Laurea Magistrale interfacoltà: LM-9, -18

Systems Biology

Biomedical applications

Prof. Gabriela Constantin M.D., Ph.D.

Dipartimento di Medicina
Sezione di Patologia Generale

“Laboratory of Neuroimmunology and Neuroinflammation”

Tel.: 045-8027102
Email: gabriela.constantin@univr.it

“Systems Biology” and biomedical applications

1) Diseases– introduction
2) Networks in biomedicine – introduction
3) Application: The human diseasome
4) Application: Comorbidity
5) Innate immunity: introduction and applications
6) Inflammation: introduction and applications
7) Tumors: introduction and applications
8) P4 medicine

DISEASE

DEFINITION:

ALTERATION (REDUCTION, INCREASE, LACK) OF CELLULAR FUNCTION OF CELLS/TISSUES/ORGANS

ALTERATION OF HOMEOSTATIC EQUILIBRIUM

WHO: "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity"

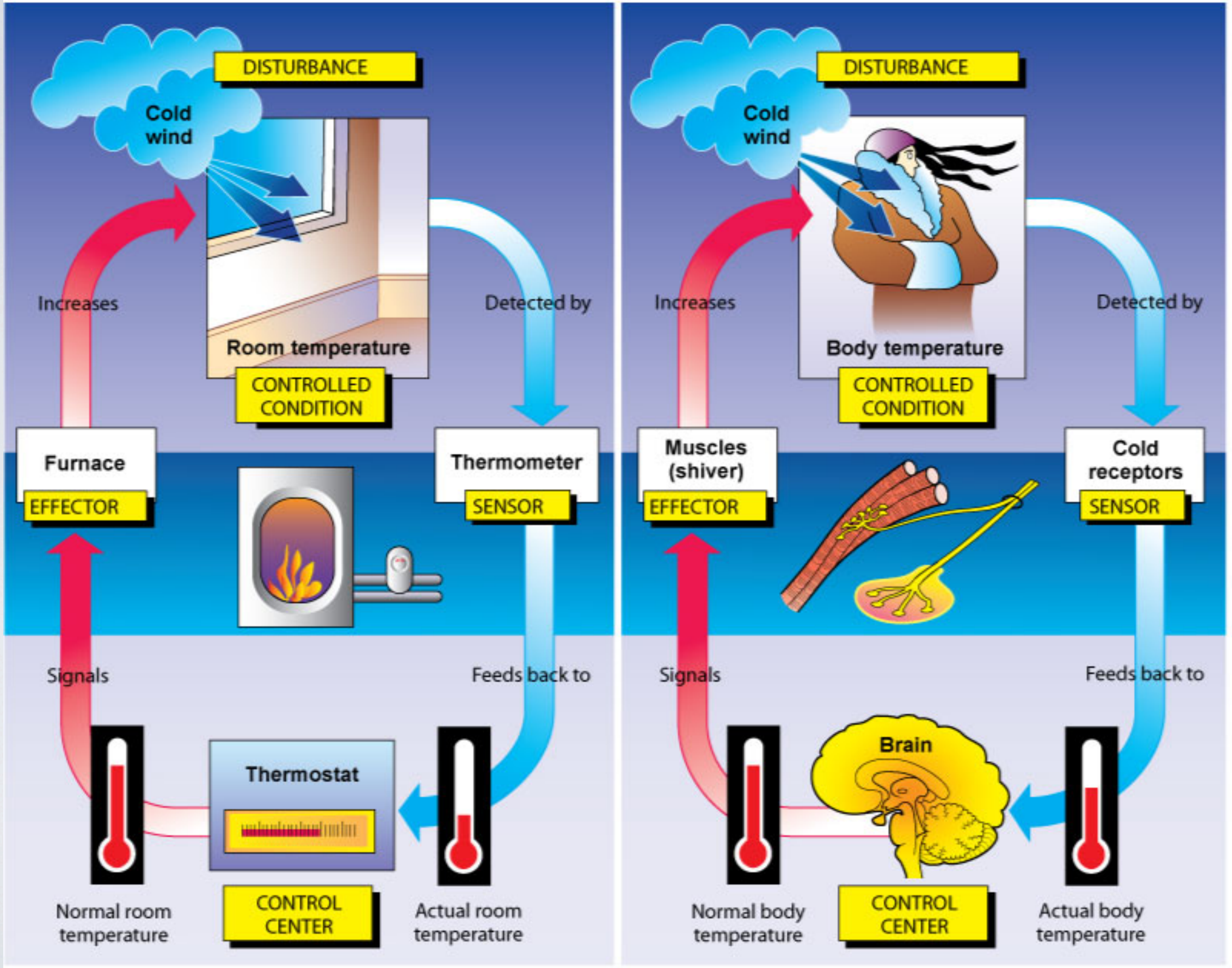
HOMEOSTASIS IN BIOLOGY

= **The tendency of an organism or a cell to regulate its internal conditions, usually by a system of feedback controls, so as to stabilize health and functioning, regardless of the outside changing conditions.**

= The ability of the body or a cell to seek and maintain a <u>condition of equilibrium</u> or stability within its internal environment when dealing with external changes.

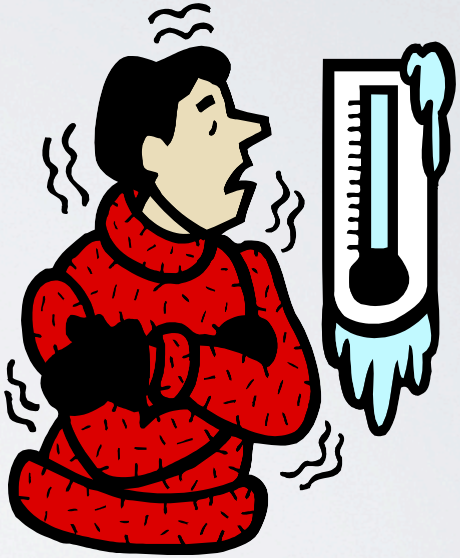
3 components are necessary: 1) receptor/sensor; 2) control center; 3) effector

Homeostasis maintainement



Maintainement of homeostasis (body temperature)

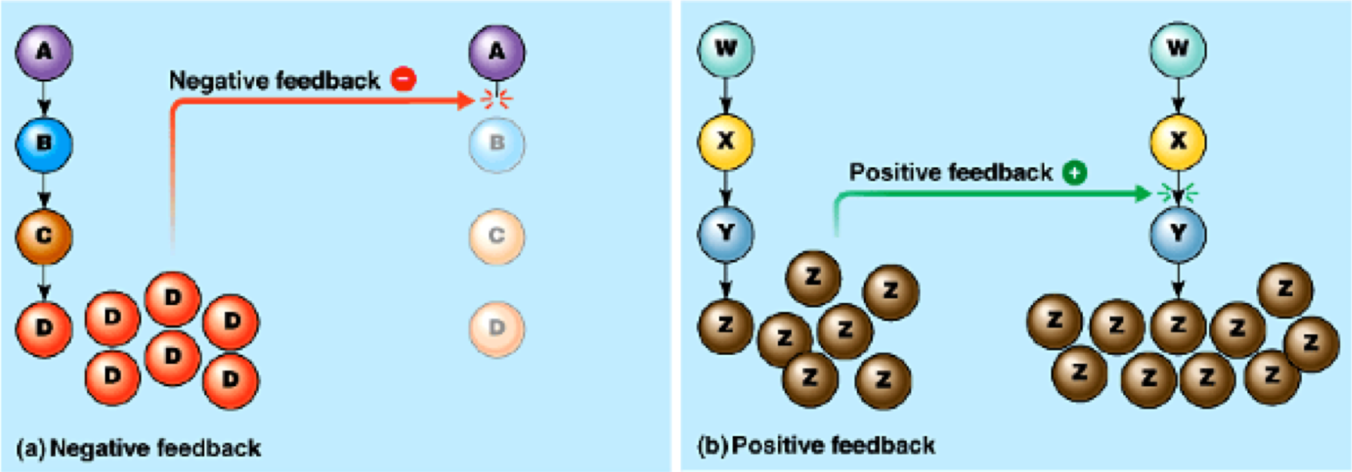
Sweat



Shiver

Regulation & Homeostasis

- Many biological processes are self-regulating, in which an output or product of a process regulates that process.
- Negative feedback or feedback inhibition slows or stops processes.
- Positive feedback speeds a process up.



(a) Negative feedback

(b) Positive feedback

Campbell Fig. 1.8

The cell as unit of health and disease

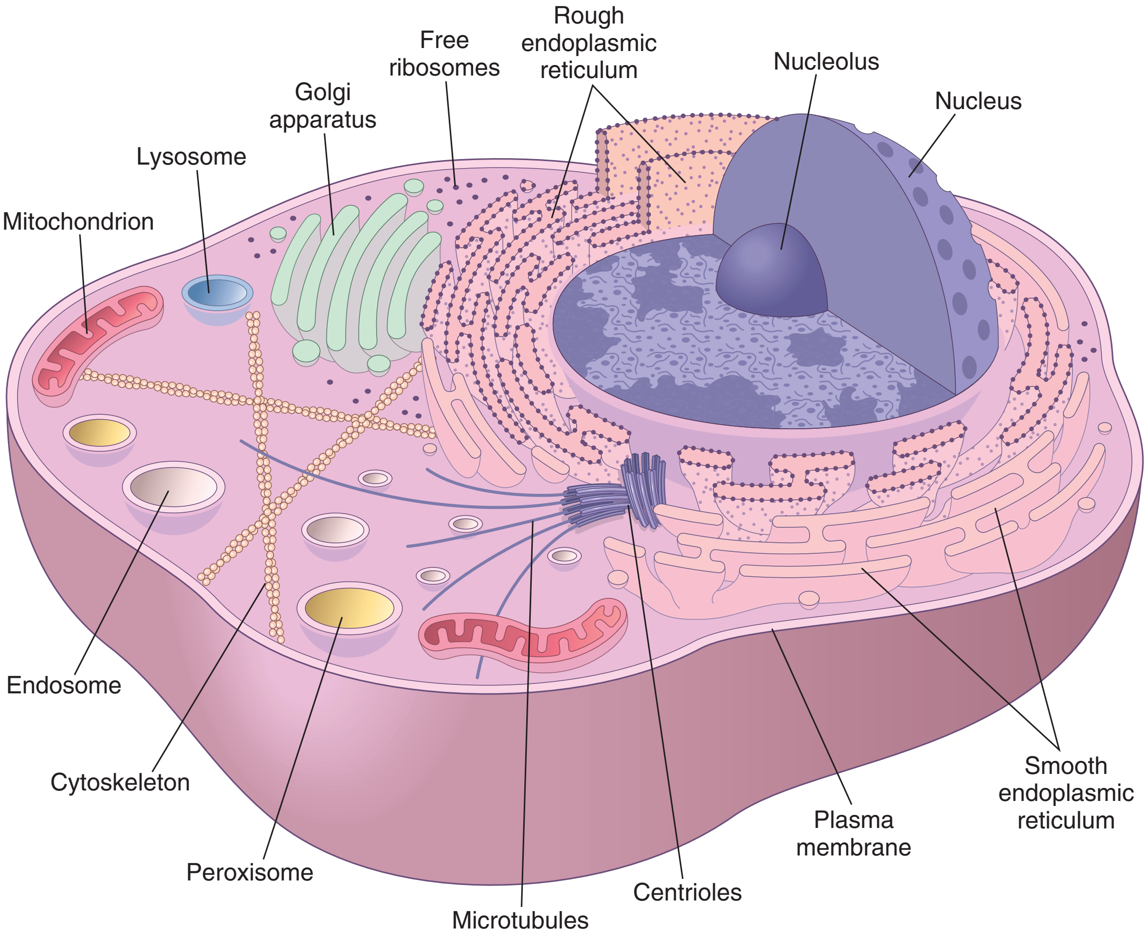
Each branch of science has an elementary unit: the athom for phyisic, the molecule for chemistry. **The elementary unit for the study of diseases by bio-medicine is the cell.** (modificato da *G. Mayno* e *I. Joris*: cellule, tessuti e malattia)

The cell as unit of health and disease

Relative volumes of intracellular organelles (hepatocyte)

Compartment	**% total volume**	**number/cell**	**role in the cell**
Cytosol	54%	1	metabolism, transport, protein translation
Mitochondria	22%	1700	energy generation, apoptosis
Rough ER	9%	1*	synthesis of membrane and secreted proteins
Smooth ER, Golgi	6%	1*	protein modification, sorting, catabolism
Nucleus	6%	1	cell regulation, proliferation, DNA transcription
Endosomes	1%	200	intracellular transport and export, ingestion of extracellular substances
Lysosomes	1%	300	cellular catabolism
Peroxisomes	1%	400	very long-chain fatty acid metabolism

Inherited or acquired alterations of specifc molecules can result in damage to an organelle, alterations of cell functions and possibly cell death



ETIOLOGY: CAUSES OF DISEASE

CONGENITAL: START BEFORE OR IN CONCOMITANCE WITH BIRTH; can be:

- **Genetical,**
- **Due to pregnancy**
- **Due to delivery**

HEREDITARY

ACQUIRED (after birth)

- CHEMICAL

  - Exogenous compounds (either natural or deriving from human activities)
  - Endogenous molecules
    - Metabolic/catabolic products (bilirubin, lactic acid)
    - Reactive oxygen and nitrogen species (ROS, RNS)
    - Modified molecules (oxidized lipoproteins, glycated proteins)

- PHYSICAL

  - temperature
  - radiations
  - pressure
  - noise
  - electrical

- BIOLOGICAL

Direct or indirect damage deriving from pathogen overgrowth

- DEFICIENCIES OR EXCESSES

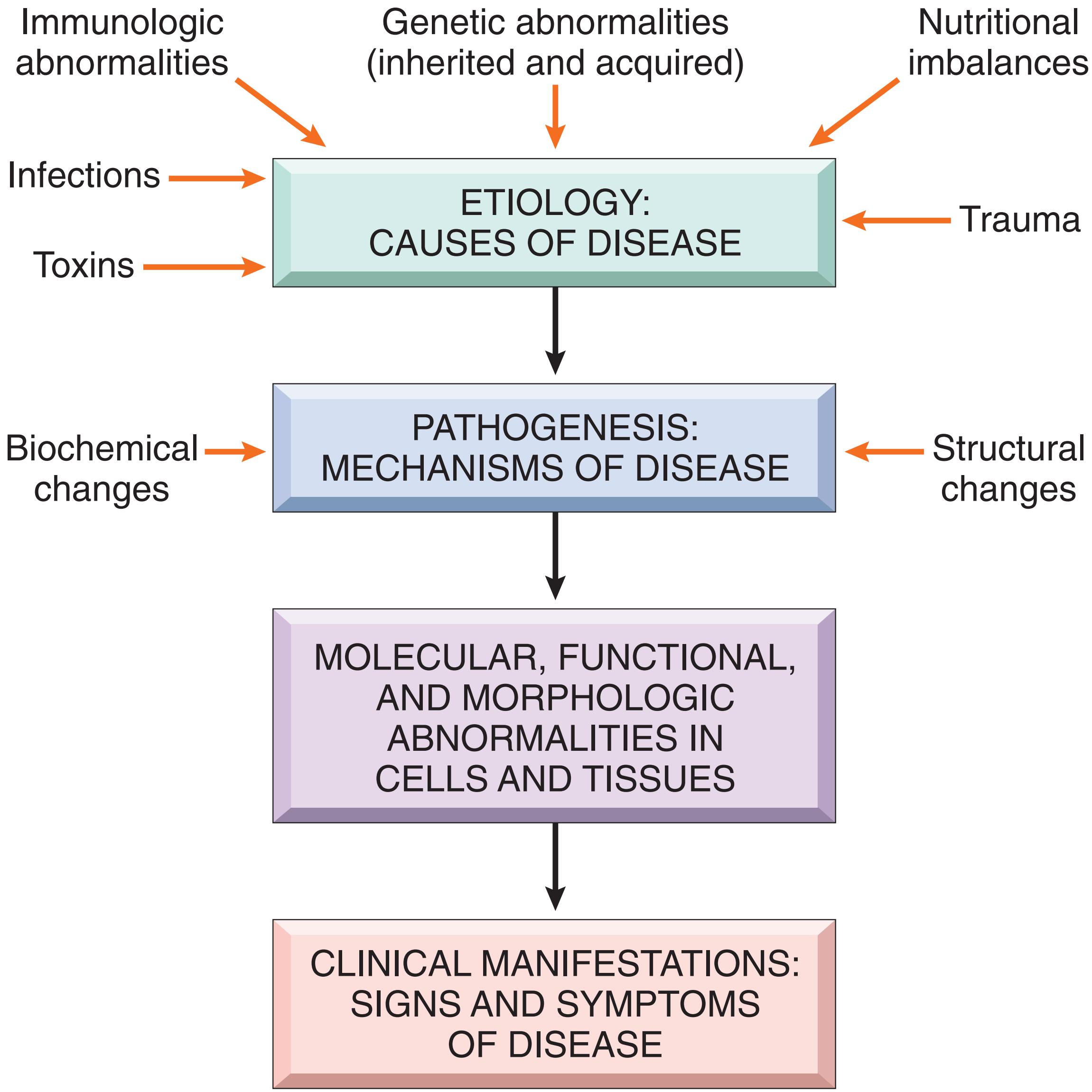
- oxygen deficiency
- vitamin deficiency
- iron deficienzy/overload
- nutriens deficiency/excess
- cholesterol accumulation

- IMMUNOLOGICAL

- IATROGENIC

- drugs

Steps in the evolution of a disease



Immunologic abnormalities

Genetic abnormalities (inherited and acquired)

Nutritional imbalances

Infections

Toxins

Trauma

ETIOLOGY:
CAUSES OF DISEASE

Biochemical changes

Structural changes

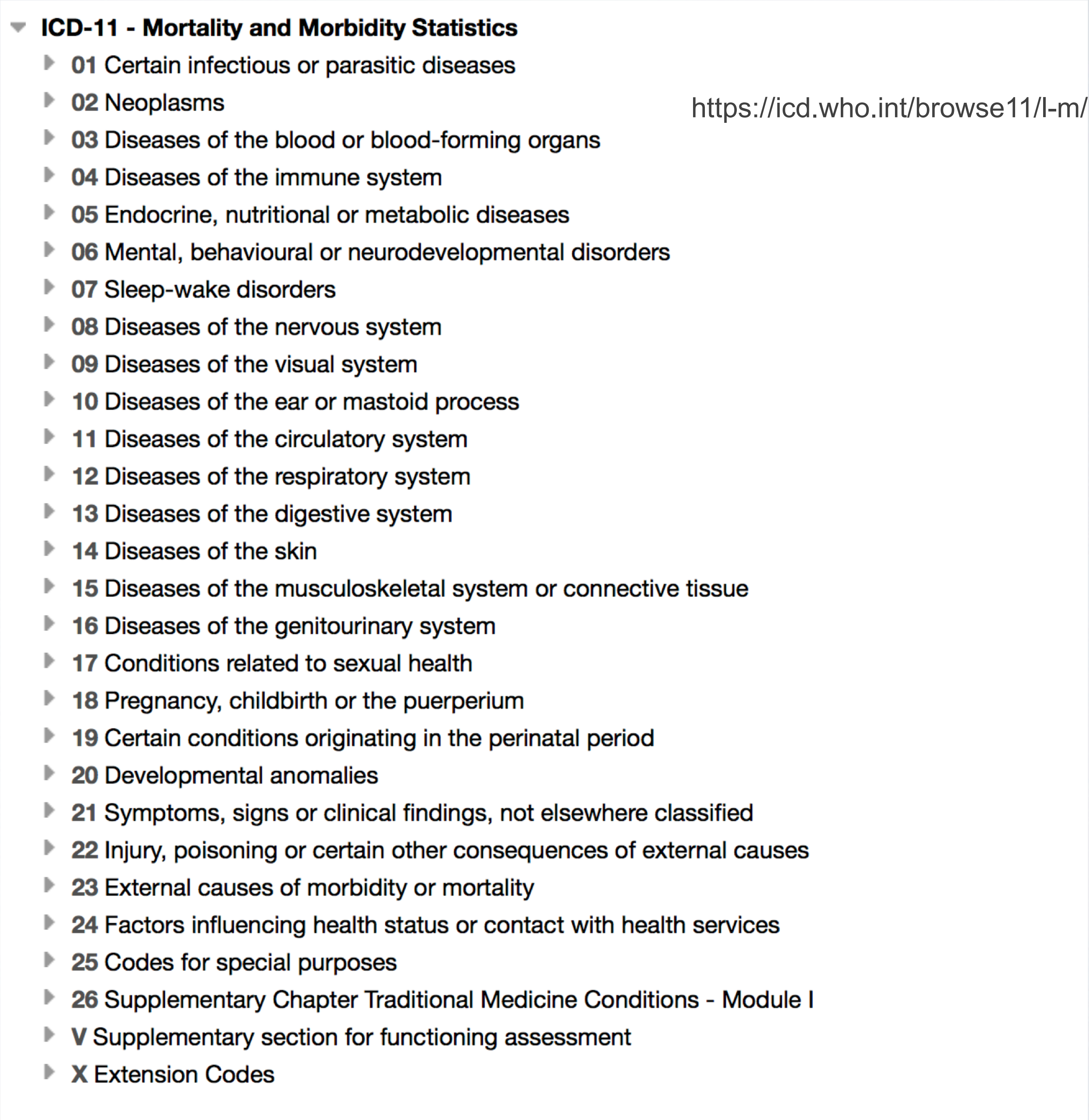
PATHOGENESIS:
MECHANISMS OF DISEASE

MOLECULAR, FUNCTIONAL, AND MORPHOLOGIC ABNORMALITIES IN CELLS AND TISSUES

CLINICAL MANIFESTATIONS:
SIGNS AND SYMPTOMS OF DISEASE

International Statistical Classification of Diseases and Related Health Problems 11th Revision

https://icd.who.int/browse11/l-m/en#/?view=G0



ICD-11 - Mortality and Morbidity Statistics

- 01 Certain infectious or parasitic diseases
- 02 Neoplasms
- 03 Diseases of the blood or blood-forming organs
- 04 Diseases of the immune system
- 05 Endocrine, nutritional or metabolic diseases
- 06 Mental, behavioural or neurodevelopmental disorders
- 07 Sleep-wake disorders
- 08 Diseases of the nervous system
- 09 Diseases of the visual system
- 10 Diseases of the ear or mastoid process
- 11 Diseases of the circulatory system
- 12 Diseases of the respiratory system
- 13 Diseases of the digestive system
- 14 Diseases of the skin
- 15 Diseases of the musculoskeletal system or connective tissue
- 16 Diseases of the genitourinary system
- 17 Conditions related to sexual health
- 18 Pregnancy, childbirth or the puerperium
- 19 Certain conditions originating in the perinatal period
- 20 Developmental anomalies
- 21 Symptoms, signs or clinical findings, not elsewhere classified
- 22 Injury, poisoning or certain other consequences of external causes
- 23 External causes of morbidity or mortality
- 24 Factors influencing health status or contact with health services
- 25 Codes for special purposes
- 26 Supplementary Chapter Traditional Medicine Conditions - Module I
- V Supplementary section for functioning assessment
- X Extension Codes

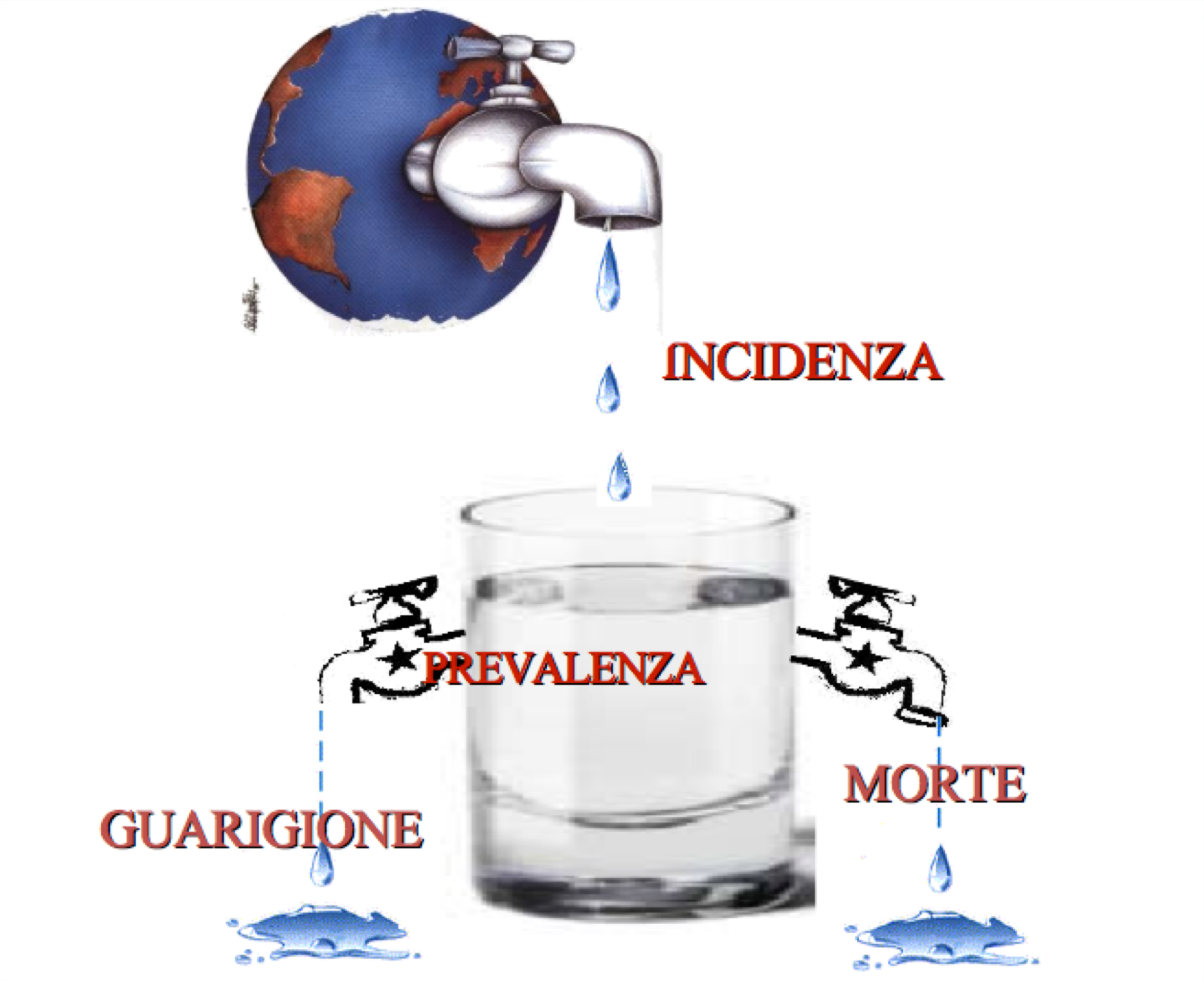
Prevalence and Incidence: definitions

Le misure di **frequenza** delle malattie possono descrivere:
- il verificarsi di nuovi casi (incidenza)
- l'insieme di tutti i casi esistenti in un determinato momento ed in una determinata popolazione (prevalenza)

Incidence = the number of new cases of a disease in a population over a period of time
- Estimates the probability/risk of a person to develop the disease

Prevalence = is a measurement of all individuals affected by the **disease** at a particular time

A prevalence rate is the total number of cases of a disease existing in a population divided by the total population. So, if a measurement of cancer is taken in a population of 40,000 people and 1,200 were recently diagnosed with cancer and 3,500 are living with cancer, then the prevalence of cancer is 0.118. (or 11,750 per 100,000 persons)



INCIDENZA

PREVALENZA

GUARIGIONE

MORTE

Mortality = indicates numbers of deaths by place, time and cause

Mortality rate: is a measure of the number of deaths (in general, or due to a specific cause) in a particular population, scaled to the size of that population, per unit of time. Mortality rate is typically expressed in units of deaths per 1,000 individuals per year.

Developed countries	Deaths in millions	% of deaths
Coronary heart disease	1.33	16.3
Stroke and other cerebrovascular diseases	0.76	9.3
Trachea, bronchus, lung cancers	0.48	5.9
Lower respiratory infections	0.31	3.8
Chronic obstructive pulmonary disease	0.29	3.5
Alzheimer and other dementias	0.28	3.4
Colon and rectum cancers	0.27	3.3
Diabetes mellitus	0.22	2.8
Breast cancer	0.16	2.0
Stomach cancer	0.14	1.8

Mortality

In the world

	Deaths in millions	% of deaths
Coronary heart disease	7.20	12.2
Stroke and other cerebrovascular diseases	5.71	9.7
Lower respiratory infections	4.18	7.1
Chronic obstructive pulmonary disease	3.02	5.1
Diarrhoeal diseases	2.16	3.7
HIV/AIDS	2.04	3.0
Tuberculosis	1.46	2.5
Trachea, bronchus, lung cancers	1.32	2.3
Road traffic accidents	1.27	2.2
Prematurity and low birth weight	1.18	2.0

Morbidity (from Latin *morbidus*, meaning "sick, unhealthy")
= is a diseased state, disability, or poor health due to any cause.
= incidence of a particular disease in a population

Comorbidity = co-existance of more pathological states in the same person

Pattern = a form or model proposed for imitation, example.

- group of traits/features of a cell, person
- design, motif, example, configuration, plan.

Genotype = The entire set of genes in a cell, an organism, or an individual. It is hereditary. What is written in the DNA. It is unchangeable.
= A set of alleles that determines the expression of a particular characteristic or trait (phenotype).

