

MONO- AND BI-CISTRONIC CHIMERIC mRNAs IN ARABIDOPSIS AND RICE GENOMES

ILHAM A. SHAHMURADOV^{1,2}, AMINA U. ABDULAZIMOVA², VICTOR V. SOLOVYEV³,
RAHEEL QAMAR¹, SHAHID NADEEM CHOCHAN^{1,4}, JALAL A. ALIYEV²

ABSTRACT. In contrast to prokaryotes, the proximity of genes in eukaryotic genomes has not previously been known to play any significant role in their expression profiles. However, recently reported phenomena in human, mouse, yeast and a few plant species indicate such role: chimeric mRNAs and proteins produced via alternative transcription termination, splicing and translation of the Tail-to-Head (T2H) neighboring genes. To further verify this evidence at the genomic scale, we scanned the rice and *Arabidopsis* genomes for the presence of T2H gene pairs at a distance of less than 1000 bp. Our studies suggest that “non-stopping” transcription and alternative splicing of some of these T2H pairs may produce chimeric transcripts. We obtained cDNA support for 100 *Arabidopsis* and 42 rice T2H pairs to be transcribed into chimeric mRNA(s). Analysis of 155 chimeric transcripts from both species revealed 71 mono-cistronic and 58 bi-cistronic mRNAs potentially encoding chimeric or separate proteins.

Keywords: *Arabidopsis*, Rice, Tail-to-Head Neighboring Genes, Chimeric mRNAs, Computer Analysis.

AMS Subject Classification: 62-07, 92-08, 92C80.

1. INTRODUCTION

The “one-gene, one-protein” rule proposed by Beadle and Tatum suggests that the number of proteins in an organism must not exceed the total number of genes in its genome. However, in eukaryotes one gene may code for many mRNAs through alternative splicing and other mechanisms [14]. For example, less than 30,000 genes in human genome code for over 100,000 different proteins [21](see also: http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml). A recently reported phenomenon of transcription induced chimeric (TIC) genes generates additional mRNA variants [1,17].

Transcription of the eukaryotic genes begins at a transcription start site (TSS) of promoter, and ends at a regulated termination point [20,24]. Neighboring genes are usually separated from each other by intergenic nonexpressed regions [12]. However, occasionally the regulated termination point of one gene fails to work effectively and results in transcription to continue until the termination point of the next downstream gene located in tandem on the same strand of the DNA (Tail-to-Head, T2H). Such read-through transcription produces mRNA comprising of exons from both the neighboring T2H. Furthermore, alternative splicing of these chimeric pre-mRNA might give various alternative chimeric mRNAs consisting of exons from both the

¹Department of Biosciences, COMSATS Institute of Information Technology, Islamabad 44000, Pakistan, e-mail: ilham@comsats.edu.pk

²Department of Molecular-Genetic Bases of Production Processes, Institute of Botany, Baku AZ1073, Azerbaijan

³Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20, UK

⁴School of Natural Sciences, University of Western Sydney, NSW, Australia

Manuscript received 25 December 2008.

genes, where intergenic spacer may serve as a part of a new exon or intron. The translation of such a chimeric mRNA will produce a chimeric protein with domains from both the original proteins [1,5,10,11,17,18,19,23].

Akiva et al [1] identified, through in silico studies, over 200 cases of intergenic splicing, involving 421 genes in the human genome and then experimentally demonstrated the existence of at least half of the resulting fusion proteins in the human tissues. They also showed that: (1) unique splicing patterns dominate the functional and regulatory nature of the resulting transcripts, and that (2) there is an intergenic distance bias in the fused genes compared to the non-fused genes. It was suggested that that the hundreds of fused genes identified are only a subset of a large number of fused genes present in human cells.

Parra et al [17], by relying on both computational and experimental analysis, is estimated that 4%–5% of the tandem gene pairs in the human genome can be eventually transcribed into a single RNA sequence encoding a putative chimeric protein. Moreover, transcription-induced chimerism followed by retroposition might result in a new and active fused gene in eukaryotes [1,6,16,17].

The transcription induced chimerism was reported also in plants. Thus, comparisons of full-length cDNAs and genomic DNA in *Arabidopsis* revealed that some adjacent loci are transcribed into long RNAs covering two annotated genes of T2H arrangement. Alternative splicing of some of these transcripts generates bi-cistronic transcripts which span both their coding and intergenic sequences and contain two complete open reading frames. Others are spliced to generate mono-cistronic transcripts coding for fused or chimer, proteins derived from both loci [22]. Recently, Muralla et al [15] reported that in *Arabidopsis* the non-stop transcription of the adjacent genes BIO3 (AT5G57600; encodes the third enzyme, dethiobiotin synthetase, in the biotin biosynthetic pathway) and BIO1 (AT5G57590; associated with the second reaction in the pathway). These genes are organized in T2H fashion and in addition to individual BIO3 and BIO1 transcripts, produces two different types of chimeric BIO3-BIO1 transcripts, via alternative splicing. One of the fused transcripts produced such is mono-cistronic and encodes a bifunctional fusion protein. The second chimeric mRNA variant is bi-cistronic, with distinct but overlapping reading frames. It was also shown that protein encoded by the mono-cistronic transcript has the dual functionality. So, in contrast to most biosynthetic enzymes in eukaryotes, encoded by separate genes, biotin biosynthesis in *Arabidopsis* involves a bifunctional locus catalyzing two sequential reactions in the same metabolic pathway [15].

The current work was aimed to explore the role of the T2H location of genes in their expression profiles, both in rice and *Arabidopsis*, at the whole genome level. Below we present the results of the comparative analysis of gene organization in these genomes and discuss the possible functional and evolutionary implications of our findings.

2. MATERIALS AND METHODS

Following sets of publicly annotated 12 chromosomes of rice (NC_008394.1, NC_008395.1, NC_008396.1, NC_008397.1, NC_008398.1, NC_008399.1, NC_008400.1, NC_008401.1, NC_008402.1, NC_008403.1, NC_008404.1, NC_008405.1 ; February 2008) and 5 chromosomes of *Arabidopsis* (NC_003070.6, NC_003071.4, NC_003074.5, NC_003075.4 and NC_003076.5; July 2008) were used for this study.

Gene ontology data for rice and *Arabidopsis* were obtained from the genome annotations and TAIR WEB-site, (ftp://ftp.Arabidopsis.org/home/tair/Genes/Gene_OntologyATH_GO.20031202.txt), respectively.

A search for cDNA/mRNA support of TIC potential of the T2H genes selected (by criteria described below) was performed by using TIGR Database of Rice Transcript Assembles (Release 2, totally 247,516 transcripts; ftp://ftp.tigr.org/pub/data/plantta/Oryza_sativa/Oryza_sativa_release_2.fasta) and 6 sets of totally 138,017 *Arabidopsis* cDNA sequences (ATH1.cdna.gz from ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/cDNA_full_reading_020509.txt.gz, cDNA_full_reading_050524.txt.gz, cDNA_full_reading_030312.txt.gz, and cDNA_full_reading_051102.txt.gz from <http://rarge.gsc.riken.jp/archives/rafl/sequence/ATcdna155.bz2> from <http://gremlin3dev.gdcb.iastate.edu/AtGDB/download.php>). To map cDNAs/transcripts on T2H genes, we applied EST_MAP program (<http://www.softberry.com>) that finds the best alignment of a transcript with a region of genome sequence and reconstructs the exon-intron gene structure taking into account splice site weight matrices.

Analysis of the genome annotations of rice and *Arabidopsis* was performed using computer programs specially developed by us for this task. The pairwise comparison of amino acid sequences has been carried out by BLAST program [3]. Search for statistically significant open reading frames (ORFs) was done by the BESTORF program (<http://www.softberry.com>).

