*Article*

# Functional Haplotypes and Evolutionary Insight into the Granule-Bound Starch Synthase II (*GBSSII*) Gene in Korean Rice Accessions (KRICE_CORE)

**Thant Zin Maung** [1] **, Sang-Ho Chu** [2] **and Yong-Jin Park** [1,2,*]

1   Department of Plant Resources, College of Industrial Science, Kongju National University, Yesan 32439, Korea; tzmaung.yau2009@gmail.com
2   Center of Crop Breeding on Omics and Artificial Intelligence, Kongju National University, Yesan 32439, Korea; sanghochu76@gmail.com
*   Correspondence: yjpark@kongju.ac.kr

**Abstract:** Granule-bound starch synthase 2 (*GBSSII*), a paralogous isoform of *GBSSI*, carries out amylose biosynthesis in rice. Unlike *GBSSI*, it mainly functions in transient organs, such as leaves. Despite many reports on the starch gene family, little is known about the genetics and genomics of *GBSSII*. Haplotype analysis was conducted to unveil genetic variations (SNPs and InDels) of *GBSSII* (*OS07G0412100*) and it was also performed to gain evolutionary insight through genetic diversity, population genetic structure, and phylogenetic analyses using the KRICE_CORE set (475 rice accessions). Thirty nonsynonymous SNPs (nsSNPs) were detected across the diverse *GBSSII* coding regions, representing 38 haplotypes, including 13 cultivated, 21 wild, and 4 mixed (a combination of cultivated and wild) varieties. The cultivated haplotypes (C_1–C_13) contained more nsSNPs across the *GBSSII* genomic region than the wild varieties. Nucleotide diversity analysis highlighted the higher diversity values of the cultivated varieties (weedy = 0.0102, landrace = 0.0093, and bred = 0.0066) than the wild group (0.0045). The cultivated varieties exhibited no reduction in diversity during domestication. Diversity reduction in the japonica and the wild groups was evidenced by the negative Tajima's *D* values under purifying selection, suggesting the domestication signatures of *GBSSII*; however, balancing selection was indicated by positive Tajima's *D* values in indica. Principal component analysis and population genetics analyses estimated the ambiguous evolutionary relationships among the cultivated and wild rice groups, indicating highly diverse structural features of the rice accessions within the *GBSSII* genomic region. $F_{ST}$ analysis differentiated most of the classified populations in a range of greater $F_{ST}$ values. Our findings provide evolutionary insights into *GBSSII* and, consequently, a molecular breeding program can be implemented for selecting desired traits using these diverse nonsynonymous (functional) alleles.

**Keywords:** granule-bound starch synthase 2 (*GBSSII*); haplotype; SNP; domestication; cultivated rice; wild rice

## 1. Introduction

Rice is a cereal grain in which the starch is mainly used as a food source for humans. Starch biosynthesis includes many enzymes, such as starch synthase, granule bound starch synthase (GBSS), starch-branching enzyme (SBE), debranching enzyme, and pullulanase, which synthesize amylose and amylopectin, the two main components of starch [1]. GBSS is the only enzyme that synthesizes amylose, which is a linear form of starch. GBSS has two similar isoforms, GBSSI which carries out amylose synthesis in the endosperm [2], and GBSSII which has the same activity but in leaves [3]. Salt-induced research has been conducted on the down-regulation of *GBSSI* and *II* genes to achieve molecular regulation of starch accumulation in rice seedling leaves [4].

Starch granules accumulate in plants as transient starch in the chloroplasts of source organs (such as leaves) or storage starch in amyloplasts of sink organs (such as seeds, tubers,

and roots) [5]. Transient starch is synthesized during photosynthesis, whereas storage starch accumulates once photosynthesis becomes inactive and later degrades to sustain the seedlings during germination [6]. Comprehensive expression analysis revealed that GBSSII protein is mainly expressed in non-storage tissues of plants, such as leaves [7], for starch biosynthesis. *GBSSII* functions as a mediating binding protein during starch granule development in rice leaves [8]. Similar to the *GBSSI* gene, which is a grain quality (eating and cooking quality) determinant of amylose content in the seed endosperm, the *GBSSII* gene expressed in rice leaves also induces nitrogen starvation by supplying ammonia or amino acids, which, in turn, affect the photosynthetic products (hexokinase inhibitor) [7].

It was recently discovered that the location of GBSS on starch granules in Arabidopsis leaves is facilitated by the presence of the protein targeting to starch 1 (PTST1) protein, which is specifically required for amylose biosynthesis [9]. PTST1 was first identified as a plastidial protein containing an N-terminal coiled-coil domain (specialized α-helices that often mediate protein-protein interactions) and a carbohydrate-binding module 48 (CBM48) at the C-terminus [9,10]. Seung et al. [11] updated their findings with two additional plastidial proteins, PTST2 and PTST3, which are homologs of PTST1, and identified their expression in chloroplasts of Arabidopsis leaves, where they control the initiation of small starch granules.

Despite the differences in specific expression patterns and protein locations of *GBSSII* and *GBSSI* [3], the two genes encoding these proteins share high sequence similarity with almost 66% identical deduced amino acid sequences in wheat [3] and about 70% homology at the nucleotide sequence level in rice [4]. However, several different functional alleles or haplotypes [12–14] have been identified in the waxy gene (*GBSSI*) region and are associated with amylose content in rice and many other crops, such as maize [15], wheat [16], barley [17], and cassava [18,19], but research on the functional properties of *GBSSII* gene in crops is limited, particularly in rice. A recent study reported novel *GBSSII* haplotypes in rice through genome-wide identification and genetic variation analyses [20], but additional evolutionary analyses on this gene have not been performed. Evolutionary insights based on SNP haplotypes of many starch synthase genes (*GBSSI*, *SSI*, *SSIIa*, *SSIIb*, *SSIIIa*, *SSIIIb*, *SSIVa*, and *SSIVb*) other than *GBSSII* have been provided by the high variability between cultivated and wild rice populations [21]. The limited findings on *GBSSII* genetics and its evolutionary history allowed us to explore the genomic region of this gene using high-throughput sequencing technology in the 475 KRICE_CORE accession set based on classified rice populations.

Next-generation sequencing (NGS) is a powerful tool to discover domestication genes in crop plants and their wild relatives [22,23]. The recent rapid development of NGS has boosted the number of genomic sequences for hundreds to thousands of rice varieties [24]. Evolutionary research on rice has been performed previously (Khush 1997) and various strategies, such as whole-genome resequencing [25] or sequencing [26,27], natural or artificial selection [12], and genotyping-by-sequencing [16], have been explored by breeders to broaden the genetic diversity of crops by examining sequence variations and either evolutionary relationships between or among populations using NGS data. Genetic variations (alleles) have been detected using various selection methods and have been increasing in number due to their distinctive patterns or signatures due to selection or other confounding effects, such as bottlenecks, expansions, or subdivided populations [28].

A breeding program to develop a new cultivar requires knowledge and understanding of the genetic diversity, genetic relationships, and population structure of the desired gene [29]. To identify the functional genetic properties of *GBSSII* gene and provide evolutionary insight, a deep sequencing analysis was performed for the whole genomes of the 475 Korean rice accessions (KRICE_CORE), and genetic variations were detected. Moreover, a series of analyses was conducted to estimate the genetic diversity, population genetic structure, and genetic relationships using classified rice populations. Notably, limited information is available about the functional properties of the *GBSSII* gene and the related evolutionary story in rice. The findings of this study will fill in missing information on the

functional properties of the *GBSSII* gene and its domestication signatures, which, in turn, will be applicable to future rice variety breeding programs.

## 2. Materials and Methods

### 2.1. Plant Materials and Experimental Site

A core set of 421 cultivated rice accessions consisting of three variety types (landrace, weedy, and bred) (Data S1) was collected previously around the globe, administered by the National GeneBank of the Rural Development Administration (RDA-GenBank, Republic of Korea) using the PowerCore program [30], and selected for whole-genome resequencing [31]. An additional set of 54 wild rice accessions was obtained from the International Rice Research Institute in 2017. These two core sets were combined to form the KRICE_CORE (Korean World Rice Collection).

Field experiments were conducted with these 421 Asian-cultivated and 54 wild-rice accessions in the departmental field of the Plant Resources Department, Kongju National University (Yesan Campus) in 2016 and 2017. Among the three varieties, there were 279 temperate japonica, 26 tropical japonica, 102 indica, 9 aus, 2 aromatic, and 3 admixed varieties (Table S1). Recommended cultural practices for crop management were carried out as necessary.

### 2.2. DNA Extraction, Resequencing and Variant Calling

Plant samples (young green leaves) were collected from all tested plants approximately 15 days after transplanting. DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method, and genomic DNA was stored at 4 °C until use [32]. High-quality DNA was used for whole-genome resequencing of the rice varieties with an average coverage of approximately 15× on the Illumina HiSeq 2000 Sequencing Systems Platform. These HiSeq 2000 sequencing data were deposited in the NCBI GenBank data (accession numbers: SAMEA4999071—SAMEA4999480, SAMN12714206—SAMN12714236 and SAMN16203520—SAMN16203712). The decoded sequences were saved in the FastQ file format. A program package, VCFtools (variant call format) version 0.1.15 [33], was used to remove missing values and heterozygotes from the raw data and the results were saved in the FastQ format. To compare the output sequences among the accessions, high-quality reads that remained after removing the missing values and heterozygotes were aligned to the International Rice Genome Sequencing Project (IRGSP) 1.0 rice genome sequence. The alignment of the reads was saved in binary alignment map (BAM) format. Duplicate reads aligned in multiple locations were removed using PICARD (version 1.88) software [34]. SNP and InDel calling was then performed using GATK tools (version 4.0.1.2) [35,36] to identify the SNP regions in the BAM file. The extracted mutations were saved in VCF file format and filtered using VCFtools (version 0.1.15) to remove false-positive SNPs/InDels. To identify the genetic variants, the specific variant files of the classified subgroups were viewed by using bcftools program package version 1.8 and their respective numbers of genetic variants were counted in TASSEL 5 (version 20210408) software [37].

### 2.3. Population Genetic Structure and Phylogenetic Study

To identify the number of populations in the 475 rice accessions, VCFtools version 0.1.15 was used to convert the previously called variants into PLINK outputs, and bed files were created again using the PLINK (version 1.07) analysis toolset. Two additional files (.bim and .fam format) were incorporated using Python script (structure.py) with the fastStructure [38] package tools and a range of increasing K values from 2 to 7. The admixture patterns of the defined populations (population structure) were inferred using average Q-values in the POPHELPER version 2.3.1 [39] analytical tool of RStudio (version 1.4.1106) software. Principal component analysis (PCA) was processed to explore the classified subpopulations on a dimensional scale in RStudio version 1.4.1106. A list of principal components (PC1 and PC2) referring to variants was generated in TASSEL 5 (version 20210408) [37], and relatedness among the classified subpopulations was investigated in

2D scatterplots using RStudio (version 1.4.1106). A phylogenetic analysis was conducted in MEGA X software [34] using the neighbor-joining method, and a tree was drawn in FigTree (version 1.4.4) software.

