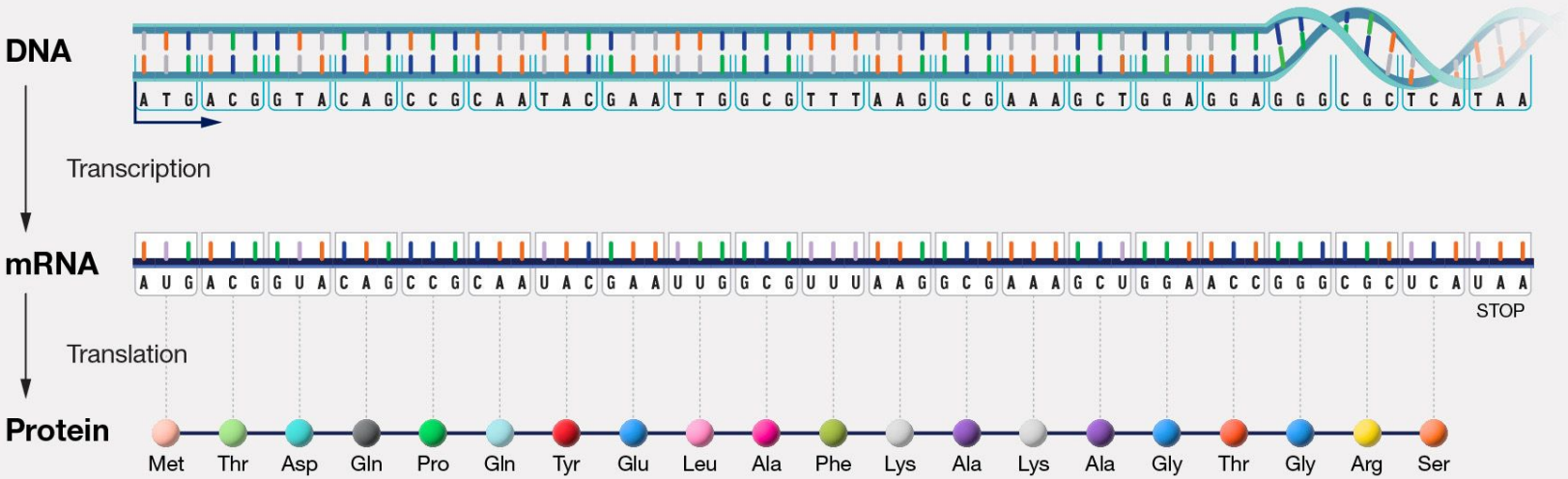
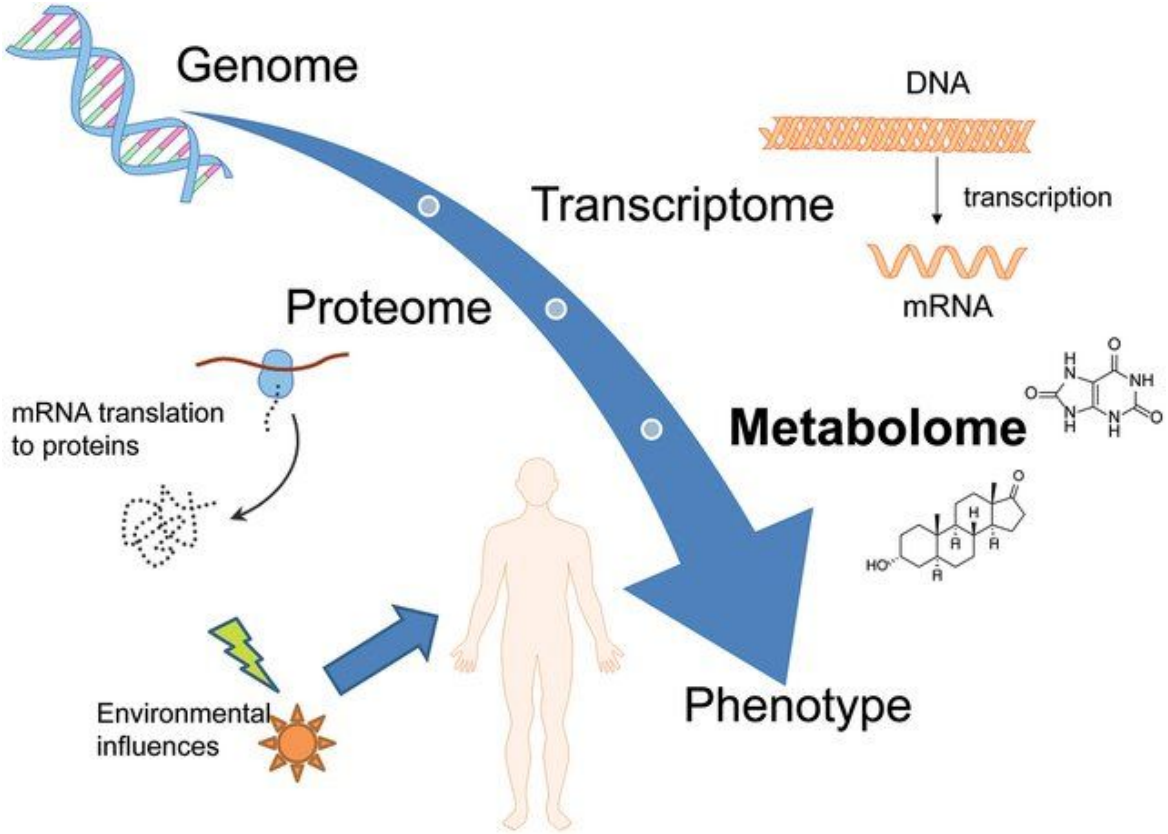
Decoding Complexity: Omics and Network Biology

Savandara Besse

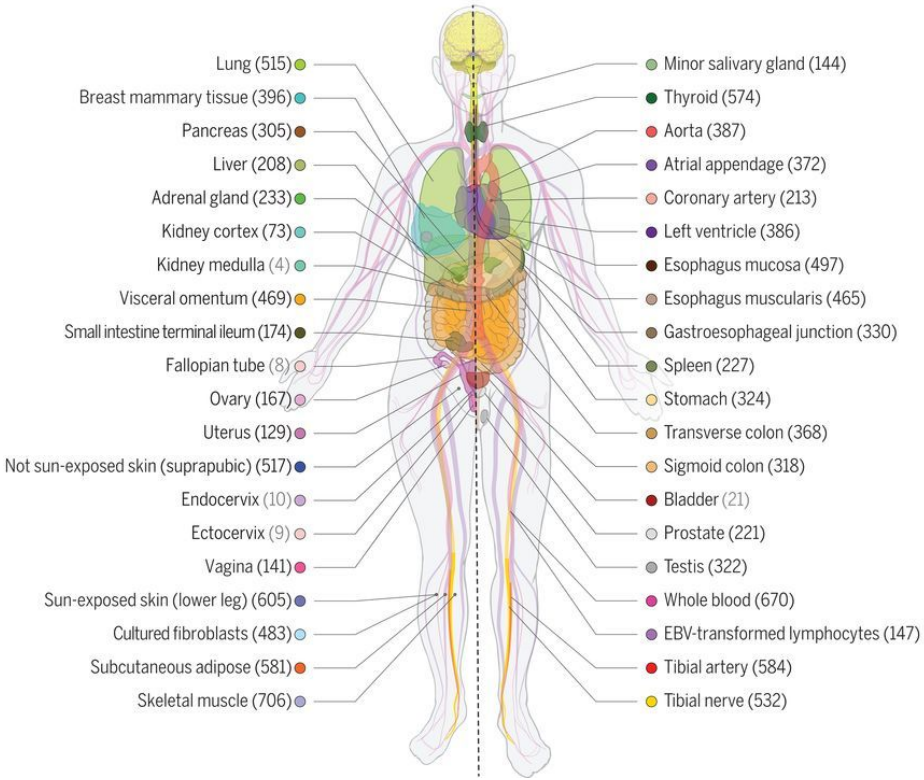

DNA
A T G A C G G T A C A G C C G C A A T A C G A A T T G G C G T T T A A G G C G A A A G C T G G A G G A G G G C G C T C A T A A
Transcription
mRNA
A U G A C G G U A C A G C C G C A A U A C G A A U U G G C G U U U A A G G C G A A A G C U G G A A C C G G G C G C U C A U A A
STOP
Translation
Protein
Met Thr Asp Gln Pro Gln Tyr Glu Leu Ala Phe Lys Ala Lys Ala Gly Thr Gly Arg Ser



(Steuer et al, Frontiers in Chemisry, 2019)

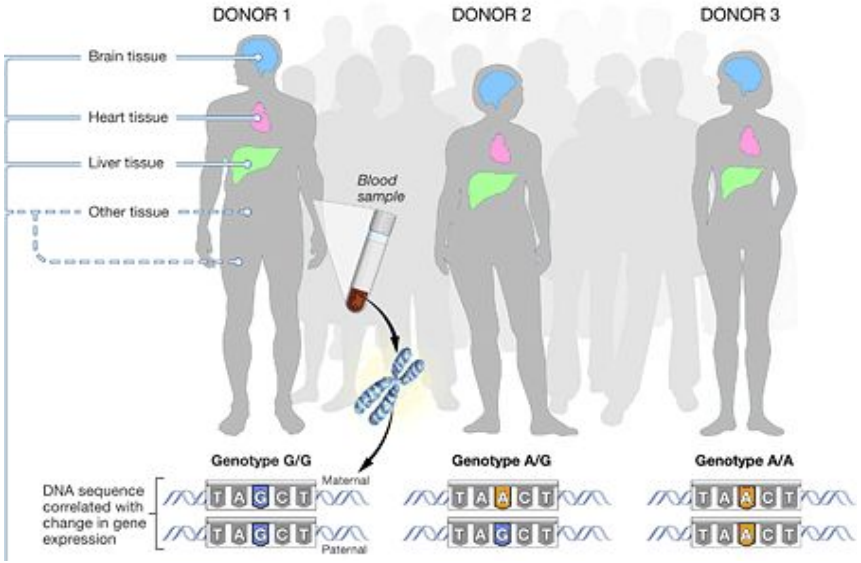
Omics Type	Key Technologies / Methods	Typical Output	Applications / Notes
Genomics	Short-read WGS (Illumina), Long-read (PacBio, Nanopore)	DNA sequence, SNPs, structural variants	Genome assembly, comparative genomics, mutation detection
Transcriptomics	RNA-seq (bulk, single-cell), Microarrays	Gene expression levels, splice variants	Development, evolution, disease profiling
Proteomics	LC-MS/MS, MALDI-TOF, Antibody arrays	Protein identity, abundance, PTMs	Pathway analysis, biomarker discovery
Metabolomics	LC-MS, GC-MS, NMR	Metabolite levels, flux	Functional phenotype, metabolic pathways
Epigenomics	ChIP-seq, ATAC-seq, Bisulfite sequencing	DNA methylation, histone modifications, chromatin accessibility	Gene regulation, development, disease

