Divergent Evolutionary Pattern of Starch Biosynthetic Pathway Genes in Grasses and Dicots

Chun Li,1,2 Qi-Gang Li,1 Jim M. Dunwell,3 and Yuan-Ming Zhang*,1
1State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Agriculture, Department of Crop Genetics and Breeding, Nanjing Agricultural University, Nanjing, People’s Republic of China
2Henan Sesame Research Center, Henan Academy of Agricultural Sciences, Zhengzhou, People’s Republic of China
3School of Biological Sciences, University of Reading, Whiteknights, Reading, United Kingdom
*Corresponding author: E-mail: soyzhang@njau.edu.cn.
Associate editor: Neelima Sinha

Abstract
Starch is the most widespread and abundant storage carbohydrate in crops and its production is critical to both crop yield and quality. In regard to the starch content in the seeds of crop plants, there is a distinct difference between grasses (Poaceae) and dicots. However, few studies have described the evolutionary pattern of genes in the starch biosynthetic pathway in these two groups of plants. In this study, therefore, an attempt was made to compare evolutionary rate, gene duplication, and selective pattern of the key genes involved in this pathway between the two groups, using five grasses and five dicots as materials. The results showed 1) distinct differences in patterns of gene duplication and loss between grasses and dicots; duplication in grasses mainly occurred before the divergence of grasses, whereas duplication mostly occurred in individual species within the dicots; there is less gene loss in grasses than in dicots, 2) a considerably higher evolutionary rate in grasses than in dicots in most gene families analyzed, and 3) evidence of a different selective pattern between grasses and dicots; positive selection may have occurred asymmetrically in grasses in some gene families, for example, ADP-glucose pyrophosphorylase small subunit. Therefore, we deduced that gene duplication contributes to, and a higher evolutionary rate is associated with, the higher starch content in grasses. In addition, two novel aspects of the evolution of the starch biosynthetic pathway were observed.

Key words: Starch synthesis pathway, gene duplication, evolutionary rate, divergent selection, grasses, dicots.

Introduction
Monocots and dicots are two main classes of angiosperms. Dicots include soybean, oilseed rape, and related species, whereas monocots include wheat, corn, and rice. The distinct differences in morphological and physiological features between monocots and dicots, such as the structure and chemical components of the seed, have been described by Raven and Johnson (2002). The evolutionary divergence of the genetic factors underlying these features is of great interest. Grasses (Poaceae), including the most important cereal crops, are typical monocots. Monocot seeds are rich in starch, whereas dicot seeds are generally rich in lipid or protein. Therefore, it can be expected that there has been evolutionary divergence in the genes controlling starch biosynthesis between these two groups of plants.

Starch is the most widespread and abundant storage carbohydrate in crops and its production is critical to both crop yield and quality (Umemoto et al. 2002). To date, most studies relating to the genes of starch biosynthesis have focused on two aspects, namely molecular genetics and evolution (Whitt et al. 2002; Tetlow et al. 2004; Han et al. 2007; Wu et al. 2008; Comparot-Moss and Denyer 2009; Zeeman et al. 2010). First, in the area of molecular genetics, the pathway of starch biosynthesis and the genes encoding the various enzymes have been shown to be remarkably conserved in dicots and monocots; and four enzymes, which play a major role in starch biosynthesis, are suggested to be unique in the two plant kingdoms. These enzymes are ADP-glucose pyrophosphorylase (AGPase), starch synthase (SS), starch branching enzyme (SBE) and debranching enzyme (DBE). Each of them is further divided into different subunits or subtypes, for example, two subunits (AGPase large subunit [AGPL] and AGPase small subunit [AGPS]) for AGPase, five classes (granule-bound starch synthase [GBSS and SSI-IV]) for SS, two main classes (classes I and II) for SBE, and two types (isoamylase [ISA] and pullulanase [PUL]) for DBE (Tetlow et al. 2004; Comparot-Moss and Denyer 2009; Zeeman et al. 2010).

