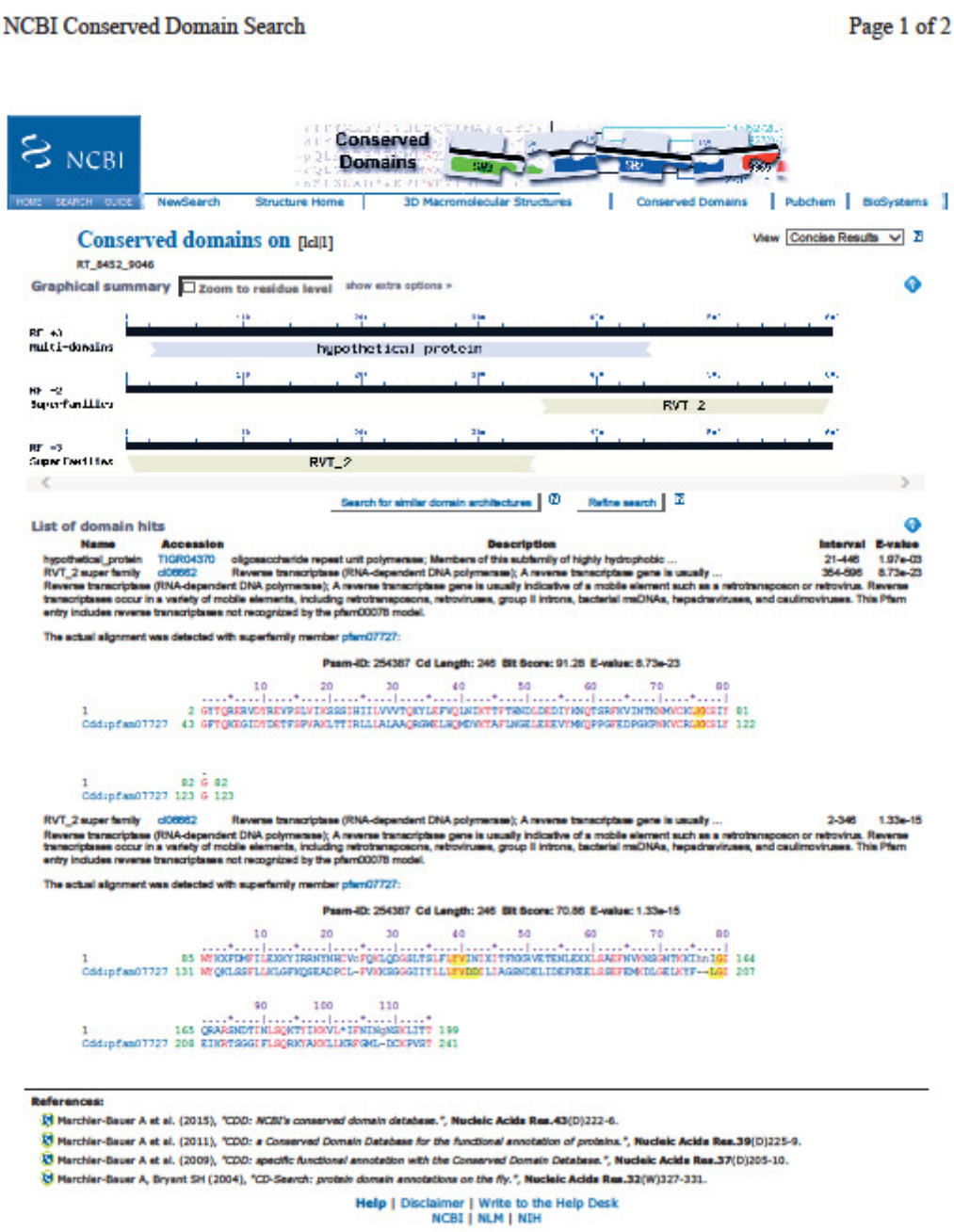

Fig. A7. Copia reversetranscriptase in the region 8452bp to 9043bp

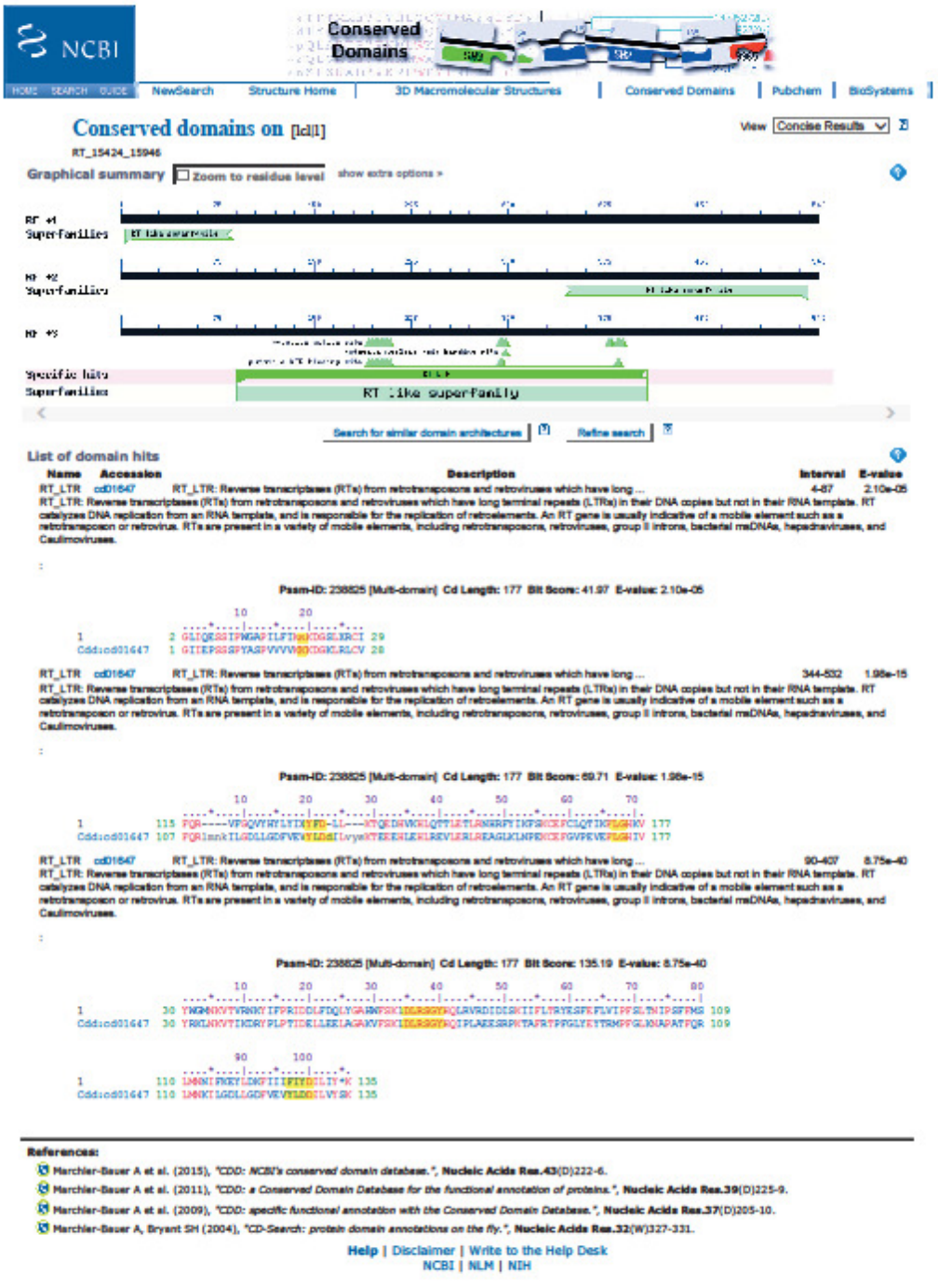

