Marseille Medical Genetics

IBENS
Institut de Biologie de l'Ecole Normale Supérieure

Aix*Marseille université

ENS
ÉCOLE NORMALE SUPÉRIEURE

PSL
RESEARCH UNIVERSITY PARIS

Multi-omics data integration mini-workshop

Laura Cantini / Anaïs Baudot

Anaïs Baudot
anais.baudot@univ-amu.fr

Laura Cantini
laura.cantini@ens.fr

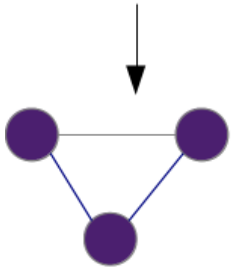
cnrs

Inserm

Nice 19/04

Systems Biology: The Complex Systems framework for Biology

Genotype

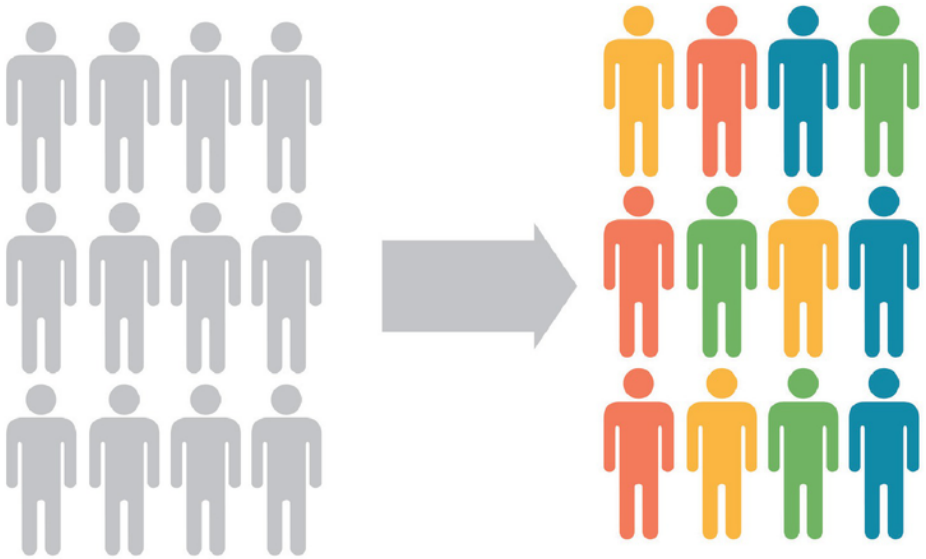


-Omics Networks

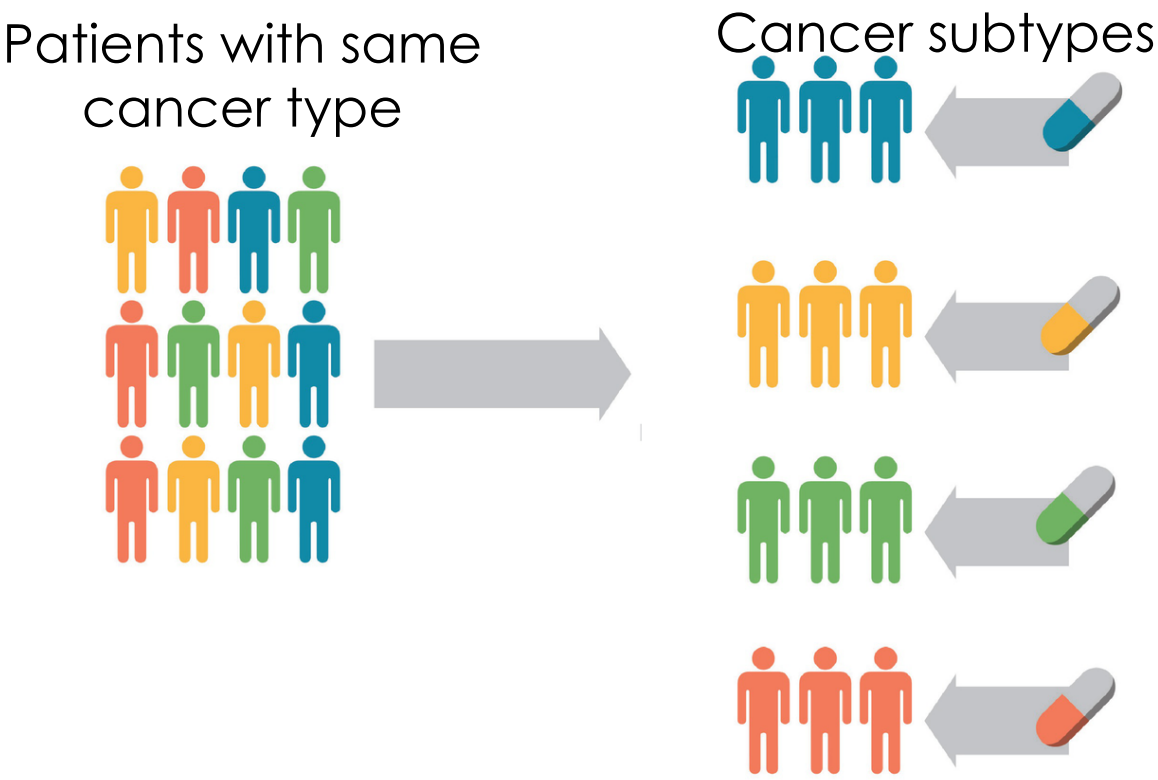
Phenotype

Personalized cancer medicine

Patients with same cancer type don't have the same survival, treatment response and molecular characteristics

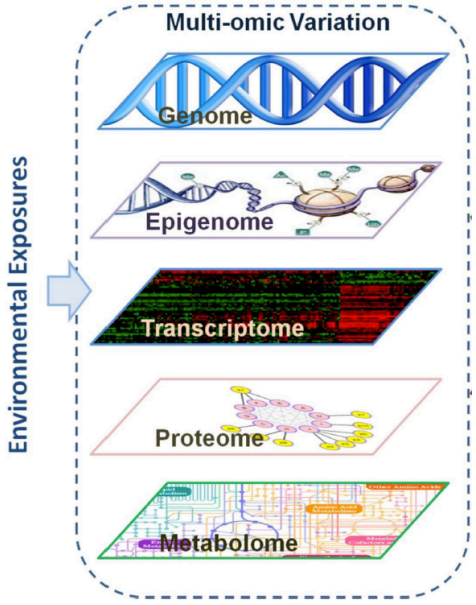


Personalized cancer medicine

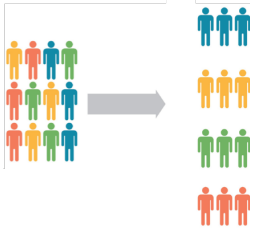


Classify cancer patients into groups with similar prognosis, drug response or molecular features

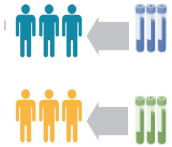
Multi-omics data improve personalized medicine



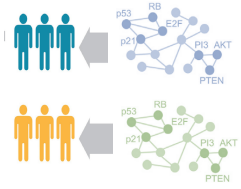
Cancer subtyping



Biomarker discovery

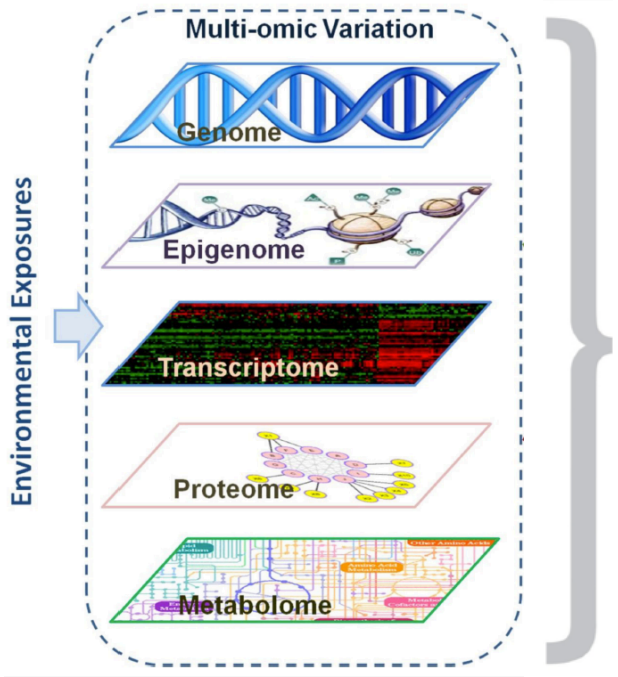


Regulatory mechanisms



Drug discovery and repositioning

Multi-omics data are easily accessible

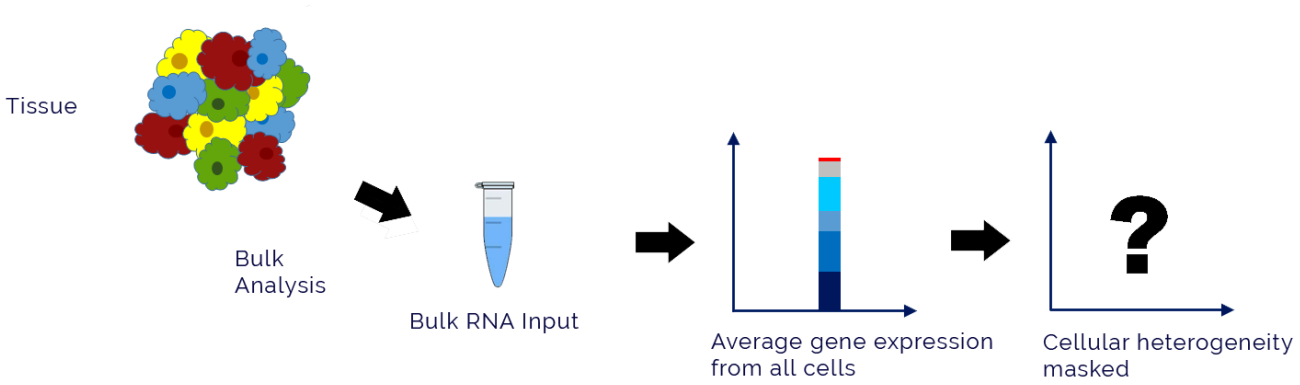


The Cancer Genome Atlas (TCGA)

contains data from 10.000 patients, 33 cancer types, 6 omics, plus clinical data

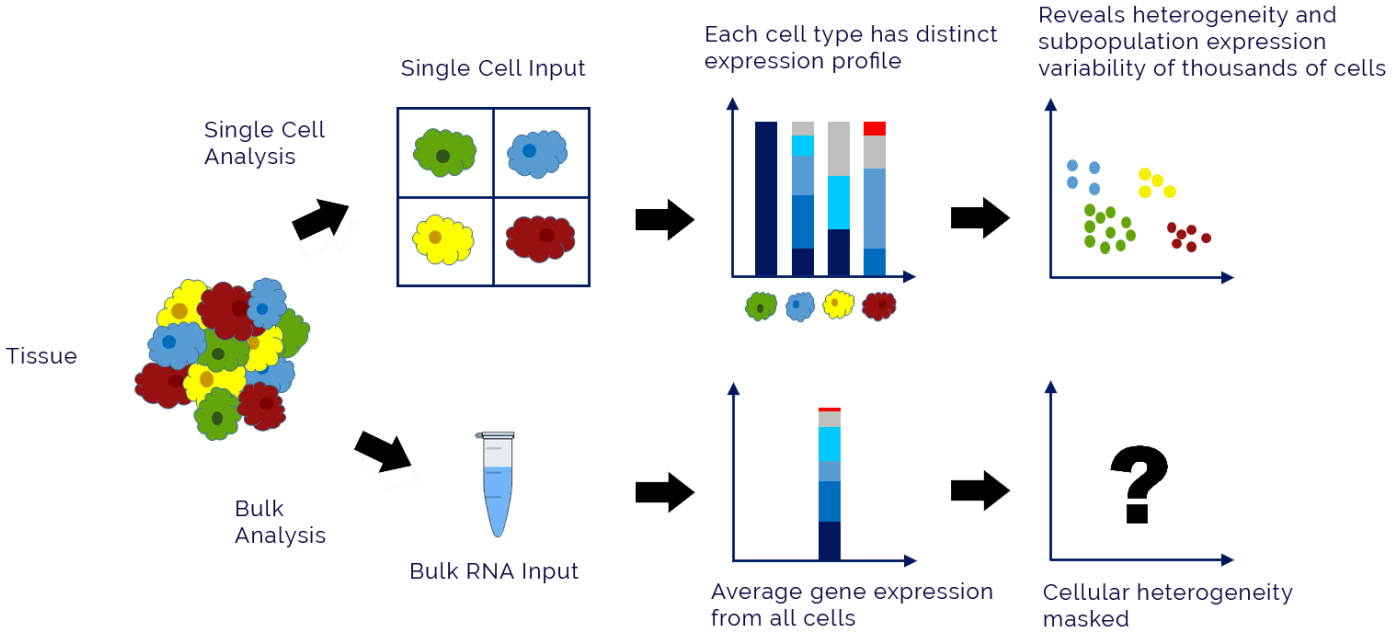
The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi:10.1038/ng.2764
Sun, Yan V., and Yi-Juan Hu. "Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases." *Advances in genetics*. Vol. 93. Academic Press, 2016. 147-190.

