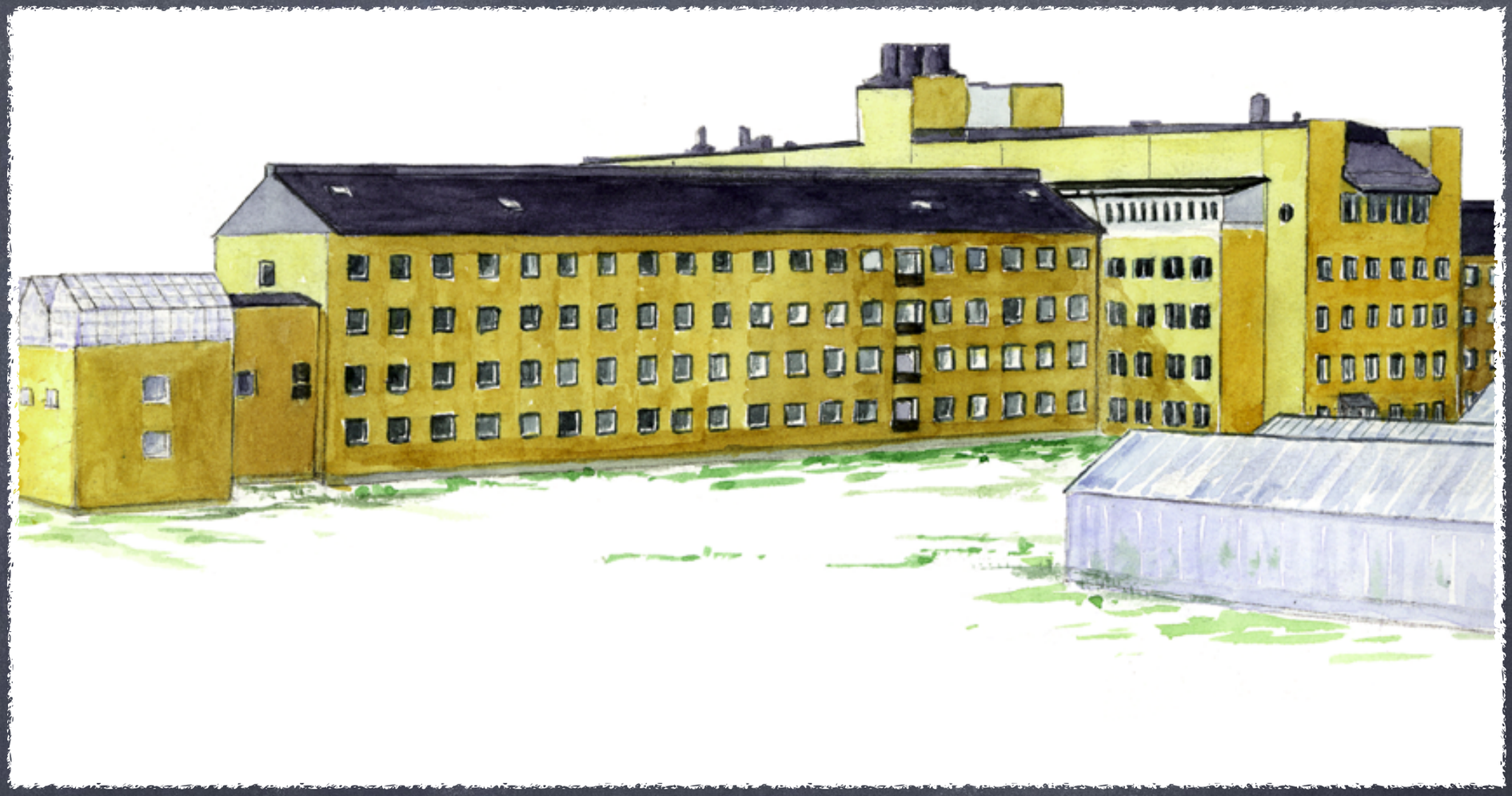
SLU

SCIENCE AND EDUCATION
FOR SUSTAINABLE LIFE

Insights into gene network inferences



Nicolas Delhomme

UMEÅ UNIVERSITY

Umeå Plant Science Centre

Context: High-Throughput Sequencing

NATURE | REVIEW

DNA sequencing at 40: past, present and future

Jay Shendure, **Shankar Balasubramanian**, **George M. Church**, **Walter Gilbert**, **Jane Rogers**, **Jeffery A. Schloss** & **Robert H. Waterston**

Affiliations | **Contributions** | **Corresponding author**

Nature (2017) | doi:10.1038/nature24286

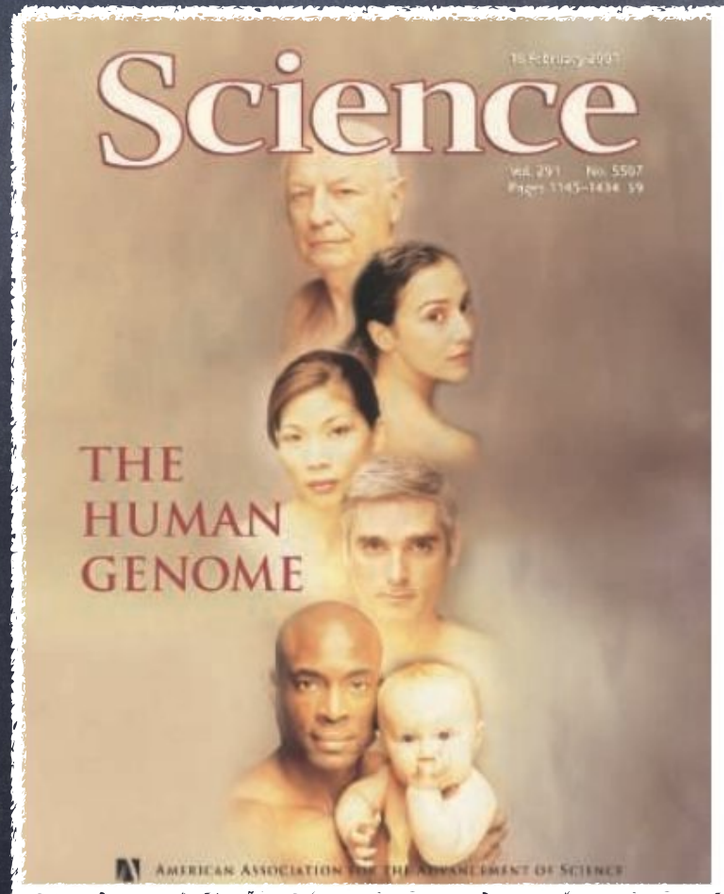
Received 13 July 2017 | Accepted 21 September 2017 | Published online 11 October 2017

http://www.nature.com/nature/journal/vaop/ncurrent/full/nature24286.html

DNA Sequencing:

- Deciphering the blueprint of an organism
- Is not a novel technology

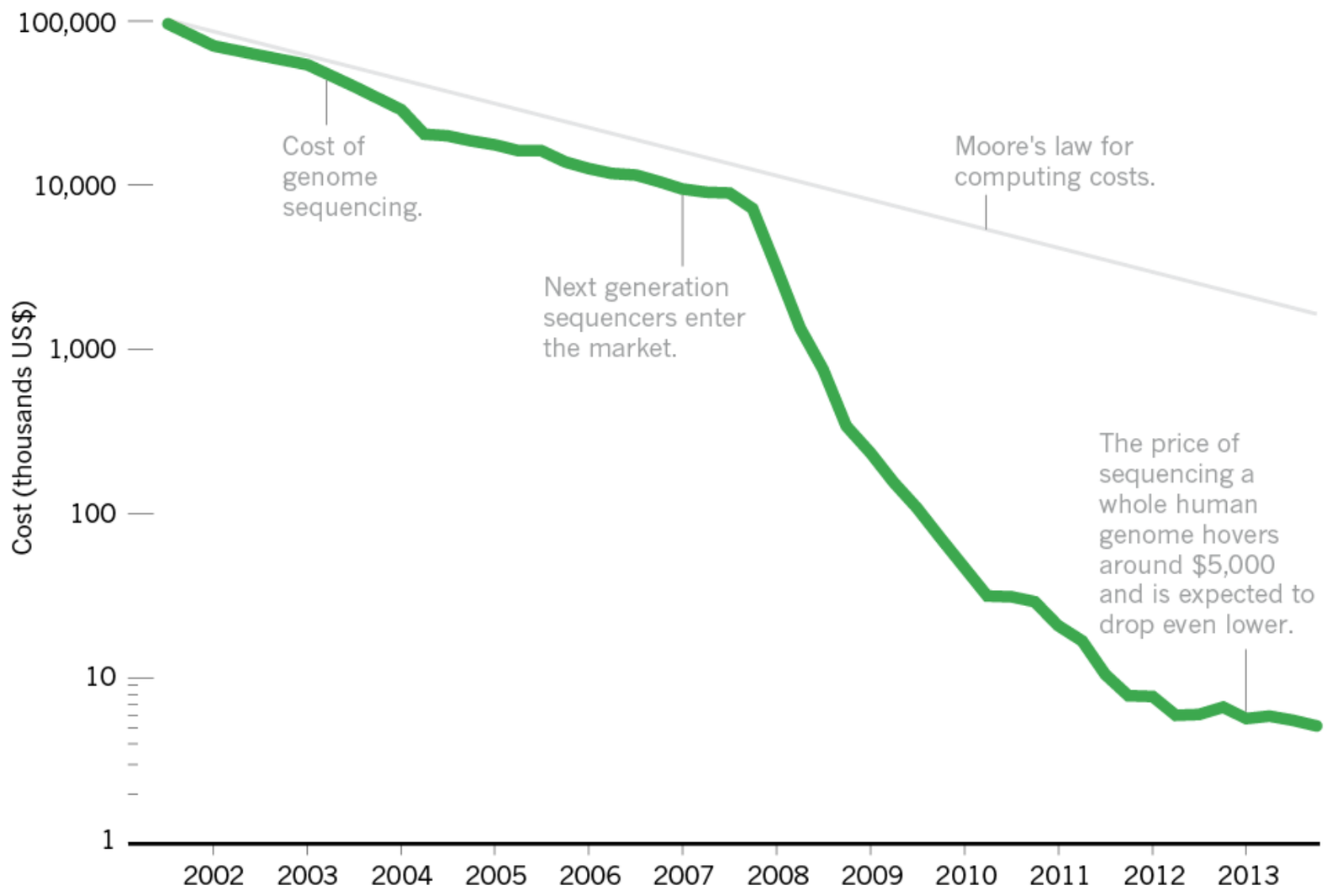
- went ballistic with the Human Genome Project
- and the contribution from Celera, led by Dr. Craig Venter



High-Throughput Sequencing (HTS)

Falling fast

In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore's law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.

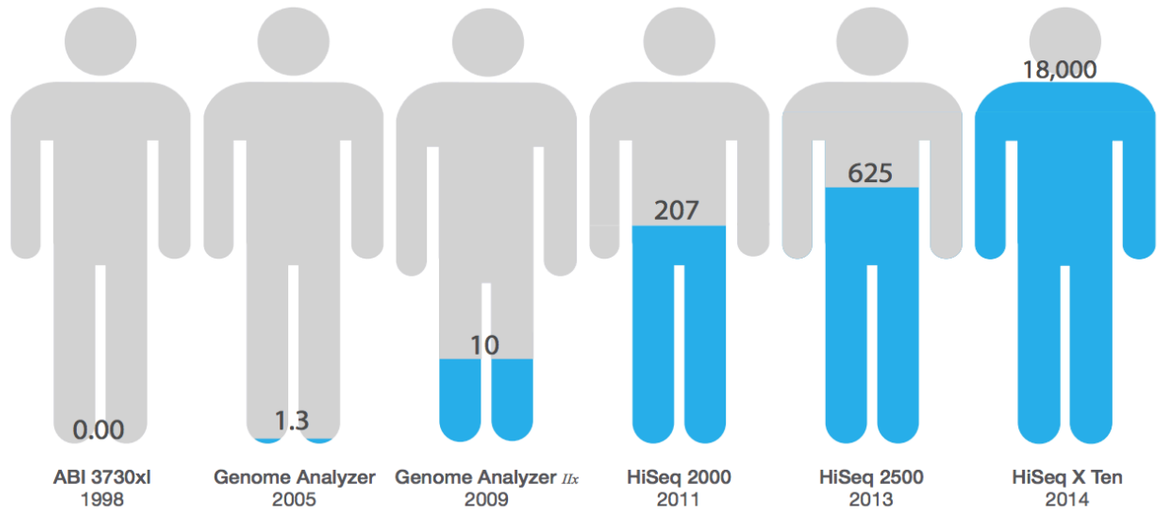


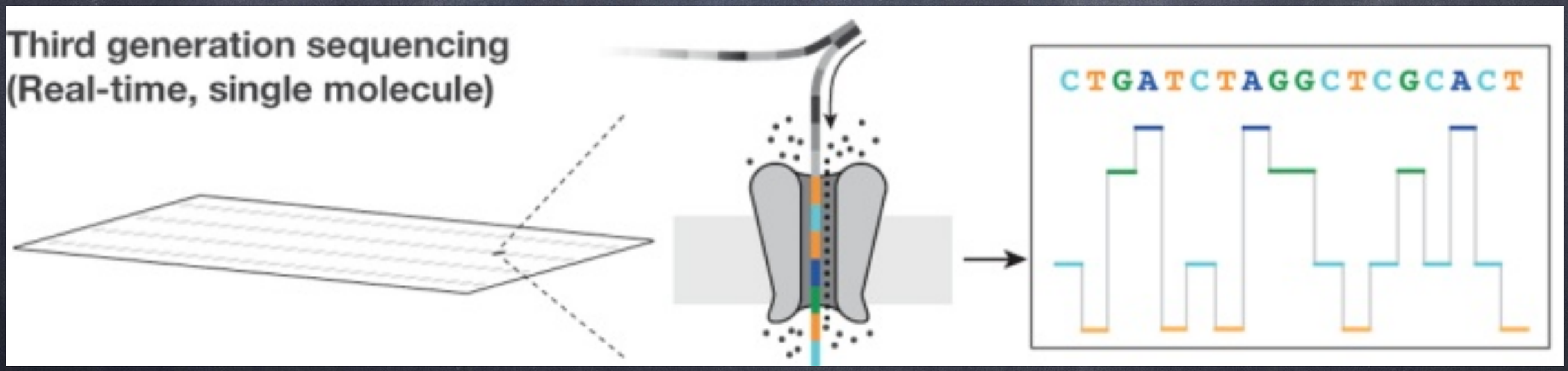
http://www.nature.com/news/technology-the-1-000-genome-1.14901

HTS - What is coming next?

Population power. Extreme throughput. $1,000 human genome.

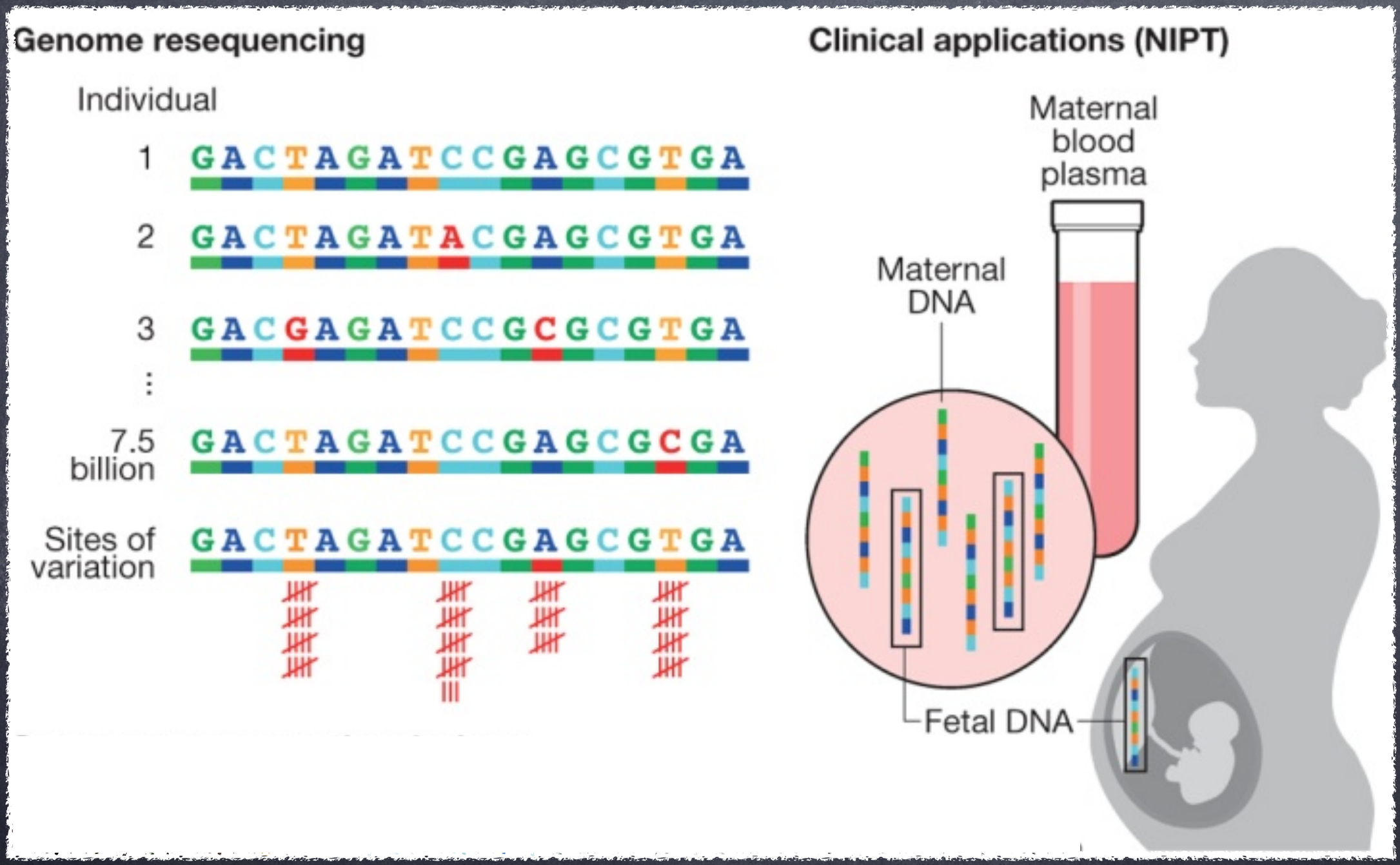
The HiSeq X Ten is a set of ten ultra-high-throughput sequencers, purpose-built for large-scale human whole-genome sequencing.





