Functional diversification of a basic helix–loop–helix protein due to alternative transcription during generation of amphidiploidy in tobacco plants

Yutaka KODAMA and Hiroshi SANO¹

Research and Education Center for Genetic Information, Nara Institute of Science and Technology, Nara 630-0192, Japan

A plastid-resident basic helix–loop–helix protein, previously identified in *Nicotiana tabacum* and designated as NtWIN4 (*N. tabacum* wound-induced clone 4), has been converted from a nuclear transcription repressor into a plastid-resident regulatory factor through replacement of the DNA-binding domain with a plastid transit sequence during evolution. *N. tabacum* is a natural amphidiploid plant derived from *Nicotiana tomentosiformis* and *Nicotiana sylvestris* and immunoblot staining using anti-NtWIN4 antibodies identified two protein species, a 26 kDa form and a 17 kDa form, in *N. sylvestris*, whereas only the 17 kDa form was found in *N. tabacum*. The 26 kDa protein is produced when translation starts from the first AUG codon of the mRNA and is predominantly localized in the cytoplasm and nucleus, whereas the 17 kDa protein is derived from a 24 kDa precursor protein, synthesized from the second AUG codon, and localizes only to plastids. Subsequent analyses revealed that the lengths of the mRNAs

vary in the two plant species. One major form lacks the first AUG, while minor populations possess variable 5'-untranslated regions prior to the first AUG codon. Translation of the two types produces the 24 kDa and 26 kDa proteins respectively. *In vitro* translation assays indicated that initiation frequency from the first AUG codon is higher in mRNAs from *N. sylvestris* than from *N. tabacum*. In contrast, initiation from the second AUG codon was found to be equally efficient in mRNAs from both species. These results suggest that both mRNA populations and translation efficiency changed during the amphidiploidization responsible for generation of *N. tabacum*. This scheme could reflect a molecular mechanism of protein evolution in plants.

Key words: alternative transcription, amphidiploid, *Nicotiana sylvestris*, *Nicotiana tabacum*, plastid localization, protein evolution.

INTRODUCTION

Polyploidization of plants by genome doubling is an ubiquitous event in nature [1], constituting one of the major factors inducing evolutionary changes such as mutations, genomic rearrangements and addition of new components [2]. Amphidiploidy is defined as a natural combining of two plant species, resulting in the formation of a new species with two sets of chromosomes. *Nicotiana tabacum* has been proposed as one such example, based on morphological and cytological observation, biochemical characteristics, composition of gene families and of repeated sequences [3–9] (see Figure 1A). Genomic *in situ* hybridization has also provided consistent evidence for the amphidiploidy, the *N. tabacum* genome comprising two subgenomes from *Nicotiana sylvestris* and *Nicotiana tomentosiformis* [10–12]. These morphological and genomic studies thus strongly suggest that *N. tabacum* is a natural amphidiploid which arose approximately 6 million years ago. However, the influence of polyploidization on protein function is as yet poorly understood.

Plastids such as chloroplasts are unique plant organelles, which originated from prokaryotic endosymbiotic bacteria and function in the biosynthesis of chlorophylls and fatty acids [13]. Chloroplasts are predicted to contain 3500–4000 polypeptides by proteomic analysis but the majority of them are considered to be transported into the plastids from the cytoplasm, since the maximal number of chloroplast genes is estimated to be approx. 200 [14,15]. Among transported proteins, several DNA-binding proteins have been identified, with origins not only from

prokaryotes but also from eukaryotes [16], albeit only few in number in the latter case. Proteins of eukaryotic origin generally possess motifs of nuclear transcription factors. For instance, a chloroplast nucleoid DNA-binding protein, CND41, and a plastid envelope DNA-binding protein, PEND, isolated from *N. tabacum* and *Pisum sativum* respectively [17,18], were found to contain a zinc-finger motif in the first and a basic domain plus leucine-zipper motif in the second [17,18]. These were suggested to be converted from nuclear transcription factors during evolution [19].

bHLH (basic helix–loop–helix) proteins, belonging to a large family of nuclear transcriptional factors, feature a bHLH domain which consists of approximately 60 amino acids with two different functional regions for DNA binding and dimerization. The basic region, which is located at the N-terminal end of the domain, binds to the consensus hexanucleotide E-box (CANNTG), while the helix–loop–helix region, located at the C-terminal end, forms homo- and/or heterodimers [20,21]. A number of bHLH proteins have been identified in both plants and animals and these are considered to be important regulatory factors in nuclear transcriptional networks [22–25]. The genome of *Arabidopsis thaliana* has been shown to contain as many as 147 genes encoding bHLH proteins, these apparently constituting the largest nuclear transcription factor family [24,25]. In our previous studies, we found a novel bHLH protein from *N. tabacum*, designated as NtWIN4 (*N. tabacum* wound-induced clone 4), localized in chloroplasts [19,26,27]. Functional and structural analyses indicated involvement of NtWIN4 in stress responses, and indicated that it has been converted from a nuclear

