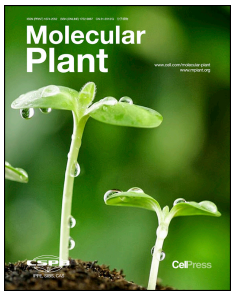
Accepted Manuscript



WheatNet: A genome-scale functional network for hexaploid bread wheat, *Triticum aestivum*

Tak Lee, Sohyun Hwang, Chan Yeong Kim, Hongseok Shim, Hyojin Kim, Pamela C. Ronald, Edward M. Marcotte, Insuk Lee

PII: S1674-2052(17)30108-9
DOI: 10.1016/j.molp.2017.04.006
Reference: MOLP 457

To appear in: *MOLECULAR PLANT*
Accepted Date: 19 April 2017

Please cite this article as: **Lee T., Hwang S., Kim C.Y., Shim H., Kim H., Ronald P.C., Marcotte E.M., and Lee I.** (2017). WheatNet: A genome-scale functional network for hexaploid bread wheat, *Triticum aestivum*. Mol. Plant. doi: 10.1016/j.molp.2017.04.006.

WheatNet: A genome-scale functional network for hexaploid bread wheat, *Triticum aestivum*

Tak Lee[1,8], Sohyun Hwang[1,2,3,8], Chan Yeong Kim[1], Hongseok Shim[1], Hyojin Kim[1], Pamela C. Ronald[4,5,6*], Edward M. Marcotte[2,7,*] and Insuk Lee[1,*]

[1]Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, Seoul 120-749, Korea

[2]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, Texas 78712, USA

[3]Department of Biomedical Science, College of Life Science, CHA University, Seongnam-si 13496, Korea

[4]Department of Plant Pathology, University of California, Davis, California 95616

[5]Joint Bioenergy Institute, Emeryville, California 94608

[6]The Genome Center, University of California, Davis, California 95616

[7]Department of Molecular Biosciences, University of Texas at Austin, Texas 78712, USA

[8]These authors contributed equally.

*Corresponding authors: I.L: insuklee@yonsei.ac.kr, E.M.M.: marcotte@icmb.utexas.edu, P.C.R.: pcronald@ucdavis.edu

Running title: A functional gene network for *Triticum aestivum*

Keywords: Bread wheat, *Triticum aestivum*, gene network, crop genetics

Dear Editor:

Gene networks provide a system-level overview of genetic organizations and enable the dissection of functional modules underlying complex traits. Integration of diverse genomics data based on the Bayesian statistics framework has been successfully applied to the construction of genome-scale functional networks for major crop species such as rice (Lee et al., 2011), soybean (Kim et al., 2017), and tomato (Kim et al., 2016), and their predictive power for gene-to-trait associations has been demonstrated. However, such a predictive gene network is not yet available for bread wheat, *Triticum aestivum,* an important staple food crop accounting for approximately 20% of the world's daily food consumption. Bread wheat also serves as a model for studying polyploidy in plants.

Some of the reasons that functional genomics studies on bread wheat have lagged behind those on other crops include the large genome of bread wheat (~17 Gb) and its polyploidy nature, which complicates genetic analysis. However, recent advances in wheat research have considerably improved genome assembly and gene models (International Wheat Genome Sequencing, 2014). Furthermore, the discovery and application of genome editing (Upadhyay et al., 2013) and TILLING technologies (Uauy et al., 2009) have enabled targeted knockout in wheat protoplasts and whole plants. These developments have set the stage for the application of reverse genetics approaches for the functional characterization of wheat genes.