Fig A8. Gypsy reversetranscriptase in the region 15424bp to 15945bp

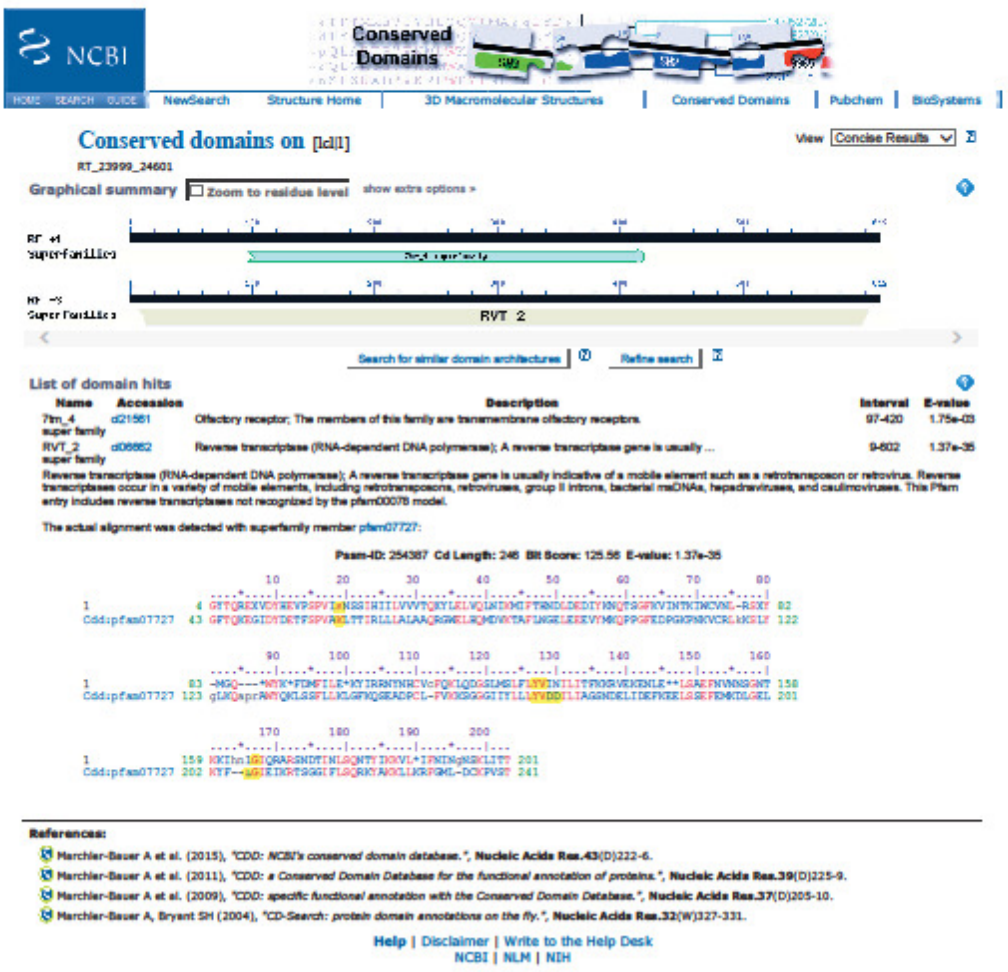

Fig. A9. Copia reversetranscriptase in the