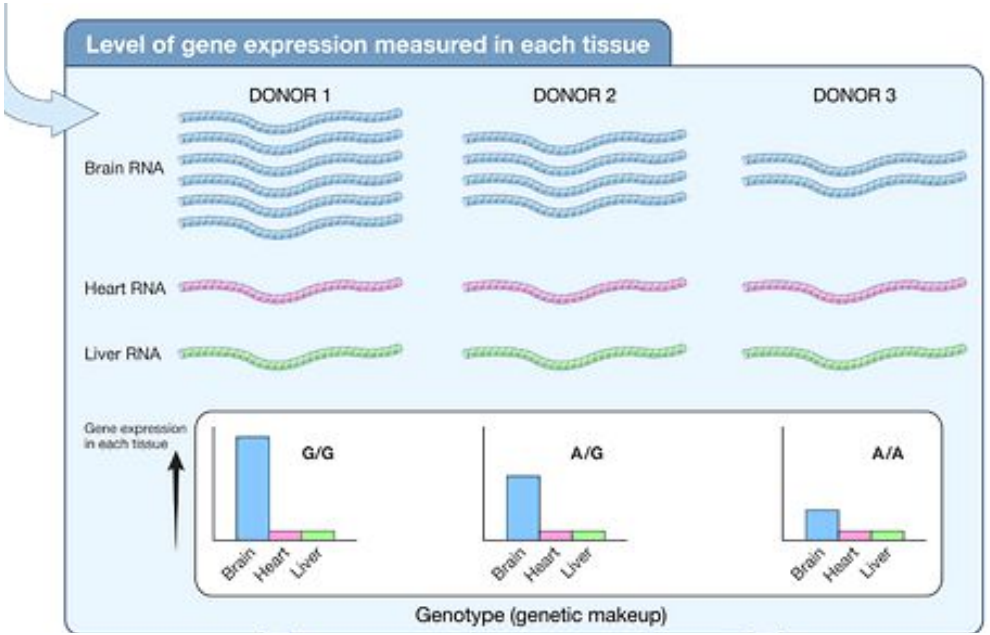
The Adult Genotype Tissue Expression (GTEx) Project



(GTeX Consortium, Science 2020)

- 54 non-diseased tissues in 1,000 deceased individuals
- Determination of genetic variation for each individual (SNPs, indels, structural variants)
  → Whole Genome Sequencing
- Measure of gene expression within each tissue
  → RNA-sequencing

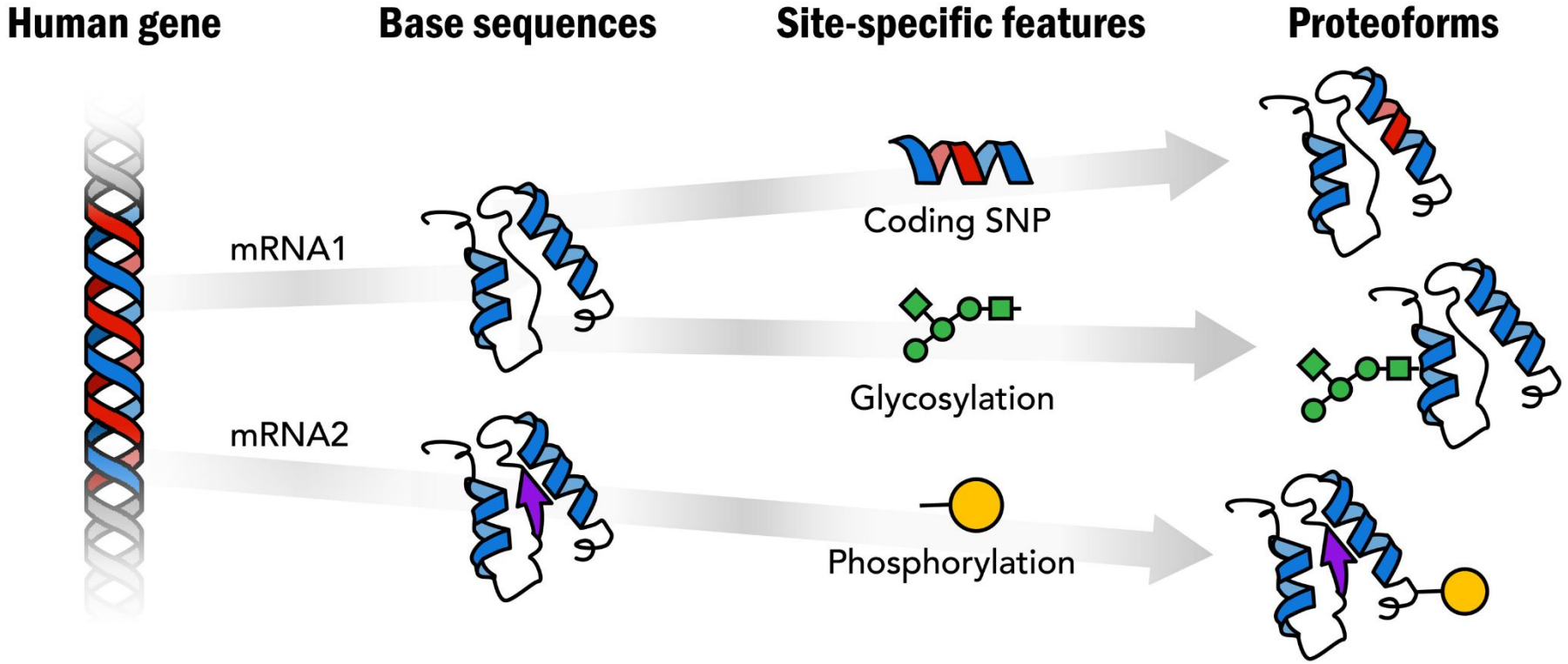




⇒ Identification of eQTLs (expression Quantitative Trait Loci)

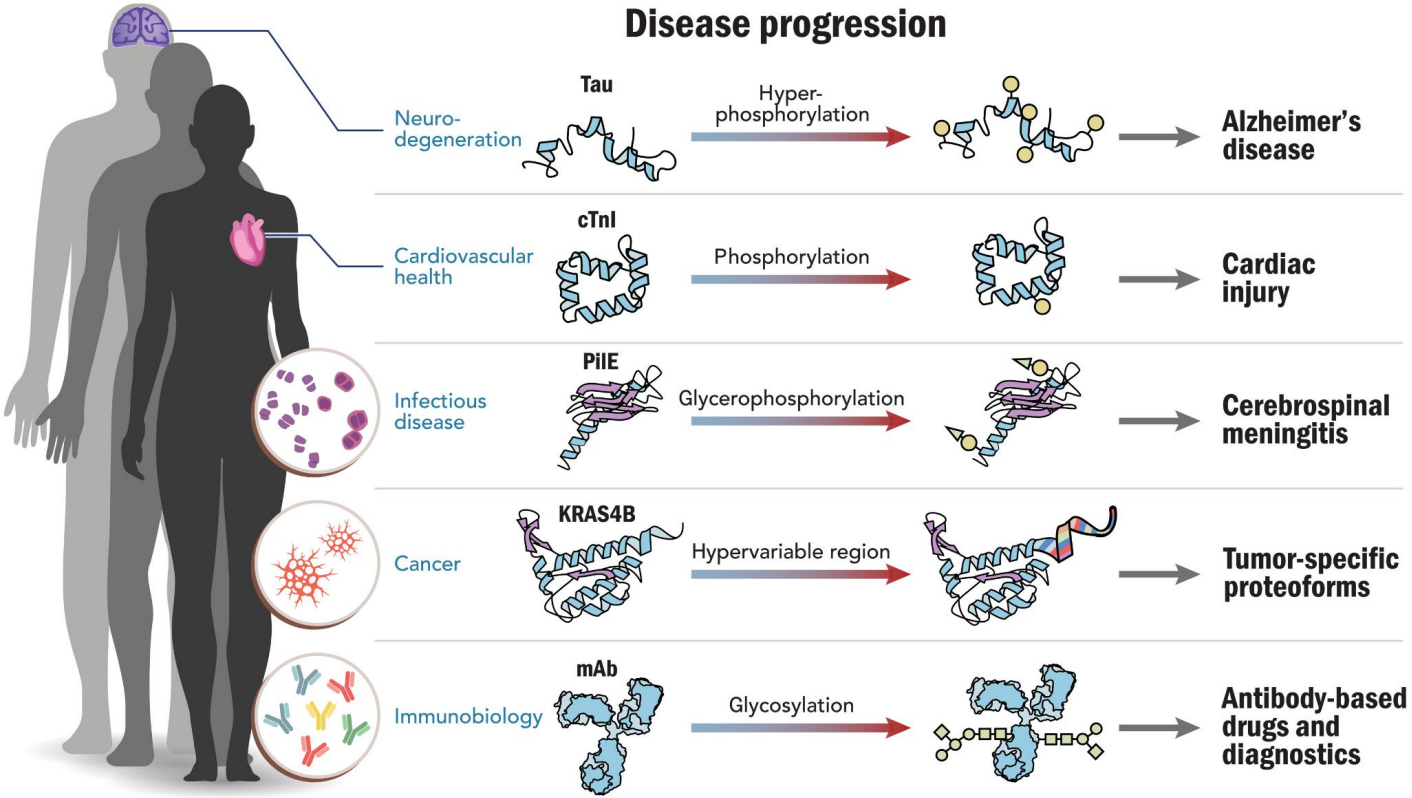
https://www.genome.gov/Funded-Programs-Projects/Genotype-Tissue-Expression-Project

