Session 1: Multiomic Analysis of Frailty

Resources

Discord

QA

Support

Interact

Amazon SageMaker

Link

Raw data

Analysis code

Visualization

GitHub

Resources

Slides

Notebooks

https://github.com/PriceLab/Aging_Workshop_24

Goals for session 1

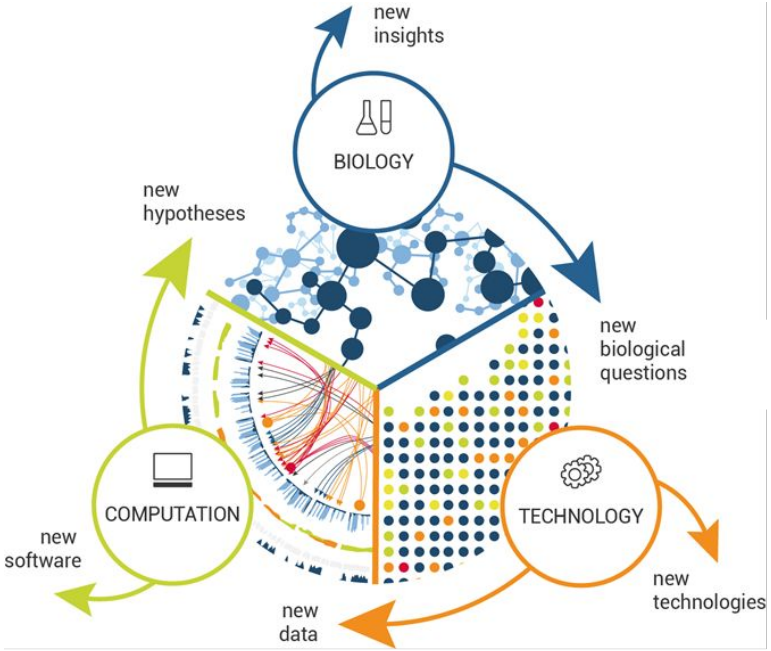
Session 1: Exploratory data analysis

- Define our aging phenotype of interest: Frailty Index
- Overview of the Arivale dataset
- Look at a systems approach to single omics
- Compare to a standard single feature analysis
- And finally- Explore a multi-omics (metabolomic, proteomic, clinical labs) network with our frailty outcome