region 23999bp to 24601bp

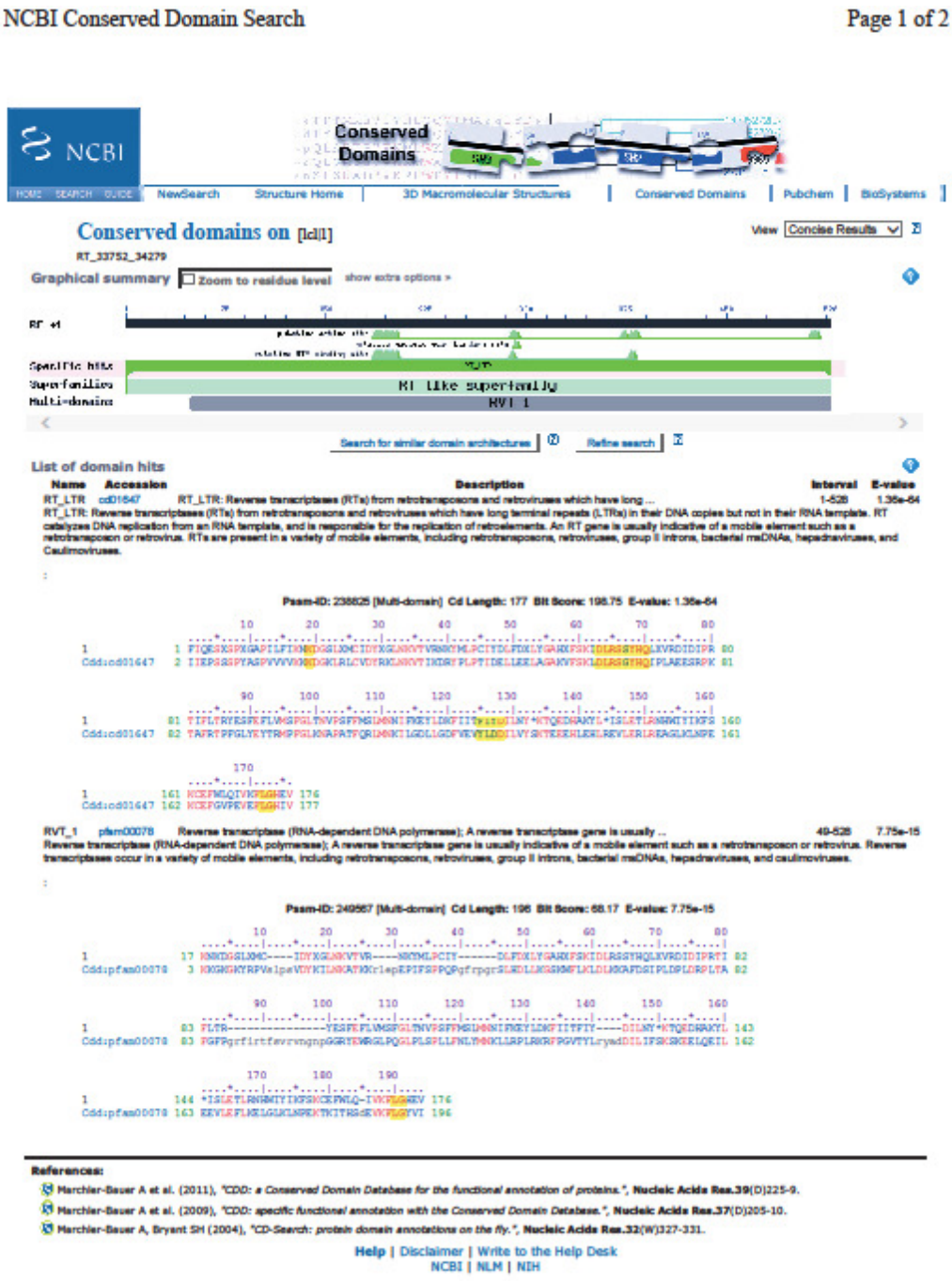

Fig. A10. Gypsy reversetranscriptase in the region 33752bp to 34279bp