Fenotype = The total characteristics displayed by an organism or cell as a result of the interaction of its genotype and the environment. It can change.
Disease phenotype: group of traits/characteristics that define the pathological process

“Systems Biology” and biomedical applications

1) Diseases– introduction
2) Networks in biomedicine – introduction
3) Application: The human diseasome
4) Application: Comorbidity
5) Innate immunity: introduction and applications
6) Inflammation: introduction and applications
7) Tumors: introduction and applications
8) P4 medicine

Network medicine

The human interactome:

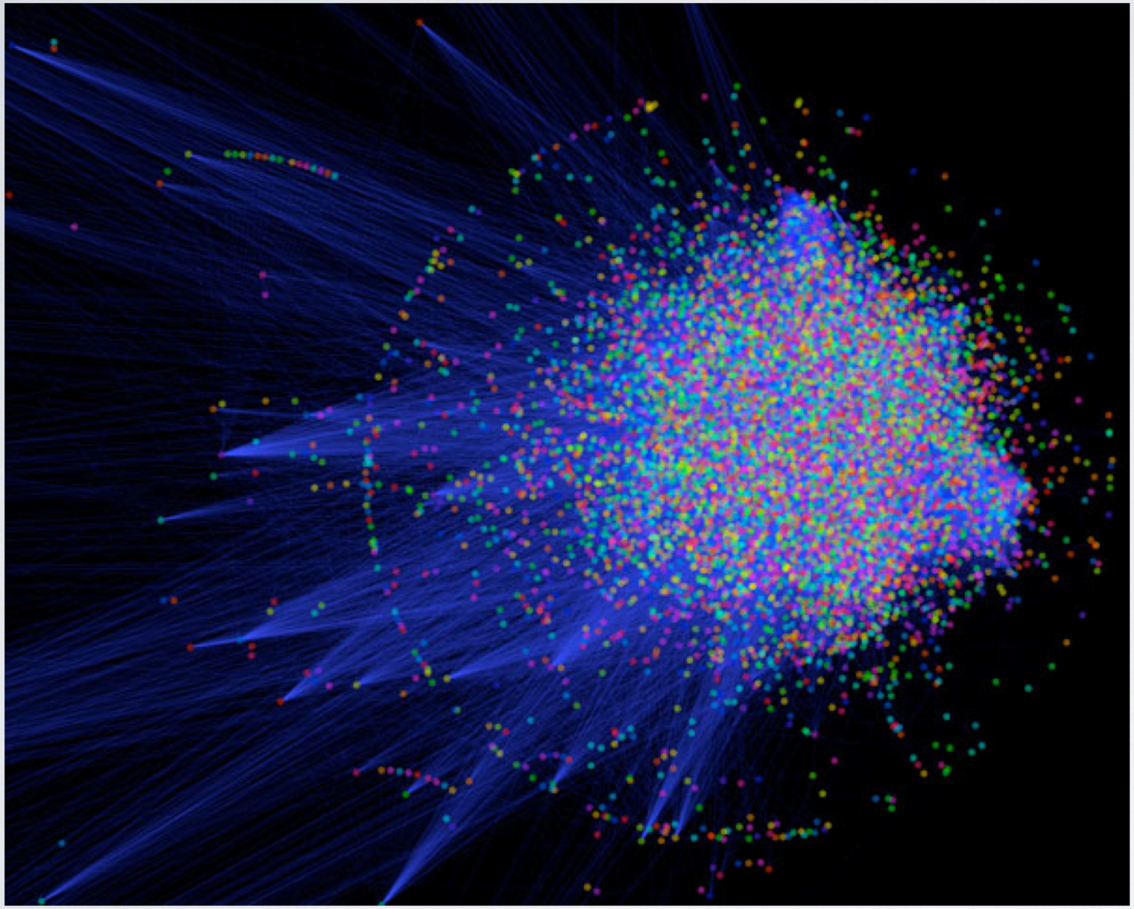
Formato da componenti cellulari che esercitano le loro azioni mediante interazioni con altri componenti cellulari nella stessa cellula, in altre cellule vicine o di altri organi

- 25,000 geni codificanti proteine
- ~30,000 proteine (HPRD)
- 1,000 metaboliti
- numero indefinito di molecole di RNA

↓

- Links: interconnettività intra- e intercellulare

The human interactome



Ray and Charles Eames: "Eventually, everything connects"

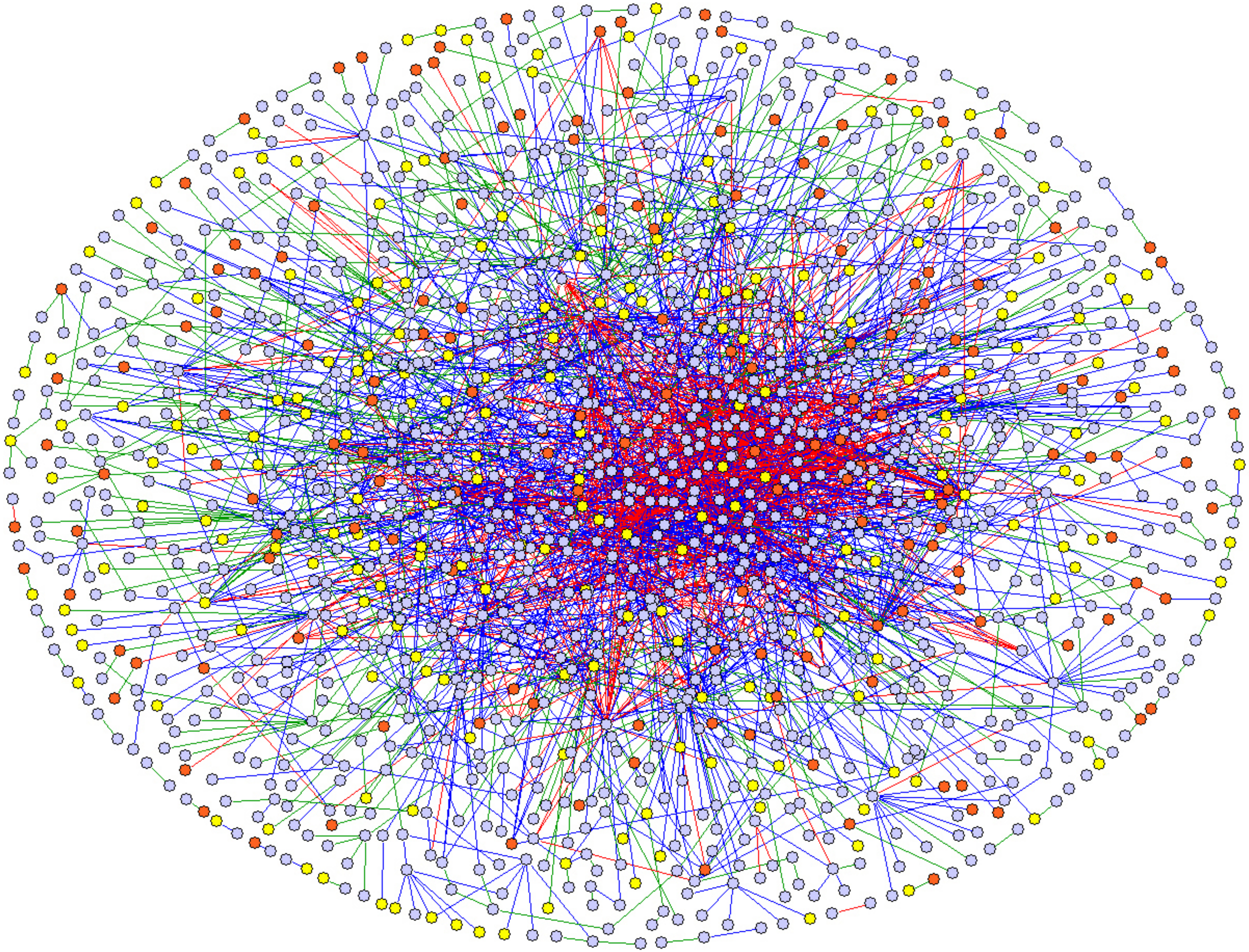
self-made. Dataset created by Andrew Garrow at Unilever UK.	Human Interactome network visualized by Cytoscape 2.5.

Understand the **context** of the gene/protein in a **network** is fundamental to understand the impact of gene/protein alteration on disease phenotype

↓

Fundamental principle:

1. The inter- and intracellular interconnectivity implies that THE IMPACT of a specific genetic abnormality is not restricted to the activity of the gene product that carries it, but CAN SPREAD ALONG THE LINKS of the network and alter the activity of gene products that otherwise carry no defects.



Understand the **context** of the gene/protein in a **network** is fundamental to understand the impact of gene/protein alteration on disease phenotype

↓

Fundamental principle:

2. A disease phenotype is rarely a consequence of an abnormality in a single effector gene product, but reflects various pathobiological processes that are connected in a complex network

The importance of networks in medicine:

1. Identification of genes, proteins and “pathways” involved in diseases
2. Predict disease mechanisms
3. Find biomarkers
4. New disease classifications
5. New pharmacological targets and in silico farmacology

Network maps in biologia

1. Protein-protein interaction (PPI) networks
2. Metabolic networks
3. Regulatory networks
4. RNA networks

1. Protein-protein interaction networks:

Nodi = proteine; Links/Edges (Archi) = interazioni fisiche proteina-proteina

- Munich Information Center for Protein Sequence (MIPS) protein interaction database
- Biomolecular Interaction Network Database (BIND)
- Database of Interacting Proteins (DIP)
- Molecular Interaction database (MINT)
- protein Interaction database (IntAct).
- Biological General Repository for Interaction Datasets (BioGRID)
- Human Protein Reference Database (HPRD) have attempted larger-scale curation of data.
- STRING database contains known and predicted protein–protein interactions.

2. Metabolic networks:

Nodi = metaboliti connessi se partecipano nelle stesse relazioni chimiche

The metabolic network maps are probably the most comprehensive of all biological networks.

Databases:

-Kyoto Encyclopedia of Genes and Genomes (KEGG)

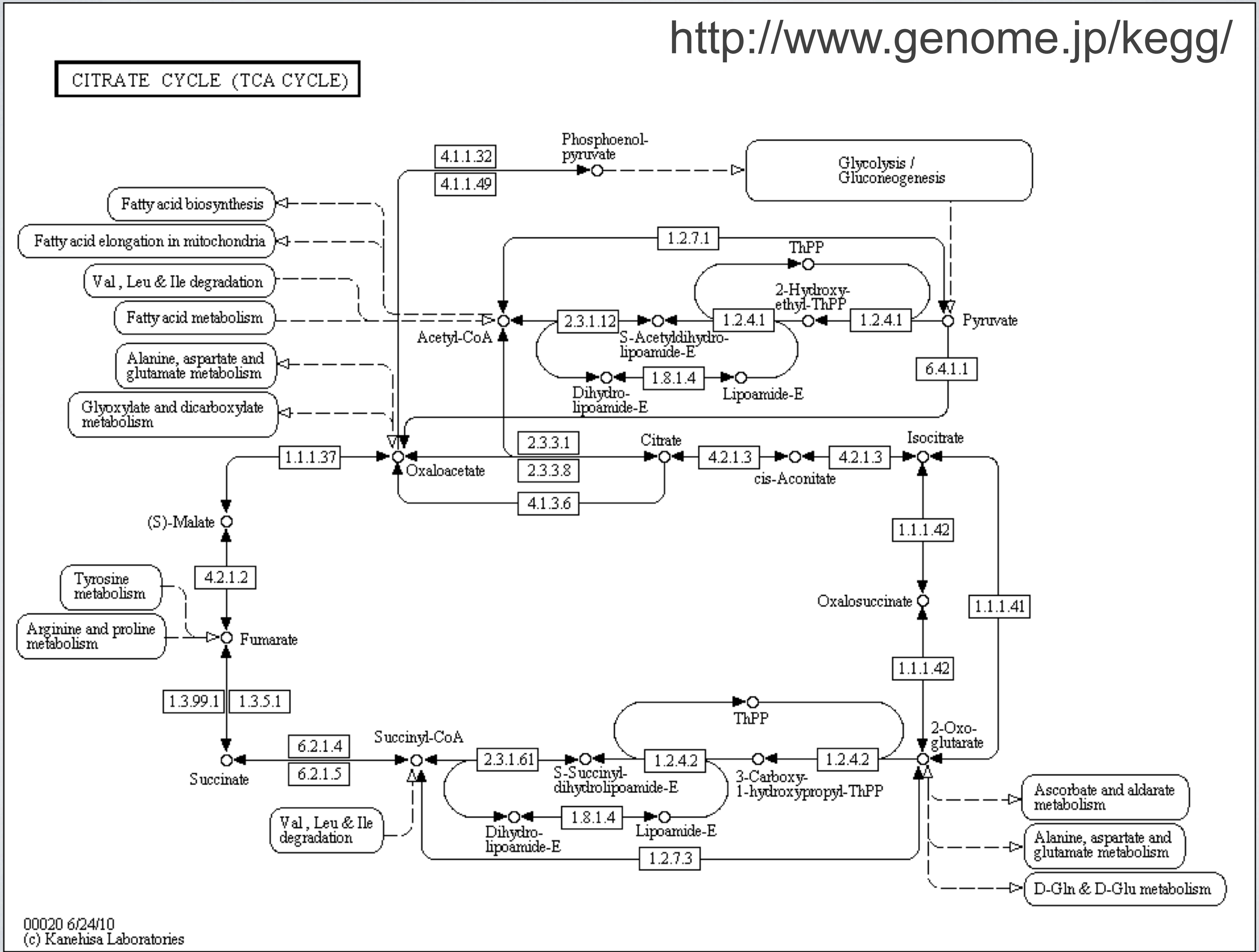
-Biochemical Genetic and Genomics knowledgebase (BIGG)

- lavoro di Duarte *et al (Proc. Natl Acad. Sci. 2007)*: a comprehensive literature-based genome-scale metabolic reconstruction of human metabolism, with 2,766 metabolites and 3,311 metabolic and transport reactions.

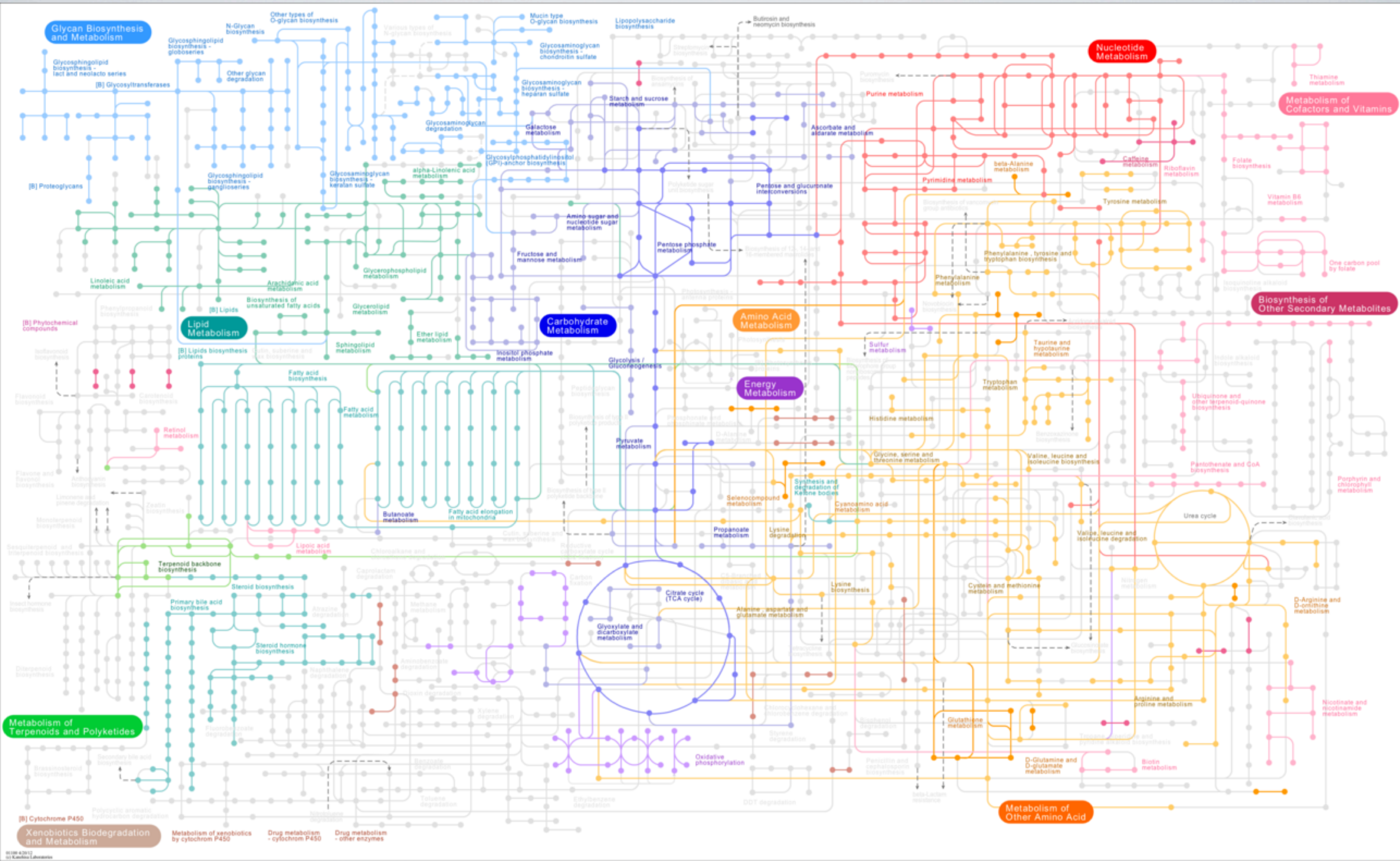
- lavoro di Ma *et al (Mol. Syst. Biol., 2007)*: an independent manual construction containing nearly 3,000 metabolic reactions, organized into about 70 human-specific metabolic pathways.

Kyoto Encyclopedia of Genes and Genomes

http://www.genome.jp/kegg/



Metabolic pathways homo sapiens (HSA)


Glycan Biosynthesis and Metabolism
N-Glycan biosynthesis
Other types of O-glycan biosynthesis
Mucin type O-glycan biosynthesis
Lipopolysaccharide biosynthesis
Butirosin and neomycin biosynthesis
Nucleotide Metabolism
Thiamine metabolism
Metabolism of Cofactors and Vitamins
Glycosphingolipid biosynthesis - globoseries
Glycosaminoglycan biosynthesis - chondroitin sulfate
Glycosphingolipid biosynthesis - lact and neolacto series
Other glycan degradation
[B] Glycosyltransferases
Glycosaminoglycan biosynthesis - heparan sulfate
Starch and sucrose metabolism
Purine metabolism
Glycosaminoglycan degradation
Galactose metabolism
Ascorbate and aldarate metabolism
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
beta-Alanine metabolism
Caffeine metabolism
Riboflavin metabolism
Folate biosynthesis
[B] Proteoglycans
Glycosphingolipid biosynthesis - ganglioseries
Glycosaminoglycan biosynthesis - keratan sulfate
alpha-Linolenic acid metabolism
Pentose and glucuronate interconversions
Pyrimidine metabolism
Tyrosine metabolism
Vitamin B6 metabolism
Amino sugar and nucleotide sugar metabolism
Pentose phosphate metabolism
Fructose and mannose metabolism
Phenylalanine , tyrosine and tryptophan biosynthesis
One carbon pool by folate
Linoleic acid metabolism
Arachidonic acid metabolism
Glycerophospholipid metabolism
Phenylalanine metabolism
Biosynthesis of Other Secondary Metabolites
[B] Phytochemical compounds
Biosynthesis of unsaturated fatty acids
[B] Lipids
Glycerolipid metabolism
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Ether lipid metabolism
Sphingolipid metabolism
[B] Lipids biosynthesis proteins
Inositol phosphate metabolism
Sulfur metabolism
Taurine and hypotaurine metabolism
Glycolysis / Gluconeogenesis
Fatty acid biosynthesis
Energy Metabolism
Tryptophan metabolism
Ubiquinone and other terpenoid-quinone biosynthesis
Fatty acid metabolism
Retinol metabolism
Histidine metabolism
Pyruvate metabolism
Glycine, serine and threonine metabolism
Valine, leucine and isoleucine biosynthesis
Pantothenate and CoA biosynthesis
Porphyrin and chlorophyll metabolism
Synthesis and degradation of Ketone bodies
Selenocompound metabolism
Cyanoamino acid metabolism
Butanoate metabolism
Fatty acid elongation in mitochondria
Urea cycle
Propanoate metabolism
Lysine degradation
Valine, leucine and isoleucine degradation
Lipoic acid metabolism
Terpenoid backbone biosynthesis
Lysine biosynthesis
Cystein and methionine metabolism
D-Arginine and D-ornithine metabolism
Steroid biosynthesis
Citrate cycle (TCA cycle)
Primary bile acid biosynthesis
Alanine, aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
Steroid hormone biosynthesis
Arginine and proline metabolism
Nicotinate and nicotinamide metabolism
Metabolism of Terpenoids and Polyketides
Glutathione metabolism
Oxidative phosphorylation
D-Glutamine and D-glutamate metabolism
Biotin metabolism
Metabolism of Other Amino Acid
[B] Cytochrome P450
Xenobiotics Biodegradation and Metabolism
Metabolism of xenobiotics by cytochrom P450
Drug metabolism - cytochrom P450
Drug metabolism - other enzymes
01100 4/20/12
(c) Kanehisa Laboratories

3. Regulatory networks:

Nodi: geni, fattori di trascrizione, ensimi; Links: tra geni e fattori di trascrizione o modificazioni posttraslazionali

The human regulatory network is the most incomplete among all biological networks.

Includes:

- Relationships between transcription factors and genes
- Post-translational modifications (es. Kinase-substrate)

Databases:

1. Data generated by experimental techniques, such as chromatin immunoprecipitation (ChIP) followed by microarrays (ChIP–chip) and ChIP followed by sequencing (ChIP–seq), have started to be collected in databases such as:

- Universal Protein Binding Microarray Resource for Oligonucleotide Binding Evaluation (UniPROBE)
- JASPAR.

3. Regulatory networks:

(continuation)

2. Literature-curated and predicted protein–DNA interactions have been compiled in various databases, such as:

-TRANSFAC

- B-cell interactome (BCI).

3. Human post-translational modifications can be found in databases such as:

- Phospho.ELM

- PhosphoSite

- Phosphorylation site database (PHOSIDA)

- NetPhorest

- CBS prediction database.

4. RNA networks:

RNA networks contain RNA–RNA or RNA–DNA interactions (siRNA, miRNAs).

microRNA–gene networks have been constructed using predicted microRNA targets available in databases such as:

- **TargetScan,**
- **PicTar,**
- **microRNA,**
- **miRBase**
- **miRDB.**

The number of experimentally supported targets is also increasing, and they are now compiled in databases such as:

- **TarBase**
- **miRecords**

Ontology = the branch of metaphysics dealing with the nature of being
= a set of concepts and categories in a subject area or domain that shows their **properties and the relations** between them.

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a **controlled vocabulary of terms for describing gene product characteristics and gene product annotation** data from GO Consortium members, as well as tools to access and process this data.

The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated: 1) **biological processes**, 2) **cellular components** and 3) **molecular functions** in a species-independent manner.

http://www.geneontology.org/

"Gene Ontology" classifies functions along three aspects:

1) **Biological process**
2) Molecular function
3) Cellular component

- **CELLULAR COMPONENT**
DEFINES where gene products are active

- **BIOLOGICAL PROCESS**
defines pathways and larger processes made up of the activities of multiple gene products.

- **MOLECOLAR FUNCTION**
molecular activities of gene products/BIOCHEMICAL ACTIVITY.

http://www.geneontology.org/

Altri tipi di rete

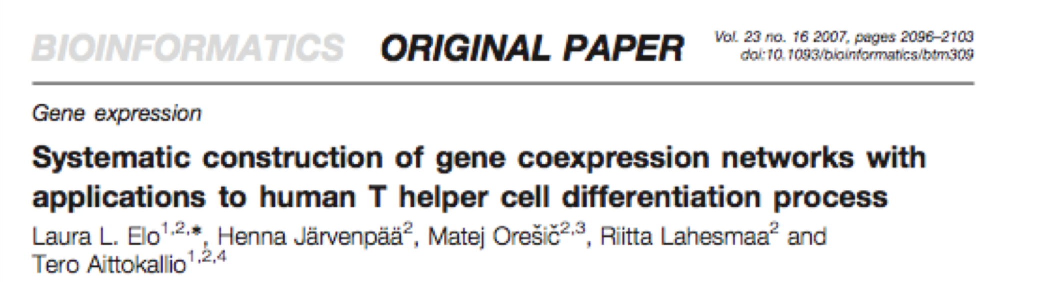
Gene coexpression networks - geni con pattern di co-espresssione simile sono linked e collegati funzionalmente. Chiarisce la funzione di un gene su scala globale

Coexpressin networks: novel holistic approaches to analyze and interpret/explain microarray data.

Coexpression networks represent an alternative to more conventional statistic analysis and clustering.

There is a cutoff - coexpression threshold (al di sopra quale le interazioni gene-gene sono considerate rilevanti). Il cutoff viene calcolato dalla topologia della rete. Si usa *PCC*.

The nodes are genes. The links are non-direct.



BIOINFORMATICS **ORIGINAL PAPER** *Vol. 23 no. 16 2007, pages 2096–2103*
doi:10.1093/bioinformatics/btm309

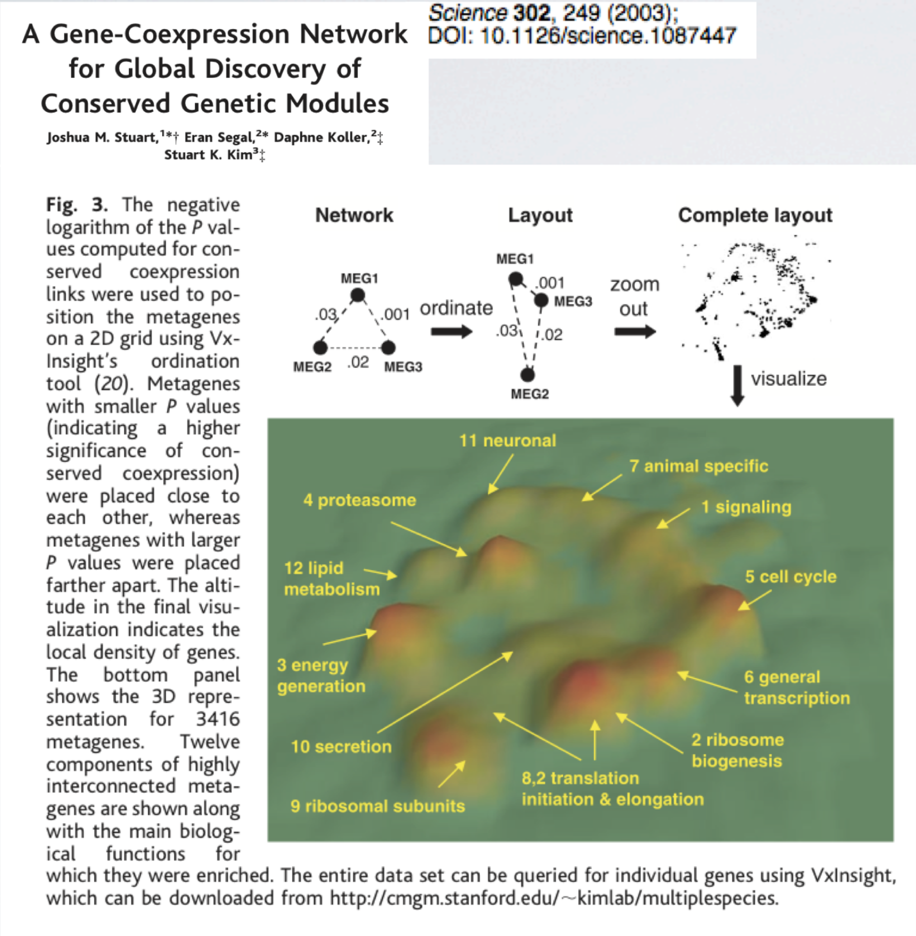
Gene expression

Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process

Laura L. Elo[1,2,*], Henna Järvenpää[2], Matej Orešič[2,3], Riitta Lahesmaa[2] and Tero Aittokallio[1,2,4]

Other types of networks

Gene coexpression networks

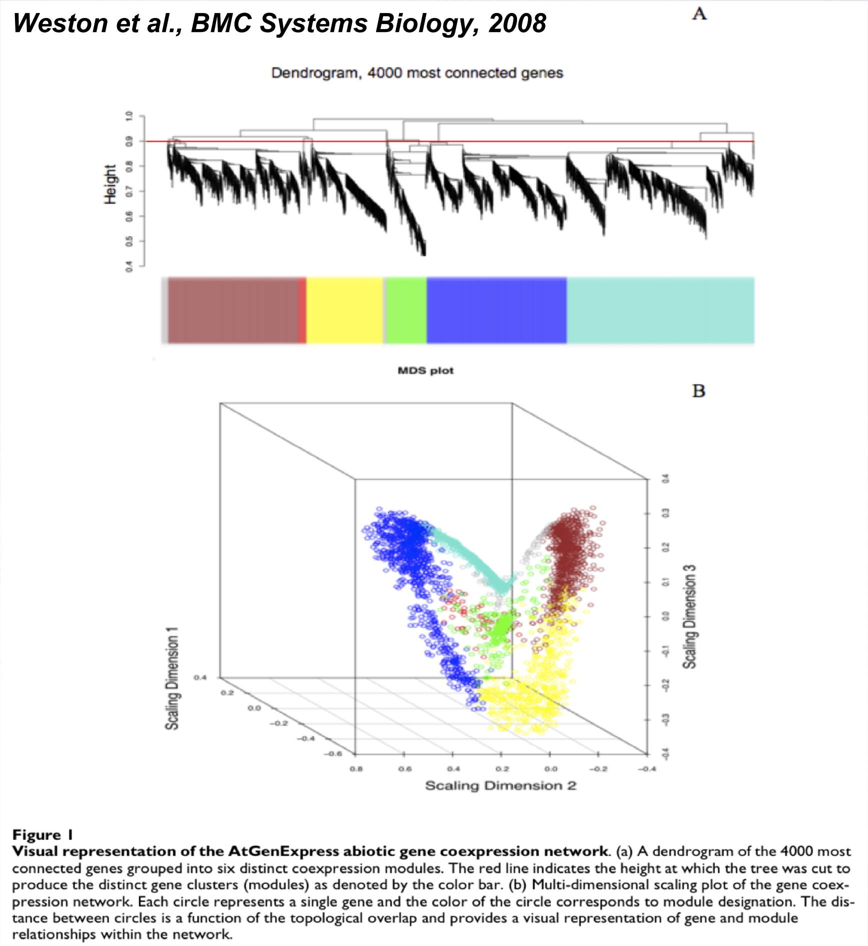


Science 302, 249 (2003);
DOI: 10.1126/science.1087447

A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules

Joshua M. Stuart,[1*†] Eran Segal,[2*] Daphne Koller,[2‡] Stuart K. Kim[3‡]

Fig. 3. The negative logarithm of the *P* values computed for conserved coexpression links were used to position the metagenes on a 2D grid using VxInsight's ordination tool (*20*). Metagenes with smaller *P* values (indicating a higher significance of conserved coexpression) were placed close to each other, whereas metagenes with larger *P* values were placed farther apart. The altitude in the final visualization indicates the local density of genes. The bottom panel shows the 3D representation for 3416 metagenes. Twelve components of highly interconnected metagenes are shown along with the main biological functions for which they were enriched. The entire data set can be queried for individual genes using VxInsight, which can be downloaded from http://cmgm.stanford.edu/~kimlab/multiplespecies.