Second, various evolutionary and domestication studies of the genes in the starch biosynthetic pathway have been presented previously. For example, Whitt et al. (2002) surveyed nucleotide diversity of six genes, namely Amylose extender1 (Ae1), Brittle2 (Bt2), Shrunken1 (Sh1), Shrunken-ken2 (Sh2), Sugary (Su), and Waxy (Wx), which play major roles in the starch biosynthetic pathway in maize, Zea mays ssp. parviglumis, and uncovered evidence of positive selection on three of these genes (Ae1, Su1, and Bt2). Han et al. (2007) analyzed the phylogeny of the SBEII gene and found different duplication patterns between grasses and dicots. Wu et al. (2008) investigated the fate of duplicated genes following the ancient whole-genome duplication (WGD) in rice and...
showed that genes involved in starch synthesis are preferentially retained in rice compared with *Arabidopsis*. In the most recent study, it was shown that the diversification of GBSS genes is mainly associated with genome-wide duplication throughout the evolutionary history of monocots and eudicots (Cheng et al. 2012).

All these evolutionary studies imply that the evolutionary pattern of the starch pathway in grasses is different from that in dicots. However, there are few comprehensive studies to confirm this viewpoint. In this study, therefore, we compared evolutionary rate, gene duplication, and selective pattern of the key genes involved in the starch biosynthetic pathway between grasses and dicots, using five grasses and five dicots as materials. The results should shed light on how evolutionary divergence contributes to differences in morphological and physiological features between these two groups of plants.

**Materials and Methods**

**Retrieval of Sequences**

Sequences were collected using the method described by Tatusov et al. (2000) with a slight modification. Briefly, our method includes the following steps:

1. Protein-coding transcripts of *Brachypodium distachyon* (JGI v1.0, International Brachypodium Initiative), *Oryza sativa* (MSU Release 6.0, OuYang et al. 2007), *Setaria italica* (version 1.1, DoE Joint Genome Institute), *Z. mays* (5a release, Maize Genome Project), Sorghum bicolor (v1.0 release, Paterson et al. 2009), *Arabidopsis thaliana* (TAIR release 9, The Arabidopsis Genome Initiative), *Glycine max* (Glyma1.0, Soybean Genome Project, DoE Joint Genome Institute), *Medicago truncatula* (Mt3.0, M. truncatula genome sequencing project), *Populus trichocarpa* (v2.0 release, The Joint Genome Institute), and *Ricinus communis* (release v0.1, Chan et al. 2010) were downloaded from JGI ([http://www.phytozone.net/](http://www.phytozone.net/)) or maize sequence ([http://www.maizequence.org](http://www.maizequence.org)) and used to construct a local BLAST database using BLAST 2.2.24;

2. An all-against-all protein sequence comparison was performed;

3. In all the comparisons carried out in step 2, the ones with *Arabidopsis AGPL1* as a query were identified, and the obvious paralogs were collapsed;

4. All interspecies Best Hits (BeTs) of AtAGPL1 and its paralogs were detected;

5. Steps 3 and 4 were repeated with the resulting sequences as secondary BLASTp queries until no new sequence was found;

6. Clusters of Orthologous Groups (COGs) were formed using the protocol of Tatusov et al. (2000); and

7. COGs with *Arabidopsis AGPS1*, GBSS, SSI, SSII, SSIII, SSIV, BEI, BEII, ISA, ISAI, ISAII, ISAIII, and PUL as a query were formed in the same way.

This approach was based on the consistency between genome-specific best hits, rather than the absolute level of similarity, and therefore it allowed the detection of orthologs among both slowly and rapidly evolving genes.

**Estimation of Gene Duplication and Loss**

Gene duplication and loss within each gene family were analyzed principally by checking each subfamily manually. Each gene family is assumed to have at least two subclades (the grass and dicot subfamily) and at least five members from each species in each subfamily (supplementary fig. S3, Supplementary Material online). If there are two or more grass or dicot subclades, or if these subclades have two or more members from the same species, it was concluded that there is one or more duplications. If there is a member missing from any species within the grass or dicot subclade, it was concluded that this is a gene loss. Every gene loss was further confirmed by searching GenBank.

**Phylogenetic Analyses**

The codon sequences were first aligned using the PRANK codon model ([Löytynoja and Goldman 2005]) and then translated into amino acid sequences. When necessary, sequences were further edited and aligned manually. Phylogenetic tree reconstruction was performed with both Neighbor Joining (NJ) and Bayesian approaches using the amino acid sequences. In the NJ method, the phylogenetic analyses were conducted using the Molecular Evolutionary Genetics Analysis 4.0 program (Tamura et al. 2007). The parameter setups were as follows: model: p-distance; bootstrap: 1,000 replicates; and gap/missing data: pairwise deletion.

