Aggregating and exploring *Arabidopsis thaliana* Gene Regulatory Networks

Rachel Woo | BCB430 | April 9th 2019

Supervisors: Professor Nick Provart and Vincent Lau

Introduction

The molecular interactions that control gene expression are collectively known as the interactome. A significant aspect of the interactome is the set of interactions that regulate levels of mRNA and proteins, known as a gene regulatory network (GRN) (Mayo M et al., 2012 and Ouma et al., 2018). GRNs can be modeled via predictive algorithms (Christodoulou et al., 2019) or generated via experiments such as yeast 1 or 2 hybrid, chip-seq or microarray (Ikeuchi et al., 2018 and Keurentjes JJ et al, 2007). Exploration of these networks can be used to elucidate important loci of control in the cell (Rowland et al., 2017). Aggregating and exploring the GRNs extant in the literature can be a useful tool to highlight important regulations and generate research hypotheses. In this project our team explores and visualizes GRNs for *Arabidopsis thaliana* by conducting network analysis and contributing to the creation of an exploratory bioinformatics tool aggregating and exploring GRNs.

GRNs can be visualized in numerous ways, however, graphs are the most widely used and easy to visualize. Graphs are sets of edges (E) and vertices (V), where a given edge (e) is composed of some source (v_i) and some target (v_j) in V. Edges can optionally have both a directionality and/or weight. In the context of GRNs the directionality of an edge would be the source acting upon the target to either activate or repress its expression and the weight of an edge would illustrate the strength of the interaction. Beyond these basic data, numerous metrics can be associated with the vertices creating a rich tapestry of information. With annotation, vertices can be augmented with information about the gene's expression level, cellular location, co-expression group, chromosome number or any other mutually exclusive categories (Karlebach and Shamir, 2008).

GRNs can also be analyzed via network theory (Alon U, 2007). One such metric is shortest path betweenness centrality (SPBC) which is based upon the shortest paths in a network or degree centrality which is the number of links incident on a node (Koschützki and Schreiber, 2008). This method involved counting the number of shortest paths a given vertex (v) is involved in. This allows a researcher to estimate which vertices are essential for network communication and flow (Koschützki and Schreiber, 2008). Several papers have noted that SPBC can be used in combination with other network statistics to illustrate significant genes. Specifically, Sonawane et al.'s 2017 paper found human tissue-specific genes had a higher median betweenness relative to non-tissue-specific genes despite a relatively lower number of edges. Bafna and Isaac’s 2017 paper further used this betweenness and closeness centrality, calculated from Cytoscape, to identify breast cancer genes as targets for drug therapies highlighting the significance of betweenness centrality and other network statistics in the identification of regulatory elements.

Significant regions of GRNs can also be identified via network motifs or recurring patterns in the regulation (Alon U, 2007). A particular motif of note is the feed-forward loop in which one transcription factor (C) is jointly regulated by 2 others (A and B). Depending upon the patterns of activation or repression within these motifs can lead to different patterns of regulation where either both expression of A and B are required to regulate C or where expression of either A or B regulates C (Mangan S, and Alon U., 2003). The consistency of this pattern in implicating regulation has caused it to be a source of interest in many different gene regulatory network studies (Chen X, Li M, Zheng R, et al., 2019; Zhiponova et al., 2014; Sakuraba et al., 2015).

Clearly there exists a need for a tool which compiles and explores *Arabidopsis thaliana* GRNs. This is the goal of the webapp Arabidopsis Gene Network Tool (aGENT) created by Vincent Lau and for which I am contributing features. This tool will be useful for researchers to explore GRNs and ultimately identify and understand regulation. aGENT is one of many tools on the Bio-Analytic Resource (BAR) which mainly facilitates exploratory analysis for genomic Arabidopsis data.

