Systematic prediction of gene function in *Arabidopsis thaliana* using a probabilistic functional gene network

Sohyun Hwang[1], Seung Y Rhee[2], Edward M Marcotte[3,4] & Insuk Lee[1]

[1]Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, Seoul, Korea. [2]Department of Plant Biology, Carnegie Institution for Science, Stanford, California, USA. [3]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, USA. [4]Department of Chemistry and Biochemistry, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, USA. Correspondence should be addressed to S.Y.R. (rhee@acoma.stanford.edu), E.M.M. (marcotte@icmb.utexas.edu) or I.L. (insuklee@yonsei.ac.kr).

Published online 25 August 2011; doi:10.1038/nprot.2011.372

AraNet is a functional gene network for the reference plant *Arabidopsis* and has been constructed in order to identify new genes associated with plant traits. It is highly predictive for diverse biological pathways and can be used to prioritize genes for functional screens. Moreover, AraNet provides a web-based tool with which plant biologists can efficiently discover novel functions of *Arabidopsis* genes (http://www.functionalnet.org/aranet/). This protocol explains how to conduct network-based prediction of gene functions using AraNet and how to interpret the prediction results. Functional discovery in plant biology is facilitated by combining candidate prioritization by AraNet with focused experimental tests.

INTRODUCTION

The engineering of plant traits holds significant promise for improving the production of food, fiber and renewable energy. Genetic engineering of plants can benefit from the identification of genes that have roles in traits of interest. Over the past few decades, plant biologists have been seeking genes that can be modified to obtain many desirable traits, such as high yield and stress tolerance. One very effective approach that has been used to identify genes for important cellular or organismal traits in other organisms is the network-guided guilt-by-association approach[1,2]. Using traditional forward genetics to identify these genes is limited because mutations for many genes may generate only moderate or weak phenotypes that may not be sensitive enough for the given screening method. Reverse genetics allows examination of more subtle phenotypic changes by directed assaying of each mutant[3]; however, genome-wide phenotyping for many plant traits is impractical.

In the network-guided guilt-by-association approach, we first construct a functional gene network and then computationally suggest new candidate genes for traits on the basis of the connectivity of the genes in the network. In a functional gene network[4], a link is made between two genes if they participate in a common biological process or pathway; therefore, the links between two genes represent functional associations between them. The functional gene network can integrate heterogeneous biological data into a single model; integrating many independent data sets enhances both model accuracy and coverage[4,5]. The network-guided screening method has been successfully applied to unicellular organisms[1,6] and *Caenorhabditis elegans*[7–9], and it has been used for identifying human disease genes[10–13]. Functional networks for a broad collection of organisms are available through the STRING[14] and GeneMANIA[15] websites.

For customizing network-guided guilt-by-association methodology to plants, we have constructed a probabilistic functional gene network for the reference plant *Arabidopsis* (AraNet) by a modified Bayesian integrating diverse 'omics' data from multiple organisms (e.g., *C. elegans, Drosophila melanogaster, Homo sapiens and Saccharomyces cerevisiae*), with each data type weighted according to how well it links genes that are known to function together in *Arabidopsis thaliana*[16] (**Box 1**). Each interaction in AraNet has an associated log-likelihood score (LLS) that measures the probability of an interaction representing a true functional linkage between two genes. AraNet consists of 1,062,222 functional associations among 19,647 genes of *Arabidopsis thaliana* (~73% of the total *Arabidopsis* genes). This genome coverage is far beyond that of genes with any Gene Ontology (GO) annotations (~45% coverage of the genome, including 33% by computational inferences only). The functional gene associations of AraNet are highly predictive for diverse biological pathways. A previous study of network prediction power showed that AraNet outperformed other published *Arabidopsis* gene networks[16–20].

We provide AraNet as a web tool (http://www.functionalnet.org/aranet) to prioritize genes for candidate-based functional screening by which plant biologists can more efficiently discover novel functions for uncharacterized *Arabidopsis* genes. Therefore, in this protocol, we describe how to use the AraNet web tool for predicting new candidate genes for plant traits or identifying new biological function of uncharacterized genes. The website provides two complementary search options: (i) 'Find new members of a pathway', which predicts new candidate genes using a set of query genes known or inferred to be involved in the same pathway; and (ii) 'Infer function from network neighbors', which predicts the candidate GO biological process terms for a query gene (**Fig. 1**). These two search procedures can be performed individually, but we suggest combining them to maximize the effectiveness of biological function discovery using the website.

New candidate genes predicted using AraNet should then be tested experimentally to confirm the relationships. For example, we experimentally tested new candidate genes associated with the set of 23 known embryo pigmentation genes. From the top 200 candidate genes suggested by AraNet, we tested 90 genes with available homozygous transfer DNA (T-DNA) insertional mutant lines. A total of 14 genes exhibited color and morphology defects in young seedlings, reminiscent of embryo pigmentation mutants. This represents a tenfold enrichment in the discovery rate of the

BOX 1 | CONSTRUCTING ARANET BY INTEGRATING OMICS DATA

Constructing a network of functional associations between *Arabidopsis* genes is essentially an exercise in appropriately weighting various types of experimental data according to how well they reconstruct a reference set of gene-gene functional associations (Gold standard–positives, GSP), composed of pairs of genes sharing the same GO biological process (GO-BP) annotation, as flagged by experimental evidence codes in The *Arabidopsis* Information Resource (TAIR; downloaded from ftp://ftp.geneontology.org/pub/go/gene-associations/gene_association.tair.gz)[23]. Pairs of annotated genes that do not share GO-BP annotation terms are considered negative examples (Gold standard–negatives, GSN).

The experimental data sets incorporated into the network are diverse, and include mRNA coexpression patterns, protein-protein interaction data, genomic contexts of orthologous proteins, protein domain co-occurrence profiles and functional linkage data transferred from other organisms by orthology relationships (identified by INPARANOID[27]). To integrate those heterogeneous data into an integrated model of functional associations, we use a scoring scheme for the linkages based on Bayesian statistics[2,4]. It measures how likely pairs of genes are to be functionally associated, on the basis of how well the relevant experimental data sets capture the set of trusted functional associations (GSP) compared with the set of negative associations (GSN). The log-likelihood score (LLS) for a pair of genes to be functionally associated can be calculated as follows:

$$\mathrm{LLS} = \ln\left(\frac{\mathrm{P(GSP}|D)/\mathrm{P(GSN}|D)}{\mathrm{P(GSP)/P(GSN)}}\right)$$

P(GSP|D) is the number of gold standard–*positive* gene pairs given the experimental data.
P(GSN|D) is the number of gold standard–*negative* gene pairs given the experimental data.
P(GSP) is the total number of gold standard–*positive* gene pairs.
P(GSN) is the total number of gold standard–*negative* gene pairs.

For data sets in which each gene pair is associated with a continuous data score (e.g., correlation coefficients from mRNA coexpression data sets), LLS scores are calculated for bins containing equal numbers of rank-ordered gene pairs. These LLS scores and their corresponding data scores (the mean score for each bin) are fit with regression models and LLS scores for each gene pair assigned from these models. LLS scores from the various data sets are then combined into an integrated score for each gene pair, using a heuristic approach—a weighted sum integration[4] method using linearly decaying weights—that has been empirically observed to perform well on our network data sets (see the weighted sum method in **Box 2**).

mutant phenotype over that observed during a forward-genetics screen of T-DNA insertion lines[16].

Experimental design

If you have connected to the AraNet website to find new member genes of a pathway, a statistically powerful way to explore AraNet is to perform a search with a set of multiple genes known or suspected to be involved in a common pathway or phenotype[7]. Here we query the network with known genes to collect new functional information of their network neighbors ('Find new members of a pathway' option). These genes are called 'query genes' in this search method. The web tool assesses the predictability of AraNet with the given set of query genes by measuring how well the query genes are connected to each other in AraNet. (Note that the assessment of prediction quality requires at least four known query genes to be statistically meaningful. Candidate genes can be identified by using even a single query gene, but the prediction quality cannot be measured using this approach.)

The predictive power of AraNet for the particular query genes is measured using receiver operator characteristic (ROC) curve analysis. ROC curve analysis is a useful technique for organizing classifiers and visualizing their performance[21]. ROC curves are 2D graphs in which the true-positive (TP) rate is plotted on the *y* axis and the false-positive (FP) rate is plotted on the *x* axis. The TP rate of a classifier is estimated as:

$$\text{TP rate} = \frac{\text{positives correctly classified}}{\text{total positives}} = \frac{\text{query genes correctly classified}}{\text{total query genes}}$$

The classifier ranks all *Arabidopsis* genes (including those that are not present in AraNet) by the sum of AraNet connections from each gene to the set of query genes with edge weights. Thus, if the query genes are well connected to one another, they will be highly ranked in the classifier. The analysis assumes that the query genes are the positives and all other *Arabidopsis* genes are the negatives (a fairly stringent condition, as we posit that there are more genes in the genome that are involved in the same process as the query genes). Consequently, the TP rate is the ratio of query genes correctly classified among the total query genes by the given classifier threshold. The FP rate of the classifier is

$$\text{FP rate} = \frac{\text{negatives incorrectly classified}}{\text{total negatives}} = \frac{\text{nonquery genes incorrectly classified}}{\text{total nonquery genes}}$$

Similarly, in our analysis, the FP rate is the ratio of nonquery genes incorrectly classified as query genes among the total nonquery genes at the given classifier threshold.

After drawing the ROC curves, we calculate the area under the ROC curve (AUC) in order to provide a numerical measure of prediction strength, ranging from ~0.5 for random expectation to 1 for perfect predictions. For example, if the AUC score is higher than 0.68 (our empirical and very approximate choice of AUC threshold score for predictable gene function) and the *P* value is lower than 0.05, AraNet can be considered to have predictive power for inferring gene function related to a query gene set. The top 200 candidate genes associated with the set of valid query genes by AraNet will be suggested as the new member genes of the pathway.