In search for putative TIC genes, only T2H genes at a distance ≤ 1000 bp have been selected for further analysis. TIC potential of these T2H genes was analyzed by the following criteria based on known cDNA:

- The presence of at least one cDNA/mRNA supporting "non-stop" transcription of a pair of adjacent T2H genes.
- Supporting cDNA/mRNA covering ≥ 500 bp or $2/3$ of the annotated gene length from every adjacent genes.
- Mapping similarity level (between putative exons and cDNA/mRNA) $\geq 95\%$.
- Mapping homology length (as exons) ≥ 10 nucleotides (nt).
- A gap is not allowed between the compared genomic and corresponding cDNA sequences.

3. MAIN RESULTS

3.1. Structural characteristics of the annotated genes of *Arabidopsis* and rice. Analysis of annotations of *Arabidopsis* and rice genomes revealed interesting features of their gene structure and organization. Some of these features relevant for this study are summarized in Table 1, while, many others have been discussed by [2]. Overall, the *Arabidopsis* genome was found to be much better annotated than the rice genome, especially in defining the gene boundaries. For example, 26,264 protein coding genes have been annotated in rice. However, the position of the ATG start codon is unknown in 3,247 genes, and such genes were not included in our analysis. By contrast, in *Arabidopsis* all 26,877 genes are annotated with known ATG start codon.

We studied how often the neighboring genes can be produced by duplications during the evolution and found that only a minor fraction of both rice and *Arabidopsis* adjacent gene pairs show significant protein sequence similarity between genes in pairs (about 2.2% and 7% of 23,005 and 26,872 pairs, respectively). At the same time, at least 6% rice and 12.6% *Arabidopsis* gene pairs share one or more GO code therefore having similar function and/or cell compartment.

For 1545 and 148 genes with cDNA/mRNA support in *Arabidopsis* and rice, respectively, the 5'-UTR of the mRNA is not known. It does actually mean that the start of these genes is not fully identified. The distribution of 5'-UTR lengths in *Arabidopsis* and rice genes is presented in Fig. 1. Hundreds of genes in both organisms are annotated with very short (≤ 20 nt) 5'-UTR.

Although the minimal length of the 5'-UTR sufficient for the normal translation of the mature mRNA is unknown, we suppose that it is not less than 10-20 nt. On the other side, in some cases the 5'-UTR is very long (even more than 1000 nt). The role of such very long 5'-UTRs is not clear, but they may contain unrecognized introns and such pre-mRNA molecules could be potential targets for the alternative splicing of the 5'non-coding exons in a tissue-specific and/or development stage-dependent manner [4,9].

3.2. The mutual location of annotated neighboring genes. The neighboring genes can occur on a genome in either H2H, T2H or T2T fashion. In eukaryotes, except the rare cases where the genes are organized in operons in some of the organisms, every gene has its own regulatory elements for transcription initiation (promoter) and termination. Analyzing the location of the genes in relation to each other in *Arabidopsis* and rice revealed some very interesting facts. Indeed, a bias in favour of one of these mutual orientation types is known and this is also observed in both studied plant genomes (Table 1). Nevertheless, a comparison of the distribution of the intergenic distances in H2H, T2H and T2T pairs reveals a significant difference: genes in T2T pairs show a tendency to be located more closely than in the other two types.

In the *Arabidopsis* genome, 2155 T2T gene pairs are located at a distance less than 100 bp. This is in striking contrast with only 145 H2H and 466 T2H pairs with genes at this distance. The similar, but less strong, tendency is observed in the rice genome: 487 T2T pairs versus 235 H2H and 65 T2H pairs. Although this difference between *Arabidopsis* and rice is probably related to the big difference in the sizes of these genomes, this observation suggest some functional restriction for the presence of the H2H and T2H genes within the proximity of each other.

One interesting observation is the existence of overlapping genes in both the genomes (Table 1). In particular, 1021 and 280 *Arabidopsis* and rice gene pairs with cDNA support were found to contain overlapping genes, where the prevalence of T2T type pairs is obvious. Moreover, in comparison to *Arabidopsis*, in rice most of the overlapping area includes coding sequences (CDS) of both the genes (124 pairs) or even, in 84 cases, one gene is completely inserted in the other one of the pair.

Pairwise BLAST comparison of 23017 rice and 26877 *Arabidopsis* proteins revealed a significant difference in evolutionary conservation for various types of neighboring gene pairs, H2H, T2H and T2T. Thus, we found only 235 interspecies conservative H2H pairs (with BLAST E-value $\leq 10^{-10}$). The similar picture is observed with T2T pairs. By contrast, T2H pairs seem more conserved: rice and *Arabidopsis* genomes have 11614 pairs with interspecies conservation. About 50% of these pairs are pairs with a tandem gene duplication.

3.3. Potential transcription-induced chimeras in *Arabidopsis* and rice genomes. The observed presence of thousands of adjacent gene pairs in a T2H pairs indicated a possibility of existence of chimeric transcripts both in rice and *Arabidopsis*. To investigate this idea we analyzed 6895 *Arabidopsis* and 1386 rice T2H genes separated by an intergenic spacer of less than 1000 bp, following criterion stated above and using the accurate mapping cDNAs to the exons on the genomic sequence by EST_MAP program (<http://www.softberry.com>). We performed the mapping procedure for 138,017 and 247,516 *Arabidopsis* and rice cDNA sequences, respectively, on the chromosomal regions of the corresponding species, including both genes of the T2H pair (hereafterin referred as Gene 1 and Gene 2) and the intergenic spacer (IGS) between them. The mapping revealed 6 different variants (classes) of potential alternative splicing in the sense of origin of their exons involved (i.e. covering Gene1, Gene2 and, in most cases, IGS, too; Table 2). Results of this study are summarized in Table 3.

The cDNA mapping results demonstrate that 100 *Arabidopsis* and 42 rice T2H pairs undergo read-through transcription and following alternative splicing which produce chimeric mRNA(s), including 82 and 11 cases when both genes have cDNA support to be transcribed also separately, 26 and 9 cases with cDNA support for additional independent transcription of only one of the two genes (Gene 1 or Gene 2) of a pair, respectively. In *Arabidopsis* alternative splicing of read-through transcripts of 100 T2H pairs gives, at least, 110 different chimeric mRNAs (9 pairs with 2 alternative splicing variants and 1 pair with 3 alternative splicing variants), among them in 90 cases only the canonical GT-AG splicing sites are used, in 5 cases mRNA consists of a single exon and the remaining portion includes non-canonical exon-intron boundaries. Alternative splicing of the read-through transcripts of 42 T2H pairs in rice results in, at least, 45 different chimeric mRNAs (2 pairs with 2 alternative splicing variants and 1 pair with 3 alternative splicing variants), where in 25 cases only the canonical GT-AG splicing sites are used, in 5 cases mRNA consists of a single exon and the remaining portion includes non-canonical exon-intron boundaries.

Another interesting result of this study is the observation of 6 different classes of alternative splicing of the read-through transcripts in the sense of origin of their exons (Table 2, 3). Out of 155 chimeric mRNAs from both organisms, 73 ones (49 in *Arabidopsis* and 24 in rice) include, at least, one exon completely or partially located in the IGS of the corresponding T2H pair. These finding indicates that in about 40% cases alternative splicing producing a chimeric mRNA captures, at least, one new (alternative) exon. In 82 (61 in *Arabidopsis* and 21 in rice) cases, an alternative splicing generating chimeric mRNA seems to be a combination of exons from Gene 1 and Gene 2 of T2H pair.

