RESEARCH ARTICLE

SeqEnrich: A tool to predict transcription factor networks from co-expressed Arabidopsis and *Brassica napus* gene sets

Michael G. Becker[1], Philip L. Walker[1], Nadège C. Pulgar-Vidal[2], Mark F. Belmonte[1]*

1 Department of Biological Sciences, University of Manitoba, Winnipeg, Manitoba, Canada, **2** Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada

* Mark.Belmonte@umanitoba.ca

OPEN ACCESS

Citation: Becker MG, Walker PL, Pulgar-Vidal NC, Belmonte MF (2017) SeqEnrich: A tool to predict transcription factor networks from co-expressed Arabidopsis and *Brassica napus* gene sets. PLoS ONE 12(6): e0178256. https://doi.org/10.1371/journal.pone.0178256

Editor: Martina Stromvik, McGill University, CANADA

Received: January 6, 2017

Accepted: May 10, 2017

Published: June 2, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by Discovery and Accelerator Program grants from the National Science and Engineering Research Council of Canada to MFB, a Vanier Scholarship to MGB and USRA to NCPV.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Transcription factors and their associated DNA binding sites are key regulatory elements of cellular differentiation, development, and environmental response. New tools that predict transcriptional regulation of biological processes are valuable to researchers studying both model and emerging-model plant systems. SeqEnrich predicts transcription factor networks from co-expressed Arabidopsis or *Brassica napus* gene sets. The networks produced by SeqEnrich are supported by existing literature and predicted transcription factor–DNA interactions that can be functionally validated at the laboratory bench. The program functions with gene sets of varying sizes and derived from diverse tissues and environmental treatments. SeqEnrich presents as a powerful predictive framework for the analysis of Arabidopsis and *Brassica napus* co-expression data, and is designed so that researchers at all levels can easily access and interpret predicted transcriptional circuits. The program outperformed its ancestral program ChipEnrich, and produced detailed transcription factor networks from Arabidopsis and *Brassica napus* gene expression data. The SeqEnrich program is ideal for generating new hypotheses and distilling biological information from large-scale expression data.

Introduction

Advances in next generation RNA sequencing (RNA-seq) technologies to investigate biological processes is becoming commonplace in research laboratories. Despite these advances, the analyses of large scale RNA-seq datasets are time consuming and publicly available tools able to analyze this information are often difficult to navigate for researchers with limited bioinformatics experience. Deriving relevant biological information from large-scale datasets represents a significant bottleneck in sequencing experiments. Thus, new user-friendly tools that facilitate analyses of large datasets are in demand from scientists studying both model and emerging-model plant systems.

A major objective in gene expression analyses is the identification of transcription factors (TFs) and cis-regulatory elements that direct cellular bioprocesses at the cellular, tissue, or whole plant level or in response to biotic or abiotic stresses. Many of the currently available tools, including Grassius Regulatory Grid Explorer hosted by the Arabidopsis Gene Regulatory Information Server (AGRIS; [1]) and HRGRN [2], are ideal for individual gene lookups but are not conducive to large datasets. Additionally, several tools can effectively identify enriched promoter motifs from Arabidopsis datasets, such as the motif analysis tool hosted by The Arabidopsis Information Server (TAIR; [3]), the Cistome tool at the Toronto the Bio-Analytic Resource (BAR; [4]), and Arabidopsis Motif Scanner [5]; however, these programs are not designed to identify enriched biological processes or generate TF networks. The Arabidopsis Interactions Viewer hosted by the Toronto BAR identifies experimentally validated protein-DNA interactions and generates a transcription network from an input query list [6]. This tool can build on this network by adding predicted or validated protein-protein interactions. Ultimately, this provides detailed and accurate information on direct interactions between genes within a dataset, but reliance on experimentally validated protein-DNA interactions limits discovery. Further, the Arabidopsis Interactions Viewer does not identify enriched cis-regulatory elements or relate networks back to biological function. Further, none of the programs above take input data derived from emerging model species *Brassica napus* (canola).

ChipEnrich was developed to perform promoter analysis, TF-DNA binding prediction, and functional enrichment in a single Java-based tool [7,8]. Developed to analyze large-scale co-expressed gene sets from the Arabidopsis ATH1 microarray, ChipEnrich takes a list of genes identified through clustering or differential gene expression analyses and identifies enriched Gene Ontology (GO) terms, TF families, and TF binding site motifs. The program also offers an analysis function that associates these terms into a TF network thus providing a predictive framework into the transcriptional programs of co-expressed gene sets. While this program has been used successfully in the analysis of large-scale microarray data [8–10], it is limited to ATH1 GeneChip sequences of the model plant Arabidopsis and publicly available information on i) gene function ii) nucleotide sequence of DNA binding sites and iii) TF-DNA sequence motif interactions.

Here, we developed SeqEnrich (S1 File) based on the ChipEnrich platform to predict TF networks from next generation RNA-seq datasets. The SeqEnrich program contains the most extensive database of TFs, TF-DNA sequence motif interactions, and gene function(s) for the efficient interrogation of Arabidopsis or *B*. *napus* gene sets. SeqEnrich was able to successfully predict TF networks supported by existing experimental data in addition to providing new insights into the underlying transcriptional circuitry controlling biological process in space and time. This program demonstrates a substantial improvement when compared to its ancestor ChipEnrich. SeqEnrich complements existing tools available for Arabidopsis, merging the functions of multiple tools into a single user-friendly program. Additionally, this serves as the first resource able to produce transcription factor networks from *B*. *napus* datasets.

Materials and methods

Program execution

The SeqEnrich software was developed using Java language and designed to be fully compatible with Windows and Linux operating systems. SeqEnrich is a user-friendly application that takes an input text format (.txt) query list of Arabidopsis Genome Initiative [11] identifiers or *B*. *napus* annotation v5 identifiers [12]. Input gene lists are developed through differential gene expression or clustering analyses performed following next generation RNA-seq experiments.

The SeqEnrich jar file (S1 File) and SeqEnrich source code (S2 File) are freely available. Updated versions of the SeqEnrich program will be deposited as they become available at http://www.belmontelab.com and updated source code deposited at the SourceForge open-source repository (https://sourceforge.net/).

