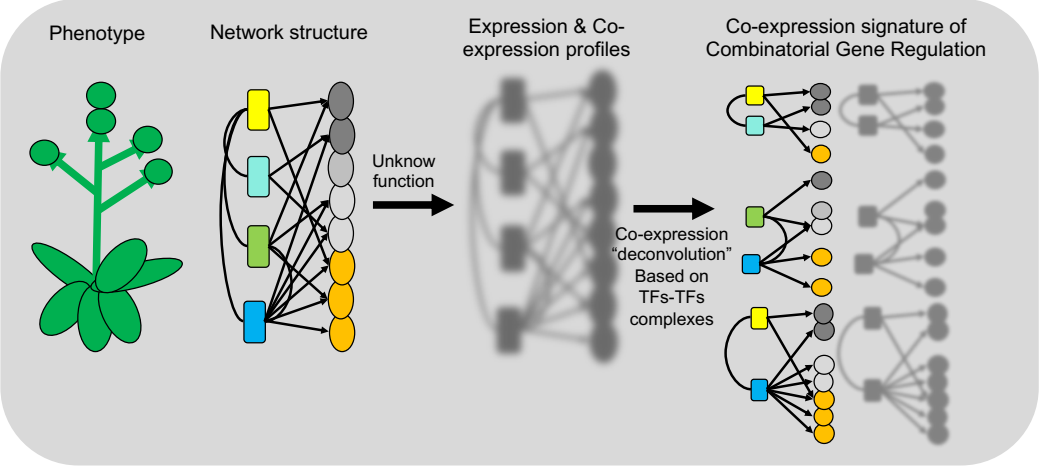
Co-expression Signature of Combinatorial Gene Regulation

Fabio Gomez-Cano[a]

[a]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48823

Graphical Abstract



Abstract

Combinations of TFs controlling different sets of targets genes is a phenomenon widely accepted. However, it is not clear how commonly TF complex formation influences the co-expression of individual TF with their targets. To reduce this conceptual gap, we performed a systematic characterization of all protein-protein interaction (PPI) reported for Arabidopsis, and combined this data with protein-DNA interactions (PDI), to test the potential contribution of TF's targets not co-expressed with its corresponding TF as a function of a third TF. We used here the REPRESSOR OF GA1-3 1 (RGA) DELLA protein as a proof of concept. Interestingly, each RGA's PPI is able to explain in average 90 targets usually not co-expressed with RGA. Similarly, comparison of the whole conditional co-expression ranks of RGA interactor vs. random interaction showed that changes induced by validated interactor is far more conservative than those caused by random interactions. We used five reported PPI prediction strategies to identify new PPI and test its support by conditional co-expression. Preliminary results suggest that RGA may have a significant control role in targets of those new PPI.

Introduction

Elucidation of gene regulatory networks (GRNs) is one of the major areas in biology, given the intrinsic relationships between phenotype and gene expression. It is widely accepted that this connection required also a well-defined multilayer system which could involve different signaling, interpretation, and control layers (cite). However, in many cases the interpretation and control layers are represent by transcription factor which interact between them in a combinatorial way to face and respond to the huge diversify of potential signals. In plants have been described a valuable number of phenomena which involve these multilayer systems. The crosstalk between Auxin and Brassinosteroid (BR) are one sample of them, in which both signaling pathways are able to induction of similar TF family members (Auxin Response Factors [ARFs]). Consequently, create and "destroy" TF complexes is a cellular daily task, which actually do have major transcriptional consequences for the cell. Accordingly, understand and predict when a set TFs form a regulatory complex is an imperative task to shed light on cellular transcriptional responses that different signals are able to trigger in the cell.

In general, GRNs have a large number of components (nodes [TFs and targets genes]) with intricate relationships between them (edges) (Bolouri 2014), where some of them has an specific hierarchy level e.g., TFs interaction with targets genes (refer here as protein-DNA interaction [PDI]). A simple interpretation of this structure is that a large set of TFs could regulate the same gene, and similarly a single TF could regulate several different genes. *Arabidopsis thaliana* has ~2500 known genes that code for TFs and co-regulators (Dai et al. 2016; Yilmaz et al. 2011). This last highlight the vast combinatorial potential of its genome (i.e., more than three hundred thousand potential TFs complexes [assuming all possible combinations of heterodimers]), of which are known only 9,503 pairs (Bemer et al. 2017; Chatr-aryamontri et al. 2017), which highlight the gap between the current knowledge and the potential regulatory plasticity of Arabidopsis.