To explore protein coding potential of mTIC genes, we performed a search for statistically significant ORFs in chimeric mRNAs by BESTORF program (<http://www.softberry.com>). Results of the analysis are summarized in Table 3. In *Arabidopsis*, 90 T2H pairs produce 95 chimeric mRNA(s) capable of encoding a putative mTIC protein(s) of 30 aa or longer. For the remaining 15 chimeric mRNAs for 10 T2H pairs we failed to identify any open reading frame (ORF) ≥ 30 aa. Fifty chimeric mRNAs contain a single ORF for a putative mTIC protein and 45 chimeric mRNAs include 2 separate ORFs (hereafter, the mRNAs with one and two ORFs will be referred as “mono-cistronic” and “bi-cistronic”, respectively; [22]). Excluding 3 chimeric mRNAs, putative chimeric proteins for mono-cistronic mRNAs are homologous to both (28 cases) or only one (19) of annotated Protein 1 and Protein 2. For 45 bi-cistronic mRNAs, in 28 cases mTIC Protein 1 and mTIC Protein 2 are homologous to Protein 1 and 2, respectively; in 5 cases both mTIC Protein 1 and mTIC Protein 2 are homologous to both Protein 1 and Protein 2; in 12 cases mTIC Protein 1 and mTIC Protein 2 show similarity to only one of Protein 1 and 2 (Table 3). The lists of mono- and bi-cistronic chimeric mRNAs are given in Tables 4a and 4b.

In rice, 42 T2H pairs produce 38 chimeric mRNA(s) encoding a putative mTIC protein(s) of 30 aa or longer; for the remaining 7 chimeric mRNAs in 5 T2H pairs we did not identify any ORF for ≥ 30 aa. 21 and 17 chimeric mRNAs are mono-cistronic and bi-cistronic, respectively. Excluding 2 chimeric mRNAs, putative chimeric proteins for mono-cistronic mRNAs are homologous to both (4 cases) or only one (15 cases) of Protein 1 and Protein 2. For 17 bi-cistronic mRNAs, in 8 cases mTIC Protein 1 and mTIC Protein 2 are homologous to Protein 1 and 2, respectively; in 2 cases both mTIC Protein 1 and mTIC Protein 2 are homologous to both Protein 1 and Protein 2; in 4 cases mTIC Protein 1 and mTIC Protein 2 show similarity to only one of both proteins; in 3 cases putative mTIC proteins show no similarity with Protein 1 and Protein 2 (Table 3, 4a, 4b).

One of the mono-cistronic chimeric mRNAs found in *Arabidopsis* is schematically illustrated in Fig. 2. It includes 2 exons (out of 5) from Gene 1 and all 3 exons from Gene 2. However, 4 (out of 5; excluding the second exon) of the chimeric mRNA only partially cover the corresponding original exons (“alternative” exons are shorter). Potential mTIC protein of 431 aa is chimera which is homologous to both annotated proteins and similar to GDSL-motif lipase/hydrolase family protein (amino acids 123-295, out of 356 aa; 98% similarity) as well as to integral membrane protein (1-283, out of 283; 90%), respectively. Therefore, the chimeric protein is likely to be a slightly changed isoform of the integral membrane protein with additional domain(s) from the GDSL-motif lipase/hydrolase family protein. Such a hybrid protein with sharing domains from both neighboring genes was recently described in *Arabidopsis*: the non-stop transcription of the adjacent genes BIO3 (At5g57600) and BIO1 (At5g57590) genes of T2H arrangement produces mono-cistronic chimeric mRNA encoding a fusion functional protein catalyzing two sequential reactions in the biotin biosynthetic pathway [15].

An example of the bi-cistronic chimeric mRNA is given in Fig. 3. It includes a complete single exon of Gene 1, IGS and all 4 exons of Gene 2. At least, 2 ORFs are encoded by this chimeric mRNA and they encode proteins identical to Protein 1 and 2, respectively.

The mono- and bi-cistronic chimeric mRNAs encode, totally, 195 putative proteins. Although our search conditions allowed any protein of 30 aa or longer, most (171) of them are ≥ 50 aa, including 150 proteins of 100 aa or longer.

4. DISCUSSION

A read-through transcription of neighbouring T2H genes gives rise to chimeric pre-mRNA combining sequences of both genes and spacer which can further undergo alternative splicing. Thus, we have found 155 (110 in *Arabidopsis* and 45 in rice) chimeric mRNAs covering both annotated genes of T2H pair. Most of these chimeric transcripts seems to be real TIC genes, because (a) they have cDNA support; (b) both annotated genes of 93 (out of 161) T2H pairs, as well as one of 2 annotated genes of 35 T2H pairs also have cDNA support (i.e. transcribed separately), and only in 24 cases (2 in *Arabidopsis* and 22 in rice) there is no cDNA support for the annotated genes; (c) no gap is available between the genomic and corresponding cDNA sequences.

Our studies also revealed that in 40% cases alternative splicing resulting in a chimeric mRNA involves, at least, one new (alternative) exon. Moreover, our mapping analysis revealed that the first exon of Gene 1 and/or the last exon of Gene 2 of 47 and 23 T2H pairs from *Arabidopsis* and rice, respectively, are incomplete in the genome annotations studied. In particular, we have identified 3 (1 first and 2 last) and 4 (3 first and 1 last) new exons in *Arabidopsis* and rice.

Exploration of the protein coding potential of 155 chimeric mRNAs revealed 71 mono-cistronic and 62 bi-cistronic transcripts. In 36 bi-cistronic mRNAs the adjacent ORFs encode proteins homologous to proteins encoded by Gene 1 and Gene 2 of a pair, respectively. 32 mono-cistronic mRNAs are capable to encode a protein of homology to both annotated genes of T2H pair (like with domains from both proteins). Here it should be mentioned that previously [22] described 60 chimeric mRNAs (29 bi-cistronic and 31 mono-cistronic) covering 58 T2H pairs in *Arabidopsis*. Now, only for 8 T2H pairs we mapped the same bi-cistronic mRNAs (AT2G25620-AT2G25610, AT3G26310-AT3G26300, AT3G45640-AT3G45650, AT4G00030-AT4G00040, AT4G35870-AT4G35880, AT4G20830-AT4G20840, AT5G57140-AT5G57150 and AT5G65360-AT5G65350). The remaining 50 and 92 T2H pairs producing chimeric RNAs, identified by [22] and us, respectively, are completely different.

For 22 chimeric mRNAs (out of 155) for 15 T2H pairs (out of 142) neither a single ORF nor double ORF (encoding ≥ 30 aa) was found. This could have arisen from multiple reasons, including: (a) such mRNAs represent incomplete splicing variants; (b) sequencing errors; and (c) a limited prediction capacity of the BESTORF program. Moreover, some of them may not even be coding any protein in which case, it would be worthwhile investigating how such mRNAs avoid decaying by cell control mechanisms to minimize the damage resulting from genome mistakes or metabolic processes, by nonsense-mediated mRNA decay, nonstop mRNA decay, and no-go mRNA decay? Along with other types of quality control that occur during the complex processes of mRNA biogenesis, these mRNA surveillance mechanisms help to ensure the integrity of protein-encoding gene expression (for a review see: [8]). However, it was also found that many such “abnormal” transcripts are stable: they do not undergo the degradation process for reasons still to be understood. Moreover, these transcripts are generally more abundant than the correctly spliced partner in various cell lines and tissues [8,13]. In this sense, TIC mRNAs may be the result of prevention of a decay of the nonstop mRNAs covering two T2H genes with retained introns and/or intergenic spacers. Intron and/or IGS retention is the least characterized event of all alternative splicing types, because this phenomenon is mostly excluded in studies, due to the difficulty to differentiate it from genomic DNA or incompletely-processed transcripts. For example, studies show that up to 15% of human genes present at least one intron retention event, and that at least 22% of all informative introns-retention events are also present in the mouse transcriptome [7].

Parra et al [17], by relying on both computational and experimental analysis, is estimated that at least 4–5% of the tandem gene pairs in the human genome can be eventually transcribed into a single RNA sequence encoding a putative chimeric protein. While the reality and functional significance of most of these chimeric transcripts remains to be determined, they provide strong evidence that this phenomenon does not correspond to mere technical artifacts and that it is a common mechanism with the potential of generating hundreds of new proteins in the human genome. In other words, tandem chimeras seem to be a means to increase protein variety and complexity in the human genome. Moreover, the non-stop transcription of adjacent genes in T2H manner raises some additional possibilities in regulation of genes, like:

- (1) transcription of the second gene of T2H pair comes under control of a promoter of the first gene;
- (2) formation of the 3'-end of the chimeric transcript will be conducted by polyadenylation signals of the second gene;
- (3) the destination compartment of a protein expressed by the second T2H gene is determined by transit signals of the first gene.

