Visualizing Protein-Protein and Gene Regulatory Networks in *Arabidopsis thaliana*

by

Vincent Lau

A thesis submitted in conformity with the requirements
for the degree of Master of Science

Department of Cell and Systems Biology
University of Toronto

Visualizing Protein-Protein and Gene Regulatory Networks in *Arabidopsis thaliana*

Vincent Lau

Master of Science

Department of Cell and Systems Biology
University of Toronto

2020

Abstract

Protein-protein interaction networks (PPNs) and gene regulatory networks (GRNs) visualize protein-binding and gene regulation respectively. Advancements in screening technologies such as enhanced yeast-one-hybrid and protein binding predictions have allowed for high throughput determination of such networks. Currently, Arabidopsis-based GRN and PPN viewers are restricted to ad-hoc networks and do not link out to external sources. I created two interactive web-based viewers, Arabidopsis Interactions Viewer 2.0 (AIV2) and Arabidopsis GEne Network Tool (AGENT) to visualize PPNs and GRNs respectively. Features include dynamic interaction filtering, uploading and downloading network data, subcellular-focused network layouts, and loading Bio-Analytic Resource's (BAR) expression data onto genes. Data visualization principles are applied such as rapid serial visual presentation (RSVP) and Shneiderman's mantra of details-on-demand. Moreover, AGENT is a framework for GRN curation which allows future expansion. Lastly, I demonstrate how to use these tools for hypothesis generation.

Acknowledgments

What a wild ride this project was in terms of being exposed to different fields and people! First, I would like to express deep gratitude to Dr. Nicholas Provart for his mentorship and giving me the opportunity to work at the Bio-Analytic Resource (BAR). I never thought I would learn so much under his wing as Dr. Provart's work has incredible breadth. To illustrate, in no particular order, I will acknowledge a list of people who gave me insights in their fields and/or guided me through my project. I thank Dr. Jamie Waese whom gave me constructive feedback on how to design user interfaces based on data viz theory. I am grateful towards my committee, Dr. Shelley Lumba and Dr. Alan Moses who supported me and gave me advice on what focus on when visualizing a network. I also thank Dr. Gary Bader for guidance on the best practices of storing biological networks. I also want to acknowledge Max Franz's wonderful web-based homegrown graph theory library, Cytoscape.js! Max was always available via e-mail to answer software questions, something you cannot always get with other programming packages. Last but not least, I thank Dr. Siobhan Brady and her lab for inviting me to UC Davis which gave me the opportunity to gain their insight of large-scale plant work.

I also cannot forget to acknowledge the Provart Lab members (Anna van Weringh, Michael Dong, Matthew Cumming, Alex Sullivan, Eddi Esteban, and Asher Pasha) for their advice and company. I would like to especially like to thank Asher for all his funny political memes and "LINUXMASTERRACE" style humour! On a more serious note, I learned a lot about server administration when solving problems with Asher. For that, I am grateful, but I will never be a full Linux guy, Asher! I would also like to thank Rachel Woo for her help in implementing some features in the gene network viewer such as the motif finder and shortest-path highlighter. You learned fast and I recognize your contributions to the project.

Last but not least, I would like to thank my family and friends for their support during completion of this project. Special thanks to Kevin Xue, Angelica Miraples, and Sonhita Chakraborty for all the bonding and drinking us grad students do. Finally, I want to thank my family again for their unwavering support through my academic journey as a first-generation university graduate.

Table of Contents

List of Tables

List of Figures

Abbreviations

AD - activation domain
AGENT - Arabidopsis GEne Network Tool
AGI IDs - Arabidopsis Gene Identifiers
AIV - Arabidopsis Interactions Viewer
AIV2 - Arabidopsis Interactions Viewer 2
AJAX - asynchronous JavaScript and XML
API - application programming interface
AuxRE - auxin response element
BAR - Bio-Analytic Resource
BIFC - bimolecular fluorescence complementation
BIP - Binding Protein
C1-FFL - coherent type-1 FFL
CF-MS - Co-fractionation MS
ChIP-Seq - chromatin immunoprecipitation sequencing
CORS - Cross-Origin Resource Sharing
DAP-Seq - DNA affinity purification sequencing
DBDs - DNA-binding domains
DOIs - digital object identifers
ER - endoplasmic reticulum
ERD - entity relationship diagram
ERSE - endoplasmic reticulum stress-responsive element
eY1H - enhanced Y1H
FFL - feed-forward loops
FFT - Fast Fourier Transform
FT - Flowering Locus T
gDNA - genomic DNA
GEO - Gene Expression Omnibus
GO - Gene Ontology
GRNs - gene regulatory networks
GTFs - general transcription factors
HTML - Hypertext Markup Language
IMEX - International Molecular Exchange
JSON - JavaScript Object Notation
MI - Molecular Interaction
miRNAs - micro-RNAs
MS - mass spectrometry
NF-Y - Nuclear Factor Y
ORC - origin recognition complex
ORFs - open-reading frames
PCC - Pearson correlation coefficient
PDB - Protein Data Bank
PDIs - protein-DNA interactions
Pol II - RNA polymerase II
PPIs - protein-protein interactions
PPNs - protein-protein interaction networks
PSICQUIC - Proteomics Standard Initiative Common QUery InterfaCe

PWMs - position weight matrices
qPCR - quantitative polymerase chain reaction
REST - Representational state transfer
RFD - random forest decision
RMSD - root mean square deviation
RSVP - rapid serial visual presentation
S-PPI - structure-based PPI
SIF - simple interaction format
SPA - single-page app
SUBA4 - SUBcellular localization database for Arabidopsis proteins
SVG - scalable vector graphics
TAIR - The Arabidopsis Information Resource
TAP - tandem affinity purification
TBP - TATA-binding protein
TF - transcription factors
TFBSs - transcription factor binding sites
TSS - transcription start site
UI - user interface
UPR - unfolded protein response
URL - Uniform Resource Locator
UX - user experience
Y1H - yeast-one-hybrid
Y2H - yeast-two-hybrid

List of Appendices

Chapter 1
Introduction

1 Introduction

1.1 Protein-Protein Interactions

1.1.1 Introduction to Protein-Protein Interactions

Our understanding of protein biochemistry is becoming increasingly more intricate as large-scale proteomics experiments have been undertaken. Proteins can also bind to other proteins to form protein complexes or activate signalling events. Such affinity events are called protein-protein interactions (PPIs) where multiple identical or non-identical protomers (protein subunits) form homo- or hetero-oligomers respectively. These interactions may drastically change the protomers' native conformation (Nooren and Thornton, 2003). Additionally, PPIs can be categorized by their lifetime in a complex (transient vs permanent associations) (Nooren and Thornton, 2003). Examples of PPIs being an important area of study in *Arabidopsis thaliana* (hereafter denoted as Arabidopsis) are when transcription factors form complexes to regulate key flavonoid biosynthetic genes or how bacterial effector proteins disrupt critical immune response PPIs (Bhattacharjee *et al.*, 2011; Xu, Dubos and Lepiniec, 2015). Moreover, PPIs are also important to those who study gene expression, as PPI formation is controlled via gene expression and localization (Nooren and Thornton, 2003). Although PPIs are often studied in isolation, say those belonging to a pathway of interest, attempts have been made to estimate their quantity in various species.

In a 2003 review, Andrei Grigoriev analyzed large-scale proteomic data and estimated that there were approximately 16,000-26,000 PPI pairs in the yeast proteome excluding PPIs which formed homo-oligomers (Grigoriev, 2003). For comparison, the Arabidopsis' "…[protein]

interactome is estimated to be larger than yeast” (Arabidopsis Interactome Mapping Consortium, 2011) at approximately 300,000 binary PPI pairs excluding homo-oligomeric PPIs. The consortium used a collection of Arabidopsis open-reading frames (ORFs) to test potential pairwise interactions between ~8,000 ORFs using a yeast-two-hybrid (Y2H) pipeline which was evaluated for accuracy against a positive reference set. They then found ~6,000 binary interactions between ~2,700 proteins.

1.1.2 Techniques to Screen Protein-Protein Interactions

As our work is focused on large-scale data visualization and as the Arabidopsis protein interactome was estimated to be relatively large, we will focus on higher screening throughput methods. Many of the experiments in PPI databases come from such methods. In fact, according to the IntAct molecular database, 77% of their PPIs are sourced from large-scale experiments (> 100 PPIs) such as Y2H (Dong and Provart, 2018).

One common high throughput technique for detecting PPIs in Arabidopsis is the Y2H method. The Y2H technique was invented in 1989 from the idea of reconstituting a split transcription factor’s activity from its two domains (DNA-binding and activation domains) by the interaction of heterologous fusion proteins (Fields and Song, 1989). In their seminal paper, Fields and Song (1989) used known yeast protein interactors, SNF1 and SNF4 fused to the DNA-binding (usually the ‘bait’) and activation domains (usually the ‘prey’) of GAL4 to upregulate the expression of β-galactosidase. Verification of the putative interacting partners binding was then verified by inspection of galactose metabolism (i.e. blue/white screening). Since then, the Y2H method has been adapted for high throughput genomic scale use by optimization, automation, commercialization of domain-tagged libraries, and miniaturization, which was pivotal in early systems biology experiments (Fields, 2005). Although the Y2H technique has been improved for speed and ease, there remains the caveat of a high error rate (Von Mering *et al.*, 2002; Huang, Jedynak and Bader, 2007). First, as this system has been optimized for yeast, Arabidopsis bait-prey interactions cannot be assumed to directly translate *in planta*, leading to a high false negative rate which is estimated to be between 75% to 90% (Huang, Jedynak and Bader, 2007). Second, some baseline transcriptional activity may occur independent of bait-prey binding. Last, some proteins cause transcription of the reporter on their own. However, Fields (2005) argues that most Y2H screens that involve a follow-up validation (i.e. another technique such as co-immunoprecipitation) can eliminate the most types of false positives and those arising from background activation are fairly reproducible and can also be removed. Vidal and Fields (2014) further argue in a follow-up commentary that "… among all highly reliable interactions published by the scientific community, upwards of three quarters are supported by at least one yeast two-hybrid experiment". In this commentary, Vidal and Fields (2014) discuss how the Y2H technique will be pivotal in ushering a new age of interactomics akin to how DNA-sequencing technology ushered in an age of genomics. Indeed, the authors specifically mention how Y2H complemented with tandem affinity purification (TAP)-mass spectrometry (MS) can give rise to large, high-confidence protein-protein interactomes. Indeed, the Arabidopsis protein-protein interactome has been extensively mapped in large part to these two techniques (Van Leene *et al.*, 2007; Arabidopsis Interactome Mapping Consortium, 2011). Therefore, Y2H can be a valuable initial screening technology. Indeed, Y2H experiments are often further validated by more focused experiments *in planta.* For example, Lumba *et al.* (2014) used bimolecular fluorescence complementation (BIFC), which is a technique whereby two proteins fluoresce when in very close contact to another to validate a putative Y2H interaction between SNRK3.15 and SNRK3.22, which are key kinases in abscisic acid signalling.

As an alternative and complement to Y2H, TAP is able to distinguish in multi-protomer protein-protein interactions unlike, Y2H which can only screen for binary partners. In 2002, TAP was first applied in a large-scale project to scan for multiprotein complexes in yeast by creating a fusion protein of interest which can be purified twice (i.e. in ‘tandem’) to elute the protomers of interest (Gavin *et al.*, 2002). The authors then used mass spectrometry to identify the protomers of 232 protein complexes. Moreover, they were able to create a PPI network which highlights proteins that can complex with another. In this seminal paper, they used data visualization techniques such as colouring-coding proteins according to their cellular roles “… for the generation of hypotheses for further investigations”, which is a central theme to this thesis. Additionally, TAP can test for the existence of multimeric complexes *in planta* unlike Y2H. However, the inventors caution that the tandem tag may not be specific to the elution beads once expressed *in vivo* and/or that the tag may interfere with binding or expression of the fusion protein (Puig *et al.*, 2001). The technique has since then been modified and optimized for Arabidopsis by modifying its tag cleavage site, in combination with better annotated Arabidopsis databases for MS analysis (Rubio *et al.*, 2005; Van Leene *et al.*, 2015). In terms of disadvantages, TAP requires *pre priori* knowledge and requires fusion constructs of the proteins of interest.

In contrast to creating fusion constructs, protein microarrays are glass slides on which proteins are orderly immobilized to allow large high-throughput protein-protein interaction analysis (Hu *et al.*, 2011). The chip is incubated with labeled proteins and a signal determines if an interaction has occurred. Popescu *et al.* (2007) used a 1113 Arabidopsis protein microarray to identify calmodulin-binding partners and found many partners, such as kinases and F-box proteins, thereby establishing that protein chips can be used for Arabidopsis. However, as the affinity events are performed *in vitro*, these results may not be as reflective of the biology as

TAP. Additionally, large-scale protein production for the microarrays can be costly and labour-intensive when dealing with larger and more complex proteomes such as in Arabidopsis, which can limit the throughput of protein microarrays (Hu *et al.*, 2011).

Reviewing these techniques reveals that each has its own advantages and disadvantages. Namely, although methods such as microarrays and Y2H are high throughput, they suffer from being performed outside of the native cellular environment. Recently, co-fractionation mass spectrometry was used in multiple plant species including Arabidopsis to uncover known and unknown plant protein complexes (McWhite *et al.*, 2019). Co-fractionation MS (CF-MS) is a method of simply applying mass spectrometry to precise biochemical protein extract fractions that have been separated by chromatographical methods (size, ion exchange, isoelectric). It is also a high-throughput PPI detection method that does not require creating constructs with protein fusions/tags and detects PPIs in the native cellular extract. McWhite *et al.* (2019) compared their CF-MS results to plant Y2H plant databases and found high concordance of CF-MS results with Y2H, especially when the interactions were seen across multiple plant species. Proximity based labeling methods such as BioID also exist where a promiscuous biotin ligase is fused to a protein of interest to biotinylate proximal (and likely interacting) proteins (Kim and Roux, 2016) to be identified via MS. Although not as high throughput as CF-MS, BioID can identify interactors in their physiological conditions in living cells before cell lysing unlike CF-MS . Recently, Khan *et al.* (2018) used BioID in Arabidopsis to validate this technique against previously identified interactors of HopF2 (a membrane-targeted, type III secreted effector).

There are also *in silico* PPI prediction methods that utilize evolutionary, sequence, or structural information to infer potential PPIs (Rao *et al.*, 2014). Evolutionary based methods can apply the similarity of phylogenetic trees across different protein families to estimate protein-

protein interactors (Pazos and Valencia, 2001). In this study, the authors hypothesized that the phylogenetic trees of protein partners would be more similar than what would be expected under the standard molecular clock due to coordinated evolution. They used multiple sequence alignment (ClustalW) across several prokaryotic species to create distance matrices which are then compared to other matrices for a similarity score. In reference to a known true positive prokaryote protein interaction set, they found that most truly interacting proteins had high similarity values.

Another evolutionary approach that is to use established interactions in other related species to infer interactions in a species of interest where the known pair of protein interactors are known as 'interologs'. Geisler-Lee *et al.* (2007) was the first to implement an interolog-based Arabidopsis interactome by identifying orthologs in Arabidopsis to yeast, fruitfly, worm and human. They then predicted 19,979 Arabidopsis interactions for 3,617 proteins based on established interactome databases for the four species. For quality control of these predicted interactions, Geisler-Lee *et al.* (2007) ranked them according to how many databases, species, and experiments within the databases supported the interologs. They then compared the results to established networks and applied co-expression & subcellular localization data (under the assumption that interaction partners would be similarly expressed and share the same localization) for further validation. Visualization of the predicted interactome was then summarized in a web-based interactions viewer, the Arabidopsis Interactions Viewer (AIV; see Figure 1.1; http://bar.utoronto.ca/interactions), which is hosted on our plant-centric bioinformatics webserver, the Bio-Analytic Resource (BAR).

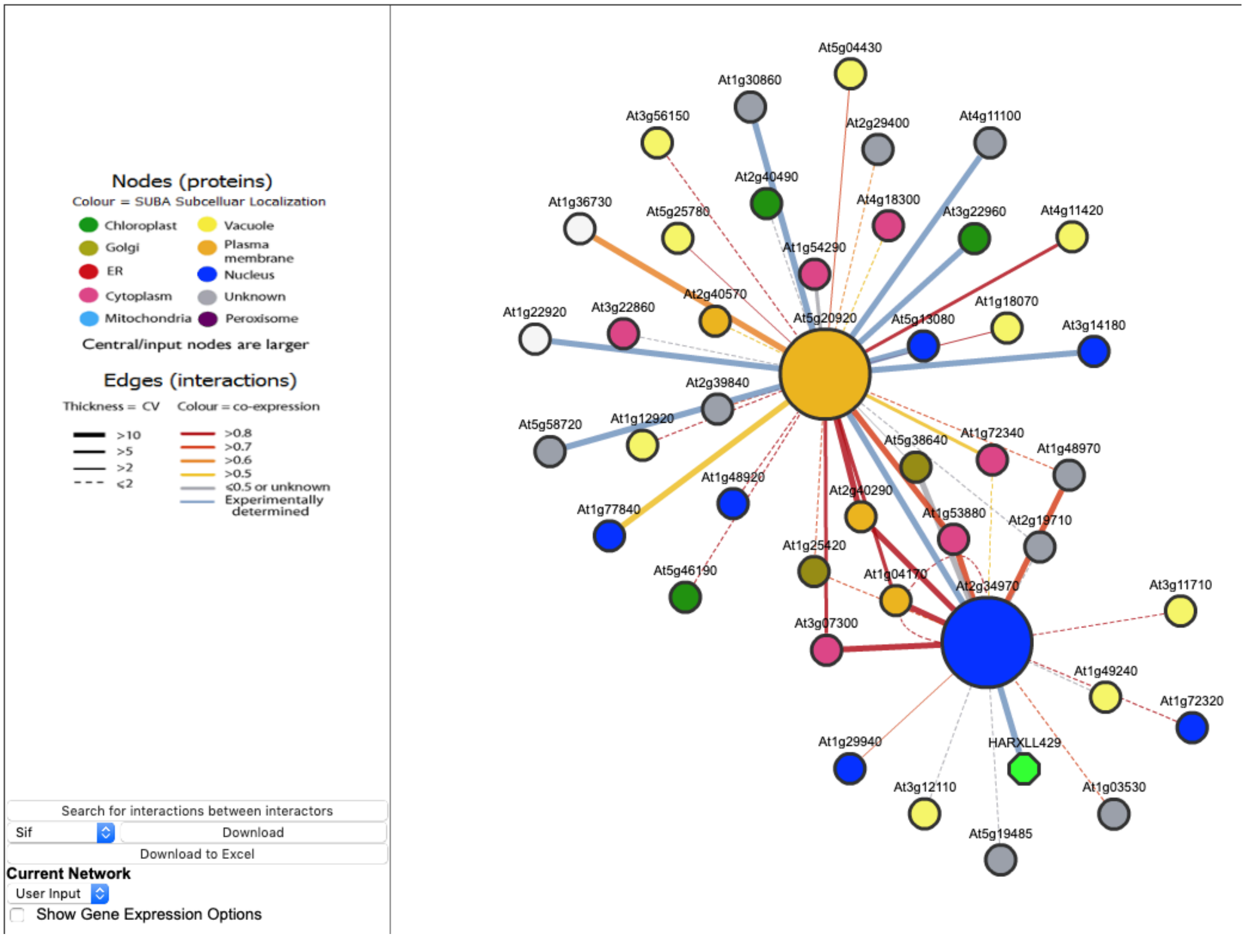


Figure 1.1. Default Arabidopsis Interactions Viewer (Geisler-Lee *et al*., 2007) output when user selects the example to visualize the PPIs of AT5G20920 (EIF2-Beta) and AT2G34970 (Trimeric LpxA-like enzyme).

Sequence based prediction approaches typically assume key sequence homology across genomes, for example key binding domains that allow protein interaction (Shoemaker and Panchenko, 2007; Rao *et al.*, 2014). One computational method utilizing domain information is to train classifiers such as the Random Forest Decision (RFDs) algorithm to distinguish between true interactors and non-interactors. Chen and Liu (2005) used yeast PPI data from the Database of Interacting Proteins and domain data from Pfam to build decision trees using RFD. Specifically, they decided that each domain can be represented by a feature vector. Each feature is represented by a ternary value (which builds a ternary tree) - that is either none-, one-, or both of the putative interaction members contain that specific domain. The RFD algorithm generates many decision trees by random sampling the domain data, which then generates a final classification (interaction or no interaction) based on voting of the trees. Using RFD, the authors were able to outperform a prior maximum likelihood approach (Deng *et al.*, 2002) in specificity (64% vs 38%) in a 8,917 protein pair data set, where the specificity was defined as the percent of predicted, true non-interactions over the total of true non-interactions. Recently, Ding and Kihara (2019) also used RFDs to predict PPIs in Arabidopsis, soybean and corn. However, they used other types of protein information for feature representation such as co-expression and protein function. By training and testing on The Arabidopsis Information Resource (TAIR) Arabidopsis PPI dataset, they were able to outperform the widely popular STRING database which curates predicted PPIs from the literature/databases and generates PPIs from orthology (Szklarczyk *et al.*, 2019).

Structure based techniques utilize the 3D structure of the proteins to predict protein interactions (Rao *et al.*, 2014). A structural technique of particular interest is protein docking whereby a protein complex is calculated from its protomers (Macindoe *et al.*, 2010). A docking program, HEX (http://hex.loria.fr/) utilizes fast Fourier Transform (FFT) to find potential docking orientations (Macindoe *et al.*, 2010). The FFT is a method of calculating molecular surface three-dimensional overlap by utilizing Fourier transformation by of a correlation function (Katchalski-Katzir *et al.*, 1992). Its algorithm speed was then improved by 45-fold when the HEX authors implemented the algorithm to be run on graphic processor units, which can perform multi-thread and matrix-operations much more quickly and using a spherical Fourier correlation to encode the protein surface shape (Ritchie and Kemp, 2000; Ritchie and Venkatraman, 2010). More generally, HEX rotates two proteins a desired number of degrees in three dimensions to find the most optimal (most low-energy) conformations. In our lab, we recently employed HEX in Arabidopsis to expand on the relatively limited Arabidopsis protein-protein interactome (Dong *et al.*, 2019). As HEX requires Protein Data Bank (PDB) files as inputs and Arabidopsis' entries in the PDB is limited (587 genes), we first constructed a predicted 'structure-ome' using Phyre2, which is a multi-stage protein folding prediction tool utilizing multiple sequence alignment and Hidden Markov Models (Kelley *et al.*, 2015). Dr. Geoffrey Fucile was able to predict ~84% (29,180 structures) of the Arabidopsis proteome. Co-author Dr. Michael Dong then filtered the 1,346 best structures by Phyre2 identity score and cross-validated the predictions to published structures by calculating the root mean square deviation (RMSD) value between them. By validating that there is a short RMSD distance (2.59Å on average) between the predicted and published models, he confirmed that the predicted structure-ome is valid for HEX processing. From these 1,346 structures, there are 906,531 possible binary interactions for which he took the top 1% best scoring interactions to generate and visualize a structure-based PPI (S-PPI) network. Like the Geisler-Lee *et al.*, (2007) paper, he then chose to validate these interactions by assessing if they are colocalized in the same subcellular compartment. He found significant enrichment for S-PPIs to both be localized in either the nucleus, Golgi, endoplasmic reticulum, peroxisome and plasma membrane. Finally, he experimentally validated a subset of the top interactions via Y2H against randomly chosen interactions. He showed that Y2H was able to confirm his HEX-predicted interactions much better than an interaction-pair selected at random (26% vs 10%). One may note that 26% is a low figure, however he showed previously published interactions determined by Y2H also show moderately low levels of Y2H confirmation as well (45%). As discussed, the Y2H methodology has its disadvantages such as differences in libraries used and interactions being tested in a non-native cellular environment. Last, my contributions as a co-author were to update the aforementioned AIV to help visualize these predicted interactions as we now can visualize ~12,000 proteins' PPIs, amongst other improvements, which I will cover in in the next chapter.

1.1.3 Visualizing Protein-Protein Interaction Networks

From publications using such techniques, we can amalgamate all the interactions determined into a large collection, which can then be analyzed and visualized. This is an important theme of this thesis as high-throughput techniques are continually generating larger data set, which Kaminski (2000) even prior to higher throughput platforms such as sequencing observed "…have created an information overload". When Y2H began producing large numbers of PPIs, those PPI collections were often pictorialized by a mathematical network where proteins are represented by nodes (circles) and interactions between them are represented by edges (lines) (Seebacher and Gavin, 2011). A large 'birds-eye view' of the PPI network allows for readers to easily estimate the size and interconnectedness (how often nodes interact with another) of the PPI network. However, researchers are often also concerned with individual members of the network, enrichment of particular protein functions/members, key subnetworks, localization, complex membership, and biological relevancy of the network (Tucker, Gera and Uetz, 2001).

For instance, in the case of the aforementioned Arabidopsis predicted interactomes from Geisler-Lee *et al.* (2007) and Dong *et al.* (2019), these networks were not only visualized as the networks themselves but with additional analyses and decorations to emphasize a particular point. For example, in both papers the authors organized and coloured the protein nodes via their subcellular localization to suggest colocalization and thus support for the predicted PPIs. Alternatively, one can integrate a given PPI collection to another PPI network by expanding an existing network. Geisler-Lee *et al.* (2007) showed the utility of importing external data such as PPI databases to provide further support for the predicted interactions. Specifically, the authors expanded a known SNARE-syntaxin network with their predicted PPIs and found expected & unexpected new protein partners. Thus, the ability to import external data and cross-validate PPIs is especially important given that techniques such as Y2H and *in silico* methods have a number of issues. Lastly, Geisler-Lee *et al.* (2007) was one of the first to discuss protein hubs in Arabidopsis. They defined hubs as highly connected nodes (i.e. proteins that have many interactions) or more formally, nodes that have a high degree centrality. Earlier definitions of Arabidopsis protein hubs deemed that the histone phosphotransfer proteins act as hubs as they were highly connected and interacted with two major groups of a cytokinin signalling network (Dortay *et al.*, 2006).

Earlier seminal work on a yeast Y2H PPI networks performed by Jeong *et al.* (2001) suggested that PPI networks are typically *scale-free*, that is that a relatively few number of nodes are highly connected and that these nodes are essential. Indeed, Geisler-Lee *et al.*, (2007) showed that only a very small fraction of proteins were highly connected (51+ interactions). However recent reviews argue that hub proteins correlate more to the number of biological functions such proteins are involved than to lethality. There seems to be two classes of hubs: date hubs help bridge functions between functional subnetworks and party hubs complex with many members of the same process (Yu *et al.*, 2007; Vallabhajosyula *et al.*, 2009; Song and Singh, 2013; Vandereyken *et al.*, 2018). Despite this controversy, degree centrality is an incredibly intuitive concept that many biologists unfamiliar with network biology can understand and that can be a guiding principle to help identify important proteins in a large network. Vallabhajosyula *et al.* (2009) cleverly suggested increasing node size in proportion to the node's to degree centrality to highlight highly connected nodes visually.

Another technique to organize a PPI network is to decorate and/or cluster proteins according to their biological function. Schwikowski, Uetz and Fields (2000) coloured protein nodes by their functional role in a large 1,548 node PPI network to demonstrate that proteins of similar functions typically cluster together, as one might expect. One can then extend this idea to hypothesizing functions to uncharacterized proteins by the function of the interaction partners (Tucker, Gera and Uetz, 2001). Yu *et al.* (2008) used a multifaceted computational approach (interologs, co-expression, enriched domains) to construct a large PPI network based on key chloroplast proteins to determine and validate that the unannotated AT1G52220 protein interacts with a key photosystem I subunit as predicted in their framework. Lastly, De Bodt *et al.* (2009) segmented network clusters on an interolog-mapped Arabidopsis interactome and performed Gene Ontology (GO) biological process enrichment on each cluster. They found that the cell cycle clusters were not only highly enriched for GO terms involved in DNA replication but the protein members were highly co-expressed with another, which they hypothesized was due to "tight regulation of proteins involved in DNA replication". This example shows that we can also integrate expression data in our visualization and analysis of PPI networks. The assumption is that interacting partners are co-expressed if a protein interaction is key for a particular biological process, i.e. co-expression could be evidence for a particular PPI (Bhardwaj and Lu, 2005). However, some nuances exist, such as the fact that interactions between constitutively expressed proteins and transiently expressed proteins would not have a high level co-expression, in contrast to tightly regulated proteins involved in the cell cycle as mentioned above (Geisler-Lee *et al.*, 2007). Nonetheless, Geisler-Lee *et al.* (2007) found that their predicted PPI members exhibited much more co-expression than random PPIs. Co-expression between protomers is often visualized by colouring edges by their correlation efficient in a bi-coloured fashion. For example, Boruc *et al.* (2010) used a red/blue heat map scale to delineate between inversely or positively co-expressed cell cycle proteins in their PPI network as shown in Figure 1.2. However, linear colour changes in colour maps are not perceived uniformly leading to biases when judging values (Moreland, 2009). As an alternative, De Bodt *et al.* (2009) used only two discrete colours to distinguish between interactions that either did or did not surpass their *ad hoc* correlation threshold.

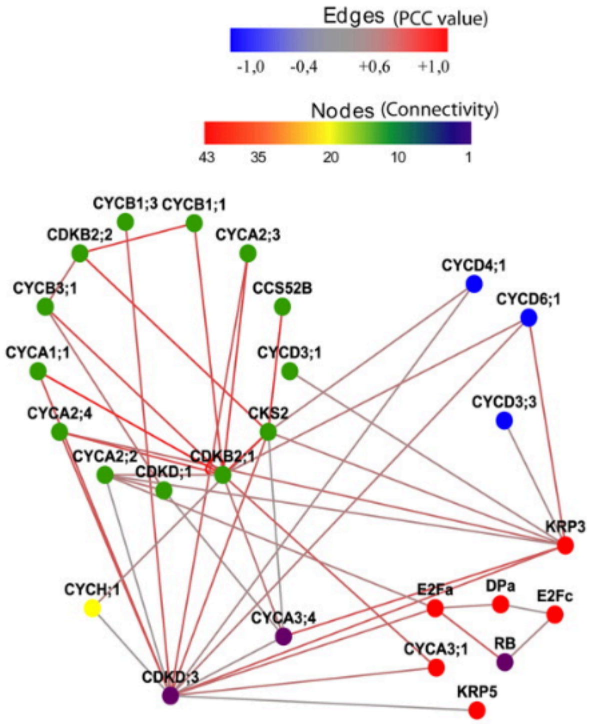


Figure 1.2. PPI network between Arabidopsis core cell cycle proteins as examined in (Boruc *et al.*, 2010).

1.1.4 Web-based Arabidopsis Protein-Protein Interaction Network Tools

Until now I have only discussed PPI networks from static figures in Arabidopsis focused studies. However, interactive bioinformatic tools exist to assist researchers in querying PPI databases and in visualizing PPI networks. As there are over 250 PPI resources as of June 2017 (Dong and Provart, 2018), this chapter will focus on Arabidopsis-data-containing resources as summarized in Table 1.1.

