**Research Paper**

**Structure and diversity of 13S globulin zero-repeat subunit, the trypsin-resistant storage protein of common buckwheat (Fagopyrum esculentum M.) seeds**

Fakhrul Islam Monshi¹², Nadar Khan¹³, Kohtaro Kimura¹, Seita Suzuki¹, Yuka Yamamoto¹ and Tomoyuki Katsube-Tanaka*¹

¹) Graduate School of Agriculture, Kyoto University, Kitashirakawa, Kyoto 606-8502, Japan
²) Faculty of Agriculture, Sylhet Agricultural University, Sylhet-3100, Bangladesh
³) Present address: Bio-Resources Conservation Institute, National Agricultural Research Centre (NARC), Park road ChakShehzad, Islamabad, Pakistan

The zero-repeat subunit of 13S globulin, which lacks tandem repeat inserts, is trypsin-resistant and suggested to show higher allergenicity than the other subunits in common buckwheat (Fagopyrum esculentum Moench). To evaluate allelic variations and find novel alleles, the diversity of the zero-repeat genes was examined for two Japanese elite cultivars and 15 Pakistani landraces. The results demonstrated that two new alleles GlbNA1 and GlbNC1, plus three additional new alleles GlbNA2, GlbNA3, and GlbND, were identified besides the already-known GlbNA, GlbNB, and GlbNC alleles. In the Pakistani landraces, GlbNA was the most dominant allele (0.60–0.88 of allele frequency) in all except one landrace, where GlbNB was the most dominant allele (0.50 of allele frequency). Similar to GlbNC, the alleles GlbNA2 and GlbNA3 had extra ~200 bp MITE-like sequences around the stop codon. Secondary structure predictions of a sense strand demonstrated that the extra ~200 bp sequences of GlbNC, GlbNA2, and GlbNA3 can form rigid hairpin structures with free energies of −78.95, −67.06, and −29.90 kcal/mol, respectively. These structures may affect proper transcription and/or translation. In the GlbNC homozygous line, no transcript of a zero-repeat gene was detected, suggesting the material would be useful for developing hypoallergenic buckwheat.

**Key Words:** allergen, common buckwheat, MITEs, 13S globulin, seed storage protein.

---

**Introduction**

Common buckwheat (Fagopyrum esculentum Moench) is considered a healthy food crop because of its well-balanced amino acid composition, high dietary fiber content, and beneficial physiological functions, such as anti-hypercholesterolemic, anti-hypertensive, anti-carcinogenic, and anti-inflammatory activities (Chen et al. 2014, Wieslander and Norbäck 2001). Reducing allergic reactions to buckwheat seed is becoming a focus of research with the goal of improving nutritional and physiological quality. This will improve the beneficial impacts of buckwheat food on human health.

To date, several allergenic proteins of buckwheat have been identified and categorized as most prevalent (24 kDa), or less prevalent (30, 43, and 67 kDa) (Park et al. 2000). More recently, Cho et al. (2015) considered 16, 24, 40, 43, and 48 kDa as major allergenic proteins of common buckwheat. 13S globulin, a salt soluble legumin-like protein, accounts for about 43% of the total seed proteins. 13S globulin is composed of disulphide-bonded heterogeneous acidic (α) and basic (β) polypeptides (Radović et al. 1996). The α and β polypeptides are biosynthesized as a larger single precursor with a signal peptide (Fig. 1). The signal peptide is processed during translocation to the endoplasmic reticulum. Then the polypeptide is further processed into α and β polypeptides. The β polypeptide of 13S globulin is 24 kDa, which is recognized as one of the major allergens, Fag e 1 (Nagata et al. 2000, Nair and Adachi 1999).

The 13S globulin subunits of common buckwheat have been categorized into methionine rich (Met-rich) and
methionine poor (Met-poor) subunits. Both the subunits contain a pair of α and β polypeptides. In our previous report, Met-poor subunits were further divided into two types; 1) with variable number of tandem repeat sequences, and 2) without tandem repeat sequences (Khan et al. 2012). The tandem repeat sequences are hydrophilic with many arginine residues which are target of trypsin. Actually α polypeptides with the tandem repeat have been found to be easier to digest by trypsin than those without tandem repeat sequences.

