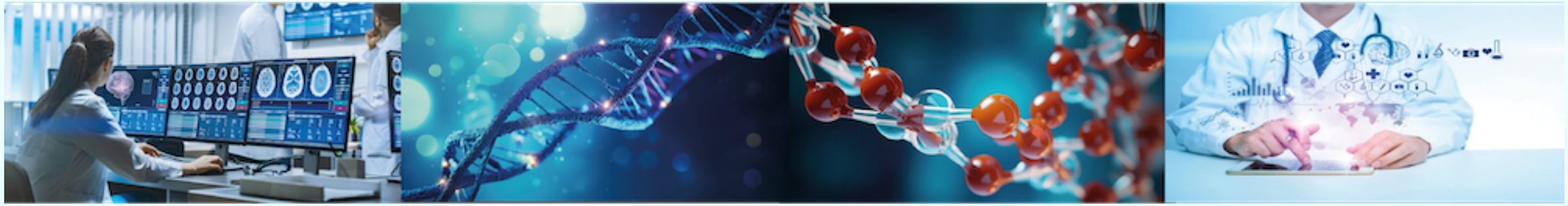


Programming for Biomedical Informatics

Lecture 13 "Essential Network Methods"

https://github.com/tisimpson/pbi

Ian Simpson
ian.simpson@ed.ac.uk

Background

THE UNIVERSITY of EDINBURGH
informatics

Common Network Analysis Approaches in Biomedical Informatics

Gene Co-expression Networks

- correlation-based or similarity-based networks, often constructed using methods like Weighted Gene Co-expression Network Analysis (WGCNA)
- used to find groups of genes that exhibit similar expression patterns, which may be co-regulated or involved in similar biological processes
- identifying biomarkers, discovering gene modules associated with diseases

Protein-Protein Interaction (PPI) Networks

- experimental data or predicted interactions to map physical interactions between proteins
- used to understand functional relationships and biological pathways
- drug target discovery, pathway enrichment analysis, studying disease mechanisms at the protein level

Metabolic Networks

- metabolic pathways and enzyme-catalysed reactions
- used to identify biomarkers, understand reaction kinetics, effects of perturbations
- metabolic alterations in disease, metabolic phenotypes

Signalling Pathway Networks

- regulatory pathways, where nodes are proteins, enzymes, or other molecules involved in signalling
- used to model signal transduction and communication processes within and between cells
- disease-specific signalling pathways and therapeutic target identification

Gene Regulatory Networks (GRNs)

- transcriptional data and prior data about transcription factor -> target gene relationships to infer regulatory networks (use data such as ENCODE, GTEx..)
- used to reveal how gene expression is regulated under different conditions, including disease states
- discovering master regulators in diseases, therapeutic targets

Integrated Multi-Omics Networks

- integrates multi-modal genomics data such as transcriptomics, proteomics, metabolomics, etc., often using network fusion techniques
- Used to gain a more holistic understanding of biological processes and disease phenotypes by combining multiple omics layers
- Disease sub-typing, biomarker discovery, systems biology, and personalised medicine

Functional Enrichment and Pathway Analysis

- biological pathways and ontologies to identify enriched pathways or gene sets, often leveraging databases like KEGG, Reactome, and GO
- used to aid interpretation of networks in the context of biological functions and pathways
- drug repositioning, identifying disease-relevant pathways, functional annotation of genes or proteins

Structural Network Analysis Techniques

- network metrics like centrality, modularity, and community detection
- used to analyse biomedical network structures, to identify influential nodes (like hub genes or proteins), and detect functional modules
- biomarkers, drug targets, disease-related gene communities, studying network robustness

Patient Networks for Autism Spectrum Disorders

SSC Collection Overview
Families: 2644
Individuals: 10474
Female probands: 352
Male probands: 2292

Simons Searchlight
155 genes, 23 CNVs
>1500 individuals

SFARI GENE

SFARI GeneBase
Scored Genes: 1028
ASD linked genes: 969
Animal Models: 276 genes
CNVs: 17

SPARK
Simons Powering Autism Research

Phenotypic data from **283,520** individuals (including more than **100,000** individuals with ASD)

94,116 children (<18y) with ASD
17,604 adults with ASD
83,897 male pro-bands
27,973 female pro-bands
43,158 unaffected siblings
7,074 enrolled twins, triplets and quadruplets
16,951 multiplex families

Types of data include basic medical screening, Developmental Coordination, Repetitive Behaviour, Social Communication, Adaptive Behaviour.

11/02/22

DECIPHER
GRCh38

Open Access Patient Records: **40,544**
Phenotpye Observations: 173,565
CNVs 41,679

DDG2P
Unique Genes: **2132**
Diseases: 2161
With Phenotype: 1550

PubMed.gov
Europe PMC

Literature corpora for Autism Spectrum Disorder
OA and UoE licensed for Data Science
Full-Text Retrieval HTML, XML, & PDF
68,329 target papers 04/02/22 (**54,591** - 86% retrieved)

Questionnaire Data

DCDQ - 17 questions with 5-point Likert scale
[scores are 0 for "not at all like this child to 5 for "extremely like this child"]

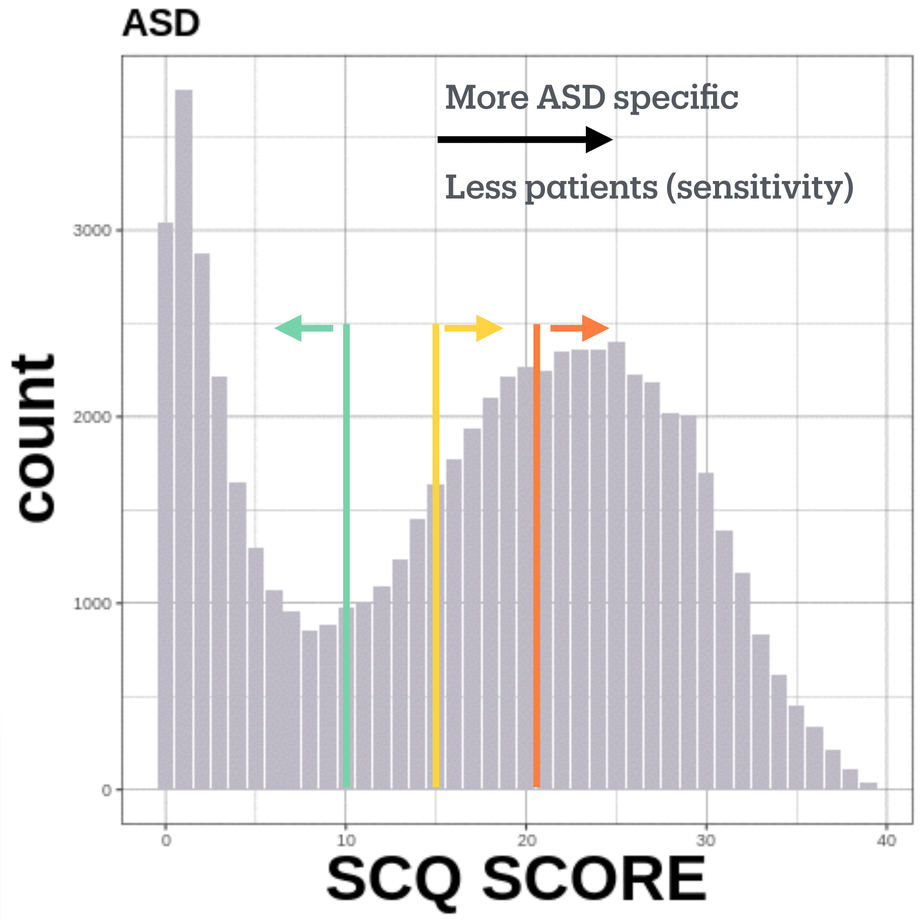
RBS-R - 43 questions 4-point Likert scale
[scores are from 0 "behaviour doesn't occur" to 3 "behaviour occurs and is a severe problem”]

SCQ - 40 yes/no questions with 0 = 0 absence of abnormal behaviour and 1 presence.
[SCQ can be asked as current or lifetime, the latter being better in very young children]

Instrument	Parents	Individuals with ASD	Unaffected Siblings
Background History Questionnaire-Adult Measure of: Demographics and social history Format: Online Questionnaire Completed by: Independent adults with ASD Regarding: Independent ASD adults over 18 years Author: SFARI		x	
Background History Questionnaire-Child/Dependent Measure of: Demographics and developmental history Format: Online Questionnaire Completed by: Parent/guardian Regarding: Dependent ASD individuals Author: SFARI		x	
Background History Questionnaire-Sibling Measure of: Demographics and developmental history Format: Online Questionnaire Completed by: Parent/guardian Regarding: Non-ASD siblings under 18 years Author: SFARI			x
Basic Medical Screening Questionnaire Measure of: Medical history Format: Online Questionnaire Completed by: Parent/guardian or self Regarding: All ASD individuals, their parents and siblings Author: SFARI	x	x	x
Developmental Coordination Disorder Questionnaire Measure of: Motor delays Format: Online Questionnaire Completed by: Parent/guardian Regarding: Dependent ASD individuals age 5–15 years Author: B.N. Wilson		x	
Social Communication Questionnaire (SCQ) — Lifetime Measure of: Screen of ASD markers Format: Online Questionnaire Completed by: Parent/guardian Regarding: ASD individuals and non-ASD siblings age 2 to 17 years 11 months Publisher: WPS		x	x
Repetitive Behaviors Scale-Revised Measure of: Repetitive behaviors Format: Online Questionnaire Completed by: Parent/guardian Regarding: Dependent ASD individuals age 3 years and up Author: Bodfish		x	

Wilson, B. N., Kaplan, B. J., Crawford, S. G., Campbell, A. & Dewey, D. Reliability and Validity of a Parent Questionnaire on Childhood Motor Skills. Am J Occup Ther 54, 484–493 (2000).
Lam, K. S. L. & Aman, M. G. The Repetitive Behavior Scale-Revised: Independent Validation in Individuals with Autism Spectrum Disorders. J Autism Dev Disord 37, 855–866 (2007).
Marvin, A. R., Marvin, D. J., Lipkin, P. H. & Law, J. K. Analysis of Social Communication Questionnaire (SCQ) Screening for Children Less Than Age 4. Curr Dev Disord Reports 4, 137–144 (2017).

