DNA Research, 2016, 23(3), 215–224
doi: 10.1093/dnares/dsw012
Advance Access Publication Date: 2 April 2016
Full Paper

OXFORD

Full Paper

Assembly of the draft genome of buckwheat and its applications in identifying agronomically useful genes

Yasuo Yasui[1,†,*], Hideki Hirakawa[2,†], Mariko Ueno[1], Katsuhiro Matsui[3], Tomoyuki Katsube-Tanaka[1], Soo Jung Yang[1], Jotaro Aii[4], Shingo Sato[4], and Masashi Mori[5,*]

[1]Graduate School of Agriculture, Kyoto University, Kitashirakawa Oiwake-cho, Sakyou-ku, Kyoto 606-8502, Japan, [2]Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan, [3]NARO Kyushu Okinawa Agricultural Research Center, 2421 Suya, Koshi, Kumamoto 861-1192, Japan, [4]Faculty of Applied Life Science, Niigata University of Pharmacy and Applied Life Science, Akiha-ku, Niigata 956-8603, Japan, and [5]Research Institute for Bioresources and Biotechnology, Ishikawa Prefectural University, 308 Suematsu, Nonoichi, Ishikawa 912-8836, Japan

*To whom correspondence should be addressed: Tel. +81 75-753-6480. Fax. +81 75-753-6146. E-mail: yasyas@kais.kyoto-u.ac.jp (Y.Y.); Tel. +81 76-227-7527. Fax. +81-76-227-7557. E-mail: mori@ishikawa-pu.ac.jp (M.M.)

†Co-first authors.

Edited by Dr Katsumi Isono

Received 15 December 2015; Accepted 20 February 2016

Abstract

Buckwheat (*Fagopyrum esculentum* Moench; $2n = 2x = 16$) is a nutritionally dense annual crop widely grown in temperate zones. To accelerate molecular breeding programmes of this important crop, we generated a draft assembly of the buckwheat genome using short reads obtained by next-generation sequencing (NGS), and constructed the Buckwheat Genome DataBase. After assembling short reads, we determined 387,594 scaffolds as the draft genome sequence (FES_r1.0). The total length of FES_r1.0 was 1,177,687,305 bp, and the N50 of the scaffolds was 25,109 bp. Gene prediction analysis revealed 286,768 coding sequences (CDSs; FES_r1.0_cds) including those related to transposable elements. The total length of FES_r1.0_cds was 212,917,911 bp, and the N50 was 1,101 bp. Of these, the functions of 35,816 CDSs excluding those for transposable elements were annotated by BLAST analysis. To demonstrate the utility of the database, we conducted several test analyses using BLAST and keyword searches. Furthermore, we used the draft genome as a reference sequence for NGS-based markers, and successfully identified novel candidate genes controlling heteromorphic self-incompatibility of buckwheat. The database and draft genome sequence provide a valuable resource that can be used in efforts to develop buckwheat cultivars with superior agronomic traits.

Key words: buckwheat, draft sequence, database usage, GBS marker, heteromorphic self-incompatibility

1. Introduction

The genomes of model plants, such as *Arabidopsis thaliana* and *Oryza sativa* (rice), were fully sequenced by the start of the 21st century, and databases containing chromosomal pseudo-molecules and gene annotation information have subsequently been developed and are widely used as tools and resources for plant genomics and genetics studies. Recently, next-generation sequencing (NGS) has emerged as a powerful technique for analysing the genomes of non-model crops

in which few molecular genetic studies have been performed. Genome sequences obtained by NGS can be used to construct databases that contain information of genes inferred from available information of genes in model plants. These genome databases in non-model crops will pave the way for the rapid identification of useful genes for crop breeding, which have already been identified in model plants. These databases will also facilitate the construction of fine genetic maps [based on single nucleotide polymorphism, simple sequence repeat (SSR), and NGS-based markers], and make it possible to identify agronomically important genes by map-based cloning. Thus, genome analyses in various non-model crops are underway, and the genomes of >50 non-model crops have already been sequenced (CoGepedia; https://genomevolution.org/). For example, NGS has been used to sequence the genomes of crops that produce beneficial secondary metabolites, such as flavonoid-producing *Viburnum trilobum* (American cranberry)[1] and capsaicin-producing *Capsicum annuum* (hot pepper),[2] and of crops that are tolerant to environmental stress, such as *Setaria italica* (foxtail millet)[3] and *Cajanus cajan* (pigeonpea),[4] which grow in semi-arid regions. NGS technology has opened the door to elucidating the molecular mechanisms that control agronomically important traits, and there is much interest in using this technology to analyse the genomes of non-model crops.

Buckwheat (*Fagopyrum esculentum* Moench; $2n = 2x = 16$) is a widely cultivated annual crop in temperate zones. This nutritionally dense non-model crop contains high levels of starch, protein, flavonoids, and dietary fibre in the grain.[5] Furthermore, buckwheat flour is gluten-free and can replace wheat flour in a coeliac diet.[6] Buckwheat, however, has two major defects as a crop. First, its outcrossing nature, caused by heteromorphic self-incompatibility (SI), makes it difficult to produce pure cultivars of buckwheat and to fix useful traits. Second, buckwheat grains contain allergens, which induce anaphylactic reactions in some people.[7] Improving the nutritional quality of the grain and removing genes responsible for SI and allergens are important breeding objectives in buckwheat, and various genetic molecular marker systems have been developed for this purpose [e.g. amplified fragment length polymorphism (AFLP) markers,[8] SSR markers,[9] expressed sequence tag (EST) markers,[10] and array-based markers[11]]. However, AFLP markers have not yet been converted to single locus markers in the buckwheat genome. SSR markers have limited utility in buckwheat due to difficulty in amplifying specific loci because of the high level of genetic diversity between buckwheat cultivars, and EST marker systems do not span the entire genome. The newest genome map of buckwheat constructed using array-based markers has sufficient markers to cover the entire genome; however, it requires a specialized instrument to interpret the fluorescence signals of the arrays.[11] Recently, a versatile NGS-based genotyping method with a low-cost, genotyping-by-sequencing (GBS) marker system was developed.[12] The GBS system utilizes redundant libraries constructed with PCR fragments that have recognition sites of two kinds of restriction enzymes on both ends. The PCR fragments sequenced using NGS technology are mapped to reference sequences for genome-wide genotyping. The GBS system has been used to genotype various crop species to date.[13] A draft genome of buckwheat could be used as a reference sequence for developing GBS markers to identify genes that control desirable breeding traits.

Here, we used NGS-based technology to sequence the buckwheat genome, and constructed the Buckwheat Genome DataBase (BGDB; http://buckwheat.kazusa.or.jp). This database can be used for the rapid detection of homologues of genes previously identified in other plants, and we present three examples of buckwheat genes identified using this approach, i.e. genes controlling flavonoid biosynthesis and genes encoding 2S albumin-type allergens and granule-bound starch synthases (GBSSs). Furthermore, to illustrate that the draft genome can be used as a reference sequence for NGS-based genotyping, we used GBS technology to identify novel candidate genes for controlling heteromorphic SI of buckwheat.

2. Materials and methods

2.1. Plant material

A single buckwheat plant with short-styled flowers, a descendant of material used in a previous study to construct a buckwheat BAC library,[14] was obtained from sib-crossing (BC_1F_6). Nuclei were extracted from leaf tissues of the single plant as described previously.[14] Subsequently DNA was extracted from the nuclei according to a previously described method.[11] To construct a training set for gene prediction using Augustus 3.0.3,[15] total RNA was prepared from the anthers of short-styled and long-styled plants, cv 'KOTO' using a previously described method.[16]

2.2. Genome sequencing of buckwheat

A paired-end (PE) library with insert sizes of 180–200 bp and a mate-pair (MP) library with insert sizes of 3, 5, 10, and 20 kb were constructed from nuclear DNA according to the manufacture's protocol (Illumina Inc., San Diego, CA, USA). A PE RNA-Seq library with insert sizes of ~275 bp was also constructed. Sequencing of genomic and RNA-Seq libraries using Illumina HiSeq 2000 was respectively carried out at Hokkaido System Science Co., Ltd and Beijing Genomics Institute. The PE and MP reads were subjected to quality trimming by PRINSEQ 0.20.4,[17] and further to adaptor trimming by the fastx_clipper program in the FASTX-toolkit 0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit). The quality value threshold used for quality trimming was 10 from the 3′ terminal, and the adaptor sequence used was 'AGATCGGAAGAGC'. Then, for the PE library with insert sizes of 180 bp, one base at the 3′ terminal was trimmed from all reads due to low quality, and PE reads shorter than 99 bp and including undetermined nucleotides (Ns) were excluded. For the MP library with insert sizes of 3, 5, and 10 kb, reads shorter than 49 bp and including Ns were excluded, and the 50 bp from the 5′ terminal were used for scaffolding. For the MP library with an insert size of 20 kb, reads shorter than 99 bp and including Ns were excluded, and the 50 bp from the 3′ terminal were used for scaffolding. For the PE RNA-Seq data, reads shorter than 89 bp and including Ns were excluded. The trimmed reads were used for further analyses.

