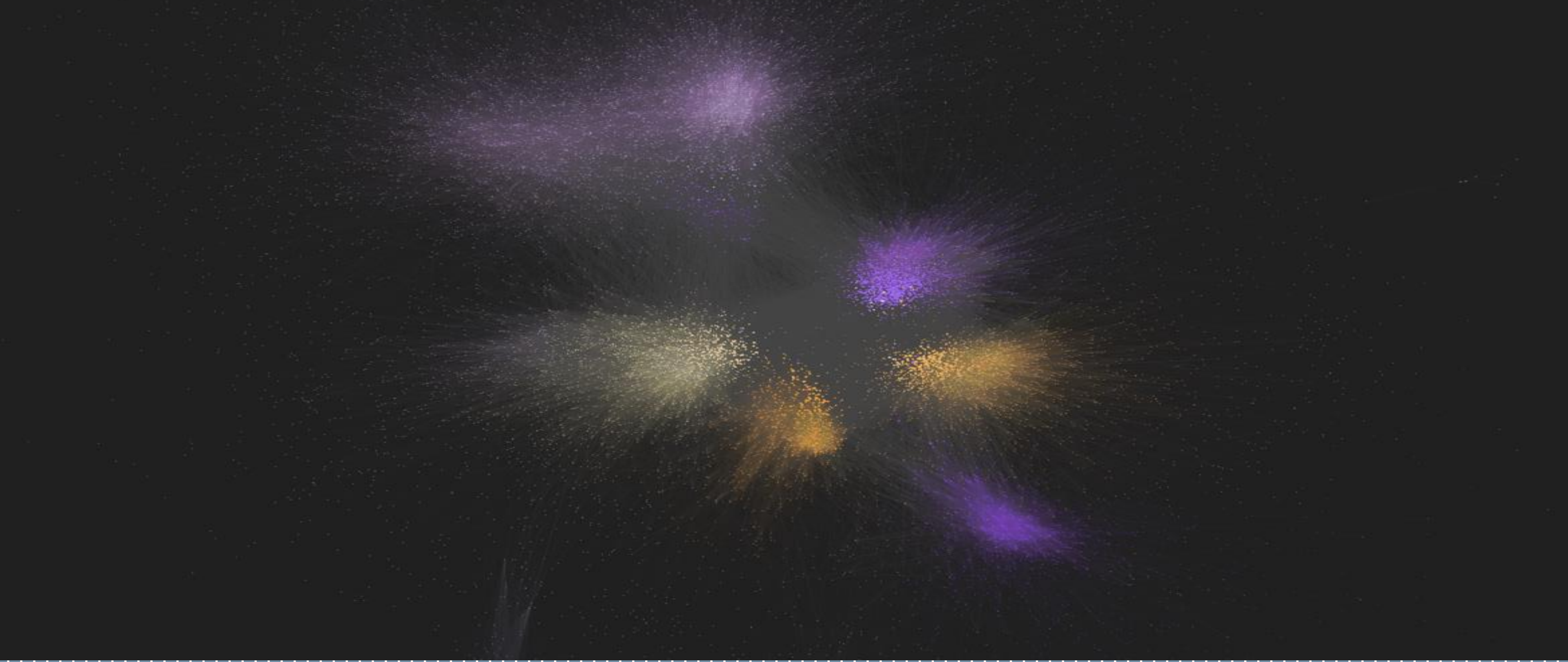
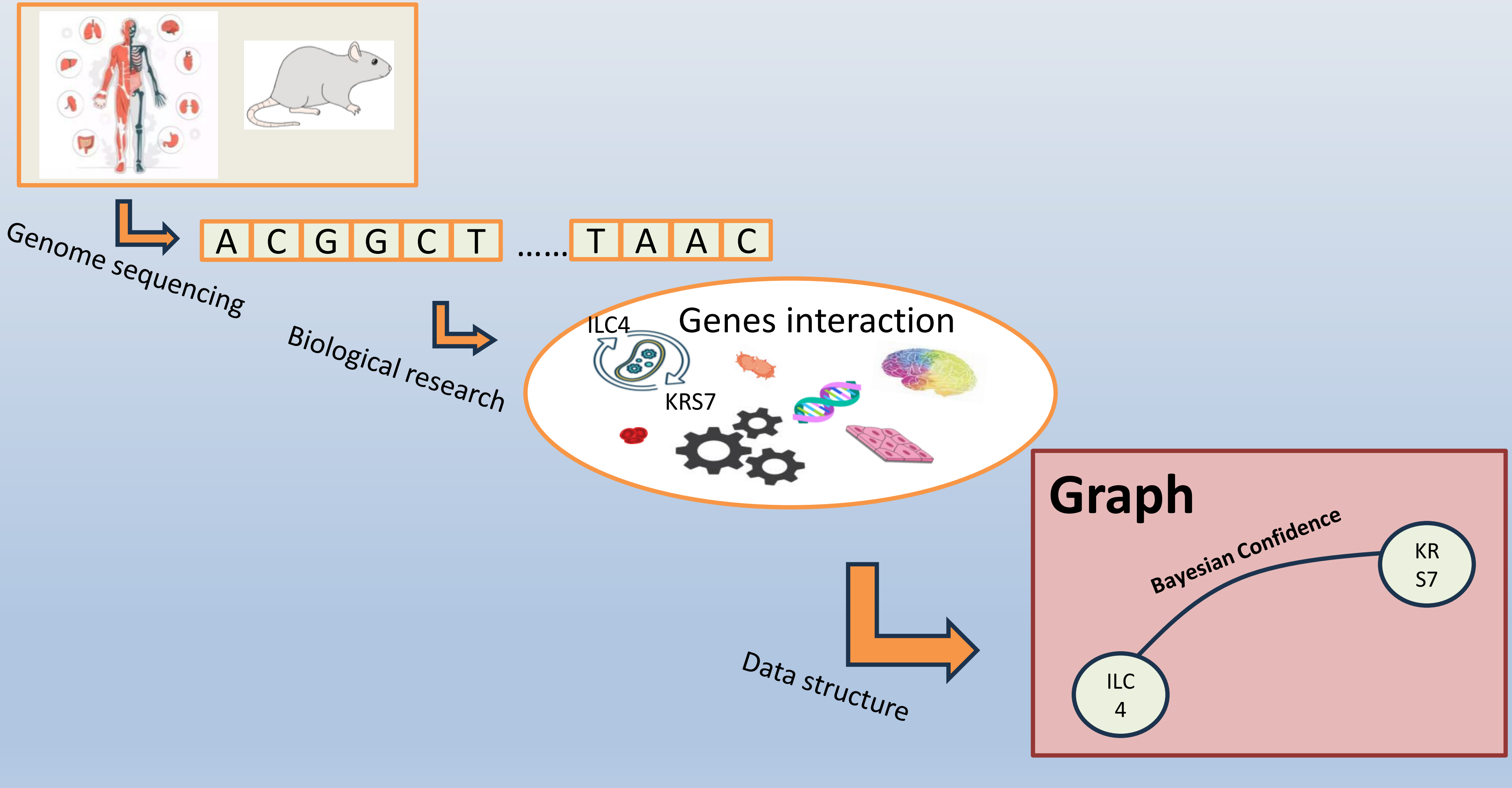


# Interactome network analysis

Giuseppe Gentile

Course: COMPLESSITÁ NEI SISTEMI E NELLE RETI

# Origin of the network



Genome sequencing

A C G G C T ...... T A A C

Biological research

Genes interaction

ILC4

KRS7

Data structure

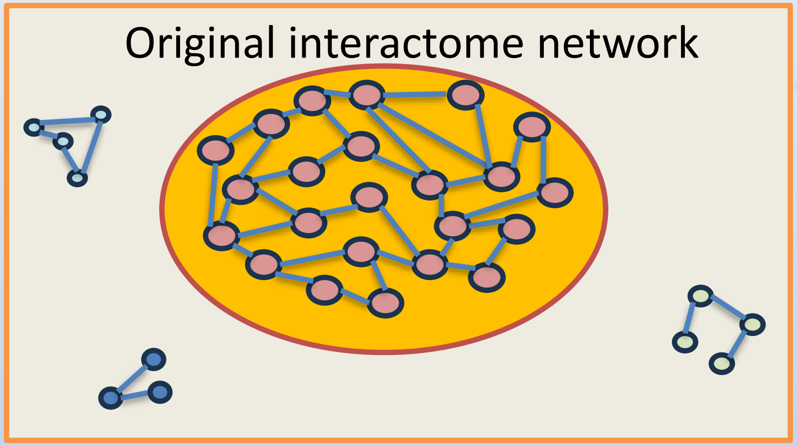
**Graph**

Bayesian Confidence

KR S7

ILC 4

# Network preprocessing



Not connected

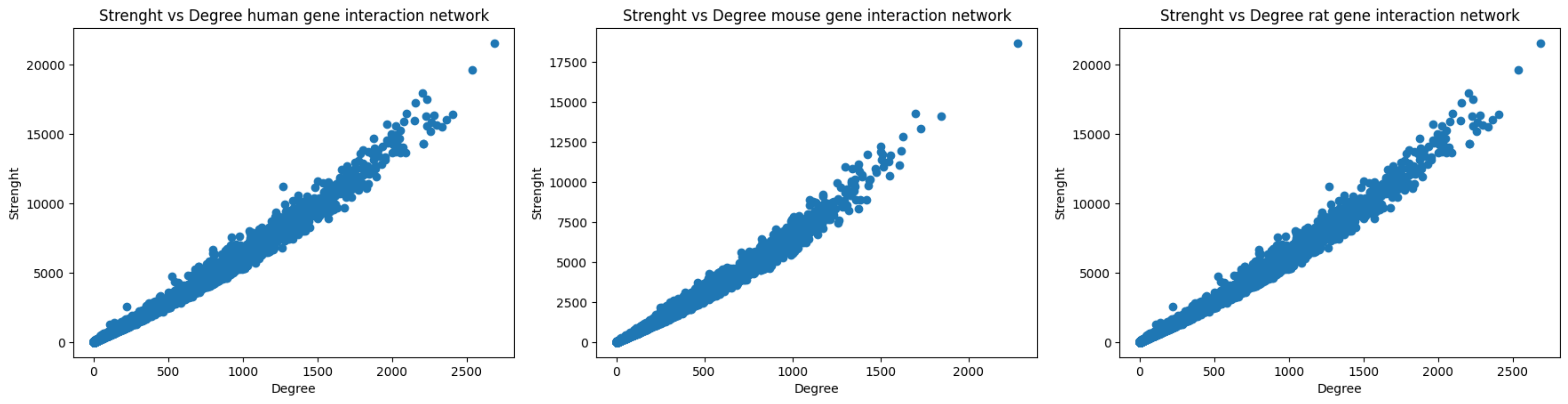
| Genome | Nodes | Edges | Avg degree |
|---|---|---|---|
| **Human** | 15796 | 1.387e+6 | 175.871 |
| **Rat** | 14360 | 9.235e+5 | 166.965 |
| **Mouse** | 10833 | 9.042e+5 | 128.647 |

**Considering only the one big connected component**

| Genome | Nodes | Edges | Avg degree | Nodes kept | Edges kept |
|---|---|---|---|---|---|
| **Human** | 15571 | 1.386e+6 | 178.063 | **99.98752%** | **98.576%** |
| **Rat** | 14300 | 9.235e+5 | 167.349 | **99.996102%** | **99.582%** |
| **Mouse** | 10806 | 9.041e+5 | 129.164 | **99.998341%** | **99.571%** |

# Network’s statistic

Strenght and degree are linearly correlated with a factor of ≈ 6


Strenght vs Degree human gene interaction network

Strenght vs Degree mouse gene interaction network

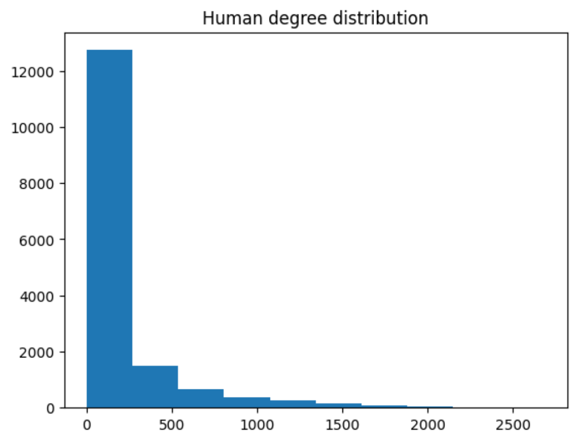
Strenght vs Degree rat gene interaction network

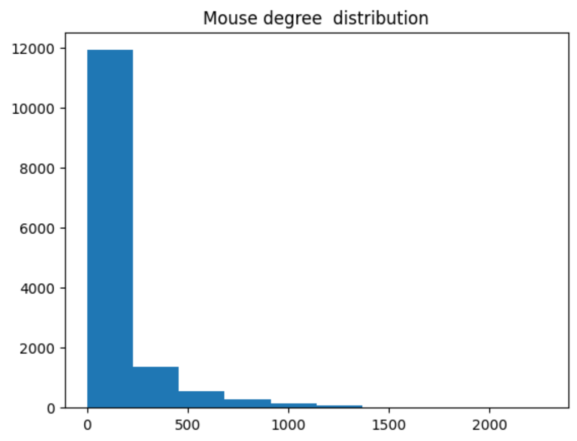
If a **gene** has **lots of links**, those links are also **more likely** to be **experimentally** true (in the next slides this concept will be more clear)

