EUGENE MICHAEL MCCARTHY

LTR_STRUC, A Novel Data-Mining Tool, and Its Application to the Rice and Mouse Genomes

(Under the direction of JOHN F. MCDONALD)

The research presented in this dissertation consists of three parts: 1) a description of the design and function of a novel data-mining program, LTR_STRUC; 2) a survey of LTR retrotransposons in the rice genome; 3) a survey of LTR retrotransposons in the mouse genome. The algorithm used by LTR_STRUC differs at a fundamental conceptual level from that employed in BLAST-type, query-based programs and thus provides an alternative, complementary method of identifying LTR retrotransposons in nucleotide sequence data. We combined LTR_STRUC and conventional techniques to thoroughly search of the rice and mouse genomes for LTR retrotransposons. In rice, we found 59 families (37 *copia*-like, 20 *gypsy*-like, 2 non-autonomous). In mouse, we found 20 (all *gypsy*-like). In both species, we were able to more than double the number of recognized LTR retrotransposon families, a testament to the efficacy of LTR_STRUC-supported retrotransposon surveys.

INDEX WORDS: *Oryza sativa, Mus musculus, copia, gypsy,* Transposable element, Retrotransposon, LTR, Data-miner, Search algorithm

LTR_STRUC, A NOVEL DATA-MINING TOOL, AND ITS APPLICATION TO THE RICE AND MOUSE GENOMES

by

EUGENE MICHAEL MCCARTHY

B.S., The University of Georgia, 1983

M.S., The University of Georgia, 1995

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2003

LTR_STRUC, A NOVEL DATA-MINING TOOL, AND ITS APPLICATION TO THE RICE AND MOUSE GENOMES

by

EUGENE MICHAEL MCCARTHY

Major Professor: John F. McDonald

Committee: Eileen Kraemer
John Avise
Daniel Promislow
Ethan W. Taylor

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2003

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

To date, most searches of nucleotide sequence databases have been query based. In such surveys the researcher compares a sequence of interest, the query, with available nucleotide data, to see whether the query can be matched with one or more sequences in the database. Various criteria have been employed in defining the degree to which two sequences "match." In general, however, when a query is aligned with a prospective match, the percentage of aligned nucleotide pairs that are identical must significantly exceed random expectation. Thus, for two nucleotide sequences selected at random from a dataset in which each of the four possible nucleotides are present in equal proportions, the expected level of identity between the two sequences is 25%. A researcher might therefore take some higher level of identity, say 90%, as the criterion of a match and then proceed to identify all sequence in a database that were 90% or more identical to a query.

An alternative method of search is to seek in nucleotide sequences defined structures characteristic of a particular class of genes. For example, if a tRNA is to form a cloverleaf, it must be composed of nucleotides that are capable of forming that particular sort of secondary structure. That is, it must contain four separate pairs of mutually complementary sequences that allow the annealing that must occur if a cloverleaf is to form. Further, these sequences must appear in the proper order and at the proper distances from each other if the tRNA is to achieve proper folding. Thus, since the primary structure of *any* tRNA gene should contain four such pairs of sequences, a

computer can be programmed to identify tRNA genes by looking for regions that contain four such pairs. Note that this approach does not require a putative tRNA gene ("hit") to be sequentially similar to any known tRNA. Instead the computer looks for a particular type of structure. Another example of a commonly used search algorithm that depends on this general approach would be the class of gene-finding programs that search for genes by looking for ORFs rather than by searching for sequences similar to known genes.

I have used a structural approach of this nature to create a new data-mining program, LTR_STRUC, for automated identification and analysis of a biologically important class of transposable elements, long terminal repeat (LTR) retrotransposons, in nucleotide sequence data. In higher eukaryotes a large fraction of the genome is typically comprised of transposable elements (TEs), repetitive DNA sequences that are able to move from one chromosome location to another (Sherratt, 1995). Such "transposition" of elements is now known to be an important source of mutation. Indeed, many plant and animal diseases have been attributed to the insertion of TEs. On the basis of their mode of replication, transposable elements are classified into two major classes. Class II (DNA) elements use an element-encoded enzyme, transposase, to cut themselves out of the host genome. These excised elements then replicate and reinsert throughout the genome. While TEs belonging to Class II are abundant in bacteria and lower eukaryotes, they are relatively rare in higher eukaryotes. All other TEs belong to Class I, the retrotransposons. Inserted copies of retrotransposons are transcribed using the host transcriptional machinery. Some of these transcripts are processed into mRNAs encoding proteins essential to element replication. Other transcripts remain full-length and serve as

a template for reverse transcription into DNA, a process carried out by reverse transcriptase, an element-encoded RNA-dependent DNA polymerase. Retrotransposons are particularly abundant in plants, where they are often a principal component of nuclear DNA. In corn, 50-80%, and, in wheat, fully 90% of the genome is made up of retrotransposons (Flavell, 1986; SanMiguel et al., 1996). In animals this percentage is generally lower than in plants but it can still be significant. For example, at least 40 percent of the human genome is composed of retrotransposons (Yoder *et al.,* 1997).

LTR retrotransposons are a major subclass of Class I elements. They have a structure and mode of replication similar to infectious retroviruses (Coffin et al., 1997). All retrotransposons are distinguished by a life cycle involving an RNA intermediate. The RNA genome of a retrotransposon is copied into a double-stranded DNA molecule by reverse transcriptase and is subsequently integrated into the host genome. The integrated, or proviral form, of an LTR retrotransposon has characteristic structural features, in particular a pair of long repeats is present, one member of the pair at each end of the element. Because they are situated at the two termini of the element they are referred to as terminal repeats. If an element is capable of autonomous replication, then the interval between its LTRs will be no shorter than about 3 kbp. The interior of the element must be large enough to contain all of the proviral genes that are essential for replication (some proviruses can avoid this constraint by obtaining the various enzymes necessary for replication from other proviruses in the same cell). The upper bound on length for a replication-competent provirus, although not precisely known, seems to be about 30 kbp (longer lengths appear to interfere with encapsulation in the capsid). The LTRs themselves can vary in length from 100 bp up into the thousands of base pairs. In

addition, because during insertion LTR retrotransposons make staggered cuts in the host DNA sequence, proviruses are flanked by target site repeats (usually 4-6 bp in length). About 8% of the human genome is now known to be specifically composed of LTR retrotransposons (Lander et al., 2001). In the mouse genome the figure is comparable, estimated at 10% (Mouse Genome Sequencing Consortium, 2002).

In Chapter 2, LTR_STRUC is described and discussed in detail, but I will briefly mention here that it searches sequence data for two long stretches of nucleotides (putative LTRs) that are 1) highly similar to each other; 2) reasonably near each other (i.e., no further apart than one would expect in the case of the two LTRs of the same element); 3) flanked by short target site repeats. These are characteristic features of an LTR retrotransposon, but again, recognition of hits by this method does not require sequential similarity to a query. LTR_STRUC does not require putative LTR retrotransposons to share sequential identity with any previously known LTR retrotransposon. The search is for members of a class of structures rather than for members of a set of sequences that exceed a specified level of identity.

Prior to the development of LTR_STRUC, workers in our laboratory had searched *Saccharomyces cerevisiae* (Jordan and McDonald, 1999) the *Caenorhabditis elegans* (Bowen and McDonald, 1999) genomes for LTR retrotransposons. Other laboratories had conducted similar projects (e.g., Kumekawa *et al.,* 1999; Tristem, 2000). These searches used the sequences of reverse transcriptases from known LTR retrotransposons as BLAST queries to identify putative transposons in available sequence data. Candidate retrotransposons were then analyzed to further characterize and evaluate the nature of the hit. This approach has two drawbacks. First there is the potential for overlooking

elements that either lack a reverse transcriptase ORF (a common phenomenon in LTR retrotransposons) or that were insufficiently similar to the query due to sequence divergence. Secondly, much of the process of analysis and characterization involved crude and labor-intensive methods such as "sequence-gazing" to find LTRs and target site repeats. LTR_STRUC reduces the bias inherent in such an approach by providing the researcher with a method of finding even transposons that entirely lack ordinary retroviral ORFs. It also automates much of the analysis process, reducing the time required for finding, characterizing, and classifying the types of LTR retrotransposon in a large genome from months to hours.

In addition to the technical description of LTR_STRUC (Chapter 2) this dissertation presents the results of two biological surveys. I used LTR_STRUC to search the rice (*Oryza sativa*) and mouse (*Mus musculus*) genomes for LTR retrotransposons. In both these species I went on to use the elements identified by LTR_STRUC, together with previously known retrotransposons from rice and mouse, as queries in BLAST searches in an attempt to identify all types of LTR retrotransposons present in these two genomes.

These two searches were intended as tests of the program that would, in addition, yield useful data. The results were more successful than we had hoped. Many new families were discovered and characterized. Our surveys more than doubled the number of recognized families for both rice and mouse. LTR_STRUC has stimulated a good deal of interest in the retroelement community, leading me to make it freely available to researchers on website (http://www.genetics.uga.edu/retrolab/data/LTR_struc.html) of our laboratory. Many of the newly discovered elements have unusual structures that have

prompted interest in conducting future studies. It is my hope that the novel families of retrotransposons uncovered by LTR_STRUC will help stimulate new hypotheses and shed new light on the biological significance of transposable elements.

CHAPTER 2

LTR_STRUC: A NOVEL SEARCH AND IDENTIFICATION PROGRAM FOR LTR RETROTRANSPOSONS [1]

[1]McCarthy, E.M. and J.F. McDonald. 2003. Bioinformatics 19: 362-367 (reprinted with permission of publisher).

ABSTRACT

Motivation: Long terminal repeat (LTR) retrotransposons constitute a substantial fraction of most eukaryotic genomes and are believed to have a significant impact on genome structure and function. Conventional methods used to search for LTR retrotransposons in genome databases are labor intensive. We present an efficient, reliable and automated method to identify and analyze members of this important class of transposable elements.

Results: We have developed a new data-mining program, LTR_STRUC (*LTR* retrotransposon *struc*ture program), which identifies and automatically analyzes LTR retrotransposons in genome databases by searching for structural features characteristic of such elements. LTR_STRUC has significant advantages over conventional search methods in the case of LTR retrotransposon families having low sequence homology to known queries or families with atypical structure (e.g., non-autonomous elements lacking canonical retroviral ORFs) and is thus a discovery tool that complements established methods. LTR_STRUC finds LTR retrotransposons using an algorithm that encompasses a number of tasks that would otherwise have to be initiated individually by the user. For each LTR retrotransposon found, LTR_STRUC automatically generates an analysis of a variety of structural features of biological interest.

INTRODUCTION

Retrotransposons are a major component of eukaryotic genomes. For example, at least 40 percent of the human genome is composed of retrotransposons (Yoder *et al.,* 1997). Long terminal repeat (LTR) retrotransposons have a structure and mode of replication similar to infectious retroviruses (Coffin et al., 1997). All retrotransposons are distinguished by a life cycle involving an RNA intermediate. The RNA genome of a retroelement is copied into a double-stranded DNA molecule by reverse transcriptase and is subsequently integrated into the host genome.

We here describe a new data-mining program, LTR_STRUC (*LTR* retrotransposon *struc*ture) that provides a rapid, automated method for finding, delineating, and analyzing LTR retrotransposons in large genome databases. In identifying LTR retroelements in nucleotide sequence data, the search algorithm used by LTR_STRUC seeks certain generic structural features of such elements. The algorithm differs from conventional search methods, in which locating and identifying transposons depends on sequence similarity to previously identified elements. LTR_STRUC identifies full-length LTR retrotransposons independent of sequence homology and, as such, is complementary to conventional search methods.

SYSTEM AND METHODS

A Search Algorithm with a Structural Basis

In scanning genomic sequence data, LTR_STRUC employs a search algorithm that first identifies putative LTRs and then analyzes them further, looking for other defining features of LTR retrotransposons. The structure of a typical LTR

retrotransposon is shown in Figure 1. Structural features important to LTR_STRUC include two sites critical to replication, the primer binding site (PBS) and polypurine tract (PPT), as well as, the presence of dinucleotides at the ends of each LTR (typically TG and CA). Particularly important are the direct or "target-site" repeats (see Figure 1). When a LTR retroelement inserts itself into host DNA, a short (4-6 bp) segment of host DNA is replicated at the site of insertion ("target site repeat" or TSR). This feature allows LTR_STRUC to make an exact demarcation of the limits of a putative element.

Step 1: Finding An Initial Pair Of Matches: For each LTR retrotransposon present in the input file, LTR_STRUC first seeks the LTR pairs present at the ends of a putative element and then searches for additional characteristic retrotransposon features to confirm the hit. The two LTRs of a particular element may fall anywhere within the contig, but the distance (D) between their 5′ ends should fall within the range dictated by the expected range of lengths characteristic of LTR retrotransposons (see below). Thus, it is possible to specify reasonable values for d_{min} and d_{max} such that the relationship $d_{min} < D < d_{max}$ will hold true for the vast majority of LTR retrotransposons (even those bearing large inserts).

Suppose (I):

1) that the first nucleotide of a sequence S_i lies at position i in the input nucleotide sequence and denote the length of S_i by $L_S(i)$ (Figure 2);

2) that both S_i and its two endpoints (the two nucleotides i and $i + L_S(i) - 1$) lie entirely within the bounds of an LTR, $L_{5'}$, situated at the 5′ end of an as yet unidentified LTR retrotransposon, R;

3) that the first nucleotide of a second sequence M_k, which is highly similar to S_i,

lies at position k of the input nucleotide sequence and denote the length of M_k by $L_M(k)$;

4) that M_k and its two endpoints (the two nucleotides k and $k + L_M(k) - 1$) lie entirely within the bounds of $L_{3'}$, the 3′ LTR of R;

5) that M_k lies at the same relative position within $L_{3'}$ as does S_i within $L_{5'}$;

If the conditions specified in (I) hold, then

$$i + d_{min} < k < i + d_{max}, \text{ where } d_{min} < D < d_{max} \qquad \text{(II)}$$

Thus, if for a given i, LTR_STRUC searches the interval $(i + d_{min}, i + d_{max} + L_S(i) - 1)$ and finds a sequence M_k which is highly similar to S_i, then it has found a pair of sequences (S_i and M_k) such that one (S_i) is likely to lie in the 5′ LTR, and the other (M_k) is likely to lie in the 3′ LTR of a retrotransposon.

Now, further suppose (III):

1) that the search of the input contig begins at its 5′ end;

2) that the search proceeds in the 3′ direction taking sample sequences S_i at intervals of length Δi and, for each such sample sequence the search has scanned the interval, $i + d_{min} < k < i + d_{max} + L_S(i) - 1$, for a match M_k to S_i;

3) that realistic values for d_{min} and d_{max} have been chosen.

If the conditions specified in (I) and (III) hold, then for any given i, the search for a match, M_k, to S_i has been exhaustive so long as the search has compared S_i to every sequence beginning with a nucleotide that falls in the range specified by (II). If, for any S_i, a sufficiently matching sequence, M_k, (that is one greater than 40 nucleotides in length and exhibiting greater than 70% homology) is found, such that the first nucleotide, k, of M_k lies in the interval $i + d_{min} < k < i + d_{max}$, then LTR_STRUC proceeds to Step 2

(alignment of putative LTRs). Otherwise i is incremented by Δi and a match for sequence $S_{i+\Delta i}$ is sought.

Step 2: Alignment of regions flanking the initial pair of matches: Alignment proceeds outward from the site of the initial match, (S_i and M_k). See Figure 2. Regions 3′ to S_i and M_k (i.e., e_1, e_2, e_3, e_4, and e_5 in Figure 2) are aligned first, then those 5′ direction (e_6 and e_7). The alignment proceeds outward by steps of size *N*. Examination of alignment results has shown that high quality alignments are obtained if $N = 100$.

The alignment steps outward because it is unknown at any given step whether the high levels of sequence similarity present in regions that have already been aligned will continue into adjacent regions. This approach to alignment saves computer time by avoiding alignment of regions outside the putative LTRs. It also provides a means of detecting their approximate ends since the expectation is that sequence similarity will fall to near random levels once the alignment passes the end of the LTRs.

The value of *N* sets an upper bound on the length of any given extension (See Figure 2). The extension is composed of the two sequences aligned in a single "step" one from the putative 5' LTR and one from the putative 3′ LTR. Figure 2 illustrates the alignment of two LTRs in seven successive extensions (e_1, e_2, e_3, e_4, e_5, e_6, e_7). The lengths of the two sequences composing an extension will usually differ.

Starting with the first extension, extensions are added in the 3′ direction until the level of similarity between the aligned sequences falls below 70 percent for two successive extensions. Then extension begins in the 5′ direction and is continued until similarity between aligned sequences again falls below the 70 percent criterion for two

successive extensions. At this point in the algorithm the LTRs are considered to be fully aligned and the analysis proceeds to Step 3.

Step 3: Identification of the approximate endpoints of the aligned LTRs: Two adjacent windows β and γ, each of length 100bp, are passed across the LTR alignment. These windows may be described as the intervals (j-100, j-1) and (j, j+99), respectively, where j is an index specifying position in the alignment. As j is incremented and the windows pass across the alignment, at each position, T_β (the total number of matches in window β) is subtracted from T_γ (the total number of matches in window γ) to obtain the difference

$$\Delta = T\gamma - T_\beta.$$

Where Δ reaches a maximum and minimum, k will be a good approximation of the respective positions in the alignment of the 5′ and 3′ ends of the LTRs; the former of these two values of Δ denote $\hat{\jmath}_{5'}$ and the latter, $\hat{\jmath}_{3'}$.

