Corfu Workshop on Medical Physics & Biomedical Engineering

Subpathway-based Computational Approaches to Network Medicine

Aristidis G. Vrahatis

Postdoctoral researcher

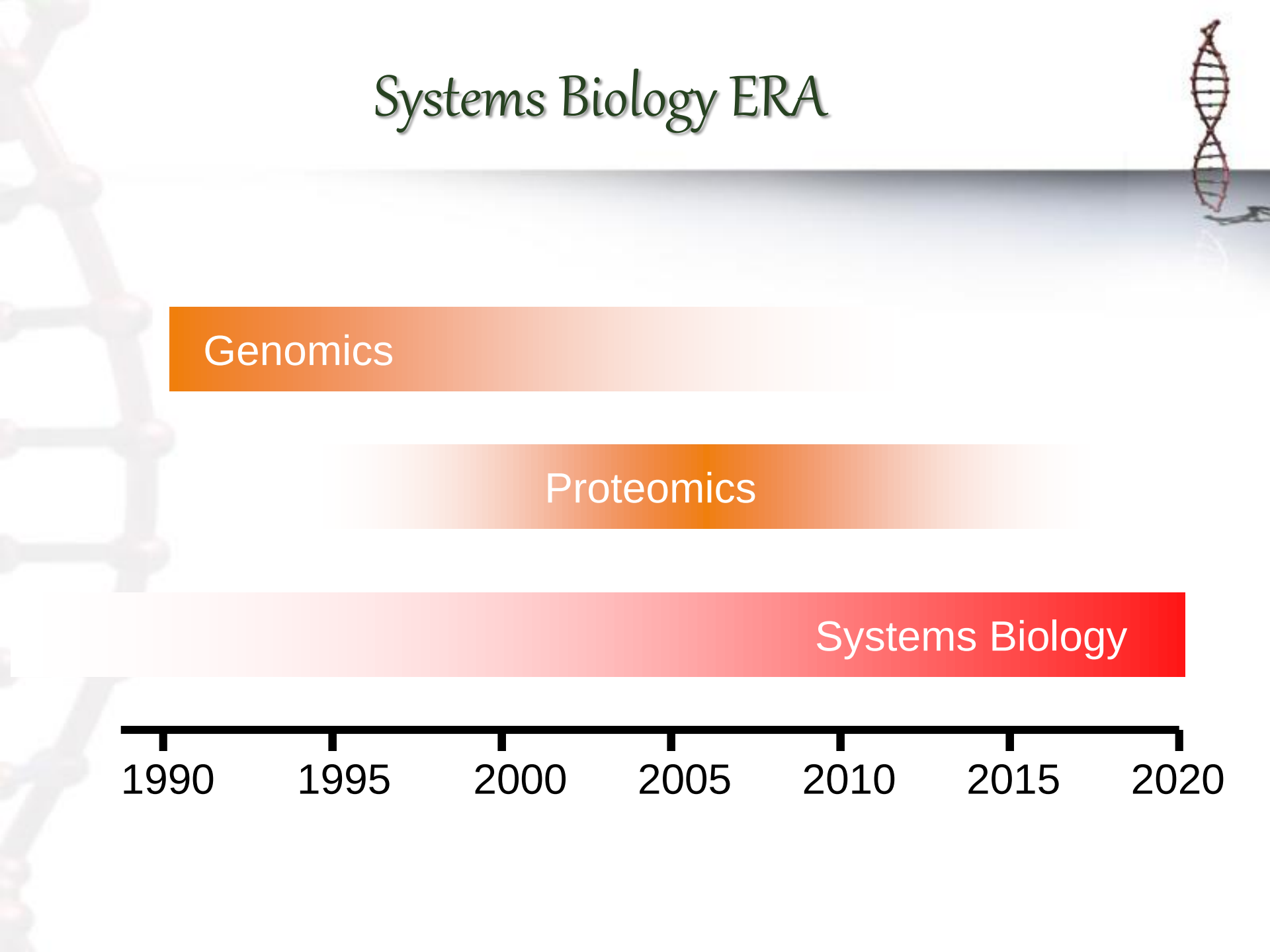
Outline

i. Systems Biology and Extensions
ii. Network-Based Approaches
iii. Subpathway Approaches
   - CHRONOS tool
   - DEsubs tool

Outline

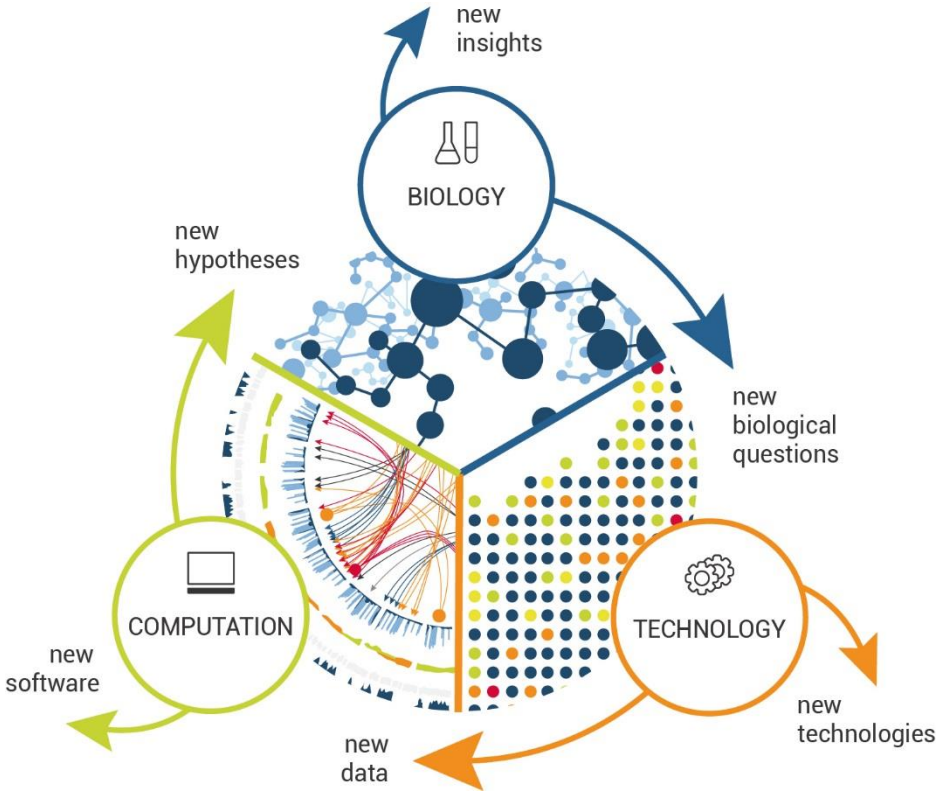
i. **Systems Biology and Extensions**
ii. Network-Based Approaches
iii. Subpathway Approaches
   - CHRONOS tool
   - DEsubs tool

Systems Biology ERA



Systems Biology

- Focuses on complex interactions within biological systems
- Based on the understanding that the whole is greater than the sum of the parts
- Integrates and analyzes complex data sets from multiple experimental sources



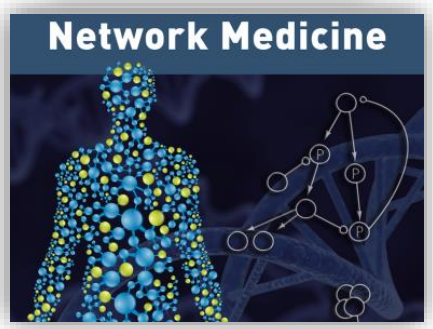
Source: www.systemsbiology.org

Systems Biology Extensions

- Systems Medicine
  - Implementation of Systems Biology approaches in medical concepts, research and practice.
  - Systems Biology + clinical practice

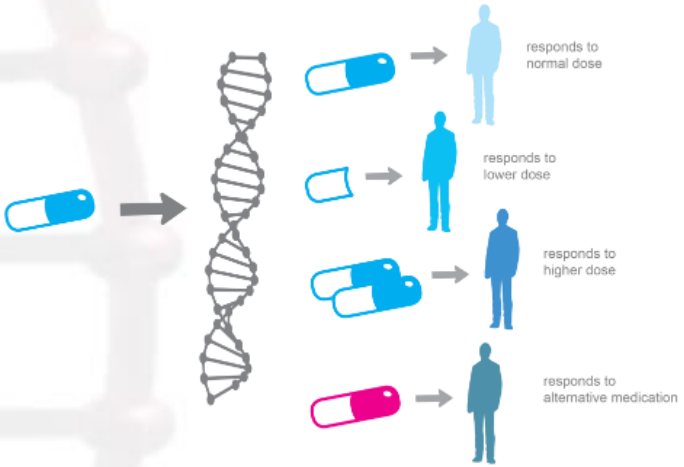
- Systems Pharmacology
  - Application of systems biology principles to the field of pharmacology.

- Network Medicine (Nature Reviews Genetics, 2011)
  - Applies network science and systems biology approaches to understand the causes of human complex diseases

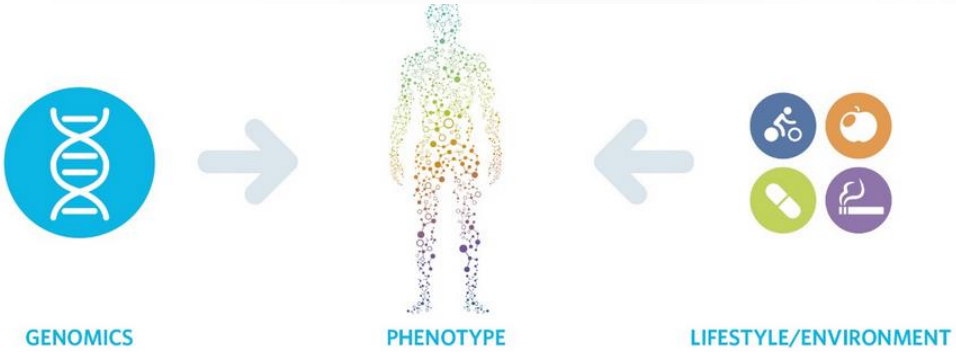


Systems Biology Extensions

— Personalized Medicine



— Precision Medicine



— P4 Medicine

Systems Biology, WHY ???

Main Problem of Diseases

- Human Cell
  - Hundreds of thousands of functional interdependencies between the molecular components in a human cell
- Diseases
  - A disease is rarely a consequence of an abnormality in a single gene, but reflects the perturbations of the complex intracellular and intercellular network
  - Typically disease-associated genes have small effects building up a significant combined effect (Nature genetics, 2007).

Systems Biology, WHY ???

Low Drug Efficacy

- The reductionists approaches in drug development has served the community and the industry for a long time,
  - however the complexity of diseases such as cancer make this approach as efficient as searching needles in haystacks.

- Drug discovery until now
  - focus on identifying single selective drugs that target a single mechanism

- Drug efficacy
  - Low due to the inherent complexity of diseases
  - 90% FDA approved drugs on the US work in about 40% of patients
  - annual cost of ineffective drugs in the US alone is estimated at US$350 billion (Genome Medicine, 2014).
  - cost of drug development now → $2.6 billion per drug

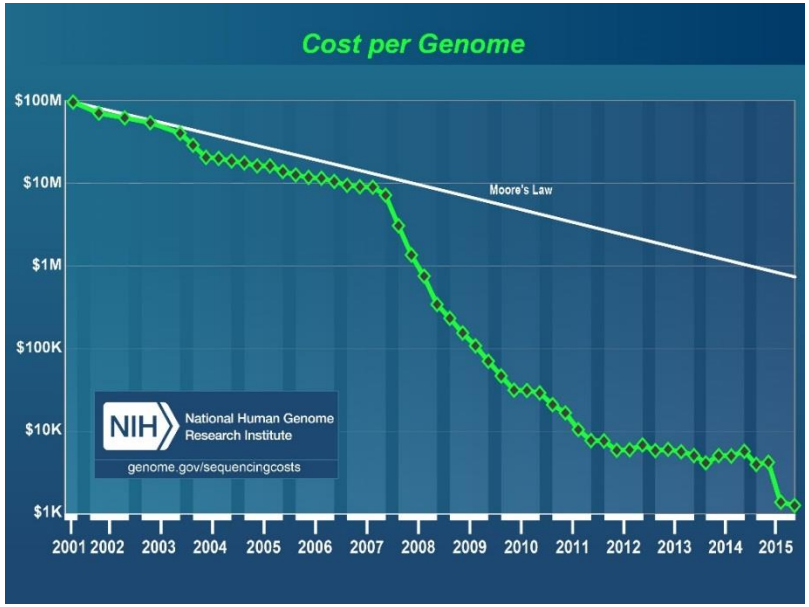
Systems Biology, WHY ???

- Disease arises from disease-perturbed networks
  - The holistic approach employed by Systems Biology analyzes how all the components of a biological system interact.

- Systems Biology approaches have shown that they can tackle part of the complexity of perturbed biological processes.
  - Any high-demand research endeavor can benefit from a proper systems biology approach tailored to the needs of the specific problem.
  - Both in the level of interpreting the results, as well as designing the experiment itself making teams with such expertise invaluable.

What will change in the near future ???

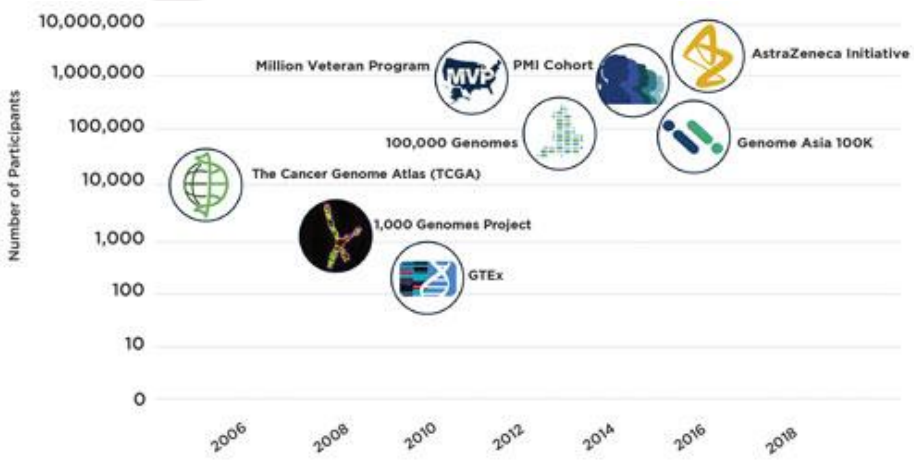
- DNA sequencing costs began at $95.263.072 in 2001 (Human Genome Project) and dropped to $1.363 in July 2015 (Wetterstrand, 2015).
  - Illunima's latest product drops this estimate to $1.000 in December 2015.
  - The final projection to the corresponding cost is a few hundred dollars.
- In the coming years researchers will move from genotyping platforms based on known genetic variants to genome sequencing of thousands of individuals.

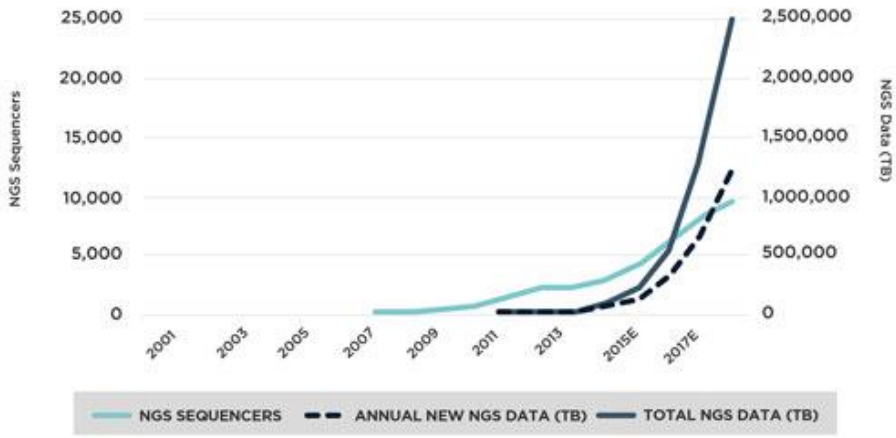


Source: NHGRI (http://www.genome.gov/sequencingcosts/)

What will change in the near future ???

- We are in the million-genome era
  - several projects now aiming to accrue over 1 million participants.
- By 2025, the annual acquisition of genomic data is anticipated to exceed 2 exabytes (2 million terabytes) and could be considerably higher.
- Getting the most from these data will require robust infrastructure and tools for large-scale analysis of multi-omic datasets.