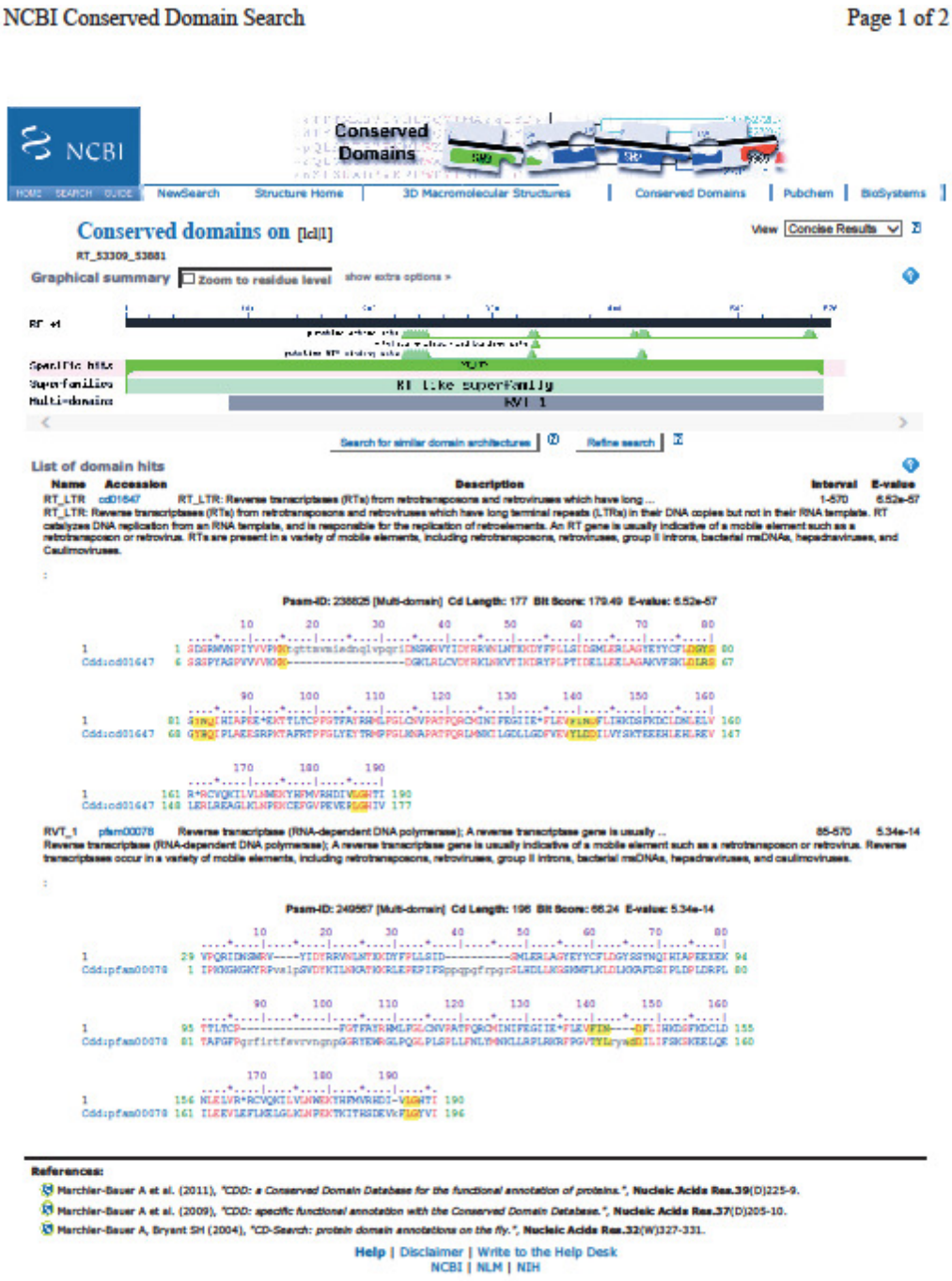

Fig A11. Gypsy reversetranscriptase in the region 53309bp to 53881bp

Fig A12. Copia RNase H in the region 7735bp to 8105bp

