



Diversification of 13S globulins, allergenic seed storage proteins, of common buckwheat



Madoka Sano, Mariko Nakagawa, Akifumi Oishi, Yasuo Yasui, Tomoyuki Katsube-Tanaka *

Graduate School of Agriculture, Kyoto University, Kitashirakawa, Kyoto 606-8502, Japan

ARTICLE INFO

Article history:

Received 10 October 2013

Received in revised form 11 December 2013

Accepted 15 January 2014

Available online 24 January 2014

Keywords:

Allergen

Common buckwheat

Secondary structure

Tandem repeat

13S globulin

ABSTRACT

The α polypeptide of the 13S globulin subunit of common buckwheat is the counterpart of the major allergenic β polypeptide. Trypsin digestibility varies between variants of the α polypeptide with and without a tandem repeat insert. To evaluate the intra-species diversity of 13S globulin, the comprehensive screening of a genomic DNA library was performed, resulting in the isolation of 14 and 3 genes for Met-poor and Met-rich subunits, respectively. Although most tandem repeat units were 45 bp in length, the two-repeat gene *Glb2B* and all one-repeat genes contained an additional 3 bp. Secondary structure predictions and polyacrylamide gel electrophoresis demonstrated that the sense strand of *Glb2B*-CCG, the additional 3 bp-deletion clone of *Glb2B*, formed a more rigid secondary structure than that of the wild-type. Thus, the large intra-species variation of 13S globulin revealed in this study and its diversification might be attributable to the unique nature of the tandem repeat sequences.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Common buckwheat (*Fagopyrum esculentum* Moench) is a nutritionally and physiologically superior food crop with good palatability traits, because it contains a well-balanced amino acid composition, favourable hypocholesterolemia and hypotensive activity (Chen, Jiao, & Ma, 2008). However, buckwheat seeds contain an allergen which induces an IgE-mediated type I reaction similar to soybeans or peanuts (Wieskabder, 1996). So far, proteins of 9, 16, 19, and 24 kDa have been shown to be the most prevalent allergens, while less prevalent allergens of 30, 43, and 67 kDa have also been identified (Park et al., 2000). Among them the 24 kDa proteins, the β polypeptide of the major seed storage protein '13S globulin', is one of the dominant allergens (Nair & Adachi, 1999), and epitope sequences of the polypeptide have been determined (Kondo, Urisu, Tokuda, Ishida, & Yasuda, 2001; Sordet, Culerrier, Granier, Didier, & Rougé, 2009; Yoshioka, Ohmoto, Urisu, Mine, & Adachi, 2004). 13S globulin, which accounts for 43% of total seed proteins (Radović, Maksimović, & Varkonji-Gasić, 1996), is a neutral, saline-soluble, legumin-like protein belonging to the 11S globulin superfamily, which also includes rice glutelin and soybean glycinin (Derbyshire, Wright, & Boulter, 1976; Dickinson, Hussein,

& Nielsen, 1989; Katsube et al., 1999). Similar to rice glutelin and soybean glycinin, 13S globulin composes a multigene family and consists of various subunits comprised of disulphide-bonded α and β polypeptides (Fig. 1). 13S globulin has a low methionine content subunit (Met-poor) and a high methionine content subunit (Met-rich). Unlike rice glutelin and soybean glycinin, some Met-poor α polypeptide subunits of 13S globulin have one to six tandemly repeated sequences of 15 amino acid residues, and show a large variation in their molecular weight.

In a previous report, we demonstrated that the tandem repeat sequences are hydrophilic with many arginine residues, and that α polypeptides with a tandem repeat are easier to digest by trypsin than those without the tandem repeat (Khan, Takahashi, & Katsube-Tanaka, 2012). In general, proteins that are difficult to digest are likely to become allergens (Sen et al., 2002); thus, it has been suggested that the allergenicity of buckwheat might be affected by the presence or absence of a tandem repeat (Khan et al., 2012). However, it is not known how variable the tandem repeats are or how that variation developed.

Therefore in this report, we first identified the intra-species diversity by comprehensive cloning of 13S globulin genes from a genomic DNA library, which is fundamental and a prerequisite information for the 13S globulin analyses. We then demonstrated the possibility that some tandem repeats are involved in their own diversification through the formation of secondary structures, which may induce DNA replication errors. We especially focused on the effect of additional 3 nucleotides within some tandem

Abbreviations: Met, methionine; MW, molecular weight; PAGE, polyacrylamide gel electrophoresis.

* Corresponding author. Tel.: +81 75 7536043; fax: +81 75 7536065.

E-mail address: tanakato@kais.kyoto-u.ac.jp (T. Katsube-Tanaka).

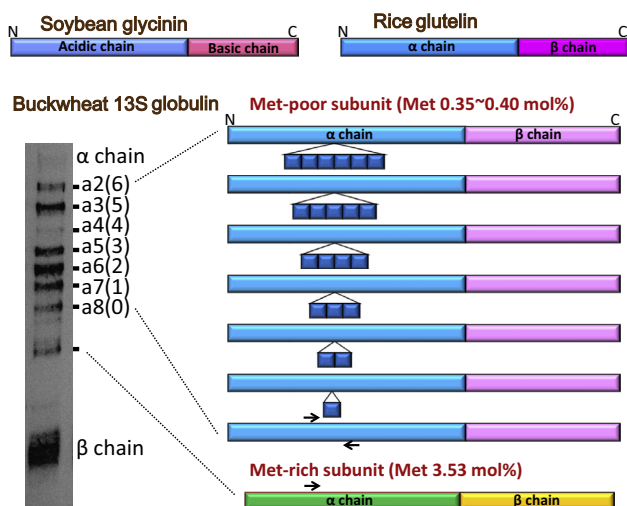


Fig. 1. Schematic representation of buckwheat 13S globulin, soybean glycinin, and rice glutelin. Subunit structures of 13S globulin (pro-form) were compared with that of soybean glycinin and rice glutelin, all of which are members of the 11S globulin superfamily. 13S globulin has a low methionine content (Met-poor) subunit and a high methionine content (Met-rich) subunit. There are seven types of Met-poor subunits, which differ in the number of tandemly repeated inserts (0–6 repeats; indicated by square boxes) residing within their α polypeptides (α chains). Met-poor α polypeptides (a2–a8) are matched to the seven subunit types. The number in the parentheses after labels a2–a8 indicates the number of tandem repeats. The positions of PCR primers designed at conserved regions before and after tandem repeat sequences were indicated with horizontal arrows.

