D66–D70 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 28 October 2006*
doi:10.1093/nar/gkl780

RetrOryza: a database of the rice LTR-retrotransposons

Cristian Chaparro, Romain Guyot, Andrea Zuccolo[1], Benoît Piégu and Olivier Panaud*

Laboratoire Génome et Développement des Plantes, UMR 5096 CNRS-IRD-Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan, France and [1]Department of Plant Sciences, Arizona Genomics Institute, University of Arizona, Tucson, AZ 85721, USA

Received August 14, 2006; Revised September 5, 2006; Accepted October 1, 2006

ABSTRACT

Long terminal repeat (LTR)-retrotransposons comprise a significant portion of the rice genome. Their complete characterization is thus necessary if the sequenced genome is to be annotated correctly. In addition, because LTR-retrotransposons can influence the expression of neighboring genes, the complete identification of these elements in the rice genome is essential in order to study their putative functional interactions with the plant genes. The aims of the database are to (i) Assemble a comprehensive dataset of LTR-retrotransposons that includes not only abundant elements, but also low copy number elements. (ii) Provide an interface to efficiently access the resources stored in the database. This interface should also allow the community to annotate these elements. (iii) Provide a means for identifying LTR-retrotransposons inserted near genes. Here we present the results, where 242 complete LTR-retrotransposons have been structurally and functionally annotated. A web interface to the database has been made available (http://www.retroryza.org/), through which the user can annotate a sequence or search for LTR-retrotransposons in the neighborhood of a gene of interest.

INTRODUCTION

Transposable elements (TEs) are ubiquitous in all eukaryotic genomes. A particular class of TE, the long terminal repeat (LTR)-retrotransposon, is the main component of large plant genomes (1). These elements transpose via mRNAs through a copy/paste mechanism and therefore have a propensity to increase their copy number while active. In some cases, the copy number of a given active element multiplies to such an extent that it causes genomic expansion (2). LTR-retrotransposons have thus been shown to be a main contributor to genome size variation in plants (3). In this regard, large plant genomes, such as that of maize (2500 Mb) or that of bread wheat (17 000 Mb) contain over 50% LTR-retrotransposons. On the functional side, LTR-retrotransposons harbor their own internal promoter as well as some regions with a strong enhancer activity (both located in the 5′ LTR of the element). Recent reports have shown that these regulatory sequences can have an impact on the expression of nearby genes (4). Therefore, the identification and characterization of LTR-retrotransposons has become a priority in crop species genome sequencing projects. Compared to other agronomically important crops, rice has a relatively small genome (380 Mb). However, we estimate that >20% of the genome (i.e. ~80 Mb) is composed of LTR-retrotransposons. The complete genomic sequence of rice has been available for more than a year and its annotation by the international consortium is ongoing (5). The annotation of all rice TEs and, in particular, of the LTR-retrotransposons is cumbersome, because it requires a knowledge of all the various families found in the genome. Although some repetitive sequence databases exist for rice, e.g. the TIGR rice repeat database (6) and the Repbase database (7), there is no resource available for the characterization of low copy number elements. The full characterization of these low copy number elements is critical since they are often mistakenly identified as plant genes by the annotation pipelines implemented in genome sequencing projects therefore leading to an overestimation of gene number (8).

One of the aims of the retrOryza database is thus to provide the community with free access to the most complete dataset of rice LTR-retrotransposons. Another goal of the retrOryza database is to provide the user with a resource for studying potential gene/LTR-retrotransposons interactions. We provide a tool in the interface that allows the identification of LTR-retrotransposons in the vicinity of any of the rice genes annotated in the TIGR Nipponbare pseudomolecules. In addition, we provide an interface that allows one to identify and annotate LTR-retrotransposons in any genomic sequence. The user can generate such annotation by uploading the sequence in a query form. The output consists of a GFF formatted file that can be visualized using a text editor or an annotation browser like Artemis (9).

