EVOLUTIONARY BIOLOGY

Multimerization variants as potential drivers of neofunctionalization

Youngwoo Lee[1,2] and Daniel B. Szymanski[1,2,3]*

Whole-genome duplications are common during evolution, creating genetic redundancy that can enable cellular innovations. Novel protein-protein interactions provide a route to diversified gene functions, but, at present, there is limited proteome-scale knowledge on the extent to which variability in protein complex formation drives neofunctionalization. Here, we used protein correlation profiling to test for variability in apparent mass among thousands of orthologous proteins isolated from diverse species and cell types. Variants in protein complex size were unexpectedly common, in some cases appearing after relatively recent whole-genome duplications or an allopolyploidy event. In other instances, variants such as those in the carbonic anhydrase orthologous group reflected the neofunctionalization of ancient paralogs that have been preserved in extant species. Our results demonstrate that homo- and heteromer formation have the potential to drive neofunctionalization in diverse classes of enzymes, signaling, and structural proteins.

INTRODUCTION

Neofunctionalization of duplicated genes can broaden the activity and connectivity of almost any cellular function and, as a result, is central to developmental innovation and models of evolution (*1*–*4*). New functions can be generated by altered patterns of gene expression that place a protein in a different cellular context (*5*, *6*). Conserved "housekeeping" genes can evolve "moonlighting" functions and interact with pathways that were previously independent of its original biochemical activity (*7*). Enzymes can neofunctionalize to alter their substrate specificity and generate new biochemical pathways (*8*). Chromosome rearrangements and the loss or gain of functional protein domains can markedly alter protein function (*9*). Protein complex formation is a cornerstone of biological systems (*10*), and when mutations generate new stable protein-protein interactions, the biochemical and/or biophysical properties of the protein are unavoidably changed. Self-interaction can generate homomers with increased catalytic efficiency or altered substrate specificity (*11*, *12*).

Some evolutionarily conserved heteromeric protein complexes have catalytic activities (*13*), force transduction mechanisms (*14*), reversible subcellular localizations (*15*), and complicated signal transduction functions (*16*), which, in many instances, could not be achieved by a single polypeptide. An important challenge in biology is to better understand the extent to which altered multimerization, used here as a general term to include homo- and heteromeric complexes, could drive neofunctionalization. Mass spectrometry (MS)–based protein correlation profiling (PCP) is well suited to this task. It is an established and valid approach to broadly analyze protein multimerization and their dynamics by quantifying the elution profiles of thousands of endogenous protein complexes as a function of their chromatographic separation (*17*–*24*). This PCP of native protein complexes in combination with machine learning–based predictions emphasizes the compositional identity of orthologous complexes across kingdoms (*18*, *25*). Here, we take an open-ended evolution proteomics approach to analyze variability in protein complex size among orthologous proteins and test for its potential evolutionary relevance.

RESULTS AND DISCUSSION

Soluble extracts from two Malvales (leaves of *Arabidopsis thaliana* and cotton fibers of *Gossypium hirsutum*), one of Fabales (leaves of *Glycine max*), and one monocot species of the Poales (the seed aleurone layers of developing *Oryza sativa*) were separated using size exclusion chromatography (SEC), and the individual column fractions were then processed for quantitative liquid chromatography (LC)/MS analysis (Fig. 1A). Biological replicates and automated peak detection algorithms (*26*) were used to make thousands of apparent mass (M_{app}) determinations that were derived from reproducible elution profiles (fig. S1, A to C). The peptide (table S1, A to D), protein (table S1, E to H), and normalized abundance profiles (table S2A) are provided. The measured M_{app} values accurately reflect protein complex formation and are not related to protein abundance, artifactual aggregation, or extensive glycosylation (*20*, *22*, *26*). Stable complex formation, defined as a protein having an M_{app} that was at least twice that of its monomeric mass (M_{mono}), was common with more than one-third of all proteins falling in this class (fig. S1, C and D). This degree of multimerization is similar to what has been reported previously for soluble proteins in other studies (*20*, *22*, *27*). This PCP method detects stable protein complexes that remain assembled during cell fractionation and column chromatography. Therefore, any variability among orthologs in unstable complexes, or those that compose a small fraction of the total cellular pool of a given protein that falls below our detection thresholds, will not be detected. In this study, 93% of the predicted multimeric proteins had a single peak on the SEC column, reflecting a cellular protein pool that is dominated by one species (table S3). Those with multiple peaks might have assembly states and protein-protein or protein-ligand interactions that are more sensitive to protein concentrations.

In this LC/MS workflow, the elution profiles of protein particles are dominated by protein-protein interactions, as this method is being successfully used to broadly predict the composition of protein complexes (*18*, *21*, *24*, *25*) and how they change in response to mutation

[1]Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907, USA. [2]Center for Plant Biology, Purdue University, West Lafayette, IN 47907, USA. [3]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA.
*Corresponding author. Email: dszyman@purdue.edu

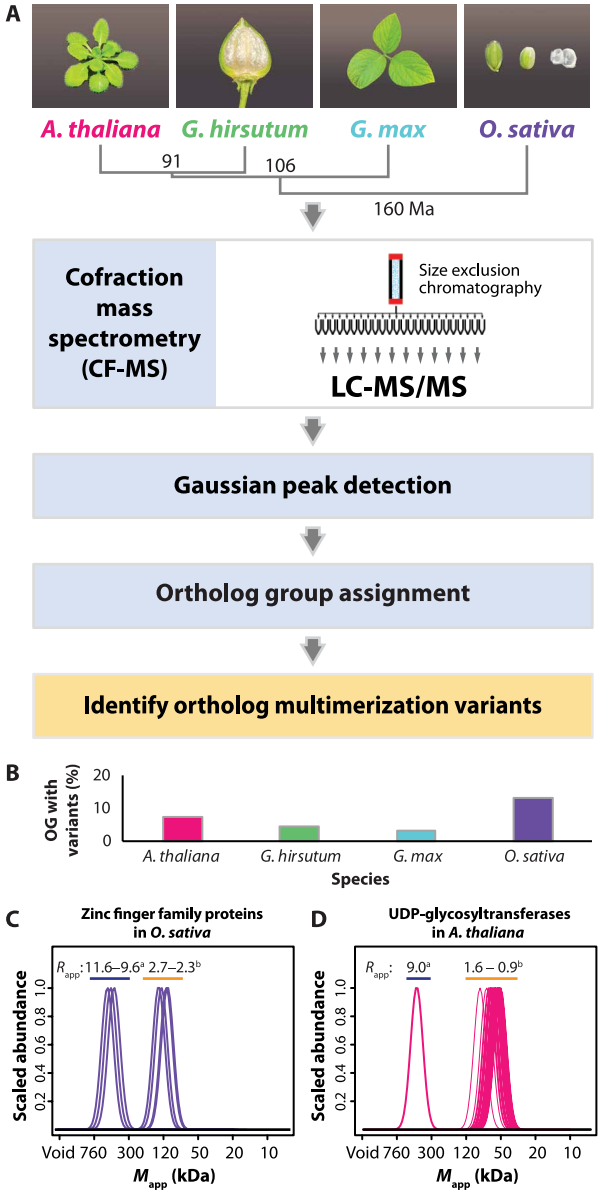

Fig. 1. An evolution proteomics pipeline to analyze multimerization variability among thousands of orthologous proteins. (**A**) Gel-free proteomics workflow in which soluble extracts are isolated from diverse plant species. Cell extracts are separated under nondenaturing conditions on a sizing column, and the column fractions are analyzed by LC-MS/MS for protein quantification. Elution profiles are fitted to Gaussian peaks to assess reproducibility and assign M_{app} values for thousands of proteins. ANOVA and pairwise *t* tests are used to identify proteins within OGs that differ in multimerization state. (**B**) Summary of the percent of the OGs for each species that contained a size variant. (**C**) The elution profiles of an *O. sativa* OG of zinc finger proteins that contained multiple size variants. (**D**) The elution profiles of an *A. thaliana* OG of UDP-glycosyltransferases that contained a detected size variant. The differently colored horizontal bars, R_{app} distributions, and statistical groupings in (C) and (D) indicate groups of mOVs with distinct multimerization states. Photo credit: Youngwoo Lee, Purdue University.

(*22*) or growth factor stimulation (*17*). The mass differences that we detected are large (>40% of the mass of the particle) and cannot be explained by most posttranslational modifications that have relatively small effects on the total mass. However, an altered mass could also arise if one ortholog acquired extensive glycosylation or RNA binding activity.