repeats on their DNA secondary structure. This report could extend our knowledge concerning the allergenic 13S globulin of common buckwheat and provide new insight into a hypoallergenic buckwheat development.

2. Material and methods

2.1. Identification of 13S globulin genes from a genomic DNA library

A genomic DNA library consisting of 142,005 BAC clones with an average insert size of ~ 76 kbp (Yasui et al., 2008) was used to identify 13S globulin genes. BAC clones containing genes for Met-poor and Met-rich subunits were isolated by PCR with primers left-1 (5'-AGGATGYCCGGAGACRTWCCA-3') and right-1 (5'-CTA ACGTTCYCATCGAGCTG-3') and with primers left-1 and right-2 (5'-CGGTGTTGAGGTTGTGGAC-3'), respectively. The primers were designed at conserved regions before and after tandem repeat sequences (Fig. 1), so as to distinguish the length of tandem repeats (Khan et al., 2012). PCR reactions were performed using a thermal cycler (Standard PCR, Takara, Japan) under the following conditions: an initial denaturing step at 94 °C for 30 s; 45 cycles of denaturing at 94 °C for 30 s, annealing at 60 °C for 60 s, and extension at 72 °C for 60 s, with a final extension step at 72 °C for 7 min. Three rounds of PCR were performed to screen BAC clones stored in 384-well plates. The 1st screening consisted of 47 super-pools, each of which contained cloned DNA from eight 384-well plates. The 2nd screening consisted of 8 pools, each of which contained cloned DNA from a positive super-pool, and the 3rd screening consisted of 24 and 16 pools, each of which contained a DNA mixture of 16 vertical and 24 horizontal clones, respectively, from a positive 384-well plate. BAC clones containing multiple 13S globulin genes (ID No. 7J5, 70M10, 75L24, 83J12, 15P2, 148G9 and 342C11) were partially digested with Sau3AI or HindIII and ligated into pBlue-script SK- to produce a shotgun clone library. Colony PCR was

employed to identify subclones from the shotgun library containing the 13S globulin genes. DNA of isolated BAC clones and subclones was purified with a Genopure Plasmid Midi Kit (Roche, Switzerland), an AxyPrep Plasmid Miniprep Kit (Corning, USA), or a FastGene™ Plasmid Mini Kit (Nippon Genetics, Japan) and sequenced. Nucleotide sequences were assembled, aligned, and edited using the computer programs Phred/Phrap (Ewing & Green, 1998; Ewing, Hillier, Wendl, & Green, 1998), Biolign, and BioEdit (Hall, 1999). Phylogenetic and molecular evolutionary analyses were conducted using the neighbour-joining method with a JTT matrix for evolutionary distance in the computer program MEGA version 5 (Tamura et al., 2011). The secondary structures of sense and antisense strands of tandem repeat inserts were analysed with the computer program Centroid Fold (Hamada, Kiryu, Sato, Mituyama, & Asai, 2009; Sato, Hamada, Asai, & Mituyama, 2009).

2.2. Construction of deletion clones for tandem repeats containing the additional 3 bp

Tandem repeats for most of the 13S globulin Met-poor subunit genes had a 45 bp repeat unit, while those of *Glb1A*, *Glb1B*, *Glb1C*, *Glb1D*, and *Glb2B* had a 48 bp repeat unit. To construct deletion clones for tandem repeats containing the additional 3 bp (*Glb1A*-GAG and *Glb2B*-CCG), the tandem repeat regions of *Glb1A* and *Glb2B* were amplified with primers M-poor3 left-1 (5'-AGGATGCCCGGA-GACGTTCCA-3') and M-poor3 right-1 (5'-CTAACGTTCTCATC-GAGCTG-3') and subcloned into pTAC-2 vector (BioDynamics Laboratory Inc., Tokyo) to produce pTAC:rep1A and pTAC:rep2B, respectively. After confirming the direction of the insert, pTAC:rep1A fragments, amplified with primers Rep1A-1-GAG3' (5'-TCCCCGCGGCTGGATTCTTCGCTCTCACTCTG-3') and RV-M (5'-GAGCGGATAACAATTTACACAGG-3'), were digested with BtgI and HindIII and inserted into the corresponding sites of pTAC:rep1A to produce pTAC:rep1A-GAG. Similarly, BtgI-digested pTAC:rep2B fragments, amplified with primers Rep2B-1-CCG3'b (5'-TCTCCACGGCTGGATTCTTCGCTCTCACTCTG-3') and T7promoter-1 (5'-TAATACGACTCACTATAGGG-3'), were digested with BtgI and SacI, and inserted into the corresponding sites of pTAC:rep2B to produce pTAC:rep2B-CCG. Sequencing was used to confirm that both deletion clones contained a 45 bp repeat unit lacking the additional 3 bp (GAG or CCG).

