High-quality rice RNA-seq-based co-expression network for predicting gene function and regulation

Hua Yu^{a,b*}, Bingke Jiao^{a,b}, Chengzhi Liang^{a,b*}

^a State Key Laboratory of Plant Genomics, Institute of Genetic and Developmental Biology, Chinese Academy of Sciences

^b University of Chinese Academy of Sciences, Beijing 100039, China

* Corresponding author: Hua Yu, huayu@genetics.ac.cn; Chengzhi Liang, cliang@genetics.ac.cn

Abstract

Inferring the genome-scale gene co-expression network is important for understanding genetic architecture underlying the complex and various biological phenotypes. The recent availability of large-scale RNA-seq sequencing data provides great potential for co-expression network inference. In this study, for the first time, we presented a novel heterogeneous ensemble pipeline integrating three frequently used inference methods, to build a high-quality RNA-seq-based Gene Co-expression Network (GCN) in rice, an important monocot species. The quality of the network obtained by our proposed method was first evaluated and verified with the curated positive and negative gene functional link datasets, which obviously outperformed each single method. Secondly, the powerful capability of this network for associating unknown genes with biological functions and agronomic traits was showed by enrichment analysis and case studies. Particularly, we demonstrated the potential applications of our proposed method to predict the biological roles of long non-coding RNA (lncRNA) and circular RNA (circRNA) genes. Our results provided a valuable data source for selecting candidate genes to further experimental validation during rice genetics research and breeding. To enhance identification of novel genes regulating important biological processes and agronomic traits in rice and other crop species, we released the source code of constructing high-quality RNA-seq-based GCN and rice RNA-seq-based GCN, which can be freely downloaded online at <https://github.com/czllab/NetMiner>.

Key words: Ensemble pipeline, RNA-seq-based GCN, Agronomic traits, lncRNA gene, CircRNA gene

32 Introduction

33 The complex cellular network formed by the interacting macromolecules underlie an
 34 organism's phenotypes (Kitano, 2002a, 2002b; Vidal et al., 2011). Reconstructing a
 35 complete map of the cellular network is crucial for understanding an organism's
 36 genetic architecture underlying phenotypes. In animals, multiple types of networks
 37 have been built based on multi-level '-omics' datasets from genome, transcriptome,
 38 proteome, epigenome, metabolome and other subcellular systems (Mitra et al., 2013).
 39 In plants, most of the current available '-omics' dataset comes from the transcriptome
 40 analysis, with relatively few studies generating other types of '-omics' datasets (Ma et
 41 al., 2013). The rapid accumulation of large-scale open access plant transcriptome data
 42 provides the great potential for identifying the molecular networks underlying diverse
 43 functions. Co-expression meta-analysis is a powerful method for reconstructing gene
 44 co-expression network using transcriptome data. This method combines expression
 45 profiles from all available experimental conditions, aims to predict the statistically
 46 significant functional associations between genes. The extensibility and easiness to
 47 apply make it a powerful tool for inferring the biological roles of uncharacterized
 48 genes (Bergmann et al., 2003; Gerstein et al., 2014; Ma et al., 2013; Mutwil et al.,
 49 2011; Stuart et al., 2003).

50 For co-expression meta-analysis, many algorithms have been proposed to construct
 51 the gene networks. However, it has been shown that the outcome of network inference
 52 varies between tools, and the single network inference approach has inherent biases
 53 and is unable to perform optimally across all experimental datasets (De Smet and
 54 Marchal, 2010; Marbach et al., 2012). In addition, how to clean-up the links occurring
 55 by accident in a gene co-expression network and select biologically significant
 56 associations is also a critical procedure for modeling the authentic gene relations
 57 (Alipanahi and Frey, 2013; Usadel et al., 2009). Moreover, the current computational
 58 methods are mainly designed for analyzing microarray dataset. Indeed, microarrays
 59 are intrinsically limited for measuring a relative small dynamic range of gene
 60 expression and only representing a subset of genomic contents (Abdullah Sayani et al.,
 61 2006; Mutwil et al., 2011). Compared with microarrays, RNA sequencing (RNA-seq)
 62 emerges as a new approach to transcriptome profiling, which provides broader
 63 dynamic range of measurements allowing genome-wide detection of novel, rare and
 64 low-abundance transcripts. However, the majority of co-expression meta-analyses
 65 have been neglected the rapid growing availability of next-generation RNA-seq data
 66 (especially in plants). Its potential capacity in co-expression network inference has not
 67 been well studied.

68 In this study, we designed a novel ensemble pipeline for inferring high-quality Gene
69 Co-expression Network (GCN) using RNA-seq data by integrating the predictions of
70 three different network inference algorithms. **Since the multiple types of networks in**
71 **the model plant, *Arabidopsis*, has been constructed and widely analyzed, we directly**
72 **applied this pipeline to the important crop species, rice, to enhance its efficiency of**
73 **molecular breeding.** We compiled a standard physical and non-physical set of positive
74 and negative functional link datasets between genes derived from 4 known biological
75 networks and evaluated the quality of our network. In the case study, bottom-up
76 subnetwork analysis revealed that the usefulness of reconstructed RNA-seq-based
77 gene co-expression network for realistic biological problems. Particularly, we showed
78 that the potential application of our method for predicting the biological roles of the
79 uncharacterized genome elements including long non-coding RNA (lncRNA) and
80 circular RNA (circRNA) genes. Our study revealed the massive genetic regulatory
81 relationships associating with cellular activities and agronomic traits, which provide a
82 valuable data source for selecting candidate genes to accelerate rice genetics research.

83 Results

84 Network construction and evaluation

85 To evaluate the quality and reliability of publicly available RNA-seq dataset, we
86 analyzed 348 RNA-seq transcriptomes of the important monocot crop species rice
87 after removing the unreliable genes and samples (for details, see Dataset 2, Materials
88 and methods section). After quality filtering and trimming, a total of 12,458,505,209
89 reads were remained in the samples, 75.2% of which were mapped to the MSU7.0
90 reference genome and 71.4% were mapped uniquely (see Dataset 2). Of the genes
91 (MSU7.0 reference set) covered with RNA-Seq reads, 98.4% have coverage of > 50%
92 of the gene length (see Supplementary Information, Fig.S1A). Despite of the large
93 difference in the number of mapped reads between samples, the percentage of
94 expressed genes is similar in most of them, ranging from 32% (10th percentile) to
95 66% (90th percentile), and as the number of mapped reads increases, the ratio of the
96 number of expressed genes is rapidly increased to saturation (see Supplementary
97 Information, Fig.S1B). We tested several normalization methods to compute the
98 expression abundance and expression correlations between genes and samples, the
99 tissue-specific expression pattern and enrichment results of rice genes showed that
100 these RNA-seq data are highly reliable (see Supplementary Text, Fig.S2-Fig.S6, Table
101 S1 and Dataset 3 for details).

102 We comprehensively analyzed whether the co-expression between genes is associated

with their biological roles, and demonstrated that functionally related genes are often to be co-expressed in our RNA-seq dataset (see Supplementary Text, Fig.S7-Fig.S8, Dataset 4 for details). Based on this, we designed a new ensemble pipeline to build RNA-seq-based gene co-expression network by integrating the predictions of three state-of-the-art network inference methods, including Weighted Gene Correlation Network Analysis (WGCNA) (Langfelder and Horvath, 2008), Graphical Gaussian Model (GGM) (Schäfer et al., 2001) and Bagging the Conservative Causal Core of Network (BC3NET) (de Matos Simoes and Emmert-Streib, 2012), based upon an un-weighted voting system and rescoring the co-expression links (see Materials and Methods for details). We here did select these three inference methods but not the other existing approaches is because of either their high computational complexity or the inconsistent data source (Feizi et al., 2013; Friedman et al., 2008; Huynh-Thu et al., 2010; Qin et al., 2014). We constructed the co-expression network of rice which included 16770 genes with 146,419 links. This network shows the small-world characteristic with an average path length between any two nodes is equal to 6.28. The distribution of connection degrees obeys the truncated power-law where most nodes have a few co-expression partners with only a small ratio of hub nodes associating with a large number of partners (see Supplementary Information, Fig.S9A). The negative correlation between degrees and clustering coefficients of genes reveal hierarchical and modular characteristics of network and the possible synergistic regulation of gene expression (Supplementary Information, Fig.S9B) (Bergmann et al., 2003).

