# INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

**The quality of this reproduction is dependent upon the quality of the copy submitted.** Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

# Studying the replication mechanism of the yeast retrotransposon Ty5 by molecular and computational approaches

by

Xiang Gao

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Majors: Molecular Cellular and Developmental Biology;
Bioinformatics and Computational Biology

Program of Study Committee:
Dr. Daniel F. Voytas, Co-major Professor
Dr. Leslie G. Miller, Co-major Professor
Dr. Linda Ambrosia
Dr. Susan Carpenter
Dr. Eric Henderson

Iowa State University

Ames, Iowa

2001

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

**Graduate College**

**Iowa State University**

**This is to certify that the doctoral dissertation of**

**Xiang Gao**

**has met the dissertation requirements of Iowa State University**

Signature was redacted for privacy.

**Co-major Professor**

Signature was redacted for privacy.

**Co-major Professor**

Signature was redacted for privacy.

**For the Co-major Program**

Signature was redacted for privacy.

**For the Co-major Program**

*For my parents,*

*Zhaoyuan Gao and Jinge Gan*

*and my sister Hui Gao*

*and for Qunfeng Dong.*

# TABLE OF CONTENTS

# ABSTRACT

The yeast retrotransposon Ty5 is a Ty1/*copia* element. Officially, it is in the *Hemivirus* genus of the *Pseudoviridae* family. The ability to genetically manipulate retrotransposons and the yeast host cell was taken advantage of to explore replication mechanisms unique to Ty5 and common to most retrotransposons. Because of the abundance and diversity of retroelement sequences, along with the fact that many retroelement enzymes have evolved unique functional specificities, computational approaches were also developed to study functional divergence in replication. By screening a randomly mutagenized Ty5 library, two mutations (Y68C, D252N) that caused higher transposition frequencies were identified. Both mutations increased Ty5 cDNA levels, but did not have dramatic effects on the steps after cDNA synthesis (i.e. integration and recombination), or protein synthesis, processing, or solubility. The D252N mutation increased the hydrogen bonding potential of the CCHC zinc finger of nucleocapsid protein (NCp), making the Ty5 NCp zinc finger more like Ty1/*copia* consensus zinc fingers in terms of hydrogen bonding potential. Other mutations that increased the hydrogen bonding potential (D252R, D252K) provided the same fold increase in Ty5 transposition. These results suggest that NCp and its CCHC domain play an important role in Ty5 reverse transcription, and natural occurring mutations in the Ty5 zinc finger repress this function. Hydrogen bonding is suggested to be a universal requirement for the function of retroviral type zinc fingers and cellular zinc fingers. A half-tRNA priming mechanism for Ty5 reverse transcription was also demonstrated.

Mutations in the anticodon of $tRNA_i^{Met}$ (IMT) and the putative PBS of Ty5 decreased transposition, but transposition was restored when complementarity between the IMT and PBS was restored. A tree-based method and supplemental *Split Tester* software were developed to study the functional divergence of reverse transcriptase (RT) with respect to half-tRNA and full-tRNA priming mechanisms. The domains identified by this computational approach were previously experimentally demonstrated to bind with the tRNA primer/template in HIV RT. Using this software, another domain related to integrase functional specificity, namely whether or not integrase carries out 3'-end processing during integration, was also consistently identified in different integrase datasets. A model describing this functional divergence is proposed.

# CHAPTER I. GENERAL INTRODUCTION

## Taxonomy of transposable elements

Transposable elements are discrete sequences in the genome, which can transport themselves directly to other locations without requiring sequence homology (Berg and Howe, 1989). Transposable elements are distributed in a wide variety of organisms, from bacteria to humans, and they are present in remarkable abundance. Because of their mobility, they perform important functions in many cellular processes, such as gene transfer, genome rearrangements, genetic regulation and telomere maintenance.

There are two types of transposable elements that differ according to their replication process. DNA transposons replicate through a DNA intermediate and include elements such as bacterial Tn elements, the maize *Ac* element and the *Drosophila melanogaster* P element (O'Hare and Rubin, 1983; Pohlman et al., 1984). At least one strand of the original DNA transposon will be transferred to a new target site in the chromosome. In contrast, retroelements use RNA as an intermediate in replication. The original retroelement DNA template is not excised from the chromosome. Instead, the complementary DNA (cDNA) copy, which is the reverse transcription product of the retroelement RNA, is inserted into other sites of the chromosome. Retroelements can be divided into two groups based on the comparisons of more than 80 reverse transcriptase (Poch et al., 1989; Xiong and Eickbush, 1990). One group is the non-LTR retrotransposons (retroposons) in eukaryotes, group II introns in bacteria and the Mauriceville plasmid of mitochondria. Non-LTR retrotransposons

do not have long terminal repeats but share a poly(A) tail at their 3'-end. They are diverse in structure and distribution, and include the I factor in *D. melanogaster* (Fawcett et al., 1986), R2 in many insects (Jakubczak et al., 1991), and long interspersed nuclear elements (LINEs) in mammals (Hattori et al., 1986). The second group of retroelements includes retroviruses and LTR-retrotransposons. Retroviruses are the most extensively studied, because of their importance in human health and disease. LTR retrotransposons are remarkably widespread and abundant in eukaryotes. According to their genome organization (see below), LTR retrotransposons can be divided into the Ty3/*gypsy* family (*Metaviridae*) and Ty1/*copia* family (*Pseudoviridae*). Ty3 (Hansen et al., 1988) in *Saccharomyces cerevisiae*, Tf1 in *Schizosaccharomyces pombe* and *gypsy* in *D. melanogaster* are extensively studied Ty3/*gypsy* elements. Ty1 (Farabaugh and Fink, 1980), and Ty5 (Zou et al., 1995) in *S.cerevisiae* and *copia* (Fouts et al., 1981) in *D. melanogaster* are several model elements of the Ty1/*copia* family. This thesis will focus on the LTR retrotransposons.

**Retroviruses and LTR retrotransposons**

Retroviruses and LTR retrotransposons are functionally and genetically analogous (Fig1)(Boeke and Sandmeyer, 1991; Brown and Varmus, 1989). Both are flanked by direct long terminal repeats (LTR), which are identical in sequence. The LTR sequences are divided into 3 regions (U3, R, and U5), based on transcription start and stop sites. U3 contains enhancer and promoter sequences for mRNA expression; R is a repeated

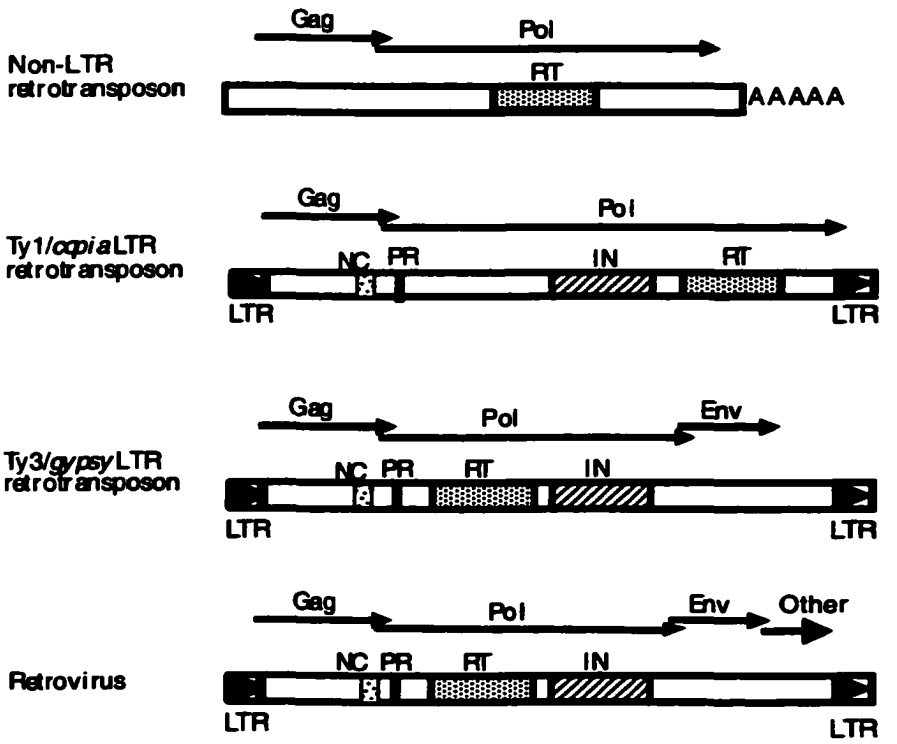



Fig 1. Genetic and structural comparisons between retroelements. The consensus structures for each group is used for comparison. Five A's in non-LTR retrotransposons represent the poly(A) tails. The staggered arrows are different open reading frames (ORF). Shaded boxes represent conserved gene subunits. Nucleocapsid (NC) protein is located at the 3'end of Gag. Reverse transcriptase (RT), integrase (IN) and protease (PR), in most cases, are in Pol. Envelope (ENV) appears in retroviruses and some Ty1/*copia* and Ty3/*gypsy* group retrotransposons. Other ORFs are accessory genes in different retroviruses. The boxes and lines are not drawn to scale.

sequence at both ends of the mRNA that is required for strand transfer in reverse transcription; signals are found in U5 for transcript termination and polyadenylation.

*Genome organization* There are three genes common to all retroviruses: *gag, pol* and *env*. *gag* encodes a polyprotein, which is processed into the 3~4 virus structural proteins: matrix (MA), capsid (CA), and nucleocapsid (NCp) and p9. Together these proteins assemble into virons. The polyprotein encoded by *pol* is the precursor of protease (PR), reverse transcriptase (RT) and integrase (IN). *PR* is related to cellular aspartate proteases and is responsible for proteolytic processing of the primary translation product of *gag* and *pol* and the maturation of the viron (Dougherty and Semler, 1993). *env* encodes the envelope glycoprotein, which is required for virus budding and infection. Other regulatory and accessory genes are encoded by diverse retroviruses. They play roles in transcriptional regulation, RNA/cDNA transport, etc.

LTR retrotransposon have *gag* and *pol* analogous genes and the functions of these genes are the same as for the retroviruses. Some *env*-like genes are found in certain lineages of the Ty3/*gypsy* and Ty1/*copia* families. The position, length and the presence of a transmembrane domain support that this ORF might have an *env*-like function. In some cases, i.e. *gypsy*, this ORF is glycosylated and cleaved like retroviral *env* proteins. Furthermore, this ORF allows *gypsy* to infect *D. melanogaster*, providing evidence that it functions as a retrovirus Env (Song et al., 1994; Song et al., 1997). Therefore, *env*-containing

retrotransposons are thought to be endogenous retroviruses. In the LTR retrotransposons, Ty1/*copia* and Ty3/*gypsy* families are distinguished by the order of RT and IN in *pol*. The order of PR-IN-RT in the Ty3/*gypsy* family is the same as the retroviruses, while the order of IN and RT is reversed in the Ty1/*copia* family.

*Open reading frame organization* In retroviruses and retrotransposons, the Gag and Pol polyproteins are translated from one mRNA transcript. Gag and Pol are typically encoded on separate ORFs, separated by a stop codon or frameshift. The relative expression level of Gag and Pol are important for the correct assembly of the viron or virus-like particle (VLP). Typically, more Gag than Pol is needed to assemble the VLP. The ratio of these two ORF products is regulated by several different mechanisms. The most common strategy is -1 translational frameshifting, such as in human immunodeficiency virus (HIV) (Jacks, 1990) and +1 frameshifting as in the yeast retrotransposons Ty1-Ty4 (Voytas and Boeke, 1993). Stop codon suppression is used by Murine leukemia virus (MuLV) to make a Gag-Pol fusion protein (Alam et al., 1999). *copia* elements rely on a post-transcriptional splicing strategy to remove the *pol* sequences from a majority of *gag-pol* transcripts (Brierley and Flavell, 1990; Miller et al., 1989), thereby resulting in an excess of Gag. Tf1 uses a posttranslational mechanism to degrade Pol proteins preferentially (Atwood et al., 1996). In addition to Gag and Pol, in some retroviruses, protease exists as separate ORF. *env* in retroviruses and *env*-like genes in Ty3/*gypsy* and Ty1/*copia* family are located downstream of *pol* and are typically expressed from a spliced genomic mRNA.

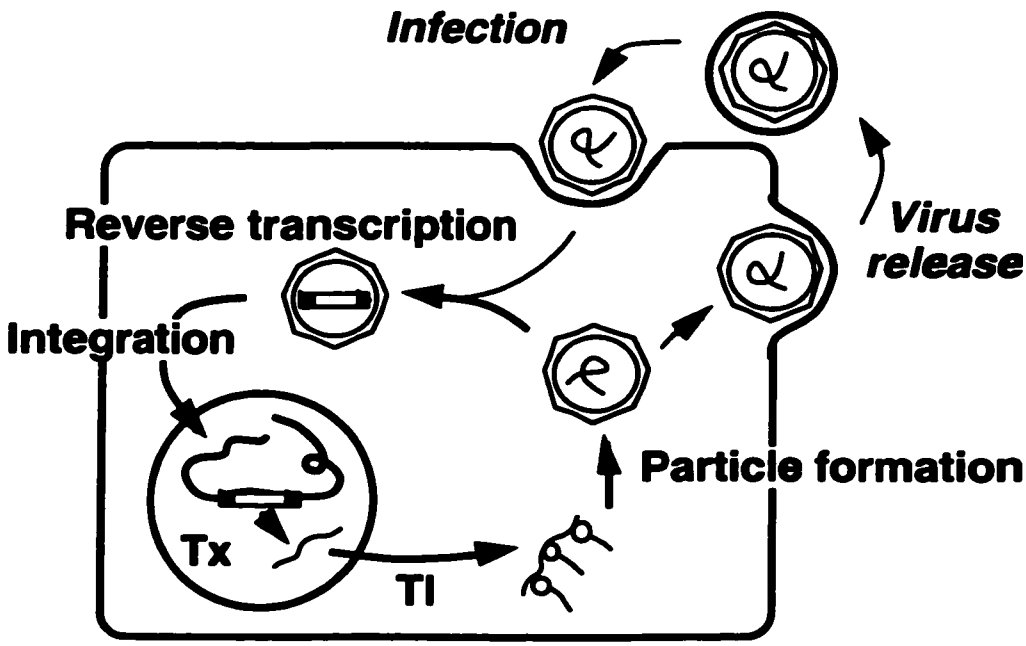



Fig 2. Life cycle of retrotransposons and retroviruses. Retrotransposons use the inner cycle, whereas the retroviruses take the larger loop, which involves release and infection. Tx, transcription. Tl, translation.

*Replication life cycle* LTR retroelements share a very similar life cycle (Fig2). RNA and protein are transcribed and translated from the retroelement in the host genome. In the cytoplasm, retroelement-encoded proteins and their RNA assemble into virons or VLPs. In these particles, double stranded cDNA is synthesized through reverse transcription. Eventually, cDNAs of retroelements integrate into the host chromosome to complete the life cycle. The retroviruses can leave the host and infect other cells, but retrotransposons can not. The similarity between these two groups of retroelements, and the fact retrotransposons are found in model organisms, such as yeast, make retrotransposons an ideal model system to study the mechanisms of their replication.

**Overview of reverse transcription**

Retroelement replication involves a complicated series of biochemical reactions that require the reverse transcription and RNaseH activities associated with RT. As shown in Fig 3, reverse transcription initiates with the annealing of a unwound tRNA primer to the primer binding site (PBS) in the retroelement RNA template. This unwinding process can not occur without help from RT and NCp (Chan and Musier-Forsyth, 1997; Lapadat-Tapolsky et al., 1995; Remy et al., 1998). The RNA-dependent DNA polymerase activity of RT extends DNA synthesis from the 3'-end of the tRNA primer to the 5' end of the genomic RNA. Because PBS is close to the 5'LTR, the accumulated short minus strands are termed minus strand strong stop DNA (-ssDNA). During -ssDNA synthesis, RNaseH activity degrades

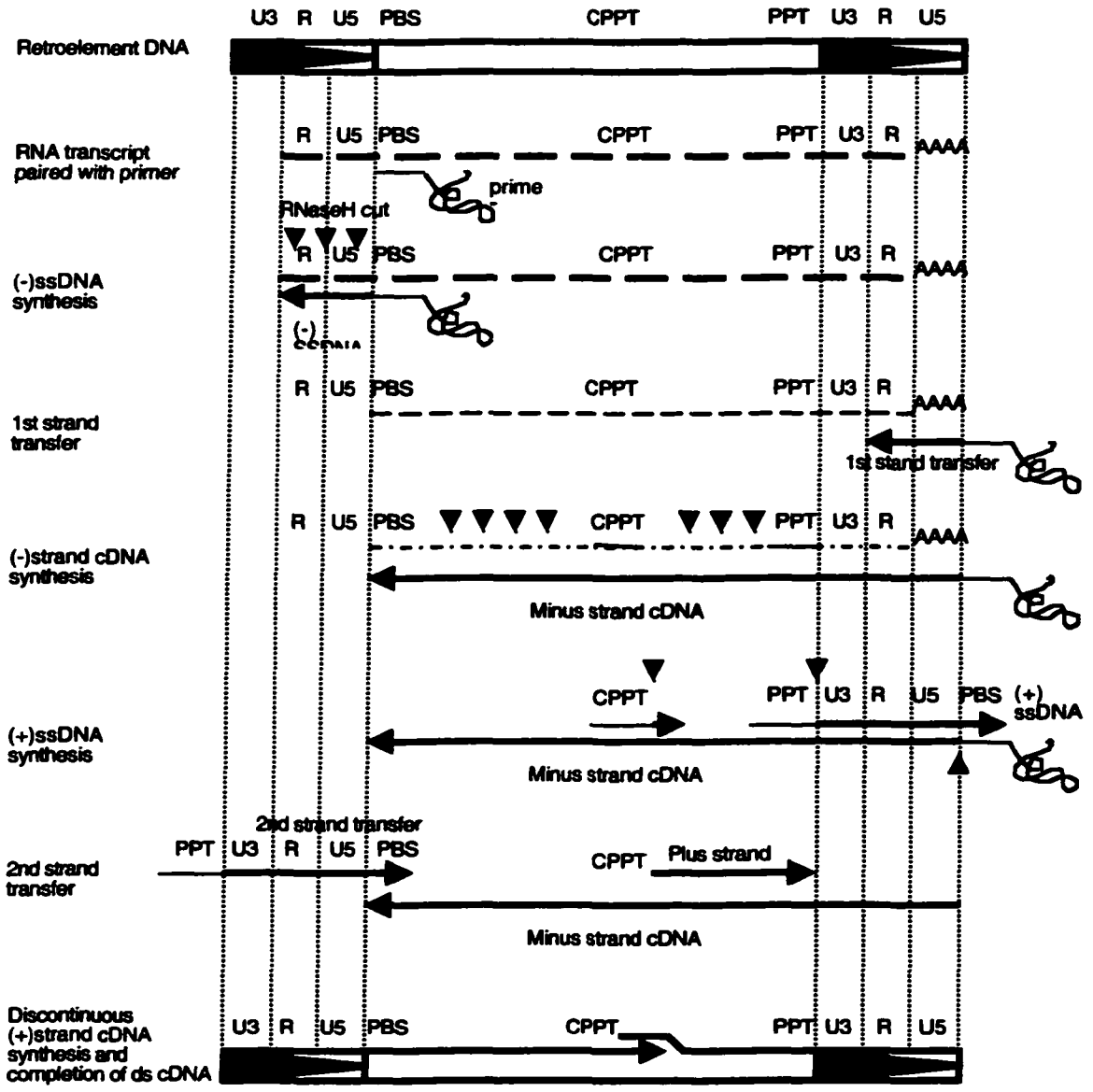



Fig 3. Reverse transcription of retroelement RNA into double stranded cDNA. RNA is represented by dotted or thin lines. cDNA is represented by bold lines. The boxes with arrowheads at the termini of DNA or cDNA denate long terminal repeats (LTR). A brief description of each step is listed on the left margin. PBS, primer binding sequence; PPT, polypurine tract; CPPT, central polypurine tract (adapted from Wilhelm, Cell. Mol. Life Sci. 58 (2001) 1246-1262.)

the RNA template in the newly formed RNA-DNA hybrid and releases the -ssDNA. Because the R region is present at both ends of the RNA template, the terminal region of -ssDNA is complementary to R in the 3'-end of the template. It might be because of this sequence complementary that minus ssDNA is transferred from the 5'-end to the 3'-end of the RNA template. The pairing to the template allows synthesis of minus strand DNA to resume. As minus strand synthesis progresses, the RNA template is incompletely digested by RNaseH activity. Specifically, polypurine tract (PPT) sequences are resistant to RNaseH activity and function as primers for plus strand synthesis. There are two PPTs: a central PPT and one close to the 3'LTR. In all retroelements, the strong stop plus strand DNA (+ssDNA) is initiated from the PPT located near the 3'LTR. Through sequence complementarity, +ssDNA is transferred to the 5'-end of the template. In Ty1 (Lauermann and Boeke, 1997) and several lentiviruses, the central PPT is also used to synthesize the second half of plus strand DNA. The plus strand DNA synthesis transferred at the 5'-end of RNA will stop elongation shortly after it reaches the second half of plus strand DNA initiated from the central PPT. The overlap between the two halves of plus strand DNA is termed the central DNA flap, which in HIV has been demonstrated to function in nuclear import (Zennou et al., 2000).

**Retroelement proteins involved in reverse transcription**

Reverse transcription takes place in the nucleocore complex in the cytoplasm. Three retroelement-encoded proteins have been identified so far that are involved in priming: RT, NCp and IN.

*Reverse transcriptase* RT has both DNA polymerase activity and RNaseH activity, but they are in separate domains of the enzyme. DNA polymerase activity can be divided into RNA dependent DNA polymerase activity and DNA dependent DNA polymerase activity. All of these activities are required in reverse transcription. RT can form a complex with the primer tRNA and facilitate unwinding of the tRNA with the help of NCp (Barat et al., 1993; Isel et al., 1995; Isel et al., 1999; Mishima and Steitz, 1995). RT activity is needed to coordinate all steps in reverse transcription, including strong stop DNA synthesis and strand transfer, RNA degradation, as well as completion of the minus and plus strands. Sequence analysis of RT revealed seven conserved domains found in all retroelement RTs (Xiong and Eickbush, 1988). YXDD is the motif that defines the active site and is required for polymerase activity (Boyer et al., 1992; Larder et al., 1987). Four residues in RNaseH have been identified that are conserved among viral and nonviral. Three of these residues are important for catalytic activity of the bacterial and retroviral enzymes (Kanaya et al., 1990; Mizrahi et al., 1990; Repaske et al., 1989). HIV RT is processed into two different forms: p66 and p51, the latter of which lacks RNaseH. These two forms of RT form a heterodimer. p66 is the subunit that performs all functions, while p51in important in maintaining structure and

probably has tRNA primer binding activity. Because RTs of retrotransposons have only recently been expressed *in vitro*, they have not been studied as extensively as the retroviral enzymes. The published results suggest that the enzymatic activity of Ty1 and Ty3 RT are similar to retroviral homologues (Cristofari et al., 1999; Rausch et al., 2000; Wilhelm et al., 2000).

*Nucleocapsid protein* NCp is the primary protein in the retroelement nucleocore. NCp binds tightly to both the genomic RNA of retroelements and to their tRNA primers (Barat et al., 1993). Binding is carried out by one or two highly conserved CCHC type zinc fingers (Vogt, 1997). The conserved sequence of retroviral zinc fingers is $CX_2CX_4HX_4C$. In Ty5, the consensus finger motif differs slightly from most retroelements ($CX_2CX_3HX_4C$ vs. $CX_2CX_3GHX_4C$). Basic amino acids flanking the zinc fingers also interact with template and primer RNAs. The requirement for the zinc finger and the surrounding basic region is different in different retroelements. Some retroelements like Ty1 of *S. cerevisiae* do not have a conserved zinc finger; rather, three stretches of basic amino acids in the C terminus of Gag perform the required nucleic acid chaperon activity (Cristofari et al., 2000). Thus, although there are exceptions, the use of zinc fingers is the most widespread means of interacting with nucleic acids. The nucleic acid binding activity of NCp is important for a number of steps in replication, including RNA dimerization (Barat et al., 1993; Feng et al., 1996; Prats et al., 1988), primer and template RNA packaging (Berkowitz et al., 1996), annealing of the tRNA primer to the template RNA (Chan and Musier-Forsyth, 1997; Lapadat-Tapolsky et al.,

1995; Remy et al., 1998), initiating reverse transcription (Cristofari et al., 2000; Rong et al., 1998), transferring strong stop DNA (Allain et al., 1994; Cristofari et al., 2000; Darlix et al., 1993; Hsu et al., 2000) and ensuring fidelity of cDNA synthesis (Gorelick et al., 1999).

*Integrase* Integrase is reported for both retrotransposon and retroviruses to be essential for efficient reverse transcription. In Ty3, deletion of the C-terminus and mutations in both the N- and C- terminus of IN have severe effects on the amount of cDNA associated with VLPs (Kirchner and Sandmeyer, 1996; Nymark-McMahon and Sandmeyer, 1999). The hypothesis based on these observations is that Ty3 polymerase might be an RT/RT-IN heterodimer and that the interaction between RT and IN has an important role in reverse transcription (Note that RT-IN indicates an unprocessed form of the Pol polyprotein). In Ty1, RT retains full activity in the RT-IN intermediate and cDNA production is at about wild type levels (Merkulov et al., 2001). In HIV, a direct interaction between the mature IN and RT was observed, and the mature IN was suggested to be essential for the efficient initiation of reverse transcription (Wu et al., 1999).

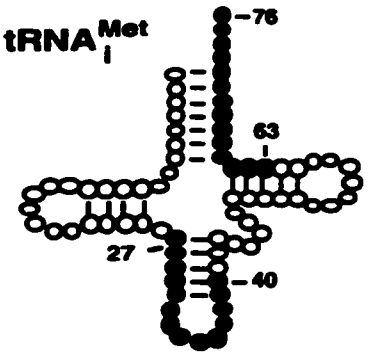
**Diversity in priming LTR retroelement reverse transcription**

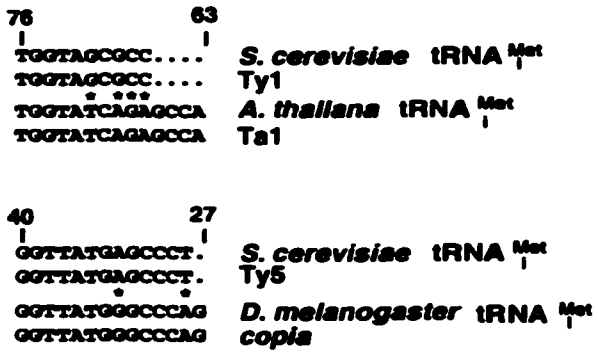
Like other polymerases, reverse transcriptase needs a primer to initiate the elongation reaction. The basic requirement for the primer is that it provides a free 3'-OH group onto which RT can transfer nucleotides and elongate the cDNA. Reverse transcription in retroviruses and LTR retrotransposons is initiated by a specific tRNA primer that pairs to

the primer binding site (PBS) of the RNA template. Based on the region of the tRNA that is paired to the template, two tRNA priming mechanisms were identified. For retroviruses and most retrotransposons, the 3'end of the tRNA (acceptor stem) pairs to the PBS of the RNA template (Chapman et al., 1992; Leis et al., 1993). Another novel mechanism is proposed for the *copia* retrotransposon of *Drosophila melanogaster* (Kikuchi et al., 1986). For the *copia* element, the anticodon stem-loop of the initiator methionine tRNA (IMT) pairs to the PBS. The initiator tRNA is cleaved in half, and cDNA extension occurs from the 3'-OH group of the tRNA cleavage product (Fig. 4). Other retrotransposons that likely share the same priming mechanism exist among a variety of organisms: 1731 from *D. melanogaster*, *Osser* from *Volvox carteri*, Tp1 from *Physarnm polycephalum* and Ty5 from *Saccharomyces cerevisiae* (Fourcade-Peronnet et al., 1988; Lindauer et al., 1993; McCurrach et al., 1990; Rothnie et al., 1991; Voytas and Boeke, 1992). These five retrotransposons are classified into a separate *genus* of the *Pseudoviridae* family, called the Hemiviruses, which are characterized by using a half-tRNA priming mechanism.

A tRNA is not the only primer used by LTR-retrotransposons. Tf1, a Ty3/*gypsy* element, undergoes self-primed reverse transcription, using a unique cleaved fragment of its RNA genome. The first 11 bases of the 5' end of the mRNA pair with the PBS, and RNaseH performs cleavage at the 12$^{th}$ base. This cleavage creates a 3'-OH group at the end of the 11 bases, allowing it to function as a primer (Levin, 1995; Levin, 1996). The PBS in all LTR retroelements is located near the 5' LTR. The length of PBS in retrotransposons varies from 8-18 bases. Other sequences in retroelement RNA were found to interact with the primer

A




B

| Retroviruses | Primer | Length of PBS |
|---|---|---|
| *S. cerevisiae retrotransposons* | | |
| Ty1/Ty2 | 3'end of tRNA$_i^{Met}$ | 10 |
| Ty3 | 3'end of tRNA$_i^{Met}$ | 8 |
| Ty4 | 3'end of tRNA$^{Asn}$ | 18 |
| Ty5 | anticodon of tRNA$_i^{Met}$ | 14 |
| *Drosophila retrotransposons* | | |
| 17.6/297/tom | 3'end of tRNA$^{Ser}$ | 18 |
| mdg1, 412 | 3'end of tRNA$^{Arg}$ | 18 |
| copia | anticodon of tRNA$_i^{Met}$ | 14 |
| *Pararetroviruses* | | |
| CaMV | 3'end of tRNA$_i^{Met}$ | 13 |
| *Plant retrotransposons* | | |
| Tnt1 | 3'end of tRNA$_i^{Met}$ | 12 |
| *Retroviruses* | | |
| HIV-1 | 3'end of tRNA$_3^{Lys}$ | 18 |
| HFV | 3'end of tRNA$^{Lys}$ | 18 |
| ASLV | 3'end of tRNA$^{Trp}$ | 18 |
| MuLV | 3'end of tRNA$^{Pro}$ | 18 |
| *gypsy* | 3'end of tRNA$^{Lys}$ | 11 |

Fig 4. The diversity of tRNA primers used in retroelement reverse transcription. A) Sequences in the 3' acceptor stem and anticodon stem-loop of tRNA$_i^{Met}$ are shown paired with PBS of Ty1, Ta1, Ty5, and *copia*. B) List of tRNA primers of selected retroviruses and retrotransposons (Wilhelm, Cell. Mol. Life Sci. 58 (2001) 1246-1262).

tRNA in several retroelements (Friant et al., 1998; Friant et al., 1996; Gabus et al., 1998; Isel et al., 1995; Isel et al., 1999; Keeney et al., 1995). The extensive interaction between the tRNA and RNA template helps stabilize the primer/template complex. This extensive binding might form a special structure for cognate RT to recognize. For example, $tRNA^{lys}$/HIV-1 RNA complex can only be recognized by RT from HIV-1, AMV and SIV, but not from HIV-2, FIV, EIAV, MLV, although they all share $tRNA^{lys}$ as a primer (Feuerbach et al., 1997).

**Integration and recombination of cDNA**

*Integration* After reverse transcription, a blunt-ended double stranded linear cDNA is the precursor to integration. IN, which specifically recognizes the LTR-end sequences, associates with RT (Lee and Coffin, 1991), NCp (Lapadat-Tapolsky et al., 1993), MA (Bukrinsky et al., 1993) and probably host factors (Farnet and Bushman, 1997; Lee and Craigie, 1994) to form the preintegration complex. For most retroelements, before the complex enters the nucleus, integrase cleaves the 3' termini of cDNA on both strands to eliminate two bases. For some plant retrotransposons, the number of cleaved bases varies from 2-7 bases (Feuerbach et al., 1997). This whole process is termed 3'-end processing. However, 3'-end processing is not conserved in retroelements. Both the Ty1/*copia* and Ty3/*gypsy* families have members that do not carry out 3'-end processing (Feuerbach et al., 1997). This feature can be deduced from the presence or absence of a space between the PBS and 5' LTR. After reverse transcription, nucleotides in this space are copied to the end of the linear cDNA, and need to be removed

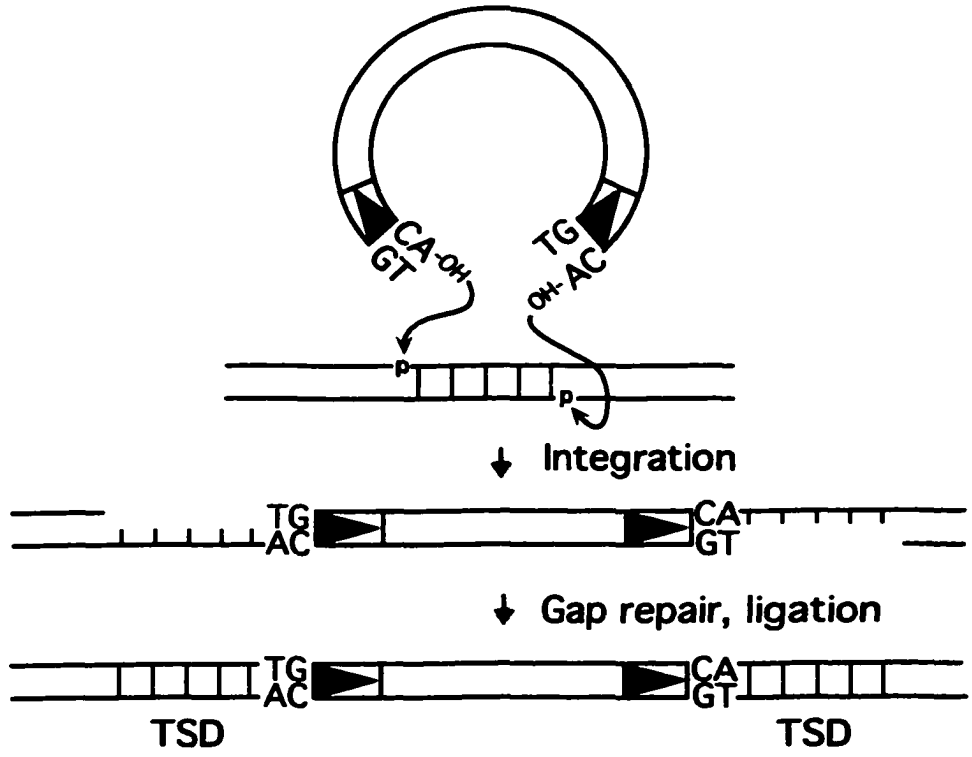



Fig 5. Integration of cDNA to host DNA. This figure shows integration without 3'-end processing, which is the case for Ty5. For integration that involves 3'-end processing, the 3' CA will be cleaved off before attacking the host DNA and the 5' TG will be cleaved off during gap repair.

before integration. Depending on the retroelement, the preintegration complex enters the nucleus either by active transport through nuclear pores during interphase, (e.g. HIV), or when the nuclear membrane is dissembled during cell division, (e.g. Moloney murine leukemia virus) (Roe et al., 1993). For the first method, a nuclear localization signal is associated with IN, MA and/or Vpr (Gallay et al., 1995a; Gallay et al., 1995b; Kukolj et al., 1998). After the complex enters the nucleus, it associates with the chromosome. The target site on the chromosome is not chosen randomly. A widely held model is that the proteins associated with host DNA guide integration to specific chromosomal regions (Bushman, 1995). Retroviruses prefer to insert into transcriptionally active regions (Mooslehner et al., 1990; Scherdin et al., 1990), DNase I hypersensitive sites (Goodenow and Hayward, 1987) or bent DNA (Milot et al., 1994). Ty1, Ty2, Ty3, and Ty4 of *S. cerevisiae* are mostly associated with RNA polymerase III transcription complexes, and they target into the upstream region of genes transcribed by RNA polymerase III (Chalker and Sandmeyer, 1992; Devine and Boeke, 1996; Kim et al., 1998). Ty5 integrase interacts with Sir proteins at the telomeres and silent mating loci, and Ty5 targets integration to these sites preferentially (Gai and Voytas, 1998; Xie et al., 2001; Zhu et al., 1999; Zou et al., 1996). Integrase catalyze the 3'-OH at each cDNA end to attack the phosphodiester bonds on both host DNA strands creating a staggered cut of 4-6 bases in the 5' direction. The 5' staggered cut in the target DNA is duplicated, and this duplication flanks the element. The 5' staggered two bases in the cDNA, which are generated by 3'-end processing, are removed. Cellular enzymes presumably mediate gap repair and ligation.