2.3. Affect of tandem repeats on secondary structures

Because the tandem repeat sequences were predicted to form secondary structures, such as hairpins, sense and antisense strands containing tandem repeats, amplified with primers M-poor3 left-1 and M-poor3 right-1, were analysed independently in terms of the secondary structure. Aliquots of the primers M-poor3 left-1 and M-poor3 right-1 were phosphorylated by T4 polynucleotide kinase (Fermentas, USA), and PCR performed with a combination of phosphorylated and non-phosphorylated primers and templates of the wild type clones *Glb1A*, *Glb2B*, and *Glb2A* and the 3 bp deletion clones *Glb1A*-GAG and *Glb2B*-CCG. The amplified products were digested with lambda exonuclease (New England Biolabs, USA) to produce single stranded DNA through decomposition of the phosphorylated strand (Takagi, Kimura, & Katsuki, 1993). The single stranded DNA was dissolved in TE buffer (Tris-EDTA) after ethanol precipitation. The sample (~ 100 ng) was electrophoresed through an 8% polyacrylamide gel containing 1x TBE at a constant voltage of 200 V for 40 min at 25 and 40 °C (with 0.5x TBE running buffer), following a denaturing step at 95 °C for 5 min and rapid cooling on ice, according to the method previously reported (Hofmann, Heinrich, Radziwill, & Moelling, 2009). Gels were stained with ethidium bromide and visualised with a Molecular Imager (BioRad, USA).

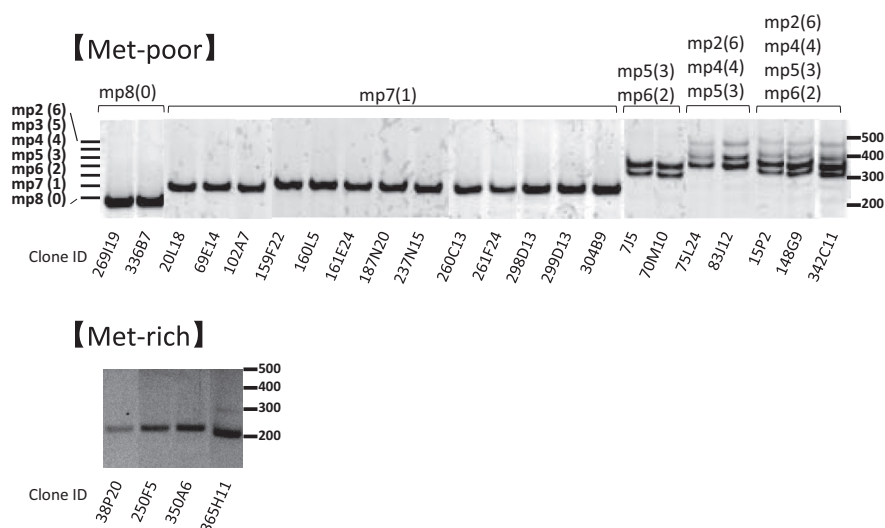


Fig. 2. BAC clones containing the 13S globulin gene identified in a genomic DNA library. Twenty-two BAC clones for Met-poor subunits and four BAC clones for Met-rich subunits were identified. The pattern of PCR bands amplified with primers left-1/right-1 and left-1/right-2 for Met-poor and Met-rich subunit, respectively, were compared. The amplified bands for Met-poor subunits were named mp2–mp8. The number in the parentheses after labels mp2–mp8 indicates the number of tandem repeats. The BAC clone ID is indicated below the figure. The position of DNA size markers (200–500 bp) is given on the right.

3. Results

3.1. 13S globulin genes identified from a genomic DNA library

Twenty-two BAC clones containing Met-poor subunit genes and four BAC clones containing Met-rich genes were identified (Fig. 2). Thirteen of the 22 BAC clones produced PCR bands containing a single tandem repeat (mp7), suggesting that the 13S globulin genes containing single repeats (hereafter termed one-repeat genes; other genes containing zero or multiple repeats are similarly named) were scattered throughout the genome. This result is in agreement our previous study in which almost all heterogeneous seeds produced the PCR band mp7 (Khan et al., 2012). Meanwhile, multiple PCR bands, mp2, mp4, mp5, and mp6, containing

six, four, three, and two repeats, respectively, were amplified from 7 single BAC clones in different combinations (i.e., mp5 and mp6; mp2, mp4, and mp5; mp2, mp4, mp5, and mp6), suggesting that such multiple-repeat genes are adjacent to each other in the genome. PCR band mp3, derived from a five-repeat gene, was not observed in this genomic DNA library. Because ten progeny seeds from the pedigree used for the library construction did not produce the PCR band mp3 (data not shown), we speculate that the pedigree did not contain the five-repeat gene from its inception.

The sequencing of 13S globulin genes using all the identified BAC clones as a template revealed that there were four types of one-repeat genes, *Glb1A*, *Glb1B*, *Glb1C*, and *Glb1D*, two types of zero-repeat genes, *GlbNA* and *GlbNB*, and three types of Met-rich subunit genes, *GlbRA*, *GlbRB*, and *GlbRC* (Table 1). Due to the high

Table 1
Characteristics of 13S globulin genes identified in a genomic DNA library.

