Supporting Information

Kalendar *et al.* 10.1073/pnas.0709698105

Supporting Text

SI Methods

Plant Material. Leaves of *Nephrolepis exaltata*, *Sphaeropteris cooperi (Cyathea cooperi)*, *Lotus corniculatus* L., *Medicago truncatula Gaertner* (var. *longeaculeata Urban*), and *Mesembryanthemum crystallinum* L. were gifts of the Gardenia (Helsinki, Finland) and of the University of Helsinki Botanic Garden. Leaves of *Prunus domestica* L., *Malus domestica Borkh.*, *Chaenomeles japonica Lindl. ex Spach*, *Rubus idaeus* L., cv. "*Muskoka*," *and* cv. "*Jatsi*," *Rosa hybrida*, *Rosa rugosa Thunb.*, and *Fragaria x ananassa Royer* were from the MTT Agrifood Research Finland collection. Seeds of *Glycine max* (L.) *Merr.* and heads of *Brassica oleracea* L. (cabbage) were purchased locally. *Solanum tuberosum* L. plants were a gift of Veli-Matti Rokka (MTT Agrifood Research Finland, Biotechnology and Food Research). Seeds of *Brassica napus* L., *B. napa*, and *Phleum pretense* were gifts of Boreal Plant Breeding Ltd.. Seeds of *Arabidopsis thaliana*, ecotype Columbia, were from the Arabidopsis Biological Resource Center, Ohio State University. Seeds for various members of the Poaceae (*Bromus sterilis*, *Agropyron cristatum*, *Amblyopyrum muticum*, *Australopyrum retrofractum*, *Australopyrum velutinum*, *Comopyrum comosum*, *Crithodium monococcum*, *Crithopsis delileana*, *Dasypyrum vilosum*, *Eremopyrum distans*, *Eremopyrum triticeum*, *Festucopsis serpentinii*, *Henrardia persica*, *Heteranthelium piliferum*, *Hordeum brachyantherum ssp. californicum*, *Hordeum erectifolium*, *Hordeum marinum ssp. Gussoneanum*, *Hordeum murinum ssp. Glaucum*, *Hordeum vulgare ssp. spontaneum*, *Lophopyrum elongatum*, *Peridictyon sanctum*, *Psathyrostachys fragilis ssp. fragilis*, *Psathyrostachys fragilis ssp. villosus*, *Psathyrostachys stoloniformis*, *Pseudoroegneria spicata*, *Taeniatherum caput-medusae*, *Trinopyrum bessarabicum*, *Elymus repens*, *Hordeum patagonicum*, *Aegilops speltoides var. speltoides*, *Aegilops tauschii var. meyeri*, *Triticum aestivum*, *Triticum durum*, *Secale strictum*, *Secale cereale*, *Oryza sativa*, *Avena sativa*, *Sorghum vulgare*) were gifts of Ole Seberg (Institute of Biology, University of Copenhagen). *Spartina alterniflora* was a gift of Malika Ainouche (University of Rennes). *Zea mays* was a gift of Pioneer Hi-Bred International.

Cloning of RNA Polymerase III Transcripts by RT-PCR. Total RNA was extracted from leaves (barley cv Kymppi), and treated with DNase I (DNA-free; Ambion). A RACE RNA adapter (taken from the FirstChoice® RLM-RACE kit, product 1700; Ambion) was ligated to the ends of the RNAs to allow amplification. The 50-μl ligation reaction contained: 1× T4 RNA ligase buffer [33 mM Tris-acetate (pH 7.8) (25°C), 66 mM sodium acetate, 10 mM MgCl$_2$, 0.5 mM DTT, 1 mM ATP], 50 units of RNasin, 5 mg of DNA-free total barley RNA, 5 mg of RACE RNA adaptor, 25 units of T4 RNA Ligase (Epicentre). The reaction was carried out for 1 hour at 37°C. The cDNA was synthesized by reverse transcription in a 50-μl reaction containing: 1× M-MuLV reverse-transcriptase buffer [50 mM Tris·HCl (pH 8.3) (25°C), 50 mM KCl, 4 mM MgCl2, 10 mM DTT, 1 mM dNTP], 50 units of RNasin, 20 μl of ligated RNA from the previous step, 6 μM 16-mer random primers, 500 units of RevertAid M-MuLV Reverse Transcriptase (Fermentas). The reaction was carried out for 1 hour at 42°C.

This was followed by two stages of nested RT-PCR. In the first stage, a 50-μl reaction mix was prepared, containing: 5 μl of cDNA from the previous step, 1× PCR buffer [75 mM Tris·HCl (pH 8.8), 20 mM (NH4)$_2$SO$_4$, 1.5 mM MgCl$_2$, 0.01% Tween-20], 200 μM dNTP, 2.7 units of *Taq* polymerase, 200 nM RNA-adaptor primer (RACE, 5′-GCTGATGGCGATGAATGAACACTG-3′), 20 nM LTR primer (for barley, primer A113: 5′-TGTAACGCCCCGGACACACC-3′, matching nucleotides 245–265 from the 5′ end of the *Cassandra* LTR). The amplification was carried out by using the following program: 94°C for 4 min, 30 cycles of 94°C for 40 sec, 55°C for 40 sec, 72°C for 30 sec and a final elongation at 72°C for 5 min. After the reaction, the mixture was diluted with 200 μl of TE to a final volume of 250 μl.

The second-stage RT-PCR mix of 50 μl contained: 5 μl from the diluted first-stage reaction, 1× PCR buffer [75 mM Tris·HCl (pH 8.8), 20 mM (NH$_4$)$_2$SO$_4$, 1.5 mM MgCl$_2$, 0.01% Tween-20], 200 μM dNTP, 2.7 units of *Taq* polymerase, 200 nM RNA adaptor primer (RACE, 5′-GCTGATGGCGATGAATGAACACTG-3′), 20 nM 5S RNA primer (for barley, primer 1,181: 5′-GGAGCAACTTCCCGGTCGGTCA-3′, located 129–151 nt from the 5′ end of the *Cassandra* LTR. The amplification was carried out with a program consisting of 94°C for 4 min; 25 cycles of 94°C for 40 sec, 55°C for 40 sec, 72°C for 30 sec, and a final elongation at 72°C for 5 min.

Determination of *Cassandra* 5S RNA Transcript Ends. Total barley RNA was treated with calf intestinal phosphatase (CIP) to remove the free 5′ phosphates from all noncapped RNA (rRNA, fragmented mRNA, tRNA) and remaining genomic DNA. The cap structure of mRNAs was then removed with tobacco acid pyrophosphatase (TAP) to produce a 5′ monophosphate. A 45-nt RNA adaptor oligonucleotide (supplied with the kit) was ligated to the mRNA by using T4 RNA ligase. The product was amplified with the one-step RT-PCR method as described above. The first round of amplifications was made with a primer complementary to the RNA adaptor (5′-GCTGATGGCGATGAATGAACACTG-3′) and a reverse primer matching the PBS motif of the *Cassandra* (5′-ACCGCGAGGGTCGGCTCTGATACCA-3′) under the conditions recommended by the manufacturer. The product was then diluted 10 times, and a second round of amplification carried out with the RACE primer and a reverse primer matching the LTR (5′-TTGTCCTCACTCATGCGCACC-3′) similar to the nested amplifications described above. The reaction was carried out as described above but with the following program: an initial denaturation at 94°C for 4 min, 25 cycles of 94°C for 40 sec, 57°C for 40 sec, and 72°C for 30 sec. Products were isolated from the gel, cloned into the pGEM-T vector, and sequenced.