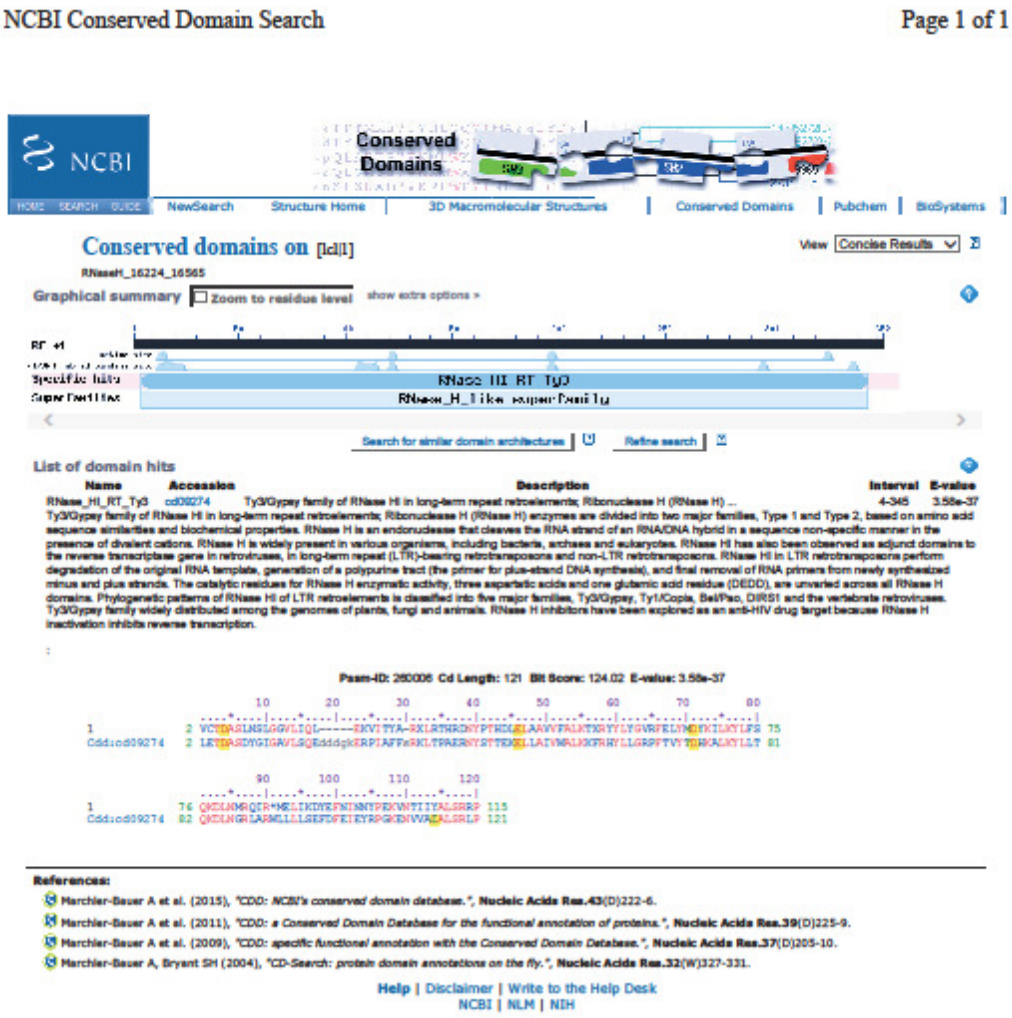

Fig A13. Gypsy RNase H in the region 16224bp to 16565bp

Fig. A14. Copia RNase H in the region 23353bp to 23723bp