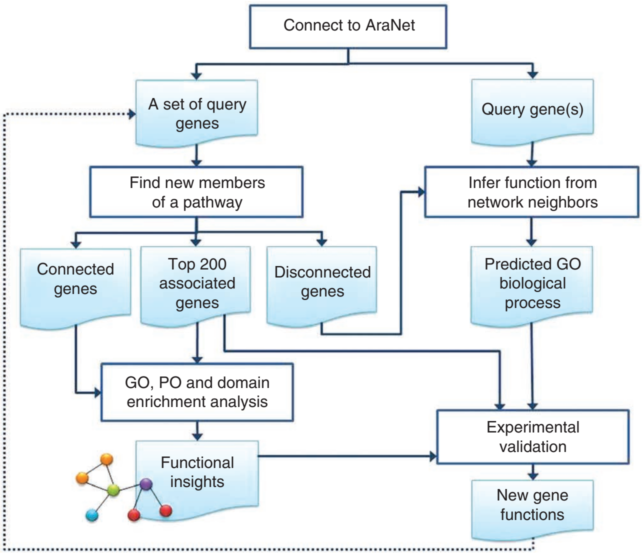


Figure 1 | Overview of the method of using AraNet to discover gene function. The AraNet web tool can be divided into two search paths for identifying new gene functions: 'Find new members of a pathway' and 'Infer function from network neighbors'. If you submit a set of query genes to the 'Find new members of a pathway' search, you can retrieve three gene sets: connected query genes, disconnected query genes and the top 200 candidate genes that connect to the query genes. The gene set enrichment analyses using connected query genes and the top 200 candidate genes can provide biological insight from enriched GO, PO and protein domain terms. The top 200 candidate genes can be tested directly for identifying new gene functions. If you submit query gene(s) to 'Infer function from network neighbors' search, you can obtain candidate GO biological process terms for each query gene. An alternative source of the query gene is the disconnected query genes from the 'Find new members of a pathway' search. Predicted GO biological process terms for each query gene can be tested to discover new gene functions. The genes with newly discovered functions may update the query genes for the next round of the 'Find new members of a pathway' search. Such iterative searching can improve the enrichment of GO, PO or protein domain terms in subsequent analysis.

As AraNet connects functionally associated genes, it can be applied to discover the common biological functions (e.g., biological pathways or GO terms) among connected genes (e.g., a set of query genes or a set of the top 200 candidate genes). For the functional analysis of gene sets rather than individual genes, AraNet provides a 'Gene set enrichment analysis' tool. This analysis can determine whether any biological process found in the reference gene set is over-represented among the listed genes. First, we calculate the hypergeometric *P* values of the intersection between the listed genes and the genes in the reference gene set annotated with the same GO term, and then we adjust the *P* value as *q* value, which is an extension of a quantity called the false discovery rate, to solve the multiple hypotheses testing problem (because we test many gene sets for the analysis)[22]. The 'gene set enrichment analysis' of AraNet uses three types of reference gene sets: GO[23], Plant Ontology (PO)[24] and InterPro[25] protein domain annotations obtained from The *Arabidopsis* Information Resource[26]. The *Arabidopsis* Information Resource (http://www.arabidopsis.org/) maintains a database of genetic and molecular biology data for the model plant *A. thaliana* and provides extensive gene annotations from manually extracted, experimentally derived data in the literature, as well as from computational predictions[23].

Alternatively, you can use the AraNet web tool to predict functions of individual genes from their network neighbors ('Infer function from network neighbors' option). This search can be used for any genes of interest, including completely uncharacterized genes or the query genes that were used for the 'Find new members of a pathway' search that were found to be disconnected from each other. This prediction method suggests multiple GO biological process terms as candidate functions for each query gene based on known GO biological process terms annotated to its neighbors in AraNet. The candidate GO terms are rank ordered by the sum of the edge weights (LLS) to all neighbors annotated with each GO term.

Searching with genes from plant species other than *Arabidopsis*. If you are interested in searching for functions of genes from plants other than *Arabidopsis*, you can try to find *Arabidopsis* homologs based on sequence analysis such as BLAST at NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) or look for homologs in databases such as INPARANOID[27] (http://inparanoid.sbc.su.se/cgi-bin/index.cgi). Please note that we have not yet tested performance of AraNet on genes from other plant species and the results may be less reliable.

Below we detail a step-by-step procedure for predicting new candidate genes for a variety of biological processes using the two AraNet search applications described above ('Find new members of a pathway' and 'Infer function from network neighbors'). These predictions should be considered as first-pass clues serving as a guide to generate hypotheses regarding gene function, which should then be tested experimentally to confirm the gene-phenotype association.

MATERIALS

EQUIPMENT

- Internet access, web browser
- Cytoscape software (optional)

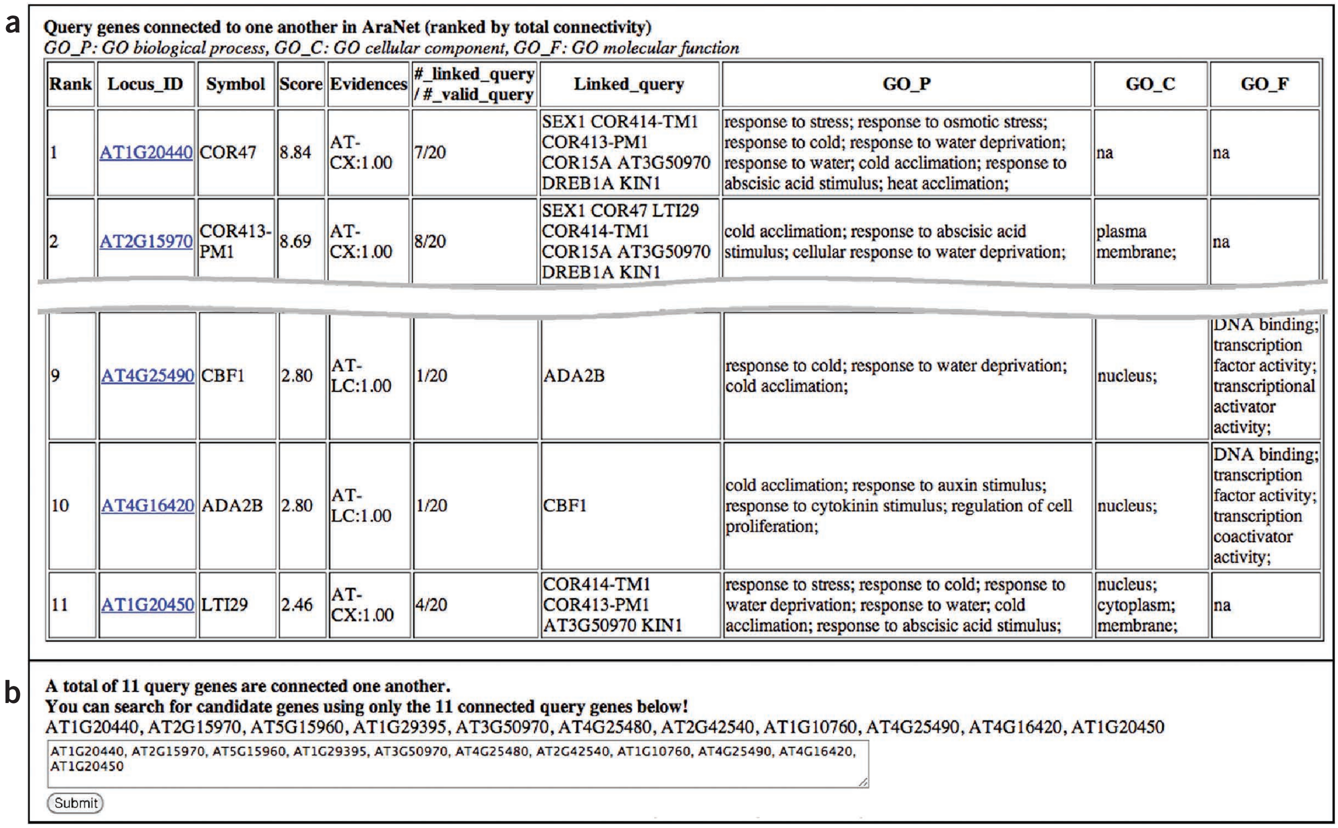
EQUIPMENT SETUP

Data An *Arabidopsis* gene of interest or a group of *Arabidopsis* genes that are known or predicted to be involved in a common pathway or phenotype. Our protocol uses the sample query genes given on the AraNet website—20 genes associated with *Arabidopsis* cold acclimation—for demonstration purposes.

PROCEDURE

Connect to the AraNet website

1| Start a web browser such as Internet Explorer, Firefox or Netscape. Go to http://www.functionalnet.org/aranet/. The AraNet homepage provides three links: 'About AraNet', 'AraNet Search' and 'Batch Downloads'. The 'About AraNet' page briefly describes AraNet and how to use it. On the 'Batch Downloads' page, you can download the benchmark set (341,821 reference functional associations between 6,487 *Arabidopsis* genes sharing GO biological process annotations[23]) and the full-size AraNet (1,062,222 functional associations among 19,647 *Arabidopsis* genes).



Query genes connected to one another in AraNet (ranked by total connectivity)
GO_P: GO biological process, GO_C: GO cellular component, GO_F: GO molecular function

Rank	Locus_ID	Symbol	Score	Evidences	#_linked_query / #_valid_query	Linked_query	GO_P	GO_C	GO_F
1	AT1G20440	COR47	8.84	AT-CX:1.00	7/20	SEX1 COR414-TM1 COR413-PM1 COR15A AT3G50970 DREB1A KIN1	response to stress; response to osmotic stress; response to cold; response to water deprivation; response to water; cold acclimation; response to abscisic acid stimulus; heat acclimation;	na	na
2	AT2G15970	COR413-PM1	8.69	AT-CX:1.00	8/20	SEX1 COR47 LTI29 COR414-TM1 COR15A AT3G50970 DREB1A KIN1	cold acclimation; response to abscisic acid stimulus; cellular response to water deprivation;	plasma membrane;	na
9	AT4G25490	CBF1	2.80	AT-LC:1.00	1/20	ADA2B	response to cold; response to water deprivation; cold acclimation;	nucleus;	DNA binding; transcription factor activity; transcriptional activator activity;
10	AT4G16420	ADA2B	2.80	AT-LC:1.00	1/20	CBF1	cold acclimation; response to auxin stimulus; response to cytokinin stimulus; regulation of cell proliferation;	nucleus;	DNA binding; transcription factor activity; transcription coactivator activity;
11	AT1G20450	LTI29	2.46	AT-CX:1.00	4/20	COR414-TM1 COR413-PM1 AT3G50970 KIN1	response to stress; response to cold; response to water deprivation; response to water; cold acclimation; response to abscisic acid stimulus;	nucleus; cytoplasm; membrane;	na

A total of 11 query genes are connected one another.
You can search for candidate genes using only the 11 connected query genes below!
AT1G20440, AT2G15970, AT5G15960, AT1G29395, AT3G50970, AT4G25480, AT2G42540, AT1G10760, AT4G25490, AT4G16420, AT1G20450

```
AT1G20440, AT2G15970, AT5G15960, AT1G29395, AT3G50970, AT4G25480, AT2G42540, AT1G10760, AT4G25490, AT4G16420,
AT1G20450
```

Figure 2 | A report from a 'Find new members of a pathway' search showing an analysis of query genes (e.g., a set of genes involved in cold acclimation) connected to one another in AraNet. (**a**) The list of connected query genes contains information about the rank on the basis of: the total connection score to other query genes (**Box 2**), locus ID, gene symbol, AraNet data types (evidence) supporting connections between the gene and all other query genes (**Table 1**), the fraction of connected query genes out of the total valid query genes, all other query genes connected to the gene, and three Gene Ontology (GO) annotations (biological process, cellular components and molecular function). The Locus ID links to the annotation page at the TAIR database[26]. (**b**) The next round of search using only connected query genes can be run by clicking the 'Submit' button at the bottom of the screen.

