New insights into the origin and evolution of α-amylase genes in green plants

Liangliang Ju1,2, Zhifen Pan1, Haili Zhang1, Qiao Li1, Junjun Liang1, Guangbing Deng1, Maoqun Yu1 & Hai Long1

Gene duplication is a source of genetic materials and evolutionary changes, and has been associated with gene family expansion. Functional divergence of duplicated genes is strongly directed by natural selections such as organism diversification and novel feature acquisition. We show that, plant α-amylase gene family (AMY) is comprised of six subfamilies (AMY1-AMY6) that fell into two ancient phylogenetic lineages (AMY3 and AMY4). Both AMY1 and AMY2 are grass-specific and share a single-copy ancestor, which is derived from grass AMY3 genes that have undergone massive tandem and whole-genome duplications during evolution. Ancestral features of AMY4 and AMY5/AMY6 genes have been retained among four green algal sequences (Chrein_08.g362450, Vocart_0021s0194, Dusali_0430s00012 and Monegl_16464), suggesting a gene duplication event following Chlorophyceae diversification. The observed horizontal gene transfers between plant and bacterial AMYs, and chromosomal locations of AMY3 and AMY4 genes in the most ancestral green body (C. reinhardtii), provide evidences for the monophyletic origin of plant AMYs. Despite subfamily-specific sequence divergence driven by natural selections, the active site and SBS1 are well-conserved across different AMY isoforms. The differentiated electrostatic potentials and hydrogen bands-forming residue polymorphisms, further imply variable digestive abilities for a broad substrates in particular tissues or subcellular localizations.

As the best known and most deeply studied amylolytic enzyme, α-amylase (AMY, α-1,4-glucan-4-gluconohydrolases, EC 3.2.1.1) is an ubiquitous hydrolase synthesized by plants, animals and microorganisms, catalyzing the cleavage of internal α-(1–4)-glycosidic linkages in starch, glycogen and other related oligosaccharides with a retaining endo-acting mechanism. Its widespread distribution reflects the use of principal energy and carbon source through exploiting environmental polysaccharides. Currently, under the sequence-based classification rules of carbohydrate-active enzymes (CAZy database, http://www.cazy.org/index.html), AMYs are classified as the main representative of the glycoside hydrolase (GH) family 13, and probably also present in families GH57, GH119 and GH126. These enzymes in GH13 are characterized by adopting the (β/α)8-barrel (TIM-barrel) catalytic domain, and display strong conservation of their tertiary conformation, although the amino acid sequences exhibit a high degree of variability, and only a few amino acids are conserved, as revealed by inter-kingdom pairwise comparisons; however, the sequence similarities are much higher within kingdoms. Our present work focuses on the AMYs belonging to GH13.

The evolution of AMY genes has been an attractive subject for more than 30 years. The original study was presented by Nakajima1, who compared 11 different AMY sequences from plants, animals and microbes, and observed four highly conserved regions necessary for enzyme functions. Subsequently, Janecek2 analyzed 37 sequences that were also from the different living organisms, and established three main phylogenetic lineages: fungi and yeasts, plants, and streptomycetes, Thermomonospora curvata, insects and mammals. The archaea AMYs showed close relatedness with their plant counterparts, and these two distinct branches should retain their own originality. Additionally, several bacterial AMYs have been reported to share the typical animal-like motifs and chloride-dependent properties. This observation raised the α-amylase model of horizontal gene transfer (HGT) between animals and bacteria. However, the origin of animal AMY genes is still debatable under alternative hypotheses. Within the family GH13, AMYs from different living groups were separately present in particular subfamilies, such as GH13_1 (fungi and yeasts), GH13_6 (plants), GH13_7 (archaeons), GH13_24

1Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, 610041, China. 2University of Chinese Academy of Sciences, Beijing, 100049, China. Correspondence and requests for materials should be addressed to H.L. (email: hailong@cib.ac.cn)
chloroplasts targeting. Unlike the monocot AMY2, At1g69830) have been well-annotated, and only At1g69830 is predicted to be leaf-containing. In Arabidopsis, different categories of signal peptides or other domain modules (Fig. S2). Among them, (CAX51374-type), AMY2 investigated green species (Table S1). The global sequence alignment clearly divided the α-amylase gene family into three AMY genes, including those from Plantago major (StAmy23 and StAmy1 genes (and lar transport, organelle targeting, and organ-specific actions. In potato tubers, two β/α)8-barrel domain constitute cereal α-amylases (AMY1 and AMY2) consisted of three domains: a central conserved (domain A), an additional domain B nested between β3 and α3 of domain A, and a five-stranded C-terminal β-sheet domain (domain C)41–43. The active site and starch granule binding surface site (SBS1) are conserved in most reported enzymes; and the surface binding site of domain C (SBS2) is unique in low-pI AMY242. Multiple isoforms encoded by different AMY genes suggested that their roles in starch degradation might depend on specific plant tissues or subcellular types. In germinating cereal seeds, AMYs are typical secretory protein molecules, that biosynthesized in the secretory tissues (the scutellar epithelium and the aleurone layer), and secreted subsequently into the starchy endosperm46. Radchuk et al.39 detected AMY4 isoforms in tissues that undergo programmed cell death (PCD) during caryopsis development. In contrast with secretory isoforms, the Arabidopsis AMY3 has a predicted N-terminal transit peptide for chloroplasts localization, and is involved in leaf transitory starch degradation45,46. Additionally, it was reported to possess two carbohydrate-binding module 45 (CBM45) repeats47. In rice, the isoform α-amylase I-1 is a typical secretory glycoproteins encoded by a large gene family, and secreted subsequently into the starchy endosperm44. Radchuk et al.52 presented an comprehensive review, providing new insights into functional roles in intracellular transport, organelle targeting, and organ-specific actions. In potato tubers, two AMY genes (StAmy1 and StAmy23) were expressed, but only StAmy23 could be induced by low temperature53. Recently, Hou et al.34 indicated that StAmy23 degraded the cytosolic phyoglycogen in sweetening tubers. Plant starches are the principal storage reserve constituents, and their degradations need the coordinated participation of categories of amylolytic enzymes in distinct tissues (i.e., leaves, tuberous tissues and endosperms). The wide-distribution nature of α-amylases in the main living organisms suggests its central importance in basal carbohydrate metabolisms, such as the glycogen and starch degradation. Previous studies conducted extensive inter-kingdom comparisons of AMY enzymes/gene, and successfully extracted conserved sequence motifs responsible for enzyme specificity and/or function. However, available data is largely lacked within kingdoms, especially in green plants. The increasing megabase genome datasets enable us to conduct the first comprehensive analysis of AMY genes with sampling from green algae to higher angiosperms. We hope this work can contribute to a better understanding of the origin, expansion dynamics, and functional diversification of this multiple gene family.

