Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 358127, 8 pages
http://dx.doi.org/10.1155/2015/358127

Research Article

The *Glycine max* cv. Enrei Genome for Improvement of Japanese Soybean Cultivars

Michihiko Shimomura,[1] Hiroyuki Kanamori,[2] Setsuko Komatsu,[3] Nobukazu Namiki,[1] Yoshiyuki Mukai,[2] Kanako Kurita,[2] Kaori Kamatsuki,[1] Hiroshi Ikawa,[1] Ryoichi Yano,[4] Masao Ishimoto,[5] Akito Kaga,[5] and Yuichi Katayose[2]

[1] *Mitsubishi Space Software Co., Ltd., Takezono, Tsukuba, Ibaraki 305-0032, Japan*
[2] *National Institute of Agrobiological Sciences, Owashi, Tsukuba, Ibaraki 305-8634, Japan*
[3] *NARO Institute of Crop Sciences, Kannondai, Tsukuba, Ibaraki 305-8518, Japan*
[4] *University of Tsukuba, Tennodai, Tsukuba, Ibaraki 305-0006, Japan*
[5] *National Institute of Agrobiological Sciences, Kannondai, Tsukuba, Ibaraki 305-8602, Japan*

Correspondence should be addressed to Yuichi Katayose; katayose@affrc.go.jp

Received 30 March 2015; Accepted 31 May 2015

Academic Editor: Graziano Pesole

We elucidated the genome sequence of *Glycine max* cv. Enrei to provide a reference for characterization of Japanese domestic soybean cultivars. The whole genome sequence obtained using a next-generation sequencer was used for reference mapping into the current genome assembly of *G. max* cv. Williams 82 obtained by the Soybean Genome Sequencing Consortium in the USA. After sequencing and assembling the whole genome shotgun reads, we obtained a data set with about 928 Mbs total bases and 60,838 gene models. Phylogenetic analysis provided glimpses into the ancestral relationships of both cultivars and their divergence from the complex that include the wild relatives of soybean. The gene models were analyzed in relation to traits associated with anthocyanin and flavonoid biosynthesis and an overall profile of the proteome. The sequence data are made available in DAIZUbase in order to provide a comprehensive informatics resource for comparative genomics of a wide range of soybean cultivars in Japan and a reference tool for improvement of soybean cultivars worldwide.

1. Introduction

Soybean (*Glycine max*) is one of the world's most important leguminous crops being a major source of edible proteins and vegetable oils. In terms of global production, soybean ranks fourth, following the major cereal crops such as rice, wheat, and corn. It is also a major source of nutritionally and physiologically active substances such as saponins, isoflavones, phytosterols, and tocopherols. Consumption of soybeans as food is largely concentrated in Asia. Soybean has been a part of the Japanese diet and eaten from ancient times as a valuable source of traditional fermented products such as miso, soy sauce, and natto and nonfermented products such as edamame (boiled soybean), kinako (toasted soybean flour), tofu, and soymilk. As in other major crops, the main targets of soybean breeding in Japan are high yield, high quality (absence of seed coat cracking, seed size, hilum color, uniformity of seed size, and food processing adaptability) to compete with imported soybean, and resistance to biotic/abiotic stress for stable production. Additionally, the chemical component of seeds including high protein content, modification of storage proteins, absence of lipoxygenases and saponin, high isoflavone content, and high sucrose have been given much consideration in many soybean breeding programs [1].

The domesticated soybean has its origin from *Glycine soja,* a wild soybean species found mainly in northern China, Japan, Korea, and the eastern part of Russia [2]. Archaeological studies indicate that the word for soybean first appeared in China about 3,700 years ago in bone inscriptions dating back to the Yin and Shang dynasties and carbonized soybean seeds found about 2,600 years ago [2]. Estimation of archaeological records indicates a widespread early association of small seeded soybean to be as old as 9,000–8,600 calibrated years before the present (cal BP) in northern China and 7,000 cal BP in Japan [3]. Direct radiocarbon dates on charred soybean seeds indicate selection resulted in large seed sizes in Japan by 5,000 cal BP (Middle Jomon) and in Korea by 3,000 cal BP (Early Mumun) [3]. Extensive genome analysis also indicates that the *G. soja*/*G. max* complex diverged from the most recent common ancestor at 0.27 Mya [4] or 0.8 Mya [5]. In a more recent study, the genetic variation and population structure among 1,603 soybean accessions indicated a clear genetic differentiation among Japanese soybean landraces, exotic and cultivated soybeans, and wild soybeans [6].

From the genomics point of view, soybean has been used as a model plant for comparative studies of legumes in terms of root nodulation, oilseed production, and secondary metabolism. It is also a valuable material for genome research because of the availability of many genomic and germplasm resources. In 2010 a great deal of effort in the USA culminated with the sequencing of the paleopolyploid soybean genome based on a soybean cultivar Williams 82 [7]. This cultivar was derived from backcrossing a *Phytophthora* root rot resistance locus from the donor parent Kingwa which was selected in 1921 from the cultivar Peking introduced from Beijing, China, in 1906 [8].

