Integrated WGCNA with an application to chronic fatigue syndrome

Angela Presson, Steve Horvath
Department of Biostatistics
Department of Human Genetics
University of California, Los Angeles
Chronic Fatigue Syndrome

6 Months or more of medically unexplained severe fatigue + 4 of the following symptoms:

Post exertional fatigue lasting > 24 hrs
Unrefreshing sleep
Difficulty concentrating or remembering
Headaches unusual in frequency or duration
Muscle pain
Joint pain
Sore throat
Tender lymph nodes
(Fukuda et al. 1994)
Outline

1. Demonstrate a weighted gene co-expression network analysis (WGCNA)
   a. Screen for ~20 candidate genes to consider for follow up analysis using:
      i. SNP
      ii. Severity
      iii. Module connectivity
   b. Check for biological relevance of module and candidate genes using gene ontology software.

2. Conduct a standard microarray analysis using the false discovery rate and ignoring the SNP data. Identify ~20 candidate genes and annotate.

3. Compare standard analysis with WGCNA.
Chronic Fatigue data set

164 Samples with the following data:

**DNA Level**: ~ 36 Pre-selected autosomal SNP’s

**mRNA Level**: ~ 20K genes/array

**Organism Level**: ~ 70 Clinical Traits

Analyzed the following subset of data:

1.) 127 fatigued samples
2.) 8966 genes with high mean and variance
Selecting a clinical trait

• Scores from diagnostic procedures used to evaluate quality of life:
  • Medical Outcomes Survey Short Form (SF-36)
  • Multidimensional Fatigue Inventory (MFI)
  • CDC Symptom Inventory Case Definition scales

• Reeves et al. (2005) clustered these 14 scores from 118 patients and identified three clusters of CFS severity: high, moderate and low.

➢ Out of 70 clinical scores, we chose to use this CFS severity trait.
Selecting a SNP marker

- CDC provided 36 autosomal SNPs from 8 candidate CFS genes: TPH2, POMC, NR3C1, CRHR2, TH, SLC6A4, CRHR1, COMT (Smith et al. 2006)

- Selected “TPH2 SNP”: rs10784941 (12q21) from the TPH2 gene because:
  - Previously found to be associated with chronic fatigue (Goertzel et al. 2006)
  - Had significant correlation with CFS severity (p-value = 0.0099).
a. IWGCNA steps

1. Construct a co-expression network and modules
2. Find clinical trait related modules
3. Prioritize module genes using disease-related SNP
4. Use the integrated model to screen for genes

b. CFS Patients analyzed in IWGCNA steps

1. Full data set: 127 CFS patients
2. Primary data set: 87 patient subset with CFS severity scores
   - Exclude heterogeneous female samples
3-5. 76 samples with CFS severity scores
Constructing a weighted gene co-expression network

1. Construct a Pearson correlation matrix from microarray data: $x_i$ and $x_j \rightarrow r(x_i, x_j)$

2. Transform via an adjacency function:
   - Step function: $a_{ij} = I(r(x_i, x_j) > \tau) \rightarrow$ Unweighted network
   - Power function: $a_{ij} = r(x_i, x_j)^\beta \rightarrow$ Weighted network
Five modules identified using hierarchical clustering

- Grey colors indicate genes outside of any module.
- MDS plot indicates separation of blue, green, brown, turquoise and yellow modules.

a) Gene Network

b) MDS view
The blue module relates to severity

\[ GS.\text{severity}(i) = |\text{cor}(x(i), \text{severity})|, \] where \( GS = \) “Gene Significance” and \( x(i) \) is the gene expression profile of the \( i^{th} \) gene. Can also define:

\[ \text{Module.Significance}(k) = E(\text{GS.severity}(i) \text{ genes in module } k) \]

Blue module (299 genes) has highest Module Significance
Correlate gene expression data with \textit{TPH2 SNP}

- Integration of WGCNA with genetic marker data: IWGCNA
  \[ GS.SNP(i) = |\text{cor}(x(i), TPH2 SNP) | \]
  where \( x(i) \) is the \( i \)-th gene expression

- Additive SNP marker coding: AA = 2, AB = 1, BB = 0
- Absolute value of the correlation ensures that this is equivalent to AA = 0, AB = 1, BB = 2
- Dominant or recessive coding is more appropriate for most Mendelian diseases
Why Consider Gender Differences?