Results

AMY genes in green plants. Our similarity searches resulted in 472 AMY or AMY-like sequences from the 78 investigated green species (Table S1). The global sequence alignment clearly divided the α-amylase gene family into six subfamilies: AMY1 (CAX51374-type), AMY2 (CAX51372-type), AMY3 (AAA34259- and AT4G25000-type), AMY4 (CAX51375- and AT1G76130-type), AMY5 (AT1G69830-type) and a novel subfamily AMY6 which has not been described before (Fig. 1). The overall amino acid sequence identities ranged approximately from 36% to 76% between matching conserved regions. These subfamilies were conserved to possess the basic catalytic and C-terminal beta-sheet domains, while their N-terminal ends exhibited larger sequence variations, and carried different categories of signal peptides or other domain modules (Fig. S2). Among them, AMY3, AMY4 and AMY5 genes were widely distributed across the entire green lineage, while both AMY1 and AMY2 were grass-specific.
AMY6 genes were scattered around species of basal land plants (Marchantia polymorpha, Physcomitrella patens, Sphagnum fallax and Selaginella moellendorffii), basal angiosperms (Amborella trichopoda) and two monocots (Elaeis guineensis and Phoenix dactylifera), however, it was found in almost all the main dicot lineages, except the Citrus and Brassicaceae. Occasionally, the AMY4 genes were absent in two species, S. moellendorffii and Beta vulgaris. We also had not detected any AMY3 gene among the five species (green algae Auxenochlorella protothecoides and Ostreococcus tauri, the monocot Spirodela polyrhiza and Zostera marina, and the dicot Tarenaya hassleriens).

**Phylogenetic relationships of plant AMYs.** We used three different methods (NJ, ML and BI) to infer the phylogenetic relationships. The general topology placed plant AMY genes onto two major groups: AMY1 + AMY2 + AMY3 and AMY4 + AMY5 + AMY6 (Fig. 2a). In each subfamily, the phylogeny (i.e., AMY3, AMY4, AMY5 and AMY6) exactly agreed to the green plant tree of life that is evolved from more ancient green algae or basal land and vascular plants, to higher angiosperms (Fig. 2b,c).

Within the subtree of AMY1 + AMY2 + AMY3, we defined six Grass_AMY3 subclades (3a to 3f) (Fig. 2b). To gain a more precise picture of these six subclades, we selected the redundant monocot and grass AMY3 genes to re-evaluate their phylogeny and detected the synteny. We found subclades 3a, 3e and 3d, and 3b and 3c were separately clustered together, and distinct from each other; the subcalde 3f situated between AMY3 and AMY1 + AMY2 (Figs 2b and 3a). All the single-copy genes (i.e., Hovulg_5Hr1G068350, Brdist_4g32140 and Traest_5A2) in diploid species belonged to the cluster of 3b + 3c (Fig. 3a).

Synteny detection indicated that 3b + 3c and 3a + 3e + 3d were extremely conserved on chromosomes, and resulted from tandem gene duplications (Fig. 3b). Nevertheless, apparent gene colinearities were also observed on the chromosomes of Os_chr08 and Sb_chr07, which were duplicated from Os_chr09 and Sb_chr02 at the whole genome scale. Subfamilies AMY1 and AMY2 were well segregated with each other, and formed a distinct branch that was embedded in the AMY3 subtree (Fig. 2b). Our previous work demonstrated that the AMY1 loci were conserved in a syntenic block, which were derived from the intermediate ancestral chromosome A2. From this subtree, we found that AMY1 and AMY2 located in an approximately equivalent position. Thus, we further detected the syntenic relationships of AMY2 genes, and the genomic segments carrying the AMY2 loci were also highly conserved and originated from A6 (Fig. S3).

There are three main clusters in the subtree of AMY4 + AMY5 + AMY6 (Fig. 2a). Unexpectedly, the latter AMY4 cluster (Green algae_AMY4) situated at a position that was distinct from AMY4 genes of basal land plants to angiosperms, but with less powerful supports (Fig. 2c). Subfamilies AMY5 and AMY6 showed a closer relationship to each other than to AMY4 (Fig. 2c). In the 11 sampled green algae species, the Green algae_AMY5 genes were present in seven of them, while AMY5 genes were absent in any of them (Fig. 1). By contrast, AMY5 genes were absent in the four basal land species, but all of them contained the AMY6 gene members (Fig. 1). Furthermore, these two subfamilies were featured by possessing the N-terminal extension of 460–570 amino acids.
Figure 2. Phylogenetic relationships of AMY gene subfamilies in green plants. The phylogeny was inferred from combined analyses of NJ, ML and BI methods. Statistical supports associated with branches (> 50%) were consecutively displayed as NJ and ML bootstrap values, and Bayesian posterior probabilities (BPP). The tree was rooted to bacterial AMY genes. (a) The overall NJ tree of AMY genes. (b) The subtree diagram of AMY1 + AMY2 + AMY3 in (a). (c) The subtree diagram of AMY4 + AMY5 + AMY6 in (a).