The Human Proteoform Project



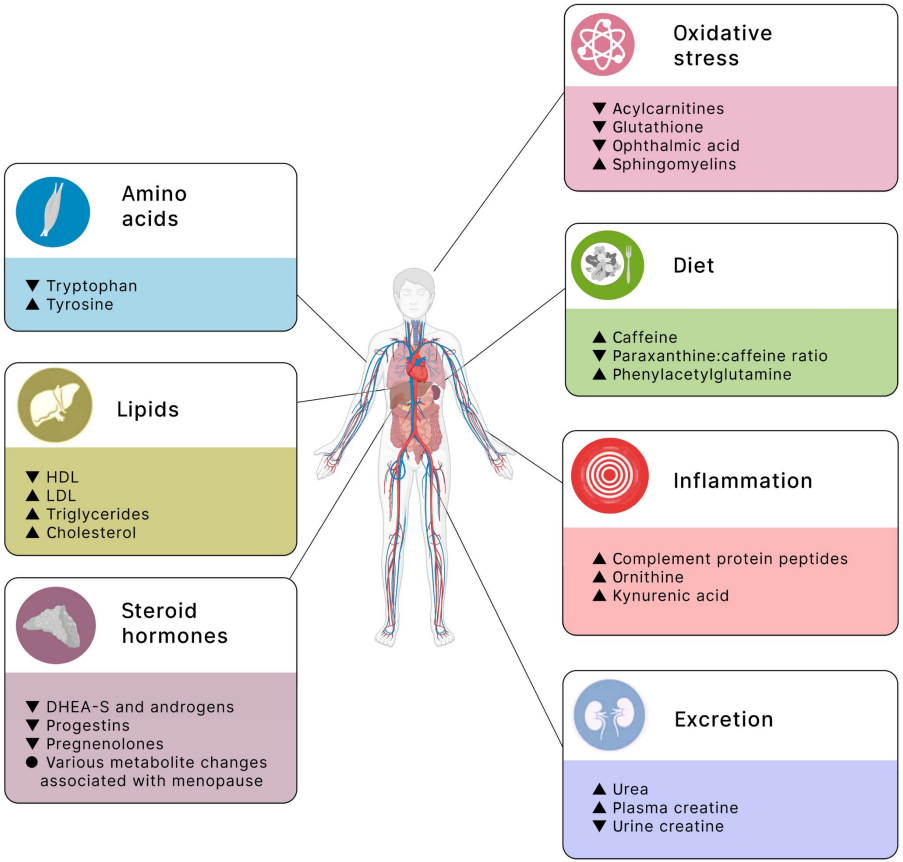
(Smith & The Consortium for Top-Down Proteomics, Science Advances 2021)

The Human Proteoform Project



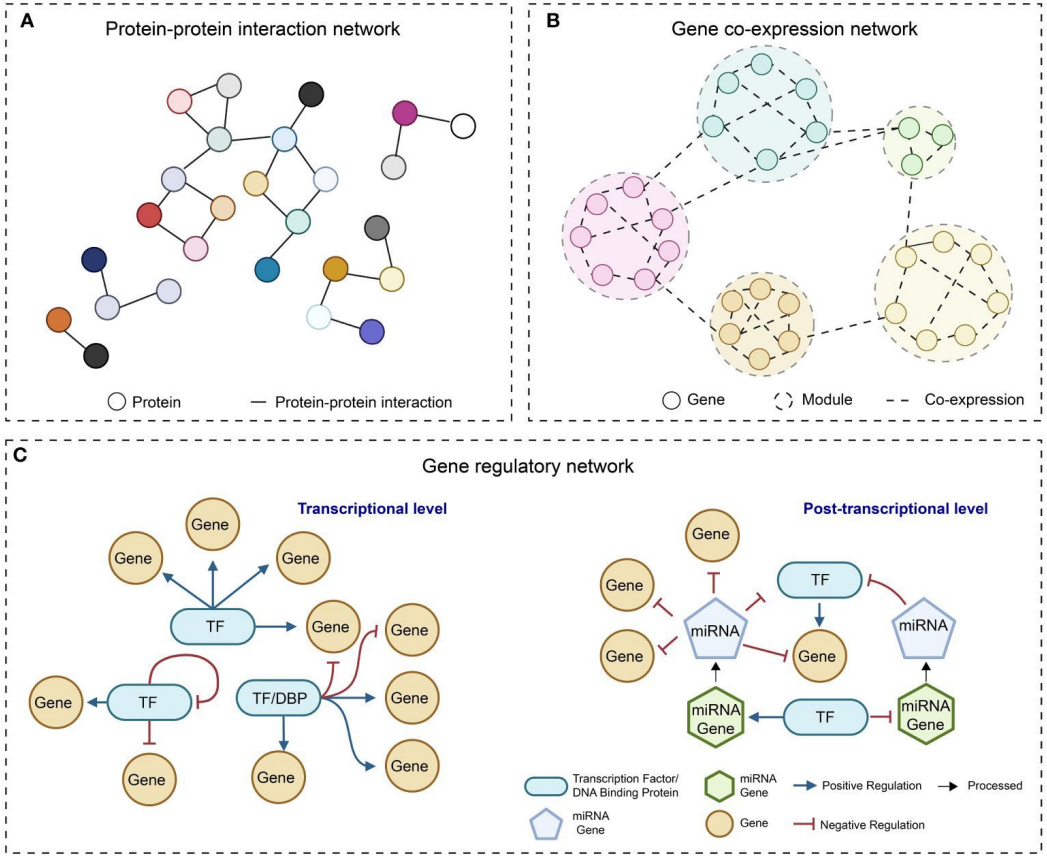
(Smith & The Consortium for Top-Down Proteomics, Science Advances 2021)

Metabolomics Pathways Involved In Human Aging



(Panyard, Yu & Snyder, Science Advances, 2022)

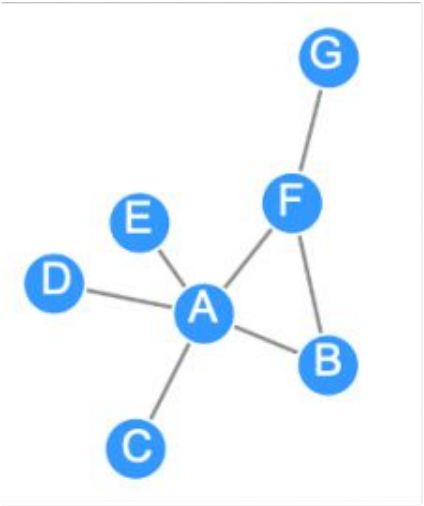
From omics to biological networks



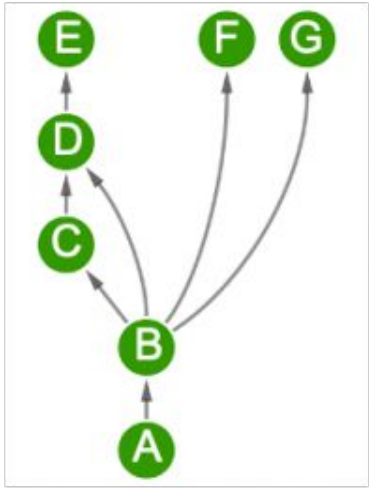
Important concepts in network biology

Different types of biological networks

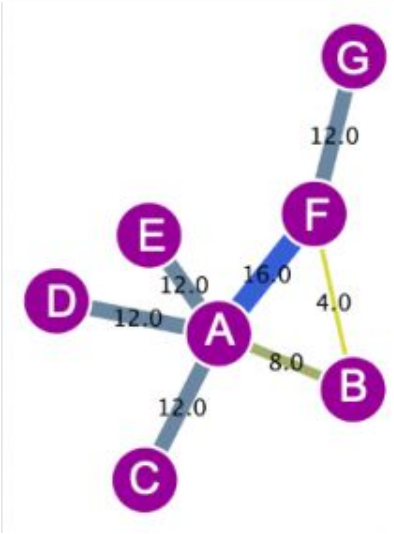
Undirected



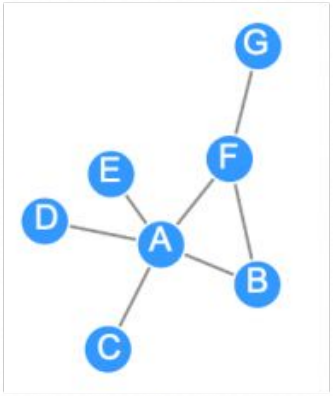
Directed



Weighted

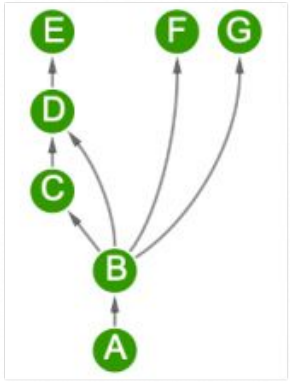


Undirected



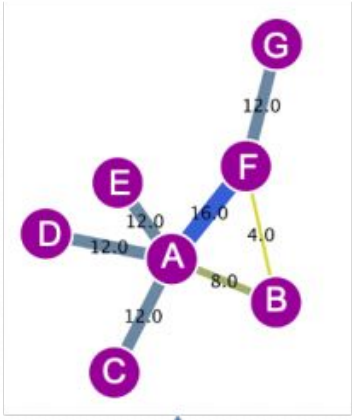
	A	B	C	D	E	F	G	Degree
A	0	1	1	1	1	1	0	5
B	1	0	0	0	0	1	0	2
C	1	0	0	0	0	0	0	1
D	1	0	0	0	0	0	0	1
E	1	0	0	0	0	0	0	1
F	1	1	0	0	0	0	1	3
G	0	0	0	0	0	1	0	1

Directed



	A	B	C	D	E	F	G	Out-degree
A	0	1	0	0	0	0	0	1
B	0	0	1	1	0	1	1	4
C	0	0	0	1	0	0	0	1
D	0	0	0	0	1	0	0	1
E	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0

Weighted



	A	B	C	D	E	F	G	Degree
A	0	8	12	12	12	16	12	72
B	8	0	0	0	0	4	0	12
C	12	0	0	0	0	0	0	12
D	12	0	0	0	0	0	0	12
E	12	0	0	0	0	0	0	12
F	16	4	0	0	0	0	12	32
G	12	0	0	0	0	12	0	24

Adjacency matrices

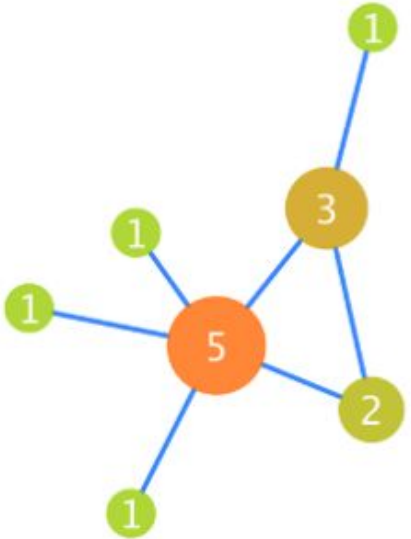
Network topological properties

In an **undirected graph**, the **degree** of a node is the **number of edges** connected to it.