*2.4. Nucleotide Diversity, Tajima's D, and the Fixation Index ($F_{ST}$ Test)*

To determine the degree of polymorphism within the pairwise comparisons of the classified populations, the resulted DNA sequences based on the populations were investigated for their diversity using nucleotide diversity values ($\pi$). Using these population-based sequences with variants, Tajima's *D* values were also calculated to measure the difference between the estimated average number of nucleotides and the observed number of segregating sites for all 475 rice accessions. The estimated measures of the fixation index ($F_{ST}$-values) were investigated to observe the performance of the genetic relationship between and among the populations. Using VCFtools (version 0.1.15), variant files were selected for the *GBSSII* gene region to compare the classified representative cultivated ecotypes/subpopulations. The sliding window size used for nucleotide diversity ($\pi$) and the Tajima's *D* calculations was 1.5 kb, and the values were compared among the classified rice groups based on their varietal types or ecotypes.

*2.5. Haplotype Network*

A haplotype network was constructed to investigate the genetic relatedness of the samples based on variants within the *GBSSII* gene region. VCFtools was used to specify the selected gene regions of the tested samples, and the reference gene sequences adapted from the Rice Annotation Project Database (RAP-DB, https://rapdb.dna.affrc.go.jp/index.html, last accessed on 20 April 2021) were aligned in Molecular Evolutionary Genetics Analysis (MEGA X) version 10.1.8 software [40]. A compiled list of aligned sequences was saved in the nexus file format and analyzed in DnaSP (version 6.0) software [41] to generate a list of haplotypes with their accession numbers. Taking the accessions with each haplotype into account, a list of the same mutated sequences was created for each trait/ecotype in Population Analysis with Reticulate Trees (PopART) software [42], and a TCS network [43] was drawn using that ecotype list.

## 3. Results

*3.1. Identification of Genetic Variations*

The *GBSSII* genomic region (*Os07g0412100*) is located between positions 12,916,883 and 12,924,202 (–strand) on chromosome 7 and is 7320 bp in length. Genetic variations among populations are important because the genetic differences are the ultimate source of useful mutations for breeding. The whole *GBSSII* genomic region of the 475 rice accessions were resequenced using VCFtools and the counted variations classified by geography (ecotypes). Five types of genetic variations were identified, including single nucleotide polymorphisms (SNPs), insertions (Ins), deletions (Dels), duplications (Dupls), and different variations (DVs) (Table 1). Among the variations, SNPs were the most frequently observed and comprised the largest number of variations in every cultivated ecotype, of which aromatic varieties had the highest number of SNPs (200) among the classified cultivated subgroups. For the specific numbers of variations, the wild rice group was subdivided into *Oryza nivara*, *Oryza rufipogon*, and others. Cultivated and wild rice have a considerable number of variations that are useful for investigating their functional responses in evolutionary analyses. Interestingly, all of the classified cultivated subgroups (ecotypes) had more SNPs, indicating differences in segregated sites, which, in turn, supported the genetic diversity values.

*3.2. Haplotype Variations*

To investigate the haplotype variations that were due to differences in alleles (SNPs and InDels) at their different segregating sites, a haplotype analysis was performed for all of the cultivated (421) and wild (54) rice accessions and the differences or similarities in terms

of polymorphic changes (SNPs,) including InDels, were determined within the *GBSSII* gene region. Using such genetically distinct sequences of the 475 rice accessions, these analyzed sequenced accessions were grouped into haplotypes after alignment with the Nipponbarae reference (Data S2). The differences among the classified haplotypes were examined. A total of 45 haplotypes was observed representing a total of 113 variants, including 58 SNPs and 55 InDels, within the gene region of *GBSSII* in chromosome 7 (Data S3). There were 30 nonsynonymous SNPs among the 58 SNPs, and as a summary, 38 functional haplotypes were verified based on these SNPs, and mainly discussed these functional SNP substitutions (Figure 1). According to the haplotypes in which they belonged to the same rice ecotypes, it was formed into three main groups, including cultivated rice (13 haplotypes: C_1–C_13), wild rice (21 haplotypes: W_1–W_21), and mixed (cultivated and wild, 4 haplotypes: M_1–M_4).

**Table 1.** Summary of genetic variations in the *GBSSII* gene region of 475 accessions from the Korean world rice collection. The *GBSSII* reference gene region was adapted from Nipponbare. Others [1]: group of wild rice accessions other than *Oryza nivara* and *Oryza rufipogon*, Ins [2]: insertion, Del [3]: deletion, Dupl [4]: duplication, and DV [5]: different variation.

| Group | Subgroup (Ecotype) | Total No. of Variations | | | | | No. of Accessions |
|---|---|---|---|---|---|---|---|
| | | SNP | Ins [2] | Del [3] | Dupl [4] | DV [5] | |
| Cultivated rice | Temperate Japonica | 193 | 16 | 21 | 1 | 1 | 279 |
| | Tropical Japonica | 122 | 3 | 7 | 0 | 0 | 26 |
| | Indica | 196 | 13 | 15 | 1 | 1 | 102 |
| | Aus | 200 | 12 | 16 | 1 | 1 | 9 |
| | Aromatic | 101 | 4 | 5 | 0 | 0 | 2 |
| | Admixture | 146 | 6 | 10 | 0 | 1 | 3 |
| Wild rice | *O. nivara* | 121 | 5 | 5 | 0 | 0 | 3 |
| | *O. rufipogon* | 120 | 6 | 9 | 1 | 0 | 3 |
| | Others [1] | 276 | 25 | 43 | 1 | 0 | 48 |

Nineteen functional substitutions (nonsynonymous SNPs) were identified in the group of cultivated haplotypes (C_1–C_13) and were present in 162 temperate japonica, 18 tropical japonica, 6 indica, 3 aus, and 2 admixture rice accessions. Among these 19 functional SNPs, there were four C to T nucleotide substitutions (two yielded the same amino acid change and the other two produced different amino acid changes). The remaining substitutions and their frequencies were three G/A and T/C; two G/C, A/G, and C/A; G/T, A/T, and T/G indicating transitions in their respective amino acids. Among the 13 cultivated haplotypes, C_1 was the predominant haplotype (149 accessions), followed by C_3, a temperate japonica-specific haplotype represented by 20 accessions. Despite C_1 being represented by the highest number of rice accessions, there was only a single functional SNP (A/G in exon 12). By contrast, C_3 contained 14 functional SNPs, which was the highest number of alleles for any of the haplotypes.

Twenty functional SNPs were detected in the wild group (21 haplotypes: W_1–W21), representing 30 wild rice accessions. Different frequencies of functional SNPs were observed with four C/T; three C/A; two G/C; two T/C; two A/G; two A/T; and two G/T, G/A, C/G, and T/C. Thirteen functional SNPs were found in the mixed group (cultivated and wild, four haplotypes: M1–M4), representing two (G/C, T/C, A/G, C/T, C/A), G/T, A/T, and T/G for 105 cultivated and 9 wild rice accessions.

Most of the functional substitutions were in the same positions in the wild and cultivated rice groups. However, some functional SNPs were found only in the cultivated or wild groups. Such functional SNPs as T/C (exon 12), G/A (exon 11), C/T (exon 8), and G/A (exons 2 and 1) were only found in the cultivated haplotypes, with different amino acid changes. Similarly, the wild haplotypes also contained various SNPs including C/A (exon 13); T/C (exon 11); A/T (exon 10); G/T and C/T (exon 9); G/A, G/T, T/C, two C/T, and C/A (exon 8); A/G and C/G (exon 6); G/C (exon 4); and A/T and C/T (exon 1).

The number of functional SNPs at different segregation sites did not differ between the cultivated and wild groups.



**Figure 1.** Gene structure and haplotype analysis of the *GBSSII* (*Os07g0412100*) gene in the 475 Korean rice accessions. Light-blue (cultivated) and pale-green-colored (wild) columns provide the list of haplotypes together with their respective number of rice accessions in each subpopulation. Blue cells indicate all SNP variations (minor alleles) found only at nonsynonymous mutation sites. Blank cells indicate major alleles or the same nucleotide to that of the reference. Cyan color, "Het", refers to "heterozygote", and the "dash (-)" indicate a position for an "unknown" nucleotide or generally refers to "N" (any nucleotide A, T, G, or C). C: cultivated rice, W: wild rice, M: a combination of cultivated and wild rice. Note: The selected number of chromosome positions and sorted haplotypes were based on the identifications where nonsynonymous substitutions (nsSNPs) or functional substitutions were observed.

Despite more functional SNP numbers in the wild group, the number of representative rice accessions were very few. In case of polymorphisms by insertions or deletions (InDels), a greater number of diverse alleles was found in the wild group compared to the cultivated rice group based on the same number of 38 haplotypes (Figure S4). In this case, some

of the minor alleles were responsible to minor haplotypes, especially to the wild and the positions for such alleles were excluded together with the positions indicated by 1 bp Ins/Del variations. Lastly, there were only 27 positions for InDels and interestingly, all the identified variations generated were deletions (Dels) (Figure S4). Similar to the haplotypes for SNP, major haplotypes (such as M_3 and C_3, except C_1) indicated the same deletion (Del) variations. Interestingly, most identified haplotypes generated deletion (Del) variations.

*3.3. Population Genetic Structure and Genetic Differentiation ($F_{ST}$ Test)*

To stratify the ancestral proportion of the classified populations due to the genetic similarity or differentiation of the *GBSSII* gene, the analysis on population structure was carried out by increasing the K value from 2 to 7 (Figure 2A). At every K value (K = 2 to 7), the wild group was clearly separated from the cultivated ecotypes, but its internal subgroups were partially admixed with each other. On the variety basis, from K = 3 to 7, the cultivated subgroups were also separated from each other but at K = 3, the internal subgroups of bred and landrace were in close ancestral history indicating a high genetic similarity. At K = 3 and 4, the identified clusters of the wild enhanced a clear separation to the wild from the cultivated subgroups, but the genetic influxes within the *GBSSII* region were connected with all of the classified varieties, especially the bred variety. Meanwhile, the structural features of landrace and bred were similar at K values 4 and 5, indicating their close associations. On the basis of ecotypes, the cultivated subgroups were also separated with each other (Figure S1A), but like the wild group, their internal subgroups, by different cluster numbers, shared their ancestral proportions, indicating their close genetic association within the *GBSSII* region. Despite the clear separation at K = 3 and 4, the major cultivated ecotypes, indica and temperate japonica, had a similar population structure, and the rice accessions shared their genetic properties in the *GBSSII* region between them. Each group was subdivided into many individual subgroups at every K value, suggesting that there might be internal interbreeding events among or between the subgroups.

