Science
AAAS

Supporting Online Material for

Evidence for Network Evolution in an *Arabidopsis* Interactome Map

Arabidopsis Interactome Mapping Consortium.

correspondence to: marc_vidal@dfci.harvard.edu; ecker@salk.edu;
pascal_braun@dfci.harvard.edu; david_hill@dfci.harvard.edu

This PDF file includes:

SOM Text

Glossary

References

Figs. S1 to S47

Other Supporting Online Material for this manuscript includes the following:

Tables S1 to S12 as zipped archives.

SUPPORTING ONLINE MATERIAL METHODS

SOM I: CONSTRUCTION OF AN *ARABIDOPSIS* ORF COLLECTION: AtORFeome2.0

The construction of the Gateway compatible (*44, 45*) AtORFeome2.0 clone collection was accomplished by transferring the "SSP/Salk pUNI51" clone collection (*2*) into a custom pENTR vector, pENTR-*Sfi*I-223. pENTR-*Sfi*I-223 contains two unique Sfi*I* restriction sites (Sfi*I*-A and Sfi*I*-B) inside the *att*L1 and *att*L2 Gateway recombination sites, respectively, flanking the *ccdB*-CAM^R negative selection cassette. This vector was designed to allow the directional subcloning of open reading frames (ORFs) situated between *Sfi*I-A and *Sfi*I-B in pUNI51. To transfer the 7,109 "SSP/Salk pUNI51" clones into pENTR-*Sfi*I-223, we first inoculated liquid Luria-Bertani (LB) media containing kanamycin (50 µg/ml) with bacterial glycerol stocks of pUNI51 ORF clones, in a 96-well format. After overnight growth at 37°C, plasmid DNA was extracted and purified using either the ChargeSwitch NoSpin Plasmid Micro Kit (Invitrogen) or the Purelink HQ 96 kit (Invitrogen). Plasmid DNA was then cut by restriction digestion with *Sfi*I, at 50°C for 2 hours. Digested pUNI51-ORF DNA was mixed with purified *Sfi*I-digested linearized pENTR-*Sfi*I-223 DNA (*ccdB*-CAM^R cassette removed) and ligated overnight at 4°C. Chemically competent *Escherichia coli* cells (DH10B-T1^R or DH5α strains) were transformed with ligation reaction products and plated on solid LB media with 100 µg/ml spectinomycin to select for transformed cells. From each transformation reaction (each pENTR-*Sfi*I-223-ORF ligation), a single transformant was picked and used to inoculate an overnight culture in 96-well format. From these cultures, archival glycerol stocks were prepared.

Liquid LB with spectinomycin (100 µg/ml) was inoculated with bacteria transformed with pENTR-*Sfi*I-223-ORF clones. After overnight growth, plasmid DNA was extracted and purified using the Purelink HQ 96 kit (Invitrogen). The identity of pENTR-*Sfi*I-223-ORF clones was verified by end-read sequencing (5' and 3') using the following primers:

pENTR-*Sfi*I_Fwd: 5'-TAAGCTCGGGCCCCAAATAAT-3'
pENTR-*Sfi*I_Rev: 5'-GGATATCAGCTGGATGGCAAA-3'.

The ORFs whose sequence could be matched and mapped to the expected genomic loci and for which in-frame SfiI junctions could be confirmed were robotically re-arrayed into new plates. In addition, 1,474 pENTR-TOPO-ORF clones constructed as part of the SSP consortium (*2*) were also included in the collection. Together these two sets of ORFs represent the AtORFeome2.0 collection (8,583 ORFs) and were used as a starting point for the interactome mapping system (fig. S2A).

SOM II: BINARY INTERACTOME MAPPING SYSTEM

All interactome mapping experiments in this manuscript were performed essentially as described before (*4, 6, 7, 9, 10*), with the following modifications (fig. S2B): (i) space 1 was screened twice to increase sampling; (ii) reproducibility was optimized by testing the phenotype of each candidate Y2H pair four times and retaining only those that scored positive at least three times (fig. S2B); and (iii) the identity of each verified Y2H pair was confirmed by DNA sequencing. We summarize the procedures below.

Preparation of Y2H reagents

ORFs from the AtORFeome2.0 collection were transferred by Gateway LR recombinational cloning (Invitrogen) into pDEST-DB and pDEST-AD-*CYH2* yeast two-hybrid (*4*) destination vectors to generate Gal4 DNA binding domain (DB)-X hybrid proteins and Gal4 activation domain (AD)-Y hybrid proteins, respectively. Recombinational products were directly used to transform *E. coli* (DH5α-T1^R strain) via a selection for ampicillin resistance in liquid LB media. After overnight growth (20 hours), plasmid DNA was extracted from bacteria in a 96-well format using a Qiagen 8000 Miniprep Biorobot.

Yeast strains and yeast transformation

We used the yeast strains Y8800 and Y8930 (*4*), of mating type *MAT***a** and *MAT*α respectively, which harbor the following genotype: *leu2-3,112 trp1-901 his3Δ200 ura3-52 gal4Δ gal80Δ GAL2::ADE2 GAL1::HIS3@LYS2 GAL7::lacZ@MET2 cyh2*^{R}. The availability of two haploid strains of opposite mating types enables the use of mating to efficiently combine large collections of DB-X and AD-Y hybrid constructs. By convention the Y8800 *MAT***a** and Y8930 *MAT*α strains are transformed with AD-Y and DB-X hybrid constructs, respectively. The reporter genes *GAL2-ADE2* and *GAL1-HIS3* are integrated into the yeast genome. Expression of the *GAL1-HIS3* reporter gene was tested with 1 mM 3-AT (3-amino-1,2,4-triazole, a competitive inhibitor of the *HIS3* gene product) to reduce background growth.

Competent Y8930 (*MAT*α) and Y8800 (*MAT***a**) yeast strains were transformed with individual DB-X and AD-Y plasmid constructs respectively and plated onto selective synthetic complete (SC) solid media without leucine (SC-Leu; DB-X) or without tryptophan (SC-Trp; AD-Y) to select transformants. Transformed yeast cells were transferred into selective SC liquid media and grown at 30℃ for 4 days. We prepared archival stocks of transformants by combining an equal volume of liquid culture and 40% (w/v) sterile glycerol and stored them at -80℃.

Space 1 definition

Space 1 is defined as the 2-dimensional combinatorial space of all ORFs for which a *GAL4*AD-ORF hybrid construct (1st dimension) or a *GAL4*DB-ORF hybrid construct (2nd dimension) were screened against each other (fig. S2B). We mapped the corresponding ORFs to TAIR7 (*46*) genomic loci (space 1, table S3). This mapping mainly consisted of exact sequence matching between ORF sequence and TAIR7 gene models (82% ORFs). For ORFs without an exact match, we performed both BLAST and BLAT (*47*) on the ORF sequences; the BLAST was against TAIR7 coding sequence with an *E*-value cutoff of 10^{-5} and the BLAT was against the TAIR7 genome (default options). We then compared the best BLAST result with the best BLAT result. In most cases these two analyses indicated the same gene model, to which the ORF was thus mapped. In less than 0.1% of cases, there was no BLAST hit for the ORF sequence, therefore the ORF was assigned a locus solely based on its best BLAT result. Space 1 is defined by 8,583 ORFs corresponding to 8,429 TAIR7 genomic loci: 8,033 AD-ORF constructs correspond to 7,895 gene loci and 7,760 DB-ORF constructs correspond to 7,644 gene loci.

Auto-activator identification and removal

Prior to Y2H screening, DB-X yeast strains were tested for auto-activation of the *GAL1-HIS3* Y2H reporter gene in the absence of any AD-Y plasmid. Individual DB-X yeast strains were mated with the Y8800 yeast strain transformed with pDEST-AD-*CYH2* (empty vector). Diploid cells were first selected on solid SC-Leu-Trp media then transferred onto solid SC-Leu-Trp media lacking histidine and containing 1mM 3-AT (SC-Leu-Trp-His + 1mM 3-AT, hereafter -His) to select for diploid cells that showed activation of the *GAL1-HIS3* Y2H reporter gene. To increase auto-activator scoring confidence, after incubation at 30℃ for 3 days, growth was scored twice by independent observers relative to the "no interaction" Y2H control (*4*). In total, 1,126 (15%) DB-X auto-activating yeast strains were identified and removed from the collection of DB-X yeast strains by robotically re-arraying the non auto-activators into new plates. Overall, 6,634 *Arabidopsis* ORFs were screened as Gal4-DB hybrid constructs against 7,896 *Arabidopsis* ORFs as Gal4-AD hybrid constructs.

AD-ORF mini-library assembly

To increase screening efficiency (number of pairs tested at once) of the Y2H screening pipeline, we used 48 AD-Y mini-libraries, each containing 192 distinct AD-Y yeast strains (two 96-well plates). This mini-library size has been experimentally determined to represent a good compromise between screening efficiency and screening sensitivity (*10*). By combining the contents of ninety-six 96-well plates, two at a time, we assembled 48 AD-Y mini-libraries. Each mini-library was tested for possible auto-activators in a manner similar to the DB-X identification scheme using AD-Y specific reagents. No AD-Y auto-activators were found.

Y2H screen (related to fig. S2)

Y2H selection, phenotyping and sequencing

We systematically mated sets of 94 individual *MAT*α Y8930 DB-X yeast strains, in a 96-well format, with *MAT***a** Y8800 AD-Y mini-libraries on solid rich medium (YEPD). Each 96-well plate of DB-X yeast strains was used for mating with each of the 48 AD-Y mini-libraries. After overnight incubation at 30℃, yeast cells were transferred onto -His media to select for diploids that could grow under selective conditions, indicating activation of the *GAL1-HIS3* Y2H reporter gene (His$^+$ phenotype). In parallel, yeast cells were also transferred onto DB-auto-activator detection media (-His + 1mg/l cycloheximide (CHX), hereafter -HisCHX, growth on this media constitutes a HisCHX$^+$ phenotype). The pDEST-AD-*CYH2* vector carries the *CYH2* counter-selectable marker, which allows for plasmid shuffling on CHX containing media. This control step is essential to identify auto-activators that can spontaneously arise during the Y2H selection process ("spontaneous auto-activators"). Auto-activating DB-X yeast strains show a His^+ / $HisCHX^+$ phenotype, whereas genuine positives show a His^+ / $HisCHX^-$ phenotype. Since each DB-X yeast strain is mated against a mini-library of 192 AD-Y yeast strains, it is possible to obtain multiple interactions per mini-library. To account for this infrequent yet possible event we picked up to three colonies (primary positives) per growth spot. In total, ~68,000 primary positive colonies that exhibited a His^+/ $HisCHX^-$ phenotype were picked from -His plates into a second-generation set of 96-well plates. Using both Y2H reporter genes (*GAL1-HIS3* and *GAL2-ADE2*), the phenotype of primary positive colonies was retested for Y2H reporter activation and auto-activation. Of those, 42,000 activated at least one reporter gene in a CHX sensitive manner, hence passing the second step of phenotypic characterization (secondary positives), and were further processed.

Yeast PCR and interaction sequence tags sequencing

The identity of secondary positives pairs was determined by end-read sequencing of PCR products amplified directly from yeast colonies. PCR amplicons were purified and used as templates in cycle-sequencing reactions to obtain two (DB-X and AD-Y) interaction sequence tags (ISTs) per secondary positive colony.

IST analysis

The quality of ISTs obtained by sequencing was evaluated by moving a sliding window of 20 nucleotides to define portions of ISTs with an average PHRED score greater than or equal to 30 over at least 10% of their lengths. The sequences were aligned against the *Arabidopsis* ORFeome resource, and unique IST pairs with a BLASTN *E*-value less than or equal to 10^{-15} were retained. When an IST could not be unambiguously assigned to a single ORF because multiple ORFs represented the same locus, we provisionally assigned the IST to all possible ORF matches, ultimately only keeping those passing the verification step. Note that ORFs from different loci, even with highly similar sequences could always be distinguished and that these ambiguities only occurred between clones representing the same TAIR7 locus.

The entire Y2H screen step (Y2H selection, phenotyping and sequencing, yeast PCR and IST sequencing and IST analysis) was completed twice on space 1 to build

AI-1_{MAIN}, yielding a total of 11,293 unique IST pairs representing 11,716 candidate Y2H pairs.

Y2H verification (related to fig. S2)

The phenotype of all candidate Y2H interaction pairs was experimentally verified to ensure reproducibility and to exclude the possibility that physiologic and genetic changes occurring during the course of the experiment might have given rise to experimental artifacts. In total, we verified the Y2H phenotype of 11,716 candidate Y2H pairs by mating on YEPD media the matching individual *MAT*α Y8930 DB-X yeast strains and *MAT***a** Y8800 AD-Y yeast strains. We selected diploid cells on solid SC-Leu-Trp selective media and tested them for activation of *GAL1-HIS3* and *GAL2-ADE2* reporter genes. To control for technical variability and to increase the reproducibility of the dataset, all pairs were processed through these steps four times independently by four separate experimenters. Only pairs that gave rise to a His^+ growth phenotype in three out of four replicates and a $HisCHX^-$ growth phenotype four out of four times were considered verified. Of the 11,716 pairs, 6,871 (59%) scored positive according to these criteria (fig. S2C). Of these, 82% of interactions scored positive in 4/4 replicates and are thus 100% reproducible. Furthermore, 15% of the remaining pairs (3% of the total) correspond to experimental failures where no phenotype was measured in one of the trials (failed mating). The phenotype of remaining 15% interactions in the dataset were reproduced in 3 of 4 replicates and are thus 75% reproducible. All interactions that were reproduced in 3 or 4 replicates were considered verified Y2H interactions and included in AI-1. Altogether, the overall reproducibility of interactions in AI-1 is thus: [(0.85*1) + (0.15*0.75)] = 96%.

Y2H confirmation (related to fig. S2)

In the last step, the phenotypes of each verified Y2H interaction pair were tested once more on selective –His and –HisCHX plates. We also assessed each individual hybrid construct (DB-X and AD-Y) separately for possible spontaneous auto-activation. Only those pairs whose phenotype could be confirmed, and whose respective hybrid constructs were not auto-activators were retained for identity confirmation by end-read sequencing of DB-X and AD-Y PCR products amplified directly from yeast cells. In total, of the 6,871 verified Y2H pairs, the phenotype and identities of 5,903 (86%) were confirmed. After collapsing individual ORFs to unique genomic loci, we obtained the 5,664 Y2H interactions that comprise the dataset AI-1_{MAIN}.

Y2H repeat screen experiment

The repeat screen experiment aims at experimentally determining the sampling sensitivity of the two Y2H selection step iterations, since this is where the greatest effect of incomplete sampling is introduced. The Y2H selection step was repeated four independent times on a subset of Y2H constructs, defined as the repeat subspace (fig. S2B, table S6). The set of IST pairs obtained from these four experiments was merged with the set of IST pairs from the two experiments previously described, prior to the verification and confirmation steps, and all pairs were processed together. The resulting dataset, representing six independent trials on the same subspace and named AI-1_{REPEAT}, contains 1,066 Y2H interactions including 525 common with AI-1_{MAIN}. The union of interactions from AI-1_{MAIN} and AI-1_{REPEAT} yielded AI-1.

SOM III: PROTEIN-PROTEIN INTERACTION VALIDATION BY WELL NUCLEIC ACID PROGRAMMABLE PROTEIN ARRAY (wNAPPA)

Protein-protein interaction reference sets (related to fig. S3)

As of April 2010, there were 4,707 literature-curated protein-protein interactions of *Arabidopsis* compiled in two public databases, TAIR (*46*) and IntAct (*48*). After filtering for interactions described in ≥ 2 publications or by ≥ 2 methods (1,054) and eliminating those interactions involving ORFs for which no reagents are available in our *Arabidopsis* ORFeome resource (540), we randomly picked 200 interactions and manually re-curated the corresponding 276 publications using established criteria to ensure maximum quality (fig. S3) (*5*). The 118 well-documented binary interactions (pairs of proteins experimentally demonstrated to physically interact) fulfilling our curation criteria constituted the *Arabidopsis thaliana* positive reference set, version 1 (AtPRS-v1, referred to as PRS in the main text for simplicity) (fig. S3). We also picked 146 protein pairs at random from the ~3.6 X 10^7 possible pairwise combinations of available ORFs to assemble the *Arabidopsis thaliana* random reference set version 1 (AtRRS-v1, referred to as RRS in the main text for simplicity) (fig. S3).

wNAPPA implementation

We determined the precision of AI-1_{MAIN} by testing a randomly chosen sample of 249 AI-1_{MAIN} interactions plus all AtPRS-v1 and AtRRS-v1 pairs in the "well-based Nucleic Acid Programmable Protein Array" assay (wNAPPA) (*8, 49*). To do so, the corresponding *Arabidopsis* ORFs were transferred by Gateway LR recombinational cloning (Invitrogen) into both pIX-GST::*ccdB* and pIX-3xHA::*ccdB* destination vectors. Competent bacteria (*E. coli*, strain DH5α-T1^R) were transformed with the resulting LR recombination products. After selection of transformants in liquid terrific broth medium containing 50 µg/ml carbenicillin, plasmid DNA was extracted and purified using Qiaprep 96 Turbo kits (Qiagen). DNA concentrations were measured using an 8-channel nanodrop and normalized to 125 ng/µl. On each assay plate, we included a set of 22 pairs of clones (normalization reference set, NRS, table S11) to control for experimental plate-to-plate variation. Each X-Y pair was tested in both wNAPPA configurations: GST-X, HA-Y and GST-Y, HA-X.

Bait and prey proteins were co-expressed using the TNT SP6 Coupled Wheat Germ Extract System (Promega) according to the recommendations of the manufacturer. Expressed proteins in wheat germ extract were added to glutathione-S-transferase (GST) detection plates (GE Healthcare), and incubated at 15℃ for 2 hours. Subsequently, wells were washed and blocked in 1X phosphate buffered saline with tween, containing 5% non-fat dry milk ("blocking buffer") and subsequently incubated with mouse anti-HA monoclonal antibody [HA.11 clone16B12, 1:5000 (Covance) in blocking buffer] for 1 hour at room temperature. After further washes using blocking buffer, wells were incubated with anti-mouse HRP-coupled secondary antibody (1:2000 in blocking buffer; GE Healthcare) for 1 hour at room temperature. Wells were then washed in 1X phosphate buffered saline before adding Supersignal ELISA Femto substrate (Pierce). Luminescence (RLU) was detected using a Gemini SpectraMax plate reader.

wNAPPA scoring

On each assay plate we included 22 normalization reference set pairs (NRS, table S11), which we used to model the null (non-interacting) case for pairs. This set of pairs was chosen at random from the set of AtRRS-v1 pairs. The wNAPPA signals from each protein pair on each plate were standardized using these normalization pairs as follows:

1. We truncated any negative luminescence values to a raw value of 1.
2. Following truncation, we log-transformed (base-2) the intensity values to remove the dependence of the intensity variance to the intensity average (fig. S4A).
3. The truncated-and-log-transformed NRS intensities were then used to find estimates of the NRS mean, μ_{NRS} and standard deviation σ_{NRS}, on each plate.

The intensities of all remaining protein pair on each plate were then transformed to z-scores (number of σ_{NRS} units away from μ_{NRS}). We normalized the scores relative to the

NRS. For each pair, the maximum z-score of the two configurations was considered and used to determine recall rates.

GST-only background control of wNAPPA assay

A set of controls was designed to assess the relative level of background signal obtained by co-expression of individual 3xHA-tagged proteins and the empty GST vector relative to the NRS threshold. A total of 72 3xHA-tagged proteins (36 from AtPRS-v1 and 36 from AtRRS-v1) were each co-expressed with GST protein using an equal concentration of pIX-GST empty vector ("GST-only" plate). The 22 NRS pairs were run alongside as usual (co-expressed pIX-GST-X fusion and pIX-3xHA-Y fusion). Two positive controls indicated normal functioning of the assay. All wNAPPA assay conditions were identical to those used for non-control plates.

The distributions of signal from the 3x-HA-AtPRS-v1 and -AtRRS-v1 against GST only pairs were significantly below that of the NRS pairs on the "GST-only" plate (*P* = 5.8 X 10^{-9}, *P* = 1.9 X 10^{-9}, respectively, one-sided KS-tests), indicating the absence of background signal due to non-specific binding to the GST tag (fig. S4B). We also verified that the signal obtained from NRS controls on GST-only plates gave similar results to those on wNAPPA AtPRS/RRS-v1 plates (fig. S4B). The signals of the NRS pairs showed no evidence of differing distributions (*P* = 0.87, two-sided KS-test).

SOM IV: ESTIMATION AND IMPLICATIONS OF THE FRAMEWORK PARAMETERS

Completeness of AI-1$_{MAIN}$ screening space

There are 27,029 predicted protein-coding genes in the TAIR7 version of the *Arabidopsis thaliana* genome annotation (*46*), and thus 365,283,420 possible protein pairs. In our experiment, we tested 8,583 constructs corresponding to 8,429 gene loci present in space 1 (fig. S2, table S3). This collection contains 8,033 AD-hybrid constructs corresponding to 7,895 loci, and 7,760 DB-hybrid constructs corresponding to 7,645 gene loci. Among these, 7,210 constructs corresponding to 7,110 gene loci were tested as both AD- and as DB-hybrid proteins (25,276,050 protein pairs). In addition, there were 823 constructs corresponding to 785 gene loci tested only as AD-X hybrid constructs against all DB-Y hybrid constructs (6,001,325 protein pairs). Lastly, there were 550 constructs corresponding to 534 loci tested only as DB-X hybrid constructs against all AD-Y hybrid constructs (4,215,930 protein pairs). In total, 35,500,949 unique protein pairs were tested in our high-throughput Y2H screens. Therefore, the completeness of our screen is 35,500,415 / 365,283,420 or 9.7%.

Estimation of Y2H assay sensitivity (related to Fig. 1B)

We estimated Y2H assay sensitivity by a pairwise Y2H test of AtPRS-v1 pairs (SOM II and fig. S3). Together with the pairs detected in the main and repeat screens, a total of 43/118 (36.4%) AtPRS-v1 pairs passed the scoring criteria of our Y2H pipeline (table S5). The assay sensitivity of our implementation of Y2H is therefore 36.4% ± 4.4% (standard error of the proportion) (Fig. 1B). This may be an overestimation due to biases in the literature from which AtPRS-v1 pairs were taken. AtPRS-v1 contains many interactions originally detected by Y2H, which may be more easily detected by our implementation of this assay than a perfectly representative sampling of interactions in the *Arabidopsis* interactome.

Experimental estimation of the precision of AI-1$_{MAIN}$ (related to Fig. 1B)

AtPRS-v1, AtRRS-v1, and a sample of the AI-1$_{MAIN}$ interactions (AI-1$_{MAIN}$ sample) were tested and scored by wNAPPA assay (protocol in SOM III). The recall rate of each of these protein-pair sets at all possible z-score thresholds was determined (fig. S4, C and E). The recall rate of AtPRS-v1 pairs was statistically indistinguishable from that of pairs from the AI-1_{MAIN} sample.

AI-1_{MAIN} precision was estimated for different wNAPPA z-score thresholds (*t*) as before (fig. S4, D and E *6*):

prec(*t*) = [recall(AI-1_{MAIN}sample,*t*) - recall(AtRRS-v1,*t*)) / (recall(AtPRS-v1,*t*) - recall(AtRRS-v1,*t*)].

We computed the precision for z-score thresholds between 1.4 and 1.6 (fig. S4E) to maximize both sensitivity and specificity of wNAPPA (high recall rate of AtPRS-v1 pairs and low recall rate of AtRRS-v1 pairs). Our final estimate for the precision of AI-1_{MAIN} was computed as the mean precision within this range, equal to 80.3% ± 8.7% (standard error of the mean) (Fig. 1B).

Estimation of the sampling sensitivity of AI-1_{MAIN} (related to fig. S5)

In addition to the two screens on space 1, there were four additional screens on a subspace of space 1 (fig. S2B, table S6). Although the screens were done in a certain order, each screen could be considered the first screen (or second, third, etc.). To avoid the particular experimental order chosen from contributing to discontinuities in the accumulation of interactions, we simulated results for all possible (6! or 720) orderings for the six screens. We calculated the average number and standard deviation of interactions detected at each step, considering all possible orders, both for verifiable and confirmed AtPRS-v1 IST pairs and for the total number of verifiable and confirmed IST pairs uncovered. For the last data point the standard deviation is zero because the total number of interactions uncovered after the sixth screen is constant. Even after six screens, saturation was not reached (fig. S5A). However, the information from these screens allowed us to build a model to estimate the fraction of interactions detectable by Y2H captured after any number of screens.

We modified the variables within the Michaelis-Menten equation to model how many interactions we would find at saturation and how close we were to this number after any number of repeat screens. Just as maximum reaction velocity is approached asymptotically with increasing substrate concentrations, we approach screen saturation with increasing numbers of experimental iterations. We can thus write:

$$N_i(R) = \frac{N_{iMAX}.R}{K_M + R}$$

where N_i is the number of interactions detected by our assay after R repeats, N_{iMax} is the number of interactions detected at screen saturation, and K_M is the Michaelis constant. We determined the parameters for each of the 720 repeat screen permutations using nonlinear weighted least squares estimates of N_{iMax} and K_M within our modified Michaelis-Menten function. Using the N_{iMax} and K_M estimates for individual repeat screen permutations, we predicted N_i for larger numbers of repeats (fig. S5B).

Within the repeat screen space, we estimated saturation to occur at 1,719 ± 309 interactions (mean ± standard deviation), therefore the 1,066 interactions detected after six repeats represent 61.9% of possible interactions. The two repeats in AI-1_{MAIN} therefore yielded 36.5% ± 7.5% (mean ± standard deviation) of the total number of Y2H detectable interactions in space 1. For AtPRS-v1 pairs, using the same reasoning, we estimated that two repeats yielded 35.7% ± 7% of the total number of detectable AtPRS-v1 pairs.

Estimation of the overall sensitivity of AI-1_{MAIN}

The overall sensitivity of AI-1_{MAIN} is the product of assay and sampling sensitivity. With an assay sensitivity of 36.4% ± 4.4% and a sampling sensitivity of 36.5% ± 7.5%, the overall sensitivity of AI-1_{MAIN} is estimated to be 13.3% ± 3.2% (mean ± standard deviation, assuming independence between assay and sampling sensitivities). We also calculated the overall sensitivity of AI-1_{MAIN} following the observation that on average 5 of the 31 AtPRS-v1 pairs were detected after two screens in our repeat experiment (fig. S5A), corresponding to an overall sensitivity of 15.7% ± 3.8% (mean ± standard deviation).

Estimation of the size of the *Arabidopsis* binary interactome

With the calculation of the precision *P*, overall sensitivity *S*, and completeness *c*, we estimated the size of the *Arabidopsis* binary interactome according to the equation:

$$size(n,c,P,S) = \frac{nP}{cS}$$

where *n* is the number of interactions in AI-1$_{\mathrm{MAIN}}$. *P* and *S* are variables that we estimated above, and n and c are observables. Our estimate of the size variance is derived using the Delta Method (with a 1st-order Taylor expansion, assuming independence between precision and sensitivity):

$$\mathrm{Var}[\mathrm{size}(n,c,\mu_P,\mu_S)] = [\frac{\partial}{\partial P} size(n,c,\mu_P,\mu_S)]^2.Var[P] + [\frac{\partial}{\partial S} size(n,c,\mu_P,\mu_S)]^2.Var[S]$$

$$= (\frac{n}{c.\mu_S})^2.Var[P] + (\frac{-n.\mu_P}{c.\mu_S^2})^2.Var[S]$$

where μ_P, *Var[P]*, μ_S, and *Var[S]* are means and variances of *P* and *S*, respectively. The following values were used:

$c = 0.097$
$n = 5664$
$\mu_P = 0.803$
Var[P] = 0.087^2
$\mu_S = 0.157$
Var[S] = 0.038^2

This calculation results in an interactome size estimate of 298,653 ± 79,197 binary protein-protein interactions (mean ± standard deviation). Accounting only for a symmetric search space where every clone is present both as AD hybrid construct and DB hybrid construct, and calculating the number of interactions, completeness, sampling sensitivity and assay sensitivity accordingly, this interactome size estimate is of 473,023 ± 123,387 binary protein-protein interactions (mean ± standard deviation).

SOM V: INTEGRATION OF AI-1 WITH EXTERNAL DATASETS

Integration of AI-1 and $\mathrm{LCI}_{\mathrm{BINARY}}$

To build a network of literature-curated binary interactions for *Arabidopsis*, we downloaded and parsed interaction information from IntAct (*48*), TAIR (*46*) and Biogrid (*50*) in April 2010. A total of 5,270 unique protein pairs were reported to interact either directly, or indirectly in a protein complex (LCI; fig. S1A; table S4). This set includes 4,252 “binary” (direct) interactions ($\mathrm{LCI}_{\mathrm{BINARY}}$; fig. S1B; tables S1, S4). The integration of $\mathrm{LCI}_{\mathrm{BINARY}}$ and AI-1 yielded a network of 10,362 interactions between 4,439 proteins. AI-1_{MAIN} detected 64 of those present in $\mathrm{LCI}_{\mathrm{BINARY}}$ single evidence, and 15 present in $\mathrm{LCI}_{\mathrm{BINARY}}$ multiple evidence (Fig. 1B).

Classification of plant-specific proteins

To identify plant-specific proteins, we identified homologs using BLASTP with an *E*-value cutoff of 0.001 for all proteins in AI-1 within the following eukaryotic lineages (*51*): fungi, animals, choanoflagellida, amoebazoans, rhodophyta, alveolata, heterokonts, haptophycae, discicristates, excavates (genomes listed in table S12). We considered proteins that did not have homologs in these lineages to be plant-specific proteins. Most of them have homologs in other plants and others are *Arabidopsis thaliana* specific proteins.