Find new members of a pathway

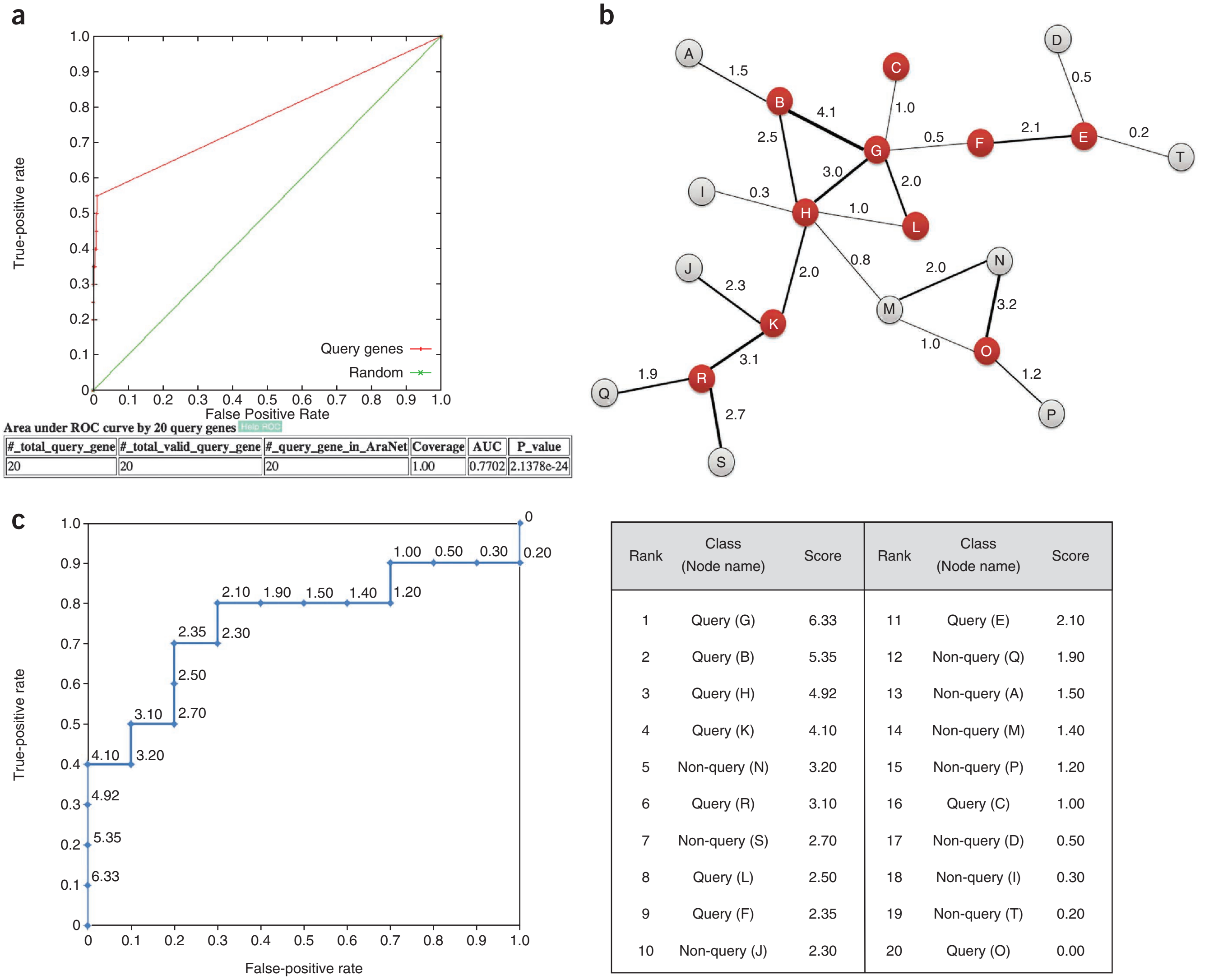
2| Click the 'AraNet Search' link, and four links come up on the page: 'Find new members of a pathway with graph layout', 'Find new members of a pathway with NO graph layout', 'Infer function from network neighbors' and 'Evidence Code'.

3| Select the search type according to the number of query genes. If the number of query genes is greater than 250, choose 'Find new members of a pathway with NO graph layout'. Otherwise, choose 'Find new members of a pathway with graph layout'. Although the latter option has a limit of 250 query genes, it provides useful graphical displays of the networks of query genes along with the predicted candidate genes. The former option does not provide network graphs, but it can search AraNet with up to 2,000 query genes. Most pathways or phenotypes have no more than 250 member genes, but some phenotypes, such as growth defects, may be caused by more than 250 genes.

4| Click on the 'Find new members of a pathway with (NO) graph layout' link, paste your query genes (e.g., the example set of 20 genes for *Arabidopsis* cold acclimation) into the text box and click the 'Submit' button.

5| Obtain a search report. The search report comprises six parts: query genes connected to one another in AraNet; disconnected query gene(s) in AraNet; the area under the ROC curve; network layouts by Cytoscape; new candidate pathway genes (only top 200 predictions are displayed); and gene set analysis using GO, PO and protein domain.

6| Interpret the report of the query genes connected to one another in AraNet. The report table of the query genes connected to one another contains basic information regarding each connected query gene (e.g., locus ID, gene symbol, total connection score to all other query genes and GO annotations, **Fig. 2a**). For all query genes present in AraNet, you can obtain information regarding the AraNet connections among them. If two genes connect to each other in AraNet, they are likely to operate in the same biological process or pathway. Moreover, genes that are connected in AraNet are four times



Area under ROC curve by 20 query genes Help ROC

#_total_query_gene	#_total_valid_query_gene	#_query_gene_in_AraNet	Coverage	AUC	P_value
20	20	20	1.00	0.7702	2.1378e-24

Rank	Class (Node name)	Score	Rank	Class (Node name)	Score
1	Query (G)	6.33	11	Query (E)	2.10
2	Query (B)	5.35	12	Non-query (Q)	1.90
3	Query (H)	4.92	13	Non-query (A)	1.50
4	Query (K)	4.10	14	Non-query (M)	1.40
5	Non-query (N)	3.20	15	Non-query (P)	1.20
6	Query (R)	3.10	16	Query (C)	1.00
7	Non-query (S)	2.70	17	Non-query (D)	0.50
8	Query (L)	2.50	18	Non-query (I)	0.30
9	Query (F)	2.35	19	Non-query (T)	0.20
10	Non-query (J)	2.30	20	Query (O)	0.00

Figure 3 | An example of ROC analysis of predictive power of query genes for the 'Find new members of a pathway' search in AraNet. (**a**) A resultant ROC curve summarizing the predictive power of AraNet for '*Arabidopsis* cold acclimation' with the 20 query genes by AUC score and *P* value. (**b**) A toy example network, in which query genes are represented by red nodes and nonquery genes by gray nodes. The link thickness reflects the log-scaled likelihood of two genes sharing a biological function. (**c**) The resultant curve by ROC analysis of the toy example network. The *x* axis and the *y* axis represent the false-positive rate and true-positive rate, respectively. Area under the curve (AUC) score is 0.75. Scores of network genes (including query genes) having the same function as query genes are calculated by integration of all network connections to query genes with weighted-sum method (**Box 2**). As expected from the high AUC score, the majority of query genes are highly ranked (e.g., all the top four most-likely candidates for query genes are indeed query genes).

more likely to be expressed in the same cell types in the root than expected by chance[16]. Therefore, if the query genes are connected to one another by AraNet, we might expect that they affect the same pathway or trait.

7| (Optional) For the search option, 'Find new members of a pathway', it is possible to improve network-guided prediction by using only connected query genes that may represent core members of the biological process or phenotype. Thus, AraNet provides an alternative option to search for function using only the connected query genes listed below the table (**Fig. 2b**).

8| Interpret the report of the AUC result. After drawing the ROC curves, we calculate the AUC providing a measure of predictability, ranging from ~0.5 for random expectation to 1 for perfect predictions. The diagonal line $y = x$ of the ROC curve for 20 cold-acclimation genes (**Fig. 3a**) represents the performance of the random expectation. **Box 2** explains how to draw ROC curves using a toy example network (**Fig. 3b**). If a classifier makes a random guess, we expect to get correct and incorrect predictions in proportion to their frequency of occurrence. This classification yields the diagonal line in the ROC space, and the area under the diagonal is 0.5. AraNet also reports the *P* values of the measured AUC score using randomized query genes and a normal distribution of the random scores. When the query genes are tightly clustered in the network, all of the query genes will be ranked higher than all the nonquery genes; this would yield an AUC of 1, indicating perfect

BOX 2 | ROC ANALYSIS

In the toy example network shown in **Figure 3b**, query genes are represented by red nodes and nonquery genes by gray nodes. The link thickness reflects the likelihood of a functional association between two genes. The likelihood of two genes interacting functionally (i.e., being involved in the same pathway or process) is measured as a log-likelihood score using Bayesian statistics[2,4] (see **Box 1**). Our network-based classifier calculates the log-likelihood score that each gene has the same function or phenotypic effect as the query genes (red nodes) as the sum of the log likelihood scores of links from the candidate gene to the query genes. The linkage scores from each candidate gene to the set of red query genes are combined using a weighted sum (WS) method[4] using linearly decaying weights for additional link scores (D = 2 for AraNet, found by optimizing performance using recall/precision analysis). The WS score for a gene is calculated as follows:

$$\mathrm{WS} = L_0 + \sum_{i=1}^{n} \frac{L_i}{D \cdot i}$$

where L represents the score for a linkage to the red query genes, L_0 is the maximum link score and i is the rank-order index of the remaining link scores (L). For example, four out of six links to gene H are connected to query genes. Therefore, the score for gene H is calculated as: 3.0 + 2.5 / (2 × 1) + 2.0 / (2 × 2) + 1.0 / (2 × 3) ~ 4.92 (**Fig. 3b**).

Unlike many other classifiers that produce only a class decision for each gene, such as decision trees or rule sets, this classifier produces a quantitative score for each gene[21]. Such a quantitative classifier can be used to produce a discrete decision (i.e., a positive (P) or negative (N) for each gene) with a threshold. If the classifier output is above the threshold, the classifier produces a P; otherwise it produces an N. Each threshold value produces a different point in ROC space. For example, if we set the threshold at 3.20, the top five genes would be classified as P (query genes) and other genes as N (nonquery genes). The four query genes are correctly classified as query genes among the total of ten query genes. Thus, the true-positive rate = 4/10 = 0.4. The one nonquery gene is incorrectly classified as a query gene among a total of ten nonquery genes. Thus, the false-positive rate = 1/10 = 0.1. Therefore, the coordinate of nonquery gene N becomes (0.1, 0.4) in ROC space (**Fig. 3c**).