The genetic relatedness of classified populations was explored by performing the analysis based on the detected principal components within the *GBSSII* gene region. There were four clear separations of different clusters by *GBSSII* genetic differentiation levels (Figure 2B). As expected, the wild group revealed the most diverse *GBSSII* genetic structure, although its accessions were clearly separated as a group (group II). Among the cultivated groups, bred was shown to be an isolated group (group II) but associated with some wild accessions while its additional individuals were closely admixed with other cultivated groups, landrace and weedy in groups I, III and IV, indicating their genetic similarity within the *GBSSII* gene region. Once PCA on the cultivated rice accessions were repeated based on their ecotypes (Figure S1B), the most admixed cluster was observed in group I, exhibiting that most of the cultivated subpopulations, except admixture and aromatic, are associated with the wild group. Group III was indica specific, indicating its close relatedness to some wild rice accessions. Group IV was clustered with a higher proportion of temperate japonica, followed by indica, aromatic, and wild. Overall, four clear separate groups revealed admixed clusters in the cultivated and wild groups, suggesting a closer genetic association within the cluster.

To determine the level of genetic differentiation between the classified subpopulations, the fixation indices ($F_{ST}$ values) were estimated using genetic variants within *GBSSII* gene region (Figures 2C and S1C and Table S2). On the basis of varieties (landrace, weedy, and bred), the differentiation levels from the wild group were signified by pairwise $F_{ST}$ values. The $F_{ST}$ values among the classified groups indicated a range from 0.0196 to 0.1992, suggesting a similar differentiation between them (Figure 2C). In case of ecotype basis, the greater genetic distances were indicated between most of the cultivated subpopulations, and the highest $F_{ST}$ value was detected between tropical japonica and aromatic (0.8318), followed by temperate japonica and aromatic (0.7567) (Table S2). As expected, a close

association tendency by narrow genetic differentiation was observed between temperate and tropical japonica (0.0498). Genetic differentiation of the wild from the cultivated subpopulations were varied: the highest was observed in indica (0.4335) and the lowest occurred in aus (0.1343). Two major cultivated subpopulations, indica and japonica (temperate and tropical), expressed a wide range of genetic differentiation of 0.6952 for indica-temperate japonica and 0.6935 for indica-tropical japonica. Statistically, aus had a greater differentiation from both japonica groups, compared to its closer relationship with indica (Figure S1C).
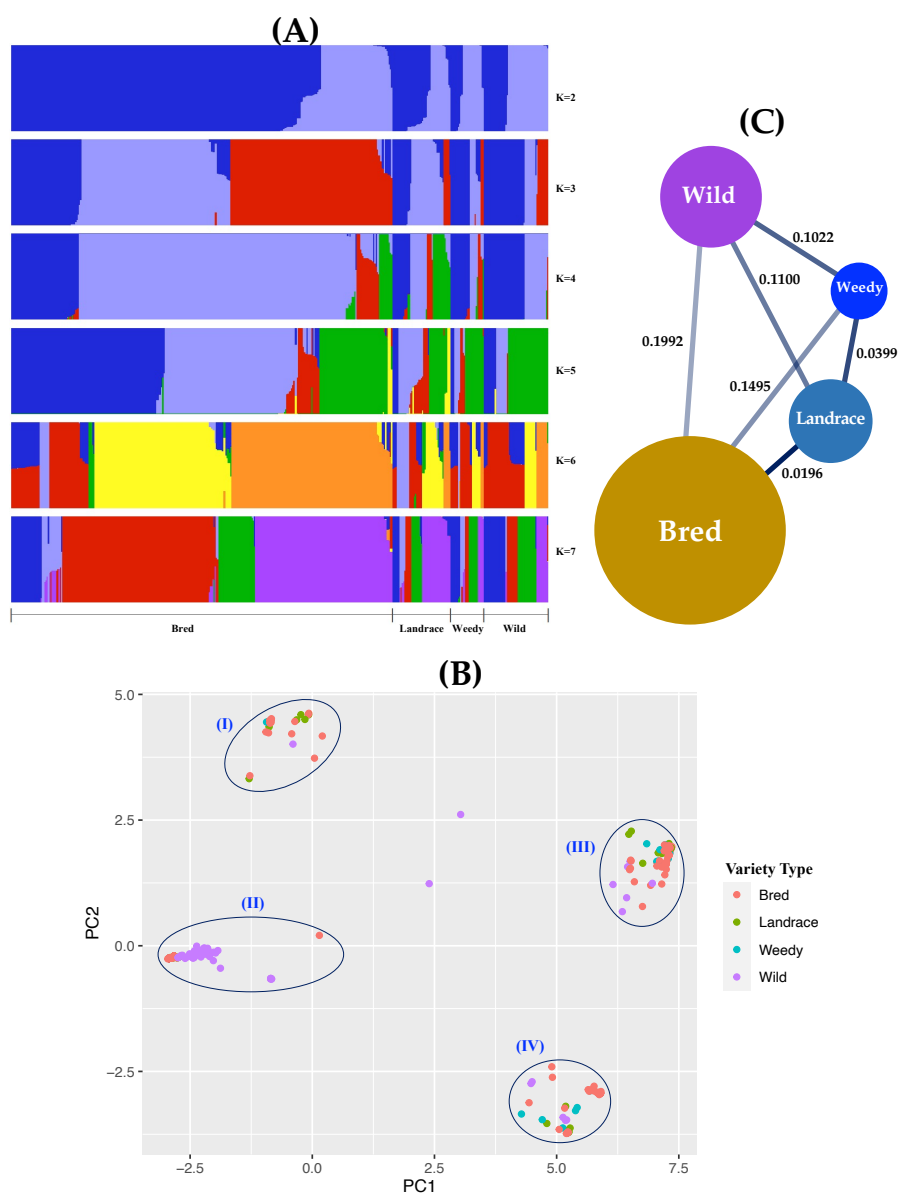


**Figure 2.** Estimate of structure and population differentiation within the gene region of *GBSSII* (*Os07g0412100*) in the Korean rice collection in terms of different varietal types (landrace, weedy, bred, and wild). (**A**) Population structure of the *GBSSII* gene in the 475 Korean rice accessions clustered by increasing K value from 2 to 7. The different colors of each K value refer to the different numbers of clustered populations. (**B**) Two-dimensional (2D) principal component analysis (PCA) of the 475 Korean rice accessions. (**C**) Pairwise estimates of genetic differentiation ($F_{ST}$ values) of the *GBSSII* gene among different varietal types of the 475 Korean rice accessions.

### 3.4. Nucleotide Diversity Analysis

To measure the degree of polymorphism due to different genetic variations such as SNPs and InDels, the nucleotide diversity values of *GBSSII* were measured based on the classified populations, and the differences investigated between and among them (Figures 3 and S2). Here, it was considered to include the upstream and downstream regions of *GBSSII* to estimate diversity. In the case of variety basis, landrace, weedy, bred, and wild, the resulted nucleotide diversity values ranged from the lowest of 0.0027 (wild) at the position 12,919,500 to the highest of 0.0213 (weedy) indicated at position 12,922,500 (Figure 3A). Once the nucleotide diversity values were compared by their average values (Table S3), it was noticed that the lowest diversity value (0.0045) was in the wild group, while the highest diversity value occurred in the weedy group (0.0102). Despite no significant difference in diversity values among the classified groups, a considerable reduction in diversity was found in the wild group compared to any of the cultivated rice groups.
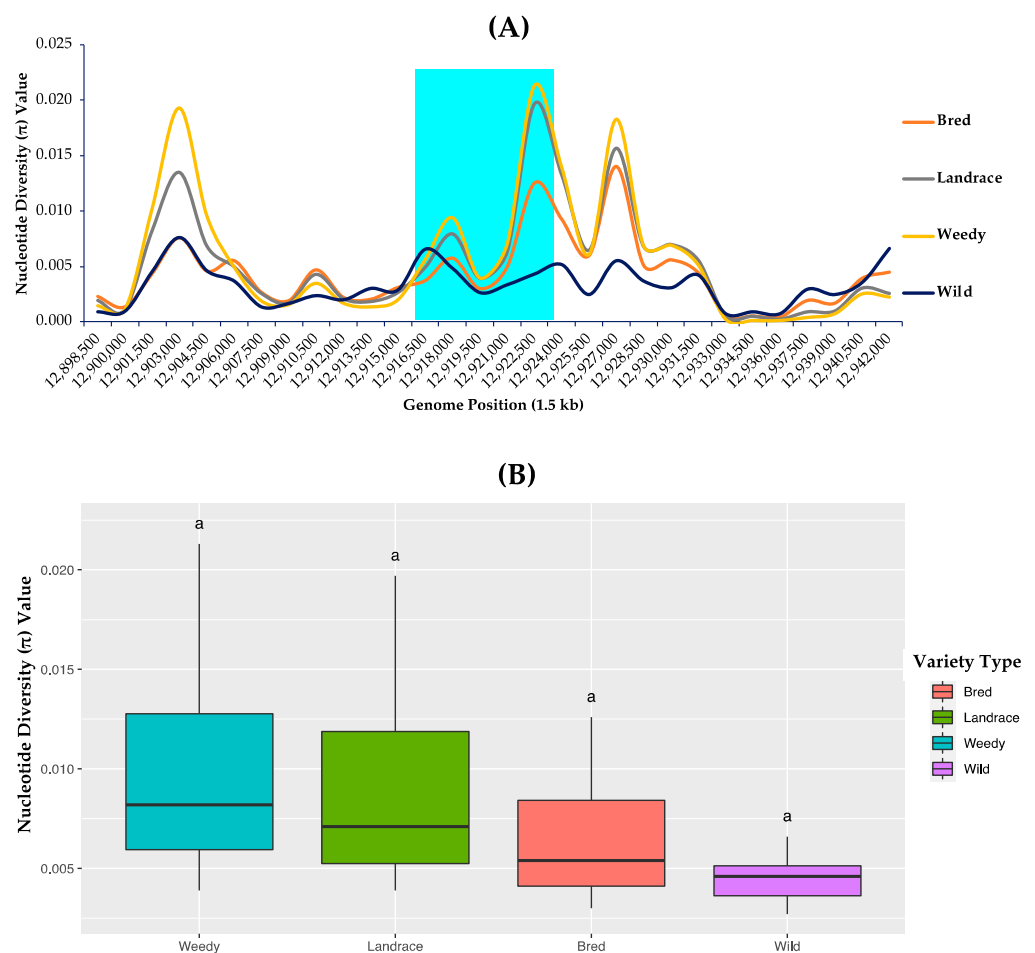


**Figure 3.** Nucleotide diversity analysis of *GBSSII* (*Os07g0412100*) in the 475 Korean rice accessions utilizing the variety type (landrace, weedy, and bred) together with the wild group. (**A**) Nucleotide diversity (π-value) representing the number of nucleotide variations within the *GBSSII* gene region at individual segregating sites in 1.5 kb sliding windows. Cyan indicates the *GBSSII* gene region, and each colored line represents a different rice varieties. (**B**) Box plots representing the different distribution patterns of *GBSSII* genetic variations based on mean nucleotide diversity values among the variety types. Lowercase letter "a" indicate a significant difference level at *p*-value < 0.05.