In Japan, however, domestic cultivars have been developed to suit a variety of conditions and applications of specific importance to Japanese growers. Although the *G. max* cv. Williams 82 reference soybean genome sequence could be useful in understanding the diversity among many cultivars, it is necessary to have genomic resources that could be directly applied to Japanese soybean cultivation. The Japanese soybean cultivar Enrei was derived from cultivars Norin number 2 and Higashiyama number 6 (also known as cv. Shiromeyutaka) and was developed in 1971 at Kikyogahara Branch of the Nagano Agricultural Experiment Station (presently known as Nagano Vegetable and Ornamental Crops Experiment Station) [9]. In this paper, we described the analysis of the genome sequence of the Japanese soybean cultivar Enrei focusing on the phylogenetic analysis and major traits for soybean breeding including anthocyanin and flavonoid biosynthesis and proteome profile.

2. Materials and Methods

2.1. Genome Sequencing. The plant material was provided by the Genebank of the National Institute of Agrobiological Sciences (NIAS). High-quality nuclear DNA with reduced organellar DNA was extracted from young leaves using a protocol designed for BAC DNA extraction with some modifications [10]. All sequencing reads were obtained using the Illumina HiSeq2000 at Operon Biotechnologies, Inc. (Eurofins Genomics). Standard short-read libraries and mate-paired libraries with 8 kbp insertion were built using the TruSeq SBS v5 for sequencing runs at 2 × 100 bp or 200 bp total. After sequencing, HiSeq Control Software v.1.4.8 and CASAVA 1.8.1 (Illumina) were utilized for base calling. Single-ended libraries and 3 kbp pair-ended libraries constructed with the GS FLX Titanium General Library Preparation Kit and Rapid Library Preparation Kit (Roche) were sequenced on Roche 454 FLX Titanium at the NIAS, and base calling was performed using the 454 FLX Titanium base caller.

2.2. Assembly and Reference Mapping. We constructed a *de novo* genome assembly (*G. max*_Enrei1) and reference genome assembly (*G. max*_Enrei2) to facilitate comprehensive analysis of the genome. The *G. max*_Enrei1 assembly was constructed from the Roche 454 FLX Titanium single-ended reads and pair-ended reads with 3 kbp insert, the Illumina HiSeq2000 pair-ended reads with 300 bp insert and mate-pair reads with 8 kbp insert, and the ABI 3730xl BAC-end reads using the Roche Newbler 2.7.

The *G. max*_Enrei2 assembly was derived from Roche 454 FLX Titanium single-ended reads and Illumina HiSeq2000 pair-ended reads were used for reference mapping with the BWA 0.7.5a (Li H. Aligning sequence reads, clone sequences, and assembly contigs with BWA-MEM, 2013; http://bio-bwa.sourceforge.net/), SAMtools 0.1.19 [11], and NIG script (NGS Surfer's wiki, http://cell-innovation.nig.ac.jp/wiki/tiki-index.php?page=samtools#mpileup_). The *G. max* cv. Williams 82, also referred to as Gmax275 genome assembly, was used for reference mapping. The pseudomolecules and scaffolds in the *G. max*_Enrei2 were searched for marker sequences by BLASTn (NCBI BLAST, ftp://ftp.ncbi.nih.gov/blast/). Then marker sequences were mapped in the regions with clear sequences, gap regions (indicated as N's in the sequence), BAC-end sequences (BES) hit position, marker hit position, and scaffold derived from *de novo* assembly hit position. Subsequently, the cutting points were identified to reconstruct the pseudomolecules and scaffolds.

2.3. Gene Models. The soybean parameter files were built from chromosome 16 region of hard masked Gmax275 genome [30,000,000–37,887,014 bps] using Augustus program [12]. The transposable elements in the scaffolds and pseudomolecules of the *G. max*_Enrei2 genome assembly were masked using RepeatMasker (Smit AFA, Hubley R., and Green P. RepeatMasker Open-3.0, 1996–2010, http://www.repeatmasker.org/) and the gene models were built using Augustus. These gene models were used as queries in BLASTn search using the soyTE as a database [13]. The filtered gene models with bit score of 100 and above were selected. Additionally, we used available RNAseq data (PRJDB3582) assembled by Trinity version 2014-07-17 [14]. A total of 172,753 gene models were extracted. For each gene, the longest ORF was identified using EMBOSS getorf [15].

2.4. Phylogenetic Analysis. The amino acid sequences of the gene models for *Arabidopsis thaliana* [16], *Arabidopsis lyrata* [17], *Medicago truncatula* [18], and *Oryza sativa* (annotation data on Os-Nipponbare-Reference-IRGSP-1.0, http://rapdb.dna.affrc.go.jp/download/archive/irgsp1/IRGSP-1.0_protein_2014-06-25.fasta.gz) were obtained. These were combined with the corresponding sequences of Gmax275 and *G. max*_Enrei2 for clustering with OrthoMCL v2.0.7 [19]. After removal of incomplete gene models, a set of single copy genes (Supplementary Table S1 in Supplementary Material

Table 1: Genome assembly and annotation of *G. max* cv. Enrei.

Reference mapping length	With gaps [bp]	Without gaps [bp]	Ratio
Chromosome	946,877,581	904,901,085	95.6
Scaffold	31,116,190	22,803,649	73.3
Total	977,993,771	927,704,734	94.9
Gene models			
Number of gene models	60,838		
Mean coding sequence length	1455.3 [bp]		
Mean number of exons per gene	4.5		
Mean exon length	323.4 [bp]		

available online at http://dx.doi.org/10.1155/2015/358127) was built from gene models that completely matched gene models in the genome and gene models derived from RNAseq. Then the fourfold degenerative sites derived from the refined single copy gene set were aligned using Clustal Omega 1.2.0 [20]. A guided tree was built by MEGA 6.06 [21] and the phylogenetic tree was constructed using PAML 4.8a [22], Multidivtime [23], and FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

2.5. Anthocyanin and Flavonoid Biosynthesis. All gene models in Gmax275 and *G. max*_Enrei2 associated with anthocyanin and flavonoid biosynthesis were extracted and clustered using OrthoMCL [19]. Then these gene models were associated by BLASTn.