Materials and Methods

aGENT is implemented in React JS app and uses Cytoscape JS (Franz et al., 2016) for GRN aggregation and exploration. My role in this project consists of GRN curation, developing exploratory tools and implementing UI features for ease of use. A summary can be seen below:

1. Aggregate and document experimentally validated *Arabidopsis thaliana* GRNs to populate a database via:

    Manually annotating GRNs

    Uploading GRNs to mySQL database

2. Develop tools for network exploration and data integration including:

    Shortest Path

    Betweenness centrality

    Motif explorer

    Find Selected Targets

    Upload and Download functionality

3. Develop UI features for ease of use including:

    Layouts

    Clean UI design

    Gene search box

Results

1. GRN Curation

Annotation

Thus far I have manually annotated 9 GRNs that can be seen in the figure below:

Table 1: Manually Annotated Networks and their aGENT links

GRN Citation	**GRN Link in aGENT**
(Ikeuchi et al., 2018)	http://www.bar.utoronto.ca/aGENT/network/20
(Keurentjes et al, 2007)	http://www.bar.utoronto.ca/aGENT/network/14
(Park et al., 2015)	http://www.bar.utoronto.ca/aGENT/network/15
(Sparks et al., 2016)	http://www.bar.utoronto.ca/aGENT/network/17
(Vidal et al., 2015)	http://www.bar.utoronto.ca/aGENT/network/19
(Zhang et al., 2019)	http://www.bar.utoronto.ca/aGENT/network/18
(Smit et al., 2020)	http://www.bar.utoronto.ca/aGENT/network/32
(Jin et al., 2015)	http://www.bar.utoronto.ca/aGENT/network/35
(Espinosa-Soto C et al., 2004)	http://www.bar.utoronto.ca/aGENT/network/36

The link for the sif annotations can be found at my git repository (https://github.com/raywoo32/grnAnnotation) or by clicking the download button at each respective GRN's aGENT link. At this link, the original paper can also be accessed by clicking on the GRN title.

Database Population

To upload the annotated .sif files to the mySQL database a script to automate the process was created. This is based on Vincent's redesigning of the BAR interactions database (https://github.com/VinLau/BAR-interactions-database) and can be found here (https://github.com/raywoo32/readSIF). This script was used to automatically upload manual annotations to the BAR's GRN database.

2. Exploratory and Data Integration Features

Shortest Path

Functionality for obtaining the shortest path has been implemented using the Cytoscape JS function aStar (Franz et al., 2016). This feature allows researchers to visualize the flow of information through the network and between nodes. Intuitively, if two genes are more closely connected they are more likely to affect one another. I demonstrate this by easily visualizing *AT5G49450* to *AT1G07640* using data from Brady et al., 2011. The user can note that the path degree is 9 and thus the two genes are unlikely to strongly be linked in the network (Figure 1).

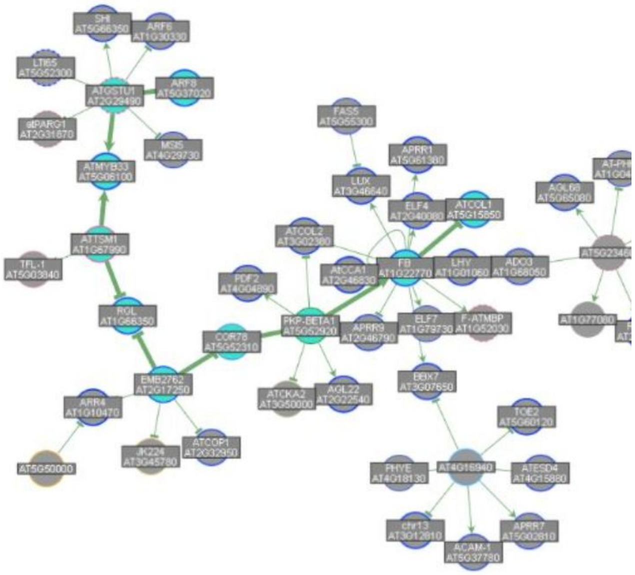


Figure 1: Visualization of the shortest path from *AT5G49450* to *AT1G07640* for the Root Steele Network, data from Brady et al., 2011.

Betweenness and Degree Centrality

Both normalized betweenness and normalized degree centrality have been implemented as exploratory features using functions from Cytoscape JS. The choice to scale by the network size was made to make it intuitive to the user and so all nodes are easily legible. The centrality is mapped to node size between 30 and 70 px where 30 is 0 and 70 is the most central node in the network. The subtle differences in centrality metrics can be seen when using this tool on the 2007 Keurentjes et al. paper (Figure 2).