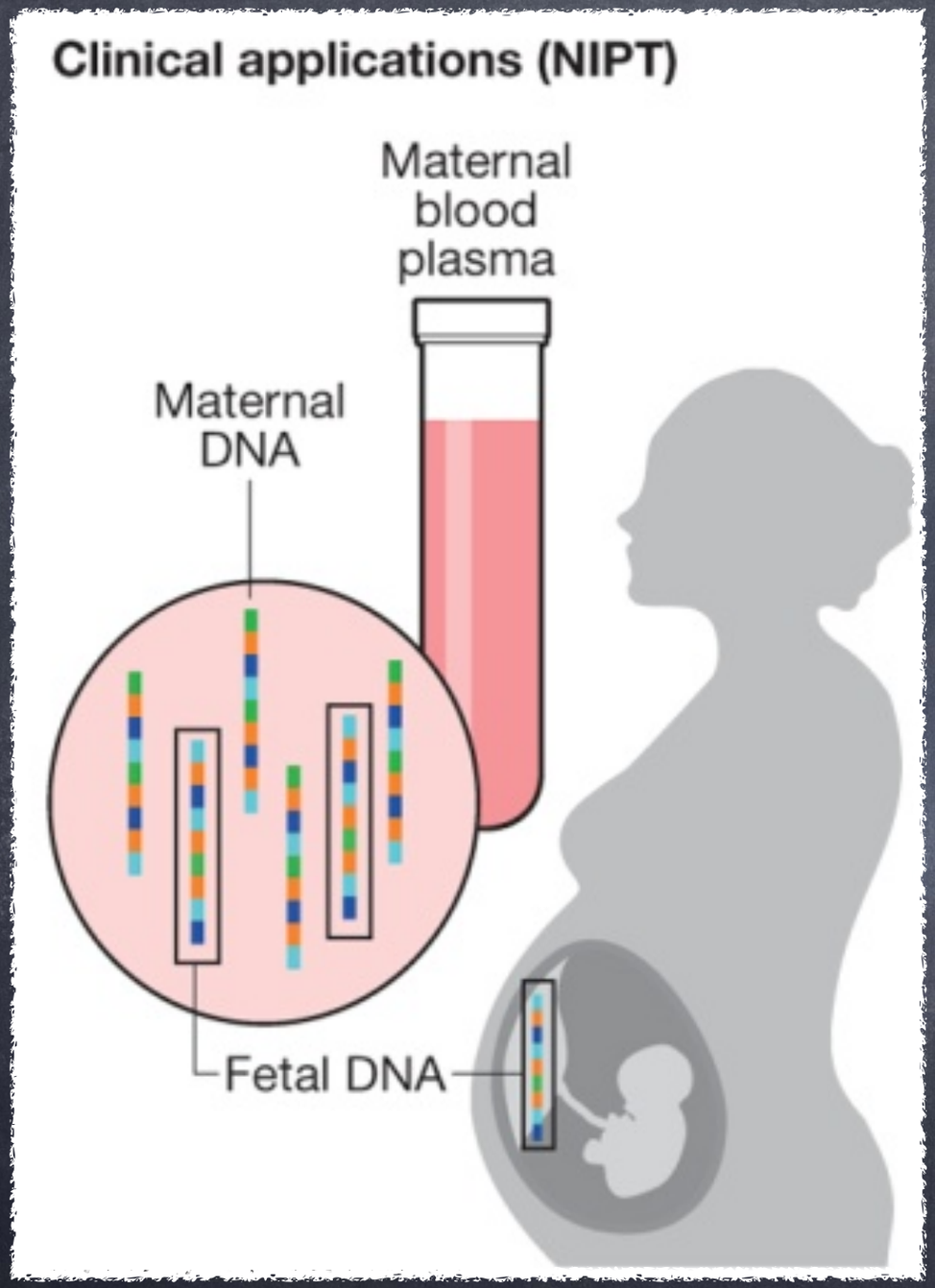
J Shendure et al. Nature 1-9 (2017) doi:10.1038/nature24286

Personalised medicine

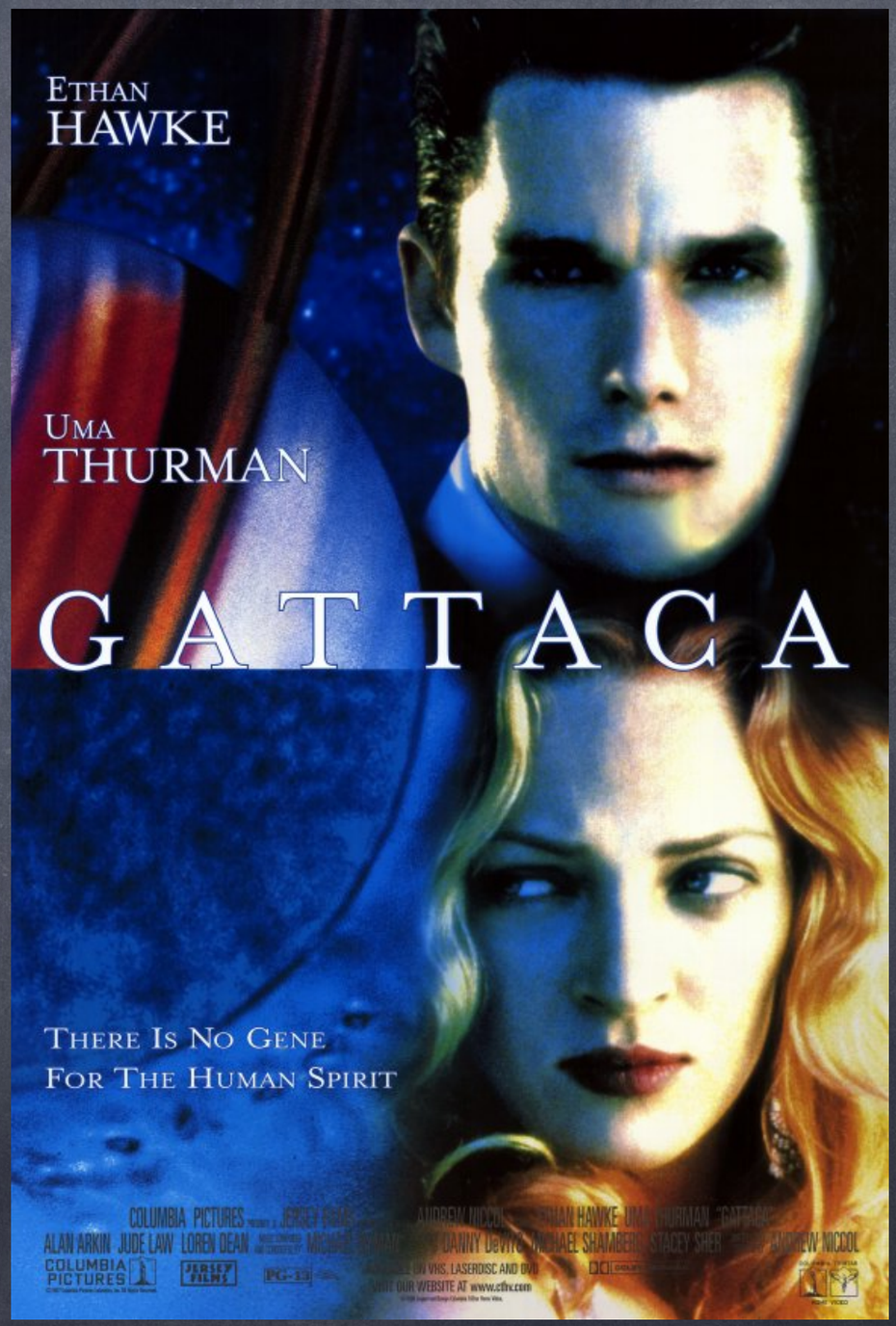


J Shendure et al. Nature 1-9 (2017) doi:10.1038/nature24286

Somewhat spooky

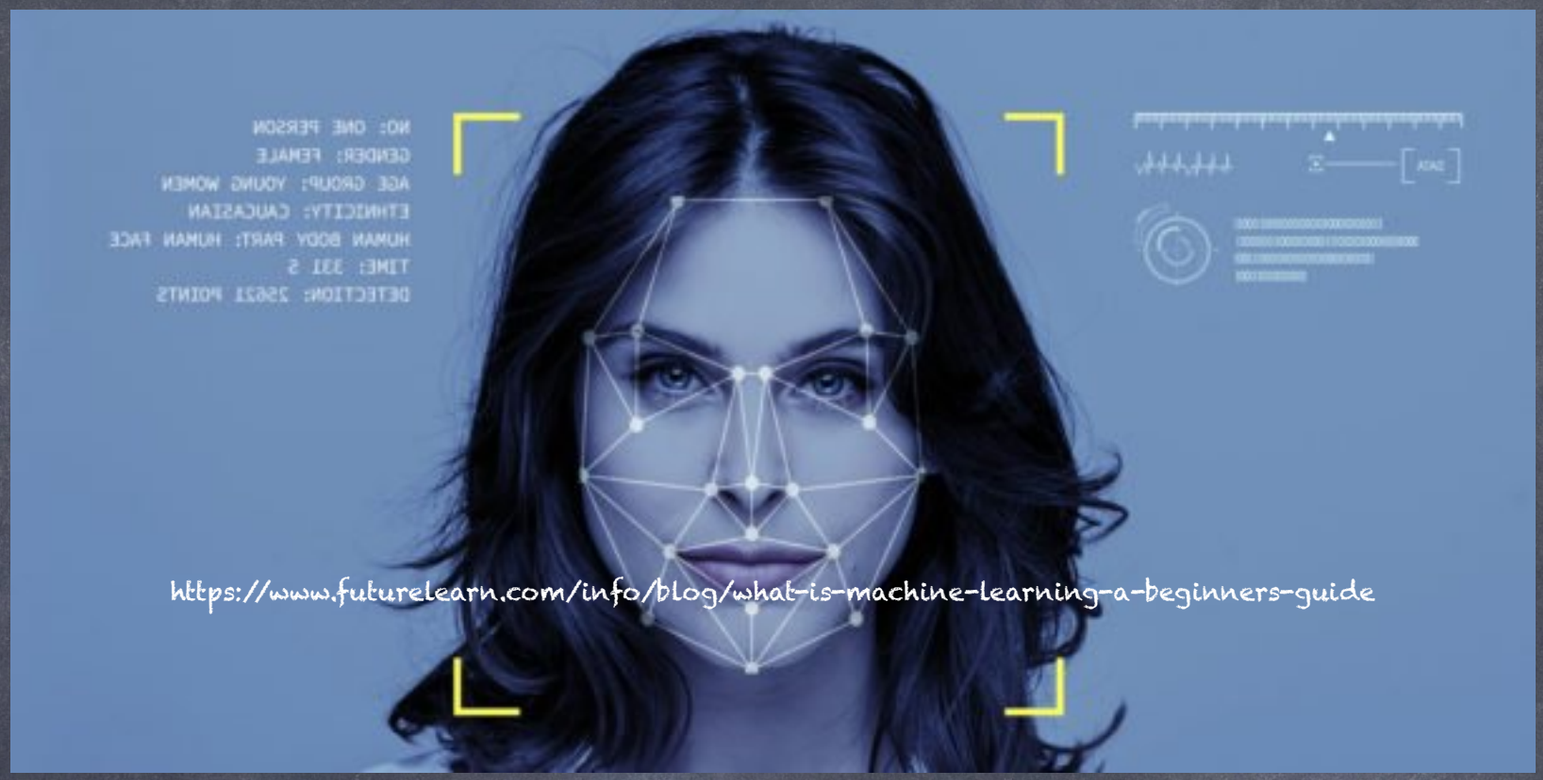


J Shendure et al. Nature 1-9 (2017) doi:10.1038/nature24286



This was "only" ~20 years ago...

HTS - What is coming

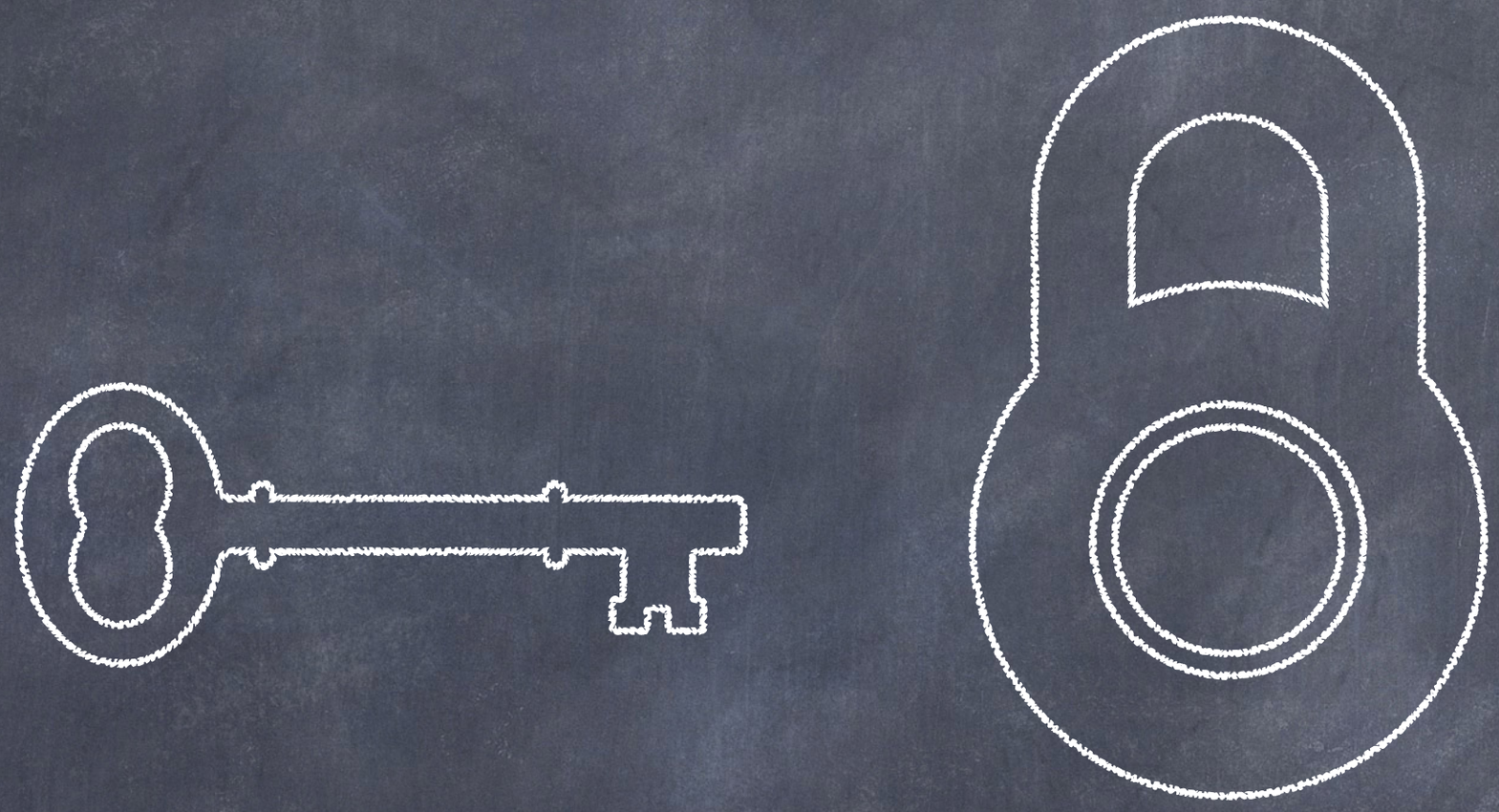


https://www.futurelearn.com/info/blog/what-is-machine-learning-a-beginners-guide

https://www.futurelearn.com/info/blog/what-is-machine-learning-a-beginners-guide

Artificial intelligence, namely machine learning

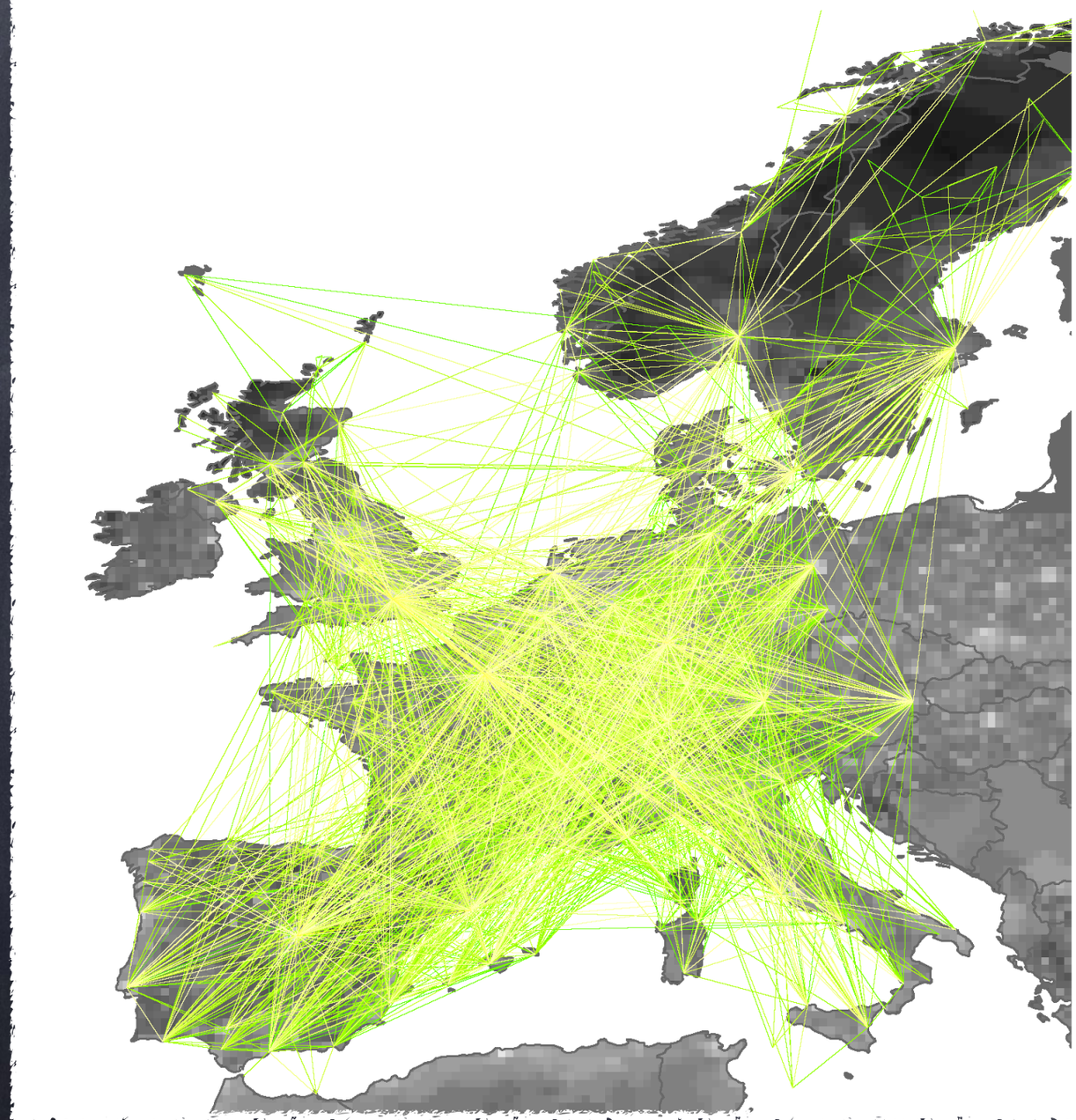
HTS - What is the promise



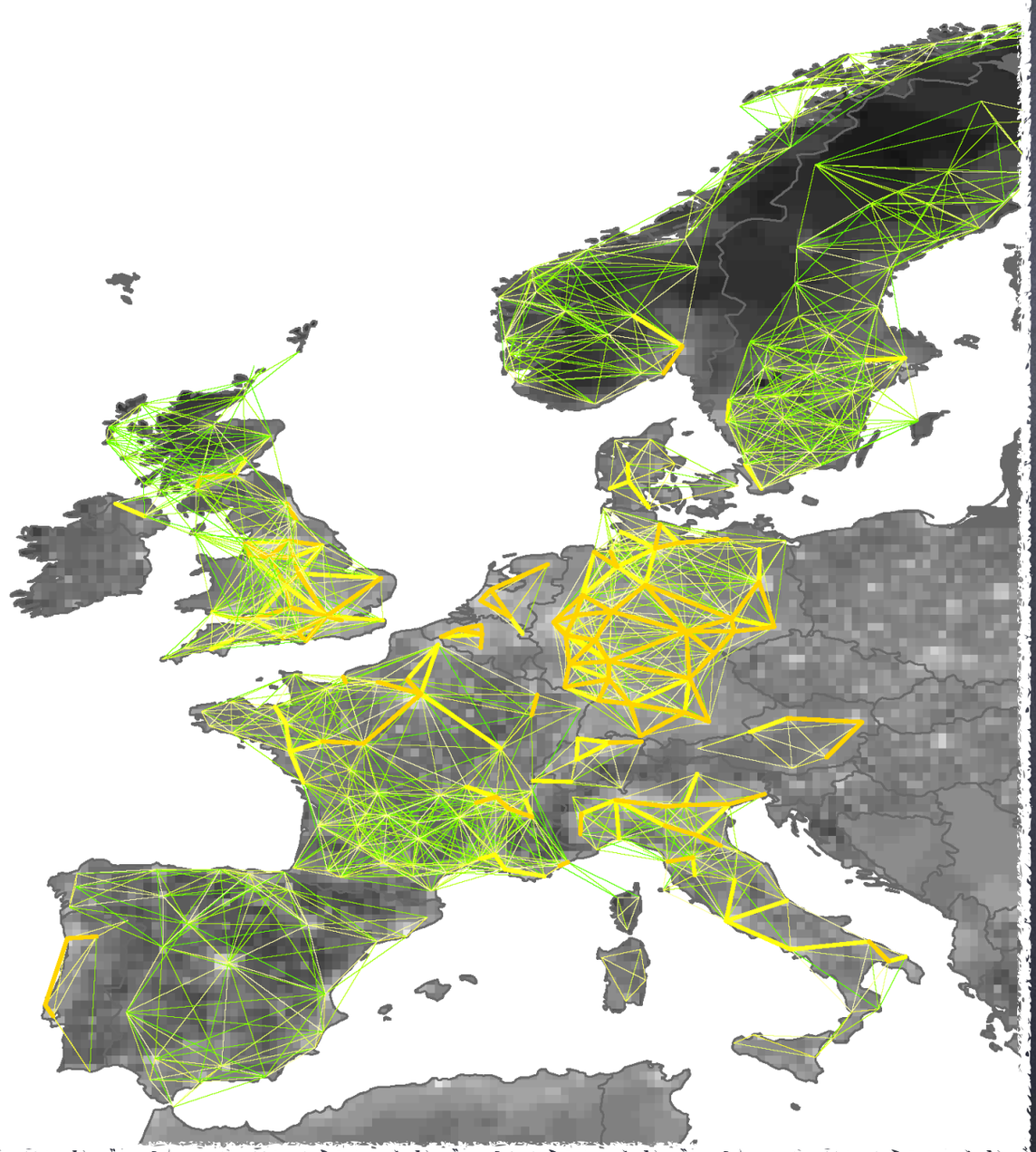
-> unlocking systems biology
by integrating all known
level of gene regulation