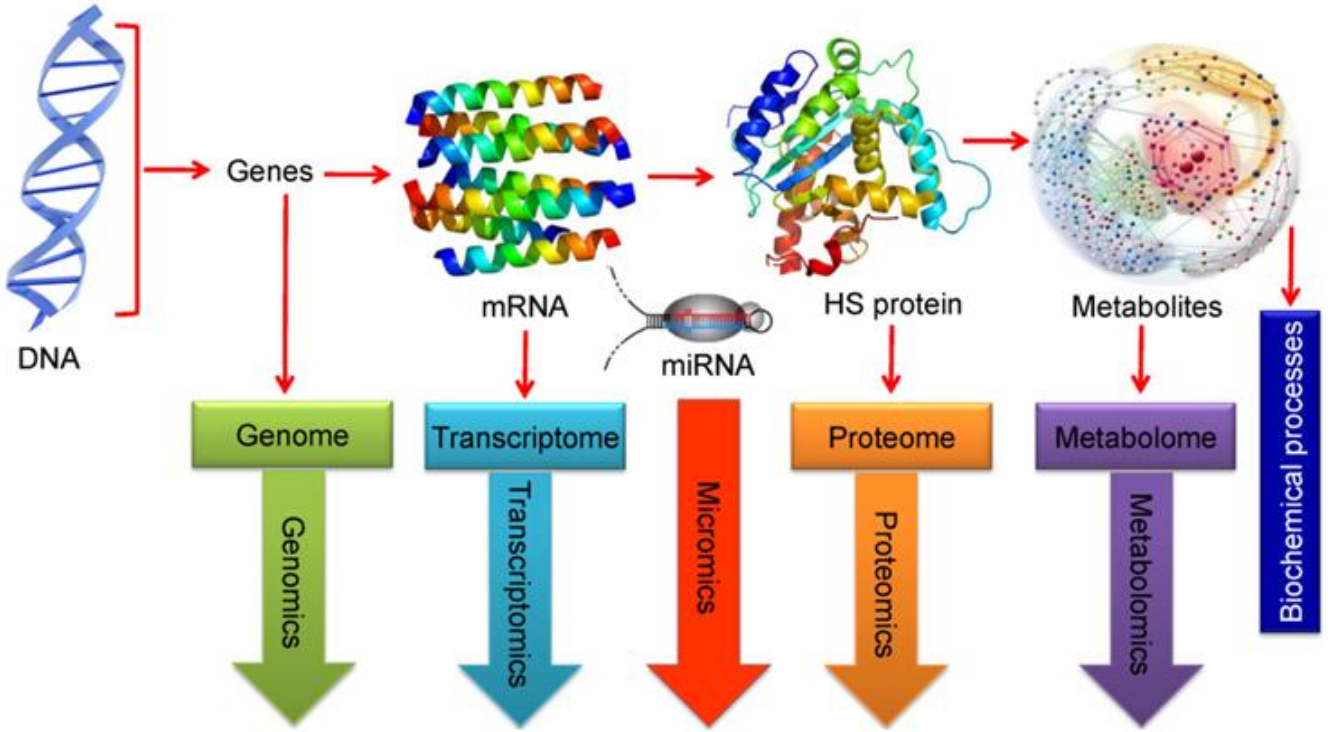




Omics Data

- A field of study in biology ending in -omics, such as genomics, proteomics or metabolomics etc.

- Omics Data Explosion Challenges → Heterogeneous Data Integration, Big Data Analysis etc.



What will happen in few years ???

- In 10 years a virtual cloud of billions of data points will surround each patient.
- These data will be of many different types and, accordingly, multistage.
- The challenge will be to convert these data into simple hypotheses about health and disease for the individual.

Rambam Maimonides Med J (2013)

Outline

i. Systems Biology and Extensions

ii. Network-Based Approaches

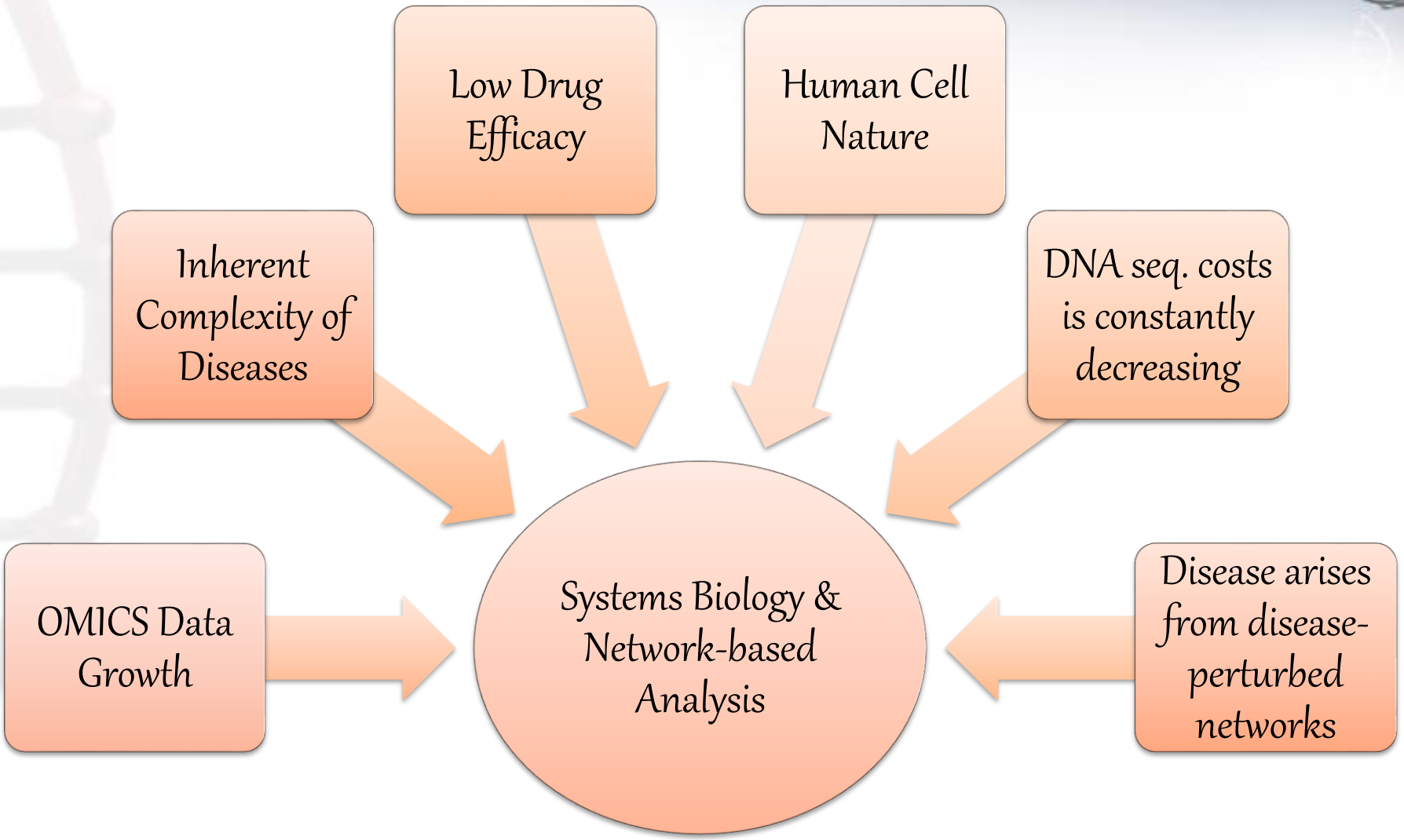
iii. Subpathway Approaches

- CHRONOS tool
- DEsubs tool

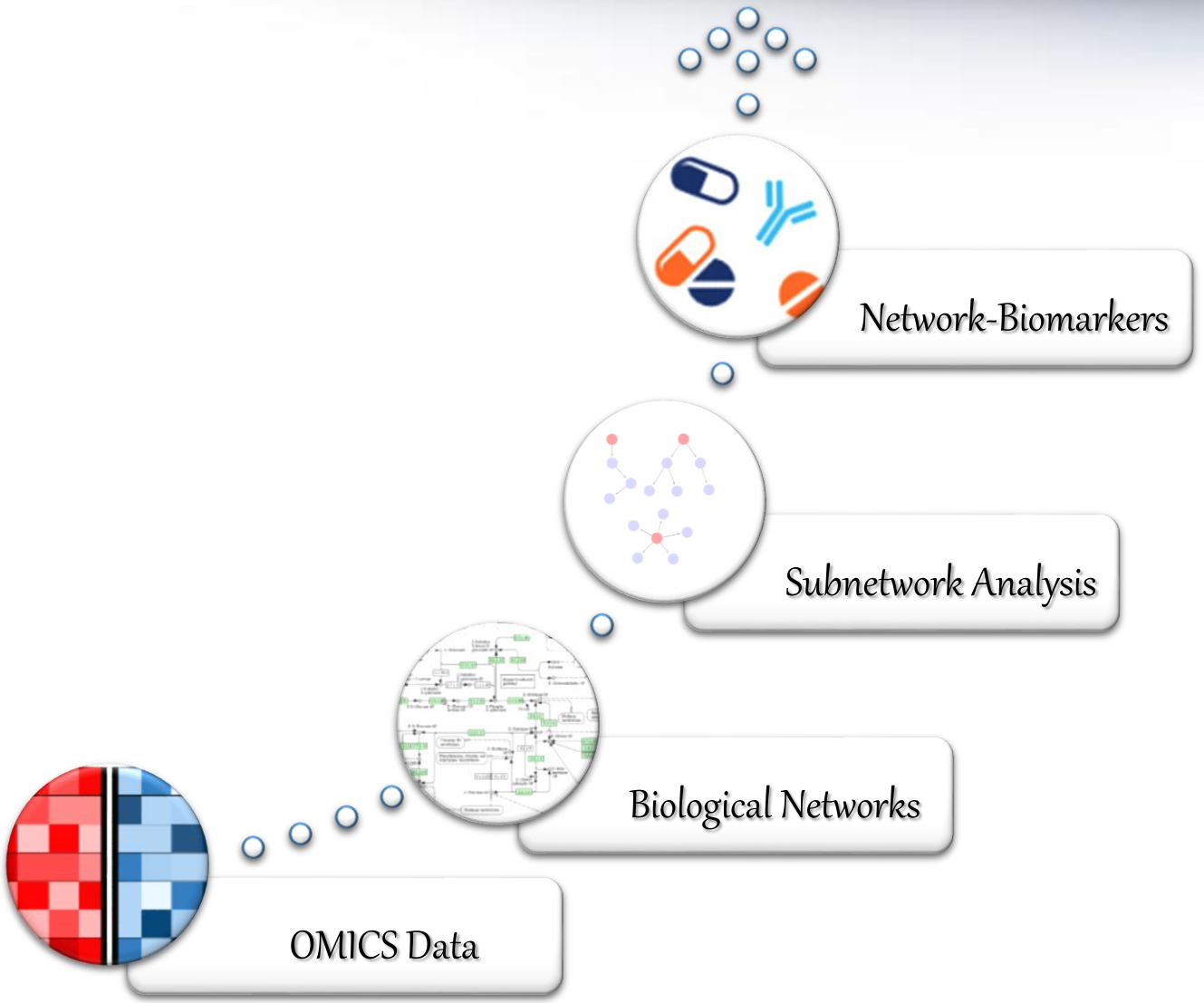
How all of the above are combined with Networks ?

- Disease arises from disease-perturbed networks
- Studying the connections among diseases may reveal more information about the underlying mechanisms (*PLoS Comput Biol, 2009*).
- Future of pharmacology
  - to restore network dysfunction by simultaneously targeting key components in disease networks (*Drug Discovery Today, 2013*).
- Network-based analysis is the main family of Systems Biology methods
- The emerging network-based tools offer a platform to explore systematically
  - not only the molecular complexity of a particular disease,
  - but also the molecular relationships among apparently distinct (patho)phenotypes

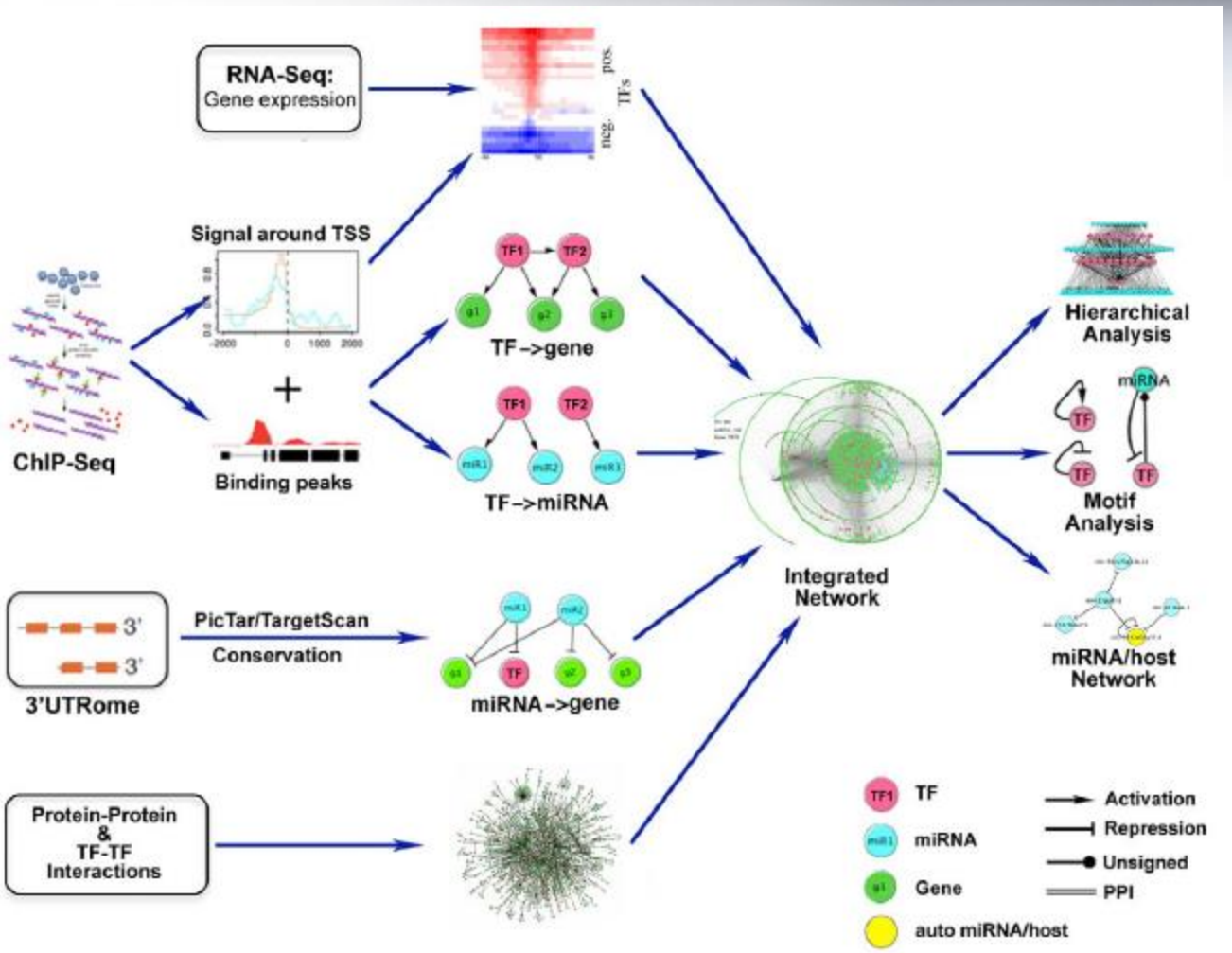
Systems Biology & Network Approaches



Network-based Analysis - Framework

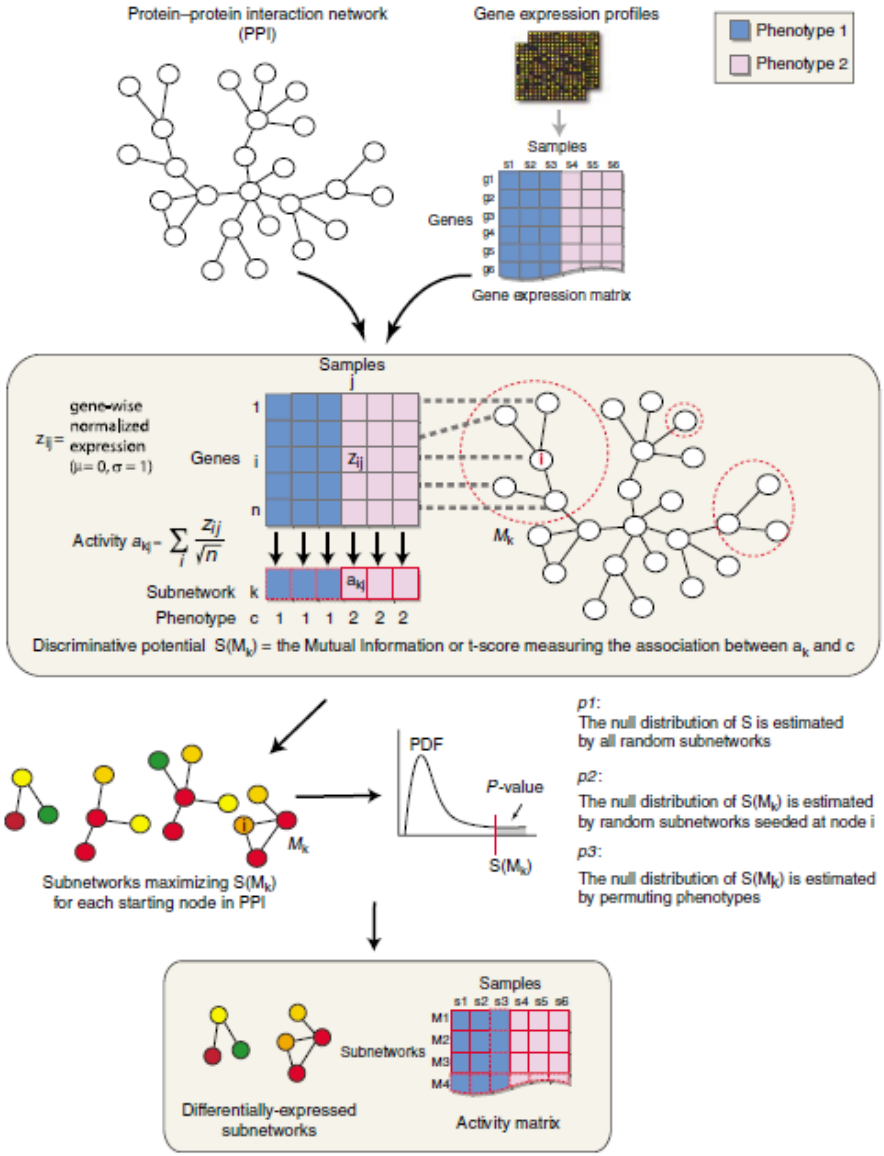


Network-based Approaches – Example



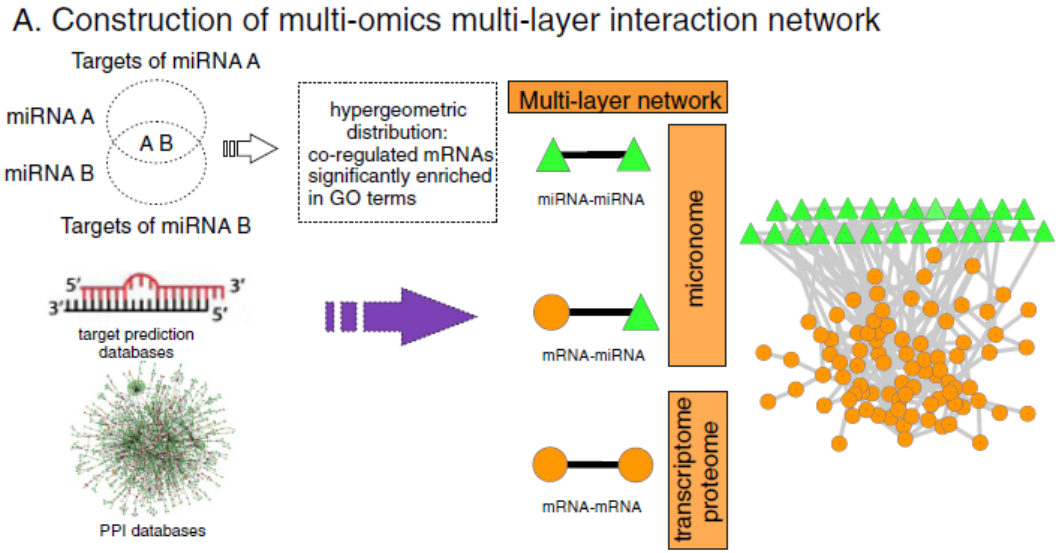
Cheng et al. (2011), Plos Computational biology