*Recombination* Because there is an efficient homologous recombination system in yeast, there is another pathway for Ty element cDNA to enter the genome after cDNA synthesis: it can recombine with an existing element using the host's recombination system (Ke et al., 1999; Melamed et al., 1992; Nevo-Caspi and Kupiec, 1996; Sharon et al., 1994). Recombination was extensively studied in the Ty1 and Ty5 retroelements. Ty1 cDNA can recombine with preexisting insertions to generate simple replacements or to form tandem repeats. In another case, through double crossing-over, chromosomes acquire the marker gene from the donor Ty1 DNA (Melamed et al., 1992; Sharon et al., 1994; Weinstock et al., 1990). Cellular DNA repair genes *RAD1, RAD51* and *RAD52* are involved in Ty1 cDNA recombination (Nevo-Caspi and Kupiec, 1996; Sharon et al., 1994). Ty5 has higher recombination frequencies, which accounts for 35% of the total transposition events. Half of these are LTR-mediated recombination events result in Ty5 tandem repeats. Another half are LTR independent recombination events resulting in selectable marker exchange between donor and recipient Ty5 elements. *RAD52* plays a major role in Ty5 cDNA recombination, because Ty5 recombination in *rad52* strains deceases to less than 1% of transposition, whereasin *rad1* strains recombination only decreases less than 2-fold (Ke and Voytas, 1997; Ke and Voytas, 1999).

**Evolution of retroelements**

Because of the wide distribution of retroelements, they are generally thought to originate from an ancient ancestor. According to the most popular model, the origin of retroelements dates back to the RNA world, which is assumed to exist before protein synthesis began and all biological processes were mediated by RNA. During the transition from the RNA world to the DNA world, the evolution of RT gave great advantage for RNA replication, primarily because DNA is much more chemically stable. Comparison of sequences (Xiong and Eickbush, 1990) and crystal structures (Telesnitsky and Goff, 1997) suggest that RT is more closely related to the RNA dependent RNA polymerases than to DNA polymerases. The phylogenetic tree of RT rooted on RNA-directed RNA polymerases shows that non-LTR and LTR retrotransposons/retroviruses diverged early (Boeke and Stoye, 1997). The Ty3/*gypsy* family is more closely related to retroviruses than to the Ty1/*copia* family. In Fig.6, the phylogenetic tree represents the phylogenetic relationship between the *Hemivirus* genus and the *Pseudovirus* genus within the Ty1/*copia* family. These two genera did not branch completely; instead, species from the same host species tend to cluster (Boeke et al., 2000a; Boeke et al., 2000b).

**The Ty5 system**

Most of the studies on retroviral replication have been restricted to *in vitro* biochemical approaches. Reverse transcription and integration can not be dissected genetically in animal cell culture systems. However, we are fortunate that retrotransposons

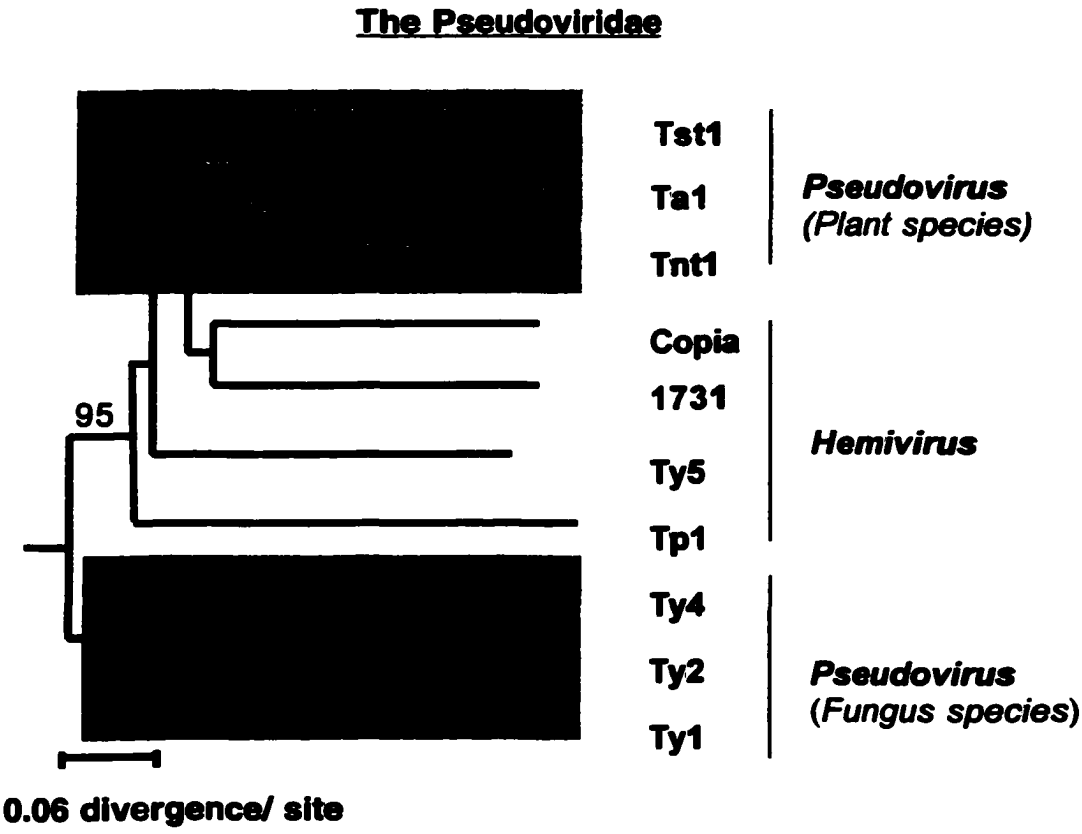



Fig 6. Phylogenetic relationships of Ty1/*copia* elements (*Pseudoviridae*). The tree is rooted on Ty3/*gypsy* (*Metaviridae*) sequences. The numbers on the branches are the bootstrap values. Values less than 50 are not shown and are less reliable.

of *S. cerevisiae* and their tRNA primers can be genetically manipulated, making yeast an ideal system to study replication mechanisms.

*S. cerevisiae* has five families of retrotransposons, Ty1 to Ty5. Ty1, Ty3 and Ty5 elements have been placed under inducible transcriptional controls to allow for regulated transposition (Boeke et al., 1985; Hansen and Sandmeyer, 1990; Zou et al., 1996). Over-expression of these elements makes it possible to purify VLPs and to analyze transposition intermediates (Eichinger and Boeke, 1988; Hansen et al., 1992; Ke et al., 1999). Selectable marker genes inserted into the elements allow for efficient quantitative measurements of the transposition frequency (Boeke and Garfinkel, 1988; Chalker and Sandmeyer, 1990; Garfinkel et al., 1988).

The primer tRNA is also amenable to genetic manipulation in yeast (von Pawel-Rammingen et al., 1992). Most *S. cerevisiae* strains have four *IMT* genes. All of these can be mutated by insertion and complemented by providing a wild type, plasmid-borne *IMT* gene (Bystrom and Fink, 1989; von Pawel-Rammingen et al., 1992). This system makes it possible to test tRNA mutants for their effect on transposition. Using this approach, Ty1 and Ty3 have been shown to prime reverse transcription with the 3'acceptor stem, which is complementary to the PBS region of the RNA template (Chapman et al., 1992; Keeney et al., 1995). Other residues in the D arm and TΨC arm are also important in priming (Friant et al., 1998; Gabus et al., 1998; Keeney et al., 1995). These residues base pair with other regions in the mRNA template and stabilize the primer–template interaction (Friant et al., 1998; Gabus et al., 1998).

The PBS of Ty5 is complementary to the anticodon stem-loop of the *S. cerevisiae IMT* (Fig. 2) (Voytas and Boeke, 1992). The region of complementarity is identical to that observed between *IMT* and *copia* in *D. melanogaster* (Kikuchi et al., 1986). To date, most of the studies of half-tRNA priming have been conducted with the *copia* element (Kikuchi et al., 1986; Kikuchi and Sasaki, 1992; Kikuchi et al., 1990). Since genetic assays for *copia* transposition have not been developed, most of the work was limited to biochemical approaches. Now, we can study half-tRNA priming with Ty5, using genetic and biochemical methods. The half-priming mechanism is likely highly conserved, since identical half-tRNAs are used by alge, protist and animal retroelements. Therefore, the half-tRNA should exist in diverse organisms.

Like any protein superfamily, retrotransposons are found in a wide variety of organisms, and most organisms typically have multiple diverse retroelements. Despite the number of diverse sequences, these proteins carry out similar roles in replication. Despite the similarity in the roles carried out by these proteins, these proteins have evolved distinct mechanisms to perform the similar functions. As we discussed above, some reverse transcriptases have evolved to use different RNA primers, most notably different cellular tRNAs or tRNA fragments (Chapman et al., 1992; Ke et al., 1999; Kikuchi et al., 1986; Leis et al., 1993). Similarly, integrase, which inserts linear cDNA, carries out its reaction with some variations. One variation is that the end of the cDNA is processed by removing some nucleotides prior to integration (3'end-processing) (Feuerbach et al., 1997). We therefore felt that the retrotransposons, both because of their number and diversity and the fact that they

have clearly documented cases of novel functional specificity, would serve as good models to develop general methods to identify sequences responsible for functional diversity.

The question we wantd to address is what are the determinant factors in the tRNA primer and Ty5 proteins for half-tRNA priming. In my dissertation, a bioinformatics approach was used to analyze the role of RT in half-tRNA primer complex recognition. A library screen was performed to identify mutations in Ty5 proteins relevant to reverse transcription.

**Dissertation Organization**

Chapter II, III and the appendix of my thesis are organized in the form of papers. In Chapter II, I identified amino acids in Ty5 protein important for reverse transcription by using molecular genetics approaches. Two mutations in Ty5 *gag* were characterized. These mutations increase Ty5 cDNA synthesis independently, and therefore result in an increase in Ty5 transposition of about 5-6 fold. I demonstrated that a mutation in the Ty5 zinc finger optimized the hydrogen bonding ability of NCp during reverse transcription, by replacing the acidic amino acid with other amino acids that have higher hydrogen bonding potential. This research suggested that the NCp protein and its zinc finger play an important role in Ty5 reverse transcription as in other retroelements. The requirement of efficient hydrogen bonding is likely universal for both retroviral CCHC and cellular zinc fingers. In addition, the hypothesis was tested that accumulated mutations in the Ty5 genome hinder the replication process and reduce the activity of transposition. Chapter III describes a new bioinformatics

method I developed to study the protein domains involved in functional divergence. Software was implemented to achieve this goal. The domains of RT that might be involved in the recognition of primer/template complexes were explored. The application of this software on other protein families (i.e. IN and the cellular Myb gene family) was also tested to identify domains of functional divergence. The results from this bioinformatics method were consistent with previous experimental results and implicated the domains of RT responsible for different primer recognition. In the appendix, determinants for half -tRNA priming are characterized on the IMT primer. Base paring between the anticodon stem-loop of the IMT and PBS of Ty5 was shown to be important for priming. Chapter IV provides a general conclusion. The accumulated data on half-tRNA priming was also discussed in this chapter.

## References

Alam, S. L., Wills, N. M., Ingram, J. A., Atkins, J. F., and Gesteland, R. F. (1999). Structural studies of the RNA pseudoknot required for readthrough of the gag-termination codon of murine leukemia virus, J Mol Biol *288*, 837-52.

Allain, B., Lapadat-Tapolsky, M., Berlioz, C., and Darlix, J. L. (1994). Transactivation of the minus-strand DNA transfer by nucleocapsid protein during reverse transcription of the retroviral genome, Embo J *13*, 973-81.

Atwood, A., Lin, J. H., and Levin, H. L. (1996). The retrotransposon Tf1 assembles virus-like particles that contain excess Gag relative to integrase because of a regulated degradation process, Mol Cell Biol *16*, 338-46.

Barat, C., Schatz, O., Le Grice, S., and Darlix, J. L. (1993). Analysis of the interactions of HIV1 replication primer tRNA(Lys,3) with nucleocapsid protein and reverse transcriptase, J Mol Biol *231*, 185-90.

Berg, D. E., and Howe, M. M. (1989). Mobile DNA (Washington, DC, American Society for Microbiology).

Berkowitz, R., Fisher, J., and Goff, S. P. (1996). RNA packaging, Curr Top Microbiol Immunol *214*, 177-218.

Boeke, J. D., Eickbush, T., Sandmeyer, S. B., and Voytas, D. F. (2000a). Metaviridae. In Virus Taxonomy: Seventh Report of the International Committee on Taxonomy of Viruses, M. H. V. van Regenmortel, C. M. Fauquet, D. H. L. Bishop, E. B. Carsten, M. K. Estes, S. M. Lemon, J. Maniloff, M. A. Mayo, D. J. McGeoch, C. R. Pringle, and R. B. Wickner, eds. (New York, Academic Press), pp. 359-67.

Boeke, J. D., Eickbush, T., Sandmeyer, S. B., and Voytas, D. F. (2000b). Pseudoviridae. In Virus Taxonomy: Seventh Report of the International Committee on Taxonomy of Viruses, M. H. V. van Regenmortel, C. M. Fauquet, D. H. L. Bishop, E. B. Carsten, M. K. Estes, S. M. Lemon, J. Maniloff, M. A. Mayo, D. J. McGeoch, C. R. Pringle, and R. B. Wickner, eds. (New York, Academic Press), pp. 349-57.

Boeke, J. D., and Garfinkel, D. J. (1988). Yeast Ty elements as retroviruses, 15-39.

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

Boeke, J. D., and Sandmeyer, S. B. (1991). Yeast transposable elements. In The Molecular and Cellular Biology of the Yeast *Saccharomyces*, J. Broach, E. Jones, and J. Pringle, eds. (Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory), pp. 193-261.

Boeke, J. D., and Stoye, J. P. (1997). Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In Retroviruses, J. M. Coffin, S. H. Hughes, and H. E. Varmus, eds. (Cold Spring Harbor, Cold Spring Harbor Laboratory Press), pp. 343-436.

Boyer, P. L., Ferris, A. L., and Hughes, S. H. (1992). Cassette mutagenesis of the reverse transcriptase of human immunodeficiency virus type 1, J Virol *66*, 1031-9.

Brierley, C., and Flavell, A. J. (1990). The retrotransposon *copia* controls the relative levels of its gene products post-transcriptionally by differential expression from its two major mRNAs, Nucleic Acids Res *18*, 2947-51.

Brown, P., and Varmus, H. (1989). Retroviruses. In Mobile DNA, D. E. Berg, and M. M. Howe, eds. (Washington D. C., American Society for Microbiology), pp. 53-108.

Bukrinsky, M. I., Sharova, N., McDonald, T. L., Pushkarskaya, T., Tarpley, W. G., and Stevenson, M. (1993). Association of integrase, matrix, and reverse transcriptase antigens of human immunodeficiency virus type 1 with viral nucleic acids following acute infection, Proc Natl Acad Sci U S A *90*, 6125-9.

Bushman, F. (1995). Targeting retroviral integration [comment], Science *267*, 1443-4.
Bystrom, A. S., and Fink, G. R. (1989). A functional analysis of the repeated methionine initiator tRNA genes (IMT) in yeast, Mol Gen Genet *216*, 276-86.

Chalker, D. L., and Sandmeyer, S. B. (1990). Transfer RNA genes are genomic targets for *de novo* transposition of the yeast retrotransposon Ty3, Genetics *126*, 837-50.

Chalker, D. L., and Sandmeyer, S. B. (1992). Ty3 integrates within the region of RNA polymerase III transcription initiation, Genes Dev *6*, 117-28.

Chan, B., and Musier-Forsyth, K. (1997). The nucleocapsid protein specifically anneals tRNALys-3 onto a noncomplementary primer binding site within the HIV-1 RNA genome in vitro, Proc Natl Acad Sci U S A *94*, 13530-5.

Chapman, K. B., Bystrom, A. S., and Boeke, J. D. (1992). Initiator methionine tRNA is essential for Ty1 transposition in yeast, Proc Natl Acad Sci U S A *89*, 3236-40.

Cristofari, G., Ficheux, D., and Darlix, J. L. (2000). The GAG-like protein of the yeast Ty1 retrotransposon contains a nucleic acid chaperone domain analogous to retroviral nucleocapsid proteins, J Biol Chem *275*, 19210-7.

Cristofari, G., Gabus, C., Ficheux, D., Bona, M., Le Grice, S. F., and Darlix, J. L. (1999). Characterization of active reverse transcriptase and nucleoprotein complexes of the yeast retrotransposon Ty3 in vitro, J Biol Chem *274*, 36643-8.

Darlix, J. L., Vincent, A., Gabus, C., de Rocquigny, H., and Roques, B. (1993). Trans-activation of the 5' to 3' viral DNA strand transfer by nucleocapsid protein during reverse transcription of HIV1 RNA, C R Acad Sci III *316*, 763-71.

Devine, S. E., and Boeke, J. D. (1996). Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III, Genes Dev *10*, 620-33.

Dougherty, W. G., and Semler, B. L. (1993). Expression of virus-encoded proteinases: functional and structural similarities with cellular enzymes, Microbiol Rev *57*, 781-822.

Eichinger, D. J., and Boeke, J. D. (1988). The DNA intermediate in yeast Ty1 element transposition copurifies with virus-like particles: cell-free Ty1 transposition, Cell *54*, 955-66.

Farabaugh, P. J., and Fink, G. R. (1980). Insertion of the eukaryotic transposable element Ty1 creates a 5-base pair duplication, Nature *286*, 352-6.

Farnet, C. M., and Bushman, F. D. (1997). HIV-1 cDNA integration: requirement of HMG I(Y) protein for function of preintegration complexes in vitro, Cell *88*, 483-92.

Fawcett, D. H., Lister, C. K., Kellett, E., and Finnegan, D. J. (1986). Transposable elements controlling I-R hybrid dysgenesis in D. melanogaster are similar to mammalian LINEs, Cell *47*, 1007-15.

Feng, Y. X., Copeland, T. D., Henderson, L. E., Gorelick, R. J., Bosche, W. J., Levin, J. G., and Rein, A. (1996). HIV-1 nucleocapsid protein induces "maturation" of dimeric retroviral RNA in vitro, Proc Natl Acad Sci U S A *93*, 7577-81.

Feuerbach, F., Drouaud, J., and Lucas, H. (1997). Retrovirus-like end processing of the tobacco Tnt1 retrotransposon linear intermediates of replication, J Virol *71*, 4005-15.

Fourcade-Peronnet, F., d'Auriol, L., Becker, J., Galibert, F., and Best-Belpomme, M. (1988). Primary structure and functional organization of *Drosophila* 1731 retrotransposon, Nucleic Acids Res *16*, 6113-25.

Fouts, D. L., Manning, J. E., Fox, G. M., and Schmid, C. W. (1981). A complex repeated DNA sequence within the Drosophila transposable element copia, Nucleic Acids Res *9*, 7053-64.

Friant, S., Heyman, T., Bystrom, A. S., Wilhelm, M., and Wilhelm, F. X. (1998). Interactions between Ty1 retrotransposon RNA and the T and D regions of the tRNA(iMet) primer are required for initiation of reverse transcription *in vivo*, Mol Cell Biol *18*, 799-806.

Friant, S., Heyman, T., Wilhelm, M. L., and Wilhelm, F. X. (1996). Extended interactions between the primer tRNAi(Met) and genomic RNA of the yeast Ty1 retrotransposon, Nucleic Acids Res *24*, 441-9.

Gabus, C., Ficheux, D., Rau, M., Keith, G., Sandmeyer, S., and Darlix, J. L. (1998). The yeast Ty3 retrotransposon contains a 5'-3' bipartite primer-binding site and encodes nucleocapsid protein NCp9 functionally homologous to HIV-1 NCp7, EMBO J *17*, 4873-80.

Gai, X., and Voytas, D. F. (1998). A single amino acid change in the yeast retrotransposon Ty5 abolishes targeting to silent chromatin, Mol Cell *1*, 1051-5.

Gallay, P., Swingler, S., Aiken, C., and Trono, D. (1995a). HIV-1 infection of nondividing cells: C-terminal tyrosine phosphorylation of the viral matrix protein is a key regulator, Cell *80*, 379-88.

Gallay, P., Swingler, S., Song, J., Bushman, F., and Trono, D. (1995b). HIV nuclear import is governed by the phosphotyrosine-mediated binding of matrix to the core domain of integrase, Cell *83*, 569-76.

Garfinkel, D. J., Mastrangelo, M. F., Sanders, N. J., Shafer, B. K., and Strathern, J. N. (1988). Transposon tagging using Ty elements in yeast, Genetics *120*, 95-108.

Goodenow, M. M., and Hayward, W. S. (1987). 5' long terminal repeats of myc-associated proviruses appear structurally intact but are functionally impaired in tumors induced by avian leukosis viruses, J Virol *61*, 2489-98.

Gorelick, R. J., Fu, W., Gagliardi, T. D., Bosche, W. J., Rein, A., Henderson, L. E., and Arthur, L. O. (1999). Characterization of the block in replication of nucleocapsid protein zinc finger mutants from moloney murine leukemia virus, J Virol *73*, 8185-95.

Hansen, L. J., Chalker, D. L., Orlinsky, K. J., and Sandmeyer, S. B. (1992). Ty3 *GAG3* and *POL3* genes encode the components of intracellular particles, J Virol *66*, 1414-24.

Hansen, L. J., Chalker, D. L., and Sandmeyer, S. B. (1988). Ty3, a yeast retrotransposon associated with tRNA genes, has homology to animal retroviruses, Mol Cell Biol *8*, 5245-56.

Hansen, L. J., and Sandmeyer, S. B. (1990). Characterization of a transpositionally active Ty3 element and identification of the Ty3 integrase protein, J Virol *64*, 2599-607.

Hattori, M., Kuhara, S., Takenaka, O., and Sakaki, Y. (1986). L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein, Nature *321*, 625-8.

Hsu, M., Rong, L., de Rocquigny, H., Roques, B. P., and Wainberg, M. A. (2000). The effect of mutations in the HIV-1 nucleocapsid protein on strand transfer in cell-free reverse transcription reactions, Nucleic Acids Res *28*, 1724-9.

Isel, C., Ehresmann, C., Keith, G., Ehresmann, B., and Marquet, R. (1995). Initiation of reverse transcription of HIV-1: secondary structure of the HIV-1 RNA/tRNA(3Lys) (template/primer), J Mol Biol *247*, 236-50.

Isel, C., Westhof, E., Massire, C., Le Grice, S. F., Ehresmann, B., Ehresmann, C., and Marquet, R. (1999). Structural basis for the specificity of the initiation of HIV-1 reverse transcription, Embo J *18*, 1038-48.

Jacks, T. (1990). Translational suppression in gene expression in retroviruses and retrotransposons, Curr Top Microbiol Immunol *157*, 93-124.

Jakubczak, J. L., Burke, W. D., and Eickbush, T. H. (1991). Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects, Proc Natl Acad Sci U S A *88*, 3295-9.

Kanaya, S., Kohara, A., Miura, Y., Sekiguchi, A., Iwai, S., Inoue, H., Ohtsuka, E., and Ikehara, M. (1990). Identification of the amino acid residues involved in an active site of Escherichia coli ribonuclease H by site-directed mutagenesis, J Biol Chem *265*, 4615-21.

Ke, N., Gao, X., Keeney, J. B., Boeke, J. D., and Voytas, D. F. (1999). The yeast retrotransposon Ty5 uses the anticodon stem-loop of the initiator methionine tRNA as a primer for reverse transcription, Rna *5*, 929-38.

Ke, N., and Voytas, D. F. (1997). High frequency cDNA recombination of the *Saccharomyces* retrotransposon Ty5: The LTR mediates formation of tandem elements, Genetics *147*, 545-56.

Ke, N., and Voytas, D. F. (1999). cDNA of the yeast retrotransposon Ty5 preferentially recombines with substrates in silent chromatin, Mol Cell Biol *19*, 484-94.

Keeney, J. B., Chapman, K. B., Lauermann, V., Voytas, D. F., Astrom, S. U., von Pawel-Rammingen, U., Bystrom, A., and Boeke, J. D. (1995). Multiple molecular determinants for retrotransposition in a primer tRNA, Mol Cell Biol *15*, 217-26.

Kikuchi, Y., Ando, Y., and Shiba, T. (1986). Unusual priming mechanism of RNA-directed DNA synthesis in copia retrovirus-like particles of *Drosophila*, Nature *323*, 824-6.

Kikuchi, Y., and Sasaki, N. (1992). Hyperprocessing of tRNA by the catalytic RNA of RNase P. Cleavage of a natural tRNA within the mature tRNA sequence and evidence for an altered conformation of the substrate tRNA, J Biol Chem *267*, 11972-6.

Kikuchi, Y., Sasaki, N., and Ando-Yamagami, Y. (1990). Cleavage of tRNA within the mature tRNA sequence by the catalytic RNA of RNase P: implication for the formation of the primer tRNA fragment for reverse transcription in *copia* retrovirus-like particles, Proc Natl Acad Sci U S A *87*, 8105-9.

Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A., and Voytas, D. F. (1998). Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence, Genome Res *8*, 464-78.

Kirchner, J., and Sandmeyer, S. B. (1996). Ty3 integrase mutants defective in reverse transcription or 3'-end processing of extrachromosomal Ty3 DNA, J Virol *70*, 4737-47.

Kukolj, G., Katz, R. A., and Skalka, A. M. (1998). Characterization of the nuclear localization signal in the avian sarcoma virus integrase, Gene *223*, 157-63.

Lapadat-Tapolsky, M., De Rocquigny, H., Van Gent, D., Roques, B., Plasterk, R., and Darlix, J. L. (1993). Interactions between HIV-1 nucleocapsid protein and viral DNA may have important functions in the viral life cycle [published erratum appears in Nucleic Acids Res 1993 Apr 25;21(8):2024], Nucleic Acids Res *21*, 831-9.

Lapadat-Tapolsky, M., Pernelle, C., Borie, C., and Darlix, J. L. (1995). Analysis of the nucleic acid annealing activities of nucleocapsid protein from HIV-1, Nucleic Acids Res *23*, 2434-41.

Larder, B. A., Purifoy, D. J., Powell, K. L., and Darby, G. (1987). Site-specific mutagenesis of AIDS virus reverse transcriptase, Nature *327*, 716-7.

Lauermann, V., and Boeke, J. D. (1997). Plus-strand strong-stop DNA transfer in yeast Ty retrotransposons, Embo J *16*, 6603-12.

Lee, M. S., and Craigie, R. (1994). Protection of retroviral DNA from autointegration: involvement of a cellular factor, Proc Natl Acad Sci U S A *91*, 9823-7.

Lee, Y. M., and Coffin, J. M. (1991). Relationship of avian retrovirus DNA synthesis to integration in vitro, Mol Cell Biol *11*, 1419-30.

Leis, J., Aiyar, A., and Cobrinik, D. (1993). Regulation of initiation of reverse transcription of retroviruses. In Reverse transcriptase, S. Goff, and A. Skalka, eds. (Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory), pp. 33-47.

Levin, H. L. (1995). A novel mechanism of self-primed reverse transcription defines a new family of retroelements, Mol Cell Biol *15*, 3310-7.

Levin, H. L. (1996). An unusual mechanism of self-primed reverse transcription requires the RNase H domain of reverse transcriptase to cleave an RNA duplex, Mol Cell Biol *16*, 5645-54.

Lindauer, A., Fraser, D., Bruderlein, M., and Schmitt, R. (1993). Reverse transcriptase families and a *copia*-like retrotransposon, *Osser*, in the green alga *Volvox carteri*, FEBS Lett *319*, 261-6.

McCurrach, K. J., Rothnie, H. M., Hardman, N., and Glover, L. A. (1990). Identification of a second retrotransposon-related element in the genome of *Physarum polycephalum*, Curr Genet *17*, 403-8.

Melamed, C., Nevo, Y., and Kupiec, M. (1992). Involvement of cDNA in homologous recombination between Ty elements in *Saccharomyces cerevisiae*, Mol Cell Biol *12*, 1613-20.

Merkulov, G. V., Lawler, J. F., Eby, Y., and Boeke, J. D. (2001). Ty1 proteolytic cleavage sites are required for transposition: all sites are not created equal, J Virol *75*, 638-44.

Miller, K., Rosenbaum, J., Zbrzezna, V., and Pogo, A. O. (1989). The nucleotide sequence of Drosophila melanogaster copia-specific 2.1- kb mRNA, Nucleic Acids Res *17*, 2134.

Milot, E., Belmaaza, A., Rassart, E., and Chartrand, P. (1994). Association of a host DNA structure with retroviral integration sites in chromosomal DNA, Virology *201*, 408-12.

Mishima, Y., and Steitz, J. A. (1995). Site-specific crosslinking of 4-thiouridine-modified human tRNA(3Lys) to reverse transcriptase from human immunodeficiency virus type I, Embo J *14*, 2679-87.

Mizrahi, V., Usdin, M. T., Harington, A., and Dudding, L. R. (1990). Site-directed mutagenesis of the conserved Asp-443 and Asp-498 carboxy- terminal residues of HIV-1 reverse transcriptase, Nucleic Acids Res *18*, 5359-63.

Mooslehner, K., Karls, U., and Harbers, K. (1990). Retroviral integration sites in transgenic Mov mice frequently map in the vicinity of transcribed DNA regions, J Virol *64*, 3056-8.

Nevo-Caspi, Y., and Kupiec, M. (1996). Induction of Ty recombination in yeast by cDNA and transcription: role of the *RAD1* and *RAD52* genes, Genetics *144*, 947-55.

Nymark-McMahon, M. H., and Sandmeyer, S. B. (1999). Mutations in nonconserved domains of Ty3 integrase affect multiple stages of the Ty3 life cycle, J Virol *73*, 453-65.

O'Hare, K., and Rubin, G. M. (1983). Structures of P transposable elements and their sites of insertion and excision in the Drosophila melanogaster genome, Cell *34*, 25-35.

Poch, O., Sauvaget, I., Delarue, M., and Tordo, N. (1989). Identification of four conserved motifs among the RNA-dependent polymerase encoding elements, Embo J *8*, 3867-74.

Pohlman, R. F., Fedoroff, N. V., and Messing, J. (1984). The nucleotide sequence of the maize controlling element Activator, Cell *37*, 635-43.

Prats, A. C., Sarih, L., Gabus, C., Litvak, S., Keith, G., and Darlix, J. L. (1988). Small finger protein of avian and murine retroviruses has nucleic acid annealing activity and positions the replication primer tRNA onto genomic RNA, Embo J *7*, 1777-83.

Rausch, J. W., Grice, M. K., Henrietta, M., Nymark, M., Miller, J. T., and Le Grice, S. F. (2000). Interaction of p55 reverse transcriptase from the Saccharomyces cerevisiae retrotransposon Ty3 with conformationally distinct nucleic acid duplexes, J Biol Chem *275*, 13879-87.

Remy, E., de Rocquigny, H., Petitjean, P., Muriaux, D., Theilleux, V., Paoletti, J., and Roques, B. P. (1998). The annealing of tRNA3Lys to human immunodeficiency virus type 1 primer binding site is critically dependent on the NCp7 zinc fingers structure, J Biol Chem *273*, 4819-22.

Repaske, R., Hartley, J. W., Kavlick, M. F., O'Neill, R. R., and Austin, J. B. (1989). Inhibition of RNase H activity and viral replication by single mutations in the 3' region of Moloney murine leukemia virus reverse transcriptase, J Virol *63*, 1460-4.

Roe, T., Reynolds, T. C., Yu, G., and Brown, P. O. (1993). Integration of murine leukemia virus DNA depends on mitosis, Embo J *12*, 2099-108.

Rong, L., Liang, C., Hsu, M., Kleiman, L., Petitjean, P., de Rocquigny, H., Roques, B. P., and Wainberg, M. A. (1998). Roles of the human immunodeficiency virus type 1 nucleocapsid protein in annealing and initiation versus elongation in reverse transcription of viral negative-strand strong-stop DNA, J Virol *72*, 9353-8.

Rothnie, H. M., McCurrach, K. J., Glover, L. A., and Hardman, N. (1991). Retrotransposon-like nature of Tp1 elements: implications for the organisation of highly repetitive, hypermethylated DNA in the genome of *Physarum polycephalum*, Nucleic Acids Res *19*, 279-86.

Scherdin, U., Rhodes, K., and Breindl, M. (1990). Transcriptionally active genome regions are preferred targets for retrovirus integration, J Virol *64*, 907-12.

Sharon, G., Burkett, T. J., and Garfinkel, D. J. (1994). Efficient homologous recombination of Ty1 element cDNA when integration is blocked, Mol Cell Biol *14*, 6540-51.

Song, S. U., Gerasimova, T., Kurkulos, M., Boeke, J. D., and Corces, V. G. (1994). An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus, Genes Dev *8*, 2046-57.

Song, S. U., Kurkulos, M., Boeke, J. D., and Corces, V. G. (1997). Infection of the germ line by retroviral particles produced in the follicle cells: a possible mechanism for the mobilization of the gypsy retroelement of Drosophila, Development *124*, 2789-98.

Telesnitsky, A., and Goff, S. P. (1997). Reverse transcriptase and the generation of retroviral DNA. In Retroviruses, J. Coffin, S. H. Hughes, and H. E. Varmus, eds. (Cold Spring Harbor, Cold Spring Harbor Laboratory Press), pp. 121-160.

Vogt, V. M. (1997). Retroviral virions and genomes. In Retroviruses, J. Coffin, S. H. Hughes, and H. E. Varmus, eds. (Cold Spring Harbor, Cold Spring Harbor Laboratory Press), pp. 27-70.

von Pawel-Rammingen, U., Astrom, S., and Bystrom, A. S. (1992). Mutational analysis of conserved positions potentially important for initiator tRNA function in *Saccharomyces cerevisiae*, Mol Cell Biol *12*, 1432-42.

Voytas, D. F., and Boeke, J. D. (1992). Yeast retrotransposon revealed, Nature *358*, 717.

Voytas, D. F., and Boeke, J. D. (1993). Yeast retrotransposons and tRNAs, Trends Genet *9*, 421-7.