Session 2: Explore the outcome

Session 3: Machine learning

Multiomics & Systems Biology



Genetics
Whole-Genome Sequencing

Social Determinants of Health
Stress and Environment

Electronic Health Records (EHRs)
Longitudinal Clinical History

Self-Tracking
Wearables, Questionnaires, & Assessments

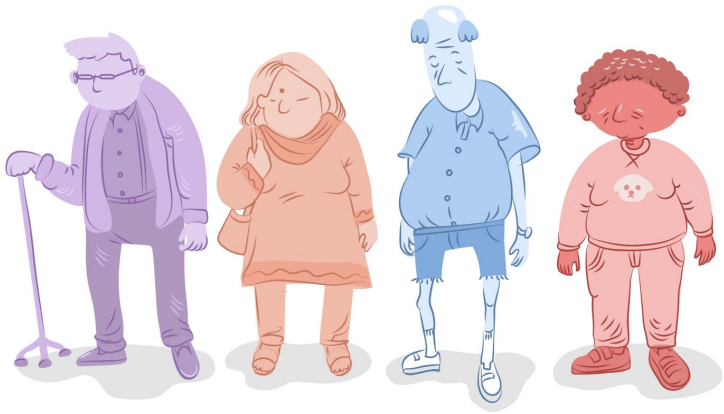
Microbiome
Gut Microbiome Composition

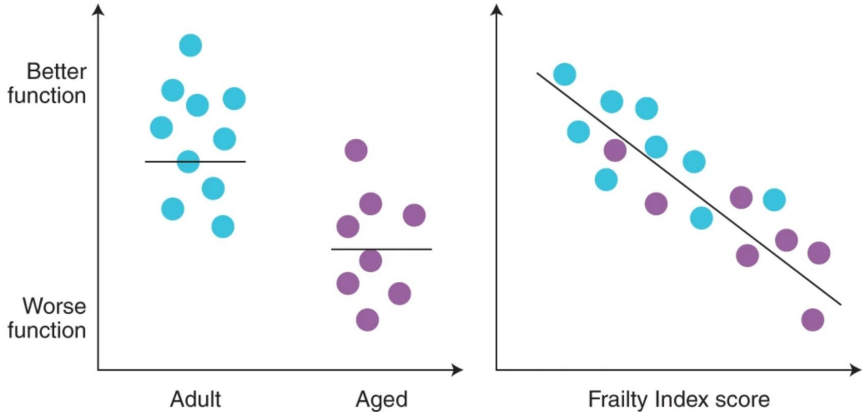
Brain
Cognitive & Psychiatric Assessments

Blood
Clinical Labs, Proteins, Metabolism, Immune System, and Much More ...

Frailty Index (FI)

- FI is used to measure the health status of older individuals;
- A proxy measure of aging and vulnerability to poor outcomes/resilience
- Males have lower mean frailty index values than females of the same age, whereas females show better mean survival than males with the same frailty index value





Howlett, Rutenberg & Rockwood, Nature Aging, 2021

How to calculate a Frailty Index (FI)?

FI score = sum of health deficits/total number of health deficits measured

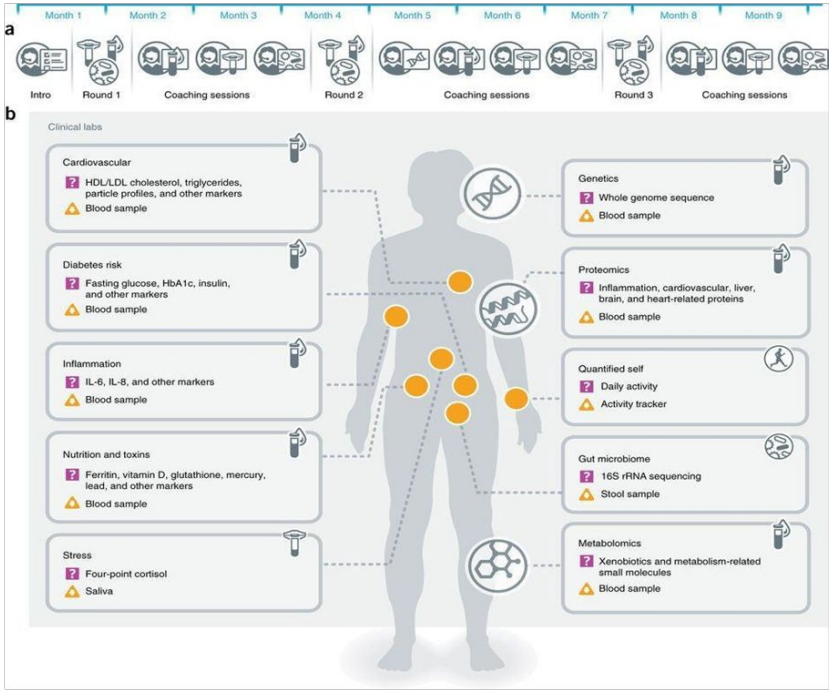
0 = no frailty, 0.7 = maximum frailty observed, 1= theoretical maximum

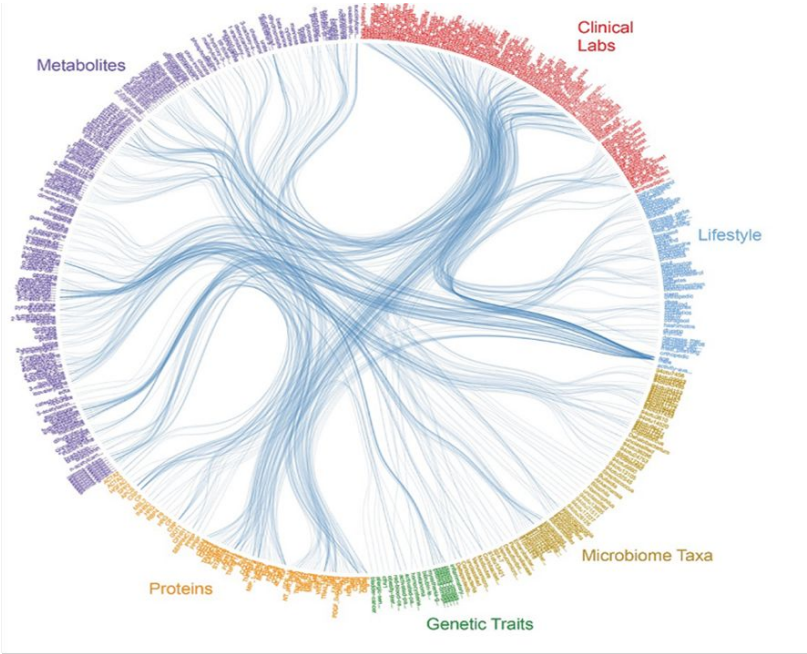
Traits of a frailty index:

1. Associated with health status
2. Prevalence increases with age, generally
3. Doesn’t saturate too early
4. As a group, cover a range of systems
5. Minimum 30 items included
6. Coded such that 0=absence of deficit, 1=presence of deficit

Searle et al, BMC Geriatrics, 2008

The Arivale Dataset





+Wellness coaching

Price et al, Nat Biotechnol, 2018

FI health deficits for the Arivale dataset:

- Self-Report FI (35 items)
  - Disease (15 items)
  - Activity (9 items)
  - Satisfaction (6 items)
  - Medication (3 items)
  - Digestion (2 items)
- Lab FI (34 items, cut-offs used to establish deficits)
  - Blood test items (29 items)
  - Blood pressure items (5 items)
- Combined FI (69 items)
  - The combination of the above two

Data cleaning

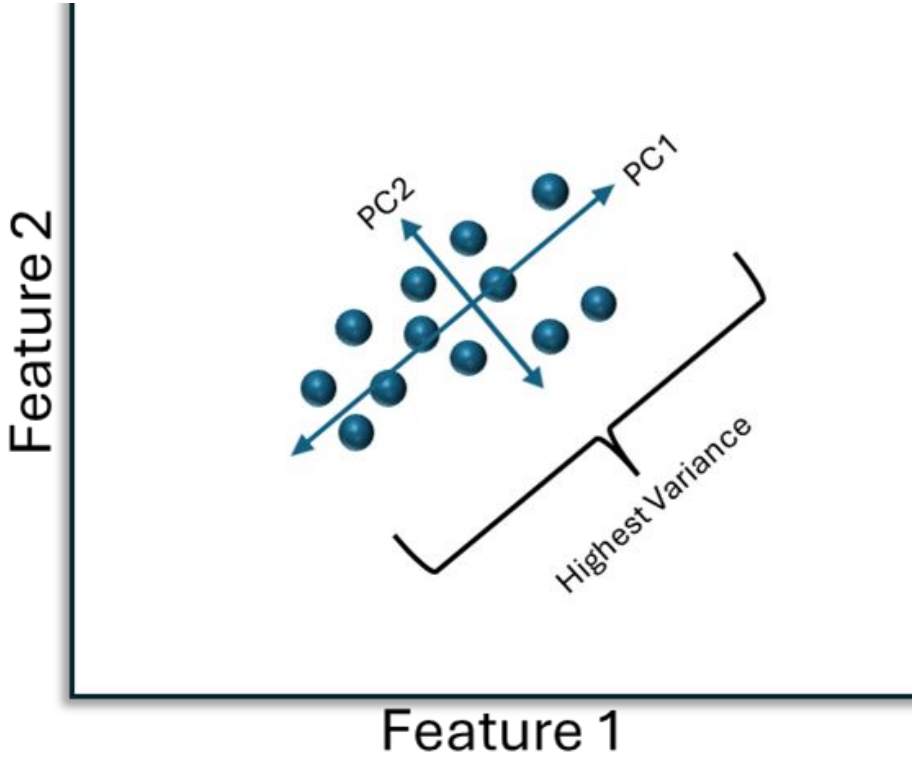
- Missingness
  - Various reasons for missingness that can be omic dependent.
  - Is missingess random? Correlated with other features?
- Normalization
- Imputation
  - Technique to deal with missing features. The method used should consider the omics.
- Removing features/participants
  - Are any features outliers? Why?
  - Are any participants outliers? Why?