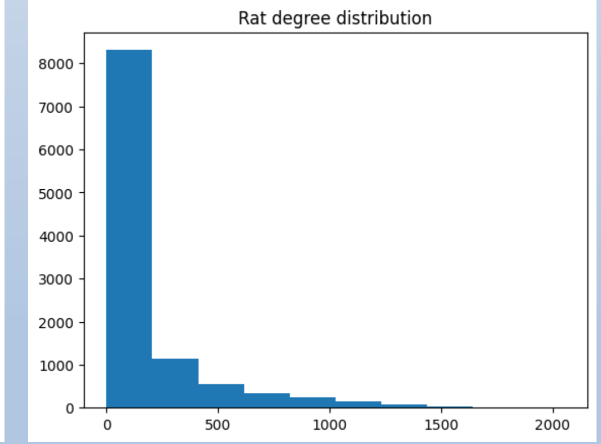
# Network’s statistic

| Genome | Density |
|---|---|
| HUMAN | 0.008961 |
| RAT | 0.011155 |
| MOUSE | 0.015417 |

**Sparse network**

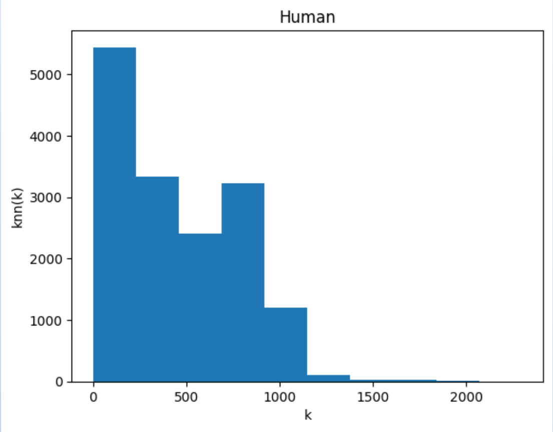


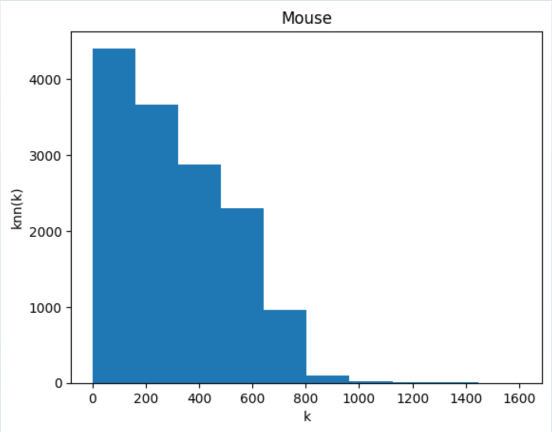


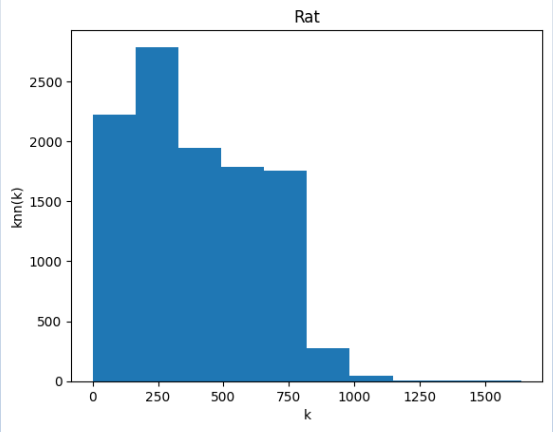


# Disassortative network

**High-degree nodes** (hubs) connected **to low-degree** nodes

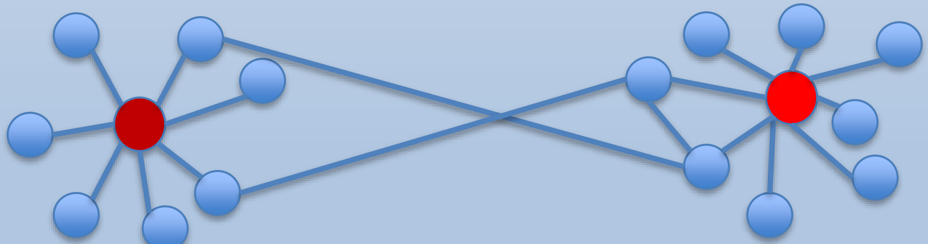






No dense core of interconnected hubs, instead, **hubs are isolated:** each **important genes for a pathway** is **independent from other** important regulatory genes

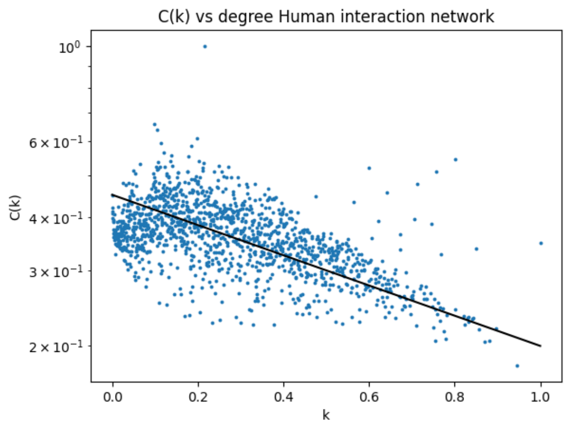
$$k_{nn} =< k > + \frac{\sigma^2}{< k >}$$

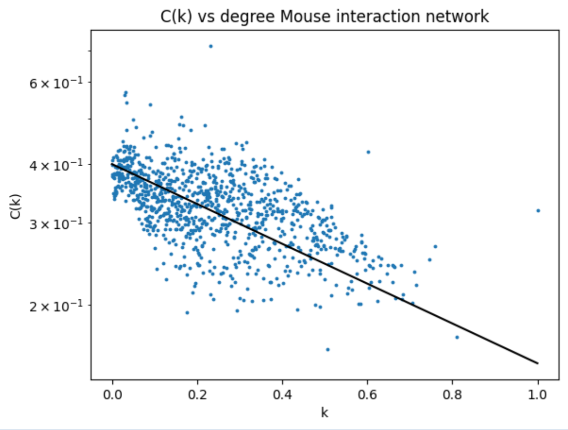


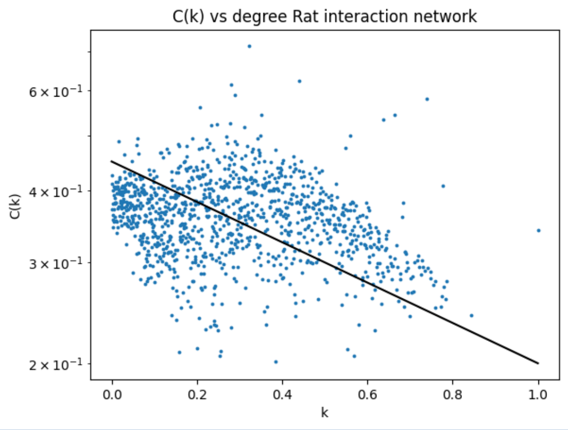
**Pathways are instrinsecally dependent due to the network structure, but important genes don’t interact with others pathway’s leaders**

# Disassortative network

**Modular Organization**: The network's modular organization is reflected in the clustering behavior. Modules (biological pathways) have high internal clustering, but the connectivity between different modules is less dense, particularly through hub genes.

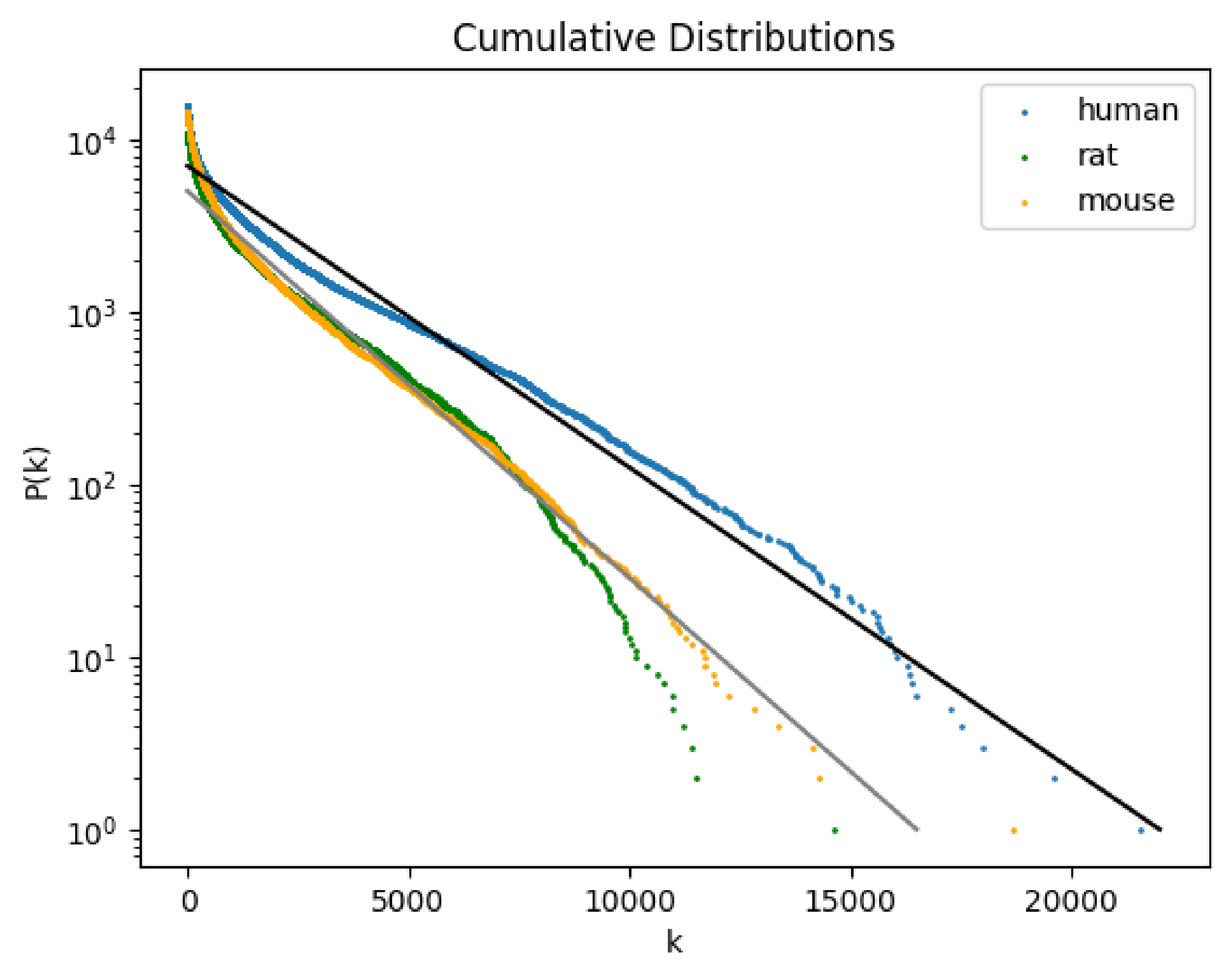






**More interactions within pathway than between pathways.**
Genes involved in the same pathway (forming cliques) are more likely to be connected, while those connecting different pathways (hubs) decrease the overall clustering coefficient as their degree increases.

# Scale free: preferential attachment



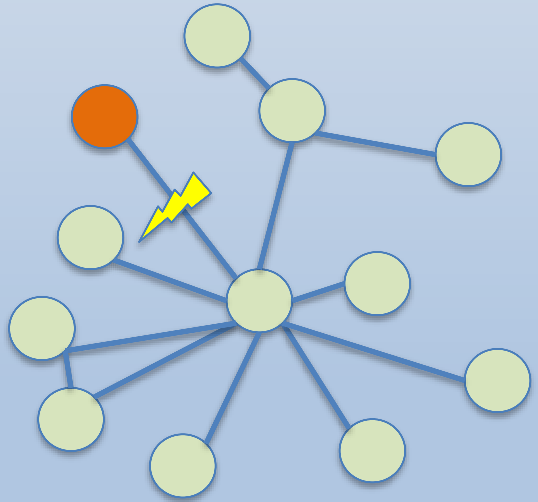
$$P(k) = k^{-\gamma}$$

| Network | $\gamma$ |
|---|---|
| Human | 5,28 |
| Rat | 12,51 |
| Mouse | 6,53 |

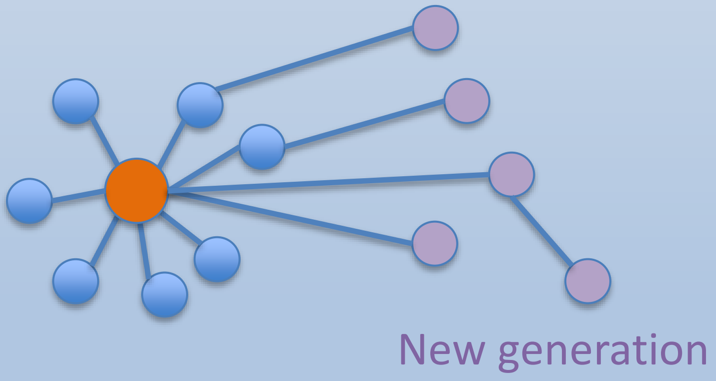
Human genome has more hubs (more primary biological functions)

# Origin of preferential attachment

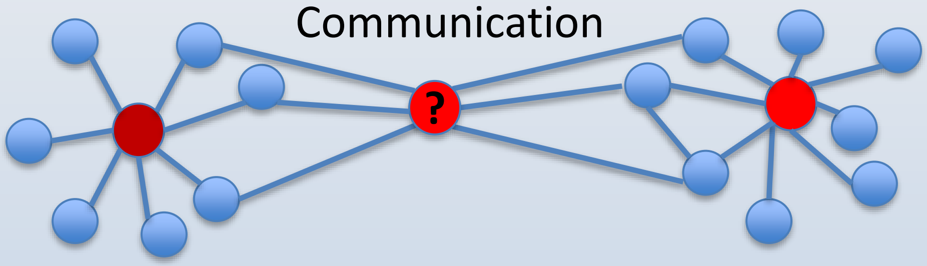
**Robust** to both **mutations** and **deletion.** (not to attacks)
Hub genes are essential for survival, so their preservation is vital: mutations on hubs are mutation that cause changes to overall network (diseases).



