Introduction to Statistical Genetics and Genomics

Umut Özbek, PhD
Associate Professor of Biostatistics
Tisch Cancer Institute
Department of Population Health Science & Policy

Outline

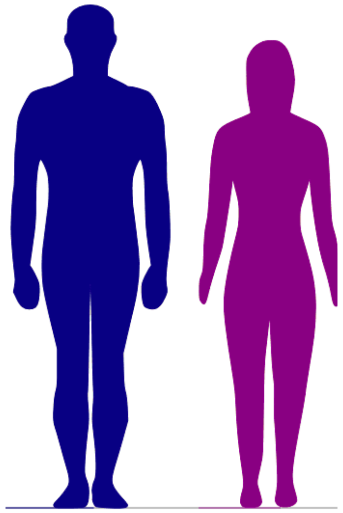
- Part I
  - Definitions & background
  - Statistical methods
- Part II
  - More advanced statistical modeling
  - Network Analysis
  - An example

Part I
Genetics and Statistics

What is statistical genetics ?

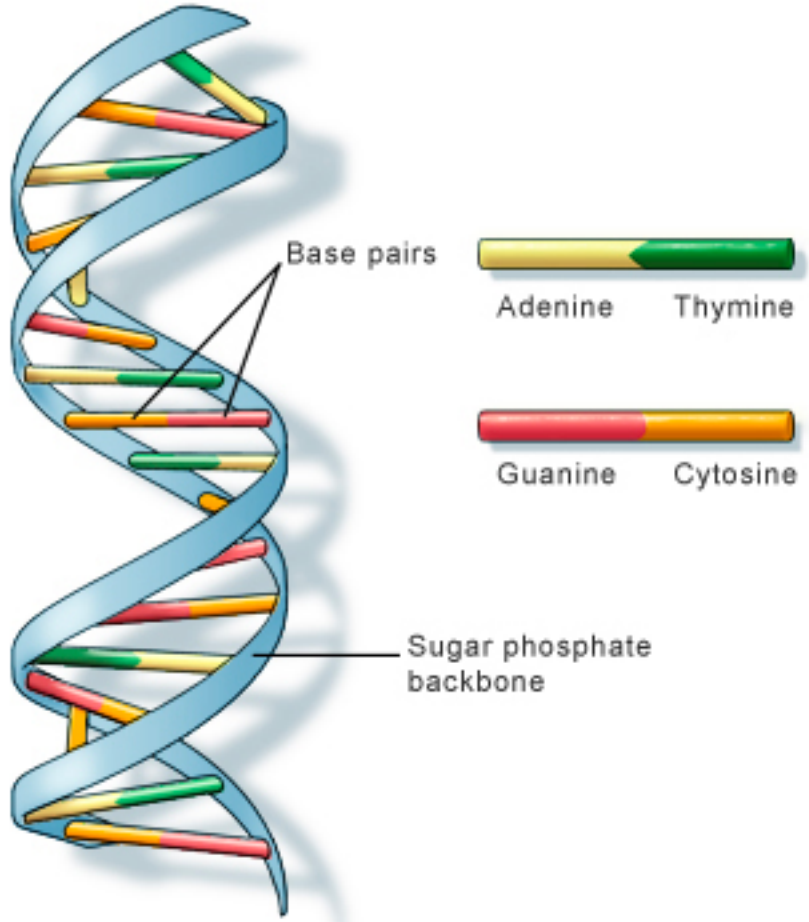
- Statistical methods to analyze and make inferences from genetic data.

  - Discover genetic variants associated with traits or diseases



What we've learnt so far

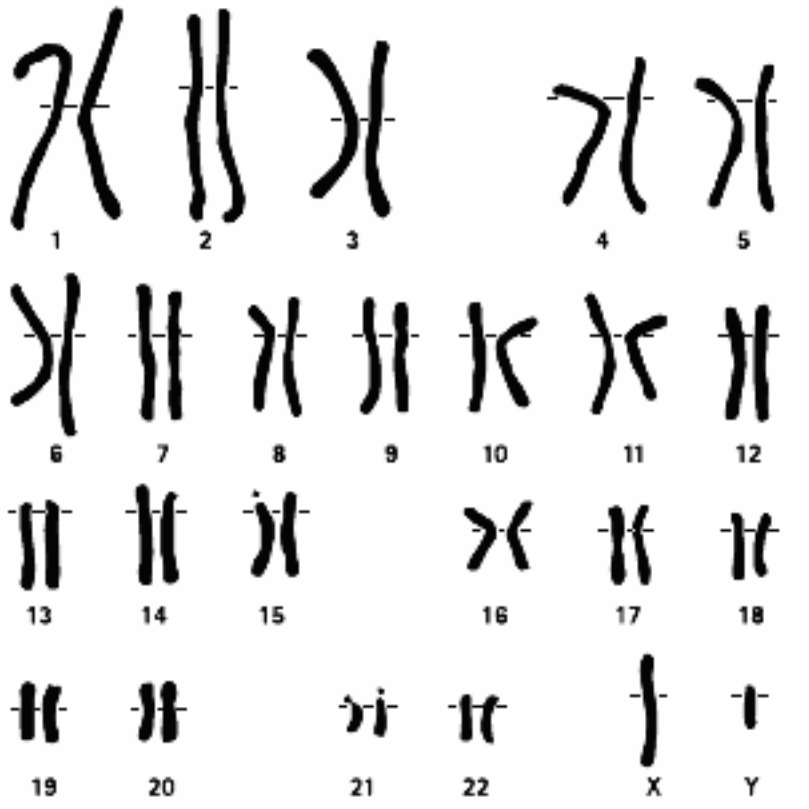
- DNA: deoxyribonucleic acid, the hereditary material in organisms



U.S. National Library of Medicine

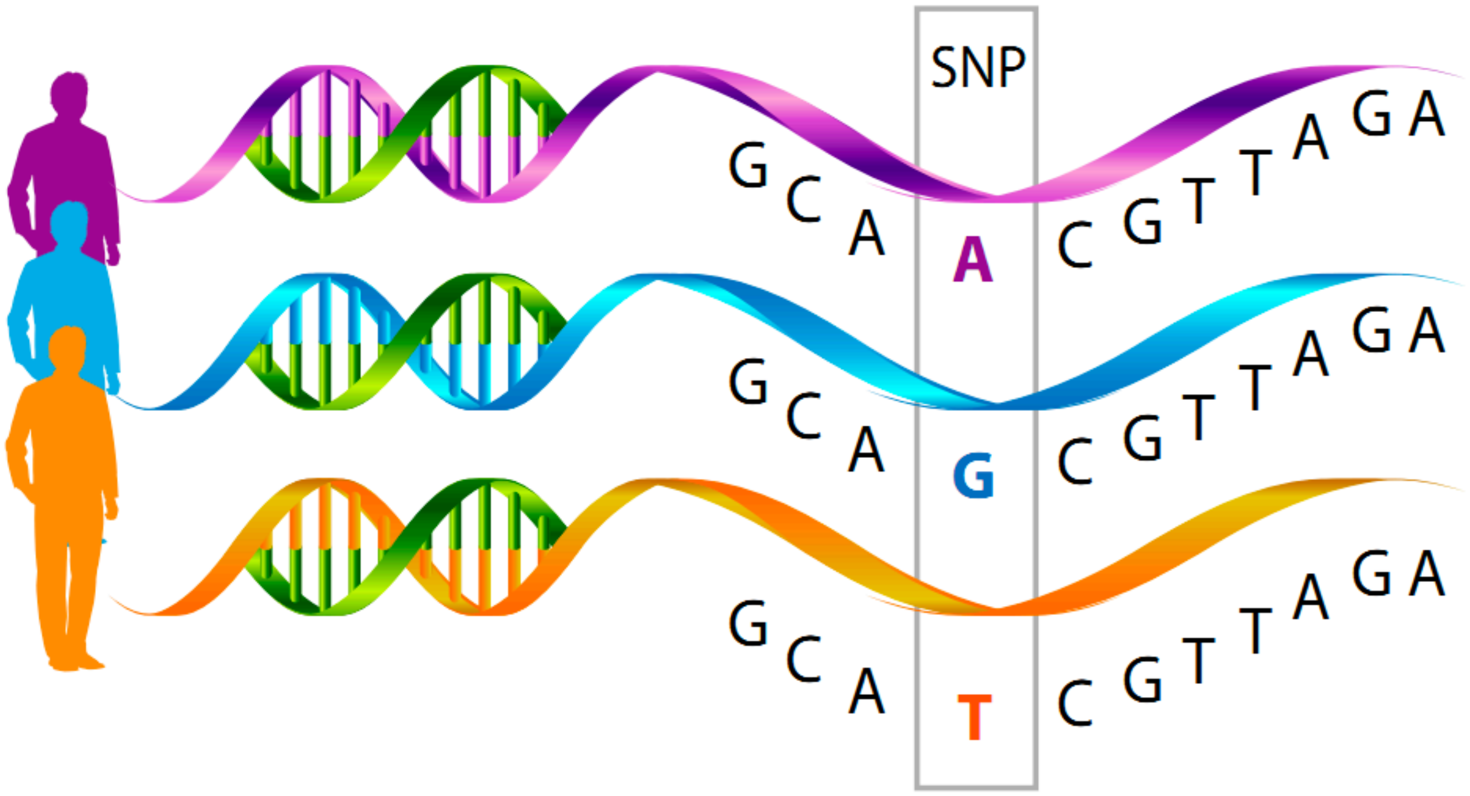
What we’ve learnt so far

- Genome: organism’s complete set of DNA
  - Human genome: 22 autosomal + sex chromosomes



What we’ve learnt so far

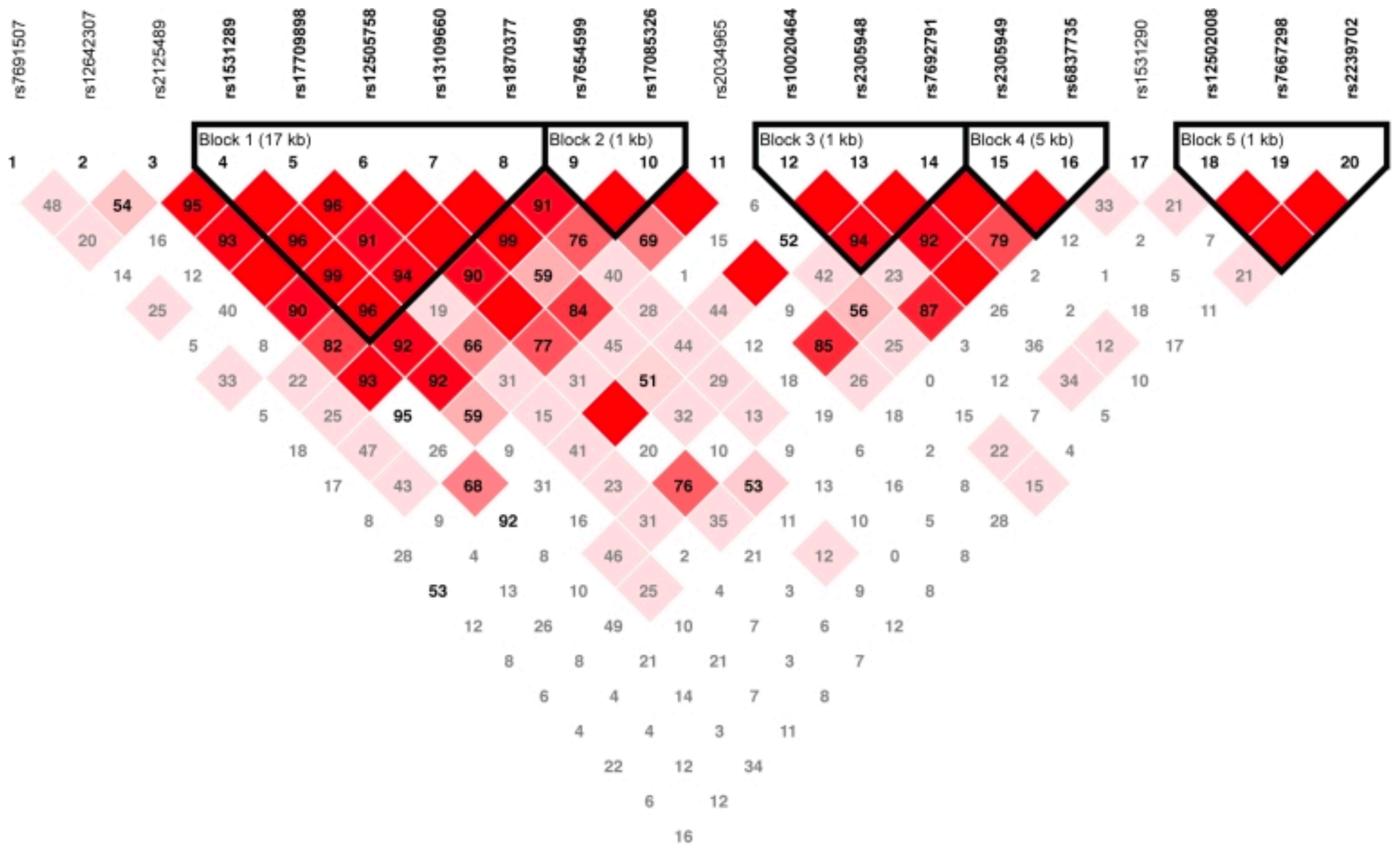
- SNP: single nucleotide polymorphisms, the most common type of genetic variation among people.



http://cahartmanfiction.com/understanding-genetics-snps/

What we've learnt so far

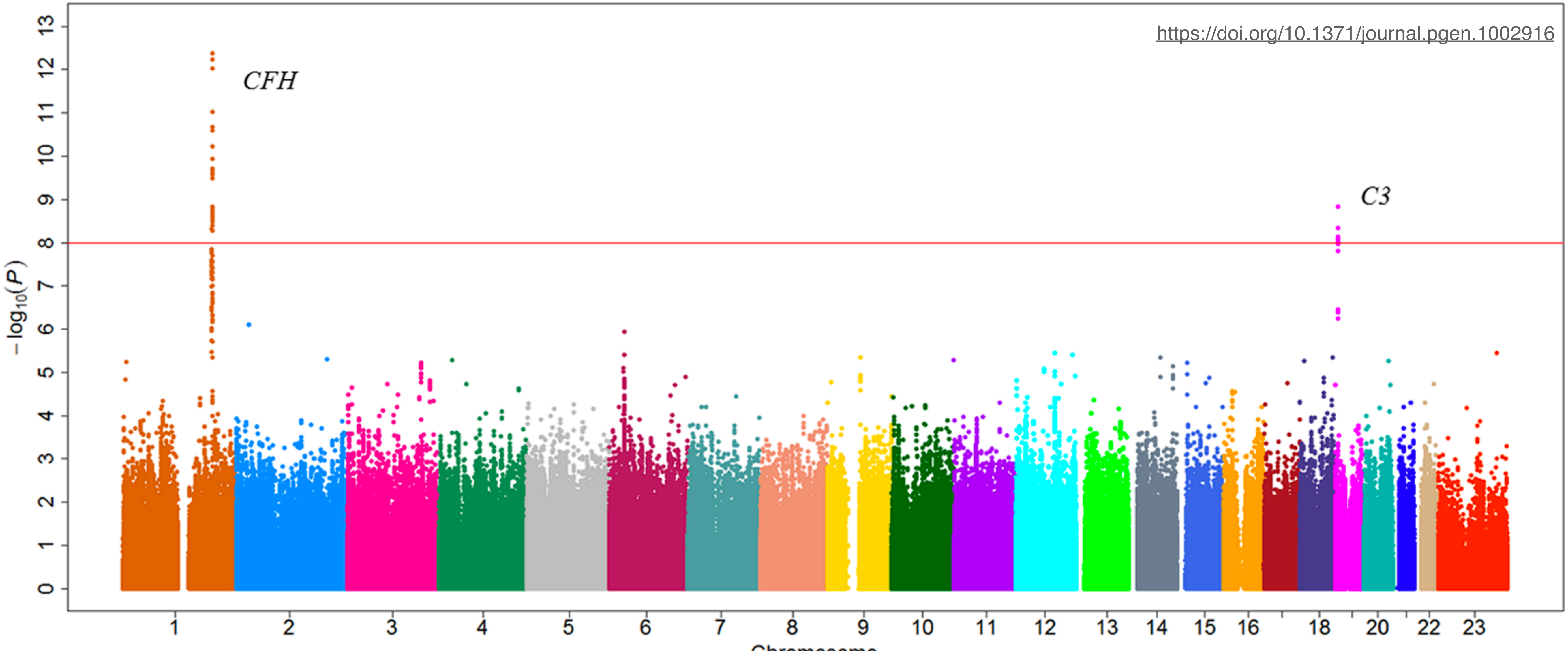
- Linkage Disequilibrium: non-random association of alleles at different loci



http://openi.nlm.nih.gov/detailedresult.php?img=2793900_mv-v15-2710-f2&req=4

What we’ve learnt so far

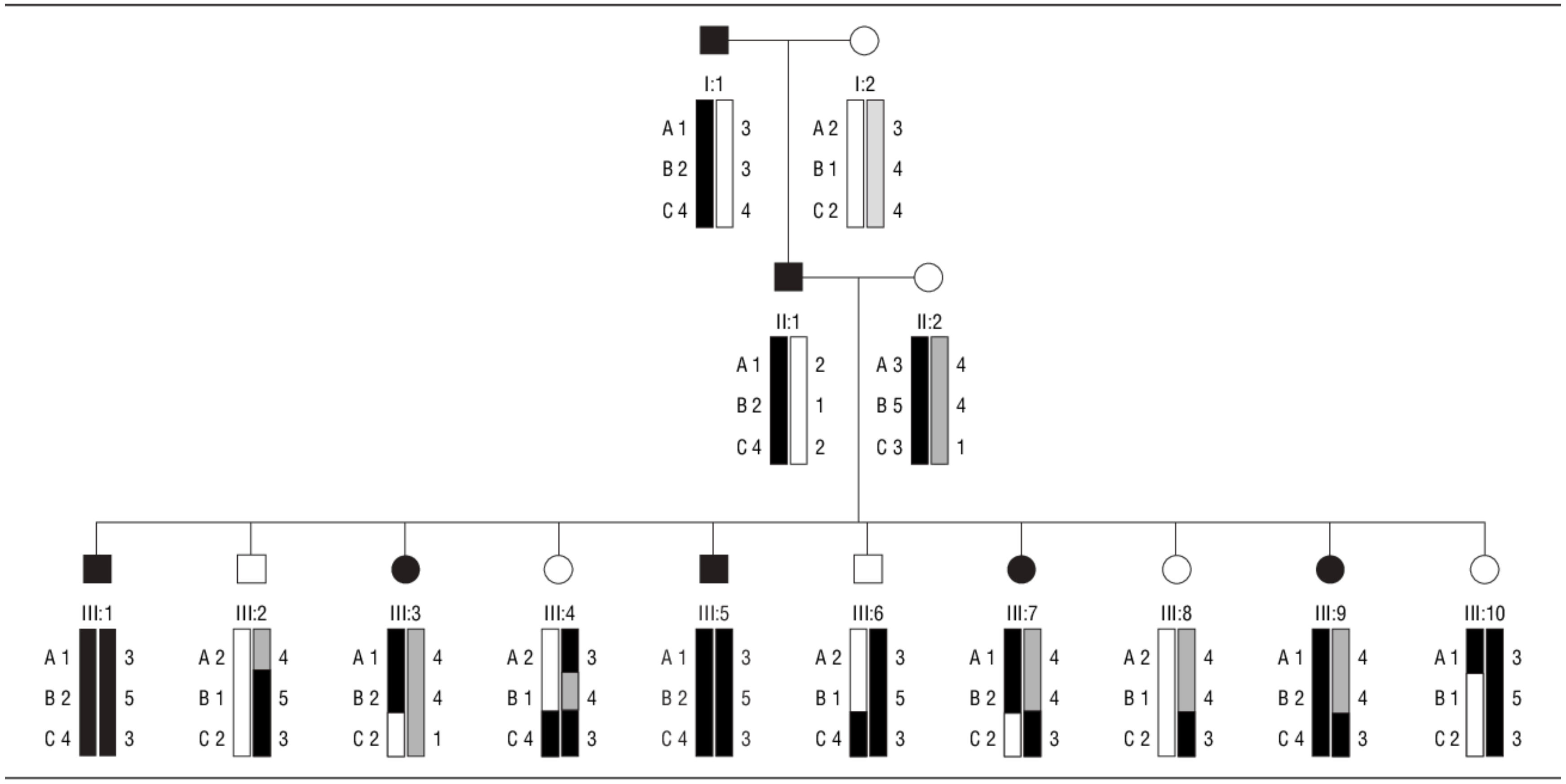
- Association analysis: statistical method to identify disease susceptibility variants that contribute to a specific disease or phenotype.



