Genome analysis

Transcriptionally active gene fragments derived from potentially fast-evolving donor genes in the rice genome

Xiangfeng Wang1,†, Zhihui Yu2,†, Xiaozeng Yang2, Xing-Wang Deng1 and Lei Li2,*

1Department of Molecular, Cell and Developmental Biology, Yale University, New Haven, CT 06520 and
2Department of Biology, University of Virginia, Charlottesville, VA 22904, USA

Received on December 18, 2008; revised on March 5, 2009; accepted on March 11, 2009
Advance Access publication March 16, 2009

ABSTRACT
The unprecedented complexity of the transcriptomic data obtained in recent years creates opportunities for new genomic studies aimed at interpolating regulatory code of gene expression and tracing genome evolution. We report here the identification and characterization of a set of 851 intergenic loci that represent transcribed gene fragments (TGFs) ectopically duplicated from 1030 non-transposable element (non-TE) donor genes in the rice genome. We analyzed the genomic context of the TGFs and donor genes. We show that the TGFs have adopted transcriptional orientation and pattern independent of the donor genes. We further show that TGFs have undergone relaxed purifying selection, consistent with their being pseudogenized. We found that the donor genes, which are biased toward certain molecular functions, exhibit an accelerated evolution rate comparing to the genome average. Our results demonstrated a large number of actively TGFs in the rice genome and shed light on the origin, mode of action and function of the TGFs.

Contact: 114jn@virginia.edu
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
One of the most exciting biological findings in recent years is the discovery of widely occurring transcriptional activity beyond the annotated genes. Transcriptome profiling efforts enabled by rapidly accumulating genome sequences and high-throughput techniques have led to the identification of numerous transcripts. One of the approaches broadly used in eukaryotic transcriptome profiling is whole-genome tiling microarray that involves the construction of a virtual ‘tile path’ consisting of progressive oligonucleotide probes to represent a target genome (Mockler and Ecker, 2005). Tiling array analyses have been carried out in yeast (Wilhelm et al., 2008), fly (Stoeck et al., 2004), worm (He et al., 2007), sea urchin (Samanta et al., 2006), human (Bertone et al., 2004; Cheng et al., 2005), Arabidopsis (Yamada et al., 2003), rice (Li et al., 2006) and legume plants (Li et al., 2008). Results from these studies have led inscrupably to the conclusion that transcripts are produced from a large number of genomic loci that do not encode proteins or structural RNA-species.

The varied properties of the newly identified novel transcripts suggest that they have diverged genomic origin and biological function. A large group of these transcripts locate in the intergenic regions and have yet to be incorporated into the molecular biology knowledge framework. One suggested mode of action for the intergenic transcripts is to provide continued transcription activity across genome regions for maintenance of an open chromatin state. Such a role is in line with results from Pol II occupancy assays where Pol II binding was found upstream of many genes that are either active or poised for rapid activation (Kim et al., 2005; Radonjic et al., 2005). In addition to influencing the structure of their surrounding chromatin, there are documented instances where intergenic transcripts act to repress transcriptional initiation at nearby genes by means of local competition between adjacent promoters (Hirschman et al., 1988) or interference by Pol II elongating from an upstream promoter (Martens et al., 2005). Thus, the function of some novel intergenic transcripts appears to require physical adjacency to their target genes.

Despite the rapid progress in our understanding of the complexity and dynamics of eukaryotic transcriptomes, relatively little attention has been paid to the interaction between physically unlinked intergenic transcripts and protein-coding genes. In the current study, we report a group of transcribed intergenic loci in the rice genome that correspond to ectopically duplicated gene fragments. Characterization of these transcribed intergenic gene fragments revealed their potential role as a mechanism contributing to functional and genomic diversity.

2 METHODS
Transcriptionally active regions (TARs) and their expression profile were obtained from a previous report (Li et al., 2007). Genome information, including intergenic sequences, protein sequences, segmental duplication, Gene Ontology (GO)-slim terms and paralogous gene families were downloaded from TIGR rice genome annotation release 5 (http://rice.tigr.org/) in October 2007. The rice mutator-like transposable elements (MULE) data have been previously described (Juretic et al., 2005) and were downloaded from http://www.genome.org/. The Arabidopsis proteome sequences were obtained from the latest version of the Arabidopsis genome annotation (TAIR8), which we downloaded in May 2008 from http://www.arabidopsis.org/.