Gene	# of repeat	Ins/Del [*] in Tandem Repeat	1st intron length (bp)	Microsatellite in 1st intron	Note ^{**}
<i>Met poor</i>					
<i>GlbNA</i>	Null	–	99	–	Ins of His at 40th residue
<i>GlbNB</i>	Null	–	108	–	6 bp Ins in 3rd exon
<i>Glb1A</i>	1	Ins (GAG)	92	(TAA) ₄ , TAT	22nd residue of β, Tyr
<i>Glb1B</i>	1	Ins (GAG)	95	(TAA) ₆	
<i>Glb1C</i>	1	Ins (GAG)	92	(TAA) ₄ , TAT	Identical to <i>Glb1A</i> in TR
<i>Glb1D</i>	1	Ins (GAG)	155	(TAA) ₂₆	Identical to <i>Glb1A</i> in TR
<i>Glb2A</i>	2	–	113	–	
<i>Glb2B</i>	2	Ins (CCG)	101	–	12 bp Ins in 3rd exon
<i>Glb3A</i>	3	–	104	–	27 bp Ins in 3rd intron
<i>Glb3B</i>	3	–	110	–	3 bp Ins in 3rd exon
<i>Glb3C</i>	3	–	160	–	9 bp Del in 4th exon
<i>Glb3D</i>	3	Del (12 bp)	103	–	36 bp Ins in 3rd intron
<i>Glb4A</i>	4	–	100	–	
<i>GlbXA</i>	<1	–	113	–	Pseudo(truncated at mid-TR)
<i>Met rich</i>					
<i>GlbRA</i>	Null	–	400	(TA) ₁₉	similar to <i>GlbRB</i>
<i>GlbRB</i>	Null	–	408	(TA) ₂₃	
<i>GlbRC</i>	Null	–	404	(TA) ₂₁	

^{*} Ins, insertion; Del, deletion.

^{**} TR, Tandem Repeat; β, β chain of 13S globulin.

degree of similarity among multiple-repeat genes, subclones (shotgun clones) instead of BAC clones were also used for sequencing. Sequencing of the shotgun clones revealed at least one type of a four-repeat gene, *Glb4A*, four types of three-repeat genes, *Glb3A*, *Glb3B*, *Glb3C*, and *Glb3D*, and two types of two-repeat genes, *Glb2A* and *Glb2B*, as well as a possible pseudo gene, *GlbXA*, which was truncated at the end of the first tandem repeat. Six-repeat genes corresponding to the PCR band mp2 could not be isolated.

So far, full cDNA sequences of one-, three-, and five-repeat genes (Acc. No. D87982, AF152003, and D87980, respectively), a genomic DNA sequence corresponding to the β chain of a one-repeat gene (Acc. No. AF216800, AF216801), and a full genomic DNA sequence of a Met-rich subunit gene (Acc. No. AY359286) have been identified. In the current study with nearly comprehensive cloning, it was demonstrated that one-repeat genes and Met-rich subunit genes have different microsatellite lengths, composed of TAA and TA, respectively, in the first intron (Table 1) and that multiple- and zero-repeat genes have short insertions or deletions in the 3rd exon, the 4th exon, and the 3rd intron, etc. (Table 1).

The neighbour-joining phylogenetic trees, with a JTT matrix for evolutionary distance, used for the deduced amino acid sequences of Met-poor subunits, showed three differentiated groups; the first group containing zero-repeat genes, the second group containing one-repeat genes, and the third group containing the other multiple-repeat genes (data not shown). The multiple-repeat genes existed in a common BAC clone as previously described, and thus it appears that the multiple-repeat subunits may have been duplicated recently in the species evolutionary history. Because the alignment of Met-rich subunits was largely different from that of Met-poor subunits, the phylogenetic analysis was performed without Met-rich subunits.

3.2. Sequences and secondary structure prediction of tandem repeat

The length of most tandem repeat units was 45 bp, while that of the first repeat of *Glb2B* (2B-1), *Glb1A*, *Glb1B*, *Glb1C*, and *Glb1D* (1A-1, 1B-1, 1C-1, and 1D-1) was 48 bp (Table 1, Fig. 3). More specifically, 2B-1 contained an additional 3 bp consisting of CCG, while the additional base pairs contained in 1A-1, 1B-1, 1C-1, and 1D-1 were GAG. In contrast, *Glb3D* had a 12 bp deletion at the junction between the first and the second tandem repeat and *GlbXA* was truncated in the middle of the first tandem repeat.

The software Centroid Fold predicted secondary structures for sense and antisense strands containing tandem repeats (Fig. 4). The sense strand of 2A had many base pairs with high base-pairing probability, resulting in an apparently rigid hairpin (stem and loop) structure (free energy: -28.9 kcal/mol), while the sense strands of 2B, 1A, 1B had a lower base-pairing probability (free energy: -11.2 to 0 kcal/mol). However, the sense strand of the tandem repeats from the deletion clones *Glb2B*-CCG or 2B-CCG, had more base pairs with high base-pairing probability and significantly decreased free energy (-19.56 kcal/mol). There was no significant change in structure for the sense strand of 1A-GAG when compared with that of 1A. Similarly it was predicted that all antisense strands of 2A, 2B, 2B-CCG, 1A, 1B, and 1A-GAG had lower base-pairing probability, higher free energy, or less variation in the same number of repeat sequences when compared with the corresponding sense strand. Accordingly, the predicted secondary structure of the sense strand was significantly altered depending on slight changes in tandem repeat sequences. In particular, the additional three nucleotides (CCG) inserted in the first repeat of 2B seemed to have a large impact on the secondary structure of the 2B sense strand, resulting in possible suppression of hairpin formation.