The differences in nucleotide diversity values were investigated among the classified groups by means of ecotype differences (temperate japonica, indica, tropical japonica, aus, aromatic, and admixture), including a group for the wild rice accessions (Figure S2 and

Table S4). It was observed that aus indicated the highest nucleotide diversity value (0.0332) at chromosome position of 12,922,500 and the lowest in temperate japonica (0.0003) at position 12,919,500. On the basis of comparing their average values, the same previous feature of the lowest and highest diversity values was represented by temperate japonica (0.0029) and aus (0.0141) and the wild (0.0045) was only higher than the japonica subpopulations (temperate and tropical japonica). Overall, despite the lowest value in the wild group the japonica group showed the lowest nucleotide diversity value among the ecotypes, indicating a wide range of genetic variation of *GBSSII* among the cultivated populations.

### 3.5. Tajima's D Test

To understand the difference between the estimated values of the nucleotide differences and the observed values of the segregating sites, the corresponded DNA sequences within the *GBSSII* gene region were analyzed to calculate Tajima's *D* values based on the classified subpopulations of the 475 rice accessions. Tajima's *D* values were computed using the same classified rice groups used for the nucleotide diversity analysis (Figure 4 and Figure S3). On the basis of variety type, all the cultivated varietal groups showed only positive Tajima's *D* values within the *GBSSII* gene region while both positive and negative values were indicated for the wild group. The highest Tajima's *D* value (3.5924) was detected at position 12,919,500 in the *GBSSII* gene region of the weedy group and the lowest (−0.6795) was found for the wild group at 12,916,500. Once the average Tajima's *D* values were calculated, the same representative of the highest and positive Tajima's *D* value was in the weedy rice set (2.7883), while the lowest and negative value (−0.0499) was still in the wild rice (Table S3). Interestingly, all the cultivated varietal groups had positive Tajima's *D* values.

In case of ecotype basis, all the classified subgroups showed both positive and negative Tajima's *D* values, except aus in which all identified positions indicated positive Tajima's *D* values (Figure S3). From their average values within the *GBSSII* gene region, negative Tajima's *D* values were observed in both the japonicas (−0.4769 for temperate and −0.3883 for tropical) and the lowest Tajima's *D* value (−0.0499) was represented by the wild group. The other ecotypes exhibited positive values in their descending order, 2.0490 (aus), 0.9794 (indica), and 0.5919 (admixture) (Table S4).

### 3.6. Haplotype Network and Phylogenetic Analyses

To visualize how the rice accessions were genetically related (different or similar) within the *GBSSII* gene region during the evolution, the informativeness of evolutionary analyses were investigated through haplotype network and phylogenetic analyses. A haplotype network was constructed to infer the level of evolutionary relationships among the different classified populations from the 475 rice accessions, using a list of generated haplotypes identified within the *GBSSII* gene region (Figure 5A). Forty-five haplotypes created a network, covering genetically different DNA sequences in the cultivated (temperate japonica, indica, tropical japonica, aus, aromatic, and admixture) and wild rice groups. As shown in Figure 5A, three major haplotypes were found. One (Hap_5) was indica specific and the other two (Hap_1 and Hap_32) were temperate japonica predominant. The temperate japonica-specific haplotypes (Hap_1 and Hap_32) were close to each other, sharing similar genetic variations of *GBSSII* with tropical japonica, indica, and the admixture including the wild rice. The indica-specific haplotype (Hap_5) was distantly associated with the temperate japonica-specific haplotypes, indicating different genetic variation of *GBSSII*, but shared its similarity with the aus, a few temperate japonica, and wild accessions. The remaining cultivated haplotypes were derivatives of these three major haplotypes at different levels of genetic variation. On the other hand, the wild haplotypes were diverse throughout the network, linking each other in closer genetic relationship under the *GBSSII* region. However, some wild haplotypes had close associations with cultivated haplotypes, particularly with temperate japonica-specific haplotypes.
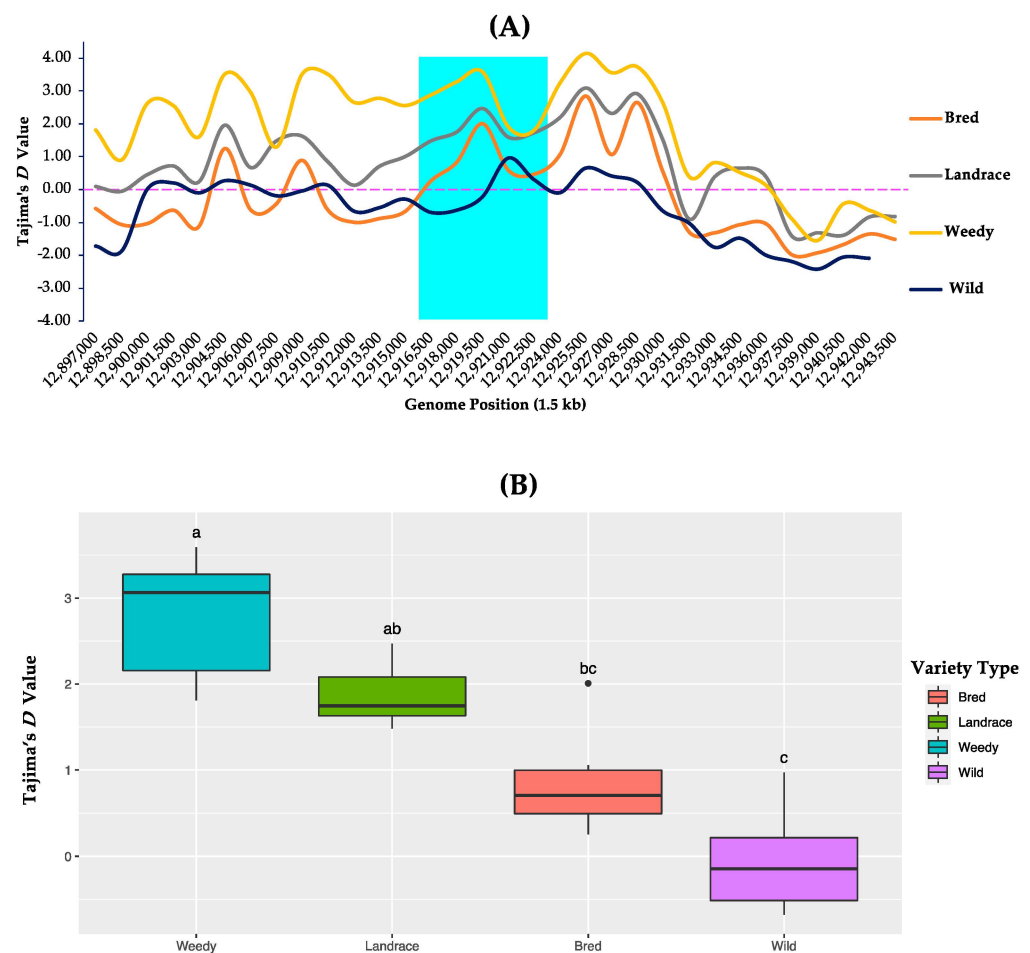
**(A)**



**(B)**



**Figure 4.** Tajima's *D* values of *GBSSII* (*Os07g0412100*) in the 475 Korean rice accessions by means of variety types (landrace, weedy, and bred) together with the wild. (**A**) Tajima's *D* values, representing different individual segregating sites within the *GBSSII* gene region in the 1.5 kb sliding window. Cyan indicates the *GBSSII* gene region, and each colored line represents a different rice variety type. (**B**) Box plots represent different distribution patterns of the *GBSSII* genetic variations according to the Tajima's *D* values among the variety types. Different lowercase letters (a, ab, bc and c) indicate a significant difference level at *p*-value < 0.05