Resistance to digestion by proteinases is one of the major characteristics of food allergens (Kopper et al. 2004, Sen et al. 2002). Therefore, the allergenicity of 13S globulin might be affected by the presence or absence of tandem repeat sequences. The allergenicity may be reduced by lowering the expression of genes with no tandem repeat sequences, also known as zero-repeat genes (Khan et al. 2012).

Recently, Sano et al. (2014) detected 17 open reading frames (ORFs) encoding 13S globulin from a genomic DNA library, indicating that the protein composes a multi-gene family, similarly as other 11S globulins such as glutein of rice (Oryza sativa L.) and glycinin of soybean (Glycine max (L.) Merr.) do. Out of the 17 ORFs, there are two zero-repeat genes GlbNA (GenBank accession no. AB828117) and GlbNB (GenBank accession no. AB828118). Katsube-Tanaka et al. (2014) identified GlbNC (GenBank accession no. LC484359), which had ~200 bp of MITE (miniature inverted-repeat transposable element)-like sequence inserted ~70 bp upstream from the stop codon.

The genes GlbNA, GlbNB, and GlbNC all belong to Met-poor subunits and show a high similarity to each other, possibly suggesting those genes are allelic. However, no other information about the diversity of zero-repeat genes in common buckwheat is available. Additionally, it is unknown whether or not the zero-repeat genes are located at a single locus. We also do not know if varietal and geographical differences in common buckwheat influence allelic frequency, or the structural variation of zero-repeat genes. Understanding the genetic variation of zero-repeat genes is important for the development of hypoallergenic buckwheat. Therefore, assessments of intra- as well as inter-varietal diversity of zero-repeat genes, including improved varieties, are needed.

The Gilgit-Baltistan region in northern Pakistan, which is surrounded by the Karakoram Range, the western Himalayas, the Pamir Mountains, and the Hindu Kush, is the western terminus of buckwheat cultivation in the Himalayan regions (Ohnishi 1994), connecting to the original birthplace of common buckwheat in the northwestern corner of Yunnan province (Ohnishi 1998). Buckwheat, as well as millet, are important crops in mountainous areas where maize cannot grow well due to the short summers (Ohnishi 1994). At least two waves of diffusion of common buckwheat have taken place in northern Pakistan; the first being pink flower genotypes which are currently growing in Hunza, Nagar, and the Hushe valley of Ghanche district, and the second being white flower genotypes in the Indus valley of Baltistan (Ohnishi 1994). Allozyme variability analysis indicated that northern Pakistan populations are

Fig. 1. Schematic representation of zero-repeat subunit gene structure and the position of PCR primers for genotyping and cloning. White, black and gray boxes indicate exon of α polypeptide, β polypeptide, and signal peptide or MITE-like sequence, respectively. Length of the boxes and lines between boxes indicate approximate length of nucleotide sequences. NA2-MITE, NA3-MITE, and NC-MITE indicate a MITE-like sequence specifically inserted in GlbNA2, GlbNA3, and GlbNC allele, respectively. Horizontal arrows indicate the position of PCR primers. Note that the primer ‘TAA-3’-NANB-rec’ was used for amplification of GlbNC but was not perfectly matched to GlbNC.
closely related with those of Kashmir (Ohnishi 1994), which are distantly related to European populations (Ohnishi 1993). However, no other research on the genotypic diversity of buckwheat from Pakistan has been reported, except for phytochemicals (Abbasi et al. 2015), SDS-PAGE pattern of seed proteins (Hussain et al. 2016b), and major and trace elements (Hussain et al. 2016a). No systematic breeding program exists for the improvement of buckwheat crops. Thus, farmers grow locally adapted genotypes and landraces, which are assumed to conserve valuable genetic resources.

In this study, we briefly confirmed the three published zero-repeat genes were alleles on single locus in common buckwheat and identified new unique zero-repeat alleles, some of which have MITE-like sequences possibly to form a rigid hairpin secondary structure. We also evaluated the inter-variety diversity in allele frequency of zero-repeat genes using Japanese elite cultivars and Pakistani landraces. The findings of this study extend our knowledge about zero-repeat genes in common buckwheat, which will be useful for developing hypoallergenic buckwheat.

