Title

Identification of gene expression logical invariants in Arabidopsis

Journal

Plant Direct, 3(3)

ISSN

2475-4455

Authors

Pandey, Sonalisa

Sahoo, Debashis

Publication Date

2019-03-01

DOI

10.1002/pld3.123

Peer reviewed

Received: 18 November 2018 | Revised: 28 December 2018 | Accepted: 1 February 2019

DOI: 10.1002/pld3.123

ORIGINAL RESEARCH

American Society of Plant Biologists
SEB SOCIETY FOR EXPERIMENTAL BIOLOGY
WILEY

Identification of gene expression logical invariants in *Arabidopsis*

Sonalisa Pandey | **Debashis Sahoo**

University of California San Diego, San Diego, California

Correspondence
Debashis Sahoo, University of California San Diego, San Diego, CA.
Email: dsahoo@ucsd.edu

Abstract

Numerous gene expression datasets from diverse tissue samples from the plant variety *Arabidopsis thaliana* have been already deposited in the public domain. There have been several attempts to do large scale meta-analyses of all of these datasets. Most of these analyses summarize pairwise gene expression relationships using correlation, or identify differentially expressed genes in two conditions. We propose here a new large scale meta-analysis of the publicly available *Arabidopsis* datasets to identify Boolean logical relationships between genes. Boolean logic is a branch of mathematics that deals with two possible values. In the context of gene expression datasets we use qualitative high and low expression values. A strong logical relationship between genes emerges if at least one of the quadrants is sparsely populated. We pointed out serious issues in the data normalization steps widely accepted and published recently in this context. We put together a web resource where gene expression relationships can be explored online which helps visualize the logical relationships between genes. We believe that this website will be useful in identifying important genes in different biological context. The web link is http://hegemon.ucsd.edu/plant/.

KEYWORDS

bioinformatics, Boolean analysis, microarray, systems biology

1 | INTRODUCTION

A large amount of microarray and RNASeq datasets has been continually deposited in public databases such as GEO (Gene Expression Omnibus) (Barrett et al., 2005, 2013; Edgar, Domrachev, & Lash, 2002) and ArrayExpress (Brazma et al., 2003; Rocca-Serra et al., 2003). It is challenging to put together all of these data from different labs or different studies in order to facilitate comparisons across labs and diversity of tissue types; however, there have been several attempts of massive large scale data analysis that provides new hypothesis and insight into biological processes. NASCArrays is one of the first few that started this revolution in the plant community (Craigon et al., 2004). Following this, several studies have put together large databases (Ball et al., 2005; He et al., 2016; Lukk et al., 2010; Schmid, Palmer, Kohane, & Berger, 2012; Zimmermann, Hirsch-Hoffmann, Hennig, & Gruissem, 2004), and web resources to investigate gene-gene relationships online (Katari et al., 2010; Manfield et al., 2006; Mutwil, Obro, Willats, & Persson, 2008; Obayashi & Yano, 2014; Srinivasasainagendra, Page, Mehta, Coulibaly, & Loraine, 2008; Toufighi, Brady, Austin, Ly, & Provart, 2005). The GeneMANIA App in Cytoscape also incorporates co-expression analysis of transcriptomic datasets (Montojo et al., 2010).

The Arabidopsis Information Resource (TAIR) provides an integrative platform where many different datatypes, including gene expression, can be effectively analyzed (Rhee et al., 2003).

Despite these largescale efforts, none of the above resources provide interfaces to analyze Boolean logical relationships between genes. Boolean logic is mathematics of two possible values. In the context of gene expression data, one might ask what other genes are highly expressed if the expression of gene A is high. Boolean logical gene-gene relationship is mathematically the simplest form of relationship between genes. We have published earlier how Boolean relationship can be explored in large microarray datasets (Sahoo, 2012; Sahoo, Dill, Gentles, Tibshirani, & Plevritis, 2008). Boolean relationship is identified by searching for at least one sparsely populated quadrant out of four possible quadrants by the BooleanNet algorithm (Sahoo et al., 2008). According to this BooleanNet algorithm, there are six potential Boolean implications of gene relationships: two symmetric Boolean implications (Equivalent and Opposite) and four asymmetric Boolean implications (Sahoo et al., 2008). Two genes are considered Boolean equivalent if they are positively correlated with only high-high and low-low gene expression values. Two genes are considered Boolean opposite if they are negatively correlated with only high-low and low-high gene expression values. Asymmetric Boolean implications result when there is only one sparsely populated quadrant.

In this paper, we put together a web resource for the plant community to explore Boolean logical gene-gene relationships. In addition, we describe special types of relationships called logical invariants in detail. An invariant is a formula that evaluates to true in a universe of sample types. A universe consists of a coherent set of samples from a particular tissue. The plant universe consists of all plant tissue types. All samples from roots can form a universe: the root-universe. Similarly, we can have a shoot, leaf, and flower universe. A logical invariant is associated with a particular universe where it evaluates to true. A Boolean logical gene-gene relationship will be called a logical invariant if all possible samples from that universe follow the same logical rule. Therefore, it is hard to call any relationship logical invariant because we do not have access to all possible samples. However, we could hypothesize a Boolean relationship logical invariant if the relationship looks strong. In this paper, we will identify several candidate logical invariants in all plant tissues as well as specific tissue types.

2 | METHODS

2.1 | Data collection and annotation

Publically available microarray databases in *Arabidopsis thaliana* ATH1 (number of microarray samples in 2014 $n = 4{,}306$, GPL198) Affymetrix platform were downloaded from the National Center for Biotechnology Information (NCBI) GEO website (Barrett et al., 2005, 2013; Edgar et al., 2002). Gene expression summarization was performed by normalizing each Affymetrix platform by Robust Multi-Array Average (RMA) (Irizarry et al., 2003). We downloaded the latest dataset where tissue type, growth conditions, and developmental stage were manually curated for each sample (GSE69995, $n = 6{,}057$) (He et al., 2016). In addition to these large datasets which are mostly bulk tissue datasets, we put together a couple of datasets where specific cell types were purified using Fluorescence-Activated Cell Sorting (FACS) method. The Yadav-2014 dataset that provides a high resolution map of the shoot apical meristem (SAM) cell types of central zone (CZ), peripheral zone (PZ), and rib meristem (RM) (GSE28109, GSE13596)(Yadav, Girke, Pasala, Xie, & Reddy, 2009; Yadav, Tavakkoli, Xie, Girke, & Reddy, 2014) was prepared. Similarly, the Benfey dataset that provides a high-resolution spatiotemporal map of the root (GSE15876, GSE16468, GSE16469, GSE21582, GSE30166, GSE35580, GSE5749, GSE61408, GSE64253, GSE7641, and GSE8934) (Bargmann et al., 2013; Birnbaum et al., 2003; Brady et al., 2007; Carlsbecker et al., 2010; Dinneny et al., 2008; Efroni, Ip, Nawy, Mello, & Birnbaum, 2015; Iyer-Pascuzzi et al., 2011; Lee et al., 2006; Long et al., 2010; Nawy et al., 2005; Sozzani et al., 2010) was prepared. Both the Yadav-2014 and Benfey dataset use specific reporter lines and purify specific cell types and profile them using the *Arabidopsis thaliana* ATH1 Affymetrix platform. We have also collected multiple RNASeq datasets ($n = 747$) using previously published tool by Zhuo, Emerson, Chang, and Di (2016) StablyExpressedGenes. We prepared these RNASeq datasets by computing TPM (Li & Dewey, 2011; Pachter, 2011) values using a custom perl script. We used log2(TPM) if TPM > 1 and (TPM − 1) if TPM < 1 as the final gene expression value.