　　　To view the relationship of the tested rice accessions during the evolution of *GBSSII*, a phylogenetic analysis was conducted using different nucleotide sequences from the core set of 475 rice accessions based on the six cultivated subpopulations and one wild group (Figure 5B). It was clearly observed for two clear separations of cultivated subpopulations in the neighbor-joining tree: the temperate japonica-specific subclade shared its similar genetic variation of *GBSSII* with, of course, many tropical japonica, indica, and admixture, derived from the same common ancestor and an indica-specific subclade displayed a close association with many wild, aus, admixture and aromatic, including temperate japonica. This indica-specific subclade was rooted in the same and most recent common ancestor of another temperate japonica subclade. Both major separate subclades were derived from different wild ancestors, indicating a distant common ancestor. Despite the clear separation of the wild subclades, the relatedness of their internal rice accessions was high, and they shared similar genetic variations in *GBSSII* with cultivated rice. By contrast, not only the two major separated groups (indica and japonica) and the other cultivated rice groups (aus, aromatic, and admixture) were separate, although most of them in the same clade were under the recent common ancestors.
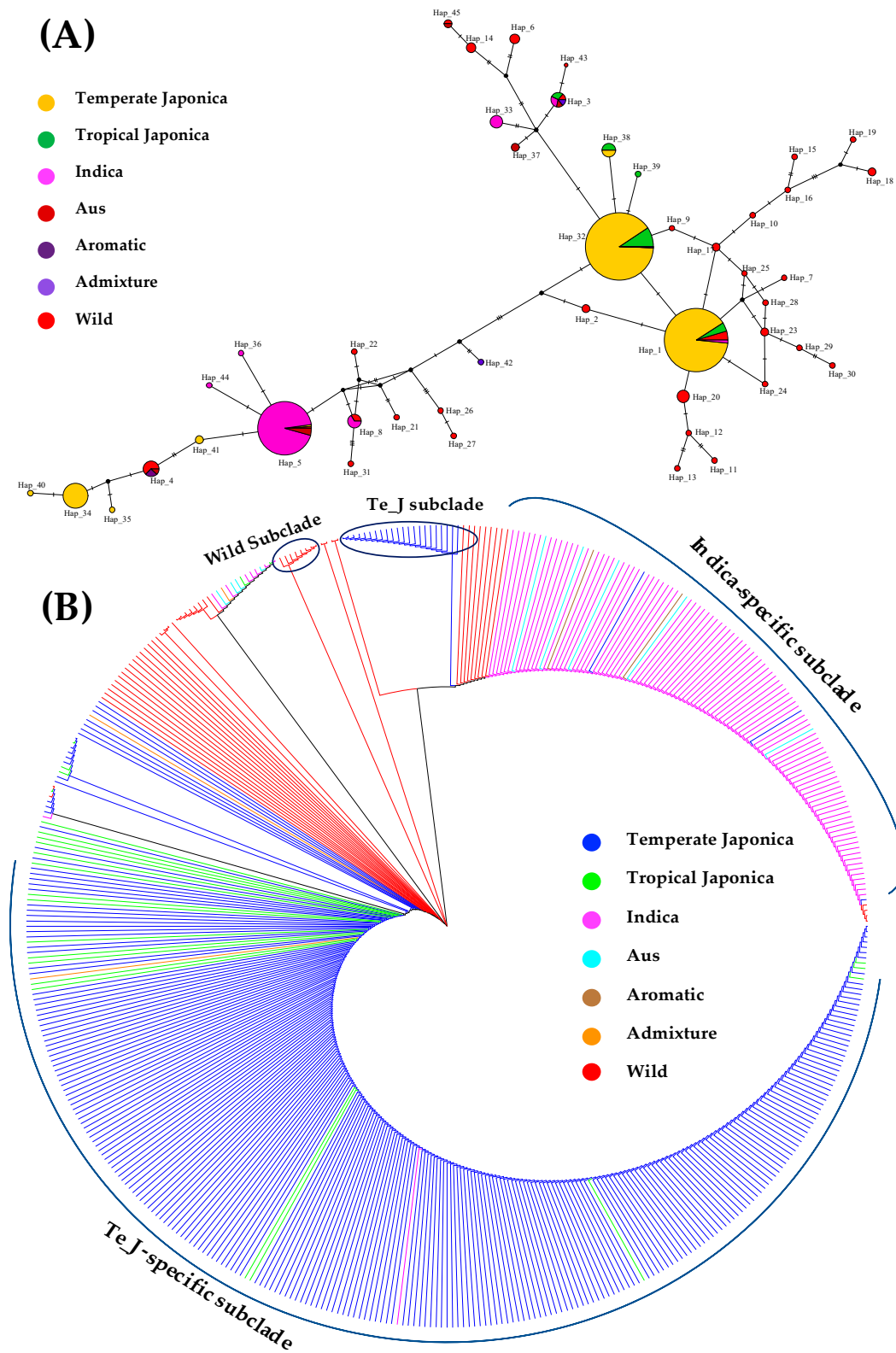
**Figure 5.** Haplotype-based analysis of the *GBSSII* (*Os07g0412100*) gene in the 475 Korean rice accessions. (**A**) Haplotype network visualizing the evolutionary relationships among the identified haplotypes. The size of each circle is proportional to the number of accessions that harbor the haplotype, and different colors refer to the different ecotypes. Black circular dots indicate median vectors and in the present analysis, there were 20 median vectors in total. (**B**) Phylogenetic tree inferring the evolutionary relationships among the tested rice accessions, indicating their ancestral history using the neighbor-joining method. Different colors refer to different ecotypes.

## 4. Discussion

Starch is composed mainly of the glucose polymers amylose and amylopectin, which are alpha-D-glucose units [44]. Most of the starch molecules are 5–35% amylose, which is synthesized by granule-bound starch synthase (GBSS) [44]. The linear chain length of amylose is elongated by GBSS. GBSS is one of the several isoforms of starch synthases (SS) of land plants with two paralogues: GBSSI, which is strictly expressed in endosperm and pollen, and another enzyme, GBSSII, which is expressed in non-storage tissues (e.g., leaves) and the pericarp [3,7,45]. GBSSI is a major determinant of cooking and eating quality based on the amylose content in the rice endosperm, whereas amylose biosynthesis is carried out in rice leaves by GBSSII [8]. Despite these differences, the two enzymes were derived from the same *Oryza species* (*O. japonica*) but belonged to different subclades of an ortholo-gous tree based on 19 plant species, including most rice and other *GBSS*-related species (Figure S5). Comparative sequence analysis is an approach to study genome function and evolution by identifying homologous or paralogous genes to interpret their relation-ships [46]. *GBSSII* in our study is a paralogous of *GBSSI*, representing 70% homology in rice by deducted amino acid [4].

Genetic variations (alleles) in crop species, such as SNPs, InDels, and intron length polymorphisms (ILPs), are essential to revealing molecular mechanisms, as they function in controlling traits [47–50]. Genetic variations of the starch synthase gene family has recently been updated in rice giving new insights into their correlation with improved amylose content and eating quality [20]. Using these identified variations, genetic markers have been developed for various analyses, including genetic diversity assessment, trait association mapping, and fine mapping of QTLs that regulate important agronomic traits [51]. In this study, a core set of 475 genetically different DNA sequences was characterized to identify the functional haplotypes of *GBSSII* and a series of analyses were performed, genetic diversity and population genetic structure analyses to gain evolutionary insight into this gene, *GBSSII*, based on the classified populations.

Haplotype analysis of a specific gene reveals its distinct major dormant and non-dormant haplotypes [52] and can also identify many functional alleles which are signifi-cantly associated with desired plant traits required for a breeding program, such as eating quality or salt tolerance and resistance [53]. Genome-wide haplotype analysis of the starch synthase gene family recently conducted in rice resulting in the identification of six haplo-types with 17 SNPs were identified in the diverse regions of *GBSSII* [20]. In our study, a total of 38 functional haplotypes was recognized, representing several nonsynonymous substitutions (30 SNPs) within the diverse regions of *GBSSII* (Figure 1). Among the three groups of haplotype collections, the cultivated group had not only major haplotypes (C1 and C_3) but also had 19 nonsynonymous SNPs across the different *GBSSII* coding regions. The most predominant haplotype, C_1 had only a single amino acid change of leucine (L) to serine (S) resulting from an A-G substitution at position 12,917,914 on chromosome 7, while the C_3 haplotype had multiple alleles with non-synonymous substitutions across multiple exons in 20 temperate japonica rice accessions. In the wild group, 21 haplotypes revealed 20 SNP positions within the *GBSSII* gene region, representing a total of 30 wild rice accessions. Another major haplotype (M_3) was found in the mixed group, including a total of 93 rice accessions among which indica predominated. Similar to C_3, M_3 exhibited nonsynonymous SNPs. These findings in our study were consistent with the previous findings from a *GBSSII* haplotype analysis that found that Hap_4 and Hap_5 were the predominant haplotypes of *GBSSII* haplotyping and the respective rice accessions were mostly from two major rice groups (indica in Hap_4 and japonica in Hap_5) compared to other groups, aus and admixture [20]. Simultaneously, that research group identified a key *GBSSII* SNP (T/G), which produces significant variations in amylose content. In our study, many functional *GBSSII* alleles (19 SNPs in 13 cultivated haplotypes and 20 SNPs in 21 wild haplotypes) were identified. These functional genetic variations (SNPs) can be assumed to be of high consideration in molecular breeding of starch-related rice varieties' improvement programs.

The levels of starch accumulation and free sugars in rice leaves (sheaths and culm) can be a maximum within 10 to 11 weeks after transplanting [54] and these sugars induce the accumulation of *OsGBSSII* transcripts through the glycolysis-dependent pathway [7]. These accumulated starches in rice leaves are then rapidly decreased during grain filling mechanism [54] and are indirectly translocated into the panicles as carbohydrate contents for grain development [55–57]. Starch content in mature grain can be grain-filling-duration dependent [58] and it can also be regulated by a starch remobilization mechanism in leaf sheaths after rice heading [59]. Therefore, starch biosynthesis in transient organs (leaves) directly or indirectly appeared to be correlated with structural and behavioral functionalities of starch storage (seeds). In our study, functional studies of *GBSSII* haplotypes and their responsible alleles were detected, and these findings can be assumed to be key enzymes for transitory starch. Despite its main expression in transient organs during the daytime, synthetic starches are mobilized at night for plant carbohydrates regulated by circadian rhythm, which indirectly supply non-structural carbohydrates (NSCs) for grain filling. One recent research reported that remobilization of stored starch from leaves to developing grains was positively correlated with a decrease of starch and starch synthase (SS) activities in leaves [58]. Understanding the starch mechanism and its synthetic genes expressed in different plant tissues is beneficial to comprehensive study of the starch synthase gene family in rice. Simultaneously, characterization of these genes in their genetic controls for functional activities also appears to be indispensable.

High-throughput RNA sequencing (RNA-Seq) provides insight into the transcriptome [60], and careful selection of the RNA-Seq quantification measure is critical for inter-sample comparisons, such as different gene expression levels between two or more conditions [61]. In this study, the measures of fragments per kilobase of exon per million (FPKM) were quantified from the RNA-Seq data of 353 rice accessions (300 cultivated and 53 wild) within the *GBSSII* genomic region for an expression analysis based on the identified haplotypes. Association testing using a generalized linear model (GLM) through FPKM measures generated 19 marker positions for *GBSSII* expression (Data S4), of which only three positions (one SNP and two InDels) overlapped from those from haplotyping (S5). The haplotypes representative of rice accessions identified from the FPKM measures were investigated for variations (Figure S6). The results revealed a significant difference between the M_3 haplotype (indica specific) and the C_3 haplotype (temperate japonica specific), indicating a higher level of *GBSSII* expression in temperate japonica. However, most of the observed variants (SNPs and InDels) found in these haplotypes (M_3 and C_3) were almost the same (Data S5). Despite the differences in the variant number among the other haplotypes, no significant difference was detected in their expression levels.

Many recent studies have shed-light on genetic diversity and molecular evolution in rice in terms of analyzing gene regions and improving varietal quality [62–65]. Of which, many starch synthesis genes have been studied on their genetics and genomics [47,66] but *GBSSII* is the only exception. Analysis of nucleotide diversity is important to highlight the functional importance of genomic regions [67] and it has also appeared to be useful to characterize local populations, representing the whole genome distribution of each population using π-values [68]. In our study, the nucleotide diversity analysis based on the classified ecotypes of 475 rice accessions revealed the lowest diversity values in japonica ecotypes (temperate japonica, 0.0029, and tropical japonica, 0.0030) despite the wild group representing its lowest diversity value (0.0045) on a variety basis, suggesting no significant differences between the wild group and any of the identified cultivated subgroups (Figure 3 and Table S3). However, the relative lowest diversity values of japonicas were assumed to be domestication signatures under selective sweeps due to genetic bottleneck effects. Previous findings suggested that lower genetic diversity was experienced in a recent bottleneck during the domestication selection of the waxy locus in the waxy rice genome [69–71]. On the other hand, the remaining cultivated subgroups, indica, aus, and admixture (representing higher diversity landrace, weedy, and bred groups in another way) were higher in nucleotide diversity than that of the wild group, suggesting

positive artificial selection due to the high π-value regions in each of these classified populations [68].