2.6. Proteome Analysis. The proteome analysis of Enrei cultivar was performed using seeds. The cotyledons from ten seeds were grounded in liquid nitrogen and purified by phase separation using standard procedures [24]. The purified proteins were digested with trypsin. For mass spectrometry analysis, the eluted peptides were analyzed on a nanospray LTQ XL Orbitrap mass spectrometer and the MS spectra were used for protein identification. Identification of proteins was performed using the Mascot search engine version 2.4.1 (Matrix Science, London, UK) and Proteome Discoverer software version 1.4.0.288 (Thermo Fisher Scientific) against 54,175 soybean peptide sequences [7]. Mascot results were filtered with Mascot Percolator software to improve the accuracy and sensitivity of the peptide identification [25]. The protein abundance was analyzed using emPAI value as described in Shinoda et al. [26]. Furthermore, the protein gene models derived from Gmax275 and *G. max*_Enrei2 genome assemblies were associated using the clustered data obtained from OrthoMCL [19]. Additionally, these gene models were associated by BLASTn.

3. Results and Discussion

3.1. Genome Sequencing and Reference Mapping. The whole genome sequence of the Japanese soybean cultivar Enrei was assembled using a total of 22.2X coverage (Table 1). Reference mapping into the Gmax275 genome assembly was performed using DNA markers in the genetic linkage map of cultivar Enrei (Supplementary Table S2) and the *de novo* genome assembly (accession numbers BBNX01000001–BBNX01092182) (Supplementary Table S3). The ratio of *G. max*_Enrei2 to Gmax275 genome length (978,495,272 bps) [7] was 99.95%. The quality of the genome was evaluated by mapping marker sequences, BES, and the 56,264 gene models without alternative splicing in the Gmax275 genome [7]. As a result a total of 56,043 (99.6%) gene models in Gmax275 were represented in the Enrei genome. Additionally, a total of 1,860 marker sequences (Kaga et al., unpublished data) were mapped with a ratio of 98.8%. A total of 87 markers were unmatched in terms of linkage order and physical position (Supplementary Table S4). Additionally, a total of 92,451 BES pairs were mapped with a ratio of 76.3% (Supplementary Table S5). The Enrei genome sequence is deposited at the DNA Data Bank of Japan (DDBJ) under accession numbers BBNX02000001–BBNX02108601.

3.2. Gene Models. In total, 60,838 Enrei gene models were predicted (Table 1). Comparison with the gene models of Gmax275 [7] showed a longer mean coding sequence length (1,455 bps in Enrei and 1,168 bps in Gmax275) and a longer mean exon length (323 bps in Enrei and 231 bps in Gmax275). To complement the annotation of the genome sequence, gene models were mapped to the longest 172,753 open reading frame (ORF) sequences from the mRNAs of young leaves (Supplementary Table S6). In total, 11 gene models had no ORF hit sequences, 20,542 gene models had more than 50% coverage, 5,950 gene models had more than 90% coverage, and 2,269 gene models had 100% coverage.

As mean coding sequence length was 1,168.1 bps, mean number of exons per gene 5.0, and mean exon length 231.5 bps of 56,044 Gmax275 gene models without variant, Enrei number of exons per gene was shorter and CDS was longer. The difference in the gene models between Gmax275 and *G. max*_Enrei2 may be attributed to SNPs between the two cultivars as well as several parameters used in building the gene models.

3.3. SNPs and INDELs. A total of 1,659,041 SNPs and 344,418 insertions and deletions (INDELs) were identified between the *G. max*_Enrei2 and Gmax275 genome assemblies (Supplementary Table S7). Both SNPs and INDELs were largely predominant in chromosome 18 and relatively less predominant in chromosome 11. The average distance between SNPs was calculated at 589.8 bp/SNP against the Gmax275 genome assembly. The minimum average distance was 320.8 bp/SNP on chromosome 18 and maximum average