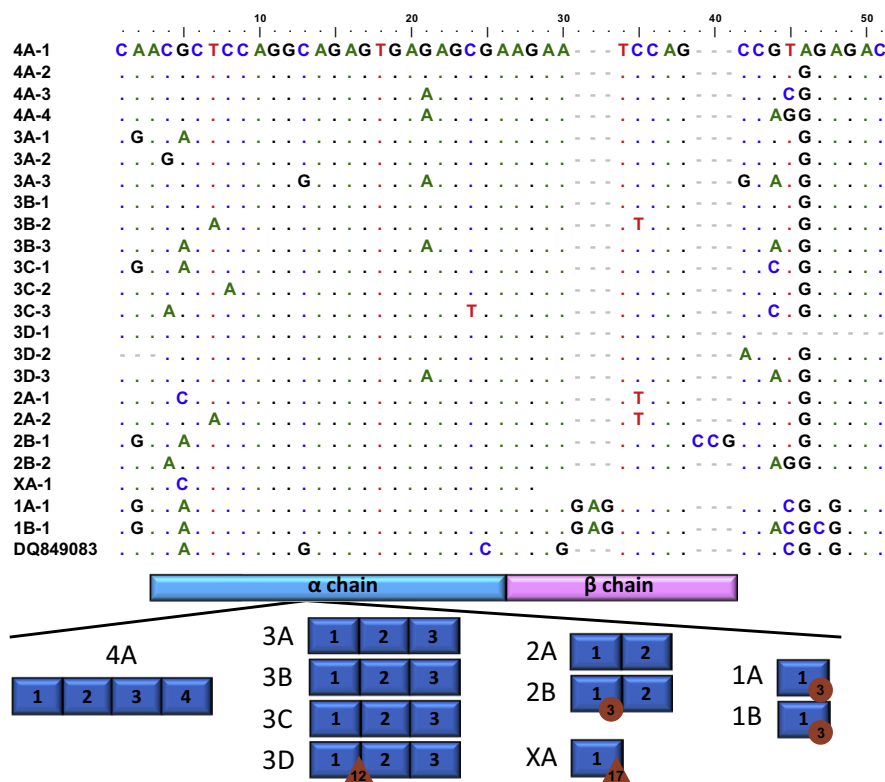


Fig. 3. Nucleotide sequence comparison among tandem repeat units of Met-poor subunits. Tandem repeat units of 4A, 3A, 3B, 3C, 3D, 2A, 2B, XA, 1A, and 1B were numbered from the N-terminus and compared with that of the tartary buckwheat allergenic protein (DQ849083). The tandem repeat units of 1C and 1D were omitted because they are the same as 1A. In the sequence alignment, a dot indicates identity with the nucleotide of the first repeat of 4A (4A-1). In the illustration, the dark blue box indicates the tandem repeat unit; the 3 bp insertion into 1A-1, 1B-1, and 2B-1 is shown as a brown circle. The 12 bp and 17 bp deletions from 3D and XA, respectively, are shown as brown triangles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

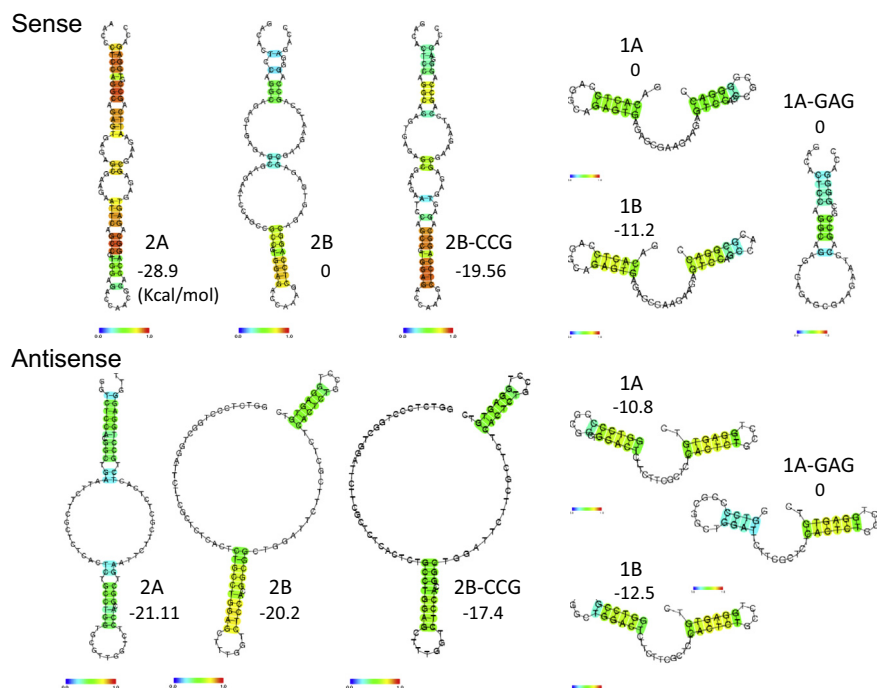


Fig. 4. Secondary structure of tandem repeats predicted by Centroid Fold. The sense and antisense strands of the tandem repeat regions of 2A, 2B, 1A, and 1B, and the 3 bp deletion clones 2B-CCG and 1A-GAG were analysed using the computer program Centroid Fold. Heat colour from blue to red indicates increasing base-pairing probability from 0 to 1. The numerical values shown at each clone label are predicted free energy (kcal/mol). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