Correlation of gene expression (related to fig. S7)

A compendium of normalized Affymetrix data from 1,436 array experiments was downloaded from TAIR (*46*) [folder affy_data_1436_10132005, (*52, 53*)] and filtered to keep only the probes unambiguously mapped to single TAIR7 gene loci. For each gene pair, the pairwise Pearson Correlation Coefficient (PCC) was calculated using R (http://www.rproject.org, (*54*)) over the entire compendium (fig. S7A).

Gene Ontology dataset and functional enrichment analyses (related to fig. S7)

Gene Ontology (GO) functional annotations and the definition of the GO Directed Acyclic Graph (DAG) were obtained directly from the Gene Ontology Annotation database (March 9, 2010) (*55*). Annotations were mapped to the subset of TAIR7 genomic loci of *Arabidopsis thaliana* that have remained stable from TAIR7 to TAIR9, *i.e.* genomic loci that have been present and neither merged nor split since TAIR7. These attributes were "up-propagated": if a locus was annotated as having a given attribute, the locus was also associated with all the less specific attributes implied by this attribute in the GO DAG.

The enrichment of interacting pairs in shared GO annotations is expressed as an odds ratio relative to all unordered pairs of proteins in AI-1_{MAIN} that were not found to be interacting in AI-1_{MAIN} and that have one or more GO annotations for each given functional specificity threshold (fig. S7B). The enrichment in shared precise GO annotations of pairs sharing more than 0%, 25%, 50% of interacting partners excludes directly interacting pairs, and is an odds ratio relative to non-interacting pairs sharing a smaller percentage of interacting partners (fig. S7D). All enrichments in fig. S7, B-D are statistically significant with *p*-values < 0.05. Significance and estimates of enrichment were calculated using Fisher's exact test. Part of the signal seen in fig. S7D can be attributed to the high number of shared interactors between paralogous proteins (Fig. 4). Enrichments remained significant in most categories when paralogous pairs were removed from this analysis (data not shown). Functional enrichment analyses relating to Fig. 3 are described in SOM VI.

Putative phosphorylation signaling subnetwork (related to fig. S8)

In AI-1, 220 proteins were found to interact with a protein kinase or phosphatase, doubling the number of known interactions within the search space for this class of signaling proteins. Such interactors likely include putative substrates as well as scaffolding partners or other regulators. To help differentiate between these potential functions, the 38 interactions involving a protein for which phosphorylation has been experimentally demonstrated represent attractive substrate candidates. The recovery of the known MKK4-MPK6 interaction indicates that this approach can indeed identify genuine kinase-substrate pairs (*56*) and provides a compelling starting point for the deeper functional characterization of the involved proteins (fig. S8A).

Novel aspects of hormone signaling (related to fig. S8C)

Jasmonic acid (JA) serves as a primary signal in the regulation of plant defense and reproductive development (*57*). Current understanding of the JA signaling pathway is limited to physical interactions between the hormone receptor COI1, twelve jasmonate ZIM-domain (JAZ) transcriptional repressors, and the transcription factor MYC2. Given the involvement of JA in a wide range of biological processes, it has been hypothesized that other transcription factors are regulated through their interaction with JAZ proteins (*58*). In agreement with this hypothesis, we found seven transcription factors binding to JAZ proteins (fig. S8C). Moreover, we identified three JAZ-related proteins (*59*) that interact with transcription factors, two of which also bind to a JAZ protein, likely pointing to a complex transcriptional regulation module (fig. S8C).

Overlap with predicted protein-protein interaction datasets (related to table S9)

Table S9 describes protein-protein interactions from AI-1 that were computationally predicted in (*60-63*).

Pfam domain assignments (related to tables S8, S9)

Pfam domains were assigned to proteins with the HMMER2 program utilizing Pfam domain family HMM profiles (*64*). We used a *p*-value cut-off of 0.001 for considering domain assignments as valid to build domain architectures. We constructed domain architectures using custom written PERL scripts and we were able to build unique domain architecture for 2,196 proteins in the network.

Most frequent domains (related to table S9)

Based on protein domain architectures, we identified the top 10 most frequent domains, in terms of number of interactions, for the proteins in the network. We only considered the number of unique domains in a protein and did not factor in multiple copies of a particular domain in a given protein. The most frequent domains were involved either in nucleic acid binding or signaling.

Ubiquitin pathway protein domain assignments (related to Fig. 2A and tables S8, S9)

In addition to information available in the literature (*12*), we added domain descriptions for E1 enzymes (UBA and UBACT domains), E2 enzymes (RWD, UEV and UQ_CON domains), E3 enzymes (zf-MIZ, zf-C3HC4, U-box and HECT domains) and de-ubiquitinating enzymes (DUBs - UCH, PPDE and OUT). To do so, we used the following procedure:

1. PFAM seed alignments of these domain families were realigned using PCMA sequence alignment program (*65*) and checked for accuracy manually.

2. The reliable PCMA generated alignments were used as a starting point for PSI-BLAST (*47*) searches to assign these domains to proteins in AI-1.

3. The domain assignments were checked using the HHpred program for consistency (http://toolkit.tuebingen.mpg.de/hhpred) (*66*).

Homology and orthology assignments (related to table S9)

We used BLASTP to assign homologs of proteins in the network (*47*). We deemed a protein from another species' proteome to be a homolog of a queried *Arabidopsis thaliana* protein if the protein was the best BLAST hit with an *E*-value ≤ 0.001 (species list in table S12). Similarly, we assigned orthologs if two proteins were reciprocal best hits at an *E*-value threshold of 0.001.

Proteome sequences (related to table S9)

Predicted proteome sequences of organisms were downloaded from www.supfam.org (May 2009, version 1.73) (*51*). Tomato and Maize proteome sequences were downloaded respectively from http://solgenomics.net/genomes/Solanum_lycopersicum/genome_data.pl (August 2010, version 2.30) and ftp://ftp.plantgdb.org/download/MaizeData/ (August 2010, version B73 RefGen_v2).

Co-evolving pairs: mutual information index (related to table S9)

We estimated significantly coevolving pairs of proteins in AI-1 by:

(1) Identification of orthologs, if any, of interacting proteins across completely sequenced plant genomes (18 species: *Chlamydomonas reinhardtii, Volvox carteri f. nagariensis, Chlorella sp. NC64A, Chlorella vulgaris, Ostreococcus tauri, Ostreococcus sp. RCC809, Ostreococcus lucimarinus CCE9901, Micromonas sp. RCC299, Micromonas sp. CCMP490, Physcomitrella patens subsp. Patens, Selaginella moellendorffii, Medicago truncatula, Populus trichocarpa, Vitis vinifera, Oryza sativa ssp. Indica, Oryza sativa ssp. Japonica, Sorghum bicolor, Carica papaya*) obtained from www.supfam.org (May 2009, version 1.73) (*51*).

(2) Based on identified orthologs, we constructed a phylogenetic profile:

Given interaction	P1	P2	P3	P4	P5	P6	P7	P8	P_i
Protein A	1	0	1	0	1	0	1	1	
Protein B	0	1	1	0	1	0	1	0	

Where Pi denotes proteome of species i, i = 1 to 18.

(3) Based on the constructed phylogenetic table, we calculated the mutual information index (*67*), defined as:

Mutual Information index = H(A)+H(B)-H(A,B)
where H(A) = summation(-Frequency of (index for protein A)*log(Frequency of index for protein A)) and H(A,B) = summation(-Frequency of (index for proteins A and B)*log(Frequency of index for proteins A and B)).

(4) To empirically estimate significance, we adopted the following procedure:

a) Generation of random phylogenetic profiles for each interacting protein. This was done by preserving the conservation profile of proteins in AI-1, *i.e.* the total number of species a given protein was originally identified to be present.

b) Calculation of mutual information index for every interacting pairs in AI-1 using these random phylogenetic profiles.

c) This procedure was repeated 1,000 times. Average and standard deviation of number of pairs within a given range of mutual information index was calculated. The distribution was plotted and compared against original mutual information index distribution.

Experimentally validated interologs outside of the plant kingdom (related to table S9)

We identified putative evolutionary conservation of interactions ("interologs"; *68, 69*) in the crown group eukaryotes (animals, plants and fungi) by using AI-1. Here, interologs are defined as interacting protein pairs in AI-1 that have interacting homologs in yeast or humans. This process involved:

(1) Identifying yeast and human homologs of AI-1 proteins.

(2) Identifying conserved interactions among homologs in yeast or human. Interaction data for yeast and human was obtained from IntAct (*48*) and from (*6, 9, 10*).

We found 239 interologs in yeast or human, of which 29 are present in all three species. Because they are present in species belonging to three kingdoms, it is possible these interactions existed in the common ancestor of the crown group eukaryotes before the plant lineage diverged from the fungal and animal lineages. However, because homology relationships do not correspond to one-to-one mapping (in opposition to orthology mapping) and because the interaction datasets used are not generated in a systematic and controlled fashion, statistical analyses of these interactions are hindered. No conserved interolog interaction between AI-1_{MAIN} and systematic and controlled interactome datasets for yeast and humans (YI-1, *9*; HI-1, *10*) was detected when using strict orthology relationships.

Read-me file for Supporting Online Material table S8: Protein annotations

Protein attributes assigned to 26,850 *Arabidopsis* protein-coding genes listed in table S8, as of March 2010 unless otherwise indicated below.

Column A: TAIR_locus_ID

Values: AGI names for loci corresponding to each gene/protein.

Description: Unique identifiers for each protein in the network (which were used as keys for all protein annotations). Only AGI names that have not disappeared or been merged or split between TAIR7 and TAIR9 are included in this table.

Source: http://www.arabidopsis.org (TAIR9) (*46*).

Column B: Mazzucotelli_Es

Values: “E1”, “E2”, “E3”.

Description: Annotations according to literature descriptions of ubiquitin activating (E1) enzymes, ubiquitin conjugating (E2) enzymes and ubiquitin ligase (E3) enzymes.

Source: (*12*).

Column C: Mazzucotelli_E3domain

Values: “PUB”, “F-Box”, “RING”, “RING and CUL4-DDB6”, “BTB”, “Cullin”, “APC complex”, “ASK”, “HECT”, “CUL4-DDB1”, “Cullin and CUL4-DDB7”, “CUL4-DDB3”, “Cullin and APC complex”, “RING and F-Box”, “RING and APC complex”, “CUL4-DDB2”, “CUL4-DDB4”, “CUL4-DDB5”.

Description: Domain or complex names assigned based on literature description of E3 enzymes.

Sources: (*12*).

Column D: E_by_Domains

Values: “E1domains”, “E2domains”, “E3domains”.

Description: Ubiquitin activating E1 enzyme, ubiquitin conjugating E2 enzyme and ubiquitin ligase E3 enzyme domain assignments based on sequence alignments.

Source: This manuscript (SOM V).

Column E: Ubiquitinated

Values: “Ubiquitinated”.

Description: Proteins determined experimentally to be ubiquitinated. List compiled from four publications.

Sources: (*70-73*)

Column F: Deubiquitinating_enzyme

Values: “DUB”.

Description: De-ubiquitinating domain assignments based on sequence alignments.

Source: This manuscript (SOM V).

Column G: Ubiquitin

Values: “Ubiquitin”.

Description: Ubiquitin domain assignments based on sequence alignments.

Source: This manuscript (SOM V).

Column H: Plant_specific

Values: “plant_specific”.

Description: Genes defined as plant-specific, absent from other eukaryotic lineages.

Source: This manuscript (SOM V).

Column I: Kinase

Values: “kinase”.

Description: Kinase enzymatic activities predicted and assigned to proteins encoded by the indicated loci.

Source: http://plantsp.genomics.purdue.edu/plantsp/html/families.html (August 2009 download) (*74*).

Column J: Phosphatase

Values: "phosphatase".

Description: Phosphatase enzymatic activities predicted and assigned to proteins encoded by the indicated loci.

Source: http://plantsp.genomics.purdue.edu/plantsp/html/families.html (August 2009 download) (*74*).

Column K: Phosphorylated

Values: "phosphorylated".

Description: Proteins experimentally shown to be phosphorylated.

Source: http://phosphat.mpimp-golm.mpg.de/ (PhosphAt database, August 2009 download) (*75*) and original sources: (*76-90*).

Column L: Abscisic acid

Values: "0" or "1".

Description: "1" denotes loci encoding proteins annotated with involvement in abscisic acid biosynthesis, signaling or response in The *Arabidopsis* Hormone Database. "0" indicates no such annotations exist.

Source: http://ahd.cbi.pku.edu.cn/ (downloaded on January 9, 2010) (*91*).

Column M: Auxin

Values: "0" or "1".

Description: "1" denotes loci encoding proteins annotated with involvement in auxin biosynthesis, signaling or response in The *Arabidopsis* Hormone Database. "0" indicates no such annotations exist.

Source: http://ahd.cbi.pku.edu.cn/ (downloaded on January 9, 2010) (*91*).

Column N: Brassinosteroid

Values: “0” or “1”.

Description: “1” denotes loci encoding proteins annotated with involvement in brassinosteroid biosynthesis, signaling or response in The *Arabidopsis* Hormone Database. “0” indicates no such annotations exist.

Source: http://ahd.cbi.pku.edu.cn/ (downloaded on January 9, 2010) (*91*).

Column O: Cytokinin

Values: “0” or “1”.

Description: “1” denotes loci encoding proteins annotated with involvement in cytokinin biosynthesis, signaling or response in The *Arabidopsis* Hormone Database. “0” indicates no such annotations exist.

Source: http://ahd.cbi.pku.edu.cn/ (downloaded on January 9, 2010) (*91*).

Column P: Ethylene

Values: “0” or “1”.

Description: “1” denotes loci encoding proteins annotated with involvement in ethylene biosynthesis, signaling or response in The *Arabidopsis* Hormone Database. “0” indicates no such annotations exist.

Source: http://ahd.cbi.pku.edu.cn/ (downloaded on January 9, 2010) (*91*).

Column Q: Gibberellin

Values: “0” or “1”.

Description: “1” denotes loci encoding proteins annotated with involvement in gibberellin biosynthesis, signaling or response in The *Arabidopsis* Hormone Database. “0” indicates no such annotations exist.

Source: http://ahd.cbi.pku.edu.cn/ (downloaded on January 9, 2010) (*91*).

Column R: Jasmonic Acid

Values: “0” or “1”.

Description: “1” denotes loci encoding proteins annotated with involvement in jasmonic acid biosynthesis, signaling or response in The *Arabidopsis* Hormone Database. “0” indicates no such annotations exist.

Source: http://ahd.cbi.pku.edu.cn/ (downloaded on January 9, 2010) (*91*).

Column S: Salicylic acid

Values: “0” or “1”.

Description: “1” denotes loci encoding proteins annotated with involvement in salicylic acid biosynthesis, signaling or response in The *Arabidopsis* Hormone Database. “0” indicates no such annotations exist.

Source: http://ahd.cbi.pku.edu.cn/ (downloaded on January 9, 2010) (*91*).

Column T: Description

Values: Text.

Description: Descriptions of gene/protein function from literature, when available.

Source: http://www.arabidopsis.org/ (TAIR9) (*46*).

Column U: Alias

Values: Text.

Description: Concatenated text listings of available aliases for each locus.

Source: http://www.arabidopsis.org/ (TAIR9) (*46*).

Column V: Transporter

Values: “0” or “1”.

Description: “1” indicates a predicted transporter function. “0” indicates that no such function was found in the study.

Source: (*92*).

Column W: Stress

Values: “0” or “1”.
Description: “1” indicates the GO-slim annotation “response to stress” is assigned to this locus. “0” indicates otherwise.
Source: http://www.geneontology.org/ (March 9, 2010) (*55*).

Column X: Abiotic_and_biotic_stimulus
Values: “0” or “1”.
Description: “1” indicates the GO-slim annotation “response to abiotic or biotic stimulus” is assigned to this locus. “0” indicates otherwise.
Source: http://www.geneontology.org/ (March 9, 2010) (*55*).

Column Y: Metabolic_enzyme
Values: “0” or “1”.
Description: “1” indicates that the encoded proteins are known or predicted metabolic enzymes according to AraCyc. “0” indicates otherwise.
Source: http://www.arabidopsis.org/biocyc/downloads.jsp (AraCyc 6.0; March 25, 2010) (*93*).

Column Z: Transcription_factor
Values: “0” or “1”.
Description: “1” indicates known or predicted transcription factors. “0” indicates otherwise.
Source: http://arabidopsis.med.ohio-state.edu/ (Jan 19, 2010 version – downloaded on March 25, 2010) (*94*).

Column AA: Membrane_domain
Values: “membrane_domain”.
Description: “Membrane_domain” indicates prediction of at least one transmembrane domain.
Source: http://www.arabidopsis.org (TAIR9) (*46*).

SOM VI: COMMUNITIES IN AI-1$_{\text{MAIN}}$

The inclusive neighbors of node (protein) *i*, $n_+(i)$, is defined as the set of node *i* and its neighbors. For a link pair e_{ik} and e_{jk} that share a node *k*, the similarity is defined as the Jaccard coefficient of inclusive neighbors:

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}$$

Based on this similarity measure, we built a dendrogram using single-linkage hierarchical clustering. We obtained link communities by cutting the dendrogram at a certain threshold. To determine relevant thresholds, we used measures of partition density. The partition density of a community *c* with n_c nodes and m_c links is defined by:

$$D_c = \frac{m_c - (n_c - 1)}{n_c(n_c - 1)/2 - (n_c - 1)}$$

or: (actual number of links - minimum possible number of links) divided by (maximum possible number of links - minimum possible number of links).

The partition density of communities used to determine relevant thresholds is the weighted sum of partition density for each community. The weight is the number of links in each community:

$$D = \frac{1}{M}\sum_c m_c D_c$$

where *M* is the number of links in the network. When applied to AI-1$_{\text{MAIN}}$, this methodology yielded 2,453 communities of at least one link at the optimal partition density (fig. S9A).

Selection and GO analyses of relevant communities

We selected relevant communities among the 2,453 detected by the algorithm. Out of 108 communities containing more than five nodes (proteins), 26 have a partition density greater than zero (internally well-connected communities). We evaluated whether these internally connected communities appeared to be more biologically relevant than other communities with a zero partition density. To this end, GO annotation enrichments (SOM V) were calculated for each of the 108 communities using the FuncAssociate R library (*95*), with a false discovery rate cutoff of 10%, using proteins in AI-1_{MAIN} as a reference set and requiring that a minimum of 10% of nodes in the tested community are annotated with the enriched GO annotation. According to these criteria, the 26 communities with partition density >0 were significantly more enriched in GO annotations than the remaining 82 (fig. S9B). We therefore selected these 26 network communities as putative biologically interesting sets of proteins of related function (table S10 and figs. S10-35).

Comparison of community relevance between AI-1_{MAIN} and random networks

AI-1_{MAIN} was randomized 100 times by degree-preserving edge shuffling. The link clustering methodology was applied to each of the resulting 100 random networks, and communities containing more than five nodes and with a link density > 0 were tested for GO annotation enrichment with the same criteria as applied to AI-1_{MAIN} (Fig. 3). Whereas AI-1_{MAIN} has 26 communities of which 90% are enriched in at least one GO term, randomized networks contained a maximum of 25 communities, of which the proportion of GO-enriched communities never surpassed 25% (empirical $P < 0.01$).

SOM VII: EVOLUTION OF PROTEIN-PROTEIN INTERACTIONS FOLLOWING GENE DUPLICATION

Paralogy relationships and time-since-duplication

Paralogy relationships were downloaded from the Gramene website as of February 2009 (http://www.gramene.org; (*96*)). These paralogy relationships were predicted by the EnsemblCompara GeneTrees pipeline (*97*) with the following genomes: *Oryza sativa Japonica Group*, *Oryza sativa Indica Group, Oryza glaberrima*, *Oryza barthii*, *Oryza punctata*, *Oryza minuta*, *Oryza officinalis*, *Brachypodium distachyon*, *Sorghum bicolor*, *Zea mays*, *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Populus trichocarpa*, *Vitis vinifera*, *Homo sapiens*, *Ciona intestinalis*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*. The EnsemblCompara GeneTrees pipeline first performs pairwise and multiple alignments to identify clusters, *i.e.* families of genes in the genomes used, and build for each one a gene tree thus containing both orthologous and paralogous genes. These gene trees are then reconciled with the species tree to produce phylogeny-based gene trees. Orthology and paralogy relationships are inferred for every pair of genes in a cluster directly from the corresponding phylogeny-based gene tree.

For each pair of paralogous genes, the timing of the duplication event is estimated based on the phylogeny-based gene tree, due to the presence of extant species past the duplication, and thus implicitly outgroup lineages before the duplication. This method proved advantageous over molecular clock methods because it allows more reliable timing of the formation of ancient paralogs. Methods based on molecular clocks face the problem of saturation of synonymous sites in mutations, the number of synonymous substitutions per synonymous sites (Ks) reaching a plateau (fig. S39).

In the Gramene dataset, the timing of the duplication is provided in the form of a common ancestor, which refers to the root taxon where the duplication happened (*97*). Root taxa were determined based on the Entrez Taxonomy browser (*98*). Pairs present in AI-1$_{\mathrm{MAIN}}$ were grouped into four age categories related to their ancestry. Pairs with an

Arabidopsis ancestor were grouped in Age 1 (326 pairs in AI-1_{MAIN}), pairs with a Rosids or Core eudycotyledons ancestor were grouped in Age 2 (141 pairs in AI-1_{MAIN}), pairs with a Magnoliophyta ancestor were grouped in Age 3 (850 pairs in AI-1_{MAIN}), and pairs with a Eukaryota ancestor were grouped in Age 4 (565 pairs in AI-1_{MAIN}). The divergence time since duplication was then estimated by combining the timing of duplication (ancestors) together with estimates of divergence time between taxa spanning over ~700 million years obtained from TimeTree (*99*). The two most extreme scenarios, most recent estimates (Fig. 4C) and most ancient estimates (fig. S42) gave similar results. In conclusion, our dating methodology is much less affected by the saturation of synonymous sites in mutations over time than Ks-based methods and allows us to identify pairs of paralogous genes from particularly ancient duplication events (fig. S39).

Protein sequence identity, co-expression and functional data for paralogous pairs
Measures of protein sequence identity between the two paralogous proteins were extracted from the Gramene paralogy relationships dataset (averaged per pair). As seen in the distributions of protein sequence identity of paralogous proteins grouped by age (fig. S46), the paralog dataset can assign a young age to divergent paralogous proteins and an old age to similar paralogous proteins.

Correlation of paralogous gene expression was assessed using the expression data in SOM V (correlation of gene expression). Correlation significantly decreases as a function of the age group (fig. S44). Given the differences of the distributions of the different age groups, we analyzed the fraction of shared interactors of paralogous proteins as a function of age and gene co-expression by comparing the most co-expressed pairs (top tertile) to the least co-expressed pairs (bottom tertile, Fig. 4E).

When comparing our data with functional relationships from Hanada and colleagues (Fig. 4B, *28*), we only considered those pairs labeled as paralogous by both approaches.

Estimation of selective pressures
Strength of purifying selection among paralogous genes was measured using estimates of synonymous (Ks) and non-synonymous (Ka) substitutions within protein coding

regions. To obtain optimal alignments we first aligned at the amino acid level using CLUSTALW version 1.83 (*100*), then mapped positions to DNA coding sequence coordinates. Gaps in the alignment were removed. Ka and Ks were estimated with codeml of the PAML package version 3.15 (*101*) using the F3X4 codon frequency model. Paralogous pairs having very low homology failed to produce accurate alignments. To aid in screening these out, we conducted an all versus all NCBI BLASTP alignment of the *Arabidopsis* proteome. Paralogous pairs having an *E*-value greater than 10^{-10} were removed from analysis. This left 325 pairs in the Age 1 group, 138 pairs in the Age 2 group, 760 pairs in the Age 3 group, and 392 pairs in the Age 4 group.

Measures of interaction profile similarities

The fraction of partners shared by two proteins, also known as the Jaccard index or Jaccard similarity coefficient (*102*), corresponds to the number of shared interactors divided by the total number of interactors (Fig. 4A). Self-interactions and interactions between the two proteins themselves are ignored.

We confirmed that our conclusions were not biased by the use of a particular similarity metric by using other measures (results not shown):

- The Dice similarity coefficient, calculated as twice the number of interactors shared by A and A' proteins divided by the sum of the degree of A and A' (*103*). Self-interactions and interactions between the A and A' proteins themselves are ignored.
- A measure by node calculated as the average of the fraction of interactors A shares with A' and of the fraction of interactors A' shares with A. Self-interactions and interactions between the A and A' proteins themselves are ignored.
- All three of the previous metrics, but including self-interactions and interactions between A and A' proteins.

Comparison between paralogous proteins originating from whole-genome duplication and other mechanisms

Two datasets listing pairs of paralogous proteins predicted to have originated from a polyploidy event (whole-genome duplication, WGD) were obtained from (*104, 105*). Analysis was only possible for the most recent WGDs because most pairs from older WGDs have been lost and genomic rearrangements drastically limit the validity of predictions (*35, 106*). We limited the analysis presented in Fig. 4D and fig. S43 to paralogous pairs of Age 1 in our dataset. Paralogous protein pairs likely originating from a recent WGD were defined as pairs of Age 1 in our dataset that were also classified coming from a *recent* WGD according to (*104, 105*). Paralogous protein pairs likely not originating from a WGD were defined as pairs of Age 1 in our dataset that were not classified as coming from any (*recent* or *old*) WGD according to either of the two datasets. Using both datasets independently gave similar results (fig. S43). More precise time-since-duplication estimates were obtained for paralogous pairs in our Age 1 group from Ks estimates (fig. S43).

Correction for the illusion of divergence due to low coverage using our empirically controlled quantitative framework

Previous comparisons of paralogous protein interaction profiles were limited by the use of protein-protein interaction datasets obtained by literature curation. Their non-systematic nature introduces biases where measures of divergence may reflect that paralogous proteins had not been tested in the same conditions and/or against the same set of potential partners. Alternatively, a hypothesis-driven approach may lead to preferential investigation of common interactors. These problems do not occur with systematically generated high-throughput datasets.

The incomplete nature of all currently available interaction datasets (both literature-curated and systematic datasets) can lead to an overestimation of divergence (fig. S40). For systematic datasets, these effects can be quantified and corrected for. We used the concepts developed in our quantitative framework to estimate and correct for possible biases due to low coverage. Three parameters defined in our framework (*6*) account for missing interactions in the dataset: *completeness*, *assay sensitivity*, and *sampling sensitivity*. Each can affect measures of interaction divergence.

- *Completeness*: Since all proteins in AI-1_{MAIN} have been tested systematically against the same set of proteins, the observed interaction divergence should not be affected by the completeness as long as space 1 can be assumed to be a representative sample of the *Arabidopsis* proteome, not biased towards specific interactors of some paralogs. The proportion of duplicated genes was not different in the genome, in space 1 and in AI-1 (70%, 73%, 73% respectively). Our observations may however be biased by the ORFeome collection used here (*2*), which mostly comes from a cDNA library and as a result may be depleted in ORFs with very low levels of transcription which may escape amplification and therefore cloning.
- *Assay sensitivity*: Reflecting the ability of a particular assay to detect certain interactions, assay sensitivity is unlikely to bias these results for two main reasons. First, given their common evolutionary origin and thus their similar biophysical properties, it is unlikely that paralogous proteins would behave distinctly in a given standardized assay, such as Y2H. Second, all paralogous proteins compared here were found to interact with at least one other protein in our assay, which indicates that the hybrid construct and other experimental conditions did not abolish their ability to biophysically interact. This said, the results presented here only reflect the divergence of binary interactions detectable in our Y2H assay. Evolutionary divergence of protein-protein interactions not detectable by Y2H might follow a different dynamic.
- *Sampling sensitivity*: Sampling sensitivity clearly affects the estimation of interaction divergence since the interactions missed at random for one or the other paralogous protein affect both the numerator and the denominator no matter which metric is used (fig. S40). While this problem cannot be solved for individual paralogous pairs without further experimentation, we were able to correct for the impact of sampling sensitivity on the overall averages of our measurements by using data from the six times screening of the subspace (fig. S2B). Since the main screen was completed twice, we generated every possible combination of two by two different screens out of the four additional iterations of the subspace screening (R1, R2, R3 and R4). Each of these combinations was

then used separately to build two networks, each coming from independent experiments (R1&R2 *vs* R3&R4; R1&R3 *vs* R2&R4; R2&R3 *vs* R1&R4). Since the differences between the two networks in each comparison only reflect the impact of the sampling sensitivity (the *completeness* and *assay sensitivity* are identical), comparing the interactors of each protein present in these two networks allowed us to empirically estimate the contribution of the sampling sensitivity to the observed interaction divergence (fig. S40). The three screen combinations either alone or together gave similar results: ~64% of identical interactors in the two networks. We observed that paralogous pairs share less interacting partners than sensitivity control pairs (fig. S47). Our data, while affected by under-sampling, therefore clearly reflect a real divergence. The average fraction of shared partners for *identical* proteins in *different* screens (64%) was then used to correct the average interaction divergence of the four paralogous pairs age groups, as it corresponds to an upper bound *in expectation* (in the absence of interaction divergence). For aggregate paralog analysis (Fig. 4, C-E; figs. S41-S43), data values were linearly rescaled from the range [0, 0.64] to [0, 1], which we then represents as the "corrected average fraction of shared interactors".

Statistical analyses for the duplication section

All statistical analyses were performed using the R statistical package (*54*). The correlation between functional divergence classification and interaction divergence (Fig. 4B) was assessed by the Kendall ranking correlation test (one-sided test). Two-sided non-parametric Mann-Whitney *U*-tests were used to assess the significance of the differences of the fraction of shared interactors between: (i) oldest (Age 4) paralogous protein pairs in AI-1_{MAIN} and non-paralogous protein pairs in AI-1_{MAIN} (Fig. 4C), (ii) identical proteins in two different screens and all paralogous protein pairs in AI- 1_{MAIN} (fig. S47, left *vs* middle), (iii) paralogous and non-paralogous protein pairs in AI-1_{MAIN} (fig. S47, middle *vs* right), and (iv) identical proteins in two different screens and non-paralogous protein pairs in AI-1_{MAIN} (fig. S47, left *vs* right), (v) recent paralogous protein pairs originating from polyploidy events or not (Fig. 4D; fig. S43). Two-sided non-parametric Mann-Whitney *U*-tests were used to assess the significance of the differences of Ks and Ka/Ks distributions between age groups (fig. S39). One-sided non-parametric Mann-Whitney *U*-tests were used to assess the significance of the differences of the fraction of shared interactors between the most co-expressed and the least co-expressed pairs (Fig. 4E).