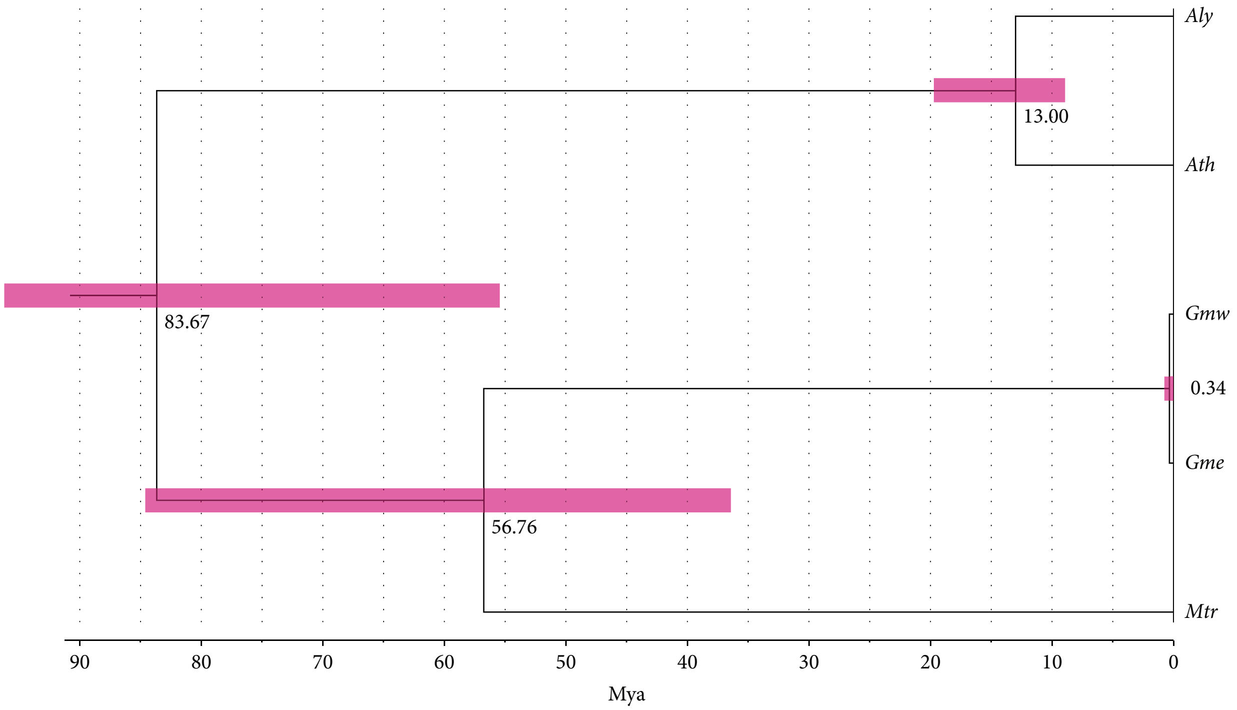


Figure 1: Phylogenetic tree of *G. max* cv. Williams 82 (*Gmw*), *G. max* cv. Enrei (*Gme*), *A. thaliana* (*Ath*), *A. lyrata* (*Aly*), and *M. truncatula* (*Mtr*). The pink bar represents the 95% probability density. Mya represents a unit in million years.

distance was 984.9 bp/SNP on chromosome 5. Moreover, the total INDELs in Enrei represent approximately 50.79 Mbp of the genome sequence. These values however merely represent an overview of the differences in genome structure of the two cultivars. An accurate analysis of the SNPs and INDELs can be obtained only from a high-quality genome sequence of both cultivars.

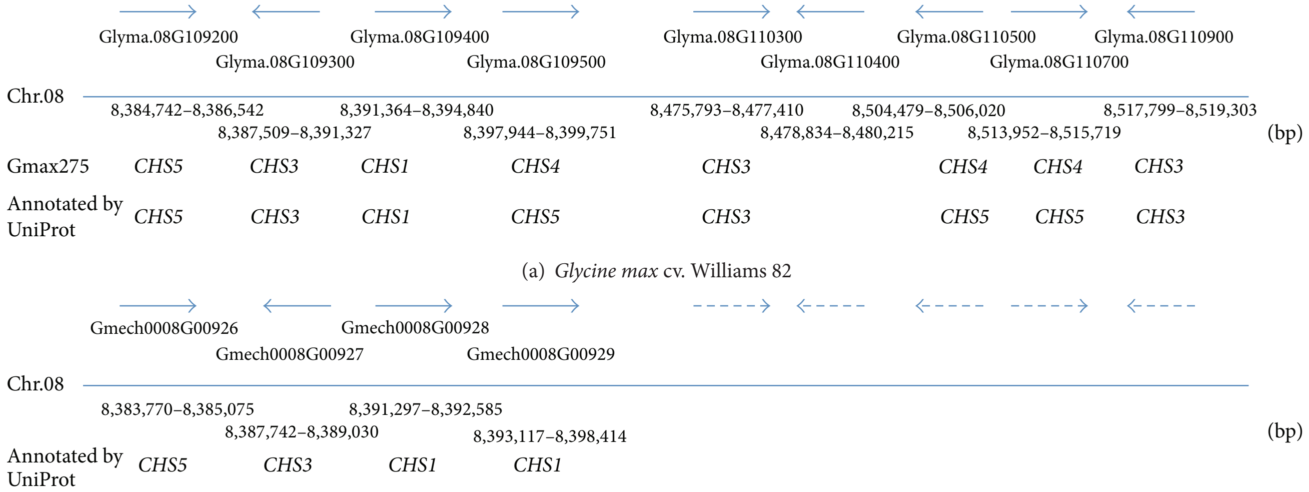
3.4. Phylogenetic Analysis. We applied OrthoMCL [19] to the clustered and aligned gene models of *A. thaliana* [16], *A. lyrata* [17], *G. max* cv. Williams 82 [7], *G. max* cv. Enrei, *M. truncatula* [18], and *O. sativa* (annotation data on Os-Nipponbare-Reference-IRGSP-1.0.). A set of filtered single copy genes was selected to calculate the phylogenetic relationships and divergence time among these species (Supplementary Table S1). Based on the phylogenetic divergence of *A. thaliana* which occurred about 13 Mya [27], the divergence between *G. max* cv. Williams 82 and cv. Enrei was estimated at 0.34 Mya (95PD: 0.78–0.10 Mya), much later than the calculated divergence time between *A. thaliana* and *A. lyrata* (95PD: 19.30–8.52 Mya; Figure 1). The divergence between the *Glycine* clade and *M. truncatula* was estimated to have occurred around 56.76 Mya (95PD: 84.54–36.99 Mya). On the other hand, the divergence between the *G. max*/*M. truncatula* clade and the *Arabidopsis* clade must have occurred much earlier around 83.67 Mya (95PD: 122.51–55.57 Mya). Previous studies have shown a divergence time of 54 Mya between *M. truncatula* and the *Glycine* clade [28]. A whole genome duplication (WGD) which occurred around 58 Mya had been a major factor in shaping the *M. truncatula* genome [18].