**Evolution** of the genome **involves** the **duplication of genes**. When a gene duplicates, its interactions with other proteins are retained, maintaining the network's connectivity. Over time **duplicates acquire new connections**, contributing to the scale-free nature.

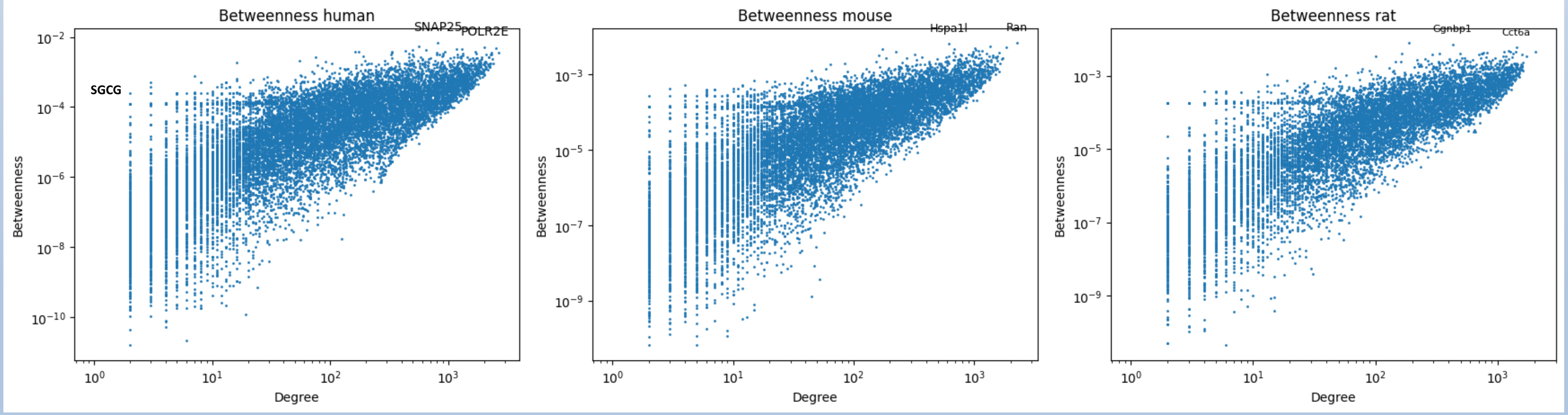


# Betweenness centrality

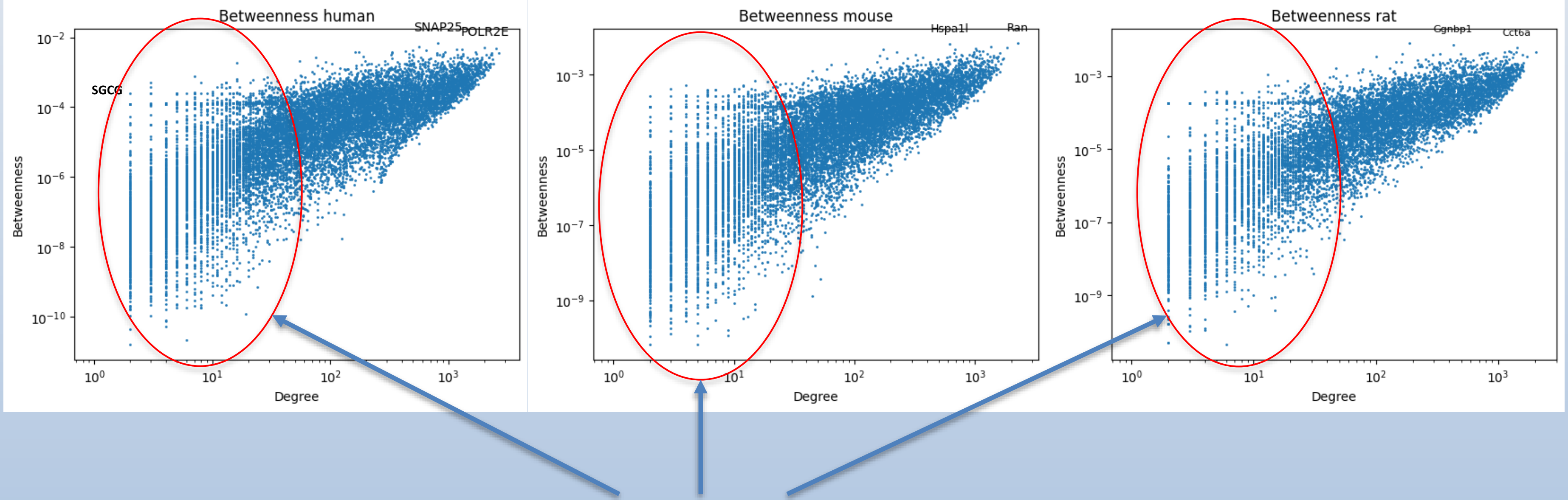


Communication

$$b_i = \sum_{j,k} \frac{shortest\ path\ from\ j\ to\ k, via\ i}{shortest\ path\ from\ j\ to\ k}$$



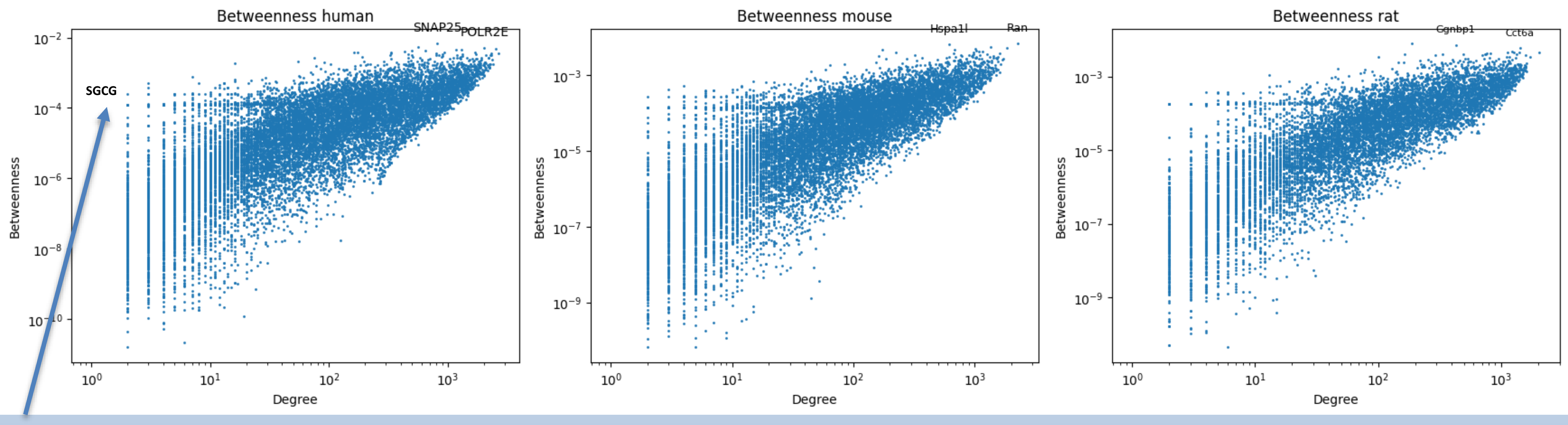
# Betweenness centrality

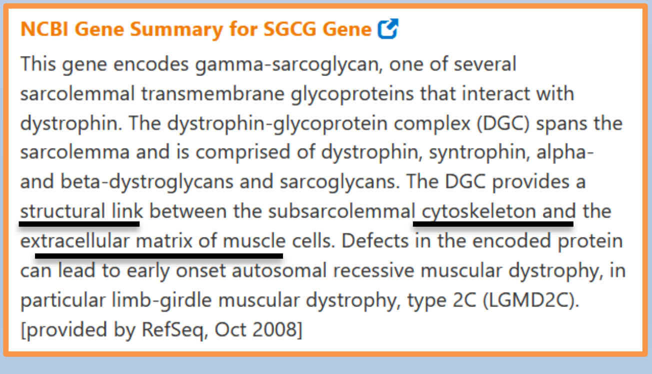


**Low degrees** nodes have **high variation of betweenness** :

**scale free**: most of the nodes have low degree

# Betweenness centrality



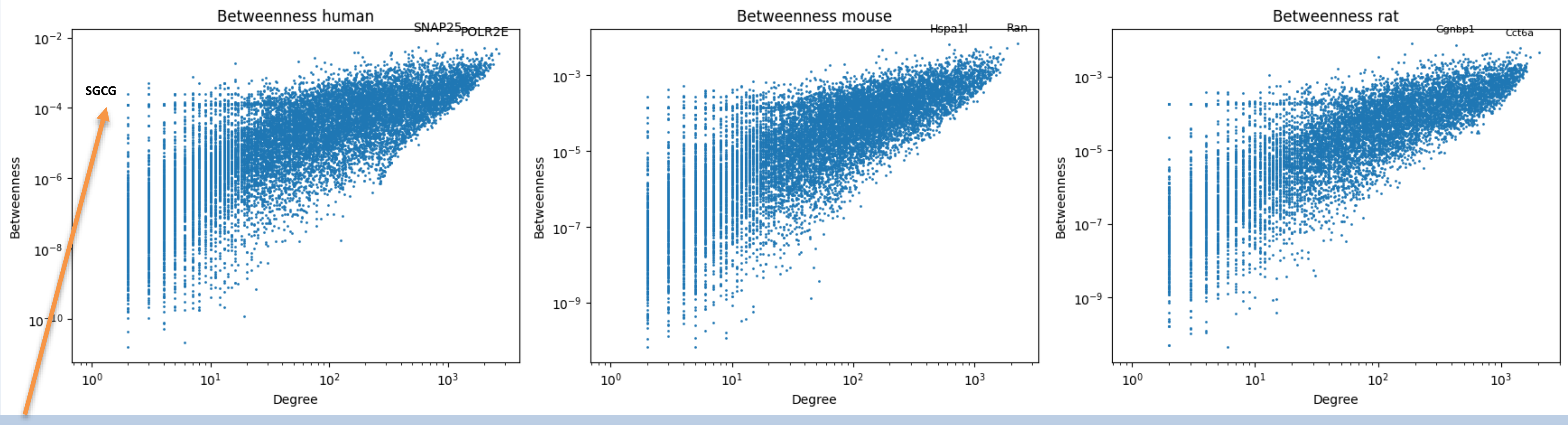


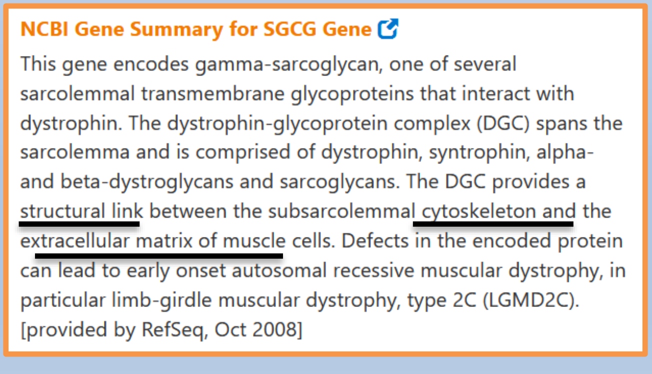
**NCBI Gene Summary for SGCG Gene**

This gene encodes gamma-sarcoglycan, one of several sarcolemmal transmembrane glycoproteins that interact with dystrophin. The dystrophin-glycoprotein complex (DGC) spans the sarcolemma and is comprised of dystrophin, syntrophin, alpha- and beta-dystroglycans and sarcoglycans. The DGC provides a structural link between the subsarcolemmal cytoskeleton and the extracellular matrix of muscle cells. Defects in the encoded protein can lead to early onset autosomal recessive muscular dystrophy, in particular limb-girdle muscular dystrophy, type 2C (LGMD2C). [provided by RefSeq, Oct 2008]