Abbreviations used: bHLH, basic helix–loop–helix; BY2, bright yellow 2; GFP, green fluorescent protein; NsWIN4, *Nicotiana sylvestris* wound-induced clone 4; NtWIN4, *Nicotiana tabacum* wound-induced clone 4; 5'-RACE, 5'-rapid amplification of cDNA ends; RT, reverse transcriptase; UTR, untranslated region.

¹ To whom correspondence should be addressed (email sano@gtc.naist.jp).

transcriptional repressor into a plastid-resident regulatory factor during plant evolution [26]. Amino acid dissection experiments showed that full length NtWIN4, synthesized from the first AUG codon of the mRNA, localizes in the nucleus and cytoplasm. In contrast, the protein synthesized from the in-frame second AUG codon lacks the first 26 amino acids from the N-terminus and localizes into plastids after removal of a transit peptide of 67 amino acids [26]. These observations indicate that not all bHLH proteins are nuclear transcription factors.

In the present study, we attempted to elucidate the differential localization mechanism of NtWIN4, and found that size diversity of mRNAs and differing translation efficiencies might be responsible for the observed phenomena. In the present study we provide an argument that altered protein localization may arise at the time of polyploidization during evolution, perhaps due to natural selection forces.

EXPERIMENTAL

Plant materials and wound treatment

Tobacco (*N. sylvestris*, *N. tomentosiformis* and *N. tabacum* cv. Xanthi NC) plants were grown on soil in a growth cabinet at 23 °C under a 14 h light/10 h dark photo-cycle. Tobacco cultured cells, BY2 (bright yellow 2), were maintained as suspension cultures in modified Linsmaier and Skoog medium [28] at 23 °C continuously in the dark. Wound stress was applied by cutting mature leaves of *N. sylvestris* and *N. tabacum* with a pair of scissors, and wounded leaves were harvested at appropriate time points.

PCR and RT (reverse transcriptase)-PCR analyses

Genomic DNA was isolated from green leaves by the cetyltrimethyl-ammonium bromide method [29] with modifications. Total RNA was isolated by the aurintricarboxylic acid method [30]. For PCR and RT-PCR analyses, *NtWIN4* genomic and cDNA fragments were amplified by PCR with ExTaq DNA polymerase (Takara) using *NtWIN4* forward (5'-GCTCTCGAGATGAATCAAAGTGCTTTTGC-3') and *NtWIN4* reverse (5'-TCA-CCCATGGGTTCTGTTCTGTGGCTGAAG-3') primer sets. The 5'-terminus sequence was determined by the 5'-RACE (5'-rapid amplification of cDNA ends) method using a 5'-RACE core set (Takara) with the primer 5'-ATGATTCGCCGATACATTTCCA-3' (nucleotide position from the 5'-end; 161–182).

Epifluorescence analyses

Full length or partial cDNA regions of *NsWIN4* (*N. sylvestris* wound-induced clone 4) and *NtWIN4* were subcloned into the SalI/NcoI sites of a CMV (cauliflower mosaic virus) 35S-sGFP(S65T)-nos vector, harbouring a synthetic gene for improved GFP (green fluorescent protein), sGFP(S65T) driven by the CaMV35S promoter and NOS terminator [31]. Onion epidermal cell layers were bombarded with gold particles (Bio-Rad) coated with appropriate vector constructs, and, after incubation at 28 °C for 6–12 h in the dark, samples were viewed under a microscope (Olympus PROVIS AX70) equipped with a fluorescence module.

