

# SCIENTIFIC REPORTS



OPEN

## Origin of the plant *Tm-1-like* gene via two independent horizontal transfer events and one gene fusion event

Received: 18 May 2016  
Accepted: 31 August 2016  
Published: 20 September 2016

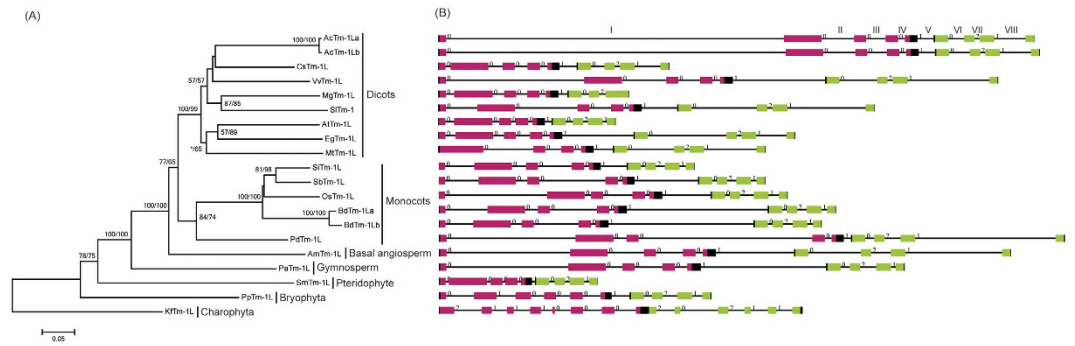
Zefeng Yang<sup>1</sup>, Li Liu<sup>1</sup>, Huimin Fang<sup>1</sup>, Pengcheng Li<sup>1</sup>, Shuhui Xu<sup>1</sup>, Wei Cao<sup>1</sup>, Chenwu Xu<sup>1</sup>, Jinling Huang<sup>2</sup> & Yong Zhou<sup>1</sup>

The *Tomato mosaic virus* (ToMV) resistance gene *Tm-1* encodes a direct inhibitor of ToMV RNA replication to protect tomato from infection. The plant *Tm-1-like* (*Tm-1L*) protein is predicted to contain an uncharacterized N-terminal UPF0261 domain and a C-terminal TIM-barrel signal transduction (TBST) domain. Homologous searches revealed that proteins containing both of these two domains are mainly present in charophyte green algae and land plants but absent from glaucophytes, red algae and chlorophyte green algae. Although *Tm-1* homologs are widely present in bacteria, archaea and fungi, UPF0261- and TBST-domain-containing proteins are generally encoded by different genes in these lineages. A co-evolution analysis also suggested a putative interaction between UPF0261- and TBST-domain-containing proteins. Phylogenetic analyses based on homologs of these two domains revealed that plants have acquired UPF0261- and TBST-domain-encoding genes through two independent horizontal gene transfer (HGT) events before the origin of land plants from charophytes. Subsequently, gene fusion occurred between these two horizontally acquired genes and resulted in the origin of the *Tm-1L* gene in streptophytes. Our results demonstrate a novel evolutionary mechanism through which the recipient organism may acquire genes with functional interaction through two different HGT events and further fuse them into one functional gene.

Green plants, also known as Viridiplantae, consist of green algae (Chlorophyta and Streptophyta) and land plants (embryophytes). The difference between Streptophyta and Chlorophyta is correlated with a remarkably conservative preference for freshwater/marine habitats, and the early freshwater adaptation of streptophyte algae was a major advantage for the colonization of land by plants<sup>1</sup>. The origin of land plants is one of the most important events in the evolution of life on Earth and is the key step in the process of developing of modern terrestrial ecosystems. From peaceably aquatic surroundings to harshly terrestrial environments, the ancestor of land plants had to face significant stresses, including dryness, ultraviolet radiation and pathogenic microorganisms other than those in water. During their colonization of land, plants gradually evolved new genes and phenotypic novelties to adapt to and radiate in terrestrial environments<sup>2,3</sup>. The genome of *Klebsormidium flaccidum* reveals that this filamentous terrestrial alga has acquired many genes, including those producing several plant hormones, and a primitive system to protect against the harmful effects of high-intensity light<sup>4</sup>. Because plants are constantly exposed to microbes, the evolution of land plants has been shaped by molecular interactions with epiphytic, symbiotic and pathogenic microbes. Thus, evolution of new genes for resistance to pathogenic microorganisms was extremely important for plants to colonize the land<sup>5</sup>.

The tomato (*Solanum lycopersicum* L.) gene *Tm-1*, conferring resistance to *Tomato mosaic virus* (ToMV), was bred from the wild tomato *S. habrochaites*<sup>6,7</sup>. Although the *Tm-1* protein does not share any functional domain with previously known resistance (R) proteins, it physically binds ToMV replication proteins and functionally inhibits the RNA-dependent RNA replication of ToMV<sup>8,9</sup>. In *S. habrochaites*, a small region of *Tm-1* was under

<sup>1</sup>Jiangsu Key Laboratory of Crop Genetics and Physiology/Co-Innovation Center for Modern Production Technology of Grain Crops, Key Laboratory of Plant Functional Genomics of the Ministry of Education, Yangzhou University, Yangzhou, 225009, China. <sup>2</sup>Department of Biology, East Carolina University, Greenville, NC, 27858, USA. Correspondence and requests for materials should be addressed to C.X. (email: qtls@yzu.edu.cn) or J.H. (email: huangj@ecu.edu) or Y.Z. (email: zhouyong@yzu.edu.cn)



**Figure 1.** The phylogenetic tree of the green plant *Tm-1L* genes (A) and their exon/intron structures (B). The numbers above the branches represent the bootstrap values for the maximum likelihood and distance analyses, respectively. The asterisks indicate values <50%. The exons are indicated by boxes, whereas introns are indicated by lines. The UPF0261 domain regions are indicated by red boxes, while the TBST domain regions are indicated by green boxes. The number above an intron indicates the phase.

positive selection during its antagonistic coevolution with ToMV, and the positively selected residues have evolved to counter ToMV infection and are important for the inhibition of ToMV RNA replication<sup>10</sup>. Further evidence also confirmed this coevolution relationship between *Tm-1* and ToMV, because a single naturally occurring amino acid change in the positively selected region of *Tm-1* enables it to inhibit the replication of a *Tm-1* resistance-breaking ToMV mutant<sup>6,10</sup>. The tomato *Tm-1* protein is predicted to contain two evolutionarily conserved domains: an uncharacterized N-terminal domain and a C-terminal TIM-barrel-like domain<sup>8</sup>. Experimental evidence has revealed that the *Tm-1* N-terminal region was responsible for its inhibitory activity, and a C-terminally truncated *Tm-1* protein still possesses inhibitory activity<sup>6,7</sup>. These findings also suggest that the full-length tomato *Tm-1* has a function other than ToMV resistance<sup>11</sup>.