Figure 3. Phylogenetic and syntenic relationships of grass AMY3 genes. (a) The NJ tree of AMY3 genes, including all the AMY3 copies from grass and other ancient monocot species. Branch supports (> 50%) were displayed. (b) The conserved genic segments carrying the archetype AMY3 loci that were highlighted with red color.
The chloroplast transit peptides were predicted (Fig. S2). Therefore, AMY6 indeed represents a novel gene subfamily, duplicated and diverged from AMY5.

Inter-kingdom sequence analysis. Phylogenetic inference indicated that AMY1 + AMY2 + AMY3, AMY4, and AMY5 + AMY6 were three main AMY lineages in plants (Fig. 2). To further understand their origination, initially, we checked the AMY genes in the more ancient red algae, Galdieria sulphuraria. Interestingly, the amino acids of Gasu_48600 (Table S2) exhibited greater sequence divergence even than the bacterial outgroup when comparing with plant AMYs. Thus, we conducted sequence comparisons across different living organisms, and successfully extracted ten plant-type bacterial AMY genes. (Fig. 4a). In addition, the AMY4 representative gene (At1g76130) showed the highest sequence similarity to plant-type bacterial AMY genes (Fig. 4b).

Evolutionary pressures. Considering the wide distribution and neutral phylogenetic position of AMY4, we hypothesized that it might be the most basic subfamily. Thus, in the predefined tree, it was constantly treated as the background branch, and the other subfamilies were consecutively used as foreground branches; AMY3 was defined to consist of Eudicot_AMY3 and Monocot_AMY3 (Fig. S1). Firstly, we estimated the whole tree using the one-ratio model (M0) and the nearly neutral site model (M1a). The estimates ($\omega_0 = 0.10958$, and $\omega_0 = 0.10276$ and $p_0 = 0.91084$, respectively) indicated that AMY genes were deeply under selection constraints or purifying selections. In two-branch tests, all the LRT comparisons generated significant statistics with the exception of two-branch (AMY6), indicating the existence of varied $\omega$-values across subfamilies (Table 1).

In contrast to the branch or site models, we carried out the branch-site tests (Model A). These tests successfully detected proportions of sites with $\omega_2$-values greater than 1, and the LRT statistics were significant for all the M1a-Model A comparisons (Table 1). For instance, when AMY1 was used as the foreground, there existed about a proportion of 4.81% sites with $P > 0.52$ under potential positive selections, and some residues (45P* / 90T* / 184 S* / 194N* / 316K**) reached significant levels under the BEB inference. Note that some sites in the test of Model A (AMY5 + AMY6), although not existing significant signs of positive selection in our analysis, were highly conserved and indeed divergent between backgrounds and foregrounds (Fig. S4).

Structural properties and expressional profiles. The electrostatic potential reflects surface properties and molecular interactions that play critical roles in protein folding, conformational stability, enzyme catalysis and binding energies. Among the resulting models, apparent surface electrostatic changes were observed between the two main lineages, AMY1 + AMY2 + AMY3 (negative surface potentials) and AMY4 + AMY5 + AMY6 (positive to neutral potentials), in the central Domain A, in which the most essential active site of each ancestral
node was consistently displayed as negative surface electrostatics (Fig. 5a). Further structural comparisons of the well-defined substrate binding sites indicated that the active site and SBS1 in Domain A shared highly similar structural folds across six AMY subfamilies, while the SBS2 displayed larger conformational variations (Fig. 5b). In each binding site, the amino acid residues involved in forming direct hydrogen bond contacts with starch-like structural folds across six AMY subfamilies, the single-copy AMY4, AMY5 genes expressed broadly in all the tissues sampled. Similar constitutive expression scenarios were observed in cases of AMY1 to AMY5 from maize, AMY3 (Solyco3g095710) to AMY6 from tomato, and AMY4 and AMY5 from rice. By contrast, the other AMY3 copy (Solyco4g078930) in tomato was hardly detected in particular tissues (i.e., young leaves, young flower buds and anthesis flower 0DPA), and the AMY1, AMY2 and AMY3 paralogs in rice also displayed strong tissue-specific patterns. Even within the AMY3 subfamily in rice, each of these four copies exhibited distinct expression profiles.

### Discussion

Alpha-amylase genes comprise four subtypes in cereal grasses and three in dicot species. Several phylogenetic studies restricted to higher model plants have been done\(^{30,34,37,40,45,58}\), however, a detailed sequence-derived comparison of these subfamilies is lacked, and little is known about how they evolved crossing the entire green lineage. With the goal of bridging this gap, we have examined the distribution, expansion dynamics, and potential functional differentiations using 472 redundant protein sequences, sampling from 78 different species or strains covering green algae to higher angiosperms.

Generally, AMY1 and AMY2 are grass-specific subfamilies. The previous defined AMY1 in dicots, such as At4g25000\(^{37}\) and StAmy1\(^{34}\), together with grass AMY3\(^{30,34,40,58}\), actually belong to plant AMY3 lineage. The same case is AMY4, which consists of the AMY2s in dicots\(^{37,54}\) and AMY6s in grass\(^{35,39,40,58}\). The AMY5 subfamily refers to the well-studied AMY3 (At1g69830) in Arabidopsis, and its orthologous in other green plants. In addition, we have identified a novel subfamily (AMY6), which presents a scattered distribution from green algae to angiosperms (Fig. 1).