https://doi.org/10.1371/journal.pgen.1002916

What we've learnt so far

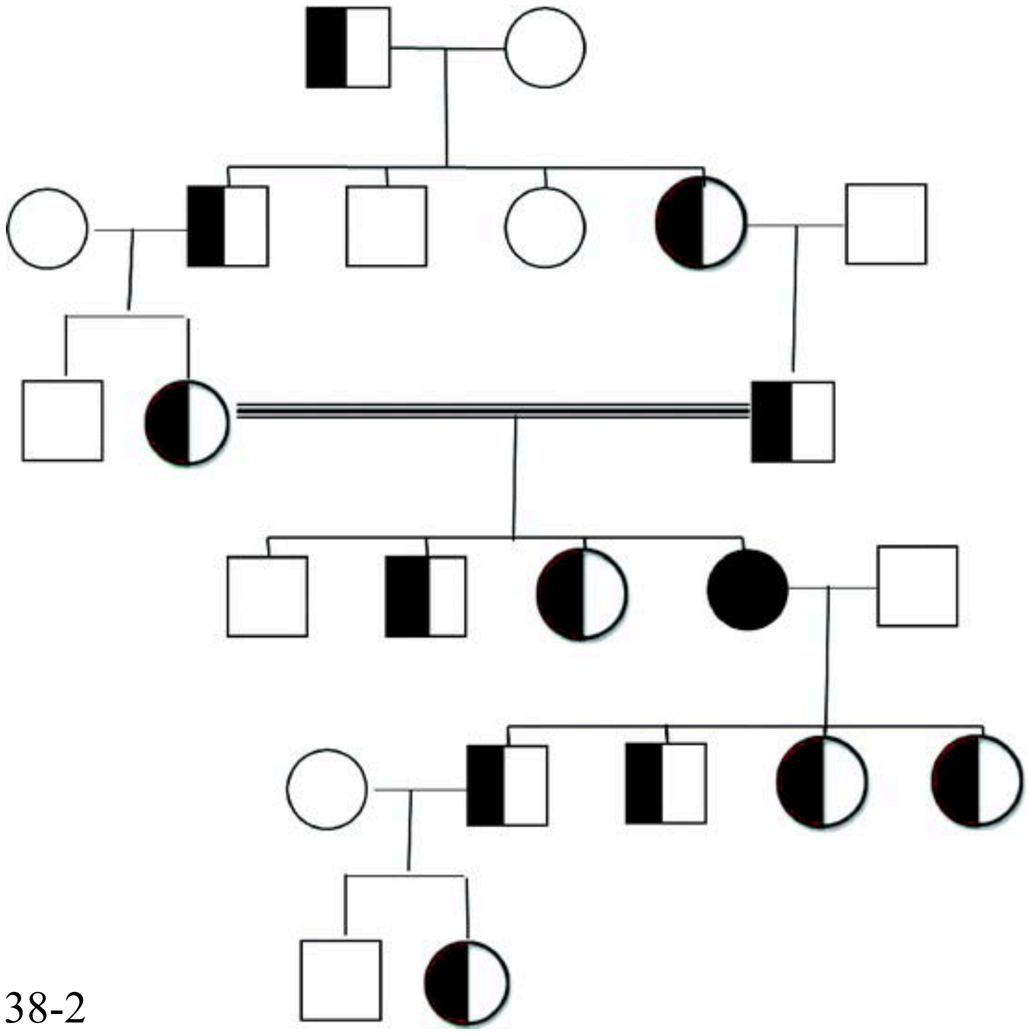
- Linkage analysis: statistical method for mapping the genes for heritable traits to their chromosome locations



10.1001/ARCHNEUR.56.6.667

What we've learnt so far

- Segregation analysis: statistical technique that attempts to explain the causes of family aggregation of disease



https://link.springer.com/book/10.1007/978-1-4419-7338-2

Association Analysis

- Based on unrelated study subjects
- Aims to find an association between a disease trait and genetic marker
- Study designs:
  - Case-control
  - Cohort

Statistical Methods for Association Analysis

Case-control study:

- Z/Chi-squared/Fisher test
- Logistic regression

Statistical Methods for Association Analysis

Case-control study:

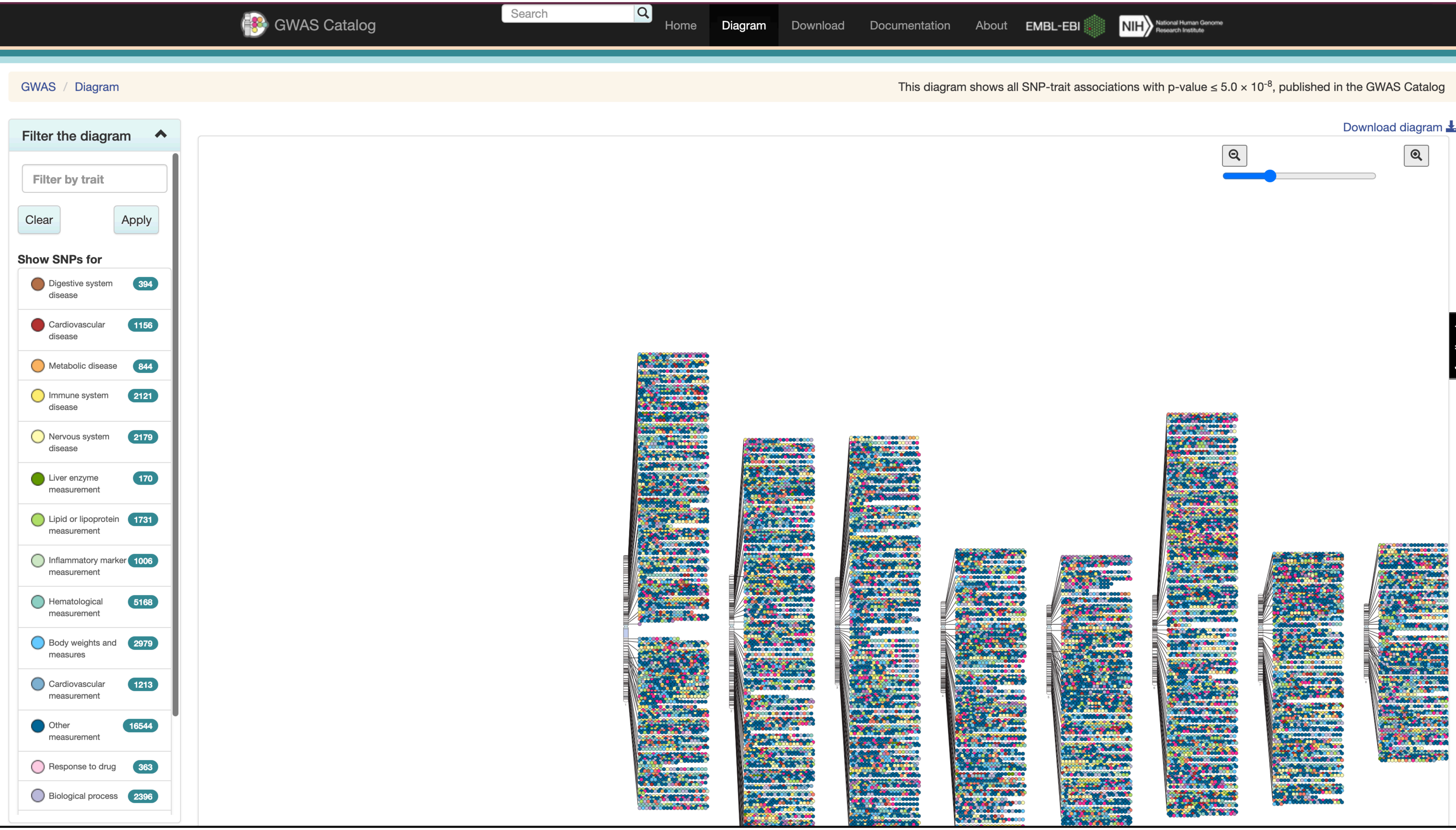
	AA	Aa	aa
Case	n_{11}	n_{12}	n_{13}
Control	n_{21}	n_{22}	n_{23}

Statistical Methods for Association Analysis

Case-control study:

	AA/Aa	aa
Case	n_{11}	n_{12}
Control	n_{21}	n_{22}

GWAS Catalog
https://www.ebi.ac.uk/gwas/diagram



GWAS / Diagram

This diagram shows all SNP-trait associations with p-value ≤ 5.0 × 10^{-8}, published in the GWAS Catalog

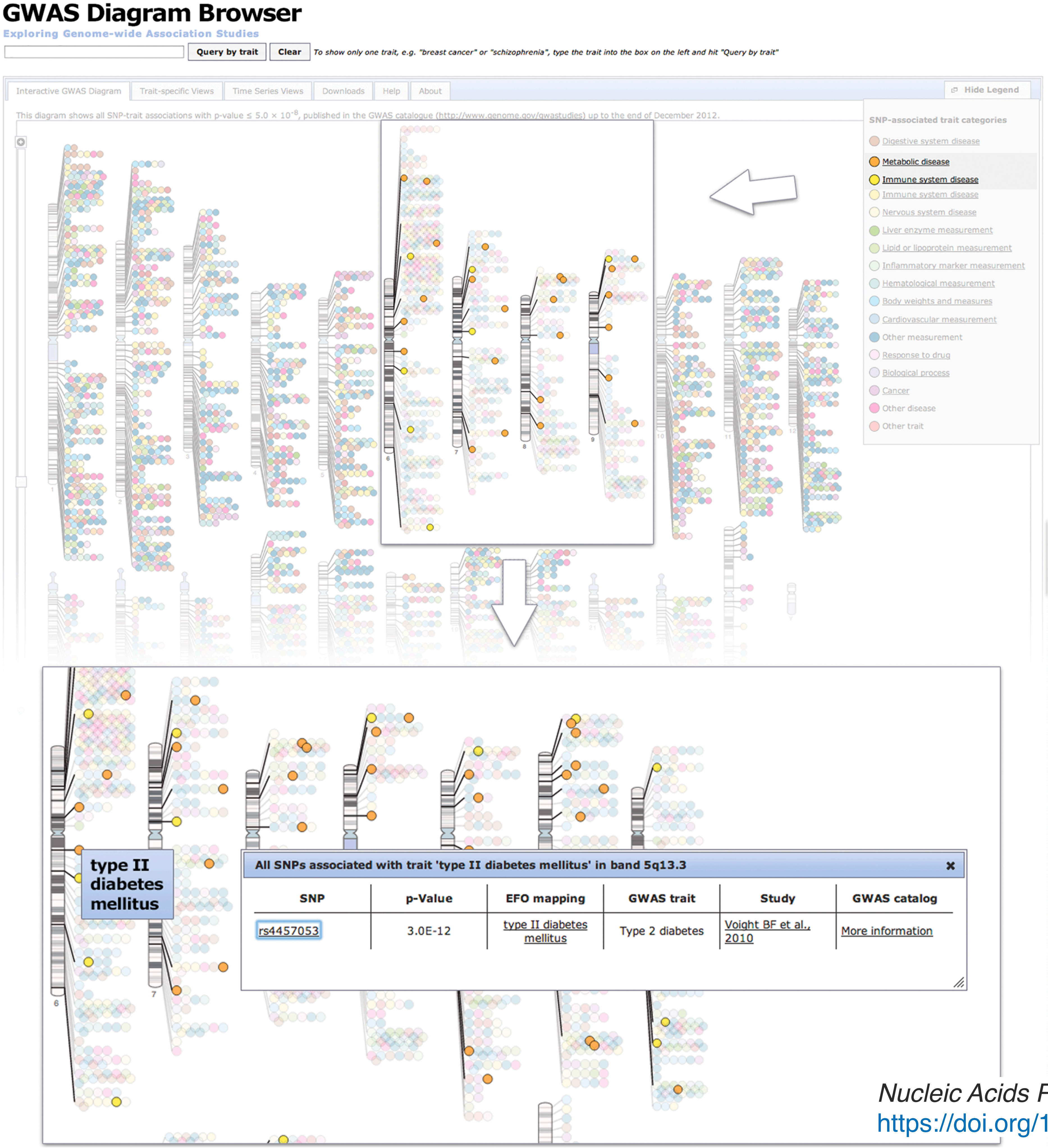
Filter the diagram

Filter by trait

Clear Apply

Show SNPs for

● Digestive system disease	394
● Cardiovascular disease	1156
● Metabolic disease	844
● Immune system disease	2121
● Nervous system disease	2179
● Liver enzyme measurement	170
● Lipid or lipoprotein measurement	1731
● Inflammatory marker measurement	1006
● Hematological measurement	5168
● Body weights and measures	2979
● Cardiovascular measurement	1213
● Other measurement	16544
● Response to drug	363
● Biological process	2396



GWAS Diagram Browser

Exploring Genome-wide Association Studies

To show only one trait, e.g. "breast cancer" or "schizophrenia", type the trait into the box on the left and hit "Query by trait"

This diagram shows all SNP-trait associations with p-value ≤ 5.0×10^{-8}, published in the GWAS catalogue (http://www.genome.gov/gwastudies) up to the end of December 2012.

All SNPs associated with trait 'type II diabetes mellitus' in band 5q13.3

SNP	p-Value	EFO mapping	GWAS trait	Study	GWAS catalog
rs4457053	3.0E-12	type II diabetes mellitus	Type 2 diabetes	Voight BF et al., 2010	More information

Nucleic Acids Research

https://doi.org/10.1093/nar/gkt1229

Population Stratification/Admixture

- Allele frequencies can differ substantially between different subpopulations
- Risk of disease can also differ substantially
- Inflates the number of false positive findings if not accounted for

Multiple Testing

- GWAS tests many markers for genetic association
- For a single marker:

$$\alpha' = P(\text{reject null hypothesis } H^{(m)} \mid H^{(m)} \text{ is true})$$

- For multiple markers:

$$\alpha = 1 - P(\text{not reject any } H^{(m)} \mid H^{(m)} \text{ is true for all } m)$$
$$= 1 - (1 - \alpha')^{M} = 1$$

Multiple Testing

- Bonferroni correction method: α/M
- Threshold of 5×10^{-8}

Some more statistics…

Simple Linear Regression

Goals:

- To identify the form of the functional relationship between two variables X and Y
- To construct a mathematical model that best fits the data

Terminology

X: the independent variable or the predictor

Y: the dependent variable or the response

Model Assumptions

- The independent variable X is fixed, i.e. its values are predetermined or chosen in advance
- The independent variable X is measured without error

This generates a FIXED EFFECT model

Model Assumptions

For each value of X there exists a sub-population of Y values with the following characteristics

Linearity
Independency
Normality
Equal Variance

Linearity

The means of the sub-populations Y all lie on the same straight line

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

m$_{Y|X}$ is the mean of the subpopulation Y given a particular value of X

Independency

- The Y values are statistically independent
- The subpopulations of Y given X are independent

Normality

- The subpopulations of Y|X are all normally distributed

Equal Variance

- The subpopulations of Y|X all have the same variance s^2

Regression Equation

All regression assumptions can be summarized by the regression equation

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

where

$$e_i \sim N(0, \sigma^2)$$

It follows that

$$y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Model

Theoretical model:

$$y = \beta_0 + \beta_1 X + e$$

Dependent or response variable

intercept

Slope

Independent or explanatory variable or predictor

error

Model

Estimated model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Expected or predicted Y value

Estimated Intercept

Estimated Slope

predictor

Interpretation

$\hat{y}$ Is the value of the dependent variable that we predict from the model

$\hat{\beta}_0$ Is the predicted value of the dependent variable for X=0

$\hat{\beta}_1$ Indicates the rate of change in Y for each unit increment of X

The sign of $\hat{\beta}_1$ indicates the direction of the change

The magnitude of $\hat{\beta}_1$ indicates the speed of change

Multiple Linear Regression

•Multiple regression is an extension of the simple regression model where more than one predictors are considered.

Example:

Simple linear model: SBP is a function of sodium intake

Multiple regression: SBP is a function of sodium intake, age, exercise, etc.

