BMC Evolutionary Biology

BioMed Central

Research article

Duplications and functional divergence of ADP-glucose pyrophosphorylase genes in plants

Nikolaos Georgelis[1], Edward L Braun[2] and L Curtis Hannah*[1]

Address: [1]Program in Plant Molecular and Cellular Biology and Horticultural Sciences, University of Florida, Gainesville, Florida 32610-0245, USA and [2]Department of Zoology, University of Florida, Gainesville, Florida 32611-8525, USA

Email: Nikolaos Georgelis - gnick@ufl.edu; Edward L Braun - ebraun68@ufl.edu; L Curtis Hannah* - hannah@mail.ifas.ufl.edu

* Corresponding author

Published: 12 August 2008

BMC Evolutionary Biology 2008, **8**:232 doi:10.1186/1471-2148-8-232

Received: 10 April 2008
Accepted: 12 August 2008

This article is available from: http://www.biomedcentral.com/1471-2148/8/232

Abstract

Background: ADP-glucose pyrophosphorylase (AGPase), which catalyses a rate limiting step in starch synthesis, is a heterotetramer comprised of two identical large and two identical small subunits in plants. Although the large and small subunits are equally sensitive to activity-altering amino acid changes when expressed in a bacterial system, the overall rate of non-synonymous evolution is ~2.7-fold greater for the large subunit than for the small subunit. Herein, we examine the basis for their different rates of evolution, the number of duplications in both large and small subunit genes and document changes in the patterns of AGPase evolution over time.

Results: We found that the first duplication in the AGPase large subunit family occurred early in the history of land plants, while the earliest small subunit duplication occurred after the divergence of monocots and eudicots. The large subunit also had a larger number of gene duplications than did the small subunit. The ancient duplications in the large subunit family raise concern about the saturation of synonymous substitutions, but estimates of the absolute rate of AGPase evolution were highly correlated with estimates of ω (the non-synonymous to synonymous rate ratio). Both subunits showed evidence for positive selection and relaxation of purifying selection after duplication, but these phenomena could not explain the different evolutionary rates of the two subunits. Instead, evolutionary constraints appear to be permanently relaxed for the large subunit relative to the small subunit. Both subunits exhibit branch-specific patterns of rate variation among sites.

Conclusion: These analyses indicate that the higher evolutionary rate of the plant AGPase large subunit reflects permanent relaxation of constraints relative to the small subunit and they show that the large subunit genes have undergone more gene duplications than small subunit genes. Candidate sites potentially responsible for functional divergence within each of the AGPase subunits were investigated by examining branch-specific patterns of rate variation. We discuss the phenotypes of mutants that alter some candidate sites and strategies for examining candidate sites of presently unknown function.

Background

ADP-glucose pyrophosphorylase (AGPase; EC 2.7.7.27) catalyses a rate-limiting step in starch synthesis, the formation of ADP-glucose from glucose-1-P and ATP. ADP-

glucose is the predominant, if not sole, precursor for starch synthesis. While AGPase is a homotetramer in bacteria (including cyanobacteria), it is a heterotetramer in angiosperms and green algae. This heterotetramer comprises two identical large and two identical small subunits. They exhibit a high degree of identity to each other and to the cyanobacterial AGPase, pointing to an origin by gene duplication early in the evolution of plants and green algae (Figure 1A) (Additional file 1) [1,2]. The two subunits have complementary rather than redundant functions, and knockout mutations in either abolish more than 90% of AGPase activity in some experimental systems [3].

Although both subunits are necessary for full AGPase activity, the angiosperm small subunit appears more conserved than the large subunit throughout its sequence (small subunits exhibit an average of 91.3% amino acid identity while large subunits an average of 70.8% identity) [2]. However, mean percent identities might be misleading since the genes encoding both subunit genes of AGPase underwent a number of duplications after the initial duplication that generated the two subunits. The potential confusion due to the comparison of paralogs rather than orthologs can be overcome by methods that incorporate phylogeny, such as the use of maximum likelihood (ML) to estimate ω (the ratio of non-synonymous substitutions per non-synonymous site [K_A] to synonymous substitutions per synonymous site [K_S]). The ML estimate of ω for the large subunit is ~2.7-fold greater than the estimate for the small subunit [2], suggesting a higher rate of amino acid replacement for the large subunit.

Although ω provides a convenient and commonly used method to examine evolutionary constraints, it has typically been used to examine sequences that have diverged relatively recently. The rate of synonymous evolution is relatively high in plant nuclear genes [4-7] and estimates of K_S appear saturated in analyses of some angiosperm gene families, even for relatively shallow evolutionary divergences [8]. Hence, the accuracy of ω estimates for ancient divergences is unclear. Another potential problem for the use of ω is the assumption that mutations at synonymous sites are neutral. It has been suggested that synonymous sites are subject to both positive and purifying selection [9-12]. The action of selection on synonymous sites may explain why adding among-sites rate variation for synonymous sites to models of codon evolution improves their fit to empirical data [13,14]. Saturation and among-sites rate variation both have the potential to cause K_S to be underestimated (and ω to be overestimated since $\omega = K_A/K_S$); biased estimates of ω will lead to incorrect inferences regarding evolutionary constraints on the proteins being analyzed. Finally, ω cannot detect changes in the evolutionary rate when rates of synonymous and non-synonymous substitution increase or decrease simultaneously [15].

The almost 3-fold difference in evolutionary rates for the AGPase subunits is a paradox because random mutagenesis revealed that maize endosperm AGPase subunits expressed in bacteria are equally susceptible to activity-altering amino acid changes [2]. Georgelis et al. [2] proposed that the difference in evolutionary rates between AGPase subunits reflected, at least in part, the differences between the subunits in their tissue-expression patterns and the fact that the small subunit has to interact with multiple large subunits in plants. Here, we establish the pattern and timing of duplications in the AGPase gene family and estimate absolute rates of AGPase sequence evolution. Functional divergence has been observed among AGPase subunits based on biochemical criteria [16-20]. One of our primary goals was to identify candidate sites for functional divergence. We identify specific AGPase sites apparently subject to either positive selection or branch-specific patterns of rate variation (types-I and -II divergence as defined by Gu [21,22]).

Results

Patterns of AGPase gene duplication

It is well known that genes can have three possible fates after duplication [23-27]: (1) nonfunctionalization, in which one duplicate is lost, (2) subfunctionalization, in which the functions of the original single-copy gene are partitioned between the duplicates, and (3) neofunctionalization, in which one duplicate gains a novel function. The latter two processes can result in paralogs that persist for a substantial length of time (although a few exceptions have been proposed, such as pseudogene resurrection [27]). Throughout this work, the term duplication will be limited to the description of the latter two processes. Although gene loss can be as important as duplication for shaping genomes [28], we have avoided making major conclusions based on gene loss since many organisms included in these analyses lack complete genome sequences.

Inclusion of AGPase sequences from the moss *Physcomitrella patens* [29], which has 7 large subunit and 4 small subunit genes, placed a major constraint on the earliest divergence within the large subunit family since the moss sequences were intermixed with angiosperm sequences. This indicates that the earliest duplication in the large subunit family occurred prior to the divergence of angiosperms and mosses, more than 400 million years (MY) ago [30]. Since the rate of synonymous evolution in angiosperms varies from $\sim 2 \times 10^{-9}$ to $\sim 10 \times 10^{-9}$ synonymous substitutions per synonymous site per year [4-7], values of K_S in excess of 2 are expected for some comparisons, which may make estimates of K_S problematic [31].