The homologs of the tomato *Tm-1* gene are present in a wide range of land plants<sup>11</sup>. Although whether these *Tm-1* homologs help to inhibit virus multiplication is unknown, the ubiquity of this gene in land plants suggests that its functions include a wide range of selectivity. In this work, we tried to obtain more insights into the origin of the plant *Tm-1-like* (*Tm-1L*) gene by performing an extensive search of its homologs in current sequence databases and by analyzing their phylogeny.

## Results

**The *Tm-1L* genes are widespread in streptophytes.** BLAST searches revealed that homologs of the tomato *Tm-1* gene are present in various streptophyte groups, including charophyte algae, bryophytes, lycophytes and seed plants. In addition, we also noticed that nearly all of the sequenced land plant genomes contained at least one *Tm-1L* gene. However, no homologs were detected in the genomes of chlorophyte algae. To investigate the origin and evolution of the land plant *Tm-1L* genes, we characterized them from species representing the main lineages of streptophytes, including the charophyte alga *K. flaccidum*, the bryophyte *Physcomitrella patens*, the lycophyte *Selaginella moellendorffii*, the gymnosperm *Picea abies*, the basal angiosperm *Amborella trichopoda* and five monocot and eight dicot angiosperms (Supplementary Table S1). Among the tested plant genomes, two (dicot *Aquilegia coerulea* and monocot *Brachypodium distachyon*) contained two *Tm-1L* genes, while all of the other genomes contained only one. In the phylogeny (Fig. 1), both of these paralogous genes were located at the termini of branches, indicating that these paralogs were formed through recent duplication events. Further investigation revealed that both paralogous pairs were results of segmental duplication because there were highly conserved genes within the flanking regions of these two pairs of paralogous *Tm-1L* genes.

Most of the land plant *Tm-1L* genes generally contained eight introns in their coding regions, suggesting a highly similar gene structure among them, and the positions and phases of these eight introns were conserved. These results also suggest that the main characteristics of the gene structure of this family were formed in the common ancestor of land plants. Some intron gain/loss events were also found in the evolution of land plant *Tm-1L* genes. For example, the *P. patens* *PpTm-1L* gene contained nine introns in the coding region, and an extra intron was inserted in the second exon. In addition, the *Medicago truncatula* *MtTm-1L* and *Mimulus guttatus* *MgTm-1L* genes contained only seven introns in the coding regions, and a gene structure comparison suggested that the *MtTm-1L* gene had lost the first intron, while the *MgTm-1L* gene had lost the last one. However, the *KfTm-1L* gene in the charophyte alga *K. flaccidum* possessed 11 introns in its coding regions, and the positions and phases of these introns were significantly different from those in land plants.

**The phylogenetic relationship of the *Tm-1L* homologs.** To investigate the origin of the streptophyte *Tm-1L* genes, we searched the *nr* and EST databases of NCBI and the available eukaryotic genome databases for *Tm-1L* homologs. However, no homologs in any other algae were detected, including glaucophytes, red algae and chlorophyte green algae. This result suggests that the plant *Tm-1L* gene might have first appeared in streptophyte algae. Blast results also reveal that homologs of plant *Tm-1L* proteins only exist in prokaryotes, fungi and *Acanthamoeba castellanii*, a species of Amoebozoa. To our surprise, only two genes in *Aspergillus*, a fungi genus, showed global similarity with plant *Tm-1L* genes, while the genes in prokaryotes and other fungi only possessed partial homeologous sequences. Some prokaryotic and fungal proteins only possess a UPF0261 domain and show

high similarity with the N terminus of plant Tm-1L proteins. However, other prokaryotic and fungal proteins only possess a TBST domain and show high similarity with the C terminus of plant Tm-1L proteins. It is difficult to build a phylogenetic tree using the whole sequences of Tm-1L homologs due to a lack of consistent sequences. Thus, two phylogenetic trees were constructed using the UPF0261 and TBST domain sequences, respectively.