Nuclear transcription assays

GBD-NsWIN4 and *GBD-NtWIN4* effector plasmids were constructed by fusing cDNAs encoding *NsWIN4* or *NtWIN4* with the GAL4 DNA-binding domain in a yy64 vector, a derivative of pMA560 [32]. *NsWIN4* and *NtWIN4* were then subcloned into the BglII and SalI sites of yy64 respectively. As an effector control, the yy64 vector alone was used. The reporter plasmid, yy96, contained a luciferase gene placed under control of the GAL4

binding site [32]. An internal control plasmid, containing a *Renilla* luciferase gene, under control of the CaMV35S promoter, was used to normalize for differences in bombardment efficiency. BY2 cells (5 days old) were plated on 1/2x Murashige–Skoog agar and bombarded with plasmids (effector/reporter/reference=2:2:1) coated on a 1.0 µm microcarrier, under a vacuum of 28 inches of mercury using a helium pressure of 7590 kPa (PDS 1000, Bio-Rad Laboratories). Cells were placed 6 cm from the stopping screen. After bombardment, they were incubated in the dark at 28 °C for 24 h. Luciferase and *Renilla* luciferase activities were assayed using a dual-luciferase assay kit (Promega) according to the manufacturer's instructions. Chemical luminescence was measured using a luminometer (Lumat LB9507, Berthold Technologies).

Protein analyses

For *in vitro* transcription/translation analysis, the TNT[®] SP6 high-yield protein expression system (Promega) was used according to the manufacturer's instructions with modifications. Sequences containing the SP6 promoter (5'-AGAATTGGACTA-TTTAGGTGACACTATAGAAA-3') were located upstream of the cDNA sequences and template cDNA fragments amplified by PCR were directly subjected to expression reactions for 2 h. *N. tabacum* and *N. sylvestris* leaves were ground in liquid nitrogen in a mortar, and homogenized in isolation buffer containing 50 mM Tris/HCl (pH 7.5), 100 mM NaCl, 0.05 % (w/v) Tween 20, 1 mM EDTA, 1 mM PMSF, 5 µg/ml leupeptin and 5 µg/ml aprotinin. After centrifugation at 20 000 g for 10 min to remove insoluble debris, crude extracts were used directly for immunoblot staining. After fractionation by SDS/PAGE, proteins were electronically transferred on to cellulose membranes (Immobilon-N, Millipore), and NtWIN4/NsWIN4 was detected using rabbit antibodies raised against a synthetic peptide (MIKERIRREKQKQSY) or a His-tagged recombinant protein (amino acid positions 27–247) and horseradish peroxidase conjugated anti-rabbit antibodies (Bio-Rad Laboratories).

RESULTS

Origin of *NtWIN4*

NtWIN4 was previously reported to exist as a single copy in the *N. tabacum* genome [26]. Since *N. tabacum* is an amphidiploid (2n = 48), derived from *N. sylvestris* (2n = 24) and *N. tomentosiformis* (2n = 24) (Figure 1A), its origin was first examined by PCR. Results showed effective amplification of *NtWIN4* sequences with genomic DNA from *N. sylvestris*, but not from *N. tomentosiformis* (Figure 1B). This indicates that *NtWIN4* originated from *N. sylvestris*, in which the corresponding gene was designated as *NsWIN4*. To examine whether both genes function similarly, their transcript accumulation profiles upon wounding were examined. In unstressed healthy leaves, neither gene was expressed (Figure 1C). When leaves were mechanically injured, transcripts began to accumulate, reaching the maximal level after 1 h, and gradually declining up to 3 h (Figure 1C). These results indicated that *NtWIN4* was transmitted from *N. sylvestris* to *N. tabacum* through amphidiploidization, and that the gene is equally involved in wound-stress responses in both species.

Structural properties

An amino acid sequence comparison between NtWIN4 and NsWIN4 revealed the two to be almost identical, except that a tryptophan residue at position 70 in the former was substituted with a glycine residue in the latter (Figure 2A). On DNA

Figure 1 Properties of NtWIN4/NsWIN4

(A) Schematic illustration of amphidiploidization. *N. tabacum* (2n = 48) was formed from two ancestors, *N. sylvestris* (2n = 24) and *N. tomentosiformis* (2n = 24). (B) Determination of the NtWIN4 ancestor. Genomic DNA was isolated from leaves of *N. sylvestris*, *N. tomentosiformis* and *N. tabacum*, and PCR analyses were performed with NtWIN4 primer sets as described in the Experimental section. Amplified DNA fragments and 500 ng of intact genomic DNA as the control were loaded. (C) Wound-responses of NtWIN4 and NsWIN4. Healthy leaves from *N. sylvestris* or *N. tabacum* were mechanically wounded and harvested at the indicated time points. Total RNA was isolated from these leaves, and cDNA was synthesized using the RT reaction with oligo-dT primers. NtWIN4/NsWIN4 and 18S rRNA genes were amplified by 35 and 20 cycles of PCR respectively.