One notices from Table 1.1 that there are few *live* Arabidopsis-centric resources available for querying PPIs. Another observation is that the Arabidopsis-centric resources pull data directly from the consortium-level databases such as BioGRID which leads to large overlaps of PPIs. This is unlike large consortium-level databases which have little overlap (Dong and

Provart, 2018) as they may collaborate under certain rules (such as IMEX – International Molecular Exchange) to curate different journals. Although some of these major databases hosts webservers which feature a network viewer, many of them are simple. For example, INTACT (https://www.ebi.ac.uk/intact/) simply shows all binary interactors for a given gene and does not integrate species-level data like expression data, functional information and/or localization (Kerrien *et al.*, 2012). Nor does the viewer have any additional functions beyond reorganizing the network layout, such as filtering on the interaction methodology. Last, many of the tools host export functionality which output a given list of interactions to be visualized in a desktop application, Cytoscape (Su *et al.*, 2014). Although Cytoscape is very well-featured and has many plug-ins, it requires installation along with JAVA which can be troublesome for users with limited experience or access to administrator-level privileges. Fortunately, the Cytoscape Consortium has created a web-based package using Flash technology, Cytoscape Web, to allow developers to create their own web applications to visualize networks (Lopes *et al.*, 2011). This package has many of the desktop-level features such as panning, zooming, and clustering networks along with decorating nodes and edges.

Indeed, AIV and INTACT use Cytoscape Web to visualize their PPIs. AIV in particular took advantage of the package's features: colouring nodes according to expression-levels or localization, decorating nodes with MapMan (functional) codes, colouring edges according to co-expression, and organizing the nodes according to their localization. Unfortunately as Flash technology will be deprecated in major browsers by 2020 (Bradshow, 2019), this means there would be no functioning Arabidopsis-focused PPI network viewer by 2020 outside of ePlant (see Figure 1.3), which we also host and is not as well-featured for viewing PPIs. Furthermore, Dr. Michael Dong's predicted PPIs were not included in AIV. Happily, the Cytoscape Consortium has created a new JavaScript-based web-package, Cytoscape.js that has new features such as more styling options for nodes/edges, smartphone compatibility, and modern user-interface (UI) elements such as tooltips to replace Cytoscape Web (Franz *et al.*, 2016). Hence, given AIV's reputation (~80,000 views to date in 2019; https://bar.utoronto.ca/awstats/awstats.pl) it is imperative to create a modern, browser-compatible update to AIV to highlight our newly predicted PPIs, following modern user-experience, software-engineering, and data-visualization principles to enable hypothesis generation with our large, extensively curated collection of PPIs.

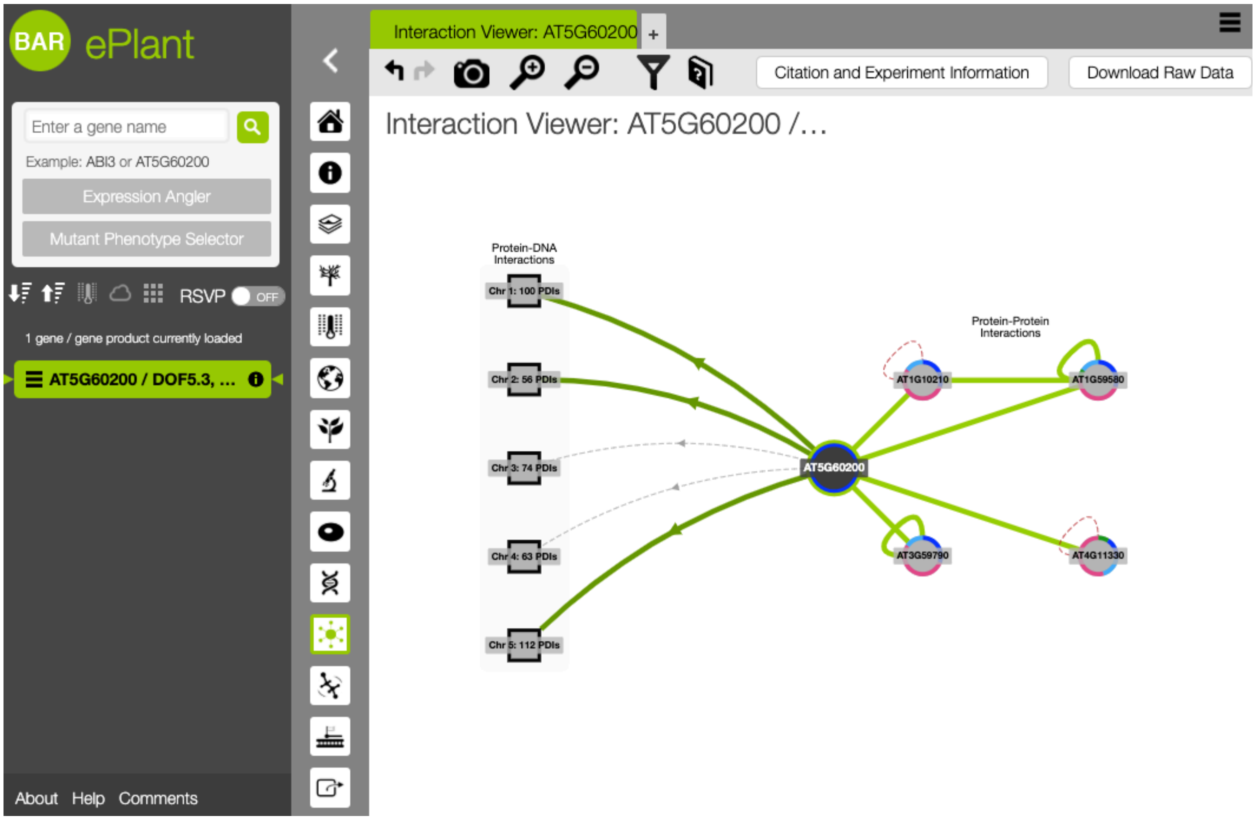


Figure 1.3. ePlant (Waese et al., 2017) output when the gene, AT5G60200 (ABI3) is entered. The interactions viewer was selected and is shown in the rightmost pane. Circular nodes represent proteins while squares are chromosomal DNA nodes which show all the protein-DNA interactions when a user clicks on the node.

1.2 Protein-DNA Interactions

1.2.1 Introduction to Protein-DNA Interactions

In Arabidopsis, it is estimated that there are approximately 2000 transcription factors (TF)s, proteins that bind to DNA and regulate transcription via protein-DNA interactions (PDIs)

(Wehner, Weiste and Dröge-Laser, 2011). Indeed, much like how proteins have specialized domains for binding to protein partners to form complexes, they can also contain DNA-binding domains (DBDs) to bind to regulatory DNA elements, such as promoters, silencers, and enhancers (Riethoven, 2010; Inukai, Kock and Bulyk, 2017). These DNA elements regulate gene expression by different mechanisms. Promoter regions can be classified by their distance from the TSS into core promoters and proximal promoters (Haberle and Stark, 2018). Core promoters contain the transcription start site (TSS; +1) and essential elements for transcriptional initiation, such as a binding platform for the transcriptional machinery comprising general transcription factors (GTFs) and RNA polymerase II (Pol II). A prominent core promoter element is the TATA box, which binds the TATA-binding protein (TBP), which then complexes other subunits to create part of the transcriptional machinery (Riethoven, 2010). The TATA box is often conserved across eukaryotes. As expected, Molina and Grotewold (2005) found that the TATA box is the most overrepresented DNA elements in Arabidopsis core promoters. Importantly, the core promoters typically have low basal activity and thus can be supressed/enhanced by other regulatory elements such as proximal promoters (Haberle and Stark, 2018). Proximal promoters are usually located further upstream from the TSS and have transcription factor binding sites (TFBSs) for sequence-specific TFs to bind (Hernandez-Garcia and Finer, 2014; Haberle and Stark, 2018). It is believed that these TFs interact with the transcriptional complex to modulate transcriptional activity. As Arabidopsis is multi-tissue organism that can be exposed to different physiological stresses, proximal promoters allow for precise spatiotemporal transcriptional control as they are sequence-specific. For example, the auxin-response factors (ARFs) are TFs that bind to the auxin response element (AuxRE), which is a pyrimidine-rich sequence typically located 2bp to 300bp upstream of the TSS of target genes in Arabidopsis (Mironova *et al.*, 2014). ARFs can either activate (upregulate) or repress (downregulate) gene expression depending on their variable middle region (Li *et al.*, 2016). Although the exact mechanism for how ARFs activate transcription is unknown, well-studied TFs such as GAL4 (yeast) recruit GTFs and thus the basal machinery after induction (Traven, Jelicic and Sopta, 2006). Auxin is a plant growth hormone involved in the auxin signalling pathway, which promotes ARFs by degrading auxin repressors which bind and repress the ARFs (Wang and Estelle, 2014). As auxin transport is highly regulated, ARFs' specificity to the AuxRE (Ulmasov, Hagen and Guilfoyle, 1999) allows more precise control of the transcription (and thus development) in a particular tissue(s) relative to basal regulation.

Enhancers are a less-studied class of DNA elements which are located far from the TSS – up to a Mbp away in higher mammals (Levine, 2010). Although the exact mechanism for gene regulation is unclear, one leading theory is that enhancers can recruit the transcriptional machinery assembly from far distances due to DNA looping (Andersson, 2015). One of the few well-characterized enhancers in Arabidopsis is an element of the Flowering Locus T (FT) gene, *Block C* which is 5,000 bp away from the TSS of the gene (Zicola *et al.*, 2019). Zicola *et al.* (2019) found that methylation of *Block C* downregulated FT expression which reduced flowering. Silencer elements are similar to enhancers in that they can be located far from the TSS as well but repress transcriptional activity (Riethoven, 2010). Even less is known about silencers and their mechanisms, hence Weber *et al.* (2016) urged plant researchers to perform more characterization studies given new methods that take advantage of high-throughput sequencing.

1.2.2 Techniques to Screen Protein-DNA Interactions

As the Y2H assay became popular due to its scalability, similar techniques were spun off the related idea of quickly capturing interactions via activation of a TF, namely the yeast-one-hybrid (Y1H) assay for screening PDIs. One of the first descriptions of Y1H were by Li and Herskowitz (1993) in which they cloned constructs that expressed a GAL4 AD that was fused to their TFs of interest (prey) along with a reporter construct that expressed β-galactosidase under the control of their promoter of interest (bait) in yeast. Therefore, TF-DNA binding events (i.e. PDIs) should upregulate galactose metabolism and turn yeast colonies blue for easy screening. The authors screened their putative uncharacterized TFs against a consensus sequence in yeast DNA replication origins. They were then able to characterize a protein that is part of the origin recognition complex (ORC) and dubbed it ORC6. Although they did not perform any further PDI validation, they validated their findings by producing deletions in ORC6 which should and were lethal deletions. Like Y2H, Y1H underwent technical modifications and optimizations to allow high-throughput processing. For instance, the Brady Lab at UC Davis has a 2,000 Arabidopsis TF library that can screen up to 6 promoters which eliminates the need for researchers to create their own cDNA libraries (UC Davis Proteomics Core, 2019). The Brady group also takes advantage of automation by having a robot mate single bait strains with the prey yeast which then reproduces diploid yeast with both constructs – termed enhanced Y1H (eY1H) (Gaudinier *et al.*, 2011; Reece-Hoyes and Marian Walhout, 2012). As a proof-of-concept for Arabidopsis, they used eY1H to test 653 Arabidopsis root stele TFs and were able to verify half of previously validated *in planta* PDIs, which is similar result to other eY1H protocols. Although the accuracy of this technique is lower than traditional transformation techniques, eY1H provides a highly standardized and throughput way of processing PDIs which can be cost- and time-saving. However, Y1H faces many similar limitations to Y2H such as false positives and that interactions are captured in yeast. To combat the former, each PDI in an eY1H assay should be tested at least four times such that only PDIs with at least two technical replicates are counted (Reece-Hoyes *et al.*, 2011; Reece-Hoyes and Marian Walhout, 2012). More importantly, eY1H's throughput (2 weeks vs 2 years) allows researchers to quickly use more intricate methods to validate putative PDIs.

Since Y1H, other techniques such as chromatin immunoprecipitation sequencing (ChIP-Seq) have been developed to take advantage of sequencing's nucleotide-level resolution and speed. ChIP-Seq is a way to assay for PDIs *in vivo* by crosslinking DNA to proteins via formaldehyde where the DNA is then sheared such that crosslinked protein-DNA complexes are protected and can be immunoprecipitated via antibodies to be sequenced (Park, 2009). Software is then used to assign ChIP-Seq reads to peaks to determine where the protein binds. Advantages of ChIP-Seq include base-pair resolution of TFBSs and ability to scan the whole genome, unlike prior ChIP-arrays which had fixed probe sequences. Moreover, ChIP-Seq is an *in vivo* technique that has been optimized to work in Arabidopsis (Cortijo *et al.*, 2018). However, disadvantages include technical artifacts such as uneven shearing, non-specific DNA binding, antibody purification/specificity and sequencing errors. Attempts to remove these artifacts have been shown to be dependent on the type of software and preprocessing used (Carroll *et al.*, 2014). Huang *et al.* (2012) used ChIP-Seq to identify PDIs for TOC1, a gene implicated in the Arabidopsis circadian rhythm which was previously shown to regulate genes depending on the circadian clock. The study found 867 potential genes that TOC1 regulated in which 40% showed differential expression across the circadian clock. Furthermore, they showed that ChIP-Seq peaks showed a rhythmic oscillation where peak binding of TOC1 to its targets occurs soon after light exposure. They also demonstrated ChIP-Seq's nucleotide resolution by finding an enriched sequence amongst TOC1's targets – the evening element which was shown to previously associated with circadian regulation (Harmer, Panda and Kay, 2001).

A more recent technique developed by O'Malley *et al.* (2016) that also utilizes sequencing technology is DNA affinity purification sequencing (DAP-Seq), which attempts to solve some of ChIP-Seq's technical limitations. Specifically, creating antibodies can be challenging or unavailable for certain proteins. Indeed, a rudimentary search performed in

December 2019 on the Gene Expression Omnibus (GEO) with the search term “chip-seq” revealed that there were only ~530 Arabidopsis results versus ~28,000 for human. DAP-Seq exposes sheared genomic DNA (gDNA) with sequencing adapters to bead-bound TFs. The gDNA-TF-bead complexes are washed and eluted until the TF-bound gDNA fraction can be sequenced. Finally, the read peaks are then mapped to the genome. The authors used this technique on 529 TFs to find ~2.7 million TFBSs which covered 9% of the Arabidopsis genome. To validate their methodology, they compared the enriched DNA motifs in their DAP-Seq peaks to curated Arabidopsis motif databases and found a large concordance. Last, they performed a ChIP-Seq experiment on 3 TFs (ABI5, ATHB5, ANAC055) to cross-validate and discovered strong concordance between motifs – especially in highly enriched motifs. Hence, DAP-Seq could be used alternative to ChIP-Seq, particularly in Arabidopsis. Indeed, initial work by Sen *et al.* (2017) performed DAP-Seq on ARFs in corn shows enriched motifs that resemble the AuxRE.

PDIs can also be predicted *in silico* by performing motif enrichment on a given set of genes. Kulkarni *et al.* (2017) developed the TF2Network algorithm by amalgamating several Arabidopsis TFBS databases which in total contain position weight matrices (PWMs) for 1,793 TFs. TF2Network takes as input a list of Arabidopsis Gene Identifiers (AGI IDs) to which enrichment analysis of the TFBSs is performed on the promoters of the AGI IDs. Based on the enrichment analysis of PWMs in those promoters, TFs suspected of regulating those via motifs are returned to the user. To validate this approach, the authors used their algorithm given a list of ChIP-Seq bound-target genes and were able to recover over 75% of the 24 TFs used in the ChIP-Seq experiment. TF2Network also showed promising specificity as randomly selecting 500 unbound ChIP-Seq genes as input returned only 0.17 TFs over multiple runs. The authors also alluded to gene regulatory networks (GRNs) and mentioned the need for interactive tools to visualize GRNs, which will be discussed next.

1.2.3 Visualizing Gene Regulatory Networks

Kulkarni et al. (2017) defines GRNs as “is a collection of regulatory interactions between transcription factors (TFs) and their target genes”. Much like a PPI network is a collection of PPIs, GRNs are synonymous with a collection of PDIs. There is some variation in the definition of a GRN where Emmert-Streib, Dehmer and Haibe-Kains (2014) define GRNs as gene-gene interaction networks strictly inferred from co-expression data. Here, I define GRNs as a *mostly* PDI network that illustrate how TFs and their target genes can regulate and be regulated by other genes no matter if they were generated via experimentation or inference.

For example, Brady *et al.* (2011) was the first to generate a tissue-specific Arabidopsis GRN. The group selected a group of 60 promoters of genes suspected to be highly involved in development of the root stele to be screened via Y1H assay against 167 TFs. They were able to map 46 PDIs along with 28 PDIs of TFs to promoters of micro-RNAs (miRNAs) that were highly expressed in root stele. They also performed a Y2H assay between the 167 TFs to generate 25 PPIs. They were then ultimately able to build an initial GRN that integrated mostly PDIs along with some PPIs. In this GRN, they used edge colouring to distinguish between PDIs and PPIs. To validate the GRN robustness, a dexamethasone assay was performed on a key TF, OBP2 and its downstream targets. OBP2 was found to bind and activate and repress PHB and PHV respectively which are HD-ZIP TFs involved in development and specification. Additionally, quantitative polymerase chain reaction (qPCR) was used to evaluate how mutating the TFs would affect target gene expression. Specifically, they investigated the regulatory relationship (i.e. activation or repression) of the PDI. With these focused experiments, Brady *et al.* (2011) were able to build an *in planta* validated GRN comprising a total of 103 PDIs, PPIs and miRNA-mRNA interactions, 59% of the initial GRN (see Figure 1.4). This GRN is seminal not only in terms of using multiple techniques to screen and validate the different types of interactions in Arabidopsis but in terms of visualizing regulatory relationships. To visualize these relationships, the authors used triangles or perpendicular lines at the end of an edge of the target gene to represent an activating or repressing relationship, respectively. Furthermore, unlike PPI networks whose nodes always represent the same biological entity (i.e. proteins), GRNs represent interactions between different entities (i.e. proteins and DNA). To solve this visualization challenge, the authors employed the "metagene". Liseron-Monfils and Ware (2015) state a metagene is "…a node [that] represents at the same time a protein, its gene and its promoter". This approach versus drawing DNA and protein nodes separately significantly reduces network clutter resulting in a cleaner data visualization. Capturing the regulatory relationships between nodes also allows calculation of in-degree (how many TFs bind to that gene; i.e. a triangle) and out-degree (how many genes a TF targets; i.e. a perpendicular line) centrality (Gaudinier and Brady, 2016). It is hypothesized that genes with high in-degree centrality are critical in their function hence the redundancy in their regulation by upstream TFs. Indeed, in the Brady *et al.* (2011) performed a literature search and found that mutations in their genes with high in-degree confer an aberrant root phenotype. Hence capturing and visualizing regulatory relationships not only determines the local dynamics between two genes but also can build the network topology for hypothesis of a gene's importance in a particular function.

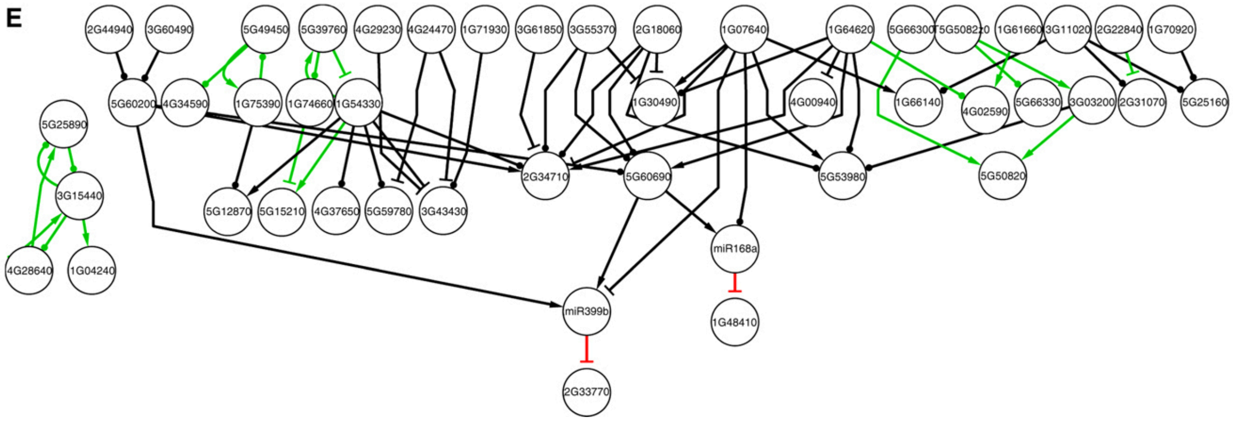


Figure 1.4. A stele-enriched GRN that was generated via Y1H and Y2H as described in (Brady et al., 2011). Black, green, and red edges represent protein-DNA-, protein-protein-, and miRNA-mRNA interactions respectively. Interactions were either repressive ("T"), activating (arrowhead), or exerting no known effort (circle).

Network topology and regulatory relationships can also be investigated in smaller, local multi-gene patterns. In particular, network motifs are small, recurring patterns of edges that are enriched in a network (Milo *et al.*, 2002). Milo *et al.* (2002) first explored network motifs in a yeast GRN to find that a 3-node patterns termed "feed-forward loops" (FFL) were highly enriched. Specifically, a FFL is where one TF (A) concomitantly regulates TF (B) and gene (C) while TF (B) also regulates target gene (C). They also found that FFLs were underrepresented in other networks such as ecological and web networks which suggests that FFL enrichment is unique to GRNs. Hence, motif enrichment can be used define and identify classes of networks. Later mathematical modeling by Alon (2007) suggests that FFLs exist to modify TF-promoter dynamics of the target gene by acting to pulse, delay, or persist its expression depending on the regulatory relationship of the TFs in the motif. The most common type of FFL is the coherent type 1 FFL (C1-FFL; see Figure 1.5, left) whereby all interactions TF (A) activates TF (B) which together activate target gene (C). If the target gene requires both TFs to activate its transcription, it is believed that C1-FFLs acts as persistence detector as TF B requires time to accumulate to sufficient levels to also upregulate the target gene (see Figure 1.5, right). Saddic *et al.* (2006) was one of the first to investigate a C1-FFL in Arabidopsis suggesting that the TFs, LMI1 and LFY (also activated by LMI1) both activate CAL that results in the transition to meristem identity. Their expression analysis of these genes fits initial suggestions that FFLs can indeed act as a persistence detector such that CAL is only expressed when both LMI1 and LFY are expressed to assist in the correct timing of flower formation. Indeed, a C1-FFL was identified in a GRN involved pancreatic specification in mice (Arda, Benitez and Kim, 2013) and modelling revealed a C1-FFL suspected to be involved in auxin-related lateral root formation in Arabidopsis (Q. Chen *et al.*, 2015). The Brady Lab was also able to find an enrichment of FFLs in another GRN created via Y1H. The FFL of VND6 and VND7 acts together to regulate many downstream genes in Arabidopsis secondary cell walls (Taylor-Teeples *et al.*, 2015). To display the enrichment of FFLs, they coloured all the edges involved in FFLs. This approach may display the relative enrichment of FFLs in an GRN but makes it challenging for users to locate individual FFLs as the same edge may be involved in several FFLs or overlap each other. I will later demonstrate in my GRN tool how I used a slider along with highlighting FFLs to remedy this visualization problem.

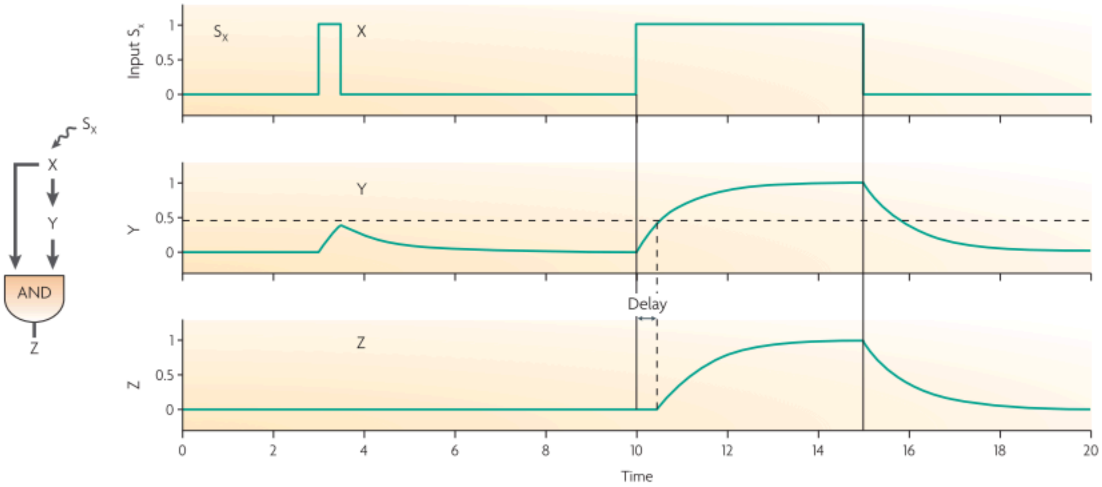


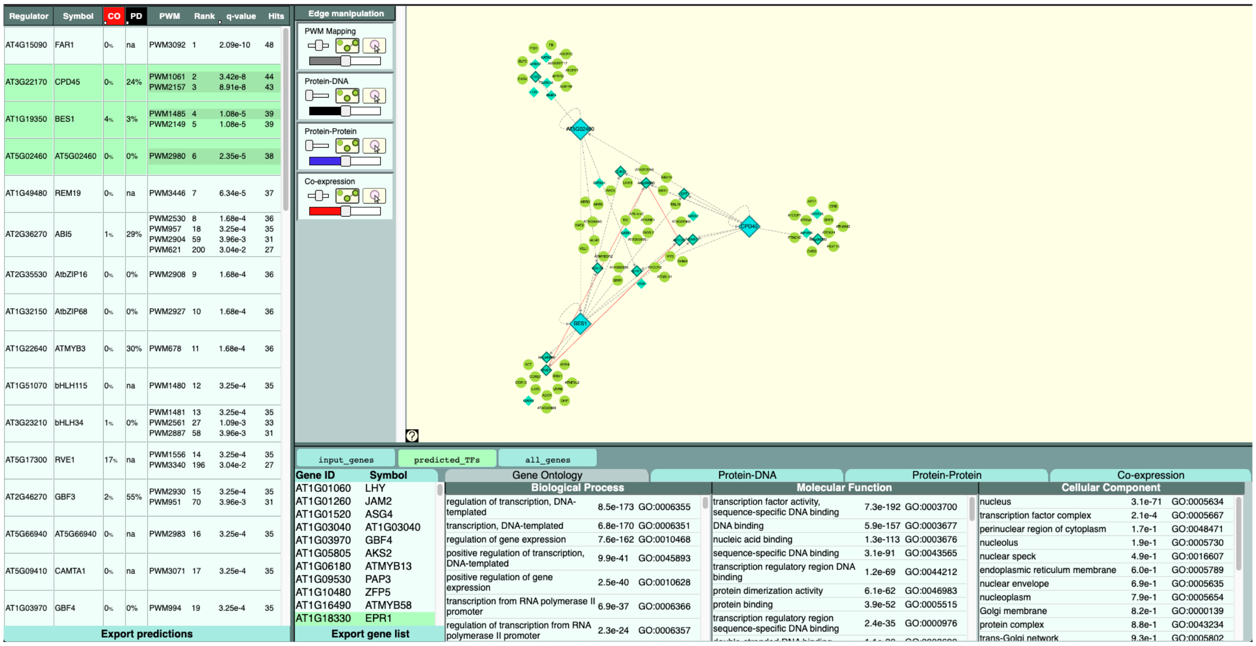
Figure 1.5. Left) Coherent type-1 feed-forward loop (C1-FFL). Right) Graphs of the C1-FFL members' protein levels whereby target gene Z is only activated after a persistence activation of TF X which upregulates TF Y to sufficient levels concomitantly results in a delay. Excerpt from (Alon, 2007)

Network topology may also be influenced by the experimental approach. For example, one may want to understand the downstream targets of a putative TF and perform ChIP-Seq on a select number of TFs. However, using this "TF-centric" method will bias the network topology to have many genes under regulation of the TFs under study (Gaudinier and Brady, 2016). Y1H is hence a "gene-centric" method in which focuses on protomers of interest and therefore will bias the network towards many different types of TFs upstream that can regulate the promoters of interest. Gaudinier and Brady (2016) therefore suggest opting for unbiased approaches such as DNase hypersensitivity assays which sequences open chromatin - such regions are typically TF-bound. However, these assays require databases of TFBSs to map reads to putative TFs, for which Arabidopsis are somewhat limited. Last, GRNs are thought to be highly tissue-specific with individual nodes not being expressed or modulated in certain tissues, which can change the network structure altogether (Emmert-Streib, Dehmer and Haibe-Kains, 2014). In summary, users who are viewing a GRN should be mindful of the techniques and experimental conditions used to map that particular network.

Lastly, GRNs can leverage external data integration from other sources, much like PPI networks can. For example, researchers can overlay expression data on a GRN to hypothesize how certain TFs or genes are perturbed under different conditions. For example, Taylor-Teeples *et al.* (2015) characterized how certain stresses (iron, salt, sulfur, pH) would alter their GRN by filtering the genes which were differently expressed and investigating highly connected TFs that regulated those genes. They hypothesized that REV was critical in depositing lignin in the secondary wall when the cell wall is under iron stress as REV had many interactions to lignin biosynthesis genes. Indeed, iron deprivation increased lignin staining and a loss-of-function REV mutant revealed altered expression in iron biosynthesis genes. Gaudinier and Brady (2016) and Liseron-Monfils and Ware (2015) also suggest integrating external PPI data in investigating additional types of regulation at the protein complex level. Last, GRNs also be integrated with functional data. Chen *et al.* (2019) constructed a yeast GRN to identify distinct subnetworks by their Gene Ontology (GO) categories to more easily reveal biological pathways implicated by the subnetwork's members.

1.2.4 Web-based Arabidopsis Protein-DNA Interaction Tools

There are a range of web-based tools for Arabidopsis researchers to query PDIs and investigate, or infer GRNs as summarized in Table 1.2. Comparing Table 1.1 to Table 1.2 reveals that there are more Arabidopsis-specific web tools for PDIs than PPIs and that these tools tend to be databases curated from high-throughput studies, particularly ChIP-Seq. Indeed, it is likely that the advent of ChIP-Seq (and now DAP-Seq) has allowed for PDI screening to overtake PPI screening. Importantly, unlike the PPI tools, there is a dedicated functional Arabidopsis web-based interactive viewer: TF2Network (Kulkarni *et al.*, 2017). Although this tool features some desired features such as external data integration for hypothesizing novel relationships (see Figure 1.6), TF2Network is limited in its functionality. Personal exploration of the tool has revealed that clicking on more than few TFs in the tool permanently deletes the generated network. Secondly, the tool requires a gene list which users may not always have. Last, the BAR can arguably offer more data integration as we host multiple expression data sets under multiple conditions/tissues along with regularly updated PDI/PPI data. Having multiple expression data sets will allow users to investigate different types of stresses in terms of they relate to their GRN, similar to like how Taylor-Teeples *et al.* (2015) investigated how their GRN was altered under several stresses with specific expression data. To conclude, TF2Network is optimal for those with a gene list to quickly create and visualize a GRN but may not be optimal for all users.