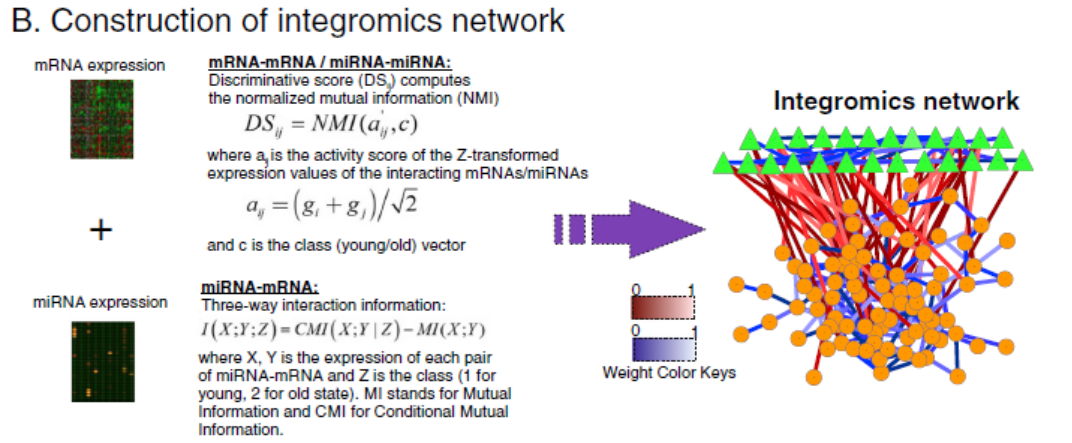
Network-based Approaches – Example

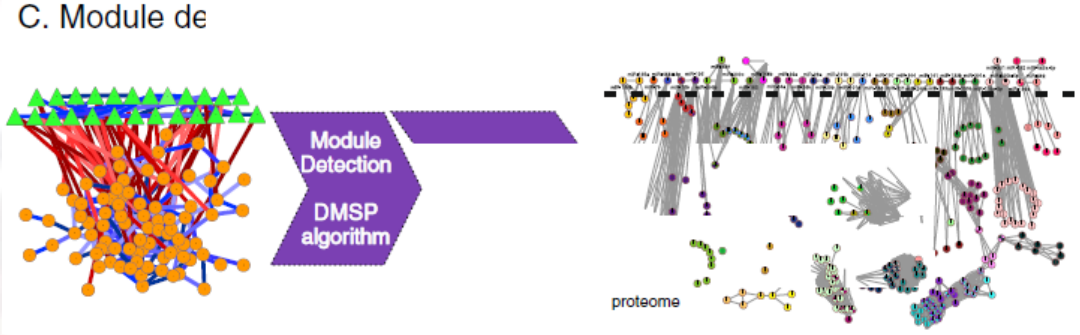


Chuang, Lee, Liu, Lee, Ideker (2011), Molecular Systems Biology

Network-based Approaches – Example

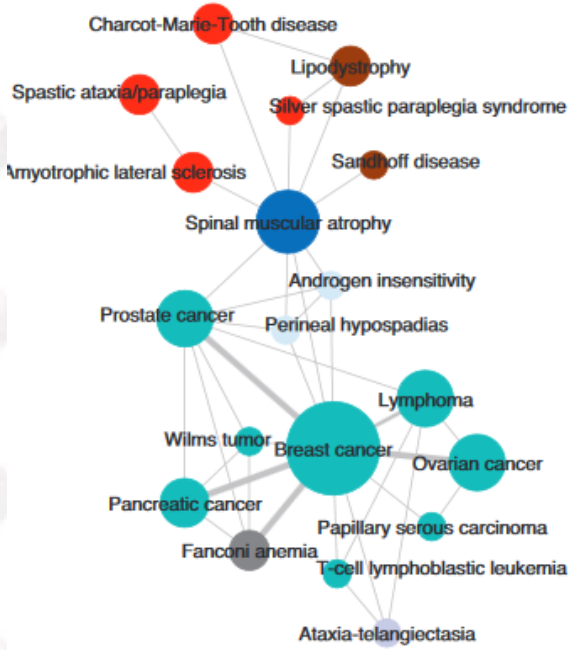




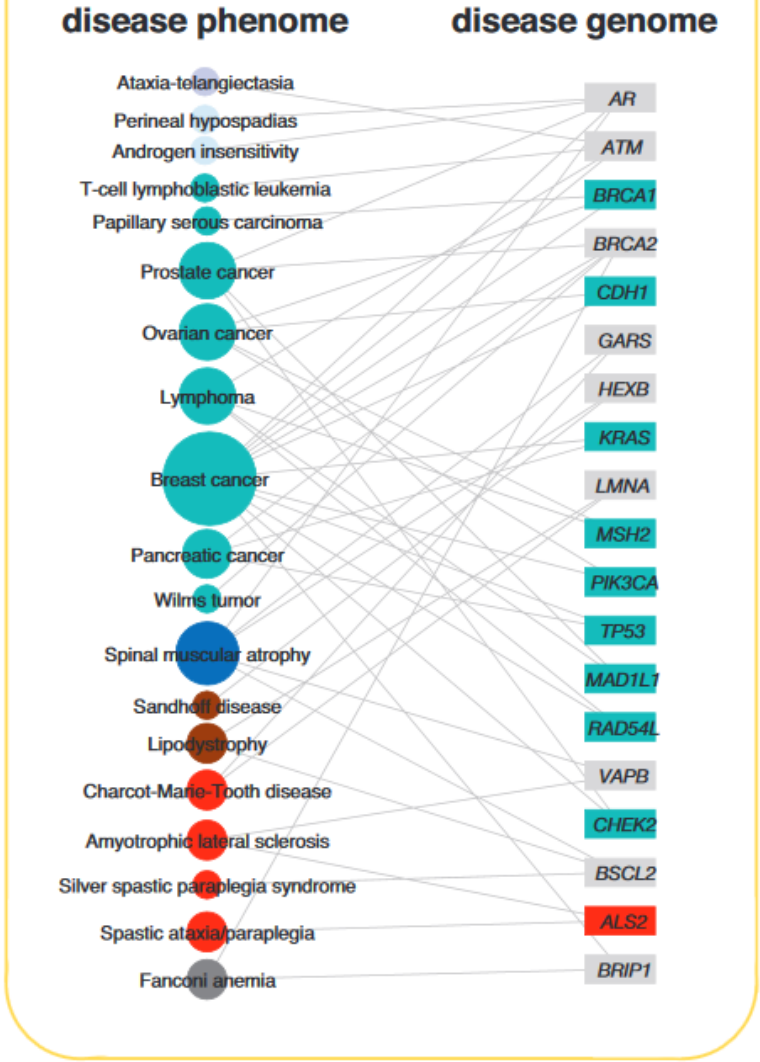


Dimitrakopoulou, Vrahatis, Bezerianos (2015), BMC Genomics

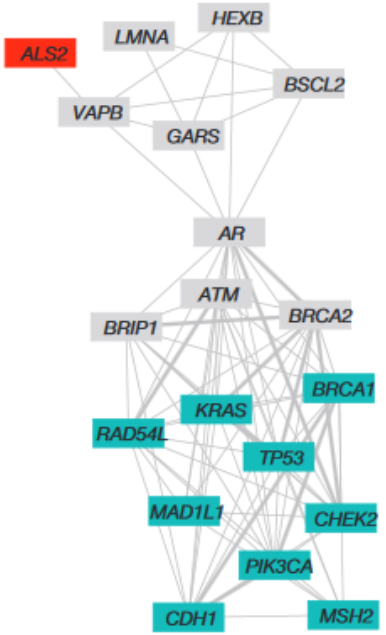
Human Disease Network (HDN)



DISEASOME

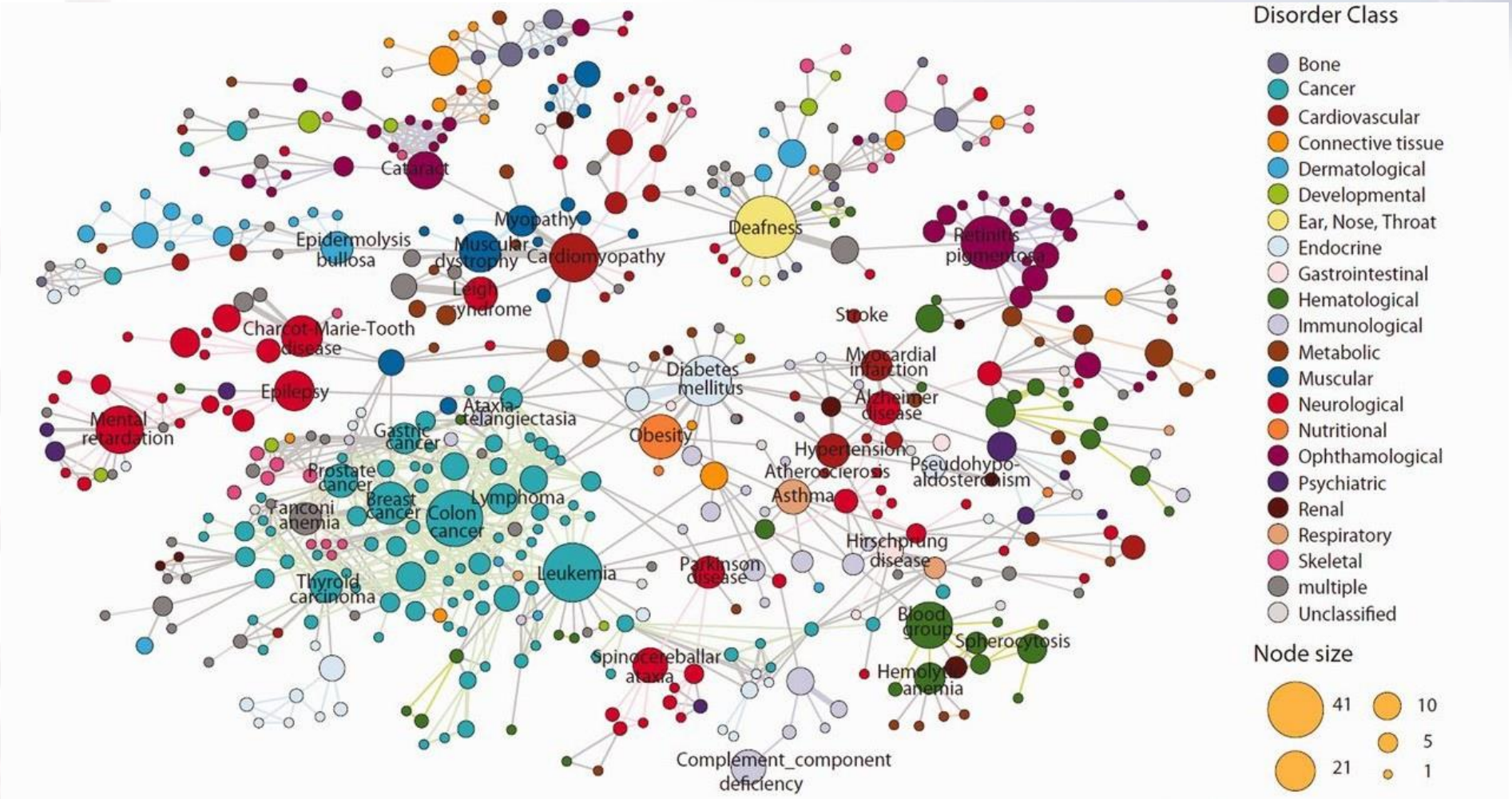


Disease Gene Network (DGN)



Goh, Cusick, Valle, Childs, Vidal, & Barabási, PNAS, (2007)

The first-published human disease network



Source: bfg.oxfordjournals.org

Outline

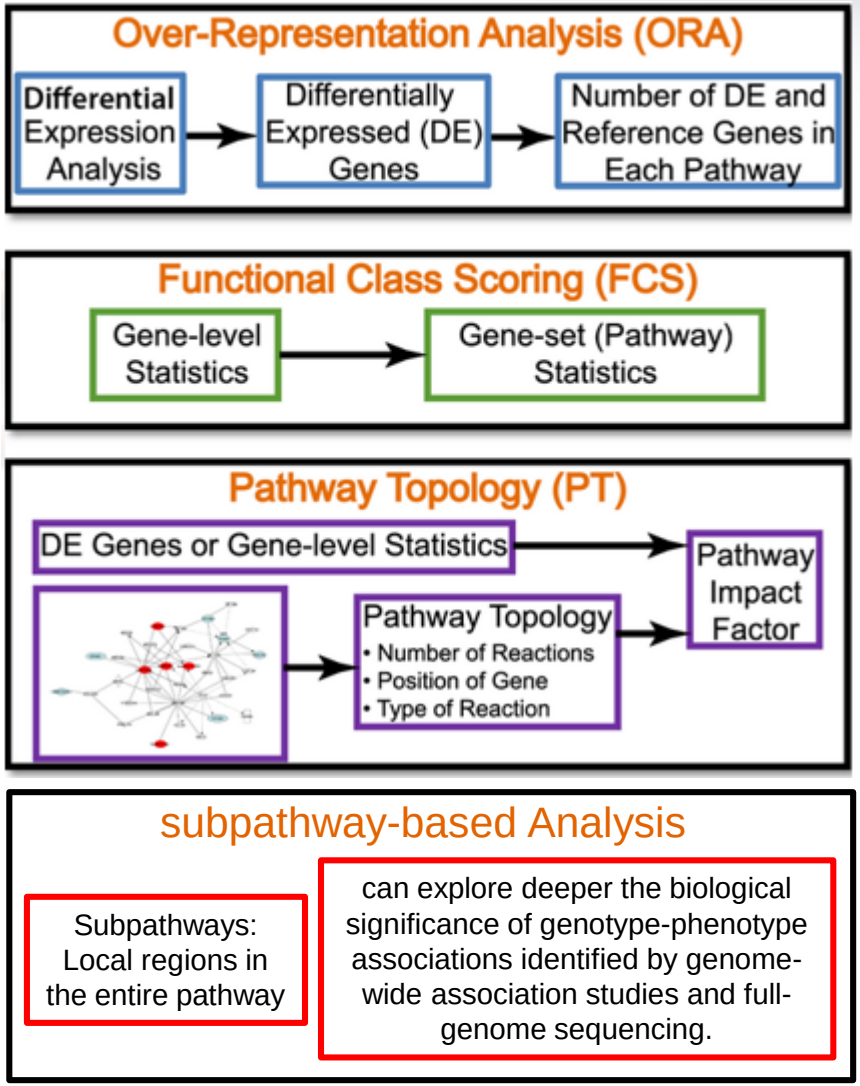
i. Systems Biology and Extensions

ii. Network-Based Approaches

iii. Subpathway Approaches

- CHRONOS tool
- DEsubs tool

Pathway Analysis Approaches



1st Generation

2nd Generation

3rd Generation

4rt Generation !!!

Khatri et al., 2012

Subpathway Approaches – WHY ???

- Recently, researchers proposed that key subpathway regions may represent the corresponding pathway and be more relevant for interpreting the associated biological phenomena [1].
- Several studies show that abnormalities in subpathway regions of metabolic pathways may contribute to the etiology of diseases [1,2]
- A pathway extracted from biological databases is the collection of specific sub-pathways or modules that perform certain functions.
- Different sub-pathways may perform the same function in the same pathway and different pathways may use the same sub-pathways in similar roles.
- Testing the whole pathway is too universal to determine which individual sub-pathways respond to a particular biological condition.

1. Li C, Li X, Miao Y, Wang Q, Jiang W, Xu C, et al. SubpathwayMiner: a software package for flexible identification of pathways. Nucleic Acids Res. 2009; 37(19):e131.
2 Li X, Li C, Shang D, Li J, Han J, Miao Y, et al. The implications of relationships between human diseases and metabolic subpathways. PLoS One. 2011; 6(6):e21131.