2.3. Estimation of genome size

For genome size estimation, we used PE reads with a k-mer size of 17, as successfully used in a previous study.[18] The k-mer distribution was investigated using Jellyfish 2.1.3.[19] The genome size and coverage (i.e. the number of base pairs sequenced as a multiple of the number of base pairs present in the genome) were estimated using the peak at 47 on the k-mer frequency distribution curve (Supplementary Fig. S1) according to a previously described method.[18]

2.4. Assembly of the buckwheat genome sequences

The trimmed PE reads were assembled using SOAPdenovo2 rev240[20] with k-mer sizes of 61, 71, 81, and 91 nt. The option used was –RF –M 1–K [k-mer size]. After the assembly, gaps in scaffolds were closed using GapCloser 1.10 (http://soap.genomics.org.cn/soapdenovo.html) ($P = 31$). The trimmed MP reads were used for scaffolding by

SSPACE2.0[21] with parameters –k 5 –x 0 –g 3 –a 0.7. Sequences homologous to bacteria, fungi, and human (hg19) genome sequences, vector sequences from UniVec (http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/), chloroplast (accession number: NC_000932.1), and mitochondrial (accession number: NC_001284.2) genome sequences from *A. thaliana*, and the PhiX sequence used in Illumina sequencing by BLASTN[22] searches with an *E*-value cut-off of 1E−10 and length coverage of ≥10% were excluded as probable contamination. Finally, scaffolds longer than 300 bp were selected and designated FES_r1.0. Repetitive sequences in FES_r1.0 were detected using RepeatScout 1.0.5[23] and RepeatMasker 4.0.3 (http://www.repeatmasker.org) according to a previously described method.[18]

2.5. Gene prediction and annotation

The RNA-Seq reads were mapped onto the draft genome sequence (FES_r1.0) with TopHat 2.0.12.[24] The bam2hints program installed in Augustus 3.0.3 was used to generate the intronhints.gff file, and Cufflinks was used to reconstruct transcripts in an exonhints.gff file. The two gff files were merged to form a HINTS file. The HINTS file was used as the buckwheat training set to predict genes for Augustus 3.0.3 (Method 1). Furthermore, genes were predicted using Augustus 3.0.2 (Method 2) and geneid 1.4.4[25] (Method 3) with an *A. thaliana* training set. Finally, the genes predicted by the three methods were merged.

The merged genes were subjected to similarity searches against NCBI's NR database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz) and amino acid sequences of *A. thaliana* from TAIR10 (https://www.arabidopsis.org) using BLASTX with an *E*-value cut-off of 1E−10. The top hit was used to assign the product name. BLAST searches against UniProt (TrEMBL + Swiss-Prot) with an *E*-value cut-off of 1E−20 were also carried out. A domain search against InterPro (http://www.ebi.ac.uk/interpro/) was conducted using InterProScan[26] with an *E*-value cut-off of 1.0. Finally, genes were classified based on NCBI's euKaryotic clusters of Orthologous Groups (KOG) database[27] by performing BLAST searches with an *E*-value cut-off of 1E−4.

Genes related to transposable elements were inferred based on a BLAST search against the NCBI's NR database, and conserved domains were identified based on a search against InterPro and GyDB 2.0[28] using hmmsearch in HMMER 3.0[29] with an *E*-value cut-off of 1.0. Transfer RNA genes (tRNAs) were predicted using tRNAscan-SE v.1.23.[30] Ribosomal RNA genes (rRNAs) were predicted in BLASTN searches with an *E*-value cut-off of 1E−10 using *A. thaliana* 5.8S and 25S rRNAs (accession number: X52320.1) and 18S rRNA (accession number: X16077.1) as queries.

2.6. Database construction

The draft genome sequence (FES_r1.0), predicted gene sequences, deduced amino acid sequences, annotations derived from BLAST searches against the TAIR10 and NCBI's NR databases, and domains identified in the search against InterPro were included in the BGDB. In addition, local BLAST searches and keywords searches for gene names and their annotations were also implemented in the BGDB.

2.7. Genotyping-by-sequencing and detection of S-allele-specific sites

Total DNA extracted from 18 short-styled and 18 long-styled buckwheat landraces from around the world (Supplementary Table S1) was used for GBS analysis. GBS was carried out according to Elshire et al.,[12] except that *Eco*RI and *Mse*I were used as restriction enzymes. Barcode adaptors are listed in Supplementary Table S2. Barcode-labeled amplicons were sequenced by Illumina HiSeq 2000 at Hokkaido System Science Co., Ltd. The PE reads were subjected to quality and adaptor trimming by Trimmomatic 0.3.2.[31] The quality value threshold used for quality trimming was 25 with a window size of 5, and the adaptor sequences used were 'CACGAC GCTCTTCCGATCT' and 'ACCGCTCTTCCGATCTGTAA'. Then, PE reads longer than 39 bp were aligned to reference sequences (FES_r1.0) using BWA 0.7.9,[32] and the mapping results were processed with SAMtools 0.1.18.[33] To minimize mismatching bases across all the reads, local realignment procedure was carried out using RealignerTargetCreator and IndelRealigner in GATK 3.4.[34] All the sites on reference sequences that mapped with reads were extracted and combined in a variant call format file using the UnifiedGenotyper in GATK 3.4 with the option of -out_mode EMIT_ALL_CONFIDENT_SITES. Sites at which >50 reads were mapped in long-styled plants but not in short-styled plants were defined as 'non-SS'. Likewise, sites at which >50 reads were mapped in short-styled plants but not in long-styled plants were defined as 'non-LS'. Then, the number of short-styled plants with mapped reads at each non-LS site and the number of long-styled plants with mapped reads at each non-SS site were counted. Non-LS sites shared by >10 short-styled plants were regarded as *S*-allele-specific sites (see Results and discussion). Then, scaffolds harbouring >39 *S*-allele-specific sites were regarded as *S*-allelic scaffolds.

2.8. Phylogenetic analyses

Alignments of amino acid sequences were carried out using CLUSTALW,[35] and the neighbor-joining (NJ) trees[36] were obtained from pairwise distances corrected by the JTT model.[37] These analyses were conducted using MEGA6.[38]

3. Results and discussion

3.1. Genome assembly of buckwheat

The k-mer frequency distribution curve (k-mer = 17) using PEs with 180 and 200 bp is shown in Supplementary Fig. S1. Based on this curve, the genome size of buckwheat was estimated to be between 1,212,021,130 and 2,424,042,260 bp using peaks at a multiplicity of 94 (coverage = 111.9) and 47 (coverage = 56.0), respectively. The genomic DNA used in this study is expected to contain heterozygous regions, due to the outcrossing nature of buckwheat; however, we used sib-mating descendant plants as material to reduce heterozygous genomic regions. The haploid genome size of 1.2 Gb calculated based on the major peak (multiplicity = 94) is almost the same as that estimated from cytometry analyses (1.34 Gb).[39]

The numbers of raw and trimmed reads are summarized in Supplementary Table S3. The trimmed reads with k-mer sizes of 61, 71, 81, and 91 nt were assembled using SOAPdenovo2. The N50 values of the assemblies using k-mer sizes of 61, 71, 81, and 91 nt were, respectively, 1,388, 1,419, 1,350, and 770 bp. The longest scaffolds, i.e. those assembled with a k-mer size of 71, were used for further analysis. Gaps in the contigs were closed using GapCloser 1.10 (http://soap.genomics.org.cn), and mate-pair reads were used for scaffolding in SSPACE2.0. The 2,693,661 scaffolds that were shorter than 299 bp and the 1,908 scaffolds that exhibited signs of contamination (identified in a BLAST search) were excluded, and the remaining 387,594 scaffolds were designated as the draft genome sequence, FES_r1.0 (Table 1). The total length of FES_r1.0 was 1,177,687,305 bp, and the N50 length was 25,109 bp. The scaffolds were named 'Fes_sc' followed by a six-digit identifier and the sequence version (e.g. Fes_sc000001.1).