Of course, these findings and the corresponding suggestions presented in this report require further theoretical and experimental studies. Moreover, some important additional aspects of functional role of the gene neighbourhood in higher eukaryotes are yet to be discovered.

5. ACKNOWLEDGEMENTS

This work, supported by the Higher Education Commission of Pakistan through Foreign-Faculty Start-up grant, is collaborative research performed by the Department of Biosciences (COMSATS Institute of Information Technology, Islamabad, Pakistan), the Department of

Molecular-genetic bases of Production Processes (Institute of Botany, Azerbaijan National Academy of Sciences, Baku, Azerbaijan) and Department of Computer Science (Royal Holloway, University of London, UK).

Table 1. Some features of the annotated protein-coding nuclear genes of *Arabidopsis* and rice.

	<i>Arabidopsis</i>		Rice	
	All ¹	Support ²	All ¹	Support ²
Total number of genes analyzed	26877	18818	23017	16091
Intron-less genes	5885	3372	4322	3092
Genes with known 5'-UTR (of any length)	18788	17273	16943	15943
Genes with ATG codon in the second or further exons	2963	2963	2972	2972
H2H gene pairs ($\leftarrow\rightarrow$) :				
all	6404	3327	5438	2763
inter-gene spacer \leq 600 bp related ³	1631	949	512	264
homologous ⁴	557	321	181	75
homologous ⁴	93	39	51	16
T2H gene pairs ($\rightarrow\rightarrow, \leftarrow\leftarrow$):				
all	14038	7040	11830	5960
inter-gene spacer \leq 1000 bp related	6895	3770	1386	842
homologous	2295	1059	1041	409
homologous	1698	639	390	134
T2T gene pairs ($\rightarrow\leftarrow$) ⁵ :				
all	6403	3361	5460	2794
inter-gene spacer \leq 100 bp related	2155	1504	487	294
homologous	522	283	163	64
homologous	105	71	57	42
Overlapping gene pairs :				
all	1388	1021	603	280
T2H	57	18	26	13
H2H	66	30	222	75
T2T	1265	973	355	192
one gene is completely included by another one	33	14	198	84
CDS with CDS	48	11	337	124

¹All annotated genes including ones without cDNA support; 3247 rice genes with unknown ATG start codon position were excluded. ²Genes with cDNA support; a gene pair of adjacent location belongs to this group, if (i) both genes have cDNA support, and (ii) there is no another gene between them; 3233 rice genes with unknown ATG start codon position were excluded. ³These genes have, at least, one common GO code of known function/process/cell compartment. ⁴BLAST E-value $\leq 10^{-10}$. ⁵T2T genes are located adjacently in a fashion of Tail-to-Tail.

Table 2. Classification of alternative splicing events found to produce chimeric mRNAs in *Arabidopsis* and rice.

Class #	Description
1	Ex[Gene 1] + Ex[Gene 1 - Gene 2]
2	Ex[Gene 1] + Ex[Gene 1 - Gene 2] + Ex[Gene 2]
3	Ex[Gene 1] + Ex[IGS] + Ex[Gene 2]
4	Ex[Gene 1] + Ex[Gene 2]
5	Ex[Gene 1 - Gene 2]
6	[Gene 1 - Gene 2] + Ex[Gene 2]

Ex[Gene 1] and **Ex[Gene 2]** : exon(s) derived from only Gene 1 and Gene 2, respectively; **Ex[IGS]**: exon (s) derived from only intergenic spacer (IGS); **Ex[Gene 1–Gene 2]**: exon(s) derived from Gene 1, IGS and Gene 2.

Table 3. Integral characteristics of the T2H pairs with cDNA support to produce chimeric mRNA(s) via the read-through transcription and consequent alternative splicing in the *Arabidopsis* and rice genomes.

Description of characteristics	AT ¹	OS ²
T2H pairs with, at least, one chimeric mRNA potential supported by cDNA	100	42
cDNA support is also available for both genes of T2H pair	82	11
cDNA support is also available for one of genes of T2H pair	26	9
cDNA support is available for only chimeric mRNA	2	22
Alternative splicing producing chimeric mRNA, Class 1 ³	8	7
Class2	16	9
Class3	3	2
Class4	61	21
Class5	8	4
Class6	14	2
T2H pairs produces mono- and/or bi- and/or tri-cistronic chimeric mRNA(s) ⁴ encoding putative mTIC protein(s) ⁵ , total	90 ⁶	37 ⁶
mTIC protein(s) without similarity ⁷ to Protein 1 ⁸ and/or Protein 2 ⁹	3	5
mTIC protein(s) with similarity to Protein 1 and/or Protein 2	87	32
Mono-cistronic chimeric mRNAs, total ¹⁰	50	21
mTIC protein with similarity to both Protein 1 and Protein 2	28	4
mTIC protein with similarity to Protein 1 or Protein 2 (not to both)	19	15
mTIC protein with similarity to neither Protein 1 nor Protein 2	3	2
Bi-cistronic chimeric mRNAs, total ¹¹	41	17
mTIC Protein 1 and mTIC Protein 2 with similarity to Protein 1 and Protein 2, respectively	27	8
mTIC Protein 1 and/or mTIC Protein 2 with similarity to both Protein 1 and Protein 2	5	2
mTIC Protein 1 and/or mTIC Protein 2 with similarity to Protein 1 or Protein 2 (not to both)	9	4
Both mTIC Protein 1 and mTIC Protein 2 without similarity to Protein 1 or Protein 2	0	3
Tricistronic chimeric mRNAs, total	4	0
mTIC protein(s) with similarity to both Protein 1 and Protein 2	2	0
mTIC protein(s) with similarity to Protein 1 or Protein 2 (not to both)	2	0

¹AT: *Arabidopsis thaliana*. ²OS: *Oryza sativa*. ³See: Table 3. ⁴mRNA containing one, two and three ORFs, respectively, where every ORF encodes a putative protein of ≥ 30 aa. ⁵Protein encoded by mapped chimeric mRNA. ⁶Both mono- and bi-cistronic chimeric mRNAs are available for some T2H pairs. ⁷BLAST E-value $\leq 10^{-10}$. ⁸Protein encoded by Gene 1 of T2H pair. ⁹Protein encoded by Gene 2 of T2H pair. ¹⁰Alternative chimeric mono-cistronic mRNAs for some T2H pairs encode a putative protein showing similarity to both or only one of the annotated proteins. ¹¹Putative proteins encoded by different ORFs of the same chimeric bi-cistronic mRNA show similarity to both or only one of the annotated proteins.

Table 4a. T2H pairs with chimeric mono-cistronic mRNAs in *Arabidopsis* and rice.