sequencing, this mutation corresponded to GGG in *N. sylvestris* instead of TGG in *N. tabacum* (Figure 2A, middle panel). When this region was compared with other plant species, TGG was found to be predominant (Figure 2A, bottom panel), indicating that the GGG mutation in *N. sylvestris* might have occurred after amphidiploid formation. To check whether or not this point mutation affected protein properties, localization and transcriptional repression activities of NsWIN4 were compared with those of NtWIN4. Cellular localization was first examined with GFP-tagged fusion proteins (Figure 2B). Since the N-terminus of NtWIN4 was found to be critical for localization [26], two different fusion constructs were prepared: one a polypeptide consisting of 111 amino acids from the N-terminus (1–111), and the other a similar polypeptide, but with the first 26 amino acids deleted (27–111). Results showed that constructs of NsWIN4-(1–111) and NtWIN4-(1–111) equally localized in the nucleus and cytoplasm, whereas NsWIN4-(27–111) and NtWIN4-(27–111) proteins were localized to plastids (Figure 2B), consistent with our previous observations [26]. Nuclear transcription assays *in vivo* were then performed using the luciferase reporter system. GBD-NsWIN4 and GBD-NtWIN4 effector constructs equally suppressed the luciferase activity to a level one-third of that of the GBD control (Figure 2C). These results indicate that the observed point

Figure 2 Cellular localization and transcriptional regulation

(A) Upper panel: Schematic illustration of full length NsWIN4/NtWIN4 consisting of 247 amino acids. A bHLH motif is located at amino acid positions 79–135. The arrow indicates the mutation at position 70. Middle panel: amino acid sequences around the mutation at position 70. Bottom panel: cDNA sequence comparison among *N. tabacum*, *N. sylvestris* and other plants, *Capsicum annuum*, *Coffea canephora*, *Arabidopsis thaliana* and *Vitis vinifera*. Arrows indicate the triplet corresponding to the mutated amino acid at position 70. (B) Cellular localization. NsWIN4-(1–111), NtWIN4-(1–111), NsWIN4-(27–111) and NtWIN4-(27–111) were fused to the N-terminus of GFP, and fusion proteins were transiently expressed in onion epidermal cells after particle bombardment. Cells were examined by microscopy with the specified filter. (C) Dual-luciferase assays. GBD-NsWIN4 and GBD-NtWIN4 effector plasmids were bombarded with reporter and internal control plasmids into 5 day old BY2 cells. As the effector control, a yy64 vector alone was used.

Figure 3 Immunoblot staining

Total proteins were extracted from leaves of *N. tabacum* or *N. sylvestris*, and crude extracts were directly used for immunoblotting. After fractionation by SDS/PAGE, proteins were electronically transferred onto cellulose membranes, and NtWIN4/NsWIN4 was detected with rabbit antibodies raised against a synthetic NtWIN4 peptide and horseradish peroxidase conjugated anti-rabbit antibodies. The arrow and arrowhead indicate the position of the 26 kDa and 17 kDa proteins respectively.

mutation does not influence protein properties, which are shared in common between NsWIN4 and NtWIN4.

Natural selection

The molecular mass of full length NtWIN4 was calculated to be 26 kDa. However, native NtWIN4 was shown by immunoblot staining to be a 17 kDa protein and exclusively located in plastids [26]. Peptide dissection analysis indicated that a polypeptide with 81 amino acids starting at position 13 from the N-terminus served as a chloroplast transit signal, and is cleaved off after translocation to yield the mature plastidic 17 kDa protein *in planta*. To see whether this was also the case for NsWIN4, immunoblot analysis was performed using total proteins extracted from leaf tissues of both *N. sylvestris* and *N. tabacum* (Figure 3). Results showed that, in the sample from *N. tabacum*, a single major signal was detected at the position of 17 kDa, while in the sample from *N. sylvestris*, two signals at positions of 17 kDa and 26 kDa were distinct (Figure 3). Thus NsWIN4 simultaneously takes two molecular forms *in planta*, one of 17 kDa localizing in plastids, and the other of 26 kDa localizing in the cytoplasm and nucleus. This implies that NsWIN4 could be bi-functional in plastids and the nucleus, whereas NtWIN4 could be mono-functional only in plastids, and that such a conversion might be the result of natural selection associated with the amphidiploidization in the distant past.