• We chose to investigate sex differences for the following reasons:
  1. CFS is 4x more prevalent in women. (Reyes et al. 2003)
  2. Possible that prevalence difference due to genetic differences between genders.

• If no gender difference, analyze male and female arrays together.

• If gender differences, exclude some female samples with expression patterns that differ most from module eigengene.
The blue module is related to severity in males, several modules relate in females.
Homogenization of Female Samples

- Based on the idea that blue module is related to severity. Uses first principal component of blue module: “module eigengene” (ME) summary measure.

- $\text{ME}_{\text{blue}} > \text{median} (\text{ME}_{\text{blue}})$ and high severity (severity $> 1$) OR $\text{ME}_{\text{blue}} < \text{median} (\text{ME}_{\text{blue}})$ and low severity (severity $< 3$).

- Reduced female samples from 64 to 53.

- Increased the module significance from 0.22 (p-value = 0.074) to 0.47 (p-value = 0.00016).
**Gender-stratified network views**

1. Calculate connectivity for a gene $x(i)$: $k_{ME}(i) = |\text{cor(MEblue}, x(i))|$
2. Blue module connectivity (membership) is highly preserved between genders
3. Less preservation for GS.severity
   - Due to GS.severity gender difference, it is useful to impose screening criteria in both males and females separately.
Gene screening procedure

Screening criteria imposed in both males and homogenized females:

1) High connectivity within blue module \((k_{ME} \text{ in top 2/3}'s)\)

2) Association with severity trait \((GS_{\text{severity}} > .2 \text{ in males and } GS_{\text{severity}} > .35 \text{ in homogenized females})\)

3) Association with TPH2 SNP (top 50%)

⇒ 20 Genes met these criteria
**IWGCNA Candidate Genes**

- 12/16 genes were a) verified as interacting and b) estimated to function in a hematological disease pathway by Ingenuity Pathways Analysis (IPA) software
- **Viral function, hematological disease and connective tissue** are consistent with previous findings.