Note that the toy example ROC curve here may look slightly different from the resultant ROC curves in AraNet reports (for example, see **Fig. 3a**) because this toy example is a special case in which all genes in a genome are modeled in the network (here we assume that there are only 20 genes in the genome), and all of them are connected to query genes except one, ‘O’. In AraNet, the majority of *Arabidopsis* genes are likely to be disconnected from query genes such as ‘O’, and may even be absent from AraNet. Those query genes are randomly ranked because they have no network-based score. Therefore, the majority of *Arabidopsis* genes would be expected to be retrieved randomly, and the ROC curve for the gene set is extrapolated toward (1, 1).

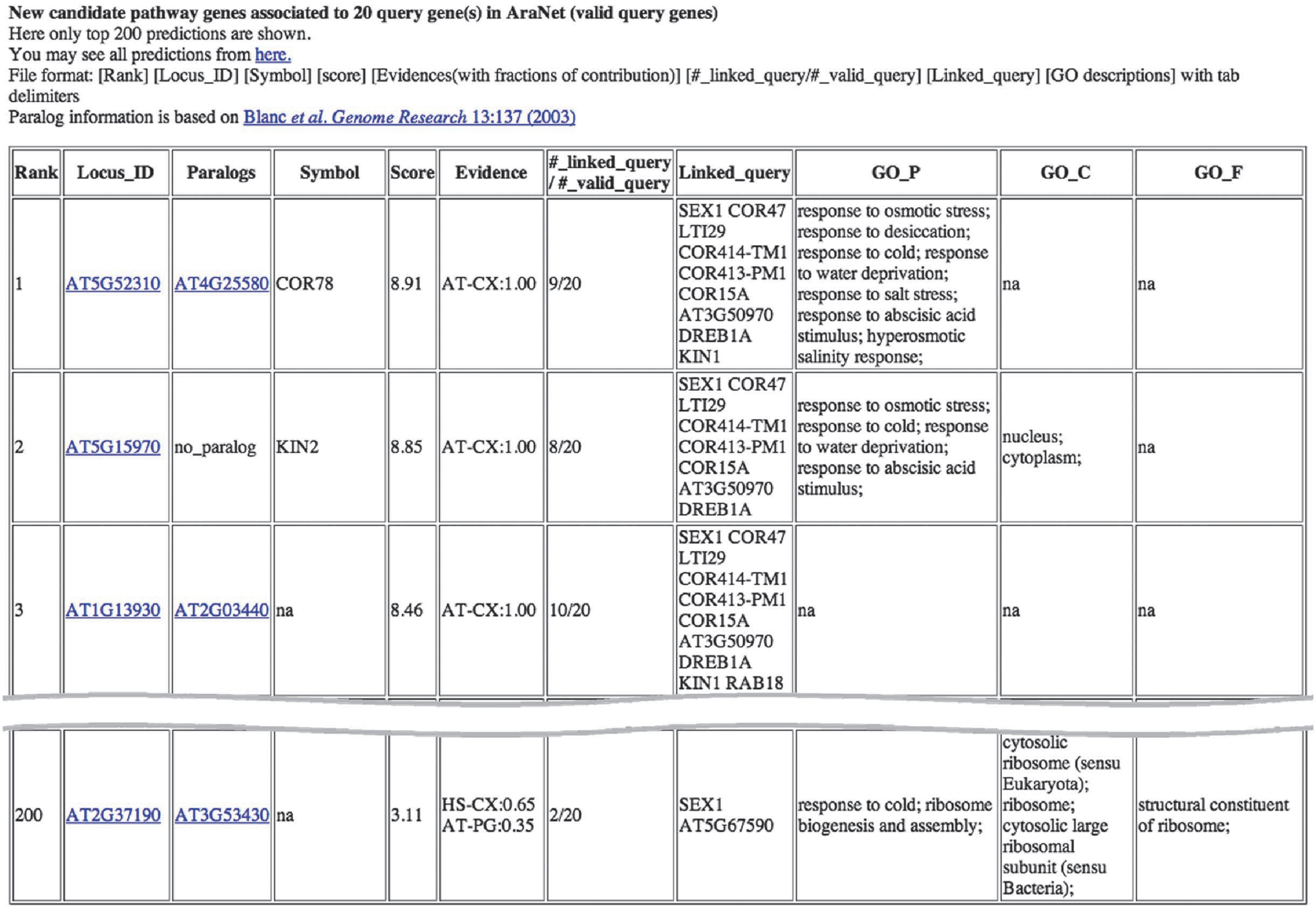
predictions. The AUC score from the 20 cold-acclimation genes is ~0.77, with a P value $<2 \times 10^{-24}$, thus indicating that AraNet is highly predictive for cold-acclimation genes (**Fig. 3a**).

We have previously estimated the proportions of relevant new genes that might be expected to be associated with the query gene set as a function of AUC score. This analysis is intrinsically conservative and estimates ‘new’ discoveries from known and with-held cases, as it is likely that we do not actually yet know the full complement of genes for any *Arabidopsis* biological process. Many known phenotypes currently have only a small number of associated genes, which also serves to make this estimate conservative. Nevertheless, as a lower bound, this approach suggested that, on average, we might expect at least four newly discovered genes among the top 200 for the query gene set with AUC > 0.6, 5 for AUC > 0.7, 6 for AUC > 0.85, 7 for AUC > 0.9, whereas no newly discovered genes are expected by random expectation[16]. The toy example network in **Box 2** (**Fig. 3b**) shows AUC = 0.75 (**Fig. 3c**), indicating that the network has enough predictive power for inferring gene function related to the query gene set.

? TROUBLESHOOTING

9| Interpret the report of the top 200 genes associated with the set of valid query genes by AraNet. A set of valid query genes is the combined set of connected query genes and disconnected query genes. In other words, all query genes that are present in AraNet are valid query genes. The default setting for network-guided prediction by AraNet uses all valid query genes. For example, the AUC score generated by the example cold-acclimation query genes is ~0.77 (**Fig. 3a**), which implies good predictability (see http://www.functionalnet.org/aranet/ROC_help.html), and thus the top 200 genes associated with the set of cold-acclimation genes are highly likely to include new cold-acclimation genes (**Fig. 4**). In fact, the top two predicted new genes, At5g52310 and At5g15970, are already known to respond to cold stress. The third predicted candidate, however, is a completely uncharacterized gene that is linked to almost the same set of query genes as the top two candidates, and thus may represent a novel gene involved in this process.

The table listing the top 200 predicted candidate genes provides seven fields of information for each candidate gene: paralogs; gene symbol; total score of AraNet links supporting the candidate gene; evidence supporting the candidate gene; the fraction of query genes linked to the candidate gene over the total valid query genes; the list of query genes that are



New candidate pathway genes associated to 20 query gene(s) in AraNet (valid query genes)
Here only top 200 predictions are shown.
You may see all predictions from here.
File format: [Rank] [Locus_ID] [Symbol] [score] [Evidences(with fractions of contribution)] [#_linked_query/#_valid_query] [Linked_query] [GO descriptions] with tab delimiters
Paralog information is based on Blanc *et al.* *Genome Research* 13:137 (2003)

Rank	Locus_ID	Paralogs	Symbol	Score	Evidence	#_linked_query / #_valid_query	Linked_query	GO_P	GO_C	GO_F
1	AT5G52310	AT4G25580	COR78	8.91	AT-CX:1.00	9/20	SEX1 COR47 LTI29 COR414-TM1 COR413-PM1 COR15A AT3G50970 DREB1A KIN1	response to osmotic stress; response to desiccation; response to cold; response to water deprivation; response to salt stress; response to abscisic acid stimulus; hyperosmotic salinity response;	na	na
2	AT5G15970	no_paralog	KIN2	8.85	AT-CX:1.00	8/20	SEX1 COR47 LTI29 COR414-TM1 COR413-PM1 COR15A AT3G50970 DREB1A	response to osmotic stress; response to cold; response to water deprivation; response to abscisic acid stimulus;	nucleus; cytoplasm;	na
3	AT1G13930	AT2G03440	na	8.46	AT-CX:1.00	10/20	SEX1 COR47 LTI29 COR414-TM1 COR413-PM1 COR15A AT3G50970 DREB1A KIN1 RAB18	na	na	na
200	AT2G37190	AT3G53430	na	3.11	HS-CX:0.65 AT-PG:0.35	2/20	SEX1 AT5G67590	response to cold; ribosome biogenesis and assembly;	cytosolic ribosome (sensu Eukaryota); ribosome; cytosolic large ribosomal subunit (sensu Bacteria);	structural constituent of ribosome;

Figure 4 | An example list of candidate pathway genes from a 'Find new members of a pathway' search. AraNet analysis returns a table of rank-ordered new candidates for the pathway of the query genes (e.g., a set of genes involved in cold acclimation). AraNet lists only the top 200 candidate genes in the HTML table and provides a list of all the candidate genes as a text file. The list contains information about the rank on the basis of the total connection score to query genes (**Box 2**), locus ID, paralogs, gene symbol, AraNet data types (evidence) supporting connections between the gene and all query genes (**Table 1**), the fraction of connected query genes out of the total valid query genes, all query genes connected to the gene and three GO annotations.

linked to the candidate gene (i.e., query genes supporting candidate prediction); and any known GO annotations for the biological process (GO-P), cellular components (GO-C) and molecular function (GO-F) terms. The 'paralogs' field lists any duplicates of the candidate gene as defined by methods described in Blanc *et al.*[28]. Because the paralogs could retain the same function as the candidate gene, this information should be considered in designing validation experiments by gene perturbation. For example, if we perturb a candidate gene with any paralog, we may detect no phenotypic defect as a result of the functional redundancy from the paralogous gene. Thus, detection of phenotypic effect may require perturbation of not only candidate genes but also paralogs. The candidate genes are ranked by their total score of connections to the set of valid query genes, measured by a weighted sum (**Box 2**) of multiple supportive AraNet links. The field 'Evidence' provides information regarding all supporting data types as evidence codes (**Table 1**) and their relative contributions toward inferring the candidate gene (e.g., HS-CX:0.65 AT-PG:0.35 for the 200th candidate AT2G37190 in the **Fig. 4** table indicates that the coexpression of *Arabidopsis* orthologs of human genes and of phylogenetic profile similarity between *Arabidopsis* homologs provides 65% and 35% of the total supporting evidence, respectively, for being involved in the same process as the query genes).