We evaluated the performance of the ensemble inference pipeline in rice. Since there are no gold standard reference co-expression networks available in rice, we compiled as replacement a standard set of positive links (9390203 interactions), by capturing gene pairs that were contained in the same Gene Ontology (GO) categories, the same pathways, interact with each other in the protein-protein interaction network or linked in the probabilistic functional gene network (RiceNet), and a standard set of negative links (272997 interactions) based on the functional dissimilarities between genes (for details, see Materials and methods section). We used fold enrichment to measure the relationship of two data sets (our network and standard positive functional links / our network and standard negative functional links): the larger the proportion of the number of shared elements divided by that expected by random chance, the closer they are (see Materials and methods for details). We found that the co-expression relationships connecting highly or frequently expressed gene pairs were positively associated with the positive standard links and were negatively associated with the negative standard links (see Supplementary Information, Fig.S10). Meanwhile, we also observed that the expression sample number of co-expression link (defined as the

total number samples which simply plus the number of gene A expressed samples and the number of gene B expressed samples) is a more reliable factor than its expression level (defined as the expression abundance summation of gene A and gene B) to affect the fold enrichment of the standard links (see Supplementary Information, Fig.S10). These outcomes indicated that the positive standard links had reliably captured the co-expression links between genes. Using the standard datasets, we found that the network structure obtained by our ensemble inference method was consistently better than the networks built by the individual method with higher enrichment for positive links and lower enrichment for negative links (Fig.1). These results suggested that the committee of different methods can reduce the bias occurring in a single inference method and provide more reliable predictions with higher sensitivity and specificity. We observed that the folds of enrichment are not obviously improved or are slightly decreased by the integrated networks from 6 data set (Fig.1A, the GGM method, line highlighted in yellow) than that of each single data set, indicating that integrating the networks built using different data normalization methods might have no obvious effects on the structure of inferred network (Fig.1). Co-expression is actually one of the inputs used to build the probabilistic functional gene network (RiceNet), which were included in the standard positive links. To examine whether this has effect on our evaluation results, we carried out the fold enrichment analysis after removing the links contained in RiceNet from the standard positive links. We found that integrating the functional links of RiceNet into the standard positive links has no effect on the results of comparing the quality of our network with the other networks obtained by the single algorithm (see Supplementary Information, Fig.S11). Based on the novel RNA-seq dataset, we also examined whether a large fraction of potential interactions was recovered by our collected RNA-seq dataset, and found that the most general transcriptional links were already established reliably with these 348 rice RNA-seq samples (see Supplementary Text for details).

Prediction of gene functions through co-expression subnetworks

We observed that our reconstructed RNA-seq-based gene co-expression network is always positive predictor of functional associations for the protein-protein interaction network and probabilistic functional gene network, GO network and pathway network (see Supplementary Text, Fig.S12). Meanwhile, we also observed that many genes under the same GO functional category are significantly more connected to each other than expected by chance (see Supplementary Text, Dataset 5). Therefore, we adopted GO enrichment analysis of a gene's co-expression neighborhood as a tool to predict its biological functions (Vandepoele et al., 2009). For each gene belonging to a given GO category, we asked whether the GO enrichment in its co-expression neighborhood

could infer its correct function: an inference is called true positive if and only if the predicted GO term is more specific than its known GO terms or is equal to the known GO terms. In the enrichment significance level of corrected p -value smaller than 0.05, we found that 15.50% (Sensitivity) annotated functions were correctly inferred based on 10545 annotated genes in rice network. If we used only the gene annotations on the second and third layers of the directed GO graph for inference, the Sensitivity was increased to 21.66%. We found that the 21.27% (Precision) of all inferred functions are true positives and this number is improved to 25.38% when we only adopted the second and third layers of directed GO graph. These results might be suggesting that the incompleteness or errors in the GO annotations of rice genes.

The relatively low Sensitivity and Precision of our network in function inference might be due to the simple scoring metrics. We here further analyzed the predictive performance of our network based on the Critical Assessment of protein Function Annotation (CAFA) metrics (Tzafrir et al., 2003) (see Materials and Methods). To eliminate the effects of the incompleteness and errors of GO annotations, we removed the genes with I) the number of known annotations smaller than 3; II) the number of predicted annotations smaller than 3 and III) the variation coefficient of the number of known annotations and the number of predicted annotations larger than 0.5. To order to produce the Receiver Operating Characteristics (ROC) and Precision-Recall (PR) curves, we calculated the sensitivities, 1-specificities and precisions under different thresholds ($-\log(\text{corrected } q\text{-value})$). For the purpose of correcting different depths of GO predictions, we also calculated the weight value of each GO term and obtained the weighted ROC and PR curves. The weighted ROC and PR curves obtain the larger AUC score (70.01%) and maximum F-measure ($F\text{-max} = 0.54$) than the not weighted ones ($\text{AUC} = 68.23\%$, $F\text{-max} = 0.53$) (see Fig.2), indicating that our gene network can effectively predict the difficult or less frequent GO terms (see Fig.2). In addition, we further compared the predictive performances of our network with RiceNet using the same evaluation criteria as employed in our study. We observed that our co-expression network is comparable or better than the RiceNet in terms of the ROC and PR curves (Fig.2). Moreover, we also found that the semantic similarities between the known GO terms and our predicted GO terms are obviously higher than the random ones ($p\text{-value} = 5.24\text{E-}10$, paired t-test). These results indicated that our RNA-seq-based gene network can be applied for inferring the potential functions of unknown genes.

In addition to the neighboring gene analysis above, we used two examples below to demonstrate the stricter and intuitive method of RNA-seq-based gene co-expression network analysis for inferring the gene functions. In flowering plants, floral organ development is a very important biological process. We therefore first selected a priori

215 guide gene *OsMADS16* involving in flower development to obtain a co-expression
 216 subnetwork consisting of 37 closely connected neighbors within two-layer links from
 217 the guide genes (see Fig.3A and Dataset 6). We found that 15 genes were involved in
 218 flower development process, with ~ 203-fold enrichment. For example, 11 members
 219 of MADS-box family, which were verified involving in the determination of floral
 220 organ identity and development, are effectively captured in this subnetwork. Moreover,
 221 this subnetwork includes the well-known genes *DL*, *Wda1* and *DPW*, which have
 222 been experimentally validated to control the floral organ identity, anther and pollen
 223 development (Jung et al., 2006; Nagasawa et al., 2003; Shi et al., 2011). Interestingly,
 224 we did not find that two YABBY domain containing genes *OsYABBY1* and *OsYABBY6*
 225 are annotated involving in floral organ development in rice, but their *Arabidopsis*
 226 homologs of *YABBY2* and *YABBY1* were associated with the inflorescence meristem
 227 growth and regulation of floral organ development (Siegfried et al., 1999). The
 228 connections between the unannotated genes (gray nodes) and known genes within a
 229 subnetwork provide clues for their associations with specific biological processes. For
 230 example, *LOC_Os07g09020* involves in the reproduction and embryo development,
 231 whose links with *OsMADS3*, *OsMADS4* and *DL* enable further targeted experimental
 232 validations.