NCBI Conserved Domain Search Page 1 of 1

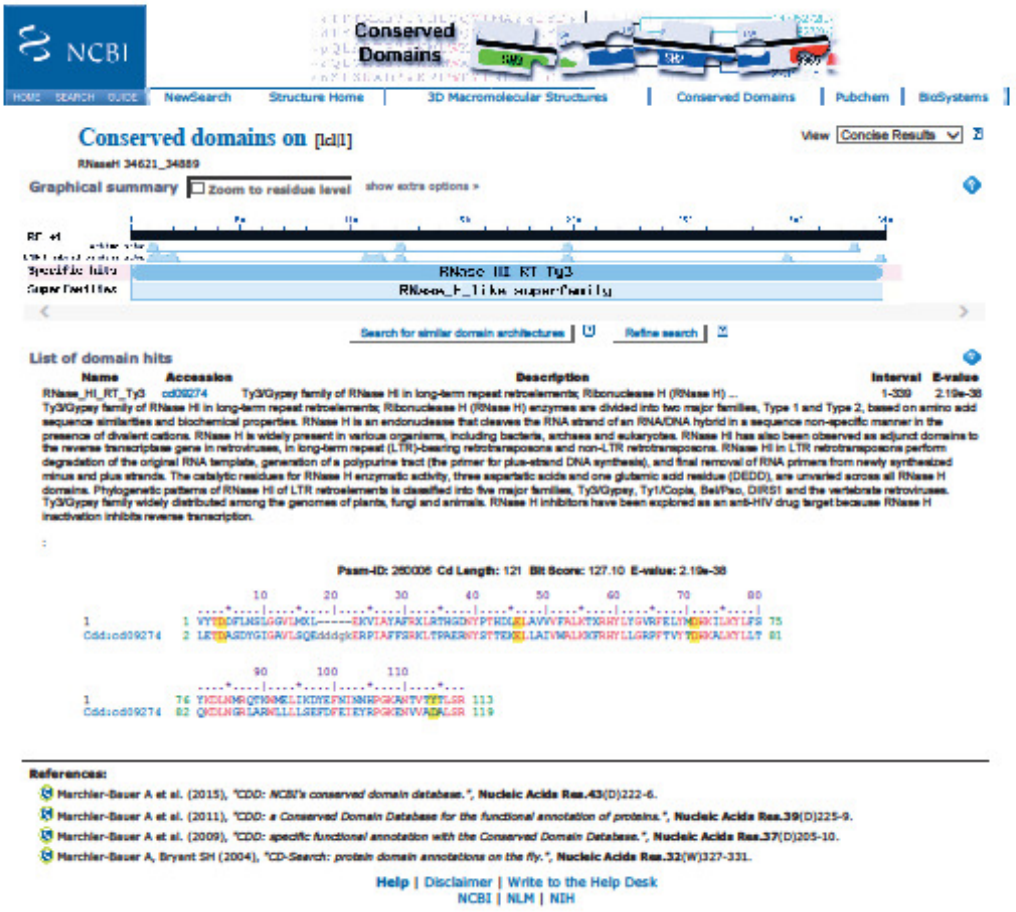

Fig. A15. Gypsy RNase H in the region 34621bp to 34889bp.