*To whom correspondence should be addressed. Tel: +33 468 661773; Fax: +33 468 668499; Email: panaud@univ-perp.fr

DESCRIPTION OF THE LTR-RETROTRANSPOSON DATABASE

Construction of the database

Our database consists of 242 retrotransposon reference molecules from rice (*Oryza sativa* ssp. *japonica*), representing previously described molecules as well as a significant number of new molecules. Forty-eight retrotransposon sequences, already characterized and published, were collected from GenBank. They include those described by McCarthy *et al.* (10) (34 *osr* sequences), *Retrosat1*, *RIRE1* (11) *RIRE2* (12), *RIRE3* and *RIRE8* (13), *RIRE7* (14), *RIRE9* (15), *RIRE10* (16), *Dagul*, *Hopi* and *Houba* (17), *Dasheng* (18), *Spip* and *Squiq* (19). All 242 molecules were submitted to an all-against-all BLAST to identify redundant sequences and contamination due to the insertion of other TEs. RepeatMasker was used to identify similarities to retrotransposons collected in RepBase (7) and to mask repetitive sequences.

Based on the results of the above described process, several of the previously published sequences (*osr9*, *osr10*, *osr23*, *RIRE5*, *RIRE7* and *RIRE10*) have been replaced by complete elements taken from genomic sequences with intact target site duplications (TSDs) and resubmitted to the cleaning process. The remaining 194 new retrotransposons were classified according to their similarity with the already described retrotransposons. In this way, the elements that possess >90% overall similarity, at the nucleotide level, to already described molecules are considered as synonyms. When the sequence identity is between 70 and 90%, the sequence is considered to be related to the already described molecule. Any sequence presenting <70% of sequence identity with a described sequence is considered as a new retrotransposon. The 194 elements, which correspond mainly to low copy number retrotransposons, have been classified accordingly generating 24 synonymous sequences, 25 sequences related to already described retrotransposons and 145 new sequences.

Annotation of the LTR-retrotransposon reference molecules

All the elements have been annotated for both structural and functional features. The LTRs were automatically identified by using a combined approach in which the candidate sequence is split into two and both halves submitted to sim4 alignment (20). If no satisfactory result was found, a Smith and Waterman alignment was performed using the Water software from the EMBOSS (21) to identify the correct borders of the LTRs. This approach significantly reduced the computation time needed to process each molecule, as the more time-consuming dynamic programming algorithm (Water) was used only when difficult cases were found.

The primer binding site (PBS), usually located downstream of the 5′ LTR, was identified by similarity searches [NCBI's blast tool (22)] against a local database of *O.sativa* ssp. *japonica* tRNA sequences, produced using tRNAscan-SE (23). Data from similarities between retrotransposons and 3′ end of tRNA were automatically evaluated and integrated into the database. The poly purine tract (PPT) is located immediately upstream of the 3′ LTR. To determine the size of the PPT we did a nucleotide frequency analysis of the 15 bases upstream of the 3′ LTR. We found that purines were overrepresented from bases −9 to −3 where adenine and guanine presented a frequency of ∼20 and 60%, respectively. Base −2 showed a frequency of 54% for adenine followed by 27% for guanine while the highest frequency for base −1 was for thymine with a 32% occurrence.

Annotation of the polyprotein genes was performed using similarity searches against a local protein databases composed of polyproteins collected from GenBank. We performed a TBLASTX (22) analysis using the retrotransposon sequences as queries against this database. The results were filtered and the most significant hit was kept. We used the PBS, PPT and the polyprotein annotations to correct the orientation of the LTR-retrotransposon reference molecule.

LTR-retrotransposon annotation in genomic sequences

Most of the tools to annotate retrotransposons are aimed at masking genomic sequences. Although this approach will identify pieces of retrotransposons, it does not attempt to search for complete elements. On the other hand, BLAST proved not to suit our needs because it produces many high-scoring segment pairs (HSPs) but does not provide a mechanism to identify the whole element, especially if there are big indels. This motivated us to develop a tool that would take advantage of the annotation effort that had been carried out and therefore from our database of reference elements.