To test for different multimerization states within and among species, all proteins with reproducible elution profiles and measured M_{app} values were assigned to specific ortholog groups (OGs) using the InParanoid algorithm (*28*) implemented at Phytozome v12 (https://phytozome.jgi.doe.gov/). For each OG, analysis of variance (ANOVA) tests at a false discovery rate (FDR) of 2% were conducted using the parameters M_{app} and R_{app} ($R_{app} = M_{app}/M_{mono}$) to identify predicted multimerization orthovariants (mOVs) that had size differences that could not be explained by variability in monomer masses. A subsequent round of Tukey's pairwise post hoc tests was used with the flagged OGs to determine which specific orthologs differed. In the four species analyzed, between 3.4 and 13.7% of the OGs contained one or more proteins with a different apparent mass, indicating that mOVs occur across species and are not rare (Fig. 1B and table S3, A to D). These percentages may not be accurate estimates of mOV frequencies, because only a small subset of the cell types of each organism was analyzed and the percentages may differ for low abundance signaling proteins that are not detected in this study.

As an example, an *O. sativa* OG encodes a zinc finger family of proteins with no known function. It existed as two distinct complex types, with three of the isoforms having an elevated M_{app} of ~500 kDa (Fig. 1C). *Arabidopsis* mOVs are being generated and maintained across broad evolutionary windows based on paralog mapping to the estimated timing of the five whole-genome duplications (WGDs) that spanned more than 500 million years (Ma) (*4*, *29*). Of the 227 known paralog pairs that were analyzed on the SEC column, there were nine instances in which a paralog clearly differed in size. Six paralog variants were generated in the two most recent WGDs that occurred after speciation. For example, a uridine 5′-diphosphate (UDP)–glycosyltransferase (OG 93524696) variant that is predicted to be involved in cell wall biosynthesis was generated from a duplication event ~25 Ma ago (Fig. 1D). The mOV was present in a 450 kDa complex, but its paralogs had an elution peak centered on ~50 kDa, which is near the calculated monomeric mass. The other three instances of paralog mOVs were mapped to ancient WGDs estimated at 320 and 400 Ma ago. Given the relatively short life span of duplicated genes (*2*, *30*), retention of these biochemical variants likely reflects a functional importance. The full supplemental datasets for OG members and the mOV identifications are provided (tables S3, A to D, and S4, A and B).

We also discovered evidence for ortholog divergence between homoeologs expressed in purified cotton fibers (Fig. 2). Cultivated cotton is an allotetraploid generated by the hybridization of the two diploid progenitor species, which contributed the A and D genomes (*31*). Homoeologs can be easily distinguished based on unique peptides (Fig. 2B). The 1861 fiber-expressed proteins that had reproducible elution profiles were distributed quite evenly between A_T- and D_T-encoded proteins (where "T" indicates tetraploid). However, in ~16% of the cases, we detected homoeolog-specific expression (Fig. 2C and table S2B), and this likely reflects the known homoeolog selectivity at the transcript level (*32*). Among the 439 homoeolog groups (HGs) in which both alleles were detected, 15 contained a size variant in which both the A_T and D_T homoeologs coeluted at the same peak location, but that apparent mass differed from other homoeologs. For example, the HG 93524686 contained 20 Rab superfamily small guanosine triphosphatases (Fig. 2D and fig. S2). The variant

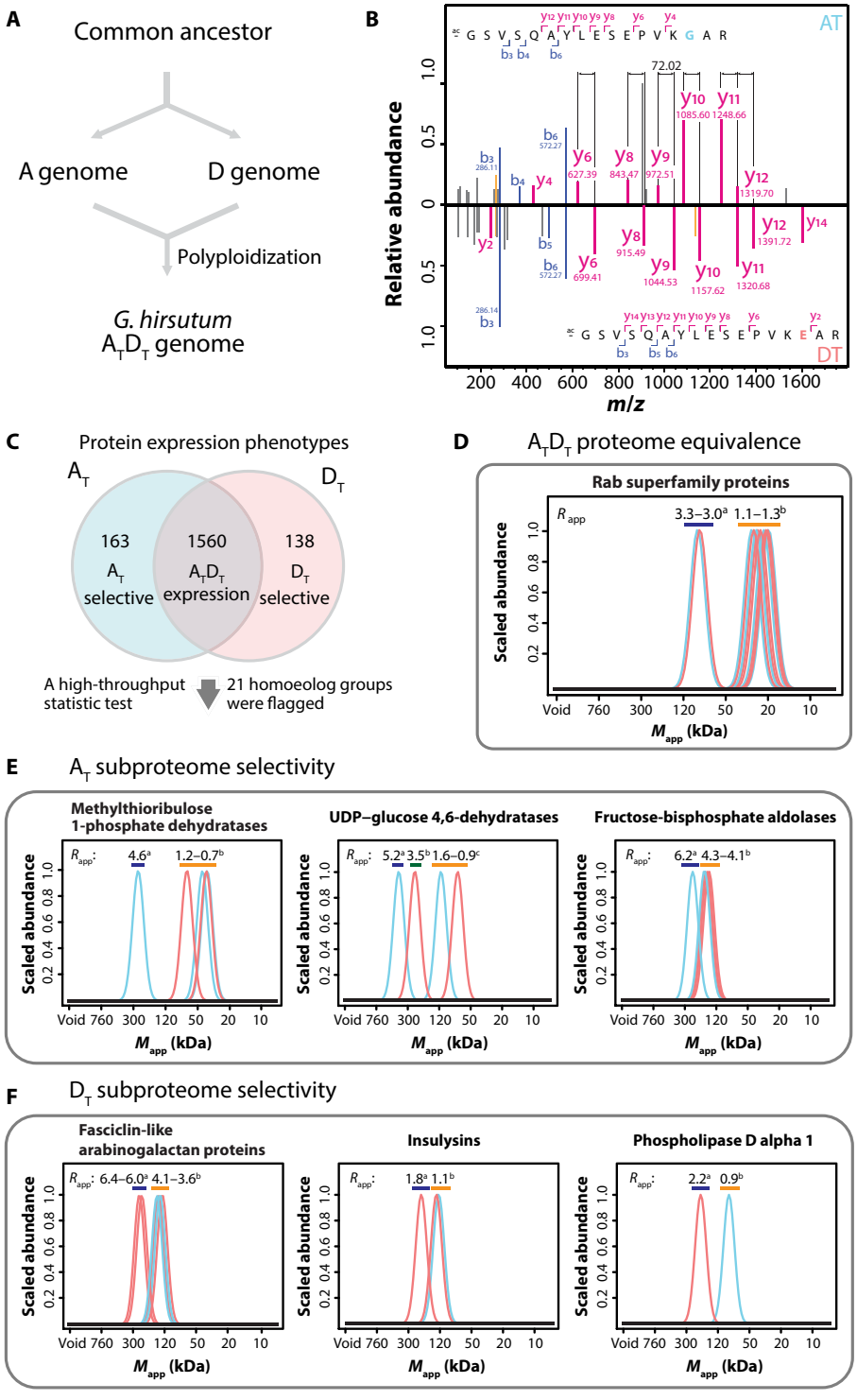

Fig. 2. Homoeolog-specific expression and multimerization in the allotetraploid fiber cells of *G. hirsutum*. (**A**) The evolutionary history of *G. hirsutum* polyploidization (*31*). (**B**) A mirror plot shows MS/MS spectrum of unique peptides derived from A_T (sky blue) and D_T (coral) methylthioribulose 1-phosphate dehydratase homoeologs on the top and bottom panels, respectively. The b ions and y ions are labeled. The lines between informative y ions of the two alleles reflect a mass difference shift of 72.02 Da, which is the mass difference of glutamic acid (E) and glycine (G) residues. *m/z*, mass/charge ratio. (**C**) A_T and D_T homoeolog–specific expression. Numbers on the Venn diagram show differentially expressed proteins originated from A_T and D_T genomes. (**D** to **F**) Classification of homoeolog-specific multimerization into two classes. A_TD_T proteome equivalence class includes homoeologs from both genomes assembly into indistinguishable protein complexes (D). Subproteome selectivity classes contain either A_T (E) or D_T (F) homoeologs that had a unique multimeric state. The differently colored horizontal bars, R_{app} distributions, and statistical groupings in (D) to (F) indicate groups of mOVs with distinct multimerization states.