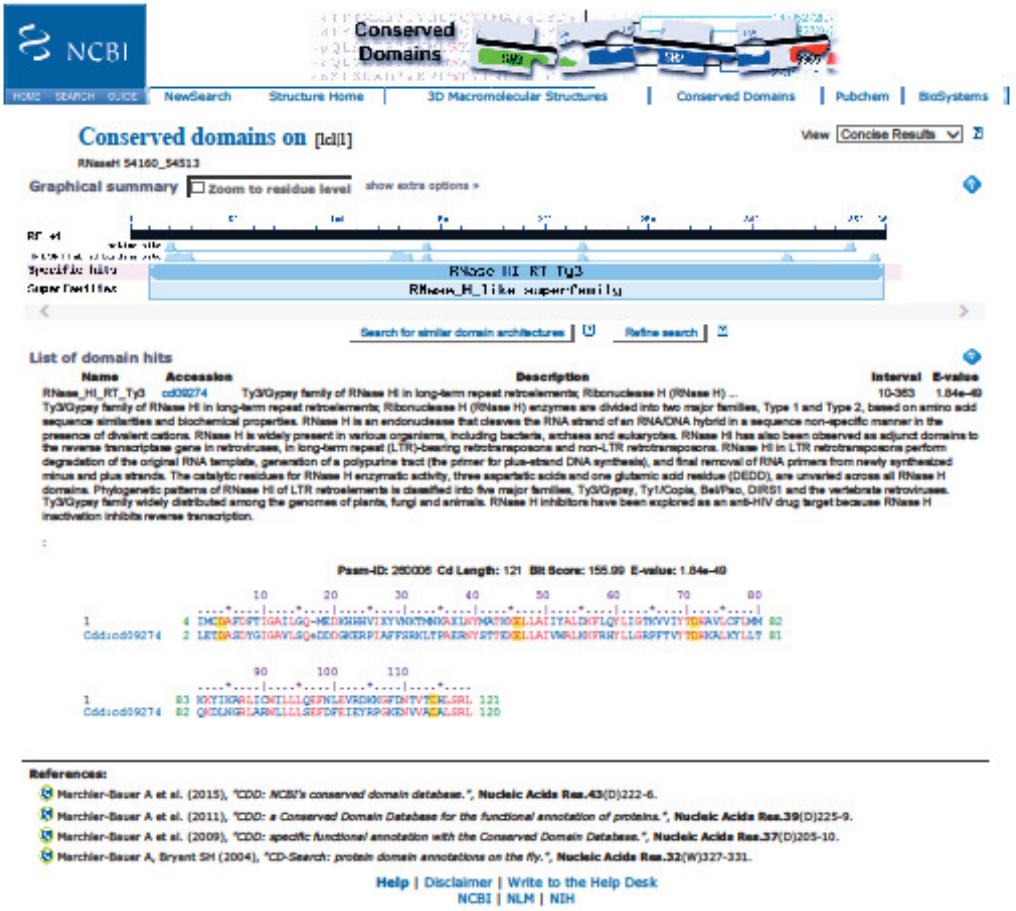

Fig. A16. Gypsy RNase H in the region 54160bp to 54513bp

Appendix 3

Sequin Table

```
>Features FBac001
Fritillaria
1     54802 source
                Organism   Fritillaria agrestis
                Mol_type   genomic DNA
                Note   Simple repeats identified using Repeat
Masker.
279   352   repeat_region
                      rpt_unit_seq    (TA)n
                      rpt_type   satellite
2436  2529  repeat_region
                rpt_unit_seq    (CA)n
                rpt_type   satellite
2530  2545  repeat_region
                Rpt_unit_seq    (TA)n

6257  6272  repeat_region
                Rpt_unit_seq    (TA)n
                rpt_type   satellite
<7735 9043> mobile_element
                mobile_element_type   remnant
retrotransposons:Ty1/copia-like

<7735 >8105 Region
                Region_name     RNase_H_like
                Db_xref         CDD:cd09272
                Note       RNase H like domain with frame shift
and truncated at the 3' end of the
domain with the last conserved residue [DE] missing.
8167  8262  repeat_region
                Rpt_unit_seq    (AT)n
                rpt_type   satellite
<8450 >9043 Region
                Region_name          RVT_2
                Note  reverse_transcriptase domain (RNA-
dependent DNA polymerase) with frame shift.
                Db_xref         CDD: pfam07727
10757 10769 repeat_region
                Rpt_unit_seq    (G)n
                Rpt_type   Satellite
11809 11868 repeat_region
                Rpt_unit_seq    (TA)n
                Rpt_type   Satellite
11891 11894 repeat_region
                Rpt_type   flanking
                Note  target site duplication
11895 20250 mobile_element
                mobile_element_type
retrotransposons:Ty3/gypsy-like
```

```
                Note  degenerated retrotransposon with frame
shifts
11895 13837 LTR
                Rpt_type    terminal
                Note  left long terminal repeat
<14805     >15068     Region
                Region_name     retropepsin_like
                Db_xref     CDD:cd00303
                Note       protease domain
<15414     >15944     Region
                Region_name      RT_LTR_like
                Note        Reversetranscriptase domain with
three frame shifts
                Db_xref     CDD:cd01647
<16224     >16565     Region
                Region_name      RNase_HI_like
                Note        RNase H domain
                Db_xref     CDD:cd09274
<17118     >17454     Region
                Region_name      rve
                Note        Integrase domain
                Db_xref     CDD:pfam00665
18211 18223 misc_feature
                Note  potential polypurine tract
18323 20250 LTR
                Rpt_type    terminal
                Note  right long terminal repeat
20210 20215 polyA_signal
                Note  potential canonical polyA signal AATAAA
20275 20278 repeat_region
                Rpt_type    flanking
                Note  target site duplication
21801 21888 repeat_region
                Rpt_unit_seq  (AT)n
                Rpt_type  Satellite