233 Second, we used another guide gene *OsCESA4* involving in cell wall metabolism to
 234 build a subnetwork (Fig.3B and Dataset 6). The resulting subnetwork was made up of
 235 139 genes with ~96-fold enrichment, including 4 homologs of *OsCESA4*: *OsCESA1*,
 236 *OsCESA3*, *OsCESA7* and *OsCESA9*, and 14 other genes associated with the cell wall
 237 metabolism. In addition, this subnetwork also captures 28 genes (pink nodes) whose
 238 *Arabidopsis thaliana* homologs were involved in cell wall metabolism. For example,
 239 *LOC_Os01g06580*, encoding a fasciclin domain containing protein, is a homologous
 240 gene to *AT5G03170* which is involved in secondary cell wall biogenesis. Two genes
 241 of *LOC_Os01g62490* and *LOC_Os03g16610* are laccase precursor proteins are both
 242 homologs to *LAC17* involved in cell wall biogenesis. *AT1G09540*, an *Arabidopsis*
 243 homolog of two rice MYB family transcription factors of *LOC_Os05g04820* and
 244 *LOC_Os01g18240*, are participating in cell wall macromolecule metabolism and
 245 xylem development. We also noted that 14 genes labeled with blue nodes, involving
 246 in carbohydrate metabolism, associating with microtubule or resembling to known
 247 cell wall metabolism genes in function domain, are recovered in this gene subnetwork.
 248 All these genes are the potential candidates for the further functional investigation.
 249 Especially, the known cell cycle genes *LOC_Os04g28620* and *LOC_Os04g53760* are
 250 also captured in this subnetwork, confirming that cell wall metabolism and cell cycle
 251 are two closely associated processes.

252 Construction of regulatory subnetworks for gene functional analysis

253 We explored the potential value of motif-guided analysis (Ma et al., 2013) in building
 254 regulatory network and finding functionally related genes using two examples. Cell
 255 cycle is a highly conserved biological process in higher eukaryotes. From G1 phase to
 256 S phase of the cell cycle is controlled by the E2F transcription factors, which bind to a
 257 conserved DNA motif WTTSSCSS (with “W” standing for “A” or “T” and “S”
 258 standing for “C” or “G”) (Vandepoele et al., 2005). We used this motif to retrieve
 259 1093 genes from the rice network. Out of the 180 cell cycle genes annotated in rice
 260 (totally 55986 genes), 33 cell cycle genes were included in these 1093 genes, resulting
 261 in 9.4-fold enrichment. We used the cell cycle genes and the genes that were directly
 262 linked to them to form a regulatory network (totally 104 genes, Fig.4A and Dataset 6).
 263 We observed that a large number of genes (red nodes in Fig.4A) encode proteins
 264 participating in regulation of cell cycle, DNA replication, chromatin dynamics and
 265 DNA repair. The currently known cell cycle genes include three cyclin genes, one E2F
 266 transcription factor, 9 DNA replication origin factors, two checkpoint regulators, 13
 267 DNA replication or repair proteins and 10 other genes with unknown biochemical
 268 functions but were annotated playing important roles during cell cycle. In addition,
 269 this subnetwork also includes 18 genes whose *Arabidopsis* homologs participate in
 270 regulation of cell cycle, DNA replication, DNA repair and chromatin dynamics. Also
 271 recovered are four genes including *LOC_Os01g64900*, *LOC_Os03g49200*, *LOC_Os*
 272 *07g18560* and *LOC_Os09g36900* whose *Arabidopsis* homologs have not annotated
 273 biochemical function but were involved in cell cycle. Although some genes are not
 274 annotated with direct participation of cell cycle, their molecular structure and function
 275 domain indicated their potential roles in it, such as the ribonuclease H2 subunit B
 276 (*LOC_Os04g40050*), ATP-dependent RNA helicase (*LOC_Os11g44910*), ribonuclease
 277 H2 subunit B (*LOC_Os04g40050*) and the BRCA1 C Terminus domain containing
 278 protein (*LOC_Os08g31930*). All these genes are the potential candidate cell cycle
 279 genes for further investigation.

280 WRKY transcription factors play important roles in regulation of plant stress response
 281 by binding the W-box sequence TTGACY (with “Y” standing for “C” or “T”) (Chen
 282 et al., 2012; Rushton et al., 2010). Similarly, we extracted a total of 1329 genes
 283 associating with W-box, from which a subset of 88 known stress response genes out
 284 of 996 genes relating to stress response in rice were found, achieving the fold
 285 enrichment of 3.72. We also constructed a regulatory network using the 88 genes and
 286 the genes with W-box that were directly linked to them (totally 389 genes, Fig.4B and
 287 Dataset 6). This subnetwork includes 172 genes that are regulated by different types
 288 of environmental stresses (red node). Among them, 138 rice genes and 34 homologs

in *Arabidopsis* are annotated in the reference genomes relating to abiotic and biotic stresses. The majority of *Arabidopsis* homologs of these genes are experimentally confirmed involving in the biological regulation of phosphate starvation, water deprivation, nitrate, hypoxia, salt, cold, heat, chitin, sugar and oxidative stresses. Particularly, 53 of 172 abiotic stress response genes whose *Arabidopsis* homologs are reacted to the ethylene (ETH), abscisic acid (ABA), salicylic acid (SA) or jasmonic acid (JA), which is in accordance with the fact that WRKYs play roles in the plant abiotic stress by invoking the ETH-, ABA-, SA- or JA-mediated signaling pathways (Chen et al., 2012). Moreover, 36 genes play important roles in regulating plant immune responses to pathogens including WRKYs, NB-ARC domain containing resistance proteins, NBS-LRR domain containing resistance proteins, kinase proteins and other verified defense members of the plant innate immune system were also contained in this network (see Dataset 6). This is completely supported by the transcriptional reprogramming network model of the WRKY-mediated plant immune responses (Eulgem and Somssich, 2007). In addition, this gene subnetwork also included 8 genes whose *Arabidopsis* homologs are associated with the seed development, dormancy and germination. In agreement with the fact that the SA and ABA antagonizes gibberellin (GA)-promoted seed germination; 6 of these genes participate in the SA- and ABA-mediated signaling pathways (Xie et al., 2007). Interestingly, three genes of *LOC_Os03g12290*, *LOC_Os01g24550* and *LOC_Os01g64470* involving in leaf senescence are also placed in this network, with *LOC_Os01g64470* involving in the SA- and JA-mediated signaling pathway, which is supported by the fact that the WRKYs function in leaf senescence by modulating the JA and SA equilibrium (Miao and Zentgraf, 2007). This subnetwork successfully captured the W-box related genes that can facilitate further studies the functions of uncharacterized genes and help us to understand the regulatory mechanisms of plant responding to various stresses.

In addition, we also used two miRNAs of osa-miR156 and osa-miR396 to capture the functionally related genes based on microRNA target enrichment analysis, which is performed similar with motif enrichment analysis (Ma et al., 2013). We observed that a large number of genes involving in cell division and organ development were captured in this gene subnetwork, for example, two TCP transcription factors of *LOC_Os01g55100* and *LOC_Os11g07460* (see Fig.S13 and Dataset 6). Meanwhile, we also found that many genes relating to stress tolerance were placed in the subnetwork of osa-miR156, for instance, a WRKY transcription factors *LOC_Os10g18099* (see Fig.S13 and Dataset 6). These obtained results well confirm the biological roles of these two miRNAs (Rodriguez et al., 2010; Stief et al., 2014; Wu et al., 2009). Taken together, all these outcomes indicated that the rice RNA-seq-based

327 gene co-expression network could be converted to highly reliable regulatory network
328 for further studying gene regulations.