T2H pair	mRNA	Supporting cDNA	ORF			D ²
			Location	Protein	H ¹	
2	3	4	5	6	7	8
AT1G53163-AT1G53165	3122 nt,22 exons	68414.m06023	5-3025	1007 aa	1,2	517
AT1G61000-AT1G60995	2923 nt,24 exons	68414.m06868	5-2926	974 aa	1,2	207
AT1G69550-AT1G69545	2601 nt, 3 exons	68414.m07998	192-2294	701 aa	1	54
AT1G72890-AT1G72900	1989 nt, 2 exons	68414.m08432	656-1744	363 aa	1,2	477
AT1G80190-AT1G80200	1836 nt, 4 exons	AY600540	1068-1682	205 aa	2	918
AT2G28100-AT2G28105	1002 nt, 4 exons	AY735590	307-804	166 aa	2	163
AT2G31450-AT2G31440	1056 nt, 7 exons	AY088444	151-900	250 aa	2	160
AT2G35040-AT2G35035	1282 nt, 8 exons	BX819257	231-1112	294 aa	2	114

Table 4a (continued)

2	3	4	5	6	7	8
AT2G35820-AT2G35830	894 nt, 3 exons	AF428270	213-779	189 aa	1,2	22
AT2G36325-AT2G36330	1295 nt, 5 exons	68415.m04459	5-1297	431 aa	1,2	240
AT2G37680-AT2G37678	941 nt, 6 exons	68415.m04621	5-943	313 aa	1,2	506
AT2G43240-AT2G43235	2576 nt,15 exons	68415.m05374	5-2365	787 aa	1,2	23
AT2G43440-AT2G43445	2464 nt, 4 exons	68415.m05399	5-2377	791 aa	1,2	766
AT2G43750-AT2G43745	1227 nt, 2 exons	AY219095	491-898	136 aa	2	193
AT3G13062-AT3G13065	2403 nt,15 exons	AY518289	144-2204	687 aa	2	307
AT3G14700-AT3G14710	1546 nt, 4 exons	BT015055	176-1501	442 aa	2	146
AT3G17668-AT3G17670	1101 nt, 9 exons	68416.m02256	5-1105	367 aa	1,2	345
AT3G20395-AT3G20390	1501 nt, 8 exons	BX824388	710-1270	187 aa	2	123
AT3G20720-AT3G20730	2294 nt, 2 exons	AK176813	204-1895	564 aa	2	89
AT3G22440-AT3G22450	1302 nt, 2 exons	DQ132688	207-1139	311 aa	2	242
AT3G26780-AT3G26782	3336 nt, 4 exons	68416.m03350	29-3163	1045 aa	1,2	97
AT3G52910-AT3G52905	1645 nt,10 exons	68416.m05831	5-1648	548 aa	1,2	275
AT3G62000-AT3G61990	719 nt, 7 exons	BX842420	1-477	159 aa	1,2	205
AT4G02150-AT4G02140	543 nt, 3 exons	BT012121	3-137	45 aa	nh	340
AT4G14310-AT4G14305	3314 nt, 9 exons	68417.m02204	5-3130	1042 aa	1,2	126
AT4G18195-AT4G18197	2287 nt, 4 exons	68417.m02705	161-2290	710 aa	1,2	518
AT4G18197-AT4G18205	2182 nt, 3 exons	68417.m02705	1-2184	728 aa	1,2	521
AT4G22290-AT4G22285	2925 nt,13 exons	68417.m03224	5-2929	975 aa	1,2	224
AT4G32272-AT4G32270	1093 nt, 8 exons	68417.m04591	5-1096	364 aa	1,2	303
AT4G32610-AT4G32605	2289 nt, 6 exons	68417.m04643	464-2134	557 aa	1,2	195
AT4G35870-AT4G35880	4950 nt, 9 exons	AY096495	3-227	75 aa	1	450
AT4G36520-AT4G36515	1184 nt, 4 exons	AK222172	3-227	75 aa	nh	292
AT5G58000-AT5G58003	2922 nt,15 exons	68418.m07256	3-2924	974 aa	1,2	205
AT5G61810-AT5G61800	2332 nt, 4 exons	68418.m07756	172-1605	478 aa	1	65
AT5G66160-AT5G66170	1726 nt, 6 exons	AK222135	1182-1589	136 aa	2	214
AT4G16560-AT4G16550	1730 nt,12 exons	68417.m02504	2-1732	577 aa	1,2	491
AT4G22285-AT4G22280 ³	1478 nt, 4 exons	68417.m03222	151-1395	415 aa	2	639
	1496 nt, 3 exons	68417.m03223	147-1142	332 aa	2	639
AT4G27740-AT4G27745	655 nt, 3 exons	68417.m03986	5-448	148 aa	1,2	612
AT5G03800-AT5G03795	4331 nt, 5 exons	68418.m00347	5-4168	1388 aa	1,2	306
AT5G57260-AT5G57250	4452 nt, 3 exons	68418.m07152	5-4456	1484 aa	1,2	112
AT1G51538-AT1G51540	3462 nt, 7 exons	68414.m05801	5-3112	1036 aa	1,2	289
AT2G41997-AT2G41000	1176 nt, 5 exons	68415.m05064	5-829	275 aa	1,2	237
AT4G28811-AT4G28815	2642 nt,12 exons	68417.m04119	5-2644	880 aa	1,2	647
AT5G40405-AT5G40410	3468 nt, 2 exons	68418.m04901	5-3472	1156 aa	1,2	220
AT5G51540-AT5G51545	2582 nt,19 exons	68418.m06391	5-2584	860 aa	1,2	12
AT4G39480-AT4G39490	3039 nt, 2 exons	68417.m05585	75-3041	989 aa	1,2	839
AT5G47077-AT5G47075	194 nt, 2 exons	AY803256	38-196	53 aa	nh	679
AT1G52310-AT1G52315	1898 nt, 2 exons	AY299255	1104-1262	53 aa	2	577
AT3G10410-AT3G10405	864 nt, 8 exons	BT000426	1-603	201 aa	2	1
AT5G10040-AT5G10050	600 nt, 1 exons	BT014910	77-337	87 aa	1	42

Table 4a (continued)

2	3	4	5	6	7	8
Os01g0142600-Os01g0142800	778 nt, 4 exons	TA39333.4530	170-781	204 aa	2	563
Os01g0918300-Os01g0918200	452 nt, 3 exons	CR284004	3-254	84 aa	1,2	787
Os02g0190800-Os02g0190700	650 nt, 3 exons	TA49881.4530	134-412	93 aa	nh	57
Os03g0324200-Os03g0324300	2695 nt, 2 exons	TA35110.4530	337-2256	640 aa	1,2	906
Os04g0632100-Os04g0631900	1616 nt, 4 exons	TA52445.4530	116-1618	501 aa	1,2	196
Os01g0843800-Os01g0843900	512 nt, 2 exons	CF961539	285-515	77 aa	2	129
Os01g0945200-Os01g0945100	476 nt, 2 exons	CI390256	95-496	134 aa	2	687
Os03g0738200-Os03g0738300	1481 nt, 7 exons	TA53879.4530	1-732	244 aa	1	808
Os04g0310800-Os04g0310700 ³	802 nt, 2 exons	TA45267.4530	3-599	199 aa	1,2	193
	516 nt, 2 exons	CI284742	110-355	82 aa	2	193
Os06g0278000-Os06g0277900	1476 nt, 7 exons	TA59386.4530	23-1249	409 aa	2	253
Os07g0495400-Os07g0495300	845 nt, 2 exons	CB629989	231-849	273 aa	2	544
Os07g0628400-Os07g0628300	840 nt, 3 exons	TA59975.4530	384-641	86 aa	2	182
Os09g0560100-Os09g0560200	2469 nt, 2 exons	TA60485.4530	887-2230	448 aa	2	439
Os11g0158800-Os11g0158700	329 nt, 2 exons	C71893	122-331	70 aa	2	86
Os01g0695600-Os01g0695500	763 nt, 5 exons	TA37575.4530	184-564	127 aa	2	305
Os02g0651100-Os02g0651200	623 nt, 1 exons	CA755595	268-588	107 aa	2	126
Os05g0485800-Os05g0485700	1018 nt, 4 exons	TA48227.4530	1-606	202 aa	1	177
Os08g0225000-Os08g0224900	1000 nt, 6 exons	TA60831.4530	2-715	238 aa	1,2	432
Os08g0436000-Os08g0435900	515 nt, 3 exons	CF987854	69-518	150 aa	2	851
Os11g0429800-Os11g0429900	509 nt, 2 exons	TA70105.4530	1-135	45 aa	nh	120

Table 4b. T2H pairs with chimeric bi-cistronic mRNAs in *Arabidopsis* and rice.

