# Network Analysis

## Integrative Genomics module

Michael Inouye
Centre for Systems Genomics
University of Melbourne, Australia

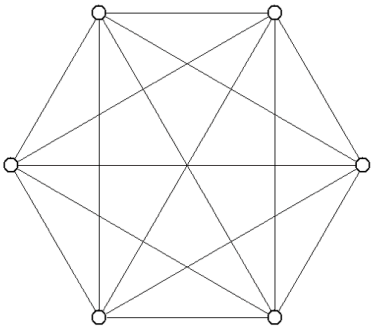
Summer Institute in Statistical Genetics 2016
Seattle, USA

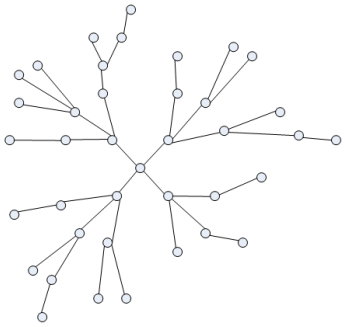
@minouye271
inouyelab.org


# This lecture

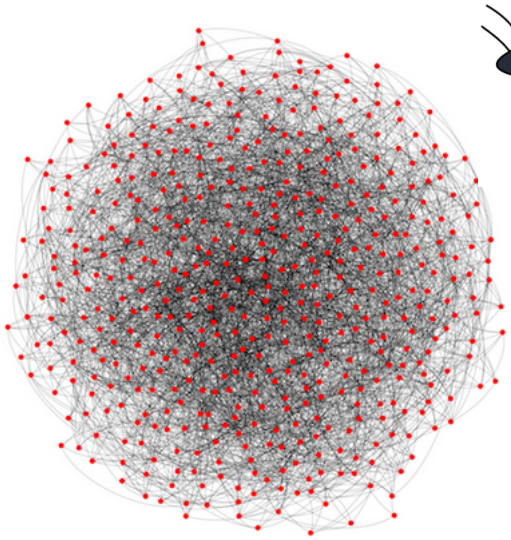
- **Network basics**
- **Properties of networks**
- **Gene co-expression networks**
- **Preservation of networks**

# What is a network?

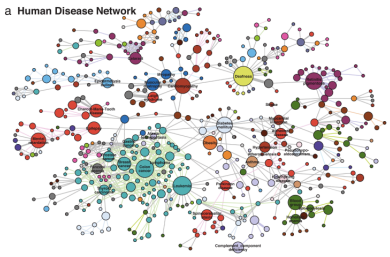


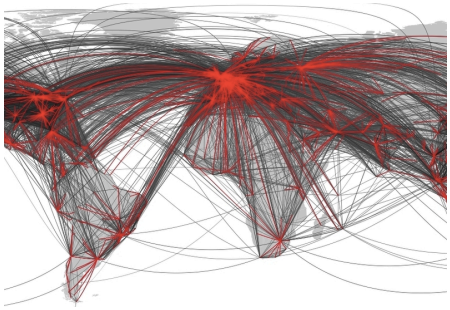


# YUCK



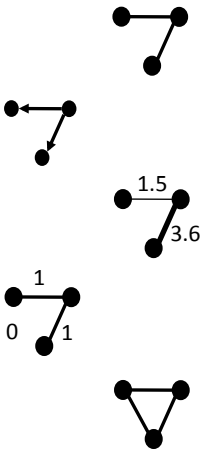
# Examples of networks



# Types of networks

- **Undirected**
- **Directed**
- **Weighted**
- **Unweighted**
- **Complete (and incomplete)**



## How are networks useful for gene expression?

- **What if probes/genes are nodes?**
  - Biological interpretation?
  - Statistically useful?
- **What if probes/genes are edges?**
  - Biological interpretation?
  - Statistically useful?

## What does a network look like in terms of data?

**Undirected & weighted: NxN symmetric matrix of relationships**

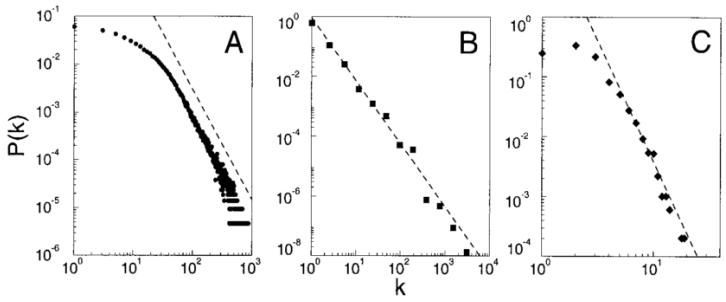
| | Gene1 | Gene2 | Gene3 |
|---|---|---|---|
| Gene1 | 1 | 0.75 | 0.95 |
| Gene2 | 0.75 | 1 | 0.04 |
| Gene3 | 0.95 | 0.04 | 1 |