EST Database Searches. The EST searches were run locally by Blastn on the EST_others database (GenBank 16.09.2007, 32, 565,588 sequences and 18,466,006, 241 nt), which includes nonmouse and nonhuman accessions. A cut-off e-value score of 0.001 was used. Taxonomy information for EST matches was derived from the taxdb databases, which were downloaded from ftp://ftp.ncbi.nih.gov/blast/db.

Prediction of Secondary Structure for *Cassandra* 5S RNA. The web interface of "alifold" (www.tbi.univie.ac.at/∼ivo/RNA/alifoldcgi.html) in the Vienna Package (www.tbi.univie.ac.at/∼ivo/RNA/) was used to visualize the consensus structure of both cellular and *Cassandra* 5S RNA. A total of 43 5S rRNAs from the 5S rRNA database (http://rose.man.poznan.pl/5SData/) were analyzed. The default settings of the Vienna package were used except for 17°C instead of 37°C as the folding temperature, a choice reflecting the plant origin of the sequences. The RNAz program of the Vienna package was used to test the probability of functionality of predicted secondary structures. Sequences were analyzed in the reverse orientation as a control.

Determination of Information Content in Cassandra 5S RNA. Two nonredundant libraries of DNA sequences were used for alignment by ClustalW software (www.ebi.ac.uk/Tools/clustalw/index.html) (1): 45 sequences of *Cassandra* 5s RNA and 11 sequences of cellular 5s rRNA. Sequence conservation of the DNA sequence was calculated as the information contribution of each base relative to its expected distribution, as previously (2). RNA secondary structures were predicted for each of the sequences with the Vienna package (3). The outputs were aligned by using the DNA multiple alignments as a guide. The conservation of RNA secondary structure at every position was computed as the information contribution of stem or loop relative to their expected distribution as before (2). The procedure was repeated for RNA structures predicted from randomized sequences shuffled along the length of each sequence. To compare the conservation of the RNA secondary structures of *Cassandra* to cellular 5S rRNA sequences, the two datasets of RNA secondary structure information content were plotted together (Fig. S3).

SI Results

Features of *Cassandra* Elements. We have assembled general features of the sequenced *Cassandra* elements in Table S2. Retrotransposon LTRs serve as promoters for transcription by RNA polymerase II and contain termination and polyadenylation signals. The long terminal inverted repeats (TIRs) in the Rosaceae and Brassicaceae are unusual; 5- to 6-nt TIRs are commonly found in retrotransposons such as *BARE1* (4). *Cassandra* integration generates target site duplications (TSDs) as does that of other retroelements (Table S5). The 5S RNA is not symmetrically situated within the LTR. In the ferns, which have the shortest LTRs found, the 5S RNA region is at the 3′ end of the LTR. In the flowering plants, it tends to be found closer to the 5′ end of the LTR.

5S RNA in *Cassandra* LTRs. The alignments for nucleotides 40–120 of the 5S RNA region (Fig. S2 and Table 3) show 77.5% (*Brassica oleracea*) to 90.8% (*Arabidopsis thaliana*) identity to the species' corresponding 5S rRNA but only 55.6–65.7% similarity when the entire 5S RNA region is aligned without gaps. The *Cassandra* elements show still higher conservation in the region corresponding to the RNA polymerase III promoter, which begins at nucleotide 51 of the alignment. Of the several types of RNA polymerase III promoters, which differ from the polymerase II promoters of most cellular genes by their position within the transcribed region, that of *Cassandra* resembles the 5S rRNA promoter rather than the tRNA promoters (5). The promoters for RNA polymerase III are multipartite. The A-Box (AGTTAAGCGTGC) is found at nucleotides 51–60, the Intermediate Element (GA) at nucleotides 71–72, and the C-Box (AGGATGGGTG) at nucleotides 81–91. In addition, the alignment shows a GGGAAGT motif between nucleotides 98 and 104 that is completely conserved between *Cassandra* and cellular 5S RNAs.

In *Arabidopsis*, additional motifs have been defined that are important for polymerase III transcription of 5S rRNA (6). A C at −1 nt from the *Arabidopsis* transcriptional start is conserved in all *Cassandra* LTRs 1 nt upstream of the aligned 5S RNA region excepting the three fern species we have investigated. A TATA box at nucleotides −23 to −28, which directs transcriptional reinitiation in *Arabidopsis* (6, 7), can be found at nucleotides 25–28 in the LTRs of *Cassandra* from the Rosaceae; the other families show A/T stretches at nucleotides −25 to −28 but not the canonical TATA. Alignments of the *Cassandra* LTRs, however, all fail to show the short motif previously described as the RNA polymerase III stop signal, a G/C pair, followed by four or more T nucleotides (6, 8).

***Cassandra* Transcription.** We used RT-PCR to detect *Cassandra* transcription by polymerase II (Fig. 1*C*), followed by sequencing of the products (data not shown). The products span from LTR to LTR. A 50-nt-shorter product represents either a well transcribed deletion or an artifact from secondary structure in the transcript. The RLM-RACE method was used with primers specific to the *Cassandra* 5S RNA domain to produce products specific to the polymerase III promoter internal to the 5S RNA domain. Sequencing enabled us to determine that these subgenomic transcripts from the TRIM element are present. The transcripts start specifically at an A nucleotide, unlike general 5S rRNAs. Furthermore, using an RLM-RACE RT-PCR method specific to the 5′ end structure, we were able to establish that the *Cassandra* subgenomic 5S RNAs are uncapped, as expected from the polymerase III promoter, and do not result from polymerase II transcription.

We also carried out EST database searches with the initial 40 nt of the 5S domain, which is specific to *Cassandra*, with the first 40 nt of the LTR, which excludes the 5S domain, and with the internal domain, which provides a discriminator of full-length *Cassandra* elements from *Cassandra* LTRs. Matching EST accessions (Table S4) include many that carry *Cassandra* full-length or solo-LTR insertions. This is unusual for retrotransposon insertions, particularly in the cereals, where retrotransposons tend to nest in intergenic regions (9–12). For comparison, our searches of the rice EST database with retrotransposon *Tos17* (accession no. D88393) of rice, which shows a strong preference towards genic sites for new insertions mobilized by tissue culture (13), reveals no elements inserted into cellular ESTs in normal nonmutagenized lines. We have also looked for insertions of Osr3–1 (AF458765), Osr38–1 (AF458766), Osr39–1 (AF458767), and Osr42–1 (AF458768), also low-copy elements of rice (14), and likewise found no insertions into cellular ESTs. Searches of the rice whole-genome database, annotated by compartment, yields a picture consistent with that from the EST database. A total of 58 *Cassandra* out of 352, or 16% of the total in the rice genome, are within genes, although only 4 of these, or 1%, are within exons.

Secondary Structure Predictions for *Cassandra* 5S RNA. The canonical structure of 5S rRNA is a T-shaped hairpin of five loops and five helices (15). However, the accuracy of predictions of this secondary structure for 5S rRNA from various sources can vary between 0 and 100%, with not all predicted structures resembling the canonical one (16, 17). We carried out structural predictions for each *Cassandra* 5S RNA region and 5S rRNA in our alignment (Fig. 3).

The prediction for the 5S RNA of *Zea mays*, *Triticum aestivum*, *Citrus limon*, and *Gingko biloba* matched the canonical structure, as did the *Cassandra* 5S RNA of members of the Roseaceae, *Chaenomeles japonica*, *Prunus domestica*, and *Rosa hybrida*. Alternative cellular structures were predicted for the other cellular 5S rRNA and *Cassandra*, and fell into several distinct groups. The *Cyathea cooperi Cassandra* 5S sequence resembled the 5S rRNA of *Metasequoia glyptostroboides*, *Spinacea oleracea*, and *Brassica napus*. The *Lactuca sativa* 5S rRNA was predicted to as a hairpin structure; similar forms were predicted for *Cassandra* products from *Avena sativa*, *Brassica oleracea*, *Spartina alterniflora*. The modeled products of many *Cassandra*, primarily from the Poaceae, fall into two internally highly similar groups, typified by *Oryza sativa* and *Hordeum vulgare* (Fig. 3). Although the accuracy of the models cannot be estimated without crystallographic data, at least some of the of the predicted *Cassandra* 5S RNA products do resemble canonical 5S rRNA structures.