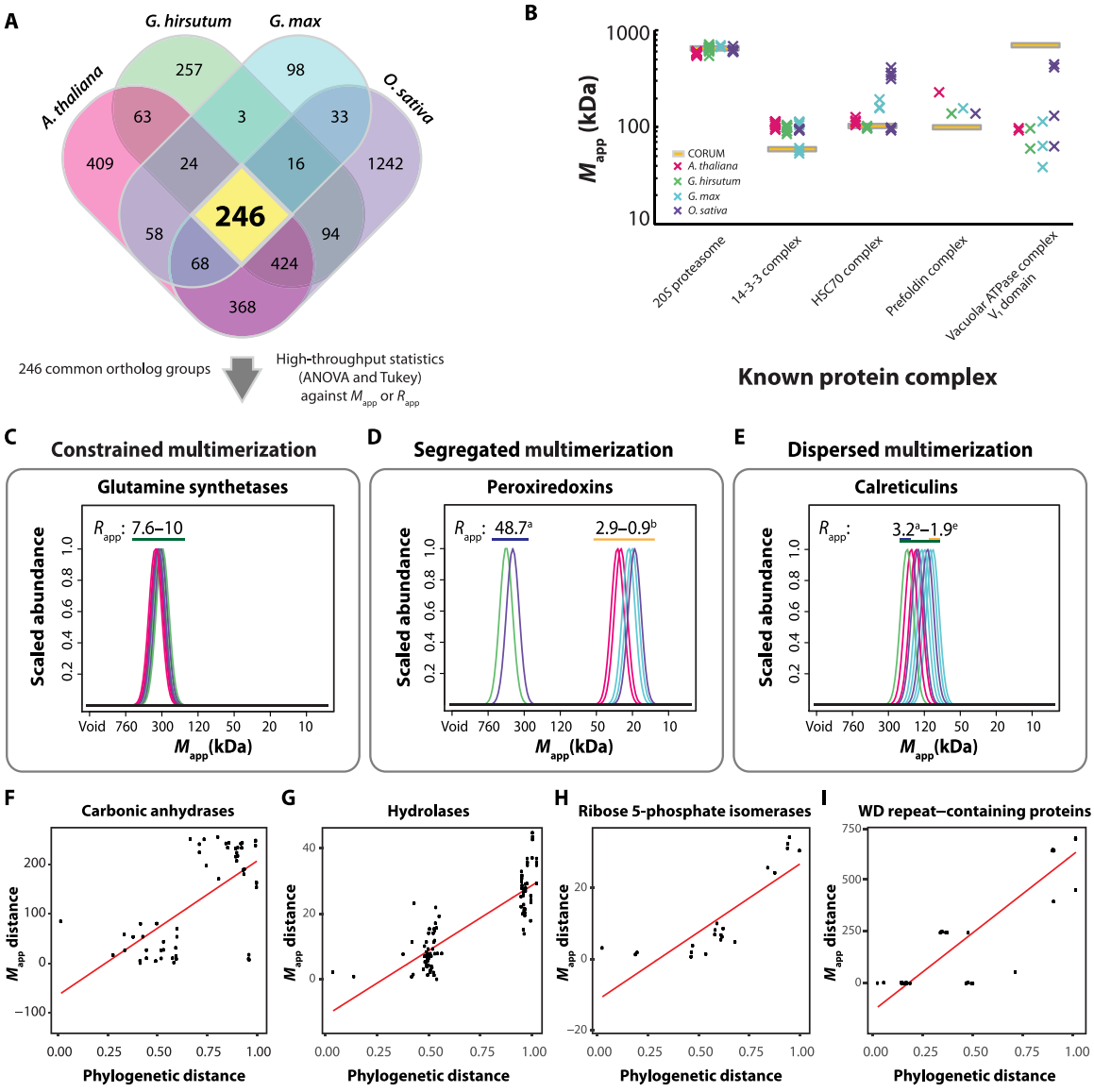

Fig. 3. mOVs are common, and, in some cases, complex size differences correlate with protein phylogeny. (**A**) The 246 OGs reproducibly profiled across all four species were tested for multimerization variability. (**B**) Variability in the assembly status of predicted known complexes. Orthologs were mapped onto CORUM database of known mammalian protein complexes (*65*). The masses of the conserved protein complexes were marked with yellow bars, and the M_{app}s (crosses) of the individual subunit orthologs from the four species were overlaid in columns. (**C** to **E**) The OG multimerization patterns were classified as constrained (C), segregated (D), or dispersed (E) based on the elution profiles. The differently colored horizontal bars, R_{app} distributions, and statistical groupings in (C) to (E) indicate mOVs with distinct multimerization states. (**F** to **I**) Scatterplots and regression lines of the pairwise differences in apparent mass and the phylogenic distance of the corresponding protein pair obtained from a multiple sequence alignment of the detected proteins in the OG.

homoeologs with an apparent mass of 80 kDa contained Rab5a subfamily members that likely reflected an ancient divergence of a gene family that controls the specificity of vesicle trafficking and has undergone extensive duplication and neofunctionalization (*33*). There were six instances in which homoeolog-specific protein complexes were detected (Fig. 2, E and F). Even though the progenitor species that contributed the A genome had a long fiber phenotype, highly assembled protein complexes with predicted cell wall functions were partitioned between the A_T and D_T genomes. The A_T-encoded UDP–glucose 4,6-dehydratases (HG 93523466) are involved in rhamnose synthesis, which becomes incorporated into rhamnogalacturonan I and II (*34*). The D_T-encoded fatty acid synthase–type arabinogalactan glycoproteins are members of an important cell wall signaling molecules (*35*).

To test for the existence of evolutionarily conserved mOVs, orthologs with reproducible M_{app} values in each of the four species were analyzed in detail (Fig. 3). Differing tissues and species were selected at the onset to maximize our chances of detecting differences among orthologs, but these choices confounded our ability to cleanly separate the contributions of cell types and speciation. However, combinations of phylogenic analysis using detected proteins and others using the full set of orthologs extracted from the proteome of each species were used to test for robust correlations between multimerization state and protein evolution (see below).

Sixty percent of the 246 proteins in this merged dataset were enzymes present in 80 different pathways (table S4A). There were a few known complexes, such as the 20*S* proteasome core particle and 14-3-3 complexes in which orthologs across all species had indistinguishable apparent masses and may reflect instances in which the compositions of orthologous complexes are broadly conserved. Other subunits of known complexes like the chaperones HSC70 and prefoldin, as well as vacuolar adenosine triphosphatase (V-ATPase) V1 domain subunits, had distinct multimerization states in one or more species (Fig. 3B). The frequent detection of subcomplexes of evolutionarily conserved complexes like the V-ATPase V1 domain in cell extracts is consistent with prior analyses from our group, which indicate that very few known complex subunits exist predominantly in a particle with a size that matches that predicted by the fully assembled state [e.g., (*20*) Fig. 4G; (*26*) fig. S3]. It is also likely that some chaperone orthologs have distinct protein-protein interactions with client or regulatory proteins.

The 246 OGs were subjected to ANOVA, pairwise *t* tests, and manual validation of the raw profile data as described in Materials and Methods. Eighty-four OGs contained one or more mOVs (Fig. 3, D and E, and table S4B). Two-thirds of the OGs contained proteins that were either monomeric or assembled into equivalently sized protein complexes. For example, glutamine synthetase (GS) orthologs eluted at mass values that are consistent with it forming a homooctamer (Fig. 3C). GS homooctamer formation creates the enzyme active sites and enables complicated allosteric control (*36*, *37*). In the cross-species comparisons, there were 69 OGs termed "segregated" that assembled into completely resolvable classes of protein complexes (Fig. 3D). Many of the mOVs in the segregated class likely reflected variability in self-interaction, as 51 of the 69 OGs were orthologous to members from other species that have been shown previously to form homomers (table S4B). The peroxiredoxins are antioxidants that interconvert between monomeric and decameric forms to enable redox-dependent chaperone activity (*38*). The high multimerization state of a subset of peroxiredoxins that we detected under reducing conditions was much larger than a predicted decamer, perhaps reflecting a neofunctionalized form that interacts uniquely with an unknown binding partner. In the second "dispersed" class of OGs, elution peaks were spread more evenly over a mass range, but the M_{app}s of the extreme groups were significantly different in the pairwise tests (Fig. 3E). For instance, calreticulin, which is a cytosol and endoplasmic reticulum lumen–localized chaperone, had peaks in the mass range from 100 to 165 kDa and may form stable complexes with distinct interaction partners, a subset of which was resolved on the sizing column.

Multiple sequence alignments from the detected proteins in each OG were generated to test for phylogenetic signals in the SEC profile data. If amino acid sequence variability in the OG was dominated by those that define the altered multimerization states, then apparent mass differences could correlate with protein phylogenies generated using protein sequence alignments of the detected proteins. For each variant-containing OG, vectors of the pairwise M_{app} and patristic phylogenetic distances were plotted and filtered on the basis of a combination of slope, R^2, and *P* values from the *t* tests to identify candidates. There were 10 segregated OGs that had a positive interaction between apparent mass and protein sequence similarity (Fig. 3, F to I, and table S4C). The instances of strong clustering of points reflected the discrete categories of M_{app} values in the segregated class. These trends were not driven by the evolutionary distances among the species. When similar control analysis was conducted on the basis of plots of M_{app} against evolutionary distances of the four species, only one rice WD-40 repeat protein was identified, and it was flagged previously in the M_{app} versus protein pair distance plots (table S4C). These correlation plots reflected true evolutionary signal because when expanded phylogenetic trees were constructed using both detected and undetected proteins in the OGs, the phylogenic topologies were segregated and reflected the measured M_{app} differences (fig. S3, A to F). The results suggest that some of the multimerization variants arose from ancient paralogs more than 100 Ma ago before the divergence of monocots and dicots. In four instances, species-specific differences drove the interaction between M_{app} and phylogenetic distance (fig. S3, G to J).

