BIO214 Lecture 9

Bioinformatics-II

Genomic knowledge representation

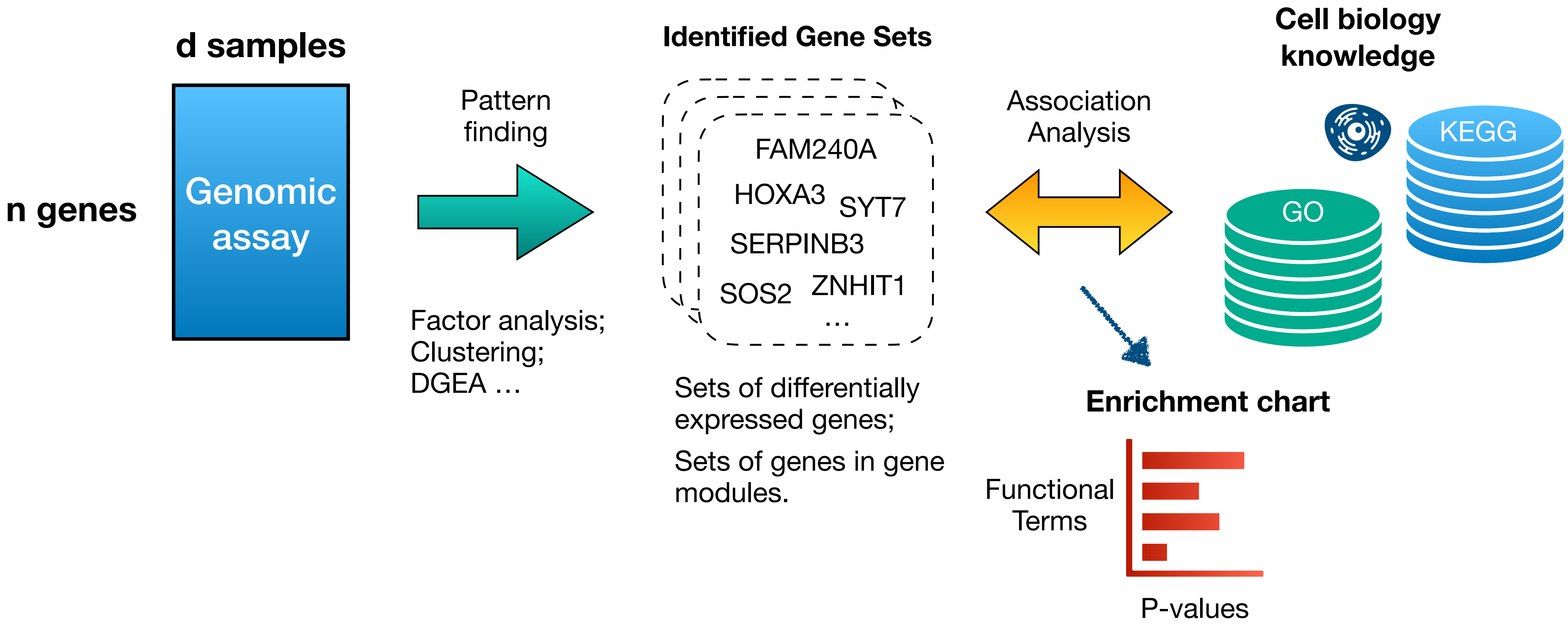
Zhen Wei; 2023-Feb-14

Outline

- Gene set annotation
- Range based annotation
- Network basics
- Co-expression network

Gene set annotation

Why we need functional annotations?



- Annotations are stored knowledge from previous biological experiments.
- Functional annotations are essential for the interpretation of gene sets obtained from the upstream analysis.
- Gene set enrichment is calculated via the statistical association between gene functions and gene sets.

Gene Ontology: use general biological knowledge to annotate genes

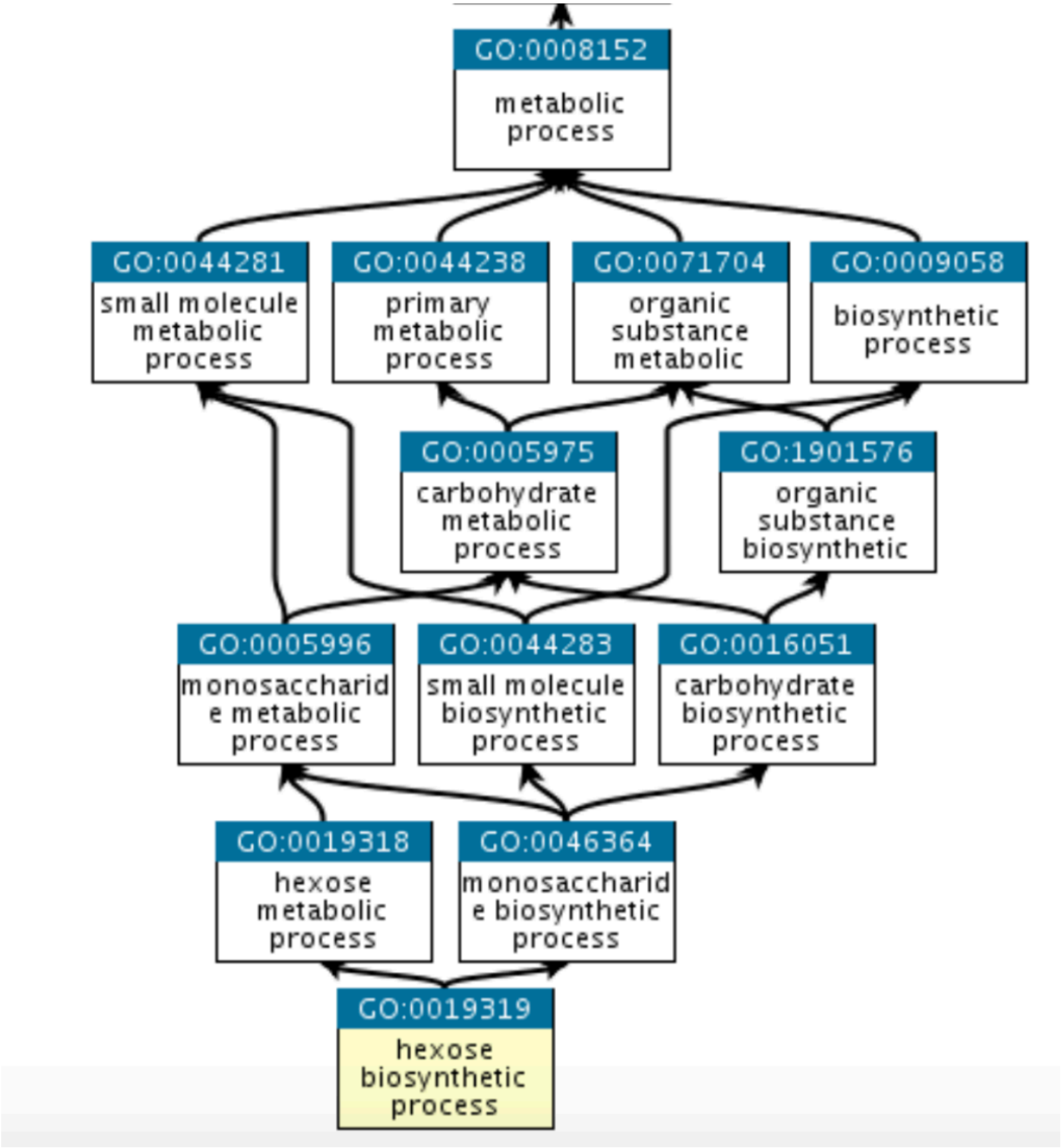
> “An ontology is a formal representation of a body
> of knowledge within a given domain.”
>
> —- From Gene Ontology website

The Gene Ontology (GO) describes our knowledge of the biological domain with respect to three aspects:

- **Molecular Function**: Molecular-level activities performed by gene products.
- **Cellular Component**: The locations relative to cellular structures in which a gene product performs a function.
- **Biological Process**: The larger processes, or ‘biological programs’ accomplished by multiple molecular activities.