10| Interpret the reported subnetworks of query genes and their neighboring genes in AraNet. AraNet provides three kinds of networks: a 'query network' (option A), an 'extended network' (option B) and an 'all-prediction network' (Option C) (**Fig. 5a**). The query network consists of only query gene nodes and the links among them. The network intuitively shows how well query genes are connected to one another. The extended network consists of query gene nodes and the top 200 gene nodes associated with the set of valid query genes. It shows how well the top 200 associated gene nodes are connected to the query nodes. The all-prediction network includes the query genes and all the genes connected to them, including the top 200 candidates. This network may be too large to calculate network coordinate values in a reasonable query time. In addition, visual inspection of a network with a large number of nodes and edges is not very practical in general. Therefore,

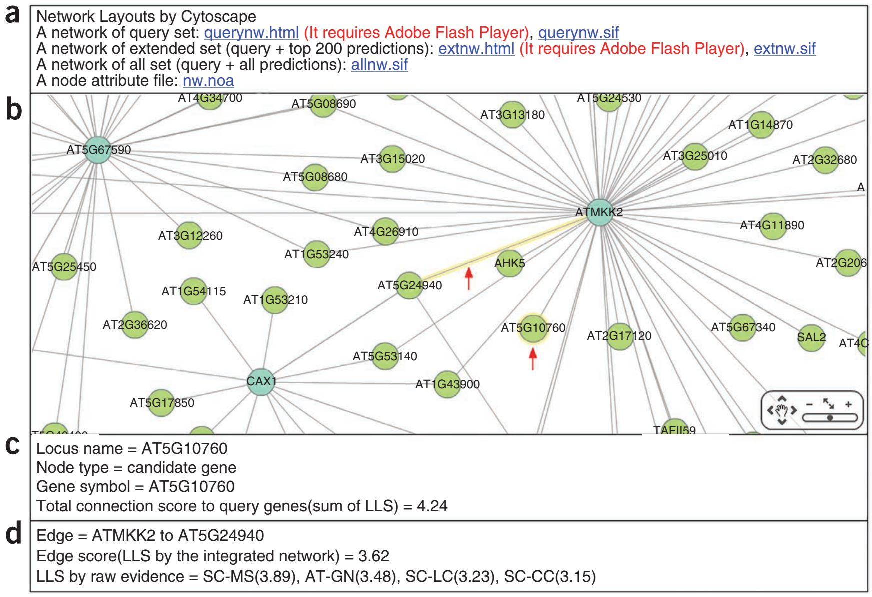
TABLE 1 | Twenty-four types of evidence incorporated into AraNet.

Evidence code	Evidence description
AT-CX	mRNA coexpression between *Arabidopsis* genes
AT-DC	Domain co-occurrence between *Arabidopsis* proteins
AT-GN	Gene neighborhoods between *Arabidopsis* orthologs in bacterial genomes
AT-LC	Literature-curated *Arabidopsis* protein interactions
AT-PG	Phylogenetic profile similarity between *Arabidopsis* homologs
CE-CC	Association between *Arabidopsis* orthologs inferred from co-citation of *C. elegans* genes in Medline abstracts
CE-CX	Association between *Arabidopsis* orthologs inferred from mRNA coexpression in *C. elegans*
CE-GT	Association between *Arabidopsis* orthologs inferred from genetic interactions in *C. elegans*
CE-LC	Association between *Arabidopsis* orthologs inferred from literature-curated *C. elegans* protein interactions
CE-YH	Association between *Arabidopsis* orthologs inferred from *C. elegans* protein interactions by high-throughput yeast two-hybrid analysis
DM-PI	Association between *Arabidopsis* orthologs inferred from *D. melanogaster* protein interactions
HS-CX	Association between *Arabidopsis* orthologs inferred from mRNA coexpression in human
HS-DC	Association between *Arabidopsis* orthologs inferred from domain co-occurrence between human proteins
HS-LC	Association between *Arabidopsis* orthologs inferred from literature-curated human protein interactions
HS-MS	Association between *Arabidopsis* orthologs inferred from human protein interactions by affinity purification/mass spectrometry analysis
HS-YH	Association between *Arabidopsis* orthologs inferred from human protein interactions by high-throughput yeast two-hybrid analysis
SC-CC	Association between *Arabidopsis* orthologs inferred from co-citation of *S. cerevisiae* genes in Medline abstracts
SC-CX	Association between *Arabidopsis* orthologs inferred from mRNA coexpression in *S. cerevisiae*
SC-DC	Association between *Arabidopsis* orthologs inferred from domain co-occurrence between *S. cerevisiae* proteins
SC-GT	Association between *Arabidopsis* orthologs inferred from genetic interactions in *S. cerevisiae*
SC-LC	Association between *Arabidopsis* orthologs inferred from literature-curated *S. cerevisiae* protein interactions
SC-MS	Association between *Arabidopsis* orthologs inferred from *S. cerevisiae* protein interactions by affinity purification/mass spectrometry analysis
SC-TS	Association between *Arabidopsis* orthologs inferred from *S. cerevisiae* protein interactions, which were also inferred from protein tertiary structures
SC-YH	Association between *Arabidopsis* orthologs inferred from *S. cerevisiae* protein interactions by high-throughput yeast two-hybrid analysis

AraNet does not provide the network view for this information but only an edge information file to allow subsequent analysis by additional Cytoscape visualization tools (optional; see Step 11), if necessary. The edge information is downloadable as Cytoscape Simple Interaction File (SIF) format for all three types of networks.

(A) Query network

(i) Click the hyperlink 'querynw.html' to see the visualized query network in a new window. We draw subnetworks using Cytoscape Web (http://cytoscapeweb.cytoscape.org/), which is modeled after the Cytoscape Java network visualization and analysis software[29]. The Adobe flash player plug-in must be installed in your web browser for this network visualization function to work. Cytoscape Web is based on the Flex/Flash technology. The edge information file (querynw.sif) is also available.
? TROUBLESHOOTING



Network Layouts by Cytoscape
A network of query set: querynw.html (It requires Adobe Flash Player), querynw.sif
A network of extended set (query + top 200 predictions): extnw.html (It requires Adobe Flash Player), extnw.sif
A network of all set (query + all predictions): allnw.sif
A node attribute file: nw.noa

Locus name = AT5G10760
Node type = candidate gene
Gene symbol = AT5G10760
Total connection score to query genes(sum of LLS) = 4.24

Edge = ATMKK2 to AT5G24940
Edge score(LLS by the integrated network) = 3.62
LLS by raw evidence = SC-MS(3.89), AT-GN(3.48), SC-LC(3.23), SC-CC(3.15)

Figure 5 | An example of a network layout view page in a new web browser window. (**a**) AraNet analysis provides downloadable network edge information files for additional network visualization and html pages that contain network view generated by Cytoscape Web (http://cytoscapeweb.cytoscape.org/). (**b**) A partial view of the network of genes known to have roles in cold acclimation and their connected genes in AraNet. A blue node represents a query gene of cold acclimation, and a green node represents a connected candidate gene. You can zoom in and out on the image by clicking the plus and minus buttons, respectively. If you click the hand icon, you can move the view window to other parts of the network. (**c**) If you click a node, the lower panel provides detailed information, including the total connection score to query genes. (**d**) If you click an edge, the lower panel provides detailed information including edge score by AraNet and supporting evidences with corresponding log-likelihood scores before weighted sum integration.

(B) Extended network

(i) Click the hyperlink 'extnw.html' to see the visualized extended network in a new window. **Figure 5b–d** shows an example of extnw.html. The network html page comprises two panels: the upper panel (**Fig. 5b**) shows the network layout view, and the lower panel (**Fig. 5c,d**) shows network information.

(ii) Click the plus or the minus button in the bottom right corner of the upper panel to zoom in or out of the network view. If you click a node or an edge, the lower panel shows its detailed information (e.g., information of a node AT5G10760 in **Fig. 5c** and an edge between ATMKK2 and AT5G24940 in **Fig. 5d**). The edge information file (extnw.sif) is also available.

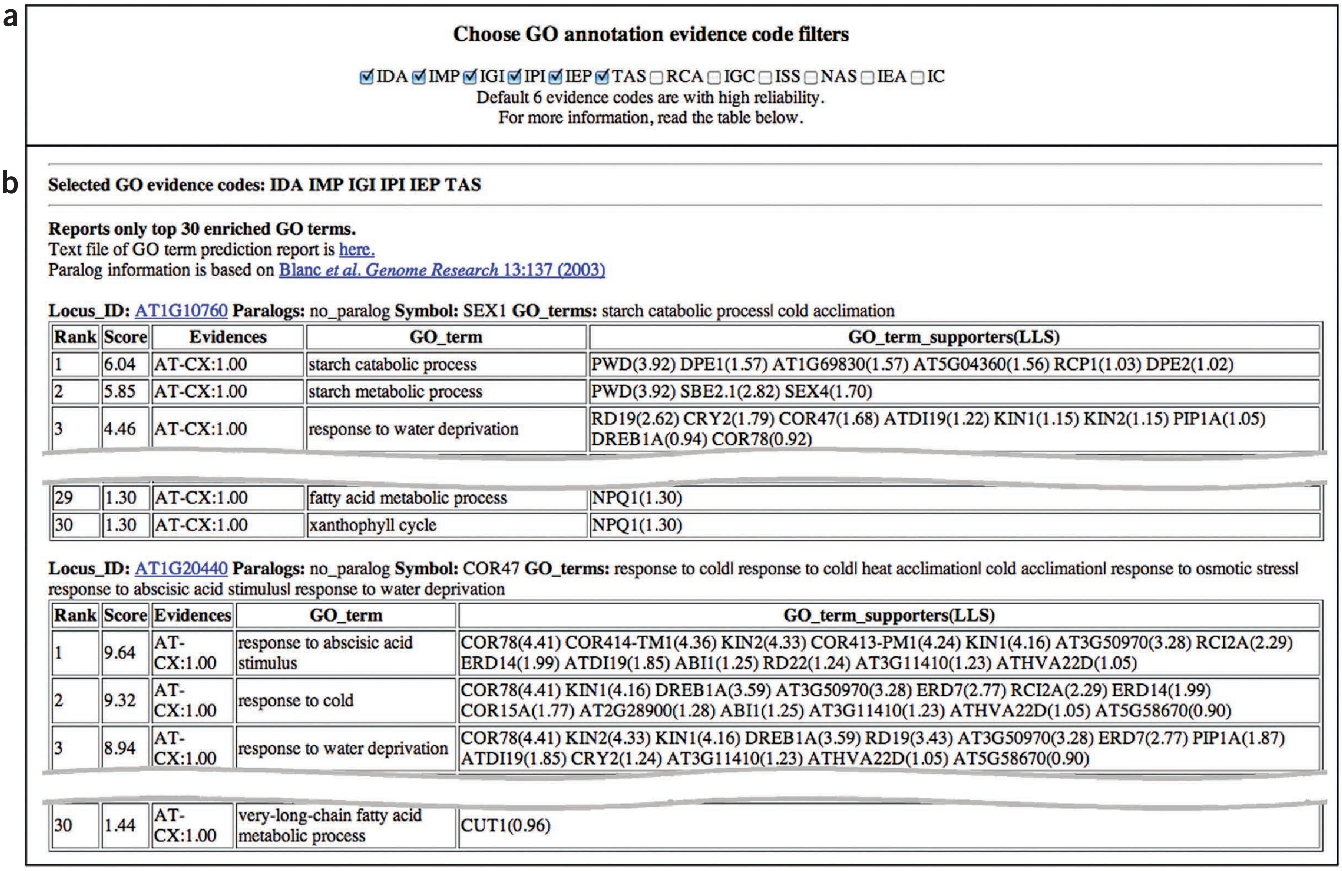
(C) All-prediction network

(i) Download the edge information file (allnw.sif) for optional analyses using local Cytoscape software, if necessary.
▲ CRITICAL STEP Although one would generally provide functionally coherent genes as a query set, it is possible to use a set of genes with functional heterogeneity (e.g., genes with similar loss-of-function phenotypes may comprise multiple pathways). The network visualization function can help identify new genes associated with each functional group, as opposed to global characteristics of all of the query genes. If there are distinct subnetworks among the query genes, querynw.html will show them as separate graphs. New query gene sets can be constructed corresponding to each of the discovered subnetworks, and separate AraNet searches performed with them. This task is not automated in the current version of the AraNet web tool, but requires only minimal manual effort using the network visualization function of AraNet.