β carbonic anhydrase (βCA) provided an opportunity to more deeply analyze the relationships between multimerization and neofunctionalization. βCA is a diffusion-limited enzyme that interconverts CO_2 and carbonate. CO_2-fixing organisms use CA as a mechanism to concentrate CO_2 and reduce biochemical futility of photorespiration caused by poor selectivity of ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO) for CO_2 over O_2. βCA has the activity, while other CAs have neofunctionalized to scaffolding proteins with no enzymatic activity (*39*). The active site for βCA involves amino acids in both polypeptides at a dimerization interface (*40*, *41*). In our samples, βCA OG fell into the segregated class, with group members divided between a low mass form that likely reflects the dimer (termed dimer hereafter) and a high mass form (Fig. 4A). The higher mass peak likely corresponded to a homooctamer based on the peak locations and solved crystal structure of octameric βCA from *Pisum sativum* (*40*). The octamer is a tetramer of dimers, and octameric forms contain a motif near the N terminus that distinguishes this clade (fig. S4, A to C). In the solved structure, the octamer has two charged amino acids in the dimerization domain and a phenylalanine residue that enables hydrophobic interactions between C termini and tetramer/octamer assembly. They are unique and highly conserved across all orthologs in the putative high mass class of βCAs (Fig. 4B and fig. S4, A and B). The dicot species expressed predicted dimeric and octameric forms of βCA. OsβCA5 was in the dimeric group, but OsβCA1 was an outlier that is part of a monocot-specific clade of C-terminally truncated βCAs (*42*). If βCA phylogenies are conducted using truncated CAs that mirror the C-terminally truncated monocot alleles, then OsβCA1 was more closely related to the octameric βCAs (fig. S5), suggesting that the C-terminal mutations reverted these orthologs back to a predicted dimeric form. This could reflect altered selection on CAs due to anatomical innovations and C4 photosynthesis pathways that evolved in the monocots and altered CO_2 gradients. βCAs from additional landmark species including the moss (*Physcomitrella patens*), fern (*Selaginella moellendorffii*), ginkgo (*Ginkgo biloba*), and the basal angiosperm *Amborella* (*Amborella trichopoda*) allowed the origin of the ancient βCA paralog to be estimated more accurately (Fig. 4C and fig. S4C). All of the moss and fern orthologs fell into the phylogenetic group that encoded dimeric βCAs. However, the ginkgo and *Amborella* orthologs included both predicted dimeric and octameric forms (Fig. 4C). H186 at the dimerization interface (Fig. 4B) was the most highly conserved octamer-defining residue. Predicted octamers from fern, gingko, and nearly all of them from the other species retained H186 but had less consistent conservation of the acidic or hydrophobic residues that were highlighted at the subunit

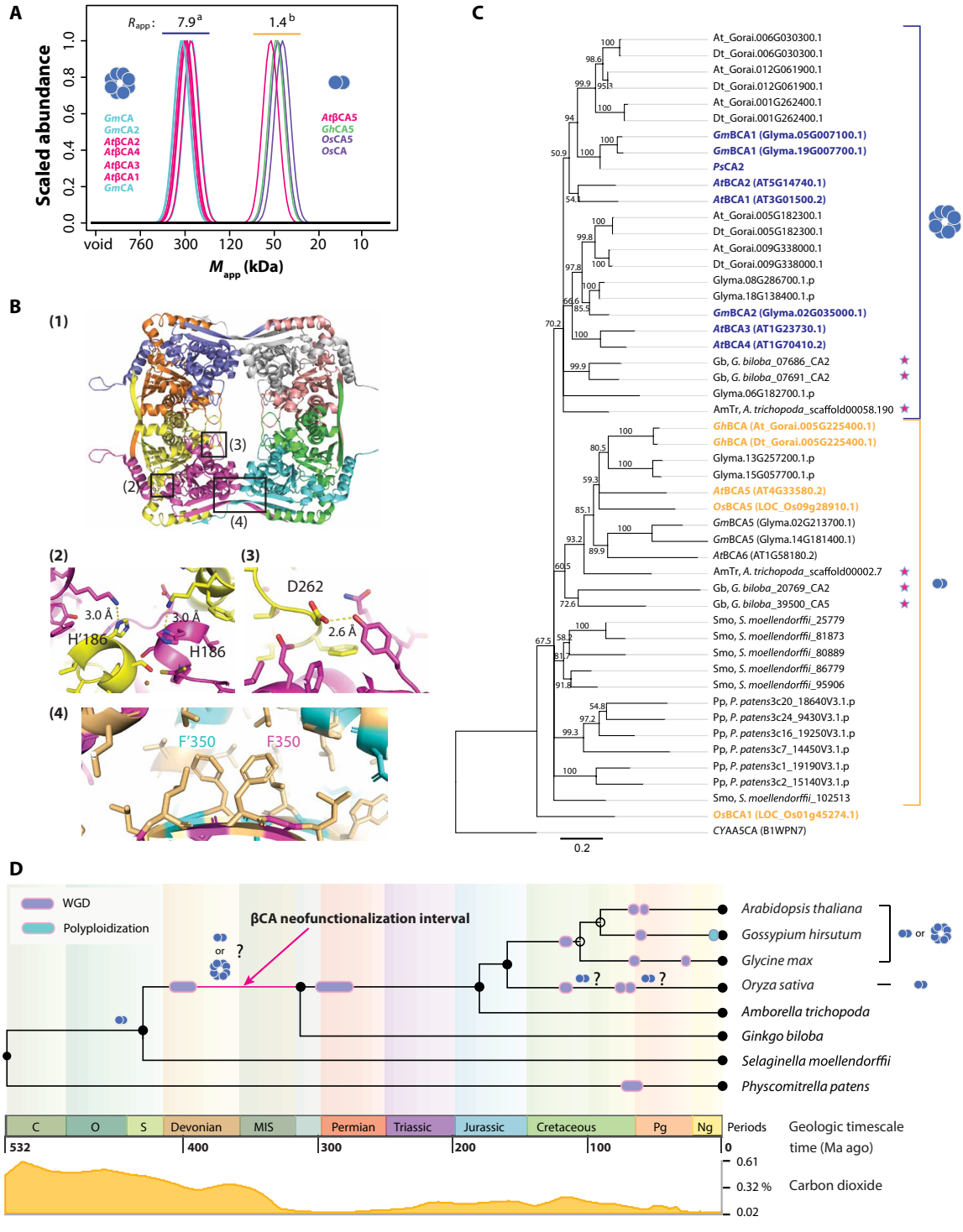

Fig. 4. Ancient paralogs of βCA neofunctionalized to increase the efficiency of CO_2 fixation. (**A**) Elution profiles of βCA orthologs were segregated into two distinct multimeric states. The differently colored horizontal bars indicate groups of βCA orthologs with distinct multimerizations based on R_{app} and statistical tests. (**B**) Conserved amino acids that mediate octamer formation. (1) βCA subunits in a homooctameric structure of *Ps*CA2 (Protein Data Bank: 1EKJ) are rendered differently. Three boxed regions depict each amino acid replacement that may affect protein-protein interaction. (2 and 3) Hydrogen bond formation by histidine (2) or aspartic acid (3) may promote βCA octamer stabilization. (4) Hydrophobic amino acids are rendered with gold color. C-terminal phenylalanine mediates dimer-dimer interactions. (**C**) Expanded phylogenetic analysis of all βCAs using the available sequences from several landmark species. Detected dimeric βCAs are colored orange, and the dimeric clade is bracketed; while detected βCAs that were octameric are rendered with dark blue, and the clade is bracketed. *Cyanothece* sp. American Type Culture Collection (ATCC) 51142 CA (*CY*AA5CA) is an outgroup. The numbers on the tree show the bootstrap statistics for clades. The scale bar indicates patristic distances. Gb, *G. biloba*; Pp, *P. patens*; Smo, *S. moellendorffii*; AmTr, *A. trichopoda*. (**D**) A timeline of WGDs (*4*) overlaid on the phylogeny of the plant kingdom was built using the Timescale of Life (*59*).

interfaces (Fig. 4B and fig. S4C). This pattern of sequence conservation at H186 may reflect an essential gatekeeping function for conversion to the octameric state.