The single-cell revolution



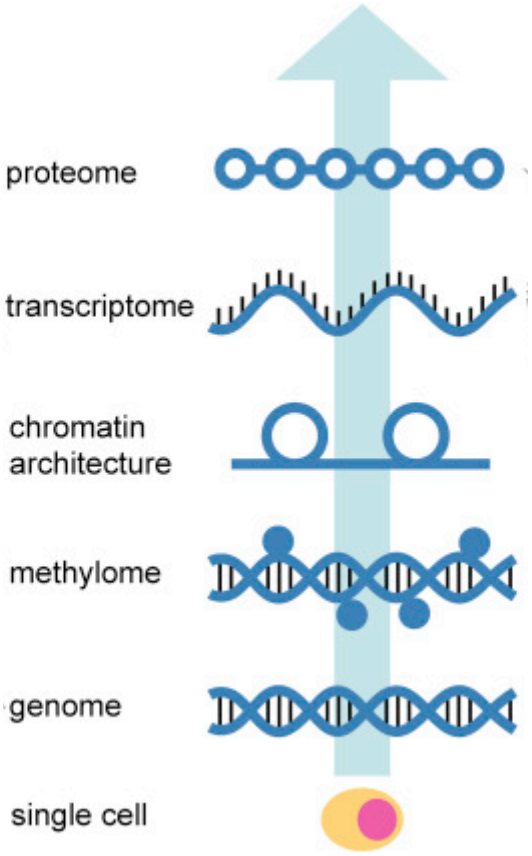
https://www.10xgenomics.com/single-cell-technology

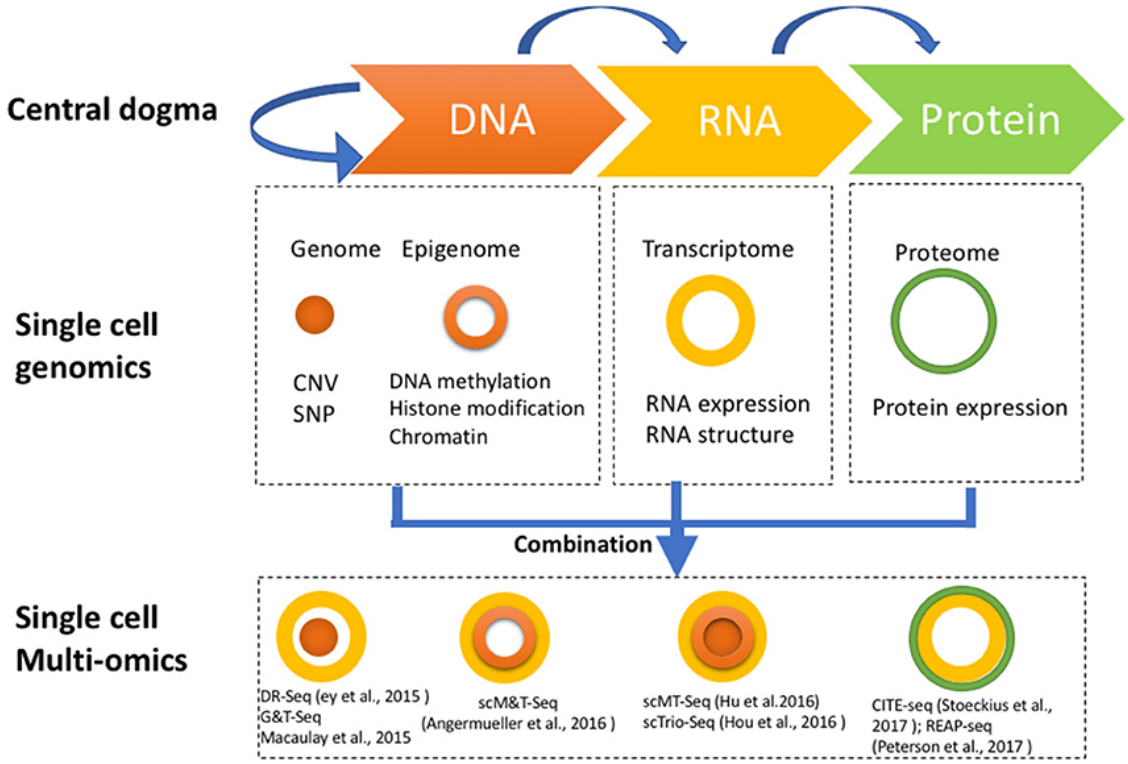
The single-cell revolution



https://www.10xgenomics.com/single-cell-technology

Single-cell multi-omics data

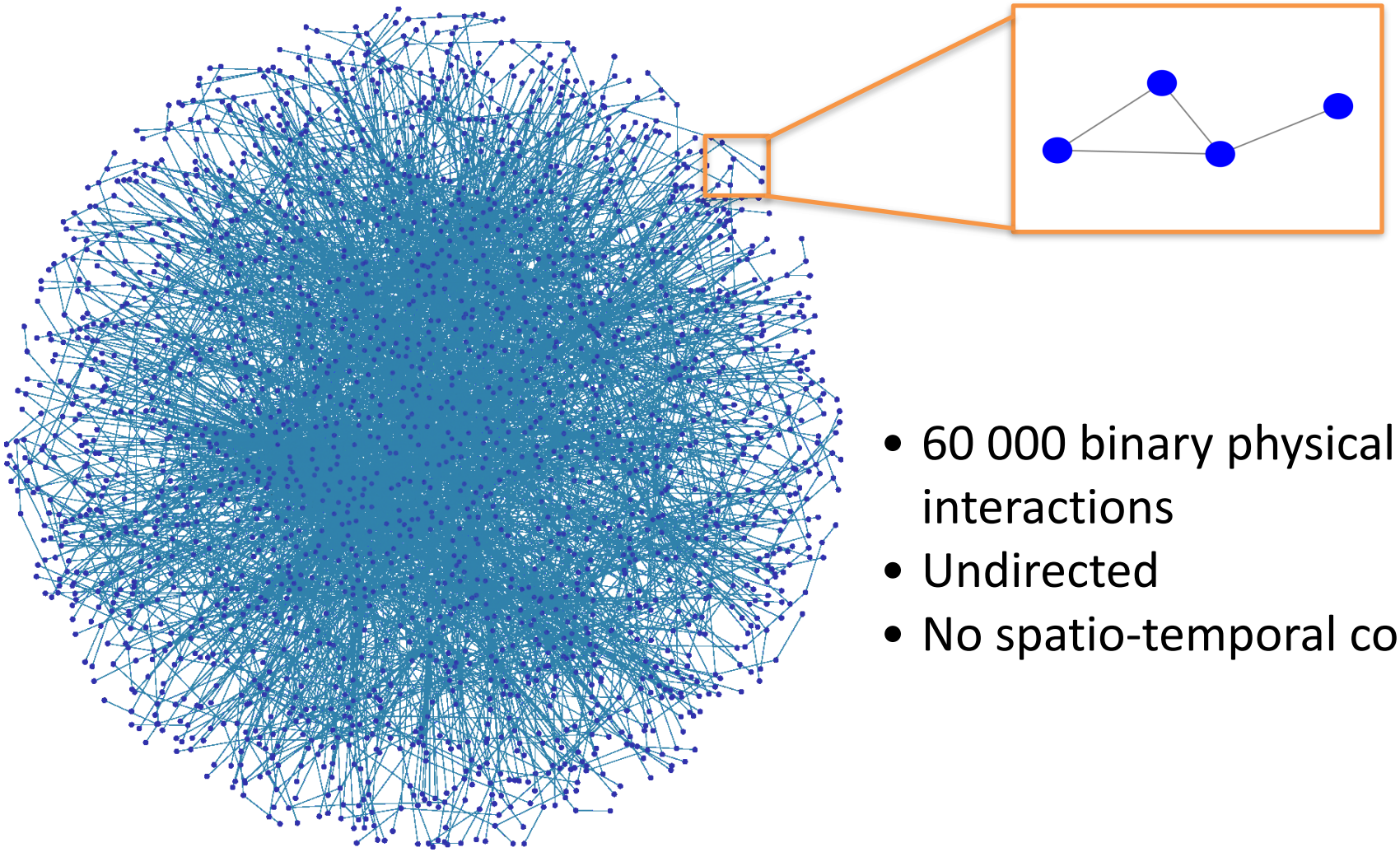




Colomé-Tatché, Maria, and Fabian J. Theis *Current Opinion in Systems Biology* 7 (2018): 54-59.

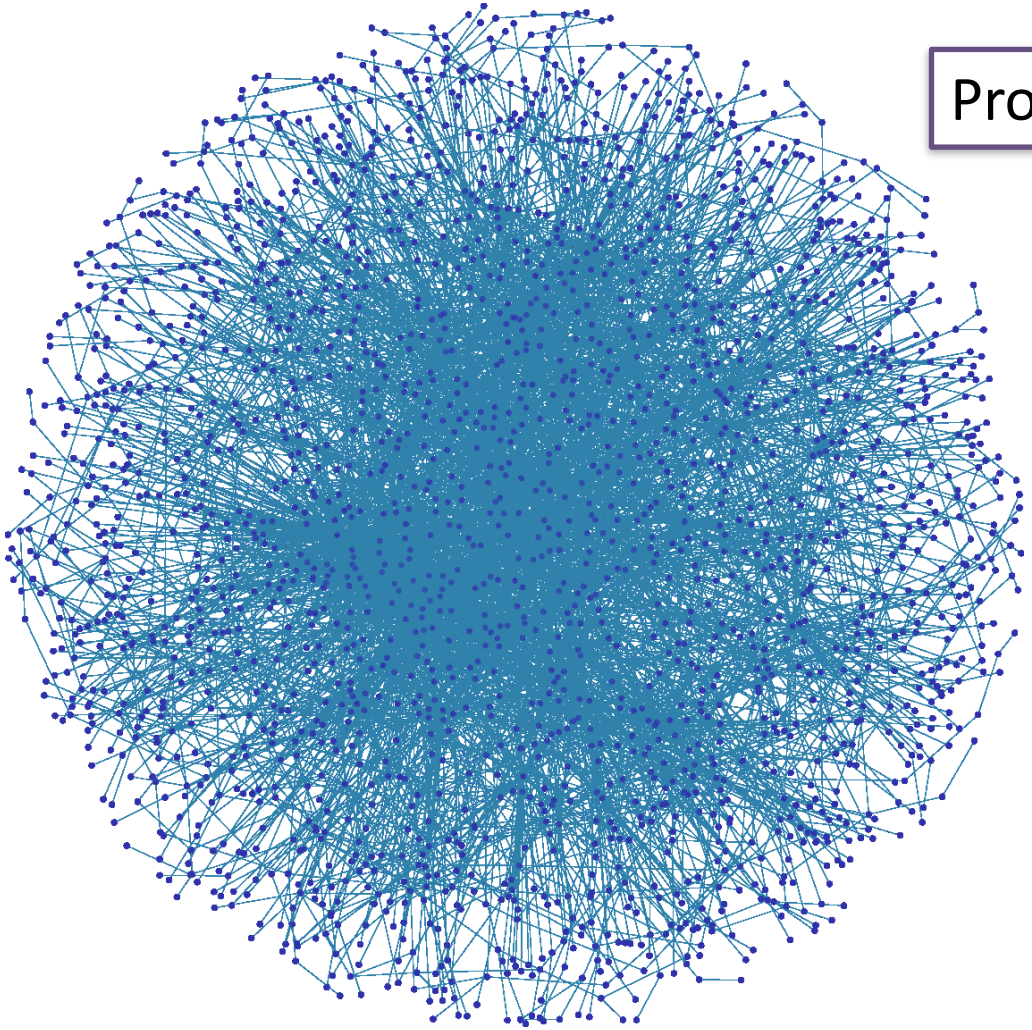
Hu, Youjin, et al. " *Frontiers in cell and developmental biology* 6 (2018): 28.