*“Data science is **80%** data cleaning and **20%** complaining about data cleaning.” -Anonymous*

What is Principal Component Analysis (PCA) anway?

Dimensionality reduction technique:

- Explore data
- Analyze outliers
- Visualize high dimensional data
- Extraction features
- Reduce Noise



Systems Analysis - Weighted Correlation Network Analysis (WGCNA)

- Unsupervised clustering method
- around for 19+ years
- Increasingly applied to proteomics and metabolomics data

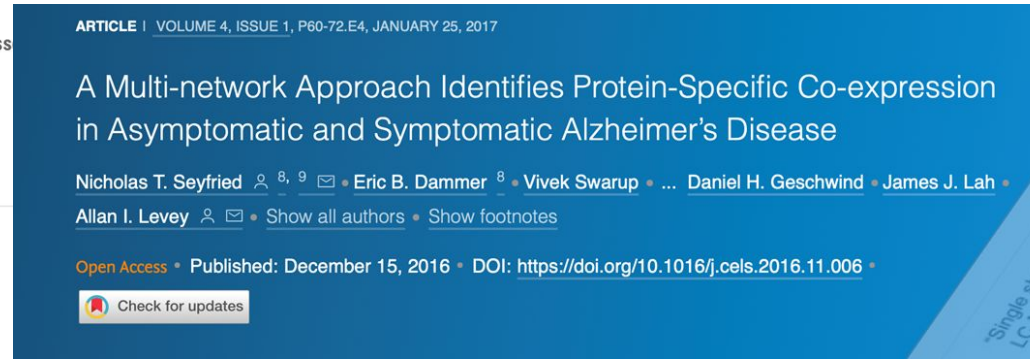
RESEARCH ARTICLE

Co-regulatory networks of human serum proteins link genetics to disease

Valur Emilsson[1,2,*,†], Marjan Ilkov[1,*], John R. Lamb[3,*,†], Nancy Finkel[4], Elias F. Gudmundss

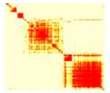
+ See all authors and affiliations

Science 03 Aug 2018:
eaaq1327
DOI: 10.1126/science.aaq1327



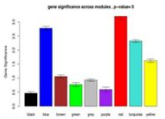
WGCNA framework

Construct a network
Rationale: make use of interaction patterns between genes

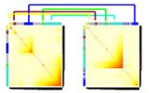


Identify modules
Rationale: module (pathway) based analysis

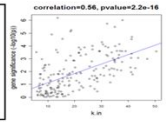
Relate modules to external information
Array Information: Clinical data, SNPs, proteomics
Gene Information: gene ontology, EASE, IPA
Rationale: find biologically interesting modules



Study Module Preservation across different data
Rationale:
- Same data: to check robustness of module definition
- Different data: to find interesting modules.

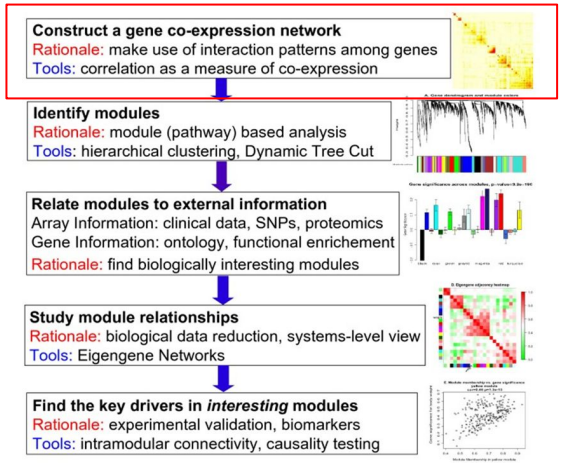


Find the key drivers in *interesting* modules
Tools: intramodular connectivity, causality testing
Rationale: experimental validation, therapeutics, biomarkers



Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics ***9,*** *559 (2008).*

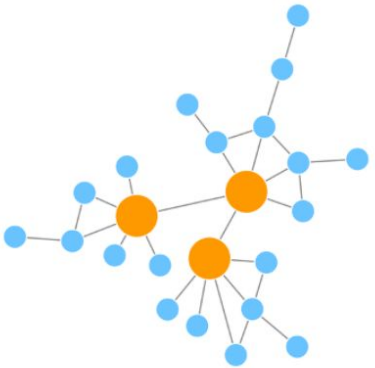
Why not just correlation?