Step 4: Determination of exact end points: The two alignment positions, $\hat{\jmath}_{5'}$ and $\hat{\jmath}_{3'}$, generated by Step 3 correspond to four indices in the original input contig, that is, to a pair of approximate endpoints for the 5′ LTR, $\alpha_{5'}(\hat{\jmath}_{5'})$, $\omega_{5'}(\hat{\jmath}_{3'})$, and an equivalent pair for the 3′ LTR, $\alpha_{3'}(\hat{\jmath}_{5'})$, and $\omega_{3'}(\hat{\jmath}_{3'})$. (Figure 4). Since these indices are good approximations of the actual endpoints, denoted $\acute{\alpha}_{5'}$, $\acute{\omega}_{5'}$, $\acute{\alpha}_{3'}$, and $\acute{\omega}_{5'}$, of the two LTRs, the inequalities

$$|\acute{\alpha}_{5'} - \alpha_{5'}(\hat{\jmath}_{5'})| < \delta, \quad |\acute{\omega}_{5'} - \omega_{5'}(\hat{\jmath}_{3'})| < \delta, \quad |\acute{\alpha}_{3'} - \alpha_{3'}(\hat{\jmath}_{5'})| < \delta, \text{ and } \quad |\acute{\omega}_{5'} - \omega_{3'}(\hat{\jmath}_{3'})| < \delta, \quad \text{(IV)}$$

should hold for some small integer δ (Figure 4).

The score given each of the quartets, ($\alpha_{5'}$, $\omega_{5'}$, $\alpha_{3'}$, $\omega_{3'}$), not only determines which quartet will be chosen as the endpoints of the posited transposon (i.e., the quartet receiving the maximum score), but also serves as a measure of how likely it is that the posited LTR retrotransposon is real. In practice we have found that an "element" for which the maximum quartet score is low is usually not an LTR retrotransposon at all. As the maximum quartet score approaches one (the highest possible score), however, it becomes a near certainty that the hit actually does represent an LTR retrotransposon. Using these scores to rank-order the output data in terms of hit quality has significantly reduced the amount of labor required in subsequent analysis of results.

IMPLEMENTATION

The LTR_STRUC is written in Visual C++ (Microsoft version 6.0) and the currently available version runs on a PC platform, but if the user has the UNIX version of MFC on his or her UNIX system, the LTR_STRUC should port with little difficulty. The program is available from the authors as a console application. LTR_STRUC reads nucleotide sequence files in FASTA format. LTR_STRUC breaks sequences containing strings of separator symbols (such as "n", "N", or "-") at the point where the separator string occurs, treating the sequences on either side of the string as separate contigs. Input files must be downloaded and scanned locally on the user's computer. The user can specify certain parameters:

1) maximum and minimum overall length of the transposon
2) maximum and minimum lengths for the LTRs
3) cutoff score

LTR_STRUC generates report files (in text format) only for hits generating a score in excess of the cutoff score. These files contain a detailed analysis of each hit. They include all the information enumerated in Table 1.

RESULTS AND DISCUSSION

The search algorithm used by LTR_STRUC consists of four steps: 1) location of an initial pair of matches (S_i,M_k), which might lie within an LTR pair (these matches need not be exact); 2) alignment of the regions adjacent to the initial match; 3) identification of the approximate end points of the putative LTRs; and 4) determination of exact end points of the putative LTRs. In addition, LTR_STRUC's reporting function provides analytic output of specific interest to researchers studying retrotransposons.

The conventional approach to identifying LTR retrotransposons in genome databases typically begins by scanning input nucleotide sequences for sequences showing similarity to known reverse transcriptases (RTs). From the standpoint of identifying LTR retrotransposons, this approach has three inherent drawbacks: 1) It biases the search toward elements containing RTs that are similar to the query RT; 2) It overlooks LTR elements that lack a reverse transcriptase – such elements are common in some species (Witte et al., 2001); 3) Even after an RT has been identified by the traditional method, completion of the analysis typically requires a number of additional steps for each putative element found. The first of the two negative aspects of the conventional approach are significant problems for researchers wishing to find *all* LTR retrotransposons in a given input data set. The third means extra labor that can be

avoided by using LTR_STRUC because it eliminates a series of additional steps required in the conventional approach.

LTR_STRUC is only effective for full-length LTR retrotransposons. To date, our tests indicate that its sensitivity and accuracy in *locating* elements in this category are comparable to established techniques. However, LTR_STRUC has been found to be superior to existing techniques with regard to *identifying* LTR retrotransposons that belong to families having novel structure or that have low sequence homology to previously recognized families. To find new families of LTR retrotransposons, a researcher has to sift through a typically lengthy list of anonymous sequences and analyze their structures before any definite identification could be made. In contrast, because LTR_STRUC relies on an algorithm that focuses on a structural analysis of the input nucleotide sequence, any hit with a high score provides the user with a high degree of assurance that the hit in question actually is an LTR retrotransposon. From the standpoint of discovering new families, BLAST searches have the inherent drawback that all of the high-score hits will be similar to the known query. Prior to the development of LTR_STRUC, our laboratory conducted a search for the presence of LTR retrotransposons in the *C. elegans* database (www.wormbase.org) using the RT from the *Cer1* elements (acc# U15406) as a query sequence. After several months of analysis, workers in our lab were able to identify and characterize 12 families of *C. elegans* LTR retrotransposons (Bowen and McDonald 1999). As a test of the accuracy and reliability of LTR_STRUC, we utilized LTR_STRUC to search the same database for LTR retrotransposons. Within two hours, LTR_STRUC had identified and characterized all

12 families of *C. elegans* LTR retrotransposons identified in our previous study, as well as, three additional elements not detected in our initial analysis.

We have also employed LTR_STRUC to search for LTR retrotransposons in the rice (*Oryza sativa)* genome (McCarthy et al., 2002). In that survey, LTR_STRUC identified 32 new families of LTR retrotransposons, increasing the number of known, named LTR retrotransposon families in the rice genome to a total 59. Among the novel families of LTR retrotransposons identified in this study were four non-autonomous families having no significant homology to any previously described retrotransposon or retroviral proteins. These elements were detected through their structural characteristics and could easily have been overlooked by conventional search methods.

Most of the transposons detected by LTR_STRUC have had LTRs that are more than 90% identical. The vast majority of full-length LTR retrotransposons found by conventional methods, also, show high levels of LTR identity, again, usually in excess of 90% (e.g., Promislow et al., 1999; Bowen and McDonald, 1999, 2000). Occasionally, in our surveys of the rice genome LTR_STRUC detected elements with LTR-LTR similarities falling into the 80% range, but these, for the most part, were elements bearing regional duplications that extend unbroken for tens of base pairs, not ones that are heavily peppered with single-nucleotide mutations. Given the set up of the algorithm, LTR_STRUC is unlikely to detect elements where levels of LTR identity fall below about 75%.

LTR_STRUC can find nested insertions of LTR retrotransposons. LTR_STRUC found examples of such nesting during the scan of the rice genome. However, to detect instances where one retroelement inserts into another, the size of the search window (i.e.,

d_{max} - d_{min}) must be increased so that it will exceed the size of the retrotransposon receiving the insertion (which will make the program run more slowly). LTR_STRUC will also find LTR retrotransposons bearing other types of insertions (so long as the LTR pair and TSRs are still recognizable).

Because it looks first for the LTRs at opposite ends of an element, LTR_STRUC cannot find elements that span contigs. The entire element (plus some flanking sequence) must be present in a single contig for LTR_STRUC to find it. For the same reason, LTR_STRUC will not locate truncated elements or solo LTRs. To date, in surveying a particular genome, our method of identifying truncated copies in a given family has been first to identify as many families as possible either, in the literature or using LTR_STRUC. We then conduct BLAST searches against available sequence data using as queries one full-length copy from each of those families. LTR_STRUC is not intended to find truncated retrotransposons directly. Its main utility is in the discovery and initial analysis of new *families* of retrotransposons.

From the user's perspective, the program rapidly generates information that otherwise requires a great deal labor. Program run times show a linear relation with the length of the scanned contig length, but the time required increases with the square of the search window size (d_{max} - d_{min}) and maximum LTR length. Experience in our lab has shown that the overall task of finding new families of retrotransposons in genomic sequence data is significantly facilitated by LTR_STRUC. We estimate that the reduction in user time is on the order of 100-1000-fold (hours vs. months).

REFERENCES

Bowen, N. and McDonald, J.F. (1999) Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. *Genome Res.,* **9**, 924-935.

Bowen, N. and McDonald, J.F. (2000) *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. *Genome Res.,* **11**, 1527-1540.

Coffin, J.M., Hughes, S.H. and Varmus, H.E. (1997) *Retroviruses*. Plainview, NY, Coldspring Harbor Laboratory Press.

McCarthy, E.M., Liu, J., Gao, L. and McDonald, J.F. (2002) LTR Retrotransposons of *Oryza sativa*. *Genome Biology*, in press.

Promislow, D.E., Jordan, I.K. and McDonald, J.F. (1999) Genomic demography, A life history analysis of transposable element evolution. *Proc. Roy. Soc. Lon. B Biol.,* **266**, 1555-1560.

Witte, C.P., Le, Q.H., Bureau, T. and Kumar, A. (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *PNAS, USA*, **98**, 13778-83.

Yoder, J.A., Walsh, C.P. and Bestor, T.H. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.,* 13, 335-340.

TABLES

TABLE 1 Information in LTR_STRUC output files:
Name of source contig
Location of element within contig
Score for current hit
Lengths of contig, element, LTRs, and largest ORF
Nucleotide sequences for the whole transposon, TSRs, LTRs, PBS, PPT
Orientation of the transposon (determined by relative positions of PBS and PPT)
Sequences for all ORFs (longer than 50 amino acids)
Intra-element percent identity of LTRs
An alignment of the putative LTRs

FIGURES

Figure 1

Generic structure of an LTR retrotransposon. The typical retrotransposon encodes genes involved in the element's replication. The *pol* (*pol*ymerase) gene encodes reverse transcriptase/integrase proteins that are packaged within a capsid protein encoded by the *gag* (group specific antigen) gene. Reverse transcription is primed by a tRNA molecule binding to a site located 5' to the *gag* gene called the "primer binding site" (PBS). Infectious retroviruses and some LTR retrotransposons encode envelope proteins encoded by the *env* (*env*elope) gene. A "polypurine tract" (PBS) is typically located just 5' to the 3' LTR. The LTR retrotransposon genome is bordered by long terminal repeats (LTRs). The ends of the LTRs typically end in the dinucleotides TG and CA.

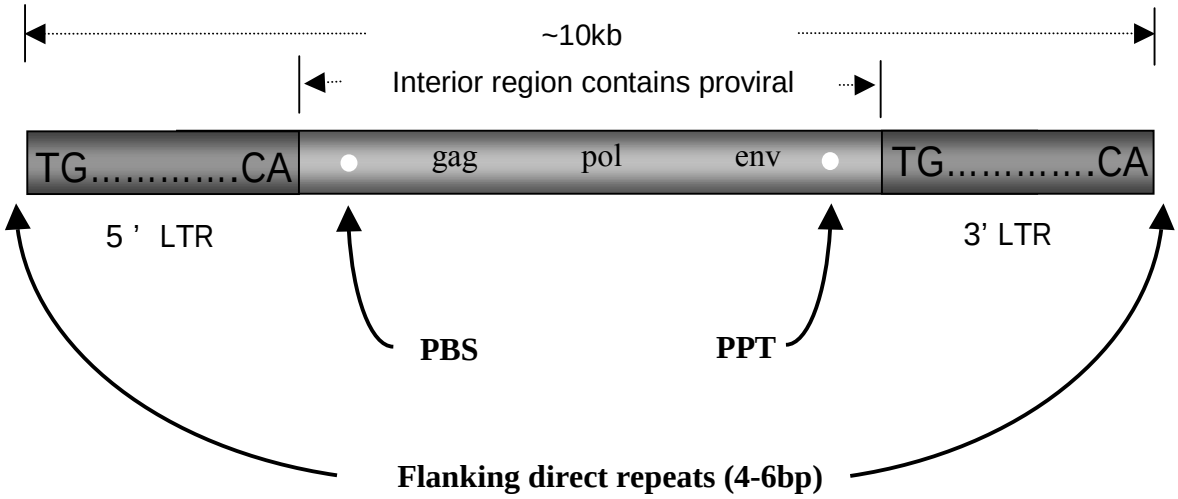

~10kb
Interior region contains proviral
TG.....CA
gag
pol
env
TG.....CA
5 ' LTR
3' LTR
PBS
PPT
Flanking direct repeats (4-6bp)

Figure 2

The alignment of an LTR proceeds in steps: LTR_STRUC extends LTR alignments in steps (e_1, e_2, e_3, e_4, e_5, e_6, and e_7), beginning each extension by finding the largest match (P_i , P_k) in the next interval (2a) and (2c) and then aligning the region between that match and the portion of the alignment that has already been completed (see Figure 3). When two matches that are of equal quality and length are identified, LTR_STRUC breaks the tie by determining which of the two possible pairs lie most nearly "opposite" each other in the alignment (2b). When LTR_STRUC reaches the 3' end of the LTR it proceeds backwards from the start point to complete the alignment (steps represented in the figure by extensions, e_6 and e_7).

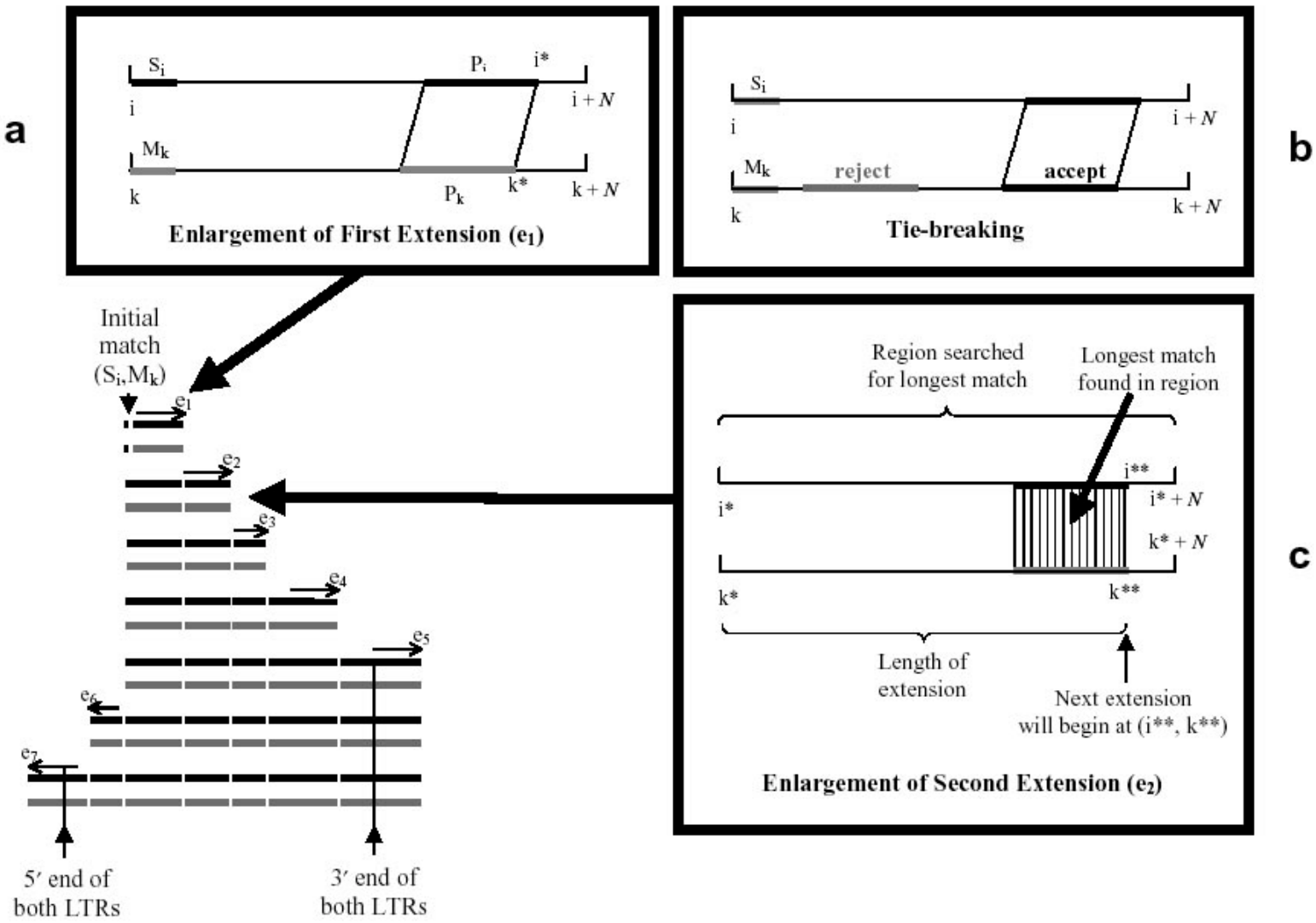

a
S_i
P_i
i^*
i
$i + N$
M_k
k
P_k
k^*
$k + N$
Enlargement of First Extension (e_1)
b
S_i
i
$i + N$
M_k
reject
accept
k
$k + N$
Tie-breaking
Initial match (S_i,M_k)
e_1
e_2
e_3
e_4
e_5
e_6
e_7
5′ end of both LTRs
3′ end of both LTRs
c
Region searched for longest match
Longest match found in region
i^{**}
$i^* + N$
i^*
$k^* + N$
k^*
k^{**}
Length of extension
Next extension will begin at (i^{**}, k^{**})
Enlargement of Second Extension (e_2)

Figure 3

LTR_STRUC uses a recursive process to complete the alignment of each extension: Once the endpoints of an extension are defined (3a), LTR_STRUC examines each remaining unaligned subinterval and then aligns the largest match in that subinterval. This process is repeated until all subintervals in the current extension are aligned (3b-3e).