11| (Optional) Draw and analyze networks using Cytoscape. AraNet also provides a SIF for each network. To draw and analyze each network using Cytoscape, follow the next three steps: first, download the Cytoscape network SIF file (querynw.sif, extnw.sif, or allnw.sif) and a network node attribute file (nw.noa); next, download Cytoscape software from http://www.cytoscape.org/ and install it on your local computer; and, finally, draw these networks using the local Cytoscape software by importing the downloaded files. For a detailed description of how to install and use Cytoscape, please refer to the published protocol[30].

Infer function from network neighbors

12| Identify candidate GO biological process terms for each query gene. The query procedure is quite similar to that for finding new members of a pathway (Steps 3–6), but the results show the GO predictions of the query genes based on their network neighbors instead of candidate genes that might be involved in the same process. The only difference in the search procedure is the option to choose an appropriate 'search filter' of GO evidence codes (http://www.geneontology.org/GO.evidence.tree.shtml). The default search filter is composed of six evidence codes based on experimental data (IDA, IMP, IGI, IPI, IEP) or authors' statements in the literature (TAS)[31] (**Fig. 6a**). Prediction with the default filter of GO evidence codes uses the neighbor's GO annotations that are supported by the experimental evidence or by authors' statements in the literature. If more evidence codes are selected (i.e., if you also use computationally predicted evidence such as ISS or IEA),



Choose GO annotation evidence code filters

☑IDA ☑IMP ☑IGI ☑IPI ☑IEP ☑TAS □RCA □IGC □ISS □NAS □IEA □IC
Default 6 evidence codes are with high reliability.
For more information, read the table below.

Selected GO evidence codes: IDA IMP IGI IPI IEP TAS

Reports only top 30 enriched GO terms.
Text file of GO term prediction report is here.
Paralog information is based on Blanc *et al. Genome Research* 13:137 (2003)

Locus_ID: AT1G10760 **Paralogs:** no_paralog **Symbol:** SEX1 **GO_terms:** starch catabolic process| cold acclimation

Rank	Score	Evidences	GO_term	GO_term_supporters(LLS)
1	6.04	AT-CX:1.00	starch catabolic process	PWD(3.92) DPE1(1.57) AT1G69830(1.57) AT5G04360(1.56) RCP1(1.03) DPE2(1.02)
2	5.85	AT-CX:1.00	starch metabolic process	PWD(3.92) SBE2.1(2.82) SEX4(1.70)
3	4.46	AT-CX:1.00	response to water deprivation	RD19(2.62) CRY2(1.79) COR47(1.68) ATDI19(1.22) KIN1(1.15) KIN2(1.15) PIP1A(1.05) DREB1A(0.94) COR78(0.92)
29	1.30	AT-CX:1.00	fatty acid metabolic process	NPQ1(1.30)
30	1.30	AT-CX:1.00	xanthophyll cycle	NPQ1(1.30)

Locus_ID: AT1G20440 **Paralogs:** no_paralog **Symbol:** COR47 **GO_terms:** response to cold| response to cold| heat acclimation| cold acclimation| response to osmotic stress| response to abscisic acid stimulus| response to water deprivation

Rank	Score	Evidences	GO_term	GO_term_supporters(LLS)
1	9.64	AT-CX:1.00	response to abscisic acid stimulus	COR78(4.41) COR414-TM1(4.36) KIN2(4.33) COR413-PM1(4.24) KIN1(4.16) AT3G50970(3.28) RCI2A(2.29) ERD14(1.99) ATDI19(1.85) ABI1(1.25) RD22(1.24) AT3G11410(1.23) ATHVA22D(1.05)
2	9.32	AT-CX:1.00	response to cold	COR78(4.41) KIN1(4.16) DREB1A(3.59) AT3G50970(3.28) ERD7(2.77) RCI2A(2.29) ERD14(1.99) COR15A(1.77) AT2G28900(1.28) ABI1(1.25) AT3G11410(1.23) ATHVA22D(1.05) AT5G58670(0.90)
3	8.94	AT-CX:1.00	response to water deprivation	COR78(4.41) KIN2(4.33) KIN1(4.16) DREB1A(3.59) RD19(3.43) AT3G50970(3.28) ERD7(2.77) PIP1A(1.87) ATDI19(1.85) CRY2(1.24) AT3G11410(1.23) ATHVA22D(1.05) AT5G58670(0.90)
30	1.44	AT-CX:1.00	very-long-chain fatty acid metabolic process	CUT1(0.96)

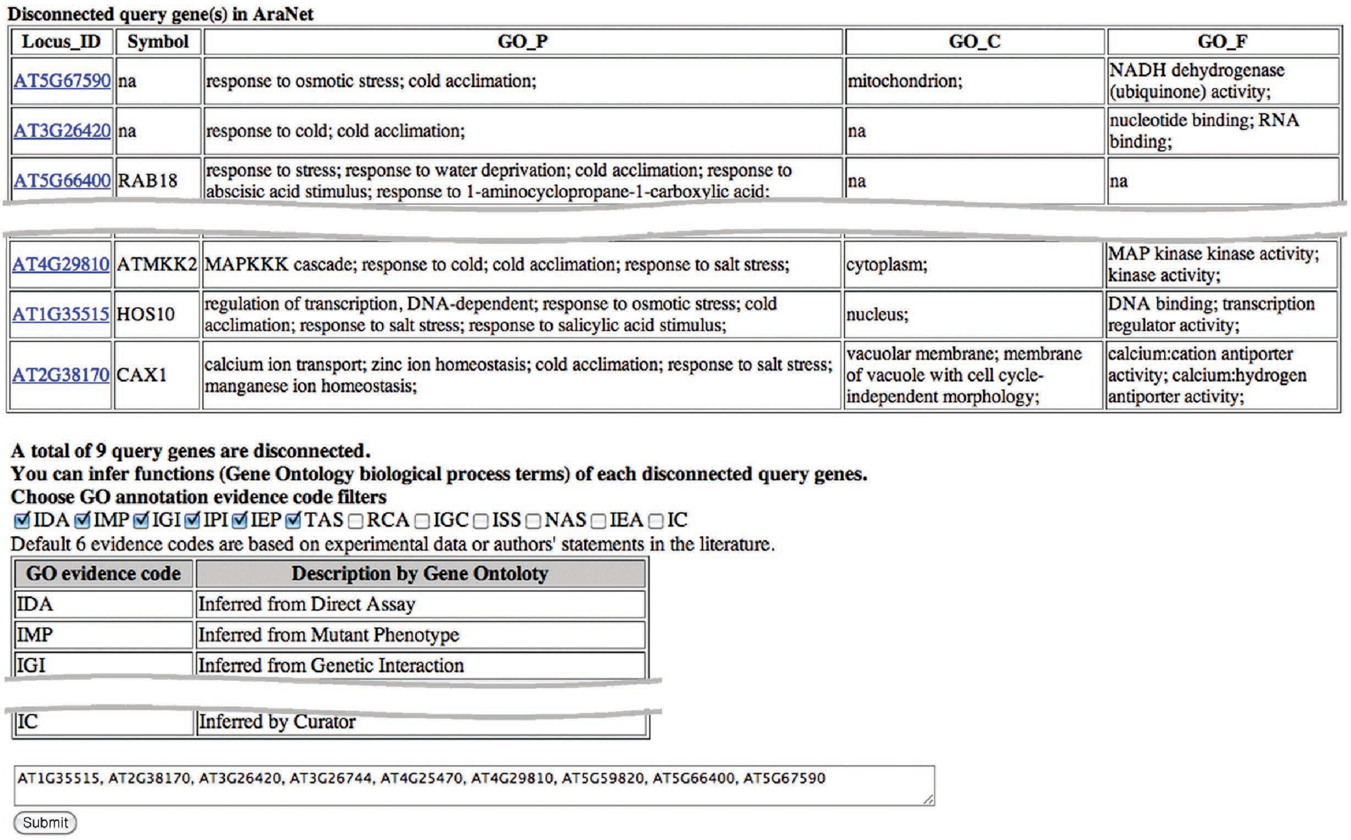
Figure 6 | The 'Infer function from network neighbors' search. (**a**) For this search option, we can filter search results by various supporting evidence codes for GO annotation. The default GO evidence types are limited to experimental data (IDA, IMP, IGI, IPI, IEP) and literature (TAS). It is possible to choose additional evidence codes with less reliability to obtain more prediction results. (**b**) Example reports of predictions of new functions. For each query gene, the report provides candidate GO biological process terms in the prediction table. The table contains five information fields: rank, total score to the neighbors annotated by the candidate GO term, evidence supporting the AraNet connections to the neighbors annotated by the candidate GO term (**Table 1**), a predicted GO term and its GO term–supporting genes connected to this query gene.

AraNet might return more predictions. However, the reliability of the predictions may become lower. If using only the experimentally derived evidence codes returns no predicted GO terms, including additional evidence codes might provide some prediction results. **Figure 6b** shows an example of the results of 'inferring function from network neighbors' search. For each query gene, the search function provides a report table listing the top 30 enriched GO biological process terms that are annotated to the genes that are directly linked to the query gene. The table includes four types of information: the total score for the candidate GO term on the basis of the links to the neighbors annotated with the term (**Box 2**); supporting evidence (**Table 1**), with the fraction of contribution of each; the predicted GO term (biological process); and neighboring genes supporting the predicted GO term, with original LLSs of each network connection indicated in parentheses.

Example queries using two genes with known GO terms as in **Figure 6b** demonstrate the power of functional inference from network neighbors in AraNet. For both genes, one involved in starch catabolic process and the other in cold-stress response, AraNet search returns their known GO terms as top predictions.

▲ CRITICAL STEP A successful validation does not always come from the top-ranked GO term. Other GO terms with relatively high ranks are also highly probable functions for the query gene. We recommend testing or at least examining the top ten candidate GO terms. For example, in our previous work, we successfully validated a gene involved in drought sensing from a GO term that was ranked third and a gene involved in lateral root formation from a GO term that was ranked fifth[16].