Variation in mRNA length

When translated from the first AUG codon on mRNA, NtWIN4 was found to be localized in the cytoplasm and nucleus. When translated from the second AUG codon, the product was localized in plastids [26]. It is thus conceivable that diversification in size and localization between NtWIN4 and NsWIN4 could have arisen from diversification of transcription and/or translation systems. To verify this hypothesis, diversity of mRNA length was first examined by isolating as many mRNAs as possible from *N. sylvestris* and *N. tabacum* by the 5'-RACE method (Figure 4A). Resulting clones were classified into three groups for *N. sylvestris*, and four groups for *N. tabacum*, and tentatively named as S1, S2 and S3 for the former and T1, T2, T3 and T4 for the latter (Figure 4A). In S1, S2, T1, T2 and T3 sequences, a distinct 5'-UTR (5'-untranslated region) was found to exist upstream of the first AUG codon. In contrast, in S3 and T4 sequences, the 5'-terminal nucleotide was guanine located in

the first AUG codon set, indicating that protein synthesis from the first AUG codon is improbable (Figure 4A). To test whether or not these sequences encode proteins, *in vitro* transcription/translation coupled assays were performed. S2, T2 and T3 sequences were found to serve as templates for the 26 kDa protein, which is the full-length protein starting from the first AUG codon (Figure 4B). In contrast, S3/T4 sequences encoded a 24 kDa protein, which was possibly synthesized from the in-frame second AUG codon as a precursor of the 17 kDa protein (Figure 4B). These experiments revealed various sizes of mRNA, which could serve as translation templates. However, translation efficiency greatly differed among the mRNA populations. The efficiency was highest with S3/T4 and lowest with S1, T1 and T2. The efficiency of S2 was lower than that of S3/T4, but appeared to be compatible with that of T3. Densitometric estimation of the signal intensity showed the translation efficiency of each sequence relative to the S3/T4 level to be less than one-tenth for T1, S1 and T2, one-fourth for T3 and a half for S2 (Figure 4C). The results showed that a 24 kDa protein was predominantly produced in both species, and that a small amount of 26 kDa protein was also produced at a higher level in *N. sylvestris* than in *N. tabacum*. Thus differential translation efficiency between the two species may partly account for the presence of the 26 kDa protein in *N. sylvestris*. However, this idea alone does not explain the absence of the 26 kDa protein in *N. tabacum*.

mRNA populations

The results described above suggested that the major protein product of 17 kDa, which is the mature form of the 24 kDa protein *in planta*, is preferentially synthesized from T4 mRNA (tentatively referred to as short mRNA) in *N. tabacum*. In contrast, two proteins of 17 kDa and 26 kDa are synthesized from S3 mRNA (short mRNA) and S2 mRNA (long mRNA) respectively, in *N. sylvestris*. However, in *N. tabacum*, the long mRNA is present and is translated *in vitro*, although the efficiency was half that of *N. sylvestris*. Since no 26 kDa protein was detected in *N. tabacum*, there must be some mechanism which interferes with production of the 26 kDa protein. One possibility concerns mRNA populations and therefore we examined the frequency of each mRNA in *N. tabacum* and *N. sylvestris* by RT-PCR using cDNA libraries from two species. The forward primer sets 1 and 2 were designed to positions upstream and downstream of the first AUG codon respectively, thus facilitating amplification of long and short mRNAs (Figure 4D). The results revealed three distinct features. First, the short mRNA was efficiently amplified in both species, indicating its dominance. Secondly, the amplification efficiency of the long mRNA was much lower than that of the short mRNA in both species, suggesting the long mRNA to be the minor form. Thirdly, its amplification was nevertheless over 2-fold higher in *N. sylvestris* than in *N. tabacum* (Figure 4D). Based on these observations, we conclude that alternative transcription might be responsible for variable mRNA populations in both plant species, and that the production of the 26 kDa protein in *N. sylvestris* is partly attributable to a higher frequency of the long mRNA in comparison with *N. tabacum*. Indeed, the ratio of the long mRNA to the short mRNA was very low in *N. tabacum*, supporting almost no production of the 26 kDa protein *in planta*.

DISCUSSION

The present study documents compelling evidence of molecular mechanisms for differential expression and cellular localization of a bHLH protein between two tobacco species. In our previous study [26], we found that NtWIN4, a typical bHLH protein from