The study of GRNs complexity has been approached from mainly three perspectives: (1) Characterizing their components, i.e., TFs, target genes or cis-regulatory regions. (2) Assaying the regulatory effect of their components, e.g., mutating TFs and testing its expression effect over potential targets. (3) Systematic identification of interactions between their components, e.g., interactions between TFs and potential targets genes (cite) or physical interaction between TFs (cite). Similar ideas were proposed from the computational area, such as the interrogation of TFs complexes based on genes co-regulated by same TFs (Wu and Lai 2016) or by co-localization of regulator binding sites (Guo and Gifford 2017). Equally, the combination of different data types (e.g., expression data, PDI, open chromatin regions) to prioritize pairs of TFs with similar behaviors (Glass et al. 2013; Song et al. 2017). Remarkably, a common characteristic of these

approaches is the use of node's co-expression as a way to prioritize or score its prediction. However, this last raised two potential restrictions/opportunities: (1) Only a selected number of organisms have enough data to use these models which limit their use. (2) There is not clear how often a TF complex formation affect the expression of its targets, and if it does how it could be implemented to prioritize and/or predict TF complexes. Several algorithms have been proposed to analyze gene expression conditionate by different type of modulator effectors (e.g., post-translational modification, signaling induction, modulators of gene-gene interactions) (Giorgi et al., 2014; Gambardella et al., 2015; Hsiao et al., 2016). Remarkable, although they were not thought to test combinatorial regulation, the differences in metrics and/or assumptions of each method, all of them shown success predicting specific combinatorial control phenomena such as the dependence of kinases in order to see co-expression between TFs and their targets (Gambardella et al., 2015).

In this work, we present a systematic and genome-wide way evaluation of TF-complex formation effect in the co-expression between TF and their targets. Our analysis takes advantage of conditional co-expression and expression average of experimental validated TF-complexes as parameter to measure the complex effect of set of also experimental predicted target genes for a set of xxx *Arabidopsis thaliana* TFs. We show that in xxx percentage of the complexes evaluated may induce affect the co-expression between TF and their target conditioned by TF which are able to interact physical between them.

Dataset and Methods

Data collection

The collection of PDI data was based on two primary sources of TFs: (1) The Arabidopsis Gene Regulatory Information Server (AGRIS) (Yilmaz et al. 2011). (2) PubMed database (https://www.ncbi.nlm.nih.gov/pubmed/) from the NCBI (searching for: Arabidopsis [AND] ChIP-seq) (Supplementary Table 1). All PDI collected were restricted to TF-centered approaches (i.e., ChIP-seq/chip and DAP-seq), which provide an unbiased selection of TF's targets. PDI datasets without binding data available were filtered out (i.e., without genomic coordinates of peaks). TF's targets were assigned by overlapping of PDI with Arabidopsis gene promoter (2 kb upstream from transcription star site [TSS]) obtained from The Arabidopsis Information Resource (TAIR, https://www.arabidopsis.org/) (annotation version TAIR10).

The PPI data was collected from the well-curated BioGRID database (V3.5.169) (Chatr-aryamontri et al. 2017). A total of ~ 56,000 pairs of PPIs were collected, of which only 845 PPIs correspond to TFs for which there was collected PDI data. Finally, all the expression data used in the TF-complex co-expression effect (see below) were carried out using 1401 RNA-seq libraries previously analyzed and collected from the ATTED-II database (version Ath-r.v15-08). Meanwhile, all normal paired TF-target comparison were carryout using already calculated weighted PCC from ATTED-II database (Version Ath-r.c2-0) using similar expression dataset (Obayashi et al. 2018).

TF-Complex's targets co-expression

Evaluation of TF-TF complexes co-expression effect over their corresponding targets was carried out based on partial correlation (Kim 2015) and by averaging the expression of TF pairs (a PPI) to then re-calculated its co-expression with their targets. The partial correlation was calculated using

the R package PPCOR (Kim 2015). The co-expression between TF pairs and targets by averaging the expression of was calculated by first averaging expression between TF pairs creating a new protein complex expression profile and then calculating the correlation between this expression profile and all the target genes of both TFs in the complex. This correlation was weighted Partial Correlation Coefficients (wPCC) using the R package wCor (v1.9.1, https://CRAN.R-project.org/package=wCorr). All the expression data used was normalized and log2 transformed (Obayashi et al. 2018), and all the TF-TF complex tested were collected as describe previously. We compare the co-expression of TFs-complex and their targets vs random complex which were create using one of the TF complex vs random TFs from the pools of TFs with PDI data as negative control. Finally, we define -for all cases- genes highly co-expressed with a TF or TF complex as the set of genes within any of the 2.5% tails of the correlation distribution.