Interactome: The set of known & possible human protein-protein interactions



- 60 000 binary physical interactions
- Undirected
- No spatio-temporal context

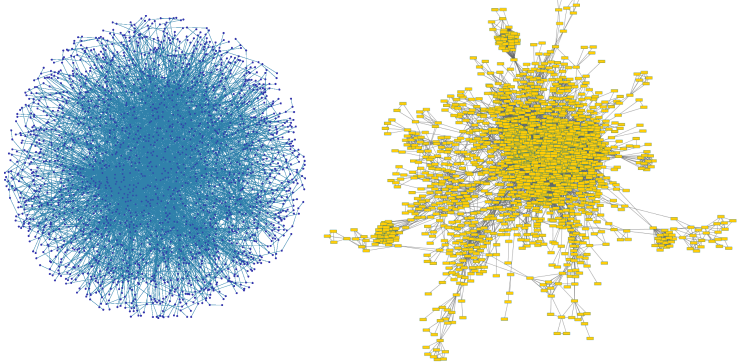
Interactome: The set of known & possible human protein-protein interactions



Protein interactions

Processes / Functions

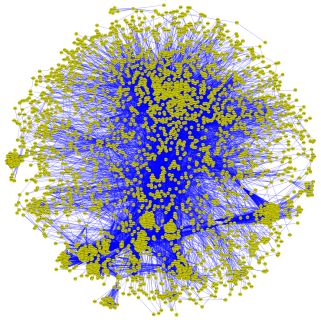
Many Biological Networks



Measured networks

Many Biological Networks

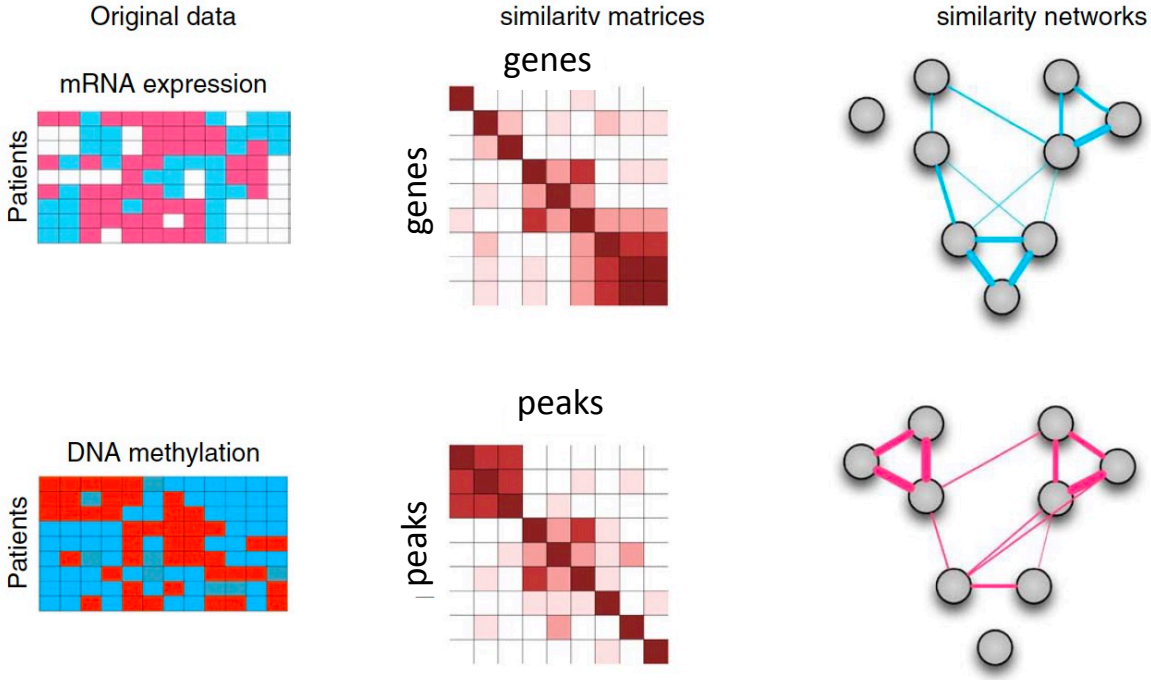
Pathways



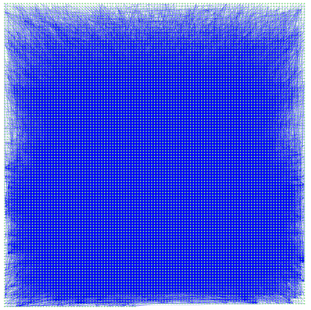
~250 000 edges

Curated networks

Many Biological Networks



Correlation of expression



~1 400 000 edges

Inferred networks

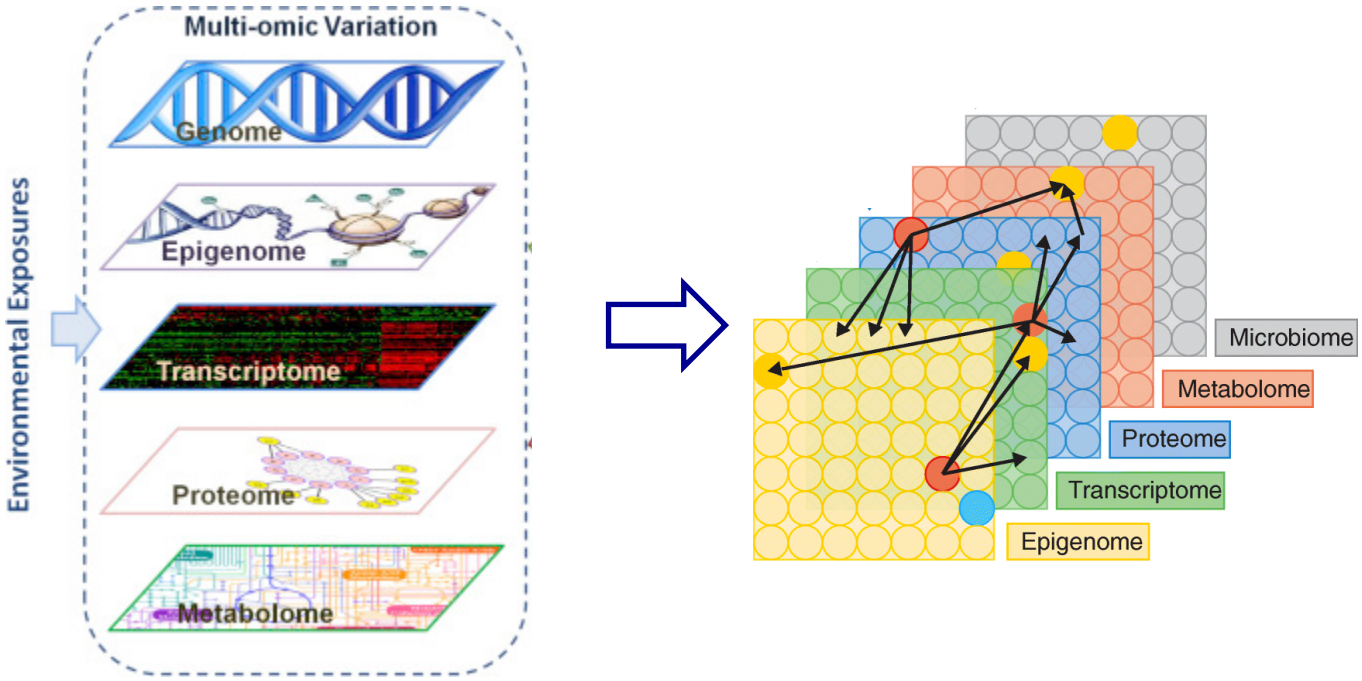
Why Integration?

Integrating multiple sources of data allows:

- To reduce the effects of experimental and biological noise
- To capture different aspects of cellular functioning
  - the different omics are complementary
  - more comprehensive overview of biological systems

The joint analysis of multiple omics/networks is required

Multi-omics data as matrices

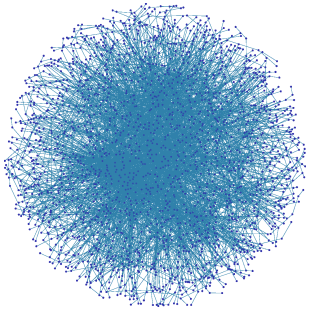
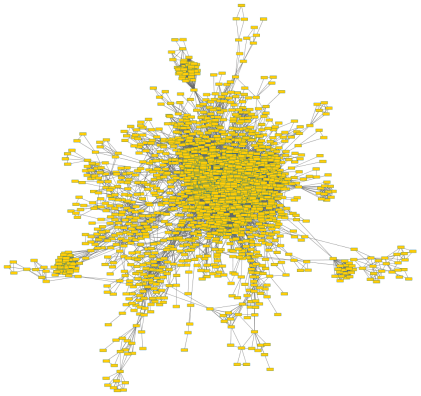
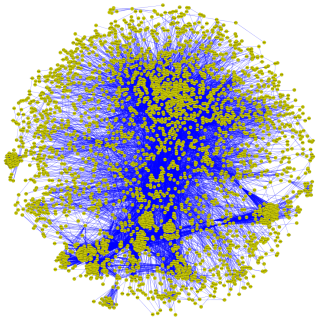
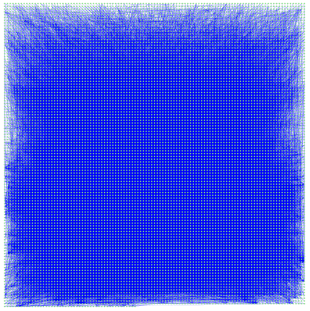


The joint analysis of multiple omics is required => joint Dimensionality Reduction

The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi:10.1038/ng.2764

Sun, Yan V., and Yi-Juan Hu. "Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases." *Advances in genetics*. Vol. 93. Academic Press, 2016. 147-190.

Multi-omics data as networks

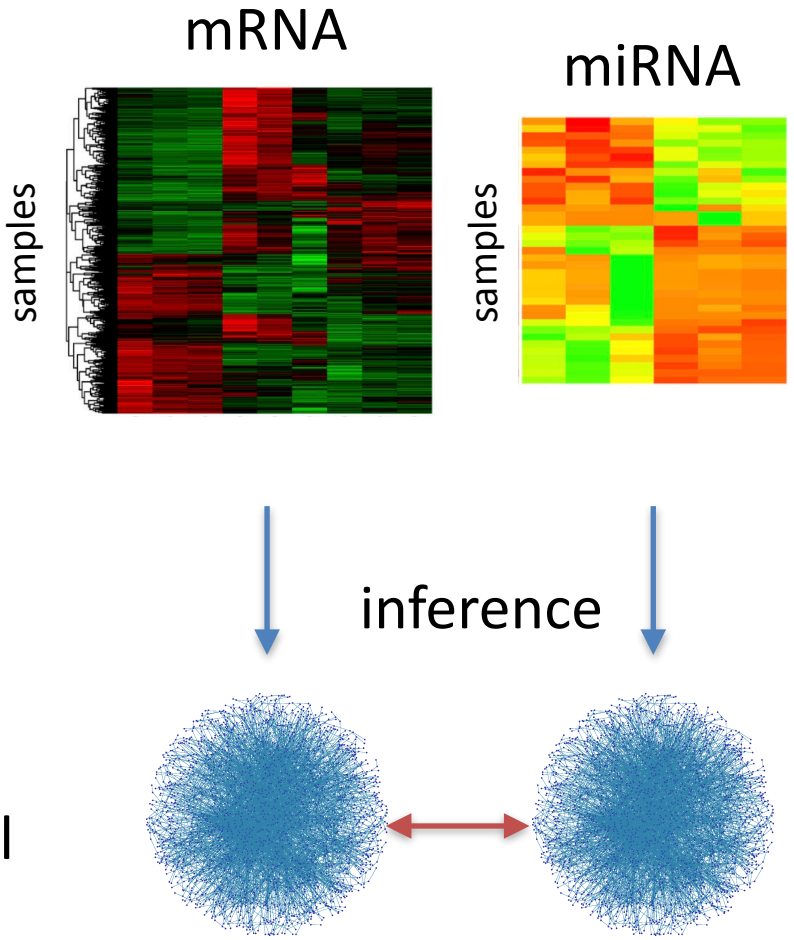
PPI	Complexes	Pathways	Correlation of expression
			
~60 000 edges	~40 000 edges	~250 000 edges	~1 400 000 edges
Measured networks		Curated networks	Inferred networks

The joint exploration of multiple networks is required => multilayer network framework

Matrices and Networks => 2 sets approaches

- Matrices
  - Works with raw data
  - Can better deal with high dimensionality
  - Information on both sample and feature-levels

- Networks
  - Interactions (ie direct access to the molecular processes)
  - Bipartites relationships (ie biological knowledge)

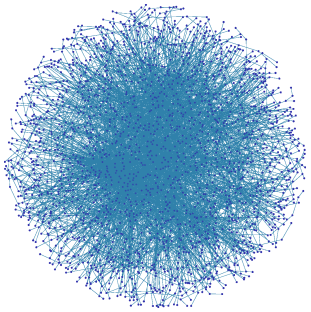


Challenges of multi-omics integration

- High-dimensionality -> Big data
- Heterogeneous variables
- Different ranges of variation
- Technical noise different for each omics