<table>
<thead>
<tr>
<th>Gene Name and Genbank Accession</th>
<th>Full gene name, Entrez Gene and/or GeneRIFs description, Chromosome Location</th>
<th>Ingenuity Pathways Gene Annotation</th>
<th>c) Causality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FOXN1 (NM_003593)</strong></td>
<td>Forkhead box N1. Mutations result in a severely compromised immune system, T-cell immunodeficiency, skin disorder congenital alopecia. 17q11-q12</td>
<td>Hematological Disease¹&lt;br&gt;Rank = 1, p-value ≈ 10⁻³²</td>
<td><strong>Cell Function⁵</strong>&lt;br&gt;Rank = 8, p-value ≈ 10⁻¹⁶&lt;br&gt;0.82&lt;br&gt;6</td>
</tr>
<tr>
<td><strong>PRDX3 (AF118073)</strong></td>
<td>Peroxiredoxin 3. Antioxidant function, regulates abundance of H₂O₂, which promotes apoptosis. 10q25-q26</td>
<td>Hematological Disease¹</td>
<td><strong>Endocrine Disorders/Inflammation⁶</strong>&lt;br&gt;Rank = 6, p-value ≈ 10⁻²⁰&lt;br&gt;0.77&lt;br&gt;8</td>
</tr>
<tr>
<td><strong>SUCLA2 (AK001458)</strong></td>
<td>Succinate-CoA ligase, ADP-forming, beta subunit. Defects associated with encephalomyopathy. 13q12.2-q13.3</td>
<td>Hematological Disease¹</td>
<td><strong>Cell Cycle⁷</strong>&lt;br&gt;Rank = 5, p-value ≈ 10⁻²²&lt;br&gt;0.77&lt;br&gt;9</td>
</tr>
<tr>
<td><strong>TFB2M (AK026314)</strong></td>
<td>Transcription factor B2, mitochondrial. 1q44</td>
<td>Hematological Disease¹</td>
<td><strong>Cell Cycle⁷</strong>&lt;br&gt;0.69&lt;br&gt;18</td>
</tr>
<tr>
<td><strong>MED8 (BC010019)</strong></td>
<td>Mediator complex subunit 8</td>
<td>Hematological Disease¹</td>
<td><strong>Amino Acid Met.⁸</strong>&lt;br&gt;Rank = 1, p-value ≈ 10⁻¹⁵&lt;br&gt;0.82&lt;br&gt;7</td>
</tr>
<tr>
<td><strong>SNURF (AF101044)</strong></td>
<td>SNRPN upstream reading frame. Alternative splicing/deletion leads to Angelman syndrome or Prader-Willi. 15q12</td>
<td>Hematological Disease¹</td>
<td><strong>Amino Acid Met.⁸</strong>&lt;br&gt;0.53&lt;br&gt;36</td>
</tr>
<tr>
<td><strong>DCTN2 (NM_006400)</strong></td>
<td>Dynactin 2 (p50), Required in peroxisome biogenesis. 12q13.2-q13.3</td>
<td>Hematological Disease¹</td>
<td><strong>Amino Acid Met.⁸</strong>&lt;br&gt;0.30&lt;br&gt;66</td>
</tr>
<tr>
<td><strong>PGK1 (AB082432)</strong></td>
<td>Phosphoglycerate kinase 1. Glycolysis. Xq13</td>
<td>Hematological Disease¹</td>
<td><strong>Amino Acid Met.⁸</strong>&lt;br&gt;-0.28&lt;br&gt;132</td>
</tr>
<tr>
<td><strong>PRKCH (BC001000)</strong></td>
<td>Protein kinase C, eta. Regulates keratinocyte differentiation. 14q22-q23</td>
<td>Hematological Disease¹</td>
<td><strong>Connective Tissue⁸</strong>&lt;br&gt;Rank = 2, p-value ≈ 10⁻¹⁴&lt;br&gt;-0.13&lt;br&gt;116</td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Function</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>RYK (NM_002958)</td>
<td>RYK receptor-like tyrosine kinase. May play a role in the development of cleft lip and/or palate. 3q22</td>
<td>Hematological Disease&lt;sup&gt;1&lt;/sup&gt; Connective Tissue&lt;sup&gt;9&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>VAMP5 (AF077197)</td>
<td>Vesicle-associated membrane protein 5 (myobrevin). Associated with myogenesis. 2p11.2</td>
<td>Hematological Disease&lt;sup&gt;1&lt;/sup&gt; Connective Tissue&lt;sup&gt;10&lt;/sup&gt; Rank = 7, p-value ≈ 10&lt;sup&gt;-18&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>PBLD (AK027673)</td>
<td>Phenazine biosynthesis-like protein domain containing. 10pter-q25.3</td>
<td>Hematological Disease&lt;sup&gt;1&lt;/sup&gt; Connective Tissue&lt;sup&gt;10&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>NPAL2 (AK024017)</td>
<td>NIPA-like domain containing 2. 8q22.2</td>
<td>Digestive System&lt;sup&gt;2&lt;/sup&gt; VIRAL FUNCTION&lt;sup&gt;11&lt;/sup&gt; Rank = 3, p-value ≈ 10&lt;sup&gt;-32&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>CD302 (BC020646)</td>
<td>C-type lectin receptor involved in cell adhesion and migration, as well as endocytosis and phagocytosis. 2q24.2</td>
<td>Carbohydrate Metabolism&lt;sup&gt;3&lt;/sup&gt; VIRAL FUNCTION&lt;sup&gt;11&lt;/sup&gt; Rank = 3, p-value ≈ 10&lt;sup&gt;-32&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>PPP1R14C (AF407165)</td>
<td>Protein phosphatase 1, regulatory (inhibitor) subunit 14C. Enriched in brain, heart and skeletal muscle. 6q24.3-q25.3</td>
<td>Cancer&lt;sup&gt;4&lt;/sup&gt; Cell Proliferation&lt;sup&gt;12&lt;/sup&gt; Rank = 2, p-value ≈ 10&lt;sup&gt;-3&lt;/sup&gt; Rank = 9, p-value ≈ 10&lt;sup&gt;-14&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>TMEM50A (AF081282)</td>
<td>Transmembrane protein 50A. May contribute to RH haplotype selection. 1p36.11</td>
<td>NA, Rank = 2 NA, Rank = 14</td>
<td></td>
</tr>
<tr>
<td>CRNKL1 (AF111802)</td>
<td>Crooked neck pre-mRNA splicing factor-like 1. 20p11.2</td>
<td>NA NA</td>
<td></td>
</tr>
<tr>
<td>LTV1 (AK027815)</td>
<td>Protein coding. 6q24.2</td>
<td>NA NA</td>
<td></td>
</tr>
<tr>
<td>AF090939</td>
<td>Discontinued record.</td>
<td>NA NA</td>
<td></td>
</tr>
<tr>
<td>XM13557</td>
<td>Unmapped.</td>
<td>NA NA</td>
<td></td>
</tr>
</tbody>
</table>