TF-TF complex prediction

The prediction new TF-TF complexes was based on widely describe methods which explore a broad spectrum of strategies. These are divided in five classes: (1) TFs co-expression (Zand and Sunter 2018). which is extrapolation of guilt-by association idea (Uygun et al., 2016). The co-expression was calculated as reported in Obayashi et al., (2018) using normalized expression data describe in data section. (2) Common co-expressed genes as describe. We use the top 150 TF's co-expressed genes to count common co-expressed genes and use this value as a metric to measure possible TF-complexes. This analysis was made for all possible TF pairs present list of TF with PPI and PDI data. The number of common co-expressed genes were centered and scale by the mean and standard deviations of all comparisons as describe by Nie et al., (2011) (3) Common target genes (Heyndrickx et al. 2014; Zand and Sunter 2018). This parameter was measure by testing the enrichment of common targets by a Fisher exact test and calculating the similarity between set of targets by Jaccard coefficient for each possible pair of TFs. (4) Network topology similitudes using CoReg R packages (v1.01) (run with default parameters) (Song et al. 2017). This strategy has into account in and out degree similarities between nodes to cluster them into modules. (5) TFs co-binding pattern. This analysis takes into account TF co-binding pattern along all chromones in order to identify modules of TF co-binding. The identification of the co-binding region was performed by the regulatory module discovery (RMD) with default parameters (Guo and Gifford 2017).

Results

Characterization of TF-TF complex effect on targets co-expression

In order to characterize the how TFs complexes formation may affect co-expression patter between TF and their targets we carry out a systematic analysis of co-expression integrating three types of regulatory related data sets: 1) Expression data, 2) Experimental identify Protein (TF) – DNA interaction data (PDI) and 3) experimental identify Protein-Protein interaction data (PPI) (*See Methods*). In total, we collected PDI data for 566 TFs corresponding to ~ 11,5 million of peaks with a median, minimal, and maximal of 4904, 20 and 6,317,204 peaks by TFs, respectively. Consequently, we filter out theses PDIs based on expression and their present in the interactome database (*See Methods & Supplemental figure 1*). Subsequently, the final data set used as our gold standard to further experiments comprised a set of 320 TFs which all least 845 experiment PPI validated PPI between them.

Using these complexes, we then ask if the co-expression between a TFx with its targets may vary or by altered after conditioned it to a third TFz. We use the REPRESSOR OF GA1-3 1 (RGA) DELLA proteins, which are functional hubs plant TFs and count of a large number of validated PPIs (28 PPI related to 16 different TFs) (Figure 1a), as proof of concept experiment. We first evaluated co-expression between RGA and whole set of Arabidopsis genes splitting them in four categories: 1) unique RGA's targets (*x*), RGA and interactors common targets (*xz*), RGA's interactors targets (*z*), and non-targets of either RGA or their interactors (*none*) (Figure 1a, b). With these categories we count the number of highly co-expressed genes (*See Methods*). Intriguingly, RGA's targets (gene set *x* and *xz*) represent only 17% (65/381) and 10% (63/652) of the total highly positive and negative RGA's co-expressed genes, respectively (Figure 1b). However, the *z* gene set (target of RGA's interactor) is able to explain 76% and 72% of the total highly positive and negative RGA co-expressed genes. Which means that adding up RGA's target and their sixteen interactor's targets we are able to explain 93% and 82% of RGA's positive and negative co-expressed genes. This last suggest that RGA may have a major regulatory role in the expression control over its interactors' targets. Consequently, by extrapolation we can hypothesize a major role of RGA's interactor over its not highly co-expressed targets.

To further evaluate the putative relationship between RGA's interactors and its non-co-expressed targets, we performed an analysis based on partial correlations and average expression between RGA and its interactors (*See Methods*). We first identify RGA target non-coexpressed under normal PCC, and then count how main of them become co-express genes under partial or average co-expression of RGA with each of its interactors. Interestingly, all RGA interactors are able to add new targets to the set of RGA co-expressed target (Figure 2a). In average, the average and partial correlation allow to identify 70 and 85 new co-expression targets (Figure2b), respectably. TCP22 was the RGA's interactor which support much larger number of new co-expressed targets (Figure2a). In total, including all targets supported by the sixteen RGA interactor, it was possible to identity 28% (630/2237) RGA's targets initially non co-expressed with it, which are co-expressed in function of any RGA interactor (Figure 2c).

Random TF-TF complex produce high variable correlation ranks

Every single RGA interactor produced alterations in the correlation between RGA and the entire genome, within them targets genes (See above). In order to test the significance of the changes, we select 10 random TFs which are not RGA interactor and used them to calculate average PCC and partial correlation between RGA and their targets similar to known interactor. To measure the change induced by these both set of TFs (known PPIs and random PPIs) we calculated the spearman correlation (ρ) between the original RGA PCC distribution vs the new distribution (Average/partial) (Figure 3). Interestingly, ρ changes induce by known interactors vary between 5.1 and 7.4 (Figure 3), meanwhile random PPI have a wide more variable range (ρ values between 3 and 7.4). Suggesting that known interactor have similar RGA expression profiles, by which they are able to alter a defined population of genes. However, further analysis is required to confirm this hypothesis.