**Materials and Methods**

**Plant materials and preparation of genomic DNA samples**

The common buckwheat Japanese indigenous cultivars, ‘Harunoibuki’ and ‘Shinano1’, were used as plant materials for preliminary genotyping. Twenty seeds from each cultivar were grown in soil filled plastic pots. Young leaves from three to four week old seedlings were collected from individual plants and stored at –80°C for genomic DNA isolation. Genomic DNA was extracted and purified from leaves using DNeasy Plant Mini Kit (QIAGEN).

Identification of 3′-Untranslated Region (3′-UTR) and downstream of 13S globulin gene GlbNC

For the isolation of the 3′-UTR of GlbNC gene, a genome walking approach was employed using Straight Genome walking approach was employed using Straight

**Table 1.** Sample numbers and accessions used for genotyping of Pakistani landraces

<table>
<thead>
<tr>
<th>Sample #</th>
<th>Accession</th>
<th>Province</th>
<th>District</th>
<th>Town</th>
<th>Altitude (m)</th>
<th>Longitude</th>
<th>Latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3716</td>
<td>Gilgit-Baltistan</td>
<td>Skardu</td>
<td>Keris</td>
<td>2,200</td>
<td>75°58'</td>
<td>35°13'</td>
</tr>
<tr>
<td>2</td>
<td>3717</td>
<td>Gilgit-Baltistan</td>
<td>Skardu</td>
<td>Dognai</td>
<td>2,370</td>
<td>76°11'</td>
<td>35°15'</td>
</tr>
<tr>
<td>3</td>
<td>3724</td>
<td>Gilgit-Baltistan</td>
<td>Skardu</td>
<td>Fiazpur</td>
<td>2,405</td>
<td>75°42'</td>
<td>35°27'</td>
</tr>
<tr>
<td>4</td>
<td>3726</td>
<td>Gilgit-Baltistan</td>
<td>Ghanche</td>
<td>Lunkha</td>
<td>2,705</td>
<td>76°27'</td>
<td>35°05'</td>
</tr>
<tr>
<td>5</td>
<td>3728</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>6</td>
<td>3732</td>
<td>Gilgit-Baltistan</td>
<td>Gilgit</td>
<td>Murtaza Abad</td>
<td>2,245</td>
<td>74°35'</td>
<td>36°16'</td>
</tr>
<tr>
<td>7</td>
<td>21079</td>
<td>Gilgit-Baltistan</td>
<td>Gilgit</td>
<td>Aliabad</td>
<td>2,140</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>8</td>
<td>21081</td>
<td>Gilgit-Baltistan</td>
<td>Gilgit</td>
<td>Aliabad</td>
<td>2,140</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>9</td>
<td>29217</td>
<td>Gilgit-Baltistan</td>
<td>Hunza Nagar</td>
<td>Nasirabad</td>
<td>2,040</td>
<td>74°21'</td>
<td>36°16'</td>
</tr>
<tr>
<td>10</td>
<td>29218</td>
<td>Gilgit-Baltistan</td>
<td>Hunza Nagar</td>
<td>Karimabad</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>11</td>
<td>29221</td>
<td>Gilgit-Baltistan</td>
<td>Hunza Nagar</td>
<td>Nagar</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>12</td>
<td>29223</td>
<td>Gilgit-Baltistan</td>
<td>Skardu</td>
<td>Skardu</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>13</td>
<td>29227</td>
<td>Gilgit-Baltistan</td>
<td>Gilgit</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>14</td>
<td>29229</td>
<td>Gilgit-Baltistan</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>15</td>
<td>E1-Hol*</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