The corrected averages of the fraction of shared interactors between protein pairs are shown in Fig. 4, C-E and figs. S41-S43. The corrected variables are products of the original variable and the correction term, which itself is treated as a random variable (with variance estimated from the collection of different pairwise re-screen combinations). Standard error (SE) terms for variable X with correction term (variable) C are computed exactly:

$$SE_{CX} = \frac{\sqrt{\mu_C^2 . SE_X^2 . n_X + \mu_X^2 . SE_C^2 . n_C}}{\sqrt{\min(n_X, n_C)}}$$

where n_C or n_X is the number of observations for C or X, μ the mean and SE, the standard error of the mean.

Fit of the observed interaction-rewiring rate to theoretical distributions

Following duplication, if a constant rate β of "duplicated" edge-copy loss occurs between two paralogous proteins sharing *k* percent interactors, the equation describing the evolution process per unit of time *t* is:

$$\frac{dk}{dt} = -\beta k$$

With boundary condition k_o, the solution is an exponential decay:

$$k(t) = k_o e^{-\beta t}$$

Because a linear combination of several exponential decay functions (averaged over many paralog pairs) remains an exponential decay function, the average percent-shared-interactors would also follow an exponential decay.

Average sequence identity and percent-shared-interactors values were computed for each of the four age categories. We then fit both a simple exponential-decay ($y=q * e^{kt}$) and power-law ($y=q^{tk}$) model to these data points (where q is the time-at-duplication intercept, 1 for sequence identity, 0.64 for shared interaction partner fraction – from the sampling sensitivity correction in SOM IV and SOM VII), with actual time estimates (millions of years) for each group. Fits were calculated with a non-linear least-squares method (Gauss-Newton method). For both the sequence-identity and percent-shared-interaction values, the power-law generated a better fit than the exponential-decay model, as evaluated from the residual sum of squares (RSS) and the Akaike Information Criterion (AIC) (*107*):

Fit quality	Exponential		Power law	
	k	RSS	k	RSS
Average sequence identity	-0.00750	107.0	-0.1942	52.2
Average fraction of shared partners	-0.01645	113.9	-0.3517	85.1

To test previously described models of network evolution that assume a constant rate of edge loss, we constructed simulations of two basic cases. In both, paralogous pairs initially sharing all interactions were generated randomly using a power-law distribution (Yule-Simon). At each time point following this theoretical duplication, an edge was deleted with fixed probability. In the first model, any edge was selected as a candidate. In the alternative model one of the “ancestral” edges must always remain, *i.e.* after deleting a “duplicated” edge, its “sister” edge is now fixed. In both cases a tight exponential-decay fit was obtained (again using a non-linear least-squares approach). From these analyses we conclude that the interaction data observed for paralogous pairs in AI-1_{MAIN} does not support the “constant edge-loss” theoretical models.

SUPPORTING ONLINE MATERIAL FIGURE LEGENDS

Supporting Online Material Fig. S1. Current knowledge of *Arabidopsis* protein-protein interactions. (**A**) As of April 2010, IntAct (*48*), TAIR (*46*) and BioGRID (*50*) reported a total of 5,270 protein-protein interactions for *Arabidopsis*, of which IntAct reported the majority (83%). (**B**) Of the 5,270 *Arabidopsis* protein-protein interactions, 4,252 (80%) are binary according to our criteria (table S1), and 82% of these are reported in IntAct. (**C**) Fraction of the *Arabidopsis*, human (*Homo sapiens*) and yeast (*Saccharomyces cerevisiae*) proteome with binary protein-protein interactions reported in IntAct as of February 2010 (table S1). (**D**) Number of binary protein-protein interactions reported from large-scale (>100 interactions per PubMed ID, red), small-scale experiments (<100 interactions per PubMed ID, gray), or both (blue) in IntAct (*11*) for *Arabidopsis*, human and yeast.

Supporting Online Material Fig. S2. (**A**) Construction of AtORFeome2.0. 7,109 pUNI-51*Sfi*I-ORF clones (*2*) were subcloned into a gateway compatible vector (pENTR-*Sfi*I). Together with 1,474 pENTR-TOPO ORFs (*3*), these clones comprise AtORFeome2.0 (8,583 ORFs). GW: gateway recombination sites. *Sfi*I: *Sfi*I restriction enzyme site. (**B**) Search space and pipeline used for interactome mapping and determination of sampling sensitivity (subspace). (**C**) Results for the Y2H verification step of the interactome mapping pipeline. The Y2H phenotype of each IST pair was tested four independent times. Left pie chart: fraction of IST pairs not verified (41%) or verified at least once (59%). Right pie chart: split of verified IST pair phenotypes by number of times their phenotype was scored positive, once (8%), twice (8%), three times (17%) or four times (67%) out of four attempts. Pairs verified more than twice underwent the final confirmation step of the pipeline (fig. S2B). As described in SOM II interactions in AI-1_{MAIN} dataset have 96% reproducibility.

Supporting Online Material Fig. S3. Assembling positive and random protein-protein interaction reference sets (AtPRS-v1 and AtRRS-v1). We collected interactions reported in two databases: TAIR and IntAct (March 2008) from which we selected 1,054 well-documented interactions described in at least two publications or described by at least two experimental methods. We manually curated (*5*) all papers describing 200 interactions out of the 540 involving ORFs available in AtORFeome2.0, resulting in 118 high confidence interactions constituting AtPRS-v1. The RRS was generated from a random selection out of all ~36 million possible *Arabidopsis* protein pairs (corresponding to the available search space that can be mapped using our ORFeome library), and is composed of 146 protein pairs (AtRRS-v1).

Supporting Online Material Fig. S4. Recall and precision of AI-1_{MAIN}. (**A**) Normalizing reference set (NRS) wNAPPA protein pair z-scores show homogeneous variance (homoscedastic behavior) after per-plate signal normalization. (**B**) Summarized results of control experiment testing for wNAPPA "auto-activation". Two plates containing the same set of 3x-HA fusion proteins were run in parallel with: (i) empty GST vector expressing only GST protein or (ii) GST-X AtPRS-v1 and AtRRS-v1 fusions (as performed for main AI-1 experiment). NRS controls on each plate were identical. Top panel shows GST-only experiment, with AtPRS-v1 (yellow) and AtRRS-v1 (blue) 3xHA-Y samples showing markedly less signal than the background NRS pairs (green). Bottom panel shows control plate with results similar to that seen in AtPRS/RRS-v1 wNAPPA plates. (**C**) Fraction of pairs scored positive (recall) of AtPRS-v1 (yellow), AtRRS-v1 (blue) and subset from AI-1_{MAIN} (green) by wNAPPA as a function of scoring thresholds (z-score, SOM III). (**D**) Corresponding precision (SOM III) as a function of scoring thresholds (z-score). The gray box indicates a range of scoring thresholds maximizing the recall rate difference between AtPRS-v1 and AtRRS-v1, while maintaining the AtRRS-v1 rate at relatively low levels. The same range of z-scores was used to compute the precision as well as corresponding error bars. (**E**) Fraction of PRS, RRS or AI-1_{MAIN} sample pairs positive in wNAPPA and associated precision across a range of scoring thresholds from 1.0 to 2.0. Gray area: range of experimental conditions maximizing sensitivity and specificity.

Supporting Online Material Fig. S5. Sampling sensitivity analyses. (**A**) Sensitivity of the interactome mapping strategy was measured by screening six times the subspace (fig. S2B). Average number of Y2H interactions detected for AtPRS-v1 pairs and for all pairs as a function of the number of screening interactions (SOM IV). Error bars: standard deviation. (**B**) Predicted saturation curve of our Y2H assay after modeling the data in fig. S5A (SOM IV). Average number of Y2H interactions predicted to be detected by the model for AtPRS-v1 pairs (yellow bars) and for all pairs (gray bars) as a function of the number of screening iterations. Error bars represent standard deviations.

Supporting Online Material Fig. S6. Comparison of the topology of AI-1_{MAIN} and LCI_{BINARY}. (**A**) Global network topology description and comparison of AI-1_{MAIN} and LCI_{BINARY}. Top left: Normalized distribution of the protein degree (number of interactors, *P* = 0.006). Top right: Normalized distribution of shortest paths ($P < 2.2 \times 10^{-6}$). Bottom left: Average degree of neighbors as a function of degree. Bottom right: Average clustering coefficient of neighbors as a function of degree ($P = 5 \times 10^{-7}$). All measurements, except degree, are calculated based on the largest connected component of both networks. For the calculation of clustering coefficient, only proteins with degree >1 were taken into account. All *p*-values are calculated using a one-sided paired Wilcoxon signed rank test. Table: topological parameters of small-world effect for AI-1_{MAIN} and LCI_{BINARY}: "λ" is ratio of average path length of the data network (AI-1_{MAIN} or LCI_{BINARY}) divided by the average path length of 30 Erdős-Rényi random networks with the same number of nodes and edges as the data networks; "γ" is the ratio of the average clustering coefficient of the data network (AI-1_{MAIN} or LCI_{BINARY}) over the average clustering coefficient of 30 Erdős–Rényi random networks with the same number of nodes and edges as the data networks; and "σ" is the ratio of γ over λ. Analyses were performed using R (*54*) and igraph (*108*). (**B**) Comparison of interactions involving at least one plant specific protein, *i.e.* plant-specific interactions (SOM V) in AI-1_{MAIN} and LCI_{BINARY}. Top: Degree distribution of plant-specific and non-plant specific proteins in AI-1_{MAIN} and LCI_{BINARY}. Bottom: AI-1_{MAIN} contains more than twice as many interactions involving plant-specific proteins than LCI_{BINARY}. Left: counts; Right: pie charts showing that ~40% of interactions in AI-1_{MAIN}, and ~20% of interactions in LCI_{BINARY} involve at least one plant-specific protein.

Supporting Online Material Fig. S7. AI-1 is enriched in biologically relevant protein-protein interactions. (**A**) Distribution of mRNA expression Pearson correlation coefficients (PCC) over 1,436 arrays (*52, 53*) for AI-1_{MAIN} interacting and non-interacting pairs. Inset: percentage of AI-1_{MAIN} interacting and non-interacting pairs with PCC > 0.75, *p*-value of one-sided Fisher's exact test; error bars: standard error of the proportion. P_{MW}: *p*-value of Mann-Whitney *U*-test. (**B**) Enrichments in shared Gene Ontology (GO) annotations (*55*) for AI-1_{MAIN} interacting pairs versus non-interacting pairs, in the three branches of the GO vocabulary and as a function of GO annotation breadth (SOM V). All enrichments are statistically significant by Fisher's exact test ($P < 0.05$). (**C**) Fraction of AI-1_{MAIN} interacting and non-interacting pairs sharing a precise GO mutant phenotype relative to pairs where both members have a GO mutant phenotype in the biological process branch with a breadth ≤50 ($n = 308$). Error bars: standard error of the proportion. (**D**) Enrichment in shared precise GO annotations (of breadth ≤50) in the indicated GO categories for protein pairs sharing interactors in AI-1_{MAIN}. All directly interacting pairs were excluded. All enrichments are statistically significant by Fisher's exact test ($P < 0.05$). (**E**) Proteins in AI-1_{MAIN} are less annotated than proteins in a network of literature-curated interactions (LCI). Ratio of number of proteins in AI-1_{MAIN} and LCI with indicated Gene Ontology GO-slim annotations, which were used as an estimate of function in each of the three branches of the GO tree: molecular function (top panel), cellular component (middle panel) and biological process (bottom panel). The red line represents a ratio of one. (**F**) Proportion of proteins in AI-1 with and without GO annotation of breadth ≤50. Proteins without such GO annotation are classified as either directly interacting or sharing more than half of their interactors with a protein with such a GO annotation, or both, or neither.

Supporting Online Material Fig. S8. Plant signaling networks in AI-1. (**A**) Putative phosphorylation signaling subnetwork extracted from $LCI_{BINARY/SPACE1}$ and AI-1. Bar plot: number of protein-protein interactions between kinases/phosphatases and phosphorylated proteins in LCI_{BINARY} and AI-1 (outside and within space 1). (**B**) Novel aspects of transcriptional co-repression mediated by TOPLESS (TPL) and TPL-related 3 (TPR3) through physical interactions with EAR-motif containing proteins in LCI_{BINARY}

and AI-1. Literature interactions from $\mathrm{LCI}_{\mathrm{BINARY}}$ and (*15, 109, 110*). (**C**) Aspects of jasmonic acid-mediated transcriptional regulation suggested by protein-protein interactions from $\mathrm{LCI}_{\mathrm{BINARY}}$ and AI-1. Literature interactions from $\mathrm{LCI}_{\mathrm{BINARY}}$ and (*110*).

Supporting Online Material Fig. S9. Community identification in AI-1_{MAIN}. (**A**) Size distribution of communities identified in AI-1_{MAIN} before applying the size and density filters. (**B**) Communities containing more than 5 proteins with density > 0 are enriched in GO annotations compared to communities containing more than 5 proteins with density = 0.

Supporting Online Material Fig. S10. "Brassinosteroid signaling and "phosphoprotein binding" community (no. 4932). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S11. "Auxin signaling" community (no. 1652). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S12. "Ubiquitin-dependent degradation" community (no. 369). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S13. "DNA repair and ubiquitination" community (no. 456). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S14. "DNA binding" community (no. 500). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S15. "Cytoskeleton organization and root hair elongation" community (no. 711). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S16. "Ubiquitination" community (no. 899). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S17. "Oxidoreductase activity" community (no. 1534). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S18. "Transcription and nitrogen metabolism" community (no. 1568). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S19. "Vesicle trafficking" community (no. 1861). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S20. "Transcription/gene expression" community (no. 1995). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S21. "Nucleosome assembly" community (no. 2535). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S22. "RNA binding" community (no. 2796). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S23. "mRNA splicing" community (no. 3963). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S24. "Calmodulin binding" community (no. 4080). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S25. "Aromatic compound metabolism" community (no. 4167). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S26. "Water transport" community (no. 4298). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S27. "Ubiquitination" community (no. 4617). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S28. "Transmembrane transport" community (no. 4716). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S29. "TCA cycle" community (no. 5027). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S30. "Ribonucloprotein complex" community (no. 5081). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S31. "Potassium transport and kinase activity" community (no. 5249). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S32. "Seed germination and gibberellin and jasmonic acid signaling" community (no. 5255). Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community.

Supporting Online Material Fig. S33. Community number 1784. Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community. There are no enriched GO annotations corresponding to this community.

Supporting Online Material Fig. S34. Community number 2706. Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community. There are no enriched GO annotations corresponding to this community.

Supporting Online Material Fig. S35. Community number 3347. Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in this community. There are no enriched GO annotations corresponding to this community.

Supporting Online Material Fig. S36. Examples of connected communities (**A**) "Transcription/gene expression" and "nucleosome assembly" communities are in contact. Nodes represent proteins. Edges represent protein-protein interactions. Edge color indicates interactions in the nucleosome community (pink), and transcription (blue). (**B**) A “ubiquitination” and “ubiquitin dependent degradation” community are in contact. Edges represent protein-protein interactions. Edge color indicates interactions in the ubiquitination community (green), and ubiquitin dependent degradation (red). (**C**) A “water transport” and “vesicle trafficking” community are in contact through four proteins, including NTL9 and ANAC089 that belong to a “transmembrane transport” community. Edges represent protein-protein interactions. Edge color indicates interactions in the water transport community (orange), vesicle trafficking community (blue) and transmembrane transport community (yellow).

Supporting Online Material Fig. S37. Interactions potentially mediating the release of NTL9 and ANAC089 from the membrane.

Supporting Online Material Fig. S38. *Arabidopsis* genome and AI-1_{MAIN} contain more paralogous genes and pairs of paralogous proteins, respectively, than human or yeast genomes and interactomes. (**A**) Number of genes with at least one paralog in the human (*Homo sapiens*), yeast (*Saccharomyces cerevisiae*) or *Arabidopsis* (*Arabidopsis thaliana*) genome. (**B**) Number of pairs of paralogous proteins present in interactome datasets of similar quality for human (HI-1, *10*), yeast (YI-1, *9*) and *Arabidopsis* (AI-1_{MAIN}). Paralogy relationships were derived using the EnsemblCompara GeneTrees pipeline (*97*), for *Arabidopsis* they were downloaded from http://www.gramene.org; *96*) in February 2009; for human and yeast they were downloaded from Ensembl in June 2009). (**C**) Distribution of the fraction of shared interactors between paralogous proteins in AI-1_{MAIN}.

Supporting Online Material Fig. S39. Phylogeny-based estimates of time-since-duplication. (**A**) Average Ks for paralogous gene pairs in each age group. Ks saturation does not allow distinction of paralogous gene pairs in age 3 and age 4 ($P = 0.52$; Mann-Whitney *U*-test). (**B**) Ratio of Ka over Ks (Ka/Ks) for paralogous gene pairs in each age group. Ka/Ks for paralogous gene pairs of age 1 is higher than for paralogous gene pairs of age 2 ($P = 2.4 \text{ X } 10^{-17}$; Mann-Whitney *U*-test), reflecting a relaxation of selective pressure.

Supporting Online Material Fig. S40. Identification and correction of biases due to dataset incompleteness. (**A**) Example of paralogous protein interaction divergence overestimation in a LCI dataset due to its non-systematic and non-standardized character. (**B**) Example of paralogous protein interaction divergence overestimation in systematic high-throughput datasets due to the sampling sensitivity. (**C**) Principle of the empirical estimation of the contribution of the sampling sensitivity to the interaction divergence observed in AI-1_{MAIN}.

Supporting Online Material Fig. S41. Average interaction divergence of paralogous pairs is not biased by the presence of large families in AI-1_{MAIN}. Corrected average fraction of shared interactors of paralogous protein pairs in each age group as a function of the size of the family of paralogs in AI-1_{MAIN}. Pairs of paralogous proteins that have one ("= 2", families of two paralogs), maximum two ("≤ 3", families of two or three paralogs), maximum three ("≤ 4", families of two to four paralogs), maximum four ("≤ 5", families of two to five paralogs), or any number ("all") of paralogs in AI-1_{MAIN}.

Supporting Online Material Fig. S42. Dynamics of interaction rewiring following gene duplication. (**A**) Corrected average fraction of shared interactors (red circles) and average protein sequence identity (blue squares) between pairs of paralogous proteins as a function of the most recent estimate of Δ time. Error bars: standard error of the mean (SOM VII). The data is plotted in log-log scale. Full lines: power-law fits; dotted red line: exponential fit to the corrected fraction of shared interactors of paralogous pairs. (**B**) Corrected average fraction of shared interactors (red circles) and average protein sequence identity (blue squares) between pairs of paralogous proteins as a function of the most ancient estimated Δ time since duplication. Error bars: standard error of the mean (SOM VII). Inset: data plotted in log-log scale. Full lines: power-law fits; dotted red line: exponential fit to the corrected fraction of shared interactors of paralogous pairs.

Supporting Online Material Fig. S43. Comparing polyploidy paralogous pairs to other paralogous pairs within age 1. (**A**) Corrected average fraction of interactors shared between pairs of paralogous proteins originating from a polyploidy event and other pairs within age 1. Polyploidy event predictions according to two datasets (*104, 105*). Error bars: standard error of the mean (SOM VII). *p*-values: Mann-Whitney *U*-test. WGD: whole genome duplication. (**B**) Average Ks of pairs of paralogous proteins originating from a polyploidy event and other pairs within age 1. The difference in average Ks for the two groups corresponds to ~55 million years (*111*). Error bars: standard error of the mean. *p*-value: Mann-Whitney *U*-test.

Supporting Online Material Fig. S44. Divergence of expression profiles following gene duplications. Distribution of mRNA expression Pearson correlation coefficients (PCC; SOM V) for paralogous pairs by age group and for non-paralogous pairs in AI-1_{MAIN}.

Supporting Online Material Fig. S45. Example of protein sequence and interaction divergence. (**A**) Sequence alignment and interaction subnetwork of IAA1 and IAA2, recent duplicates (~10 million years) encoding auxin-responsive proteins that share 75% of sequence identity and 31% of their interactors in AI-1_{MAIN}. (**B**) Evolution of protein-protein interactions within the actin family. Top: duplication tree (colored lines with respect to the time-since-duplication for a pair) and interactions of six actin proteins in AI-1_{MAIN}. Bottom: sequence identity and fraction of shared interactors for each of the 15 pairs of actin proteins as a function of the estimated Δ time-since-duplication. Colored arrows: groups of proteins of similar time-since-duplication. myrs: million years.

Supporting Online Material Fig. S46. Paralogous pair age category is not strictly related to protein sequence identity. Distributions of pairs of paralogous proteins in AI-1_{MAIN} for each age category as a function of their average protein sequence identity.

Supporting Online Material Fig. S47. Paralogous pairs in AI-1_{MAIN} present a substantial interaction divergence yet exhibit retention of ancestral history in the interactome. Box-plot distribution of the fraction of interactions of a protein commonly found in two independent screens (left), of the fraction of interactors shared by pairs of paralogous (middle), and of non-paralogous proteins (right) in AI-1_{MAIN}. The plot shows the median (red horizontal line), inner-quartile (box), and 95% range (error bars) of data and outliers (open circles). All distributions statistically different from one another by Mann-Whitney *U*-test (*P*-values $< 2.2 \times 10^{-16}$).

GLOSSARY

Space 1: The search space used for interactome mapping.

Y2H: Yeast two-hybrid.

AI-1_{MAIN}: *Arabidopsis* Interactome version 1, main screen. The dataset of 5,664 interactions produced by screening the complete space 1 twice using the Y2H assay.

wNAPPA: well Nucleic Acid Programmable Protein Array.

PRS: *Arabidopsis thaliana* positive reference set (*At*PRS-v1).

RRS: *Arabidopsis thaliana* random reference set (*At*RRS-v1).

AI-1: The composite dataset corresponding to the union of AI-1_{MAIN} and interactions identified in repeated screens on the subspace indicated in fig. S2, as performed to estimate sampling sensitivity.

GO: Gene Ontology.

$\mathrm{LCI}_{\mathrm{BINARY}}$: Binary subset of Literature Curated Interactions, split into 587 supported by multiple evidences high-quality binary interactions and 3,665 pairs supported by a single evidence.

EAR motif: Ethylene-response-factor-associated amphiphilic repression motif.

WGD: Whole Genome Duplication.

SUPPORTING ONLINE MATERIAL TABLE NAMES

Supporting Online Material Table 1. Binary protein-protein interactions detection methods codes.

Supporting Online Material Table 2. IntAct binary protein-protein interactions for yeast, human and *Arabidopsis*.

Supporting Online Material Table 3. Space 1.

Supporting Online Material Table 4. Protein-protein interactions from AI-1_{MAIN}, AI-1_{REPEAT}, LCI, LCI_{BINARY} and LCI_{BINARY} multiple evidence.

Supporting Online Material Table 5. Reference sets.

Supporting Online Material Table 6. Subspace.

Supporting Online Material Table 7. Detailed results of the repeat screen experiment.

Supporting Online Material Table 8. Protein attributes.

Supporting Online Material Table 9. Selected interactions from AI-1.

Supporting Online Material Table 10. Functional description of AI-1_{MAIN} communities.

Supporting Online Material Table 11. Normalization Reference Set (NRS).

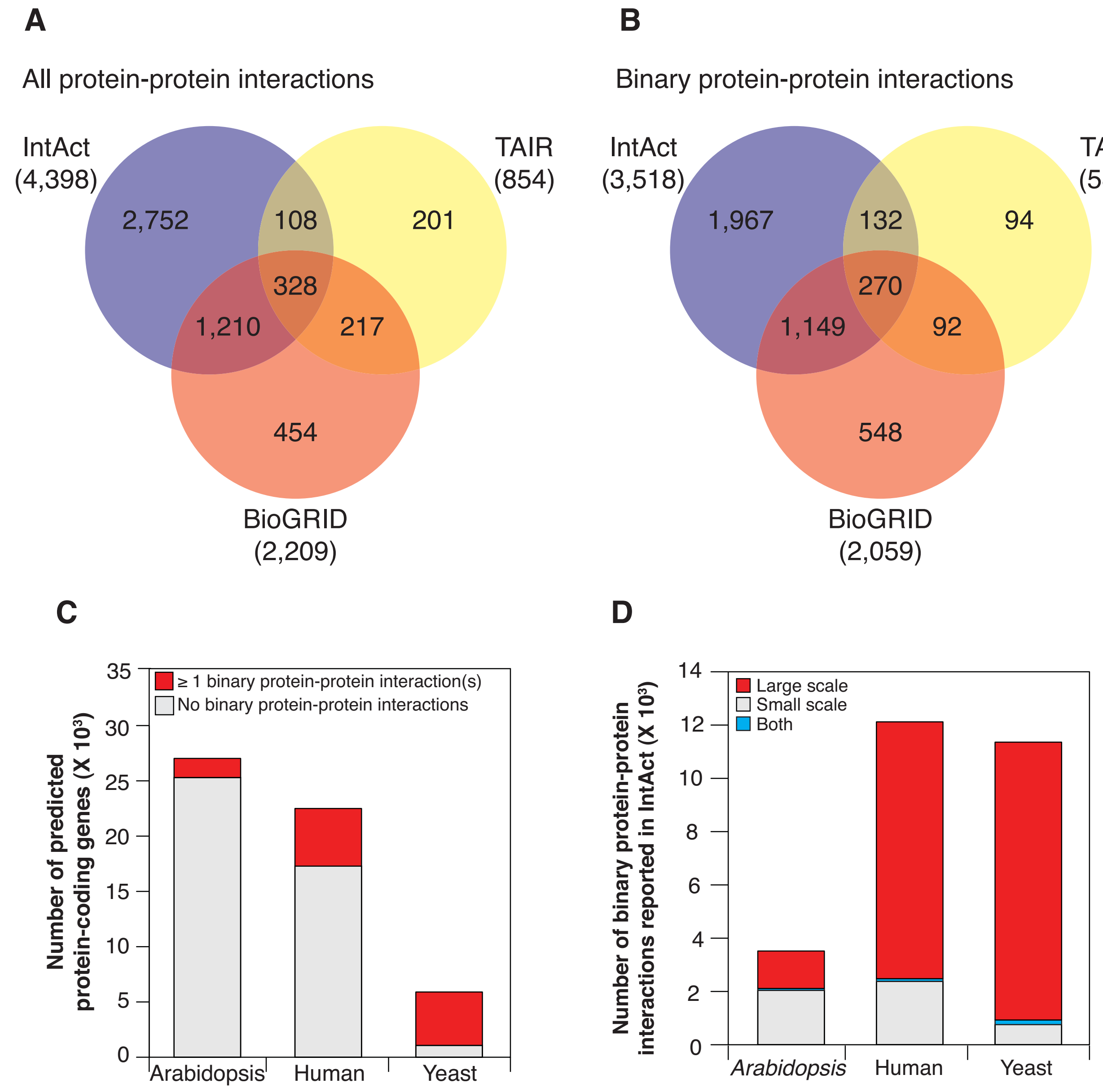
Supporting Online Material Table 12. List of genomes from non-plant (non-viridiplantae) lineages.

SUPPORTING ONLINE MATERIAL REFERENCES

44. J. L. Hartley, G. F. Temple, M. A. Brasch, *Genome Res.* **10**, 1788 (2000).
45. A. J. Walhout *et al.*, *Methods Enzymol.* **328**, 575 (2000).
46. D. Swarbreck *et al.*, *Nucleic Acids Res.* **36**, D1009 (2008).
47. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
48. S. Kerrien *et al.*, *Nucleic Acids Res.* **35**, D561 (2007).
49. N. Ramachandran *et al.*, *Nat. Methods* **5**, 535 (2008).
50. C. Stark *et al.*, *Nucleic Acids Res.* **34**, D535 (2006).
51. J. Gough, K. Karplus, R. Hughey, C. Chothia, *J. Mol. Biol.* **313**, 903 (2001).
52. D. J. Craigon *et al.*, *Nucleic Acids Res.* **32**, D575 (2004).
53. M. Schmid *et al.*, *Nat. Genet.* **37**, 501 (2005).
54. R Development Core Team. (R Foundation for Statistical Computing, Vienna, Austria, 2011).
55. M. Ashburner *et al.*, *Nat. Genet.* **25**, 25 (2000).
56. H. Wang, N. Ngwenyama, Y. Liu, J. C. Walker, S. Zhang, *Plant Cell* **19**, 63 (2007).
57. A. Santner, M. Estelle, *Nature* **459**, 1071 (2009).
58. L. Katsir, H. S. Chung, A. J. Koo, G. A. Howe, *Curr. Opin. Plant Biol.* **11**, 428 (2008).
59. B. Vanholme, W. Grunewald, A. Bateman, T. Kohchi, G. Gheysen, *Trends Plant Sci.* **12**, 239 (2007).
60. J. Cui *et al.*, *Nucleic Acids Res.* **36**, D999 (2008).
61. S. De Bodt, S. Proost, K. Vandepoele, P. Rouze, Y. Van de Peer, *BMC Genomics* **10**, 288 (2009).
62. J. Geisler-Lee *et al.*, *Plant Physiol.* **145**, 317 (2007).
63. M. Lin, X. Shen, X. Chen, *Nucleic Acids Res.* **39**, database issue (2010).
64. E. L. Sonnhammer, S. R. Eddy, E. Birney, A. Bateman, R. Durbin, *Nucleic Acids Res.* **26**, 320 (1998).
65. J. Pei, R. Sadreyev, N. V. Grishin, *Bioinformatics* **19**, 427 (2003).
66. J. Soding, *Bioinformatics* **21**, 951 (2005).
67. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, *Proc. Natl. Acad. Sci. USA* **96**, 4285 (1999).
68. L. R. Matthews *et al.*, *Genome Res.* **11**, 2120 (2001).
69. A. J. Walhout *et al.*, *Science* **287**, 116 (2000).
70. S. A. Saracco *et al.*, *Plant J.* **59**, 344 (2009).
71. R. Maor *et al.*, *Mol. Cell Proteomics* **6**, 601 (2007).
72. C. Manzano, Z. Abraham, G. Lopez-Torrejon, J. C. Del Pozo, *Plant Mol. Biol.* **68**, 145 (2008).
73. T. Igawa *et al.*, *J. Exp. Bot.* **60**, 3067 (2009).
74. M. Gribskov *et al.*, *Nucleic Acids Res.* **29**, 111 (2001).
75. P. Durek *et al.*, *Nucleic Acids Res.* **38**, D828 (2009).
76. T. S. Nuhse, A. Stensballe, O. N. Jensen, S. C. Peck, *Plant Cell* **16**, 2394 (2004).
77. F. Wolschin, W. Weckwerth, *Plant Methods* **1**, 9 (2005).
78. S. de la Fuente van Bentem *et al.*, *Nucleic Acids Res.* **34**, 3267 (2006).
79. J. J. Benschop *et al.*, *Mol. Cell. Proteomics* **6**, 1198 (2007).

