Insights into functional classification via gene co-expression networks in sunflower using public transcriptomic datasets

Andrés I. Ribone[1], Sergio Gonzales[1], Norma Paniego[1], Verónica Lía[1], Máximo Rivarola[1]

1 IABIMO UE Conicet-INTA. N. Repetto y Los Reseros s/n, Bs As., Argentina.
ribone.andres@inta.gob.ar, gonzalez.sergio@inta.gob.ar,
paniego.norma@@inta.gob.ar, lia.veronica@@inta.gob.ar,
rivarola.maximo@@inta.gob.ar.

Abstract. We present the first preliminary sunflower gene co-expression network using public transcriptome data in *Helianthus annuus* and show its utility in identifying and classifying uncharacterized genes involved in stress response. The locus HanXRQChr09g0248321 was identified and linked to several WRKY transcription factors in an enriched "stressed-response" module. Moreover, the homologue in *Arabidopsis thaliana* was shown to be differentially expressed in multiple "stress" conditions. We present our work and validate our methodology to existing knowledge and show its capability to identify/rank new candidates for crop breeding programs. Our future goal is to link genetic variation with gene networks to understand phenotypic variability in sunflower stress responses

Keywords: Gene Co-expression Networks, RNAseq, Functional Annotation, Big Data, Sunflower

1 Introduction

Sunflower is one of the most important crops for the production of high-quality oil and seeds consumed by both humans and livestock. Resistance or tolerance to stress is a complex trait usually determined by many interacting loci. Over the past decade, many groups have performed genomic and transcriptomic experiments in different stages and conditions to understand the regulatory and genetic basis for many stress related responses in plants. Knowledge gained from every experiment is key to a better understanding of the biological process, yet integrating and analyzing in conjunction different experiments can lead to a more systemic insight into the complex mechanisms of sunflower response [1]. A popular approach in systems biology is the construction and analysis of gene networks. Such networks are often used for genome-wide representation of the complex functional organization of biological systems [2]. Networks based on similarity in gene expression are called gene co-expression networks and can be used to associate genes of unknown function with biological processes, to prioritize candidate disease genes or to discern transcriptional regulatory programs. In addition, networks constructed linking variant

positions in the genome to gene expression, also known as eQTL analysis, can provide information on variant loci which control specific gene expression [3]. Together these network analysis can provide new knowledge for the domain studied. Transcriptomic meta-analysis aims at re-analysing existing data to derive novel biological hypotheses, and is motivated by the public availability of a large number of independent studies [4]. Here we construct a gene co-expression network to associate genes of unknown function with specific biological processes in *Helianthus annuus* and prioritize those involved in biotic stress resistance. Our long-term goals are to produce and analyze gene co-expression and eQTL networks in different tissues and conditions using all public transcriptomic data.

2 Methodology and Results

2.1 Data collection

To date, we have downloaded all available transcriptome datasets (39 in total) from the Sequence Read Archive (NCBI), which comprised 1021 RNAseqs runs corresponding to 49 tissue-age-treatment combinations, from 355 different genotypes (~8 TB of raw data).

To perform differential co-expression studies, samples were grouped by tissue and treatment according to the metadata. Our first group coined “photosynthetic tissue” includes samples from whole stem to leaf, and from seedlings to mature plants. We classified 693 samples as “control photosynthetic tissue” and 140 samples as “stressed photosynthetic tissue”. Moreover, we also grouped root samples in control (125) versus stressed (65). The “stress” definition, at this stage, included all biotic and abiotic stresses. To check and validate our methods, we will begin focusing on the larger group samples, specifically the “control photosynthetic tissue” where we can contrast our results to existing previous knowledge.

2.2 Quality control and read mapping

Adaptors and low quality reads were removed with Trimmomatic [5]. We quantified gene expression via Salmon [6] using the genome HanXRQ r1.2 as our reference to produce 1 transcript per gene (coding and non-coding), totaling 58138 genes [7]. The mapping rates ranged from 1% to 97%, with a median of 75% (sup. fig. 1). As expected, samples from wild-cultivar species had lower mapping percentages, as well as samples from infected tissues probably due to low efficiency in RNA extraction from sunflowers.