We developed an algorithm to search for retrotransposons in the genomic sequences, which is based on BLAST and uses the reference molecules as queries against the TIGR pseudomolecules which have previously been cut in 300 kb contiguous sequences with a 30 kb overlap. The results are filtered and the original coordinates are assigned to each HSP. The next step groups the HSPs to identify the retrotransposons. The algorithm will include an HSP in an existing group if the addition of the matching sequence improves the representation of the reference molecule attached to the region, if not, a new group is created. The process is iterated

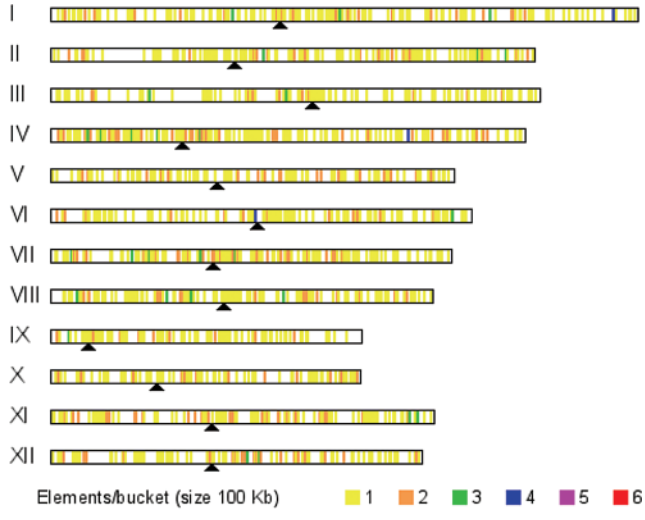

Figure 1. The 12 pseudomolecules of rice are represented with the centromere position indicated by a black triangle. Each bucket represents 100 Mb of sequence and the number of hits per bucket is color coded as indicated in the legend.

until all HSPs are assigned to a group or to its own single hit group. TSDs are then searched by extracting six bases flanking the putative element and the matching bases are counted. The TSDs that match over the first five bases are kept.

The parameters that control the grouping and the presentation of the results are the degree of similarity of the HSP, the minimum length of the HSP to include, and the minimum size of the grouped molecule to keep. We have used a minimum HSP length of 50 bases, 80% similarity and kept the groups that were >7% of the size of the complete reference retrotransposon represented by the group. With these settings, we found 19 908 groups of which 2433 represent complete elements whereas 5990 represent regions that span both LTRs and the internal region but lack well-defined borders. We also found 2886 truncated elements, composed of an LTR and part of the internal region. Solo LTRs account for 738 groups while 1348 correspond to partial LTRs. Furthermore, we found 5792 groups that represent the internal region but without flanking LTRs. Although we include some LArge Retrotransposon Derivative elements (LARDs) in the reference molecules, we found 75 putative LARDs (LTRs with an internal region that has no or limited similarity to the reference molecule) and 646 truncated LARDs. In total, 77 Mb (20.6%) of the rice genome match our reference elements. The percentage of each chromosome that matches LTR-retrotransposon is as follows: Chr1, 15%; Chr2, 15%; Chr3, 14%; Chr4, 26%; Chr5, 23%; Chr6, 20%; Chr7, 22%; Chr8, 25%; Chr9, 23%; Chr10, 25%; Chr11, 22% and Chr12, 25%. Their overall genomic distibution is shown in Figure 1.