Outline

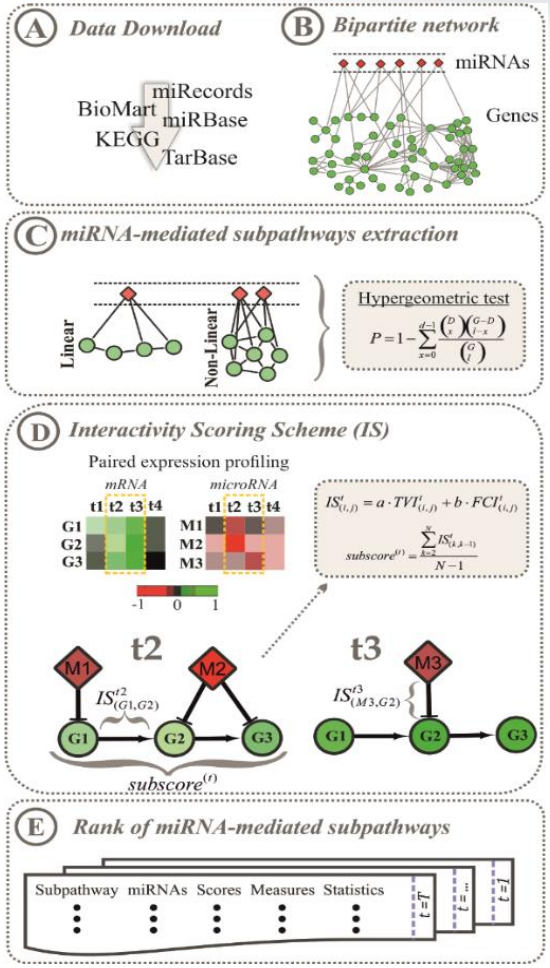
i. Systems Biology and Extensions

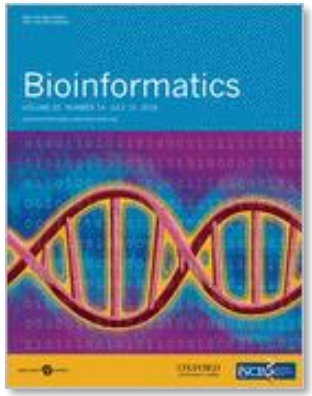
ii. Network-Based Approaches

iii. Subpathway Approaches

- **CHRONOS tool**
- DEsubs tool

CHRONOS: A time-varying method for microRNA-mediated subpathway enrichment analysis





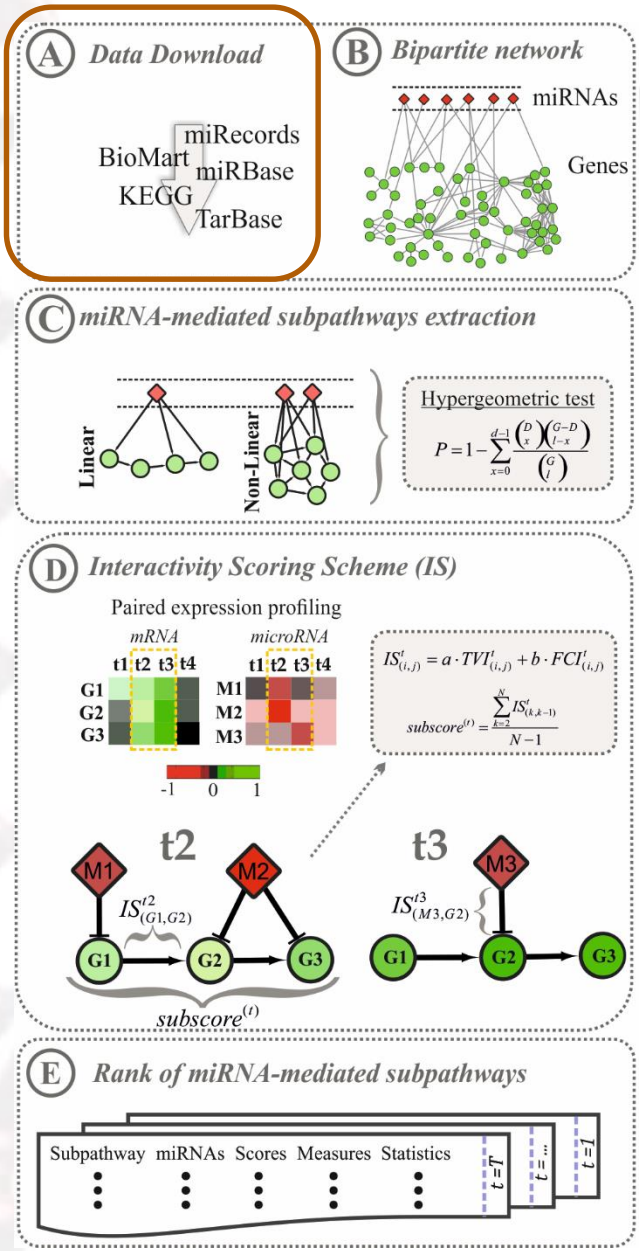
Aristidis .G. Vrahatis et al. (2016) CHRONOS: A time-varying method for microRNA-mediated subpathway enrichment analysis. *Bioinformatics*, 2016, 32(6).

CHRONOS: a time-varying method for microRNA-mediated subpathway enrichment analysis

- **Motivation:**
  - rapid growth of paired time series mRNA/microRNA expression experiments
  - An approach able to capture the time-specific 'active parts' of the biological circuitry
  - microRNA impact in group of nodes

- **Idea:**
  - Extraction of time-dependent subpathways
  - CHRONOS Development
    - integrates time series mRNA/microRNA expression data with
    - KEGG pathway maps and microRNA-target interactions
  - Our key assumption is that during the time course, different subareas of the pathway topology are active

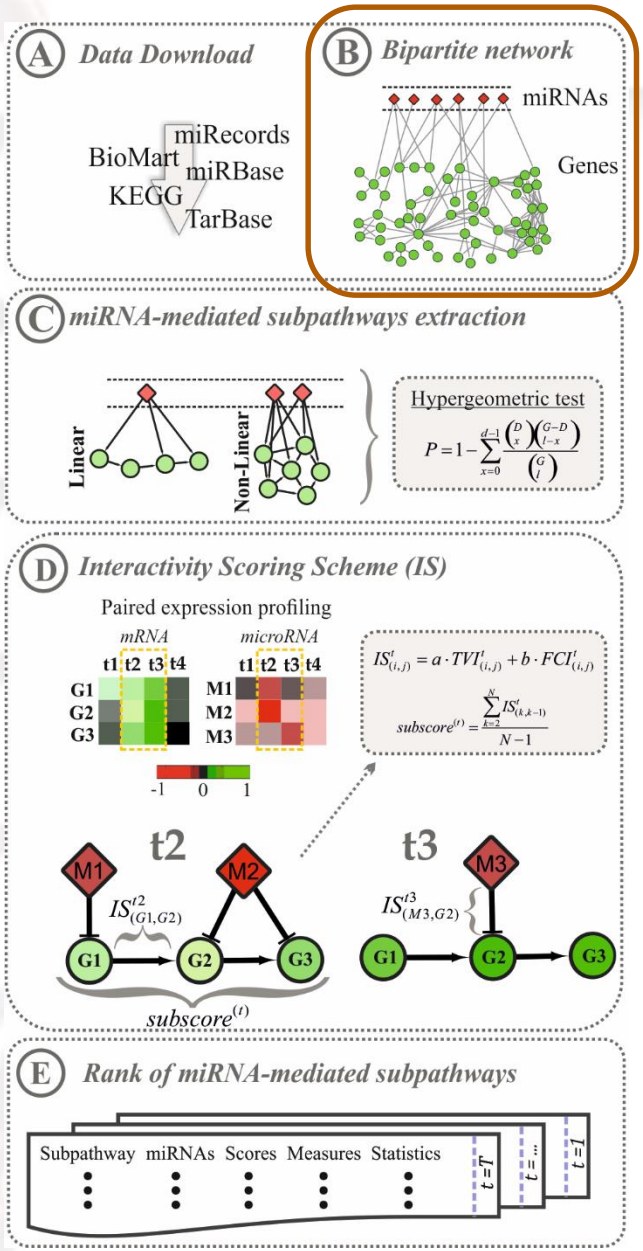
CHRONOS – Data



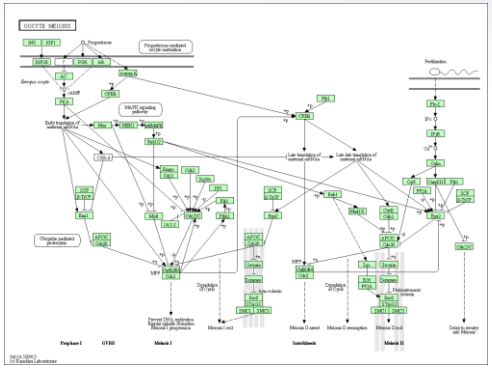
- **User Input**
  - mRNA/miRNA expression data
- **Online Data**

KEGG	Gene-Gene Pathway Graph
miRecords	miRNA/mRNA interactions
TarBase	miRNA/mRNA interactions
mirBase	miRNA Nomenclature
BioMart	Gene Nomenclature

CHRONOS – Network Construction



KEGG maps with Signaling Pathways



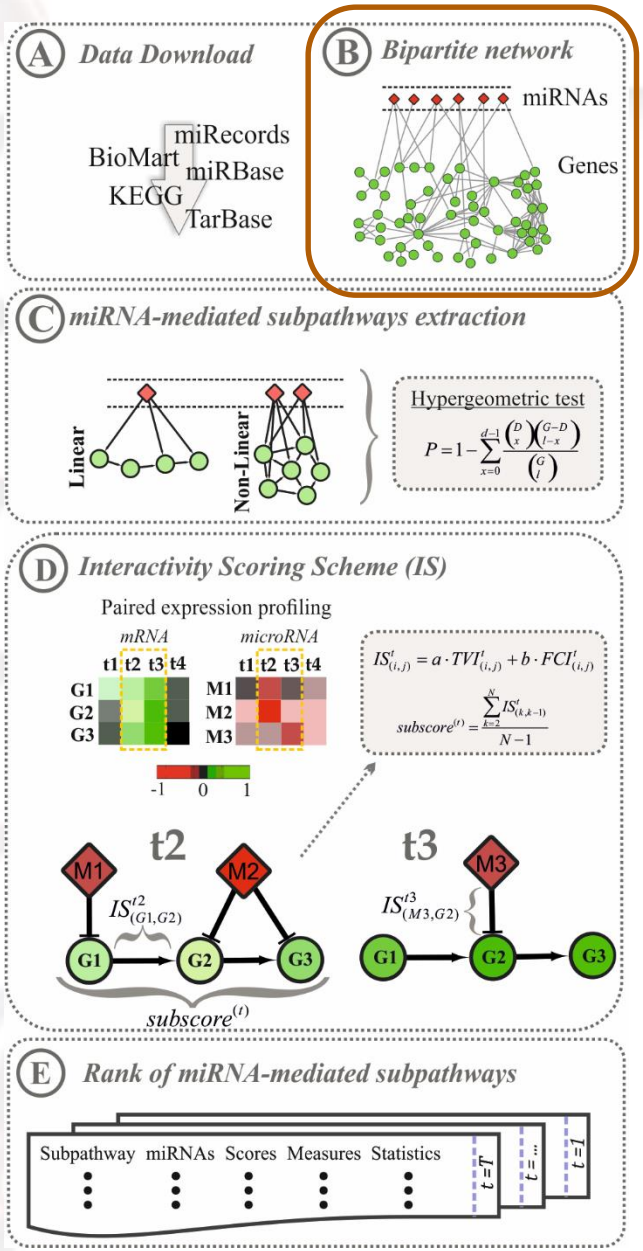
Nodes

- Proteins
- Compounds
- Genes
- Protein Complex
- Enzymes
- …

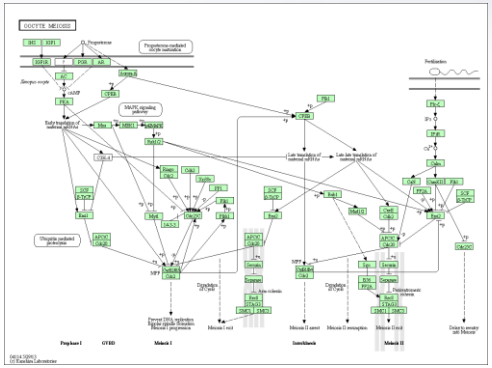
Edges

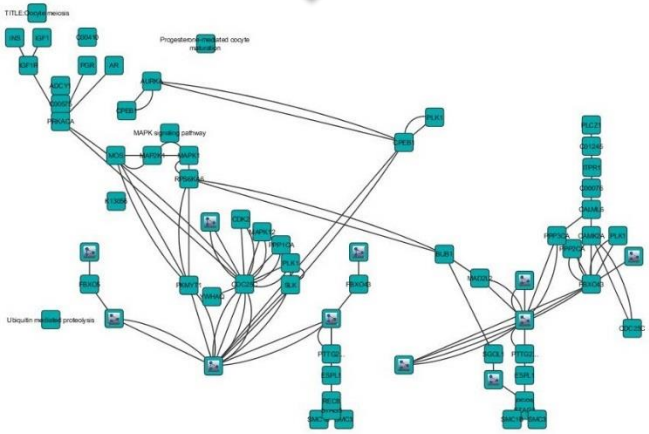
- Activation
- Inhibition
- Dissociation
- Phosphorylation
- Ubiquitination
- …

CHRONOS – Network Construction



KEGG maps with Signaling Pathways

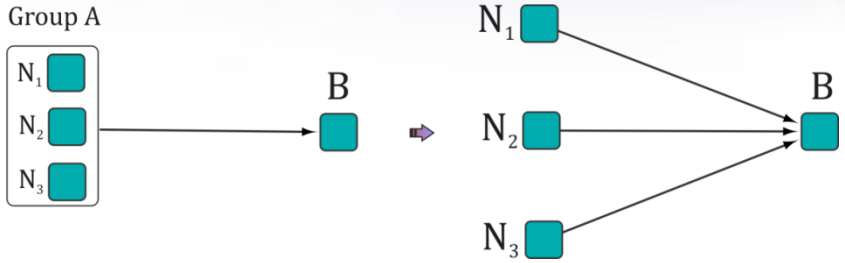


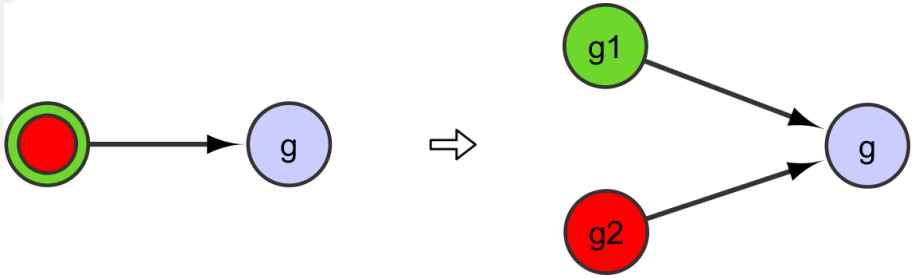


Gene-Gene Network

CHRONOS – Network Construction

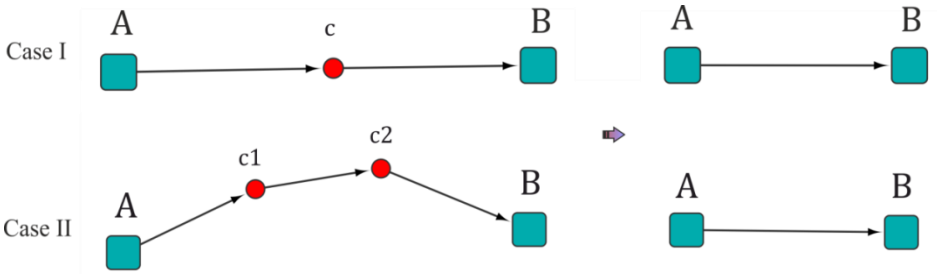
- **Group Nodes with**
  - Protein complexes
  - Multiple gene products