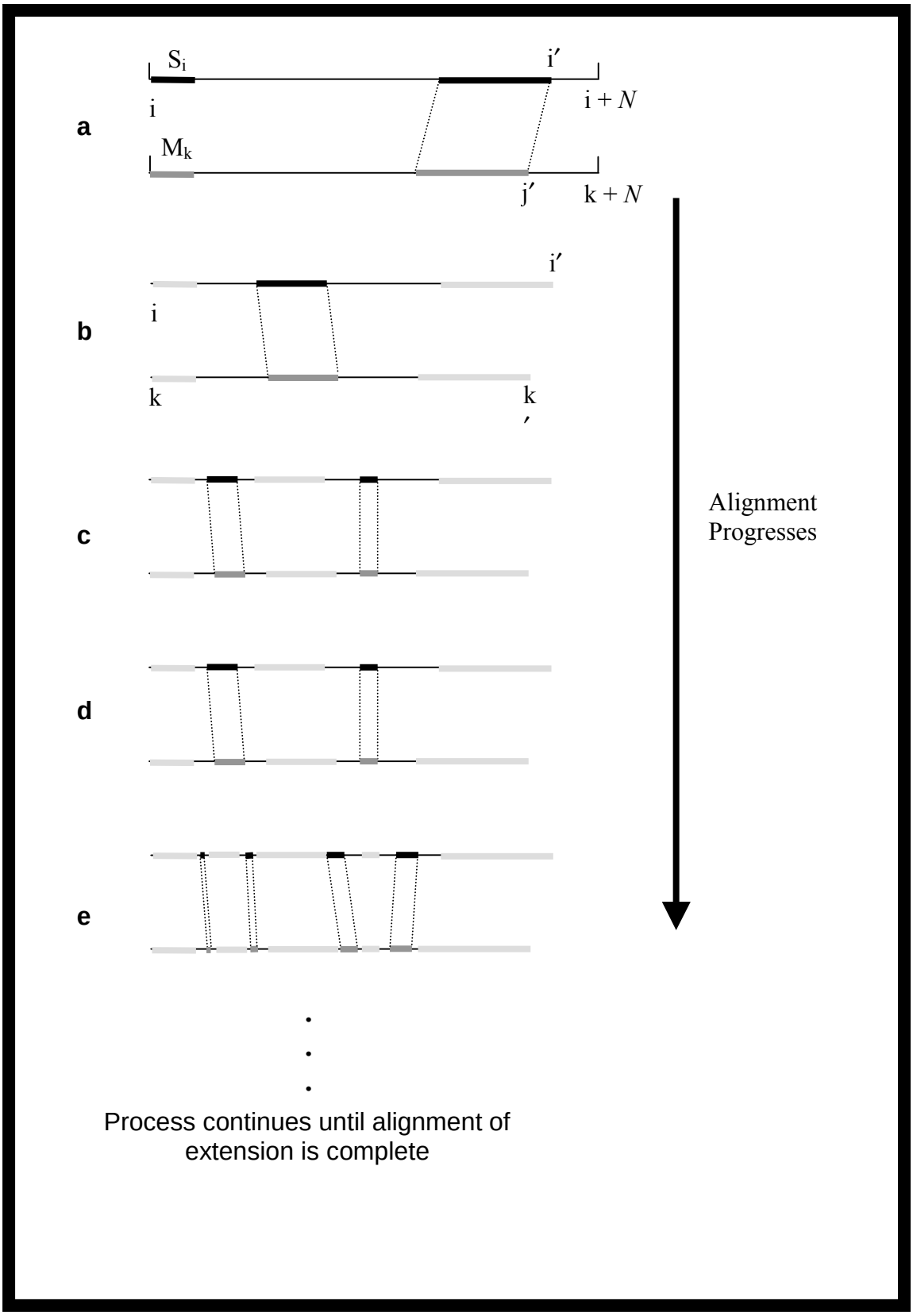
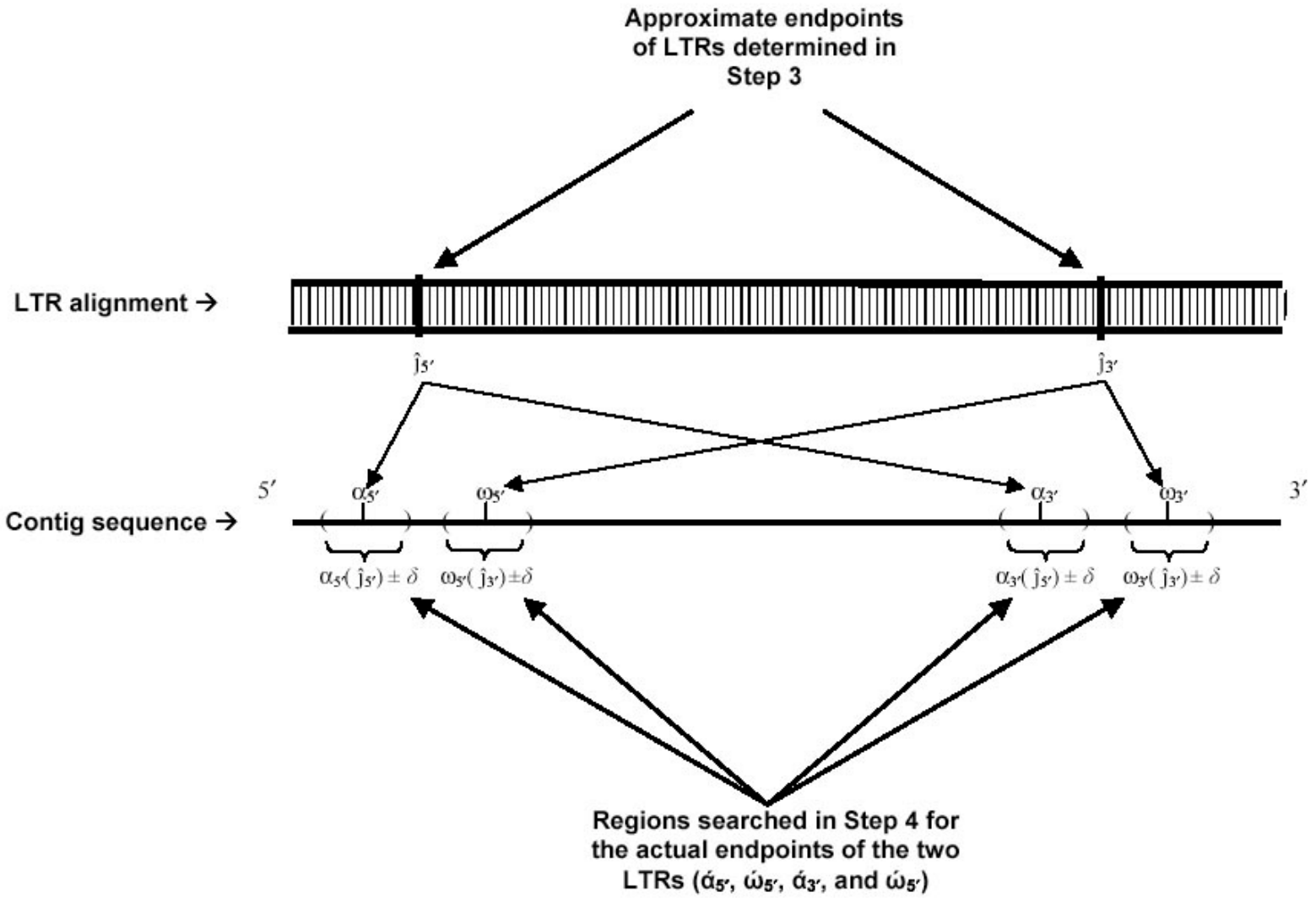

a
S_i
i′
i
i + *N*
M_k
j′
k + *N*
b
i′
i
k
k′
c
d
e
Alignment Progresses
Process continues until alignment of extension is complete

Figure 4

Endpoints of the LTRs are first approximated in the alignment and then analyzed further: In Step 3 (see text) estimated alignment positions ($\hat{j}_{5'}$ and $\hat{j}_{3'}$) of the endpoints are obtained for the LTRs. In Step 4 (see text) these positions are translated into a quartet of estimated endpoints in the contig (i.e., $\alpha_{5'}(\hat{j}_{5'})$, $\omega_{5'}(\hat{j}_{3'})$, $\alpha_{3'}(\hat{j}_{5'})$, and $\omega_{3'}(\hat{j}_{3'})$). LTR_STRUC then scores all possible endpoint combinations in the vicinity of this quartet in order to find the actual endpoints.


Approximate endpoints of LTRs determined in Step 3
LTR alignment →
Contig sequence →
5′
3′
$\hat{j}_{5'}$
$\hat{j}_{3'}$
$\alpha_{5'}$
$\omega_{5'}$
$\alpha_{3'}$
$\omega_{3'}$
$\alpha_{5'}(\hat{j}_{5'}) \pm \delta$
$\omega_{5'}(\hat{j}_{3'}) \pm \delta$
$\alpha_{3'}(\hat{j}_{5'}) \pm \delta$
$\omega_{3'}(\hat{j}_{3'}) \pm \delta$
Regions searched in Step 4 for the actual endpoints of the two LTRs ($\acute{\alpha}_{5'}$, $\acute{\omega}_{5'}$, $\acute{\alpha}_{3'}$, and $\acute{\omega}_{5'}$)

CHAPTER 3

LTR RETROTRANSPOSONS OF *ORYZA SATIVA*[1]

---

[1]McCarthy, E.M., J. Liu, L. Gao, and J.F. McDonald. 2002. Genome Biology 3(10): RESEARCH0053. (reprinted with permission of publisher).

ABSTRACT

Using a new data-mining program, LTR_STRUC, (*LTR* retrotransposon *struc*ture program), we have mined the GenBank rice (*Oryza sativa*) database (GBRD) as well as the more extensive (259 Mb) Monsanto rice data set (MRD) for LTR retrotransposons. Almost two-thirds (37) of the 59 families identified consist of *copia*-like elements, but *gypsy*-like elements outnumber *copia*-like elements by a ratio of approximately two-to-one. At least 17% of the rice genome is comprised of LTR retrotransposons. In addition to the ubiquitous *gypsy*- and *copia*-like classes of LTR retrotransposons, the rice genome contains at least two novel families of unusually small, non-encoding (non-autonomous) LTR retrotransposons. Each of the major clades of rice LTR retrotransposons is more closely related to elements present in other species than to the other clades of rice elements, suggesting that horizontal transfer may have occurred over the evolutionary history of rice LTR retrotransposons. Like LTR retrotransposons in other species with relatively small genomes, many rice LTR retrotransposons are relatively young, indicating a high rate of turnover.

INTRODUCTION

Retrotransposons are mobile genetic elements that make up a large fraction of most eukaryotic genomes. They are particularly abundant in plants, where they are often a principal component of nuclear DNA. In corn, 50-80%, and, in wheat, fully 90% of the genome is made up of retrotransposons [1,2]. In animals this percentage is generally lower than in plants but can still be large. For example, more than 40% of the human genome is now known to be composed of retroelements [3, 4].

All retrotransposons are distinguished by a life cycle involving an RNA intermediate. The RNA genome of a retroelement is copied into a double-stranded DNA molecule by reverse transcriptase and is subsequently integrated into the host's genome. Retrotransposons fall into two main categories, those with long terminal repeats (LTRs), such as retroviruses and LTR retrotransposons, and those that lack such repeats, (*e.g.,* long interspersed nuclear elements or "LINEs").

Our laboratory is in the process of screening the GenBank rice (*Oryza sativa*) database (GBRD) and the Monsanto rice dataset (MRD) for the presence of LTR retrotransposons. We have chosen to scan the rice genome because, as the most important food crop in the world, much of its sequence data is already available. With a haploid content of 430 Mb, the rice genome is the smallest among cultivated cereals [5,6] and only about three times larger than the smallest known genome among angiosperms, that of *Arabidopsis thaliana* (~130 Mb). *O. sativa* has one of the smallest genomes among grasses as a whole [6]. Genomes of other cereals are far larger. For example, the maize (*Zea mays*) genome is 2.5 Gb and that of wheat (*Triticum aestivum*), 16 Gb. The molecular genetic resources for rice are excellent, including detailed physical and genetic

maps, large YAC and BAC libraries, an efficient transformation system, and an extensive collection of expressed sequence tags (ESTs).

We have employed a new search program, LTR_STRUC (*LTR* retrotransposon *struc*ture program; McCarthy and McDonald, submitted), as the initial data-mining tool in our survey. Structural features important to the algorithm on which LTR_STRUC is based include two sites critical to replication, the primer binding site (PBS) and polypurine tract (PPT), as well as the presence of canonical dinucleotides at the ends of each LTR (typically TG and CA). Particularly important are the direct or "target-site" repeats (TSRs). When an LTR retrotransposon inserts itself into host DNA, a short (usually 4-6 bp) segment of host DNA is replicated at the site of insertion. This feature allows LTR_STRUC to make an exact demarcation of the limits of a putative element. Because it searches for retroelements on the basis of their generic structure, LTR_STRUC eliminates much of the bias inherent in BLAST searches based on a known retroelement query. After elements were initially identified using LTR_STRUC, sequence analyses were performed to identify ORFs encoding reverse transcriptase (RT) and other retrotransposon proteins. Subsequent RT sequence alignments were carried out, followed by construction of phylogenetic trees.

Reverse transcriptases (RTs) from elements identified in our survey fall into numerous distinct families, where “family” is defined as a group of elements with RTs having mutual similarity of at least 90% at the amino acid level [7]. In addition, four types of non-autonomous elements discussed here lack RT sequences (*Osr25*, *Osr37/Rire4*, *Osr43*, and *Osr44*), and were classified as distinct families on the basis of their unique structures (see below).

Currently there is no consensus with respect to rice retrotransposon nomenclature. In our method of nomenclature, rice LTR retrotransposons are specified by the appellation *Osr* (*Oryza sativa* retrotransposon). Distinct families are indicated by number (*e.g., Osr1*, *Osr2*, *Osr3,* . . .). There have been four different nomenclatures previously used in reference to rice LTR retrotransposons: *Tos* (transposon *Oryza sativa*) [8], *Rire* (rice retrotransposon) [9] *Rrt* (rice retrotransposon) [10], and *Osr* (*Oryza sativa* retrotransposon) [11]. We have chosen to adopt the *Osr* nomenclature in this study because it is consistent with the systematic logic (indicative of genus and species of host organism) employed in previous genomic studies of LTR retrotransposons and includes the letter "r" to indicate retrotransposon. However, in every case where we use the *Osr* acronym in this paper to refer to a previously named family, we also include will also present any pre-existing name(s) for the family (e.g., *Osr15/Tos12*, *Osr26/Rire2*, etc.).

RESULTS AND DISCUSSION

As is the case for most eukaryotic species analyzed to date, rice LTR retrotransposons fall, for the most part, into two major categories, *gypsy*-like and *copia*-like (two exceptions are discussed below). *Copia*-like elements in the rice genome are usually 5-6 kb in length; however, certain families are composed of longer elements so that the mean length is around 6.2 kb. For example, elements in *Osr7* and *Osr8* are about 9000 bp in length. Results of our study indicate that the target site repeats (TSRs) of all rice LTR retrotransposons are five base pairs long (See Table 1). The dinucleotides terminating the LTRs are similarly invariant: across all families, the 5' nucleotide pair is consistently TG, and the 3' end, consistently CA (except for a few mutated copies). In

the rice genome, normal *gypsy*-like elements (*i.e.*, those that lack a deletion or insertion) are typically in the 10- to 13-kb range, but some do bear large insertions or internal deletions. Their mean length of 11.7 kb is larger than that of typical *gypsy*-like elements in other species, which are usually in the range of 7-8 kb [*e.g.,* 7, 12]. The reason for this larger mean length of *O. sativa* LTR retrotransposons is presently unknown. Duplication of retroelement sequences during the process of reverse transcription have been previously observed in mammalian systems [e.g., 13] and nested insertions of transposons into LTR retrotransposons are not uncommon in plants [14]. However, none of the full-length LTR retrotransposons reported here have a sub-structure consistent with nested LTR retrotransposon insertions. For example, none of the elements we have examined encode more than one region of RT homology and none of the elements contain nested pairs of putative LTRs. Of course, we cannot eliminate the possibility that the larger size of *O. sativa gypsy*-like elements is, at least in part, due to insertions of unrecognized elements or ancient insertions of known elements that can no longer be recognized. Whatever, the reason for the exceptional size of *O. sativa gypsy*-like elements, it apparently does not inhibit function for sequence analysis (see below) indicates that the majority of these elements have transposed in the recent evolutionary past. *Gypsy*-like elements in *O. sativa* also have larger LTRs than *copia*-like elements, many with lengths in excess of 3000 bp (mean ~1000 bp), whereas the typical *copia*-like LTR is around 500 base pairs long.