In the Bayesian method, the analyses were conducted using MrBayes v3.1 ([Ronquist and Huelsenbeck 2003]) with the Jones, Taylor, Thornton substitution model (Jones et al. 1992), four chains, one million generations, and two runs. Trees were sampled every 100 generations, discarding a burn-in of 250,000 generations.

**Estimation of dS/dN Ratios**

The codon sequences were aligned using the PRANK codon model ([Löytynoja and Goldman 2005]), alignment gaps were manually deleted, and then used for the following calculations. The ratio of nonsynonymous substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous site (dS) (ω value) of homologous gene pairs was computed with the maximum likelihood method in Codeml from the PAML package (v4.4; Yang 2007). Saturation effects were avoided by discarding the gene pairs for which dS > 2 (Ramsay et al. 2009).

To test for variation of the ω ratio among different branches in gene trees, a branch-specific model was used and conducted in Codeml. This type of branch-specific model allows the ω ratio to vary among branches in the phylogeny, and it can be used to test whether there are different ω values on particular lineages (Yang 1998); thus, this model can be compared with the one-ratio model that assumes a constant ω value across all branches using the likelihood ratio test (LRT). In these calculations, truncated sequences were removed and tree topologies were adjusted manually.
Detection of Positively Selected Sites

The data sets used in the selection analysis were further used in the selection analysis conducted in the program Fitmodel 0.5.3 (Guindon et al. 2004). Fitmodel is similar to PAML (Yang 2007) and uses maximum likelihood for estimating parameters of sequence evolution. Moreover, it also estimates switching parameters, such that Fitmodel allows changes between selection patterns to occur through time (i.e., switches between selection patterns in the phylogenetic tree at individual sites) (Guindon et al. 2004). The models M3 and M3 + 1, described in Yang and Nielsen (2002) and Guindon et al. (2004), were employed in this analysis.

Results

Data Collection and Phylogenetic Analyses

Fourteen COGs were obtained and are shown in figure 1 and supplementary figure S1 (Supplementary Material online). Homologs of AGPS, AGPL, GBSS, SSII, SSIII, and SSIV are included in a single COG (fig. 1 and supplementary fig. S1, Supplementary Material online); homologs of BE are separated into three COGs (supplementary fig. S1, Supplementary Material online), corresponding to classes I and II and a third class (Wang et al. 2010), respectively; and homologs of DBE are separated into four COGS (supplementary fig. S1, Supplementary Material online), corresponding to ISAI-III and PUL, respectively.

Phylogenetic tree reconstruction was performed by the Bayesian and NJ methods for each of the COGs; both methods yielded identical topologies (data not shown). Genes from grasses (or dicots) form distinct subfamily clades with high support (mostly >90% bootstrap value) in all the phylogenetic trees (fig. 2 and supplementary fig. S2, Supplementary Material online). In the trees of AGPL type 3, AGPS, GBSS, SSII, SSIII, and SBEII, different gene classes also form distinct subfamily clades. For example, SBEIIa and SBEIIb form grass subfamilies 1 and 2, respectively (supplementary fig. S2, Supplementary Material online). These results are supported by previous results from different species and quantities of sequence, for example, five clades in the AGPL tree in higher plants (Georgelis et al. 2007). In each grass or dicot subfamily clade, the gene tree is largely concordant with the species tree (Soltis 1999) (supplementary fig. S3, Supplementary Material online).

Distinct Differences for Gene Duplication and Patterns of Loss between Grasses and Dicots

Gene duplication can be distinguished into recent and old duplications according to the time it occurred (Ohno 1970). In this study, the grass- (duplication before the grasses divergence) and dicot-specific duplications (duplication before the dicots divergence) were termed as old duplications, and the species-specific (duplication within a species) ones were termed as recent duplications. Among 14 gene families, each one was separately analyzed in detail for the recent and old duplications. The results are shown in table 1, figure 2, and supplementary figure S2 (Supplementary Material online). As for the old duplication, it occurred approximately seven times in six gene families (AGPL, AGPS,
GBSS, SSII, SSIII, and SBEII) in the grasses and once in the AGPS gene family in the dicots. Considering the recent duplication, it occurred approximately 3 times in 3 gene families (AGPS, SSIII, and SSIV) in the grasses and more than 20 times in most of the gene families in the dicots (except SBEIII, ISAIII, and PUL). Note that in the dicots, most evidence of recent duplication was found in G. max.