***Cassandra* copy number in the sequenced rice genome.** Chromosome-sized contiguous regions of the rice genome (18), which are referred to as pseudochromosomes or pseudomolecules, have been assembled. These allow nonredundant searches for sequence similarities. The *Cassandra* LTR, internal region, and entire element were used as query strings to search the rice genome. The results (Table S6) show a total of 352 elements in the rice genome, of which 76% are solo LTRs. Of these, 16% are within genes. Of the intact elements, which are those that contain two LTRs and an internal domain, 21% are within genes, whereas 15% of the solo LTRs are found in this compartment. By comparison, in rice lines where *Tos17* has been induced to integrate at high levels (13), 11% of the insertions were within exons, and 9% within introns, for a total of 20% (70.4% of the insertion sites could not be characterized). The difference between the relative distribution into exons for *Tos17* and *Cassandra* can be correlated with the mutagenicity of *Tos17* and, in turn, with its somatic inactivity. Consistent with this, *Tos17* is found in only one to five copies in unmutagenized rice lines and cultivars, whereas *Cassandra* elements are in the hundreds. Hence, *Cassandra* is able to insert into noncoding regions of genes apparently without being mutagenic.

1. Thompson JD, Higgins DG, Gibson TJ (1994) *Nucleic Acids Res* 22:4673–4680.
2. Peleg O, Brunak S, Trifonov EN, Nevo E, Bolshoy A (2002) *AIDS Res Hum Retroviruses* 18:867–878.
3. Hofacker IL, *et al.* (1994) *Monatshefte Chemie* 125:167–188.
4. Suoniemi A, Schmidt D, Schulman AH (1997) *Genetica* 100:219–230.
5. Schramm L, Hernandez N (2002) *Genes Dev* 15:2593–2620.
6. Cloix C, Yukawa Y, Tutois S, Sugiura M, Tourmente S (2003) *Plant J* 35:251–261.
7. Yukawa Y, Sugita M, Choisne N, Small I, Sugiura M (2000) *Plant J* 22:439–447.
8. Hallenberg C, Frederiksen S (2001) *Biochim Biophys Acta* 1520:169–173.
9. Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P (2000) *Genome Res* 10:908–915.
10. SanMiguel P, Gaut BS, Tikhoniv A, Nakajima Y, Bennetzen JL (1998) *Nat Genet* 20:43–45.
11. Rostocks N, *et al.* (2002) *Funct Integr Genomics* 2:51–59.
12. Wei F, Wing,RA, Wise RP (2002) *Plant Cell* 14:1903–1917.
13. Miyao A, *et al.* (2003) *Plant Cell* 15:1771–1780.
14. McCarthy EM, Liu J, Lizhi G, McDonald JF (2002) *Genome Biol* 3:research0053.1–0053.11.
15. Szyman'ski M, Barciszewska MZ, Erdmann VA, Barciszewski J (2003) *Biochem J* 371(Pt 3):641–651.
16. Mathews DH (2005) *Bioinformatics* 21:2246–2253.
17. Mathews DH, *et al.* (2004) *Proc Natl Acad Sci USA* 101:7287–7292.
18. International Rice Genome Sequencing Project (2005) *Nature* 436:793–800.
19. Wicker T, *et al.* (2007) *Nat Rev Genet* 8:973–982.
20. Kalendar R, *et al.* (2004) *Genetics* 166:1437–1450.
21. Jiang N, *et al.* (2002) *Genetics* 161:1293–1305.
22. Jiang N, Jordan IK, Wessler SR (2002) *Plant Physiol* 130:1697–1705.