Multiple Linear Regression

- To improve the predicting ability of the model
- To test interactions among predictors
- To correct for the effect of confounders
- To improve goodness of fit

Multiple Linear Regression

1.More difficult to choose – more candidate predictors are available

2.More difficult to visualize – multidimensional space

3.More difficult to interpret – many predictors

4.More difficult notation – matrix notation needed for calculations

5.Computations: cannot do by hand – need computer

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_p X_p + e$$

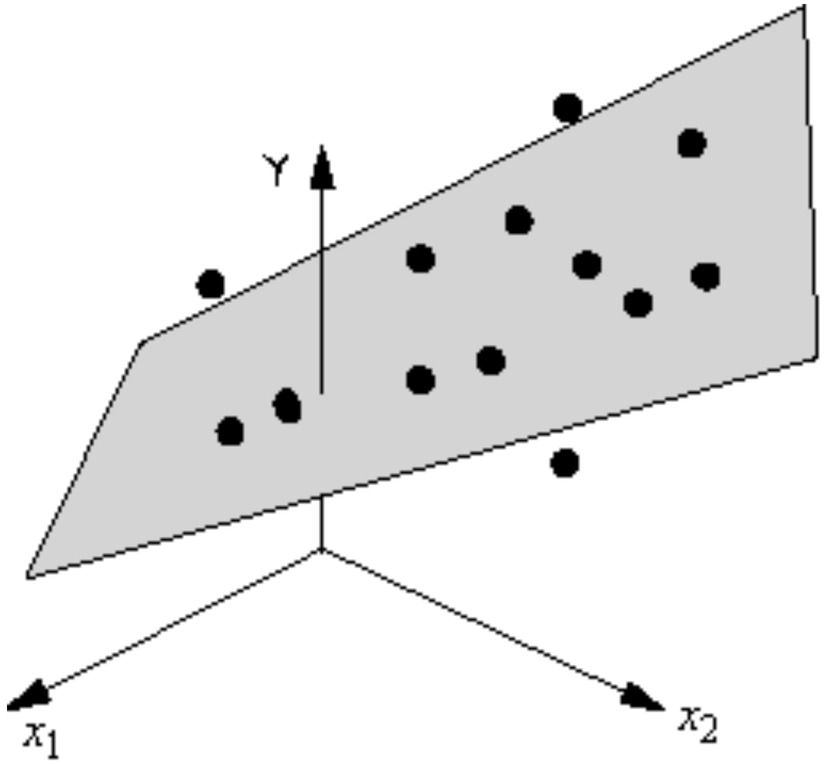
where $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the regression coefficients,

and $X_1, X_2, \ldots, X_p$ are the predictors

Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

Model with two predictors
The model defines a plane in a three dimensional space (*Y*, *X₁*, *X₂*)



Multiple Linear Regression

EXAMPLE: Y = SBP, X_1 = sodium intake, X_2 = Age

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

$\hat{\beta}_0$ Is the value of SBP when BOTH sodium intake and age are =0

$\hat{\beta}_1$ Is the change in SBP for each unit change in sodium intake when age IS HELD CONSTANT

$\hat{\beta}_2$ Is the change in SBP for each unit change in age when sodium intake IS HELD CONSTANT

Model Assumptions

As for the simple linear model

1. Linearity
2. Independency
3. Normality
4. Equal variance

Logistic Regression

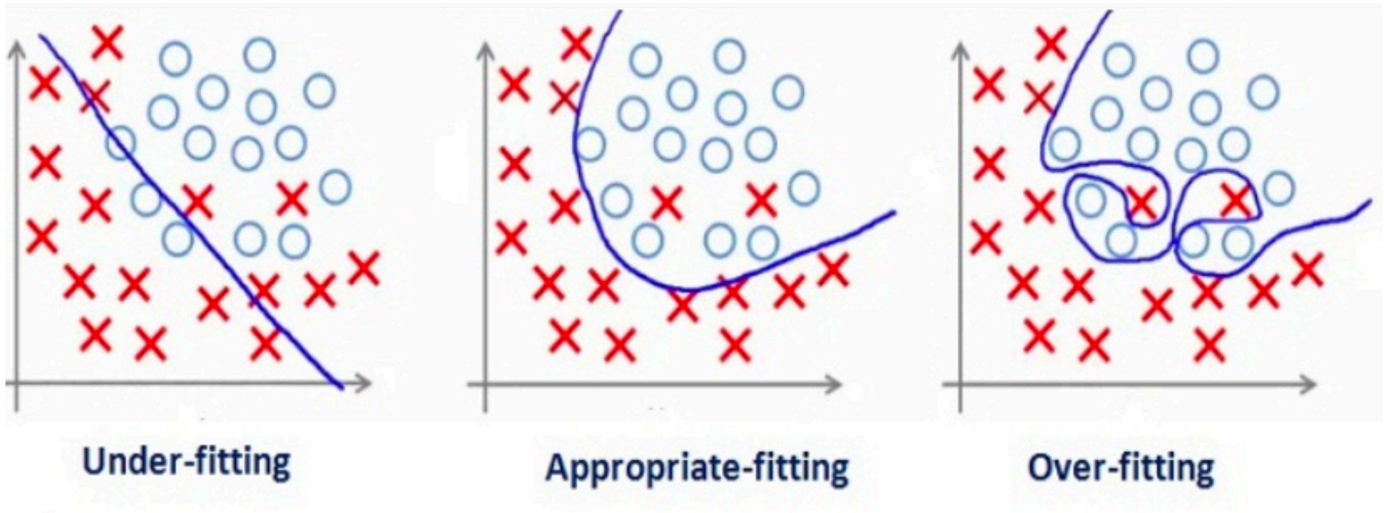
- Outcome is discrete, not continuous
- Involves a more probabilistic view of classification

$$z = \alpha + \boldsymbol{\beta} \cdot \mathbf{x} = \alpha + \beta_1 x_1 + \cdots + \beta_d x_d$$

$$z = \log\left(\frac{p}{1-p}\right) \qquad \text{logit function}$$

Regularization

- Shrink the coefficients in the resulting model to avoid overfit by penalizing certain values of the weights
- Helps the computational problem
- Helps with generalization



Part II
Genomics and Network Analysis

Introduction to Statistical Genetics and Genomics

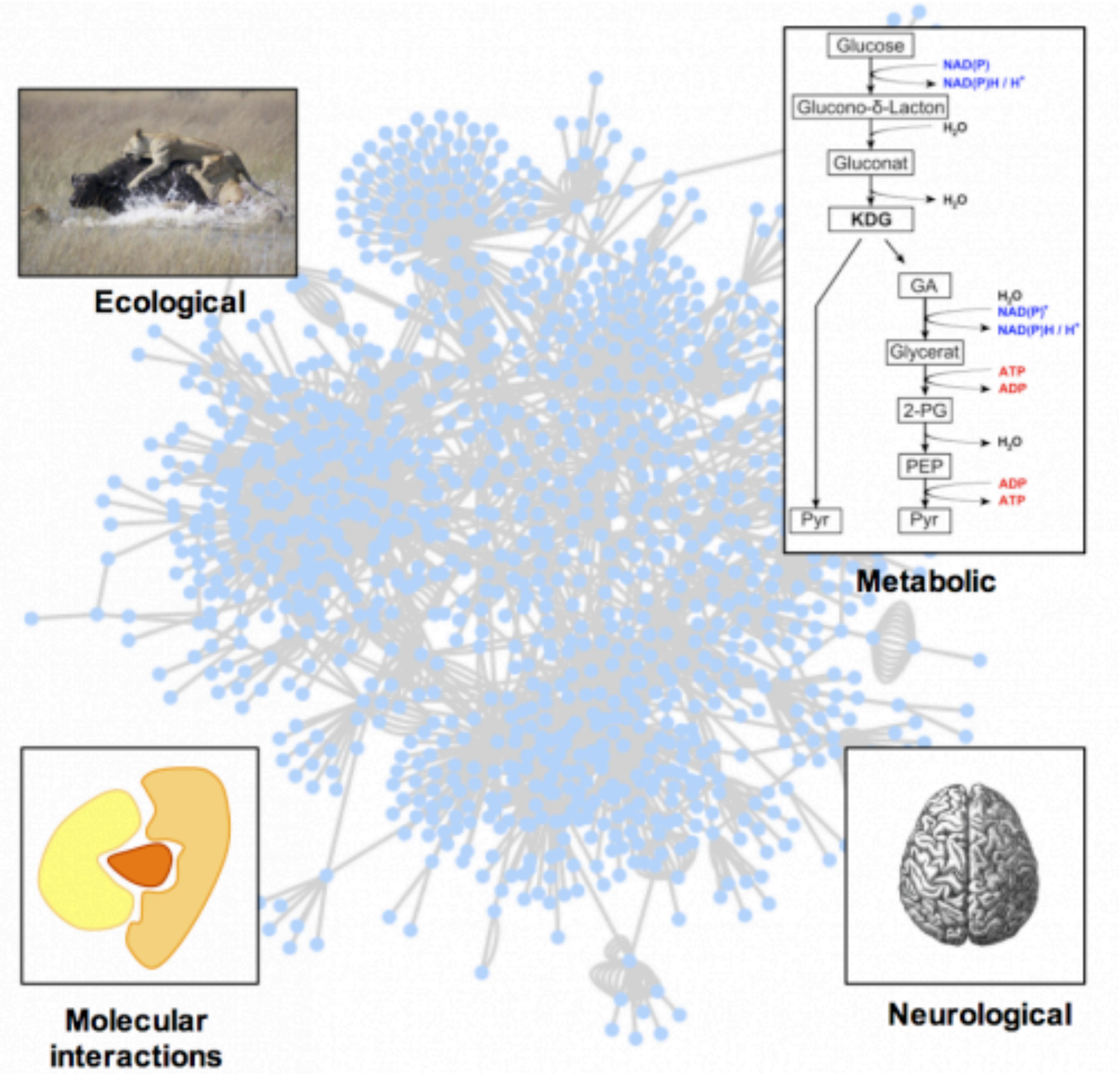
Umut Özbek, PhD
Associate Professor of Biostatistics
Tisch Cancer Institute
Department of Population Health Science & Policy

Part II
Genomics and Network Analysis

Network Analysis

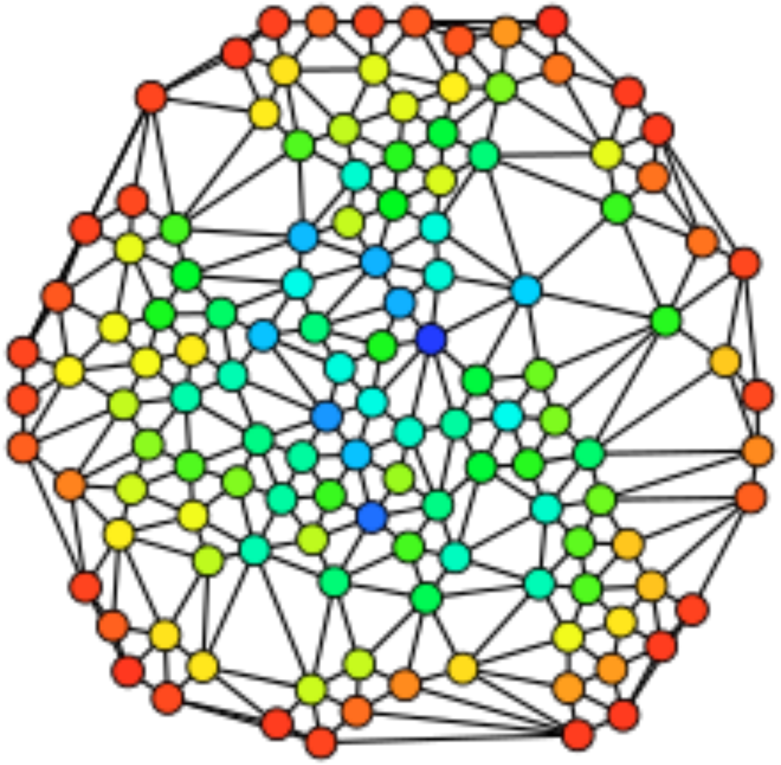
Network analysis in biology

Biological systems are often represented as networks which are complex sets of binary interactions between entities.



https://www.ebi.ac.uk

Graph Theory



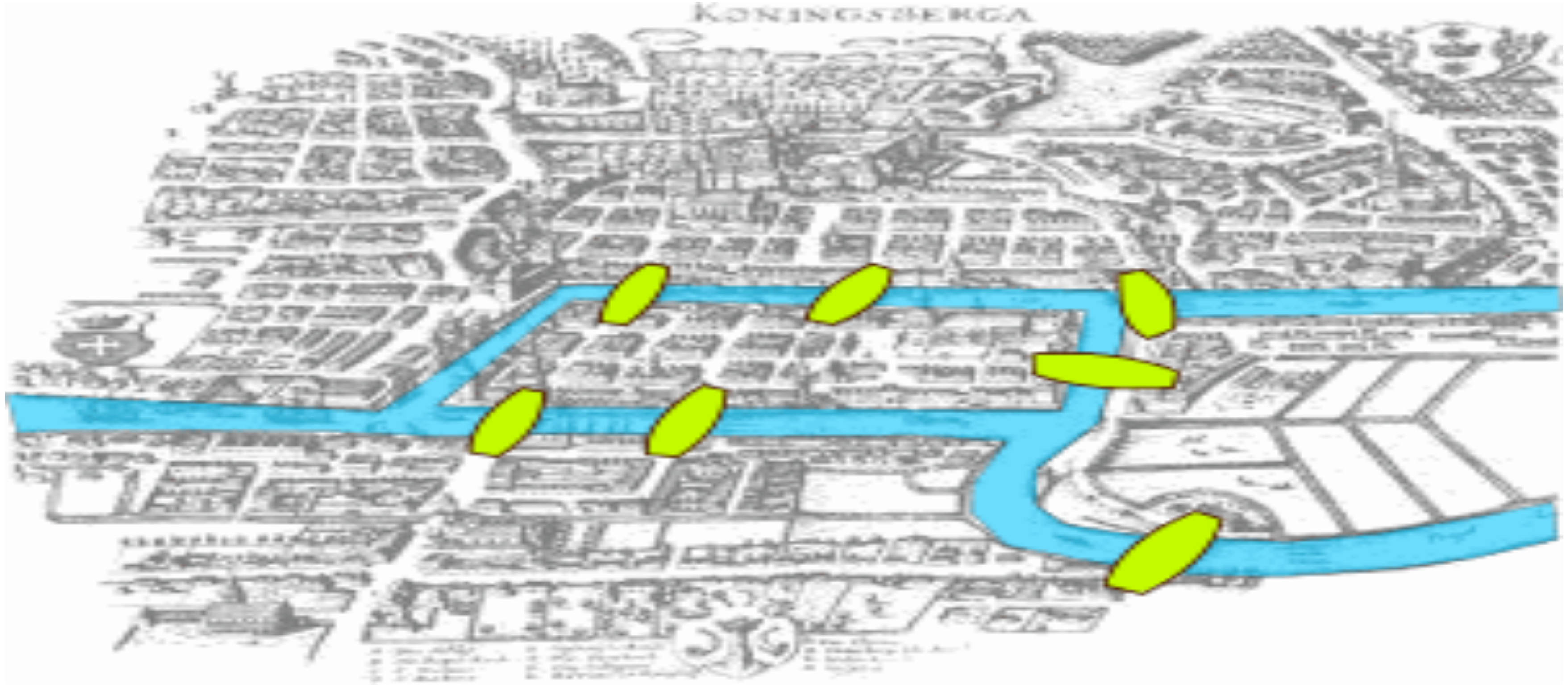
- Set of abstract concepts that can be used to visualize and analyze networks.
- Made up of nodes which are connected by edges.
- Topology is the way in which the nodes and edges are arranged within a network.

Some history…

- The idea of topology first described by the Swiss mathematician Leonhard Euler: Seven bridges of Königsberg

Seven bridges of Königsberg

Four islands connected by seven bridges



Is there a path that visited all four islands and crossed each of the bridges only once?

Seven bridges of Königsberg

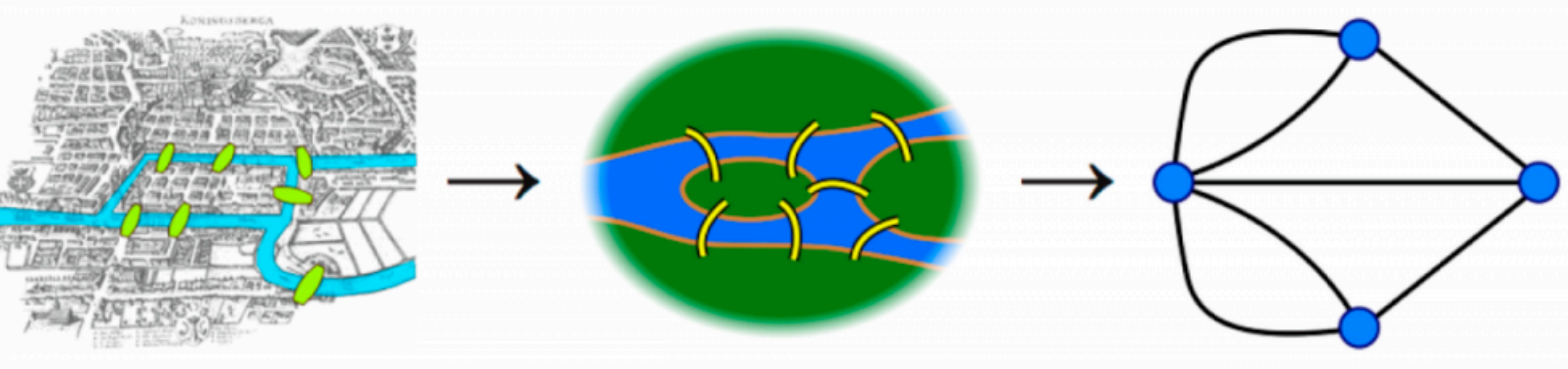
Euler:

-only the relations between the land masses are relevant.

-land masses are the nodes.

-bridges are the edges.