For example, the gene product “cytochrome c” can be described by the **molecular function** *oxidoreductase activity*, the **biological process** *oxidative phosphorylation*, and the **cellular component** *mitochondrial matrix*.

GO Graph

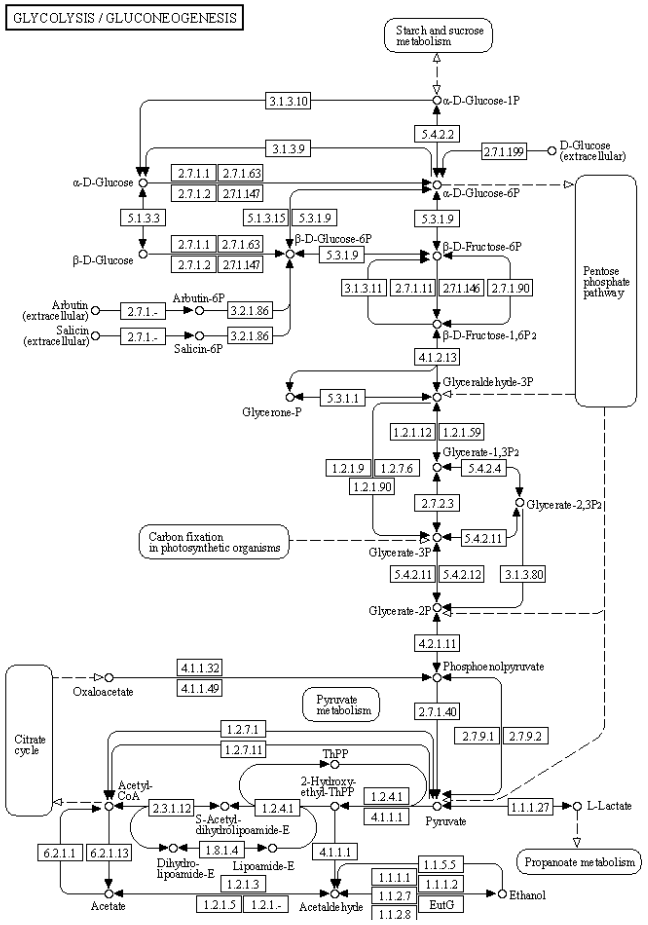

GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

- The Gene Ontology (GO) is represented as a graph with terms as nodes and relationships between terms as edges.
- GO is hierarchical, with more specific child terms and more general parent terms.
- Terms can have multiple parent terms.

KEGG: gene annotation via signaling pathway

- KEGG is a database for understanding biological systems.
- KEGG pathway maps are molecular interaction/reaction networks represented in terms of KEGG Orthology groups.
- These maps can help generalize experimental evidence from one organism to others based on genomic information.

The KEGG pathway map for glycolysis / gluconeogenesis



Calculating statistical association between annotations

	User Genes	Genome
In Pathway	10	30
Not Pathway	4	60

Can be viewed as

	Balls drawed	Balls in urn
Black balls	10	30
White balls	4	60

Hypergeometric pmf

$$P(X) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

- **Fisher's exact test** is often used to calculate p-value of association between gene sets and functional terms.

- The p-value is calculated by the hypergeometric distribution:

  1. Enumerate all possible 2 by 2 tables that are as or more associated than the observed given fixed margins (column and row sums).

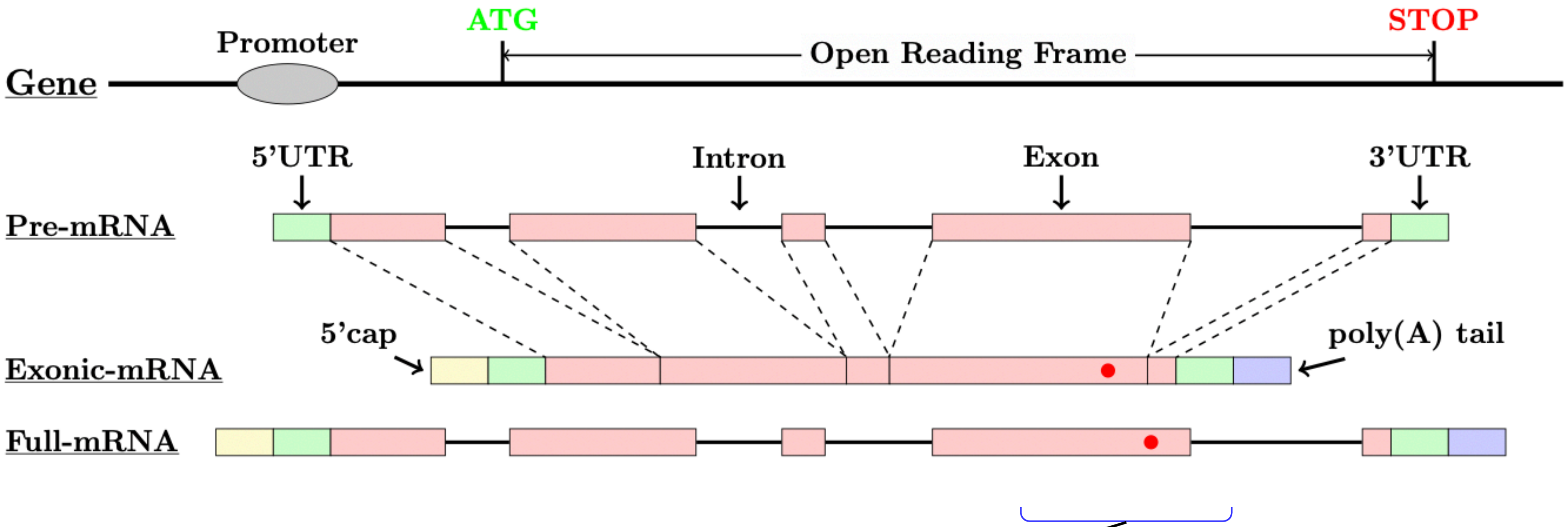
  2. Use hypergeometric distribution to calculate the probabilities of each table, sum them up and you will get the p-value.

Range based annotations

Range based annotations

Transcript annotation

Information stored in transcript annotations (GTF/GFF files)



The genome coordinate of a locus:
(Not necessarily a gene)