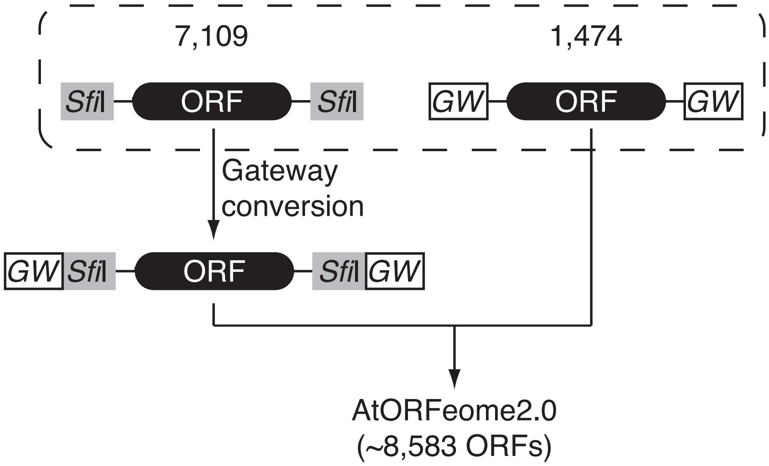
80. T. Niittyla, A. T. Fuglsang, M. G. Palmgren, W. B. Frommer, W. X. Schulze, *Mol. Cell. Proteomics* **6**, 1711 (2007).
81. T. S. Nuhse, A. R. Bottrill, A. M. Jones, S. C. Peck, *Plant J.* **51**, 931 (2007).
82. A. J. Carroll, J. L. Heazlewood, J. Ito, A. H. Millar, *Mol. Cell. Proteomics* **7**, 347 (2008).
83. N. Sugiyama *et al.*, *Mol. Syst. Biol.* **4**, 193 (2008).
84. S. de la Fuente van Bentem *et al.*, *J. Proteome Res.* **7**, 2458 (2008).
85. S. A. Whiteman *et al.*, *Proteomics* **8**, 3536 (2008).
86. A. M. Jones *et al.*, *J. Proteomics* **72**, 439 (2009).
87. H. Li *et al.*, *Proteomics* **9**, 1646 (2009).
88. S. Reiland *et al.*, *Plant Physiol.* **150**, 889 (2009).
89. H. Nakagami *et al.*, *Plant Physiol.* **153**, 1161 (2010).
90. Z. Wang, G. Dong, S. Singh, H. Steen, J. Li, *J. Proteomics* **72**, 831 (2009).
91. Z. Y. Peng *et al.*, *Nucleic Acids Res.* **37**, D975 (2009).
92. K. W. Bock *et al.*, *Plant Physiol.* **140**, 1151 (2006).
93. K. Saito *et al.*, in *Plant Metabolomics*. (Springer Berlin Heidelberg, 2006), vol. 57, pp. 141-154.
94. S. K. Palaniswamy *et al.*, *Plant Physiol.* **140**, 818 (2006).
95. G. F. Berriz, J. E. Beaver, C. Cenik, M. Tasan, F. P. Roth, *Bioinformatics* **25**, 3043 (2009).
96. C. Liang *et al.*, *Nucleic Acids Res.* **36**, D947 (2008).
97. A. J. Vilella *et al.*, *Genome Res.* **19**, 327 (2009).
98. E. Sayers *et al.*, *Nucleic Acids Res.* **37**, D5 (2009).
99. S. B. Hedges, J. Dudley, S. Kumar, *Bioinformatics* **22**, 2971 (2006).
100. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res* **22**, 4673 (1994).
101. Z. Yang, *Comput Appl Biosci* **13**, 555 (1997).
102. P. Jaccard, *Bulletin de la Société Vaudoise des Sciences Naturelles* **37**, 547 (1901).
103. L. R. Dice, *Ecology* **26**, 297 (1945).
104. G. Blanc, K. Hokamp, K. H. Wolfe, *Genome Res.* **13**, 137 (2003).
105. J. E. Bowers, B. A. Chapman, J. Rong, A. H. Paterson, *Nature* **422**, 433 (2003).
106. G. Blanc, K. H. Wolfe, *Plant Cell* **16**, 1667 (2004).
107. H. Akaike, *Automatic Control, IEEE Transactions on* **19**, 716 (1974).
108. G. Csardi, T. Nepusz, *InterJournal* Complex Systems, 1695 (2006).
109. H. Szemenyei, M. Hannon, J. A. Long, *Science* **319**, 1384 (2008).
110. H. S. Chung, G. A. Howe, *Plant Cell* **21**, 131 (2009).
111. M. A. Koch, B. Haubold, T. Mitchell-Olds, *Mol Biol Evol* **17**, 1483 (2000).

Supplementary Figure 1

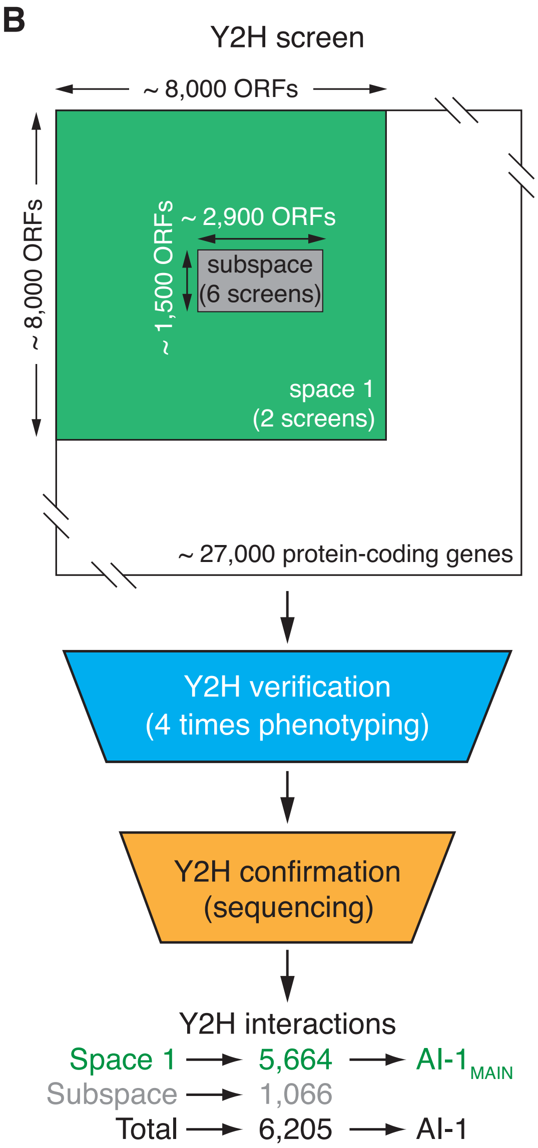


Supplementary Figure 2

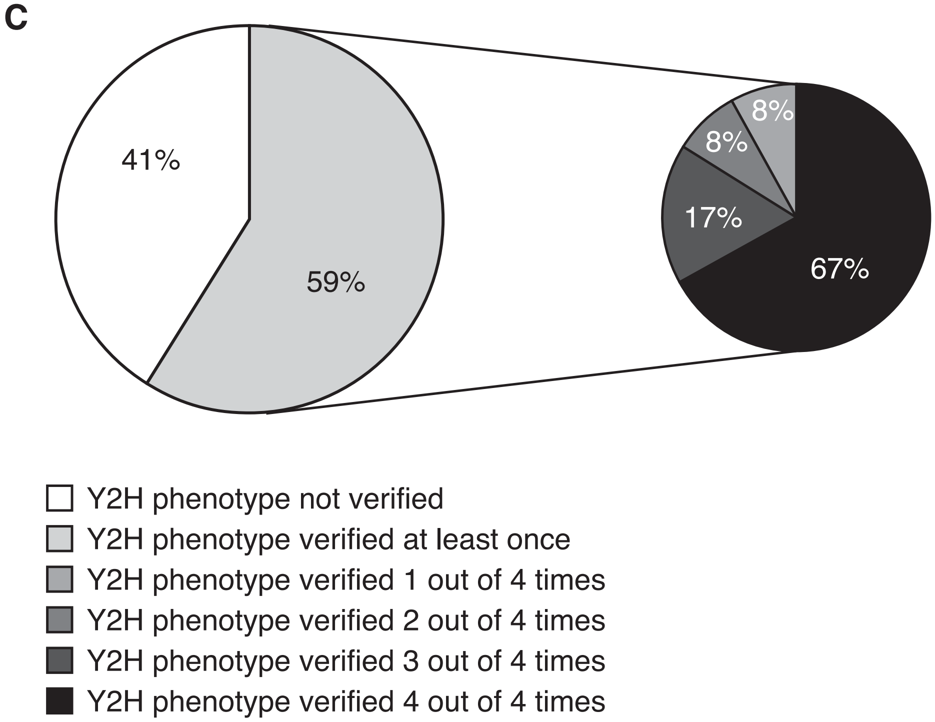
A



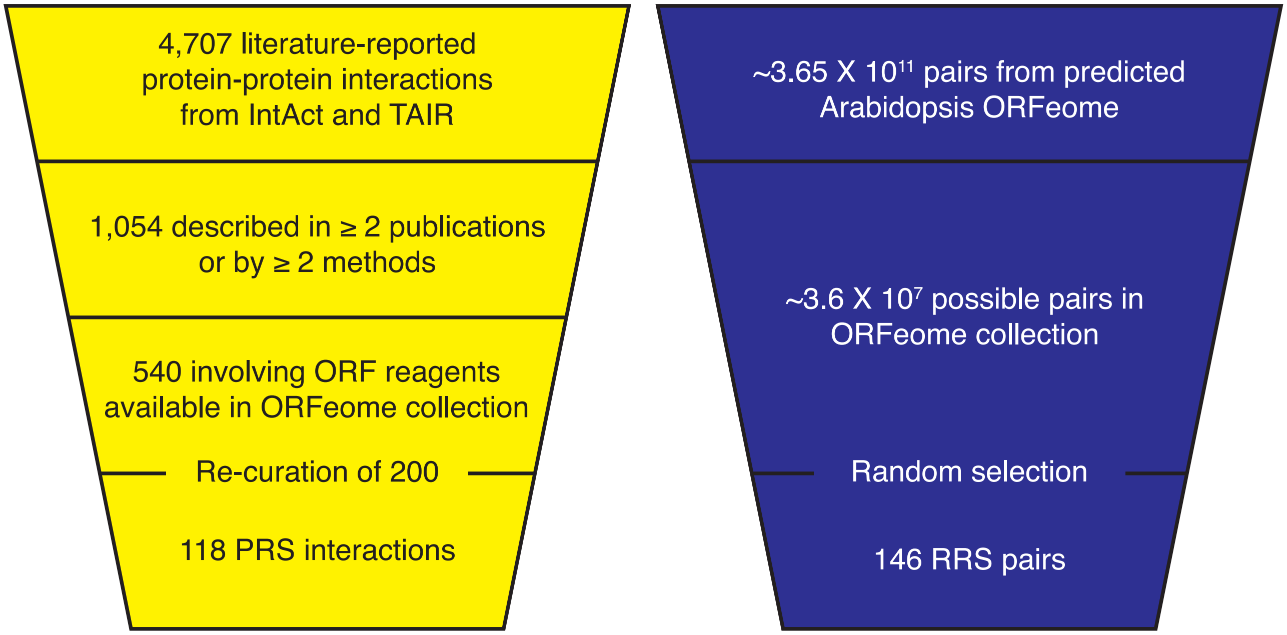
Supplementary Figure 2



Supplementary Figure 2



Supplementary Figure 3

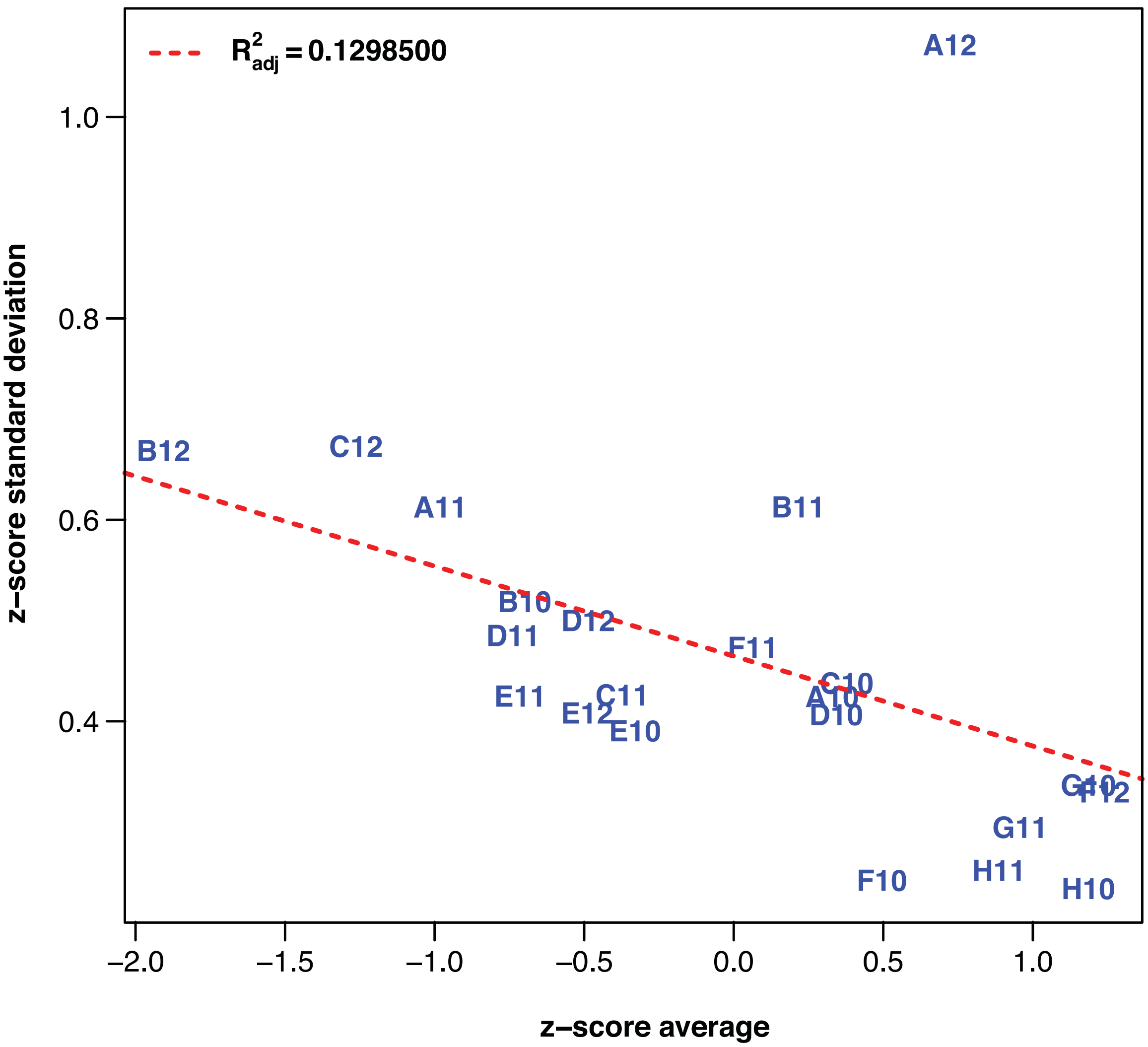

4,707 literature-reported protein-protein interactions from IntAct and TAIR
1,054 described in ≥ 2 publications or by ≥ 2 methods
540 involving ORF reagents available in ORFeome collection
Re-curation of 200
118 PRS interactions
~3.65 X 10^{11} pairs from predicted Arabidopsis ORFeome
~3.6 X 10^{7} possible pairs in ORFeome collection
Random selection
146 RRS pairs

Arabidopsis Positive Reference Set
version 1 (AtPRS-v1)

Arabidopsis Random Reference Set
version 1 (AtRRS-v1)