Diagnostic Decisions by SCQ Score



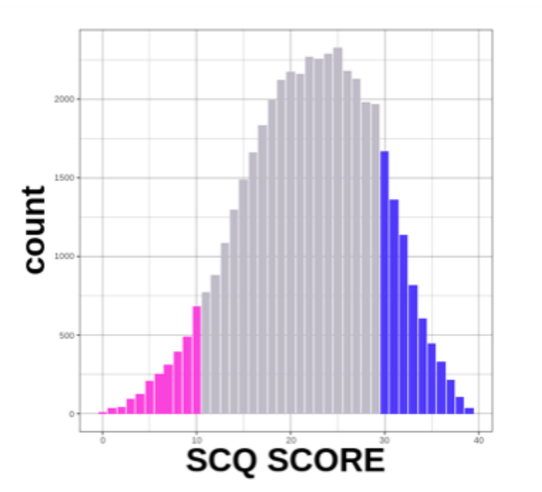
- Low SCQ scores indicates individuals less likely to be diagnosed with ASD.
- High SCQ scores indicates individuals likely to be diagnosed with ASD.
- Typically SCQ score > 15 used as threshold for ASD diagnosis

10%	20%	30%	40%	50%	60%	70%	80%	90%
1	4	10	15	18	21	24	27	30

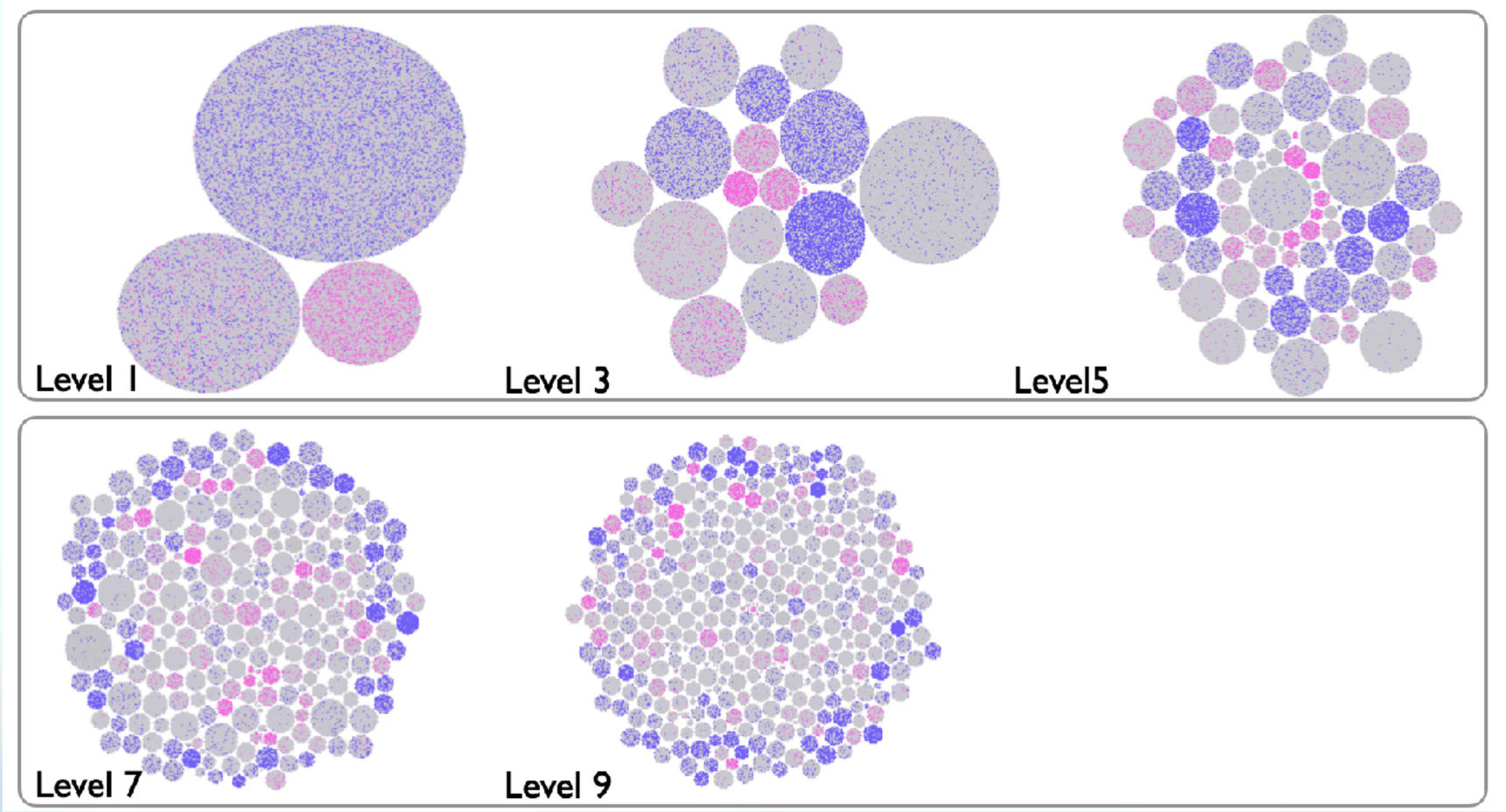
Marvin AR, Marvin DJ, Lipkin PH, Law JK. Analysis of Social Communication Questionnaire (SCQ) Screening for Children Less Than Age 4 Curr Dev Disord Rep. 2017;4(4):137-144. doi: 10.1007/s40474-017-0122-1

Individuals with Extreme SCQ Scores are Concentrated in Clusters

Social Communication Questionnaire Scores

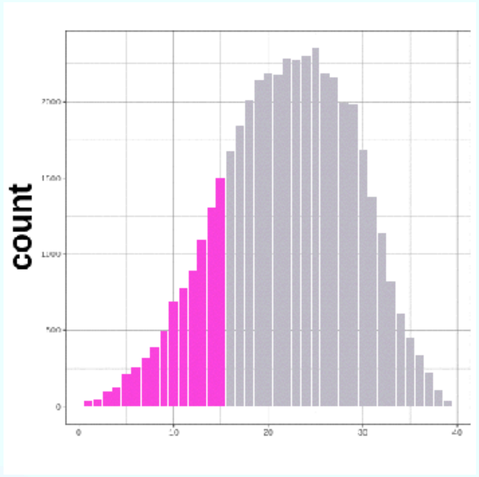


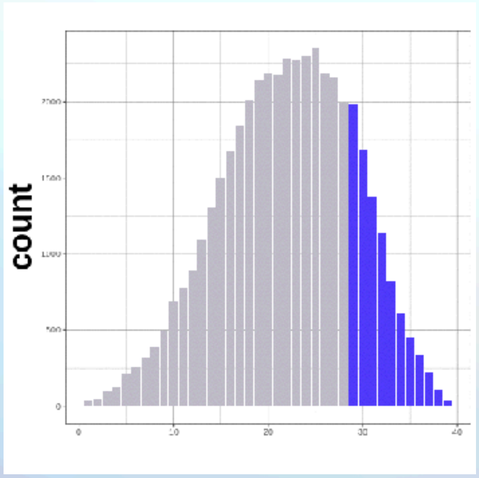
- scaled weighting in data fusion
- fast-greedy clustering
- unequal SCQ score clustering