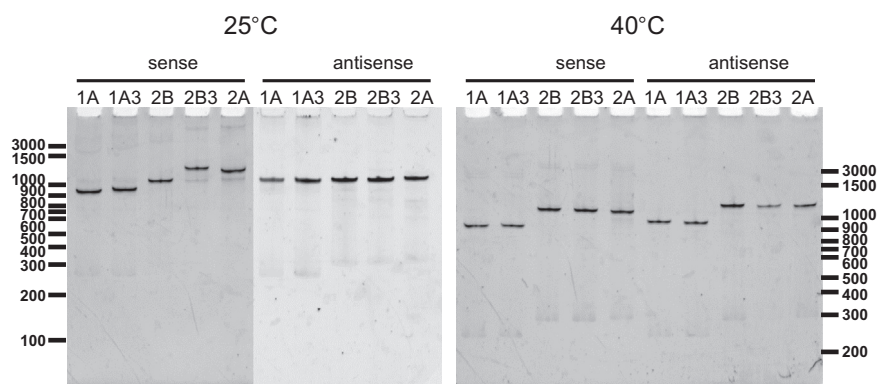


Fig. 5. Analysis of the secondary structure of tandem repeat regions by PAGE. The sense and antisense single stranded DNA containing tandem repeats was amplified with primers M-poor3 left-1 and M-poor3 right-1 and analysed by PAGE at 25 °C and 40 °C. 1A, 1A-GAG (indicated by 1A3), 2B, 2B-CCG (indicated by 2B3), and 2A were compared. Differences in secondary structures, such as hairpins, and in sequence length affect the band migration. Note that thick bands are single stranded DNA and fast-migrating faint bands are double stranded DNA formed from the undigested complementary strand. The position of DNA size markers (100–3000 bp) is given on each side of the figure.

3.3. PAGE analysis for the secondary structure of tandem repeats

To elucidate the actual differences in possible secondary structures of tandem repeats, sense and antisense strands prepared with phosphorylated, non-phosphorylated primers and lambda exonuclease were compared using PAGE (Fig. 5). On the gel, thick single strand bands migrated differently based on the varied secondary structure along with faint bands of undigested complementary strands and their matching double stranded DNA. At 40 °C, both sense and antisense strands of one-repeat genes migrated faster than those of two-repeat genes, despite there being no apparent difference in the same number of repeat sequences; this result is likely due to the destructuring of the secondary structure by heat (Fig. 5). At 25 °C, sense strands migrated in the following order from fastest to slowest: 1A > 1A-GAG >> 2B >> 2A >> 2B-CCG. All

the antisense strands migrated an equal distance and were similar to the sense strand of 2B. As a consequence, the above results at 25 °C largely corresponded to the predictions made with Centroid Fold, that is, the sense strands of 2A and 2B-CCG most likely forming a rigid secondary structure, which slowed migration during PAGE.

4. Discussion

4.1. Uniqueness and diversity of tandem repeats in buckwheat 13S globulin

The common buckwheat plant is allogamous, and its seeds may have a different genetic constitution even if they are from the same variety. Furthermore, some 13S globulin Met-poor subunit genes

contain tandem repeats of differing lengths, resulting in complex and uncharacterised gene diversity. Therefore, the intra- and inter-specific diversity within landraces, relatives and wild species, as well as modern improved varieties should be elucidated, and the underlying mechanisms increasing diversity should be uncovered to facilitate the development of hypoallergenic buckwheat plants. Here, we examined intra-specific diversity by screening a genomic DNA library and attempted to identify possible mechanisms for generating diversity by comparing the secondary structures of single stranded DNA of some tandem repeats.

The number of tandem repeats may increase or decrease through replication slippage, mismatch repair impairment, homologous recombination (Aguilera & Gómez-González, 2008), or unequal crossing-over (Brookes, 2013). Two common types of tandem repeats are microsatellites (<10 nucleotides repeat unit) and minisatellites (>10 nucleotides repeat unit) (Gemayel, Vences, Legendre, & Verstrepen, 2010). A considerable number of human genetic disorders with unusual hereditary patterns (genetic anticipation) are caused by triplet repeat amplification (Mirkin, 2007). Although microsatellites are found in both coding and non-coding regions, triplet repeats residing in a translational region can cause changes in protein structure by forming polyglutamine or polyalanine; these changes in turn may result in Huntington's disease, oculopharyngeal muscular dystrophy and so on (Mirkin, 2007). Molecular mechanisms involving the repeat instability of microsatellites, and the related balancing act hypothesis, have been described in budding yeast cells (Kim & Mirkin, 2013).

Observations of minisatellites residing in translational regions are limited. Some instances have been found in cell surface proteins (in benign brewer's yeast) and extracellular protein mucins (in mammals including humans and bovines) (Gemayel et al., 2010; Sando et al., 2009). Therefore, the minisatellite 45 bp repeat unit inserted into the α chain of 13S globulin, which is a storage protein with temporary, heavy, orderly accumulation in subcellular organelles (i.e., vacuolar protein bodies), is a unique research material, which may be used to determine the protein's structure–function relationship, and to elucidate the mechanism underlying the increased, but structurally sound, diversity.

Three-dimensional structures of 11S proglobulins, i.e., soybean pro-A1aB1b and pro-A3B4 (Adachi, Takenaka, Gidamis, Mikami, & Utsumi, 2001; Adachi et al., 2003), pumpkin pro-11S (Protein Data Bank entry 2E9Q), pea prolegumin and rapeseed procruciferin (Tandang-Silvas et al., 2010), amaranthus 11S proglobulin (Tandang-Silvas et al., 2012), and the major peanut allergen Ara h 3 (Jin, Guo, Chen, Howard, & Zhang, 2009), have been resolved as a trimer structure by X-ray crystallography. Thus, it is likely that buckwheat 13S proglobulin also has a trimeric structure. When comparing the structure-based sequence alignment of the 11S globulins with the sequence of buckwheat 13S proglobulin, the tandem repeats seem to reside within the variable region II identified in a previous study (Khan et al., 2008). Because the length of the variable region II demonstrated in the above-mentioned 11S proglobulins (except in buckwheat 13S proglobulin) is at most 41 amino acid residues, observed in rapeseed procruciferin (Tandang-Silvas et al., 2010), determining the number of repeat inserts that can be maintained in the 13S proglobulin is of interest. It is notable that one repeat insertion corresponds to an increase in MW of ~ 1.8 kDa, compared with a MW of 28.5 kDa for the zero-repeat α chain.