Considering gene losses, those specific to the grass species only occurred in the AGPL and SSII gene families and those specific to the dicot species occurred in nearly all the families analyzed and the majority of the missing members are from the species M. truncatula (table 1, fig. 2, and supplementary fig. S2, Supplementary Material online). Obviously, there are fewer gene losses in the grasses.

As a result, there are distinct differences in patterns of gene duplication and loss between grasses and dicots. These results are consistent with the observations that there are more copies of starch synthesis genes in Z. mays and O. sativa than in A. thaliana (Yan et al. 2009), the genes involved in synthesis and catabolism of saccharides are preferentially retained in rice (Wu et al. 2008), and the proposition that the genes encoding the starch pathway originated from a WGD event in an early ancestor of the grasses (Comparot-Moss and Denyer 2009; Yan et al. 2009).

Starch Pathway Genes Evolve Rapidly in Grasses

To determine whether the starch pathway genes are under different evolutionary constraints in grasses and dicots, the \( \omega \) values for the above genes were calculated. We applied two methods: the pairwise comparison approach and the branch-specific model.

In the pairwise comparison approach, we calculated the pairwise \( \omega \) values within each grass or dicot subfamily and compared their averages between two subfamilies. The result showed that the average \( \omega \) values of nine grass subfamilies (subfamily of AGPL type 1, subfamily 2 of AGPL type 3, subfamilies 1 and 2 of AGPS, subfamilies 2 and 3 of SSII, subfamily 1 of SSIII and SSIV, and subfamily 2 of SBEII) are significantly higher than that of the corresponding subfamilies composed of the homologs from the dicots; and three grass subfamilies (subfamily 1 of AGPL type 3 and 4 and subfamily 2 of GBSS) show the opposite result (fig. 3).
However, there is a shortcoming in the pairwise comparison approach: It does not take into account the evolutionary rate of the most recent common ancestor of the subfamily (e.g., branch A in fig. 2). To overcome this issue, we applied the branch-specific model to the gene trees to check the changes in the evolutionary constraints across different subfamilies. In this approach, the model assumes two or more $\omega$ ratios, which correspond to the number of subfamilies. As a result, this model was favored over the one-ratio model by the LRT ($P < 0.05$) in all the gene trees except for $SSI$ and $ISAi$, and the estimated $\omega$ values of the subfamilies confirm the result of the pairwise comparison approach. More importantly, these results show that some of the grass subfamilies, which did not exhibit a higher $\omega$ value than their dicot subfamilies in the pairwise comparison approach, exhibit a higher $\omega$ value when the most recent common ancestor is taken into account (table 2).

Therefore, all the grass subfamilies examined, except for subfamily 1 of $AGPl$ type 3 and $SBEiI$ and subfamilies of $AGPl$ type 4, $GBSS$, and $ISAi$, have a considerably higher $\omega$ value than their corresponding dicot equivalent (table 2). This result suggests that these genes evolve rapidly in grasses.

**Divergent Selection between Grasses and Dicots**

To assess the potential for selection on gene families in the starch synthesis pathway, we performed a selection analysis using the Fitmodel of Guindon et al. (2004), and compared the sites under positive selection between the grass and the dicot subfamilies in each gene tree. The advantage of the Fitmodel approach, rather than the PAML model, is that it does not require a priori knowledge of the positions in the tree that are potentially experiencing positive selection, and thus, it is more appropriate for this analysis. For every data set analyzed, the M3+1 model was favored over the M3 model. The parameter estimates of the M3+1 model suggest that most sites (≥ 87%, $p_1 + p_2$) are identified to be under purifying selection with $\omega_1 \leq 0.012$ and $\omega_3 \leq 0.391$ in all the genes (table 3); and 1% to 9% ($p_3$) of sites within the genes in type 1 and 3 of $AGPl$, $AGPS$, $SSI$-IV, and $SBEiII$ and two $DEB$ ($ISAi$ and $PUL$) gene families are identified to be under positive selection with an $\omega_3$ value considerably larger than one. Among these positive selection sites, the number of sites for genes in $AGPS$, $SSIv$, $SBEiII$, and $PUL$ gene families is significantly asymmetric across grasses and dicots, suggesting that the selection is mainly located either in the grass or in the dicot subclade of the gene trees. This result provides evidence of divergent selection between grasses and dicots (table 3). In addition, positive selection may also have driven evolution in the $AGPl$ group 4, $GBSS$, $ISAii$, and $ISAiiii$ gene families, although no or slightly positive selection sign ($\omega_3 \leq 1.040$) was observed (table 3). The probable reason is due to the low power derived from the small number of sequences.

