Weighted Correlation Network Analysis (WGCNA) & Multiscale Embedded Gene Co-expression Network Analysis (MEGENA)

Brayan Gutierrez

September 2025

Overview

- WGCNA, hdWGCNA, MEGENA
  - Why was it developed
  - Core Concepts
  - Advantages
  - Limitations
  - Relevance

WGCNA

Why WGCNA?

- Traditional clustering methods (like hierarchical clustering or k-means) group genes based on similarity but may miss higher-level *network structure.*
- WGCNA (Weighted Gene Correlation Network Analysis) extends clustering by building a **network** where:
  - Nodes = genes
  - Edges = weighted correlations between expression profiles
- This makes it possible to find **modules** of tightly co-expressed genes rather than isolated clusters.
- **Modules:** functionally related groups of genes, often identified by similar expression patterns, which work together to contribute to a specific cellular function or phenotype

- **Soft-thresholding:**
  - Instead of hard-cutting correlation values (e.g., 0.8 cutoff), WGCNA uses a power function to emphasize strong correlations while keeping weaker ones.
  - Leads to networks that approximate a scale-free topology (a property seen in many biological systems).
- **Modules:**
  - Defined as groups of genes with high topological overlap (not just correlation).
  - Identified via hierarchical clustering of the Topological Overlap Measure (TOM).

Core Concepts Cont.

- **Module eigengenes:**
  - Each module is summarized by its first principal component (the “eigengene”), capturing the main expression pattern of the module.
  - Modules can be correlated with external traits (e.g., AMD vs. control, severity stages, age).

Advantages

- Goes beyond simple grouping, captures relationships between clusters/modules.
- Allows testing how modules relate to biological or clinical variables (e.g., AMD phenotype).
- Identifies hub genes within modules (genes with the highest intramodular connectivity, potential key drivers).
- Applicable to any high-dimensional data (not just gene expression).

Limitations

- Requires relatively large sample sizes for stable module detection.
- Sensitive to preprocessing (normalization, batch effects).
- Module definitions can vary between datasets — often needs consensus networks across cohorts.
- WGCNA is exploratory: modules suggest hypotheses but need biological validation (pathway analysis, wet lab experiments).

Relevance

- Use of different clustering methods (k-means, spectral, hierarchical), WGCNA is a network-based complement:
  - Clustering $\rightarrow$ finds groups of genes with similar expression.
  - WGCNA $\rightarrow$ finds network modules with biological relevance and highlights candidate hub genes that might drive AMD pathways.
- WGCNA can be useful for diseases like AMD where complex pathways and interactions are involved (oxidative stress, complement system, lipid metabolism, angiogenesis).

Relevance Cont.

- Can test whether WGCNA modules associate with:
  - AMD case/control status
  - Disease severity/stages
  - Age (a key covariate)

hdWGCNA

Why hdWGCNA?

- Traditional WGCNA works well for bulk RNA-seq, but struggles with:
  - Single-cell & spatial transcriptomics (sparse, noisy, high-dimensional).
  - Datasets with hundreds of thousands to millions of cells.
- hdWGCNA adapts WGCNA for these data types, scaling network analysis to modern high-throughput transcriptomics.

Core Concepts

- High-dimensional scalability: Efficient enough to handle $\sim$1 million single cells.
- Seurat integration: Built to work seamlessly with the widely used single-cell analysis framework.
- Module detection in single-cell/spatial data: Allows discovery of cell-type-specific or spatially restricted co-expression modules.
- Downstream tools included: Gene set enrichment, statistical testing, visualization.
- Like WGCNA, hdWGCNA constructs gene co-expression networks and identifies modules (clusters of genes with highly correlated expression).

Core Concepts Cont. 1

- Unlike k-means, hierarchical, or spectral clustering, which directly group genes based on similarity, hdWGCNA:
  - Builds a weighted network first.
  - Clusters genes into modules based on topological overlap (shared network connections).
  - Summarizes each module with a module eigengene (main expression pattern).
- This approach helps reveal biological structure that standard clustering might miss, especially in complex diseases.
- Applied to single-cell datasets from neurological disorders (ASD, Alzheimer's).

Core Concepts Cont. 2

- Identified disease-relevant gene modules linked to pathology.
- Demonstrated that hub genes (genes with highest connectivity within modules) can highlight potential drivers of disease pathways.

Advantages

- Network-based, scalable, trait-integrated $\rightarrow$ ideal for identifying biologically relevant modules and hub genes in AMD

Limitations

- Requires careful preprocessing (normalization, batch correction) in single-cell data.
- Modules may differ across datasets → need replication or consensus analysis.
- As with WGCNA, results are hypothesis-generating and need biological validation.

Relevance

- AMD is complex: involves oxidative stress, inflammation, lipid metabolism, angiogenesis, likely requiring multi-gene modules rather than isolated genes.
- hdWGCNA could help:
  - Detect gene co-expression modules specific to retinal cell types (e.g., RPE cells, photoreceptors, microglia).
  - Correlate modules with AMD traits (case/control, severity, age).
  - Highlight hub genes that may play key roles in AMD progression.
- Especially powerful if your AMD data includes single-cell or spatial transcriptomics.

MEGENA

Why MEGENA?

- Traditional clustering and network methods often:
  - Miss hierarchical/multiscale structures in gene networks.
  - Identify only flat modules (one level of grouping).
- MEGENA is designed to capture multiscale modular organization of gene co-expression networks.
- Relevant for complex diseases (like AMD) where gene interactions may occur at multiple scales (small functional clusters within larger pathways).

Core Concepts

- Planar Filtered Network (PFN):
  - Starts by constructing a co-expression network filtered into a planar graph to reduce noise and redundancy.
- Multiscale clustering:
  - Detects modules within modules (hierarchical communities).
  - Allows analysis at different scales of network resolution.
- Hub gene detection:
  - Identifies genes with high connectivity inside each module (candidate key drivers).

Core Concepts Cont.

- Functional enrichment:
  - Tests modules for enrichment in biological pathways and traits.

Advantages

- Captures hierarchical network structure (multiscale).
- Identifies both global and local patterns in co-expression.
- Good for uncovering complex disease pathways with multiple interacting modules.

Limitations

- Computationally more intensive than simple clustering.
- Requires careful preprocessing (normalization, batch correction).
- As with WGCNA/hdWGCNA, results are exploratory → need external validation.

Relevance

- AMD involves multiple biological layers:
  - Inflammation, angiogenesis, oxidative stress, complement system, lipid metabolism.
- MEGENA could:
  - Detect small, specialized gene modules (e.g., complement activation genes).
  - Reveal how these nest inside larger functional modules (e.g., immune regulation networks).
  - Prioritize hub genes within AMD modules as candidate biomarkers or therapeutic targets.
- Particularly useful if you want to highlight multi-level gene interactions instead of just flat clusters.

References

References

- Langfelder, P., & Horvath, S. (2008). *WGCNA: An R package for weighted correlation network analysis*. BMC Bioinformatics, 9, 559.
- Morabito, S., Reese, F., Rahimzadeh, N., Swarup, V., & McManus, M. (2023). *hdWGCNA: High-dimensional weighted gene co-expression network analysis*. Cell Reports Methods, 3(6), 100498.
- Song, W.-M., & Zhang, B. (2015). *Multiscale Embedded Gene Co-expression Network Analysis*. PLoS Computational Biology, 11(11), e1004574.