Our survey has identified numerous LTR retrotransposon families that have not been described previously. These findings demonstrate that at least 59 distinct LTR retrotransposon families exist in the rice genome. This result compares with an earlier

family estimate of 32 based on screening of genomic libraries [8]. *Copia*-like elements are less numerous than *gypsy-like* elements in the rice genome, but they still comprise more than half the families, a total of 37. In addition to 57 families of *copia*- and *gypsy*-like elements, we have identified two families of LTR retrotransposons (*Osr43* and *Osr44*) that show no significant sequence similarity to any known transposon.

For purposes of this analysis, a "full-length element" is defined as one that has two complete and recognizable LTRs**.** Any other LTR retrotransposon sequence is here defined as a "fragment." The results of our survey of the GBRD and MRD suggest there are on the order of 450 full-length *copia*-like elements in the entire rice genome. We found full-length *copia*-like elements both with and without RT domains. We estimate the total copy number (including fragmentary copies) at 3500, or about 3% of the genome. BLAST searches with representative LTR queries from each of the rice LTR-retrotransposon families against the MDR indicate that *gypsy*-like elements are twice as common (total copy number of ~7000, about 1400 of them full-length). Previous estimates of this ratio have been somewhat higher [15]. Due in part to their large LTRs, *gypsy*-like elements in rice are twice as long as *copia*-like elements (11.7 kb versus 6.2 kb) and so make up a proportionately larger fraction of the genome (~14%). That is, a total of about 17% of the genome is composed of LTR retrotransposon sequences. This estimate exceeds those of previous workers [8, 15, 16, 17]. For example, using a variety of RT probes, Wang *et al.* [15] estimated that ~100 copies of *copia*-like elements are present in the entire haploid genome. This estimate did not discriminate between full-length and fragmentary copies. From our examination of the searchable portion of the GBRD alone (which represented at the time approximately 10% of the rice genome), we

have identified the actual sequences for 46 separate full-length *copia*-like elements. This implies that the number of full-length *copia*-like elements in the whole genome should be about ten times higher, that is, around 450-500 elements. In an analysis of 340 kb around the *Adh1-Adh2* region of the rice genome, Tarchini *et al.* [15] reported that 14.4 % of this region consisted of LTR retrotransposons. This value is in reasonably good agreement with our estimate of about 17%. *Mao et al.* [17] give a lower figure (9.3%) but we suspect our higher figure is more accurate because their study sought homology to known retrotransposon sequences and such homology would be undetectable for the many new families of retrotransposons presented here. Similarly, they give a higher ratio of *gypsy*- to *copia*-like elements, but they may not have been aware that gypsy-like elements are significantly larger in rice, which would inflate their estimate of this ratio.

The previous low estimates of copy number given for rice LTR retrotransposons are likely attributable to three factors: 1) these earlier studies used an incomplete set of RTs as probes for hybridization (or as queries for BLAST). For example, *Osr8* a high copy copia-like family was not recognized in previous studies; 2) a number of rice LTR retrotransposons lack an RT ORF and would thus go undetected in studies using RT probes. In particular, no member of families *Osr25* and *Osr37/Rire4* seem to have an RT (yet these two families have a total copy number of around 900 elements); and 3) Data-mining with LTR_STRUC (See METHODS) allows a higher degree of assurance that the putative RTs detected in the survey actually are RTs because it places putative polyproteins in the context of a canonical retroviral structure. Such is not the immediate result of a simple BLAST with an RT query. Our estimate that LTR retrotransposons make up 17% of the rice genome is conservative inasmuch as our study was based

primarily on euchromatic sequences and did not include elements present within the traditionally retrotransposon-rich heterochromatin [16,19]. Thus, our results bring the rice genome closer to the LTR retrotransposon densities reported for other cereals.

Percent LTR-LTR nucleotide identity. Due to the replication process characteristic of LTR retrotransposons, the LTRs of a given retroelement are sequentially identical at the time the element inserts into the host genome [20]. Thereafter, as an element accumulates mutations, its LTRs become increasing different from each other as substitutions specific for each of the two LTRs increase in number. The level of nucleotide identity seen between LTRs of a particular element, usually referred to as percent LTR-LTR nucleotide identity (%LNI), can be used in determining the relative ages of LTR retrotransposon families [7]. In rice, comparison of the two LTRs of the same element often showed the presence of a 10- to 30-bp-long regional duplication present in one LTR but not the other. In calculating %LNI, we have considered such duplications as single mutation events.

Since the neutral nucleotide substitution rate has yet to be computed for rice, we cannot presently equate %LNI with a divergence time in years. However, the generally low level of sequence divergence between flanking LTRs of rice LTR retrotransposons (1.7%) indicates that most of the euchromatic full-length LTR retrotransposons in rice are relatively young, although significantly older elements were also identified. The seeming preponderance of young full-length LTR retrotransposons in the euchromatin of rice is similar to what has been previously reported in yeast [21, 22] *Caenorhabditis elegans* [7],

Arabidopsis thaliana [23] and *Drosophila melanogaster* [12]. This contrasts with what has been observed in *Zea mays* [14] and humans [24].

COPIA-LIKE FAMILIES

To date, twenty-three families of copia-like elements have been reported for rice [8, 9, 10, 11, 21, 25, 26]. Several have been described under more than one name. For example, the amino acid sequence given for *Tos4* in Hirochika *et al.* 1996 [25] is the same as that given for *Tos1* in the GBD (acc# S22455) so they are really the same. *Rire5* described by Kumekawa *et al.* 1999 [27] is the same family as *Tos14* previously described by Hirochika *et al.* 1996 [25]. The equivalence between *Tos14* and *Rire5* became evident when we found the LTR sequence reported by Kumekawa *et al.* in elements that also contained the RT sequence given by Hirochika for *Tos14*. In our survey of the GBD and MRDB, we have identified an additional 16 copia-like families that have not been described by previous workers. In addition, exemplars for each of the previously identified families were found (except in the case of certain families that exist at such low copy numbers that no full-length element exists in the GBD or MRDB).

The largest copia-like family. One of the most interesting new finds in our survey was *Osr8,* one of the oldest families of LTR retrotransposons in the rice genome. Based on a survey of the available portion of the GBRD and MRD, we estimate the copy number of *Osr8* to be ~1100 (more than any other *copia*-like family). *Osr8* elements exist far more frequently as fragments (ratio of 10:1) and they display relatively low levels of %LNI in their full-length copies (mean %LNI for the five full-length *Osr8* elements present in the

GBRD is 97.2%). The RT of *Osr8* is 60% similar to an unnamed polyprotein in *Z. mays* (acc# AAD20307). A closely related family, *Osr10* has two full-length copies in the GBRD but scans of the MRD suggest this element, also previously unrecognized, has the third highest copy number (~400) among *copia*-like elements. Outside rice, the RT of *Osr10* shows highest similarity (~65%) to that of the maize retrotransposon *Opie-2* (acc# T04112). The broader clade that includes *Osr7*, *Osr8*, *Osr9*, and *Osr10* is closely related to *Endovir1-1* (acc# AAG52949) of *Arabidopsis* (Figure 1; Table 3). These elements are also related (~60% similar) to maize's *PREM-2* as well as to tomato's *ToRTL1*. Both *Osr7* and *Osr9* are present in very low copy number (one full-length and a few fragments in the GBRD).

Osr14/Tos1/Tos4; Osr15/Tos12; Osr53/Tos18. Although it is present at only about one-quarter the copy number of *Osr8,* the unrelated *Osr14*/*Tos1*/*Tos4* is also composed primarily of highly fragmented elements. Those that are full-length have low %LNI (family mean 97.6%). Thus, *Osr14/Tos1/Tos4* and *Osr8* seem to be of similar age and to have followed a similar evolutionary pattern, albeit with less intense amplification in the case of *Osr14Tos1/Tos4*. *Osr14Tos1/Tos4*, *Osr15/Tos12*, and *Osr53/Tos18* form a well-defined clade and are more closely related to *Ta1-2* (acc# S23315) of *Arabidopsis* than to any other rice retroelement family outside their clade (Figure 1; Table 3). *Osr15/Tos12* and *Osr53* are only just sufficiently different to constitute distinct families.

A quartet of closely-allied families. *Osr1/Tos14/Rire5*, *Osr13*/*Tos5*, *Osr51/Tos15*, and *Osr52*/*Tos16* have been described as distinct families but, inasmuch as their RTs are all 85% similar to each other these groups are only marginally distinct. Searches of GenBank show that elements in this group are much more closely related to (75-80% at the amino acid level) to maize retrotransposon *Fourf* (acc# AAK73108) than to any rice LTR retrotransposon outside their clade. If the elements belonging to this group were considered to be a single family, it would be almost as large (~900 elements) as *Osr8*. In the GBRD the majority of these elements are fragmentary, but the estimated copy number of full-length elements in the rice genome for this quartet still exceeds 100.

A Hopscotch-like clade of fragmented elements. *Osr18*, *Osr19*, *Osr20*, *Osr22*, *Osr23, Osr24, Osr45/Tos7*, and *Osr46*/*Tos8* form a clade of low-copy families composed primarily of fragmentary copies. Our results suggest each of these families have copy numbers in the range of 50-100 elements. Members of this clade are closely related to maize's *Hopscotch* element (acc# T04112) (Figure 1; Table 3).

Low-copy copia-like families. *Osr2* and *Osr12* are low-copy families and are represented in the GBRD by two and three copies respectively, all of which are full-length (although one copy of *Osr12* contains a large internal deletion), suggesting that these elements may have recently invaded the rice genome. The high level of LTR nucleotide identity (99%+) seen in these elements is consistent with this recent invasion hypothesis. Members of *Osr12* and *Osr2* are potentially active because they have large, intact polyprotein ORFs, usually in excess of 1000 amino acids. All three *Osr12*

elements detected in the GBRD are on chromosome 10. Similarly, both *Osr2* elements are inserted within 50kb of each other on chromosome 4. Nonetheless, these two families are not closely related (their reverse transcriptase sequences are only ~50 % similar at the amino acid level). *Osr12* RTs differ from those of all other rice *copia*-like elements by 50%. And yet, RT sequences of elements in *Osr12* are 60% similar to certain elements in the maize genome (*Zmr1*, acc# S27768; *mzecopia*, acc# M94481.1).

One full-length, and one fragmented copy of *Osr6* are present in the GBRD. *Osr5* is slightly more common that *Osr6*, to which it is most closely related, but it is currently represented in the GBRD by only a single full-length copy and a few fragments. *Osr5* is 60% similar to the tobacco retrotransposon *Tnt1-94* at the amino acid level (RT comparison). *Osr4* is another low-copy family. It has several fragmented representatives in the GBRD, and is probably somewhat older than *Osr12* and *Osr2*, but it has only 3 full-length copies in the GBRD, *Osr4* elements have an exceptionally large polyprotein ORF (~1600 amino acids). The RT of *Osr4* shows 50% similarity to that of retroelements in the *Arabidopsis* genome (*e.g.,* acc#s BAB01972, NP_175303).

Although the RT of *Osr3* was detected during our survey, elements in this family are fragments with ill-defined LTRs. TBLASTN reveals the RT of *Osr3* to be the single representative of its type in the GBRD. Both *Osr3* and the equally aberrant *Osr21*/*Tos17* differ from those of other *copia*-like elements found in our study by about 55%. *Osr11*/*Rire1* is a low-copy family closely related (75% similarity) to a retroelement in the *Arabidopsis* genome (*Atr-2*, acc# T01860). Two other closely related families are *Osr16*/*Tos6* and *Osr17*, both of which are both similar to *Sto-4* (acc# T17429) of maize (Figure 1; Table 3). Eleven additional low-copy families identified by earlier workers are

Osr47/*Tos9*, *Osr48/Tos10*, *Osr49/Tos11*, *Osr50* /*Tos13*, *Osr54/Tos19*, *Osr55*/*Tos20*, *Osr57*/*Rtr3*, *Osr58*/*Rrt5*, and *Osr59*/*Rrt8.* Source references for each of these nine families are given in Table 2.

GYPSY-LIKE FAMILIES

Osr27/Rire9 [28] is the third largest family in the rice genome, with an estimated copy number of 900 elements, mostly full-length (Li *et al.* [28] estimated the copy number of this family at 1,600). The typical *Osr27/Rire9* element is quite large (~12.8 kb total length). Having intact polyprotein ORFs and high mean %LNI (99%), these elements likely are, or recently have been, actively transposing. Yet, the presence of a few members of this family that are more mutated (short ORFs, low LTR-LTR nucleotide identity) suggests that this may also be an ancient family. Two other families, *Osr40* and *Osr41,* are also members of the same clade as *Osr27/Rire9, Osr25* and *Osr26/Rire2* (Osr25 and *Osr26/Rire2* are discussed below), but both have RTs that are about 30% different from those of *Osr26/Rire2* and *Osr27/Rire9*. Neither *Osr40* nor *Osr41* have been previously identified, but with approximate copy numbers of 600 and 300, respectively, these are both large families. The RTs of members of this clade show about 60% similarity to that of *Retrosor1* (*Sorghum bicolor;* acc# AAD19359).

With approximately 1500 elements, *Osr30* constitutes 14% of all LTR retrotransposons in the rice genome. Although *Osr30* is the largest family of LTR retroelements in the genome, it has not been previously named. These elements are slightly larger (~13.1 kb) than those of *Osr27/Rire9.* A higher proportion of fragmented copies and lower level of LTR-LTR nucleotide identity suggest that *Osr30* is older than

Osr27/Rire9. *Osr29*, which is closely allied to *Osr30*, is also a large family with more than 500 member elements. Taken together, the elements of the *Osr29* and *Osr30* clade are unusual, because they are as closely related to other major rice clades as they are to any elements outside rice. *Osr28* is a low-copy family that is most closely related to Osr29 and *Osr*30 (Figure 2).

Two other large *gypsy*-like families are *Osr33*/*Rire8* [27] and *Osr34*. These two families each have copy numbers of approximately 500. Two low-copy families belonging to the same clade are *Osr32* and *Osr56/Rire3* [29] (Figure 2). Members of these families have large LTRs, typically in the range of 3000-3500 bp. RTs of families in this clade show high sequence similarity to an LTR retrotransposon in pineapple (~70% to *Acr-1*; acc# CAA73042) and to one in *Sorghum bicolor* (~77% to *Retrosor3*acc# AAD221153) (Figure 2).

Low-copy *gypsy*-like elements. *Osr31/Rire7* is an aberrant low-copy family that is much more closely related (77% similarity) to an *Arabidopsis* element, *Atr-4* (see Table 3), than to any other LTR retroelement families in the rice genome (Figure 2). In the clade of five low-copy families, composed of *Osr35, Osr36, Osr38, Osr39,* and *Osr42,* an RT was found in the GBRD for only two families, *Osr35* and *Osr36*. The other elements were identified in scans of the MRD and their full sequences have since been submitted to GenBank (for accession numbers, see Table 1). This clade is closely related to *Arabidopsis* element *Atr-5* (Figure 2; Table 3).

Families of non-autonomous elements. Members of family *Osr25* are all internally deleted and thus non-autonomous (mean length 4.3 kb). Although *Osr25* elements have typical LTRs, PBS, and PPT, the inter-LTR region contains only non-coding, repetitive DNA. The LTRs of *Osr25* display 65-70% sequence similarity to the autonomous elements of the *gypsy*-like family *Osr26*/*Rire2*. Elements with LTRs having such a high degree of similarity are usually considered members of the same family. Nevertheless, because 1) members of *Osr26/Rire2* have the usual coding structure typical of other *gypsy*-like elements (while *Osr25* elements entirely lack typical retroviral genes) and 2) members of these two families fall into two sharply distinct, non-overlapping clades, we report these two types of elements as separate families. Estimates based on scans of the MRD and the GBRD suggest that the rice genome contains about 500 copies each of *Osr25* and *Osr26/Rire2*. *Osr25* and *Osr26/Rire2* display 98.9 and 97.9% LNI respectively.

Osr37/Rire4 is also aberrant compared to other rice LTR retrotransposon families. The typical element in this family is 4.4 kb long, about the same length as *Osr25* elements. Members of *Osr37/Rire4* usually carry a large ORF (up to 600 amino acids long) just upstream of the 3' LTR. This ORF shows no significant similarity to any known RT sequence. Up to the present in the GBRD, where these ORFs are generally identified simply as hypothetical proteins, the large ORF of *Osr37/Rire4* seems not to have been recognized as a retroviral gene. This ORF may serve an integrase function since BLAST searches show it has low homology to a putative integrase in *A. thaliana* (28%; acc# AC005171). There are about 600 copies of *Osr37/Rire4* in the entire rice genome.

In addition to the foregoing copia- and gypsy-like families, our scans identified two families, Osr43 and Osr44, of small elements (overall length < 2000 bp). With LTRs only 148-bp-long and an overall length of 1207 bp, Osr44 elements are especially small. Members of Osr43 and Osr44 are unique because, although they possess all of the canonical LTR-retrotransposon structural features (LTRs, PBS, PPT, and TSRs), they are internally deleted and either completely lack or encode only very small ORFs with no similarity to any known protein. Both families contain on the order of 100 copies genome-wide.

