Reconstructing Gene Networks of Forest Trees from Gene Expression Data: Toward Higher-Resolution Approaches

Matt Zinkgraf[1], Andrew Groover[2], and Vladimir Filkov[3](✉)

[1] Western Washington University, Bellingham, WA, USA
[2] United States Forest Service, Davis, USA
[3] University of California, Davis, CA, USA

Abstract. In two of our recent systems biology studies of forest trees we reconstructed gene networks active in wood tissue development for an undomesticated tree genus, *Populus*. In the first study, we used time series data to determine gene expression dynamics underlying wood formation in response to gravitational stimulus. In the second study, we integrated data from newly generated and publicly available transcriptome profiling, transcription factor binding, DNA accessibility and genome-wide association mapping experiments, to identify relationships among genes expressed during wood formation. We demonstrated that these approaches can be used for dissecting complex developmental responses in trees, and can reveal gene clusters and mechanisms influencing poorly understood developmental processes. Combining orthogonal approaches can yield better resolved gene networks, but the resulting network modules may contain large numbers of genes. This limitation reflects the difficulty in creating a variety of experimental conditions that can reveal expression and functional differences among genes within a module, thus imposing limits on the resolving power of network models in practice. To resolve networks at a finer level we are now adding a complementary approach to our work: using cross-species gene network inference. In this approach, transcriptome assemblies of two or more species are considered together to identify expression responses common to all species and also responses that are species specific. To that end here we present a new tool, fastOC, for identifying gene co-expression networks across multiple species. We provide initial evidence that the tool works effectively in calculating co-expression modules with minimal computing requirements, thus making cross-species gene network comparison practical.

Keywords: Bioinformatics · Gene networks · Tree genomics

1 Introduction

Transcription is a primary regulatory point in gene regulation. The expression level of a gene reflects in part the regulation from upstream genes encoding

© Springer Nature Switzerland AG 2018
S. Kalajdziski and N. Ackovska (Eds.): ICT 2018, CCIS 940, pp. 3–12, 2018.
https://doi.org/10.1007/978-3-030-00825-3_1

transcription factors (TFs), that bind to the regulatory cis-elements in the promoters of downstream gene targets [8,15]. The specific and combinatorial binding of transcription factors to target gene promoters contributes to spatial, and temporal regulation, as well as rates of transcription [8,11]. Additional levels of regulation include the stability and rates of degradation of gene mRNA transcripts. When RNA sequencing is used to measure mRNA transcripts, the combined influence of gene expression regulation and transcript stability properties is being observed.

Gene regulatory networks describe the links between the regulator genes and their targets. In aggregate these networks are extremely complex and in practice incompletely modeled, even for model species [11]. However, it appears that basic properties of these networks are shared by all organisms surveyed to date. For example, similar to what has been described in animals, yeast and bacteria, a gene regulatory network model in *Arabidopsis*was characterized by hierarchical relationships among regulators, with top-level master regulator TFs controlling expression of lower order TFs that in turn regulate expression of structural genes encoding proteins involved in building new cell components [5,24]. Additional regulatory features such as *feed forward loops* are also shared between the *Arabidopsis*and other gene regulatory networks [24]. Together these results suggest that concepts and approaches developed for modeling gene-based regulatory networks in model eukaryotic and prokaryotic species can be extended to plants [19]. We are exploring how to further extend gene regulatory network concepts and approaches further still, to undomesticated tree species.

Developmental processes in trees are complex, and are controlled through the interactions of thousands of genes [23]. Network modeling concepts and approaches thus hold promise to transcend simple gene-gene interactions to more comprehensively and realistically describe functional processes. Advances in network science and network analytics allow the understanding of gene networks in terms of their modularity [1,13], stability [6,26], controllability [18], and other emergent properties [2,3], some of which can be directly mapped onto phenotypic traits [9,21]. However, the difficulty has been in learning those networks in forest trees and other non-model organisms. Indeed for forest trees, there are thousands of species of ecological or economic importance, presenting the additional challenge of understanding how traits of interest vary across many species.