Chr9; [1653482–1654482]; + strand

Genomic Feature Extraction

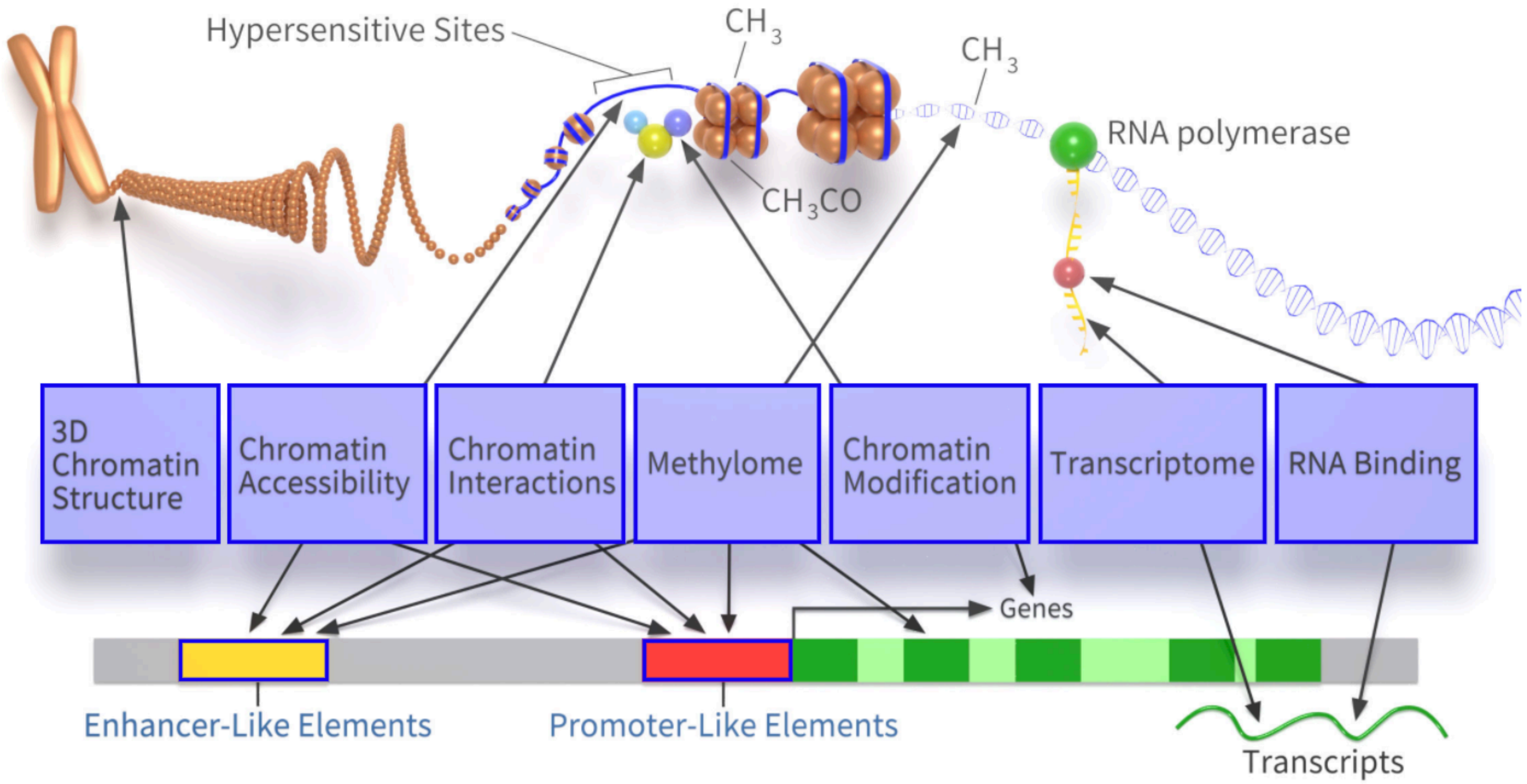
Overlapping with 5'UTR, CDS, 3'UTR, intron, exon or not?

Properties of the overlapped features, e.g. length, GC content, state of evolutionary conservation.

- Gene & Transcript annotations from GTF/GFF files are often used to annotate range based genomic experiments (e.g. peaks from CHIP-Seq).

Other range based annotations

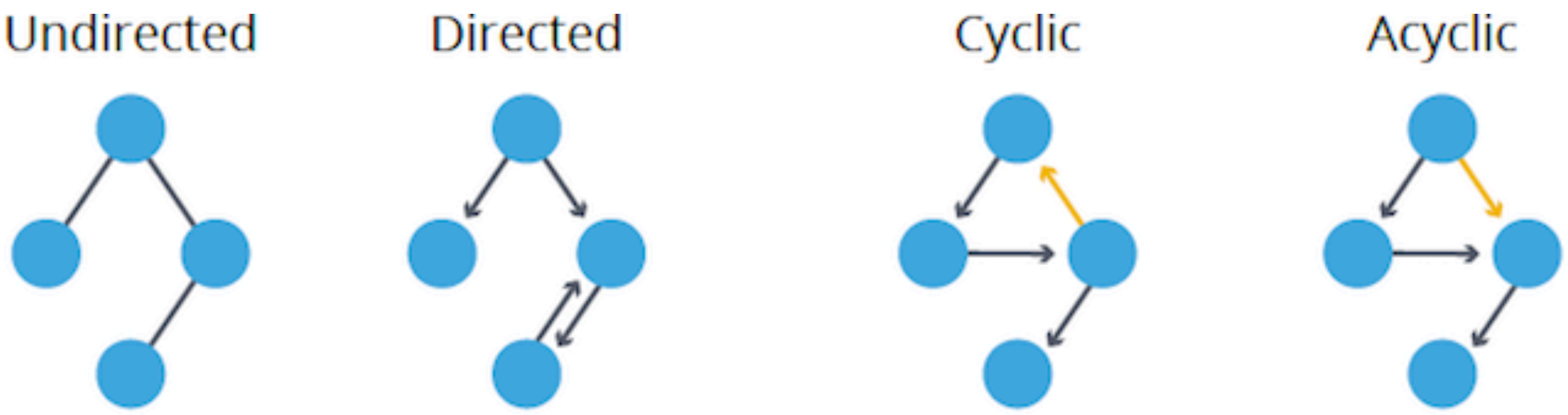
Epigenetic markers



- **ENCODE** stands for ENCyclopedia Of DNA Elements.
- It's a database that collects high-quality data about epigenetic markers, expressed transcripts, and epitranscriptomic markers.
- ENCODE uses strict and well-documented data processing pipelines to ensure data quality.
- Researchers can use the epigenetic markers from ENCODE to annotate their own experiments.

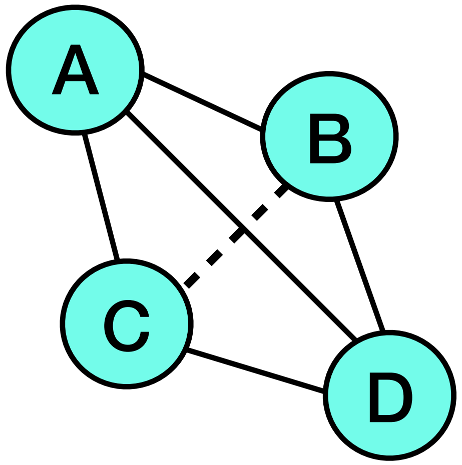
Introduction to biological networks

Correlational v.s. causal gene networks



Types	Description	Example
Correlational graphs (undirected graph)	Represent the positive / negative correlation between genes. The significantly correlated genes are linked by an undirected edge.	PPI network, gene co-expression network
Cause-effect graph (directed graph)	Describe the relationship of causality between genes, such as a gene is changed upon the action of another gene. The direction of the arrowed edge represents cause and effect.	Cell signaling network, epigenetic regulatory network

Representation: adjacency matrix



Adjacency matrix: $\mathbf{A}$

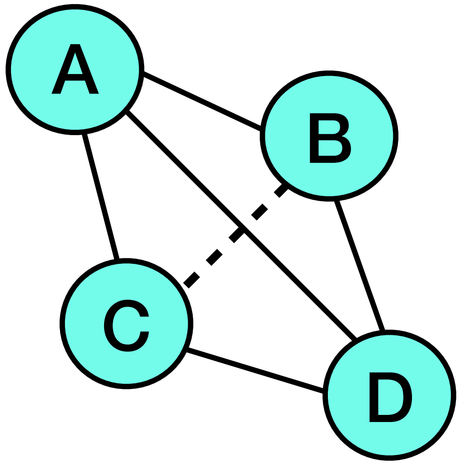
From
(node i)

$$\begin{array}{c|cccc|} & \text{A} & \text{B} & \text{C} & \text{D} \\ \text{A} & 0 & 1 & 1 & 1 \\ \text{B} & 1 & 0 & 1 & 1 \\ \text{C} & 1 & 1 & 0 & 1 \\ \text{D} & 1 & 1 & 1 & 0 \end{array}$$

To
(node j)

- **Nodes:** A, B, C, D
- **Edges**: A <-> B, B <-> C, C <-> D, D <-> A
- The whole network structure can be specified by an $n \times n$ **adjacency matrix**, where n is the number of nodes.
- $\mathbf{A}_{ij} = 1$ if nodes i and j are connected from i to j.
- Can be softly weighted by probabilities, i.e. $\mathbf{A}_{ij} \in [0,1]$

Degree of node



$$\begin{array}{ccccc} & A & B & C & D \\ A & 0 & 1 & 1 & 1 \\ B & 1 & 0 & 1 & 1 \\ C & 1 & 1 & 0 & 1 \\ D & 1 & 1 & 1 & 0 \end{array}$$

Degree of node C = 3

- **Neighbors** are pairs of nodes connected by an edge.
- **Degree** (k) of a node counts the number of edges connecting its neighbors to it.
- Degrees for an undirected graph can be calculated by the row sums of the adjacency matrix.