329 **Co-expression analysis of genes controlling the important agronomic** 330 **traits**

331 For the perspective of system biology, the phenotype of an organism was controlled
332 by functionally linked genes involving in the related biological processes. Given the
333 co-expressed genes tend to have the related biochemical functions; we next want to
334 use the co-expression relationships between genes to assign the agronomic traits for
335 unknown genes. This is especially important for identifying the candidate genes in
336 Quantitative Trait Loci (QTL) mapping, Genome-Wide Association Study (GWAS) or
337 in reverse genetic studies. We collected 1031 known rice genes with the well-studied
338 functions through wet lab experiments. For these genes, we found that 934 genes were
339 expressed in our collected RNA-seq datasets and 623 genes were in network with
340 12125 connections. To examine the potential capacity of our RNA-seq-based gene
341 co-expression network for associating genes with the agronomic traits, we analyzed
342 the density of co-expression links between genes of within and between agronomic
343 traits. We found that 262 co-expression links out of 88041 all possible links within the
344 common agronomic traits and that 252 co-expression links out of 982302 all possible
345 links between the different agronomic traits were captured in network, with ~11-fold
346 enrichment of links within the agronomic traits. In details, we found that several
347 agronomic traits whose genes were tightly clustered together relative to the average
348 link density of whole co-expression network (Supplementary Text, Table S2). For
349 example, an agronomic trait, source activity, measuring the capacity of making
350 photosynthetic products; whose genes was highly aggregated in network with the
351 enrichment fold of 47.81 and the corrected p -value of $3.96E-117$. Besides, genes
352 associating with culm leaf, panicle flower, eating quality and tolerance are also
353 significantly clustered together. Moreover, we performed the permutation test,
354 discovering found that co-expression link densities between genes of same agronomic
355 traits were significantly larger than random control gene set (Supplementary Text,
356 Table S2). These results indicated that our gene networks can be used to discover the
357 gene related to important agronomic traits by co-expression links.

358 **Function discovering for lncRNA genes**

359 Long non-coding RNAs (lncRNAs) have been shown to play important roles in the
360 kingdoms of plants and animals (Ranzani et al., 2015; Zhang et al., 2014). Given that
361 the reconstructed RNA-seq-based co-expression network can successfully associate
362 genes with biological functions and phenotypes of interest, we next wish to discover

the functions for uncharacterized lncRNA genes using network-based method. We downloaded the known lncRNAs of rice identified in previous studies (Zhang et al., 2014). We then combined these lncRNA genes with MSU7.0 reference genes to establish co-expression network based on the ensemble inference pipeline. The obtained network is composed of 24875 genes, containing 24014 protein-coding gene and 861 lncRNA genes connected by 1357039 edges. Compared with the previous protein-coding gene network, 7692 novel protein-coding genes were captured and linked with 817 lncRNA genes. As there is no gold standard available to evaluate the predictive performance, we adopted gene-guide subnetwork analysis to illustrate the potential capacity of this network for lncRNA function discovering. We selected a well-studied lncRNA gene of *XLOC_057324*, which was verified involving in panicle development and fertility, to establish a gene subnetwork consisting of the two-step co-expression neighborhoods (Fig.5 and Dataset 7). In this subnetwork, 4 genes including *SSD1*, *PLA1*, *DEP1* and *GSD1* related to panicle development or fertility. In addition, we also found that 7 genes (pink nodes) whose *Arabidopsis* homologs participate in meiosis, embryo development or reproductive process. According to the functional annotation, some genes (blue nodes) might be also involved in pollen development, such as two cyclin genes *CYCA2* and *CYCD2*. Interestingly, 3 lncRNAs of *XLOC_061753*, *XLOC_006119* and *XLOC_031878* expressed in the reproductive organs are contained in this subnetwork. These results are in good agreement with the experimentally verified role of *XLOC_057324*.

CircRNA gene identification and function analysis

CircRNA is an RNA molecule forming a covalently closed continuous loop that has been discovered in various species across the domains of life with distinct sizes (Memczak et al., 2013; Ye et al., 2015). The functions of circRNAs are largely unknown and hard to investigate. Therefore, we try to classify them through gene co-expression network. We first identified 14325 circRNAs in rice derived from 5284 genes including 4609 protein-coding genes, 675 noncoding genes (see Materials and Methods for details). 43 of these genes including 27 protein-coding genes and 16 non-coding genes produce the circRNAs with the percentage larger than 90% in at least one sample. We analyzed the distribution of the number of detected circRNAs and found that a majority of circRNAs were identified in one sample with relative small number of circRNAs were detected in more than 3 samples (Fig.S14A). Though a large number of circRNAs were detected in relative small number of RNA-seq samples, 63 circRNAs (transcribed from the protein-coding genes), identified in more than 10 samples and supported by more than 26 junction reads, were captured in the gene co-expression network. Moreover, we found that the primary genes transcribing

these circRNAs were not contained in the co-expression network. We predicted the functions of these circRNAs using GO enrichment analysis of their co-expression neighborhoods. Indeed, these circRNAs are related to a broad range of biological functions, for example protein phosphorylation, ATP binding and photosynthesis (Fig.S14B). These results indicated that a great number of circRNAs play important biological roles but not are the transcriptional noise.

Discussion

The phenotypes of an organism are determined by the coordinated activity of many genes and gene products. To gain insight into the genetic foundation underlying the complex biological processes and phenotypes, we developed a novel analytic pipeline for constructing high-quality RNA-seq-based co-expression network and predicting gene function and regulations. we applied this pipeline to the important crop species rice. The obtained co-expression links between genes were ranked by confidence score, expression level and expression sample number. The thresholds of these measures can be selected as the indicators of co-expression reliability for the further targeted experimental validation. The detailed analysis of the topology properties of network demonstrates that the degree frequency distribution follows the truncated power-law and network structure is highly modular. Using the rice gold standards and bottom-up co-expression subnetwork analysis, we showed that this analysis pipeline can be effectively applied to study the gene function and regulation. Particularly, the potential application value of RNA-seq gene network for predicting biological roles of lncRNA and circRNA genes are well demonstrated. Overall, our analysis provides new insights into the regulatory code underlying transcription control and a starting point for understanding the complex regulatory system.

Compared with the sequence-based functional annotation, a great advantage of gene co-expression-based inference approach is that homologs are not required for a gene to receive a prediction. Actually, it is the case when a novel function appears for a particular species and the genes participating in the new biological process do not have corresponding homologues in other species. This is especially interesting for the non-coding RNAs because only short regions of non-coding RNA transcripts are limited by sequence- or structure-specific interactions, compared to the protein-coding gene; this difference in selection pressure makes it very difficult to find orthologous non-coding RNAs by their sequences. Indeed, using the BLAST search against NCBI Reference Sequence Database (RSD), we found that 87% and 89% of unannotated genes and lncRNA genes do not have homologous genes in other species, respectively. The functional analysis of rice lncRNA gene of *XLOC_057324* suggested that our

436 RNA-seq-based gene network can be effectively applied to annotate the functions of
437 non-coding genome elements.

438 For RNA-seq-based gene co-expression network investigators, the creation of novel
439 computational methods for building high-quality network poses a future fundamental
440 challenge. According to our best knowledge, only four existing methods including
441 Pearson's Correlation Coefficient (PCCs), WGCNA, Canonical Correlation Analysis
442 (CCA) and SpliceNet have been used to establish the RNA-seq gene co-expression
443 networks (Giorgi et al., 2013; Hong et al., 2013; Iancu et al., 2012; Yalamanchili et al.,
444 2014). Moreover, some of these inference tools are unable to be applied to the
445 large-scale expression dataset owing to their high computational complexity. For the
446 uncertainty and complexity of mechanism models underlying the RNA-seq data, we
447 designed a novel ensemble-based inference pipeline to establish the high confidence
448 RNA-seq gene co-expression network. Our outcomes demonstrate that the committee
449 of three inference methods provides more robust and less false positive and false
450 negative results than single algorithm. The improved performance of our ensemble
451 inference method depends on the voting and rescoring scheme which can reduce the
452 bias occurring in a single learning method and assign a higher confidence to the
453 interactions that are repeatedly retrieved by different methods. Indeed, the standpoint
454 of aggregating the results of different algorithms has been adopted in various contexts
455 and it has proven to be effective in a variety of applications (Lertampaiporn et al.,
456 2013; Liu et al., 2007; Yang et al., 2010).