n/a, not applicable; * The accession is now unavailable in the genebank collection.
Walk Kit (BEX Co., Ltd, Tokyo, Japan). About 400 ng of genomic DNA extracted from a seed that had the GlbNC allele, but not the GlbNA and GlbNB alleles, was digested with XbaI restriction enzyme. After one nucleotide (dCTP) elongation, the DNA was ligated to SWA-2 adaptor using T4 DNA ligase (Nippon Gene, Tokyo, Japan). The primary PCR amplification was performed with walking primer-1 (5′-CGCAGGGCTGGACGTCTCTTTAG-3′) and sequence-specific primer-1 (Mpoor-TAA-200: 5′-ATTGGAGTGGGT GGAGTTGGAAGACC-3′) along with KOD Plus DNA polymerase (Toyobo, Japan). PCR conditions were as follows: denaturation at 94°C for 2 min, followed by 35 cycles at 94°C for 30 s, 65°C for 30 s, and 68°C for 5 min. Following the primary PCR, a 100-fold diluted primary PCR product was used as a template for the secondary PCR. The secondary PCR was performed with walking primer-2 (5′-ATGCGGCCTCTCTTTAGGTTACAGTTGCTT-3′) and sequence-specific primer-2 (Shina_S03UNQ: 5′-CTGACCCCAACCATAATTAAAGC-3′). The PCR conditions were as same as for the primary PCR except that only 30 thermal cycles were performed. After the secondary PCR amplification, the PCR product was electrophoresed on agarose gel and the specific band was isolated and purified with FastGene Gel/PCR Extraction kit (Nihon Genetics Co. Ltd). Finally, the purified PCR product was cloned to pTAC-2 vector after the A-attachment reaction to acquire overhanging dA at the 3′-ends (Toyobo, Japan).

Genotyping of 13S globulin zero-repeat genes

Genotyping of 13S globulin zero-repeat genes (GlbNA, GlbNB, GlbNC, and GlbND) was conducted for single seed genomic DNA using common forward primer (Mpoor-TAA-200: 5′-ATTGGAGTGGGT GGAGTTGGAAGACC-3′) and allele-specific (allele group-specific) reverse primers designed at 3′-UTR and downstream region of 13S globulin gene containing genes of 13S globulin of succeeding PCRs. This resulted in the development of specific primers for GlbNC-homozygous line and the ‘Harunoibuki’ (non GlbNC-homozygous) were sampled periodically, frozen in liquid nitrogen, and stored in a freezer at –80°C. Total RNA was extracted with RNAs-oci!-P (Rizo, Japan) and treated with DNasel that was followed by column purification (Total RNA Extraction Column, Favorgen). RT-PCR was employed with TaKaRa RNA PCR Kit (AMV) Ver. 3.0 (TaKaRa, Japan) using a random 9 mer primer for cDNA synthesis and primers of Fw: MetPoor 0 repeat SP, ATGTCAACTAA CTCC-3′, and allele-specific reverse primers TAA 3′ 1A2A3D4A rc, CGGAGCTCTTTAAGTGGAGAAGCCTC for repeat-containing genes of 13S globulin of succeeding PCRs.

Results

3′-UTR and downstream region of 13S globulin gene GlbNC and development of specific primers

Because the coding region homology is high between GlbNA, GlbNB, and GlbNC, with the exception of the MITE-like sequence of GlbNC, the 3′-UTR and downstream region of GlbNC was isolated to develop specific primers to distinguish the three genes. The alignment of the 3′-UTR and downstream region of GlbNC, plus the regions of GlbNA and GlbNB, which were determined by primer walking of BAC clones 269I19 and 336B7 that were isolated by Sano et al. (2014), respectively, demonstrated large downstream sequence variations more than 250 bp away from stop codon (Supplemental Fig. 1). Evolutionary divergence between GlbNA and GlbNB, GlbNB and GlbNC, GlbNA and GlbNC were 0.026, 0.022, 0.022, respectively. This resulted in the development of specific primers for amplifying GlbNA, GlbNB, and GlbNC.

Preliminary genotyping of zero-repeat genes in Japanese elite cultivars

Genotyping of zero-repeat genes was performed using Japanese indigenous cultivars ‘Harunoibuki’ and the salt correction of Peyret (2000) (http://unafold.rna.albany.edu/?q=mfold/DNA-Folding-Form). Evolutionary divergence analysis was conducted using the Kimura 2-parameter in MEGA6 (Tamura et al. 2013).

Development of GlbNC-homozygous line

GlbNC-containing plants were identified using the allele-specific and GlbNC-MITE specific primer (Shina_S03_UNQ: 5′-CTGACCCCAACCATAATTAAAGC-3′) from the cultivar ‘Harunoibuki’. The GlbNC containing plants were naturally crossed with each other in the isolated environment of a phytotron to avoid cross pollination with other genotypes. The genotype of the GlbNC-homozygous line was confirmed with the allele group-specific primers and the size of PCR amplified products of the coding region.