T2H pair	mRNA	Supporting cDNA	ORF1			ORF2			D^2
			Location	Protein	H^1	Location	Protein	H^1	
1	2	3	4	5	6	7	8	9	10
AT1G51360-AT1G51370	2691 nt,4 exons	AK228894	29-658	210 aa	1,2	1121-2425	435 aa	2	311
AT1G69550-AT1G69545	4629 nt,3 exons	AK226752	75-2177	701 aa	1	2322-4430	703 aa	2	54
AT1G74630-AT1G74640	3522 nt,4 exons	AK220733	1-1911	637 aa	1	2312-3421	370 aa	2	310
AT2G20810-AT2G20815	2182 nt,7 exons	AY735579	19-357	113 aa	1	738-2099	454 aa	2	207
AT2G25620-AT2G25610	2722 nt,7 exons	AF436827	276-1451	392 aa	1	2115-2648	178 aa	2	411
AT2G28830-AT2G28820	2775 nt,6 exons	AK226821	56-2017	654 aa	1	2345-2779	145 aa	2	130
AT2G30890-AT2G30900	1707 nt,6 exons	AY262050	3-251	83 aa	1	83-1583	367 aa	2	664
AT2G32240-AT2G32235 ³	2090 nt,5 exons	AY219116	2-544	181 aa	1	962-1891	310 aa	2	99
	2165 nt,5 exons	AY219117	2-544	181 aa	1	963-1619	219 aa	2	99
AT3G23255-AT3G23260	1423 nt,8 exons	BX824147	3-668	222 aa	1	1059-1427	123 aa	2	247
AT3G23340-AT3G23350	2203 nt,9 exons	AY735606	1-768	256 aa	1	1373-2005	211 aa	2	144
AT3G26310-AT3G26300	3362 nt,4 exons	AY139766	905-1468	188 aa	1,2	1694-3193	500 aa	1,2	222
AT3G45640-AT3G45650	3088 nt, 11 exons	AY090981	59-1168	370 aa	1	1253-2926	558 aa	2	553
AT3G49060-AT3G49055	2329 nt,6 exons	AY735613	21-554	178 aa	1	643-1710	356 aa	2	245
AT3G58990-AT3G58980	2016 nt,2 exons	AY142548	21-779	253 aa	1	1159-2019	287 aa	2	51
AT3G59330-AT3G59320	2224 nt,16 exons	AK227436	607-1008	134 aa	1,2	1027-2043	339 aa	1,2	191
AT4G00030-AT4G00040	2259 nt,4 exons	BT001993	43-678	212 aa	2	968-2122	385 aa	1	213
AT4G19410-AT4G19400	2224 nt,17 exons	AK228936	118-1290	391 aa	1	1573-2016	148 aa	2	31
AT4G28025-AT4G28020	1747 nt,13 exons	AK226675	20-517	166 aa	1	626-1678	351 aa	2	288

Table 4b (continued)

1	2	3	4	5	6	7	8	9	10
AT4G28390-AT4G28380	1960 nt,4 exons	68417.m04063	78-1214	379 aa	1	1117-1215	33 aa	nh	356
AT5G02680-AT5G02690	1553 nt,4 exons	AK228657	3-197	65 aa	1	1431-1556	42 aa	nh	594
AT5G26880-AT5G26870 ³	1552 nt,7 exons	BX842289	1-792	264 aa	1	928-1290	121 aa	2	30
	1702 nt,6 exons	AK227520	24-812	263 aa	1	1100-1462	121 aa	2	30
AT5G27390-AT5G27395	1649 nt,17 exons	68418.m03270	1099-1353	85 aa	2	1390-1653	88 aa	2	166
AT5G44650-AT5G44660	2825 nt,7 exons	AK228589	196-1035	280 aa	1	1607-2731	375 aa	2	406
AT5G48080-AT5G48090	2924 nt,17 exons	AY735708	623-1426	268 aa	2	1910-2725	272 aa	2	261
AT5G50645-AT5G50640 ⁴	2074 nt,15 exons	AK230278	248-964	239 aa	2	1011-1895	295 aa	2	177
	960 nt,2 exons	BT011575	262-366	35 aa	nh	619-771	51 aa	nh	177
AT5G51960-AT5G51970	1687 nt,7 exons	AK176901	20-328	103 aa	1	478-1569	364 aa	2	96
AT5G57140-AT5G57150	3073 nt,8 exons	AY136380	258-1241	328 aa	1	1443-2120	226 aa	2	815
AT5G60370-AT5G60380	1831 nt,11 exons	AY735742	2-508	169 aa	1	764-1687	308 aa	2	886
AT5G63400-AT5G63390	2718 nt,12 exons	AY735745	23-760	246 aa	1	222-2520	433 aa	2	409
AT5G65360-AT5G65350	1123 nt,2 exons	AK117157	80-487	136 aa	1,2	390-488	33 aa	nh	33
AT4G22285-AT4G22280	1495 nt,4 exons	BX842369	147-1025	293 aa	2	1310-1456	49 aa	2	639
AT4G21065-AT4G21070	4013 nt,15 exons	68417.m03047	3-1112	370 aa	1	1236-3923	896 aa	2	270
AT2G07671-AT2G07777	1728 nt,3 exons	AY648320	192-446	85 aa	1	931-1545	205 aa	2	119
AT2G07701-AT2G07702	1498 nt,1 exon	AY131999	392-619	76 aa	2	844-1299	152 aa	nh	544
AT2G18210-AT2G18200	1556 nt,1 exon	DQ069798	21-392	124 aa	1,2	1216-1518	101 aa	1,2	538
AT2G31680-AT2G31670	1272 nt,1 exon	BT011004	368-487	40 aa	2	486-1160	225 aa	nh	89
AT2G35840-AT2G35850	1396 nt,9 exons	AY735594	2-937	12 aa	1	960-1199	80 aa	2	747
AT3G59090-AT3G59100	1553 nt,15 exons	BX823658	414-518	35 aa	nh	414-1223	270 aa	1	651
AT4G20830-AT4G20840	4231 nt,1 exon	AY062595	31-1650	540 aa	1,2	2462-4078	539 aa	1,2	109
AT2G43750- AT2G43745	1441 nt,1 exon	AY219096	491-955	155 aa	2	1096-1221	42 aa	nh	193
AT3G49060- AT3G49055	2229 nt,7 exons	AY735614	21-554	178 aa	1	643-2082	480 aa	2	245
AT5G63810-AT5G63820	3839 nt,22 exons	AK228789	173-793	207 aa	1	888-2483	532 aa	1	46
AT2G07707- AT2G07708	1877 nt,1 exon	DQ069799	1048-1368	107 aa	nh	1311-1622	104 aa	2	82
Os03g0839000-Os03g0838900	831 nt,6 exons	CB621473	423-521	33 aa	nh	514-687	58 aa	2	732
Os06g0715400-Os06g0715500	2280 nt,10 exons	TA31927_4530	3-284	94 aa	1	1171-1941	257 aa	2	325
Os09g0458400-Os09g0458700	3851 nt,2 exons	TA48792_4530	123-716	198 aa	1	3098-3652	185 aa	2	304
Os01g0250400-Os01g0250300	1464 nt,3 exons	TA40716_4530	13-105	31 aa	nh	14-106	31 aa	nh	441
Os01g0121700-Os01g0121600	2980 nt,12 exons	TA49554_4530	123-2126	668 aa	1	2702-2983	94 aa	2	226
Os02g0151500-Os02g0151600	812 nt,7 exons	CB656280	2-151	50 aa	nh	629-817	63 aa	2	246
Os04g0212100-Os04g0212200	2314 nt,13 exons	TA46072_4530	64-1605	514 aa	1,2	1508-1606	33 aa	nh	319
Os06g0155500-Os06g0155400	1173 nt,3 exons	TA66723_4530	1-750	250 aa	1	1048-1176	43 aa	2	293
Os06g0625900-Os06g0625800	5446 nt,9 exons	TA43613_4530	298-2733	812 aa	1	3028-4302	425 aa	2	130
Os12g0298800-Os12g0298700	848 nt,3 exons	CB634952	2-376	125 aa	1	691-852	54 aa	2	310
Os01g0142100-Os01g0142000	3954 nt,14 exons	TA29627_4530	64-1686	541 aa	1	3519-3947	143 aa	2	607
Os02g0760600-Os02g0760700	1392 nt,2 exons	TA62667_4530	162-548	129 aa	1	755-1036	94 aa	nh	68
Os03g0757300-Os03g0757500	1934 nt,1 exon	TA46207_4530	357-953	199 aa	2	1027-1869	281 aa	2	263
Os04g0250000-Os04g0250100	1099 nt,4 exons	TA37582_4530	158-571	138 aa	nh	501-608	36 aa	nh	147
Os04g0250100-Os04g0250200	1642 nt,4 exons	TA37582_4530	85-192	36 aa	nh	466-660	65 aa	nh	225
Os04g0255500-Os04g0255600	697 nt,1 exon	TA68096_4530	58-342	95 aa	nh	540-701	54 aa	2	194
Os07g0685500-Os07g0685600	361 nt,4 exons	CF992329	29-658	210 aa	1,2	1121-2425	435 aa	2	505