Coexpression network, measure of correlation between features

Scale-free network, a network whose connections follow a power law

Topological Overlap, Indirect associations



Scale-free network

Source

Transform coexpression into adjacency network

Unsigned network, absolute value of coefficient

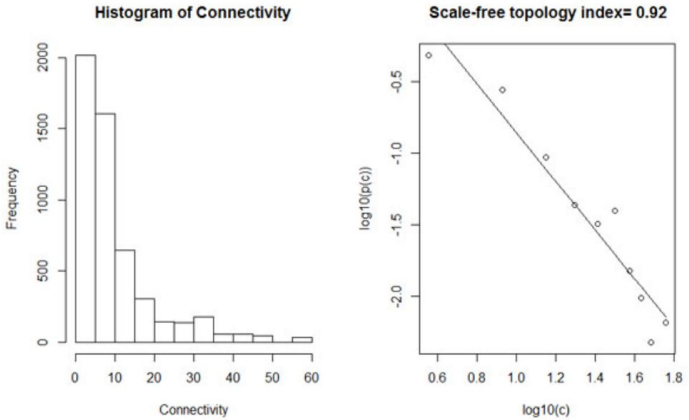
$$a_{ij} = | cor(x_i, x_j) |^{\beta}$$

Signed network, preserves sign

$$a_{ij} = | 0.5 + 0.5 \times cor(x_i, x_j) |^{\beta}$$

β is identified using the correlation of node connectivity and the log transformation of connectivity frequency