Random v.s. Scale-free Network

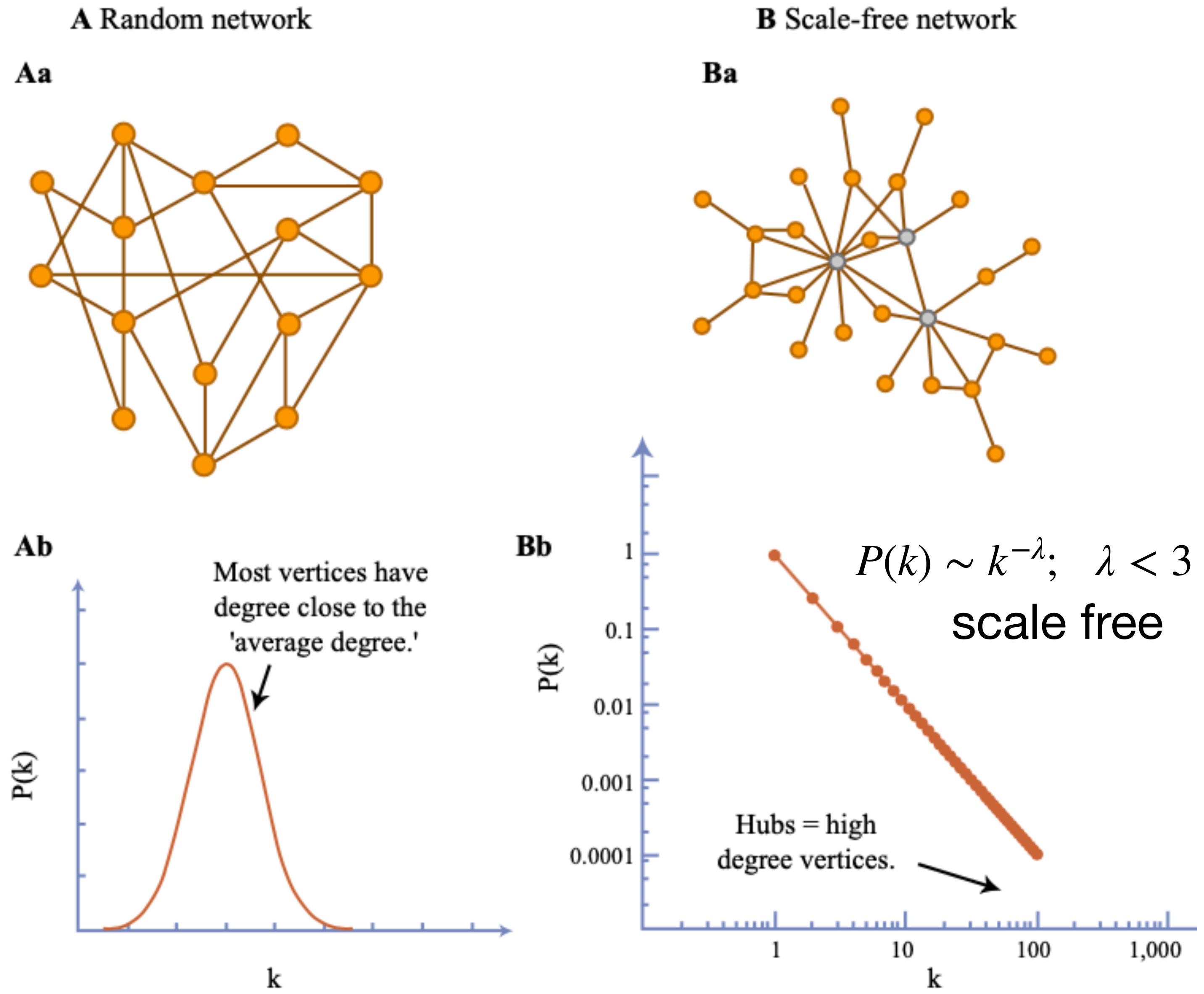
- The distribution of degrees over a graph reveals essential network properties.

- In **random network**, edges are added to node pairs with equal probabilities.

- The degree distribution for random network is Poisson distribution.

- In **scale-free network**, the probability of adding a new edge from node i to a new node increases as the degree of node i increases.

- The degree distribution for scale free network is power distribution.



$P(k) \sim k^{-\lambda}; \quad \lambda < 3 \Rightarrow$ scale free

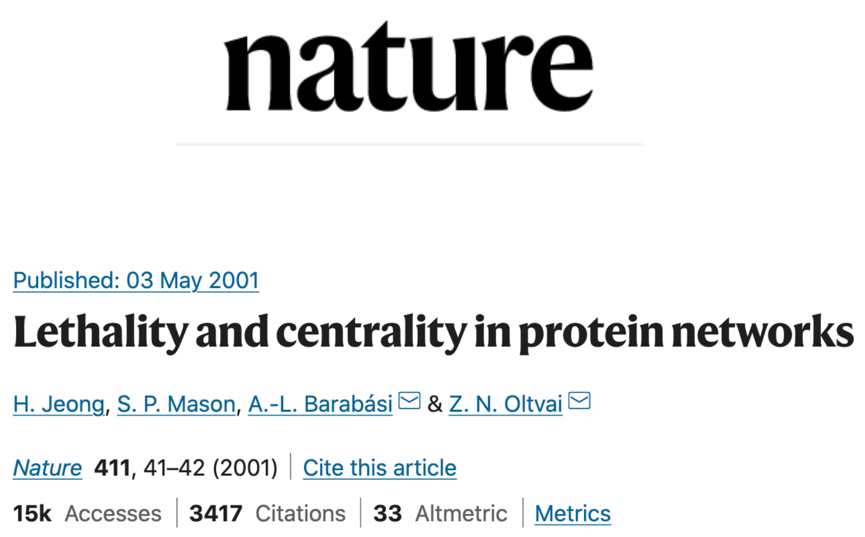
Properties of scale-free network

- Average steps between a random pair of nodes in a graph of size n:
  - For a random network, the average path length is $\sim \log(n)$
  - For a scale-free network, the average path length is $\sim \log(\log(n))$

There by, information transfer is more efficient on a scale-free network.

- When “attacks” are made by removing nodes from the graph:
  - If the failures happened randomly, the scale-free network is more likely to survive than the random network.
  - If the failures are targeted toward the **hub nodes** (the nodes with highest degree), then the scale-free network is more vulnerable than the random network.