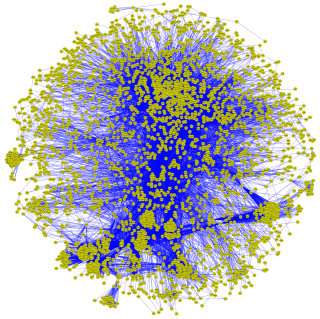
Many Biological Networks

PPI



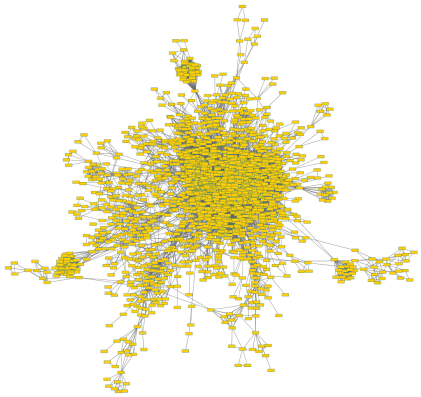
~60 000 edges

Pathways



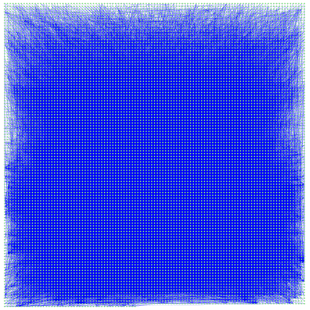
~250 000 edges

Complexes



~40 000 edges

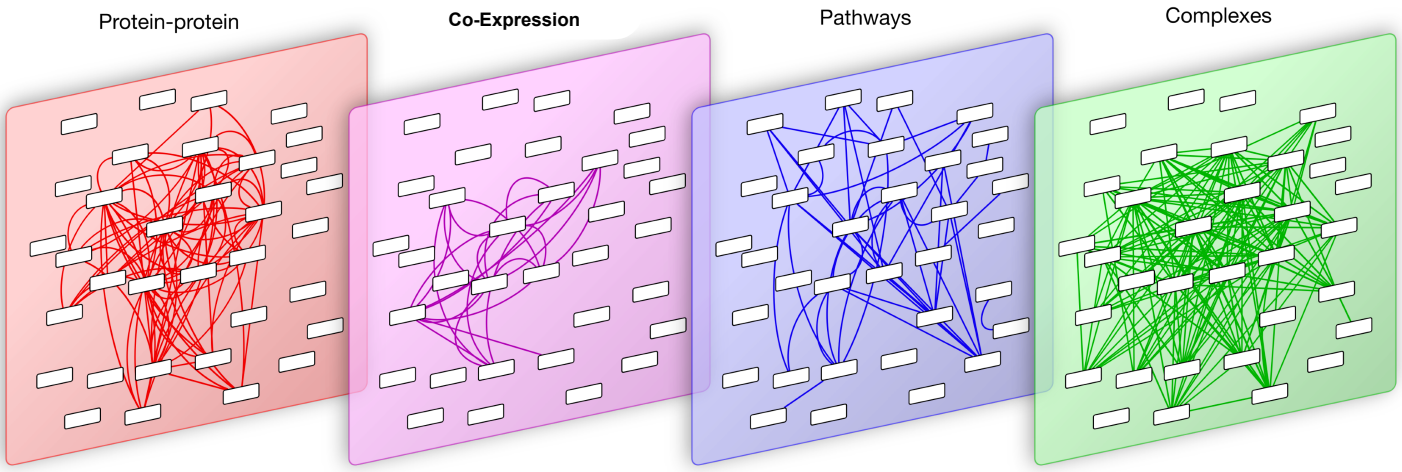
Correlation of expression



~1 400 000 edges

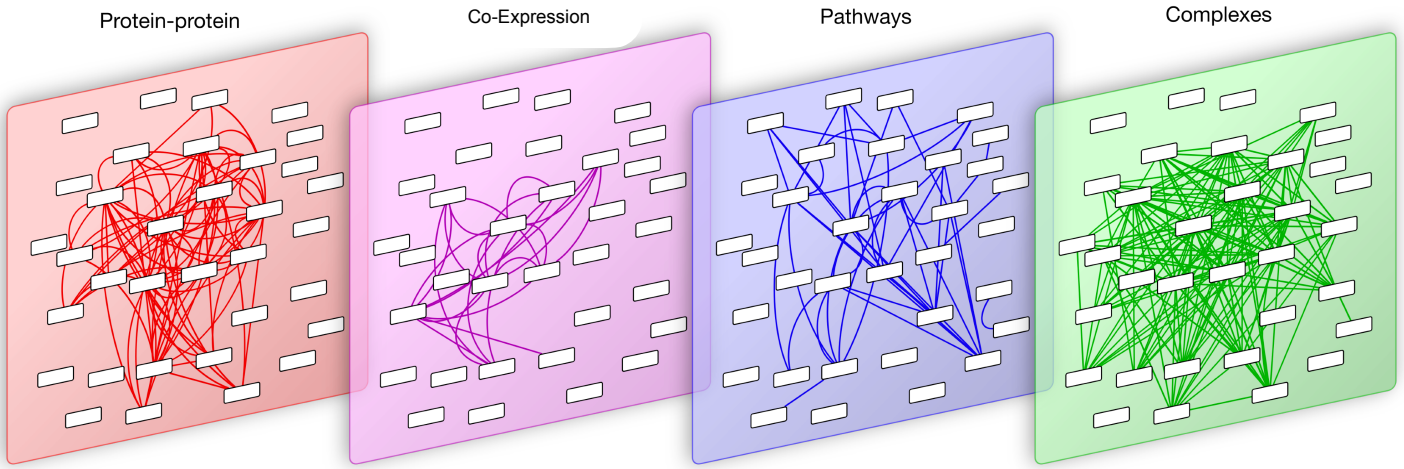
Huge quantity of information on gene/ protein cellular functions → How to extract this information ? → Multiplex framework

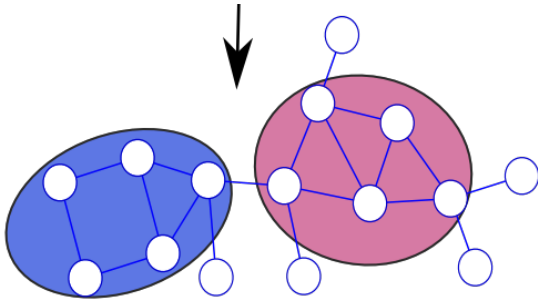
Multiplex Network framework and associated algorithms



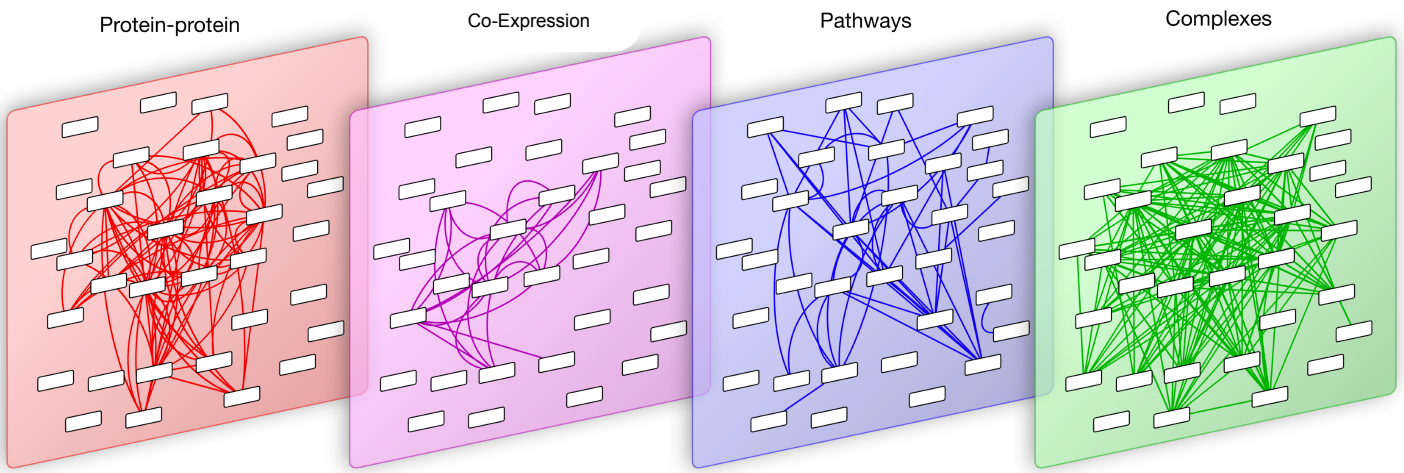
- Multiplex network framework: same nodes, different edges
- Different from merging the networks!
- Extending network analysis algorithms to multiplex networks
  - Clustering
  - Random Walk with Restart

Community identification from multiplex networks



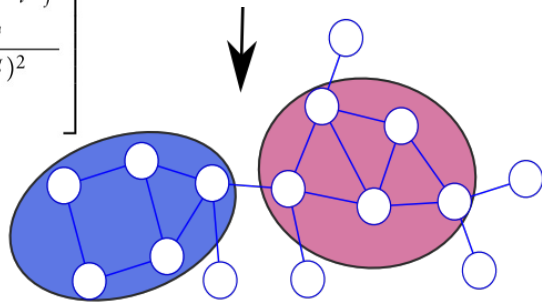


Community identification from multiplex networks



Multiplex-modularity

$$\mathcal{Q}^{M}((X^{(g)})_g, \mathbf{c}) = \sum_a \left[\sum_g \frac{\sum_{\{i,j\}} X_{i,j}^{(g)}}{2m^g} \right]_{i \neq j, \mathbf{c}_j = \mathbf{c}_j = a} - \sum_g \frac{\sum_{\{i,j\}} k_i^g k_j^g}{(2m^g)^2}_{i \neq j, \mathbf{c}_j = \mathbf{c}_j = a}$$



