Supplementary Information

Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*

Insuk Lee, Bindu Ambaru, Pranjali Thakkar, Edward M. Marcotte, and Seung Y. Rhee

METHODS

Reference and benchmark sets

The Gene Ontology (GO) biological process (BP) annotation from TAIR7[1] served as the major reference set for training and benchmarking the network. For the best trade-off between reliability and comprehensiveness, we selected BP annotations that have been manually checked (*i.e.*, supported by GO evidence codes IDA, IMP, IGI, IPI, TAS, RCA). We excluded annotations supported by evidence codes IEP and ISS to deemphasize gene-expression and sequence similarity based functional information in the training data. To minimize training bias, we excluded annotations made directly to the following terms: 1) 2 over-dominant terms (these 2 terms out of more than 1,200 BP terms account for >40% of total training gene pairs and were thus removed), “regulation of transcription, DNA-dependent” (GO:0006355) and “regulation of transcription” (GO:0045449); 2) 4 additional phosphorylation terms that have highly diverse biological roles, “protein amino acid phosphorylation” (GO:0006468), “protein amino acid autophosphorylation” (GO:0046777), “protein amino acid dephosphorylation” (GO:0006470), and “phosphorylation” (GO:0016310); and 3) direct children (with annotations) of the BP root term, “metabolic process” (GO:0008152), “growth” (GO:0040007), and “reproduction” (GO:0000003). The resulting dataset of reference gene annotations for training contained 341,821 pairs covering 6,487 Arabidopsis genes (~24% of 27,029 protein-coding loci).

To validate AraNet using independent annotation, we employed two reference sets, GO cellular component (CC) annotations (based on TAIR7) and the Kyoto-based KEGG database (omitting isozymes)[2], with the former annotating sub-cellular protein locations and the latter annotating biochemical pathways. Links generated between genes sharing these annotation terms overlap only 0.4-2.5% with the GO BP training set.

Log likelihood scoring scheme for heterogeneous data standardization

Different biological data sets support gene-gene associations with differing levels of confidence due to variation in the data quality or the innate value for inferring functional associations. Thus, to integrate heterogeneous data into a composite model of functional associations, we first evaluated each data set using a common scoring scheme, allowing the relative merits of each to be measured prior to integration. Specifically, using the log likelihood score (*LLS*) scheme described in[3], we estimated the strength of functional coupling between each pair of genes, defined as the likelihood of participating in the same process, conditioned on each dataset. We then combined functional linkages derived from the various datasets to construct an overall integrated gene network.

In this scheme, $LLS = \ln\left(\frac{P(L \mid E)/P(\neg L \mid E)}{P(L)/P(\neg L)}\right)$, where P(L|E) and P(¬L|E) are the frequencies of linkages (L) observed in the given experiment (E) between annotated genes operating in the *same* pathway and in *different* pathways, respectively, while P(L) and P(¬L) represent the prior expectations (*i.e.*, the total frequency of linkages between all annotated *Arabidopsis thaliana* genes operating in the *same* pathway and operating in *different* pathways, respectively). Scores greater than zero indicate the data set tends to link genes in the same pathway, with higher scores indicating more confident linkages and stronger support for the genes operating in the same pathway.

To monitor and avoid overtraining the network model, we employed 0.632 bootstrapping for all *LLS* evaluations. 0.632 bootstrapping has been shown to provide a robust estimate of classifier accuracy, generally out-performing cross-validation, especially for very small datasets (e.g., see[4]). The data evaluation and integration strategy we describe below is therefore appropriate even for less well annotated genomes such as for crop species. 0.632 bootstrapping employs sampling with replacement, constructing the training set from data sampled with replacement and the test set from the remaining data that were not sampled. Each linkage has a probability of 1-1/*n* of not being sampled, resulting in ~63.2% of the data in the training set and ~36.8% in the test set. The final *LLS* is the weighted average of results on the two sets, equal to $0.632 * LLS_{test} + (1-0.632) * LLS_{train}$.

Integration of log likelihood scores from different data sets

To combine the *LLS* score from each dataset, we modified the previously described weighted sum method[3] to employ linearly decaying weights for additional datasets, and by incorporating a free parameter, *T*, which represents a minimum *LLS* threshold on the data sets to be integrated. The weighted sum (*WS*) integration of multiple *LLS* scores for a given gene-pair was thus calculated as:

$WS = L_0 + \sum_{i=1}^{n} \frac{L_i}{D \cdot i}$, *for all* $L \geq T$, where L_0 represents the maximum *LLS* score for a given gene pair, *D* is a free parameter determining the decay rate of the weight for secondary evidence, and *i* is the rank order index of *LLS* scores associated with a given gene pair, ranking starting from the second highest *LLS* with descending magnitude for all *n* remaining *LLS* scores. For integration, we consider only the *LLS* scores above the empirically chosen threshold *T*, thereby excluding noisy low scoring linkages. The free parameter *D* ranges from 1 to +∞, and is optimized to maximize overall performance of the integrated model, measured as the area under a recall-precision curve for recovery of training set gene pairs. Note that all gene pairs in a given integration share the same value of *D*; the relative independence of the datasets being integrated is thus captured with only a single parameter. As the optimal value of *D* approaches +∞, *WS* approaches $L_{0,}$ and lower scoring *LLS* scores do not provide any additional likelihood, as appropriate when all data sets are completely dependent. We independently test the performance of a *naïve* Bayesian integration of the *LLS* scores (which is simply the sum of the *LLS* scores for each given gene pair), then select the integration model that maximizes the area under a plot of *LLS* versus coverage of genes incorporated in the network. Note that because *LLS* scores for a given gene-gene association are first sorted by decreasing magnitude prior to calculation of the WS scores, each individual gene-gene association may have a different data type as its primary line of support, with additional datasets/types contributing in a fashion weighted according to the extent of inter-dataset dependency.

The resulting network represents a unified model of functional coupling between *Arabidopsis thaliana* genes, estimated from the corpus of large-scale, predominantly systematically collected, data. We describe in detail how each data set was analyzed and used in building the network below.

Inferring functional linkages from mRNA expression data

Gene functional linkages were inferred from co-expression patterns of mRNA as described[5], in particular restricting analysis to sets of experiments assaying similar biological processes. Data from 1,074 DNA microarrays (468 dual channel and 606 single channel experiments) were downloaded from the Arabidopsis Information Resource (TAIR) and organized into sets according to publication, with each set representing multiple microarray experiments from a lab or a consortium in which experiments were focused on a particular biological process, e.g. abiotic stress. Among 116 dual channel DNA microarray experiment sets and 65 single channel experiment sets, we considered those with at least 10 array experiments, corresponding to 5 from single channel and 10 from dual channel arrays, comprising 308 DNA microarray experiments in total. Out of these 15 sets, 2 dual channel sets and 9 single channel sets (comprising 242 microarray experiments in total) exhibited significant correlations between the Pearson correlation coefficient (PCC) between genes' expression vectors and the likelihood of functional coupling between the genes (LLS, as described above) and were analyzed further (**Supplementary Table 1A** and **Supplementary Figure 13**). Linkages were derived from each of these 11 DNA microarray experiment sets, then were integrated by the weighted-sum method as described above. We also tested for signal using mRNA expression vectors derived by concatenating all 245 experiments for those 11 experiment sets, and found no significant regression between co-expression and functional association (**Supplementary Figure 13**), consistent with previous observations regarding the importance of considering sets of microarray experiments with related biological contexts[3].

Inferring functional linkages from physical interactions between proteins

Protein-protein interaction data were collected from the literature and the online databases IntAct[6], BIND[7], TAIR[1], and de Falter *et al.*[8]. Protein sequence IDs were mapped to AGI locus names, and redundant entries from each database were merged to create a non-redundant data set, wherein each interaction was supported by published literature. The final interaction set included 751 unique interactions among 691 proteins. We calculated a single LLS for the entire protein-protein interaction set using annotated genes (LLS = 3.55), and assigned it to all gene pairs in the protein-protein interaction set, including unannotated ones.

Inferring functional linkages from the genomic context of orthologous proteins

Functional linkages were also inferred from comparative analyses of genome sequences. We found that phylogenetic profiling[9-11] and gene neighbors[12-14] among prokaryotic orthologs of *Arabidopsis* genes show reasonable performance for linking functionally related *Arabidopsis* genes. We analyzed a total of 424 completely sequenced bacterial genomes (downloading 31 archaeal and 393 eubacterial genome sequences from NCBI on Dec. 18, 2006). Briefly, each *A. thaliana* protein sequence was compared to every bacterial protein sequence using the program BLASTP with default settings, then the alignment scores analyzed to calculate functional linkages as described[15]. We benchmarked inferred linkages from three different genome sets—all 424 bacterial genomes, a subset of 313 genomes obtained by selecting one from each unique species, and a subset of 184 genomes selecting one from each unique genus. Representative genomes for each unique species or genus were chosen by the maximum number of BLASTP hit proteins to the *Arabidopsis* proteome. We found that the 184 unique-genus genome set maximized the performance for inferring functional linkages by both the phylogenetic profiling and gene neighbor algorithms. Based on the 184 genome subset, we assigned log likelihood scores to each *A. thaliana* gene pair, based upon a regression model relating the LLS to the mutual information between the phylogenetic profiles, calculated as described[15]. Similarly, we assigned log likelihood scores to each *A. thaliana* gene pair based upon a regression model relating the LLS to the log of the probability of observing gene neighbors by chance, calculated as described[15] (**Supplementary Figure 14**).

Inferring functional linkages from protein domain co-occurrence profiles

Functional association between proteins can also be inferred by their sharing of defined protein domains. This is an intuitive approach but requires appropriate training data, both to avoid circularity and because the quality of functional inference varies for different types of domains. We modified the mutual information scoring method employed for phylogenetic profiles to instead identify functional associations based on domain co-occurrence between protein pairs as follows: We first retrieved the set of InterPro database[16] domains for all *A. thaliana* proteins from TAIR (v. TAIR7). A total of 47,771 InterPro domain mappings for 21,605 *A. thaliana* proteins were identified, spanning 4,129 unique domains. We then generated a matrix of all proteins versus all InterPro domains, filling the matrix with binary scores such that 1 indicates presence of a given domain in a given protein and 0 indicates absence. Tests with functional linkages derived directly from similarities between pairs of proteins' domain vectors indicated that common domains carried significantly less value for inferring functional linkages than rare domains. We thus generated a weighted version of the domain occurrence matrix in which each domain occurrence was scored instead as the inverse of its frequency in the proteome. Similarities between these weighted domain occurrence vectors were calculated as the mutual information of the vectors, which accounts for vector complexity and performed better than correlation measures at identifying functionally related proteins due to the presence of many vectors with low complexity.

Specifically, we calculated the mutual information score for each protein pair as:

$MI(A,B) = H(A) + H(B) – H(A,B)$, where $H(A) = -\sum p(a)\ln p(a)$ represents the marginal entropy of the probability distribution of $p(a)$ of gene A, and $H(A,B) = -\sum\sum p(a,b)\ln p(a,b)$ represents the relative entropy of the joint probability distribution $p(a,b)$ of genes A and B. To minimize trivial associations, we excluded homologous protein pairs with BLASTP scores of $E < 10^{-3}$. The remaining associations showed significant enrichment for high LLS scores (**Supplementary Figure 14**).

Inferring functional linkages from associalogs

AraNet includes many functional linkages transferred from other organisms by orthology relationships. These datasets were scored as for any other *A. thaliana* dataset (e.g., assigning LLS scores to the transferred linkages using the *A. thaliana* annotation benchmarks), but involved the additional step of calculating orthologs and weighting linkages by confidence in the orthology assignments.

Orthologs were identified using INPARANOID[17]. In many cases, we might expect 1-to-many or many-to-many orthology relationships between species. To better handle such cases, we weighted orthology-based functional inferences by the confidence scores in the in-paralogs (paralogs retaining functional similarity) identified by INPARANOID. We inferred *A. thaliana* functional linkages based on linkages from version 3 of YeastNet[5], version 2 of WormNet[15], and a functional network of human genes (I.L., E.M.M., manuscript in preparation). For each organism, each type of evidence (mRNA co-expression, yeast two-hybrid interactions, etc.) was treated as an individual data set. A total of 19 linkage sets were generated (dubbed associalogs[15], for conserved functional associations between organisms): 5 from worm, 1 from fly, 5 from human, and 8 from yeast. To minimize effects of errors in ortholog assignments and to better handle effects of in-paralogs, we weighted transferred functional linkages by the INPARANOID confidence scores (ranging from 0 to 1) in the in-paralogs. We observed improved performance (judged by recall-precision analysis at recovering *A. thaliana* functional linkages) using a heuristically defined INPARANOID-Weighted Log Likelihood Score (IWLLS) for each transferred linkage, which equals the LLS score of the gene pair in the orthology source organism + log(INPARANOID score for gene A) + log(INPARANOID score for gene B). Each such associalog dataset was then scored as for *A. thaliana* datasets, e.g., using a regression model between the assigned IWLLS scores and the LLS for sharing *A. thaliana* functional annotation (**Supplementary Figure 14**). Another set of functional linkages was transferred from fly protein-protein interactions derived from BIOGRID[18], IntAct[6], and MINT[19], downloaded on March 2007. We divided those interactions into literature-based low-throughput data and high-throughput yeast two hybrid data[20], and then measured a global LLS for linkages in each subset (2.74 for the low-throughput subset and 1.79 for the high-throughput yeast two hybrid subset).

The 19 associalog sets (8 from yeast, 1 from fly, 5 from worm, and 5 from human) were integrated with 5 linkage sets derived from *Arabidopsis* to construct the final AraNet (**Supplementary Table 2**).

ROC analysis of gene function identification

The predictive power of AraNet for inferring gene function was tested by measuring the tendency for genes annotated with the same function to cluster in the network. We evaluated clustering of genes annotated with GO biological process terms, as well as those sharing GO cellular compartment annotations, or KEGG pathway annotations.

For each set of genes annotated with the same term (the 'seed set' of genes), clustering was evaluated by rank-ordering genes in the network by each genes' sum of linkage LLS scores to the seed gene set, using cross-validation (*i.e.*, omitting each seed gene in turn from the seed set for the purposes of its evaluation). For cases in which genes annotated to have the same function cluster in the network, we expect a higher retrieval rate for genes that are involved in the seed gene set (positives) than for genes that are not annotated with that function (negatives) in a Receiver Operating Characteristic (ROC) plot, resulting in a ROC curve above the plot diagonal. However, if the genes known to be involved in the same function are not clustered in the network, the retrieval rate of positive and negative genes will be similar, resulting in a diagonal ROC curve, indicating random expectation **(Figure 2A)**. Each such ROC analysis was summarized by the area under the ROC curve (AUC), which ranges from near 0.5 (*i.e.*, the area under the diagonal, indicating random performance) to 1 (genes with this function are tightly clustered in the network). We compared the predictive power of a randomized network and AraNet for 318 GO biological process terms with at least 5 annotated genes, with AraNet showing significantly higher predictive ability than random (examples are shown in **Figs. 2E-F and Supplementary Figure 2)**. Similar analyses of 86 GO cellular compartment terms and 82 isozyme-free KEGG pathways (the KEGG annotation set after exclusion of genes with isozymes) are shown in **Figs. 3A-B**. In all cases, we considered only annotations with at least 5 associated genes.

Detailed procedure for reconstructing the *Arabidopsis thaliana* gene network

To more clearly define the procedure for generating the network, we provide the full procedure as pseudo-code. Regression models are plotted in **Supplementary Figures 13** and **14. Supplementary Table 1** lists specific DNA microarray experimental data sets evaluated for AraNet. **Supplementary Table 2** lists the contributions of different datasets to the final network.

```
1. Identify Arabidopsis orthologs of yeast, worm, human, and fly proteins using INPARANOID
2. For Arabidopsis DNA microarray data
    2.1. For each set of Arabidopsis DNA microarray experiments (corresponding to all arrays from a given TAIR data set)
        2.1.1. Calculate the mean-centered Pearson correlation coefficient (PCC) between all pairs of genes' expression profiles
            2.1.1.1.Calculate (by t-test) the minimum correlation coefficient for 99% confidence given the number of experiments in the set. For further analyses, consider only those gene pairs meeting this criterion.
            2.1.1.2.Evaluate the regression between PCC and the log likelihood score (LLS) of sharing pathway annotations
                2.1.1.2.1. Reject set if no relationship is evident between PCC and LLS
            2.1.1.3.Filter genes considered in the correlation analysis by requiring each gene to exhibit significant expression changes (e.g., >x-fold, typically ~1.2-fold) in some minimal number y of experiments across the dataset. Optimize the parameters x and y by recall-precision analysis, maximizing the area under a plot of LLS versus the number of genes participating in the linkages.
            2.1.1.4.Fit a regression model (typically sigmoidal) between PCC and LLS, considering only genes passing the optimized filtering criteria (2.1.1.3) and only gene pairs whose correlation exceeds the 99% confidence level (2.1.1.1).
            2.1.1.5.Using the regression model, assign LLS scores to all gene pairs whose correlation exceeds the 99% confidence level, including unannotated gene pairs.
            2.1.1.6.Select a minimum LLS threshold from the inflection point of the regression model. Retain only LLS scores/gene pairs surpassing threshold.
    2.2. Integrate LLS scores from complete collection of sets of DNA microarrays
        2.2.1. Calculate the weighted sum of LLS scores for each gene pair across the analyzed DNA microarray experiment sets
        2.2.2. Optimize the choice of the weighting parameters D and T using recall-precision analysis by maximizing the area under a plot of LLS versus # of genes participating in the linkages. Compare to naïve Bayesian integration, and choose from weighted integration versus naïve Bayes by recall-precision analysis.
3. For Arabidopsis protein-protein interaction (PPI) data
    3.1. Measure the LLS score for all pairs in the set
    3.2. Assign this LLS score to all interacting pairs in the set, including unannotated pairs
4. For Arabidopsis protein domain co-occurrence, phylogenetic profiles, and gene neighbors data
```

4.1. Fit regressions between LLS and data-intrinsic scores (log(weighted mutual information) of domain co-occurrence, mutual information of phylogenetic profiles, and –log(random probability of being gene neighbors), respectively)

4.2. Using regression fit(s), assign LLS scores to all domain co-occurring (or co-inherited or co-neighboring) gene pairs, including unannotated gene pairs

5. For fly PPI data
    - 5.1. Considering *Arabidopsis* gene pairs corresponding to interacting fly proteins, fit regression between LLS and fly PPI confidence scores provided with fly PPIs
    - 5.2. Using regression fit, assign LLS scores to all *Arabidopsis* gene pairs corresponding to interacting fly proteins, including unannotated pairs
6. For yeast, worm, human functional network data
    - 6.1. Analyze each data type (e.g., DNA microarrays, affinity purification/mass spec, etc.) separately, considering *Arabidopsis* gene pairs whose yeast (or worm, human) orthologs are linked by the given data type.
        - 6.1.1. Fit regression models between LLS for *Arabidopsis* gene pairs and IWLLS (INPARANOID-weighted LLS) associated with the orthologous yeast gene pairs in the yeast (or worm, human) network
        - 6.1.2. Using the regression model, assign LLS scores to all *Arabidopsis* gene pairs corresponding to linked yeast (or worm, human) genes, including unannotated pairs
    - 6.2. Integrate yeast (or worm, human)-derived linkages by calculating the weighted sum of LLS scores for each gene pair across the set of yeast (or worm, human) data types, optimizing the choice of D and T parameters by recall-precision analysis as in (2.2). Compare to *naïve* Bayesian integration, and choose from weighted integration versus *naïve* Bayes by recall-precision analysis.
    - 6.3. Fit regression between LLS and weighted sum (or *naïve* Bayes sum), then assign LLS scores to all *Arabidopsis* gene pairs corresponding to linked yeast (or worm, human) genes, including unannotated pairs
7. Integrate all linkages using the weighted sum method, optimizing the choice of D and T parameters by recall-precision analysis as in (2.2). Compare to *naïve* Bayesian integration, and choose from weighted integration versus *naïve* Bayes by recall-precision analysis

Topological analysis of network model

We examined the topological properties of AraNet. **Supplementary Figure 15A** plots the node degree distribution of the *Arabidopsis thaliana* gene network. Many network models derived from complex biological systems are characterized by scale-free degree distributions[21]. However, the core AraNet gene network is not scale-free. Instead, we find the degree distribution is well fit (r^2 = 0.99) by a combined power-law/exponential decay model. This distribution follows a power-law for genes with lower connectivity, then exponential decay for genes with degrees higher than a characteristic threshold (β = 185, **Supplementary Figure 15A**). This may stem from having an upper bound on the size of typical pathways, resulting in systematic under-representation of genes with the highest connectivity. We also observe extensive clustering in the network, consistent across various sizes of network coverage (clustering coefficient ~ 0.3, calculated as in [22]), indicating a highly structured network, with many clusters (connected subnetworks) likely representing pathways or functional modules (**Supplementary Figure 15B**). This trend likely underlies the correct identifications of genes associated with specific biological processes (**Figures 2-4** and **Supplementary Figure 2**).

Comparison of AraNet with previous *Arabidopsis* networks

The general predictive power of AraNet was compared to 4 previous gene networks for *Arabidopsis* (described in **Supplementary Table 3**). To compare the networks fairly, we employed the two annotation sets that are most independent from all 5 networks: GO cellular compartment annotations and KEGG pathway annotations. (Note that some overlap with these test sets was unavoidable, as Multinetwork[23] employed KEGG; nonetheless the performance of Multinetwork on these datasets was not notably elevated). AraNet showed higher predictive performance than previous networks across all tests **(Supplementary Figure 3)**.

Analysis of cell-type specific expression specificity

For each cell type among the 20 root cell types profiled by Brady *et al.*[24], transcripts with DNA microarray-based integrated signal intensities >1200 were defined as well-expressed, resulting in roughly 3,000 genes observed to be strongly expressed in each root cell type. We determined the enrichment for co-expressed genes for each cell type as an odds ratio, calculated as posterior odds / prior odds. Here, the posterior odds equals the number of gene pairs that are linked and co-expressed in a cell type divided by the number of gene pairs that are linked but not co-expressed in a cell type. The prior odds was calculated as the number of gene pairs that are co-expressed in a cell type and linked in a randomized network generated with the same number of genes and linkages as AraNet, divided by the number of gene pairs linked in the randomized network but not co-expressed.

Validation of AraNet using independent seed phenotype test sets

The predictive power of AraNet for associating genes with phenotypic traits was also tested using sets of genes associated with two seed phenotypes as reported by the *Arabidopsis* SeedGenes Project[25]. This database reports essential genes causing embryonic lethality when disrupted by mutation, as well as genes whose disruption caused changes in seed (embryo) pigmentation. A version of the database downloaded from http://www.seedgenes.org/ on December, 2007 (Release 7) reported 245 confirmed genes with embryonic lethality genes and 23 confirmed genes with seed pigmentation phenotypes. Using these phenotypic gene sets, the predictive power of AraNet and 4 previous gene networks for Arabidopsis **(Supplementary Table 3)** was compared by ROC analysis**.** AraNet was the only network to predict gene essentiality substantially better than random expectation; it was also the strongest predictor of seed pigmentation **(Supplementary Figure 4)**.

Confirmation of T-DNA insertions

The genotype of each T-DNA insertion allele was confirmed by PCR using a pair of primers against the gene and a primer against the right border of the T-DNA (LBb1.3:ATTTTGCCGATTTCGGAAC), as recommended by the SALK (http://signal.salk.edu/tdnaprimers.2.html). Gene-specific primers are listed in **Supplementary Table 16**. Genotypes of 3-8 plants of each mutant line were tested. For each line, four PCR reactions were performed with genomic DNA extracted from leaf tissue: T-DNA primer with either forward or reverse primer of the gene, gene-specific primers, gene-specific primers for another gene (positive control). Selfed progeny of confirmed homozygote lines were collected and used for further analysis.