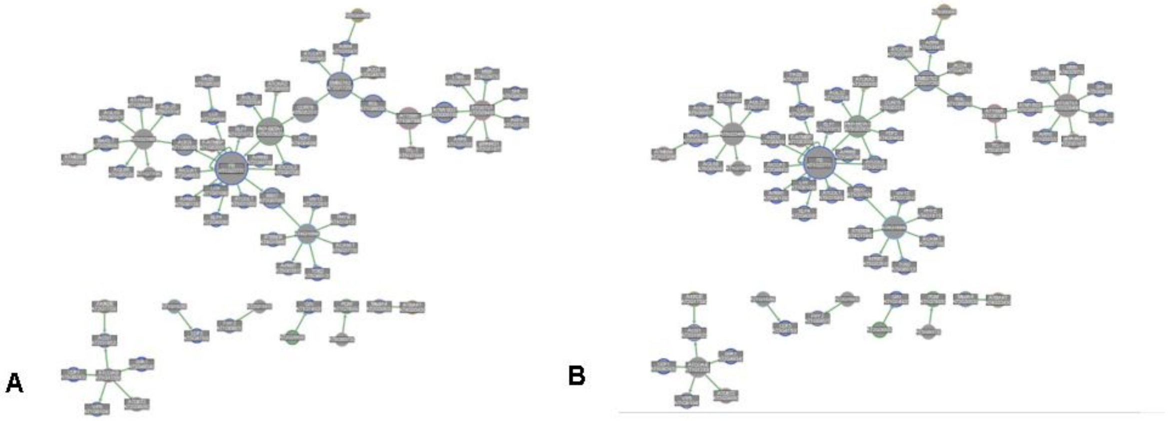


Figure 2: Centrality metrics for the Flowering Time Network, data from Keurentjes et al., 2007. (A) Normalized Betweenness (B) Normalized Degree

Motif Explorer

The motif explorer uses Encode's mfinder API and previous work done by Vincent Lau to interactively display network motifs (Kashtan et al., 2004). When the motif tab is clicked the API is queried and the user can select the motif type by mfinder ID and then step through the motifs of that type using a slider. The usefulness can of this tool can be seen below in Figure 3 where a regulatory pattern (feed forward loop) has been identified.

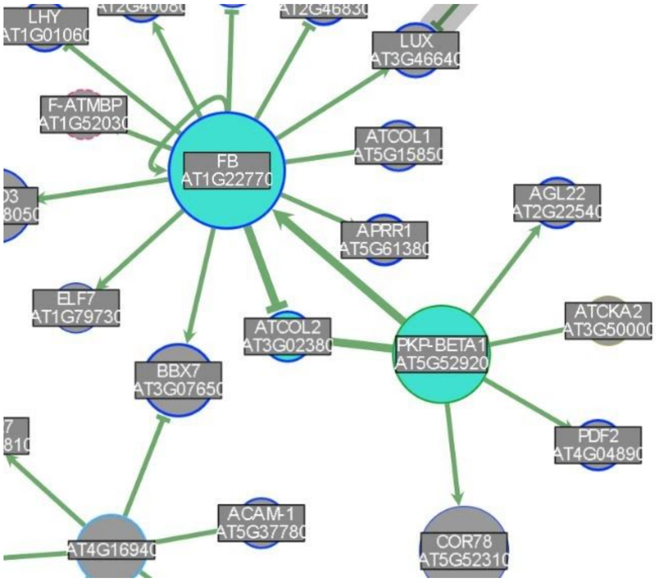
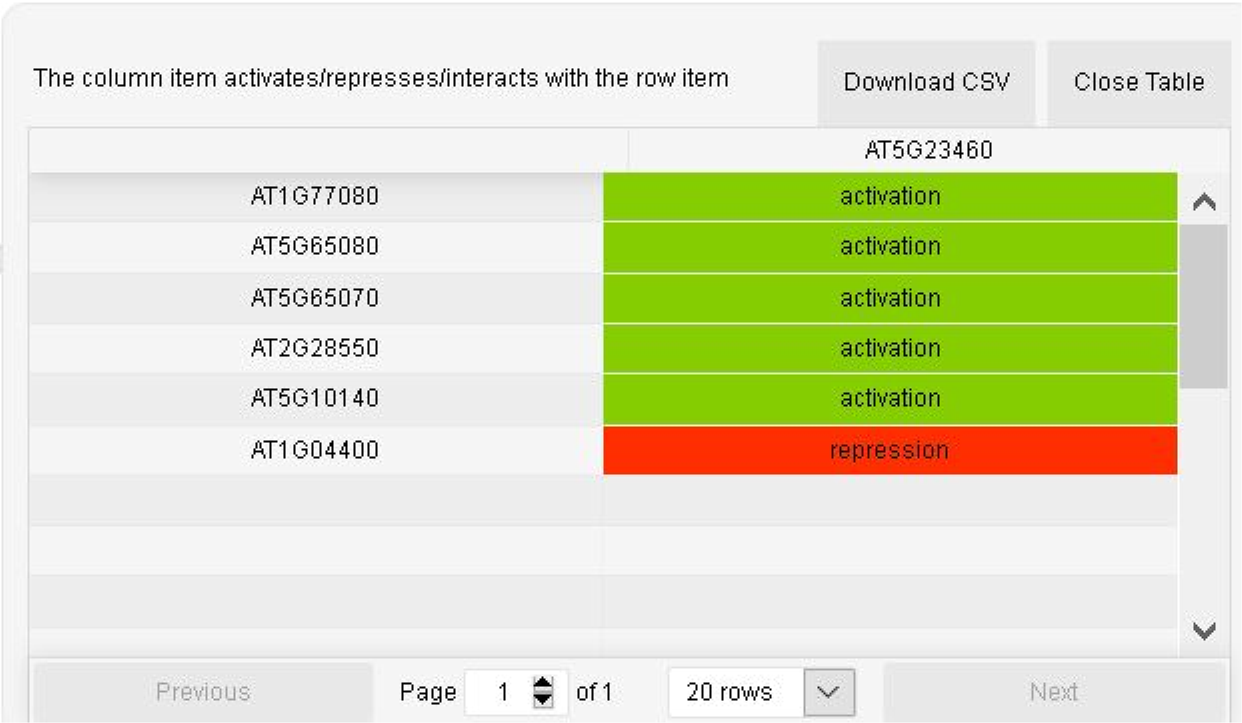