Tajima's *D*-values together with nucleotide diversity and linkage disequilibrium (LD), are the estimators of the domestication signatures of the crops compared to their wild progenitors [72–75]. Referring to Tajima's *D* values, the molecular evolution of the starch synthetic genes, such as *AGPL2*, *AGPS2b*, *SSIIa*, *SBEIIb*, *GBSSI*, and *ISA1*, has been estimated in rice endosperm under a standard neutral model [70,76–79]. The Tajima's *D* results in our study revealed a negative value for the wild group compared to the other cultivated variety groups of landrace, weedy, and bred. In the case of ecotype-based analysis, the wild group, as well as both japonicas (temperate and tropical), revealed negative Tajima's *D* estimates, compared to the other classified ecotypes of indica, aus, and the admixture. These negative *D* values in the japonica and wild groups occurred because of an excess of rare alleles within this gene region compared to expectations, indicating evidence for purifying selection. This consistent feature of selective force has been shown in a molecular evolution analysis of the *ALK* (*SSIIa*) gene using 321 rice accessions [77]. However, the landrace, weedy, and bred varieties (in terms of ecotypes by indica, admixture, and aus) expressed positive Tajima's *D* values within the *GBSSII* region, subjected to balancing selection through signifying higher observed diversity compared to those of the expected values.

Population structure is a systematic pattern used to measure the relatedness between individuals in large genomic data sets together with the dimension reduction technique such as principal component analysis (PCA) to visualize the principal components [80]. The structure clustering analysis demonstrated a clear separation of the classified populations (varieties or ecotypes, including the wild group) at most of K values (3–7). However, all of their internal subgroups were admixed, suggesting a diverse relationship among them due to their probable differing ancestral sources. PCA and the phylogenetic analysis also indicated similar features of the relationships among the classified populations where cultivated populations were interbred. Despite the clear separation of the four major groups by PCA, each group was composed of many classified populations. However, in variety-based PCA, group II showed a close association between the wild and bred groups which appeared to be in recent ancestors in the phylogenetic analysis by means of temperate japonica subclade between the two wild subclades. Forty-five haplotypes formed a network that infers their relatedness and estimates the recent evolutionary history of these diverse populations within the *GBSSII* region. Like the wild haplotypes, the cultivated haplotypes also expressed a diverse pattern, indicating close and similar genetics among their internal subgroups. Despite the vector-mediated connection between the wild and cultivated haplotypes, they were inferred to have a recent evolutionary history due to their associations with fewer mutational steps.

The empirical Bayesian inference of pairwise $F_{ST}$ and its distribution in the genome has been used to estimate the locus-specific pairwise $F_{ST}$ [81], and it is a widely used descriptive statistic in evolutionary genetics to provide insights based on genetic variations within and among populations [82]. In our study, both the highest and lowest genetic differentiation values of the *GBSSII* locus were observed between the cultivated ecotypes (0.8313 for the tropical japonica-aromatic pair and 0.0498 for temperate-tropical pair). However, the bred group was markedly distant from the wild group based on an $F_{ST}$-value of 0.1992, compared to the other varieties. A population structure and diversity analysis of red rice germplasm indicated a similar genetic differentiation feature among the populations, with pairwise $F_{ST}$ values of 0.108–0.207 [83]. Our findings also indicate a variety of greater differentiations among most of the pairs (60%) of classified ecotypes.

Identifying genes of interest is vital for conducting a successful breeding program and improving agronomic traits [84]. Functional alleles are very helpful to such breeding programs, for example, a functional SNP (G/A in exon 3) of the *SSIIIa* gene is useful for eating and cooking quality performance [84] and two waxy SNPs (G/T in intron 1 and A/C in exon 6) are useful for selecting the desired amylose content in rice varieties [13].

The haplotype analysis in our study provided many functional *GBSSII* alleles potentially suitable for future breeding studies by estimating the diverse structures of *GBSSII* and providing evolutionary insight among or between the classified rice groups.

## 5. Conclusions

Our findings provide collective and supportive information for the functional haplotypes of *GBSSII* in cultivated (13 haplotypes with 19 nsSNPs) and wild (21 haplotypes with 20 nsSNPs) rice accessions. These functional haplotypes identifying their specific alleles together with their detected regions can be of applicable and valuable genetic information for further studies of this gene performance in amylose biosynthesis (transient starch). Estimations of domestication signatures through a series of evolutionary analyses (genetic diversity, population genetics structure, and evolutionary relationships in phylogeny and differentiation analyses) provide different selective sweeps (balancing selection and purifying selection), which appear to be useful in molecular breeding programs for this gene's functions. These findings of our study can promote further research in a variety development with desired traits activated by this gene.

gene among the different ecotypes of the 475 Korean rice accessions, Figure S2: Nucleotide diversity analysis of the *GBSSII* (*Os07g0412100*) gene in the 475 Korean rice accessions by means of ecotypes (temperate japonica, indica, tropical japonica, aus, aromatic and admixture) together with the wild group. (A) Nucleotide diversity (π-value) representing the number of nucleotide variations within the *GBSSII* gene region at individual segregating sites in 1.5 kb sliding window. Cyan indicates the *GBSSII* gene region, and each colored line represents an individual ecotype. (B) Box plots representing the different distribution patterns of *GBSSII* genetic variations based on mean nucleotide diversity values among the classified ecotypes, Figure S3: Tajima's *D* values of *GBSSII* (*Os07g0412100*) in the 475 Korean rice accessions by means of ecotypes (temperate japonica, indica, tropical japonica, aus, aromatic and admixture) together with the wild. (A) Tajima's *D* values representing different individual segregating sites within the *GBSSII* gene region in 1.5 kb sliding window. Cyan indicates the *GBSSII* gene region, and different colored lines represent different rice ecotypes. (B) Box plots represent different distribution patterns of *GBSSII* genetic variations according to Tajima's *D* values among the ecotypes, Figure S4: A list of 38 haplotypes representing 55 InDel variations within *GBSSII* region of 475 rice accessions. Most of the variations were deletions (Dels) and very few positions were identified for insertions (Ins), Figure S5: Phylogenetic tree for the orthologous genes *GBSS1* and *GBSSII*. Scale bar indicates the proportion of sites changing along each branch. To characterize the relationship between *GBSS1* and *GBSSII*, these homologues in rice and other plant species were inferred using ORTHOFINDER [85]. Nineteen different plant species (most were from rice) were used from these public databases, https://rice-genome-hub.southgreen.fr/node/70/53; Ensembl Plants; http://rice.hzau.edu.cn; and https://rapdb.dna.affrc.go.jp, last accessed on 20 April 2021, Figure S6: Statistical analysis on the association between significant marker positions by a generalized linear model (GLM) using fragments per kilobase of exon per million (FPKM) measures and those detected under selected haplotypes by haplotype analysis. The hypothesis was performed by Scheffé test at alpha value by 0.05 (5%) and the selection of haplotypes were based on the rice accession numbers (≥5) each haplotype belonged to.

**Author Contributions:** Conceptualization, T.Z.M. and Y.-J.P.; methodology, formal analysis, and writing—original draft, T.Z.M.; resources, S.-H.C.; writing—review and editing, S.-H.C. and Y.-J.P.; supervision and project administration, Y.-J.P. All authors have read and agreed to the published version of the manuscript.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data is available within the article or Supplementary Materials.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Cho, Y.-G.; Kang, K.-K. *Functional Analysis of Starch Metabolism in Plants*; Multidisciplinary Digital Publishing Institute: Basel, Switzerland, 2020.
2. Martin, C.; Smith, A.M. Starch Biosynthesis. *Plant Cell* **1995**, *7*, 971.
3. Vrinten, P.L.; Nakamura, T. Wheat Granule-Bound Starch Synthase I and II Are Encoded by Separate Genes That Are Expressed in Different Tissues. *Plant Physiol.* **2000**, *122*, 255–264. [CrossRef]
4. Chen, H.-J.; Chen, J.-Y.; Wang, S.-J. Molecular Regulation of Starch Accumulation in Rice Seedling Leaves in Response to Salt Stress. *Acta Physiol. Plant.* **2008**, *30*, 135–142. [CrossRef]
5. Fasahat, P.; Rahman, S.; Ratnam, W. Genetic Controls on Starch Amylose Content in Wheat and Rice Grains. *J. Genet.* **2014**, *93*, 279–292. [CrossRef] [PubMed]
6. D'Hulst, C.; Wattebled, F.; Szydlowski, N. Starch biosynthesis in leaves and its regulation. In *Starch*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 211–237.
7. Dian, W.; Jiang, H.; Chen, Q.; Liu, F.; Wu, P. Cloning and Characterization of the Granule-Bound Starch Synthase II Gene in Rice: Gene Expression Is Regulated by the Nitrogen Level, Sugar and Circadian Rhythm. *Planta* **2003**, *218*, 261–268. [CrossRef] [PubMed]

8.  Wang, W.; Wei, X.; Jiao, G.; Chen, W.; Wu, Y.; Sheng, Z.; Hu, S.; Xie, L.; Wang, J.; Tang, S. GBSS-BINDING PROTEIN, Encoding a CBM48 Domain-containing Protein, Affects Rice Quality and Yield. *J. Integr. Plant Biol.* **2020**, *62*, 948–966. [CrossRef]

9.  Seung, D.; Soyk, S.; Coiro, M.; Maier, B.A.; Eicke, S.; Zeeman, S.C. PROTEIN TARGETING TO STARCH Is Required for Localising GRANULE-BOUND STARCH SYNTHASE to Starch Granules and for Normal Amylose Synthesis in Arabidopsis. *PLoS Biol.* **2015**, *13*, e1002080. [CrossRef] [PubMed]

10. Lohmeier-Vogel, E.M.; Kerk, D.; Nimick, M.; Wrobel, S.; Vickerman, L.; Muench, D.G.; Moorhead, G.B. Arabidopsis At5g39790 Encodes a Chloroplast-Localized, Carbohydrate-Binding, Coiled-Coil Domain-Containing Putative Scaffold Protein. *BMC Plant Biol.* **2008**, *8*, 1–14. [CrossRef]

11. Seung, D.; Boudet, J.; Monroe, J.; Schreier, T.B.; David, L.C.; Abt, M.; Lu, K.-J.; Zanella, M.; Zeeman, S.C. Homologs of PROTEIN TARGETING TO STARCH Control Starch Granule Initiation in Arabidopsis Leaves. *Plant Cell* **2017**, *29*, 1657–1677. [CrossRef]

12. Zhang, C.; Yang, Y.; Chen, S.; Liu, X.; Zhu, J.; Zhou, L.; Lu, Y.; Li, Q.; Fan, X.; Tang, S. A Rare Waxy Allele Coordinately Improves Rice Eating and Cooking Quality and Grain Transparency. *J. Integr. Plant Biol.* **2021**, *63*, 889–901. [CrossRef]