Preliminary result in PPI prediction

Given the predictive potential and biological impact that the conditional correlation may have in the identification of TF- complexes or targets genes, we apply different widely used PPI methods based on the current available data to then characterized them by mean of conditional correlation. First, we apply co-binding methods originally thought to predict regulatory combinatorial modules (Guo and Gifford 2017). Here was assume that TF present in the same regulatory module are potential PPIs. Under this assumption, we identify 64 modules, with had wide rank of TF in each of them (Figure 4a) and predict 272,669 possible PPIs (Figure 4). Second, we compared the number of common targets between each possible pair of TFs along with its co-expression values as reported by Zand et. al., (2018). In total, it was identified 11,812 putative PPIs (Figure 5). Third, we count the number of common co-expressed genes between all TF pairs between the top 150 highly co-expressed gene of each TF. We use the number of common co-expressed gene of all pairs assayed to filter out pairs below 1.5 standard deviation from the mean, which leach a total of 2699 possible PPI (Figure 6)

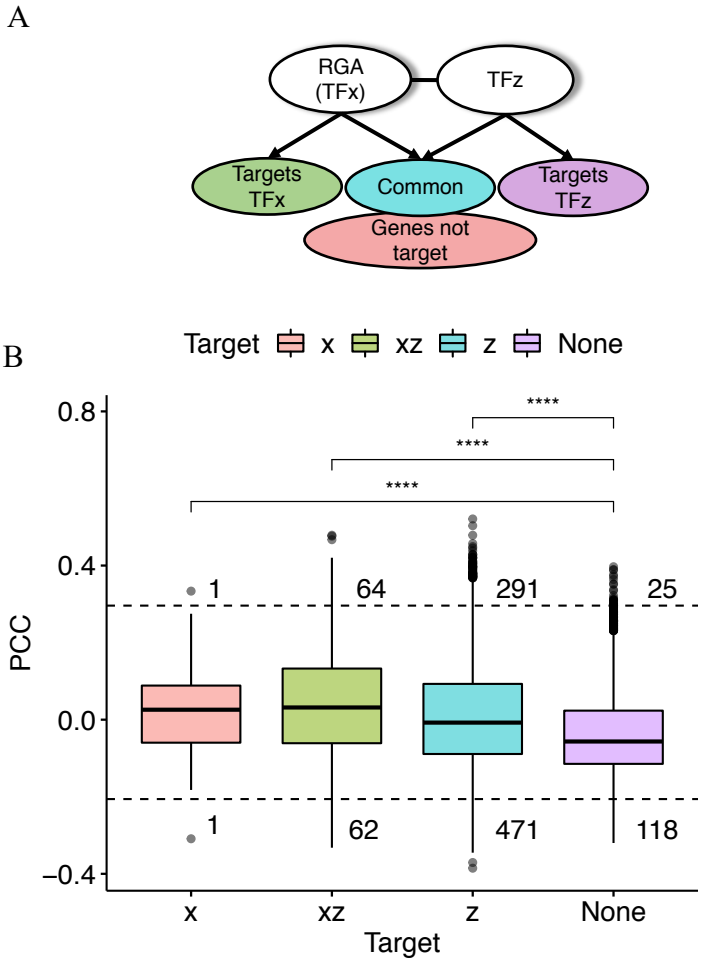


Figure 1. Arabidopsis genes correlation (PCC) profile with the transcription factor RGA. A) Network example showing the classification rules used to characterize RGA genes highly co-expressed with RGA. B) PCC distribution of RGA targets (*x*) (light red), RGA and its interactor common targets (*xz*) (light green), RGA's interactor targets (*z*), and gene non-targets either of RGA

or its interactors (*None*) (light purple). Asterisk represent P value significance of based on a Wilcoxon test (p > 0.05. *: p <= 0.05. **: p <= 0.01. ***: p <= 0.001. ****: p <= 0.0001).