Gene Expression

RT-PCR was performed to confirm lack of gene expression in the confirmed homozygote lines. Real-time PCR was performed to determine expression of the genes in different tissues and conditions. To determine expression in different tissues and developmental stages, RNA was isolated from 100 mg of leaf or flower tissues of 4 week old plants in soil, root or shoot tissues from 12 days old plants grown on MS plates, and seedling tissue from 3 days old plants grown on MS plants (**Supplementary Figure 5**). RNA was isolated using Qiagen RNeasy plant mini kit (Catalog #74904). Potential contaminating genomic DNA was removed with a DNA free kit (Applied Biosystems #AM1906). 2 µg of RNA was used in two-step RT-PCR kit (Ambion #AM1710) according to manufacturer's directions. Real time PCR was performed using a Roche Lightcycler480 with the Lightcycler DNA master SYBR green I reporter from Roche Applied Science (Catalog #12015099001). For all RT-PCR experiments, primers against actin were included as a positive control. Relative expression quantification was performed using the ΔΔCT method[26] using actin as the reference gene. Gene-specific primers used for RT-PCR experiments are listed in **Supplementary Table 17**.

Genetic Analysis

Linkage tests: Homozygote lines of *drs1-1* and *lrs1-1* were crossed to wild type Col-0 and the ensuing F1 plants were selfed to generate an F2 population. Genotypes of 259 and 128 F2 plants of *drs1-1* x Col-0 and *lrs1-1* x Col-0 crosses, respectively, were determined by PCR using the T-DNA primer LBb1.3 and gene-specific primers in **Supplementary Table 16**. To determine linkage between the mutant phenotypes and the *drs1-1* allele, half of the F2 population of the *drs1-1* x Col-0 cross were subjected to the relative water content assay (see Drought response assay for details) in which half of the plants were treated with drought and the other half watered. The other half of the plants were subjected to the leaf transpiration assay in the presence and absence of 10 µM ABA (see Hormone response assays for details). Phenotypes for plants in each genotype were averaged. To determine the linkage between the root phenotype and *lrs1-1* allele, F2 plants were grown on MS agar plates and the number and length of the primary and lateral roots of 10 day old seedlings were measured using the ImageJ software (http://rsbweb.nih.gov/ij/) on digital images of the plants. Unpaired Student's t-tests were used to determine significance between genotypes and treatments, and Chi square tests were used to determine deviation from the expected segregation ratio.

Functional Complementation and Overexpression: An Entry clone G22154 (ABRC) containing the full-length cDNA of *Lrs1* was introduced to a Gateway-compatible binary vector (pGWB2) containing a 35S CaMV promoter[27] to generate expression clone pGWB2-LRS1. This clone was transformed into *Agrobacterium* strain C58C1 pGV3101 pMP90. Mutant plants carrying the *lrs1-1* allele and Col-0 wild type plants were transformed with the transgenic *Agrobacterium* using the floral dip method[28]. Although the *lrs1-1* allele contains a T-DNA insertion that contains the kanamycin resistance marker, the *lrs1-1* plants were not resistant to kanamycin. Therefore, transgenic plants were selected on agar plates with 50mg/L kanamycin. Seven independent transgenic lines were obtained from each transformation. Representative lines were tested for segregation of kanamycin resistance. For both the complemented and overexpresed lines, kanamycin resistance segregated as a single locus (Kan^R:Kan^S::50:19 for complementation lines and 60:12 for overexpression lines, $p >= 0.1$ of 3:1 expection ration, chi-square test). Selfed progeny of the tranformants were grown on agar plates containing kanamycin to assay for root phenotypes.

DISCUSSION

To determine how much of the Arabidopsis data contributes to the predictability of AraNet, we constructed a version of AraNet with no plant-derived data but including plant-domain-based links, and tested the performance of this network by ROC analysis. If the prediction power depended heavily upon plant domain annotation, we would see significantly better AUCs with the version lacking plant data but including domain-based linkages than the version lacking both. In fact, prediction power improves only modestly and in proportion to the expected minor contribution of the plant-domain-based (AT-DC) linkages (**Supplementary Fig. 16**). This confirms that the other plant-derived datasets are the critical ones. *Arabidopsis* protein domain annotations play a relatively minor role in AraNet performance compared to other plant datasets. To assess how much each data set contributes to AraNet's performance, we tested the predictive power of each individual data set in isolation by ROC analysis, plotting median AUC versus coverage (**Supplementary Fig. 17**). Individual data sets show much poorer predictive ability than the integrated AraNet. Among those individual data set, plant gene co-expression links shows the best predictive power.

To determine the relative contribution of incorporating diverse data types versus combining different evidences for inferring function to the performance of AraNet, we compared the guilt-by-association (GBA) method to 1-nearest neighborhood (1NN) method to predict biological roles. GBA method infers biological roles of a gene based on *all* of the neighbors of the gene, whereas 1NN method only uses the closest neighbor information. We tested the performance of a 1-NN classifier on AraNet, scoring each gene for its association with a trait according to its single strongest network edge (*i.e.*, testing whether consideration of different data types (data integration) alone is the primary driver of performance or whether combining evidence across multiple network edges is also a significant contributor). 1-NN performs significantly worse than the GBA approach we employ, indicating that both data integration *and* the combination of lines of evidence across the network edges are important to performance (**Supplementary Fig. 18**).

TABLES

Supplementary Table 1A. DNA microarray experiment sets exhibiting significant correlation between mRNA co-expression and LLS scores.

TAIR expression set name	Experiment name	Num. expts	Authors (data set URL)
ExpressionSet_93	Circadian rhythm (dual)	17	Schaffer, Robert (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1005823568)
ExpressionSet_203	Circadian rhythm in Col & Lan WT and mutants (dual)	29	Barak, Simon (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1005823573)
ExpressionSet_ME00313	R gene induced gene expression profile (single)	20	Dangl, J and Eulgem, T (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1006710792)
ExpressionSet_ME00332	Response to bacterial-(LPS, HRPZ, FLG22) and oomycete-(NPP1) derived elicitors (single)	36	Brunner, F and Nürnberger, T. AtGenExpress (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1008080727)
ExpressionSet_ME00335	Brassinolide time course study (single)	12	Goda, H, Yoshida, S and Shimada, Y. AtGenExpress (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1007966053)
ExpressionSet_ME00343	GA3 time course study (single)	12	Goda, H, Yoshida, S and Shimada, Y. AtGenExpress (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1007966175)
ExpressionSet_ME00345	Light treatments (single)	42	Kretsch, T. AtGenExpress (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1007966126)
ExpressionSet_ME00352	Effect of Brassinosteroids in seedlings (single)	22	Goda, H. and Shimada, Y. AtGenExpress (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1007999438)
ExpressionSet_ME00354	Response to *Erysiphe orontii* infection (single)	24	Ausubel, F. and Dewdney, J. AtGenExpress (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1008031468)
ExpressionSet_ME00357	Effect of Gibberellic acid biosynthesis inhibitors on seedlings (single)	16	Goda, H., Yoshida, S., Asami, T., and Shimada, Y. AtGenExpress (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1008080692)
ExpressionSet_ME00359	Effect of Brassinosteroid inhibitors on seedlings (single)	12	Goda, H., Yoshida, S., Asami, T., and Shimada, Y. AtGenExpress (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1008205330)

Supplementary Table 1B. DNA microarray experiment sets lacking correlation between mRNA co-expression and LLS scores.

TAIR expression set name	**Experiment name**	**# expts**	**Authors (data set URL)**
ExpressionSet_ 231	Sulfate in plant growth and defense (dual)	12	Bones, A. AFGC (http://www.arabidopsis.org/servlets/TairObject?type =expression_set&id=1005823598)
ExpressionSet_ 237	White light time course (dual)	32	Wu, S-H. AFGC (http://www.arabidopsis.org/servlets/TairObject?type =expression_set&id=1005823603)
ExpressionSet_ 239	Chitin elicitation time course (dual)	12	Zhang, B.,Ramonell, K.,Somerville, S.,Stacey, G. (2002) Characterization of early, chitin-induced gene expression in Arabidopsis. Molecular Plant Microbe Interactions. 15(9):963.
ExpressionSet_ ME00377	In vitro tracheary element transdifferentiation (single)	10	Fukuda, H., Kubo, M., and Demura, T. (http://www.arabidopsis.org/servlets/TairObject?type =expression_set&id=1008805373)

Supplementary Table 2 Twenty-four types of evidence incorporated into AraNet.

Evidence code	Evidence description	# unique genes	# unique gene pairs
AT-CX	Co-expression among Arabidopsis genes	13,821	308,320
AT-DC[16]	Co-occurrence of domains among *A. thaliana* proteins	9,334	51,562
AT-GN[12-14]	Gene neighbourhoods of bacterial and archaeal orthologs of *A. thaliana* genes	5,100	109,479
AT-LC[1, 6-8]	Literature curated *A. thaliana* protein physical interactions	691	751
AT-PG[9-11]	Co-inheritance of bacterial and archaeal orthologs of *A. thaliana* genes	3,971	134,076
CE-CC[15]	Co-citation of worm orthologs in Medline abstracts	1,020	7,936
CE-CX[15]	mRNA co-expression of worm orthologs	3,164	131,328
CE-GT[15]	Genetic interactions between worm orthologs	553	2,741
CE-LC[15]	Literature curated worm protein physical interactions	1,274	2,920
CE-YH[15]	High-throughput yeast 2-hybrid interactions among worm orthologs	1,241	3,007
DM-PI[6, 18-20]	Fly protein physical interactions	3,920	18,163
HS-CX	mRNA co-expression between human orthologs	4,035	72,211
HS-DC	Co-occurrence of domains among human proteins	4,013	27,410
HS-LC[6, 7, 18, 19, 29]	Literature curated human protein physical interactions	4,510	115,036
HS-MS[6]	Human protein complexes from affinity purification/mass spectrometry	857	2,880
HS-YH[30, 31]	High-throughput yeast 2-hybrid interactions among human orthologs	870	6,667
SC-CC[5]	Co-citation of yeast orthologs in Medline abstracts	4,125	91,656
SC-CX[5]	mRNA co-expression among yeast orthologs	3,510	164,746
SC-DC[5]	Co-occurrence of domains among yeast proteins	3,292	40,220
SC-GT[5]	Genetic interactions between yeast orthologs	3,629	42,110
SC-LC[5]	Literature curated yeast protein physical interactions	3,908	36,588
SC-MS[5]	Yeast protein complexes from affinity purification/mass spectrometry	3,960	253,226
SC-TS[5]	Yeast protein interactions inferred from tertiary structures of complexes	1,451	13,549
SC-YH[5]	High-throughput yeast 2-hybrid interactions among yeast orthologs	2,163	7,324

Supplementary Table 3 Comparison of network models for *A. thaliana*.

Network model	**Scale**	**Description**
Multinetwork[23]	203,586 linkages among 4,339 genes (16% of genome)	No confidence scores. Linkages were collected from metabolic pathway database, protein-DNA database, protein-protein database, and interologs.
Interolog network[32]	19,368 linkages among 3,565 genes (13% of genome)	Scored by confidence values (CV). Only interolog based linkages are included.
AtPID[33]	24,418 linkages among 11,706 genes (43% of genome)	Scored by the likelihood of protein-protein interactions. Seven data sets (interologs, shared biological function, co-expression, gene fusions, gene neighbors, phylogenetic profiling, and enriched domain pair) were integrated using a *naïve* Bayesian approach.
GGM network[34]	17,476 linkages among 6,374 genes (24% of genome)	Scored by partial correlation (pcor). Used a graphical Gaussian model (GGM) to infer co-regulated gene pairs.
AraNet (this study)	1,062,222 linkages among 19,647 genes (73% of genome)	Scored by the log likelihood of functional association between gene pairs. 24 data sets (**Supp. Table 2**) were integrated using a modified *naïve* Bayesian method.

Supplementary Table 4. Genes that show seed pigmentation phenotype and defects in early seedling development from SeedGenes (www.seedgenes.org).

Locus	Symbol	Source of Mutant	Predicted Function	Refs
At1g02090	CSN7/ FUS5	S. Misera	Component of COP9 Signalosome	[35]
At1g05750	PDE247	Meinke/Syngenta	PPR Protein	[36]
At1g06570	PDS1	D. DellaPenna	p-Hydroxyphenylpyruvate Dioxygenase	[37]
At1g08520	CHLD/PDE166	Meinke/Syngenta	Magnesium Chelatase (CHLD)	[36, 38]
At2g24120	PDE319	Meinke/Syngenta, Micol/Salk	Chloroplast DNA-Dependent RNA Polymerase	[36, 39]
At2g28800	ALB3	E. Sundberg	Chloroplast Protein Translocase (Oxa1p)	[40]
At2g30950	VAR2	S. Rodermel	Chloroplast Homolog of FtsH	[41]
At2g32950	COP1	Deng/Feldmann	Nuclear Protein that Represses Photomorphogenesis in the Dark	[42]
At2g48120	PAC	Meinke/Syngenta, Scolnik/Feldmann	Uncertain	[43]
At3g03710	PDE326	Meinke/Syngenta	Uncertain	[36, 44]
At3g04260	PDE324	Meinke/Syngenta	A component required for plastid gene expression	[36, 45]
At3g11670	DGD1	C. Benning, Meinke/Syngenta	Digalactosyl Diacylglycerol Synthase	[46]
At3g48500	PDE312	Meinke/Syngenta	A component required for plastid gene expression	[36, 45]
At3g51820	CHLG/PDE325	Meinke/Syngenta	Chlorophyll synthase	[36, 47]
At3g61140	CSN1/FUS6	Meinke/Feldmann, Meinke/Syngenta	Component of COP9 Signalosome	[48, 49]
At3g62910	APG3	Meinke/Syngenta	Translation Releasing Factor RF-1	[50]
At4g10180	DET1	J. Chory, Meinke/Syngenta	Nuclear-Localized Protein	[51]
At4g14110	CSN8/COP9	Deng/Feldmann	Component of COP9 Signalosome	[52]
At4g15560	DXS/CLA1	Mandel/Feldmann, Meinke/Syngenta	1-Deoxyxylulose 5-Phosphate Synthase	[53]
At4g18480	CHL1/CH42	J. Relichova, Meinke/Syngenta	Magnesium Chelatase (CHLI)	[54]
At4g22260	IM	S. Rodermel	Chloroplast Homolog of Mitochondrial Alternative Oxidase	[55, 56]
At5g42970	CSN4/COP8	Deng/Feldmann	Component of COP9 Signalosome	[57]
At5g62790	PDE129	Meinke/Syngenta	1-Deoxyxylulose 5-Phosphate Reductoisomerase	[36, 58]

Supplementary Table 5. Top 200 candidates for seed pigmentation and early seedling development defective mutants predicted by AraNet and the 23 known genes.

Locus	Rank	Symbol	LLS	Evidences	Linked genes	GO terms	Screened
AT5G14250	1	COP13	6.55	HS-LC:0.38 AT-DC:0.32 HS-DC:0.30	FUS5 FUS6 COP9 COP8	photomorphogenesis;	no
AT3G57290	2	EIF3E	6.44	AT-DC:0.35 HS-DC:0.33 HS-LC:0.22 AT-LC:0.11	FUS5 FUS6 COP9 COP8	transcription initiation;	yes
AT2G26990	3	FUS12	6.06	HS-LC:0.32 AT-DC:0.25 HS-DC:0.17 CE-YH:0.07 CE-LC:0.07 DM-PI:0.06 HS-CX:0.06	FUS5 FUS6 COP9 COP8	photomorphogenesis; protein catabolic process;	yes
AT3G02200	4	na	6.01	AT-DC:0.41 HS-DC:0.39 CE-YH:0.10 CE-LC:0.10	FUS5 FUS6 COP8	na	no
AT5G15610	5	na	6.01	AT-DC:0.41 HS-DC:0.39 CE-YH:0.10 CE-LC:0.10	FUS5 FUS6 COP8	na	yes
AT4G11420	6	EIF3A	5.98	AT-DC:0.51 HS-DC:0.49	FUS5 FUS6 COP8	translational initiation;	no
AT4G19006	7	na	5.85	HS-DC:0.59 AT-DC:0.41	FUS5 FUS6 COP8	ubiquitin-dependent protein catabolic process; protein catabolic process;	no
AT5G45620	8	na	5.84	HS-DC:0.59 AT-DC:0.41	FUS5 FUS6 COP8	ubiquitin-dependent protein catabolic process;	yes
AT5G13630	9	GUN5	5.74	AT-GN:0.53 AT-PG:0.34 AT-CX:0.12	AT1G08520 CHLI1	biosynthetic process;	yes
AT5G07590	10	na	5.7	HS-MS:1.00	FUS6 COP8	na	yes
AT1G29150	11	ATS9	5.39	HS-DC:0.54 AT-DC:0.37 DM-PI:0.09	FUS5 FUS6 COP8	ubiquitin-dependent protein catabolic process; protein catabolic process;	no
AT1G71230	12	AJH2	5.37	HS-LC:0.83 DM-PI:0.17	FUS5 FUS6 COP9 COP8	protein deneddylation; photomorphogenesis; response to auxin stimulus; negative regulation of photomorphogenesis;	yes
AT1G75990	13	na	5.14	HS-DC:0.50 AT-DC:0.50	FUS5 FUS6 COP8	ubiquitin-dependent protein catabolic process; protein catabolic process;	yes
AT3G56150	14	EIF3C	5.14	HS-DC:0.50 AT-DC:0.50	FUS5 FUS6 COP8	translational initiation;	no

AT1G20200	15	na	5.13	HS-DC:0.50 AT-DC:0.50	FUS5 FUS6 COP8	ubiquitin-dependent protein catabolic process; embryonic development ending in seed dormancy;	no
AT4G24820	16	na	5.12	AT-DC:0.51 HS-DC:0.49	FUS5 COP8	ubiquitin-dependent protein catabolic process; protein catabolic process;	no
AT5G64760	17	na	5.12	AT-DC:0.51 HS-DC:0.49	FUS5 FUS6	ubiquitin-dependent protein catabolic process; protein catabolic process;	yes
AT5G09900	18	na	5.12	AT-DC:0.51 HS-DC:0.49	FUS5 FUS6	ubiquitin-dependent protein catabolic process; embryonic development ending in seed dormancy;	yes
AT2G19560	19	na	5.05	HS-DC:1.00	FUS5 FUS6 COP8	na	no
AT1G17220	20	na	5.03	AT-GN:0.79 AT-CX:0.21	VAR2 AT3G03710 CHLI1	translation; translational initiation;	yes
AT3G22860	21	TIF3C2	5.02	HS-DC:0.50 AT-DC:0.50	FUS5 FUS6 COP8	translational initiation;	yes
AT4G11160	22	na	4.89	AT-GN:0.73 CE-CX:0.27	VAR2 AT3G03710 APG3	translation; translational initiation;	yes
AT1G22920	23	AJH1	4.87	HS-LC:0.83 DM-PI:0.17	FUS5 FUS6 COP9 COP8	protein deneddylation; photomorphogenesis; response to auxin stimulus; specification of floral organ identity; negative regulation of photomorphogenesis;	no
AT5G01230	24	na	4.82	AT-GN:0.73 CE-CX:0.27	VAR2 AT3G03710 APG3	na	yes
AT1G76810	25	na	4.75	AT-GN:1.00	VAR2 AT3G03710 DXR	translation;	no
AT1G80620	26	na	4.73	AT-GN:1.00	VAR2 AT3G03710	translation;	no
AT5G56280	27	CSN6A	4.72	HS-LC:0.85 DM-PI:0.15	FUS6 COP9 COP8	ubiquitin-dependent protein catabolic process; multicellular organismal development; photomorphogenesis; protein catabolic process;	no
AT4G39040	28	na	4.7	AT-GN:0.77 AT-CX:0.23	VAR2 AT3G03710 CHLI1	na	yes
AT1G29070	29	na	4.62	AT-GN:0.54 AT-CX:0.46	ALB3 CHLI1	translation; ribosome biogenesis and assembly;	no
AT4G34730	30	na	4.6	AT-GN:1.00	VAR2 AT3G03710	rRNA processing;	no
AT1G49530	31	GGPS 6	4.59	AT-GN:0.70 AT-PG:0.30	CLA1 DXR	isoprenoid biosynthetic process;	no

AT2G23800	32	GGPS2	4.59	AT-GN:0.70 AT-PG:0.30	CLA1 DXR	isoprenoid biosynthetic process;	no
AT1G12800	33	na	4.59	AT-CX:0.66 AT-DC:0.34	AT1G08520 AT3G03710 CHLI1 DXR	na	yes
AT5G40950	34	na	4.57	AT-CX:1.00	CHLI1	translation;	no
AT4G26430	35	CSN6B	4.56	HS-LC:0.85 DM-PI:0.15	FUS6 COP9 COP8	protein deneddylation; ubiquitin-dependent protein catabolic process; multicellular organismal development;	yes
AT2G13440	36	na	4.53	AT-GN:0.63 CE-CX:0.37	ALB3 APG3	electron transport; tRNA processing;	yes
AT3G04770	37	na	4.53	AT-PG:0.57 AT-GN:0.43	ALB3 AT3G51820 DXR	translation;	yes
AT4G25730	38	na	4.52	AT-GN:1.00	VAR2 AT3G03710	na	no
AT5G13830	39	na	4.51	AT-GN:1.00	VAR2 AT3G03710	na	no
AT2G21350	40	na	4.47	AT-GN:1.00	VAR2 AT3G03710	na	yes
AT3G57150	41	NAP57	4.46	AT-GN:1.00	AT3G03710 DXR	RNA processing;	yes
AT1G21160	42	na	4.45	AT-GN:1.00	VAR2 AT3G03710	translation;	yes
AT1G76825	43	na	4.45	AT-GN:1.00	AT3G03710 DXR	translational initiation;	no
AT2G27700	44	na	4.44	AT-GN:1.00	VAR2 AT3G03710	translation;	yes
AT1G76720	45	na	4.42	AT-GN:1.00	AT3G03710 DXR	translational initiation;	yes
AT1G13270	46	MAP1C	4.41	AT-GN:0.60 AT-CX:0.40	AT3G48500 DXR	proteolysis; N-terminal protein amino acid modification;	yes
AT2G40490	47	na	4.3	AT-CX:1.00	CHLI1	porphyrin biosynthetic process;	no
AT3G08740	48	na	4.29	AT-CX:1.00	CHLI1	translational elongation;	yes
AT1G72370	49	P40	4.24	AT-GN:0.52 AT-CX:0.24 AT-PG:0.24	AT3G51820 CHLI1 DXR	translation; mature ribosome assembly;	no
AT1G51580	50	na	4.23	AT-DC:0.61 AT-GN:0.39	VAR2 AT3G03710	na	no
AT2G44520	51	na	4.21	AT-DC:1.00	AT3G51820	heme biosynthetic process;	no
AT4G23660	52	ATPPT1	4.21	AT-DC:1.00	AT3G51820	biosynthetic process;	no
AT2G32480	53	na	4.15	AT-GN:1.00	DXR	proteolysis;	no
AT1G05140	54	na	4.15	AT-GN:1.00	DXR	proteolysis;	yes
AT3G13882	55	na	4.15	AT-GN:1.00	ALB3	translation;	no
AT5G58770	56	na	4.15	AT-GN:1.00	DXR	dolichol biosynthetic process;	no
AT5G58780	57	na	4.14	AT-GN:1.00	DXR	dolichol biosynthetic process;	yes
AT2G23400	58	na	4.14	AT-GN:1.00	DXR	dolichol biosynthetic process;	no
AT3G09310	59	na	4.14	AT-GN:1.00	ALB3	na	yes