Table 1. Statistics of the draft genome sequence (FES_r1.0)

Number of sequences	387,594
Cumulative length of sequences (bases)	1,177,687,305
Average length of sequences per contig (bases)	3,038
Max length of sequences (bases)	1,053,114
Min length of sequences (bases)	300
N50 length (bases)	25,109
Number of undetermined bases	309,030,247
G ± C% (GC/ATGC)	39.1

Considering the genome size, the total length of the assembled genome sequence was close to the estimated size; therefore, the draft genome sequence (FES_r1.0) was considered as the haploid genome sequence. The draft genome sequence spanned 98.3% of the genome size estimated in the k-mer frequency distribution analysis, and 88.1% of that estimated by flow cytometry analysis.[39] The high coverage rates, low N50 value (25,109 bp), and large total number of scaffolds (387,594) obtained in the present study may be due to the high proportion of heterozygous genomic regions present in the plant materials used, as indicated by the k-mer frequency distribution curve (Supplementary Fig. S1). Long-read data generated by PacBio RS was found to increase N50 and reduce the total number of scaffolds in the draft genome of *Primula veris* (cowslip), a heterozygous plant.[40] A study of *Raphanus sativus* (radish)[41] indicated that the lengths of scaffolds assembled based only on short-read data were drastically increased by constructing super-scaffolds with SSPACE2.0 using BAC-end sequences. We have already constructed a BAC library[14] using parental plants of the material used in this study. Long reads generated by PacBio RS and BAC-end sequencing will be used to expand the scaffold length of the present draft genome of buckwheat.

3.2. Gene prediction and annotation

Gene predictions were performed using Augustus 3.0.3 with the buckwheat training set (Method 1), Augustus 3.0.2 with the *A. thaliana* training set (Method 2), or geneid with the *A. thaliana* training set (Method 3), and the results obtained using the three methods (Methods 1–3) are summarized in Supplementary Table S4. If the genes were located at the same locus when using Methods 1–3, the longest gene was selected. After the results were integrated, the total length of the CDSs (FES_r1.0_cds) was 212,917,911 bp composed of 286,768 CDSs, and N50 was 1,101 bp. The gene name was prefixed with a six-digit identifier followed by the prediction method and scaffold number (i.e. auf: Augustus 3.0.3, buckwheat training set, Method 1; aua: Augustus 3.0.2, *A. thaliana* training set, Method 2; and gia, geneid 1.4.4, *A. thaliana* training set, Method 3), as in the following example: Fes_sc0012271.1.g000001.aua.1.

Genes related to transposable elements (TEs) were inferred according to BLAST searches against the NCBI's NR database (Supplementary Table S5). The total length of known repeats was 133,362,886 bp (11.32% of FES_r1.0, i.e. 1,177,687,305 bp) and Class I long terminal repeat (LTR) elements were frequently found (8.79% of FES_r1.0). We identified unique repeats that had not previously been sequenced in this analysis, and these had a total length of 475,367,120 bp and accounted for 40.36% of FES_1.0. Genes annotated as transposons were tagged 'TE' in the database.

Based on BLAST searches against the NR database, the genes were tagged as intrinsic (including both of a start codon and stop codon), partial (including a start codon or stop codon, or lacking both start and stop codons), pseudo (pseudogenes; including a stop codon in the coding region), and short (<49 amino acids). Based on BLAST searches against UniProt (TrEMBL + Swiss-Prot) with an E-value cut-off of 1E−20, the genes were tagged 'f' (hit region of ≥70% in query length), 'p' (hit region of <70% in query length), or 'n' (no hits against UniProt). The tags assigned to the genes are listed in Supplementary Table S6; 35,816 predicted genes tagged as full length or partial were annotated by BLAST searches against NR and/or UniProt. The predicted genes were classified based on NCBI's KOG database for *F. esculentum*, *Beta vulgaris* (which is classified in the same order, Caryophyllales, as *F. esculentum*), and *A. thaliana* (the most advanced model plant). KOGs for the predicted genes in the three species were assigned to the 25 functional categories, which were classified into four large groups (Groups 1–4; Supplementary Table S7), and the percentage of KOGs in each category was calculated for each plant species. Figure 1 shows the percentage of KOGs in Groups 1–3. Note that the number of KOGs assigned to the functional categories of N (cell motility) in Group 2 was <10 for all three species and was excluded in Fig. 1. The distribution of KOGs in each large group was similar among these species. For instance, KOGs from all three species were enriched in categories K, O, and G, and were poor in categories B, W, and H. In addition to the protein-coding genes, we identified 1,374 genes for tRNAs in total, and the numbers of genes for each tRNA are summarized in Supplementary Table S8.

Consequently, we were able to identify >35,000 annotated genes including genes for tRNA. Many of them were classified into similar functional categories as in the other two plant species. The BGDB constructed in this study on the basis of FES_r1.0 is expected to serve as a useful tool for identifying genes to develop buckwheat cultivars with improved agronomic properties. The BGDB is available from the Kazusa DNA Research Institute (http://buckwheat.kazusa.or.jp). To demonstrate the utility of this database, we then conducted four test analyses focusing on agronomically important genes.

3.2.1. Example I: identifying buckwheat genes that regulate flavonoid biosynthesis

Buckwheat contains several kinds of flavonoids, such as flavonols, proanthocyanidins, and anthocyanins. The flavonol rutin is present at high levels in buckwheat seeds[42] and seems to be beneficial for human health. Several genes encoding enzymes related to flavonoid biosynthesis in buckwheat have been reported[43,44] and are presumed to be regulated by transcription factors (TFs), such as MYB, bHLH, and WD40, as in other plant species.[45–48] However, little is known about such TFs in buckwheat. The R2R3-MYB TFs are thought to play central roles in plant-specific processes, based on their specific gene expression patterns.[49–51] To provide an overview of genes that regulate plant-specific processes, including flavonoid synthesis in buckwheat, we searched for candidate genes encoding R2R3-MYB TFs using the BGDB.

By conducting a keyword search using the term 'MYB', we identified 274 genes predicted to encode MYB TFs. From these, we excluded partial sequences, pseudogenes, and genes that did not contain fully conserved R2R3 regions. The remaining 71 putative R2R3-MYB TFs obtained from the database are listed in Supplementary Table S9. Phylogenetic analyses based on the R2R3 domain often reveals functionally characterized groups that are present in a wide range of plant species.[52,53] In the present study, six putative R2R3-MYBs were assigned within known functional groups consisting of representatives from other plant species (Supplementary Fig. S2). Though functional analyses would need to be conducted to determine the role of