For RNA experiments, immature seeds of the GlbNC-homozygous line and the ‘Harunoibuki (non GlbNC-homozygous)’ were sampled periodically, frozen in liquid nitrogen, and stored in a freezer at –80°C. Total RNA was extracted with RNA-extraction Column, Favorgen). RT-PCR was employed with TaKaRa RNA PCR Kit (AMV) Ver. 3.0 (TaKaRa, Japan) using a random 9 mer primer for cDNA synthesis and primers of Fw: MetPoor 0 repeat SP, ATGTCAACTAA CTCC-3′, and allele-specific reverse primers TAA 3′ 1A2A3D4A rc, CGGAGCTCTTTAAGTGGAGAAGCCTC for repeat-containing genes of 13S globulin of succeeding PCRs.

Secondary structure prediction of DNA and molecular evolutionary analysis

The secondary structures of MITE-like sequences of sense and antisense strands were analyzed with the computer programs Centroid Fold (Hamada et al. 2009, Sato et al. 2009) (http://rtools.cbrc.jp/centroidfold/), and Mfold (DNA folding form) that was developed by Zuker (2003) using the free energies rules by SantaLucia (1998) and the
‘Shinano1’, with the specific primers designed at 3'-UTR and downstream of GlbNA, GlbNB, and GlbNC. Because of the cross-pollination and diploid natures of common buckwheat, individual seed is expected to have different gene composition, thus 20 individual plants from each variety were analyzed separately. In ‘Harunoibuki’, the fragment amplification with GlbNA-specific primer resulted in two bands with the expected size of 500 bp, and a smaller one by 24 bp (Fig. 2). Thus, the corresponding genes were designated as GlbNA and GlbNA1, respectively. The amplification with GlbNB-specific primer showed a single band of 600 bp corresponding to the authentic GlbNB gene. The PCR amplification with GlbNC-specific primer resulted in two fragments that were approximately 760 bp and smaller 550 bp. Thus, the corresponding genes for the smaller band was named GlbNC1. The difference in fragment size was due to the insertion of the MITE-like sequence of GlbNC. The genotyping results demonstrated that all the seeds of two different cultivars showed only one or two fragments, suggesting the allelism of the five genes. Thus, hereafter, the five genes are treated as allele. Because GlbNA plus GlbNA1, and GlbNC plus GlbNC1, were amplified with the same specific primer, respectively, the alleles within each combination were treated as belonging to the same allele group.

When the varietal differences and allele frequencies in the two Japanese indigenous cultivars were compared, the allele frequencies of GlbNA were highest in both cultivars (0.33 and 0.53), followed by the GlbNC1 and GlbNC alleles in ‘Harunoibuki’, and the GlbNB and GlbNC1 alleles in ‘Shinano1’ (Table 2). The GlbNA1 allele was observed only in ‘Harunoibuki’. Thus, the allele frequencies of GlbNB and GlbNC, as well as GlbNA1, differed between the two cultivars, suggesting that this genotyping methodology seems to be useful for a varietal characterization.

**Genotyping of zero-repeat genes in Pakistani landraces**

A similar genotyping was conducted for 20 seeds from each of the 15 Pakistani landraces, most of which were collected in Gilgit-Baltistan region in northern Pakistan (Fig. 3). Because two seeds of #4_6 and #6_9 could not be genotyped with the three allele group-specific primers, the zero-repeat genes of the two seeds were isolated and sequenced. Subsequently, the new allele designated as GlbND was identified and a new allele group-specific primer for GlbND was developed and added to the genotyping routine.