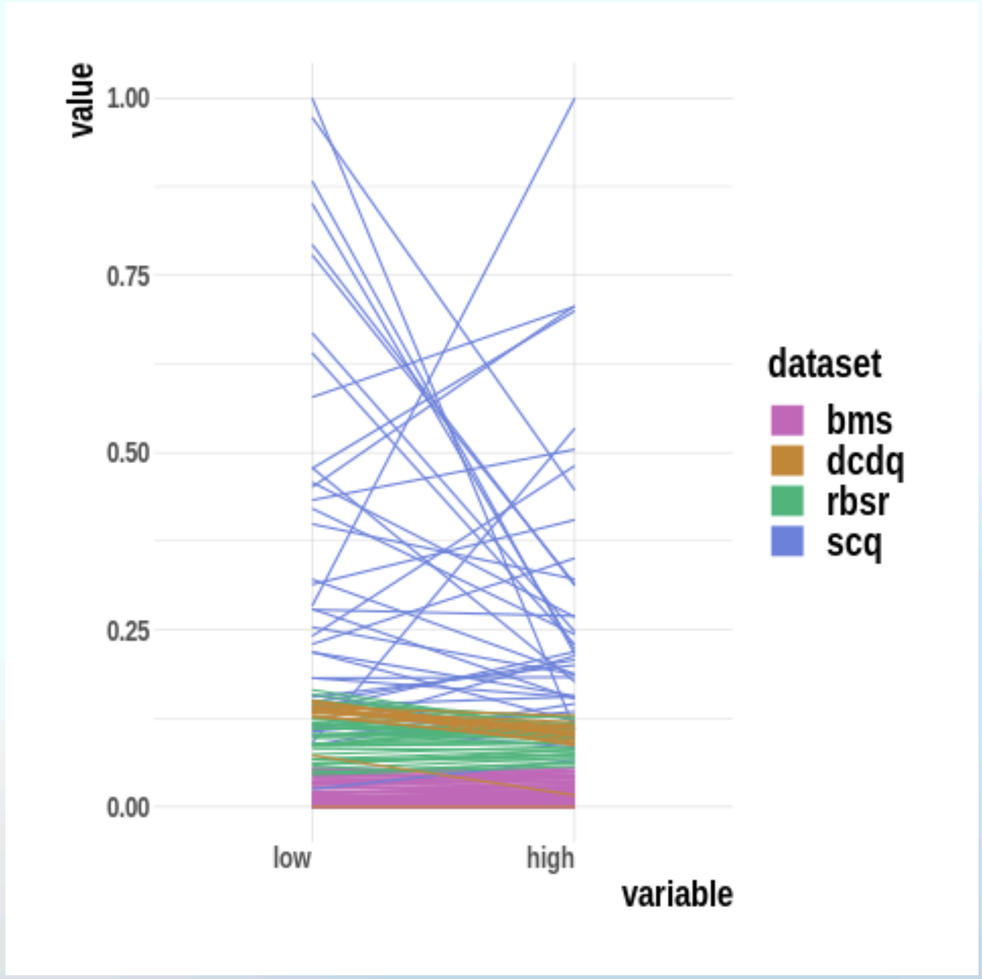
Social Coordination is the Strongest Predictor of ASD Status in SPARK

SCQ Scores

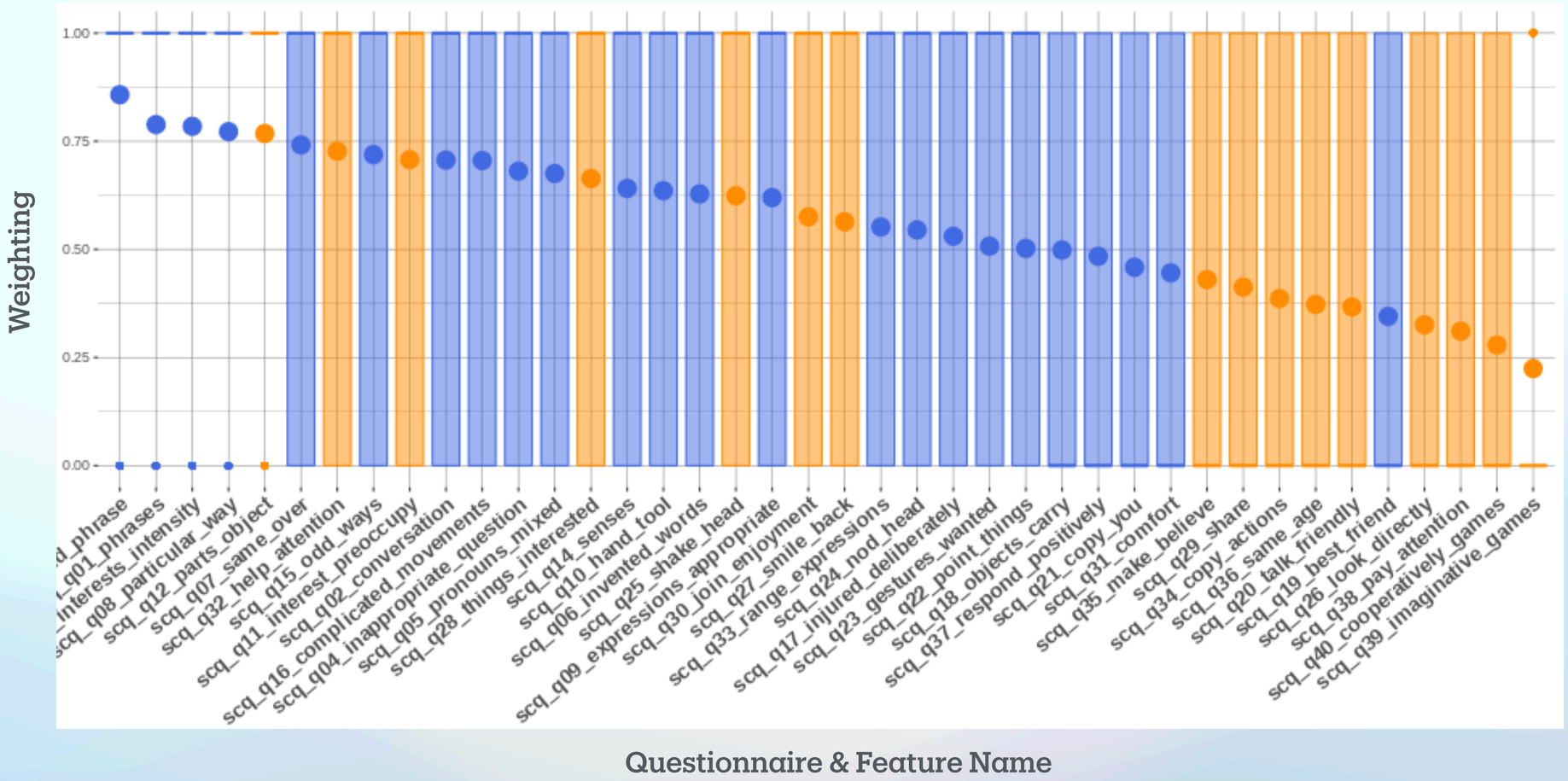




Feature Importance



Diagnostic Features



Feature Guided RF Predicts ASD Status

- Prediction of ASD status using as few as 9 features (out of c.200) used in the parental questionnaires given to parents upon joining the scheme.
- Raises possibility of new low impact approach for ASD triage
- Further work can stratify performance by other factors such as age
- Network clusters can be evaluated using this schema to identify sub-groups of clinical features.
- Establishes joint modelling framework for additional data fusion using genomics (exome and genome/GWAS - Clustering of genes and variants with genomic features)

low extreme gamma = 1

	Joint	non-joint
N Patients	23990	31915
N features	9	9

Joint	ASD	nonASDext
precision	0.95	0.62
recall	0.90	0.79
fscore	0.92	0.7

	ASD	nonASDext
ASD	6562	239
nonASDext	438	740

non-joint	ASD	nonASDext
precision	0.96	0.63
recall	0.94	0.76
fscore	0.95	0.67

low extreme gamma = 0.6

	Joint	non-joint
N Patients	23990	31915
N features	35	35

Joint	ASD	nonASDext
precision	0.98	0.83
recall	0.96	0.93
fscore	0.97	0.88

	ASD	nonASDext
ASD	6738	78
nonASDext	239	924

non-joint	ASD	nonASDext
precision	0.98	0.80
recall	0.97	0.92
fscore	0.98	0.85

high extreme gamma = 1

	Joint	non-joint
N Patients	23990	31915
N features	6	6

Joint	ASD	nonASDext
precision	0.93	0.57
recall	0.90	0.69
fscore	0.92	0.63

	ASD	nonASDext
ASD	6876	191
nonASDext	477	435

non-joint	ASD	nonASDext
precision	0.97	0.48
recall	0.94	0.70
fscore	0.95	0.57

high extreme gamma = 0.6

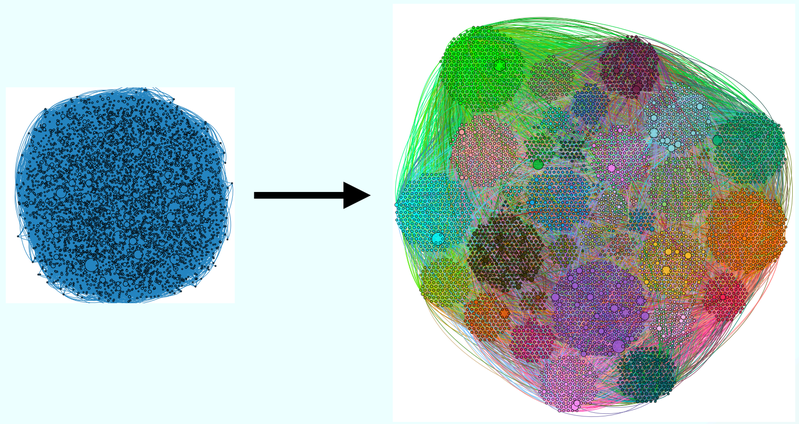
	Joint	non-joint
N Patients	23990	31915
N features	32	32

