Gene Co-expression Network Design from RNA-seq data in Arabidopsis thaliana

Ximena Contreras, Daniel Nichol, Michelle Parker

August 17, 2012
Introduction

Project Overview

1. Clean and process RNA-Seq data collected from the plant *Arabidopsis thaliana*
2. Construct weighted gene co-expression network
3. Identify groups of highly interconnected genes
4. Assess statistical significance of these modules
5. Associate modules with biological functions through Gene Ontology annotation
Data Cleaning and Processing

Data

- Recombinant inbred lines of the MAGIC population: 128 samples in which there are 21,359 transcripts per sample.
- Calculated expression levels for all genes using the Cufflinks transcript assembly software
- We removed genes:
  1. with an expression level of zero across all samples
  2. with at least one missing sample value.
- We removed samples which had significantly different gene expression values than the rest by using average linkage hierarchical clustering algorithm
Sample clustering to detect outliers
Network Construction

Similarity Matrix

- Similarity matrix $S = [s_{ij}]$
- Measures the level of concordance between gene expression profiles across the experiments.
- Unsigned

\[ s_{ij} = |\text{cor}(i, j)| \]

- The resulting symmetric matrix has values between 0 and 1.
Network Construction

**Adjacency Function**

- Adjacency matrix $A = [a_{ij}]$
- Transforms co-expression similarities into connection strengths.
- Soft Thresholding (Weighted network)

  \[ a_{ij} = |s_{ij}|^\beta \]

- Continuous nature of the co-expression information is retained as network has weighted edges
- The resulting symmetric matrix has values between 0 and 1.
Scale-free topology criterion is an important characteristic of many biological and social networks.

The probability that a node is connected with \( k \) other nodes decays as a power law.

\[
p(k) \sim k^{-\gamma}
\]

The parameter values must generate scale-free topology networks.

Plot \( \log_{10}(p(k)) \) versus \( \log_{10}k \).
Network Construction

Scale independence

Mean connectivity

Scale Free Topology Model Fit, signed $R^2$

Soft Threshold (power)

Mean Connectivity

Soft Threshold (power)
Module Detection

To find modules we will define a measure of similarity between the expressions of pairs of genes

Let $A = [a_{i,j}]$ be the weighted adjacency matrix. Given two genes $i$ and $j$ we define a measure of topological overlap by

$$\omega_{i,j} = \frac{l_{i,j} + a_{i,j}}{\min\{k_i, k_j\} + 1 - a_{i,j}}$$

Where

$$l_{i,j} = \sum_u a_{i,u}a_{u,j}$$

$$k_i = \sum_u a_{i,u}$$

$\omega_{i,j}$ gives a measure of how closely the genes $i$ and $j$ are clustered together.
Module Detection

Topological Overlap Heatplot
Module Detection
Eigengene Analysis

- Each module was summarized by a single gene known as an *eigengene*.

- The eigengene accounts for as much of the variability in the data as possible and is defined as the first principal component of the module expression matrix.

Module Merging

- The average linkage hierarchical clustering method is used to cluster the 50 modules by their eigengene similarity.

- Modules are merged if they have an expression profile correlation greater than 0.8.

- This resulted in 33 modules.
Eigengene Analysis

Cluster Dendrogram

hclust (*, "average")

Height

50 Modules

Merged Modules
Eigengene Analysis

Cluster Dendrogram

Dynamic Tree Cut

Merged dynamic
Statistical Significance of Modules

The permutation test was used in order to test the significance of each of the modules:

▶ 100 random networks of the same size as each of the modules was synthesized and the values on the edges were permuted.

▶ Other network properties such as mean connectivity, mean scaled connectivity, density, heterogeneity, centralization and mean clustering coefficient were calculated.

Results

▶ Density, heterogeneity, centralization and mean clustering coefficient of each module lie over 3 standard deviations away from the mean of the distribution of the sampled networks.