Regulator	Symbol	CO	PD	PWM	Rank	q-value	Hits
AT4G15090	FAR1	0%	na	PWM3092	1	2.09e-10	48
AT3G22170	CPD45	0%	24%	PWM1061 PWM2157	2 3	3.42e-8 8.91e-8	44 43
AT1G19350	BES1	4%	3%	PWM1485 PWM2149	4 5	1.08e-5 1.08e-5	39 39
AT5G02460	AT5G02460	0%	0%	PWM2980	6	2.35e-5	38
AT1G49480	REM19	0%	na	PWM3446	7	6.34e-5	37
AT2G36270	ABI5	1%	29%	PWM2530 PWM957 PWM2904 PWM621	8 18 59 200	1.68e-4 3.25e-4 3.96e-3 3.04e-2	36 35 31 27
AT2G35530	AtbZIP16	0%	0%	PWM2908	9	1.68e-4	36
AT1G32150	AtbZIP68	0%	0%	PWM2927	10	1.68e-4	36
AT1G22640	ATMYB3	0%	30%	PWM678	11	1.68e-4	36
AT1G51070	bHLH115	0%	na	PWM1480	12	3.25e-4	35
AT3G23210	bHLH34	1%	0%	PWM1481 PWM2561 PWM2887	13 27 58	3.25e-4 1.09e-3 3.96e-3	35 33 31
AT5G17300	RVE1	17%	na	PWM1556 PWM3340	14 196	3.25e-4 3.04e-2	35 27
AT2G46270	GBF3	2%	55%	PWM2930 PWM951	15 70	3.25e-4 3.96e-3	35 31
AT5G66940	AT5G66940	0%	na	PWM2983	16	3.25e-4	35
AT5G09410	CAMTA1	0%	na	PWM3071	17	3.25e-4	35
AT1G03970	GBF4	0%	0%	PWM994	19	3.25e-4	35

Gene ID	Symbol
AT1G01060	LHY
AT1G01260	JAM2
AT1G01520	ASG4
AT1G03040	AT1G03040
AT1G03970	GBF4
AT1G05805	AKS2
AT1G06180	ATMYB13
AT1G09530	PAP3
AT1G10480	ZFP5
AT1G16490	ATMYB58
AT1G18330	EPR1

Biological Process		
regulation of transcription, DNA-templated	8.5e-173	GO:0006355
transcription, DNA-templated	6.8e-170	GO:0006351
regulation of gene expression	7.6e-162	GO:0010468
positive regulation of transcription, DNA-templated	9.9e-41	GO:0045893
positive regulation of gene expression	2.5e-40	GO:0010628
transcription from RNA polymerase II promoter	6.9e-37	GO:0006366
regulation of transcription from RNA polymerase II promoter	2.3e-24	GO:0006357

Molecular Function		
transcription factor activity, sequence-specific DNA binding	7.3e-192	GO:0003700
DNA binding	5.9e-157	GO:0003677
nucleic acid binding	1.3e-113	GO:0003676
sequence-specific DNA binding	3.1e-91	GO:0043565
transcription regulatory region DNA binding	1.2e-69	GO:0044212
protein dimerization activity	6.1e-62	GO:0046983
protein binding	3.9e-52	GO:0005515
transcription regulatory region sequence-specific DNA binding	2.4e-35	GO:0000976

Cellular Component		
nucleus	3.1e-71	GO:0005634
transcription factor complex	2.1e-4	GO:0005667
perinuclear region of cytoplasm	1.7e-1	GO:0048471
nucleolus	1.9e-1	GO:0005730
nuclear speck	4.9e-1	GO:0016607
endoplasmic reticulum membrane	6.0e-1	GO:0005789
nuclear envelope	6.9e-1	GO:0005635
nucleoplasm	7.9e-1	GO:0005654
Golgi membrane	8.2e-1	GO:0000139
protein complex	8.8e-1	GO:0043234
trans-Golgi network	9.3e-1	GO:0005802

Figure 1.6. TF2Network (Kulkarni et al., 2017) output when the demo data set is used. The predicted TFs, BES1, CPD45, and AT5G02460 were selected to create the gene regulatory network as shown in the top right panel.

Currently, ePlant (Waese *et al.*, 2017) can only visualize PDIs for a single TF but not a GRN as the tool was specifically made to show multiple levels of biological information for a single gene (see Figure 1.3). However, as GRNs contains multiple genes we can create an additional plug-in for ePlant such to show all GRNs a particular a gene is involved in. This tool could help understand the regulatory relationships of a gene of interest and in which particular conditions interactors play a role, much like how Taylor-Teeples *et al.* (2015) explored the role of REV in iron stress. As ePlant is our most popular tool (~125,000 views to date in 2019), it would also introduce the relatively new concept of GRNs to the plant community. Hence my second objective is to create a GRN viewer for users to generate hypotheses regarding the regulatory nature of their gene of interest.

1.3 Applying Data Visualization to Biological Networks

In the seminal data visualization book, The Visual Display of Quantitative Information (Tufte, 2001), Tufte discussed the utility of visualizing the data explored versus conventional summarization statistics. As an example, he used the Anscombe's quartet (see Figure 1.7), where Anscombe (1973) visualized datasets that were vastly different but had the same correlation coefficient. Visualizing the dataset allows one to quickly see if there are any outliers to their dataset, which can be seen in the most rightmost panel in Figure 1.7. At a recent network visualization symposium, Chen et al. (2018) also made a similar analogy to networks in that statistics such as average path length can summarize widely different graphs. Indeed, as seen in Figure 1.8, some graphs can be vastly different topologically even when they share the exact same graphical statistics, such as number of edges/nodes, and clustering coefficient. Therefore, network visualization can be a useful complement to traditional statistical approaches in examining key genes and regulatory relationships.

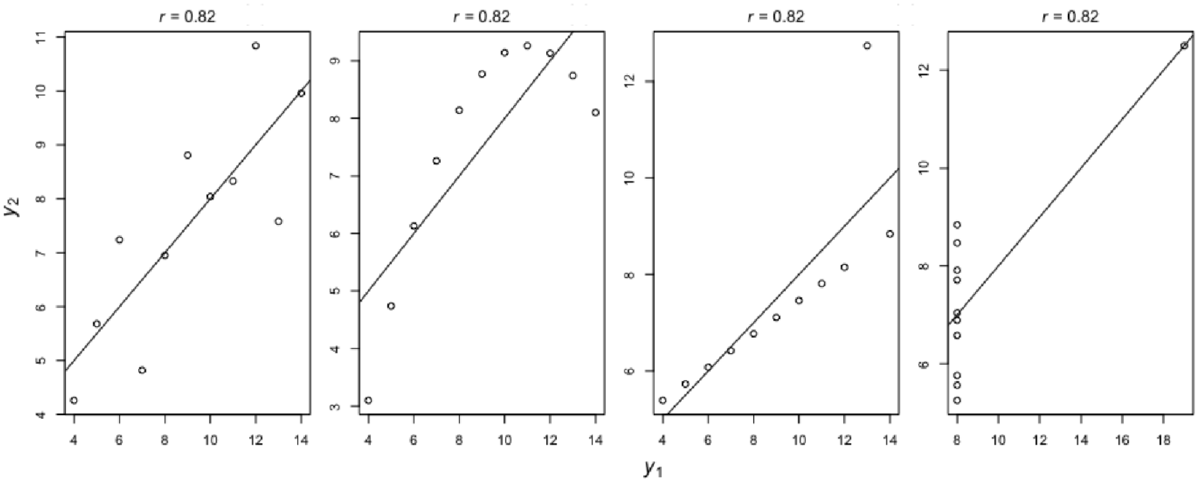


Figure 1.7. Anscombe's quartet where vastly different datasets have the same descriptive statistic (correlation coefficient). Graph was created using R with the Anscombe dataset via the plot function.

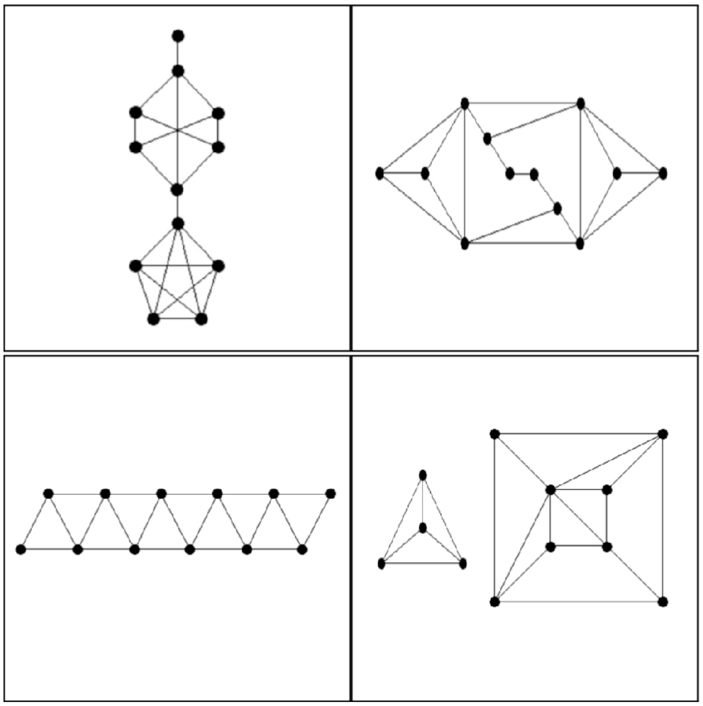


Figure 1.8. These graphs share the same properties: number of edges, nodes, number of triangles, girth and global clustering coefficient. However, they are structurally very different. Excerpt from Chen et al. (2018)

Nesbitt and Friedrich (2002) made the analogy of applying Gestalt psychology’s “the whole is more than the sum of the parts” to graphs by stating “…[by] looking at individual nodes we don’t necessarily learn much about the overall structure of the graph”. The authors describe the need to apply Gestalt’s principles of organization for users to contain a consistent mental map of the network topology. For example, they discuss applying the law of familiarity (where things form groups if the groups appear meaningful) by shading and highlighting groupings to identify key subnetworks. Indeed, although visualizing all the edges without any shading follows Tufte's (2001) rule of maximizing data-ink (where data-ink is defined as “the non-erasable core of a graphic”), laying out all the edges of a large network remains a challenge (Albrecht *et al.*, 2009; Dogrusoz *et al.*, 2018). Specifically, default layouts often result in hairball-like structures which are difficult to navigate. Moreover, these often force-directed graphs have little semblance to biological organization, as researchers most likely choose an efficient, default layout for graphs. Another approach is to summarize the data instead of developing an optimal layout.

Here, Ahnert (2013) developed a compression algorithm to collapse networks based on common edges between nodes to create power nodes. He then summarized a large *E. coli* GRN (889 nodes, 1465 edges) into a graph of 124 power nodes (see Figure 1.9) and showed that the most compressible components exhibited significant enrichment for certain GO terms, which suggests this method retains biological relevance. Moreover, many of the power nodes correspond to the genes that regulate the same particular operon. Unfortunately, this technique has not been widely adopted such that the challenge of visualizing large networks remains. In AIV2, I will demonstrate an alternative layout which uses subcellular localization to organize a graph efficiently and applies Gestalt principles. I will also demonstrate how I summarize PDIs in a fashion akin to Ahnert's (2013) power nodes.

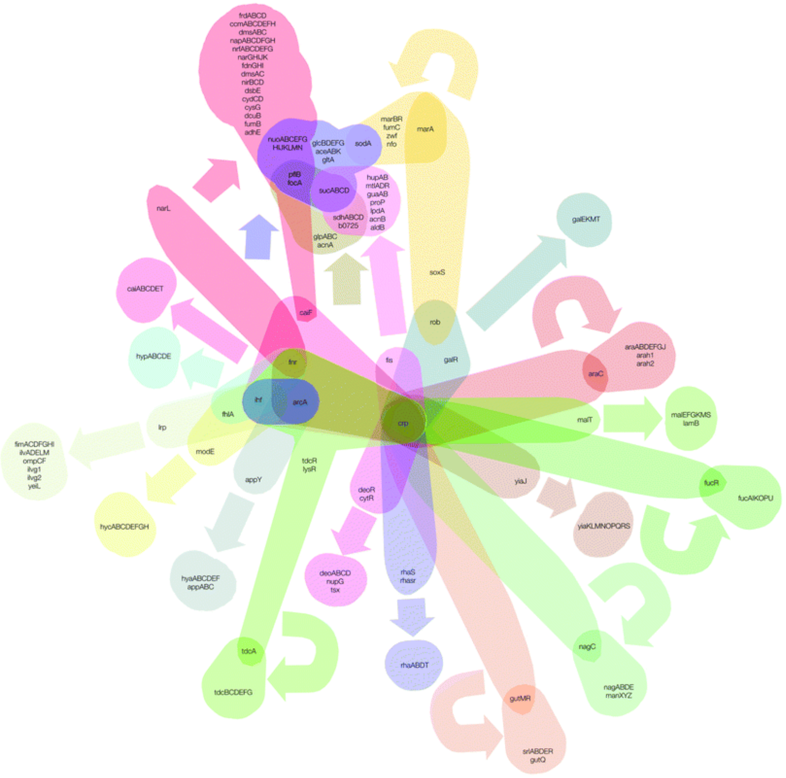


Figure 1.9. Ahnert (2013) developed an algorithm to compress large GRNs. Here are the most compressible 20 power nodes of the compressed E. coli GRN that originally composed of 889 nodes and 1465 edges.

Although users can identify key regulatory relationships in a network via its topology, users can also focus on (a set of) individual nodes/edges. I discussed earlier how colour was used in PPI networks to either display subcellular localization and expression data. However, users may not always be interested in such information. For example, a user may be only interested in identifying the targets of a particular TF. In this instance, coloured localization data could be considered what Tufte (2001) termed “chartjunk” – unnecessary data-ink that does not assist in conveying a narrative. Currently, in AIV, there are no options to hide such colouring. In the following chapter, I show AIV2’s host of features which include filters to hide task-specific colouring. I will also discuss applying data visualization techniques such as Gestalt’s principles, rapid serial visual presentation (Spence, 2002) and Shneiderman's (1996) mantra of “overview-first, zoom-and-filter, then details-on-demand”.

1.4 Research Goals in Summary

I aim to create two modern, interactive web applications to visualize PPI networks and GRNs respectively. The PPI viewer, Arabidopsis Interactions Viewer 2 (AIV2) will host all the prior features of AIV1 along with newer data sets including our predicted structure-based interactome. Additional features will be included such as dynamic filtering of nodes, and external linkouts to other bioinformatic tools for further validation. My GRN viewer, Arabidopsis GEne Network Tool (AGENT) will database and visualize curated Arabidopsis GRNs to assist users in generating hypotheses. The features in AGENT will integrate our expression and interaction data at the BAR along with having a simple user interface to allow it to easily plug-in to ePlant. Together these two tools will serve as two of the few fully-featured web-based network viewers in Arabidopsis which respect modern principles of web design, software engineering, and data visualization.

1.5 Tables

Table 1.1. Summary of web-based tools for querying and/or visualizing Arabidopsis PPIs. PPI numbers were taken from the resource's statistics directly in Dec 2019 or the paper if the former was not available.

Tool	Description
AIV (Geisler-Lee *et al.*, 2007)	Arabidopsis-centric Flash-based interactive viewer for 70,000 predicted interolog PPIs and 36,000 literature-based PPIs. Integrates expression, localization, and functional (MapMan) information. See Figure 1.1.
ATPID (Lv *et al.*, 2017)	Database curated 28,062 Arabidopsis PPIs from major databases and text mining along with *in silico* predictions (interolog, enriched domains/GO-terms). **Defunct as of December 2019**
ATPIN (Brandão, Dantas and Silva-Filho, 2009)	Database curated 96,276 PPIs from other major databases and Geisler-Lee *et al.'s* (2007) interactome. Calculated *ad hoc* values for co-localization of protomers. Authors suggest exporting an XGMML file to visualize PPIs.
BioGRID (Oughtred *et al.*, 2019)	Large, species-agnostic curator-based database that has 48,981 Arabidopsis PPIs. No visualization provided, but users are able to export large datasets. Now collaborates with BAR in reciprocally sharing PPIs.
ePlant (Waese *et al.*, 2017)	Before BAR and AIV2 update (discussed later), displayed same set of AIV PPIs for a given gene without Flash technology. Now has 140,353 PPIs. Simpler and more modern interactive viewer to AIV. See Figure 1.3.
INTACT (Kerrien *et al.*, 2012)	Very large, species-agnostic curator-based database that has 50,415 Arabidopsis interactions (PPIs inclusive). Each interaction requires a molecular interaction (MI) term which represents the method used to uncover the interaction. Webserver hosts a basic Flash-based viewer without any additional decorations.
PAIR (Lin, Shen and Chen, 2011)	Database of 145,949 predicted PPIs and 5,990 validated PPIs from other databases. Used a support vector machine (SVM) model that used expression, localization, interolog, and functional information to predict PPIs. **Defunct as of December 2019**
MIND (Jones *et al.*, 2014)	12,102 high-confidence Arabidopsis PPIs that were screened using the split-ubiquitin method (proteins of interest are fused to a split ubiquitin which reconstitute when in close contact to activate a transcription factor). Webserver of PPIs contain confidence scores for each PPI (authors re-screened candidates). Present in aggregating databases such as BioGRID.
STRING (Szklarczyk *et al.*, 2019)	Large, species agnostic multi-source (text mining, other databases, interologs, reprocessing large-scale experiments) database that hosts PPI networks for 22463 Arabidopsis proteins. Contains a simple network viewer for each protein

	of interest with edges coloured by evidence type but no Arabidopsis-specific features.
TAIR (Lamesch *et al.*, 2012)	TAIR hosts the Geisler-Lee *et al.* (2007) predicted interactome as well as their own tab-delimited file of 2,655 PPIs that were curated from the literature and BioGRID.

Table 1.2. Summary of web-based tools for querying and/or visualizing Arabidopsis PDIs and/or GRNs. Statistics were taken from the resource's statistics directly in Dec 2019 or the paper if the former was not available.

Tool	**Description**
AIV (Geisler-Lee *et al.*, 2007)	Arabidopsis-centric Flash-based viewer for 1,784 PDIs.
AthaMap (Bülow, Brill and Hehl, 2010)	Generates a list of potential target genes for 211 TFs by looking for enriched patterns and PWMs upstream of TSSs.
AtRegNet/AGRIS (Yilmaz *et al.*, 2011)	Curated database of 1,770 TFs with detailed summaries. AtRegNet is an interactive GRN viewer based on 1,638,778 PDIs curated from published high-throughput studies (mostly ChIP-Seq). **AtRegNet is defunct as of Dec 2019**
CressInt (X. Chen *et al.*, 2015)	Web form that allowed users to search through 575 TFs that bind to a given promoter. Curated multiple types of high-throughput data to build database of PDIs. **Defunct as of Dec 2019**
ePlant (Waese *et al.*, 2017)	Modern gene-centred viewer that databases ~3 million PDIs. Visualizes multiple gene targets for a given gene as a single "chromosomal" nodes.
Expresso (Aghamirzaie *et al.*, 2017)	Web form that displays potential downstream targets for 20 TFs that were used in publicly available ChIP-Seq experiments. Used MEME suite to identify potential targets of TFs.
PlantTFDB (Jin *et al.*, 2017)	Webserver for curated database of 2296 Arabidopsis TFs with detailed summaries of annotations, domains, expression, and publications for each TF. Can retrieve targets for TF via motif searching but network viewer is non-functional.
PlnTFDB (Pérez-Rodríguez *et al.*, 2009)	Webserver for 2,451 predicted TFs. TFs were predicted from Arabidopsis genome based on known DBD data. No predicted targets provided.

TF2Network (Kulkarni *et al.*, 2017)	Interactive GRN viewer that builds a GRN based on user-submitted gene list. Integrates external PPI, PDI, and GO data.

Chapter 2
Arabidopsis Interactions Viewer 2 (AIV2)

2 Arabidopsis Interactions Viewer 2, published as part of Dong *et al.* (2019)

2.1 Materials and Methods

2.1.1 Data Sources and Libraries

The AIV database is powered by MySQL 8 which is hosted on the Bio-Analytic Resource (BAR) Linux server. It has been curated and updated to include a total of 140,353 PPIs and 3 million PDIs including the predicted interactomes inferred by Dong *et al.* (2019) and Geisler-Lee *et al.* (2007). Of the 3 million PDIs, 2.7 million PDIs were curated from O'Malley *et al.*'s (2016) DAP-Seq pipeline and 2,967 from Y1H experiments (Brady *et al.*, 2011; Gaudinier *et al.*, 2011; Li *et al.*, 2014; Taylor-Teeples *et al.*, 2015; De Lucas *et al.*, 2016; Murphy *et al.*, 2016; Porco *et al.*, 2016; Sparks *et al.*, 2016). Approximately, 40,000 Predicted PDIs have also been recently added from the Yu, Lin and Li (2016) pipeline where they predicted TFBSs for 400 TFs using the FIMO (MEME suite) tool. The BAR also has an active collaboration with BioGRID to mutually share interactions regularly. We have imported 42,605 BioGRID Arabidopsis PPIs (Oughtred *et al.*, 2019) to date. Due to the sheer volume of PPI and PDI references in our database, a list is not given. However, we can send an up-to-date MySQL database dump of publications to those who are interested.

The database schema also contains columns for the Pearson correlation coefficient (PCC) between PPIs as calculated in Geisler-Lee *et al.* (2007), Molecular Interaction (MI) ontology terms, interolog hits for several species (yeast, worm, fly, human, mouse, *E. coli*) and any values for predicted interactions if applicable. A web service has been written by the BAR's bioinformatic technician Asher Pasha to retrieve all the interactions and the above information which will be returned in JSON (JavaScript Object Notation) format when queried with an AGI ID(s). The application programming interface (API) can be accessed via https://bar.utoronto.ca/interactions2/cgi-bin/get_interactions_dapseq.php. For an example of the JSON structure returned, see Appendix 1. I also use another API that Mr. Pasha wrote that connects to our expression database and powers many applications including ePlant (Waese *et al.*, 2017). This API (https://bar.utoronto.ca/interactions2/cgi-bin/getSample.php) returns expression information when queried with a list of AGI IDs and a tissue/condition, as seen in Appendix 2. Last, Mr. Pasha has also written an API (https://bar.utoronto.ca/interactions2/cgi-bin/gene_summaries_POST.php) that provides gene annotations and names for list of AGI IDs.

Additional APIs are also used to retrieve interactions, integrate additional biological information, and create external links. The BAR, INTACT, and BioGRID subscribe to PSICQUIC (Proteomics Standard Initiative Common QUery InterfaCe) which is a protocol for PPI databases to draft their web services. This protocol allows for a standardized format to import additional Arabidopsis PPIs. To enable Cross-Origin Resource Sharing (CORS) support on my web application, I created two proxies in PHP to their web services: https://bar.utoronto.ca/interactions2/cgi-bin/psicquic_intact_proxy.php?request=AGI_ID and https://bar.utoronto.ca/interactions2/cgi-bin/psicquic_biogrid_proxy.php?request=AGI_ID where AGI_ID is an AGI ID (hereafter used as a placeholder). These proxies simply call the PSICQUIC APIs through our server and return a PSICQUIC compliant tab-delimited text file of interactions for an AGI ID. Another proxy was made for the MapMan API (https://mapman.gabipd.org; Usadel *et al.* (2009)), which returns MapMan codes, which are functional categories along with gene descriptions for a given AGI ID(s). The API is accessible at https://bar.utoronto.ca/interactions2/cgi-bin/bar_mapman.php?request=["AGI_ID"], see Appendix 3 for an example JSON structure.

As the SUBcellular localization database for Arabidopsis proteins (SUBA4; Hooper *et al.* (2017)) does not host a publicly available API, they have kindly dumped their MySQL database for us to host a separate copy. This database curates evidence for subcellular localization from the literature and databases for an AGI ID and also notes if the evidence is predicted or validated. Our database copy is also hosted via MySQL 8.0. I used PHP to create an API that queries this database and returns *ad hoc* localization scores for each subcellular compartment for a gene. This score is dependent on how much predicted and experimental evidence for a compartment exists (experimental evidence is scored five times as much) and is similar in logic to the ePlant SUBA3 API (Waese *et al.*, 2017). The code for this API is accessible from https://github.com/VinLau/BAR-SUBA4-Webservice and an example JSON structure is given in Appendix 4.

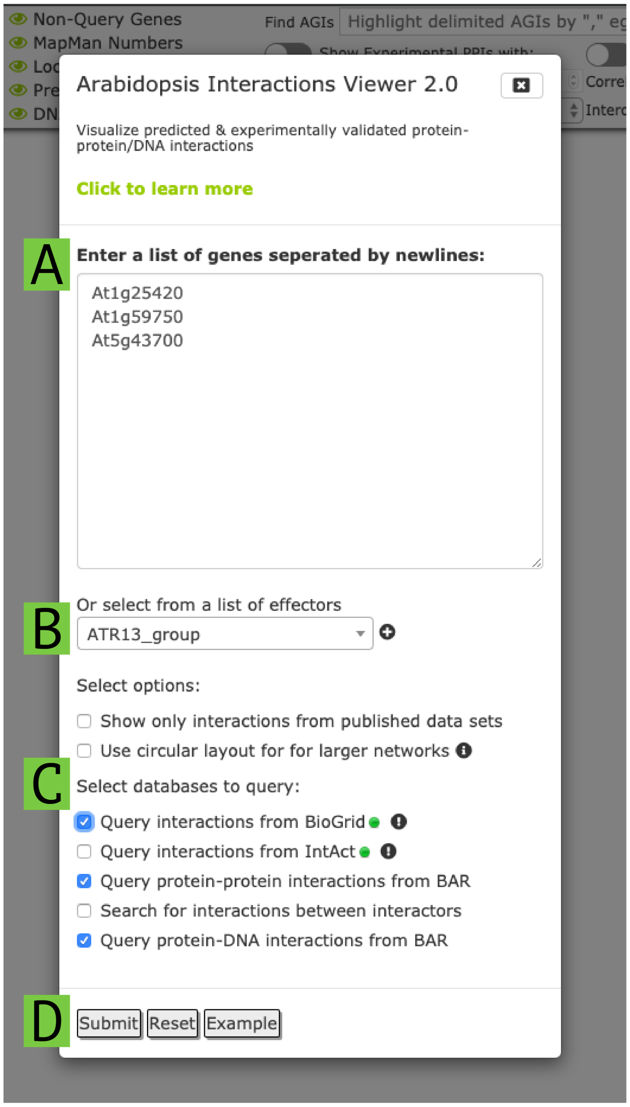
External linkouts to other tools and references are also utilized. To reference O'Malley *et al.*'s (2016) work I used the JavaScript webserver library, Express.js (https://expressjs.com) to create an API that dynamically links out to their genome viewer for visualizing DAP-Seq peaks (http://neomorph.salk.edu/aj2/pages/hchen/dap_ath_pub.php). This small app takes in two genes as an input, one TF and one target gene, to return a dynamic URL for the DAP-Seq tool. To do this, I created a dictionary of map TFs to the correct unique DAP-Seq identifiers and used an internal BAR API to locate the TSS of the target genes to create the Uniform Resource Locator (URL). For example, to find if TF At1g44830 (ERF14) has a DAP-Seq peak near the TSS of the target gene At2g44160 (MTHFR2), one would use the following link http://bar.utoronto.ca/DAP-Seq-API?target=At2g44160&tf=At1g44830. ERF14 maps to "3_11" and the TSS of MTHFR2 is located at position 18262210 on the 2nd chromosome. The API will then return the following URL: http://neomorph.salk.edu/aj2/pages/hchen/dap_ath_pub.php?active=DAP data&location=**2**:**18262210**:600:20&hide=["1_2","**3_11**"]&config=[{id:"1_2",height:350,scale:1.25},{id:"**3_11**",height:250,scale:1}]&settings={yaxis:250,accordion:"collapsed"}. Clicking on this link returns the DAP-Seq genome browser which has a moderate DAP-Seq peak in a region ~700bp upstream of the TSS of MTHFR2 for ERF14 (see Appendix 6). The DAP-Seq browser options (scale, height) were configured to optimize space on a typical modern screen resolution (1080p) and to distinguish large peaks. For our 9,065 top ranking predicted structural-PPIs (S-PPIs), Richard Song, an intern of the Provart Lab has created a web tool which visualizes the docking frequency for 500 runs. The tool URL (http://bar.utoronto.ca/protein_docker/?id1=AT5G27670&id2=AT3G53650) will load the two 3D structures with a heat maps denoting areas of contact frequency (see Appendix 5). Last, I created dynamic URLs to link to PubMed, digital object identifers (DOIs), MIND, the landing page of the Arabidopsis Interactome Mapping Consortium (2011), BIND, and BioGRID.

The AIV2 User Interface (UI) was built using Hypertext Markup Language (HTML) elements and jQuery (https://jquery.com). The tablefilter.js (https://www.tablefilter.com/) library was used to filter, tabulate, and export a table summarizing the network. Version 3 of Cytoscape.js library (Franz *et al.*, 2016) was used to create and manipulate networks. Additional plug-ins for Cytoscape.js were used such as Cytoscape-qTip.js which creates tooltips for nodes when clicked, Cytoscape-context-menus.js to allow right-clicking on nodes, and Cytoscape-canvas.js which allows one to draw an image behind the network interface. Last, the I used the Cytoscape.js layout libraries, CerebralWeb (Frias *et al.*, 2015), which I heavily modified, and CoSE-Bilkent for additional layout options.

The source code for AIV2 can be accessed via https://github.com/VinLau/AIV-v2-cytoscapeJS. The two major scripts are aiv.js and aiv_ui.js. While aiv_ui.js adds functionality to HTML elements, the asynchronous JavaScript and XML (AJAX) requests to the aforementioned APIs and application state/logic is mostly contained in the aiv.js script. The logic for building the dynamic URLs to linkout to the external resources (PubMed, MIND, etc.) is found in the aiv.js file. The application is hosted via an Apache HTTP server (BAR's server) and is publicly accessible via https://bar.utoronto.ca/interactions2.

2.1.2 Application Logic

Upon opening the app, the user is greeted with a modal that prompts the user to complete the form which will query AIV2 (see Figure 2.1). Options include the databases to query (BAR, BioGRID, INTACT), whether to query PPIs and/or PDIs, and whether to only show experimentally-validated interactions. Behind-the-scenes checks such as live verification of AGI IDs with regular expressions allow a thorough user-experience to complete the form. The form also follows modern user experience (UX) principles such as a single column, chunked structure that allows offers clear delineation between form categories as eye-tracking experiments shown this format is most effective for an optimal UX (Baginski, 2019). For example, the optional list of Effectors (bacterial proteins that invade Arabidopsis cells) has its own distinct 'chunk' or section which allows for users to easily distinguish this option and whether to ignore or select it (see Figure 2.1, Item B). As AIV2 will be superseding AIV which is one of our most popular tools, I also chose to include similar design patterns from AIV to ease transition for legacy users. I chose a moderate redesign as UX research performed on focus groups show that incremental design changes are usually preferred over major overhauls unless major architectural changes are required (Hoa, 2015). An example of maintaining legacy design is keeping the ordering of database checkboxes the same, likewise for the ordering of the buttons (see Figure 2.1, Item C and Appendix 7). Once the user selects the appropriate options and has entered in a list of AGI IDs delimited by newlines the user will press the submit button (see Figure 2.1, Item A, D).