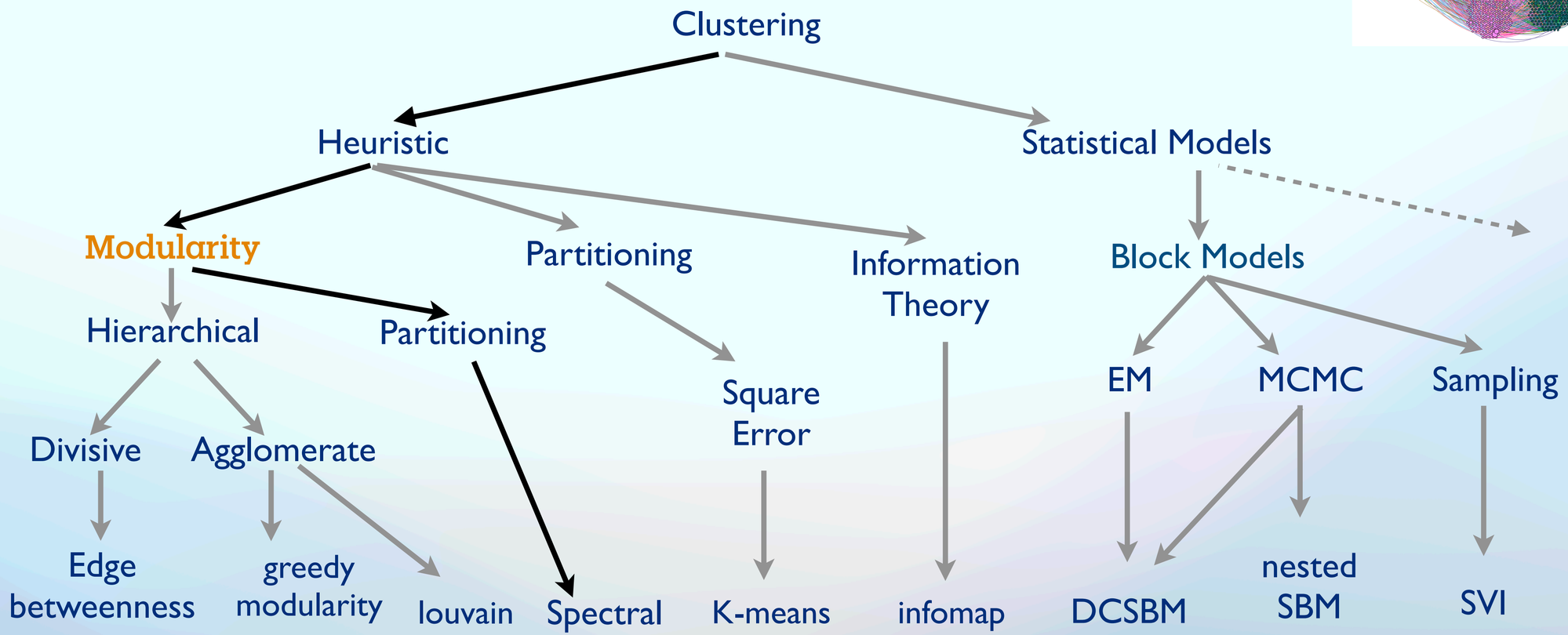
Joint	ASD	nonASDext
precision	0.98	0.83
recall	0.96	0.92
fscore	0.97	0.87

	ASD	nonASDext
ASD	6917	94
nonASDext	174	794

non-joint	ASD	nonASDext
precision	0.99	0.82
recall	0.98	0.89
fscore	0.98	0.86

Clustering Networks



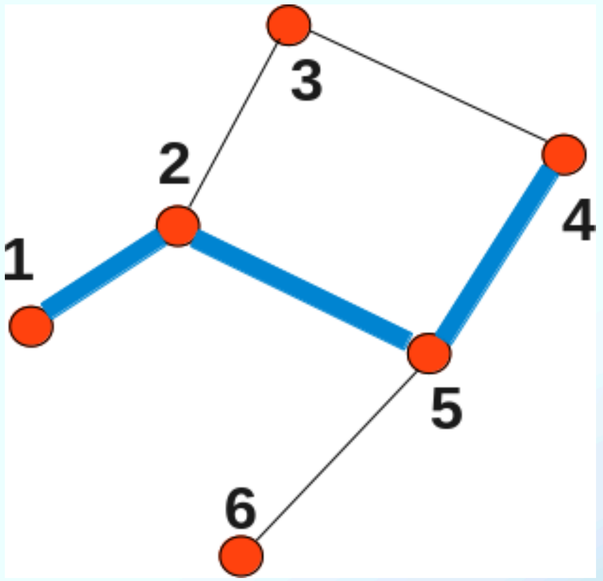


Network - Representations

$A_{ij} \in \mathbb{R}$ (unweighted = 0,1)

$A_{ij} = A_{ji}$ (undirected)

$A_{ii} = 2$ (if self-edge)



Adjacency Matrix

$$\mathbf{A} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 & 1 & 0 \\ 3 & 0 & 1 & 0 & 1 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 & 1 & 0 \\ 5 & 0 & 1 & 0 & 1 & 0 & 1 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}$$

$$k_i = \sum_{j=1}^{n} A_{ij}$$ (degree of node i)

(edges in network)

$$\frac{1}{2}\sum_{i=1}^{n} k_i = 6$$

Degree Matrix

$$\mathbf{D} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 3 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 2 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 2 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 3 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Laplacian Matrix

$$\mathbf{L} = \mathbf{D} - \mathbf{A} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 2 & -1 & 3 & -1 & 0 & -1 & 0 \\ 3 & 0 & -1 & 2 & -1 & 0 & 0 \\ 4 & 0 & 0 & -1 & 2 & -1 & 0 \\ 5 & 0 & -1 & 0 & -1 & 3 & -1 \\ 6 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Converting Adjacency Matrix to Distance Matrix

Binary Inversion for Unweighted Graphs

- for an unweighted adjacency matrix:
  - set connected nodes (1s in the adjacency matrix) to a small distance, such as 1
  - set unconnected nodes (0s in the adjacency matrix) to a large distance, such as infinity or a high number, to indicate no direct path
- this conversion results in a binary distance matrix where 1 represents adjacent nodes and infinity (or a high value) represents disconnected nodes

Invert Edge Weights for Weighted Graphs

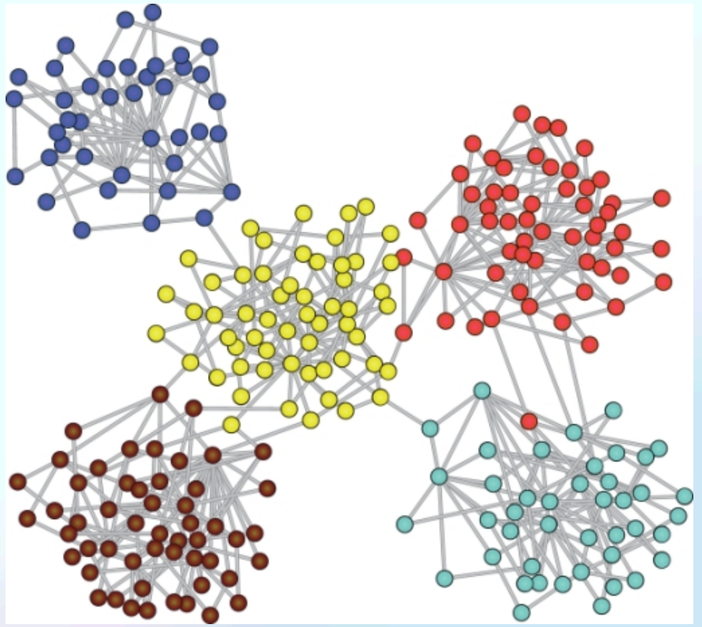
- if the adjacency matrix has weighted edges, where weights represent some form of proximity or similarity, you can invert these values to obtain distances:
  - for each entry in the adjacency matrix A(i, j), calculate $D(i, j) = 1 / A(i, j)$
  - ensure that the diagonal values (self-distances) are set to 0
- this method is useful when weights are proportional to closeness or similarity between nodes

Shortest Path Algorithm for Indirect Distances

- for sparse graphs or graphs with nodes that are indirectly connected, compute a distance matrix based on the shortest path between each pair of nodes:
  - use algorithms like Floyd-Warshall (for dense graphs) or Dijkstra's algorithm (for sparse graphs) on the adjacency matrix to compute shortest path distances
  - set self-distances to 0, and unreachable nodes to infinity (or a high number)
- this approach provides the geodesic distance between nodes, which is useful in community detection or network analysis