The complex of *G. max* and its wild relative, *G. soja,* diverged from the most recent common ancestor around 0.27 Mya [4] or 0.8 Mya [5]. Assuming that *G. max* diverged from *G. soja* at around 0.8 Mya, the divergence of the branch for both Williams 82 and Enrei at around 0.34 Mya was much later than previously estimated. As Li et al. [5] pointed out, divergent selection may have contributed to the differentiation of *G. soja* and *G. max* before domestication of *G. max*. The divergent selection as adaptation to different environments must have contributed to the differentiation of both cv. Williams 82 and cv. Enrei from the most recent common ancestor.

3.5. Anthocyanin and Flavonoid Biosynthesis. In soybeans, several chalcone synthesis genes, namely, *CHS3* (P19168), *CHS1* (P24826), *CHS7* (P30081), *CHS4* (Q6X0N0), and *CHS8* (AY237728), are associated with seed coat pigmentation [29]. The physical position of these *CHS* genes was determined using BAC assembly [30, 31] for loci associated with RNA silencing and WGS assembly [7]. The corresponding genes in the Gmax275 genome assembly are as follows: *CHS1* (Glyma.08G109400), *CHS2* (Glyma.05G153200), *CHS3* (Glyma.08G110300 and Glyma.08G110900), *CHS4* (Glyma.08G110500 and Glyma.08G110700), *CHS5* (Glyma.08G109200, Glyma.08G109300, and Glyma.08G110400), *CHS6* (Glyma.09G075200), *CHS7* (Glyma.01G228700), *CHS8* (Glyma.11G011500), and *CHS9* (Glyma.08G109500). Most of the genes in the pathway were commonly represented in both cultivars (Figure 2).

Intermediate	Enzyme	Chr. number	Williams 82 gene name (Gmax275)	Enrei gene name
L-Phenylalanine				
	PAL	10	Glyma.10G209800	Gmech0010G03487
	PAL	20	Glyma.20G180800	Gmech0020G03393
Cinnamic acid				
	4CL	1	Glyma.01G232400	Gmech0001G04494
	4CL	7	Glyma.07G112700	
	4CL	11	Glyma.11G010500, Glyma.11G091600	Gmech0011G00081, Gmech0011G00767
	4CL	13	Glyma.13G095600, Glyma.13G372000	Gmech0013G01703, Gmech0013G04132
	4CL	15	Glyma.15G001700	Gmech0015G00022
	4CL	17	Glyma.17G064400, Glyma.17G064500, Glyma.17G064600	Gmech0017G00575, Gmech0017G00576, Gmech0017G00577
4-Coumaric acid				
	C4H	2	Glyma.02G236500	Gmech0002G03627
	C4H	10	Glyma.10G275600	Gmech0010G04048
	C4H	14	Glyma.14G205200	Gmech0014G03800
	C4H	20	Glyma.20G114200	Gmech0020G02809
4-Coumaroyl-CoA				
	CHS	1	Glyma.01G073600, Glyma.01G091400, Glyma.01G228700	Gmech0001G01135, Gmech0001G02138, Gmech0001G04461
	CHS	2	Glyma.02G130400	Gmech0002G01230
	CHS	5	Glyma.05G153100, Glyma.05G153200	Gmech0005G02640
	CHS	6	Glyma.06G118500, Glyma.06G118600	Gmech0006G00999
	CHS	8	Glyma.08G109200, Glyma.08G109300, Glyma.08G109400, Glyma.08G109500	Gmech0008G00926, Gmech0008G00927, Gmech0008G00928, Gmech0008G00929
	CHS	8	Glyma.08G110300, Glyma.08G110400, Glyma.08G110500, Glyma.08G110700, Glyma.08G110900	
	CHS	9	Glyma.09G074900, Glyma.09G075200	Gmech0009G00747, Gmech0009G00751
	CHS	11	Glyma.11G011500, Glyma.11G097900	Gmech0011G00090, Gmech0011G00823
	CHS	12	Glyma.12G023800	Gmech0012G00205
	CHS	13	Glyma.13G034300	Gmech0013G00877
	CHS	19	Glyma.19G105100	Gmech0019G02509
Chalcone				
	CHI	1	Glyma.01G166300	Gmech0001G03908
	CHI	2	Glyma.02G048700	Gmech0002G00430
	CHI	3	Glyma.03G154600	Gmech0003G02851
	CHI	4	Glyma.04G222400	Gmech0004G03876
	CHI	6	Glyma.06G143000	Gmech0006G01203
	CHI	10	Glyma.10G292200	Gmech0010G04176
	CHI	11	Glyma.11G077200	Gmech0011G00642
	CHI	13	Glyma.13G262500	Gmech0013G03207
	CHI	14	Glyma.14G098100	
	CHI	15	Glyma.15G242900	
	CHI	16	Glyma.16G128800	Gmech0016G02252
	CHI	17	Glyma.17G226600	Gmech0017G03319
	CHI	19	Glyma.19G156900	
	CHI	20	Glyma.20G241500, Glyma.20G241600, Glyma.20G241700	Gmech0020G03913, Gmech0020G03914, Gmech0020G03915
Flavanone				
	F3H	1	Glyma.01G166200	Gmech0001G03907
	F3H	2	Glyma.02G048400	Gmech0002G00428
	F3H	2	Glyma.02G048600	Gmech0002G00429
	F3H	16	Glyma.16G128700	Gmech0016G02250
Dihydroflavonol			→	Flavanol
	FLS	5	Glyma.05G088100, Glyma.06G110600	Gmech0005G00932, Gmech0006G00935
	FLS	13	Glyma.13G082300	Gmech0013G01540
	FLS	14	Glyma.14G163300	
	DFR	2	Glyma.02G158700	
	DFR	13	Glyma.13G203800	Gmech0013G02695
	DFR	13	Glyma.13G355600	Gmech0013G03984
	DFR	14	Glyma.14G072700	
	DFR	14	Glyma.14G072800	
	DFR	14	Glyma.14G072900	
	DFR	15	Glyma.15G018500	
	DFR	17	Glyma.17G173200	Gmech0017G01714
	DFR	17	Glyma.17G252200	Gmech0017G03547
	DFR	17	Glyma.17G252300	
Flavan-3,4-diol				
	ANS	1	Glyma.01G214200	Gmech0001G04339
	ANS	11	Glyma.11G027700	Gmech0011G00224
Anthocyanidin				