¹Homologous to Protein 1 (1), to Protein 2 (2), to both Protein 1 and (1,2), to neither Protein 1 nor Protein 2 (nh). ²Distance (bp) between neighboring genes. ³For these T2H pairs 2 alternative chimeric bi-cistronic mRNAs are available.

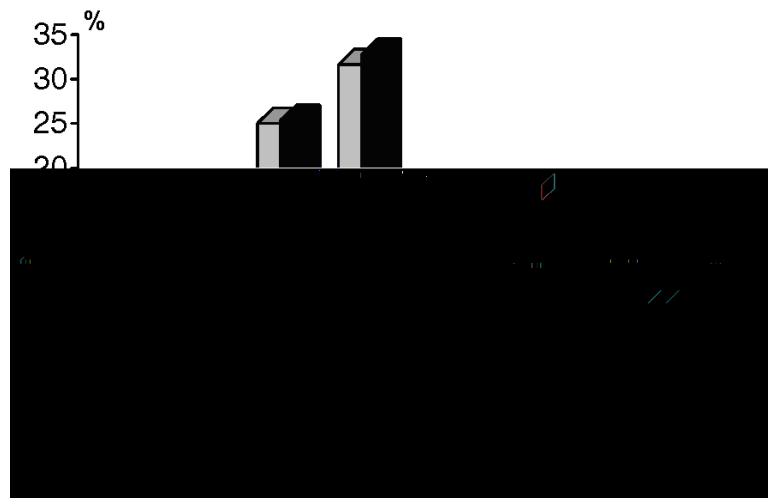


Figure 1. The Distribution of 5'-UTR lengths in *Arabidopsis* (grey) and rice (black) genes with cDNA/mRNA support. Abscise axis – 5'-UTR length in bp; ordinate axis – a fraction of the corresponding genes (%).

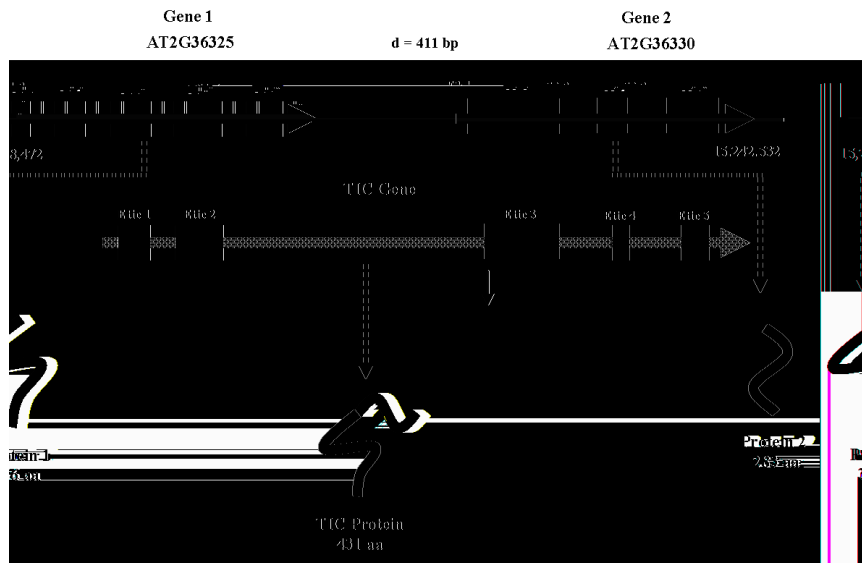


Figure 2. Potential TIC gene which is supported by cDNA (At2g36330.1 68415.m04459; ATH1.cdna.gz) covering both genes of the T2H pair. Gene with cDNA support (gb AK2216761) encodes protein similar to GDSL-motif lipase/hydrolase family protein. Gene 2 with cDNA support (gb BT002934) encodes protein similar to integral membrane protein. Correspondence between exons of the neighboring genes (E1.1 – E1.5 and E2.1 – E2.4) and putative TIC gene (Etic1 – Etic5): Etic1 – E1.3 (partial), Etic2 – E1.4, Etic3 – E2.1 (partial), Etic4 – E2.2 (partial), Etic5 – E2.3 (partial). Similarity between putative TIC protein and genes of the T2H pair: amino acids 1-172 and 173-431 from chimeric protein and amino acids 123-295 (98% identities) and 1-283 (90% identities) from AT2G36325 and AT2G36330, respectively.

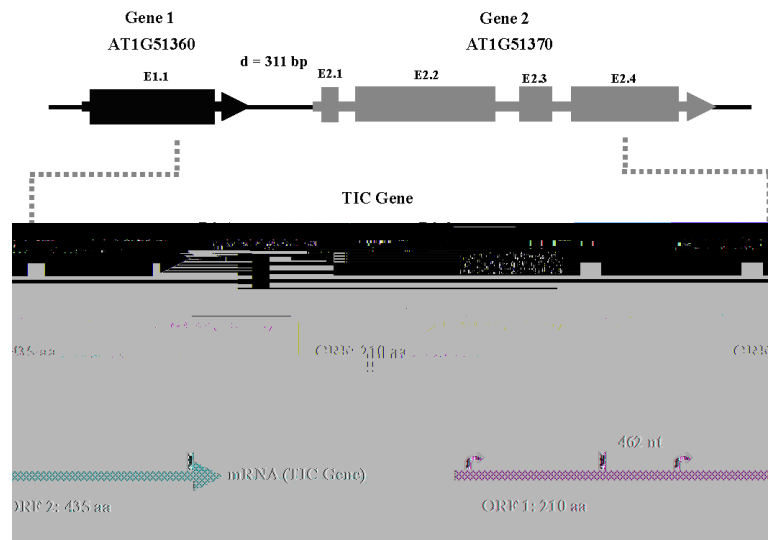


Figure 3. A potential TIC gene which is supported by cDNA (gb AK228894.1) covering both genes of the T2H pair, the Gene 1 and the Gene 2. This TIC gene pair produces a chimeric mRNA which contains 2 ORFs separated by 462 nucleotides (bi-cistronic mRNA). Gene 1 with cDNA support (gb BX817596) encodes protein similar to unknown proteins from *Arabidopsis*, *Vitis vinifera* and *Populus trichocarpa*, containing InterPro domain stress responsive dimeric alpha-beta barrel (InterPro:IPR013097). Gene 2 with ESTsupport (gb EL178293) encodes protein similar to *Arabidopsis* F-box family protein, FBL13 (TAIR:AT5G53840.1). Correspondence between exons of the neighboring genes (E1.1 and E2.1 – E2.4) and putative TIC gene (Etic1 – Etic4): Etic1 – E1.1 + [Intergenic spacer] + E2.1, Etic2 – E2.2, Etic3 – E2.3, Etic4 – E2.4. ORFs for the Gene 1 and the Gene 2 encode proteins identical to proteins encoded by ORF 1 and ORF 2, respectively.