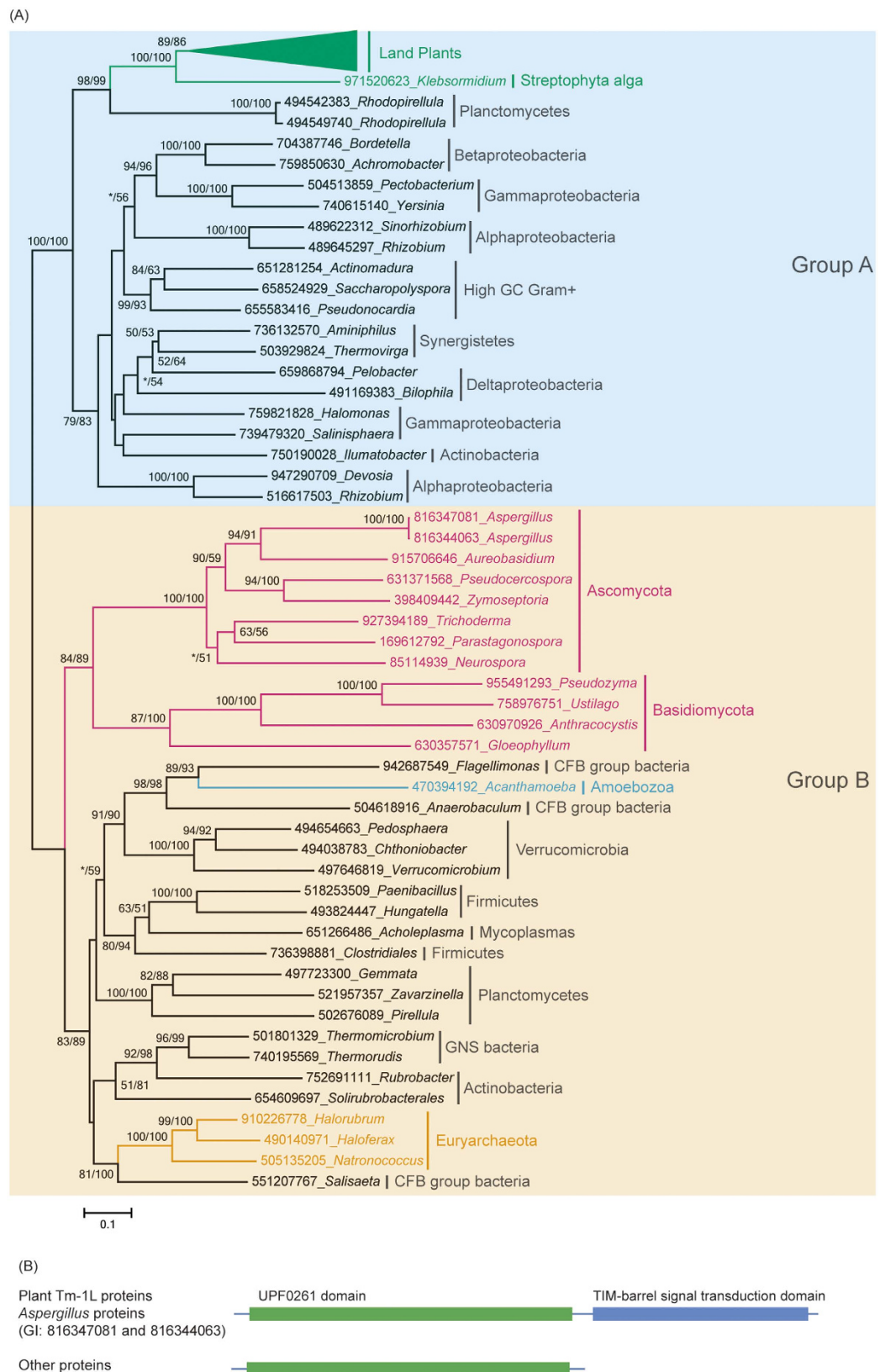
We first selected representative homologs of proteins possessing a UPF0261 domain from each taxonomic group of cellular organisms in the *nr* database to build a phylogenetic tree (Fig. 2). The genes encoding proteins with UPF0261 domains are distributed widely in bacteria, suggesting that this gene first emerged in bacteria. In addition, the molecular phylogeny of the UPF0261 domain consists of two distinct groups with high bootstrap support values. All of the genes in streptophyte plants were in group A, while their homologs in fungi were in group B. Group A only contained genes from bacteria and plants, and all of the plant UPF0261-domain-containing proteins formed a single clade with high bootstrap support. The monophyly of the plant UPF0261-domain-containing proteins strongly suggests that they have a single origin and are derived from a unique gene that was already present in streptophyte algae. The plant UPF0261-domain-containing proteins fell within the branch of planctomycete genes, showing high bootstrap support values in both maximum likelihood and distance analyses. These observations indicate that the origin of plant UPF0261-domain-containing protein genes is a result of horizontal gene transfer and that the putative donor might be a planctomycete.

Group B contained genes from fungi, amoebozoa, bacteria and archaea. The fungal genes encoding a UPF0261-domain-containing protein formed a single clade with high bootstrap support. However, blast searches revealed that fungal UPF0261-domain-containing proteins are only present in the subkingdom Dikarya. The monophyly of the fungal UPF0261-domain-containing protein orthologs strongly indicates that they have a single origin and are derived from a gene that was already present at least in the ancestor of Dikarya. These results also suggest that the origin of this gene in fungi is another HGT event from prokaryotes. The genome of the free-living amoeba species *Acanthamoeba castellanii* also possesses an UPF0261-domain-containing protein gene that is located in group B and falls into the branch of bacteria. There were no homologs in close relatives of *A. castellanii*. This result suggests that the origin of the UPF0261-domain-containing protein gene in *A. castellanii* is another horizontal gene transfer event from bacteria.