Weinstock, K. G., Mastrangelo, M. F., Burkett, T. J., Garfinkel, D. J., and Strathern, J. N. (1990). Multimeric arrays of the yeast retrotransposon Ty, Mol Cell Biol *10*, 2882-92.

Wilhelm, M., Boutabout, M., and Wilhelm, F. X. (2000). Expression of an active form of recombinant Ty1 reverse transcriptase in *Escherichia coli*: a fusion protein containing the C-terminal region of the Ty1 integrase linked to the reverse transcriptase-RNase H domain exhibits polymerase and RNase H activities, Biochem J *2*, 337-42.

Wu, X., Liu, H., Xiao, H., Conway, J. A., Hehl, E., Kalpana, G. V., Prasad, V., and Kappes, J. C. (1999). Human immunodeficiency virus type 1 integrase protein promotes reverse transcription through specific interactions with the nucleoprotein reverse transcription complex, J Virol *73*, 2126-35.

Xie, W., Gai, X., Zhu, Y., Zappulla, D. C., Sternglanz, R., and Voytas, D. F. (2001). Targeting of the Yeast Ty5 Retrotransposon to Silent Chromatin Is Mediated by Interactions between Integrase and Sir4p, Mol Cell Biol *21*, 6606-14.

Xiong, Y., and Eickbush, T. H. (1988). Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns, Mol Biol Evol *5*, 675-90.

Xiong, Y., and Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences, Embo J *9*, 3353-62.

Zennou, V., Petit, C., Guetard, D., Nerhbass, U., Montagnier, L., and Charneau, P. (2000). HIV-1 genome nuclear import is mediated by a central DNA flap, Cell *101*, 173-85.

Zhu, Y., Zou, S., Wright, D., and Voytas, D. (1999). Tagging chromatin with retrotransposons: target specificity of the *Saccharomyces* Ty5 retrotransposon changes with the chromosomal localization of Sir3p and Sir4p, Genes & Dev *13*, 2738-49.

Zou, S., Ke, N., Kim, J. M., and Voytas, D. F. (1996). The *Saccharomyces* retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci, Genes Dev *10*, 634-45.

Zou, S., Wright, D. A., and Voytas, D. F. (1995). The *Saccharomyces* Ty5 retrotransposon family is associated with origins of DNA replication at the telomeres and the silent mating locus *HMR*, Proc Natl Acad Sci U S A *92*, 920-4.

# CHAPTER II. Ty5 *gag* MUTATIONS INCREASE RETROTRANSPOSITION AND SUGGEST A ROLE FOR HYDROGEN BONDING IN THE FUNCTION OF THE NUCLEOCAPSID ZINC FINGER

**A manuscript submitted to *Journal of Virology***

**Xiang Gao[1], Daniel J. Rowley[2], Xiaowu Gai[3] and Daniel F. Voytas[4]**

## ABSTRACT

The Ty5 retrotransposon of *Saccharomyces paradoxus* transposes in *S. cerevisiae* at frequencies 1000-fold lower than the native Ty1elements. The low transposition activity of Ty5 could be due to differences in cellular environments between these yeast species or to naturally occurring mutations in Ty5. By screening a Ty5 mutant library, two single mutants (D252N and Y68C) were each found to increase transposition approximately 6-fold. When combined, transposition increased 36-fold, implying that the two mutations act independently. Neither mutation affected Ty5 protein synthesis, processing, cDNA recombination or target site choice. However, cDNA levels in both single mutants and the

---

[1] Primary researcher and author.
[2] Undergraduate student who mapped the mutation sites.
[3] Graduate student who screened the library.
[4] Professor and corresponding author, Department of Zoology and Genetics, Iowa State University, Ames, IA 50011.

double mutant were significantly higher than wild type. The D252N mutation resides in the zinc finger of nucleocapsid and increases the potential for hydrogen bonding with nucleic acids. We generated other mutations that increase the hydrogen bonding potential (i.e. D252R and D252K) and found that they similarly increased transposition. This suggests that hydrogen bonding within the zinc finger motif is important for cDNA production, and builds upon previous studies implicating basic amino acids flanking the zinc finger as important for zinc finger function. Although NCp zinc fingers differ from the zinc finger motifs of cellular enzymes, the requirement for efficient hydrogen bonding is likely universal.

## INTRODUCTION

Retrotransposons and retroviruses (collectively referred to as retroelements) replicate through a mRNA intermediate. Retroelements have a conserved genomic organization. They encode Gag and Pol polyproteins from ORF(s) between two long terminal repeats (LTRs). In the retroviruses, Gag is processed into capsid (CA), nucleocapsid (NCp) and matrix proteins, which assemble into a virus particle; most retrotransposons encode CA and NCp homologues. The retroelement Pol polyprotein has the enzymatic functions required for replication, including protease (PR), integrase (IN) and reverse transcriptase (RT) activities. RT synthesizes a cDNA copy of the retroelement, and IN inserts the cDNA into the chromosome of the host.

Retrotransposons serve as important models for understanding retroelement replication. This is particularly true of the yeast retrotransposons, namely those of *S. pombe* (Tf1) and *S. cerevisiae* (Ty1, Ty3, and Ty5), where considerable genetic resources and tools are available to facilitate their

study. Among the yeast retrotransposons, Ty5 has become an important model for understanding integration specificity, because of its preference to integrate into silent regions of the yeast genome (Xie et al., 2001; Zhu et al., 1999; Zou et al., 1996). Ty5 is unusual among retroelements in that it encodes a single open reading frame that is cleaved by protease into PR, RT, IN, and two forms of Gag, Gag-p27 and Gag-p37 (Irwin and Voytas, 2001). The 10 kD fragment released during Gag processing is likely NCp. Ty5 is also one of a small group of retrotransposons (the Hemiviruses) that prime reverse transcription with a half-tRNA (Boeke et al., 2000; Ke et al., 1999). Studies of Ty5, however, have been hampered by its low transposition frequency ($\sim 10^{-5}$) compared to other yeast retrotransposons (e.g. $\sim 10^{-2}$ for Ty1) (Curcio and Garfinkel, 1991; Zou et al., 1996). The Ty5 element used in all studies to date (Ty5-6p) comes from *S. paradoxus*, a sibling species of *S. cerevisiae* (Zou et al., 1996). All *S. cerevisiae* elements are either solo LTRs or degenerate elements (Kim et al., 1998; Zou et al., 1995).

Although Ty5-6p is transposition-competent, considering its low transposition frequency, it may carry mutations that impede high-efficiency replication. Alternatively, Ty5 could be partially incompatible with the cellular environment of its surrogate host, *S. cerevisiae*, and differences in host factors between species might negatively impact Ty5 transposition. In this regard, a number of host genes have been identified that affect retrotransposition of yeast elements such as Ty1 (Curcio and Garfinkel, 1999; Voytas and Boeke, In Press). High frequency transposition is important for undertaking genetic and biochemical studies directed at understanding unique aspects of Ty5 biology. In this study, we screened for Ty5 mutants with increased transposition frequency. In addition to making Ty5 a more tractable model system, we felt that mutations that increase transposition offer

the opportunity to understand better the element/host relationship and to identify amino acids critical for Ty5 replication.

## MATERIL AND METHODS

**Strain and plasmid construction.** Ty5 transposition was measured in the *S. cerevisiae* strains YPH499 (*MATa ura3-52 lys2-801 ade2-101 trp1Δ63 his3Δ200 leu2Δ1*), *W303 (MATα ade2-1 can1-100 his3-11 leu2,3,112 trp1-1 ura3-1*), and their *rad52::TRP1* derivatives. pXW97 and pXW98 are *GAL*-Ty5 plasmids recovered from the library screen (see below) that show elevated transposition. pDR3 through pDR6 are derivatives generated by swapping a *Xho*I-*Bsp*MII fragment with a wild type *GAL*-Ty5. pDR4 and pDR5 carry the D252N and Y68C mutations, respectively. To make the double mutant, the *Xho*I-*Hpa*I fragment of pDR5 was cloned into pSP72 (Promega) to generate pDR10. The *Bam*HI-*Sph*I fragment of pDR4 and pDR10 were swapped to generate a clone (pDR12) with both mutations. The *Xho*I-*Hpa*I fragment from pDR12 was excised and used to replace the wild type fragment of Ty5 in pNK254. This generated the doubly mutant plasmid pDR14. Other plasmids were generated as follows:

i) PCR-based mutagenesis was used to insert N-terminal epitope tags ($RGSH_6$, Qiagen) into wild type Ty5 and the single mutants (Ausubel et al., 1987). Two overlapping primers were used: DVO557 (5'-ATG-AGA-GGA-TCG-CAT-CAC-CAT-CAC-CAT-CAC-ACA-TAT-AAG-CTA-GAT-CG-3') and DVO558 (5'-GTG-ATG-GTG-ATG-

GTG-ATG-CGA-TCC-TCT-CAT-AAT-GTT-GTA-AGT-TTA-TTG-G-3'). This resulted in the plasmids pNK520 (wild type Ty5), pXG21 (Y68C), and pXG23 (D252N).

ii) pXG50 carries a Ty5 element with the D252R mutation and was constructed by PCR mutagenesis using overlapping primers DVO1509 (5'-GGG-GCT-CGG-CAT-CGC-TTA-AGC-3') and DVO1510 (5'-GCG-ATG-CCG-AGC-CCC-ACA-AAT-3') (Ausubel et al., 1987). Plasmids with Ty5 double mutants include pXG48 (Y68C, D252R) and pXG49 (Y68C, D252K), and they were constructed by PCR mutagenesis using pDR5 (Y68C) as a template and primers DVO1509, DVO1510, DVO1511 (5'-GGG-GCT-AAG-CAT-CGC-TTA-AGC-3'), and DVO1512 (5'-GCG-ATG-CTT-AGC-CCC-ACA-AAT-3').

iii) The competitor template for cDNA quantification by PCR was generated by first cloning a 820 bp *Kpn*I-*Sac*II fragment from Ty5 into pBluescript (Stratagene) to generate pXG24. pXG28 was constructed by deleting a 60 bp *Xma*I fragment from pXG24.

**Mutagenesis and library screening.** The *GAL*-Ty5 plasmid (pNK254) was mutagenized by growing for two days in the *E. coli* strain XL-1 Red, which has mutations in multiple DNA repair pathways (Stratagene) (Gai and Voytas, 1998). The mutagenized library was transformed into YPH499, and patch assays were performed with independent transformants to assess transposition (Zou et al., 1996). Of 3000 transformants evaluated, two showed higher levels of transposition. Plasmids from these transformants (pXW97 and pXW98) were purified and retransformed into YPH499 and W303 to confirm that the plasmids conferred the increase in transposition.

**Assays for integration, recombination and target specificity.** Quantitative transposition assays were conducted as previously described (Zou et al., 1996), with the exception that the induction of transposition on galactose media was carried out for 3 days. Throughout this manuscript, transposition refers to the total number of $His^+$ cells generated by Ty5 after growth on galactose. This includes both integration and cDNA recombination events. To calculate the frequency of integration, transposition assays were carried out in a *rad52Δ* strain to eliminate recombination events (Ke and Voytas, 1997). To determine the relative levels of integration and recombination, one hundred $His^+$ colonies were randomly selected from synthetic complete media lacking histidine (SC-H) plates, patched to new SC-H plates and allowed to grow for 3 days at 30°C. Cell patches were replica-plated onto SC-H plates with 5-fluoroorotic acid (5-FOA). The percentage of integration was calculated as the number of colonies that grew on SC-H/5-FOA plates divided by the number of colonies on SC-H plates. Our assay that measures targeting of Ty5 to a plasmid-borne *HMR* locus was carried out as previously described (Gai and Voytas, 1998).

**Protein preparation and immunoblot analysis.** The conditions used for cell growth and the induction of Ty5 transcription were as previously described (Irwin and Voytas, 2001). Harvested cells were disrupted by the glass bead method (Ausubel et al., 1987). The supernatant was collected from the cell lysate after centrifugation (20,000 x *g*, 60 min, 4°C). The remaining pellet was extracted with sample loading buffer (Ausubel et al., 1987). An equivalent volume of supernatant and pellet was used to compare proteins in the soluble and insoluble fractions. Proteins were subjected to 10% SDS-polyacrylamide gel

electrophoresis and electrophoretically transferred to nitrocellulose membranes (NitroBind; Micron Separations Inc.). The protocols for transfer and western blot analysis were as previously described (Irwin and Voytas, 2001).

**Assaying Ty5 cDNA.** Yeast total DNA was purified by the glass-bead method from cells grown to O.D. 3.0. All DNA was quantified spectrophotometrically (Ausubel et al., 1987). PCR primers 1 and 2 were DVO200 (5'-CAT-TAC-CCA-TAT-CAT-GCT-3') and DVO208 (5'-CAG-CCG-GAA-TGC-TTG-GCA-3'), respectively. Serially diluted competitor template was added to reactions with the same amount of yeast DNA. PCR conditions were as follows: 94°C 1 min, 54 °C 1 min, 72°C 1 min, for 30 cycles. After electrophoresis of the PCR reactions, bands in each lane were quantified using NIH image software (version 1.62, http://rsb.info.nih.gov/nih-image/). Lanes were selected for calculating Ty5 cDNA levels in which amounts of products derived from cDNA and competitor were nearly equal. cDNA levels were considered to be the amount of competitor DNA added to that reaction, with adjustments made for slight differences in amounts of the two products.

**Calculating hydrogen bonding potential.** Hydrogen bonding potential refers to the number of observed hydrogen bonds between a given amino acid and all four DNA bases (Mandel-Gutfreund et al., 1995). For each retroelement in the Ty1/*copia* group, including the wild type and mutant Ty5 elements, the hydrogen bonding potential was summed for amino acids in each interval of the $CX_2CX_{3-4}HX_4C$ motif (i.e. the C-C, C-H and H-C intervals). When considering the hydrogen bonding potential of the Ty1/*copia* group as a whole, the average potential for each interval was calculated.

## RESULTS

**Two mutations in *gag* increase Ty5 transposition frequency.** The *S. paradoxus* Ty5 element (Ty5-6p) transposes in *S. cerevisiae* at frequencies 1000-fold lower than Ty1 (Zou et al., 1996). To test whether there are mutations in Ty5-6p that negatively impact transposition, we screened for Ty5 mutants with increased transposition frequencies. Plasmids with a Ty5 element were mutagenized and transformed into the YPH499 strain of *S. cerevisiae*. Transposition was assayed for more than 3000 independent transformants. Our transposition assay measures the frequency by which $His^+$ cells are generated after inducing transcription of a *GAL*-Ty5 by growth on galactose. Ty5 carries a non-functional *his3* marker gene interrupted by an artificial intron (*his3*AI), and a functional *HIS3* gene is generated upon reverse transcription of spliced Ty5 mRNA (Curcio and Garfinkel, 1991; Zou et al., 1996). A $His^+$ phenotype results when Ty5 cDNA enters a target DNA molecule by integration or recombination. Transposition frequency, as defined here, is therefore the sum of the integration and recombination frequencies.

Two mutants, pXW98 and pXW97, were identified with approximately 6-fold higher frequencies of transposition (Table 1). To identify the mutations responsible for the increased transposition frequency, restriction fragments from pXW98 and pXW97 were swapped with the wild type element Ty5-6p. Transposition of the chimeric elements was retested, and the mutations were localized to a 3 kb *Xho*I/*Bsp*MII fragment encompassing the 5' half of Ty5. DNA sequencing revealed a single base change in pXW98 that resulted in a missense mutation, D252N. This mutation is located just before the conserved H residue in

the CCHC zinc finger domain of Gag ($CX_2CX_3HX_4C$). This zinc finger is the defining feature of nucleocapsid (NCp) proteins (Vogt, 1997). In pXW97, a single mutation resulted in the missense mutation Y68C. We combined the two mutations into one Ty5 element and found that it transposed approximately 36-fold higher than wild type. Because the fold increase in transposition of the double mutant is the product of the fold increase of the two single mutants, it is likely that the mutations affect different steps of Ty5 transposition.

**Effects of *gag* mutations on integration and recombination.** We were interested in identifying the steps of transposition affected by the two single *gag* mutations. One possibility is that the mutations affect interactions with a host factor(s) critical for transposition. We previously noted several-fold lower levels of transposition in the lab strain W303 relative to YPH499. This strain difference was also observed for the Ty5 mutants (Table 2), suggesting that the genetic differences between strains act independently from the *gag* mutations. Because our transposition assay measures the frequency of $His^+$ cells, which includes both integration and recombination events, we quantified integration in strains with mutations in the recombination/repair gene *RAD52*. Homologous recombination by Ty5 cDNA does not occur in *rad52_ _*strains(Ke and Voytas, 1997; Ke and Voytas, 1999). The frequency of $His^+$ cells, therefore, represents the integration frequency, which is typically about 70% of the total $His^+$ frequency in wild type strains. Our results showed that for each of the three Ty5 mutants, the fold increase in $His^+$ frequency (integration plus recombination) in wild type strains was comparable to the fold increase in integration in the *rad52_* strains

(Table 2). This suggests that the increase in total $His^+$ cells observed in the mutants is not due to an increase in the efficiency of integration.

We next tested the effect of the *gag* mutations on Ty5 cDNA recombination. Because chromosomal Ty5 elements are degenerate, Ty5 cDNA almost always recombines with the plasmid-borne donor element (Ke and Voytas, 1997; Ke and Voytas, 1999). This replaces the *his3*AI marker with the wild type *HIS3* gene, and confers a $His^+$ phenotype. The plasmid with the donor element also carries a *URA3* gene, which prevents growth in the presence of 5-fluoroorotic acid (5-FOA). Recombinants, therefore, can be identified by their $His^+$, $5\text{-}FOA^s$ phenotype. By this selection strategy, we calculated the percentage recombination for each of the mutants in two different wild type strains (Fig. 1). The results showed no distinguishable difference in recombination between the wild type and mutants. Integration and recombination, therefore, contribute almost equally to the increase in $His^+$ frequency. It is likely that the mutants affect steps prior to integration and recombination.

**Effects of *gag* mutations on cDNA levels.** Because the *gag* mutations cause an increase in both integration and recombination, they may act by increasing cDNA synthesis. We were previously unable to detect Ty5 cDNA by Southern hybridization experiments (N. Ke and D.F. Voytas, unpublished data), and so we developed a more sensitive, PCR-based assay. This assay was designed to amplify Ty5 cDNA and not DNA from the plasmid-borne donor element. To do this, we took advantage of the artificial intron (AI), which is present in the donor element but not in the cDNA. One PCR primer spans the AI, and therefore can only anneal to cDNA and not to Ty5 DNA (Primer 1, Fig. 2A). Amplification

with this primer and a second, downstream primer (Primer 2) should yield a 590 bp product using cDNA as a template. Strains with a *GAL*-Ty5 were grown in the presence or absence of galactose to test the specificity of the PCR assay. Ty5 cDNA was only detected in DNA isolated from strains with a *GAL*-Ty5 and after galactose induction (Fig. 2B). We also did not observe a product when the PCR assay was used with Ty5 donor plasmid as a template (data not shown). A control DNA template (the competitor template described below) was added to all reaction mixtures to ensure that the PCR reactions were working.

To quantify relative levels of Ty5 cDNA, we designed a second template to use in competitive PCR experiments. The competitor template can be amplified by both primers and yields a product 60 bp shorter than the product from cDNA (Fig. 2A). The amount of cDNA in a sample should equal the amount of added competitor when the amount of PCR product generated from both templates is equal. We showed that this was the case in control experiments using our competitor and a cloned copy of Ty5 cDNA (data not shown). PCR reactions were carried out with total genomic DNA prepared from induced cells with wild type or mutant Ty5 elements (Fig. 3A). Ty5 cDNA was quantified from reactions in which the products of the two templates were equal. The zinc finger mutant increased Ty5 cDNA 2.8-fold, the Gag mutant increased cDNA 2.2-fold, and the double mutant increased cDNA levels 4.8-fold (Fig. 3B). These results indicate that both mutants increase transposition by affecting Ty5 cDNA levels.

**Effects of *gag* mutations on protein processing and solubility.** Wild type Ty5 Gag is processed by protease into 27 Kd and 37 Kd proteins (Irwin and Voytas, 2001). Ty5 Gag is also

largely insoluble and typically requires ionic detergents to go into solution. To monitor Gag processing and solubility, an epitope tag ($RGSH_6$) was inserted into the Gag-Pol N-terminus of wild type and mutant Ty5 elements. Gag processing in the single mutants was comparable to wild type, as measured in immunoblot experiments by the ratios of the 27 and 37 Kd species (Fig. 4). The solubility of the mutants was also comparable to wild type, as evidenced by the levels of Gag in the soluble and insoluble (pellet) fractions (Fig. 4). We extracted Gag from the pellets with various concentrations of urea (from 1 to 8 M), and both the wild type and mutant proteins were solubilized to approximately the same extent by the various urea concentrations (data not shown). We therefore concluded that the Y68C and D252N mutations did not significantly affect Gag processing and solubility.

**Effects of *gag* mutations on target bias.** We previously demonstrated that a short domain in the integrase C-terminus is required for Ty5 to integrate into silent regions of the yeast genome (Gai and Voytas, 1998; Xie et al., 2001). Mutations in this domain also decrease transposition frequency ~ 4-fold. We were curious as to whether the *gag* mutations identified here that increase transposition frequency also affect target choice. Target bias was measured by our assay that monitors integration into a plasmid with an *HMR* locus, a preferred Ty5 target (Gai and Voytas, 1998). Plasmid insertions give rise to white colonies relative to the red or red-sectored colonies that result from chromosomal integration events. A change in the percentage of white colonies reflects an alteration in target specificity. The Y68C and D252N mutations did not change the target bias compared to wild type (Fig. 5).

**Features of the zinc finger important for transposition.** In sequence comparisons of NCp zinc fingers, two significant differences were observed between Ty5 and 22 related Ty1/*copia* retrotransposons (Fig. 6A). First, Ty5 is the only retrotransposon with three (rather than four) amino acids in the C-H interval of the CCHC motif. Furthermore, in most retrotransposons (18 out of 22 or 82%), G is located just before H; this is not the case for Ty5. Because the D252N mutation inserts a residue with a bulkier side chain, we were curious to know whether changes in the spatial organization of the zinc finger underlie the increase in transposition frequency caused by this mutation. To test this, we inserted G before H in a wild type Ty5 element so that the zinc finger domain matched the evolutionarily conserved consensus sequence. The G insertion did not affect Ty5 transposition frequency (data not shown). We therefore reasoned that the D252N mutation does have its primary impact on Ty5 transposition by altering the spatial organization of the C-H interval.

The second significant difference between Ty5 and the other retrotransposons concerns the potential for hydrogen bonding between the zinc finger motif and nucleic acids (Fig. 6B). The average probability for hydrogen bond formation was calculated for each interval in the finger motif. The Ty5 zinc finger has significantly less capacity for hydrogen bonding relative to the other retrotransposons. This is especially true for the C-C and C-H intervals. The D252N mutation significantly increases the potential to form hydrogen bonds. In addition, the acidic properties of D in wild type Ty5 might repel nucleic acids, whereas N is neutral. To test whether the chemical properties of the zinc finger affect its biological

activity, we created a D252R mutation. Among amino acids, R has the highest hydrogen bonding potential. The D252R mutation increased Ty5 transposition about 7-fold (Table 3). The double mutant – D252R, Y68C – increased Ty5 transposition about 33-fold. We also generated a D252K mutation, which when combined with Y68C, caused an approximately 40-fold increase in transposition. These results suggest that D inhibits the normal function of the Ty5 zinc finger motif, likely by preventing interactions with nucleic acids. Similar to our observation with the D252N, Y68C double mutant, the fold increase in transposition for other double mutants was the multiple of the fold increase for each single mutant. This indicates that the mutations in the zinc finger domain affect transposition independently of the Y68C mutation.

## DISCUSSION

Most eukaryotic genomes harbor retrotransposon families ranging from those that are highly successful to those that are likely extinct. We previously characterized the number and diversity of Ty5 elements in various *S. cerevisiae* strains, most of which have several degenerate insertions. For example, seven Ty5 insertions are present in the completed sequence of strain S288C, all of which are either truncated elements, solo LTRs or LTR fragments (Kim et al., 1998). The number of these degenerate insertions suggests that Ty5 elements were once active in *S. cerevisiae*. In the closely related species, *S. paradoxus*, many strains have a few, apparently full-length elements. Of two such elements that were sequenced, only one, Ty5-6p, showed transposition activity after being expressed in *S.*

*cerevisiae* (Zou et al., 1996). However, the transposition frequency of Ty5-6p was 1000-fold lower than Ty1. We wanted to understand the reasons for this low transposition activity.

Host-encoded factors are important regulators of retrotransposition. A number of host genes, for example, affect transposition of the yeast Ty1 elements (Curcio and Garfinkel, 1999; Voytas and Boeke, In Press). Host factors also influence Ty5 transposition. Genetic differences between the *S. cerevisiae* strains YPH499 and W303 result in a 10-fold difference in Ty5 transposition frequency. Because the functional Ty5-6p element was recovered from *S. paradoxus*, it is possible that *S. cerevisiae* host factors negatively impact transposition. However, in preliminary experiments, we found that Ty5 transposition frequencies are comparable in these two yeast species (X. Gao and D.F. Voytas, unpublished data). Ty5-6p may itself be inefficient in transposition. We tested this latter hypothesis by identifying Ty5 mutations that increase transposition. We hoped that by identifying residues that negatively affect transposition, we could gain insight into mechanisms of Ty5 transposition or interactions between this retroelement and its host's cellular environment.

***gag* mutations increase cDNA synthesis.** Two Ty5 Gag mutants (Y68C and D252N) were identified that increase transposition approximately 6-fold. Our transposition assay measures the $His^+$ phenotype generated when Ty5 cDNA enters the *S. cerevisiae* genome. cDNA has two pathways to insert into target DNA molecules: it can integrate into the chromosome using the Ty5-encoded integrase or it can recombine with the Ty5 element on the donor plasmid (Ke and Voytas, 1997; Ke and Voytas, 1999). The recombination

pathway requires the host gene *RAD52* (Ke and Voytas, 1999). We first tested which of these pathways are affected by the *gag* mutations. A comparable fold increase in transposition frequency was found in both wild type and *rad52 S. cerevisiae* strains, and this was observed for each Gag mutant compared to wild type Ty5. In addition, the percentage of integration and recombination relative to the total number of $His^+$ cells did not change. This suggests that the mutations did not alter the efficiency of either the integration or recombination pathways. It is very likely that steps prior to integration and recombination (e.g. cDNA synthesis) were affected by the mutations. An effect on cDNA synthesis is supported by the observation that protein levels and processing, as well as integration specificity, were unchanged by the mutations.

We have not been able to detect Ty5 cDNA by southern hybridization (N. Ke and D.F. Voytas, unpublished data), suggesting that very low amounts of cDNA are present in cells expressing Ty5. This might be one reason for Ty5's low transposition activity. A more sensitive PCR method, therefore, was developed to detect Ty5 cDNA. The specificity of the assay was confirmed by several controls: 1) The PCR assay failed to amplify the Ty5 donor plasmid; 2) The assay also failed to amplify DNA extracted from yeast strains in which Ty5 transcription was not induced. These first two controls ruled out the possibility of non-specific amplification from Ty5 in the donor plasmid; 3) Amplification was not simply due to growth on galactose, which induces Ty5 transcription, as extracts from galactose-grown strains without Ty5 failed to yield an amplification product; 4) An internal control was amplified in each reaction, indicating that PCR was taking place. PCR products originating

from cDNA were only amplified from genomic DNA extracted from yeast strains with Ty5 that had been grown on galactose; 5) The PCR products were of the size predicted for templates from which the artificial intron was removed. Using a competitor template, we were able to determine that the D252N mutation caused a 2.6-fold increase in cDNA levels. Y68C caused a 2.2-fold increase, whereas the double mutant increased cDNA 4.8-fold. Our direct physical measurements of cDNA, therefore, support the idea that cDNA levels are increased in the mutants. Note that the increase in levels of cDNA is not commensurate with the observed increase in transposition. This may be because the assays measure different steps in replication or because there are limitations in the sensitivity of the PCR assay. However, the results of both assays are consistent.

RT carries out cDNA synthesis; however, cDNA levels do not depend on RT alone. Other factors encoded by the retrotransposon and host cell are part of the replication complex and can affect cDNA levels. NCp, for example, participates in cDNA synthesis, and one of the Ty5 mutations, D252N, is located in the zinc finger of NCp. This suggests that Ty5 NCp carries out a role in cDNA synthesis similar to the roles of NCp's from other retroelements (Vogt, 1997). As described below, the D252N mutation might optimize the zinc finger domain of NCp for nucleic acid binding, making it more efficient in RNA template packaging, tRNA primer annealing or strand transfer, and thereby increasing its effectiveness in cDNA synthesis.

The Y68C mutation is located near the N-terminus of Gag. The role of this mutation in cDNA synthesis is less clear. One possibility is that it positively affects virus-like

particle formation, thereby increasing the total amount of cDNA synthesized. However, this mutation did not alter the processing or solubility of Ty5 Gag. The Y68C mutation replaces an aromatic, polar uncharged residue (Y) with a hydrophobic residue (C). This could change the local structure of Gag. Post-translational modifications could also be affected: the hydroxyl group of Y could be phosphorylated or sulfated, whereas C can undergo cysteinylation, oxidation, and glutathionylation or can form disulfide bonds. Changes in post-translational modifications might make Gag better at particle formation, and thereby result in the synthesis of more cDNA.

Host factors that are part of the retroelement replication complex are still not well defined. Strain differences between YPH499 and W303 result in a several-fold difference in Ty5 transposition. Because all three Ty5 mutants displayed the same fold difference in transposition between the two strains, host factor differences and the mutations in *gag* appear to act at independent steps.

**Mutations in the zinc finger implicate a role for hydrogen bonding in NCp function.** NCp is the primary protein in the retroelement nucleocore. NCp binds tightly to both the genomic RNA of retroelements and their tRNA primer (Barat et al., 1993). Binding is carried out by one or two highly conserved CCHC type zinc fingers. The zinc fingers are flanked by basic amino acids that also interact with template and primer RNAs. In Ty5, the consensus finger motif differs slightly from most retroelements ($CX_2CX_3HX_4C$ vs. $CX_2CX_3GHX_4C$). Some retroelements like Ty1 of *S. cerevisiae* do not have a conserved zinc finger; rather, three stretches of basic amino acids in the C terminus of Gag perform the

required nucleic acid chaperon activity (Cristofari et al., 2000). Thus, although there are exceptions, the use of zinc fingers is the most widespread means of interacting with nucleic acids. The nucleic acid binding activity of NCp is important for a number of steps in replication, including RNA dimerization (Barat et al., 1993; Feng et al., 1996; Prats et al., 1988), primer and template RNA packaging (Berkowitz et al., 1996), annealing of the tRNA primer to the template RNA (Chan and Musier-Forsyth, 1997; Lapadat-Tapolsky et al., 1995; Remy et al., 1998), initiating reverse transcription (Cristofari et al., 2000; Rong et al., 1998), transferring strong stop DNA (Allain et al., 1994; Cristofari et al., 2000; Darlix et al., 1993; Hsu et al., 2000) and ensuring fidelity of cDNA synthesis (Gorelick et al., 1999).

The D252N mutation is located just before the conserved H in the zinc finger domain of Ty5 NCp. This mutation, therefore, occurs in the interval in which spacing differs in Ty5 compared to other retroelements. In particular, the conserved G, which is located just before H in the zinc finger of most retrotransposons and retroviruses, is missing in Ty5 (Fig. 6A) (Summers, 1991). We suspected that in the D252N mutation, the additional $NH_2$ group from the N side chain might make the Ty5 zinc finger resemble and function like other retroelement zinc fingers that have the G insertion. To test this hypothesis, we inserted a G before H in Ty5 NCp; however, this consensus zinc finger did not increase Ty5 transposition (Y. Chin and D.F. Voytas, unpublished data). The effect of the D252N mutation, therefore, is likely not due to local structural changes, and the conserved G does not appear to have a critical role in substrate binding, at least in Ty5 NCp.

An alternative explanation for the increase in transposition of the D252N mutation is that it changed a chemical property of the zinc finger. Ty5 has only one R and N in the H-C interval (Fig. 6A). The zinc fingers of other Ty1/*copia* retrotransposons often have in their C-C and C-H intervals R, K or N – three amino acids with a strong potential to form hydrogen bonds with nucleic acids. The hydrogen bonding potential of the Ty5 zinc finger is only 64.6% that of other Ty1/*copia* retroelements (calculated as the sum of Ty5 hydrogen bonding potential across all intervals divided by the sum of hydrogen bonding potential in the Ty1/*copia* group elements) (Fig. 6B). The D252N mutation increases the hydrogen bonding potential in the C-H interval to 71.3%. This might strengthen the zinc finger's ability to bind RNA, and consequently, the D252N mutation might make RNA template packaging and cDNA synthesis more efficient.

To test our hypothesis that hydrogen bonding plays a role in NCp function, we mutated D to R as an alternative means of increasing the hydrogen bonding potential in the C-H interval. As predicted, the D252R mutation increased Ty5 transposition about 7-fold, and the D252R, Y68C double mutant transposed 33-fold more efficiently. Another double mutation, D252K, Y68C, which also has a higher hydrogen bonding potential in the zinc finger, showed a 40-fold increase in Ty5 transposition. The transposition increase caused by these D substitutions is likely due to hydrogen bonding with bases and not phosphate group contacts. R and K, which are basic, should make stronger phosphate contacts than neutral amino acids such as N. If phosphate contacts were important, the D252R and D252K mutations should have the most impact on zinc finger function. However, all three

substitutions increased Ty5 transposition by the same magnitude. The D in the wild type Ty5 zinc finger provides no potential to hydrogen bond with G and U bases and a very low bonding potential for C and A bases. Its acidic property might even repel nucleic acids. The transposition increase caused by substituting D with other strong hydrogen bonding residues suggests an important role of hydrogen bonding in forming complexes between the zinc finger and RNA.

**Relationship of the Ty5 zinc finger to other zinc finger motifs.** Our observations regarding the Ty5 NCp zinc finger have bearing on the function of other zinc finger motifs. These include the CCHH motifs ($CX_{2-5}CX_{12}HX_{3-5}H$) found in proteins such as the mouse transcription factor Zif268 and the CCCC motifs ($CX_2CX_{13}CX_2C$) found in proteins such as the glucocorticoid receptor (Iuchi, 2001; Klug and Schwabe, 1995). These motifs have large middle intervals of 12-13 amino acids compared to the 3-4 amino acids in the CCHC motif, and their C-terminal halves form α-helices (Luisi et al., 1991; Pavletich and Pabo, 1991). In contrast, the crystal structure of retroviral NCp zinc fingers do not reveal any obvious secondary structure (Morellet et al., 1998; Schuler et al., 1999; Summers et al., 1992). This structural difference might distinguish the retroviral zinc fingers from other finger domains, and reflect a role in binding single versus double-stranded nucleic acids.