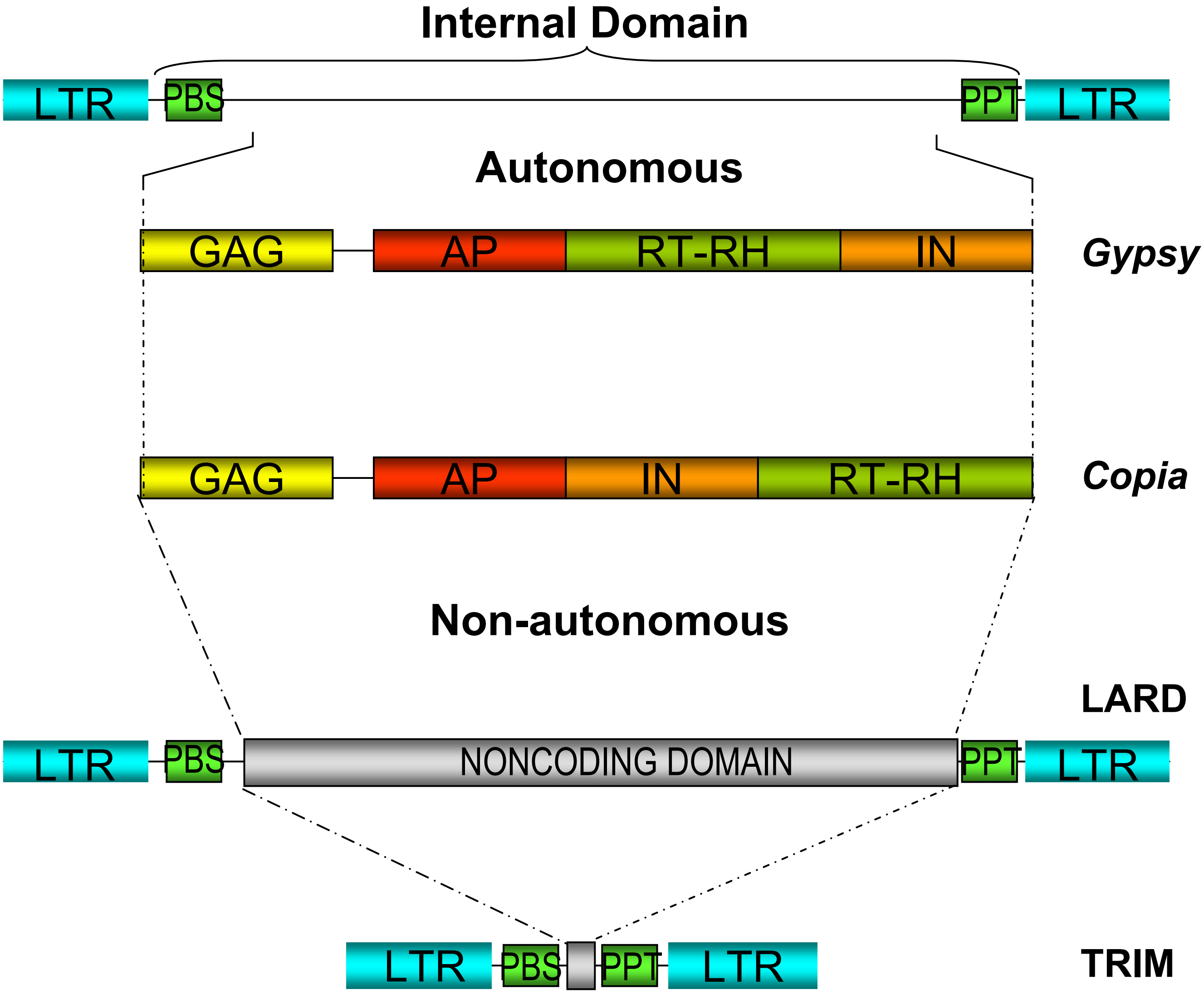

Fig. S1. Retrotransposon organization. (*A*) Retrotransposons of the *Copia* and *Gypsy* Superfamilies (19). The long terminal repeats (LTRs) are flanked internally by the primer binding site (PBS) (*Left*) and polypurine tract (PPT) (*Right*), which serve respectively as the priming sites for (−)-strand and (+)-strand cDNA synthesis by reverse transcriptase. Normally, two ORFs are encoded, one for Gag, the structural protein that forms the virus-like particle. The second encodes a polyprotein that is later cleaved into function domains specifying: AP, aspartic proteinase, which cleaves the polyprotein into functional units; IN, integrase, which inserts the cDNA copy into the genome; RT, reverse transcriptase, which copies the RNA transcript into cDNA; RNaseH, which digests the template during reverse transcription and thereby provides fragments that serve as the (+)-strand primer. The domain order differs in *Copia* and *Gypsy* elements. (*B*) Nonautonomous elements. LARDs contain long noncoding core domains that are conserved in secondary structure (20–22). TRIMs have very short LTRs and also short internal domains. *Cassandra* belongs to the TRIM group of retroelements.

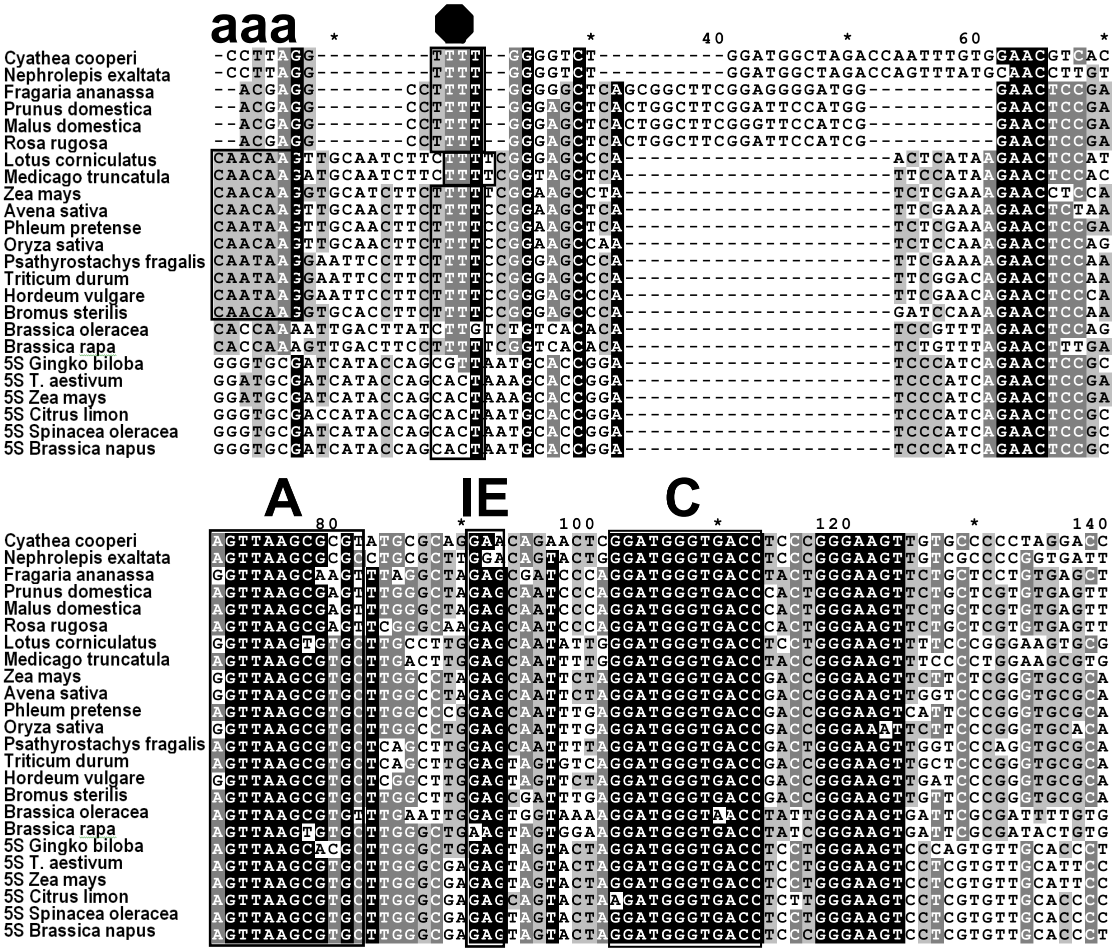

Fig. S2. Alignment of *Cassandra* 5S regions with cellular 5S rRNAs. Nucleotides are shaded according to the proportion that are identical at each position in the alignment: white on black, ≥90%; white on gray, ≥70%; black on gray, ≥50%; black on white, <50%. Elements that have been identified as important for transcription (5, 6, 8) are labeled: A-Box; IE, Intermediate Element; C-Box. The predicted pol III terminator is marked as a black octagon, and the putative polyadenlyation signal as ''aaa.'' This alignment was used for the phylogenetic prediction shown in Fig. 2. The accessions are organized by plant family.