GO term enrichment

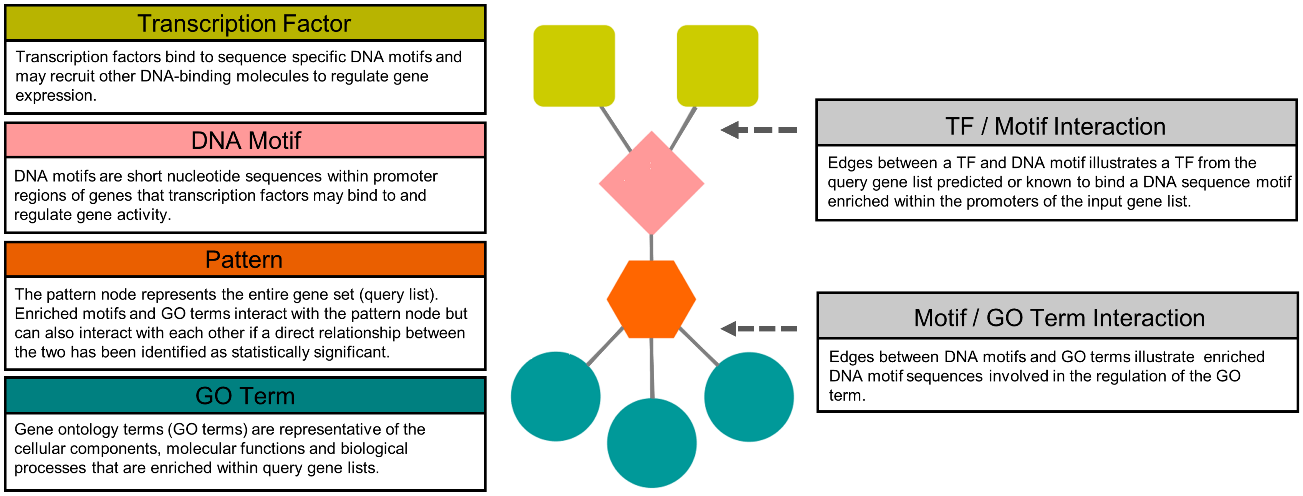
GO term enrichment was defined as the ratio between (a) the number of genes in the query list belonging to the GO term and (b) the total number of genes belonging to the GO term within the genome compared to the ratio of (c) the total number of genes within the query list to (d) the total number of genes within the genome. Significantly enriched GO terms are reported as p-values calculated from the hypergeometric distribution using the Apache Commons Math Library (http://jakarta.apache.org/commons/math). SeqEnrich reports all GO term enrichment data and produces a summary file containing significantly enriched data with a p-value of < 0.001.

DNA sequence motif enrichment

For each gene within the input query list, known TF binding sites are called from a lookup array within the program database. This array was created by identifying known TF binding sites within all gene promoters in the *B*. *napus* and Arabidopsis genomes. Here, we focus on motif enrichment within the 1 kb upstream region from the transcription start site, capturing the majority of Arabidopsis TF binding sites [13]. Background motif distributions are determined from the TF binding sites identified within all promoters across the genome. Statistical enrichment for each DNA sequence motif within the query list was determined using the hypergeometric distribution. In addition, a subanalysis identifies DNA sequence motifs enriched within promoters of query genes associated with each enriched GO term. SeqEnrich reports all motif enrichment data and produces a summary file containing significantly enriched motifs with a p-value of < 0.001.

Prediction of transcription factor networks

Fig 1 shows a conceptual TF network generated using the SeqEnrich program. To produce TF networks, the SeqEnrich program performs an analysis and subanalysis of the input query list.



Transcription Factor

Transcription factors bind to sequence specific DNA motifs and may recruit other DNA-binding molecules to regulate gene expression.

DNA Motif

DNA motifs are short nucleotide sequences within promoter regions of genes that transcription factors may bind to and regulate gene activity.

Pattern

The pattern node represents the entire gene set (query list). Enriched motifs and GO terms interact with the pattern node but can also interact with each other if a direct relationship between the two has been identified as statistically significant.

GO Term

Gene ontology terms (GO terms) are representative of the cellular components, molecular functions and biological processes that are enriched within query gene lists.

TF / Motif Interaction

Edges between a TF and DNA motif illustrates a TF from the query gene list predicted or known to bind a DNA sequence motif enriched within the promoters of the input gene list.

Motif / GO Term Interaction

Edges between DNA motifs and GO terms illustrate enriched DNA motif sequences involved in the regulation of the GO term.

Fig 1. Conceptual description of a transcription factor network. TFs are represented as green rounded squares, DNA motifs as pink diamonds, gene patterns as orange hexagons, and gene ontology (GO) terms as blue circles. Connections between TFs and motifs, and between motifs and patterns/GO terms are represented by a grey connecting line.

https://doi.org/10.1371/journal.pone.0178256.g001

In the analysis, DNA sequence motifs significantly ($p < 0.001$) enriched in promoters of query genes are associated with TFs within the same query gene list capable of binding to that DNA sequence motif. In the subanalysis, DNA sequence motifs significantly ($p < 0.001$) enriched in promoters of query genes of individual significantly enriched GO terms ($p < 0.001$) are associated with TFs within the query gene list capable of binding to that sequence motif. Together, these analyses produce connections between TFs, enriched DNA motifs, query genes (gene pattern), and GO terms and is presented in output network files. SeqEnrich produces two network files: i) TF networks from the analysis identifying DNA sequence motifs and GO terms enriched within the entire query list, and the associated TFs (extension.analysis.networks.txt) and ii) TF networks from the subanalysis identifying motifs enriched within each individual enriched GO term, and associated TFs (extension.subanalysis.networks.txt). Output network files are compatible with the Cytoscape visualization tool (http://www.cytoscape.org/), and p-values imported as edge attributes. A separate attributes file is produced and labels each node presented in the network file as a "pattern", "GO term", "motif", or "TF".

Database construction

The database for the SeqEnrich program was built using information from publicly available sources (Fig 2). The GO term database and genome annotations were derived from data curated by the TAIR consortium (TAIR10, https://www.arabidopsis.org) and *B*. *napus* annotation v5 [12]. Arabidopsis and *B*. *napus* TFs were identified using information from several databases to ensure inclusion of all TFs: i) JASPAR (http://jaspar.genereg.net/) ii) Plant TFDB (http://planttfdb.cbi.pku.edu.cn/) iii) Plant Transcription Factor Database (http://plntfdb.bio.uni-potsdam.de/) and iv) Database of Arabidopsis Transcription Factors (http://datf.cbi.pku.edu.cn/). A database of experimentally validated TF binding sites was built using information from the following sources: i) JASPAR ii) CIS-BP (http://cisbp.ccbr.utoronto.ca/) iii) Arabidopsis protein binding microarray [14] iv) Yeast-one hybrid experiments [15] and v) Arabidopsis DAP-sequencing experiments [16].