Amino acids in the CCHH and CCCC motifs also hydrogen bond to nucleic acids. In the CCHH zinc finger of Zif268 (Pavletich and Pabo, 1991), hydrogen bonding occurs predominantly between the α-helix of the zinc finger and the G-rich DNA strand in the major groove. Nine out of twelve of these interactions are hydrogen bonds with bases, and five of

them involve interactions between R and guanine. Therefore, hydrogen bonding between residues within the zinc finger and nucleic acids are important for in the function of both the CCHH and retroelement CCHC motifs. Other studies of CCHC zinc fingers of retroviral NCp revealed that flanking basic amino acids are also important for function (Takahashi et al., 2001). These basic residues may contact phosphates in the DNA, as has been shown for basic residues in the linker region of the CCHH type zinc finger (Wolfe et al., 2000). It is interesting to note that there is a conserved D in the C-H interval of the three zinc fingers repeats of Zif268 and other CCHH finger motifs (Iuchi, 2001). Although D has a low hydrogen bonding potential with nucleic acids, D plays a role in specifically binding adenine or cytosine residues in CCHH fingers (Wolfe et al., 2000). In contrast, nucleic acid binding by NCp zinc fingers is relatively non-specific and of weaker affinity than the other finger motifs (Klug and Schwabe, 1995). The D in C-H interval that was mutated in Ty5 is not conserved in the retroviruses (Summers, 1991) or in retrotransposon zinc fingers (Fig 6). The D at this position in Ty5 likely has a minimal role in DNA binding.

Previous work has implicated amino acids flanking the zinc finger in binding nucleic acids. Our study suggests that the hydrogen bonding potential of amino acids inside the CCHC motif also play an important role in zinc finger function. Although there are structural differences between the zinc fingers of cellular enzymes and the retroviral NCp zinc finger, the underlying mechanism of nucleic acid binding is likely conserved. Our findings suggest ways to increase the binding affinity of zinc finger domains, which is one of the current challenges in engineering nucleic acid binding proteins.

## ACKNOWLEDGEMENTS

We thank Yvette Chin for generating the Ty5 G insertion mutant. This work was supported by NIH grant GM51425. This is Journal Paper no. J-12345 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, project no. 3383 and was supported by Hatch Act and State of Iowa funds.

## REFERENCES

Allain, B., Lapadat-Tapolsky, M., Berlioz, C., and Darlix, J. L. (1994). Transactivation of the minus-strand DNA transfer by nucleocapsid protein during reverse transcription of the retroviral genome, Embo J *13*, 973-81.

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K. (1987). Current Protocols in Molecular Biology (New York, Greene/Wiley Interscience).

Barat, C., Schatz, O., Le Grice, S., and Darlix, J. L. (1993). Analysis of the interactions of HIV1 replication primer tRNA(Lys,3) with nucleocapsid protein and reverse transcriptase, J Mol Biol *231*, 185-90.

Berkowitz, R., Fisher, J., and Goff, S. P. (1996). RNA packaging, Curr Top Microbiol Immunol *214*, 177-218.

Boeke, J. D., Eickbush, T., Sandmeyer, S. B., and Voytas, D. F. (2000). Pseudoviridae. In Virus Taxonomy: Seventh Report of the International Committee on Taxonomy of Viruses, M. H. V. van Regenmortel, C. M. Fauquet, D. H. L. Bishop, E. B. Carsten, M. K. Estes, S. M. Lemon, J. Maniloff, M. A. Mayo, D. J. McGeoch, C. R. Pringle, and R. B. Wickner, eds. (New York, Academic Press), pp. 349-57.

Chan, B., and Musier-Forsyth, K. (1997). The nucleocapsid protein specifically anneals tRNALys-3 onto a noncomplementary primer binding site within the HIV-1 RNA genome in vitro, Proc Natl Acad Sci U S A *94*, 13530-5.

Cristofari, G., Ficheux, D., and Darlix, J. L. (2000). The GAG-like protein of the yeast Ty1 retrotransposon contains a nucleic acid chaperone domain analogous to retroviral nucleocapsid proteins, J Biol Chem *275*, 19210-7.

Curcio, M. J., and Garfinkel, D. J. (1991). Single-step selection for Ty1 element retrotransposition, Proc Natl Acad Sci U S A *88*, 936-40.

Curcio, M. J., and Garfinkel, D. J. (1999). New lines of host defense: inhibition of Ty1 retrotransposition by Fus3p and NER/TFIIH, Trends Genet *15*, 43-5.

Darlix, J. L., Vincent, A., Gabus, C., de Rocquigny, H., and Roques, B. (1993). Trans-activation of the 5' to 3' viral DNA strand transfer by nucleocapsid protein during reverse transcription of HIV1 RNA, C R Acad Sci III *316*, 763-71.

Feng, Y. X., Copeland, T. D., Henderson, L. E., Gorelick, R. J., Bosche, W. J., Levin, J. G., and Rein, A. (1996). HIV-1 nucleocapsid protein induces "maturation" of dimeric retroviral RNA in vitro, Proc Natl Acad Sci U S A *93*, 7577-81.

Gai, X., and Voytas, D. F. (1998). A single amino acid change in the yeast retrotransposon Ty5 abolishes targeting to silent chromatin, Mol Cell *1*, 1051-5.

Gorelick, R. J., Fu, W., Gagliardi, T. D., Bosche, W. J., Rein, A., Henderson, L. E., and Arthur, L. O. (1999). Characterization of the block in replication of nucleocapsid protein zinc finger mutants from moloney murine leukemia virus, J Virol *73*, 8185-95.

Hsu, M., Rong, L., de Rocquigny, H., Roques, B. P., and Wainberg, M. A. (2000). The effect of mutations in the HIV-1 nucleocapsid protein on strand transfer in cell-free reverse transcription reactions, Nucleic Acids Res *28*, 1724-9.

Irwin, P. A., and Voytas, D. F. (2001). Expression and processing of proteins encoded by the *Saccharomyces* retrotransposon Ty5, J Virol *75*, 1790-97.

Iuchi, S. (2001). Three classes of C2H2 zinc finger proteins, Cell Mol Life Sci *58*, 625-35.
Ke, N., Gao, X., Keeney, J. B., Boeke, J. D., and Voytas, D. F. (1999). The yeast retrotransposon Ty5 uses the anticodon stem-loop of the initiator methionine tRNA as a primer for reverse transcription, Rna *5*, 929-38.

Ke, N., and Voytas, D. F. (1997). High frequency cDNA recombination of the *Saccharomyces* retrotransposon Ty5: The LTR mediates formation of tandem elements, Genetics *147*, 545-56.

Ke, N., and Voytas, D. F. (1999). cDNA of the yeast retrotransposon Ty5 preferentially recombines with substrates in silent chromatin, Mol Cell Biol *19*, 484-94.

Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A., and Voytas, D. F. (1998). Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence, Genome Res *8*, 464-78.

Klug, A., and Schwabe, J. W. (1995). Protein motifs 5. Zinc fingers, Faseb J *9*, 597-604.
Lapadat-Tapolsky, M., Pernelle, C., Borie, C., and Darlix, J. L. (1995). Analysis of the nucleic acid annealing activities of nucleocapsid protein from HIV-1, Nucleic Acids Res *23*, 2434-41.

Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R., and Sigler, P. B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA, Nature *352*, 497-505.

Mandel-Gutfreund, Y., Schueler, O., and Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA- complexes: in search of common principles, J Mol Biol *253*, 370-82.

Morellet, N., Demene, H., Teilleux, V., Huynh-Dinh, T., de Rocquigny, H., Fournie-Zaluski, M. C., and Roques, B. P. (1998). Structure of the complex between the HIV-1 nucleocapsid protein NCp7 and the single-stranded pentanucleotide d(ACGCC), J Mol Biol *283*, 419-34.

Pavletich, N. P., and Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A, Science *252*, 809-17.

Prats, A. C., Sarih, L., Gabus, C., Litvak, S., Keith, G., and Darlix, J. L. (1988). Small finger protein of avian and murine retroviruses has nucleic acid annealing activity and positions the replication primer tRNA onto genomic RNA, Embo J *7*, 1777-83.

Remy, E., de Rocquigny, H., Petitjean, P., Muriaux, D., Theilleux, V., Paoletti, J., and Roques, B. P. (1998). The annealing of tRNA3Lys to human immunodeficiency virus type 1 primer binding site is critically dependent on the NCp7 zinc fingers structure, J Biol Chem *273*, 4819-22.

Rong, L., Liang, C., Hsu, M., Kleiman, L., Petitjean, P., de Rocquigny, H., Roques, B. P., and Wainberg, M. A. (1998). Roles of the human immunodeficiency virus type 1 nucleocapsid protein in annealing and initiation versus elongation in reverse transcription of viral negative-strand strong-stop DNA, J Virol *72*, 9353-8.

Schuler, W., Dong, C., Wecker, K., and Roques, B. P. (1999). NMR structure of the complex between the zinc finger protein NCp10 of Moloney murine leukemia virus and the single-stranded pentanucleotide d(ACGCC): comparison with HIV-NCp7 complexes, Biochemistry *38*, 12984-94.

Summers, M. F. (1991). Zinc finger motif for single-stranded nucleic acids? Investigations by nuclear magnetic resonance, J Cell Biochem *45*, 41-8.

Summers, M. F., Henderson, L. E., Chance, M. R., Bess, J. W., Jr., South, T. L., Blake, P. R., Sagi, I., Perez-Alvarado, G., Sowder, R. C., 3rd, Hare, D. R., and et al. (1992). Nucleocapsid zinc fingers detected in retroviruses: EXAFS studies of intact viruses and the solution-state structure of the nucleocapsid protein from HIV-1, Protein Sci *1*, 563-74.

Takahashi, K., Baba, S., Koyanagi, Y., Yamamoto, N., Takaku, H., and Kawai, G. (2001). Two basic regions of NCp7 are sufficient for conformational conversion of HIV-1 dimerization initiation site from kissing-loop dimer to extended-duplex dimer, J Biol Chem *276*, 31274-8.

Vogt, V. M. (1997). Retroviral virions and genomes. In Retroviruses, J. Coffin, S. H. Hughes, and H. E. Varmus, eds. (Cold Spring Harbor, Cold Spring Harbor Laboratory Press), pp. 27-70.

Voytas, D. F., and Boeke, J. D. (In Press). Ty1 and Ty5. In Mobile DNA II (Washington, DC, American Society for Microbiology).

Wolfe, S. A., Nekludova, L., and Pabo, C. O. (2000). DNA recognition by Cys2His2 zinc finger proteins, Annu Rev Biophys Biomol Struct *29*, 183-212.

Xie, W., Gai, X., Zhu, Y., Zappulla, D. C., Sternglanz, R., and Voytas, D. F. (2001). Targeting of the Yeast Ty5 Retrotransposon to Silent Chromatin Is Mediated by Interactions between Integrase and Sir4p, Mol Cell Biol *21*, 6606-14.

Zhu, Y., Zou, S., Wright, D., and Voytas, D. (1999). Tagging chromatin with retrotransposons: target specificity of the *Saccharomyces* Ty5 retrotransposon changes with the chromosomal localization of Sir3p and Sir4p, Genes & Dev *13*, 2738-49.

Zou, S., Ke, N., Kim, J. M., and Voytas, D. F. (1996). The *Saccharomyces* retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci, Genes Dev *10*, 634-45.

Zou, S., Wright, D. A., and Voytas, D. F. (1995). The *Saccharomyces* Ty5 retrotransposon family is associated with origins of DNA replication at the telomeres and the silent mating locus *HMR*, Proc Natl Acad Sci U S A *92*, 920-4.

## FIGURE LEGENDS

**Fig 1. The effect of Ty5 mutations on integration and recombination.** Transposition (e.g. as presented in Table 1 and 2) is the sum of the integration and recombination frequencies. Bars in the graph represent the percentage of integration relative to total transposition for each Ty5 genotype. Grey and white bars denote integration in YPH499 and W303 strains, respectively. Stippled bars denote recombination.

**Fig 2. A PCR assay for cDNA quantification.** (A) The Ty5 donor element and Ty5 cDNA differ by the presence of an artificial intron (AI) in the *HIS3* marker gene. Note that *HIS3* is inserted in the opposite orientation relative to Ty5. Primer 1 (DVO208) spans the AI in the Ty5 donor element and can not generate a PCR amplification product. However, Primer 1 pairs to Ty5 cDNA, and in conjunction with Primer 2 (DVO200), yields a PCR product of 580 bp. A cloned Ty5 DNA fragment in pXG28 serves as an internal control. This DNA fragment has a 60 bp deletion, such that the size of the product amplified by Primers 1 and 2 is shorter (520 bp vs. 580 bp for Ty5 cDNA). (B) Ty5 cDNA was detected by PCR in yeast strains expressing Ty5. Lane 1, marker; lane 2, PCR products amplified from DNA extracted from a galactose-induced strain without Ty5; lane 3, PCR products amplified from DNA extracted from an uninduced strain with Ty5; lane 4, PCR products amplified from DNA extracted from a galactose-induced strain with Ty5. Competitor (0.01

ng) was added to each PCR reaction. When Ty5 cDNA was present in the cellular extract (i.e. lane 4), less competitor amplification product was observed. This is likely because the primers are competed away by the cDNA.

**Fig 3. The effect of Ty5 mutations on cDNA levels.** (A) Results of quantitative, competitive PCR assays using DNA extracted from strains with wild type or mutant Ty5 elements. The pXG28 internal control (ranging from 0.01 to 0.08 ng) was added to each PCR reaction. All PCR reactions in a given experiment contained the same amount of yeast DNA. (B) Quantification of cDNA levels determined in part A.

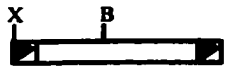
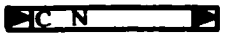
**Fig 4. Ty5 Gag expression, processing and solubility.** An epitope tag ($RGSH_6$) was inserted into the N-terminus of Gag-Pol in wild type and mutant Ty5 elements. Cells were lysed by the glass bead method, and the lysate was separated by centrifugation. The supernatant (S) and pellet (P) fractions were loaded onto a 10% polyacrylamide gel. Whole cell lysates were also analyzed. The separated proteins were transferred to a nitrocellulose membrane and probed with an antibody that recognizes the $RGSH_6$ epitope. The antibodies non-specifically cross-react to a protein of slightly lower molecular weight than Gag-p37; this band is particularly evident in the supernatant (S) lanes.

**Fig 5. Target specificity of the Ty5 mutants.** A plasmid-based targeting assay was used to measure Ty5 integration specificity (Gai and Voytas, 1998). The plasmid carries an *HMR*

locus, one of Ty5's preferred integration sites (Zou et al., 1996). No significant difference in target specificity was observed between wild type Ty5 and the mutants.

**Fig 6. Features of the NCp zinc finger domain of Ty5 and other Ty1/*copia* retrotransposons.** (A) Amino acid sequence alignment of the NCp CCHC motif of Ty5 and 22 other retrotransposons in the Ty1/*copia* group. (B) The hydrogen bonding potential for amino acids with bases was calculated for each interval of the zinc finger motif (see Materials and Methods). An average value is provided for the Ty1/*copia* retrotransposons.

**Table 1. Two *gag* mutations (Y68C, D252N) increase Ty5 transposition frequency.**

| Plasmid [a] | | Transposition [b] frequency | Fold increase |
|---|---|---|---|
| pNK254 |   | 1.9E-5 | 1.0 |
| pXW98 |  | 11.6E-5 | 6.1 |
| pDR3 |  | 1.9E-5 | 1.0 |
| pDR4 |  | 9.6E-5 | 5.6 |
| pXW97 |  | 10.6E-5 | 5.6 |
| pDR6 |  | 1.9E-5 | 1.0 |
| pDR5 |  | 9.2E-5 | 4.8 |
| pDR14 |   | 67.9E-5 | 35.7 |

[a] Ty5 elements are drawn schematically. Arrowheads represent the LTRs; the wild-type element pNK254, is white and the mutant elements are black or gray. The region swapped in pDR3, pDR4, pDR5 and pDR6 represents the 3.0 kb *Xho*I / *Bsp*MII fragment. For the double mutant, pDR14, C and N represent the amino acid sequence changes at Y68 and D252, respectively.

[b] Values represent the average of results obtained from three independent transformants.

**Table 2. Transposition frequencies of wild type and mutant Ty5 elements in different yeast strains.**

| Strain | Ty5 genotype | Transposition or integration[a] | Fold increase |
|---|---|---|---|
| YPH499 | wild type | $(1.90 \pm 0.70) \times 10^{-5}$ | 1.00 |
| | D252N | $(10.69 \pm 0.67) \times 10^{-5}$ | 5.63 |
| | Y68C | $(9.21 \pm 0.95) \times 10^{-5}$ | 4.84 |
| | D252N Y68C | $(67.89 \pm 6.96) \times 10^{-5}$ | 35.73 |
| W303 | wild type | $(0.73 \pm 0.04) \times 10^{-5}$ | 1.00 |
| | D252N | $(4.76 \pm 1.51) \times 10^{-5}$ | 6.52 |
| | Y68C | $(1.98 \pm 0.09) \times 10^{-5}$ | 2.71 |
| | D252N Y68C | $(12.33 \pm 1.01) \times 10^{-5}$ | 16.88 |
| YPH499 *rad52Δ* | wild type | $(0.97 \pm 0.52) \times 10^{-6}$ | 1.00 |
| | D252N | $(7.14 \pm 0.82) \times 10^{-6}$ | 7.36 |
| | Y68C | $(3.02 \pm 0.63) \times 10^{-6}$ | 3.11 |
| | D252N Y68C | $(42.17 \pm 17.00) \times 10^{-6}$ | 43.47 |
| W303 *rad52Δ* | wild type | $(0.15 \pm 0.04) \times 10^{-6}$ | 1.00 |
| | D252N | $(1.31 \pm 0.26) \times 10^{-6}$ | 8.91 |
| | Y68C | $(0.59 \pm 0.45) \times 10^{-6}$ | 4.01 |
| | D252N Y68C | $(4.72 \pm 0.84) \times 10^{-6}$ | 32.11 |

[a] Transposition is the sum of the integration and recombination in the wild type strains. In *rad52Δ* strains, recombination is eliminated and only Ty5 integration is measured (Ke and Voytas, 1997).

**Table 3. Transposition frequencies of NCp zinc finger mutants.**

| Strain | Ty5 genotype | Transposition Frequency | Fold increase |
|---|---|---|---|
| YPH499 | wild type | $(0.35 \pm 0.04) \times 10^{-5}$ | 1.0 |
| | D252R | $(2.39 \pm 0.37) \times 10^{-5}$ | 6.8 |
| | D252R,Y68C | $(11.18 \pm 4.84) \times 10^{-5}$ | 32.9 |
| | D252K, Y68C | $(13.93 \pm 5.09) \times 10^{-5}$ | 39.8 |

Fig. 1

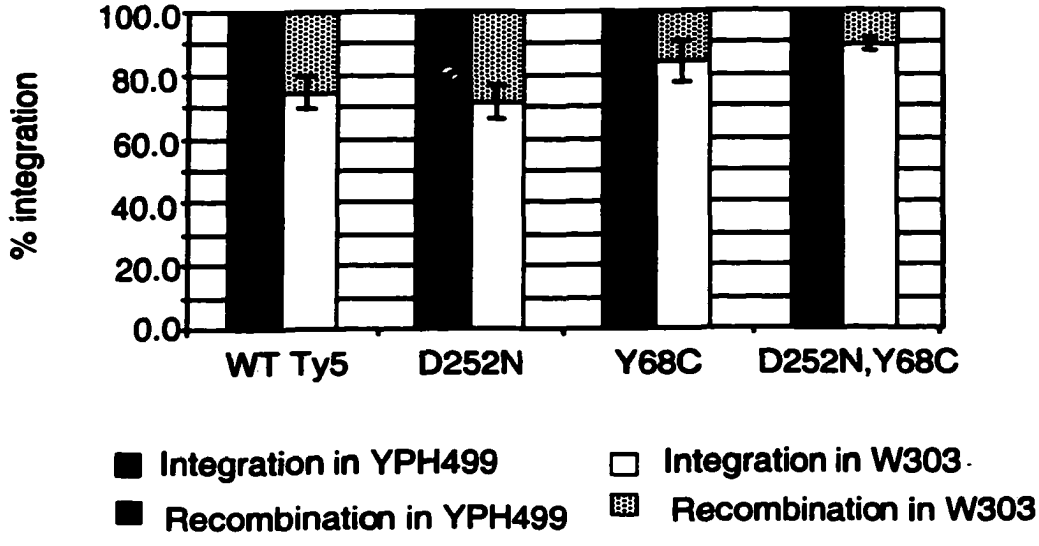
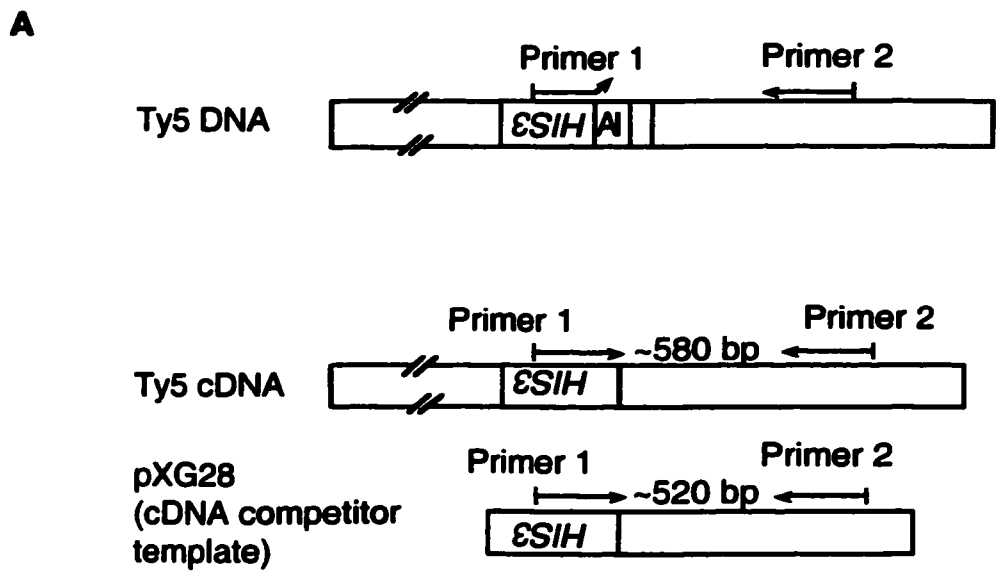

% integration
100.0
80.0
60.0
40.0
20.0
0.0
WT Ty5
D252N
Y68C
D252N,Y68C
Integration in YPH499
Recombination in YPH499
Integration in W303
Recombination in W303

Fig. 2


A
Primer 1
Primer 2
Ty5 DNA
εSIH
AI
Primer 1
Primer 2
~580 bp
Ty5 cDNA
εSIH
pXG28
(cDNA competitor
template)
Primer 1
Primer 2
~520 bp
εSIH

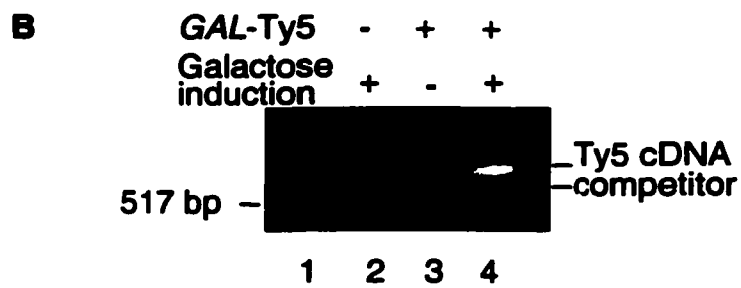

B
GAL-Ty5 - + +
Galactose
induction + - +
Ty5 cDNA
competitor
517 bp
1 2 3 4

Fig. 3

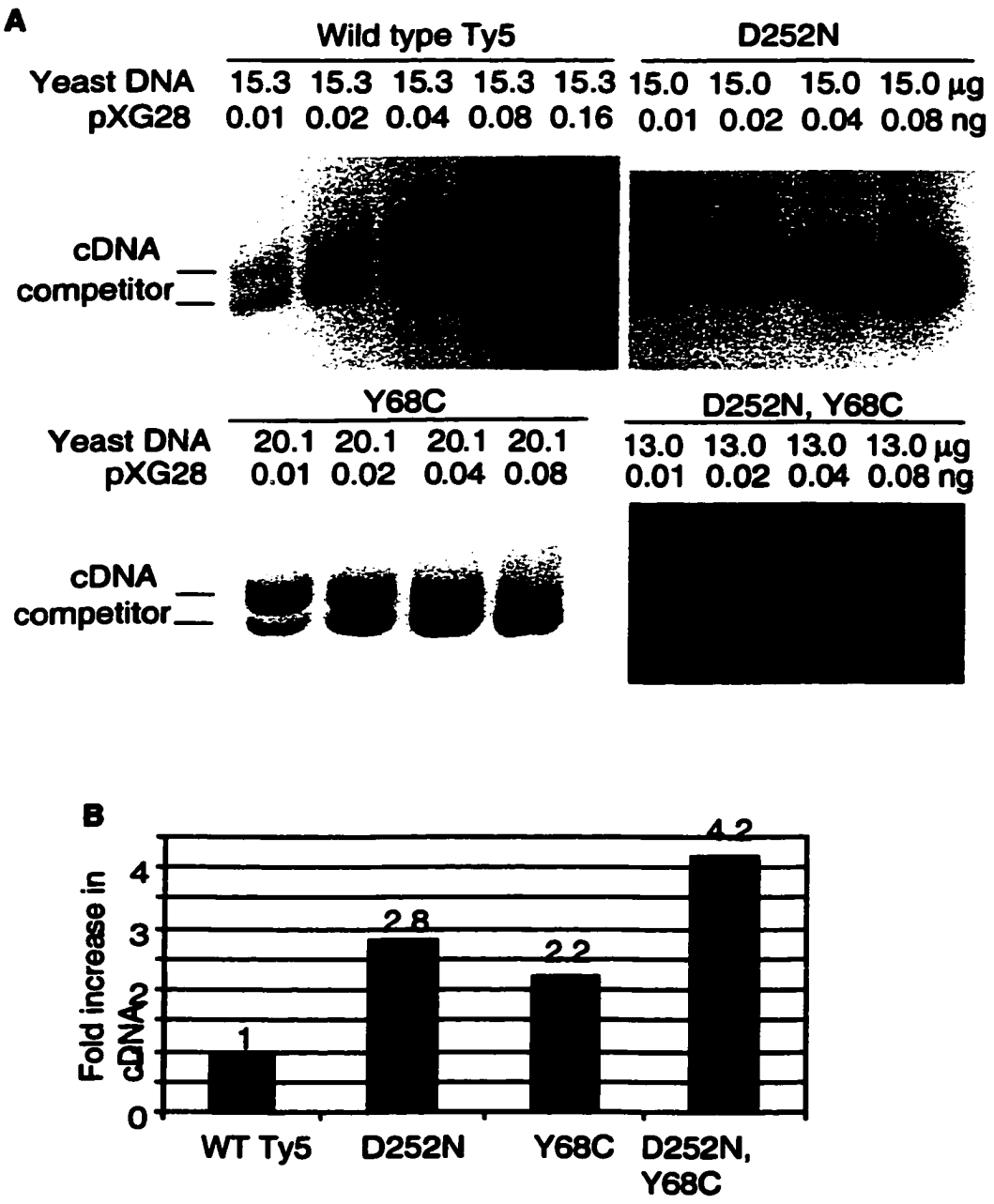

A
Wild type Ty5
D252N
Yeast DNA 15.3 15.3 15.3 15.3 15.3 15.0 15.0 15.0 15.0 µg
pXG28 0.01 0.02 0.04 0.08 0.16 0.01 0.02 0.04 0.08 ng
cDNA
competitor
Y68C
D252N, Y68C
Yeast DNA 20.1 20.1 20.1 20.1 13.0 13.0 13.0 13.0 µg
pXG28 0.01 0.02 0.04 0.08 0.01 0.02 0.04 0.08 ng
cDNA
competitor
B
Fold increase in cDNA
0
1
2
3
4
2.8
2.2
4.2
WT Ty5
D252N
Y68C
D252N, Y68C

Fig. 4

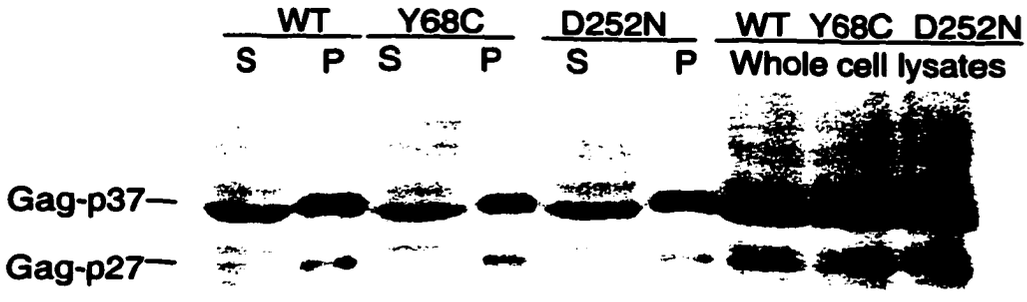

WT
Y68C
D252N
WT Y68C D252N
S P S P S P Whole cell lysates
Gag-p37—
Gag-p27—

Fig. 5

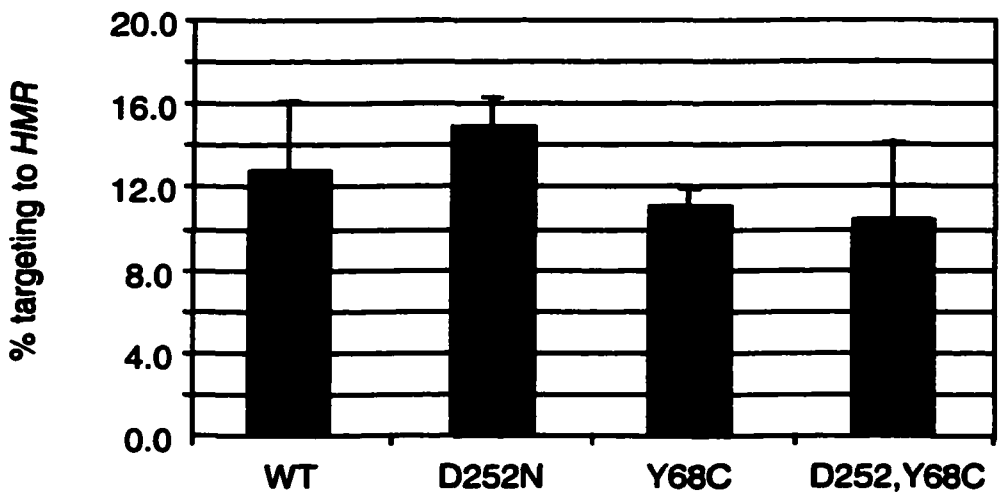
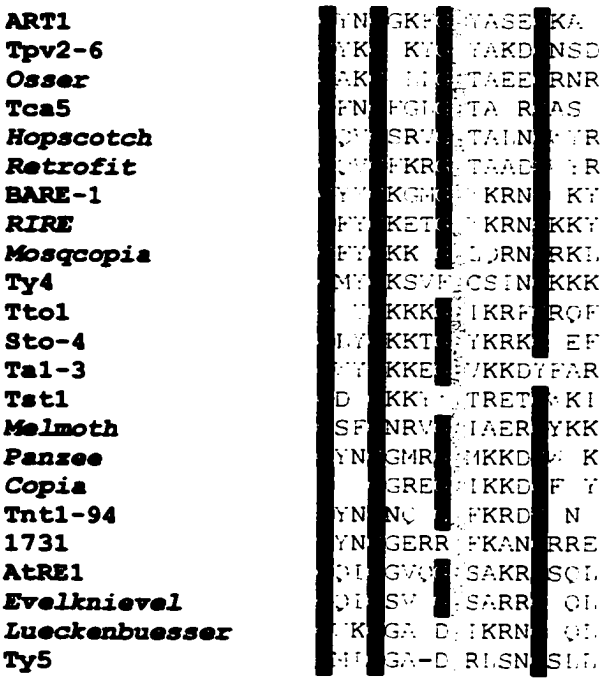
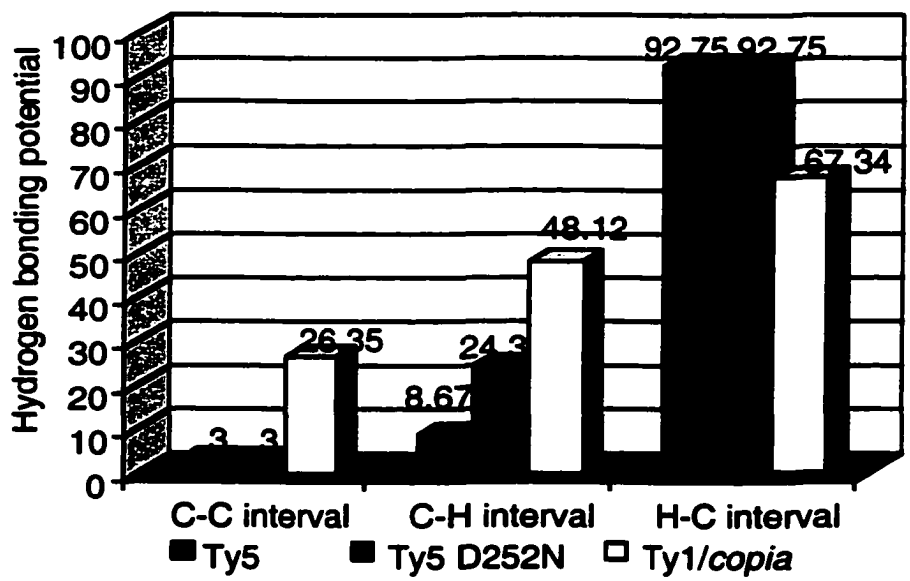


Fig. 6

A


ART1
Tpv2-6
Osser
Tca5
Hopscotch
Retrofit
BARE-1
RIRE
Mosqcopia
Ty4
Tto1
Sto-4
Ta1-3
Tst1
Melmoth
Panzee
Copia
Tnt1-94
1731
AtRE1
Evelknievel
Lueckenbuesser
Ty5

B


Hydrogen bonding potential
100
90
80
70
60
50
40
30
20
10
0
3
3
26.35
8.67
24.3
48.12
92.75
92.75
67.34
C-C interval
C-H interval
H-C interval
Ty5
Ty5 D252N
Ty1/copia

# CHAPTER III. TREE-BASED METHOD TO IDENTIFY PROTEIN FUNCTIONAL DOMAINS: CASE STUDIES USING DIVERGENT RETROTRANSPOSON PROTEINS AND THE CONSERVED Myb PROTEIN FAMILY

**A manuscript to be submitted to *Proc. Natl. Acad. Sci. U.S.A.***

**Xiang Gao[1], Kent Vander Velden[2] and Daniel F. Voytas[3]**

## ABSTRACT

Many protein families have undergone functional divergence, such that sublineages of the family now carry out unique biological roles. To identify the amino acids or sequence domains responsible for these unique functions, we developed a phylogenetic tree-based method that analyzes sequence alignments in sliding windows of different length. Only trees derived from windows supporting a predefined functional split are considered as candidates for determining functional specificity. The strongest candidates are those domains that show strong signals at the same position for windows of different length. We used our strategy to identify the sequence domains that are likely responsible for the recognition of different primers utilized in retroelements reverse transcription. Two domains surrounding the

---

[1] Primary researcher and author.
[2] Graduate student who implement the algorithm.
[3] Professor and corresponding author, Department of Zoology and Genetics, Iowa State University, Ames, IA 50011.

conserved reversed transcriptase 'primer grip', which interacts directly with the 3'-end of the HIV-1 primer were identified. Surrounding domains were identified that likely contact the primer/template complex. We also tested retroelement integrase sequences to identify the domain likely responsible for cDNA 3'-end processing. The same region was consistently identified from two independent datasets, namely Ty1/*copia* and Ty3/*gypsy* integrases. The candidate region spans the catalytic core domain and the C-terminal domain. Based on the HIV-1 integrase structure and results from our bioinformatics approach, a model for functional divergence in 3'-end processing is proposed. The broad application of this method was tested using highly conserved sequences that define the Myb gene family. The method identified amino acids responsible for DNA binding specificity.