Node with multiple gene products

- **Chemical compounds removal**

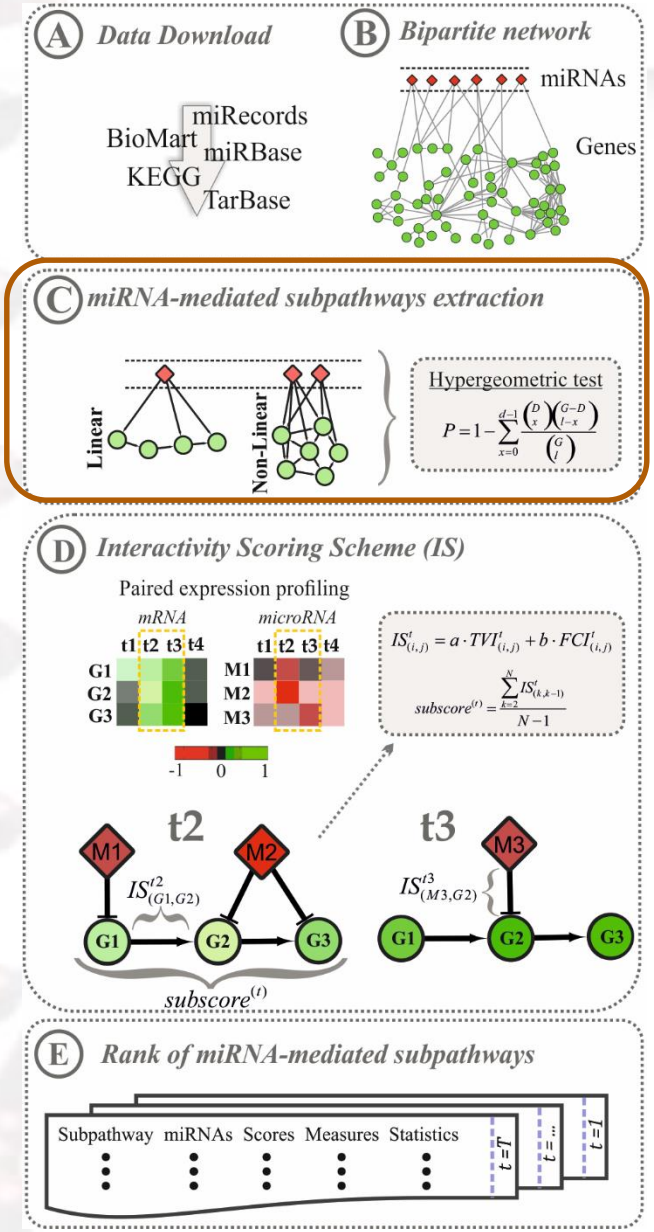


CHRONOS – Network Construction

- User-defined categorization

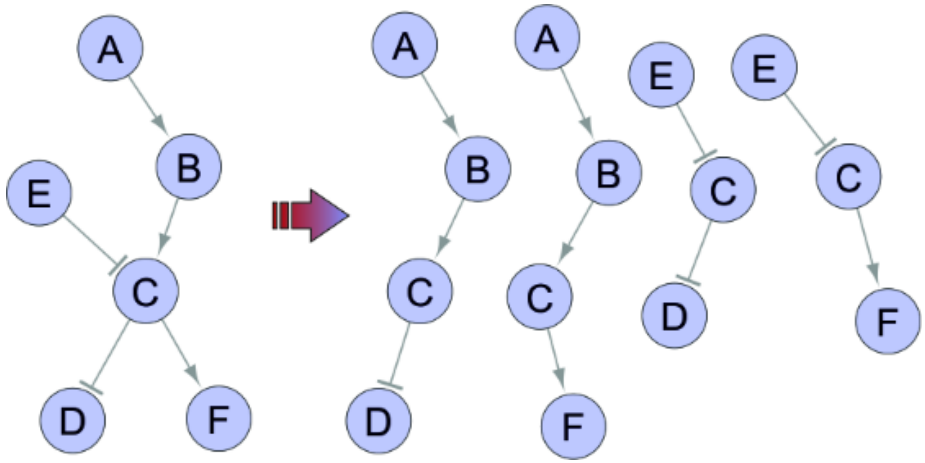
Interaction Type	
expression	Activation
activation + phosphorylation	Activation
activation + methylation	Activation
activation + binding/association	Activation
inhibition	Inhibition
repression	Inhibition
inhibition + methylation	Inhibition
ubiquitination + inhibition	Inhibition
binding/association	Unknown
activation + inhibition	No-Interaction
.....	.....

CHRONOS – Subpathway Extraction

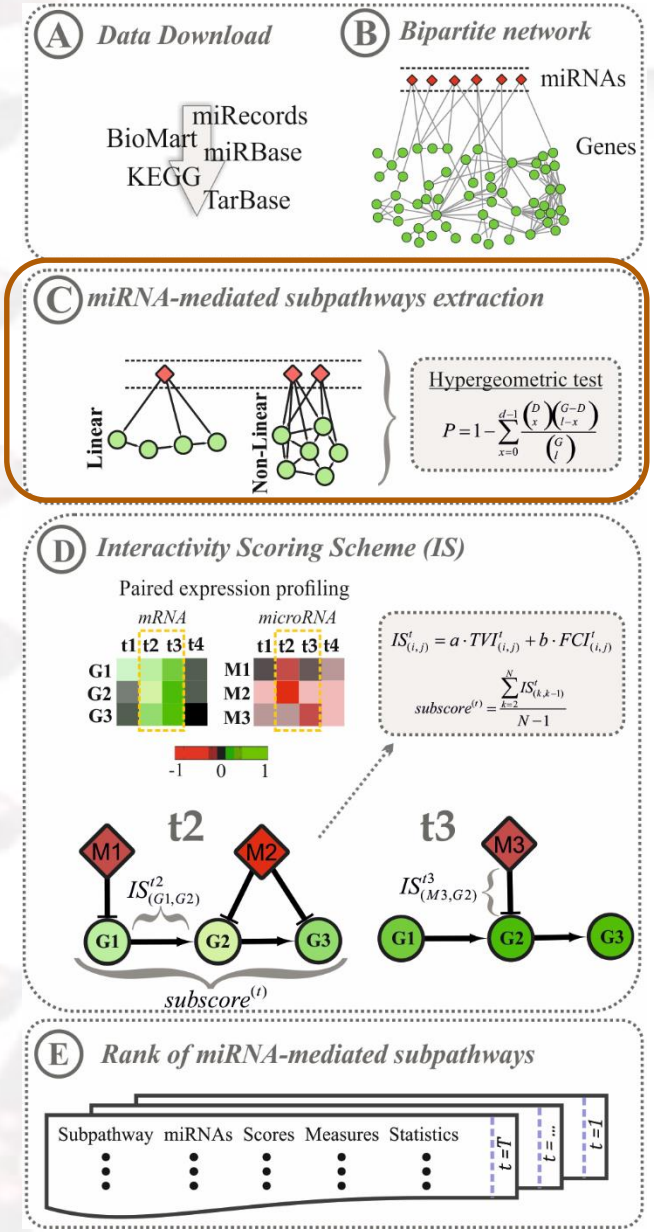


Linear Subpathways

– all ordered linear sequences of gene interactions from biologically meaningful start-nodes to end-nodes

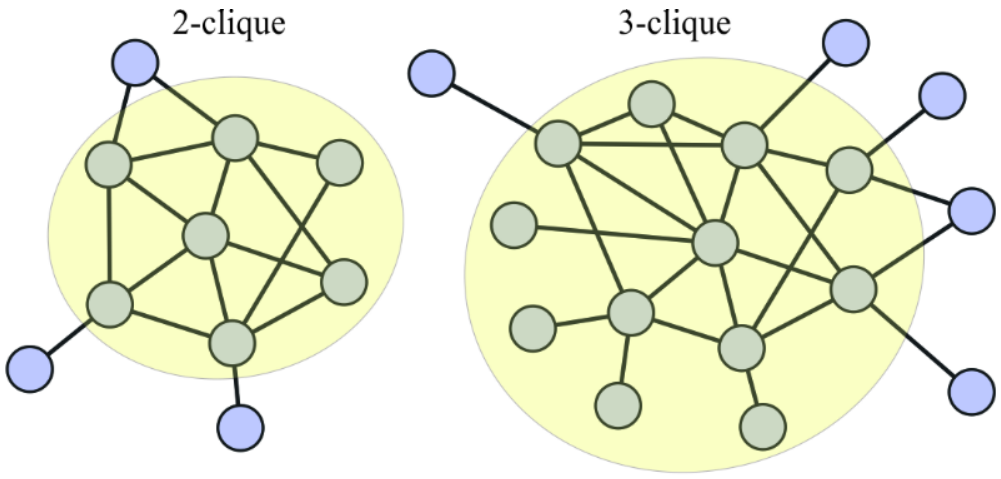


CHRONOS – Subpathway Extraction

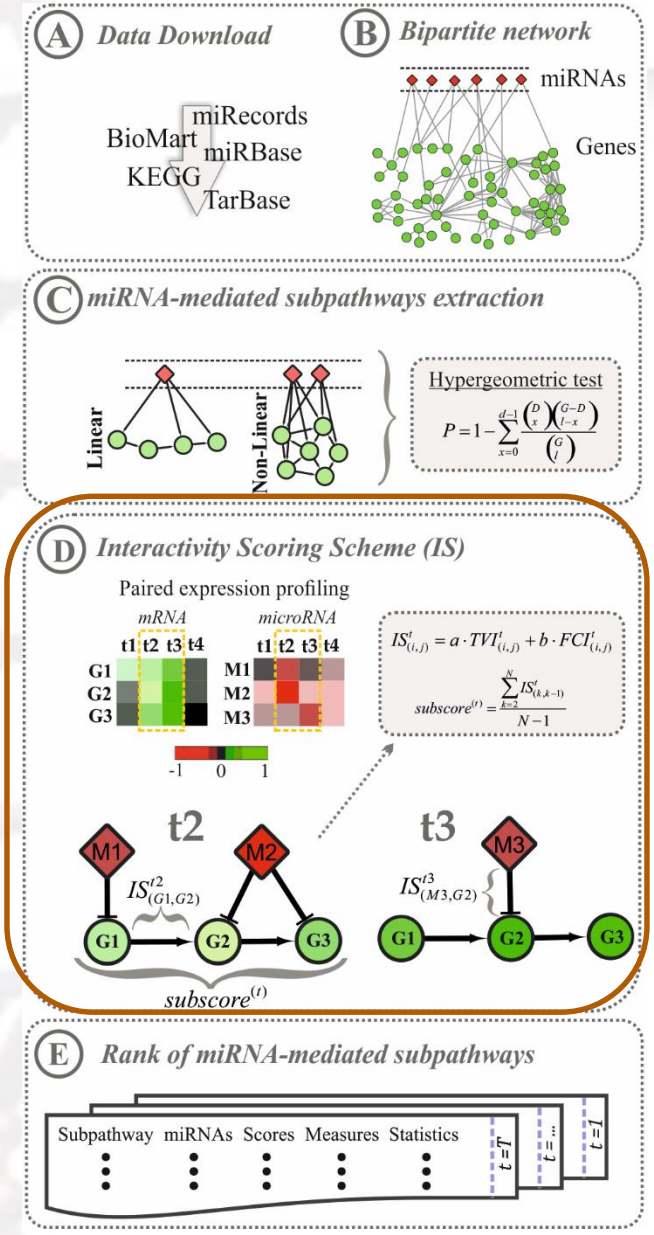


non-linear subpathways

- Biological assumption:
  - the functional similarity between two genes increases as their distance in pathways decreases
- k-clique: a subgraph where the distance between any two nodes is no greater than k.



CHRONOS – Evaluation of Subpathways



— For each edge e, at time t, IS weight is:

$$IS_e^t = a \cdot |TVI_e^t| + b \cdot |FCI_e^t|$$

Fold Change Interactivity (FCI) score*

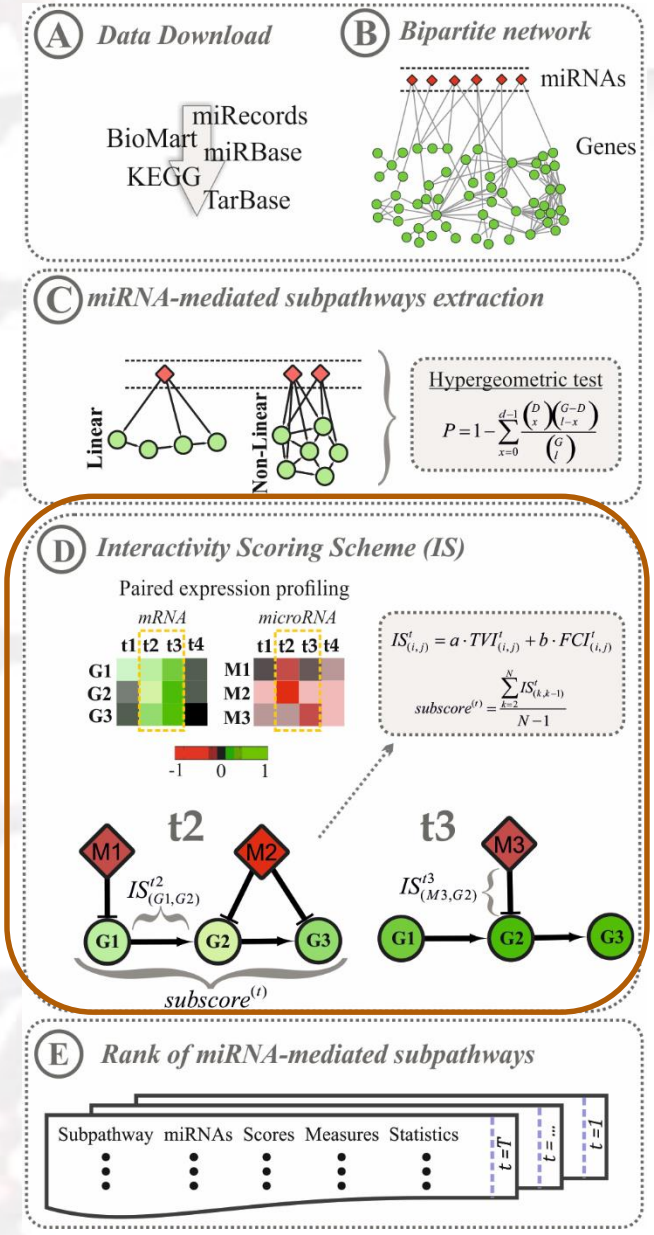
— captures pairs of nodes with high absolute fold change values and with high positive or negative correlation

$$FCI_e^t = \left(1 + c \sum_{k=(i,j)\in e} e^{-K(f_k^t - T)}\right)^{-1} - \left(1 + c \sum_{k=(i,j)\in e} e^{-K(-f_k^t - T)}\right)^{-1}$$

- f_i^t , f_j^t are the log2-fold change values of nodes i and j
- C, K denote the parameters controlling the shape of the
- multivariate logistic distribution
- T is a shifting parameter

* Kim, Y. et al. (2011). Principal network analysis: identification of subnetworks representing major dynamics

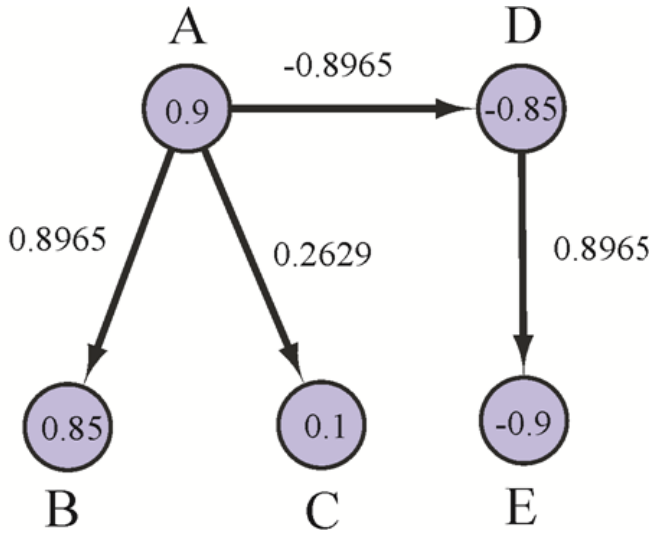
CHRONOS – Subpathway Evaluation



— For each edge e, at time t, IS weight is:

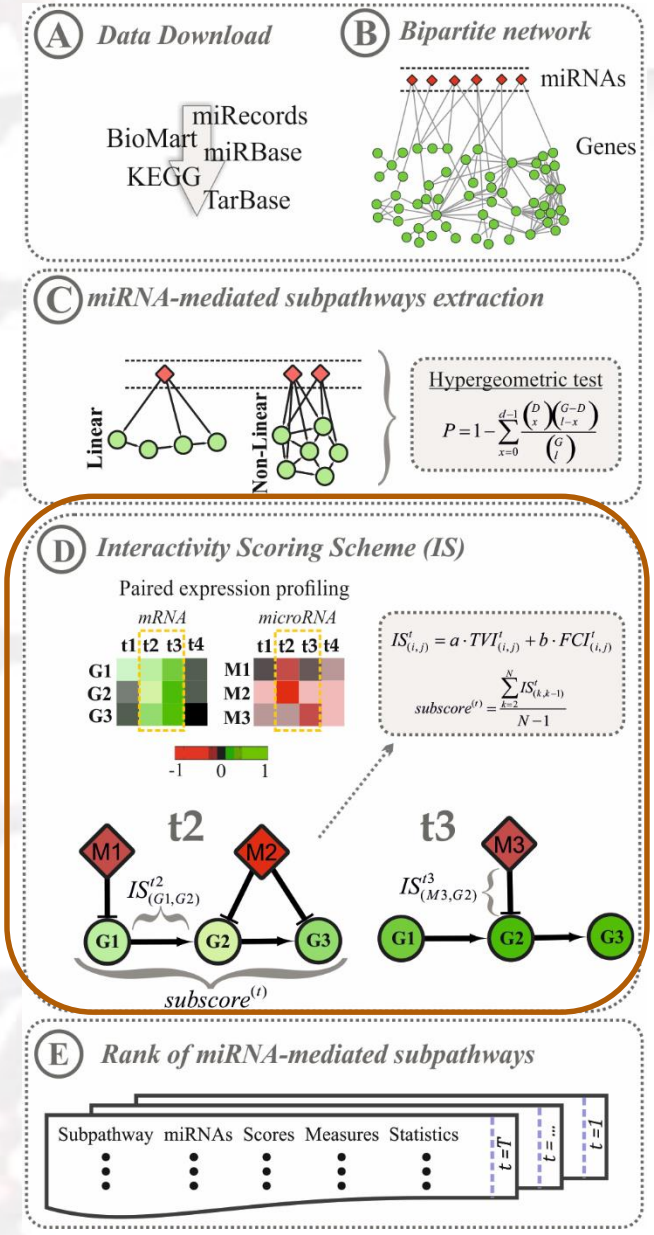
$$IS_e^t = a \cdot |TVI_e^t| + b \cdot |FCI_e^t|$$

Fold Change Interactivity (FCI) score



FCI_e^t	Genes Correlation Type
-1	High negative correlation
0	No correlation
+1	High Positive correlation