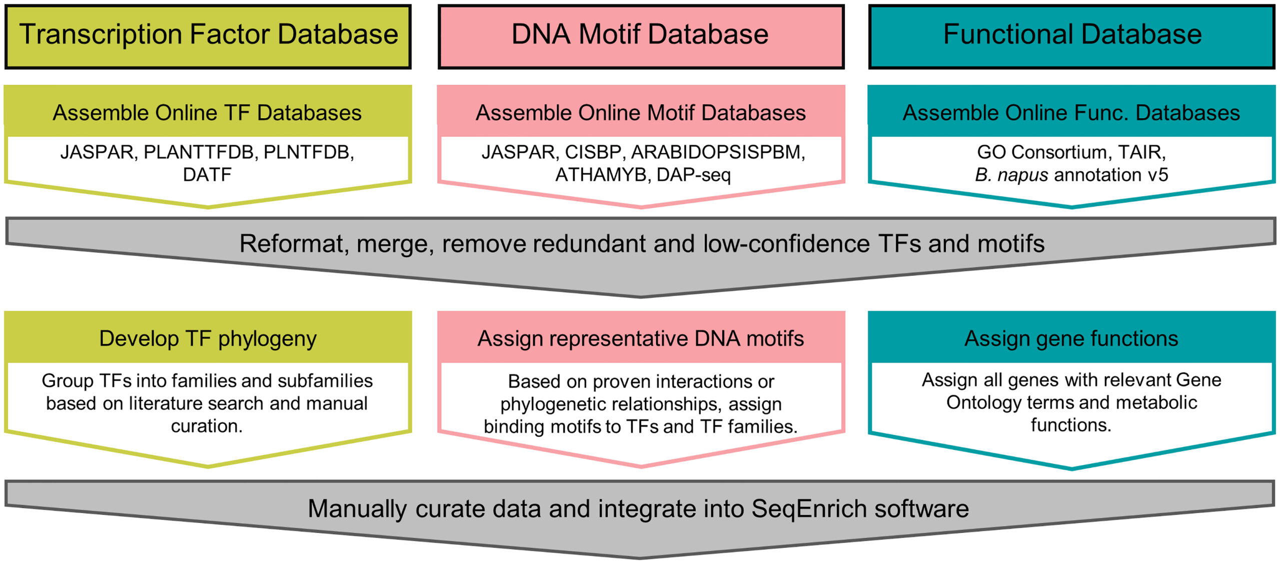

Transcription Factor Database
DNA Motif Database
Functional Database
Assemble Online TF Databases
JASPAR, PLANTTFDB, PLNTFDB, DATF
Assemble Online Motif Databases
JASPAR, CISBP, ARABIDOPSISPBM, ATHAMYB, DAP-seq
Assemble Online Func. Databases
GO Consortium, TAIR, B. napus annotation v5
Reformat, merge, remove redundant and low-confidence TFs and motifs
Develop TF phylogeny
Group TFs into families and subfamilies based on literature search and manual curation.
Assign representative DNA motifs
Based on proven interactions or phylogenetic relationships, assign binding motifs to TFs and TF families.
Assign gene functions
Assign all genes with relevant Gene Ontology terms and metabolic functions.
Manually curate data and integrate into SeqEnrich software

Fig 2. Design and assembly of the SeqEnrich database. Information on transcription factors, DNA binding site motifs, and gene functions were collected from publically available sources and integrated into the SeqEnrich program.

https://doi.org/10.1371/journal.pone.0178256.g002

Table 1. IUPAC codes used for representation of nucleotides in motifs and corresponding likelihood of each nucleotide at position.

IUPAC code	IUPAC identity	Nucleotide probabilities
A	Adenine	$[p(A) \geq 0.1, p(T) < 0.1, p(C) < 0.1, p(G) < 0.1]$
C	Cytosine	$[p(A) < 0.1, p(T) < 0.1, p(C) \geq 0.1, p(G) < 0.1]$
G	Guanine	$[p(A) < 0.1, p(T) < 0.1, p(C) < 0.1, p(G) \geq 0.1]$
T	Thymine	$[p(A) < 0.1, p(T) \geq 0.1, p(C) < 0.1, p(G) < 0.1]$
R	A or G	$[p(A) \geq 0.1, p(T) < 0.1, p(C) < 0.1, p(G) \geq 0.1]$
Y	C or T	$[p(A) < 0.1, p(T) \geq 0.1, p(C) \geq 0.1, p(G) < 0.1]$
S	G or C	$[p(A) < 0.1, p(T) < 0.1, p(C) \geq 0.1, p(G) \geq 0.1]$
W	A or T	$[p(A) \geq 0.1, p(T) \geq 0.1, p(C) < 0.1, p(G) < 0.1]$
K	G or T	$[p(A) < 0.1, p(T) \geq 0.1, p(C) < 0.1, p(G) \geq 0.1]$
M	A or C	$[p(A) \geq 0.1, p(T) < 0.1, p(C) \geq 0.1, p(G) < 0.1]$
B	C or G or T	$[p(A) < 0.1, p(T) \geq 0.1, p(C) \geq 0.1, p(G) \geq 0.1]$
D	A or G or T	$[p(A) \geq 0.1, p(T) \geq 0.1, p(C) < 0.1, p(G) \geq 0.1]$
H	A or C or T	$[p(A) \geq 0.1, p(T) \geq 0.1, p(C) \geq 0.1, p(G) < 0.1]$
V	A or C or G	$[p(A) \geq 0.1, p(T) < 0.1, p(C) \geq 0.1, p(G) \geq 0.1]$

https://doi.org/10.1371/journal.pone.0178256.t001

All motif databases presenting motifs in MEME format were converted into degenerate IUPAC codes using a custom Java script (S3 File) to ensure consistency and clarity amongst motifs. If the probability of any given nucleotide at each position within a motif is <0.10, it is not recorded at that position during generation of the IUPAC code (Table 1). This probability cut-off provided an ideal balance between specificity and sensitivity. Lower thresholds that were tested generated overly general motifs, and higher thresholds eliminated putative binding sites.