2.3 Weighted Gene Co-expression Network Analysis

As a proof of concept, the first group analyzed were samples from healthy photosynthetic tissues, initially a cohort of 693 samples. Samples with less than 3 million read counts were removed; then genes with less than 2 counts per million (CPM) in ¾ of samples or more were filtered out, resulting in a subset of 15.865 genes, across 673 remaining samples.

Moreover, outlier samples were filtered out in a recursive manner. Briefly, read counts were normalized via variance stabilizing transformation; batch effects arising from sample origin were adjusted via an empirical Bayesian method described by Johnson *et al.* [8] and implemented in the R package sva [9]; normalized and adjusted samples were hierarchically clustered using Pearson distance and visual outliers were removed (in this case, corresponding to a Pearson distance greater than 0.45) (sup. fig. 2); after this, normalization and batch effect correction was re-run for the remaining 653 samples. In our first approach we performed a gene co-expression network analysis (WGCNA) [10], using an unsigned correlation network with soft-threshold power beta of 6, a minimum module size of 30, and a signed topological overlap measure (TOM). Merge cut height, and deep split parameters were 0.25 and 2 respectively. This run resulted in 22 modules with a degree distribution that followed a power law (R^2=0.83, given by the soft-threshold 6), suggesting a scale free topology as expected. Approximately 5.5% of total genes remained unconnected and the largest module connected 19.5% of all genes. Moreover, we observed 82 "hub-genes" with a scaled intramodular-connectivity (SIC) of 0.9 or greater, with module3 and module4 having the highest number of hub-genes, with 10 and 14 respectively.

2.4 Gene Ontology Analysis

In order to relate and integrate modules within a biological context, we performed an enriched Gene Ontology (GO) term test on all 22 modules [11]. We observed that all modules were enriched significantly with one or more GO terms (padj-value <0.05). We first looked at the largest module, and found it contained enriched GO terms in "growth", "cell division", "plant-type cell wall biogenesis", and "microtubule-based process" among others. In order to validate our network, we chose module20 which contained 77 genes and was enriched for several interesting GOs, such as "response to stress", "response to fungus", "regulation to response to stress", and "signaling" among others (sup. fig. 3). Remarkably this module contained 7 WRKY transcription factors which are known to be involved in several stress response mechanisms [12]. Three out of seven WRKY are members of a module in a previous sunflower gene-network study published by Moschen *et al.* [13]. Besides, 24 of the 77 genes are also associated with the same module of their network. Interestingly, out of the 77 genes, 7 are of unknown/uncharacterized protein function yet conserved in plant genome(s) and supported by expression data (annotation Badouin *et al.* [7]). From these 7 genes HanXRQChr09g0248321 has a SIC=0.94, suggesting a main role in the module (sup. fig. 4, crowned blue node). The homologous gene in *Arabidopsis thaliana* AT4G29780 (blastX e-value ~0.0 and with 64% identity across a query coverage of 57%) has been described to have differential expression in response to several stress conditions [14]. Overall there are 45 interesting uncharacterized genes which belong to a module and have a SIC greater than 0.75, which makes them interesting candidates to follow up. We are currently examining these genes and other modules in the network.