Essential proteins are hub-nodes



nature

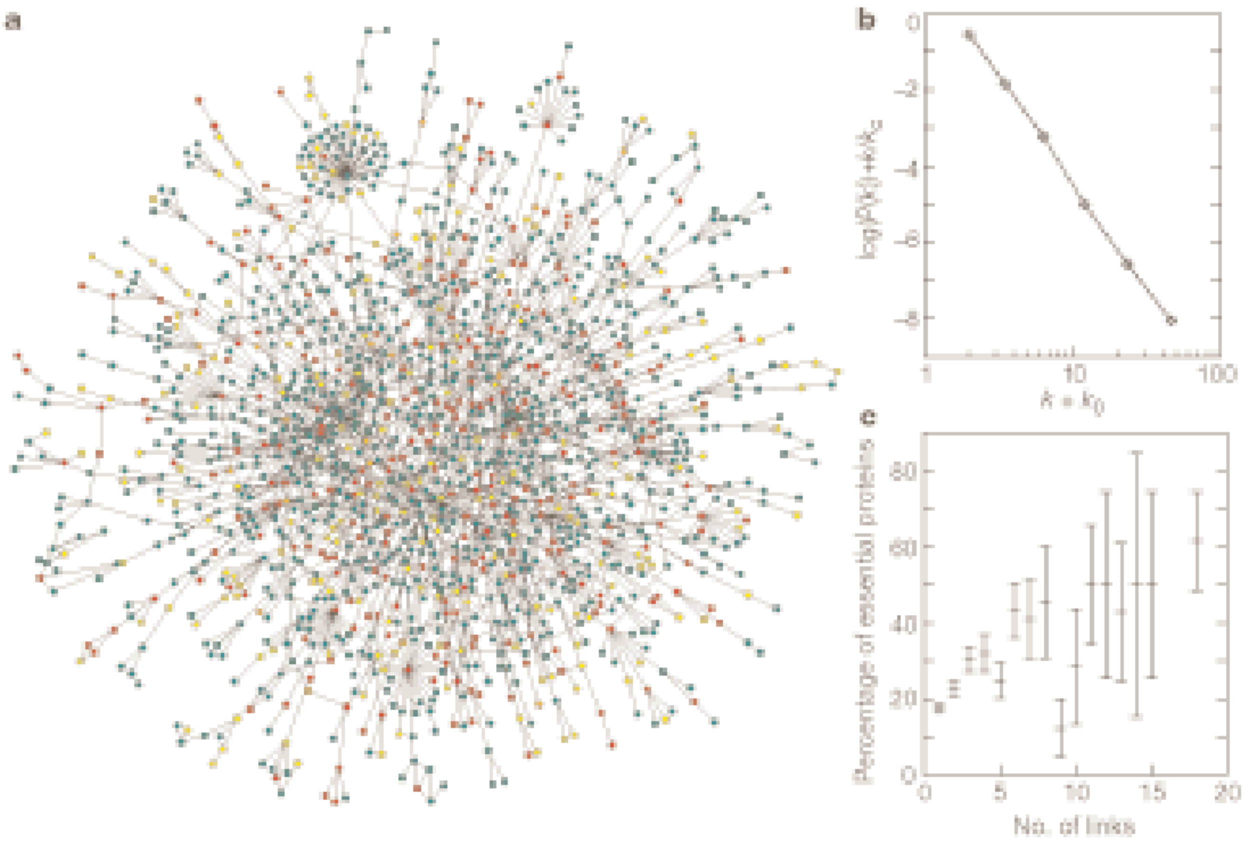
Published: 03 May 2001

Lethality and centrality in protein networks

H. Jeong, S. P. Mason, A.-L. Barabási & Z. N. Oltvai

Nature **411**, 41–42 (2001) | Cite this article

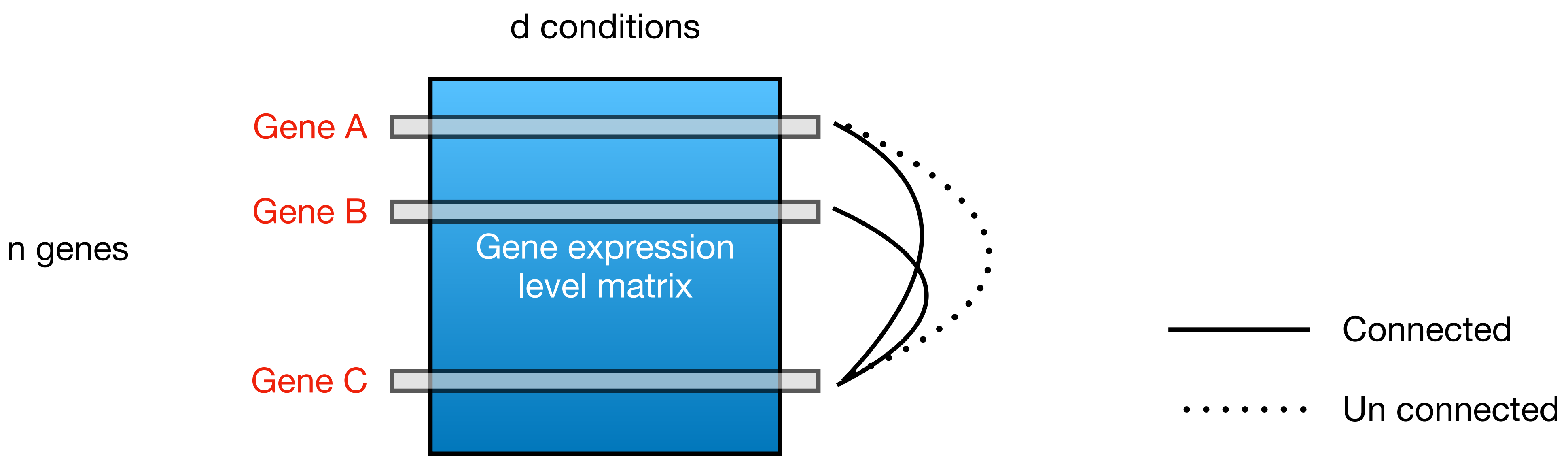
15k Accesses | **3417** Citations | **33** Altmetric | Metrics



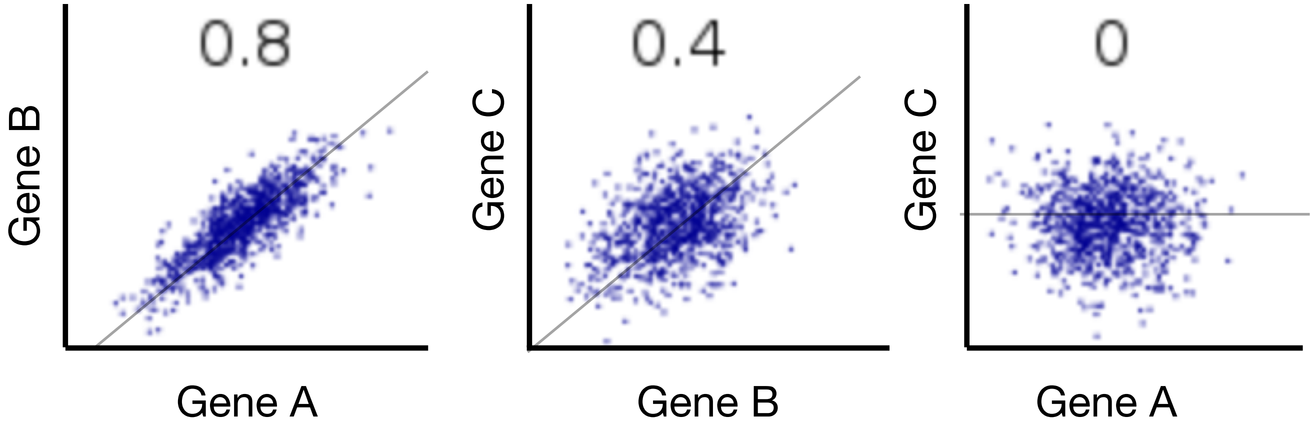
- A protein is essential if its knock-down is lethal.
- In yeast PPI network, the proteins with higher degree (more direct interactions with other proteins) are more likely to be essential proteins.
- 2240 edges are formed among 1870 nodes (proteins) in yeast PPI network.
- 93% of proteins have degrees < 3, among them, 21% are essential to yeast survival.
- 0.7 % of proteins have > 15 degree, and 62% of those are essential.
- The overall correlation coefficient between lethality and connectivity is 0.76.