2.2 | Duplicate CEL files identification

Following the 330 experiments from the previously published dataset GSE69995, 6,535 CEL files were downloaded from GEO. 6,057 CEL files were used to build the published dataset GSE69995. Quality control steps from the "simplyaffy" and "affyPLM" data packages (Bolstad, Irizarry, Astrand, & Speed, 2003) were used before to identify these 6,057 CEL files that exclude 478 files (6,535–6,057) from our list. We computed a MD5 hash of each file and compared them to check if there were duplicate entries under a different file name. If two files were identical, their MD5 hash was matched even if the file names were different. In 6,535 CEL files, we found a total of 87 duplicated entries (Supporting Information Table S1) and 85 duplicated entries were present in the published dataset GSE69995. We created a new dataset based on this after removing all 85 duplicates from GSE69995, with a total of 5,972 files (6,057 − 85 = 5,972). Our dataset is available at GEO using the accession no GSE118579.

2.3 | Boolean analysis of datasets

The expression values of each gene were ordered from low to high and a rising step function was computed to define a threshold by StepMiner algorithm in the individual dataset (Sahoo, Dill, Tibshirani, & Plevritis, 2007). If the assigned threshold for a gene was t, then expression levels above $t + 0.5$ were classified as "high", and the expression levels below $t - 0.5$ were classified as "low". Expression levels between $t - 0.5$ and $t + 0.5$ were classified as "intermediate". A previously

published BooleanNet algorithm was performed to determine Boolean Implication relationships between genes. Briefly, the BooleanNet algorithm searches for at least one sparsely populated quadrant in a scatterplot between two genes. The "intermediate" expression values were ignored by the BooleanNet algorithm. There were six possible scenarios: one of the four quadrants was sparse (four asymmetric Boolean implications) and two diagonally opposite quadrants were sparse (Equivalent and Opposite Boolean implications). We used the same thresholds as in our previously published algorithm: statistic > 3 and error-rate < 0.1.

2.4 | Web-based visualization

Boolean implication relationships were visualized using two dimensional scatterplots between two genes. The scatterplot shows the normalized expression values of two genes along with the thresholds that separate the high and low values. Sparsely populated quadrants can be immediately spotted by visual inspection. Each individual point in the scatterplot belongs to a particular sample that can be traced to its original source at GEO with a GEO accession number. The samples in the plot can be selected using a mouse by dragging a rectangle in the plot. A group was created with the number of the sample shown on the right side of the scatterplot. The interface lets the user supply two genes at the top in two different textboxes. The textbox can be used to input a set of genes separated by whitespace. When the user clicks on "getPlots", all possible pairs of probesets derived from the two sets of genes are plotted.

2.5 | Comparison of Boolean networks between GSE69995 and our (Pandey 2018) dataset

A direct head-to-head comparison with a large dataset identified by GEO accession number GSE69995 was performed to check if data processing steps influence the discovery of logical relationships (He et al., 2016). We re-processed the same dataset using our Boolean analysis pipeline. We used RMA to normalize the dataset while the GSE69995 dataset was normalized using MAS 5.0(Hubbell, Liu, & Mei, 2002). We matched the probeset IDs of the two datasets using the Affymetrix annotation for GPL198 which is the GEO accession number of the *Arabidopsis thaliana* ATH1 Affymetrix platform. We computed the full Boolean implication network in both datasets. For each probeset ID A we discovered six different possible Boolean implication relationships: A low ⇒ X high (lohi), A low ⇒ X low (lolo), A high ⇒ X high (hihi), A high ⇒ X low (hilo), A equivalent X (eqv), and A opposite X (opo). We plotted the number of relationships identified in both datasets in a scatter plot with a log-log scale to compare the approaches.

3 | RESULTS

3.1 | Identification of duplicate entries in previous datasets

To gain more insight into gene function in plants, it is important to study tissue-specific gene activity under a variety of conditions. However, the analysis of public expression data by the plant research community is hampered by the lack of consistent sample annotation. Searching keywords in the metadata fields for each expression sample in the GEO, such as "Characteristics," "Description", and "Source name" is not reliable because of inconsistent annotations. We discovered a largescale meta-analysis of previously published datasets in GEO (GSE69995) (He et al., 2016). This dataset consists of a carefully annotated description of each sample. Therefore, it was relevant for our study to investigate logical relationships between genes. It is important to understand that accurate annotation is key to success. It is also important to remove any technical biases which may hamper further interpretation from the dataset. For example, if a sample is duplicated several times in a particular dataset, it may lead to unintended consequences in the analysis and interpretation. We discovered 85 duplicated entries in this dataset. While these duplicates may not be highly significant relative to the scale of the dataset, they should be removed before any meta-analysis is performed. Supporting Information Table S1 lists all the duplicated entries in this dataset. The plant community should be aware of such samples in the dataset.