Non-Query Genes
MapMan Numbers
Loc
Pre
DN
Find AGIs
Highlight delimited AGIs by "," eg
Show Experimental PPIs with:
Corre
Interc
A
B
C
D

Arabidopsis Interactions Viewer 2.0 ☒

Visualize predicted & experimentally validated protein-protein/DNA interactions

Click to learn more

Enter a list of genes seperated by newlines:

At1g25420
At1g59750
At5g43700

Or select from a list of effectors

ATR13_group

Select options:

- ☐ Show only interactions from published data sets
- ☐ Use circular layout for for larger networks

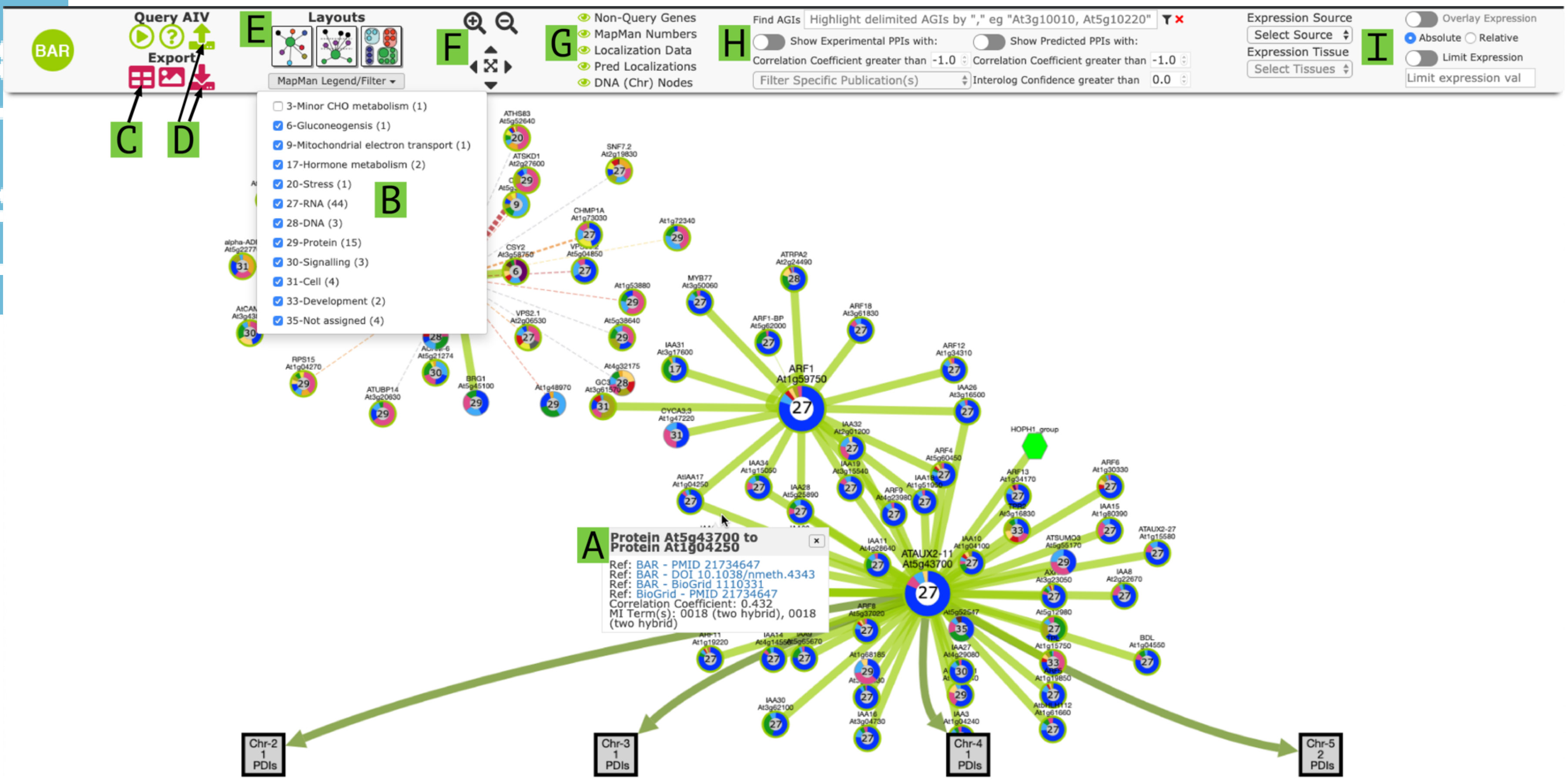
Select databases to query:

- ☑ Query interactions from BioGrid
- ☐ Query interactions from IntAct
- ☑ Query protein-protein interactions from BAR
- ☐ Search for interactions between interactors
- ☑ Query protein-DNA interactions from BAR

Submit Reset Example

Figure 2.1. Splash page when user launches AIV2. A) Users enter query genes by their AGI IDs delimited by newlines B) Clear demarcations between sections allows easier 'chunking' of information and form flow C) Database selection shows similar design and choice structure to AIV1 for legacy users D) Submit button to initiate the application.

When the submit button is clicked, the form is validated again and an AJAX request is made to the AIV2 interactions database to fetch the PPIs and PDIs for the user's given form. Once the AJAX request is received, it is parsed manually by a JavaScript function. This function loops through each interaction to build a list of PPIs and PDIs. A unique list of genes for PDIs and PPIs is then stored in memory. From this list, Cytoscape.js then adds protein nodes and DNA nodes for each chromosome. The library then lays out the protein nodes with a “force-directed” algorithm in which the layout simulates a physical system where nodes repel each other like electrons until the whole energy of the layout is minimized (Dogrusoz *et al.*, 2018). In most cases, this produces a bicycle-spoke like layout for the protein nodes. See for the output of AIV2 in Figure 2.2 with given the form input in Figure 2.1. To condense the sheer number of PDIs, we summarized the PDIs in distinct, square ‘chromosomal’ DNA nodes where a table of PDIs appears when a user hovers over the node. The user can linkout to the reference or O’Malley *et al.*'s (2016) DAP-Seq browser that support each PDI by clicking on the icons (see Figure 2.3, Item A). This multi-layer design follows Shneiderman's (1996) mantra of “[o]verview first, zoom and filter, then details-on-demand”. The chromosomal nodes also take inspiration from Ahnert's (2013) compression nodes. The script then styles the edges according to the PCC (redder edges have higher PCCs) and experimental support (lime green lines denote validated PPIs). For our predicted S-PPIs, the top-ranking S-PPIs are relatively thicker than lower-ranking edges.



Query AIV

Export

Layouts

MapMan Legend/Filter ▾

☐ 3-Minor CHO metabolism (1)
☑ 6-Gluconeogensis (1)
☑ 9-Mitochondrial electron transport (1)
☑ 17-Hormone metabolism (2)
☑ 20-Stress (1)
☑ 27-RNA (44)
☑ 28-DNA (3)
☑ 29-Protein (15)
☑ 30-Signalling (3)
☑ 31-Cell (4)
☑ 33-Development (2)
☑ 35-Not assigned (4)

Non-Query Genes
MapMan Numbers
Localization Data
Pred Localizations
DNA (Chr) Nodes

Find AGIs: Highlight delimited AGIs by "," eg "At3g10010, At5g10220"

Show Experimental PPIs with:
Correlation Coefficient greater than: -1.0
Filter Specific Publication(s)

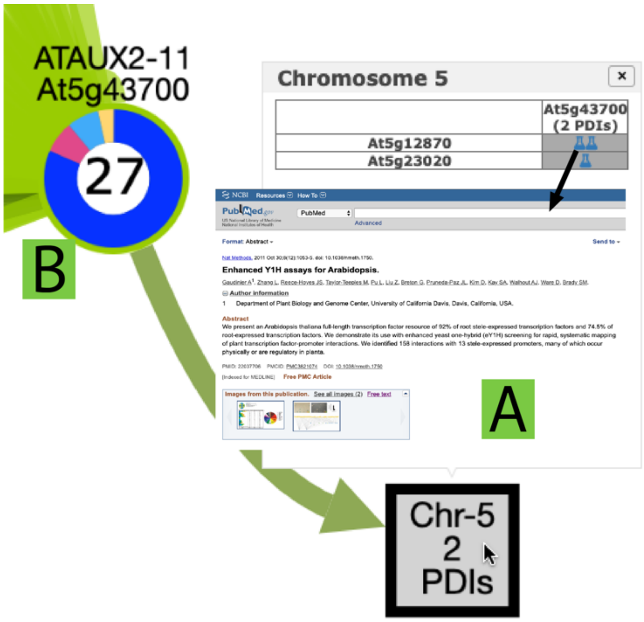
Show Predicted PPIs with:
Correlation Coefficient greater than: -1.0
Interolog Confidence greater than: 0.0

Expression Source: Select Source
Expression Tissue: Select Tissues

Overlay Expression
● Absolute ○ Relative
Limit Expression
Limit expression val

Protein At5g43700 to Protein At1g04250
Ref: BAR - PMID 21734647
Ref: BAR - DOI 10.1038/nmeth.4343
Ref: BAR - BioGrid 1110331
Ref: BioGrid - PMID 21734647
Correlation Coefficient: 0.432
MI Term(s): 0018 (two hybrid), 0018 (two hybrid)

Figure 2.2. Output and overview of AIV2 when At1g25420, At1g59750, and At5g43700 are queried with the form options listed in Figure 2.1. A) Detailed information (references, PCC, MI terms) regarding each interaction are shown when hovered. B) MapMan legend which also can act as a filter C) Table export button which opens table summarizing network (see) D) Import and export options for network E) Alternative layout buttons for visualizing the network F) Zoom and pan options for navigating the network G) Filter for specific nodes H) Filters for edges in the network based on certain interaction criteria I) Expression overlay UI



	At5g43700 (2 PDIs)
At5g12870	⚗⚗
At5g23020	⚗

Nat Methods. 2011 Oct 30;8(12):1053-5. doi: 10.1038/nmeth.1750.

Enhanced Y1H assays for Arabidopsis.

Gaudinier A[1], Zhang L, Reece-Hoyes JS, Taylor-Teeples M, Pu L, Liu Z, Breton G, Pruneda-Paz JL, Kim D, Kay SA, Walhout AJ, Ware D, Brady SM.

Author information

1 Department of Plant Biology and Genome Center, University of California Davis, Davis, California, USA.

Abstract

We present an Arabidopsis thaliana full-length transcription factor resource of 92% of root stele-expressed transcription factors and 74.5% of root-expressed transcription factors. We demonstrate its use with enhanced yeast one-hybrid (eY1H) screening for rapid, systematic mapping of plant transcription factor-promoter interactions. We identified 158 interactions with 13 stele-expressed promoters, many of which occur physically or are regulatory in planta.

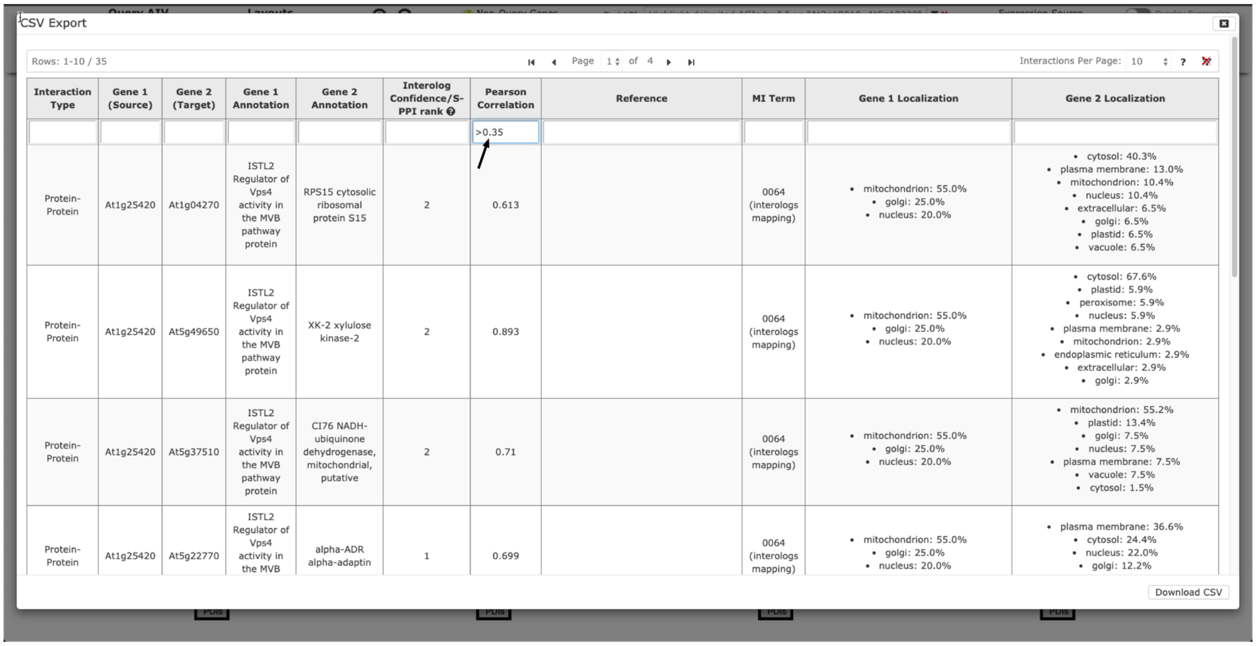
PMID: 22037706 PMCID: PMC3821074 DOI: 10.1038/nmeth.1750

[Indexed for MEDLINE] Free PMC Article

Figure 2.3. Zoomed in image of chromomsal DNA node. Hovering over the node creates a table that summarizes all the PDIs for that particular chromosome. A) Each protein node in AIV2 is decorated with a 'donut' pie-chart that summarizes the relative localization values derived from SUBA4. The MapMan code fills the centre of the donut. B) Clicking on an icon within the PDI table linkouts to the resource that validated the interaction.

As genes in the PPI network have now been retrieved, the aiv.js script then sends a gene list to the BAR server via AJAX queries to retrieve the gene annotations, SUBA4 localization data, and MapMan codes for each gene. The script then loops through every protein node and stores the above data in memory (or 'data property' in Cytoscape.js) for each particular node. Once complete, the script re-styles the protein nodes such that the gene names are appended to the AGI IDs, and 'donut' pie-charts which denote the MapMan code and localization data are overlaid on the node body (see Figure 2.3, Item B). Users may now hover over the nodes and edges to reveal additional information such as the gene description, PCC of the gene expression pattern of the partners or references that support the PPI (see Figure 2.2, Item A).

After the above AJAX requests are complete, AIV2 is fully loaded and the user may now select additional options to modify the network, search for genes, or integrate additional biological data. These functionalities are hosted in the upper task bar which follows responsive web design principles and was optimized for modern screen resolutions (pixel width above 1600). All of these UI elements are HTML elements with event listeners that run custom functions when the appropriate event is executed. For example, the MapMan legend maps each present MapMan number to the functional category (see Figure 2.2, Item B). However, the legend also has a bound function which executes when a category is by clicked. This function analyzes the unchecked categories and then hides the nodes in the network as each node hosts their MapMan code in their ‘data property’. Quickly filtering the network via MapMan categories allows for user-enabled rapid serial visual presentation (RSVP; Spence (2002)) of the magnitude of that functional category and whether that functionality is localized in a particular subnetwork. Another UI option is to tabulate all the interactions in a dynamic table which can be exported (see Figure 2.4). This table allows for focused filtering of the interactions in cases where a visual approach would be inadequate. For example, one can filter the table to create a subset of edges that have a PCC score above 0.35. Last, a user can hide task-irrelevant elements in the network. For instance, users only interested in PPIs can hide the chromosomal nodes to eliminate “chartjunk” (Tufte, 2001). A summary of the UI elements and their functionality is summarized in Figure 2.2.



CSV Export

Rows: 1-10 / 35 Page 1 of 4 Interactions Per Page: 10

Interaction Type	Gene 1 (Source)	Gene 2 (Target)	Gene 1 Annotation	Gene 2 Annotation	Interolog Confidence/S-PPI rank	Pearson Correlation	Reference	MI Term	Gene 1 Localization	Gene 2 Localization
						>0.35				
Protein-Protein	At1g25420	At1g04270	ISTL2 Regulator of Vps4 activity in the MVB pathway protein	RPS15 cytosolic ribosomal protein S15	2	0.613		0064 (interologs mapping)	• mitochondrion: 55.0% • golgi: 25.0% • nucleus: 20.0%	• cytosol: 40.3% • plasma membrane: 13.0% • mitochondrion: 10.4% • nucleus: 10.4% • extracellular: 6.5% • golgi: 6.5% • plastid: 6.5% • vacuole: 6.5%
Protein-Protein	At1g25420	At5g49650	ISTL2 Regulator of Vps4 activity in the MVB pathway protein	XK-2 xylulose kinase-2	2	0.893		0064 (interologs mapping)	• mitochondrion: 55.0% • golgi: 25.0% • nucleus: 20.0%	• cytosol: 67.6% • plastid: 5.9% • peroxisome: 5.9% • nucleus: 5.9% • plasma membrane: 2.9% • mitochondrion: 2.9% • endoplasmic reticulum: 2.9% • extracellular: 2.9% • golgi: 2.9%
Protein-Protein	At1g25420	At5g37510	ISTL2 Regulator of Vps4 activity in the MVB pathway protein	CI76 NADH-ubiquinone dehydrogenase, mitochondrial, putative	2	0.71		0064 (interologs mapping)	• mitochondrion: 55.0% • golgi: 25.0% • nucleus: 20.0%	• mitochondrion: 55.2% • plastid: 13.4% • golgi: 7.5% • nucleus: 7.5% • plasma membrane: 7.5% • vacuole: 7.5% • cytosol: 1.5%
Protein-Protein	At1g25420	At5g22770	ISTL2 Regulator of Vps4 activity in the MVB	alpha-ADR alpha-adaptin	1	0.699		0064 (interologs mapping)	• mitochondrion: 55.0% • golgi: 25.0% • nucleus: 20.0%	• plasma membrane: 36.6% • cytosol: 24.4% • nucleus: 22.0% • golgi: 12.2%

Figure 2.4. Table output of AIV2. This table has multiple dynamic options such as searching and filtering through columns (with combinatory logic). Here the original set of 95 interactions is reduced to 35 as the user sets a filter to only show interactions that have a PCC above 0.35.

2.2 AIV2 Use Cases

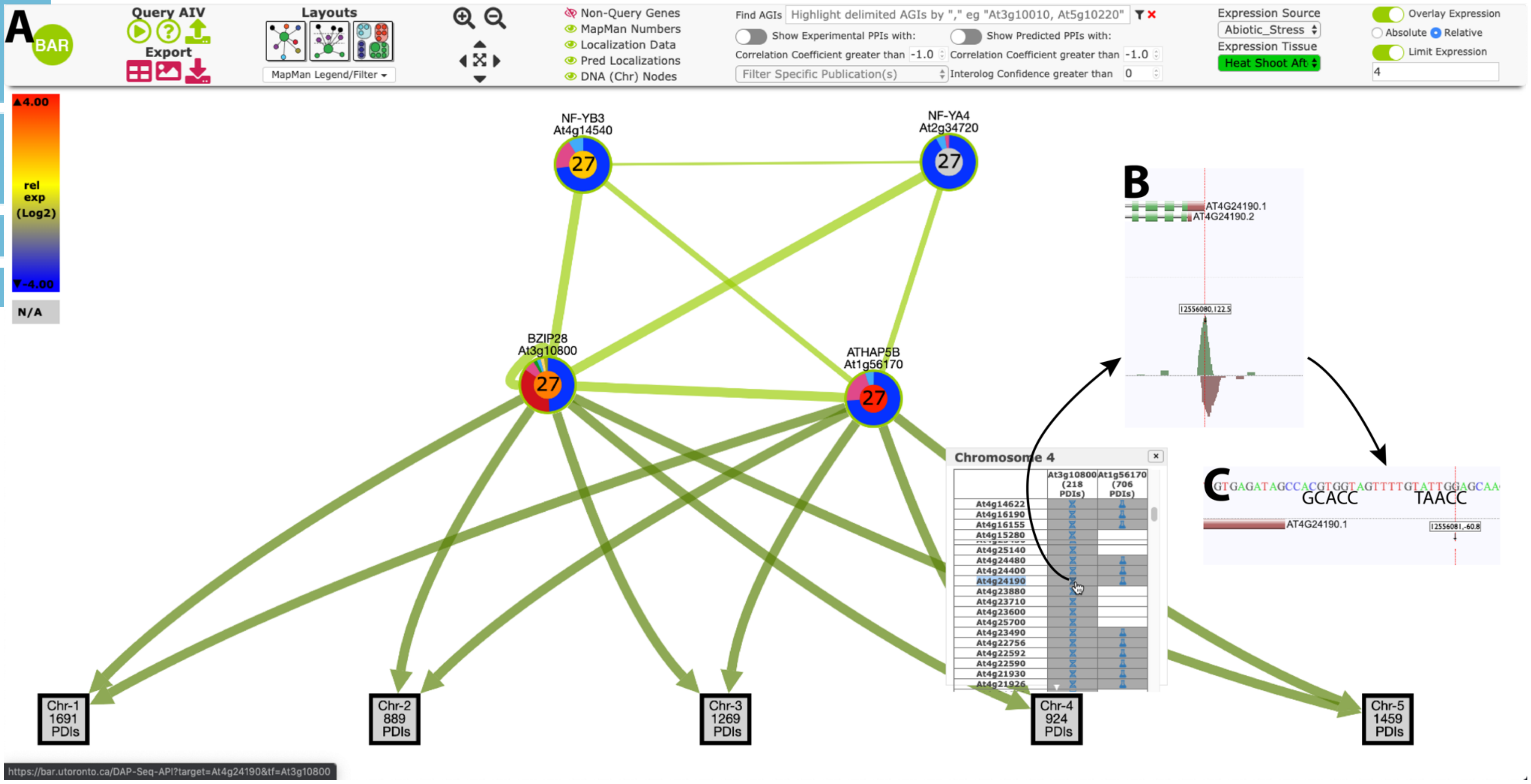
2.2.1 Characterizing Transcription Factor Complexes

TFs can complex with another to regulate target genes that they otherwise could not alone. For example, MYB TFs complex with bHLH TFs and TTG1 proteins to form a ternary complex to regulate flavonoid biosynthesis genes via binding to specific regulatory sequences, such as G-boxes (Xu, Dubos and Lepiniec, 2015). Other sets of genes are activated via other specific *cis*-acting elements and their respective TF complexes. For example, when endoplasmic reticulum (ER) stress accumulates via unfolded proteins, the unfolded protein response (UPR) is induced to activate a set of genes required for efficient transport and degradation of unfolded proteins (Iwata and Koizumi, 2012). UPR gene regulation is mediated via the endoplasmic reticulum stress-responsive element (ERSE; consensus sequence CCAAT-N9-CCACG). The UPR is initiated when environmental stresses (heat, chemical) accumulate sufficient levels of unfolded proteins which then outcompetes Binding Protein’s (BIP) binding to bZIP28 (AT3G10800). BIP then releases the TF from the ER to be transported the nucleus (Srivastava, Deng and Howell, 2014; see Appendix 8). bZIP28 can then target genes necessary for the UPR. bZIP28 unlikely acts alone, as bZIP28 was to shown to complex with the Nuclear Factor Y (NF-Y) subunits NF-YA4 (AT2G34720), NF-YB3 (AT4G14540), and NF-YC2 (AT1G56170) in a yeast hybrid system and via BIFC based on Geisler-Lee *et al.*'s (2007) predictions (Liu and Howell, 2010). This complex binds to the ERSE *in vitro* but not when any of the protomers were tested independently. The NF-Y subunits (A,B,C) form a heterotrimeric NF-Y complex which bind to CCAAT boxes to regulate target genes (Zhao *et al.*, 2017). NF-Y complexes can also complex with other proteins to confer additional specificity, such as bZIP28. There are up to 1000 potential NF-Y combinations from the 30 predicted NF-Y members in Arabidopsis, but only a few NF-Y combinations have been verified. As the bZIP28-NF-Y complex has been relatively unstudied since Liu and Howell's (2010) work, AIV2 can assist in predicting potential NF-Y combinations, and finding target genes for this complex.

To test this use case, bZIP28 and NF-Y members were entered as query genes into AIV2 with default settings along with “Search for interactions between interactors” checked. The output is seen in Figure 2.5A where the non-query genes are hidden for simplicity and the expression profiles for “Heat Shoot After 1 Hour” was overlaid on the protein nodes. Unsurprisingly, this TF complex was already documented in AIV2 as each TF has been validated to bind to another (lime green edges distinguish experimental support). As expected, bZIP28 shows high (~4-fold) induction during heat stress while NF-YC2 shows even higher (~11-fold) induction. Interestingly, NF-YB3 only shows a ~2-fold increase (NF-YA4 did not have any data available) relative to the 0h unstressed, control (expression data derived from Kilian *et al.* (2007)). Although these TFs all show upregulation which suggests coordination, the magnitude of induction may suggest some members such as NF-YC2 complex with other NF-Y subunits and TFs to regulate sets of genes in response to heat. To investigate further, I looked into multiple time points after heat shock (see Appendix 9). We see that there is still strong concomitant expression 3 hours after heat shock, however the protomers' expression return to baseline levels at 6 hours after treatment which suggests this complex undergoes coordinated regulation. It also hints that this TF complex is highly active only during the beginning of the UPR.

To identify new members of this complex, we can use the 'guilt-by-association' co-expression approach to find potential interactors. After overlaying the 1h-post heat shock on top of the whole PPI network (see Figure 2.6), a few (highlighted) nodes are of particular interest. As discussed, BIP is shown to interact with bZIP28, localizes (see Appendix 10 for localization layout) in the ER, and shows high expression during heat shock – likely to manage unfolded proteins. ATTBP2 (TATA Binding Protein 2) is also predicted to interact with NF-YB3 (interolog confidence of 8), which suggests that NF-YB3 recruits TBP2 and thus the transcription initiation machinery to ERSE containing genes as TBP2 is functionally equivalent to TBP1 (Heard, Kiss and Filipowicz, 1993). NF-YA7 is also highly expressed and interacts with NF-YB3 and NF-YC2. However, based on data available in AIV2, NF-YA7 does not interact with bZIP28. As the B and C domains confer protein-protein binding to other TFs (Zhao *et al.*, 2017), it is possible that NF-YA7 complexes with bZIP28 indirectly. This remains a possibility as Liu and Howell (2010) did not screen for interactions with NF-YA7. Moreover, NF-Y subunits have been shown to be functionally redundant as a double NF-YA mutants were shown to be lethal, but single mutants were not (Fornari *et al.*, 2013). Last, NF-YA7 shows a similar expression profile to the other bZIP28-TF-Y members when looking at other post-heat shock timepoints. Hence, I propose that NF-YA7 is a potential functional substitute for NF-YA4 in the bZIP28-NF-Y TF complex.

To find or validate a TF’s targets, we can scan through AIV2’s ~3 million PDIs. Liu and Howell (2010) tested specific genes such as SHD to be targeted by bZIP28 by looking at their expression in bZIP28 mutants after chemical treatment. However, the authors did not validate whether bZIP28 physically binds to the promoter of SHD. In AIV2, one can quickly scroll through a TF’s targets and if any are available, linkout to O’Malley *et al.*'s (2016) DAP-Seq genome browser for further investigation. In Figure 2.5A, AIV2 shows that bZIP28 does indeed bind with SHD’s promoter amongst many other genes. Interestingly, NF-YC2 also showed a PDI with SHD – further validating that the bZIP-NF-Y complex regulates SHD. Following Shneiderman's (1996) mantra of details-on-demand, clicking on the DAP-Seq linkout displays a large DAP-Seq peak near the TSS of SHD (see Figure 2.5B). Last, zooming in near the peak shows that the ERSE is also located near the TSS of SHD (see Figure 2.5C). Therefore, it is likely the bZIP28-TF-Y complex binds to ERSE near the TSS and recruits TBP2 to help initiate transcription of SHD after heat shock.



Chromosome 4

	At3g10800 (218 PDIs)	At1g56170 (706 PDIs)
At4g14622	⧗	⚗
At4g16190	⧗	⚗
At4g16155	⧗	⚗
At4g15280	⧗	
[illegible]	⧗	
At4g25140	⧗	
At4g24480	⧗	⚗
At4g24400	⧗	⚗
At4g24190	⧗	⚗
At4g23880	⧗	
At4g23710	⧗	
At4g23600	⧗	
At4g25700	⧗	
At4g23490	⧗	⚗
At4g22756	⧗	⚗
At4g22592	⧗	⚗
At4g22590	⧗	⚗
At4g21930	⧗	⚗
At4g21926	⧗	⚗

Figure 2.5. A) Output of AIV2 when queried for bZIP28, NF-YA4, NF-YB3, and NF-YC2 with default settings along with searching between interactors checked. Non-query genes were hidden via the "Non-query Genes" feature and the expression profiles for " Heat Shoot After 1 Hour" were overlaid onto the protein nodes with a maximum of Log216 (4; 16 fold increase) set. When hovering on the fourth chromosomal node, a tooltip showing all the PDIs for the query genes is shown. Note that the table was modified to see the TFs (column headers). Clicking on the helix-icon for bZIP28 and SHD (AT4G24190) redirects the user to the DAP-Seq genome browser as seen in B. B) Cropped image of the DAP-Seq genome browser near the TSS of the SHD. C) Nucleotide resolution of the TSS of SHD, with the ERSE sequence reverse transcribed and overlaid. Browser was zoomed in from the DAP-Seq peak in B.