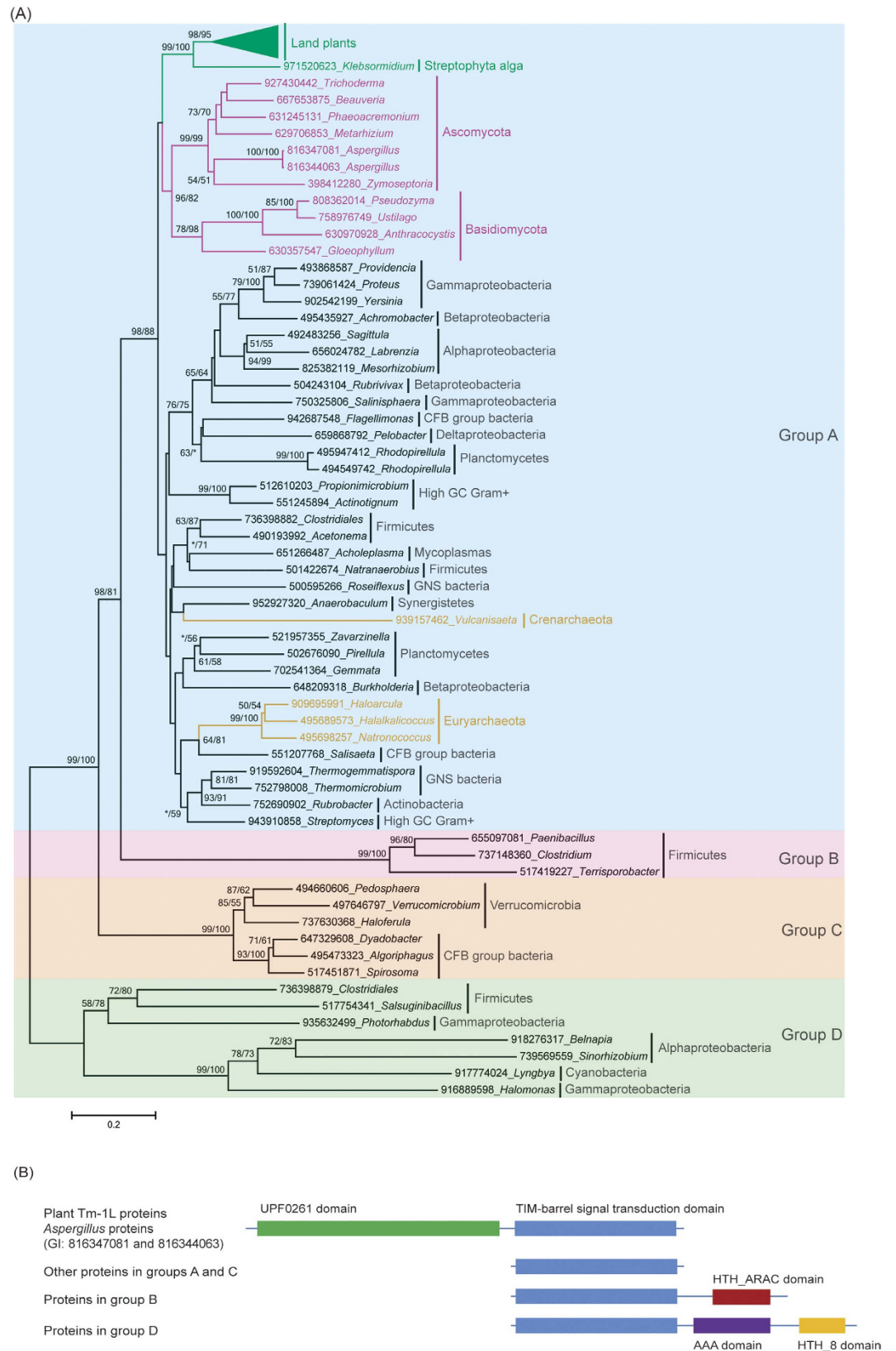
Second, we also selected representative homologs of proteins possessing a TBST domain from each taxonomic group of cellular organisms in the *nr* database to build a phylogenetic tree (Fig. 3). According to the phylogenetic tree and the domain structure, the TBST-domain-containing proteins were divided into four evolutionarily distinct groups. Group A contained genes from bacteria, archaea and eukaryotes, while the other groups only possessed genes from bacteria. The protein sequences within each group showed high similarity, while those between groups were quite divergent. On the phylogenetic tree of TBST-domain-containing proteins, the eukaryotic cluster exclusively contained orthologs from green plants and fungi but no other eukaryotic lineage, forming two well-supported monophyletic sister groups. The genes encoding TBST-domain-containing proteins were widely distributed in prokaryotes and mainly in bacteria. The universality of their distribution in bacteria also suggests that this gene first emerged in bacteria. Both the plant and fungal TBST-domain-containing proteins formed a single clade with high bootstrap support, suggesting that they have a single origin in these two lineages, respectively. The evolutionary relationship of TBST homologs from plants and fungi clearly indicates a common origin. However, the kingdom Plantae does not share a specific ancestor with fungi except for the common ancestor of eukaryotes<sup>12,13</sup>. The first scenario for the origin of the TBST gene in eukaryotes is that it was present in the ancestor of all eukaryotes. However, this scenario requires too many independent gene loss events, which seems unlikely. The second scenario is that a TBST gene originated either in the ancestor of green plants or in the ancestor of at least Dikarya fungi through one HGT event and then was transferred via another HGT event between these two subkingdoms (Fig. 4A). The third scenario is that two independent HGT events transferred the TBST gene from prokaryotes to the ancestor of land plants and to the ancestor of Dikarya fungi, respectively. Although we cannot confirm which of the last two scenarios is more parsimonious, at least one HGT event is suggested to contribute to the origin of the TBST domain in streptophyte plant *Tm-1L* genes.

**Co-evolution between the genes encoding UPF0261 and TBST proteins.** To evaluate the possible interaction between the proteins containing UPF0261 and TBST domains, the Mirrortree server<sup>14</sup> was used to access the co-evolution between these two protein families in bacteria, archaea and fungi. When we used the sequences of the UPF0261 and TBST domain regions of the *Arabidopsis* Tm-1L protein as the reference sequences, a total of 159 pairs of homologs were selected by the Mirrortree server. An analysis of the full-length UPF0261 and TBST sequences from bacteria, archaea and fungi showed that the phylogenies of these two protein families are topologically similar (Supplementary Figure S1). Mirrortree analysis revealed that the correlation coefficient of these trees was 0.886 ( $P < 0.0001$ ) (Table 1), suggesting the co-evolution of both proteins across bacteria, archaea and fungi.