CONCLUSIONS

Rice LTR retrotransposons are a significant component of the rice genome. We estimate that LTR retrotransposons constitute at least 17% of the *O. sativa* genome. Although this value is lower than the estimated percentage of LTR retrotransposons in the genomes of other cereal plants [2, 14], it is more than 10-fold greater than the estimated percentage of LTR retrotransposons in *Arabidopsis thaliana*, a species with a genome one-third the size of the rice genome [23]. This disproportionate increase in the percentage of LTR retrotransposons as a function of genome size is consistent with the view that genome size variability in plants is often heavily dependent on variation in LTR retrotransposon content [29, 30].

We have determined that individual full-length LTR-retrotransposons present in the sequenced euchromatic regions of the rice genome are all relatively young, displaying, on average, greater than 98% sequence identity between their LTRs. Comparative genomic studies of LTR retrotransposons in both plants and animals have

revealed that species with smaller genomes [7, 12, 21-23] do not harbor older families of LTR retrotransposons, as do species with larger genomes [14, 24]. It has been hypothesized that the rate of turnover of retroelements may be higher in small genomes due to the presence of less effective epigenetic silencing mechanisms [12]. It remains to be determined whether or not this hypothesis is an adequate explanation of the apparent lack of older full-length LTR retrotransposons in the euchromatic portion of the rice genome.

In general, the major clades of rice LTR retrotransposons are more closely related to elements present in other species than to the other clades of rice elements, suggests that horizontal transfer may have occurred over the evolutionary history of rice LTR retrotransposons. Further analysis is required to definitively test the horizontal transfer hypothesis.

The newly developed search algorithm (LTR_STRUC) we have employed in this study to initially identify LTR retrotransposons in the rice genome is not dependent upon sequence homology, as are standard search methods (*e.g.,* BLAST). As a consequence, we identified several previously unreported families of rice LTR retrotransposons consisting of non-encoding and, in some cases, repeating sequence motifs. LTR retrotransposons of similar structure have recently been identified within the genomes of both mono- and dicotyledonous plants [31]. Preliminary evidence suggests that these elements may play a significant role in restructuring plant genomes over evolutionary time [31].

METHODS

Automated characterization of LTR retrotransposons. LTR_STRUC, identifies new LTR retrotransposons based on the presence of characteristic retroelement features (McCarthy and McDonald, submitted). It scans nucleotide sequence data for putative LTR pairs, aligns the putative pairs, and scores them on the basis of the presence/absence of expected motifs such as TSRs, canonical dinucleotides, PBS, PPT, etc. When a given pair receives a score above a (user-specified) cutoff, an output record is generated that specifies salient information about the putative element, such as the length of the transposon and its LTRs, its position within the contig, an alignment of its LTRs, the nucleotide sequence of the transposon, its LTRs and target site repeats, as well as a file listing all ORFs. In our study, once putative elements were identified, sequence analysis was carried out on the individual output files to identify those that described actual LTR retrotransposons. Additional elements were identified by BLAST searches using elements located by LTR_STRUC as queries.

Data sets scanned. Initial scans with LTR_STRUC were conducted on a data set consisting of the 29.8 Mb of *O. sativa* BAC-derived sequence data available in the GenBank database at the time of the initial scan (Dec. 2000). This data set (TDS) was obtained from the TIGR web site [32]. Subsequently, LTR_STRUC was used to scan the non-redundant Monsanto rice dataset (MRD), a product of the Monsanto Rice Genome Sequencing Project. The MRD is based on an initial dataset of 3391 BACs distributed across the genome of *O. sativa cv*. Nipponbare — the same cultivar used by the International Rice Genome Sequencing Project. Removal of contaminants and

redundancies from this initial dataset produced the MRD (consisting of 52,202 contigs, totaling 259 Mb of the 430-Mb rice genome). More recently, in an effort to determine the relative copy numbers of the various families and identify additional elements not picked up in our initial survey with LTR_STRUC, we have used representative sequences from each retrotransposon family identified in this study as queries to conduct BLAST searches against both the MRD and the GenBank rice database (GBRD) Thus, the results reported here constitute a reasonably unbiased survey of LTR-retrotransposon diversity in rice. Both the MRD and GBRD are heavily weighted toward euchromatic sequences. The amount of data scanned was significantly less than the total amount of nucleotide sequence contained in the MRD and GBRD. Much of the MRD (~36%) is composed of contigs that are less than 10 kb in length and are therefore of limited utility for the LTR_STRUC program, which finds only full-length elements (rice *gypsy*-like elements are typically longer than 10kb and are not entirely contained in such short contigs). In the case of the GBRD, the amount of rice nucleotide sequence available for search was less than one-third of the 174 Mb released to the public (due to a 15% redundancy, the GBRD sequences amounted to a total of only about 150 Mb, of which only some 50 Mb were actually available for BLAST search because most of these sequences were in the process of being "finished"). RT sequences were identified according to previously described criteria [33,34].

Multiple sequence alignments and phylogenetic analyses. The RT domains of the *Osr* elements were aligned with previously reported RT sequences (Table 3). The `ClustalW` analysis [35] extension to `MacVector 7.0` was used to generate two amino acid

alignments, one for *gypsy*-like, and one for *copia*-like elements. Draw N-J Tree and Bootstrap N-J commands of ClustalW were then used to generate non-bootstrapped and bootstrapped trees, respectively.

REFERENCES

1. SanMiguel P, Tikhanov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL: Nested retrotransposons in the intergenic regions of the maize genome. Science 1996, 274:765-768.

2. Flavell RB: Repetitive DNA and chromosome evolution in plants. Phil. Trans. R. Soc. Lond. B. Biol. Sci. 1986, 13: 335-340.

3. Yoder JA, Walsh CP, Bestor TH: Cytosine methylation and the ecology of intragenomic parasites. Trends Genet.1997, 13:335-340.

4. Smit AF: Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev.1999, 9: 657-663.

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

6. Plant DNA C-Values Database [http://www.rbgkew.org.uk/cval/searchguide.html]

7. Bowen N, McDonald JF: Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res. 2000, 11:1527-1540.

8. Hirochika H, Fukuchi A, Kikuchi F: Retrotransposon families in rice. Mol. Gen. Genet. 1992, 233: 209-216.

9. Nakajima R, Noma K, Ohtsubo H, Ohtsubo E: Identification and characterization of two tandem repeat sequences (TrsB and TrsC) and a retrotransposon (Rire1) as genome-general sequences in rice. Genes Genet. Syst. 1996, 71: 373-82.

10. Wang S: Direct submissions to EMBL: Rtr3 (acc# T03666), Rrt5 (acc# T03669), and Rrt8 (acc# T03671).

11. Jwa N: GenBank, Direct Submission: Osr1 (acc# AB046118).

12. Bowen N, McDonald JF: Genomic analysis of Caenorhabditis elegans reveals ancient families of retroviral-like elements. Genome Res. 1999, 9:924-935.

13. Burns DP, Temin, HM: High rates of frameshift mutations within homo-oligomeric runs during a single cycle of retroviral replication. J. Virol. 1994, 68: 4196-4203.

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

15. Turcotte K, Srinivasan S, Bureau T: Survey of transposable elements from rice genomic sequences. Plant J. 2001, 25: 169-180.

16. Wang SP, Liu N, Peng KM, Zhang, QF: The distribution and copy number of copia-like retrotransposons in rice (Oryza sativa L.) and their implications in the organization and evolution of the rice genome. PNAS, USA 1999, 96: 6824-6828.

17. Mao L, Wood, TC, Yu,Y, Budiman MA, Tomkins J, Woo S, Sasinowski M, Presting G, Frisch D, Goff S, Dean RA, Wing RA: Rice Transposable Elements: A Survey of 73,000 Sequence-Tagged-Connectors Genome Res. 2000, 10: 982-990.

18. Tarchini R, Biddle P, Wineland R, Tingey S, Rafalski A: The complete sequence of 340 kb of DNA around the rice Adh1-Adh2 region reveals interrupted co-linearity with maize chromosome 4. Plant Cell 2000, 12: 381-392.

19. Heslop-Harrison JS, Brandes A, Taketa S, Schmidt T, Vershinin AV, Alkhimova EG, Karum A, Doudrick RL, Scwarzacher T, Katsiotis A, Kubis S, Kumar A, Pearce SR, Flavell AJ, Harrison GE: The chromosomal distributions of Ty1-copia group retrotransposable elements in higher plants and their implications for genome evolution. Genetica 1997, 100: 197-204.

20. Boeke JD, Stoye JP: Retrotransposons, endogenous retroviruses and the evolution of retroviruses. In Retroviruses, Edited by Coffin J, Hughes S, Varmus H, Cold Spring Harbor, NY: Cold Springs Harbor Laboratory Press, 1997, 343-435.

21. Jordan IK, McDonald JF: Tempo and mode of evolution in Saccharomyces cerevisiae genome. Genetics 1999, 151: 1341-1351.

22. Promislow DE, Jordan, IK, McDonald JF: Genomic demography: A life history analysis of transposable element evolution. Proc. Roy. Soc. Lon. B Biol. 1999, 266:155-1560.

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

24. Tristem M: Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J. Virol. 2000, 74: 3715-3730.

25. Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M: Retrotransposons of rice involved in mutations induced by tissue culture. PNAS, USA 1996, 93:7783-8.

26. Noma K, Nakajima R, Ohtsubo H, Ohtsubo E: Rire1, a retrotransposon from wild rice *Oryza australiensis*. Genes Genet Syst. 1997, 72: 131-40.

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

28. Li ZY, Chen SY, Zheng XW, Zhu LH: Identification and chromosomal localization of a transcriptionally active retrotransposon of Ty3-gypsy type in rice. Genome 2000, 43: 404-8.

29. Kumar A, Bennetzen JL: Plant retrotransposons. Ann. Rev. Genet. 1999, 33: 497-532.

30. Wendel JF, Wessler SR: Retrotransposon-mediated genome evolution on a local ecological scale. PNAS, USA 2000, 97: 6250-6252.

31. Witte CP, Le QH, Bureau T, Kumar A: Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. PNAS, USA 2001, 98:13778-83.

32. TIGR Web Site [http://www.tigr.org/tdb/ogi/].

33. Xiong Y, Eickbush TH: Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol. Biol. Evol.1988, 5: 675-690.

34. Xiong Y, Eickbush TH: Origin and evolution of retroelements based upon their reverse-transcriptase sequences. EMBO 1990, 9: 3353-3362.

35. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: The CLUSTAL X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 1997, 25: 4876-4882.

Table 1. Summary of Rice LTR Retrotransposons Characterized in this Study

Family	Pre-existing name(s)	Accession # of Exemplar	Location	Ch.	LTR length (bp)	Inserted Element length	TSR	%LNI	Appr. Copy Num.
Osr1	*Tos14/Rire15*	AC023240	100410-106807	10	965	6398	AGTCC	98.1	250
Osr2		AL442110	95121-100070	4	267	4950	ATATT	98.5	<50
Osr3		AF458765	51- 5250	?	146	5200	CATTC	99.3	50-100
Osr4		AB026295	160208-165872	6	350	5665	GTTAC	98.9	<50
Osr5		AC021891	56044-62135	X	477	6092	TACAG	96.2	<50
Osr6		AP001366	57569 -62773	1	440	5205	ACCTG	99.8	<50
Osr7		AP002538	44996-53915	1	1608	8920	AGTTT	98.8	<50
Osr8		AC021891	65191-74406	X	1220	9216	TAAAT	97.2	1100
Osr9[a]		AP000969	25869 -28634	1	ND	ND	ND	ND	50-100
Osr10[a]		AC069324	137920 -139740	10	ND	ND	ND	ND	400
Osr11[a]	*Rire1*	AP003853	96975-98088	1	ND	ND	ND	ND	<50
Osr12		AC073166	104289-109024	10	221	4736	AGAAG	99.7	<50
Osr13	*Tos5*	AC073405	72924-79364	5	968	6441	TATGT	99.6	650
Osr14	*Tos1/Tos4*	AC069324	8821-17191	10	319	8371	CTCCC	97.6	350
Osr15	*Tos12*	AP002867	127118-132180	1	262	5062	GCTTC	94.5	250
Osr16	*Tos6*	AP002845	42644-49551	1	300	6908	TGCTT	97.9	<50
Osr17		AC018727	102539-96583	10	501	5957	TCATC	99.6	50-100
Osr18		AC068654	23423-25036	X	ND	ND	ND	ND	<50
Osr19		AC069300	73013-77731	10	205	4719	GGGAC	99.5	50-100
Osr20		AC084406	8749-14200	3	286	5452	TTATA	97.9	50-100
Osr21[a]	*Tos17*	AC087545	81711-84269	10	ND	ND	ND	ND	50-100
Osr22		AC074283	24546- 19810	10	191	4647	GAACC	97.9	50-100
Osr23		AP002843	144255-139782	1	209	4774	AGGAT	99.5	50-100
Osr24		AC016781	25997-30858	ND	221	4852	CCGAG	98.6	<50
Osr25		AP001278	28729 35569	1	417	6841	TCGAG	98.9	500[d]
Osr26	*Rire2*	AP001111	59274-70587	5	440	11314	GATAT	97.9	500
Osr27	*Rire9*	AP000399	75139-88038	6	1087	12900	AATAT	99.0	900
Osr28		AP002539	139654-121650	1	2195	18005	GTTAT	99.0	<50
Osr29		AP002747	78609- 87615	1	656	9007	GGAAC	96.0	550
Osr30		AC078891	52683- 65684	10	1507	13002	ACTTT	97.2	1500
Osr31	*Rire7*	AP003054	102778-110180	1	787	7403	AAACC	99.9	<50
Osr32[a]		AP002820	111559-12278	1	ND	ND	ND	ND	50-100
Osr33	*Rire8*	AP002864	35539- 47557	6	3009	12009	CACAC	99.1	550
Osr34		AF111709	25889-38685	5	3292	12797	AGAAA	99.4	450
Osr35		AC068924	94924- 100611	10	423	5688	CTAAT	98.3	<50
Osr36		AP001551	59722-64876	1	319	5155	GGTCA	98.4	<50
Osr37	*Rire4?*	AC068654	2534-6969	X	794	4436	CTTGA	98.9	600
Osr38[b]		AF458766	31-5535	?	332	5525	TGAGG	96.2	<50
Osr39		AF458767	51-5267	?	368	5217	CAAAG	97.6	<50
Osr40		AC020666	65731- 77151	10	564	11421	ACATG	98.3	600
Osr41		AP003631	27347-43001	1	518	15655	GGTTC	97.7	300
Osr42		AF458768	51-5655	?	358	5605	ATGTC	99.9	<50
Osr43		AP000815	77117-78910	1	291	1794	CTGAT	98.6	<50
Osr44		AP000364	41541-42747	8	148	1207	AACAA	99.9	<50

a. Location given is for an example RT in the GBRD (no full-length element was identified for this family)
b. Since a full-length element is known in the MRD, the TSR and lengths of the LTR and element (columns 5-7) are taken from an element in the MRD while the location (if given) in columns 2-4 refer to an RT in the GBRD.
c. Percentages based on number of hits using a sample LTR from each family as query to search the MRD
d. Jiang and Wessler (in preparation) suggest that if pericentric DNA (which is largely heterochromatic is taken into account, *Osr25* elements exist at a higher copy number (*i.e.,* ~1000 copies in the entire genome) than our survey, based largely on euchromatic sequences, would suggest.

Ch: Chromosome number. ND: not determined.

Table 2. Previously named low-copy families for which a full-length exemplar has not been presented in this paper:

Family	Pre-existing family name	Accession number (or source) of sequence
Osr45	*Tos7*	T03709
Osr46	*Tos8*	T03704
Osr47	*Tos9*	T03705
Osr48	*Tos10*	T03706
Osr49	*Tos11*	T03707
Osr50	*Tos13*	Hirochika *et al.* 1999
Osr51	*Tos15*	T03711
Osr52	*Tos16*	T03712
Osr53	*Tos18*	T03716
Osr54	*Tos19*	T03721
Osr55	*Tos20*	T03723
Osr56	*Rire3*	Kumekawa *et al.* 1999
Osr57	*Rtr3*	T03666
Osr58	*Rrt5*	T03669
Osr59	*Rrt8*	T03671

Table 3 Non-rice RTs used in phylogenies

Name of retrotransposon	Accession Number	Host organism
Opie-2	**T04112**	***Zea***
Hopscotch	**T02087**	***Zea***
Fourf	**AAK73108**	***Zea***
Sto-4	**T17429**	***Zea***
Zmr-1*	**S27768**	***Zea***
Endovir1-1	**AAG52949**	***Arabidopsis***
Ta1-2	**S23315**	***Arabidopsis***
Atr-1*	**NP_175303**	***Arabidopsis***
Atr-2*	**T01860**	***Arabidopsis***
Atr-3*	**NP_178752**	***Arabidopsis***
Atr-4*	**NP_174802.1**	***Arabidopsis***
Atr-5*	**AAF13073.1**	***Arabidopsis***
Atr-6*	**NP_179047**	***Arabidopsis***
Retrosor1	**AAD19359**	***Sorghum***
Retrosor3	**AAD22153**	***Sorghum***
Daniela	**AF326781**[†]	***Triticum***
Acr-1*	***CAA73042***	***Ananas comosus***

*Previously unnamed RT found by BLAST searches of the GBRD, using rice RTs found in our study as queries.
Acr: Ananas comosus retrotransposon
Atr: Arabidopsis thaliana retrotransposon
Zmr: Zea mays retrotransposon

FIGURES

Figure 1

RT-based neighbor-joining tree for *copia*-like retrotransposons. Distances (uncorrected "p") appear next to each of the branches. RT sequences from plant species other than rice are included for comparison.