The genotyping results showed the allele frequencies of GlbNA were highest in all landrace accessions (0.600–0.875), except for the E1-Hol landrace accession where the GlbNB was the highest (0.500), and the GlbNA allele frequency was the second highest (0.48) (Table 3). With the exception of sample #15 (acc. E1-Hol), the maximum allele frequencies of GlbNA1, GlbNB, GlbNC, GlbNC1, and GlbND were 0.075 (#1), 0.150 (#3, 10, 14), 0.050 (#1, 2), 0.150 (#3), and 0.225 (#4), respectively. The combined allele frequencies of non-GlbNA were high in most accessions from Skardu and Ghanche, which are located at the east side of the Gilgit-Baltistan (0.325: #1, 0.400: #2, 0.375: #3, 0.400: #4). The lowest allele frequencies of non-GlbNA (0.125–0.250) were found in the accessions from Hunza-Nagar and Gilgit, which are located at the middle and the north side of the Gilgit-Baltistan (#6, 7, 8, 9, 10, 11, 13) (Fig. 4).

**Secondary structure prediction of MITE-like sequences**

During the genotyping and preliminary genotyping of GlbNA4(NA1) in sample #11, two larger fragments with MITE-like sequences were amplified and the alleles were named GlbNA2 and GlbNA3 (Fig. 1). The MITE-like
sequence of GlbNA3 (229 bp) was inserted 70 bp upstream of the stop codon, at exact position of the MITE-like sequence of GlbNC (208 bp), with the same direct repeat sequence of 5′-TGGTATTTTCC-3′. Meanwhile the MITE-like sequence of GlbNA2 (198 bp) was inserted downstream of stop codon, with direct repeat sequence of 5′-GTTTAAA GG-3′.

The predicted secondary structures of the MITE-like sequences of GlbNC, GlbNA3, and GlbNA2 had significant numbers of hydrogen bonds with high base-pairing probability, resulting in a rigid hairpin structure with free energies at 37°C of −78.95, −67.06, and −29.90 kcal/mol for the sense strand, and −76.58, −71.65, and −24.24 kcal/mol for the antisense strand (Fig. 5). The MITE-like sequences forming rigid hairpin secondary structures might affect the expression or function of GlbNC, GlbNA3, and GlbNA2 genes and may have evolutionary implications.

Development of the GlbNC homozygous line and expression of the zero-repeat gene

Because the allele frequency of GlbNC was five-fold higher in ‘Harunoibuki’ than that in ‘Shinano1’, we isolated GlbNC containing plants from ‘Harunoibuki’ and cross pollinated the plants to enrich the GlbNC allele. In the next generation, GlbNC homozygous plants with long-style and short-style flowers were isolated to propagate the GlbNC homozygous seeds. Within the maturing seeds of the GlbNC homozygous line, no transcripts of a zero-repeat gene were detected, whereas transcripts of repeat-containing genes were detected (data not shown). In the maturing seeds of ‘Harunoibuki (non GlbNC homozygous)’, both transcripts of the zero-repeat gene and the repeat-containing gene were detected.
Discussion

Importance of characterization of zero-repeat genes and identification of new alleles

The zero-repeat genes (without tandem repeat sequences) of 13S globulin of common buckwheat encode a trypsin resistant protein that is, therefore, considered to show higher allergenicity than the other repeat-containing, trypsin-digestible 13S globulins (Khan et al. 2012). Because the genes have not been fully described, a characterization of the zero-repeat genes is crucially important in understanding allergenic allelic variation. Thus, the primary focus of this study was to explore a novel allele of the zero-repeat gene, which could be useful for developing a hypoallergenic buckwheat using the diversified sequence information at 3'-UTR and downstream. We identified two new alleles, GlbNA1 and GlbNC1, from the Japanese cultivars and three new alleles GlbNA2, GlbNA3, and GlbND, from the Pakistani germplasms. Yasui et al. (2016) developed the Buckwheat Genome DataBase (BGDB) that is composed of 387,594 scaffolds, 286,768 predicted coding sequences (CDSs), and 36,763 annotated CDSs, which were generated from a draft assembly of the buckwheat genome using short reads of 264.5 Gb obtained by Next-Generation Sequencing (NGS) of Illumina HiSeq 2000. A BLAST search of the BGDB with the GlbNA sequence as a query resulted in finding a complete, but intron mis-predicted, sequence of GlbNC1 (Fes_sc0022676.1.g000001.aua.1), plus an incomplete and divided sequence of GlbNA (Fes_sc0203109.1.g000001.aua.1, Fes_sc0028260.1.g000002.aua.1), an indication of the difficulty in the analysis of complicated 13S globulin sequence structures with short reads of NGS. Meanwhile, an amplicon deep sequence for zero-repeat gene coding regions divided into quarters, which were amplified from pooled genomic DNA in several cultivars, showed large numbers of diversified sequences but did not detect a novel allele containing a large insert similar to the GlbNC sequence (data not shown). Thus, the traditional genotyping and cloning methodology, such conducted in this study, are still effective for the analysis of diversified 13S globulin genes in common buckwheat.