The N-terminal presequences of α-amylases have been well-described previously, and could be used as good signatures for discriminating genes belonging to specific subfamilies. In most cases, α-amylases encoded by AMY1, AMY2 and AMY3 genes are typical secretory proteins found in plastids (chloroplasts in leaves and amyloplasts in starch cells)\(^{40,56,57}\), including a length of 24 to 27 signal peptides. However, those AMY3 genes from the basal land plants (Liverworts, Mosses and Ferns) are predicted to carry the chloroplast signals. This clear signal shift from basal land plants to angiosperms may reflect some evolutionary novelties, such as the occurrence of seeds in plant bodies. Generally, the AMY4 isoform do not possess any signal peptide\(^{39}\). Our predictions indicate that three categories of signal patterns (secretory signals, no signals and chloroplast transit peptides) are overwhelmingly dominated the N-terminal regions of α-amylase genes in higher plants (Fig. S2). The most remarkable are AMY5 and AMY6, both of which own the N-terminal stretches of greater than 460 amino acids, and are predicted to carry the chloroplast signals. In green algae, the N-terminal signal divisions are less clear, because

### Table 1. Positive selection analysis using the Maximum likelihood method. BEB significance levels: *P < 0.05, **P < 0.01. l: log likelihoods. NA: not allowed. Amino acids referred to 1st sequence: Arlyra_7G28170.

| Models                  | l       | LRT   | Positive Selected Sites |
|------------------------|---------|-------|-------------------------|
| M0: one-ratio          | −26184.75 | NA    | NA                      |
| Two branches (AMY1)    | −26174.68 | 7.20E-06 | NA                      |
| Two branches (AMY2)    | −26166.64 | 1.76E-09 | NA                      |
| Two branches (AMY3)    | −26179.54 | 1.25E-03 | NA                      |
| Two branches (AMY1 + AMY2 + AMY3) | −26179.54 | 1.25E-03 | NA                      |
| Two branches (AMY5)    | −26174.38 | 5.26E-06 | NA                      |
| Two branches (AMY6)    | −26184.57 | 0.55  | NA                      |
| Two branches (AMY5 + AMY6) | −26181.63 | 1.43E-02 | NA                      |
| M1a: nearly neutral    | −26017.13 | NA    | NA                      |
| Model A (AMY1)         | −25988.90 | 6.07E-13 | NA                      |
| Model A (AMY2)         | −25975.19 | 0.00E+00 | NA                      |
| Model A (AMY3)         | −25993.51 | 6.09E-11 | NA                      |
| Model A (AMY1 + AMY2 + AMY3) | −25979.01 | 0.00E+00 | NA                      |
| Model A (AMY5)         | −25987.02 | 9.36E-14 | NA                      |
| Model A (AMY6)         | −25981.39 | 3.33E-16 | NA                      |
| Model A (AMY5 + AMY6)  | −26004.65 | 3.79E-06 | NA                      |

In green algae, the N-terminal presequences of α-amylases have been well-described previously, and could be used as good signatures for discriminating genes belonging to specific subfamilies. In most cases, α-amylases encoded by AMY1, AMY2 and AMY3 genes are typical secretory proteins found in plastids (chloroplasts in leaves and amyloplasts in starch cells)\(^{40,56,57}\), including a length of 24 to 27 signal peptides. However, those AMY3 genes from the basal land plants (Liverworts, Mosses and Ferns) are predicted to carry the chloroplast signals. In green algae, the N-terminal signal divisions are less clear, because...
of A4/A6 breakages and fusions. So α-amylases with different signals suggest their different subcellular locations for starch digestion.

Colinearity identifications indicate that the grass AMY lineage includes AMY1, AMY2, AMY3, and AMY4. Previous studies well characterized the none-signal bearing genes. They unanimously agree that each of them represented a distinct AMY subfamily in angiosperms. Differently, Mascher et al. put them together as a single subfamily (AMY4 as AMY4-I, and AMY5 as AMY4-II). In the present work, we provide more detailed identifications across the entire green lineage. In the closely-related Chlamydomonadales (Chlamydomonas reinhardtii, Dunaliella salina, Volvox carteri) and Sphaeropleales (Monoraphidium neglectum) AMY5 genes are absent. Interestingly, their corresponding AMY4 gene members (Green algae_AMY4 in Fig. 2c) situate at the common ancestral position of AMY4 and AMY5 subfamilies. The poorly resolved or conflicting relationships of this ancestral branch may represent the ancestral status of AMY4 and AMY5 genes. The shared common ancestor is additionally explained by their equivalent chromosomal locations in genomes of four Brassicaceae species (A. thaliana, A. lyrata, B. oleracea and B. rapa) and Prunus persica. Thus, AMY4 and AMY5 should have resulted from an ancient gene duplication event following the radiation of Chlamydomonadales species, and the divergent direction is AMY5 from AMY4. The newly described AMY6 is distinct from AMY5 based on their N-terminal features (Fig. S2). And from its distribution, we see that species

**Figure 5.** 3D analyses of AMY protein structure. (a) Evolution of the divergent ancestral models across AMY1 to AMY6. Surfaces were colored with red, white and blue, separately indicating the negative (−10 kcal/mol·e) neutral (0) and positive (10 kcal/mol·e) electrostatic potentials. (b) Structure comparison of substrate binding sites (the active site, SBS1 and SBS2).
with AMY6 genes must have corresponding AMY5 gene members, except those four basal land and vascular plants (Common liverwort, Moss, Bog moss and Spikemoss) (Fig. 1).

Inter-kingdom comparisons provide important evidences for evolution of α-amylase genes in plants. Similar to the case of animals4,9, our inter-kingdom analysis indicates frequent horizontal gene transfers between plant and bacteria AMY genes (Fig. 4a), and the AMY4 subfamily is more similar to bacterial AMYs than AMY3 (Fig. 4b). That is to say, the plant AMY genes share a common ancestor, and AMY4 genes have retained more sequence features of the ancestral AMYs in plants. In species of C. reinhardtii, Aquilegia coerulea, Theobroma cacao and Cucumis sativus, the equivalent chromosomal locations of AMY3 and AMY4 genes suggest gene duplications predating the diversification of green plants. Among those, we propose a diagram for the monophyletic origin of plant AMYs (Fig. 6).