3 Conclusions and Perspectives

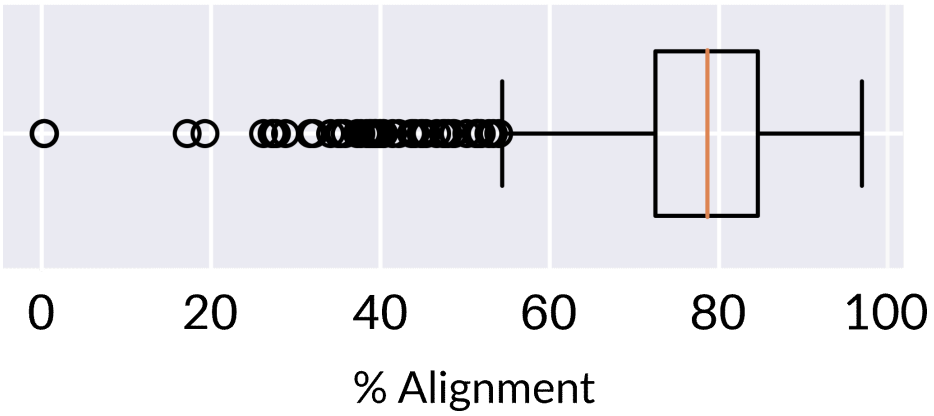
In this preliminary work we established and validated our large-scale network analysis and found promising evidence about the uncharacterized locus HanXRQChr09g0248321 being involved in stress response. We are currently analyzing other modules, particularly those with uncharacterized genes as hubs, as well as constructing networks with the other major groups of samples. In addition, we are working on eQTL analysis derived from these same sample groups. We have successfully genotyped all samples with high accuracy and expect to run eQTL analysis in the near future. Our final aim is to integrate all new knowledge coming from new candidates from network analysis, eQTLs, previous candidate loci (e.g. from GWAS) to stress response and integrate them to create a more systemic approach to better understand stress response in sunflower. In addition, any links we find between previous and new candidates will be followed up at the Sunflower Genomics laboratory at IABIMO with molecular assays and variant identification in local germplasm. Finally, we plan to implement a publicly available web page of networks for the scientific community to explore.

References

1. Hassani-Pak, K. et al. Developing integrated crop knowledge networks to advance candidate gene discovery. Appl. Transl. Genomics 11, 18–26 (2016).
2. Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M. & Ligterink, W. Learning from Co-expression Networks: Possibilities and Challenges. Front. Plant Sci. 7, 444 (2016).
3. Civelek, M. & Lusis, A. J. Systems genetics approaches to understand complex traits. Nat. Rev. Genet. 15, 34–48 (2014).
4. Caldas, J. & Vinga, S. Global meta-analysis of transcriptomics studies. PLoS ONE 9, e89318 (2014).
5. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).
6. Patro, R. et al. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419 (2017).
7. Badouin, H. et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 546, 148–152 (2017).
8. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127 (2007).
9. Leek, J. T. et al.. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28, 882–883 (2012).
10. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).
11. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology. (2019).
12. Eulgem, T., Rushton, P. J., Robatzek, S. & Somssich, I. E. The WRKY superfamily of plant transcription factors. Trends Plant Sci. 5, 199–206 (2000).
13. Moschen, S. et al. Exploring gene networks in two sunflower lines with contrasting leaf senescence phenotype using a system biology approach. BMC Plant Biol. 19, 446 (2019).
14. AT4G29780 - www.arabidopsis.org/servlets/TairObject?type=locus&id=127857.