Genomic loci corresponding to intergenic TARs were translated in all six possible open reading frames and a BLASTX search was performed to
establish the relationship between the deduced peptides and the annotated rice non-transposable element (non-TE) proteins. An intergenic locus was identified as a duplicated fragment of a donor gene if its encoded peptide (longest ORF) has an identity above 50% over a 50 amino acid span with the BLASTX expected value $E \leq 1 \times e^{-3}$. All hits were mapped against the rice genome annotation release 5 (Ouyang et al., 2007) again to remove loci that intersect with annotated exons or introns. The published genomic coordinates of MULEs were based on the International Rice Genome Sequencing Project annotation version 2.0 (Juretic et al., 2005). We remapped all the MULEs onto TIGR annotation release 5. After mapping these sequences, transcribed gene fragments (TGFs) that intersect with genomic regions flanked by the MULE target-site duplications were identified. In addition, we also used these sequences to identify Pack-MULEs that associate with the TGFs.

In the rice genome database, Plant GO-Slim Ontologies, which are a simple version of Gene Ontologies, were assigned to annotation rice non-transposable element (non-TE) proteins. Hypothetical proteins, TE-related proteins and proteins with GO IDs of “unknown” definitions were excluded from this analysis. A total of 19102 genes have been assigned GO functions with a total of 106119 associations made in the dataset used in the current study (Ouyang et al., 2007). This dataset was used as the benchmark against which the donor genes were compared. For a given GO-Slim term in the molecular function (MF), cellular component (CC) and biological process (BP) categories, we counted the number of all rice non-TE genes and TGF donor genes linked at least once with that term. We determined whether proportions were equivalent between the two groups of genes using Pearson’s $\chi^2$-test with the Bonferroni’s correction for multiple tests.

To carry out Ka/Ks analysis, pairs of homologous protein regions were determined for two datasets using the same criteria for identifying TGFs: (1) TGFs and donor genes in rice and (2) rice genes and Arabidopsis genes. For dataset (2), the identification cutoff was increased to 60% to reduce the number of homologous regions to facilitate downstream analysis. These pairs of homologous protein regions were aligned with CLUSTALW using default options. The alignments were subsequently superimposed on the coding region of nucleotide sequences using a Perl script based on Bioperl functions. For all pairwise alignments, the Ka and Ks substitution rate was calculated using the modified Yang–Nielsen (YN) method as previously described (Zhang et al., 2006).

### 3 RESULTS AND DISCUSSION

#### 3.1 Identification of TGFs

In a previous tiling microarray analysis of the *japonica* rice genome, we identified a total of 39018 TARs including 15472 in the intergenic regions (Li et al., 2007). To understand the origin and function of the intergenic TARs, we compared their deduced peptide sequences with the annotated non-TE rice protein-coding genes using a set of stringent criteria (Section 2). This analysis identified 851 TAR-containing loci that each encodes a partial open reading frame highly similar to one or more of 1030 annotated genes (Supplementary Table 1). These 851 partial open reading frames are referred to as putative TGFs derived from the corresponding protein-coding genes (TGF donor genes). It should be noted that the TARs are identified from a pool of four tissue types (Li et al., 2007) and not likely to be exhaustive. Consequently, our estimation of the number of TGFs in the rice genome is likely conservative.

To further validate TGFs and to discern their regulation, we used previously reported transcriptome data to examine the transcription pattern of TGFs and donor genes. In this dataset, a microarray containing five optimized probes for each non-TE gene and TARs in rice was used to measure gene expression in 10 different rice tissue types (Li et al., 2007). We detected expression of 7965 (18%) genes in all 10 tissue types and 28265 (64%) genes in at least one of the 10 tissues. From the same experiment, we detected transcription of 290 (34%) and 729 (86%) TGFs in all and at least one tissue type, respectively. These results indicate that TGFs are transcriptionally active as the protein-coding genes.

Furthermore, we found that the transcript levels of the TGFs and the corresponding donor genes are not correlated across the 10 tissue types (Pearson correlation coefficient $r = 0.20 \pm 0.13$). In contrast, transcript levels of the 2874 genes in the two-member paralogous gene families in rice exhibit a modest correlation ($r = 0.48 \pm 0.09$). Assuming cross-hybridization leads to elevated transcriptional correlation, our result suggests that cross-hybridization between a TGF and its donor genes does not have a predominant effect on the observed transcription pattern. This conclusion in turn suggests that the TGFs are regulated independently of the donor genes.