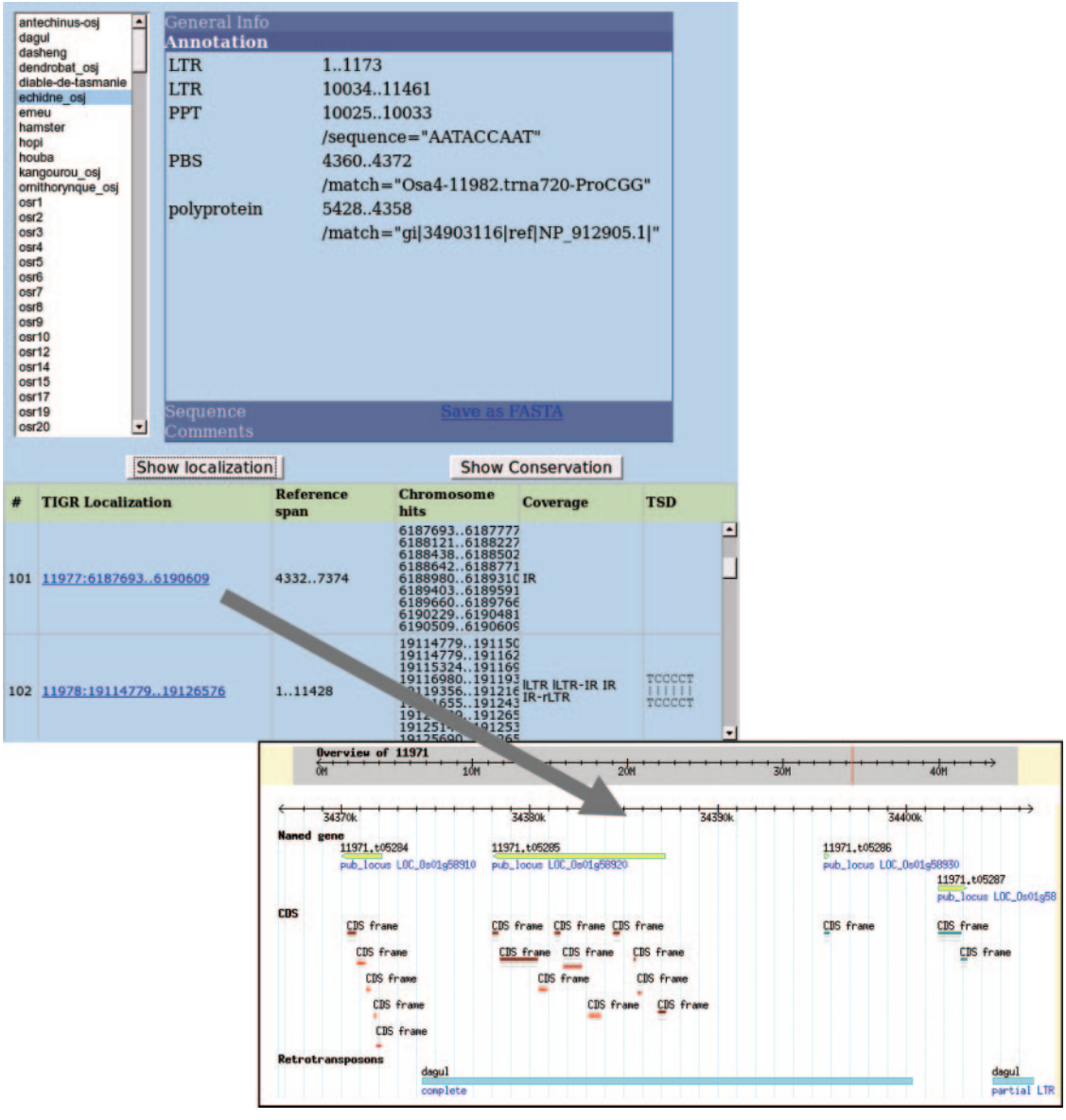

Figure 2. The database can be browsed through this interface. The widget on the upper right consists of four tabs which are used to present the data to the user. One opens a tab by clicking on the name of the tab. When localizations are shown, the TIGR localization link can be used to browse the region using Gbrowse. The complete genome annotation version 4 is available.