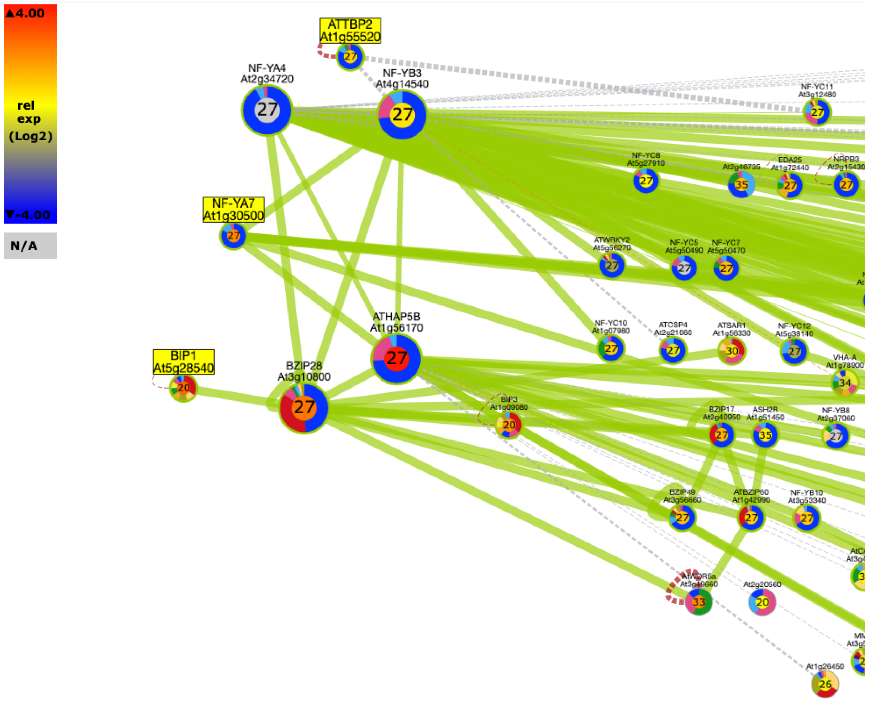
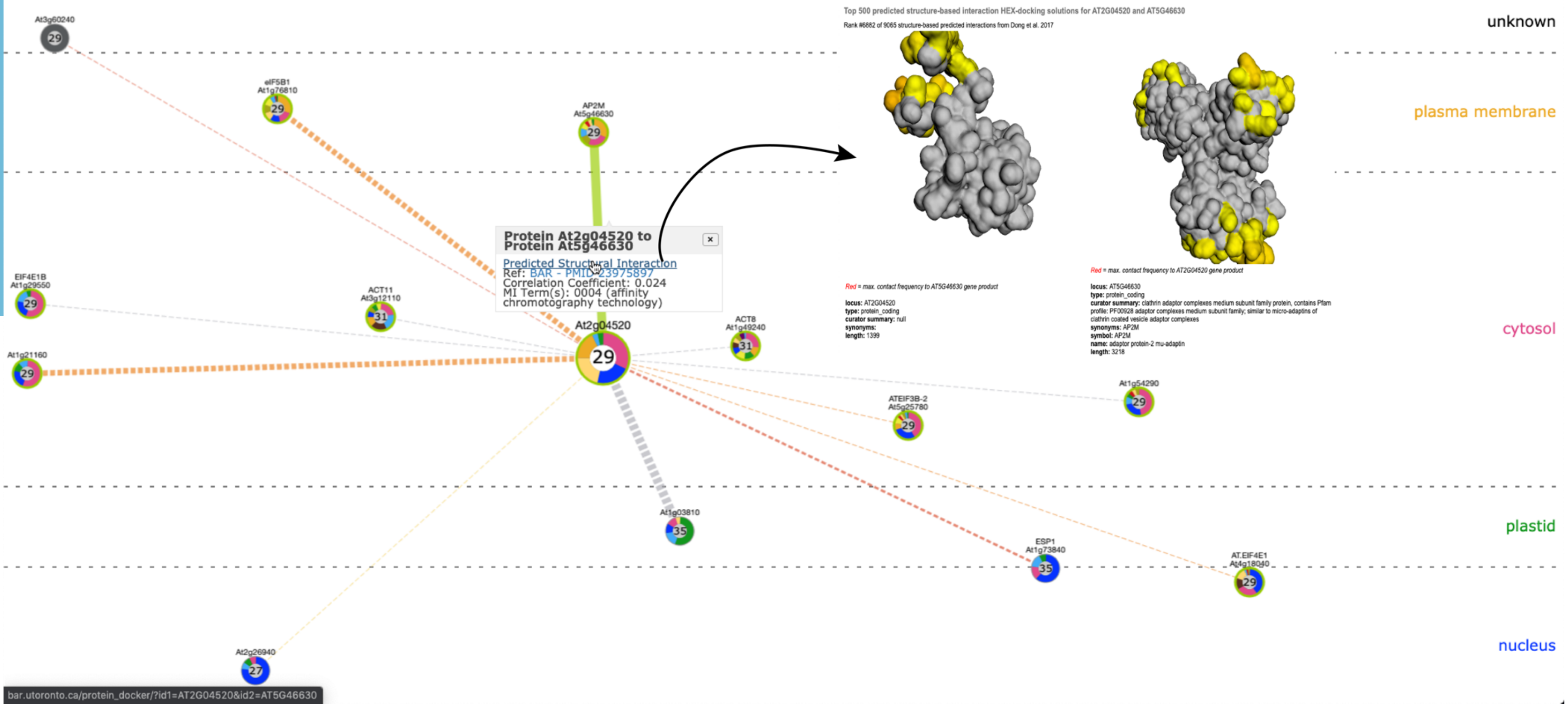


Figure 2.6. Output of AIV2 when queried for bZIP28, NF-YA4, NF-YB3, and NF-YC2 with settings as mentioned in Figure 2.5. DNA nodes are hidden and proteins of interest are highlighted with the find gene feature (yellow rectangles). Other genes were partially cropped out.

2.2.2 Predicting Protein-Protein Interactions

Although the previous example required *a priori* knowledge regarding the protomers, AIV2 can aid researchers in finding PPIs regarding their protein of interest by visualizing our predicted PPIs. For example, AT2G04520 is an uncharacterized protein which has several predicted PPIs (see Figure 2.7A) from interolog inference (Geisler-Lee *et al.*, 2007) and our structure-ome pipeline (Dong *et al.*, 2019). In addition to the examples of using expression profiles and localization to select the most probable PPIs, I have created filters to hide predicted PPIs by their predictive value (see Figure 2.2; Item H) which allow users to dynamically find strongly-predicted interactions in an RSVP fashion. Furthermore, several of the PPIs share a similar MapMan ontology (protein synthesis initiation) to AT2G04520 which support our predicted PPIs. Additionally, a user may also hover on the PPI and linkout to Richard Song's web tool which visualizes the predicted structures and their predicted docking interfaces (see Figure

2.7B). As shown, in the tooltip one of our predicted interactions between AT2G04520 and AP2M which is involved in plant growth, floral organ development, and Effector-triggered immunity (Hatsugai *et al.*, 2016) was validated via immunoprecipitation (Yamaoka *et al.*, 2013) thereby linking AT2G04520 to the processes AP2M is involved in. Last, Figure 2.7A demonstrates AIV2's subcellular localization 'cake' layout which applies Gestalt's Law of Familarity (Nesbitt and Friedrich, 2002) where "things are more likely to form groups if the groups appear familiar or meaningful" by grouping nodes by their subcellular localization in a layered, ordered fashion.



Protein At2g04520 to
Protein At5g46630

Predicted Structural Interaction
Ref: BAR - PMID 23975897
Correlation Coefficient: 0.024
MI Term(s): 0004 (affinity chromotography technology)

bar.utoronto.ca/protein_docker/?id1=AT2G04520&id2=AT5G46630

Top 500 predicted structure-based interaction HEX-docking solutions for AT2G04520 and AT5G46630

Rank #6882 of 9065 structure-based predicted interactions from Dong et al. 2017

Red = max. contact frequency to AT5G46630 gene product

locus: AT2G04520
type: protein_coding
curator summary: null
synonyms:
length: 1399

Red = max. contact frequency to AT2G04520 gene product

locus: AT5G46630
type: protein_coding
curator summary: clathrin adaptor complexes medium subunit family protein, contains Pfam profile: PF00928 adaptor complexes medium subunit family; similar to micro-adaptins of clathrin coated vesicle adaptor complexes
synonyms: AP2M
symbol: AP2M
name: adaptor protein-2 mu-adaptin
length: 3218

Figure 2.7. A) Output of AIV2 when given AT2G04520 as a query gene with default settings. The layered subcellular localization layout was utilized. Tooltip shown appears when a PPI edge is hovered. When an edge contains a predicted structural prediction, the user can linkout to an external page as seen in B. B) Minified web page of the predicted structural interaction where the interactors- predicted structure is visualized. The coloured portions of the proteins represent areas of likely contact.

2.3 Discussion and Future Directions

I have shown how to use AIV2 for those who wish to further characterize a fairly well-studied protein (bZIP28) by integrating its PPI members and expression data from biological functions bZIP28 is involved in. Not only did I validate prior knowledge regarding the bZIP28-TF-Y complex but I have predicted a novel interactor and its transcriptional mechanism (recruiting TBP). This use case highlights that by following Shneiderman's (1996) mantra, we can elicit optimal results for hypothesis generation. Specifically, instead of displaying all the PDIs as separate nodes as in AIV, condensing the PDIs into a tabular format allows users to easily scan for potential combinatory TFs that bind to a target of interest (see Figure 2.5B). The linkout to O'Malley *et al.*'s (2016) tool also allows nucleotide resolution of protein-DNA binding to investigate potential regulatory sequences. Additional options such as exporting the data will also allow researchers to analyze the PPI data in other ways. For example, performing a GO enrichment analysis on the PDI targets of an uncharacterized TF will allow one to begin to understand what role(s) that TF is involved in.

Although AIV2 is still under beta-testing, it has received 6,500 views in 2019, which is much lower than AIV (~80,000; https://bar.utoronto.ca/awstats/awstats.pl). This discrepancy is likely due to external links to our older tool. In preparation for Flash's obsolesce by the end of 2020 and thus AIV1, the BAR must update the AIV URL to redirect to AIV2. Furthermore, to allow a streamlined UX (Hoa, 2015), I designed AIV2 to share a few elements of AIV1 such as the form input. However, I intentionally incorporated newer features such as the condensed chromosomal nodes and donut pie-charts that also exist in our newer apps such as ePlant (Waese *et al.*, 2017), as we will want our users to be accustomed to our newer design language. Most importantly to those who are migrating, AIV2 hosts all the features AIV had. New features such as including Dong *et al.*'s (2019) predicted S-PPIs, importing networks, linkouts, filtering options, descriptive tooltips, and offering multiple layouts such as the stacked localization layout (see Figure 2.7) will allow researchers to have additional tools to generate hypotheses.

However, when I began exploring AIV2, I noticed there were certain quality-of-life features which would make for a more enjoyable UX:

- Create a filter based on expression level to easily hide nodes that are not highly expressed (or attenuated). Currently the only option is to visually scan for highly expressed nodes which can be difficult when the network is large for those who with colour vision deficiencies.
- Use an alternate tooltip library as the existing one is fairly slow.
- Allow filtering of the table to also subsequently filter the network.
- Decrease edge size dynamically based on network to avoid ‘hairball’-like layout.
- Further optimize the layered subcellular localization layout such that the nodes are not randomly sorted with in a layer.

Major improvements that I could implement that would allow for alternate methods of hypothesis generation could include:

- Integrate a GO analysis tool to give users a sense of what their functions their PPI network is involved in
- Include other types of interactions (genetic, metabolic) between genes.
- Begin collaborating with other databases to mutually share PPIs like BioGRID (Oughtred *et al.*, 2019).

Once the above features are implemented, the BAR's AIV tool will even further become the premier web-based Arabidopsis PPI query tool.

Chapter 3
Arabidopsis Gene Network Tool (AGENT)

3 Arabidopsis Gene Network Tool

3.1 Materials and Methods

3.1.1 Data Sources, Libraries, and Tools

To overhaul and integrate our original AIV2 interactions database with new interactions that are present in gene regulatory networks (GRNs), I used MySQL Workbench to design a MySQL 8.0 entity relationship diagram (ERD). See Appendix 11 for the ERD and Table 3.1 for rationale for each database table. To summarize, the new database design reduces redundancy of the older AIV2 database (for example, I normalized the interolog scores to a single table) while also integrating new features to summarize a network of interactions as a whole. For example, I created an attribute (See Appendix 11, cyjs_layout in the external_source table) that stores the preferred default layout for a network which allows us to load an aesthetically pleasing, efficient network layout based on the network's size and dynamics. Another key new feature is storing the modality of an interaction (i.e. activation or repression) which is increasingly able to be captured, as in the Brady *et al.* (2011) stele-enriched GRN.

To add GRNs to our refactored database, Rachel Woo (a Provart Lab undergraduate student) and I performed a literature search of Arabidopsis-focused GRNs searching for publications with keywords such as "gene regulatory network", "transcriptional network", "gene network" along with "Arabidopsis" in PubMed. 12 GRNs were then digitally transformed into a simple interaction format (SIF) file (see https://github.com/raywoo32/grnAnnotation and https://github.com/VinLau/AGENT-GRNs for the files). Under my supervision, Rachel Woo used nodeJS libraries (knexjs, shelljs) to create a script which deposits the interaction and network data into the database. The source code is publicly available and can be accessed publicly at https://github.com/raywoo32/readSIF. Note that the SIF files have custom headers for curator data such as the reference, network description, and custom tags which will also be stored in the database. An example of our customized SIF file can be accessed via https://github.com/raywoo32/readSIF/blob/master/test/example.sif (see Appendix 14).

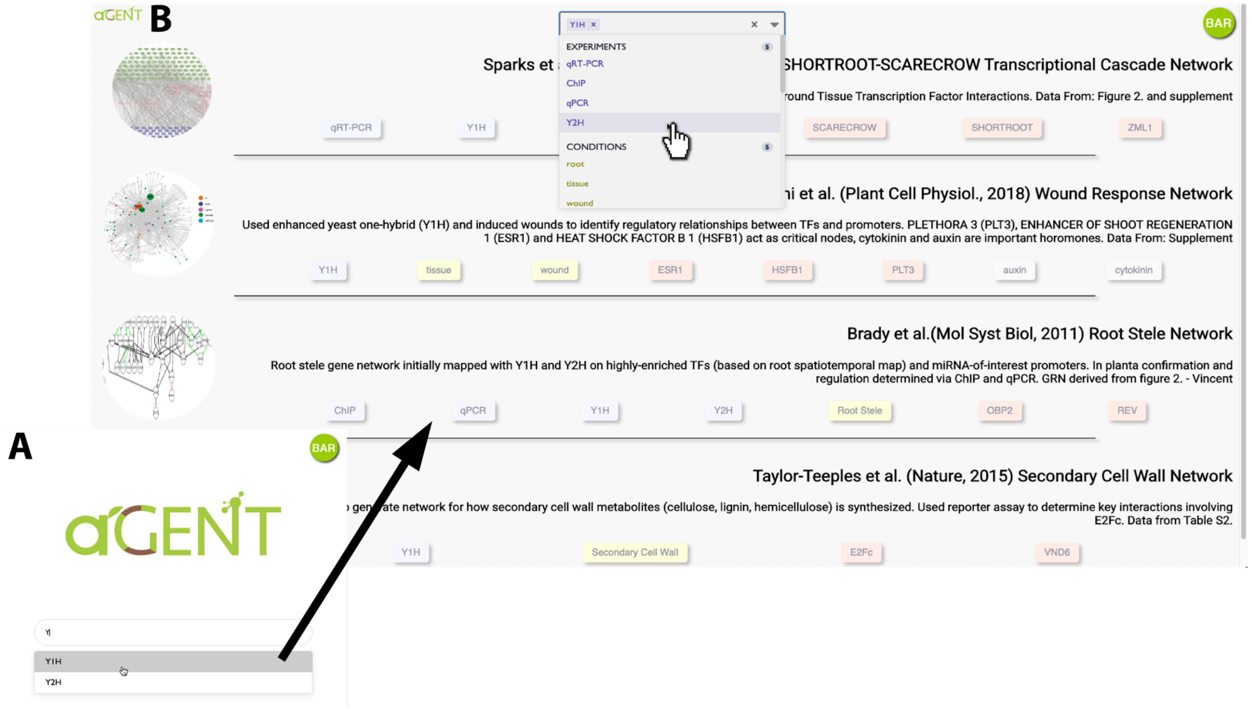
To design the API to fetch the interactions from GRNs, I used knexjs as a database connector along with Express.js to create a REST (Representational state transfer) API with multiple endpoints that executes different MySQL queries. For example, the API endpoint https://bar.utoronto.ca/interactions_api/tags/flower retrieves all the GRNs that are tagged with 'flower' (i.e. networks related to flower development) using a specialized MySQL query. See Appendix 13 for an example JSON output from such a query. Similarly, this API can return GRNs based on the nodes (i.e. genes) contained in them. For instance, https://bar.utoronto.ca/interactions_api/gene/AT1G01010 returns all GRNs that contain the AGI ID "AT1G01010". This endpoint allows us to link out networks based on a gene search, and therefore will enable ePlant integration. A GRN's interactions can then be retrieved via https://bar.utoronto.ca/interactions_api/papers/14/interactions (14 is a unique identifier for a specific GRN). Other APIs used so far in AGENT are the SUBA4, AIV2, expression and gene annotation APIs as mentioned in the previous chapter.

To create an API that highlights motifs in a given network, I used Alon's (2007) mfinder tool which finds enriched motifs in a network when compared to randomized networks. To do so, I used Express.js along with shelljs to create an API which executes mfinder 1.21 on the BAR server via command line. It then returns a list of enriched motifs based on the network and parameters set. Configurations for mfinder parameters (directed, r = 100, s = 3, u = 4, z = 2) were recommended by the authors and were also optimized for speed as this API is made for the web. The API can be accessed via https://bar.utoronto.ca/mfinder.

To build AGENT's frontend, I used React (https://reactjs.org/) as a framework to build the UI. React was chosen as it is a mature modern framework with many resources and is well maintained by Facebook. The React framework efficiently renders HTML elements when users interact with the UI via a 'diffing' algorithm of the current webpage and the 'future' webpage. Additionally, it has many plugins which allow for extended functionality such as the React-Router library which allows app pagination/redirection (essential for ePlant integration). As with AIV2, Cytoscape.js was chosen to visualize networks. The following Cytoscape.js plugins were also used: cytoscape-context-menus, cytoscape-popper, cytoscape-klay, and cytoscape-cose-bilkent. For tooltips, I used Tippy.js which also integrates with React and Cytoscape.js. To implement AGENT as a standalone application along with ePlant integration, I used HTML5's iframe technology. The source code for AGENT can be accessed via https://github.com/VinLau/AGENT.

3.1.1.1 Application Logic and Design

Unlike AIV2, AGENT does not have a legacy application. Therefore, I had the freedom to design the UI without consideration to prior designs as long as it fits BAR's universal design language. However, I had to consider a design and technological architecture to build a standalone app that also can integrate with ePlant. Therefore, I chose a single-page app (SPA) architecture that features pagination such that URLs that are returned by a gene query can be redirected to ePlant. To illustrate this, see for the landing page of standalone AGENT (Figure 3.1A; https://bar.utoronto.ca/AGENT/). Once a user launches AGENT, s/he can query a GRN(s) via a direct gene search by AGI ID, interaction pair, or have a list of autocomplete suggestions based on our list of curated tags. My implementation of autocompletion follows modern UX standards such as avoiding scrollbars, visual distinction between suggestions, and reduced visual noise (Appleseed, 2019). Once a user enters a query or selects a suggestion, the app initiates an AJAX request to the AGENT API and is then redirected to different URL within the app which then displays a list of GRNs (see Figure 3.1B; https://bar.utoronto.ca/AGENT/list/tag/y1h). Additionally, this app can directly load the required data when given a URL, which allows ePlant integration (see Figure 3.2; http://bar.utoronto.ca/~vlau/eplant). Specifically, when a user searches for a gene in ePlant, it will load the https://bar.utoronto.ca/AGENT/list/gene/AGI_ID (where AGI_ID is an AGI ID) URL inside ePlant as a separate webpage using iFrame technology (Figure 3.2). Thus, I also designed AGENT with responsive web design principles (i.e. to fit many screen resolutions) as ePlant reduces the available screen resolution when AGENT is running in an iFrame.



Sparks et ; SHORTROOT-SCARECROW Transcriptional Cascade Network

round Tissue Transcription Factor Interactions. Data From: Figure 2. and supplement

ni et al. (Plant Cell Physiol., 2018) Wound Response Network

Used enhanced yeast one-hybrid (Y1H) and induced wounds to identify regulatory relationships between TFs and promoters. PLETHORA 3 (PLT3), ENHANCER OF SHOOT REGENERATION 1 (ESR1) and HEAT SHOCK FACTOR B 1 (HSFB1) act as critical nodes, cytokinin and auxin are important horomones. Data From: Supplement

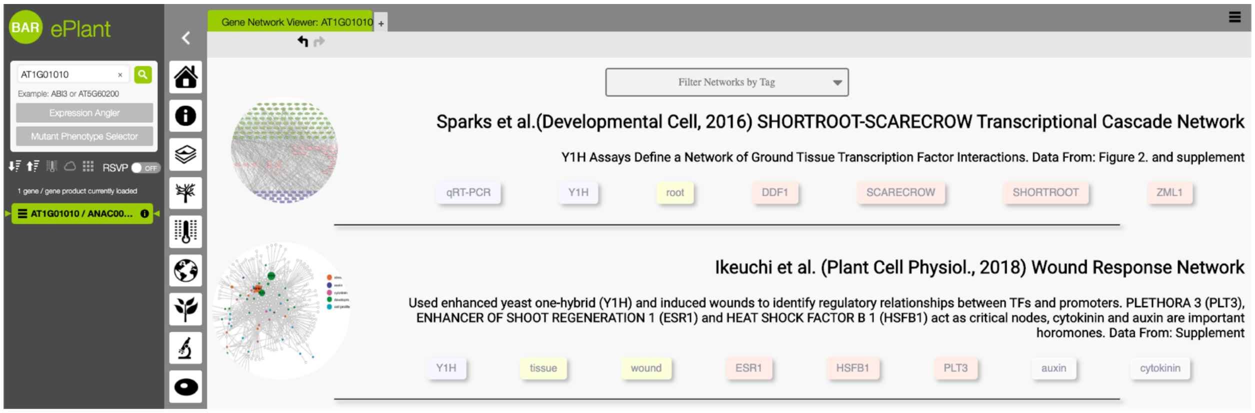
Brady et al.(Mol Syst Biol, 2011) Root Stele Network

Root stele gene network initially mapped with Y1H and Y2H on highly-enriched TFs (based on root spatiotemporal map) and miRNA-of-interest promoters. In planta confirmation and regulation determined via ChIP and qPCR. GRN derived from figure 2. - Vincent

Taylor-Teeples et al. (Nature, 2015) Secondary Cell Wall Network

gen ate network for how secondary cell wall metabolites (cellulose, lignin, hemicellulose) is synthesized. Used reporter assay to determine key interactions involving E2Fc. Data from Table S2.

Figure 3.1. The resultant webpages shown when navigating AGENT when the user A) Opens AGENT as a standalone application and enters 'Y' in the search box. B) Selects a recommended, curated tag such as ‘Y1H’.


BAR ePlant
AT1G01010
Example: ABI3 or AT5G60200
Expression Angler
Mutant Phenotype Selector
RSVP
OFF
1 gene / gene product currently loaded
AT1G01010 / ANAC00...
Gene Network Viewer: AT1G01010
Filter Networks by Tag

Sparks et al.(Developmental Cell, 2016) SHORTROOT-SCARECROW Transcriptional Cascade Network

Y1H Assays Define a Network of Ground Tissue Transcription Factor Interactions. Data From: Figure 2. and supplement

Ikeuchi et al. (Plant Cell Physiol., 2018) Wound Response Network

Used enhanced yeast one-hybrid (Y1H) and induced wounds to identify regulatory relationships between TFs and promoters. PLETHORA 3 (PLT3), ENHANCER OF SHOOT REGENERATION 1 (ESR1) and HEAT SHOCK FACTOR B 1 (HSFB1) act as critical nodes, cytokinin and auxin are important horomones. Data From: Supplement

Figure 3.2. Cropped ePlant prototype which renders AGENT's list of GRNs based on the search query submitted to ePlant.

When the user then arrives on the GRN list webpage (see Figure 3.1B), s/he can explore the GRNs by their curated descriptions and tags. Moreover, tags are colour coded and categorized into either: experiment (types), condition, (important) genes, or miscellaneous. Colour coding these categories captures the user's attention (Brown, 1999) and allows one to perform quick estimations. For example, a user can quickly estimate the number of techniques used to generate a GRN by the number of red "experiment" tags. Last, users can also filter the GRNs displayed using combinatory logic of the tags (see Figure 3.1B, cursor).

A user then clicks on their desired GRN, AGENT will make a request to the AGENT interactions API at https://bar.utoronto.ca/interactions_api/papers/X/interactions where X is a unique identifier for each GRN. The user is now redirected to the URL: https://bar.utoronto.ca/AGENT/network/X. AGENT will then parse through an array of interactions and layout the metagene nodes along with additional information such as localization and gene names in a similar fashion to AIV2. AGENT also initializes the mFinder AJAX request soon after, as it takes some time to execute mFinder. AGENT will also forward the GRN description data from the previous webpage for additional processing such as applying the default layout. See Figure 3.3 for AGENT's default output when a user selects on Keurentjes *et al.*'s (2007) flowering-time GRN (see Appendix 12 for original reference). Note that there are slight design deviations from AIV2 in that SUBA4 data is not visualized as a donut pie-chart. Instead, I summarized the data in the node-border with the most prominent localization representing the colour and the width representing the relative prominence. However, following the details-on-demand principle (Shneiderman, 1996) users can hover on the gene node to reveal a stacked bar-chart instead of the donut pie-chart. As we store the modality of the network, we can visualize whether a TF activates or represses its target by representing its edge with an arrowhead or a “T” respectively, which follow typical GRN conventions.

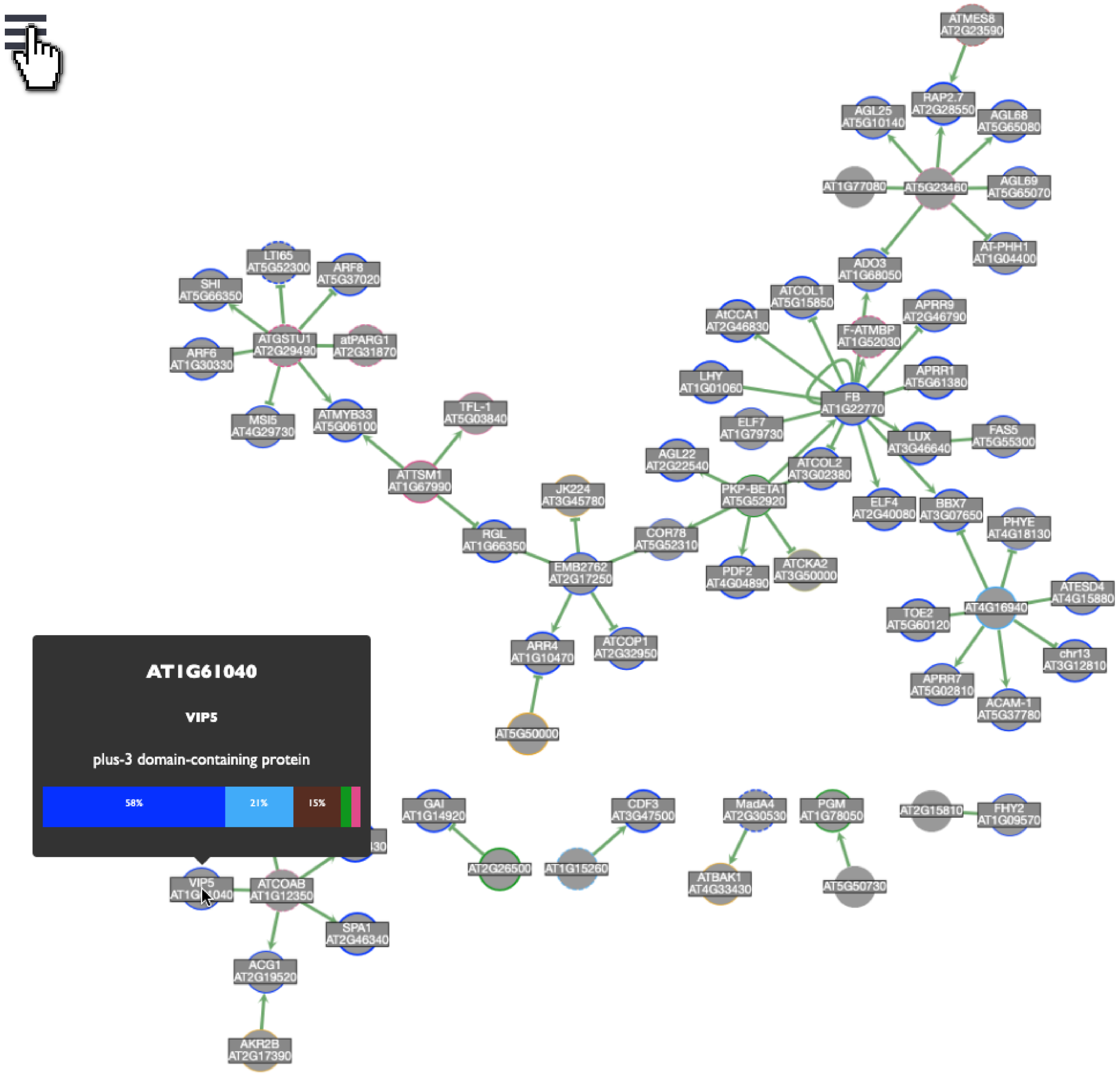


Figure 3.3. Default output of AGENT when user selects on Keurentjes et al.'s (2007) flowering-time network. The user has hovered over a gene node (VIP5) which displays a tooltip with gene annotation and localization data summarized in a stacked bar chart.

Users can then select on the "burger icon" (see Figure 3.3, cursor) to reveal additional investigative tools (see Figure 3.4 for summary). This minimal sidebar can be hidden to allow for maximal use of screen space to render the network, which is key for visualizing larger networks in ePlant. As an alternative to the event-listener architecture used in AIV2, I used React's native Context feature which allows UI elements to directly modify the network via a global 'state'. Cytoscape.js can also communicate to UI elements by using React Context's publish-subscribe functionality (Brudnicki, 2018). For example, when a node is deleted, Cytoscape.js triggers an event that changes the global state, which then updates specified UI components, such as removing that node from a dropdown menu. This is an experimental feature as most UI elements in web development are unidirectional in their behaviour. Novel features that are not present in AIV2 include resizing nodes based on degree centrality (Figure 3.4, Item D), similar to what Vallabhajosyula *et al.* (2009) have suggested, calculating the shortest path between two nodes (Figure 3.4, Item G), and "scrubbing" through motifs that mFinder finds (Figure 3.4, Item H). To calculate the (normalized) degree centrality and shortest path, we used Cytoscape.js' native centrality and shortest path (Dijkstra's algorithm) functions. To display mFinder's output, we used the rc-slider React component which has an event listener to fire an animation to highlight the nodes which are members of the same motif. Last, we can also load interactions from AIV2 or delete nodes (right-click menu, not shown).