**Discussion**

In this study, the major genes involved in starch biosynthesis, that is $AGPl$, $AGPS$, $GBSS$, $SBEiIV$, $ISAiIII$, $ISAiIII$, and $PUL$, were used as materials to compare the evolutionary pattern of these genes between grasses and dicots. As a result, two distinct evolutionary features were observed: more old duplications and a higher evolutionary rate of the genes in the grasses. Given the fact that the grass seeds are rich in starch, while the dicot seeds are rich in protein or lipid, we hypothesized that the old duplication contributes to, and the higher evolutionary rate is associated with, the higher starch content in the grasses.

**Hypothesis 1, Gene Duplication Contributes to Higher Starch Content in Grass Seed**

In this study, the grass subfamilies of $AGPl$ type 3, $AGPS$, $GBSS$, $SSI$, $SSIi$, and $SBEiII$ gene families were shown to have experienced one or two rounds of duplication before the grasses divergence (table 1, fig. 2, and supplementary fig. S2, Supplementary Material online). Of these duplications, the ones that gave rise to the subfamilies 1 and 2 of $AGPl$ type 3, $AGPS$, $GBSS$, $SSI$, and $SBEiII$ and subfamilies 2 and 3 of $SSI$ (fig. 2 and supplementary fig. S2, Supplementary Material online) were suggested to have been created through the ancient WGD event that occurred approximately 70 Ma in grasses (Paterson et al. 2004; Comparot-Moss and Denyer 2009; Yan et al. 2009; Pan et al. 2011). Following the WGD, the duplicates underwent a process of subfunctionalization: For the genes of $AGPl$ type 3, the one from subfamily 1 is assumed to encode plastidal AGPLs, whereas the one from subfamily 2 is assumed to encode cytosolic AGPLs (Comparot-Moss and Denyer 2009); for the $AGPS$ genes, those from subfamily 1 are known to encode plastidal subunits in leaf and cytosolic subunits in seed endosperm by the use of the alternate first exon, whereas the genes from grass

---

**Table 1. Gene Duplication and Loss in the Gene Families of the Starch Biosynthetic Pathway.**

<table>
<thead>
<tr>
<th>Types</th>
<th>Gene Duplication</th>
<th>Gene Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>$AGPl$ type 3*; $AGPS*$; $GBSS*$; $SSIIII*$; $SBEII*$</td>
<td>$AGPl$ type 3*</td>
</tr>
<tr>
<td>Dicot specific</td>
<td>$AGPS*$</td>
<td>$AGPS$</td>
</tr>
<tr>
<td>Grass species specific</td>
<td>$AGPS*$; $SSIIII*$; $SSIV*$</td>
<td>$AGPl$ type 4*; $AGPS$; $GBSS$; $SSI$; $SSII$; $SSIII$; $SSIV$; $SBEiI$; $SBEiiII$; $ISAiI$; $ISAiiii$</td>
</tr>
<tr>
<td>Dicot species specific</td>
<td>$AGPl$ type 1*; 2** and 3***; $AGPS***$; $GBSS***$; $SSI$; $SSII*$; $SSIII*$; $SBEiI*$; $SBEiiII*$; $ISAi$; $ISAiiii$</td>
<td>$AGPl$ type 4*; $AGPS$; $GBSS$; $SSI$; $SSII$; $SSIII$; $SSIV$; $SBEiI$; $SBEiiII$; $ISAi$; $PUL$</td>
</tr>
</tbody>
</table>

*aThe number of asterisks indicates the number of replications.*
FIG. 3. Pairwise estimation of $\omega$ values in each grass or dicot subfamily of gene family: (A) AGPL type 1, 3, 4 and AGPS; (B) GBSS and SSI-IV; (C) SBEI-III; and (D) ISAI-III and PUL. In each of the gene families, a significance test on the difference between grass subfamily and dicot subfamily was carried out, and its result, listed in supplementary table S1 (Supplementary Material online), is indicated by n.s. (no significance). *0.01 < $P$ ≤ 0.05 and **$P$ ≤ 0.01.