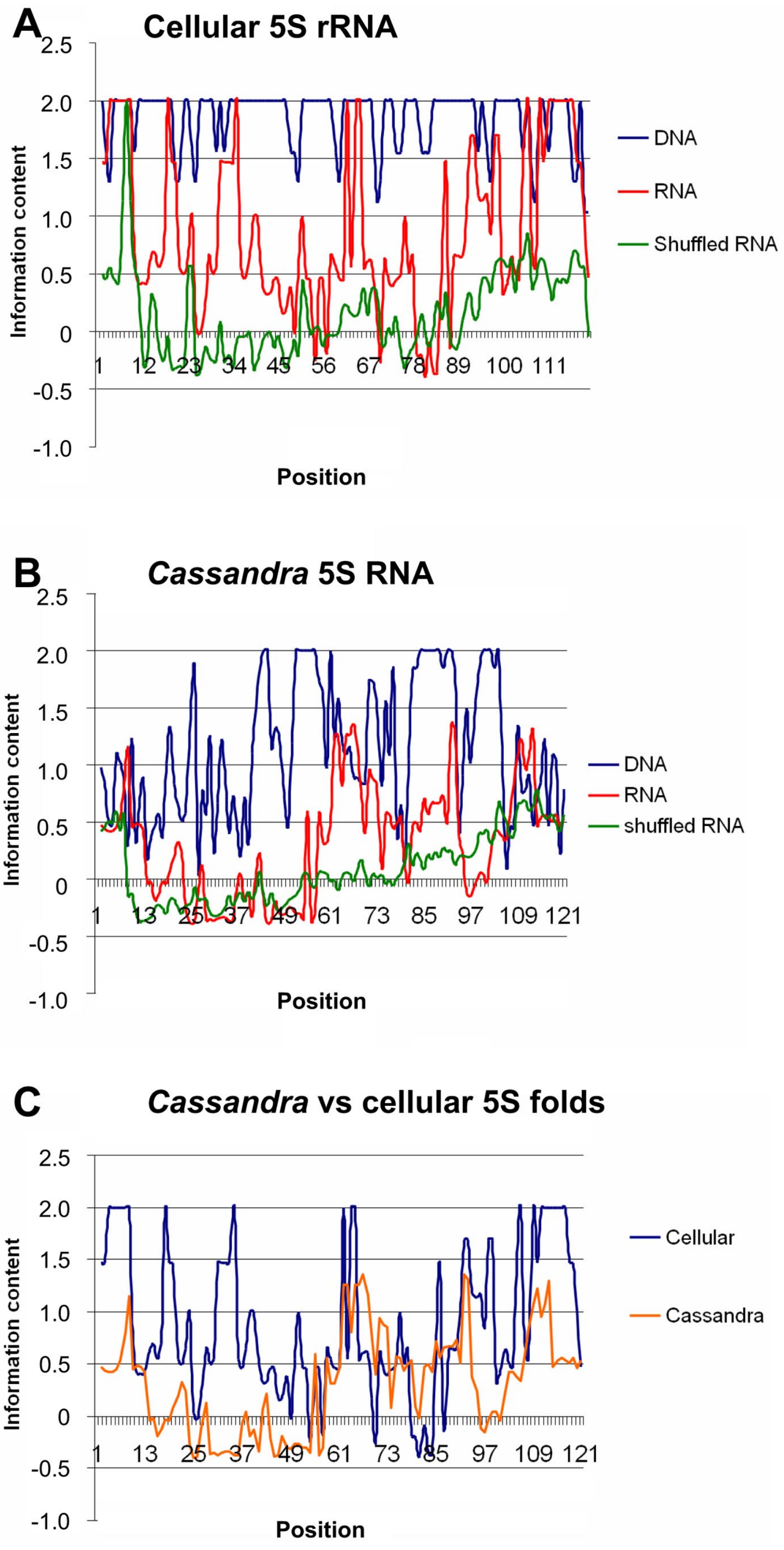

Fig. S3. Information content in *Cassandra* and cellular 5S RNA. Information content is plotted as a function of the position in alignments of *Cassandra* and cellular 5S sequences and in their predicted secondary structures. (*A*) Information content in cellular 5S rRNA sequences (DNA), secondary structures (RNA), and in the secondary structures after shuffling of each aligned sequence (shuffled RNA). (*B*) *Cassandra* 5S RNA. Details as for *A*. (*C*) Comparison at each aligned position of the information content in *Cassandra* and cellular 5S secondary structure.

Table S1. Plants investigated

- Division: Pteridophyta (ferns):
  - Class: Filicopsida
    - Order: Polypodiales
      - Family: Dryopteridaceae
        - *Didymochlaena truncatula*
        - *Nephrolepis exaltata*
      - Family: Cyatheaceae
        - *Sphaeropteris cooperi* (*Cyathea cooperi*)
- Division: Magnoliophyta (flowering plants)
  - Class: Magnoliopsida, Dicotyledons
    - Subclass: Dilleniidae
      - Order: Salicales
        - Family: Salicaceae-willow family
          - *Populus balsamifera* L. ssp. trichocarpa
          - *Populus tremula* L
    - Subclass: Magnoliids
      - Order Piperales
        - Family: Aristolochiaceae
          - *Saruma henryi* Oliv
    - Subclass: Rosidae
      - Order: Rosales
        - Family: Rosaceae
          - *Prunus domestica* L.
          - *Malus domestica Borkh.*
          - *Chaenomeles japonica Lindl. ex Spach*
          - *Rosa hybrida*
          - *Rosa rugosa Thunb.*
          - *Rubus idaeus* L.
            - cv. "*Muskoka*"
            - cv. "*Jatsi*"
          - *Fragaria x ananassa Royer*
      - Order: Linales
        - Family: Linaceae
          - *Linum usitatissimum* L.
      - Order: Fabales
        - Family: Fabaceae
          - *Lotus corniculatus* L.
          - *Medicago truncatula Gaertner*
          - *Glycine max* (L.) *Merr.*
          - *Pisum sativum* L.
          - *Robinia pseudoacacia* L.
    - Subclass: Asteridae
      - Order: Solanales
        - Family: Solanaceae (potato family)
          - *Solanum tuberosum* L.
    - Subclass: Dilleniidae
      - Order: Capparales
        - Family: Brassicaceae (mustard family)
          - *Brassica rapa* L.
          - *Brassica oleraceae* L.
          - *Arabidopsis thaliana*
      - Order: Ericales
        - Family: Ericaceae
          - *Vaccinium corymbosum* L.
    - Subclass: Caryophyllidae
      - Order: Caryophyllales
        - Family: Aizoaceae
          - *Mesembryanthemum crystallinum* L.
  - Class: Liliopsida, Monocotyledons
    - Subclass: Zingiberidae
      - Order: Zingiberales
        - *Zingiber officinale* Roscoe
    - Subclass: Commelinidae
      - Order: Cyperales
        - Family: Poaceae (grass family)

Bromus sterilis, *Agropyron cristatum*, *Amblyopyrum muticum*, *Australopyrum retrofractum*, *Australopyrum velutinum*, *Comopyrum comosum*, *Crithodium monococcum*, *Crithopsis delileana*, *Dasypyrum vilosum*, *Eremopyrum distans*, *Eremopyrum triticeum*, *Festucopsis serpentinii*, *Henrardia persica*, *Heteranthelium piliferum*, *Hordeum brachyantherum ssp. californicum*, *Hordeum erectifolium*, *Hordeum marinum ssp. Gussoneanum*, *Hordeum murinum ssp. Glaucum*, *Hordeum vulgare ssp. spontaneum*, *Lophopyrum elongatum*, *Peridictyon sanctum*, *Psathyrostachys fragilis ssp. fragilis*, *Psathyrostachys fragilis ssp. villosus*, *Psathyrostachys stoloniformis*, *Pseudoroegneria spicata*, *Taeniatherum caput-medusae*, *Trinopyrum bessarabicum*, *Elymus repens*, *Hordeum patagonicum*, *Aegilops speltoides var. speltoides*, *Aegilops tauschii var. meyeri*, *Triticum aestivum*, *Triticum durum*, *Secale strictum*, *Secale cereale*, *Spartina alterniflora*, *Oryza sativa*, *Avena sativa*, *Zea mays* var. B73 X MO17, *Sorghum vulgare*, *Eragrostis tef*, *Brachypodium distachion*, *Saccharum officinarum* L.

Plants from which *Cassandra* retrotransposons were identified in this study. The classification scheme is according to the "Plants Classification" online system maintained by the U.S. Department of Agriculture (http://plants.usda.gov/cgi_bin/topics.cgi?earl = plant_profile.cgi&symbol = BROL).