Advances in sequencing technologies now make gene expression and co-expression analyses tractable in non-model organisms. Projects such as oneKP are generating transcriptome assemblies for 1,000 undomesticated plants across the plant kingdom [20]. A major challenge is how to now leverage these types of data sets from individual and even multiple species to enable new analysis that provide insight into the regulation of complex traits both within and across species.

Co-expression networks provide a framework for integrating data types from multiple experiments [22,25]. Conceptually, gene transcript levels assayed across multiple conditions and developmental tissues can be clustered into groups of

genes (i.e. modules) that show highly correlated expression. Overlaying such modules with functional annotations (e.g. Gene Ontology [7]) and correlations with phenotypes can provide insight into the biological pathways influencing developmental processes and the dissection of complex traits [29].

Calculating co-expression networks across species within a phylogenetic framework presents additional opportunities and challenges. On the one hand, the analysis of co-expression modules that show specific phylogenetic relationship can be extremely informative. For example, co-expression modules that are conserved across wide phylogenetic distance could represent signal from evolutionarily ancient regulatory mechanisms. On the other hand, these analyses are conceptually and computationally demanding. For example, the homologous relationships among genes must be established across all species being analyzed, and can include one to one, one to many, or many to many relationships among orthologs. A very practical challenge is thus to establish both orthologous relationships as well as co-expression relationships across multiple species in a computationally tractable way.

2 Our Recent Work on Trees

Plants modify their growth and development in response to external stimuli. As trees grow, they integrate environmental and developmental signals using complex but poorly defined transcriptional gene networks, allowing trees to produce woody tissues appropriate to diverse environmental conditions. Here, we summarize two of our recent studies that illustrate the different approaches we have undertaken to elucidate gene networks of forest trees [28,29].

2.1 Recent Study #1: Time Series of Gene Expression

Plants respond to gravity to produce new growth that is properly oriented in space. For example, gravistimulation of leaning stems in angiosperm trees such as *Populus* results in modifications of wood development, to produce tension wood that pulls leaning stems upright (see Fig. 1, top row) [12]. This response provides an experimental system to perturb gene expression, and can be temporally calibrated against tissue development and stem movements.

In our recent work, we used gravistimulation and tension wood response to dissect the temporal changes in gene expression underlying wood formation in *Populus* stems [28].

Using time series analysis of transcriptome sequences at seven time points over a 14-day experiment, we identified 8,919 genes that were differentially expressed between tension wood (upper) and opposite wood (lower) sides of leaning stems. Clustering of differentially expressed genes showed four major transcriptional responses, including gene clusters whose transcript levels were associated with two types of tissue-specific impulse responses that peaked at about 24 to 48 h, and gene clusters with sustained changes in transcript levels that persisted until the end of the 14-day experiment. Our approach is illustrated in Fig. 1.