## INTRODUCTION

Analyses of the genome sequences of a number of organisms have revealed that many proteins fall within well-defined superfamilies. More often than not, multiple members of each superfamily are found in different organisms, and in many cases, the root of these proteins can be traced to prokaryotes. The amplification of these gene sequences in evolution suggests that they evolved new functions and that they carry out a wide variety of cellular roles. Clearly, one goal of functional genomics is to identify the specific cellular roles carried out by given members of a protein superfamily. Many approaches have been undertaken, most of which involve high throughput microarray, or proteomic analysis, or involve genetic manipulation and mutant screens. However, it is likely that the key to understanding protein

manipulation and mutant screens. However, it is likely that the key to understanding protein function resides in the primary amino acid sequence. One bioinformatics goal is to develop tools to identify the amino acids responsible for functional diversity.

Our laboratory works on the ubiquitous group of mobile genetics elements, known as retrotransposons. Like any protein superfamily, retrotransposons are found in a wide variety of organisms and typically most organisms have multiple diverse retroelements. Despite the number of these diverse sequences, they by and large undergo a similar mechanism of replication. Retroelement replication involves reverse transcribing the element mRNA into cDNA and then integrating the cDNA into new sites in the genome. Despite the similarity in the roles carried out by retroelement proteins, these proteins have evolved some distinct functions. For example, some reverse transcriptase have evolved to use a wide variety of RNA primers, most notably different cellular tRNAs or tRNA fragments (Chapman et al., 1992; Ke et al., 1999; Kikuchi et al., 1986; Leis et al., 1993). Similarly, integrase, which inserts the linear cDNA into the host chromosome, carries out its reaction with some variations (Feuerbach et al., 1997). One variation is that the end of the cDNA is processed by removing some nucleotides prior to integration. We therefore felt that the retrotransposons, both because of the number and diversity of their sequences and the fact that they have evolved several examples of functional specificity, serve as good models to develop general methods to identify determinants of functional diversity.

In this manuscript, we describe a phylogenetic approach for identifying protein functional domains. As indicated above, we reason that some of the relevant information for

determining novel protein function resides in primary amino acid sequences. Therefore, we felt that a phylogenetic approach may be used to identify relevant amino acid sequences. We described a tree-based algorithm and accompanying software that can be utilized to dissect the functionally relevant domains in groups of related of proteins.

## RESULTS

**Phylogenetic method to discern functional diversity and *Split Tester* software.** We hypothesize that conserved amino acid sequences should distinguish proteins that carry out slightly different roles. Although there are many potential ways to identify these conserved sequences, we have chosen a phylogenetic approach. Our method begins with an amino acid sequence alignment and assumes that conservation of sequences involved in a particular function will be reflected in phylogenetic relationships among the aligned proteins. However, trees based on complete amino acid sequence alignments rarely separate proteins by known function. We predict that if phylogenetic trees are constructed from windows of the alignment, certain regions may show the expected grouping. Figure 1 illustrates a hypothetical case for reverse transcriptase amino acid sequences responsible for priming with full or half-tRNAs (the actual case is described below). We know that different reverse transcriptases within the collection recognize one of these two primers, and this forms the basis of the predefined function tree. Sequences responsible for primer choice are not known, but by identifying windows of amino acid sequences that group the reverse transcriptases in a

manner that matches the predefined function tree, we can identify candidate regions responsible for functional diversity.

To implement our algorithm, we wrote a software package called *Split Tester* (Fig 2). As an input, *Split Tester* uses amino acid sequence alignments of related proteins generated using standard methods (e.g. ClustalX) (Jeanmougin et al., 1998). A function tree is predefined, based on the different known functions carried out by proteins within the group. The program uses the neighbor-joining method for tree construction, and the user can select one of several amino acid substitution matrices. The program then begins generating trees for different windows of the aligned sequences. The procedure is iterative and starts with very small windows (i.e. one amino acid), which slide along the length of the alignment. Window size gradually increases until it equals the full length of the aligned proteins. The tree generated for each window is compared to the predefined function tree, and if it matches, this window is marked on a plot. For each matching window, an assessment is made regarding the strength of the phylogenetic signal. Because each window generates multiple trees, the strength of the signal is based on the number of trees that support the function tree. The degree of confidence is represented by different colors in the plot: red means that all the best trees derived from this window split the taxa according to the function tree; green indicates that half of the trees match the function tree; blue indicates that 25% of the best trees match the predefined function tree. A color gradient is used to represent degrees of confidence in between the above intervals. The partitioning of the sequences is limited to two groups, but

by starting with crude partitions and progressively refining pairs of groups, one could analyze groups with more than two functional types.

**Statistical method to identify residues important for functional diversity.** As a complementary approach to the tree based method, we use a statistical approach to identify residues that may be important for functional diversity. Amino acid properties for each position in the alignment (columns) were evaluated for shared biochemical properties. The Chi-square test was performed on each amino acid position, using the null hypothesis that the amino acid properties at the same position in the two functional groups have no significant difference. If the confidence was lower than 5% ($P<0.05$), the amino acid property at this site was considered significantly different between the two groups:

$$\chi^2 = (\Sigma_{n=1..m} T_n)^2 / ((\Sigma_{n=1..m} G1_n \times \Sigma_{n=1..m} G2_n) \times (\Sigma_{n=1..m}(G1_n/T_n) - (\Sigma_{n=1..m} G1_n)^2/\Sigma_{n=1..m} T_n)$$

m is the total number of classes for the 20 amino acids, which is 9 (see Materials and Methods).

$1 <= n <= m$. $T_n$ is the number of species that have the amino acids in class n at this position. $G1_n$ and $G2_n$ are the number of species in functional group1 and functional group2, respectively, that have the amino acid in class n at this position.

**Test case 1: primer utilization by retroelement reverse transcriptases.** We applied our *Split Tester* software and statistical methods to understand functional diversity among retroelement proteins. As described above, one case we considered was primer utilization by retroelement reverse transcriptases. Reverse transcriptases of retroviruses (*Retroviridae*), the Metaviruses (*Metaviridae*) as well as members of the genus Pseudovirus

(*Pseudoviridae*) use the 3'acceptor stem of the host tRNA as a primer for DNA synthesis. This region pairs with the retroelement RNA template, and DNA synthesis extends from the 3'-OH of the tRNA. Members of the Hemiviruses (*Pseudoviridae*) use a half-tRNA primer, and cDNA synthesis initiates from nucleotide 40 of the tRNA, which resides within the anticodon stem-loop and constitutes the 5'-end of the half tRNA molecule. It is likely that the primer template complexes for the two groups of elements have different conformations or properties, and that reverse transcriptases from different groups of retroelements have evolved to recognize these distinct primer/template complexes. By searching for domains responsible for differences in primer utilization, we hoped to define candidate regions related to primer selection or binding that could then be tested by molecular genetics or structural studies. This analysis may also give us clues as to what properties of the primer binding domains of reverse transcriptase contribute to functional specialization.

We focused our analysis on members of the *Pseudoviridae*, which include both half-tRNA priming elements (namely the Hemiviruses: Ty5-6p, *Osser*, 1731 and *copia)* and elements that use full tRNAs (namely the Pseudoviruses: Ty1, *Opie*, Tnt1 and SIRE-1) (Fourcade-Peronnet et al., 1988; Lindauer et al., 1993; McCurrach et al., 1990; Rothnie et al., 1991; Voytas and Boeke, 1992). Two regions were identified by *Split Tester* when the mutation distance matrix was used to compute the cost of amino acid substitution and gaps were considered as potentially informative characters. These regions correspond to positions 144 to 245 and 269 to 314 in the aligned amino acid sequence (Fig 3A). The hydrophobicity matrix also identified these regions, and in addition, identified sequences at the N-terminus

(amino acids 1-29). Sequences corresponding to these two regions were identified in the HIV RT amino acid sequence (Xiong and Eickbush, 1990) and mapped onto the crystal structure (Huang et al., 1998) (Fig 3B). HIV RT p66 can be viewed as a 'right hand' structure. The residues 147-245 in the alignment correspond to the "palm" of the protein that encompasses the polymerase active site, which is responsible for adding nucleotides to the 3' end of the primer. Residues 269 - 314 correspond to two α-helices in the 'thumb' region of HIV RT, which contact with the primer/template directly as determined by cross-linking experiments (Jacobo-Molina et al., 1993; Peletskaya et al., 2001). The region (246-268) between the two identified domains (147-245, 269-314) form β-sheets, called the primer grip, which cross-links to the HIV primer, $tRNA^{lys}$ (Arnold et al., 1995; Setlik et al., 1994; Wohrl et al., 1995). The primer grip is conserved between the two groups, suggesting that it serves the same conserved function. This implies that the regions identified by *Split Tester* relate to primer binding, and this is consistent with our prediction that reverse transcriptase recognizes different primer-template complexes.

Our statistical method was used to characterize fifteen reverse transcriptases – seven from the Hemiviruses and eight from the Pseudoviruses. The reason we chose more sequences for the statistical analysis was to meet the requirement of the Chi-square test. Amino acids were divided into 8 clusters based on their chemical properties. Three residues, 246, 290 and 292 in the alignments between the full and half-tRNA priming elements were identified as significantly different in amino acid properties. All three residues are located in the regions identified by *Split Tester*. The statistical method, therefore, supports the

conclusion that *Split Tester* is capable of identifying regions of amino acid sequence that are different between the two functional groups of proteins.

**Test case 2: cDNA 3'-end processing by retroelement integrases.** Integrase carries out two reactions during the insertion of retroelement cDNA into the host genome: 3'-end processing is a reaction wherein a few nucleotides are cleaved from the 3'-end of cDNA; this occurs prior the second reaction, namely strand transfer, in which processed cDNA is inserted into the DNA target. Some retrotransposon integrases do not carry out 3'-end processing and insert unprocessed, blunt-ended cDNA (Feuerbach et al., 1997). The central catalytic or DD35E domain, a conserved feature of every integrase, is responsible for 3'-end processing. The C-terminus of integrase is also required (Coffin et al., 1997) and is joined to the catalytic domain by a linker region. Integrase acts as a dimer or multimer (Chen et al., 2000) (Fig 4B). Based on the structure of HIV integrase, the catalytic domain residues includes D64, D116 and E152. A region of positive charge begins at the catalytic domain of monomer A, includes K159, K186, R187, K188 and extends to K211 K215 and K218 in the α-helix linker region of monomer B. This positively charged strip likely binds DNA to integrase. The docking of DNA to this platform is thought to allow K159 to interact with the adenine at the terminus of HIV cDNA. Furthermore, residues R263 and K264 in the C-terminus of HIV integrase cross-link viral DNA (Chen et al., 2000). As indicated by the structure of SIV and RSV integrase, the spatial arrangement between the core domain and C-terminus are different (Yang et al., 2000). However, positively charged amino acids are found around the active site and periodically every 3-5 amino acids in the linker region in most

retroelement integrases, including SIV and RSV. This suggests that the charged pocket exists for DNA binding.

We used *Split Tester* to identify regions of integrase responsible for cDNA 3'-end processing (Fig 4A). Retroelements were sorted into functional classes based on whether or not they carry out 3'-end processing. It is possible to determine whether or not a retroelement uses 3'-end processing by the presence of extra bases between the tRNA primer and the 5' long terminal repeat. After cDNA synthesis, these extra bases result in 3' extensions to the cDNA, which are removed by integrase. Using this criterion, the integrases of Ty1, 1731, *Osser* and *Hopscotch* (members of the *Pseudoviridae*), were designated as not carrying out 3'-end processing. Other members of the *Pseudoviridae*, namely Endovir1-1, ToRTL1, Ta13 and Tto1, process two bases from the 3'-end of cDNA (Feuerbach et al., 1997). *Split Tester* was applied to aligned integrase amino acid sequences for these elements using the mutation distance matrix. Three regions were identified that distinguish the 3'-end processing elements when windows grow larger than 10 amino acids (210-228, 232-254, 239 - 261). The hydrophobicity matrix delineated a broader region for this data set (145-299), which originates from two shorter regions (199-240 and 238-280) and encompasses roughly the same region identified by the mutation distance matrix (residue 200-280).

We carried out a similar analysis with integrases from the *Metaviridae* (Fig 4A). Ty3 is known to engage in 3'-end processing. Based on the organization of the primer binding site with respect to the 5' LTR, *RIRE*, Athila4-3 and *Retrosor* also likely process their cDNA and mdg1, 412, Cer1 and Tf1 likely do not (Feuerbach et al., 1997). Using the *Split Tester*

software, a common region of the alignment was identified with the mutation distance (243-315) and hydrophobicity matrices (251-315). The slight difference in position of the signal between the *Pseudoviridae* and *Metaviridae* is because of differences in sequence alignments between the two groups. Because the *Metaviridae* are closely related to the retroviruses, we mapped the signal onto the HIV-1 integrase crystal structure (Fig 4C). The signal corresponds to amino acids 165-243 in HIV integrase, which spans the catalytic core domain (residues 52-210) and the C-terminal DNA binding domain (residues 220-288).

Statistical analysis was performed on thirty-three *Pseudoviridae* integrases. Fifteen carry out 3'-end processing and eighteen do not. Four amino acids in the aligned sequences were found to have distinct properties between the two groups. Two of the four are in the region identified by *Spilt Tester*: residue 252 and 255 correspond to residue 216 and 219 in HIV integrase (Fig 4C). Two other amino acids, 164 and 167 are located in the core domain and correspond to residue 128 and 132 in HIV integrase. Because most *Metaviridae* engage in 3'-end processing, we did not have enough integrase sequences from the non-processing group to carry out statistical analysis.

The results from our *Split Tester* and statistical analyses are consistent with what is known about the biological roles of the integrase domains. As described above, the linker region in conjunction with the core domain and C-terminus likely provide a platform for the binding of cDNA. In considering the dimeric structure of integrase, the two amino acids, 128 and 132, in the core domain from one monomer and the other two amino acids, 216 and 219, in the linker region from the second monomer form a grove that might bind the end of the

cDNA. The conformation of this grove, therefore, could affect 3'-end cDNA processing. Based on our results, we predict that conformation of the linker region affects access of the catalytic domain to the 3'-end of cDNA. The residues identified by our analysis may be important for cDNA access to the active site and thereby may define the functional difference between these two classes of integrases.

**Test case 3: The two- and three-repeat Myb protein family**. Because retroelement proteins are highly variable, we evaluated *Split Tester* on a more conserved set of protein sequences. We chose the Myb proteins, which are DNA binding, transcriptional regulators found in diverse organisms and are involved in diverse cellular functions (Thompson and Ramsay, 1995). Myb proteins are characterized by the Myb domain, which usually consists of two (R2R3) or three direct repeats (R1R2R3). Each repeat is about 50 amino acids long and forms three α-helices. In plants and fungi, both R2R3 and R1R2R3 Myb proteins are present, but in animals, only the R1R2R3 Myb class has been identified to date. The details of Myb protein function remain unclear, as do the functional differences between the two and three repeat proteins. One possible difference concerns DNA binding specificity; the plant R2R3 Myb proteins recognize the consensus CCT/AACC and the animal R1R2R3 Myb proteins recognize C/TAACGG (Grotewold and Peterson, 1994; Williams and Grotewold, 1997).

We used 78 Myb protein sequences: 53 3-repeat proteins and 20 2-repeat proteins (Jiang, C. 2002). Viral Myb proteins only have the R2R3 repeats, but they are more similar to animal three repeat proteins and recognize the same consensus sequence; we therefore

grouped them with the animal sequences. In contrast to the retroelement proteins, the MyB protein sequences are very similar. Between the two subtypes, the sequence identity is 70% in the R2 and R3 domains. Within each subtype, the amino acid sequence identity can reach 90% within the Myb domain. *Split Tester* found 11 residues that differentiate the two and three repeat Myb proteins (Fig 5A). All of these sites were marked on the NMR structure of the mouse c-Myb DNA binding domain (Fig 5B) (Ogata et al., 1994). Most sites face the major groove of DNA, implying that the difference between two types of Myb domains is related to DNA binding activity. Furthermore, most sites are located in the third α-helix of a repeat unit. This is consistent with experimental results indicating that the third α-helix of each repeat may play a role in DNA recognition (Rabinowicz et al., 1999). Among the 11 sites identified, three have been experimentally related to DNA binding: an insertion at Leu126 or a substitution at Ala180 compromise DNA interactions in the three-repeat proteins; in the two-repeat proteins, a mutations in Leu55 (which corresponds to Glu132 in the three-repeat proteins) similarly affects interactions with DNA (Williams and Grotewold, 1997). However, single amino acid substitutions that make one subtype more like another are not sufficient to switch binding specificity. It is very possible that multiple site changes are required to produce a change in DNA specificity. Further testing is required to evaluate the relevance of other sites in DNA binding.

# DISCUSSION

Here we describe a phylogenetic tree-based method to identify amino acids or amino acid sequence domains responsible for protein functional specificity. The method uses primary amino acid sequences of a protein family, which can range considerably in their degree of divergence. We tested our method on highly heterogeneous retroelement reverse transcriptases and integrase sequences to identify domains responsible for primer choice and 3'-end processing, respectively. We also tested the conserved DNA binding domain of eukaryotic Myb proteins. In all cases, the protein domains predicted by our method appear to be biological relevant; they are consistent with experimental data suggesting that these domains carry out a specific activity. Predicting domains of functional specificity should facilitate and guide wet-lab experimentation. Also, our method should be useful for drug design. For example, to target one protein of a superfamily without affecting the function of its paralogues, it is common to design drugs that recognize its unique features. The *Split Tester* software can help identify these unique domains.

**Evolution and functional divergence.** From the point of view of evolution, different proteins with the same function could share the same ancestors and their relationships could be described in a variety of ways: plesiomorphy (primate state), synapomorphy (shared derived state), or homoplasy (derived independently). Phylogenetic trees derived from full-length protein sequences can distinguish between these different relationships (Fig 6). New functions are derived, and if species with the same derived function are in the same clade, they are synapomorphious. If species with the same derived

function are randomly distributed in different clades of the phylogenetic tree, they are homoplasious. For plesiomorphy and synapomorphy, proteins sharing the same function might share a domain(s) with similar sequences responsible for that function, because the function was derived from the same ancestor and functional constraint kept the sequence conserved. In homoplasy, the new function evolved independently by functional convergence. However, in some cases, proteins may evolve amino acids or sequence domains with the same features because of the requirements of the function. For example, Therefore, if one functional group is homoplasy, our hypothesis that proteins carrying out similar functions will have conserved amino acid sequences will hold at least for the other function group, which is plesiomorphious. The conserved sequences in one of the functional groups will split different function types in the phylogenetic tree from this domain. The domain identified by this software corresponds to the function that shares the same origin in that functional group. If the derived function is a synapomorphy, the hypothesis will hold for that function subtype too. In both cases, the detailed tree topology for the functional domain should be studied to find out to which subtype of function it corresponds.

In the phylogenetic tree of full-length RT, the retroelements from Hemivirus and Pseudovirus are not split into two clades. It is very possible that the tRNA priming function of Pseudovirus is plesiomorphious, because it is shared by retrovirus and most of retrotransposons, and Hmiviruses are homoplasious. Two regions were found by our bioinformatics tool for priming divergence in reverse transcriptase. Studied in detail, the phylogenetic relationships of these regions reveal that the splits were not derived only from

the conservation of full tRNA priming members (*Pseudoviruses*) in these regions. Actually, sequence conservation defines two functional specificities. In the structures of the HIV-1 RT/template/primer complex, these two regions contact primers directly and this is consistent with the previous experimental result that identified regions surrounding the 'primer grip', which contact with the primer 3'-end. These results together suggest that the 'primer grip' is conserved among all the RTs for its common function in tRNA 3'-end contacts, and that the surrounding regions are specific for primer recognition.

To explain why the domain conservation also apply to of the half-tRNA priming function. Two reasons are proposed here: domains with the same chemical function were evolved independently in different retrotransposons as mentioned above; or it is possible that the evolutionary relationship from the full-length sequences can not reflect the evolutionary relationship of every functional domain in these proteins. Some functional domains have their own constraints. They might derive from the same origin before host speciation. After host speciation, the sequences outside this functional domain evolved under different constraints. Some retroelement proteins with different functions probably share the same evolutionary pressure, e.g. they are in the same host cell after speciation. Therefore, the primary region of the proteins reflects the evolution after this functional divergence. This is especially possible for the cases like priming divergence in RT and 3'-end processing divergence in IN, because the functions we studied are not dominant in these proteins. The evolutionary signals from these domains are easy to be overwhelmed by other functional

domains and noise. Therefore, homoplasy might not reflect the true relationship of functional domain divergence.

**Sequence divergence in the dataset.** The degree of sequence divergence in the input data impacts the results obtained with *Split Tester*. The retroelement proteins we analyzed were all highly divergent, and in fact, it was because of this sequence diversity that we could not find other software suitable for identifying functional domains. We wanted a method to distinguish functional classes of retroelement proteins based on subtle sequence differences. The level of pairwise sequence similarity in our RT dataset ranged from 20% to 60%. However, the RT sequences shared non-random patterns of sequence conservation, as any two sequences had BLASTP E-values much greater than the cutoff value of 10 ($e^{-10}$). In general, we found that the sequence alignment is critical in evaluating divergent datasets. Relevant signals can be obscured in alignments of divergent sequences. For example, an exceptionally divergent sequence can obscure a relevant signal, and this effect is more likely to happen in large datasets of diverse sequences. When the diverse sequences are applied to *Split Tester*, the user may have to experiment with alignment parameters and the choice of sequences in order to identify a clean alignment for input. In the case of reverse transcriptase and integrase, we evaluated our alignments based on whether known conserved domains of the proteins were aligned in all members of the dataset. The alignment is not as critical for the complementary statistical method. The statistical method evaluates each column in the amino acid sequence alignment and calculates significance for whether or not the amino acid separates the data into two functional classes. If the alignment is imperfect, the significance

will be lowered, but the user can still evaluate the likelihood that the given residue plays a functional role. Because of the nature of the Chi-square test, the statistical method is more accurate with larger datasets.

For sequences with a high degree of sequence similarity, there is no significant difference between large and small datasets. For conserved sequences, the phylogenetically informative windows appear when the window size is small (e.g. one amino acid). Between signal windows, the residues are typically identical between both functional classes. As the windows grow, these additional sequences do not obscure the signal arising from the informative sites. In general, the small windows usually identify the real informative sites. This contrasts with divergent datasets in which the informative signal typically appears within a larger window over multiple residues (e.g. greater than 10 amino acids).

**Validation.** Correct sequence alignments are a critical pre-requirement for users of this method. Suggestions for appropriate selection of data sets and evaluation of the alignments are described above. We used clustalX in default conditions to align all three datasets. Our alignment for the reverse transcripase dataset was confirmed by the fact that it is corresponded to the alignment published by Xiong & Eickbush (Xiong and Eickbush, 1990). The alignment of integrase had all known domains in integrase perfectly aligned. Because the Myb domains are so conserved, the correct alignment of this domain is obvious. We validated the correctness of our coded neighbor-joining function. The neighbor-joining function from Phylip (http://evolution.genetics.washington.edu/phylip.html) and ClustalW packages (Thompson et al., 1994) were used as a subroutine in our software to compute the

tree from each sequence window. The results were the same when Dayhoff PAM matrices (http://www.cmbi.kun.nl/bioinf/tools/pam.shtml) were used.

**The effect of optional functions in the software.** There are several substitution matrices available to compute the distance between protein sequences. The mutation distance matrix and hydrophobic distance matrix (Levitt, 1976) (http://www.sb.fsu.edu/imb/facilities/software/msi/insight970/homology/970TOC.doc.html) are the matrices used in this paper. Other choices are also available, i.e. PAM10-500 (http://www.cmbi.kun.nl/bioinf/tools/pam.shtml), BLOSUM 30-100 (Henikoff and Henikoff, 1992). For the three data sets tested in this paper, there is no significant difference in the results derived from different matrices. The same regions of functional divergence are identified by different matrices, only the details of window position and length are slightly changed.

Some researchers would like to ignore positions with gaps in aligned sequences, if they know this gap is not meaningful for the comparison. The software provides a choice for users to ignore gaps. However, the user should be aware that sites with a deletion only in one function type might relate to functional specificity.

Since we use a sliding window, only consecutive sites are examined. While this makes the problem tractable, the scattered sites in the primary sequence related to the same function will also be picked up in the small windows. The consecutive sites enforce the signal because they are dominant in the window. But the scattered sites will not enforce each other because they don't have enough sites to dominate the windows they reside in, and the signal is soon

lost when the noisy sites are added in the window. Therefore, the signal from small windows should not be considered as noise all the time. They are also good candidates for determinants of functional diversity, particularly if they are close to other identified regions in the 3D structure. However, the sparse sites have less chance to align against each other in groups of proteins as do consecutive sites, because of the weak signal. This is a common challenge in data mining methods that require an alignment as input.

We have described one ranking method that is based on the percentage of trees derived from a particular window that contain the sought split. This is not the only method that we tried, but it is the method enabled in the public version of the software. Other methods that we tried included bootstrapping and comparison of the branch lengths. Both of these methods used a non-exhaustive neighbor-joining implementation, i.e. only one tree was extracted from a window even though multiple trees may be implied (this is fairly standard for phylogenetic tools). If this single tree contained the sought split, then the ranking method was applied. The bootstrap method performed randomized sampling of the sites in the window. The tree for this re-sampled window was then extracted and tested whether the sought split still exists. This re-sampling occurs a few hundred times, and the original window is ranked depending on the percentage of the re-sampled windows that imply a tree with the sought split. This method has been found not reliable with small window lengths. The branch length method (actually a class of a number of different methods) was an attempt to build rules for ranking the tree based on branch lengths within partitions, between

partitions, and by separating partitions. All these methods yielded overall similar results as the bootstrap.

**Comparison to other methods.** There are several methods that have been developed to find sequences specific to subgroups of a protein family. Casari et al. used a vector to represent the sequence information in space (Casari et al., 1995). By analyzing the vectors' projection on various dimensions, they could find the residues important for conserved function in the whole group and for specific subgroups through their location in the vector space. Lichtarge et al. developed a phylogenetic trace method, which combines the evolutionary history, sequence alignment and structural information of proteins to find the binding surface. The adjustment of the phylogenetic tree resolution helped to find the residues conserved for all groups and for specific subgroups (Lichtarge et al., 1996). Gu developed a site-specific profile based on the Hidden Markov model to identify the residues responsible for functional divergence after gene duplication (Gu, 1999). Sridhar compared the relative entropy for a subtype with the rest of their family at each position in the alignment. The positions with significantly high relative entropy correlate with those residues responsible for functional specificity (Hannenhalli and Russell, 2000).

The existing methods have relatively strict requirements on data similarity compared to the tree-based method described in this paper. Some of the methods are good at finding residues from the alignment of conserved protein sequences, at least conserved subtypes (Gu, 1999; Hannenhalli and Russell, 2000). Others can tolerate modest degrees of sequence divergence (Casari et al., 1995; Lichtarge et al., 1996). However, our tree-based method,

because it takes advantage of the phylogenetic tree construction methods, can manipulate both conserved and divergent sequences. The informative sites will become the primary signal in certain windows, and the background noise will be reduced because the windows are narrow.

Some methods require that the clustering in the phylogenetic tree reflects functional diversity (Casari et al., 1995; Lichtarge et al., 1996; Sjolander, 1998). Ideally, the evolutionary history reflected from the phylogenetic tree corresponds to the same inherent functional hierarchy. However, if the evolutionary process is long enough after functional divergence or the mutation rate is very high (as is the case for retroelement proteins), the sequences could be too divergent and beyond the limit that a reliable phylogenetic tree can be constructed (Hannenhalli and Russell, 2000). Also, there often exists more than one functional domain in a protein. The tree can not reflect the evolutionary relationship of every functional domain, which may have different functional constraints and evolutionary rates. The method we describe in this proposal can separate the evolutionary relationships among proteins and the functional constraint. The former is usually reflected from the phylogenetic tree based on the full protein sequences, while the latter is based on the tree constructed from fragments of sequences. All that is needed is a predefined functional tree based on known functional diversity. The detailed phylogenetic relationships of proteins in each subtype can be ignored. We feel therefore that this method will be an important addition to tools that seek protein function.

## MATERIALS AND METHODS

**Sequence sources.** Ty1 (Boeke et al., 1988), 1731 (Fourcade-Peronnet et al., 1988), *copia* (Mount and Rubin, 1985), Tnt1 (Grandbastien et al., 1989), Ty5-6p (Zou et al., 1996), *Osser* (Lindauer et al., 1993), Opie-2 (SanMiguel et al., 1996), SIRE-1 (Laten et al., 1998). Ty3 (Hansen et al., 1988), RIRE2 (Ohtsubo et al., 1999), Athila4-3 (Chao, access ID in gene bank AC007534), *Retrosor* (Llaca,V, aceess ID in gene bank AF061282), 412 (Yuki et al., 1986), mdg1 (Avedisov et al., 1990), Cer1 (Britten, 1995), Tf1 (Weaver et al., 1993), Endovir1-1 (Kaneko et al., 1999), TORTL1 (Daraselia et al., 1996), Tal1-3 (Voytas and Ausubel, 1988), Tto1 (Hirochika et al., 1996), *Hopscotch* (White et al., 1994).

All the sequences were aligned in ClustalX by using the default parameters (Thompson et al., 1997). The Ty1 RT sequences were manually adjusted to align against other sequences to make it consistent with Xiong & Eickbush's domain alignment (Xiong and Eickbush, 1990).

**Detail of implementation.** To guide our search for regions of proteins responsible for functional divergence we developed *Split Tester*. This software tool, when supplied with an alignment of the protein sequences in question and a user defined partitioning of these sequences based on functional knowledge, will search for regions of the alignment that may be responsible for functional divergence.

*Split Tester's* searching is performed by extracting the phylogenetic signal from every possible portion of the alignment. Examination of all portions is done by using a sliding window of increasing size. Starting with a window length of one and testing all sites, moving

to a window length of two and testing all neighbors of sites, and so forth until at the end a single window is examined which includes all the sites. With each window an exhaustive neighbor-joining search is performed. Since the windows are often rather small there are generally multiple trees implied by the sequences and the neighbor-joining method. Trees that are identical under rotation of branches are discarded. The remaining unique trees are examined to see if any internal node perfectly splits the sequences into the partition the user selected. Windows that generate trees supporting the sought partitioning are ranked based on the strength of that signal. If a window contains only trees that contain the split then it receives a high ranking. If a window contains a small percentage of trees that contain the split then it receives a low rank. The ranking can be visualized by different colors (100%-red, 75%-yellow, 50%-green, 25%-blue). The color is agradient to represent the rankings between the listed percentage above.

*Split Tester* is available as precompiled binary in a distribution package for Microsoft Windows from the following URL: http://www.public.iastate.edu/~voytas/SplitTester.html. Both a zip file of the installation files and a self-extracting installer are available. The only difference between these two files is how the program and support files are installed. The final installation will be the same regardless of the method used. Documentation as well as example files are included in the distribution packages.

**Statistical method: how to group amino acids in the Chi-square test:** The standard Chi-square test was implemented in C++. The amino acid grouping is based on the hydrophobicity matrix: Group1: R, K (positive charge), Group2: D, E (negative charge),

Group3: S, N, Q, G, Group4: T, H, A, Group5: C, Group6: M, P, V, L, I, Group7: F, Y,

Group8: W, Group9: -.

## REFERENCES

Arnold, E., Ding, J., Hughes, S. H., and Hostomsky, Z. (1995). Structures of DNA and RNA polymerases and their interactions with nucleic acid substrates, *Curr. Opin. Struct. Biol. 5*, 27-38.

Avedisov, S. N., Cherkasova, V. A., and Il'in, I. V. (1990). [Features of the structural organization of the MDG1 retrotransposon of Drosophila, revealed during its sequencing], *Genetika 26*, 1905-14.

Boeke, J. D., Eichinger, D., Castrillon, D., and Fink, G. R. (1988). The Saccharomyces cerevisiae genome contains functional and nonfunctional copies of transposon Ty1, *Mol. Cell. Biol. 8*, 1432-42.

Britten, R. J. (1995). Active gypsy/Ty3 retrotransposons or retroviruses in Caenorhabditis elegans, *Proc. Natl. Acad. Sci. U. S. A. 92*, 599-601.

Casari, G., Sander, C., and Valencia, A. (1995). A method to predict functional residues in proteins, *Nat. Struct. Biol. 2*, 171-8.

Chapman, K. B., Bystrom, A. S., and Boeke, J. D. (1992). Initiator methionine tRNA is essential for Ty1 transposition in yeast, *Proc. Natl. Acad. Sci. U. S. A. 89*, 3236-40.

Chen, J. C., Krucinski, J., Miercke, L. J., Finer-Moore, J. S., Tang, A. H., Leavitt, A. D., and Stroud, R. M. (2000). Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding, *Proc. Natl. Acad. Sci. U. S. A. 97*, 8233-8.

Coffin, J., Hughes, S. H., and Varmus, H. E., eds. (1997). Retroviruses (Cold Spring Harbor, Cold Spring Harbor Laboratory Press).

Daraselia, N. D., Tarchevskaya, S., and Narita, J. O. (1996). The promoter for tomato 3-hydroxy-3-methylglutaryl coenzyme A reductase gene 2 has unusual regulatory elements that direct high-level expression, *Plant Physiol. 112*, 727-33.

Feuerbach, F., Drouaud, J., and Lucas, H. (1997). Retrovirus-like end processing of the tobacco Tnt1 retrotransposon linear intermediates of replication, *J. Virol. 71*, 4005-15.

Fourcade-Peronnet, F., d'Auriol, L., Becker, J., Galibert, F., and Best-Belpomme, M. (1988). Primary structure and functional organization of *Drosophila* 1731 retrotransposon, *Nucleic Acids Res. 16*, 6113-25.

Grandbastien, M. A., Spielmann, A., and Caboche, M. (1989). Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics, *Nature 337*, 376-80.

Grotewold, E., and Peterson, T. (1994). Isolation and characterization of a maize gene encoding chalcone flavonone isomerase, *Mol. Gen. Genet. 242*, 1-8.

Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication, *Mol. Biol. Evol. 16*, 1664-74.

Hannenhalli, S. S., and Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments, *J. Mol. Biol. 303*, 61-76.

Hansen, L. J., Chalker, D. L., and Sandmeyer, S. B. (1988). Ty3, a yeast retrotransposon associated with tRNA genes, has homology to animal retroviruses, *Mol. Cell. Biol. 8*, 5245-56.

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci. U. S. A. 89*, 10915-9.

Hirochika, H., Otsuki, H., Yoshikawa, M., Otsuki, Y., Sugimoto, K., and Takeda, S. (1996). Autonomous transposition of the tobacco retrotransposon Ttol in rice, *Plant Cell 8*, 725-34.