Here we present WheatNet, a genome-scale functional gene network for *T. aestivum* and a companion web server (www.inetbio.org/wheatnet), which provides network information and generates network-based functional hypotheses. WheatNet was constructed by integrating 20 distinct genomics datasets (**Supplemental Table 1**), including 156,000 wheat-specific co-

expression links mined from 1,929 DNA microarray datasets (**Supplemental Table 2**). A unique feature of WheatNet compared with previously constructed crop functional networks is that each network node in WheatNet represents either a single gene or a group of genes to reduce complexity. An allopolyploid wheat genome contains three homeologous chromosome sets—A, B, and D—that originate from three closely related species *Triticum urartu*, *Aegilops speltoides*, and *Aegilops tauschii*, respectively (International Wheat Genome Sequencing, 2014). Therefore, the wheat genome contains many homologous genes between the three ancestral chromosome sets. Because homeologs are likely to have redundant functions, collapsing homeologs into a single network node would facilitate the network analysis by reducing network complexity. Unfortunately, comprehensive definitions of wheat homeologous relationships are not yet available. Therefore, we computationally partitioned "gene groups" mimicking homeologous genes by clustering 99,386 wheat genes, resulting in 20,248 gene groups comprising 63,401 genes, and 35,985 individual genes. WheatNet was thus constructed using 56,233 nodes; the final network has 20,230 nodes (13,430 gene groups and 16,800 individual genes) and 567,000 edges, integrating 20 sources of functional evidence linking pairs of genes (**Supplemental Methods**). The edge information of the integrated WheatNet and all 20 component networks are available for download.

To assess WheatNet, we used biological process annotations by agriGO (Du et al., 2010), which are moderately distinct from the dataset used for network training (~38% gene pairs by shared agriGO annotations overlap the training data) and one of the few other large-scale wheat annotation sets available for testing. To help reduce bias, we excluded agriGO terms that annotate more than 300 wheat genes. Next, the accuracy of functional gene pairs by WheatNet or by random chance was measured using the proportion of gene pairs that share

agriGO annotations for different coverage of the coding genome. We observed strong performance by WheatNet, in which a network covering approximately 20% of all genes map functional gene pairs with about 40% accuracy (**Supplemental Figure 1**). The quality of WheatNet was further evaluated by the degree of connectivity among genes involved in a particular biological process. Considering that genes for the same complex traits are more likely to be functionally coupled, high connectivity among known genes for a trait would support the quality of functional networks. We tested network connectivity for a group of genes based on two measures: (i) the number of edges among gene members (i.e., within-group edge count) and (ii) the number of network neighbors that overlap among group members (i.e., network neighbor overlap). We used genes for two complex traits derived from proteomics studies: 45 genes with differential protein expression after *Blumeria graminis* f. sp. *tritici* infection (Mandal et al., 2014) and 17 genes with differential protein expression under drought conditions (Cheng et al., 2015). The significance of network connectivity was also measured based on a null distribution from 1000 random gene sets of the same size. We found that the connectivity among each trait's genes was significantly higher than by random chance (**Figure 1A-B**). We consistently observed network communities of genes for both traits (**Figure 1C-D**). We conclude that WheatNet successfully predicts additional genes that are involved in a given trait.

The WheatNet web server provides two options for prioritizing genes for wheat traits: (i) direct neighbors in the gene network and (ii) context-associated hubs (CAHs). In the first approach, a user submits genes known for a trait that can guide network searches for new candidate genes. New genes are then ranked by the strength of evidence connecting them to the "guide genes," measured for each candidate gene as the sum of network edge scores from

that gene to the guide genes. The result page provides the ranked list of candidates and a visualization of the local guide gene network (**Figure 1E**). To provide functional clues for candidate genes, WheatNet provides available wheat and *Arabidopsis* gene annotations from the Gene Ontology biological process (GOBP) (**Supplemental Methods**).

In the second approach, users exploit gene expression data related to a trait of interest. Gene expression profiles are one of the most common types of genomic data, and differential expression analysis provides many genes that are potentially associated with given traits such as abiotic and biotic stresses. However, many genes that are associated with stress conditions are not differentially expressed. Hypothesizing that a gene associated with many differentially expressed genes (DEGs) in stress (i.e., CAHs) is likely to be responsible for responses to the given stress condition, we prioritized genes by connections to the context-associated DEGs. To conduct CAH prioritization, we first defined a subnetwork that comprises a hub gene and all of its network neighbors in WheatNet. For the gene prioritization, we considered only subnetworks with hub genes that have at least 50 network neighbors. Assuming that DEGs are representative genes for a relevant biological context, we prioritized hub genes based on the enrichment of their network neighbors for the DEGs, measured using Fisher's exact test. The hub genes with significant enrichment ($P < 0.01$) of network neighbors for the DEGs are considered as CAHs and are presented as candidate genes for the context-associated trait. Similar to the network direct neighborhood search, all candidate genes are appended by GOBP annotations for wheat genes and for Arabidopsis orthologs. In addition, users can access a network view of a CAH and its connected DEGs by clicking each candidate gene (**Figure 1F**).