13| (Optional) Interpret the AraNet report of the disconnected query genes from the 'Find new members of a pathway' search (**Fig. 7**). Some query genes from the 'Find new members of a pathway' search may not be connected to one another by AraNet or may not even exist in AraNet. For the query genes that are not included in AraNet, the network-guided prediction is not applicable. For disconnected query genes, we provide the known GO[32] annotation information to help reason possible functional hypotheses. However, these genes are often functionally uncharacterized. In this case, we recommend using the



Disconnected query gene(s) in AraNet

Locus_ID	Symbol	GO_P	GO_C	GO_F
AT5G67590	na	response to osmotic stress; cold acclimation;	mitochondrion;	NADH dehydrogenase (ubiquinone) activity;
AT3G26420	na	response to cold; cold acclimation;	na	nucleotide binding; RNA binding;
AT5G66400	RAB18	response to stress; response to water deprivation; cold acclimation; response to abscisic acid stimulus; response to 1-aminocyclopropane-1-carboxylic acid;	na	na
AT4G29810	ATMKK2	MAPKKK cascade; response to cold; cold acclimation; response to salt stress;	cytoplasm;	MAP kinase kinase activity; kinase activity;
AT1G35515	HOS10	regulation of transcription, DNA-dependent; response to osmotic stress; cold acclimation; response to salt stress; response to salicylic acid stimulus;	nucleus;	DNA binding; transcription regulator activity;
AT2G38170	CAX1	calcium ion transport; zinc ion homeostasis; cold acclimation; response to salt stress; manganese ion homeostasis;	vacuolar membrane; membrane of vacuole with cell cycle-independent morphology;	calcium:cation antiporter activity; calcium:hydrogen antiporter activity;

A total of 9 query genes are disconnected.
You can infer functions (Gene Ontology biological process terms) of each disconnected query genes.
Choose GO annotation evidence code filters

☑IDA ☑IMP ☑IGI ☑IPI ☑IEP ☑TAS □RCA □IGC □ISS □NAS □IEA □IC

Default 6 evidence codes are based on experimental data or authors' statements in the literature.

GO evidence code	Description by Gene Ontoloty
IDA	Inferred from Direct Assay
IMP	Inferred from Mutant Phenotype
IGI	Inferred from Genetic Interaction
IC	Inferred by Curator

AT1G35515, AT2G38170, AT3G26420, AT3G26744, AT4G25470, AT4G29810, AT5G59820, AT5G66400, AT5G67590

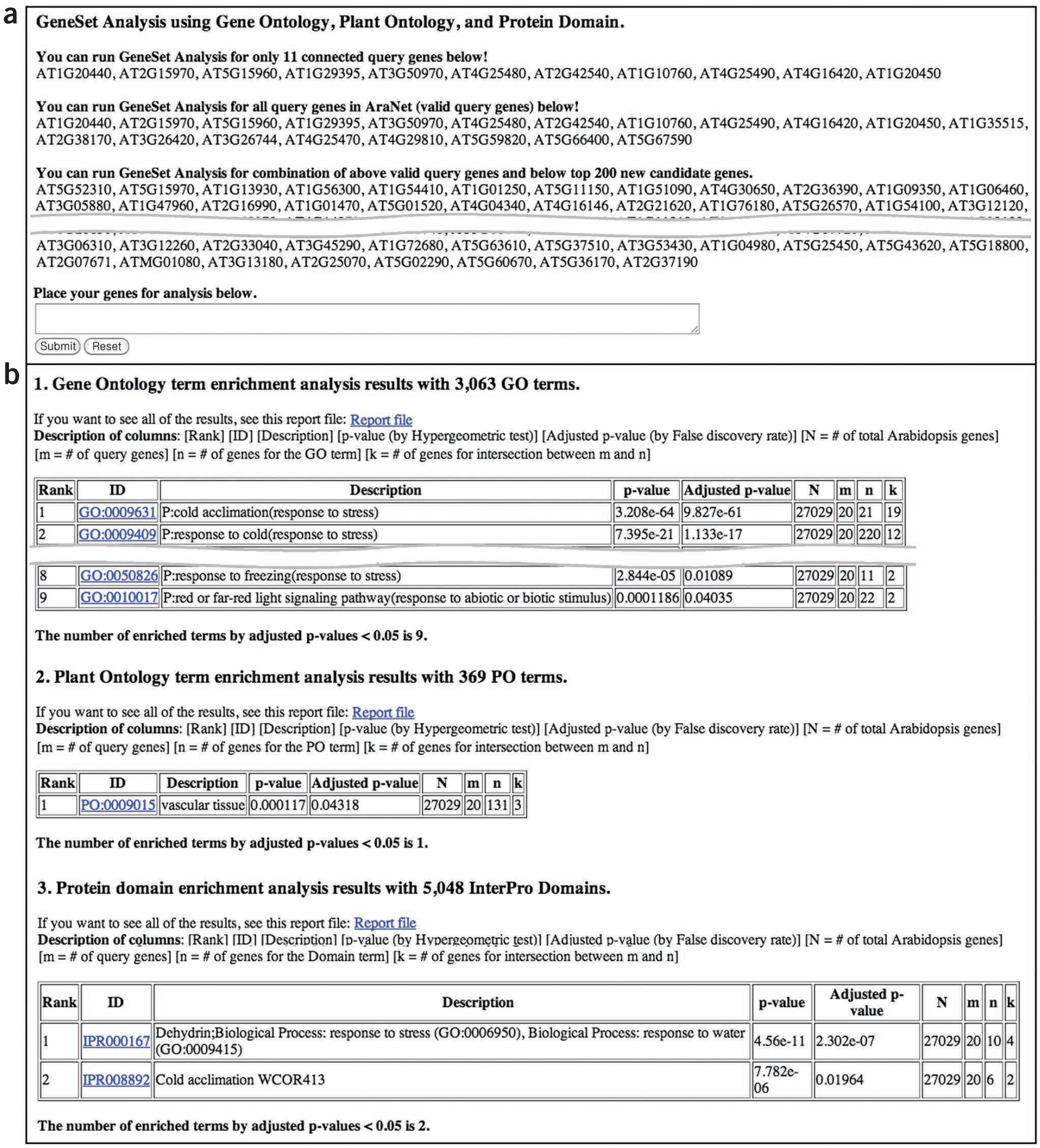
Figure 7 | A report from a 'Find new members of a pathway' search. The report shows an analysis of disconnected query genes in AraNet (e.g., a set of genes involved in cold acclimation). The list of disconnected query genes contains information such as locus ID, gene symbol and three GO annotations. You may directly submit the disconnected genes to an 'Infer function from network neighbors' search by clicking the 'Submit' button.

other search option, 'Infer function from network neighbors'. Inferring function from network neighbors provides candidate GO biological process terms for each query gene based on its neighboring genes' known GO biological process annotations in AraNet. To do this, you only need to click the 'Submit' button below the box containing the disconnected gene list (see Step 12 for interpreting the results).

GO, PO and protein domain enrichment analysis

14| Perform gene set enrichment analysis with the valid query genes. The 'gene set enrichment analysis' query form is shown on the last part of the 'Find new members of a pathway' search report page (**Fig. 8a**). Three types of gene lists for the gene set enrichment analysis are provided by AraNet: connected query genes, valid query genes and the top 200 genes associated with the valid query genes. Copy one of these gene lists into the text box below the gene lists. Click the 'Submit' button to obtain the gene set enrichment analysis results. AraNet simultaneously performs gene set analysis for three reference gene sets—all *Arabidopsis* genes annotated with GO, PO and protein domains—for each submitted query gene set. **Figure 8b** shows an example gene set analysis result for the 20 cold-acclimation query genes. As expected, we obtain the GO term for cold acclimation and other closely related terms (e.g., response to cold and response to freezing). An enriched PO term for vascular tissue suggests this tissue may be important in response to cold stress. We also find enrichment for protein domain family IPR008892, a domain that is contained in several WCOR413-like plant cold-acclimation proteins[33], in addition to the dehydrin domain. The HTML tables display only terms with multiple hypothesis test-adjusted *P* values < 0.05 (a commonly used cutoff for statistical significance), but the complete lists of enriched terms can be downloaded as text files.

15| (Optional) Analyze the gene set enrichment for a subset of the query gene set or a combined set of query genes and new candidate genes. For example, it might be useful to analyze different gene lists, such as valid query genes and top ten associated genes. If you compare the results of this analysis with those from Step 14, you might find new enriched functions with the addition or subtraction of query or candidate genes.



GeneSet Analysis using Gene Ontology, Plant Ontology, and Protein Domain.

You can run GeneSet Analysis for only 11 connected query genes below!
AT1G20440, AT2G15970, AT5G15960, AT1G29395, AT3G50970, AT4G25480, AT2G42540, AT1G10760, AT4G25490, AT4G16420, AT1G20450

You can run GeneSet Analysis for all query genes in AraNet (valid query genes) below!
AT1G20440, AT2G15970, AT5G15960, AT1G29395, AT3G50970, AT4G25480, AT2G42540, AT1G10760, AT4G25490, AT4G16420, AT1G20450, AT1G35515, AT2G38170, AT3G26420, AT3G26744, AT4G25470, AT4G29810, AT5G59820, AT5G66400, AT5G67590

You can run GeneSet Analysis for combination of above valid query genes and below top 200 new candidate genes.
AT5G52310, AT5G15970, AT1G13930, AT1G56300, AT1G54410, AT1G01250, AT5G11150, AT1G51090, AT4G30650, AT2G36390, AT1G09350, AT1G06460, AT3G05880, AT1G47960, AT2G16990, AT1G01470, AT5G01520, AT4G04340, AT4G16146, AT2G21620, AT1G76180, AT5G26570, AT1G54100, AT3G12120,

AT3G06310, AT3G12260, AT2G33040, AT3G45290, AT1G72680, AT5G63610, AT5G37510, AT3G53430, AT1G04980, AT5G25450, AT5G43620, AT5G18800, AT2G07671, ATMG01080, AT3G13180, AT2G25070, AT5G02290, AT5G60670, AT5G36170, AT2G37190

Place your genes for analysis below.

1. Gene Ontology term enrichment analysis results with 3,063 GO terms.

If you want to see all of the results, see this report file: Report file
Description of columns: [Rank] [ID] [Description] [p-value (by Hypergeometric test)] [Adjusted p-value (by False discovery rate)] [N = # of total Arabidopsis genes] [m = # of query genes] [n = # of genes for the GO term] [k = # of genes for intersection between m and n]

Rank	ID	Description	p-value	Adjusted p-value	N	m	n	k
1	GO:0009631	P:cold acclimation(response to stress)	3.208e-64	9.827e-61	27029	20	21	19
2	GO:0009409	P:response to cold(response to stress)	7.395e-21	1.133e-17	27029	20	220	12
8	GO:0050826	P:response to freezing(response to stress)	2.844e-05	0.01089	27029	20	11	2
9	GO:0010017	P:red or far-red light signaling pathway(response to abiotic or biotic stimulus)	0.0001186	0.04035	27029	20	22	2

The number of enriched terms by adjusted p-values < 0.05 is 9.