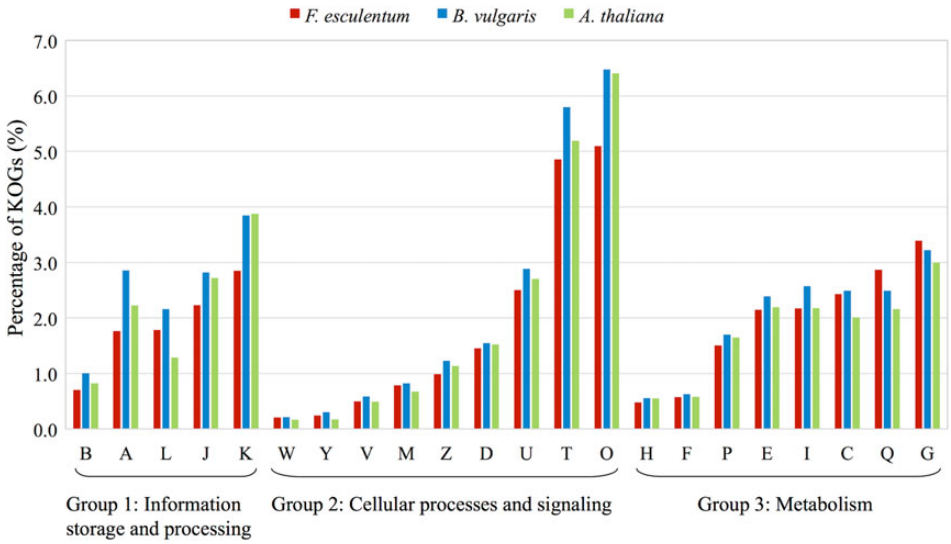

Figure 1. Assignment of proteins to KOG functional categories in the three plant species. Genes from *F. esculentum* (red), *B. vulgaris* (blue), and *A. thaliana* (green) were classified based on NCBI's KOG database by performing BLAST searches with an *E*-value cut-off of 1E−4. KOGs were classified into functional categories. The percentage of KOGs in each functional category is plotted, and percentages are arranged in ascending order within each group. (A) RNA processing and modification; (B) chromatin structure and dynamics; (C) energy production and conversion; (D) cell cycle control, cell division, and chromosome partitioning; (E) amino acid transport and metabolism; (F) nucleotide transport and metabolism; (G) carbohydrate transport and metabolism; (H) coenzyme transport and metabolism; (I) lipid transport and metabolism; (J) translation, ribosomal structure, and biogenesis; (K) transcription; (L) replication, recombination, and repair; (M) cell wall/membrane/envelope biogenesis; (O) posttranslational modification, protein turnover, and chaperones; (P) inorganic ion transport and metabolism; (Q) secondary metabolites biosynthesis, transport, and catabolism; (R) general function prediction only; (S) function unknown; (T) signal transduction mechanisms; (U) intracellular trafficking, secretion, and vesicular transport; (V) defense mechanisms; (W) extracellular structures; (Y) nuclear structure; and (Z) cytoskeleton. Note that KOGs in Groups 1–3 are shown, and that fewer than 10 KOGs were assigned to category N (cell motility) in the three species and were excluded.

Table 2. Fag e 2 and its homologues obtained by BLASTP search for buckwheat genome database

Gene ID of the BGDB	Scaffold ID	Similarity with reported allergen of *F. esculentum*	*E*-value
Fes_sc0000087.1.g000011.aua.1	Fes_sc0000087.1	97% (BW 8 kDa allergen protein)	8e−74
Fes_sc0000087.1.g000013.aua.1	Fes_sc0000087.1	79% (Fag e 2)	3e−26
Fes_sc0000087.1.g000014.aua.1	Fes_sc0000087.1	100% (Fag e 2)	7e−104
Fes_sc0000087.1.g000028.aua.1	Fes_sc0000087.1	98% (BW 8 kDa allergen protein)	5e−85
Fes_sc0007211.1.g000003.aua.1	Fes_sc0007211.1	43% (BW 8 kDa allergen protein)	4e−21

Scaffold ID, ID number of the scaffold in which the predicted gene is situated; BW 8 kDa, buckwheat 8 kDa allergen of 2S albumin; Fag e 2, buckwheat 16 kDa allergen of 2S albumin.

The GenBank accession numbers of BW 8 kDa and Fag e 2 are AB055892 and DQ304682, respectively.

each gene, this finding shows that R2R3-MYB genes, which likely have different roles, can successfully be obtained from the BGDB.

To initiate the transcription of genes encoding enzymes in the flavonoid biosynthetic pathway, a TF such as MYB or the MYB-bHLH-WD40 (MBW) complex must bind to TF binding sites (TFBSs) in the promoter region of each gene. Mutation of TFBSs alters the expression of genes.[54,55] Therefore, MYB TFs as well as the TFBSs of target genes can be manipulated to improve flavonoid production. To identify promoter sequences in a non-model plant species, genome walking, which is time-consuming and expensive, would usually be performed. In this study, we tried to identify the promoter sequences of genes in the flavonoid biosynthetic pathway and estimated the TFBSs using PLACE (http://www.dna.affrc.go.jp/PLACE/). cDNA sequences of nine genes in the pathway have been already registered in GenBank. For three of these genes (chalcone isomerase, CHI; flavonoid 3′-hydroxylase, F3′H; anthocyanidin synthase, ANS), we successfully determined the 1,000–2,000 bp upstream region after a BLASTN search against the BGDB. DNA motifs relating to the MYB or MBW complex predicted by PLACE analysis are shown in Supplementary Fig. S3. These results can be confirmed by molecular techniques such as gel-shift assays, as reported for other plant species (e.g. apple,[56] persimmon,[53] and soybean[57]). The BGDB is thus a powerful tool for isolating promoter sequences and accelerating molecular-based analyses of TFs. In this study, however, we could not identify sufficiently long promoter regions (over 1,000 bp) for the remaining six genes, mainly because of gaps between contigs. Gap closing using long reads generated by a PacBio sequencer will greatly improve the ability to search for the promoter regions of target genes.

3.2.2. Example II: identifying a duplicate of a buckwheat allergen gene, Fag e 2

Buckwheat seeds contain allergens. For instance, Fag e 2 (16 kDa protein) is a pepsin-resistant 2S albumin that causes an immediate allergic reaction.[58] Although *Fag e 2* cDNA has been sequenced,[58] no further genomic information is available. Efforts to develop hypoallergenic buckwheat and establish inspection techniques to minimize allergen contamination in food products require detailed genomic information of *Fag e 2*.

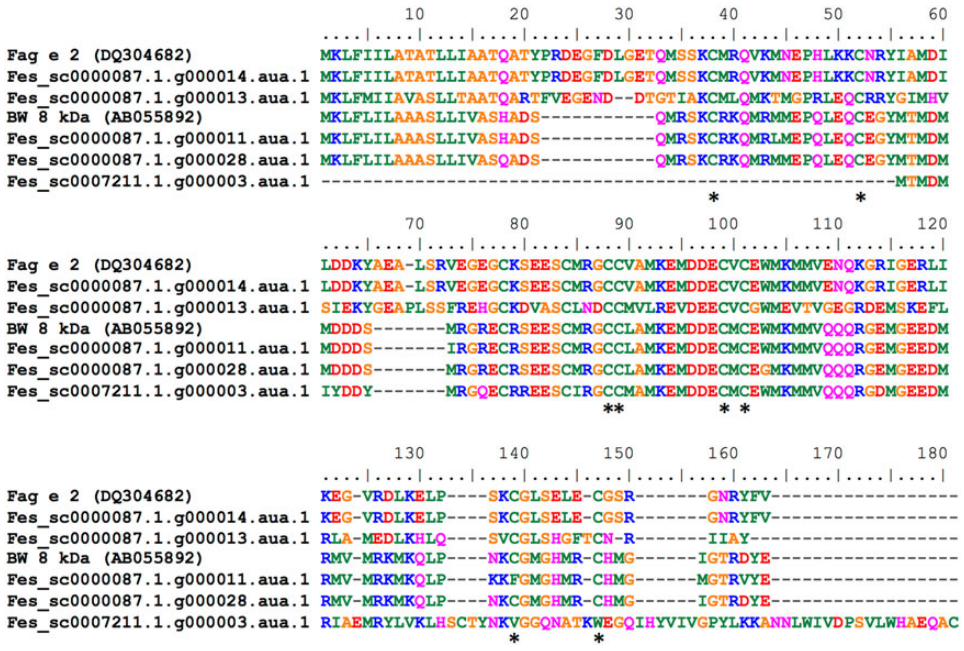

Figure 2. Alignment of amino acid sequences of Fag e 2 (buckwheat 16 kDa allergen of 2S albumin; GenBank accession number: DQ304682), BW 8 kDa (buckwheat 8 kDa allergen of 2S albumin; GenBank accession number: AB055892) and predicted sequences of homologues obtained from the BGDB. The following amino acid colour code is used: orange, small nonpolar (Gly, Ala, Ser, and Thr); green, hydrophobic (Cys, Val, Ile, Leu, Pro, Phe, Tyr, Met, and Trp); magenta, polar (Asn, Gln, and His); red, negatively charged (Asp and Glu); and blue, positively charged (Lys and Arg). Asterisks indicate the eight characteristic Cys residues present in Fag e 2 and 2S albumin family proteins, as described by Satoh et al.[60] Note that Fes_sc0007211.1.g000003.aua.1 had low similarity with other amino acid sequences and lacked four of the eight characteristic Cys residues.