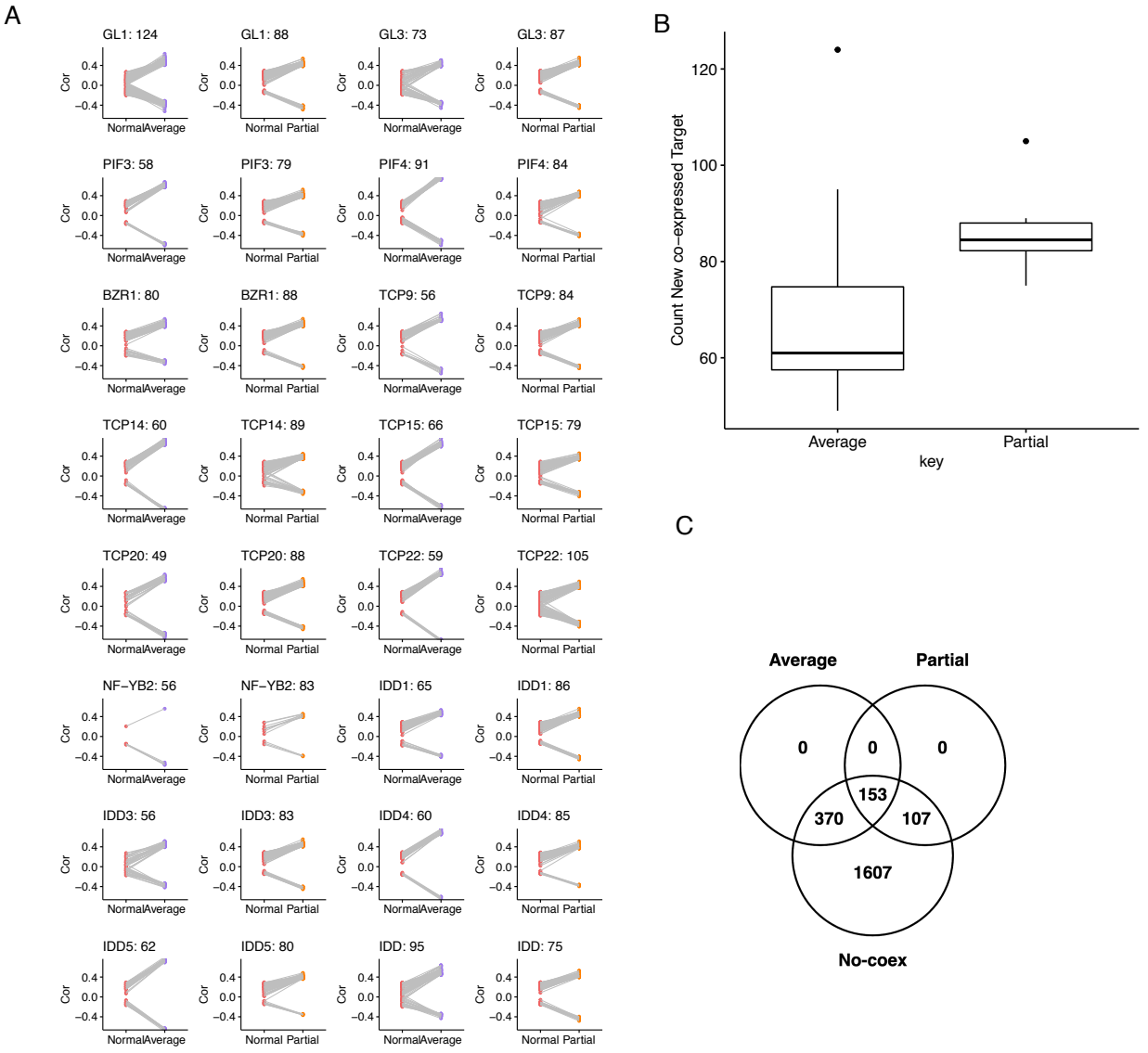


Figure 2. RGA target co-expression in function of its interactors. A) Dot paired plot of RGA targets co-expressed in function of each RGA interactor. Light red, represent normal PCC, purple dots resent new PCC in based on TF expression average, and orange dots represent new correlation based on partial correlation. Highly co-expressed genes base on average and partial correlation are defined as the 2.5% genes in each tail (positive and negative) of the distribution. B) Distribution of new targets co-expressed by all ARG interactor based on average and partial correlation. C) Total list of new co-expressed targets compared to the indicial set of non-coexpressed RGA targets.