Co-expression network

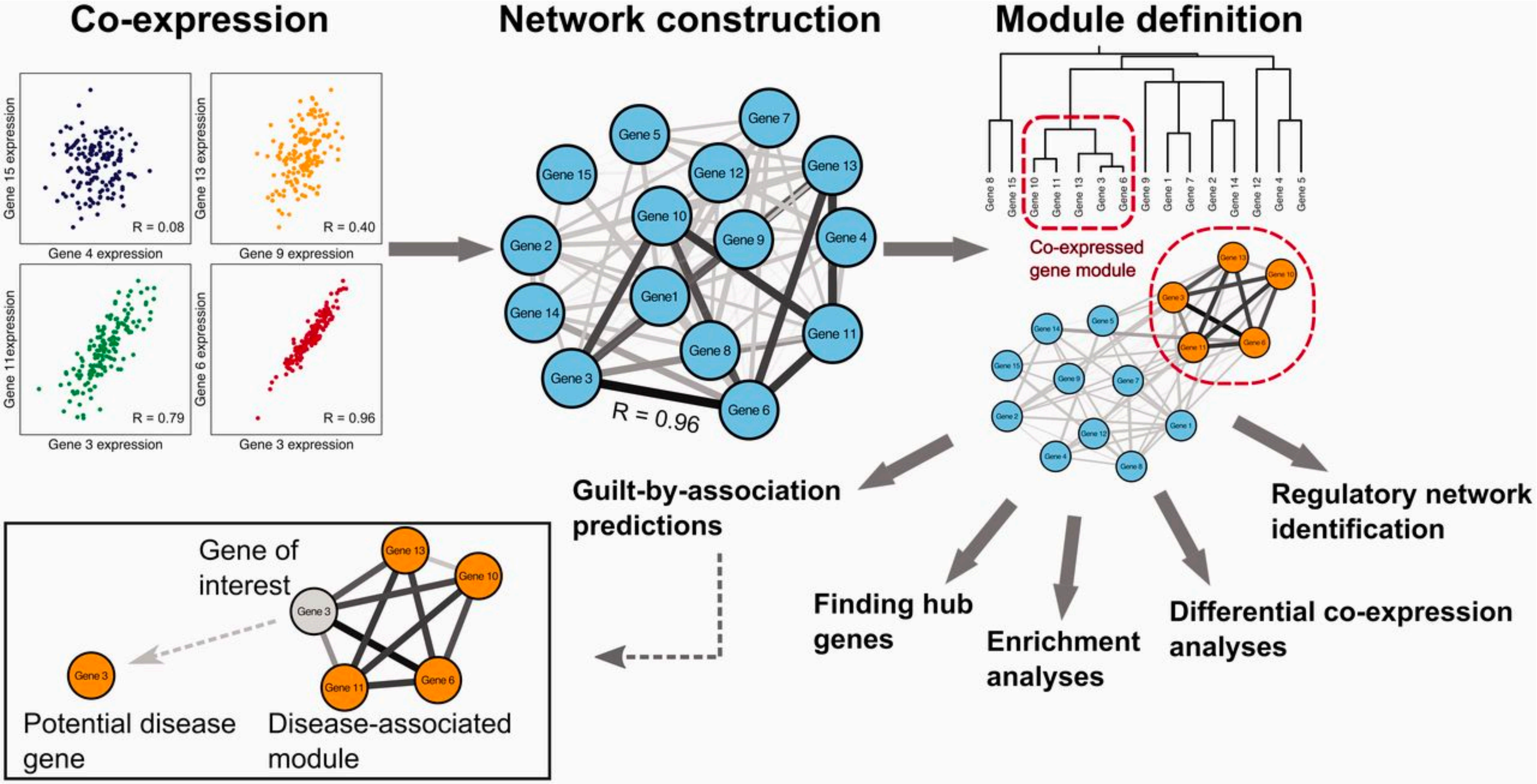
How to construct gene network from gene expression levels?



Pearson correlation values:

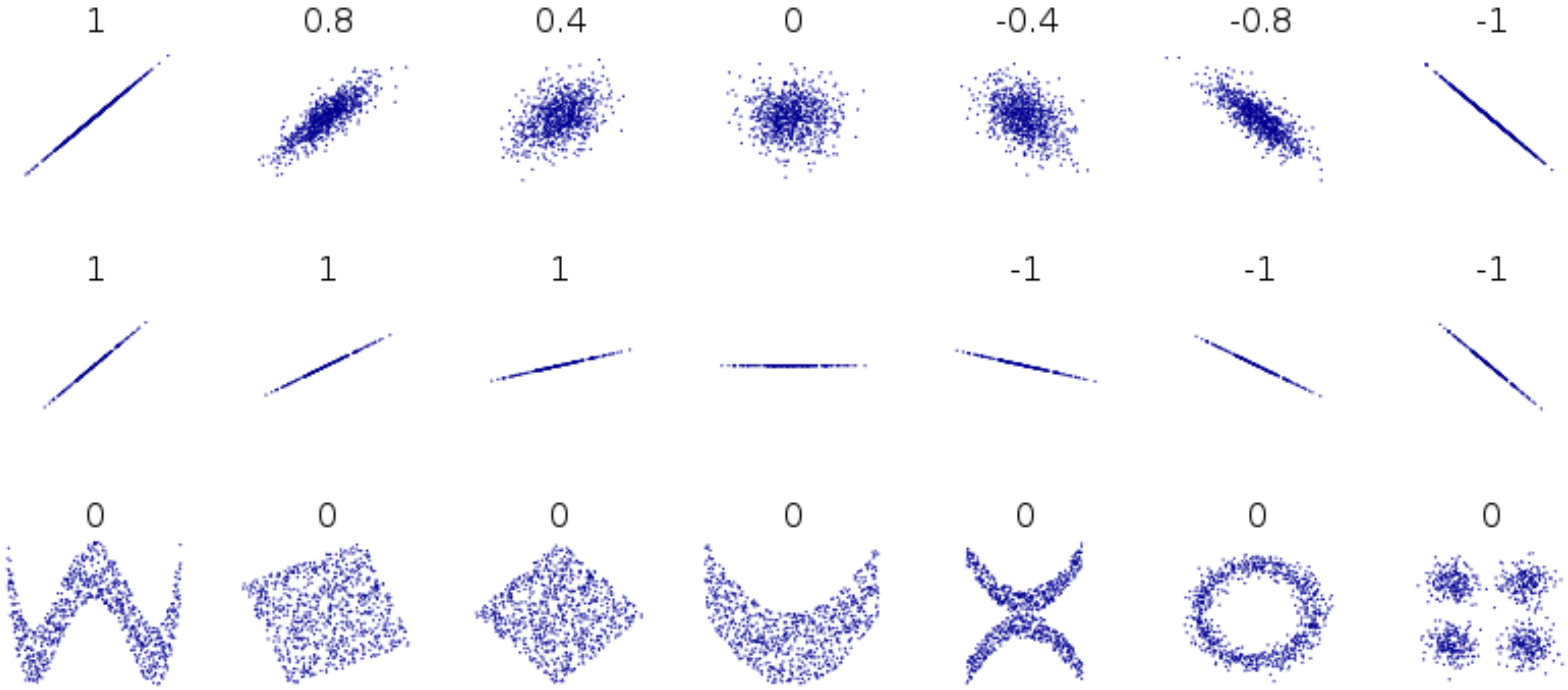


Workflow of co-expression network analysis



- Pairwise correlation used to construct network
- Clustering identifies modules
- Differential co-expression analysis identifies regulatory genes
- Guilt-by-association approach identifies potential disease genes

Limitation of Pearson correlation



- Pearson correlation cannot capture non-linear interaction (last row).