Weston et al., BMC Systems Biology, 2008

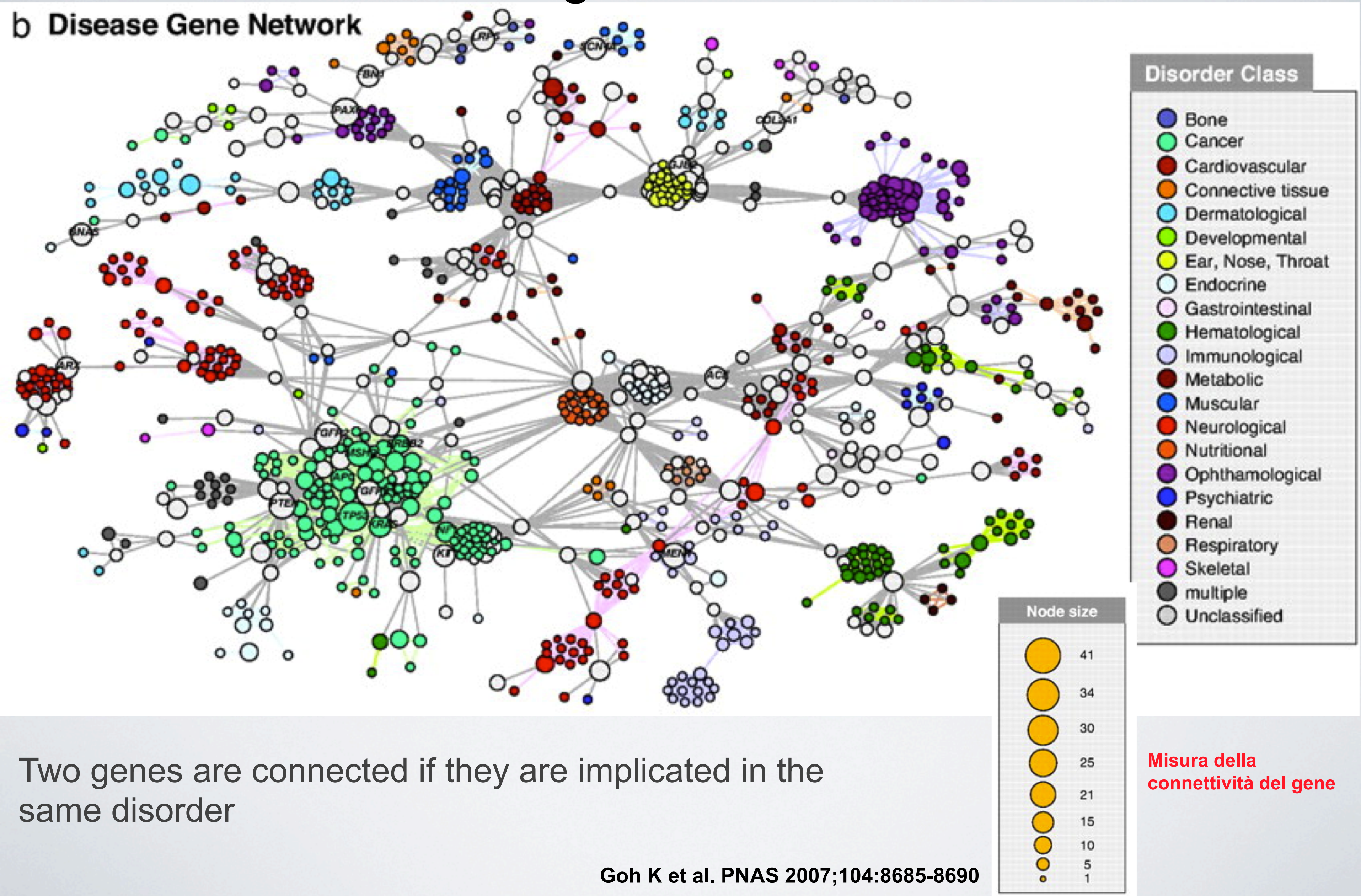
Figure 1
Visual representation of the AtGenExpress abiotic gene coexpression network. (a) A dendrogram of the 4000 most connected genes grouped into six distinct coexpression modules. The red line indicates the height at which the tree was cut to produce the distinct gene clusters (modules) as denoted by the color bar. (b) Multi-dimensional scaling plot of the gene coexpression network. Each circle represents a single gene and the color of the circle corresponds to module designation. The distance between circles is a function of the topological overlap and provides a visual representation of gene and module relationships within the network.

>3.182 DNA micorarrays human, flies, worms, yeast; found >20.000 coexpressions conserved during evolution

Each gene = spot;
Short distance = genes functionally relared

Other types of networks

Disease gene network



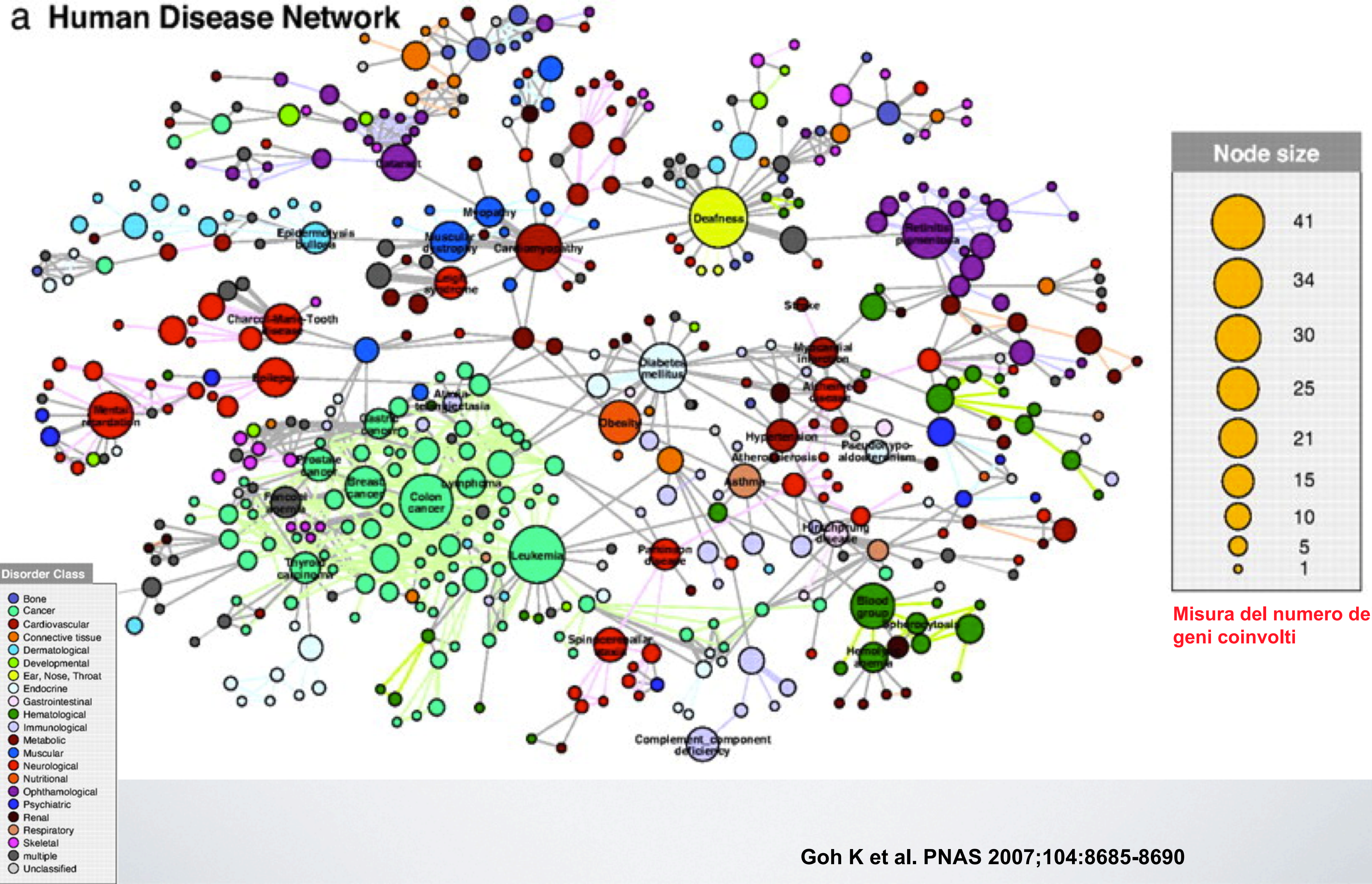
Misura della connettività del gene

Two genes are connected if they are implicated in the same disorder

Goh K et al. PNAS 2007;104:8685-8690

Other types of networks

Human disease network



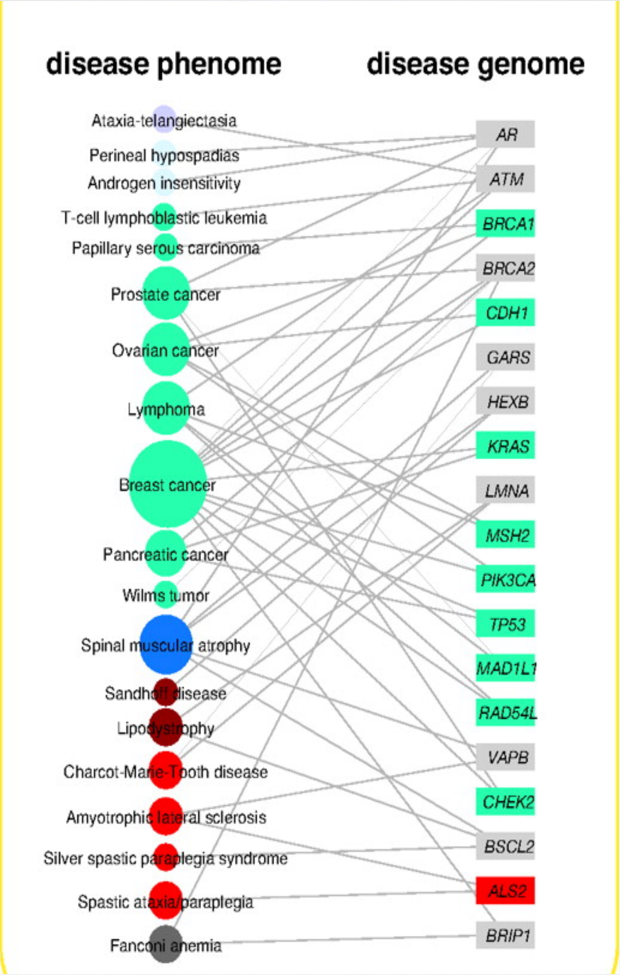
Misura del numero dei geni coinvolti

Goh K et al. PNAS 2007;104:8685-8690

Other types of networks

BIPARTITE Network

Fenoma = **THE SET OF ALL PHENOTYPES EXPRESSED BY A CELL, TISSUE, ORGAN, ORGANISM, OR SPECIES**



Goh K et al. PNAS 2007;104:8685-8690

Network medicine

Applications:
The human diseasome

Background:

- **Derives from the simple observation of inter- and intracellular interconnectivity**
- **Disease modules overlap**
- **Perturbations in a module influences other modules**
- **Estensive studies (per es. GWAS) have shown numerous gene-disease associations**
- **Many genetic diseases have more than one loci involved (locus heterogeneity)**
- **Different mutations in the same genes may give birth of different phenotypes; ex. TP53, Ras**
- **The linkage of one gene to diferent phenotypes suggest that different disease phenotypes may have a common origin and that the diseases are linked (Goh et al., PNAS 2007).**

The concept of **Diseasome**

Basic concept:

- The involvment of a gene in more diseases suggest these diseases are linked and may have a common origin (Goh et al., PNAS 2007).

Disease 1 Common Gene Disease 2

Disease 1 and 2 are linked

The Diseasome concept

Diseasome = disease network in which:

- the nodes are diseases
-the links between two nodes are due to comon components

Organizzazione globale, sistemica delle malattie “genetiche”, un “framework” che unisce le conoscenze sui “disease genes” (disease genome) e sulle malattie genetiche (disease phenome)

Why study the diseasome?

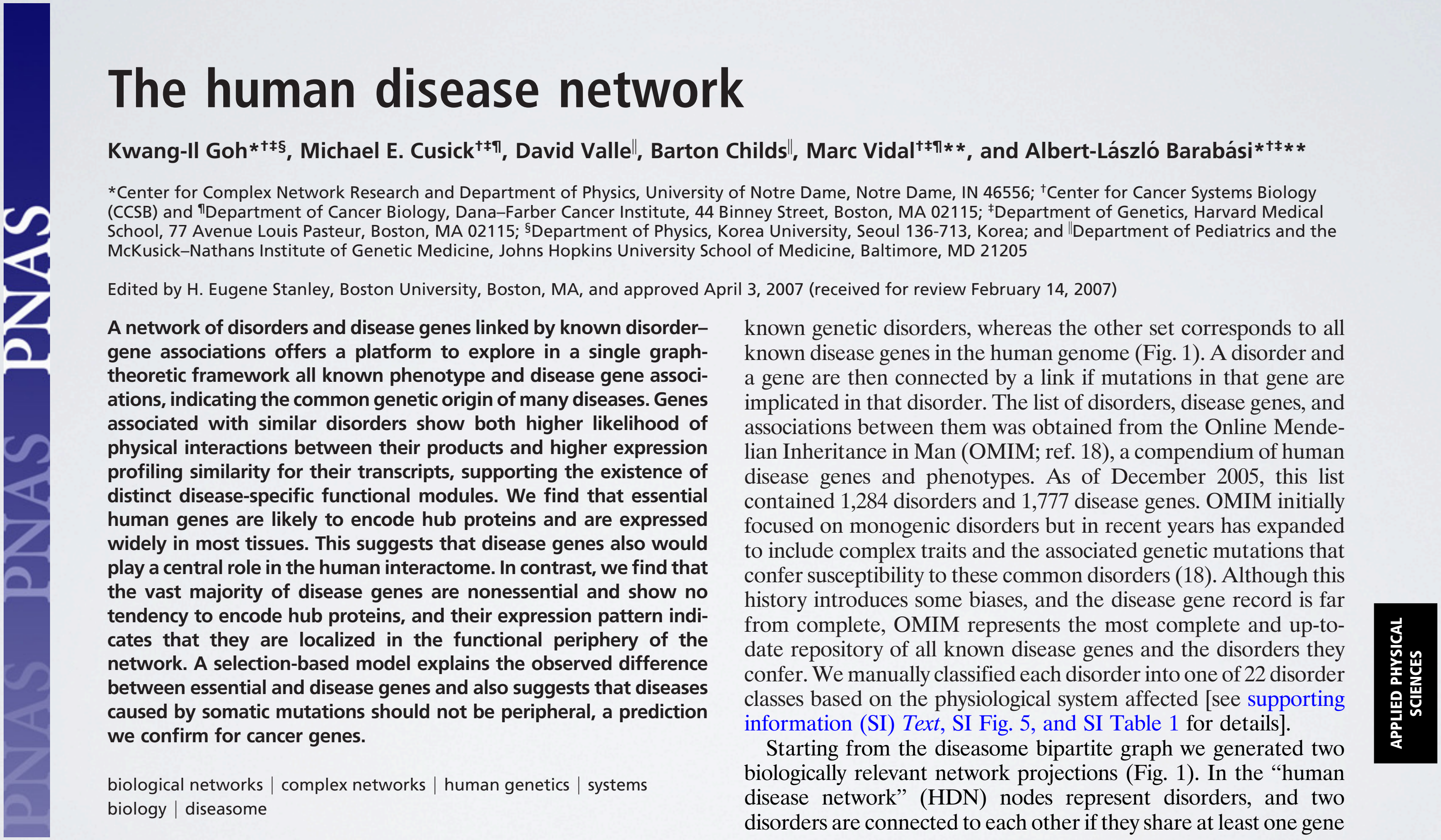
1. Understand why some diseases develop in parallel

2. Extract “comorbidity”

3. Drug discovery (use drugs already approved for diseases with similar molecular basis)

4. New approaches for prevention, diagnosis and treatment

Building the human diseasome



The human disease network

Kwang-Il Goh[*†‡§], Michael E. Cusick[†‡¶], David Valle[||], Barton Childs[||], Marc Vidal[†‡¶**], and Albert-László Barabási[*†‡**]

[*]Center for Complex Network Research and Department of Physics, University of Notre Dame, Notre Dame, IN 46556; [†]Center for Cancer Systems Biology (CCSB) and [¶]Department of Cancer Biology, Dana–Farber Cancer Institute, 44 Binney Street, Boston, MA 02115; [‡]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115; [§]Department of Physics, Korea University, Seoul 136-713, Korea; and [||]Department of Pediatrics and the McKusick–Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205

Edited by H. Eugene Stanley, Boston University, Boston, MA, and approved April 3, 2007 (received for review February 14, 2007)

A network of disorders and disease genes linked by known disorder–gene associations offers a platform to explore in a single graph-theoretic framework all known phenotype and disease gene associations, indicating the common genetic origin of many diseases. Genes associated with similar disorders show both higher likelihood of physical interactions between their products and higher expression profiling similarity for their transcripts, supporting the existence of distinct disease-specific functional modules. We find that essential human genes are likely to encode hub proteins and are expressed widely in most tissues. This suggests that disease genes also would play a central role in the human interactome. In contrast, we find that the vast majority of disease genes are nonessential and show no tendency to encode hub proteins, and their expression pattern indicates that they are localized in the functional periphery of the network. A selection-based model explains the observed difference between essential and disease genes and also suggests that diseases caused by somatic mutations should not be peripheral, a prediction we confirm for cancer genes.

biological networks | complex networks | human genetics | systems biology | diseasome

known genetic disorders, whereas the other set corresponds to all known disease genes in the human genome (Fig. 1). A disorder and a gene are then connected by a link if mutations in that gene are implicated in that disorder. The list of disorders, disease genes, and associations between them was obtained from the Online Mendelian Inheritance in Man (OMIM; ref. 18), a compendium of human disease genes and phenotypes. As of December 2005, this list contained 1,284 disorders and 1,777 disease genes. OMIM initially focused on monogenic disorders but in recent years has expanded to include complex traits and the associated genetic mutations that confer susceptibility to these common disorders (18). Although this history introduces some biases, and the disease gene record is far from complete, OMIM represents the most complete and up-to-date repository of all known disease genes and the disorders they confer. We manually classified each disorder into one of 22 disorder classes based on the physiological system affected [see supporting information (SI) *Text*, SI Fig. 5, and SI Table 1 for details].

Starting from the diseasome bipartite graph we generated two biologically relevant network projections (Fig. 1). In the "human disease network" (HDN) nodes represent disorders, and two disorders are connected to each other if they share at least one gene

APPLIED PHYSICAL SCIENCES

Albert-László Barabási

https://www.google.com/search?client=safari&rls=en&q=albert+laszlo+barabasi&ie=UTF-8&oe=UTF-8

energy range of the majority spin d-band of Co. This is also observed in Fig. 3A.

For ALO and ALO/STO barriers, a predominant tunneling of s-character electrons (see arrow in Fig. 2B) is the usual explanation of the positive polarization (*6–8*). The rapid drop with bias (Fig. 3B) is similar to what has been observed in most junctions with ALO barriers, and completely different from what is obtained when the tunneling is predominantly by d-character electrons (Fig. 3A). The origin of this rapid decrease of the TMR at relatively small bias has never been clearly explained. This is roughly consistent with the energy dependence of the DOS induced by sp-d bonding effects on the first atomic layer of ALO in the calculation of Nguyen-Mahn *et al*. (*8*) for the Co-ALO interface. But Zhang *et al*. (*13*) have also shown that a large part of the TMR drop can be attributed to the excitation of spin waves.

The experiments reported here and in several recent publications (*3, 4*) demonstrate the important role of the electronic structure of the metal-oxide interface in determining the spin polarization of the tunneling electrons. The negative polarization for the Co-STO interface has been ascribed to d-d bonding effects between Al and Ti (*4*). This interpretation is similar to

Emergence of Scaling in Random Networks

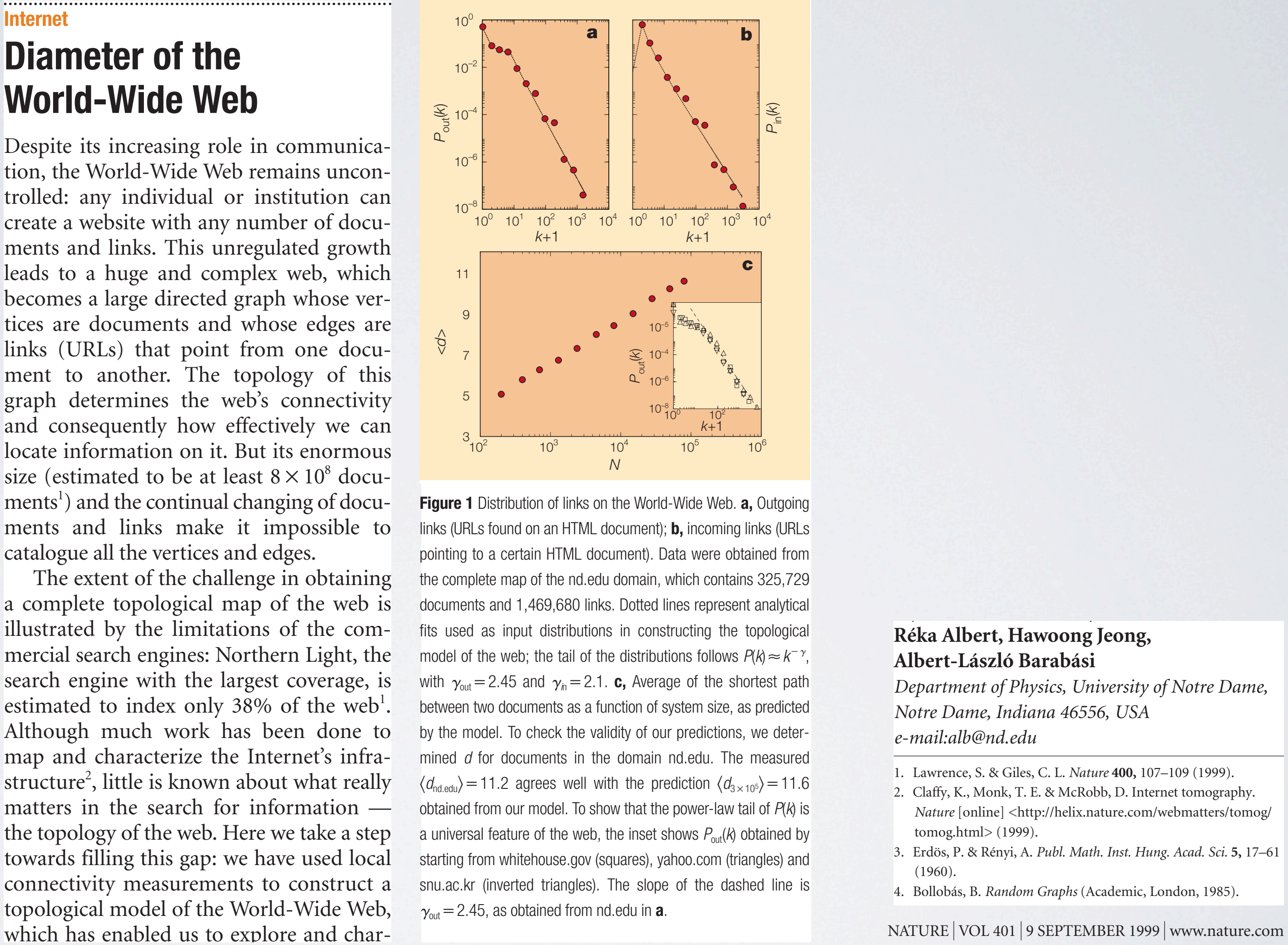
Albert-László Barabási* and Réka Albert

Systems as diverse as genetic networks or the World Wide Web are best described as networks with complex topology. A common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution. This feature was found to be a consequence of two generic mechanisms: (i) networks expand continuously by the addition of new vertices, and (ii) new vertices attach preferentially to sites that are already well connected. A model based on these two ingredients reproduces the observed stationary scale-free distributions, which indicates that the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems.

The inability of contemporary science to describe systems composed of nonidentical elements that have diverse and nonlocal interactions currently limits advances in many disciplines, ranging from molecular biology to computer science (*1*). The difficulty of describing these systems lies partly in their topology: Many of them form rather complex networks whose vertices are the elements of the system and whose edges represent the interactions between them. For example, liv-

Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA.

*To whom correspondence should be addressed. E-mail: alb@nd.edu



Internet

Diameter of the World-Wide Web

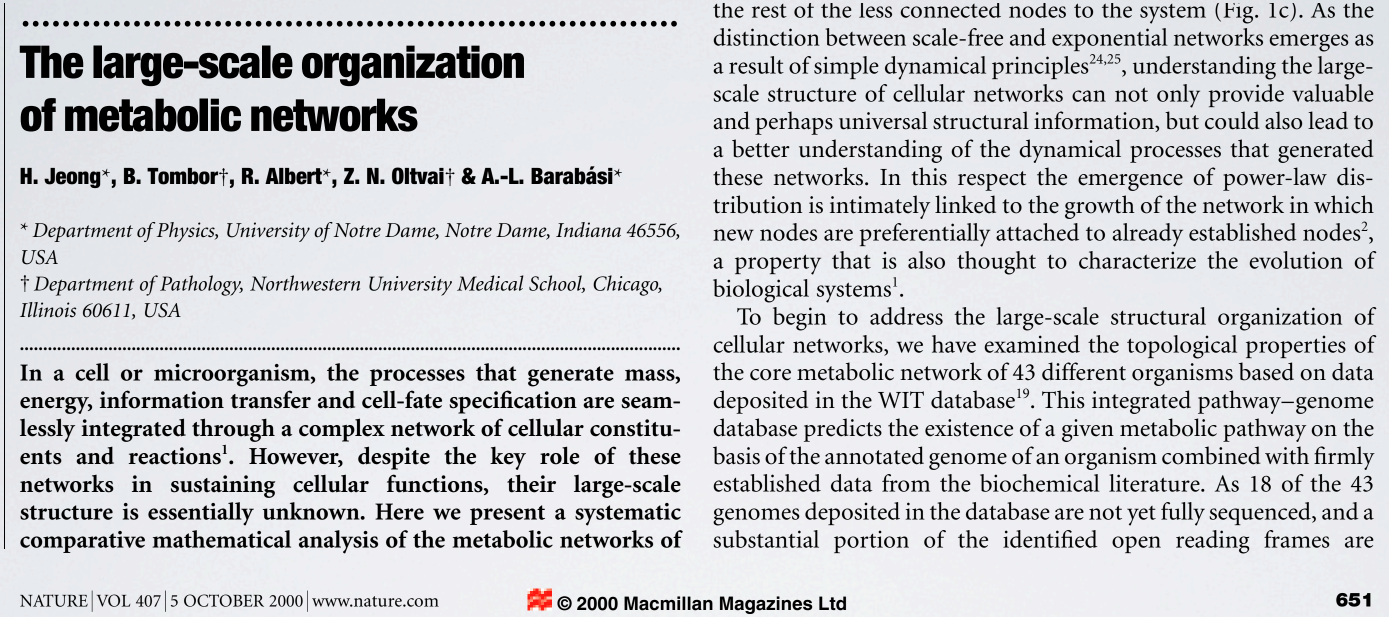
Despite its increasing role in communication, the World-Wide Web remains uncontrolled: any individual or institution can create a website with any number of documents and links. This unregulated growth leads to a huge and complex web, which becomes a large directed graph whose vertices are documents and whose edges are links (URLs) that point from one document to another. The topology of this graph determines the web's connectivity and consequently how effectively we can locate information on it. But its enormous size (estimated to be at least 8×10^8 documents[1]) and the continual changing of documents and links make it impossible to catalogue all the vertices and edges.