Figure 3: Motif 38 (Feed forward loop) for the Flowering Time Network, data from Keurentjes et al., 2007.

Find Selected Targets

The Find Selected Targets feature allows the user to visualize interactions and their mode of action in a chart view for user selected genes. The source of the interaction appears on the top row and the target appears on the leftmost columns. A csv of these interactions can also be downloaded from the interface. An example usage can be seen in Figure 4, where all targets and modes of regulation for the transcription factor AT5G23460 can easily and intuitively be seen.



The column item activates/represses/interacts with the row item

	AT5G23460
AT1G77080	activation
AT5G65080	activation
AT5G65070	activation
AT2G28550	activation
AT5G10140	activation
AT1G04400	repression

Figure 4: The interactions and their mode of action for all genes connected to *AT5G23460* in the Flowering Time Network from Keurentjes et al., 2007.

Upload and Download

Upload and download functionality has been implemented for .sif files (where the format of the .sif file is as described by cytoscape). The upload is destructive, allowing the user to replace the network with one of their own annotations. The download button gives the user access to the .sif files used to populate the database on the backend.

3. UI Features

Default Display and Optional Layout

Several Cytoscape JS layouts such as cose-blinket, breadth layout, spread and klay (Dogrusoz et al., 2009; Franz et al., 2016) have been implemented. Depending upon the network size, different default layouts have been implemented such that small networks use the force directed layout (Cose-Blinket) and the large networks use the breadth layout. This choice is to optimize initial load time and ease of viewing for the user. The other two optional layouts, spread and klay can respectively be used to make edges more accessible and to see interactions visualized in a pseudo-bipartite manner. A demonstration of the layouts feature can be seen below in Figure 5.

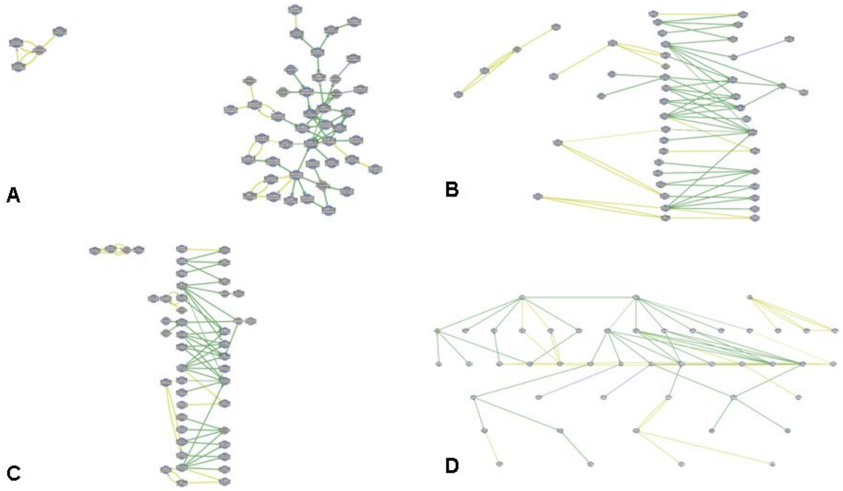


Figure 5: Different Layouts of data from Brady et al., 2011. (A) Cose-Blinket (B) Spread and Klay (C) Klay (D) Breadth

Sidebar Design

The menu side bar was redesigned with Figma. The design as seen in Figure 6 was implemented in the aGENT alpha.