13. Chen, M.-H.; Bergman, C.J.; Pinson, S.R.; Fjellstrom, R.G. Waxy Gene Haplotypes: Associations with Pasting Properties in an International Rice Germplasm Collection. *J. Cereal Sci.* **2008**, *48*, 781–788. [CrossRef]

14. Zhang, C.; Zhu, J.; Chen, S.; Liu, Q. Wxlv, the Ancestral Allele of Rice Waxy Gene. *Multidiscip. Digit. Publ. Inst. Proc.* **2020**, *36*, 140. [CrossRef]

15. Luo, M.; Shi, Y.; Yang, Y.; Zhao, Y.; Zhang, Y.; Shi, Y.; Kong, M.; Li, C.; Feng, Z.; Fan, Y. Sequence Polymorphism of the Waxy Gene in Waxy Maize Accessions and Characterization of a New Waxy Allele. *Sci. Rep.* **2020**, *10*, 1–10. [CrossRef] [PubMed]

16. Mishra, A.; Sharma, V.; Rahim, M.S.; Sonah, H.; Pal, D.; Mantri, S.; Sharma, T.R.; Roy, J. Genotyping-by-Sequencing Based QTL Mapping Identified a Novel Waxy Allele Contributing to High Amylose Starch in Wheat. *Euphytica* **2021**, *217*, 1–14. [CrossRef]

17. Chen, X.; Shao, S.; Chen, M.; Hou, C.; Yu, X.; Xiong, F. Morphology and Physicochemical Properties of Starch from Waxy and Non-waxy Barley. *Starch-Stärke* **2020**, *72*, 1900206. [CrossRef]

18. Karlström, A.; Calle, F.; Salazar, S.; Morante, N.; Dufour, D.; Ceballos, H. Biological Implications in Cassava for the Production of Amylose-Free Starch: Impact on Root Yield and Related Traits. *Front. Plant Sci.* **2016**, *7*, 604. [CrossRef]

19. Aiemnaka, P.; Wongkaew, A.; Chanthaworn, J.; Nagashima, S.K.; Boonma, S.; Authapun, J.; Jenweerawat, S.; Kongsila, P.; Kittipadakul, P.; Nakasathien, S. Molecular Characterization of a Spontaneous Waxy Starch Mutation in Cassava. *Crop Sci.* **2012**, *52*, 2121–2130. [CrossRef]

20. Zhang, H.; Jang, S.-G.; Lar, S.M.; Lee, A.-R.; Cao, F.-Y.; Seo, J.; Kwon, S.-W. Genome-Wide Identification and Genetic Variations of the Starch Synthase Gene Family in Rice. *Plants* **2021**, *10*, 1154. [CrossRef]

21. Singh, N.; Singh, B.; Rai, V.; Sidhu, S.; Singh, A.K.; Singh, N.K. Evolutionary Insights Based on SNP Haplotypes of Red Pericarp, Grain Size and Starch Synthase Genes in Wild and Cultivated Rice. *Front. Plant Sci.* **2017**, *8*, 972. [CrossRef] [PubMed]

22. Henry, R.J. Next-Generation Sequencing for Understanding and Accelerating Crop Domestication. *Brief. Funct. Genom.* **2012**, *11*, 51–56. [CrossRef]

23. Shavrukov, Y.; Suchecki, R.; Eliby, S.; Abugalieva, A.; Kenebayev, S.; Langridge, P. Application of Next-Generation Sequencing Technology to Study Genetic Diversity and Identify Unique SNP Markers in Bread Wheat from Kazakhstan. *BMC Plant Biol.* **2014**, *14*, 258. [CrossRef] [PubMed]

24. Zhou, X.; Bai, X.; Xing, Y. A Rice Genetic Improvement Boom by Next-Generation Sequencing. *Curr. Issues Mol. Biol.* **2018**, *27*, 109–126. [CrossRef]

25. Feng, P.; Zeng, T.; Yang, H.; Chen, G.; Du, J.; Chen, L.; Shen, J.; Tao, Z.; Wang, P.; Yang, L. Whole-Genome Resequencing Provides Insights into the Population Structure and Domestication Signatures of Ducks in Eastern China. *BMC Genom.* **2021**, *22*, 1–13. [CrossRef]

26. Trung, K.H.; Nguyen, T.K.; Khuat, H.B.T.; Nguyen, T.D.; Khanh, T.D.; Xuan, T.D.; Nguyen, X.-H. Whole Genome Sequencing Reveals the Islands of Novel Polymorphisms in Two Native Aromatic Japonica Rice Landraces from Vietnam. *Genome Biol. Evol.* **2017**, *9*, 1816–1820. [CrossRef] [PubMed]

27. Bindusree, G.; Natarajan, P.; Kalva, S.; Madasamy, P. Whole Genome Sequencing of Oryza Sativa L. Cv. Seeragasamba Identifies a New Fragrance Allele in Rice. *PLoS ONE* **2017**, *12*, e0188920. [CrossRef] [PubMed]

28. Sabeti, P.C.; Schaffner, S.F.; Fry, B.; Lohmueller, J.; Varilly, P.; Shamovsky, O.; Palma, A.; Mikkelsen, T.S.; Altshuler, D.; Lander, E.S. Positive Natural Selection in the Human Lineage. *Science* **2006**, *312*, 1614–1620. [CrossRef] [PubMed]

29. Kim, H.R.; Sa, K.J.; Nam-Gung, M.; Park, K.J.; Ryu, S.-H.; Mo, C.Y.; Lee, J.K. Genetic Characterization and Association Mapping in Near-Isogenic Lines of Waxy Maize Using Seed Characteristics and SSR Markers. *Genes Genom.* **2021**, *43*, 79–90.

30. Kim, K.-W.; Chung, H.-K.; Cho, G.-T.; Ma, K.-H.; Chandrabalan, D.; Gwag, J.-G.; Kim, T.-S.; Cho, E.-G.; Park, Y.-J. PowerCore: A Program Applying the Advanced M Strategy with a Heuristic Search for Establishing Core Sets. *Bioinform.* **2007**, *23*, 2155–2162. [CrossRef]

31. Kim, T.-S.; He, Q.; Kim, K.-W.; Yoon, M.-Y.; Ra, W.-H.; Li, F.P.; Tong, W.; Yu, J.; Oo, W.H.; Choi, B. Genome-Wide Resequencing of KRICE_CORE Reveals Their Potential for Future Breeding, as Well as Functional and Evolutionary Studies in the Post-Genomic Era. *BMC Genom.* **2016**, *17*, 408. [CrossRef]

32. Doyle, J.J.; Doyle, J.L. Isolation Ofplant DNA from Fresh Tissue. *Focus* **1990**, *12*, 39–40.

33. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T. The Variant Call Format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef]

34. Picard Toolkit. Broad Institute, GitHub Repository. 2019. Available online: http://broadinstitute.github.io/picard (accessed on 20 April 2021).

35. DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; Del Angel, G.; Rivas, M.A.; Hanna, M. A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data. *Nat. Genet.* **2011**, *43*, 491. [CrossRef] [PubMed]

36. Van der Auwera, G.A.; O'Connor, B.D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*; O'Reilly Media: Sebastopol, CA, USA, 2020.

37. Bradbury, P.J.; Zhang, Z.; Kroon, D.E.; Casstevens, T.M.; Ramdoss, Y.; Buckler, E.S. TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples. *Bioinformatics* **2007**, *23*, 2633–2635. [CrossRef]

38. Raj, A.; Stephens, M.; Pritchard, J.K. FastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics* **2014**, *197*, 573–589. [CrossRef] [PubMed]

39. Francis, R.M. Pophelper: An R Package and Web App to Analyse and Visualize Population Structure. *Mol. Ecol. Resour.* **2017**, *17*, 27–32. [CrossRef]

40. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef] [PubMed]

41. Rozas, J.; Ferrer-Mata, A.; Sánchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.E.; Sánchez-Gracia, A. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* **2017**, *34*, 3299–3302. [CrossRef]

42. Leigh, J.W.; Bryant, D. POPART: Full-Feature Software for Haplotype Network Construction. *Methods Ecol. Evol.* **2015**, *6*, 1110–1116. [CrossRef]

43. Clement, M.; Posada, D.; Crandall, K.A. TCS: A Computer Program to Estimate Gene Genealogies. *Mol. Ecol.* **2000**, *9*, 1657–1659. [CrossRef]

44. Seung, D. Amylose in Starch: Towards an Understanding of Biosynthesis, Structure and Function. *New Phytol.* **2020**, *228*, 1490–1504. [CrossRef]

45. Nakamura, T.; Vrinten, P.; Hayakawa, K.; Ikeda, J. Characterization of a Granule-Bound Starch Synthase Isoform Found in the Pericarp of Wheat. *Plant Physiol.* **1998**, *118*, 451–459. [CrossRef] [PubMed]

46. Dufayard, J.-F.; Duret, L.; Penel, S.; Gouy, M.; Rechenmann, F.; Perrière, G. Tree Pattern Matching in Phylogenetic Trees: Automatic Search for Orthologs or Paralogs in Homologous Gene Sequence Databases. *Bioinformatics* **2005**, *21*, 2596–2603. [CrossRef]

47. Kharabian-Masouleh, A.; Waters, D.L.; Reinke, R.F.; Henry, R.J. Discovery of Polymorphisms in Starch-related Genes in Rice Germplasm by Amplification of Pooled DNA and Deeply Parallel Sequencing. *Plant Biotechnol. J.* **2011**, *9*, 1074–1085. [CrossRef]

48. Lai, K.; Duran, C.; Berkman, P.J.; Lorenc, M.T.; Stiller, J.; Manoli, S.; Hayden, M.J.; Forrest, K.L.; Fleury, D.; Baumann, U. Single Nucleotide Polymorphism Discovery from Wheat Next-generation Sequence Data. *Plant Biotechnol. J.* **2012**, *10*, 743–749. [CrossRef]

49. Lu, K.; Wei, L.; Li, X.; Wang, Y.; Wu, J.; Liu, M.; Zhang, C.; Chen, Z.; Xiao, Z.; Jian, H. Whole-Genome Resequencing Reveals Brassica Napus Origin and Genetic Loci Involved in Its Improvement. *Nature Commun.* **2019**, *10*, 1–12. [CrossRef]

50. Jayaswall, K.; Sharma, H.; Bhandawat, A.; Sagar, R.; Yadav, V.K.; Sharma, V.; Mahajan, V.; Roy, J.; Singh, M. Development of Intron Length Polymorphic (ILP) Markers in Onion (Allium Cepa L.), and Their Cross-Species Transferability in Garlic (*A. Sativum* L.) and Wild Relatives. *Genet. Resour. Crop Evol.* **2019**, *66*, 1379–1388. [CrossRef]