Table 2. LRT Statistic and Parameters from Branch Model of PAML.

<table>
<thead>
<tr>
<th>Gene Family/Subfamily</th>
<th>Null Hypothesis</th>
<th>$-\ln L$</th>
<th>$\omega$</th>
<th>Alternative Hypothesis</th>
<th>$-\ln L$</th>
<th>$\omega$</th>
<th>Statistic</th>
<th>LRT</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGPL</td>
<td>Type 1</td>
<td>6,556.04</td>
<td>0.074</td>
<td>Grass = 0.095; Dicot = 0.061</td>
<td>6,551.64</td>
<td>0.095</td>
<td>8.8</td>
<td>3.01e-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type 3</td>
<td>10,741.28</td>
<td>0.113</td>
<td>Grass1 = 0.055; Grass 2 = 0.262; Dicot = 0.103</td>
<td>10,737.01</td>
<td>0.113</td>
<td>88.96</td>
<td>&lt;1.0e-12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type 4</td>
<td>5,908.11</td>
<td>0.0097</td>
<td>Grass = 0.073; Dicot = 0.125</td>
<td>5,902.29</td>
<td>0.073</td>
<td>11.64</td>
<td>6.46e-4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGPS</td>
<td>8,243.16</td>
<td>0.033</td>
<td>Grass 1a = 0.045; Grass 2a = 0.064; Dicot = 0.019</td>
<td>8,218.80</td>
<td>0.033</td>
<td>48.72</td>
<td>2.63e-11</td>
<td></td>
</tr>
<tr>
<td>SSI</td>
<td>GBSS</td>
<td>13,938.81</td>
<td>0.111</td>
<td>Grass1 = 0.106 M; Grass2 = 0.077; Dicot = 0.133</td>
<td>13,927.50</td>
<td>0.111</td>
<td>22.62</td>
<td>1.23e-5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSI</td>
<td>7,779.76</td>
<td>0.100</td>
<td>Grass = 0.102; Dicot = 0.099</td>
<td>7,779.71</td>
<td>0.102</td>
<td>0.1</td>
<td>0.752</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSII</td>
<td>12,886.63</td>
<td>0.114</td>
<td>Grass 1 = 0.120; Grass 2 = 0.159; Grass 3 = 0.176; Dicot = 0.076</td>
<td>12,856.93</td>
<td>0.114</td>
<td>59.4</td>
<td>7.90e-13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSIII</td>
<td>23,127.17</td>
<td>0.169</td>
<td>Grass 1 = 0.248; Grass 2 = 0.169; Dicot = 0.111</td>
<td>23,076.27</td>
<td>0.169</td>
<td>101.8</td>
<td>&lt;1.0e-12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSIV</td>
<td>15,032.69</td>
<td>0.118</td>
<td>Grass = 0.181; Dicot = 0.134</td>
<td>15,027.01</td>
<td>0.118</td>
<td>11.36</td>
<td>7.50e-4</td>
<td></td>
</tr>
<tr>
<td>SBE</td>
<td>SBEI</td>
<td>9,291.88</td>
<td>0.011</td>
<td>Grass = 0.138; Dicot = 0.077</td>
<td>9,280.23</td>
<td>0.011</td>
<td>23.3</td>
<td>1.39e-6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SBEII</td>
<td>13,201.58</td>
<td>0.070</td>
<td>Grass 1 = 0.061; Grass 2 = 0.100; Dicot = 0.060</td>
<td>13,191.18</td>
<td>0.070</td>
<td>20.8</td>
<td>3.04e-5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SBEIII</td>
<td>11,671.75</td>
<td>0.120</td>
<td>Grass = 0.140; Dicot = 0.106</td>
<td>11,668.39</td>
<td>0.120</td>
<td>6.72</td>
<td>9.53e-3</td>
<td></td>
</tr>
<tr>
<td>DBE</td>
<td>ISAI</td>
<td>11,634.27</td>
<td>0.111</td>
<td>Grass4 = 0.116; Dicot = 0.107</td>
<td>11,634.01</td>
<td>0.111</td>
<td>0.52</td>
<td>0.471</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ISAII</td>
<td>10,935.65</td>
<td>0.162</td>
<td>Grass = 0.195; Dicot = 0.147</td>
<td>10,932.39</td>
<td>0.162</td>
<td>6.52</td>
<td>1.07e-2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ISAIII</td>
<td>10,181.83</td>
<td>0.115</td>
<td>Grass = 0.142; Dicot = 0.102</td>
<td>10,177.50</td>
<td>0.115</td>
<td>8.66</td>
<td>3.25e-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PUL</td>
<td>14,383.27</td>
<td>0.158</td>
<td>Grass4 = 0.184; Dicot = 0.143</td>
<td>14,379.68</td>
<td>0.158</td>
<td>7.18</td>
<td>7.37e-3</td>
<td></td>
</tr>
</tbody>
</table>