457 In principle, gene co-expression meta-analysis can only detect co-regulations between
458 genes which are co-expressed constantly or are sometimes co-expressed but otherwise
459 silent. However, many activation patterns of gene groups appear only under the
460 specific experimental conditions but behave independently under the other conditions,
461 which might not be captured by our method. Especially, for lncRNA and circRNA
462 genes, their expression patterns demonstrated highly spatiotemporal specificity. To
463 overcome this problem, the high-efficiency bi-clustering methods can be integrated
464 into our model to reveal the transcriptional gene interactions presented only under a
465 specific subset of the experimental conditions (Madeira and Oliveira, 2004). Our
466 approach can further improved by I) expanding our ensemble pipeline with other
467 high-efficiency inference methods (Hase et al., 2013), II) employing more reasonable
468 voting and rescoring schemes to generate the consensus networks.

469 Materials and methods

470 Dataset preprocessing

471 We downloaded 456 rice primary RNA-seq samples from the NCBI Sequence Read
472 Archive (see Dataset 1 and 2 for details), with the keywords of “*Oryza sativa*”
473 [Organism] AND “platform illumina” [Properties] AND “strategy rna seq” [Properties]
474 (accessed on May 29, 2014). These RNA-seq samples contained a wide spread of
475 experimental conditions, tissue types and developmental stages. After the SRA files
476 were gathered, the archives were extracted and saved in FASTQ format using the SRA
477 Toolkit. The FASTQ files were firstly trimmed using Trimmomatic software (version
478 0.32) (Bolger et al., 2014) with the default settings, except for an additional parameter
479 of minimum read length at least 70% of original size. Then, the fastq_quality_filter
480 program included in FASTX Toolkit was adopted to further filtrate the FASTQ files,
481 with the minimum quality score 10 and minimum percent of 50% bases that have a
482 quality score larger than this cutoff value. Surviving RNA-seq samples were mapped
483 to the MSU7.0 reference genomes (55986 genes) using TopHat v2.0.4 with the default
484 settings except for “--max-multihits 1” (Trapnell et al., 2009). The PCR and
485 optical/sequencing-driven duplicate reads were removed using the Picard tools. After
486 reads mapping, the uniquely aligned reads count (RAW) and Fragments Per Kilobase
487 Of Exon Per Million Fragments Mapped (FPKM) of each gene was calculated relative
488 to the reference gene model using the HTSeq-count (v0.5.4) and Cufflinks software
489 (v2.1.1), respectively (Anders et al., 2014; Trapnell et al., 2012). The unreliable
490 samples and genes were filtered according to the following three criteria: I) The
491 samples, in which the percentage of the number of genes with expression value
492 smaller than 10 reads is larger than 90%, were not considered for further analysis; II)
493 We did not consider the genes whose expression value is less than 10 reads in more
494 than 80% samples; III) Genes with the variation coefficient of expression values
495 smaller than 0.5 were excluded from subsequent analysis. After filtering, we got two
496 expression datasets composed of 348 RNA-seq samples and 24775 genes were. The
497 filtered RAW dataset were further corrected using four normalization methods: I)
498 Upper Quartile (UQ) (Robinson et al., 2010); II) Trimmed Mean of M values (TMM)
499 (Robinson et al., 2010); III) Relative Log Expression (RLE) (Robinson et al., 2010)
500 and IV) Variance Stabilizing Transformation (VST) (Anders and Huber, 2010).

501 The microarray gene expression data were extracted from both ATTED-II database
502 and Rice Oligonucleotide Array Database (ROAD) (Cao et al., 2012; Obayashi et al.,
503 2009). The Gene Ontologies (GOs) were downloaded from Plant GeneSet Enrichment
504 Analysis Toolkit (PlantGSEA) (Yi et al., 2013). We downloaded biological pathways
505 from two data sources including PlantGSEA database and Plant Metabolic Network
506 (PMN) (<http://pmn.plantcyc.org/>). The gene sets of transcription factor family were
507 downloaded from Plant Transcription Factor Database (PlantTFDB) (Jin et al., 2013).
508 MicroRNAs and their related targets were collected from the Plant MicroRNA Target

509 Expression database (PMTED) and Plant MicroRNA database (PMRD) (Zhang et al.,
510 2010). Known agronomic trait genes were collected from both Q-TARO database
511 (Yonemaru et al., 2010) and literatures. Tos17 mutant phenotypes were extracted from
512 Rice Tos17 Insertion Mutant Database (Hirochika et al., 1996). The phenotypes were
513 associated with MSU7.0 gene locus identifiers through BLASTN alignments of Tos17
514 flanking sequences obtained from NCBI website. Protein-protein interaction network
515 of rice were downloaded from PRIN (Gu et al., 2011). Probabilistic functional gene
516 network of rice was obtained from RiceNet data portal (Lee et al., 2011).

517 **Gene co-expression network construction**

518 We developed an ensemble-based inference pipeline for constructing the high-quality
519 RNA-seq-based Gene Co-expression Network (GCN) based upon combining multiple
520 inference algorithms, then aggregating their predictions through an unweighted voting
521 system and rescoring co-expression links. Our ensemble-based inference system was
522 designed based on the hypothesis that the different network inference methods have
523 complementary advantages and limitations under the different contexts. To select base
524 inference methods for constructing an ensemble system, five algorithms were initially
525 tested and evaluated, including the weighted gene co-expression network analysis
526 (Langfelder and Horvath, 2008), graphical Gaussian model (Schäfer et al., 2001),
527 bagging statistical network inference (de Matos Simoes and Emmert-Streib, 2012),
528 graphical lasso model (Friedman et al., 2008) and tree-based method (Huynh Thu et
529 al., 2010). Since graphical lasso and tree-based method have high computational
530 complexity and are infeasible for large number of RNA-seq dataset, we did not adopt
531 these two algorithms for subsequent network construction. The flowchart for building
532 high confidence RNA-seq-based gene co-expression network was depicted in Fig.6.
533 In details, our procedure for producing the high-quality gene co-expression network
534 was started from 6 RNA-seq datasets as described in Dataset preprocessing. Based on
535 the 6 RNA-seq expression datasets, the weighted co-expression network inference,
536 graphical Gaussian model and bagging statistical network inference were adopted to
537 obtain 18 initial gene co-expression networks using the R packages of WGCNA,
538 GeneNet and BC3NET, respectively (available from the CRAN repository). Since the
539 outputs of WGCNA and GeneNet produced the long ordered list of confidence scores
540 (topological overlap for WGCNA and partial correlation coefficient for GeneNet) for
541 an enormous amount of gene pairs, we designed a random permutation model to
542 choose the restrict threshold that roughly identifies functional co-expression links. We
543 repeatedly created 100 times random datasets to obtain a series of background
544 distributions, by randomly shuffling the associations from genes to expression profiles,
545 and used the average of 99.99th percentile of these distributions (corresponding to the

546 probability of 10^{-4} that two genes are connected by chance) to define the threshold.
 547 After obtaining initial networks, we employed two-step voting procedure, including
 548 voting within inference method and voting among the inference methods, to construct
 549 the high-quality gene co-expression network. In the first step of voting procedure, we
 550 selected the links included in more than two networks of all 6 initial co-expression
 551 networks, which were built by applying the single network inference algorithm to 6
 552 RNA-seq datasets, to establish a consensus gene network (i.e. intra-method consensus
 553 network). In second step of voting procedure, we pick the co-expression relationships
 554 contained in more than one network of three intra-method consensus networks to
 555 establish the final co-expression network.