**What might a matrix for a unweighted network look like?**

**What might a matrix for a directed network look like?**

# A few properties of networks

- **Connectivity distribution**
  - Scale-free topology $P(k) \sim k^{-\gamma}$

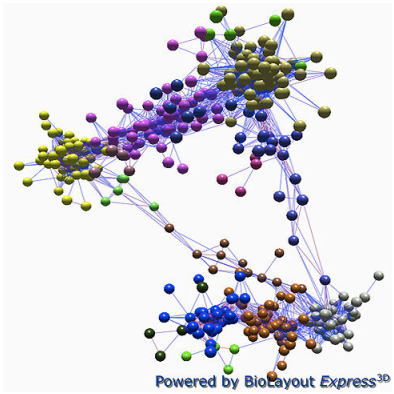



| Actors | www | Powergrid |
|---|---|---|
| $\gamma = 2.3$ | $\gamma = 2.1$ | $\gamma = 4$ |

Barabasi, *Science* 1999

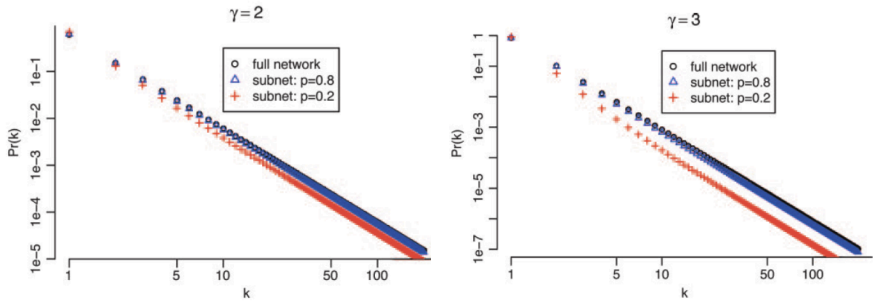
# A few properties of networks

- **Sub-networks**
  - Clusters of nodes which are more tightly related to each other than the rest of the network



# A few properties of networks

- **Subnets of scale-free networks**
  - **Are not necessarily accurate representations of the overall network**
  - **True for both random and especially nonrandom sampling of nodes**




**Why is this important for gene expression?**
**What effects might a scale-free assumption have?**

Stumpf et al, *PNAS* 2005

# Control of networks

- **What is control?**
  - Predict and test…

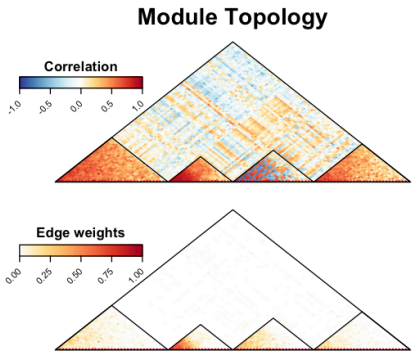
**State_A** —*Pertubation*→ **State_B**

- **What the main tools to control gene networks?**
  - Gene KOs and KDs

- **Are these tools adequate for useful control?**
  - Do our tools effect the predictions we make? Should they?

# Gene co-expression networks

- **Weighted, undirected complete gene network**
  - **Nodes**: genes/probes
  - **Edges**: $|cor(node\_i, node\_j)|^{\gamma}$
    - Scale-free assumption and [0,1]
- **Identify subnets (modules/ clusters)**
  - Typically subnets represent known biological pathways
  - Various methods and tools for clustering




# Strategies for testing association of a subnet with a phenotype

- **Univariate**
  - For each subnet gene, perform a test
- **Eigenvector**
  - Calculate 1st principal component
  - With vector of PC1 sample loadings, perform a test
- **Multivariate**
  - Simultaneously test for association of phenotype with all genes
  - Example: Canonical correlation analysis (CCA)
- **Considerations**
  - Multiple testing burden
  - Sensitivity and specificity

# Interpretation of subnets

- **Pathway analysis and gene set statistics**
- **If subnet is small enough, manual interpretation is possible (with proper literature support)**
- **Correlation vs Causation**
  - Confounding, causality and reactivity
    - It is more useful (and more difficult) to know the underlying structure of relationships b/n genes than *clusters* of co-regulation
  - How can causality be tested?
    - Perturbation techniques
    - Mendelian randomisation

# Preservation of subnets

- **Given a subnet (nodes, edges), is to preserved in a separate dataset?**
- **Examples**
  - Replication
    - Given N datasets generated under identical/similar settings, does a subnet 'replicate'?
  - Cross-tissue gene network preservation
    - Is a subnet derived from liver data preserved in adipose data?
  - Microbial communities between body sites
    - Is an operational taxonomic unit (OTU) subnet preserved between skin and upper airway samples?