The extent of the challenge in obtaining a complete topological map of the web is illustrated by the limitations of the commercial search engines: Northern Light, the search engine with the largest coverage, is estimated to index only 38% of the web[1]. Although much work has been done to map and characterize the Internet's infrastructure[2], little is known about what really matters in the search for information — the topology of the web. Here we take a step towards filling this gap: we have used local connectivity measurements to construct a topological model of the World-Wide Web, which has enabled us to explore and char-

Figure 1 Distribution of links on the World-Wide Web. **a,** Outgoing links (URLs found on an HTML document); **b,** incoming links (URLs pointing to a certain HTML document). Data were obtained from the complete map of the nd.edu domain, which contains 325,729 documents and 1,469,680 links. Dotted lines represent analytical fits used as input distributions in constructing the topological model of the web; the tail of the distributions follows $P(k) \approx k^{-\gamma}$, with $\gamma_{out} = 2.45$ and $\gamma_{in} = 2.1$. **c,** Average of the shortest path between two documents as a function of system size, as predicted by the model. To check the validity of our predictions, we determined d for documents in the domain nd.edu. The measured $\langle d_{nd.edu} \rangle = 11.2$ agrees well with the prediction $\langle d_{3\times10^5} \rangle = 11.6$ obtained from our model. To show that the power-law tail of $P(k)$ is a universal feature of the web, the inset shows $P_{out}(k)$ obtained by starting from whitehouse.gov (squares), yahoo.com (triangles) and snu.ac.kr (inverted triangles). The slope of the dashed line is $\gamma_{out} = 2.45$, as obtained from nd.edu in **a**.

**Réka Albert, Hawoong Jeong,
Albert-László Barabási**
*Department of Physics, University of Notre Dame,
Notre Dame, Indiana 46556, USA
e-mail:alb@nd.edu*

1. Lawrence, S. & Giles, C. L. *Nature* **400,** 107–109 (1999).
2. Claffy, K., Monk, T. E. & McRobb, D. Internet tomography. *Nature* [online] <http://helix.nature.com/webmatters/tomog/tomog.html> (1999).
3. Erdös, P. & Rényi, A. *Publ. Math. Inst. Hung. Acad. Sci.* **5,** 17–61 (1960).
4. Bollobás, B. *Random Graphs* (Academic, London, 1985).

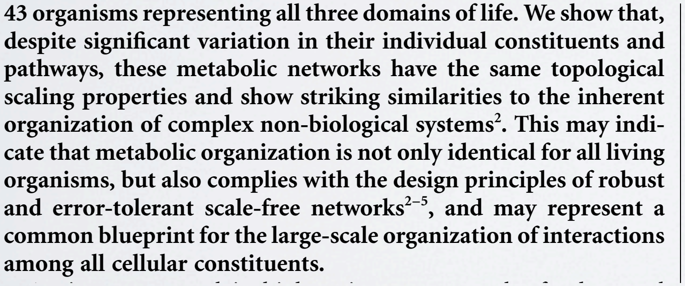


The large-scale organization of metabolic networks

H. Jeong*, B. Tombor†, R. Albert*, Z. N. Oltvai† & A.-L. Barabási*

* *Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA*
† *Department of Pathology, Northwestern University Medical School, Chicago, Illinois 60611, USA*

In a cell or microorganism, the processes that generate mass, energy, information transfer and cell-fate specification are seamlessly integrated through a complex network of cellular constituents and reactions[1]. However, despite the key role of these networks in sustaining cellular functions, their large-scale structure is essentially unknown. Here we present a systematic comparative mathematical analysis of the metabolic networks of 43 organisms representing all three domains of life. We show that, despite significant variation in their individual constituents and pathways, these metabolic networks have the same topological scaling properties and show striking similarities to the inherent organization of complex non-biological systems[2]. This may indicate that metabolic organization is not only identical for all living organisms, but also complies with the design principles of robust and error-tolerant scale-free networks[2–5], and may represent a common blueprint for the large-scale organization of interactions among all cellular constituents.

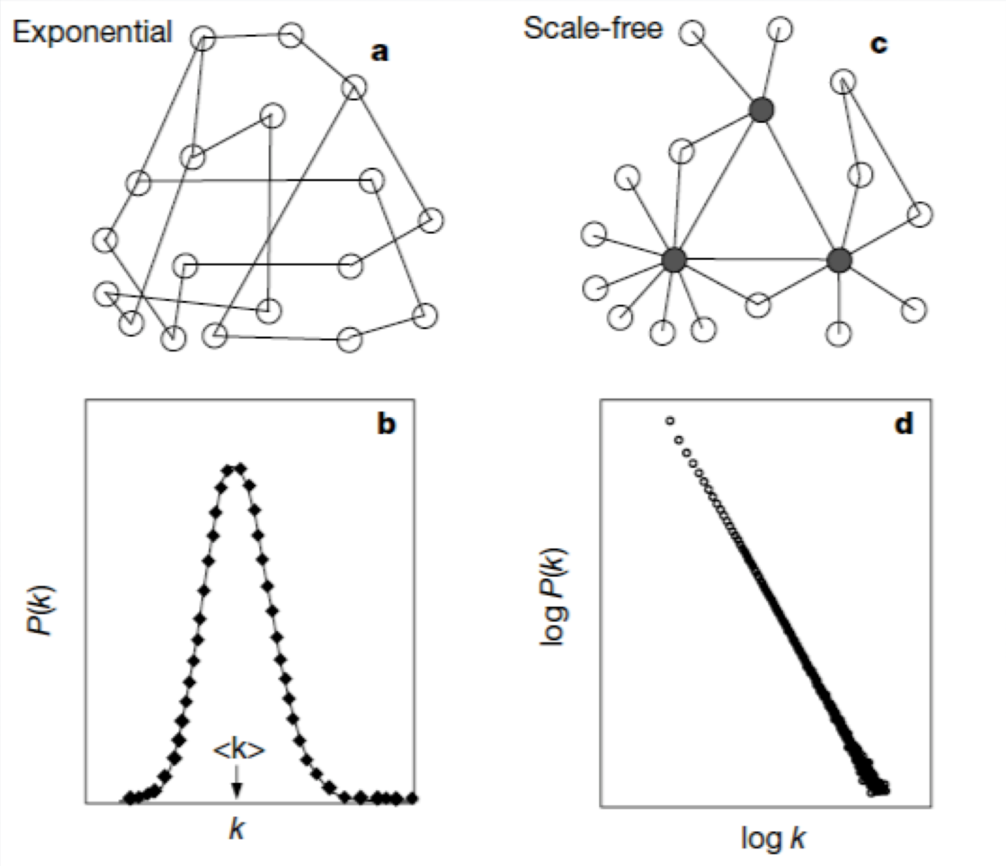


the rest of the less connected nodes to the system (Fig. 1c). As the distinction between scale-free and exponential networks emerges as a result of simple dynamical principles[24,25], understanding the large-scale structure of cellular networks can not only provide valuable and perhaps universal structural information, but could also lead to a better understanding of the dynamical processes that generated these networks. In this respect the emergence of power-law distribution is intimately linked to the growth of the network in which new nodes are preferentially attached to already established nodes[2], a property that is also thought to characterize the evolution of biological systems[1].

To begin to address the large-scale structural organization of cellular networks, we have examined the topological properties of the core metabolic network of 43 different organisms based on data deposited in the WIT database[19]. This integrated pathway–genome database predicts the existence of a given metabolic pathway on the basis of the annotated genome of an organism combined with firmly established data from the biochemical literature. As 18 of the 43 genomes deposited in the database are not yet fully sequenced, and a substantial portion of the identified open reading frames are

Random network

Scale-free network



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

y= degree exponent

Generalmente molto omogenei

La connettivita' dei nodi segue la distribuzione di Poisson.

La probabilita' di trovare nodi fortemente connesi (con molti links) decade in modo **esponenziale.**
In genere k = 2 o 3

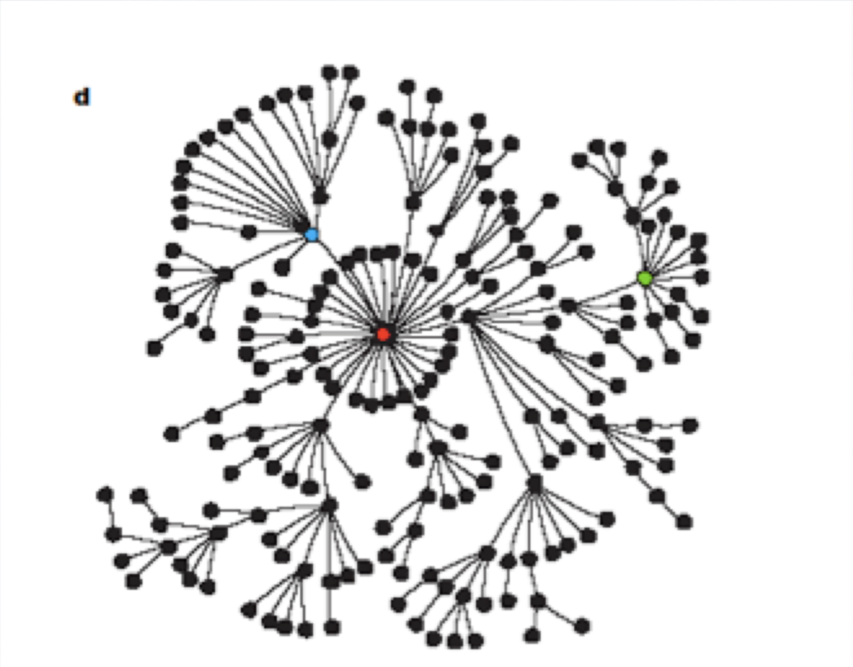
Generalmente molto eterogenei

Hanno topologia dominata da pochi nodi fortemente connessi (hubs) a tutti gli latri nodi poco connessi

La probabilita' di trovare nodi fortemente connesi (con molti links) segue un andamento logaritmico
(power law)

1. Degree distribution e hubs

Scale-free network
(Barabasi)



Party hubs

Date hubs

OMIM

Online Mendelian Inheritance in Man

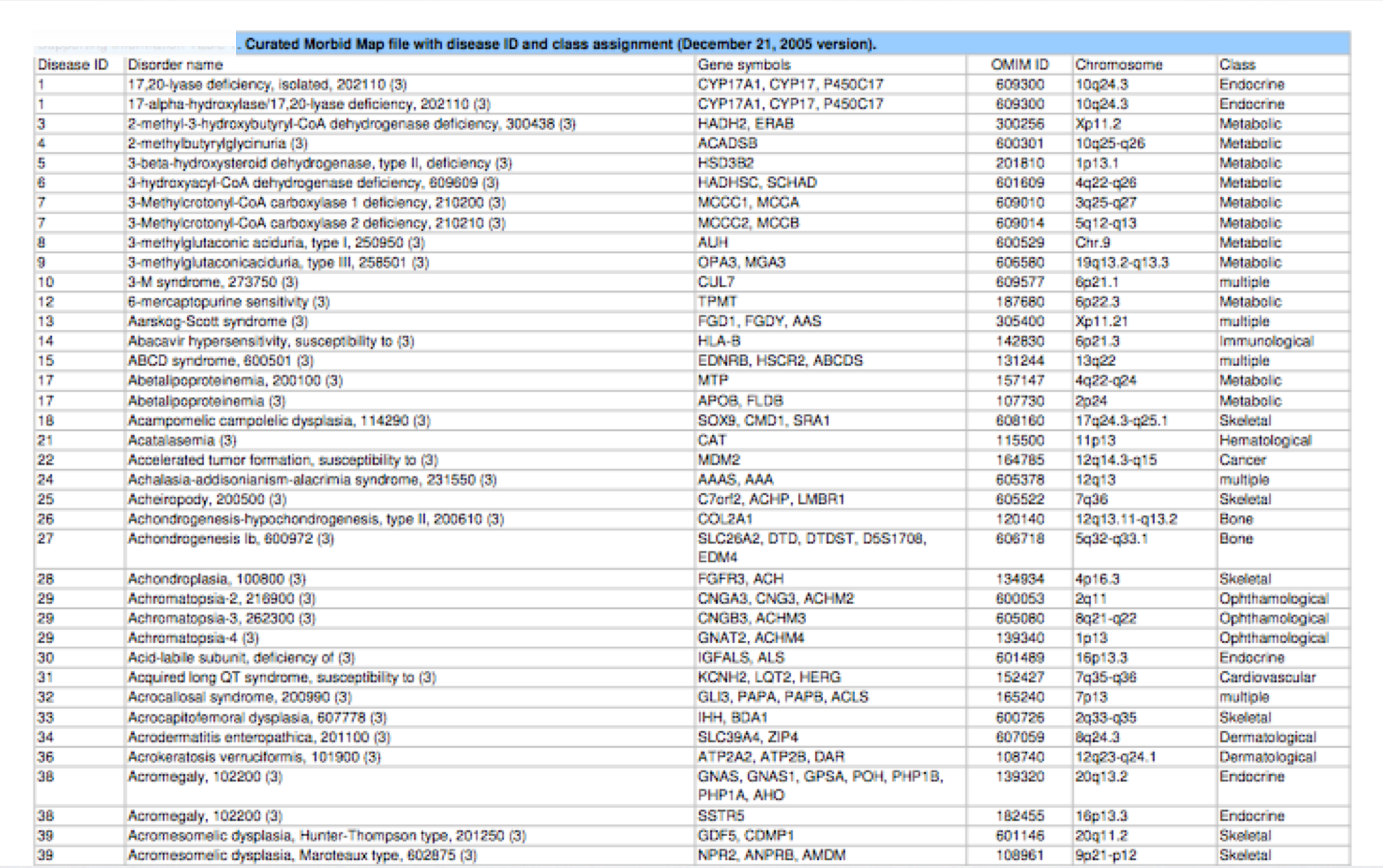
Johns
Hopkins
University

OMIM is a comprehensive, authoritative, and timely compendium of human **genes** and **genetic phenotypes**. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources.

On line ("O") catalog of mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM).

http://www.ncbi.nlm.nih.gov/omim

OMIM database



. Curated Morbid Map file with disease ID and class assignment (December 21, 2005 version).

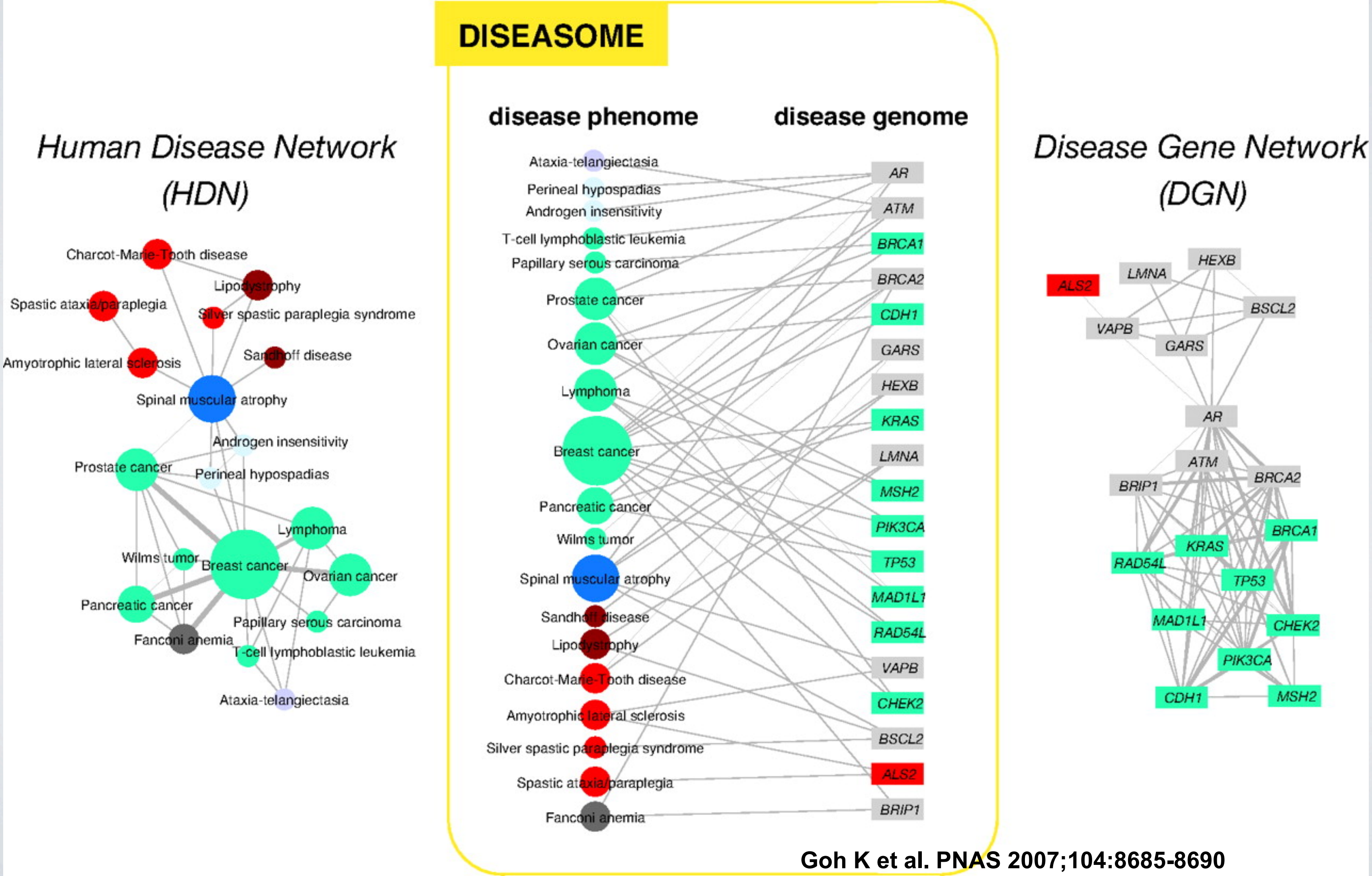
Disease ID	Disorder name	Gene symbols	OMIM ID	Chromosome	Class
1	17,20-lyase deficiency, isolated, 202110 (3)	CYP17A1, CYP17, P450C17	609300	10q24.3	Endocrine
1	17-alpha-hydroxylase/17,20-lyase deficiency, 202110 (3)	CYP17A1, CYP17, P450C17	609300	10q24.3	Endocrine
3	2-methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency, 300438 (3)	HADH2, ERAB	300256	Xp11.2	Metabolic
4	2-methylbutyrylglycinuria (3)	ACADSB	600301	10q25-q26	Metabolic
5	3-beta-hydroxysteroid dehydrogenase, type II, deficiency (3)	HSD3B2	201810	1p13.1	Metabolic
6	3-hydroxyacyl-CoA dehydrogenase deficiency, 609609 (3)	HADHSC, SCHAD	601609	4q22-q26	Metabolic
7	3-Methylcrotonyl-CoA carboxylase 1 deficiency, 210200 (3)	MCCC1, MCCA	609010	3q25-q27	Metabolic
7	3-Methylcrotonyl-CoA carboxylase 2 deficiency, 210210 (3)	MCCC2, MCCB	609014	5q12-q13	Metabolic
8	3-methylglutaconic aciduria, type I, 250950 (3)	AUH	600529	Chr.9	Metabolic
9	3-methylglutaconicaciduria, type III, 258501 (3)	OPA3, MGA3	606580	19q13.2-q13.3	Metabolic
10	3-M syndrome, 273750 (3)	CUL7	609577	6p21.1	multiple
12	6-mercaptopurine sensitivity (3)	TPMT	187680	6p22.3	Metabolic
13	Aarskog-Scott syndrome (3)	FGD1, FGDY, AAS	305400	Xp11.21	multiple
14	Abacavir hypersensitivity, susceptibility to (3)	HLA-B	142830	6p21.3	Immunological
15	ABCD syndrome, 600501 (3)	EDNRB, HSCR2, ABCDS	131244	13q22	multiple
17	Abetalipoproteinemia, 200100 (3)	MTP	157147	4q22-q24	Metabolic
17	Abetalipoproteinemia (3)	APOB, FLDB	107730	2p24	Metabolic
18	Acampomelic campolelic dysplasia, 114290 (3)	SOX9, CMD1, SRA1	608160	17q24.3-q25.1	Skeletal
21	Acatalasemia (3)	CAT	115500	11p13	Hematological
22	Accelerated tumor formation, susceptibility to (3)	MDM2	164785	12q14.3-q15	Cancer
24	Achalasia-addisonianism-alacrimia syndrome, 231550 (3)	AAAS, AAA	605378	12q13	multiple
25	Acheiropody, 200500 (3)	C7orf2, ACHP, LMBR1	605522	7q36	Skeletal
26	Achondrogenesis-hypochondrogenesis, type II, 200610 (3)	COL2A1	120140	12q13.11-q13.2	Bone
27	Achondrogenesis Ib, 600972 (3)	SLC26A2, DTD, DTDST, D5S1708, EDM4	606718	5q32-q33.1	Bone
28	Achondroplasia, 100800 (3)	FGFR3, ACH	134934	4p16.3	Skeletal
29	Achromatopsia-2, 216900 (3)	CNGA3, CNG3, ACHM2	600053	2q11	Ophthamological
29	Achromatopsia-3, 262300 (3)	CNGB3, ACHM3	605080	8q21-q22	Ophthamological
29	Achromatopsia-4 (3)	GNAT2, ACHM4	139340	1p13	Ophthamological
30	Acid-labile subunit, deficiency of (3)	IGFALS, ALS	601489	16p13.3	Endocrine
31	Acquired long QT syndrome, susceptibility to (3)	KCNH2, LQT2, HERG	152427	7q35-q36	Cardiovascular
32	Acrocallosal syndrome, 200990 (3)	GLI3, PAPA, PAPB, ACLS	165240	7p13	multiple
33	Acrocapitofemoral dysplasia, 607778 (3)	IHH, BDA1	600726	2q33-q35	Skeletal
34	Acrodermatitis enteropathica, 201100 (3)	SLC39A4, ZIP4	607059	8q24.3	Dermatological
36	Acrokeratosis verruciformis, 101900 (3)	ATP2A2, ATP2B, DAR	108740	12q23-q24.1	Dermatological
38	Acromegaly, 102200 (3)	GNAS, GNAS1, GPSA, POH, PHP1B, PHP1A, AHO	139320	20q13.2	Endocrine
38	Acromegaly, 102200 (3)	SSTR5	182455	16p13.3	Endocrine
39	Acromesomelic dysplasia, Hunter-Thompson type, 201250 (3)	GDF5, CDMP1	601146	20q11.2	Skeletal
39	Acromesomelic dysplasia, Maroteaux type, 602875 (3)	NPR2, ANPRB, AMDM	108961	9p21-p12	Skeletal

Dicembre 2005: 1,284 malattie “genetiche” e 1,777 geni implicati nelle malattie
Classificazione “manually-curated” in 22 classi di malattie

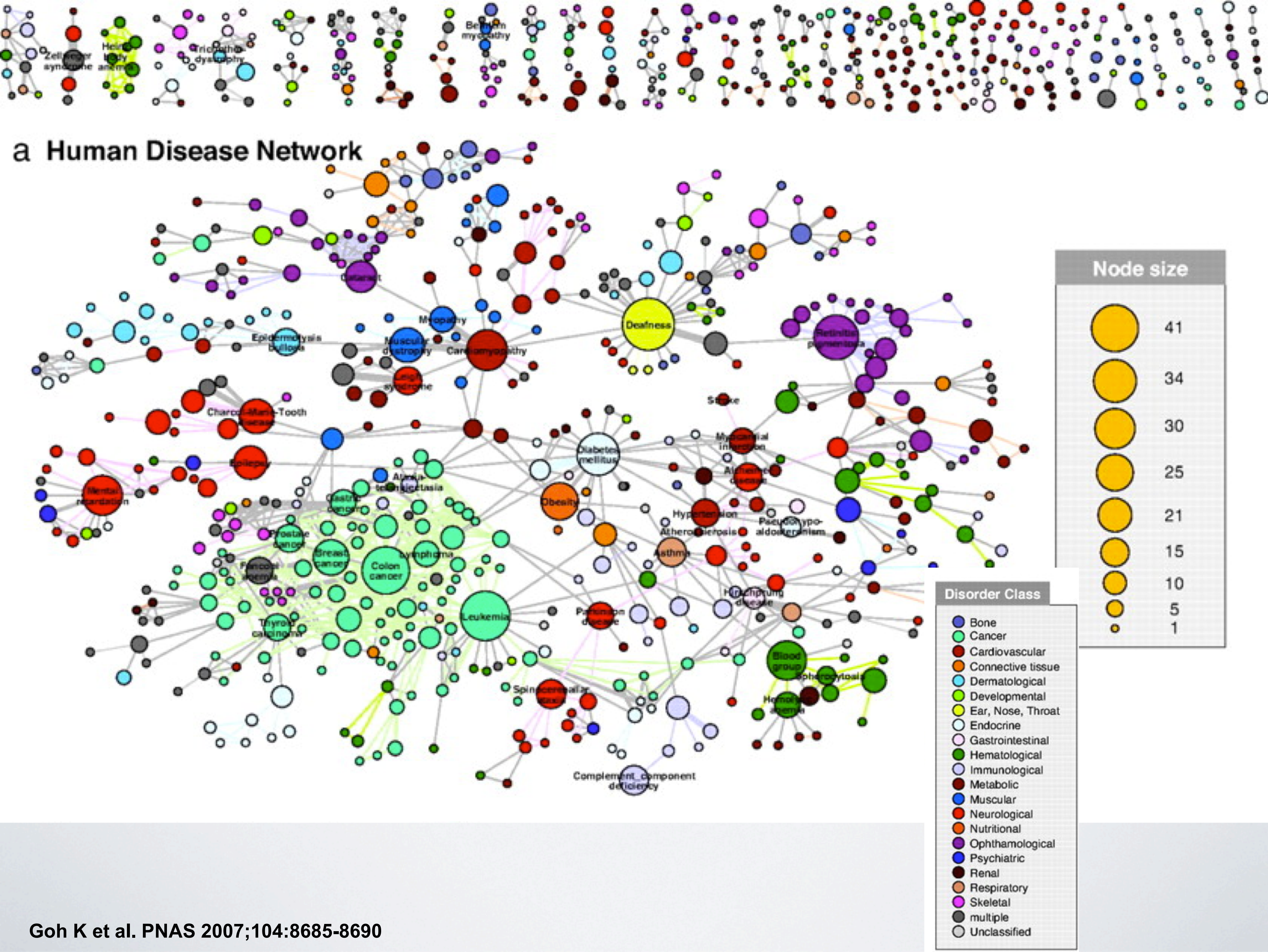
Construzione del diseasoma

HDN: Due malattie sono connesse se hanno in comune almeno un gene mutato

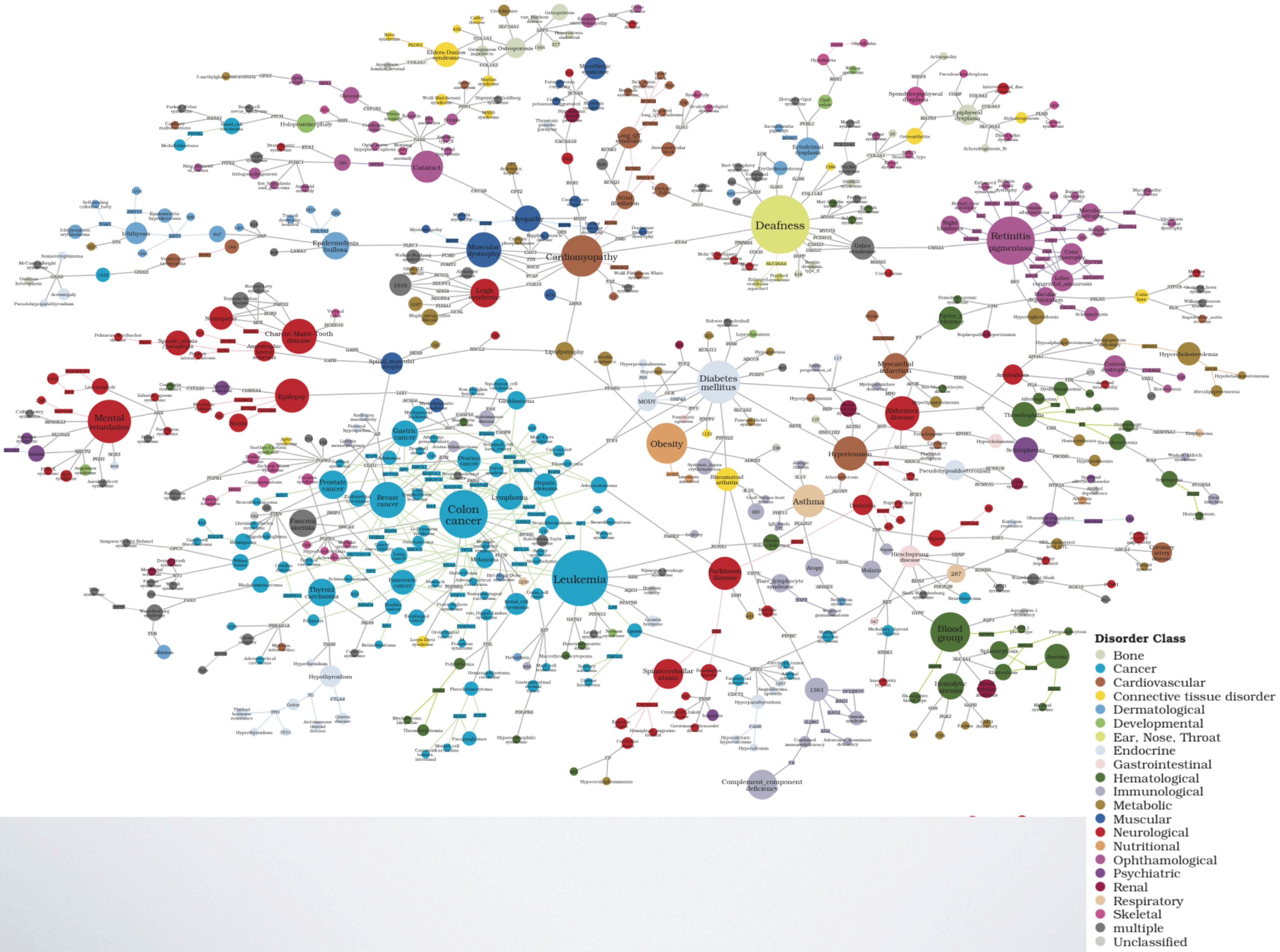
DGN: due geni sono interconnessi se sono coinvolti nella stessa malattia