▶ Mean clustering coefficient of each module was extremely different from the sample networks; lying over 10 standard deviations away from the mean of the samples.
Gene Ontology: Biological Process

- Functional enrichment analysis to test whether the modules identified are biologically meaningful
- amiGO term enrichment webtool and Gene Ontology annotation database
- Found statistically significant (p-value < 0.01) shared GO terms used to describe the hub genes for each module
- Discovered 10 independent modules with statistically significant association to a biological process
  1. Global processes: translation, programmed cell death, circadian rhythm and nutrient reservoir activity
  2. Plant related processes: response to light, photorespiration, fungi response, UV-B radiation response
Gene Ontology: Biological Process
Gene Ontology: Biological Process

Table 1: A table showing the different network measures for those modules with and without known functionality

<table>
<thead>
<tr>
<th>Name</th>
<th>Mean Connectivity</th>
<th>Mean Clust Coef</th>
<th>Density</th>
<th>Centralization</th>
<th>Heterogeneity</th>
<th>Number of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photosynthesis</td>
<td>1.130</td>
<td>0.048</td>
<td>0.010</td>
<td>0.038</td>
<td>1.074</td>
<td>112</td>
</tr>
<tr>
<td>Programmed Cell Death</td>
<td>1.160</td>
<td>0.037</td>
<td>0.013</td>
<td>0.026</td>
<td>0.807</td>
<td>91</td>
</tr>
<tr>
<td>Glycosinolate Biosynthesis</td>
<td>1.490</td>
<td>0.102</td>
<td>0.022</td>
<td>0.066</td>
<td>1.090</td>
<td>69</td>
</tr>
<tr>
<td>Light Response</td>
<td>0.760</td>
<td>0.021</td>
<td>0.004</td>
<td>0.019</td>
<td>1.020</td>
<td>200</td>
</tr>
<tr>
<td>Translation</td>
<td>0.260</td>
<td>0.019</td>
<td>0.002</td>
<td>0.015</td>
<td>1.330</td>
<td>109</td>
</tr>
<tr>
<td>Fungi Response</td>
<td>1.050</td>
<td>0.094</td>
<td>0.012</td>
<td>0.039</td>
<td>1.110</td>
<td>89</td>
</tr>
<tr>
<td>Nutrient Reservoir Activity</td>
<td>2.590</td>
<td>0.318</td>
<td>0.041</td>
<td>0.143</td>
<td>1.310</td>
<td>54</td>
</tr>
<tr>
<td>Photorespiration</td>
<td>1.510</td>
<td>0.033</td>
<td>0.010</td>
<td>0.029</td>
<td>0.862</td>
<td>152</td>
</tr>
<tr>
<td>Circadian Rhythm</td>
<td>0.574</td>
<td>0.028</td>
<td>0.010</td>
<td>0.024</td>
<td>0.786</td>
<td>57</td>
</tr>
<tr>
<td>UV-B radiation Response</td>
<td>0.704</td>
<td>0.171</td>
<td>0.012</td>
<td>0.059</td>
<td>1.400</td>
<td>58</td>
</tr>
<tr>
<td>bisque4</td>
<td>1.002</td>
<td>0.512</td>
<td>0.026</td>
<td>0.104</td>
<td>1.626</td>
<td>40</td>
</tr>
<tr>
<td>brown4</td>
<td>0.337</td>
<td>0.226</td>
<td>0.008</td>
<td>0.060</td>
<td>1.891</td>
<td>42</td>
</tr>
<tr>
<td>darkolivegreen</td>
<td>0.500</td>
<td>0.070</td>
<td>0.007</td>
<td>0.039</td>
<td>1.301</td>
<td>69</td>
</tr>
<tr>
<td>darkorange2</td>
<td>0.538</td>
<td>0.261</td>
<td>0.013</td>
<td>0.070</td>
<td>1.634</td>
<td>43</td>
</tr>
<tr>
<td>darkslateblue</td>
<td>0.402</td>
<td>0.264</td>
<td>0.010</td>
<td>0.078</td>
<td>2.085</td>
<td>40</td>
</tr>
<tr>
<td>floralwhite</td>
<td>0.484</td>
<td>0.158</td>
<td>0.011</td>
<td>0.064</td>
<td>1.731</td>
<td>208</td>
</tr>
<tr>
<td>ivory</td>
<td>0.746</td>
<td>0.396</td>
<td>0.016</td>
<td>0.093</td>
<td>1.919</td>
<td>47</td>
</tr>
<tr>
<td>lightcyan1</td>
<td>1.100</td>
<td>0.453</td>
<td>0.022</td>
<td>0.118</td>
<td>1.982</td>
<td>50</td>
</tr>
<tr>
<td>lightsteelblue1</td>
<td>0.405</td>
<td>0.112</td>
<td>0.008</td>
<td>0.054</td>
<td>1.964</td>
<td>51</td>
</tr>
<tr>
<td>lightyellow</td>
<td>0.449</td>
<td>0.023</td>
<td>0.004</td>
<td>0.020</td>
<td>1.209</td>
<td>127</td>
</tr>
<tr>
<td>mediumpurple3</td>
<td>0.323</td>
<td>0.260</td>
<td>0.006</td>
<td>0.052</td>
<td>2.143</td>
<td>52</td>
</tr>
<tr>
<td>paleturquoise</td>
<td>0.723</td>
<td>0.281</td>
<td>0.010</td>
<td>0.044</td>
<td>1.433</td>
<td>72</td>
</tr>
<tr>
<td>plum1</td>
<td>0.238</td>
<td>0.085</td>
<td>0.004</td>
<td>0.024</td>
<td>1.412</td>
<td>56</td>
</tr>
<tr>
<td>plum2</td>
<td>0.434</td>
<td>0.247</td>
<td>0.013</td>
<td>0.082</td>
<td>1.792</td>
<td>34</td>
</tr>
<tr>
<td>saddlebrown</td>
<td>0.996</td>
<td>0.187</td>
<td>0.013</td>
<td>0.078</td>
<td>1.604</td>
<td>75</td>
</tr>
<tr>
<td>sienna3</td>
<td>0.901</td>
<td>0.243</td>
<td>0.014</td>
<td>0.066</td>
<td>1.502</td>
<td>65</td>
</tr>
<tr>
<td>skyblue</td>
<td>0.762</td>
<td>0.042</td>
<td>0.010</td>
<td>0.029</td>
<td>0.959</td>
<td>76</td>
</tr>
<tr>
<td>tan</td>
<td>4.150</td>
<td>0.353</td>
<td>0.027</td>
<td>0.133</td>
<td>1.820</td>
<td>155</td>
</tr>
<tr>
<td>violet</td>
<td>0.474</td>
<td>0.307</td>
<td>0.007</td>
<td>0.051</td>
<td>1.957</td>
<td>69</td>
</tr>
</tbody>
</table>
Conclusions