Euclidean Distance

- if each node has coordinates (e.g., in a spatial network), compute Euclidean distances instead of using the adjacency matrix directly:
  - calculate pairwise Euclidean distances between nodes based on their coordinates, forming a distance matrix
- this is especially applicable in spatial networks or if you have a node embedding with coordinates in a latent space



Clustering Networks - Community Clustering

General Approach

Graph Representation

- represent the data as a graph, where nodes represent entities and edges represent connections or relationships between them

Initialise Clusters

- start with each node as its own individual cluster or use an initial partitioning method to assign nodes to initial clusters

Calculate Modularity

- compute the initial modularity score, which quantifies the density of connections within clusters compared to connections between clusters

Optimise Modularity Locally

- iteratively merge or reassign nodes or small groups of nodes to maximise modularity
- for each iteration, assess the change in modularity resulting from moves or merges

Evaluate Modularity Gain

- accept moves or merges that improve the modularity score, and reject those that do not

Repeat Until Convergence

- continue optimising the modularity score until no further significant improvements can be made

Refine Clusters

- optionally, reapply the process at different resolutions (e.g., hierarchical clustering) to refine the clusters

Output Final Clusters

- when modularity optimisation reaches a stable maximum, finalise and output the clusters

Parameters

Resolution Parameter

- adjusts the size of the communities detected
- higher values of gamma often yield smaller localised communities, lower values yield larger, broader communities
- important for tuning clusters to the desired level of granularity

Initial Partitioning (Starting Clusters)

- determines the starting point for the modularity optimisation process
- some methods start with each node as its own cluster, while others use pre-defined partitions to initialise the process
- can affect the convergence and final clustering results

Edge Weights

- defines the weight or strength of the connections between nodes, where available
- higher weights indicate stronger connections, which can increase likelihood of nodes staying in the same community
- edge weights are particularly useful for networks with varying strengths of relationships, such as social networks

Maximum Iterations

- sets a limit on the number of iterations for the optimisation algorithm
- ensures that the algorithm does not run indefinitely and helps manage computational cost

Stopping Criterion (Convergence Threshold)

- specifies the minimum change in modularity required to continue iterations
- when the change in modularity between iterations falls below this threshold, the algorithm stops

Random Seed (for stochastic methods)

- controls the random initialisation for algorithms that have a stochastic component, such as the Louvain method
- ensuring a consistent random seed can make results reproducible

Algorithm Type

- choice of modularity-based algorithm, such as Louvain, Leiden, or Girvan-Newman
- different algorithms have unique approaches to modularity optimisation, affecting speed, accuracy, and scalability

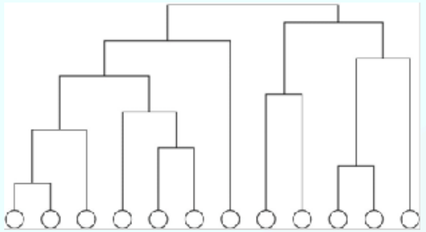
Modularity Function Type

- specifies the formula for modularity, as there can be variations
- the choice depends on the network structure and the specific goals of the clustering

Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the National Academy of Sciences 99.12 (2002): 7821-7826

Clustering Networks - Hierarchical Clustering

agglomerative - (bottom-up)



classic hierarchical clustering

General Approach

Initialise clusters
- treat each data point as its own cluster (single cluster for each data point)

Compute initial distances
- calculate the distance between every pair of clusters
- with average linkage, the distance between clusters is the average distance between each point in one cluster and each point in the other

Merge closest clusters
- identify the two clusters with the smallest average linkage distance and merge them to form a new cluster

Update distances
- recalculate the distances between the new cluster and all remaining clusters using average linkage

Repeat merging
- continue merging the closest clusters and updating distances until only one cluster remains or until a desired number of clusters is reached

Build dendrogram
- track the order of merges to create a dendrogram, a tree-like structure that shows the hierarchical relationships between clusters

Determine clusters
- based on the dendrogram, decide on the final clustering by “cutting” the dendrogram at the desired level

Distance Metrics

Euclidean Distance
- measures the straight-line distance between two points in Euclidean space
- commonly used in continuous, numeric data

Manhattan Distance
- computes the sum of absolute differences between points
- useful when dealing with high-dimensional data

Cosine Similarity
- measures the cosine of the angle between two non-zero vectors, indicating how aligned they are
- often used in text analysis and high-dimensional spaces

Jaccard Similarity
- measures the proportion of common elements in two sets over the union of the sets.
- commonly used for binary and categorical data, especially in set-based data.

Pearson Correlation
- measures the linear relationship between two variables, ranging from -1 to 1
- useful for continuous data where the focus is on correlation rather than direct distance

Distance
- counts the number of positions at which two binary strings differ
- ideal for categorical or binary data, such as in DNA sequence analysis

Biological Interpretation Using Enrichment Analysis

Functional enrichment analysis on biological networks aims to identify and interpret the biological significance of groups of genes, proteins, or other biomolecules that are highly interconnected or co-expressed

Identify Overrepresented Functions
detect biological functions, pathways, or processes that are significantly enriched among the interconnected biomolecules in a network, suggesting shared functional roles

Underlying Biological Mechanisms
uncover the molecular mechanisms that drive specific biological responses, conditions, or diseases by analysing clusters within the network

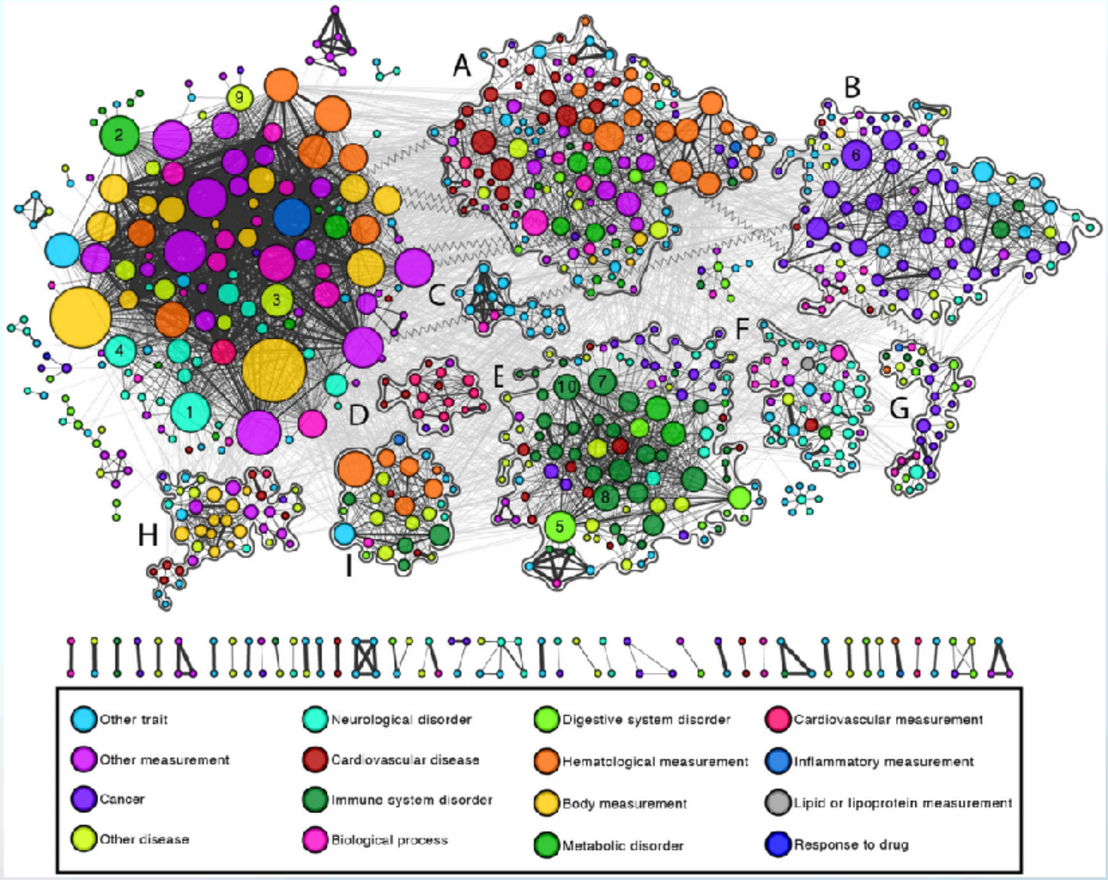
Network Modules
assign biological meaning to distinct network modules or communities, linking them to known biological processes, cellular functions, or molecular pathways

Novel Functional Associations
predict new functional associations or pathways by examining less characterised nodes within clusters, potentially leading to hypotheses about unknown gene or protein functions

Biomarker Discovery
identify key molecules within enriched pathways that could serve as biomarkers for diagnostics or therapeutic targets in specific conditions or diseases

Hypothesis Generation
provide insights for further experimental work by highlighting functionally enriched pathways or gene sets worth exploring in more detail

Data Interpretation
aid in the interpretation of high-throughput biological data (like gene expression or proteomics data) by linking molecular network components to established biological knowledge