Goh K et al. PNAS 2007;104:8685-8690



Goh K et al. PNAS 2007;104:8685-8690

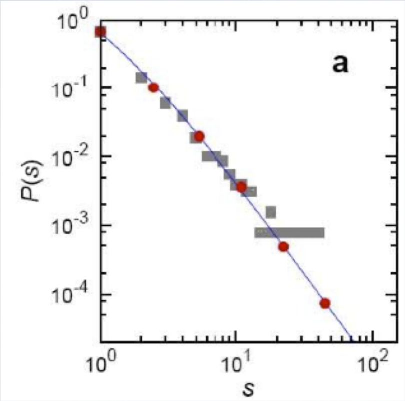

Ehlers-Danlos syndrome
Osteoporosis
Cataract
Holoprosencephaly
Ichthyosis
Epidermolysis bullosa
Myasthenic syndrome
Long QT syndrome
Atrial fibrillation
Myopathy
Muscular dystrophy
Cardiomyopathy
Leigh syndrome
1016
Deafness
Usher syndrome
Ectodermal dysplasia
Spondyloepiphyseal dysplasia
Epiphyseal dysplasia
Retinitis pigmentosa
Macular dystrophy
Cone dystrophy
Leber congenital amaurosis
Macular degeneration
Neuropathy
Charcot-Marie-Tooth disease
Spastic ataxia/paraplegia
Amyotrophic lateral sclerosis
Spinal muscular atrophy
Lipodystrophy
Diabetes mellitus
MODY
Obesity
Myocardial infarction
Hypercholesterolemia
Corneal dystrophy
Amyloidosis
Alzheimer disease
Thrombophilia
Hypertension
Schizophrenia
Mental retardation
Epilepsy
Ataxia
Glioblastoma
Gastric cancer
Ovarian cancer
Prostate cancer
Breast cancer
Colon cancer
Lymphoma
Hepatic adenoma
Fanconi anemia
Melanoma
Leukemia
Pancreatic cancer
Thyroid carcinoma
Rheumatoid arthritis
Asthma
Dementia
Parkinson disease
Atopy
Malaria
Hirschsprung disease
287
Blood group
Spherocytosis
Anemia
Hemolytic anemia
Spinocerebellar ataxia
1383
Complement_component deficiency
Hypothyroidism
Coronary artery disease
Pseudohypoaldosteronism
Disorder Class
● Bone
● Cancer
● Cardiovascular
● Connective tissue disorder
● Dermatological
● Developmental
● Ear, Nose, Throat
● Endocrine
● Gastrointestinal
● Hematological
● Immunological
● Metabolic
● Muscular
● Neurological
● Nutritional
● Ophthamological
● Psychiatric
● Renal
● Respiratory
● Skeletal
● multiple
● Unclassified

The properties of disease network (HDN)

1. Diseases and disease classes are strongly interconnected) (from 1284 diseases, 867 have at least one link)

2. 516 diseases form a giant cluster, suggesting shared genetical bases for numerous diseases

The properties of disease network

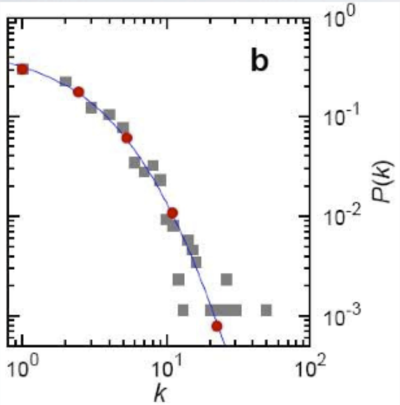


Distribution of the size (s) in the HDN

s = number of genes associated to a disease

- Large distribution
- The majority of diseases have few genes involved
- deafness (*s*=41), leukemia (*s*=37); colon cancer (s=34)

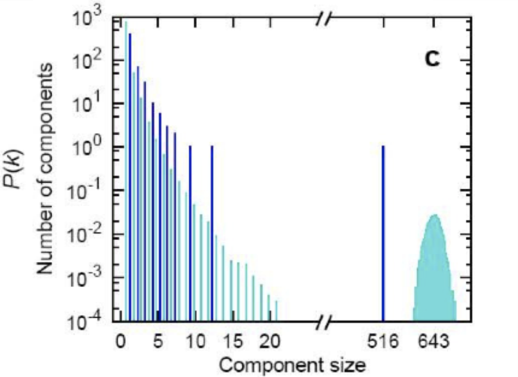
P(s) = probability distribution



Distribution of *K* = degree

- The majority of diseases has few links
- **hubs**: colon cancer (k=50); breat cancer (k= 30)
- most comnnected nodes are specific for oncologic diseases (especially through *TP53* e *PTEN genes*)

$f(x) = c(x + a)^b$



Distribution of the cluster sizes in the HDN

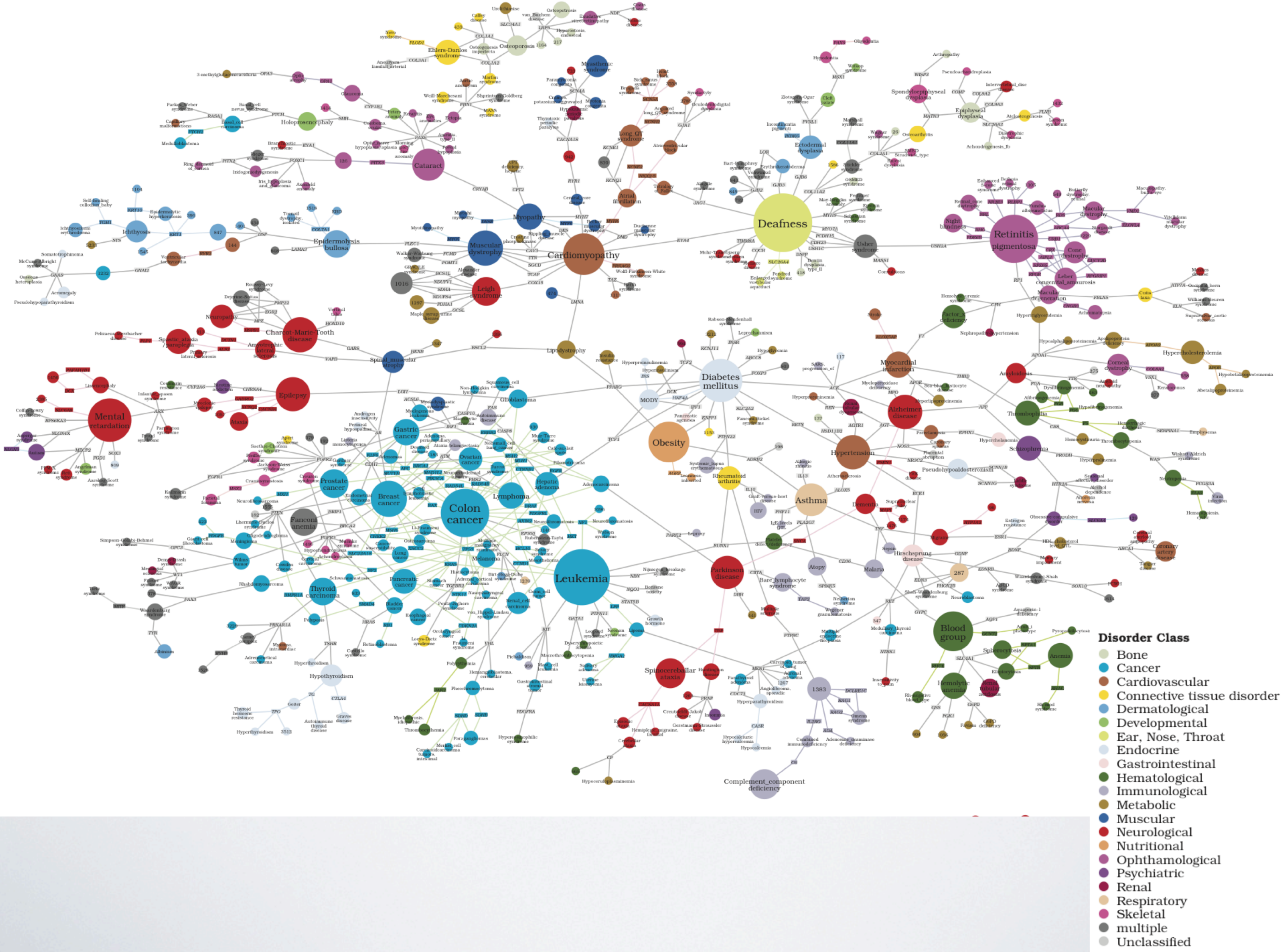
- the isolated peak at 516 corresponds to the size of the giant component

DATA BINNING IS A DATA PRE-PROCESSING TECHNIQUE USED TO REDUCE THE EFFECTS OF MINOR OBSERVATION ERRORS.

THE ORIGINAL DATA VALUES WHICH FALL IN A GIVEN SMALL INTERVAL, A BIN, ARE REPLACED BY A VALUE REPRESENTATIVE OF THAT INTERVAL, OFTEN THE CENTRAL VALUE.

QUANTIZATION IS THE PROCEDURE OF CONSTRAINING SOMETHING FROM A RELATIVELY LARGE OR CONTINUOUS SET OF VALUES (SUCH AS THE REAL NUMBERS) TO A RELATIVELY SMALL DISCRETE SET.

Data binning IS A FORM OF QUANTIZATION = QUANTIZZAZIONE IN ITALIANO


Deafness
Retinitis pigmentosa
Cardiomyopathy
Muscular dystrophy
Myopathy
Cataract
Ehlers-Danlos syndrome
Osteoporosis
Holoprosencephaly
Ichthyosis
Epidermolysis bullosa
Leigh syndrome
Charcot-Marie-Tooth disease
Spinal muscular atrophy
Lipodystrophy
Mental retardation
Epilepsy
Diabetes mellitus
MODY
Obesity
Colon cancer
Breast cancer
Prostate cancer
Gastric cancer
Ovarian cancer
Lymphoma
Fanconi anemia
Thyroid carcinoma
Pancreatic cancer
Leukemia
Melanoma
Glioblastoma
Hepatic adenoma
Asthma
Rheumatoid arthritis
Hypertension
Alzheimer disease
Myocardial infarction
Parkinson disease
Spinocerebellar ataxia
Blood group
Hemolytic anemia
Anemia
Schizophrenia
Macular degeneration
Hypercholesterolemia
Coronary artery disease
Hypothyroidism
Complement_component deficiency
Usher syndrome
Long_QT syndrome
Atrial fibrillation
Disorder Class
● Bone
● Cancer
● Cardiovascular
● Connective tissue disorder
● Dermatological
● Developmental
● Ear, Nose, Throat
● Endocrine
● Gastrointestinal
● Hematological
● Immunological
● Metabolic
● Muscular
● Neurological
● Nutritional
● Ophthamological
● Psychiatric
● Renal
● Respiratory
● Skeletal
● multiple
● Unclassified

The properties of disease network

The network shows connections between diseases and between disease classes

Clusters tend to be formed based on disease class

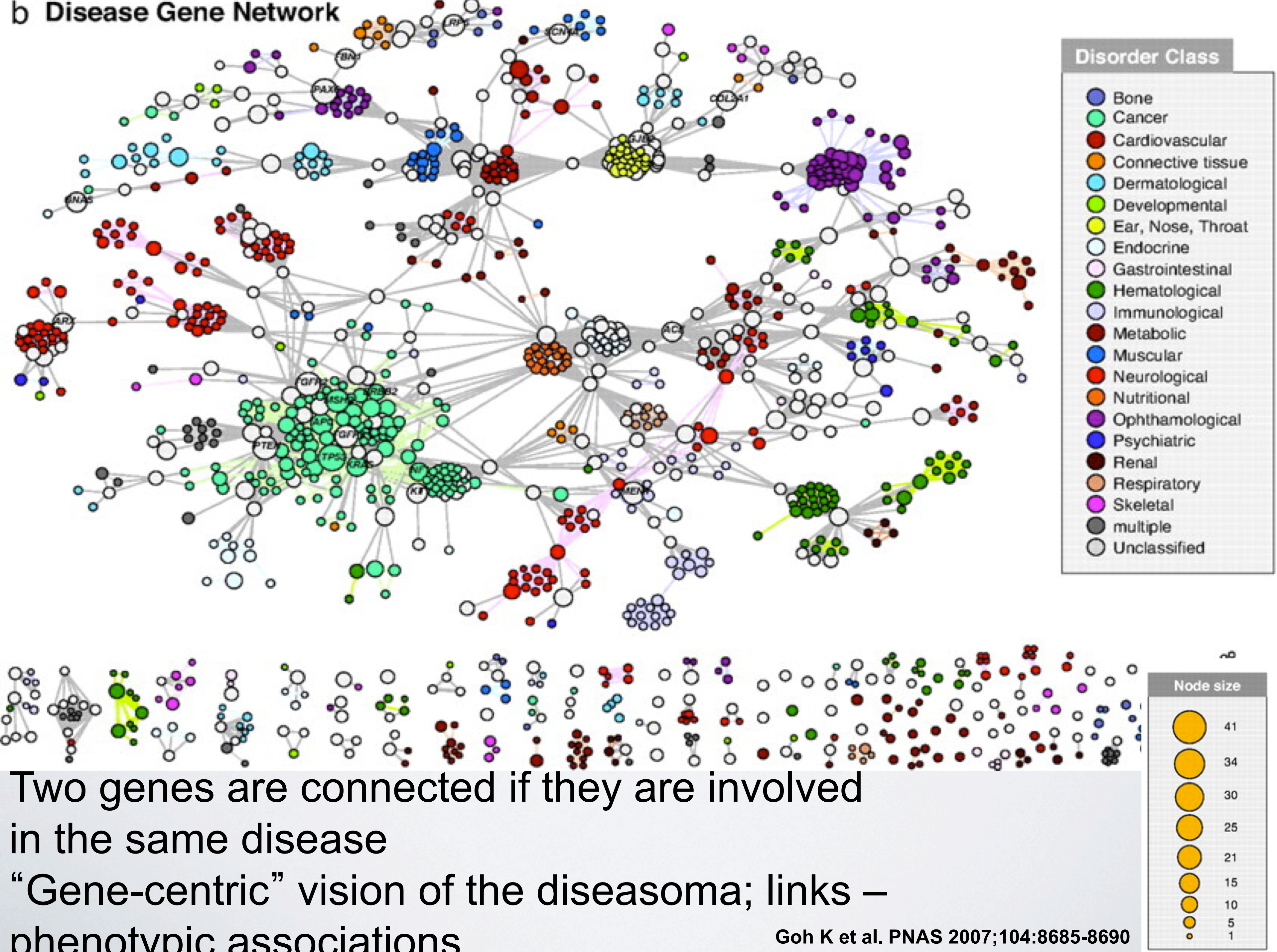
The cluster of cancer is highly interconnected (several genes are associated with many types of cancers: *TP53, KRAS, ERBB2, NF1*)

The cluster of cancer includes also diseases such as ataxia-teleangiectasia and Fanconi anemia

Metabolic diseases do not form a single cluster, but are distributed between other clusters

Oncologic diseases and neurologic diseases have highest locus heterogeneity and are the most interconnected.

(LOCUS HETEROGENEITY: same disease phenotype is generated by mutations in different genes)



Two genes are connected if they are involved in the same disease

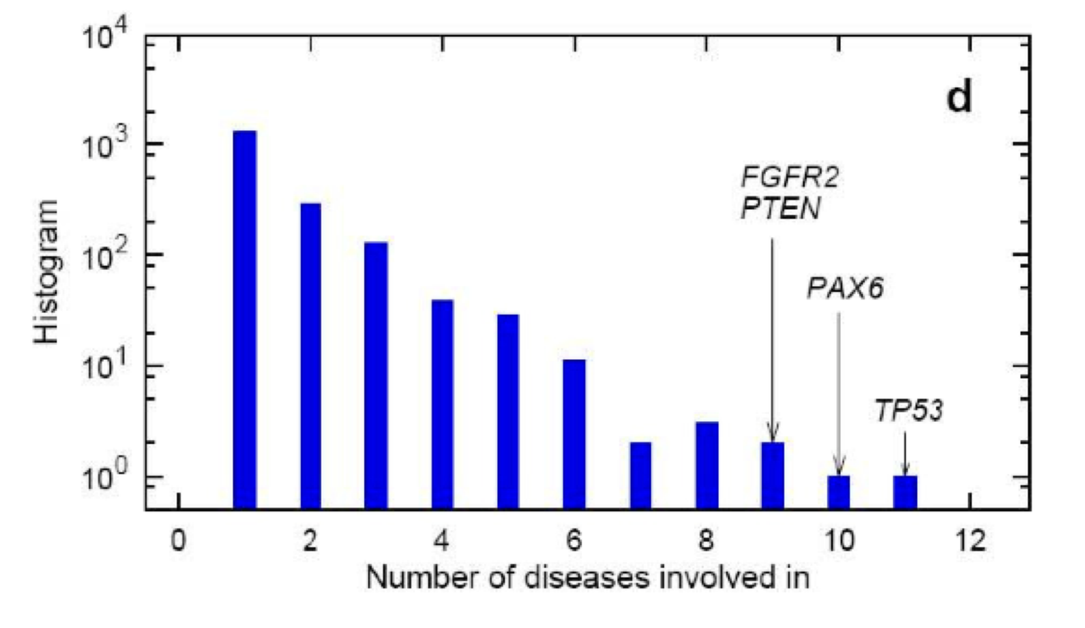
“Gene-centric” vision of the diseasoma; links – phenotypic associations

Goh K et al. PNAS 2007;104:8685-8690

Le proprietà della rete di disease genes

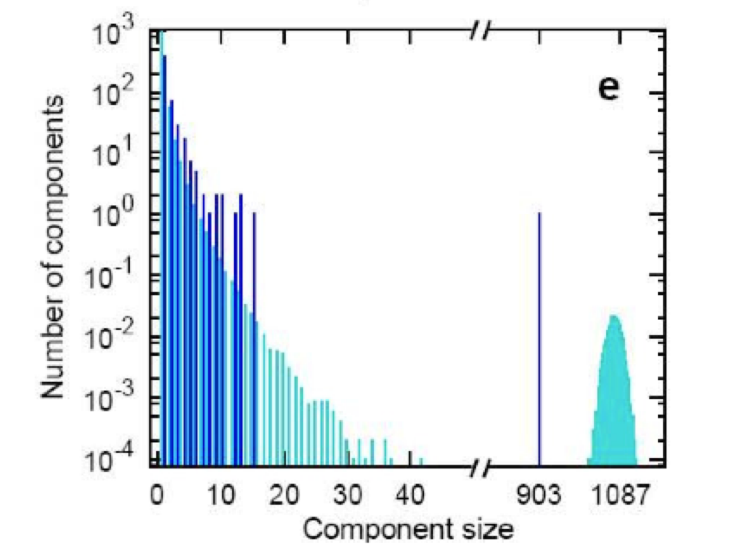
1. Disease genes are strongly interconnected I (from **1777** genes considered, 1377 have at least one link)
2. 903 genes form a giant cluster, suggesting common genetic bases for many diseases

The properties of disease gene network

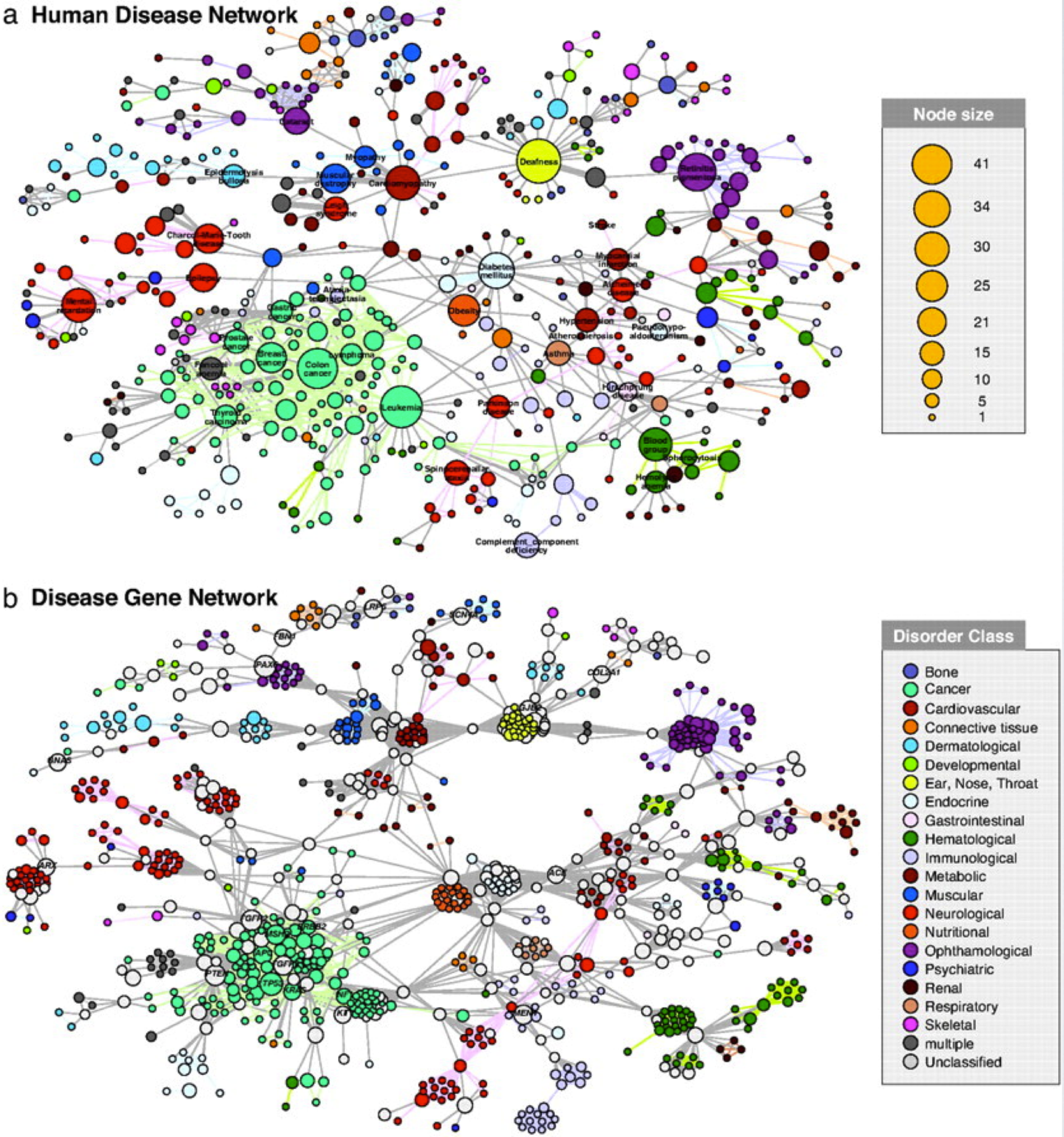


Histogram of the number of disorders a gene is involved in

- The majority of disease genes are involved in few diseases
- The 4 genes involved in more diseases (major hubs) are shown



Distribution of the connected component sizes in the DGN (blue) versus randomized network (light blue)


a Human Disease Network
Cataract
Myopathy
Epidermolysis bullosa
Muscular dystrophy
Cardiomyopathy
Leigh syndrome
Deafness
Retinitis pigmentosa
Charcot-Marie-Tooth disease
Stroke
Myocardial infarction
Diabetes mellitus
Alzheimer disease
Epilepsy
Mental retardation
Ataxia-telangiectasia
Gastric cancer
Obesity
Hypertension
Pseudohypo-aldosteronism
Atherosclerosis
Prostate cancer
Breast cancer
Colon cancer
Lymphoma
Asthma
Fanconi anemia
Hirschsprung disease
Parkinson disease
Thyroid carcinoma
Leukemia
Blood group
Spherocytosis
Spinocerebellar ataxia
Hemolytic anemia
Complement_component deficiency
Node size
41
34
30
25
21
15
10
5
1
b Disease Gene Network
LRP5
SCN4A
FBN1
PAX6
COL2A1
GJB2
GNAS
ARX
FGFR3
MSH2
ERBB2
APC
PTEN
TP53
KRAS
NF1
KIT
MEN1
Disorder Class
● Bone
● Cancer
● Cardiovascular
● Connective tissue
● Dermatological
● Developmental
● Ear, Nose, Throat
● Endocrine
● Gastrointestinal
● Hematological
● Immunological
● Metabolic
● Muscular
● Neurological
● Nutritional
● Ophthamological
● Psychiatric
● Renal
● Respiratory
● Skeletal
● multiple
● Unclassified

The common charateristics of disease network and disease gene network

1. Clusterization of diseases and disease genes based on pathological phenotype

2. Diseases and disease genes tend to cluster based on common pathological processes

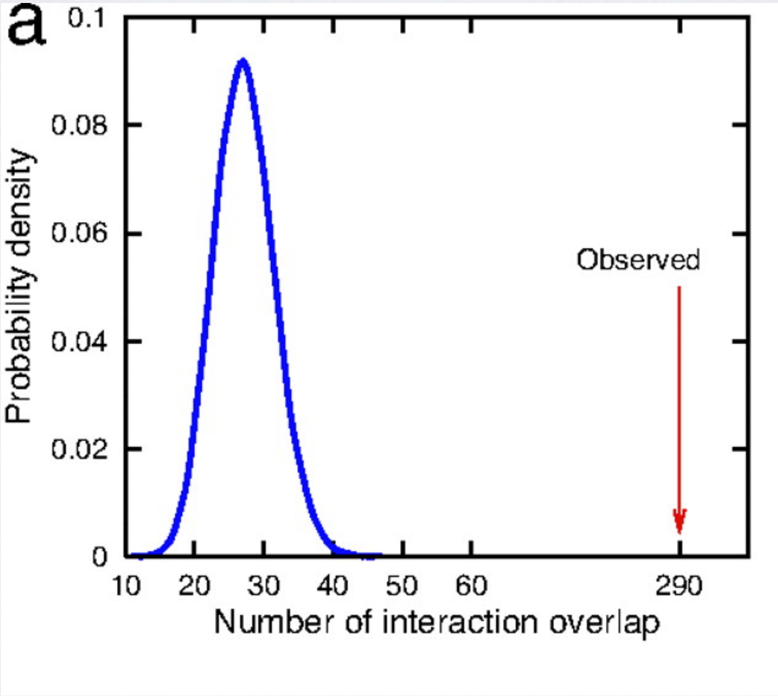
Ex.: in the disease network there are 812 links between diseases of the same class compared to 107 links in the randomized network (8-fold increase).

Does disease gene network correspond to the PPI network?

Hypothesis: Disease genes → proteins interacting in functional modules

Overlap between disease gene network and PPI network

Limiti dello studio: non conosciamo tutti geni coinvolti nelle malattie e non conosciamo tutte le interazioni proteiche



Number of observed physical interactions between the products of genes within the same disorder (red arrow) and the distribution of the expected number of interactions for the random control (blue) ($P < 10^{-6}$).

10-fold increase of interactions compared to the random control

Goh K et al. PNAS 2007;104:8685-8690

Una distribuzione di probabilità è, in sostanza, una funzione matematica che, per ogni valore della variabile, fornisce la probabilità che venga osservato quel valore.

La distribuzione di probabilità *continua*: il risultato cade in un certo intervallo finito di valori, compreso, ad esempio, fra *a* e *b*. Una tale probabilità, *P(a, b)* si esprime come un integrale:

$$P(a,b) = \int_a^b \phi(x)\, dx$$

$\phi(x)$ = densità di probabilità = probabilità che il risultato cada in un intervallo *infinitesimamente piccolo* attorno al valore *x* divisa per l'ampiezza di questo intervallo.

http://www.thch.unipg.it/~franc/i/node4.html

VALIDATION OF DISEASE GENE NETWORK

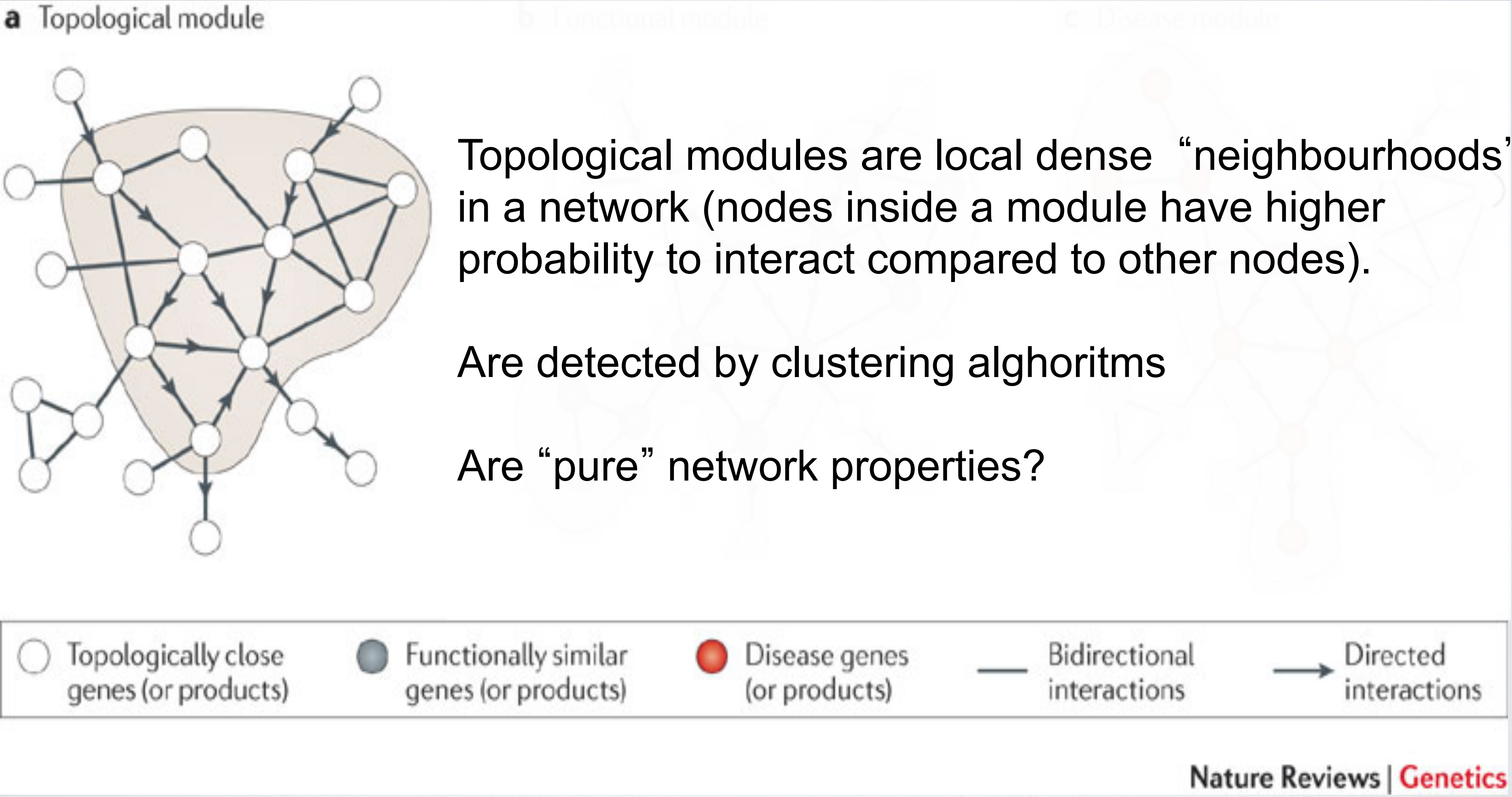
How we validate a network or a disease module?