DATABASE ACCESS AND SERVICES

We provide access to the database through a web interface (http://www.retroryza.org/). This interface allows the data pertaining to individual retroelements to be consulted. General information is presented, which includes the name and the length of the retrotransposon, the name of the author that described the molecule, as well as a reference if available. In the same widget the sequence is presented in a tab with the option to download it in the FASTA format. Furthermore, an annotation tab is presented where the positions of the LTRs, the PBS, the PPT or the polyprotein are given when available. Finally, the comment tab is used to provide further data on the retroelement (Figure 2).

We used the TIGR rice annotation (release 4) (6) to annotate the retrotransposons. The genomic localizations for a retroelement are shown on demand by pressing the 'show location' button. In the table (Figure 2), the column named 'TIGR localization' shows the genomic location which is hyperlinked so that the region can be browsed using Gbrowse (24). The following column, 'span', shows the area of the reference molecule that is represented. The next column shows the HSPs that comprise the group and is followed by the region of the reference molecule that is covered. Here 'lLTR' represents the 5′ LTR, 'IR' the internal region and 'rLTR' the 3′ LTR. When a hit spans a LTR and the internal region, it is represented by 'lLTR-IR' or 'IR-rLTR' and when a HSP covers from the 'lLTR' to the 'rLTR' it is represented by 'lLTR-rLTR'. The last column gives the TSD when one is found. A graph showing the conservation of the sequence is also shown. This graph is constructed by accumulating all HSPs that are longer than 50 bp and possess >90% identity.

One of the services offered in our site is the ability to search for LTR-retrotransposons in the vicinity of a gene of interest. The database can be queried with a genomic location, a TIGR locus or TIGR gene model identifier. Another option is offered to the user who can query the database by searching all gene descriptions for a keyword and selecting one of the results. The size of the region surrounding the gene to be searched can be defined from 10 to 60 kb. The results are displayed on the same page and can be browsed using Gbrowse.

Another service is the automatic annotation of retrotransposons in a sequence of interest (Figure 3). We use the annotation algorithm described earlier and provide the user with the possibility of controlling the three parameters that will influence the sensibility or sensitivity of the annotation tool. A default of 50/80/100 (minimum HSP length/minimum similarity percentage/minimum group length to report),