In a **directed graph**:

- **In-degree** = number of incoming edges
- **Out-degree** = number of outgoing edges

⇒ Important for understanding whether a network is **homogeneous** (most nodes have similar degree) or **heterogeneous** (few nodes have very high degree).



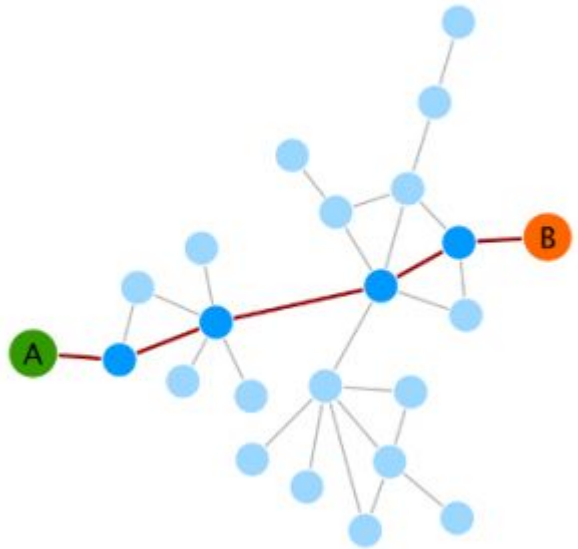
Network topological properties

The **shortest path** between two nodes in a graph is defined as the path that has the minimum total distance among all possible paths connecting those two nodes.

- If the graph is **unweighted**, the shortest path is simply the path with the fewest edges.

- If the graph is **weighted**, the shortest path is the one for which the **sum of the weights of the edges along the path is minimized**.

⇒ Reflects how “small” or “efficient” the network is in terms of connectivity.



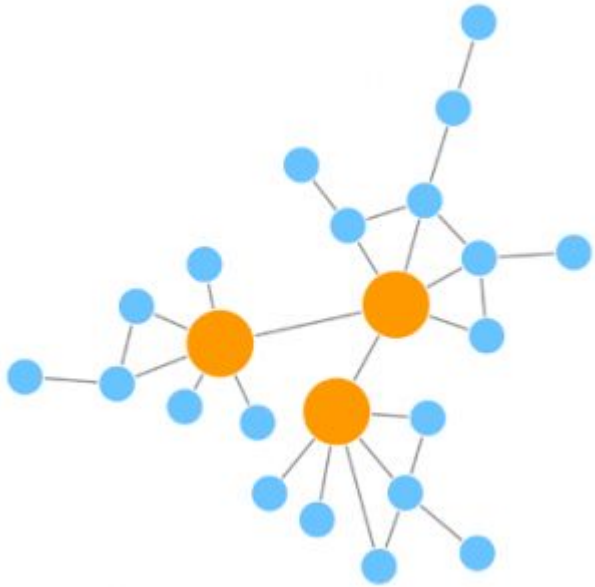
Network topological properties

A **hub** is a **node with a significantly higher number of connections (degree) than most other nodes in the network**.

- Hubs are **highly connected nodes** that often play a critical role in the structure and function of the network.

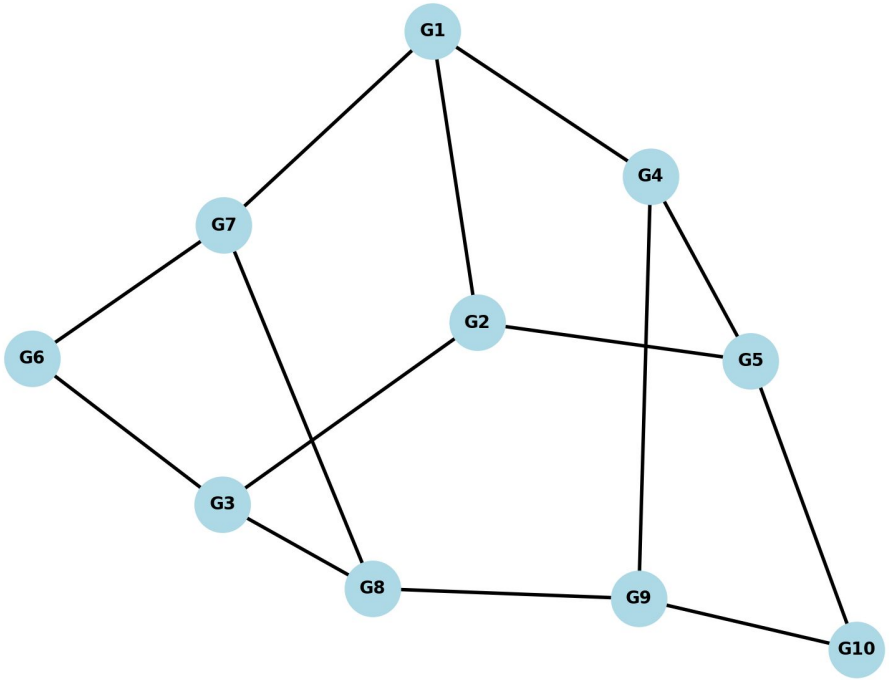
- In **biological networks**, hubs can represent essential proteins whose disruption may strongly affect the system.

⇒ In **scale-free networks**, hubs emerge naturally, following a **power-law degree distribution**, where a few nodes have very high connectivity, and most nodes have few connections.



Exercice

1. Représentez ce graphe sous la forme d'une **matrice d'adjacence** A de taille 10×10.
   - Rappel : $A_{ij} = 1$ si le gène G_i est connecté à G_j, et 0 sinon.
   - Comme le graphe est **non dirigé**, la matrice doit être **symétrique**.
   - Les diagonales sont nulles car aucun gène n'est connecté à lui-même.



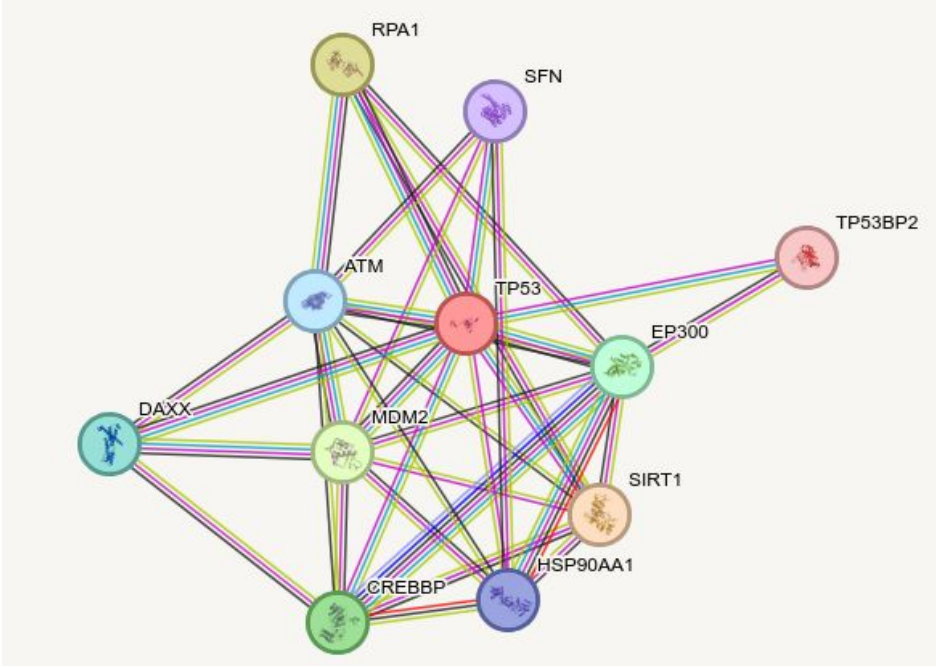
Inference of biological networks from database

String

Purpose: Curates known and predicted **protein–protein interactions**.

Data Sources: Experimental data, computational prediction, co-expression, text mining.

Confidence Scores: Each interaction has a **score indicating reliability**.