Ferolito B, do Valle IF, Gerlovin H, Costa L, Casas JP, Gaziano JM, Gagnon DR, Begoli E, Barabási AL, Cho K. Visualizing novel connections and genetic similarities across diseases using a network-medicine based approach. Sci Rep. 2022 Sep 1;12(1):14914. doi: 10.1038/s41598-022-19244-y

Over Representation Analysis (ORA)

- consideration of foreground and background lists is crucial
  - genome ?
  - array content ?
  - mappable elements ?

contingency table	foreground	background	Row Total
genes with Term in list	a 10	b 50	60
genes with Term not in list	c 190	d 13950	14140
Column Total	200	14000	n 14200

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

probability given by the hypergeometric distribution

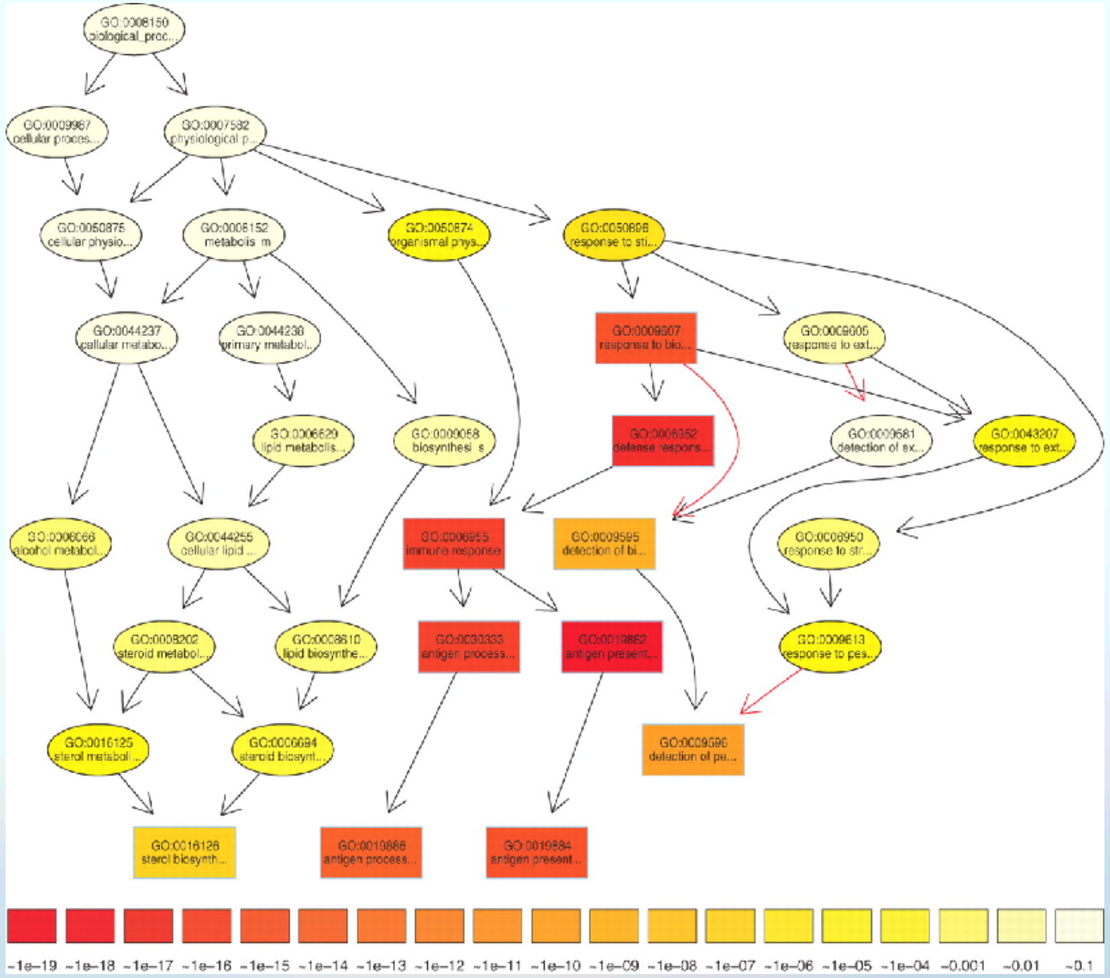
in order to calculate significance we need to

- one-tailed test
  sum all probabilities as extreme or more extreme
- two-tailed test
  sum as above but also all that are as extreme or more extreme in both directions

p-value = 1.003×10^{-8} odds ratio = 14.68

Over Representation Analysis (ORA)

Transitivity Means Terms are not Statistically Independent



Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by de-correlating GO graph structure. Bioinformatics. 2006 Jul 1;22(13):1600-7. doi: 10.1093/bioinformatics/btl140.

Over Representation Analysis (ORA) - Topology Aware

Algorithm 1 elim

markedGenes ← Ø: *nodeSig* ← Ø
get the DAG levels list *dagLevels*
for *i* **from** max(*dagLevels*) **to** 1
  for *u* **in** *nodes*(*dagLevels*, *i*)
    genes[*u*] ← *genes*[*u*]*markedGenes*[*u*]
    nodeSig[*u*] ← *FisherTest*(*genes*[*u*], *sigGenes*)
    if *nodeSig*[*u*] ≤ *thresheold* **then**
      for *x* **in** *upperInducedGraph*(*u*)
        markedGenes[*x*] ← *markedGenes*[*x*] ∪ *genes*[*u*]
  end
end
return *nodeSig*

Problem

- In classical analysis terms are inherited by parents from children
- conditional dependence between nodes invalidates test
- end-up with a “chain” of probability

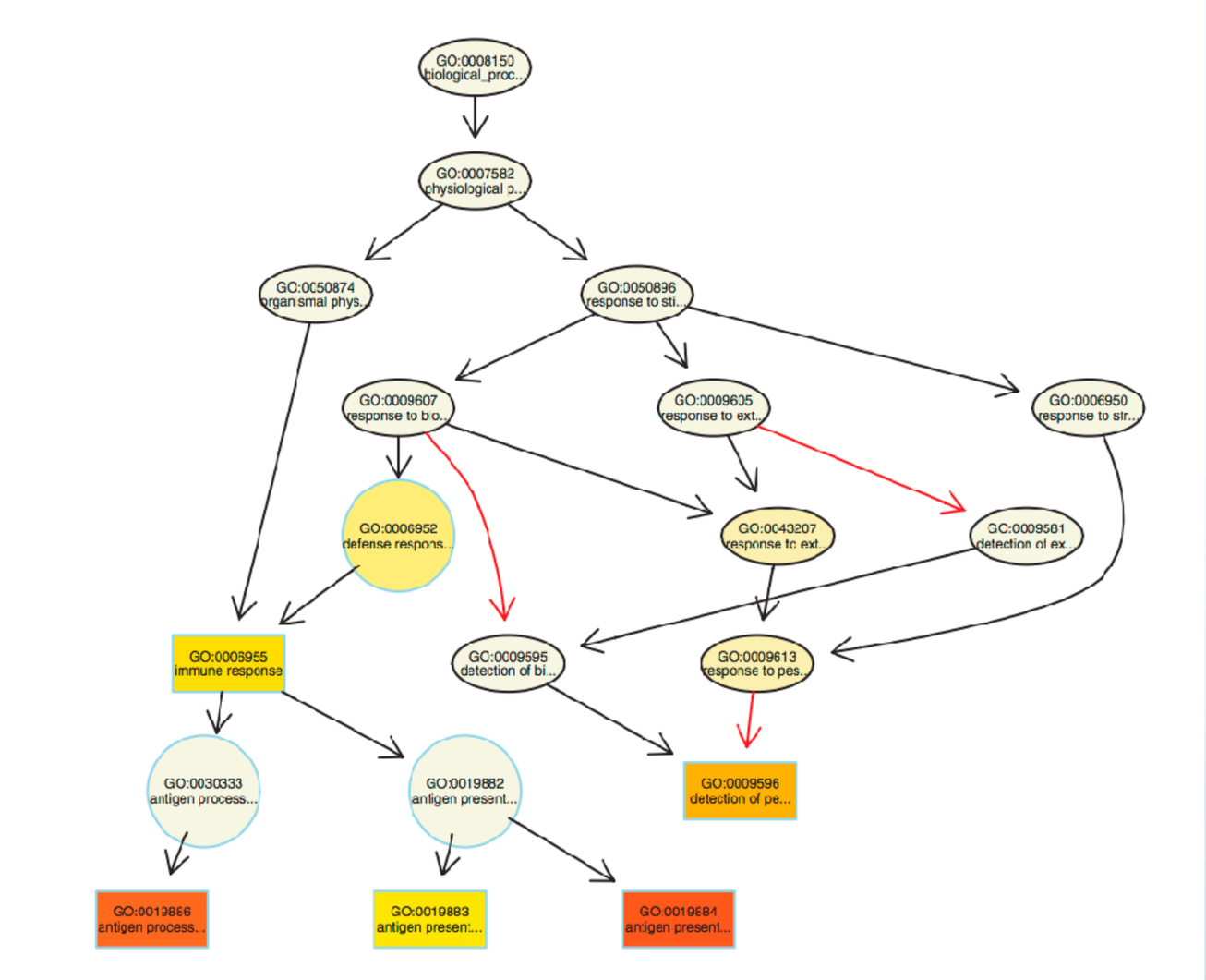
Solution

- de-correlate the GO-graph by removing genes associated with significant nodes from parent term - elimination