Note.—$P$-values were calculated from “Twice the log-likelihood difference between the two models compared to a distribution with $df$ = the difference for the number of parameters between the two models.”

*The genes from these subfamilies were suggested to be mainly expressed in the endosperm of the grasses.
### Table 3. Likelihood Ratio Test Statistic and Parameters Estimated from the M3 and M3 + 1 Models in Fitmodel.

<table>
<thead>
<tr>
<th>Gene Family Subfamily</th>
<th>M3 Model</th>
<th>M3 + 1 Model</th>
<th>Positive Sites (M3 + 1 Model)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$-\ln L$</td>
<td>$p_1$, $p_2$, $p_3$</td>
<td>$-\ln L$</td>
</tr>
<tr>
<td>AGPL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 1 AGPL</td>
<td>6,519.57</td>
<td>0.48 0.35 0.17</td>
<td>6,510.01**</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 3 AGPL</td>
<td>10,679.19</td>
<td>0.55 0.43 0.02</td>
<td>10,659.58**</td>
</tr>
<tr>
<td>Type 4 AGPL</td>
<td>5,862.01</td>
<td>0.31 0.32 0.07</td>
<td>5,857.56**</td>
</tr>
<tr>
<td>AGPS</td>
<td>8,185.64</td>
<td>0.84 0.15 0.01</td>
<td>8,173.32**</td>
</tr>
<tr>
<td>SS</td>
<td>13,787.51</td>
<td>0.51 0.35 0.14</td>
<td>13,735.55**</td>
</tr>
<tr>
<td>SSI</td>
<td>7,645.76</td>
<td>0.65 0.34 0.01</td>
<td>7,633.18**</td>
</tr>
<tr>
<td>SSIi</td>
<td>12,662.74</td>
<td>0.61 0.31 0.08</td>
<td>12,599.16**</td>
</tr>
<tr>
<td>SSSII</td>
<td>22,727.50</td>
<td>0.43 0.44 0.13</td>
<td>22,680.34**</td>
</tr>
<tr>
<td>SSIV</td>
<td>14,858.01</td>
<td>0.44 0.48 0.08</td>
<td>14,826.22**</td>
</tr>
<tr>
<td>SBE</td>
<td>9,179.54</td>
<td>0.69 0.29 0.02</td>
<td>9,174.00**</td>
</tr>
<tr>
<td>SBEI</td>
<td>12,964.60</td>
<td>0.66 0.29 0.05</td>
<td>12,947.2**</td>
</tr>
<tr>
<td>SBEII</td>
<td>11,554.07</td>
<td>0.53 0.42 0.05</td>
<td>11,527.44**</td>
</tr>
<tr>
<td>DBE</td>
<td>11,437.12</td>
<td>0.62 0.33 0.05</td>
<td>11,409.15**</td>
</tr>
<tr>
<td>ISAI</td>
<td>10,758.39</td>
<td>0.45 0.42 0.13</td>
<td>10,748.98**</td>
</tr>
<tr>
<td>ISAII</td>
<td>9,947.50</td>
<td>0.52 0.35 0.13</td>
<td>9,939.22**</td>
</tr>
<tr>
<td>PUL</td>
<td>14,124.90</td>
<td>0.46 0.38 0.16</td>
<td>14,090.21**</td>
</tr>
</tbody>
</table>

$^a$ $p_i$ proportion of sites that fall into the site class, $i = 1, 2, 3$.  