Browse interactions in tabular form:

node1	node2	node1 accession	node2 accession	node1 annotation	node2 annotation	score
ATM	CREBBP	ENSP00000278616	ENSP00000262367	Serine-protein kinase ATM; Seri...	CREB-binding protein; Acetylate...	0.497
ATM	DAXX	ENSP00000278616	ENSP00000266000	Serine-protein kinase ATM; Seri...	Death domain-associated prote...	0.813
ATM	EP300	ENSP00000278616	ENSP00000263253	Serine-protein kinase ATM; Seri...	Histone acetyltransferase p300;...	0.904
ATM	HSP90AA1	ENSP00000278616	ENSP00000335153	Serine-protein kinase ATM; Seri...	Heat shock protein HSP 90-alph...	0.785
ATM	MDM2	ENSP00000278616	ENSP00000258149	Serine-protein kinase ATM; Seri...	E3 ubiquitin-protein ligase Mdm...	0.995
ATM	RPA1	ENSP00000278616	ENSP00000254719	Serine-protein kinase ATM; Seri...	Replication protein A 70 kDa DN...	0.928
ATM	SFN	ENSP00000278616	ENSP00000340989	Serine-protein kinase ATM; Seri...	14-3-3 protein sigma; Adapter p...	0.604
ATM	SIRT1	ENSP00000278616	ENSP00000212015	Serine-protein kinase ATM; Seri...	NAD-dependent protein deacety...	0.856
ATM	TP53	ENSP00000278616	ENSP00000269305	Serine-protein kinase ATM; Seri...	Cellular tumor antigen p53; Act...	0.999
CREBBP	ATM	ENSP00000262367	ENSP00000278616	CREB-binding protein; Acetylate...	Serine-protein kinase ATM; Seri...	0.497
CREBBP	DAXX	ENSP00000262367	ENSP00000266000	CREB-binding protein; Acetylate...	Death domain-associated prote...	0.842
CREBBP	EP300	ENSP00000262367	ENSP00000263253	CREB-binding protein; Acetylate...	Histone acetyltransferase p300;...	0.999
CREBBP	HSP90AA1	ENSP00000262367	ENSP00000335153	CREB-binding protein; Acetylate...	Heat shock protein HSP 90-alph...	0.415

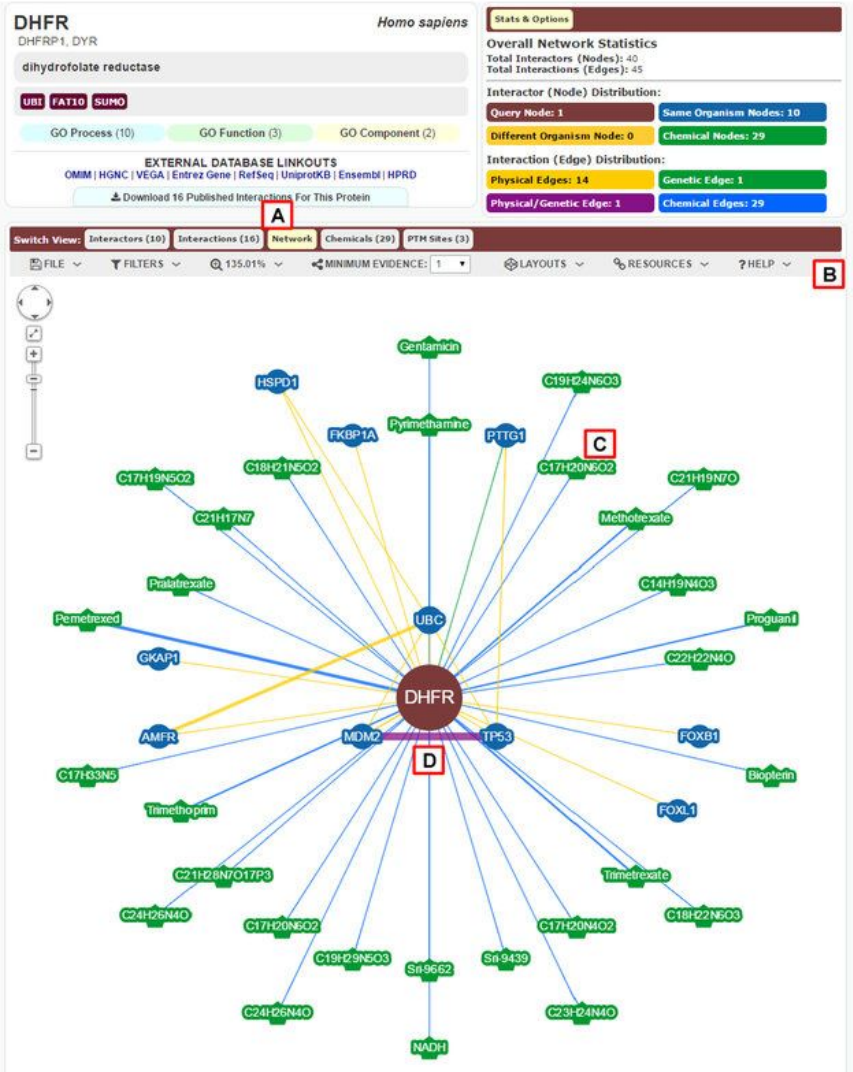
https://string-db.org/

BioGRID

Purpose: Curates **physical and genetic interactions** in multiple organisms.

Data Sources: Published **experimental studies** (manual curation from literature).

- Node and edge colour indicates the interaction type and node size is proportional to its connectivity.
- In this example, green nodes represent chemicals and blue nodes represent proteins.
- Yellow edges represent protein interactions, green edges represent genetic interactions, blue edges represent chemical interactions and purple edges represent both protein and genetic interactions.



https://thebiogrid.org/

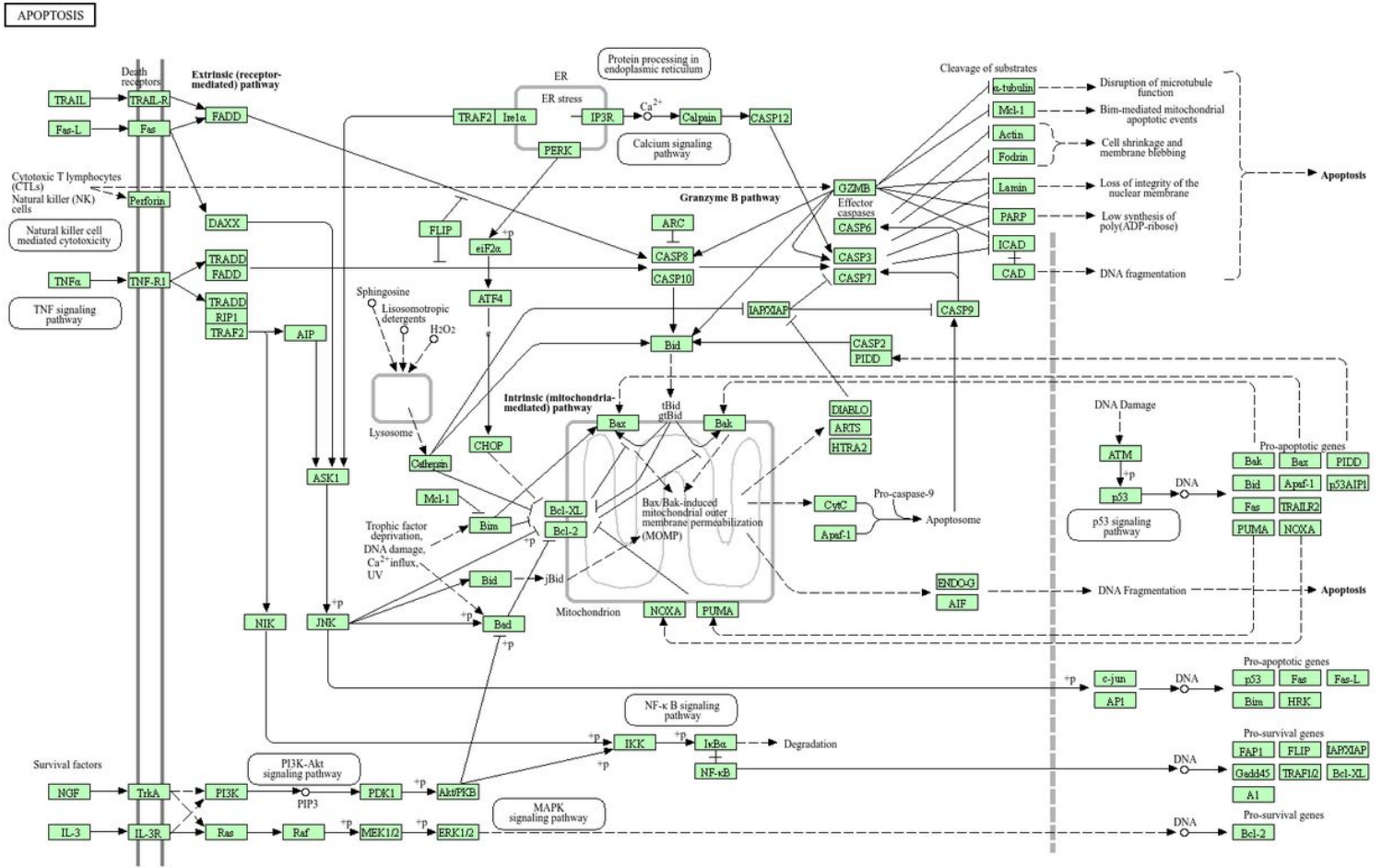
(Chatr-aryamontri et al, Nucleic Acid Research, 2017)