Figure 2: Enzymes involved in the major pathway for anthocyanin and flavonoid biosynthesis and the corresponding genes in Gmax275 and *G. max* cv. Enrei. *PAL* (phenylalanine ammonia-lyase), *4CL* (4-coumaroyl-CoA-ligase), *C4H* (cinnamate-4-hydroxylase), *CHS* (chalcone synthase), *CHI* (chalcone isomerase), *F3H* (flavanone 3-hydroxylase), *FLS* (flavonol synthase), *DFR* (dihydroflavonol 4-reductase), and *ANS* (anthocyanidin synthase).



(a) *Glycine max* cv. Williams 82

(b) *Glycine max* cv. Enrei

FIGURE 3: A region in soybean chromosome 8 showing the position of *CHS* gene clusters. The region encompassing 8.3~8.5 Mb of soybean chromosome 8 of Williams 82 Gmax275 (a) and Enrei cultivar (b) is characterized by *CHS* gene clusters. Most of the *CHS* genes correspond in both cultivars as indicated by the position and UniProt annotation of identified genes. However many *CHS* genes could not be localized in Enrei cultivar due to fragmented sequence.

However, one 4CL gene in chromosome 7, 5 *CHS* genes in chromosome 8, 3 CHI genes in 14, 15, and 19, respectively, 1 FLS gene in chromosome 14, and 6 DFR genes on chromosomes 2, 14, 15, and 17 were not found in the Enrei genome. Most of the *CHS* genes correspond in both cultivars as indicated by the position and UniProt annotation of identified genes except for those genes that could not be localized in the Enrei cultivar due to fragmented sequence (Figure 3). As most of these genes are involved in pigmentation of seed and hilum color in soybean, the absence of these genes in the Enrei genome in relation to anthocyanin and flavonoid biosynthesis pathway may be associated with gene silencing due to siRNA activity [32, 33].

3.6. Protein. Using gene models associated with seed proteome data, a total of 164 protein gene models corresponding to storage proteins, lipid synthesis/degradation enzymes, sorting/folding-related proteins, late embryogenesis abundant (LEA) protein, glycolysis pathway enzymes, protease/protease inhibitors, and others were identified (Supplementary Table S8). The protein content of dry seeds was 35–42% [34, 35] of the dry weight, and 70% of protein consists of 7S and 11S globulins [34, 36], which are part of the cupin superfamily (http://www.ebi.ac.uk/interpro/entry/IPR006045), corresponding to beta-conglycinin and glycinin, respectively [37]. To identify the proteins associated with grain filling of soybean seeds, we conducted a proteome analysis of the cotyledon. A total of 160 protein gene models in Gmax189 correspond to Enrei protein gene models ranging from 7.87 mol% to 0.03 mol% (Supplementary Table S8). Most of these proteins are storage proteins and cupin including beta-conglycinin and glycinin representing about 42% of total mol% and about 55% of total weight (sum of mass ∗ mol) (Table 2). Genes controlling the content of seed storage proteins were also highly represented [38, 39]. Genes associated with lipid metabolism such as lipoxygenase 1, peroxygenase 2, and oleosin family protein genes [40]; gene associated with sorting/folding-related protein such as HSP20-like chaperone, PDI-like, SNF7 family, and vacuolar sorting receptor proteins; and LEA protein genes which may be important in protecting other proteins from aggregations were highly represented in the Enrei genome. In addition, some genes involved in glycolysis pathway, enzymes, and proteinase/protease inhibitors, which may play an important role in germination stage, were also found. This proteome profile may provide the basis for understanding cultivar diversity and adaptation to cultivation condition.

TABLE 2: Composition of storage proteins in *G. max* cv. Enrei.