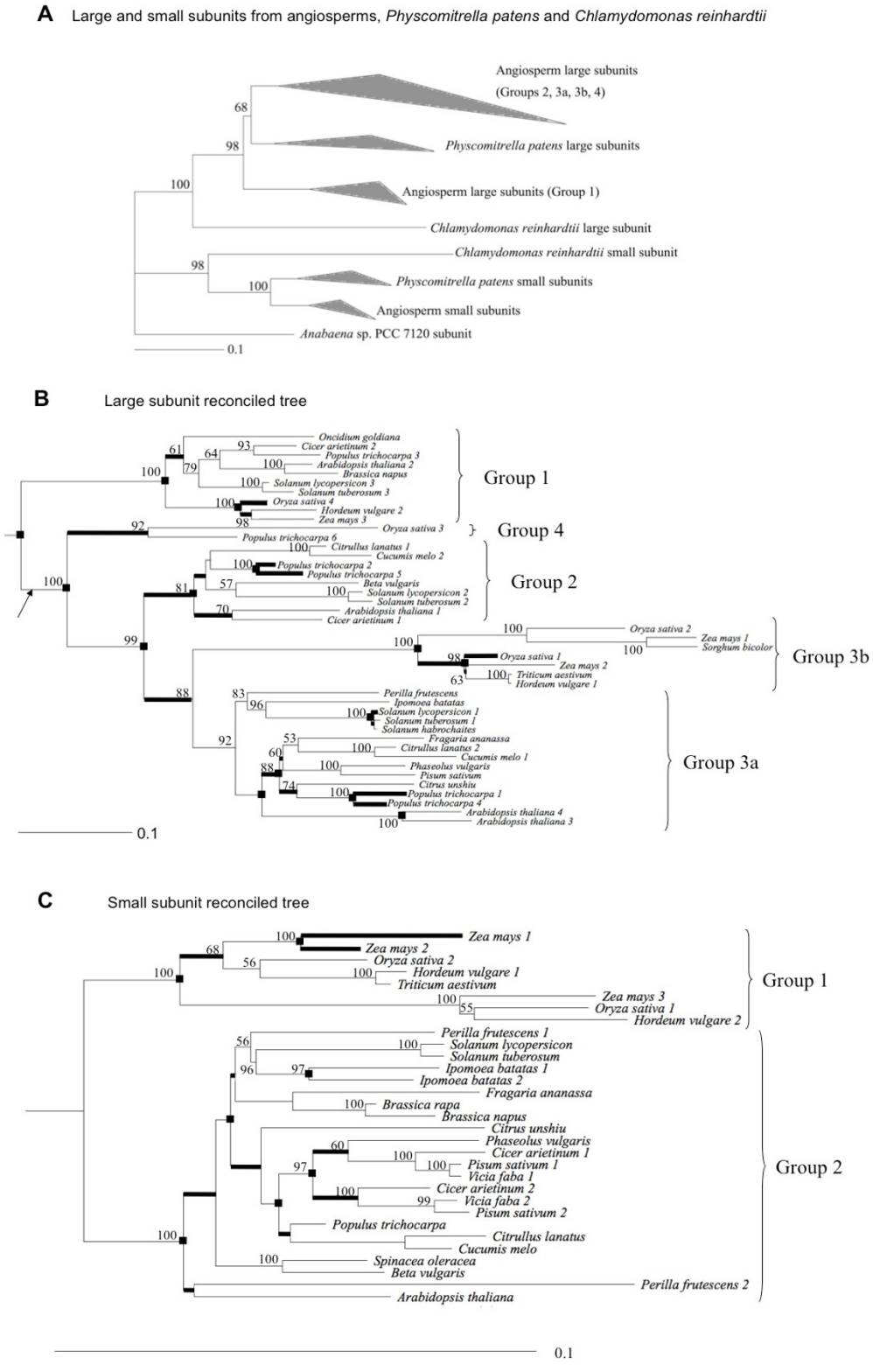
A
Large and small subunits from angiosperms, Physcomitrella patens and Chlamydomonas reinhardtii
Angiosperm large subunits
(Groups 2, 3a, 3b, 4)
Physcomitrella patens large subunits
Angiosperm large subunits (Group 1)
Chlamydomonas reinhardtii large subunit
Chlamydomonas reinhardtii small subunit
Physcomitrella patens small subunits
Angiosperm small subunits
Anabaena sp. PCC 7120 subunit
0.1
B
Large subunit reconciled tree
Oncidium goldiana
Cicer arietinum 2
Populus trichocarpa 3
Arabidopsis thaliana 2
Brassica napus
Solanum lycopersicon 3
Solanum tuberosum 3
Oryza sativa 4
Hordeum vulgare 2
Zea mays 3
Oryza sativa 3
Group 1
Group 4
Populus trichocarpa 6
Citrullus lanatus 1
Cucumis melo 2
Populus trichocarpa 2
Populus trichocarpa 5
Beta vulgaris
Solanum lycopersicon 2
Solanum tuberosum 2
Arabidopsis thaliana 1
Cicer arietinum 1
Group 2
Oryza sativa 2
Zea mays 1
Sorghum bicolor
Oryza sativa 1
Zea mays 2
Triticum aestivum
Hordeum vulgare 1
Group 3b
Perilla frutescens
Ipomoea batatas
Solanum lycopersicon 1
Solanum tuberosum 1
Solanum habrochaites
Fragaria ananassa
Citrullus lanatus 2
Cucumis melo 1
Phaseolus vulgaris
Pisum sativum
Citrus unshiu
Populus trichocarpa 1
Populus trichocarpa 4
Arabidopsis thaliana 4
Arabidopsis thaliana 3
Group 3a
0.1
C
Small subunit reconciled tree
Zea mays 1
Zea mays 2
Oryza sativa 2
Hordeum vulgare 1
Triticum aestivum
Zea mays 3
Oryza sativa 1
Hordeum vulgare 2
Group 1
Perilla frutescens 1
Solanum lycopersicon
Solanum tuberosum
Ipomoea batatas 1
Ipomoea batatas 2
Fragaria ananassa
Brassica rapa
Brassica napus
Citrus unshiu
Phaseolus vulgaris
Cicer arietinum 1
Pisum sativum 1
Vicia faba 1
Cicer arietinum 2
Vicia faba 2
Pisum sativum 2
Populus trichocarpa
Citrullus lanatus
Cucumis melo
Spinacea oleracea
Beta vulgaris
Perilla frutescens 2
Arabidopsis thaliana
Group 2
0.1

Figure 1

Figure 1
Reconciled large and small subunit trees. A) Amino acid tree of the large and small subunits from angiosperms, *Physcomitrella patens* and *Chlamydomonas reinhardtii*. The topology of the tree was determined by ML using aligned amino acid sequences using PhyML. Branch lengths reflect numbers of amino acid substitutions per site. The branches within groups have been replaced by the grey triangles. ML bootstrap values are indicated above branches, and the bar shows the number of amino acid substitutions per site. B) Angiosperm large subunit reconciled tree. C) Angiosperm small subunit reconciled tree. The topology of the trees shown in B) and C) was determined by ML using aligned cDNA sequences using GARLI. Branch lengths reflect numbers of amino acid substitutions per site as estimated by AAML and the scale bar shows the number of amino acid substitutions per site. ML bootstrap values > 50% are indicated above branches. Reconciled tree analyses (using the gene trees shown and the species tree shown in Additional file 1) were conducted using GENETREE. Black boxes at nodes indicate duplication events. The arrow in B) indicates the divergence of *Physcomitrella patens* from angiosperms. The trees in B) and C) were rooted with the AGPase large and small subunit from *Chlamydomonas reinhardtii* respectively. Thicker lines indicate branches that follow duplication events and have $K_S < 0.1$ (based upon ML estimates of synonymous branch lengths).

Since many divergence times for plants can be constrained to reasonable ranges, it should be possible to estimate absolute rates of AGPase subunit amino acid evolution and establish whether they correlate with estimates of ω. However, this requires differentiating between speciation and gene duplication events in AGPase phylogeny. Gene family phylogenies reflect both speciation and duplication events, and these events can be distinguished by reconciled tree analyses if the gene and species trees are known. Gene tree parsimony [32] is the most commonly used reconciled tree method, and the only approach practical for even moderately sized phylogenies at this time. Reconciling the AGPase gene trees with the best available estimate of the land plant species tree (Additional file 2) revealed 11–14 large subunit duplications (Figure 1B) and 5–7 small subunit duplications (Figure 1C). It may be appropriate to view the lower estimates, which are based upon well-supported nodes, as the primary results since they are based on modified versions of the gene trees (Additional file 3A, B) in which the topology was rearranged near poorly supported nodes to increase congruence with the species tree (Methods). In contrast, the higher estimates were based only on reconciling the optimal estimates of the gene trees (Figure 1B, C) with the species tree. Regardless, both analyses indicate that the large subunit genes underwent a larger number of duplications than the small subunit genes.

AGPase large subunits have narrower tissue-specificity than small subunits [33-38], and the large subunit phylogeny appears more complex (with four major clades some of which include both monocots and eudicots; Figure 1B) than the small subunit phylogeny (Figure 1C). Large subunit group 1 genes are predominantly expressed in leaves, group 2 genes are expressed both in source and sink tissues, group 3 genes are expressed sink tissues (these genes are subdivided into group 3a in eudicots and group 3b in monocots), and group 4 corresponds to a clade of two sequences that have not been characterized yet in terms of function and expression patterns (Figure 1B). Some of the major large subunit clades arose prior to the divergence of monocots and eudicots, and the optimal placement of the AGPase large subunit sequences from *Physcomitrella* suggests that the first duplication in the large subunit happened around 400 MY ago (Figure 1A). In contrast, there is no evidence that angiosperm small subunits underwent a duplication prior to the divergence of monocots and eudicots, and we have divided them into a monocot clade (group 1) genes and a eudicot clade (group 2). These results emphasize that the large subunit underwent a larger number of duplications than did the small subunit and that only large subunit duplications began before the divergence of monocots and eudicots.

Absolute rates of AGPase evolution

Absolute rates of amino acid evolution for AGPase subunits were estimated by examining terminal branch lengths for divergences that reflect speciation events with known divergence times (these divergence times are presented in Additional file 2). This approach is called the tip procedure since it involves only terminal branches (Methods), and it revealed that the average rate of evolution for the large subunit was 2.7-fold faster than that of the small subunit (Figure 2). This rate difference was both congruent with the difference in ML estimates of ω [2] and highly significant ($P = 0.0006$ by Student's unpaired *t*-test). Our conclusions were unchanged if we limited consideration to strongly supported duplication events (those retained when bootstrap support was considered; see Additional file 3).

Estimates of the absolute rate of amino acid substitution for AGPase subunits obtained by the penalized likelihood (PL) method (Figure 3) were very similar to those obtained using the tip procedure (Figure 2). Using gene trees in which poorly supported nodes were rearranged to minimize number of duplications yielded similar results (Additional file 4, Figure 2). Thus, very similar estimates of the absolute rate of amino acid substitution were