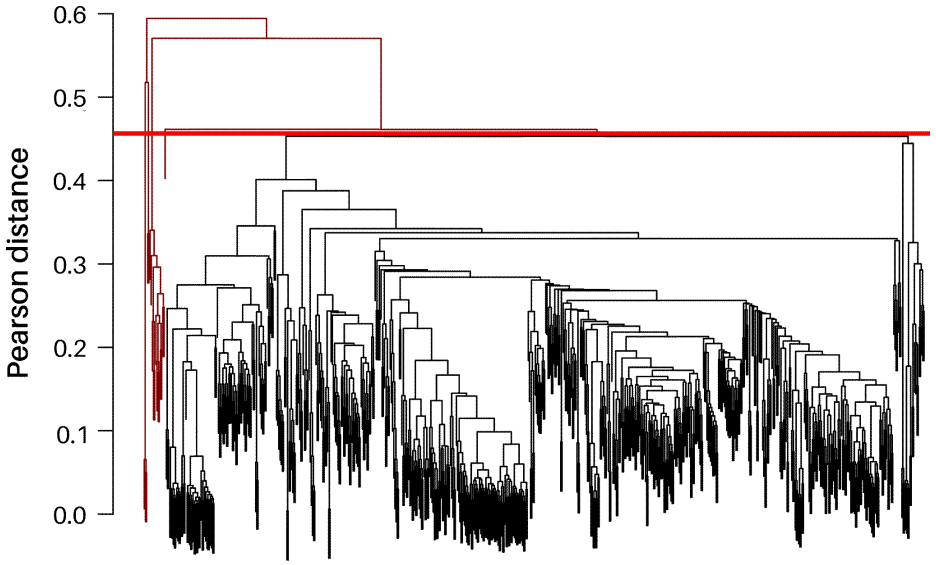
Supplementary Materials

Supplementary figure 1



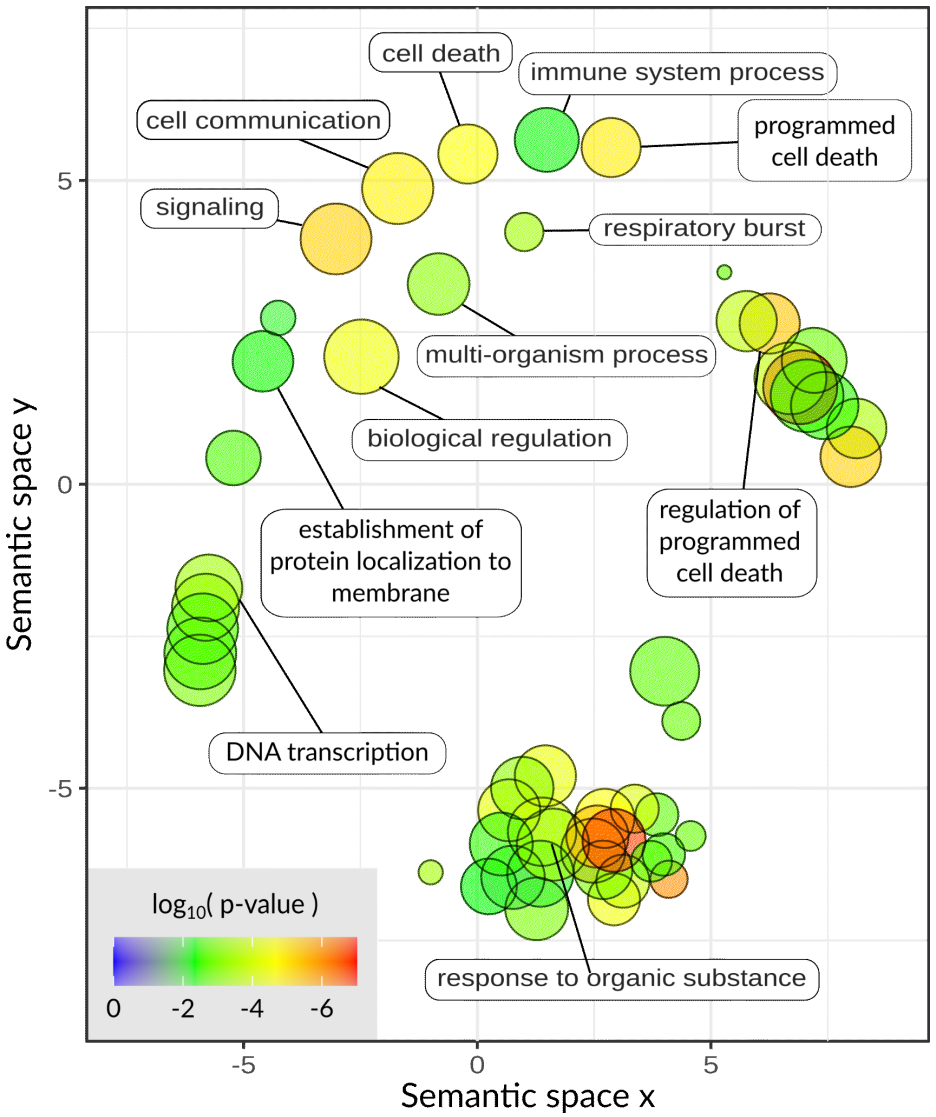
Supplementary figure 1. Whisker boxplot of Salmon alignment rates of all 1021 downloaded samples.