51. Mammadov, J.; Aggarwal, R.; Buyyarapu, R.; Kumpatla, S. SNP Markers and Their Impact on Plant Breeding. *Int. J. Plant Genom.* **2012**, *2012*. [CrossRef]

52. Shorinola, O.; Balcárková, B.; Hyles, J.; Tibbits, J.F.; Hayden, M.J.; Holušova, K.; Valárik, M.; Distelfeld, A.; Torada, A.; Barrero, J.M. Haplotype Analysis of the Pre-Harvest Sprouting Resistance Locus Phs-A1 Reveals a Causal Role of TaMKK3-A in Global Germplasm. *Front. Plant Sci.* **2017**, *8*, 1555. [CrossRef] [PubMed]

53. Min, M.-H.; Maung, T.Z.; Cao, Y.; Phitaktansakul, R.; Lee, G.-S.; Chu, S.-H.; Kim, K.-W.; Park, Y.-J. Haplotype Analysis of BADH1 by Next-Generation Sequencing Reveals Association with Salt Tolerance in Rice during Domestication. *Int. J. Mol. Sci.* **2021**, *22*, 7578. [CrossRef]

54. Perez, C.M.; Palmiano, E.P.; Baun, L.C.; Juliano, B.O. Starch Metabolism in the Leaf Sheaths and Culm of Rice. *Plant Physiol.* **1971**, *47*, 404–408. [CrossRef] [PubMed]

55. Lian, S.; Tanaka, A. Behaviour of Photosynthetic Products Associated with Growth and Grain Production in the Rice Plant. *Plant Soil* **1967**, *26*, 333–347. [CrossRef]

56. TOGARI, Y.; SATO, K. Studies on the Production and Behavior of Carbohydrates in Rice Plant: II. On the Accumulation and Distribution of Starches in the Organs of Rice Plant with Its Development of Growth. *Jpn. J. Crop Sci.* **1954**, *22*, 98–99. [CrossRef]

57. Yoshida, S.; Ahn, S.B. The Accumulation Process of Carbohydrate in Rice Varieties in Relation to Their Response to Nitrogen in the Tropics. *Soil Sci. Plant Nutr.* **1968**, *14*, 153–161. [CrossRef]

58. Prathap, V.; Ali, K.; Singh, A.; Vishwakarma, C.; Krishnan, V.; Chinnusamy, V.; Tyagi, A. Starch Accumulation in Rice Grains Subjected to Drought during Grain Filling Stage. *Plant Physiol. Biochem.* **2019**, *142*, 440–451.

59. Sugimura, Y.; Michiyama, H.; Hirano, T. Involvement of α-Amylase Genes in Starch Degradation in Rice Leaf Sheaths at the Post-Heading Stage. *Plant Prod. Sci.* **2015**, *18*, 277–283. [CrossRef]

60. Kukurba, K.R.; Montgomery, S.B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015**, *2015*. [CrossRef]

61. Zhao, Y.; Li, M.-C.; Konaté, M.M.; Chen, L.; Das, B.; Karlovich, C.; Williams, P.M.; Evrard, Y.A.; Doroshow, J.H.; McShane, L.M. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-Seq Data from the NCI Patient-Derived Models Repository. *J. Translat. Med.* **2021**, *19*, 1–15. [CrossRef]
62. Cuevas, R.P.; Fitzgerald, M.A. Genetic Diversity of Rice Grain Quality. *Genet. Divers. Plants* **2012**, 286–310.
63. Moonsap, P.; Laksanavilat, N.; Sinumporn, S.; Tasanasuwan, P.; Kate-Ngam, S.; Jantasuriyarat, C. Genetic Diversity of Indo-China Rice Varieties Using ISSR, SRAP and InDel Markers. *J. Genet.* **2019**, *98*, 1–11. [CrossRef]
64. Muto, C.; Ishikawa, R.; Olsen, K.M.; Kawano, K.; Bounphanousay, C.; Matoh, T.; Sato, Y.-I. Genetic Diversity of the Wx Flanking Region in Rice Landraces in Northern Laos. *Breed. Sci.* **2016**, 16032. [CrossRef] [PubMed]
65. Eris, F.R.; Kartina, A.M.; Maryani, Y.; Aryani, T. Genetic Diversity of Red Rice Varieties Originating from West Java and Banten Based on SSR Marker Related to Palatability. In *Earth and Environmental Science*; IOP Conference Series; IOP Publishing: Philadelphia, PA, USA, 2020; Volume 482, p. 012037.
66. Bao, J.S.; Corke, H.; Sun, M. Nucleotide Diversity in Starch Synthase IIa and Validation of Single Nucleotide Polymorphisms in Relation to Starch Gelatinization Temperature and Other Physicochemical Properties in Rice (*Oryza Sativa* L.). *Theor. Appl. Genet.* **2006**, *113*, 1171–1183. [CrossRef]
67. Tatarinova, T.V.; Chekalin, E.; Nikolsky, Y.; Bruskin, S.; Chebotarov, D.; McNally, K.L.; Alexandrov, N. Nucleotide Diversity Analysis Highlights Functionally Important Genomic Regions. *Sci. Rep.* **2016**, *6*, 1–12. [CrossRef]
68. Tam, N.T.; Dwiyanti, M.S.; Koide, Y.; Nagano, A.J.; Ky, H.; Tin, H.Q.; Hien, N.L.; Dung, L.V.; Kishima, Y. Profiling SNP and Nucleotide Diversity to Characterize Mekong Delta Rice Landraces in Southeast Asian Populations. *Plant Genome* **2019**, *12*, 190042. [CrossRef]
69. Olsen, K.M.; Purugganan, M.D. Molecular Evidence on the Origin and Evolution of Glutinous Rice. *Genetics* **2002**, *162*, 941–950. [CrossRef] [PubMed]
70. Wei, X.; Qiao, W.-H.; Chen, Y.-T.; Wang, R.-S.; Cao, L.-R.; Zhang, W.-X.; Yuan, N.-N.; Li, Z.-C.; Zeng, H.-L.; Yang, Q.-W. Domestication and Geographic Origin of O Ryza Sativa in C Hina: Insights from Multilocus Analysis of Nucleotide Variation of O. Sativa and O. Rufipogon. *Mol. Ecol.* **2012**, *21*, 5073–5087. [CrossRef] [PubMed]
71. Yamanaka, S.; Nakamura, I.; Watanabe, K.N.; Sato, Y.-I. Identification of SNPs in the Waxy Gene among Glutinous Rice Cultivars and Their Evolutionary Significance during the Domestication Process of Rice. *Theor. Appl. Genet.* **2004**, *108*, 1200–1204. [CrossRef]
72. Huang, X.; Kurata, N.; Wang, Z.-X.; Wang, A.; Zhao, Q.; Zhao, Y.; Liu, K.; Lu, H.; Li, W.; Guo, Y. A Map of Rice Genome Variation Reveals the Origin of Cultivated Rice. *Nature* **2012**, *490*, 497–501. [CrossRef]
73. Reif, J.C.; Zhang, P.; Dreisigacker, S.; Warburton, M.L.; van Ginkel, M.; Hoisington, D.; Bohn, M.; Melchinger, A.E. Wheat Genetic Diversity Trends during Domestication and Breeding. *Theor. Appl. Genet.* **2005**, *110*, 859–864. [CrossRef]
74. Wright, S.I.; Bi, I.V.; Schroeder, S.G.; Yamasaki, M.; Doebley, J.F.; McMullen, M.D.; Gaut, B.S. The Effects of Artificial Selection on the Maize Genome. *Science* **2005**, *308*, 1310–1314. [CrossRef]
75. Lam, H.-M.; Xu, X.; Liu, X.; Chen, W.; Yang, G.; Wong, F.-L.; Li, M.-W.; He, W.; Qin, N.; Wang, B. Resequencing of 31 Wild and Cultivated Soybean Genomes Identifies Patterns of Genetic Diversity and Selection. *Nat. Genet.* **2010**, *42*, 1053–1059. [CrossRef] [PubMed]
76. Olsen, K.M.; Caicedo, A.L.; Polato, N.; McClung, A.; McCouch, S.; Purugganan, M.D. Selection under Domestication: Evidence for a Sweep in the Rice Waxy Genomic Region. *Genetics* **2006**, *173*, 975–983. [CrossRef]
77. Zhou, Y.; Zheng, H.; Wei, G.; Zhou, H.; Han, Y.; Bai, X.; Xing, Y.; Han, Y. Nucleotide Diversity and Molecular Evolution of the ALK Gene in Cultivated Rice and Its Wild Relatives. *Plant Mol. Biol. Rep.* **2016**, *34*, 923–930. [CrossRef]
78. Yu, G.; Olsen, K.M.; Schaal, B.A. Molecular Evolution of the Endosperm Starch Synthesis Pathway Genes in Rice (Oryza Sativa L.) and Its Wild Ancestor, O. Rufipogon L. *Mol. Biol. Evol.* **2011**, *28*, 659–671. [CrossRef]
79. Zhou, H.; Xia, D.; Zhao, D.; Li, Y.; Li, P.; Wu, B.; Gao, G.; Zhang, Q.; Wang, G.; Xiao, J. The Origin of Wxla Provides New Insights into the Improvement of Grain Quality in Rice. *J. Integr. Plant Biol.* **2021**, *63*, 878–888. [CrossRef] [PubMed]
80. Li, H.; Ralph, P. Local PCA Shows How the Effect of Population Structure Differs along the Genome. *Genetics* **2019**, *211*, 289–304. [CrossRef]
81. Kitada, S.; Kitakado, T.; Kishino, H. Empirical Bayes Inference of Pairwise F ST and Its Distribution in the Genome. *Genetics* **2007**, *177*, 861–873. [CrossRef]
82. Holsinger, K.E.; Weir, B.S. Genetics in Geographically Structured Populations: Defining, Estimating and Interpreting F ST. *Nat. Rev. Genet.* **2009**, *10*, 639–650. [CrossRef]
83. Islam, M.Z.; Khalequzzaman, M.; Prince, M.; Siddique, M.A.; Rashid, E.; Ahmed, M.S.U.; Pittendrigh, B.R.; Ali, M.P. Diversity and Population Structure of Red Rice Germplasm in Bangladesh. *PLoS ONE* **2018**, *13*, e0196096. [CrossRef] [PubMed]
84. Kim, M.-S.; Yang, J.-Y.; Yu, J.-K.; Lee, Y.; Park, Y.-J.; Kang, K.-K.; Cho, Y.-G. Breeding of High Cooking and Eating Quality in Rice by Marker-Assisted Backcrossing (MABc) Using KASP Markers. *Plants* **2021**, *10*, 804. [CrossRef]
85. Emms, D.M.; Kelly, S. OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biol.* **2019**, *20*, 1–14. [CrossRef] [PubMed]