These extended protein phylogenies allowed the estimated origin of the land plant octameric orthologs to be placed at a WGD that occurred before the speciation of ginkgo (Fig. 4D). The timing of this multimerization innovation coincides with a period of decreasing atmospheric CO_2 concentrations at around 360 to 400 Ma ago that was due to the expansion of terrestrial plant communities (*43*). Under CO_2-limiting conditions, locally concentrating CA can provide an advantage for CO_2-fixing organisms. Prokaryotic species take this to an extreme generating encapsulated organelles, containing CA and the CO_2-fixing enzyme RuBisCO (*44*). In plants, CO_2 fixation is compartmentalized in chloroplasts, and any biochemical or morphological innovation that increases CO_2 flux to chloroplasts can increase photosynthetic rates and plant productivity (*45*). Octamer formation could increase local CA concentrations in the cell and increase CO_2 flux. Mutation of the octameric βCA1 causes hypersensitivity to reduced CO_2 concentrations (*46*). βCA4 is octameric and plasma membrane localized (*47*), and the sterol-rich plasma membrane in plants may present a diffusion barrier for CO_2 (*48*). In this context, octamer formation and subcellular localization could more efficiently exploit a local CO_2 gradient. The increased mass of the octamer would also reduce the rate of CA diffusion away from the membrane and enable more efficient clustering at the cell periphery. All of the above experimental data are consistent with the hypothesis that falling CO_2 levels in the Devonian selected for octamer formation and neofunctionalized CAs.

Duplicated genes can be subjected to strong positive selection (*2*). Interacting networks of protein complexes that assemble in a protein concentration–dependent manner likely underlie positive selection to maintain duplicated genes and preserve existing functionality (*49*). Diversification of physical and genetic interactions among the duplicates is a distinct path to gene retention and functional novelty (*50*). The PCP approach here provides broad access to the outcomes of numerous natural selection, evolution, and multimerization experiments that have played out in the land plants. Hundreds of orthologous enzymes, structural, and signaling proteins display divergent protein-protein or protein-ligand interaction profiles that have been generated and maintained over hundreds of millions of years and constitute a new group of candidate genes that may drive developmental adaptation and increased fitness. Self-interaction is a common mode of structural and functional diversification, and the evolutionary trajectories of mOVs in the segregated class appear to be defined by this process. Homomer formation can enable metabolic pathway integration via novel allosteric control and create binding surfaces and diffusive properties that can drive novel subcellular localizations and stable binding interactions. This evolution proteomics approach provides a knowledge framework to discover how systems of protein complexes neofunctionalize and respond to changing selective pressures and developmental processes.

MATERIALS AND METHODS

Plant materials and isolation of soluble extracts

Arabidopsis thaliana

Two grams of leaf tissue from *A. thaliana* (Columbia-0) at 21 days after germination was used for cell fractionation as described previously (*20*, *22*, *26*, *27*). Briefly, fresh tissues were ground using a Polytron homogenizer (Brinkmann Instruments) under ice-cold microsome isolation buffer solution [50 mM Hepes/KOH (pH 7.5), 250 mM sorbitol, 50 mM KOAc, 2 mM $Mg(OAc)_2$, 1 mM EDTA, 1 mM EGTA, 1 mM dithiothreitol (DTT), 2 mM phenylmethylsulfonyl fluoride, and 1% (v/v) protease inhibitor cocktail (160 mg/ml benzamidine-HCl, 100 mg/ml leupeptin, 12 mg/ml phenanthroline, 0.1 mg/ml aprotinin, and 0.1 mg/ml pepstatin A)]. The homogenate was filtered with four layers of cheesecloth to remove debris and enriched by centrifugation at 1000*g* on a Beckman Avanti 30 (Beckman Coulter Life Sciences) for 10 min at 4°C. The supernatant was obtained by ultracentrifugation at 200,000*g* for 20 min at 4°C using a Beckman Optima Ultracentrifuge (Beckman Coulter Life Sciences). The volume of the soluble fraction was concentrated to 1 ml using an Amicon Ultra-4 10K Centrifugal Filter (Millipore). Seppro Rubisco spin columns (Sigma-Aldrich) were used for the removal of RuBisCO from the concentrated samples.

Gossypium hirsutum

Cotton plants (*G. hirsutum* cv. Deltapine 90) were grown in the Whistler Greenhouse at Purdue University. The greenhouse was controlled to retain 60 to 80% of relative humidity and remain at 26°C/22°C for a day/night with 16 hours of supplementary lighting. Flowers were marked as 0 day post-anthesis (0 DPA) at anthesis and harvested at 10 DPA, representing the maximum fiber elongation period. The harvested cotton bolls pooled from five plants were dissected immediately to obtain fiber tissue (*51*). Three grams of fibers was used for soluble protein isolation as described above, except for the cheesecloth filtration. RuBisCO depletion was not performed in this sink type of tissue.

Glycine max

Williams 82 cultivar was grown at 26°C/22°C for a day/night with 16 hours supplementary lighting and 60 to 80% of relative humidity in the Lilly Greenhouses at Purdue University. Two grams of trifoliate leaves at the fourth node was harvested from 2-month-old plants. Crude cytosolic protein extraction was conducted as described above, except for the depletion of RuBisCO.

Oryza sativa

Kitaake (*O. sativa* L. ssp. *japonica*) cultivar was grown in a Conviron E15 growth chamber (Conviron) with a day/night setting of 26°C/22°C and 12/12 hours. One hundred developing seeds at 10 to 12 days after flowering were peeled to isolate sub- and aleurone layers as described before (*52*). The peeled layers were homogenized and differentially centrifuged to obtain crude cytosolic proteins as mentioned above, except for the filtration step through cheesecloth and sample concentration by the 10-kDa cutoff column. The depletion of RuBisCO was not conducted in this sink type of tissue.

Size exclusion chromatography

The soluble protein samples were fractionated by an ÄKTA fast protein LC system (Amersham Biosciences) as described previously (*20*, *22*, *27*). Approximately 1 mg of total proteins in the *A. thaliana*, *G. hirsutum*, and *O. sativa* samples was loaded onto a Superdex 200 10/300 GL column (GE Healthcare). The SEC elution was carried out with the mobile phase [50 mM Hepes/KOH (pH 7.8), 100 mM NaCl, 10 mM $MgCl_2$, and 5% glycerol] at a flow rate of 0.65 ml/min. Protein elution was monitored by absorption at 280 nm. For LC-MS/MS analysis, 500 μl of SEC fractions was collected. In each replicate for all species, an equivalent of at least 24 fractions were analyzed across the full mass range of the column. For *G. max* SEC separation, about 4 mg of cytosolic proteins was fractionated by a HiLoad 16/600

Superdex 200 pg column (GE Healthcare) at a flow rate of 1 ml/min. For LC-MS/MS analysis, 500 µl of 36 SEC fractions was collected. Both columns were calibrated using dextran blue to measure the void volume and the gel filtration marker proteins MWGF1000 (Sigma-Aldrich) with a mass range from 669 to 29 kDa.

Sample preparation for LC-MS/MS analysis

Fractionated samples were prepared for LC-MS/MS analysis as described previously (*22*, *26*). The mobile phase was removed, and proteins in each fraction were precipitated by acetone precipitation. Dried pellets were denatured in 8 M urea for 1 hour at room temperature, reduced in 10 mM DTT for 45 min at 60°C, and then alkylated with 20 mM iodoacetamide for 45 min at room temperature in the dark. The urea concentration of the peptide solution was brought down to below 1.5 M by additional ammonium bicarbonate for trypsin digestion. The digested peptides were purified using Pierce C18 Spin Columns (Thermo Fisher Scientific Inc.) according to the manufacturer's protocol. Peptide concentration for each SEC fraction was measured by bicinchoninic acid assay following the manufacturer's protocol. All the peptide samples were adjusted to equal volume. The most concentrated sample had a peptide concentration of 0.2 µg/µl. Five microliters of each sample was injected onto LC-MS/MS. Duplicates were analyzed.

LC-MS/MS data acquisition

LC-MS/MS analysis was performed as described previously (*26*). In brief, a Q Exactive HF Hybrid Quadrupole-Orbitrap mass spectrometer in conjunction with a reverse-phase high-performance LC (HPLC)–electrospray ionization (ESI)–MS/MS using a Dionex UltiMate 3000 RSLCnano System (Thermo Fisher Scientific Inc.) was used for the analysis of *A. thaliana*, *G. hirsutum*, and *O. sativa* peptide samples. Peptides were resolved over a 125-min gradient with a flow rate of 300 nl/min. An MS survey scan was acquired from 350 to 1600 mass/charge ratio range. MS/MS spectra were acquired by selecting the 20 most abundant precursor ions for sequencing with high-energy collisional dissociation normalized collision energy of 27%. A 15-s dynamic exclusion window was applied for reducing the number of times the same ion was sequenced. For *G. max*, samples were analyzed as described previously (*27*). Briefly, peptides were resolved over a 90-min gradient by a reverse-phase HPLC-ESI-MS/MS using an Eksigent ekspert NanoLC 425 HPLC system (Dublin, CA, USA). MS/MS acquisition was conducted on an AB Sciex quadrupole time-of-flight (Q-TOF) TripleTOF 5600 plus mass spectrometer (AB Sciex Pte. Ltd.) operated in a data-dependent mode.