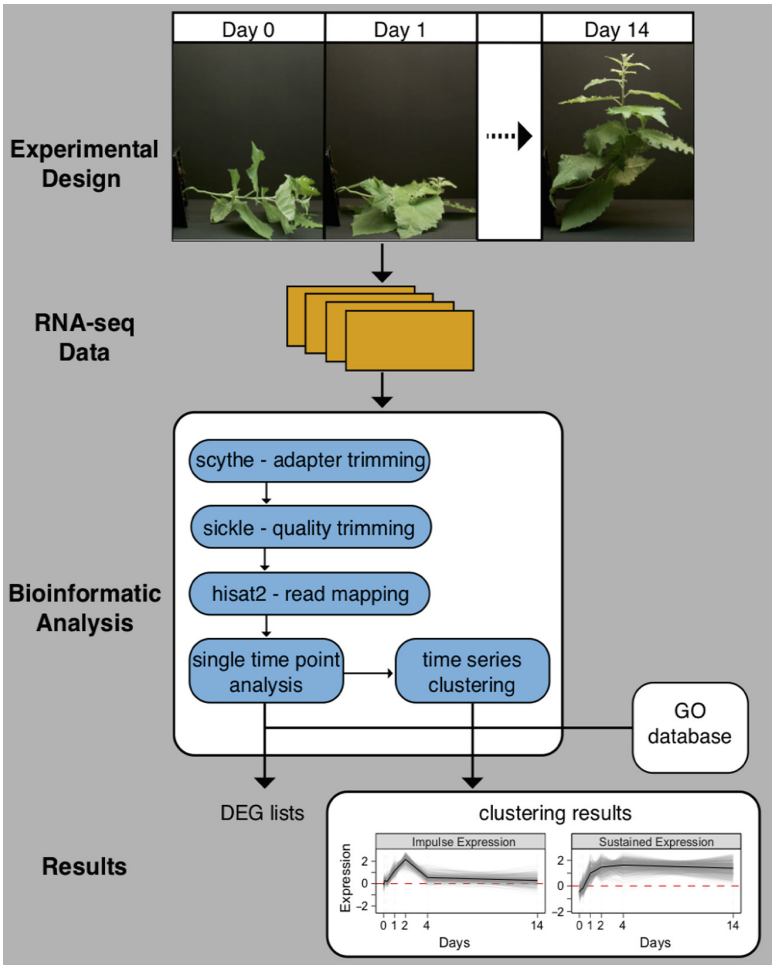

Fig. 1. The steps in our experimental and computational analysis of gene expression of different types of woody tissue in poplar trees, over 14 days [28]. The results include lists of differentially expressed genes (DEGs) over time, clusters of genes coexpressed over time, and functional annotation of those clusters from the GO database [7].

Functional enrichment analysis of those clusters suggested they reflect temporal changes in pathways associated with hormone regulation, protein localization, cell wall biosynthesis and epigenetic processes. Time series analysis of gene expression is an underutilized approach for dissecting complex developmental responses in plants, and can reveal gene clusters and mechanisms influencing development.

2.2 Recent Study #2: Integrating Data from Different Sources

Trees modify wood formation through integration of environmental and developmental signals in complex but poorly defined transcriptional networks, allowing trees to produce woody tissues appropriate to diverse environmental conditions. Basic, conceptual questions include whether multiple environmental inputs impinge upon common regulatory mechanisms, or if individual environmental inputs connect directly to independent regulatory mechanisms controlling growth.

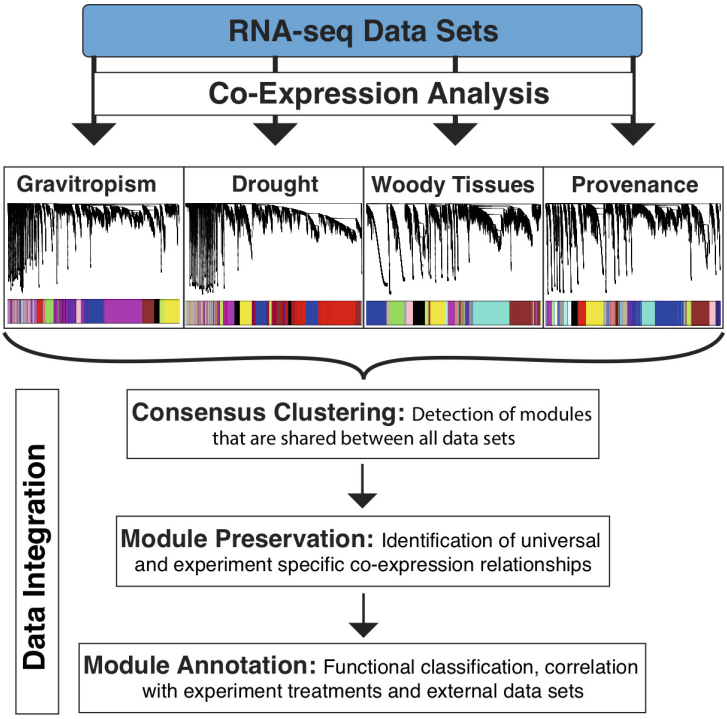

Fig. 2. Flow chart depicting the experimental approach for modeling co-expression networks underlying wood development using data integration and consensus clustering. Original figure appeared in New Phytologist [29].