AT2G18640	60	GGPS4	4.14	AT-GN:1.00	CLA1	isoprenoid biosynthetic process;	no
AT5G58782	61	na	4.14	AT-GN:1.00	DXR	dolichol biosynthetic process;	no
AT3G20160	62	na	4.14	AT-GN:1.00	CLA1	isoprenoid biosynthetic process;	yes
AT3G32040	63	na	4.14	AT-GN:1.00	CLA1	isoprenoid biosynthetic process;	no
AT3G14510	64	na	4.14	AT-GN:1.00	CLA1	isoprenoid biosynthetic process;	no
AT3G29430	65	na	4.14	AT-GN:1.00	CLA1	isoprenoid biosynthetic process;	no
AT5G58784	66	na	4.14	AT-GN:1.00	DXR	dolichol biosynthetic process;	no
AT4G38460	67	na	4.14	AT-GN:1.00	CLA1	isoprenoid biosynthetic process;	no
AT2G18620	68	na	4.14	AT-GN:1.00	CLA1	isoprenoid biosynthetic process;	no
AT2G23410	69	ACPT	4.14	AT-GN:1.00	DXR	dolichol biosynthetic process;	yes
AT3G14530	70	na	4.14	AT-GN:1.00	CLA1	isoprenoid biosynthetic process;	no
AT4G36810	71	GGPS1	4.13	AT-GN:1.00	CLA1	isoprenoid biosynthetic process;	no
AT3G60620	72	na	4.13	AT-GN:1.00	DXR	phospholipid biosynthetic process;	no
ATCG01120	73	na	4.13	AT-GN:1.00	AT3G03710	translation;	no
AT3G01800	74	na	4.13	AT-GN:1.00	DXR	translation;	no
AT1G78010	75	na	4.13	AT-GN:1.00	ALB3	tRNA modification;	yes
AT3G14550	76	GGPS3	4.13	AT-GN:1.00	CLA1	isoprenoid biosynthetic process;	no
AT2G45150	77	na	4.13	AT-GN:1.00	DXR	phospholipid biosynthetic process;	yes
AT5G60500	78	na	4.13	AT-GN:1.00	DXR	metabolic process;	no
AT5G64150	79	na	4.13	AT-GN:0.70 DM-PI:0.30	APG3	protein amino acid methylation;	no
AT3G63190	80	na	4.13	AT-GN:1.00	DXR	translation;	no
AT5G60510	81	na	4.13	AT-GN:1.00	DXR	metabolic process;	yes
AT4G05420	82	DDB1A	4.13	HS-LC:1.00	COP1 DET1	negative regulation of photomorphogenesis; negative regulation of transcription;	yes
AT2G17570	83	na	4.13	AT-GN:1.00	DXR	metabolic process;	no
AT3G18680	84	na	4.12	AT-GN:1.00	DXR	pyrimidine nucleotide biosynthetic process;	no
AT3G10030	85	na	4.12	AT-GN:1.00	DXR	amino acid biosynthetic process;	no
AT4G11120	86	na	4.11	AT-GN:1.00	DXR	translational elongation;	yes
AT4G22340	87	na	4.1	AT-GN:1.00	DXR	phospholipid biosynthetic process;	yes
AT4G26770	88	na	4.1	AT-GN:1.00	DXR	phospholipid biosynthetic process;	yes
AT1G62430	89	ATCDS1	4.1	AT-GN:1.00	DXR	phospholipid biosynthetic process;	yes

AT1G69190	90	na	4.09	AT-GN:1.00	VAR2	folic acid and derivative biosynthetic process;	no
AT4G30000	91	na	4.08	AT-GN:1.00	VAR2	folic acid and derivative biosynthetic process;	no
AT3G03600	92	na	4.08	AT-GN:1.00	DXR	translation;	no
AT4G29060	93	na	4.07	AT-GN:1.00	DXR	translational elongation; embryonic development ending in seed dormancy;	no
AT5G05520	94	na	4.07	AT-GN:1.00	DXR	na	no
ATCG00160	95	na	4.07	AT-GN:1.00	DXR	translation;	no
AT3G24560	96	RSY3	4.06	AT-GN:1.00	VAR2	chloroplast organization and biogenesis; embryonic development ending in seed dormancy; suspensor development;	no
AT5G63460	97	na	4.06	AT-DC:1.00	AT3G04260	na	yes
AT5G66840	98	na	4.06	AT-DC:1.00	AT3G04260	na	no
AT4G39680	99	na	4.05	AT-DC:1.00	AT3G04260	na	yes
AT5G10160	100	na	4.04	AT-GN:1.00	DXR	fatty acid biosynthetic process;	no
AT1G09940	101	HEMA2	4.04	AT-GN:1.00	APG3	porphyrin biosynthetic process;	yes
AT2G31250	102	na	4.04	AT-GN:1.00	APG3	porphyrin biosynthetic process;	yes
AT1G58290	103	HEMA1	4.04	AT-GN:1.00	APG3	porphyrin biosynthetic process; heme biosynthetic process; response to light stimulus; chlorophyll biosynthetic process;	yes
AT2G22230	104	na	4.03	AT-GN:1.00	DXR	fatty acid biosynthetic process;	yes
AT5G14460	105	na	4.03	AT-GN:1.00	AT3G03710	RNA modification;	yes
AT1G60600	106	na	4.02	AT-DC:1.00	AT3G51820	photosynthetic electron transport in photosystem II; plastoquinone biosynthetic process; phylloquinone biosynthetic process;	yes
AT1G62850	107	na	4.01	AT-DC:1.00	APG3	na	no
AT1G26830	108	na	4	AT-LC:0.51 HS-LC:0.49	FUS6 COP8	ubiquitin-dependent protein catabolic process; cell cycle; response to red or far red light; embryonic development ending in seed dormancy; positive regulation of flower development; endosperm development;	yes
AT4G21100	109	DDB1B	4	HS-LC:1.00	COP1 DET1	embryonic development ending in seed dormancy;	yes

AT5G18070	110	DRT101	4	AT-GN:1.00	VAR2	photoreactive repair; response to UV;	yes
AT5G50110	111	na	3.99	AT-GN:1.00	ALB3	cell cycle;	yes
AT4G29540	112	na	3.99	AT-GN:1.00	DXR	na	no
AT2G04560	113	na	3.97	AT-GN:1.00	DXR	na	no
AT5G58370	114	na	3.92	AT-GN:1.00	ALB3	na	yes
AT3G25470	115	na	3.92	AT-GN:1.00	CLA1	hemolysis by symbiont of host red blood cells;	no
AT5G04130	116	na	3.9	AT-GN:1.00	ALB3	DNA metabolic process; DNA topological change;	yes
AT5G60410	117	na	3.89	AT-DC:1.00	AT3G04260	na	no
AT3G13440	118	na	3.85	AT-GN:1.00	APG3	na	no
AT4G05210	119	na	3.85	AT-GN:1.00	DXR	na	no
AT3G23890	120	TOPII	3.84	AT-GN:1.00	ALB3	DNA metabolic process; DNA topological change;	no
AT4G21220	121	na	3.79	AT-GN:1.00	DXR	na	no
AT1G68590	122	na	3.79	AT-CX:1.00	CHLI1	translation;	yes
AT2G41460	123	ARP	3.78	AT-DC:1.00	AT3G04260	positive regulation of transcription;	yes
AT1G60080	124	na	3.78	AT-DC:1.00	AT3G03710	RNA processing;	no
AT4G27490	125	na	3.78	AT-DC:1.00	AT3G03710	RNA processing;	no
AT3G60500	126	na	3.78	AT-DC:1.00	AT3G03710	RNA processing;	no
AT3G07750	127	na	3.78	AT-DC:1.00	AT3G03710	RNA processing;	no
AT3G12990	128	na	3.78	AT-DC:1.00	AT3G03710	RNA processing;	no
AT4G02390	129	APP	3.78	AT-DC:1.00	AT3G04260	protein amino acid ADP-ribosylation;	yes
AT4G08170	130	na	3.74	HS-LC:1.00	FUS6 COP9	na	yes
AT3G11070	131	na	3.71	AT-GN:1.00	DXR	na	yes
AT2G25100	132	na	3.71	AT-GN:1.00	DXR	na	no
AT4G28706	133	na	3.68	HS-CX:1.00	PDS1	D-ribose metabolic process;	no
AT3G25740	134	MAP1B	3.68	AT-GN:1.00	DXR	proteolysis; N-terminal protein amino acid modification;	no
AT1G16970	135	na	3.67	AT-DC:1.00	AT3G04260	telomere maintenance; DNA repair;	yes
AT2G45240	136	MAP1A	3.66	AT-GN:1.00	DXR	protein processing; N-terminal protein amino acid modification;	yes
AT1G76990	137	ACR3	3.64	AT-GN:0.68 AT-PG:0.32	DXR	metabolic process;	yes
AT5G08280	138	na	3.62	AT-CX:1.00	CHLI1	porphyrin biosynthetic process; chlorophyll biosynthetic process;	no
AT4G10070	139	na	3.61	AT-DC:1.00	AT3G03710	na	yes
AT5G53060	140	na	3.61	AT-DC:1.00	AT3G03710	na	yes
AT4G18375	141	na	3.61	AT-DC:1.00	AT3G03710	na	yes
AT5G09560	142	na	3.61	AT-DC:1.00	AT3G03710	na	no
AT2G03110	143	na	3.61	AT-DC:1.00	AT3G03710	na	no

AT4G26000	144	na	3.61	AT-DC:1.00	AT3G03710	shoot development; gynoecium development;	no
AT5G46190	145	na	3.61	AT-DC:1.00	AT3G03710	na	yes
AT2G22600	146	na	3.61	AT-DC:1.00	AT3G03710	na	yes
AT5G64390	147	HEN4	3.61	AT-DC:1.00	AT3G03710	mRNA processing; specification of floral organ identity;	no
AT5G15270	148	na	3.61	AT-DC:1.00	AT3G03710	na	yes
AT1G33680	149	na	3.61	AT-DC:1.00	AT3G03710	na	yes
AT1G14170	150	na	3.61	AT-DC:1.00	AT3G03710	na	no
AT5G04430	151	na	3.61	AT-DC:1.00	AT3G03710	RNA splicing;	yes
AT2G25970	152	na	3.61	AT-DC:1.00	AT3G03710	na	no
AT3G49870	153	na	3.6	AT-GN:1.00	AT3G03710	rRNA processing; intracellular protein transport; small GTPase mediated signal transduction; protein transport; ribosome biogenesis and assembly;	yes
AT2G24580	154	na	3.58	HS-CX:1.00	PDS1	tetrahydrofolate metabolic process;	no
AT5G54080	155	HGO	3.58	AT-GN:0.62 AT-PG:0.38	PDS1	L-phenylalanine catabolic process; tyrosine catabolic process;	no
AT4G32520	156	SHM3	3.58	HS-CX:0.65 AT-GN:0.35	PDS1 APG3	glycine metabolic process; L-serine metabolic process;	no
AT5G04110	157	na	3.56	AT-GN:1.00	ALB3	DNA metabolic process; DNA topological change;	yes
AT4G37040	158	MAP1D	3.56	AT-GN:1.00	DXR	proteolysis; N-terminal protein amino acid modification;	no
AT3G12130	159	na	3.55	AT-DC:1.00	AT3G03710	regulation of transcription;	yes
AT5G06770	160	na	3.55	AT-DC:1.00	AT3G03710	regulation of transcription;	yes
AT5G45930	161	CHLI2	3.55	AT-GN:1.00	AT1G08520 CHLI1	chlorophyll biosynthetic process;	yes
AT4G02510	162	TOC159	3.51	AT-GN:1.00	ALB3	protein targeting to chloroplast;	no
AT1G15810	163	na	3.5	AT-GN:0.64 AT-CX:0.36	AT3G03710 CHLI1	translation;	yes
AT5G67560	164	na	3.49	AT-GN:1.00	AT3G03710	rRNA processing; intracellular protein transport; small GTPase mediated signal transduction; protein transport; ribosome biogenesis and assembly;	yes
AT5G11480	165	na	3.48	AT-GN:1.00	ALB3	na	yes

AT3G49860	166	na	3.45	AT-GN:1.00	AT3G03710	rRNA processing; intracellular protein transport; small GTPase mediated signal transduction; protein transport; ribosome biogenesis and assembly;	no
AT4G28660	167	na	3.44	AT-CX:1.00	CHLI1	photosynthesis;	no
AT4G00090	168	na	3.43	HS-DC:0.42 AT-PG:0.29 CE-CX:0.28	COP1 APG3	na	no
AT2G25910	169	na	3.43	AT-DC:1.00	AT3G03710	na	yes
AT3G10270	170	na	3.42	AT-GN:1.00	ALB3	DNA metabolic process; DNA topological change;	no
AT1G49880	171	na	3.41	HS-LC:1.00	FUS6 COP9	na	no
AT4G14090	172	na	3.39	HS-CX:1.00	PDS1	metabolic process;	no
AT2G36780	173	na	3.39	HS-CX:1.00	PDS1	metabolic process;	no
AT3G46670	174	na	3.39	HS-CX:1.00	PDS1	metabolic process;	no
AT1G05560	175	UGT1	3.39	HS-CX:1.00	PDS1	response to salicylic acid stimulus; cell plate formation involved in cellulose and pectin-containing cell wall biogenesis;	no
AT5G59580	176	na	3.39	HS-CX:1.00	PDS1	metabolic process;	no
AT3G55700	177	na	3.39	HS-CX:1.00	PDS1	metabolic process;	yes
AT2G36770	178	na	3.39	HS-CX:1.00	PDS1	metabolic process;	no
AT5G05860	179	UGT76C2	3.39	HS-CX:1.00	PDS1	metabolic process;	yes
AT1G24100	180	na	3.39	HS-CX:1.00	PDS1	glucosinolate biosynthetic process;	no
AT5G59590	181	na	3.39	HS-CX:1.00	PDS1	metabolic process;	no
AT4G15260	182	na	3.39	HS-CX:1.00	PDS1	metabolic process;	yes
AT2G15480	183	na	3.39	HS-CX:1.00	PDS1	response to other organism;	yes
AT3G04610	184	na	3.37	AT-DC:1.00	AT3G03710	positive regulation of flower development;	no
AT1G01910	185	na	3.34	HS-CX:1.00	FUS6	anion transport;	no
AT5G42270	186	VAR1	3.29	AT-LC:0.62 AT-GN:0.38	VAR2 AT3G03710	PSII associated light-harvesting complex II catabolic process;	no
AT1G03360	187	na	3.29	AT-DC:1.00	AT3G03710	na	no
AT3G23700	188	na	3.29	AT-DC:1.00	AT3G03710	response to cold;	no
AT1G71720	189	na	3.29	AT-DC:1.00	AT3G03710	na	yes
AT4G24830	190	na	3.29	HS-CX:1.00	PDS1	arginine biosynthetic process;	no
AT5G37680	191	na	3.27	AT-GN:1.00	AT3G03710	rRNA processing; intracellular protein transport; small GTPase mediated signal transduction; protein transport; ribosome biogenesis and	yes

						assembly;	
AT2G24790	192	na	3.25	AT-LC:0.65 AT-CX:0.35	COP1 CHLI1	regulation of photomorphogenesis; red light signaling pathway;	yes
AT2G15490	193	na	3.21	HS-CX:1.00	PDS1	response to other organism;	yes
AT1G20560	194	na	3.18	HS-CX:0.56 CE-CX:0.44	PDS1	metabolic process;	no
AT4G14520	195	na	3.16	AT-DC:1.00	AT3G03710	na	no
AT2G42220	196	na	3.14	AT-CX:1.00	CHLI1 DXR	na	yes
AT4G33770	197	na	3.14	HS-LC:1.00	FUS6 COP9	na	yes
AT4G22780	198	ACR7	3.12	AT-GN:1.00	DXR	metabolic process;	yes
AT3G46660	199	na	3.09	HS-CX:1.00	PDS1	metabolic process;	yes
AT1G52280	200	na	3.07	AT-GN:1.00	ALB3	intracellular protein transport; small GTPase mediated signal transduction; protein transport;	no

Supplementary Table 6. SALK T-DNA lines of the seed pigmentation candidate genes tested.

Seed Stock	insertion site	Gene	LLS	Rank	seed pigmentation defect?
SALK_113234C	~15bp downstream from 3' UTR	AT3G57290	6.44	2	no
SALK_054763C	5' Promoter ~ 30bp upstream from 5' UTR	AT2G26990	6.06	3	no
SALK_151350C	8th exon	AT5G15610	6.01	5	yes
SALK_147710C	1st exon - exactly at ATG	AT5G45620	5.84	8	yes
SALK_018378C	4th intron	AT5G45620	5.84	8	yes
SALK_152096C	5' Promoter ~300bp upstream from 5' UTR	AT5G13630	5.74	9	yes
SALK_093768C	4th exon	AT5G07590	5.7	10	no
SALK_036658C	5' UTR	AT1G71230	5.37	12	no
SALK_007134C	2nd exon	AT1G71230	5.37	12	no
SALK_049248C	7th exon - last exon	AT1G75990	5.14	13	no
SALK_088176C	1st exon	AT1G75990	5.14	13	yes
SALK_133892C	9th exon towards 3' end	AT5G64760	5.12	17	no
SALK_017454C	5' Promoter ~150bp upstream from 5' UTR	AT5G09900	5.12	18	no
SALK_015320C	5' promoter ~125 bp upstream from 5' UTR	AT1G17220	5.03	20	no
SALK_136612C	5' promoter ~ 300bp upstream from start codon	AT3G22860	5.02	21	no
SALK_011380C	8th exon - last exon	AT4G11160	4.89	22	no
SALK_128966C	2nd exon	AT4G11160	4.89	22	no
SALK_035918C	5' promoter ~ 100bp from UTR	AT5G01230	4.82	24	no
SALK_036405C	5' Promoter ~275bp upstream from 5' UTR	AT4G39040	4.7	28	no
SALK_046738C	1st exon	AT1G12800	4.59	33	no
SALK_030714C	5' promoter ~ 100bp from UTR	AT1G12800	4.59	33	yes
SALK_036965C	3rd intron	AT4G26430	4.56	35	yes
SALK_049514C	5' UTR	AT4G26430	4.56	35	yes
SALK_100713C	2nd exon	AT2G13440	4.53	36	no
SALK_131338C	5' Promoter ~ 175bp upstream from 5' UTR	AT3G04770	4.53	37	no
SALK_135983C	1st exon ~ 150bp downstream from start codon	AT2G21350	4.47	40	no
SALK_023066C	5' Promoter ~200bp upstream from 5' UTR	AT3G57150	4.46	41	no
SALK_047254C	5th intron	AT1G21160	4.45	42	no
SALK_109541C	5' Promoter ~ 175bp upstream from start codon	AT2G27700	4.44	44	no
SALK_143304C	3rd intron	AT1G76720	4.42	45	no
SALK_124755C	2nd exon	AT1G76720	4.42	45	no
SALK_027575C	5' promoter ~ 200bp upstream from UTR	AT1G13270	4.41	46	no
SALK_064599C	5' UTR	AT1G13270	4.41	46	no
SALK_120844C	5' Promoter ~ 90bp upstream from 5' UTR	AT3G08740	4.29	48	no

SALK_071288C	5' UTR	AT1G05140	4.15	54	no
SALK_147556C	5' Promoter ~ 175bp upstream from 5' UTR	AT5G58780	4.14	57	no
SALK_057096C	1st exon ~ 50bp downstream from start codon	AT3G09310	4.14	59	no
SALK_038548C	~410bp upstream from start codon	AT3G20160	4.14	62	yes
SALK_100795C	3' end - ~ 50bp from 3' UTR	AT2G23410	4.14	69	no
SALK_032276C	2nd exon	AT2G23410	4.14	69	no
SALK_076607C	6th exon - middle of the gene	AT1G78010	4.13	75	no
SALK_106720C	4th exon - middle of the gene	AT1G78010	4.13	75	no
SALK_115705C	1st exon	AT2G45150	4.13	77	no
SALK_106884C	5' promoter ~ 200bp upstream from UTR	AT5G60510	4.13	81	no
SALK_055584C	4th intron	AT4G05420	4.13	82	no
SALK_007854C	8th intron	AT4G11120	4.11	86	no
SALK_106246C	5' UTR	AT4G22340	4.1	87	no
SALK_082197C	Polymorphism site in Gene AT4G26780	AT4G26770	4.1	88	no
SALK_001496C	11th intron - towards 3' end	AT1G62430	4.1	89	no
SALK_088268C	1st exon	AT1G62430	4.1	89	no
SALK_081993C	1st intron	AT5G63460	4.06	97	no
SALK_132910C	5' promoter ~ 200bp from UTR	AT5G63460	4.06	97	yes
SALK_047712C	1st exon	AT4G39680	4.05	99	no
SALK_061742C	5' uTR	AT1G09940	4.04	101	no
SALK_084047C	5' promoter ~ 200bp upstream from start codon	AT2G31250	4.04	102	no
SALK_032256C	3rd exon - last exon just before stop codon	AT2G31250	4.04	102	no
SALK_053036C	3rd exon- middle of gene	AT1G58290	4.04	103	no
SALK_026580C	3rd exon- middle of gene	AT2G22230	4.03	104	no
SALK_086767C	2nd exon	AT2G22230	4.03	104	no
SALK_082735C	5' Promoter ~50bp upstream from 5' UTR	AT5G14460	4.03	105	no
SALK_021962C	5th intron – middle of gene	AT1G60600	4.02	106	no
SALK_050756C	2nd exon; towards the 3' UTR	AT1G26830	4	108	no
SALK_061944C	19th exon - last exon	AT4G21100	4	109	no
SALK_096148C	5' promoter ~ 15bp upstream from UTR	AT5G18070	4	110	no
SALK_039132C	5' promoter ~ 40bp upstream from UTR	AT5G18070	4	110	no
SALK_027109C	1st intron	AT5G50110	3.99	111	yes
SALK_086197C	5' UTR	AT5G50110	3.99	111	yes
SALK_036661C	3rd exon - in the middle of the gene	AT5G58370	3.92	114	no
SALK_060321C	5' Promoter ~150bp upstream from 5' UTR	AT5G04130	3.9	116	no
SALK_104063C	5' utR	AT1G68590	3.79	122	no
SALK_021009C	5' promoter -200 bp upstream of ATG	AT2G41460	3.78	123	no
SALK_140400C	6th intron	AT4G02390	3.78	129	no
SALK_097261C	13th intron	AT4G02390	3.78	129	no
SALK_123871C	1st intron	AT4G08170	3.74	130	no

SALK_120653C	3rd intron	AT4G08170	3.74	130	no
SALK_048769C	5' promoter -100 bp from 5' UTR	AT3G11070	3.71	131	no
SALK_106654C	5' promoter 8bp upstream from UTR	AT1G16970	3.67	135	no
SALK_123114C	8th exon - middle of the gene	AT1G16970	3.67	135	no
SALK_097303C	14th exon - towards 3' end	AT2G45240	3.66	136	no
SALK_021985C	1st exon	AT2G45240	3.66	136	yes
SALK_032604C	1st intron	AT1G76990	3.64	137	no
SALK_064756C	5' UTR	AT1G76990	3.64	137	no
SALK_000033C	7th exon	AT4G10070	3.61	139	no
SALK_013918C	3rd exon - towards 5' end	AT5G53060	3.61	140	no
SALK_016188C	6th exon	AT4G18375	3.61	141	no
SALK_051182C	3rd exon - towards 3' end	AT5G46190	3.61	145	no
SALK_047259C	1st exon	AT5G46190	3.61	145	no
SALK_048634C	5' Promoter ~ 75bp upstream from start codon	AT2G22600	3.61	146	no
SALK_126569C	1st intron	AT5G15270	3.61	148	no
SALK_121893C	5' Promoter ~ 240bp upstream from 5' UTR	AT1G33680	3.61	149	no
SALK_117242C	5' UTR	AT5G04430	3.61	151	no
SALK_059077C	5' Promoter ~250bp upstream from 5' UTR	AT3G49870	3.6	153	no
SALK_108979C	1st intron	AT5G04110	3.56	157	yes
SALK_057095C	3rd exon – last exon	AT3G12130	3.55	159	no
SALK_057355C	5' Promoter ~275bp upstream from 5' UTR	AT3G12130	3.55	159	no
SALK_014716C	2nd intron – middle of the gene	AT5G06770	3.55	160	no
SALK_105370C	5' promoter ~ 250bp upstream from UTR	AT5G45930	3.55	161	no
SALK_010288C	5' promoter ~ 250bp upstream from UTR	AT1G15810	3.5	163	no
SALK_077021C	3rd intron	AT1G15810	3.5	163	no
SALK_081093C	5' UTR	AT5G67560	3.49	164	no
SALK_018461C	3' end after stop codon	AT5G67560	3.49	164	yes
SALK_029559C	1st exon	AT5G11480	3.48	165	yes
SALK_080472C	7th exon	AT2G25910	3.43	169	no
SALK_046282C	2nd exon – last exon	AT3G55700	3.39	177	no
SALK_010205C	1st exon	AT5G05860	3.39	179	no
SALK_135793C	2nd exon	AT5G05860	3.39	179	no
SALK_094287C	only 1 exon - towards 3' end	AT4G15260	3.39	182	no
SALK_039472C	only 1 exon - towards 3' end	AT4G15260	3.39	182	yes
SALK_078055C	5' Promoter ~ 15bp upstream from 5' UTR	AT2G15480	3.39	183	no
SALK_127604C	5' promoter ~100bp upstream from start codon	AT1G71720	3.29	189	no
SALK_107226C	7th exon - towards 3 ' end	AT1G71720	3.29	189	no
SALK_129296C	5' promoter 93bp upstream from 5' UTR	AT5G37680	3.27	191	no
SALK_040211C	5' Promoter ~ 330bp upstream from 5' UTR	AT2G24790	3.25	192	no
SALK_061595C	5' Promoter ~ 75bp upstream from 5'	AT2G15490	3.21	193	no

	UTR				
SALK_045769C	3' UTR ~20bp downstream from stop codon	AT2G42220	3.14	196	no
SALK_147144C	7th exon	AT4G33770	3.14	197	no
SALK_019532C	5' UTR - just two bases before start codon	AT4G22780	3.12	198	no
SALK_021844C	5' Promoter ~75bp upstream from 5' UTR	AT3G46660	3.09	199	no

Supplementary Table 7. Mutant lines with pale or purple leaves and seedling morphology defects in 1% sucrose agar plates.

mutant #	Seed stock	gene	LLS	Rank	Expressivity	Number of alleles with phenotype/ total tested
21	SALK_151350C	AT5G15610	6.01	5	100.00%	1/1
31	SALK_147710C	AT5G45620	5.84	8	50.00%	2/2
120	SALK_018378C	AT5G45620	5.84	8	66.70%	2/2
134	SALK_152096C	AT5G13630	5.74	9	50.00%	1/1
196	SALK_088176C	AT1G75990	5.14	13	57.10%	1/2
18	SALK_030714C	AT1G12800	4.39	33	55.60%	1/2
67	SALK_036965C	AT4G26430	4.56	35	100.00%	2/2
14	SALK_049514C	AT4G26430	4.56	35	100.00%	2/2
2	SALK_038548C	AT3G20160	4.14	62	55.60%	1/1
256	SALK_132910C	AT5G63460	4.06	97	57.10%	1/2
197	SALK_027109C	AT5G50110	3.99	111	44.40%	2/2
225	SALK_086197C	AT5G50110	3.99	111	60.00%	2/2
150	SALK_021985C	AT2G45240	3.66	136	100.00%	1/2
258	SALK_108979C	AT5G04110	3.56	157	75.00%	1/1
104	SALK_018461C	AT5G67560	3.49	164	88.90%	1/2
26	SALK_029559C	AT5G11480	3.48	165	83.30%	1/1
15	SALK_039472C	AT4G15260	3.39	182	55.60%	1/2

Supplementary Table 8. Survival rate of mutants in soil (two weeks after transfer from agar plates to soil)**.**

Gene	***Seed Stock number***	***Total number transplanted to soil***	***Number of abnormal phenotype in soil***	***Number died***	***% abnormal***	***% survival***
AT4G26430	SALK_049514C	4	1	1	25.0%	75.0%
AT4G26430	SALK_036965C	11	0	0	0.0%	100.0%
AT5G45620	SALK_147710C	8	0	3	0.0%	62.5%
AT5G45620	SALK_018378C	13	0	1	0.0%	92.3%
AT5G50110	SALK_027109C	15	3	6	20.0%	60.0%
AT5G50110	SALK_086197C	6	1	4	16.7%	33.3%
AT3G20160	SALK_038548C	10	0	6	0.0%	40.0%
AT5G04110	SALK_108979C	2	0	0	0.0%	100.0%
AT5G11480	SALK_029559C	8	0	0	0.0%	100.0%
AT5G13630	SALK_152096C	6	0	0	0.0%	100.0%
AT5G15610	SALK_151350C	6	0	1	0.0%	83.3%

Supplementary Table 9. Annotations of newly identified and known seed pigmentation genes that are connected into subnetworks.