Following amalgamation of motif databases, data filtering was used to remove duplicate entries and non-informative motifs. A non-informative motif was defined as a motif that lacked sequence specificity and had a $p < 0.02$ of occurring at any individual locus within a promoter by chance. To improve the resolution and accuracy of SeqEnrich, TF phylogenetic relationships were used to predict TF-DNA sequence motif interactions. For example, TFs with no experimentally validated binding information or TFs with conflicting experimental data could be assigned putative DNA binding site motifs based on information from closely related TFs. These TF phylogenetic relationships were determined by combining existing TFF phylogenetic footprinting data [15,17–26] and protein clustering performed with MUSCLE v3.8.31 (http://www.drive5.com/muscle/muscle.html). Using TF phylogenetic relationships to predict TF binding sites represents a new feature of SeqEnrich and dramatically improves the number of TFs with known or predicted binding sites, with 902 new predicted TF-DNA sequence motif interactions as compared to the original ChipEnrich program.

Results and discussion

The SeqEnrich database

The SeqEnrich database contains a total of 2,263 Arabidopsis TFs and 240 Arabidopsis degenerate DNA binding site motifs, representing a 44.8% and 135% increase respectively, as compared to the original ChipEnrich software. As an additional comparison, the AGRIS databases [1] currently contain 1,773 TFs and 99 DNA binding site motifs (http://arabidopsis.med.ohio-state.edu). The majority (85.4%) of TFs in the AGRIS transcription factor database overlapped with SeqEnrich. For the remaining ~15%, many were incorrectly classified as C3H family TFs due to their zinc finger, RING-type domains. Experimental evidence has suggested that they are more likely involved in ubiquitination [27,28]. The newly developed *B*. *napus* database contains 8,306 TFs and 228 degenerate motifs. The greater number of TFs in *B*. *napus* can be explained by a genome duplication event since its divergence from Arabidopsis and its large allotetraploid genome of ~90,000 genes [12]. The decline in the number of binding site motifs identified in *B*. *napus* is likely due to its comparative lack of experimental TF binding data.

Since little information about *B*. *napus* TF binding sites are available, the majority of *B*. *napus* binding site motifs are predicted from closely related Arabidopsis homologs. Using homology to predict TF binding sites is common in mammalian systems [29,30], and is supported by multiple independent experiments that identify conserved TF target sequences between plant species as distally related as Arabidopsis and rice [21,22,31]. Our SeqEnrich program represents the largest database for *B*. *napus* TFs and motifs currently available, and is accessible within the program source code (S2 File) and as a separate download at http://www.belmontelab.com.

Validation of the SeqEnrich program

To demonstrate how the SeqEnrich program improves detection of TF-DNA sequence motif interactions, we compared its output to previously published Arabidopsis TF networks from publicly available datasets [8,32]. A total of 661 genes co-expressed in the chalazal endosperm of the mature seed (S4 File) were used as input into both the ChipEnrich (Fig 3A) and SeqEnrich programs (Fig 3B). This comparison identified 2.6-fold more TFs, 7.1-fold more enriched GO terms, and 9.3-fold more enriched motifs using the SeqEnrich program with the same gene query list. As the SeqEnrich database contains only 44.8% more motifs than ChipEnrich, the observed increase in motif detection is likely due to the improved accuracy of new databases and not solely a result of increase in database size. Similarly, the larger number of enriched GO terms in SeqEnrich can be attributed to the improvements in TAIR curated GO annotations.

We also identified predictive TF networks in mature chalazal endosperm using the subanalysis function of SeqEnrich (Fig 3C). This transcriptional circuit predicts the regulation of a number of seed-related bioprocesses, including “nutrient reservoir activity”, “embryo development to seed dormancy”, “lipid storage”, “response to abscisic acid”, and “seed dormancy” by TFs binding to the BZIP, MYB, BES, and BHLH binding sites. As the subanalysis function is a new function in the program a direct comparison between subanalyses is not possible; however, several of the DNA binding sites identified, including PHL2, BBR-BPC, BES, HHO are found exclusively within the SeqEnrich database. Several of these TFs identified in this circuit, including SPATULA and BZR1, are already associated with seed maturation and have delayed development phenotypes [33,34]. Other TFs identified in the subanalysis, including PHL2, CIB4, and BPC6, have no known biological function. Based on the late timing of their expression, they may serve as transcriptional regulators of seed maturation active in tissue-specific subregions of the seed.

To confirm that the SeqEnrich output is reflective of *in vivo* gene regulation, we tested the program on a dataset containing TFs and DNA targets with proven interactions. We used the Taylor-Teeples et al. (2015) dataset of coexpressed genes involved in secondary cell wall biosynthesis of the Arabidopsis steele, which includes TF-motif interactions confirmed with yeast one hybrid experiments [32]. We produced a network with these data containing 58 TFs targeting genes associated with cell wall, xylem, lignin, cellulose and hemicellulose, root development, and regulation of transcription (S1 Fig). Our analysis identified enriched bZIP, homeobox domain (HD), and zinc-finger HD transcription factor families, which agrees with the findings in Taylor-Teeples et al. (2015). We also identified TFs PHABULOSA and

