Gene networks

Gene networks can be useful tools to explore and generate hypotheses about gene function. In wheat there are many gene networks which have been used for specific research questions, but there are also several networks which could be used for more generic questions about gene function. These include co-expression networks that identify genes expressed in a similar pattern across different samples and a GENIE3 network that identifies putative targets of transcription factors.

a) Co-expression networks

Co-expression networks incorporating 850 RNA-Seq samples from diverse tissues, developmental stages and stress conditions have been developed using the RefSeqv1.0 transcriptome annotation (see Gene models for details on the annotation). All high confidence genes expressed >0.5 tpm were included for the networks, which were created using the R package WGCNA (Langfelder and Horvath, 2008).

One of the networks integrates all 850 RNA-Seq samples, which is referred to as "WGCNA_850" in the associated data files. Further details on the network construction and the samples included can be found in IWGSC (2018). The WGCNA_850 network has been incorporated into KnetMiner for rapid analysis and visualisation.

There are an additional four tissue-specific networks that were generated using grain (n = 119), leaf (n = 245), root (n = 45), and spike (n = 128) samples.

Additionally, stress-specific networks were created which included 12 distinct abiotic and disease stress conditions. The samples were divided into abiotic stress and disease stress (along with their relevant control samples) to produce an abiotic stress network and a disease stress network. More details for the tissue and stress-specific networks are available in the associated publication (Ramírez-González, Borrill *et al.*, 2018).

All data associated with these networks is publicly available through https://opendata.earlham.ac.uk/wheat/under_license/toronto/Ramirez-Gonzalez_etal_2018-06025-Transcriptome-Landscape/

All scripts used to develop the networks and for downstream analysis are available here: https://github.com/Uauy-Lab/WheatHomoeologExpression

A workflow for a potential data analysis is described in section c) below.

b) GENIE3 network (transcription factor target prediction)

A second type of network has been developed using the GENIE3 algorithm (Huynh-Thu *et al.*, 2010). This algorithm takes a Random Forest approach to predict targets of transcription factors based on expression datasets. The wheat GENIE3 network was built expression data for high confidence genes expressed >5 tpm in 850 RNA-seq samples from diverse tissues, developmental stages and stress conditions. The network incorporates 4,956 annotated transcription factors, derived from the RefSeqv1.0 gene annotation. The complete network is available to download. Further details are available in the associated publication (Ramírez-González, Borrill *et al.*, 2018). This GENIE3 network

has been incorporated into KnetMiner for rapid analysis and visualisation (Figure 1), and global analyses and validation have been carried out in Harrington *et al.*, (2019).