Figure 4 Properties of mRNAs

(A) Diversity in mRNA length. cDNAs corresponding to *NsWIN4* and *NtWIN4* were isolated from *N. sylvestris* and *N. tabacum* respectively, and classified according to length: S1–S3 are from *N. sylvestris* and T1–T4 are from *N. tabacum*. The first and in-frame second AUG codons are indicated by bold characters. (B) *In vitro* transcription/translation coupled assays. Each cDNA clone shown above was subjected to assays using the TNT[®] SP6 high-yield protein expression system according to the manufacturer's instructions as described in the text. Resulting crude extracts were fractionated by SDS/PAGE and products visualized by immunoblot staining with rabbit antibodies raised against a recombinant *NtWIN4* protein and horseradish peroxidase-conjugated anti-rabbit antibodies. Molecular masses of migrating proteins are indicated on the left. The position of the 26 kDa protein is indicated by an open arrowhead, and that of the 24 kDa protein by a closed arrowhead. The asterisk indicates non-specific signals. (C) Densitometric estimation of signal intensities shown in (B). Values are relative to that of S3/T4. (D) Population analysis of long and short mRNAs. Primers were designed to position before (primer 1) and after (primer 2) the first AUG codon (upper panel), and RT-PCR was performed (lower panel) using cDNA templates prepared from *N. sylvestris* (S) or from *N. tabacum* (T).

N. tabacum, was translocated into plastids due to conversion of its DNA-binding domain into a plastid-transit peptide. In the present study, we identified its counterpart *NsWIN4* from the ancestral *N. sylvestris*, and found both mRNA populations and translation efficiencies to have diverged between two species upon polyploidization.

Variable mRNAs

The size diversity of mRNA can be best explained by the idea of multiple transcription start sites, which was recently proposed based on studies of mammalian cells [33]. In mammals, promoters can be classified into two types, those enriched with TATA-box motifs and those enriched with CpG dinucleotides. The former possess one transcription start site, whereas the latter possess

multiple transcription start sites. CpG-rich promoters are much more abundant than TATA-box enriched promoters, and are considered to have rapidly evolved through epigenetic control [33]. In the case of *NtWIN4/NsWIN4*, however, no apparent CpG islands were found within their approximately 1 kb promoter regions (results not shown). This may suggest the presence of a so far unknown mechanism for multiple transcription starting in plants.

The question then arises as to why transcription start sites differ between *N. sylvestris* and *N. tabacum*. Three causes are conceivable: mutation in genomic DNA, genomic rearrangement and addition of new components. The first probability was examined with approximately 1 kb promoter regions of *NtWIN4* and *NsWIN4*, but no differences in nucleotide sequence were identified between the two (results not shown). This finding makes the first idea unlikely and also the second idea of

genomic rearrangement unlikely as well. The third possibility that new regulatory components have been added, however, is worthy of consideration. *N. sylvestris* and *N. tomentosiformis* had independently evolved until the amphidiploid *N. tabacum* was formed. This means that each had developed its own transcription system, and that *N. tabacum* inherited both. Although *N. tomentosiformis* does not possess an *NtWIN4*-like gene, it is highly probable that the transcription system derived from *N. tomentosiformis* partly and differently might control expression of *N. sylvestris*-derived *NtWIN4*, resulting in production of unique mRNA sizes in *N. tabacum*.

In addition to size diversity, changes in translation efficiency of each mRNA also appeared to be responsible for preferential production of certain protein species. Previously we proposed that the nucleotide length of the 5'-UTR might be critical for translation initiation from the first AUG codon [26]. This idea was based on the finding that *NtWIN4* mRNA possesses only a 17-base UTR, considered to be insufficient for recognition of the first AUG codon. The present finding is partly consistent with this speculation in regard to translation initiation being undetectable with the T1 mRNA (17-base UTR), but partly inconsistent with reference to translation efficiency, showing that shorter mRNAs, such as S2 (7-base UTR) and T3 (2-base UTR), than the T1 mRNA were relatively efficiently translated from the first AUG codon. These observations suggest that translation efficiency is not necessarily determined by nucleotide length of the 5'-UTR, but perhaps by higher conformation of the 5'-terminal regions. Structural analysis of these mRNA species is necessary to determine the mechanism.

Biological significance

Another question to be addressed concerns the biological functions of *NtWIN4* and particularly of *NsWIN4*. *NtWIN4* was initially identified as a wound-responsive gene, and shown to be up-regulated not only by wounding but also by pathogen attack, possibly mediated through jasmonate and hydrogen peroxide [26]. Transgenic approaches indicated induction of hypersensitive cell death by disrupting chloroplasts and it was concluded that *NtWIN4*, initially a nuclear-resident transcription regulator, evolved into a plastid-resident protein, resulting in acquisition of novel function(s) other than transcriptional regulation. In this context, *NtWIN4* can be regarded as a bi-functional protein on the evolutionary time scale, although it is currently mono-functional in *N. tabacum*. In contrast, *NsWIN4* is apparently bi-functional at present, serving as a transcription repressor in the nucleus and as a chlorosis-inducer in *N. sylvestris*. To verify the activity as a transcription repressor, we performed gel-shift experiments with bacterially expressed *NtWIN4* for binding to the G-box (CACGTG) sequence, a common *cis*-element for bHLH proteins. However, results were negative, and subsequent yeast two-hybrid assays indicated that the observed failure to bind was due to its inability to form homodimers, which are generally required for bHLH proteins to function (Y. Kodama, H. Ohya and H. Sano, unpublished work). It is conceivable that *NtWIN4*/*NsWIN4* might be a negative regulator of expression of so far unidentified wound-/pathogen-responsive genes by forming heterodimers with other bHLH protein(s).