MITE-like sequences of GlbNA2, GlbNA3, and GlbNC

Three types of MITE-like sequences of ~200 bp were
found around the stop codon of the alleles GlbNA2, GlbNA3, and GlbNC. A NCBI-BLAST search for the three MITE-like sequences resulted in the identification of only 25–33 bp similar sequences in other organisms, whereas a BLAST search (BLASTN 2.2.26) in the BGDB showed many hits in other scaffolds of the BGDB. For example, GlbNC-MITE yielded 152 sequences with E-values better than 1.0e-01, with many having high scores with extra 8–9 bases of thymine in the middle of sequence. GlbNA3-MITE gave 30 sequences with E-values better than 1.0e-01, with many having high scores that lack approximately 80 bases in the middle. The sequence of these 80 bases was not found in the BGDB. GlbNA2-MITE gave 50 sequences with E-values better than 1.0e-01. These results suggest that the three MITE-like sequences are unique in common buckwheat and are widely dispersed in the genome.

The secondary structure predictions demonstrated that the extra 200 bp sequence of GlbNC might form a rigid hairpin structure. MITEs play important roles in gene regulation and genome evolution (El Amrani et al. 2002, Kuang et al. 2009, Naito et al. 2006, Oki et al. 2008, Yang et al. 2005). MITEs may down regulate gene expression through MITE-derived small RNAs by either disrupting or altering gene structure (Kuang et al. 2009). The down regulation of genes by MITEs insertion has been reported by several researchers. For example, genome-wide analysis in rice showed that genes associated with MITEs have significantly lower expression than genes away from MITEs, and all genes with MITEs insertions have a larger proportion of weakly expressed genes than the genes with no MITEs insertions (Chen et al. 2012, Lu et al. 2012).

Formation of hairpin or stem-and-loop structures affects gene expressions, DNA recombination, and DNA transposition (Lah et al. 2011), plus inhibition of molecular biology techniques such as PCR and sequencing (Nelms and Labosky 2011). mRNA pseudoknots mediate ribosomal frameshfts or acts as roadblocks, whereby synthesis of multiple proteins is controlled from a single polycistronic mRNA (Tholstrup et al. 2012). Information carried by RNA in their primary, secondary, and tertiary structures influence the transcription, splicing, cellular localization, translation and turnover of RNA (Wan et al. 2011). Hairpins, stem-loops or triplexes formed during the lagging strand synthesis would disrupt DNA replication causing slippage or blockage (Aguilera and Gómez-González 2008). Thus, the possible secondary structure formation in the MITE-like sequences should be noted from this study.

The reason the MITE-like sequences were found only around the stop codons is noteworthy. A possible explanation is that the primers used for genotyping amplified the region between 200 bp upstream to 300–400 bp downstream from a stop codon. If we focused on other regions with other primers, we may have detected other MITE-like sequences being inserted. Another possible explanation may be that the C-terminus of β polypeptide or the 3′-UTR may be structurally tolerant to short polypeptide insertions and truncations or changes in mRNA secondary structure. Before verifying such hypotheses, expression and accumulation of mRNA and polypeptides that coincide with MITE-like sequences should be carefully studied.