The WheatNet predictions by each of the network-based gene prioritization methods were

validated as follows: For the network direct neighborhood method, we evaluated the new candidate genes for drought stress response that were predicted by submitting 17 genes with differential protein expression under drought conditions (Cheng et al., 2015) as guide genes. We hypothesized that novel candidate genes for drought response are also likely to be expressed differentially under drought conditions. Thus, we investigated the enrichment of candidate drought response genes from DEGs under drought conditions. We generated a set of 2,346 DEGs under drought condition based on genes that showed more than 4-fold changes in expression levels at $P < 0.01$ (SRP045409 of NCBI Sequence Read Archive) (Liu et al., 2015). We found 15 drought-condition DEGs among the top 50 candidate genes by the network direct neighborhood method, which indicates more than 7-fold enrichment over predictions by random chance (15/50 = 0.3 by WheatNet vs. 2346/56233 = 0.042 by random chance). For the CAH method, we evaluated the candidate genes for *Fusarium graminearum* infection response that were predicted by submitting 837 DEGs after infection with *F. graminearum* (GSE54551 of NCBI Gene Expression Omnibus database) (Wojcik et al., 2015) as user input data. We found that the top 100 candidates by CAHs were significantly enriched for GOBP annotations relevant to fungus infection based on Arabidopsis orthologs: 'response to chitin' (GO:0010200, $P = 9.72 \times 10^{-31}$), 'regulation of plant-type hypersensitive response' (GO:0010363, $P = 8.20 \times 10^{-21}$), 'defense response to fungus' (GO:0050832, $P = 1.73 \times 10^{-20}$), 'response to fungus' (GO:0009620, $P = 1.03 \times 10^{-8}$), and 'detection of biotic stimulus' (GO:0009595, $P = 3.43 \times 10^{-5}$).

These results indicate that WheatNet can effectively prioritize novel candidate genes for complex traits, including those governing abiotic and biotic stress responses, by using multiple network-based methods, which can be easily performed by simple submission of

input data in the companion web server. WheatNet complements other types of knowledge mining systems (Hassani-Pak et al., 2016) and provides a useful resource for systems biology and predictive genetics analysis of wheat.

Author contributions

T.L. and S.H. developed the network model and conducted bioinformatics analysis. C.Y.K., H.S., and H.K. assisted data analysis for network modeling. P.C.R., E.M.M., and I.L. designed and supervised the study. T.L. and I.L. drafted the manuscript. P.C.R., and E.M.M., edited the manuscript.

Acknowledgments

This work was supported by grants from the National Research Foundation of Korea (2012M3A9B4028641, 2012M3A9C7050151, and 2015R1A2A1A15055859) to I.L. This work was supported by a grant to PCR and EMM from NSF (1237975) and from the Welch Foundation (F1515) to EMM. The work conducted by the US Department of Energy Joint Genome Institute was supported by the Office of Science of the US Department of Energy under Contract no. DE-AC02-05CH11231.

We thank Jorge Dubcovsky, Ksenia Krasileva, Kelly Eversole and Catherine Feuillet for helpful discussions.

References

Cheng, Z., Dong, K., Ge, P., Bian, Y., Dong, L., Deng, X., Li, X., and Yan, Y. (2015). Identification of Leaf Proteins Differentially Accumulated between Wheat Cultivars Distinct in Their Levels of Drought Tolerance. PLoS One 10:e0125302.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 38:W64-70.

Hassani-Pak, K., Castellote, M., Esch, M., Hindle, M., Lysenko, A., Taubert, J., and Rawlings, C. (2016). Developing integrated crop knowledge networks to advance candidate gene discovery. Appl Transl. Genom. 11:18-26.