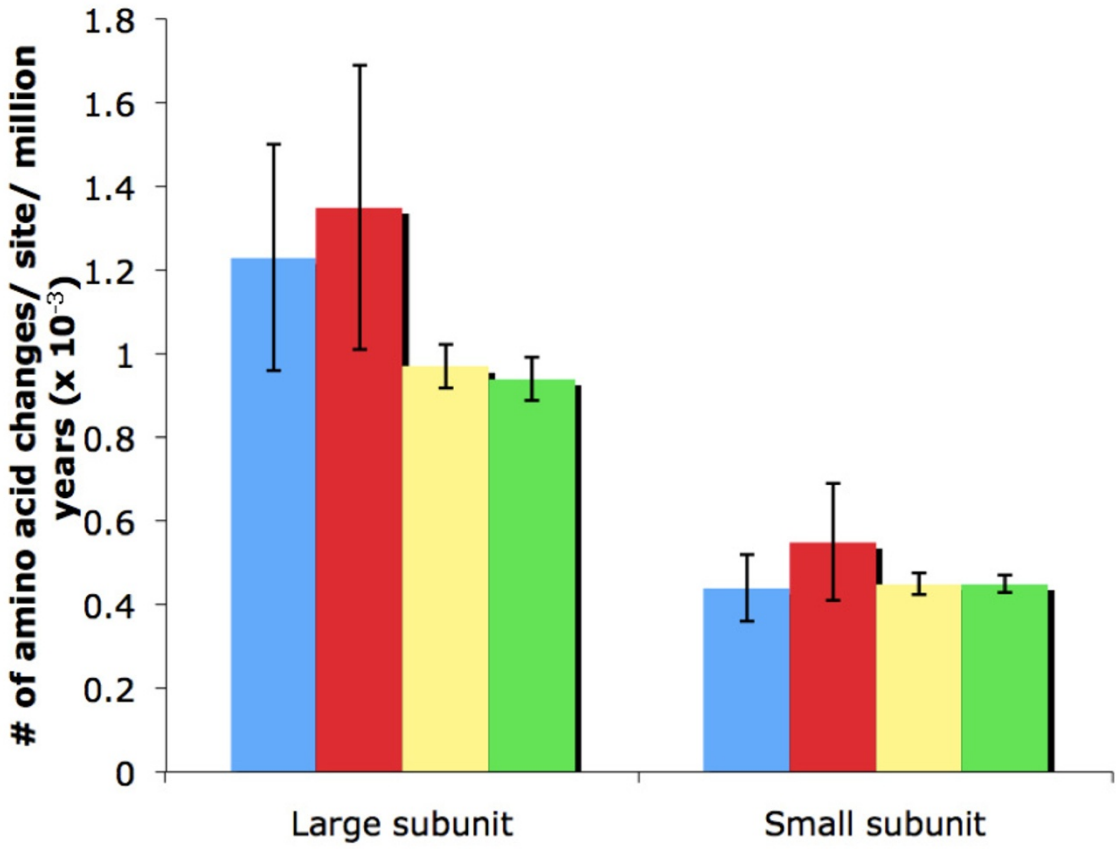

Figure 2
Absolute rate of evolution of the large and the small subunit of AGPase from angiosperms (measured in aass MY^{-1}). The blue bars indicate the average rates (amino acid substitutions per site per million years; aass MY^{-1}) estimated from the most recent dated speciation events to present sequences of the trees shown in Figure 1B and 1C. The red bars indicate the average aass MY^{-1} estimated from the most recent dated speciation events to present sequences of the trees shown in Additional file 3A and 2B. The yellow bars indicate the average aass MY^{-1} estimated from all branches in Figure 3A and 3B. The green bars indicate the average aass MY^{-1} estimated from all branches in Additional file 4A and 3B. The error bars indicate 2× standard error.

obtained despite the different assumptions made by the tip procedure and the PL method.

The absolute rate of synonymous evolution was estimated using ML estimates of K_S (Additional file 5A, B). The tip procedure resulted in virtually identical rates for both large and small subunit genes (6.5×10^{-9} synonymous substitutions per synonymous site per year; Additional file 5C). PL rate estimates were $5.5 \times 10^{-9} \pm 0.3 \times 10^{-9}$ and $6.3 \times 10^{-9} \pm 0.2 \times 10^{-9}$ (mean ± standard error) synonymous substitutions per synonymous site per year for the large and small subunit, respectively (data not shown). The slightly lower estimates based upon PL are consistent with saturation being a problem, but presumably only for the deepest branches in the tree. All of these values are well within the range of previous estimates for a variety of angiosperm genes (which range from approximately 2×10^{-9} to 10×10^{-9} synonymous substitutions per synonymous site per year and exhibit some variation among lineages [4-7,39,40]). This suggests that there are little to no constraints on the synonymous sites of angiosperm AGPase genes and, when combined with estimate of the absolute rate of sequence evolution, that there was minimal bias in our estimates of ω.

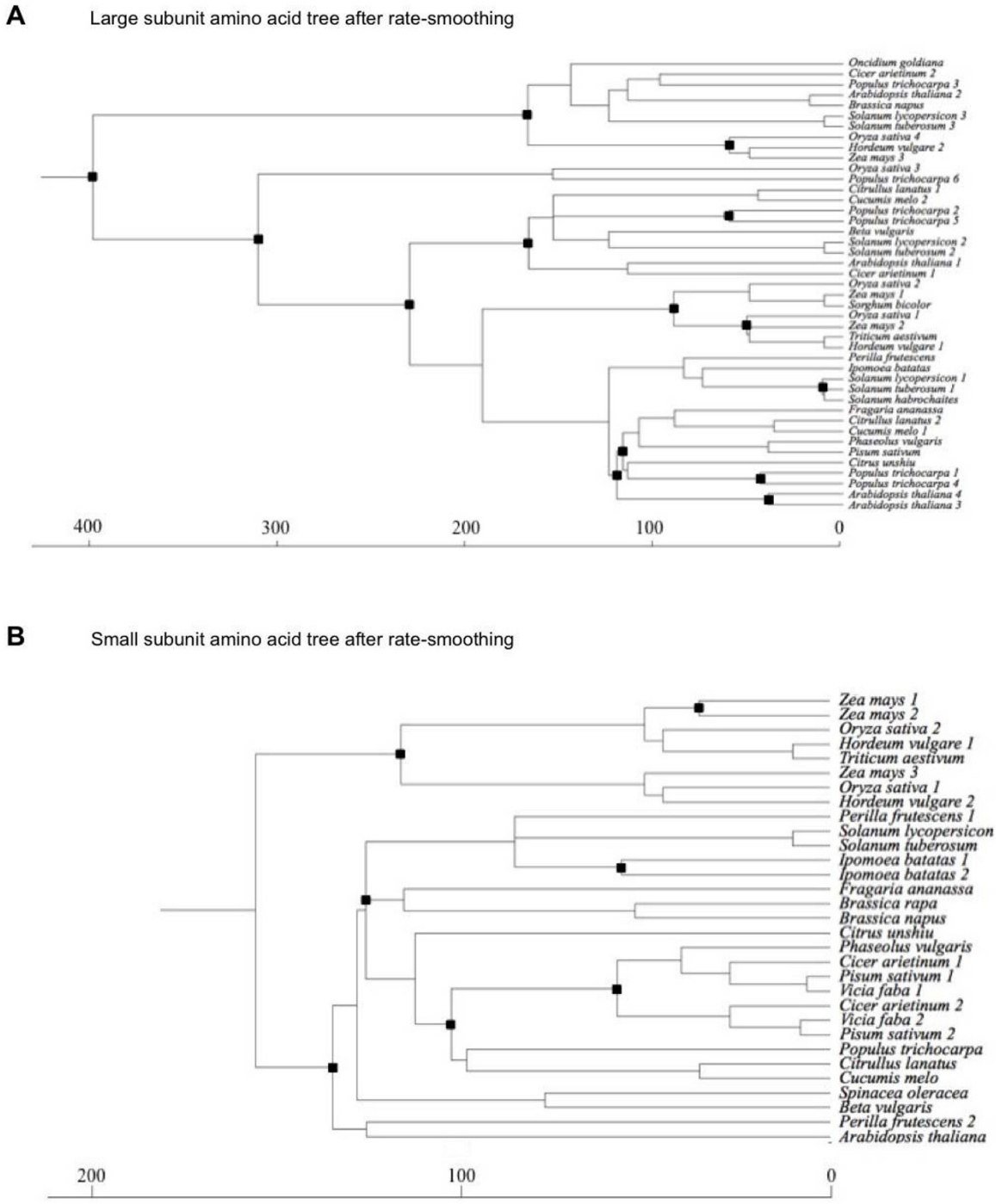

Figure 3

Phylogenetic trees of the large and the small subunits from angiosperms after rate-smoothing. The trees in A) and B) are the trees shown in Figure 1B and 1C respectively after rate-smoothing. Rate-smoothing was done by using penalized likelihood as implemented in the r8s software [79]. Black boxes indicate duplication events.

Does the large subunit show temporary or permanent elevation of ω?

Estimates of the mean rate of evolution for AGPase, whether based upon ω [2] or the absolute rate of amino acid substitution (Figure 2), show a substantially higher rate for the large subunit. The existence of these rate differences despite identical sensitivities to mutations when expressed in bacteria suggests that there are important differences *in planta*. Transient increases in the evolutionary rate might explain the observed rate differences if they are more common for the large subunit. Large-scale analyses provide evidence for transient increase in the rate of non-synonymous evolution. Indeed, paralogs with a recent origin (defined as those with $K_S < 0.1$) accumulate more non-synonymous mutations per non-synonymous site (and thus have a higher K_A) relative to the number of the synonymous mutation per synonymous site than older duplicates [23]. Thus, ω is expected to be elevated for paralogs with $K_S < 0.1$ relative to those with $K_S > 0.1$. Lynch and Conery [23] interpreted this phenomenon as reflecting a temporary relaxation of constraints, positive selection, or a combination of both phenomena. Thus, it is important to consider the potential impact of the elevation of ω on our analyses of the evolutionary processes that shaped the AGPase gene family.

The large subunit underwent more gene duplications than did the small subunit (Figure 1). Thus, the higher mean estimates of ω for the large subunit might reflect a larger number of periods during which ω is elevated (due to temporary relaxation of constraints and/or positive selection) rather than permanent relaxation of constraints for the large subunit. To distinguish between transient and permanent relaxation of constraint we tested whether the non-synonymous rate was increased after duplication and if the increase is sufficient to explain the observed differences in the mean rate. Branches in both large and small subunit gene trees (Figure 1) were placed into two groups, the first of which (class 1) contained branches that follow a duplication event with K_S = ~0.1 (these branches are shown in Figure 1B, C and Additional file 3). The second group (class 2) contained all other branches (branches that follow either a speciation event or a duplication and have $K_S > 0.1$). We examined two nested models using the likelihood ratio test (LRT) [41-45] to determine whether ω for class 1 branches is higher than ω for class 2 branches for either subunit using CODEML (included in PAML software). The more complex model, which assumes two different ω values (one ω for class 1 and one ω for class 2), was favored over the null hypothesis model, which assumes a single ω for both classes, for both the small subunit ($2\delta lnL = 18.4$; $P < 0.001$) and the large subunit ($2\delta lnL = 7.48$; $P = 0.005$) (for details see Table 1). The ω estimate for short branches following duplications was 1.3 to 1.5-fold greater than the ω estimate for all the other branches when the large subunit was examined and 2 to 2.8-fold greater for the small subunit (Table 1). These results support periods of increased ω after duplications in the AGPase gene family, due to the temporary relaxation of constraints, positive selection, or both. However, these results also indicate that this phenomenon cannot explain the differential rates of amino acid sequence divergence of the two AGPase subunits, since estimates of ω for the large subunit are 2.6-fold greater than estimates for the small subunit (Table 1). Instead these results suggest that the small subunit is permanently subject to greater purifying selection than is the large subunit.