In order to identify relationships among genes expressed during wood formation, in another recent study [29] we integrated data from our own and publicly available data sets in *Populus*. These data sets were generated from woody tissues and include transcriptome profiling, transcription factor binding, DNA accessibility and genome-wide association mapping experiments. Co-expression modules were calculated, each of which contains genes showing similar expression patterns across experimental conditions, genotypes and treatments. Conserved

gene co-expression modules (four modules totaling 8,398 genes) were identified that were highly preserved across diverse environmental conditions and genetic backgrounds. Figure 2 illustrates our approach.

Functional annotations as well as correlations with specific experimental treatments associated individual conserved modules with distinct biological processes underlying wood formation, such as cell-wall biosynthesis, meristem development and epigenetic pathways. Module genes were also enriched for DNase I hypersensitivity footprints and binding from four transcription factors associated with wood formation. The conserved modules are excellent candidates for modeling core developmental pathways common to wood formation in diverse environments and genotypes, and serve as testbeds for hypothesis generation and testing for future studies.

3 Toward Higher Resolution Network Inference: Multi-species Approaches

More systemically, comparison of co-expression networks across multiple species can identify gene modules in common to all of them (i.e. ancestral) and modules unique to specific lineages. This approach allows additional inference of the evolutionary history and other features of co-expression modules through consideration of their orthologous features. Functional annotation of modules and association with phenotypic traits that vary across the lineages surveyed can provide important insights into the evolutionary history of modules, associated molecular mechanisms, and traits that they regulate. This type of comparison is emerging as an important task in systemic analyses, with existing approaches developed to align protein-protein interaction networks [10,14,17] and co-expression networks [27] across species.

Computationally, there are some important challenges that need to be addressed to allow adoption of evolutionary co-expression network approaches. Network comparison is computationally expensive because of the tens of thousands of genes expressed within each species and the super-linear growth of possible relationships among them. Similarly, determination of orthologous relationships across species is a challenging task. Additionally, there can be complications working with multiple non-model species, for which high quality transcriptomes must be established without the aid of guiding genomic sequence.

An existing state-of-the art approach, OrthoClust [27], e.g., is prohibitively slow on mammalian or plant size genomes, and can only work with two species at a time. Here we present a modified approach, fastOC, that eliminates those limitations and enables the construction and comparison of co-expression networks of multiple tree species, in tractable time and on typical hardware.

3.1 Results

We implemented fastOC, an extension of OrthoClust v1.0 [27], which works on two or more species, and is orders of magnitude faster. Both packages use a

multi-layer network approach to compute co-expression networks across multiple species. These methods cluster genes based on the correlation of expression patterns within species and align networks across species using orthologous gene relationships. The within species co-expression relationships are defined by each gene and the top-N gene neighbors based on Pearson correlations.

The interactions of genes across species are defined as an orthologous weight that accounts for complex orthologous relationships, such as one-to-one, one-to-many and many-to-many relationships. The main difference between fastOC and OrthoClust v1.0, is that fastOC performs clustering of the entire multi-layer network using the Louvain community detection algorithm [4], instead of the original approach of simulated annealing. The Louvain algorithm is a heuristic approach that assigns genes (nodes) to communities to optimize community modularity. Using many Louvain runs, we calculate how often genes co-appear in the same Louvain communities, and identify both the gene modules conserved across species and those unique to specific lineages. Gene modules represent groups of genes that display high co-appearance in Louvain communities. We provide functions for module detection using dynamic tree cutting of hierarchical dendrograms [16]. Figure 3 illustrates the main steps in the fastOC program.