	GO ID	Term	Observed	Expected	Annotated	*p*-values				
						classic	elim	weight.log	weight.ratio	all.M
1	GO:0019882	antigen presentation	22	2.287	41	1.6e−17	0.2821	1.6e−17	1.6e−17	1.8e−13
2	GO:0006952	defense response	107	47.143	845	8.3e−17	0.0065	1.4e−09	1.1e−06	5.4e−09
3	GO:0030333	antigen processing	20	2.12	38	7.8e−16	1.0000	7.8e−16	7.8e−16	4.7e−12
4	GO:0006955	immune response	98	43.293	776	2.7e−15	5.9e−06	3.0e−05	0.024	3.3e−07
5	GO:0019884	antigen presentation, exogenou...	14	1.004	18	5.9e−15	5.9e−15	2.2e−10	0.054	1.4e−10
6	GO:0009607	response to biotic stimulus	112	53.949	967	9.5e−15	0.6873	1.0e−05	0.404	1.3e−05
7	GO:0019886	antigen processing, exogenous ...	14	1.116	20	6.8e−14	6.8e−14	1.5e−11	0.054	2.5e−10
8	GO:0009596	detection of pest, pathogen or...	9	0.725	13	2.9e−09	2.9e−09	2.9e−09	2.9e−09	2.9e−09
9	GO:0009595	detection of biotic stimulus	9	0.893	16	3.9e−08	1.0000	1.0e−05	0.107	0.00046
10	GO:0016126	sterol biosynthesis	9	1.395	25	4.5e−06	0.0015	4.5e−06	4.5e−06	1.9e−05

Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by de-correlating GO graph structure. Bioinformatics. 2006 Jul 1;22(13):1600-7. doi: 10.1093/bioinformatics/btl140.

Over Representation Analysis (ORA) - Topology Aware

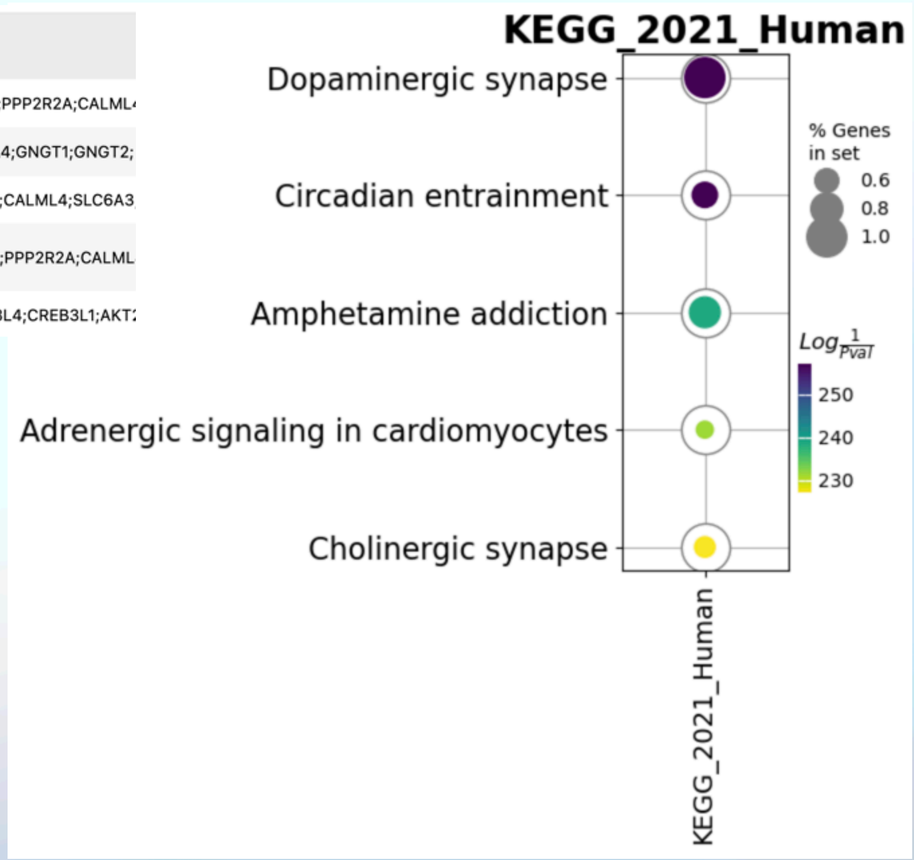


Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by de-correlating GO graph structure. Bioinformatics. 2006 Jul 1;22(13):1600-7. doi: 10.1093/bioinformatics/btl140.

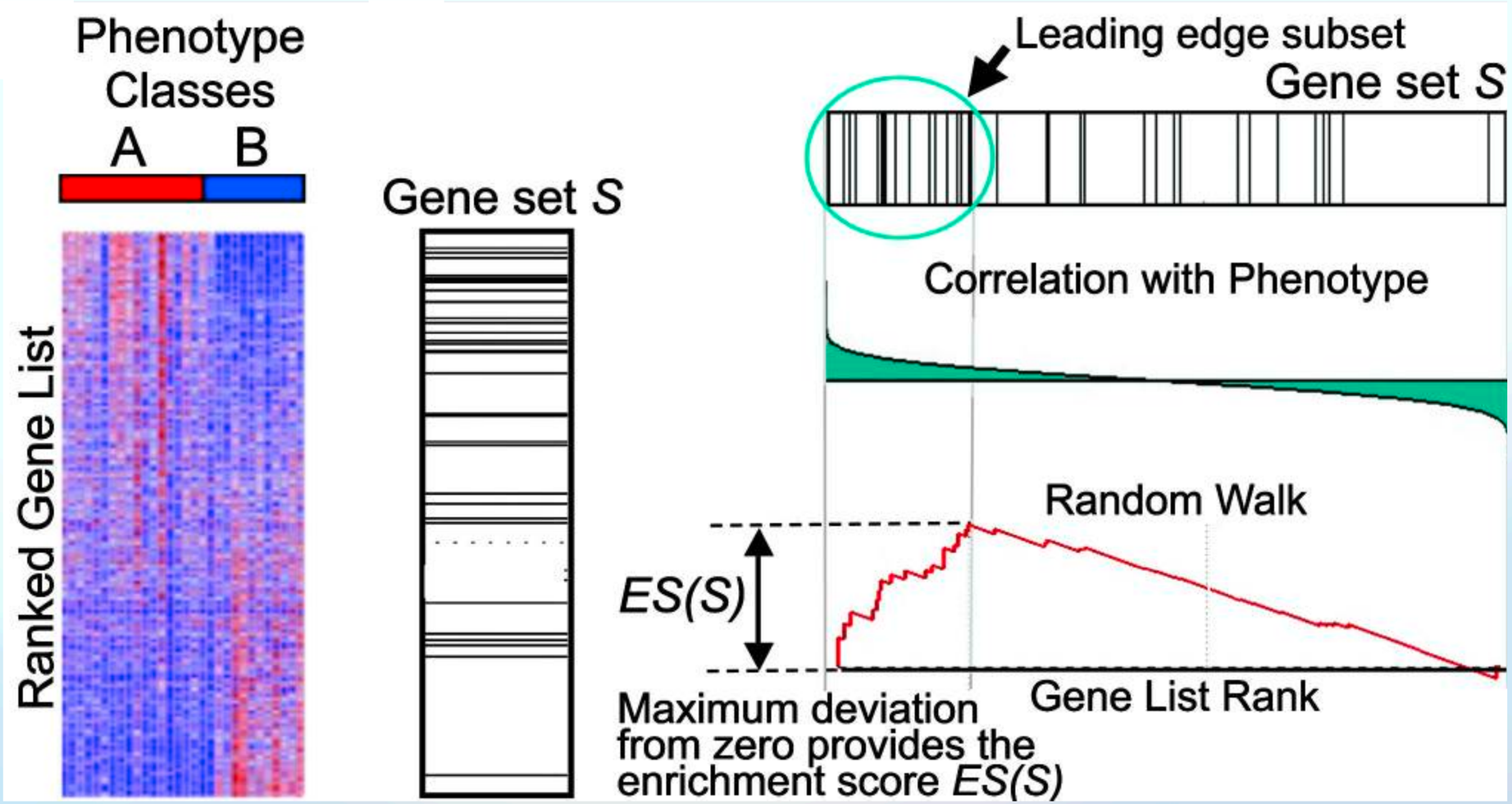
Over Representation Analysis (ORA Example)

ORA example testing the Dopaminergic Synapse gene list against the KEGG pathway database

	Gene_set	Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Odds Ratio	Combined Score	
0	KEGG_2021_Human	Dopaminergic synapse	132/132	0.000000e+00	0.000000e+00	0	0	2.622576e+06	inf	ATF2;CALML5;CALML6;CALML3;PPP2R2A;CALML
1	KEGG_2021_Human	Circadian entrainment	61/97	2.715484e-114	2.634020e-112	0	0	4.732989e+02	1.237656e+05	CALML5;CALML6;CALML3;CALML4;GNGT1;GNGT2;
2	KEGG_2021_Human	Amphetamine addiction	53/69	2.598422e-106	1.680313e-104	0	0	8.324019e+02	2.023728e+05	ATF2;CALML5;CALML6;CALML3;CALML4;SLC6A3
3	KEGG_2021_Human	Adrenergic signaling in cardiomyocytes	63/150	7.726235e-103	3.747224e-101	0	0	2.075967e+02	4.881048e+04	ATF2;CALML5;CALML6;CALML3;PPP2R2A;CALML
4	KEGG_2021_Human	Cholinergic synapse	58/113	7.593595e-101	2.946315e-99	0	0	2.823474e+02	6.509062e+04	GNGT1;GNGT2;CREB3L3;CREB3L4;CREB3L1;AKT2



Gene Set Enrichment Analysis (GSEA)



Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.
Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50

Gene Set Enrichment Analysis (GSEA)

Input data

- Expression data set D with N genes and k samples
- Ranking procedure to produce Gene List L against a phenotype or profile of interest C.
- Independently derived Gene Set S of NH genes (e.g., a pathway, a cytogenetic band, or a GO category).