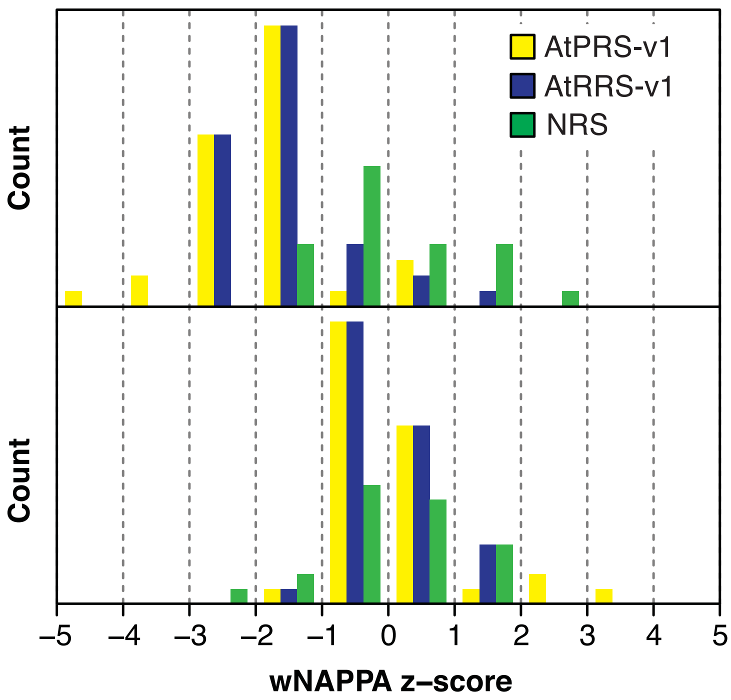
Supplementary Figure 4

A

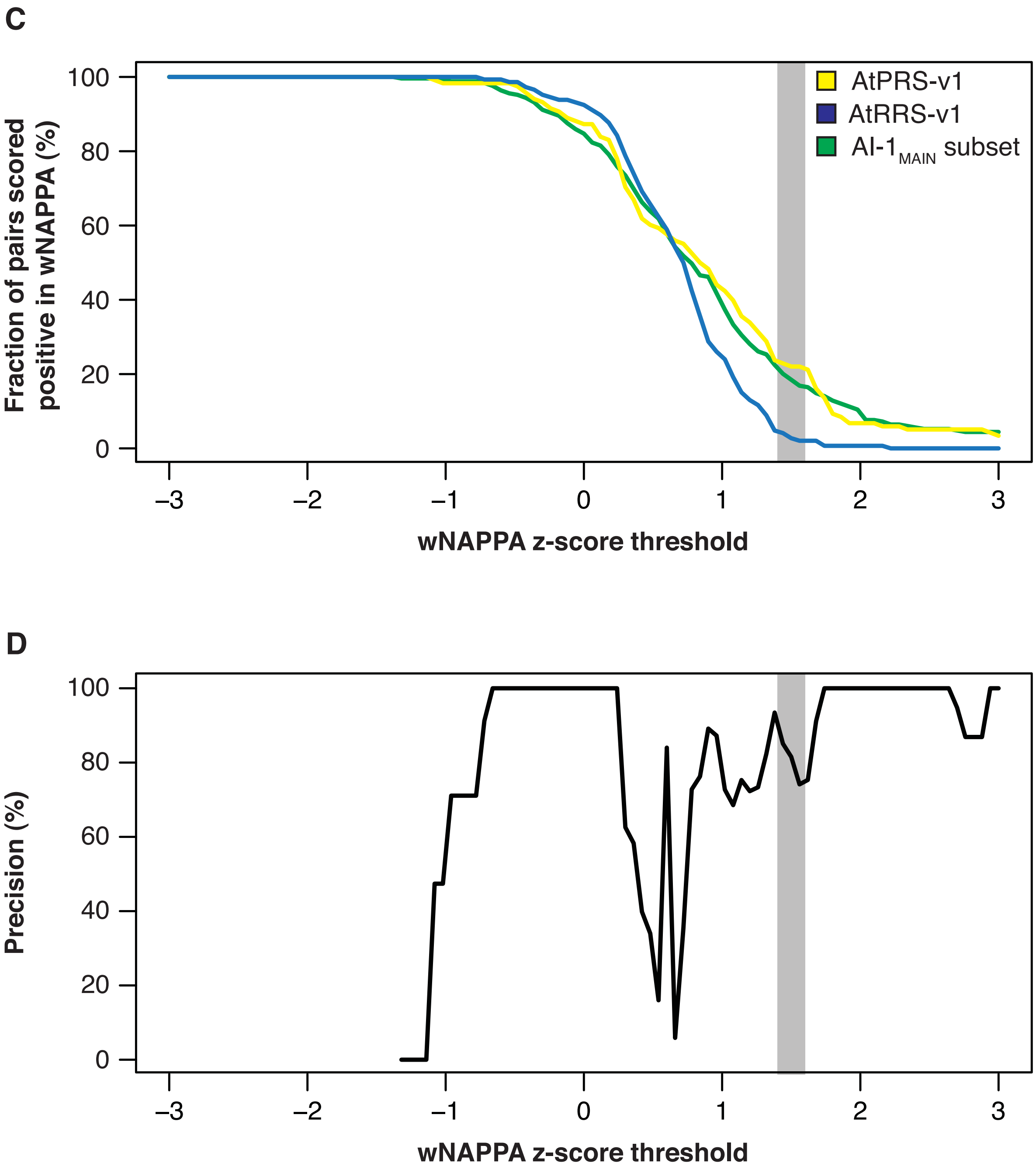


Supplementary Figure 4

B

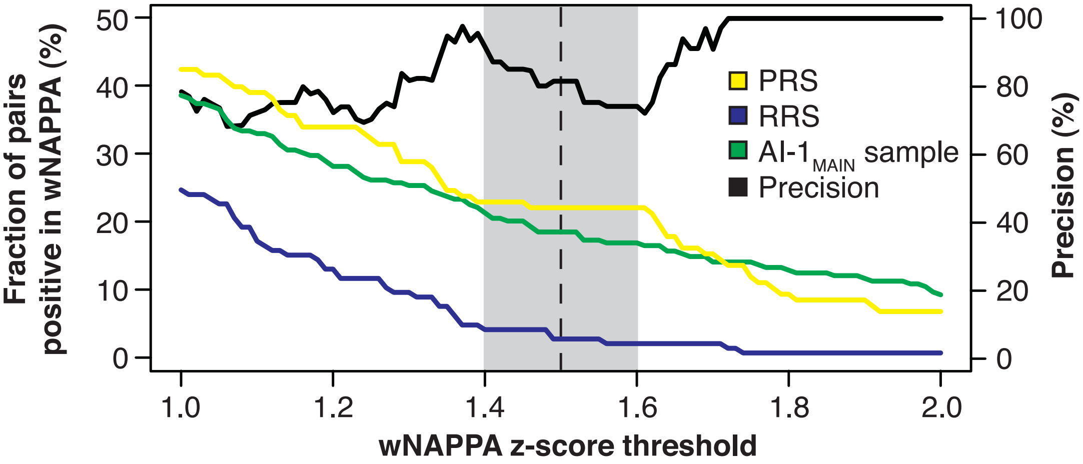


Supplementary Figure 4

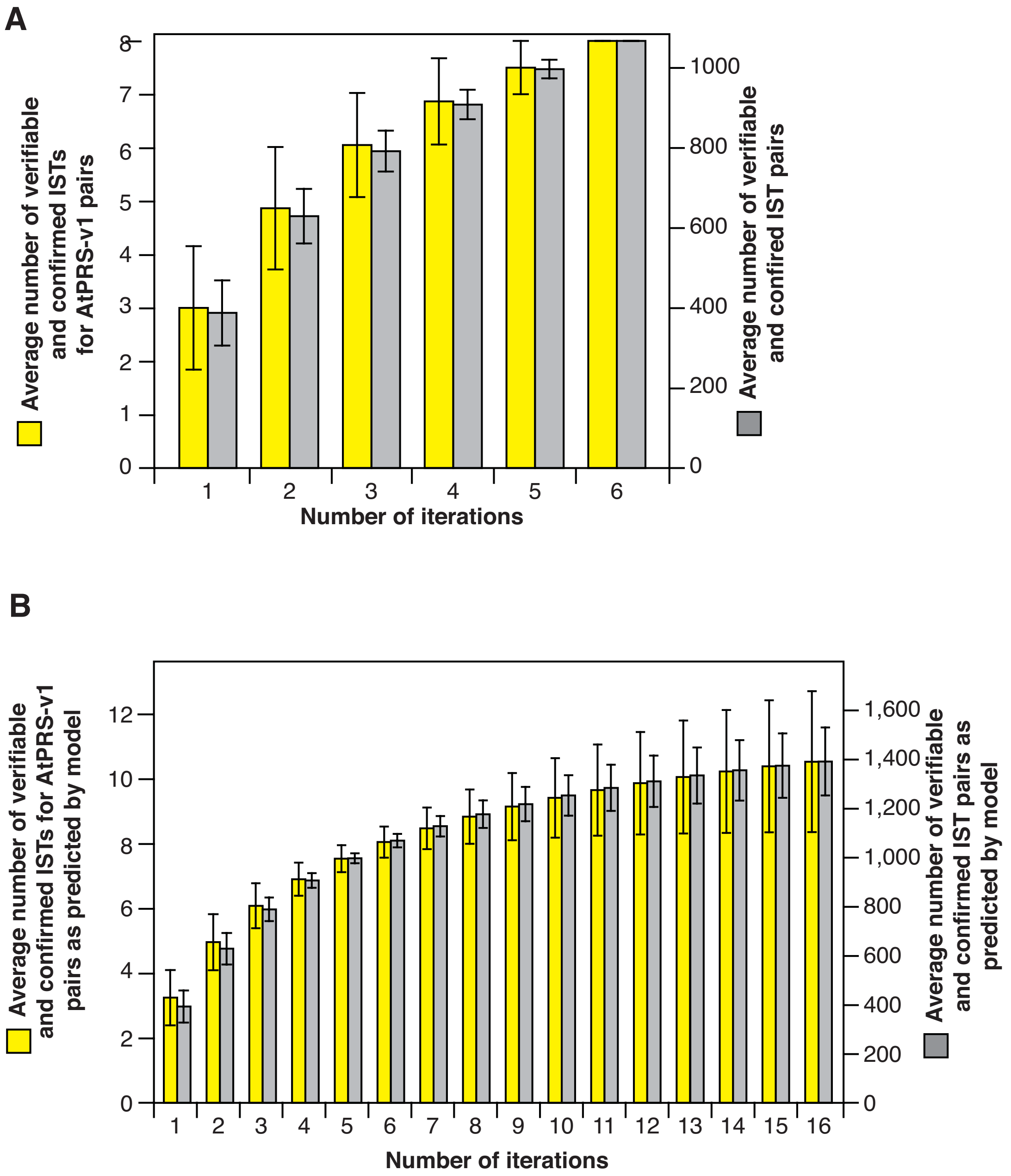


Supplementary Figure 4

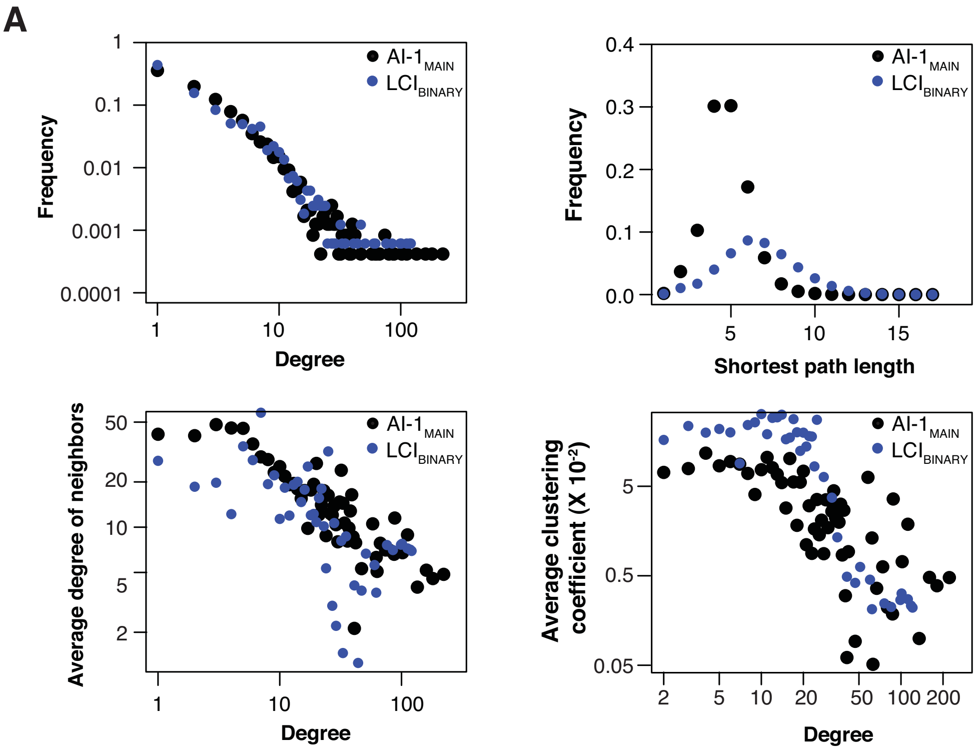
E



Supplementary Figure 5

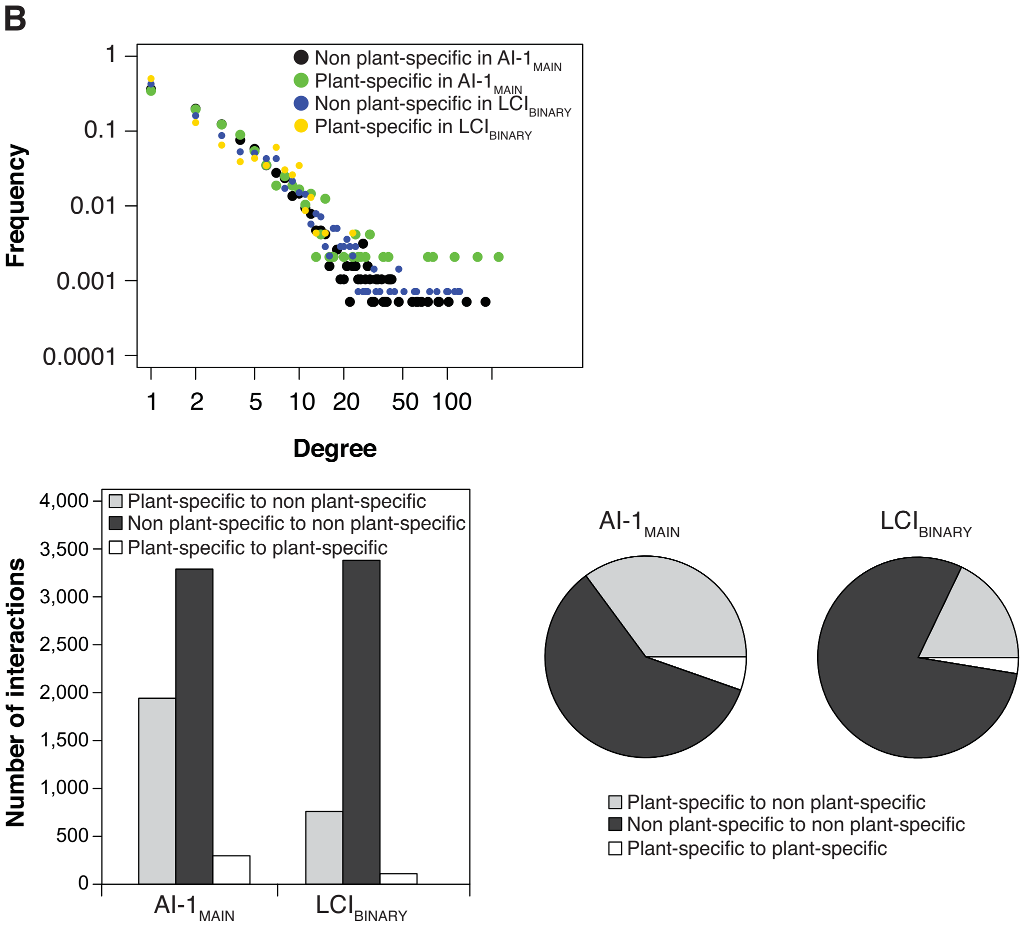


Supplementary Figure 6

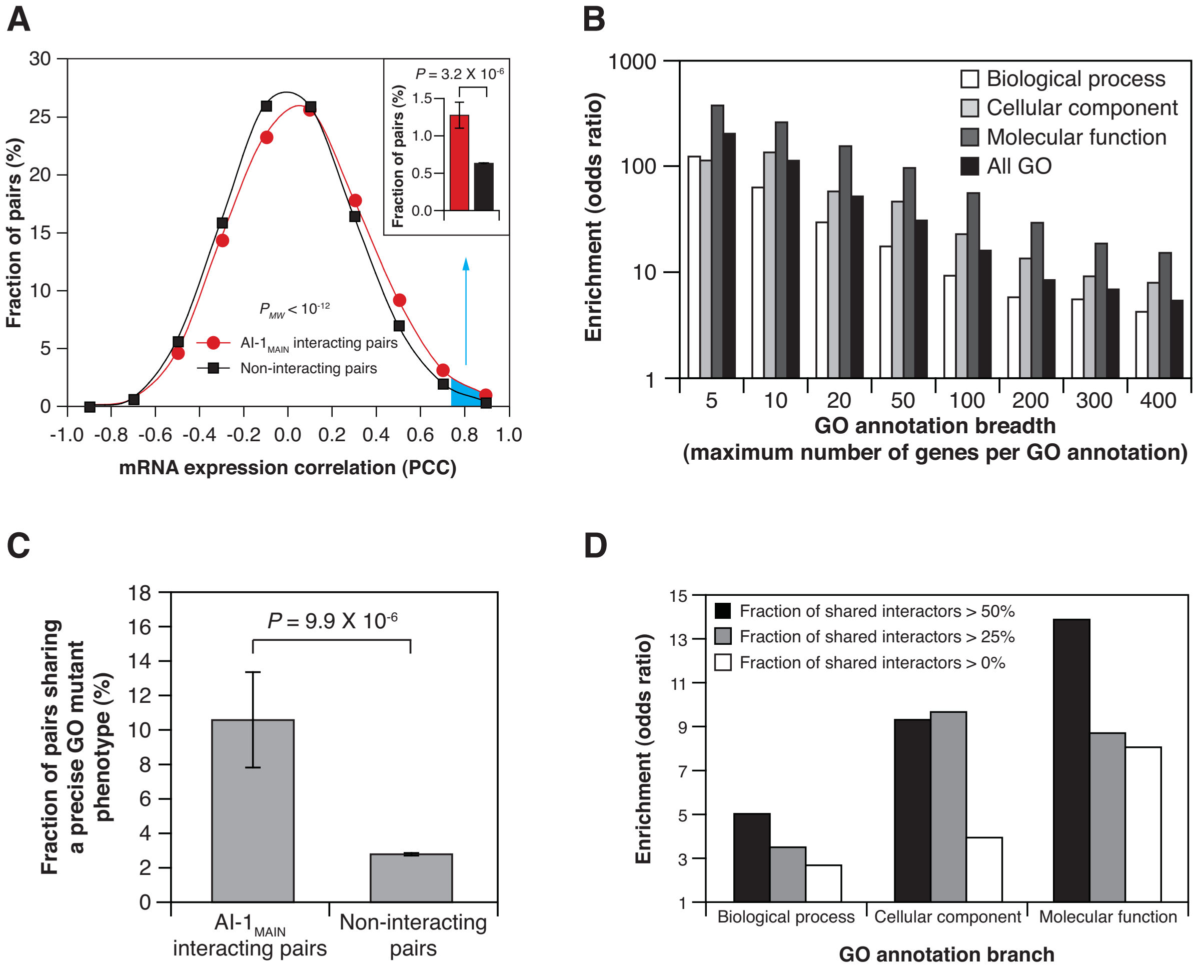


	λ	γ	σ
$AI\text{-}1_{MAIN}$	0.844	32.018	37.936
LCI_{BINARY}	1.138	70.524	61.972

Supplementary Figure 6

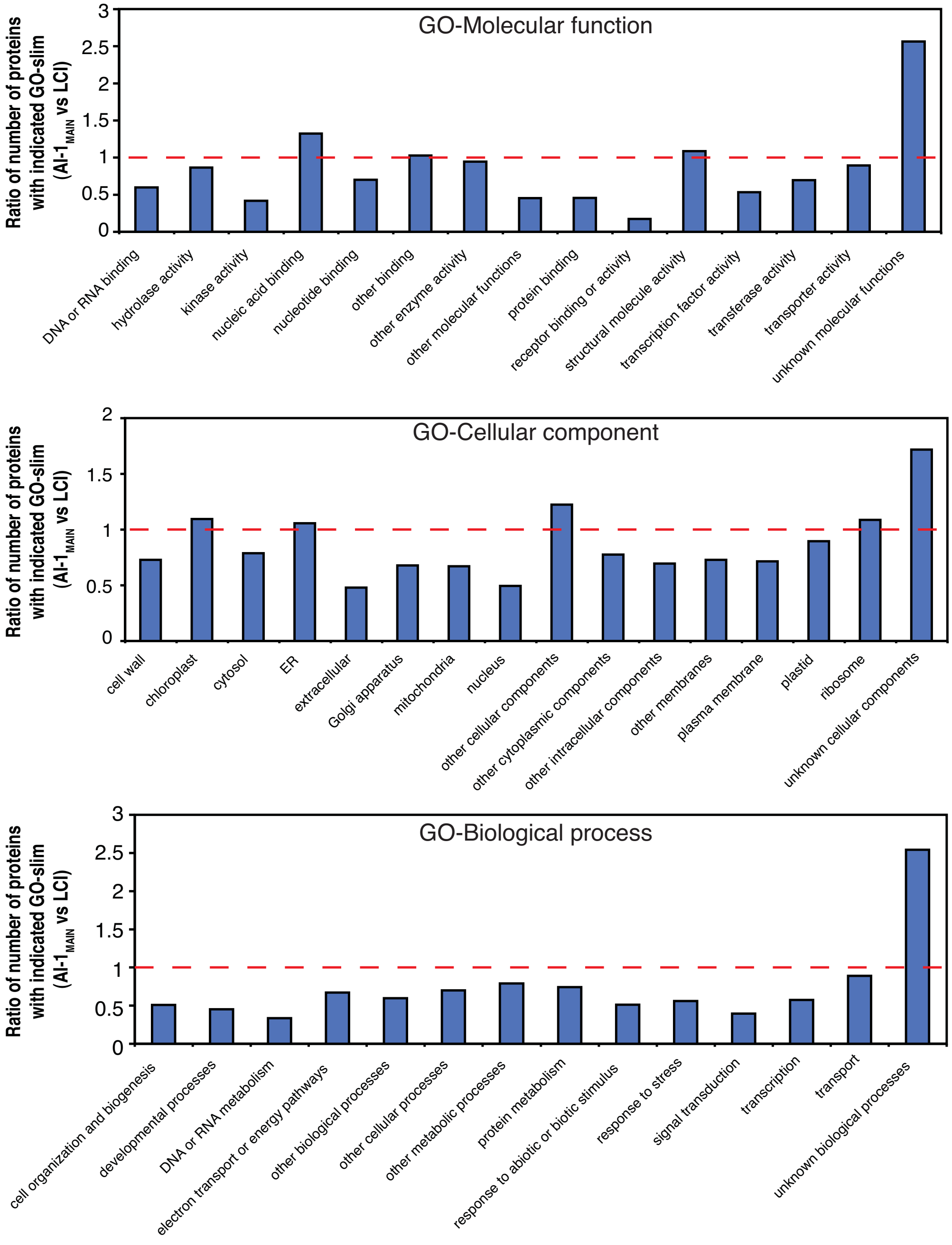


Supplementary Figure 7

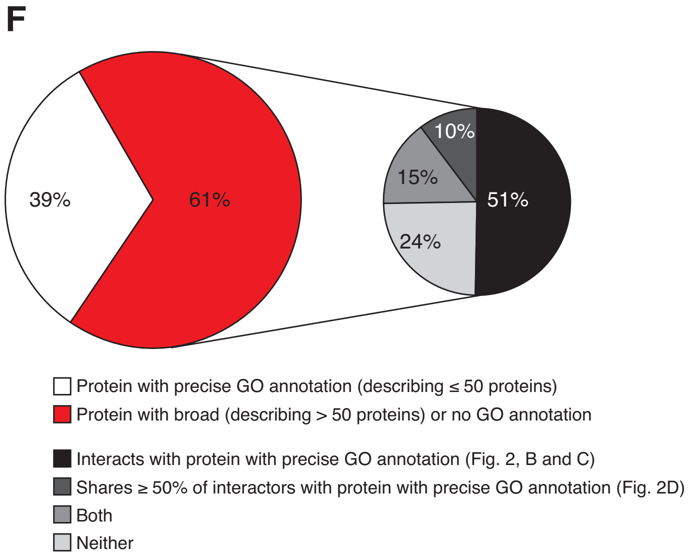


Supplementary Figure 7

E

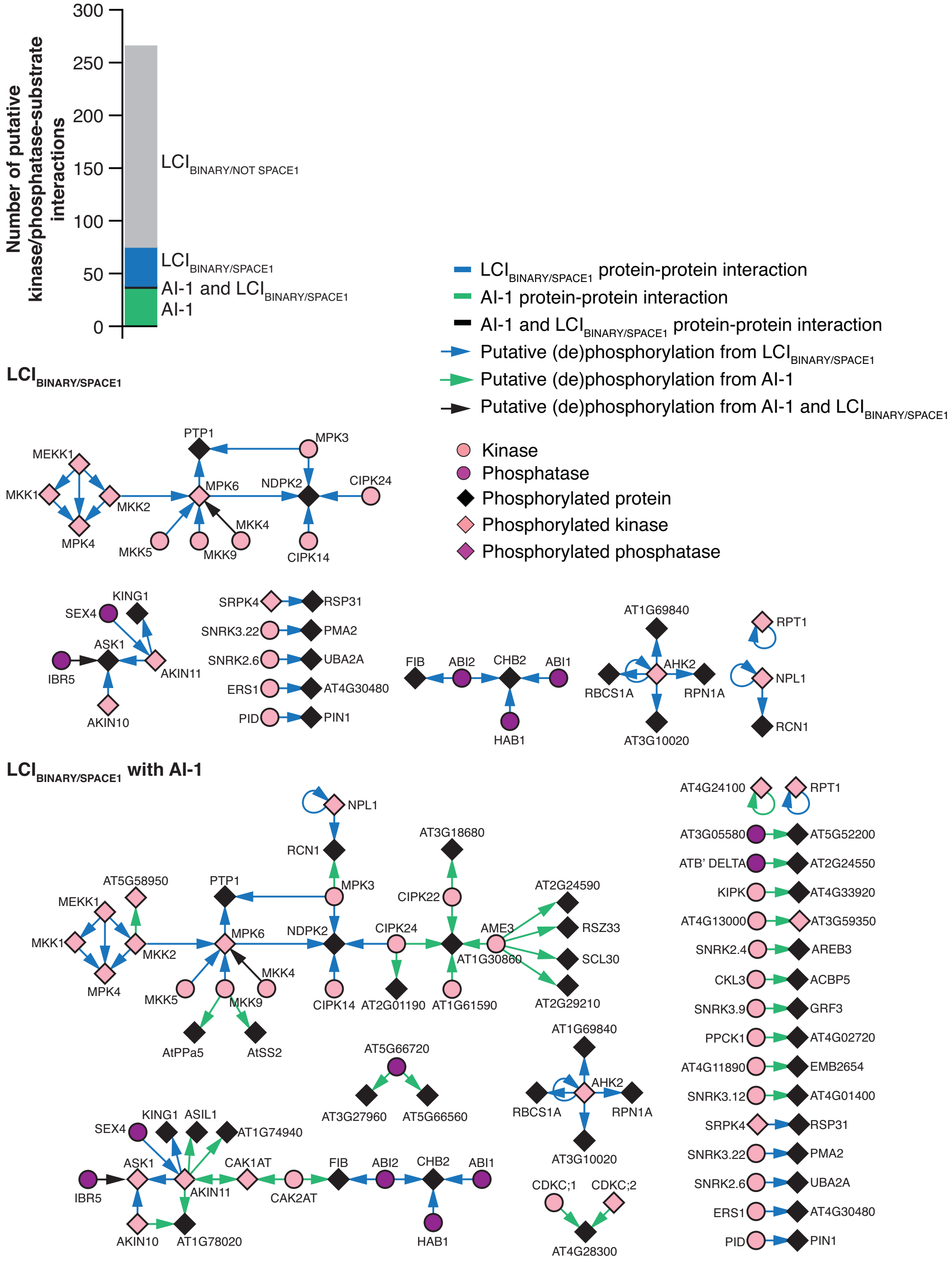


Supplementary Figure 7

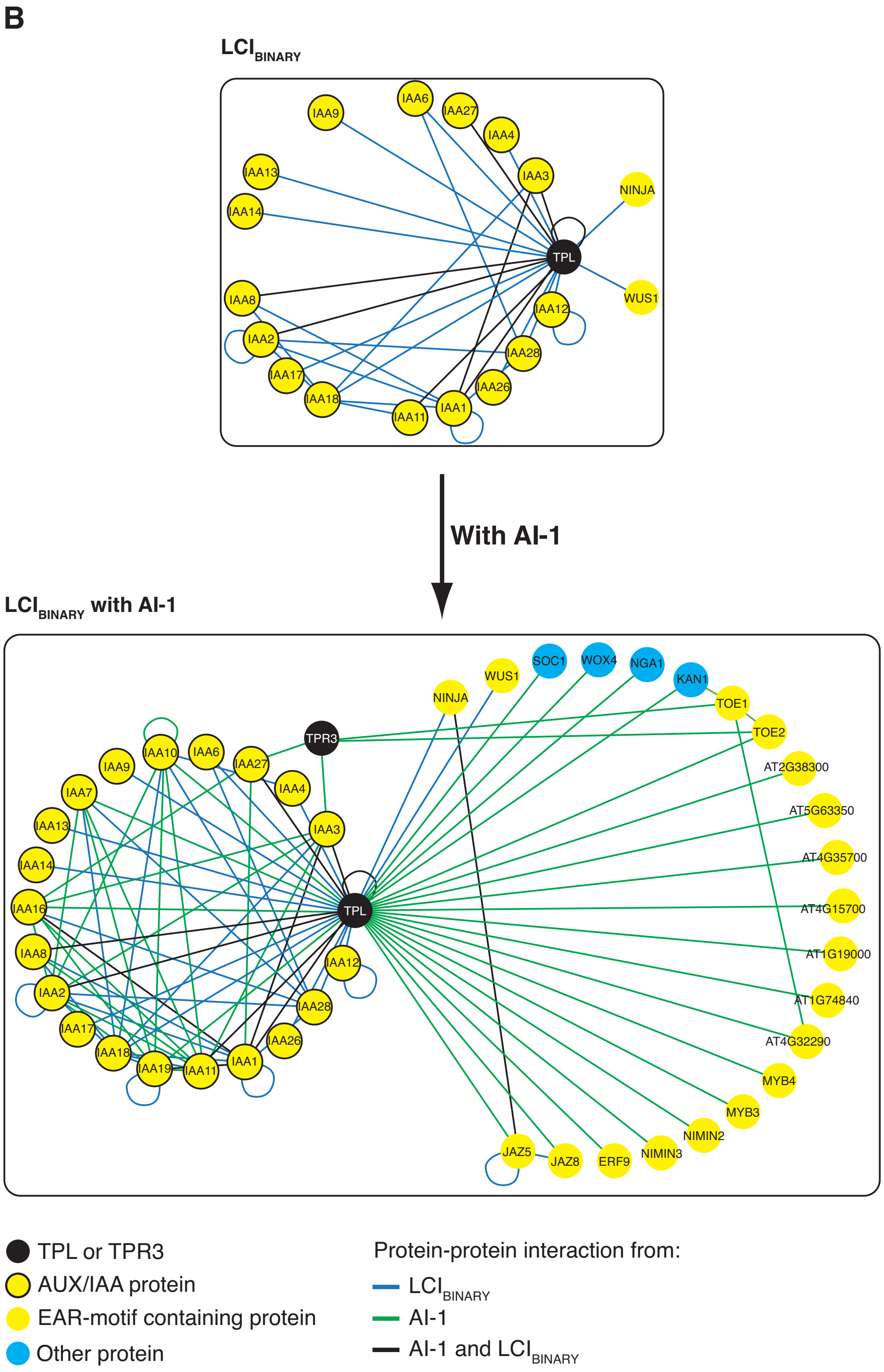


Supplementary Figure 8

A

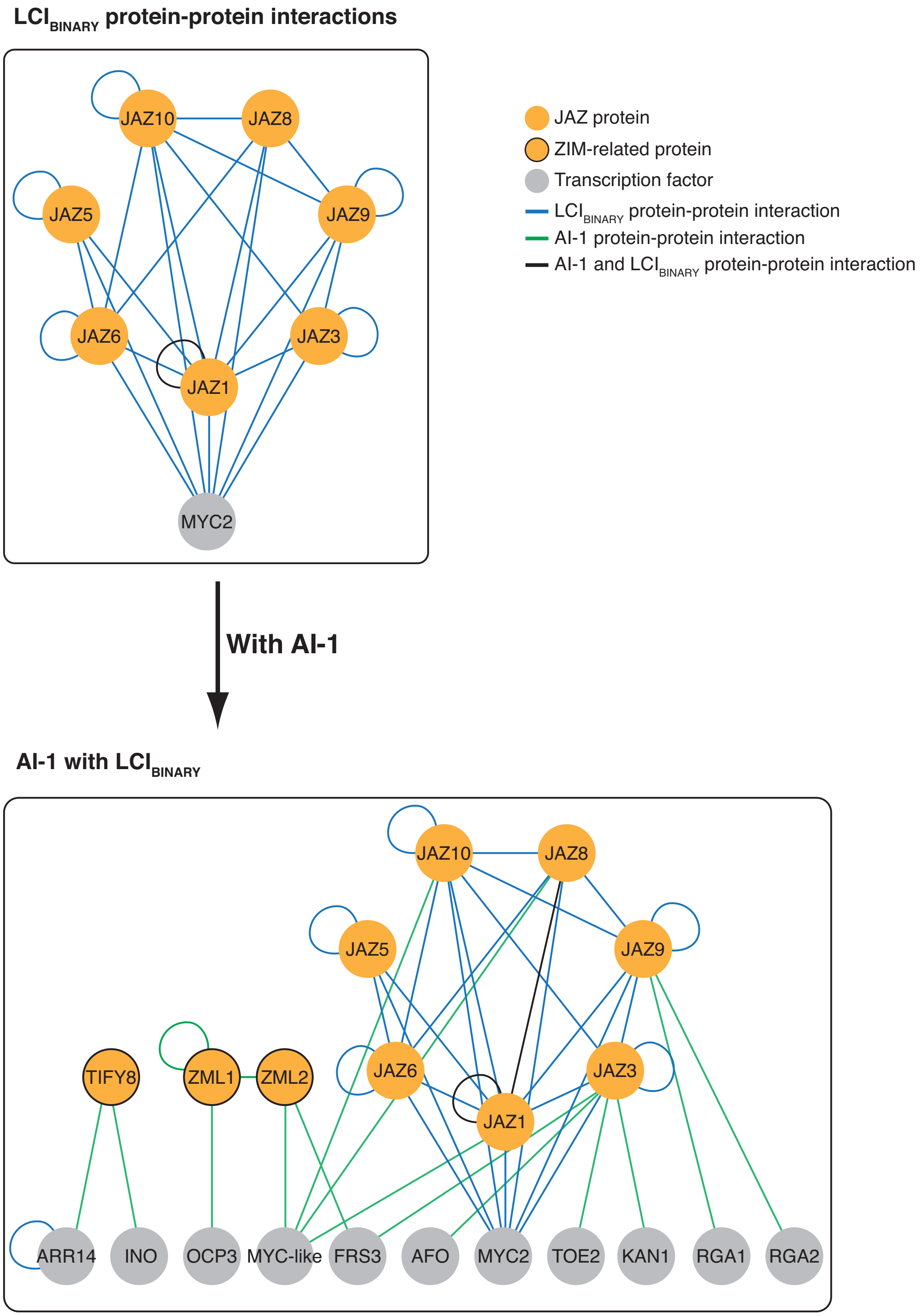


Supplementary Figure 8



Supplementary Figure 8

C



LCI$_{BINARY}$ protein-protein interactions

With AI-1

AI-1 with LCI$_{BINARY}$

● JAZ protein

○ ZIM-related protein

● Transcription factor

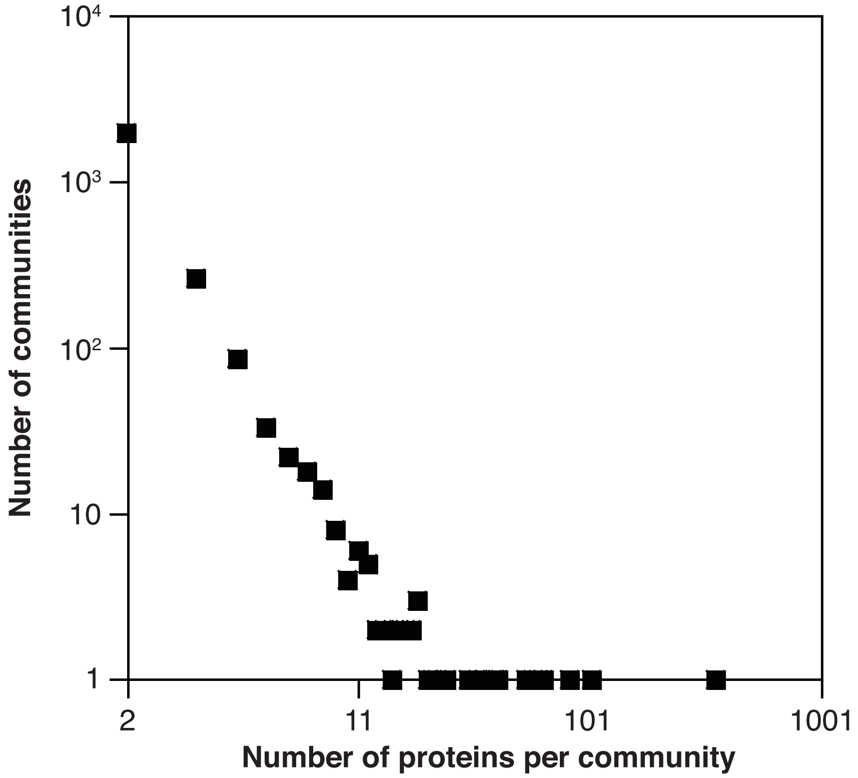
— LCI$_{BINARY}$ protein-protein interaction

— AI-1 protein-protein interaction

— AI-1 and LCI$_{BINARY}$ protein-protein interaction

Supplementary Figure 9

A

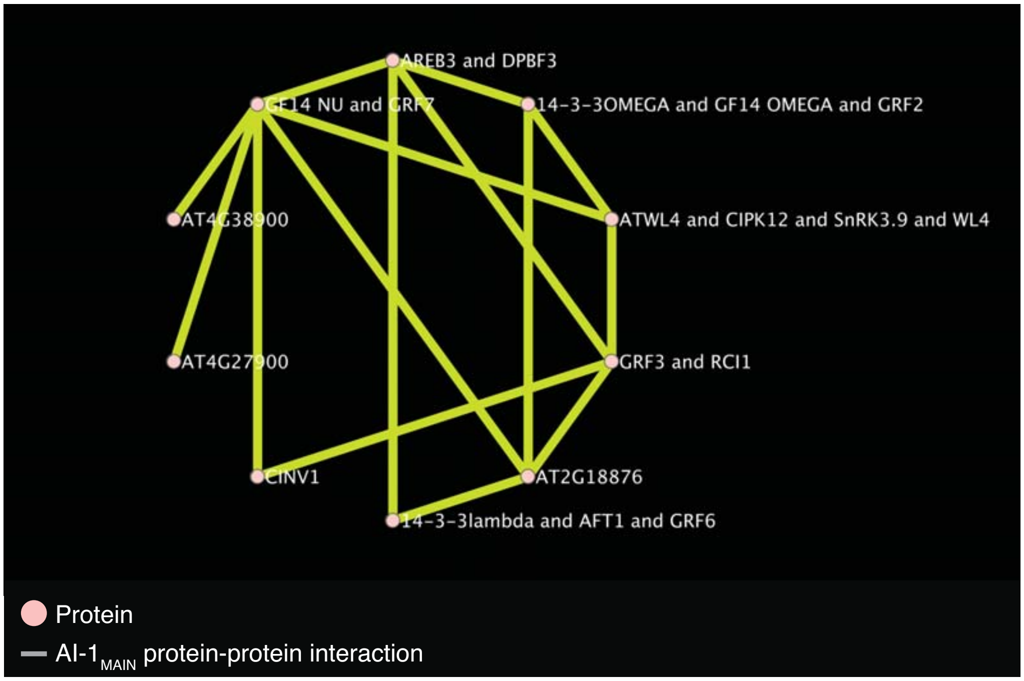


B

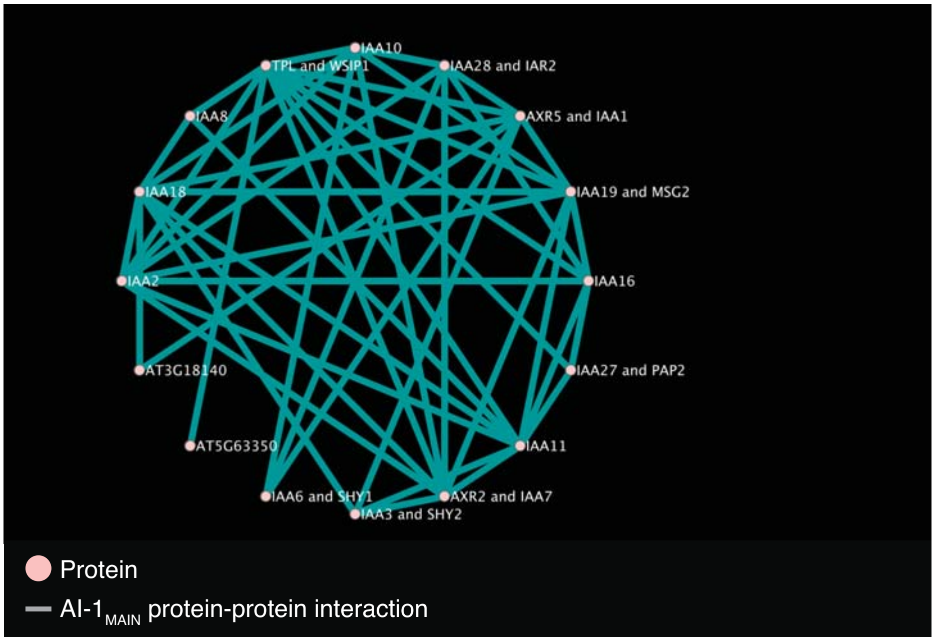
Community > 5 proteins	Density > 0	Density = 0
GO enriched	23	29
Not GO enriched	3	53