**Low degree:** specialized function related to muscle

# Betweenness centrality





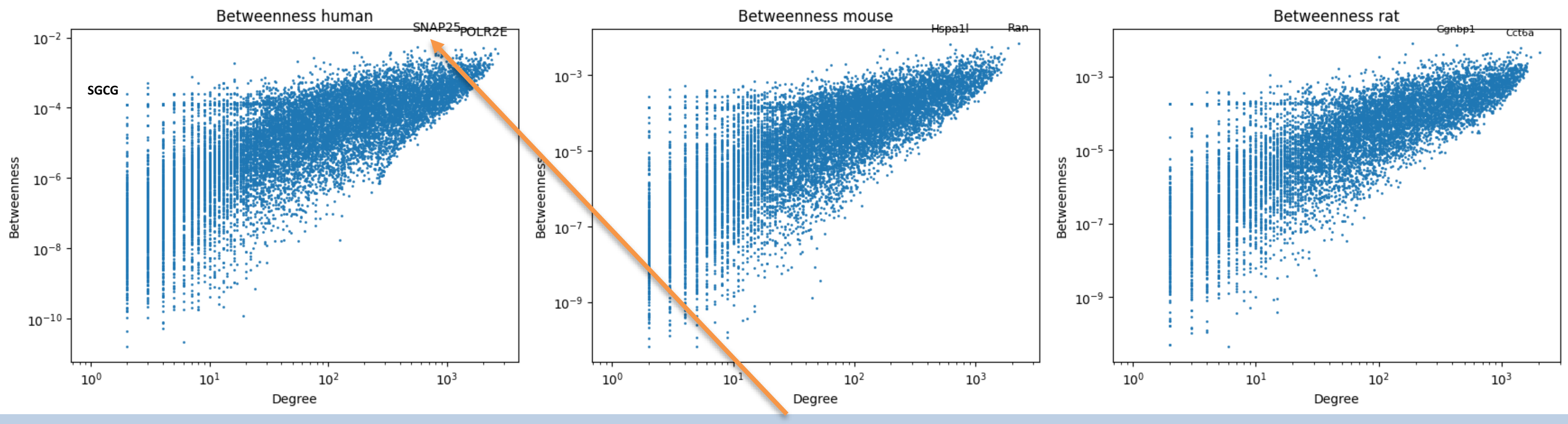
**NCBI Gene Summary for SGCG Gene**

This gene encodes gamma-sarcoglycan, one of several sarcolemmal transmembrane glycoproteins that interact with dystrophin. The dystrophin-glycoprotein complex (DGC) spans the sarcolemma and is comprised of dystrophin, syntrophin, alpha- and beta-dystroglycans and sarcoglycans. The DGC provides a structural link between the subsarcolemmal cytoskeleton and the extracellular matrix of muscle cells. Defects in the encoded protein can lead to early onset autosomal recessive muscular dystrophy, in particular limb-girdle muscular dystrophy, type 2C (LGMD2C). [provided by RefSeq, Oct 2008]

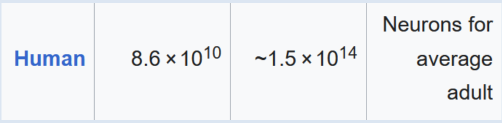
**Low degree:** specialized function related to muscle

**High betweenness:** mediates communication between cells.

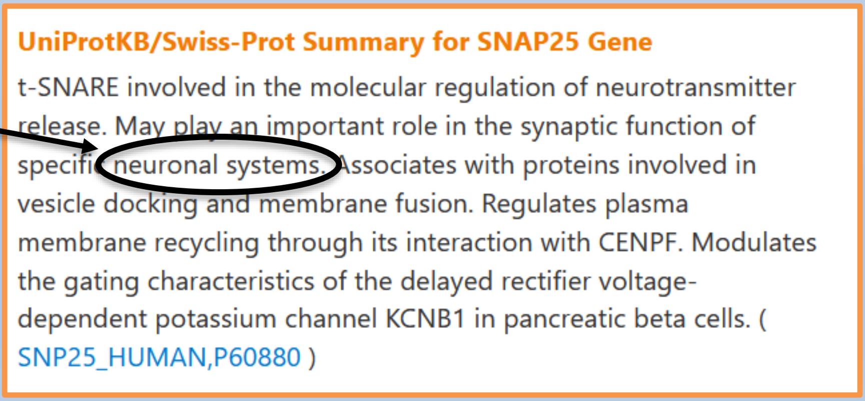
# Betweenness centrality



## High degree



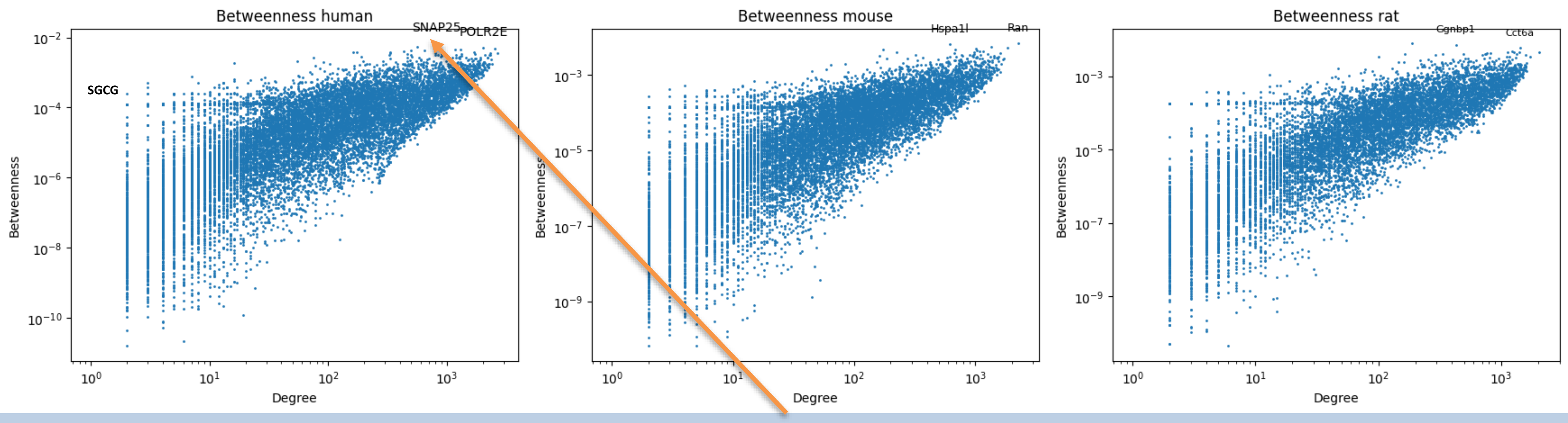
| | | | |
|---|---|---|---|
| Human | $8.6 \times 10^{10}$ | $\sim 1.5 \times 10^{14}$ | Neurons for average adult |



**UniProtKB/Swiss-Prot Summary for SNAP25 Gene**

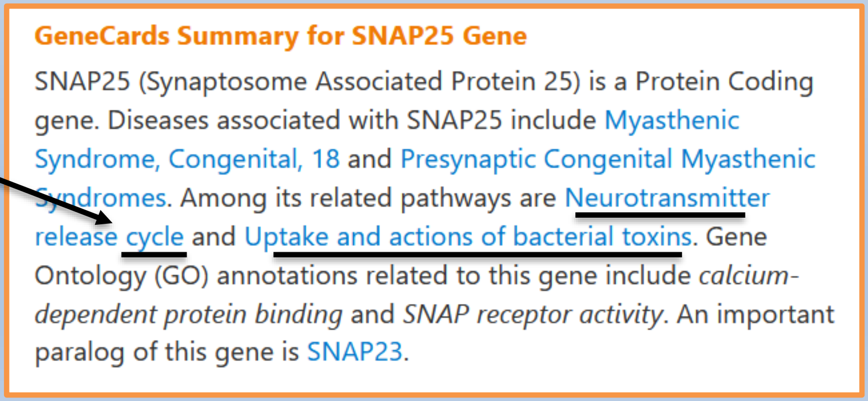
t-SNARE involved in the molecular regulation of neurotransmitter release. May play an important role in the synaptic function of specific neuronal systems. Associates with proteins involved in vesicle docking and membrane fusion. Regulates plasma membrane recycling through its interaction with CENPF. Modulates the gating characteristics of the delayed rectifier voltage-dependent potassium channel KCNB1 in pancreatic beta cells. ( SNP25_HUMAN,P60880 )

# Betweenness centrality



**High degree & betweenness**

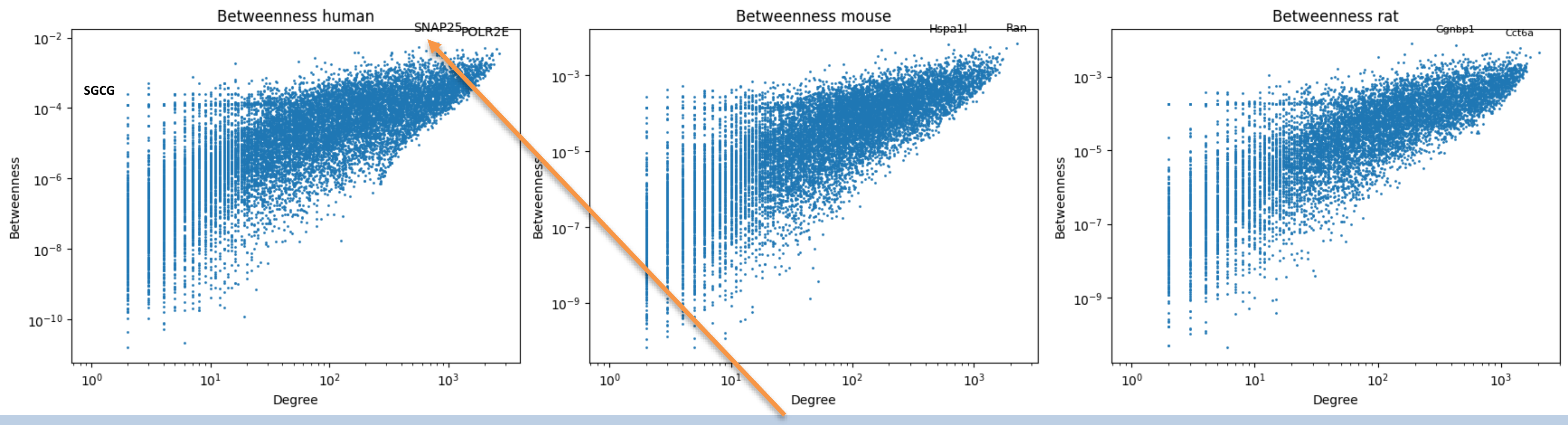
Involved in 2 biological processes



**GeneCards Summary for SNAP25 Gene**

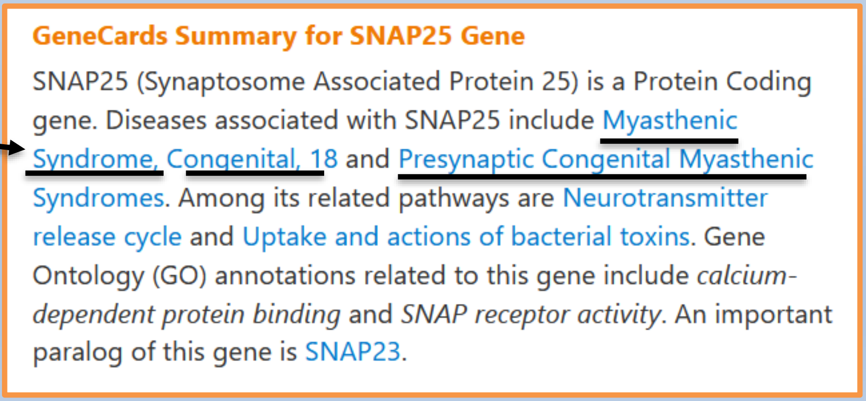
SNAP25 (Synaptosome Associated Protein 25) is a Protein Coding gene. Diseases associated with SNAP25 include Myasthenic Syndrome, Congenital, 18 and Presynaptic Congenital Myasthenic Syndromes. Among its related pathways are Neurotransmitter release cycle and Uptake and actions of bacterial toxins. Gene Ontology (GO) annotations related to this gene include *calcium-dependent protein binding* and *SNAP receptor activity*. An important paralog of this gene is SNAP23.