3.2 | Identification of six possible types of logical invariants

A full Boolean implication network was created using the new dataset. The analysis identified six possible types of Boolean implication relationships. Figure 1 shows an example of each that might be associated with some known gene functions in plants. For example, ARABIDOPSIS THIOREDOXIN Y2 (ATY2) and PHOTOTROPIN 2 (PHOT2) have a logical equivalent relationship as the top-left and bottom-right quadrants are sparse (Figure 1a). PHOT2 is a membrane-bound protein serine/threonine kinase that functions as a blue light photoreceptor (Sakai et al., 2001). ATY2 is mainly expressed in leaves and induced by light (Collin et al., 2004). ARABIDOPSIS THIOREDOXIN Y1 (ATY1) and ATY2 have clear opposite relationship (Figure 1d). ATY1 is mainly expressed in non-photosynthetic organs including seeds (Collin et al., 2004). When the APETALA 3 (AP3) expression level is low, the LIPID TRANSFER PROTEIN 12 (LTP12) expression level is also low (Figure 1b, AP3 low ⇒ LTP12 low, LTP12 high ⇒ AP3 high). AP3 is mainly expressed in flower petal and stamen (Bowman, Smyth, & Meyerowitz, 1989), while LTP12 is expressed specifically in anther and pollen (Li et al., 2014). This is consistent with the logical relationship demonstrated by LTP12 expression in a subset of tissues in flower, anther, pollen, and stamen. FER-LIKE REGULATOR OF IRON UPTAKE (FRU) is mainly expressed in the root (Bauer et al., 2004), therefore it is consistent with the logical relationship of AP3 high ⇒ FRU low (Figure 1c). Figure 1e shows the relationship between GLUTAMATE DEHYDROGENASE 2 (GDH2) and SHORT HYPOCOTYL IN WHITE LIGHT1 (SHW1): GDH2 low ⇒ SHW1 high. The three different quadrants in the scatterplot between GDH2 and SHW1 are populated with three different tissues: roots (bottom right, GDH2 high SHW1 low), seedlings (top right, GDH2 high SHW1 high), and leaves (top left, GDH2 low SHW1

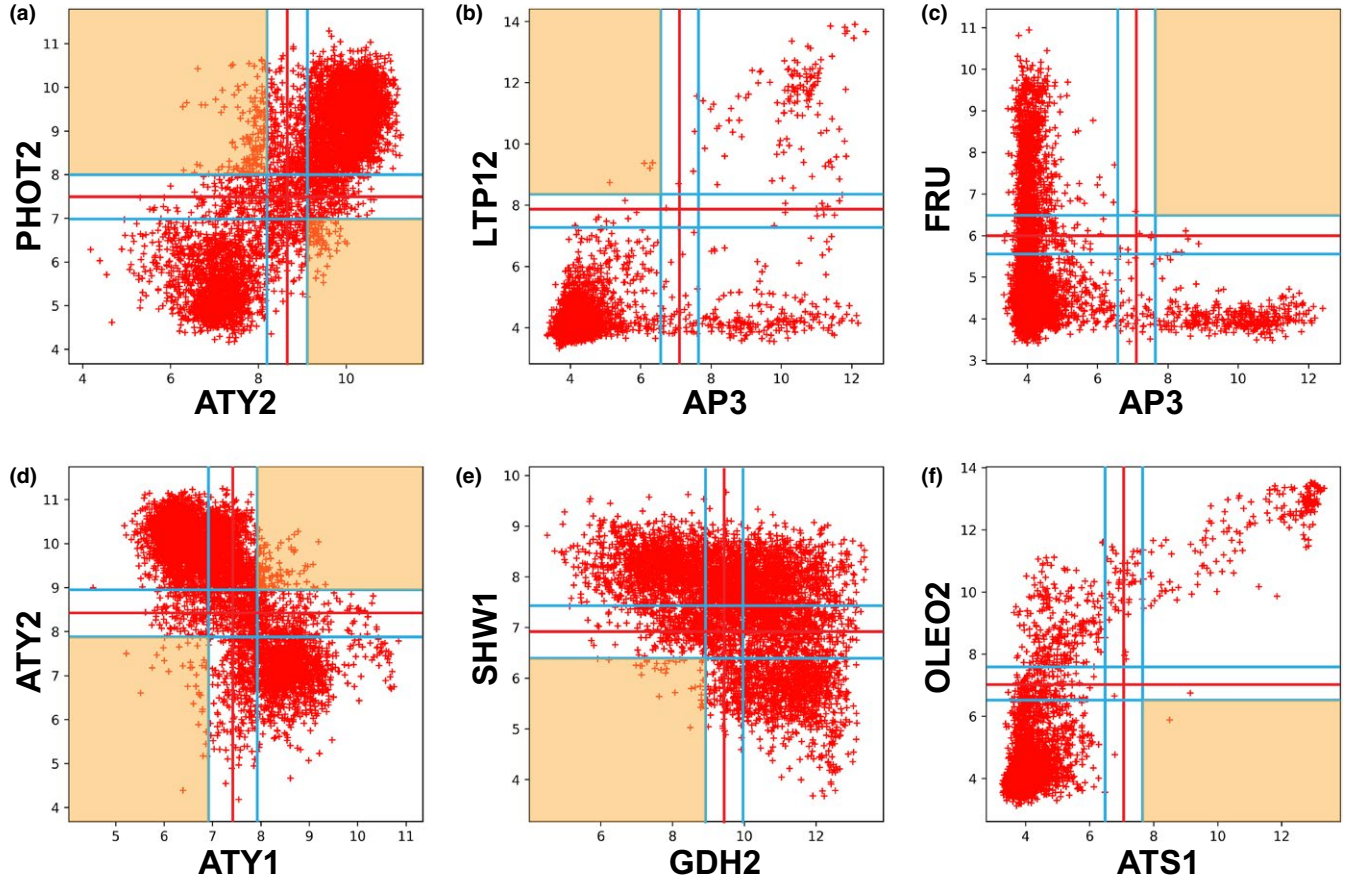

FIGURE 1 Six possible logical gene-gene relationships. Every point in the plot is a microarray experiment performed in the ATH1 Affymetrix platform. The *x*- and *y*-axes represent log2 normalized gene expression values. Sparsely populated quadrants are highlighted with orange. (a, d) Symmetric relationships. (b, c, e, f) Asymmetric relationships. (a) ATY2 equivalent PHOT2. (b) AP3 low ⇒ LTP12 low. (c) AP3 high ⇒ FRU low. (d) ATY1 opposite ATY2. (e) GDH2 low ⇒ SHW1 high. (f) ATS1 high ⇒ OLEO2 high

high). When ARABIDOPSIS THALIANA SEED GENE 1 (ATS1) expression is high, OLEOSIN 2 (OLEO2) expression is also high (Figure 1f, ATS1 high ⇒ OLEO2 high, OLEO2 low ⇒ ATS1 low). Both OLEO2 and ATS1 are mainly expressed in seeds (Kim, Hsieh, Ratnayake, & Huang, 2002; Nuccio & Thomas, 1999). All of the six possible relationships described above are strong in all data points. In other words, almost all of the data points follow the Boolean formula. Therefore, they are candidates for logical invariants in plants.