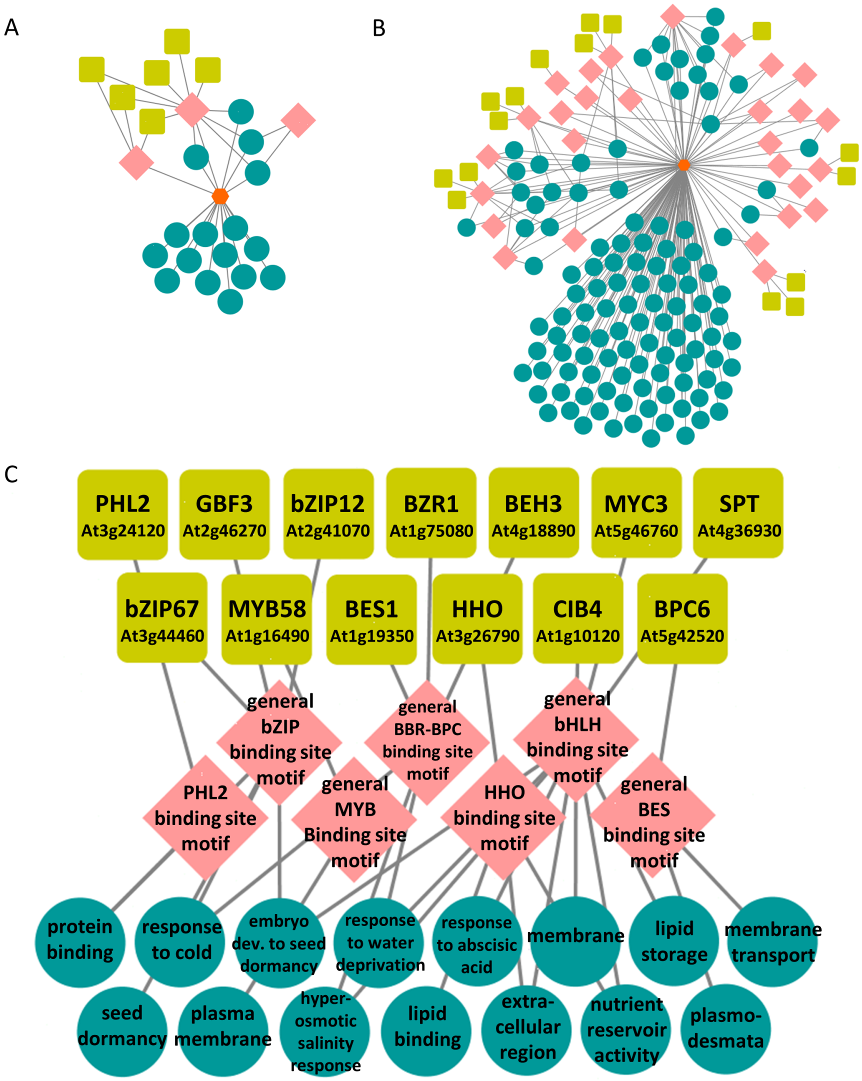


Fig 3. Predicted transcription factor networks from the chalazal endosperm of Arabidopsis. A) Predicted transcriptional module developed from the ChipEnrich program; B) predicted transcriptional module using the SeqEnrich program; C) subset of the transcriptional module produced from the subanalysis function of the SeqEnrich program. A predicted bZIP, bHLH, MYB, and BES transcriptional module controlling biological processes within the mature endosperm of Arabidopsis.

https://doi.org/10.1371/journal.pone.0178256.g003

REVOLUTA in our networks, which were a center of focus in Taylor-Teeples et al. (2015) and proven central regulators of xylem cell specification and secondary cell wall biosynthesis. Together, these data support the ability of SeqEnrich to report meaningful biological predictions.

SeqEnrich predicts transcription factor networks in *Brassica napus*

We used two different datasets (S4 File) to test the efficacy of the SeqEnrich program on large-scale *B. napus* RNA-seq co-expressed gene lists: i) a tissue-specific dataset of 684 genes activated within the vasculature of the *B. napus* funiculus [35] (Fig 4A); and ii) a defense response dataset of 3,234 genes activated specifically in cotyledons resistant to the fungal pathogen *Leptosphaeria maculans* [36] (Fig 4B–4C). These datasets were selected because of i) the

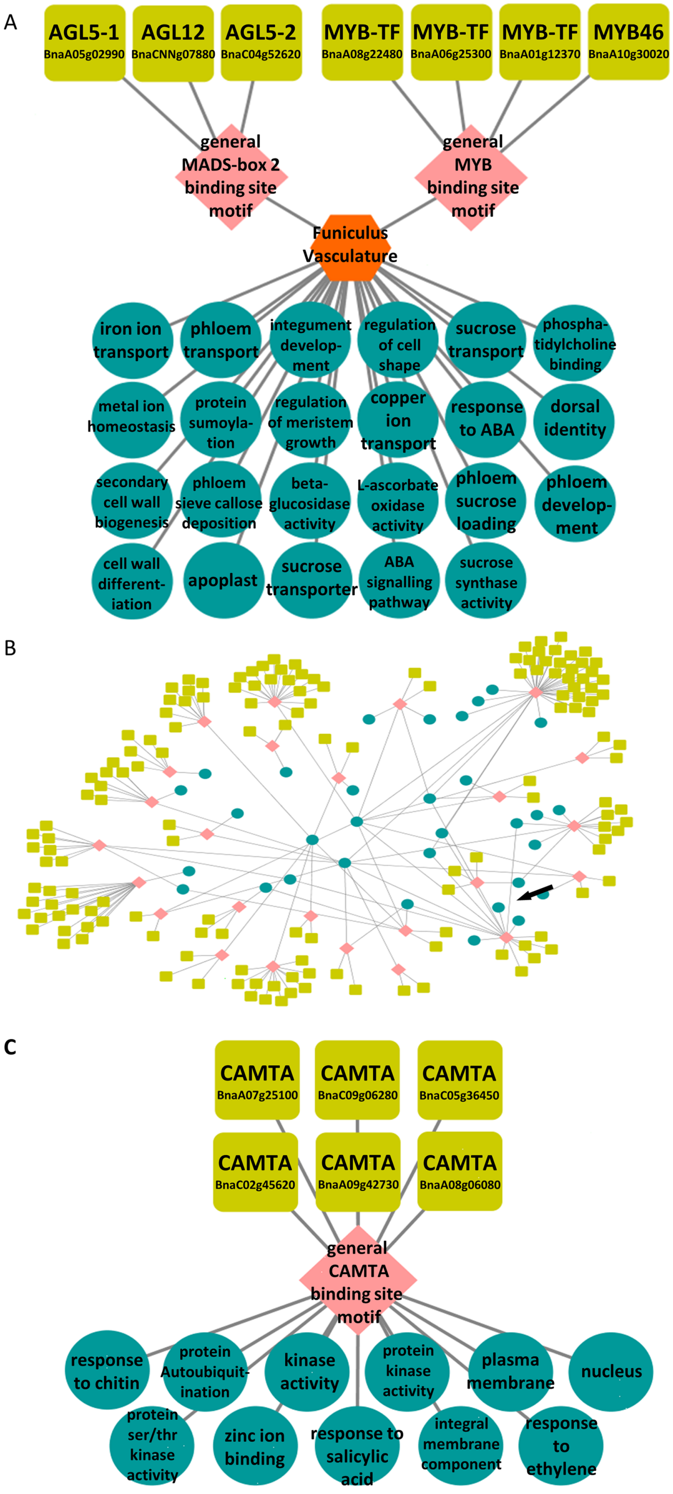


Fig 4. Predicted transcription factor networks from *Brassica napus* gene sets. A) Subset of transcription factor network identified from funiculus vasculature dataset. MYB and MADS-box TFs are predicted to regulate genes associated with transport, metal ion homeostasis, and cell wall modification; B) subset of transcriptional module identified from the SeqEnrich subanalysis function predicted to be operative in seedlings infected with fungal pathogen *Leptosphaeria maculans*; C) transcriptional module depicted by arrow in (B), showing regulation of genes associated with defense bioprocesses by a family of calmodulin-binding transcriptional activators (CAMTAs).