Back to the topic: Gene Network Inference

http://www.gleamviz.org

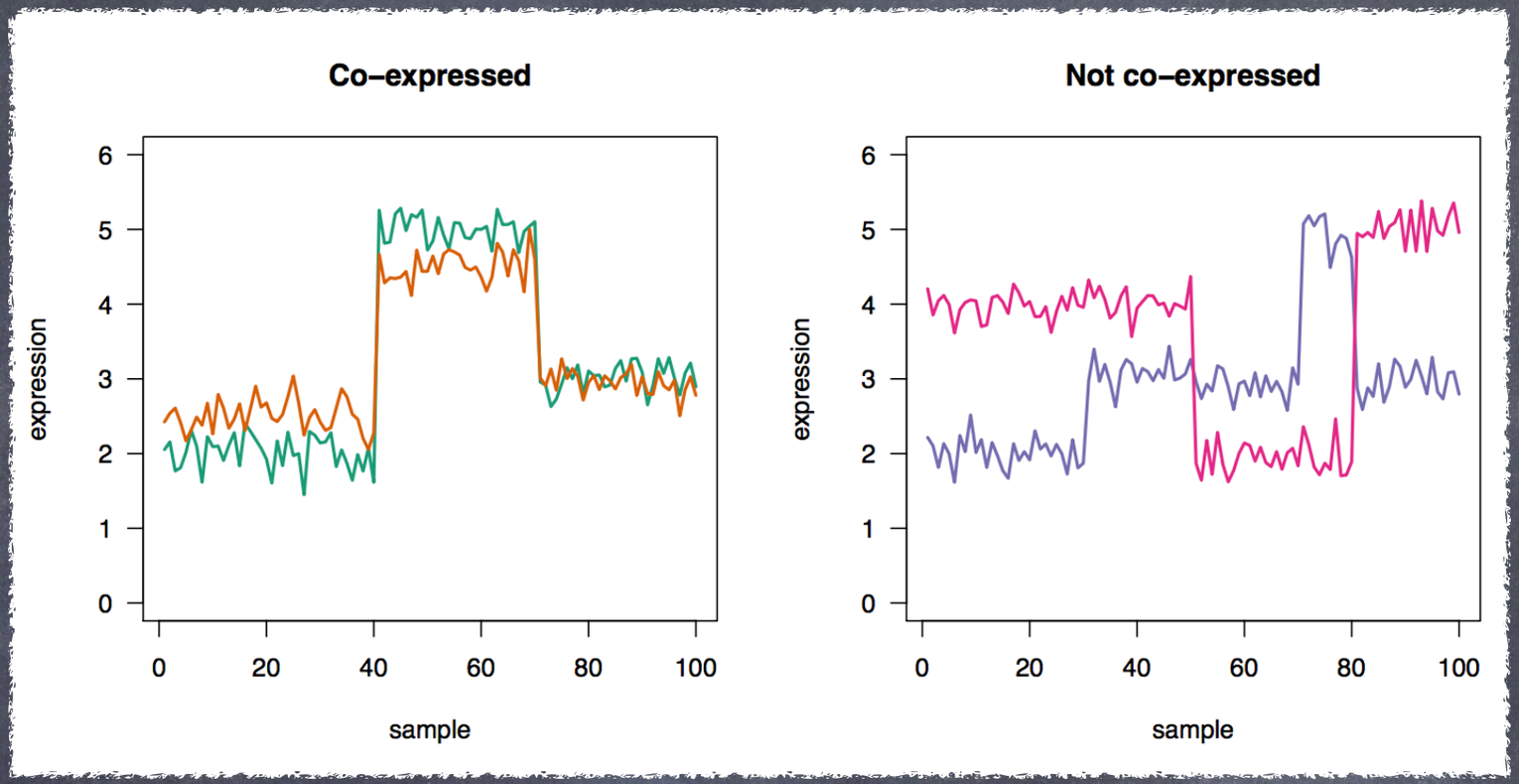


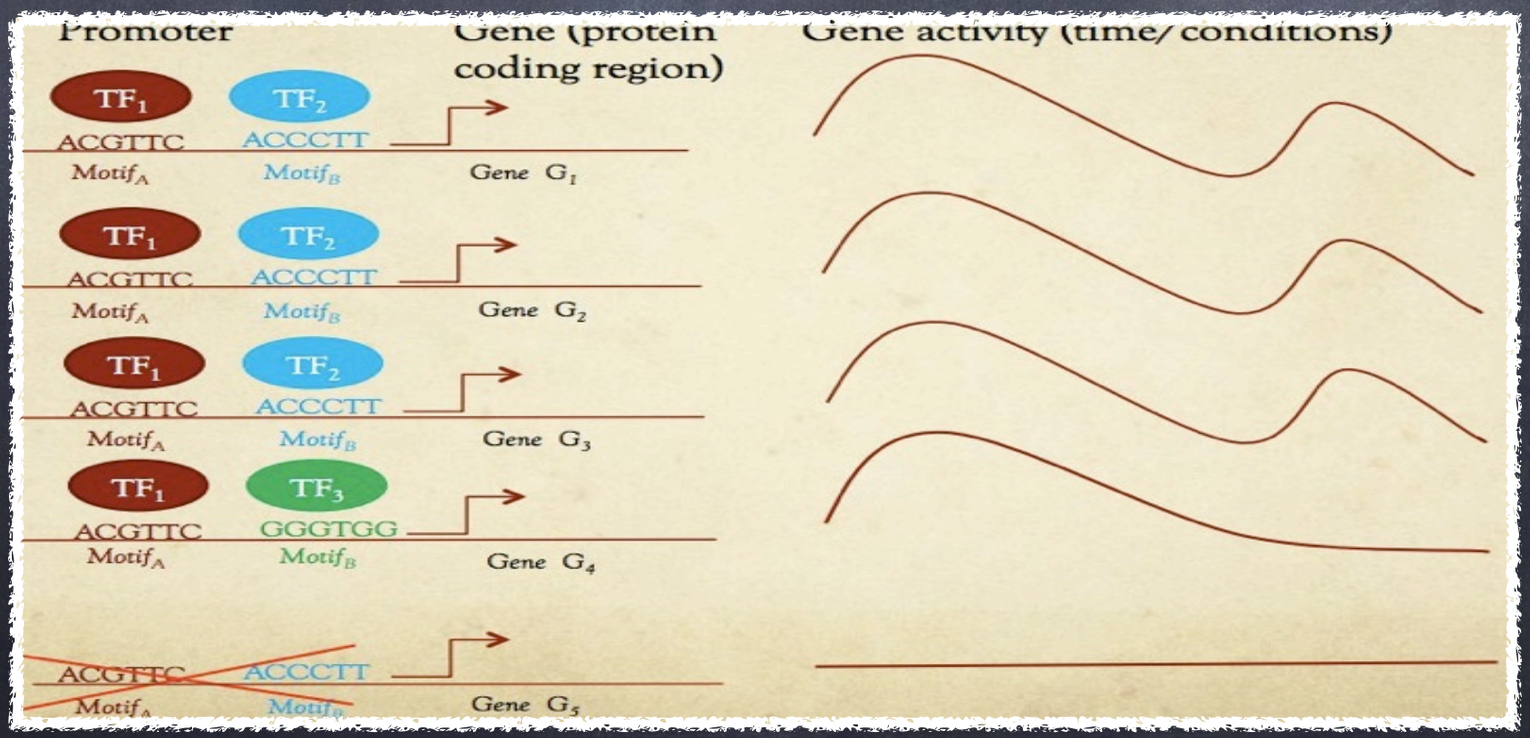
Airport



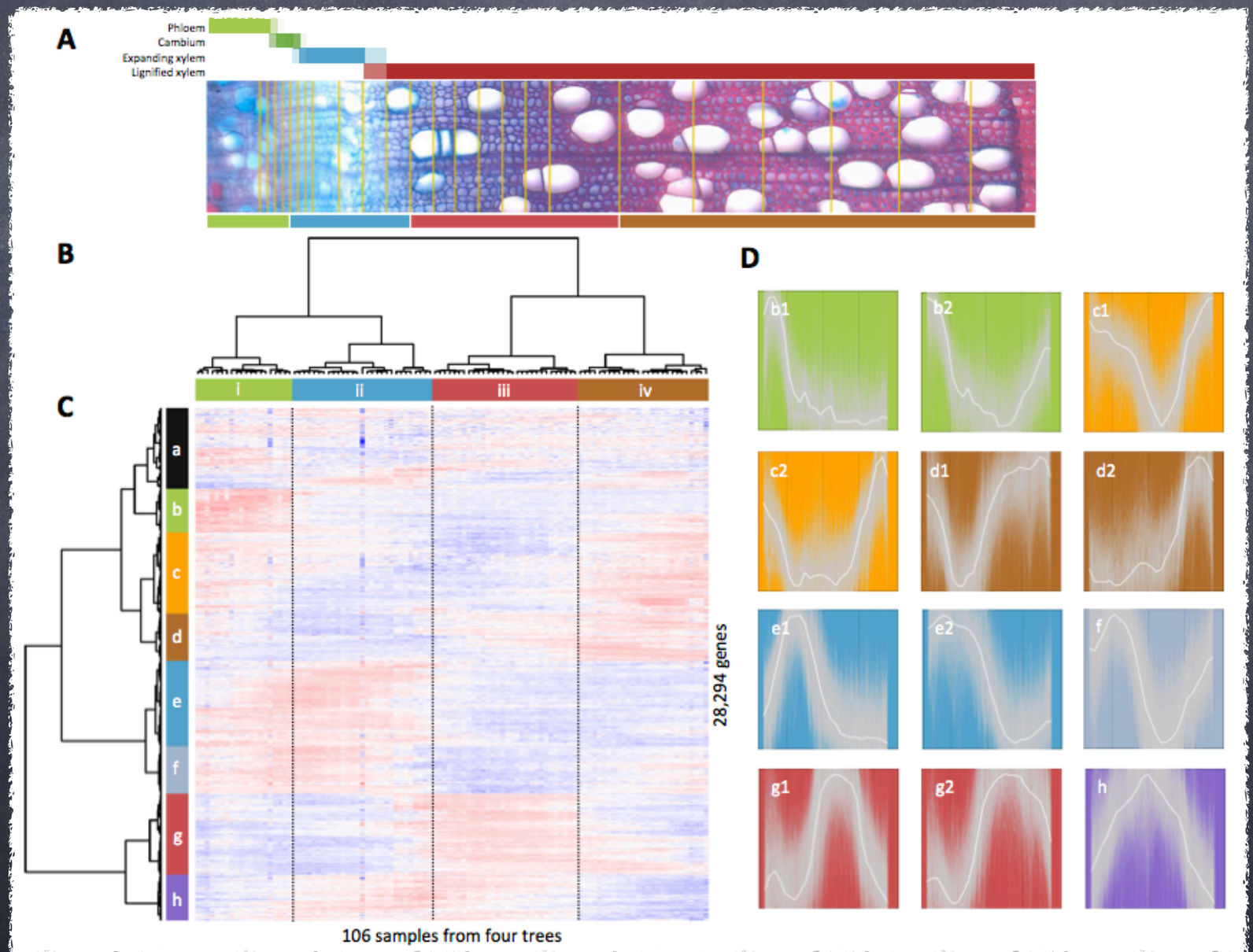
Commute

Gene regulation



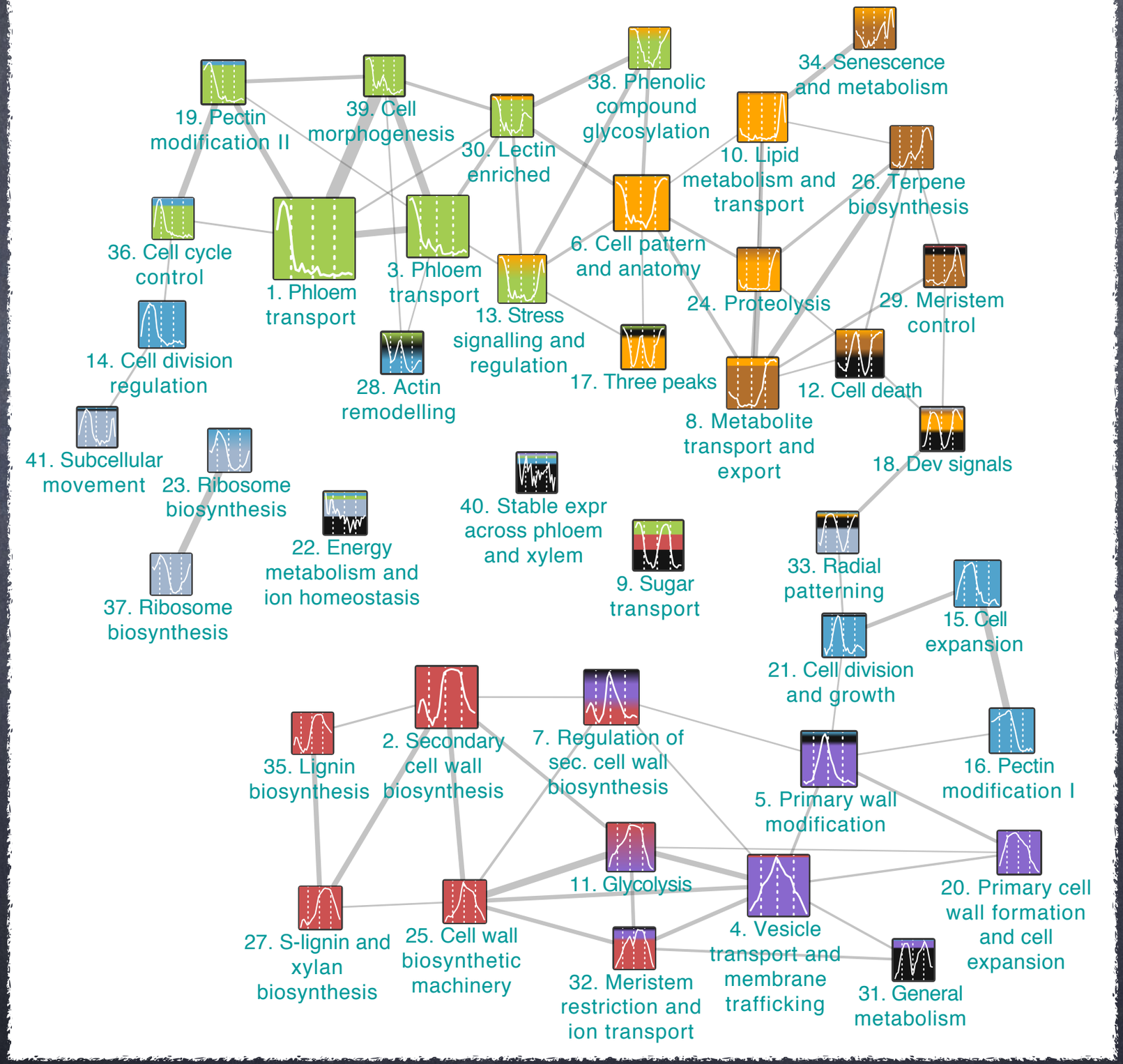


Towards a more sustainable forestry industry - a case study



Sundell et al., Plant Cell, 2017

For us: a lecture in wood formation



Sundell et al., Plant Cell, 2017

Why networks?

- Data lends itself to network analysis (time series, developmental, large numbers of samples)
- Complementary to DE
- Centrality metrics can provide other insights
- Gene clusters with associations

A word on preprocessing

- Data is expected to be homoscedastic -> VST / RLD
- Low variance genes should be filtered
- Filter by transformed expression
- Alternatively filter by P-value (LRT) or through independent filtering (DESeq2)

Likelihood test ratio

- https://hbctraining.github.io/DGE_workshop/lessons/08_DGE_LRT.html

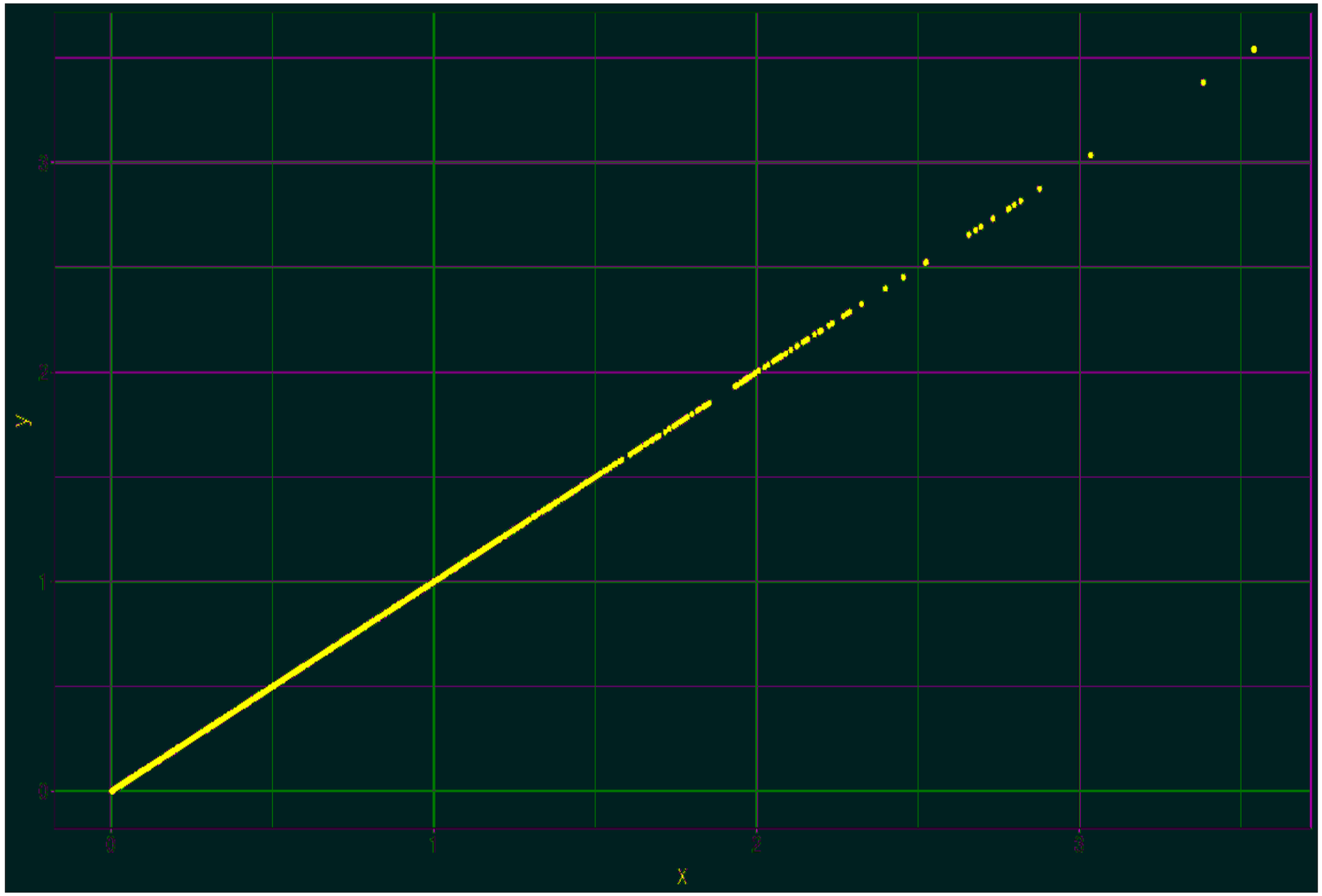
How are networks inferred?

- Vast number of options:
  - Correlation (Pearson, Spearman)
  - Mutual Information and extensions (CLR, ARACNE)
  - Regression (TIGRESS, GENIE3)
  - Other (PLSNET, ANOVERENCE)

Pearson correlation

Linear relationship (parametric)