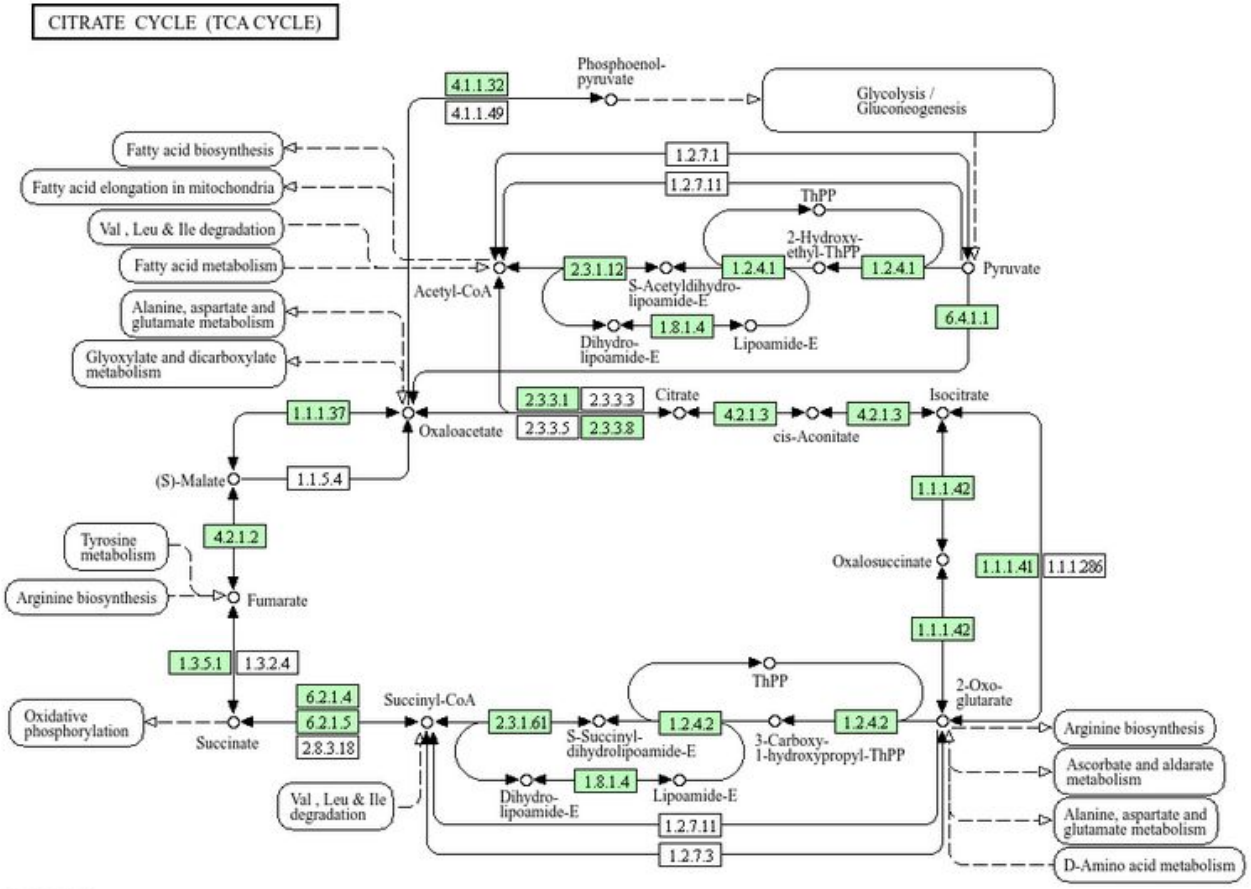
KEGG

Purpose: Integrates **genes, proteins, and metabolites** into **pathways and functional networks**.

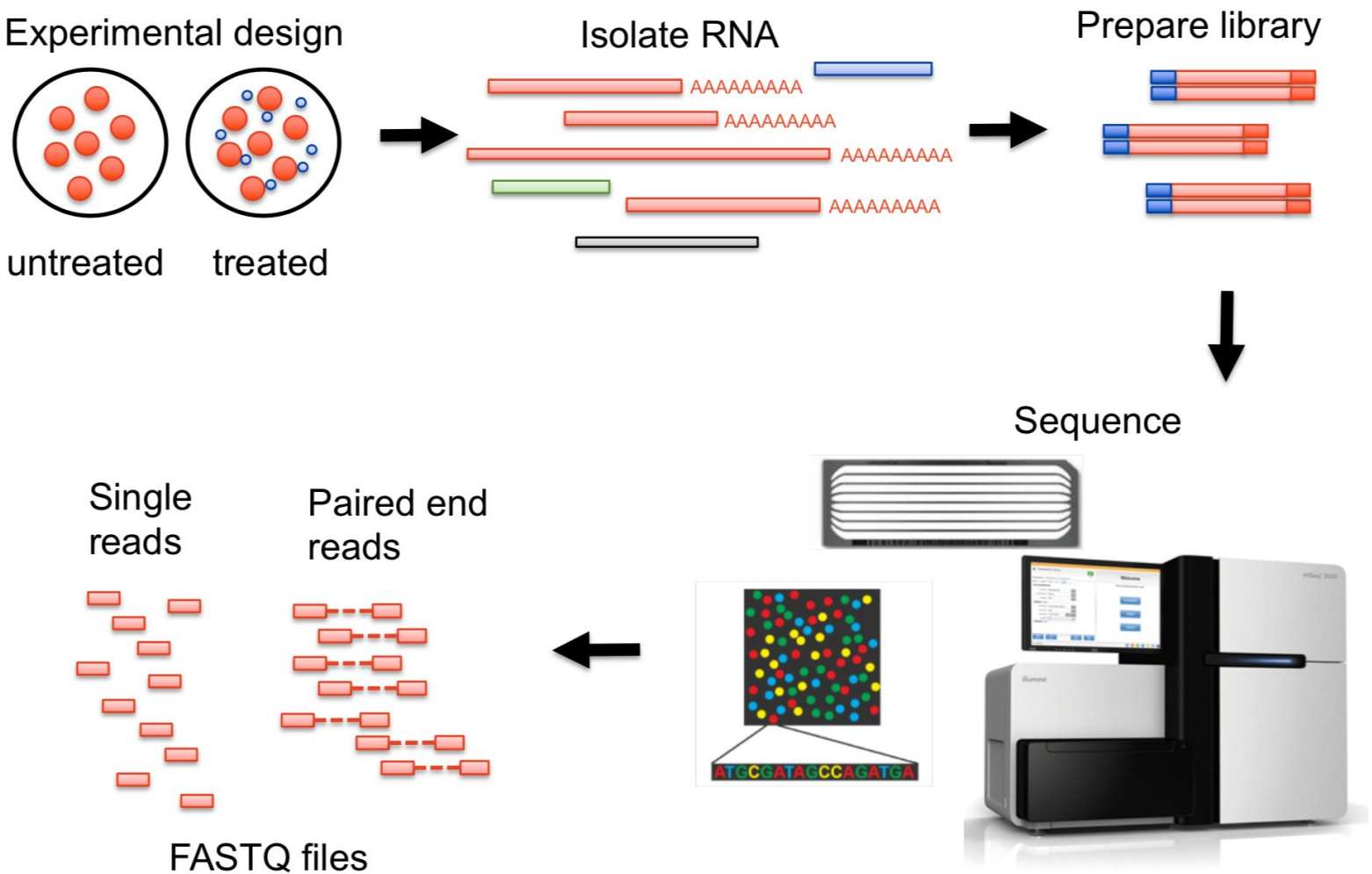
Data Types: Metabolic pathways, signaling pathways, disease pathways, and drug information.