https://doi.org/10.1371/journal.pone.0178256.g004

differences in the size of the query list, ii) the tissue type from where RNA sequencing was performed, and iii) the environmental conditions in which the plants were reared based on their cellular response.

The newly developed SeqEnrich program identified TFs binding to the MYB and MADS-box binding site motifs enriched within the promoters of genes specific to funiculus vasculature, and are associated with "phloem sieve callose deposition", "secondary cell wall biogenesis", "phloem sucrose loading", "integument development", and "metal ion homeostasis" (Fig 4A). As the only connection between maternal tissue and the growing seed, the funiculus is responsible for nutrient transfer during seed filling [35]. Two of the TFs identified in this transcriptional circuit have proven roles in plant root tissue [37,38], which may be due to the similar functions of root and funicular tissues in nutrient transport [39]. For example, MYB46 is concentrated within the xylem of growing roots and a master regulator of secondary cell wall formation [37]. The identification of MYB46 in funicular vasculature (Fig 4A) clearly demonstrates the ability of SeqEnrich to generate new scientific hypotheses that can be functionally validated at the bench.

The SeqEnrich subanalysis predicted a large transcriptional network active in resistant *B. napus* cotyledons, suggesting that coordination of the defense response is complex and involves large numbers of TFs to control expression of genes responsible for plant defense and highlights the sensitivity of the program (Fig 4B). Most TFs identified converged on similar binding site motifs. For example, 30 WRKY TFs were identified in a transcriptional module and are predicted to bind to a similar WRKY binding site motif. The large number of cognate TFs is likely due to gene duplication in the *B. napus* genome and redundant TF functions. A subsection of this network is presented in Fig 4C and predicts a family of calmodulin binding transcriptional activators (CAMTAs) binding to the CAMTA binding site motif enriched in gene promoters associated with "response to chitin", "kinase activity", "response to salicylic acid", and "response to ethylene" during defense. As calmodulin has been shown to accumulate in resistant plants upon pathogen exposure [40], the activity of calmodulin-binding TFs is logical and provides additional evidence into the predictive framework of the SeqEnrich analysis using *B. napus* data. As the binding sites of these CAMTAs were unknown until recent DAP-seq experiments [16] this clearly demonstrates the sensitivity of our newly developed program afforded by its carefully constructed DNA sequence motif databases.

Program utility

Tools to identify enriched GO terms [41] or DNA sequence motifs [5,42–44] from Arabidopsis gene sets are available, however, none exist that combine these functions into a complete analysis that identifies TF-DNA sequence interactions within biological processes encoded by gene sets of RNA-seq data. Here, SeqEnrich identifies enriched DNA sequence motifs within the promoters of query gene sets or a smaller subset of genes belonging to a biological process with predicted or known TF-DNA sequence motif interactions. Further, the SeqEnrich program interrogates current databases of GO terms, DNA sequence motifs, and TFs thus reducing manual data curation by the user. Current user-friendly programs that facilitate rapid analysis of RNA-seq data and elucidation of regulatory networks are limited to humans and other mammalian systems [45–48]. Our SeqEnrich program serves the plant biology community and is designed for the model organism Arabidopsis and economically important *B. napus*. Given the utility and applications of the program and the continued generation of large scale RNA-seq datasets from model and emerging model plant systems, we are currently developing similar data analysis platforms for globally-important monocot crop systems such as rice and corn. The hypotheses generated from this program can then be transferred to the laboratory bench to functionally characterize the underlying molecular mechanisms of the plant.

Conclusion

The SeqEnrich program represents a powerful user-friendly platform for the analyses of large scale RNA-seq datasets and the prediction of TF networks from differential gene expression or co-expression analyses. We show the program can be applied broadly across query gene lists of different sizes, tissue types, and experimental treatments. The program's ease of use allows researchers without a background in computational biology to perform comprehensive analysis of RNA-seq data and identify TF-DNA motif interactions orchestrating complex biological processes.

Supporting information

S1 File. SeqEnrich is a Java based program and fully compatible with Windows or Linux operating systems. Detailed user instructions are provided with the SeqEnrich file. Updated versions of the SeqEnrich program will be deposited as they become available at http://www.belmontelab.com.
(ZIP)

S2 File. SeqEnrich source code. Updated versions of the SeqEnrich source code will be deposited as they become available at the SourceForge open-source repository (https://sourceforge.net/).
(ZIP)

S3 File. File contains both the Java file and usage information.
(ZIP)

S4 File. Arabidopsis and *B. napus* query lists used as test data for the SeqEnrich program.
(ZIP)

S1 Fig. Transcription factor network from experimentally validated TF-DNA interactions from a root xylem secondary cell wall biosynthesis dataset.
(TIFF)

Author Contributions

Conceptualization: MFB MGB.

Data curation: MGB PLW NCPV.

Formal analysis: MGB PLW NCPV MFB.

Funding acquisition: MFB.

Investigation: MGB PLW NCPV MFB.

Methodology: MGB PLW NCPV MFB.

Project administration: MFB.

Resources: MFB.

Software: MGB NCPV.

Supervision: MFB.

Validation: MGB PLW NCPV MFB.

Visualization: MGB PLW NCPV MFB.

Writing – original draft: MGB PLW NCPV MFB.

Writing – review & editing: MGB PLW NCPV MFB.