What is the role of positive selection in AGPase evolution?

To examine the potential role of positive selection in AGPase evolution, we used ML to compare two distinct sets of models. The first model set contains a neutral (null) model M1a allowing two categories of sites, one with ω = 0 and one with ω = 1, and model M2a that adds an extra category of sites with ω > 1. The second includes a neutral (null) model M7 assuming that ω is β-distributed among sites and model M8 that adds an extra category of sites with ω > 1 [46]. Neither of the models with

Table 1: Temporary relaxation of purifying selection, positive selection, or both after duplications in the large and small subunit of AGPase from angiosperms.

Branches	Large subunit	p-value	Small subunit	p-value
Class 1 (Figure 1)	0.114	0.005	0.058	< 0.001
Class 2 (Figure 1)	0.086		0.029	
All branches (Figure 1)	0.090		0.033	
Class 1 (Additional file 3)	0.125	< 0.001	0.081	< 0.001
Class 2 (Additional file 3)	0.086		0.029	
All branches (Additional file 3)	0.089		0.034	

The branches of the trees shown in Figure 1B and Additional file 3A (large subunit) or Figure 1C and Additional file 3B (small subunit) were classified into two groups. One group includes the short branches ($K_S < 0.1$) following a duplication event (post-duplication; class 1) and the other group includes all the other branches (class 2). The overall ω value of each group and the overall ω value of all branches were estimated by CODEML. The LRT was used to compare the model with two different ω values (one for class1 branches and one for class 2 branches) to the null hypothesis model with a single ω value.

positive selection was significantly better than the null model based upon the LRT for either of the subunits when the tests were applied to the complete trees for the small subunit (Figure 1C) or large subunit (Figure 1B) (data not shown). Likewise, neither of the models that include positive selection was significantly better when the tests were applied to the individual groups (Figure 1B, C) within each subunit (groups 1, 2, 3A, and 3B for the large subunit as well as groups 1, and 2 for the small subunit) (data not shown). These results indicate either that positive selection has not played a role in the evolution of the large and small subunit of AGPase in the angiosperms or that these tests have insufficient power.

To increase the power of our tests for positive selection, we used branch-site models to examine the potential for positive selection at specific sites in all tree branches separately (Methods). There are branches in the large or small subunit tree on which specific sites may be subject to positive selection (Figure 4), with a total of 0.8 and 0.2 sites/branch potentially affected by positive selection for the large and the small subunit respectively (Additional file 6). However, the limited number of sites potentially affected by positive selection suggests that purifying selection is the major force in the evolution of both AGPase subunits in angiosperms. Thus, positive selection cannot explain the different rates of amino acid evolution for the AGPase subunits.

Functional divergence of AGPase subunits

Gu [21,22,47] proposed that specific sites in proteins have the potential to undergo two distinct types of divergences after gene duplication (e.g., divergence among the different groups within the large and small subunits), and these types were designated type-I and type-II divergence. Sites that have undergone type-I divergence are conserved in one group but variable in another group while type-II sites are fixed in both groups but differ between groups [21,22].

Tests for types-I and -II divergence based upon the relevant coefficients of divergence (θ), which correspond to the probability that a specific site has undergone type-I or -II divergence in a pairwise comparison, were proposed by Gu [21,22,48]. These tests for types-I and -II establish whether the relevant θ value is significantly greater than zero. We conducted all possible pairwise comparisons among small (groups 1 and 2) and large (groups 1,2, 3a and 3b) subunits with the exception of group 4 of large subunits (that were excluded because the group included only two sequences). All of type-I coefficients (θ_I) of functional divergence were significantly greater than zero while none of the type-II coefficients (θ_{II}) were significantly greater than zero (Table 2). The estimates of θ_I are also much larger than the estimates of θ_{II}, suggesting that type-I divergence is the dominant pattern of sequence evolution for AGPase large and small subunit groups.

Sites contributing to functional divergence among AGPase groups

We identified the sites likely to be involved in the changes of functional constraints between groups revealed by the significant values of θ_I using a posterior probability analysis (Additional file 7). A greater number of large subunit sites appear to have undergone type-I divergence than the number of small subunit sites. Specifically, for an alignment with 453 amino acids we found evidence that 78 large subunit sites and 13 small subunit sites show evidence for type-I divergence (Additional file 8). These sites appear randomly distributed with respect to secondary structure (data not shown), but pairwise comparisons among groups in the large and the small subunit reveal some non-random patterns in the distribution of sites that are conserved in one group but not another. For instance, large subunit group 1 proteins (leaf isoforms) had more conserved sites in the N-terminus while group 3a proteins (sink isoforms) exhibited more conserved in the C-terminus. The large subunit crystal structure has not been elucidated yet. However, the high degree of amino acid sequence identity (~43%) and similarity (~61%) [2] (Additional file 1) to the small subunit along with structure modelling (unpublished data) strongly suggest that the 3D structure of the large subunit is almost identical to the known structure of the small subunit. The N-terminal domain includes the active site that resembles a Rossman fold while the C-terminal domain is a β-helix that extensively interacts with the N-terminal domain based on the structure of the potato tuber small subunit elucidated by Jin et al. [49]. Both domains are important for stability and catalytic/allosteric properties of the enzyme [49-52], but the non-random spatial distribution of type I sites clearly suggests these sites should be targeted in mutational studies focused on the analysis of AGPase structure/function relationships.

In contrast to the significant estimates of θ_I, estimates of θ_{II} (the type-II coefficient) were not significantly greater than zero. However, the power of the test for type-II divergence is unclear. To determine whether there was any evidence for type-II divergence we examined the sequences to determine whether any specific sites show evidence for this type of divergence. The likelihood that these sites reflect *bona fide* instances of type-II divergence would be increased if they correspond to sites that have been identified in mutagenesis studies. The posterior ratio test, using a cut-off 2 (i.e., a posterior probability of 0.5), identified a few potential type-II sites (Additional file 9). Some sites were also identified by analyses of positive selection and one has an important function revealed by mutagenesis (see Discussion), so it is reasonable to postulate that

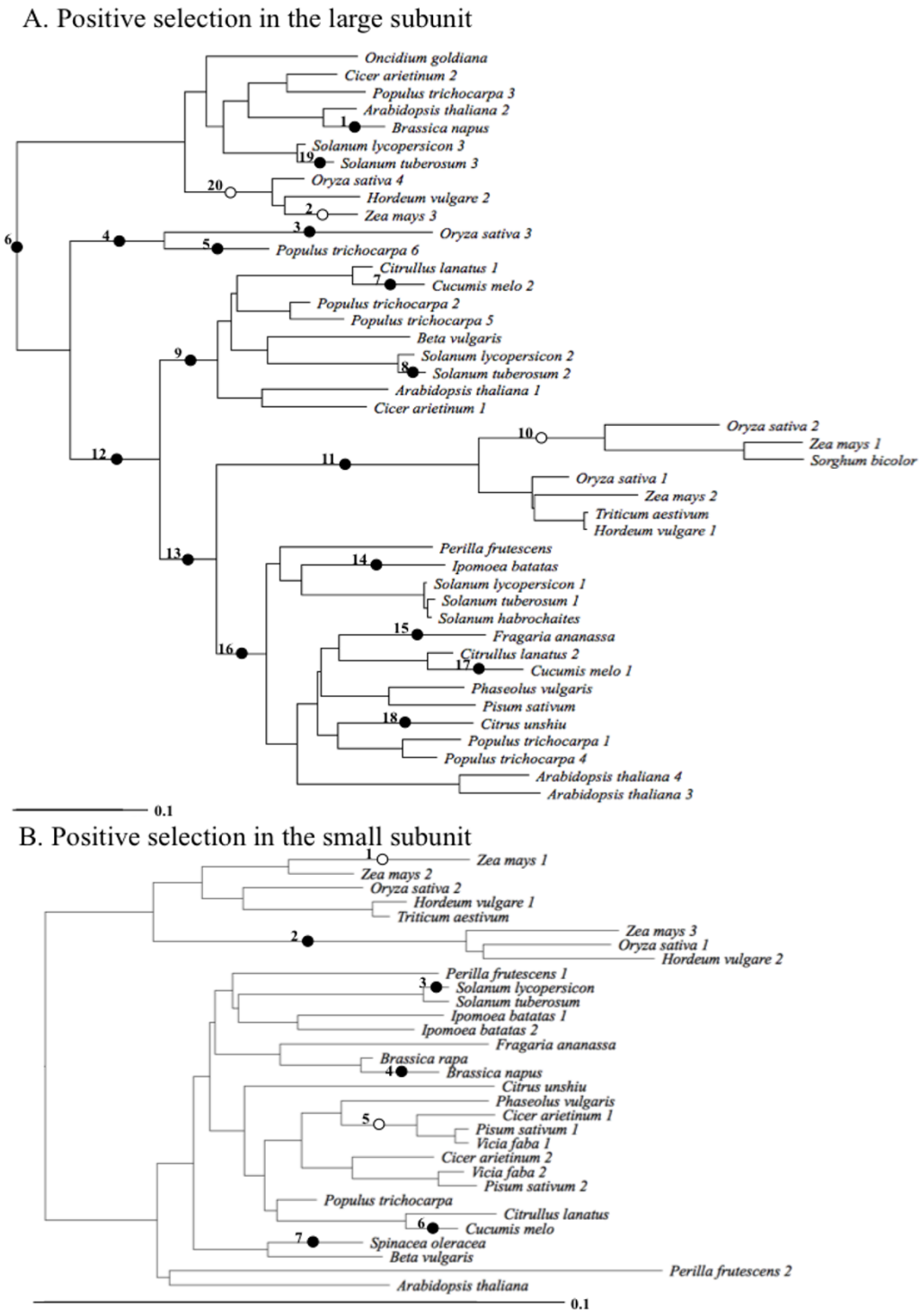

Figure 4

Positive selection in the large and the small subunit of the angiosperms. The trees shown in A) and B) have the same topology as the trees shown in Figure 1B and 1C, although the shown here are unrooted. White circles indicate branches where positive selection was detected only by Test 1 but not Test 2 (described in Methods). Black circles indicate branches where positive selection was detected by both Test 1 and Test 2. The branches where positive selection is detected are numbered.