Peptide identification and quantification

Acquired MS/MS files were submitted simultaneously to MaxQuant (version 1.5.6.5) for protein identification and relative quantification using the Andromeda search engine (*53*). The search was conducted with all the fractions obtained from the biological duplicates in one species at a single search as described previously (*26*). Each MS/MS spectrum in each species was searched against TAIR Athaliana_167_TAIR10.protein_primaryTranscriptOnly.fa (*54*), the cotton FASTA file provided from J. Wendel, Joint Genome Institute (JGI) Gmax_275_Wm82.a2.v1.protein_primaryTranscriptOnly.fa (*55*), and JGI Osativa_323_v7.0.protein_primaryTranscriptOnly.fa (*56*) for *Arabidopsis*, cotton, soybean, and rice, respectively. The search parameters were set as follows: Cysteine carbamidomethylation was chosen as a fixed modification; oxidation on methionine and acetylation on protein N terminus were chosen as variable modifications; trypsin was selected as an enzyme specificity, and up to two missed cleavages were accepted for digestion error; the precursor mass tolerance and the fragment mass tolerance were set at 10 and 40 parts per million, respectively; one percent of the protein and peptide FDR was chosen; a reverse decoy database was enabled; for identification, proteins should have at least one peptide that had a minimum of seven–amino acid length; peptide abundance was calculated from the extracted ion current for both razor and unique peptides, and protein level signals were aggregated from peptide intensities; and "match between runs" function was set with a maximum matching time window of 0.7 min as default. After MaxQuant searching, proteins identified by at least one unique peptide were selected, and all the protein groups were filtered out when they had zero intensity value and zero MS/MS count.

Reproducibility, Gaussian peak fitting, and R_{app} calculations

Reproducibility between two biological SEC profile replicates was determined as described previously (*22*, *26*). An optimized Gaussian fitting algorithm was previously developed to convert raw profile data into fitted peaks and to deconvolve profiles with multiple peaks into individual profiles (*22*, *26*). The algorithm identified protein peaks when they had more than three nonzero fractions, with two being adjacent. Multiple reproducible peaks from a protein were split into multiple entries by labeling with a "_peak number" on their locus identifications (IDs). If a protein elution profile was not fitted to a Gaussian curve, then the global maximum (G_{max}) was replaced to its peak location. The reproducible peaks were selected from the two replicates if they were present within two fraction shifts or three fraction shifts for soybean, considering the rate of mass increment between fractions in both columns. All nonreproducible peaks were not used for the following analysis. The fraction locations of the fitted peaks were used to determine the apparent masses (M_{app}) of proteins using the SEC calibration curve obtained above. The protein multimerization state (R_{app}) was defined as the ratio of the M_{app} of a protein to the theoretical monomer mass (M_{mono}) of the protein. Peaks greater than M_{app} of 760 kDa or less than R_{app} of 0.62 were filtered out as unresolved or degraded.

Protein assignments into OGs

Orthologous proteins were searched against Phytozome v12.1 database (*57*). The gene ancestry file (cluster_members_Angiosperm_5197.tsv.gz), which includes all the ancestral families with memberships belonging to the family, was directly downloaded from Phytozome. To generate ortholog datasets, first, an OG converter was made by filtering out those memberships from species that are not experimental plants in this project. Cluster IDs in the gene ancestry file were used as OG identification. Reproducible proteins from each species were assigned into OGs using the converter. OGs with orthologs from all four species were referred to as "common OGs" and used for cross-species comparison. For within a species comparison, ortholog searching was performed within each species as described above, and OGs that contained at least two proteins were subjected to further analysis.

mOV detection and OG classification

ANOVA and Tukey's post hoc test were performed to find mOVs in OGs. Multimerization variations in the common OGs were tested by

the ANOVA test against M_{app}. Some orthologous proteins have incredibly long sequences, which, in turn, prevent a proper multimerization prediction. A monomeric ortholog with twice longer amino acid sequences cannot be distinguished from its dimeric ortholog in terms of M_{app}, and vice versa. Thus, ANOVA against both M_{app} and R_{app} of orthologs in an OG was run. This multiple comparison test was set at a 2% FDR to minimize likely false positives and negatives based on manual inspection of the profile patterns and pairwise tests.

Then, Tukey's post hoc test was performed at 0.05 level to group mOVs in the flagged OGs. These statistic test results were confirmed manually by investigating elution profiles of two replicates from each ortholog to answer whether or not both profiles of an mOV showed significant difference with its orthologs at the Tukey's post hoc test. In most cases, Tukey's results clearly explained two or three distinct multimerization states within each OG. However, in some cases, the results defined several groups within a small range of M_{app} or R_{app}, which could not be resolved by the SEC column due to its limitation of resolving power. For instance, when all orthologs in the same OG had $R_{app} < 1.6$, these proteins were defined as monomers and conserved with respect to multimeric states, considering the column resolution. In addition, a flagged OG was saved if the estimated value of the averaged R_{app} or M_{app} minimum multiplied by the increase rate of mass per SEC fraction (26%) is overlapped with the calculated value of the averaged maximum multiplied by the decrease rate of mass per SEC fraction (20%) in the OG.

A profile pattern–based classification scheme was defined to categorize the ortholog comparison results. OGs consisting of orthologous proteins with either statistically insignificant M_{app} or R_{app}, compared to each other, were classified as "class I." This category is consistent with the hypothesis that orthologs have conserved binding partners. OGs with mOVs were further sorted into two classes based on the pattern of the peaks among orthologs using Tukey's grouping results. Thus, flagged OGs were defined as "class II (segregated)" if mOVs resolved into two (52 OGs) or three (17 OGs) clearly discrete groups. Another class includes orthologs with a significant difference by ANOVA test and highly dispersed pattern of multimerization. Therefore, we referred to "class III (dispersed)." Class III was defined if Tukey's test were not able to show significance between the majority of orthologs in an OG, but Tukey's result distinctly distinguished the biggest mOV and the smallest mOV. The characteristics of this class III are that mOVs showed continuous and dispersed patterns. So, if an OG did not follow this definition and there is a distinct gap between Tukey's grouping, then this OG was recategorized into class II.

Protein sequence alignments and phylogeny methods

Multiple sequences were aligned using MUltiple Sequence Comparison by Log-Expectation (MUSCLE) algorithm (*58*) using default program parameters on Geneious Prime software (version 2019.0.3). The maximum number of iterations was eight as default. The optimized profile-dependent parameters were selected for the alignment optimization. Geneious tree builder on Geneious Prime software (version 2019.0.3) was used to generate phylogenetic trees based on the multiple sequence alignment results. The neighbor-joining phylogenetic trees were built using the Jukes-Cantor genetic distance model. The trees were resampled with the bootstrap method under default parameters as follows: The number of replicates was 1000, and the support threshold percentage was 50. The resulting trees were saved as .pdf files, and the patristic distance matrices were exported as .csv files for further analysis.

Correlation analyses between mOVs and phylogenetic signals

We analyzed whether the phylogenetic distance of orthologs reflected M_{app} and whether the evolutionary distance of species reflected M_{app} distribution. For the comparison of phylogenetic distances to M_{app} distributions, M_{app} distance matrices were made and then compared to the patristic distance matrices exported from the resulting trees above. Both distance matrices were subjected to linear regression analysis. Diagnostic scatter plots were simultaneously created to evaluate how well the models fit the data. To analyze the influence of the evolutionary distance of species on M_{app} distributions, the evolutionary distance of species was estimated. The TimeTree search option on The Timescale of Life was used to build a timetree among the four species (*59*). The written tree was exported to a Newick file, which was uploaded onto Geneious Prime software (version 2019.0.3) to estimate patristic distances between species. A linear regression analysis between the evolutionary distance of species and M_{app} distributions was performed. A positive slope, with an adjusted $R^2 > 0.28$ and the $P < 0.05$ from the t test, was used to judge whether the phylogeny or evolution of species significantly affected M_{app}.

Extended OG phylogenies

The resulting phylogenetic trees were also used to predict a driving force for mOVs. OGs, where mOVs with similar multimerization states were clustered into the same phylogenetic clades, were referred to as "ancient paralog." The orthology relationship between mOVs in each OG was determined using the paralog information (*29*, *60*, *61*). Paralogous proteins for *Arabidopsis*, soybean, and rice are denoted in tables S1 and S2. For example, βCA5, which was clustered into the dimeric clade, has no paralogous relationship with other βCAs in *Arabidopsis*, implying that the βCA5 was diversified from ancestral βCA at least before *Arabidopsis* speciation (Fig. 4). This can support that the multimerization diversification was achieved during ancient ortholog evolution. When a single species has distinct mOVs, the phylogenetic tree assigned the mOVs into a unique clade, and any more paralogs of the mOVs were not found from the paralog information; we called this category as "species-specific ortholog evolution." Although some OGs had similar parameters, we defined them as "potential species-specific ortholog evolution." This is because those OGs had relatively small numbers of detected orthologs, which could limit the prediction of the driving force on multimerization selection. Thus, we used all undetected ortholog proteins descended from a common ancestral protein in each OG, which showed a correlation between phylogeny and M_{app}. The phylogenetic trees were generated with all members of proteins in OGs, and the selection forces were inferred with paralog information (tables S3 and S4 and fig. S3).