https://en.wikipedia.org/wiki/Correlation

Performances of different network inference methods

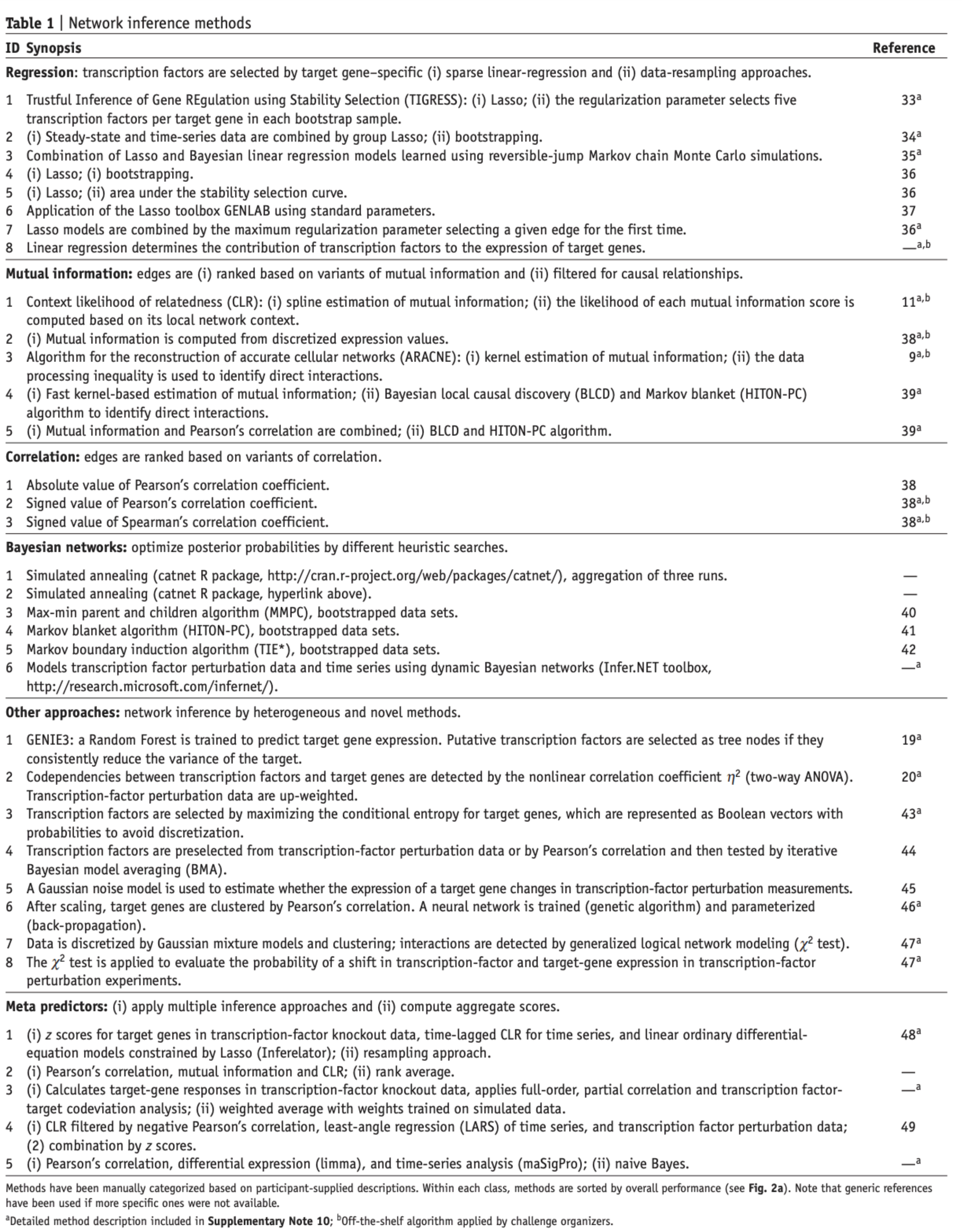


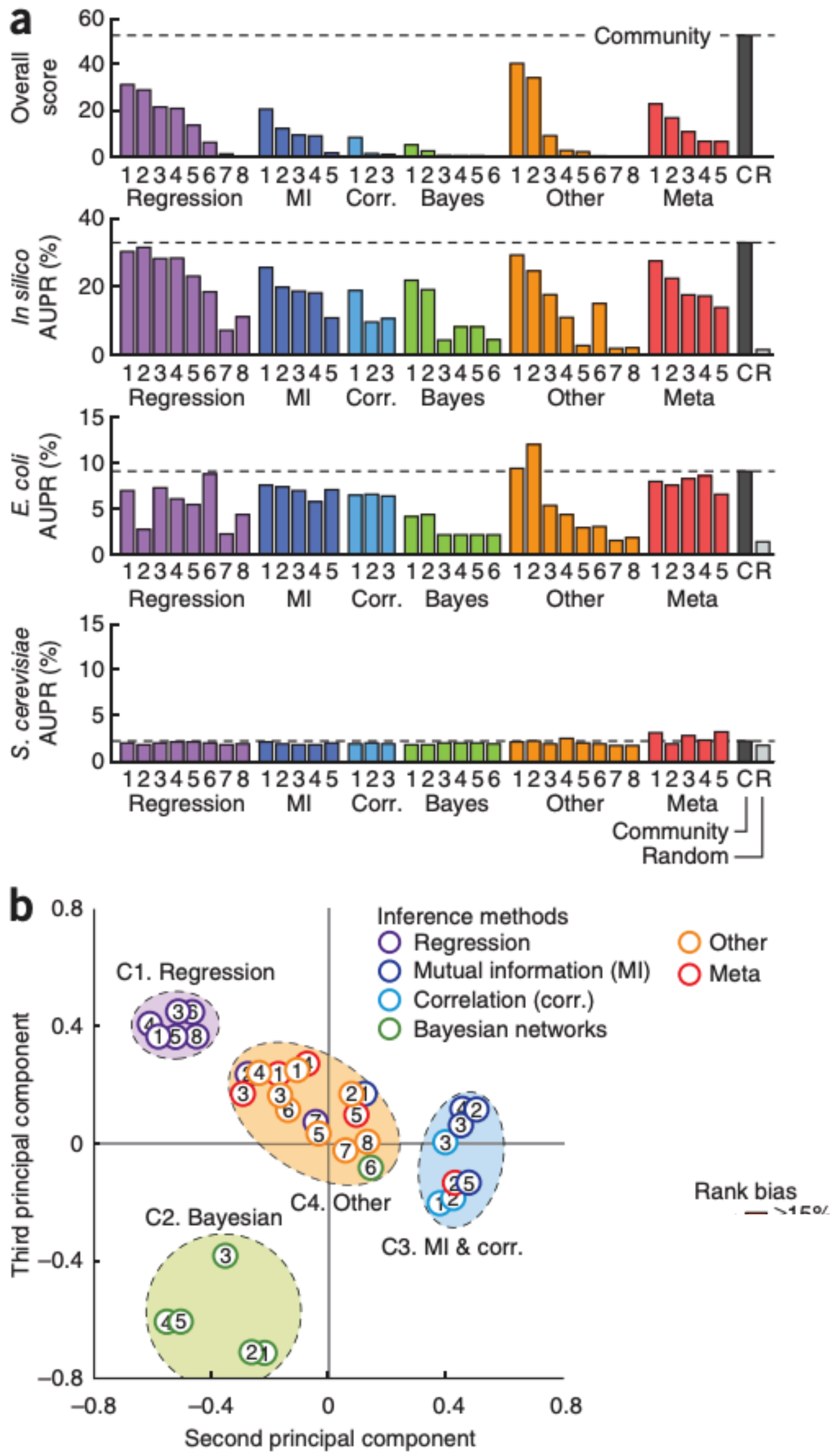
Table 1 | Network inference methods

ID	Synopsis	Reference
	Regression: transcription factors are selected by target gene–specific (i) sparse linear-regression and (ii) data-resampling approaches.	
1	Trustful Inference of Gene REgulation using Stability Selection (TIGRESS): (i) Lasso; (ii) the regularization parameter selects five transcription factors per target gene in each bootstrap sample.	33[a]
2	(i) Steady-state and time-series data are combined by group Lasso; (ii) bootstrapping.	34[a]
3	Combination of Lasso and Bayesian linear regression models learned using reversible-jump Markov chain Monte Carlo simulations.	35[a]
4	(i) Lasso; (i) bootstrapping.	36
5	(i) Lasso; (ii) area under the stability selection curve.	36
6	Application of the Lasso toolbox GENLAB using standard parameters.	37
7	Lasso models are combined by the maximum regularization parameter selecting a given edge for the first time.	36[a]
8	Linear regression determines the contribution of transcription factors to the expression of target genes.	—[a,b]
	Mutual information: edges are (i) ranked based on variants of mutual information and (ii) filtered for causal relationships.	
1	Context likelihood of relatedness (CLR): (i) spline estimation of mutual information; (ii) the likelihood of each mutual information score is computed based on its local network context.	11[a,b]
2	(i) Mutual information is computed from discretized expression values.	38[a,b]
3	Algorithm for the reconstruction of accurate cellular networks (ARACNE): (i) kernel estimation of mutual information; (ii) the data processing inequality is used to identify direct interactions.	9[a,b]
4	(i) Fast kernel-based estimation of mutual information; (ii) Bayesian local causal discovery (BLCD) and Markov blanket (HITON-PC) algorithm to identify direct interactions.	39[a]
5	(i) Mutual information and Pearson's correlation are combined; (ii) BLCD and HITON-PC algorithm.	39[a]
	Correlation: edges are ranked based on variants of correlation.	
1	Absolute value of Pearson's correlation coefficient.	38
2	Signed value of Pearson's correlation coefficient.	38[a,b]
3	Signed value of Spearman's correlation coefficient.	38[a,b]
	Bayesian networks: optimize posterior probabilities by different heuristic searches.	
1	Simulated annealing (catnet R package, http://cran.r-project.org/web/packages/catnet/), aggregation of three runs.	—
2	Simulated annealing (catnet R package, hyperlink above).	—
3	Max-min parent and children algorithm (MMPC), bootstrapped data sets.	40
4	Markov blanket algorithm (HITON-PC), bootstrapped data sets.	41
5	Markov boundary induction algorithm (TIE*), bootstrapped data sets.	42
6	Models transcription factor perturbation data and time series using dynamic Bayesian networks (Infer.NET toolbox, http://research.microsoft.com/infernet/).	—[a]
	Other approaches: network inference by heterogeneous and novel methods.	
1	GENIE3: a Random Forest is trained to predict target gene expression. Putative transcription factors are selected as tree nodes if they consistently reduce the variance of the target.	19[a]