# Betweenness centrality



**Highest betweenness:**

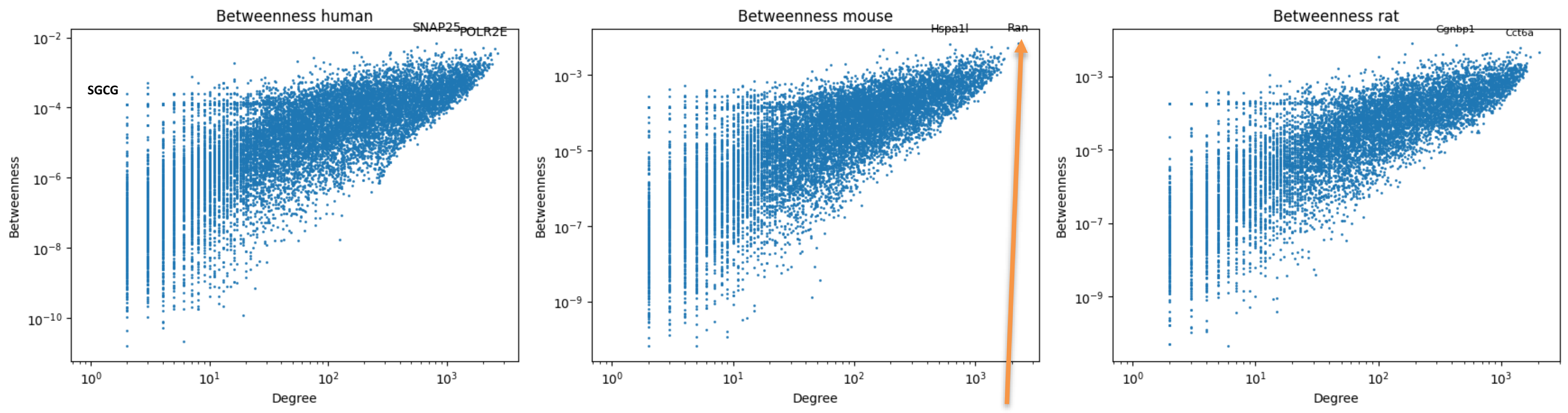
Mutations causes several diseases



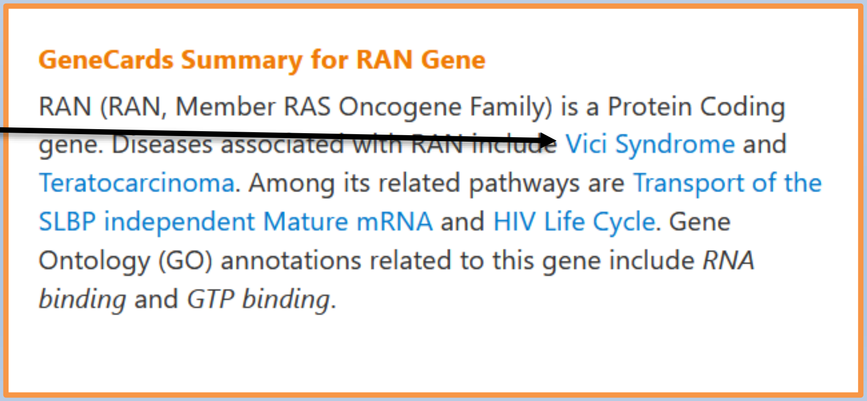
**GeneCards Summary for SNAP25 Gene**

SNAP25 (Synaptosome Associated Protein 25) is a Protein Coding gene. Diseases associated with SNAP25 include Myasthenic Syndrome, Congenital, 18 and Presynaptic Congenital Myasthenic Syndromes. Among its related pathways are Neurotransmitter release cycle and Uptake and actions of bacterial toxins. Gene Ontology (GO) annotations related to this gene include *calcium-dependent protein binding* and *SNAP receptor activity*. An important paralog of this gene is SNAP23.

# Betweenness centrality



A rare multisystem **disorder**: **lack of** the corpus callosum, that **link the two hemispheres**

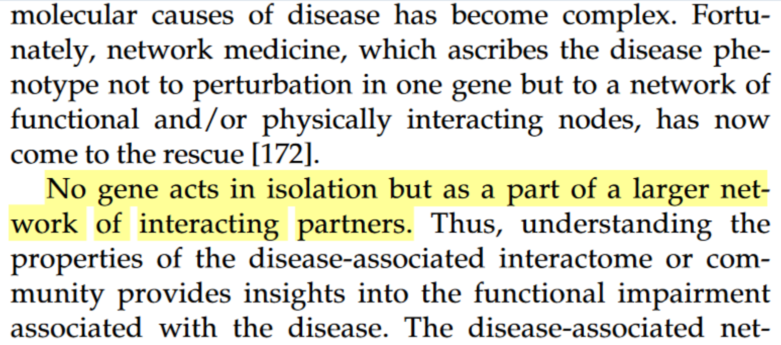


**GeneCards Summary for RAN Gene**

RAN (RAN, Member RAS Oncogene Family) is a Protein Coding gene. Diseases associated with RAN include Vici Syndrome and Teratocarcinoma. Among its related pathways are Transport of the SLBP independent Mature mRNA and HIV Life Cycle. Gene Ontology (GO) annotations related to this gene include *RNA binding* and *GTP binding*.

https://www.telethon.it/cosa-facciamo/ricerca/malattie-studiate/sindrome-di-vici/

# The need of a community based analysis



molecular causes of disease has become complex. Fortunately, network medicine, which ascribes the disease phenotype not to perturbation in one gene but to a network of functional and/or physically interacting nodes, has now come to the rescue [172].

No gene acts in isolation but as a part of a larger network of interacting partners. Thus, understanding the properties of the disease-associated interactome or community provides insights into the functional impairment associated with the disease. The disease-associated net-

"TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 20, JANUARY/FEBRUARY 2023"

## **Single** gene **mutation ≠ disease**

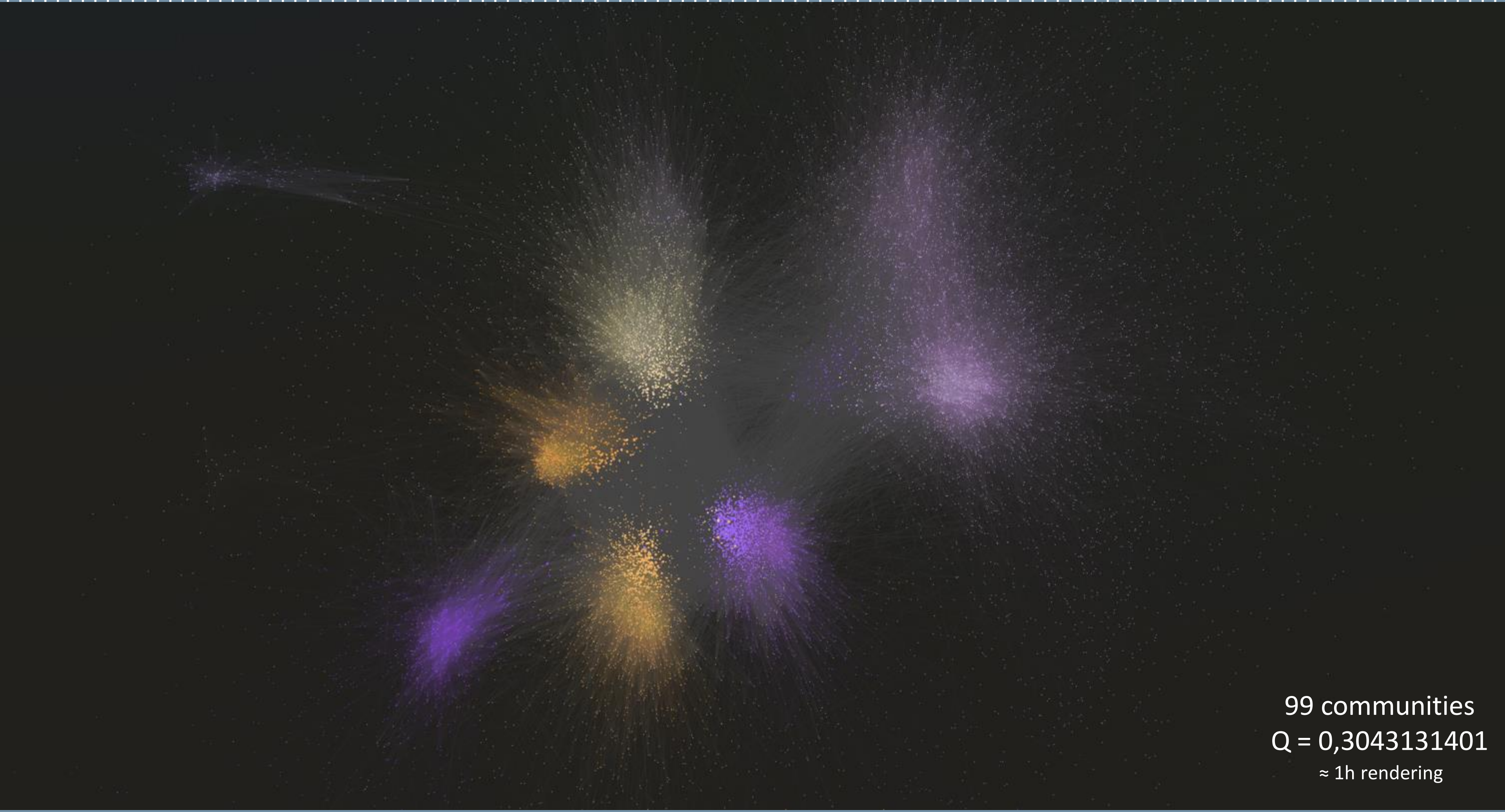
(Not necessarly)

Community-based analysis allow to gain **insights about functional impairment of the disease**

Community detection = Disease module detection

(network jargon) (medicine jargon)

# Human interactome network communities



99 communities

Q = 0,3043131401

≈ 1h rendering

# Properties of the most important communities

| # of nodes | <k> | knn |
|---|---|---|
| 5389 | 28.09317 | 272.2270 |
| 2706 | 87.4077 | 199.6563 |
| 2350 | 253.0825 | 503.5698 |
| 2142 | 117.9299 | 77.73041 |

| ANOVA highest pvalue | 1.1102e-16 |
|---|---|

# Properties of the most important communities

| # of nodes | <k> | knn |
|---|---|---|
| 5389 | 28.09317 | 272.2270 |
| 2706 | 87.4077 | 199.6563 |
| 2350 | **253.0825** | **503.5698** |
| 2142 | 117.9299 | 77.73041 |

| ANOVA highest pvalue | 1.11e-16 |
|---|---|

core biological **module in which genes coordinate cohesely**!

**Dense community**

**nodes are likely to connect to nodes with high degree**

**10% density: 1/10 of all possible edges are actually present!**

# Database for community interpretation

reactome

An open-source, open access, manually curated and reviewed pathway database. For pathway visualization and interpretation.

GenAge

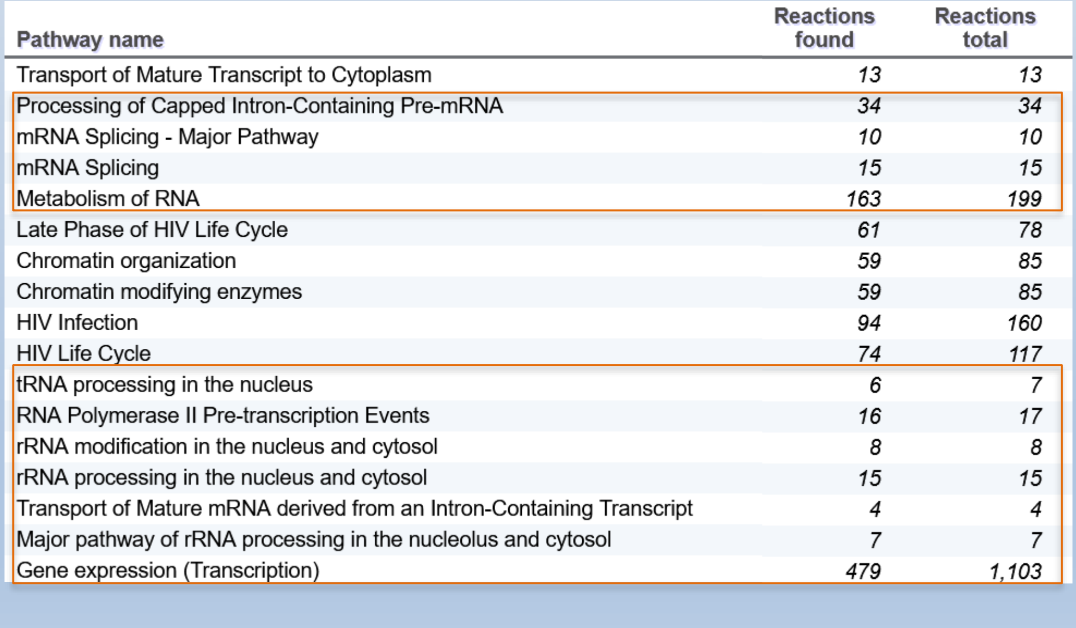
DB of genes related to human ageing. Result of extensive review of the literature with manually-curated annotation.