In general, genes with constitutive expressions are more conserved than those exhibiting tissue-specific patterns63,64. The broad expressions of AMY4 and AMY5 genes among representative taxa (Fig. S5) suggest that they are functionally-essential in plants. Based on signal peptide predictions, we guess that these two subfamilies possibly target different subcellular localizations for initiating starch degradation, which is partly reflected by previous studies37,54,65. The novel AMY6 with the closest phylogenetic relationship with AMY5 also shares the similar constitutive expression scenario, however, it is scattered distributed, and absent or loss in particular green lineages (Fig. 1). These observations truly reflect the taxa-specific characteristic of AMY6 genes. Cereal α-amylases, especially those encoded by AMY1, AMY2 and AMY3 genes, are typical secretory proteins during seed germination or grain development. In wheat and barley, AMY1 and AMY2 showed clear expression divergence in terms of transcript abundance, specific tissues or developmental stages29, and the wheat AMY3 (Traest_5A1 in Fig. 3a) shared a similar pattern with AMY2 in developing grains30,31. Unexpectedly, the expression of archetype AMY3 genes (Traest_5A2, Traest_5B3, Traest_5D2 and Hovulg_5Hr1g068350) were hardly detected in all the tissues sampled (Data collection from https://wheat-urgi.versailles.inra.fr/ and http://webblast.ipk-gatersleben.de/barley_ibsc/).

Figure 6. A proposal for the monophyletic origin of plant AMY genes. The ancestral archetype AMY gene in plants was initially duplicated into two primary lineages (AMY3 and AMY4), which possibly exist predating green algal diversification. Circles represent three critical nodes for the birth of AMY5, AMY6, AMY1 and AMY2, respectively. Rectangles along each line indicate different categories of signals, such as ‘c’ for chloroplast transit, ‘m’ for mitochondrial targeting, and ‘s’ for secretory pathways. Ages (mya: million years ago) of critical nodes for the evolution of green plants were based on previous publications50–52. The sister species were listed in parentheses.
respectively), implying their nonfunctionalization during evolution. In rice, the expression divergence is also evident between AMY1 and AMY2, and the four duplicated AMY3 paralogs have more diversified expression profiles than single- or two-copy status in other representative organisms (Fig. S5).

The combination of sequence comparison, selection simulation and expression analysis leads to a definition of functional divergent between subfamilies. Hypothetically, the presence of multiple divergent isoforms could enable plant organisms to exploit categories of starch-like polysaccharides in a broad environmental conditions. Structural comparison further indicate that the active site and SBS1 are well-conserved (Fig. S), which is not restricted to the cereal isoforms AMY1 and AMY2, but among these six divergent enzymes (AMY1-AMY6). Generally, the active site is responsible for binding and catalyzing substrates whereas the secondary binding site (SBS1) and the AMY2-specific binding site (SBS2) are just critical for binding different starch-like polysaccharides. This conservation reflect the ability in starch digestion across AMY subfamilies. However, the electrostatic changes between AMY1 + AMY2 + AMY3 and AMY4 + AMY5 + AMY6 (Fig. 5), together with the substituted residues involved in forming direct hydrogen bonds (Table S3), may suggest the evolutionary diversification of enzyme specificity, such as substrate preference and product specificity. Furthermore, in the N-terminal presequences of AMY5 genes, there also exist another kind of noncatalytic carbohydrate-binding module (CBM45), which was reported to be associated with plastidial starch metabolism.

Methods

Sequence data. As previously reported by some authors and demonstrated by Yu et al., AMY genes were categorized into AMY1 to AMY4 among cereal crops and AtAMY1 to AtAMY5 in A. thaliana. For systematically identification of AMYs in green plants, we selected seven well-characterized AMY genes as queries (four GenBank accessions: CAX51372, CAX51374, AAC4289 and CAX51375 in barley and wheat, and three Araport11 entries: AT4G25000, AT1G76130 and AT1G69830 in A. thaliana) to perform blastp searches against databases of Phytozome v12.1 (https://phytozome.jgi.doe.gov/pz/portal.html), KEGG (http://www.genome.jp/), IPK barley (http://webblast.ipk-gatersleben.de/barley_ibsc/), and blastn searches against Wheat URGII (https://wheat-urgi.versailles.inra.fr/) with default setting details, separately including 63, 13, 1 and 1 green plant species (Table S1). The deduced amino acid sequences were subjected to multiple sequence alignment by ClustalX and Bioedit. Sequences not belonging to GH13.6, and/or with apparent sequence erosion (poor coverage and qualities) were discarded. For copy number determination, the primary peptide was chosen if alternative splicing peptides per copy were available. Since the N- and C-terminal end sections showed poor alignment across different AMY subfamilies, they were cutoff before phylogenetic reconstruction. And when several loci per species and several copies per locus were available, a single protein/copy sequence was kept if these loci/copies were shown to cluster in the same lineage. Subsequently, we selected 19 protein sequences representing all the analyzed AMY subfamilies across major plant lineages, to detect protein domain architecture using HMMER. All the searches run with default parameters. Signal peptides were also predicted using the TargetP 1.1 server.

Phylogenetic reconstructions. In the first step, we analyzed the phylogenetic relationships between different subfamilies. We conducted Neighbor-Joining (NJ), Maximum Likelihood (ML) and Bayesian inference (BI) analyses using the single-copy orthologous dataset, including 509 protein sequences of functional AMY genes and 5 bacteria outgroups retrieved from NCBI (https://www.ncbi.nlm.nih.gov/) according to Da Lage et al.. The NJ tree was analyzed using Jones-Taylor-Thornton (JTT) model with Gamma Distributed (G) rate across sites in MEGA7. Branch confidence levels were estimated by using 1000 bootstrap replications. The ML trees were conducted using the best-fitting amino-acid model LG + G with the lowest Bayesian Information Criterion (BIC) score in MEGA7. Supports values were estimated from 500 non-parametric bootstrap iterations. The BI analyses were performed using MrBayes 3.2. The preset Poisson substitution model was invoked by ‘Whelan and Goldman’ (WAG) through estimating the fixed rate models implemented in MrBayes 3.2. The option for rates was set to invgamma, and all the other parameters of the likelihood model were default values. Four simultaneous Markov chains (3 heated and 1 cold) were run starting from a random tree for 3 million generations requested and trees were sampled per 1,000 generations. The standard deviation of split frequencies fell below 0.05. Using a relative burn-in of 25% for diagnostics, the consensus tree was based on the remaining 1,500 trees. The confidence level of the tree topologies was estimated according to Bayesian posterior probabilities (BPPs). Gaps or missing data were treated as complete deletion.