References

1. Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L, Grotewold E. AGRIS: The arabidopsis gene regulatory information server, an update. Nucleic Acids Res. 2011; 39: 1118–1122. https://doi.org/10.1093/nar/gkq1120 PMID: 21059685
2. Dai X, Li J, Liu T, Zhao PX. HRGRN: A graph search-empowered integrative database of arabidopsis signaling transduction, metabolism and gene regulation networks. Plant Cell Physiol. 2016; 57: e12. https://doi.org/10.1093/pcp/pcv200 PMID: 26657893
3. Rhee SY. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res. 2003; 31: 224–228. https://doi.org/10.1093/nar/gkg076 PMID: 12519987
4. Austin RS, Hiu S, Waese J, Ierullo M, Pasha A, Wang TT, et al. New BAR tools for mining expression data and exploring Cis-elements in Arabidopsis thaliana. Plant J. 2016; 88: 490–504. https://doi.org/10.1111/tpj.13261 PMID: 27401965
5. Mele G. Arabidopsis Motif Scanner. BMC Bioinformatics. BMC Bioinformatics; 2016; 17: 50. https://doi.org/10.1186/s12859-016-0896-x PMID: 26817596
6. Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar a H, Geisler M. A Predicted Interactome for Arabidopsis. Plant Physiol. 2007; 145: 317–329. https://doi.org/10.1104/pp.107.103465 PMID: 17675552
7. Orlando DA, Brady SM, Koch JD, Dinneny R, Benfey PN. Manipulating Large-Scale Arabidopsis Microarray Expression Data: Identifying Dominant Expression Patterns and Biological Process Enrichment. In: Belostotsky D, editor. Plant Systems Biology. Humana Press; 2009. pp. 57–77.
8. Belmonte MF, Kirkbride RC, Stone SL, Pelletier JM, Bui AQ, Yeung EC, et al. Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. Proc Natl Acad Sci U S A. 2013/01/16. 2013; 110: E435–44. https://doi.org/10.1073/pnas.1222061110 PMID: 23319655
9. Khan D, Millar JL, Girard IJ, Chan A, Kirkbride RC, Pelletier JM, et al. Transcriptome atlas of the Arabidopsis funiculus—a study of maternal seed subregions. Plant J. 2015; 82: 41–53. https://doi.org/10.1111/tpj.12790 PMID: 25684030
10. Silva AT, Ribone PA, Chan RL, Ligterink W, Hilhorst HWM. A Predictive Coexpression Network Identifies Novel Genes Controlling the Seed-to-Seedling Phase Transition in Arabidopsis thaliana 1 [OPEN]. 2016; 170: 2218–2231. https://doi.org/10.1104/pp.15.01704 PMID: 26888061
11. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000; 408: 796–815. https://doi.org/10.1038/35048692 PMID: 11130711
12. Chalhoub B, Denoeud F, Liu S, Parkin I a. P, Tang H, Wang X, et al. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science (80-). 2014; 345: 950–953. https://doi.org/10.1126/science.1253435 PMID: 25146293
13. Yu C-P, Lin J-J, Li W-H. Positional distribution of transcription factor binding sites in Arabidopsis thaliana. Sci Rep. Nature Publishing Group; 2016; 6: 25164. https://doi.org/10.1038/srep25164 PMID: 27117388
14. Franco-Zorrilla JM, López-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R. DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci U S A. 2014; 111: 2367–72. https://doi.org/10.1073/pnas.1316278111 PMID: 24477691
15. Kelemen Z, Sebastian A, Xu W, Grain D, Salsac F, Avon A, et al. Analysis of the DNA-binding activities of the arabidopsis R2R3-MYB transcription factor family by one-hybrid experiments in yeast. PLoS One. 2015; 10: 1–22. https://doi.org/10.1371/journal.pone.0141044 PMID: 26484765
16. O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, et al. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell. Elsevier Inc.; 2016; 165: 1280–1292. https://doi.org/10.1016/j.cell.2016.04.038 PMID: 27203113
17. Stender EG, O'Shea C, Skriver K. Subgroup-specific intrinsic disorder profiles of arabidopsis NAC transcription factors: Identification of functional hotspots. Plant Signal Behav. 2015; 10: e1010967. https://doi.org/10.1080/15592324.2015.1010967 PMID: 26107850

18. Ferrari S, Plotnikova JM, De Lorenzo G, Ausubel FM. Arabidopsis local resistance to Botrytis cinerea involves salicylic acid and camalexin and requires EDS4 and PAD2, but not SID2, EDS5 or PAD4. Plant J. 2003; 35: 193–205. https://doi.org/10.1046/j.1365-313X.2003.01794.x PMID: 12848825

19. Englbrecht CC, Schoof H, Böhm S. Conservation, diversification and expansion of C2H2 zinc finger proteins in the Arabidopsis thaliana genome. BMC Genomics. 2004; 5: 39. https://doi.org/10.1186/1471-2164-5-39 PMID: 15236668

20. Reyes JC, Muro-Pastor MI, Florencio FJ. The GATA family of transcription factors in Arabidopsis and rice. Plant Physiol. 2004; 134: 1718–1732. https://doi.org/10.1104/pp.103.037788 PMID: 15084732

21. Nakano T, Suzuki K, Fujimura T, Shinshi H. Genome-Wide Analysis of the ERF Gene Family. Plant Physiol. 2006; 140: 411–432. https://doi.org/10.1104/pp.105.073783 PMID: 16407444

22. Guo J, Wu J, Ji Q, Wang C, Luo L, Yuan Y, et al. Genome-wide analysis of heat shock transcription factor families in rice and Arabidopsis. J Genet Genomics. 2008; 35: 105–118. https://doi.org/10.1016/S1673-8527(08)60016-8 PMID: 18407058

23. Wang D, Guo Y, Wu C, Yang G, Li Y, Zheng C. Genome-wide analysis of CCCH zinc finger family in Arabidopsis and rice. BMC Genomics. 2008; 9: 44. https://doi.org/10.1186/1471-2164-9-44 PMID: 18221561

24. Kaplan-Levy RN, Brewer PB, Quon T, Smyth DR. The trihelix family of transcription factors—light, stress and development. Trends Plant Sci. Elsevier Ltd; 2012; 17: 163–171. https://doi.org/10.1016/j.tplants.2011.12.002 PMID: 22236699

25. Du HAI, Wang YONGIN, Xie YI, Liang ZHE, Jiang SANIE, Zhang SHH, et al. Genome-Wide Identification and Evolutionary and Expression Analyses of MYB-Related Genes in Land Plants. DNA Res. 2013; 1: 1–12.