x and *y* behave the same in a linear manner

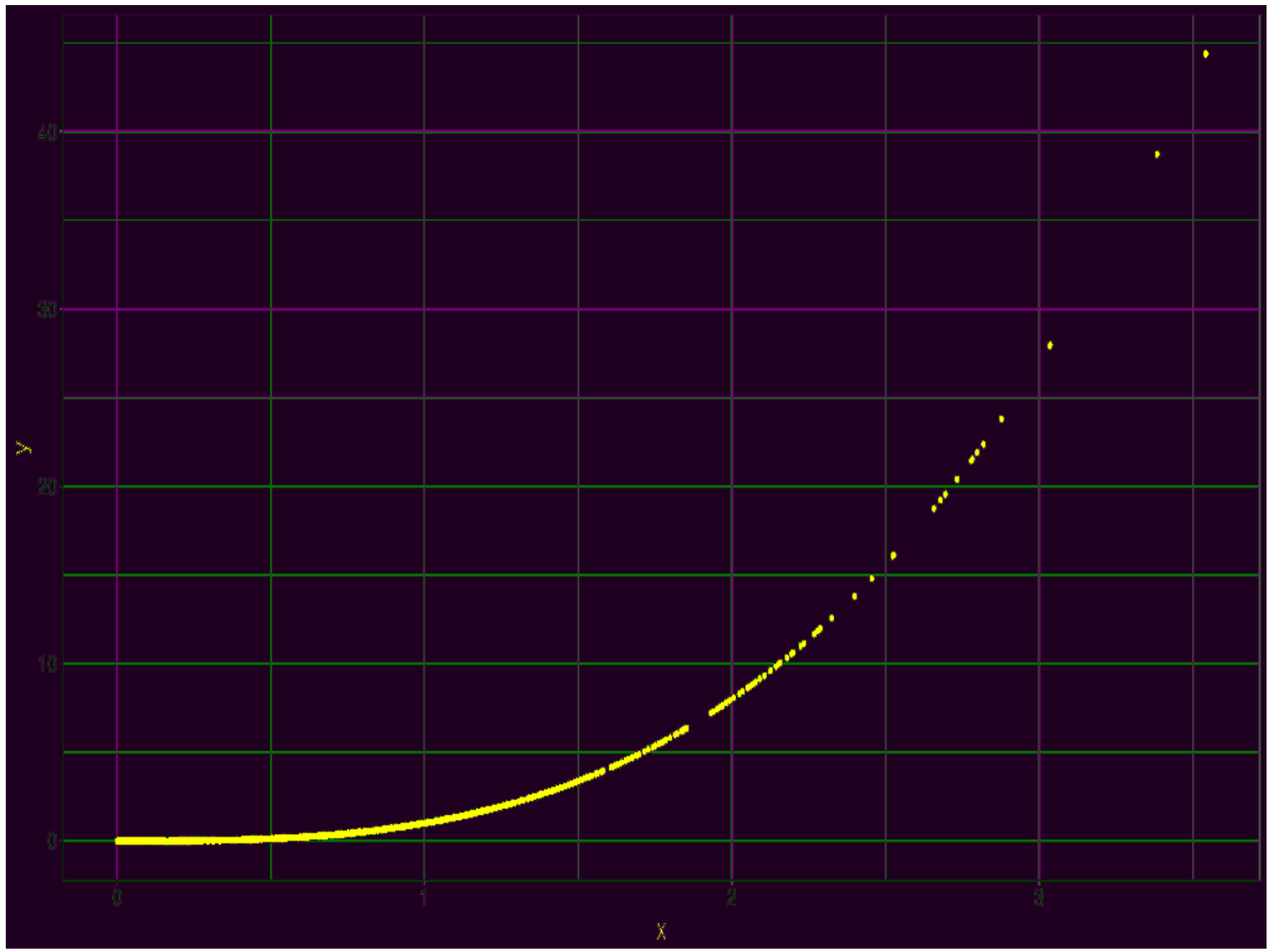


Spearman correlation

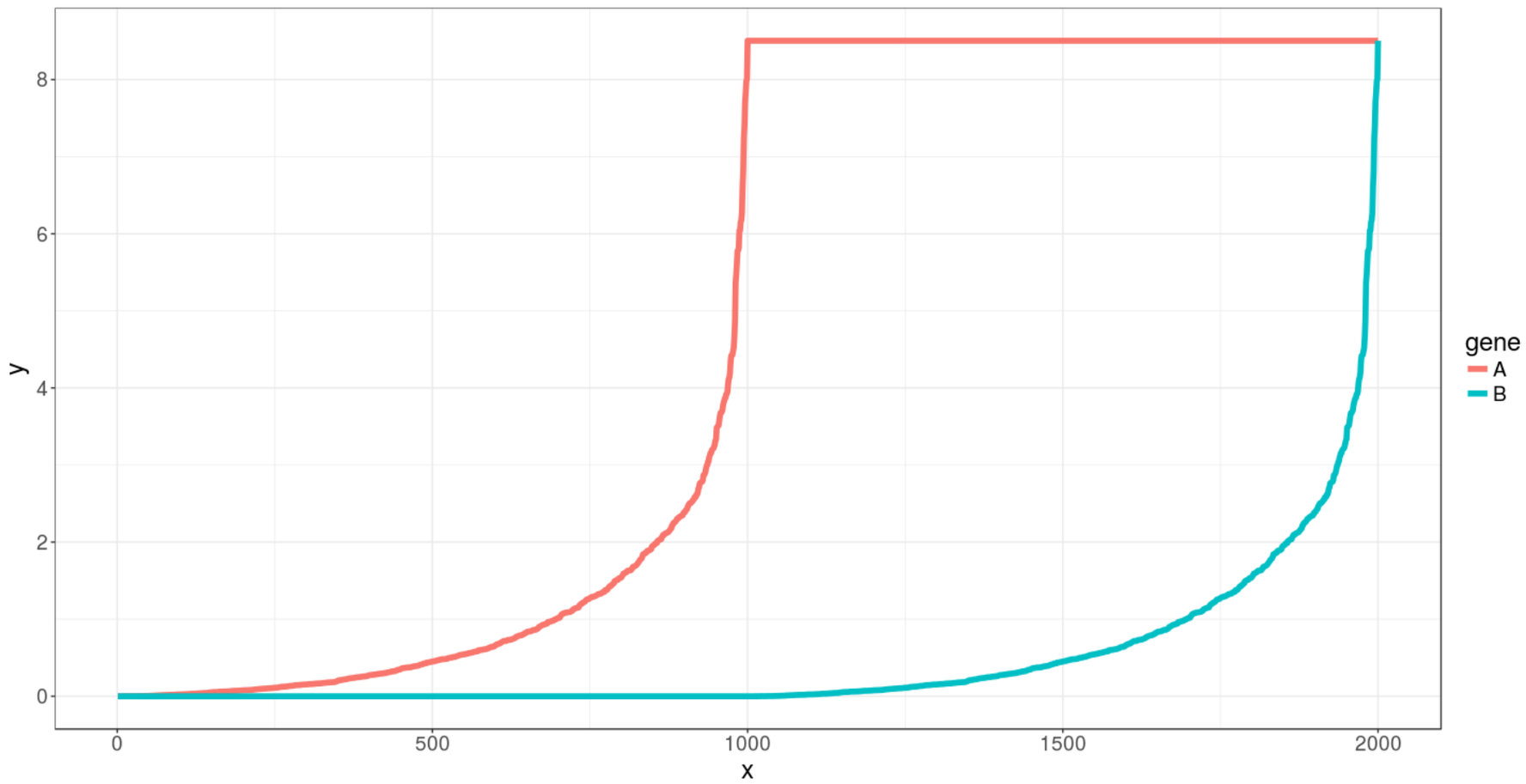
As Pearson, but computed on ranks

Also detects nonlinear relationships

x and *y* behave the same in a linear or nonlinear manner



What about these two genes

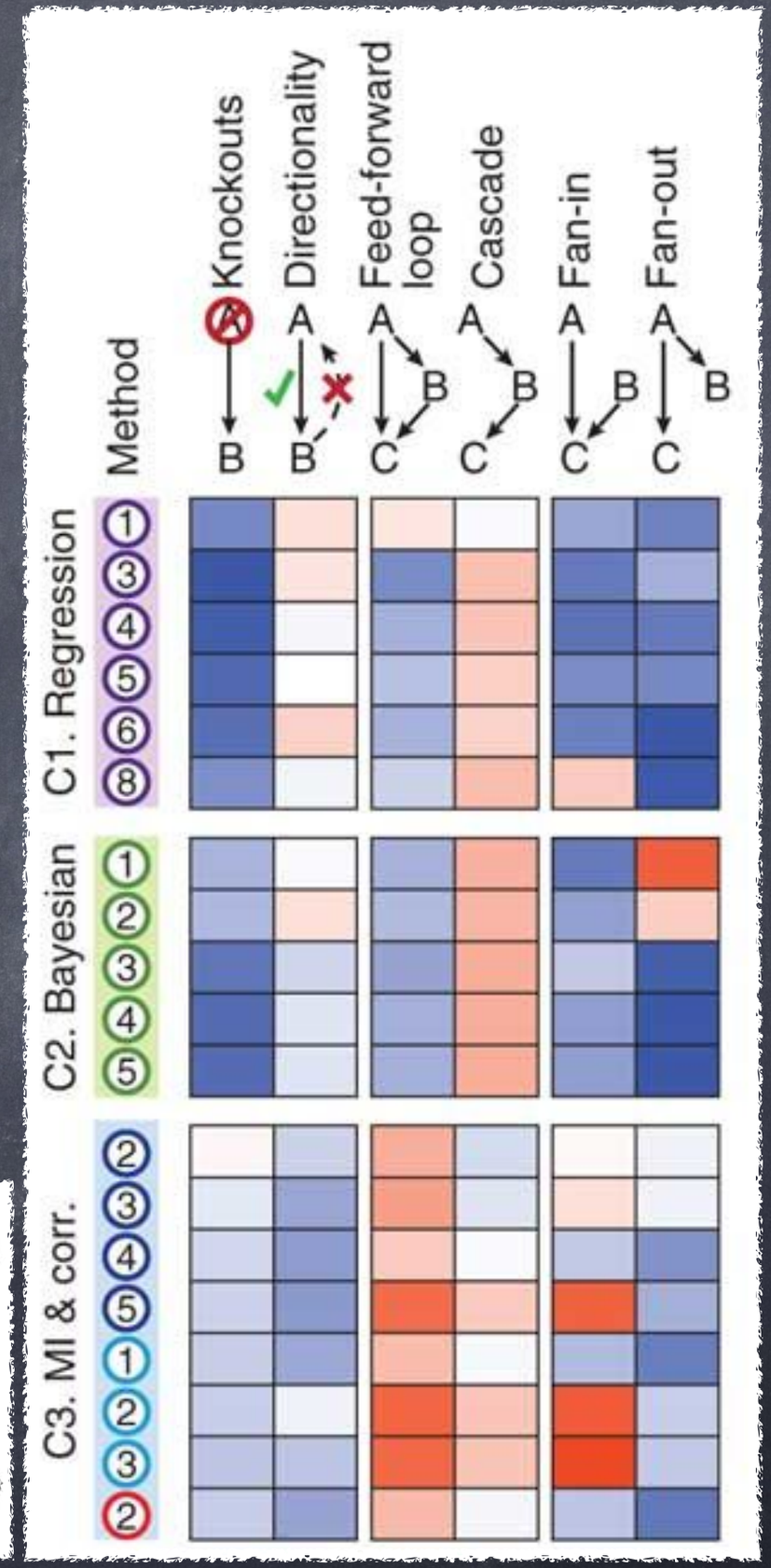


Complex interactions are harder to detect

Network inferences

- Not recent: over 30 methods reviewed in 2012
- Methods identify interactions preferentially
- Methods using the same principles identify the same interactions

=> Crowd network



Marbach et al., Nature Methods, 2012

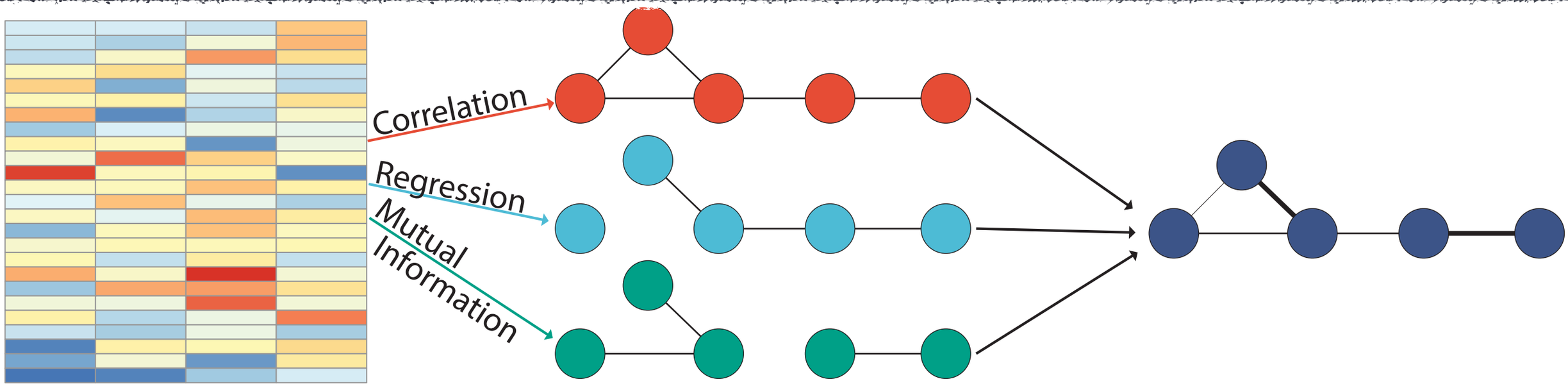
Seidr: Crowd network inference

- Seidr is a software toolkit to infer crowd networks of gene-gene relationships in expression data.
- Infer network using published algorithms
- Compute ranks of all inferred networks
- Aggregate ranked representation networks into a crowd-network

Seidr makes life easier-ish

ANOVER-ENCE	Küffner et al., 2012	ANOVA	C++	C++	No	No
ARACNE	Margolin et al., 2006	MI + DPI	C++	C++	Yes	Yes
CLR	Faith et al., 2007; Daub et al., 2004	MI + CLR	MATLAB / C / C++	C++	No	Yes
Elastic Net ensemble	Ruyssinck et al., 2014	Elastic Net Regression	R (glmnet)	C++ (glmnet)	No	Yes
GENIE3	Huynh-Thu et al., 2010	Random Forest Regression	R (randomForest)	C++ (ranger)	No	Yes
NARROMI	Zhang et al., 2013	MI + Linear Programming	MATLAB	C++ (glpk)	No	Yes
Partial Correlation	Schäfer and Strimmer, 2005	Correlation	R	C++	No	No
Pearson Correlation	NA	Correlation	NA	C++	No	No
PLSNET	Guo et al., 2016	PLS	MATLAB	C++	No	Yes
Spearman Correlation	NA	Correlation	NA	C++	No	No
SVM ensemble	Ruyssinck et al., 2014	SVM regression	R (libsvm) / C	C++ (libsvm, liblinear)	No	Yes
TIGRESS	Haury et al., 2012	LASSO Regression	MATLAB / R	C++ (glmnet)	No	Yes

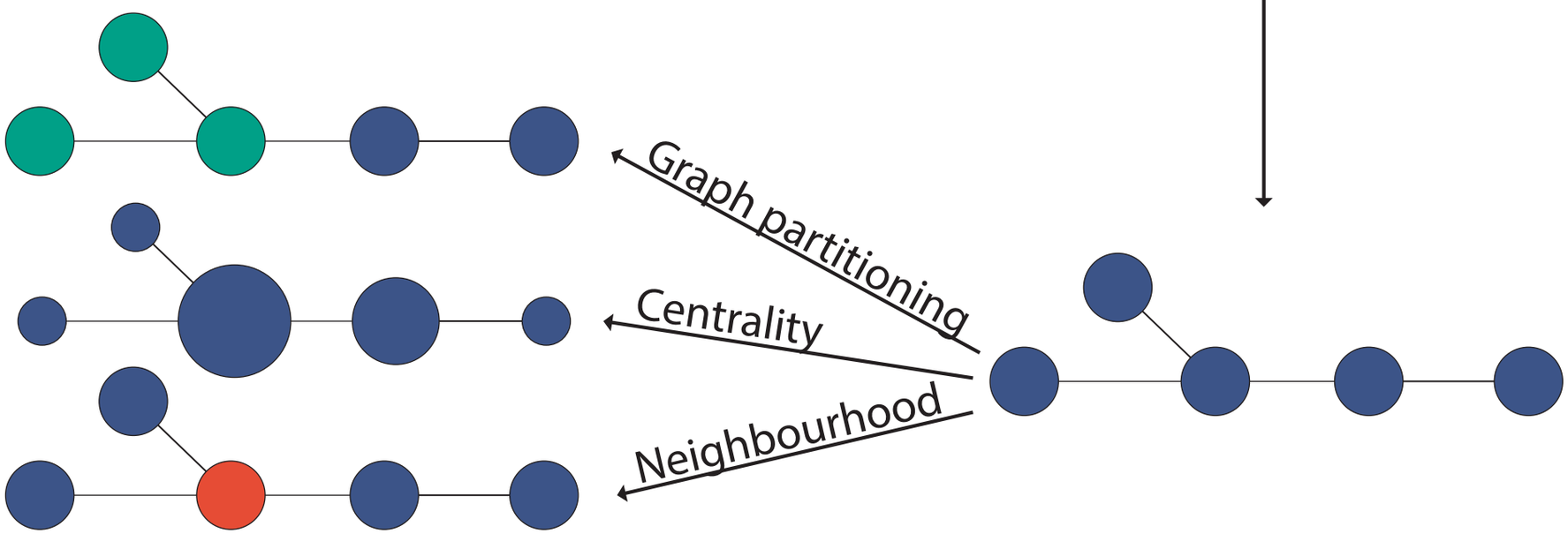
Seidr flow



Preprocessed expression data

Create subnetworks

Rank and aggregate scores

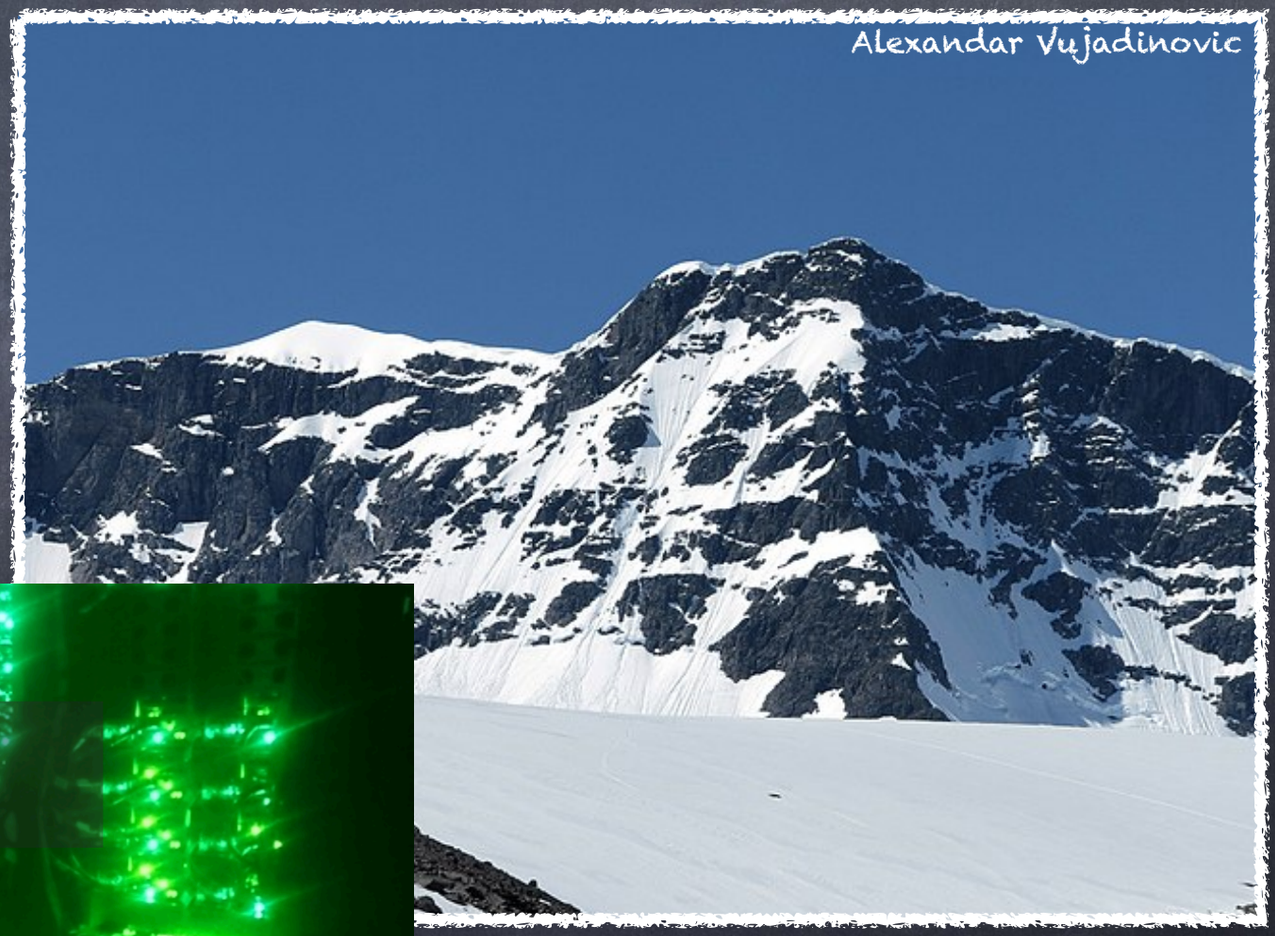


Example downstream analyses

Prune low-scoring edges

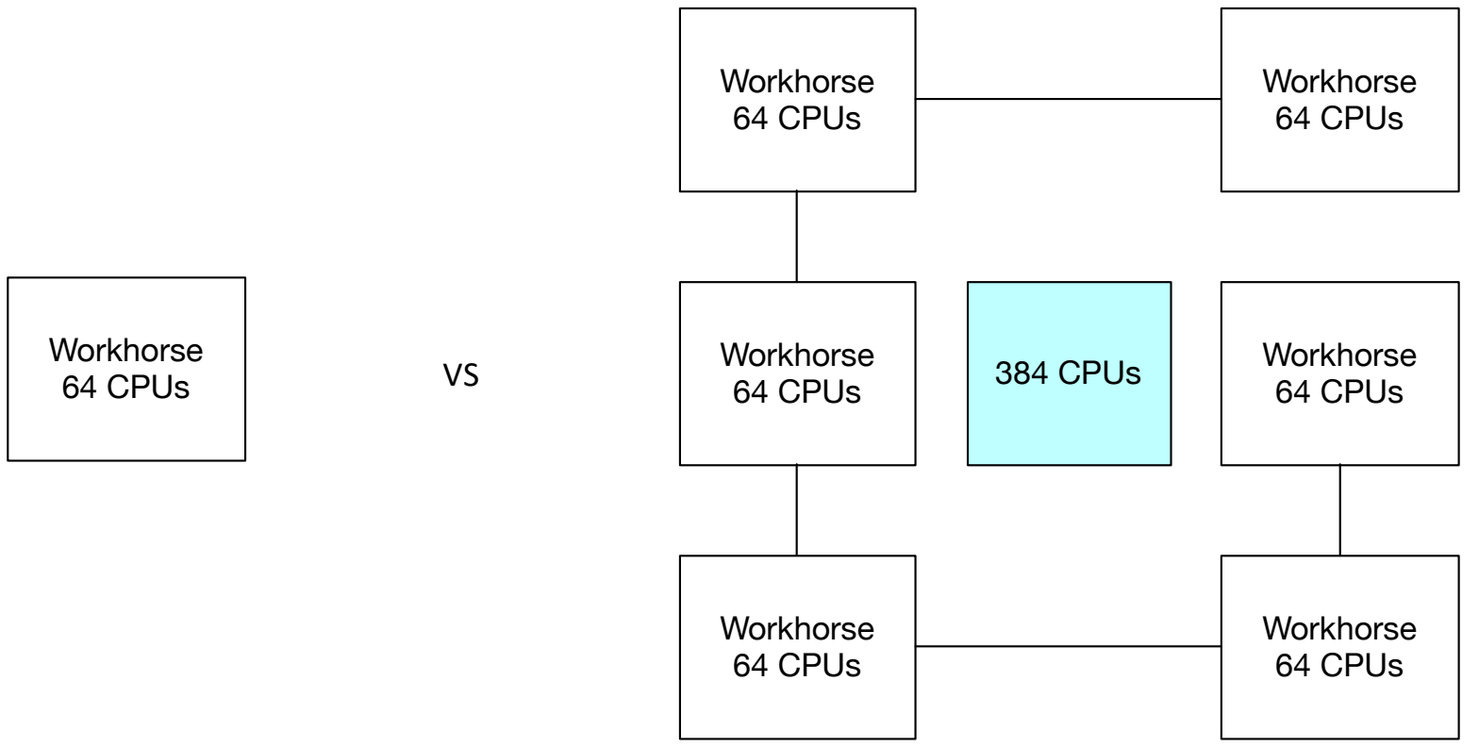
82 years and 70 days

- Not "just" wood in poplar
  - 3,000 samples for 3 species: Arabidopsis thaliana, Picea abies and Populus tremula