556 The confidence score calculation procedure for each gene pair of the final RNA-seq
 557 gene co-expression network was performed as following: I) Firstly, we normalized the
 558 confidence scores of each co-expression link of each initial network to the interval
 559 range from 0 to 1. II) Then, we assigned a confidence score to each association of the
 560 intra-method consensus gene networks by averaging the normalized confidence scores
 561 of all 6 initial networks. III) Finally, we defined the confidence score for each edge of
 562 final high confidence co-expression network by averaging the confidence scores of
 563 three intra-method consensus gene networks. Note that for the co-expression links not
 564 listed in a co-expression network were assigned a confidence score of 0.

565 **Performance evaluation**

566 As the information about gold standard *Oryza sativa* reference gene network is
 567 unavailable, we compiled as replacement a standard set of positive and negative links
 568 for the performance evaluation. The gold standard of positive functional links was
 569 obtained by capturing gene pairs that were contained in the same GO categories, the
 570 same pathways, interact with each other in protein-protein interaction network or
 571 linked in probabilistic functional gene network. To construct the gold standard of
 572 negative functional links, we firstly selected all the biologically unrelated GO pairs
 573 (semantic similarity score = 0) that have the number of genes greater than 5 and less
 574 than 50, coupling all possible gene pairs of each partnership in remainder GO terms as
 575 initial non-functional relationships. Subsequently, we established 10000 background
 576 distributions of functional similarity, by 10000 times randomly sampling of 1000 gene
 577 pairs and calculating the functional similarities. We selected a subset of gene pairs
 578 from the initial non-functional links as final non-functional links using the criterion
 579 that the functional similarity between gene pair that are smaller than the average of
 580 5th percentiles of these simulated background distributions. The semantic similarities
 581 between the GO terms were calculated using the R package of GOSim (Fröhlich et al.,

2007). Functional similarities between genes in terms of the GO space were calculated using the metric adopted from (Chabalier et al., 2007).

Since our gold standards included only a subset of true functional and non-functional link, we evaluated the predictive performance of our method for gene co-expression network inference using the fold enrichment measure. The fold of enrichment was calculated as a function of the confidence score cutoff (k) in the edge list of the inferred network by the following formula:

$$\frac{n_k}{m_k} \times \frac{M}{N} \quad (1),$$

where, n_k is the number of true positive or true negative functional links in the k th cutoff of the edge list; m_k is the number of edges of the inferred network in the k th cutoff; M denotes the number of true positive or true negative functional links in the gold standards and N represents the number of all possible interactions in the genome space. The network visualization was carried out using both Cytoscape (Cline et al., 2007) and BioLayout Express3D (Theocharidis et al., 2009).

The function enrichment of co-expression neighborhoods was calculated as the ratio of the relative occurrence in gene set of co-expression neighborhoods to the relative occurrence in genome using Fisher's exact test. The p -value was further adjusted by Benjamini-Hochberg correction for multiple hypotheses testing. The corrected p -value smaller than 0.05, was considered as enriched. To evaluate the predictive performance of our RNA-seq-based network for inferring gene function using the co-expression neighborhoods, we adopted the gene-centric evaluation, which were provided in the Critical Assessment of protein Function Annotation (CAFA) project (Tzafrir et al., 2003). For this metric, the GO terms of each gene (gold and predicted) are propagated up the GO hierarchy to the root, obtaining a set of terms. In this process, for each scored GO term, we propagated its score ($-\log(q\text{-value})$ of Fisher's exact test) toward the root of the ontology such that each parent term received the highest score among its children. The Sensitivity (Recall), 1-specificity, Precision and maximum F-measure (F-max) was calculated using the same method as in the CAFA project. The Receiver Operating Characteristics (ROC) curve was drawn by changing the threshold and plotting the Sensitivity versus the 1-specificity and then calculated the score of Area Under Curve. Similarly, we plotted the Precision-Recall (PR) curve by altering the threshold and plotting the Precision versus the Recall. Semantic similarity scores between the GO term pairs were calculated using the R package of GOSim.

Analysis of circRNA genes

616 The circular RNA (circRNA) genes were predicted using 618 novel rice RNA-seq
617 samples downloaded from the NCBI Sequence Read Archive (accessed on February
618 15, 2016) by CIRI software (Gao et al., 2015). We calculated the counts of junction
619 reads of a circRNA as its relative expression abundance. Then, we integrated the
620 aligned reads number of known rice genes using HTSeq-count program (v0.5.4) and
621 expression values of circRNAs into a numeric expression matrix. We removed the
622 circRNAs from the matrix if it was identified in less than 3 RNA-seq samples. Using
623 the filtered matrix, we built three initial gene co-expression networks by WGCNA,
624 GGM and BC3NET. Based on this, we selected the co-expression links contained in
625 more than one network of the three initial networks to obtain the final co-expression
626 network. Although only the numbers of junction reads were adopted to measure the
627 expression abundances of circRNAs, this method is simple and effective for building
628 co-expression network, given the reads were distributed uniformly along circRNA.

629 References

- 630 Abdullah Sayani, A., Bueno de Mesquita, J.M., and van de Vijver, M.J. (2006). Technology Insight: tuning into the
631 genetic orchestra using microarrays-limitations of DNA microarrays in clinical practice. *Nat. Clin. Pract. Oncol.* 3,
632 501-516.
- 633 Alipanahi, B., and Frey, B.J. (2013). Network cleanup. *Nat. Biotechnol.* 31, 714-715.
- 634 Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11,
635 R106.
- 636 Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq-A Python framework to work with high-throughput sequencing
637 data. *Bioinformatics* 31, 166-169.
- 638 Bergmann, S., Ihmels, J., and Barkai, N. (2003). Similarities and differences in genome-wide expression data of
639 six organisms. *PLoS Biol.* 2, e9.
- 640 Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data.
641 *Bioinformatics* 2114-2120.
- 642 Cao, P., Jung, K.H., Choi, D., Hwang, D., Zhu, J., and Ronald, P.C. (2012). The rice oligonucleotide array
643 database: an atlas of rice gene expression. *Rice* 5, 1-9.
- 644 Chabalier, J., Mosser, J., and Burgun, A. (2007). A transversal approach to predict gene product networks from
645 ontology-based similarity. *BMC Bioinf.* 8, 235.
- 646 Chen, L., Song, Y., Li, S., Zhang, L., Zou, C., and Yu, D. (2012). The role of WRKY transcription factors in plant
647 abiotic stresses. *Biochim. Biophys. Acta, Gene Regul. Mech.* 1819, 120-128.
- 648 Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I.,
649 Creech, M., and Gross, B. (2007). Integration of biological networks and gene expression data using Cytoscape.
650 *Nat. Protoc.* 2, 2366-2382.
- 651 de Matos Simoes, R., and Emmert-Streib, F. (2012). Bagging statistical network inference from large-scale gene
652 expression data. *PLoS One* 7, e33624.
- 653 De Smet, R., and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nat. Rev.*
654 *Microbiol.* 8, 717-729.
- 655 Eulgem, T., and Somssich, I.E. (2007). Networks of WRKY transcription factors in defense signaling. *Curr. Opin.*
656 *Plant Biol.* 10, 366-371.
- 657 Feizi, S., Marbach, D., Médard, M., and Kellis, M. (2013). Network deconvolution as a general method to
658 distinguish direct dependencies in networks. *Nature biotechnology* 31, 726-733.
- 659 Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso.
660 *Biostatistics* 9, 432-441.