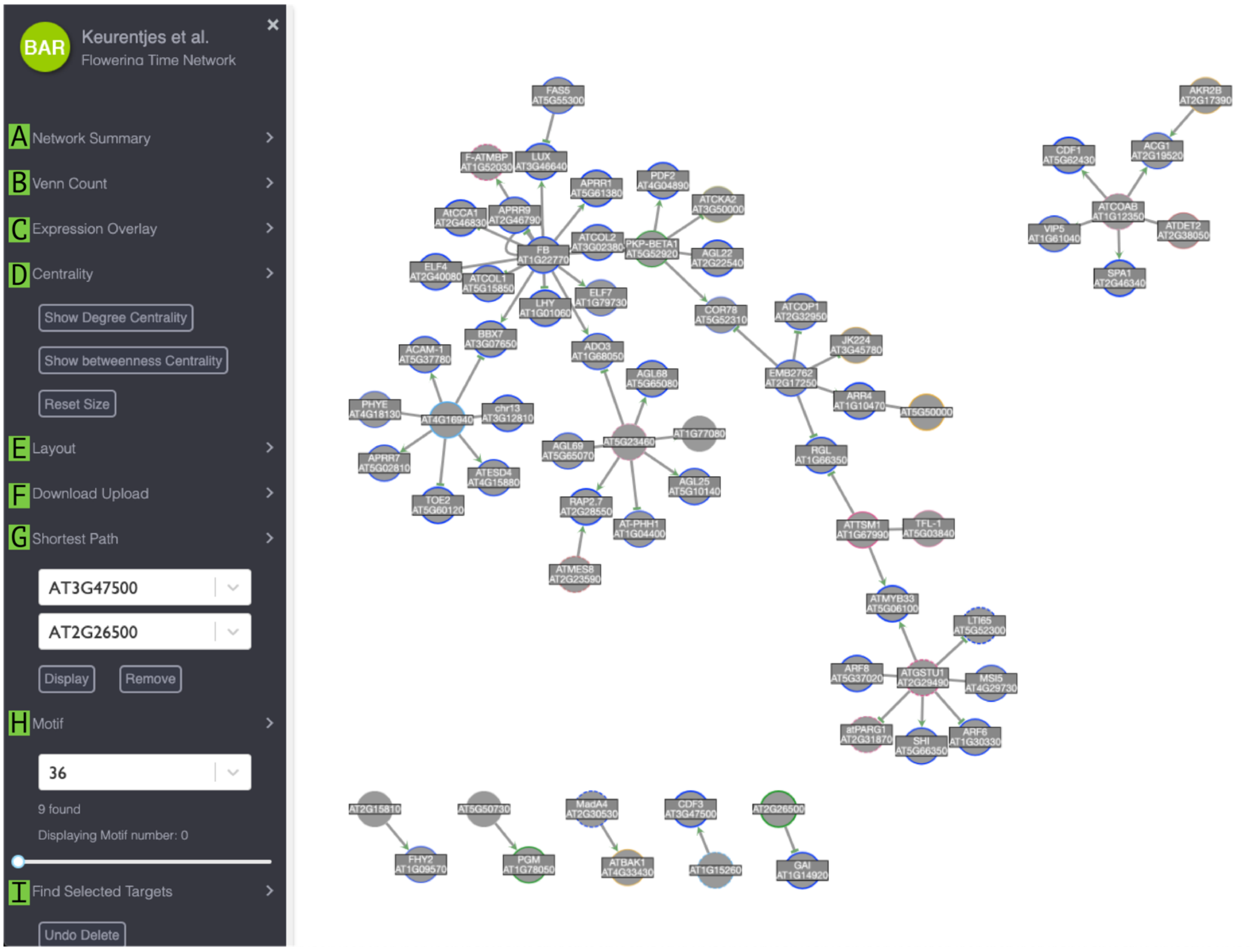


Figure 3.4. Output of AGENT when the sidebar is launched which features: A) A description of the network. B) A venn diagram displaying of nodes the overlap between two networks when loaded (feature still in development) C) Expression overlay options similar to AIV2 D) Buttons resizing the nodes in the network based on centrality measures. E) Different layout options. F) Export and import abilities. G) A tool to calculate the shortest path between two nodes of interest. H) A motif finder feature which hosts a slider to quickly scan through the numerous motifs. I) A tool to build a dynamic table similar to AIV's chromosomal node tooltip.

3.2 AGENT Use Cases

3.2.1 Identifying Feed-Forward Loops to Identify Key Regulatory Controls

In addition to using AGENT to quickly identify a gene of interest's regulators and targets, it can also search for smaller subnetwork motifs such as feed-forward loops (FFLs) which have been associated with floral commitment (Adrian, Torti and Turck, 2009). A FFL is relationship between three genes where TF A that regulates another TF B, which together regulate a target gene C (Alon, 2007). The most well-studied FFL is the coherent type-1 FFL (C1-FFL) in which TF A activates TF B which together upregulate gene C. C1-FFLs have been experimentally shown in *E. coli* to delay upregulation of the target gene as time is required for sufficient levels of the TF B to accumulate after TF A is activated (Mangan, Zaslaver and Alon, 2003). This network motif is believed to act as a persistence detector against noisy input signals. That is, there the target gene's upregulation is tightly controlled to protect against fluctuating inputs. Saddic *et al.* (2006) first explored FFLs in Arabidopsis by using ChIP along with expression profiling to identify that the TF, LEAFY (LFY) upregulates another TF, LMI1, to which both upregulate CAL, a known meristem identity regulator (see Appendix 15). They suggested that these genes exist in a C1-FFL as meristem identity transition was significantly delayed when LMI1 levels are reduced in a mutant with slightly reduced LFY activity. Indeed, they discuss how a higher threshold of LFY levels is needed for Arabidopsis to flower when the photoperiod is decreased (Blázquez *et al.*, 1997).

To look for important regulators of secondary cell wall biosynthesis in Arabidopsis, (Taylor-Teeples *et al.*, 2015) employed Y1H to generate a large GRN. In addition to finding a highly interconnected network, they also discovered a large number (96) of FFL motifs. Specifically, they found that the TFs, ATE2F2 (E2Fc) acts upstream with ATHB9 (PHV) amongst other TFs to regulate many genes, thereby creating FFLs involved in secondary cell wall biosynthesis. Although the authors provided a web-based network viewer for their GRN (https://gturco.github.io/trenzalore/stress_network), they did not include a tool to highlight their aforementioned FFLs. Highlighting all the FFLs can be useful in identifying key regulatory controls in secondary cell wall synthesis similar to how Saddic *et al.* (2006) reasoned that CAL expression is tightly regulated. Therefore, Rachel Woo and I have created a motif search tool in AGENT to quickly "scrub" through motifs such as FFLs. Users can quickly "scrub" through the playlist until their gene(s) of interest is highlighted as shown in Figure 3.5, where a FFL between E2Fc, PHV, and ATC4H is shown. A user may then hypothesize the role of his/her gene in relation to the other FFL members. For example, a dually targeted gene in the FFL could be under tight regulation as it is involved in turning on a developmental switch. It is possible that ATC4H (cinnamate 4-hydroxylase) is required in the commitment to secondary cell wall synthesis. However, as Taylor-Teeples *et al.* (2015) did not determine the modality of the interactions, the exact type of FFL and thus the mechanistic control cannot be presumed. Nonetheless, it seems like C1-FFLs are the vastly dominant form, at least in yeast and *E. coli* (Alon, 2007).

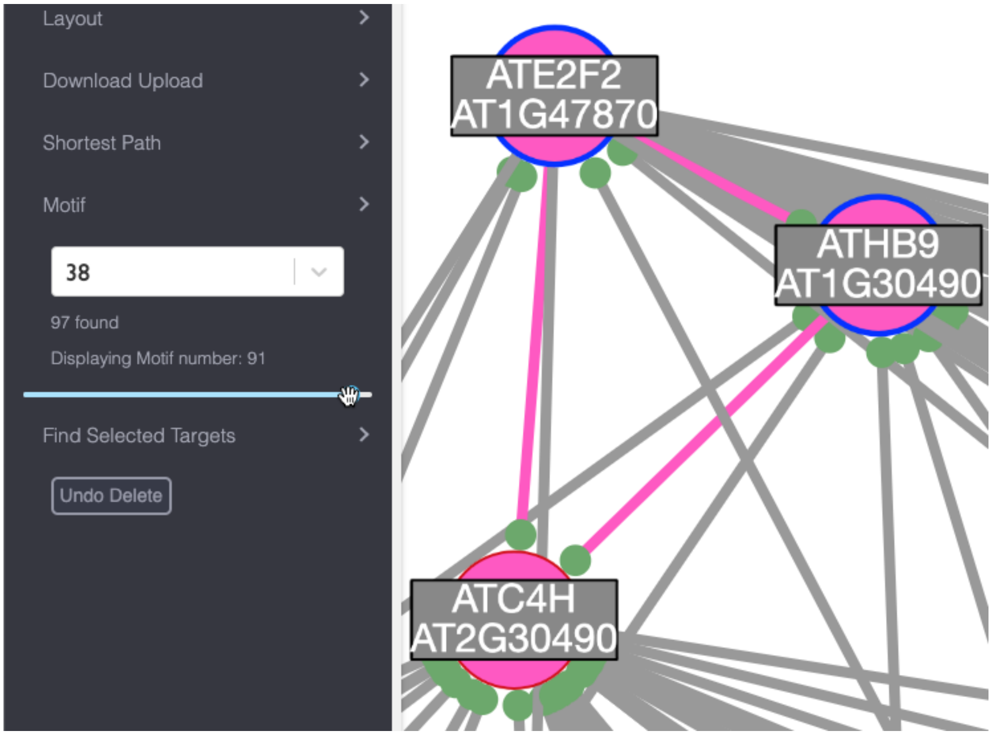


Figure 3.5. Cropped AGENT output of Taylor-Teeples et al.'s (2015) GRN when user searches for feed-forward loops (FFLs; which has an ID of 38) by using the ‘scrubbing’ tool. The user stopped at FFL 91 which identifies an FFL between E2Fc (ATE2F2), PHV (ATHB9), and ATC4H.

FFLs can be useful to researchers generating hypotheses regarding the tight control of key developmental genes. Although research involved in studying Arabidopsis-based FFLs is limited, other researchers are beginning to reference FFLs (Zhiponova *et al.*, 2014; Sakuraba *et al.*, 2015). Indeed, Jin *et al.* (2015) found an enrichment of FFLs amongst other network motifs in Arabidopsis developmental subnetworks which suggests that these motifs are critical in cell fate decision-making. Interestingly, earlier work performed Vidal *et al.* (2010) showed that miRNAs can be implicated in FFLs in Arabidopsis. Fortunately, AGENT can display miRNA interactions and FFLs between miRNAs.

3.2.2 Hypothesizing Novel Interactions by Integrating AIV2 Data

To contrast to AIV2 where the network is displayed without a context, AGENT displays a network of nodes for a particular tissue/condition that a researcher is interested in. From here, s/he can expand the network to infer additional novel interactions in that particular tissue/condition. Indeed, AGENT features the ability to load interactions dynamically from a right-click menu. As Gaudinier and Brady (2016) noted how certain methods can bias a GRN's network topology, it may be in a user's interest to load in AIV2 interactions to expand or validate a GRN. To demonstrate this, I investigated the Brady *et al.* (2011) root stele Y1H and Y2H GRN. To quickly narrow down important genes, I looked for 'hub' nodes by using the degree centrality tool to find well-connected nodes. I then overlaid the 'stele, standard conditions' expression profile onto the network. See Figure 3.6 for results. OBP2 (AT2G34710) was chosen as a candidate gene due to its high expression and degree centrality. I then loaded in the AIV2 interactions for PHB to reveal any uncaptured interactions. While only 6 PPIs/genes were returned, I was interested in seeing if any of these genes could be involved in interacting with other network members. I again loaded the AIV2 interactions for DRNL (AT1G24590) and found that DRNL not only interacts (via a PPI) with PHB but also PHV and REV. See Figure 3.7 for results (pink edges denote AIV2 interactions). As the authors did not investigate DRNL, it is possible that these DRNL dimerizes with these TFs in the root stele which can be investigated with methods such as immunoprecipitation or BIFC. Indeed, REV has been shown to bind to DRNL to upregulate meristem formation (Shimotohno and Scheres, 2019). Last, loading interactions dynamically validate a GRN's predictions by adding an additional edge between the two nodes (image not shown).

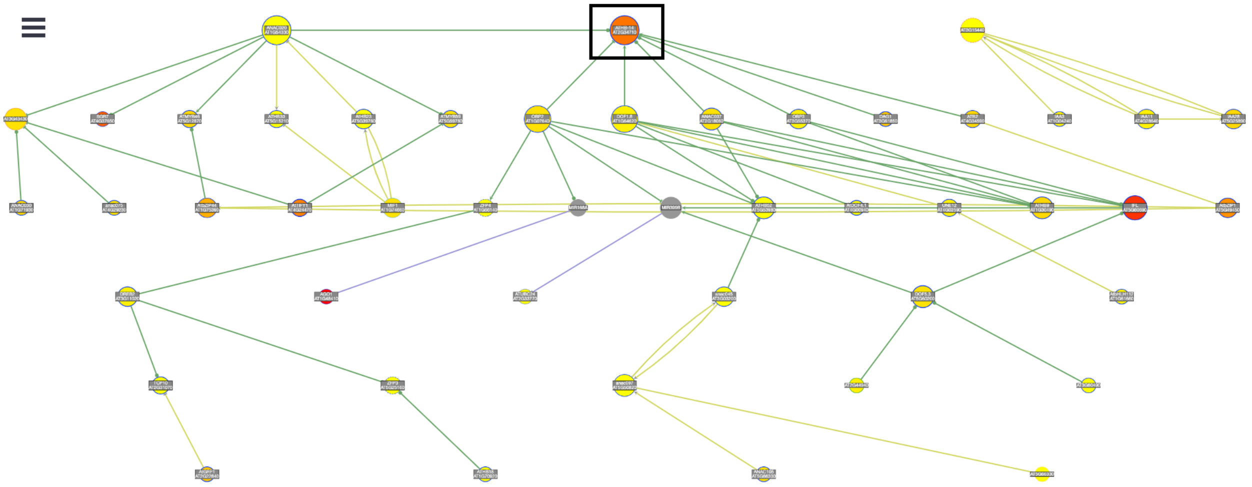


Figure 3.6. Brady *et al.*'s (2011) root stele GRN as rendered by AGENT. Nodes are resized according to their degree centrality. Node bodies are also coloured according to their gene expression under the tissue 'stele, standard conditions'. Black box: PHB (AT2G34710) is highlighted as a candidate gene for further exploration.

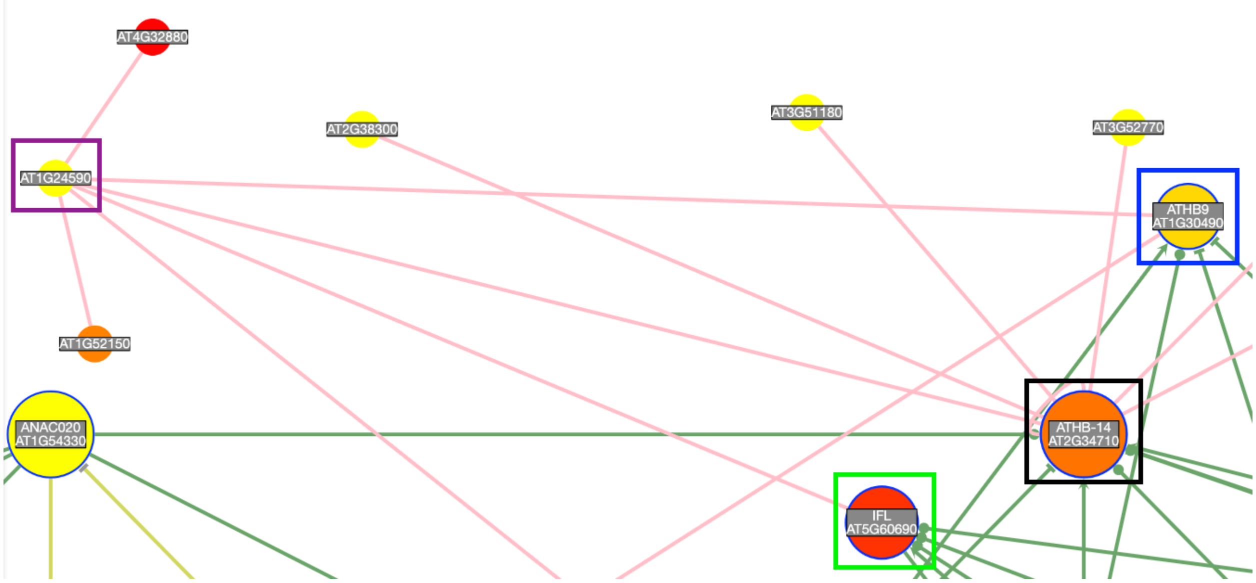


Figure 3.7. Cropped image of AGENT output when PHB's (black box) and DRNL's (purple box) interactions are loaded from AIV2 (pink edges). Blue box: PHV. Green Box: REV.

3.3 Discussion and Future Directions

Here I present a web-based Arabidopsis GRN viewer that hosts curated GRNs unlike TF2Network which requires a gene list (Kulkarni *et al.*, 2017). AGENT can be useful in quickly identifying a gene of interest's targets or regulators in a particular condition/tissue. Indeed, one can use our shortest path tool to see if two genes regulate one another. Additional features include centrality measures, expression overlays, importing and exporting GRNs, and finding motifs. To my knowledge this is the only web-based network viewer that has integrated a motif search. Although MotifNet (Smoly *et al.*, (2017); http://netbio.bgu.ac.il/motifnet) exists, it is a standalone tool that requires an input network. Furthermore, MotifNet does not highlight the motifs in a network viewer like AGENT but simply lists the motifs and their abundance/significance. By using a slider element along with highlighting the motifs (see Figure 3.5), users can quickly "scrub" through numerous motifs in the network to find potential motif hotspots in a network. Scrubbing through many motifs in a network enables rapid serial visual presentation (RSVP) of the data (Spence, 2002). AGENT is also valuable in that is a catalogue and viewer of published Arabidopsis GRNs versus traditional webservers which simply store a large collection of interactions (see Introduction). Although nDEX (Pratt *et al.*, (2015); http://www.ndexbio.org) is a species-agnostic web repository of user-submitted networks that allows users to search networks (like AGENT), the tool is much slower and does not feature network tools or species-specific annotations. Furthermore, AGENT users can filter through a network via combinatory logic of curated tags (see Figure 3.1B) to quickly find a network of interest, unlike nDEX. These tags can help identify the condition a GRN was studied in to highlight its context as Chudasama *et al.* (2018) has shown how different cancer types give rise to different GRNs. Indeed, Windram and Denby (2015) suggest that network topology can be modified in Arabidopsis in response to different environmental signals. Additionally, researchers can use AGENT's URLs to link to specific networks to share and reference networks. These links can also be displayed inside other webpages such as ePlant (see Figure 3.2) due to the SPA redirection architecture. With continual curation and improvements, AGENT can become the "next generation eFP Browser" (Winter *et al.*, 2007) in the sense that the eFP Browser is integrated into other Arabidopsis resources such as TAIR.

Previously, I showed how to utilize AGENT's motif tool to quickly highlight FFLs to hypothesize the role a gene(s) play in developmental switches. As AGENT also stores the modality of interactions, we can also determine the type of FFLs. From here, a user can investigate the role of the regulatory dynamics between the FFL members based on the interaction type. However, I again stress that such motifs are a niche field and have not been studied extensively in Arabidopsis. To remedy this, I could include a small tutorial explaining FFLs and their hypothetical roles in AGENT in the final version of AGENT's user manual. Another feature I plan to implement is to displayed what type of FFL is found when a user highlights it. In the future, once we compile enough networks, we can explore the frequency of occurrence of different kinds of FFLs as Milo *et al.* (2002) did in yeast and *E. coli*. Furthermore, Alon (2007) argued that network motifs undergo convergent evolution as species rewire their network motifs according to their unique environments, while the motif's genes themselves undergo conservative evolution. To test this hypothesis, another future direction could be to see if the same FFLs in Arabidopsis also in exist in yeast (when comparing homologues).

I also showed how loading AIV2's interactions into AGENT can allow researchers to hypothesize novel interactions to a network member, especially when a node can interact with many members of the network. To allow users to easily identify candidate genes for further investigation, tools such the degree centrality resizing and expression overlays exist. However, other measures of important genes in a network are currently under discussion, such as the "bottleneck-ness" (how well a node communicates between two large subnetworks) of a node in a network (Yu *et al.*, 2007). I am considering including other options for users to choose their preferred method to select candidate genes, such as bottleneck-ness. Furthermore, currently loading AIV2 interactions does not visualize the predictive strength of the interaction as in AIV2. I hope to implement this in a similar fashion to AIV2, using edge width to distinguish an interaction's predictive value or the number of publications that have measured an edge (interaction). Such a feature will give confidence to users who are investigating a novel interaction and wish to choose the best candidates.

Last, as AGENT is still in alpha and requires extensive user-testing, I will begin honing in the current design instead of adding too many features to avoid "featuritis" as Norman (2013) describes. Indeed, users seem to be more dissatisfied when faced with a large range of features and options (Schwartz, 2004; Gócza, 2015). To illustrate, consider how our most popular tool, eFP Browser simply does one thing (display expression data on a pictograph) but does it extremely well. To contrast, NAP (Theodosiou *et al.* (2017); http://bioinformatics.med.uoc.gr/NAP) has a host network analysis features that are native to Cytoscape.js but its UI is incredibly overwhelming, especially to those unfamiliar with graph theory. Below are some future optimizations I have for AGENT in addition to the ones mentioned above:

- Allow users to upload species-agnostic networks and analyze them with our motif search tool
- Allow users to only view PDIs or PPIs to simplify a network view
- Display expression profiles as a percentage of their maximum potential expression
- Include gene annotation (MapMan, GO) data

Of course, I should also foresee the decision to delete features if user-testing shows that users are overwhelmed by AGENT. Such a decision is not always a bad thing as it will make AGENT more lightweight and thus easier to integrate into other webpages.

3.4 Tables

Table 3.1. Summary of tables and their design rationale in AGENT and future interactions database. See Appendix 11 for the ERD.

Table Name	Design Rationale
interactions_lookup_table	A lookup table which standardizes and references the different types of interactions we host in the database. The most canonical example is a protein-protein interaction wherein one protein binds to another, therefore the alias of entity_1_alias and entity_2_alias would be 'protein'. Having such a lookup table will allow future expandability for future novel interaction types (protein-lncRNAs interactions in Arabidopsis for example).
interactions	A table summarizing how one entity is related to another entity (i.e. interact), doesn't usually refer interactions of a metagene but usually does. Surrogate key chosen over a 3-column composite key for simplicity. Note there is a unique index on the entity_1, entity_2 and interaction_type_id column to restrict redundant interactions from different sources.
interolog_confidence_subset_table	To normalize our previous database, I decided to create a subset table for the interactions (~70k) that have an interolog score. This vastly improves redundancy as we had >3M rows that had 9 columns of NULLs (i.e. we had a lot of NULL redundancy). This table and its columns represent the interolog score for a particular interaction for several species.
algorithms_lookup_table	This reference table stores the necessary explanation of algorithmic rankings/scores as we had multiple predicted PPIs in the prior AIV2 database.
interactions_algo_score_join_table	Since every interaction could potentially have more than one algorithm ran on it, this is a join table where the algorithm name is the foreign key.

modes_of_action_lookup_table	A table which will store how a gene modulates the activity of another (i.e. does it repress or activate the target?).
external_source	Table which will store all the unique sources (usually a paper) with extra columns for curation of GRNs for AGENT. Note that the source_name column is unique and thus will primarily store a pubmed ID.
interactions_source_mi_join_table	A table that offers a join for interactions and sources (along with MI terms). That is, one interaction can be referenced by more than one paper. This table allows efficient querying of (1) How many interactions exist in a paper and (2) Fetching the individual interactions of a paper for visualization.
tag_lookup_table	Lookup table for each tag so we can categorize them for our front-end app. I.e. group tags like 'Y1H, Y2H, CHIP' under 'experiment' for categorization for the user. Note that since the tag_name is the PK and MySQL is case-insensitive by default, we won't get 'chIP' and 'CHIP' duplicated.
source_tag_join_table	As one paper may have many tags, create a join table where the FKs are the source id and tag_name.

Summary

I have developed two web-based Arabidopsis interaction viewers, AIV2 and AGENT to display PPI networks and gene regulatory networks (GRNs) respectively. In continuing the legacy of the AIV, AIV2 maintains all the previous features of the older application easily transition users. AIV2 used Cytoscape.js (Franz *et al.*, 2016) instead of other graph libraries such as D3.js as Cytoscape.js hosts functionalities that were seen in AIV1 in addition to many plug-ins and developer-defined customizations. For example, I overlaid SUBA4 (Hooper *et al.*, 2017) localization data as a donut pie-chart scalable vector graphics (SVG) file on top of the node bodies which was enabled by a built-in Cytoscape.js feature. Other improvements include a layered ‘cake’ localization layout, linkouts to other tools such as O’Malley *et al.*'s (2016) DAP-Seq genome browser, vastly improved loading times, and dynamic filtering of nodes/edges. These new features apply data visualization principles such as the Shneiderman’s mantra (Shneiderman, 1996) of overview-first by implementing tooltips and rapid serial visual presentation (RSVP) (Spence, 2002) when users filter nodes/edges dynamically. Dynamic filtering of nodes and edges also allows users to simplify a network depending on their goals which follows Tufte's (2001) principle of minimizing ‘chartjunk’. I showed AIV2 can predict generate TF complex members by combining *a priori* knowledge of NF-Y TFs (Zhao *et al.*, 2017) and bZIP28’s binding to the ER-responsive elements (Srivastava, Deng and Howell, 2014) with AIV2’s results. Additionally, I demonstrated AIV can visually highlight Dong *et al.*'s (2019) structurally predicted PPIs when given an unannotated gene. Currently, AIV2 is under beta. However, as the tool is feature-rich, user-testing should be focused the new features instead of legacy features. For example, I can perform A/B testing if users prefer the current search box which highlights genes versus another user interface (UI) component such as a dropdown. With user-testing, AIV2 will remain as the premier web-based Arabidopsis PPI tool as it remains the only tool to visually display PPIs versus traditional text-based databases.

AGENT also uses Cytoscape.js along with React to visualize its GRNs. I designed AGENT to be lightweight and simpler than AIV so that it can be integrated into other webpages such as ePlant (Waese *et al.*, 2017). The viewer's features are largely inspired by systems biology. For example, Rachel Woo and I implemented a motif searching slider which easily highlights motifs such as feed-forward loops (FFLs) (Alon, 2007) which can allow users to hypothesize the regulatory dynamics between the FFL members. Additional systems biology approaches include resizing nodes according to their degree centrality as genes with higher degree centrality can be used to identify protein hubs (Gaudinier and Brady, 2016). For example, a large number of TFs binding to a gene suggest functional redundancy which implies that gene is critical in that particular GRN. AGENT is still under development but is currently released in the alpha stage. I foresee that some of the mentioned features may be novel to researchers unfamiliar with graph theory. Therefore, instead of adding new features, I will continue to refine the initial graph layout of the GRN until I perform user-testing at the International Conference on Arabidopsis Research 2020. I intend to follow the user-testing guidelines as listed in Waese *et al.* (2017) of free exploration, task completion, and questionnaire. Currently, AGENT has a feedback form for those who wish to request a feature (http://bar.utoronto.ca/AGENT/feedback). Last, AGENT is also a repository of GRNs which can easily be referenced and linked. Hopefully this ability will entice Arabidopsis researchers to deposit their GRNs into AGENT to share with others much like eFP browser (Winter *et al.*, 2007) displays many colloborators' expression data. As GRNs are continually being generated via newer high-throughput methods such as DAP-Seq (O'Malley *et al.*, 2016) and *in silico* methods, AGENT will be an essential tool for cataloguing and viewing GRNs.

References

Adrian, J., Torti, S. and Turck, F. (2009) 'From decision to commitment: the molecular memory of flowering', *Molecular plant*. Elsevier, 2(4), pp. 628–642.

Aghamirzaie, D. *et al.* (2017) 'Expresso: A database and web server for exploring the interaction of transcription factors and their target genes in Arabidopsis thaliana using ChIP-Seq peak data', *F1000Research*, 6(0), p. 372. doi: 10.12688/f1000research.10041.1.

Ahnert, S. E. (2013) 'Power graph compression reveals dominant relationships in genetic transcription networks', *Molecular BioSystems*, 9(11), pp. 2681–2685. doi: 10.1039/c3mb70236g.

Albrecht, M. *et al.* (2009) 'On open problems in biological network visualization', in *International Symposium on Graph Drawing*, pp. 256–267.

Alon, U. (2007) 'Network motifs: Theory and experimental approaches', *Nature Reviews Genetics*, 8(6), pp. 450–461. doi: 10.1038/nrg2102.

Andersson, R. (2015) 'Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model', *BioEssays*, 37(3), pp. 314–323. doi: 10.1002/bies.201400162.

Anscombe, F. J. (1973) 'Graphs in statistical analysis', *The american statistician*. Taylor & Francis Group, 27(1), pp. 17–21.

Appleseed, J. (2019) *13 Design Patterns for Autocomplete Suggestions (27% Get it Wrong) - Articles - Baymard Institute*. Available at: https://baymard.com/blog/autocomplete-design (Accessed: 29 December 2020).

Arabidopsis Interactome Mapping Consortium (2011) 'Evidence for Network Evolution in an Arabidopsis Interactome Map', *Science*, 333(6042), pp. 601–607.

Arda, H. E., Benitez, C. M. and Kim, S. K. (2013) 'Gene regulatory networks governing pancreas development', *Developmental Cell*. Elsevier Inc., 25(1), pp. 5–13. doi: 10.1016/j.devcel.2013.03.016.

Baginski, D. (2019) *UX Checklist Series: Form Design | Seer Interactive*. Available at: https://www.seerinteractive.com/blog/ux-checklist-series-form-design/ (Accessed: 13 December 2019).

Bhardwaj, N. and Lu, H. (2005) ‘Correlation between gene expression profiles and protein-protein interactions within and across genomes’, *Bioinformatics*, 21(11), pp. 2730–2738. doi: 10.1093/bioinformatics/bti398.

Bhattacharjee, S. *et al.* (2011) ‘Pathogen effectors target Arabidopsis EDS1 and alter its interactions with immune regulators’, *Science*, 334(6061), pp. 1405–1408.

Blázquez, M. A. *et al.* (1997) ‘LEAFY expression and flower initiation in Arabidopsis’, *Development*, 124(19), pp. 3835–3844.

De Bodt, S. *et al.* (2009) ‘Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression.’, *BMC genomics*, 10, p. 288. doi: 10.1186/1471-2164-10-288.

Boruc, J. *et al.* (2010) ‘Functional modules in the Arabidopsis core cell cycle binary protein-protein interaction network’, *Plant Cell*, 22(4), pp. 1264–1280. doi: 10.1105/tpc.109.073635.

Bradshow, K. (2019) *Chrome will warn users ahead of Flash Player’s deprecation - 9to5Google*. Available at: https://9to5google.com/2019/03/22/chrome-warn-flash-player-deprecation-july/ (Accessed: 3 December 2019).

Brady, S. M. *et al.* (2011) ‘A stele-enriched gene regulatory network in the Arabidopsis root’, *Molecular Systems Biology*. Nature Publishing Group, 7(459), pp. 1–9. doi: 10.1038/msb.2010.114.

Brandão, M. M., Dantas, L. L. and Silva-Filho, M. C. (2009) ‘AtPIN: Arabidopsis thaliana protein interaction network’, *BMC Bioinformatics*, 10, pp. 1–7. doi: 10.1186/1471-2105-10-454.