Soft-thresholding preserves information and tends to be more robust

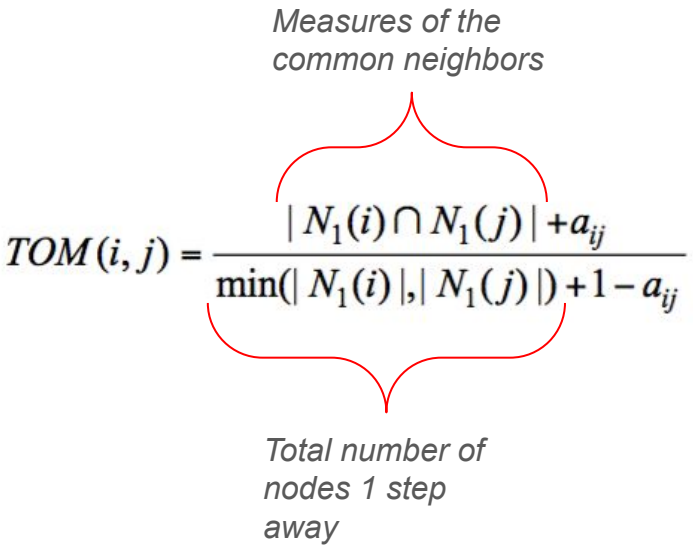


Source

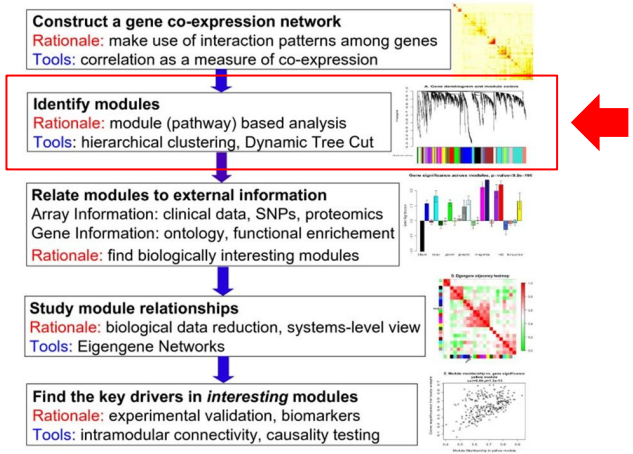
Reduce noise by topological overlap

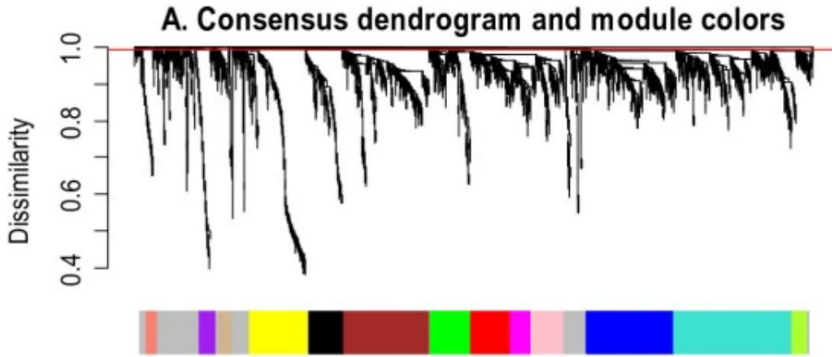
What about spurious or missing correlations?

The topological overlap is a measure of shared connectivity that normalizes the adjacency matrix based on shared nodes.



Identify modules by clustering



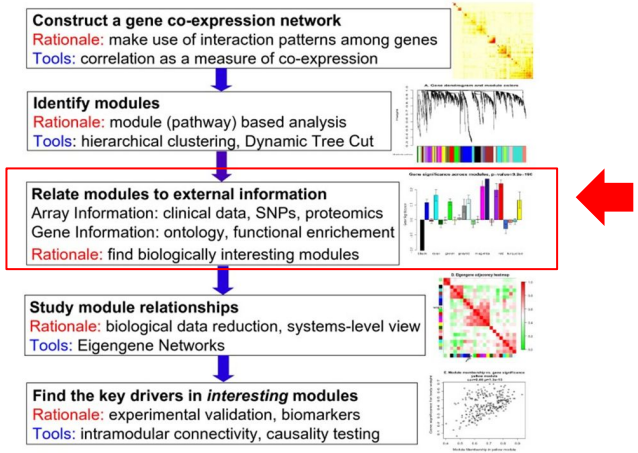


1. Calculate pairwise distances between nodes
2. Combine nodes with smallest distance
3. Repeat for combined nodes

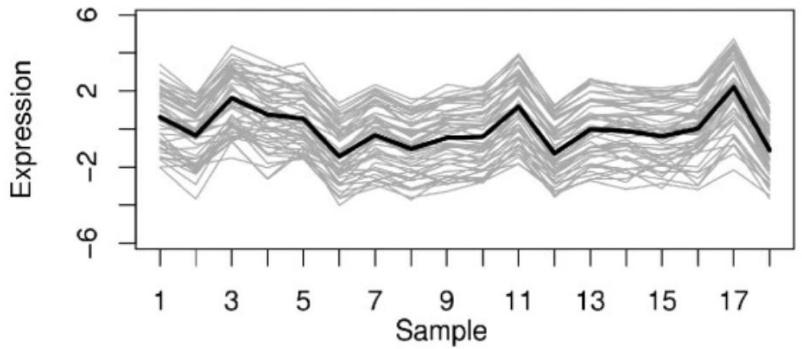
A dendrogram identifies cluster distances and cut-height determines modules. WGCNA uses a Dynamic Tree Cut.

Langfelder and Horvath, BMC System Biology, 2007

How to summarize a module? The eigengene



The module eigengene is defined as the first principal component of a given module. It can be considered a representative of the expression profiles in a module.