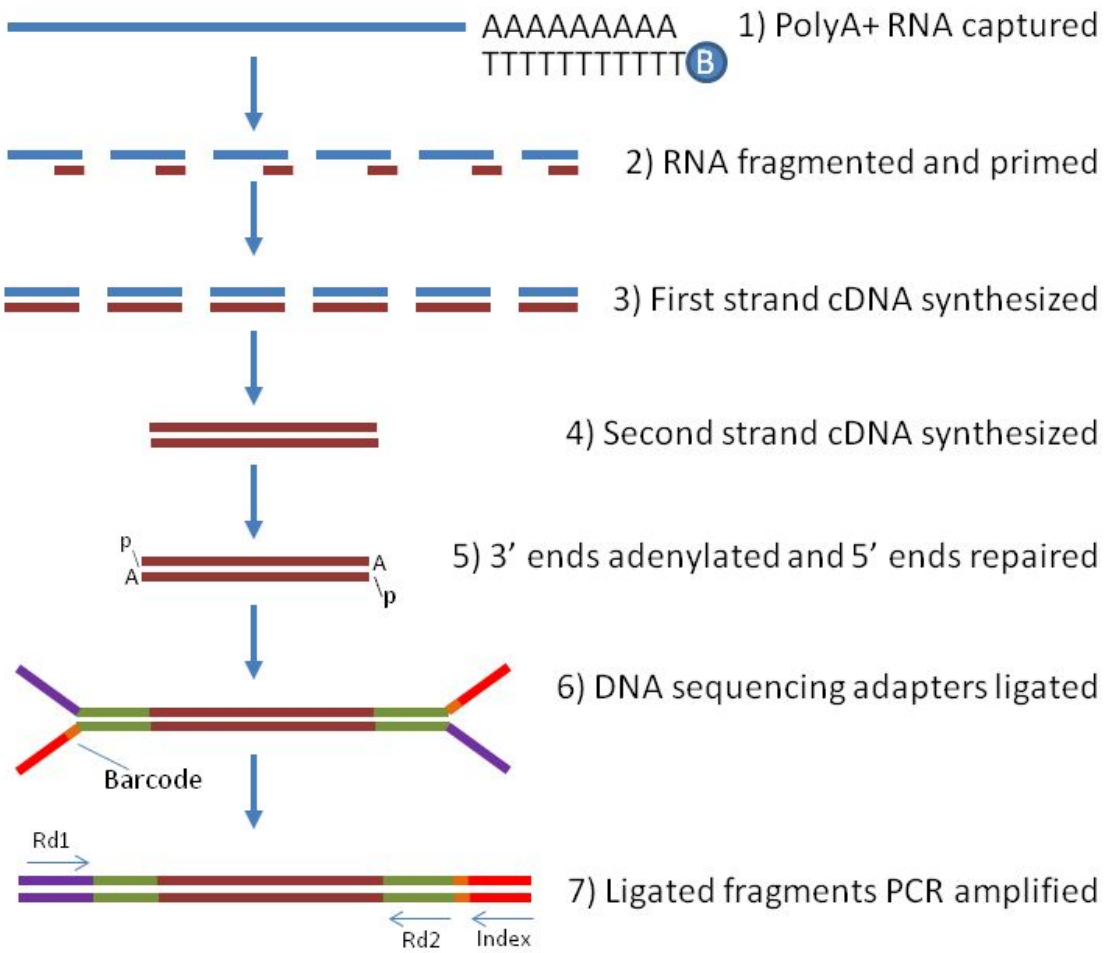
https://www.kegg.jp/


APOPTOSIS
Death receptors
Extrinsic (receptor-mediated) pathway
TRAIL
TRAIL-R
Fas-L
Fas
FADD
Cytotoxic T lymphocytes (CTLs)
Natural killer (NK) cells
Perforin
Natural killer cell mediated cytotoxicity
DAXX
TNFα
TNF-R1
TNF signaling pathway
TRADD
FADD
TRADD
RIP1
TRAF2
AIP
ASK1
NIK
JNK
ER
Protein processing in endoplasmic reticulum
ER stress
TRAF2
Ire1α
IP3R
Ca2+
Calpain
CASP12
PERK
Calcium signaling pathway
FLIP
eiF2α
ATF4
CHOP
Sphingosine
Lisosomotropic detergents
H2O2
Lysosome
Cathepsin
Intrinsic (mitochondria-mediated) pathway
Bax
tBid
gtBid
Bak
Mcl-1
Bcl-XL
Bcl-2
Bim
Trophic factor deprivation, DNA damage, Ca2+ influx, UV
Bid
jBid
Bad
Mitochondrion
NOXA
PUMA
Bax/Bak-induced mitochondrial outer membrane permeabilization (MOMP)
DIABLO
ARTS
HTRA2
CytC
Apaf-1
Pro-caspase-9
Apoptosome
ENDO-G
AIF
DNA Fragmentation
Apoptosis
Granzyme B pathway
ARC
CASP8
CASP10
IAP/XIAP
CASP9
CASP2
PIDD
GZMB
Effector caspases
CASP6
CASP3
CASP7
Cleavage of substrates
α-tubulin
Disruption of microtubule function
Mcl-1
Bim-mediated mitochondrial apoptotic events
Actin
Fodrin
Cell shrinkage and membrane blebbing
Lamin
Loss of integrity of the nuclear membrane
PARP
Low synthesis of poly(ADP-ribose)
ICAD
CAD
DNA fragmentation
Apoptosis
DNA Damage
ATM
p53
DNA
p53 signaling pathway
Pro-apoptotic genes
Bak
Bax
PIDD
Bid
Apaf-1
p53AIP1
Fas
TRAILR2
PUMA
NOXA
NF-κ B signaling pathway
IKK
IκBα
Degradation
NF-κB
c-jun
AP1
DNA
Pro-apoptotic genes
p53
Fas
Fas-L
Bim
HRK
Survival factors
NGF
TrkA
IL-3
IL-3R
PI3K-Akt signaling pathway
PI3K
PIP3
PDK1
Akt/PKB
Ras
Raf
MEK1/2
ERK1/2
MAPK signaling pathway
Pro-survival genes
FAP1
FLIP
IAP/XIAP
Gadd45
TRAF1/2
Bcl-XL
A1
Pro-survival genes
Bcl-2

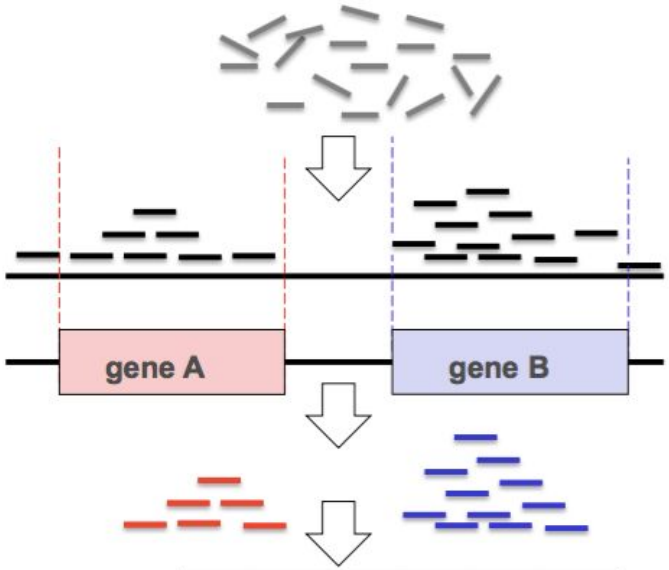

CITRATE CYCLE (TCA CYCLE)
Phosphoenol-pyruvate
4.1.1.32
4.1.1.49
Glycolysis / Gluconeogenesis
Fatty acid biosynthesis
Fatty acid elongation in mitochondria
Val , Leu & Ile degradation
Fatty acid metabolism
Alanine, aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
1.2.7.1
1.2.7.11
ThPP
2-Hydroxy-ethyl-ThPP
2.3.1.12
1.2.4.1
1.2.4.1
Pyruvate
Acetyl-CoA
S-Acetyldihydro-lipoamide-E
1.8.1.4
Dihydro-lipoamide-E
Lipoamide-E
6.4.1.1
1.1.1.37
2.3.3.1
2.3.3.3
2.3.3.5
2.3.3.8
Citrate
4.2.1.3
4.2.1.3
Isocitrate
Oxaloacetate
cis-Aconitate
(S)-Malate
1.1.5.4
1.1.1.42
Tyrosine metabolism
4.2.1.2
Oxalosuccinate
1.1.1.41
1.1.1.286
Arginine biosynthesis
Fumarate
1.1.1.42
1.3.5.1
1.3.2.4
ThPP
2-Oxo-glutarate
6.2.1.4
Succinyl-CoA
Oxidative phosphorylation
6.2.1.5
2.8.3.18
Succinate
2.3.1.61
1.2.4.2
1.2.4.2
S-Succinyl-dihydrolipoamide-E
3-Carboxy-1-hydroxypropyl-ThPP
Arginine biosynthesis
Ascorbate and aldarate metabolism
Alanine, aspartate and glutamate metabolism
D-Amino acid metabolism
1.8.1.4
Dihydro-lipoamide-E
Lipoamide-E
Val , Leu & Ile degradation
1.2.7.11
1.2.7.3

Network inference from a transcriptomic dataset


Experimental design
untreated
treated
Isolate RNA
AAAAAAAAA
AAAAAAAAA
AAAAAAAAA
AAAAAAAAA
Prepare library
Sequence
ATGCGATAGCCAGATGA
Single reads
Paired end reads
FASTQ files


AAAAAAAAA
TTTTTTTTTTB
1) PolyA+ RNA captured
2) RNA fragmented and primed
3) First strand cDNA synthesized
4) Second strand cDNA synthesized
p
A
A
p
5) 3’ ends adenylated and 5’ ends repaired
6) DNA sequencing adapters ligated
Barcode
Rd1
Rd2
Index
7) Ligated fragments PCR amplified

RNA-seq data analysis: typical steps



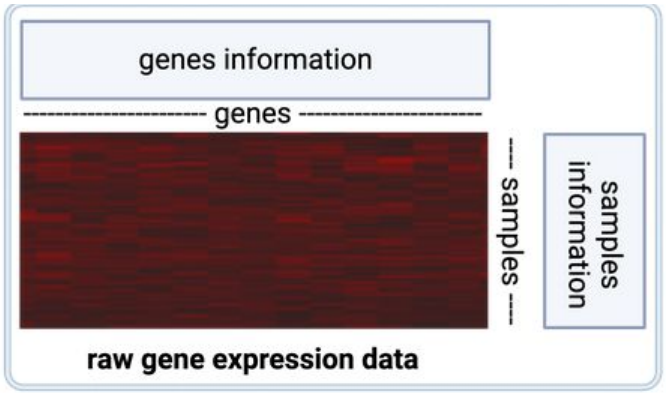
	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3
Gene A	6	5	7	17	10	11
Gene B	11	11	10	3	4	2
Gene C	200	150	355	50	1	3
Gene D	0	1	0	2	0	1

STEP:	TOOLS:	FILE:
Quality control	FastQC	FASTQ
Pre-processing	Trimmo-matic	FASTQ
Alignment	TopHat	BAM
Quality control	RSeQC	
Quantitation	HTSeq	Read count file (TSV)
Combine count files to table	Define NGS experiment	Read count table (TSV)
Quality control	PCA, clustering	
Differential expression analysis	DESeq2, edgeR	Gene lists (TSV)

CSC

Weighted Gene Co-Expression Network Analysis

Initial inputs



1. calculating correlation matrix

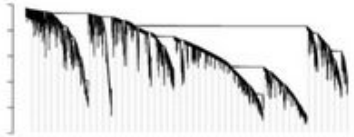
	Gene A	Gene B	Gene C	Gene D	Gene E
Gene A	0	0.98	1	0.31	0.31
Gene B	0.98	0	0.98	0.22	0.22
Gene C	1	0.98	0	0.31	0.31
Gene D	0.31	0.22	0.31	0	0.99
Gene E	0.31	0.22	0.31	0.99	0

2. calculating adjacency matrix

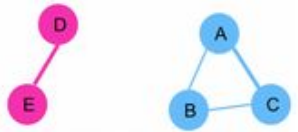
$|cor(t(matrix))|^{\beta}$, β = high power

	Gene A	Gene B	Gene C	Gene D	Gene E
Gene A	0	0.90	1	0	0
Gene B	0.90	0	0.90	0	0
Gene C	1	0.90	0	0	0
Gene D	0	0	0	0	0.98
Gene E	0	0	0	0.98	0