Statistic tests and data analysis

Statistical analysis was performed using R version 3.5.1 (*62*) on RStudio 1.1.463 (*63*). Microsoft Excel on Office 365 for Mac was used for all data management. PyMOL (*64*) was used to visualize the carbonic anhydrase structure.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/13/eabf0984/DC1

REFERENCES AND NOTES

1. S. Ohno, *Evolution By Gene Duplication* (Springer, Verlag, 1970).
2. M. Lynch, J. S. Conery, The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
3. R. C. Hardison, Evolution of hemoglobin and its genes. *Cold Spring Harb. Perspect. Med.* **2**, a011627 (2012).
4. J. W. Clark, P. C. J. Donoghue, Whole-genome duplication and plant macroevolution. *Trends Plant Sci.* **23**, 933–945 (2018).
5. M. W. Giorgianni, N. L. Dowell, S. Griffin, V. A. Kassner, J. E. Selegue, S. B. Carroll, The origin and diversification of a novel protein family in venomous snakes. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10911–10920 (2020).
6. J. Doebley, A. Stec, L. Hubbard, The evolution of apical dominance in maize. *Nature* **386**, 485–488 (1997).
7. M. A. Sirover, On the functional diversity of glyceraldehyde-3-phosphate dehydrogenase: Biochemical mechanisms and regulatory control. *Biochim. Biophys. Acta* **1810**, 741–751 (2011).
8. J. K. Weng, Y. Li, H. Mo, C. Chapple, Assembly of an evolutionarily new pathway for α-pyrone biosynthesis in *Arabidopsis*. *Science* **337**, 960–964 (2012).
9. K. Hashimoto, A. R. Panchenko, Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 20352–20357 (2010).
10. B. Alberts, The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998).
11. J. Monod, J. Wyman, J. P. Changeux, On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12**, 88–118 (1965).
12. J. Kuriyan, D. Eisenberg, The origin of protein interactions and allostery in colocalization. *Nature* **450**, 983–990 (2007).
13. D. Voges, P. Zwickl, W. Baumeister, The 26S proteasome: A molecular machine designed for controlled proteolysis. *Annu. Rev. Biochem.* **68**, 1015–1068 (1999).
14. M. Kaksonen, C. P. Toret, D. G. Drubin, A modular design for the clathrin- and actin-mediated endocytosis machinery. *Cell* **123**, 305–320 (2005).
15. C. Aguilar-Gurrieri, A. Larabi, V. Vinayachandran, N. A. Patel, K. Yen, R. Reja, I. O. Ebong, G. Schoehn, C. V. Robinson, B. F. Pugh, D. Panne, Structural evidence for Nap1-dependent H2A–H2B deposition and nucleosome assembly. *EMBO J.* **35**, 1465–1482 (2016).
16. D. Basu, J. Le, T. Zakharova, E. L. Mallery, D. B. Szymanski, A SPIKE1 signaling complex controls actin-dependent cell morphogenesis through the heteromeric WAVE and ARP2/3 complexes. *Proc. Natl. Acad. Sci.* **105**, 4044–4049 (2008).
17. A. R. Kristensen, J. Gsponer, L. J. Foster, A high-throughput approach for measuring temporal changes in the interactome. *Nat. Methods* **9**, 907–909 (2012).
18. P. C. Havugimana, G. T. Hart, T. Nepusz, H. Yang, A. L. Turinsky, Z. Li, P. I. Wang, D. R. Boutz, V. Fong, S. Phanse, M. Babu, S. A. Craig, P. Hu, C. Wan, J. Vlasblom, V. Dar, A. Bezginov, G. W. Clark, G. C. Wu, S. J. Wodak, E. R. M. Tillier, A. Paccanaro, E. M. Marcotte, A. Emili, A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012).
19. K. J. Kirkwood, Y. Ahmad, M. Larance, A. I. Lamond, Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics. *Mol. Cell. Proteomics* **12**, 3851–3873 (2013).
20. U. K. Aryal, Y. Xiong, Z. McBride, D. Kihara, J. Xie, M. C. Hall, D. B. Szymanski, A proteomic strategy for global analysis of plant protein complexes. *Plant Cell* **26**, 3867–3882 (2014).
21. C. Wan, B. Borgeson, S. Phanse, F. Tu, K. Drew, G. Clark, X. Xiong, O. Kagan, J. Kwan, A. Bezginov, K. Chessman, S. Pal, G. Cromar, O. Papoulas, Z. Ni, D. R. Boutz, S. Stoilova, P. C. Havugimana, X. Guo, R. H. Malty, M. Sarov, J. Greenblatt, M. Babu, W. B. Derry, E. R. Tillier, J. B. Wallingford, J. Parkinson, E. M. Marcotte, A. Emili, Panorama of ancient metazoan macromolecular complexes. *Nature* **525**, 339–344 (2015).
22. Z. McBride, D. Chen, Y. Lee, U. K. Aryal, J. Xie, D. B. Szymanski, A label-free mass spectrometry method to predict endogenous protein complex composition. *Mol. Cell. Proteomics* **18**, 1588–1606 (2019).
23. D. Salas, G. R. Stacey, M. Akinlaja, L. J. Foster, Next-generation Interactomics: Considerations for the use of co-elution to measure protein interaction networks. *Mol. Cell. Proteomics* **19**, 1–10 (2020).
24. T. W. M. Crozier, M. Tinti, M. Larance, A. I. Lamond, M. A. J. Ferguson, Prediction of protein complexes in *Trypanosoma brucei* by protein correlation profiling mass spectrometry and machine learning. *Mol. Cell. Proteomics* **16**, 2254–2267 (2017).
25. C. D. McWhite, O. Papoulas, K. Drew, R. M. Cox, V. June, O. X. Dong, T. Kwon, C. Wan, M. L. Salmi, S. J. Roux, K. S. Browning, Z. J. Chen, P. C. Ronald, E. M. Marcotte, A pan-plant protein complex map reveals deep conservation and novel assemblies. *Cell* **181**, 460–474. e14 (2020).
26. Z. McBride, D. Chen, C. Reick, J. Xie, D. B. Szymanski, Global analysis of membrane-associated protein oligomerization using protein correlation profiling. *Mol. Cell. Proteomics* **16**, 1972–1989 (2017).
27. U. K. Aryal, Z. McBride, D. Chen, J. Xie, D. B. Szymanski, Analysis of protein complexes in *Arabidopsis* leaves using size exclusion chromatography and label-free protein correlation profiling. *J. Proteomics* **166**, 8–18 (2017).
28. E. L. L. Sonnhammer, G. Östlund, InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **43**, D234–D239 (2015).
29. G. Blanc, K. H. Wolfe, Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678 (2004).
30. T. Mandáková, S. Joly, M. Krzywinski, K. Mummenhoff, M. A. Lysak, Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *Plant Cell* **22**, 2277–2290 (2010).
31. J. F. Wendel, New world tetraploid cottons contain old world cytoplasm. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 4132–4136 (1989).
32. T. Zhang, Y. Hu, W. Jiang, L. Fang, X. Guan, J. Chen, J. Zhang, C. A. Saski, B. E. Scheffler, D. M. Stelly, A. M. Hulse-Kemp, Q. Wan, B. Liu, C. Liu, S. Wang, M. Pan, Y. Wang, D. Wang, W. Ye, L. Chang, W. Zhang, Q. Song, R. C. Kirkbride, X. Chen, E. Dennis, D. J. Llewellyn, D. G. Peterson, P. Thaxton, D. C. Jones, Q. Wang, X. Xu, H. Zhang, H. Wu, L. Zhou, G. Mei, S. Chen, Y. Tian, D. Xiang, X. Li, J. Ding, Q. Zuo, L. Tao, Y. Liu, J. Li, Y. Lin, Y. Hui, Z. Cao, C. Cai, X. Zhu, Z. Jiang, B. Zhou, W. Guo, R. Li, Z. J. Chen, Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* **33**, 531–537 (2015).
33. V. Vernoud, A. C. Horton, Z. Yang, E. Nielsen, Analysis of the small GTPase gene superfamily of *Arabidopsis*. *Plant Physiol.* **131**, 1191–1208 (2003).
34. J. Kamsteeg, J. V. Brederode, G. V. Nigtevecht, The formation of UDP-L-rhamnose from UDP-D-glucose by an enzyme preparation of red campion (*Silene Dioica* (L) clairv) leaves. *FEBS Lett.* **91**, 281–284 (1978).
35. A. M. Showalter, Arabinogalactan-proteins: Structure, expression and function. *Cell. Mol. Life Sci.* **58**, 1399–1417 (2001).
36. O. Llorca, M. Betti, J. M. González, A. Valencia, A. J. Márquez, J. M. Valpuesta, The three-dimensional structure of an eukaryotic glutamine synthetase: Functional implications of its oligomeric structure. *J. Struct. Biol.* **156**, 469–479 (2006).
37. X. Wang, Y. Wei, L. Shi, X. Ma, S. M. Theg, New isoforms and assembly of glutamine synthetase in the leaf of wheat (Triticum aestivum L.). *J. Exp. Bot.* **66**, 6827–6834 (2015).
38. L. Puerto-Galán, J. M. Pérez-Ruiz, J. Ferrández, B. Cano, B. Naranjo, V. A. Nájera, M. González, A. M. Lindahl, F. J. Cejudo, Overoxidation of chloroplast 2-Cys peroxiredoxins: Balancing toxic and signaling activities of hydrogen peroxide. *Front. Plant Sci.* **4**, (2013).
39. M. Perales, H. Eubel, J. Heinemeyer, A. Colaneri, E. Zabaleta, H. P. Braun, Disruption of a nuclear gene encoding a mitochondrial gamma carbonic anhydrase reduces complex I and supercomplex I+III2 levels and alters mitochondrial physiology in *Arabidopsis*. *J. Mol. Biol.* **350**, 263–277 (2005).
40. M. S. Kimber, E. F. Pai, The active site architecture of *Pisum sativum* beta-carbonic anhydrase is a mirror image of that of *alpha*-carbonic anhydrases. *EMBO J.* **19**, 1407–1418 (2000).
41. S. Huang, T. Hainzl, C. Grundström, C. Forsman, G. Samuelsson, A. E. Sauer-Eriksson, Structural studies of β-carbonic anhydrase from the green alga *Coccomyxa*: Inhibitor complexes with anions and acetazolamide. *PLOS ONE* **6**, e28458 (2011).
42. R. J. DiMario, H. Clayton, A. Mukherjee, M. Ludwig, J. V. Moroney, Plant carbonic anhydrases: Structures, locations, evolution, and physiological roles. *Mol. Plant* **10**, 30–46 (2017).
43. G. Le Hir, Y. Donnadieu, Y. Goddéris, B. Meyer-Berthaud, G. Ramstein, R. C. Blakey, The climate change caused by the land plant invasion in the Devonian. *Earth Planet. Sci. Lett.* **310**, 203–212 (2011).
44. T. O. Yeates, C. A. Kerfeld, S. Heinhorst, G. C. Cannon, J. M. Shively, Protein-based organelles in bacteria: Carboxysomes and related microcompartments. *Nat. Rev. Microbiol.* **6**, 681–691 (2008).
45. T. Ren, S. M. Weraduwage, T. D. Sharkey, Prospects for enhancing leaf photosynthetic capacity by manipulating mesophyll cell morphology. *J. Exp. Bot.* **70**, 1153–1165 (2019).
46. F. J. Ferreira, C. Guo, J. R. Coleman, Reduction of plastid-localized carbonic anhydrase activity results in reduced *Arabidopsis* seedling survivorship. *Plant Physiol.* **147**, 585–594 (2008).
47. N. Fabre, I. M. Reiter, N. Becuwe-Linka, B. Genty, D. Rumeau, Characterization and expression analysis of genes encoding α and β carbonic anhydrases in *Arabidopsis*. *Plant Cell Environ.* **30**, 617–629 (2007).
48. V. Endeward, M. Arias-Hidalgo, S. Al-Samir, G. Gros, CO_2 permeability of biological membranes and role of CO_2 channels. *Membranes* **7**, 61 (2017).
49. J. A. Birchler, R. A. Veitia, Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14746–14753 (2012).
50. E. Kuzmin, B. V. Sluis, A. N. N. Ba, W. Wang, E. N. Koch, M. Usaj, A. Khmelinskii, M. M. Usaj, J. van Leeuwen, O. Kraus, A. Tresenrider, M. Pryszlak, M. C. Hu, B. Varriano, M. Costanzo, M. Knop, A. Moses, C. L. Myers, B. J. Andrews, C. Boone, Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. *Science* **368**, eaaz5667 (2020).
51. G. Hu, J. Koh, M. J. Yoo, K. Grupp, S. Chen, J. F. Wendel, Proteomic profiling of developing cotton fibers from wild and domesticated *Gossypium barbadense*. *New Phytol.* **200**, 570–582 (2013).