Brown, C. M. (1999) *Human-computer interface design guidelines*. Intellect Books.

Brudnicki, D. (2018) *PubSub for communicating between React Components - Coding Stuff - Medium*. Available at: https://medium.com/coding-stuff/pubsub-for-communicating-between-react-components-999159d59a77 (Accessed: 29 December 2020).

Bülow, L., Brill, Y. and Hehl, R. (2010) 'AthaMap-assisted transcription factor target gene identification in Arabidopsis thaliana.', *Database : the journal of biological databases and curation*, 2010, pp. 2–5. doi: 10.1093/database/baq034.

Carroll, T. S. *et al.* (2014) 'Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data', *Frontiers in Genetics*, 5(APR), pp. 1–11. doi: 10.3389/fgene.2014.00075.

Chen, C. *et al.* (2019) 'Inferring Gene Regulatory Networks from a Population of Yeast Segregants', *Scientific Reports*, 9(1), pp. 1–9. doi: 10.1038/s41598-018-37667-4.

Chen, H. *et al.* (2018) 'Same Stats, Different Graphs', in *International Symposium on Graph Drawing and Network Visualization*, pp. 463–477.

Chen, Q. *et al.* (2015) 'A coherent transcriptional feed-forward motif model for mediating auxin-sensitive PIN3 expression during lateral root development', *Nature Communications*, 6. doi: 10.1038/ncomms9821.

Chen, X. *et al.* (2015) 'CressInt: A user-friendly web resource for genome-scale exploration of gene regulation in Arabidopsis thaliana', *Current Plant Biology*. Elsevier B.V., 3–4, pp. 48–55. doi: 10.1016/j.cpb.2015.09.001.

Chen, X. W. and Liu, M. (2005) 'Prediction of protein-protein interactions using random decision forest framework', *Bioinformatics*, 21(24), pp. 4394–4400. doi: 10.1093/bioinformatics/bti721.

Chudasama, D. *et al.* (2018) 'Identification of cancer biomarkers of prognostic value using specific gene regulatory networks (GRN): A novel role of RAD51AP1 for ovarian and lung cancers', *Carcinogenesis*, 39(3), pp. 407–417. doi: 10.1093/carcin/bgx122.

Cortijo, S. *et al.* (2018) 'Chromatin Immunoprecipitation Sequencing (ChIP-Seq) for Transcription Factors and Chromatin Factors in Arabidopsis thaliana Roots: From Material Collection to Data Analysis', in Ristova, D. and Barbez, E. (eds) *Root Development: Methods and Protocols*. New York, NY: Springer New York, pp. 231–248. doi: 10.1007/978-1-4939-7747-5_18.

Deng, M. *et al.* (2002) ‘Inferring Domain – Domain Interactions From Protein – Protein Interactions’, pp. 1540–1548. doi: 10.1101/gr.153002.2.

Ding, Z. and Kihara, D. (2019) ‘Computational identification of protein-protein interactions in model plant proteomes’, *Scientific Reports*. Springer US, 9(1), pp. 1–13. doi: 10.1038/s41598-019-45072-8.

Dogrusoz, U. *et al.* (2018) ‘Efficient methods and readily customizable libraries for managing complexity of large networks’, *PLoS ONE*, 13(5), pp. 1–18. doi: 10.1371/journal.pone.0197238.

Dong, S. *et al.* (2019) ‘Proteome-Wide, Structure-Based Prediction of Protein-protein Interactions / New Molecular Interactions Viewer’, *Plant Physiology*, 179(April), p. pp.01216.2018. doi: 10.1104/pp.18.01216.

Dong, S. and Provart, N. J. (2018) ‘Analyses of Protein Interaction Networks Using Computational Tools’, in *Two-Hybrid Systems*. Springer, pp. 97–117.

Dortay, H. *et al.* (2006) ‘Analysis of protein interactions within the cytokinin-signaling pathway of Arabidopsis thaliana’, *FEBS Journal*, 273(20), pp. 4631–4644. doi: 10.1111/j.1742-4658.2006.05467.x.

Emmert-Streib, F., Dehmer, M. and Haibe-Kains, B. (2014) ‘Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks’, *Frontiers in Cell and Developmental Biology*, 2(August), pp. 1–7. doi: 10.3389/fcell.2014.00038.

Fields, S. (2005) ‘High-throughput two-hybrid analysis: The promise and the peril’, *FEBS Journal*, 272(21), pp. 5391–5399. doi: 10.1111/j.1742-4658.2005.04973.x.

Fields, S. and Song, O. K. (1989) ‘A novel genetic system to detect protein-protein interactions’, *Nature*, 340(6230), pp. 245–246. doi: 10.1038/340245a0.

Fornari, M. *et al.* (2013) ‘The Arabidopsis NF-YA3 and NF-YA8 Genes Are Functionally Redundant and Are Required in Early Embryogenesis’, *PLOS ONE*. Public Library of Science, 8(11), pp. 1–13. doi: 10.1371/journal.pone.0082043.

Franz, M. *et al.* (2016) ‘Cytoscape.js: A graph theory library for visualisation and analysis’, *Bioinformatics*, 32(2), pp. 309–311. doi: 10.1093/bioinformatics/btv557.

Frias, S. *et al.* (2015) ‘CerebralWeb: A cytoscape.js plug-in to visualize networks stratified by subcellular localization’, *Database*, 2015, pp. 1–4. doi: 10.1093/database/bav041.

Gaudinier, A. *et al.* (2011) ‘Enhanced Y1H assays for Arabidopsis’, *Nature Methods*, 8(12), pp. 1053–1056. doi: 10.1038/nmeth.1750.

Gaudinier, A. and Brady, S. M. (2016) ‘Mapping Transcriptional Networks in Plants: Data-Driven Discovery of Novel Biological Mechanisms’, *Annual Review of Plant Biology*, 67(1), pp. 575–594. doi: 10.1146/annurev-arplant-043015-112205.

Gavin, A. C. *et al.* (2002) ‘Functional organization of the yeast proteome by systematic analysis of protein complexes’, *Nature*, 415(6868), pp. 141–147. doi: 10.1038/415141a.

Geisler-Lee, J. *et al.* (2007) ‘A Predicted Interactome for Arabidopsis’, *Plant Physiology*, 145(2), pp. 317–329. doi: 10.1104/pp.107.103465.

Gócza, Z. (2015) *Myth #12: More choices and features result in higher satisfaction - UX Myths*. Available at: https://uxmyths.com/post/712569752/myth-more-choices-and-features-result-in-higher-satisfac (Accessed: 31 December 2020).

Grigoriev, A. (2003) ‘On the number of protein-protein interactions in the yeast proteome’, *Nucleic Acids Research*, 31(14), pp. 4157–4161. doi: 10.1093/nar/gkg466.

Haberle, V. and Stark, A. (2018) ‘Eukaryotic core promoters and the functional basis of transcription initiation’, *Nature Reviews Molecular Cell Biology*. Springer US, 19(10), pp. 621–637. doi: 10.1038/s41580-018-0028-8.

Harmer, S. L., Panda, S. and Kay, S. A. (2001) ‘Molecular bases of circadian rhythms’, *Annual review of cell and developmental biology*. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, 17(1), pp. 215–253.

Hatsugai, N. *et al.* (2016) ‘The μ subunit of Arabidopsis adaptor protein-2 is involved in effector-triggered immunity mediated by membrane-localized resistance proteins’, *Molecular Plant-Microbe Interactions*, 29(5), pp. 345–351. doi: 10.1094/MPMI-10-15-0228-R.

Heard, D. J., Kiss, T. and Filipowicz, W. (1993) ‘Both Arabidopsis TATA binding protein (TBP) isoforms are functionally identical in RNA polymerase II and III transcription in plant cells: evidence for gene-specific changes in DNA binding specificity of TBP.’, *The EMBO Journal*, 12(9), pp. 3519–3528. doi: 10.1002/j.1460-2075.1993.tb06026.x.

Hernandez-Garcia, C. M. and Finer, J. J. (2014) ‘Identification and validation of promoters and cis-acting regulatory elements’, *Plant Science*. Elsevier Ireland Ltd, 217–218, pp. 109–119. doi: 10.1016/j.plantsci.2013.12.007.

Hoa, L. (2015) *Radical Redesign or Incremental Change?* Available at: https://www.nngroup.com/articles/radical-incremental-redesign/ (Accessed: 13 December 2019).

Hooper, C. M. *et al.* (2017) ‘SUBA4: The interactive data analysis centre for Arabidopsis subcellular protein locations’, *Nucleic Acids Research*, 45(D1), pp. D1064–D1074. doi: 10.1093/nar/gkw1041.

Hu, S. *et al.* (2011) ‘Functional protein microarray technology’, *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(3), pp. 255–268. doi: 10.1002/wsbm.118.

Huang, H., Jedynak, B. M. and Bader, J. S. (2007) ‘Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps’, *PLoS computational biology*. Public Library of Science, 3(11).

Huang, W. *et al.* (2012) ‘Mapping the core of the Arabidopsis circadian clock defines the network structure of the oscillator’, *Science*. American Association for the Advancement of Science, 336(6077), pp. 75–79.

Inukai, S., Kock, K. H. and Bulyk, M. L. (2017) ‘Transcription factor–DNA binding: beyond binding site motifs’, *Current Opinion in Genetics and Development*. Elsevier Ltd, 43, pp. 110–119. doi: 10.1016/j.gde.2017.02.007.

Iwata, Y. and Koizumi, N. (2012) ‘Plant transducers of the endoplasmic reticulum unfolded protein response’, *Trends in Plant Science*. Elsevier Ltd, 17(12), pp. 720–727. doi: 10.1016/j.tplants.2012.06.014.

Jeong, H. *et al.* (2001) ‘Lethality and centrality in protein networks’, *Nature*, 411(6833), pp. 41–

42. doi: 10.1038/35075138.

Jin, J. *et al.* (2015) ‘An Arabidopsis transcriptional regulatory map reveals distinct functional and evolutionary features of novel transcription factors’, *Molecular biology and evolution*. Oxford University Press, 32(7), pp. 1767–1773.

Jin, J. *et al.* (2017) ‘PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants’, *Nucleic Acids Research*, 45(D1), pp. D1040–D1045. doi: 10.1093/nar/gkw982.

Jones, A. M. *et al.* (2014) ‘Border control - A membrane-linked interactome of Arabidopsis’, *Science*, 344(6185), pp. 711–716. doi: 10.1126/science.1251358.

Kaminski, N. (2000) ‘Bioinformatics. A user’s perspective.’, *American journal of respiratory cell and molecular biology*. United States, 23(6), pp. 705–711. doi: 10.1165/ajrcmb.23.6.4291.

Katchalski-Katzir, E. *et al.* (1992) ‘Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques’, *Proceedings of the National Academy of Sciences of the United States of America*, 89(6), pp. 2195–2199. doi: 10.1073/pnas.89.6.2195.

Kelley, L. A. *et al.* (2015) ‘The Phyre2 web portal for protein modeling, prediction and analysis’, *Nature Protocols*, 10(6), pp. 845–858. doi: 10.1038/nprot.2015.053.

Kerrien, S. *et al.* (2012) ‘The IntAct molecular interaction database in 2012’, *Nucleic Acids Research*, 40(D1), pp. 841–846. doi: 10.1093/nar/gkr1088.

Keurentjes, J. J. B. *et al.* (2007) ‘Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci’, *Proceedings of the National Academy of Sciences of the United States of America*, 104(5), pp. 1708–1713. doi: 10.1073/pnas.0610429104.

Khan, M. *et al.* (2018) ‘In planta proximity dependent biotin identification (BioID)’, *Scientific reports*. Nature Publishing Group, 8(1), pp. 1–8.

Kilian, J. *et al.* (2007) ‘The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses', *The Plant Journal*. Wiley Online Library, 50(2), pp. 347–363.

Kim, D. I. and Roux, K. J. (2016) 'Filling the void: proximity-based labeling of proteins in living cells', *Trends in cell biology*. Elsevier, 26(11), pp. 804–817.

Kulkarni, S. R. *et al.* (2017) 'TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information', *Nucleic Acids Research*, pp. 1–28. doi: 10.1093/nar/gkx1279.

Lamesch, P. *et al.* (2012) 'The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools', *Nucleic Acids Research*, 40(D1), pp. 1202–1210. doi: 10.1093/nar/gkr1090.

Van Leene, J. *et al.* (2007) 'A tandem affinity purification-based technology platform to study the cell cycle interactome in Arabidopsis thaliana', *Molecular and Cellular Proteomics*, 6(7), pp. 1226–1238. doi: 10.1074/mcp.M700078-MCP200.

Van Leene, J. *et al.* (2015) 'An improved toolbox to unravel the plant cellular machinery by tandem affinity purification of Arabidopsis protein complexes', *Nature Protocols*, 10(1), pp. 169–187. doi: 10.1038/nprot.2014.199.

Levine, M. (2010) 'Transcriptional enhancers in animal development and evolution', *Current Biology*. Elsevier, 20(17), pp. R754--R763.

Li, B. *et al.* (2014) 'Promoter-based integration in plant defense regulation', *Plant Physiology*, 166(4), pp. 1803–1820. doi: 10.1104/pp.114.248716.

Li, J. J. and Herskowitz, I. (1993) 'Isolation of ORC6, a component of the yeast origin recognition complex by a one-hybrid system', *Science*. American Association for the Advancement of Science, 262(5141), pp. 1870–1874.

Li, S. B. *et al.* (2016) 'A review of auxin response factors (ARFs) in plants', *Frontiers in Plant Science*, 7(FEB2016), pp. 1–7. doi: 10.3389/fpls.2016.00047.

Lin, M., Shen, X. and Chen, X. (2011) 'PAIR: The predicted Arabidopsis interactome resource', *Nucleic Acids Research*, 39(SUPPL. 1), pp. 1134–1140. doi: 10.1093/nar/gkq938.

Liseron-Monfils, C. and Ware, D. (2015) ‘Revealing gene regulation and associations through biological networks’, *Current Plant Biology*. Elsevier B.V., 3–4, pp. 30–39. doi: 10.1016/j.cpb.2015.11.001.

Liu, J. X. and Howell, S. H. (2010) ‘bZIP28 and NF-Y transcription factors are activated by ER stress and assemble into a transcriptional complex to regulate stress response genes in Arabidopsis’, *Plant Cell*, 22(3), pp. 782–796. doi: 10.1105/tpc.109.072173.

Lopes, C. T. *et al.* (2011) ‘Cytoscape Web: An interactive web-based network browser’, *Bioinformatics*, 27(13), pp. 2347–2348. doi: 10.1093/bioinformatics/btq430.

De Lucas, M. *et al.* (2016) ‘Transcriptional regulation of arabidopsis polycomb repressive complex 2 coordinates cell-type proliferation and differentiation’, *Plant Cell*, 28(10), pp. 2616–2631. doi: 10.1105/tpc.15.00744.

Lumba, S. *et al.* (2014) ‘A mesoscale abscisic acid hormone interactome reveals a dynamic signaling landscape in arabidopsis’, *Developmental Cell*, 29(3), pp. 360–372. doi: 10.1016/j.devcel.2014.04.004.

Lv, Q. *et al.* (2017) ‘AtPID: A genome-scale resource for genotype-phenotype associations in Arabidopsis’, *Nucleic Acids Research*, 45(D1), pp. D1060–D1063. doi: 10.1093/nar/gkw1029.

Macindoe, G. *et al.* (2010) ‘HexServer: An FFT-based protein docking server powered by graphics processors’, *Nucleic Acids Research*, 38(SUPPL. 2), pp. 445–449. doi: 10.1093/nar/gkq311.

Mangan, S., Zaslaver, A. and Alon, U. (2003) ‘The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks’, *Journal of Molecular Biology*, 334(2), pp. 197–204. doi: 10.1016/j.jmb.2003.09.049.

McWhite, C. D. *et al.* (2019) ‘A pan-plant protein complex map reveals deep conservation and novel assemblies’, *bioRxiv*. Cold Spring Harbor Laboratory. doi: 10.1101/815837.

Von Mering, C. *et al.* (2002) ‘Comparative assessment of large-scale data sets of protein--protein interactions’, *Nature*. Nature Publishing Group, 417(6887), pp. 399–403.

Milo, R. *et al.* (2002) ‘Network Motifs: Simple Building Blocks of Complex Networks’, *Science*, 298(October), pp. 824–827.

Mironova, V. V. *et al.* (2014) ‘Computational analysis of auxin responsive elements in the Arabidopsis thaliana L. genome’, *BMC Genomics*, 15(Suppl 12), pp. 1–14. doi: 10.1186/1471-2164-15-S12-S4.

Molina, C. and Grotewold, E. (2005) ‘Genome wide analysis of Arabidopsis core promoters’, *BMC Genomics*, 6, pp. 1–12. doi: 10.1186/1471-2164-6-25.

Moreland, K. (2009) ‘Diverging color maps for scientific visualization’, in *International Symposium on Visual Computing*, pp. 92–103.

Murphy, E. *et al.* (2016) ‘RALFL34 regulates formative cell divisions in Arabidopsis pericycle during lateral root initiation’, *Journal of Experimental Botany*, 67(16), pp. 4863–4875. doi: 10.1093/jxb/erw281.

Nesbitt, K. V and Friedrich, C. (2002) ‘Applying gestalt principles to animated visualizations of network data’, in *Proceedings Sixth International Conference on Information Visualisation*, pp. 737–743.

Nooren, I. M. A. and Thornton, J. M. (2003) ‘Diversity of protein–protein interactions’, *The EMBO Journal*, 22(14), pp. 3486–3492. doi: 10.1093/emboj/cdg359.

Norman, D. (2013) *The design of everyday things: Revised and expanded edition*. Basic books.

O’Malley, R. C. *et al.* (2016) ‘Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape’, *Cell*. Elsevier Inc., 165(5), pp. 1280–1292. doi: 10.1016/j.cell.2016.04.038.

Oughtred, R. *et al.* (2019) ‘The BioGRID interaction database: 2019 update’, *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D529–D541. doi: 10.1093/nar/gky1079.

Park, P. J. (2009) ‘ChIP-seq: Advantages and challenges of a maturing technology’, *Nature Reviews Genetics*. Nature Publishing Group, 10(10), pp. 669–680. doi: 10.1038/nrg2641.

Pazos, F. and Valencia, A. (2001) ‘Similarity of phylogenetic trees as indicator of protein-protein interaction’, *Protein Engineering*, 14(9), pp. 609–614. doi: 10.1093/protein/14.9.609.

Pérez-Rodríguez, P. *et al.* (2009) ‘PlnTFDB: Updated content and new features of the plant transcription factor database’, *Nucleic Acids Research*, 38(SUPPL.1), pp. 822–827. doi: 10.1093/nar/gkp805.

Popescu, S. C. *et al.* (2007) ‘Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays’, *Proceedings of the National Academy of Sciences of the United States of America*, 104(11), pp. 4730–4735. doi: 10.1073/pnas.0611615104.

Porco, S. *et al.* (2016) ‘Lateral root emergence in Arabidopsis is dependent on transcription factor LBD29 regulation of auxin influx carrier LAX3’, *Development (Cambridge)*, 143(18), pp. 3340–3349. doi: 10.1242/dev.136283.

Pratt, D. *et al.* (2015) ‘NDEx, the Network Data Exchange’, *Cell Systems*. Elsevier Inc., 1(4), pp. 302–305. doi: 10.1016/j.cels.2015.10.001.

Puig, O. *et al.* (2001) ‘The tandem affinity purification (TAP) method: A general procedure of protein complex purification’, *Methods*, 24(3), pp. 218–229. doi: 10.1006/meth.2001.1183.

Rao, V. S. *et al.* (2014) ‘Protein-protein interaction detection: methods and analysis’, *International journal of proteomics*. Hindawi, 2014.

Reece-Hoyes, J. S. *et al.* (2011) ‘Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping’, *Nature Methods*, 8(12), pp. 1059–1068. doi: 10.1038/nmeth.1748.

Reece-Hoyes, J. S. and Marian Walhout, A. J. (2012) ‘Yeast one-hybrid assays: A historical and technical perspective’, *Methods*. Elsevier Inc., 57(4), pp. 441–447. doi: 10.1016/j.ymeth.2012.07.027.

Riethoven, J.-J. M. (2010) ‘Regulatory regions in DNA: promoters, enhancers, silencers, and insulators’, in *Computational biology of transcription factor binding*. Springer, pp. 33–42.

Ritchie, D. W. and Kemp, G. J. L. (2000) ‘Protein docking using spherical polar Fourier correlations’, *Proteins: Structure, Function and Genetics*, 39(2), pp. 178–194. doi: 10.1002/(SICI)1097-0134(20000501)39:2<178::AID-PROT8>3.0.CO;2-6.

Ritchie, D. W. and Venkatraman, V. (2010) ‘Ultra-fast FFT protein docking on graphics processors’, *Bioinformatics*, 26(19), pp. 2398–2405. doi: 10.1093/bioinformatics/btq444.

Rubio, V. *et al.* (2005) ‘An alternative tandem affinity purification strategy applied to Arabidopsis protein complex isolation’, *Plant Journal*, 41(5), pp. 767–778. doi: 10.1111/j.1365-313X.2004.02328.x.

Saddic, L. A. *et al.* (2006) ‘The LEAFY target LMI1 is a meristem identity regulator and acts together with LEAFY to regulate expression of CAULIFLOWER’, *Development*, 133(9), pp. 1673–1682. doi: 10.1242/dev.02331.

Sakuraba, Y. *et al.* (2015) ‘The arabidopsis transcription factor NAC016 promotes drought stress responses by repressing AREB1 transcription through a trifurcate feed-forward regulatory loop involving NAP’, *Plant Cell*, 27(6), pp. 1771–1787. doi: 10.1105/tpc.15.00222.

Schwartz, B. (2004) *The paradox of choice: Why more is less*.

Schwikowski, B., Uetz, P. and Fields, S. (2000) ‘A network of protein-protein interactions in yeast’, *Nature Biotechnology*, 18(12), pp. 1257–1261. doi: 10.1038/82360.

Seebacher, J. and Gavin, A. C. (2011) ‘SnapShot: Protein-protein interaction networks’, *Cell*. Elsevier, 144(6), pp. 1000-1000.e1. doi: 10.1016/j.cell.2011.02.025.

Sen, S. *et al.* (2017) ‘Development of an informatics analytics workflow for DAP-seq data exploration and validation for auxin response factors in maize’, *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*, 2017-Janua, pp. 2300–2301. doi: 10.1109/BIBM.2017.8218034.

Shimotohno, A. and Scheres, B. (2019) ‘Topology of regulatory networks that guide plant meristem activity: similarities and differences’, *Current Opinion in Plant Biology*. Elsevier Ltd, 51, pp. 74–80. doi: 10.1016/j.pbi.2019.04.006.

Shneiderman, B. (1996) ‘The eyes have it: A task by data type taxonomy for information visualizations’, in *Proceedings 1996 IEEE symposium on visual languages*, pp. 336–343.

Shoemaker, B. A. and Panchenko, A. R. (2007) ‘Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners’, *PLoS Computational Biology*, 3(4), pp. 595–601. doi: 10.1371/journal.pcbi.0030043.

Smoly, I. Y. *et al.* (2017) ‘MotifNet: A web-server for network motif analysis’, *Bioinformatics*, 33(12), pp. 1907–1909. doi: 10.1093/bioinformatics/btx056.

Song, J. and Singh, M. (2013) ‘From Hub Proteins to Hub Modules : The Relationship Between Essentiality and Centrality in the Yeast Interactome at Different Scales of Organization’, *PLoS Computational Biology*, 9(2). doi: 10.1371/journal.pcbi.1002910.

Sparks, E. E. *et al.* (2016) ‘Establishment of Expression in the SHORTROOT-SCARECROW Transcriptional Cascade through Opposing Activities of Both Activators and Repressors’, *Developmental Cell*, 39(5), pp. 585–596. doi: 10.1016/j.devcel.2016.09.031.

Spence, R. (2002) ‘Rapid, Serial and Visual: a presentation technique with potential’, *Information Visualization*, 1(1), pp. 13–19. doi: 10.1057/palgrave/ivs/9500008.

Srivastava, R., Deng, Y. and Howell, S. H. (2014) ‘Stress sensing in plants by an ER stress sensor/transducer, bZIP28’, *Frontiers in Plant Science*, 5(FEB), pp. 1–6. doi: 10.3389/fpls.2014.00059.

Su, G. *et al.* (2014) ‘Biological network exploration with Cytoscape 3’, *Current protocols in bioinformatics*. Wiley Online Library, 47(1), pp. 8–13.

Szklarczyk, D. *et al.* (2019) ‘STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets’, *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D607–D613. doi: 10.1093/nar/gky1131.

Taylor-Teeples, M. *et al.* (2015) ‘An Arabidopsis gene regulatory network for secondary cell wall synthesis’, *Nature*. Nature Publishing Group, 517(7536), pp. 571–575. doi: 10.1038/nature14099.

Theodosiou, T. *et al.* (2017) ‘NAP: The Network Analysis Profiler, a web tool for easier topological analysis and comparison of medium-scale biological networks’, *BMC Research Notes*. BioMed Central, 10(1), pp. 1–9. doi: 10.1186/s13104-017-2607-8.

Traven, A., Jelicic, B. and Sopta, M. (2006) ‘Yeast Gal4: A transcriptional paradigm revisited’, *EMBO Reports*, 7(5), pp. 496–499. doi: 10.1038/sj.embor.7400679.

Tucker, C. L., Gera, J. F. and Uetz, P. (2001) ‘Towards an understanding of complex protein networks’, *Trends in Cell Biology*, 11(3), pp. 102–106. doi: 10.1016/S0962-8924(00)01902-4.

Tufte, E. R. (2001) *The visual display of quantitative information*. Graphics press Cheshire, CT.

UC Davis Proteomics Core (2019) *Yeast One Hybrid Services (Brady Lab) | UC Davis Proteomics Core*. Available at: https://proteomics.ucdavis.edu/services-and-prices/yeast-one-hybrid-services-brady-lab/ (Accessed: 8 December 2019).

Ulmasov, T., Hagen, G. and Guilfoyle, T. J. (1999) ‘Dimerization and DNA binding of auxin response factors’, *Plant Journal*, 19(3), pp. 309–319. doi: 10.1046/j.1365-313X.1999.00538.x.

Usadel, B. *et al.* (2009) ‘A guide to using MapMan to visualize and compare Omics data in plants: A case study in the crop species, Maize’, *Plant, Cell and Environment*, 32(9), pp. 1211–1229. doi: 10.1111/j.1365-3040.2009.01978.x.

Vallabhajosyula, R. R. *et al.* (2009) ‘Identifying Hubs in protein interaction networks’, *PLoS ONE*, 4(4), pp. 1–10. doi: 10.1371/journal.pone.0005344.

Vandereyken, K. *et al.* (2018) ‘Hub Protein Controversy: Taking a Closer Look at Plant Stress Response Hubs’, *Frontiers in Plant Science*, 9(June), pp. 1–24. doi: 10.3389/fpls.2018.00694.

Vidal, E. A. *et al.* (2010) ‘Nitrate-responsive miR393/AFB3 regulatory module controls root system architecture in Arabidopsis thaliana’, *Proceedings of the National Academy of Sciences of the United States of America*, 107(9), pp. 4477–4482. doi: 10.1073/pnas.0909571107.

Vidal, M. and Fields, S. (2014) ‘The yeast two-hybrid assay: Still finding connections after 25 years’, *Nature Methods*. Nature Publishing Group, 11(12), pp. 1203–1206. doi: 10.1038/nmeth.3182.

Waese, J. *et al.* (2017) ‘ePlant: Visualizing and Exploring Multiple Levels of Data for Hypothesis Generation in Plant Biology’, *The Plant Cell*, 29(8), pp. 1806–1821. doi: 10.1105/tpc.17.00073.

Wang, R. and Estelle, M. (2014) ‘Diversity and specificity: Auxin perception and signaling through the TIR1/AFB pathway’, *Current Opinion in Plant Biology*. Elsevier Ltd, 21, pp. 51–58. doi: 10.1016/j.pbi.2014.06.006.

Weber, B. *et al.* (2016) ‘Plant Enhancers: A Call for Discovery’, *Trends in Plant Science*. Elsevier Ltd, 21(11), pp. 974–987. doi: 10.1016/j.tplants.2016.07.013.

Wehner, N., Weiste, C. and Dröge-Laser, W. (2011) ‘Molecular screening tools to study arabidopsis transcription factors’, *Frontiers in Plant Science*, 2(NOV), pp. 1–7. doi: 10.3389/fpls.2011.00068.

Windram, O. and Denby, K. J. (2015) ‘Modelling signaling networks underlying plant defence’, *Current Opinion in Plant Biology*. Elsevier Ltd, 27(Table 1), pp. 165–171. doi: 10.1016/j.pbi.2015.07.007.

Winter, D. *et al.* (2007) ‘An “electronic fluorescent pictograph” Browser for exploring and analyzing large-scale biological data sets’, *PLoS ONE*, 2(8), pp. 1–12. doi: 10.1371/journal.pone.0000718.

Xu, W., Dubos, C. and Lepiniec, L. (2015) ‘Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes’, *Trends in Plant Science*, 20(3), pp. 176–185. doi: 10.1016/j.tplants.2014.12.001.

Yamaoka, S. *et al.* (2013) ‘Identification and dynamics of arabidopsis adaptor protein-2 complex and its involvement in floral organ development’, *Plant Cell*, 25(8), pp. 2958–2969. doi: 10.1105/tpc.113.114082.

Yilmaz, A. *et al.* (2011) ‘AGRIS: The arabidopsis gene regulatory information server, an update’, *Nucleic Acids Research*, 39(SUPPL. 1), pp. 1118–1122. doi: 10.1093/nar/gkq1120.

Yu, C. P., Lin, J. J. and Li, W. H. (2016) ‘Positional distribution of transcription factor binding sites in Arabidopsis thaliana’, *Scientific Reports*. Nature Publishing Group, 6(April), pp. 1–7.

doi: 10.1038/srep25164.

Yu, H. *et al.* (2007) ‘The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics’, *PLoS Computational Biology*, 3(4), pp. 713–720. doi: 10.1371/journal.pcbi.0030059.

Yu, Q. B. *et al.* (2008) ‘Construction of a chloroplast protein interaction network and functional mining of photosynthetic proteins in Arabidopsis thaliana’, *Cell Research*, 18(10), pp. 1007–1019. doi: 10.1038/cr.2008.286.

Zhao, H. *et al.* (2017) ‘The Arabidopsis thaliana nuclear factor Y transcription factors’, *Frontiers in Plant Science*, 7(January), pp. 1–11. doi: 10.3389/fpls.2016.02045.

Zhiponova, M. K. *et al.* (2014) ‘Helix--loop--helix/basic helix--loop--helix transcription factor network represses cell elongation in Arabidopsis through an apparent incoherent feed-forward loop’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 111(7), pp. 2824–2829.

Zicola, J. *et al.* (2019) ‘Targeted DNA methylation represses two enhancers of FLOWERING LOCUS T in Arabidopsis thaliana’, *Nature Plants*. Springer US, 5(3), pp. 300–307. doi: 10.1038/s41477-019-0375-2.