Enrichment Score (ES)

- Rank order the N genes in D to form L = {g1,...,gN} according to the correlation of their expression profiles with C.
- Evaluate the fraction of genes in S (“hits”) weighted by their correlation and the fraction of genes not in S (“misses”) present up to a given position i in L.
- The ES is the maximum deviation from zero of $P_{hit} - P_{miss}$.

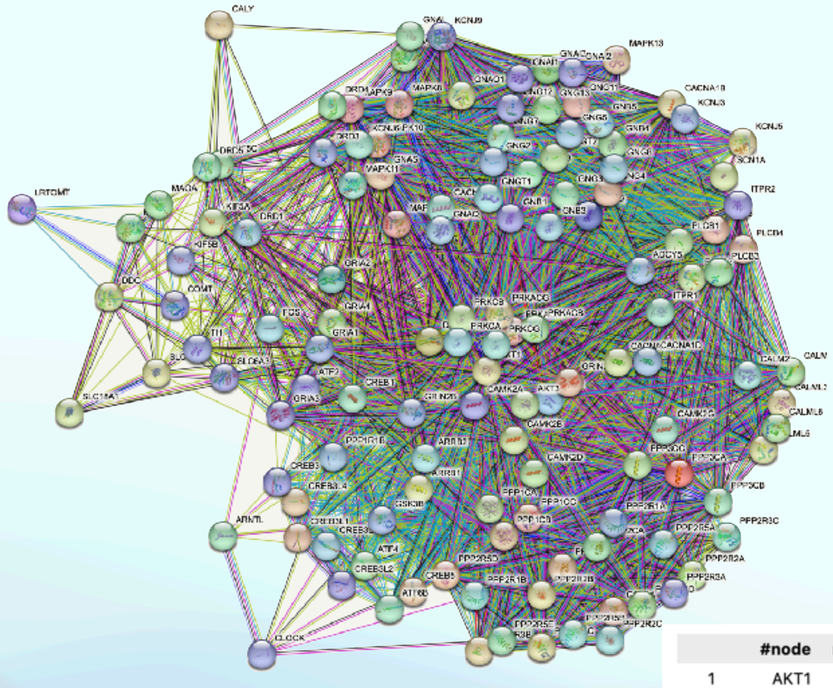
Estimating significance

- Randomly assign the original phenotype labels to samples, reorder genes, and re-compute ES(S).
- Repeat for 1,000 permutations, and create a histogram of the corresponding enrichment scores ES_{Null}.
- Estimate nominal P value for S from ES_{Null} by using the positive or negative portion of the distribution corresponding to the sign of the observed ES(S).

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.
Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50

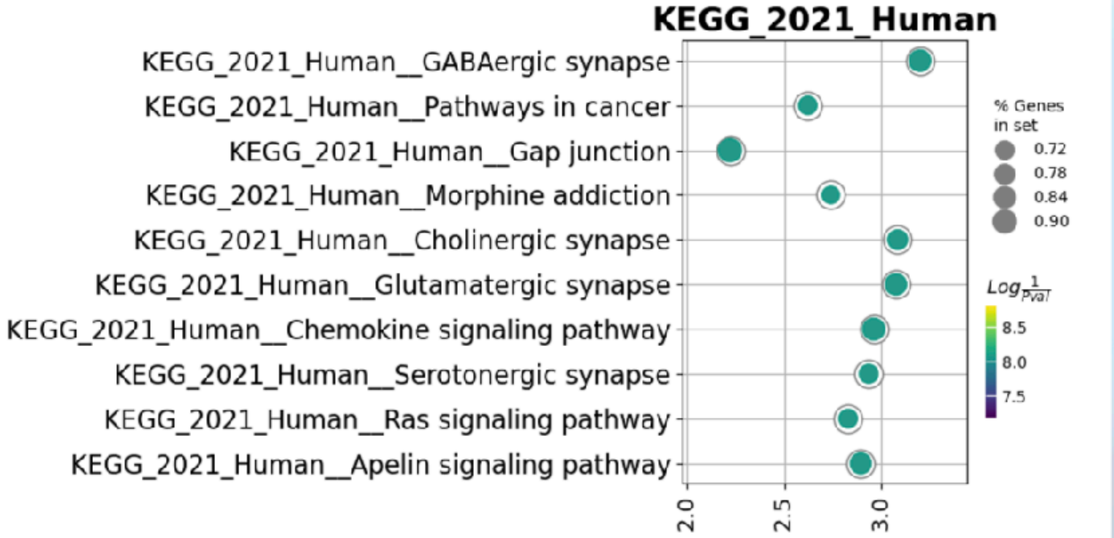
Gene Set Enrichment Analysis (GSEA Example)

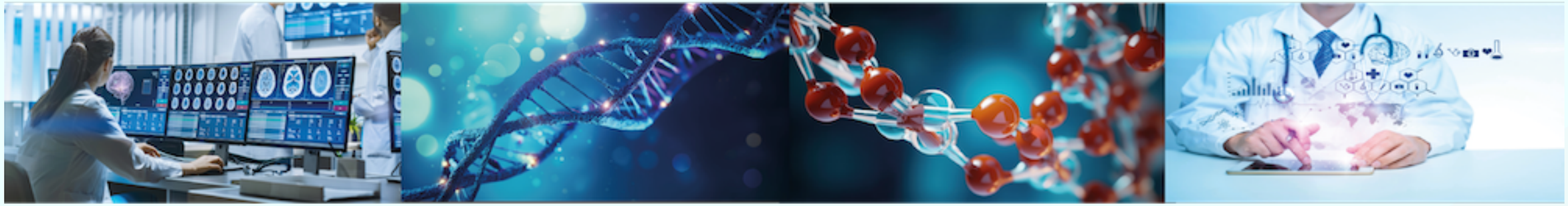
Created a protein-protein interaction network from the Dopaminergic Synapse protein list and used node-degree to rank the genes - then we do a pre-rank GSEA analysis



	#node	node_degree
1	AKT1	82
12	CACNA1C	75
13	CACNA1D	68
35	DRD2	66
120	PRKACB	64
123	PRKCB	63
121	PRKACG	63
122	PRKCA	62
119	PRKACA	62
72	ITPR1	61

	Name	Term	ES	NES	NOM p-val	FDR q-val	FWER p-val	Tag %	Gene %	Lead_genes
0	prerank	KEGG_2021_Human__GABAergic synapse	0.843099	3.197717	0.0	0.0	0.0	28/33	24.62%	CACNA1C;CACNA1D;PRKACB;PRKCB;PRKACG;PRKCA;PRKA...
1	prerank	KEGG_2021_Human__Cholinergic synapse	0.77439	3.083102	0.0	0.0	0.0	45/58	40.00%	AKT1;CACNA1C;CACNA1D;PRKACB;PRKCB;PRKACG;PRKCA...
2	prerank	KEGG_2021_Human__Glutamatergic synapse	0.77133	3.073702	0.0	0.0	0.0	42/50	40.00%	CACNA1C;CACNA1D;PRKACB;PRKCB;PRKACG;PRKCA;PRKA...
3	prerank	KEGG_2021_Human__Chemokine signaling pathway	0.769659	2.96008	0.0	0.0	0.0	32/37	35.38%	AKT1;PRKACB;PRKCB;PRKACG;PRKACA;GNB1;GNB3;GNG3...
4	prerank	KEGG_2021_Human__Serotonergic synapse	0.741352	2.9343	0.0	0.0	0.0	37/50	35.38%	CACNA1C;CACNA1D;PRKACB;PRKCB;PRKACG;PRKCA;PRKA...





Programming for Biomedical Informatics

Next Lecture this Thursday - “Network Analysis in Practice”

Please Bring your Laptop!

Ask Questions on the EdStem Discussion Board

https://github.com/tisimpson/pbi

Coding