3.3 | Comparison of Boolean network with previous dataset

A detailed comparison was performed between our dataset (Pandey 2018) and the previously published He-2016 dataset GSE69995 to check if discovery of logical relationships was sensitive to the data processing steps. He et al. (2016) used MAS 5.0 normalization, while we used RMA. Figure 2 shows the number of relationships for each probesets in both dataset using scatterplots. *X*-axes represent GSE69995, *y*-axes represent our approach, and the different scatterplots correspond to the six possible logical relationships. As can be seen in the figure, our approach identified significantly more logical relationships in all other comparisons except A low ⇒ X high (Figure 2e). The *p*-value was less than 0.001 for equivalent, opposite, lolo, hihi, and hilo. Boolean approach discovered more A low ⇒ X high in GSE69995 compared to our dataset. We conclude that the Boolean approach is best suited for data processing steps using RMA. Below we describe a few reasons why our algorithm did not find many statistically significant relationships in GSE69995.

To get a deeper insight into the structure of the Boolean network, we show four scatterplots that demonstrate the discrepancy between datasets. Figure 3a shows a scatterplot between AP3 and FRU in three datasets including GSE69995, Pandey 2018, and Zhuo RNASeq. There is no significant logical relationship in the GSE69995 dataset, whereas our dataset and the RNASeq dataset show very clear AP3 high ⇒ FRU low relationship. In the scatterplots, root and flower tissue samples are highlighted in dark blue and red, respectively. The GSE69995 dataset shows that FRU is highly expressed in root samples, AP3 is high in flower samples, but some of the flower samples may have high levels of FRU, and many root samples may have high levels of AP3. However, both our dataset and the RNASeq dataset shows that all of the flower samples have low to intermediate levels of FRU, and all root samples have low levels of AP3. Our data suggest that FRU and AP3 expression levels are clearly mutually

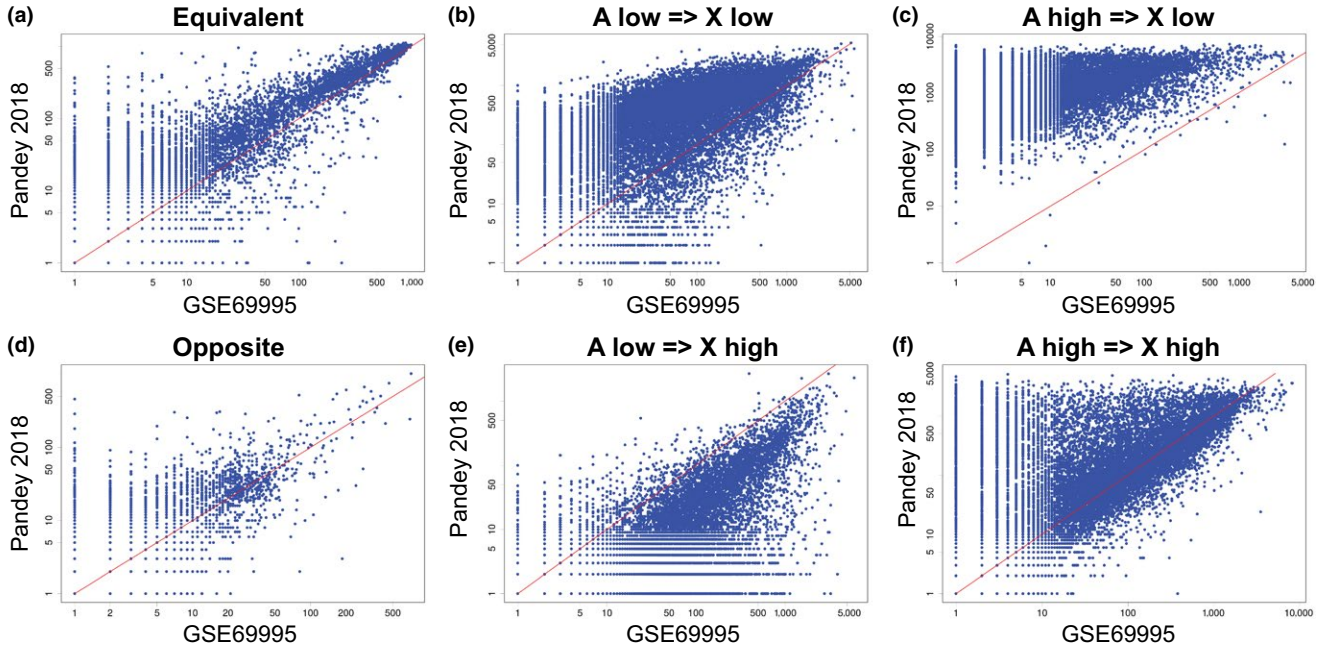

FIGURE 2 Comparison of Boolean network between GSE69995 and Pandey 2018 dataset. Every point in the plot is a probeset ID in the ATH1 Affymetrix platform. The *x*- and *y*-axes represent log2 count of the respective logical relationships. *x* = *y* is plotted with a red line. (a, d) Symmetric relationships. (b, c, e, f) Asymmetric relationships. (a) A equivalent X. (b) A low ⇒ X low. (c) A high ⇒ X low. (d) A opposite X. (e) A low ⇒ X high. (f) A high ⇒ X high. Our approach discovered more significant logical relationships than the previously published dataset in all other cases except A low ⇒ X high. Gene A and Gene X are candidate probeset IDs in each dataset

exclusive which is consistent with the literature (Jakoby, Wang, Reidt, Weisshaar, & Bauer, 2004; Kramer & Irish, 1999). These scatterplots also highlight that there may be discrepancy in identifying threshold in many genes as the threshold of AP3 is remarkably different in both datasets. Figure 2c shows a clear increase in the frequency of A high ⇒ X low in our dataset compared to the GSE69995 dataset. We hypothesize that this increase is due to the discrepancy in identifying threshold in two datasets.

Figure 2e suggested that the frequency of A low ⇒ X high is higher in the GSE69995 dataset compared to our dataset and the Zhuo RNASeq dataset. We hypothesize that there may be technical issues associated with the discovery of A low ⇒ X high in the GSE69995 dataset. Figure 3b shows a scatterplot between CONSTANS-LIKE 4 (COL4) and ARABIDOPSIS THALIANA RECEPTOR KINASE 1 (ARK1). In the GSE69995 dataset the relationship is COL4 low ⇒ ARK1 high. It also suggests that ARK1 expression levels are higher in root samples compared to other samples. In contrast, both our dataset and the RNASeq dataset suggest that the relationship is COL4 low ⇒ ARK1 low, and the expression levels of ARK1 in root samples are low. Literature is consistent with our observation, since Northern blot analyses prepared from various tissues including floral bud, leaf, root and stem only detected ARK1 expression in leaf and floral bud (Tobias, Howlett, & Nasrallah, 1992). In the TAIR database, ARK1 is annotated as "not expressed" in root. We observed that A low ⇒ X high is usually rare in human and mouse datasets. The overwhelmingly high frequency of A low ⇒ X high in GSE69995 may be due to a technical bias.