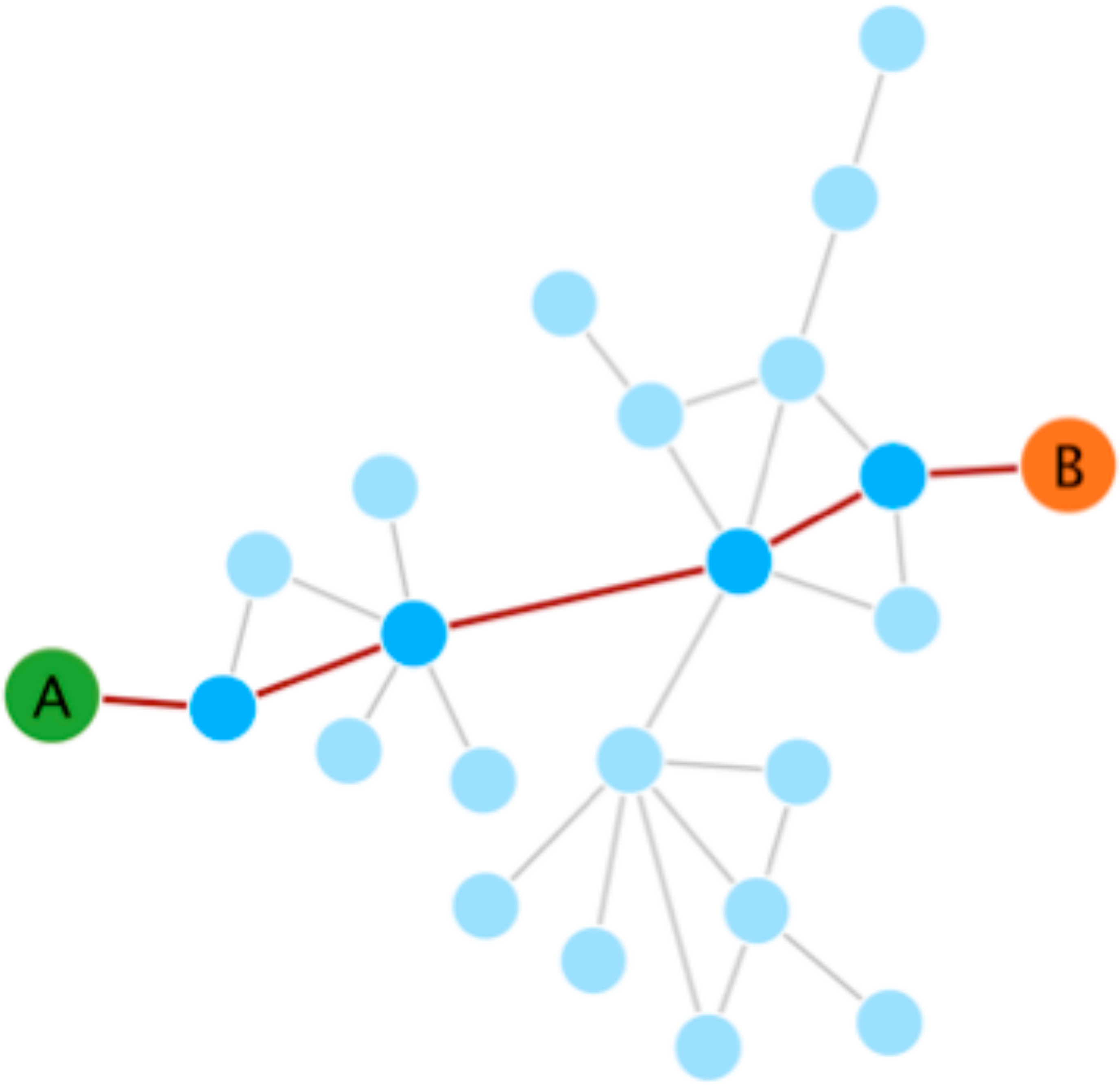
-the path does not exist.



Network topology

Topological properties:

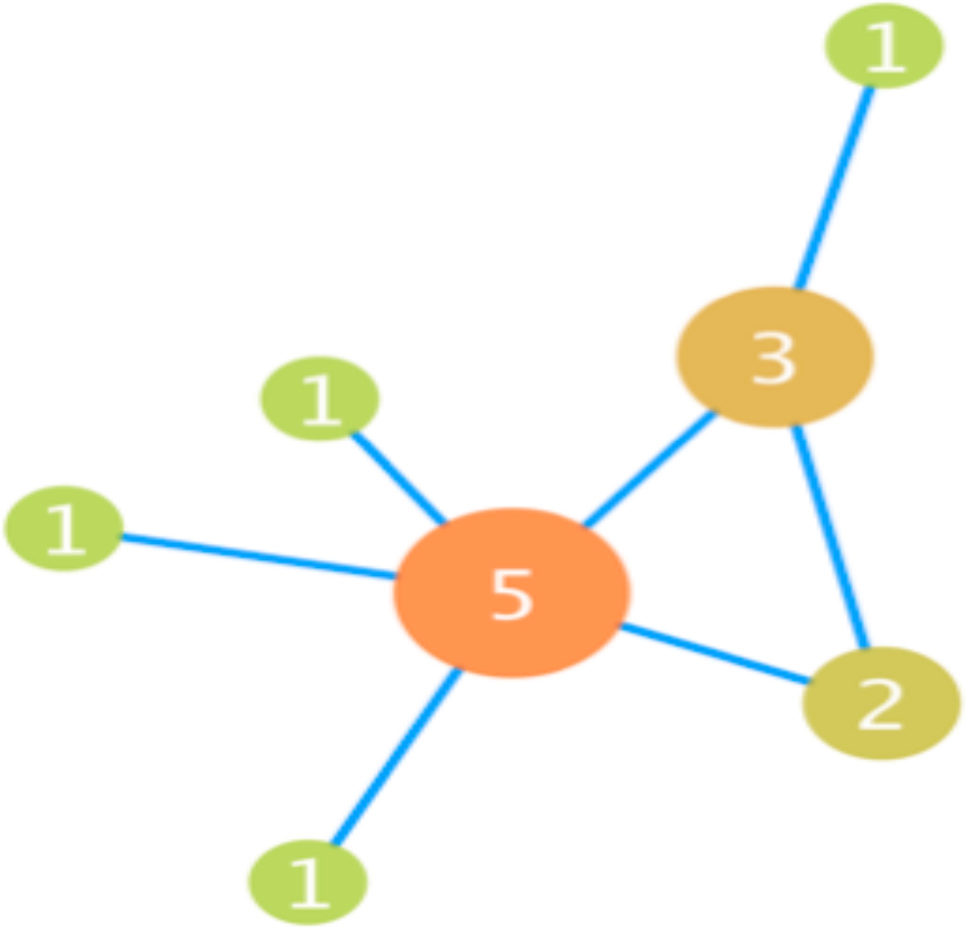
- Shortest path: shortest distance between any two node.



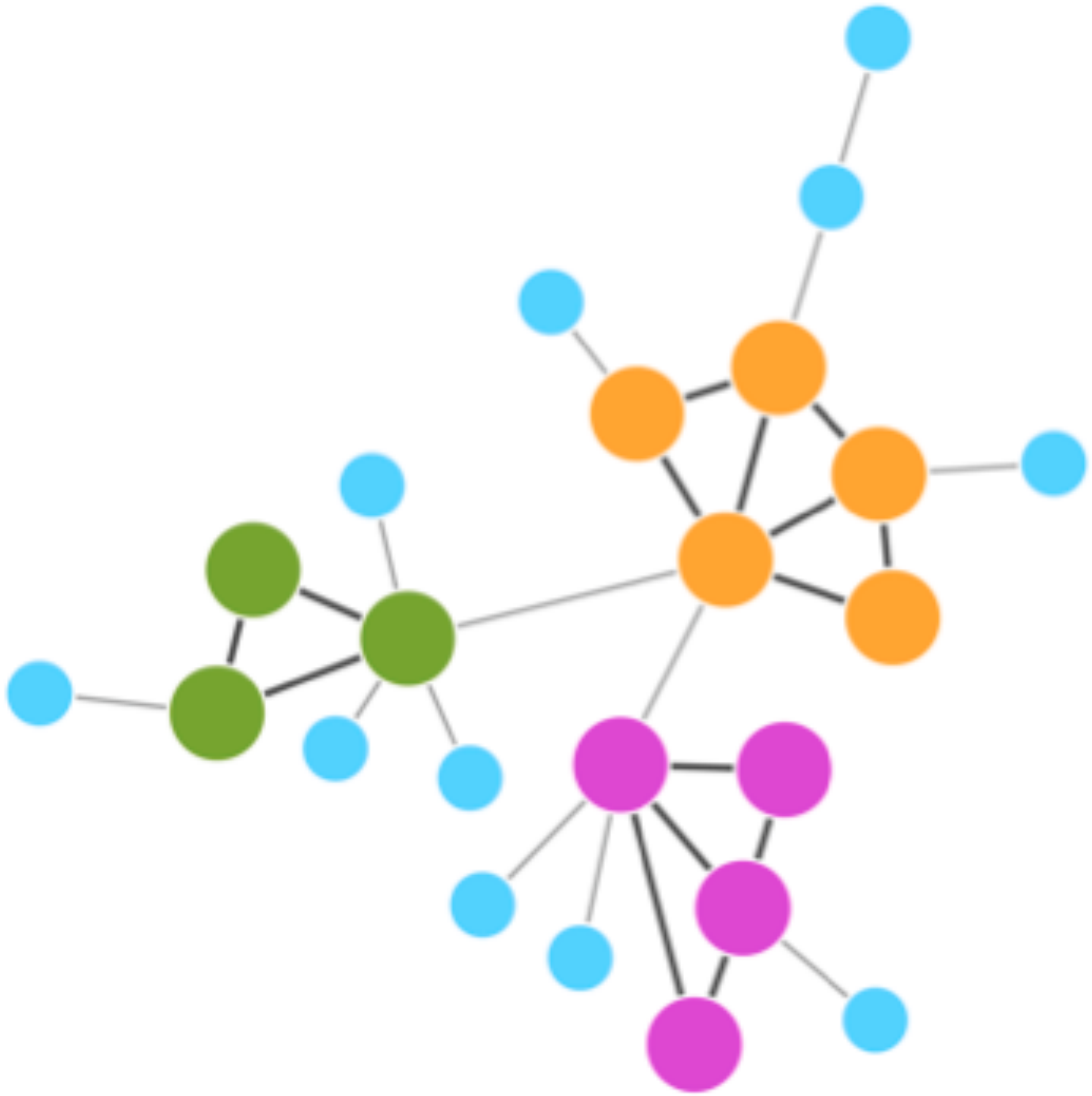
Network topology

Topological properties:

- Degree: number of edges that connect to a node.



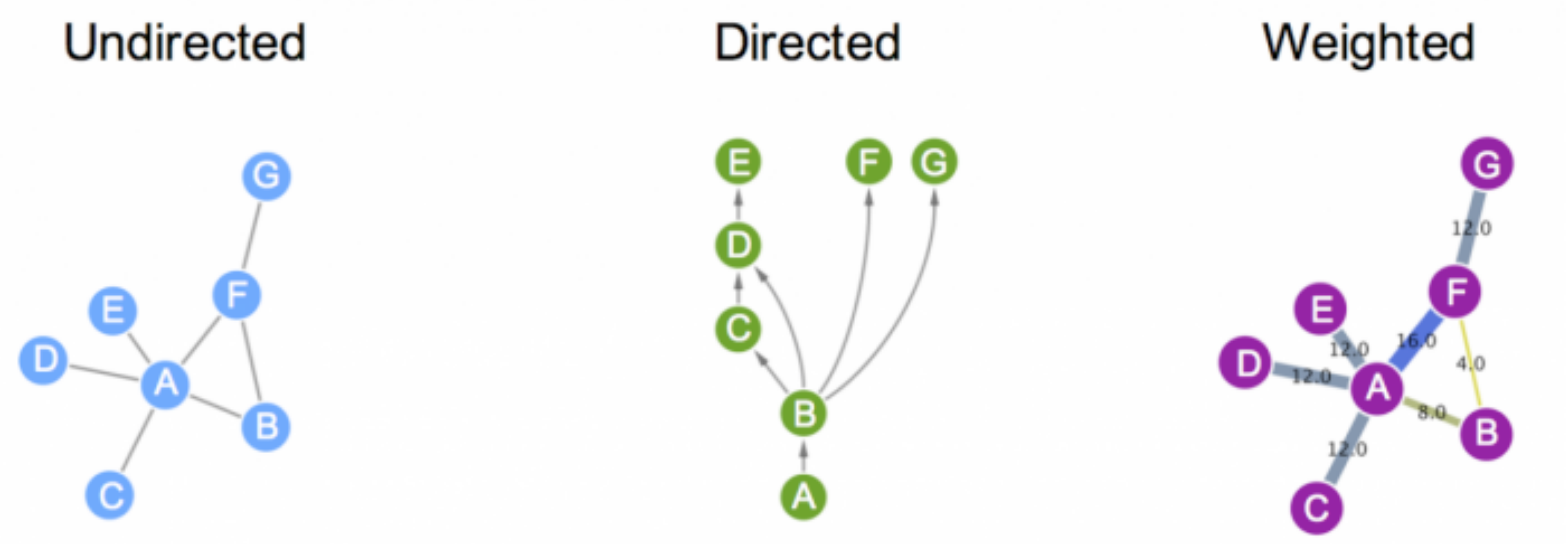
Network topology



Topological properties:

- Topological clusters: groups of nodes that are more internally connected than they are with the rest of the network.

Types of network edges

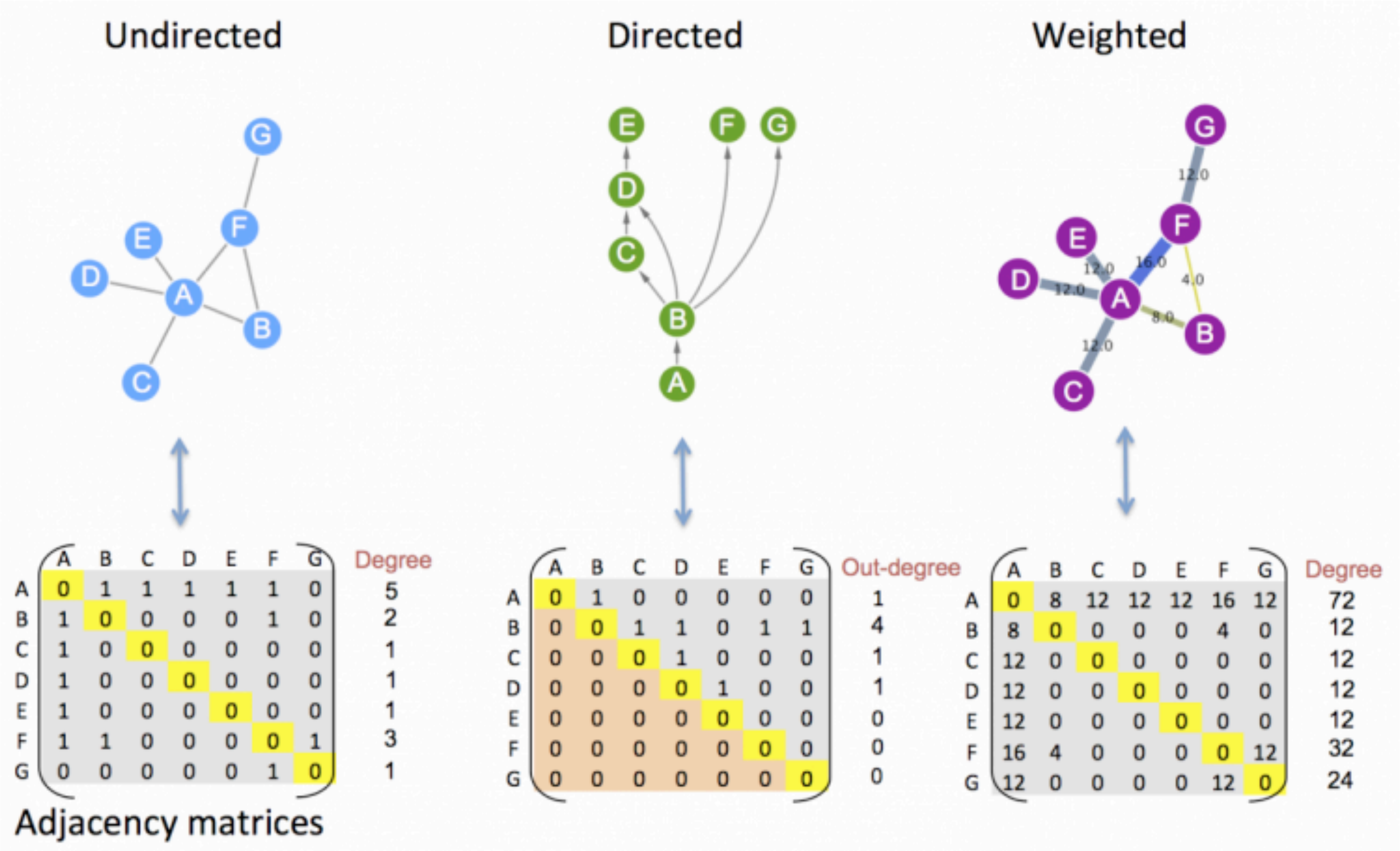


Undirected

Directed

Weighted

Adjacency matrices



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

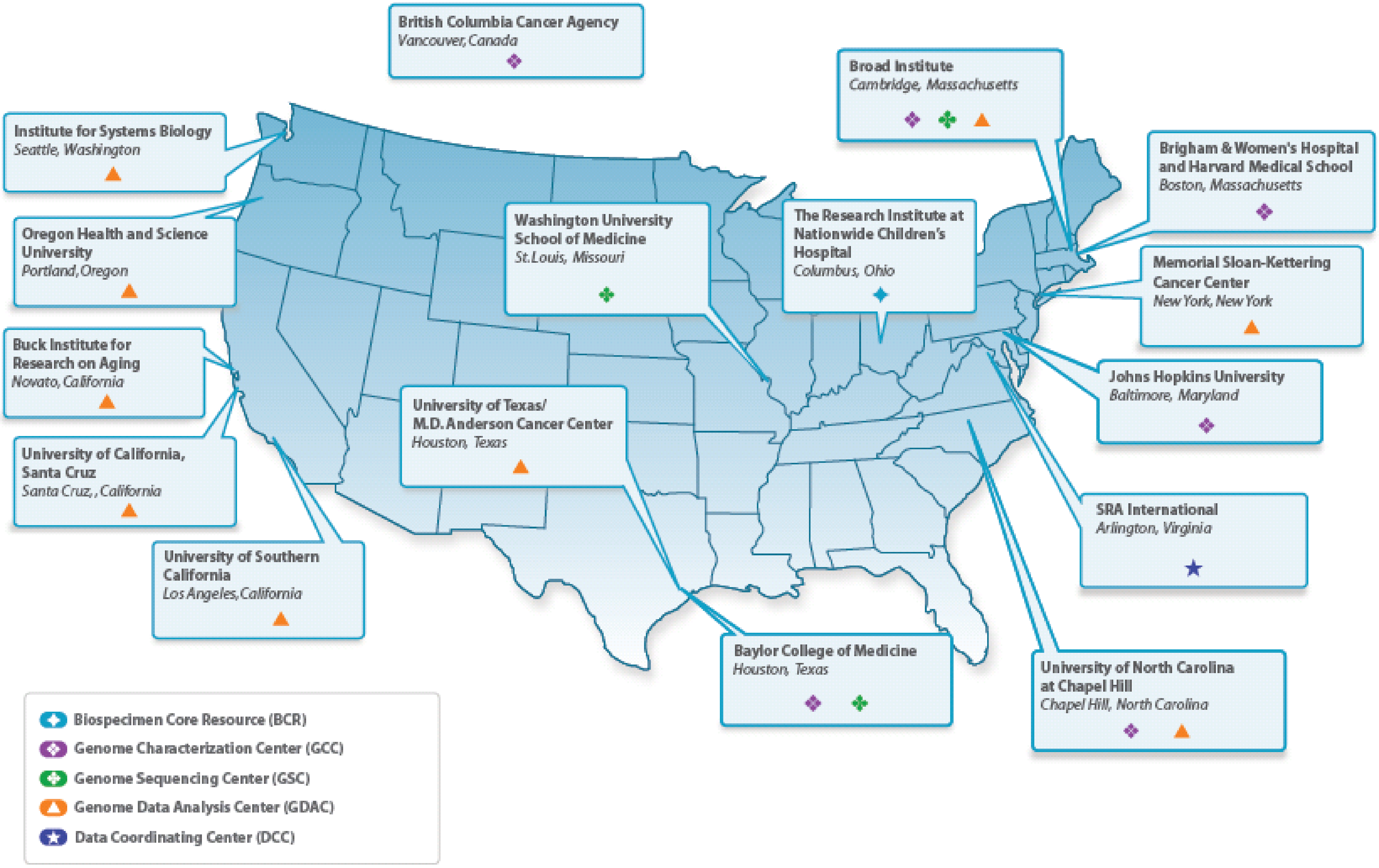
	A	B	C	D	E	F	G	Degree
A	0	8	12	12	12	16	12	72
B	8	0	0	0	0	4	0	12
C	12	0	0	0	0	0	0	12
D	12	0	0	0	0	0	0	12
E	12	0	0	0	0	0	0	12
F	16	4	0	0	0	0	12	32
G	12	0	0	0	0	12	0	24