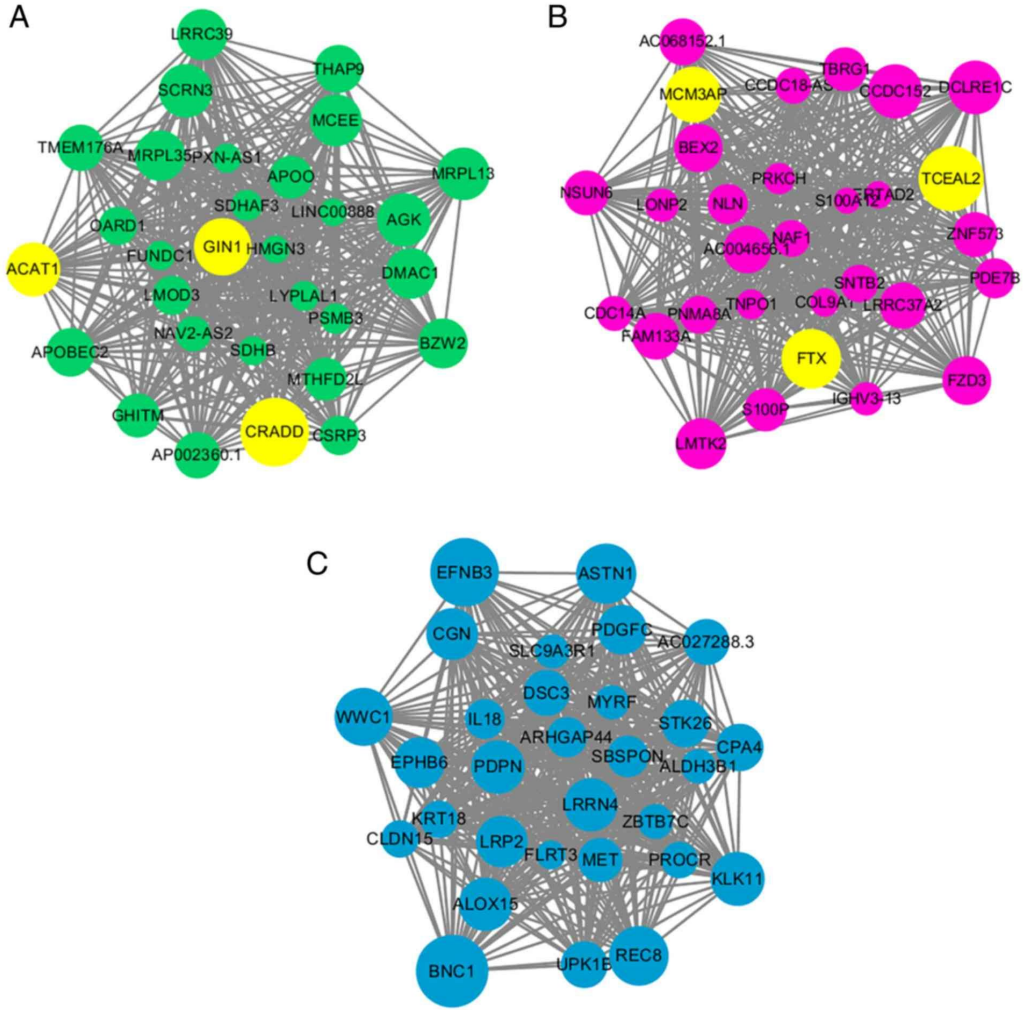
3. calculating dissimilarity matrix and hierarchical clustering



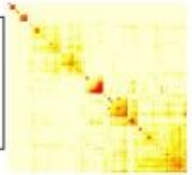
4. Identifying gene modules



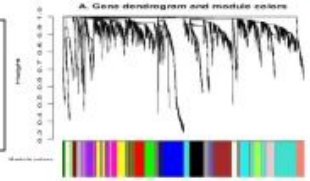
(Langfelder & Horvath, BMC Bioinformatics, 2008)


A
LRRC39
THAP9
SCRN3
MCEE
TMEM176A
MRPL35
PXN-AS1
APOO
MRPL13
SDHAF3
LINC00388
AGK
OARD1
GIN1
HMGN3
FUNDC1
ACAT1
DMAC1
LMOD3
LYPLAL1
PSMB3
NAV2-AS2
BZW2
APOBEC2
SDHB
MTHFD2L
GHITM
CRADD
CSRP3
AP002360.1
B
AC068152.1
TBRG1
MCM3AP
CCDC18-AS
CCDC152
DCLRE1C
BEX2
TCEAL2
PRKCH
NSUN6
LONP2
NLN
S100A12
ERVAD2
NAF1
ZNF573
AC004656.1
SNTB2
PDE7B
CDC14A
PNMA8A
TNPO1
COL9A1
LRRC37A2
FAM133A
FTX
FZD3
IGHV3-13
S100P
LMTK2
C
EFNB3
ASTN1
CGN
PDGFC
AC027288.3
SLC9A3R1
DSC3
MYRF
WWC1
IL18
STK26
ARHGAP44
CPA4
SBSPON
EPHB6
PDPN
ALDH3B1
LRRN4
KRT18
ZBTB7C
CLDN15
LRP2
FLRT3
MET
PROCR
KLK11
ALOX15
BNC1
UPK1B
REC8

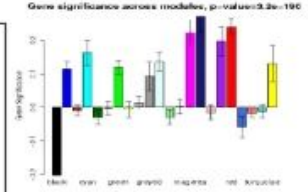
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



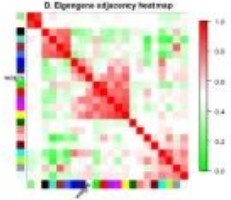
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



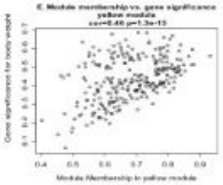
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



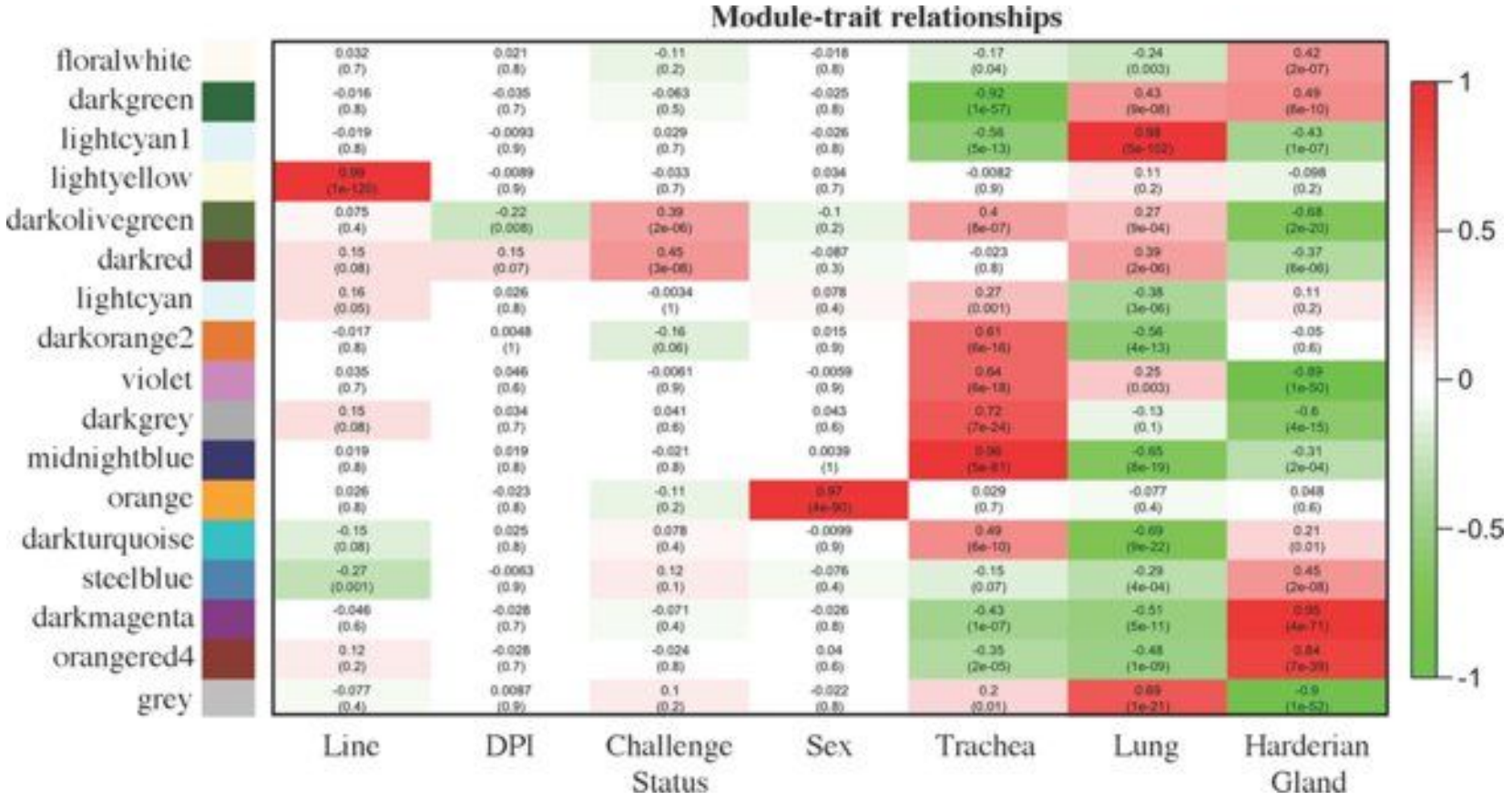
Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks



Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



(Langfelder & Horvath, BMC Bioinformatics, 2008)


Module-trait relationships
floralwhite
darkgreen
lightcyan1
lightyellow
darkolivegreen
darkred
lightcyan
darkorange2
violet
darkgrey
midnightblue
orange
darkturquoise
steelblue
darkmagenta
orangered4
grey
Line
DPI
Challenge Status
Sex
Trachea
Lung
Harderian Gland
1
0.5
0
-0.5
-1

Gene	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
G1	1	0.8	0.7	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
G2	0.8	1	0.75	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
G3	0.7	0.75	1	0.1	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0
G4	0.2	0.2	0.1	1	0.8	0.75	0.0	0.0	0.0	0.0	0.0	0.0
G5	0.1	0.1	0.1	0.8	1	0.7	0.0	0.0	0.0	0.0	0.0	0.0
G6	0.1	0.1	0.2	0.75	0.7	1	0.0	0.0	0.0	0.0	0.0	0.0
G7	0.0	0.0	0.0	0.0	0.0	0.0	1	0.85	0.8	0.0	0.0	0.0
G8	0.0	0.0	0.0	0.0	0.0	0.0	0.85	1	0.75	0.0	0.0	0.0
G9	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.75	1	0.0	0.0	0.0
G10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	0.85	0.8
G11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.85	1	0.75
G12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.75	1

Conclusion

- **Omics revolution**: From genomics to metabolomics, technologies now allow us to measure biology at scale.
- **Integration across omics**: Genotype → Transcriptome → Proteome → Metabolome → Epigenome.
- **From data to systems**: Systems biology links these layers to understand biology as a whole.
- **Networks as a unifying language**: Genes, proteins, and metabolites can be studied as **interacting networks**, revealing emergent properties.
- **Applications**: Evolution, development, disease progression, personalized medicine.