International Wheat Genome Sequencing, C. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345:1251788.

Kim, E., Hwang, S., and Lee, I. (2017). SoyNet: a database of co-functional networks for soybean Glycine max. Nucleic Acids Res. 45:D1082-D1089.

Kim, H., Kim, B.S., Shim, J.E., Hwang, S., Yang, S., Kim, E., Iyer-Pascuzzi, A.S., and Lee, I. (2016). TomatoNet: A Genome-wide Co-functional Network for Unveiling Complex Traits of Tomato, a Model Crop for Fleshy Fruits. Molecular Plant Advance Access published Nov 29, 2016, doi: 10.1016/j.molp.2016.11.010.

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.

Liu, Z., Xin, M., Qin, J., Peng, H., Ni, Z., Yao, Y., and Sun, Q. (2015). Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (Triticum aestivum L.). BMC Plant Biol. 15:152.

Mandal, M.S., Fu, Y., Zhang, S., and Ji, W. (2014). Proteomic analysis of the defense response of wheat to the powdery mildew fungus, Blumeria graminis f. sp. tritici. Protein J. 33:513-524.

Uauy, C., Paraiso, F., Colasuonno, P., Tran, R.K., Tsai, H., Berardi, S., Comai, L., and Dubcovsky, J. (2009). A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. BMC Plant Biol. 9:115.

Upadhyay, S.K., Kumar, J., Alok, A., and Tuli, R. (2013). RNA-guided genome editing for target gene mutations in wheat. G3 3:2233-2238.

Wojcik, P.I., Ouellet, T., Balcerzak, M., and Dzwinel, W. (2015). Identification of biomarker genes for resistance to a pathogen by a novel method for meta-analysis of single-channel microarray datasets. J. Bioinform. Comput. Biol. 13:1550013.

Figure legend

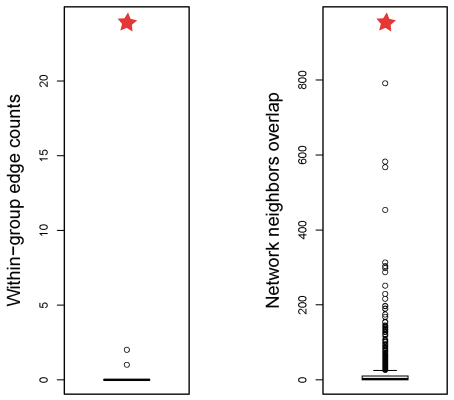
Figure 1. **Overview of WheatNet**

Degree of connectivity (A) among 45 genes for response to *Blumeria graminis f.* sp. *Tritici*

infection and (B) among 17 genes for response to drought stress were measured by determining the number of edges among group members (i.e., within-group edge count) or the number of network neighbors that overlapped among group members (i.e., network neighbor overlap) by using WheatNet (red stars) or 1000 random gene sets having the same size (black circles). Largest components of networks of (C) the genes for response to *B. graminis f.* sp. *Tritici* infection and (D) those for response to drought stress by WheatNet. (E) Results of gene prioritization by direct neighborhood method. The top 100 candidate genes and associated Gene Ontology terms are listed in a table. In addition, the network of guide genes and candidate genes is shown. (F) The results of gene prioritization by the context-associated hub method. The top 100 predictions and associated Gene Ontology terms are listed in a table. By clicking each candidate gene, users can view a network composed of the hub gene and connected differentially expressed genes.