CHRONOS – Subpathway Evaluation



— For each edge e, at time t, IS weight is:

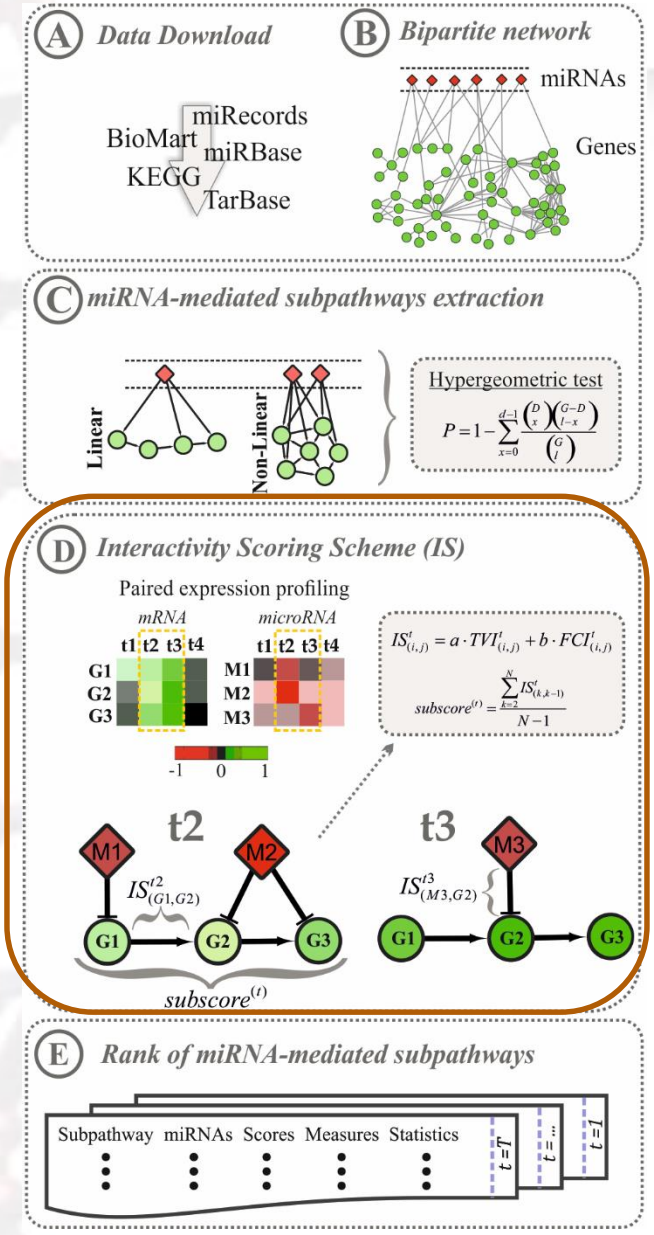
$$IS_e^t = a \cdot |TVI_e^t| + b \cdot |FCI_e^t|$$

Time-Varying Interactivity (TVI) Score*

TVI_e^t	Genes Correlation Type
-2	High negative correlation
-1	Weak negative correlation
0	No correlation
+1	Weak Positive correlation
+2	High Positive correlation

* Jethava, et al. (2011). Netgem: Network embedded temporal generative model for gene expression data. BMC bioinformatics, 12(1), 1.

CHRONOS – Subpathway Evaluation



— For each edge e, at time t, IS weight is:

$$IS_e^t = a \cdot |TVI_e^t| + b \cdot |FCI_e^t|$$

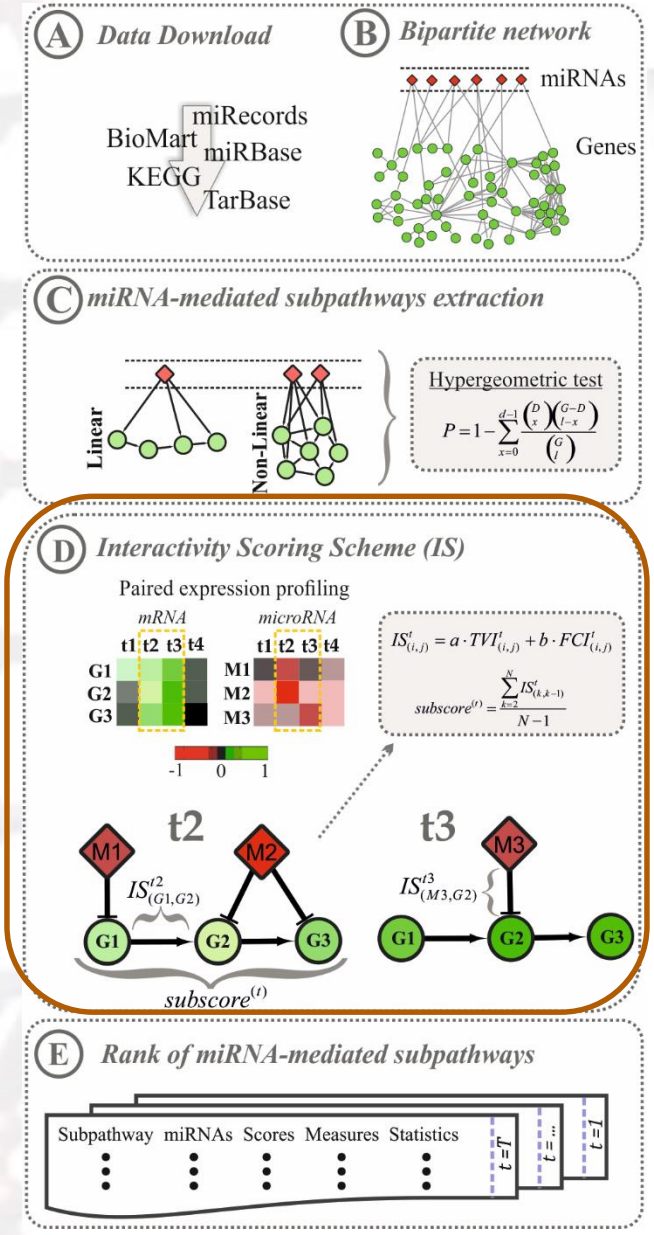
Acceptance Rule

IF edge type = 'activation' AND FCI, TVI Indicate 'activation' **THEN** APPROVED

IF edge type = 'inhibition' AND FCI, TVI Indicate 'inhibition' **THEN** APPROVED

IF ALL THE OTHER CASES **THEN** REJECTED

CHRONOS – Subpathway Evaluation



— For each edge e, at time t, IS weight is:

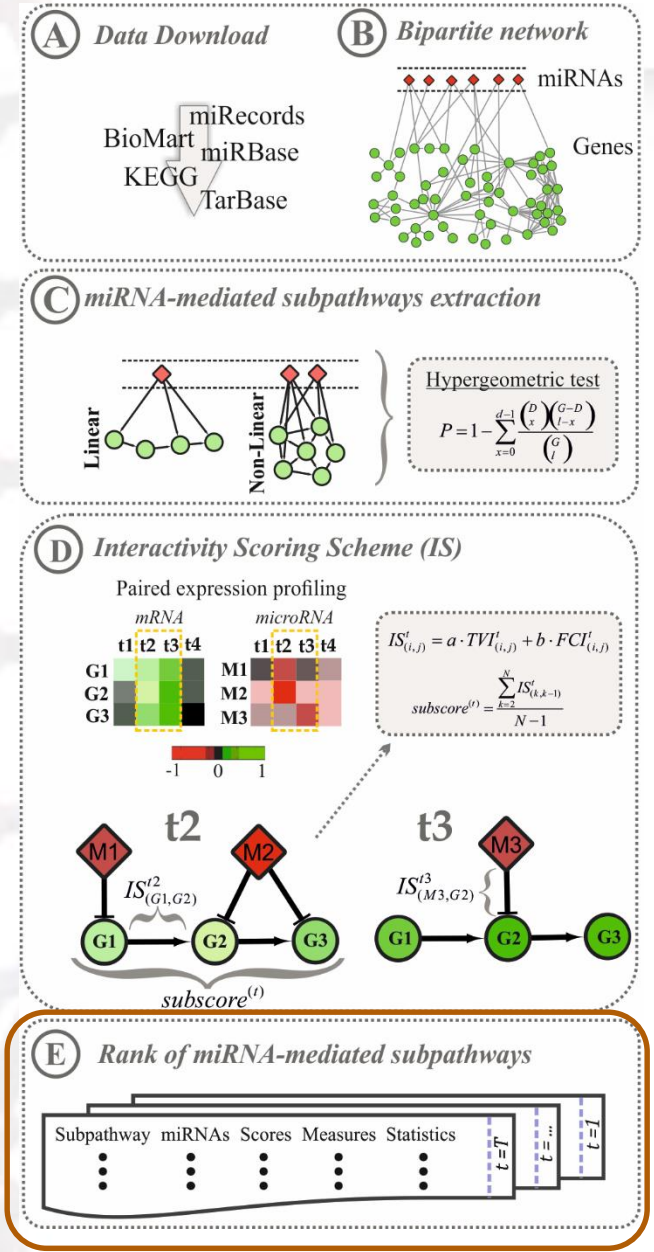
$$IS_e^t = a \cdot |TVI_e^t| + b \cdot |FCI_e^t|$$

Subpathway Score

- For a subpathway with N gene members and N-1 interactions at time t, subscore is:

$$subscore^t = \frac{1}{N-1} \sum_{e=1}^{N-1} IS_e^t$$

CHRONOS – Visualization and Output Results



List of Output Results

		Subscores					Measures			
Sub Id	Pathway	T1	T2	T3	T4	T5	q-value	subPathness	subBC	subDEG
S1	(hsa04630) Jak-STAT signaling pathway	0,951	0,961	0,29	0,442	0,47	0.01	0,1	0,1	0,1
S19	(hsa05215) Prostate cancer	0,435	0,39	0,367	0,628	0,242	0.02	0,1	0,1	0,3
S307	(hsa04622) RIG-I-like receptor signaling pathway	0,442	0,449	0,099	0,551	0,373	0.03	0,1	0,1	0,3
S5596	(hsa04910) Insulin signaling pathway	0,31	0,301	0,229	0,414	0,259	0.01	3,8	2,2	0,3
S13937	(hsa04910) Insulin signaling pathway	0,129	0,056	0,316	0,419	0,193	0.05	0,3	0,5	0,4

	miRNA Mediated Subpathway Members		
Sub Id	Members (HGNC Symbol)	Members (EntrezGene ID)	miRNA
S1	PIAS3 STAT2 PIM1	10401 6773 5292	hsa-miR-18a(t1,t3,t5)
S19	PDGFD PDGFRB PIK3R3 AKT1 FOXO1	80310 5159 8503 207 2308	hsa-miR-19b(t1,t4) hsa-miR-29a(t4)
S307	TRIM25 DDX58 ATG5 MAVS TRAF3 TANK IKBKG	3665 7706 23586 9474 57506 7187 10010 8517 3665	hsa-miR-671-5p(t5)
S5596	INPPL1 PDPK1 AKT2 PPP1R3E CALM1 PYGM	3636 5170 208 90673 801 5837	hsa-miR-196a(t4) hsa-miR-19b(t4)
S13937	SOCS4 INSR IRS1 PIK3CA PRKCZ SLC2A4	122809 3643 3667 5290 5590 6517	hsa-miR-196a(t3) hsa-miR-27b(t4,t5)

- For each time
  - ✓ *subscore* value
  - ✓ FDR corrected P-values
  - ✓ Values of 3 topological measures
  - ✓ miRNA interactions

Outline

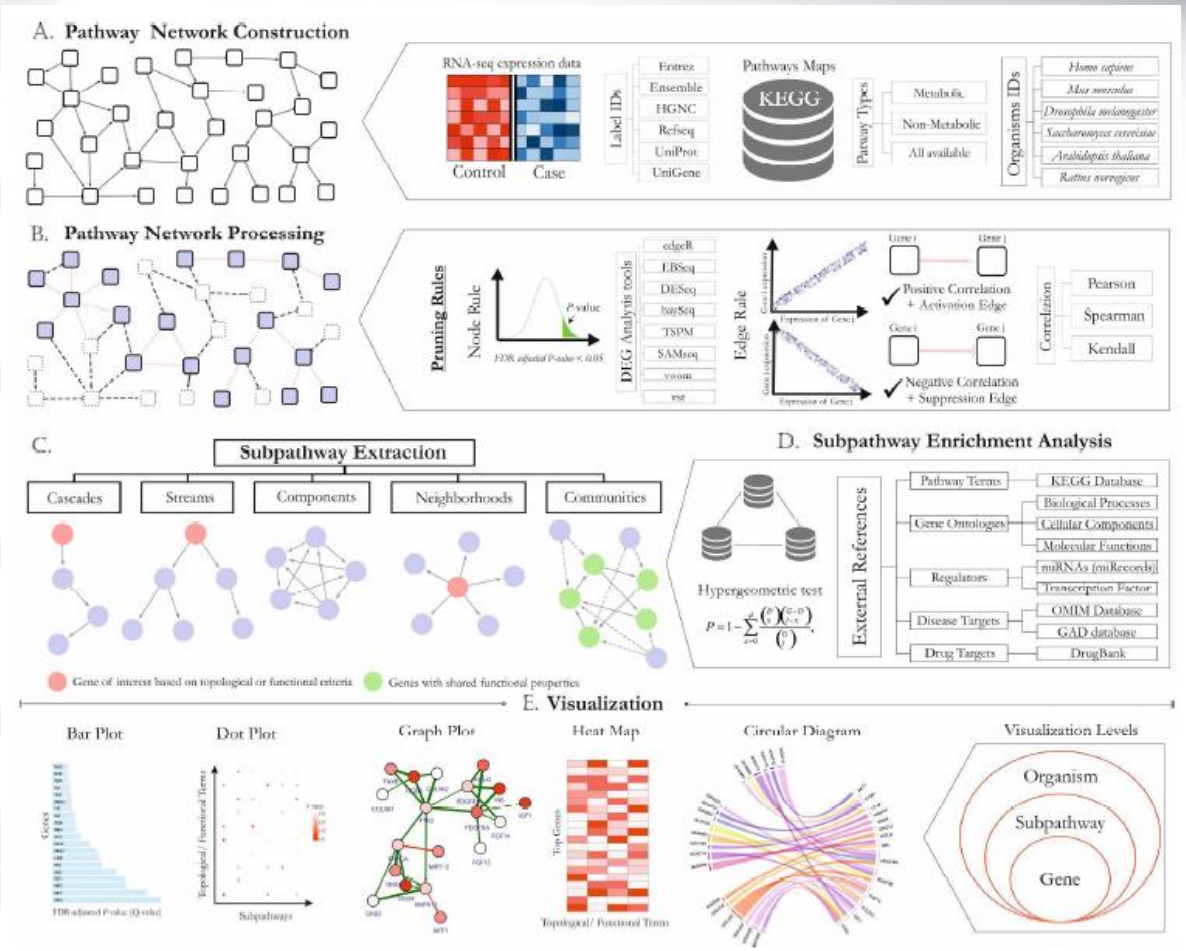
i. Systems Biology and Extensions

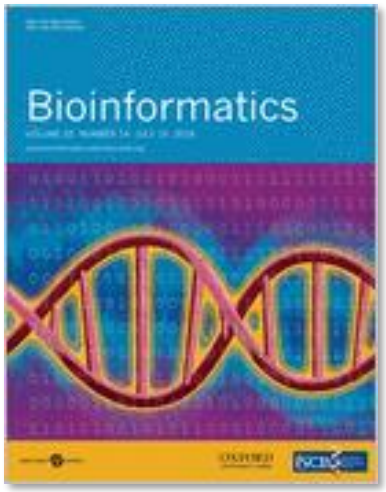
ii. Network-Based Approaches

iii. Subpathway Approaches

- CHRONOS tool
- **DEsubs tool**

DEsubs – Tool for the identification of differentially expressed subpathways using RNA-seq experiments





Aristidis G. Vrahatis, et al. (2016) DEsubs: an R package for flexible identification of differentially expressed subpathways using RNA-seq experiments. Bioinformatics, 32(24), 3844-3846.

DEsubs

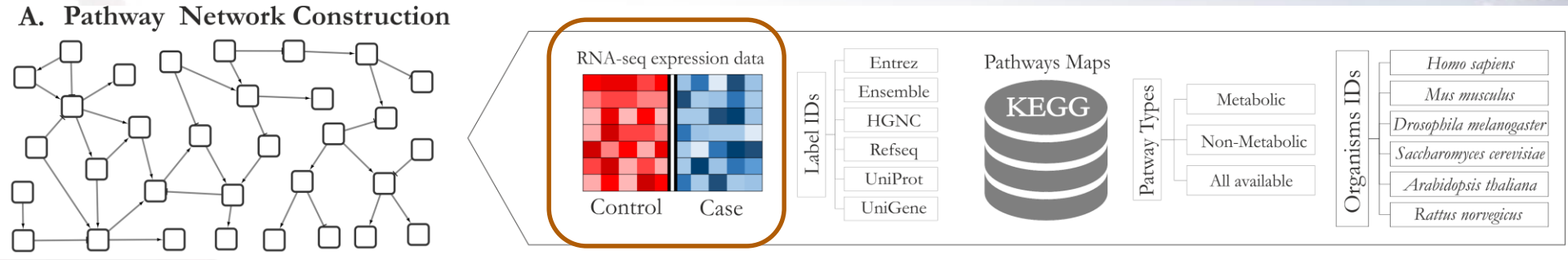
—Motivation

- systems-level network-based approaches have gained ground in the research field of systems biology
- RNA-seq transcriptome studies are increasing rapidly year by year

—Idea

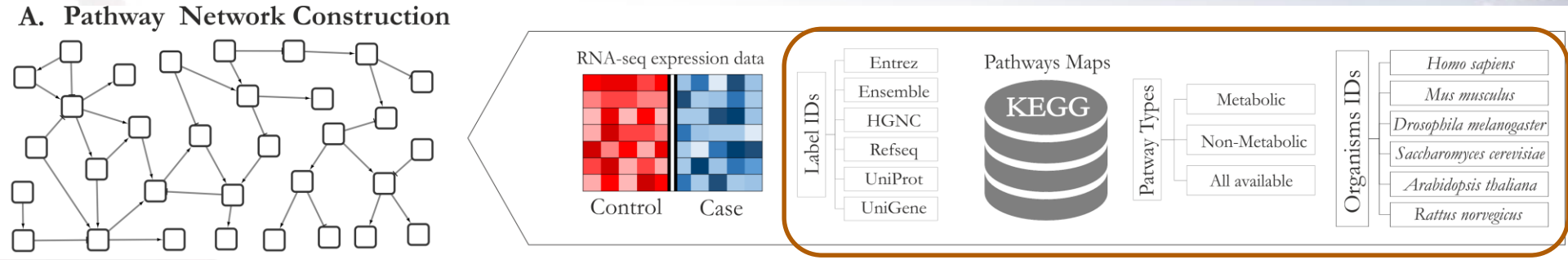
- Development of an R package, called DEsubs,
  - which extracts differentially expressed subpathways based on RNA-seq expression profiles
  - which enables a customized analysis to the problem under investigation through numerous operation modes.