We determined the transcriptional orientation of the TGFs relative to the ORF of the donor genes. By mapping the strandedness of all the intergenic TARs intersecting with a given TGF, we were able to determine whether a TGF is transcribed from the same strand (sense configuration), or the antisense strand (antisense configuration), or bidirectionally relative to the donor gene. From this analysis, we found that 606 (71.2%) and 414 (48.6%) TGFs are transcribed in the antisense and sense configuration, respectively, indicating antisense configuration is the dominant form of TGF transcription. A significant portion of the TGFs (169 or 19.9%) is either bidirectionally transcribed or match with more than one donor gene. Taken together, our results suggest that the TGFs have adopted transcriptional orientation and regulation independent of the donor genes. Assuming TGFs and their donor genes are transcribed in the same cell, our results suggest that they have the potential to form complex transcript networks. Furthermore, independent transcription from the TGFs and donor genes may entail a combinatorial effect on the steady-state levels of the donor gene transcripts.

#### 3.2 Chromosomal organization of TGFs

The rice genome has undergone whole-genome duplications and more recently short segmental and individual gene duplications (Yu et al., 2005). To understand how TGFs relate to these duplication events, we examined whether TGFs are derived from tandem duplication of their donor genes. This analysis revealed that only seven (0.7%) TGF/donor gene pairs are physically adjacent (separated by zero intervening spacer genes). In contrast, it was estimated that about 1291 (3.1%) genes are tandem duplicates in rice using the same criteria (Rizzon et al., 2006). Thus, frequency of tandemly arrayed TGF and donor gene is significantly lower than the genome average ($\chi^2$-test, $P < 0.001$).

We next tested whether TGFs are generated from segmental duplication by mapping TGFs and the donor genes to the known segmentally duplicated blocks in the rice genome (Ouyang et al., 2007). From this analysis, we found that 20 (2.0%) TGF/donor gene pairs and 2321 (11.3%) non-TE gene pairs locate in one of the corresponding pairs of duplicated segments. Thus, TGF and donor genes are significantly depleted from segmentally duplicated blocks ($\chi^2$-test, $P < 0.001$), indicating that the generation of TGFs occurred either independently from segmental duplications or after the duplication events.
TGFs that overlap with 38 Pack-MULEs (Fig. 1). This result indicates that a significant portion of the TGFs (13%) may arise from MULE-mediated gene fragment duplication. To evaluate whether the donor genes are biased toward particular classes of Class I TEs, we identified 109 TGFs whose corresponding region in the donor gene regions span introns. The majority of these TGFs (76 or 94%) retained the intron sequences. Further examination of the genomic context of the corresponding donor genes revealed that 38 (50%) of them are in fact embedded in clusters of TEs (≥ 2 TEs on each side) that consist mostly of Class I TEs, suggesting that TE-mediated unequal DNA exchange may have produced some of the TGFs. Mutator-like TEs (MULEs) in higher plants are known to capture and carry fragments of cellular genes. Such elements are referred to as Pack-MULEs (Jiang et al., 2004; Juretic et al., 2005). In rice, whole-genome scan has identified 8274 MULEs including 1968 Pack-MULEs. We mapped TGFs against the genomic regions flanked by the MULE target-site duplications and identified 109 (13%) TGFs that overlap with 71 MULEs, including 55 (50%) TGFs that overlap with 38 Pack-MULEs (Fig. 1). This result indicates that a significant portion of the TGFs (13%) may arise from MULE-mediated gene fragment duplication.

3.3 Functional and evolutionary implication of TGFs on donor genes

To evaluate whether the donor genes are biased toward particular functions or processes, we analyzed their assignment to the plant GO-slim terms. We used GO-slim terms in three categories: MF, BP and CC. In the predicted rice proteome, a total of 19058 proteins have been assigned to at least one GO term with a total of 106 515 assignments (Ouyang et al., 2007). In rice, whole-genome scan has identified 8274 MULEs including 1968 Pack-MULEs. We mapped TGFs against the genomic regions flanked by the MULE target-site duplications and identified 109 (13%) TGFs that overlap with 71 MULEs, including 55 (50%) TGFs that overlap with 38 Pack-MULEs (Fig. 1). This result indicates that a significant portion of the TGFs (13%) may arise from MULE-mediated gene fragment duplication.

We further examined the distribution of TGFs in the 12 rice chromosomes by comparing the number and density of TEs, non-TE genes and TGFs to the size of each chromosome (Supplementary Table 2). We found TGF distribution is biased toward smaller chromosomes and TGF density is strongly correlated with TE density (unit per Mb) among the 12 chromosomes (Spearman rank correlation coefficient ρ = 0.83, P = 0.0058). In contrast, TGF density and non-TE gene density is inversely correlated (ρ = −0.89, P = 0.0032).