Adjacency matrices

The Cancer Genome Atlas (TCGA) Research Network NCI/NHGRI

- Comprehensive genomic data to understand the molecular basis of cancer
- 33 cancer types
- Tumor and matched normal tissues from 11,000 patients
- Publicly available

The Cancer Genome Atlas Research Network



British Columbia Cancer Agency
Vancouver, Canada
❖

Broad Institute
Cambridge, Massachusetts
❖ ✤ ▲

Institute for Systems Biology
Seattle, Washington
▲

Brigham & Women's Hospital and Harvard Medical School
Boston, Massachusetts
❖

Oregon Health and Science University
Portland, Oregon
▲

Washington University School of Medicine
St. Louis, Missouri
✤

The Research Institute at Nationwide Children's Hospital
Columbus, Ohio
✦

Memorial Sloan-Kettering Cancer Center
New York, New York
▲

Buck Institute for Research on Aging
Novato, California
▲

Johns Hopkins University
Baltimore, Maryland
❖

University of Texas/ M.D. Anderson Cancer Center
Houston, Texas
▲

University of California, Santa Cruz
Santa Cruz, , California
▲

SRA International
Arlington, Virginia
★

University of Southern California
Los Angeles, California
▲

Baylor College of Medicine
Houston, Texas
❖ ✤

University of North Carolina at Chapel Hill
Chapel Hill, North Carolina
❖ ▲

✦ Biospecimen Core Resource (BCR)
❖ Genome Characterization Center (GCC)
✤ Genome Sequencing Center (GSC)
▲ Genome Data Analysis Center (GDAC)
★ Data Coordinating Center (DCC)

Courtesy of Dr. Boris Reva

Achievements of TCGA

- Large scale view on genomic landscapes in cancer
  - Major driver genes
  - Major altered cancer pathways
- Insights into complexity and inherent diversity of cancer
- Understanding necessity of personalized approach to treat cancer

Giant piles of TCGA data

“Clinical researchers use the Atlas to match patient molecular profiles to a specific tumor subtype.

Biotech companies use the Atlas to extract potential drug targets and new indications.

Progress in DNA sequencing, IT, and analytical technologies adds more detail to the Atlas, **and lower prices make genomic characterization available to more patients**.”

Linda Chin, TCGA pioneer

Targeted therapy for cancer

"**Targeted therapies** are drugs that interfere with a specific biochemical pathway that is central to the development, growth and spread of that particular cancer."

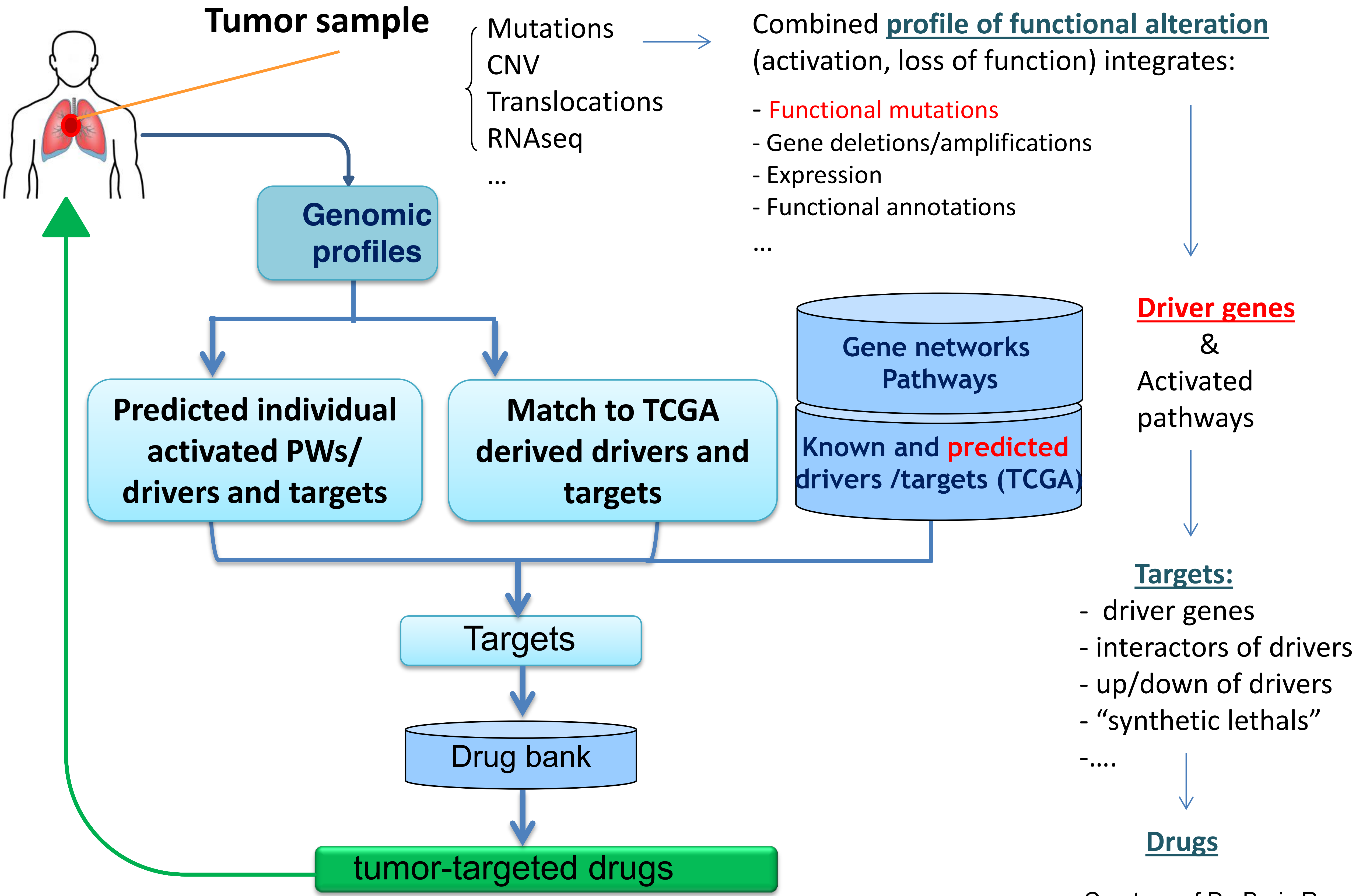
Then, a general **computational task** for targeted cancer therapy:

Given a molecular profile of a tumor and a set of molecular profiles of tumors with characterized driving alterations, determine

- activated cancer genes, gene modules and pathways
- targets for therapeutic interventions
- targets for known drug

Setting "targeted therapy " as a computational problem implies application of powerful approaches developed in data analysis, network modeling, pattern recognition and machine learning….

Targeted therapy as a “target finding” problem



Tumor sample

Mutations
CNV
Translocations
RNAseq
…

Combined **profile of functional alteration** (activation, loss of function) integrates:

- Functional mutations
- Gene deletions/amplifications
- Expression
- Functional annotations

…

Genomic profiles

Predicted individual activated PWs/ drivers and targets

Match to TCGA derived drivers and targets

Gene networks Pathways

Known and predicted drivers /targets (TCGA)

Driver genes
&
Activated pathways

Targets:

- driver genes
- interactors of drivers
- up/down of drivers
- “synthetic lethals”

-….

Drugs

Targets

Drug bank

tumor-targeted drugs

Courtesy of Dr. Boris Reva

Integrative Analysis

Motivation

- Integrating multiple molecular information leads to higher level discoveries of cell biology

Motivation

- Integrating multiple molecular information leads to higher level discoveries of cell biology
- Models that integrate omic data types rather than one model per type

Motivation

- Integrating multiple molecular information leads to higher level discoveries of cell biology
- Models that integrate omic data types rather than one model per type
- Model scope: N << P

Motivation

- Using only a subset of all possible interactions

Motivation

- Using only a subset of all possible interactions
- Data types cannot be modeled by Normal distributions

Sparse Conditional Gaussian Graphical Model (SCGGM), Zhang and Kim, 2014

- Conditional Gaussian graphical model
- Fits models through an l_1 penalized conditional log-likelihood

SCGGM

$(Y_i, X_i) \sim_{i.i.d.} \mathcal{N}(0, \Sigma), i = 1, \ldots, N$

joint covariance: $\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$

concentration matrix: $\Sigma^{-1} = \begin{pmatrix} \Theta_{xx} & \Theta_{xy} \\ \Theta_{yx} & \Theta_{yy} \end{pmatrix}$

The conditional p.d.f for the CGM is:

$$Y_i | X_i \sim N(-\Theta_{yy}^{-1}\Theta_{yx}X_i, \Theta_{yy}^{-1})$$

Like the GGM, the goal is to learn the zero (and non-zero) entries of Θ_{xy}, Θ_{yy}.

spaceMap

Conley et al. 2018

- Conditional graphical model
- Learns the conditional dependencies between two types of nodes through a penalized multivariate regression framework
- Cross-validation and model aggregation for reproducibility

spaceMap

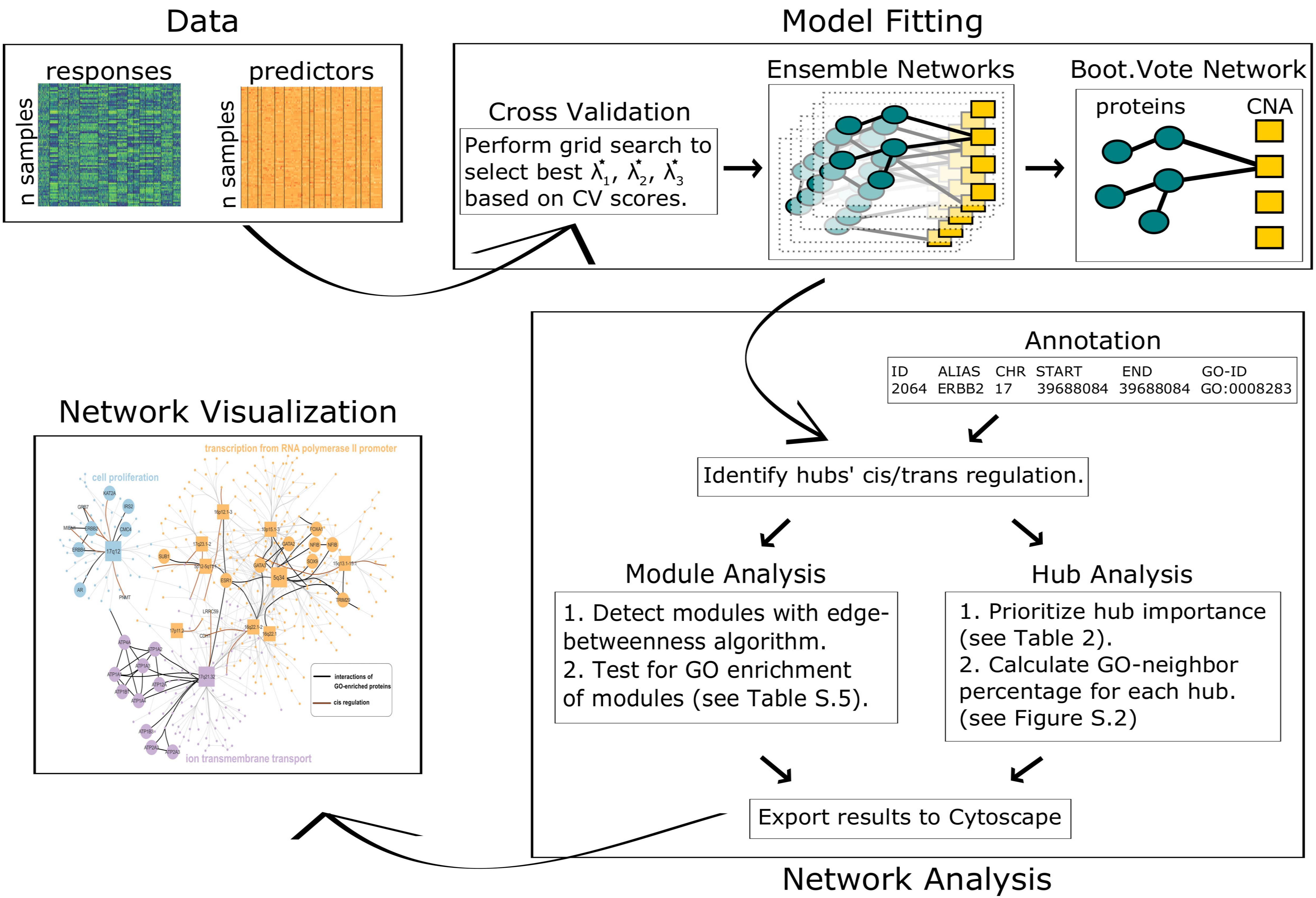
For $j = 1, \ldots, Q$: regress Y_j on predictors $\{Y_{-(j)}, X_l : l = 1, \ldots, P\}$

$$Y_j = \sum_{j \neq k} \rho^{jk} \sqrt{\sigma_{kk}/\sigma_{jj}} Y_k + \sum_{l=1}^{P} \gamma_{jl} X_l + \epsilon_j$$

Minimize penalized least-squares criterion:

$$L_{N,\lambda}(\theta, \sigma, \Gamma) = \frac{1}{2} \sum_{j=1}^{Q} \Big(Y_j - \sum_{j \neq k} \rho^{jk} \sqrt{\sigma_{kk}/\sigma_{jj}} Y_k - \sum_{l=1}^{P} \gamma_{jl} X_l \Big)^2 + \lambda_1 \sum_{1 \leq j < k \leq Q} |\rho^{jk}| + \lambda_2 \sum_{l=1}^{P} ||\Gamma_l||_1 + \lambda_3 \sum_{l=1}^{P} ||\Gamma_l||_2$$

The β_{jk}'s & γ_{jl}'s are proportional to the partial correlations $\mathrm{Cor}(y_j, y_k | y_{-(j,k)}, x)$ & $\mathrm{Cor}(y_j, x_l | y_{-(j)}, x_{-(l)}\ )$, respectively.



Data

Model Fitting

Cross Validation

Perform grid search to select best $\lambda_1^\star$, $\lambda_2^\star$, $\lambda_3^\star$ based on CV scores.

Ensemble Networks

Boot.Vote Network

Network Analysis

Annotation

ID	ALIAS	CHR	START	END	GO-ID
2064	ERBB2	17	39688084	39688084	GO:0008283

Identify hubs' cis/trans regulation.