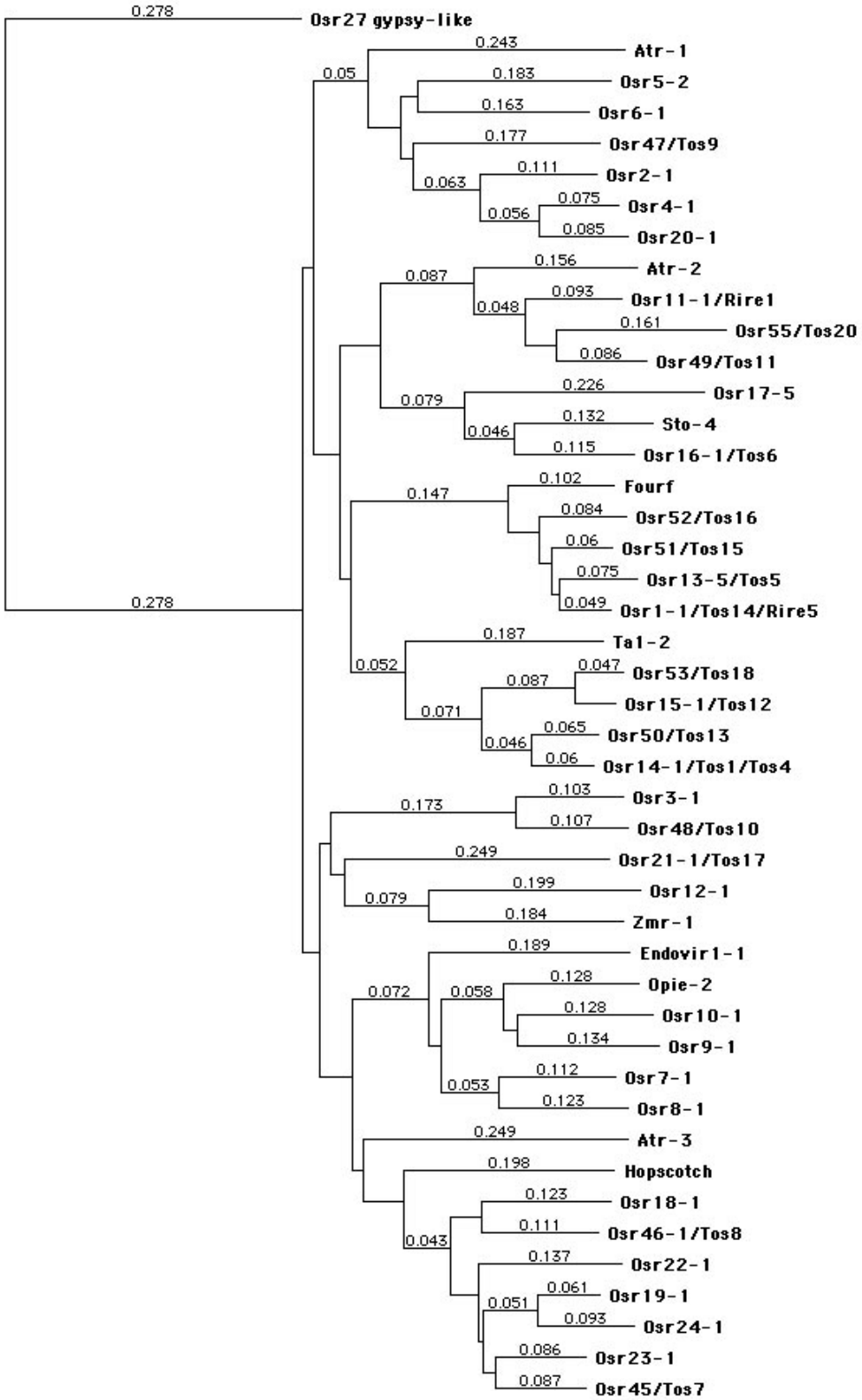
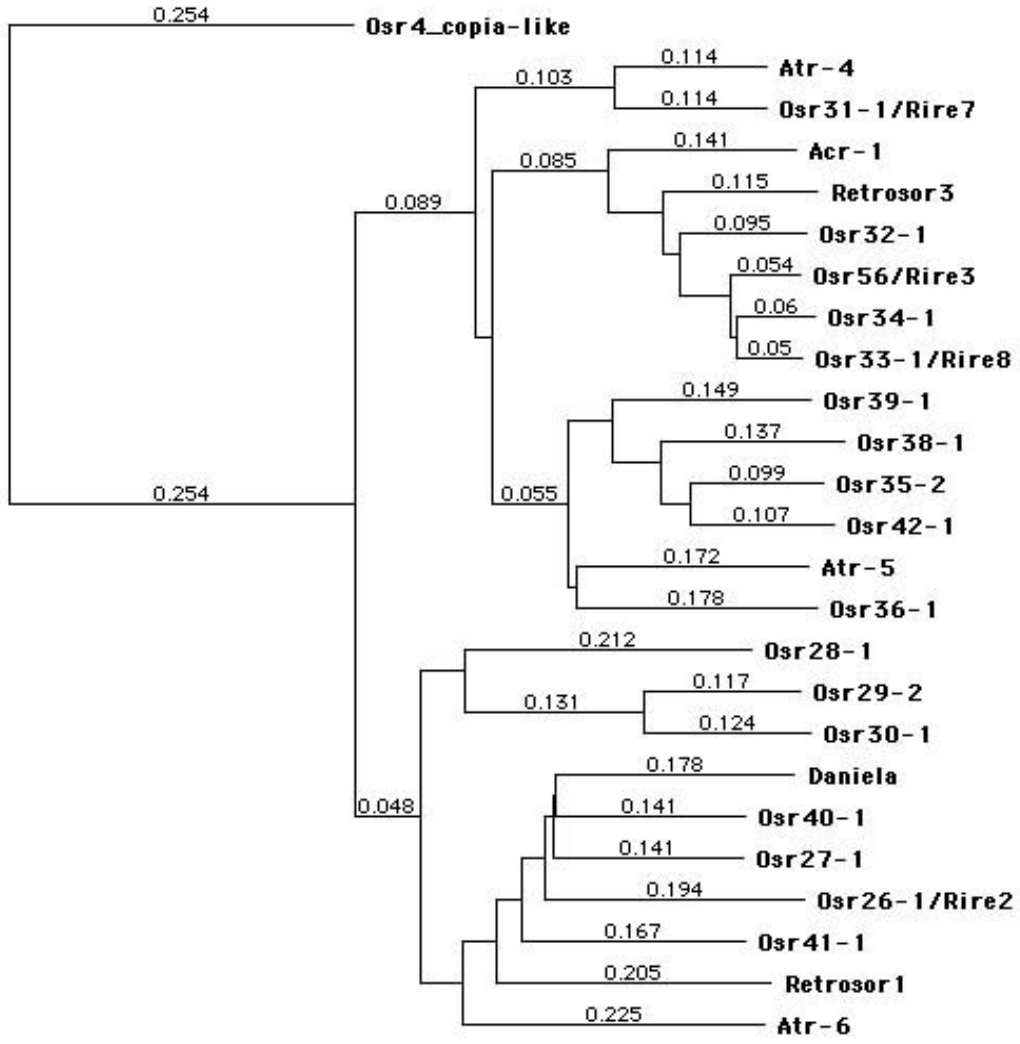

0.278
Osr27 gypsy-like
0.243
Atr-1
0.05
0.183
Osr5-2
0.163
Osr6-1
0.177
Osr47/Tos9
0.111
Osr2-1
0.063
0.075
Osr4-1
0.056
0.085
Osr20-1
0.156
Atr-2
0.087
0.093
Osr11-1/Rire1
0.048
0.161
Osr55/Tos20
0.086
Osr49/Tos11
0.226
Osr17-5
0.079
0.132
Sto-4
0.046
0.115
Osr16-1/Tos6
0.102
Fourf
0.147
0.084
Osr52/Tos16
0.06
Osr51/Tos15
0.075
Osr13-5/Tos5
0.278
0.049
Osr1-1/Tos14/Rire5
0.187
Ta1-2
0.052
0.047
Osr53/Tos18
0.087
Osr15-1/Tos12
0.071
0.065
Osr50/Tos13
0.046
0.06
Osr14-1/Tos1/Tos4
0.103
Osr3-1
0.173
0.107
Osr48/Tos10
0.249
Osr21-1/Tos17
0.199
Osr12-1
0.079
0.184
Zmr-1
0.189
Endovir1-1
0.128
Opie-2
0.072
0.058
0.128
Osr10-1
0.134
Osr9-1
0.112
Osr7-1
0.053
0.123
Osr8-1
0.249
Atr-3
0.198
Hopscotch
0.123
Osr18-1
0.111
Osr46-1/Tos8
0.043
0.137
Osr22-1
0.061
Osr19-1
0.051
0.093
Osr24-1
0.086
Osr23-1
0.087
Osr45/Tos7

Figure 2

RT-based neighbor-joining tree for *gypsy*-like retrotransposons. Distances (uncorrected "p") appear next to each of the branches. RT sequences from plant species other than rice are included for comparison.


0.254
Osr4_copia-like
0.114
Atr-4
0.103
0.114
Osr31-1/Rire7
0.141
Acr-1
0.085
0.115
Retrosor3
0.089
0.095
Osr32-1
0.054
Osr56/Rire3
0.06
Osr34-1
0.05
Osr33-1/Rire8
0.149
Osr39-1
0.137
Osr38-1
0.254
0.099
Osr35-2
0.055
0.107
Osr42-1
0.172
Atr-5
0.178
Osr36-1
0.212
Osr28-1
0.117
Osr29-2
0.131
0.124
Osr30-1
0.178
Daniela
0.048
0.141
Osr40-1
0.141
Osr27-1
0.194
Osr26-1/Rire2
0.167
Osr41-1
0.205
Retrosor1
0.225
Atr-6

CHAPTER 4

LTR RETROTRANSPOSONS OF *MUS MUSCULUS*[1]

[1]McCarthy, E.M., and J.F. McDonald (to be submitted to Genome Biology).

ABSTRACT

Using a new data-mining program, LTR_STRUC, (*LTR* retrotransposon *struc*ture program), we have mined the ENSEMBLE mouse (*Mus musculus*) data set for LTR retrotransposons. We have used elements found in this initial search, as well as murine LTR retrotransposons identified by previous workers, to conduct BLAST searches of the GenBank mouse database. They break down into 20 separate families, 13 of which have not been previously described.

INTRODUCTION

Retrotransposons are mobile genetic elements that make up a large fraction of most eukaryotic genomes. All retrotransposons are distinguished by a life cycle involving an RNA intermediate. The RNA genome of a retroelement is copied into a double-stranded DNA molecule by reverse transcriptase, which is subsequently integrated into the host's genome. Retrotransposons fall into two main categories: those with long terminal repeats (LTRs), such as retroviruses and LTR retrotransposons, and those that lack such repeats (*e.g.,* long interspersed nuclear elements or "LINEs").

Retrotransposons are particularly abundant in plants, where they are often a principal component of nuclear DNA. In corn, 50-80%, and, in wheat, fully 90% of the genome is made up of retrotransposons [1,2]. In animals this percentage is generally lower than in plants but it can still be significant. For example, about 8% of the human genome is now known to be composed of LTR retrotransposons [3]. In the mouse genome this figure has been estimated at 10% [4].

This paper presents the results of a recent (Dec., 2002) survey of the GenBank mouse (*Mus musculus*) database (GBMD) and the 2.9 Gbp Ensemble (http://www.ensembl.org) mouse dataset (EMD) for the presence of LTR retrotransposons. We have employed a new search program, LTR_STRUC (*LTR* retrotransposon *struc*ture program) as the initial data-mining tool in our survey [5]. Identified elements were subjected to sequence analyses to identify open reading frames (ORFs) encoding reverse transcriptase (RT) and other retroviral proteins. LTR_STRUC finds only full-length elements, that is, ones having two LTRs and a pair of target site repeats (TSRs). We therefore augmented our search approach by conducting BLAST

searches using reverse transcriptase queries. These queries are of two types: 1) previously known RTs in the public database from mouse and other mammals; 2) RTs obtained from our initial scan of the EMD with LTR_STRUC. Subsequent RT sequence alignments were carried out, followed by construction of phylogenetic trees.

An LTR retrotransposon "family" is defined as a group of elements with RTs at least 90% similar at the amino acid level [6]. Experience has shown that when two elements have RTs that are 90% similar, their LTRs are usually about 60% similar. Thus, non-autonomous elements, lacking an RT ORF, are assigned to the same family if their LTRs are at least 60% similar. Many LTR retrotransposons replicate non-autonomously. Four different families of murine LTR retrotransposons have non-autonomous members. (*MalR* elements, *ETn* elements, and two types identified in this study, one related to *IAP* elements, the other to *MmERV* elements). These are discussed below. Non-autonomous elements can reach a high copy number even though they lack an RT ORF [4, 7, 8, 9].

Currently there is no standard mouse retrotransposon nomenclature. In our system of classification for mouse, LTR retrotransposons are specified by the acronym *Mmr* (*Mus musculus* retrotransposon). Distinct families are indicated by number (*e.g., Mmr1*, *Mmr2*, *Mmr3, . . .*). We have chosen to adopt the *Mmr* nomenclature in this study because it is consistent with the systematic logic used in previous papers ("Mm" indicative of the genus and species of the host organism; r indicates retrotransposon). In each case where we use the *Mmr* acronym in this paper to refer to a previously named family, we also include any pre-existing name for the family.

RESULTS AND DISCUSSION

RTs from elements identified in our survey fall into numerous distinct families. All autonomous LTR retrotransposons identified were of the *gypsy*-like elements (Classes I, II, and III). Autonomous retroviral-like elements in the mouse genome usually have an overall length of between 6000 and 9000 bp. Results of our study indicate that the target site repeats (TSRs) of mouse LTR retrotransposons are 4-6 base pairs long and that within each the three major classes of these elements a single TSR length is characteristic. With the exception of a few mutated copies, mouse LTR retrotransposons seem to have the same canonical dinucleotides terminating the LTRs as are typically found in other species (TG/CA). The LTRs of murine retroviral-like elements are generally 300-600 bp long, with the exception of mouse mammary tumor virus (MMTV) where the LTRs are some 1300 bp in length. Our survey shows that at least 20 distinct LTR retrotransposon families exist in the mouse genome, 13 of which that have not been described previously (see below).

LTR Retrotransposon Families Of The Murine Genome

Overview: To date, LTR retrotransposon diversity has been rigorously classified into families for only a few organisms (*Oryza sativa* [7], *Drosophila melanogaster* [6], and *Caenorhaditis. elegans* [10]). This paper represents a first attempt to establish a similar uniform classification and nomenclature for the domestic mouse. Past studies have classified murine retrotransposons only into broad categories, which ignore the standard definition of "family" (see above). For example, the term "intracisternal type A particle" (IAP) has been used to refer to elements that belong to several distinct LTR-

retrotransposon phylogenetic groups. The autonomous elements identified in our survey of the GBMD and EMD fall into 20 families on the basis of degree of RT divergence (>10% = family). In addition, we have classified *MalR* elements, which are non-autonomous in the same family as *MuERV-L* elements because these two types of transposons have similar LTRs. *MusD* and *ETn* elements form an additional pair of related autonomous and non-autonomous elements. These two pairs of families are discussed in more detail below.

Three distinct classes of murine LTR retrotransposons are recognized [4]: Class I, containing elements related to retroviral leukemia viruses in mouse (*MuLV*) and other species (e.g., *GALV*, *FeLV*); Class II which contains the *IAP* elements, mouse mammary tumor virus (MMTV), and the *MusD2*/*ETn* family; and Class III which comprises the *MalR* and *MuERV-L* elements.

Class I (families 1-4): Members of this class make up 0.68% of the mouse genome (copy number ~34,000) [4]. They have 4-bp TSRs and are related to murine leukemia virus (*MuLV;* AF033811), a C-type retrovirus that occurs only in mice, a major cause of cancer in that species. Class I, to which *MuLV* belongs, contains at least three other families: *Mmr1_MmERV, Mmr3_MuRRS,* and *Mmr4*. In this paper *MuLV* is referred to as family *Mmr2_MuLV*. Class I endogenous retroviruses are more closely related to elements in other species than to mouse retroelements belonging to classes II or III. RTs from endogenous retrovirus in pig (*PK15;* AF038601) and koala (*KoRV;* AAF15098), as well as from leukemia viruses in gibbon (*GALV;* AAA466810) and cat (*FeLV;* L06140), group with this class and their RTs are all about 80% similar at the amino acid level to

those of murine Class I elements. One member of Class I is found in two different mouse species, *Mus musculus* and *M. dunni*, and has previously been referred to either as *MmERV* (in *M. musculus*) or *MDEV* (in *M. dunni*) [11]. It is here referred to as *Mmr1_MmERV*. The identity of this family in these two species, is demonstrated by the presence of an element (AAC31805) in the *M. dunni* (Indian pigmy mouse) genome, which is 96% similar (at the amino acid level) to members of *Mmr1_MmERV* resident in *M. musculus* (Figure 1). This finding is consistent either with a recent common origin of these two mouse species or with a horizontal transfer of this retrovirus. This virus may be infectious, since an envelope protein sequence is present in the GenBank database (AAC31806) for the *M. dunni* retrovirus and has been detected as well in copies of this family during our own survey of *M. musculus*. *Mmr4* is a previously unrecognized Class I family, with members about 80% similar to those of *Mmr2_MuLV*. Family *Mmr3_MuRRS* includes the so-called murine retroviral related sequences (*MuRRS)*. A known human endogenous retrovirus type C oncoviral sequence (AAA73090) is approximately 56% similar at the amino acid level to members of Class I. BLAST searches with RT queries from Class I indicate that at least some elements in the human genome are even more similar (65%+) to Class I elements in mouse (e.g., HSAP_2, Figure 1, Table 3).