Fröhlich, H., Speer, N., Poustka, A., and Beißbarth, T. (2007). GOSim—an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinf.* 8, 166.

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* 16.

Gerstein, M.B., Rozowsky, J., Yan, K.K., Wang, D., Cheng, C., Brown, J.B., Davis, C.A., Hillier, L., Sisu, C., and Li, J.J. (2014). Comparative analysis of the transcriptome across distant species. *Nature* 512, 445-448.

Giorgi, F.M., Del Fabbro, C., and Licausi, F. (2013). Comparative study of RNA-seq-and Microarray-derived coexpression networks in *Arabidopsis thaliana*. *Bioinformatics* 29, 717-724.

Gu, H., Zhu, P., Jiao, Y., Meng, Y., and Chen, M. (2011). PRIN: a predicted rice interactome network. *BMC Bioinf.* 12, 161.

Hase, T., Ghosh, S., Yamanaka, R., and Kitano, H. (2013). Harnessing diversity towards the reconstructing of large scale gene regulatory networks. *PLoS Comput. Biol.* 9, e1003361.

Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H., and Kanda, M. (1996). Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl. Acad. Sci. U.S.A.* 93, 7783-7788.

Hong, S., Chen, X., Jin, L., and Xiong, M. (2013). Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Res.* 41, e95-e95.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5, e12776.

Huynh Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5, e12776.

Iancu, O.D., Kawane, S., Bottomly, D., Searles, R., Hitzemann, R., and McWeeney, S. (2012). Utilizing RNA-Seq data for de novo coexpression network inference. *Bioinformatics* 28, 1592-1597.

Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. (2013). PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.* gkt1016.

Jung, K.-H., Han, M.-J., Lee, D.-y., Lee, Y.-S., Schreiber, L., Franke, R., Faust, A., Yephremov, A., Saedler, H., and Kim, Y.-W. (2006). Wax-deficient *anther1* is involved in cuticle and wax production in rice anther walls and is required for pollen development. *Plant Cell* 18, 3015-3032.

Kitano, H. (2002a). Computational systems biology. *Nature* 420, 206-210.

Kitano, H. (2002b). Systems biology: a brief overview. *Science* 295, 1662-1664.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.* 9, 559.

Lee, I., Seo, Y.-S., Coltrane, D., Hwang, S., Oh, T., Marcotte, E.M., and Ronald, P.C. (2011). Genetic dissection of the biotic stress response using a genome-scale gene network for rice. *Proc. Natl. Acad. Sci. U.S.A.* 108, 18548-18553.

Lertampaiporn, S., Thammarongtham, C., Nukoolkit, C., Kaewkamnerdpong, B., and Ruengjitchachawalya, M. (2013). Heterogeneous ensemble approach with discriminative features and modified-SMOTEBagging for pre-miRNA classification. *Nucleic Acids Res.* 41, e21-e21.

Liu, J., Kang, S., Tang, C., Ellis, L.B., and Li, T. (2007). Meta-prediction of protein subcellular localization with reduced voting. *Nucleic Acids Res.* 35, e96.

Ma, S., Shah, S., Bohnert, H.J., Snyder, M., and Dinesh-Kumar, S.P. (2013). Incorporating motif analysis into gene co-expression networks reveals novel modular expression pattern and new signaling pathways. *PLoS Genet.* 9, e1003840.

Madeira, S.C., and Oliveira, A.L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 1, 24-45.

Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Kellis, M., Collins, J.J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796-804.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., and Munschauer, M. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333-338.

Miao, Y., and Zentgraf, U. (2007). The antagonist function of *Arabidopsis* WRKY53 and ESR/ESP in leaf senescence is modulated by the jasmonic and salicylic acid equilibrium. *Plant Cell* 19, 819-830.

Mitra, K., Carvunis, A.R., Ramesh, S.K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* 14, 719-732.

714 Mutwil, M., Klie, S., Tohge, T., Giorgi, F.M., Wilkins, O., Campbell, M.M., Fernie, A.R., Usadel, B., Nikoloski,
715 Z., and Persson, S. (2011). PlaNet: combined sequence and expression comparisons across plant networks derived
716 from seven species. *Plant Cell* 23, 895-910.

717 Nagasawa, N., Miyoshi, M., Sano, Y., Satoh, H., Hirano, H., Sakai, H., and Nagato, Y. (2003). SUPERWOMAN1
718 and DROOPING LEAF genes control floral organ identity in rice. *Development* 130, 705-718.

719 Obayashi, T., Hayashi, S., Saeki, M., Ohta, H., and Kinoshita, K. (2009). ATTED-II provides coexpressed gene
720 networks for Arabidopsis. *Nucleic Acids Res.* 37, D987-D991.

721 Qin, J., Hu, Y., Xu, F., Yalamanchili, H.K., and Wang, J. (2014). Inferring gene regulatory networks by integrating
722 ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods* 67, 294-303.

723 Ranzani, V., Rossetti, G., Panzeri, I., Arrigoni, A., Bonnal, R.J., Curti, S., Gruarin, P., Provati, E., Sugliano, E., and
724 Marconi, M. (2015). The long intergenic noncoding RNA landscape of human lymphocytes highlights the
725 regulation of T cell differentiation by linc-MAF-4. *Nat. Immunol.* 16, 318-325.

726 Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential
727 expression analysis of digital gene expression data. *Bioinformatics* 26, 139-140.

728 Rodriguez, R.E., Mecchia, M.A., Debernardi, J.M., Schommer, C., Weigel, D., and Palatnik, J.F. (2010). Control of
729 cell proliferation in Arabidopsis thaliana by microRNA miR396. *Development* 137, 103-112.

730 Rushton, P.J., Somssich, I.E., Ringler, P., and Shen, Q.J. (2010). WRKY transcription factors. *Trends Plant Sci.* 15,
731 247-258.

732 Schäfer, J., Opgen-Rhein, R., and Strimmer, K. (2001). Reverse engineering genetic networks using the GeneNet
733 package. *J. Am. Stat. Assoc.* 96, 1151-1160.

734 Shi, J., Tan, H., Yu, X.-H., Liu, Y., Liang, W., Ranathunge, K., Franke, R.B., Schreiber, L., Wang, Y., and Kai, G.
735 (2011). Defective pollen wall is required for anther and microspore development in rice and encodes a fatty acyl
736 carrier protein reductase. *Plant Cell* 23, 2225-2246.

737 Siegfried, K.R., Eshed, Y., Baum, S.F., Otsuga, D., Drews, G.N., and Bowman, J.L. (1999). Members of the
738 YABBY gene family specify abaxial cell fate in Arabidopsis. *Development* 126, 4117-4128.

739 Stief, A., Altmann, S., Hoffmann, K., Pant, B.D., Scheible, W.-R., and Bäurle, I. (2014). Arabidopsis miR156
740 regulates tolerance to recurring environmental stress through SPL transcription factors. *Plant Cell* 26, 1792-1807.

741 Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of
742 conserved genetic modules. *Science* 302, 249-255.

743 Theodoridis, A., Van Dongen, S., Enright, A.J., and Freeman, T.C. (2009). Network visualization and analysis of
744 gene expression data using BioLayout Express3D. *Nature protocols* 4, 1535-1550.

745 Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq.
746 *Bioinformatics* 25, 1105-1111.

747 Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and
748 Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and
749 Cufflinks. *Nat. Protoc.* 7, 562-578.