4.2. Evaluation of the diversity and secondary structure of tandem repeats

Sequence variation of the tandem repeats was elucidated by comprehensive cloning of 13S globulin genes from a genomic DNA library. We focused on the 3-nucleotide additional sequences inserted into the first tandem repeat of the one-repeat genes *Glb1A*,

Glb1B, *Glb1C*, and *Glb1D* (GAG) and the two-repeat gene *Glb2B* (CCG). Because the other tandem repeat containing genes essentially have a repeat unit of 45 bp, we hypothesised that the additional 3 nucleotides act as a type of stabilising interruption (Mirkin, 2007) to suppress changes in tandem repeats. Although the deletion clone of 1A (1A-GAG) was not significantly different from the wild type in terms of predicted secondary structure and PAGE using the sense and antisense strands, the deletion clone of 2B (2B-CCG) had a predicted structure with much lower free energy, slower migration in PAGE at 25 °C of the sense strand (the antisense strand showed no clear and consistent difference). The above results suggest that the 3 nucleotides CCG inserted into 2B may contribute to the stability of the gene. In other words, the loss or the gain of those 3 nucleotides from *Glb2B*-like or into *Glb2A*-like genes, respectively, may change the evolutionary rate of the tandem repeats by altering the stability of the secondary structure.

The newly identified gene *Glb3D* lacked 12 bp around the boundary between the first and second tandem repeat. The newly found pseudogene *GlbXA*, the tandem repeat of which is similar to that of *Glb2A* (Fig. 3), was truncated in the middle of a tandem repeat. These modifications may be accounted for by replication error or double strand breaks (Mirkin, 2007) caused by the formation of the secondary structure. In consequence, the above-mentioned circumstantial evidence coupled with our results about 3 bp deletion clones suggest the secondary structure of tandem repeats would affect their stability and might cause genome instability.

4.3. Zero-repeat genes and attempts to lower allergenicity

There would be two strategies to reduce the accumulation of zero-repeat subunit; one is loss of zero-repeat gene function and the other is altering in the 13S globulin subunit composition. Because only two types of zero-repeat genes were identified thus far, it might be less difficult to explore and develop the zero-repeat defective genotype (although it is not known whether the zero-repeat genes are allelic or non-allelic). In contrast, multiple-repeat genes are likely to exhibit more rapid evolutionary expansion, suggesting that further expansion and accumulation of these genes and gene products may reduce the relative accumulation level of zero-repeat proteins. In this context, the mechanism underlying the diversification of the 13S globulin genes should be minutely examined.

The full sequence of the zero-repeat genes was determined for the first time in this study. Previously we thought the low trypsin-digestibility of zero-repeat subunits was simply attributable to a lack of the hydrophilic tandem repeat containing arginine residues. However, the identity of amino acid sequences between tandem repeat containing subunits was greater than 75%, while that between subunits with and without tandem repeats was as low as 51–61% (data not shown). Therefore, regions of the subunits other than the tandem repeat may affect variability in trypsin-digestibility. Accordingly, protein chemistry at a subunit level should be exploited for further characterisation of 13S globulin. Not only the inter-species diversity and the environment-dependent variation in gene product accumulation but also more detailed clarification of topological relationships among adjacent multiple-repeat genes are needed to describe the molecular evolution of the 13S globulin gene, which may be important issues for a hypoallergenic buckwheat production.

Note

Nucleotide sequences of newly identified 13S globulin genes were deposited at the DNA Data Bank of Japan with the following

accession numbers: *Glb4A* (AB828105), *Glb3A* (AB828106), *Glb3B* (AB828107), *Glb3C* (AB828108), *Glb3D* (AB828109), *Glb2A* (AB828110), *Glb2B* (AB828111), *GlbXA* (AB828112), *Glb1A* (AB828113), *Glb1B* (AB828114), *Glb1C* (AB828115), *Glb1D* (AB828116), *GlbNA* (AB828117), *GlbNB* (AB828118), *GlbRA* (AB828119), *GlbRB* (AB828120), *GlbRC* (AB828121)

Conflict of interest

The authors have declared no conflict of interest.

Acknowledgement

This work was supported in part by the Ministry of Education, Culture, Sports, Science, and Technology, Japan [Scientific Research (C) 23580020, 2011–2013 to T.K.-T.].