Table 2. Features of *Cassandra* retrotransposons

Species	Accession no.	LTR size, bp	5S rRNA location in LTR, nt	TIR 5′-TG. . . CA-3′		Internal domain, bp	Sequence between LTR and PBS	PPT sequence
Nephrolepis exaltata	AY860313	187	65–182	TGttgg	tttaCA	191	ACT	TTAAGGGGGCGAT
Didymochlaena trunculata	AY860311	208	86–204	TGttgg	cctaCA	190	AG	TTAAGGGGGCGAT
Cyathea cooperi	AY860310	208	86–204	TGttgg	cctaCA	190	AC	TTAAGGGGGCGGT
Populus trichocarpa	EF125877	169	?	TGtaatatccca	tggggtgttaCA	75	TT	TAAGGGGGATGGAT
Populus tremula	EF125876	202	?	TAtaagatccc	ggggtgttaca	44	TT	TAAGGTGGAGGGAT
Garcinia mangostana	EU140956	439	247–366	tgtaacacccg	ggactgttaca	71	GTT	AGAAGCTGGTGGGCA
Saruma henryi	EF125873	234	73–195	TGtggcgtccca	tggggtgttaca	95	AA	TAAGAGGGGGTGAT
Prunus domesticus	AY860314	270	72–192	TGtaacatccc	gggatgtgaCA	75	ATT	AAGGGGGGTGGAT
Malus domestica	AY603366, AY603367	306	78–199	TGtaacatccc	gggatgtgaCA	71	AAA,ATT	AAGGGGGCTAGAT
Chaenomoles japonica	AY860309	297	72–192	TGtgagatccc	cgagatgtgaCA	71	GTT,ATT	AAGGGGGGTGGAT
Rubus idaeus	AY860317	299	75–195	TGtgagatccc	gggatgtgaCA	71	ATT	AAGGGGGGTGGAT
Rosa rugosa	AY860316	300	75–195	TGtaacatccc	gggatgtgaCA	75	ATT	AAGGGGGGTGGAT
Rosa hybrid	AY860315	299	73–193	TGtgagatccca	tgggatgtgaCA	71	ATT	AAGRGRGGTGGAT
Fragaria ananassa	AY860312	267	72–192	TGtaatatccca	tgggatgtgaCA	75	AAT,AT	AAGGGGGGTGGAT
Linum usitatissimum	DQ767972	276	87. . .204	Tgtaatg	tgttaCA	80	AG	TATGGGGGG
Lotus corniculatus	AY603364, AY603365	394	205–324	TGtgacgccc	gggatgttaCA	70	AR	GGAAGTTGGTGGGCC
Medicago truncatula	AY603369	389	187–306	TGtaacaccc	gggatgttaCA	70	AA	GGAAGCTGGTGGGCA
Pisum sativum	DQ673669	421	220...340	TGtaacaccc	gggtgttaCA	71	AA	GGAAGCTGGTGGGCA
Robinia pseudoacacia	EF125871	?		?	tttca	?	?	?
Glycine max	EF125870	310	107–128	Tgtaa	ttaca	?	AAT	?
Solanum tuberosum		295	?	?	?	75	ATT	AAGGAGGGTGGAT
Brassica rapa	AY860308	350	62–182	TGtaacatccc	ggggcattaCA	104–105	AR	GAGAGGGGGTGAAT
Brassica olereaceae	AY860307	350	62–182	TGtaacatcct	agaacgttaCA	104–105	AR	GAGAGGGGGTGAAT, GAGAGATGGTGAAT
Arabidopsis thaliana	AY923749	356	65–185	TGtaacacccc	ggggcgttaCA	112	AA	TAGTGTGGGTGAAT
Vaccinium corymbosum	DQ788719	235	74...194	Tgtgac	gttaCA	180	GT	TAAGGTGGGGAGAA
Mesembryanthemum crystallinum	AY603370	299	67–186	TGtaatagccc	ggggtgttaCA	65	AA	GCCTGGTCAGCCC
Zingiber officinale	EF125874	309	69–191	Tgtaatacccc	ggggcgttaca	147	AG	TAAGTGCGGGTGAT
Sorghum bicolor	AF538605	290	75–193	TGcatcatc	gatgttaCA	143	AGT	ATGGCGGTGAGGA
Oryza sativa	AY271961, AF538611	281–296	49–168	TGtaacat	acgttaCA	202–206	AAA	AAGGGGGTGAGGG
Zea mays	AY271958, AY271959, AF538618	280–306	72–191	TGtgata	tgttaCA	197–202	AR	AAGGGGGGTGGAT
Saccharum officinarum	EF125872	?		?	ttaca	?	?	?
Spartina alterniflora	AY603377	253–279	45–164	TGtaacatc	gatgttaCA	258–260	AAAG,GA	AAGGGGGGTGGAT
Avena sativa	AY271960	262–287	49–169	TGtgagacc	ggtgttaCA	191–200	AAT,AAA,AG	AAGGGTGGGTGTA
Bromus sterilis	AY271957	266	42–162	TGtgacat	ggcgttaCA	157	AA	GAGTGGGGGTGTA
Phleum pratense	AF538603, AF538607-AF538610, AF538612-AF538617	280–285	49–169	TGtgata	tgttaCA	163–167	ATT,TT	AAGGGGGGAAGGA
Peridictyon sanctum	AY603376	267–289	42–162	TGtgacatcc	ggcgttaCA	185–196	ATT,AT	GAGTGGGGGTGTA
Brachypodium distachion	DQ094839-DQ094843	247–280	32...152	TGtr	gttaCA	154–169	GAT, AT, AAT	AAGGGGGGGTGTG
Eragrostis tef		280		?	gggacgctaCA	230	G	
Secale cereale	AY359471	264	42–162	TGtga	ttaCA	196	AGT,AAA	GAGTGGGGGTGTA
Amblyopyrum muticum	AY603371	264	42–162	TGtga	ttaCA	196	AGT,AAT,ATT	GAGTGGGGGTGTA

Species	Accession no.	LTR size, bp	5S rRNA location in LTR, nt	TIR 5′-TG. . . CA-3′		Internal domain, bp	Sequence between LTR and PBS	PPT sequence
Eremopyrum distans	AY603372	244–264	42–162	TGtga	ttaCA	198	AGT,AG	TGAGTGGGGGTGT
Psathyrostachys fragilis	AY271962	268	42–162	TGtga	ttaCA	195	AGT	GAGTGGGGGTGTA
Triticum aestivum	AY271963	267–272	42–162	TGtga	ttaCA	195–196	AGT	GAGTGGGGGTGTA
Henrardia persica	AY603374	264	42–162	TGtga	ttaCA	198	AGT	AAGTGGGGGTGTA
Hordeum vulgare	AY164585	240–264	42–162	TGtga	ttaCA	196–197	AGTT,AGT	GAGTGGTGGTGTA
Hordeum marinum	AY603375	264	42–162	TGtga	ttaCA	196	AGT,AT	GAGTGGCGGTGTA
Hordeum brachyantherum	AY603373	264	42–162	TGtgatag	cattaCA	196	AGT,AT,AA	AGGTGGGGGTGTA

Properties of the cloned *Cassandra* elements from seed plants. A range of lengths is given for the long terminal repeat (LTR) where database sequences or cloned examples varied. The nucleotide at which the 5S RNA domain begins and ends is listed. The terminal inverted repeat (TIR) refers to the ends of the LTR. The internal domain includes the primer binding site (PBS) and polypurine tract (PPT) as well as the intervening segment. The question mark denotes current lack of a complete sequence.

Table 3. *Cassandra* and 5S RNA comparisons

	Aligned segment of 5S rRNA region, %	
Genus and species	Entire, ungapped	Nucleotides 40–100
Hordeum vulgare	61.7	85.4
Hordeum marinum	61.7	85.4
Henrardia persica	59.6	82.1
Secale cereale	60.2	83.3
Amblyopyrum muticum	60.8	84.6
Triticum durum	60.4	83.8
Eremopyrum distans	57.7	79.2
Psathyrostachys fragalis	59.8	83.3
Hordeum brachyantherum	60.2	85.4
Bromus sterilis	59.8	81.7
Peridictyon sanctum	57.3	79.2
Phleum pratense	62.9	83.8
Avena sativa	61.7	85.8
Zea mays	61.8	84.2
Oryza sativa	59.8	80.8
Spartina alterniflora	57.7	80
Nephrolepis exaltata	56.5	78.3
Didymochlaena truncala	56.7	80.4
Cyathea cooperi	55.8	80.8
Rubus idaeus	63.3	80.8
Rosa hybrid	65.6	83.3
Chaenomelesomeles	67.7	85.4
Malus domestica	66.3	82.5
Rosa rugosa	66.7	83.3
Prunus domestica	64.2	84.2
Fragaria ananassa	60	78.8
Arabidopsis thaliana	61.7	90.8
Brassica oleracea	55.6	77.5
Brassica rapa	58.5	84.6
Mesembryanthemum	64.8	87.9
Medicago truncatula	61.7	83.8
Lotus corniculatus	61.9	82.5

Cassandra and 5S RNA comparisons. Pairwise comparisons between the 5S RNA regions of *Cassandra* (Table S2) and the corresponding 5S rRNA for these species. The similarity of the alignment over the entire length is lower than that over nucleotides 40–100 due to *Cassandra*- and species-specific segments at the 5′ end of this region.