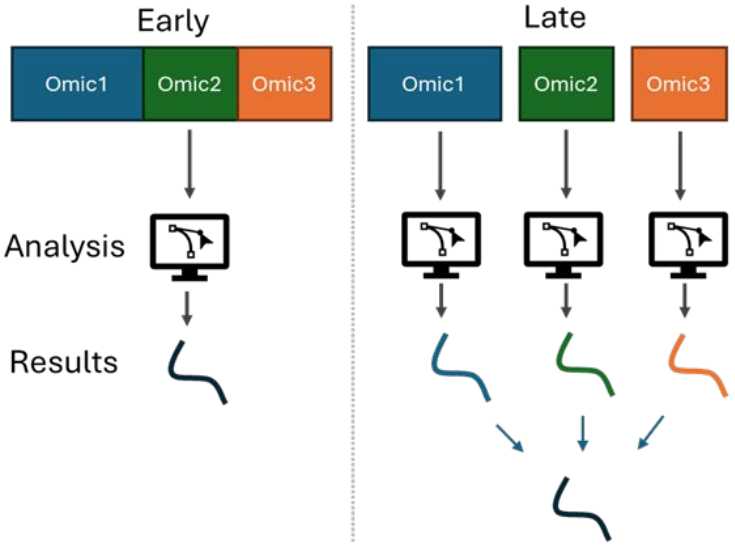


- Relate modules to each other
- Relate modules to phenotypes of interest
- Define module membership measure

Langfelder and Horvath, BMC System Biology, 2007

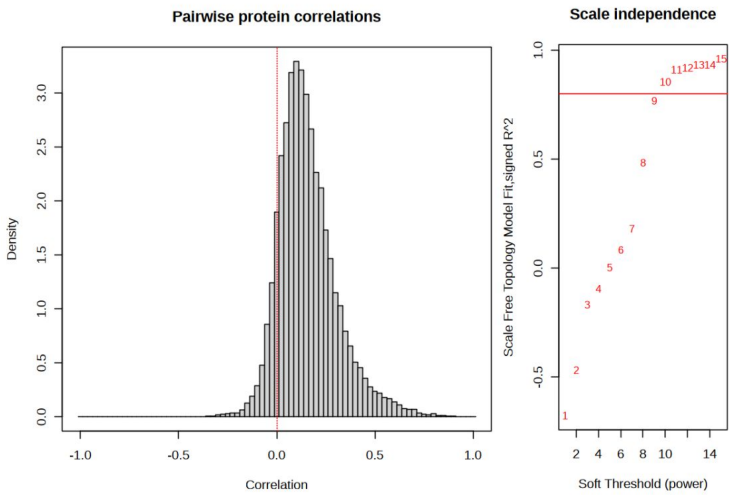
Multiomic integration

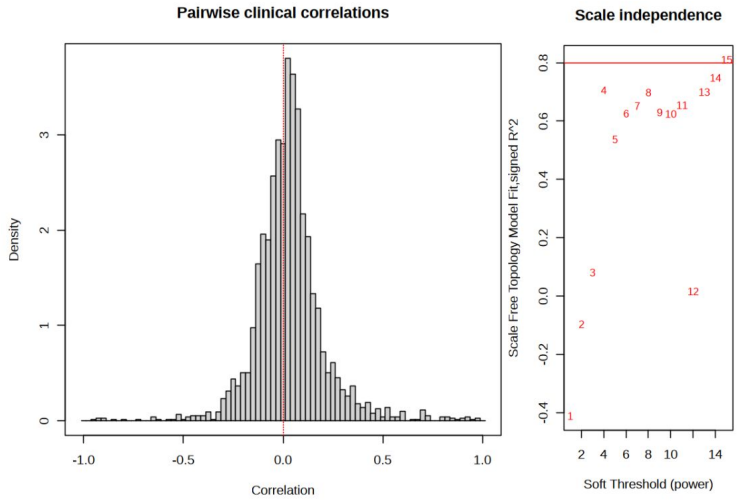
- Early integration: Concatenate omics and analyze
  - Requires consideration of data distribution
  - Preserves preserving correlation between omics
- Late integration: Analyze each omic separately and merge results
  - Straightforward by modeling each omic type
  - Does not capture interomic relationships
  - Correlation between omic eigengenes is commonly used to identify multiomic signatures



Why do we need to transform the distributions?

Which β to pick?





Attempting a scale-free network after concatenation would find a β that poorly fits all the omics, resulting in different adjacencies for interomic feature pairs.

Transform the distributions

1. Model each correlation matrix as a beta distribution.
2. Adjust the model to capture the peak, leaving room for positive correlation outliers
3. Center all models to a standard beta distribution (z-score)
4. Compare modeling results