Table 2: Type I and II functional divergence between large and small subunit groups.

	Large subunit						Small subunit
	Group 3b/ Group 3a	Group 3b/ Group 2	Group 3b/ Group 1	Group 3a/ Group 2	Group 3a/ Group 1	Group 2/ Group 1	Group 1/ Group 2
				Type I divergence			
Theta-I	0.261*	0.383*	0.448*	0.220*	0.331*	0.273*	0.263*
SE	0.069	0.095	0.097	0.060	0.072	0.076	0.084
				TypeII divergence			
Theta-II	0.016	0.072	0.080	0.016	0.058	0.001	0.028
SE	0.061	0.052	0.054	0.059	0.061	0.053	0.033

Coefficients of type I (θ_I) and II (θ_{II}) functional divergence were estimated using DIVERGE. * denotes statistical significance at 5% level of confidence. SE: standard error.

at least some of these sites reflect genuine instances of type-II divergence that have contributed to the specialization among the large and small subunits.

Discussion

The large and the small subunit of AGPase in plants exhibit considerable sequence identity [1,2] and they reflect a gene duplication that occurred prior to the divergence of land plants and green algae. Both subunits are equally sensitive to activity-altering amino acids, at least when expressed in a bacterial system [2]. However, the small subunit of angiosperms is more conserved than large subunit based upon estimates of the rate of evolution (both estimates of ω [2] and estimates of the absolute rate of amino acid evolution). These results suggest saturation has not had a major impact upon the estimation of ω for angiosperm AGPases, and this may reflect, at least in part, the limited codon bias of these genes (data not shown) since codon bias can have a major impact on the estimation of K_S [8]. They also suggest that estimation of the absolute rate of amino acid evolution provides a valid method that can be used as an alternative to ω analysis when the use of ω is not appropriate (e.g., ancient divergences, especially for gene families with strong codon bias, and for gene families where there is evidence for strong selection on synonymous sites). Taken as a whole, these results confirm that plant AGPases represent a genuine paradox: the large and small subunits exhibit similar sensitivities to activity-altering changes but differ almost 3-fold in their rates of non-synonymous evolution.

Although a temporary elevation of ω after duplication was observed for both large and small subunit gene families, this transient elevation cannot account for the overall difference in ω value (and the related difference in the overall rate of amino acid substitutions) between the two subunits. Additionally, although both subunits appear to have been subject to positive selection the observed rate differences are too large to be explained by postulating that they reflect greater differences in positive selection. Based upon the falsification of these hypotheses, we conclude that the small subunit has been evolving under stronger purifying selection than the large subunit.

Consistent with the numbers of large and small subunit genes found in the sequenced plant genomes [2,35,38], reconciled tree analyses indicated that large subunit genes underwent more duplications than small subunit genes. Both phylogeny and molecular clock analyses indicate that the initial duplication in the large subunit family of angiosperms occurred *ca*. 400 MY ago, close to the divergence of angiosperms and bryophytes [30] (Figure 3A and Additional file 4A). In contrast, the oldest retained small subunit duplicates date back to 120–140 MY ago, after the divergence of monocots and eudicots (Figure 3B and Additional file 4B). The reason why the large subunit had more ancient and duplications than did the small subunit remains enigmatic. Macroevolutionary models that can be applied to gene families are poorly developed, so it remains possible that the observed difference is coincidental. Alternatively, the large subunit might have a greater ability to undergo subfunctionalization after duplication. The fact that 7 large subunit genes and 4 small subunit genes can be identified in the *Physcomitrella* genome suggests similar patterns in both mosses and angiosperms. However, more rigorous tests to distinguish among these scenarios must await both the acquisition of additional data from additional deep-branching land-plant lineages (e.g., liverworts and hornworts) and the development of better models of gene family macroevolution.

Studies of expression patterns of AGPase genes in several species, including rice, *Arabidopsis*, potato, tomato and

barley, have shown that the large subunit is tissue specific while the small subunit is more broadly expressed [33-38]. Based on these studies, the major large subunit groups (Figure 1B) are likely to be expressed in different tissues in most or all plants. The tissue-specificity of large subunit genes suggests the expression patterns of these genes might undergo subfunctionalization after duplication, as predicted by the "DDC model" of gene duplication [53]. The DDC model predicts that duplicated genes are preserved by complementary changes in their expression pattern (e.g., a broadly expressed gene might undergo duplication and have one duplicate expressed in a specific tissue like leaves while the other duplicate is expressed elsewhere). Although the potential for subfunctionalization due to changes in gene expression to preserve duplicated genes is generally accepted [54,55], it also remains possible that distinct AGPase genes have specialized in terms of protein function (e.g., their pH optima might have shifted based upon the specific tissues in which the paralogs are expressed). The relative contributions of subfunctionalization and specialization or neofunctionalization to gene family evolution are open questions [56,57], and it is unclear that there is any reason why the large subunit would be more likely to undergo specialization at either the protein or gene expression level. Such a model, which postulates that subfunctionalization of gene expression was followed by specialization, is similar to a combined model called subneofunctionalization [57]. The subneofunctionalization model postulates that subfunctionalization occurs shortly after duplication while neofunctionalization is a more prolonged process [56,57]. If the combined model were applied to AGPases, the initial preservation of paralogous AGPase genes immediately after duplication might reflect subfunctionalization but this process would be followed by adaptation to the more specialized domains of expression in which each of the paralogs are expressed. Either a neofunctionalization or a subneofunctionalization model would be consistent with the evidence that different large subunit groups have functionally diverged from each other at the protein level (Figure 4, Table 2). However, subneofunctionalization provides a means to directly link the divergence of expression patterns and divergence at the protein level. Corroborating the subneofunctionalization will require correlating gene expression and sequence divergence for a large number of plants.

Differences in their patterns of expression represent the major difference between the large and small subunits that could explain the differences in their rates of evolution, since broadly expressed genes are more conserved than tissue-specific genes [58-60]. However, this raises the question of why broadly expressed genes, like AGPase small subunit genes, exhibit slower rates of evolution. Although it may simply be that mutations in broadly expressed genes have a greater impact on fitness or because these genes have to function in multiple cellular environments [58], a simpler explanation might be that small subunit genes have to function with multiple large subunit genes. Georgelis et al. [2] presented data consistent with this possibility, since they showed that the effects of several amino acid changes in the maize endosperm small subunit on enzyme activity depended on the identity of the large subunit [maize endosperm large subunit (SHRUNKEN-2)(SH2) and maize embryo large subunit (AGPLEMZM) were used]. Both SH2 and AGPLEMZM are members of group 3b (Figure 1B), so these results suggest that even fairly similar large subunit genes can interact differently with small subunits.

In addition to the potential for subfunctionalization due to changes in gene expression, the observation that estimates of θ_I for large subunit groups were significantly greater than zero suggests that it will be possible to attribute differences among AGPase genes to specific amino acid changes. We found 99 candidate sites of the large subunit likely to have been involved in rate shifts (either type-I or -II divergence; Additional files 8 and 9). At least some of these putative rate shift residues are likely to have contributed to functional changes among the different large subunit groups. The estimate of θ_I for the small subunit groups 1 (monocot) and 2 (eudicot) is also significantly greater than zero. It was possible to find evidence for 13 type-I candidate sites in the small subunit alignment. Like the large subunit, the estimate of θ_{II} for the small subunit was not significantly greater than zero. Nonetheless, there were two potential type-II sites could be identified (Additional file 9).

A total of 21 candidate sites for positive selection could be identified in the large subunit branches following duplications that led to different groups (Branch numbers: 4,6,9,11,12,13,16 shown in Additional file 6), and six of these sites overlapped with the set of sites that appear to have undergone either type-I (sites 341, 364, 445) or type-II (sites 106, 114, 382) divergence. Biochemical and genetic studies confirm that at least one of the sites (sites 106) is important for AGPase activity. This site is a threonine (T) in large subunit groups 3a and 3b but a lysine (K) in groups 1 and 2 and in all small subunits. The potato tuber large subunit, which falls into group 3a, has a T at site 106 and it forms an inactive complex if it is combined with an inactivated potato tuber small subunit [61]. Changing this T in the potato tuber large subunit to a K actually results (the T106K mutant) in a complex with some activity with the same inactivated potato tuber small subunit [61]. These results were interpreted as evidence that the large subunit lost its catalytic ability partly because of the K to T change. Although the K residue at site 106 may be necessary for catalysis if the small subunit is

inactive, another model that explains the data would be one in which the wild-type large subunits (which have a T at site 106) require prior catalysis by the small subunit before they perform catalysis by themselves. Such a catalytic mechanism has been proposed for the *Escherichia coli* AGPase [62]. The T residue at site 106 is absolutely conserved in large subunit groups 3 and the branch-site model suggests positive selection for the T immediately following the duplication that generated group 3 (Additional file 6). This suggests that this change was important for enzymatic activity and beneficial for the plants. Indeed, the overall activity of a complex that includes the T106K mutant of the potato tuber large subunit and the wild-type potato tuber small subunit showed significantly reduced activity relative to wild-type potato tuber AGPase [63].

One potential type-II site (site 507) and four sites that represent candidates for positive selection on branches that immediately follow duplications (sites 104, 230, 441, 445) have been shown to be important for the allosteric properties of AGPase [[49,64,65], Hannah personal communication]. However, most of the candidate sites for either rate shifts or positive selection do not have a known function.