2. Plant Ontology term enrichment analysis results with 369 PO terms.

If you want to see all of the results, see this report file: Report file
Description of columns: [Rank] [ID] [Description] [p-value (by Hypergeometric test)] [Adjusted p-value (by False discovery rate)] [N = # of total Arabidopsis genes] [m = # of query genes] [n = # of genes for the PO term] [k = # of genes for intersection between m and n]

Rank	ID	Description	p-value	Adjusted p-value	N	m	n	k
1	PO:0009015	vascular tissue	0.000117	0.04318	27029	20	131	3

The number of enriched terms by adjusted p-values < 0.05 is 1.

3. Protein domain enrichment analysis results with 5,048 InterPro Domains.

If you want to see all of the results, see this report file: Report file
Description of columns: [Rank] [ID] [Description] [p-value (by Hypergeometric test)] [Adjusted p-value (by False discovery rate)] [N = # of total Arabidopsis genes] [m = # of query genes] [n = # of genes for the Domain term] [k = # of genes for intersection between m and n]

Rank	ID	Description	p-value	Adjusted p-value	N	m	n	k
1	IPR000167	Dehydrin;Biological Process: response to stress (GO:0006950), Biological Process: response to water (GO:0009415)	4.56e-11	2.302e-07	27029	20	10	4
2	IPR008892	Cold acclimation WCOR413	7.782e-06	0.01964	27029	20	6	2

The number of enriched terms by adjusted p-values < 0.05 is 2.

Figure 8 | Functional gene set enrichment analysis (e.g., a set of 20 query genes involved in cold acclimation). (**a**) Optional function enrichment analysis is available. Using the listed valid query genes, the top 200 new candidate genes, or the combined gene set, AraNet reports the enriched GO, PO and protein domain terms. (**b**) A GO, PO and InterPro protein domain enrichment analysis report. This enrichment analysis tool provides three analysis results for each reference gene set: GO, PO and protein domain. The report table shows the following nine fields of information: rank based on an adjusted *P* value; GO, PO or protein domain ID; a brief description of the ID; hypergeometric *P* value of the ID; adjusted hypergeometric *P* value by false discovery rate; the number of total *Arabidopsis* genes; the number of query genes; the number of genes annotated with the ID; and the number of genes common to both the query genes and the genes annotated with the ID.

Experimental validation

16| Candidate gene functions from the above searches can be assayed by investigating the phenotypes of mutants for the genes. A genome-wide library of T-DNA insertional mutant seeds is available from the *Arabidopsis* Biological Resource Center (http://abrc.osu.edu/)[34]. Similar resources are available for other plants such as rice (http://www.rgrc.dna.affrc.go.jp/index.html.en).
▲ CRITICAL STEP Mutants in many candidate genes may show weak, but detectable, phenotypic effects of the given mutant allele. Thus, closer examination using quantitative analysis, for example, may be required for successful experimental validation. In addition, testing at least two independent alleles for each candidate gene is highly recommended.

Iteration of AraNet searches

17| (Optional) Run the next round of the 'Find new members of a pathway' search with updated query genes, including any newly discovered genes for the same function. Additional query genes should be selected using conservative criteria (e.g., only validated candidate genes from experimental tests) to improve the results of the guilt-by-association approach. Therefore, iterative searching through AraNet may gradually extend the understanding of a given pathway. Additional pathway genes found by iterative searching may improve enrichment scores for the pathway gene set at subsequent steps.

? TROUBLESHOOTING

Low AUC value (Step 8)

If the query genes are not well connected to one another, the AUC value will be low and the prediction power of the query gene set will also be low. There are two major reasons for low AUC scores. First, AraNet may not be a good model for the particular function. Although we observe that the general prediction power of AraNet is excellent, we do not expect AraNet to be highly predictive for every gene function. The 'help ROC' page (http://www.functionalnet.org/aranet/ROC_help.html) shows examples of ROC curves for query genes with good (AUC ≥ 0.7), reasonable (0.7 > AUC ≥ 0.6) and poor (0.6 > AUC) predictabilities. Excessively low AUC values indicate that the known query genes are not well connected to one another within AraNet, and thus the prediction of new genes for this biological process is unlikely to be successful. In this case, another network-based prediction method, Gaussian field label propagation[15], may improve the predictive performance because it takes into account both direct and indirect connections in the network. A general improvement of predictive power using this approach has been observed in yeast and worm[35]. Second, if query genes are involved in a process that is not modulated by one group of functionally coherent genes, the given query genes would not be expected to be well connected to one another in AraNet; the query genes are derived from different functional modules that could be separated as local subnetworks in AraNet. However, the functional association between each gene pair is still predictive for suggesting good candidate genes or functions. Thus, predictions based on disconnected query genes are often still worth examining.

Not seeing network html pages in Linux (Step 10)

As mentioned above, AraNet plots subnetworks using Cytoscape Web, which is based on Flex/Flash technology. If you cannot see the networks (especially on a Linux client), you must try to install the Adobe Flash player plug-in into your web browser.

● TIMING

The search procedure described above takes less than 10 min if you do not execute several of the optional steps. However, for beginners to fully understand, interpret and follow up on the search results and design experiments, more time will be required. Executing only one search option (e.g., 'Find new members of a pathway' or 'Infer function from network neighbors') on the website takes ~1 min, regardless of the number of query genes.

ANTICIPATED RESULTS

'Find new members of a pathway' search results

If you execute the search option 'Find new members of a pathway', you can obtain a report that includes six subsections: (i) query genes connected to one another in AraNet; (ii) disconnected query gene(s) in AraNet; (iii) the area under ROC curve; (iv) network layouts by Cytoscape; (v) new candidate pathway genes (only the top 200 predictions are displayed); and (vi) gene set analysis using GO, PO and protein domain. The first four parts analyze the query genes in four different views. These parts of the results list query genes that are connected to one another (**Fig. 2a**) as well as disconnected query genes (**Fig. 7**) in AraNet. The AUC score allows the user to assess the predictive power of AraNet by using the query gene set to infer new members of a pathway (**Fig. 3**). The network files provide graphical linkage information among the query genes and their neighbors (**Fig. 5**). On the basis of the top 200 genes connected to the valid query genes, you can identify new candidate genes that may have the same function as the query genes (**Fig. 4**). In the last part of the report, you can analyze enriched GO terms, PO terms (growth and structure) and protein domains in both your query genes and the new candidate genes (**Fig. 8**).

'Infer function from network neighbors' search results

If you execute 'Infer function from network neighbors', you can obtain the candidate GO biological process terms predicted by neighbors of each query gene in AraNet. AraNet provides the top 30 enriched GO terms among the neighboring genes of each query gene (**Fig. 6b**).

ACKNOWLEDGMENTS This work was supported by the National Research Foundation of Korea grant funded by the Korean government (no. 20100017649); Pohang Iron and Steel Company (POSCO) TJ Park Science Fellowship (to I.L.); grants from the National Science Foundation, National Institutes of Health, Welch (F1515) and Packard Foundations (to E.M.M.); the Carnegie Institution for Science; and a grant from the National Science Foundation (to S.Y.R.).

AUTHOR CONTRIBUTIONS I.L., E.M.M., S.Y.R. and S.H. conceived the protocol. S.H. constructed the current version of the AraNet web application and wrote the protocol. I.L., E.M.M. and S.Y.R. edited the protocol.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Published online at http://www.natureprotocols.com/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. McGary, K.L., Lee, I. & Marcotte, E.M. Broad network-based predictability of *Saccharomyces cerevisiae* gene loss-of-function phenotypes. *Genome Biol.* **8**, R258 (2007).
2. Lehner, B. & Lee, I. Network-guided genetic screening: building, testing and using gene networks to predict gene function. *Brief Funct. Genomic Proteomic* **7**, 217–227 (2008).
3. Alonso, J.M. *et al.* Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**, 653–657 (2003).
4. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
5. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
6. Fraser, H.B. & Plotkin, J.B. Using protein complexes to predict phenotypic effects of gene mutation. *Genome Biol.* **8**, R252 (2007).
7. Lee, I. *et al.* A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat. Genet.* **40**, 181–188 (2008).
8. Zhong, W. & Sternberg, P.W. Genome-wide prediction of *C. elegans* genetic interactions. *Science* **311**, 1481–1484 (2006).
9. Lee, I. *et al.* Predicting genetic modifier loci using functional gene networks. *Genome Res.* **20**, 1143–1153 (2010).
10. Franke, L. *et al.* Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* **78**, 1011–1025 (2006).
11. Huttenhower, C. *et al.* Exploring the human genome with functional maps. *Genome Res.* **19**, 1093–1106 (2009).
12. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).
13. Linghu, B., Snitkin, E.S., Hu, Z., Xia, Y. & Delisi, C. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol.* **10**, R91 (2009).
14. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011).
15. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* **9** (Suppl 1): S4 (2008).
16. Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M. & Rhee, S.Y. Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nat. Biotechnol.* **28**, 149–156 (2010).
17. Cui, J. *et al.* AtPID: *Arabidopsis thaliana* protein interactome database—an integrative platform for plant systems biology. *Nucleic Acids Res.* **36**, D999–D1008 (2008).
18. Geisler-Lee, J. *et al.* A predicted interactome for *Arabidopsis*. *Plant Physiol.* **145**, 317–329 (2007).
19. 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).
20. Ma, S., Gong, Q. & Bohnert, H.J. An *Arabidopsis* gene network based on the graphical Gaussian model. *Genome Res.* **17**, 1614–1625 (2007).
21. Fawcett, T. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers (Hewlett-Packard Company, 2003).
22. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
23. Berardini, T.Z. *et al.* Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol.* **135**, 745–755 (2004).
24. Avraham, S. *et al.* The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Res.* **36**, D449–D454 (2008).
25. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
26. Rhee, S.Y. *et al.* The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* **31**, 224–228 (2003).
27. 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).
28. Blanc, G., Hokamp, K. & Wolfe, K.H. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* **13**, 137–144 (2003).
29. Lopes, C.T. *et al.* Cytoscape Web: an interactive web-based network browser. *Bioinformatics* **26**, 2347–2348 (2010).
30. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
31. Bard, J.B. & Rhee, S.Y. Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.* **5**, 213–222 (2004).
32. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
33. Danyluk, J., Carpentier, E. & Sarhan, F. Identification and characterization of a low temperature regulated gene encoding an actin-binding protein from wheat. *FEBS Lett.* **389**, 324–327 (1996).
34. Scholl, R. & Anderson, M. *Arabidopsis* Biological Resource Center. *Plant Mol. Bio. Rep.* **12**, 242–244 (1994).
35. Wang, P.I. & Marcotte, E.M. It's the machine that matters: predicting gene function and phenotype from protein networks. *J. Proteomics* **73**, 2277–2289 (2010).