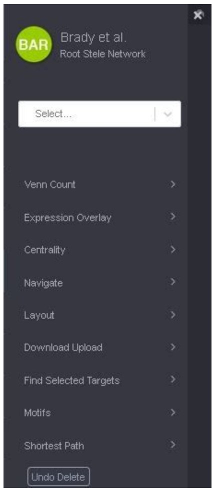


Figure 6: UI design for the menu sidebar.

Gene Search Box

The gene search box allows the user to search for a gene by name so that it can be selected in the network. Informal names that exist in the BAR's gene summaries API (https://bar.utoronto.ca/interactions2/cgi-bin/gene_summaries_POST.php) are supported.

Discussion

With the successful curation of GRNS and implementation of exploratory features, researchers are easily able to explore networks and generate hypotheses. The potential usefulness of these features can be seen when walking through a potential average use case for aGENT. In this case the user has found the GRN by searching for a particular gene of interest in the ePlant interface. The GRN will be laid out in an optimal layout and in future the gene search box will be used to highlight the desired gene. From here the user may easily quantify the gene’s connectedness in the network (via centrality), view the types of interactions (via Find Selected Targets), or step through the motifs to see the regulation patterns. If the network is dense, and the user wants to see if another gene is interacting with their gene of interest they could use the shortest path feature. If results of interest are found, the user can also download interactions in .sif or csv format to perform their own external analysis. Clearly, these aGENT features allow for exploratory analysis and increase the accessibility of GRN data.

One continuing challenge of this project is maintainability and library dependencies. aGENT depends upon several libraries and APIs that are external to the BAR. Therefore, should support for the libraries stop being updated or the APIs go down, several features or even the base functionality of aGENT could be affected. Thus, there must be continual work to maintain the integrity of the codebase. Another way to increase ease of maintainability is to add unit testing and to maintain good commenting practises for the future developers. In addition, the data curated by aGENT needs to be modern and therefore, continual manual annotation and upload to the database is needed to stay apace with the literature.

Furthermore, although the user interface has been improved through certain design choices and including default layout and gene search box, more work can be done to be done to improve the user experience. The creation of demonstration content and tooltips explaining the biological contexts and use cases of particular features would lower the barrier of entry for researchers who may not be familiar with graph theory. In addition, refining the information the user is allowed to access could potentially make the interface more usable. In particular, this includes cleaning up the motif viewer such that only motifs with biological implications with a particular significance are explorable. The next step in the development process is to refine the user interface through user testing. This will allow aGENT to be lightweight and to focus on effectively delivering a smaller number of highly refined features.

Clearly visualization of gene regulatory networks has great potential to facilitate research hypotheses generation. Although this work was focused on Arabidopsis in the use of GRNs in plant biology this tool is fairly general and could be extended for use in other model organisms, which has implications for other fields of study.

Value in science is in the aggregate and the body of literature as a whole. However, the large amount of individual papers and methods for documenting results can make it difficult to see the broader and important patterns. Thus, there is great value in development of tools to standardize, aggregate and organize similar data. A fully developed tool to aggregate and explore *Arabidopsis thaliana* GRNs will have important implications and aid in identifying significant regulatory nodes in *Arabidopsis* and across plant biology, particularly when alongside other BAR tools.

References

Alon U. Network motifs: theory and experimental approaches. Nat Rev Genet. 2007;8(6):450-61.

Bafna D and Isaac AE. Identification of Target Genes in Breast Cancer Pathway using Protein-Protein Interaction Network. International Journal of Cancer Research. 2017;13 (2): 51-58.

Brady SM, Zhang L, Megraw M, et al. A stele-enriched gene regulatory network in the Arabidopsis root. Mol Syst Biol. 2011;7:459.

Chen X, Li M, Zheng R, et al. D3GRN: a data driven dynamic network construction method to infer gene regulatory networks. BMC Genomics. 2019;20:929.

Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22.

Dogrusoz U, Giral E, Cetintas A, Civril A, Demir E. A Layout Algorithm for undirected compound graphs. Information Sciences. 2009;82(2):980-994.

Espinosa-Soto C, Padilla-Longoria P, Alvarex-Buylla ER et al. A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. Plant Cell. 2004;16(11):2923-39.

Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: a graph theory library for visualisation and analysis. Bioinformatics. 2016;32(2):309-11.

Ikeuchi M, Shibata M, Rymen B, et al. A Gene Regulatory Network for Cellular Reprogramming in Plant Regeneration. Plant Cell Physiol. 2018;59(4):765-777.

Keurentjes JJ, Fu J, Terpstra IR, et al. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc Natl Acad Sci USA. 2007;104(5):1708-13.

Jin J, He K, Tang K et al. An Arabidopsis Transcriptional Regulatory Map Reveals

Distinct Functional and Evolutionary Features of Novel Transcription Factors. Mol Biol Evol. 2015;32(7):1767-73.

Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol. 2008;9(10):770-80.

Kashtan N, Itzkovitz S, Milo R, et al. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics. 2004;22;20(11):1746-58.

Keurentjes JJ, Fu J, Terpstra IR, et al. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc Natl Acad Sci USA. 2007;104(5):1708-13.

Koschützki D, Schreiber F. Centrality analysis methods for biological networks and their application to gene regulatory networks. Gene Regul Syst Bio. 2008;2:193-201.

Mangan S, and Alon U. Structure and Function of the feed-forward loop network motif. Proc Natl Acad Sci. 2003;100(21): 11980-5.

Mayo M, Abdelzaher AF, Perkins EJ, Ghosh P. Motif Participation by Genes in E. coli Transcriptional Networks. Front Physiol. 2012;3:357.

Ouma WZ, Pogacar K, Grotewold E. Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties. PLoS Comput Biol. 2018;14(4).

Park S, Lee CM, Doherty CJ, Gilmour SJ, Kim Y, Thomashow MF. Regulation of the Arabidopsis CBF regulon by a complex low-temperature regulatory network. Plant J. 2015;82(2):193-207.

Rowland MA, Abdelzaher A, Ghosh P, Mayo ML. Crosstalk and the Dynamical Modularity of Feed-Forward Loops in Transcriptional Regulatory Networks. Biophys J. 2017;112(8):1539-1550.

Sakuraba, Y. et al. The arabidopsis transcription factor NAC016 promotes drought stress responses by repressing AREB1 transcription through a trifurcate feed-forward regulatory loop involving NAP. Plant Cell. 2015;27(6):1771–1787.

Sonawane AR, Platig J, Fagny M, et al. Understanding Tissue-Specific Gene Regulation. Cell Rep. 2017;21(4):1077-1088.

Smit M, McGregor S, Sun H, et al. A PXY-Mediated Transcriptional Network Integrates Signaling Mechanisms to Control Vascular Development in Arabidopsis. Plant Cell. 2020;32(2):319-335.

Sparks EE, Drapek C, Gaudinier A, et al. Establishment of Expression in the SHORTROOT-SCARECROW Transcriptional Cascade through Opposing Activities of Both Activators and Repressors. Dev Cell. 2016;39(5):585-596.

Vidal EA, Álvarez JM, Moyano TC, Gutiérrez RA. Transcriptional networks in the nitrate response of Arabidopsis thaliana. Curr Opin Plant Biol. 2015;27:125-32.

Waese J, Fan J, Pasha A, et al. ePlant: Visualizing and Exploring Multiple Levels of Data for Hypothesis Generation in Plant Biology. Plant Cell. 2017;29(8):1806-1821.

Zhang F, Liu X, Zhang A, Jiang Z, Chen L, Zhang X. Genome-wide dynamic network analysis reveals a critical transition state of flower development in Arabidopsis. BMC Plant Biol. 2019;19(1):11.

Zhang J, Eswaran G, Alonso-serra J, et al. Transcriptional regulatory framework for vascular cambium development in Arabidopsis roots. Nat Plants. 2019;5(10):1033-1042.

Zhiponova, M. K. et al. Helix--loop--helix/basic helix--loop--helix transcription factor network represses cell elongation in Arabidopsis through an apparent incoherent feed-forward loop. Proc Natl Acad Sci. 2014;111(7): 2824–2829.