Odds ratio = 13 *P* = 0.000002

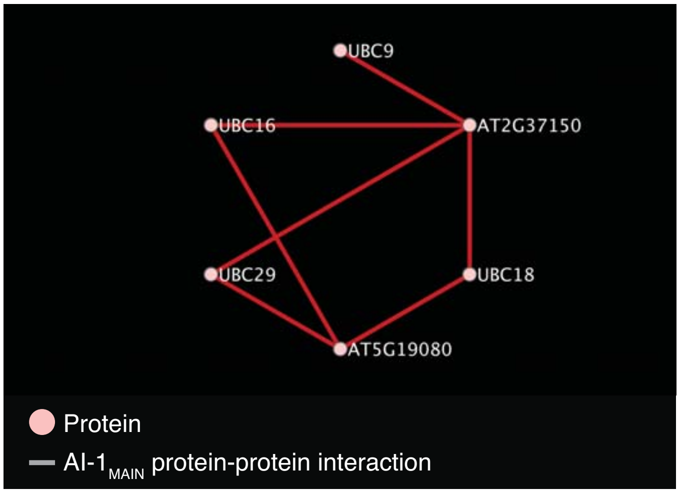
Supplementary Figure 10


AREB3 and DPBF3
GF14 NU and GRF7
14-3-3OMEGA and GF14 OMEGA and GRF2
AT4G38900
ATWL4 and CIPK12 and SnRK3.9 and WL4
AT4G27900
GRF3 and RCI1
CINV1
AT2G18876
14-3-3lambda and AFT1 and GRF6
Protein
AI-1$_{MAIN}$ protein-protein interaction

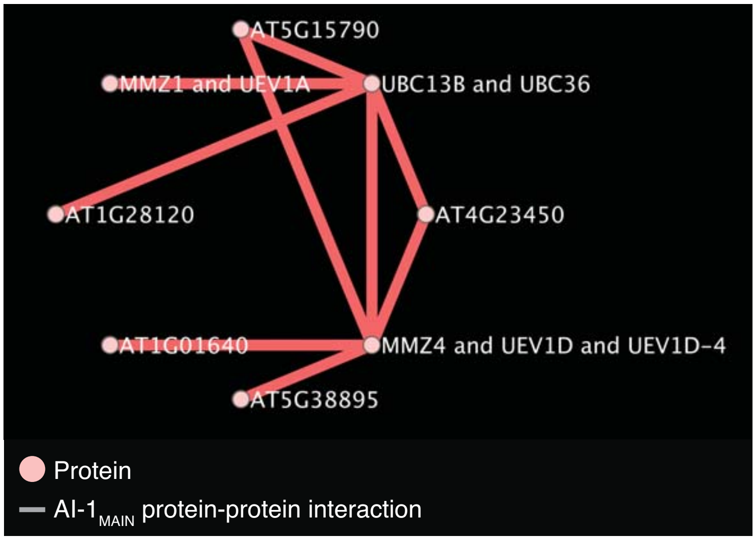
Supplementary Figure 11



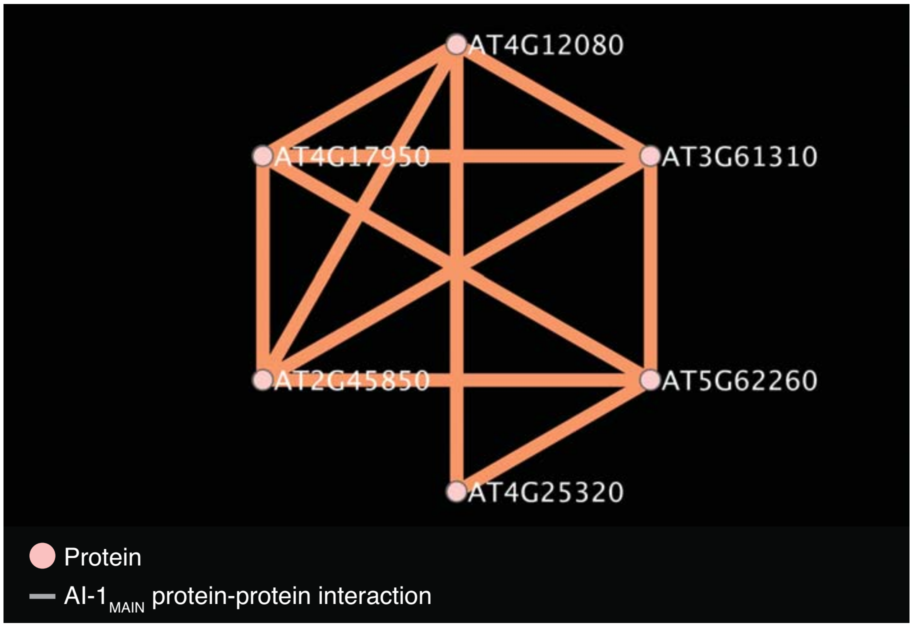
Supplementary Figure 12



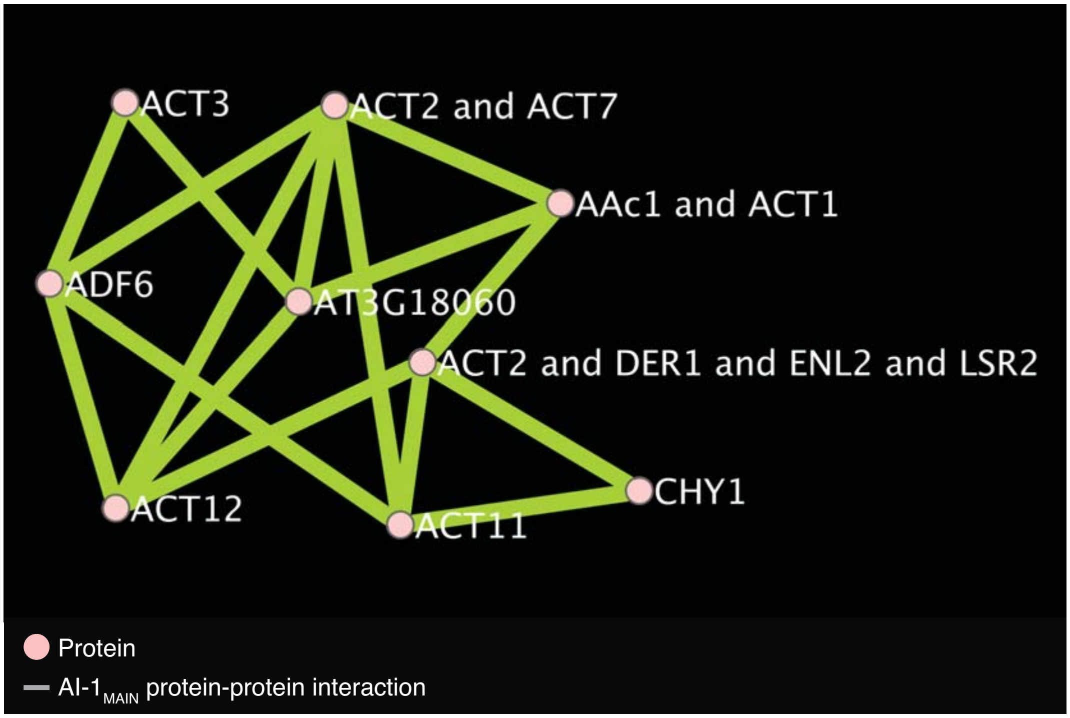
Supplementary Figure 13



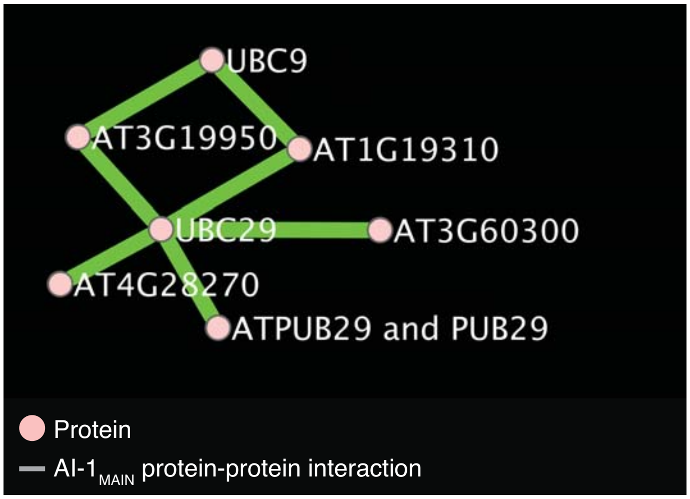
Supplementary Figure 14



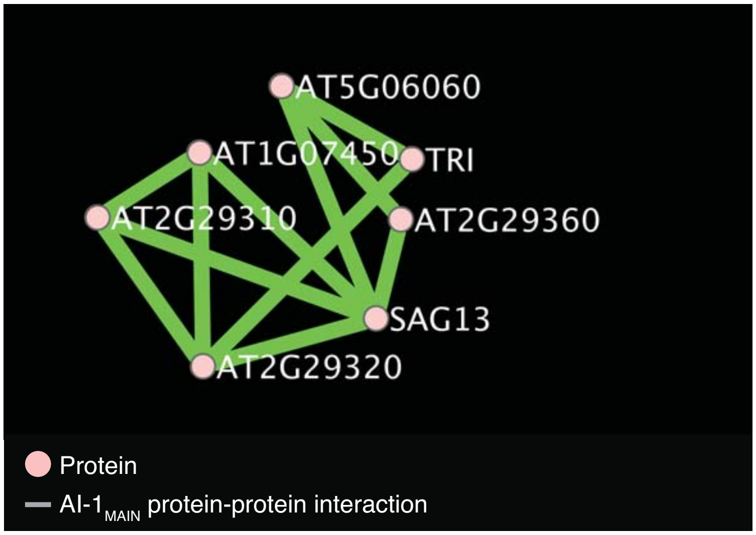
Supplementary Figure 15


ACT3
ACT2 and ACT7
AAc1 and ACT1
ADF6
AT3G18060
ACT2 and DER1 and ENL2 and LSR2
CHY1
ACT12
ACT11
Protein
AI-1$_{MAIN}$ protein-protein interaction

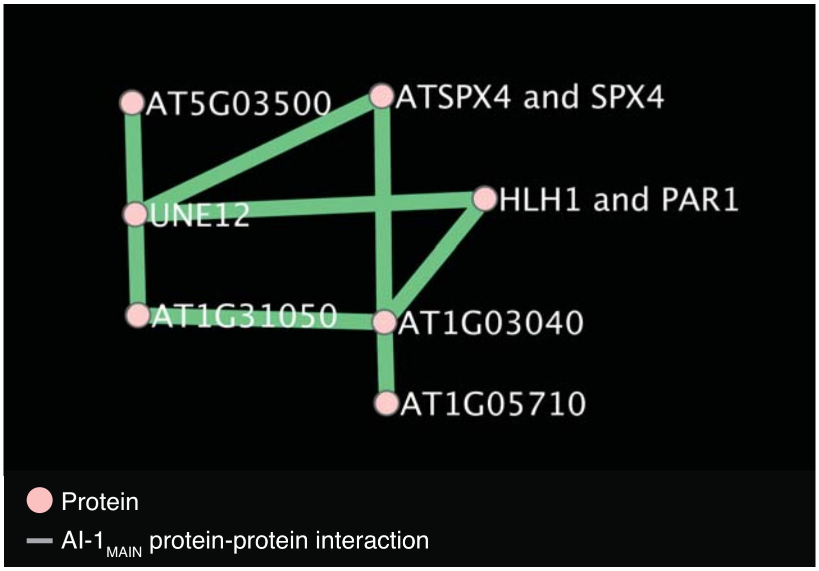
Supplementary Figure 16



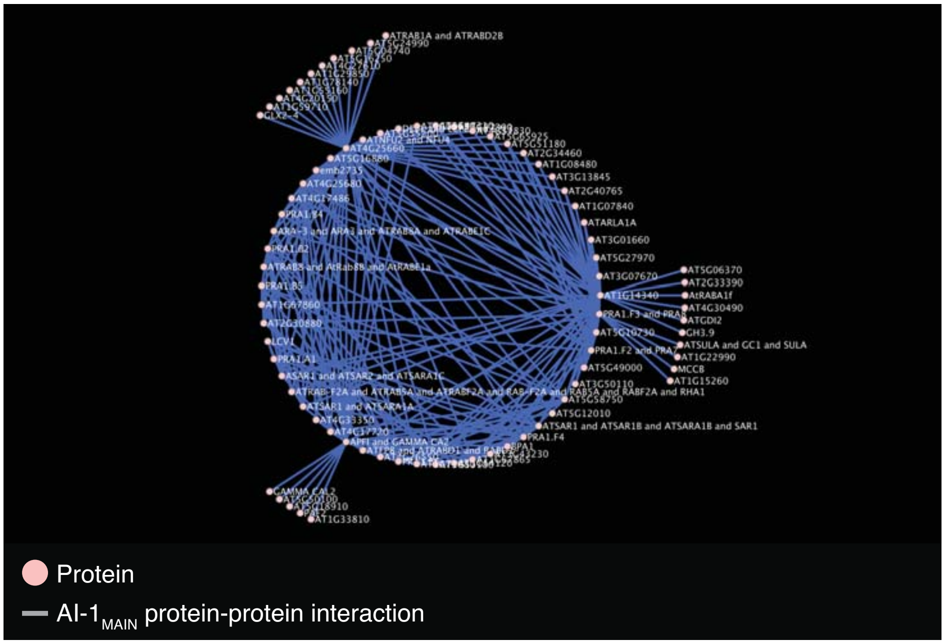
Supplementary Figure 17



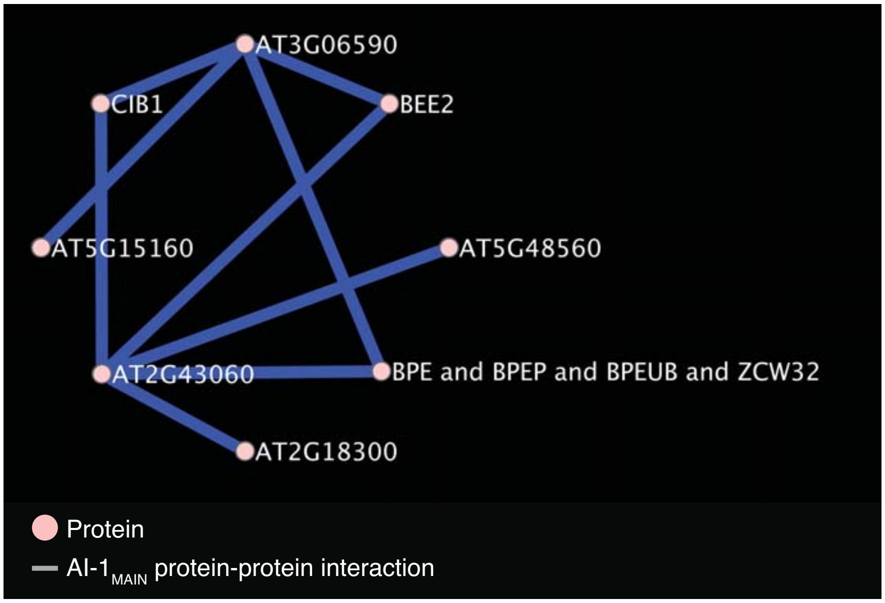
Supplementary Figure 18



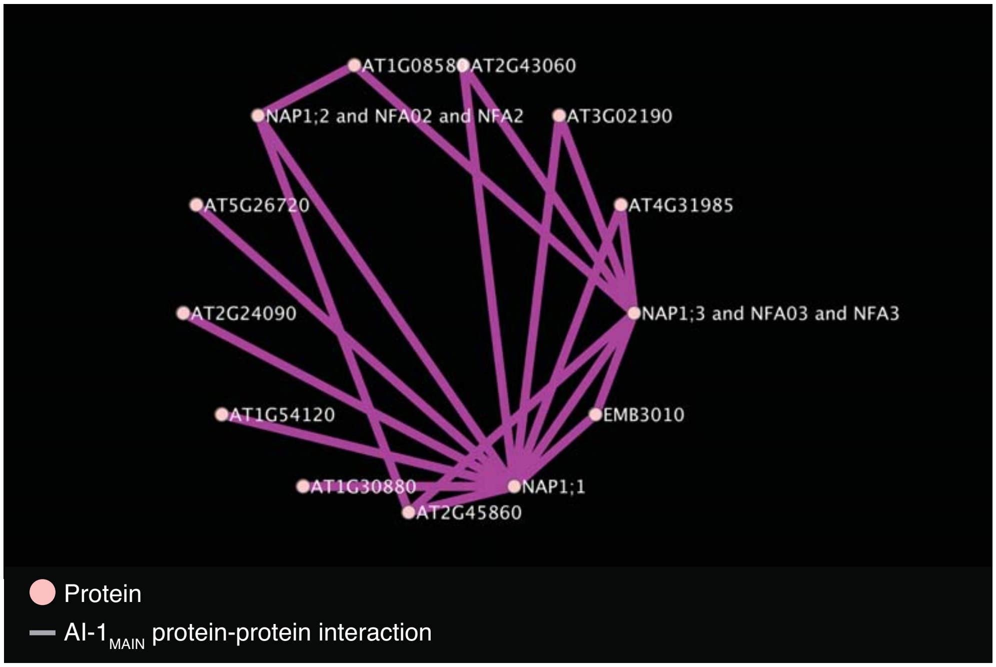
Supplementary Figure 19



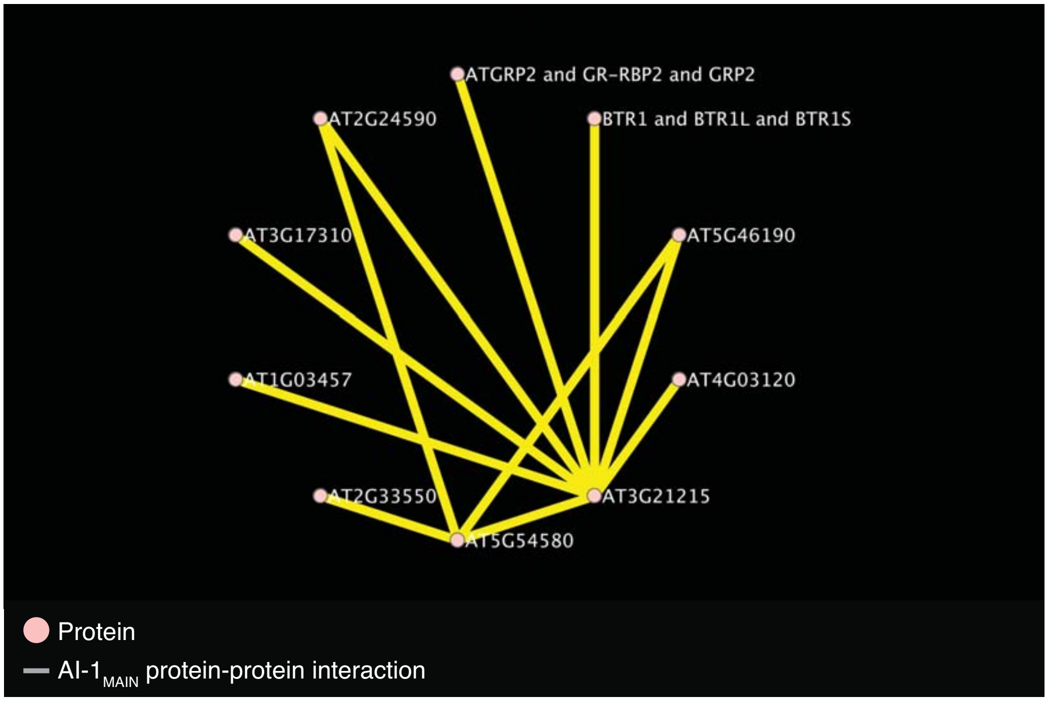
Supplementary Figure 20


AT3G06590
CIB1
BEE2
AT5G15160
AT5G48560
AT2G43060
BPE and BPEP and BPEUB and ZCW32
AT2G18300
● Protein
— AI-1$_{MAIN}$ protein-protein interaction

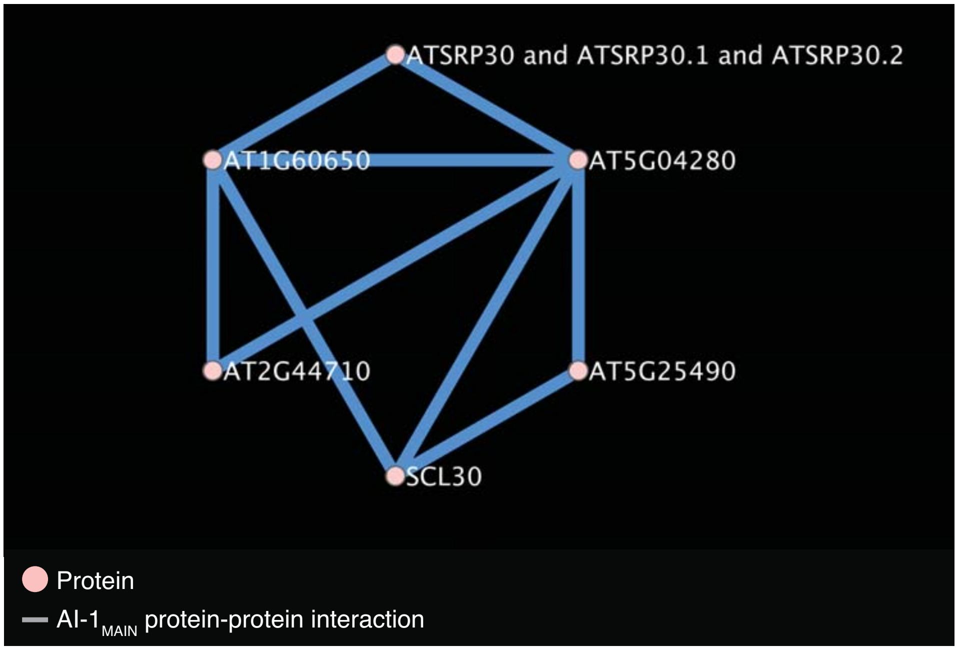
Supplementary Figure 21



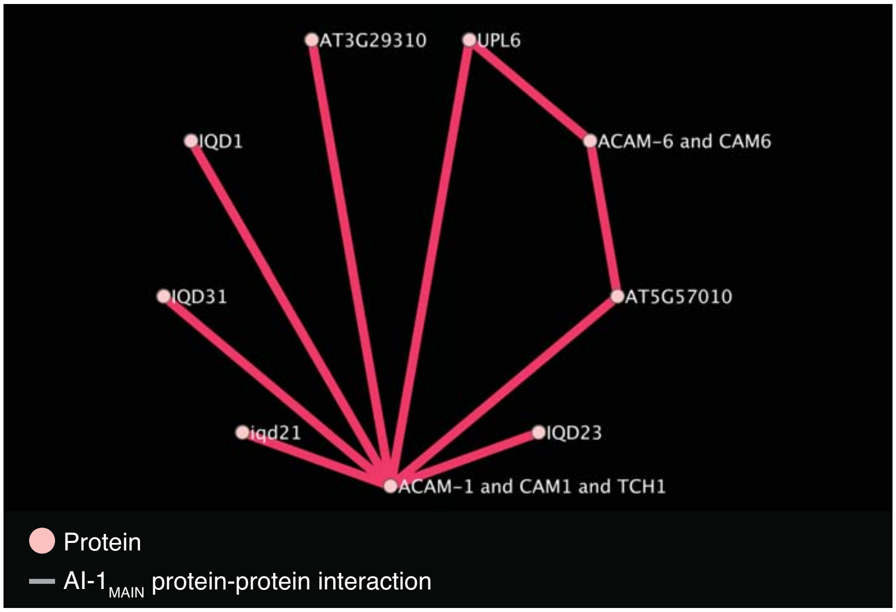
Supplementary Figure 22



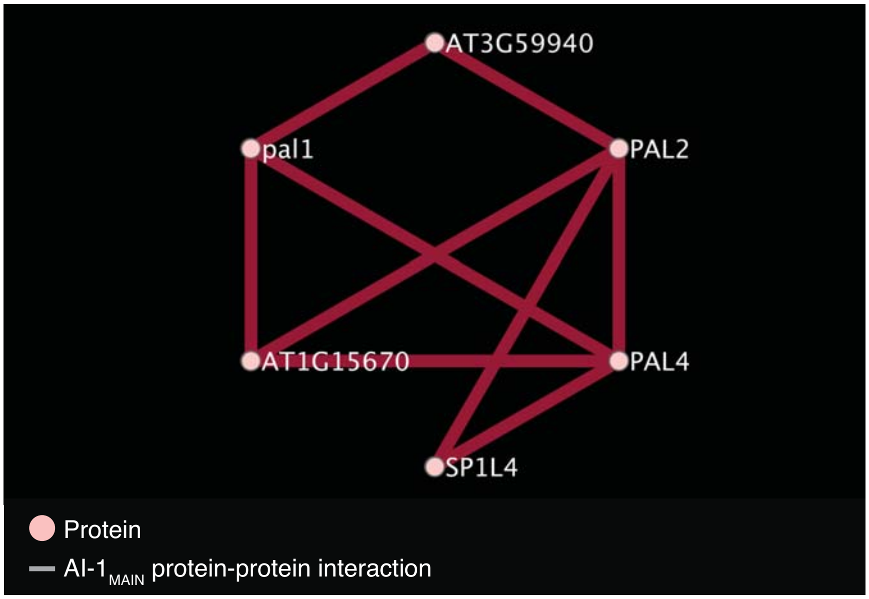
Supplementary Figure 23


ATSRP30 and ATSRP30.1 and ATSRP30.2
AT1G60650
AT5G04280
AT2G44710
AT5G25490
SCL30
● Protein
— AI-1$_{MAIN}$ protein-protein interaction

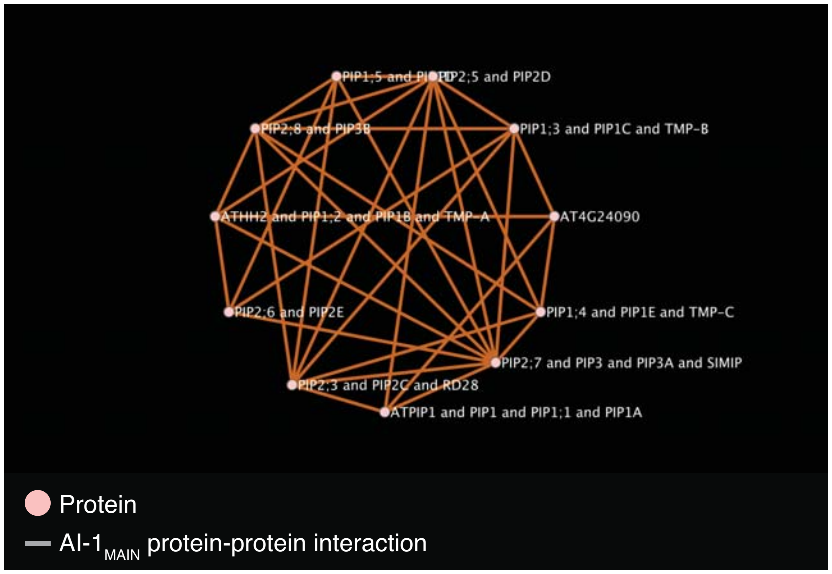
Supplementary Figure 24



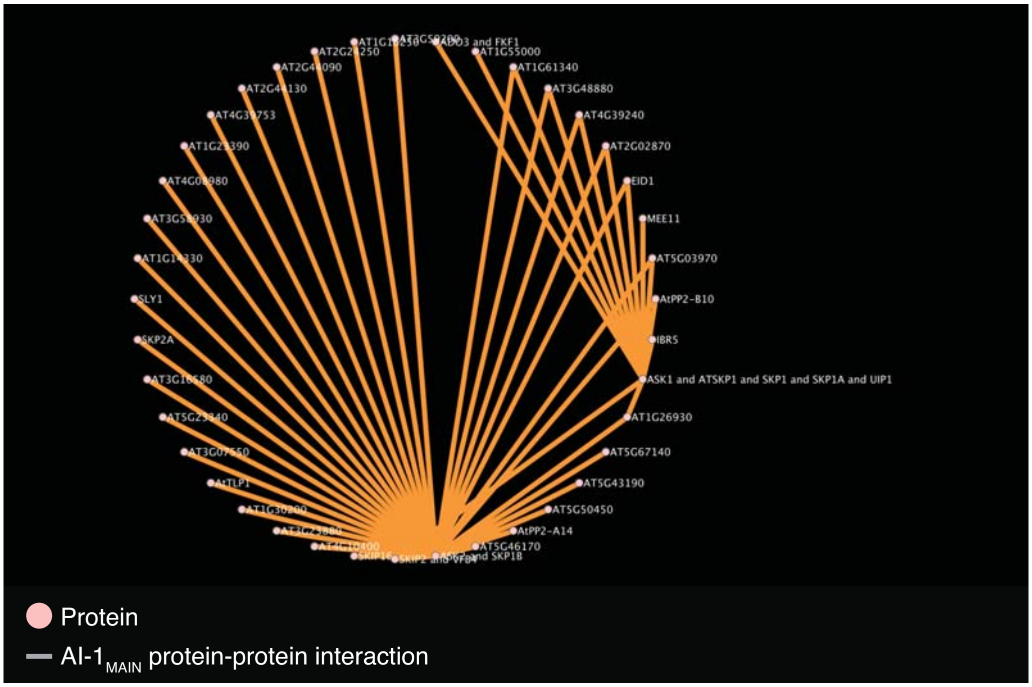
Supplementary Figure 25



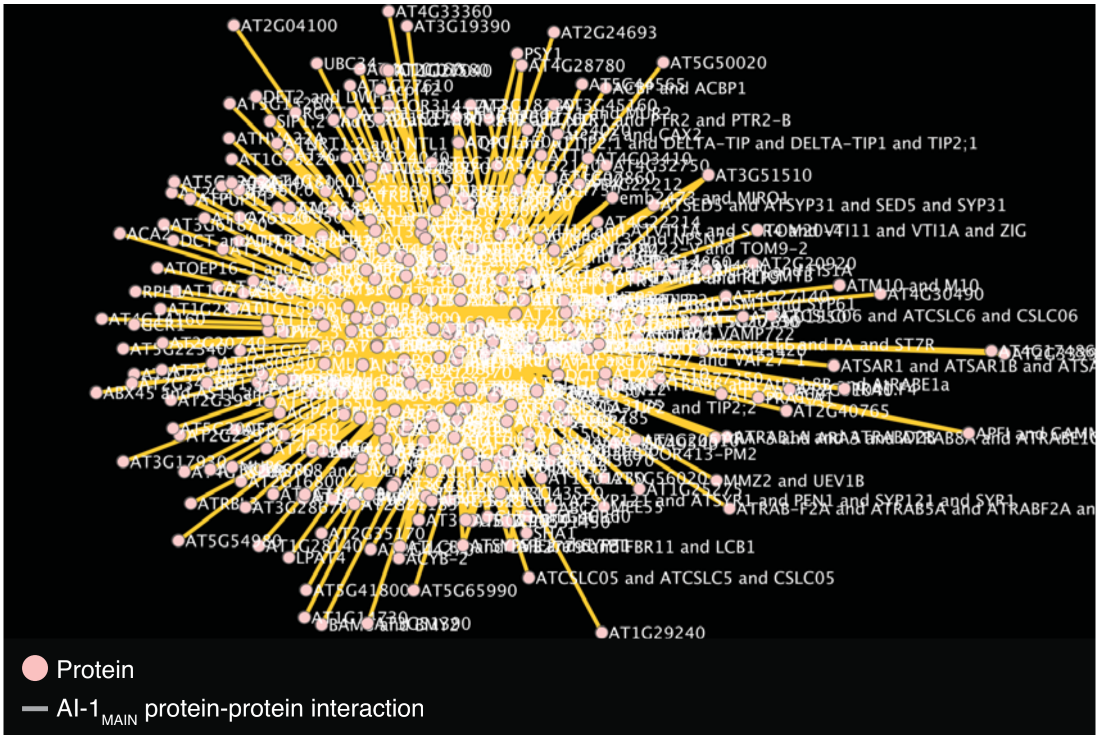
Supplementary Figure 26


PIP1;5 and PI
PIP2;5 and PIP2D
PIP2;8 and PIP3B
PIP1;3 and PIP1C and TMP-B
ATHH2 and PIP1;2 and PIP1B and TMP-A
AT4G24090
PIP2;6 and PIP2E
PIP1;4 and PIP1E and TMP-C
PIP2;7 and PIP3 and PIP3A and SIMIP
PIP2;3 and PIP2C and RD28
ATPIP1 and PIP1 and PIP1;1 and PIP1A
● Protein
— AI-1$_{MAIN}$ protein-protein interaction