750 Tzafrir, I., Dickerman, A., Brazhnik, O., Nguyen, Q., McElver, J., Frye, C., Patton, D., and Meinke, D. (2003). The
751 Arabidopsis seedgenes project. *Nucleic Acids Res.* 31, 90-93.

752 Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F.M., Bassel, G.W., Tanimoto, M., Chow, A., Steinhäuser, D.,
753 Persson, S., and Provart, N.J. (2009). Co-expression tools for plant biology: opportunities for hypothesis
754 generation and caveats. *Plant, Cell & Environment* 32, 1633-1651.

755 Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., and Van de Peer, Y. (2009). Unraveling transcriptional
756 control in Arabidopsis using cis-regulatory elements and coexpression networks. *Plant Physiol.* 150, 535-546.

757 Vandepoele, K., Vlieghe, K., Florquin, K., Hennig, L., Beemster, G.T., Gruissem, W., Van de Peer, Y., Inzé, D., and
758 De Veylder, L. (2005). Genome-wide identification of potential plant E2F target genes. *Plant Physiol.* 139,
759 316-328.

760 Vidal, M., Cusick, M.E., and Barabasi, A.-L. (2011). Interactome networks and human disease. *Cell* 144, 986-998.

761 Wu, G., Park, M.Y., Conway, S.R., Wang, J.-W., Weigel, D., and Poethig, R.S. (2009). The sequential action of
762 miR156 and miR172 regulates developmental timing in Arabidopsis. *Cell* 138, 750-759.

763 Xie, Z., Zhang, Z.L., Hanzlik, S., Cook, E., and Shen, Q.J. (2007). Salicylic acid inhibits gibberellin-induced
764 alpha-amylase expression and seed germination via a pathway involving an abscisic-acid-inducible WRKY gene.
765 *Plant Mol. Biol.* 64, 293-303.

766 Yalamanchili, H.K., Li, Z., Wang, P., Wong, M.P., Yao, J., and Wang, J. (2014). SpliceNet: recovering splicing
767 isoform-specific differential gene networks from RNA-Seq data of normal and diseased samples. *Nucleic Acids*

768 Res., gku577.

769 Yang, P., Hwa Yang, Y., B Zhou, B., and Y Zomaya, A. (2010). A review of ensemble methods in bioinformatics.

770 Curr. Bioinform. 5, 296-308.

771 Ye, C.Y., Chen, L., Liu, C., Zhu, Q.H., and Fan, L. (2015). Widespread noncoding circular RNAs in plants. New

772 Phytol. 208, 88-95.

773 Yi, X., Du, Z., and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic

774 Acids Res. 41, W98-W103.

775 Yonemaru, J.I., Yamamoto, T., Fukuoka, S., Uga, Y., Hori, K., and Yano, M. (2010). Q-TARO: QTL annotation rice

776 online database. Rice 3, 194-203.

777 Zhang, Y.C., Liao, J.Y., Li, Z.Y., Yu, Y., Zhang, J.P., Li, Q.F., Qu, L.H., Shu, W.S., and Chen, Y.Q. (2014).

778 Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the

779 sexual reproduction of rice. Genome Biol. 15, 512.

780 Zhang, Z., Yu, J., Li, D., Zhang, Z., Liu, F., Zhou, X., Wang, T., Ling, Y., and Su, Z. (2010). PMRD: plant

781 microRNA database. Nucleic Acids Res. 38, D806-D813.

782

783 Figure Legends

784 **Fig.1** Enrichment folds of different algorithms for co-expression network inference. A) Comparing to GGM with

785 positive links. B) Comparing to WGCNA with positive links. C) Comparing with BC3NET with positive links. D)

786 Comparing with GGM with negative links. E) Comparing with WGCNA with negative links. F) Comparing with

787 BC3NET with negative links. In the legends, the RAW, FPKM, UQ, TMM, RLE and VST represent the networks

788 obtained by the single RNA-seq dataset; INT indicates intra-method consensus networks established by integrating

789 the predictions of different RNA-seq datasets, EBM denotes high-quality gene co-expression network obtained by

790 integrating all intra-method consensus networks

791 **Fig.2** Performance evaluation of our network for predicting gene function. A) Receiver Operating Characteristics

792 (ROC) curve. B) Precision-Recall (PR) curve. In the legends, Not-weighted indicates the evaluation parameters

793 were calculated by the standard method of CAFA project; Weighted indicates the evaluation parameters were

794 calculated by the weighted method of CAFA project

795 **Fig.3** Subnetworks derived from the gene-guide approach. The subnetworks include all other nodes within two

796 layer connections from guide genes. A) *OsMADS16* involved in flower development; B) *OsCESA4* involved in cell

797 wall biosynthesis. Within each subnetwork, red nodes represent the experimentally verified genes related to

798 corresponding biological functions. Pink nodes indicate the genes whose *Arabidopsis* homologs are experimentally

799 verified relating to the corresponding biological processes. Blue nodes represent potential function-related genes,

800 and the gray nodes denote that the genes with unknown functions or annotated with irrelevant functions. The size

801 of node is proportional to the number of connected genes

802 **Fig.4** Subnetworks derived from the known *cis*-regulatory motif-guide approach. A) WTTSSCSS combined with

803 the E2F transcription factors involved in cell cycle. B) TTGACY combined with the WRKY transcription factors

804 involved in stress response. Within each subnetwork, red nodes represent the experimentally verified genes related

805 to corresponding biological functions. Pink nodes indicate the genes whose *Arabidopsis thaliana* homologs are

806 experimentally verified to associate with the corresponding biological functions. Blue nodes denote potential

807 function-related genes. Gray nodes indicate that the genes with unknown functions or annotated with irrelevant
808 functions. The size of node is proportional to the number of connected genes

809 **Fig.5** Co-expression subnetwork derived from guide-gene approach for *XLOC_057324* associated with panicle
810 development and fertility. Within the subnetwork, red nodes represent the experimentally verified genes related to
811 corresponding biological functions; chrysoidine nodes represent transcription factors; pink nodes indicate the
812 genes whose *Arabidopsis thaliana* homologues are experimentally verified to related to corresponding biological
813 functions; blue nodes represent that the genes are potential function-related, and the gray nodes indicate that the
814 genes are function unknown or annotated with unrelated functions

815 **Fig.6** Flowchart of high-quality RNA-seq-based gene co-expression network inference

816 **Acknowledgements**

817 This work is financially supported by The Strategic Priority Research Program of the
818 Chinese Academy of Sciences (Grant No. XDA08020302). The funders had no role
819 in study design, data collection and analysis, decision to publish, or preparation of the
820 manuscript.

821 **Author Contributions**

822 H.Y. conceived the original screening and research plans; H.Y. and C.Z.L conceived
823 the project; C.Z.L. and H.Y. supervised the experiments; H.Y. performed most of the
824 experiments, analyzed the data and wrote the paper; B.K.J analyzed the phenotype
825 data and revised the paper.

826 **Additional Information**

827 **Competing financial interests:** The authors declare no competing financial interests.

[illegible]

Figure B is a network diagram illustrating the relationships between 1000 genes. The nodes are colored blue, red, or purple, and their size represents their degree. The network is highly interconnected, with a central cluster of large red nodes and many smaller nodes on the periphery.

Large-scale RNA-seq Samples

Reads Filtering and
Mapping

Outlier Detection,
Filtering and Data
Normalization

6 RNA-seq Expression Datasets

WGCNA, Hard
Threshold
Determination Using
Permutation Test

BC3NET

GGM, Hard Threshold
Determination Using
Permutation Test

Candidate Gene Co-expression Networks

Ensemble-based
Network Construction

High-quality Gene Co-expression Network