Horizontal gene transfer (HGT) is the exchange of genetic material between organisms that are not in a parent-offspring relationship. HGT was observed between animals and bacteria AMY genes. For understanding the origin of green plant AMY genes, we compared protein sequences from eukaryotes (animals, plants, fungi and red algae) with those from bacteria. Firstly, we tested the key word ‘alpha-amylase’ against the protein fields of NCBI, and filtered with Swissprot entries. This strategy produced 87 animal and 79 fungal AMYs, and based on the global alignment, we selected 13 representative sequences from animals and 10 from fungi. Three AMYs in A. thaliana were selected as representatives in green plants. In red algae, we selected the single periplasmic sequence (Gsu_48600) in G. sulphuraria from KEGG as the representative. Then, we used all these representative sequences to blast against the KEGG prokaryote genomes. The top 10 hits with the annotations of ‘Alpha-amylase, EC: 3.2.1.1’ were kept. Finally, a dataset of 48 sequences from different living kingdoms (Table S2) were produced and used for NJ phylogenetic analysis with MEGA7. Supports values were estimated by using 1000 bootstrap replications.

Synteny detection. Based on the well-established model of cereal genome evolution, chromosome-scale pseudomolecules carrying AMY targets were downloaded from available resources for local database construction. Reciprocal blast was carried out to confirm the orthologous relationships. By manual chromosomal walking,
genic markers flanking AMY targets were used as queries to blast against the local database using the basic tool NCBI-BLAST-2.4.0+
. Genomic segments covering these markers were selected for gene order detection.

**Maximum likelihood analyses of positive selection.** To measure variation in functional constraints and to test whether positive selection was involved among evolutionary AMY gene lineages, we estimated the omega values ($\omega = dN/dS$), the ratio of the rate of nonsynonymous substitution per nonsynonymous site [dn] to the rate of synonymous substitution per synonymous site [ds], through using a maximum-likelihood based CODEML program implemented in the PAML version 4.9
. The branch models allowed the $\omega$ ratio to vary among lineages, and were specified for detecting adaptive selection acting on particular branches (variable 'model' and NSsites = 0)
. The site analysis allowed $\omega$ to vary among codon sites but kept it constant over the tree topologies (model = 0 and variable 'NSsites')
. The alternative Model A, also the recommended branch-site test of positive selection, assume that sites in predefined foreground branches are allowed to evolve under positive selection ($\omega_1 > 1$), whereas the background branches evolve neutrally ($\omega_1 = 1$) or under purifying selection ($0 < \omega_1 < 1$)
. Model A was tested against the nearly neutral site model M1a (Chi-square test, degree of freedom 2) using the likelihood ratio test (LRT)
. The Bayes Empirical Bayes (WEB) approach was used to calculate posterior probabilities and to select sites from the class with $\omega > 86$. The codon dataset consisted of 41 sequences with 1137 characters. It was output from the PAL2NAL server (http://www.bork.embl.de/pal2nal/), and then used for model simulations. The phylogenetic subtree was presented in Fig. S1.

**3D structure analysis and expression level comparison.** To evaluate potential functional changes among different AMY subfamilies, we carried out 3D structure studies for each ancestral node back to the time point of gene divergence or duplication. A total of 10 ancestral amino acid sequences (AMY1 to AMY6, AMY123, AMY56 and AMY456) were generated using parsimony state reconstruction in Mesquite 3.31 (http://mesquiteproject.org). We modeled protein crystal structures using the ancestral state sequences in the work-place of SWISS-MODEL
 based on their respective best matching templates from RCSB Protein Database Bank (PDB), such as template entries 1BG9
, 1RPK
, 2QPU
 and 3WN6
. The resulting models were then subjected to UCSF Chimera
 for electrostatic potential mapping (Coulombic surface coloring defaults: $\epsilon = 4\epsilon_0$, thresholds $\pm 10\, \text{kcal/mol}$), domain comparison and visualization.

Four species (tomato, rice, Arabidopsis and maize) were selected for expressionanalyses. Transcripts pro-files from various tissues in different developmental stages were retrieved from their corresponding databases: TFGD (http://ted.bti.cornell.edu/), Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/), and Arabidopsis and maize eFP Browser (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi, http://bar.utoronto.ca/efp/maize/cgi-bin/efpWeb.cgi). The normalized expression values were log2 transformed. Heatmaps were generated by hierarchically clustering with the software R version 3.4.3 (https://www.R-project.org/).