Module Analysis

1. Detect modules with edge-betweenness algorithm.
2. Test for GO enrichment of modules (see Table S.5).

Hub Analysis

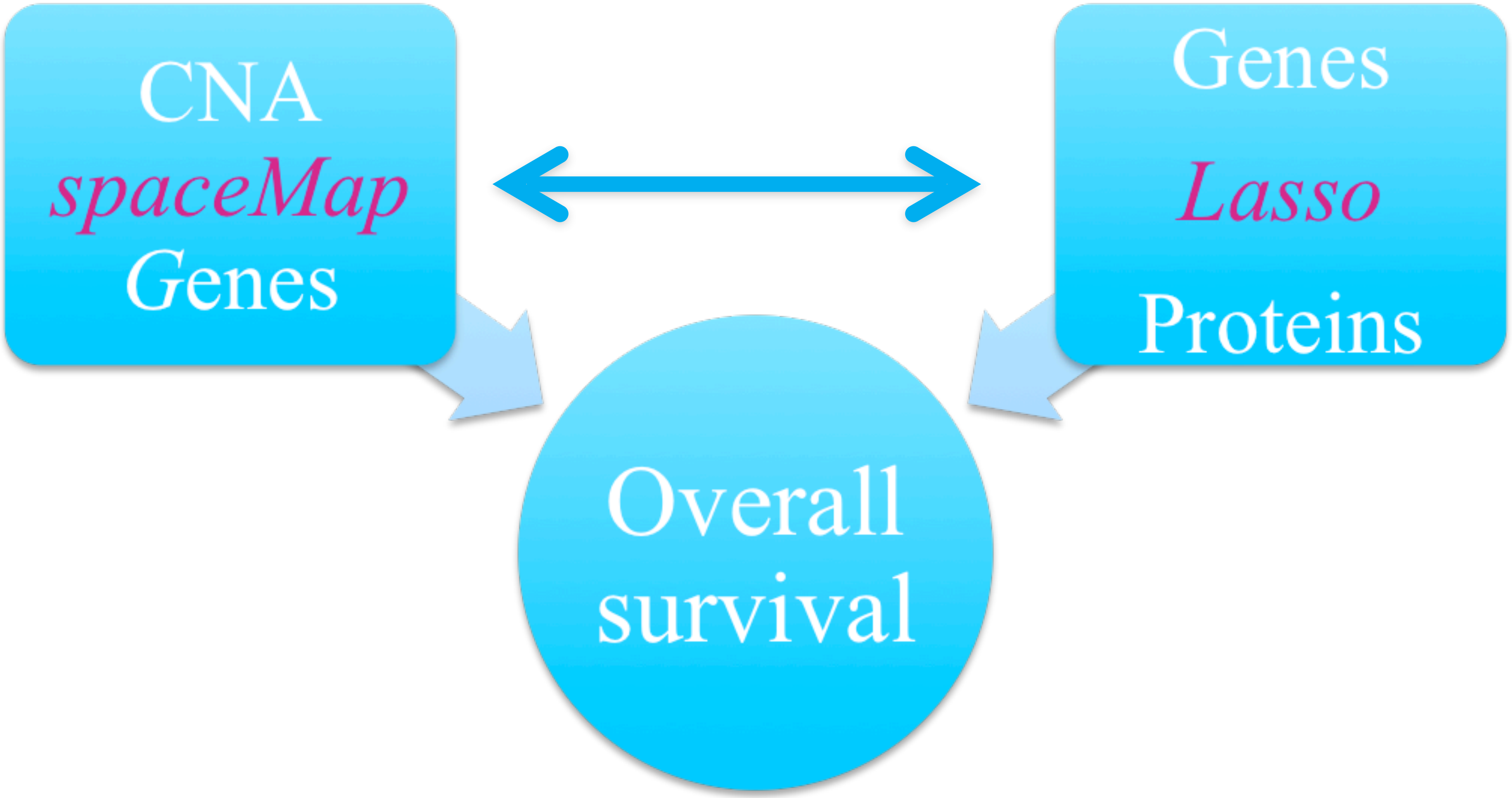
1. Prioritize hub importance (see Table 2).
2. Calculate GO-neighbor percentage for each hub. (see Figure S.2)

Export results to Cytoscape

Network Visualization

Hypothesis

Genes that are driven by CNAs would affect the protein activities and should be more informative for ovarian cancer survival.

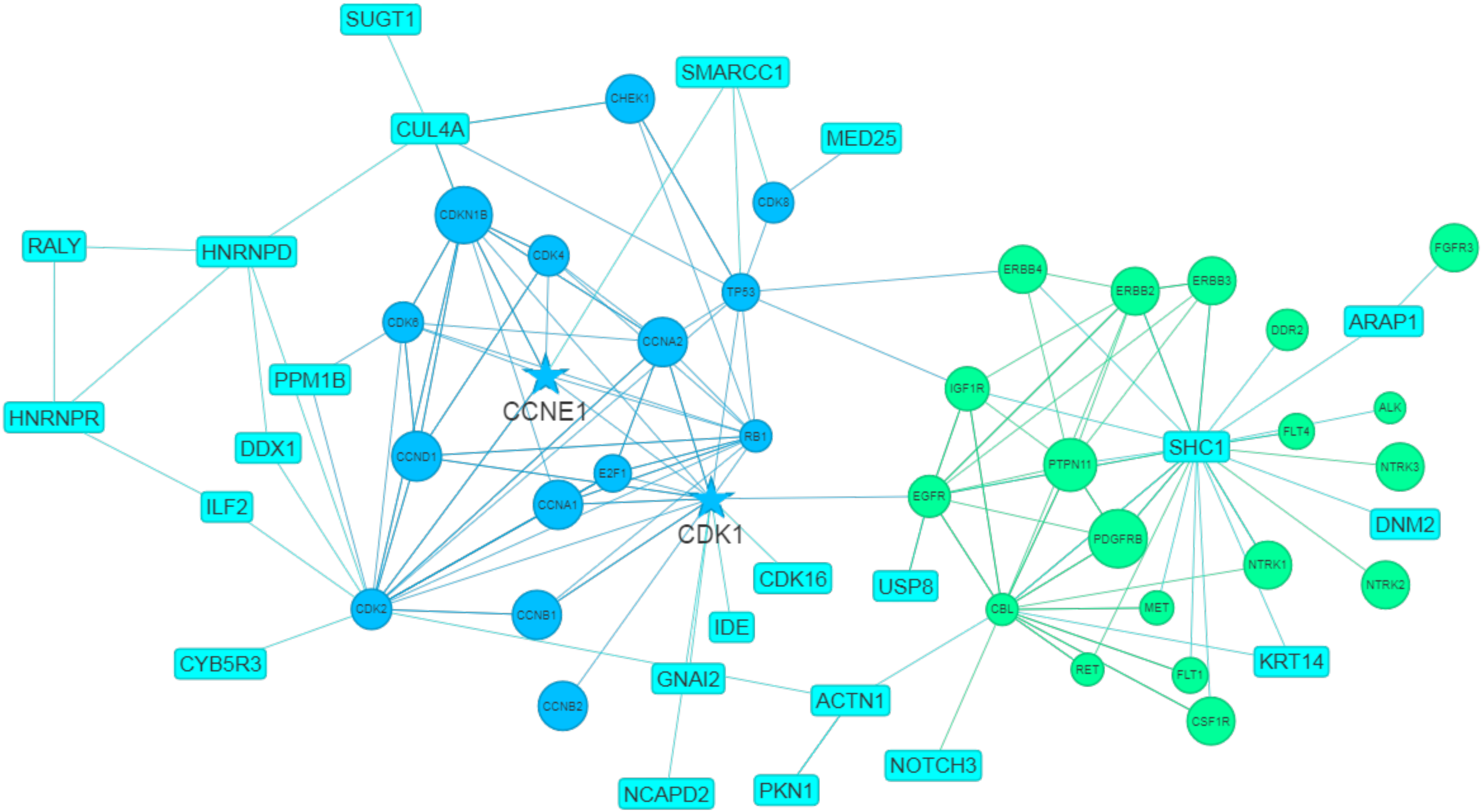


CNA *spaceMap* Genes

Genes *Lasso* Proteins

Overall survival

Cell-cycle and RTK pathways

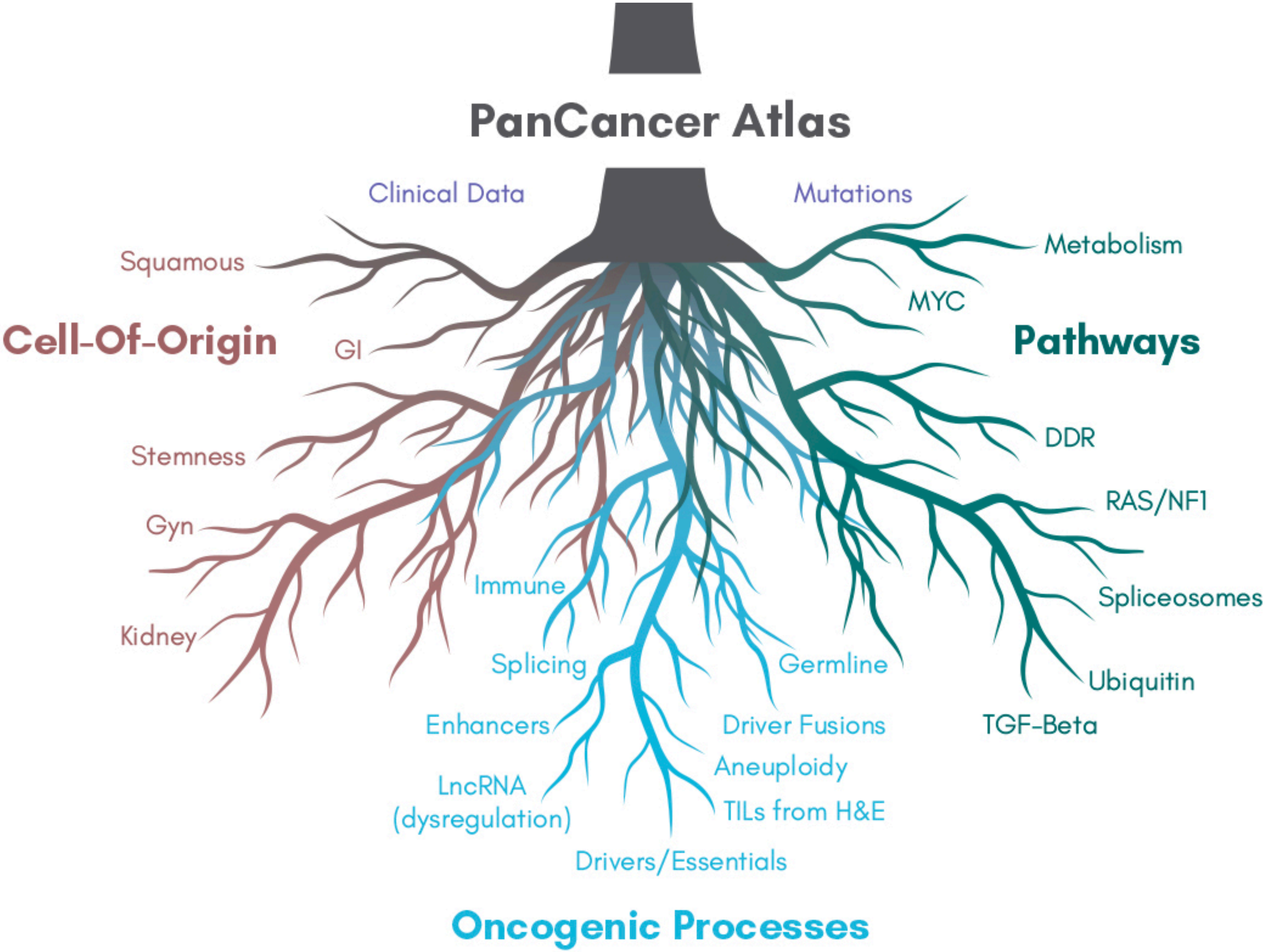


The Pan-Cancer Atlas

- Extension of TCGA
- Completed in 2018
- Alterations across different tumor types

https://www.cell.com/pb-assets/consortium/PanCancerAtlas/PanCani3/index.html

PanCancer Atlas



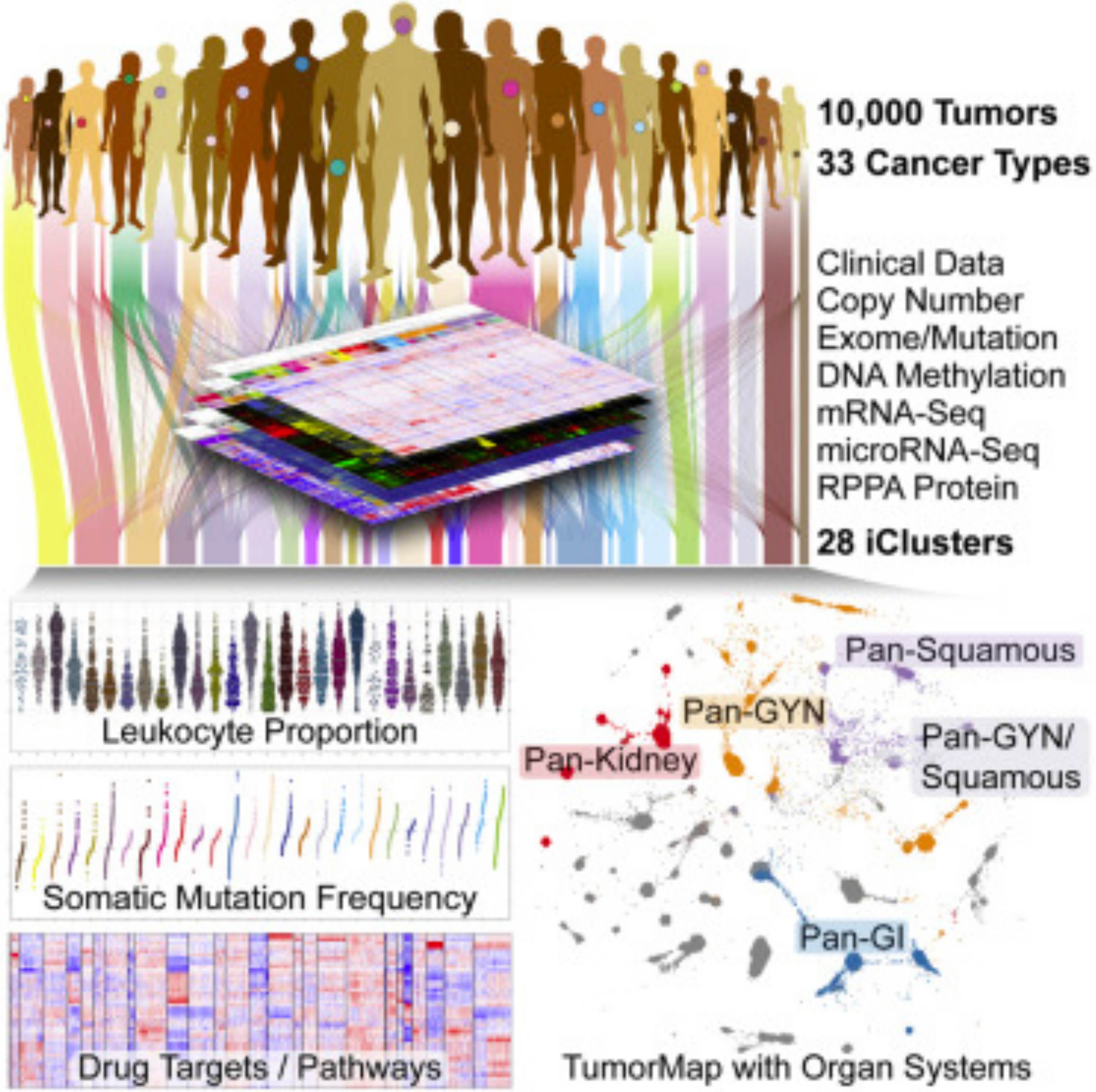
Cell of Origin Patterns

Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer

- Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation
- A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers
- Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas

http://www.cell.com/cell/fulltext/S0092-8674(18)30302-7?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC

Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer (2018)



http://www.cell.com/cell/fulltext/S0092-8674(18)30302-7?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC

Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer (2018)

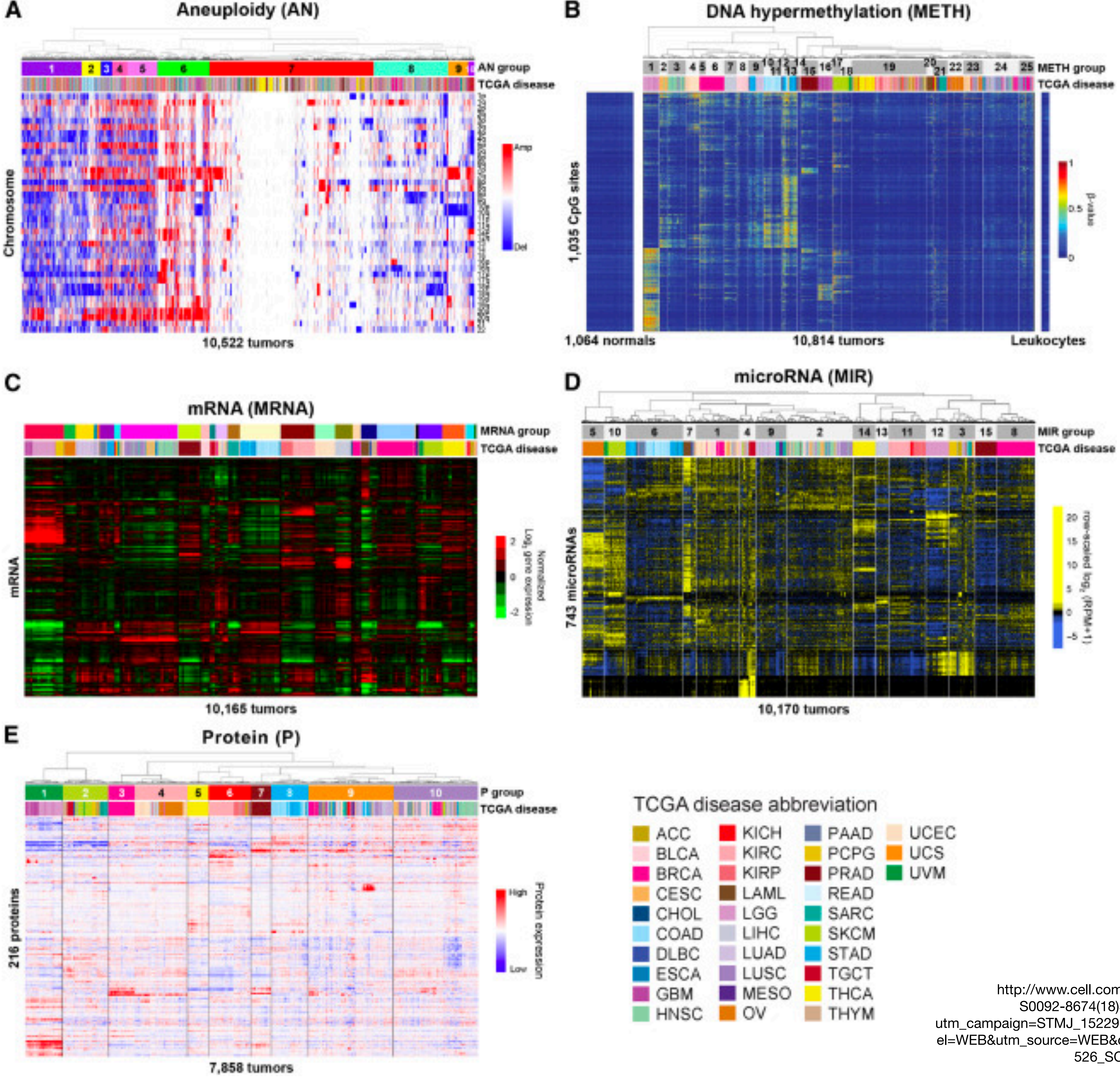
iCluster:

- Joint latent variable model for integrative clustering.
- Incorporates flexible modeling of the associations between different data types and the variance-covariance structure within data types in a single framework, while simultaneously reducing the dimensionality of the datasets.
- Likelihood-based inference is obtained through the Expectation-Maximization algorithm.