3.4 | Web resource for easy exploration

We provide a web interface where the gene expression data can be explored using two dimensional scatterplots. Using this interface, the user can start with two well defined gene names and query the database to plot the normalized expression values in a scatterplot. Each individual data point in the scatterplot is linked to the GSE accession number. The website provides a link to the GEO website where details of the experiment are found. The website has several features to explore Boolean logical relationships between genes. Using a mouse, the user can select a set of experiments from the scatterplot by dragging a rectangle. The groups of experiments can be manipulated using various sets of operations: union, intersection, and difference. Previously defined manual annotations can be explored on the right side of the window using drop down options and several buttons. The website also provides an annotation browser where GEO annotations can be searched conveniently using mouse clicks.

4 | DISCUSSION

Integrative data analysis platforms where all publicly available databases with diverse data types are coherently put together to propose novel hypotheses for ongoing deep investigation of biological processes is key for success in this new era of genomic data revolution. Tools that are developed in both plant community and human disease studies will significantly benefit each other. StepMiner

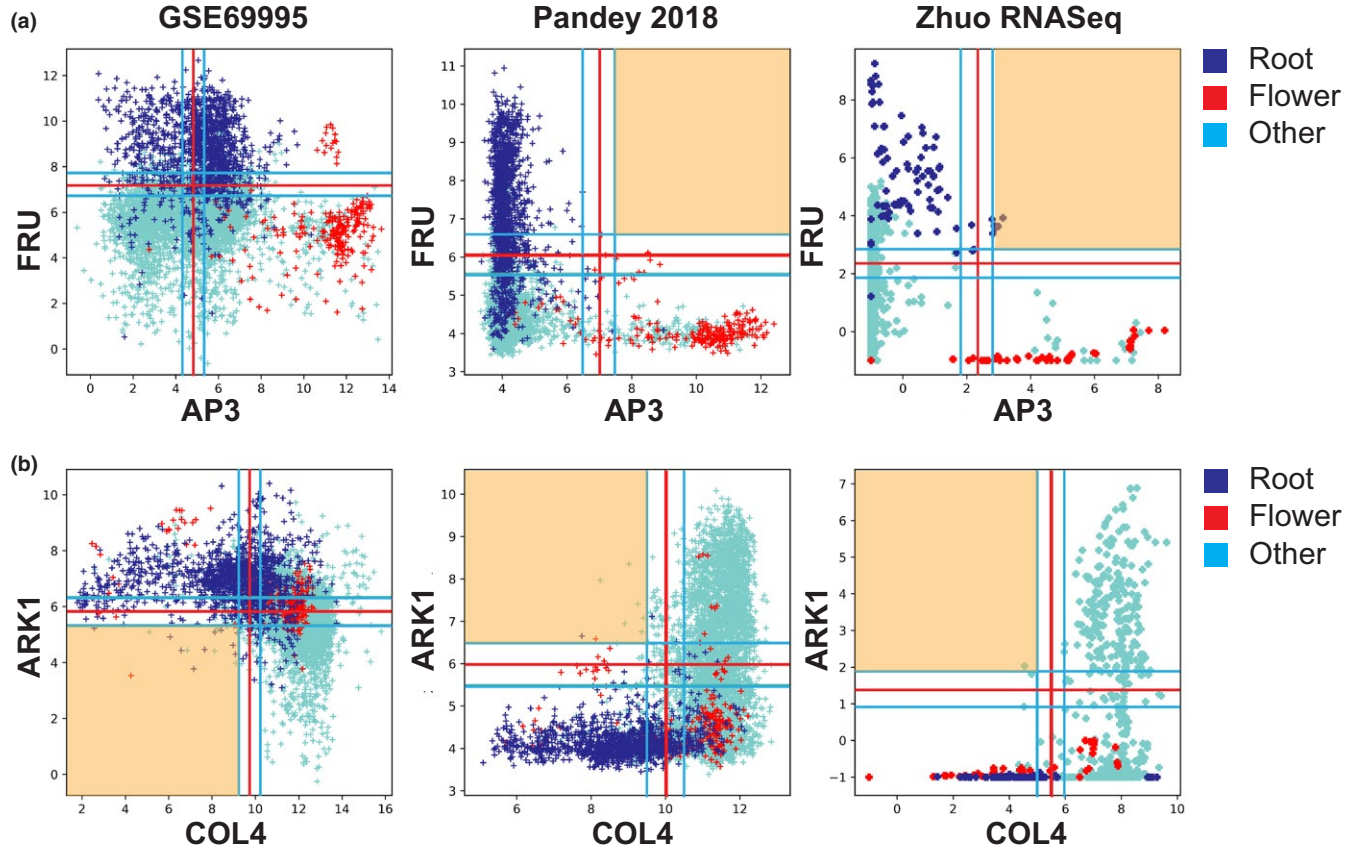

FIGURE 3 Data normalization and Boolean implication. In GSE69995 and Pandey 2018 datasets every point in the plot is a microarray experiment in the ATH1 Affymetrix platform. In Zhuo RNASeq datasets every point is an RNASeq experiment in a Arabidopsis tissue sample. The *x*- and *y*-axes represent log2 normalized expression values. Root and flower tissue samples are highlighted with dark blue and red, respectively. (a) GSE69995: no relationship, some flower samples may express high levels of FRU; Pandey 2018, Zhuo RNASeq: AP3 high ⇒ FRU low, FRU is expressed in roots and AP3 is expressed in flower samples. (b) GSE69995: COL4 low ⇒ ARK1 high; ARK1 is expressed in root samples; Pandey 2018, Zhuo RNASeq: COL4 low ⇒ ARK1 low, ARK1 is not expressed in root samples

(Sahoo et al., 2007), BooleanNet (Sahoo et al., 2008) and MiDReG (Sahoo et al., 2010) are examples of computational tools that were developed primarily to analyze human normal and cancer tissues are directly applicable in plant studies because the data characteristics are similar. Microarray data in human tissues and the plant tissues can be processed similarly. In this paper, we identify Boolean relationships between *Arabidopsis* genes; some of the genes are homologous to human genes. The comparison of data processing steps that we performed here will also benefit human studies.