2	Codependencies between transcription factors and target genes are detected by the nonlinear correlation coefficient η^2 (two-way ANOVA). Transcription-factor perturbation data are up-weighted.	20[a]
3	Transcription factors are selected by maximizing the conditional entropy for target genes, which are represented as Boolean vectors with probabilities to avoid discretization.	43[a]
4	Transcription factors are preselected from transcription-factor perturbation data or by Pearson's correlation and then tested by iterative Bayesian model averaging (BMA).	44
5	A Gaussian noise model is used to estimate whether the expression of a target gene changes in transcription-factor perturbation measurements.	45
6	After scaling, target genes are clustered by Pearson's correlation. A neural network is trained (genetic algorithm) and parameterized (back-propagation).	46[a]
7	Data is discretized by Gaussian mixture models and clustering; interactions are detected by generalized logical network modeling (χ^2 test).	47[a]
8	The χ^2 test is applied to evaluate the probability of a shift in transcription-factor and target-gene expression in transcription-factor perturbation experiments.	47[a]
	Meta predictors: (i) apply multiple inference approaches and (ii) compute aggregate scores.	
1	(i) z scores for target genes in transcription-factor knockout data, time-lagged CLR for time series, and linear ordinary differential-equation models constrained by Lasso (Inferelator); (ii) resampling approach.	48[a]
2	(i) Pearson's correlation, mutual information and CLR; (ii) rank average.	—
3	(i) Calculates target-gene responses in transcription-factor knockout data, applies full-order, partial correlation and transcription factor-target codeviation analysis; (ii) weighted average with weights trained on simulated data.	—[a]
4	(i) CLR filtered by negative Pearson's correlation, least-angle regression (LARS) of time series, and transcription factor perturbation data; (2) combination by z scores.	49
5	(i) Pearson's correlation, differential expression (limma), and time-series analysis (maSigPro); (ii) naive Bayes.	—[a]

Methods have been manually categorized based on participant-supplied descriptions. Within each class, methods are sorted by overall performance (see **Fig. 2a**). Note that generic references have been used if more specific ones were not available.

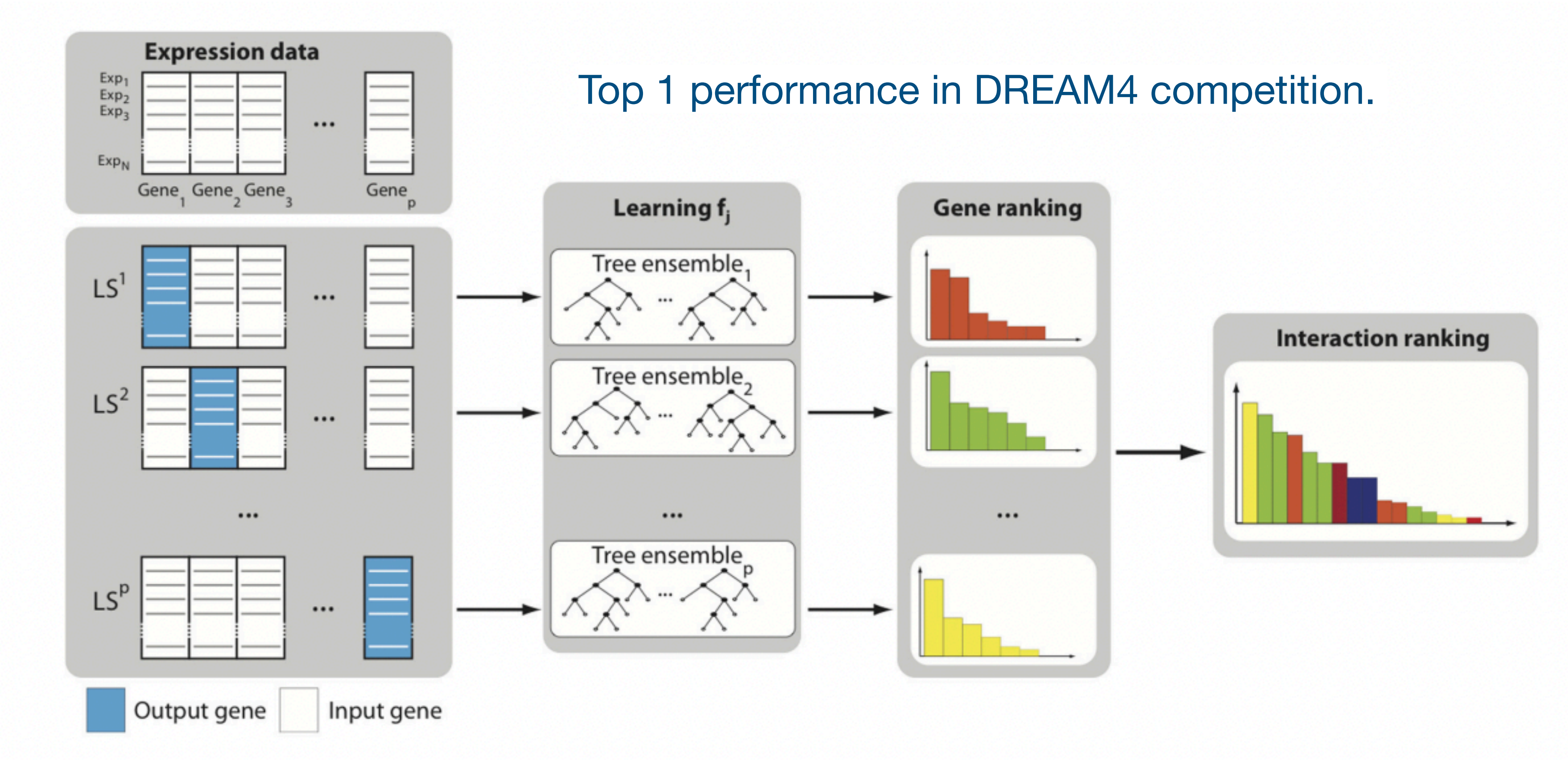
[a]Detailed method description included in **Supplementary Note 10**; [b]Off-the-shelf algorithm applied by challenge organizers.

AUPR: area under precision-recall curve



Marbach, Daniel, et al. "Wisdom of crowds for robust gene network inference." *Nature methods* 9.8 (2012): 796-804.

GINIE3: a high performing network inference algorithm



Top 1 performance in DREAM4 competition.

To create a gene regulatory network in GINIE3:

- For each gene, train Random Forest predictors (f_j) with its expression levels as output and other genes' levels as input.
- For each predictors, rank all input genes by feature importance.
- Combine the rankings of all predictors to get the edge scores for network's regulatory links.

Huynh-Thu, Vân Anh, et al. "Inferring regulatory networks from expression data using tree-based methods." PloS one 5.9 (2010): e12776.