<sup>1</sup>Cell Cycle, Cancer, Hematological Disease
<sup>2</sup>Digestive System D&F, Hepatic System D&F, Organ Dev.
<sup>3</sup>Carbohydrate Metabolism, Gene Expression, Genetic Disorder
<sup>4</sup>Cancer, Cellular Movement, Skeletal and Muscular Disorders
<sup>5</sup>Cell Fun. and Main., Small Molecule Biochem., Molecular Transport
<sup>6</sup>Endocrine System Disorders, Infectious Disease, Inflammatory Disease
<sup>7</sup>Cell Assembly and Org., Cell Cycle, DNA Replication/Recomb./Repair
<sup>8</sup>Post-Translational Modification, Amino Acid Metabolism, Molecular Transport
<sup>9</sup>Organ Morphology, Cell Morphology, Connective Tissue D&F
<sup>10</sup>Gene Expression, Cellular Development, Connective Tissue D&F
<sup>11</sup>Viral Function, Cell. Assembly and Org., Cell Fun. and Maintenance
<sup>12</sup>Post-Translational Modification, Cancer, Cellular Growth/Proliferation
<sup>13</sup>LEO.NB.SingleMarker scores (converted to fold changes).
Centrality of candidate gene pathway reflects use of connectivity in gene screening strategy.
Repeat IPA with **TPH2** gene:

Does including **TPH2** SNP in screening procedure result in genes that interact with **TPH2**?

- Yes, it is part of the large pathway.
- The p-value improves slightly and the functions stay the same.
Pathways are very similar

20 + TPH2

20 only
Results from previous CFS studies

1. Associated with other conditions: fibromyalgia, connective tissue disease and mitochondrial deficiency (Bains 2008; Hench 1989)

2. Affects the endocrine, muscular and immune systems and some cases may be triggered by viruses (Lloyd et al. 1991; Holmes et al. 1987; Torpy and Chrousos 1996; Kaushik et al. 1987)

3. Evidence for immune and hypothalamic-pituitary-adrenal (HPA) axis abnormalities have been observed at the symptom, molecular and genetic level of CFS patients (Klimas and Koneru 2007)