Disease modules

- Cellular components associated with a specific disease tend to Cluster in the same network “neighbourhood”

- Each disease is characterized by a module (“neighbourhood”) Inside the interactome

2. Disease modules


a Topological module
b Functional module
c Disease module
○ Topologically close genes (or products)
● Functionally similar genes (or products)
● Disease genes (or products)
— Bidirectional interactions
→ Directed interactions
Nature Reviews | Genetics

Topological modules are local dense “neighbourhoods” in a network (nodes inside a module have higher probability to interact compared to other nodes).

Are detected by clustering alghoritms

Are “pure” network properties?

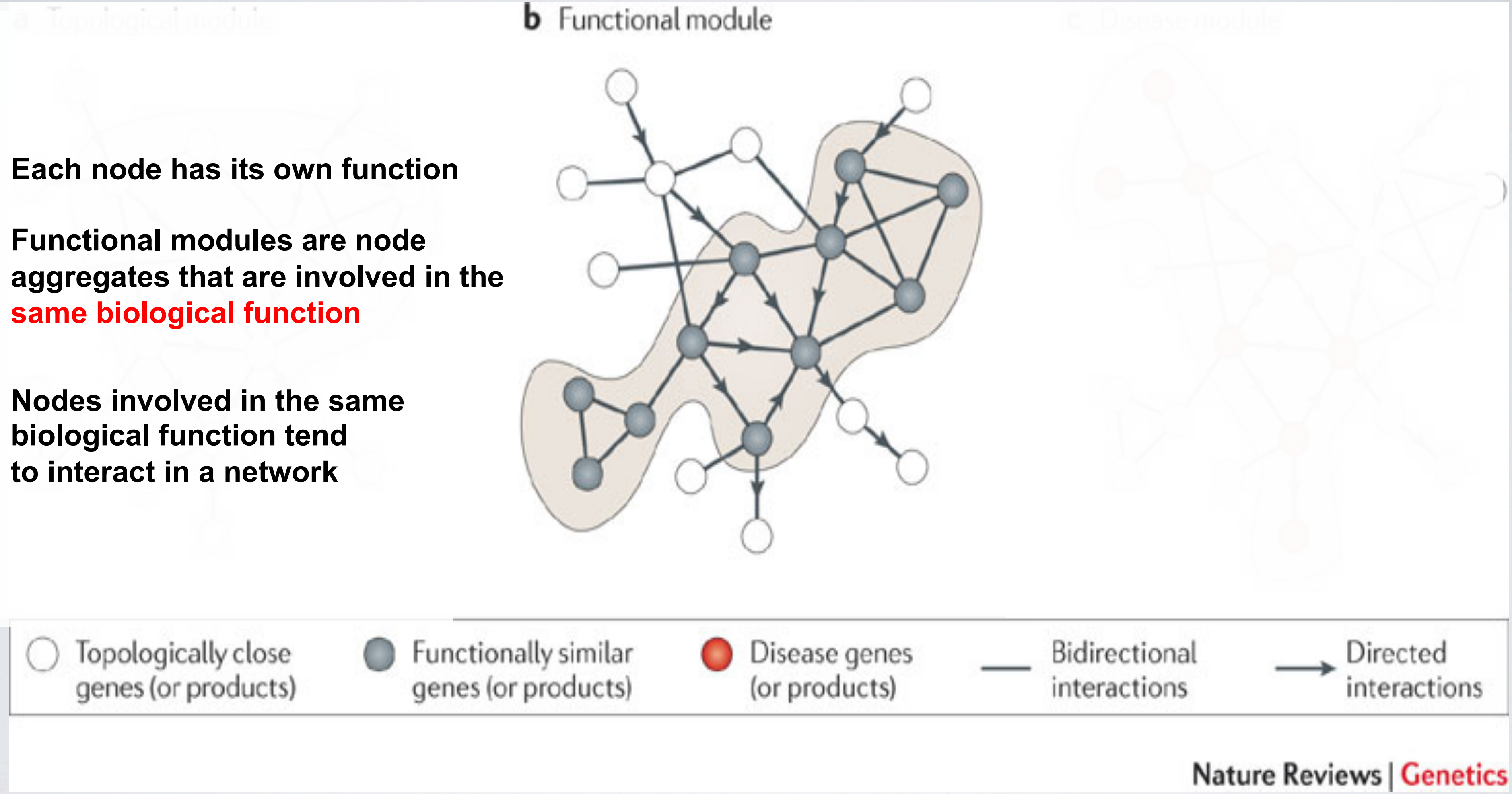
Barabasi et al., 2011

2. Disease modules

Each node has its own function

Functional modules are node aggregates that are involved in the same biological function

Nodes involved in the same biological function tend to interact in a network

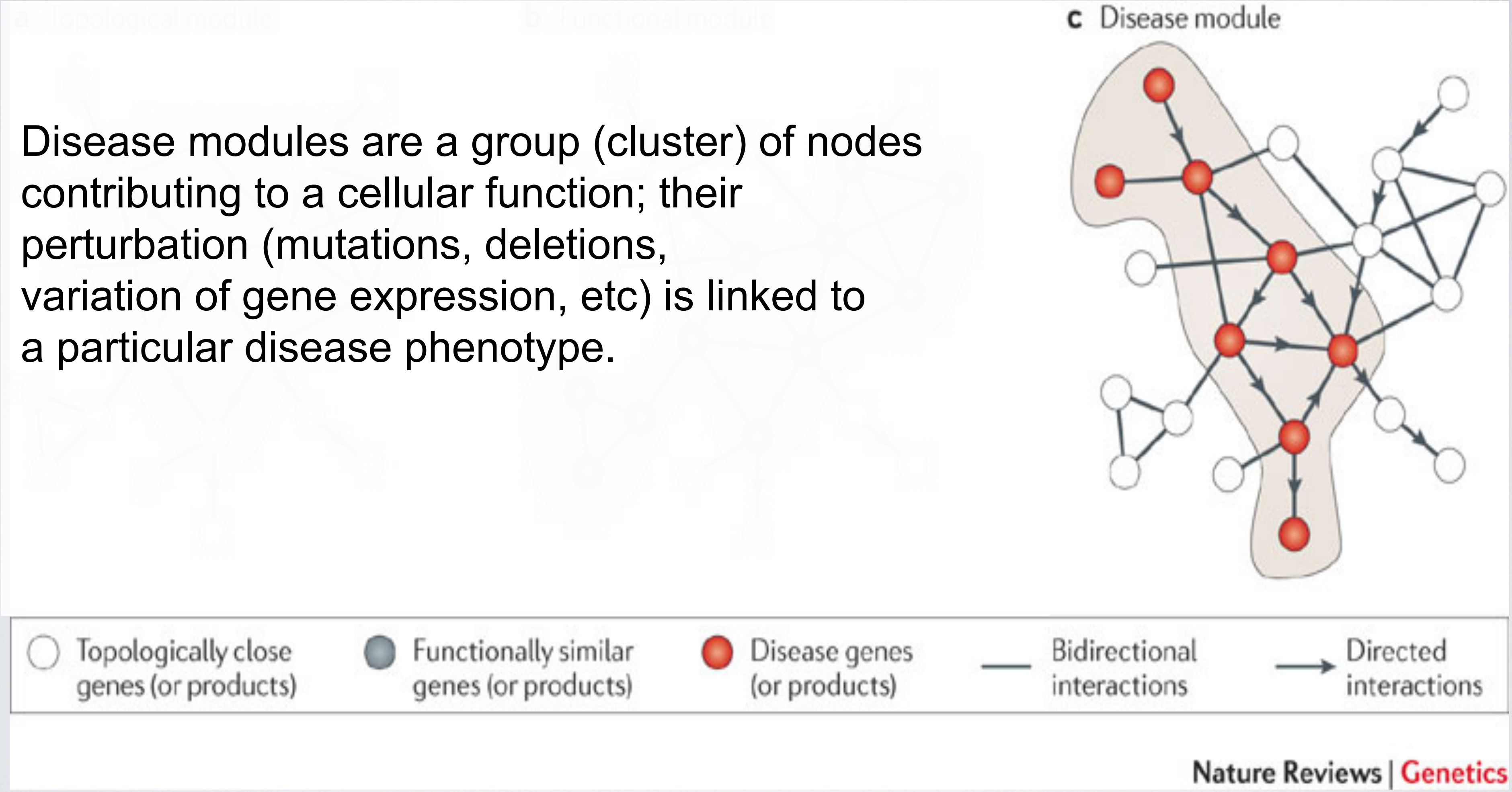

a Topological module
b Functional module
c Disease module
Topologically close genes (or products)
Functionally similar genes (or products)
Disease genes (or products)
Bidirectional interactions
Directed interactions

Nature Reviews | Genetics

Barabasi et al., 2011

2. Disease modules

Disease modules are a group (cluster) of nodes contributing to a cellular function; their perturbation (mutations, deletions, variation of gene expression, etc) is linked to a particular disease phenotype.


a Topological module
b Functional module
c Disease module
○ Topologically close genes (or products)
● Functionally similar genes (or products)
● Disease genes (or products)
— Bidirectional interactions
→ Directed interactions

Nature Reviews | Genetics

Barabasi et al., 2011

2. Disease modules

Hypotheses in network medicine:

- **Topological modules correspond with functional modules**

- **The disease is viewed as a perturbation of a functional module;**
Thus, functional modules (often) are also disease modules

- **Topological, functional and disease modules often OVERLAPP**

Disease module validation

Remember!!

Functional homogeneity

- **Gene ontology (**1) **biological processes**, 2) **cellular components** and 3) **molecular functions)**

Tissue specificity

Dynamic homogeneity

- **Co-expression**

Other tests:

- **GWA (SNPs identified for the nodes/genes)**

Predictions:

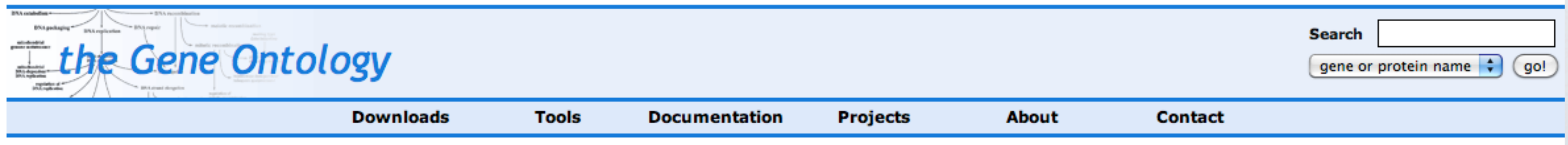
- **Disease genes**
- **Disease pathways**
- **Drug targets**

Ontologia = rappresentazione formale di una concettualizzazione di un dominio di interesse

= schema concettuale di classificazione, un glossario di base

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data.

The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated: 1) **biological processes**, 2) **cellular components** and 3) **molecular functions** in a species-independent manner.



http://www.geneontology.org/

Functional homogeneity of disease genes on Gene Ontology

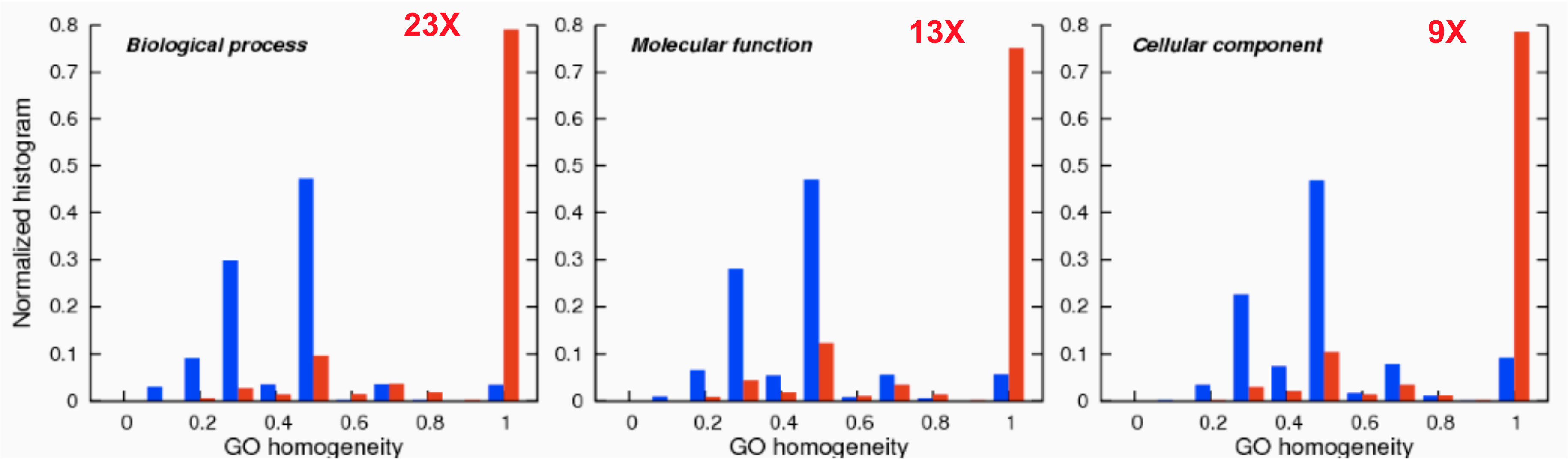
Hypothesis: the groups of genes associated to each disease share cellular and functional similarities (annotations in Gene Ontology)

Measure GO homogeneity (GH) for each disease= maximum fraction of genes in the same disorder that have the same GO terms

$GH_i = \max_j [n^j_i/n_i]$,

n_i = numero di geni nella malattia “i” che ha qualsiasi annotazione GO (number of genes in the disorder i that have any GO annotations,)
n^j_i = numero di geni che ha uno specifico termine j in GO (number of genes that have the specific GO term j)

«To obtain the random control of the GO homogeneity distribution for each disorder we picked the same number of genes randomly in the GO annotation data and calculated their GO homogeneity. 10^4 random instances were generated to reach statistical significance.»



Tissue-specific distribution of disease genes

Consideration: the proteins encoded by disease genes that interact in the same functional module tend to be expressed in the same tissue

Network validation: are the proteins encoded by disease genes, which are specific for each disease, expressed in the same tissues?

Genomics. 2005 Aug;86(2):127-41.

Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues.

Ge X, Yamamoto S, Tsutsumi S, Midorikawa Y, Ihara S, Wang SM, Aburatani H.

Abstract

A critical and difficult part of studying cancer with DNA microarrays is data interpretation. Besides the need for data analysis algorithms, integration of additional information about genes might be useful. **We performed genome-wide expression profiling of 36 types of normal human tissues** and identified 2503 tissue-specific genes. We then systematically studied the expression of these genes in cancers by reanalyzing a large collection of published DNA microarray datasets. We observed that the expression level of liver-specific genes in hepatocellular carcinoma (HCC) correlates with the clinically defined degree of tumor differentiation. Through unsupervised clustering of tissue-specific genes differentially expressed in tumors, we extracted expression patterns that are characteristic of individual cell types, uncovering differences in cell lineage among tumor subtypes. We were able to detect the expression signature of hepatocytes in HCC, neuron cells in medulloblastoma, glia cells in glioma, basal and luminal epithelial cells in breast tumors, and various cell types in lung cancer samples. We also demonstrated that tissue-specific expression signatures are useful in locating the origin of metastatic tumors. Our study shows that integration of each gene's breadth of expression (BOE) in normal tissues is important for biological interpretation of the expression profiles of cancers in terms of tumor differentiation, cell lineage, and metastasis.

Tissue –specific distribution of disease genes

Network validation: are the proteins encoded by disease genes, which are specific for each disease, expressed in the same tissues?

The tissue homogeneity (TH) coefficient quantifies whether genes that are implicated in the same disorders tend to be expressed in similar human tissues.

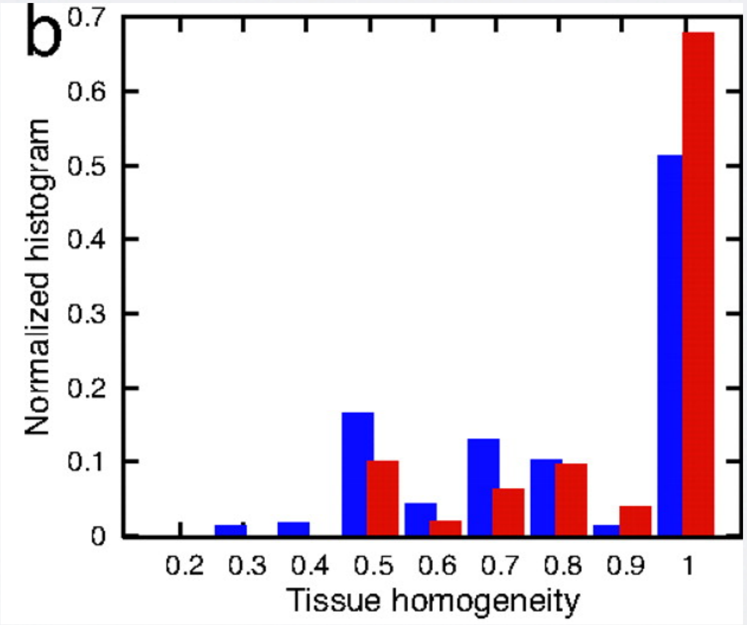
$T_{H_i} = \max_j [n^j_i/n_i]$

n_i = numero di geni nella malatia “i” che e’ espresso in almeno un tessuto; n^j_i = numero di geni espressi nel tessutto j tra i geni di malattia;

TH = 1 se tutti I geni sono espressi assieme in almeno un tessuto ed ha valore minimo 1/n quando tutti i geni sono espressi in tessutti diversi

Dataset of 10.594 geni in 36 healthy tissues (Ge et al., Genomics, 2005)

68% delle malattie connesse presentano omogeneità’ tessutale *P*<10^{-5}



In rosso-il coefficiente dei disease genes
In blu-il coefficiente dello stesso numero di geni
scelti in maniera random dal micorrray

Goh K et al. PNAS 2007;104:8685-8690

Disease module validation

Remember!!

Functional homogeneity

- **Gene ontology** (1) **biological processes**, 2) **cellular components** and 3) **molecular functions)**

Tissue specificity

Dynamic homogeneity

- **Co-expression**

Other tests:

- **GWA (SNPs identified for the nodes/genes)**

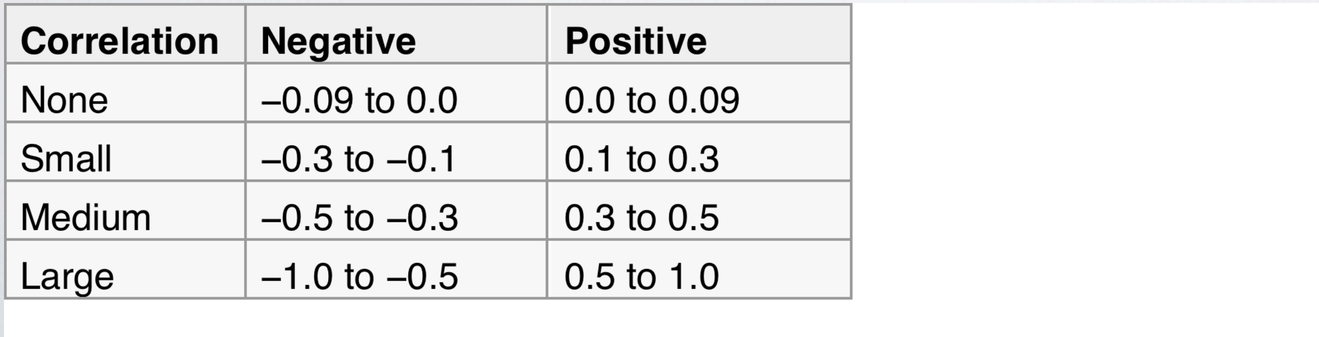
Predictions:

- **Disease genes**
- **Disease pathways**
- **Drug targets**

Network validation: are the disease genes co-expressed? Does the network show dynamic homogeneity?

Pearson correlation coefficient (PCC) = descriptor of the degree of linear association between two variables.

The correlation coefficient ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables.

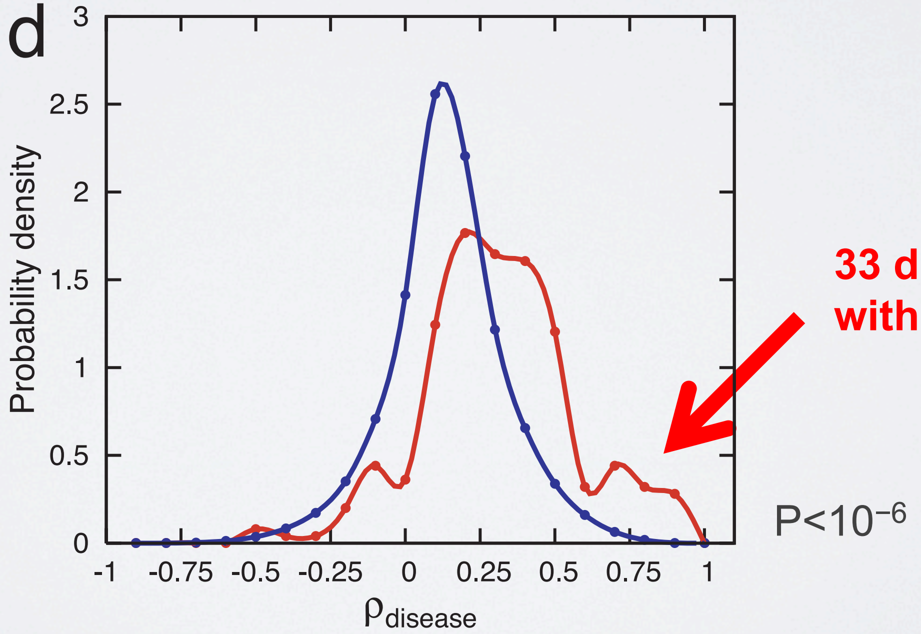


Correlation	Negative	Positive
None	–0.09 to 0.0	0.0 to 0.09
Small	–0.3 to –0.1	0.1 to 0.3
Medium	–0.5 to –0.3	0.3 to 0.5
Large	–1.0 to –0.5	0.5 to 1.0

Network validation: are the disease genes co-expressed? Does the network show dynamic homogeneity?

DISTRIBUTION OF PEARSON CORRELATION COEFFICIENTS (PCCS) FOR THE COEXPRESSION PROFILES OF PAIRS OF GENES ASSOCIATED WITH THE SAME DISORDER

Average PCC of all gene pairs belonging to the same disease
(blu: randomly chosen genes)



33 diseases with PCC >0.6

The genes involved in the **same disease**:

1. Their products tend to interact in PPI networks

2. Have the tendency to be expressed together in the same tissues

3. Have high level of coexpression (dynamic homogeneity)

4. Tend to share the same GO terms

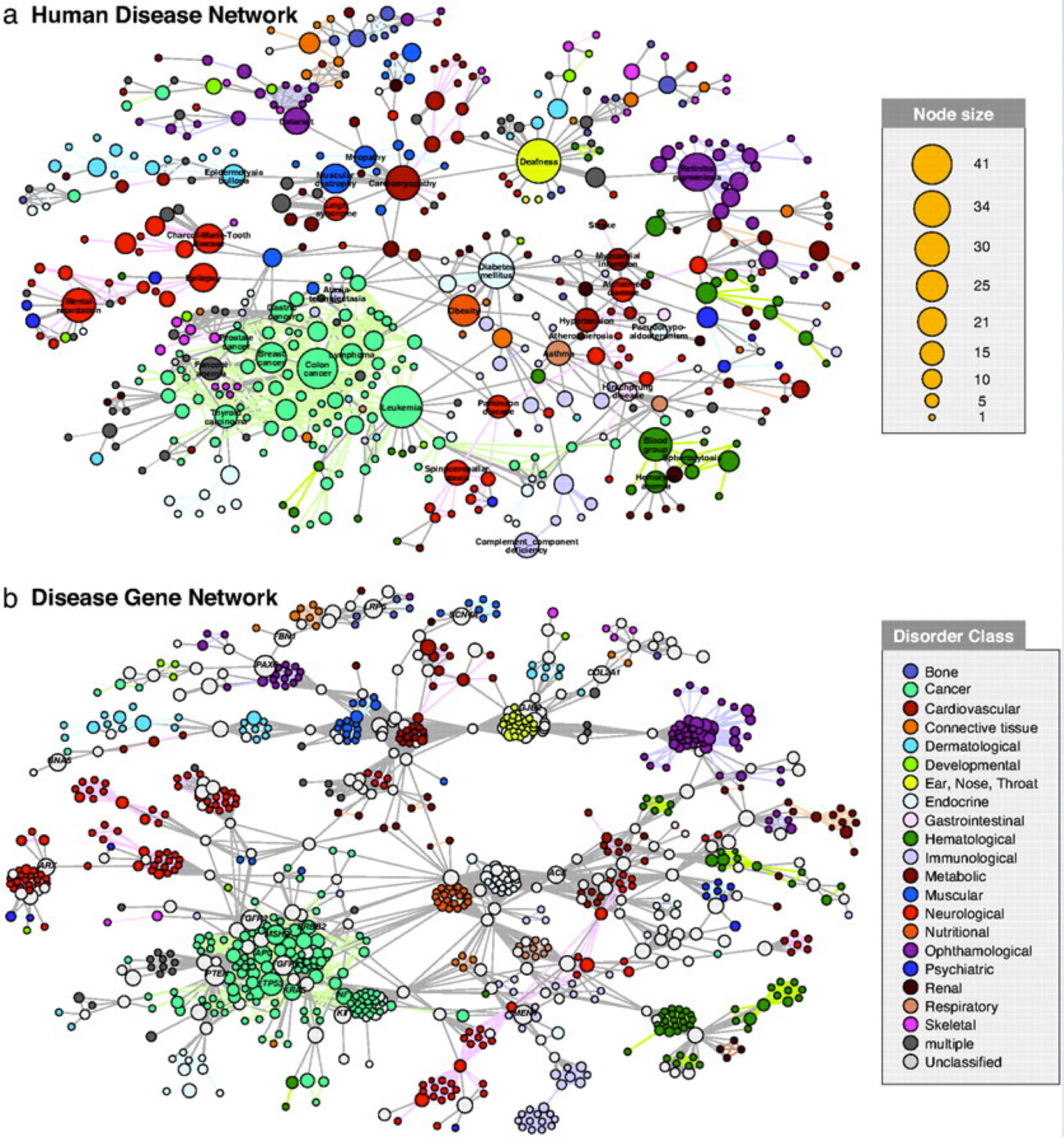
→

The analysis suggests:

GLOBAL FUNCTIONAL INTERRELATION of disease genes and their products offering a network-based functional explanation for complex and polygenic diseases

▼

Validity of network model for the diseasome


a Human Disease Network
Cataract
Myopathy
Epidermolysis bullosa
Muscular dystrophy
Cardiomyopathy
Long-QT syndrome
Deafness
Retinitis pigmentosa
Charcot-Marie-Tooth disease
Epilepsy
Mental retardation
Ataxia telangiectasia
Gastric cancer
Diabetes mellitus
Obesity
Stroke
Myocardial infarction
Alzheimer disease
Hypertension
Atherosclerosis
Pseudohypoaldosteronism
Asthma
Prostate cancer
Breast cancer
Colon cancer
Lymphoma
Fanconi anemia
Thyroid carcinoma
Leukemia
Parkinson disease
Hirschsprung disease
Blood group
Spherocytosis
Hemolytic anemia
Spinocerebellar ataxia
Complement_component deficiency
Node size
41
34
30
25
21
15
10
5
1
b Disease Gene Network
LRP5
SCN4A
FBN1
PAX6
GJB2
COL2A1
GNAS
AR
ACE
FGFR2
MSH2
ERBB2
APC
FGFR3
PTEN
TP53
KRAS
NF1
KIT
MEN1
Disorder Class
Bone
Cancer
Cardiovascular
Connective tissue
Dermatological
Developmental
Ear, Nose, Throat
Endocrine
Gastrointestinal
Hematological
Immunological
Metabolic
Muscular
Neurological
Nutritional
Ophthamological
Psychiatric
Renal
Respiratory
Skeletal
multiple
Unclassified

Centrality and peripherality in the diseasome

Observations in *Sacharomyces cerevisiae: hub proteins Tend to be encoded by essential genes*

▼

Does disease genes encode hub proteins?