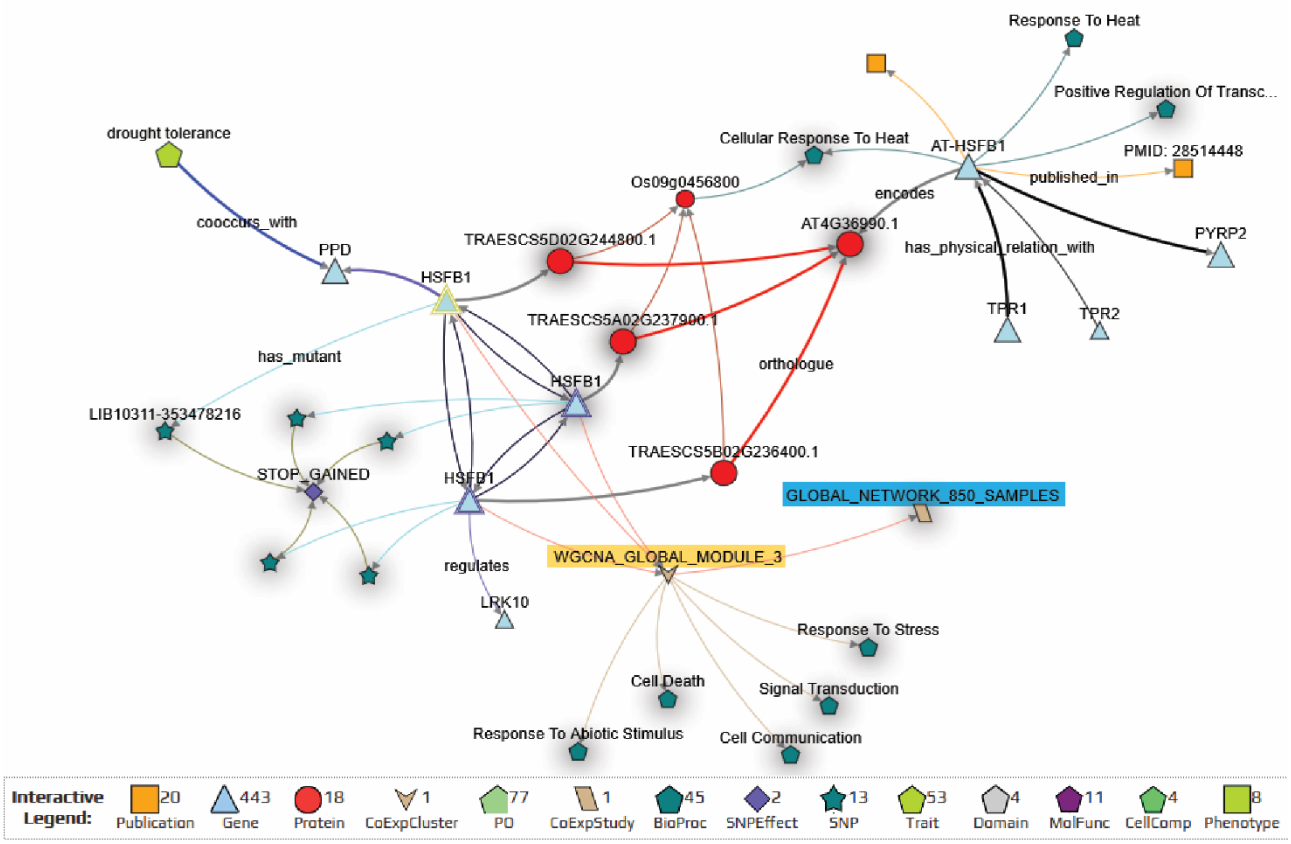

Figure 1. Visualization of networks using Knetminer
The WGCNA_850 network has been integrated into Knetminer, allowing for its rapid analysis and visualization. As an example, the WGCNA_850 network (blue highlight) contains a sub-group (called a module; yellow highlight), which further contains a gene (*TaHSFB1*) that is associated with regulating drought tolerance. There are TILLING mutants available for all three homoeologs of this gene, and each homoeolog is linked to the orthologous gene in rice and *Arabidopsis*, which contain further links to putative roles, interacting partners and peer-reviewed publications.

c) Investigating gene network information about a candidate gene

In this case study we will use a transcription factor *TaHSF1* which is the ortholog of a known *Arabidopsis* gene (also known as *AtHsfB1*) which is critical to the growth to defence transition (Pajerowska-Mukhtar *et al.*, 2012).

TaHSF1 has three homoeologs: *TraesCS5A01G237900*, *TraesCS5B01G236400* and *TraesCS5D01G244800*.

We can download the information about which co-expression module each homoeolog is in: WGCNA_table.csv . Using this we filter the table to identify which module the homoeologs are in.

Table 1. Abiotic and disease network modules for *TaHSF1*.

Gene	Module	Number_genes_in_module	Network
TraesCS5A01G237900	2	4485	WGCNA_abiotic
TraesCS5B01G236400	2	4485	WGCNA_abiotic
TraesCS5D01G244800	2	4485	WGCNA_abiotic
TraesCS5A01G237900	12	1693	WGCNA_disease
TraesCS5B01G236400	12	1693	WGCNA_disease
TraesCS5D01G244800	8	2637	WGCNA_disease

Using this information, we can see that the three homoeologs are largely in the same modules.
The next step is to identify stresses these modules are associated with. This information is available in TableS29 and TableS30 in Ramírez-González, Borrill *et al*., 2018. This analysis reveals that module 2 in the abiotic network is associated with drought, and drought and heat. Module 12 in the disease network is associated with chitin and flagellin 22, whilst module 8 is associated with responses to yellow rust. To further explore these modules we could identify hub genes using this data. This may reveal genes which are hubs in the module which may be important for controlling the module as a whole. We could also identify which GO terms are enriched within these modules using the ontologies file (filter for the ontology "IWGSC+Stress" or "slim_IWGSC+Stress" for the most comprehensive GO term assignments or the GO Slim version). GO term enrichment for all the genes in each module can then be calculated using an enrichment program such as goseq. It is necessary to use the universe table to select which genes to use as a background for the enrichment because not all genes in the annotation will have be included in the network due to the expression cut off. For example, if you are calculating the GO term enrichment for modules from the abiotic network, select only genes from the universe "WGCNA_abiotic" as a background for the enrichment calculation.

Taken together this information can generate new hypotheses about the function of candidate genes, which can be then be tested for example using knock-out mutants generated using the TILLING mutants, transgenics, viral induced gene silencing or segregating populations.

References

Harrington, S.A., Backhaus, A.E., Singh, A., Hassani-Pak, K., Uauy, C. Validation and characterisation of a wheat GENIE3 network using an independent RNA-Seq dataset. *BioRxiv*. doi:10.1101/684183

Huynh-Thu V, Irrthium A, Wehenkel L and Geurts P, 2010. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, **5**, e12776. doi:10.1371/journal.pone.0012776

IWGSC 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 361: eaar7191 doi:10.1126/science.aar7191

Langfelder P and Horvath S, 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559 doi:10.1186/1471-2105-9-559

Pajerowska-Mukhtar KM, Wang W, Tada Y, Oka N, Tucker CL, Fonseca JP, Dong XN, 2012. The HSF-like transcription factor TBF1 is a major molecular switch for plant growth-to-defense transition. *Current Biology* 22, 103. doi: 10.1016/j.cub.2011.12.015

Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, Davey M, Jacobs J, van Ex F, Pasha A, Khedikar Y, Robinson S, Cory A, Florio T, Concia L, Juery C, Schoonbeek H, Steuernagel B, Xiang D, Ridout CJ, Chalhoub B, Mayer KFX, Benhamed M, Latrasse D, Bendahmane A, International Wheat Genome Sequencing Consortium, Wulff BBH, Appels R, Tiwari V, Datla R, Choulet F, Pozniak C, Provart NJ, Sharpe AG, Paux E, Spannagl M, Bräutigam A, Uauy C. 2018. The transcriptional landscape of polyploid wheat. *Science* 361: eaar6089. doi:10.1126/science.aar6089