- Built a weighted gene coexpression network for *A. thaliana*, an organism for which such a network has not been previously built.
- Used co-expression values found using RNA-Seq technology which as a new technology has not been used in coexpression network construction previously.
- Able to identify 33 statistically significant and independent modules.
- Able to identify the biological functionality for ten of them.
  1. Global processes: translation, programmed cell death, circadian rythm and nutrient reservoir activity
  2. Plant related processes: response to light, photorespiration, fungi response, UV-B radiation response
Conclusions

- 23 of the modules were significant by the permutation test but were not biologically significant.
- T-test: modules with functionality have a significantly lower ($p<0.01$) mean clustering coefficient than those without.
- This suggests that topological overlap may not be a sufficient similarity measure to use for WGCNA a more accurate method which merges several network properties could be developed.
Future Work

- Use samples from *A. thaliana* under different environmental conditions to build co-expression networks and validate those modules with associated functionality.
- Identify eQTLs for gene hubs from modules with biological function.
- Find regulatory signals for gene hubs.
- Compare these networks with other from different species.
- Develop and test clustering methods accounting for more measures than only the topological overlap.
Thank you to the following people for their help and support with this project:

- Martha Imprialou
- James Anderson
- OSSCB12
- Jotun Hein
- Richard Mott
- University of Oxford Plant Sciences Department
- University of Oxford Statistics Department
References