A BLAST search of Fag e 2 (accession number: DQ304682) among the predicted proteins in the BGDB yielded one identical gene (Fes_sc0000087.1.g000014.aua.1) and four homologues (Fes_sc0000087.1.g000011.aua.1, Fes_sc0000087.1.g000013.aua.1, Fes_sc0000087.1.g000028.aua.1, and Fes_sc0007211.1.g000003.aua.1) (Table 2). The results of a BLAST search against the NCBI's NR database indicated high similarities of four homologues with previously reported allergens of buckwheat. As shown in Fig. 2, the predicted amino acid sequences of Fes_sc0000087.1.g000011.aua.1 and Fes_sc0000087.1.g000028.aua.1 showed high levels of similarity (97 and 98%, respectively) with the buckwheat 8 kDa allergen, which is a member of the 2S-albumin multi-gene family.[59] In contrast, the predicted proteins of Fes_sc0007211.1.g000003.aua.1 and Fes_sc0000087.1.g000013.aua.1 did not have high levels of amino acid sequence similarity (43 and 79%, respectively) with known buckwheat allergens. Fes_sc0007211.1.g000003.aua.1 had a 55-amino acid deletion at the N terminal and lacked four of eight characteristic Cys residues present in Fag e 2 and 2S albumin family.[60] Thus, the gene product is not likely to be allergenic. On the other hand, Fes_sc0000087.1.g000013.aua.1, which shows similarity with Fag e 2, might be a novel allergen, because its predicted protein retained the eight characteristic Cys residues. Further immunoblotting analysis will clarify whether or not the protein encoded by Fes_sc0000087.1.g000013.aua.1 is allergenic. It is notable that four genes that retain conserved Cys residues are estimated to be located within a genomic region of 108 kb on single scaffold (Fes_sc0000087.1). Particularly, the *Fag e 2* gene (Fes_sc0000087.1.g000014.aua.1), and the Fes_sc0000087.1.g000011.aua.1 and Fes_sc0000087.1.g000013.aua.1 homologues, which have similarities with buckwheat allergens, are located within a 17 kb region of the scaffold. Therefore, this region would be a prime candidate for silencing in studies aimed at producing hypoallergenic buckwheat.

3.2.3. Example III: identifying *GBSS* gene in buckwheat

Starch, which is present in many plant seeds, legumes, and tuber crops, is an important part of the human diet and has industrial applications. Starch contains two types of glucose polymer, i.e. amylopectin and amylose.[61] Amylopectin is the major component of starch (60–90%),[62] whereas the amylose content varies among plant species.[63] Since the amylose content affects the properties of starch, modulating the amylose content has been an important breeding objective in crops.[64] GBSS catalyses amylose synthesis.[65] Thus, we searched for genes encoding starch synthases (SSs) from the BGDB, and aimed to identify *GBSS* genes using phylogenetic approaches.

To identify buckwheat *GBSS* genes, we conducted a keyword search using 'starch synthase' and obtained 42 hits. We then filtered these hits using modified five conserved sequence motifs proposed by Cao et al.[66] [i.e. P(2)K(1)GGL(1)D(4)L, VS(5)E, G(2)NG(7)P(2)D, R(3)QKG, D(5)S(2)EPC(1)L(1)Q(5)YG(8)GGL (numbers in parentheses represent numbers of amino acids)]. Of the eight resulting sequences, one was excluded, as it was annotated as a pseudogene. The seven remaining sequences were each derived from a different scaffold.

To assign these seven putative SSs of buckwheat to previously proposed phylogenetic groups,[67] we performed NJ analysis using the deduced amino acid sequences of SSs from various plant species. A keyword search using 'starch synthase' was also conducted in Phytozome 10.3 (http://phytozome.jgi.doe.gov/) analysing 36 angiosperm species (Supplementary Table S10). Of the 27,852 sequences identified, 238 sequences remained after the same filtering procedure as mentioned earlier and were subjected to phylogenetic analysis. Two GBSS sequences of *Fagopyrum* species, one from *F. esculentum* deposited at the EMBL/GENBANK/DDBJ (accession number: HW041459) and the other from *F. tataricum* (AHA36967.1),[68] were also included in the analyses. A NJ-tree based on the alignment was suggested to

contain the following five known phylogenetic groups: SSI, SSII, SSIII, SSIV, and GBSS (Supplementary Fig. S4). The SS sequences obtained from the BGDB belonged to four of the five classes. This suggests that the BGDB can be used to identify agronomically important genes. However, the previously deposited *GBSS* (HW041459) sequence is not identical to any of the seven sequences identified in the BGDB, and a BLASTN search using the coding region of HW041459 as a query detected only partially identical sequences over 360-bp and three scaffolds (Fes_sc0195744.1, Fes_sc0059460.1, and Fes_sc0005470.1). This is a shortcoming of the short scaffold size of the assemblies in the BGDB.

The GBSS clade contained four phylogenetically distinguishable sequences in total: HW041459 and three from BGDB. To clarify the detailed phylogenetic relationship among these four *GBSS* genes, we performed NJ analysis based on 71 aligned amino acid sequences of GBSS including those from *Physicomitrella patens* as outgroup (Fig. 3). The copy number of *GBSS* genes varies in plants; two diverged groups exist in the rosids[69] and several copies of *GBSS* in buckwheat seem to also have diverged, at least in two lineages. In the cladogram, a *GBSS* sequence, Fes sc0004292.1.g000004, clustered with a *GBSS* sequence from *F. tataricum*, which belongs to the same genus. *GBSS* of *F. tataricum* was confirmed to be expressed in the endosperm,[68] thus Fes sc0004292.1.g000004 is likely to be expressed in the endosperm too. If more than one *GBSS* is active within endosperms, the amylose content of buckwheat flour can be controlled by altering their copy number. In hexaploid wheat, Yamamori and Quynh[70] evaluated the dosage effects of three *GBSS* genes. Loss of function of these genes is expected to differentiate the starch properties of the grain; moreover, distinct proportions of amylose might be produced according to copy number of active *GBSS* genes. Studies are underway to examine the expression of the three buckwheat *GBSS* genes identified here, and also the previously identified one (HW041459).

3.2.4. Example IV: isolation of heteromorphic SI genes

Finally, we screened for candidate genes that control heteromorphic SI in buckwheat. Buckwheat is a heteromorphic self-incompatible crop with dimorphic flowers (i.e. short-styled and long-styled flowers). Short-styled flowers have short styles and long stamens, whereas long-styled flowers have long styles and short stamens. The SI response is expressed between plants bearing the same flower morph, but not between plants bearing different flower morphs. Flower morph and SI response are determined by a diallelic system at the *SELF-INCOMPATIBILITY* supergene complex locus (*S* locus); *S*/*s* heterozygotes and *s*/*s* recessive homozygotes bear short-styled and long-styled flowers, respectively.[71,72] Recently, *S-LOCUS EARLY FLOWERING 3* (*S-ELF3*), which controls the short-styled phenotype of buckwheat, was isolated.[16] Furthermore, it was suggested that recombination is strongly suppressed around *S-ELF3*.[16] Based on these findings, we predicted that *S*-allelic scaffolds exist in which heteromorphic SI-related genes other than *S-ELF 3* are located. Thus, we tried to detect *S*-allelic scaffolds in the draft genome and to identify novel candidate genes involved in the SI response in these.

To obtain *S*-allele-linked scaffolds from the draft genome, we used GBS reads obtained from each of 18 short-styled and 18 long-styled landraces of buckwheat originating from various countries (Supplementary Table S1). Briefly, GBS reads from 36 plants were mapped to scaffolds of >1,000 bp, and we subsequently extracted the non-LS sites (i.e. sites not present in all the long-styled plants) in which no reads from all of the 18 long-styled plants were mapped. The number