Overall, our work points to one series of molecular events which have occurred during plant evolution. Three steps took place to change protein function in this case: first, diversification of the mRNA population caused by varied translation initiation; secondly, conversion of polypeptide function from DNA-binding into plastid-transit signalling; and thirdly, formation of amphidiploid species from two independent ancestral species.

NsWIN4 might have originally been bi-functional due to long and short mRNA populations respectively encoding large (26 kDa) and small (24 kDa) proteins. The N-terminus of the small protein was, by chance, suited to function as a transit peptide after translocation into plastids while that of the large protein was not due to masking effects by the first 26 amino acid residues [26]. Upon amphidiploidization, *NsWIN4* was transmitted to *N. tabacum*, as a consequence of which the frequency of long mRNA production and efficiency of its translation were reduced. This resulted in decline of the 26 kDa form, leaving *NtWIN4* as a mono-functional plastid protein. The absence of *NtWIN4* as a transcription regulator might not have been a major disadvantage for *N. tabacum*, possibly due to functional redundancy with similar proteins derived from the other ancestor, *N. tomentosiformis*. On the contrary, such a functional conversion might have been of advantage to cope better with environmental stresses, so that it was stably fixed as a new trait in *N. tabacum*. This model proposes the hypothesis that functional alteration of a protein is one of the factors that induces fitness to survive severe environmental conditions.

The authors thank Dr T. Hashimoto and Mr K. Inai (Nara Institute of Science and Technology, Nara, Japan) for providing plant materials and genomic DNA of *N. sylvestris* and *N. tomentosiformis*, and Dr M. Moore (Internal, Nagoya) for critical reading of the manuscript. This work was partly supported by a grant for Scientific Research for Plant Graduate Students (to Y. K.) from the Nara Institute of Science and Technology, supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES

- Wendel, J. F. (2000) Genomic evolution in polyploids. *Plant Mol. Biol.* **42**, 225–249
- Soltis, P. S. (2005) Ancient and recent polyploidy in angiosperms. *New Phytol.* **166**, 87–110
- Goodspeed, T. H. (1954) The genus *Nicotiana*. Chronica Botanica Company, Waltham, Massachusetts
- Gray, J. C., Kung, S. D., Wildman, S. G. and Sheen, S. J. (1974) Origin of *Nicotiana tabacum* L. detected by polypeptide composition of Fraction I protein. *Nature* **252**, 226–227
- Bland, M. M., Matzinger, D. F. and Levings, III, C. S. (1985) Comparison of the mitochondrial genome of *Nicotiana tabacum* with its progenitor species. *Theor. Appl. Genet.* **69**, 535–541
- Okamuro, J. K. and Goldberg, R. B. (1985) Tobacco single-copy DNA is highly homologous to sequences present in the genomes of its diploid progenitors. *Mol. Gen. Genet.* **198**, 290–298
- Matassi, G., Melis, R., Macaya, G. and Bernardi, G. (1991) Compositional bimodality of the nuclear genome of tobacco. *Nucleic Acids Res.* **19**, 5561–5567
- Skalická, K., Lim, K. Y., Matyásek, R., Koukalová, B., Leitch, A. R. and Kovarik, A. (2003) Rapid evolution of parental rDNA in a synthetic tobacco allotetraploid line. *Am. J. Bot.* **90**, 988–996
- Skalická, K., Lim, K. Y., Matyásek, R., Matzke, M., Leitch, A. R. and Kovarik, A. (2005) Preferential elimination of repeated DNA sequences from the paternal, *Nicotiana tomentosiformis* genome donor of a synthetic, allotetraploid tobacco. *New Phytol.* **166**, 291–303
- Kenton, A., Parokonny, A. S., Gleba, Y. Y. and Bennett, M. D. (1993) Characterization of the *Nicotiana tabacum* L. genome by molecular cytogenetics. *Mol. Gen. Genet.* **240**, 159–169
- Kitamura, S., Inoue, M., Ohmido, N. and Fukui, K. (2000) Quantitative chromosome maps and rDNA localization in the T subgenome of *Nicotiana tabacum* L. and its putative progenitors. *Theor. Appl. Genet.* **101**, 1180–1188
- Lim, K. Y., Matyásek, R., Lichtenstein, C. P. and Leitch, A. R. (2000) Molecular cytogenetic analyses and phylogenetic studies in the *Nicotiana* section *Tomentosae*. *Chromosoma* **109**, 245–258
- Buchanan, B. B., Gruissem, W. and Jones, R. L. (2000) *Biochemistry and Molecular Biology of Plants*, pp. 37–45, American Society of Plant Physiologists, Rockville, Maryland
- Race, H. L., Herrmann, R. G. and Martin, W. (1999) Why have organelles retained genomes? *Trends Genet.* **15**, 364–370
- Leister, D. (2003) Chloroplast research in the genomic age. *Trends Genet.* **19**, 47–56