The existence of type-I and type-II divergence among AGPase subunit groups along with the detection of positively selected sites after duplications that led to different groups in the large subunit provide evidence for functional divergence especially among the large subunit groups. Our data are consistent with biochemical studies showing that the four possible AGPase complexes in *Arabidopsis*, which have a single functional small subunit gene and four distinct large subunit genes (belonging to different groups in Figure 1B), have different kinetic and allosteric properties [16]. There is further evidence for functional divergence among plant AGPases, since the maize and barley endosperm AGPases are less dependent than potato tuber AGPase on the allosteric activator 3-PGA for activity and the maize endosperm AGPase is more heat labile than potato tuber AGPase [17-20]. Functional divergence among the different subunit groups was also suggested by Georgelis et al. [2], who showed that all groups of large subunit genes have ω values (which range from 0.073 to 0.132) that exceed those for small subunit genes (which range from 0.027 to 0.054). These results are consistent with the rate shifts within the large and the small subunit families that were observed in the present study and they further imply that the various groups of plant AGPases have undergone functional divergence. The present study also identifies specific residues that are likely to have contributed towards that divergence. Site directed mutagenesis of these candidate sites is likely to shed some light on their functions and reveal the proportion of candidate sites that reflect type II error for tests to identify sites subject to rate shifts and positive selection.

A number of angiosperm AGPases have been successfully expressed in *E. coli* and purified, including maize endosperm AGPase [66], potato tuber AGPase [67], and all possible *Arabidopsis* AGPase complexes [16]. The most straightforward candidate sites to test are the type-II sites and those subject to positive selection, where it is possible to change the relevant residue either to the amino acid present in the other group (for type-II sites) or the ancestral amino acid (for positively selected sites). Testing type-I sites may be more challenging, since there is not a clear way to swap the residues present in a member of the focal group of proteins with that in a different group. However, it is reasonable to predict that any of the residues present in the group of proteins for which the site is variable should alter the biochemical activity of a protein in which the site is invariant. Regardless of the specific strategies for generating mutants, mutant large or small subunits could be expressed in *E. coli* along with a wild-type version of the other subunit, the relevant complex purified, and the properties of the enzyme determined to allow the impact of the mutations to be studied. This will allow elucidation of the importance of the sites in enzyme activity. Ultimately, it will be interesting to determine whether these sites are important for the kinetic and allosteric properties of AGPase, for enzyme stability for the pH optimum, or for multiple properties.

Conclusion

Herein, we validated and extended the observation, initially based upon estimates of ω [2], that the AGPase large subunit accumulated non-synonymous substitutions more rapidly than the small subunit in angiosperms by estimating absolute rates of amino acid change. The earliest duplication in the large subunit family of angiosperms was close to the time that angiosperms and mosses diverged (~400 MY ago;[30]). The large subunit underwent a larger number of duplications than the small subunit, which only began to duplicate after the divergence of monocots and eudicots. We suggest that the large subunit evolved faster due to permanently relaxed constraints since positive selection and sporadic episodes of relaxed constraints cannot account for the different rates of evolution between the large and the small subunit.

Large subunit genes exhibit narrower tissue specificity than small subunit genes in terms of their gene expression patterns [33-38], and they are likely to have experienced subfunctionalization in terms of expression patterns. However, we use analyses of rate shifts and positive selection to demonstrate that different groups of both large and small subunits are likely to have diverged at the protein level. We have identified candidate amino acid sites

with the potential to account for the functional divergence and described strategies for site-directed mutagenesis experiments that could shed light into the specific roles of these sites.

Methods

Sequence Retrieval and Alignment

Full-length AGPase sequences from plants were retrieved from the NCBI database and the DOE Joint Genome Institute (JGI) web site, and the source and accession numbers of all sequences are presented in Additional file 10. DNA and protein sequence alignments were obtained using the MEGA software [68] with BLOSUM matrix followed by manual inspection. The poorly aligned N-termini (~70–80 amino acids for the large subunit and ~40 amino acids for the small subunit) were excluded from alignment. The large subunit amino acid numbers used correspond to the protein encoded by the maize *Shrunken-2* (*Sh2*) gene (Accession #: P55241) while the small subunit amino acid numbers used correspond to the protein, encoded by the maize *Brittle-2* (*Bt2*) gene (Accession #: AAQ14870).

Phylogenetic analysis

Estimates of AGPase gene trees based upon nucleotide data were obtained using the GARLI (Genetic Algorithm for Rapid Likelihood Inference) software [69]. Estimates of phylogeny estimated for protein sequence alignments were obtained either using neighbor joining in the MEGA software or ML in the RAxML software [70], using the JTT model [71] in both cases (to estimate distances or calculate likelihoods). Bootstrap support [72] was calculated using 100 replicates.

Branch lengths were estimated by ML, with those form amino acid trees based estimated using AAML, using the Dayhoff (PAM) model [73] with Γ-distributed rates. Nonsynonymous substitutions per nonsynonymous site (K_A) and synonymous substitutions per synonymous site (K_S) and the ratio of these values ($\omega = K_A/K_S$) were estimated using CODEML. AAML and CODEML are programs in the PAML (Phylogenetic Analysis by Maximum Likelihood) package [74].

Reconciled tree analysis

Reconciled tree analyses map a gene tree onto a species tree [75,76], and most commonly used procedure for doing this is gene tree parsimony, which minimizes the number of duplication events needed to explain a specific gene tree given the species tree [32]. However, some error is typically associated with estimates of phylogeny for individual genes, and accommodating error in gene trees is difficult in reconciled tree analyses [75]. Chen et al. [77] suggested an algorithm that rearranged gene trees to increase congruence with the species tree when nodes in the gene tree were poorly supported to limit the impact of error in the gene tree. We implemented this idea manually, by rearranging nodes in the gene tree with limited support (those with bootstrap support < 70%; see [78]) to increase the congruence with the species tree. This yielded two estimates of the numbers of duplications, one based on the optimal gene trees and a conservative estimate based on well-supported nodes.

Estimation of the absolute rate of evolution

Estimates of the absolute rate of amino acid evolution and synonymous site evolution for each subunit were obtained using the "tip procedure", which uses the average number of amino acid substitutions per site along unique paths from each tip (extant sequence) to a dated speciation event (allowing us to avoid pseudoreplication of rate estimates). In addition to this method, absolute rates of amino acid substitution were also obtained by estimating the age of each node after smoothing rates using penalized likelihood in r8s ("PL method") [79]. The PL method was used because the tip method is biased towards recent branches, although the PL method also has the potential to be affected by saturation, especially for synonymous sites.

Detection of branch-specific patterns of rate variation among sites and positive selection

Type-I and type-II functional divergence among large or small subunit groups was examined using the DIVERGE software [48], which implements the tests suggested by Gu [22,47] that can be used to determine whether the coefficients of divergence (θ_I and θ_{II}) are significantly greater than zero. Amino acid sites likely to have undergone types-I or -II divergence were detected as those with a posterior probability > 0.5–0.6.

Sites subject to positive selection were identified using the site, branch and branch-site models implemented in CODEML, using the model comparisons recommended by Yang et al. [80]. The first comparison was between model M1a, which includes two ω values (one for sites subject to purifying selection with $\omega < 1$ and one for neutral sites with $\omega = 1$) and model M2a (which adds positively selected sites [with $\omega > 1$] to model M1a). The second comparison was between model M7, which assumes values of ω at different sites are β-distributed, and model M8 (which adds positively selected sites to model M7). We also searched for positive selection by estimating different values of ω for each branch and using branch-site models. For the first test, we compared a model specified as model = 2 NSsites = 2 to model M1a [81]. For the second test, we compared a model specified as model = 2 NSsites = 2 to a model whose specifications are model = 2 NSsites = 2 fix_omega = 1 and omega = 1. When the likelihood ratio test was significant, the Bayes empirical Bayes

method was used to calculate posterior probabilities that sites were subject to positive selection [82].

Authors' contributions

NG conducted the experiments and carried out the analyses. NG, ELB and LCH conceived and designed the experiments. NG, ELB and LCH wrote the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Alignment of large and small subunits AGPases from angiosperms with protein domains highlighted. *The blue domain indicates the hypervariable N terminus of the large and the small subunit. The pink and green domains indicate the catalytic domain and the β-helix domain respectively. The yellow domain indicates the loop that connects the catalytic to the β-helix domain.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-232-S1.pdf]

Additional file 2

Species tree. *Times of divergence are indicated in million of years (MY) at nodes [83-92]. All divergence times were examined for consistency with the fossil record [93].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-232-S2.pdf]

Additional file 3

Reconciled large and small subunit trees. *A) Angiosperm large subunit reconciled tree. B) Angiosperm small subunit reconciled tree. The topology of the trees shown in A) and B) was determined by ML using aligned cDNA sequences analyzed by GARLI. Nodes with bootstrap values < 70% (Figure 1) were then rearranged to minimize the number of duplications (to increase congruence with the species tree). Branch lengths reflect numbers of amino acid substitutions per site, estimated AAML (with the scale bar showing the number of amino acid substitutions per site). Reconciled tree analyses were conducted using GENETREE and the species tree in Additional file 1. Black boxes indicate duplication events. The trees in A) and B) were rooted with the AGPase large and small subunit from* Chlamydomonas reinhardtii *respectively. Thicker lines indicate branches with* $K_S < 0.1$ *following duplication events (using ML estimates of synonymous branch lengths).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-232-S3.pdf]

Additional file 4

Phylogenetic trees of the large and the small subunits from angiosperms after rate-smoothing. *The trees in parts A) and B) of this figure are rate-smoothed versions of the gene trees shown in Additional file 3A and 3B that were rearranged to increase congruence with the species tree. Rate-smoothing was done by using the PL method implemented in the r8s software. Black boxes indicate duplication events.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-232-S4.pdf]