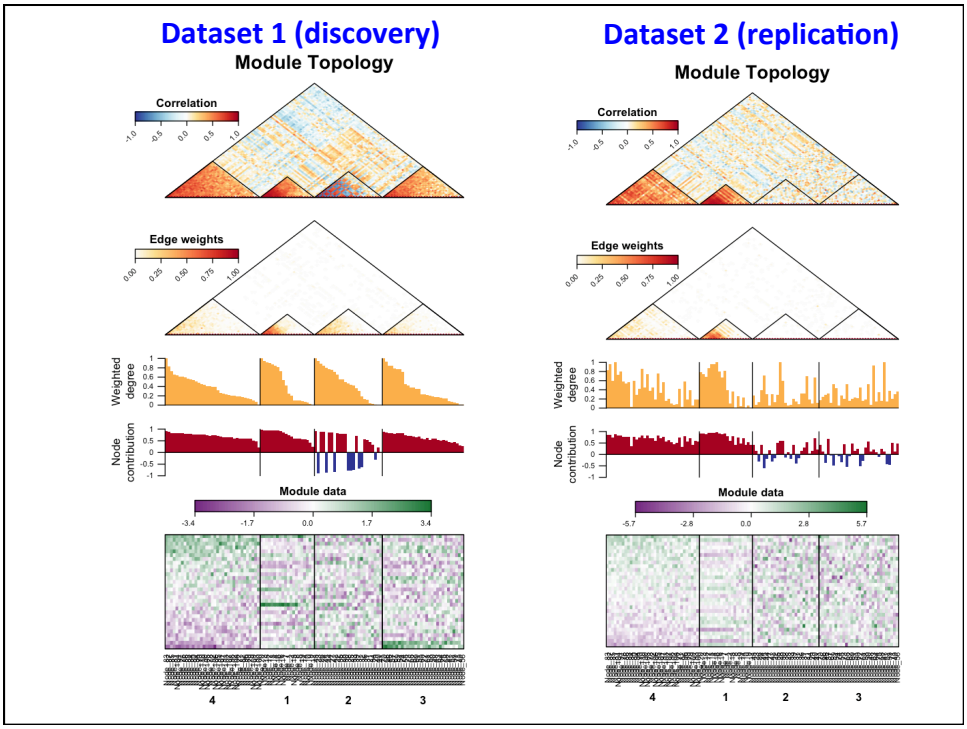
# Approaches to subnet preservation

- **Tabulation**
  - Make a table of features in a given subnet and those not. Test for deviation from null (e.g. Fisher Exact Test).

| | | Dataset 1 subnet A: IN | Dataset 1 subnet A: OUT |
|---|---|---|---|
| Dataset 2 subnet A | IN | a | b |
| | OUT | c | d |

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

- **Topological properties**
  - Edge patterns (for simplicity, assume no missing nodes)



# Preservation of topology

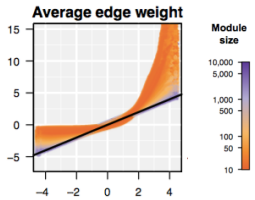
- Langfelder & Horvath, *PLOS Comp Bio* 2011
- Ritchie et al, *Cell Systems* 2016

| | General name of test statistic | WGCNA | Calculation |
|---|---|---|---|
| (1) | Module coherence | Proportion of variance explained | $mean\left(\left(cor\left(g_i^{[t](w)}, Eig_1^{[t](w)}\right)\right)^2\right)$ |
| (2) | Average node contribution | Mean sign-aware module membership | $mean\left(sign\left(cor\left(g_i^{[d](w)}, Eig_1^{[d](w)}\right)\right) \cdot cor\left(g_i^{[t](w)}, Eig_1^{[t](w)}\right)\right)$ |
| (3) | Concordance of node contributions | Correlation of module membership | $cor\left(cor\left(g_i^{[d](w)}, Eig_1^{[d](w)}\right), cor\left(g_i^{[t](w)}, Eig_1^{[t](w)}\right)\right)$ |
| (4) | Density of correlation structure | Mean sign-aware coexpression | $mean\left(sign\left(C^{[d](w)}\right) \cdot C^{[t](w)}\right)$ |
| (5) | Concordance of correlation structure | Correlation of coexpression | $cor_{i \neq j}\left(C^{[d](w)}, C^{[t](w)}\right)$ |
| (6) | Average edge weight | Mean adjacency | $mean_{i \neq j}\left(a_{ij}^{[t](w)}\right)$ |
| (7) | Concordance of weighted degree | Correlation of intramodular connectivities | $cor\left(\left(\sum_{i \neq j}^{J} a_i\right)^{[d](w)}, \left(\sum_{i \neq j}^{J} a_i\right)^{[t](w)}\right)$ |

| | |
|---|---|
| a | adjacency |
| cor | correlation |
| Sign | + / - |
| Eig | 1st principal component |