<23353     24580>     mobile_element
                Mobile_element_type
retrotransposons:Ty1/copia-like
                Note  remnant retrotransposon with frame shifts
and no structural features
<23353     >23723     Region
                Region_name      RNase_H_like
                Db_xref     CDD:cd09272
                Note       RNase H domain with a frame shift and
truncated at the 3' end missing the
last conserved residue [D/E]
23782 23811 repeat_region
                Rpt_unit_seq     (AT)n
                Rpt_type    Satellite
```

```
<23995     >24580     Region
                region_name      RVT_2
                Note   Reversetranscriptase domain
                Db_xref     CDD:pfam07727
28707 28720 repeat region
                Rpt_unit_seq     (G)n
                Rpt_type  Satellite
29772 29825 repeat_region
                Rpt_unit_seq  (AT)n
                Rpt_type  Satellite
29826 29861 repeat_region
                Rpt_unit_seq  (GT)n
                Rpt_type  Satellite
29876 38225 mobile_element
                Mobile_element_type
retrotransposons:Ty3/gypsy-like
29876 31448 LTR
                Rpt_type  flanking
                Note  left long terminal repeat
<33132     >33395     Region
                region_name      retropepsin_like
                Note   peptidase family A2
                Db_xref     CDD:cd00303
<33752     >34279     Region
                region_name     RT_LTR/RVT_1
                Note   reversetranscriptase domain
                Db_xref     CDD:cd01647;pfam00078

<34621     >34889     Region
                Region_name      RNase_HI
                Db_xref     CDD:cd09274

<35499     >35792     Region
                Region_name      rve
                Note   Integrase core domain
                Db_xref     CDD:pfam006655
36653 38225 LTR
                Rpt_type    terminal
                Note   right long terminal repeat
37327 37333 polyA_signal
                Note  potential canonical polyA signal AATAAA
39090 39106 repeat region
                Rpt_unit_seq     (G)n
                Rpt_type  Satellite
39529 39576 repeat_region
                Rpt_unit_seq  A-rich

41314 41377 repeat_region
                Rpt_unit_seq     A-rich
```

```
<50364     54513>     mobile_element
                Mobile_element_type   retrotransposons:Ty3/gypsy-
like
                Note  Truncated retrotransposon element

<50364     >51009     Region
                Region_name       retrotrans_gag
                Note   gag or capsid like protein domain
                Db_xref     CDD:pfam03732
<52344     >52529     Region
                Region_name       gag-asp_protease
                Note   pepsin_retropepsin_like superfamily
                Db_xref     CDD:pfam13975
<53309     >53881     Region
                Region_name       RT_LTR/RVT_1
                Note   reversetranscriptase domain
                Db_xref     CDD:cd01647/pfam00078

<54160     >54513     Region
                Region_name       RNase_HI
                Db_xref     CDD:cd09274
                Note       RNase H domain
```