Both Class I (retrotransposons) and II (DNA transposons) TEs are known to amplify host DNA segments. To investigate whether Class I TEs contribute to the generation of TGFs, we identified 81 TGFs whose corresponding region in the donor gene regions span introns. The majority of these TGFs (76 or 94%) retained the intron sequences. Further examination of the genomic context of the corresponding donor genes revealed that 38 (50%) of them are in fact embedded in clusters of TEs (≥ 2 TEs on each side) that consist mostly of Class I TEs, suggesting that TE-mediated unequal DNA exchange may have produced some of the TGFs. Mutator-like TEs (MULEs) in higher plants are known to capture and carry fragments of cellular genes. Such elements are referred to as Pack-MULEs (Jiang et al., 2004; Juretic et al., 2005). In rice, whole-genome scan has identified 8274 MULEs including 1968 Pack-MULEs. We mapped TGFs against the genomic regions flanked by the MULE target-site duplications and identified 109 (13%) TGFs that overlap with 71 MULEs, including 55 (50%) TGFs that overlap with 38 Pack-MULEs (Fig. 1). This result indicates that a significant portion of the TGFs (13%) may arise from MULE-mediated gene fragment duplication.

We found that 462 (34.7%) TGFs contain at least one in-frame stop codon. Thus, a large portion of TGFs has been pseudogenized. To gain further insight into the evolutionary fate of TGFs, we determined the synonymous and non-synonymous substitutions in the TGFs in comparison with the donor genes. In total, we were able to make 1333 pairwise alignments between a TGF-coding sequence and that of a donor gene. In 79 (5.9%) alignments, the TGF contained no synonymous substitution of amino acids, suggesting that these TGFs are fairly recent duplications. These alignments were not included in subsequent analysis.

We estimated the evolutionary rate of the TGFs on the basis of the ratio of non-synonymous substitutions per non-synonymous site (Ka) to synonymous substitutions per synonymous site (Ks). The average Ka/Ks value for the 1254 TGF and donor gene pairs was 0.892 with a SD of 0.702. Notably, the overall distribution of Ka/Ks values was trimodal with the three peaks locating at approximately 0.20 (mode I), 0.65 (mode II) and 1.1 (model III), respectively (Fig. 2A). When TGFs in the sense and antisense configuration were examined separately, it was found that antisense TGFs contributed mainly to modes I and III (Fig. 2B). This result suggests that while a small portion of TGFs might have been under strong purifying selection, most antisense TGFs were under neutral selection. On the other hand, the distribution of Ka/Ks values for TGFs in the sense configuration contained a major peak at about 0.55 and a secondary peak around 1 (Fig. 2B). This result suggests that the sense TGFs were mainly under relaxed purifying selection and to a less extent neutral selection.

To investigate whether the presence of a TGF affects the evolution rate of the corresponding donor gene, we calculated the Ka/Ks ratio of the homologous regions between rice and Arabidopsis. When all homologous regions between the two species are considered, a bimodal Ka/Ks distribution with the major peak at 0.03 and a minor

Table 1. Overrepresented MF groups in TGF donor genes

<table>
<thead>
<tr>
<th>Rank</th>
<th>N</th>
<th>X</th>
<th>GO-slim attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>221</td>
<td>617</td>
<td>Receptor activity</td>
</tr>
<tr>
<td>2</td>
<td>350</td>
<td>1103</td>
<td>Carbohydrate binding</td>
</tr>
<tr>
<td>3</td>
<td>457</td>
<td>2753</td>
<td>Kinase activity</td>
</tr>
<tr>
<td>4</td>
<td>461</td>
<td>2893</td>
<td>Nucleotide binding</td>
</tr>
<tr>
<td>5</td>
<td>482</td>
<td>3283</td>
<td>Protein binding</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>155</td>
<td>Nuclease activity</td>
</tr>
<tr>
<td>7</td>
<td>37</td>
<td>786</td>
<td>Structural molecule activity</td>
</tr>
<tr>
<td>8</td>
<td>72</td>
<td>1746</td>
<td>MF unknown</td>
</tr>
<tr>
<td>9</td>
<td>142</td>
<td>3520</td>
<td>Hydrolase activity</td>
</tr>
<tr>
<td>10</td>
<td>27</td>
<td>2360</td>
<td>Transcription factor activity</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
<td>1740</td>
<td>DNA binding</td>
</tr>
</tbody>
</table>

Overrepresentation was determined by χ²-test with Bonferroni correction for multiple tests at P < 0.01. N, number of donor genes; X, number of genes in the rice genome from the same GO category.