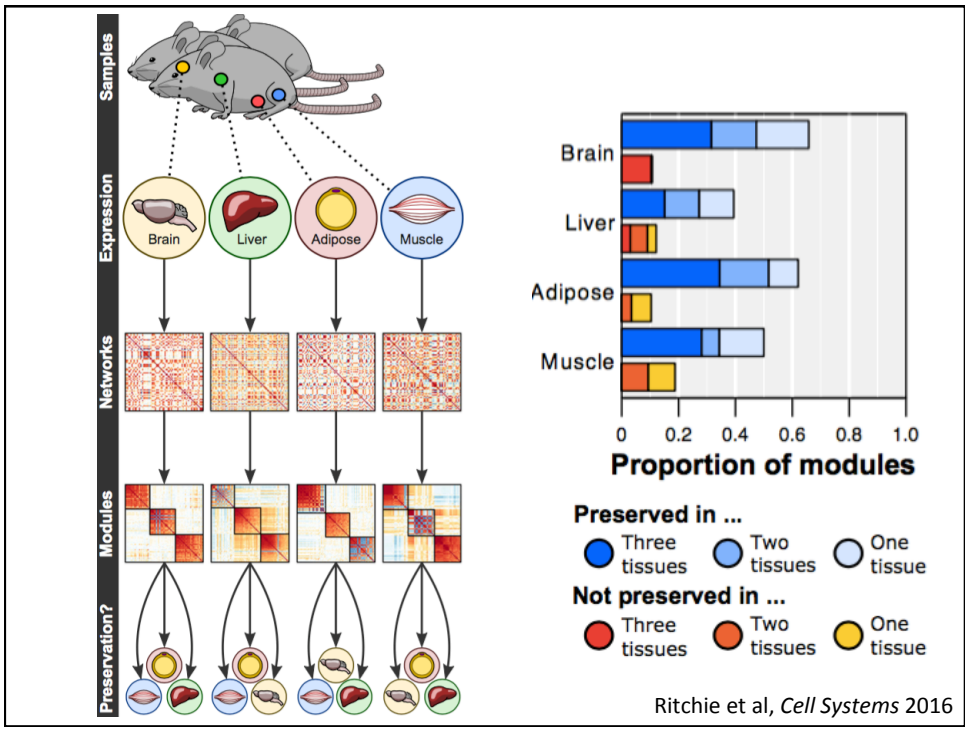
# When in doubt, permute the data

- **In network analysis, the complex relationships amongst nodes can make it difficult to assume a given test statistic follows a particular distribution**