- 16 Sato, N. (2001) Was the evolution of plastid genetic machinery discontinuous? *Trends Plant Sci.* **6**, 151–155
- 17 Nakano, T., Murakami, S., Shoji, T., Yoshida, S., Yamada, Y. and Sato, F. (1997) A novel protein with DNA-binding activity from tobacco chloroplast nucleoids. *Plant Cell* **9**, 1673–1682
- 18 Sato, N., Albrieux, C., Joyard, J., Douce, R. and Kuroiwa, T. (1993) Detection and characterization of a plastid envelope DNA-binding protein which may anchor plastid nucleoids. *EMBO J.* **12**, 555–561
- 19 Kodama, Y. (2007) Plastidic proteins containing motifs of nuclear transcription factors. *Plant Biotechnol.* **24**, 165–170
- 20 Murre, C., McCaw, P. S. and Baltimore, D. (1989) A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD and myc proteins. *Cell* **56**, 777–783
- 21 Ferre-D'Amare, A. R., Pognonec, P., Roeder, R. G. and Burley, S. K. (1994) Structure and function of the b/HLH/Z domain of USF. *EMBO J.* **13**, 180–189
- 22 Garrell, J. and Modolell, J. (1990) The *Drosophila* extramacrochaetae locus, an antagonist of proneural genes that, like these genes, encodes a helix–loop–helix protein. *Cell* **61**, 39–48
- 23 Quail, P. H. (2000) Phytochrome interacting factors. *Semin. Cell Dev. Biol.* **11**, 457–466
- 24 Heim, M. A., Jacoby, M., Werber, M., Martin, C., Weisshaar, B. and Bailey, P. C. (2003) The basic helix–loop–helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. *Mol. Biol. Evol.* **20**, 735–747
- 25 Toledo-Ortiz, G., Huq, E. and Quail, P. H. (2003) The *Arabidopsis* basic/helix–loop–helix transcription factor family. *Plant Cell* **15**, 1749–1770
- 26 Kodama, Y. and Sano, H. (2006) Evolution of a basic helix–loop–helix protein from the transcriptional repressor to a plastid-resident regulatory factor – involvement in hypersensitive cell death in tobacco plants. *J. Biol. Chem.* **281**, 35369–35380
- 27 Kodama, Y. and Sano, H. (2007) A comparative analysis of basic helix–loop–helix proteins, AtPTF1 and NtWIN4, with reference to plastid localization. *Plant Biotechnol.*, in the press
- 28 Nagata, T., Okada, K., Takabe, I. and Matsui, C. (1981) Delivery of tobacco mosaic virus RNA into plant protoplasts mediated by reverse-phase evaporation vesicles (liposomes). *Mol. Gen. Genet.* **184**, 161–165
- 29 Murray, M. G. and Thompson, W. F. (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4325
- 30 Gonzalez, R. G., Haxo, R. S. and Schleich, T. (1980) Mechanism of action of polymeric aurintricarboxylic acid, a potent inhibitor of protein–nucleic acid interactions. *Biochemistry* **19**, 4299–4303
- 31 Chiu, W., Niwa, Y., Zeng, W., Hirano, T., Kobayashi, H. and Sheen, J. (1996) Engineered GFP as a vital reporter in plants. *Curr. Biol.* **6**, 325–330
- 32 Yamamoto, Y. Y. and Deng, X. Y. (1998) A new vector set for GAL-4-dependent transactivation assay in plants. *Plant Biotechnol.* **15**, 217–220
- 33 Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P. G., Frith, M. C. et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635

Received 2 January 2007/5 February 2007; accepted 8 February 2007

Published as BJ Immediate Publication 8 February 2007, doi:10.1042/BJ20070011