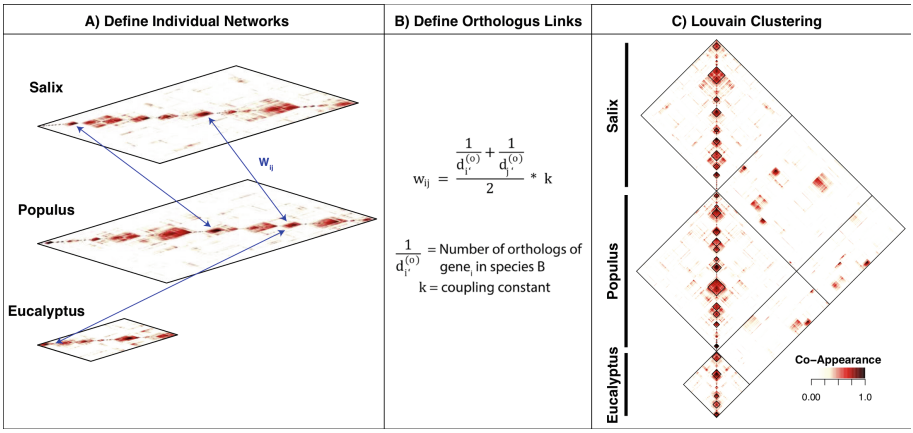

Fig. 3. The three steps in the cross-species co-expression analysis in fastOC.

We implemented fastOC as an R package and the source code is available at https://github.com/mzinkgraf/fastOC under a GPL-3 license. The package includes functions for (1) generating co-expression networks from expression data, (2) calculating orthologous weights, (3) summarizing expression of module genes, and (4) visualization of results. In addition, we developed functions to include parallel processing options to leverage the multi-core functionality of many modern computers. For example, the Louvain clustering function allows for parallel processing using the foreach and doParallel packages in R, to further increase computational efficiency and decrease run times.

Table 1 shows the comparison of running times between our tool, fastOC, and the existing OrthoClust using a single thread on typical hardware. In addition to being between two and three orders of magnitude faster, our examination of the results showed that fastOC finds similar modules in the worm-fly data set as those described in Yan et al. [27]. Specifically, the modules found by fastOC had statistically significant overlap with those found by OrthoClusts ($\chi^2 = 711.8$; $df = 462$; p-value $= 6.244 \times 10^{-13}$).

Table 1. Comparison of run times for OrthoClust and fastOC. Run times for each method were calculated using 20 runs on a single CPU thread on MacBook Pro with a 3.0 GHz Intel Core i7, and 16 GB RAM. The Simulated and Worm-Fly data sets were obtained from https://github.com/gersteinlab/OrthoClust/. The 3 Woody Species data set is available at https://github.com/mzinkgraf/fastOC/.

	Simulated	Worm vs. Fly	3 Woody Species
# Genes	800	34K	83K
# Edges	13K	147K	625K
OrthoClust v.1	6.5 s	2,308.3 s	Not implemented
fastOC	0.1 s	1.8 s	7.6 s

4 Conclusion

Our past work demonstrated that combining different data types can help resolve gene networks, as can time series gene expression experiments. However, the sizes of resulting functional modules can still be large.

Here we described a tool, fastOC, to complement those approaches by enabling co-expression analyses across related species. fastOC is an almost real-time interactive tool for multi-species co-expression network comparison. It can make possible studies of the evolution of genomic function based on how genes interact in complex regulatory networks. Furthermore, fastOC can efficiently work with more than two species. As of this writing, we are using this tool on 13 tree species containing 291,375 genes and 4,642,738 edges in a multi-layer network.