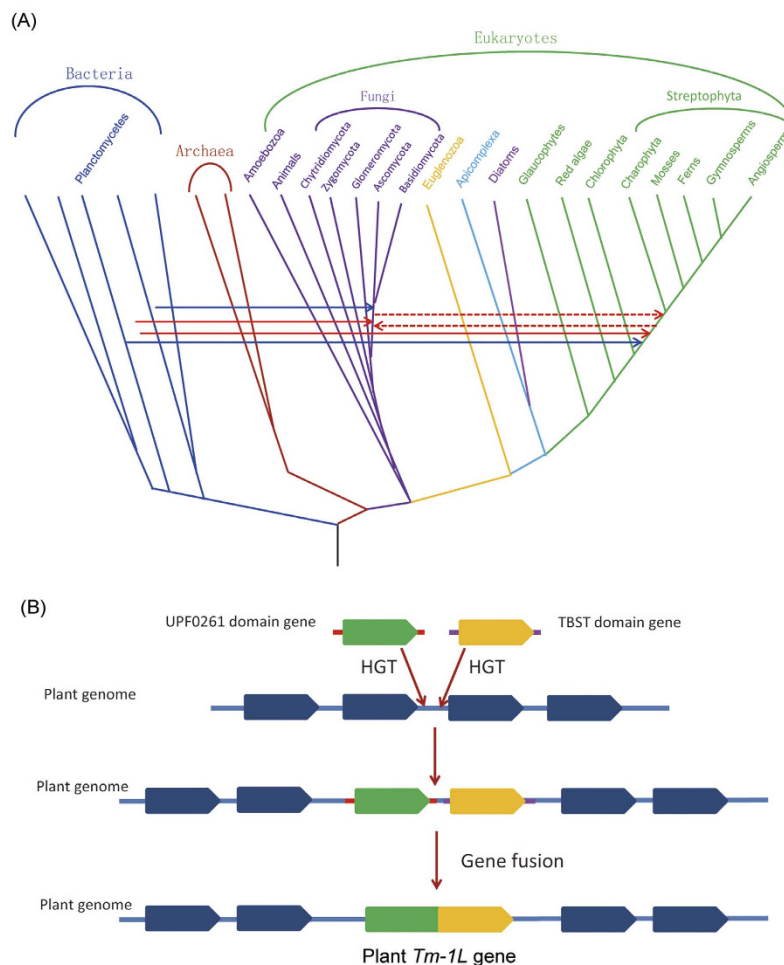
In the NCBI RefSeq protein database, at least 714 TBST-domain-containing proteins were carried by 504 bacterial genomes. Among these genomes, 491 (97.4%) contained genes encoding UPF0261 proteins. A total of 338 genomes contained only one TBST-domain-encoding gene and only one UPF0261-domain-encoding gene. We selected the 333 genomes to evaluate the co-evolution between TBST and UPF0261 genes in bacteria because the other five genomes carried genes with truncated TBST and/or UPF0261 domains (Supplementary Table S2). The pairwise distances between sequences of the TBST genes were strongly correlated with those between UPF0261 genes (Table 1), and the estimated correlation coefficient  $r$  was 0.7735 ( $P < 0.0001$ ). Furthermore, a correlation between the TBST and UPF0261 genes was also found in archaea and fungi. In the NCBI RefSeq database, 24 archaea genomes were found to contain TBST genes, and a total of 22 genomes (91.67%) had only one TBST gene and only one UPF0261 gene. The estimated  $r$  was 0.7185, which was also significantly higher than zero at the level of  $P < 0.0001$ . In fungi, 68 genomes contained TBST genes, and 61 (89.7%) carried only one TBST gene and one UPF0261 gene. The estimated  $r$  was 0.8067. When we used all of these bacteria, archaea and fungi genomes to



**Figure 2. Phylogenetic analyses and the domain structures of the genes encoding UPF0261-domain-containing proteins.** (A) Phylogenetic analyses of the proteins containing a UPF0261 domain. The numbers above the branches represent the bootstrap values for the maximum likelihood and distance analyses. All of the sequences were obtained from NCBI, except for those in the green plants, and each protein is indicated by the GI numbers in NCBI and its genus. (B) The domain structure of the proteins that were used in the phylogeny.



**Figure 3. Phylogenetic analyses and the domain structures of the genes encoding TBST-domain-containing proteins.** (A) Phylogenetic analyses of the proteins containing a TBST domain. The numbers above the branches represent the bootstrap values for the maximum likelihood and distance analyses. All of the sequences were obtained from the NCBI, except for those in the green plants, and each protein is indicated by the GI numbers in NCBI and its genus. (B) The domain structure of the proteins that were used in the phylogeny.



**Figure 4. An evolutionary scenario for the origin of the plant *Tm-1L* gene.** (A) Horizontal gene transfer events associated with the origin of *Tm-1* homologs in land plants and fungi. Solid blue arrows indicate the HGT of the UPF0261-domain-containing gene. Two solid red arrows indicate the HGT of the TBST-domain-containing gene from prokaryotes to plants and fungi, respectively. Dashed red arrows indicate the HGT of the TBST-domain-containing gene between plants and fungi. (B) Evolutionary history of the green plant *Tm-1L* gene.

Lineage	RefSeq proteins			Mirrortree		
	Genomes	<i>r</i>	<i>P</i>	Genomes	<i>r</i>	<i>P</i>
Common	416	0.7192	<0.0001	159	0.886	<0.0001
Bacteria	333	0.7735	<0.0001	136	0.906	<0.0001
Archaea	22	0.7185	<0.0001	3	0.990	<0.05
Fungi	61	0.8067	<0.0001	19	0.867	<0.0001

**Table 1. Pearson correlation coefficients for the co-evolution of the UPF0261 and TBST proteins using the RefSeq protein dataset and Mirrortree method.**

perform a correlation analysis for the distance matrices between TBST and UPF0261 proteins, the estimated *r* was 0.7192. These results indicate that the TBST and UPF0261 proteins underwent highly correlated co-evolution in bacteria, archaea and fungi. In addition, the TBST and UPF0261 domains in plant *Tm-1* proteins were also used to investigate their possible domain-domain coevolution. The estimated *r* between the distances of the two domains was 0.8675 ( $P < 0.0001$ ), suggesting that highly correlated co-evolution between the TBST and UPF0261 domains in *Tm-1* proteins occurred during plant evolution. However, the correlation coefficient in plants was not statistically larger than the average value in bacteria, archaea and fungi ( $P = 0.0778$ , one-sample *t*-test).

## Discussion

**Horizontal gene transfer and the origin of plant *Tm-1L* genes.** Horizontal gene transfer (HGT), also known as lateral gene transfer (LGT), refers to the transformation of genetic material between organisms with reproductive isolation<sup>15</sup>. HGT has been thought to be one of the most important evolutionary forces and to be frequent only within prokaryotes and certain unicellular eukaryotes<sup>16,17</sup>. Recent genome analyses also detected horizontally acquired genes in all major lineages of multicellular eukaryotes, including plants, fungi and animals<sup>18–21</sup>, suggesting that HGT was also critical for adaptive evolution throughout eukaryotic evolution. In green plants, many horizontally acquired genes were also found to be of adaptive and evolutionary importance<sup>22</sup>. Because intimate physical association may facilitate HGT, parasitic plants have been used to detect HGT<sup>23,24</sup>. For instance, based on a thorough transcriptome screening, a *strictosidine synthase-like* (*SSL*) gene was found to be independently transferred from Brassicaceae to the root parasitic plant *Orobanche aegyptiaca* and to the shoot parasitic plant *Cuscuta australis*. Expression analyses also suggested that foreign *SSL* genes may retain certain functions in the recipient species<sup>25</sup>.  $C_4$  photosynthesis, a complex trait that confers higher productivity under warm and arid conditions, evolved independently many times from  $C_3$  ancestors. Christin and his colleagues<sup>26</sup> revealed that the grass lineage *Alloteropsis* acquired two essential genes of the  $C_4$  pathway *via* a minimum of four independent lateral gene transfers from  $C_4$  taxa in the past 10 million years.