Data processing steps strongly influence the downstream analysis. In this context, the choice of normalization steps has been debated before (Harr & Schlotterer, 2006; Lim, Wang, Lefebvre, & Califano, 2007). In this study, we conclude that RMA is more appropriate than MAS 5.0 for the investigation of Boolean logical gene-gene relationships. ATH1 Affymetrix platform was extremely popular for initial transcriptomics studies within the Arabidopsis community. However, it contains a set of 22K probeset IDs, whereas the latest annotation of genes in this species is around 32K (Swarbreck et al., 2008). Therefore, it should be noted that the ATH1 platform may be missing well over 20% of Arabidopsis transcriptomic information. ATH1 experiments are being replaced by RNASeq alternatives nowadays. We show that RNASeq studies are also amenable for Boolean analysis by using log transformed TPM values.

A recent study focused on microarray and RNA-seq based global and targeted co-expression networks in *Arabidopsis* (Liesecke et al., 2018). This study identified Pathway Level Co-expression using a set of guide genes, and compared how Pearson Correlation Coefficient (PCC), Spearman Correlation Coefficient (SCC), their respective ranked values (Highest Reciprocal Rank (HRR)), Mutual Information (MI) and Partial Correlations (PC) performed on global networks. In another study, a co-expression database for plant species ATTED-II (http://atted.jp) was published to aid in the discovery of relationships of unknown genes within a species (Obayashi, Aoki, Tadaka, Kagaya, & Kinoshita, 2018). ATTED-II (version 9) provides 16 co-expression platforms for nine plant species, including seven species supported by both microarray- and RNA sequencing (RNAseq)-based co-expression data (Obayashi et al., 2018). Similarly, co-expression networks have been a popular tool in the literature to understand regulatory pathways in *Arabidopsis* (He & Maslov, 2016; Van Bel & Coppens,

2017; Zheng et al., 2017). All of the above studies focused on symmetric relationships between genes. However, our approach suggests that majority of the interesting biological information is present in the asymmetric relationships between genes which are often blurred in the co-expression network investigations. Boolean relationships have been used to understand cell fate decisions in both normal (Inlay et al., 2009; Sahoo et al., 2010) and cancer tissues (Dalerba et al., 2011; Sahoo, 2012; Volkmer et al., 2012). Boolean relationships have been used to identify important biomarkers in colon cancer which was published in the New England Journal of Medicine (Dalerba, Sahoo, & Clarke, 2016; Dalerba, Sahoo, Paik, et al., 2016).

In summary, largescale global network analyses have tremendous potential in influencing the way plant biological investigations are approached. Co-expression networks have been influential in this process. We sincerely believe that Boolean implication networks will benefit the ongoing investigations of plant biological processes by the plant community. We have provided several useful webservers and software packages to help biologists to systematically analyze their high-throughput transcriptome data. We will constantly revise these software packages to make them more user-friendly and effective based on users' suggestions, comments, and recommendations.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (NIH) grant #R00-CA151673 to DS, 2017 Padres Pedal the Cause/Rady Children's Hospital Translational PEDIATRIC Cancer Research Award to DS, 2017 Padres Pedal the Cause/C3 Collaborative Translational Cancer Research Award to DS. We thank Martin Yanofsky, Juan-José Ripoll, Brian Crawford, and Yanofsky lab researcher and staff for help and support during various phases of the study. We thank Jouni M. Vesa for excellent critical review at the early stages of the manuscript.

AUTHORS CONTRIBUTION

DS collected data and processed them for analysis. SP, DS analyzed data, made figures, and wrote manuscript.

DATA ACCESSIBILITY

GEO Accession No: GSE118579. http://hegemon.ucsd.edu/plant.

REFERENCES

Ball, C. A., Awad, I. A., Demeter, J., Gollub, J., Hebert, J. M., Hernandez-Boussard, T., … Sherlock, G. (2005). The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Research*, *33*, D580–D582.

Bargmann, B. O., Vanneste, S., Krouk, G., Nawy, T., Efroni, I., Shani, E., … Birnbaum, K. D. (2013). A map of cell type-specific auxin responses. *Molecular Systems Biology*, *9*, 688.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., … Edgar, R. (2005). NCBI GEO: Mining millions of expression profiles–database and tools. *Nucleic Acids Research*, *33*, D562–D566.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., … Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets–update. *Nucleic Acids Research*, *41*, D991–D995.

Bauer, P., Thiel, T., Klatte, M., Bereczky, Z., Brumbarova, T., Hell, R., & Grosse, I. (2004). Analysis of sequence, map position, and gene expression reveals conserved essential genes for iron uptake in Arabidopsis and tomato. *Plant Physiology*, *136*, 4169–4183. https://doi.org/10.1104/pp.104.047233

Birnbaum, K., Shasha, D. E., Wang, J. Y., Jung, J. W., Lambert, G. M., Galbraith, D. W., & Benfey, P. N. (2003). A gene expression map of the Arabidopsis root. *Science*, *302*, 1956–1960. https://doi.org/10.1126/science.1090022

Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, *19*, 185–193. https://doi.org/10.1093/bioinformatics/19.2.185

Bowman, J. L., Smyth, D. R., & Meyerowitz, E. M. (1989). Genes directing flower development in Arabidopsis. *Plant Cell*, *1*, 37–52. https://doi.org/10.1105/tpc.1.1.37

Brady, S. M., Orlando, D. A., Lee, J. Y., Wang, J. Y., Koch, J., Dinneny, J. R., … Benfey, P. N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science*, *318*, 801–806. https://doi.org/10.1126/science.1146265

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., … Sansone, S. A. (2003). ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, *31*, 68–71. https://doi.org/10.1093/nar/gkg091

Carlsbecker, A., Lee, J. Y., Roberts, C. J., Dettmer, J., Lehesranta, S., Zhou, J., … Benfey, P. N. (2010). Cell signalling by microRNA165/6 directs gene dose-dependent root cell fate. *Nature*, *465*, 316–321. https://doi.org/10.1038/nature08977

Collin, V., Lamkemeyer, P., Miginiac-Maslow, M., Hirasawa, M., Knaff, D. B., Dietz, K. J., & Issakidis-Bourguet, E. (2004). Characterization of plastidial thioredoxins from Arabidopsis belonging to the new y-type. *Plant Physiology*, *136*, 4088–4095. https://doi.org/10.1104/pp.104.052233

Craigon, D. J., James, N., Okyere, J., Higgins, J., Jotham, J., & May, S. (2004). NASCArrays: A repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Research*, *32*, D575–D577. https://doi.org/10.1093/nar/gkh133

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., … Quake, S. R. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, *29*, 1120–1127. https://doi.org/10.1038/nbt.2038

Dalerba, P., Sahoo, D., & Clarke, M. F. (2016). CDX2 as a Prognostic Biomarker in Colon Cancer. *New England Journal of Medicine*, *374*, 2184.