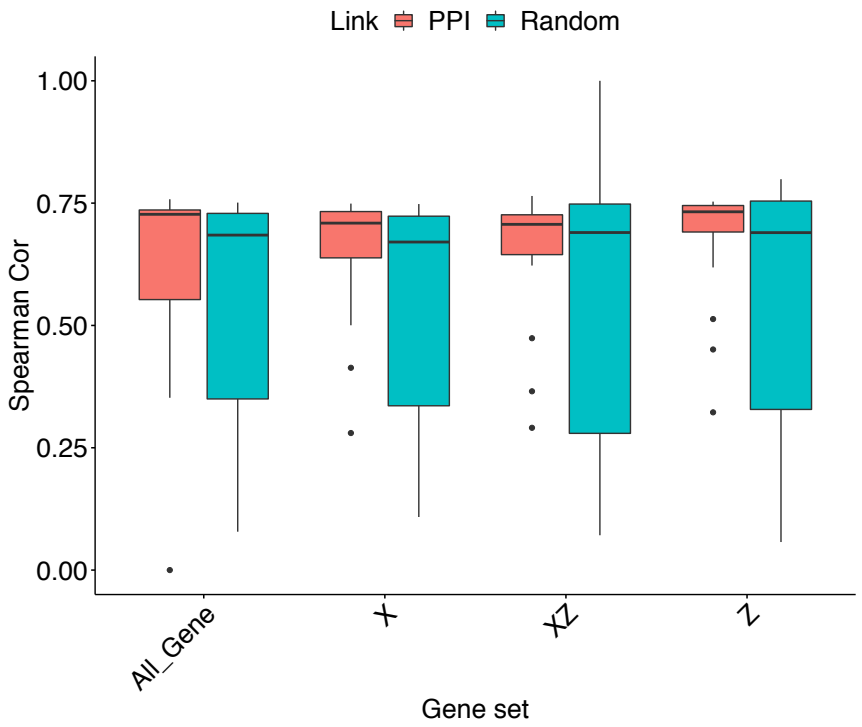


Figure 3. Changes induce by random PPI are more variable than known PPI. Comparison of known and random PPI based in the normal RGA co-expression profile and the new correlation profile (based on average/partial) of different gene sets.

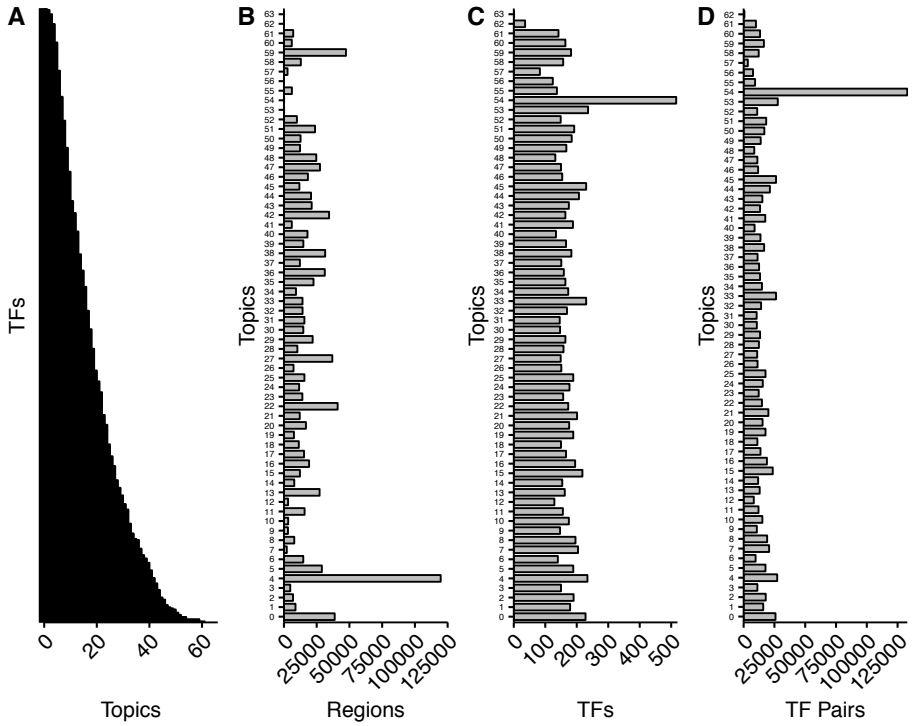


Figure 4. PPI prediction based on co-binding. A) Distribution of each TF along each topic. B) Distribution of topics along genomic regions. C) Distribution of number of TF by topic. D) Predicted PPI by topic.