One of the most momentous events in the evolution of green plants is the colonization of land, which contributed tremendously to the establishment of the modern terrestrial environment. During their transition from aquatic to terrestrial environments, plants evolved some complex regulatory systems, body plans and other phenotypic novelties to conquer tremendous challenges, such as increased UV irradiation, drought, heat and microbial infection<sup>13</sup>. Many of the genes that were horizontally acquired by early land plants are involved in many plant-specific activities, including vascular development, plant defense, nitrogen recycling and the biosynthesis of starch, polyamines, hormones and glutathione<sup>27–30</sup>, suggesting the importance of HGT in adaptation during plant colonization of land. Land plants evolved from charophycean green algae *c.* 480–490 million years ago<sup>31</sup>. The *Tm-1L* gene is present in all of the lineages of land plants and *Klebsormidium*, which belongs to the charophyte alga that comprise streptophytes with land plants. However, this gene is absent from chlorophyte green algae. These results suggest that green plants acquired the *Tm-1L* gene before the split of land plants with charophytic algae. The phylogenetic relationship between *Tm-1L* homologs revealed that green plants gained this gene through HGT events. Surprisingly, at least two different HGT events, which had respectively contributed to the origin of two domains of the protein encoded by the *Tm-1L* gene, were included in the origin of this gene. The charophytic algae *Klebsormidium* species have primitive body plans, and most that have adapted to land also can survive in fresh water. In addition, some *Klebsormidium* species also show tolerance to typical terrestrial stresses, such as drought, freezing and high-intensity light<sup>4,32,33</sup>. The ubiquity of the *Tm-1L* gene in land plants and in the charophytic algae *Klebsormidium* suggests that the functions of this gene include a wide range of selectivity and might have been involved in the adaptive evolution from aquatic to land environments.

**Putative interaction between UPF0261- and TBST-domain-containing proteins.** In cell biology, proteins rarely perform functions alone. Many molecular and biological processes within a cell are carried out by the synergy of a large number of protein components that are organized by their protein-protein interactions, which is a fundamental mechanism that underlies virtually all biological processes. Because many important interactions between proteins are conserved across lineages and/or species, the maintenance of these relationships would lead to a high degree of co-evolution between genetic matrices of proteins with interactive relationships<sup>34</sup>. A correlation between genetic matrices of two proteins can quantitatively measure the co-evolution between them and, therefore, make specific and impartial inferences regarding probable protein-protein interactions<sup>35</sup>. In this work, we noticed that if there is a TBST-domain-encoding gene in a bacterial, archaeal or fungal genome, the genome generally has at least one UPF0261-domain-encoding gene. In addition, most genomes containing these two genes only possess one copy of each. More importantly, both the methods of Mirrortree and distance revealed that these two genes co-evolved in bacteria, archaea and fungi. Although our results do not provide structural evidence of the direct binding between the UPF0261- and TBST-domain-containing proteins, high evolutionary correlation and putative interaction between them were suggested. When the genome acquired one of the UPF0261- and TBST-domain-encoding genes through HGT, the interaction between them may have forced the plant genome to gain the other.

**Evolutionary scenarios.** Two independent HGT events contributed to the acquisition of the UPF0261 and TBST domains in the *Tm-1L* gene of the land plant ancestor, respectively. However, unlike the genes in other species, the plant *Tm-1L* gene encodes a protein with both of these domains. Thus, a gene fusion event is suggested between the genes encoding UPF0261- and TBST-domain-containing proteins in the genome of the land plant ancestor. The story of the evolution of land plant *Tm-1L* genes is dramatic and fascinating, and at least two steps were included in the origin of this functionally important gene (Fig. 4B). In the first step, the ancestor of streptophytes acquired two bacterial genes through two independent HGT events. These two genes encode UPF0261- and TBST-domain-containing proteins, respectively, and are located in adjacent regions in the plant genome. In the second step, the ancestor of streptophytes evolved the *Tm-1L* gene through the fusion of the UPF0261-domain-encoding gene with the TBST-domain-encoding gene. In our investigation, the origin of the plant *Tm-1L* gene is the result of a combination of two independent HGT events and one gene fusion event.

Both HGT and gene fusion are of evolutionary importance in creating new genes with novel functions<sup>36</sup>. It has been suggested that the benefit and the driving force behind gene fusion is to lower the regulational load of multiple interacting gene products<sup>37</sup>. The accretion of multiple domains appears to be one of the important routes for increasing functional complexity in the evolution of multicellular eukaryotes<sup>38</sup>. In addition, gene fusion is suggested to be one of the main principles of functional design in signal transduction systems, such as sugar

phosphotransferase systems and receptor kinases<sup>39,40</sup>. The fusion of UPF0261 and TBST domains led to the origin of the Tm-1L protein in plants. The origin of the plant Tm-1L protein may also be included the functional design in signal transduction systems because the C-terminal domain TBST has been suggested to possess a potential role in signal recognition/receiving and signal transduction.

The combination of HGT and gene fusion has been previously reported, and two types of mechanisms have been suggested. One mechanism is the HGT of fused genes. For example, the origin of the fern neochrome involved the retrotransposition of a phototropin gene, fusion with a phytochrome and HGT from hornworts to ferns<sup>41</sup>. The other mechanism is the fusion of the horizontally acquired gene with a pre-existing gene in the acceptor genome. Phylogenetic evidence has revealed that plant expansin genes that encode cell-wall-loosening proteins were transferred from plants to bacteria. Subsequently, expansin genes have been suggested to be independently fused to the genes that code for an endoglucanase domain or for a carbohydrate-binding module in bacteria<sup>42</sup>. Here, we report a rare case of the origin of plant genes through the combination of HGT and gene fusion. In this case, the ancestral genome of land plants independently acquired two genes with a putatively functional interaction through two HGT events, and a subsequent gene fusion event that joined these two genes led to the origin of the *Tm-1L* gene in plants.