Appendices

```
{
    "AT2G44940": [
        {
            "source": "At2g44940",
            "target": "At1g01020",
            "index": "2",
            "interolog_confidence": 0,
            "correlation_coefficient": "None",
            "published": true,
            "reference": "doi:10.1016/j.cell.2016.04.038",
            "mi": "2288"
        },
        {
            "source": "At2g44940",
            "target": "At4g12240",
            "index": "2",
            "interolog_confidence": 0,
            "correlation_coefficient": "None",
            "published": true,
            "reference": "doi:10.1016/j.cell.2016.04.038",
            "mi": "2288"
        }...
}
```

Appendix 1. Sample JSON when a POST request is submitted to https://bar.utoronto.ca/interactions2/cgi-bin/get_interactions_dapseq.php with the following parameters: {"loci":"AT2G44940","recursive":false,"published":true,"querydna":true}. The JSON is shortened for brevity.

```
{
    "At2g34970": {
        "mean": 72.39500000000001,
        "sd": 0.49499999999999744
    }
}
```

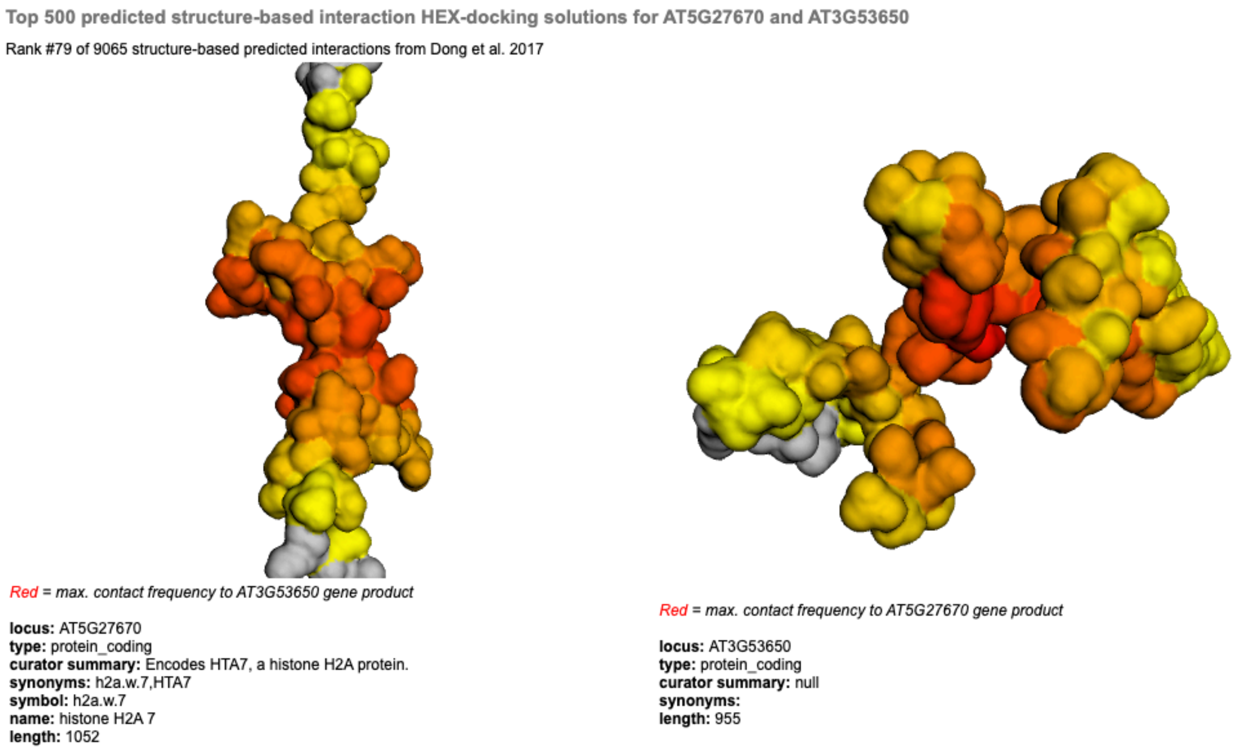
Appendix 2. Sample JSON when a POST request is submitted to https://bar.utoronto.ca/interactions2/cgi-bin/getSample.php with the following parameters: {"geneIDs":["At2g34970"],"species":"arabidopsis","inputMode":"absolute","dataSource":"Chemical","tissue":"Control"," tissuesCompare":""}.

```
[
    {
        "request": {
            "agi": "At2g34970"
        },
        "result": [
            {
                "code": "29.2.3",
                "name": "protein.synthesis.initiation",
                "description": "no description",
                "parent": {
                    "code": "29.2",
                    "name": "protein.synthesis",
                    "description": "no description",
                    "parent": {
                        "code": "29",
                        "name": "protein",
                        "description": "no description",
                        "parent": null
                    }
                }
            }
        ]
    }
]
```

Appendix 3. Sample JSON when a GET request is submitted to https://bar.utoronto.ca/interactions2/cgi-bin/bar_mapman.php?request=["At2g34970"].

```
{
    "At2g34970": {
        "includes_predicted": "yes",
        "includes_experimental": "yes",
        "data": [
            {
                "cytosol": 32
            },
            {
                "nucleus": 14
            },
            {
                "golgi": 10
            },
            {
                "mitochondrion": 6
            },
            {
                "plastid": 2
            }
        ]
    }
}
```

Appendix 4. Sample JSON when a POST request is submitted to https://bar.utoronto.ca/~vlau/suba4.php with the following parameters: {"AGI_IDs":["At2g34970"],"include_predicted":true}.



Top 500 predicted structure-based interaction HEX-docking solutions for AT5G27670 and AT3G53650

Rank #79 of 9065 structure-based predicted interactions from Dong et al. 2017

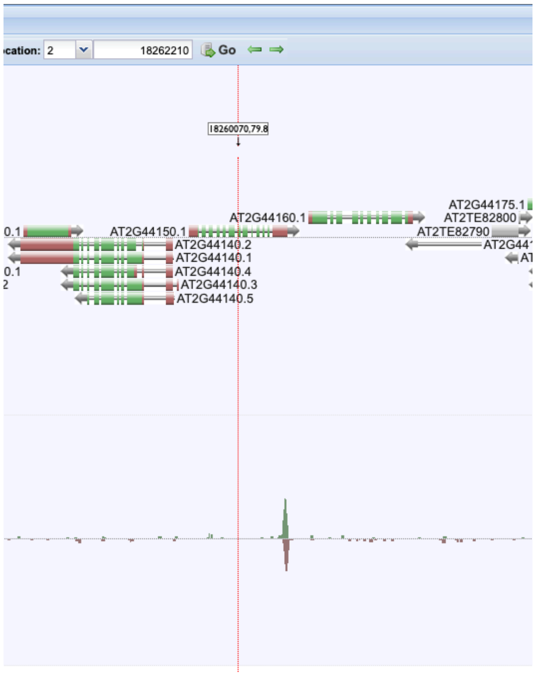
Red = max. contact frequency to AT3G53650 gene product

locus: AT5G27670
type: protein_coding
curator summary: Encodes HTA7, a histone H2A protein.
synonyms: h2a.w.7,HTA7
symbol: h2a.w.7
name: histone H2A 7
length: 1052

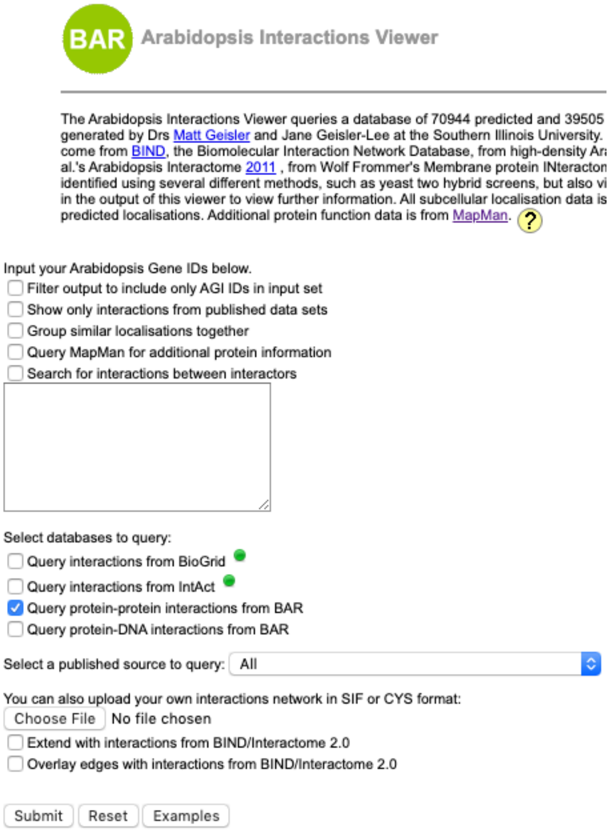
Red = max. contact frequency to AT5G27670 gene product

locus: AT3G53650
type: protein_coding
curator summary: null
synonyms:
length: 955

Appendix 5. Heat map of structurally predicted interaction between AT3G53650 and AT5G27670 where red indicates higher contact frequency as predicted by HEX. Accessible from http://bar.utoronto.ca/protein_docker/?id1=AT5G27670&id2=AT3G53650.



Appendix 6. Cropped output of the DAP-Seq Arabidopsis genome browser (http://neomorph.salk.edu/aj2/pages/hchen/dap_ath_pub.php) that was redirected from our custom API (http://bar.utoronto.ca/DAP-Seq-API?target=At2g44160&tf=At1g44830) when given the target gene is At2g44160 and the TF is At1g44830.



The Arabidopsis Interactions Viewer queries a database of 70944 predicted and 39505
generated by Drs Matt Geisler and Jane Geisler-Lee at the Southern Illinois University.
come from BIND, the Biomolecular Interaction Network Database, from high-density Ar
al.'s Arabidopsis Interactome 2011 , from Wolf Frommer's Membrane protein INteracton
identified using several different methods, such as yeast two hybrid screens, but also vi
in the output of this viewer to view further information. All subcellular localisation data is
predicted localisations. Additional protein function data is from MapMan. ?

Input your Arabidopsis Gene IDs below.

- □ Filter output to include only AGI IDs in input set
- □ Show only interactions from published data sets
- □ Group similar localisations together
- □ Query MapMan for additional protein information
- □ Search for interactions between interactors

Select databases to query:

- □ Query interactions from BioGrid ●
- □ Query interactions from IntAct ●
- ☑ Query protein-protein interactions from BAR
- □ Query protein-DNA interactions from BAR

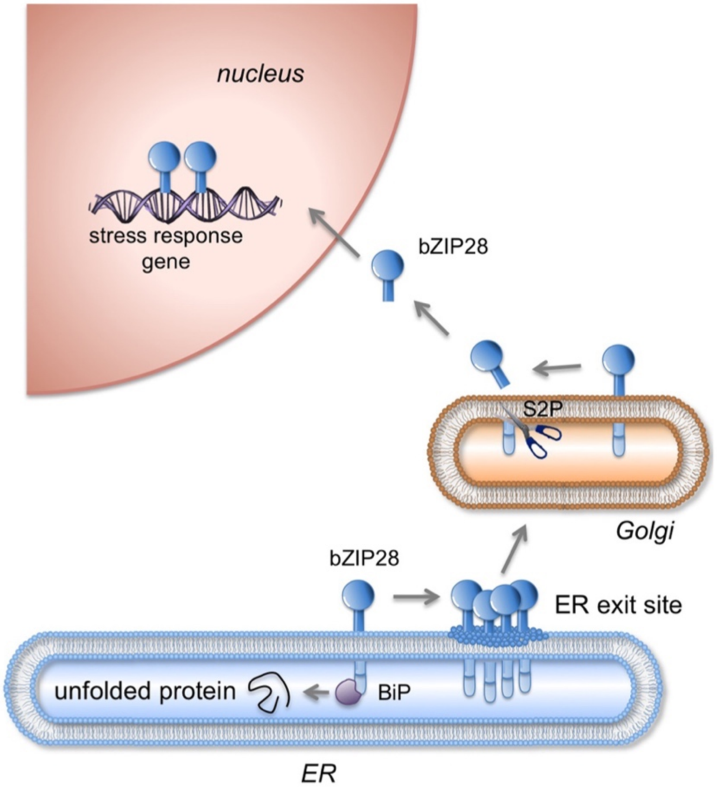
Select a published source to query: All

You can also upload your own interactions network in SIF or CYS format:

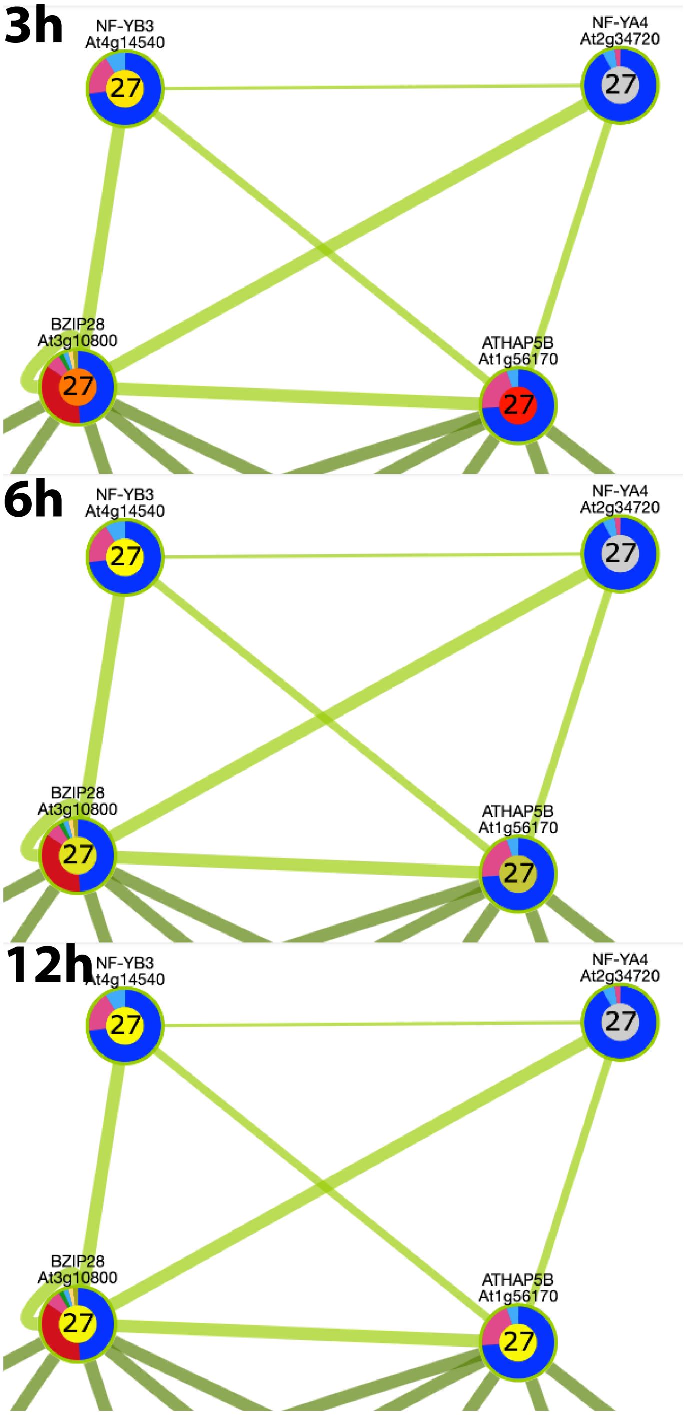
Choose File No file chosen

- □ Extend with interactions from BIND/Interactome 2.0
- □ Overlay edges with interactions from BIND/Interactome 2.0

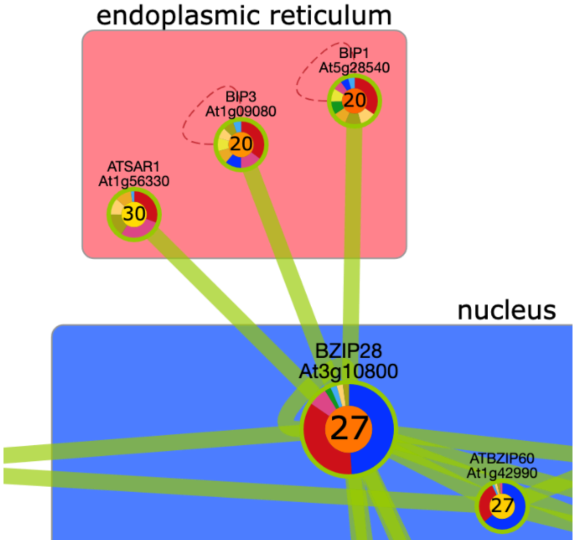
Appendix 7. Old splash page of AIV.



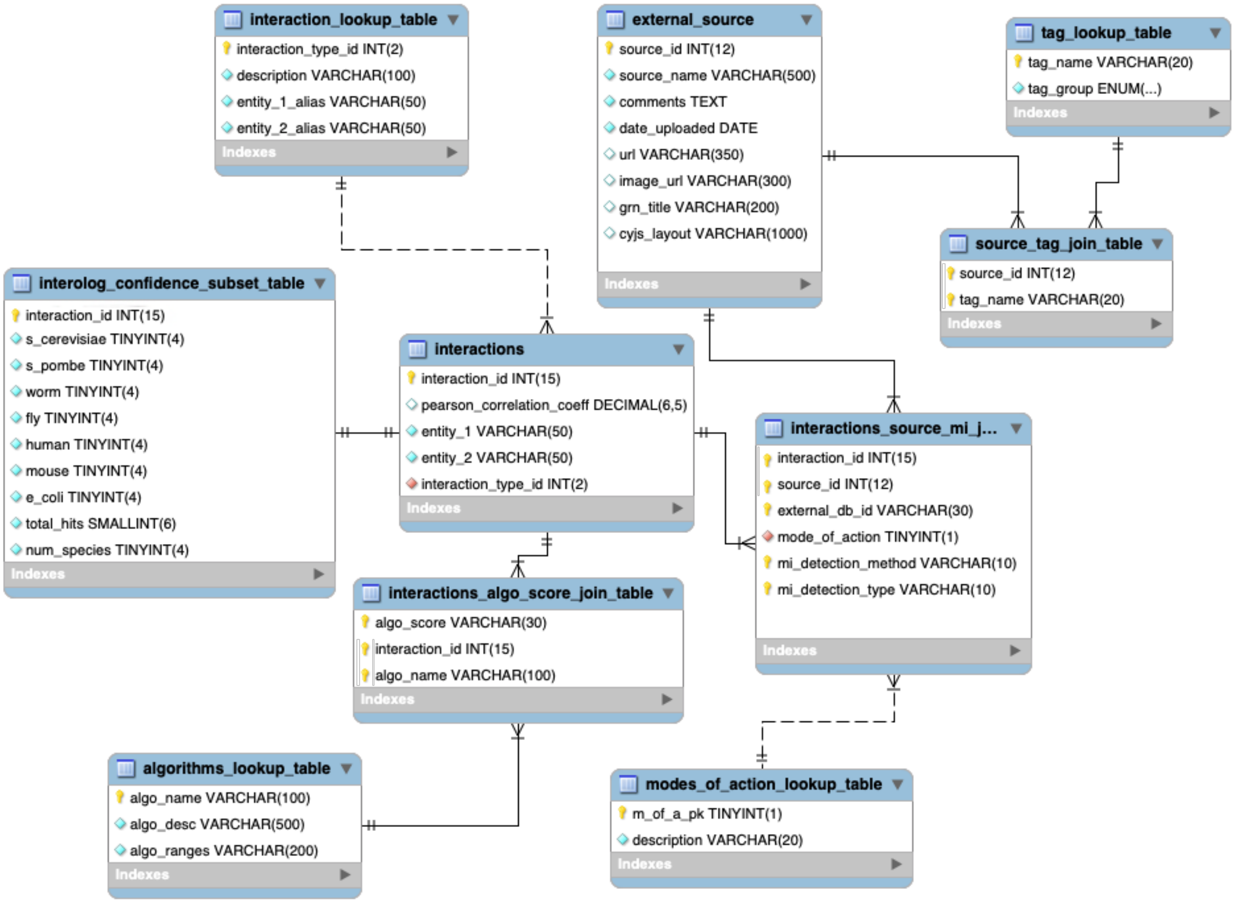
Appendix 8. Mobilization of bZip28 in response to overaccumulation of unfolded proteins. When unstressed, bZIP28 is tethered to the endoplasmic reticulum (ER) by its interaction with Binding Protein (BIP) via its lumen interface. When unfolded proteins accumulate in the ER due to environmental conditions, BIP is outcompeted from bZIP28. bZIP28 is then transported to the Golgi which is processed by Site-2-Protease (S2P) to which catalyzes and releases bZIP28's cytoplasmic domain for nuclear transport and ultimately gene regulation. Adapted from Srivastava, Deng and Howell (2014).



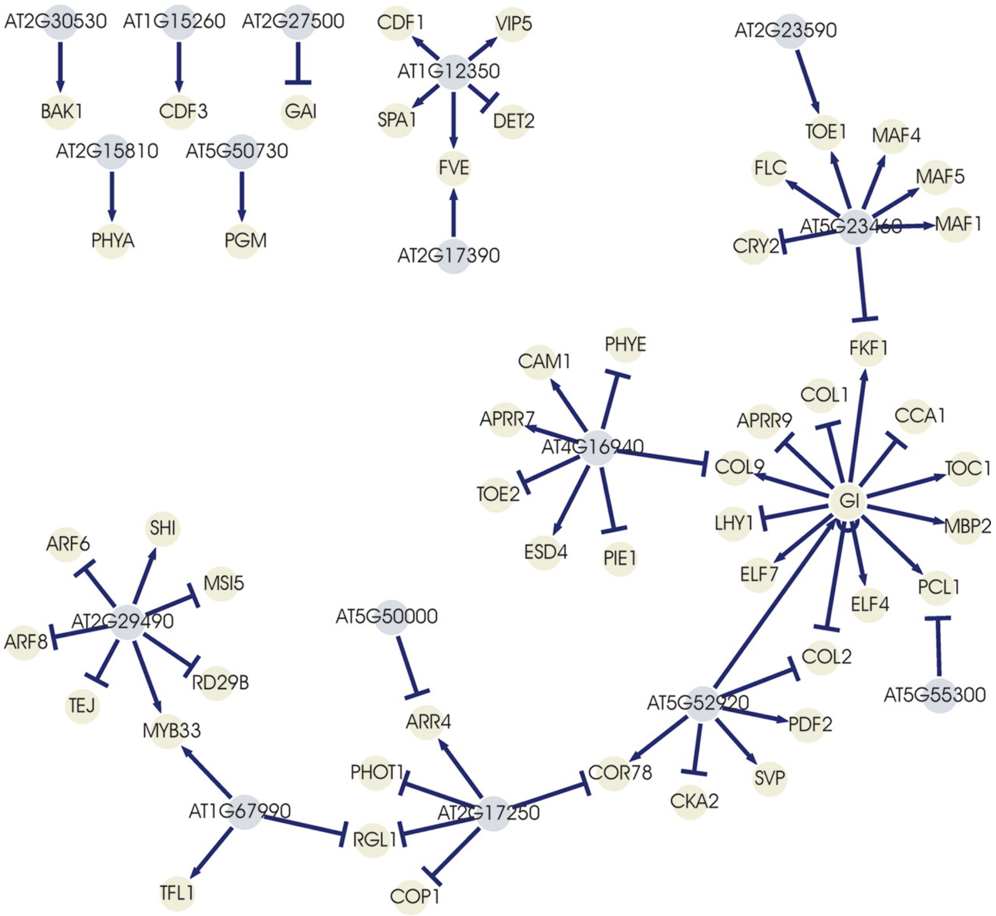
Appendix 9. Expression overlays for bZIP28, NF-YA4, NF-YB3, and NF-YC2 at certain timepoints after heat shock. Scales were set as mentioned in Figure 2.5.



Appendix 10. Localization layout applied to the PPI network as mentioned in Figure 2.6. Image focused on bZIP28 and BIP.



Appendix 11. Entity relationship diagram (ERD) of the new interactions database which integrates prior AIV2 data with curated GRN data. Yellow keys, cyan diamonds, unfilled diamonds, and red diamonds represent primary keys, NOT NULL attributes, NULL attributes, and NOT NULL foreign keys respectively. Solid and dashed lines represent identifying and non-identifying relationships respectively. Single lines and crow's feet represent one-to-one and one-to-many relationships respectively. This ERD was designed in MySQL WorkBench 8.



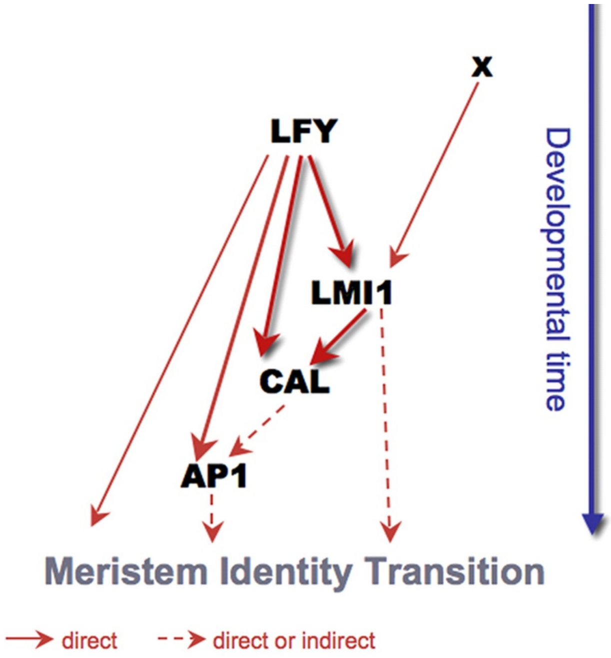
Appendix 12. Regulatory network involved in flowering patterning as predicted by eQTL analysis by Keurentjes et al. (2007). Activating and repressive relationships are represented by arrowheads and tees respectively. Copyright Note: Copyright (2007) National Academy of Sciences

```
{
    "status": "success",
    "data": [
        {
            "source_id": 14,
            "source_name": "17237218",
            "comments": "Flowering Time analysis with genome-
wide expression variation analysis (combining eQTL mapping and regulator candidate gen
e selection) in an RIL population of Arabidopsis thaliana. Data From: manual annotatio
n of Figure 2.",
            "date_uploaded": "2019-11-12T05:00:00.000Z",
            "url": "www.ncbi.nlm.nih.gov/pubmed/17237218",
            "image_url": "https://bar.utoronto.ca/GRN_Images/17237218.jpg",
            "grn_title": "Keurentjes et al. (Proc Nat Acad Sci, 2007) Flowering Time N
etwork",
            "cyjs_layout": "{\"name\": \"cose\", \"animate\" : \"true\"}",
            "tag_name": "eQTL mapping",
            "tag_group": "Experiment",
            "tags": "eQTL mapping:Experiment|ERECTA:Gene|Flower:Condition|GIGANTEA:Gen
e|Linkage Mapping:Experiment|RIL:Experiment"
        },
        {
            "source_id": 18,
            "source_name": "30616516",
            "comments": "Existing time-
course gene expression data for flower development was used to find dynamical network 
biomarker to create a gene regulatory network.",
            "date_uploaded": "2019-11-12T05:00:00.000Z",
            "url": "www.ncbi.nlm.nih.gov/pubmed/30616516",
            "image_url": "https://bar.utoronto.ca/GRN_Images/30616516.jpg",
            "grn_title": "Zhang et al. (Bmc Plant Biology, 2019) Flowering Development
 Network",
            "cyjs_layout": "{\"name\": \"cose\", \"animate\" : \"true\"}",
            "tag_name": "APETALA2",
            "tag_group": "Misc",
            "tags": "APETALA2:Misc|CBC:Misc|CBP20:Gene|DNB:Experiment|Flower:Condition
|LEA:Gene|NAP12:Gene|RIE1:Gene"
        }
    ]
}
```

Appendix 13. Sample JSON output when a GET request is submitted to https://bar.utoronto.ca/interactions_api/tags/flower.

```
#grn_title:Brady et al.(MOL SYST BIOL, 2011) Root Stele Network
#source_name:21245846
#comments:Root stele gene network initially mapped with Y1H and Y2H on highly-enriched TFs (based on root spatiotemporal map) and miRNA-of-interest promoters. In planta confirmation and regulation determnined via ChIP and qPCR. - Vincent
#tags:Y1H|Y2H|Root Stele|qPCR|ChIP|OBP2|REV
---
AT2G44940 pdi AT5G60200|407|432
AT3G60490 pdi AT5G60200|407|432
AT3G00000 pdi-r AT5G00000|407|432
```

Appendix 14. General structure of our customized simple interactions format (SIF) file for uploading GRNs to our database. The format follows the gene-A <interaction-type> gene-B format where appended letters after the '-' represents the modality (r represents repression). Additional custom MI terms are added for curation.



Appendix 15. Model of meristem identity transition as identified by Saddic et al. (2006) where LFY activates LM1, which both activate CAL through a feed-forward loop (FFL). Other meristem identity genes are implicated such as AP1 which is downstream of CAL. Solid and dashed lines represent direct or indirect effects respectively.

Copyright Acknowledgements

Figure 1.2 (Boruc *et al.*, 2010) used with permission: “Permission is granted for the life of the current edition and all future editions, in all languages and in all media.” by the American Society of Plant Biologists Terms and Conditions.

Figure 1.4 (Brady et al., 2011) used with permission: “This article is available under the terms of the Creative Commons Attribution Non-Commercial License CC BY-NC (which may be updated from time to time) and permits non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited.”

Figure 1.5 (Alon, 2007) used with permission: Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature Nature Reviews Genetics Network motifs: theory and experimental approaches, Uri Alon, 2007

Figure 1.8 (Chen *et al.*, 2018) used with permission: Reprinted/adapted by permission from Springer Nature: Springer Nature Same Stats, Different Graphs by Hang Chen, Utkarsh Soni, Yafeng Lu et al. Copyright (2018)

Figure 1.9 (Ahnert, 2013) used with permission: Ahnert, S. E. (2013) ‘Power graph compression reveals dominant relationships in genetic transcription networks’, Molecular BioSystems, 9(11), pp. 2681–2685. doi: 10.1039/c3mb70236g. - Reproduced by permission of The Royal Society of Chemistry

Appendix 8 (Srivastava, Deng and Howell, 2014) used with permission: “In most cases, adaptation and reuse of figures is permitted provided that the authors and original source are appropriately credited and that no third-party licenses apply (please see the citation on the article on-line page). Frontiers does not provide any formal permissions for reuse.”

Appendix 12 (Keurentjes et al. (2007) used with permission: “Permission is not required to use original figures or tables for noncommercial and educational use (i.e., in a review article, in a book that is not for sale) if the article published under the exclusive PNAS License to Publish. Please include a full journal reference and, for articles published in volumes 90–105 (1993–2008), include "Copyright (copyright year) National Academy of Sciences" as a copyright note.

Commercial reuse of figures and tables (i.e., in promotional materials, in a textbook for sale) requires permission from PNAS."

Appendix 15 Saddic et al. (2006) used with permission: Reproduced / adapted with permission from Development. DOI: 10.1242/dev.02331