Chromosome	Related number of gene	Weight % (mass ∗ mol)	mol %
Chr10	6	19.8	14.27
Chr20	4	15.7	14.58
Chr03	1	6.6	4.31
Chr13	1	4.6	3.06
Chr19	1	2.8	1.88
Chr04	1	2.4	1.99
Chr02	1	2.1	1.36
Chr11	1	1.2	0.85
Chr01	1	0.1	0.07
Total	17	55.4	42.4

3.7. Enrei Genome Database. All sequencing data can be accessed in DAIZUbase (http://daizu.dna.affrc.go.jp/enrei/), an informatics resource for soybean genomics focusing on the Japanese soybean cultivar Enrei. The database is provided with a GBrowse [41] with interactive pages for displaying the Enrei genome sequence as well all aligned Enrei BAC clones and accompanying annotations. DAIZUbase also includes a unified map, which indicates the relationship between the linkage map and the physical map of the Enrei cultivar.

4. Conclusion

The genome sequence of the Japanese cultivar Enrei will provide valuable information for improvement of soybean cultivars adapted to domestic cultivation. The genome sequence will complement emerging strategies for effective soybean breeding through analysis of the genome structure of Japanese (domestic) soybean, development of DNA markers serving as landmarks of agronomically important traits, development of research resources for the identification of important genes in soybean, and isolation of genes controlling important traits such as disease and pest resistance, productivity, and regional adaptability. Detailed knowledge of the genes controlling specific traits will allow for more efficient soybean improvement enabling researchers to develop plant types adaptable to various environmental conditions.

Conflict of Interests

The authors have declared that no conflict of interests exists.

Authors' Contribution

Michihiko Shimomura and Yuichi Katayose contributed equally to this work.

Acknowledgments

The authors thank Professor Takuji Sasaki for technical discussions, Dr. Kyuya Harada for helpful discussion on domestic soybean, and Ms. Michie Shibata for technical assistance in assembling the genome. This work was supported by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Genomics for Agricultural Innovation, DD-1020, Soy1001).

References

[1] M. Hajika, "Present state and prospect of soybean production and soybean breeding in Japan," in *Proceedings of the 14th NIAS International Workshop on Genetic Resources and Comparative Genomics of Legumes (Glycine and Vigna)*, N. Tomooka and D. A. Vaughan, Eds., pp. 49–52, National Institute of Agrobiological Sciences, 2011.

[2] L. J. Qiu and R. Z. Chang, "The origin and history of *Soybean*," in *The Soybean: Botany, Production and Uses*, G. Singh, Ed., pp. 1–23, CABI Publishing, Bengaluru, India, 2010.

[3] G. A. Lee, G. W. Crawford, L. Liu, Y. Sasaki, and X. Chen, "Archaeological soybean (*Glycine max*) in East Asia: does size matter?" *PLoS ONE*, vol. 6, no. 11, Article ID e26720, 2011.

[4] M. Y. Kim, S. Lee, K. Van et al., "Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 51, pp. 22032–22037, 2010.

[5] Y. H. Li, G. Zhou, J. Ma et al., "*De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits," *Nature Biotechnology*, vol. 32, no. 10, pp. 1045–1052, 2014.

[6] A. Kaga, T. Shimizu, S. Watanabe et al., "Evaluation of soybean germplasm conserved in NIAS genebank and development of mini core collections," *Breeding Science*, vol. 61, no. 5, pp. 566–592, 2011.

[7] J. Schmutz, S. B. Cannon, J. Schlueter et al., "Genome sequence of the palaeopolyploid soybean," *Nature*, vol. 463, no. 7278, pp. 178–183, 2010.

[8] R. L. Bernard and C. R. Cremeens, "Registration of 'Williams 82' soybean," *Crop Science*, vol. 28, no. 6, pp. 1027–1028, 1988.

[9] National Institute of Agrobiological Sciences, *Plant Genetic Resources Used for Food and Agriculture in Japan*, National Institute of Agrobiological Sciences, 2010.

[10] D. G. Peterson, J. P. Tomkins, D. A. Frisch, R. A. Wing, and A. H. Paterson, "Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide," *Journal of Agricultural Genomics*, vol. 5, pp. 1–100, 2000.

[11] H. Li, B. Handsaker, A. Wysoker et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[12] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern, "AUGUSTUS: ab initio prediction of alternative transcripts," *Nucleic Acids Research*, vol. 34, pp. W435–W439, 2006.

[13] J. Du, D. Grant, Z. Tian et al., "SoyTEdb: a comprehensive database of transposable elements in the soybean genome," *BMC Genomics*, vol. 11, no. 1, article 113, 2010.

[14] M. G. Grabherr, B. J. Haas, M. Yassour et al., "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology*, vol. 29, no. 7, pp. 644–652, 2011.

[15] P. Rice, L. Longden, and A. Bleasby, "EMBOSS: the european molecular biology open software suite," *Trends in Genetics*, vol. 16, no. 6, pp. 276–277, 2000.

[16] P. Lamesch, T. Z. Berardini, D. Li et al., "The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools," *Nucleic Acids Research*, vol. 40, no. 1, pp. D1202–D1210, 2012.

[17] T. T. Hu, P. Pattyn, E. G. Bakker et al., "The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change," *Nature Genetics*, vol. 43, no. 5, pp. 476–481, 2011.

[18] N. D. Young, F. Debellé, G. E. D. Oldroyd et al., "The *Medicago* genome provides insight into the evolution of rhizobial symbioses," *Nature*, vol. 480, no. 7378, pp. 520–524, 2011.

[19] L. Li, C. J. Stoeckert Jr., and D. S. Roos, "OrthoMCL: identification of ortholog groups for eukaryotic genomes," *Genome Research*, vol. 13, no. 9, pp. 2178–2189, 2003.