**References**

1. Nakajima, T., Isakamura, T. & Aiba, S. Comparison of amino acid sequences of eleven different α-amylases. *Applied Microbiology and Biotechnology* 23(5), 355–360 (1986).
2. Janecek, S. Sequence similarities and evolutionary relationships of microbial, plant and animal alpha-amylases. *FEBS Journal* 224(2), 519–524 (1994).
3. Pujadas, G. & Palau, J. Evolution of -Amylases: Architectural Features and Key Residues in the Stabilization of the (β/8 Scaffold. *Molecular Biology and Evolution* 18(1), 38–54 (2001).
4. Da Lage, J. L., Danchin, E. G. & Casane, D. Where do animal alpha-amylases come from? An interkingdom trip. *FEBS Letters* 581(21), 3927–35 (2007).
5. Janecek, S., Svensson, B. & MacGregor, E. A. alpha-Amylase: an enzyme specificity found in various families of glycoside hydrolases. *Cell Mol Life Sci* 71(7), 1149–70 (2014).
6. Janecek, S. & Gabrisko, M. Remarkable evolutionary relatedness among the enzymes and proteins from the alpha-amylase family. *Cell Mol Life Sci* 73(14), 2705–26 (2016).
7. Lombard, V. et al. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res*, 2014; 42(Database issue). D490–5.
8. Janecek, S. & Sevcik, J. The evolution of starch-binding domain. *FEBS Letters* 456(1), 119–125 (1999).
9. Da Lage, J. L., Feller, G. & Janecek, S. Horizontal gene transfer from Eukarya to bacteria and domain shuffling: the alpha-amylase model. *Cell Mol Life Sci* 61(1), 97–109 (2004).
10. Janecek, S. et al. Close Evolutionary Relatedness of α-Amylases from Archaea and Plants. *Journal of Molecular Evolution* 48(4), 421–426 (1999).
11. Petricek, M., Tichy, P. & Kuncova, M. Characterization of the α-amylase-encoding gene from Thermomonospora curvata. *Gene* 112(1), 77–83 (1992).
12. Feller, G. et al. Structural and Functional Aspects of Chloride Binding to Alteromonas Haloplanctis Alpha-Amylase. *Journal of Biological Chemistry* 271(39), 23836–23841 (1996).
13. Sumitani, J. et al. Bacillus animal type α-amylase: Cloning and sequencing of the gene, and comparison of the deduced amino acid sequence with that of other amylases. *Journal of Fermentation and Bioengineering* 85(4), 428–432 (1998).
14. Coronado, M. et al. The α-amylase gene amyH of the moderate halophile Halomonas meridiana: cloning and molecular characterization. *Microbiology* 146(4), 861–868 (2000).
15. Damico, S., Gerdau, C. & Feller, G. Structural similarities and evolutionary relationships in chloride-dependent alpha-amylases. *Gene* 253(1), 95–105 (2000).
16. Da Lage, J.-L. An optional C-terminal domain is ancestral in α-amylases of bilaterian animals. *Amylase* 1, 1 (2017).
17. Samue1on, L. C. et al. Retrovirral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. *Molecular and Cellular Biology* 10(6), 2513–2520 (1990).
18. Groot, P. C. et al. Evolution of the human α-amylase multigene family through unequal, homologous, and inter- and intrachromosomal crossovers. *Genomics* 8(1), 97–105 (1990).
19. Meister, M. H. & Ting, C. The Remarkable Evolutionary History of the Human Amylase Genes. *Critical Reviews in Oral Biology & Medicine* 4(3), 303–509 (1993).
20. Benkel, B. F. et al. Structural organization and chromosomal location of the chicken alpha-amylase gene family. *Gene* 362, 117–24 (2005).
21. Butterworth, P. J., Warren, F. J. & Ellis, P. R. Human α-amylase and starch digestion: An interesting marriage. *Starch - Stärke* 63(7), 395–405 (2011).
22. Perry, G. H. et al. Diet and the evolution of human amylase gene copy number variation. Nature Genetics 39(10), 1256–1260 (2007).
23. Mandel, A. L. et al. Individual differences in AMY1 gene copy number, salivary alpha-amylase levels, and the perception of oral starch. PLoS One 5(10), e13352 (2010).
24. Lage, J. D. et al. Amyrel, a paralogous gene of the amylase gene family in Drosophila melanogaster and the Sophophora subgenus. Proceedings of the National Academy of Sciences of the United States of America 95(12), 6848–6853 (1998).
25. Da Lage, J. L., Maczkowiak, F. & Cariou, M. L. Molecular characterization and evolution of the amylase multigene family of Drosophila melanogaster. J Mol Evol 51(4), 391–403 (2000).
26. Jacobsen, J. V. & Higgins, T. J. V. Characterization of the α-Amylase Synthesized by Aleurone Layers of Himalaya Barley in Response to Gibberellic Acid. Plant Physiology 70(6), 1647–1653 (1982).
27. Gale, M. D., Choejecki, A. J. & Kempton, R. A. Genetic control of a Amylase production in wheat. Theoretical and Applied Genetics 64(4), 309–316 (1983).
28. Rogers, J. C. & Millikan, C. Isolation and sequence analysis of a barley alpha-amylase cDNA clone. Journal of Biological Chemistry 258(13), 8169–8174 (1983).
29. Lazarus, C. M., Baulcombe, D. C. & Martienssen, R. A. α-amylase genes of wheat are two multigene families which are differentially expressed. Plant Molecular Biology 5(1), 13–24 (1985).
30. Baulcombe, D. C. et al. A novel wheat α-amylase gene (α-Amy3). Molecular Genetics and Genomics (1987).
31. Watanabe, A. et al. Engineering alpha-amylase levels in wheat grain suggests a highly sophisticated level of carbohydrate regulation during development. J Exp Bot 57(18), 5437–57 (2016).
32. Huang, N. et al. Structural organization and differential expression of rice α-amylase genes. Nucleic Acids Research 18(23), 7007–7014 (1990).
33. Huang, N. et al. Classification and characterization of the rice α-amylase multigene family. Plant Molecular Biology 14(5), 655–668 (1990).
34. Huang, N., Stebbins, G. L. & Rodriguez, R. L. Classification and evolution of alpha-amylase genes in plants. Proceedings of the National Academy of Sciences of the United States of America 89(16), 7526–7530 (1992).
35. Meier, J. C., Š. Janeček & J.-P. Ral, New insight in cereal starch degradation: identification and structural characterization of four α-amylases in bread wheat. Amylase, 1(1) (2017).