## Materials and Methods

**Sequence data sources.** To identify the genes encoding Tm-1L proteins in plants, the protein sequence of the tomato gene *SITm-1* was used as a query to in the Phytozome<sup>43</sup>, NCBI *nr* protein and spruce genome project<sup>44</sup> databases. If a protein sequence satisfied  $E \leq 1E-10$ , it was selected as a candidate protein. Then, Pfam<sup>45</sup> was used to predict the UPF0261 and TIM-barrel signal transduction (TBST) domains. The new *Tm-1L* sequences that were detected in plants were used reiteratively to search the respective sequence database.

To identify the homologs of plant *Tm-1L* genes, BLAST searches against the non-redundant (*nr*) protein sequence, NCBI EST and available eukaryotic genome databases were performed using the plant Tm-1L protein sequences as queries. The obtained hits were further analyzed *via* a Pfam search to confirm the presence of the UPF0261 and/or TBST domains in the protein structure. Protein sequences were sampled for a further combined phylogenetic analysis from representative groups within each domain of life (bacteria, archaea and eukaryotes) based on the BLASTP results.

**Phylogenetic tree reconstruction.** All of the selected representative protein sequences were aligned using Clustal X<sup>46</sup>. The gaps and ambiguously aligned sites were removed manually. Phylogenetic analyses were performed using a maximum likelihood (ML) approach with PhyML version 3.0<sup>47</sup> and a neighbor-joining (NJ) method using MEGA version 6.0<sup>48</sup>. The model of protein evolution that best fit the protein alignment, the proportion of invariable sites and the alpha parameter of the gamma distribution were estimated using the ModelGenerator program<sup>49</sup>. A total of 100 non-parametric bootstrap samplings were carried out to estimate the support level for each internal branch for both the ML and NJ trees. Phylogenetic trees were visualized using the explorer program in MEGA.

**Evaluation of genetic distance and co-evolution analyses.** The Mirrortree web server<sup>14</sup> was used to estimate the degree of co-evolution between TBST and UPF0261 proteins. In this analysis, the sequences of the UPF0261 and TBST domain regions in the *Arabidopsis* Tm-1L protein (AT5G66420) were used as two references.

All of the bacterial, archaeal and fungal genes encoding TBST-domain-containing proteins in the NCBI CDD database were retrieved. In this analysis, only the genes with RefSeq genomes were selected. Blast tools were further used to verify the existence of the genes encoding UPF0261-domain-containing proteins in these genomes. To test the putative co-evolution, the linear correlation coefficient between the genetic distance matrices of TBST and UPF0261 proteins was determined<sup>50</sup>. The TBST and UPF0261 protein sequences in these three lineages were aligned using Clustal X<sup>46</sup>. Distance matrices were computed for each of the alignments by MEGA using the Jones-Taylor-Thornton (JTT) model for amino acid replacement per site, which was also estimated by using the ModelGenerator program. For TBST and UPF0261 proteins within *n* genomes in common in their multiple sequence alignment, the correlation coefficient *r* could be calculated as follows:

$$r = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (X_{ij} - \bar{x})(Y_{ij} - \bar{y})}{\sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (X_{ij} - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (Y_{ij} - \bar{y})^2}} \quad (1)$$

where  $X_{ij}$  is the genetic distance for the TBST protein between genomes *i* and *j*, while  $Y_{ij}$  is the genetic distance for protein UPF0261 between genomes *i* and *j*.

## References

1. Becker, B. & Marin, B. Streptophyte algae and the origin of embryophytes. *Annals of botany* **103**, 999–1004 (2009).
2. Graham, L. E., Cook, M. E. & Busse, J. S. The origin of plants: body plan changes contributing to a major evolutionary radiation. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 4535–4540 (2000).
3. Pires, N. D. & Dolan, L. Morphological evolution in land plants: new designs with old genes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **367**, 508–518 (2012).
4. Hori, K. *et al.* Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nature communications* **5**, 3978 (2014).
5. Ponce de Leon, I., Hamberg, M. & Castresana, C. Oxylinins in moss development and defense. *Frontiers in plant science* **6**, 483 (2015).
6. Ishibashi, K. *et al.* Structural basis for the recognition-evasion arms race between Tomato mosaic virus and the resistance gene Tm-1. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E3486–E3495 (2014).