Human Ageing Genomic Resources: new and updated databases. (597 citations)

# Community interpretation

| # of nodes | <r> | knn |
|---|---|---|
| 5389 | 28.09317 | 272.2270 |
| 2706 | 87.4077 | 199.6563 |
| 2350 | 253.0825 | 503.5698 |
| 2142 | 117.9299 | 77.73041 |

**Core biological module**



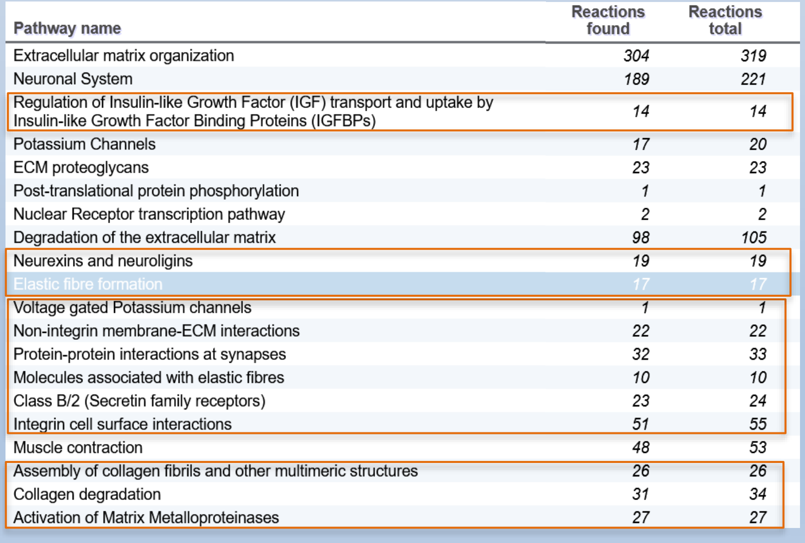
| Pathway name | Reactions found | Reactions total |
|---|---|---|
| Transport of Mature Transcript to Cytoplasm | 13 | 13 |
| Processing of Capped Intron-Containing Pre-mRNA | 34 | 34 |
| mRNA Splicing - Major Pathway | 10 | 10 |
| mRNA Splicing | 15 | 15 |
| Metabolism of RNA | 163 | 199 |
| Late Phase of HIV Life Cycle | 61 | 78 |
| Chromatin organization | 59 | 85 |
| Chromatin modifying enzymes | 59 | 85 |
| HIV Infection | 94 | 160 |
| HIV Life Cycle | 74 | 117 |
| tRNA processing in the nucleus | 6 | 7 |
| RNA Polymerase II Pre-transcription Events | 16 | 17 |
| rRNA modification in the nucleus and cytosol | 8 | 8 |
| rRNA processing in the nucleus and cytosol | 15 | 15 |
| Transport of Mature mRNA derived from an Intron-Containing Transcript | 4 | 4 |
| Major pathway of rRNA processing in the nucleolus and cytosol | 7 | 7 |
| Gene expression (Transcription) | 479 | 1,103 |

reactome

# Community interpretation

| # of nodes | <r> | knn |
|---|---|---|
| 5389 | 28.09317 | 272.2270 |
| 2706 | 87.4077 | 199.6563 |
| 2350 | 253.0825 | 503.5698 |
| 2142 | 117.9299 | 77.73041 |

reactome



| Pathway name | Reactions found | Reactions total |
|---|---|---|
| Extracellular matrix organization | 304 | 319 |
| Neuronal System | 189 | 221 |
| Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs) | 14 | 14 |
| Potassium Channels | 17 | 20 |
| ECM proteoglycans | 23 | 23 |
| Post-translational protein phosphorylation | 1 | 1 |
| Nuclear Receptor transcription pathway | 2 | 2 |
| Degradation of the extracellular matrix | 98 | 105 |
| Neurexins and neuroligins | 19 | 19 |
| Elastic fibre formation | 17 | 17 |
| Voltage gated Potassium channels | 1 | 1 |
| Non-integrin membrane-ECM interactions | 22 | 22 |
| Protein-protein interactions at synapses | 32 | 33 |
| Molecules associated with elastic fibres | 10 | 10 |
| Class B/2 (Secretin family receptors) | 23 | 24 |
| Integrin cell surface interactions | 51 | 55 |
| Muscle contraction | 48 | 53 |
| Assembly of collagen fibrils and other multimeric structures | 26 | 26 |
| Collagen degradation | 31 | 34 |
| Activation of Matrix Metalloproteinases | 27 | 27 |

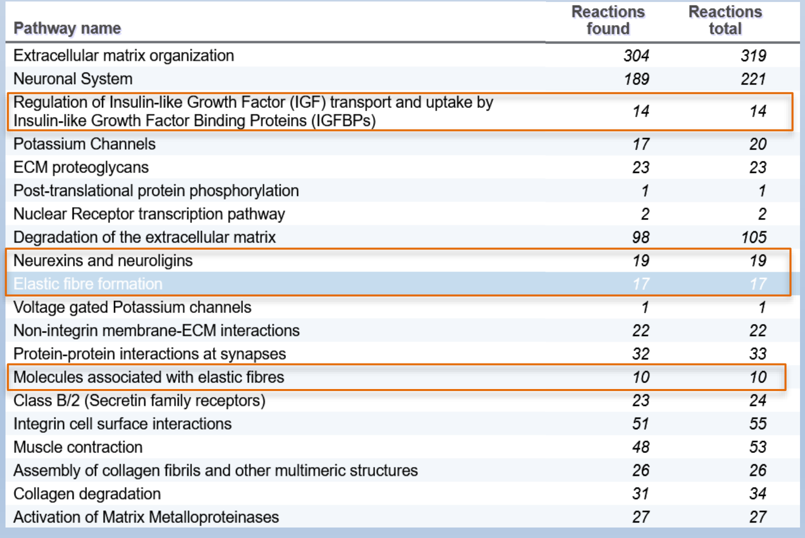
GenAge

**All** “entries with evidence **linking** the gene or its product **to longevity**”

# Community interpretation

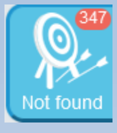
| # of nodes | <r> | knn |
|---|---|---|
| 5389 | 28.09317 | 272.2270 |
| 2706 | 87.4077 | 199.6563 |
| 2350 | 253.0825 | 503.5698 |
| 2142 | 117.9299 | 77.73041 |

reactome



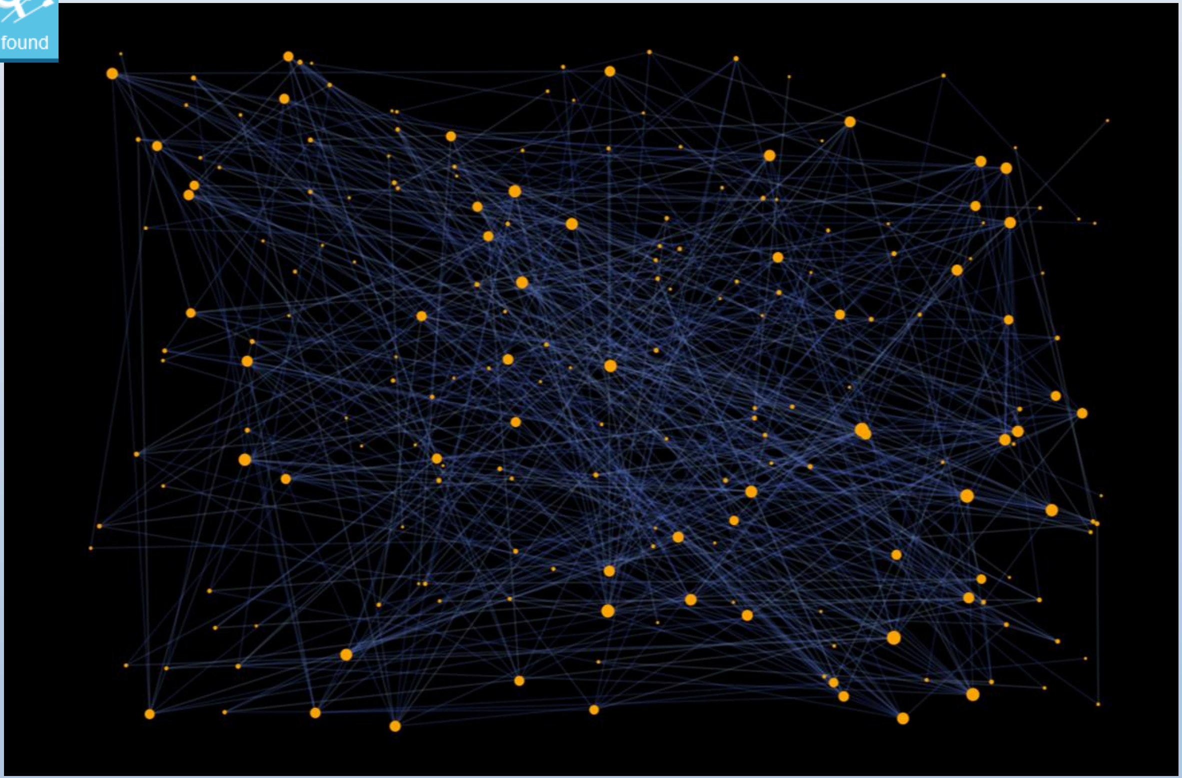
| Pathway name | Reactions found | Reactions total |
|---|---|---|
| Extracellular matrix organization | 304 | 319 |
| Neuronal System | 189 | 221 |
| Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs) | 14 | 14 |
| Potassium Channels | 17 | 20 |
| ECM proteoglycans | 23 | 23 |
| Post-translational protein phosphorylation | 1 | 1 |
| Nuclear Receptor transcription pathway | 2 | 2 |
| Degradation of the extracellular matrix | 98 | 105 |
| Neurexins and neuroligins | 19 | 19 |
| Elastic fibre formation | 17 | 17 |
| Voltage gated Potassium channels | 1 | 1 |
| Non-integrin membrane-ECM interactions | 22 | 22 |
| Protein-protein interactions at synapses | 32 | 33 |
| Molecules associated with elastic fibres | 10 | 10 |
| Class B/2 (Secretin family receptors) | 23 | 24 |
| Integrin cell surface interactions | 51 | 55 |
| Muscle contraction | 48 | 53 |
| Assembly of collagen fibrils and other multimeric structures | 26 | 26 |
| Collagen degradation | 31 | 34 |
| Activation of Matrix Metalloproteinases | 27 | 27 |

**However…**



# The uncategorized gene network

347
Not found



**Not all genes have been studied.**

They are too much!

Without experimental evidence it's challenging to categorize those genes into pathways.

# Maximize modularity for gene classification

**All genes experimentally observed to link with longevity are actually in this community**

Source: GenAge

(A weak) **Hypothesis:** this community is capturing some mechanism of longevity

This assumption can’t be actually checked

**From now on consider only this community as the whole network (longevity network)**

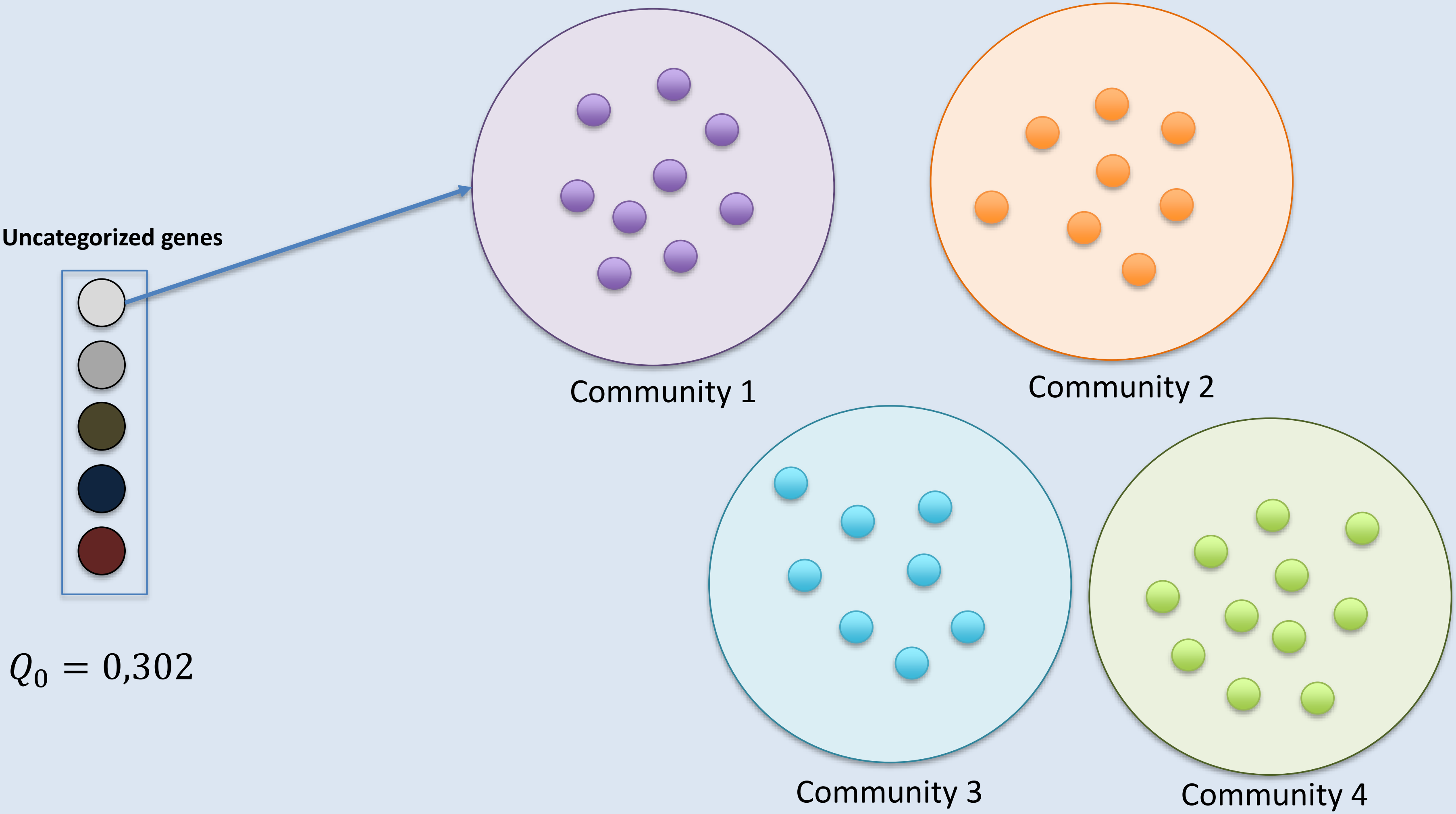
1: *Find communities of the longevity network*