- It would still take 55 hours on:

OpenMPI enables cluster computing



Reducing noise

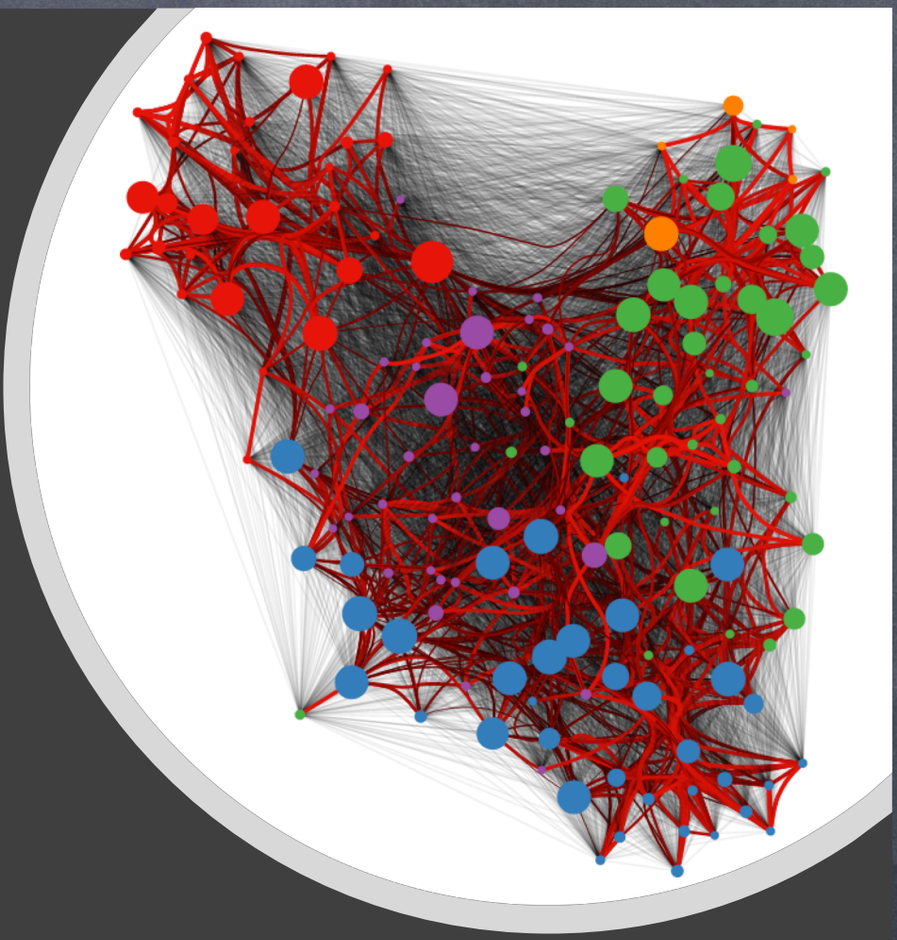
Community networks will be fully connected

Most (99%) of edges are probably noise

How do we determine which to cut

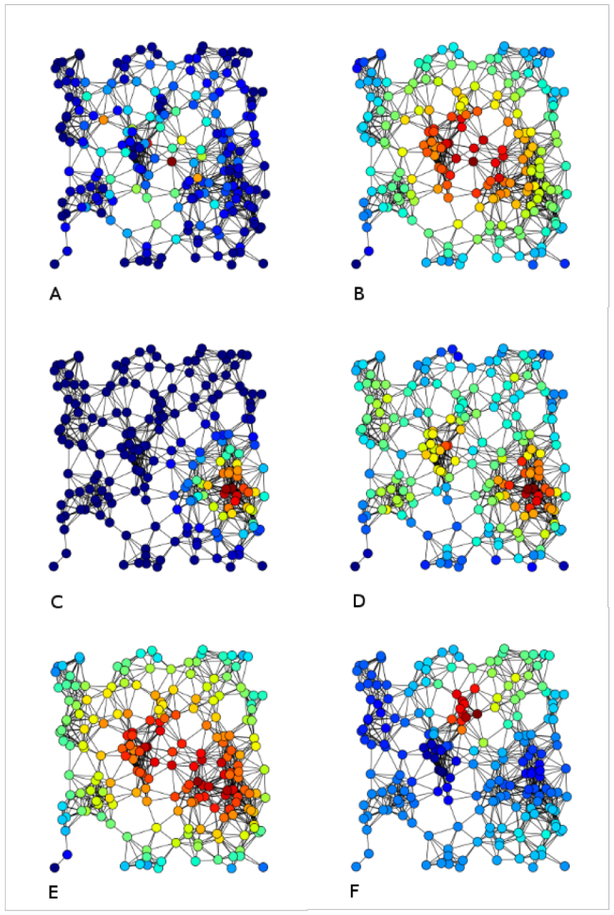
Backboning (Coscia et al.: Network Backboning with Noisy Data): dynamic, statistical method

Hard threshold: easy to implement, produces still valid results



Network centrality

A) **Betweenness** – Nodes that are placed in between others. Control flow
B) **Closeness** – Nodes that are placed most central. Can act quickly on other nodes
C) **Eigenvector** – Nodes that are connected to other important nodes and can influence the entire network
D) **Degree** – Nodes that have the most direct connections
E) **Harmonic** – Variant of closeness defined for disconnected graphs
F) **Katz** – Variant of eigenvector centrality as a measure of influence



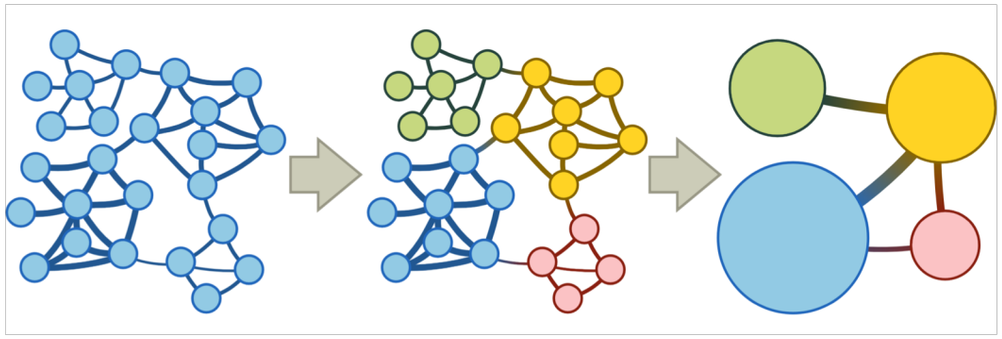
https://en.wikipedia.org/wiki/Katz_centrality

Graph partitioning

Objective is to divide the graph into meaningful clusters

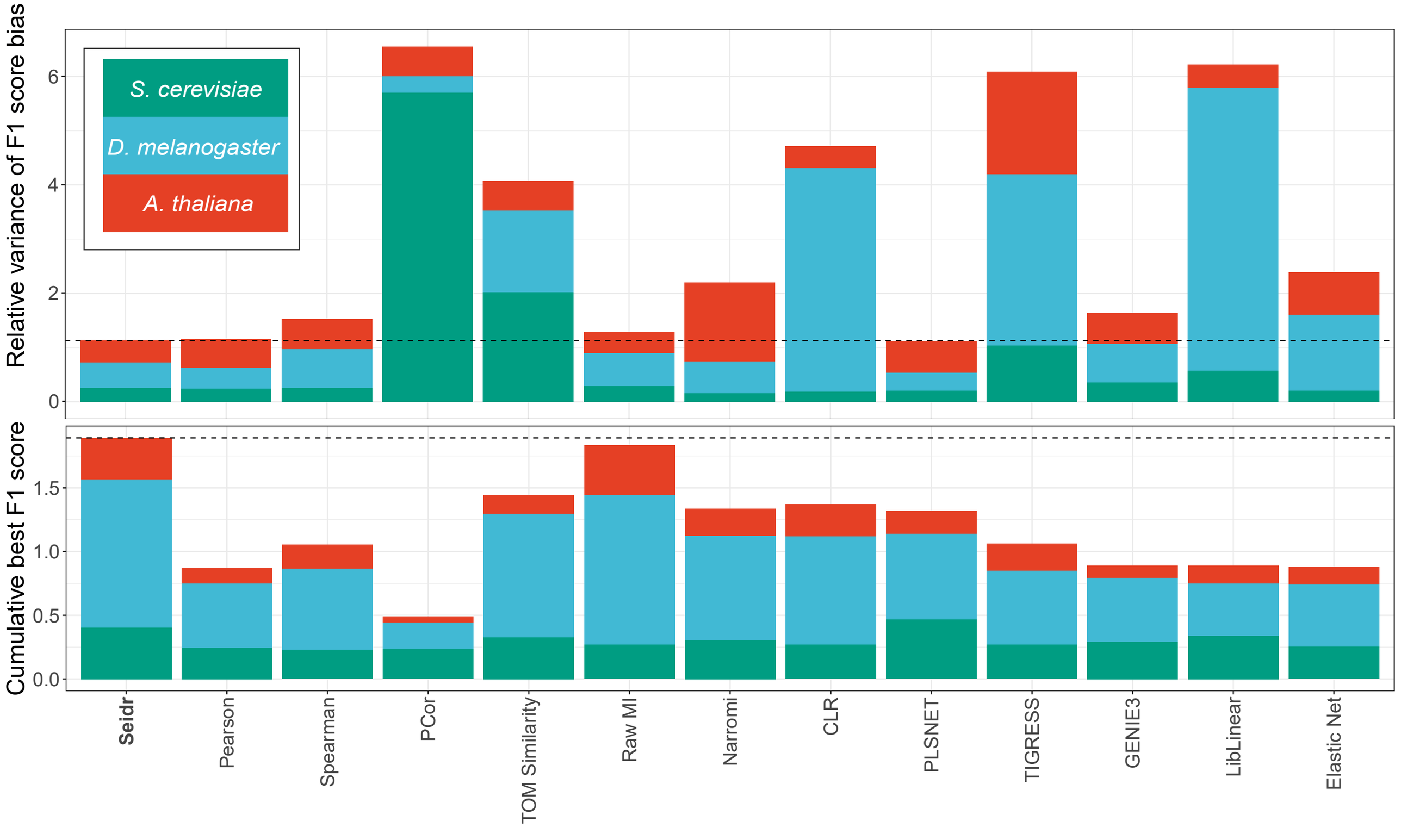
Only topology as input, not underlying data

We use InfoMap, which partitions via random walks

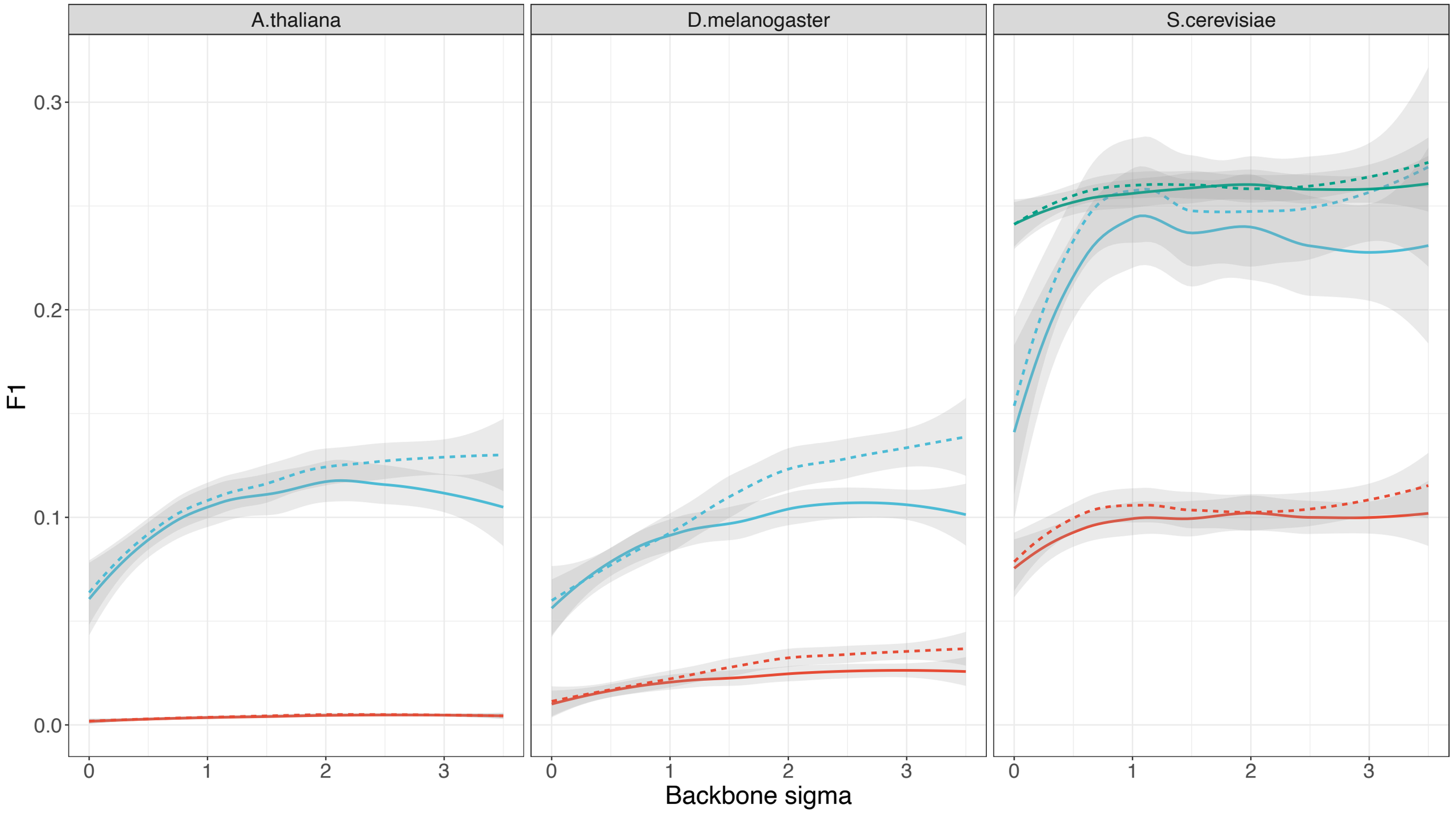


Rosvall, M., & Bergstrom, C. . (2008). Maps of Random Walks on Complex Network Reveal Community Structure. In *Proceedings of the National Academy of Sciences,* (p. 105(4), 1118-1123.).

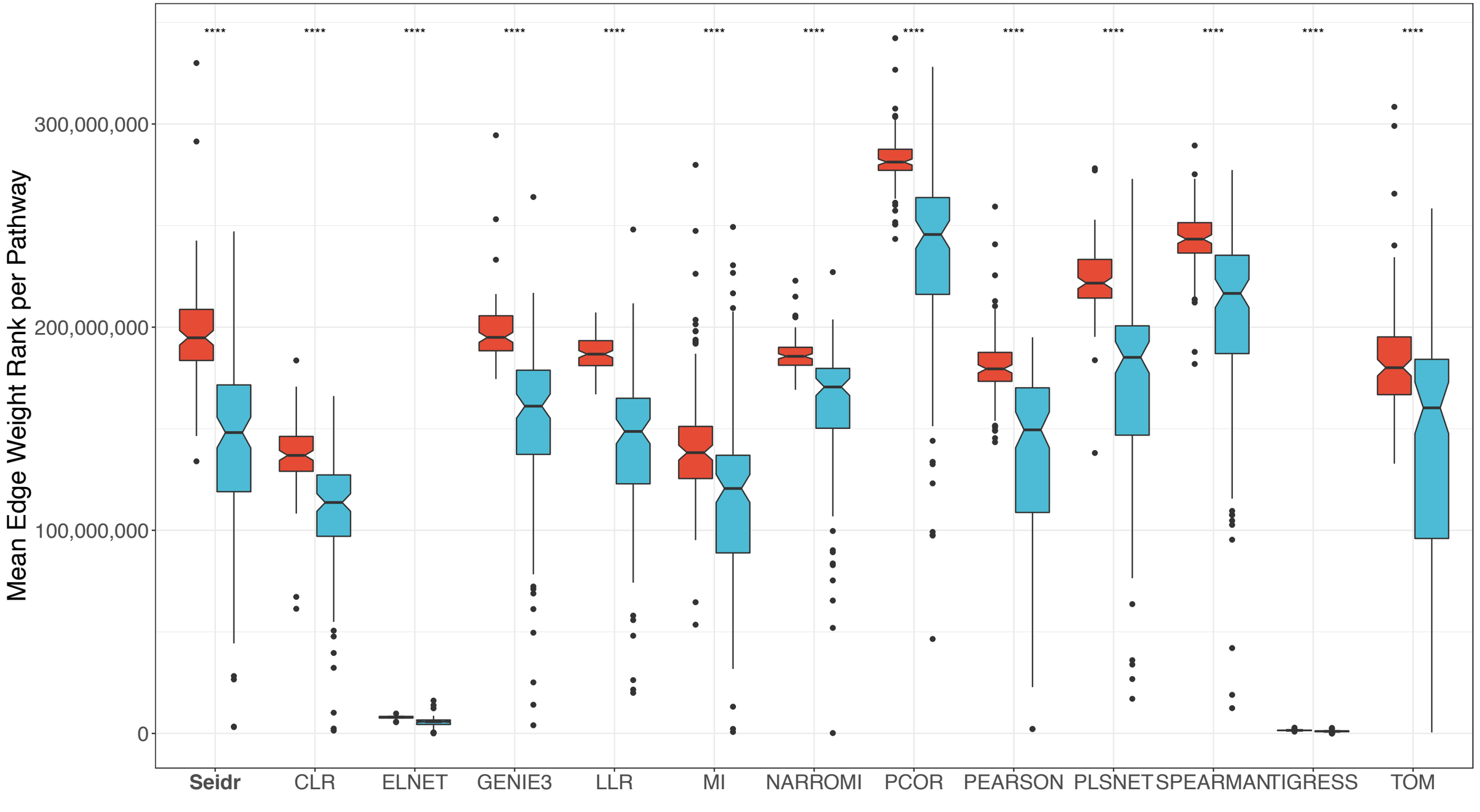
Is an ensemble really performing better?