Additional file 5

Average number of synonymous substitutions per site per year. *The trees in A) and B) have the topology of the trees shown in Figure 1B and 1C respectively. The length of the branches represents the number of synonymous substitutions per site as estimated by the free model of CODEML. The bars correspond to the number of synonymous substitutions per site. The numbers of synonymous substitutions per site per year, shown in C), were estimated from the most recent dated speciation events to present sequences of the trees shown in A) and B). The error bars represent 2× standard error.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-232-S5.pdf]

Additional file 6

Amino acid sites in the large and the small subunit of AGPase from angiosperms under positive selection. *Large subunit site numbers correspond to the amino acid sequence encoded by* Shrunken-2 *(NCBI accession number: P55241). Small subunit site numbers correspond to the amino acid sequence encoded by* Brittle-2 *(NCBI accession number: AAQ14870).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-232-S6.pdf]

Additional file 7

Distribution of type I sites along the large (A) and the small (B) subunit. *The cut-off value of posterior probability is empirical and it was set to 0.5 for all group comparisons except for group 1-group 3b and group 2-group 3b where the cut-off value was set to 0.6, since theta was greater for these pairs. The Y-axis corresponds to posterior probability. The X-axis corresponds to the number of the amino acid site based on the subunits encoded by* Shrunken-2 *(A) (NCBI accession number: P55241) and* Brittle-2 *(B) (NCBI accession number: AAQ14870).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-232-S7.pdf]

Additional file 8

Type-I sites in the large and the small subunit of AGPase from angiosperms. *Type-I functional divergence between large and small subunit groups was estimated by DIVERGE. Large subunit site numbers correspond to the amino acid sequence encoded by* Shrunken-2 *(NCBI accession number: P55241). Small subunit site numbers correspond to the amino acid sequence encoded by* Brittle-2 *(NCBI accession number: AAQ14870).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-232-S8.pdf]

Additional file 9

Type-II sites in the large and the small subunit of AGPase from angiosperms. *Type-II functional divergence between large and small subunit groups was estimated by DIVERGE. Large subunit site numbers correspond to the amino acid sequence encoded by* Shrunken-2 *(NCBI accession number: P55241). Small subunit site numbers correspond to the amino acid sequence encoded by* Brittle-2 *(NCBI accession number: AAQ14870).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-232-S9.pdf]

Additional file 10

AGPase subunit accession numbers
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-232-S10.pdf]

Acknowledgements

We thank Rebecca Kimball and members of the Kimball, Braun and Hannah laboratories for many useful comments and discussions. This research was supported by National Science Foundation Grants IOB-0444031, DBI-0077676, DBI-0606607, and IOB-9982626 and USDA Grant 2006-35100-17220.

References

1. Smith-White BJ, Preiss J: **Comparison of proteins of ADP-glucose pyrophosphorylase from diverse sources.** *J Mol Evol* 1992, **34:**449-464.
2. Georgelis N, Braun EL, Shaw JR, Hannah LC: **The two AGPase subunits evolve at different rates in angiosperm, yet they are equally sensitive to activity altering amino acid changes when expressed in bacteria.** *Plant Cell* 2007, **19:**1458-1472.
3. Hannah LC, Nelson OE Jr: **Characterization of ADP-glucose pyrophosphorylase from *shrunken-2* and *brittle-2* mutants of maize.** *Biochem Genet* 1976, **14:**547-560.
4. Wolfe KH, Li WH, Sharp PM: **Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs.** *Proc Natl Acad Sci USA* 1987, **84:**9054-9058.
5. Gaut BS, Morton BR, McCaig MM, Clegg MT: **Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*.** *Proc Natl Acad Sci USA* 1996, **93:**10274-10279.
6. Gaut BS: **Molecular clocks and nucleotide substitution rates in higher plants.** In *Evolutionary Biology* Edited by: Hecht MK. New York, Plenum Press; 1998:93-120.
7. White SE, Doebley JF: **The molecular evolution of terminal ear1, a regulatory gene in the genus *Zea*.** *Genetics* 1999, **153:**1455-1462.
8. Rabinowicz PD, Braun EL, Wolfe AD, Bowen B, Grotewold E: **Maize R2R3 Myb genes: Sequence analysis reveals amplification in the higher plants.** *Genetics* 1999, **153:**427-444.
9. Eyre-Walker A: **Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy?** *Mol Biol Evol* 1996, **13:**864-872.
10. Chamary JV, Parmley JL, Hurst LD: **Hearing silence: non-neutral evolution at synonymous sites in mammals.** *Nat Rev Genet* 2006, **7:**98-108.
11. Parmley JL, Chamary JV, Hurst LD: **Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers.** *Mol Biol Evol* 2006, **23:**301-309.
12. Resch AM, Carmel L, Mariño-Ramírez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV: **Widespread positive selection in synonymous sites of mammalian genes.** *Mol Biol Evol* 2007, **24:**1821-1831.
13. Pond SK, Muse SV: **Site-to-site variation of synonymous substitution rates.** *Mol Biol Evol* 2005, **22:**2375-2385.
14. Mayrose I, Doron-Faigenboim A, Bacharach E, Pupko T: **Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates.** *Bioinformatics* 2007, **23:**319-327.
15. Seo TK, Kishino H, Thorne JL: **Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences.** *Mol Biol Evol* 2004, **21:**1201-1213.
16. Crevillen P, Ballicora MA, Merida A, Preiss J, Romero JM: **The different large subunit isoforms of Arabidopsis thaliana ADP-glucose pyrophosphorylase confer distinct kinetic and regulatory properties to the heterotetrameric enzyme.** *J Biol Chem* 2003, **278:**28508-28515.
17. Burger BT, Cross JM, Shaw JR, Caren JR, Greene TW, Okita TW, Hannah LC: **Relative turnover numbers of maize endosperm and potato tuber ADP-glucose pyrophosphorylases in the absence and presence of 3-phosphoglyceric acid.** *Planta* 2003, **217:**449-456.
18. Doan DN, Rudi H, Olsen OA: **The allosterically unregulated isoform of ADP-glucose pyrophosphorylase from barley endosperm is the most likely source of ADP-glucose incorporated into endosperm starch.** *Plant Physiol* 1999, **121:**965-975.
19. Linebarger CR, Boehlein SK, Sewell AK, Shaw J, Hannah LC: **Heat stability of maize endosperm ADP-glucose pyrophosphorylase is enhanced by insertion of a cysteine in the N terminus of the small subunit.** *Plant Physiol* 2005, **139:**1625-1634.
20. Boehlein SK, Shaw JR, Stewart JD, Hannah LC: **Heat stability and allosteric properties of the maize endosperm ADP-glucose pyrophosphorylase are intimately intertwined.** *Plant Physiol* 2008, **146:**289-299.
21. Gu X: **Statistical methods for testing functional divergence after gene duplication.** *Mol Biol Evol* 1999, **16:**1664-1674.
22. Gu X: **Maximum likelihood approach for gene family evolution under functional divergence.** *Mol Biol Evol* 2001, **18:**453-464.
23. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151-1155.
24. Ohno S: *Evolution by Gene Duplication* New York, Springer; 1970.
25. Zhang J: **Evolution by gene duplication: an update.** *Trends Ecol Evol* 2003, **18:**292-298.
26. Hurles M: **Gene duplication: the genomic trade in spare parts.** *PLoS Biol* 2004, **2:**E206.
27. Sassi SO, Braun EL, Benner SA: **The evolution of seminal ribonuclease: pseudogene reactivation or multiple gene inactivation events?** *Mol Biol Evol* 2007, **24:**1012-1024.
28. Braun EL: **Innovation from reduction: gene loss, domain loss and sequence divergence in genome evolution.** *Appl Bioinformatics* 2003, **2:**13-34.
29. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, Tanahashi T, Sakakibara K, Fujita T, Oishi K, Shin-I T, Kuroki Y, Toyoda A, Suzuki Y, Hashimoto S, Yamaguchi K, Sugano S, Kohara Y, Fujiyama A, Anterola A, Aoki S, Ashton N, Barbazuk WB, Barker E, Bennetzen JL, Blankenship R, Cho SH, Dutcher SK, Estelle M, Fawcett JA, Gundlach H, Hanada K, Heyl A, Hicks KA, Hughes J, Lohr M, Mayer K, Melkozernov A, Murata T, Nelson DR, Pils B, Prigge M, Reiss B, Renner T, Rombauts S, Rushton PJ, Sanderfoot A, Schween G, Shiu SH, Stueber K, Theodoulou FL, Tu H, Peer Y Van de, Verrier PJ, Waters E, Wood A, Yang L, Cove D, Cuming AC, Hasebe M, Lucas S, Mishler BD, Reski R, Grigoriev IV, Quatrano RS, Boore JL: **The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants.** *Science* 2008, **319:**64-69.
30. Kenrick P, Crane PR: **The origin and early evolution of plants on land.** *Nature* 1997, **389:**33-39.
31. Schneider A, Gonnet GH, Cannarozzi GM: **SynPAM–A Distance Measure Based on Synonymous Codon Substitutions.** *IEEE/ACM TCBB* 2007, **4:**553-560.
32. Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G: **Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences.** *Syst Zool* 1979, **28:**132-168.
33. La Cognata U, Willmitzer L, Müller-Röber B: **Molecular cloning and characterization of novel isoforms of potato ADP-glucose pyrophosphorylase.** *Mol Gen Genet* 1995, **246:**538-548.
34. Park SW, Chung WI: **Molecular cloning and organ-specific expression of three isoforms of tomato ADP-glucose pyrophosphorylase gene.** *Gene* 1998, **206:**215-221.
35. Akihiro T, Mizuno K, Fujimura T: **Gene expression of ADP-glucose pyrophosphorylase and starch contents in rice cultured cells are cooperatively regulated by sucrose and ABA.** *Plant Cell Physiol* 2005, **46:**937-946.
36. Ohdan T, Francisco PB Jr, Sawada T, Hirose T, Terao T, Satoh H, Nakamura Y: **Expression profiling of genes involved in starch synthesis in sink and source organs of rice.** *J Exp Bot* 2005, **56:**3229-3244.
37. Rosti S, Rudi H, Rudi K, Opsahl-Sorteberg HG, Fahy B, Denyer K: **The gene encoding the cytosolic small subunit of ADP-glucose pyrophosphorylase in barley endosperm also encodes**