[20] F. Sievers, A. Wilm, D. Dineen et al., "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Molecular Systems Biology*, vol. 7, article 539, 2011.

[21] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "MEGA6: molecular evolutionary genetics analysis version 6.0," *Molecular Biology and Evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.

[22] Z. Yang, "PAML 4: phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, 2007.

[23] J. L. Thorne and H. Kishino, “Divergence time and evolutionary rate estimation with multilocus data,” *Systematic Biology*, vol. 51, no. 5, pp. 689–702, 2002.

[24] S. Komatsu, C. Han, Y. Nanjo et al., “Label-free quantitative proteomic analysis of abscisic acid effect in early-stage soybean under flooding,” *Journal of Proteome Research*, vol. 12, no. 11, pp. 4769–4784, 2013.

[25] M. Brosch, L. Yu, T. Hubbard, and J. Choudhary, “Accurate and sensitive peptide identification with mascot percolator,” *Journal of Proteome Research*, vol. 8, no. 6, pp. 3176–3181, 2009.

[26] K. Shinoda, M. Tomita, and Y. Ishihama, “emPAI Calc—for the estimation of protein abundance from large-scale identification data by liquid chromatography-tandem mass spectrometry.,” *Bioinformatics*, vol. 26, no. 4, pp. 576–577, 2010.

[27] M. A. Beilstein, N. S. Nagalingum, M. D. Clements, S. R. Manchester, and S. Mathews, “Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 43, pp. 18724–18728, 2010.

[28] M. Lavin, P. S. Herendeen, and M. F. Wojciechowski, “Evolutionary rates analysis of leguminosae implicates a rapid diversification of lineages during the tertiary,” *Systematic Biology*, vol. 54, no. 4, pp. 575–594, 2005.

[29] Y. B. Cho, S. I. Jones, and L. Vodkin, “The transition from primary siRNAs to amplified secondary siRNAs that regulate chalcone synthase during development of *Glycine max* seed coats,” *PLoS ONE*, vol. 8, no. 10, Article ID e76954, 2013.

[30] S. J. Clough, J. H. Tuteja, M. Li, L. F. Marek, R. C. Shoemaker, and L. O. Vodkin, “Features of a 103-kb gene-rich region in soybean include an inverted perfect repeat cluster of CHS genes comprising the I locus,” *Genome*, vol. 47, no. 5, pp. 819–831, 2004.

[31] J. H. Tuteja, G. Zabala, K. Varala, M. Hudson, and L. O. Vodkin, “Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in *Glycine max* seed coats,” *The Plant Cell*, vol. 21, no. 10, pp. 3063–3077, 2009.

[32] M. Senda, C. Masuta, S. Ohnishi et al., “Patterning of virus-infected *Glycine max* seed coat is associated with suppression of endogenous silencing of chalcone synthase genes,” *Plant Cell*, vol. 16, no. 4, pp. 807–818, 2004.

[33] J. H. Tuteja, S. J. Clough, W.-C. Chan, and L. O. Vodkin, “Tissue-specific gene silencing mediated by a naturally occurring chalcone synthase gene cluster in *Glycine max*,” *Plant Cell*, vol. 16, no. 4, pp. 819–835, 2004.

[34] Y. Tsukada, K. Kitamura, K. Harada, and N. Kaizuma, “Genetic analysis of subunits of two major storage proteins (β-conglycinin and glycinin) in soybean seeds,” *Japanese Journal of Breeding*, vol. 36, no. 4, pp. 390–400, 1986.

[35] H. B. Krishnan, S. S. Natarajan, A. A. Mahmoud, and R. L. Nelson, “Identification of glycinin and β-conglycinin subunits that contribute to the increased protein content of high-protein soybean lines,” *Journal of Agricultural and Food Chemistry*, vol. 55, no. 5, pp. 1839–1845, 2007.

[36] R. W. Yaklich, “β-conglycinin and glycinin in high-protein soybean seeds,” *Journal of Agricultural and Food Chemistry*, vol. 49, no. 2, pp. 729–735, 2001.

[37] A. D. Shutov, I. A. Kakhovskaya, A. S. Bastrygina, V. P. Bulmaga, C. Horstmann, and K. Müntz, “Limited proteolysis of β-conglycinin and glycinin, the 7S and 11S storage globulins from soybean [*Glycine max* (L.) Merr.]. Structural and evolutionary implications,” *European Journal of Biochemistry*, vol. 241, no. 1, pp. 221–228, 1996.

[38] T. Yoshikawa, S. Utsumi, T. Fukuda, Y. Okumoto, T. Sayama, and T. Tanisaka, “Identification of genes controlling the contents of seed storage proteins in soybean—identification and functional analysis of the quantitative trait locus *qPro1*,” *Soy Protein Research*, vol. 12, pp. 27–32, 2009.

[39] Y. Tsubokura, M. Hajika, H. Kanamori et al., “The β-conglycinin deficiency in wild soybean is associated with the tail-to-tail inverted repeat of the α-subunit genes,” *Plant Molecular Biology*, vol. 78, no. 3, pp. 301–309, 2012.

[40] K. McGlew, V. Shaw, M. Zhang et al., “An annotated database of Arabidopsis mutants of acyl lipid metabolism,” *Plant Cell Reports*, vol. 34, no. 4, pp. 519–532, 2015.

[41] L. D. Stein, C. Mungall, S. Shu et al., “The generic genome browser: a building block for a model organism system database,” *Genome Research*, vol. 12, no. 10, pp. 1599–1610, 2002.