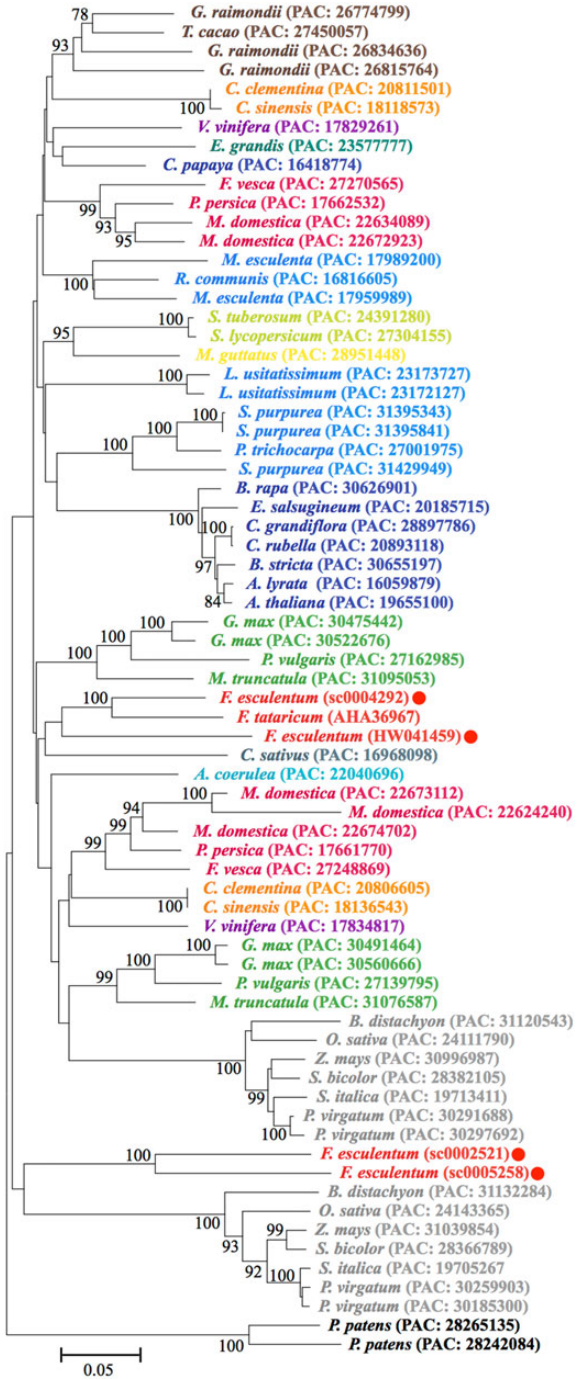

Figure 3. NJ tree based on amino acid sequences of GBSS from buckwheat and other plant species. The bootstrap values (500 replicates) not <50 are shown next to the branches. The scale bar corresponds to 0.05 substitutions per site. Two GBSSs from *Physcomitrella patens* were used as outgroup sequences. Species names are coloured according to their order: Poales, grey; Ranunculales, cyan; Vitales, purple; Cucurbitales, blue grey; Fabales, green; Malpighiales, blue; Rosales, pink; Myrtales, teal; Brassicales, indigo; Malvales, brown; Sapindales, orange; Caryophyllales, red; Lamiales, yellow; and Solanales, lime. Four GBSSs from *F. esculentum* are indicated by red circles next to the sequence names. Sequences obtained from BGDB are abbreviated (sc0002521: Fes_sc0002521.1.g000007; sc0004292: Fes_sc0004292.1.g000004; and sc0005258: Fes_sc0005258.1.g000004). Two sequences of *Fagopyrum* species (AHA36967 and HW041459) were obtained from GenBank. Sequences excluding those from *Fagopyrum* species were obtained using Phytozome 10.3 and the accession numbers are in parentheses.

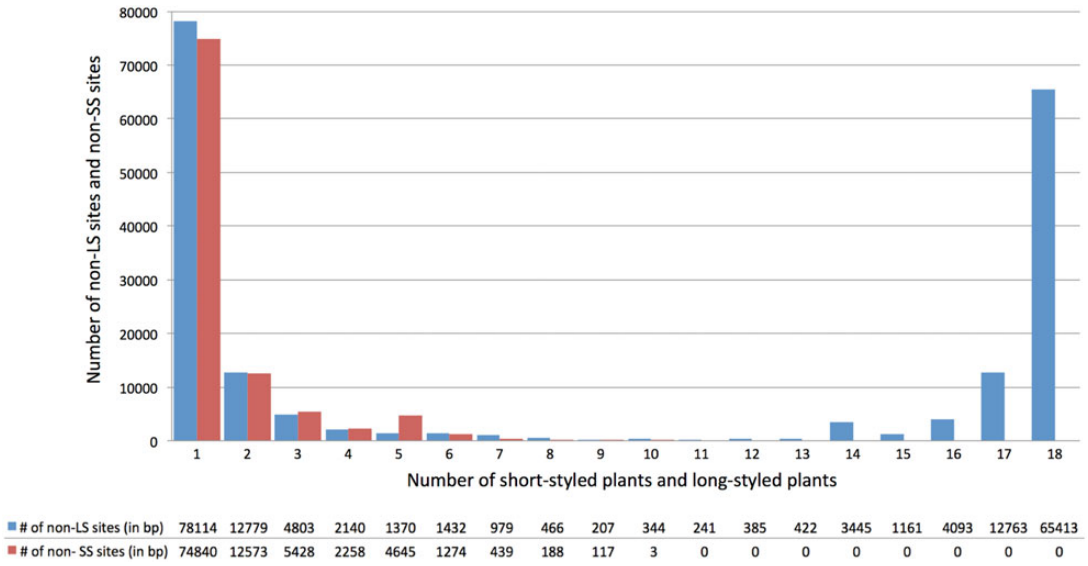

Figure 4. The number of non-LS sites and non-SS sites identified in GBS reads. Sites at which >50 reads were mapped in long-styled plants but not in short-styled plants were defined as 'non-SS'. Sites at which >50 reads were mapped in short-styled plants but not in long-styled plants were defined as 'non-LS'. The number of non-LS (blue bar) and non-SS sites (red bar) was plotted against the number of short-styled plants sharing the non-LS sites and of long-styled plants sharing the non-SS sites, respectively.

of mapped reads for each plant ranged from 1,620,260 to 2,821,338 (Supplementary Table S11). The number of mapped reads per long-styled plant was not significantly different from that per short-styled plant (*P*-value is 0.536, *t*-test). We determined how many short-styled plants share each non-LS site, and the number ranged from 1 to 18 (Fig. 4). As a control, non-SS sites (i.e. sites not present in short-styled plants) were also counted as above, and the number of long-styled plants sharing non-SS sites ranged from 1 to 10 (Fig. 4).

As shown in Fig. 4, a striking U-shaped distribution was obtained when the number of non-LS sites was plotted against the number of short-styled plants sharing the non-LS sites. In contrast, non-SS sites shared by >10 long-styled plants were not detected at all. Considering that there was no significant difference in the number of mapped reads obtained from short-styled and long-styled plants, we regarded the non-LS sites shared by >10 short-styled plants as '*S*-allele-specific sites'. In total, 88,031 of the *S*-allele-specific sites were detected and found to be located on the *S*-allelic region, consisting of 332 of scaffolds encompassing 5,393,196 bp (Supplementary Table S12). Of the 332 scaffolds, Fes_sc0003500.1 and Fes_sc0015090.1 harbour *S-ELF3* and *SSG2*, respectively. Since we previously established that *S-ELF3* and *SSG2* existed only in the genomes of short-styled plants,[16] this result shows the effectiveness of our procedure in discovering the *S*-allelic region in the buckwheat genome. However, it should be noted that we used only one restriction enzyme combination (*Eco*RI and *Mse*I) to obtain the GBS reads; we did not detect *S*-allelic scaffolds harbouring no recognition sites for these enzymes. Thus, the total length of the *S*-allelic region in the buckwheat genome obtained in the present study might be an underestimation.

In the *S*-allelic region, repeat sequences were abundant; the ratio of repeat length to the total length of the *S*-region was 71.43%, which is 1.4-times higher than that of the ratio of repeat length to the total scaffold length (51.69%, Supplementary Table S5). Gypsy elements were particularly abundant in the *S*-region; the Gypsy elements accounted for 12.15% of the *S*-region, which is 1.9 times higher than that of the repeat length in the total scaffold length (6.41%, Supplementary Table S5). Excluding TEs, only 32 predicted genes were successfully annotated by our database analyses in the 332 scaffolds (Supplementary Tables S12 and S13). Among the 32 predicted genes, two were candidates for heteromorphic SI related genes; Fes_sc0024869.1.g000002 and Fes_sc0006594.1.g000005 were predicted to encode proteins with similarity to a RING/U-box superfamily protein and an exoribonuclease 4, respectively. The RING/U-box protein is an E3 type ubiquitin ligase that functions in the 26S proteasome.[73] It has been suggested that degradation of cytotoxic S-RNases by the 26S proteasome in pollen tubes occurs during compatible pollination in a self-incompatible species, *Petunia inflata*.[74] Ongoing studies currently evaluating the expression pattern of these two genes and the effect of mutations in these two genes will clarify their roles in heteromorphic SI of buckwheat.