REFERENCES

- [1] Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A., Sorek, R. Transcription-mediated gene fusion in the human genome, *Genome Res.*, V.16, 2006, pp. 30-36.
- [2] Alexandrov, N.N., Troukhan, M.E., Brover, V.V., Tatarinova, T., Flavell, R.B., Feldmann, K.A. Features of *Arabidopsis* genes and genome discovered using full-length cDNAs, *Plant Mol. Biol.*, V.60, 2006, pp.69-85.
- [3] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, V.25, 1997, pp.3389-3402.
- [4] Cai, X.L., Wang, Z.Y., Xing, Y.Y., Zhang, J.L., Hong, M.M. Aberrant splicing of intron 1 leads to the heterogeneous 5' UTR and decreased expression of waxy gene in rice cultivars of intermediate waxy content, *Plant Jour.*, V.14, 1998, pp.459-465.
- [5] Calvanese, V., Mallya, M., Campbell, R.D., Aguado, B. Regulation of expression of two Ly-6 family genes by intron retention and transcription induced chimerism, *BMC Mol. Biol.*, V.9, 2008, 81 p.
- [6] Chen, J.J., Janssen, B.J., Williams, A., Sinha, N. A Gene Fusion at a Homeobox Locus: Alterations in Leaf Shape and Implications for Morphological Evolution, *Plant Cell*, V.9, 1997, pp.1289-1304.
- [7] Galante, P.A., Sakabe, N.J., Kirschbaum-Slager, N., de Souza, S.J. Detection and evaluation of intron retention events in the human transcriptome, *RNA*, Vol.10, 2004, pp.757-765.
- [8] Isken, O., Maquat, L.E. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes Dev*, V.21, 2007, pp.1833-1856.
- [9] Isshiki, M., Nakajima, M., Satoh, H., Shimamoto, K. *dull*: rice mutants with tissue-specific effects on the splicing of the waxy pre-mRNA, *Plant Jour.*, V.23, 2000, pp.451-460.
- [10] Kato, M., Khan, S., Gonzalez, N., O'Neill, B.P., McDonald, K.J., Cooper, B.J., Angel, N.Z., Hart, D.N. Hodgkin's lymphoma cell lines express a fusion protein encoded by intergenically spliced mRNA for the

- multilectin receptor DEC-205 (CD205) and a novel C-type lectin receptor DCL-1, *Jour. Biol. Chem.*, V. 278,2003, pp.34035-34041.
- [11] Kowalski, P.E., Freeman, J.D., Mager, D.L. Intergenic splicing between a HERV-H endogenous retrovirus and two adjacent human genes, *Genomics*, V.57, 1999, pp.371-379.
- [12] Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et all. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome, *Nature*, V.409, 2001, pp.860-921.
- [13] Mallya, M., Campbell, R.D., Aguado, B. Transcriptional analysis of a novel cluster of LY-6 family members in the human and mouse major histocompatibility complex: five genes with many splice forms, *Genomics*, V.80, 2002, pp.113-123.
- [14] Modrek, B., Lee, C. A genomic view of alternative splicing, *Nat. Genet.* Vol.30, 2002, pp.13-19.
- [15] Muralla, R., Chen, E., Sweeney, C., Gray, J.A., Dickerman, A., Nikolau, B.J., Meinke, D. A Bifunctional Locus (BIO3-BIO1) Required for Biotin Biosynthesis in *Arabidopsis*, *Plant Phys.*, 2008, V.146, pp. 60-73.
- [16] Nakamura, Y., Itoh, T., Martin, W. Rate and Polarity of Gene Fusion and Fission in *Oryza sativa* and *Arabidopsis thaliana*, *Mol. Biol. Evol.*, V.24, 2007, pp.110-121.
- [17] Parra, G., Reymond, A., Dabbouseh, N., Dermitzakis, E.T., Castelo, R., Thomson, T.M., Antonarakis, S.E., Guigo, R. Tandem chimerism as a means to increase protein complexity in the human genome, *Genome Res.*, V.16, 2006, pp.37-44.
- [18] Poulin, F., Brueschke, A., Sonenberg, N. Gene fusion and overlapping reading frames in the mammalian genes for 4E-BP3 and MASK, *Jour. Biol. Chem.*, Vol.278, 2003, pp.52290-52297.
- [19] Pradet-Balade, B., Medema, J.P., Lopez-Fraga, M., Lozano, J.C., Kolfschoten, G.M., Picard, A., Martinez, A.C., Garcia-Sanz, J.A., Hahne, M. An endogenous hybrid mRNA encodes TWE-PRIL, a functional cell surface TWEAK-APRIL fusion protein, *EMBO Jour.*, V.21, 2002, pp.5711-5720.
- [20] Proudfoot, N.J., Furger, A., Dye, M.J. Integrating mRNA processing with transcription, *Cell*, V.108, 2002, pp.501-512.
- [21] The International Human Genome Consortium. Finishing the euchromatic sequence of the human genome, *Nature*, V.431, 2004, pp.931-945.
- [22] Thimmapuram, J., Duan, H., Liu, L., Schuler, M.A. Bi-cistronic and fused mono-cistronic transcripts are derived from adjacent loci in the *Arabidopsis* genome. *RNA*, V.11, 2005, pp.128-138.
- [23] Thomson, T.M., Lozano, J.J., Loukili, N., Carrio, R., Serras, F., Cormand, B., Valeri, M., Diaz, V.M., Abril, J., Burset M, et al. Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene, *Genome Res.*, V.10, 2000, pp.1743-1756.
- [24] Zhao, J., Hyman, L., Moore, C. Formation of mRNA 3'-ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mRNA synthesis, *Microbiol. Mol. Biol. Rev.*, V.63, 1999, pp.405-445.

Ilham A. Shahmuradov, for a photography and biography, see page 33.



Amina U. Abdulazimova - graduated from Baku State University in 1998 and has got her BSc in Medical biology. In 2000 she got her MSc degree on molecular biology from Baku State University. In 2003 A.Abdulazimova was admitted as a junior researcher to the Bioinformatics Laboratory, Institute of Botany, Azerbaijan National Academy of Sciences. Since 2006 she serves a researcher in the same Laboratory. She published more than 5 research papers.

Victor V. Solovyev, for a photography and biography, see page 33.



In 2003 he was promoted to the post of Associate Professor also appointed the Director of PCR Labs. In 2004 he was appointed as the Research Director of Shifa College of Medicine. Then in 2005 he was appointed as the Professor of Biosciences at the COMSATS Institute of Information Technology, where he subsequently served as the Chairman of Biosciences and then the Dean of the Faculty of Science. He was awarded the Tamgha-i-Intiaz by the President of Pakistan and last year he was elected as a Fellow of the Pakistan Academy of Sciences. His research specialization includes Enzymology, Population Molecular Genetics, Molecular Biology and Pathology. He has published 7 monographs and more than 44 journal papers.

Raheel Qamar - received his Ph.D. in 1992 from the University of North Texas, Denton, Texas, USA. He then joined Dr. A. Q. Khan Research Laboratories as a Senior Scientific Officer and was subsequently promoted to the position of Principal Scientific Officer. From 1998 to 2001 he worked for a while, at the Department of Biochemistry, University of Oxford, Oxford, UK in the lab of Prof. E.M. Southern. In 2002 he joined Shifa College of Medicine as Assistant Professor of Biochemistry.



Shahid Nadeem Chohan - graduated from Manchester in the area of Plant Molecular Biology in 1991. Since then he has completed several post-doctoral assignments in Europe and Australia in the fields of Molecular Biology and Enzymology. Currently he is serving as HEC Foreign Faculty at CIIT, Islamabad since September 2005 as well as Senior Research Fellow (adjunct) in the University of Western Sydney (UWS), Australia. Genomics and bioinformatics along with plant molecular diagnostics are areas of his teaching and research interests in the department of Biosciences.

Jalal A. Aliyev, for a photography and biography, see page 82.