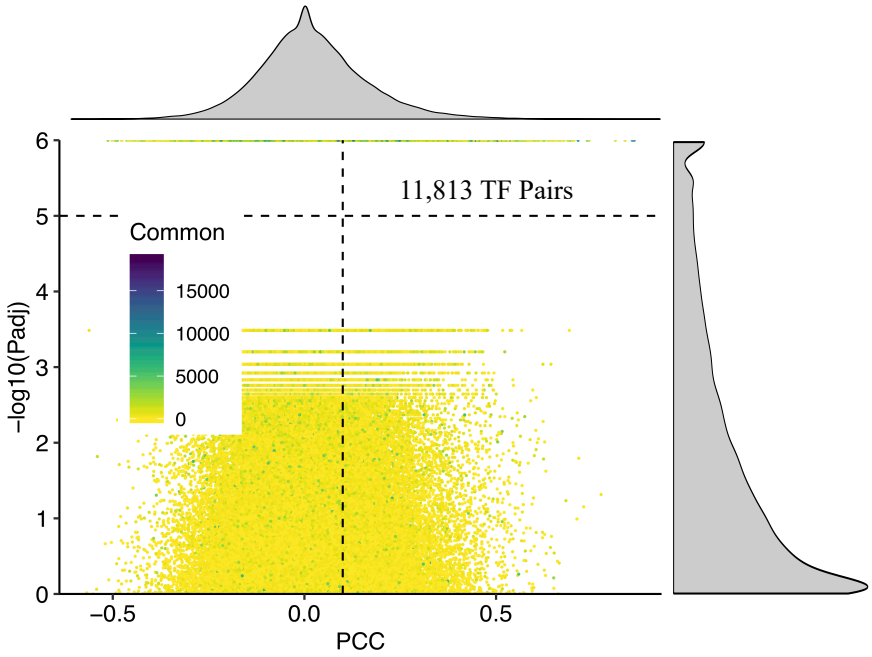


Figure 5. PPI prediction based on co-expression and common targets. PCC Distribution of predicted PPI and -log10 P value of common targets between TFs.

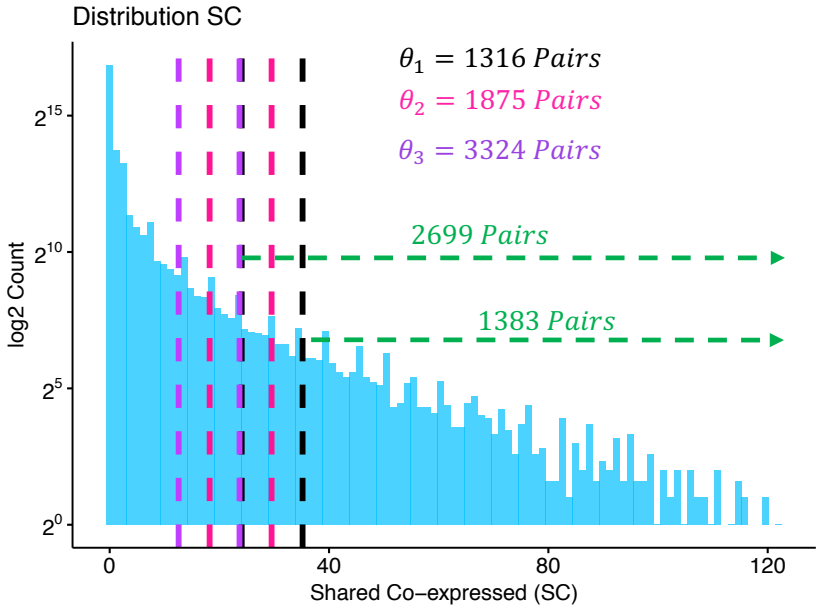
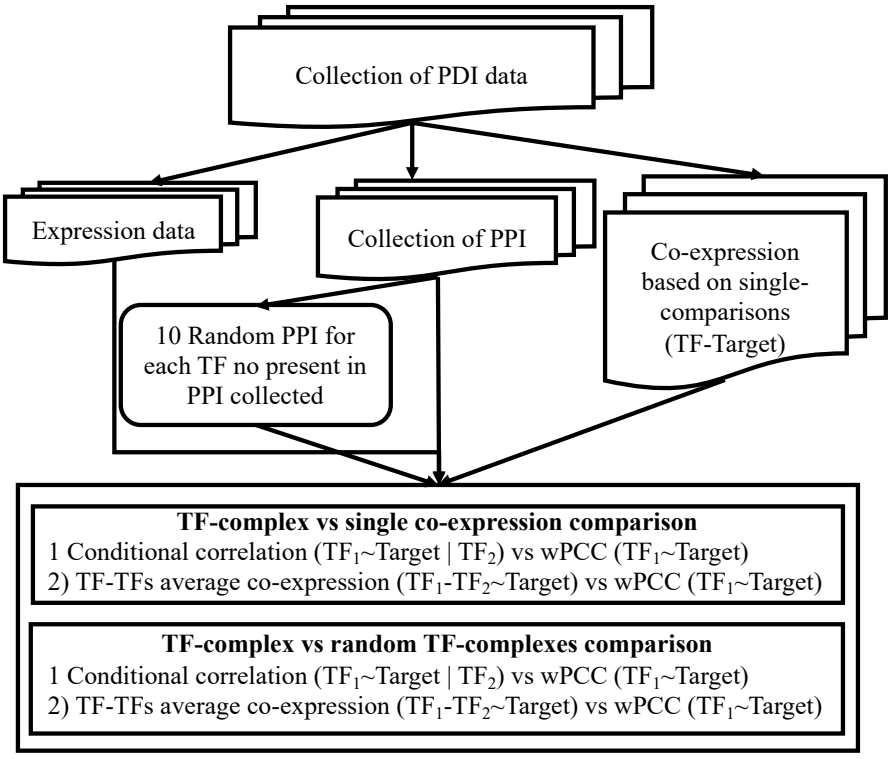


Figure 6. PPI prediction based on common co-expressed genes. Shared co-expression genes between all possible TF pairs. Vertical lines represent reported ranges reported in Nie et al., (2011).