52. Y. Yang, H. L. Chou, A. J. Crofts, L. Zhang, L. Tian, H. Washida, M. Fukuda, T. Kumamaru, O. J. Oviedo, S. R. Starkenburg, T. W. Okita, Selective sets of mRNAs localize to extracellular paramural bodies in a rice glup6 mutant. *J. Exp. Bot.* **69**, 5045–5058 (2018).
53. J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, M. Mann, Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
54. P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel, E. Huala, The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).
55. J. Schmutz, S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng, D. Xu, U. Hellsten, G. D. May, Y. Yu, T. Sakurai, T. Umezawa, M. K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X. C. Zhang, K. Shinozaki, H. T. Nguyen, R. A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R. C. Shoemaker, S. A. Jackson, Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
56. S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R. L. Malek, Y. Lee, L. Zheng, J. Orvis, B. Haas, J. Wortman, C. R. Buell, The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Res.* **35**, D883–D887 (2007).
57. D. M. Goodstein, S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, D. S. Rokhsar, Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2011).
58. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
59. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
60. H. Lin, S. Ouyang, A. Egan, K. Nobuta, B. J. Haas, W. Zhu, X. Gu, J. C. Silva, B. C. Meyers, C. R. Buell, Characterization of paralogous protein families in rice. *BMC Plant Biol.* **8**, 18 (2008).
61. C. Xu, B. D. Nadon, K. D. Kim, S. A. Jackson, Genetic and epigenetic divergence of duplicate genes in two legume species. *Plant Cell Environ.* **41**, 2033–2044 (2018).
62. R Core Team. (R Foundation for Statistical Computing, 2018).
63. RStudio Team. (RStudio Inc., 2018).
64. L. L. C. Schrödinger, The PyMOL Molecular Graphics System, Version 1.8 (2015).
65. A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, H. W. Mewes, CORUM: The comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* **38**, D497–D501 (2009).
66. U. K. Aryal, S. J. Callister, B. H. McMahon, L. A. McCue, J. Brown, J. Stöckel, M. Liberton, S. Mishra, X. Zhang, C. D. Nicora, T. E. Angel, D. W. Koppenaal, R. D. Smith, H. B. Pakrasi, L. A. Sherman, Proteomic profiles of five strains of oxygenic photosynthetic cyanobacteria of the genus *Cyanothece*. *J. Proteome Res.* **13**, 3262–3276 (2014).
67. M. Higurashi, T. Ishida, K. Kinoshita, PiSite: A database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res.* **37**, D360–D364 (2009).
68. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

Acknowledgments: We thank J. Wendel at Iowa State University, J. Banks at Purdue University, and Z. Zhao at CSIRO for the helpful discussions and the anonymous reviewers for the helpful suggestions. U. Aryal generated the LC/MS data from soybean leaves, and the Purdue Proteomics Facility ran the LC/MS samples. **Funding:** This work was supported by the National Science Foundation (PGRP grant nos. 1444610 and 1951819 to D.B.S.). Y.L. was a grad student fellow with Bilsland Graduate Dissertation Fellowship from the Graduate School at Purdue University and CPB Graduate Research Award from the Center for Plant Biology at Purdue University. **Author contributions:** Y.L. and D.B.S. conceived the project, analyzed the data, and wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. The MS raw files have been deposited to the ProteomeXchange Consortium via the PRIDE with accession codes PXD021454 (*A. thaliana*), PXD021451 (*G. hirsutum*), PXD021452 (*G. max*), and PXD021450 (*O. sativa*). The mass spectra are available at the ProteinProspector with search key 7uukknykfi (*A. thaliana*), xf02eieqif (*G. hirsutum*), r0aolnft13 (*G. max*), and dyerrueilh (*O. sativa*).

Submitted 3 October 2020
Accepted 29 January 2021
Published 26 March 2021
10.1126/sciadv.abf0984

Citation: Y. Lee, D. B. Szymanski, Multimerization variants as potential drivers of neofunctionalization. *Sci. Adv.* **7**, eabf0984 (2021).

Multimerization variants as potential drivers of neofunctionalization

Youngwoo Lee and Daniel B. Szymanski

Sci Adv **7** (13), eabf0984.
DOI: 10.1126/sciadv.abf0984