Component	AGI locus	Gene symbol	known or new (number of supporting alleles)	Biological process/role	Protein function description
1	AT3G61140	CSN1	Known	Photomorphogenesis, derubylation of CRL families of ubiquitin E3 ligase complexes	CSN complex subunit containing PCI domain[59]
1	AT5G42970	CSN4	Known	Photomorphogenesis, derubylation of CRL families of ubiquitin E3 ligase complexes	CSN complex subunit containing PCI domain[59]
1	AT1G02090	CSN7	Known	Photomorphogenesis, derubylation of CRL families of ubiquitin E3 ligase complexes	CSN complex subunit containing PCI domain[59]
1	AT4G14110	CSN8	Known	Photomorphogenesis, derubylation of CRL families of ubiquitin E3 ligase complexes	CSN complex subunit containing PCI domain[59]
1	AT4G26430	CSN6B	New (2)	Photomorphogenesis, derubylation of CRL families of ubiquitin E3 ligase complexes	CSN complex subunit containing PCI domain[60]
1	AT5G45620	AT5G45620	New (2)	Unknown	Unknown, contains PCI domain and has sequence homology to RPN9, which is a subunit of the lid subcomplex of 26S proteosome
1	AT5G15610	AT5G15610	New (1)	Unknown	Unknown, contains PCI domain (TAIR)
2	AT2G32950	COP1	Known	Photomorphogenesis, ubiquitin-mediated protein degradation	Ubiquitin E3 ligase[61]
2	AT4G10180	DET1	Known	Photomorphogenesis, ubiquitin-mediated protein degradation	Forms a complex with COP10 and DDB1 and regulates the activity of ubiquitin E2 congugating enzymes[62]
3	AT2G28800	ALB3	Known	Thylakoid membrane biogenesis, translocation of membrane proteins into the thylakoid	A membrane-bound translocase that interacts with chloroplast signal recognition particle complex to insert membrane-bound proteins into the thylakoid membrane[63]

				membrane	
3	AT5G50110	AT5G50110	New (2)	Unknown	Unknown, has sequence similarity to methyltransferases and has a domain found in bacterial glucose inhibited division proteins, is computationally predicted to be a thylakoid luminal protein[64]
3	AT5G11480	AT5G11480	New (1)	Unknown	Unknown, has GTP-binding domain and sequence similarity to an embryo defective mutant EMB2001 (TAIR)
3	AT5G04110	AT5G04110	New (1)	Unknown	Unknown, has sequence similarity to DNA topoisomerase II (TAIR)
4	AT4G18480	CHL1	Known	Chlorophyll biosynthesis, Mg branch	Mg chelatase subunit, which inserts Mg into protoporphyrin IX[65]
4	AT1G08520	CHLD	Known	Chlorophyll biosynthesis, Mg branch	Mg chelatase subunit, which inserts Mg into protoporphyrin IX[65]
4	AT3G51820	CHLG	Known	Chlorophyll biosynthesis, Mg branch	Chlorophyll synthase[65]
4	AT5G13630	CHLH	New (1)	Chlorophyll biosynthesis, Mg branch	A multifunctional protein that binds to abscisic acid[66], regulates plastid-to-nucleus signaling[67], and is a Mg chelatase subunit, which inserts Mg into protoporphyrin IX[65]
5	AT4G15560	DXS	Known	2-C-methyl-D-erythritol 4-phosphate (MEP) pathway, chlorophyll biosynthesis	1-deoxy-d-xylulose 5-phosphate synthase, which catalyzes the first committed step of MEP pathway that produces precursors for chlorophyll biosynthesis[65]
5	AT3G20160	AT3G20160	New (1)	Unknown	Has sequence similarity to geranylgeranyl pyrophosphate (GGPP) synthase (TAIR). GGPP is made via MEP pathway and is the source of the phytol that is used to make chlorophyll a[65]

Supplementary Table 10. AraNet-predicted Gene Onotology annotations for the 3 uncharacterized genes tested. The genes were entered in the text box of 'Advance Search' of AraNet website (http://www.functionalnet.org/aranet/cgi-perl/AraNet.v1_apn_form.cgi) to retrieve these predictions.

AT1G80710				
Rank	Score	Evidence	GO_term	GO_term_supporters (LLS)
1	4.78	SC-MS:1.00	Gene silencing	AT2G24490(3.19)
2	4.78	SC-MS:1.00	DNA repair	AT2G24490(3.19)
3	3.33	SC-MS:1.00	response to water deprivation	HIS4(3.33)
4	3.28	SC-MS:1.00	double-strand break repair via homologous recombination	AT1G10930(2.19)
5	3.28	SC-MS:1.00	response to DNA damage stimulus	AT1G10930(2.19)
6	3.19	SC-MS:1.00	DNA replication	AT2G24490(3.19)
7	2.71	AT-PG:0.49 SC-MS:0.35 SC-DC:0.15	ER to Golgi vesicle-mediated transport	AT3G52190(1.68) STL2P(1.14) ATRAB1B(0.92)
8	2.66	SC-DC:0.54 AT-PG:0.32 SC-MS:0.14	trichome differentiation	FAS2(1.52) MSI1(1.51) TTG1(1.51)
9	2.27	SC-DC:0.53 AT-PG:0.26 SC-MS:0.21	leaf development	FAS2(1.52) MSI1(1.51)
10	2.27	SC-DC:0.53 AT-PG:0.26 SC-MS:0.21	heterochromatin formation	FAS2(1.52) MSI1(1.51)
AT2G17900				
Rank	Score	Evidences	GO_term	GO_term_supporters (LLS)
1	7.89	HS-MS:1.00	phosphorylation	ATSK11(4.52) ATSK12(4.49)
2	6.77	HS-MS:1.00	meristem organization	ATSK11(4.52) ATSK12(4.49)
3	5.91	HS-MS:0.58 AT-DC:0.42	leaf morphogenesis	BIN2(3.94) CLF(1.78)
4	4.03	HS-MS:1.00	hyperosmotic salinity response	GSK1(4.03)
5	3.94	HS-MS:1.00	brassinosteroid mediated signaling	BIN2(3.94)
6	3.94	HS-MS:1.00	multidimensional cell growth	BIN2(3.94)
7	3.94	HS-MS:1.00	response to auxin stimulus	BIN2(3.94)
8	3.94	HS-MS:1.00	detection of brassinosteroid stimulus	BIN2(3.94)
9	3.94	HS-MS:1.00	protein amino acid phosphorylation	BIN2(3.94)
10	2.68	AT-DC:1.00	imprinting	CLF(1.78) EZA1(1.78)
AT3G05090				
Rank	Score	Evidences	GO_term	GO_term_supporters (LLS)
1	3.97	SC-DC:0.52 AT-PG:0.48	trichome differentiation	FAS2(2.40) MSI1(1.52) TTG1(1.51)

2	3.34	SC-MS:1.00	ubiquitin-dependent protein catabolic process	AT4G24820(1.42) ATS9(1.35) AT5G09900(1.33) AT5G23540(1.24) AT5G64760(1.23) AT1G64520(1.19) AT3G11270(1.17) AT1G75990(1.13) AT1G20200(1.13) AT-MCB1(1.12) ATHMOV34(1.11) ATSUG1(1.02) AT5G45620(1.00) AT1G04810(0.94) UBQ4(0.89) AT2G32730(0.89) AT2G20580(0.86)
3	3.15	AT-PG:0.51 SC-DC:0.49	leaf development	FAS2(2.40) MSI1(1.52)
4	3.15	AT-PG:0.51 SC-DC:0.49	heterochromatin formation	FAS2(2.40) MSI1(1.52)
5	3.15	AT-PG:0.51 SC-DC:0.49	cell proliferation	FAS2(2.40) MSI1(1.52)
6	3.01	SC-MS:1.00	protein catabolic process	AT4G24820(1.42) ATS9(1.35) AT5G23540(1.24) AT5G64760(1.23) AT1G64520(1.19) AT3G11270(1.17) AT1G75990(1.13) AT-MCB1(1.12) AT4G19006(1.04) AT1G04810(0.94) AT2G20580(0.86)
7	2.95	AT-PG:0.75 SC-DC:0.25	ER to Golgi vesicle-mediated transport	AT3G52190(1.97) STL2P(0.93)
8	2.4	AT-PG:0.54 SC-DC:0.46	meristem organization	FAS2(2.40)
9	2.4	AT-PG:0.54 SC-DC:0.46	nucleosome assembly	FAS2(2.40)
10	2.27	SC-DC:0.69 AT-PG:0.31	regulation of flower development	MSI1(1.52) FY(1.37)

Supplementary Table 11. Confirmation of homozygote plants by polymerase chain reaction.

Gene name	Stock number	Approximate insert position (from ATG)	Insertion site (e.g. first exon)	Portion of homozygotes
At1g80710	Salk_001238C	639bp	1st exon	5/5
At1g80710	Salk_149366C	149bp upstream	5' UTR	5/5
At3g05090	Salk_059570C	2833bp	13th exon	3/8
At2g17900	Salk_127952C	2323bp	12th intron	4/4
At1g15772	Salk_118634C	167 bp	2nd exon	3/3
At2g34170	Salk_099804C	986bp	2nd exon	5/5
At5g50110	Salk_027109C	315bp upstream	1st intron	3/3
At5g50110	Salk_086197C	477bp upstream	5' UTR	1 / 2
At4g26430	Salk_036965C	663bp	3rd intron	8/10
At4g26430	Salk_049514C	120bp upstream	5' UTR	2/2
At5g45620	Salk_147710C	267 bp	1st intron	5/5
At5g45620	Salk_018378C	1468 bp	4th exon	5/5
At3g20160	Salk_038548C	between 120-670 bp upstream	5' intergenic	3/4
At5g15610	Salk_151350C	2330 bp	3' UTR	4/5
At5g11480	Salk_029559C	between 220-470 bp	1st exon or 1st intron	ND*
At5g13630	Salk_152096C	between 180-710 bp upstream	5' intergenic	5/5
At5g04110	Salk_108979C	560 upstream	1st intron	2/3

*No homozygotes were recovered from plants transferred to soil from plates. Of the 8 plants tested, 4 were heterozygotes and 4 were homozygotes for the wild type allele.

Supplementary Table 12. Segregation of *drs1-1* (At1g80710) and *lrs1-1* (At3g05090) knock-out alleles in F2 population. Homozygote mutants were crossed with Col-0 wild type and F1 plants were selfed and F2 populations were genotyped by PCR amplification using gene-specific primers and the T-DNA primer (see Supplementary Methods).

Gene	**Genotype**			**Chi square test**			
	-/-	-/+	+/+	Expected ratio	X^2	P-value	df
At1g80710	70	128	61	1:2:1	0.660	0.7188	2
At3g05090	27	70	31	1:2:1	1.375	0.5028	2

Supplementary Table 13. AraNet predictive power measured by the area under cross-validated ROC curves (AUC) for Gene Ontology biological process terms.

Gene Ontology biological process terms	AUC	network coverage	# genes
histidine biosynthetic process	0.9999	1	6
intra-Golgi vesicle-mediated transport	0.9996	1	6
leucine biosynthetic process	0.9996	1	6
protein deneddylation	0.9996	1	5
protein import into nucleus	0.9996	1	6
acetyl-CoA biosynthetic process	0.9994	1	5
porphyrin biosynthetic process	0.9991	1	13
toxin catabolic process	0.999	1	44
ATP-dependent proteolysis	0.9988	1	13
water transport	0.998	1	5
actin filament-based movement	0.9968	1	17
N-terminal protein amino acid modification	0.9965	1	5
calcium ion transport	0.9941	1	5
nuclear mRNA splicing, via spliceosome	0.9928	1	14
membrane fusion	0.9814	1	28
Translation	0.9758	1	54
protein catabolic process	0.9597	1	17
Intracellular protein transport	0.9576	1	24
iron-sulfur cluster assembly	0.955	0.9167	12
cellular respiration	0.9432	1	18
starch catabolic process	0.9413	0.8889	9
protein folding	0.934	1	20
tryptophan biosynthetic process	0.9332	1	15
translational initiation	0.9283	1	8
cytokinin mediated signaling	0.928	0.9429	35
pentose-phosphate shunt	0.9184	1	7
Imprinting	0.9166	1	6
vesicle-mediated transport	0.9151	1	6
negative regulation of photomorphogenesis	0.9129	1	6
glucosinolate biosynthetic process	0.9116	1	12
actin filament organization	0.909	1	11
photosynthesis	0.9044	1	11
proline biosynthetic process	0.8984	1	5
Isopentenyl diphosphate biosynthetic process, mevalonate-independent pathway	0.8978	1	5
starch metabolic process	0.8978	1	5
RNA-mediated posttranscriptional gene silencing	0.8935	1	5
DNA repair	0.8882	1	15
two-component signal transduction system (phosphorelay)	0.8882	0.7778	9
negative regulation of abscisic acid mediated signaling	0.886	0.9	10
regulation of progression through cell cycle	0.8785	1	9
ubiquitin-dependent protein catabolic process	0.877	0.9407	135
brassinosteroid mediated signaling	0.8733	0.875	8
response to gamma radiation	0.8732	1	5
electron transport	0.8645	0.8667	30

Peroxisome organization and biogenesis	0.8569	1	7
nitrate assimilation	0.8545	1	7
sterol biosynthetic process	0.8476	0.95	20
nitrogen compound metabolic process	0.8333	1	6
vacuole organization and biogenesis	0.8332	1	6
regulation of seed germination	0.8323	1	6
sulfate assimilation	0.8314	1	9
photosystem II assembly	0.8284	1	6
cadmium ion transport	0.8281	1	6
photosynthesis, light reaction	0.8281	1	6
response to DNA damage stimulus	0.827	1	6
response to copper ion	0.8228	1	6
response to oxidative stress	0.8198	0.9863	73
phosphorylation	0.8181	1	6
calcium-mediated signaling	0.8151	1	11
ovule development	0.8143	0.8636	22
signal transduction	0.8132	0.8889	9
vernalization response	0.8128	1	11
chloroplast fission	0.8115	1	8
starch biosynthetic process	0.8115	1	8
zinc ion transport	0.8112	1	8
actin cytoskeleton organization and biogenesis	0.8105	1	11
miRNA-mediated gene silencing, production of miRNAs	0.8101	0.875	8
response to iron ion	0.8061	0.875	8
Microtubule cytoskeleton organization and biogenesis	0.8038	1	8
epidermal cell fate specification	0.7992	1	5
ATP synthesis coupled proton transport	0.7991	0.8	5
phosphate transport	0.7973	1	5
glucose mediated signaling	0.7967	1	5
response to heat	0.7965	0.9737	76
response to hypoxia	0.7961	1	5
metal ion transport	0.7934	1	5
ER to Golgi vesicle-mediated transport	0.7909	1	5
pollen tube growth	0.7886	0.9412	17
protein amino acid dephosphorylation	0.7886	1	8
response to high light intensity	0.7886	0.9655	29
Meiosis	0.7881	1	5
fatty acid beta-oxidation	0.7877	1	10
photomorphogenesis	0.787	0.8519	27
Cytokinesis	0.786	1	13
photoinhibition	0.7841	1	7
mRNA processing	0.783	1	5
photosynthesis, light harvesting in photosystem II	0.7829	1	7
vitamin E biosynthetic process	0.7826	1	7
cold acclimation	0.773	1	18
regulation of stomatal movement	0.7711	0.95	20
response to hydrogen peroxide	0.7682	1	28
trichome morphogenesis	0.7629	0.8462	13

phototropism	0.7616	0.9231	13
response to UV-B	0.7603	0.9167	24
response to virus	0.7603	1	11
cell proliferation	0.7577	1	12
chlorophyll biosynthetic process	0.7553	1	22
actin nucleation	0.7494	0.875	8
carpel development	0.7487	0.6	10
DNA methylation	0.7468	1	8
cell morphogenesis	0.7453	1	6
negative regulation of flower development	0.7419	1	25
trichome differentiation	0.7415	1	8
oxygen and reactive oxygen species metabolic process	0.7412	1	8
response to water deprivation	0.7406	0.925	80
lignin biosynthetic process	0.7387	1	9
phenylpropanoid biosynthetic process	0.7384	1	8
response to cold	0.7359	0.944	125
response to nematode	0.7344	0.98	50
abscisic acid mediated signaling	0.7315	0.9722	36
isoprenoid biosynthetic process	0.7298	1	10
photorespiration	0.7244	0.8519	27
systemic acquired resistance, salicylic acid mediated signaling pathway	0.7214	1	11
response to desiccation	0.7211	1	13
histone methylation	0.7202	1	9
defense response to fungus	0.7199	0.9062	32
jasmonic acid biosynthetic process	0.7177	1	20
stomatal complex morphogenesis	0.7141	1	9
protein targeting to mitochondrion	0.7109	0.8571	7
fatty acid biosynthetic process	0.7089	0.913	23
lipid transport	0.7078	1	9
systemic acquired resistance	0.6988	0.8	10
RNA interference, production of ta-siRNAs	0.6986	0.8	5
chlorophyll catabolic process	0.6986	0.8	5
RNA interference, production of siRNA	0.6965	0.6	5
cytoskeleton organization and biogenesis	0.6939	0.8	10
carotenoid biosynthetic process	0.6935	0.75	12
embryonic development ending in seed dormancy	0.6931	0.8529	68
defense response to bacterium	0.6905	1	23
response to light stimulus	0.6904	0.9474	38
abscisic acid biosynthetic process	0.687	1	8
response to bacterium	0.6864	0.9615	26
response to osmotic stress	0.6859	0.9574	47
jasmonic acid and ethylene-dependent systemic resistance	0.6857	1	5
hypersensitive response	0.6848	0.9	20
response to stress	0.6834	0.9643	28
positive gravitropism	0.6785	1	8
Chloroplast organization and biogenesis	0.6717	0.95	20
DNA endoreduplication	0.6709	0.8571	14
PSII associated light-harvesting complex II catabolic process	0.6656	1	6

brassinosteroid homeostasis	0.6655	1	6
regulation of meristem organization	0.6616	1	12
aromatic amino acid family biosynthetic process, shikimate pathway	0.6615	1	12
sugar mediated signaling	0.6606	0.8889	18
seed germination	0.6582	0.7273	11
cell differentiation	0.6564	1	12
protein ubiquitination	0.6563	0.9091	11
root epidermal cell differentiation	0.6554	1	6
thylakoid membrane organization and biogenesis	0.6426	1	7
cell fate specification	0.6425	0.7143	7
response to salt stress	0.6424	0.9167	132
regulation of transcription, DNA-dependent	0.6415	0.7143	49
stamen development	0.6407	0.7143	7
protein amino acid phosphorylation	0.6403	1	28
negative regulation of ethylene mediated signaling pathway	0.6351	1	11
defense response to fungus, incompatible interaction	0.6334	1	10
response to wounding	0.6308	0.9688	64
Cytokinesis by cell plate formation	0.6299	1	7
positive regulation of flower development	0.6263	0.8947	19
embryo sac development	0.6262	0.8571	7
flavonoid biosynthetic process	0.623	0.875	8
root development	0.6188	0.8621	29
Indoleacetic acid biosynthetic process	0.6167	1	8
cuticle development	0.6134	1	13
pollen development	0.6118	1	10
heat acclimation	0.6099	0.6154	13
response to chitin	0.6096	1	8
protein targeting to vacuole	0.6094	1	9
circadian rhythm	0.6072	1	17
defense response	0.6061	0.9016	61
brassinosteroid biosynthetic process	0.6058	1	9
purine transport	0.6044	0.9474	19
wax biosynthetic process	0.6024	1	9
Multicellular organismal development	0.6022	0.8333	12
protein amino acid autophosphorylation	0.6006	1	9
multidimensional cell growth	0.5998	1	15
red, far-red light phototransduction	0.5995	0.8	10
unidimensional cell growth	0.5988	0.8974	39
response to abscisic acid stimulus	0.5978	0.8774	155
cellulose and pectin-containing cell wall biogenesis	0.596	0.9412	17
response to auxin stimulus	0.5952	0.8421	133
auxin biosynthetic process	0.5901	1	11
jasmonic acid mediated signaling pathway	0.5894	0.8182	11
leaf senescence	0.5886	0.9091	11
cellular response to phosphate starvation	0.5874	1	11
negative regulation of transcription	0.5854	0.9375	16
auxin polar transport	0.5847	0.9333	30
trichome branching	0.5826	0.9167	12