References

1. Barabasi, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nat. Rev. Genet. **5**(2), 101 (2004)
2. Bergman, A., Siegal, M.L.: Evolutionary capacitance as a general feature of complex gene networks. Nature **424**(6948), 549 (2003)
3. Bhalla, U.S., Iyengar, R.: Emergent properties of networks of biological signaling pathways. Science **283**(5400), 381–387 (1999)

4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech.: Theory Exp. **2008**(10), P10008 (2008)
5. Brady, S.M., et al.: A stele-enriched gene regulatory network in the Arabidopsis root. Mol. Syst. Biol. **7**(1), 459 (2011)
6. Ciliberti, S., Martin, O.C., Wagner, A.: Robustness can evolve gradually in complex regulatory gene networks with varying topology. PLoS Comput. Biol. **3**(2), e15 (2007)
7. Gene Ontology Consortium: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. **32**(Suppl_1), D258–D261 (2004)
8. Djebali, S., et al.: Landscape of transcription in human cells. Nature **489**(7414), 101 (2012)
9. Ellis, T., Wang, X., Collins, J.J.: Diversity-based, model-guided construction of synthetic gene networks with predicted functions. Nat. Biotechnol. **27**(5), 465 (2009)
10. Flannick, J., Novak, A., Srinivasan, B.S., McAdams, H.H., Batzoglou, S.: Graemlin: general and robust alignment of multiple large interaction networks. Genome Res. **16**(9), 1169–1181 (2006)
11. Gerstein, M.B., et al.: Architecture of the human regulatory network derived from ENCODE data. Nature **489**(7414), 91 (2012)
12. Gerttula, S., et al.: Transcriptional and hormonal regulation of gravitropism of woody stems in Populus. Plant Cell **27**, 2800–2813 (2015). pp. tpc-15
13. Han, J.D.J., et al.: Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Nature **430**(6995), 88 (2004)
14. Kalaev, M., Bafna, V., Sharan, R.: Fast and accurate alignment of multiple protein networks. In: Vingron, M., Wong, L. (eds.) RECOMB 2008. LNCS, vol. 4955, pp. 246–256. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78839-3_21
15. Kellis, M., et al.: Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. **111**(17), 6131–6138 (2014)
16. Langfelder, P., Zhang, B., Horvath, S.: Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics **24**(5), 719–720 (2007)
17. Liao, C.S., Lu, K., Baym, M., Singh, R., Berger, B.: IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics **25**(12), i253–i258 (2009)
18. Liu, Y.Y., Slotine, J.J., Barabási, A.L.: Controllability of complex networks. Nature **473**(7346), 167 (2011)
19. Long, T.A., Brady, S.M., Benfey, P.N.: Systems approaches to identifying gene regulatory networks in plants. Annu. Rev. Cell Dev. Biol. **24**, 81–103 (2008)
20. Matasci, N., et al.: Data access for the 1,000 plants (1KP) project. GigaScience **3**(1), 17 (2014)
21. Schadt, E.E.: Molecular networks as sensors and drivers of common human diseases. Nature **461**(7261), 218 (2009)
22. Serin, E.A., Nijveen, H., Hilhorst, H.W., Ligterink, W.: Learning from co-expression networks: possibilities and challenges. Front. Plant Sci. **7**, 444 (2016)
23. Shinozaki, K., Yamaguchi-Shinozaki, K.: Gene networks involved in drought stress response and tolerance. J. Exp. Bot. **58**(2), 221–227 (2007)
24. Taylor-Teeples, M., et al.: An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature **517**(7536), 571 (2015)
25. Usadel, B., et al.: Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ. **32**(12), 1633–1651 (2009)

26. Von Dassow, G., Meir, E., Munro, E.M., Odell, G.M.: The segment polarity network is a robust developmental module. Nature **406**(6792), 188 (2000)
27. Yan, K.K., Wang, D., Rozowsky, J., Zheng, H., Cheng, C., Gerstein, M.: OrthoClust: an orthology-based network framework for clustering data across multiple species. Genome Biol. **15**(8), R100 (2014)
28. Zinkgraf, M., Gerttula, S., Zhao, S., Filkov, V., Groover, A.: Transcriptional and temporal response of Populus stems to gravi-stimulation. J. Integr. Plant Biol. (2018). https://doi.org/10.1111/jipb.12645
29. Zinkgraf, M., Liu, L., Groover, A., Filkov, V.: Identifying gene coexpression networks underlying the dynamic regulation of wood-forming tissues in Populus under diverse environmental conditions. New Phytol. **214**(4), 1464–1478 (2017)