**Evaluation of the allele frequency of zero-repeat gene**

The second focus of this study was to identify allele frequencies of GlbNC and other zero-repeat alleles. Although the number of zero-repeat gene loci have not been determined and is necessary to be examined with the progeny produced by the cross between each allele, the maximum number of alleles detected in a single seed did not exceed two, suggesting the locus is only one in a diploid common buckwheat genome. The number of zero-repeat alleles detected in the BGDB (Yasui et al. 2016) and the BAC clone library (Sano et al. 2014, Yasui et al. 2008) was also two. Meanwhile, we categorized eight zero-repeat alleles, with four allele group-specific primers. The four primers amplified at least one fragment in all of the single seeds examined, but the possibility of an unknown allele exist has not been eliminated. The current genotyping methodology will identify an unknown allele only when the unknown allele is homozygous. We examined more than 300 seeds, so at least one unknown allele-homozygous seed would be detected if the allele frequency of the unknown allele exceeded 0.058, which is estimated by the following calculation: 0.058^2 × 300 = 1.009, which is greater than one.

The observed heterozygosity at the zero-repeat locus averaged 0.36 in the 15 Pakistani germplasms (108 heterozygous seeds out of 300 examined seeds), while the expected heterozygosity was 0.44 (Table 3), suggesting a reduced diversity due to interbreeding.

Thus, when considering the small number of genotyped seeds (20 seeds from each line), the accuracy in the evaluation of allele frequency may have been limited in this study. Nonetheless, the detection of the GlbNC allele was without ambiguity and the allele frequency of GlbNC seemed reliable. Consequently ‘Harunoibuki’ showed the highest allele frequency of GlbNC, with a value of 0.25 and this cultivar was used for development of GlbNC homozygous lines.

**Development of the GlbNC homozygous lines**

Seven out of 15 accessions of Pakistani landraces showed amplified fragments derived from the allele GlbNC, with the allele frequency of GlbNC at a maximum of 0.05 in the accessions. Meanwhile the allele frequency of GlbNC in ‘Harunoibuki’ was 0.25, indicating that the cultivar would be good for developing a GlbNC homozygous line. After the enrichment of GlbNC-alleles in isolated populations, we successfully developed a GlbNC homozygous line.

Because the inserted MITE-like sequence of GlbNC was predicted to form a rigid hairpin structure, the transcription and/or translation of the GlbNC was expected to be affected. At a minimum, the translational product would be truncated because a new stop codon appeared in the MITE-
like sequence. At the moment, RT-PCR experiments demonstrated no zero-repeat gene transcripts were detected in the *GlbNC* homozygous line, although repeat-containing gene transcripts were detected in both the *GlbNC* homozygous line and ‘Haranuibuki’ (non *GlbNC* homozygous). Future work should be conducted, including western blot experiments with zero-repeat subunit-specific antibodies to determine the accumulation of zero-repeat subunits in *GlbNC* homozygous lines.

The accumulation of zero-repeat subunits may vary across other types of alleles and genotypes. The MITE-like sequences containing alleles *GlbNA2* and *GlbNA3* will be interesting for analysis. Even though the alleles *GlbNA2* and *GlbNA3* were found in only one of the 300 seeds, both alleles were identified in the accession ‘3728’, suggesting this accession is a promising germplasm to further analyze for finding a new allele.

Further genotyping using a wide range of buckwheat genetic resources, including land races as well as improved varieties, might offer an opportunity for discovering more diversity in zero-repeat genes. In our study, we only assessed the zero-repeat gene diversity in a limited number of Pakistani landraces and Japanese elite cultivars; nonetheless, our results will be helpful for better understanding zero-repeat genes. The findings of our current study will contribute to the efforts of developing hypoallergenic buckwheat.

### Author Contribution Statement

T.K.-T. designed the project, and wrote the initial draft of the manuscript. F.I.M. and N.K. contributed to analysis and interpretation of data, and assisted in the preparation of the manuscript. The other authors have contributed to data collection and interpretation.

### Acknowledgments

This work was supported by a fellowship to N.K. by the Japan Society for the Promotion of Science (JSPS) and a grant by the Ministry of Education, Culture, Sports, Science, and Technology, Japan [Scientific Research (C) 23580020, 2011-2013 to T.K.-T.]. Technical assistance provided by Ms. Madoka Sano, Mr. Naoto Mukainishi, and Ms. Mariko Nakagawa is acknowledged.

### Literature Cited


Liu, Z., W. Ishikawa, X. Huang, H. Tomotake, J. Kayashita, H. Watanabe and N. Kato (2001) A buckwheat protein product suppresses 1, 2-dimethylhydrazine-induced colon carcinogenesis in...