response to fungus	0.581	0.8421	19
regulation of transcription	0.5808	0.6264	522
flower development	0.5807	0.7647	34
ethylene biosynthetic process	0.5805	1	12
response to sucrose stimulus	0.5748	1	21
stomatal movement	0.5746	1	10
auxin mediated signaling pathway	0.5715	0.8947	19
meristem organization	0.5701	1	11
response to cytokinin stimulus	0.5655	0.8421	38
defense response to bacterium, incompatible interaction	0.5624	1	15
very-long-chain fatty acid metabolic process	0.5568	1	15
hyperosmotic salinity response	0.5547	1	26
response to UV	0.5545	0.95	20
leaf development	0.5537	0.7447	47
response to other organism	0.5477	0.7826	23
cellulose biosynthetic process	0.5473	0.8	15
response to salicylic acid stimulus	0.5459	0.8298	94
response to ethylene stimulus	0.5451	0.9012	81
response to cadmium ion	0.5399	0.8667	45
ethylene mediated signaling pathway	0.5362	0.5833	36
response to jasmonic acid stimulus	0.5289	0.9	100
response to gibberellin stimulus	0.5212	0.8772	57
D-xylose metabolic process	0.5	1	5
L-ascorbic acid biosynthetic process	0.5	0.8333	6
RNA processing	0.5	1	7
abaxial cell fate specification	0.5	0.7143	7
Aging	0.5	1	20
anatomical structure morphogenesis	0.5	1	5
anther development	0.5	0.5	6
anthocyanin biosynthetic process	0.5	1	5
auxin homeostasis	0.5	0.5	10
auxin metabolic process	0.5	1	7
blue light signaling pathway	0.5	0.8333	6
carotene biosynthetic process	0.5	0.8	5
cell death	0.5	1	13
cell division	0.5	1	8
cell growth	0.5	0.7	10
cell tip growth	0.5	0.8571	7
cellulose and pectin-containing secondary cell wall biogenesis	0.5	0.4545	11
chromatin assembly or disassembly	0.5	1	5
coenzyme A biosynthetic process	0.5	1	6
cotyledon development	0.5	0.1667	6
cytokinin biosynthetic process	0.5	1	10
cytokinin catabolic process	0.5	1	5
defense response signaling pathway, resistance gene-dependent	0.5	1	6
defense response to virus	0.5	0.8333	6
defense response, incompatible interaction	0.5	0.75	8
dolichol biosynthetic process	0.5	1	6

embryonic development	0.5	0.6364	11
Endosperm development	0.5	0.9	10
fatty acid elongation	0.5	1	8
fatty acid metabolic process	0.5	1	10
floral organ abscission	0.5	0.8	5
floral organ development	0.5	0.8571	7
fruit development	0.5	1	7
Galactolipid biosynthetic process	0.5	1	6
gibberellic acid mediated signaling	0.5	0.9091	22
gibberellin biosynthetic process	0.5	1	13
gibberellin catabolic process	0.5	1	5
glucose metabolic process	0.5	1	5
glucosinolate catabolic process	0.5	1	5
Gravitropism	0.5	1	13
Growth	0.5	1	5
lateral root development	0.5	0.9	10
lateral root morphogenesis	0.5	0.75	8
leaf morphogenesis	0.5	0.8	25
meiotic recombination	0.5	0.8889	9
meristem initiation	0.5	0.875	8
methionine biosynthetic process	0.5	1	5
microsporogenesis	0.5	0.5714	7
negative gravitropism	0.5	1	7
negative regulation of cyclin-dependent protein kinase activity	0.5	0.5714	7
nitrate transport	0.5	0.8	5
Oligopeptide transport	0.5	1	5
organ morphogenesis	0.5	0.6	10
pattern specification process	0.5	0.8	10
Pentacyclic triterpenoid biosynthetic process	0.5	1	8
pentose-phosphate shunt, oxidative branch	0.5	1	6
petal development	0.5	0.6667	9
phenylpropanoid metabolic process	0.5	1	6
photoperiodism, flowering	0.5	1	5
plastid organization and biogenesis	0.5	0.7143	7
polarity specification of adaxial/abaxial axis	0.5	1	7
pollen germination	0.5	1	7
pollen maturation	0.5	1	5
positive regulation of cell proliferation	0.5	0.8333	6
positive regulation of transcription	0.5	0.8235	17
primary shoot apical meristem specification	0.5	0.8333	12
protein import into chloroplast stroma	0.5	1	5
protein import into chloroplast thylakoid membrane	0.5	1	5
protein targeting to chloroplast	0.5	1	9
Proteolysis	0.5	1	5
radial pattern formation	0.5	0.8333	6
red or far red light signaling pathway	0.5	1	8
regulation of cell proliferation	0.5	1	5
regulation of circadian rhythm	0.5	0.875	8

regulation of flower development	0.5	1	17
regulation of gene expression, epigenetic	0.5	0.875	8
regulation of meristem size	0.5	1	6
regulation of timing of transition from vegetative to reproductive phase	0.5	0.6667	6
response to abiotic stimulus	0.5	1	6
response to blue light	0.5	1	7
response to brassinosteroid stimulus	0.5	0.6364	11
response to glucose stimulus	0.5	0.6	5
response to hormone stimulus	0.5	0.8333	6
response to insect	0.5	0.875	8
response to mechanical stimulus	0.5	1	7
response to molecule of bacterial origin	0.5	1	6
response to ozone	0.5	1	10
response to reactive oxygen species	0.5	1	5
response to red light	0.5	1	5
response to red or far red light	0.5	0.9412	17
response to starvation	0.5	0.8	5
response to temperature stimulus	0.5	0.875	8
rhamnogalacturonan II biosynthetic process	0.5	1	5
root hair cell differentiation	0.5	0.7778	9
seed development	0.5	0.8	10
sexual reproduction	0.5	0.8333	6
shoot development	0.5	1	8
specification of floral organ identity	0.5	0.7778	9
stem cell maintenance	0.5	0.6	5
syncytium formation	0.5	1	10
transcription factor import into nucleus	0.5	0	6
transcription initiation	0.5	0.875	8
vascular tissue development (sensu Tracheophyta)	0.5	0.9	10
vascular tissue pattern formation (sensu Tracheophyta)	0.5	0.6923	13
vegetative phase change	0.5	1	6
vegetative to reproductive phase transition	0.5	1	5
xanthophyll biosynthetic process	0.5	0.8	5
xylem histogenesis	0.5	0.5	6

Supplementary Table 14. AraNet predictive power measured by the area under cross-validated ROC curves (AUC) for Gene Ontology cellular component terms.

Gene Ontology cellular component term	AUC	network coverage	# genes
proteasome_regulatory_particle,_base _subcomplex_(sensu_Eukaryota)	0.9992	1	11
mitochondrial_intermembrane_space	0.9984	1	6
cytosolic_ribosome_(sensu_Eukaryota)	0.9971	1	86
cytochrome_b6f_complex	0.996	1	5
mitochondrial_small_ribosomal_subunit	0.9952	1	6
plastid_small_ribosomal_subunit	0.9608	1	16
Arp2/3_protein_complex	0.95	1	10
plastid_large_ribosomal_subunit	0.9216	1	14
chloroplast_stromal_thylakoid	0.9108	1	6
nuclear_envelope	0.906	1	14
SCAR_complex	0.8996	0.8	5
spindle	0.8975	1	20
photosystem_II_reaction_center	0.8969	1	5
nucleoplasm	0.8965	1	5
large_ribosomal_subunit	0.8916	1	5
endosome	0.8864	1	9
mitochondrial_inner_membrane	0.8695	1	16
protein_complex	0.8553	1	7
ubiquitin_ligase_complex	0.8523	1	14
nuclear_speck	0.8455	1	17
retromer_complex	0.8314	1	6
multivesicular_body	0.8314	1	6
mitochondrial_envelope	0.8228	0.889	9
trans-Golgi_network_transport_vesicle	0.8104	1	8
chloroplastic_endopeptidase_Clp_complex	0.8073	1	8
chloroplast_thylakoid_membrane	0.8035	0.947	281
integral_to_membrane	0.8015	1	13
chloroplast_photosystem_I	0.798	1	5
endoplasmic_reticulum_membrane	0.7935	0.929	14
cytoskeleton	0.7914	1	5
phragmoplast	0.7888	0.943	35
mitochondrial_matrix	0.7887	1	13
mitochondrial_outer_membrane	0.7808	1	7
chloroplast	0.7501	0.971	240
mitochondrial_inner_membrane_presequence _translocase_complex	0.7494	0.9	10
late_endosome	0.7481	1	8
peroxisome	0.7474	1	29
Cajal_body	0.7465	1	6
nucleolus	0.7407	0.918	49
microtubule	0.7407	0.9	10
cortical_microtubule,_transverse_to_long_axis	0.7391	0.75	8
cell_plate	0.7348	1	8
microsome	0.7269	1	14

signalosome_complex	0.7199	1	25
thylakoid	0.7155	0.909	11
cytoplasm	0.7131	0.946	260
plastoglobule	0.7105	0.933	105
plastid	0.7096	1	23
membrane_of_vacuole_with_cell_cycle-independent_morphology	0.7093	1	30
endoplasmic_reticulum	0.705	0.961	51
extracellular_matrix	0.6977	0.8	5
plasma_membrane	0.696	0.979	145
ribosome	0.6909	1	5
chloroplast_thylakoid_lumen	0.6873	0.977	85
trans-Golgi_network	0.6848	1	16
mitochondrial_membrane	0.6833	0.81	21
proteasome_core_complex_(sensu_Eukaryota)	0.6809	1	58
chloroplast_thylakoid	0.6637	1	9
nucleus	0.6591	0.901	567
Golgi_apparatus	0.6478	1	24
chloroplast_stroma	0.6459	1	70
mitochondrion	0.6439	0.942	771
vacuole,_cell_cycle_independent_morphology	0.6422	1	7
cell_wall	0.6407	0.929	14
cytosol	0.6384	0.889	126
SCF_ubiquitin_ligase_complex	0.607	1	9
membrane	0.6065	0.971	68
vacuolar_membrane	0.606	1	18
cellulose_and_pectin-containing_cell_wall	0.5921	0.946	110
chloroplast_inner_membrane	0.5912	0.862	29
plastid_chromosome	0.5724	0.944	18
chloroplast_envelope	0.5589	0.849	33
intracellular	0.553	0.867	15
anchored_to_membrane	0.5333	0.77	473
vacuole	0.5	1	9
peroxisomal_membrane	0.5	0.8	5
mitochondrial_outer_membrane_translocase_complex	0.5	0.8	5
Golgi_transport_complex	0.5	1	9
Golgi_stack	0.5	1	5
extracellular_region	0.5	0.607	28
endoplasmic_reticulum_lumen	0.5	0.833	6
chromatin	0.5	1	5
chloroplast_photosystem_II	0.5	1	5
chloroplast_outer_membrane	0.5	0.95	20
cell_surface	0.5	1	5
apical_plasma_membrane	0.5	1	7

Supplementary Table 15. AraNet predictive power measured by the area under cross-validated ROC curves (AUC) for isozyme-free KEGG pathway terms.

Isozyme-free KEGG metabolic pathway terms	AUC	network coverage	# genes
Proteasome	0.9997	1	44
Citrate cycle (TCA cycle)	0.9984	1	6
Fatty acid biosynthesis	0.9981	1	8
Regulation of autophagy	0.998	1	9
Phenylalanine, tyrosine and tryptophan biosynthesis	0.9932	1	11
Ribosome	0.9888	1	204
SNARE interactions in vesicular transport	0.9775	0.9792	48
Basal transcription factors	0.975	1	22
Photosynthesis - antenna proteins	0.9682	1	18
Glutathione metabolism	0.9621	1	14
Aminoacyl-tRNA biosynthesis	0.9532	1	12
Pyrimidine metabolism	0.9438	0.963	27
Valine, leucine and isoleucine biosynthesis	0.9429	1	10
Oxidative phosphorylation	0.9297	0.9688	64
Butanoate metabolism	0.9287	1	8
RNA polymerase	0.9258	1	7
Pantothenate and CoA biosynthesis	0.9149	1	6
DNA polymerase	0.9148	0.9167	12
Purine metabolism	0.9073	0.9667	30
N-Glycan biosynthesis	0.9067	1	18
Alkaloid biosynthesis II	0.9051	1	16
Glutamate metabolism	0.9049	1	17
Folate biosynthesis	0.9004	1	6
Sulfur metabolism	0.8966	1	5
Ubiquinone biosynthesis	0.8964	1	5
Benzoate degradation via CoA ligation	0.8831	1	62
Urea cycle and metabolism of amino groups	0.8719	1	12
Pyruvate metabolism	0.8715	1	10
Phosphatidylinositol signaling system	0.8714	1	89
Glycan structures - biosynthesis 1	0.8711	0.9412	17
Nicotinate and nicotinamide metabolism	0.8651	1	65
Inositol phosphate metabolism	0.8648	0.9875	80
Valine, leucine and isoleucine degradation	0.8558	1	11
Protein export	0.8502	1	26
Cysteine metabolism	0.8494	1	7
Metabolism of xenobiotics by cytochrome P450	0.8291	1	18
Biosynthesis of steroids	0.8247	0.9524	21
Alkaloid biosynthesis I	0.7984	1	5
Selenoamino acid metabolism	0.7783	0.9	10
Photosynthesis	0.7765	1	21
Arginine and proline metabolism	0.7607	1	13
Alanine and aspartate metabolism	0.7585	1	13
Starch and sucrose metabolism	0.7533	0.973	37
Histidine metabolism	0.7382	1	10

beta-Alanine metabolism	0.7197	1	9
Carotenoid biosynthesis - General	0.716	0.8889	9
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	0.7141	0.8571	7
Propanoate metabolism	0.703	1	9
Flavonoid biosynthesis	0.6978	0.8	5
Glycerophospholipid metabolism	0.6942	1	12
Lysine degradation	0.6918	1	5
Nitrogen metabolism	0.6657	0.8333	6
Aminosugars metabolism	0.6633	1	12
Glycerolipid metabolism	0.6604	0.8696	23
Carbon fixation	0.6601	1	9
Tyrosine metabolism	0.6461	1	20
Tryptophan metabolism	0.6254	0.9286	14
ABC transporters - General	0.6232	1	8
Bile acid biosynthesis	0.6192	1	8
Fructose and mannose metabolism	0.6192	0.8333	18
Glycolysis / Gluconeogenesis	0.6039	1	14
Glycan structures - biosynthesis 2	0.5999	0.7333	15
Limonene and pinene degradation	0.5783	1	66
gamma-Hexachlorocyclohexane degradation	0.571	1	64
Fatty acid metabolism	0.5651	1	14
Ascorbate and aldarate metabolism	0.5643	0.9286	14
Fluorene degradation	0.5545	1	62
Naphthalene and anthracene degradation	0.5507	1	67
Phenylalanine metabolism	0.5409	1	71
Phenylpropanoid biosynthesis	0.5255	0.9862	145
Methane metabolism	0.5134	1	69
1- and 2-Methylnaphthalene degradation	0.5	1	7
Androgen and estrogen metabolism	0.5	1	5
Cyanoamino acid metabolism	0.5	0.9091	11
Galactose metabolism	0.5	0.8571	7
Glycan structures - degradation	0.5	1	5
Glycine, serine and threonine metabolism	0.5	1	9
Indole and ipecac alkaloid biosynthesis	0.5	1	6
Methionine metabolism	0.5	1	5
Pentose and glucuronate interconversions	0.5	0.2	5
Porphyrin and chlorophyll metabolism	0.5	1	7
Terpenoid biosynthesis	0.5	1	5

Supplementary Table 16. Oligonucleotide sequences used as PCR primers in this study.

Mutant number	gene name	Stock name	Oligonucleotide sequences	Tm
#35 (unknown)	At1g80710	Salk_001238C	Fwd : GCTTACCTGATGGCTTTTCA Rev : AACTGGTGCTGAGTGAGGAG	58
#35-1 (unknown)	At1g80710	Salk_149366C	Fwd: TGCAAATCCCAAAACAGAGAG Rev: CGTTCTCATCCTTAACCACTCC	60
#36 (unknown)	At2g17900	Salk_127952C	Fwd : AACTGCAGTCCAATCAAAGGAT Rev : TGAGAACCCGTGAAAAACTTC	60
#39 (unknown)	At3g05090	Salk_059570C	Fwd : CAAGAACTTGGGGTTTTGG Rev : AGGGAGAGTGTTTTGCTGTG	58
#47 (unknown)	At1g15770	Salk_118634C	Fwd : TGCTCTATGTTTGTCTTCATGC Rev : AAATGAAAATGGAGATGATTGG	58
#51 (unknown)	At2g34170	Salk_099804C	Fwd : AAGAAAGCGAGGAGGATTCA Rev : ACACTGCGATACGGTGACAT	59
#197 (seed pigmentation)	At5g50110	Salk_027109C	Fwd: ACACAGCCCCATTACATTAGC Rev: TTTGTAAGCCCGGTAACATTC	59
#225 (seed pigmentation)	At5g50110	Salk_086197C	Fwd: ACACAGCCCCATTACATTAGC Rev: TTTGTAAGCCCGGTAACATTC	59
#67 (seed pigmentation)	At4g26430	Salk_036965C	Fwd: GGTGATGCTTAACATATCCGATCA Rev: ATTCTTTCTTCCCCGTAGGAAAAC	62
#14 (seed pigmentation)	At4g26430	Salk_049514C	Fwd: TGGTTCCAAACTCAAAACTAATTG Rev: CCAATTTTTACCGCTTTCGT	60
#31 (seed pigmentation)	At5g45620	Salk_147710C	Fwd: TCTCGGTTGTCTCTCTCACCA Rev: CCTTGATCTGTGGAATCCCTA	60
#120 (seed pigmentation)	At5g45620	Salk_018378C	Fwd: TCTCGGTTGTCTCTCTCACCA Rev: CCTTGATCTGTGGAATCCCTA	60
#2 (seed pigmentation)	At3g20160	Salk_038548C	Fwd: CTCATGAAGATGCTTGTGAAGAC Rev: AGTGCTTTATTGACGGACTTAGC	59
#258 (seed pigmentation)	At5g04110	Salk_108979C	Fwd: TGCTCTTTGATTTCCATGGTT Rev: ACAACAACATCCAGATGAAGCAAT	61
#26 (seed pigmentation)	At5g11480	Salk_029559C	Fwd: GAGCGGGCAAATTGTAATATAAGG Rev: AACAATACCCCATCTTCAAAAGT	60
#134 (seed pigmentation)	At5g13630	Salk_152096C	Fwd: AAACTTTTCGTGGGGCTTTT Rev:TTGGTACTGTTAGTGAGCGAAGAG	60
#21 (seed pigmentation)	At5g15610	Salk_151350C	Fwd: TCAAATGCGTAAGATTTTTGC Rev: AGCTTTCTGACGCGAATCAA	60

Supplementary Table 17. Gene-specific primers used for RT-PCR experiments.

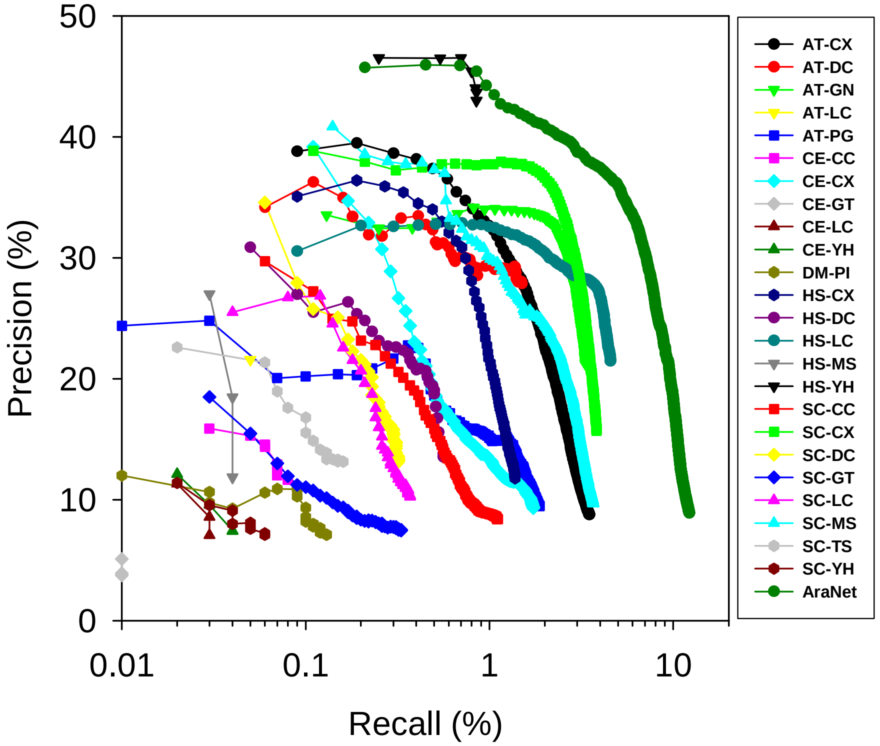
Gene	Primers	Tm	Product length
At1g80710	Fwd: TTAATCATTCTAGGGCTGTGC Rev: CCATCACTGTTCGCTTTAGTT	57	899 258
At3g05090	Fwd: AGAAGGACTTCCCATTGTGG Rev: TCCTCCTGAAACACTTGCTG	59	469 248
Actin	Fwd: TGGTCGTACAACCGGTATTGTGCTGG Rev: TGTCTCTTACAATTTCCCGCTCTGCTG	60	~220

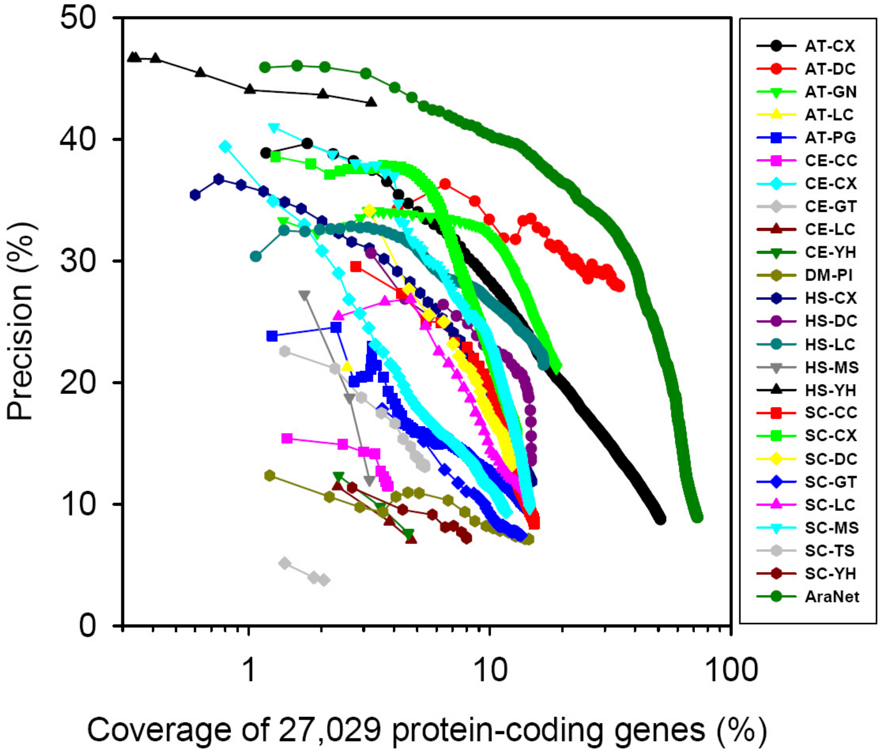
FIGURES

Supplementary Figure 1.

(Top) An alternate representation of the data in Figure 1A, reporting precision of GO-BP functional linkage reconstruction versus recall with 0.632 bootstrapping. Legend abbreviations are as in Figure 1A. (Bottom) Same as in top, but plotting linkage precision versus recall of genes.

Supplementary Figure 1.