▼

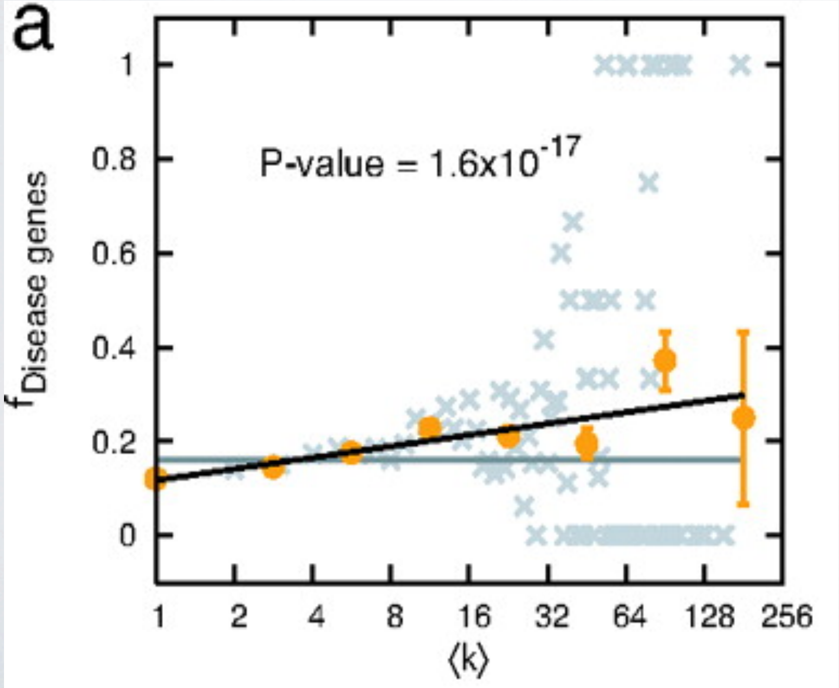
Disease genes have higher number of interactions than non-disease genes - higher average k of ≈ 30%
(*Rual JF Nature 2005; Stelzl et al., Cell 2005*)

Do disease genes encode hub proteins?

2. Disease genes have the tendency to encode proteins with high k

The fraction of disease genes among those whose protein products that interact with k other proteins (a measure of the dependency of degree)

Proteins encoded by all disease genes (1.777)



Goh K et al. PNAS 2007;104:8685-8690

Linear regression model

c^2 test

Gray symbols are the linearly binned data points, whereas color corresponds to the statistically more uniform log-binned data

In statistics a **bin** — sometimes called a class interval — is a way of sorting data in a histogram. It's very similar to the idea of putting data into categories.

**Fraction of disease nodes from the total of nodes (proteins) with that particular k.
Ex.: 40% of nodes with k=32 are disease genes**

Explanation for the previous slide

Scelto un grado K, sul punto corrispondente nell’asse Y c’e’ la frazione di nodi “disease” rispetto al totale che hanno quel grado K (la crocetta grigia). Cioè quanti sono i nodi disease che hanno grado K rispetto al totale dei nodi che hanno grado K. Per questo i valori dell’asse Y vanno da 0 a 1. Ad esempio in figura 4a con grado K=32 c’e’ una crocetta grigia in corrispondenza del valore 0,4. Significa che presi tutti i nodi di grado 32, 0,4 è la frazione (cioè il 40%) di quelli che sono disease nodes. **Cioè il 40% dei nodi di grado 32 sono disease nodes.**

Le crocette grige indicano i valori reali, i puntini colorati sono invece il logaritmo degli stessi valori, in modo da rendere tutto più uniforme.

Do disease genes encode hub proteins?

Are disease genes also essential proteins?

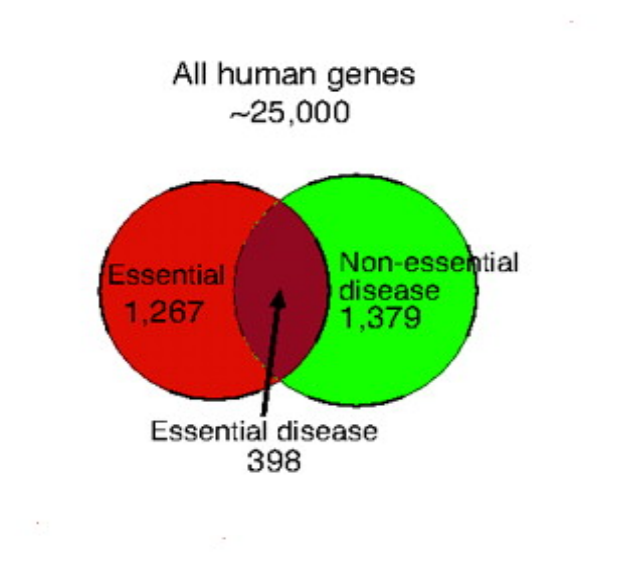
! **Geni essenziali** per lo sviluppo embrionale se sono modificati

possono portare ad aborti spontanei nel primo trimestre di gravidanza.

Quindi non conosciamo tanti geni essenziali che sono anche "disease genes" *in utero* nel primo trimestre di gravidanza !

Considerare i geni umani ortologhi dei geni murini che sono "embryonic lethal" o inducono mortalità perinatale (mouse genome informatics - www.informatics.jax.org)

Jackson Labs: 1.267 ortologhi murini letali di cui 398 disease genes umani (in rosso)



Human disease genes (1.777) (green):

1.379 non-essential

398 essential

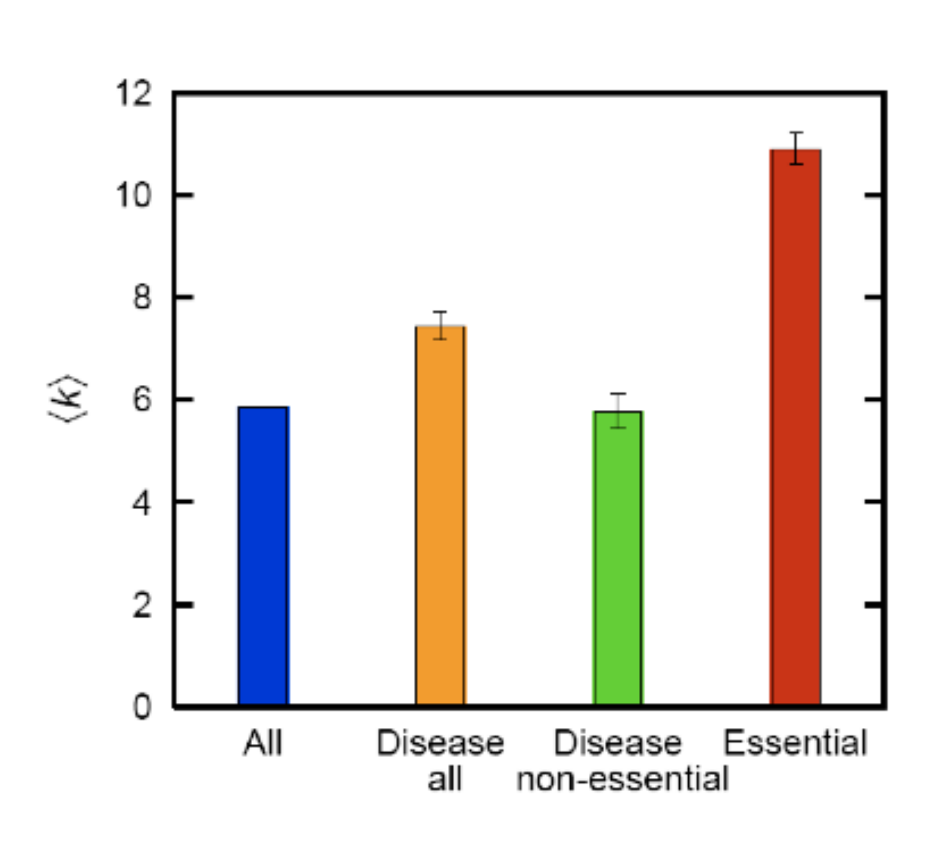
Goh K et al. PNAS 2007;104:8685-8690

Orthologs are **genes** in different species that evolved from a common ancestral **gene** by speciation. Normally, **orthologs** retain the same function in the course of evolution.

Does disease genes encode hub proteins?

Which are the differences between essential and non-essential disease genes?

Essential disease genes have higher *k* comparing to non-essential genes



Does disease genes encode hub proteins?

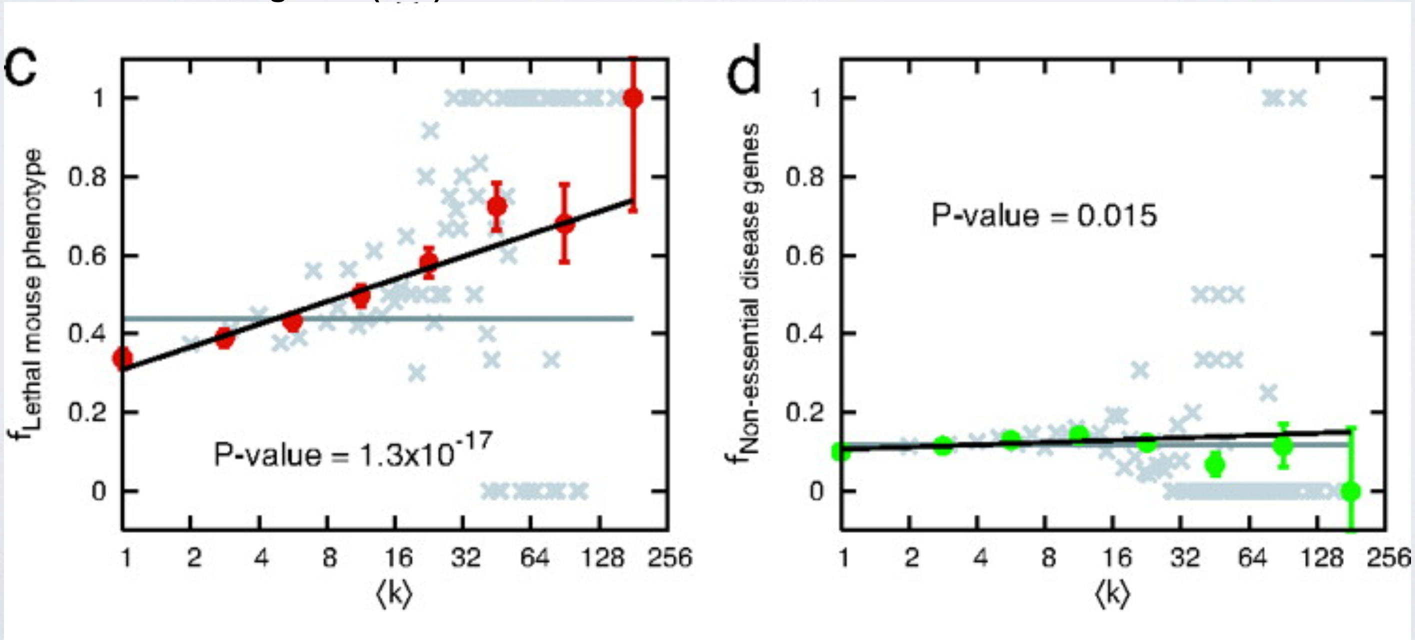
Which are the differences between essential and non-essential disease genes?

3. Essential proteins encoded by disease genes tend to be hubs

Fractions of essential and non-essential genes from the total of genes encoding proteins with k interactions

Proteins encoded by essential genes (398)

Proteins encoded by non- essential genes (1379)



Goh K et al. PNAS 2007

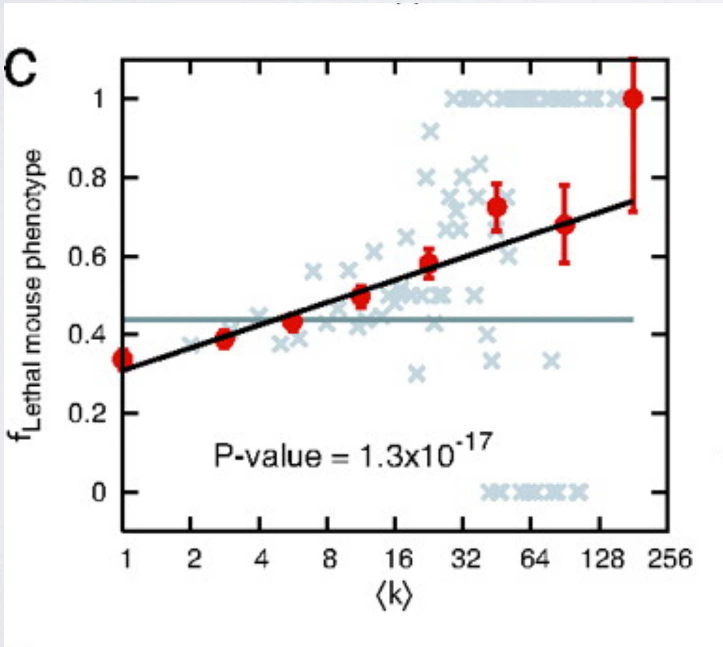
Does disease genes encode hub proteins?

Which are the differences between essential and non-essential disease genes?

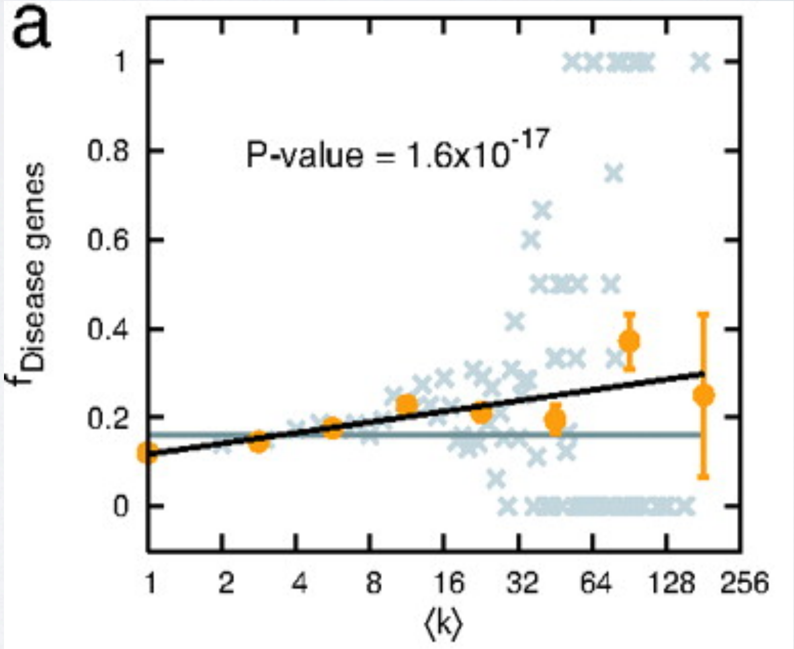
3. Essential proteins encoded by disease genes tend to be hubs

Fractions of essential and non-essential genes from the total of genes encoding proteins with k interactions

Proteins encoded by essential genes (398)



Proteins encoded by all disease genes (1.777)



Goh K et al. PNAS 2007;104:8685-8690

Gene synchronization

To carry on its basic functions, the cell needs to maintain the coordinated activity of important functional modules, driving in a relatively synchronized manner the expression patterns of the most important genes.

Therefore, one expects that the expression pattern of both essential and disease genes will be synchronized with a significant number of other genes.

Does disease genes encode hub proteins?

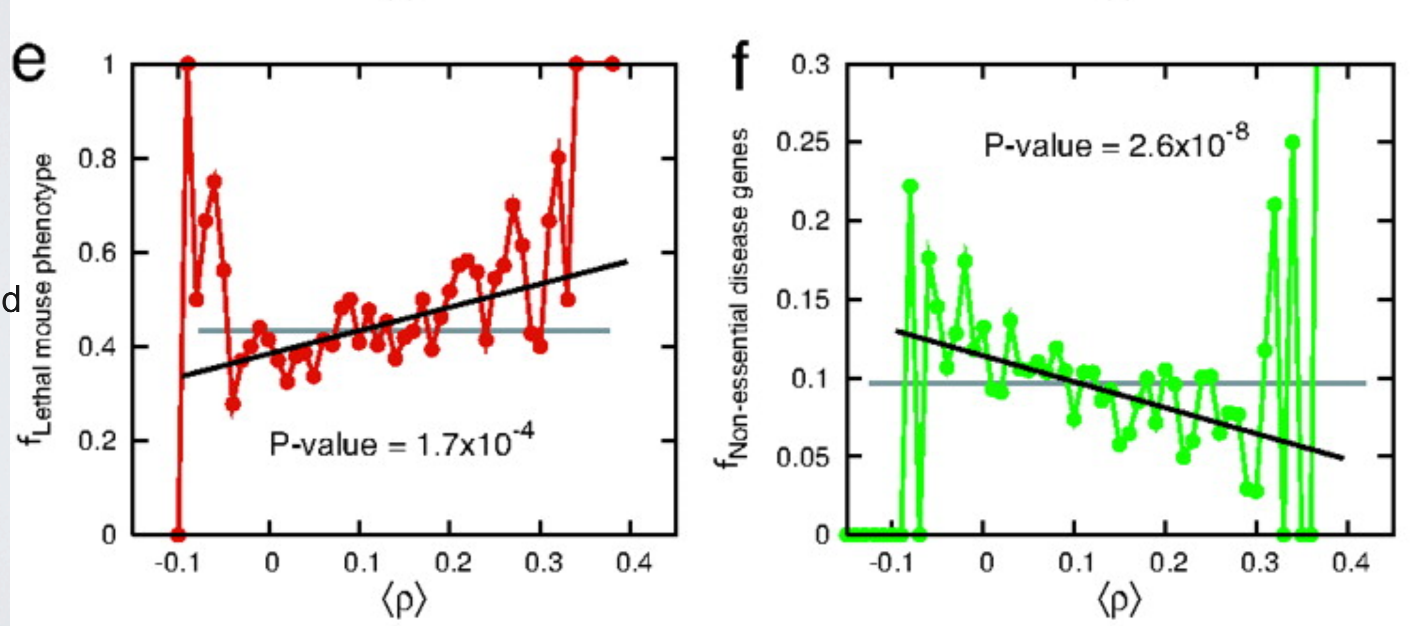
Are disease genes synchronized with other cell genes?

It was calculated the **average gene coexpression coefficient** ρ_i **(**$\rho_i = \Sigma_j PCC_{ij}$) between an essential (or nonessential disease) gene *i* and all other genes in the cell by calculating the PCC_{ij} values from healthy human tissue microarray measurements (Ge et al., 2005)

essential genes (398)

non-essential genes (1379)

$\rho > 0.2$ for highly synchronized genes



The fraction of essential genes (e) and nonessential disease genes (f) among those whose average PCC with other genes is 〈ρ〉 .

Goh K et al. PNAS 2007;104:8685-8690

Useful published study to determine gene coexpression

Genomics. 2005 Aug;86(2):127-41.
Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues.
Ge X, Yamamoto S, Tsutsumi S, Midorikawa Y, Ihara S, Wang SM, Aburatani H.

Abstract
A critical and difficult part of studying cancer with DNA microarrays is data interpretation. Besides the need for data analysis algorithms, integration of additional information about genes might be useful. **We performed genome-wide expression profiling of 36 types of normal human tissues** and identified 2503 tissue-specific genes. We then systematically studied the expression of these genes in cancers by reanalyzing a large collection of published DNA microarray datasets. We observed that the expression level of liver-specific genes in hepatocellular carcinoma (HCC) correlates with the clinically defined degree of tumor differentiation. Through unsupervised clustering of tissue-specific genes differentially expressed in tumors, we extracted expression patterns that are characteristic of individual cell types, uncovering differences in cell lineage among tumor subtypes. We were able to detect the expression signature of hepatocytes in HCC, neuron cells in medulloblastoma, glia cells in glioma, basal and luminal epithelial cells in breast tumors, and various cell types in lung cancer samples. We also demonstrated that tissue-specific expression signatures are useful in locating the origin of metastatic tumors. Our study shows that integration of each gene's breadth of expression (BOE) in normal tissues is important for biological interpretation of the expression profiles of cancers in terms of tumor differentiation, cell lineage, and metastasis.

Considerations on previous slide:

Confirming our expectation, for essential genes we find that genes that display high average coexpression ρ with all other genes are more likely to be essential than those that show small or negative ρ (*P* = $1.7 < 10^{-4}$).

ρ > 0.2 means highly synchronized genes

However, nonessential disease genes show the opposite effect, being associated with genes whose expression pattern is anticorrelated or not-correlated with other genes, and underrepresented among the genes that are highly synchronized (ρ > 0.2) (*P* = $2.6 = 10^{-8}$).

Thus, the expression pattern of nonessential disease genes appears to be decoupled from the <u>overall expression pattern of all other genes, whereas essential genes have a tendency to be coupled to the rest of the cell.</u>

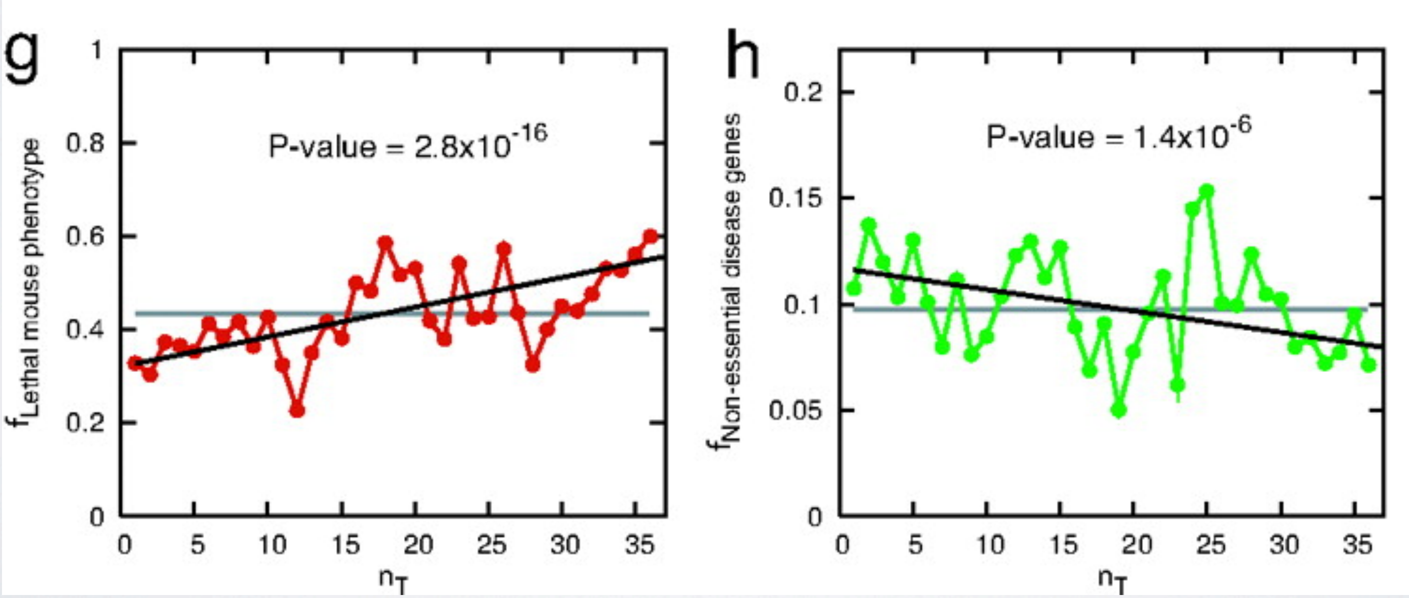
Does disease genes encode hub proteins?

Are “housekeeping” genes (constitutive, present in all cells) also “disease genes”?

Th**e FRACTION OF ESSENTIAL GENES AND NONESSENTIAL DISEASE GENES AMONG THOSE WHOSE TRANSCRIPT IS EXPRESSED IN N TISSUES** (nT = nr. di tessuti)

Essential genes

Non-essential genes



Goh K et al. PNAS 2007;104:8685-8690

Housekeeping genes are **genes** required for the maintenance of basal cellular functions that are essential for the existence of a cell, regardless of its specific role in the tissue or organism.

They are expected to be expressed in all cells of an organism under normal conditions, irrespective of tissue type, developmental stage, cell cycle state, or external signal.

Generalmente, essi codificano proteine ed enzimi fondamentali per la vita della cellula, e che pertanto devono essere sempre presenti. Sono presenti in tutte le cellule.

Alcuni esempi comuni di geni costitutivi sono, per esempio i geni che codificano la proteina actina, o enzimi quali le hexokinase.

Conclusions of diseasome application:

Non-essential disease genes:

1. Tend to non associate with hubs

2. Correlate less with other genes expressed in a cell

3. Have the tendency to be expressed in less tissues

4. Are the majority of disease genes

Essential disease genes:

1. Tend to associate with hubs

2. Correlate significantly with other genes expressed in the cells

3. Have the tendency to be expressed in more tissues

4. Are highly expressed as “housekeeping” genes

5. Are a minority of disease genes

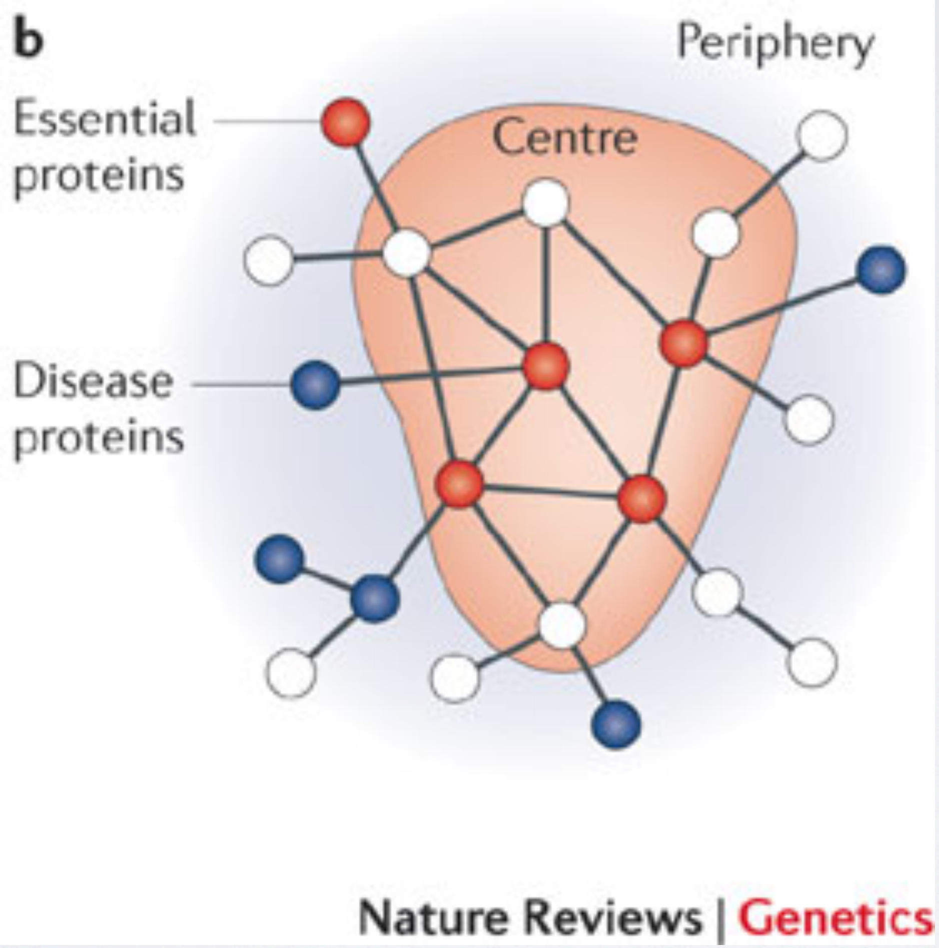
Observations and hypotheses on HUBS (= nodes with high degree _):_

- Essential genes *in utero* tend to associate with hubs.

- Genes encoding hubs are older and evolve more slowly than genes encoding non-hub proteins

- The absence of a hub is expected to affect many more other proteins than would the absence of a non-hub protein

Localization of essential *versus* non-essential *disease genes* in the interactome



Essential genes:

1. **Tend to be hubs**

2. Are localized in the **functional center** of the interactome

Barabasi et al., 2011

Conclusions

- Not all essential genes are “disease genes” (are a minority)
- Mutations in genes which are important for embrionic development do not propagate in the human population and are not “disease genes”
- The majority of “disease genes” are NON-essential genes

- Essential genes are associated to hubs, are expressed in many tissues and tend to be localized at the center of the interactome
- Non-essential disease genes are not hubs, are tissue-specific and tend to be localized at the periphery of the interactome

IMPORTANT FINAL CONCLUSIONS

1. Non-essential disease genes are more **peripheral** in the cellular network from the functional and topological point of view.
“**Neutral**” position

2. Essential disease genes have a more **central** position in the cellular network from the functional and topological point of view.
“**Central**” position

IMPORTANT GENERAL CONCLUSIONS

1. The majority of disease genes are **peripheral** and **non-essential**. Their mutations are less important for the survival.

2. **Essential** disease genes are important *in utero*, have a **central** position and are largely expressed in the tissues. Their mutations are lethal.