Class II (families 5-19): Class II retroviral-like elements make up 3.14% of the mouse genome (copy number ~127,000) [4]. This class contains 15 of the 20 murine LTR families. Its members have 6bp TSRs and are related to MMTV (NC_001503), an oncogenic B-type retrovirus that causes breast cancer in mice. Our survey has revealed

only three full-length copies of a member of this family (*Mmr11_MMTV*) in the mouse genome, only two of which could be assigned to a specific chromosome (Table 1, *Mmr11_MMTV-1* and *–2*). *MMTV* contains an ORF coding for envelope protein (BAA03768) so the life cycle of this virus in mice may be predominantly infectious, rather than hereditary. *Mmr11_MMTV* RTs are also 75% similar to those of a separate endogenous mouse family, *Mmr16*. For the most part, *Mmr16* seems to be represented in the mouse genome by fragmentary elements, but the full-length element *Mmr16-1* described in Table 1 has a full complement of retroviral genes, including an envelope ORF.

Another family in Class II, *Mmr19_MusD*, has been previously described under the name *MusD*. Mager and Freeman (2000) [8], who discovered this family, showed that the non-autonomous mouse *ETn* retroelements (early transposons) are deletion derivatives of *Mmr19_MusD*. They are so closely related to *MusD* elements that we have assigned them to the same family. Most copies of the former are around 5500 bp long, while those of the latter are usually around 7400 bp in length. *ETn* elements (Y17107; AB033509), first reported by Brulet et al. (1983) [12], are a moderately repetitive family of murine retrotransposons that lack most of the usual retroviral ORFs. Our survey with LTR_STRUC suggests that full-length copies of ETn elements are about half again as common as full-length MusD elements. Family *Mmr12* is about 80% similar to *Mmr19_MusD*. Both of these families are 70% similar to Mason-Pfizer Monkey Virus (MPMV; NC_001550). The RTs of MusD elements have an unusual active site sequence: F*T*DDVLM ("T" is not canonical for an active site) [13].

Class II contains an additional clade (See Figure 2), comprising at least eight additional families (*Mmr6, Mmr7, Mmr9, Mmr10_IAP, Mmr14, Mmr15, Mmr17,* and *Mmr18*) with no two families differing from any other by more than 70%. The major constituents of this clade are the IAP retrotransposons, the second most abundant family in the mouse genome, here referred to as family *Mmr10_IAP*. They lack complete *env* genes [14] and thus are considered non-infective. Murine elements identified in GenBank as "IAP" (e.g., GNPSIP, GNMSIA) are restricted to family *Mmr10_IAP*. Nevertheless, members of any of the eight families listed above have been described as "IAP" by various authors. In addition, a family of retroelements in golden hamster (GH_G18; Figure 2) have been described as "IAP" but do not actually belong to the *Mmr10_IAP* family (their RT ORFs differ from those of *Mmr10_IAP* by about 18% at the amino acid level). Thus, in mice, the term "*IAP"* might best be restricted to *Mmr10_IAP*. Numerous IAP elements share a common, 1800-bp deletion that includes the upstream end of the RT. Yet these elements were, and perhaps still are, capable of transposing as evidenced by the fact that copies with the same deletion were found on many different chromosomes. Even shorter, internally deleted elements, with two LTRs and ostensibly capable of transposition, can be assigned to *Mmr10_IAP* on the basis of LTR similarity (down to about 2,700 bp in overall length).

Class III (family 20): Members of this class make up 5.40% of the mouse genome (copy number ~ 442,500) [4]. They have 5bp TSRs Class III has two constituents: 1) murine *ERV-L* elements, which have an estimated copy number of 37,000 [4] and the non-autonomous *MalR*s (mammalian apparent LTR retrotransposons), which are the most

common retroviral element in the mouse genome, making up 4.8% of the mouse genome [4]. *MuERV-L* elements are closely related to human endogenous retrovirus L (*HERV-L*). In BLAST searches we have identified a human element (HSAP_1, Table 3, Figure 3) that is 85% similar at the amino acid level to *MuERV-L* RTs. Because alignments of their LTRs show that the first and last 100 bp of the LTRs are about 80% similar, we have assigned murine *MalR*s and *MuERV-L* elements to the same family, referred to in this paper as *Mmr20_ MuERV-L*. Like *MalR*s in other species, murine MalRs are all internally deleted. The internal region contains only non-coding repetitive DNA. Nevertheless they have typical LTRs, PBS, and PPT. Murine *MalR*s are of two types: 1) MT MalRs the most common type of LTR retrotransposon in the mouse genome (mean length ~1980bp); and 2) ORR1 MalRs (mean length ~2460 bp). Our survey suggests that in the mouse genome MT MalRs are about 10 times as common as their longer relatives, the ORR1 MalRs. The vast majority of non-truncated *Mmr20_ MuERV-L* elements, have an overall length of about 6400 bp.

Length variation in murine LTR retrotransposons

Although all copies of family *Mmr10_IAP* found by LTR_STRUC have two LTRs and recognizable TSRs (as required by the search algorithm employed by the program), the individual members of this abundant family vary widely in overall length (2700-7200bp) due to the presence of internal deletions of varying length. On the other hand, the two, abundant types of non-autonomous Class III elements (MT and ORR1 *MalR*s) exhibit a markedly different pattern of variation from that of *Mmr10_IAP* elements. Lengths of ORR1 *MalR*s peak sharply at 2,300bp, those of MT *MalR*s, at

1980bp, with very few elements in either case differing from these peak frequencies by more than 100 bp (<1%). Moreover, most copies of *Mmr10_IAP*, from the shortest to the longest, are preponderantly represented by copies with a high level of LTR-LTR identity (99%+), a finding consistent with recent transposition. The ability of internally truncated *Mmr10_IAP*s to complete their replication cycle is confirmed by the fact that a number of *Mmr10_IAP* copies bearing the same 1800-bp deletion (affecting the polyprotein ORF) were found in our survey on a variety of different mouse chromosomes. A similar dispersed distribution of lengths was observed in two other families *Mmr19_MusD* and *Mmr1_MmERV*.

Interspecific considerations

Certain families of mouse LTR retrotransposons are more closely related to elements present in other species than to other classes of mouse elements. For example, murine Class I elements are more similar to viruses in gibbon, pig, cat, and koala, than to murine retrotransposons of Classes II or III (Figure 1). Among Class II murine endogenous retroviruses (Figure 2), family *Mmr10_IAP* is more closely related to the golden hamster element GH_G18 than it is to any other family of murine retroviral elements. Similarly, the amino acid sequences (RT ORFs) of members of *Mmr20_MuERV_L* (mouse Class III elements, Figure 3) differ from a human element (e.g., *HSAP_1*, Table 2) by only 15%, but differ from those of any non-Class III element by more than 60%. Such findings suggest that cross-species infection has been a source of new mouse LTR retrotransposon families over evolutionary time.

CONCLUSIONS

All autonomous retrotransposons identified in our study were retroviral-like elements (of classes I, II, and III). At least 20 distinct families of murine LTR retrotransposons exist. Families *Mmr4*, *Mmr5*, *Mmr6*, *Mmr7*, *Mmr8*, *Mmr9*, *Mmr12*, *Mmr13*, *Mmr14*, *Mmr15*, *Mmr16*, *Mmr17*, and *Mmr18* have not been previously recognized, 13 families in all. These new families are all Class II elements (with the exception of *Mmr4,* which belongs to Class I) and are thus akin to immune deficiency viruses such as simian retrovirus SRV-1, to mouse mammary tumor virus (MMTV), and to IAP elements.

Our purpose in using LTR_STRUC to begin our survey of the mouse genome was to obtain a broadly representative sample of murine retrotransposons. Since the algorithm it employs is not dependent upon sequence homology, as are standard search methods (*e.g.,* BLAST), the initial results of our survey presumably were not biased toward a particular set of queries. Also, since the current version of LTR_STRUC now categorizes the elements it locates and assigns a new name to any element that differs sufficiently from any found earlier in the search, the chances of overlooking low-copy families has been reduced. The thoroughness of our BLAST search can only have been augmented by using LTR_STRUC because, in the BLAST phase of our survey, the queries used were a combination of those element types already recognized, prior to our investigation, with those found by LTR_STRUC. We believe this approach is the reason we were able to identify the 13 previously unreported families listed above.

METHODS

Using a new data-mining program, LTR_STRUC [5], (*LTR* retrotransposon *struc*ture program), we have mined the ENSEMBLE mouse (*Mus musculus*) data set for LTR retrotransposons. We have used elements found in this initial search, as well as murine LTR retrotransposons identified by previous workers, to conduct BLAST searches of the GenBank mouse database.

Automated characterization of LTR retrotransposons: The methods used in our survey of the mouse genome are essentially the same as those used in our earlier study of the rice genome and are described elsewhere [7]. Briefly, we began our survey by using a new computer program, LTR_STRUC, which identifies new LTR retrotransposons based on the presence of characteristic retroelement features, LTR_STRUC [5]. Additional elements were identified by BLAST searches using the RTs, both of elements located by LTR_STRUC and of ones previously recognized in earlier studies by previous researchers.

Data sets scanned: Initial scans with LTR_STRUC were conducted on a data set consisting of the 2.9 Gbp of *M. musculus* sequence data available in the Ensemble database at the time of the initial scan (Dec., 2002). This data set (EMD) was obtained from the Ensemble web site [15]. In an effort to identify additional elements not picked up in the initial survey with LTR_STRUC, we have used representative sequences from each retrotransposon family identified in this study as queries to conduct BLAST searches against the GenBank mouse database (GBMD). Thus, the results reported here constitute a reasonably unbiased survey of LTR-retrotransposon diversity in mouse. RT sequences were identified according to previously described criteria [13,16].

Multiple sequence alignments and phylogenetic analyses: The RT domains of the various *Mmr* elements were aligned, as described elsewhere [7], with previously reported RT sequences (Table 2).

REFERENCES

1. SanMiguel P, Tikhanov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL: Nested retrotransposons in the intergenic regions of the maize genome. Science 1996, 274: 765-768.

2. Flavell RB: Repetitive DNA and chromosome evolution in plants. Phil. Trans. R. Soc. Lond. B. Biol. Sci. 1986, 13: 335-340.

3. Lander ES, et al.: Initial sequencing and analysis of the human genome. Nature 2001, 409: 860-921.

4. Mouse Genome Sequencing Consortium: Initial Sequencing and Comparative analysis of the mouse genome. Nature 2002, 420: 520-562.

5. McCarthy EM, McDonald JF: LTR_STRUC: A Novel Search And Annotation Program for LTR Retrotransposons. Bioinformatics 2003, 19: 362-367.

6. Bowen N, McDonald JF: Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res. 2001, 11:1527-1540.

7. McCarthy, EM, Liu J, Gao L, McDonald JF: Long terminal repeat retrotransposons of Oryza sativa. Genome Biology 2002, 3(10): RESEARCH0053.

8. Mager DL, Freeman JD: Novel mouse Type D endogenous proviruses and ETn Elements share long terminal repeat and internal sequences. J. Virology 2000, 74: 7221-7229.

9. Jiang N, Jordan IK, Wessler SR: Dasheng and RIRE2: A non-autonomous long terminal repeat element and its putative autonomous partner in the rice genome. Plant Physiol. 2002, 130:1697-1705.

10. Bowen N, McDonald JF: Genomic analysis of Caenorhabditis elegans reveals ancient families of retroviral-like elements. Genome Res. 1999, 9:924-935.

11. Bromham L, Clark, F, and McKee,JJ: Discovery of a novel murine type C retrovirus by data mining. J. Virol. 2001, 75: 3053-3057.

12. Brulet P, Kaghad M, Xu YS, Croissant O, Jacob F: Early differential tissue expression of transposon-like repetitive DNA sequences of the mouse. PNAS USA 1983, 80: 5641-5645.

13. Xiong Y, Eickbush TH: Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol. Biol. Evol.1988, 5: 675-690.

14. Kuff EL, Lueders KK The intracisternal A-particle gene family: Structure and functional aspects. Adv. Cancer. Res. 1988, 51:183-276.

15. Ensemble Web Site [http://www.ensembl.org/].

16. Xiong Y, Eickbush TH: Origin and evolution of retroelements based upon their reverse-transcriptase sequences. EMBO 1990, 9: 3353-3362.

TABLES

Table 1

Exemplars of Mouse LTR Retrotransposon Families Characterized in this Study

Family	Accession Number	Location	Chrm. num.	LTR length	Element length	TSR	LTR-LTR %ID
Mmr1_MmERV-1	ND	ND	3	561	8797	TAAC	ND
Mmr1_MmERV-2	ND	ND	14	555	6942	TTAG	99.5
Mmr1_MmERV-3	AC116580	60869-69866	18	562	8998	GATG	99.1
Mmr1_MmERV-4	ND	ND	9	554	7669	TTAT	99.8
Mmr2_MuLV-1	ND	ND	13	547	8692	GTAC	ND
Mmr2_MuLV-2	ND	ND	9	ND	ND	ND	ND
Mmr2_MuLV-3	ND	ND	8	523	8730	AGCT	99.8
Mmr3_MuRRS-1	ND	ND	9	482	5747	CATC	99.6
Mmr3_MuRRS-2	ND	ND	11	483	5800	AGGG	97.7
Mmr3_MuRRS-3	ND	ND	5	483	5468	TGTG	97.6
Mmr3_MuRRS-4	ND	ND	10	482	5687	CTAT	94.1
Mmr4-1	ND	ND	X	436	8776	ATGC	97.4
Mmr4-2	ND	ND	4	431	8444	CTAC	99.3
Mmr4-3	AC129291	52257-60643	6	431	8391	GCTG	ND
Mmr4-4	ND	ND	4	430	8439	CCTAT	99.5
Mmr4-5	ND	ND	12	432	8437	ATGC	98.1
Mmr5-1	AL953900	63071-63586†	2	ND	ND	ND	ND
Mmr5-2	ND	ND	15	408	14327*	GTAAGC	90.1
Mmr6-1	ND	ND	17	345	8287	GTCATA	94.6
Mmr6-2	ND	ND	5	435	7077	GTTCTG	99.8
Mmr6-3	AL645686	82031-82609†	13	ND	ND	ND	ND
Mmr7-1	AL669907	109127-109663†	11	ND	ND	ND	ND
Mmr7-2	AC121947	7642-8223†	ND	ND	ND	ND	ND
Mmr8-1	ND	ND	17	385	8980	CTCAGT	87.7
Mmr9-1	AC093445	57410--58100†	1	ND	ND	ND	ND
Mmr10_IAP_1	GNMSIA	NA	4	ND	ND	ND	ND
Mmr10_IAP_2	AC066688	63525-70600	6	336	7076	ATAACT	99.7
Mmr10_IAP_3	ND	ND	19	324	7063	CCTTGC	97.1
Mmr11_MMTV-1	NW_000207.1	3688585- 3689262†	4	1330	9914	GCTCCC	98.5
Mmr11_MMTV-2	BAA03767.1	1-278	ND	ND	ND	ND	ND
Mmr12-1	AL645683	6610- 5843†	13	ND	ND	ND	ND
Mmr11-1	AL928538	44096-44690	X	ND	ND	ND	ND
Mmr13-1	AC122304	117988-118560	18	ND	ND	ND	ND
Mmr14-1	AL669827	49044-57291	11	306	8248	CAGAGA	96.0
Mmr15-1	AC127274	11141-11509	17	380	8968	AGAAAG	ND
Mmr16-1	Mm6_WIFeb01_115	5320195-5329185	6	973	8982	GAGTTT	ND
Mmr17-1	AL928539	62443-63024†	ND	ND	ND	ND	ND
Mmr17-2	ND	ND	16	374	13434	GACAGT	96.1
Mmr18-1	AC093341	96667- 101604	5	359	4938	GGGATC	94.4
Mmr18-2	ND	ND	12	357	6305	GACTG*	95.0
Mmr18-3	ND	ND	10	356	6056	ATTGGC	94.4
Mmr19_MusD-1	ND	ND	14	319	7404	GTCACA	ND
Mmr19_MusD-2	ND	ND	16	319	7485	TTTGGG	98.8
Mmr19_MusD-3	AC24426	12212-13012†	13	ND	ND	ND	ND
Mmr19_MusD-4	ND	ND	6	319	7479	ATCATG	98.3
Mmr19_MusD-5	ND	ND	17	319	7450	TTTCAC	99.1
Mmr19_MusD-6	ND	ND	12	320	6022	GGATGG	97.8
Mmr20_MuERV_L-1	ND	ND	12	493	6335	GTCGG	100.0
Mmr20_MuERV_L-2	ND	ND	13	491	6442	CTGCC	99.8
Mmr20_MuERV_L-3	ND	ND	15	492	6396	GTTTG	100.0