Table 4. *Cassandra* EST matches

Query source	EST source	Matches			Library types	Best e-value	Total EST accessions
		LTR	5S	Core			
Amblyopyrum muticum	*Hordeum vulgare*	9	9	1	Adult top leaves, seedling shoot (dehydration stress), embryo + scutellum, callus, leaf epidermis, adult top leaves	8 E-05	461874
	Triticum aestivum		6		Late flowering spikelet, shoot grown with desiccation, grain, embryos 14 DAP, developing inflorescence, infected leaf	5 E-06	1049875
Avena sativa	*Eragrostis tef*			1	3-week old seedling	4 E-05	2816
	Hordeum vulgare			1	Leaf epidermis	1 E-05	547805
	Oryza sativa (indica)		3		Whole-life-cycle cDNA library, infected leaf	5 E-06	172331
	Oryza sativa (japonica)		11	3	3rd week immature panicle, 100 ppm $ZnSO_4$ stressed callus, stressed seedling, 100 ppm $ZnSO_4$ stressed callus, mixed callus and mixed shoot, infected leaf anther (2-nuclei stage), infected leaf, anther (2-nuclei stage), infected leaf	5 E-06	977811
	Sorghum bicolor		1		Wounded leaves	3 E-04	204308
	Triticum aestivum			4	Shoot with ABA treatment, infected seedling	7 E-07	1049875
	Zea mays			3	Ear leaf	4 E-05	1159264
Arabidopsis thaliana	*Arabidopsis thaliana*		3	19	Pooled infected and stressed libraries	2 E-08	1276692
Brassica oleracea	*Raphanus sativus*			1	Whole seedling (with 1 set of true leaves), buds, and anthers	1 E-09	17770
Brassica rapa	*Brassica napus*	2			Not specified	2 E-05	567177
	Raphanus sativus			1	Whole seedling (with 1 set of true leaves), buds, and anthers	1 E-09	17770
Bromus sterilis	*Eragrostis tef*			1	3-week old seedling	9 E-09	2816
	Hordeum vulgare			1	Leaf epidermis	1 E-07	547805
	Oryza sativa (indica)		1		Whole-life-cycle cDNA library	8 E-05	172331
	Oryza sativa (japonica)		10	1	3rd week immature panicle, stressed seedling, 100 ppm $ZnSO_4$ stressed callus, mixed callus and mixed shoot, infected leaf	8 E-05	977811
	Triticum aestivum	2		2	Shoot with ABA treatment, late flowering spikelet, root	3 E-05	1049875
	Zea mays		1	1	Vegetative Shoot Apical Meristem (SAM) and leaf primordia staged P1-P4, mixed (silks, husks, ears, pollen, shoot tips, leaf, root tips, whole seed, embryo)	8 E-05	1159264
Eremopyrum distans	*Hordeum vulgare*	8	9	1	Adult top leaves, callus, leaf epidermis, Seedling shoot (dehydration stress), embryo + scutellum	8 E-05	547805
	Triticum aestivum		6		Late flowering spikelet, shoot grown with desiccation, grain, embryos 14 DAP, developing inflorescence, infected leaf	5 E-06	1049875
Fragaria ananassa	*Fragaria vesca*		1		Stressed seedlings	3 E-13	10945
	Malus x domestica			1	Fruit (seeds removed)	8 E-04	255091
	Prunus persica			6	Shoot	2 E-20	70972
Hordeum brachyantherum	*Hordeum vulgare*	7	4	1	Seedling shoot (dehydration stress),embryo + scutellum, adult top leaves, callus	1 E-09	547805
	Triticum aestivum	1	4	2	LATE flowering spikelet, infected seedling, root	8 E-05	1049875
Hordeum marinum	*Hordeum vulgare*	9	9	1	Adult top leaves, Seedling shoot (dehydration stress), embryo + scutellum, callus, leaf epidermis	8 E-05	547805
	Triticum aestivum		6		Late flowering spikelet, shoot grown with desiccation, grain, embryos 14 DAP, developing inflorescence, infected leaf	5 E-06	1049875
Henrardia persica	*Hordeum vulgare*	9	9	1	Seedling shoot (dehydration stress), embryo + scutellum, adult top leaves	8 E-05	547805
	Triticum aestivum		6	2	Late flowering spikelet, shoot grown with desiccation, grain, embryos 14 DAP, developing inflorescence, infected leaf, shoot with ABA treatment	5 E-06	1049875
Hordeum vulgare	*Hordeum vulgare*	9	3	1	Seedling shoot (dehydration stress),leaf (14 days old), leaf epidermis, embryo + scutellum, adult top leaves	2 E-08	461874
	Triticum aestivum			2	Shoot with ABA treatment	7 E-04	1049875

		Matches					
Query source	EST source	LTR	5S	Core	Library types	Best e-value	Total EST accessions
Oryza sativa	*Eragrostis tef*			1	3-week old seedling	2 E-32	2816
	Oryza sativa (indica)	5	4	1	Infected leaf, panicle, whole-life-cycle cDNA library	3 E-13	172331
	Oryza sativa (japonica)	18	21	6	3rd week immature panicle, stressed seedling, 100 ppm $ZnSO_4$ stressed callus, mixed callus and mixed shoot, gamma-irradiated(45Gy), 100 ppm $ZnSO_4$ stressed callus, leaf, anther (2-nuclei stage), mixed shoot, infected leaf, ABF3-overexpressing transgenic, UVB irradiated callus	3 E-13	977811
	Zea mays		1	14	Vegetative Shoot Apical Meristem (SAM) and leaf primordia staged P1-P4, ear leaf, pericarp	3 E-04	1159264
Psathyrostachys fragilis	*Hordeum vulgare*	8	4	1	Seedling shoot (dehydration stress), callus, leaf epidermis, embryo + scutellum, adult top leaves,	3 E-13	547805
	Triticum aestivum		1		Late flowering spikelet	5 E-06	1049875
Phleum pratense	*Oryza sativa (indica)*	7	2	1	Whole-life-cycle cDNA library, infected leaf, panicle	3 E-10	172331
	Oryza sativa (japonica)		19		3rd week immature panicle, stressed seedling, 100 ppm $ZnSO_4$ stressed callus, mixed callus and mixed shoot, infected leaf, gamma-irradiated(45Gy), 100 ppm ZnSO4 stressed callus, uninfected leaf, anther (2-nuclei stage), ABF3-overexpressing transgenic	3 E-10	977811
	Zea mays			1	Vegetative Shoot Apical Meristem (SAM) and leaf primordia staged P1-P4	4 E-05	1159264
Peridictyon sanctum	*Eragrostis tef*			1	3-week old seedling	2 E-04	2816
	Hordeum vulgare			1	Leaf epidermis	3 E-06	547805
	Triticum aestivum	2	3	2	Infected seedling, root, late flowering spikelet	3 E-07	1049875
	Zea mays		1		Vegetative Shoot Apical Meristem (SAM) and leaf primordia staged P1-P4	2 E-05	1159264
Rosa rugosa	*Malus x domestica*		16	23	Young root, young fruit, fruit (seeds removed), Partially senescing leaves, young fruit, young expanding leaf	1 E-09	255091
	Prunus dulcis		1		Developing seed	2 E-05	3864
	Prunus persica		4	5	Shoot	5 E-09	70972
Spartina alterniflora	*Eragrostis tef*			1	3-week old seedling	9 E-07	2816
	Hordeum vulgare			1	Leaf epidermis	6 E-05	85931
	Oryza sativa (indica)		2		Whole-life-cycle cDNA library, infected leaf	5 E-06	172331
	Oryza sativa (japonica)		11	4	3rd week immature panicle, stressed seedling, 100 ppm $ZnSO_4$ stressed callus, mixed callus and mixed shoot, infected leaf, 100 ppm $ZnSO_4$ stressed callus, anther (2-nuclei stage)	5 E-06	977811
	Zea mays			4	Apex, mixed (silks, husks, ears, pollen, shoot tips, leaf, root tips, whole seed, embryo), Vegetative Shoot Apical Meristem (SAM) and leaf primordia staged P1-P4	6 E-11	1159264
Sorghum bicolor	*Eragrostis tef*			1	3-week old seedling	1 E-11	2816
	Oryza sativa (indica)		1	1	Infected leaf	8 E-05	172331
	Oryza sativa (japonica)			5	Anther (2-nuclei stage), infected leaf	8 E-05	977811
	Zea mays			4	Mixed (silks, husks, ears, pollen, shoot tips, leaf, root tips, whole seed, embryo), ear leaf, apex	7 E-07	1159264
Secale cereale	*Hordeum vulgare*	9	4	1	Seedling shoot (dehydration stress), callus, leaf epidermis, embryo + scutellum,adult top leaves	3 E-13	461874
	Triticum aestivum		1	2	Late flowering spikelet, shoot with ABA treatment	5 E-06	1049875
Triticum durum	*Hordeum vulgare*	8	4	1	Seedling shoot (dehydration stress), leaf epidermis, embryo + scutellum, adult top leaves, callus	3 E-13	461874
	Triticum aestivum		1	2	Late flowering spikelet, shoot with ABA treatment	5 E-06	1049875
Zea mays	*Eragrostis tef*			2	3-week old seedling	1 E-11	2816
	Hordeum vulgare			1	Leaf epidermis	4 E-05	85931
	Mesembryanthemum crystallinum		4		Leaf	8 E-05	27348
	Oryza sativa (japonica)			4	Anther (2-nuclei stage), infected leaf	3 E-06	977811
	Zea mays	4	7	29	Vegetative Shoot Apical Meristem (SAM) and leaf primordia staged P1-P4, shoot apical meristem, sperm, whole plant, apex, mixed (silks, husks, ears, pollen, shoot tips, leaf, root tips, whole seed, embryo)	3 E-10	1159264