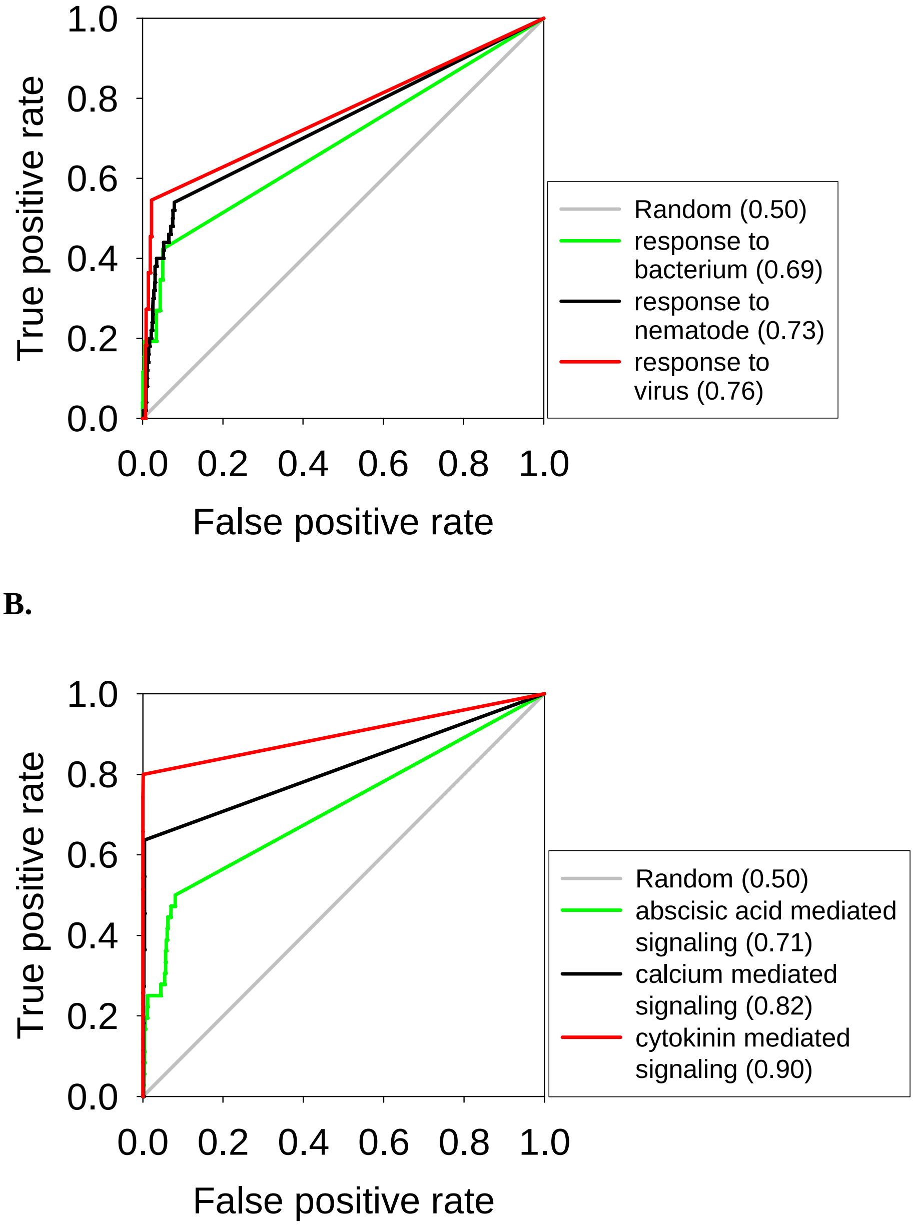




Supplementary Figure 2.

Cross-validated ROC curve analysis for AraNet (excluding literature based protein-protein interactions)-based prediction of selected sets of GO biological process terms for (**A**) biotic response and (**B**) hormonal signaling. AUC values are reported in parentheses.

Supplementary Figure 2.

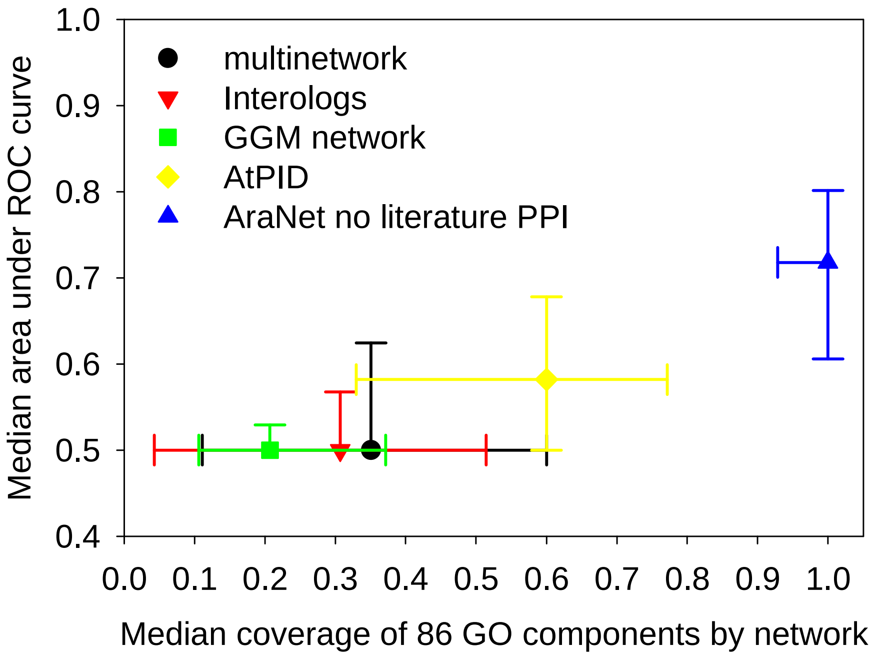


A.

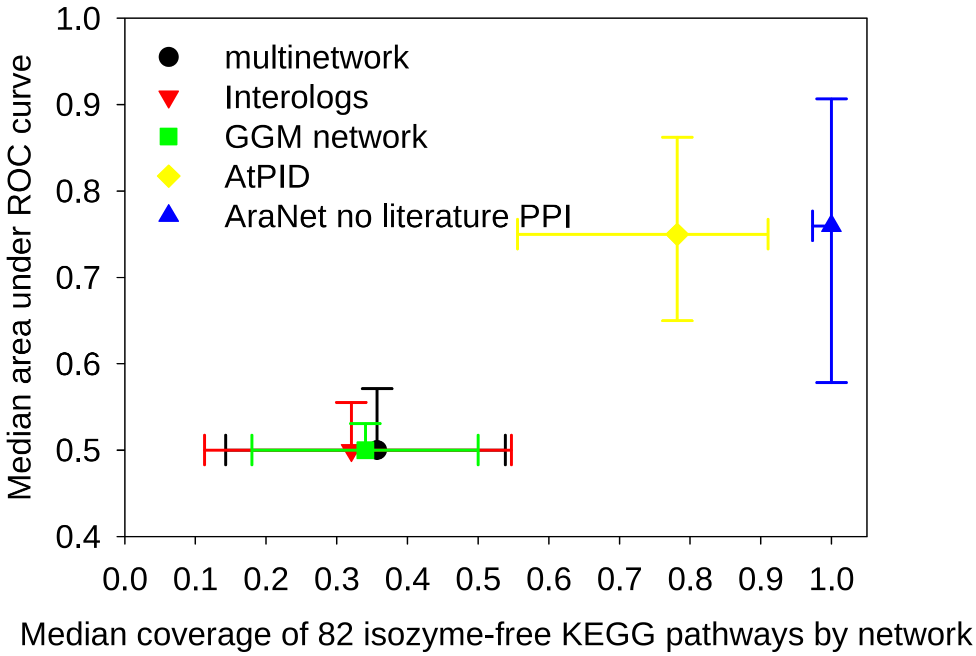
B.

Supplementary Figure 3.
Comparison of the predictive power of AraNet (excluding literature derived linkages) with previous network models (described in **Supplementary Table 3**) for (**A**) GO cellular components (86 sets with >= 5 member genes), (**B**) isozyme-free KEGG pathways (82 sets with >= 5 member genes). Each symbol indicates median predictive performance across pathways. Error bars indicate 1st, 3rd quartiles.

Supplementary Figure 3.
A.

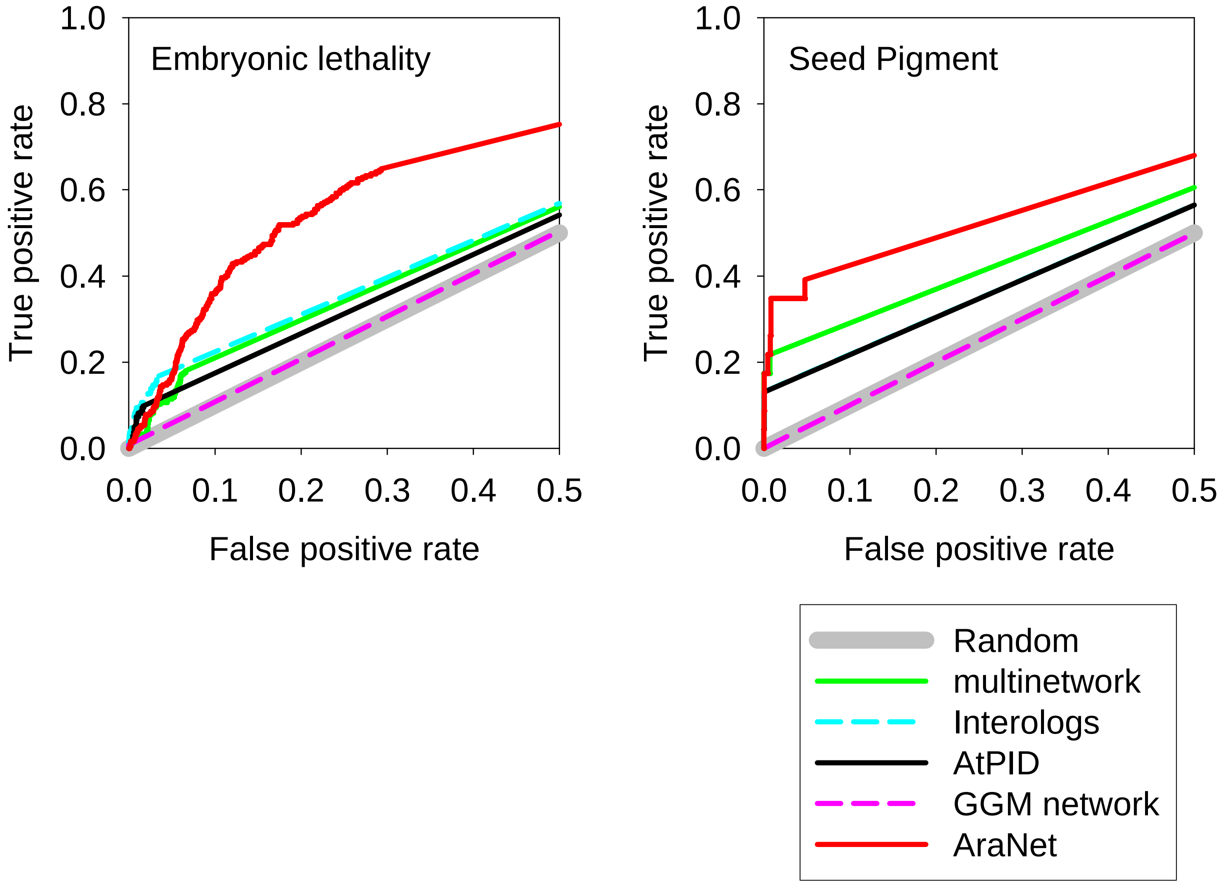


B.



Supplementary Figure 4.
Cross-validated ROC curve analysis for AraNet-based predictions of genes associated with 2 independent test sets of mutant phenotypes—embryonic lethality and seed pigmentation[25], and comparison with predictions using previous *A. thaliana* networks.

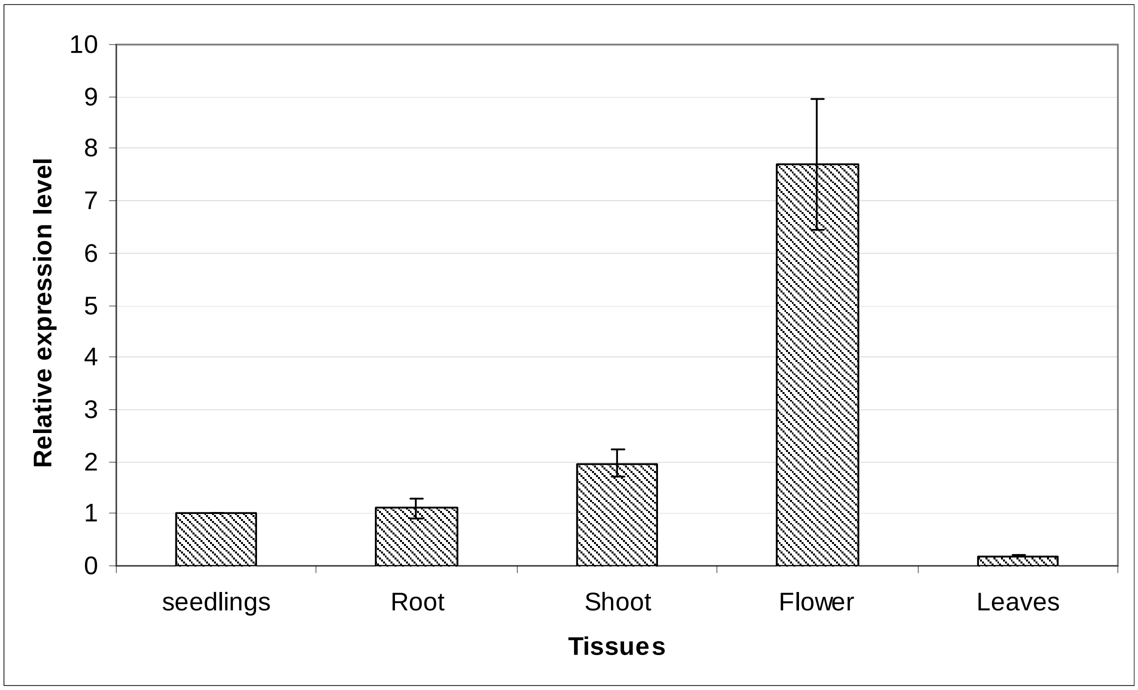
Supplementary Figure 4.



Supplementary Figure 5.

Real-time RT-PCR of *Drs1*. Relative expression quantification was performed using the ΔΔCT method[26] with actin as the reference gene, which was expressed at a constant level in all conditions. Expression in each tissue was normalized against that in seedlings. Histograms and error bars indicate mean relative expression and standard error (n = 12).

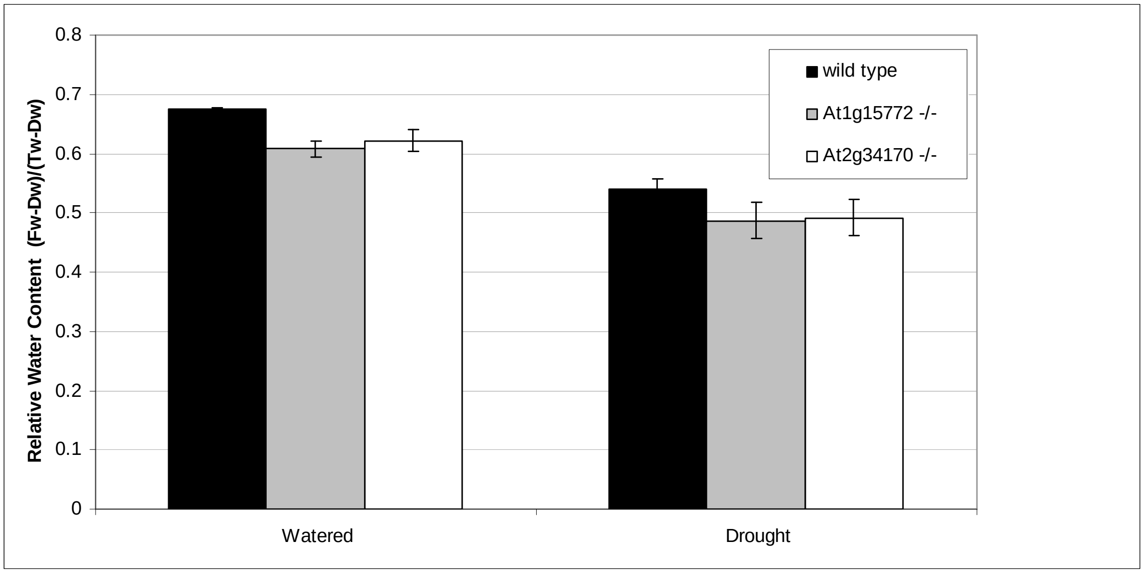
Supplementary Figure 5.



Supplementary Figure 6.

Relative water content between wild type and two randomly chosen genes, At1g15772 and At2g34170, is indistinguishable in watered and drought conditions. Four-week old wild type and mutant plants were treated for drought (no watering) for 7 days. Relative water loss was calculated as (Fw-Dw)/(Tw-Dw) (Fw, fresh weight; Dw, dry weight; Tw, turgor weight). Three plants from each genotype for each treatment condition were measured. There was no significant difference between the relative water loss neither in wild type and mutant plants ($p \geq 0.5$, unpaired t-test) nor between watered and drought conditions of the same genotype ($p \geq 0.1$, unpaired t-test). For each condition, three plants from each genotype were assayed in each experiment and two independent experiments were conducted. Error bars indicate standard error.

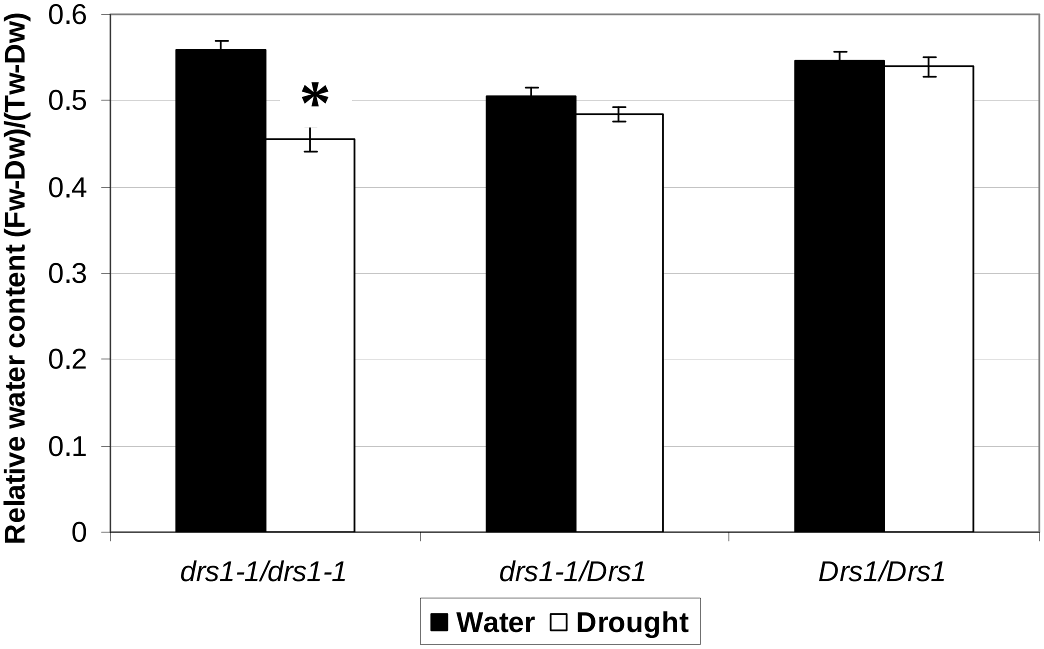
Supplementary Figure 6.



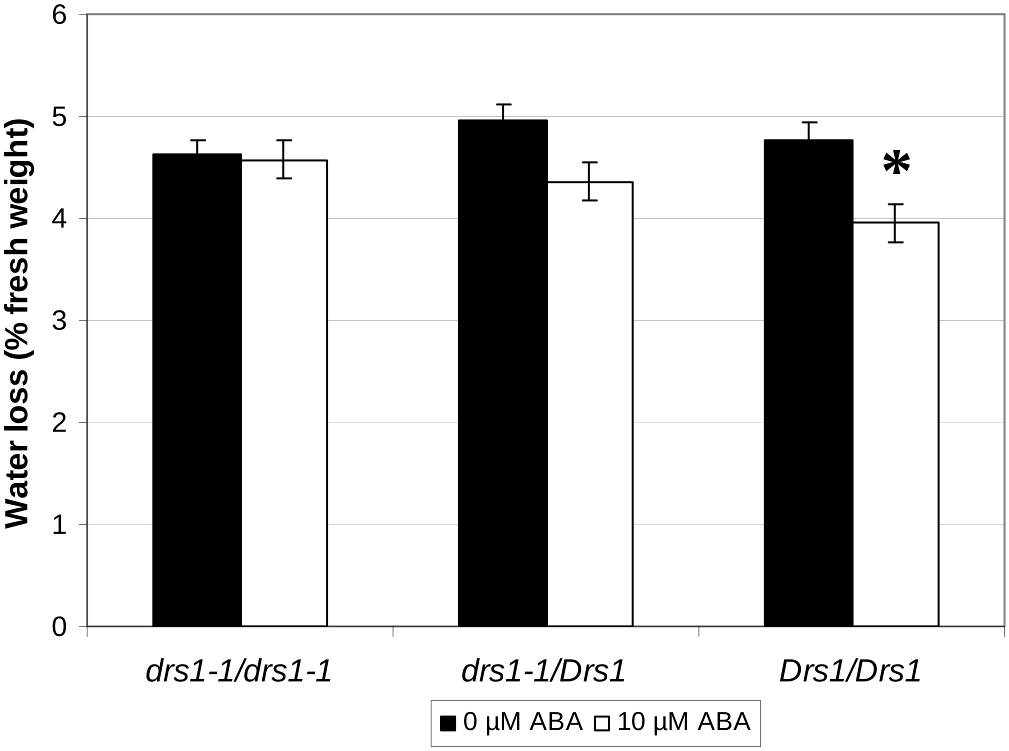
Supplementary Figure 7.

F2 linkage test of the *drs1-1* mutant shows that the water-retention and abscisic acid (ABA) response phenotypes are linked to the mutant allele. (**A**) The relative water content of the F2 segregating population shows a significant reduction in relative water content in drought-treated plants that are homozygous for the mutant allele (p = 0.007, unpaired t-test, n = 29). Four-week-old F2 plants were genotyped using PCR (Methods). Half of the plants in each genotype were treated for drought (no watering) for 7 days, and half were watered. Relative water content was calculated as (Fw-Dw)/(Tw-Dw) (Fw, fresh weight; Dw, dry weight; Tw, turgor weight). There was no significant difference between the relative water content of drought-treated and watered plants for either *Drs1/Drs1* (p = 0.848, unpaired t-test, n = 27) or for *drs1-1/Drs1* (p = 0.410, unpaired t-test, n = 69). (**B**) The excised leaf transpiration assay of the F2 segregating population shows a significant reduction in transpiration in the presence of 10 µM ABA only in the plants that are *Drs1/Drs1* (p = 0.067, unpaired t-test, n = 24). Four-week-old F2 plants were genotyped using PCR (Methods). Mature leaves from 4-week old plants were detached and immersed in sap solution containing either no ABA or 10 µM ABA for 22 hours. Heterozygotes and homozygotes for the mutant allele were insensitive to ABA (p = 0.23, unpaired t-test, n = 21 for *drs1-1/Drs1* and p = 0.92, unpaired t-test, n = 16 for *drs1-1/drs1-1*). Asterisk indicates significant difference between conditions of the same genotype. Error bars indicate standard error.

Supplementary Figure 7A.



Supplementary Figure 7B.

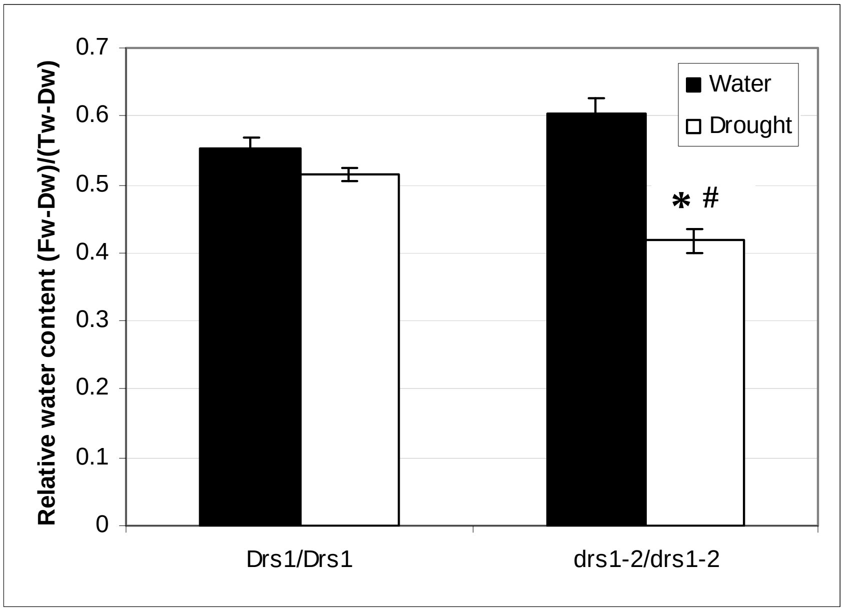


Supplementary Figure 8.