References

- Adachi, M., Kanamori, J., Masuda, T., Yagasaki, K., Kitamura, K., Mikami, B., et al. (2003). Crystal structure of soybean 11S globulin: Glycinin A3B4 homohexamer. *Proceedings of the National Academy of Sciences of the United States of America*, 100, 7395–7400.
- Adachi, M., Takenaka, Y., Gidamis, A. B., Mikami, B., & Utsumi, S. (2001). Crystal structure of soybean proglycinin A1aB1b homotrimer. *Journal of Molecular Biology*, 305, 291–305.
- Aguilera, A., & Gómez-González, B. (2008). Genome instability: A mechanistic view of its causes and consequences. *Nature Reviews Genetics*, 9, 204–217.
- Brookes, K. J. (2013). The VNTR in complex disorders: The forgotten polymorphisms? A functional way forward? *Genomics*, 101, 273–281.
- Chen, Z.-Y., Jiao, R., & Ma, K. Y. (2008). Cholesterol-lowering nutraceuticals and functional foods. *Journal of Agricultural and Food Chemistry*, 56, 8761–8773.
- Derbyshire, E., Wright, D. J., & Boulter, D. (1976). Legumin and vicilin, storage proteins of legume seeds. *Phytochemistry*, 15, 3–24.
- Dickinson, C. D., Hussein, H. A., & Nielsen, N. C. (1989). Role of posttranslational cleavage in glycinin assembly. *The Plant Cell*, 1, 459–469.
- Ewing, B., & Green, P. (1998). Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8, 186–194.
- Ewing, B., Hillier, L., Wendl, M., & Green, P. (1998). Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, 8, 175–185.
- Gemayel, R., Vences, M. D., Legendre, M., & Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics*, 44, 445–477.
- Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, 95–98.
- Hamada, M., Kiryu, H., Sato, K., Mituyama, T., & Asai, K. (2009). Predictions of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, 25, 465–473.
- Hofmann, M. H., Heinrich, J., Radziwill, G., & Moelling, K. (2009). A short hairpin DNA analogous to miR-125b inhibits C-Raf expression, proliferation, and survival of breast cancer cells. *Molecular Cancer Research*, 7, 1635–1644.
- Jin, T., Guo, F., Chen, Y., Howard, A., & Zhang, Y.-Z. (2009). Crystal structure of Ara h 3, a major allergen in peanut. *Molecular Immunology*, 46, 1796–1804.
- Katsube, T., Kurisaka, N., Ogawa, M., Maruyama, N., Ohtsuka, R., Utsumi, S., et al. (1999). Accumulation of soybean glycinin and its assembly with the glutelins in rice. *Plant Physiology*, 120, 1063–1074.
- Khan, N., Katsube-Tanaka, T., Iida, S., Yamaguchi, T., Nakano, J., & Tsujimoto, H. (2008). Diversity of rice glutelin polypeptides in wild species assessed by the higher temperature SDS-PAGE and subunit-specific antibodies. *Electrophoresis*, 29, 1308–1316.
- Khan, N., Takahashi, Y., & Katsube-Tanaka, T. (2012). Tandem repeat inserts in 13S globulin subunits, the major allergenic storage protein of common buckwheat (*Fagopyrum esculentum* Moench) seeds. *Food Chemistry*, 133, 29–37.
- Kim, J. C., & Mirkin, S. M. (2013). The balancing act of DNA repeat expansions. *Current Opinion in Genetics & Development*, 23, 280–288.
- Kondo, Y., Urisu, A., Tokuda, R., Ishida, N., & Yasuda, T. (2001). Molecular characterization of a 24-kDa buckwheat protein, one of the major allergens of buckwheat seed. *Fagopyrum*, 18, 21–25.
- Mirkin, S. M. (2007). Expandable DNA repeats and human disease. *Nature*, 447(7147), 932–940.
- Nair, A., & Adachi, T. (1999). Immunodetection and characterization of allergenic proteins in common buckwheat (*Fagopyrum esculentum*). *Plant Biotechnology*, 16, 219–224.
- Park, J. W., Kang, D. B., Kim, C. W., Ko, S. H., Yum, H. Y., Kim, K. E., et al. (2000). Identification and characterization of the major allergens of buckwheat. *Allergy*, 55, 1035–1041.
- Radović, S. R., Maksimović, V. R., & Varkonji-Gasić, E. I. (1996). Characterization of buckwheat seed storage proteins. *Journal of Agricultural and Food Chemistry*, 44, 972–974.
- Sando, L., Pearson, R., Gray, C., Parker, P., Hawken, R., Thomson, P. C., et al. (2009). Bovine *Muc1* is a highly polymorphic gene encoding an extensively glycosylated mucin that binds bacteria. *Journal of Dairy Science*, 92, 5276–5291.
- Sato, K., Hamada, M., Asai, K., & Mituyama, T. (2009). CentroidFold: A web server for RNA secondary structure prediction. *Nucleic Acids Research*, 37, W277–W280.
- Sen, M., Kopper, R., Pons, L., Abraham, E. C., Burks, A. W., & Bannon, G. A. (2002). Protein structure plays a critical role in peanut allergen stability and may determine immunodominant IgE-binding epitopes. *The Journal of Immunology*, 169, 882–887.
- Sordet, C., Culerrier, R., Granier, C., Didier, A., & Rougé, P. (2009). IgE-binding epitope mapping on a three-dimensional model built for the 13S globulin allergen of buckwheat (*Fagopyrum esculentum*). *Peptides*, 30, 1021–1027.
- Takagi, S., Kimura, M., & Katsuki, M. (1993). Direct sequencing of PCR products using unlabeled primers. *Biotechniques*, 14, 218–221.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28, 2731–2739.
- Tandang-Silvas, M. R., Cabanos, C., Peña, L. D. C., Rosa, A. P. B. D. L., Osuna-Castro, J. A., Utsumi, S., et al. (2012). Crystal structure of a major seed storage protein, 11S proglubulin, from *Amaranthus hypochondriacus*: Insight into its physico-chemical properties. *Food Chemistry*, 135, 819–826.
- Tandang-Silvas, M. R., Fukuda, T., Fukuda, C., Prak, K., Cabanos, C., Kimura, A., et al. (2010). Conservation and divergence on plant seed 11s globulins based on crystal structures. *Biochimica et Biophysica Acta – Proteins and Proteomics*, 1804, 1432–1442.
- Wieskabder, G. (1996). Review on buckwheat allergy. *Allergy*, 51, 661–665.
- Yasui, Y., Mori, M., Matsumoto, D., Ohnishi, O., Campbell, C. G., & Ota, T. (2008). Construction of a BAC library for buckwheat genome research – An application to positional cloning of agriculturally valuable traits. *Genes & Genetic Systems*, 83, 393–401.
- Yoshioka, H., Ohmoto, T., Urisu, A., Mine, Y., & Adachi, T. (2004). Expression and epitope analysis of the major allergenic protein Fag e 1 from buckwheat. *Journal of Plant Physiology*, 161, 761–767.