Cluster-of-cluster assignments (COCA):

- Takes as input the binary vectors that represent each of the platform-specific cluster groups.
- Reclusters the samples according to those vectors.
- Data across platforms are combined without the need for normalization.
- Each platform influences the final integrated result with weight proportional to the number of distinct subtypes reproducibly found by consensus clustering.

http://www.cell.com/cell/fulltext/S0092-8674(18)30302-7?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC



http://www.cell.com/cell/fulltext/S0092-8674(18)30302-7?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC

Conclusions

- Identified 10 to 25 platform-specific molecular subsets within ~10,000 tumors, each showing significant compositional heterogeneity based on classical tumor taxonomy.

- These iCluster assignments have potential clinical utility, and their multi-platform basis suggests that this new subclassification system might further improve the management of the 1%–3% of all cancer patients newly diagnosed with cancer of unknown primary (CUP).

- Interrogation of individual iClusters for their differentiating PARADIGM pathway features, canonical pathways, and gene programs amenable to drug targeting identified strong immune-related signaling features, suggesting that they may share potential susceptibility to immunotherapy.

- Integrated molecular tumor profiling may improve basket-trial design by considering both mutations and oncogenic signaling pathways along with consideration of each tumor’s tissue-specific or cell-of-origin context.

http://www.cell.com/cell/fulltext/S0092-8674(18)30302-7?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC

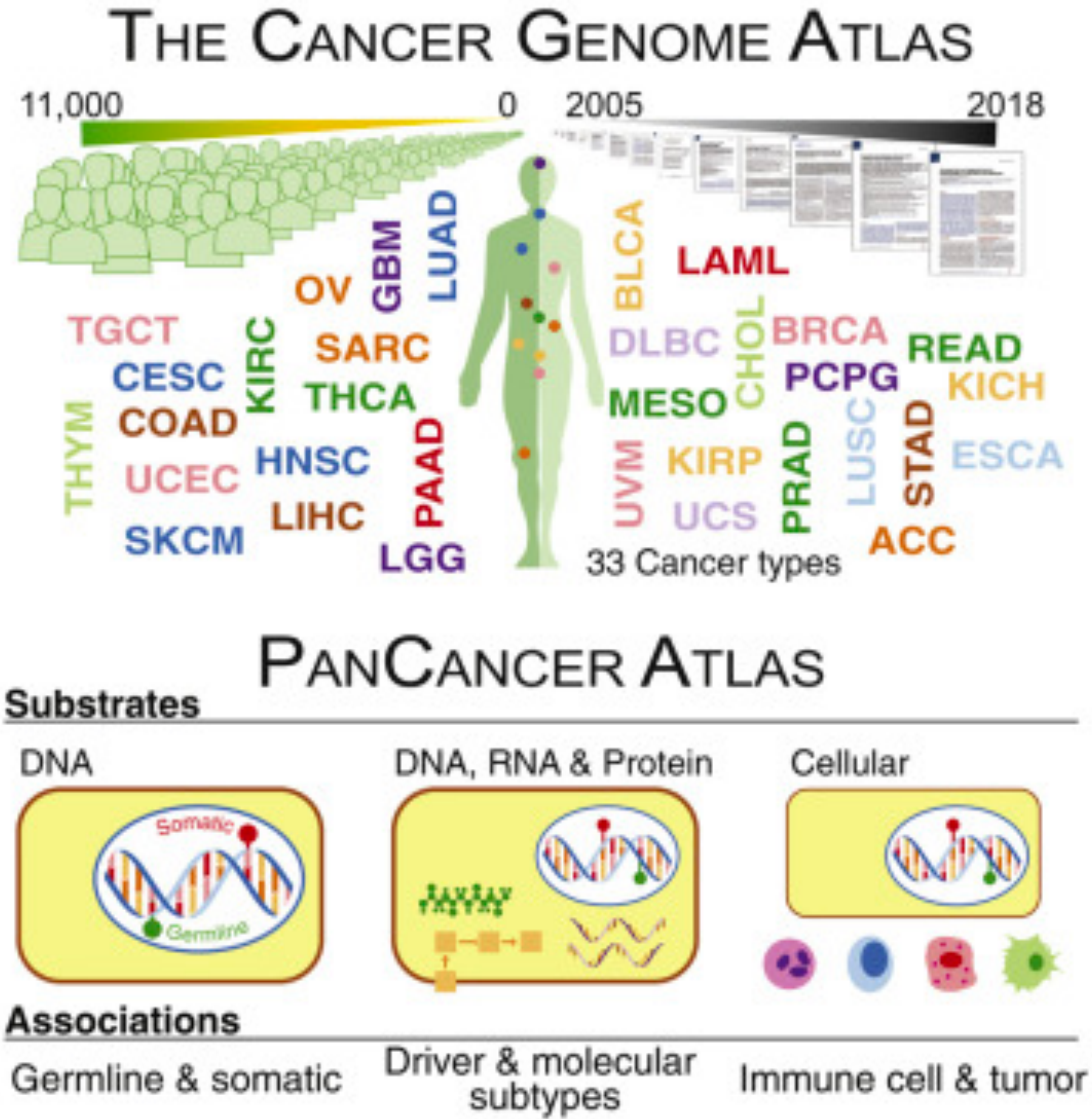
Oncogenic Processes

Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics

- Pathogenic Germline Variants in 10,389 Adult Cancers
- Comprehensive Characterization of Cancer Driver Genes and Mutations
- Driver Fusions and Their Implications in the Development and Treatment of Human Cancers

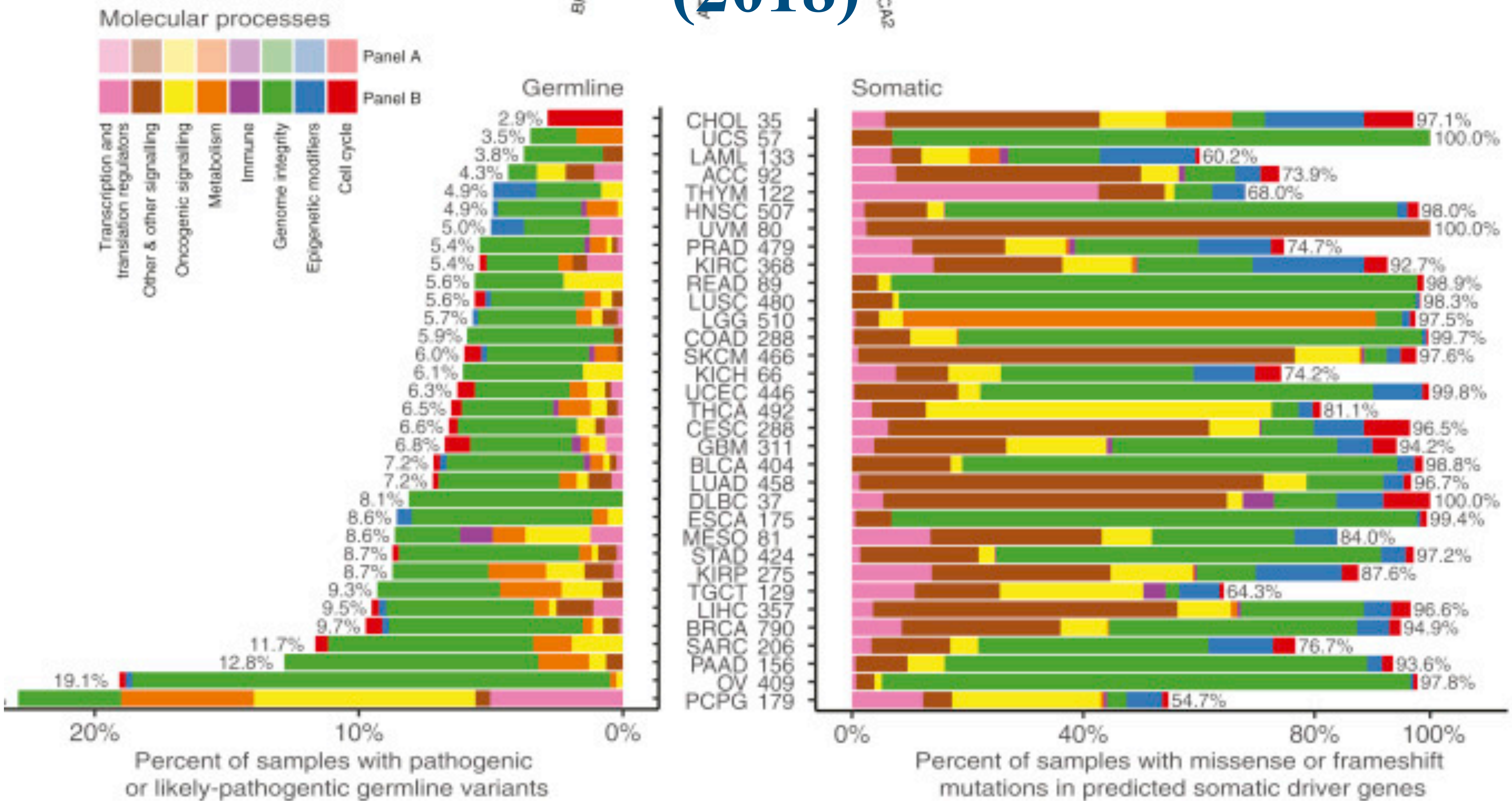
http://www.cell.com/cell/fulltext/S0092-8674(18)30313-1?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC

Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics (2018)



http://www.cell.com/cell/fulltext/S0092-8674(18)30313-1?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC

Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics (2018)



http://www.cell.com/cell/fulltext/S0092-8674(18)30313-1?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC

Conclusions

- The germline genome has far-ranging, pathway-dependent influences on the somatic landscape, often promoting somatic mutations.

- Interactions between driver genes and the transcriptome are context dependent, as is the impact of driver mutations in both *cis*- and *trans*-expression.

- Some oncogenic processes that tend to be deregulated in few cancer types are more related to specific genes rather than to prominent drivers.

- Findings suggest drastic changes in clinical practice and drug development:
  - Molecular treatments will increasingly be developed with multi-omics.
  - Bioinformatic systems will help efficiently design optimized treatment plans lurking within large combinatorial spaces with respect to dosage, efficacy, side effects, etc.

Could some somatic mutations be tolerated in normal development?
How does this impact our understanding of oncogenic mutations?
Can we find the alterations that drive the process from primary tumor to metastases?

http://www.cell.com/cell/fulltext/S0092-8674(18)30313-1?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC

Signaling Pathways

Oncogenic Signaling Pathways in The Cancer Genome Atlas

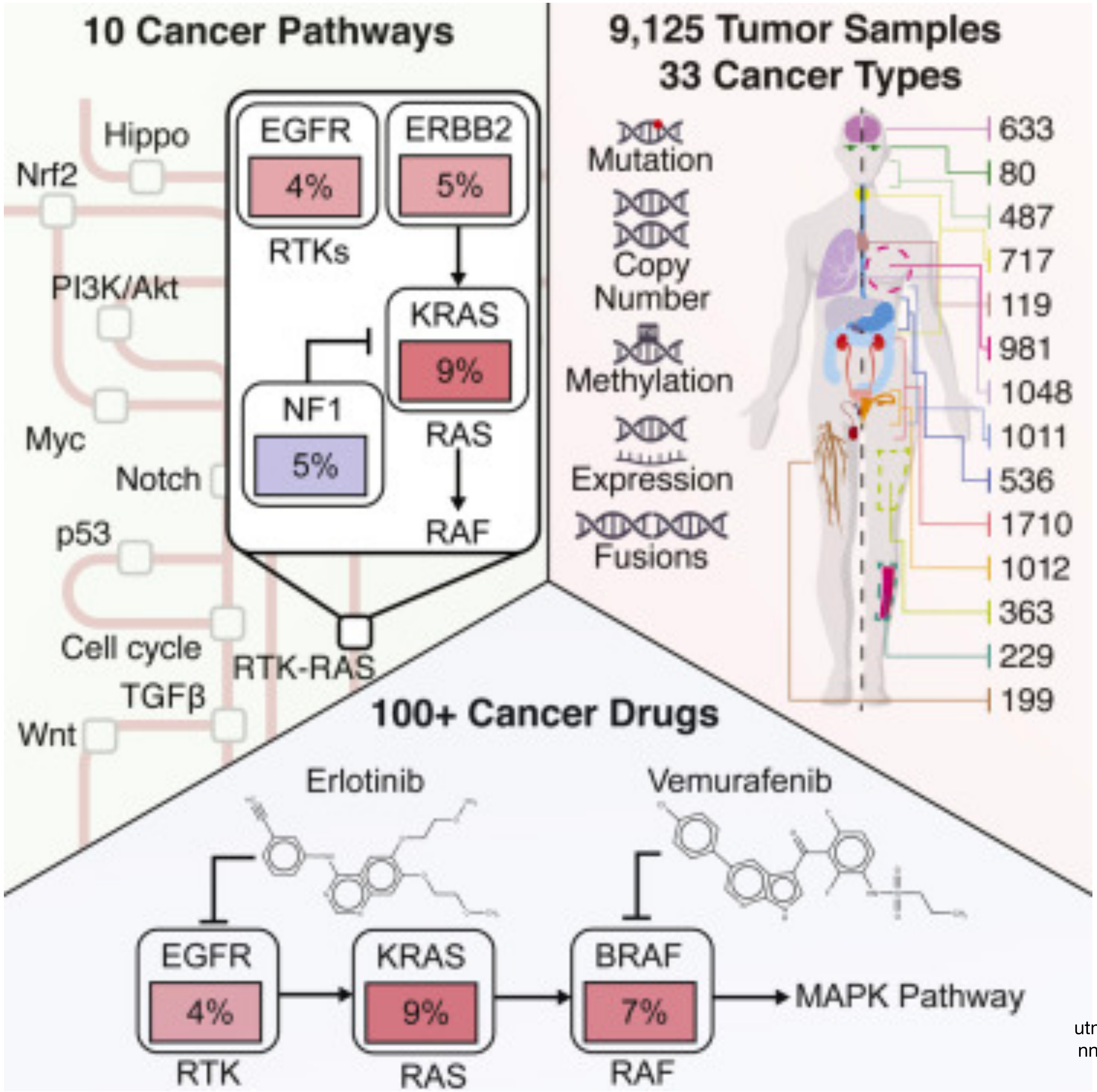
- Pan-cancer Alterations of the MYC Oncogene and Its Proximal Network across the Cancer Genome Atlas

- Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas

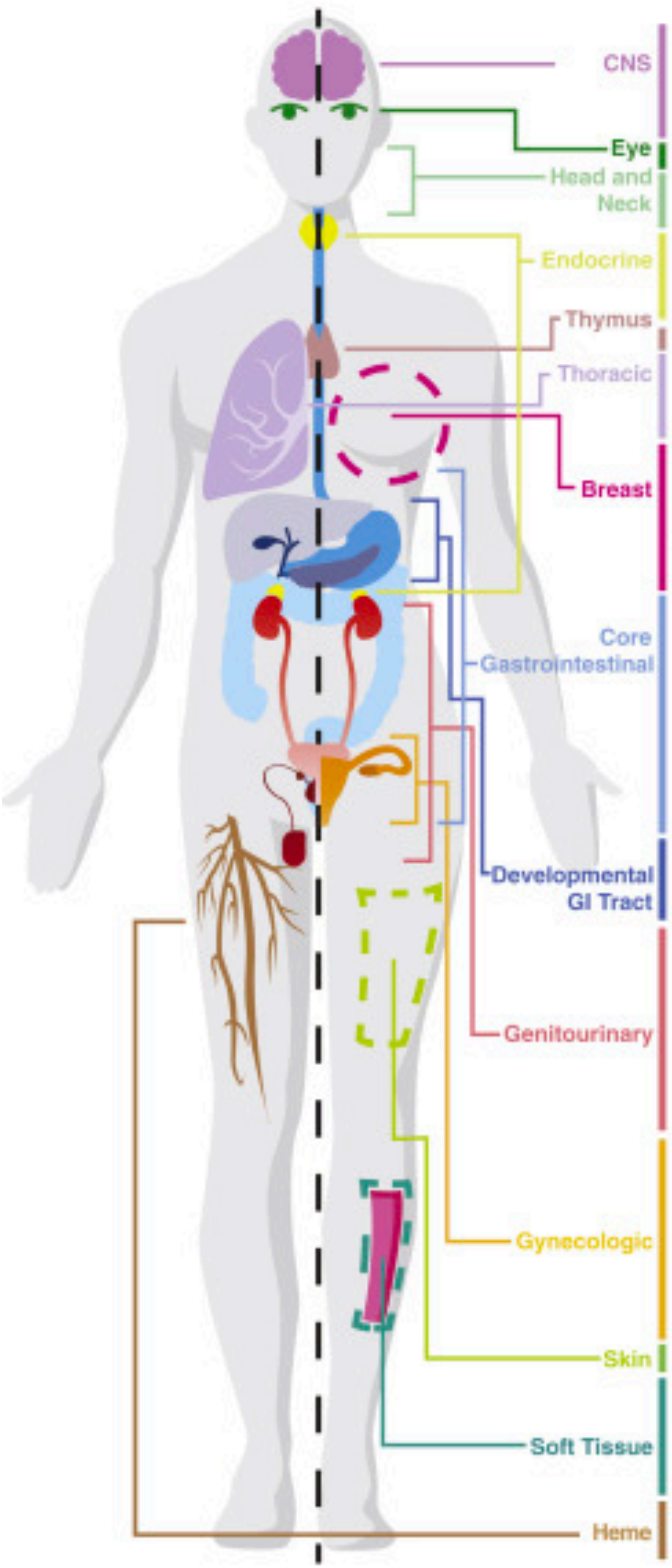
- Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas

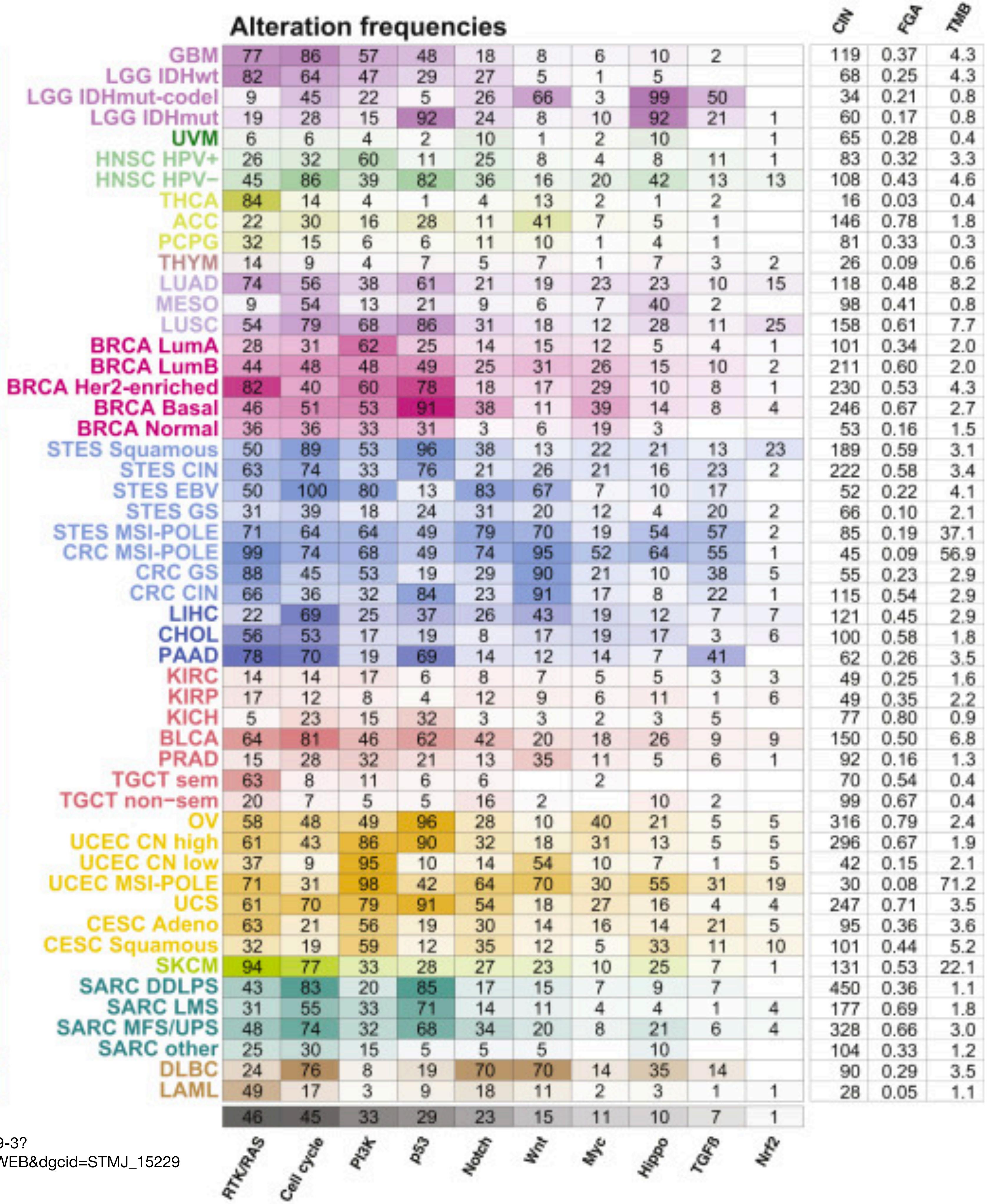
http://www.cell.com/cell/fulltext/S0092-8674(18)30359-3?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC

Oncogenic Signaling Pathways in The Cancer Genome Atlas (2018)



http://www.cell.com/cell/fulltext/S0092-8674(18)30359-3?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC





Alteration frequencies

	RTK/RAS	Cell cycle	PI3K	p53	Notch	Wnt	Myc	Hippo	TGFβ	Nrf2	CIN	FGA	TMB
GBM	77	86	57	48	18	8	6	10	2		119	0.37	4.3
LGG IDHwt	82	64	47	29	27	5	1	5			68	0.25	4.3
LGG IDHmut-codel	9	45	22	5	26	66	3	99	50		34	0.21	0.8
LGG IDHmut	19	28	15	92	24	8	10	92	21	1	60	0.17	0.8
UVM	6	6	4	2	10	1	2	10		1	65	0.28	0.4
HNSC HPV+	26	32	60	11	25	8	4	8	11	1	83	0.32	3.3
HNSC HPV−	45	86	39	82	36	16	20	42	13	13	108	0.43	4.6
THCA	84	14	4	1	4	13	2	1	2		16	0.03	0.4
ACC	22	30	16	28	11	41	7	5	1		146	0.78	1.8
PCPG	32	15	6	6	11	10	1	4	1		81	0.33	0.3
THYM	14	9	4	7	5	7	1	7	3	2	26	0.09	0.6
LUAD	74	56	38	61	21	19	23	23	10	15	118	0.48	8.2
MESO	9	54	13	21	9	6	7	40	2		98	0.41	0.8
LUSC	54	79	68	86	31	18	12	28	11	25	158	0.61	7.7
BRCA LumA	28	31	62	25	14	15	12	5	4	1	101	0.34	2.0
BRCA LumB	44	48	48	49	25	31	26	15	10	2	211	0.60	2.0
BRCA Her2-enriched	82	40	60	78	18	17	29	10	8	1	230	0.53	4.3
BRCA Basal	46	51	53	91	38	11	39	14	8	4	246	0.67	2.7
BRCA Normal	36	36	33	31	3	6	19	3			53	0.16	1.5
STES Squamous	50	89	53	96	38	13	22	21	13	23	189	0.59	3.1
STES CIN	63	74	33	76	21	26	21	16	23	2	222	0.58	3.4
STES EBV	50	100	80	13	83	67	7	10	17		52	0.22	4.1
STES GS	31	39	18	24	31	20	12	4	20	2	66	0.10	2.1
STES MSI-POLE	71	64	64	49	79	70	19	54	57	2	85	0.19	37.1
CRC MSI-POLE	99	74	68	49	74	95	52	64	55	1	45	0.09	56.9
CRC GS	88	45	53	19	29	90	21	10	38	5	55	0.23	2.9
CRC CIN	66	36	32	84	23	91	17	8	22	1	115	0.54	2.9
LIHC	22	69	25	37	26	43	19	12	7	7	121	0.45	2.9
CHOL	56	53	17	19	8	17	19	17	3	6	100	0.58	1.8
PAAD	78	70	19	69	14	12	14	7	41		62	0.26	3.5
KIRC	14	14	17	6	8	7	5	5	3	3	49	0.25	1.6
KIRP	17	12	8	4	12	9	6	11	1	6	49	0.35	2.2
KICH	5	23	15	32	3	3	2	3	5		77	0.80	0.9
BLCA	64	81	46	62	42	20	18	26	9	9	150	0.50	6.8
PRAD	15	28	32	21	13	35	11	5	6	1	92	0.16	1.3
TGCT sem	63	8	11	6	6		2				70	0.54	0.4
TGCT non−sem	20	7	5	5	16	2		10	2		99	0.67	0.4
OV	58	48	49	96	28	10	40	21	5	5	316	0.79	2.4
UCEC CN high	61	43	86	90	32	18	31	13	5	5	296	0.67	1.9
UCEC CN low	37	9	95	10	14	54	10	7	1	5	42	0.15	2.1
UCEC MSI-POLE	71	31	98	42	64	70	30	55	31	19	30	0.08	71.2
UCS	61	70	79	91	54	18	27	16	4	4	247	0.71	3.5
CESC Adeno	63	21	56	19	30	14	16	14	21	5	95	0.36	3.6
CESC Squamous	32	19	59	12	35	12	5	33	11	10	101	0.44	5.2
SKCM	94	77	33	28	27	23	10	25	7	1	131	0.53	22.1
SARC DDLPS	43	83	20	85	17	15	7	9	7		450	0.36	1.1
SARC LMS	31	55	33	71	14	11	4	4	1	4	177	0.69	1.8
SARC MFS/UPS	48	74	32	68	34	20	8	21	6	4	328	0.66	3.0
SARC other	25	30	15	5	5	5		10			104	0.33	1.2
DLBC	24	76	8	19	70	70	14	35	14		90	0.29	3.5
LAML	49	17	3	9	18	11	2	3	1	1	28	0.05	1.1
	46	45	33	29	23	15	11	10	7	1			

http://www.cell.com/cell/fulltext/S0092-8674(18)30359-3?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC

Conclusions

- Signaling pathways are somatically altered in cancer at varying frequencies and in varying combinations across different organs and tissues.

- There is a complex interplay of co-occurring and mutually exclusive alterations within and across pathways.

- Standardized set of pathway templates, curated through a combination of computational methods and expert review are reported and publicly available (http://pathwaymapper.org/).

- Most hematologic cancers are not included.

http://www.cell.com/cell/fulltext/S0092-8674(18)30359-3?utm_campaign=STMJ_1522958526_SC&utm_channel=WEB&utm_source=WEB&dgcid=STMJ_1522958526_SC

OK, great but …

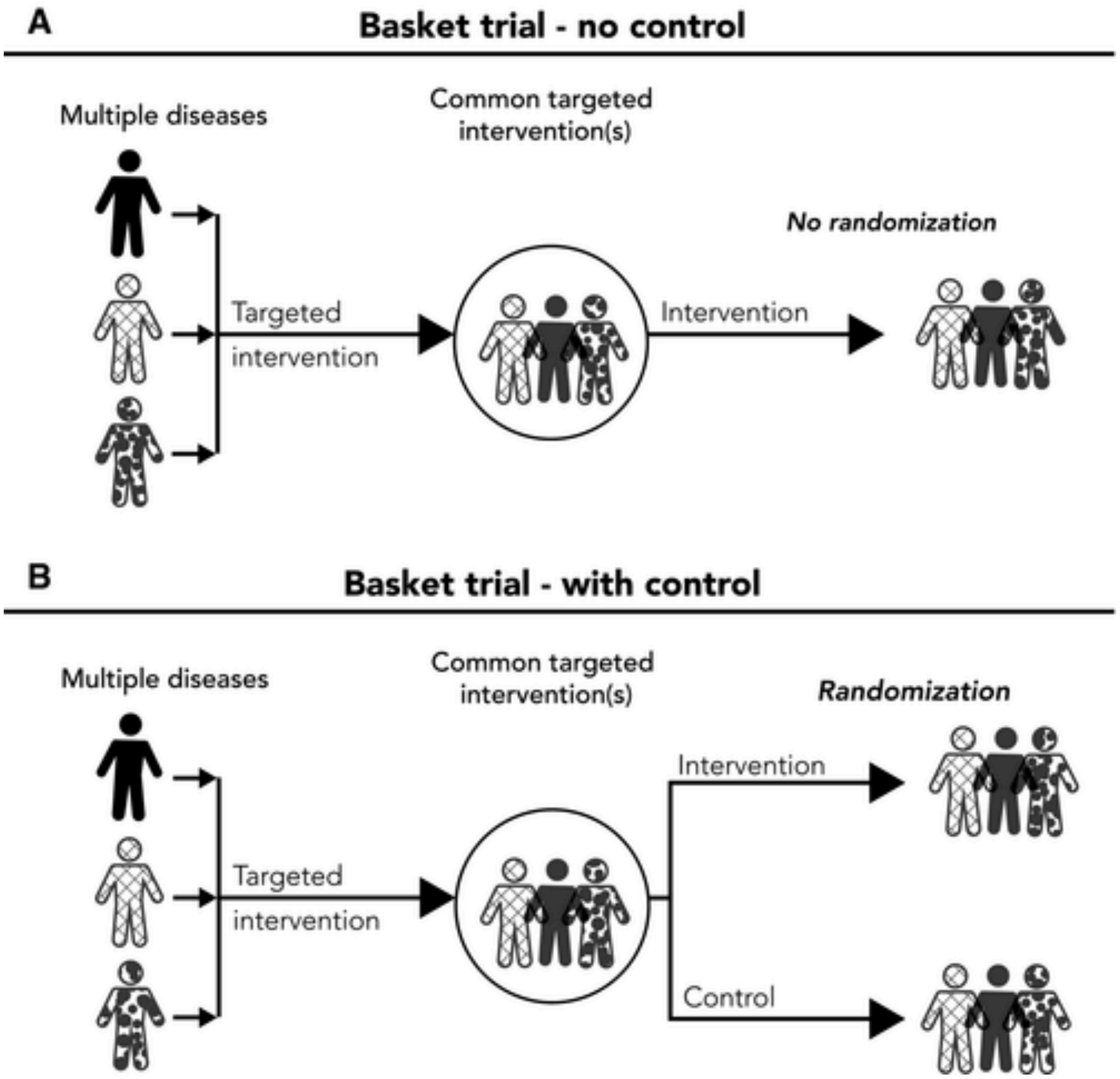
- How can we improve medicine with genetics and genomics research?
- Is it possible to implement the findings?

Basket Trials

- Basket trials are prospective clinical trials that test one or more targeted interventions across multiple types of diseases.
- There are unifying eligibility criteria usually based on a patient's predictive risk factor.

https://acsjournals.onlinelibrary.wiley.com/doi/epdf/10.3322/caac.21600

Basket Trials



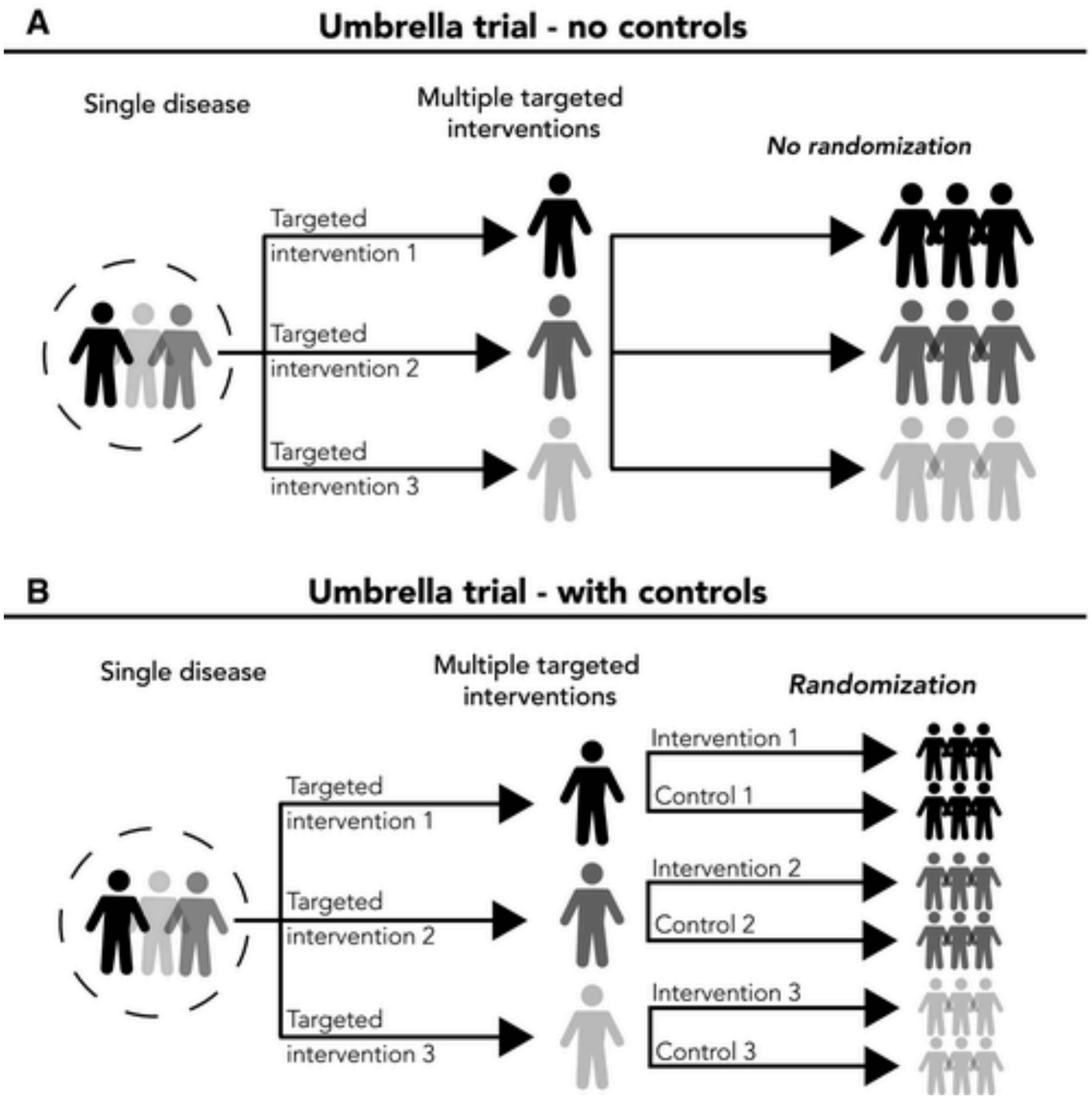
https://acsjournals.onlinelibrary.wiley.com/doi/epdf/10.3322/caac.21600

Umbrella Trials

- Umbrella trials are prospective clinical trials that test multiple targeted interventions for a single disease based on predictive biomarkers or other predictive patient risk factors.
- In umbrella trials, a single disease (eg, advanced breast cancer) is stratified into multiple subgroups, with eligibility for each intervention arm defined by the intervention's mechanism of action.

https://acsjournals.onlinelibrary.wiley.com/doi/epdf/10.3322/caac.21600

Umbrella Trials



https://acsjournals.onlinelibrary.wiley.com/doi/epdf/10.3322/caac.21600

Basket and Umbrella Trials

KEY CHARACTERISTICS	BASKET TRIALS	UMBRELLA TRIALS
Eligibility criteria	• Patients enrolled in a basket trial have multiple diseases with common unifying risk factor(s)	• Patients in an umbrella trial usually have the same disease
Patient subgroups	• Patient subgroups may be defined based on disease subtypes	• Risk factors are used to stratify patients into multiple subgroups (*patient stratification*)
Intervention assignment	• It is common for basket trials to have a single intervention that is targeted based on the unifying risk factor	• Umbrella trials have multiple interventions, with intervention assignment being determined based on their risk factor
	• Intervention assignment may or may not be determined using randomization	• Similar to basket trials, intervention assignment may or may not be determined using randomization
Choice in a control group	• Determining the choice in the control group can be difficult because there are multiple diseases being studied	• Compared with basket trials, it may be easier to pick the choice in the control group for umbrella trials because there is one disease being studied
	• If there are different established standards of care between multiple diseases being studied, a common control group may not be feasible	• The existing standard of care (or placebo, if there is no established care) for the disease being studied may be used as the control for all of the subgroups

https://acsjournals.onlinelibrary.wiley.com/doi/epdf/10.3322/caac.21600

THANK YOU!