**P ≤ 0.01, and the probabilities are obtained by LRT (df = 2) of the models M3 and M3 + 1.
The above results showed that most genes examined exhibit a considerably higher evolutionary rate in grasses than in dicots (fig. 3 and table 2). This conclusion is favored when we consider only the genes mainly expressed in the endosperm in grasses. Based on the previous studies, the genes from grass subfamily of 1 and 2 in AGPL type 3 and AGPS, 1 in GBSS and SSIII, and 2 in SSI and SBEII, ISAI, and PUL, were suggested to be mainly expressed in the endosperm (Beatty et al. 1999; Mizuno et al. 2001; Hirose and Terao 2004; Dian et al. 2005; Ohdan et al. 2005; Yan et al. 2009; Chen et al. 2011).

Hypothesis 2, the Higher Evolutionary Rate Is Associated with Higher Starch Content in Grass Seed

The above results showed that most genes examined exhibit a considerably higher evolutionary rate in grasses than in dicots (fig. 3 and table 2). This conclusion is favored when we consider only the genes mainly expressed in the endosperm in grasses. Based on the previous studies, the genes from grass subfamily of 1 and 2 in AGPL type 3 and AGPS, 1 in GBSS and SSIII, and 2 in SSI and SBEII, ISAI, and PUL, were suggested to be mainly expressed in the endosperm (Beatty et al. 1999; Mizuno et al. 2001; Hirose and Terao 2004; Dian et al. 2005; Ohdan et al. 2005; Yan et al. 2009; Chen et al. 2011) (table 2). All these subfamilies, except those in AGPL type 3 and GBSS, exhibit a higher evolutionary rate in the grasses than the corresponding subfamilies comprising their homologs from the dicots.

The above results suggest that the higher evolutionary rate is associated with the higher starch content of grass seed. However, the reason for this association is not obvious. The possible explanation is that strong positive selection on a large number of residues leads to a rapid rate of evolution; and the supporting evidence is that we found a significant positive selection signature in genes from the grasses (table 3).

Possible Explanation for the Evolutionary Features of the Starch Biosynthetic Genes in Grasses

Two basic questions of evolutionary biology are “How does it happen?” and “What is the consequence?” The above two hypotheses partly answered the latter. It is also tempting to speculate how it happens. We provide a possible explanation in terms of why the grass starch biosynthetic genes exhibit two such distinct features.

Our speculation involves protein–protein interactions. More and more evidences support the concept that enzymes involved in starch biosynthesis form a functional multiprotein complex. For example, in rice, the amylase extender (ae) mutation lacking SBEIIb caused a dramatic reduction in the activity of soluble SSI (Nishi et al. 2001); in maize, the ae mutant lines lacking SBEIIb also show loss of activity of SBEI and altered properties of an ISA-type DBE (Colleoni et al. 2003), and mutant lines lacking both PUL and ISA1 cause a loss in SBEIIa activity, although the amount of SBEIIa protein is apparently unchanged (James et al. 2003); and in wheat, it was confirmed by mass spectrometry that SSI, SSIa, and SBEII form as components of one or more protein complexes in amyloplasts (Tetlow et al. 2008).

Provided that enzymes involved in starch biosynthesis form functional multiprotein complexes, if a gene encoding the enzymes had a sister through duplication and consequently conferred a selective advantage to the plant, the sister would be maintained; and subsequently, the duplicates of the other genes in the pathway could be created through convergent evolution. Similarly, if a gene in the complex evolves more rapidly, other genes were also expected to show a similar pattern.

In conclusion, we deduced that there are two novel aspects of the evolution of the starch biosynthesis pathway. First, the results of the gene duplication analyses suggest that the whole-starch biosynthetic pathway was duplicated and then diverted in grasses. Second, the evolutionary rate of the two grass subfamilies of the AGPS gene family at the top of the pathway is greater than that of the dicot subfamily (table 2), contrary to the previous finding that upstream genes in a biosynthetic pathway evolve more slowly than downstream genes (Rausher et al. 1999; Lu and Rausher 2003). The reason for the two novel findings needs to be addressed in future studies.

Supplementary Material

Supplementary figures S1–S3 and supplementary table S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Dr Stéphane Guindon (Department of Statistics, University of Auckland, New Zealand) for his helpful discussion in the process of this study. This work was supported by National Key Basic Research Program of China.
References