36. Wegezyn, T. et al. A novel α-amylase gene is transiently upregulated during low temperature exposure in apple fruit. FEBS Journal 267(5), 1313–1322 (2000).
37. Lloyd, J. R., Kossmann, J. & Ritte, G. Leaf starch degradation comes out of the shadows. Trends in Plant Science 10(3), 130–137 (2005).
38. Pommerening, B. et al. Common Plantain. A Collection of Expressed Sequence Tags from Vascular Tissue and a Simple and Efficient Transformation Method. Plant Physiology 142(4), 1427–1441 (2006).
39. Radvukh, V. V. et al. Spatiotemporal Profiling of Starch Biosynthesis and Degradation in the Developing Barley Grain. Plant Physiology 150(1), 190–204 (2009).
40. Mascher, M. et al. A chromosome conformation capture ordered sequence of the barley genome. Nature 544(7651), 427–433 (2017).
41. Kadziola, A. et al. Molecular structure of a barley alpha-amylase inhibitor complex: implications for starch binding and catalysis. Journal of Molecular Biology 278(1), 205–217 (1998).
42. Robert, X. et al. The Structure of Barley α-Amylase Isozyme 1 Reveals a Novel Role of Domain C in Substrate Recognition and Binding. Structure 11(8), 973–984 (2003).
43. Ochiai, A. et al. Crystal structure of α-amylase from Oryza sativa: molecular insights into enzyme activity and thermostability. Bioscience, Biotechnology, and Biochemistry 78(6), 989–997 (2014).
44. Mitsui, T. et al. Biosynthesis of rice seed α-amylase: two pathways of amylase secretion by the scutellum. Archives of Biochemistry and Biophysics 241(1), 315–328 (1985).
45. Streu, S., Eiche, S. & Zeeman, S. C. The Simultaneous Abolition of Three Starch Hydrolases Blocks Transient Starch Breakdown in Arabidopsis. Journal of Biological Chemistry 287(50), 41745–41756 (2012).
46. Seung, D. et al. Arabidopsis thaliana AMY1 is a unique redox-regulated chloroplastic alpha-amylase. J Biol Chem 288(47), 33620–33 (2013).
47. Glaring, M. A. et al. Starch-binding domains in the CBM45 family–low-affinity domains from glucan, water dikinase and alpha-amylase involved in plastidial starch metabolism. FEBS J 278(7), 1175–85 (2011).
48. Mitsui, T., Yamaguchi, J. & Akazawa, T. Physicochemical and Serological Characterization of Rice [alpha]-Amylase Isoforms and Identification of Their Corresponding Genes. Plant Physiology 110(4), 1395–1404 (1996).
49. Asatsuma, S. et al. Involvement of alpha-amylase 1-1 in starch degradation in rice chloroplasts. Plant Cell Physiol 46(6), 858–69 (2005).
50. Kitaizumi, A. et al. The rice alpha-amylase glycoprotein is targeted from the Golgi apparatus through the secretory pathway to the plastids. Plant Cell 21(9), 2844–58 (2009).
51. Hakata, M. et al. Suppression of alpha-amylase genes improves quality of rice grain ripened at high temperature. Plant Biotechnol J 10(9), 1110–7 (2012).
52. Mitsui, T. et al. Novel molecular and cell biological insights into function of rice α-amylase. Amylase 2(1), 30–38 (2018).
53. Zhang, H. et al. Amylase Analysis in Potato Starch Degradation During Cold Storage and Sprouting. Potato Research 57(1), 47–58 (2014).
54. Hsu, J. et al. Amylases StAmy23, StBM1 and StBM9 regulate cold-induced sweetening of potato tubers in distinct ways. J Exp Bot 68(9), 2317–2331 (2017).
55. International Brachypodium, I. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463(7282), 763–8 (2010).
56. Salse, J. et al. Identification and characterization of shared duplications between rice and wheat reveal new insights into grass genome evolution. Plant Cell 23(1), 11–24 (2008).
57. Jiao, J. et al. Structural organization and functional divergence of high isoelectric point Iα–amylase genes in bread wheat (Triticum aestivum L.) and barley (Hordeum vulgare L.). BMC Genetics 20(1), (2019).
58. Zhang, Q. & Li, C. Comparisons of Copy Number, Genomic Structure, and Conserved Motifs for alpha-Amylase Genes from Barley, Rice, and Wheat. Front Plant Sci 8, 1727 (2017).
59. Merchant, S. S. et al. The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions. Science 318(5848), 245–250 (2007).
60. Kellogg, E. A. Evolutionary History of the Grasses. Plant Physiology 125(3), 1198–1205 (2001).
61. Ming, R. et al. The pineapple genome and the evolution of CAM photosynthesis. Nat Genet 47(12), 1435–42 (2015).
62. Leliaert, F. et al. Phylogeny and Molecular Evolution of the Green Algae. Critical Reviews in Plant Sciences 31(1), 1–46 (2012).
63. Akashi, H. Gene expression and molecular evolution. Current Opinion in Genetics & Development 11(6), 660–666 (2001).
64. Wright, S. I. et al. Effects of Gene Expression on Molecular Evolution in Arabidopsis thaliana and Arabidopsis lyrata. Molecular Biology and Evolution 21(9), 1719–1726 (2004).
65. Zeeman, S. C. et al. A starch-accumulating mutant of Arabidopsis thaliana deficient in a chloroplastic starch-hydrolysing enzyme. Plant Journal 15(3), 357–365 (1998).
66. Robert, X. et al. Oligosaccharide binding to barley α-amylase 1. Journal of Biological Chemistry 280(38), 32968–32978 (2005).
Acknowledgements
This work was supported by the Major State Basic Research Development Program of China (2014CB138104), and the National S&T Key Project of China on GMO Cultivation for New Varieties (2016ZX08009-003-004-005), and Science and Technology Support Project of Sichuan Province, China (2016NZ0103).

Author Contributions
I.J. conducted sequence collection, data analysis and interpretation, and wrote the manuscript. Q.L., H.Z. and J.L. analysed data. H.L., P.Z.F. and M.Y. planned and designed the research, and finalized the paper. All the authors have read through the manuscript and agree to the submission of the final version.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-41420-w.

Competing Interests: The authors declare no competing interests.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019