Huang, H., Chopra, R., Verdine, G. L., and Harrison, S. C. (1998). Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance, *Science 282*, 1669-75.

Jacobo-Molina, A., Ding, J., Nanni, R. G., Clark, A. D., Jr., Lu, X., Tantillo, C., Williams, R. L., Kamer, G., Ferris, A. L., Clark, P., and et al. (1993). Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 A resolution shows bent DNA, *Proc. Natl. Acad. Sci. U. S. A. 90*, 6320-4.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., and Gibson, T. J. (1998). Multiple sequence alignment with Clustal X, *Trends Biochem. Sci. 23*, 403-5.

Jiang C., Gu J., Gu X., Peterson T. Ordered origin of the typical two-and three-repeat Myb genes. Submitted to *J. Mol. Evol.* 2001.

Kaneko, T., Katoh, T., Sato, S., Nakamura, Y., Asamizu, E., Kotani, H., Miyajima, N., and Tabata, S. (1999). Structural analysis of Arabidopsis thaliana chromosome 5. IX. Sequence features of the regions of 1,011,550 bp covered by seventeen P1 and TAC clones, *DNA Res. 6*, 183-95.

Ke, N., Gao, X., Keeney, J. B., Boeke, J. D., and Voytas, D. F. (1999). The yeast retrotransposon Ty5 uses the anticodon stem-loop of the initiator methionine tRNA as a primer for reverse transcription, *RNA 5*, 929-38.

Kikuchi, Y., Ando, Y., and Shiba, T. (1986). Unusual priming mechanism of RNA-directed DNA synthesis in copia retrovirus-like particles of *Drosophila*, *Nature 323*, 824-6.

Laten, H. M., Majumdar, A., and Gaucher, E. A. (1998). SIRE-1, a copia/Ty1-like retroelement from soybean, encodes a retroviral envelope-like protein, *Proc. Natl. Acad. Sci. U. S. A. 95*, 6897-902.

Leis, J., Aiyar, A., and Cobrinik, D. (1993). Regulation of initiation of reverse transcription of retroviruses. In Reverse transcriptase, S. Goff, and A. Skalka, eds. (Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory), pp. 33-47.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding, *J. Mol. Biol. 104*, 59-107.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families, *J. Mol. Biol. 257*, 342-58.

Lindauer, A., Fraser, D., Bruderlein, M., and Schmitt, R. (1993). Reverse transcriptase families and a *copia*-like retrotransposon, *Osser*, in the green alga *Volvox carteri*, *FEBS Lett. 319*, 261-6.

McCurrach, K. J., Rothnie, H. M., Hardman, N., and Glover, L. A. (1990). Identification of a second retrotransposon-related element in the genome of *Physarum polycephalum*, *Curr. Genet. 17*, 403-8.

Mount, S. M., and Rubin, G. M. (1985). Complete nucleotide sequence of the Drosophila transposable element copia: homology between copia and retroviral proteins, *Mol. Cell. Biol. 5*, 1630-8.

Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S., and Nishimura, Y. (1994). Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices, *Cell 79*, 639-48.

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

Peletskaya, E. N., Boyer, P. L., Kogon, A. A., Clark, P., Kroth, H., Sayer, J. M., Jerina, D. M., and Hughes, S. H. (2001). Cross-linking of the fingers subdomain of human immunodeficiency virus type 1 reverse transcriptase to template-primer, *J. Virol. 75*, 9435-45.

Rabinowicz, P. D., Braun, E. L., Wolfe, A. D., Bowen, B., and Grotewold, E. (1999). Maize R2R3 Myb genes: Sequence analysis reveals amplification in the higher plants, *Genetics 153*, 427-44.

Rothnie, H. M., McCurrach, K. J., Glover, L. A., and Hardman, N. (1991). Retrotransposon-like nature of Tp1 elements: implications for the organisation of highly repetitive, hypermethylated DNA in the genome of *Physarum polycephalum*, *Nucleic Acids Res. 19*, 279-86.

SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z., and Bennetzen, J. L. (1996). Nested retrotransposons in the intergenic regions of the maize genome [see comments], *Science 274*, 765-8.

Setlik, R. F., Meyer, D. J., Shibata, M., Roskwitalski, R., Ornstein, R. L., and Rein, R. (1994). A full-coordinate model of the polymerase domain of HIV-1 reverse transcriptase and its interaction with a nucleic acid substrate, *J. Biomol. Struct. Dyn. 12*, 037-60.

Sjolander, K. (1998). Phylogenetic inference in protein superfamilies: analysis of SH2 domains, *Proc. Int. Conf. Intell. Syst. Mol. Biol. 6*, 165-74.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res. 25*, 4876-82.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res. 22*, 4673-80.

Thompson, M. A., and Ramsay, R. G. (1995). Myb: an old oncoprotein with new roles, *Bioessays 17*, 341-50.

Voytas, D. F., and Ausubel, F. M. (1988). A copia-like transposable element family in Arabidopsis thaliana, *Nature 336*, 242-4.

Voytas, D. F., and Boeke, J. D. (1992). Yeast retrotransposon revealed, *Nature 358*, 717.

Weaver, D. C., Shpakovski, G. V., Caputo, E., Levin, H. L., and Boeke, J. D. (1993). Sequence analysis of closely related retrotransposon families from fission yeast, *Gene 131*, 135-9.

White, S. E., Habera, L. F., and Wessler, S. R. (1994). Retrotransposons in the flanking regions of normal plant genes: a role for copia-like elements in the evolution of gene structure and expression, *Proc. Natl. Acad. Sci .U. S. A. 91*, 11792-6.

Williams, C. E., and Grotewold, E. (1997). Differences between plant and animal Myb domains are fundamental for DNA binding activity, and chimeric Myb domains have novel DNA binding specificities, *J. Biol. Chem. 272*, 563-71.

Wohrl, B. M., Tantillo, C., Arnold, E., and Le Grice, S. F. (1995). An expanded model of replicating human immunodeficiency virus reverse transcriptase, *Biochemistry 34*, 5343-56.

Xiong, Y., and Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences, *Embo. J. 9*, 3353-62.

Yang, Z. N., Mueser, T. C., Bushman, F. D., and Hyde, C. C. (2000). Crystal structure of an active two-domain derivative of Rous sarcoma virus integrase, *J. Mol. Biol. 296*, 535-48.

Yuki, S., Inouye, S., Ishimaru, S., and Saigo, K. (1986). Nucleotide sequence characterization of a Drosophila retrotransposon, 412, *Eur. J. Biochem. 158*, 403-10.

Zou, S., Ke, N., Kim, J. M., and Voytas, D. F. (1996). The *Saccharomyces* retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci, *Genes. Dev. 10*, 634-45.

## FIGURE LEGENDS

**Fig 1. The algorithm for our tree-based method to identify protein functional domains.** The aligned sequences are used as an input file. Phylogenetic trees from different windows of the alignment are generated by the neighbor-joining method. For each window, the program determines whether the tree matches the predefined function tree. If the phylogenetic tree can split the protein families according to their functional specificity, the sequences supporting this split are candidates for carrying out that function. The program is iterative and starts with very small windows along the alignment (i.e. 1 amino acid); window size gradually increases until it equals the full length of the protein.

**Fig 2. A snapshot of the *Split-Tester* software.** The input file is a file of aligned protein sequences. Three different matrices (identity distance matrix, hydrophobicity distance matrix and mutation distance matrix) can be selected to compute the phylogenetic relationship of the input data. The positions of those sequences that split two functional groups are plotted in the top window. The right lower two windows show all the trees that have the same best score from your selected signal region. If all of these best trees can split taxa according to the functional tree, the signal in the top window will be red. The green signal has lower confidence because some of the best trees from this region do not support the split. The lower left window shows the actual sequences that give the predefined phylogenetic relationship.

**Fig 3. The results of the RT dataset.** (A) Two domains identified by *Split Tester* in the RT sequence alignment are boxed. All the windows supporting the functional split are shown as colored lines in the plot. From the two axes, the position and the length of certain windows can be identified. The sequences in that window (line) are also shown by *Split Tester*. The red lines represent the sequences in this window that support the functional tree with high confidence. Green and blue colored lines have decreasing confidence levels, respectively. (B) The X-ray structure of HIV reverse transcriptase-template-primer complex (1RTD.pdb). The RT protein is represented by the yellow strand. The two green regions are domains identified by *Split Tester*. Three red balls are position 228, 278 and 280 and were identified by the statistical method.

**Fig 4. The functional divergence results of integrase.** (A)The upper plot displays the results from integrases of Ty1/*copia* retrotransposons. The lower plot shows the results from the integrases of the Ty3/*gypsy* group. The slightly different position of the identified region is due to differences in the alignment of the two groups of integrases. They correspond to the same region when sequences from Ty1/*copia* and Ty3/*gypsy* group are aligned together. (B) Three red balls on each monomer are the active site residues D64, D116 and D152, respectively. All the blues sites compose the positive strip, which start from one monomer (K159, K186, K187, K188) and extends to another monomer (K211, K215, K218, K263, K264). (C) The region identified by *Split Tester* related to 3'-end cDNA processing is

labeled as green. The two orange sites (Ala128 and Trp132) and two blue sites (N216 and K219) on each monomer were identified by the statistical method.

**Fig 5. The amino acids specifying functional divergence between 2- and 3-repeats of the Myb protein family.** (A) The protein sequence alignment of Myb protein family. The sequences in the upper box are the proteins containing 2-repeat Myb domains (R2R3) and the proteins with 3-repeat Myb domains are in the lower box. Only sequences in R2 and R3 domains are shown in the figure. The amino acid sites identified by *Split Tester* are indicated by arrows below the alignment, and the sites proved experimentally are pointed to by red arrows. (B) The NMR structure of mouse Myb-c (R1R2R3) and DNA complex. The blue and green strands are the DNA helix. The gray strand is the R2 and R3 domain of the mouse Myb protein (1MSF.pdb). The red balls are the 10 sites with different residues from R2R3 Myb. Leu126 is a deletion in R1R2R3 Myb, not shown in this protein.

**Fig 6. Different functional evolution from the same ancestor.** The black and blank circles represent the derived new state and the ancestral state, respectively. This figure is from Molecular Evolution-A phylogenetic approach (Page & Holmes Chap 2.)

Fig. 1

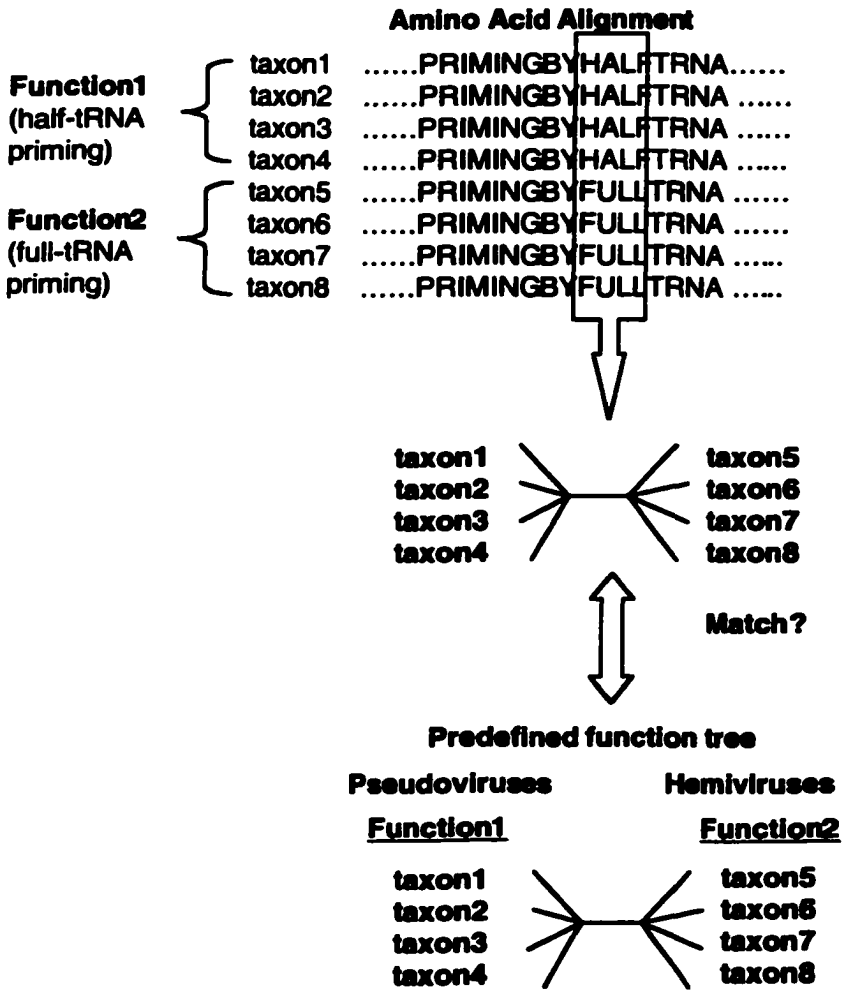
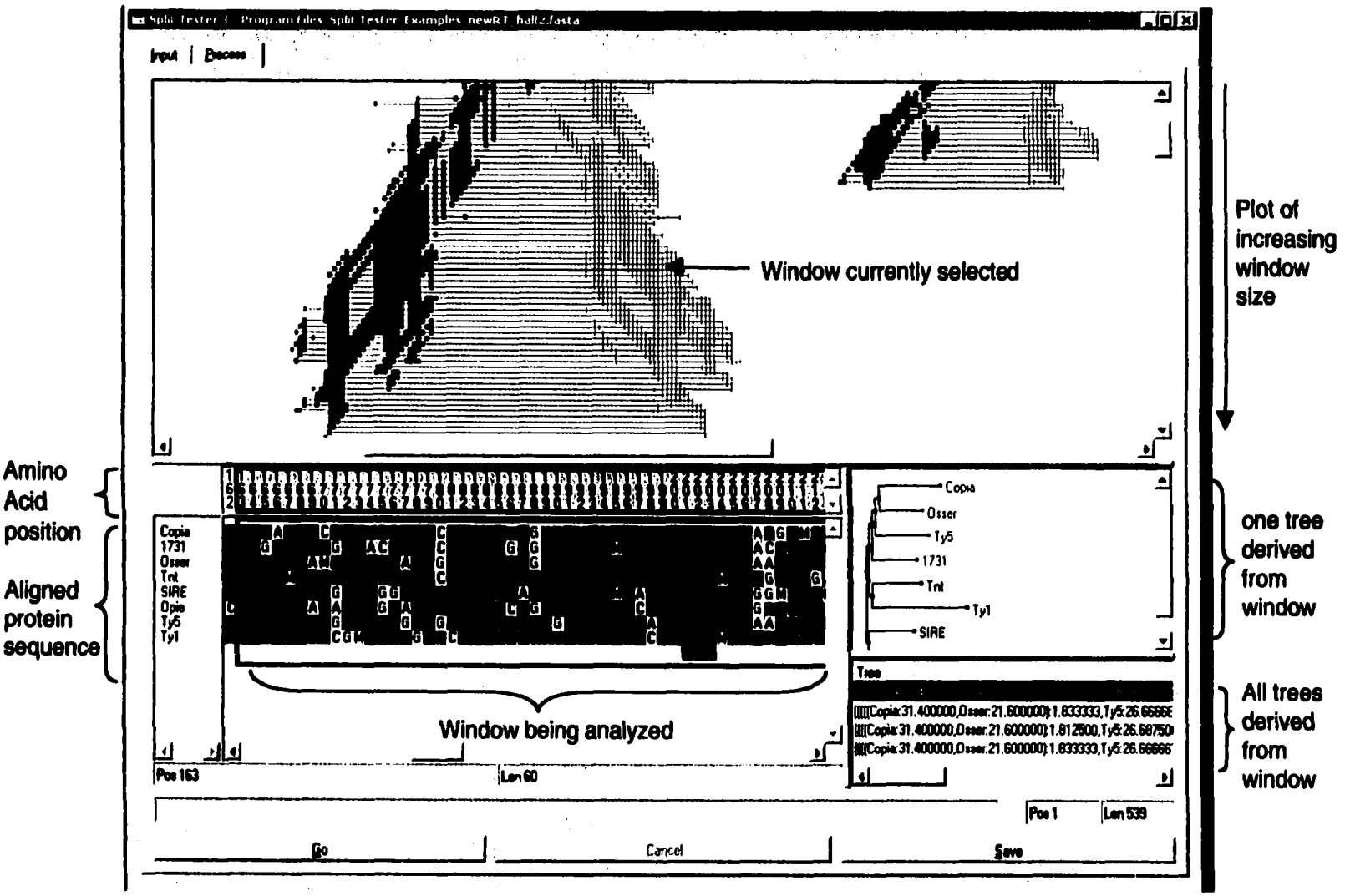

Amino Acid Alignment
Function1
(half-tRNA
priming)
taxon1 .....PRIMINGBYHALFTRNA.....
taxon2 .....PRIMINGBYHALFTRNA .....
taxon3 .....PRIMINGBYHALFTRNA .....
taxon4 .....PRIMINGBYHALFTRNA .....
Function2
(full-tRNA
priming)
taxon5 .....PRIMINGBYFULLTRNA .....
taxon6 .....PRIMINGBYFULLTRNA .....
taxon7 .....PRIMINGBYFULLTRNA .....
taxon8 .....PRIMINGBYFULLTRNA .....
taxon1
taxon2
taxon3
taxon4
taxon5
taxon6
taxon7
taxon8
Match?
Predefined function tree
Pseudoviruses
Function1
Hemiviruses
Function2
taxon1
taxon2
taxon3
taxon4
taxon5
taxon6
taxon7
taxon8

Fig. 2


Input
Process
Window currently selected
Plot of increasing window size
Amino Acid position
Aligned protein sequence
Copia
1731
Osser
Tnt
SIRE
Opie
Ty5
Ty1
Window being analyzed
one tree derived from window
All trees derived from window
Tree
Pos 163
Len 60
Pos 1
Len 539
Go
Cancel
Save

106

Fig. 3A

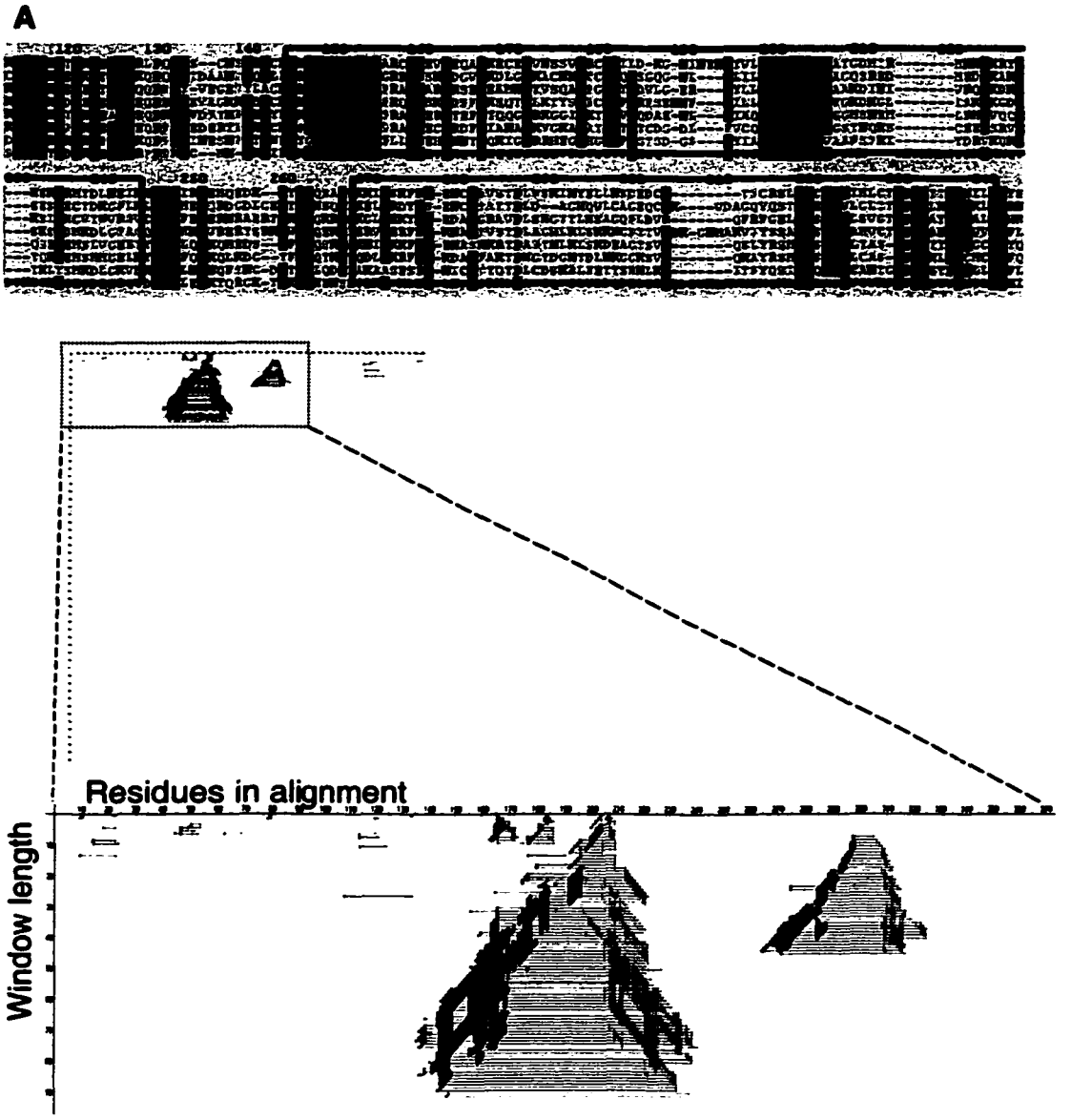


Fig. 3B

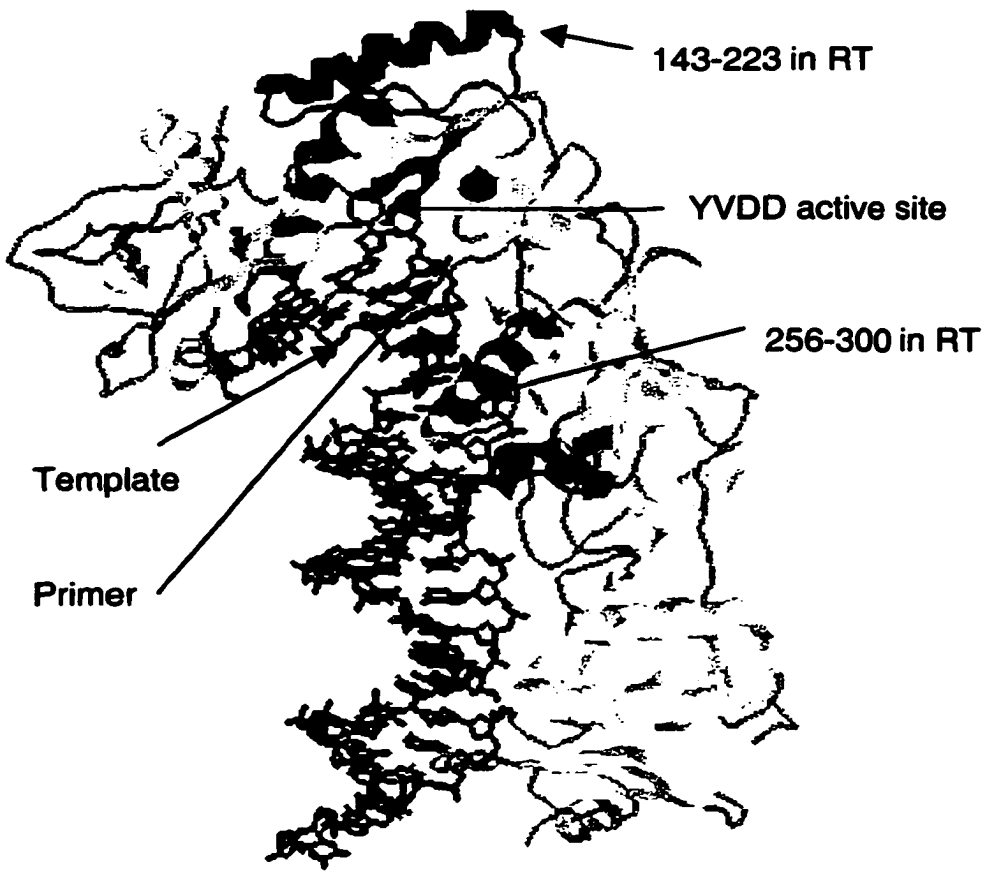
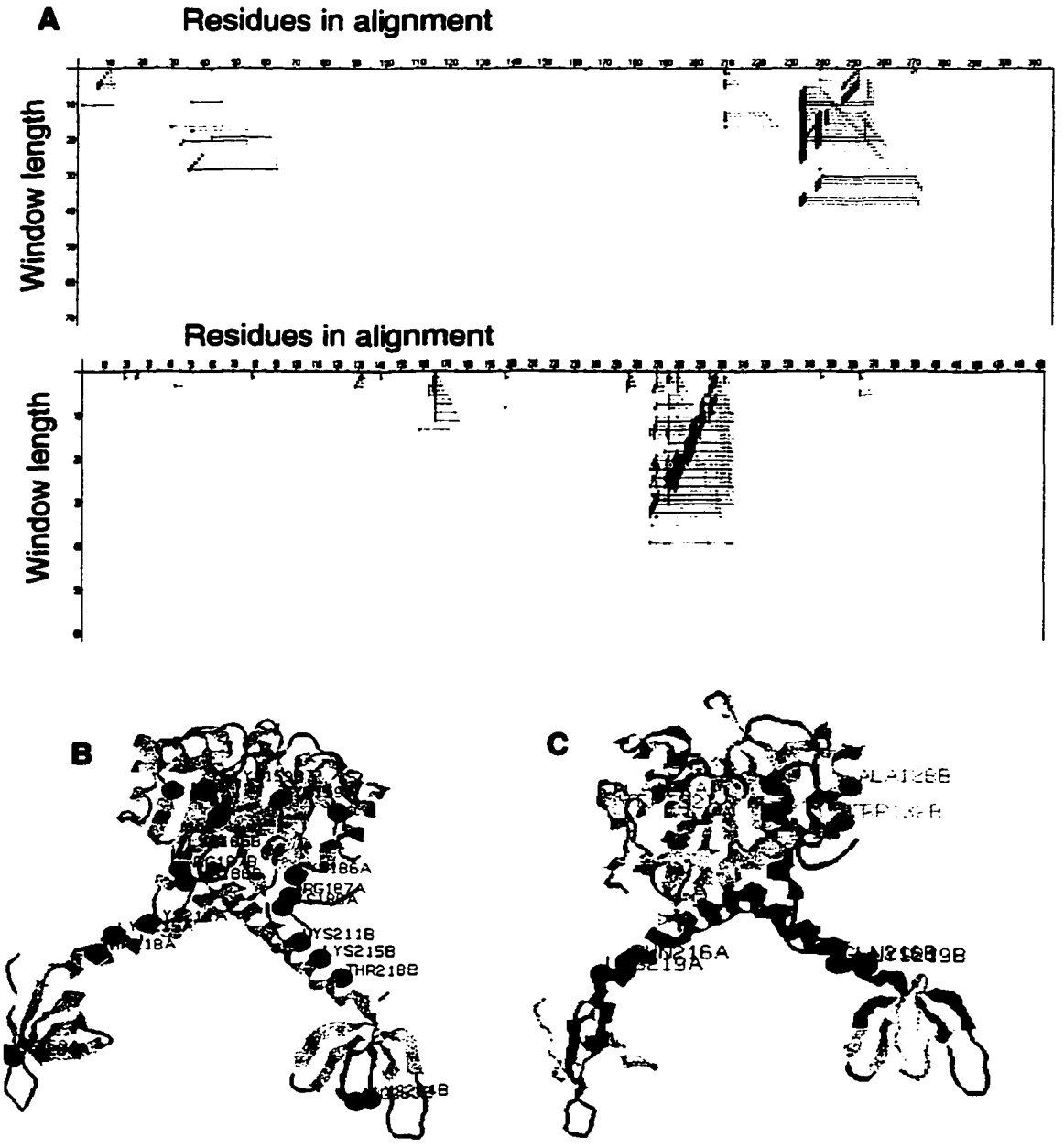


Fig. 4


A
Residues in alignment
Window length
Residues in alignment
Window length
B
C

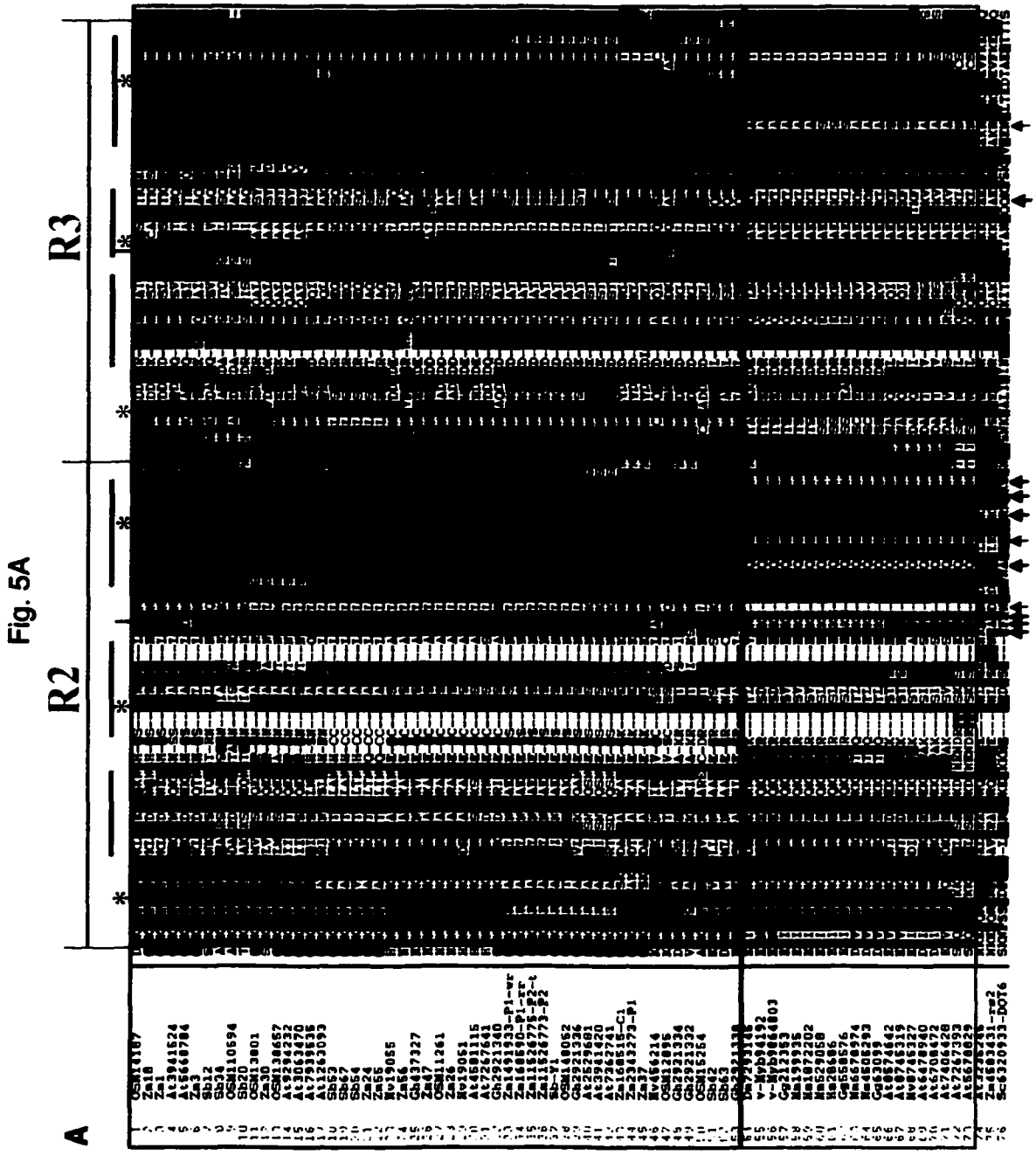



Fig. 5A

Fig. 5B

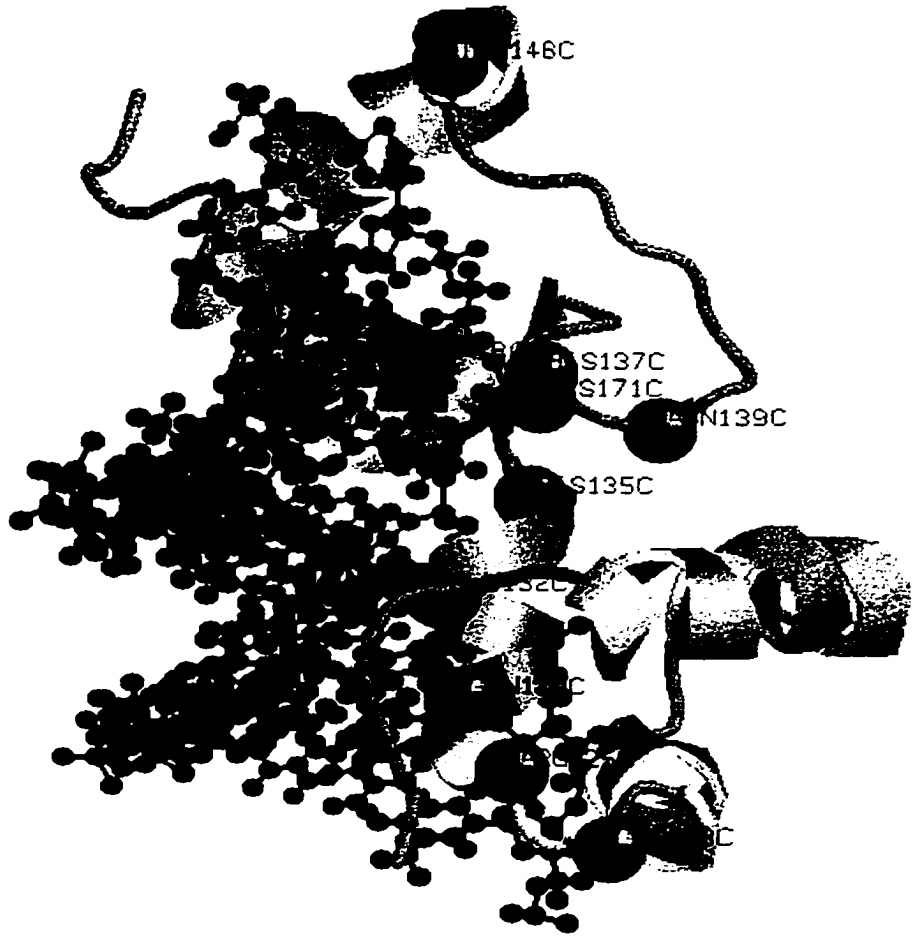
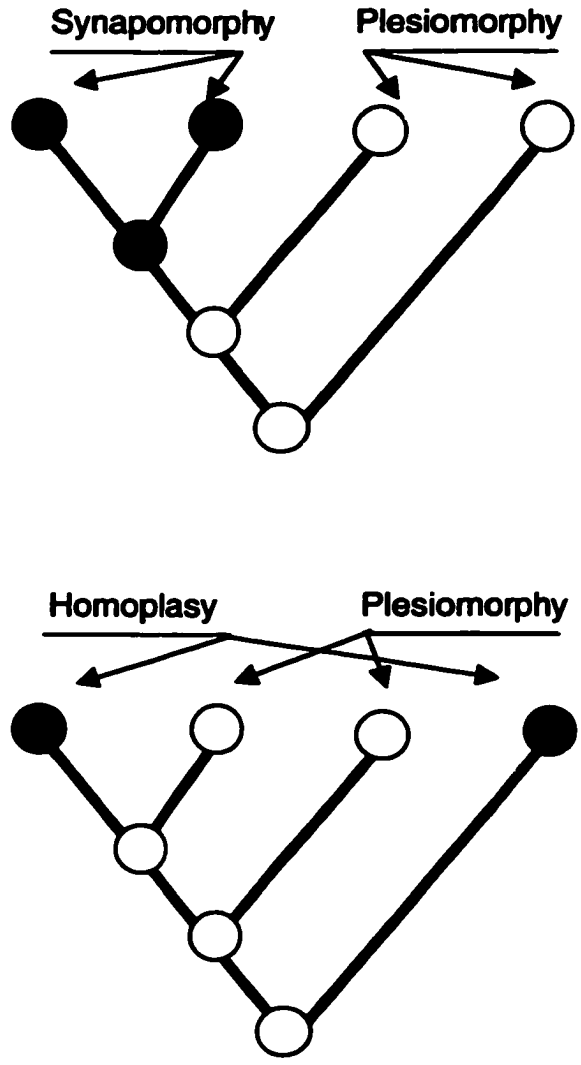


Fig. 6



# CHAPTER IV. GENERAL CONCLUSION

The yeast retrotransposon Ty5 evolved some interesting features that are used during the replication process. These features are different from the majority of retrotransposons, although they are not unique to Ty5. My dissertation research is focused on revealing the mechanism of Ty5 replication; specifically half-tRNA initiated priming, reverse transcription and 3'-end processing of cDNA. Both molecular genetics and bioinformatics approaches are used. This study will help us to understand the diverse replication mechanisms in retroelements.