Pathway network construction



	CONTROL STATE				DISEASE STATE			
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
Gene_1	1879	2734	2369	2636	2188	9743	9932	10099
Gene_2	24	40	22	27	31	118	108	144
Gene_3	3291	3259	3214	3407	3298	1058	960	679
Gene_4	97	124	146	114	126	33	19	31
Gene_5	485	485	469	428	475	128	135	103
Gene_6	113	92	64	96	137	39	16	23
Gene_7	886	687	771	786	768	3002	2768	2861
Gene_8	84	25	67	62	61	277	246	297
…	…	…	…	…	…	…	…	…
Gene_N	120	312	78	514	210	324	95	102

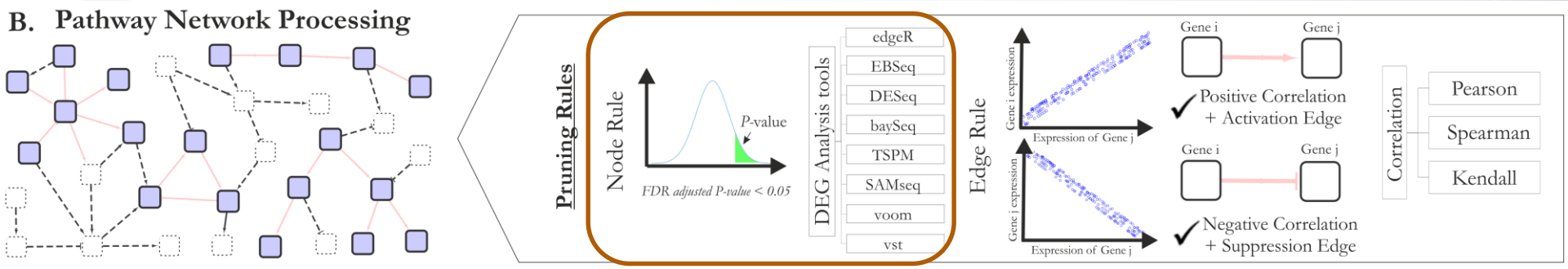
Pathway network construction



- Construction Process→ (Vrahatis et al., 2016)

- Supports
  - 6 Gene Nomenclatures
  - 3 KEGG Pathway Types
  - 6 Organisms

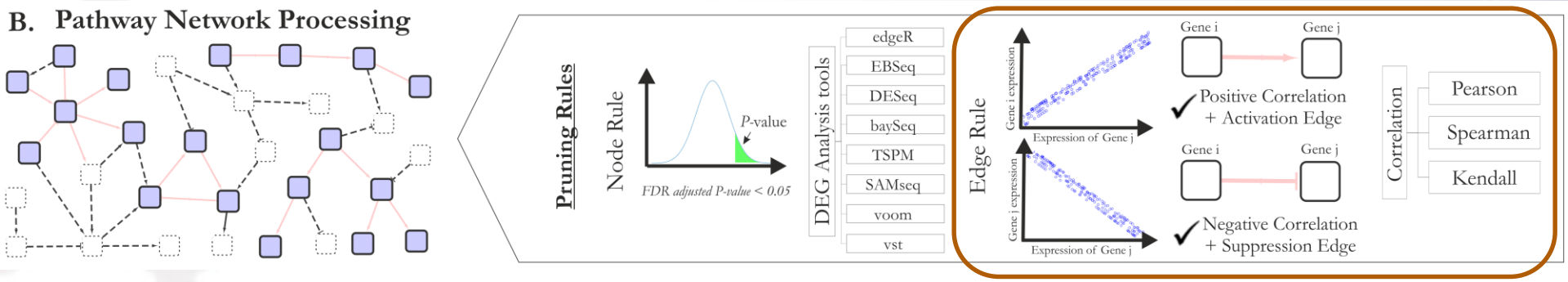
Pathway network processing - «Pruning Rules»



$$NodeRule:\ Qvalue(i) < \Theta, \qquad \forall i \in V(G)$$

- selects the core set of genes with statistically significant differential expression
- FDR-adjusted P-value ($Q\text{-}value$)

Pathway network processing - «Pruning Rules»



EdgeRule: $cor(i,j) * reg(i,j) > threshold, \quad \forall\,(i,j) \in E(G)$

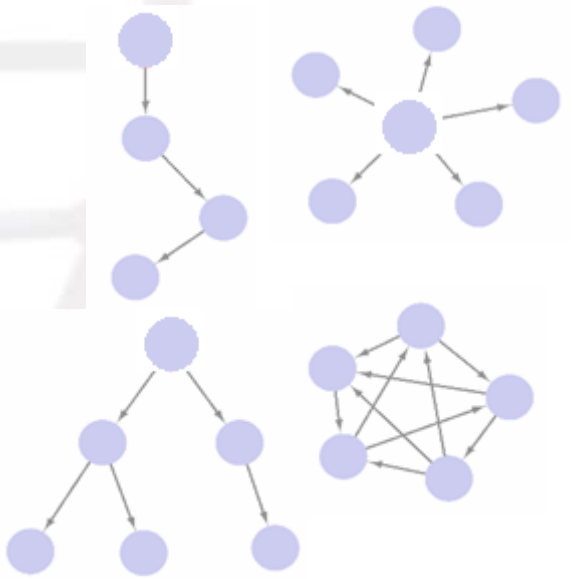
- Highlight those where the expression profiles of the adjacent genes comply with prior biological knowledge.
- These are edges with highly positive or negative correlated adjacent gene expression profiles and are considered a priori as edges with activating $reg = 1$ or suppressing $reg = -1$ regulation role based on KEGG pathway maps (*Vrahatis et al., 2016*).
- Correlation measures: Pearson, Spearman, Kendall

Subpathway extraction

— Until Now:
  — Linear & Non-Linear
  — Perturbation in a subpathway
  — NOT: Gene-based control

— Idea:

✓ Different Topologies



✓ Starting from a Gene of Interest

— Gene
  - User-defined or
  - with topological role or
  - with functional role

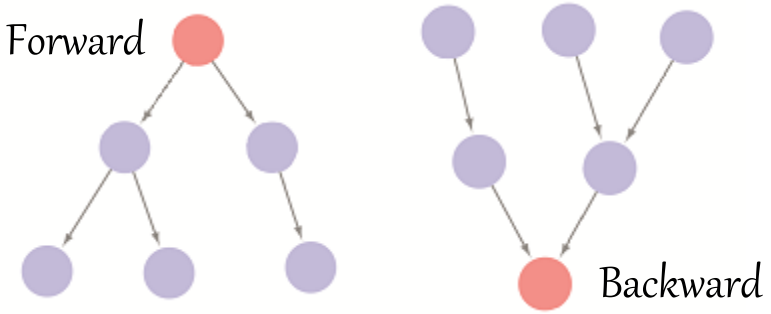
✓ up & down stream

UPstream
(is controlled)

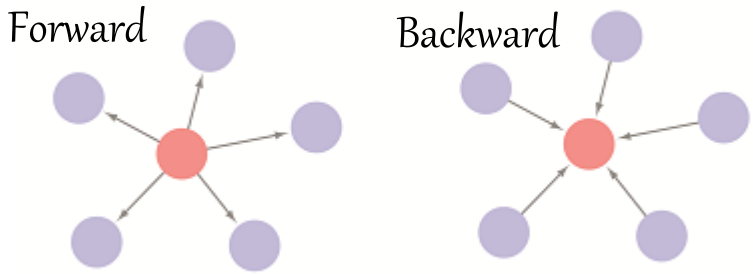
DOWNstream
(Controls)

Subpathway Extraction Types

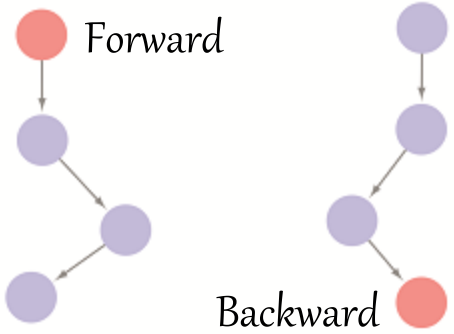
Streams



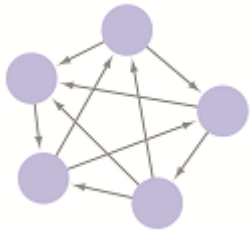
Neighborhoods



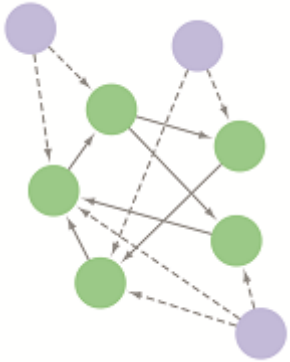
Cascades



Components



Communities

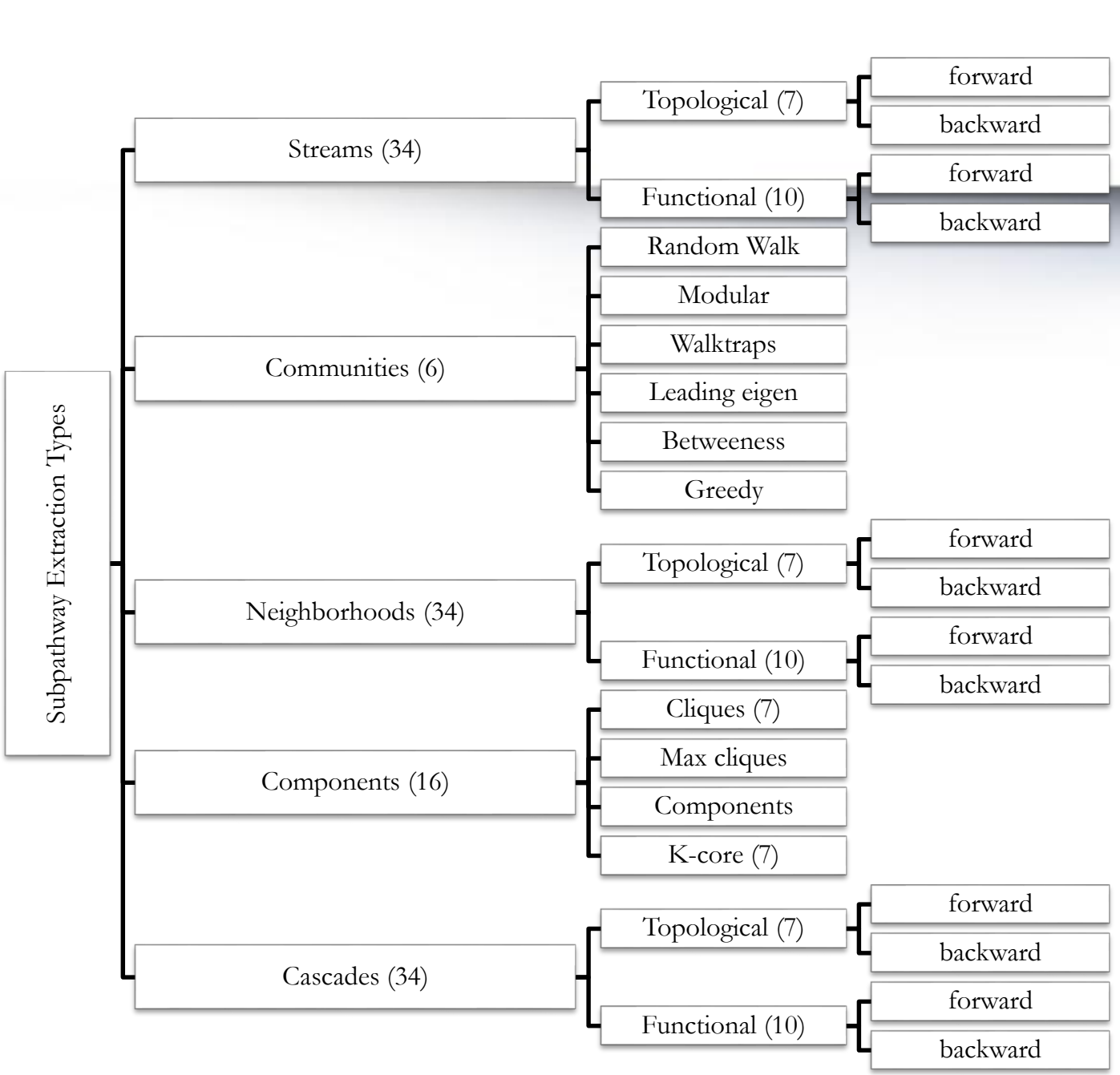


Gene of Interest (GoI)

– Genes having crucial topological or functional roles within the network are considered as GOIs

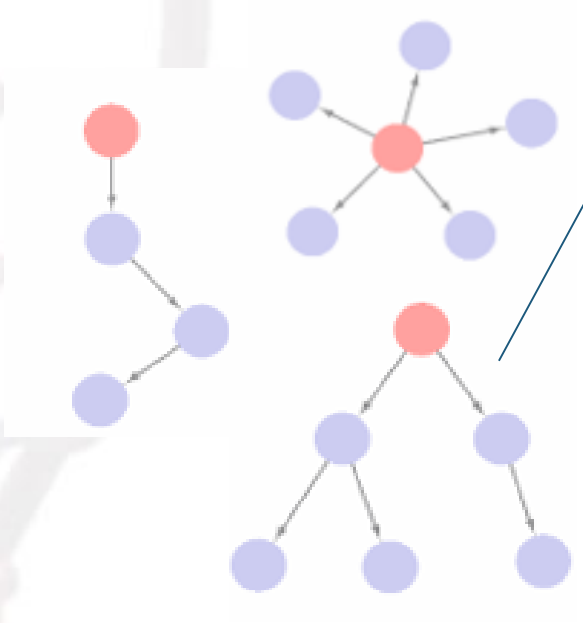
Topological Role
Degree
Betweenness
Closeness
Hub score
Eccentricity
Page rank
Start Nodes

Functional Role
DEG
Pathways
Biological Process
Cellular Component
Molecular Function
Disease OMIM
Disease GAD
Drug - DrugBank
microRNA
Transcription Factors
DEG

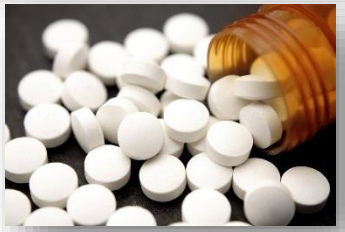

Subpathway Extraction Types
Streams (34)
Topological (7)
forward
backward
Functional (10)
forward
backward
Communities (6)
Random Walk
Modular
Walktraps
Leading eigen
Betweeness
Greedy
Neighborhoods (34)
Topological (7)
forward
backward
Functional (10)
forward
backward
Components (16)
Cliques (7)
Max cliques
Components
K-core (7)
Cascades (34)
Topological (7)
forward
backward
Functional (10)
forward
backward

Subpathway enrichment analysis

- We found the DE subpathways
- Are they related with biological or pharmacological features ???