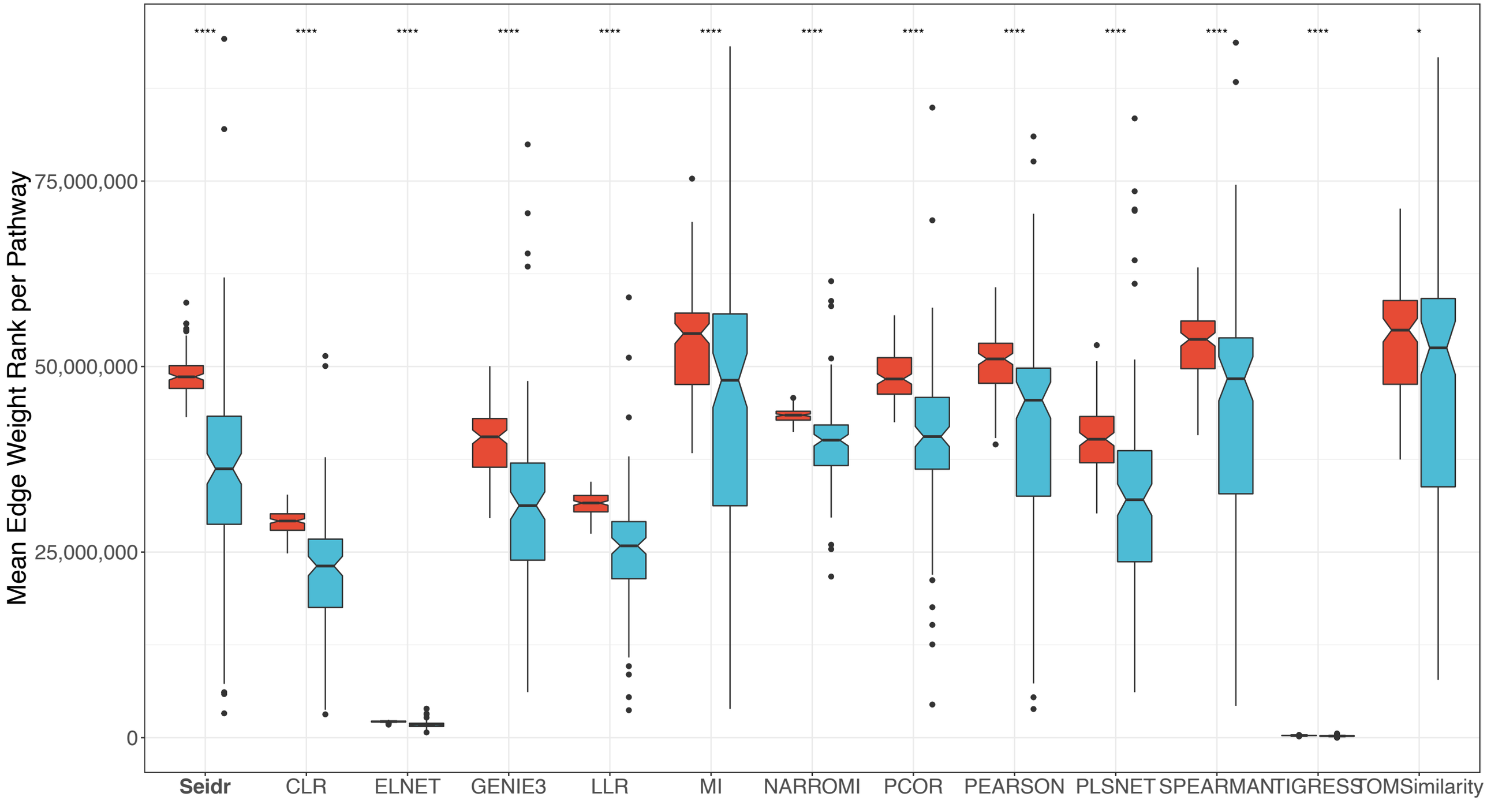
Pruning, local vs. global



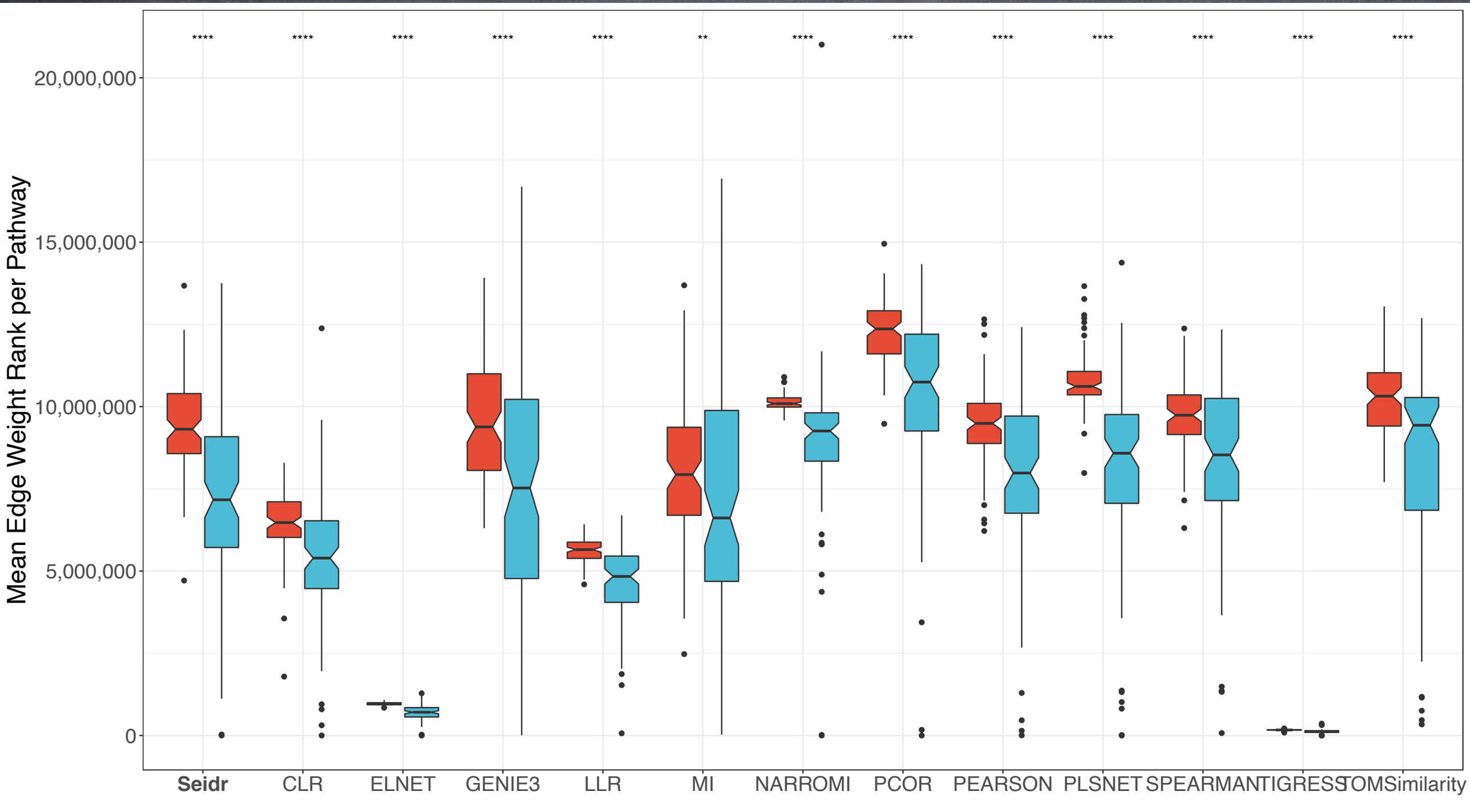
Performance on KEGG: Arabidopsis thaliana



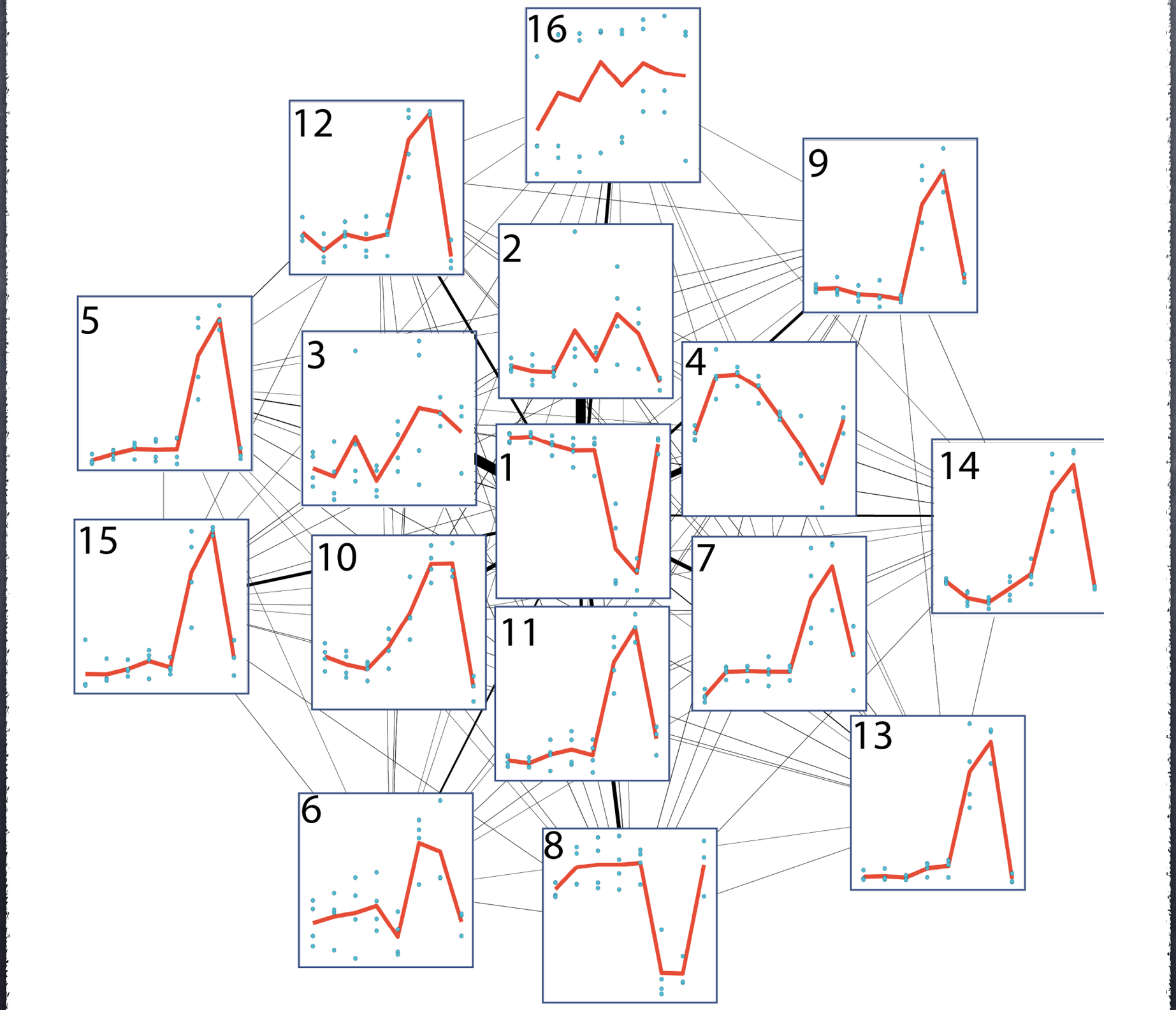
Performance on KEGG: Drosophila melanogaster



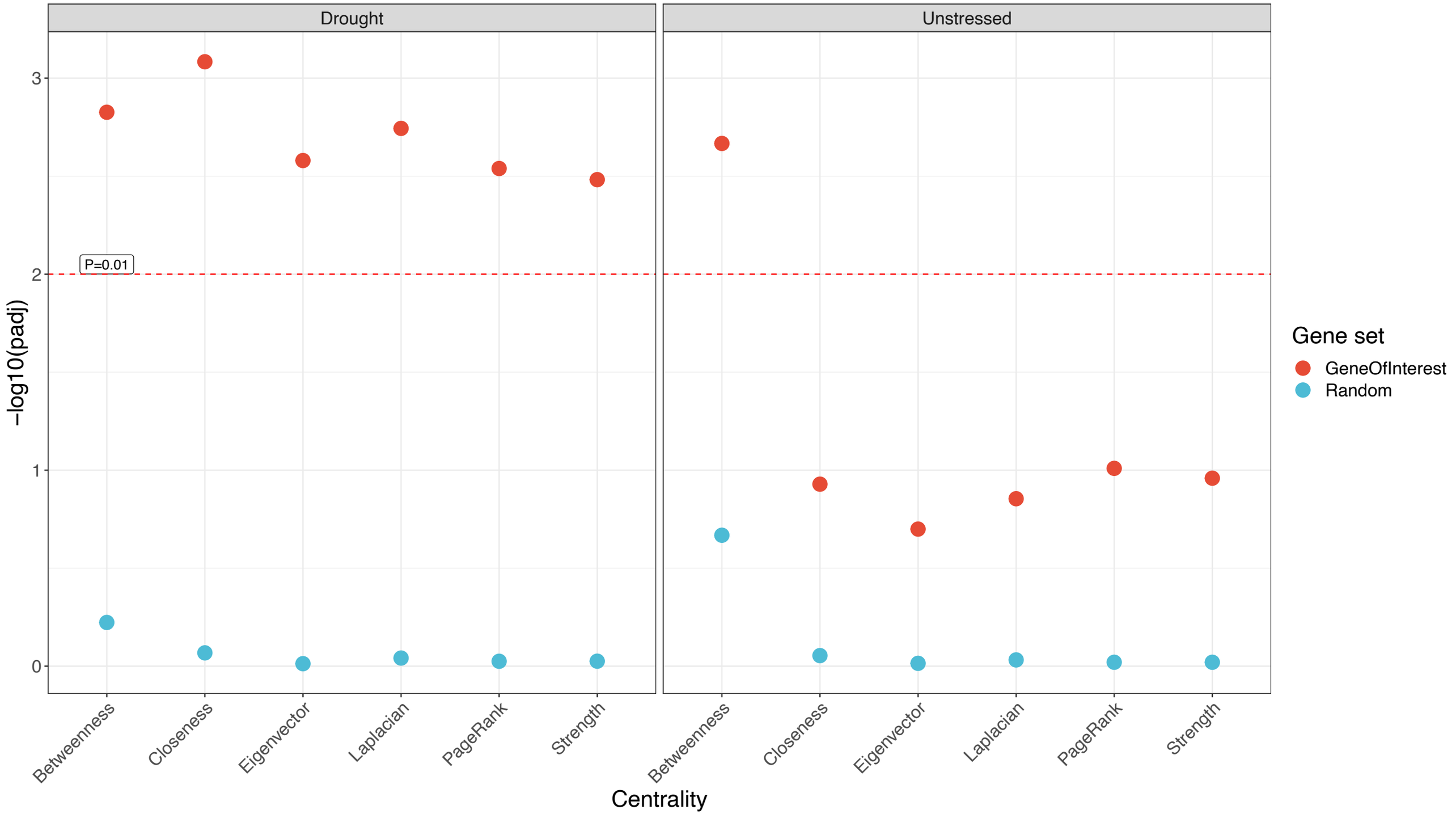
Performance on KEGG: Saccharomyces cerevisiae



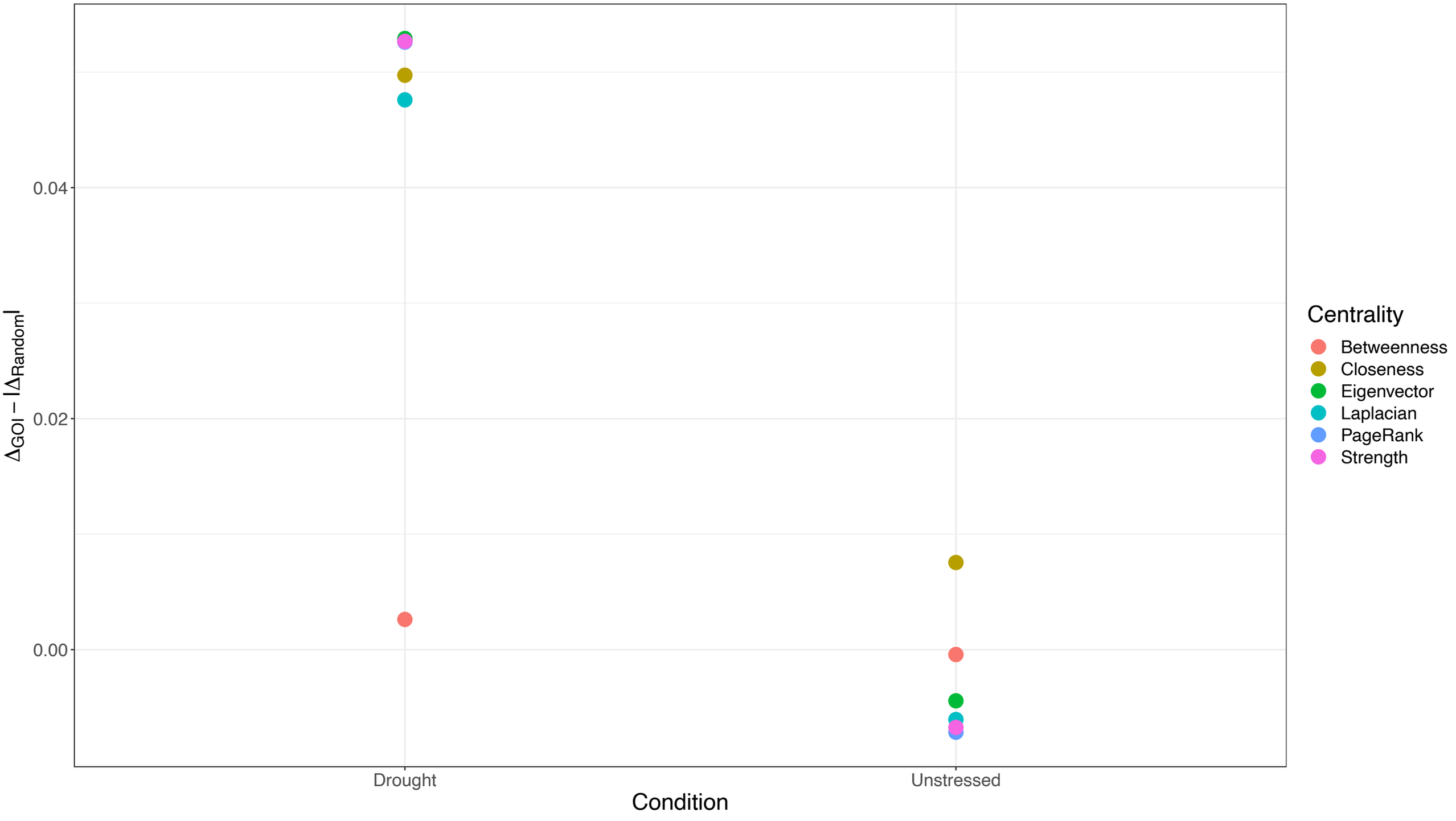
Comparative application: Picea abies drought vs. unstressed needles networks



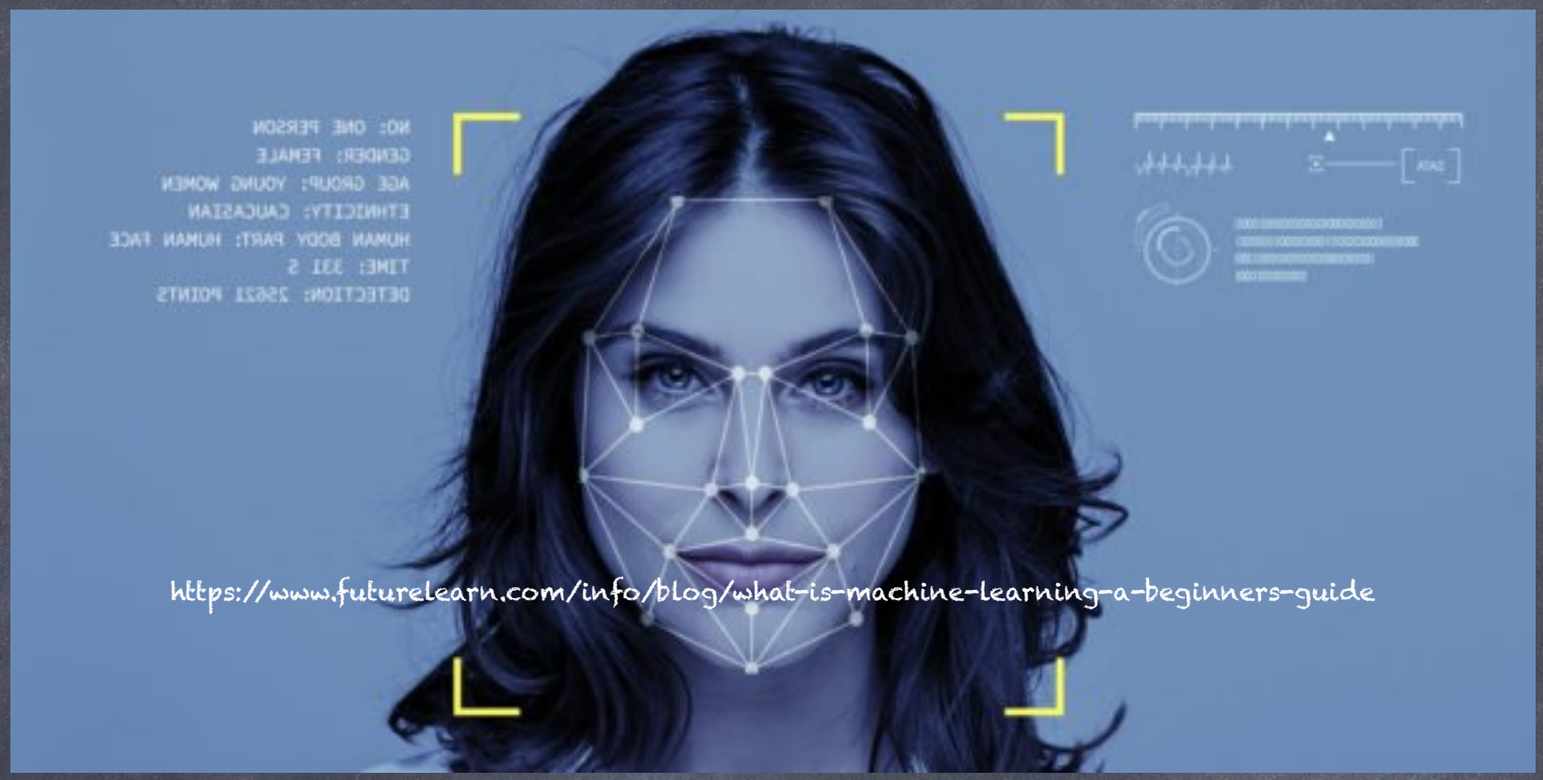
Gene of interest (GOI) centrality measure comparison



Gene of interest (GOI) signal amplitude comparison



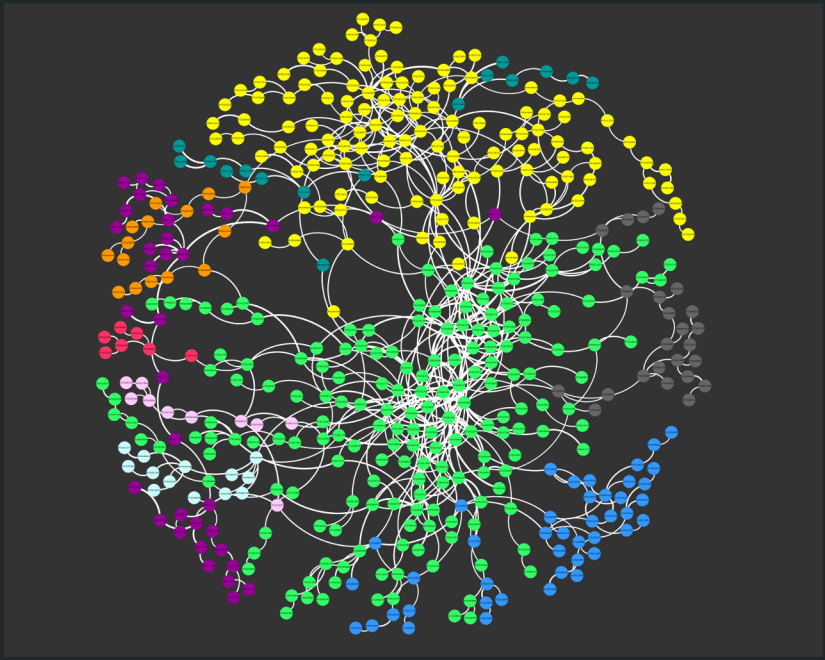
Where is the ML?



https://www.futurelearn.com/info/blog/what-is-machine-learning-a-beginners-guide

https://www.futurelearn.com/info/blog/what-is-machine-learning-a-beginners-guide

Gene networks



- To describe the complex phenomena
- Three types:
  - Co-expression networks
  - Association networks
  - Regulatory networks
- Multiple applications
  - **Functional gene prediction**
  - Gene prioritization
  - Gene Clustering
  - ...