**Hydrogen bonding in NCp zinc finger plays a role in Ty5 reverse transcription**

Fifteen Ty5 sequences were discovered in both *S. cerevisiae* and *S. paradoxus* strains. Only Ty5-6p actively transposes when expressed under the *GAL* promoter. Other Ty5 elements are either solo-LTRs or truncated fragments with accumulated stop codons. Compared to Ty1, the recovered Ty5-6p has 1000-fold lower transposition frequency. By screening the Ty5-6p mutant library, two mutations Y68C and D252N were identified that independently promote Ty5 transposition 5-6 fold. The combined double mutant increases transcription ~36 fold. Neither of the two mutations affected Ty5 protein quantity, solubility, processing or Ty5 targeting specificity. The mutations did not affect the recombination and integration efficiency. By doing competitive PCR of Ty5 cDNA, both mutations showed higher cDNA levels.

The D252N mutation is located in the conserved CCHC zinc finger region. Bioinformatics approaches were used to reveal that the zinc finger in Ty5 NCp has some differences compared to other members of the Ty1/*copia* family. Two significant differences were found. The more obvious difference is the Gly deletion in wild type Ty5 ($CX_2CX_3HX_4C$) relative to the consensus sequence ($CX_2CX_3GHX_4C$) in Ty1/*copia* elements. The D252N mutation is right before His in the Ty5 NCp zinc finger. The second difference is that there is low hydrogen bonding potential in the zinc finger of Ty5 relative to the Ty1/*copia* group, and to the mutant Ty5. I proved experimentally that the D252N substitution increased the hydrogen bonding potential of the zinc finger and made it more like the consensus zinc finger sequence in the Ty1/*copia* family. Other mutations, D252K and D252R, which increased the hydrogen bonding potential also increased Ty5 transposition to the same level. This suggests that NCp has a role in Ty5 reverse transcription similar to its role in other retroelements. In addition, hydrogen bonding likely plays an important role in Ty5 zinc finger function. The principle of nucleic acid binding appears to be the same for retroelement zinc fingers found in other cellular enzymes.

By comparing the conserved zinc finger sequence, I demonstrated that mutations accumulated in the Ty5 genome during evolution lower its ability to function actively. By studying the reasons for this low activity of Ty5, we can identify the sites important for retroelement replication and explore the evolutionary relationship between host and retroelements.

**Half-tRNA primed reverse transcription**

A fragment of cellular initiator methionine tRNA (IMT) is used as a primer for Ty5 reverse transcription. There are 14 bases (27-40) in the anticodon stem-loop of IMT that are complementary to primer binding sites in Ty5 RNA. The requirement of the free 3'-OH at the end of primer suggests that IMT is specifically cleaved between 40 and 41 residues. Mutations in the IMT stem-loop can abolish Ty5 transposition. There is a more severe effect on Ty5 activity, when the mutation is closer to the 3'end of the primer. The reason might be that mutations at the 3'end of the primer make the annealing between primer and template not complete at the initiation point, and therefore extension can not happen. Mutations in the Ty5 PBS region also abolish Ty5 transposition. However, when mutations in Ty5 PBS restore the complementary between PBS and IMT, Ty5 transposition is also restored. Therefore, the base pairing between the tRNA and the Ty5 PBS is essential for transposition, but not the sequence itself. Mutations in the acceptor stem did not change Ty5 transposition. A mutant tRNA library was constructed and the chance that there is at least one single mutation in this library for every nucleotide in the IMT D-arm and TψC-arm was over 95%. No single mutations affecting Ty5 transposition were recovered. Therefore, the bases outside of the IMT anticodon stem-loop do not have major effects on Ty5 transposition. Although extensive binding between IMT and element RNA was discovered in Ty1 and Ty3, the complementarity between Ty5 and IMT is restricted to PBS/anticodon interaction only.

**Bioinformatics approach to study functional diversity in reverse transcriptases and other protein families**

Reverse transcriptase (RT) is the most conserved protein of retroelements and therefore is used to study the evolutionary relationships of retroelements. In the phylogenetic tree of RT, the Hemiviruses and Pseudoviruses are not split into two clades. Instead, the Hemivirus members are homoplasious. Half-tRNA priming likely evolved independently, and full tRNA priming is the feature derived from their common ancestors. We developed a computational approach to identify candidate domains in protein sequences responsible for functional divergence in a group of homologous proteins. The hypothesis is that proteins sharing the same subfunction might share a domain(s) with similar sequences responsible for that function. More often than not, synapomorphious functional divergence can be identified, because the function was derived from the same ancestor and the functional constraint kept the sequence conserved. Homoplasious functions evolved independently through functional convergence. However, in some cases, they evolved the same sites because of the requirement of the function.

We checked phylogenetic trees from windows of different length at different positions along the protein sequence alignments. If the phylogenetic tree from certain windows matched the functional split, the sequences in this window should be conserved for either or both subgroups. Therefore, the sequence in that window might be responsible for the conserved subtype function(s). A program was developed to accomplish this idea.

Two regions were found by this bioinformatics tool for priming divergence in reverse transcriptase. Study of the detailed phylogenetic relationship of these regions revealed that the split was not derived solely from the conservation of full-tRNA priming members in these regions. Actually, sequence conservation defines two functional specificities. In the HIV RT/template/primer complex structure, these two regions contact primers directly, and this is consistent with previous experimental results that identified regions just surrounding the 'primer grip' as contacting the primer 3'-end. These results together suggest that the 'primer grip' is conserved in all RTs for its common function in contacting the tRNA 3'-end, and that the surrounding regions are specific for primer/template complex recognition.

Functional diversity in cDNA 3'-end processing in both families was also examined by this bioinformatics tool. Although the integrase sequences are divergent between these two families, the characteristic domains (e.g. CCHH, DDE) are still conserved. I tested Ty3/*gypsy* and Ty1/*copia* datasets independently and the same region was found in both families that is likely responsible for 3'-end processing. According to the crystal structure of HIV integrase, which is very similar to Ty3/*gypsy* family, the identified region was found in the linker region between the core domain and the 3'domain that binds with cDNA. Because the DDE domain, which executes 3'-end processing, exists in all the integrases even if they do not have the 3'-end processing, the diversity of 3'-end processing is not because of the presence or absence of this catalytic domain. The model we proposed is that the processing could occur in the integrase if the conformation in the linker region allows the 3'-end of cDNA to access to DDE catalytic domain.

To test the application range of this computational approach, we used this bioinformatics tool on sequences for cellular Myb protein family. Myb proteins are much conserved in the Myb domain. Eleven amino acid sites were identified and three of them were confirmed experimentally as being important for DNA binding and specificity.

Based on above results, the bioinformatics approach I developed provides a biological meaningful predictive tool. More than that, the results from this approach give more insight about the mechanism of the function we are studying. Because there is no strict requirement of divergent degree of input data set, it will have a broad application in understanding function of both cellular and retroelement protein families.

# APPENDIX I. THE YEAST RETROTRANSPOSON Ty5 USES THE ANTICODON STEM-LOOP OF THE INITIATOR METHIONINE tRNA AS A PRIMER FOR REVERSE TRANSCRIPTION

**A paper published in *RNA*[1]**

**Ning Ke[2], Xiang Gao[3], Jill B. Keeney[4], Jef D. Boeke[5] & Daniel F. Voytas[6]**

## ABSTRACT

Retrotransposons and retroviruses replicate by reverse transcription of an mRNA intermediate. Most retroelements initiate reverse transcription from a host-encoded tRNA primer. DNA synthesis typically extends from the 3'-OH of the acceptor stem, which is complementary to sequences on the retroelement mRNA (the primer binding site, PBS). However, for some retrotransposons, including the yeast Ty5 elements, sequences in the anticodon stem-loop of the initiator methionine tRNA (IMT) are complementary to the PBS.

---

[1] Reprinted with permission of *RNA* (1999), 5:929-938.

[2] Primary researcher and author.

[3] Thesis author who constructed the assay system for *imt* mutants that can not support translation, tested effects of mutation in the anticodon stem-loop on Ty5 transposition and proved the stability of tRNA mutant.

[4] Associate professor, who provide strains and tRNA mutants, Department of Biology, Juniata College, Huntingdon, PA 16652

[5] Professor, who provide strains and instructions, Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205

[6] Professor and corresponding author, Department of Zoology and Genetics, Iowa State University, Ames, IA 50011.

We took advantage of the genetic tractability of the yeast system to investigate the mechanism of Ty5 priming. We found that transposition frequencies decreased at least 800 fold for mutations in the Ty5 PBS that disrupt complementarity with the IMT. Similarly, transposition was reduced at least 200 fold for IMT mutations in the anticodon stem-loop. Base pairing between the Ty5 PBS and IMT is essential for transposition, as compensatory changes that restored base pairing between the two mutant RNAs restored transposition to near wild-type levels (4 fold lower). An analysis of 12 *imt* mutants with base changes outside of the region of complementarity failed to identify other residues important for transposition. In addition, assays carried out with heterologous IMTs from *Schizosaccharomyces pombe* and *Arabidopsis thaliana* indicated that residues outside of the anticodon stem-loop have at most a five-fold effect on transposition. Our genetic system should make it possible to further define the components required for priming and to understand the mechanism by which Ty5's novel primer is generated.

## INTRODUCTION

Retrotransposons are a class of genetic elements that replicate through an mRNA intermediate. They are structurally and functionally analogous to retroviruses and therefore provide an important model for understanding retroviral replication (Boeke and Sandmeyer, 1991; Brown and Varmus, 1989). During reverse transcription, both retrotransposons and retroviruses typically use a host-encoded tRNA as a primer for first strand cDNA synthesis. Immediately downstream of the 5' long terminal repeat (LTR) is a region called the minus

strand primer binding site (PBS), which base pairs with sequences at the 3' end of the primer tRNA. Reverse transcriptase initiates cDNA synthesis from the 3' OH of the tRNA and extends the cDNA to the 5' end of the element mRNA. Reverse transcription continues through a series of two strand transfers, ultimately resulting in a double stranded linear cDNA. This cDNA is integrated into the genome by the element-encoded integrase.

Whereas most retroviruses and retrotransposons, including HIV-1 and the *Saccharomyces cerevisiae* Ty1 elements, use the 3' end of a tRNA as a primer for reverse transcription (Chapman et al., 1992; Leis et al., 1993), other novel priming mechanisms have been described. For example, Hepatitis B virus uses a protein primer, and DNA synthesis initiates from an OH provided by a tyrosine residue (Tavis and Ganem, 1993; Wang and Seeger, 1992). The Tf1 retrotransposon of *Schizosaccharomyces pombe* uses a self-priming mechanism (Levin, 1995; Levin, 1996). The 5' end of the template mRNA folds back and anneals to the PBS; the template mRNA is then cleaved to release an RNA fragment that serves as a primer. For the *Drosophila melanogaster copia* element, the PBS pairs to the anticodon stem-loop of the *D. melanogaster* initiator methionine tRNA. The tRNA is cleaved by some unknown mechanism and a half tRNA molecule is used to initiate DNA synthesis (Kikuchi et al., 1986).

*Saccharomyces cerevisiae* and its Ty retrotransposons have become an important model system for understanding mechanisms by which tRNAs prime reverse transcription. In *S. cerevisiae*, both the retrotransposons and their primer tRNA can be genetically manipulated. Efficient transposition assays have been developed for Ty1 and Ty3, and both

elements have PBSs complementary to the 3' end of an initiator methionine tRNA (IMT) (Boeke et al., 1985; Chapman et al., 1992; Hansen et al., 1988; Keeney et al., 1995). A yeast strain has been developed in which all four copies of the *IMT* genes are disrupted, and cells survive by carrying a functional *IMT* gene on a plasmid (Bystrom and Fink, 1989). This makes it possible to test *imt* mutants for their effect on transposition. Using this system, Ty1 and Ty3 have been shown to use the 3' acceptor stem of the IMT as a primer for reverse transcription (Chapman et al., 1992; Keeney et al., 1995). Additional IMT residues important in priming have also been identified in the D and TψC arms (Friant et al., 1998; Gabus et al., 1998; Keeney et al., 1995). These residues base pair with other regions of the retroelement mRNA and thereby help to stabilize primer/template interactions (Friant et al., 1998; Gabus et al., 1998).

The putative PBS of the yeast Ty5 retrotransposon is complementary to the anticodon stem-loop of the *S. cerevisiae* IMT (Fig. 1) (Voytas and Boeke, 1992). Strikingly, the region of complementarity is identical to that observed between the *copia* element mRNA and the *D. melanogaster* IMT (Kikuchi et al., 1986). Other retrotransposons from a variety of organisms, including 1731 from *D. melanogaster*, *Osser* from *Volvox carteri*, and Tp1 and Tp2 from the slime mold *Physarum polycephalum* have PBSs that are complementary to the same region of the IMT anticodon stem-loop (Fourcade-Peronnet et al., 1988; Lindauer et al., 1993; McCurrach et al., 1990; Rothnie et al., 1991). This suggests that the mechanism of half-tRNA priming is highly conserved among these fungal, protist and animal

retrotransposons. In this study we have exploited the *S. cerevisiae* system to better understand how half-tRNAs are used by Ty5 to prime reverse transcription.

## RESULTS

The sequence of the putative Ty5 PBS is complementary to fourteen bases within the anticodon stem-loop of the *S. cerevisiae* initiator methionine tRNA (IMT) (Fig. 1). To test the significance of this complementarity in Ty5 transposition, we adopted an assay previously used to evaluate the role of the IMT in priming Ty1 reverse transcription (Keeney et al., 1995). In this assay, Ty5 and the *IMT* gene are carried on plasmids to facilitate the testing of a variety of Ty5 and *imt* mutants for their effect on transposition. The plasmid-borne Ty5 element can be transcriptionally induced by growth on galactose, and it also carries a *HIS3* marker gene (*his3*AI) that is non-functional due to the presence of an artificial intron. The *HIS3* marker becomes functional after intron loss through Ty5 transcription, intron splicing and reverse transcription. Integration of Ty5 cDNA into the genome confers a $His^+$ phenotype (Zou et al., 1996).

The assay system uses a yeast strain with disruptions in all four copies of the initiator methionine tRNA genes (Bystrom and Fink, 1989)(Fig. 2). Translation is supported by a wild-type *IMT* gene on a *URA3*-based plasmid. Mutant *imt* genes (on *LEU2*-based plasmids) are introduced into this strain by plasmid shuffling: the strain is transformed with a plasmid carrying a mutant *imt* gene, and the plasmid with the wild type *IMT* is lost by growing the cells on medium containing 5-fluoroorotic acid (5-FOA), which selects against the

*URA3* marker (Boeke et al., 1987). This plasmid shuffling strategy requires, however, that the mutant *imt* can support translation. The effect of given *imt* mutants on Ty5 transposition is measured by introducing the Ty5-containing plasmid and carrying out our standard transposition assay. To ensure that only transposition and not cDNA recombination is evaluated, we disrupted the *RAD52* gene, which is responsible for high frequency homologous recombination of Ty5 cDNA (Ke and Voytas, 1999).

**Mutations in the putative Ty5 PBS abolish transposition.** To test whether the putative PBS is important for transposition, two PBS mutations were generated by site-directed mutagenesis that disrupt complementarity with the IMT (*pbs-1*, five bases altered; *pbs-2*, four bases altered) (Fig. 3). Since the PBS lies within the Ty5 coding region, bases were changed that did not affect the derived amino acid sequence in these mutants. Quantitative transposition assays were carried out in a strain with a wild type, plasmid-borne *IMT* gene. For both PBS mutants, transposition frequencies were at least ~800 fold lower compared to a wild type Ty5 element. The most severe defect was observed for *pbs-1*, in which the 5'-most base of the putative PBS was no longer complementary to the IMT. These experiments indicate that the putative Ty5 PBS is important for Ty5 transposition.

**Transposition is abolished by mutations in the IMT anticodon stem-loop that disrupt complementarity with the Ty5 PBS.** Because the Ty5 PBS is complementary to fourteen bases in the IMT anticodon stem-loop (positions 27-40), we next assayed transposition in strains with mutant *imt* genes that reduce this complementarity. Our initial experiments focused on *imt* mutants with base changes near the 3'-end of the putative Ty5

primer, including two mutants that had previously been characterized (*imt4-U31,U39* and *imt4-A29,U41,U31,U39*) (Keeney et al., 1995; von Pawel-Rammingen et al., 1992)(Fig. 4). Positions 29, 31 and 39 are predicted to base pair with the Ty5 PBS, and because position 39 is the penultimate 3'-base of the putative primer, we predicted that disruption of base pairing at this position would impair DNA synthesis. Since the above two *imt* mutants have a C to T transition at position 39 and could potentially form a G-U pair with the Ty5 PBS (Fig. 4), two additional *imt* mutants (*imt4-U31,A39* and *imt4-C31,G39*) were made that disrupted this G-U pair. All four mutant tRNA genes supported translation, as cells carrying only these *imt* genes grew at rates similar to cells with a wild type *IMT* (von Pawel-Rammingen et al., 1992 and data not shown).

Ty5-containing plasmids were introduced into the four strains with the mutant *imt* genes, and transposition frequencies were determined (Fig. 4). For *imt4-U31, U39*, transposition dropped three fold relative to wild type. Further destabilizing primer/PBS complementarity with a mutation at position 29 (*imt4-A29,U41,U31,U39*) caused transposition to drop 19 fold. Because both *imt4-U31,U39* and *imt4-A29,U41,U31,U39* can form a G-U base pair at position 39 that may stabilize the primer/template complex, the role of G-U pairing was directly tested by changing U39 to either A39 (*imt4-U31,A39*) or G39 (*imt4-C31,G39*). In strains carrying these tRNA genes, transposition dropped more than 200 fold and 500 fold, respectively. These data indicate that the IMT anticodon stem-loop is important for transposition, and transposition is particularly sensitive to mutations that disrupt base pairing at the region near the 3' end of the putative primer.

**Restoring complementarity between the PBS and the IMT restores Ty5 transposition.** We next tested whether the transposition defect caused by a mutant *imt* could be suppressed by restoring base pairing between the IMT and the putative PBS. We focused on mutant *imt4-C31,G39*, since it had the most severe effect on transposition. Two PBS mutants were made that were complementary to *imt4-C31,G39* at either position 39 (*pbs-4*) or at positions 39 and 31 (*pbs-3*) (Fig. 5A). Because the PBS lies within the Ty5 coding region, these changes alter the derived amino acid sequences: *pbs-4* has a Val to Leu substitution at amino acid 14, and *pbs-3* has a Val to Leu change at position 14 and a Ser to Arg change at position16. Plasmids carrying Ty5 with either wild type or mutant PBSs were tested in strains with either a wild type *IMT* or *imt4-C31,G39*. Consistent with our previous observations, mutations in either the PBS or the IMT largely abolished transposition (Fig. 5B). However, in the strains carrying *imt4-C31,G39*, transposition was restored to almost wild type levels for *pbs-4* and to some extent for *pbs-3*. The difference in the extent of restoration was probably due to the type and number of changes in the Ty5 amino acid sequence, which may affect protein function. The single Val-Leu substitution in *pbs-4* is conservative and less likely to compromise protein function compared to the two changes in *pbs-3*. The near wild-type levels of transposition conferred by *pbs-4* indicate that base pairing between the tRNA and the Ty5 PBS is essential for transposition.

**The effect of mutations in the anticodon stem-loop on Ty5 transposition.** To identify tRNA residues important for Ty5 transposition, we tested other *imt* genes with mutations throughout the anticodon stem-loop for their affect on transposition. Six of these

mutants could support translation (Table 1): *imt4-C33* and *imt4-U38* had no significant effect on transposition, and *imt4-U41*, *imt4-A29,U41*, *imt4-C29,G41* and *imt4-U29,A41* had less than a four fold effect on transposition, despite the fact that position 41 is the putative cleavage site. Because many *imt* genes with anticodon stem-loop mutations had translation defects, a modified assay was developed (Fig. 2B). This assay used two plasmid-borne *imt* genes: *imt4-C31,G39* supports translation but cannot support transposition, and this *imt* gene was cloned into a Ty5 containing plasmid; a second *imt* mutant that cannot support translation was introduced to test its effect on transposition. The mutant *imt* genes tested were previously found to be stable and could support Ty1 transposition (Keeney et al., 1995). Using our modified assay, mutant *imt* genes with translation defects were found, in general, to have a more severe effect on transposition when the altered base was close to the 3' end of the tRNA primer (Table 1).

**Mutations in regions other than the anticodon stem-loop have no effect on Ty5 transposition.** To initiate reverse transcription, the primer tRNA needs to be packaged into virus or virus-like particles, loaded onto the template mRNA, and, in the case of Ty5, it may be processed by cleavage. We next looked at mutations in other regions of the IMT to determine whether they influence Ty5 transposition by affecting steps other than primer annealing (Table 2). Most of these mutants were made previously to study features that distinguish the IMT from the elongator methionine tRNA (EMT) and to identify residues important for Ty1 transposition (Keeney et al., 1995; von Pawel-Rammingen et al., 1992). All mutants can support translation with the exception of *imt4-C60,U54*.

Mutations in the acceptor stem had little effect on Ty5 transposition. *imt4-9*, which has nine mutations in the acceptor stem and reduces Ty1 and Ty3 transposition frequency more than 100 fold (Chapman et al., 1992; Keeney et al., 1995), supports Ty5 transposition to approximately wild type levels (Table 2). Two mutations in the D arm were also tested: *imt4-+A17* has an A inserted at position 17 that enlarges the D loop by one nucleotide; *imt4-U12, A23* has a G-C to U-A base pair change in the D stem, which is found in the *EMT* gene. Neither of the D stem-loop mutations affected transposition significantly. Several residues in the T_C arm have been implicated in distinguishing the IMT from the EMT, and some are critical for priming Ty1 reverse transcription (Astrom and Bystrom, 1994; Keeney et al., 1995; von Pawel-Rammingen et al., 1992). For example, A54 and A60 are conserved among all cytoplasmic IMTs. A64 (of the U50/A64 pair) is ribosylated at the 2' position, a modification unique to IMTs from plants and fungi. This modification has been shown to be critical in preventing the IMT from being used in elongation both *in vivo* and *in vitro* (Astrom and Bystrom, 1994; Kiesewetter et al., 1990). We tested mutations at these and other positions in the T_C arm, including mutations at position 52 and 62, which we generated by changing the G-C base pair to a U-A pair (*imt4-U52,A62*). The transposition frequencies for all mutants tested ranged from 0.48 to 1.65 (compared to 1.0 for wild type). It appears, therefore, that only mutations in the anticodon stem-loop have a significant effect on Ty5 transposition.

**The effect of heterologous *IMT* genes on Ty5 transposition.** Because *A. thaliana* and *S. pombe IMT* genes can support translation in *S. cerevisiae*, we assayed their effects on

Ty5 transposition (Fig. 6) (Keeney et al., 1995). Seven and ten fold lower transposition frequencies were observed for these heterologous tRNAs, respectively (Table 3). To identify the regions responsible for the decreases, we changed their anticodon stem-loops or their acceptor stems to match the sequence of the *S. cerevisiae* IMT. The acceptor stem changes did not significantly affect transposition (Table 3). The anticodon stem-loop changes, however, restored transposition to some extent (from ten fold to five fold lower for the *S. pombe IMT*; from seven to two fold lower for the *A. thaliana IMT*). This again indicates that the anticodon stem-loop sequences are the most important determinants in Ty5 priming, and base changes elsewhere among the heterologous IMTs have at most a five fold effect on transposition.

## DISCUSSION

During retroelement replication, a tRNA is typically used to prime reverse transcription. Priming involves multiple steps: the tRNA is first packaged into virus or virus-like particles, then loaded onto the messenger RNA, and finally reverse transcriptase initiates cDNA synthesis (Voytas and Boeke, 1993). Each step may involve multiple element or host-encoded proteins. *Saccharomyces cerevisiae* provides an attractive system to dissect retroelement priming mechanisms. Transposition assays have been developed for the yeast Ty1, Ty3 and Ty5 retrotransposons (Boeke et al., 1985; Hansen et al., 1988; Zou et al., 1996). In addition, the gene that encodes the initiator methionine tRNA (IMT) can be genetically manipulated to identify residues important in priming (Bystrom and Fink, 1989;

Keeney et al., 1995). We have taken advantage of this system to investigate the mechanism by which the Ty5 element initiates cDNA synthesis.

The Ty5 PBS is complementary to fourteen bases within the anticodon stem-loop of the initiator methionine tRNA (positions 27-40) (Voytas and Boeke, 1992). Our initial experiments focused on testing the effect of mutations in either the Ty5 PBS or the IMT anticodon stem-loop that disrupt the complementarity between these two RNAs. We observed at least an 800 fold decrease in transposition frequencies for strains carrying a PBS with four mismatched bases. A range of transposition defects was observed for *imt* mutants that disrupt the G-C base pair at position 39, the penultimate 3' base in the primer. When position 39 was changed to either a G-A pair (*imt4-U31,A39*) or a G-G pair (*imt4-C31,G39*), transposition decreased at least 200 fold. However, for *imt4-U31,U39* and *imt4-A29, U41, U31, U39*, which can form a G-U base pair with the Ty5 PBS at position 39, transposition frequency was only three and nineteen fold lower, respectively (in the latter case, the base pairing was further destabilized by a C-A mismatch). A G-U pair between the Ty1 PBS and IMT was previously shown to have little effect on Ty1 transposition (Keeney et al., 1995).

Base pairing between the Ty5 PBS and the IMT anticodon sequences is essential for Ty5 transposition. Two Ty5 PBS mutants were made (*pbs-3* and *pbs-4*) that allow for base pairing with *imt4-C31,G39* at either position 39 or at both positions 31 and 39. Individual mutations in either the IMT or PBS had dramatic effects on Ty5 transposition, causing at least a 486 fold decrease. Restoring the base pairing, however, by combining the *imt* and *pbs* mutants resulted in transposition frequencies of only 3.36 and 64.4 fold lower for *pbs-4* and

*pbs-3*, respectively. Since the Ty5 PBS lies within the coding region of Ty5, the transposition defect observed for *pbs-3, imt4-C31,G39* combination is likely due to the changes in the amino acid sequences caused by the PBS mutation. Whereas *pbs-4* has a conserved Val to Leu change, *pbs-3* has this change and a Ser to Arg change.

Although we have demonstrated that Ty5 uses the IMT anticodon stem-loop as the primer for reverse transcription, the mechanism by which priming occurs remains to be determined. For example, we cannot distinguish whether priming is initiated from a 2' OH at position 40 or from a 3' OH at position 40 after a cleavage between position 40 and 41. We also do not know whether the cleavage event occurs specifically between position 40 and 41 or whether the tRNA is digested from the 3' end by an exonuclease. Distinguishing among these possibilities is the goal of ongoing investigations.

Several closely related elements from diverse organisms appear to use the same region of the initiator tRNA as primer, suggesting that half-tRNA priming is a conserved mechanism for initiating reverse transcription (Fourcade-Peronnet et al., 1988; Kikuchi et al., 1986; Lindauer et al., 1993; McCurrach et al., 1990; Rothnie et al., 1991). Among these elements, *copia* of *D. melanogaster* has been studied most extensively. *Copia* uses the identical region of the anticodon stem-loop as a primer, and sequence analysis of the initial product of reverse transcription (strong stop DNA) has indicated that a tRNA fragment is the *bone fide copia* primer (Kikuchi et al., 1986). There is *in vitro* evidence that the catalytic RNA of RNase P can cut IMTs between positions 39 and 40, as well as at other positions, suggesting that overprocessing by RNase P produces the *copia* primer (Kikuchi and Sasaki, 1992; Kikuchi et

al., 1990). Priming of the *S. pombe* Tf1 element also involves cleavage of its primer, and this cleavage requires Tf1 RNase H (Levin, 1996). An RNase H dependent scission event may also give rise to the Ty5 primer.

For those IMT mutations that do not support translation, a modified assay was developed to test their effect on transposition. For this assay, *imt4-C31,G39*, which supports translation but not transposition, was cloned into a Ty5-containing plasmid. A second tRNA that cannot support translation was introduced on a high copy plasmid and tested for its effect on transposition. Using both our original and this modified assay, we tested mutations in residues throughout the IMT. The anticodon stem-loop mutations in *imt4-U41*, *imt4-A29,U41*, *imt4-C29,G41* and *imt4-U29,A41* had at most a four fold effect on transposition. It is interesting to note that mutations at position 41 had only a slight effect when changed to all three other nucleotides, even though this position is at the putative cleavage site. Other anticodon mutations generally had a greater effect on transposition when the mutations were close to the 3' end of the primer. For example, *imt4-_A38* and *imt4-U37* affected transposition frequencies dramatically (0.02), whereas *imt4-G32* and *imt4-C35* had more modest effects (0.15 and 0.32, respectively). An exception to this trend is *imt4-U38*, which had a minimal effect on transposition (0.64), even though it is near the 3' end of the primer. *imt4-U38* may form a U-U base pair between the PBS and tRNA; U-U pairs have been observed in other RNAs (Cech et al., 1994; Gutell, 1994; Gutell et al., 1993). However, *imt4-U37* (0.02) should also be able to form a U-U pair, suggesting that if U-U pairing occurs, it may be context dependent. Alternatively, because A37 is the only modified base in the

anticodon stem-loop (Basavappa and Sigler, 1991), the U37 mutation may affect transposition indirectly (i.e. not through base pairing).

In contrast, none of the 12 mutants with mutations outside the IMT anticodon region effected transposition more than two fold. The altered bases include nine residues in the acceptor stem (positions 1, 2, 3, 6, 67, 70, 71, 72, 73), six in the T_C arm (positions 50, 52, 54, 60, 62, 64) and three in the D-arm (12, 17, 23) (Fig. 6). This suggests that only the anticodon stem-loop region is essential for transposition. The lack of importance for bases outside of the region of PBS complementarity is also supported by our experiments with heterologous tRNAs. The *S. pombe* and *A. thaliana* IMT each differ from the *S. cerevisiae* IMT at 20 positions and therefore they are useful as probes to determine globally which regions or residues of the IMT are important for priming (Fig. 6). Relative transposition frequencies for these IMTs, however, were 0.10 for the *S. pombe* and 0.14 for the *A. thaliana* IMT. Among the bases that differ are positions within the anticodon stem-loop (two for *S. pombe* and three for *A. thaliana*). When these anticodon sequences were changed to match the *S. cerevisiae* IMT, the hybrid IMTs supported transposition at relative frequencies of 0.18 for the *S. pombe* hybrid IMT and 0.60 for the *A. thaliana* hybrid IMT. Therefore, when the anticodon sequences can base pair with the Ty5 PBS, other differences in the heterologous IMT have at most a five fold effect on Ty5 transposition.

Although IMT anticodon stem-loop sequences are necessary for priming, they are not sufficient. For example, we have found that a hybrid elongator methionine tRNA that carries the IMT anticodon stem-loop cannot support Ty5 transposition (data not shown). What are

the other structural features of the IMT that are important for transposition? Ty1 and Ty3 mRNAs pair with regions other than the 3' acceptor stem sequences, and this interaction helps stabilize the primer/template complex (Friant et al., 1998; Gabus et al., 1998; Keeney et al., 1995). Multiple interactions between element mRNAs and their primers occur among other retroelements, including Tf1 and Rous sarcoma virus (Aiyar et al., 1992; Lin and Levin, 1997). Ty5 mRNA is also complementary to other regions within the IMT: nine bases within the D-arm and 8 bases within the T_C stem are complementary to sequences within the Ty5 coding region (GAG) and U3, respectively (data not shown). Some of the mutant tRNAs tested disrupt this complementarity at a single base without consequences on transposition. It may be that more extensive disruption of pairing is required before a phenotype can be observed. Interactions between the Ty5 mRNA with other regions of the IMT are currently being tested more rigorously. We hope that a comprehensive understanding of tRNA bases required for priming, coupled with biochemical assays for each step in the priming reaction, will ultimately enable us to obtain a more comprehensive understanding of half-tRNA priming mechanisms.

## MATERIALS AND METHODS

**DNA plasmids.** Many of the *imt4* mutants were previously used to identify residues important for translation and for Ty1 transposition (Keeney et al., 1995; von Pawel-Rammingen et al., 1992). These include pKC35 (*IMT4*), pIMT116 (*imt4-+A17*), pKC74 (*imt4-U38*), pIMT114 (*imt4-U31,U39*), pIMT115 (*imt4-A29,U41,U31,U39*), pKC75 (*imt4-*

*C36*), pKC76 (*imt4-A33*), pKC77 (*imt4-G32*), pKC81 (*imt4-G34*), pKC79 (*imt4-_A38* ), pKC78 (*imt4-U37*), pKC80 (*imt4-C35*), pIMT118 (*imt4-C54*), pIMT119a (*imt4-G60*), pIMT118a (*imt4-G54*), pIMT119 (*imt4-C60*), pIMT120 (*imt4-C60,U54*), pIMT121(*imt4-U60,C54*), pIMT123 (*imt4-U64,A50*), pVIT83 (*imt4-C64,G50*), pJK258 (*S. pombe* IMT with *S. cerevisiae* IMT acceptor stem), pJK244 (*Arabidopsis thaliana* IMT with *S. cerevisiae* IMT acceptor stem). Other *imt4* mutants in this study were made by a two step PCR-based mutagenesis method with wild type *IMT4* as template (Chen and Przybyla, 1994): pNK494 contains *imt4-U12,A23* (made with DVO490, 5'-CGCCGTGGCTCAGTGGAAGAGCGCAGGGC-3'); pNK344 contains *imt4-U31,A39* (made with DVO291, 5'-GGACATCAGGTTTATGAGACCTGCGCGCT-3'); pNK346 contains *imt4-C31,G39* (made with DVO280, 5'-ACATCAGGCTTATGAGGCCTGCGCG-3'); pNK493 contains *imt4-U41* (made with DVO489, 5'-CTCATAACCTTGATGTCC-3'); pNK496 contains *imt4-U41,A29* (made with DVO488, 5'-GAAGCGCGCAAGGCTCATAACCTTGATGTCC-3'); pNK547 contains *imt4-C29,G41* (made with DVO725, 5'-GGAAGCGCGCACGGCTCATAACCGTGATGTCCTCG-3'); pNK548 contains *imt4-U29,A41* (made with DVO726, 5'-GGAAGCGCGCATGGCTCATAACCATGATGTCCTCG-3'); pNK540 contains *imt4-C33* (made with DVO708, 5'-GCGCAGGGCCCATAACCCTGATG-3'); pNK495 contains *imt4-U52,A62* (made with DVO491, 5'-GATGTCCTCTGATCGAAACAGAGCGGCGC-3').