3.4 Evolutionary and biological implications of TGFs

To determine whether the donor genes are biased toward particular GO-slim terms, we analyzed their assignment to the plant GO-slim terms. We used GO-slim terms in three categories: MF, BP and CC. In the predicted rice proteome, a total of 19058 proteins have been assigned to at least one GO term with a total of 106 515 assignments (Ouyang et al., 2007). In rice, whole-genome scan has identified 8274 MULEs including 1968 Pack-MULEs. We mapped TGFs against the genomic regions flanked by the MULE target-site duplications and identified 109 (13%) TGFs that overlap with 71 MULEs, including 55 (50%) TGFs that overlap with 38 Pack-MULEs (Fig. 1). This result indicates that a significant portion of the TGFs (13%) may arise from MULE-mediated gene fragment duplication.

We found that 462 (34.7%) TGFs contain at least one in-frame stop codon. Thus, a large portion of TGFs has been pseudogenized. To gain further insight into the evolutionary fate of TGFs, we determined the synonymous and non-synonymous substitutions in the TGFs in comparison with the donor genes. In total, we were able to make 1333 pairwise alignments between a TGF-coding sequence and that of a donor gene. In 79 (5.9%) alignments, the TGF contained no synonymous substitution of amino acids, suggesting that these TGFs are fairly recent duplications. These alignments were not included in subsequent analysis.

We estimated the evolutionary rate of the TGFs on the basis of the ratio of non-synonymous substitutions per non-synonymous site (Ka) to synonymous substitutions per synonymous site (Ks). The average Ka/Ks value for the 1254 TGF and donor gene pairs was 0.892 with a SD of 0.702. Notably, the overall distribution of Ka/Ks values was trimodal with the three peaks locating at approximately 0.20 (mode I), 0.65 (mode II) and 1.1 (model III), respectively (Fig. 2A). When TGFs in the sense and antisense configuration were examined separately, it was found that antisense TGFs contributed mainly to modes I and III (Fig. 2B). This result suggests that while a small portion of TGFs might have been under strong purifying selection, most antisense TGFs were under neutral selection. On the other hand, the distribution of Ka/Ks values for TGFs in the sense configuration contained a major peak at about 0.55 and a secondary peak around 1 (Fig. 2B). This result suggests that the sense TGFs were mainly under relaxed purifying selection and to a less extent neutral selection.

To investigate whether the presence of a TGF affects the evolution rate of the corresponding donor gene, we calculated the Ka/Ks ratio of the homologous regions between rice and Arabidopsis. When all homologous regions between the two species are considered, a bimodal Ka/Ks distribution with the major peak at 0.03 and a minor
The minor peak of Ka/Ks is still at 1.4, while the major peak shifted
counterparts are considered, a bimodal distribution of
Arabidopsis. For instance, the TGFs may exert dosage effect on the
genes between rice and the Ka/Ks values for the homologous non-TE genes and homologous donor
configuration against the corresponding donor genes. (Ka/Ks values for pairwise comparisons of TGFs and the donor genes.
∼ the homologous regions between the TGF donor genes and their
selection, most are under strong purifying selection. When only
Fig. 2.
Given the process by which gene duplication drives genome
innovation, it is conceivable that TGFs have the capability to
by which TGFs impact donor gene function and lead to genetic
diversity.

**REFERENCES**


Cheng,J. et al. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide

He,H. et al. (2007) Mapping the C. elegans noncoding transcriptome with a whole-

Hirschman,J.E. et al. (1988) Genetic evidence for promoter competition in


Juntic,N. et al. (2005) The evolutionary fate of MULE-mediated duplications of donor

Kim,T.H. et al. (2005) A high-resolution map of active promoters in the human genome.


L.L. et al. (2007) Global identification and characterization of transcriptionally active

L.L. et al. (2008) Transcriptional analysis of highly syntenic regions between Medicago

Martens,J.A. et al. (2005) Regulation of an intergenic transcript controls adjacent gene

Mockler,T.C. and Ecker,J.R. (2005) Applications of DNA tiling arrays for whole-


Radaic,M. et al. (2005) Genome-wide analyses reveal RNA polymerase II located
upstream of genes poised for rapid response upon S. cerevisiae stationary phase

Rizzini,C. et al. (2006) Striking similarities in the genomic distribution of tandemly

960–962.


genome. Science, 302, 842-846.

3, e53.

substitutions. BMC Evol. Biol., 6, 44.