4. Higher cytotoxic T-cell counts and impaired T-cell function in CFS patients (Rasmussen et al. 1994; Patarca 2001)

5. Evidence for higher rates of immune cell apoptosis in CFS patients, specifically neutrophils and peripheral blood lymphocytes (Vojdani et al. 1997; Kennedy et al. 2004)
Outline

1. Demonstrate a weighted gene co-expression network analysis (WGCNA)
   a. Screen for ~20 candidate genes to consider for follow up analysis using:
      i. SNP
      ii. Severity
      iii. Module connectivity
   b. Check for biological relevance of module and candidate genes using gene ontology software.

2. Conduct a standard microarray analysis using the false discovery rate and ignoring the SNP data.
   • Identify ~20 candidate genes and annotate.

3. Compare standard analysis with WGCNA.
Standard analysis results in 29 candidate genes

• Starting from 8966 most varying genes, computed p-values for Pearson correlation test of gene expression profiles with severity.

• For each p-value, we computed the corresponding local false discovery rate (q-value) using the qvalue package in R.

• 346 genes achieved minimum fdr = 0.081; and 241 eligible for IPA network construction. Top 3 IPA pathways:

  1. Viral Function, Molecular Transport, RNA Trafficking (p-value \( \sim 10^{-52} \), focus molecules = 29)

  2. Connective Tissue Development and Function, Cell Signaling, Molecular Transport (p-value \( \sim 10^{-31} \), focus molecules = 20)

  3. Cell Morphology, Cellular Assembly and Organization, Cancer (p-value \( \sim 10^{-29} \), focus molecules = 19)

- Selected 29 genes from Viral Function pathway as candidate genes for standard analysis.
IPA of 29 standard analysis genes with and without TPH2

- Analysis of 29 genes alone results in 2 networks.

- TPH2 is not involved in either network.
Outline

1. Demonstrate a weighted gene co-expression network analysis (WGCNA)
   a. Screen for ~20 candidate genes to consider for follow up analysis using:
      i. SNP
      ii. Severity
      iii. Module connectivity
   b. Check for biological relevance of module and candidate genes using gene ontology software.

2. Conduct a standard microarray analysis using the false discovery rate and ignoring the SNP data.
   • Identify ~20 candidate genes and annotate.

3. Compare standard analysis with WGCNA.
IPA network comparison between 20 IWGCNA and 29 standard analysis genes

- No overlap between these two lists.
- But, overlap between Hematological Disease and Viral Function networks.

Light blue = 20 IWGCN genes

Dark blue = 29 standard genes
Correlation results: IWGCNA vs. standard analysis

- $r(\text{IWGCNA, TPH2 SNP}) > r(\text{Std, TPH2 SNP})$
- $r(\text{IWGCNA, MEblue}) > r(\text{Std, MEblue})$
- $r(\text{Std, Severity}) > r(\text{IWGCNA, Severity})$
Conclusions

1. Weighted gene co-expression networks:
   a) Useful for selecting patient samples with similar gene expression profiles.
   b) Can be easily integrated with genetic marker, clinical, and other types of data.

2. Both IWGCNA and a standard analysis of CFS microarray data identify clinically interesting pathways and genes.

3. While the 20 and 29 cg lists do not overlap, IPA finds overlap between networks.

4. Integrating genotypes from a SNP marker with WGCNA identifies candidate genes that:
   • Functionally interact with the SNP-containing gene

5. Whereas a standard analysis excluding SNP data does not find expression correlations with the SNP genotypes nor does the SNP-containing gene interact with these candidate genes.
WGCNA Software:
stand alone and R package

http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork
Acknowledgements

UCLA Human Genetics/Biostatistics

• Steve Horvath
• Eric Sobel
• Jeanette Papp
• Jason Aten
• Charlyn Suarez

Centers for Disease Control

• Suzanne Vernon
• Toni Whistler
References