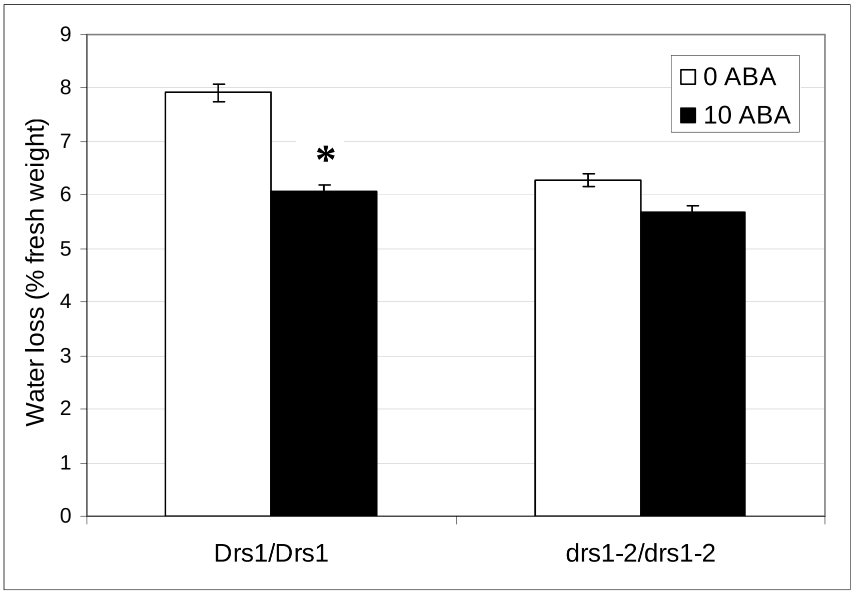
Plants carrying an independent knock-out allele of Drs1, hereby named *drs1-2* (SALK_149366C) showed similar phenotypes as those carrying *drs1-1* allele. (**A**) Plants carrying *drs1-2* retained significantly less water than wild type under drought. Relative water loss was calculated as (Fw-Dw)/(Tw-Dw) (Fw, fresh weight; Dw, dry weight; Tw, turgor weight). Significant differences between the relative water loss of wild type and mutant plants are indicated by * ($p \leq 0.01$, unpaired t-test, n = 27), significant differences between watered and drought conditions of the same genotype by # ($p \leq 0.005$, unpaired t-test, n = 21). Results are from one experiment. (**B**) Transpiration was reduced in wild type plants in the presence of 10 µM abscisic acid (ABA) whereas mutant plants were insensitive to ABA. Significant differences between treatments in each genotype are indicated by * ($p = 1.75 \times 10^{-05}$, unpaired t-test, n = 77). Results are from three independent experiments.

Supplementary Figure 8.

A.



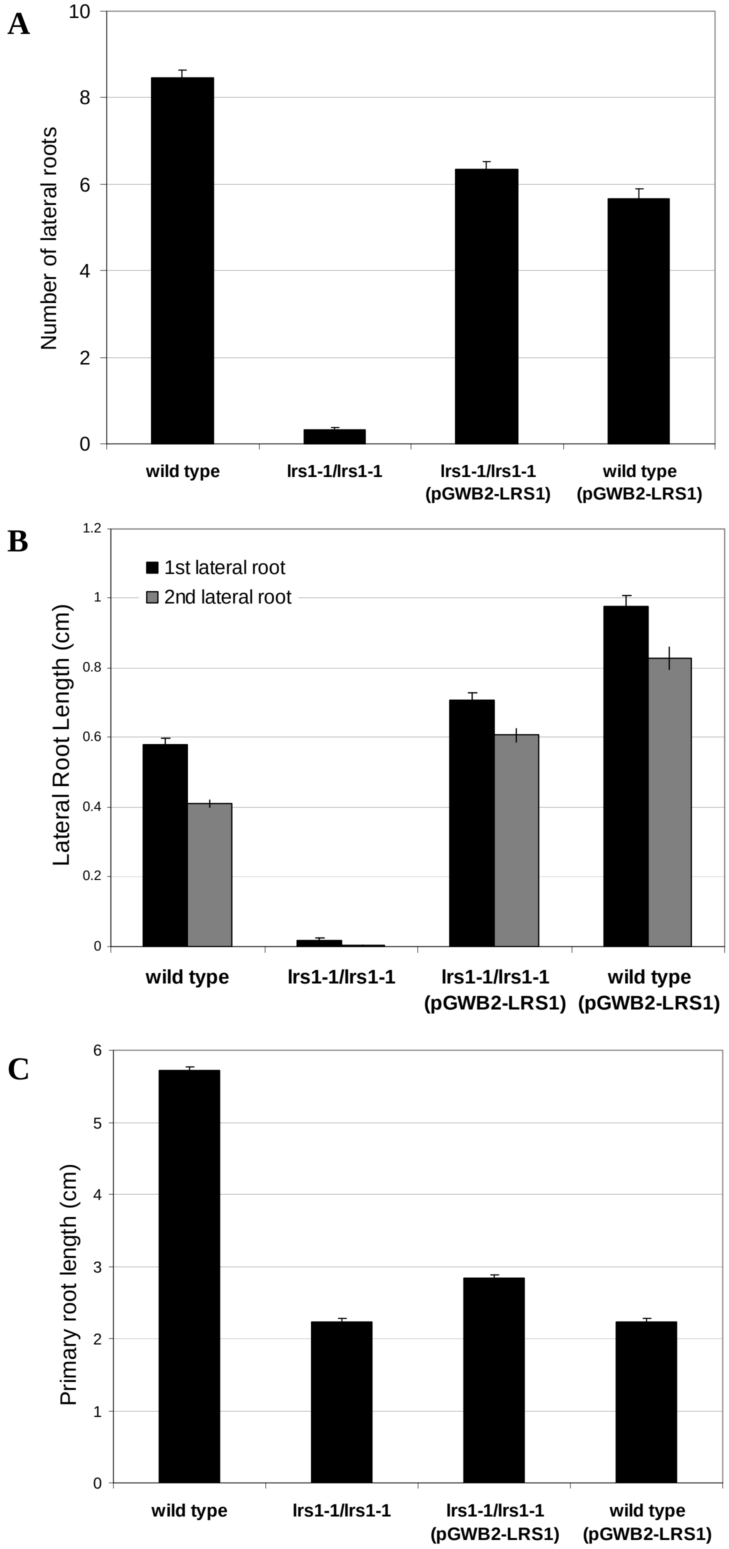
B.



Supplementary Figure 9.

The number of lateral roots (LR) is strongly reduced in *lrs1-1* mutants. This phenotype can be complemented by reintroduction of the functional gene. 11-day old seedlings grown on MS media. (**A**) The number of LR is significantly reduced in the mutant (p = 6 x 10^{-37}, unpaired t-test, n = 137). When the wild type allele is introduced to lines that are *lrs1-1/lrs1-1*, the number of LR is significantly increased compared to the mutant (p = 6 x 10^{-30}, unpaired t-test, n = 121). (**B**) When additional copies of the gene are expressed in a wild type strain, lateral roots increase in length. Only the first and second oldest lateral roots were measured. (**C**) The primary root is shorter in *lrs1-1* than wild type but this phenotype is not complemented, showing that the primary root phenotype is separable and independent from the lateral root phenotype. Fifty to 77 plants from each genotype were tested.

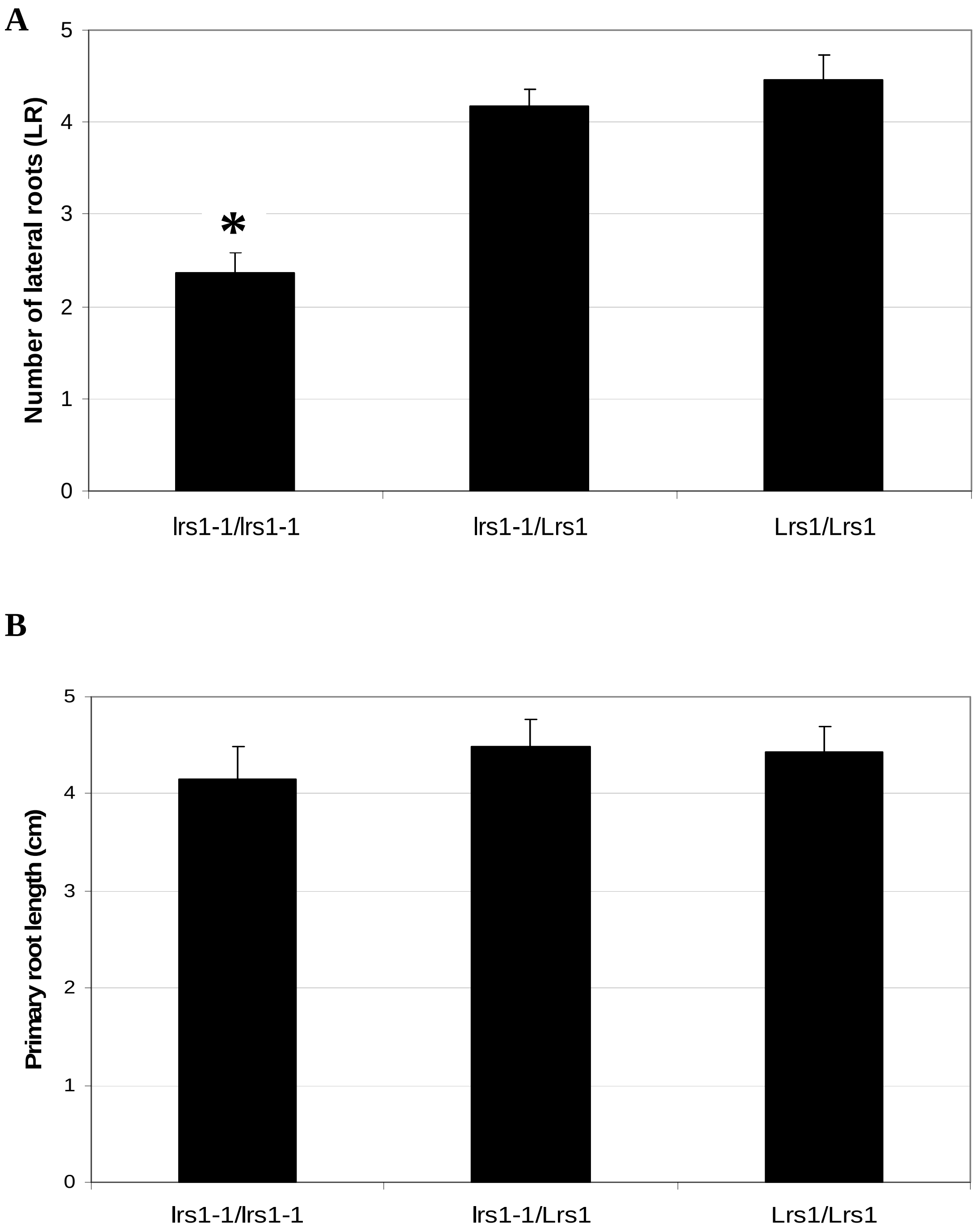
Supplementary Figure 9.



Supplementary Figure 10.

F2 linkage test of the lateral root (LR) number and primary root length of *lrs1-1* x Col-0 crosses. (**A**) The number of LR is significantly reduced in the F2 lines that are *lrs1-1/lrs1-1* compared to heterozygotes (*p* = 0.006, unpaired t-test, *n* = 57) or to homozygous wild type plants (*p* = 0.005, unpaired t-test, *n* = 58). There is no significant difference in the number of lateral roots between lines that are *lrs1-1/Lrs1-1* and *Lrs1-1/Lrs1-1* (*p* = 0.67, unpaired t-test, *n* = 101) showing that *lrs1-1* is recessive. (**B**) The primary root length is indistinguishable in all three genotypes (p > 0.2, unpaired t-test, n = 128), showing that this phenotype is not linked to the *lrs1-1* allele. 10-day old seedlings of F2 plants were photographed and the number of lateral roots was counted from the photographs. The genotypes were determined by PCR (Methods). Error bars indicate standard error.

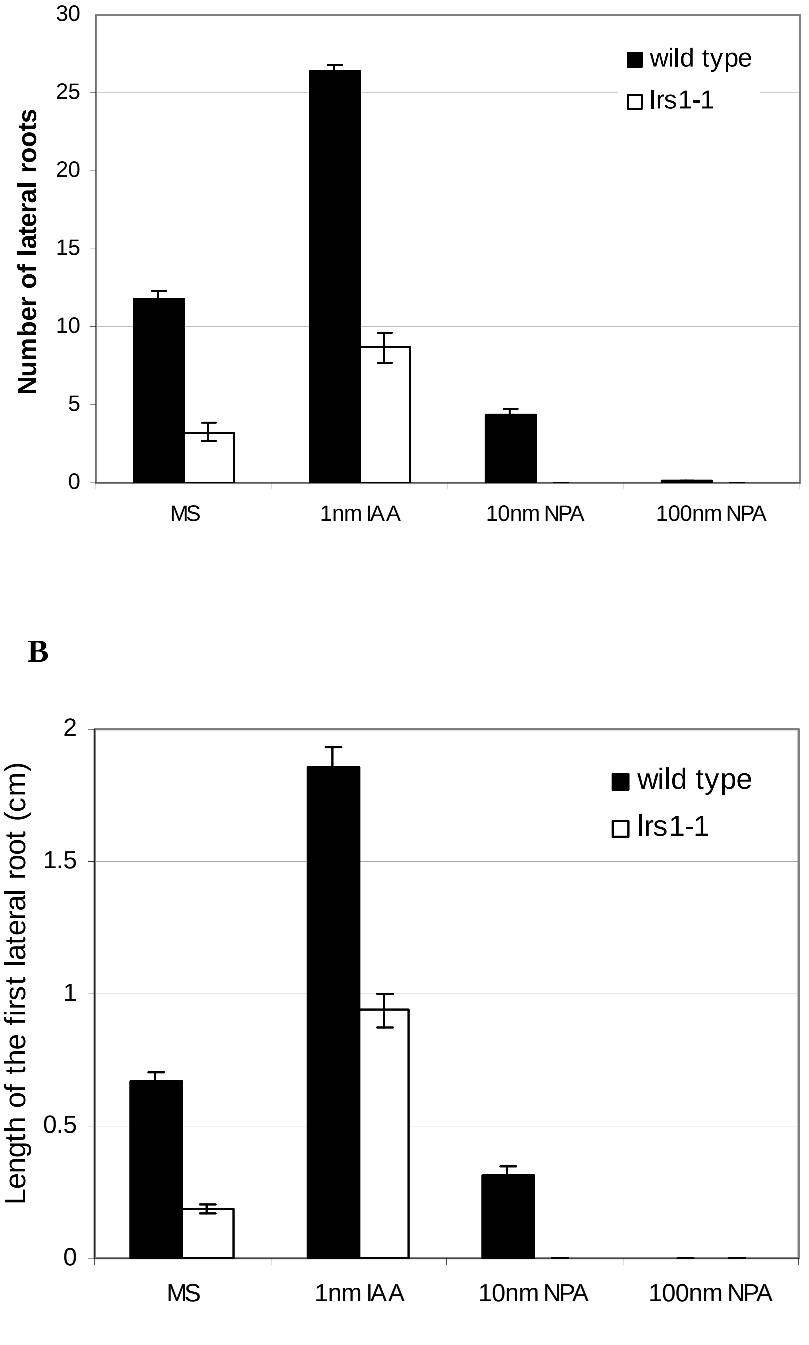
Supplementary Figure 10.



Supplementary Figure 11.

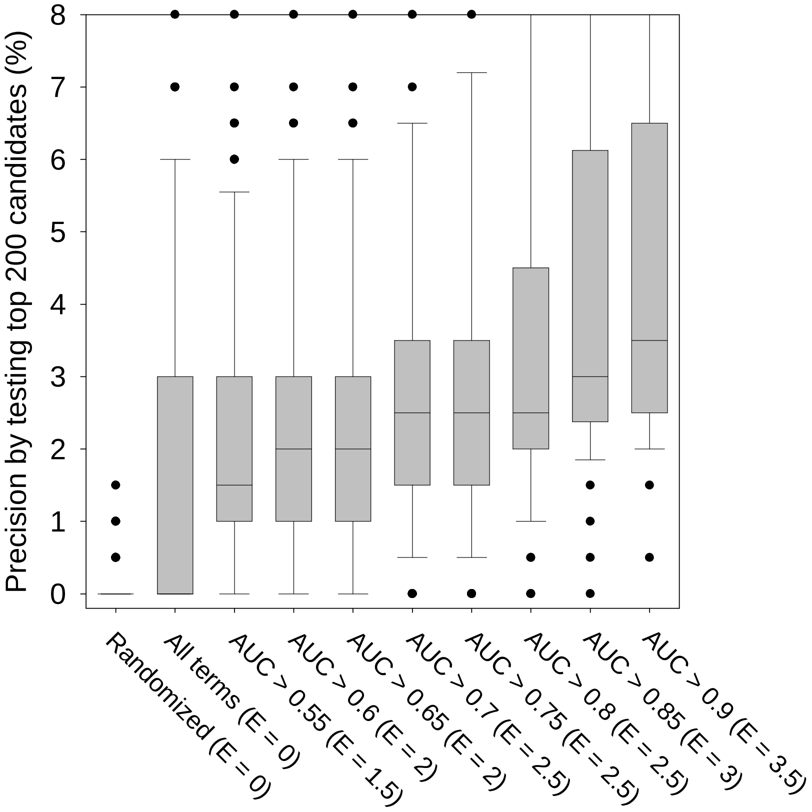
1 nM IAA (native auxin) increases the number and length of lateral roots (LR) in both the wild type and *lrs1-1* seedlings. Auxin transport inhibitor (NPA) decreases both the number and length of LR in both genotypes. Four-day old seedlings were transferred to a medium containing MS (control), 1 nM IAA, 10 nM NPA or 100 nM NPA. The number of LR (**A**) and the length of the oldest LR (**B**) were measured 8 days after the transfer. Seven to 45 plants were measured for each genotype per condition per experiment. Each experiment was repeated at least three times. Error bars indicate standard error**.**

Supplementary Figure 11.



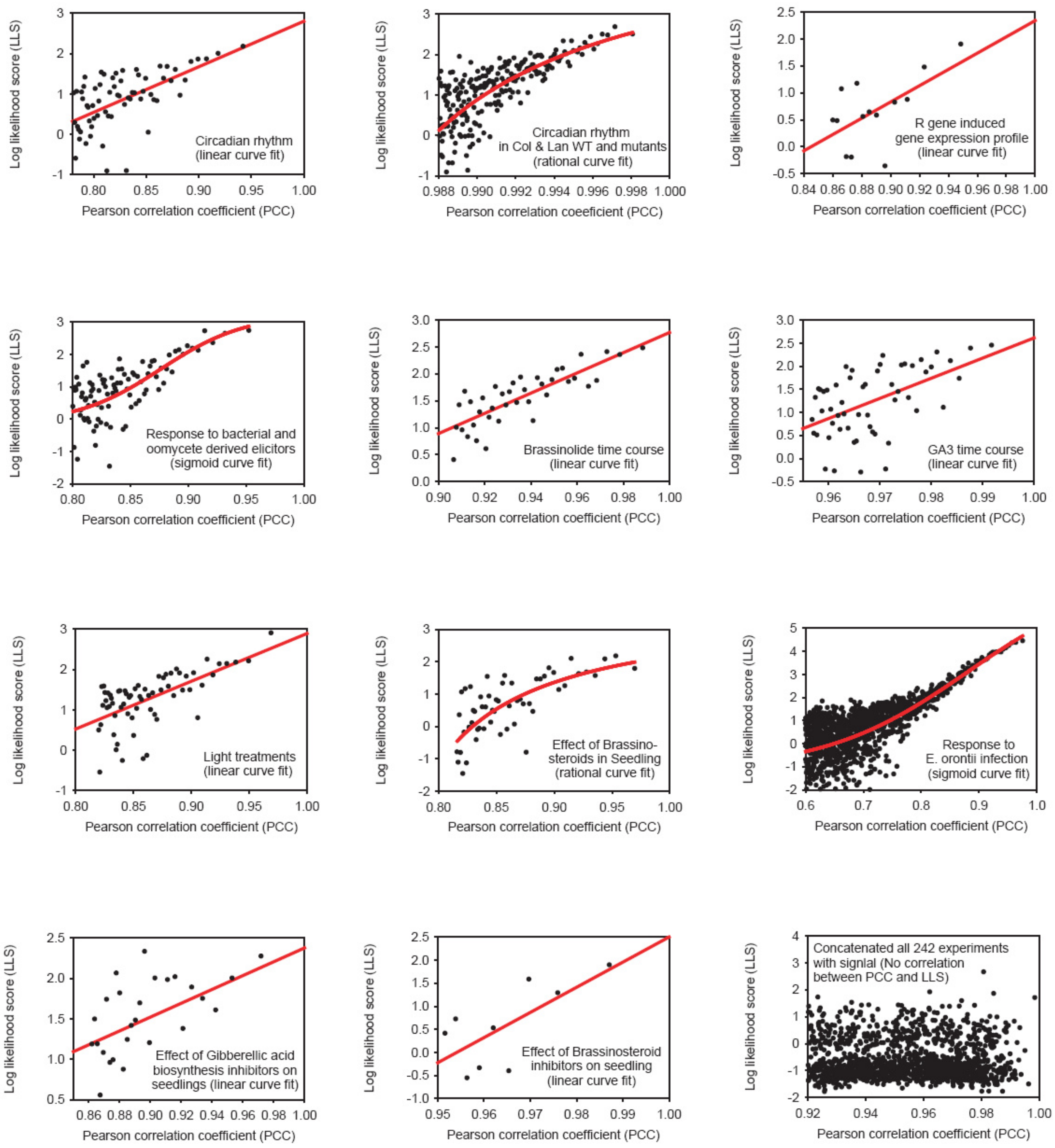
Supplementary Figure 12. Estimates of the proportion of genes newly discovered to be associated with a trait or process as a function of AUC score. For each GO-BP annotation (with >5 genes) predicted above a given AUC threshold (Supplementary Table 14), we calculate the prediction precision for the top 200 new candidate genes for that annotation, assessed using bootstrapping as in Supplementary Figure 15. The median precision (E) across the annotation terms provides an estimate of the number of new genes expected from a focused screen of the top 200 candidate genes, and suggests that one might expect ~4 new genes among the top 200 for the 175 GO-BP terms (out of 317 total) with AUC >0.6. The expected number of hits roughly doubles (~7) for the 32 terms with AUC > 0.9. In contrast, no hits are expected from random sets of 200 genes, on average.

Supplementary Figure 12.



Supplementary Figure 13.
Regression models derived between mRNA co-expression (measured as the Pearson correlation coefficients (PCC) between pairs of genes mRNA expression vectors) and log likelihood scores (*LLS*) for participating in the same biological processes. Each plot represents results for a different set of DNA microarray experiments incorporated into AraNet. The bottom right plot shows the results of concatenating all individual experiment sets into composite expression vectors. Due to the lack of correlation between PCC and LLS, this latter set was not incorporated into AraNet. In all plots, each point represents a bin of 1,000 individual observations, while the red curve indicates the regression model. Datasets are described in detail in Supplementary Table 1A.