2: *Measure modularity when placing an uncategorized gene in all possible communities*

3: *Place the uncategorized gene in the community maximizing modularity*

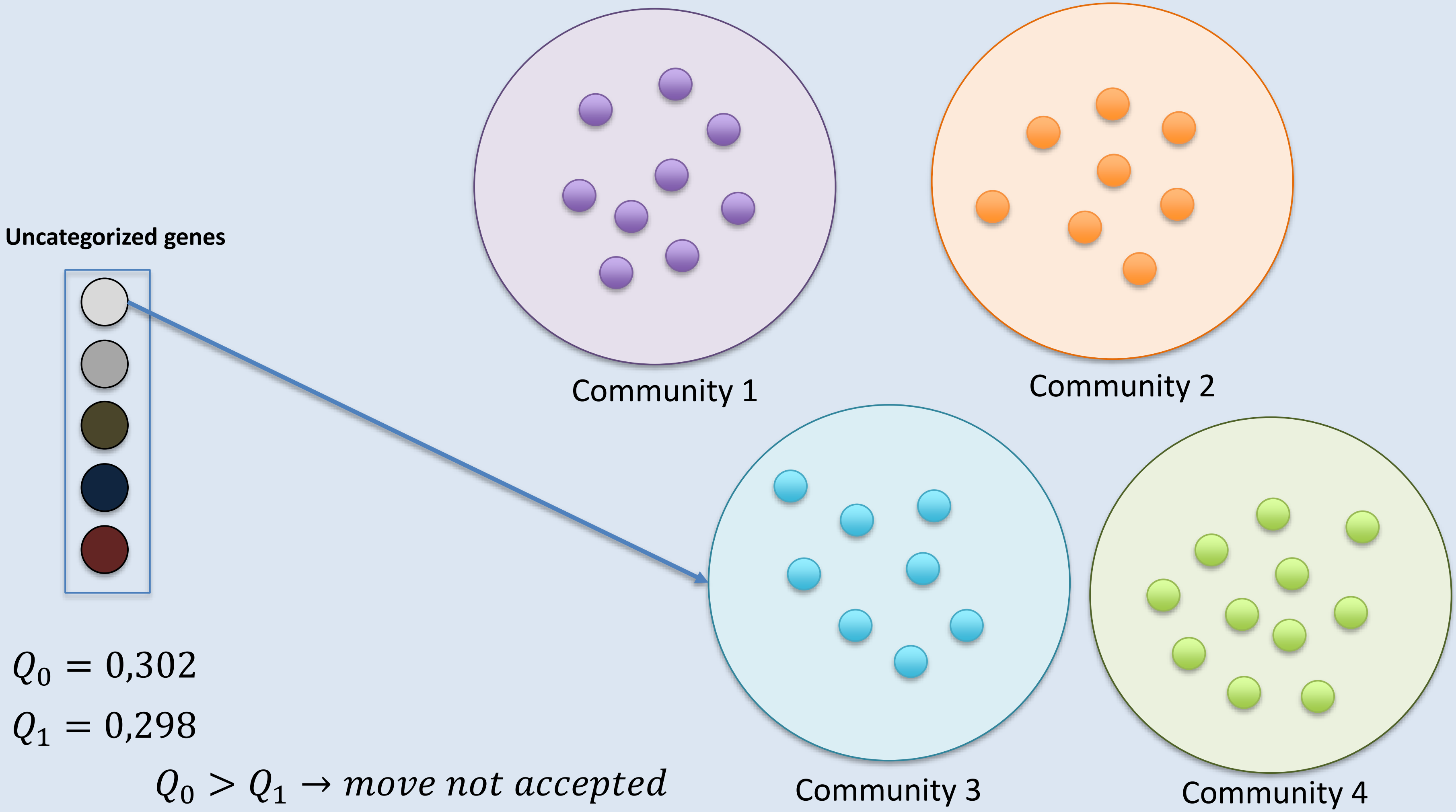
4: *Repeat for each uncategorized gene*

# Maximize modularity



$Q_0 = 0{,}302$

# Maximize modularity

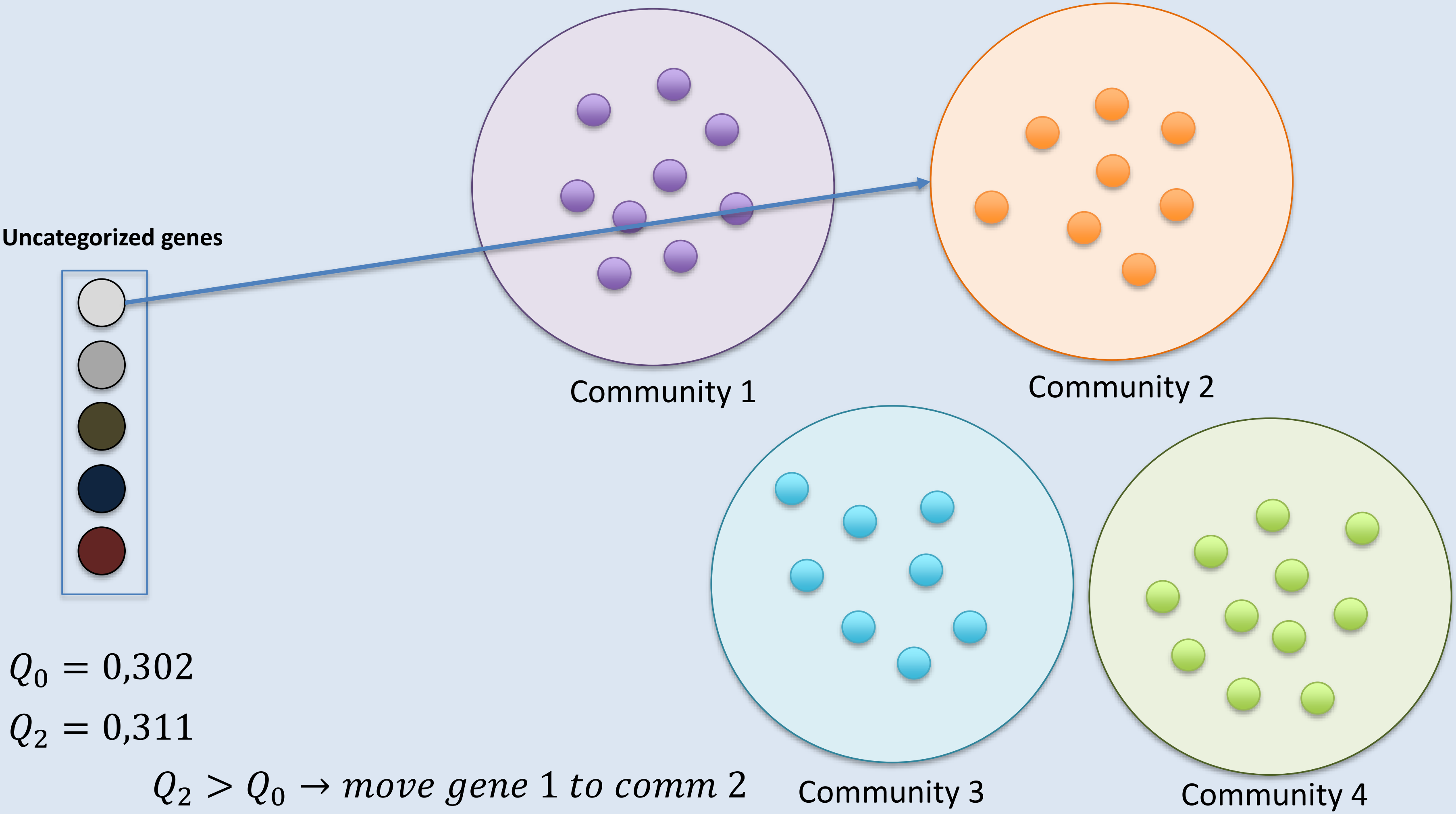


$Q_0 = 0{,}302$

$Q_1 = 0{,}298$

$Q_0 > Q_1 \rightarrow move\ not\ accepted$

# Maximize modularity

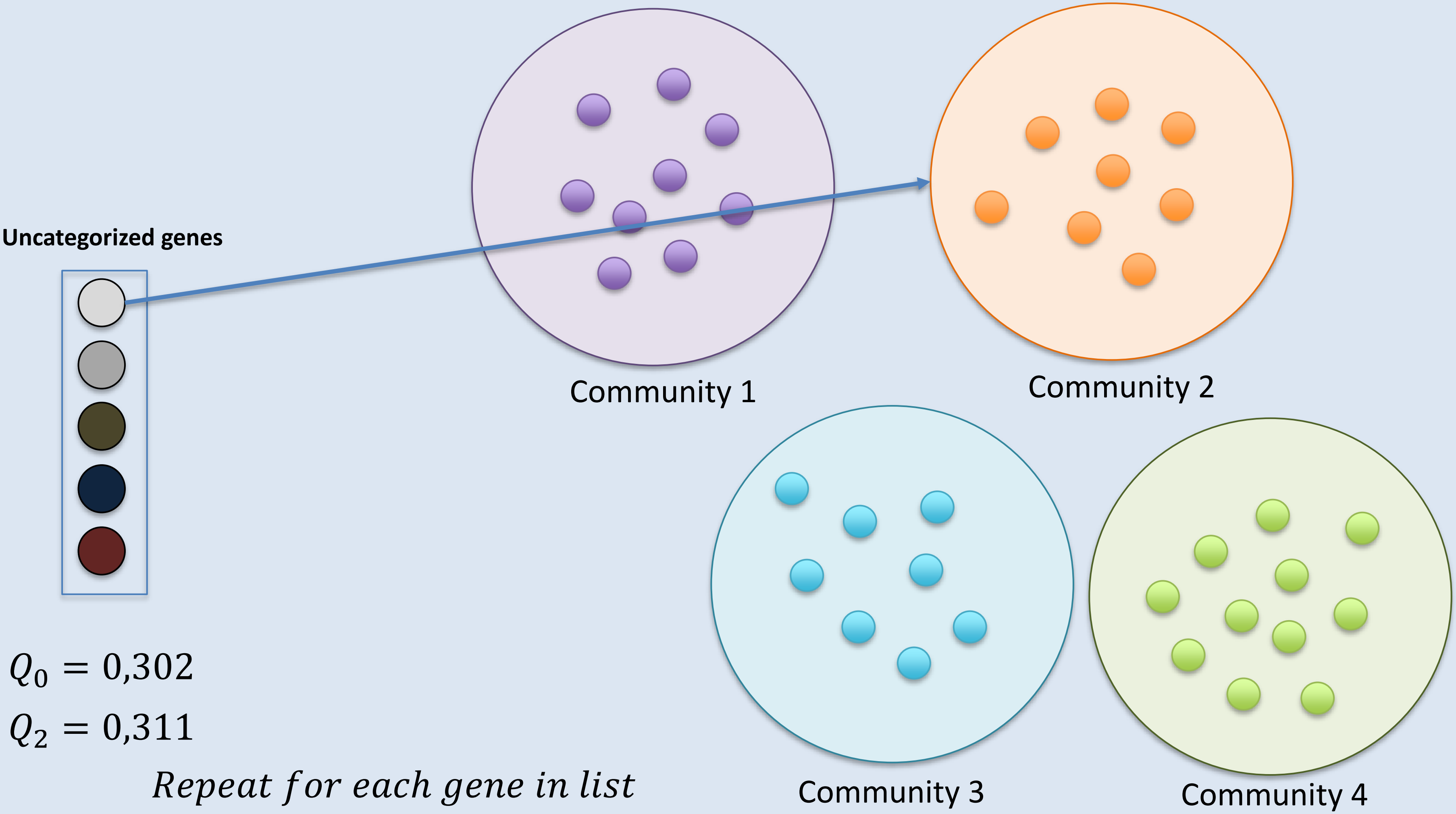


$Q_0 = 0{,}302$

$Q_2 = 0{,}311$

$Q_2 > Q_0 \rightarrow move\ gene\ 1\ to\ comm\ 2$

# Maximize modularity



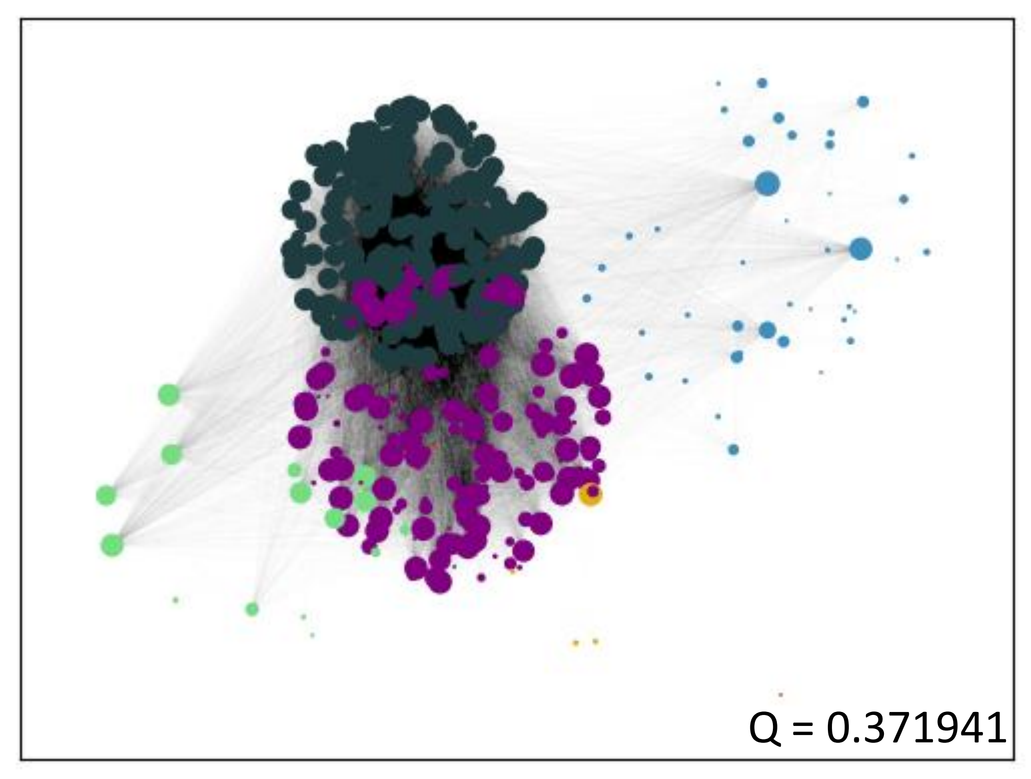
$Q_0 = 0{,}302$

$Q_2 = 0{,}311$

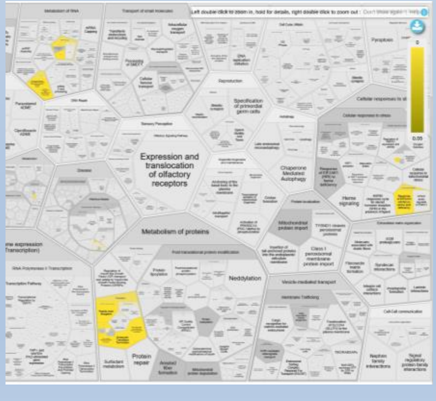
$Repeat\ for\ each\ gene\ in\ list$

# The final longevity network community

Heuristic runned on a very small subset (5) of uncategorized genes and thresholded original community

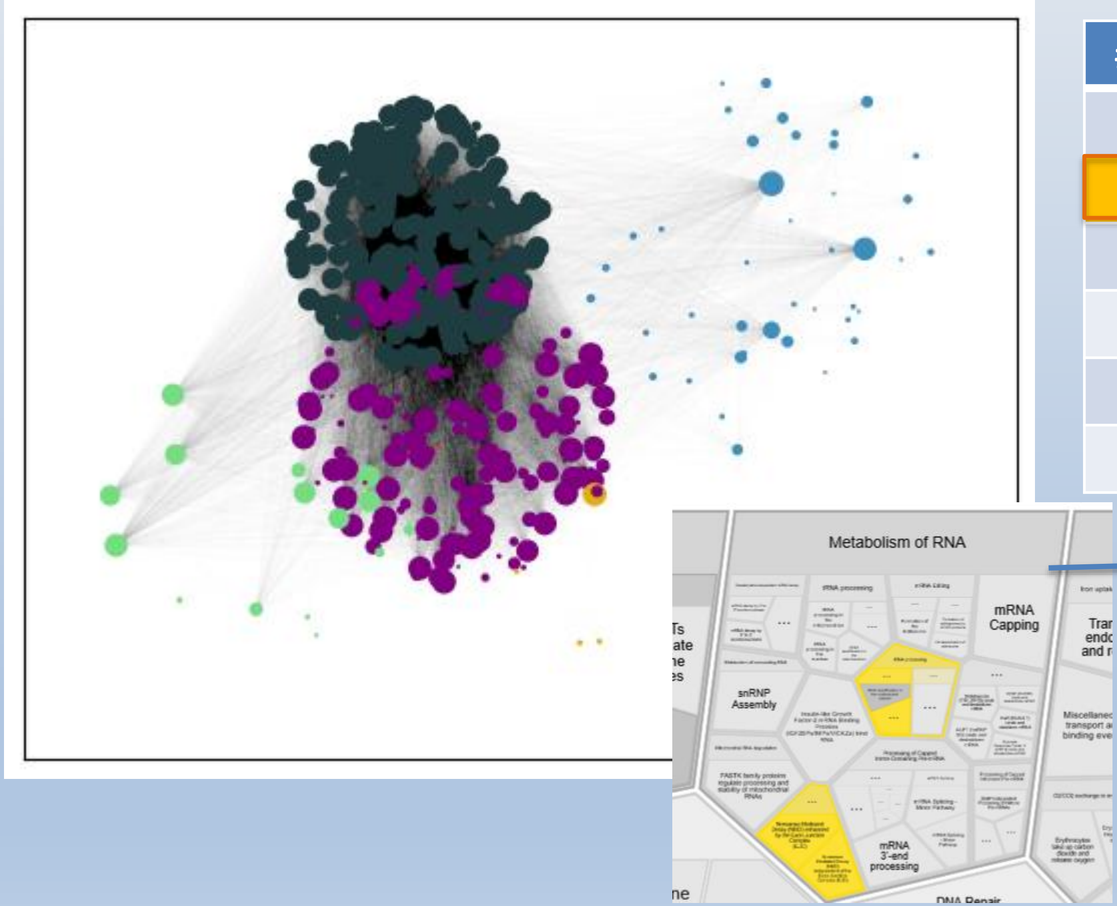


| # of nodes | <r> | knn |
|---|---|---|
| 197 | 148.467 | 162.0642 |
| 151 | 67.3245 | 89.7882 |
| 4 | 3.5 | 3.44999 |
| 41 | 9.46341 | 17.6569 |
| 15 | 6.133333 | 8.04566 |
| 2 | 1.0 | 1.0 |


Expression and translocation of olfactory receptors
Metabolism of proteins
Neddylation
Chaperone Mediated Autophagy
Mitochondrial protein import

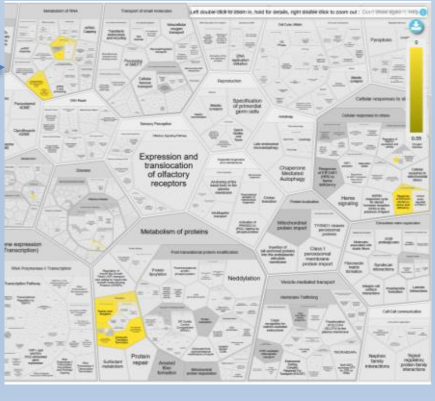
# The final longevity network community

Heuristic runned on a very small subset (5) of uncategorized genes and thresholded original community



Metabolism of RNA
tRNA processing
mRNA Editing
mRNA Capping
snRNP Assembly
Insulin-like Growth Factor-2 mRNA Binding Proteins (IGF2BPs/IMPs/VICKZs) bind RNA
Processing of Capped Intron-Containing Pre-mRNA
FASTK family proteins regulate processing and stability of mitochondrial RNAs
mRNA Splicing - Minor Pathway
mRNA 3'-end processing
DNA Repair
Miscellaneous transport and binding events