Conifers

- Few number of species
- Dominant species in boreal forest
- Important role in carbon cycle
- Commercial interest to produce:
  - Raw material for paper
  - Solid fuels
  - Biomaterials
- Challenging organisms
  - Huge genomes (~20Gb)
  - Slow growth (generation time of 40 years, plots harvested after 100 y)

Norway spruce (*Picea abies*)

Norway spruce (*Picea abies*)

Gene overview

40,8% annotated coding genes in spruces

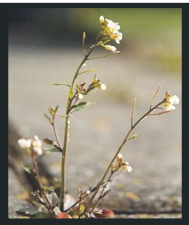
Species	#Genes	Coding	RNA	Pseudo	TE	GO	Interpro	Genes in non singleton gf	Genes in multi-species gf
Picea abies	66,632	66,632	0	0	0	27,222	37,386	56,442	54,230

gf:gene families ; 0 erroneous gene models where detected in this species

Species	#Genes	Coding	RNA	Pseudo	TE	GO	Interpro	Genes in non singleton gf	Genes in multi-species gf
Arabidopsis thaliana	33,602	27,416	1,359	924	3,903	27,123	23,129	25,589	23,543

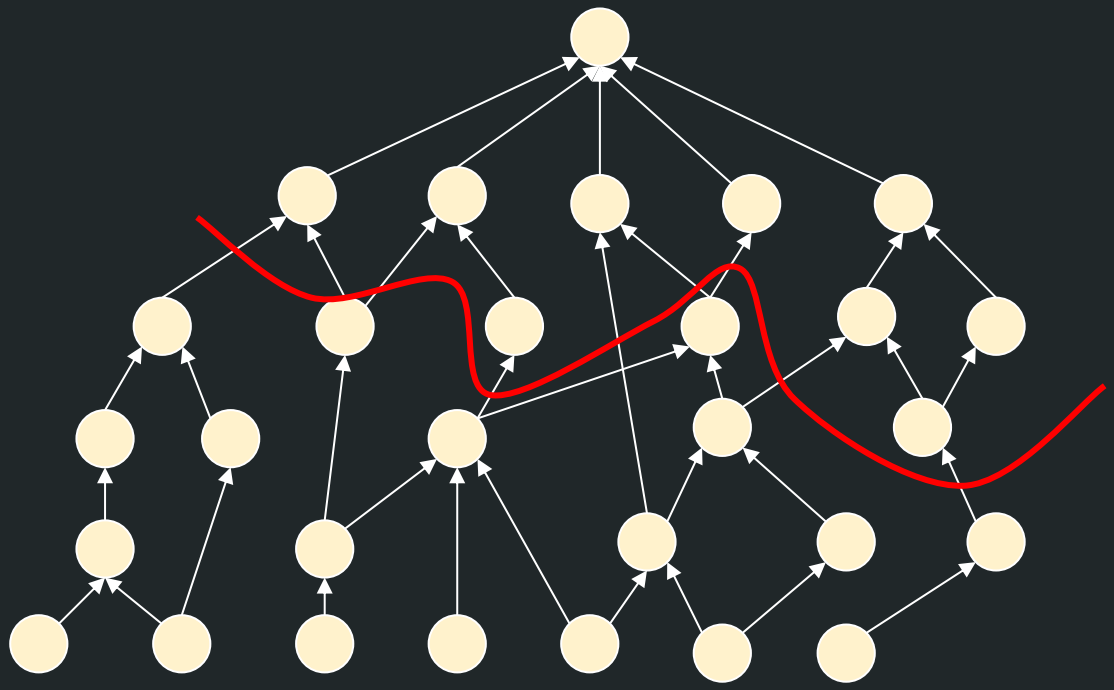
gf:gene families ; 46 erroneous gene models where detected in this species

98.9% annotated coding genes in arabidopsis

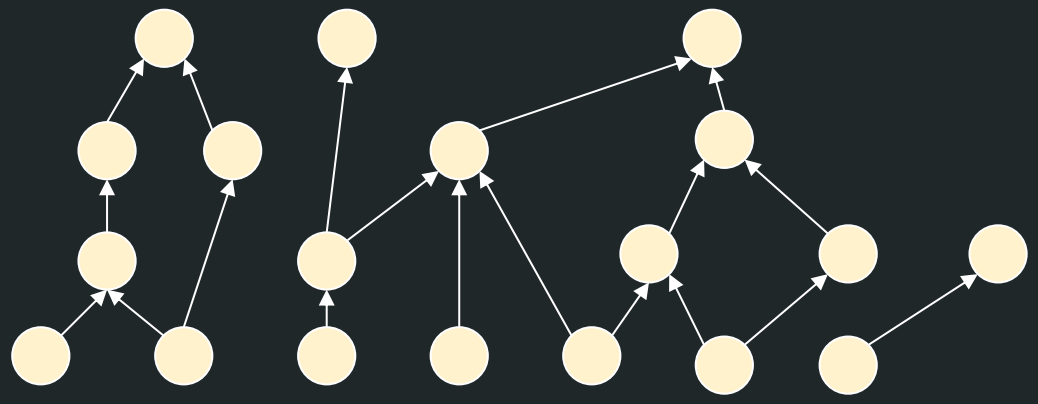


Source: https://bioinformatics.psb.ugent.be/plaza/versions/plaza_v4_dicots/

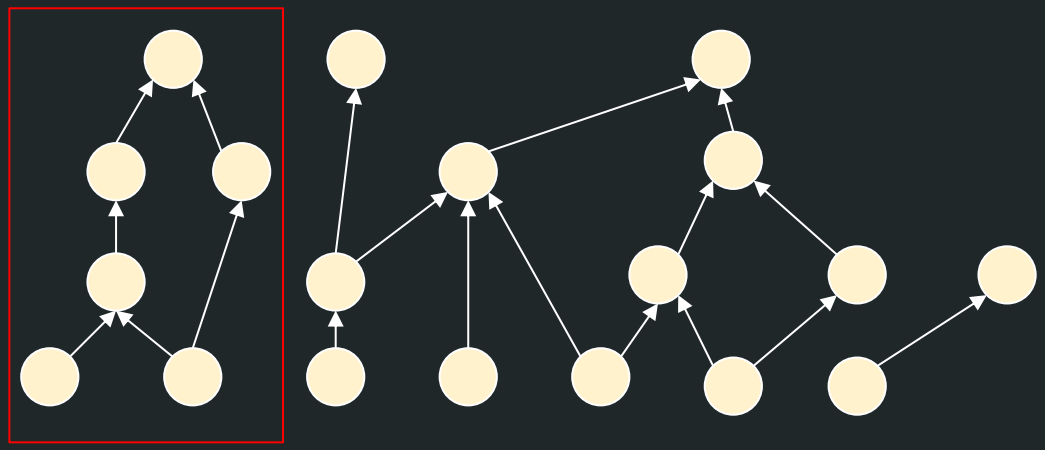
To provide a new gene predictor



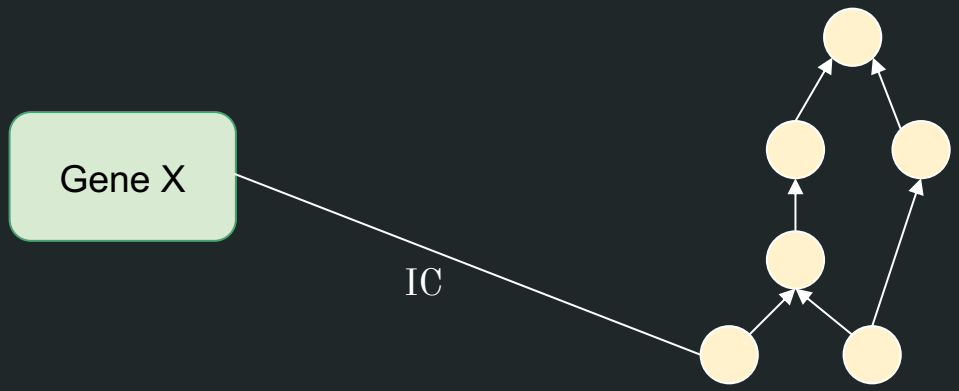
To provide a new gene predictor



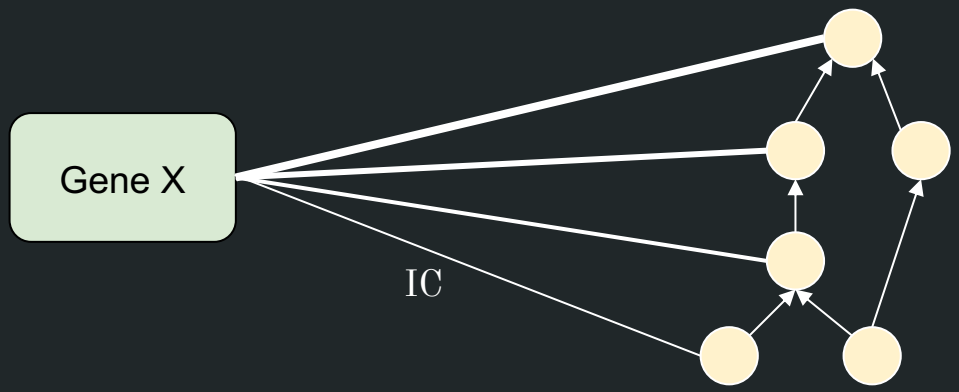
To provide a new gene predictor



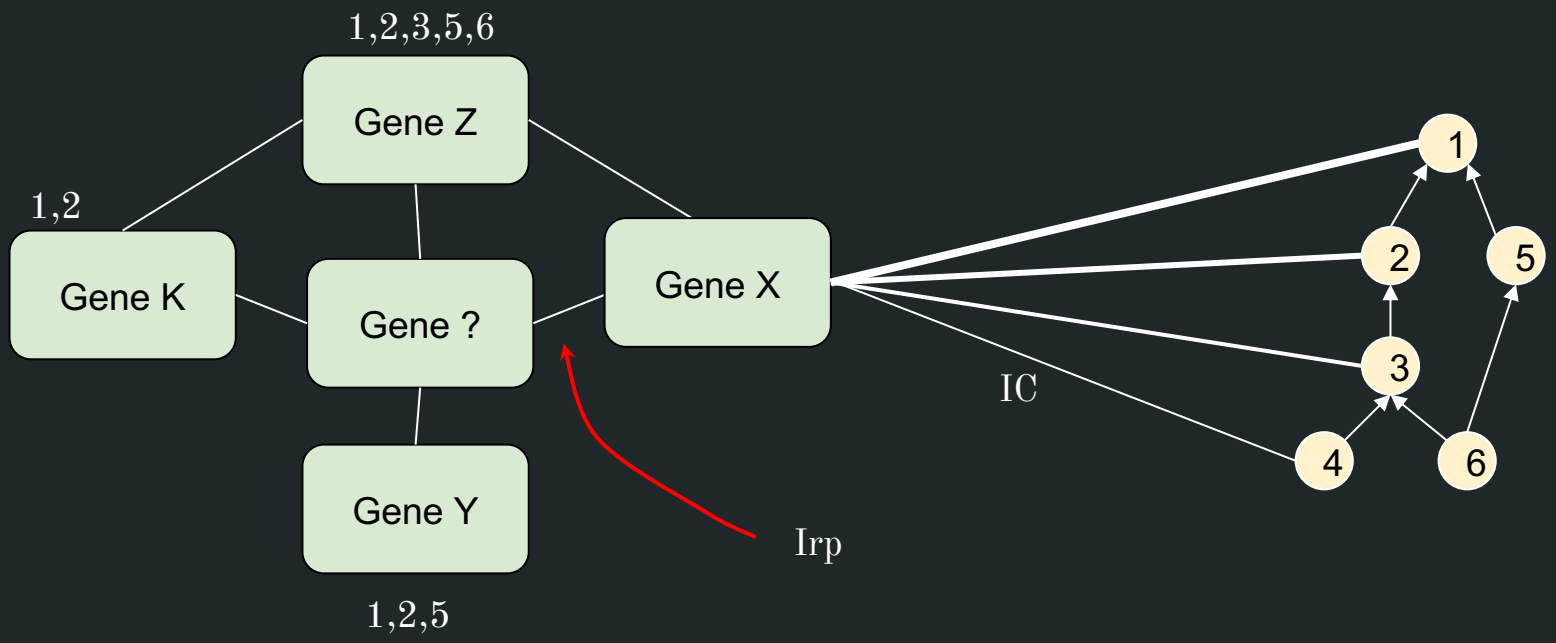
To provide a new gene predictor



To provide a new gene predictor



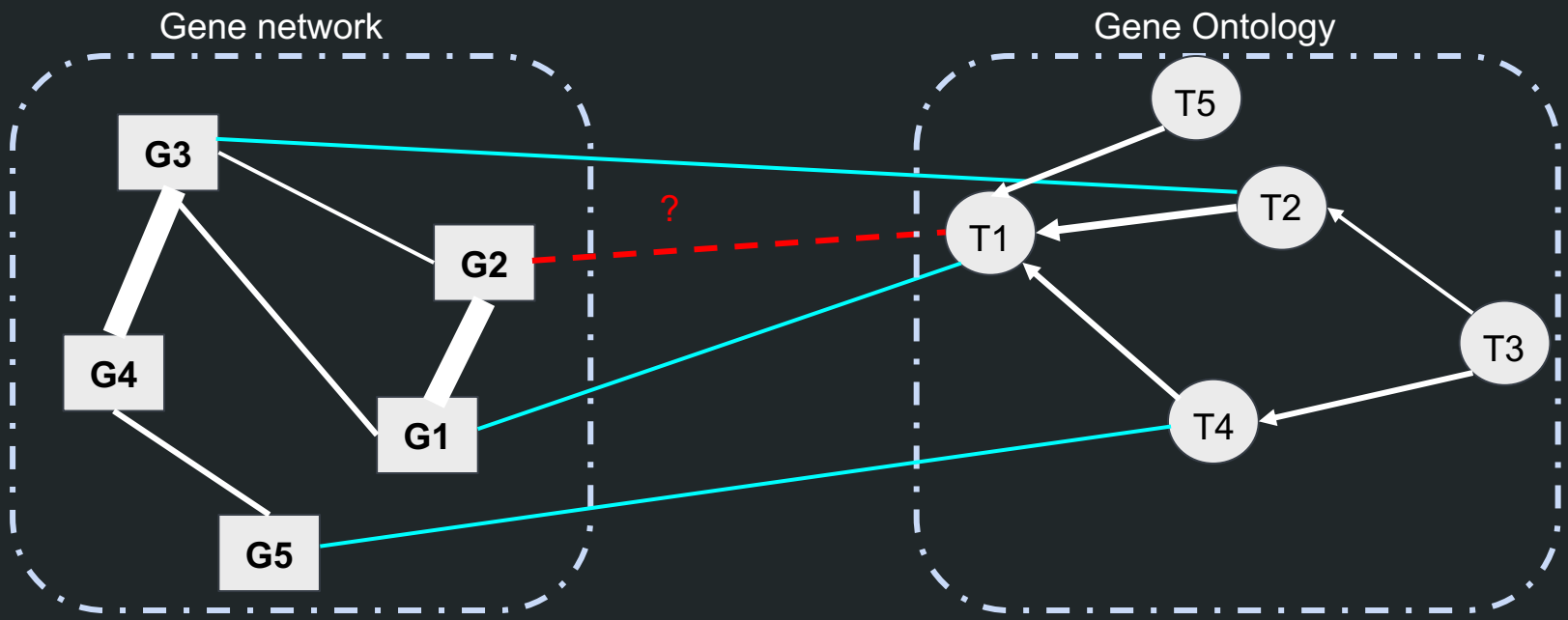
To provide a new gene predictor



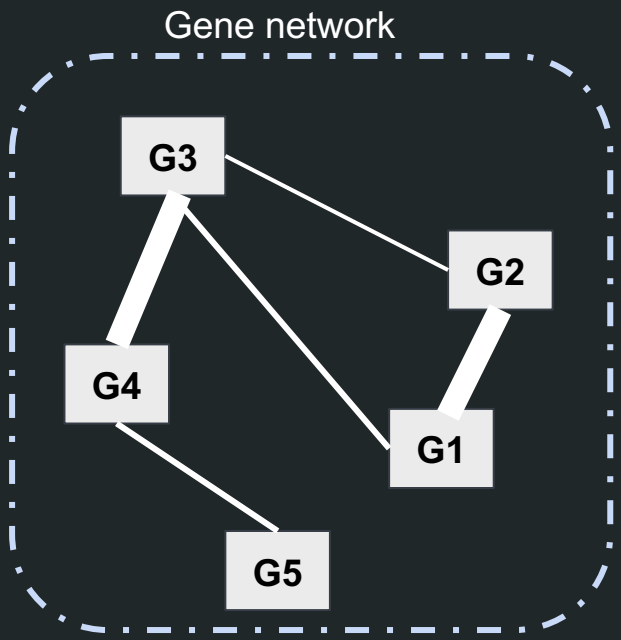
Gene ? will have GO terms 1, 2 (whose IC>p25)

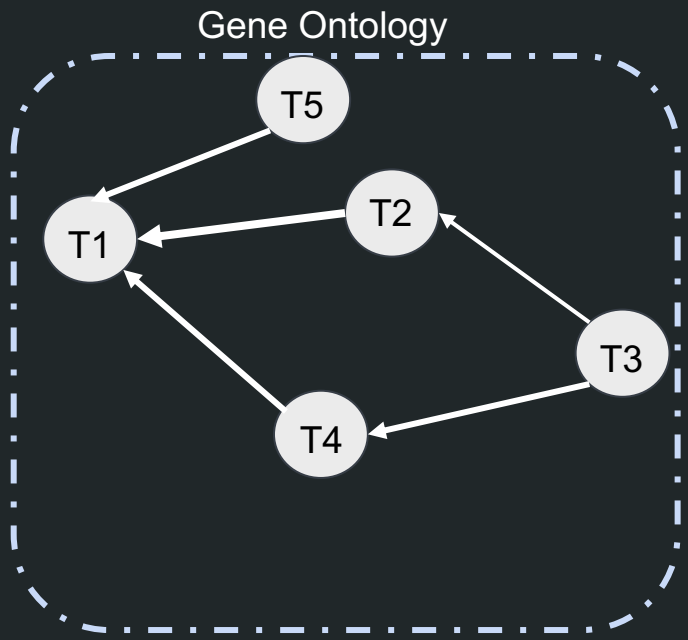
Gene predictor

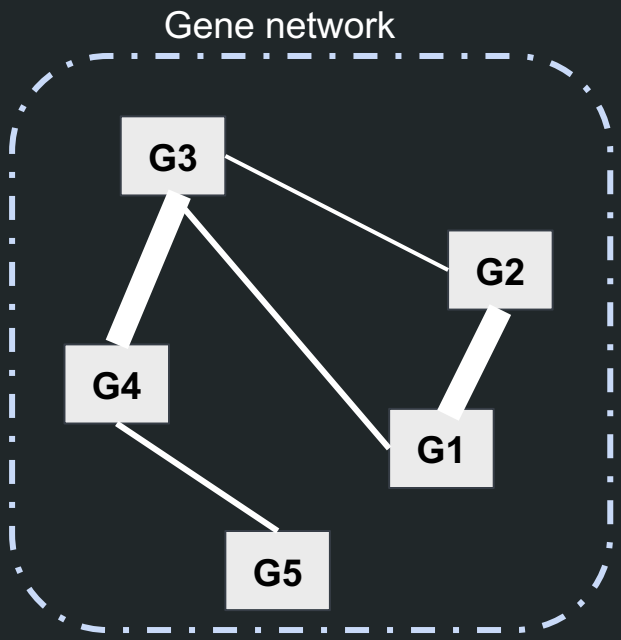
newGOA:

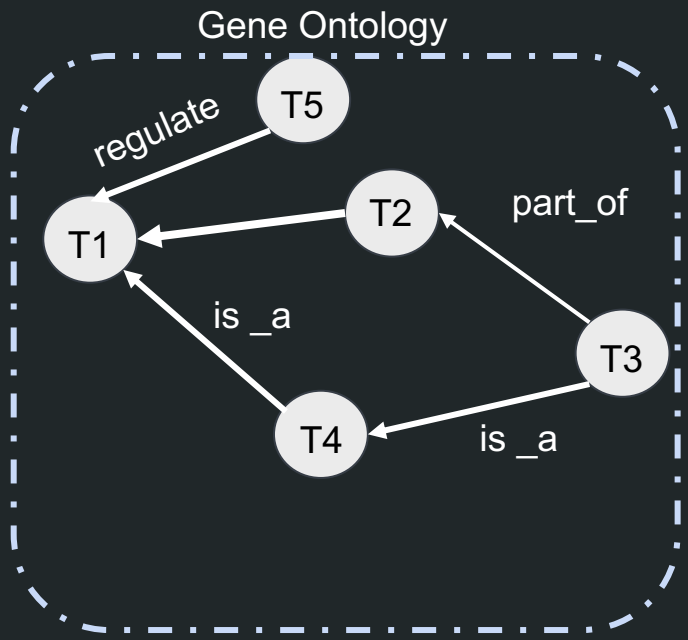


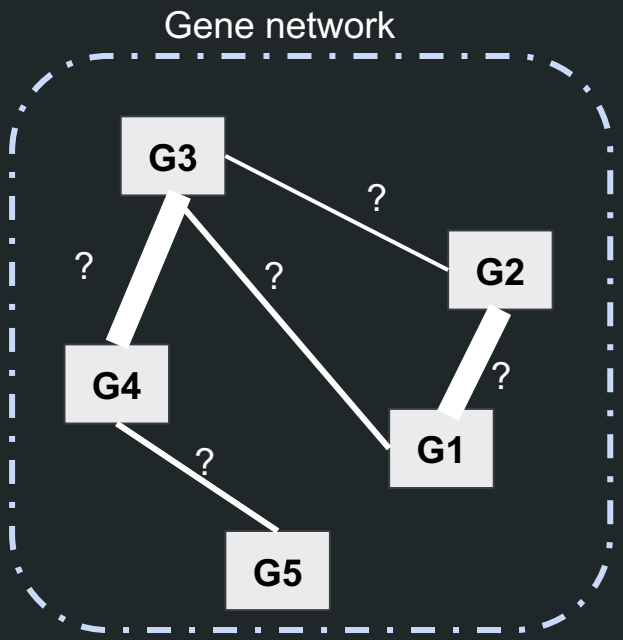
(Yu *et al. IEEE, 2018*)