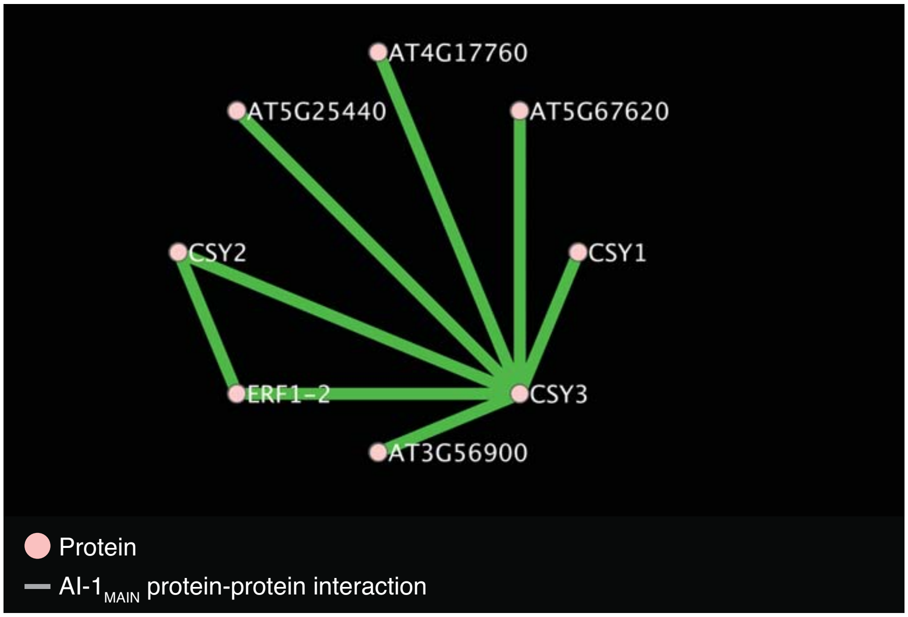
Supplementary Figure 27



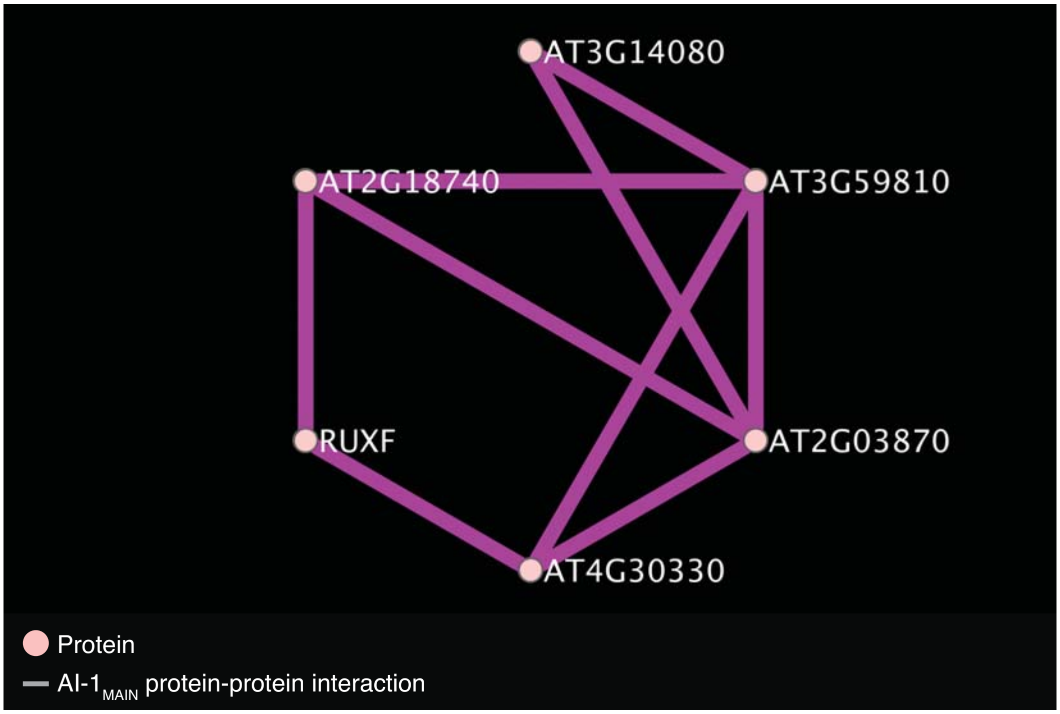
Supplementary Figure 28



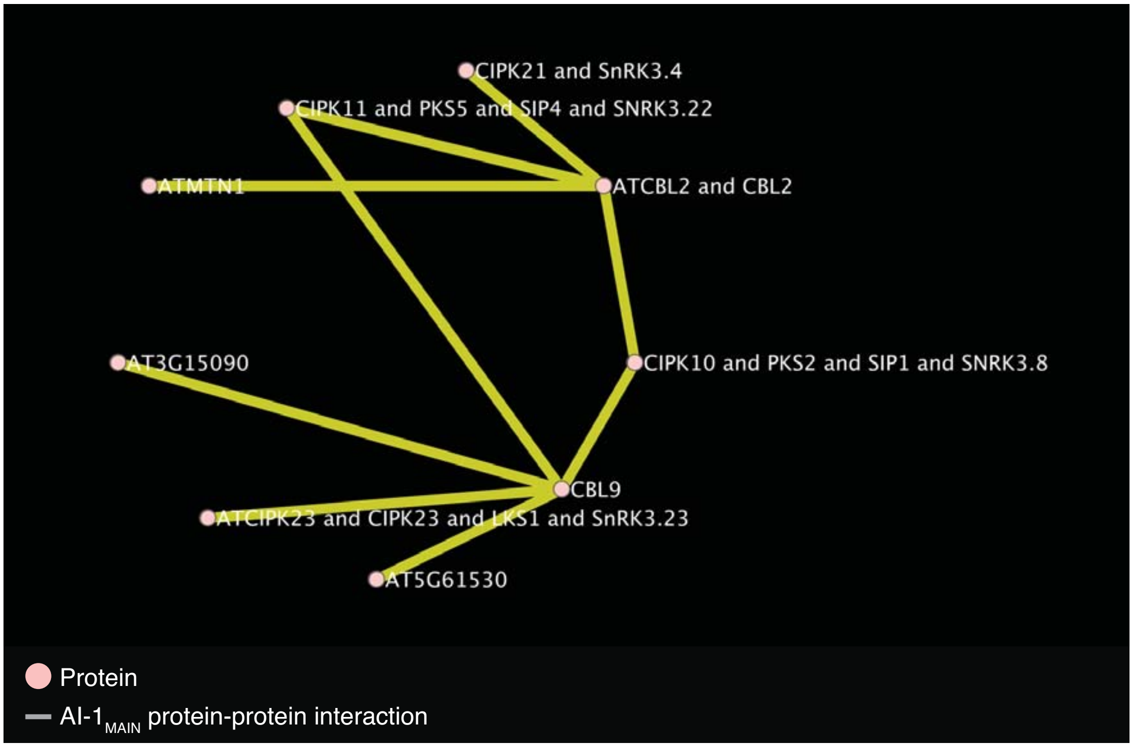
Supplementary Figure 29



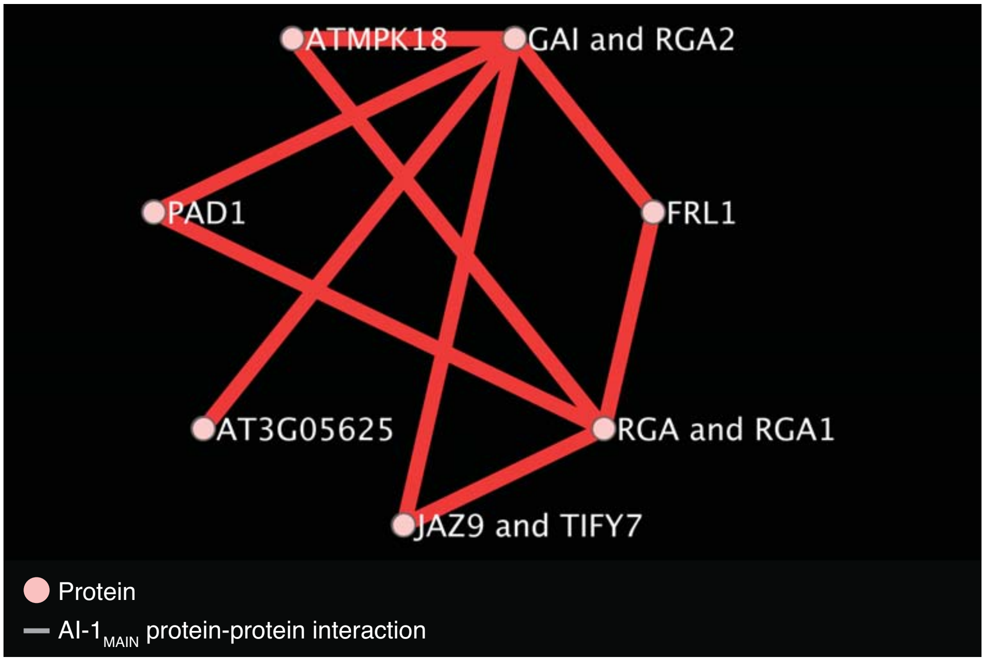
Supplementary Figure 30



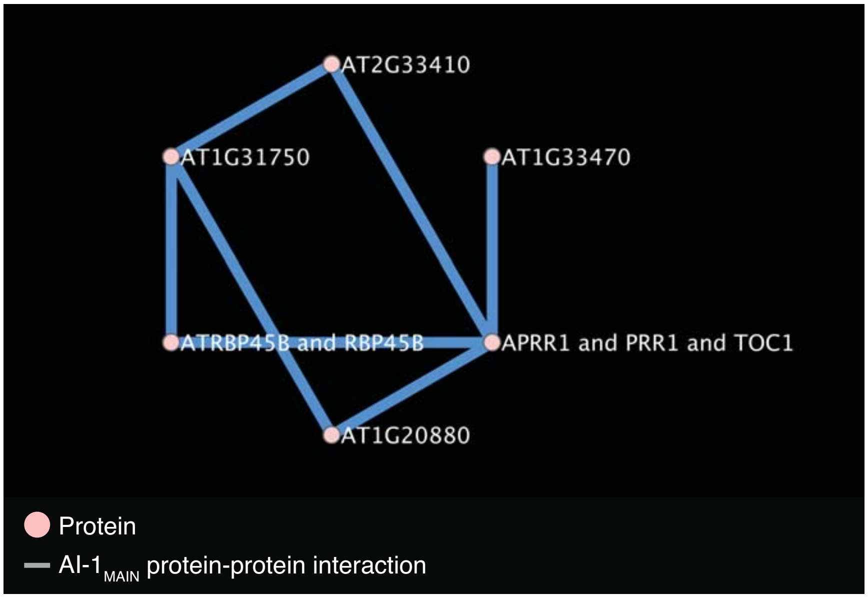
Supplementary Figure 31


CIPK21 and SnRK3.4
CIPK11 and PKS5 and SIP4 and SNRK3.22
ATMTN1
ATCBL2 and CBL2
AT3G15090
CIPK10 and PKS2 and SIP1 and SNRK3.8
CBL9
ATCIPK23 and CIPK23 and LKS1 and SnRK3.23
AT5G61530
● Protein
— AI-1$_{MAIN}$ protein-protein interaction

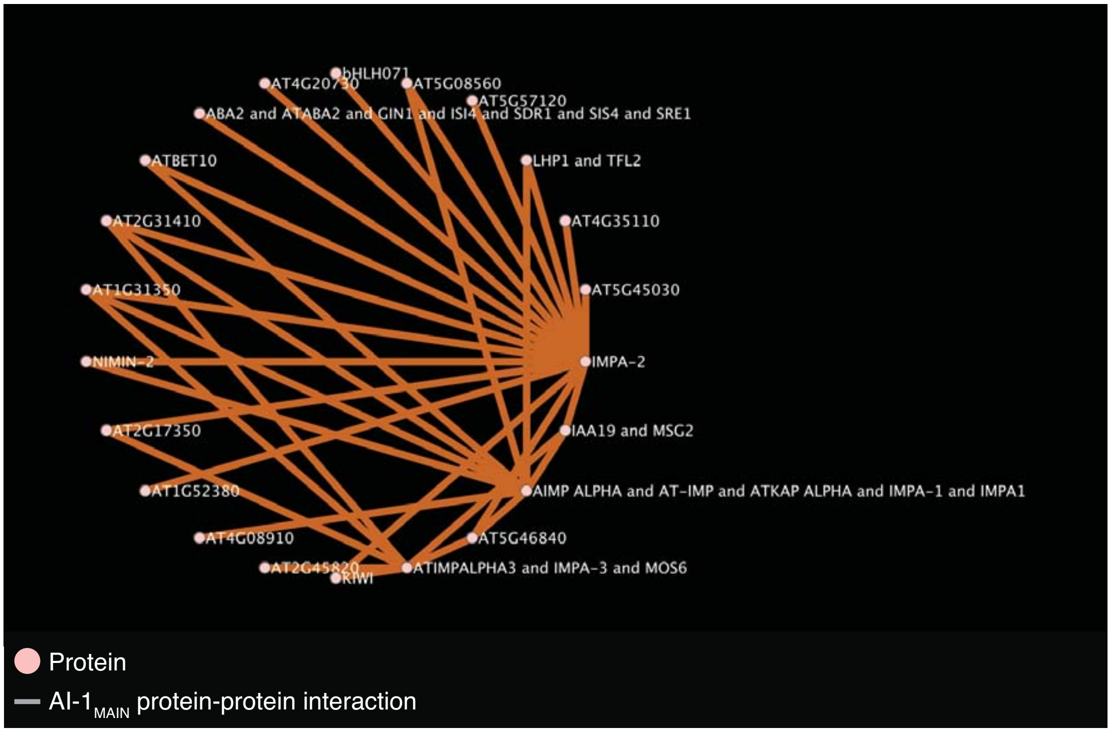
Supplementary Figure 32



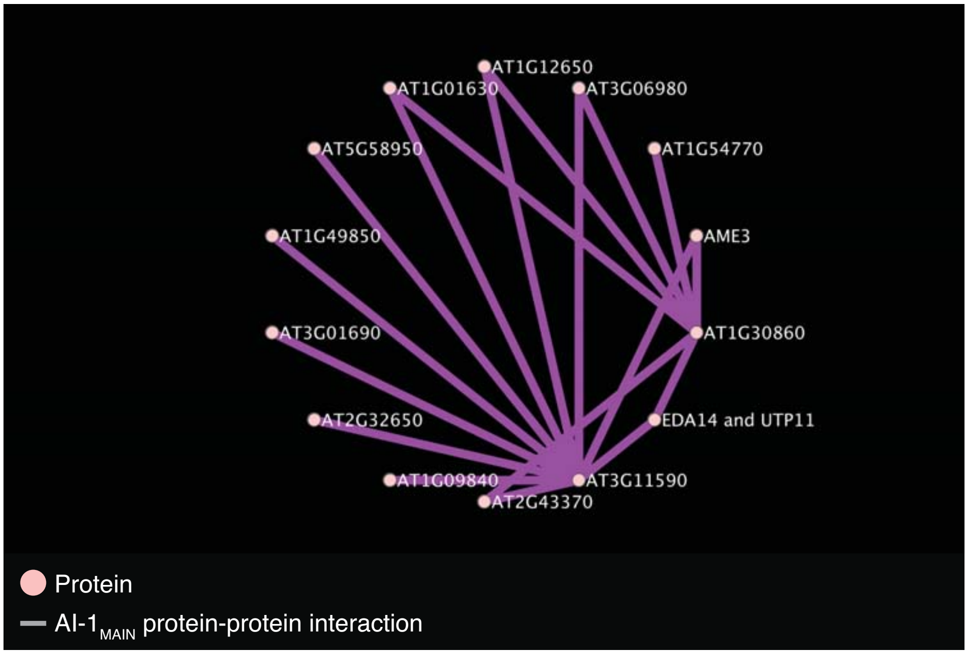
Supplementary Figure 33



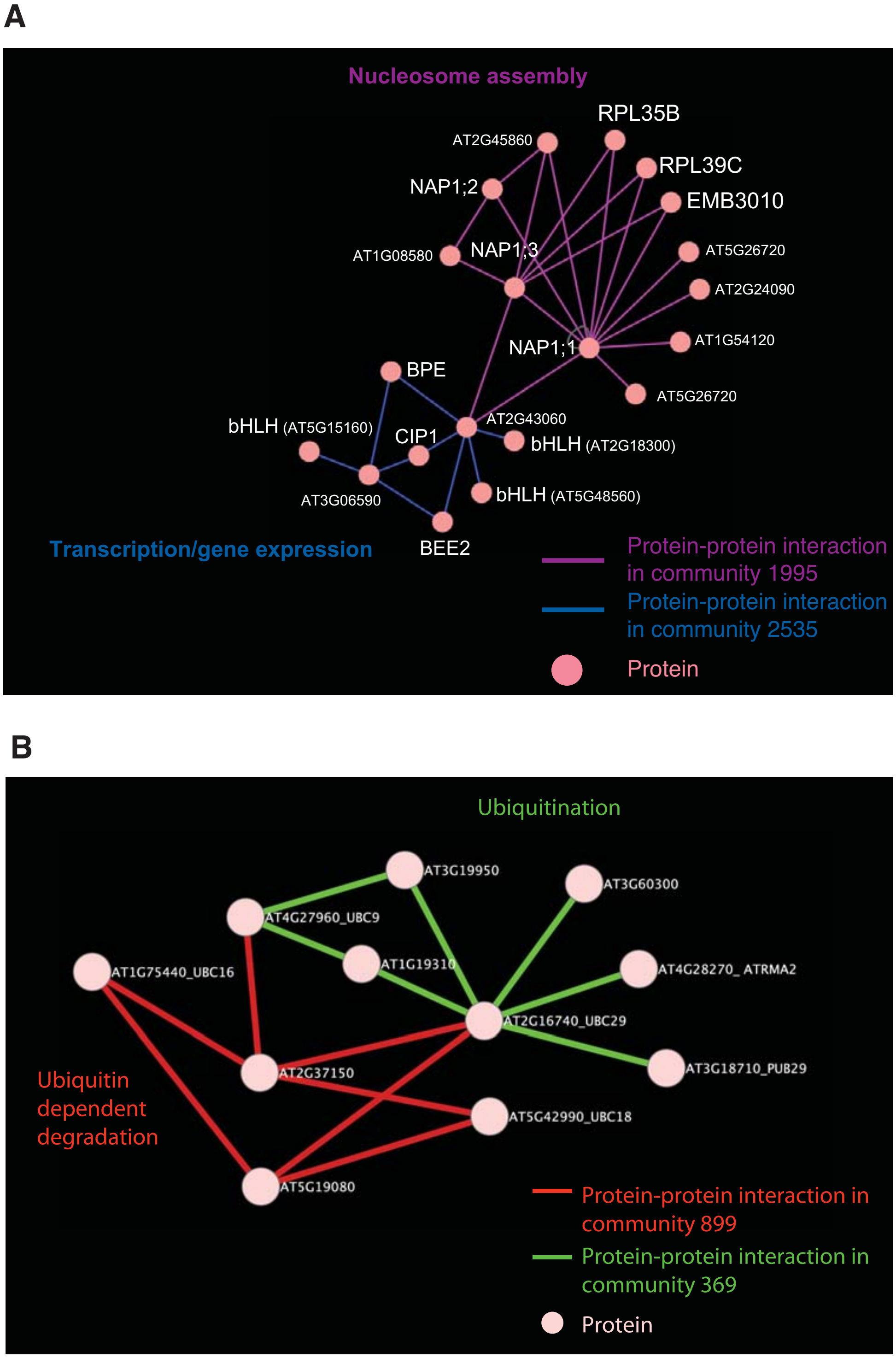
Supplementary Figure 34


bHLH071
AT4G20730
AT5G08560
AT5G57120
ABA2 and ATABA2 and GIN1 and ISI4 and SDR1 and SIS4 and SRE1
ATBET10
LHP1 and TFL2
AT2G31410
AT4G35110
AT1G31350
AT5G45030
NIMIN-2
IMPA-2
AT2G17350
IAA19 and MSG2
AT1G52380
AIMP ALPHA and AT-IMP and ATKAP ALPHA and IMPA-1 and IMPA1
AT4G08910
AT5G46840
AT2G45820
KIWI
ATIMPALPHA3 and IMPA-3 and MOS6
● Protein
— AI-1$_{MAIN}$ protein-protein interaction

Supplementary Figure 35

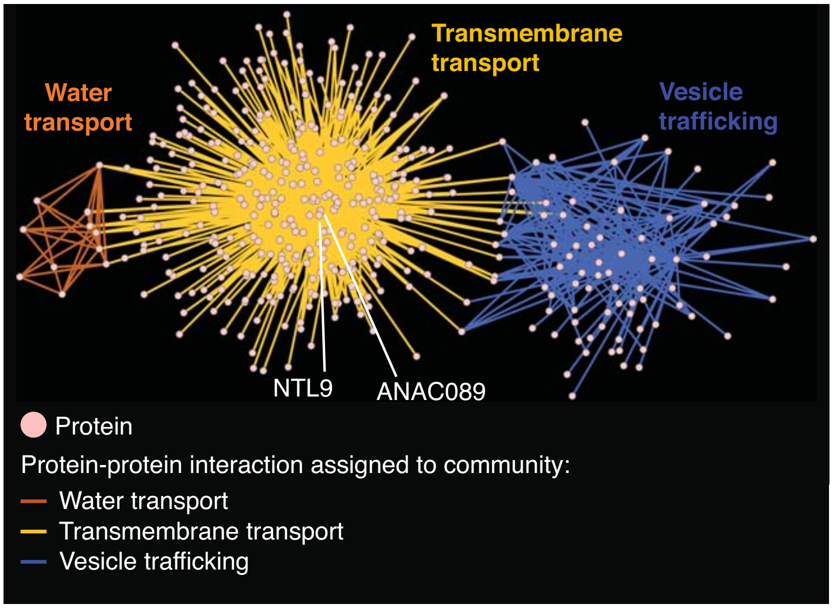


Supplementary Figure 36

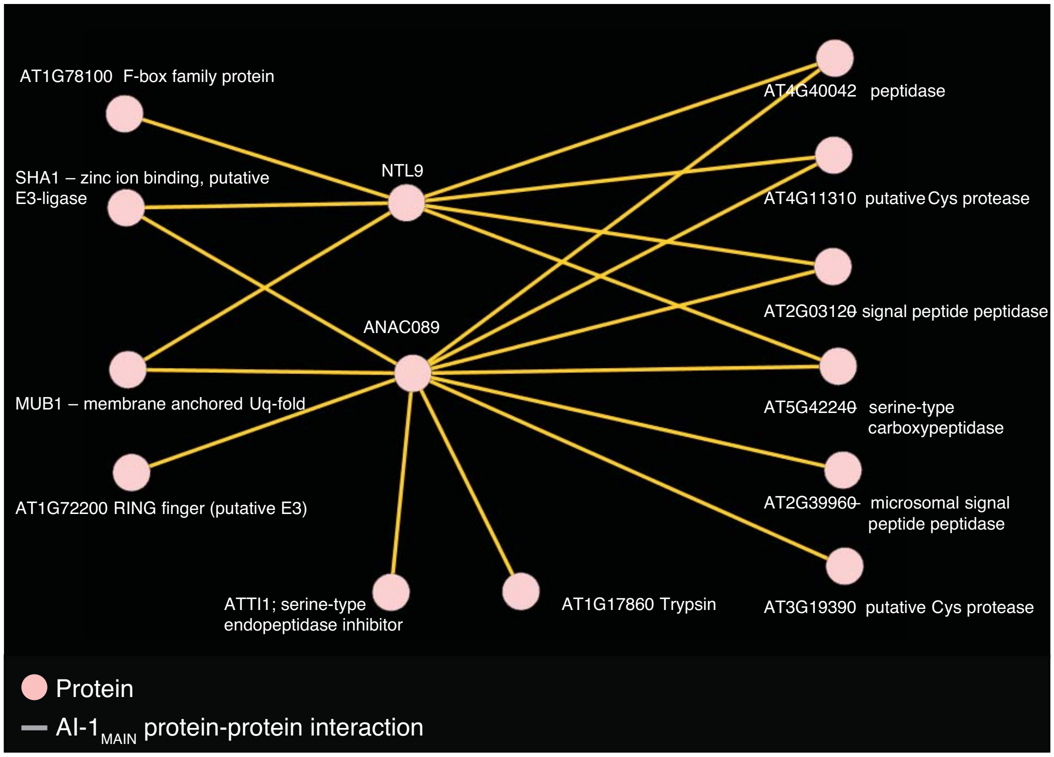


Supplementary Figure 36

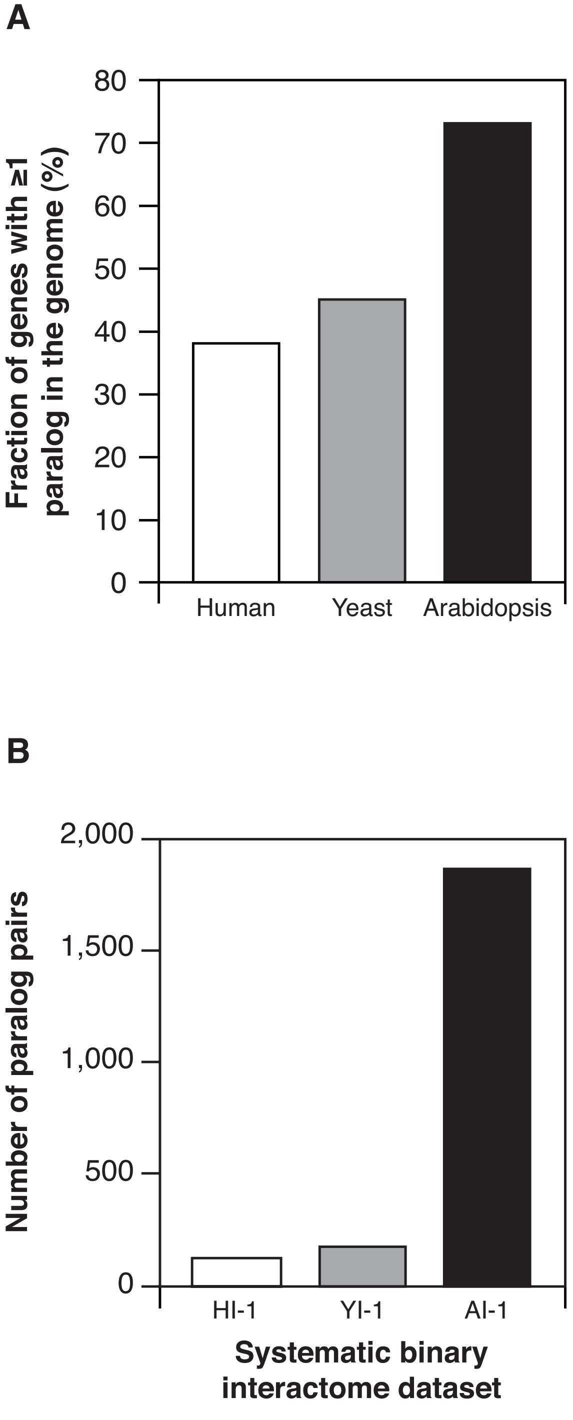
C



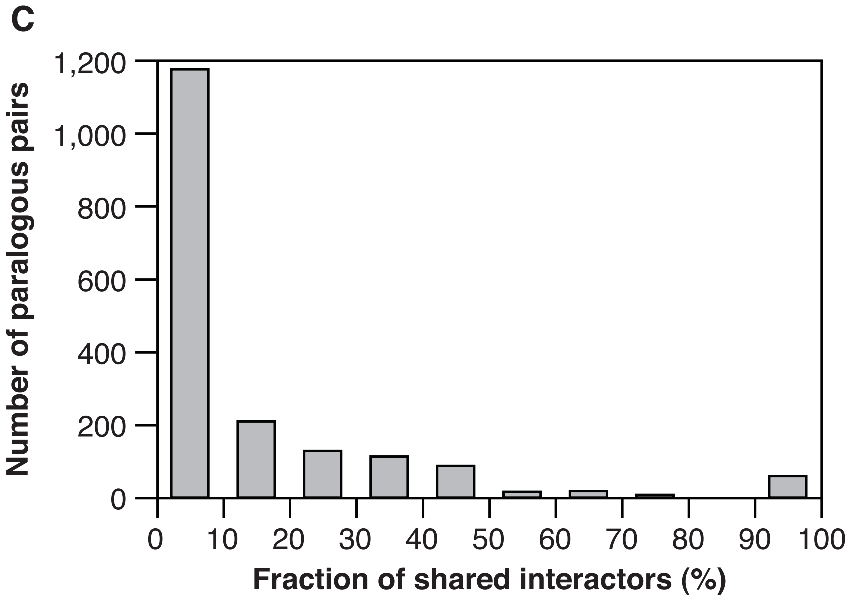
Supplementary Figure 37


AT1G78100 F-box family protein
SHA1 – zinc ion binding, putative E3-ligase
NTL9
ANAC089
MUB1 – membrane anchored Uq-fold
AT1G72200 RING finger (putative E3)
ATTI1; serine-type endopeptidase inhibitor
AT1G17860 Trypsin
AT4G40042 peptidase
AT4G11310 putative Cys protease
AT2G03120 signal peptide peptidase
AT5G42240 serine-type carboxypeptidase
AT2G39960 microsomal signal peptide peptidase
AT3G19390 putative Cys protease
● Protein
— AI-1$_{MAIN}$ protein-protein interaction