4. Conclusion and future perspective

The genome size of buckwheat is relatively large (~1.2 Gb), and some genomic regions are expected to be in the heterozygous state due to its outcrossing nature. These factors would reduce the lengths of our scaffolds in the buckwheat draft genome, which was assembled using only Illumina short reads. Though the draft genome was trancated and divided into a large number of scaffolds (387,594 scaffolds), we have successfully identified genes that control agronomically important traits using gene predictions and subsequent annotations in the BGDB. Furthermore, we also identified novel candidate genes involved in heteromorphic SI of buckwheat using the draft genome as a reference sequence for GBS mapping. Even if the scaffolds in a draft genome are truncated, they can nonetheless be used for database construction and as a reference sequence for NGS-based genetic markers. We are now preparing induced mutant pools of buckwheat using heavy-ion beams and chemicals such as ethyl methanesulfonate. Using NGS-based multi-dimensional screening,[75] mutants of the genes identified in this study will be rapidly identified from the pool and will be used to develop superior varieties of buckwheat.

5. Data availability

The Illumina reads used in this study are available from DDBJ/EMBL/NCBI under the accession numbers listed in Supplementary Table S3. The DRA accession number of the Illumina reads used in GBS analysis is DRA004489. The scaffold sequences are available under the accession numbers BCYN01000001-BCYN01387594 (387,594 entries). The draft genome sequence FES_r1.0, CDS and protein sequences, and annotation file (gff file) are also available from the Buckwheat Genome DataBase (BGDB; http://buckwheat.kazusa.or.jp).

Acknowledgements

We thank S. Tabata, T. Ota, K. Isono, and the anonymous reviewer for valuable suggestions and comments; R. Kajitani for kind information on genome assemblers; and K. L. Farquharson for language-editing support of the manuscript. We also thank Y. Nakajima for assistance in database construction and management. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Supplementary data

Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This study was supported by a grant from the Ministry of Agriculture, Forestry, and Fisheries of Japan (Genomics-based Technology for Agricultural Improvement, SFC3001) and JSPS KAKENHI (grant numbers 22580003, 25292009, 25450011, and 25660006).

References

1. Polashock, J., Zelzion, E., Fajardo, D., et al. 2014, The American cranberry: first insights into the whole genome of a species adapted to bog habitat, *BMC Plant Biol.*, **14**, 165.
2. Kim, S., Park, M., Yeom, S-I., et al. 2014, Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species, *Nat. Genet.*, **46**, 270–8.
3. Zhang, G., Liu, X., Quan, Z., et al. 2012, Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential, *Nat. Biotechnol.*, **30**, 549–54.
4. Varshney, R.K., Chen, W., Li, Y., et al. 2012, Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers, *Nat. Biotechnol.*, **30**, 83–9.
5. Giménez-Bastida, J.A. and Zieliński, H. 2015, Buckwheat as a functional food and its effects on health, *J. Agric. Food Chem.*, **63**, 7896–913.
6. Comino, I., de Lourdes Moreno, M., Real, A., Rodríguez-Herrera, A., Barro, F. and Sousa, C. 2013, The gluten-free diet: testing alternative cereals tolerated by celiac patients, *Nutrients*, **5**, 4250–68.
7. Heffler, E., Pizzimenti, S., Badiu, I., Guida, G. and Rolla, G. 2014, Buckwheat allergy: an emerging clinical problem in Europe, *J. Allergy Ther.*, **5**, 168.
8. Yasui, Y., Wang, Y., Ohnishi, O. and Campbell, C.G. 2004, Amplified fragment length polymorphism linkage analysis of common buckwheat (*Fagopyrum esculentum*) and its wild self-pollinated relative *Fagopyrum homotropicum*, *Genome*, **47**, 345–51.
9. Konishi, T. and Ohnishi, O. 2006, A linkage map for common buckwheat based on microsatellite and AFLP markers, *Fagopyrum*, **23**, 1–6.
10. Hara, T., Iwata, H., Okuno, K., Matsui, K. and Ohsawa, R. 2011, QTL analysis of photoperiod sensitivity in common buckwheat by using markers for expressed sequence tags and photoperiod-sensitivity candidate genes, *Breed. Sci.*, **61**, 394–404.
11. Yabe, S., Hara, T., Ueno, M., et al. 2014, Rapid genotyping with DNA micro-arrays for high-density linkage mapping and QTL mapping in common buckwheat (*Fagopyrum esculentum* Moench), *Breed. Sci.*, **64**, 291–9.
12. Elshire, R.J., Glaubitz, J.C., Sun, Q., et al. 2011, A robust, simple Genotyping-by-Sequencing (GBS) Approach for high diversity species, *PLoS One*, **6**, e19379.
13. He, J., Zhao, X., Laroche, A., Lu, Z-X., Liu, H. and Li, Z. 2014, Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding, *Front. Plant Sci.*, **5**, 484.
14. Yasui, Y., Mori, M., Matsumoto, D., Ohnishi, O., Campbell, C.G. and Ota, T. 2008, Construction of a BAC library for buckwheat genome research—an application to positional cloning of agriculturally valuable traits, *Genes Genet. Syst.*, **83**, 393–401.
15. Stanke, M. and Waack, S. 2003, Gene prediction with a hidden Markov model and a new intron submodel, *Bioinformatics*, **19**, ii215–25.
16. Yasui, Y., Mori, M., Aii, J., et al. 2012, *S-LOCUS EARLY FLOWERING 3* is exclusively present in the genomes of short-styled buckwheat plants that exhibit heteromorphic self-incompatibility, *PLoS One*, **7**, e31264.
17. Schmieder, R. and Edwards, R. 2011, Quality control and preprocessing of metagenomic datasets, *Bioinformatics*, **27**, 863–4.
18. Hirakawa, H., Shirasawa, K., Kosugi, S., et al. 2014, Dissection of the octoploid strawberry genome by deep sequencing of the genomes of *Fragaria* species, *DNA Res.*, **21**, 169–81.
19. Marçais, G. and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics*, **27**, 764–70.
20. Li, R., Zhu, H., Ruan, J., et al. 2010, De novo assembly of human genomes with massively parallel short read sequencing, *Genome Res.*, **20**, 265–72.
21. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W. 2010, Scaffolding pre-assembled contigs using SSPACE, *Bioinformatics*, **27**, 578–9.
22. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–10.
23. Price, A.L., Jones, N.C. and Pevzner, P.A. 2005, De novo identification of repeat families in large genomes, *Bioinformatics*, **21**(Suppl 1), i351–8.
24. Trapnell, C., Pachter, L. and Salzberg, S.L. 2009, TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, **25**, 1105–11.
25. Blanco, E., Parra, G. and Guigó, R. 2007, Using geneid to identify genes, *Curr. Protoc. Bioinformatics*, **Chapter 4**, Unit 4.3.
26. Quevillon, E., Silventoinen, V., Pillai, S., et al. 2005, InterProScan: protein domains identifier, *Nucleic Acids Res.*, **33**, W116–20.
27. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., et al. 2003, The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **4**, 41.
28. Llorens, C., Futami, R., Covelli, L., et al. 2011, The Gypsy Database (GyDB) of mobile genetic elements: release 2.0, *Nucleic Acids Res.*, **39**, D70–4.
29. Eddy, S.R. 2009, A new generation of homology search tools based on probabilistic inference, *Genome Inform.*, **23**, 205–11.
30. Lowe, T.M. and Eddy, S.R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res.*, **25**, 955–64.
31. Bolger, A.M., Lohse, M. and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, **30**, 2114–20.
32. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, **25**, 1754–60.
33. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, **25**, 2078–9.
34. DePristo, M.A., Banks, E., Poplin, R., et al. 2011, A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nat Genet.*, **43**, 491–8.
35. Larkin, M.A., Blackshields, G., Brown, N.P., et al. 2007, ClustalW and ClustalX version 2, *Bioinformatics*, **23**, 2947–8.
36. Saitou, N. and Nei, M. 1987, The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, **4**, 406–25.
37. Jones, D.T., Taylor, W.R. and Thornton, J.M. 1992, The rapid generation of mutation data matrices from protein sequences, *Comput. Appl. Biosci.*, **8**, 275–82.

38. Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. 2013, MEGA6: molecular evolutionary genetics analysis version 6.0, *Mol. Biol. Evol.*, **30**, 2725–9.
39. Nagano, M., Aii, J., Campbell, C., Kawasaki, S. and Adachi, T. 2000, Genome size analysis of the genus *Fagopyrum*, *Fagopyrum*, **17**, 35–9.
40. Nowak, M.D., Russo, G., Schlapbach, R., Huu, C.N., Lenhard, M. and Conti, E. 2015, The draft genome of *Primula veris* yields insights into the molecular basis of heterostyly, *Genome Biol.*, **16**, 12.
41. Kitashiba, H., Li, F., Hirakawa, H., et al. 2014, Draft sequences of the radish (*Raphanus sativus* L.) genome, *DNA Res.*, **21**, 481–90.
42. Cough, J.F., Naghski, J. and Krewson, C.F. 1946, Buckwheat as a source of rutin, *Science*, **15**, 197–8.
43. Li, X., Park, N.I., Xu, H., Woo, S-H., Park, C.H. and Park, S.U. 2010, Differential expression of flavonoid biosynthesis genes and accumulation of phenolic compounds in common buckwheat (*Fagopyrum esculentum*), *J. Agric. Food Chem.*, **58**, 12176–81.
44. Gupta, N., Sharma, S.K., Ranab, J.C. and Chauhan, R.S. 2011, Expression of flavonoid biosynthesis genes vis-à-vis rutin content variation in different growth stages of *Fagopyrum* species, *J. Plant Physiol.*, **168**, 2117–23.
45. Quattrocchio, F., Wing, J.F., van der Woude, K., Mol, J.N.M. and Koes, R. 1998, Analysis of bHLH and MYB domain proteins: species-specific regulatory differences are caused by divergent evolution of target anthocyanin genes, *Plant J.*, **13**, 475–88.
46. Walker, A.R., Davison, P.A., Bolognesi-Winfield, A.C., et al. 1999, The *TRANSPARENT TESTA GLABRA1* locus, which regulates trichome differentiation and anthocyanin biosynthesis in Arabidopsis, encodes a WD40 repeat protein, *Plant Cell*, **11**, 1337–49.
47. Baudry, A., Heim, M.A., Dubreucq, B., Caboche, M., Weisshaar, B. and Lepiniec, L. 2004, TT2, TT8, and TTG1 synergistically specify the expression of *BANYULS* and proanthocyanidin biosynthesis in *Arabidopsis thaliana*, *Plant J.*, **39**, 366–80.
48. Gonzalez, A., Zhao, M., Leavitt, J.M. and Lloyd, A.M. 2008, Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings, *Plant J.*, **53**, 814–27.
49. Martin, C. and Paz-Ares, J. 1997, MYB transcriptiofnactors in plant, *Trends Genet.*, **13**, 67–73.
50. Stracke, R., Werber, M. and Weisshaar, B. 2001, The *R2R3-MYB* gene family in *Arabidopsis thaliana*, *Curr. Opin. Plant Biol.*, **4**, 447–56.
51. Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C. and Lepiniec, L. 2010, MYB transcription factors in *Arabidopsis*, *Trends Plant Sci.*, **15**, 573–81.
52. Takos, A.M., Jaffé, F.W., Jacob, S.R., Bogs, J., Robinson, S.P. and Walker, A.R. 2006, Light-induced expression of a MYB gene regulates anthocyanin biosynthesis in red apples, *Plant Physiol.*, **142**, 1216–32.
53. Akagi, T., Ikegami, A., Tsujimoto, T., et al. 2009, DkMyb4 is a Myb transcription factor involved in proanthocyanidin biosynthesis in persimmon fruit, *Plant Physiol.*, **151**, 2028–45.
54. Nishihara, M., Yamada, E., Saito, M., Fujita, K., Takahashi, H. and Nakatsuka, T. 2014, Molecular characterization of mutations in white-flowered torenia plants, *BMC Plant Biol.*, **14**, 86.
55. Zhu, Z., Wang, H., Wang, Y., et al. 2015, Characterization of the *cis* elements in the proximal promoter regions of the anthocyanin pathway genes reveals a common regulatory logic that governs pathway regulation, *J. Exp. Bot.*, **66**, 3775–89.
56. Ban, Y., Honda, C., Hatsuyama, Y., Igarashi, M., Bessho, H. and Moriguchi, T. 2007, Isolation and functional analysis of a MYB transcription factor gene that is a key regulator for the development of red coloration in apple skin, *Plant Cell Physiol.*, **48**, 958–70.
57. Yi, J., Derynck, M.R., Li, X., Telmer, P., Marsolais, F. and Dhaubhadel, S. 2010, A single-repeat MYB transcription factor, GmMYB176, regulates *CHS8* gene expression and affects isoflavonoid biosynthesis in soybean, *Plant J.*, **62**, 1019–34.
58. Tanaka, K., Matsumoto, K., Akasawa, A., et al. 2002, Pepsin-resistant 16-kD buckwheat protein is associated with immediate hypersensitivity reaction in patients with buckwheat allergy, *Int. Arch. Allergy Immunol.*, **129**, 49–56.
59. Matsumoto, R., Fujino, K., Nagata, Y., et al. 2004, Molecular characterization of a 10-kDa buckwheat molecule reactive to allergic patients' IgE, *Allergy*, **59**, 533–8.
60. Satoh, R., Koyano, S., Takagi, K., Nakamura, R., Teshima, R. and Sawada, J. 2008, Immunological characterization and mutational analysis of the recombinant protein BWp16, a major allergen in buckwheat, *Biol. Pharm. Bull.*, **31**, 1079–85.
61. Sonnewald, U. and Kossmann, J. 2013, Starches—from current models to genetic engineering, *Plant Biotechnol. J.*, **11**, 223–32.
62. Copeland, L., Blazek, J., Salman, H. and Tang, M.C. 2009, Form and fuctionality of starch, *Food Hydrocoll.*, **23**, 1527–34.
63. Srichuwong, S., Sunarti, T.C., Mishima, T., Isono, N. and Hisamatsu, M. 2005, Starches from different botanical sources I: contribution of amylopectin fine structure to thermal properties and enzyme digestibility, *Carbohydr. Polym.*, **60**, 529–38.
64. Jobling, S. 2004, Improving starch for food and industrial applications, *Curr. Opin. Plant Biol.*, **7**, 210–8.
65. Tsai, C-Y. 1974, The function of the waxy locus in starch synthesis in maize endosperm, *Biochem. Genet.*, **11**, 83–96.
66. Cao, H., Imparl-Radosevich, J., Guan, H., Keeling, P.L., James, M.G. and Myers, A.M. 1999, Identification of the soluble starch synthase activities of maize endosperm, *Plant Physiol.*, **120**, 205–15.
67. Ball, S.G. and Morell, M.K. 2003, From bacterial glycogen to starch: understanding the biogenesis of the plant starch granule, *Annu. Rev. Plant Biol.*, **54**, 207–33.
68. Wang, X., Feng, B., Xu, Z., et al. 2014, Identification and characterization of granule bound starch synthase I (GBSSI) gene of tartary buckwheat (*Fagopyrum tataricum* Gaertn.), *Gene*, **534**, 229–35.
69. Cheng, J., Khan, M.A., Qiu, W-M., et al. 2012, Diversification of genes encoding granule-bound starch synthase in monocots and dicots is marked by multiple genome-wide duplication events, *PLoS One*, **7**, e30088.
70. Yamamori, M. and Quynh, N.T. 2000, Differential effects of Wx-A1, -B1 and -D1 protein deficiencies on apparent amylose content and starch pasting properties in common wheat, *Theor. Appl. Genet.*, **100**, 32–8.
71. Garber, R.J. and Quisenberry, K.S. 1927, The inheritance of length of style in buckwheat, *J. Agric. Res.*, **34**, 181–3.
72. Lewis, D. and Jones, D.A. 1992, The genetics of heterostyly. In: Barrett, S.C. H. (Ed.), *Evolution and function of heterostyly*. Springer-Verlag: Berlin, pp. 129–50.
73. Vierstra, R.D. 2003, The ubiquitin/26S proteasome pathway, the complex last chapter in the life of many plant proteins, *Trends Plant Sci.*, **8**, 135–42.
74. Hua, Z. and Kao, T. 2006, Identification and characterization of components of a putative *Petunia S*-Locus F-box–containing E3 ligase complex involved in S-RNase–based self-incompatibility, *Plant Cell*, **18**, 2531–53.
75. Tsai, H., Howell, T., Nitcher, R., et al. 2011, Discovery of rare mutations in populations: TILLING by sequencing, *Plant Physiol.*, **156**, 1257–68.