EST database matches to *Cassandra* 5S regions. The EST-other section of the GenBank DNA database was searched by BLASTN for matches to the LTRs, internal domains, and 5S RNA regions of *Cassandra*. Queries were made with sequenced *Cassandra* for which large-scale EST projects have deposited significant numbers of accessions. The search strings consisted of the first 40 nt of the LTR, the first 40 nt of the 5S rRNA region, and the core domain. The first 40 nt of the 5S region differs from the cellular 5S rRNAs; a search with the region 40 to 120 nt could also match cellular 5S transcripts. The table shows the origin of the query sequences, the source of the matching ESTs, and the type of tissue from which the libraries containing matches were made. The total number of matches having scores below 10^{-4} (indicating the probability of random occurrence at 1 in 10^4 sequences) is listed as well as the best scores for the matches above this limit.

[a] GenBank release of September 16, 2007.

[b] Accession no. AJ475506.

Table 5. Target site duplications (TSDs)

Species	Accession no.	TSD	Mismatch	Element type
Hordeum vulgare	AF474072	gagcg/gagcc	1	Whole
	HVCH4	tgaag/tgaag	0	Truncated
Medicago trunculata	AC140033	ctatt/ctatt	0	Solo LTR
	AC142526	atttc/atttc	0	Solo LTR
	AC144516	gaaga/caaga	1	Solo LTR
	AC144375	gaaat/aacat	2	Solo LTR
	AC144516	tttat/atttt	2	Solo LTR
	AC140032	atgac/attct	3	Whole
Lotus corniculatus var. japonicus	AP006095	tcccc/tcccc	0	Whole
	AP004483	atgct/tccct	3	Whole
	AP006403	aagga/aagga	0	Two nested solo LTRs
	AP004956	ggtag/ggtag	0	Whole
	AP006691	gttgg/gttgg	0	Solo LTR
	AP006143	gttac/gttac	0	Solo LTR
	AP004895	gaatc/gaatc	0	Solo LTR
		aggcc/ctgaa	4	Whole
	AP006076	tgatg/tgatg	0	Whole
Mesembryanthum crystallinum	AF326722	aaata/ataag	3	Solo LTR
Sorghum bicolor	AY542311	agtaa/agaaa	1	Whole
Zea mays	AF061282	tctac/tctac	0	Whole
	AF546188	tagga/tagga	0	Whole
	AY455286	cctta/cctta	0	Whole
Arabidopsis thaliana	AC007178	atctg/aggtg	2	Whole
	AC006161	ccagt/tcagt	1	Whole
	ATF26B15	acttt/actag	2	Two nested solo LTRs

Target site duplications (TSDs). The TSDs generated by repair of the staggered cuts made by integrase during insertion of retrotransposons are generally 5 nt in length in plants and perfect upon insertion of the element. The TSDs of available genomic sequences reveal many subsequent mutation events. Lotus shows many solo LTRs that retain perfect repeats, indicating a relatively rapid recombinational loss of the internal domains of the elements. *Arabidopsis* data are only as an example; the total population of *Arabidopsis* elements is undergoing separate analysis.

Table 6. *Cassandra* in the rice genome.

Chromosome	Segment	Total	Complete	Solo LTRs
1	Intergenic	32	8	24
	Genic	7	2	5
	Cds	2	0	2
	Total	39	10	29
2	Intergenic	21	1	20
	Genic	4	2	2
	Cds	1	0	1
	Total	25	3	22
3	Intergenic	38	7	31
	Genic	11	8	3
	Cds	0	0	0
	Total	49	15	34
4	Intergenic	32	9	23
	Genic	7	2	5
	Cds	0	0	0
	Total	39	11	28
5	Intergenic	14	5	9
	Genic	1	0	1
	Cds	0	0	0
	Total	15	5	10
6	Intergenic	29	8	21
	Genic	3	1	2
	Cds	1	0	1
	Total	32	9	23
7	Intergenic	24	9	15
	Genic	6	0	6
	Cds	0	0	0
	Total	30	9	21
8	Intergenic	27	6	21
	Genic	3	0	3
	Cds	0	0	0
	Total	30	6	24
9	Intergenic	12	1	11
	Genic	1	0	1
	Cds	0	0	0
	Total	13	1	12
10	Intergenic	22	5	17
	Genic	6	1	5
	Cds	0	0	0
	Total	28	6	22
11	Intergenic	23	6	17
	Genic	5	1	4
	Cds	0	0	0
	Total	28	7	21
12	Intergenic	20	1	19
	Genic	4	1	3
	Cds	0	0	0
	Total	24	2	22
All	Intergenic	294	66	228
	Genic	58	18	40
	Cds	4	0	4
	Total	352	84	268

Cassandra in the rice genome. Searches were carried out on the rice pseudomolecules and the compartments thereof, as annotated by TIGR. Complete elements refer to matches long enough to include two LTRs and an internal domain, and at least 500 nt, whereas solo LTRs are matches to LTR query sequences of >100 nt. Elements with additional, small insertions and deletions with respect to the consensus element are therefore included in each set. AIntergenic@ refers to areas between genic regions, Genic to gene sequences, including exons, introns, and upstream or downstream untranslated regions. The Cds regions are the protein-coding nucleotide sequence of the gene model used by the annotation system, and excludes introns and untranslated regions. The total number of Cassandra LTRs is the sum of twice the number of complete elements and the number of solo LTRs.