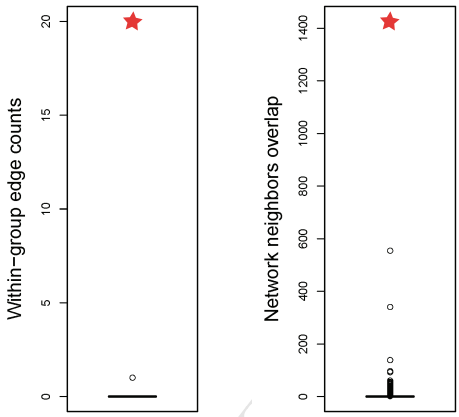
A

Bgt infection

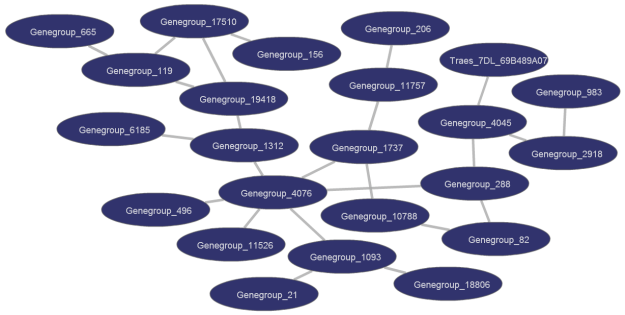


B

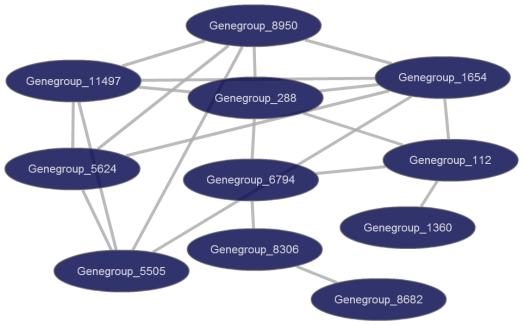
drought stress



C



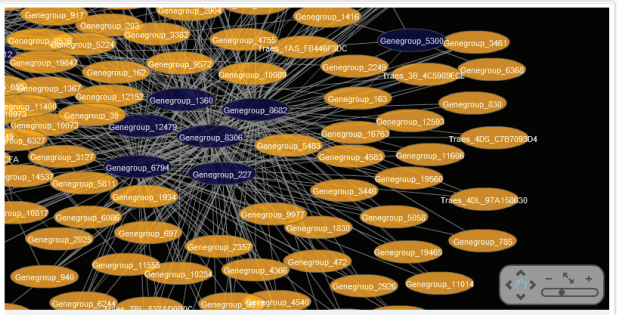
D



E

* Here only top 100 predictions are shown.

Rank	IWGSC_MIPSv2.2 id or gene group id	Score	Evidences:Contribution	# connected guide gene / # valid guide gene	Connected guide gene
		Triticum aestivum GO biological process			
1	Genegroup_12152	59.97	SC-CX:0.17 SC-CC:0.17 TA-GN:0.15 HS-HT:0.14 ZM-CX:0.08 DR-CX:0.07 CE-CX:0.06 AT-CX:0.05 HS-CX:0.05 SC-HT:0.04	7/17	Genegroup_112 Genegroup_1360 Genegroup_1654 Genegroup_288 Genegroup_6794 Genegroup_8306 Genegroup_8682
		• Gramene GO BP: glycolytic process;phosphorylation • Arabidopsis GO BP: glycolytic process;response to endoplasmic reticulum stress;response to arsenic-containing substance;response to cadmium ion			
2	Genegroup_917	59.18	AT-CX:0.28 HS-HT:0.20 ZM-CX:0.16 TA-CX:0.12 DM-CX:0.07 SC-CC:0.06 SC-CX:0.05 SC-HT:0.03 SC-GT:0.03	10/17	Genegroup_112 Genegroup_11497 Genegroup_1654



Node type = candidate gene
Gene symbol = Genegroup_5407
Total connection score to query genes(sum of LLS) = 28.13
Locus name = Genegroup_5407

F

List of candidate context-associated hubs.

Full results are found in this report file: *Report file*

* If the predicted hub is the user submitted differentially expressed gene, they are marked as 'DEG'

* **Clicking the locus ID will visualize the network connections of hubs and query DEGs. This takes a few minutes so please be patient.**

Rank	IWGSC_MIPSv2.2 ID Gene group ID	submitted DEG?	biological process terms of wheat	p-value
1	Genegroup_917 View Network		Gramene GObp: metabolic process [+] Open Arabidopsis GObp terms	7.302e-15
2	Genegroup_4041 View Network		Gramene GObp: chlorophyll biosynthetic process; cofactor biosynthetic process; metabolic process; oxidation-reduction process; photosynthesis	1.377e-13