Didier et al. 2015

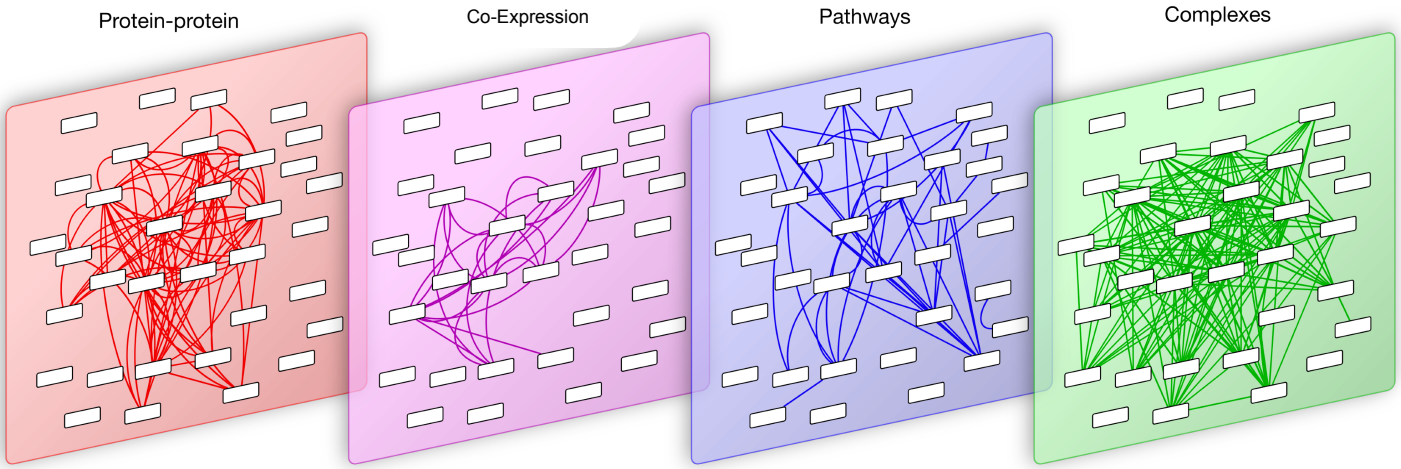
Didier et al. 2018

DREAM challenge Consortium paper. 2018

Consensus clustering

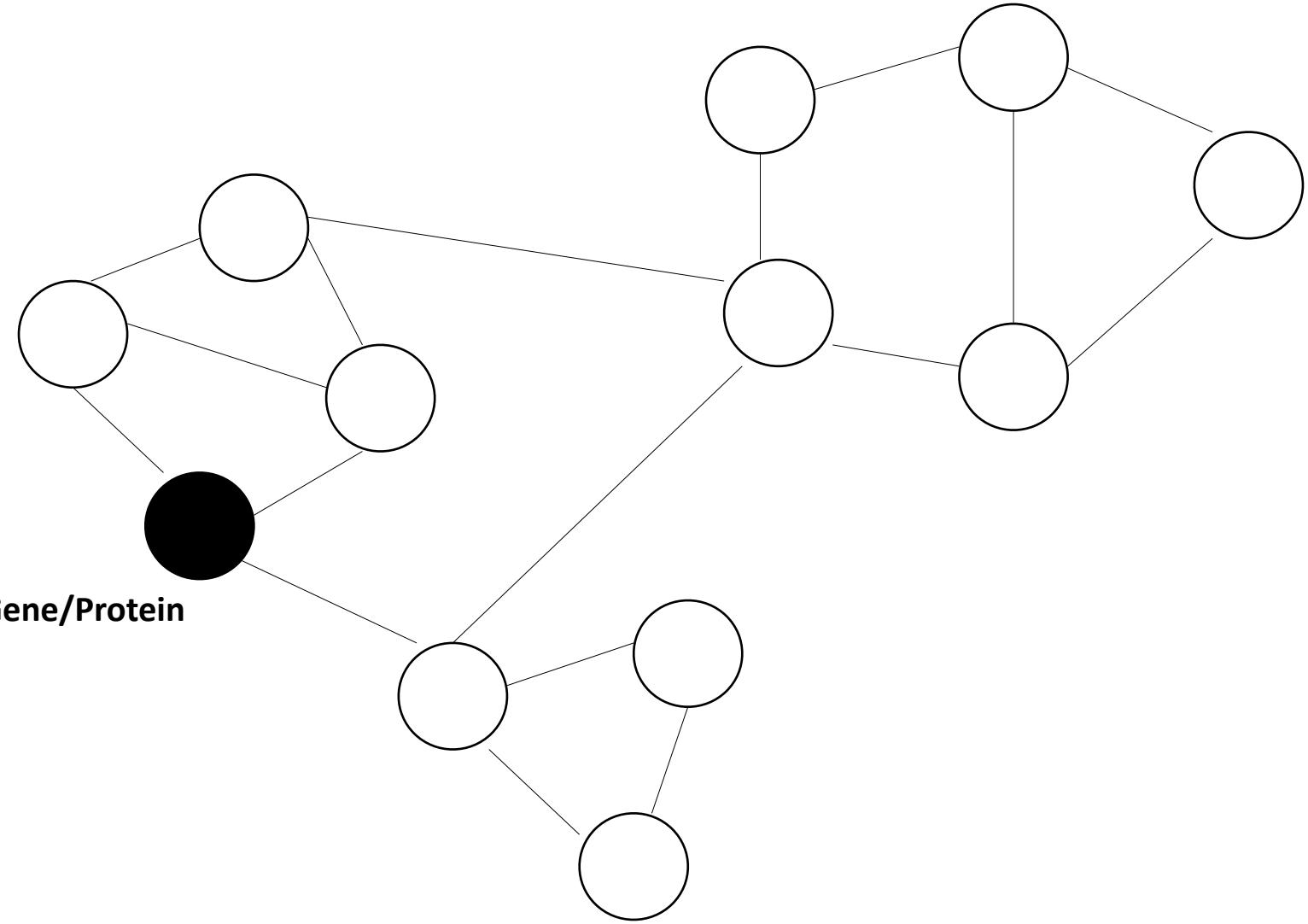
Cantini et al. 2015

Exploring multiplex networks

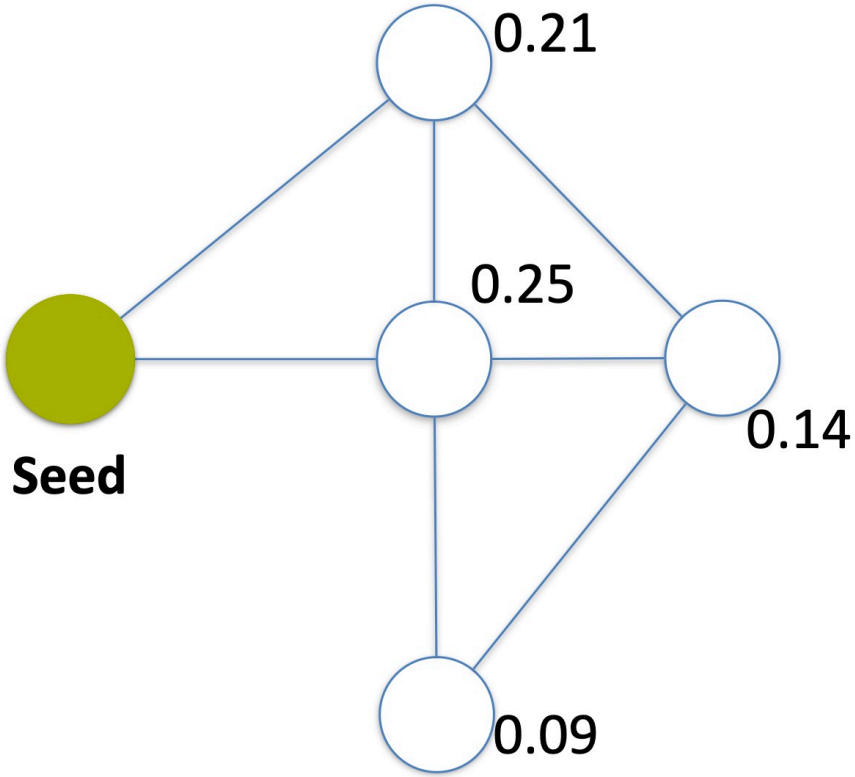


Random Walk with Restart

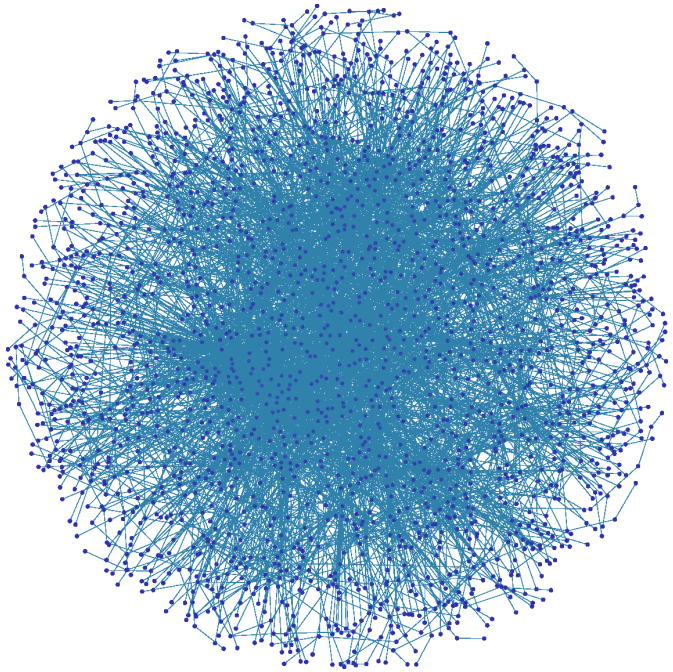
The Random Walk with Restart algorithm



The Random Walk with Restart algorithm

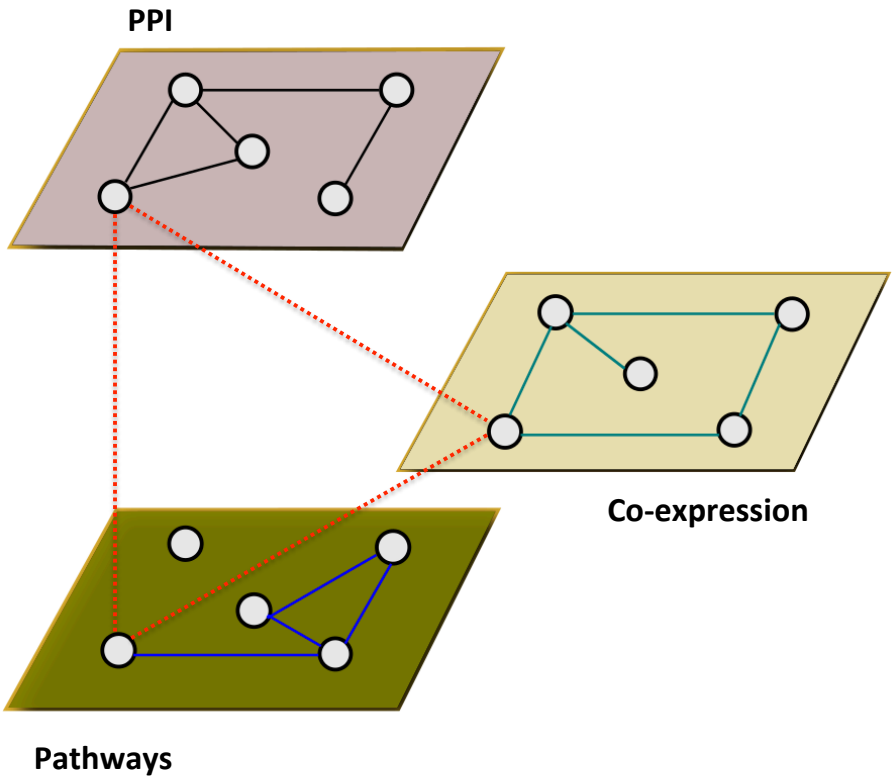


RWR score: Proximity/ pertinence *wrt* the seed

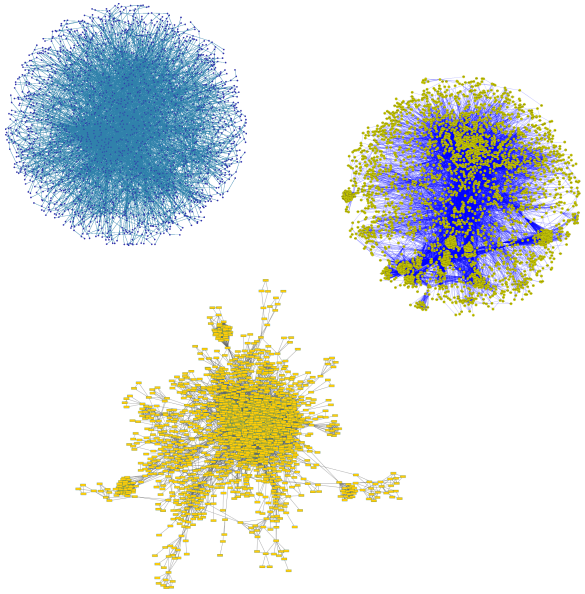


Random Walk with Restart on Multiplex Networks (RWR-M)

Alberto Valdeolivas



- Walk one layer
- Jump across layers

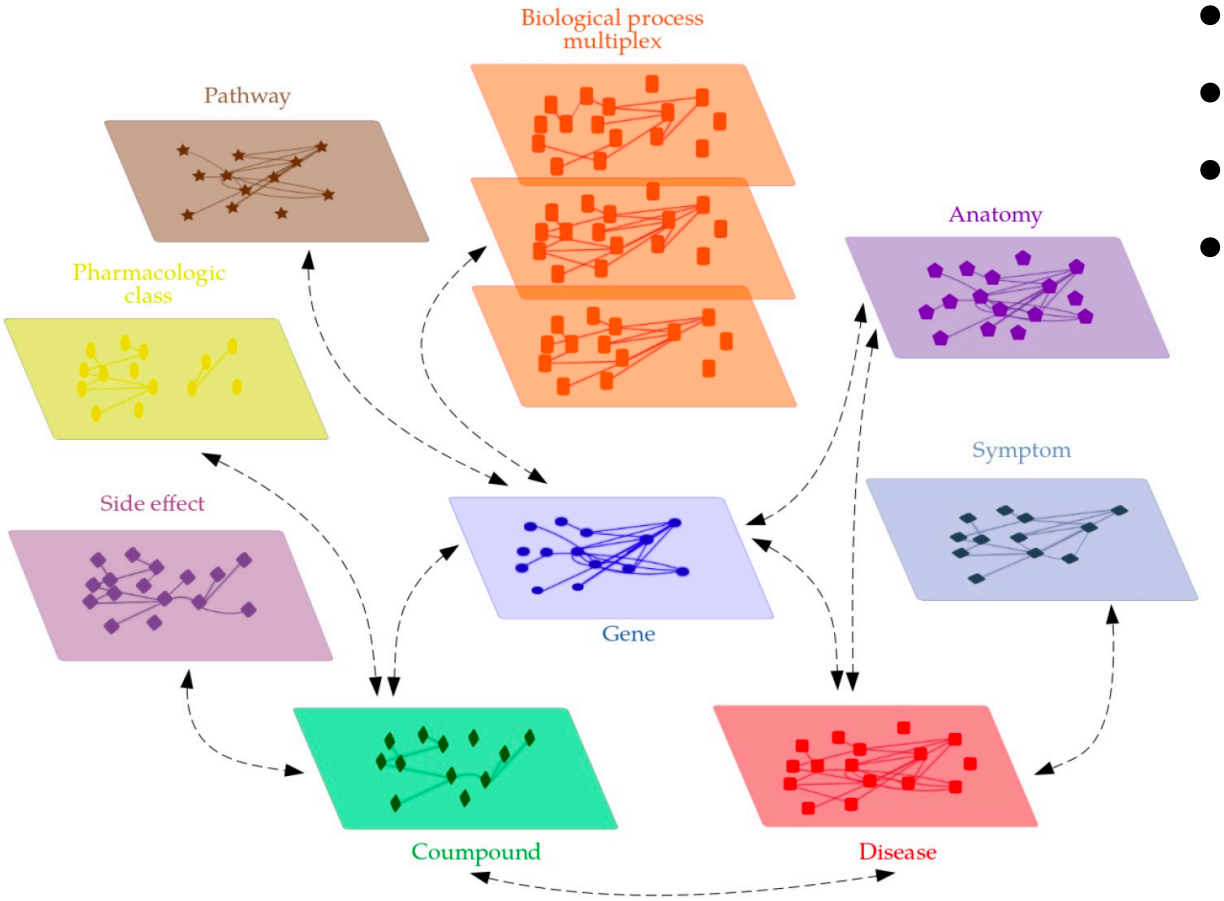


https://github.com/alberto-valdeolivas/RWR-MH

Bioconductor

Valdeolivas et al. Bioinformatics (2018)