Dalerba, P., Sahoo, D., Paik, S., Guo, X., Yothers, G., Song, N., … Clarke, M. F. (2016). CDX2 as a Prognostic Biomarker in Stage II and Stage III Colon Cancer. *New England Journal of Medicine*, *374*, 211–222. https://doi.org/10.1056/NEJMoa1506597

Dinneny, J. R., Long, T. A., Wang, J. Y., Jung, J. W., Mace, D., Pointer, S., … Benfey, P. N. (2008). Cell identity mediates the response of Arabidopsis roots to abiotic stress. *Science*, *320*, 942–945. https://doi.org/10.1126/science.1153795

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*, 207–210. https://doi.org/10.1093/nar/30.1.207

Efroni, I., Ip, P. L., Nawy, T., Mello, A., & Birnbaum, K. D. (2015). Quantification of cell identity from single-cell gene expression profiles. *Genome Biology*, *16*, 9. https://doi.org/10.1186/s13059-015-0580-x

Harr, B., & Schlotterer, C. (2006). Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Research*, *34*, e8. https://doi.org/10.1093/nar/gnj010

He, F., & Maslov, S. (2016). Pan- and core- network analysis of co-expression genes in a model plant. *Scientific Reports*, *6*, 38956. https://doi.org/10.1038/srep38956

He, F., Yoo, S., Wang, D., Kumari, S., Gerstein, M., Ware, D., & Maslov, S. (2016). Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis. *Plant Journal*, *86*, 472–480. https://doi.org/10.1111/tpj.13175

Hubbell, E., Liu, W. M., & Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics*, *18*, 1585–1592. https://doi.org/10.1093/bioinformatics/18.12.1585

Inlay, M. A., Bhattacharya, D., Sahoo, D., Serwold, T., Seita, J., Karsunky, H., … Weissman, I. L. (2009). Ly6d marks the earliest stage of B-cell specification and identifies the branchpoint between B-cell and T-cell development. *Genes & Development*, *23*, 2376–2381. https://doi.org/10.1101/gad.1836009

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, *31*, e15. https://doi.org/10.1093/nar/gng015

Iyer-Pascuzzi, A. S., Jackson, T., Cui, H., Petricka, J. J., Busch, W., Tsukagoshi, H., & Benfey, P. N. (2011). Cell identity regulators link development and stress responses in the Arabidopsis root. *Developmental Cell*, *21*, 770–782. https://doi.org/10.1016/j.devcel.2011.09.009

Jakoby, M., Wang, H. Y., Reidt, W., Weisshaar, B., & Bauer, P. (2004). FRU (BHLH029) is required for induction of iron mobilization genes in *Arabidopsis thaliana*. *FEBS Letters*, *577*, 528–534. https://doi.org/10.1016/j.febslet.2004.10.062

Katari, M. S., Nowicki, S. D., Aceituno, F. F., Nero, D., Kelfer, J., Thompson, L. P., … Gutierrez, R. A. (2010). VirtualPlant: A software platform to support systems biology research. *Plant Physiology*, *152*, 500–515. https://doi.org/10.1104/pp.109.147025

Kim, H. U., Hsieh, K., Ratnayake, C., & Huang, A. H. (2002). A novel group of oleosins is present inside the pollen of Arabidopsis. *Journal of Biological Chemistry*, *277*, 22677–22684. https://doi.org/10.1074/jbc.M109298200

Kramer, E. M., & Irish, V. F. (1999). Evolution of genetic mechanisms controlling petal development. *Nature*, *399*, 144–148. https://doi.org/10.1038/20172

Lee, J. Y., Colinas, J., Wang, J. Y., Mace, D., Ohler, U., & Benfey, P. N. (2006). Transcriptional and posttranscriptional regulation of transcription factor expression in Arabidopsis roots. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 6055–6060. https://doi.org/10.1073/pnas.0510607103

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323. https://doi.org/10.1186/1471-2105-12-323

Li, Y., Liu, T., Duan, W., Song, X., Shi, G., Zhang, J., … Hou, X. (2014). Instability in mitochondrial membranes in Polima cytoplasmic male sterility of Brassica rapa ssp. chinensis. *Functional & Integrative Genomics*, *14*, 441–451. https://doi.org/10.1007/s10142-014-0368-1

Liesecke, F., Daudu, D., Duge de Bernonville, R., Besseau, S., Clastre, M., Courdavault, V., … Duge de Bernonville, T. (2018). Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Scientific Reports*, *8*, 10885. https://doi.org/10.1038/s41598-018-29077-3

Lim, W. K., Wang, K., Lefebvre, C., & Califano, A. (2007). Comparative analysis of microarray normalization procedures: Effects on reverse engineering gene networks. *Bioinformatics*, *23*, i282–i288. https://doi.org/10.1093/bioinformatics/btm201

Long, T. A., Tsukagoshi, H., Busch, W., Lahner, B., Salt, D. E., & Benfey, P. N. (2010). The bHLH transcription factor POPEYE regulates response to iron deficiency in Arabidopsis roots. *Plant Cell*, *22*, 2219–2236. https://doi.org/10.1105/tpc.110.074096

Lukk, M., Kapushesky, M., Nikkila, J., Parkinson, H., Goncalves, A., Huber, W., … Brazma, A. (2010). A global map of human gene expression. *Nature Biotechnology*, *28*, 322–324. https://doi.org/10.1038/nbt0410-322

Manfield, I. W., Jen, C. H., Pinney, J. W., Michalopoulos, I., Bradford, J. R., Gilmartin, P. M., & Westhead, D. R. (2006). Arabidopsis Co-expression Tool (ACT): Web server tools for microarray-based gene expression analysis. *Nucleic Acids Research*, *34*, W504–W509. https://doi.org/10.1093/nar/gkl204

Montojo, J., Zuberi, K., Rodriguez, H., Kazi, F., Wright, G., Donaldson, S. L., … Bader, G. D. (2010). GeneMANIA Cytoscape plugin: Fast gene function predictions on the desktop. *Bioinformatics*, *26*, 2927–2928. https://doi.org/10.1093/bioinformatics/btq562