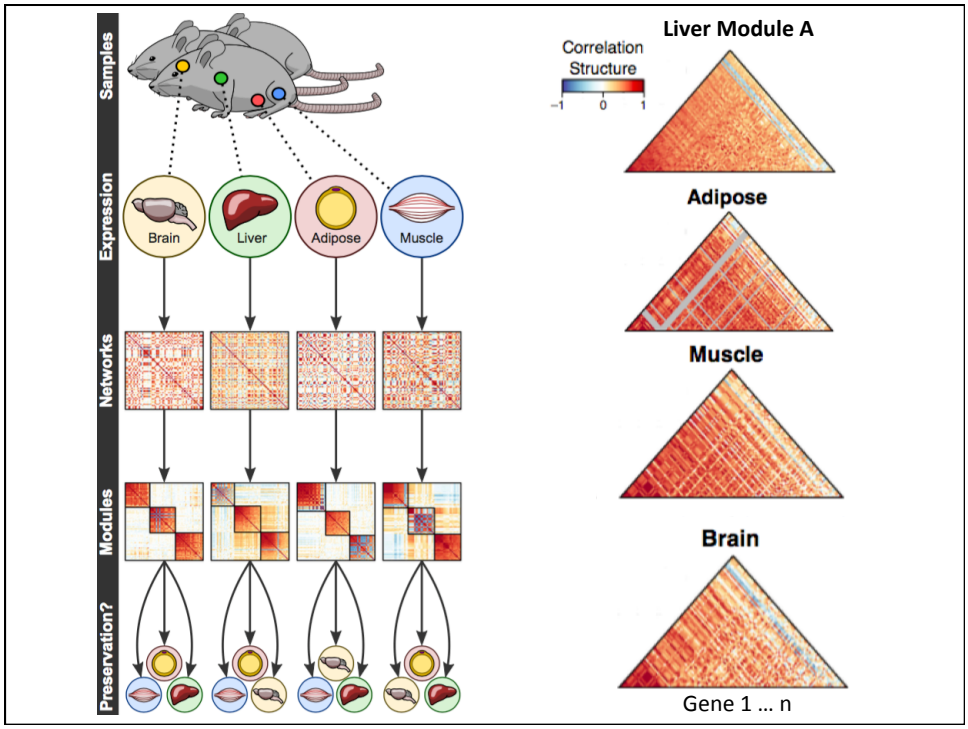



- **It is common (and good practice) to create an empirical (permuted) distribution of the test statistic to assess the original observation's significance**

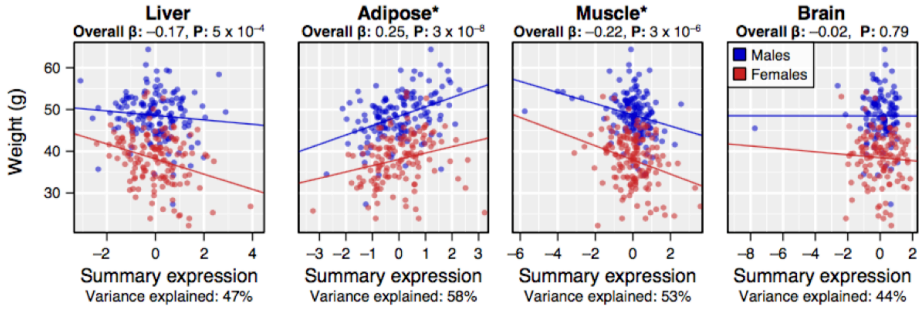
- **E.g. for a given module of with M nodes, with a given test statistic…**
  - Randomly draw M nodes from the overall network
  - Compute the test statistic of these random M nodes
  - Repeat many times
  - Compare the observed module value to the distribution of permuted values




Ritchie et al, *Cell Systems* 2016



# Phenotypic association (body weight)




| Test tissue | Trait | Effect size | 95% confidence interval | *P*-value | *Q*-value |
|---|---|---|---|---|---|
| Adipose | **Weight** | **0.25** | **0.16– 0.33** | **$3 \times 10^{-8}$** | - |
| | **Insulin** | **0.23** | **0.14– 0.32** | **$1 \times 10^{-6}$** | **$2 \times 10^{-5}$** |
| | **Glucose/Insulin** | **-0.21** | **-0.30– -0.12** | **$7 \times 10^{-6}$** | **$7 \times 10^{-5}$** |
| | Other fat | 0.23 | 0.11– 0.35 | $1 \times 10^{-4}$ | $8 \times 10^{-4}$ |
| | **Total fat** | **0.19** | **0.081– 0.30** | **$7 \times 10^{-4}$** | **0.004** |
| | Length | 0.17 | 0.069– 0.27 | 0.001 | 0.004 |
| | MCP-1 (CCL2) | 0.18 | 0.064– 0.29 | 0.002 | 0.007 |
| | Glucose | 0.18 | 0.064– 0.30 | 0.003 | 0.007 |
| | **Unesterified cholesterol** | **0.18** | **0.061– 0.29** | **0.003** | **0.007** |
| Muscle | **Weight** | **-0.21** | **-0.30– -0.13** | **$3 \times 10^{-6}$** | - |
| | **Unesterified cholesterol** | **-0.21** | **-0.34– -0.092** | **$6 \times 10^{-4}$** | **0.01** |
| | **Insulin** | **-0.16** | **-0.25– -0.061** | **0.001** | **0.01** |
| | **Total fat** | **-0.19** | **-0.31– -0.072** | **0.002** | **0.01** |
| | Abdominal fat | -0.17 | -0.27– -0.061 | 0.002 | 0.01 |
| | **Glucose/Insulin** | **0.14** | **0.048– 0.24** | **0.003** | **0.01** |
| | Free fatty acids | -0.18 | -0.31– -0.059 | 0.004 | 0.01 |
| | LDL+VLDL | -0.18 | -0.30– -0.056 | 0.005 | 0.01 |
| | HDL/LDL+VLDL | 0.17 | 0.051– 0.29 | 0.005 | 0.01 |
| | Total cholesterol | -0.17 | -0.29– -0.049 | 0.006 | 0.01 |