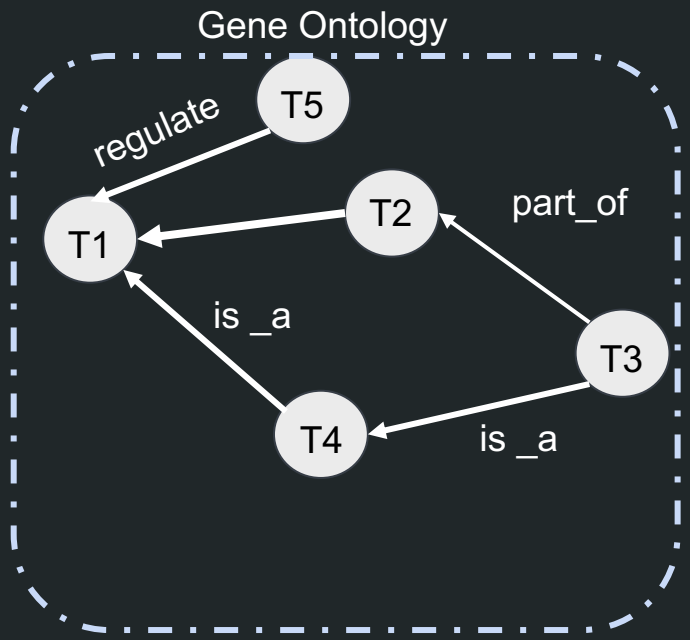

Gene network
G3
G2
G4
G1
G5

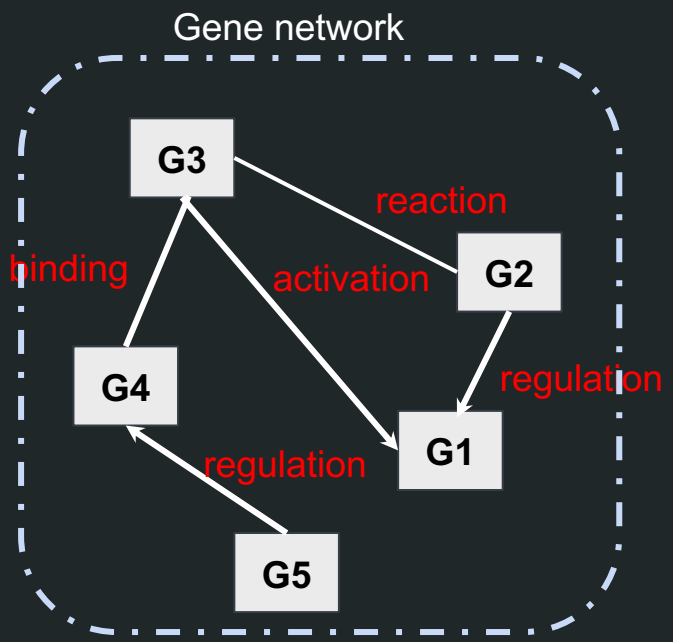

Gene Ontology
T5
T2
T1
T3
T4

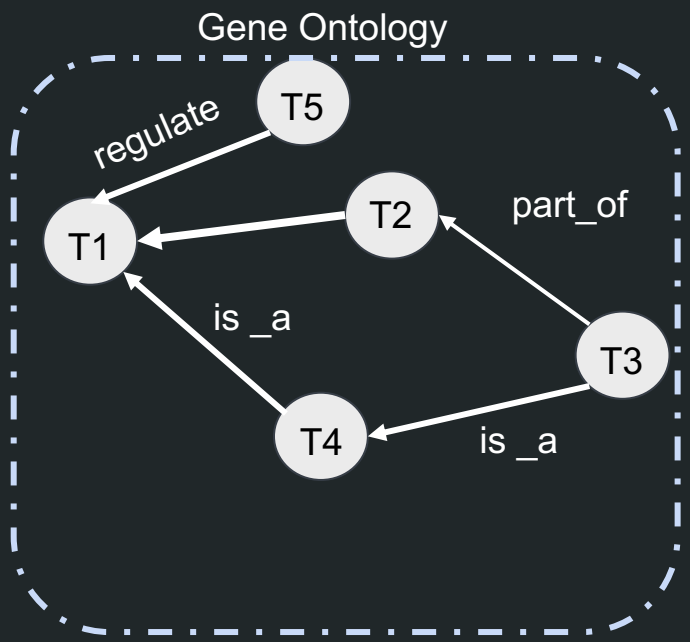

Gene network
G3
G2
G4
G1
G5

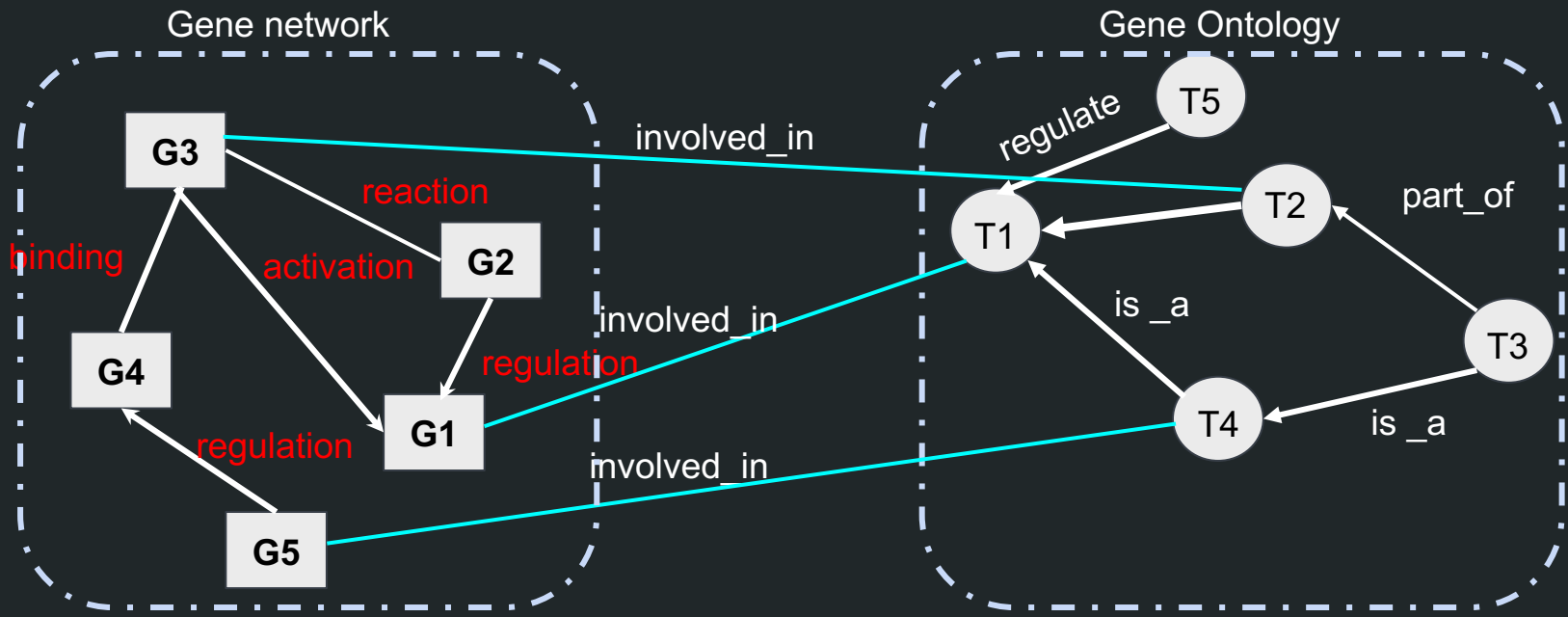

Gene Ontology
T5
regulate
T2
part_of
T1
is _a
T3
T4
is _a

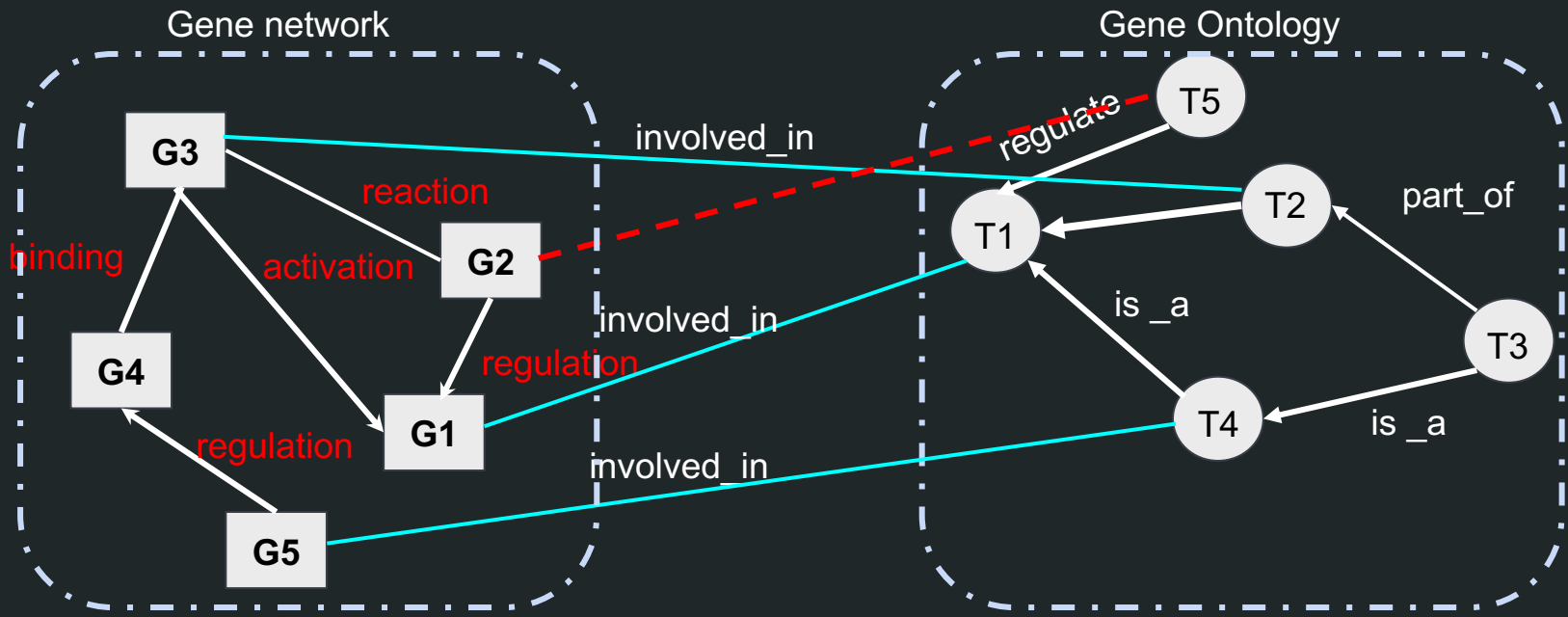

Gene network
G3
G2
G4
G1
G5
?
?
?
?
?


Gene Ontology
T5
regulate
T1
T2
part_of
is _a
T3
T4
is _a

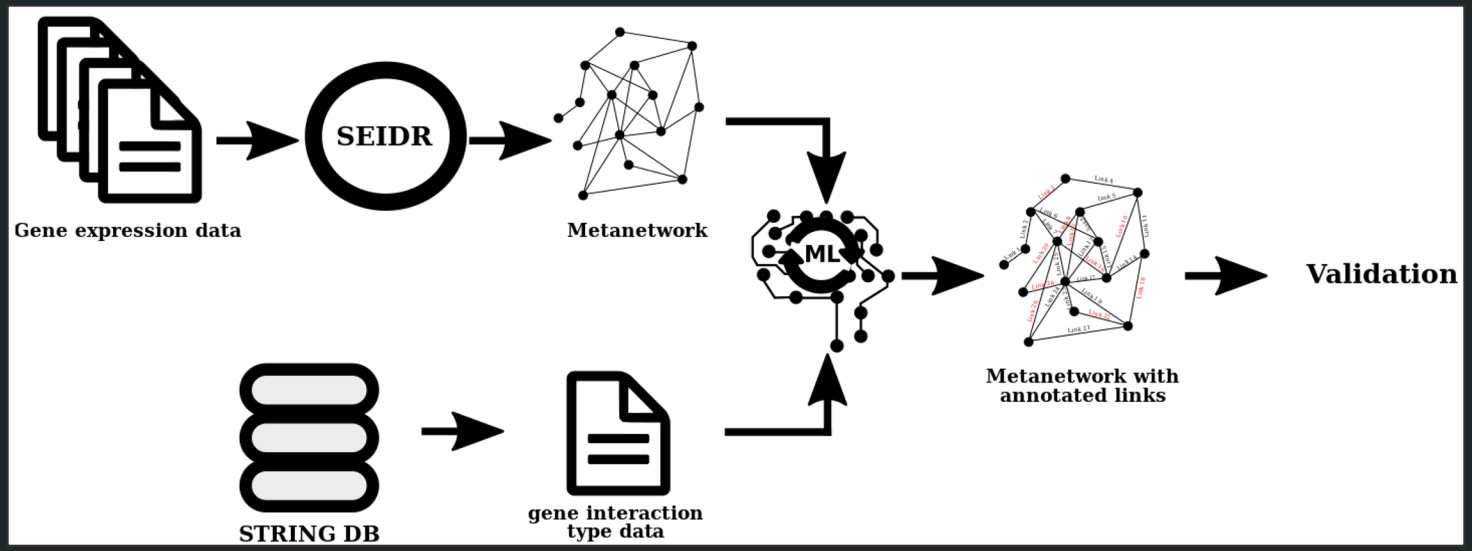

Gene network
G3
reaction
G2
binding
activation
G4
regulation
G1
regulation
G5


Gene Ontology
T5
regulate
T1
T2
part_of
is _a
T3
T4
is _a

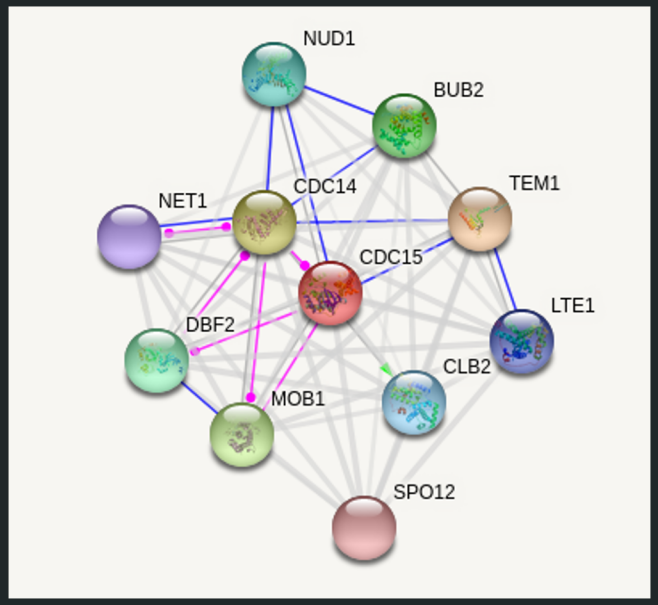

Gene network
Gene Ontology
G3
G2
G4
G1
G5
reaction
binding
activation
regulation
regulation
involved_in
involved_in
involved_in
T5
regulate
T2
part_of
T1
is _a
T3
T4
is _a


Gene network
Gene Ontology
G3
G2
G4
G1
G5
reaction
binding
activation
regulation
regulation
involved_in
involved_in
involved_in
T5
regulate
T2
part_of
T1
is _a
T3
T4
is _a

To predict link annotation



STRING



Action Types

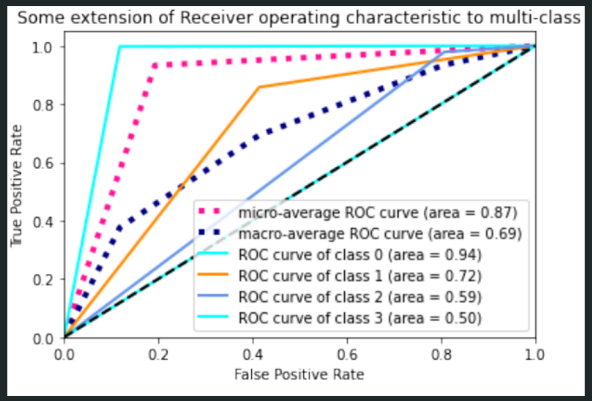
- activation
- inhibition
- binding
- catalysis
- phenotype
- posttranslational modification
- reaction
- transcriptional regulation

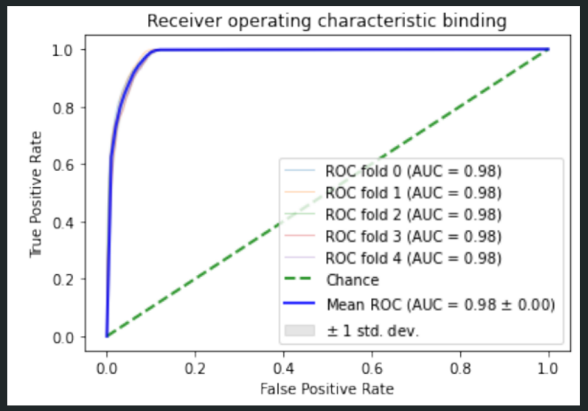
Action effects

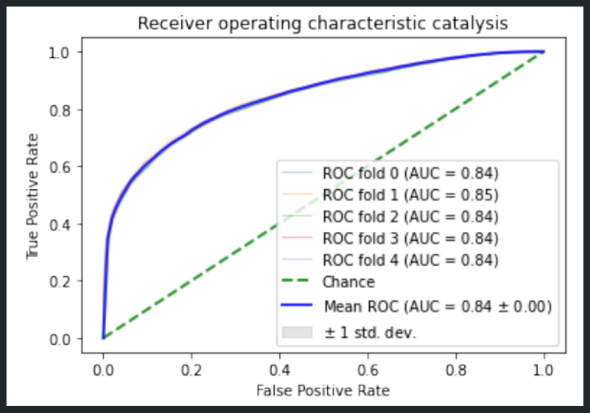
- positive
- negative
- unspecified

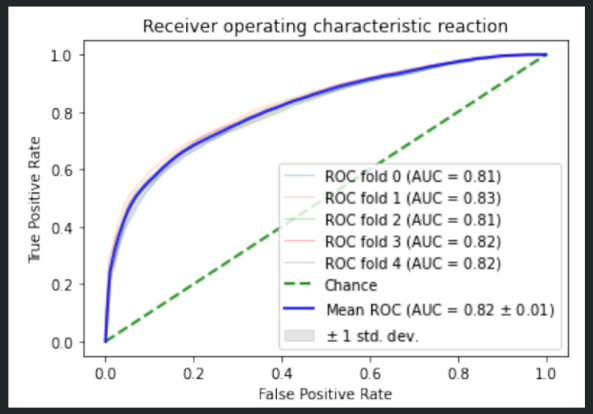
(Franceschini *et al. NAR, 2013*)

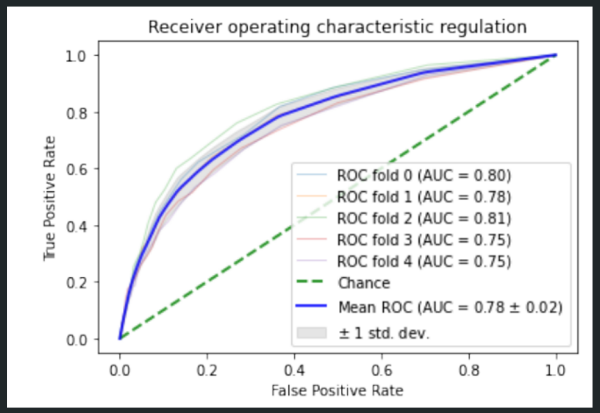
Project link annotation











AUC - F1

	Binding		Catalysis		Reaction		Regulation	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
0.6	0.94	0.97	0.76	0.80	0.64	0.88	0.5	0
0.7	0.93	0.97	0.76	0.75	0.71	0.85	0.5	0
0.8	0.93	0.96	0.73	0.66	0.74	0.77	0.5	0
0.85	0.92	0.94	0.70	0.60	0.73	0.71	0.5	0
0.9	0.90	0.91	0.68	0.54	0.70	0.62	0.5	0
0.95	0.84	0.82	0.63	0.43	0.64	0.46	0.5	0
0.99	0.67	0.52	0.55	0.20	0.55	0.19	0.5	0

Homework

- In linux
  - run one (or a few) inference method(s)
  - aggregate and prune the results
- In R
  - explore the network neighbourhood of a GOI
  - investigate it (GO, network properties, annotation, etc.)