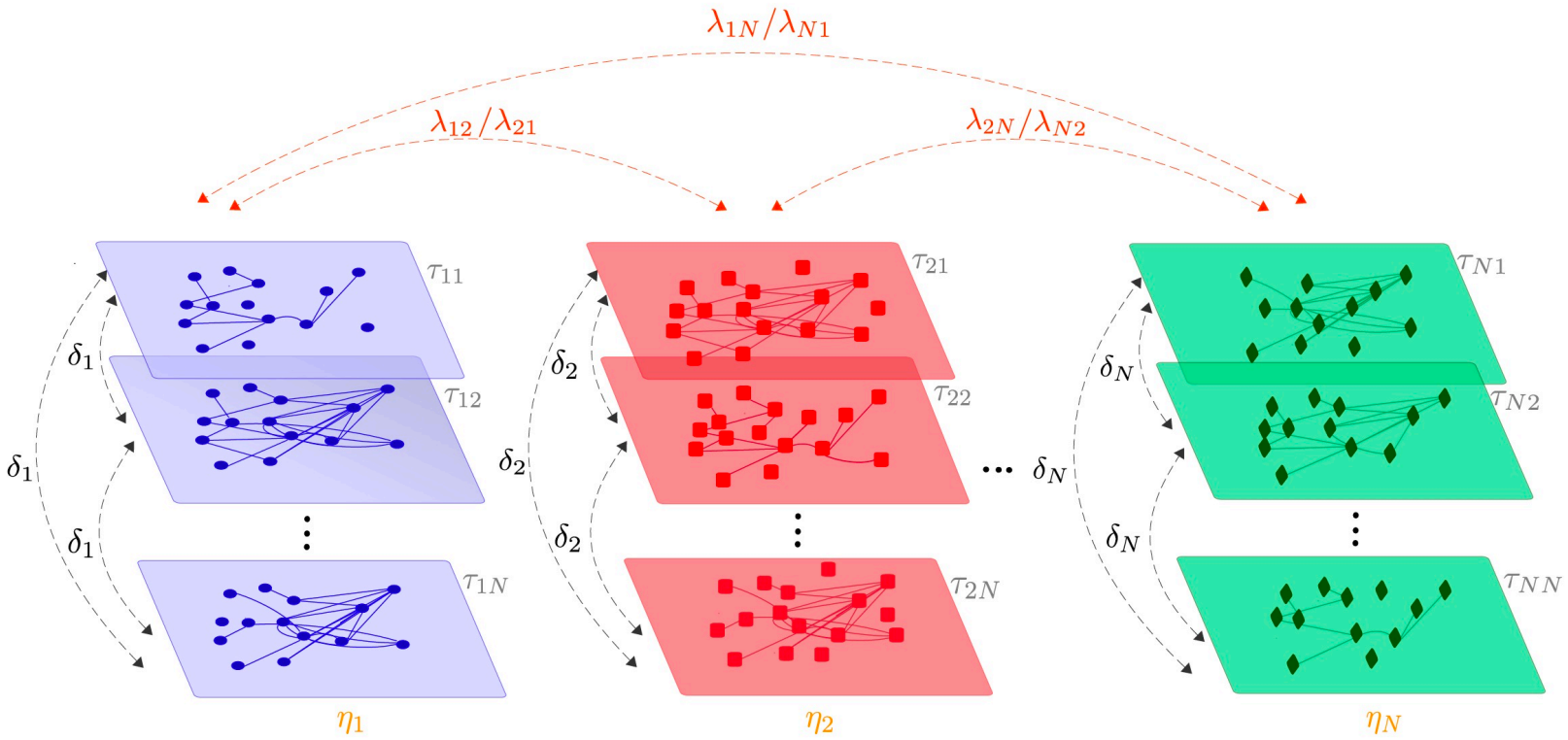
Extension of RWR to universal multilayer networks

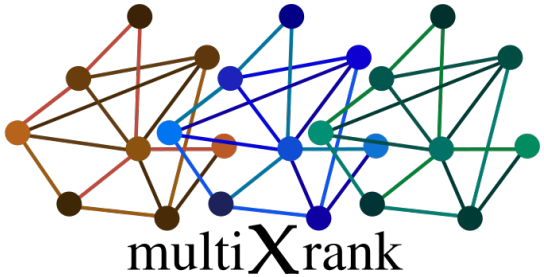


- Different nodes
- Different edges
- Bipartites
- Different multiplexes

Hetionet, adapted from Himmelstein et al. 2017

Extension of RWR to universal multilayer networks: algorithm





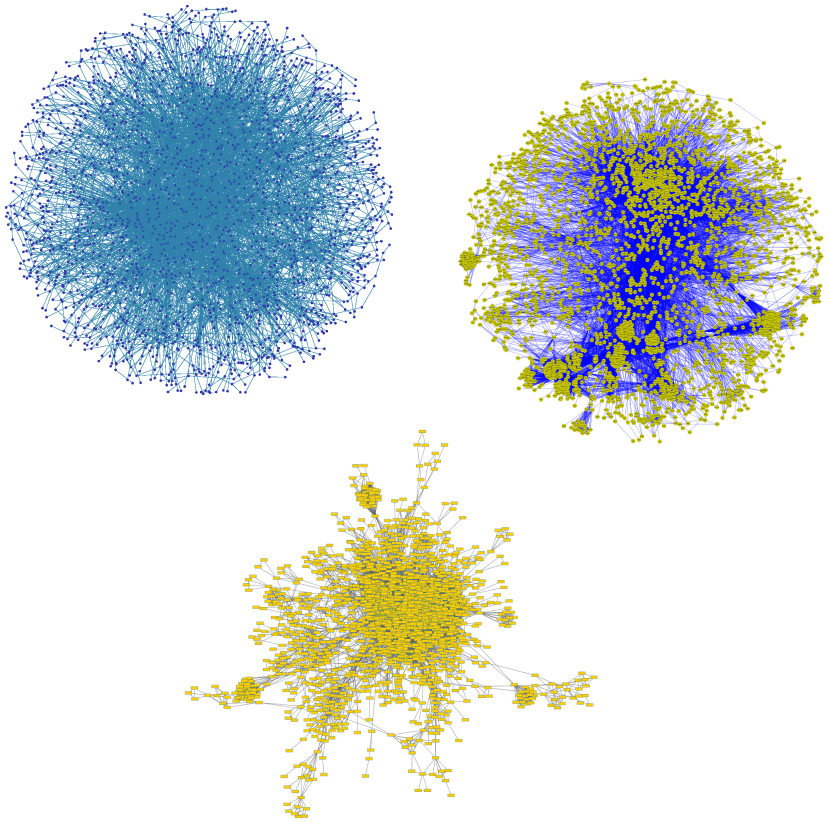
Anthony Baptista

Baptista et al. In prep

Random Walk with Restart on Multiplex networks in a nutshell

RWR output node scores for:

- Node prioritisation/ranking
- Subnetwork extraction
- Clustering
- Embedding
- …



Active module identification

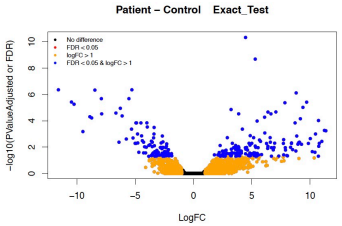
RNA-seq transcriptomics data

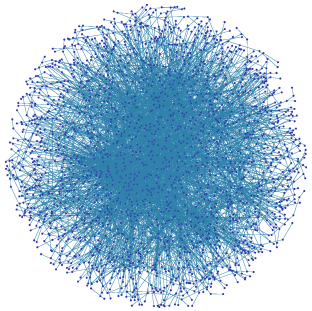
+

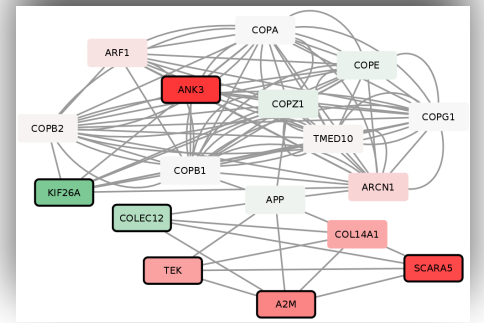
Biological Network

↓

Find “active” subnetworks







Algorithms: Greedy searches (PinnacleZ), Simulated Annealing (jActiveModules), Genetic Algorithms (COSINE) (Ideker et al. 2002, Chuang et al. 2007, Ma et al. 2011, Ozisik et al. 2017…)

Active module identification

- Few methods consider the density of interactions
- Methods are using only one (usually protein-protein) interaction networks

=> We propose a **Multi-Objective Genetic Algorithm** to identify active modules from Multiplex Networks

2 objectives to maximize

1

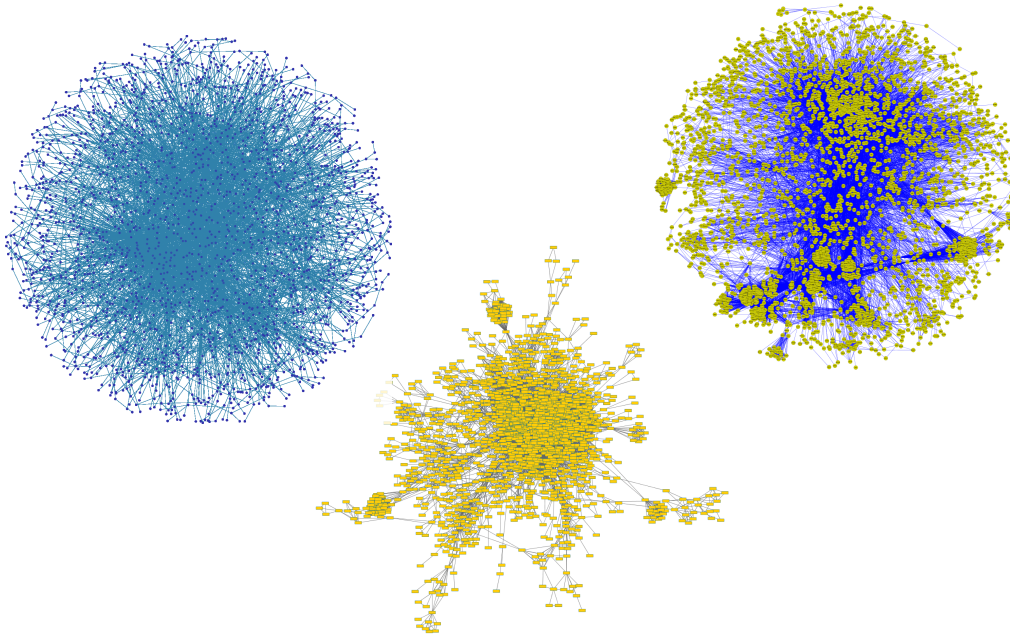
Average nodes score $\overline{NodesScore} = \frac{1}{n}\sum_{i=1}^{n}(Score_i^{norm})$

$$Score_i^{norm} = \frac{Score_i - min(Score)}{max(Score) - min(Score)} \qquad Score_i = \Phi^{-1}(1 - p_i)$$

2

Density

$$D_{norm} = \sum_{l=1}^{L} \frac{d_s}{d_l}$$



2 objectives to maximize

1

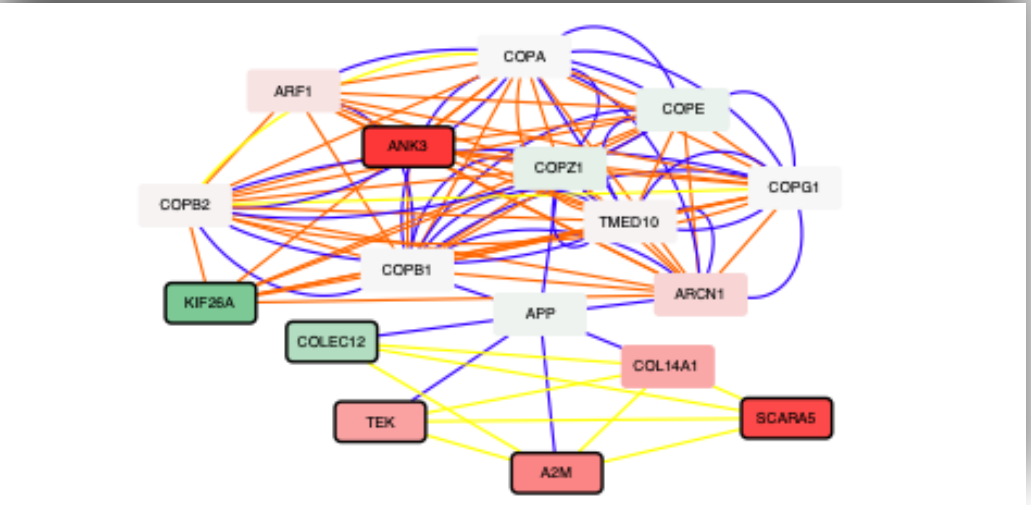
Average nodes score $\overline{NodesScore} = \frac{1}{n}\sum_{i=1}^{n}(Score_i^{norm})$

$$Score_i^{norm} = \frac{Score_i - min(Score)}{max(Score) - min(Score)} \qquad Score_i = \Phi^{-1}(1 - p_i)$$