Erythrocytes take up carbon dioxide and release oxygen

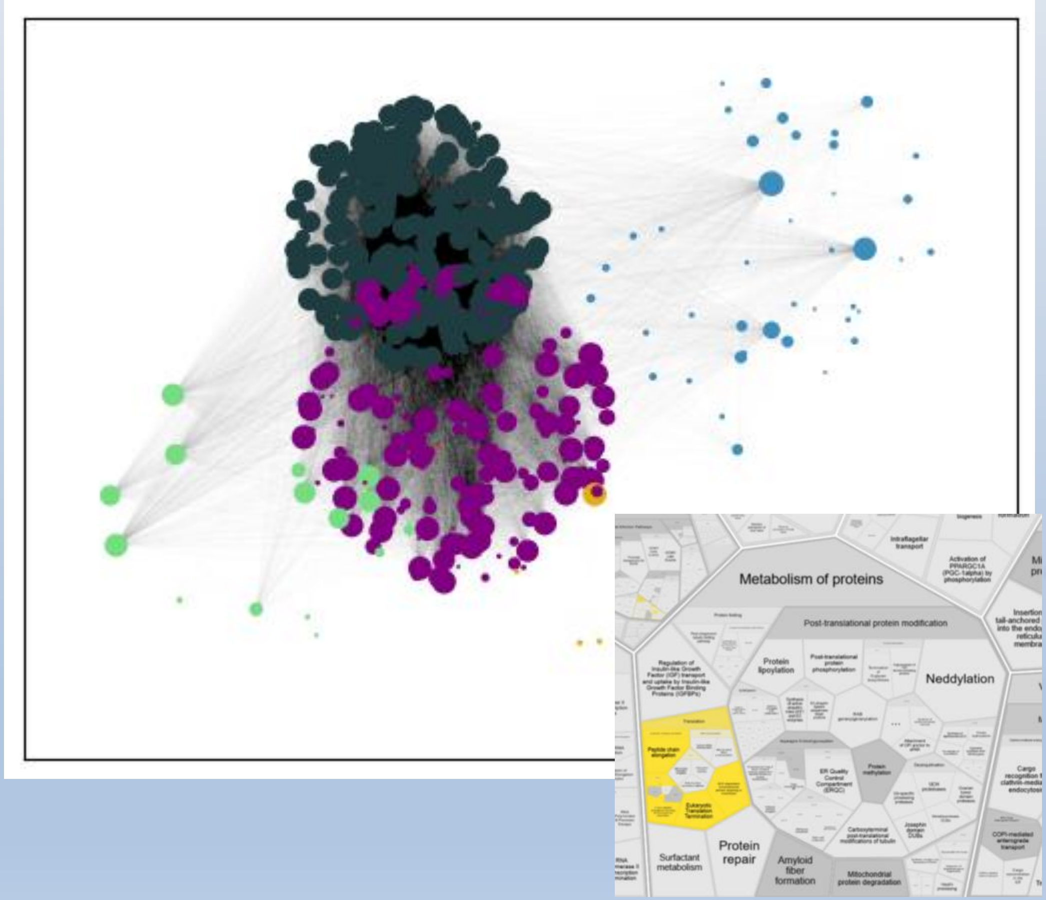
| # of nodes | <r> | knn |
|---|---|---|
| 197 | 148.467 | 162.0642 |
| 151 | 67.3245 | 89.7882 |
| 4 | 3.5 | 3.44999 |
| 41 | 9.46341 | 17.6569 |
| 15 | 6.133333 | 8.04566 |
| 2 | 1.0 | 1.0 |



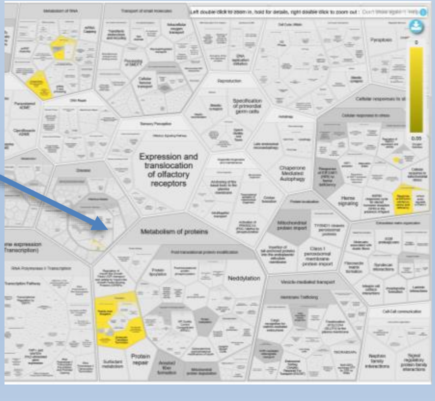
Expression and translocation of olfactory receptors
Metabolism of proteins
Neddylation
Chaperone Mediated Autophagy
Protein repair

# The final longevity network community

Heuristic runned on a very small subset (5) of uncategorized genes and thresholded original community


Metabolism of proteins
Post-translational protein modification
Neddylation
Protein lipoylation
Post-translational protein phosphorylation
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
Intraflagellar transport
Activation of PPARGC1A (PGC-1alpha) by phosphorylation
Peptide chain elongation
Eukaryotic Translation Termination
ER Quality Control Compartment (ERQC)
Protein methylation
Carboxyterminal post-translational modifications of tubulin
Josephin domain DUBs
Cargo recognition for clathrin-mediated endocytosis
COPI-mediated anterograde transport
Surfactant metabolism
Protein repair
Amyloid fiber formation
Mitochondrial protein degradation

| # of nodes | <r> | knn |
|---|---|---|
| 197 | 148.467 | 162.0642 |
| 151 | 67.3245 | 89.7882 |
| 4 | 3.5 | 3.44999 |
| 41 | 9.46341 | 17.6569 |
| 15 | 6.133333 | 8.04566 |
| 2 | 1.0 | 1.0 |



# Conclusions

- Genome interactome is a scale free, disassortative network
- This guarantee robustness to failure (not to attacks)
- Communities are crucial to understand intrinsic behaviour of such a complex network
- Modularity maximization (or other heuristics) are nowadays crucial as a pre-screening of genome to further study on laboratory
- Complex network analysis is more and more used also for medical purposes