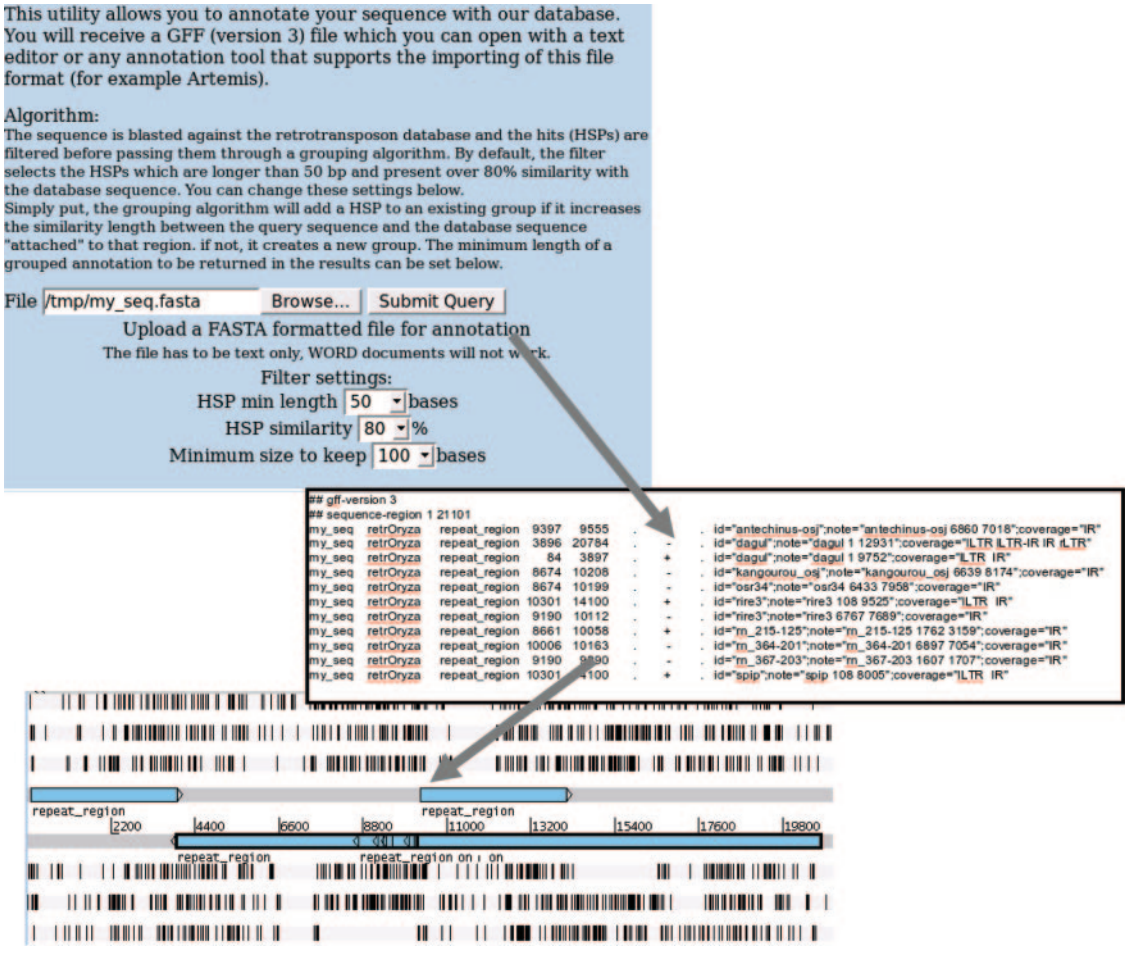

Figure 3. The annotation process consists of uploading a FASTA formatted file to the server which will return a file to the user after the analysis. This file is in the GFF format which can be opened with a text editor or loaded into an annotation editor such as Artemis.

which will detect most elements similar to the reference molecules and can be used as a basis for further annotation, is presented. When using a 50/50/50 setting, the detection of degenerated borders is improved and weakly related molecules will be better identified. However, these criteria produce more false positives, i.e. small sequences with low sequence identity to reference sequences which do not represent retrotransposons. Although by using a setting of 80/90/300 will reduce the false positive rate, distantly related elements will not be detected. Thus the user can impose conditions most suitable to his needs.

CONCLUSIONS AND FUTURE DEVELOPMENTS

We have generated a rice LTR-retrotransposon reference molecule database representing 242 elements which have been structurally and functionally annotated. These reference molecules have been used to annotate the genomic sequence of rice. As a result, the search of retroelements inserted near genes of interest is facilitated. We have also developed an annotation tool that can be used to identify and annotate retrotransposons in a sequence. This tool will allow us to extend our retrotransposon annotation to other species of *Oryza* and will provide the foundation for a comparative study of retrotransposon evolution.

Plans for future development include the integration of retrotransposon data for other *Oryza* species. Furthermore, this database has been developed within the French ANR ITEGE project which aims to identify retrotransposon-gene co-transcripts. This project includes several transcriptomic analyses from plants subjected to various stress. The results will be implemented in retrOryza when they become available. We also have made custom microarrays in order to study the expression of retroelements under stress in order to correlate them with gene expression. The experiments have been started and the data will eventually be made available on retrOryza. Future development of the database will also include the implementation of the annotation of the Rice Annotation Project III, as soon as they are publicly available.