† Endpoints given are for RT not the whole element
*TSR of his copy is not of the expected length for the family
ND: Not determined; NA: Not applicable

Table 2

Known RTs used for comparison in phylogenies

Symbol	Name of retrovirus	Accession Number/Citation	Host genus
GALV	Gibbon ape leukemia virus	AAA46810	*Hylobates*
PERV	Porcine endogenous retrovirus ERV-PK15	AF038601	*Sus*
BLV	Bovine Leukemia Virus	P03361	*Bos*
HERV-K	Human endogenous retrovirus K	P10266	*Homo*
*HBCA**	Human breast cancer associated	AAG18012	*Homo*
HERV-L	Human endogenous retrovirus L	Z72519	*Homo*
*GH_H18**	Golden hamster intracisternal A-particle H18	GNHYIH	*Mesocricetus*
FeLV	Feline leukemia virus	L06140	*Felis*
RERV	Rabbit endogenous retrovirus	AAM81191	*Oryctolagus*
*IAP-H**	Hamster intracisternal type-A	P04026	*Cricetus*
SRV-1	Simian SRV-1 type D retrovirus	M11841	*Macaca*
MPMV	Mason-Pfizer Monkey Virus	GNLJMP	*Macaca*
MuLV	Moloney murine leukemia virus	AF033811	*Mus*
MuERV-L	Murine endogenous retrovirus ERV-L	T29097	*Mus*
MusD	Murine type D-like endogenous retrovirus MusD1	AF246632	*Mus*
HERV-C	Human endogenous retrovirus type C oncovirus	AAA73090	*Homo*
*Phasco**	Koala type C endogenous virus	AAF15098	*Phascolarctos*
*MDEV**	*Mus dunni* endogenous virus	AAC31805	*Mus*
MMTV	Mouse mammary tumor virus	NC_001503	*Mus*
MmERV	*M. musculus* endogenous retrovirus	Bromham et al., 2001	*Mus*

*Symbol used only in this study

Table 3

RTs obtained from translating BLAST used in phylogenies

Symbol	Name of retrotransposon	Accession Number	Position of RT in file	Host genus
*HSAP_1**	Human endogenous retrovirus L	AL590235	114430-115010	*Homo*
*HSAP_2**	Human endogenous C type retrovirus	AC078899	151820-152410	*Homo*

*Symbol used only in this study

FIGURES

Figure 1

RT-based neighbor-joining tree for Class I murine retrotransposons. Distances (uncorrected "p") appear next to each of the branches. RT sequences from plant species other than mouse are included for comparison. The outgroup is the Class II element GH_H18 (from golden hamster, *Mesocricetus auratus*; see Table 2 and Figure 2).

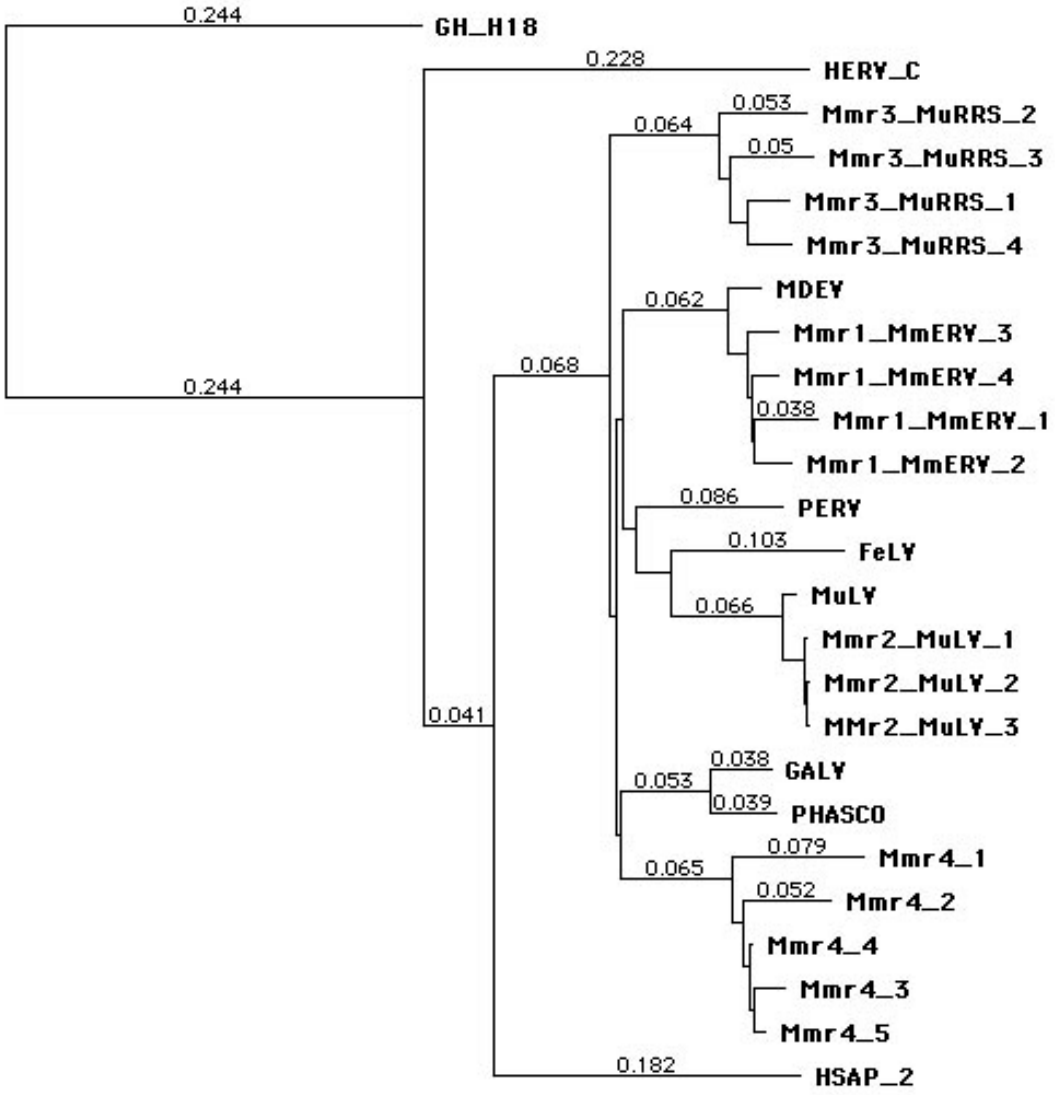

0.244
GH_H18
0.228
HERV_C
0.053
Mmr3_MuRRS_2
0.064
0.05
Mmr3_MuRRS_3
Mmr3_MuRRS_1
Mmr3_MuRRS_4
MDEV
0.062
Mmr1_MmERV_3
0.068
Mmr1_MmERV_4
0.244
0.038
Mmr1_MmERV_1
Mmr1_MmERV_2
0.086
PERV
0.103
FeLV
MuLV
0.066
Mmr2_MuLV_1
Mmr2_MuLV_2
0.041
MMr2_MuLV_3
0.038
GALV
0.053
0.039
PHASCO
0.079
Mmr4_1
0.065
0.052
Mmr4_2
Mmr4_4
Mmr4_3
Mmr4_5
0.182
HSAP_2

Figure 2

RT-based neighbor-joining tree for Class II murine retrotransposons. Distances (uncorrected "p") appear next to each of the branches. RT sequences from plant species other than mouse are included for comparison. The outgroup is the Class I element MDEV (from house/rice field mouse, *Mus dunni*; see Table 2 and Figure 1).

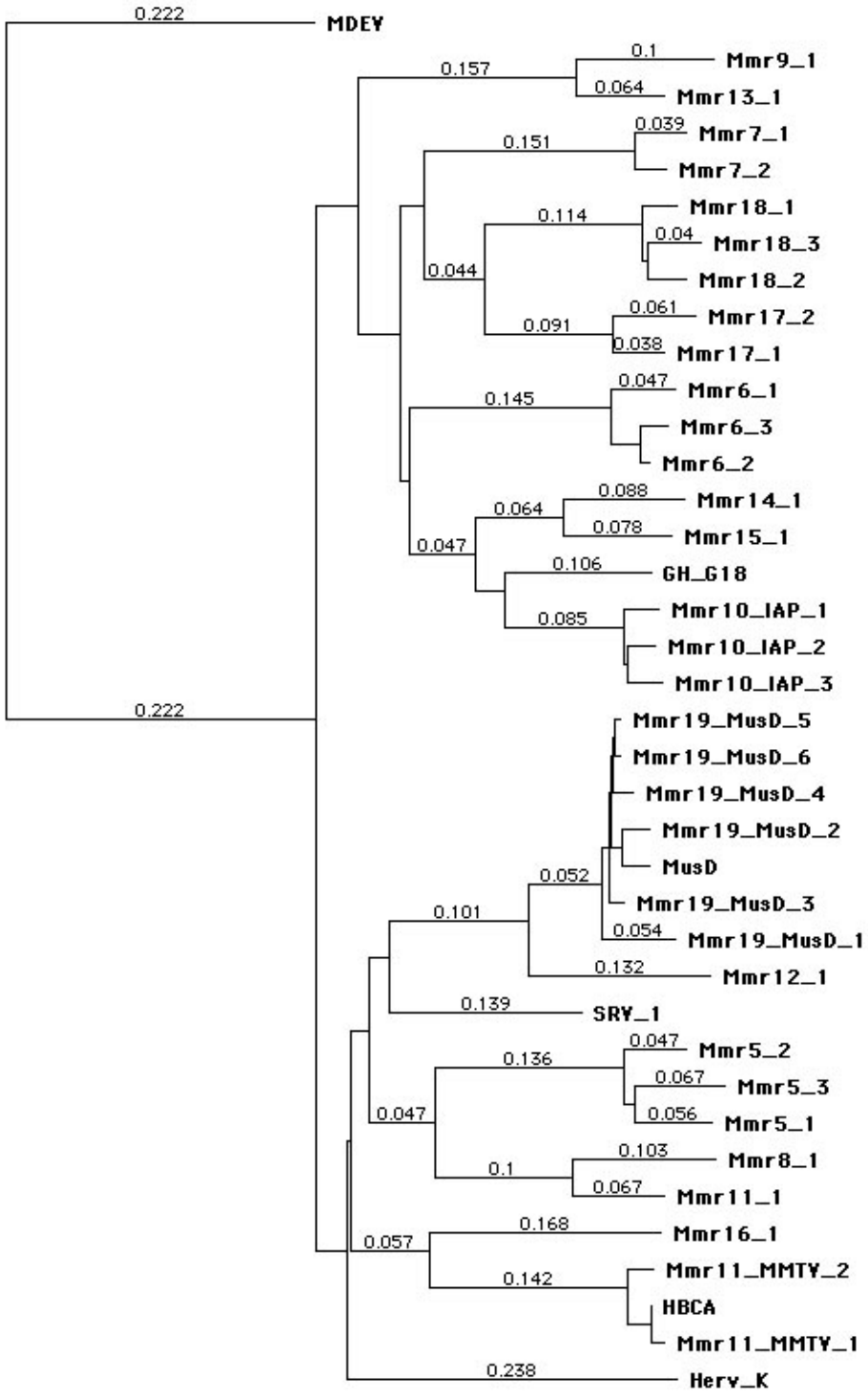
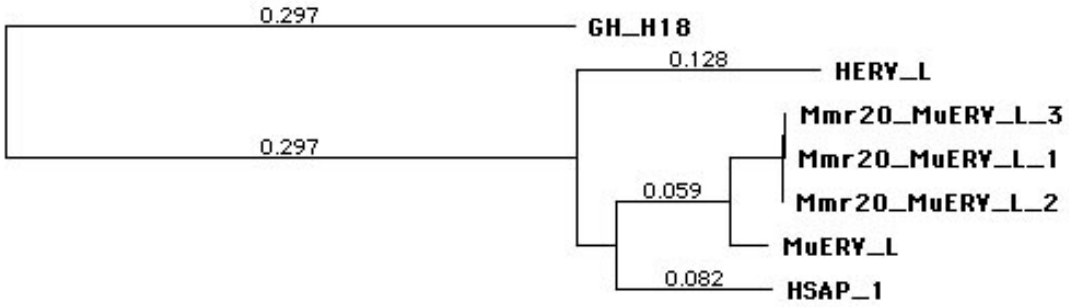

0.222
MDEV
0.1
Mmr9_1
0.157
0.064
Mmr13_1
0.039
Mmr7_1
0.151
Mmr7_2
Mmr18_1
0.114
0.04
Mmr18_3
0.044
Mmr18_2
0.061
Mmr17_2
0.091
0.038
Mmr17_1
0.047
Mmr6_1
0.145
Mmr6_3
Mmr6_2
0.088
Mmr14_1
0.064
0.078
Mmr15_1
0.047
0.106
GH_G18
Mmr10_IAP_1
0.085
Mmr10_IAP_2
Mmr10_IAP_3
0.222
Mmr19_MusD_5
Mmr19_MusD_6
Mmr19_MusD_4
Mmr19_MusD_2
MusD
0.052
Mmr19_MusD_3
0.101
0.054
Mmr19_MusD_1
0.132
Mmr12_1
0.139
SRY_1
0.047
Mmr5_2
0.136
0.067
Mmr5_3
0.047
0.056
Mmr5_1
0.103
Mmr8_1
0.1
0.067
Mmr11_1
0.168
Mmr16_1
0.057
Mmr11_MMTV_2
0.142
HBCA
Mmr11_MMTV_1
0.238
Herv_K

Figure 3

RT-based neighbor-joining tree for Class III murine retrotransposons. Distances (uncorrected "p") appear next to each of the branches. RT sequences from plant species other than mouse are included for comparison. The outgroup is the Class II element GH_H18 (from golden hamster, *Mesocricetus auratus*; see Table 2 and Figure 2).


0.297
GH_H18
0.128
HERV_L
Mmr20_MuERV_L_3
Mmr20_MuERV_L_1
Mmr20_MuERV_L_2
0.297
0.059
MuERV_L
0.082
HSAP_1

CONCLUSION

The research presented in this dissertation, has a bipartite thrust. The intention has been first to supply a description of a new data-mining program, LTR_STRUC. The second purpose was to demonstrate the utility of LTR_STRUC by using it to identify and characterize LTR retrotransposons in two large nucleotide sequence databases, those of *Oryza sativa* and *Mus musculus*. These two searches were conceived as tests of the program that would concurrently yield useful data. The tests have been highly successful; in both of these species we more than doubled the number of recognized families.

Publication of our results for rice (McCarthy, et al., 2002) and a description of the algorithm (McCarthy and McDonald, 2003) have prompted numerous requests for LTR_STRUC from other scientists. To meet this demand we have made the latest version of the program available on the McDonald lab website. I plan to continue updating LTR_STRUC, not only to increase its speed and accuracy, but also to satisfy any user requests concerning improvement of program function. Already, since the publication of our article on LTR_STRUC in *Bioinformatics* last month, I have enhanced the code so that the program now categorizes elements all of the elements that it finds on the basis of RT, PBS, and LTRs. This new feature greatly reduces the labor involved in creating alignments and phylogenies. In conclusion, our experience using LTR_STRUC encourages us to believe that it will be a useful addition to the repertory of genetic data-mining programs.

REFERENCES

Bowen N., McDonald J.F.: Genomic analysis of Caenorhabditis elegans reveals ancient families of retroviral-like elements. Genome Res. 1999, 9:924-935.

Coffin J., Hughes S., and Varmus H. Retroviruses Cold Spring Harbor, NY: Cold Springs Harbor Laboratory Press, 1997, 343-435.

Flavell R.B.: Repetitive DNA and chromosome evolution in plants. Phil. Trans. R. Soc. Lond. B. Biol. Sci. 1986, 13: 335-340.

Jordan, I.K. and McDonald, J.F. Comparative genomics and evolutionary dynamics of Saccharomyces cerevisiae Ty elements Genetica 1999; 107(1-3):3-13 Lander ES, et al.: Initial sequencing and analysis of the human genome. Nature 2001, 409: 860-921.

McCarthy E.M., McDonald J.F.: LTR_STRUC: A Novel Search And Annotation Program for LTR Retrotransposons. Bioinformatics 2003, 19: 362-367.

McCarthy, E.M., Liu J., Gao L., McDonald J.F.: Long terminal repeat retrotransposons of Oryza sativa. Genome Biology 2002, 3(10): RESEARCH0053.

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

Mouse Genome Sequencing Consortium: Initial Sequencing and Comparative analysis of the mouse genome. Nature 2002, 420: 520-562.

SanMiguel P., Tikhanov A., Jin Y.K., Motchoulskaia N., Zakharov D., Melake-Berhan A., Springer P.S., Edwards K.J., Lee M., Avramova Z., Bennetzen J.L.: Nested retrotransposons in the intergenic regions of the maize genome. Science 1996, 274:765-768.

Sherratt, D.J. 1995. Mobile genetic elements. New York: Oxford Univ. Press.

Tristem M. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J. Virol. 2000, 74: 3715-3730.

Yoder J.A., Walsh C.P., Bestor T.H.: Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. 1997, 13:335-340.