Supplementary figure 2



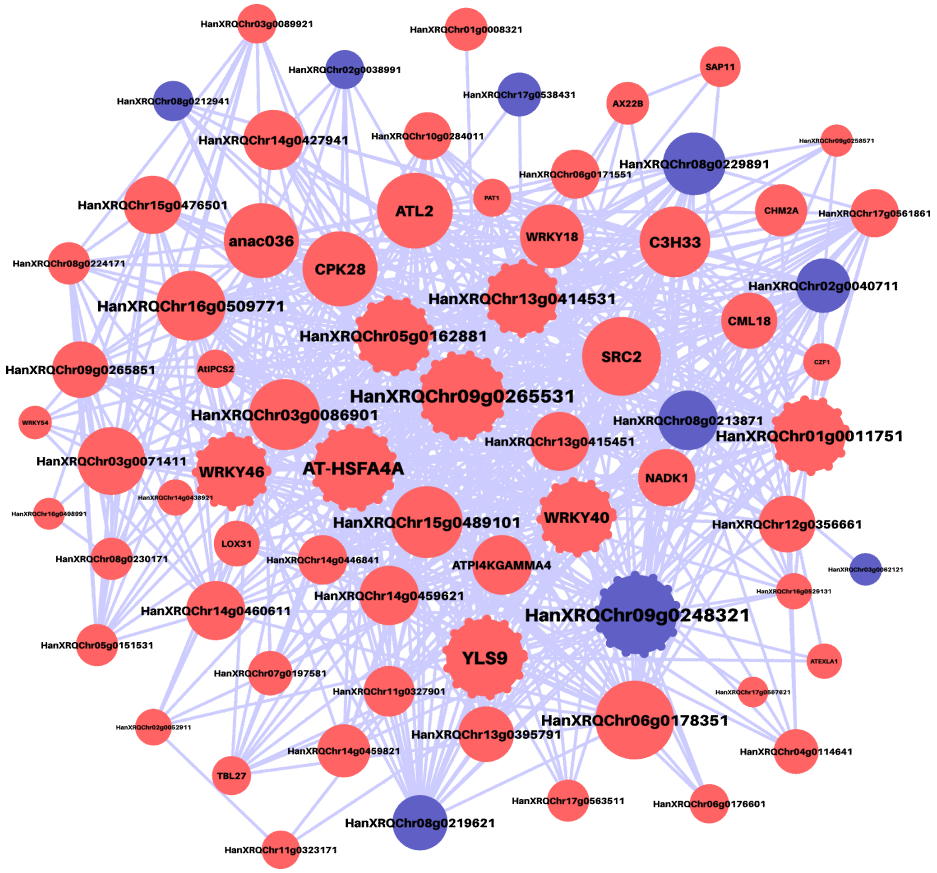
Supplementary figure 2. Hierarchical clustering by Pearson distance of batch adjusted samples corresponding to "healthy photosyntetic" group. Samples with Pear-son distance > 0.45 (red line) were excluded.

Supplementary figure 3



Supplementary figure 3. GO enrichment analysis of module20 summarized and visualized using REVIGO [1]. GO terms are represented by circles and are clustered according to semantic similarities in the gene ontology. More general terms are represented by larger size circles, and adjoining circles are most closely related. Circle color indicates the log10(p-value) for the enrichment derived from the Fisher test. Terms with dispensability < 0.15 are labeled.

Supplementary figure 4



Supplementary figure 4. Cytoscape network visualization of genes in module20 [2]. Node size indicates intramodular connectivity (IMC), with crowned nodes having IMC>0.8. Node color indicates whether the gene has a known function (red) or not (blue).

Supplementary references

1. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE. 6 , e21800 (2011).
2. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–504 (2003).