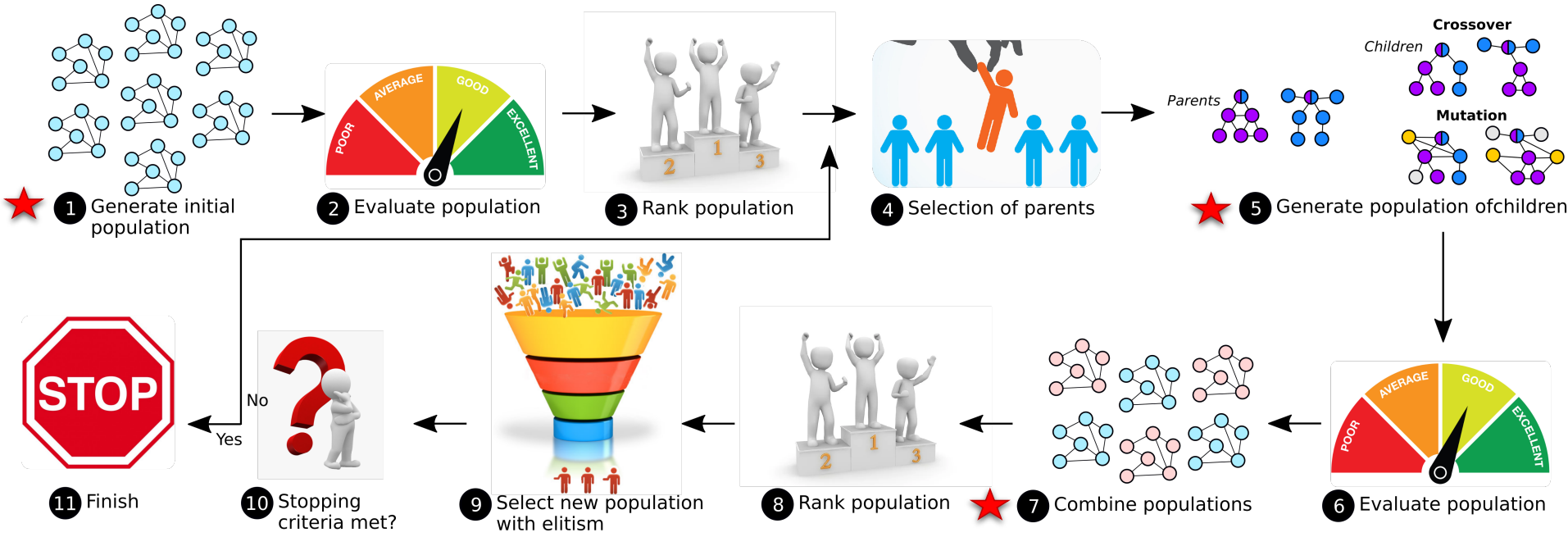
2

Density

$$D_{norm} = \sum_{l=1}^{L} \frac{d_s}{d_l}$$



Multi-Objectives Genetic Algorithm

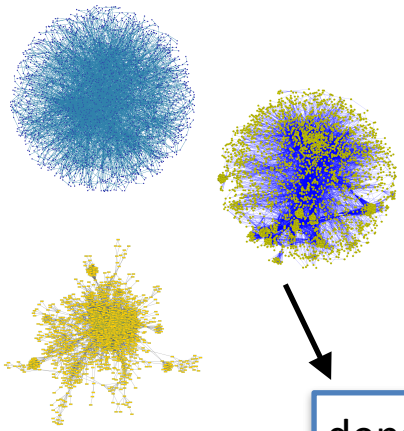


Elva Novoa

Based on NSGA-II (Deb et al., 2002)

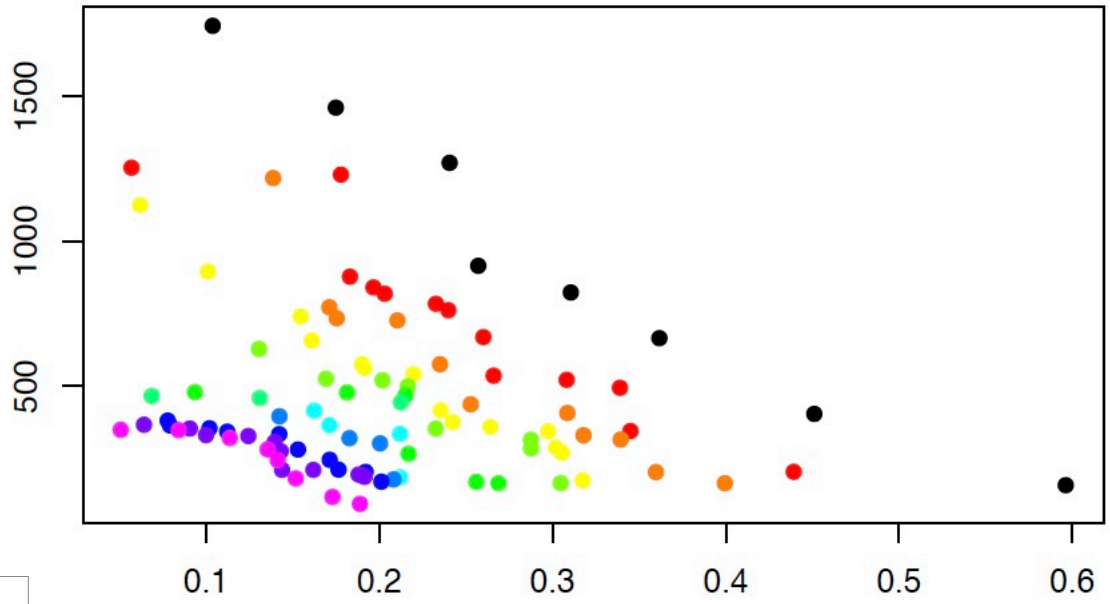
MOGAMUN

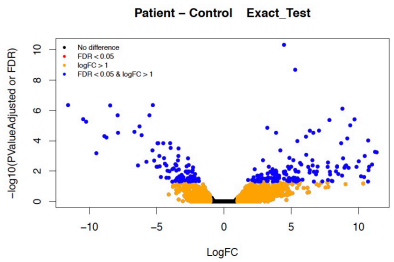
Multi-objective Genetic Algorithm to find Active Modules in Multiplex Biological Networks



density of interactions from the multiplex network

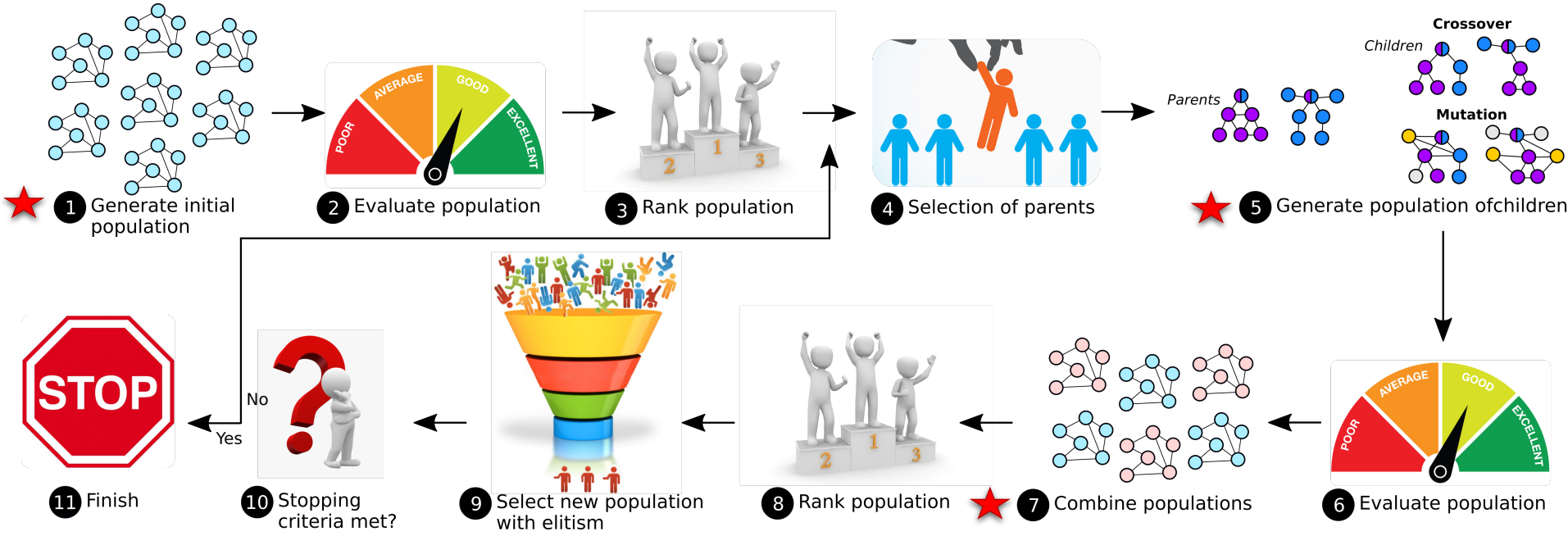
Maximize 2 objectives





Differential expression (average node-scores)

Multi-Objectives Genetic Algorithm



Elva Novoa

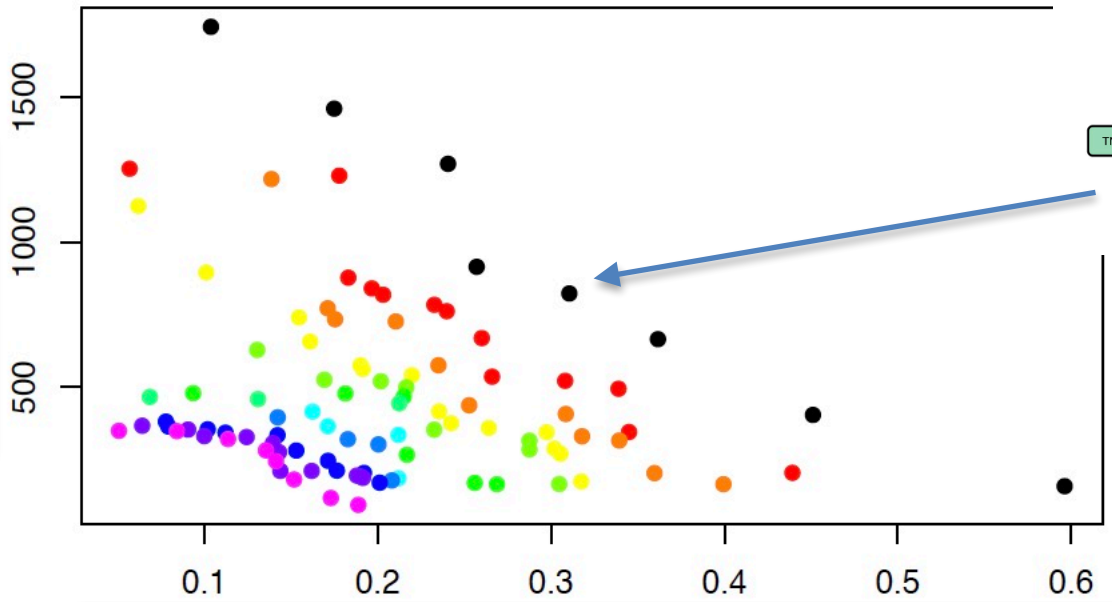
Based on NSGA-II (Deb et al., 2002)

MOGAMUN

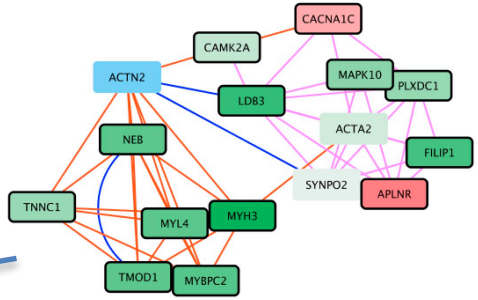
Multi-objective Genetic Algorithm to find Active Modules in Multiplex Biological Networks

MOGAMUN

density of interactions from the multiplex network



Differential expression (average node-scores)

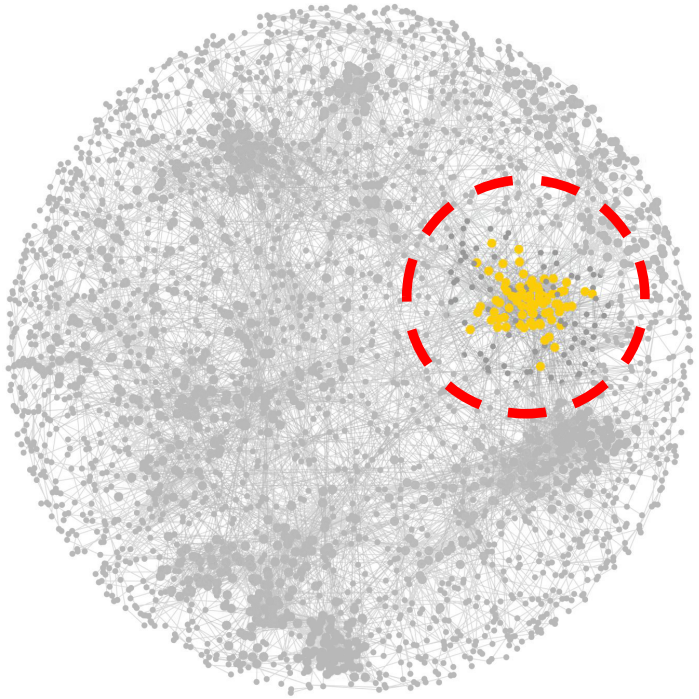


Elva Novoa

Novoa et al. BioRxiv, 2020
Bioconductor
https://github.com/elvanov/MOGAMUN

Evaluations - Comparison with other state-of-the-art methods (COSINE, PinnacleZ, jActiveModules)