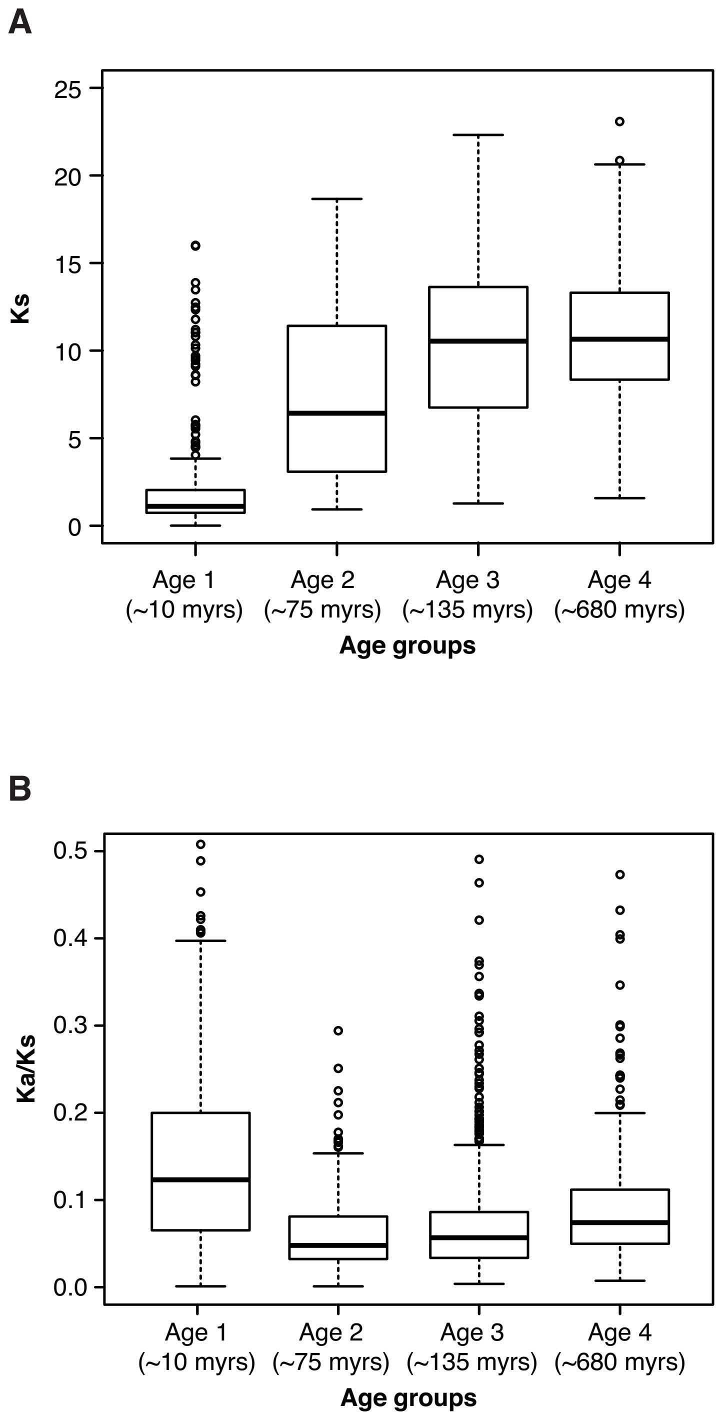
Supplementary Figure 38



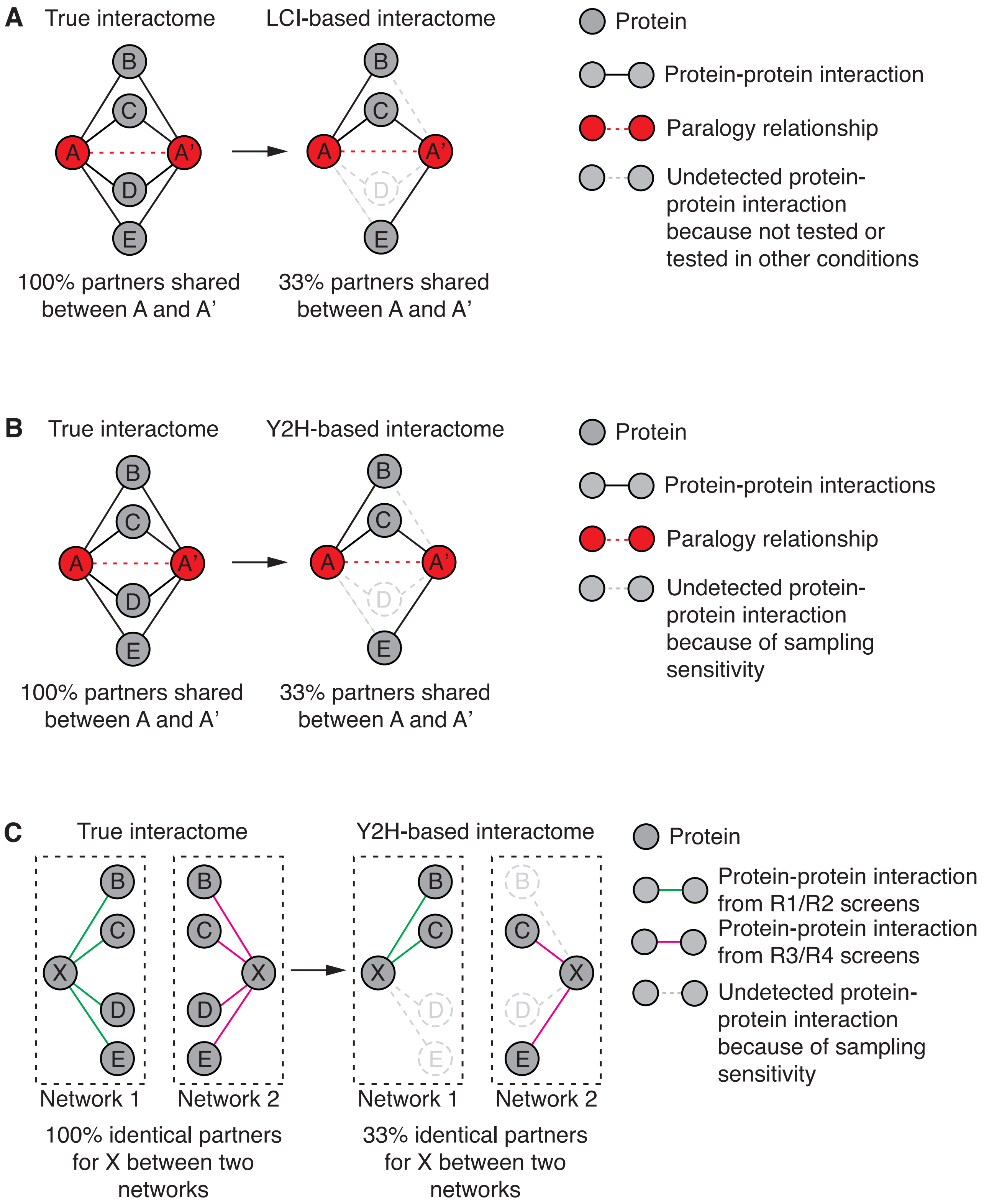
Supplementary Figure 38



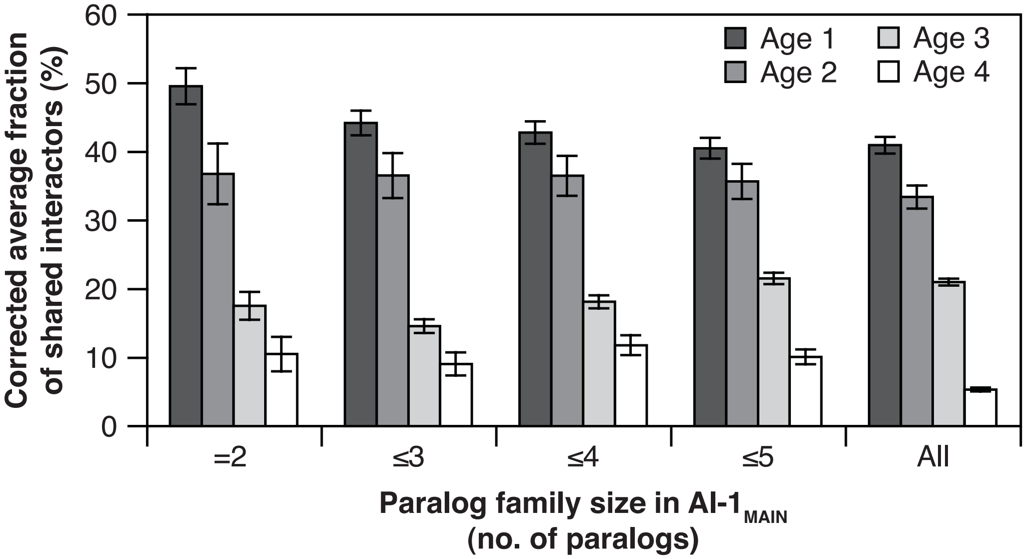
Supplementary Figure 39



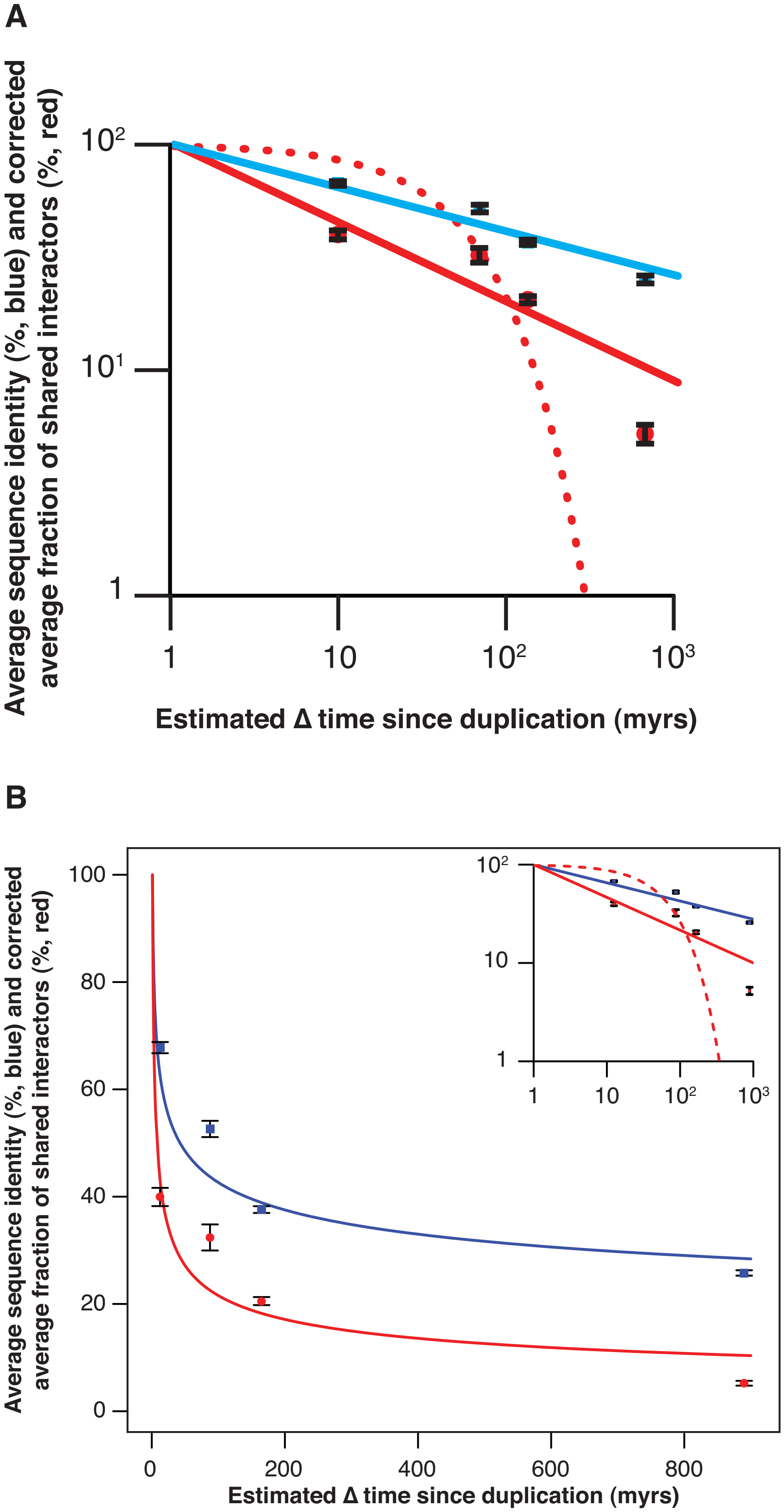
Supplementary Figure 40


A
True interactome
LCI-based interactome
B
C
A
A'
D
E
100% partners shared between A and A'
33% partners shared between A and A'
Protein
Protein-protein interaction
Paralogy relationship
Undetected protein-protein interaction because not tested or tested in other conditions
B
True interactome
Y2H-based interactome
B
C
A
A'
D
E
100% partners shared between A and A'
33% partners shared between A and A'
Protein
Protein-protein interactions
Paralogy relationship
Undetected protein-protein interaction because of sampling sensitivity
C
True interactome
Y2H-based interactome
B
C
X
D
E
Network 1
Network 2
100% identical partners for X between two networks
33% identical partners for X between two networks
Protein
Protein-protein interaction from R1/R2 screens
Protein-protein interaction from R3/R4 screens
Undetected protein-protein interaction because of sampling sensitivity

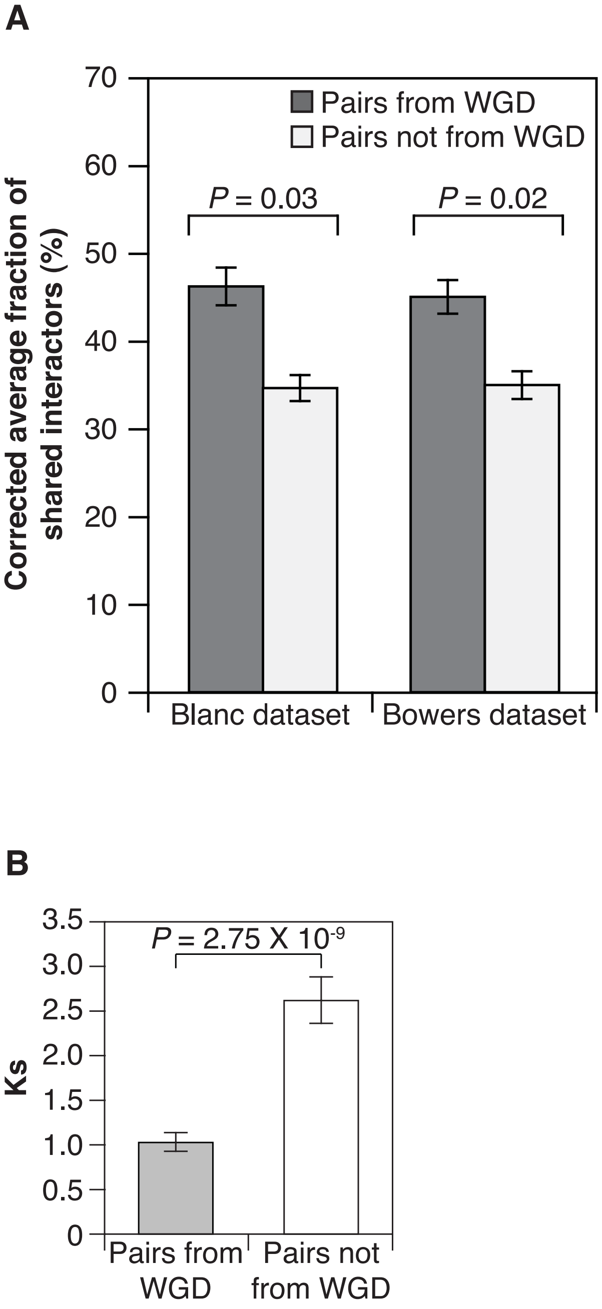
Supplementary Figure 41



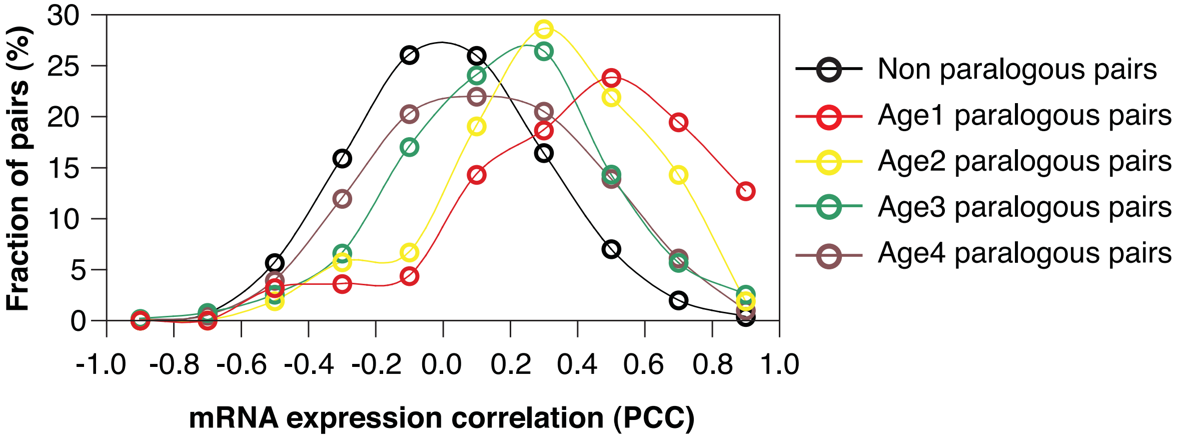
Supplementary Figure 42



Supplementary Figure 43



Supplementary Figure 44



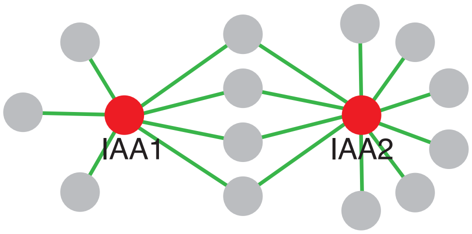
Supplementary Figure 45

A

```
IAA1  --MEVTNGLNLKDTELRLGLPG--AQEEQQLELSCVRSNNKRKNNDSTEESAPPPAKTQI
IAA2  MAYEKVNELNLKDTELCLGLPGRTEKIKEEQEVSCVKSNNKRLFEETRDEEESTPPTKTI
          * .* ******** *****   : ::: *:***:*****  ::: :*. ..*... *

IAA1  VGWPPVRSNRKNNNNKNVSYVKVSMDGAPYLRKIDLKMYKNYPELLKALENMFKFTVGEY
IAA2  VGWPPVRSSRKNNNS--VSYVKVSMDGAPYLRKIDLKTYKNYPELLKALENMFKVMIGEY
       ********.*****.  ******************** ****************. :***

IAA1  SEREGYKGSGFVPTYEDKDGDWMLVGDVPWDMFSSSCQKLRIMKGSEAPTAL---
IAA2  CEREGYKGSGFVPTYEDKDGDWMLVGDVPWDMFSSSCKRLRIMKGSDAPALDSSL
       .************************************::*******:**:
```



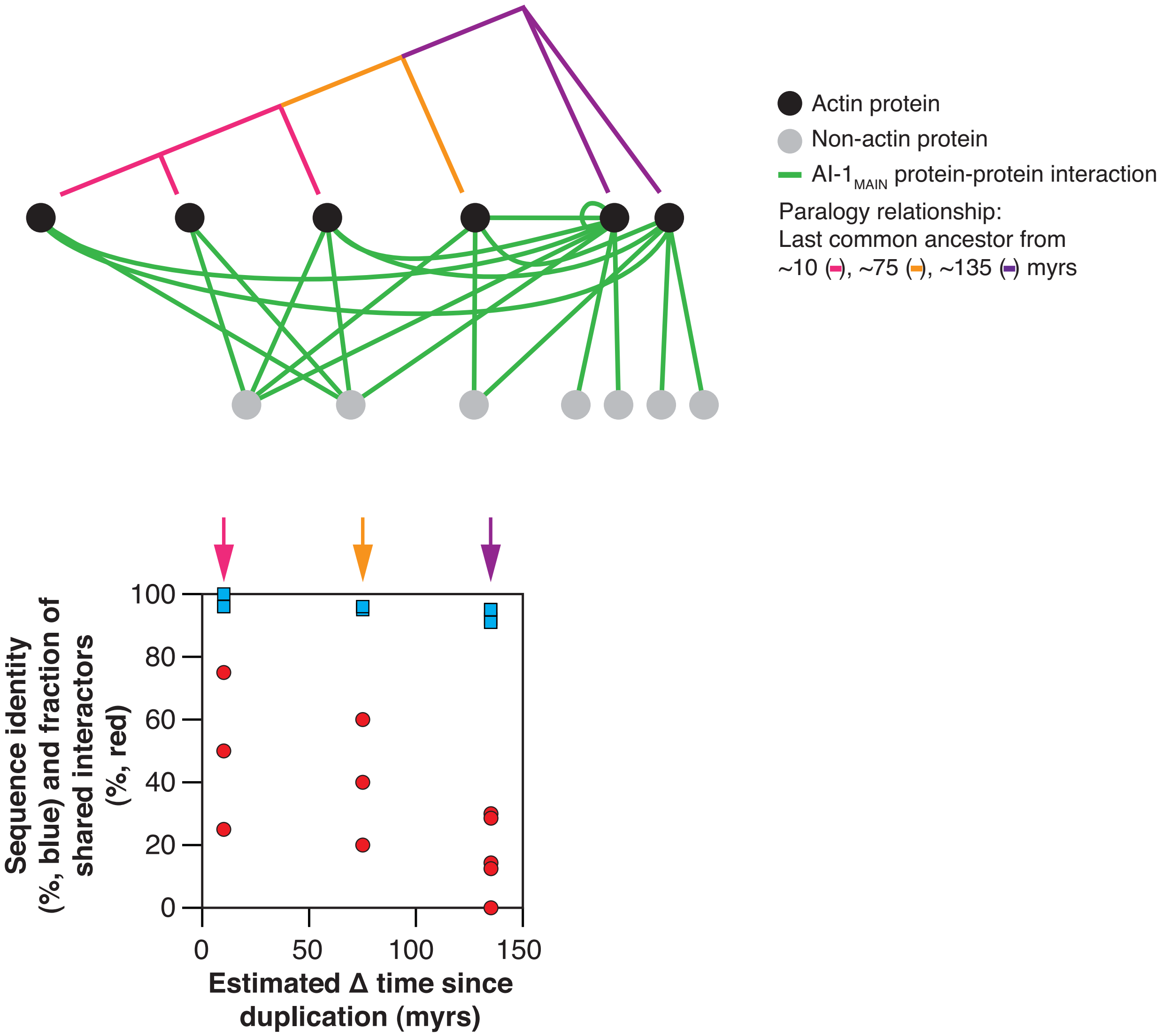
● IAA1 and IAA2

● Other proteins

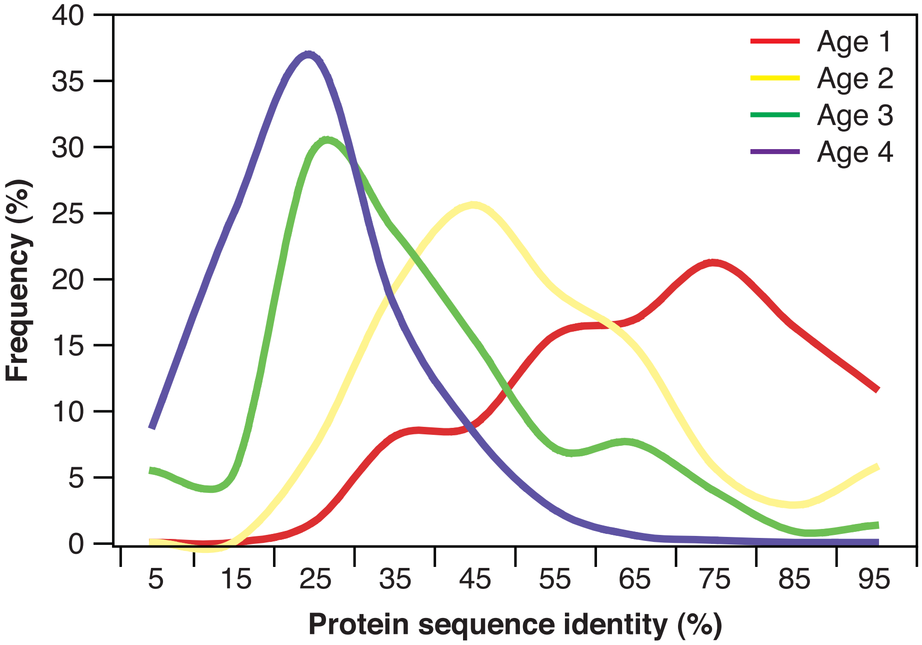
▬ AI-1 protein-protein interaction

Supplementary Figure 45

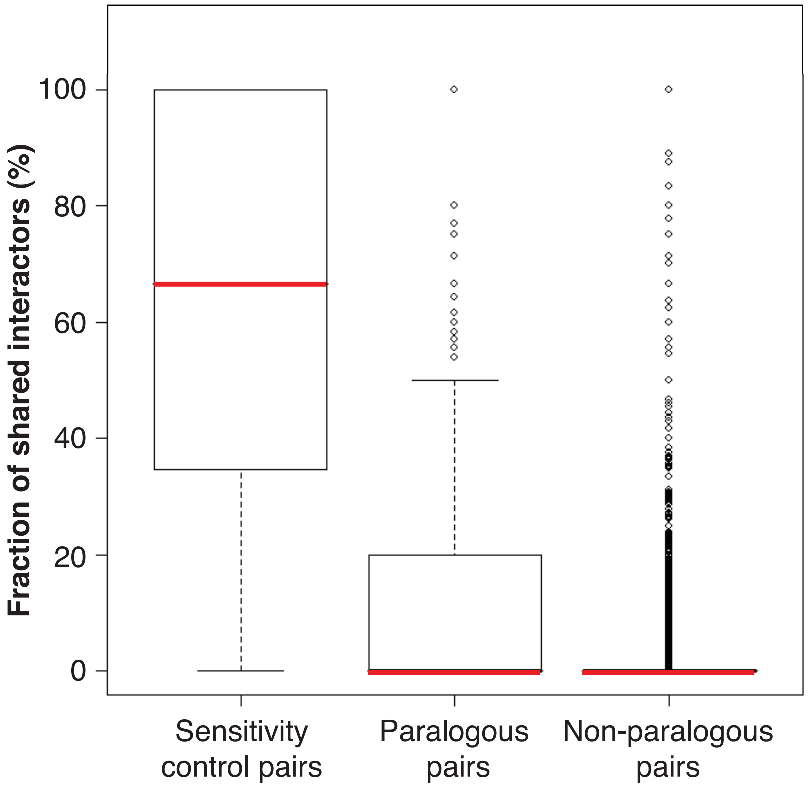
B



Supplementary Figure 46



Supplementary Figure 47


Fraction of shared interactors (%)
Sensitivity control pairs
Paralogous pairs
Non-paralogous pairs