The plasmids containing heterologous *IMT* genes were constructed as follows: pDV111 carries the *S. pombe IMT* gene on a 1.3 kb *Hin*dIII fragment in YEp351. The 237 bp *IMT*-containing *Dra*I fragment of pDV111 was cloned into the *Eco*RV site of pBluescript (Stratagene) to generate pNK507. The *S. pombe IMT* was then cloned into YEp351 using *Bam*HI and *Hin*dIII to generate pNK514. For pNK515, the anticodon sequence of the *S. pombe IMT* was changed to the corresponding sequences of *S. cerevisiae IMT* by a PCR-based site-directed mutagenesis strategy using primer DVO569 (5'-GGAACTCCGCAGGGCTCATAACCCTGAGGTCCCAG-3') (Chen and Przybyla, 1994). pSZ1 contains the *A. thaliana IMT* in YEp351. A 600 bp *Hin*dIII-*Sma*I from pSZ1 was cloned into YEp351 to generate pNK518. pNK519 was generated by changing the *S. pombe* anticodon sequences to match the corresponding *S. cerevisiae IMT* sequences using DVO568 (5'-GGAAGCGTGCAGGGCTCATAACCCTGAGGTCCCAG-3').

Ty5 PBS mutants were also made by PCR-based site-directed mutagenesis (Chen and Przybyla, 1994). pLG1 contains the wild type *GAL*-Ty5*his3*AI in pRS426; it has *GAL1-10* upstream activation sequences (UASs) fused in front of Ty5 5' long terminal repeat (LTR) and a *his3*AI marker inserted between the end of the Ty5 open reading frame and the 3' LTR (Zou et al., 1996). pLG2 contains *pbs-1*, a Ty5 element with five mutations in the PBS that was constructed using primer DVO285 (5'-ACTACGTCAACAAGTAATGTCACCTGAGAGCAAT-3'). pLG3 contains *pbs-2*, a Ty5 element with four mutations in the PBS that was constructed using primer DVO286 (5'-ACGTCAACAGGTAATGTCACCTGAGAGCAAT-3'). pLG2 and pLG3 were

constructed using pLG1 as template. pNK488 contains *pbs-3*, which restores base pairing with *imt4-C31,G39* at position 39, and was constructed using primer DVO436 (5'-ACGTCAACAGCTTATGAGCCCTG-3'). pNK469 contains *pbs-4*, which restores the base pairing with *imt4-C31,G39* at both positions 31 and 39, and was constructed using primer DVO435 (5'-ACGTCAACAGCTTATGAGGCCTGAGAGCAATG-3'). pNK488 and pNK469 were constructed using as a template pNK254, which carries the *GAL*-Ty5*his3*AI on pRS416 (Ke and Voytas, 1997).

Plasmids used in the two-*IMT* assay were constructed as follows: *imt4-C31,G39* was cloned into the Ty5 containing plasmid pNK254 by first cloning a *Bam*HI and *Hin*dIII fragment from pNK346 into the corresponding sites in pBluescript; this generated pNL2. PCR-based mutagenesis was carried out (using primer DVO474, 5'-TCCCCGCGGGACGGTATCGATAAGCTT-3') to create *Sac*II restriction sites flanking *imt4-C31,G39* (Chen and Przybyla, 1994). The resultant *Sac*II fragment was cloned into pNK254 to generate pNK502.

**Transposition assays.** All strains used in this study are isogenic derivatives of JKc543 (*MAT_ ura3-52 trp1_1 leu2-3 ,112 his3_200 ade2-BglII imt1-imt4::TRP1/* YEp351-*IMT2*). To facilitate plasmid shuffling, pKC1, which carries *IMT3* on a *URA3*-based plasmid, was introduced into JKc543. The preexisting, plasmid-borne *IMT2* gene (YEp351-*IMT2*, (Keeney et al., 1995)) was lost by selecting colonies that grew on synthetic complete media lacking uracil (SC-U) but not on SC-L media; this generated YNK611. YNK611 was transformed with *Bam*HI-digested pNK437, which contains *rad52::ADE2*. It was

constructed by cloning the *ADE2*-containing *Bam*HI fragment from pJK204 into the *Bgl*II site of pSM20. This replaced the *LEU2* marker that interrupts *RAD52* gene in pSM20 with *ADE2*. $Ade^+$ colonies were picked and confirmed to be *rad52* by Southern blot analysis; this strain was designated YNK616 and was used for all subsequent transposition assays.

To determine the importance of PBS sequences in Ty5 transposition, the *IMT4*-containing plasmid, pKC35, was introduced into YNK616. The preexisting *IMT3* gene carried on pKC1 was lost by selecting on SC-L, 5-fluoroorotic acid media (SC-L/5-FOA) (Boeke et al., 1987). The Ty5-containing plasmids pLG1 (wildtype Ty5), pLG2 (*pbs-1*), and pLG3 (*pbs-2*) were introduced, and three $Ura^+$ colonies resulting from each transformation were picked and used for transposition assays. The cells were grown as patches on SC-U/glucose plates for two days before being replica-plated onto SC-U/galactose media to induce Ty5 transcription. After three days of induction at room temperature, the cells were replica-plated to SC-H media to select for transposition events. For quantitative assays, the cells from the SC-U/galactose plates were scraped and resuspended in $ddH_2O$. Serial dilutions of the resuspended cells were made and plated on either SC-U plates to determine the total cell number or on SC-H plates to determine the number of transposition events.

To test transposition in strains with IMT mutations that support translation, *imt* mutants on *LEU2*-based YEp351 were transformed into strain YNK616. The *IMT3* gene carried on pKC1 was then lost by selecting cells on SC-L/5-FOA media. Ty5 elements carried on the *URA3*-based plasmid pRS416 (pNK254, wildtype Ty5; pNK488, *pbs-3*;

pNK469, *pbs-4*) were then introduced into the strains with either the wildtype or mutant *imt* genes. Three Ura$^+$ colonies were picked and used in transposition assays as described above.

For *imt* mutants that cannot support translation, the two-*IMT* assay was used. pKC35 (*IMT4*) was shuffled into YNK616. Plasmid (pNK502), which contains *imt4-C31,G39* and *GAL*-Ty5*his3*AI ,was then introduced, and pKC35 (YEp351-*IMT4*) was lost by plasmid shuffling prior to introducing plasmids with mutant *imt* gene. Transposition assays were then conducted and transposition frequencies calculated as described above.

## ACKNOWLEDGMENTS

We thank Liang Guo and Nianzhen Li for helping making plasmid constructs and testing transposition frequencies. This work was supported by NIH grant GM51400 to D.F.V. This is Journal Paper No. J-17884 of the Iowa Agriculture and Home Economics Experiment Station, Ames, IA, Project No. 3383, and was supported by Hatch Act and State of Iowa Funds.

## REFERENCES

Aiyar A, Cobrinik D, Ge Z, Kung HJ, and Leis J 1992. Interaction between retroviral U5 RNA and the T psi C loop of the tRNA(Trp) primer is required for efficient initiation of reverse transcription. *J Virol 66*: 2464-72.

Astrom SU, and Bystrom AS 1994. Rit1, a tRNA backbone-modifying enzyme that mediates initiator and elongator tRNA discrimination. *Cell 79*: 535-46.

Basavappa R, and Sigler PB 1991. The 3 A crystal structure of yeast initiator tRNA: functional implications in initiator/elongator discrimination. *Embo J 10*: 3105-11.

Boeke JD, Garfinkel DJ, Styles CA, and Fink GR 1985. Ty elements transpose through an RNA intermediate. *Cell 40*: 491-500.

Boeke JD, and Sandmeyer SB. 1991. Yeast transposable elements. In: Broach J, Jones E, Pringle J, eds. *The molecular and cellular biology of the yeast Saccharomyces*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory. pp 193-261.

Boeke JD, Trueheart J, Natsoulis G, and Fink GR 1987. 5-Fluoroorotic acid as a selective agent in yeast molecular genetics. *Methods Enzymol 154*: 164-75.

Brown P, and Varmus H. 1989. Retroviruses. In: Berg DE, Howe MM, eds. *Mobile DNA*. Washington D. C.: American Society for Microbiology. pp 53-108.

Bystrom AS, and Fink GR 1989. A functional analysis of the repeated methionine initiator tRNA genes (*IMT*) in yeast. *Mol Gen Genet 216*: 276-86.

Cech TR, Damberger SH, and Gutell RR 1994. Representation of the secondary and tertiary structure of group I introns. *Nat Struct Biol 1*: 273-80.

Chapman KB, Bystrom AS, and Boeke JD 1992. Initiator methionine tRNA is essential for Ty1 transposition in yeast. *Proc Natl Acad Sci U S A 89*: 3236-40.

Chen B, and Przybyla AE 1994. An efficient site-directed mutagenesis method based on PCR. *Biotechniques 17*: 657-9.

Fourcade-Peronnet F, d'Auriol L, Becker J, Galibert F, and Best-Belpomme M 1988. Primary structure and functional organization of *Drosophila* 1731 retrotransposon. *Nucleic Acids Res 16*: 6113-25.

Friant S, Heyman T, Bystrom AS, Wilhelm M, and Wilhelm FX 1998. Interactions between Ty1 retrotransposon RNA and the T and D regions of the tRNA(iMet) primer are required for initiation of reverse transcription *in vivo*. *Mol Cell Biol 18*: 799-806.

Gabus C, Ficheux D, Rau M, Keith G, Sandmeyer S, and Darlix JL 1998. The yeast Ty3 retrotransposon contains a 5'-3' bipartite primer-binding site and encodes nucleocapsid protein NCp9 functionally homologous to HIV-1 NCp7. *EMBO J 17*: 4873-80.

Gutell RR 1994. Collection of small subunit (16S- and 16S-like) ribosomal RNA structures: 1994. *Nucleic Acids Res 22*: 3502-7.

Gutell RR, Gray MW, and Schnare MN 1993. A compilation of large subunit (23S and 23S-like) ribosomal RNA structures: 1993. *Nucleic Acids Res 21*: 3055-74.

Hansen LJ, Chalker DL, and Sandmeyer SB 1988. Ty3, a yeast retrotransposon associated with tRNA genes, has homology to animal retroviruses. *Mol Cell Biol 8*: 5245-56.

Ke N, and Voytas DF 1999. cDNA of the yeast retrotransposon Ty5 preferentially recombines with substrates in silent chromatin. *Mol Cell Biol in press*.

Ke N, and Voytas DF 1997. High frequency cDNA recombination of the *Saccharomyces* retrotransposon Ty5: The LTR mediates formation of tandem elements. *Genetics 147*: 545-56.

Keeney JB, Chapman KB, Lauermann V, Voytas DF, Astrom SU, von Pawel-Rammingen U, Bystrom A, and Boeke JD 1995. Multiple molecular determinants for retrotransposition in a primer tRNA. *Mol Cell Biol 15*: 217-26.

Kiesewetter S, Ott G, and Sprinzl M 1990. The role of modified purine 64 in initiator/elongator discrimination of tRNA(iMet) from yeast and wheat germ. *Nucleic Acids Res 18*: 4677-82.

Kikuchi Y, Ando Y, and Shiba T 1986. Unusual priming mechanism of RNA-directed DNA synthesis in copia retrovirus-like particles of Drosophila. *Nature 323*: 824-6.

Kikuchi Y, and Sasaki N 1992. Hyperprocessing of tRNA by the catalytic RNA of RNase P. Cleavage of a natural tRNA within the mature tRNA sequence and evidence for an altered conformation of the substrate tRNA. *J Biol Chem 267*: 11972-6.

Kikuchi Y, Sasaki N, and Ando-Yamagami Y 1990. Cleavage of tRNA within the mature tRNA sequence by the catalytic RNA of RNase P: implication for the formation of the primer tRNA fragment for reverse transcription in copia retrovirus-like particles. *Proc Natl Acad Sci U S A 87*: 8105-9.

Leis J, Aiyar A, and Cobrinik D. 1993. Regulation of initiation of reverse transcription of retroviruses. In: Goff S, Skalka A, eds. *Reverse transcriptase*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory. pp 33-47.

Levin HL 1995. A novel mechanism of self-primed reverse transcription defines a new family of retroelements. *Mol Cell Biol 15*: 3310-7.

Levin HL 1996. An unusual mechanism of self-primed reverse transcription requires the RNase H domain of reverse transcriptase to cleave an RNA duplex. *Mol Cell Biol 16*: 5645-54.

Lin JH, and Levin HL 1997. A complex structure in the mRNA of Tf1 is recognized and cleaved to generate the primer of reverse transcription. *Genes Dev 11*: 270-85.

Lindauer A, Fraser D, Bruderlein M, and Schmitt R 1993. Reverse transcriptase families and a *copia*-like retrotransposon, *Osser*, in the green alga *Volvox carteri*. *FEBS Lett 319*: 261-6.

McCurrach KJ, Rothnie HM, Hardman N, and Glover LA 1990. Identification of a second retrotransposon-related element in the genome of *Physarum polycephalum*. *Curr Genet 17*: 403-8.

Rothnie HM, McCurrach KJ, Glover LA, and Hardman N 1991. Retrotransposon-like nature of Tp1 elements: implications for the organisation of highly repetitive, hypermethylated DNA in the genome of *Physarum polycephalum*. *Nucleic Acids Res 19*: 279-86.

Tavis JE, and Ganem D 1993. Expression of functional hepatitis B virus polymerase in yeast reveals it to be the sole viral protein required for correct initiation of reverse transcription. *Proc Natl Acad Sci U S A 90*: 4107-11.

von Pawel-Rammingen U, Astrom S, and Bystrom AS 1992. Mutational analysis of conserved positions potentially important for initiator tRNA function in *Saccharomyces cerevisiae*. *Mol Cell Biol 12*: 1432-42.

Voytas DF, and Boeke JD 1992. Yeast retrotransposon revealed. *Nature 358*: 717.

Voytas DF, and Boeke JD 1993. Yeast retrotransposons and tRNAs. *Trends Genet 9*: 421-7.

Wang GH, and Seeger C 1992. The reverse transcriptase of hepatitis B virus acts as a protein primer for viral DNA synthesis. *Cell 71*: 663-70.

Zou S, Ke N, Kim JM, and Voytas DF 1996. The *Saccharomyces* retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. *Genes Dev 10*: 634-45.

## FIGURE LEGENDS

**FIGURE 1. Sequences of the Ty5 PBS and the initiator methionine tRNA. On the left is shown the sequence and secondary structure of the *Saccharomyces cerevisiae* IMT.** The underlined IMT sequences can base pair with the Ty5 primer binding site (PBS).

The numbering of the IMT residues is based on the system for the elongator methionine tRNA (von Pawel-Rammingen et al., 1992). On the right is shown the sequence of the first 300 bases of the Ty5 retrotransposon (GenBank accession number: U19264). The arrows indicate the inverted repeats at the ends of the Ty5 LTR. The asterisk at base 176 denotes the Ty5 transcription start site. The derived amino acid sequences are shown above the nucleotide sequence. The region of complementarity with the IMT (the PBS) is underlined, and base pairing is illustrated below the sequence.

**FIGURE 2. Assay systems used to determine the effect of IMT mutations on Ty5 transposition.** A) The assay system used for *imt* mutants that support translation. The strain shown has all four of its *IMT* genes disrupted by *TRP1* and carries a mutant *imt4-x* on a *LEU2*-based plasmid. The mutant *imt* supports translation and is tested for its effect on transposition. Ty5 is carried on a *URA3*-based plasmid. B) The assay system used for *imt* mutants that cannot support translation. The Ty5 element and *imt4-C31,G39* (which supports translation but not transposition) are carried on a *URA3*-based plasmid. A second *imt-x* that cannot support translation is introduced to test its effect on transposition.

**FIGURE 3. The effect of Ty5 PBS mutations on transposition.** Base pairings between the IMT and the wildtype PBS, *pbs-1* and *pbs-2* are shown. Transposition frequencies are calculated as described in Materials and Methods.

**FIGURE 4. The effect of IMT mutations on Ty5 transposition.** Base pairings between the Ty5 PBS and the wildtype and mutant IMTs are shown. Also shown are the overall and the relative transposition frequencies.

**FIGURE 5. Complementarity between the IMT and the Ty5 is essential for Ty5 transposition.** A) Base pairings between Ty5 PBSs (wildtype, *pbs-3* and *pbs-4*) and the IMTs (wildtype and *imt4-C31,G39*) are shown. B) Transposition assay results for the strains with different combinations of the Ty5 elements and *IMT* genes. The numbers in parenthesis indicate fold decrease compared to the strain with a wildtype Ty5 and *IMT4*.

**FIGURE 6. The alignment of the *S. cerevisiae*, the hybrid *A. thaliana* and the hybrid *S. pombe* initiator methionine tRNA sequences.** The underlined sequences indicate the region that base pairs with the Ty5 PBS. The bold sequences in the *S. cerevisiae* IMT indicate the residues that were changed and tested in this study (Table 2). The bold sequences in the hybrid IMTs indicate the residues that differ from the *S. cerevisiae* IMT; bold sequences within the anticodon stem-loop were changed to match the *S. cerevisiae* sequence and tested for their effect on transposition (Table 3).

Table 1. The effect of *imt* anticodon stem-loop mutations on Ty5 transposition.

| Strains | Plasmids | Translation | Transposition | *IMT* |
|---|---|---|---|---|
| DVe67 | pKC35 | + | 1.00 | *IMT4* |
| NKe540 | pNK540 | + | 0.53 | *imt4-C33* |
| DVe90 | pKC74 | + | 0.64 | *imt4-U38* |
| NKe493 | pNK493 | + | 1.32 | *imt4-U41* |
| NKe496 | pNK496 | + | 0.64 | *imt4-A29, U41* |
| NKe547 | pNK547 | + | 0.27 | *imt4-C29, G41* |
| NKe548 | pNK548 | + | 0.41 | *imt4-U29, A41* |
| DVe369 | pKC77 | - | 0.15[a] | *imt4-G32* |
| DVe370 | pKC81 | - | 0.08[a] | *imt4-G34* |
| DVe373 | pKC80 | - | 0.32[a] | *imt4-C35* |
| DVe367 | pKC75 | - | 0.05[a] | *imt4-C36* |
| DVe372 | pKC78 | - | 0.03[a] | *imt4-U37* |
| DVe371 | pKC79 | - | 0.02[a] | *imt- A38* |

[a] Transposition frequency was calculated by the two *IMT* assay shown in Fig. 2B. All mutant *imt* genes tested are stable and support Ty1 transposition (Keeney et al., 1995).

Table 2. The effect of *IMT* mutations in regions other than the anticodon stem-loop on Ty5 transposition.

| Strains | Plasmids | Translation | Transposition | *IMT* |
|---|---|---|---|---|
| DVe67 | pKC35 | + | 1.00 | *IMT4* |
| Dve17 | pKC10 | + | 0.74 | *imt4-9* |
| NKe494 | pNK494 | + | 0.77 | *imt4-U12, A23* |
| DVe453 | pIMT116 | + | 0.66 | *imt4-+A17* |
| DVe454 | pIMT123 | + | 1.65 | *imt4-A50, U64* |
| DVe455 | pVIT83 | + | 0.65 | *imt4-G50, C64* |
| Nke495 | pNK495 | + | 0.90 | *imt4-U52, A62* |
| DVe447 | pIMT118 | + | 1.44 | *imt4-C54* |
| DVe449 | pIMT118a | + | 1.41 | *imt4-G54* |
| DVe451 | pIMT120 | - | 0.62[a] | *imt4-U54, C60* |
| DVe452 | pIMT121 | + | 1.02 | *imt4-C54, U60* |
| DVe450 | pIMT119 | + | 0.48 | *imt4-C60* |
| DVe448 | pIMT119a | + | 1.15 | *imt4-G60* |

[a] Transposition frequency was calculated by the two *IMT* assay shown in Fig. 2B.

Table 3. The effect of heterologous *IMT* genes on Ty5 transposition.

| Strains | Plasmids | Translation | Transposition | *IMT* |
|---|---|---|---|---|
| DVe67 | pKC35 | + | 1.00 | *IMT4* |
| NKe514 | pNK514 | + | 0.10 | *S. pombe IMT* |
| NKe515 | pNK515 | + | 0.18 | *S. pombe IMT* with *IMT4* anticodon stem-loop |
| DVe376 | pJK258 | + | 0.10 | *S. pombe IMT* with *IMT4* acceptor stem |
| NKe518 | pNK518 | + | 0.14 | *A. thaliana IMT* |
| NKe519 | pNK519 | + | 0.60 | *A. thaliana IMT* with *IMT4* anticodon stem-loop. |
| DVe375 | pJK244 | + | 0.10 | *A. thaliana IMT* with *IMT4* acceptor stem |

Fig. 1

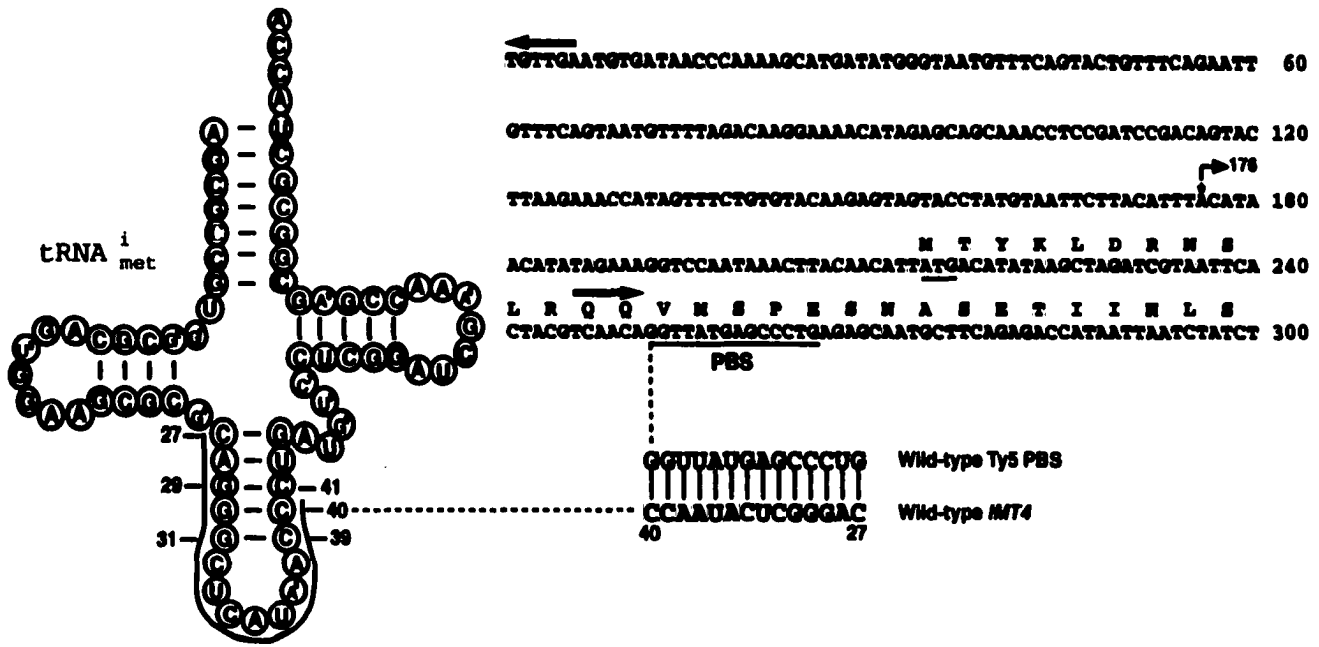

tRNA i met
27
29
31
41
40
39
TGTTGAATGTGATAACCCAAAAGCATGATATGGGTAATGTTTCAGTACTGTTTCAGAATT 60
GTTTCAGTAATGTTTTAGACAAGGAAAACATAGAGCAGCAAACCTCCGATCCGACAGTAC 120
176
TTAAGAAACCATAGTTTCTGTGTACAAGAGTAGTACCTATGTAATTCTTACATTTACATA 180
M T Y K L D R N S
ACATATAGAAAGGTCCAATAAACTTACAACATTATGACATATAAGCTAGATCGTAATTCA 240
L R Q Q V N S P E S N A S E T I I N L S
CTACGTCAACAGGTTATGAGCCCTGAGAGCAATGCTTCAGAGACCATAATTAATCTATCT 300
PBS
GGUUAUGAGCCCUG
Wild-type Ty5 PBS
CCAAUACUCGGGAC
Wild-type MT4
40
27

Fig. 2

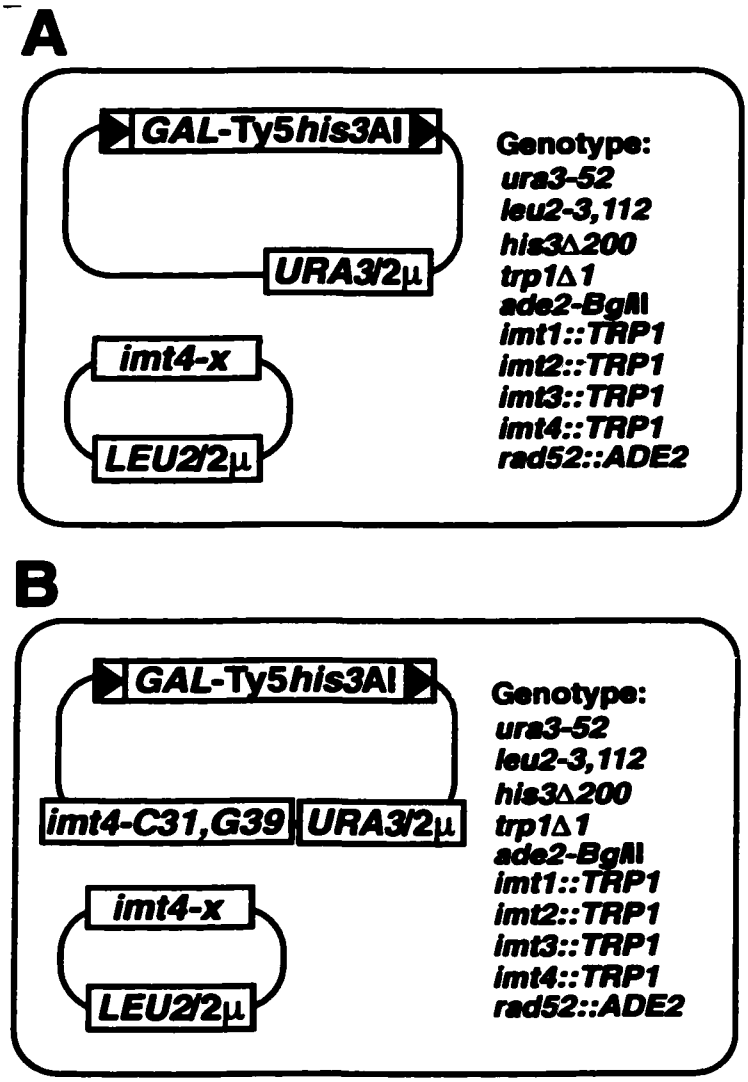

A
GAL-Ty5his3AI
URA3/2μ
imt4-x
LEU2/2μ
Genotype:
ura3-52
leu2-3,112
his3Δ200
trp1Δ1
ade2-BglII
imt1::TRP1
imt2::TRP1
imt3::TRP1
imt4::TRP1
rad52::ADE2
B
GAL-Ty5his3AI
imt4-C31,G39
URA3/2μ
imt4-x
LEU2/2μ
Genotype:
ura3-52
leu2-3,112
his3Δ200
trp1Δ1
ade2-BglII
imt1::TRP1
imt2::TRP1
imt3::TRP1
imt4::TRP1
rad52::ADE2

Fig. 3

| Ty5 PBS mutants | | Transposition frequencies ($X10^{-4}$) | Fold decrease relative to wild type Ty5 |
|---|---|---|---|
| GGUUAUGAGCCCUG<br>CCAAUACUCGGGAC<br>40 27 | WT PBS<br>WT *IMT4* | 2.70 ± 0.98 | 1 |
| AGUAAUGUCACCUG<br>CCAAUACUCGGGAC<br>40 27 | *pbs-1*<br>WT *IMT4* | 0.01 ± 0.00 | 3150 |
| GGUAAUGUCACCUG<br>CCAAUACUCGGGAC<br>40 27 | *pbs-2*<br>WT *IMT4* | 0.03 ± 0.02 | 794 |

Fig. 4

| *imt4* mutants | | Transposition frequencies ($X10^{-5}$) | Fold decrease relative to wild type *IMT4* |
|---|---|---|---|
| GGUUAUGAGCCCUG<br>CCAAUACUCGGGAC<br>40 27 | WT PBS<br>WT *IMT4* | 11.30 ± 4.11 | 1 |
| GGUUAUGAGCCCUG<br>CUAAUACUCUGGAC<br>40 27 | WT PBS<br>*imt4-U31,U39* | 3.76 ± 0.60 | 3 |
| GGUUAUGAGCCCUG<br>CUAAUACUCUGAAC<br>40 27 | WT PBS<br>*imt4-A29,U31,U39* | 0.61 ± 0.78 | 19 |
| GGUUAUGAGCCCUG<br>CAAAUACUCUGGAC<br>40 27 | WT PBS<br>*imt4-U31,A39* | 0.05 ± 0.01 | 226 |
| GGUUAUGAGCCCUG<br>CGAAUACUCCGGAC<br>40 27 | WT PBS<br>*imt4-C31,G39* | 0.02 ± 0.01 | 565 |

Fig. 5

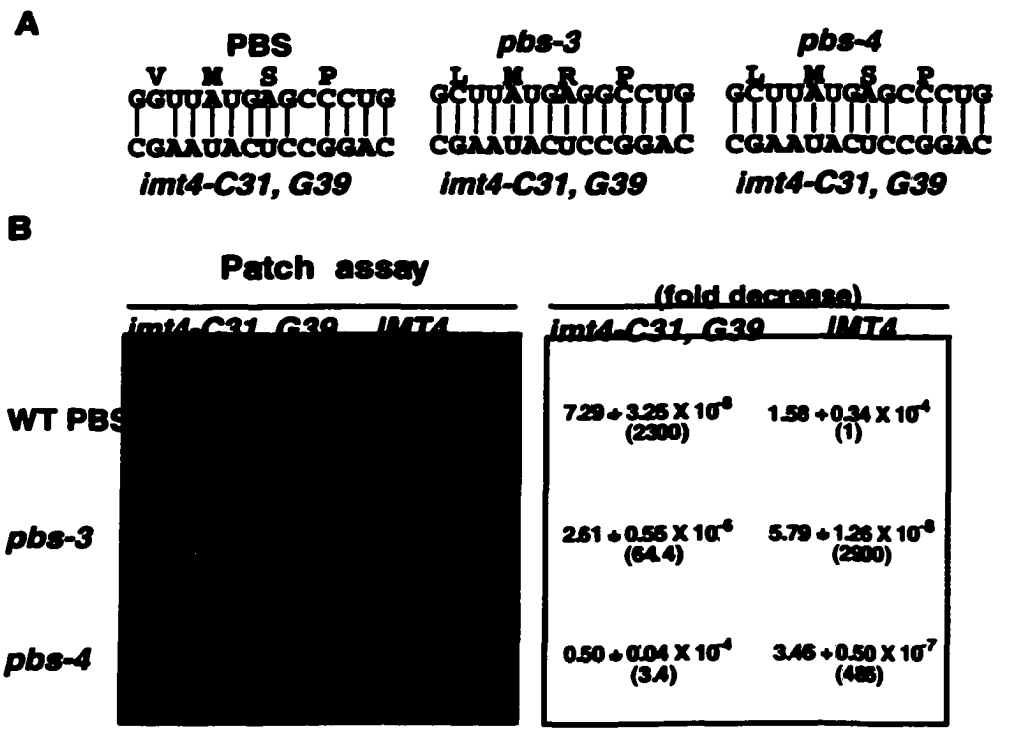
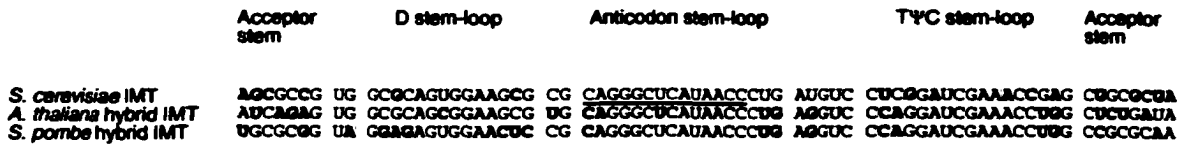

A
PBS
V M S P
GGUUAUGAGCCCUG
CGAAUACUCCGGAC
imt4-C31, G39
pbs-3
L M R P
GCUUAUGAGGCCUG
CGAAUACUCCGGAC
imt4-C31, G39
pbs-4
L M S P
GCUUAUGAGCCCUG
CGAAUACUCCGGAC
imt4-C31, G39
B
Patch assay
(fold decrease)
imt4-C31, G39
IMT4
imt4-C31, G39
IMT4
WT PBS
7.29 ± 3.26 X 10^-6 (2300)
1.58 ± 0.34 X 10^-4 (1)
pbs-3
2.51 ± 0.55 X 10^-5 (64.4)
5.79 ± 1.26 X 10^-6 (2900)
pbs-4
0.50 ± 0.04 X 10^-4 (3.4)
3.46 ± 0.50 X 10^-7 (485)

Fig. 6


Acceptor stem
D stem-loop
Anticodon stem-loop
TΨC stem-loop
Acceptor stem
S. cerevisiae IMT
A. thaliana hybrid IMT
S. pombe hybrid IMT

# ACKNOWLEDGEMENTS

I am grateful to Dr. Daniel F. Voytas for his guidance and generosity throughout my graduate studies. I would like to give my thanks to Dr. Leslie Miller who is so kind to provide great chances and helps for me to study computational knowledge. I am also thankful to my collaborators Kent Vander Velden, Dr. Ning Ke and lab members, Yevette Chin, Dr. Xiaowu Gai, Daniel J. Rowley, Brooke Peterson-Burch, Dr. Dave Wright, Phillip Irwin, Weiwu Xie, Peter Fuerst, Ericka Havecker and Junbiao Dai for their support and assistance.