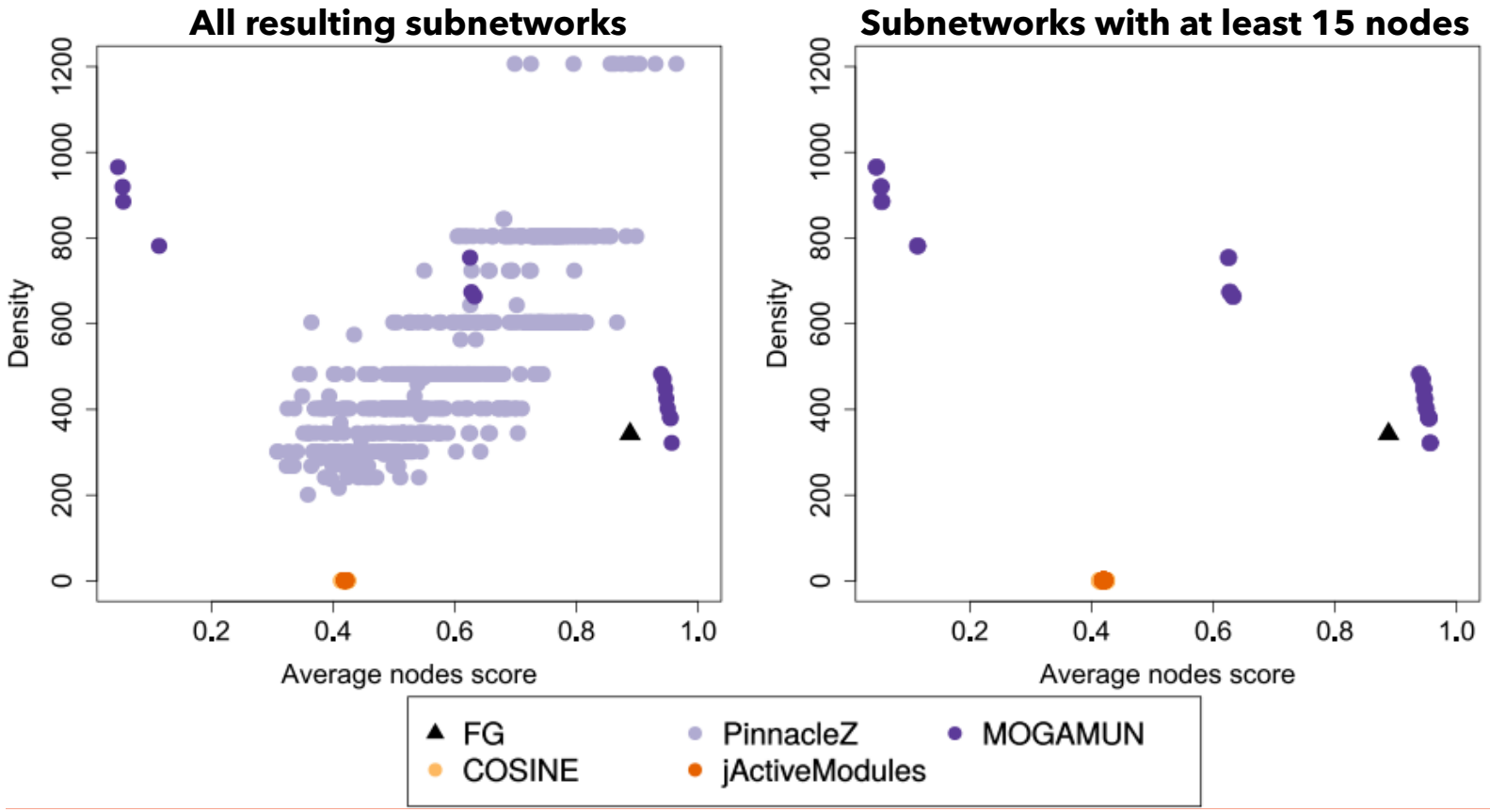
- 20 foreground genes (connected)
- Artificial expression data generated to highlight the foreground genes



	Network	|Nodes|	|Edges|	Density	Data	Cases	Controls	|DEG|
Scenario 1	PPI_1	9,425	36,811	8.28x10^(-4)	Sim_normal	100	10	483
Scenario 2	PPI_2	12,621	66,971	8.41x10^(-4)	Samp_TCGA	1,102	112	20

Batra et al. 2017

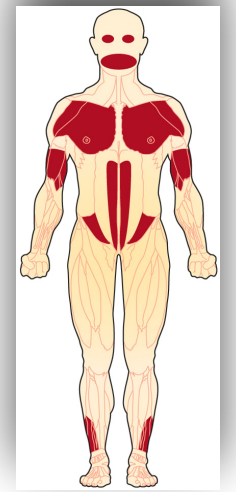
Comparison with state-of-the-art methods, Scenario 1



F1-scores close to 0 for all methods and > 0.4 for MOGAMUN

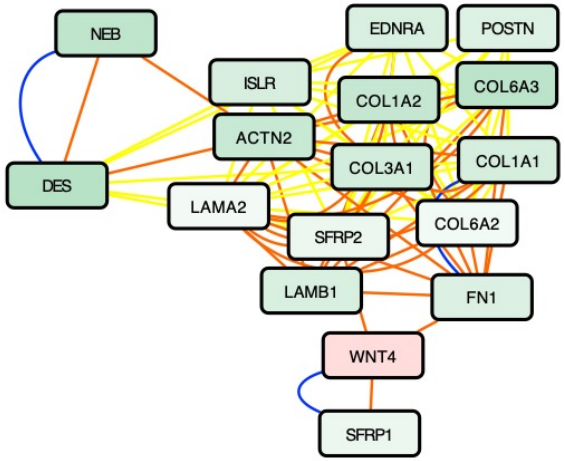
Active modules in Facio Scapulo Humeral Dystrophy

MOGAMUN



FSHD1 => high genetic complexity (hypomethylation of a repeated genomic region), causative pathways unknown

=> RNA-seq in muscle fibers derived from patient iPSCs



Collab. F. Magdinier

Laberthonière et al. in prep

Anthony Baptista
Judith Lambert
Ozan Ozisik
Morgane Terezol
Nadine Ben Boina

Alumni
Pooya Zakeri
Alberto Valdeolivas
Elva Novoa

Alfonso Valencia

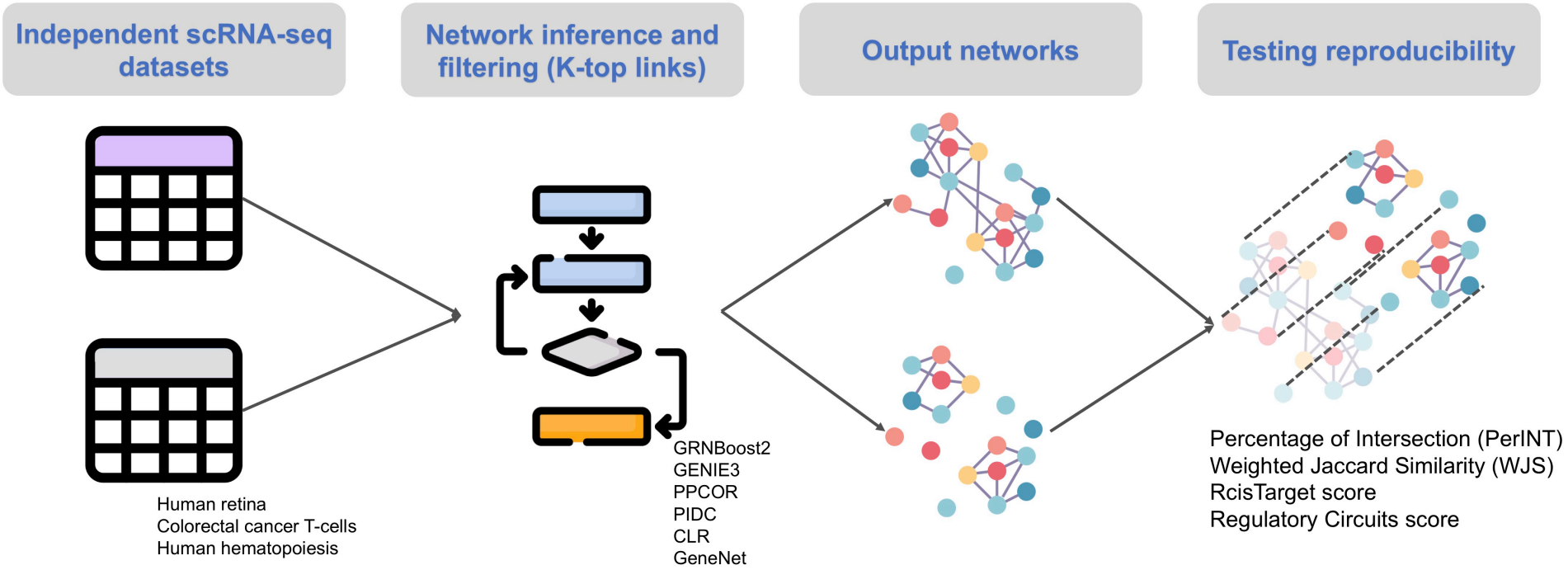
Pierre Cau
Claire Navarro
Sophie Perrin

Nicolas Lévy
Frédérique Magdinier
Marc Bartoli

Elisabeth Remy
Laurent Tichit
Léo Pio-Lopez

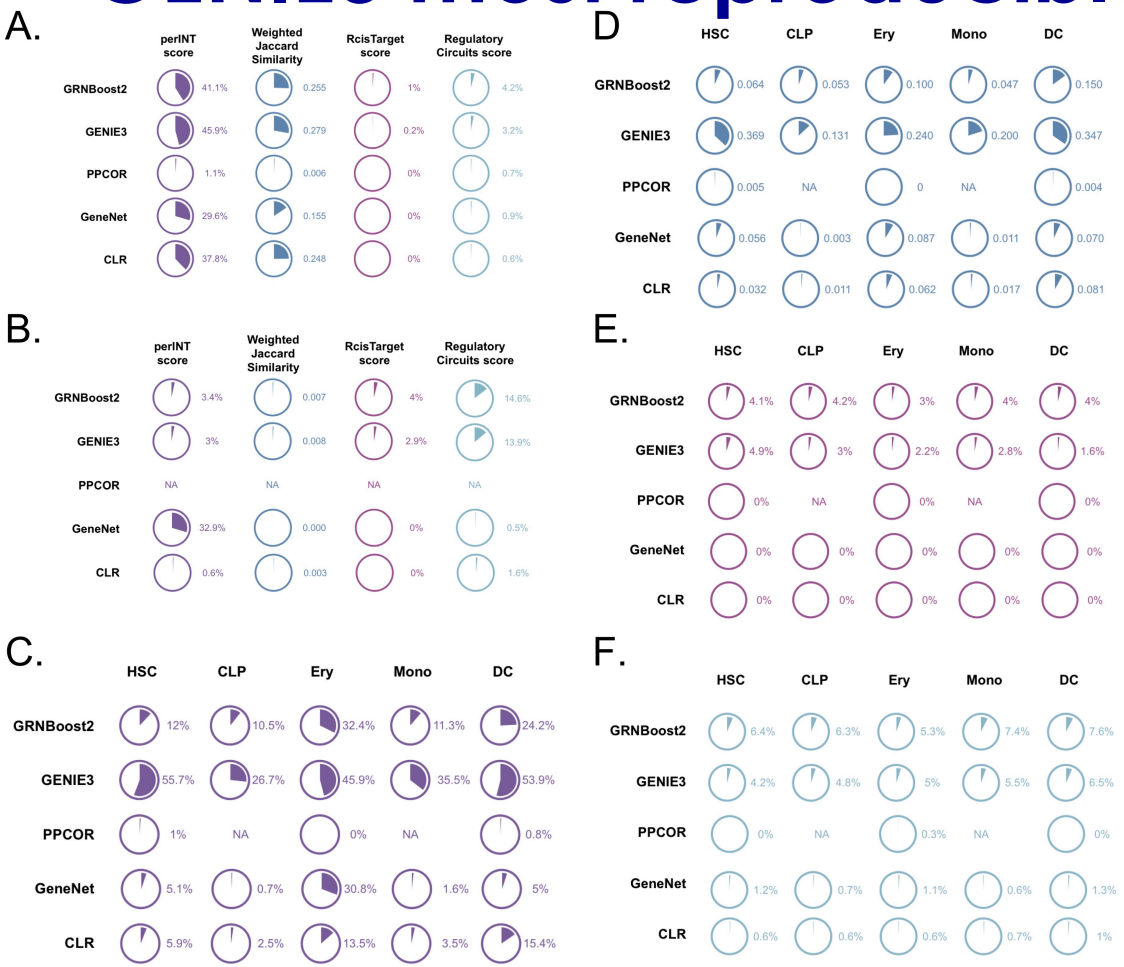
Laura Cantini

Evaluating the reproducibilty of network inference in single-cell



Kang, Yoonjee, Denis Thieffry, and Laura Cantini. "Evaluating the reproducibility of single-cell gene regulatory network inference algorithms." *Frontiers in genetics* 12 (2021): 362.

GENIE3 most reproducible algorithm



- GENIE3 results to be the most reproducible algorithm
- GENIE3 and GRNBoost2 show higher intersection with ground-truth biological interactions.
- Results are independent from:
  - the single-cell sequencing