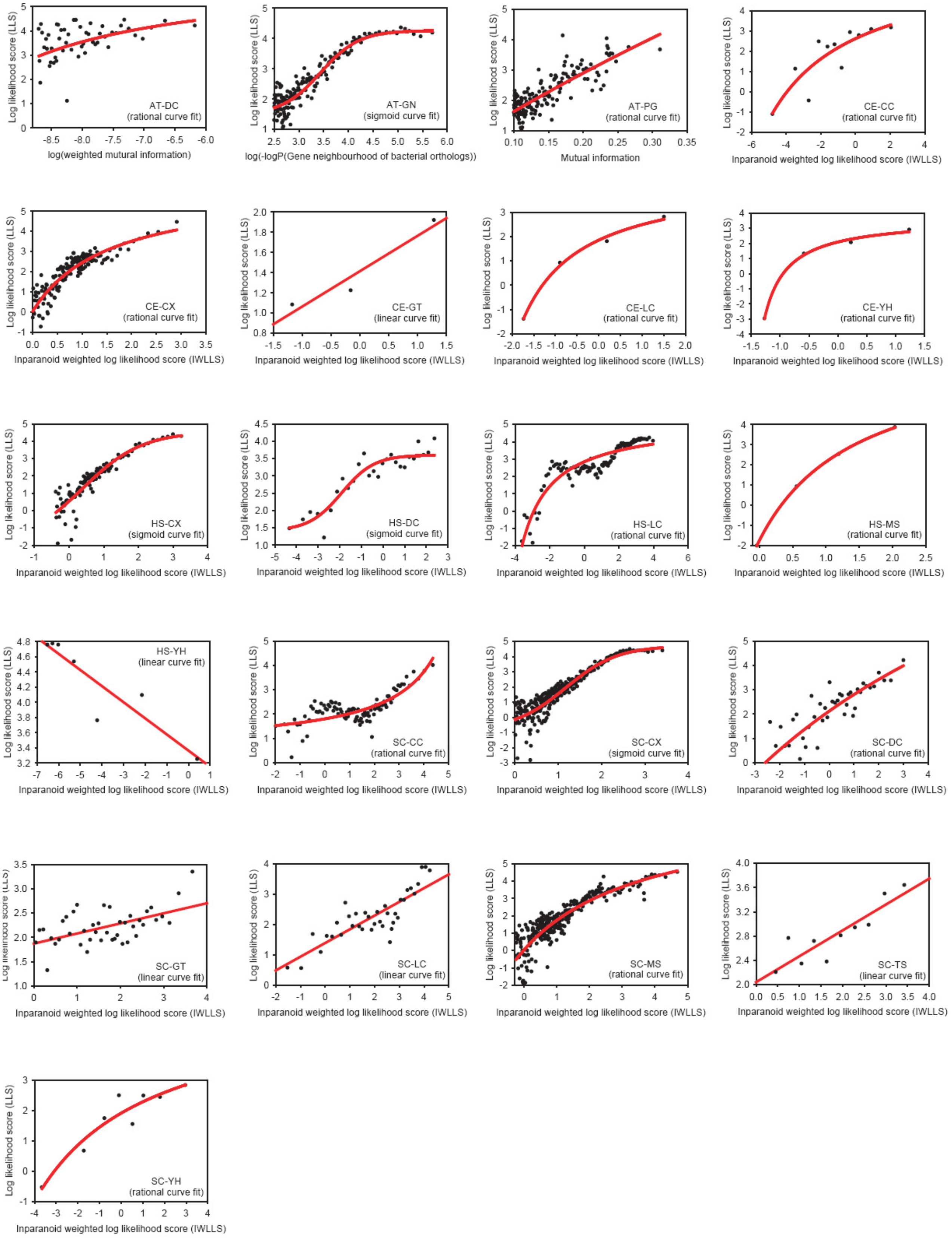
Supplementary Figure 13.



Supplementary Figure 14.

Regression models derived for each functional genomics data set incorporated into AraNet. Each plot shows the relationship between the confidence scores associated with a particular dataset (e.g., INPARANOID-weighted log likelihood scores for datasets transferred by orthology, mutual information for phylogenetic profiles, etc.) and the log likelihood scores (*LLS*) for participating in the same biological processes. Points indicate bins of 1,000 observations; red lines indicate regression models. Datasets are defined in Supplementary Table 2.

Supplementary Figure 14.

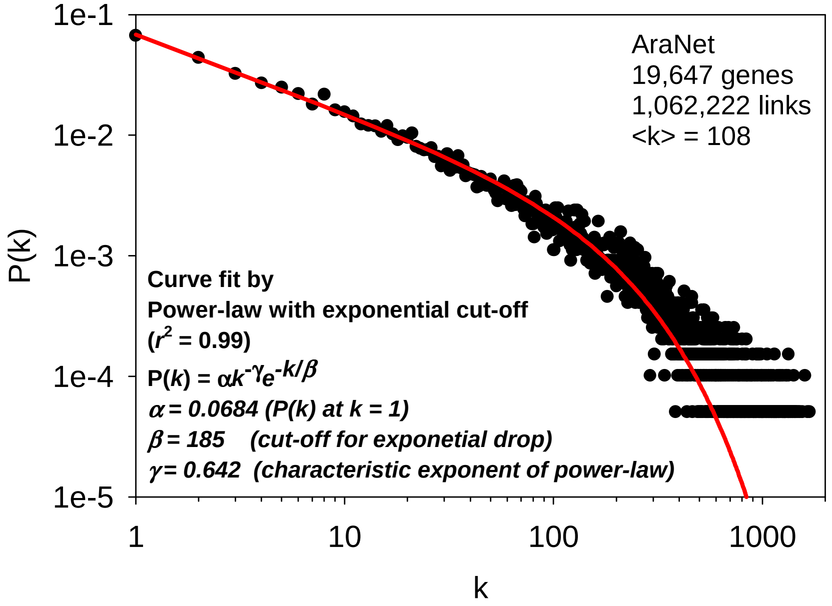


Supplementary Figure 15.

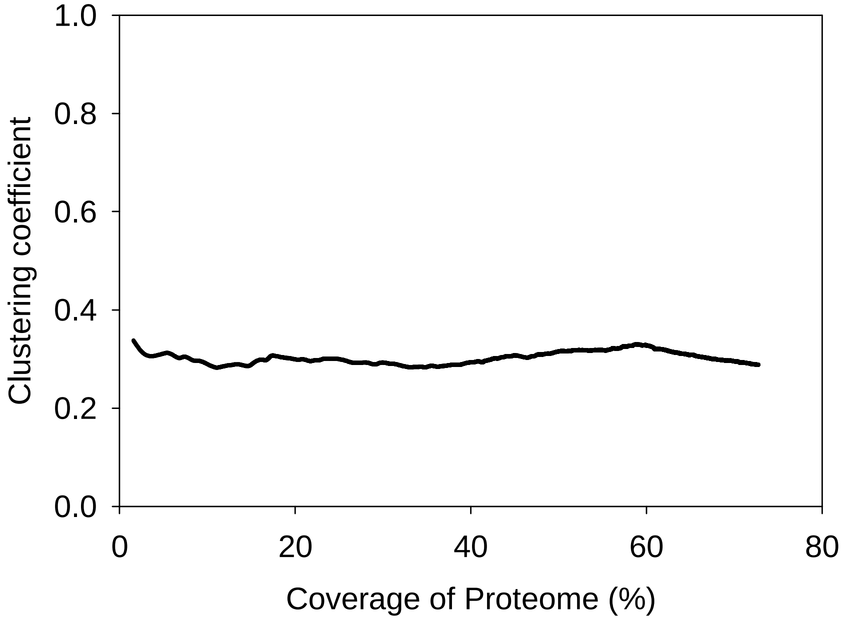
Topological properties of AraNet. (**A**) plots the AraNet degree distribution, plotting P(*k*), the frequency of observing genes in AraNet connected to *k* other genes. (**B**) plots the clustering coefficient of AraNet, calculated as in [22], as a function of network coverage (*i.e.*, rank-ordering network edges by LLS scores and plotting clustering coefficient as a function of decreasing network edge LLS scores).

Supplementary Figure 15.

A.



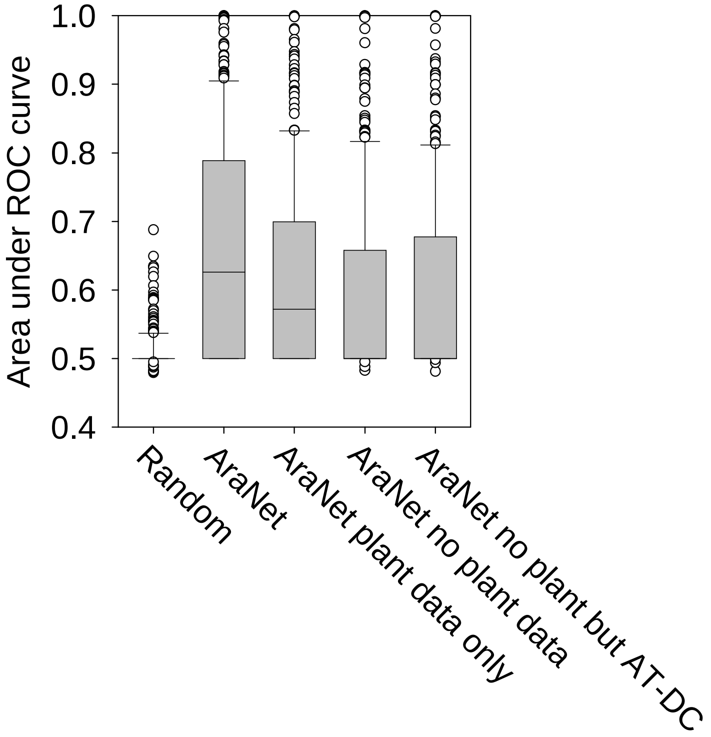
B.



Supplementary Figure 16.

Arabidopsis protein domain annotations play a relatively minor role in AraNet performance compared to other plant datasets. To test this, we constructed a version of AraNet with no plant-derived data but including plant-domain-based links, and tested the performance of this network by ROC analysis as in Figure 2B. If the prediction power depended heavily upon plant domain annotation, we might expect to see significantly better AUCs with the network including domain-based linkages but lacking other plant datasets compared to the network lacking both. In fact, prediction power improves only modestly and in proportion to the expected minor contribution of the plant-domain-based (AT-DC) linkages.

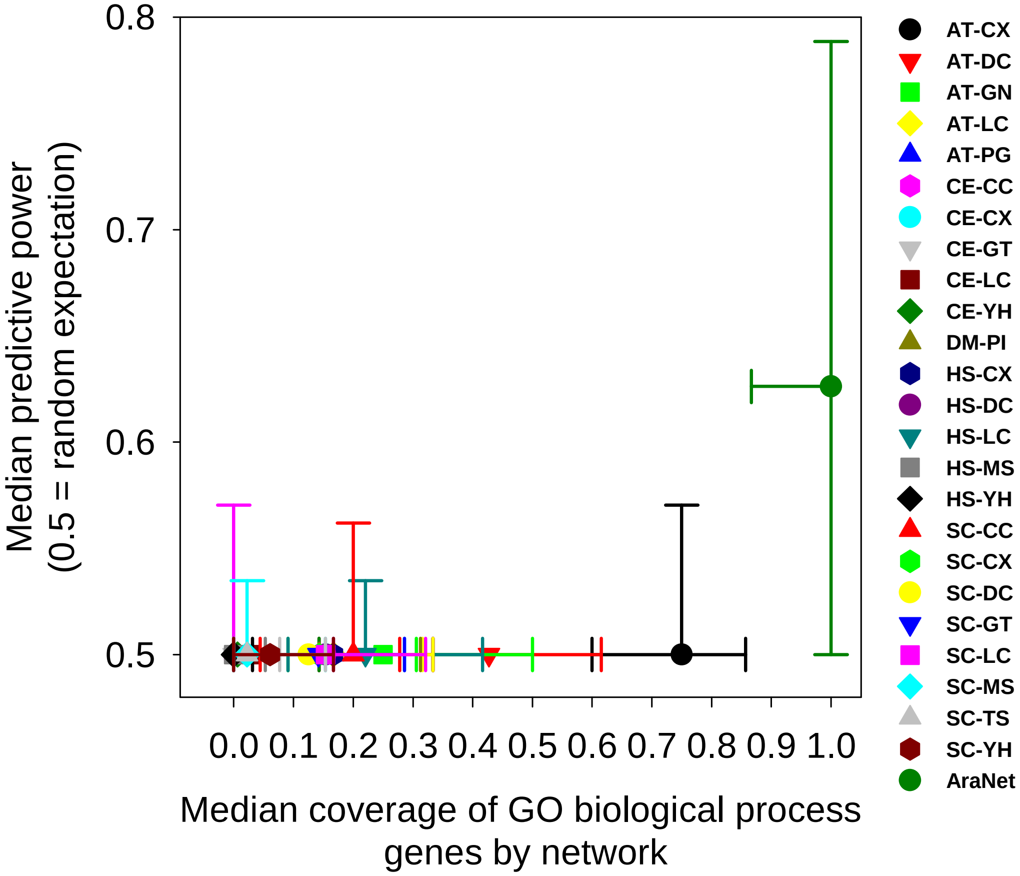
Supplementary Figure 16.



Supplementary Figure 17.

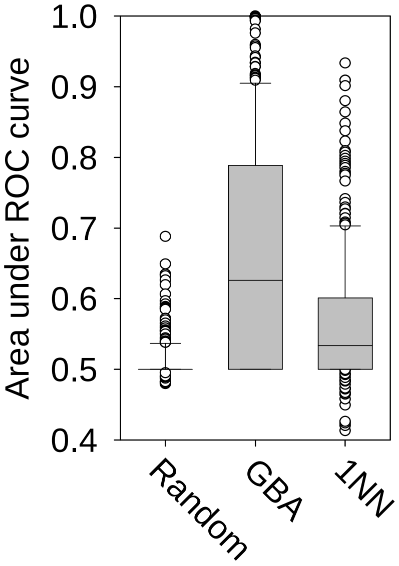
Predictive power of each individual data type, as tested in isolation by ROC analysis similarly to Figure 2B and plotting median AUC versus coverage. Individual data types show much poorer predictive ability than the integrated AraNet. Among the individual data types, plant gene co-expression links shows the strongest predictive power.

Supplementary Figure 17.



Supplementary Figure 18.
Both data integration and the combination of lines of evidence across network edges are important to AraNet performance, as tested by comparing the guilt-by-association analysis of AraNet (as in Figure 2B) to a simple 1-nearest neighbor (1-NN) algorithm using the network, in which each gene was scored for its association with a GO biological process term according to its single strongest network edge. This effectively tests whether consideration of different data types (data integration) alone is the primary driver of performance or whether combining evidence across multiple network edges is also a significant contributor. 1-NN performs significantly worse than the GBA approach, indicating that both data integration and the combination of support from multiple lines of evidence for each gene pair contributes to performance.

Supplementary Figure 18.



References

1. Swarbreck, D. et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* **36**, D1009-1014 (2008).
2. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res* **30**, 42-46 (2002).
3. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555-1558 (2004).
4. Braga-Neto, U.M. & Dougherty, E.R. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **20**, 374-380 (2004).
5. Lee, I., Li, Z. & Marcotte, E.M. An Improved, Bias-Reduced Probabilistic Functional Gene Network of Baker's Yeast, Saccharomyces cerevisiae. *PLoS ONE* **2**, e988 (2007).
6. Hermjakob, H. et al. IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32**, D452-455 (2004).
7. Alfarano, C. et al. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* **33**, D418-424 (2005).
8. de Folter, S. et al. Comprehensive interaction map of the Arabidopsis MADS Box transcription factors. *Plant Cell* **17**, 1424-1433 (2005).
9. Huynen, M., Snel, B., Lathe, W., 3rd & Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* **10**, 1204-1210 (2000).
10. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-4288 (1999).
11. Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S. & Koonin, E.V. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* **11**, 356-372 (2001).
12. Bowers, P.M. et al. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* **5**, R35 (2004).
13. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**, 324-328 (1998).
14. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896-2901 (1999).
15. Lee, I. et al. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nature genetics* **40**, 181-188 (2008).
16. Mulder, N.J. et al. New developments in the InterPro database. *Nucleic Acids Res* **35**, D224-228 (2007).
17. Remm, M., Storm, C.E. & Sonnhammer, E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**, 1041-1052 (2001).
18. Breitkreutz, B.J. et al. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* (2007).
19. Chatr-aryamontri, A. et al. MINT: the Molecular INTeraction database. *Nucleic Acids Res* **35**, D572-574 (2007).
20. Giot, L. et al. A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727-1736 (2003).

21. Barabasi, A.L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509-512 (1999).

22. Watts, D.J. & Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440-442 (1998).

23. Gutierrez, R.A. et al. Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis. *Genome Biol* **8**, R7 (2007).

24. Brady, S.M. et al. A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science* **318**, 801-806 (2007).

25. Tzafrir, I. et al. The Arabidopsis SeedGenes Project. *Nucleic Acids Res* **31**, 90-93 (2003).

26. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-408 (2001).

27. Nakagawa, T. et al. Development of series of gateway binary vectors, pGWBs, for realizing efficient construction of fusion genes for plant transformation. *Journal of bioscience and bioengineering* **104**, 34-41 (2007).

28. Clough, S.J. & Bent, A.F. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J* **16**, 735-743 (1998).

29. Mishra, G.R. et al. Human protein reference database--2006 update. *Nucleic Acids Res* **34**, D411-414 (2006).

30. Rual, J.F. et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178 (2005).

31. Stelzl, U. et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957-968 (2005).

32. Geisler-Lee, J. et al. A predicted interactome for Arabidopsis. *Plant Physiol* **145**, 317-329 (2007).

33. Cui, J. et al. AtPID: Arabidopsis thaliana protein interactome database an integrative platform for plant systems biology. *Nucleic Acids Res* (2007).

34. Ma, S., Gong, Q. & Bohnert, H.J. An Arabidopsis gene network based on the graphical Gaussian model. *Genome Res* **17**, 1614-1625 (2007).

35. Karniol, B., Malec, P. & Chamovitz, D.A. Arabidopsis FUSCA5 encodes a novel phosphoprotein that is a component of the COP9 complex. *The Plant cell* **11**, 839-848 (1999).

36. Meinke, D., Muralla, R., Sweeney, C. & Dickerman, A. Identifying essential genes in Arabidopsis thaliana *Trends in Plant Science* **13**, 483-491 (2008).

37. Norris, S.R., Barrette, T.R. & DellaPenna, D. Genetic dissection of carotenoid synthesis in arabidopsis defines plastoquinone as an essential component of phytoene desaturation. *The Plant cell* **7**, 2139-2149 (1995).

38. Stephenson, P.G. & Terry, M.J. Light signalling pathways regulating the Mg-chelatase branchpoint of chlorophyll synthesis during de-etiolation in Arabidopsis thaliana. *Photochem Photobiol Sci* **7**, 1243-1252 (2008).

39. Hricova, A., Quesada, V. & Micol, J.L. The SCABRA3 nuclear gene encodes the plastid RpoTp RNA polymerase, which is required for chloroplast biogenesis and mesophyll cell proliferation in Arabidopsis. *Plant Physiol* **141**, 942-956 (2006).

40. Long, D. et al. The maize transposable element system Ac/Ds as a mutagen in Arabidopsis: identification of an albino mutation induced by Ds insertion. *Proceedings of*

the National Academy of Sciences of the United States of America **90**, 10370-10374 (1993).

41. Martinez-Zapater, J.M. Genetic analysis of variegated mutants in Arabidopsis. *Journal of Heredity* **84**, 138-140 (1993).
42. Deng, X.W., Caspar, T. & Quail, P.H. cop1: a regulatory locus involved in light-controlled development and gene expression in Arabidopsis. *Genes & development* **5**, 1172-1182 (1991).
43. Reiter, R.S., Coomber, S.A., Bourett, T.M., Bartley, G.E. & Scolnik, P.A. Control of leaf and chloroplast development by the Arabidopsis gene pale cress. *The Plant cell* **6**, 1253-1264 (1994).
44. Sauret-Gueto, S. et al. Plastid cues posttranscriptionally regulate the accumulation of key enzymes of the methylerythritol phosphate pathway in Arabidopsis. *Plant Physiol* **141**, 75-84 (2006).
45. Pfalz, J., Liere, K., Kandlbinder, A., Dietz, K.J. & Oelmuller, R. pTAC2, -6, and -12 are components of the transcriptionally active plastid chromosome that are required for plastid gene expression. *The Plant cell* **18**, 176-197 (2006).
46. Dormann, P., Hoffmann-Benning, S., Balbo, I. & Benning, C. Isolation and characterization of an Arabidopsis mutant deficient in the thylakoid lipid digalactosyl diacylglycerol. *The Plant cell* **7**, 1801-1810 (1995).
47. Gaubier, P., Wu, H.J., Laudie, M., Delseny, M. & Grellet, F. A chlorophyll synthetase gene from Arabidopsis thaliana. *Mol Gen Genet* **249**, 58-64 (1995).
48. Castle, L.A. & Meinke, D.W. A FUSCA gene of Arabidopsis encodes a novel protein essential for plant development. *The Plant cell* **6**, 25-41 (1994).
49. Errampalli, D. et al. Embryonic Lethals and T-DNA Insertional Mutagenesis in Arabidopsis. *The Plant cell* **3**, 149-157 (1991).
50. Motohashi, R. et al. An essential role of a TatC homologue of a Delta pH- dependent protein transporter in thylakoid membrane formation during chloroplast development in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10499-10504 (2001).
51. Chory, J., Peto, C., Feinbaum, R., Pratt, L. & Ausubel, F. Arabidopsis thaliana mutant that develops as a light-grown plant in the absence of light. *Cell* **58**, 991-999 (1989).
52. Wei, N. & Deng, X.W. COP9: a new genetic locus involved in light-regulated development and gene expression in arabidopsis. *The Plant cell* **4**, 1507-1518 (1992).
53. Mandel, M.A., Feldmann, K.A., Herrera-Estrella, L., Rocha-Sosa, M. & Leon, P. CLA1, a novel gene required for chloroplast development, is highly conserved in evolution. *Plant J* **9**, 649-658 (1996).
54. Koncz, C. et al. Isolation of a gene encoding a novel chloroplast protein by T-DNA tagging in Arabidopsis thaliana. *The EMBO journal* **9**, 1337-1346 (1990).
55. Redei, G.P. Regulation of plastid differentiation in mutant im by visible light. **2**, 26 (1965).
56. Wetzel, C.M., Jiang, C.Z., Meehan, L.J., Voytas, D.F. & Rodermel, S.R. Nuclear-organelle interactions: the immutans variegation mutant of Arabidopsis is plastid autonomous and impaired in carotenoid biosynthesis. *Plant J* **6**, 161-175 (1994).
57. Wei, N., Chamovitz, D.A. & Deng, X.W. Arabidopsis COP9 is a component of a novel signaling complex mediating light control of development. *Cell* **78**, 117-124 (1994).

58. Schwender, J., Muller, C., Zeidler, J. & Lichtenthaler, H.K. Cloning and heterologous expression of a cDNA encoding 1-deoxy-D-xylulose-5-phosphate reductoisomerase of Arabidopsis thaliana. *FEBS Lett* **455**, 140-144 (1999).
59. Wei, N., Serino, G. & Deng, X.W. The COP9 signalosome: more than a protease. *Trends in biochemical sciences* **33**, 592-600 (2008).
60. Gusmaroli, G., Figueroa, P., Serino, G. & Deng, X.W. Role of the MPN subunits in COP9 signalosome assembly and activity, and their regulatory interaction with Arabidopsis Cullin3-based E3 ligases. *The Plant cell* **19**, 564-581 (2007).
61. Osterlund, M.T., Hardtke, C.S., Wei, N. & Deng, X.W. Targeted destabilization of HY5 during light-regulated development of Arabidopsis. *Nature* **405**, 462-466 (2000).
62. Yanagawa, Y. et al. Arabidopsis COP10 forms a complex with DDB1 and DET1 in vivo and enhances the activity of ubiquitin conjugating enzymes. *Genes & development* **18**, 2172-2181 (2004).
63. Asakura, Y., Kikuchi, S. & Nakai, M. Non-identical contributions of two membrane-bound cpSRP components, cpFtsY and Alb3, to thylakoid biogenesis. *Plant J* **56**, 1007-1017 (2008).
64. Peltier, J.B. et al. Central functions of the lumenal and peripheral thylakoid proteome of Arabidopsis determined by experimentation and genome-wide prediction. *The Plant cell* **14**, 211-236 (2002).
65. Masuda, T. & Fujita, Y. Regulation and evolution of chlorophyll metabolism. *Photochem Photobiol Sci* **7**, 1131-1149 (2008).
66. Shen, Y.Y. et al. The Mg-chelatase H subunit is an abscisic acid receptor. *Nature* **443**, 823-826 (2006).
67. Mochizuki, N., Brusslan, J.A., Larkin, R., Nagatani, A. & Chory, J. Arabidopsis genomes uncoupled 5 (GUN5) mutant reveals the involvement of Mg-chelatase H subunit in plastid-to-nucleus signal transduction. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 2053-2058 (2001).
68. Schmid, M. et al. A gene expression map of Arabidopsis thaliana development. *Nature genetics* **37**, 501-506 (2005).