7. Kato, M., Ishibashi, K., Kobayashi, C., Ishikawa, M. & Katoh, E. Expression, purification, and functional characterization of an N-terminal fragment of the tomato mosaic virus resistance protein Tm-1. *Protein expression and purification* **89**, 1–6 (2013).
8. Ishibashi, K., Masuda, K., Naito, S., Meshi, T. & Ishikawa, M. An inhibitor of viral RNA replication is encoded by a plant resistance gene. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 13833–13838 (2007).
9. Lee, J. M., Oh, C.-S. & Yeam, I. Molecular Markers for Selecting Diverse Disease Resistances in Tomato Breeding Programs. *Plant Breeding and Biotechnology* **3**, 308–322 (2015).
10. Ishibashi, K. *et al.* Coevolution and hierarchical interactions of Tomato mosaic virus and the resistance gene Tm-1. *PLoS pathogens* **8**, e1002975 (2012).
11. Ishibashi, K. & Ishikawa, M. Mechanisms of tomato mosaic virus RNA replication and its inhibition by the host resistance factor Tm-1. *Current opinion in virology* **9**, 8–13 (2014).
12. Baldauf, S. L. The deep roots of eukaryotes. *Science* **300**, 1703–1706 (2003).
13. Emiliani, G., Fondi, M., Fani, R. & Gribaldo, S. A horizontal gene transfer at the origin of phenylpropanoid metabolism: a key adaptation of plants to land. *Biology Direct* **4**, 1–12 (2009).
14. Ochoa, D. & Pazos, F. Studying the co-evolution of protein families with the Mirrortree web server. *Bioinformatics* **26**, 1370–1371 (2010).
15. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nature reviews. Genetics* **16**, 472–482 (2015).
16. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. *Molecular biology and evolution* **19**, 2226–2238 (2002).
17. Andersson, J. O. Lateral gene transfer in eukaryotes. *Cellular and molecular life sciences: CMLS* **62**, 1182–1197 (2005).
18. Bock, R. The give-and-take of DNA: horizontal gene transfer in plants. *Trends in plant science* **15**, 11–22 (2010).
19. Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nature reviews. Genetics* **9**, 605–618 (2008).
20. Hotopp, J. C. D. Horizontal gene transfer between bacteria and animals. *Trends in Genetics* **27**, 157–163 (2011).
21. Fitzpatrick, D. A. Horizontal gene transfer in fungi. *FEMS microbiology letters* **329**, 1–8 (2012).
22. Wang, Q., Sun, H. & Huang, J. The evolution of land plants: a perspective from horizontal gene transfer. *Acta Societatis Botanicorum Poloniae* **83** (2014).
23. Davis, C. C. & Wurdack, K. J. Host-to-parasite gene transfer in flowering plants: phylogenetic evidence from Malpighiales. *Science* **305**, 676–678 (2004).
24. Xi, Z. *et al.* Horizontal transfer of expressed genes in a parasitic flowering plant. *BMC genomics* **13**, 227 (2012).
25. Zhang, D. *et al.* Root parasitic plant *Orobanche aegyptiaca* and shoot parasitic plant *Cuscuta australis* obtained Brassicaceae-specific strictosidine synthase-like genes by horizontal gene transfer. *BMC plant biology* **14**, 1 (2014).
26. Christin, P.-A. *et al.* Adaptive evolution of C 4 photosynthesis through recurrent lateral gene transfer. *Current Biology* **22**, 445–449 (2012).
27. Yang, Z. *et al.* Evolution of land plant genes encoding L-Ala-D/L-Glu epimerases (AEEs) via horizontal gene transfer and positive selection. *BMC plant biology* **13**, 34 (2013).
28. Yang, Z. *et al.* Ancient horizontal transfer of transaldolase-like protein gene and its role in plant vascular development. *The New phytologist* **206**, 807–816 (2015).
29. Yue, J., Hu, X. & Huang, J. Horizontal gene transfer in the innovation and adaptation of land plants. *Plant signaling & behavior* **8**, e24130 (2013).
30. Yue, J., Hu, X., Sun, H., Yang, Y. & Huang, J. Widespread impact of horizontal gene transfer on plant colonization of land. *Nature communications* **3**, 1152 (2012).
31. Sanderson, M. J., Thorne, J. L., Wikstrom, N. & Bremer, K. Molecular evidence on plant divergence times. *American journal of botany* **91**, 1656–1665 (2004).
32. Rindi, F., Mikhailyuk, T. I., Sluiman, H. J., Friedl, T. & Lopez-Bautista, J. M. Phylogenetic relationships in Interfilum and Klebsormidium (Klebsormidiophyceae, Streptophyta). *Molecular phylogenetics and evolution* **58**, 218–231 (2011).
33. Karsten, U. & Holzinger, A. Light, temperature, and desiccation effects on photosynthetic activity, and drought-induced ultrastructural changes in the green alga *Klebsormidium dissectum* (Streptophyta) from a high alpine soil crust. *Microbial ecology* **63**, 51–63 (2012).
34. Feinauer, C., Szurmant, H., Weigt, M. & Pagnani, A. Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon. *Plos One* **11**, e0149166 (2016).
35. Goh, C.-S. & Cohen, F. E. Co-evolutionary analysis reveals insights into protein–protein interactions. *Journal of molecular biology* **324**, 177–192 (2002).
36. Long, M., Betran, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nature reviews. Genetics* **4**, 865–875 (2003).
37. Skrabanek, L., Saini, H. K., Bader, G. D. & Enright, A. J. Computational prediction of protein–protein interactions. *Molecular biotechnology* **38**, 1–17 (2008).
38. Yanai, I., Wolf, Y. I. & Koonin, E. V. Evolution of gene fusions: horizontal transfer versus independent events. *Genome biology* **3**, research0024 (2002).
39. Reizer, J. & Saier, M. H. Jr. Modular multidomain phosphoryl transfer proteins of bacteria. *Current opinion in structural biology* **7**, 407–415 (1997).
40. Wiesner, T. *et al.* Kinase fusions are frequent in Spitz tumours and spitzoid melanomas. *Nature communications* **5**, 3116 (2014).
41. Li, F. W. *et al.* Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 6672–6677 (2014).
42. Nikolaidis, N., Doran, N. & Cosgrove, D. J. Plant expansins in bacteria and fungi: evolution by horizontal gene transfer and independent domain fusion. *Molecular biology and evolution* **31**, 376–386 (2014).
43. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* **40**, D1178–D1186 (2012).
44. Nystedt, B. *et al.* The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**, 579–584 (2013).
45. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic acids research* **44**, D279–D285 (2016).
46. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
47. Guindon, S., Delsuc, F., Dufayard, J. F. & Gascuel, O. Estimating maximum likelihood phylogenies with PhyML. *Methods in molecular biology* **537**, 113–137 (2009).
48. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* **30**, 2725–2729 (2013).
49. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McLnerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC evolutionary biology* **6**, 29 (2006).
50. Goh, C.-S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. Co-evolution of proteins with their interaction partners. *Journal of molecular biology* **299**, 283–293 (2000).

## Acknowledgements

This work was supported by grants from the National Natural Science Foundations of China (31391632 and 91535103), the Priority Academic Program Development of Jiangsu Higher Education Institutions, the National High-tech R&D Program (863 Program, 2014AA10A601-5), the Natural Science Foundations of Jiangsu Province (BK20150010), the Natural Science Foundation of the Jiangsu Higher Education Institutions (14KJA210005) and the Innovative Research Team of Universities in Jiangsu Province.

## Author Contributions

Z.Y., C.X., J.H. and Y.Z. designed all the analyses and wrote the manuscript. Z.Y., L.L., H.F. and P.L. performed the analyses and participated in manuscript writing. Z.Y., L.L., S.X. and W.C. contributed to the data analyses. All authors read and approved the final manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Yang, Z. *et al.* Origin of the plant *Tm-1-like* gene *via* two independent horizontal transfer events and one gene fusion event. *Sci. Rep.* **6**, 33691; doi: 10.1038/srep33691 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016