26. Parenicova L. Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in Arabidopsis: New Openings to the MADS World. Plant Cell Online. 2003; 15: 1538–1551. https://doi.org/10.1105/tpc.011544 PMID: 12837945

27. Lechner E, Goloubinoff P, Genschik P, Shen WH. A gene trap Dissociation insertion line, associated with a RING-H2 finger gene, shows tissue specific and developmental regulated expression of the gene in Arabidopsis. Gene. 2002; 290: 63–71. https://doi.org/10.1016/S0378-1119(02)00556-5 PMID: 12062802

28. Pan R, Satkovich J, Hu J. E3 ubiquitin ligase SP1 regulates peroxisome biogenesis in Arabidopsis. Proc Natl Acad Sci U S A. 2016; https://doi.org/10.1073/pnas.1613530113

29. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, et al. MatInspector and beyond: Promoter analysis based on transcription factor binding sites. Bioinformatics. 2005; 21: 2933–2942. https://doi.org/10.1093/bioinformatics/bti473 PMID: 15860560

30. Hemberg M, Kreiman G. Conservation of transcription factor binding events predicts gene expression across species. Nucleic Acids Res. 2011; 39: 7092–7102. https://doi.org/10.1093/nar/gkr404 PMID: 21622661

31. Yanhui C, Xiaoyuan Y, Kun H, Meihua L, Jigang L, Zhaofeng G, et al. The MYB transcription factor superfamily of Arabidopsis: Expression analysis and phylogenetic comparison with the rice MYB family. Plant Mol Biol. 2006; 60: 107–124. https://doi.org/10.1007/s11103-005-2910-y PMID: 16463103

32. Taylor-Teeples M, Lin L, de Lucas M, Turco G, Toal TW, Gaudinier A, et al. An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature; 2014; 517: 571–575. https://doi.org/10.1038/nature14099 PMID: 25533953

33. Li QF, He JX. BZR1 Interacts with HY5 to Mediate Brassinosteroid- and Light-Regulated Cotyledon Opening in Arabidopsis in Darkness. Mol Plant. Elsevier; 2016; 9: 113–125. https://doi.org/10.1016/j.molp.2015.08.014 PMID: 26363272

34. Vaistij FE, Gan Y, Penfield S, Gilday AD, Dave A, He Z, et al. Differential control of seed primary dormancy in Arabidopsis ecotypes by the transcription factor SPATULA. Proc Natl Acad Sci U S A. 2013; 110: 10866–71. https://doi.org/10.1073/pnas.1301647110 PMID: 23754415

35. Chan AC, Khan D, Girard IJ, Becker MG, Millar JL, Sytnik D, et al. Tissue-specific laser microdissection of the *Brassica napus* funiculus improves gene discovery and spatial identification of biological processes. J Exp Bot. 2016; 67: erw179. https://doi.org/10.1093/jxb/erw179 PMID: 27194740

36. Becker MG, Zhang X, Walker PL, Wan JC, Millar JL, Khan D, et al. Transcriptome analysis of the *Brassica napus-Leptosphaeria maculans* pathosystem identifies receptor, signalling and structural genes underlying plant resistance. Plant J. 2017; https://doi.org/10.1111/tpj.13514 PMID: 28222234

37. Kim WC, Kim JY, Ko JH, Kang H, Han KH. Identification of direct targets of transcription factor MYB46 provides insights into the transcriptional regulation of secondary wall biosynthesis. Plant Mol Biol. 2014; 85: 589–599. https://doi.org/10.1007/s11103-014-0205-x PMID: 24879533

38. Tapia-López R, García-Ponce B, Dubrovsky JG, Garay-Arroyo A, Pérez-Ruíz R V, Kim S-H, et al. An AGAMOUS-related MADS-box gene, XAL1 (AGL12), regulates root meristem cell proliferation and flowering transition in Arabidopsis. Plant Physiol. 2008; 146: 1182–1192. https://doi.org/10.1104/pp.107.108647 PMID: 18203871

39. Chan A, Belmonte M. Histological and ultrastructural changes in canola (Brassica napus) funicular anatomy during the seed lifecycle. Botany. 2013; 91: 671–679. https://doi.org/10.1139/cjb-2013-0141

40. Kawai-Yamada M, Hori Z, Ogawa T, Ihara-Ohori Y, Tamura K, Nagano M, et al. Loss of calmodulin binding to bax inhibitor-1 affects pseudomonas-mediated hypersensitive response-associated cell death in Arabidopsis thaliana. J Biol Chem. 2009; 284: 27998–28003. https://doi.org/10.1074/jbc.M109.037234 PMID: 19674971

41. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: A GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010; 38: 64–70. https://doi.org/10.1093/nar/gkq310 PMID: 20435677

42. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, et al. RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Res. 2015; 43: W50–6. https://doi.org/10.1093/nar/gkv362 PMID: 25904632

43. Yan T, Yoo D, Berardini TZ, Mueller LA, Weems DC, Weng S, et al. PatMatch: A program for finding patterns in peptide and nucleotide sequences. Nucleic Acids Res. 2005; 33: 262–266. https://doi.org/10.1093/nar/gki368 PMID: 15980466

44. Blatti C, Sinha S. Motif enrichment tool. Nucleic Acids Res. 2014; 42: 1–6.

45. Kwon AT, Arenillas DJ, Worsley Hunt R, Wasserman WW. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. G3 (Bethesda). 2012; 2: 987–1002. https://doi.org/10.1534/g3.112.003202 PMID: 22973536

46. Babur O, Dogrusoz U, Cakır M, Aksoy BA, Schultz N, Sander C, et al. Integrating biological pathways and genomic profiles with ChiBE 2. BMC Genomics. 2014; 15: 642. https://doi.org/10.1186/1471-2164-15-642 PMID: 25086704

47. Haynes B, Maier E, Kramer M. Mapping Functional Transcription Factor Networks from Gene Expression Data. Genome Res. 2013; 23: 1319–1328. https://doi.org/10.1101/gr.150904.112 PMID: 23636944

48. Roider HG, Kanhere A, Manke T, Vingron M. Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics. 2007; 23: 134–141. https://doi.org/10.1093/bioinformatics/btl565 PMID: 17098775