ACKNOWLEDGEMENTS

The authors wish to thank Amy Frary for her critical reading of the manuscript. C.C. is funded by the ITEGE project (ANR-05-BLAN-0244-03). The work is supported by both ANR and CNRS. Funding to pay the Open Access publication charges for this article was provided by CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Kumar,A. and Bennetzen,J.L. (1999) Plant retrotransposons. *Annu. Rev. Genet.*, **33**, 479–532.
2. Piegu,B., Guyot,R., Picault,N., Roulin,A., Saniyal,A., Kim,H., Collura,K., Brar,D.S., Jackson,S., Wing,R.A. *et al.* (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.*, **16**, 1262–1269.
3. Vitte,C. and Panaud,O. (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenet. Genome Res.*, **110**, 91–107.
4. Kashkush,K., Feldman,M. and Levy,A.A. (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nature Genet.*, **33**, 102–106.
5. IRGSP consortium (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
6. Yuan,Q., Ouyang,S., Wang,A., Zhu,W., Maiti,R., Lin,H., Hamilton,J., Haas,B., Sultana,R., Cheung,F. *et al.* (2005) The institute for genomic research Osa1 rice genome annotation database. *Plant Physiol.*, **138**, 18–26.
7. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
8. Bennetzen,J.L., Coleman,C., Liu,R., Ma,J. and Ramakrishna,W. (2004) Consistent over-estimation of gene number in complex plant genomes. *Curr. Opin. Plant Biol.*, **7**, 732–736.
9. Berriman,M. and Rutherford,K. (2003) Viewing and annotating sequence data with Artemis. *Brief Bioinformatics*, **4**, 124–132.
10. McCarthy,E.M., Liu,J., Lizhi,G. and McDonald,J.F. (2002) Long terminal repeat retrotransposons of *Oryza sativa*. *Genome Biol.*, **3**, RESEARCH0053.
11. Nakajima,R., Noma,K., Ohtsubo,H. and Ohtsubo,E. (1996) Identification and characterization of two tandem repeat sequences (TrsB and TrsC) and a retrotransposon (*RIRE1*) as genome-general sequences in rice. *Genes Genet. Syst.*, **71**, 373–382.
12. Ohtsubo,H., Kumekawa,N. and Ohtsubo,E. (1999) *RIRE2*, a novel *Gypsy*-type retrotransposon from rice. *Genes Genet. Syst.*, **74**, 83–91.
13. Kumekawa,N., Ohtsubo,H., Horiuchi,T. and Ohtsubo,E. (1999) Identification and characterization of novel retrotransposons of the *Gypsy* type in rice. *Mol. Gen. Genet.*, **260**, 593–602.
14. Kumekawa,N., Ohmido,N., Fukui,K., Ohtsubo,E. and Ohtsubo,H. (2001) A new *Gypsy*-type retrotransposon, *RIRE7*: preferential insertion into the tandem repeat sequence TrsD in pericentromeric heterochromatin regions of rice chromosomes. *Mol. Genet. Genomics*, **265**, 480–488.
15. Li,Z.Y., Chen,S.Y., Zheng,X.W. and Zhu,L.H. (2000) Identification and chromosomal localization of a transcriptionally active retrotransposon of Ty3-*Gypsy* type in rice. *Genome*, **43**, 404–408.
16. Wang,R., Hong,G. and Han,B. (2003) [Characterization of the copy number of *RIRE10* retrotransposon and transcriptional activity of its LTR in rice genome]. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai)*, **35**, 768–773.
17. Panaud,O., Vitte,C., Hivert,J., Muzlak,S., Talag,J., Brar,D. and Sarr,A. (2002) Characterization of transposable elements in the genome of rice (*Oryza sativa* L.) using Representational Difference Analysis (RDA). *Mol. Genet. Genomics*, **268**, 113–121.
18. Jiang,N., Jordan,I.K. and Wessler,S.R. (2002) *Dasheng* and *RIRE2*. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. *Plant Physiol.*, **130**, 1697–1705.
19. Vitte,C., Quesneville,H. and Panaud,O. (2006) *Spip* and *Squiq*, two novel rice non-autonomous LTR retro-element families related to *RIRE3* and *RIRE8*. *Plant Sci.,* doi:10.1016/j.plantsci.2006.07.008.
20. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
21. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
22. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.
23. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
24. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.