Mutwil, M., Obro, J., Willats, W. G., & Persson, S. (2008). GeneCAT–novel webtools that combine BLAST and co-expression analyses. *Nucleic Acids Research*, *36*, W320–W326. https://doi.org/10.1093/nar/gkn292

Nawy, T., Lee, J. Y., Colinas, J., Wang, J. Y., Thongrod, S. C., Malamy, J. E., … Benfey, P. N. (2005). Transcriptional profile of the Arabidopsis root quiescent center. *Plant Cell*, *17*, 1908–1925. https://doi.org/10.1105/tpc.105.031724

Nuccio, M. L., & Thomas, T. L. (1999). ATS1 and ATS3: Two novel embryo-specific genes in Arabidopsis thaliana. *Plant Molecular Biology*, *39*, 1153–1163. https://doi.org/10.1023/A:1006101404867

Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., & Kinoshita, K. (2018). ATTED-II in 2018: A plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant and Cell Physiology*, *59*, e3. https://doi.org/10.1093/pcp/pcx191

Obayashi, T., & Yano, K. (2014). Plant and cell physiology 2014 online database issue. *Plant and Cell Physiology*, *55*, 1–2. https://doi.org/10.1093/pcp/pct193

Pachter, L. (2011). *Models for transcript quantification from RNA-Seq*. arXiv e-prints [Online]. Retrieved from https://ui.adsabs.harvard.edu/\#abs/2011arXiv1104.3889P. Accessed April 01, 2011.

Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., … Zhang, P. (2003). The Arabidopsis Information Resource (TAIR): A model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research*, *31*, 224–228. https://doi.org/10.1093/nar/gkg076

Rocca-Serra, P., Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Contrino, S., … Sansone, S. A. (2003). ArrayExpress: A public database of gene expression data at EBI. *Comptes Rendus Biologies*, *326*, 1075–1078. https://doi.org/10.1016/j.crvi.2003.09.026

Sahoo, D. (2012). The power of boolean implication networks. *Frontiers in Physiology*, *3*, 276.

Sahoo, D., Dill, D. L., Gentles, A. J., Tibshirani, R., & Plevritis, S. K. (2008). Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biology*, *9*, R157. https://doi.org/10.1186/gb-2008-9-10-r157

Sahoo, D., Dill, D. L., Tibshirani, R., & Plevritis, S. K. (2007). Extracting binary signals from microarray time-course data. *Nucleic Acids Research*, *35*, 3705–3712. https://doi.org/10.1093/nar/gkm284

Sahoo, D., Seita, J., Bhattacharya, D., Inlay, M. A., Weissman, I. L., Plevritis, S. K., & Dill, D. L. (2010). MiDReG: A method of mining developmentally regulated genes using Boolean implications. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 5732–5737. https://doi.org/10.1073/pnas.0913635107

Sakai, T., Kagawa, T., Kasahara, M., Swartz, T. E., Christie, J. M., Briggs, W. R., … Okada, K. (2001). Arabidopsis nph1 and npl1: Blue light receptors that mediate both phototropism and chloroplast relocation. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 6969–6974. https://doi.org/10.1073/pnas.101137598

Schmid, P. R., Palmer, N. P., Kohane, I. S., & Berger, B. (2012). Making sense out of massive data by going beyond differential expression. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 5594–5599. https://doi.org/10.1073/pnas.1118792109

Sozzani, R., Cui, H., Moreno-Risueno, M. A., Busch, W., van Norman, J. M., Vernoux, T., … Benfey, P. N. (2010). Spatiotemporal regulation of cell-cycle genes by SHORTROOT links patterning and growth. *Nature*, *466*, 128–132. https://doi.org/10.1038/nature09143

Srinivasasainagendra, V., Page, G. P., Mehta, T., Coulibaly, I., & Loraine, A. E. (2008). CressExpress: A tool for large-scale mining of expression data from Arabidopsis. *Plant Physiology*, *147*, 1004–1016. https://doi.org/10.1104/pp.107.115535

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., … Huala, E. (2008). The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Research*, *36*, D1009–D1014.

Tobias, C. M., Howlett, B., & Nasrallah, J. B. (1992). An *Arabidopsis thaliana* gene with sequence similarity to the S-Locus receptor kinase of Brassica oleracea: Sequence and expression. *Plant Physiology*, *99*, 284–290. https://doi.org/10.1104/pp.99.1.284

Toufighi, K., Brady, S. M., Austin, R., Ly, E., & Provart, N. J. (2005). The Botany Array Resource: e-Northerns, expression angling, and promoter analyses. *Plant Journal*, *43*, 153–163. https://doi.org/10.1111/j.1365-313X.2005.02437.x

Van Bel, M., & Coppens, F. (2017). Exploring plant co-expression and gene-gene interactions with CORNET 3.0. *Methods in Molecular Biology*, *1533*, 201–212. https://doi.org/10.1007/978-1-4939-6658-5

Volkmer, J. P., Sahoo, D., Chin, R. K., Ho, P. L., Tang, C., Kurtova, A. V., … Chan, K. S. (2012). Three differentiation states risk-stratify bladder cancer into distinct subtypes. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 2078–2083. https://doi.org/10.1073/pnas.1120605109

Yadav, R. K., Girke, T., Pasala, S., Xie, M., & Reddy, G. V. (2009). Gene expression map of the Arabidopsis shoot apical meristem stem cell niche. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 4941–4946. https://doi.org/10.1073/pnas.0900843106

Yadav, R. K., Tavakkoli, M., Xie, M., Girke, T., & Reddy, G. V. (2014). A high-resolution gene expression map of the Arabidopsis shoot meristem stem cell niche. *Development*, *141*, 2735–2744. https://doi.org/10.1242/dev.106104

Zheng, H. Q., Wu, N. Y., Chow, C. N., Tseng, K. C., Chien, C. H., Hung, Y. C., … Chang, W. C. (2017). EXPath tool-a system for comprehensively analyzing regulatory pathways and coexpression networks from high-throughput transcriptome data. *DNA Research*, *24*, 371–375. https://doi.org/10.1093/dnares/dsx009

Zhuo, B., Emerson, S., Chang, J. H., & Di, Y. (2016). Identifying stably expressed genes from multiple RNA-Seq data sets. *PeerJ*, *4*, e2791. https://doi.org/10.7717/peerj.2791

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., & Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiology*, *136*, 2621–2632. https://doi.org/10.1104/pp.104.046367

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Pandey S, Sahoo D. Identification of gene expression logical invariants in *Arabidopsis*. *Plant Direct*. 2019;3:1–9. https://doi.org/10.1002/pld3.123