Discussion

In addition to the some of the current ideas developed during the first section of the report, in term of the potential use of combinatorial regulation a its impact in the co-expression between TFs and their corresponding targets. I would like to highly the importance of the BIGRID database, as well as the concept of partial correlation, prediction evaluation, as well as regulatory network concept in the development of this project.

In general, I think the project goes well and currently there is an interesting set of results to interpret.



Supplement Figure 1. Collection, filter and analysis pipeline to test the effect of a TF-TF complex formation over TF's targets co-expression profiles.

Literature

Bemer, Marian, Aalt D. J. van Dijk, Richard G. H. Immink, and Gerco C. Angenent. 2017. "Cross-Family Transcription Factor Interactions: An Additional Layer of Gene Regulation." *Trends in Plant Science* 22(1):66–80.

Bolouri, Hamid. 2014. "Modeling Genomic Regulatory Networks with Big Data." *Trends in Genetics* 30(5):182–91.

Chatr-aryamontri, Andrew, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K. Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, Chris Stark, Bobby-joe Breitkreutz, Kara Dolinski, and Mike Tyers. 2017. "The BioGRID Interaction Database : 2017 Update." *Nucleic Acids Research* 45:D369–79.

Dai, Xinbin, Zhangjun Fei, Patrick X. Zhao, Gregory B. Martin, Marina A. Pombo, Chen Jiao, Honghe Sun, Seung Y. Rhee, Yi Zheng, Hernan G. Rosli, Michael Banf, Peifen Zhang, and James J. Giovannoni. 2016. "ITAK: A Program for Genome-Wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases." *Molecular Plant* 9(12):1667–70.

Guo, Yuchun and David K. Gifford. 2017. "Modular Combinatorial Binding among Human Trans-Acting Factors Reveals Direct and Indirect Factor Binding." *BMC Genomics* 18(1):1–16.

Heyndrickx, K. S., K. Vandepoele, D. Weigel, J. V. de Velde, and C. Wang. 2014. *A Functional and Evolutionary Perspective on Transcription Factor Binding in Arabidopsis Thaliana*. Vol. 26.

Obayashi, Takeshi, Yuichi Aoki, Shu Tadaka, Yuki Kagaya, and Kengo Kinoshita. 2018. "ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index." *Plant & Cell Physiology* 59(1):e3.

Song, Qi, Ruth Grene, Lenwood S. Heath, and Song Li. 2017. "Identification of Regulatory Modules in Genome Scale Transcription Regulatory Networks." *BMC Systems Biology* 11(1):1–18.

Yilmaz, Alper, Maria Katherine Mejia-guerra, Kyle Kurz, Xiaoyu Liang, Lonnie Welch, and Erich Grotewold. 2011. "AGRIS : The Arabidopsis Gene Regulatory Information Server , an Update." *Nucleic Acids Research* 39(November 2010):1118–22.

Zand, Maryam and Garry Sunter. 2018. "An Integrative Approach to Transcriptional Co-Regulatory Network Construction and Characterization in Arabidopsis." *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)* (61772288):1–6.

Supplementary Material - Co-expression Signature of Combinatorial Gene Regulation

Supplementary Table 1. Summary of PDI data available based on literature search.

Author	PDIM	Number of TFs reported	Methods
Albihlal et al., 2018	29697803	1	ChIP
Besbrugge et al., 2018	29678859	1	ChIP
Brandt et al., 2012	22578006	1	ChIP
Chen et al., 2018	30382087	15	ChIP
Gregis et al., 2013	23759218	1	ChIP
Hendrickx et al., 2014	25361952	27	ChIP
Jensen et al., 2014	23951554	1	ChIP
Li et al., 2016	27469166	1	ChIP
Liu et al., 2015	26076231	1	ChIP
Liu et al., 2016	26586835	1	ChIP
Merelo et al., 2013	24155946	1	ChIP
Nagel et al., 2015	26261339	1	ChIP
Omalley et al., 2016	27203113	522	DAP-seq
Omaoileidigh et al., 2013	23821642	1	ChIP
Shanks et al., 2018	29763523	1	ChIP
Song et al., 2016	27811239	21	ChIP
VanLeene et al., 2016	27660483	1	ChIP
Verkest et al., 2014	24453163	1	ChIP
Wang et al., 2010	20144209	1	ChIP
Xu et al., 2018	29121271	1	ChIP
Yant et al., 2010	20675573	1	ChIP
Total unique TFs*		566	

* Some TFs has more than an experiment reported, or their peaks are published as a re-analysis by a second paper. In those cases, both sets of peaks were collected as well as their respective references. The number reported as Total unique TFs represent the final list of TFs without duplications.