the major plastidial small subunit in the leaves. *J Exp Bot* 2006, **57:**3619-3626.
38. Crevillen P, Ventriglia T, Pinto F, Orea A, Merida A, Romero JM: **Differential pattern of expression and sugar regulation of *Arabidopsis thaliana* ADP-glucose pyrophosphorylase-encoding genes.** *J Biol Chem* 2005, **280:**8143-8149.
39. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Peer Y Van de, Rokhsar D: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).** *Science* 2006, **313:**1596-1604.
40. Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW: **Widespread genome duplications throughout the history of flowering plants.** *Genome Res* 2006, **16:**738-749.
41. Felsenstein J: **Evolutionary trees from DNA sequences: A maximum likelihood approach.** *J Mol Evol* 1981, **17:**368-376.
42. Goldman N: **Statistical tests of models of DNA substitution.** *J Mol Evol* 1993, **36:**182-198.
43. Yang Z, Goldman N, Friday A: **Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem.** *Syst Biol* 1995, **44:**384-399.
44. Huelsenbeck JP, Rannala B: **Phylogenetic methods come of age: Testing hypotheses in an evolutionary context.** *Science* 1997, **276:**227-232.
45. Hileman LC, Baum DA: **Why do paralogs persist? Molecular evolution of *CYCLOIDEA* and related floral symmetry genes in Antirrhineae (Veronicaceae).** *Mol Biol Evol* 2003, **20:**591-600.
46. Wong WS, Yang Z, Goldman N, Nielsen R: **Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected site.** *Genetics* 2004, **168:**1041-1051.
47. Gu X: **A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences.** *Mol Biol Evol* 2006, **23:**1937-1945.
48. Gu X, Velden K Vander: **DIVERGE: Phylogeny-based analysis for functional-structural divergence of a protein family.** *Bioinformatics* 2002, **18:**500-501.
49. Jin X, Ballicora MA, Preiss J, Geiger JH: **Crystal structure of potato tuber ADP-glucose pyrophosphorylase.** *EMBO J* 2005, **24:**694-704.
50. Wu MX, Preiss J: **Truncated forms of the recombinant Escherichia coli ADP-glucose pyrophosphorylase: the importance of the N-terminal region for allosteric activation and inhibition.** *Arch Biochem Biophys* 2001, **389:**159-165.
51. Kavakli IH, Greene TW, Salamone PR, Choi SB, Okita TW: **Investigation of subunit function in ADP-glucose pyrophosphorylase.** *Biochem Biophys Res Commun* 2001, **281:**783-787.
52. Laughlin MJ, Chantler SE, Okita TW: **N- and C-terminal peptide sequences are essential for enzyme assembly, allosteric, and/or catalytic properties of ADP-glucose pyrophosphorylase.** *Plant J* 1998, **14:**159-168.
53. Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations.** *Genetics* 1999, **151:**1531-1545.
54. Li WH, Yang J, Gu X: **Expression divergence between duplicate genes.** *Trends Genet* 2005, **21:**602-607.
55. Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW: **Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis.** *Mol Biol Evol* 2006, **23:**469-478.
56. MacCarthy T, Bergman A: **The limits of subfunctionalization.** *BMC Evol Biol* 2007, **7:**213.
57. He X, Zhang J: **Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution.** *Genetics* 2005, **169:**1157-1164.
58. Akashi H: **Gene expression and molecular evolution.** *Curr Opin Genet Dev* 2001, **11:**660-666.
59. Duret L, Mouchiroud D: **Determinants of substitution rates in mammalian genes, expression pattern affects selection intensity but not mutation rate.** *Mol Biol Evol* 2000, **17:**68-74.
60. Wright SI, Yau CB, Looseley M, Meyers BC: **Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*.** *Mol Biol Evol* 2004, **21:**1719-1726.
61. Ballicora MA, Dubay JR, Devillers CH, Preiss J: **Resurrecting the ancestral enzymatic role of a modulatory subunit.** *J Biol Chem* 2005, **280:**10189-10195.
62. Haugen TH, Preiss J: **Biosynthesis of bacterial glycogen. The nature of the binding of substrates and effectors to ADP-glucose synthase.** *J Biol Chem* 1979, **254:**127-136.
63. Hwang SK, Hamada S, Okita TW: **ATP binding site in the plant ADP-glucose pyrophosphorylase large subunit.** *FEBS Lett* 2006, **580:**6741-6748.
64. Ballicora MA, Fu Y, Nesbitt NM, Preiss J: **ADP-Glucose pyrophosphorylase from potato tubers. Site-directed mutagenesis studies of the regulatory sites.** *Plant Physiol* 1998, **118:**265-274.
65. Kavakli IH, Park JS, Slattery CJ, Salamone PR, Frohlick J, Okita TW: **Analysis of allosteric effector binding sites of potato ADP-glucose pyrophosphorylase through reverse genetics.** *J Biol Chem* 2001, **276:**40834-40840.
66. Giroux MJ, Shaw J, Barry G, Cobb BG, Greene TW, Okita TW, Hannah LC: **A single mutation that increases maize seed weight.** *Proc Natl Acad Sci USA* 1996, **93:**5824-5829.
67. Iglesias A, Barry GF, Meyer C, Bloksberg L, Nakata P, Greene T, Laughlin MJ, Okita TW, Kishore GM, Preiss J: **Expression of the potato tuber ADP-glucose pyrophosphorylase in *Escherichia coli*.** *J Biol Chem* 1993, **268:**1081-1086.
68. Kumar S, Tamura K, Nei M: **MEGA 3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment.** *Brief Bioinf* 2004, **5:**150-163.
69. Zwickl DJ: **Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.** In *Ph.D. dissertation* The University of Texas at Austin; 2006.
70. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22:**2688-2690.
71. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *CABIOS* 1992, **8:**275-282.
72. Felsenstein J: **Confidence-limits on phylogenies: An approach using the bootstrap.** *Evolution Int J Org Evolution* 1985, **39:**783-791.
73. Dayhoff MO, Schwartz RM, Orcutt BC: **A model of evolutionary change in proteins.** In *Atlas of Protein Sequence and Structure Volume 5*. Edited by: Dayhoff MO. Silver Springs, MD, National Biomedical Research Foundation; 1978:345-352.
74. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *CABIOS* 1997, **13:**555-556.
75. Page RDM, Cotton JA: **GeneTree: a tool for exploring gene family evolution.** In *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families* Edited by: Sankoff D, Nadeau J. Dordrecht, Kluwer Academic Press; 2000:525-536.
76. Page RDM: **GeneTree: comparing gene and species phylogenies using reconciled trees.** *Bioinformatics* 1998, **14:**819-820.
77. Chen K, Durand D, Farach-Colton M: **NOTUNG: A program for dating gene duplications and optimizing gene family trees.** *Journal of Computational Biology* 2000, **7:**429-447.
78. Hillis DM, Bull JJ: **An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis.** *Syst Biol* 1993, **42:**182-192.
79. Sanderson MJ: **r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock.** *Bioinformatics* 2003, **19:**301-302.

80. Yang Z: **Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A.** *J Mol Evol* 2000, **51:**423-432.
81. Yang Z, Nielsen R: **Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.** *Mol Biol Evol* 2002, **19:**908-917.
82. Yang Z, Wong WSW, Nielsen R: **Bayes empirical Bayes inference of amino acid sites under positive selection.** *Mol Biol Evol* 2005, **22:**1107-1118.
83. de Sa M, Drouin G: **Phylogeny and substitution rates of angiosperm actin genes.** *Mol Biol Evol* 1996, **13:**1198-1212.
84. Grant D, Cregan P, Shoemaker RC: **Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis.** *Proc Natl Acad Sci USA* 2000, **97:**4168-4173.
85. Ku HM, Vision T, Liu J, Tanksley SD: **Comparing sequenced segments of the tomato and Arabidopsis genomes: Large-scale duplication followed by selective gene loss creates a network of synteny.** *Proc Natl Acad Sci USA* 2000, **97:**9121-9126.
86. Wikstrom N, Savolainen V, Chase MW: **Evolution of the angiosperms: calibrating the family tree.** *Proc Biol Sci* 2001, **268:**2211-2220.
87. Gaut BS: **Evolutionary dynamics of grass genomes.** *New phytologist* 2002, **154:**15-28.
88. Choi HK, Mun JH, Kim DJ, Zhu H, Baek JM, Mudge J, Roe B, Ellis N, Doyle J, Kiss GB, Young ND, Cook DR: **Estimating genome conservation between crop and model legume species.** *Proc Natl Acad Sci USA* 2004, **101:**15289-15294.
89. Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC: **Mining EST databases to resolve evolutionary events in major crop species.** *Genome* 2004, **47:**868-876.
90. Soltis P, Soltis D, Edwards C: **Angiosperms, Flowering Plants.** [http://tolweb.org/Angiosperms/20646/2005.06.03]. in The Tree of Life Web Project, http://tolweb.org/ Version 03, 2005
91. Bar-Or C, Bar-Eyal M, Gal TZ, Kapulnik Y, Czosnek H, Koltai H: **Derivation of species-specific hybridization-like knowledge out of cross-species hybridization results.** *BMC Genomics* 2006, **7:**110.
92. Yang TJ, Kim JS, Kwon SJ, Lim KB, Choi BS, Kim JA, Jin M, Park JY, Lim MH, Kim HI, Lim YP, Kang JJ, Hong JH, Kim CB, Bhak J, Bancroft I, Park BS: **Sequence-Level Analysis of the Diploidization Process in the Triplicated FLOWERING LOCUS C Region of *Brassica rapa*.** *Plant Cell* 2006, **18:**1339-1347.
93. Crepet WL, Nixon KC, Gandolfo MA: **Fossil evidence and phylogeny: the age of major angiosperm clades based on mesofossil and macrofossil evidence from Cretaceous deposits.** *Am J Botany* 2004, **91:**1666-1682.