Biological Features



Pharmacological Features

Subpathway enrichment analysis

- The extracted subpathways are further examined for enrichment in various biological and pharmacological features
- hypergeometric test is used to estimate the subpathway associations

I.	Pathway Terms
II.	Gene Ontology Terms
III.	Disease Terms
IV.	Pharmaceuticals Terms
V.	Regulation Terms

Databases
KEGG
Gene Ontology
OMIM
GAD
DrugBank
miRecords
Transfac
Jaspar

Subpathway enrichment analysis

- The extracted subpathways are further examined for enrichment in various biological and pharmacological features
- hypergeometric test is used to estimate the subpathway associations

I. Pathway Terms

II. Gene Ontology Terms

III. Disease Terms

IV. Pharmaceuticals Terms

V. Regulation Terms

$$P = 1 - \sum_{x=0}^{d} \frac{\binom{D}{x}\binom{G-D}{l-x}}{\binom{G}{l}},$$

- ✓ G: number of user-defined genes
- ✓ l: number of genes that belongs to the subpathway.
- ✓ D: number of related terms
- ✓ d: terms that related with the subpathway

Subpathway enrichment analysis

- The extracted subpathways are further examined for enrichment in various biological and pharmacological features
- hypergeometric test is used to estimate the subpathway associations

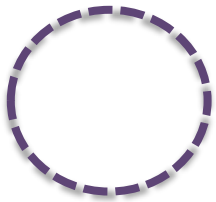
I. Pathway Terms

II. Gene Ontology Terms

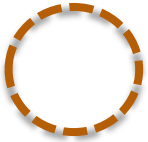
III. Disease Terms

IV. Pharmaceuticals Terms

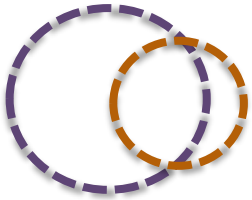
V. Regulation Terms



Genes that belongs with the Disease A



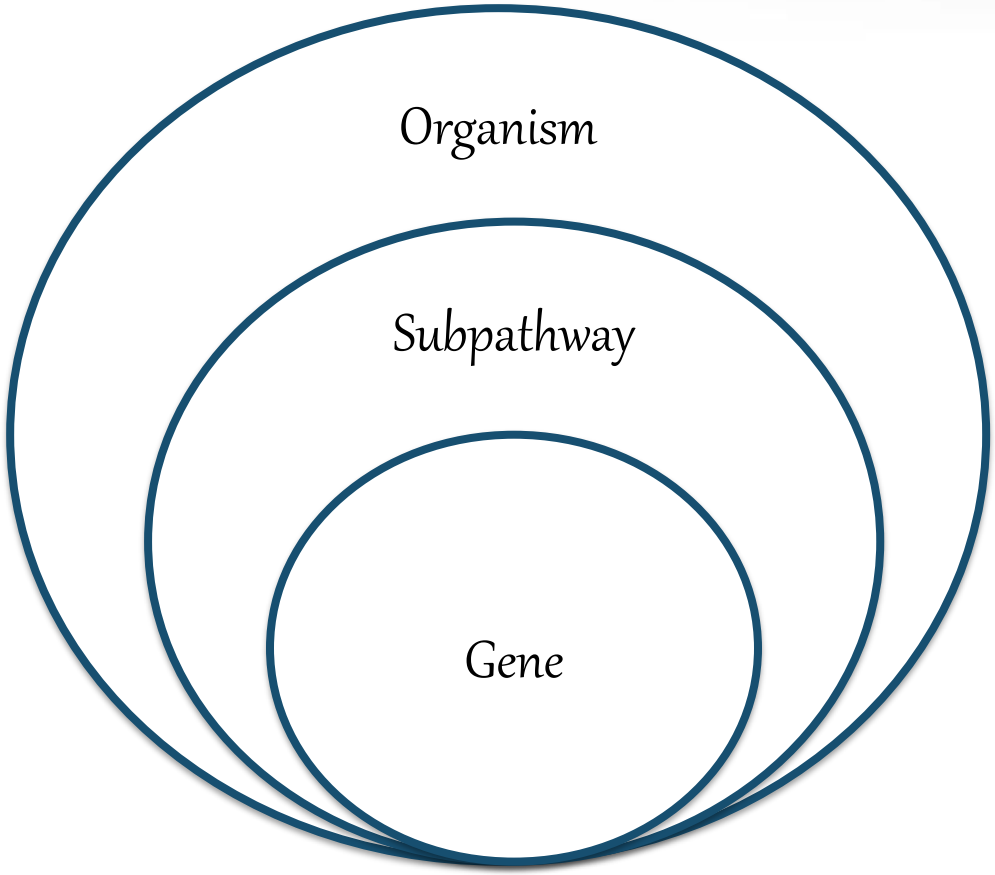
Genes that are members of subpathway B



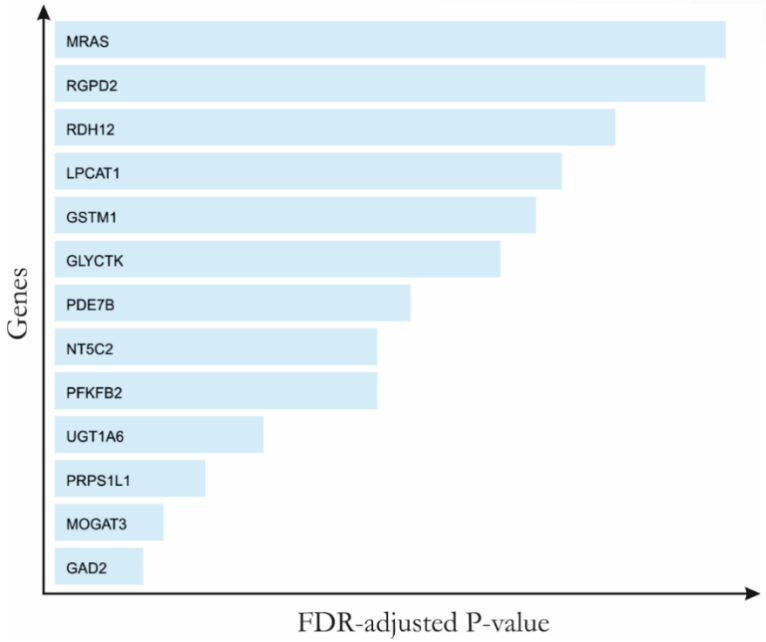
IF $P < 0.05$ THEN

Subpathway B is related with Disease A

Visualization - Levels

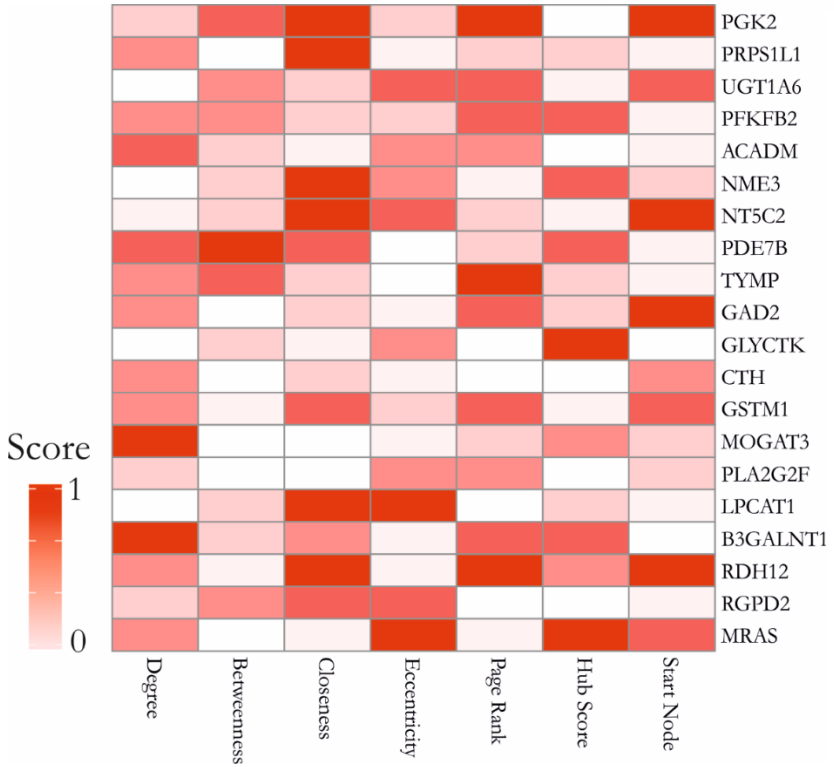


Visualization – Gene Level



Gene level visualization example – Bar plot.
Bars illustrate the genes with the highest FDR-adjusted P-value (*Qvalue*).
Defined number of genes was thirteen.

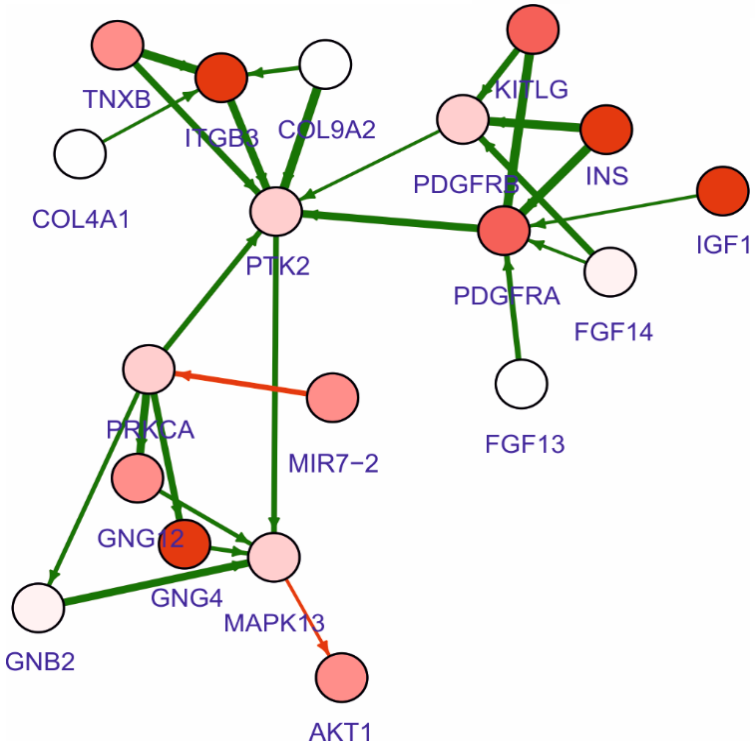
Visualization – Gene Level



Gene level visualization example – Heat map.

Heat map represents the first twenty genes that are included in the pruned graph and have the highest values from seven topological measures. The values are scaled and the red graduation indicates the value degree.

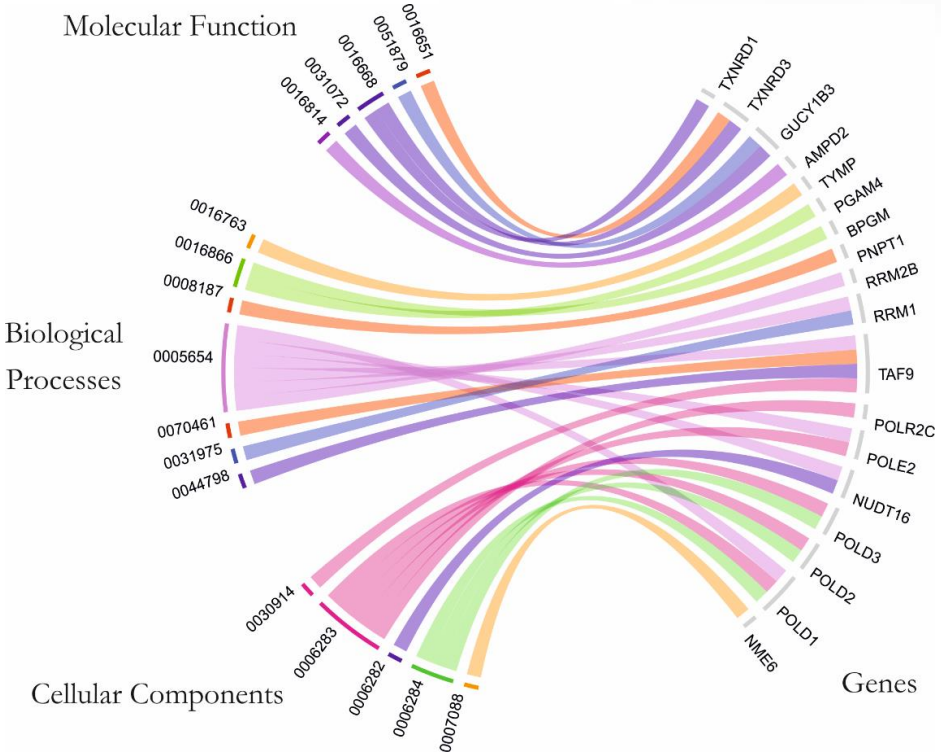
Visualization – Subpathway Level



Subpathway level visualization example – Directed weighted graph

Red graduation in nodes indicate the *Qvalue* degree, the edge width indicates the correlation degree between the respective genes. Green or red color in edges indicates the positive or negative correlation respectively.

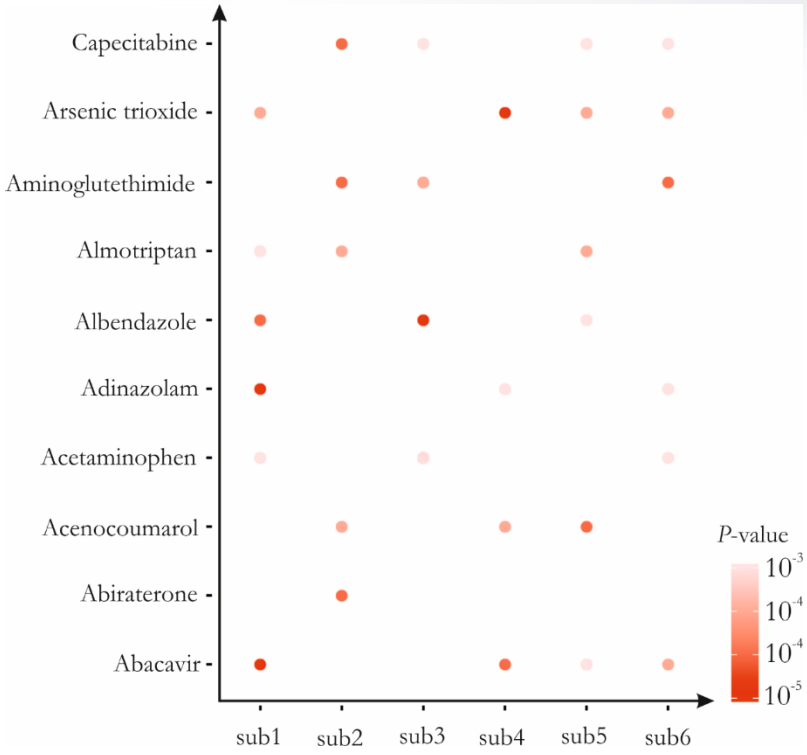
Visualization – Subpathway Level



Circular Diagram

Circular Diagram shows the associations among genes including in a subpathway and Gene Ontology terms where are enriched

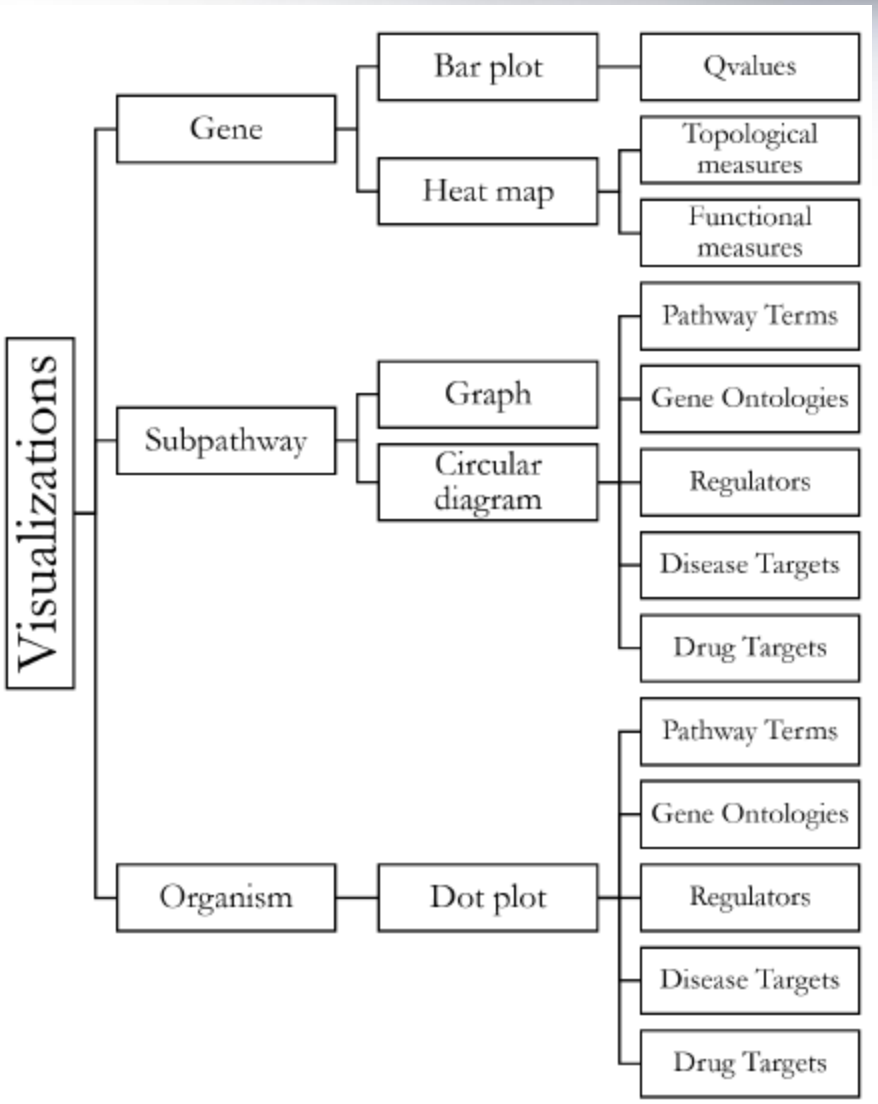
Visualization – Organism Level



Dot plot

Dot plot shows the enriched associations among experiment-specific extracted subpathways and drug substances from DrugBank database. Ten substances were selected as desired number.

Visualization Scheme



Conclusion

- Network-based and Personalized-based Medicine is the future of Medicine
- Network-based approaches have gained ground in the research field of systems biology
- OMICS data explosion make the systems biology computational tools, an imperative need
- Subpathway-based Analysis is the state-of-the-art of Pathway analysis and can explore deeper the complex diseases and complex biological processes.


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
спасибо
dank je
tapadh leat
bedankt
dziękuję
hvala
mauruuru
obrigado
sukriya
kop khun krap
go raibh maith agat
arigatō
takk
dakujem
merci
merci
рахмат
Баярлалаа
mochchakkeram
grazie
terima kasih
감사합니다
ευχαριστώ
мерси