FINE

THE END

Network theory → network medicine

Nods: proteins, metabolites, diseases
Links: protein-protein interactions; metabolic reactions; shared genes

Elements of network theory:

1. Degree distribution e hubs
2. Modules
3. Phenomenon of small-world
4. Motifs
5. Betweeness centrality

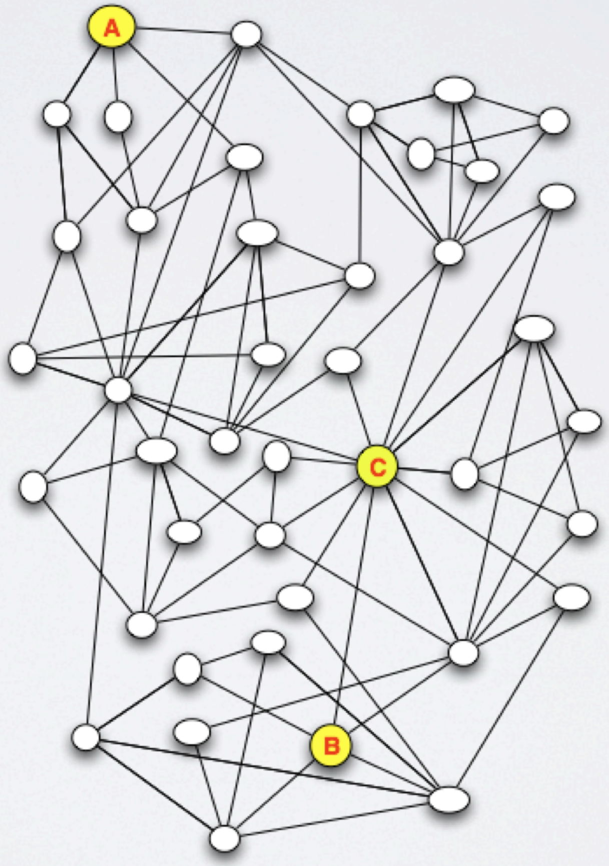
1. Degree distribution e hubs

"Node degree" *k* = numero di connessioni

A = *k* 4 B = *k* 5 C = *k* 12

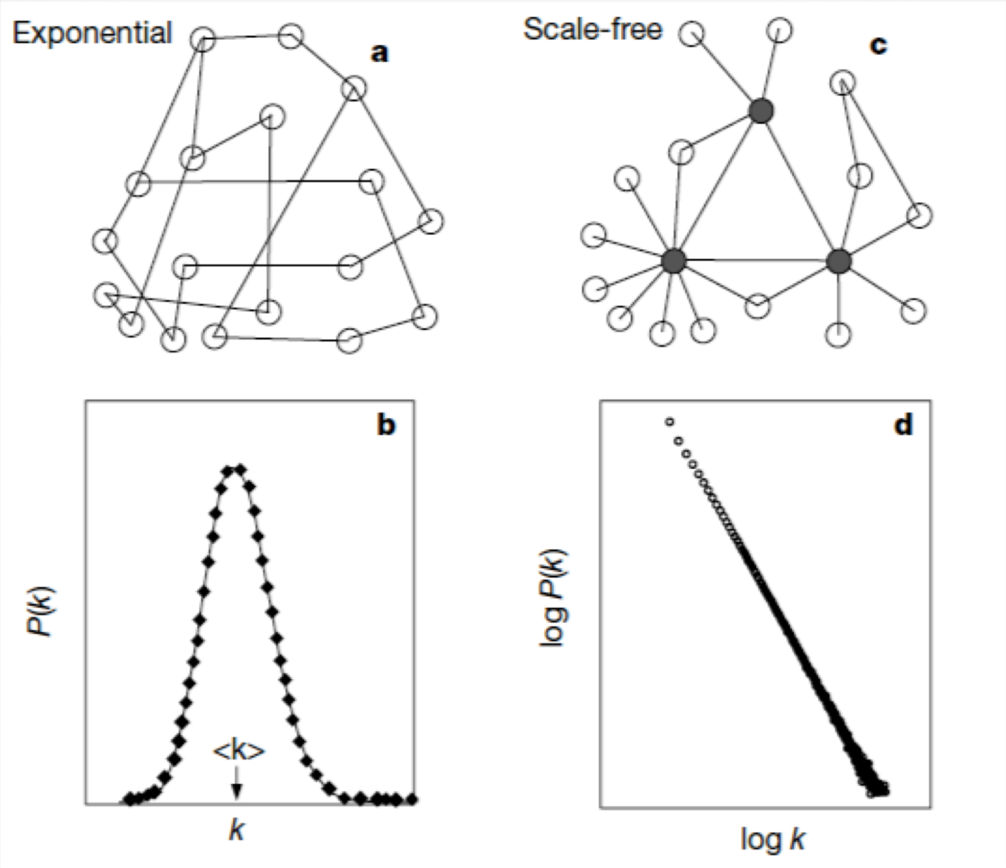
"Average degree" <*k*> = numero medio di connessioni/nodo

*k*4 + *k*5 + *k*12 = <k> 7



Random network

Scale-free network



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

y= degree exponent

Generalmente molto omogenei

La connettivita' dei nodi segue la distribuzione di Poisson.

La probabilita' di trovare nodi fortemente connesi (con molti links) decade in modo **esponenziale.**
In genere k = 2 o 3

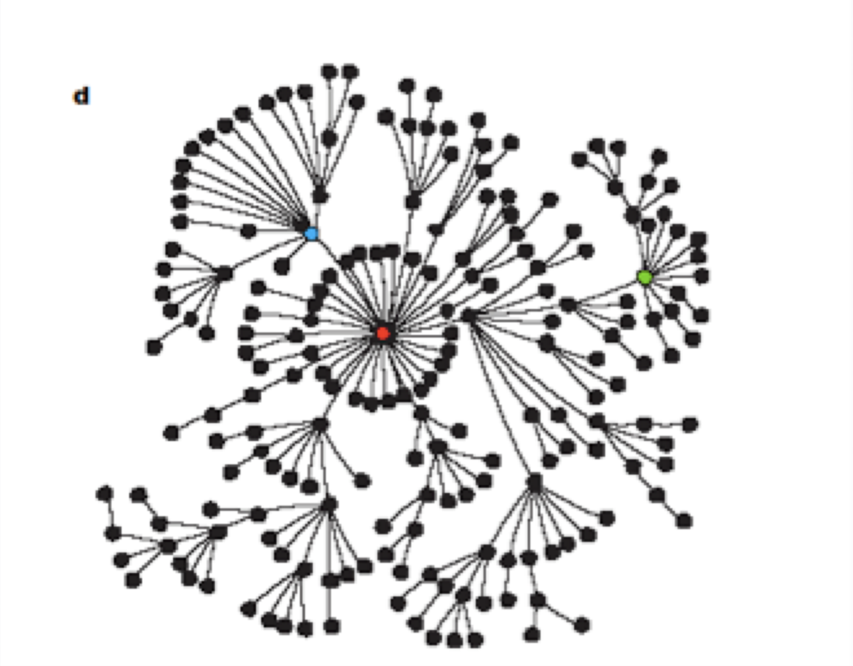
Generalmente molto eterogenei

Hanno topologia dominata da pochi nodi fortemente connessi (hubs) a tutti gli latri nodi poco connessi

La probabilita' di trovare nodi fortemente connesi (con molti links) segue un andamento logaritmico **(power law)**

1. Degree distribution e hubs

Scale-free network
(Barabasi)



Party hubs

Date hubs

Albert-László Barabási

https://www.google.com/search?client=safari&rls=en&q=albert+laszlo+barabasi&ie=UTF-8&oe=UTF-8

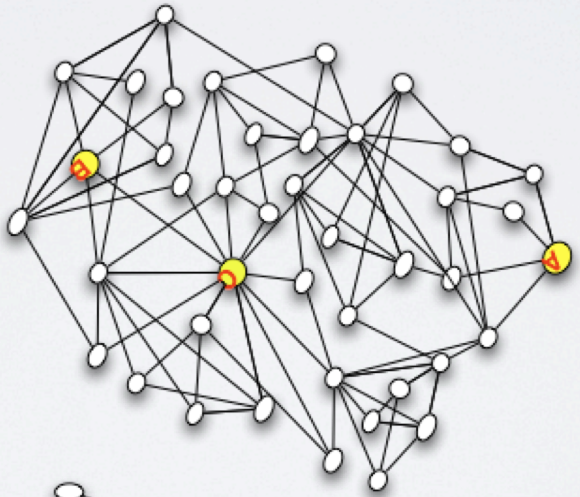
2. Moduls = highly interconected areas in a network

"Clustering coefficient" *C* =
probabilità che se A -- B e B -- C allora A -- C

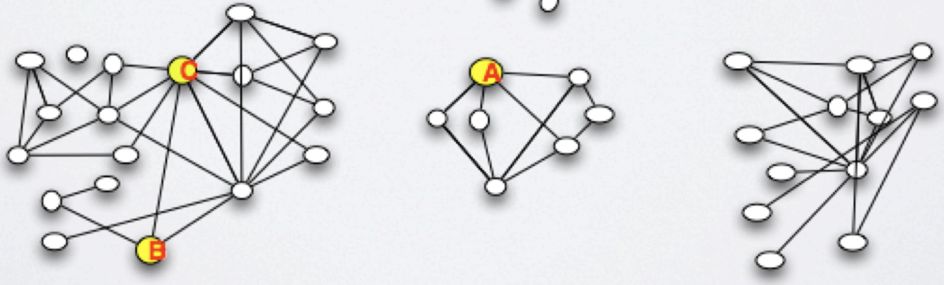
"Average Clustering coefficient" *<C>* =

Probabilità media della rete di formare clusters.

Indica la tendenza globale dei nodi
ad organizzarsi in gruppi (clusters)



Decomposizione modulare
(ad es. MCODE)

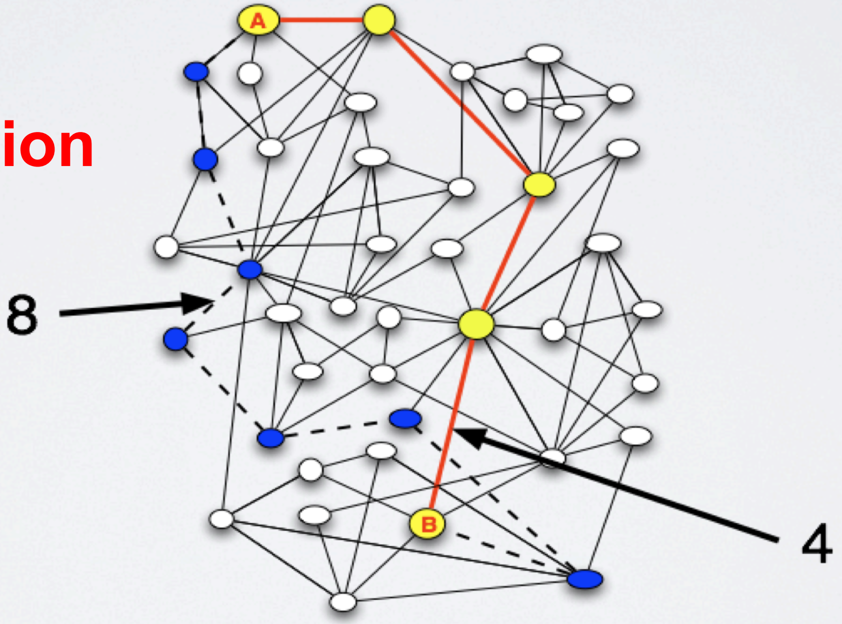


3. Small-world phenomena
= there are relatively “short paths” between the nodes

"Average shortest path" =
il tragitto minimo "medio" fra tutti i nodi

Misura la "navigabilità" globale della rete,
cioè la facilità con cui si si sposta da un
nodo all'altro

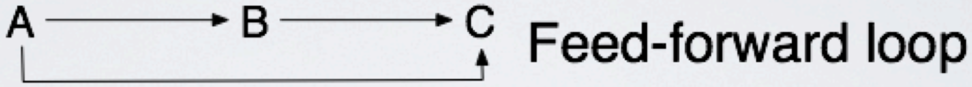
One node dysfunction can influence the global network

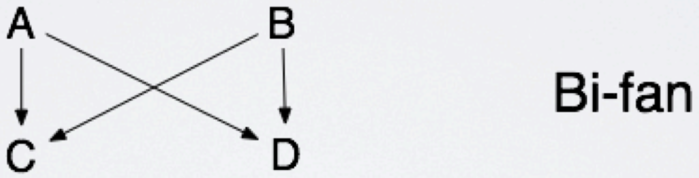


"Average path length" = distanza media fra i nodi
(non minima);
"Diameter" = shortest path (distanza) massima fra i nodi;
Indici di connettività

4. Motifs

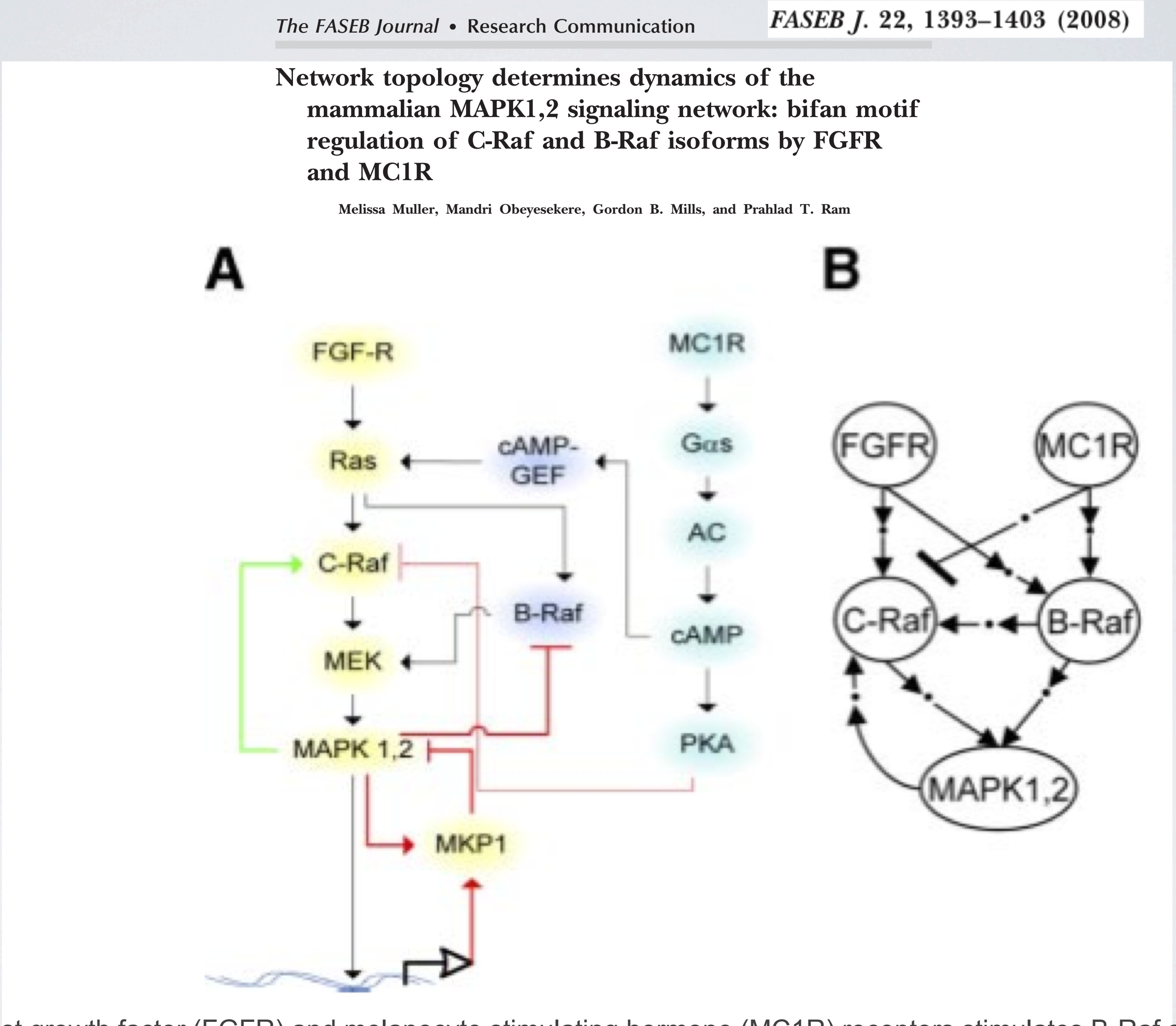
= patterns (<u>subgraphs</u>) that occur more frequently in real networks than in random networks generated at computer





<u>are likely to be associated with some optimized biological function</u>

4. Motifs – esempio di bifan motif



The FASEB Journal • Research Communication

FASEB J. 22, 1393–1403 (2008)

Network topology determines dynamics of the mammalian MAPK1,2 signaling network: bifan motif regulation of C-Raf and B-Raf isoforms by FGFR and MC1R

Melissa Muller, Mandri Obeyesekere, Gordon B. Mills, and Prahlad T. Ram

partially incoherent bifan motif

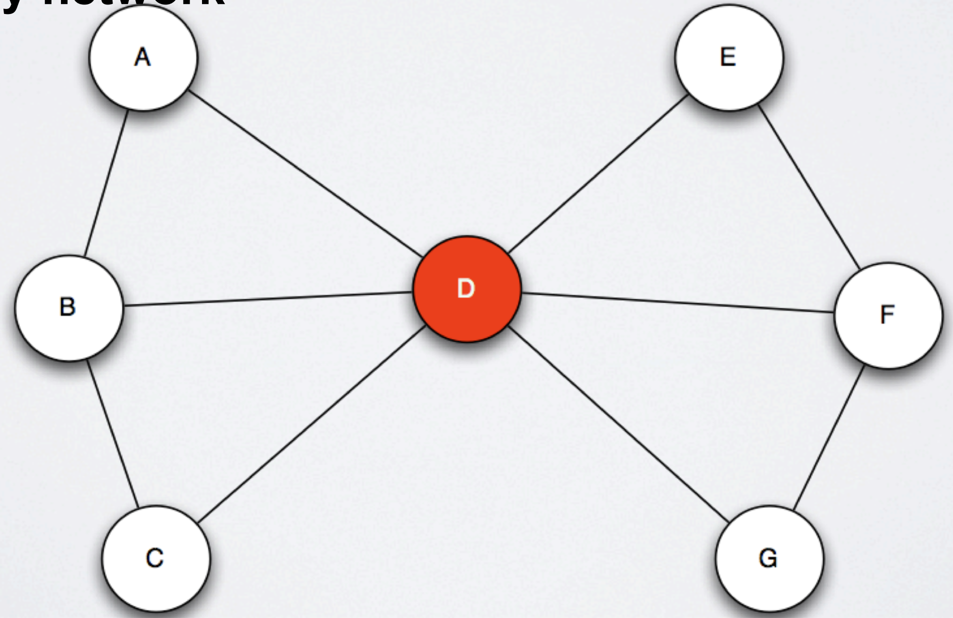
Activation of the fibroblast growth factor (FGFR) and melanocyte stimulating hormone (MC1R) receptors stimulates B-Raf and C-Raf isoforms that regulate the dynamics of MAPK1,2 signaling in human melanoma cells.
A) Detailed network model of the FGFR-MC1R-B-Raf-C-Raf-MAPK1,2 network based on the experimental data. B) Connection diagram shows the reduced network structure of the FGFR-MC1R-B-Raf-C-Raf-MAPK1,2 network dynamics of MAPK1,2 activity.

5. Betweeness centrality

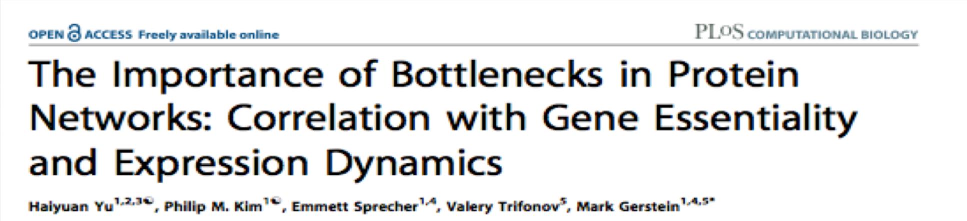
= nodo capace di mettere in comunicazione nodi o zone distinte della rete stessa.

High betweeness centrality means high number of shortest paths (Un valore alto di betweenness indica la capacità del nodo di funzionare come nodo comunicatore o "collo di bottiglia" - "bottleneck").

Bottleneck niodes are crossesd by many "shortest paths"; are like bridges or tunnels in a highway network



Are bottlenecks important because they are hubs or because they have high betweeness?



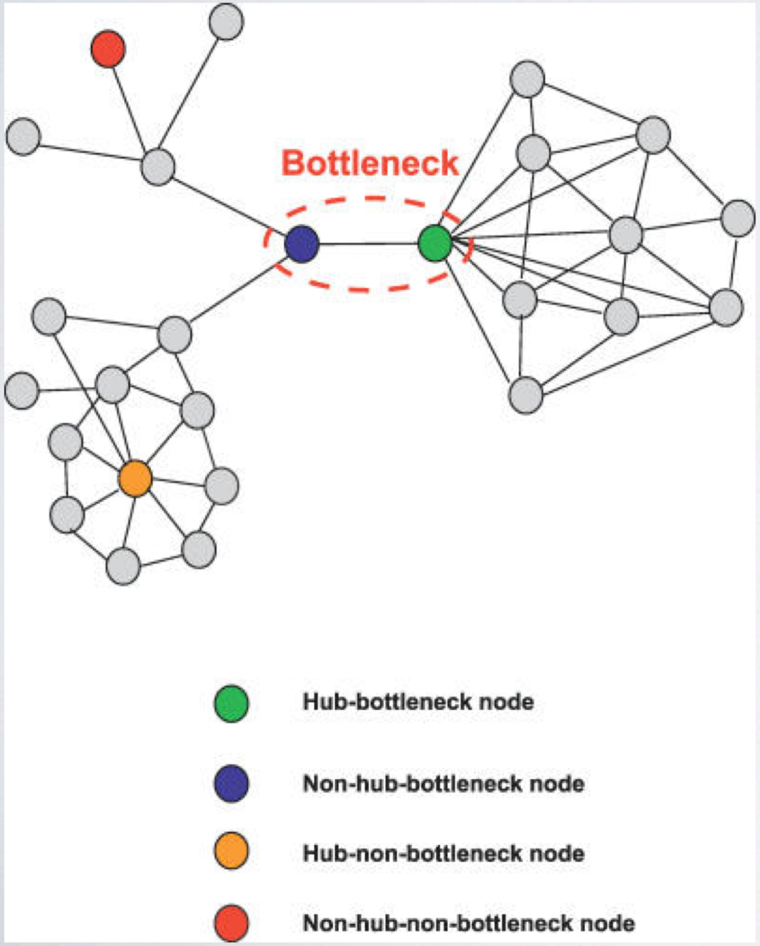
OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics

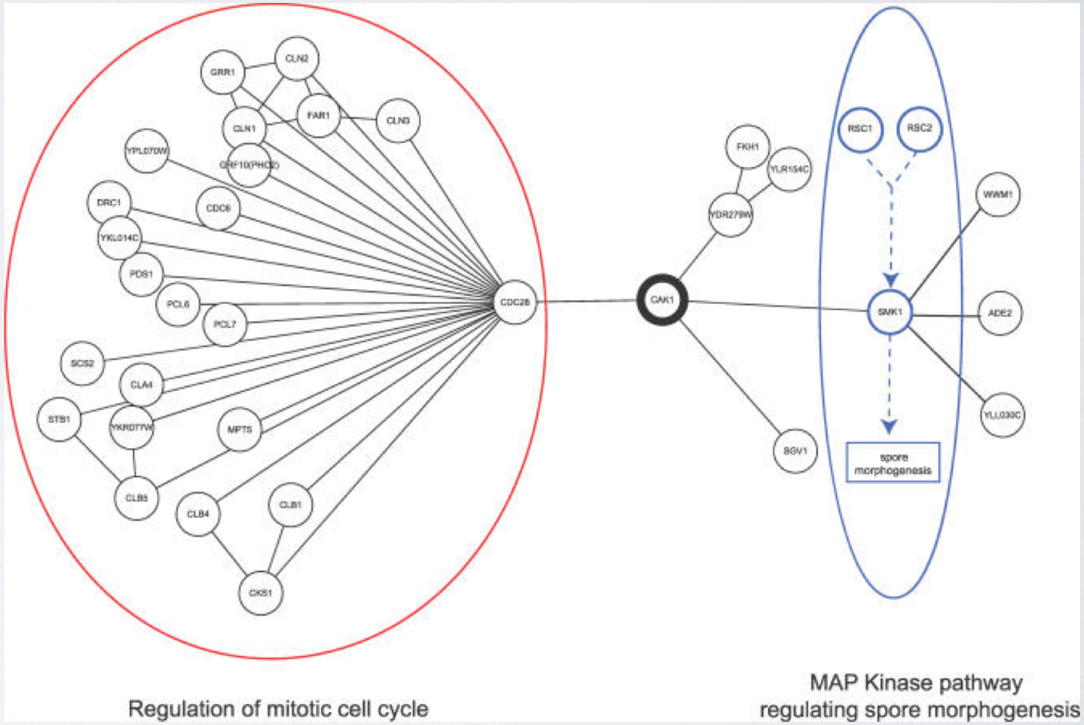
Haiyuan Yu[1,2,3☯], Philip M. Kim[1☯], Emmett Sprecher[1,4], Valery Trifonov[5], Mark Gerstein[1,4,5*]

2007

Tyoes of “bottleneck” nodes


Bottleneck
● Hub-bottleneck node
● Non-hub-bottleneck node
● Hub-non-bottleneck node
● Non-hub-non-bottleneck node

Cyclin-dependent kinase bottleneck



The topological position of Cak1 (cyclin-dependent kinase-activating kinase) suggest this gene/molecule is essential for the cell

5. Betweeness centrality

Yu et al.,PLoS Computational Biology, 2007:

In biological networks “bottleneck-ness” is a much more significant indicator of essentiality than degree (i.e., "hub-ness").

Bottlenecks are, in fact, key connector proteins with functional and dynamic properties.
They are more likely to be essential proteins

Bottlenecks correspond to the dynamic components of the interaction network; they are significantly less well coexpressed with their neighbors than non-bottlenecks

Hypotheses and principles of network medicine

1. Hubs
2. Disease module
3. Local hypothesis
4. Network parsimony principle
5. Shared components hypothesis

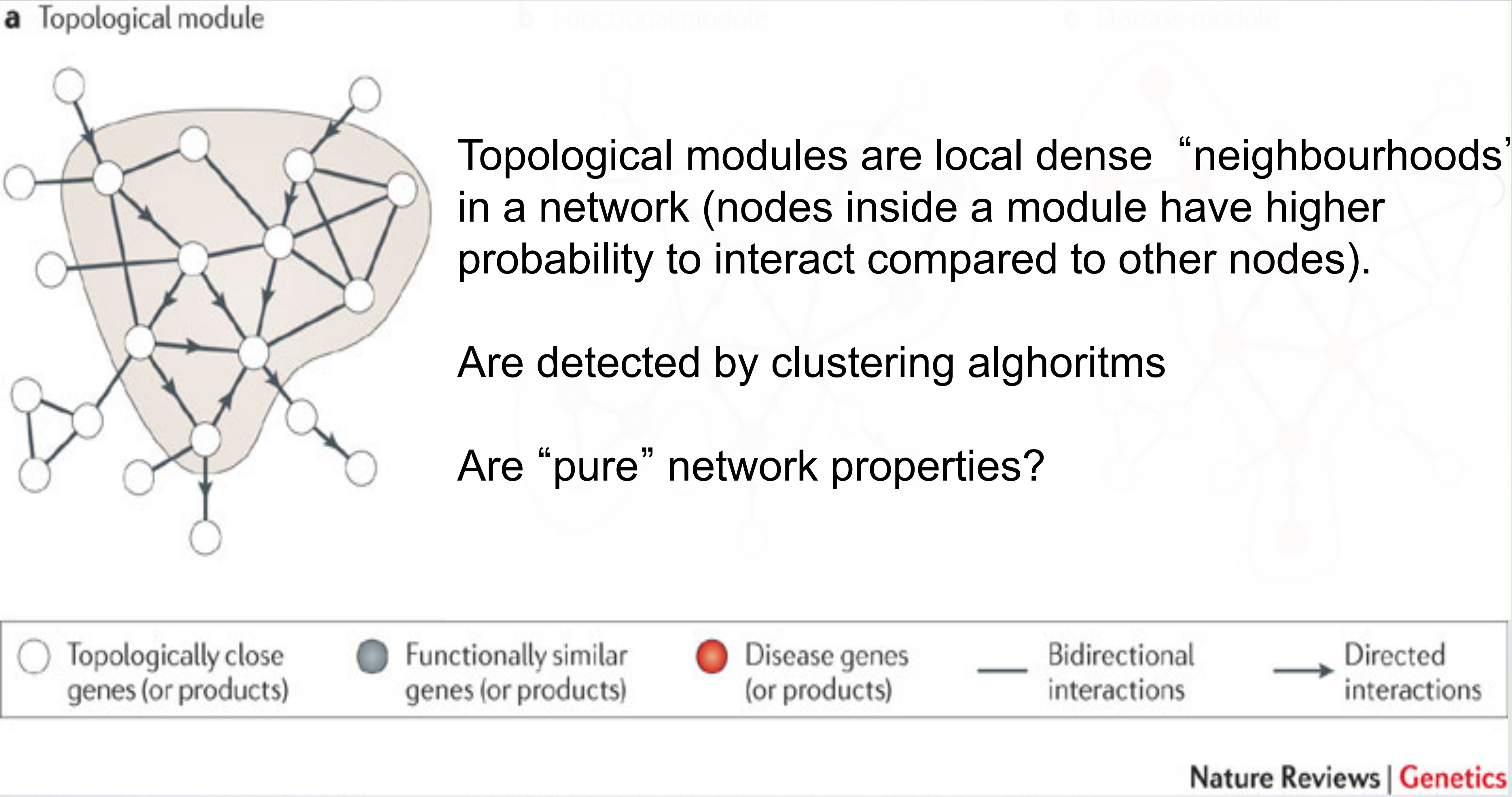
Hypotheses and principles of network medicine:

2. Disease modules

- Cellular components associated with a specific disease tend to Cluster in the same network “neighbourhood”

- Each disease is characterized by a module (“neighbourhood”) Inside the interactome

2. Disease modules



a Topological module
b Functional module
c Disease module
○ Topologically close genes (or products)
● Functionally similar genes (or products)
● Disease genes (or products)
— Bidirectional interactions
→ Directed interactions
Nature Reviews | Genetics

Topological modules are local dense "neighbourhoods" in a network (nodes inside a module have higher probability to interact compared to other nodes).

Are detected by clustering alghoritms

Are "pure" network properties?

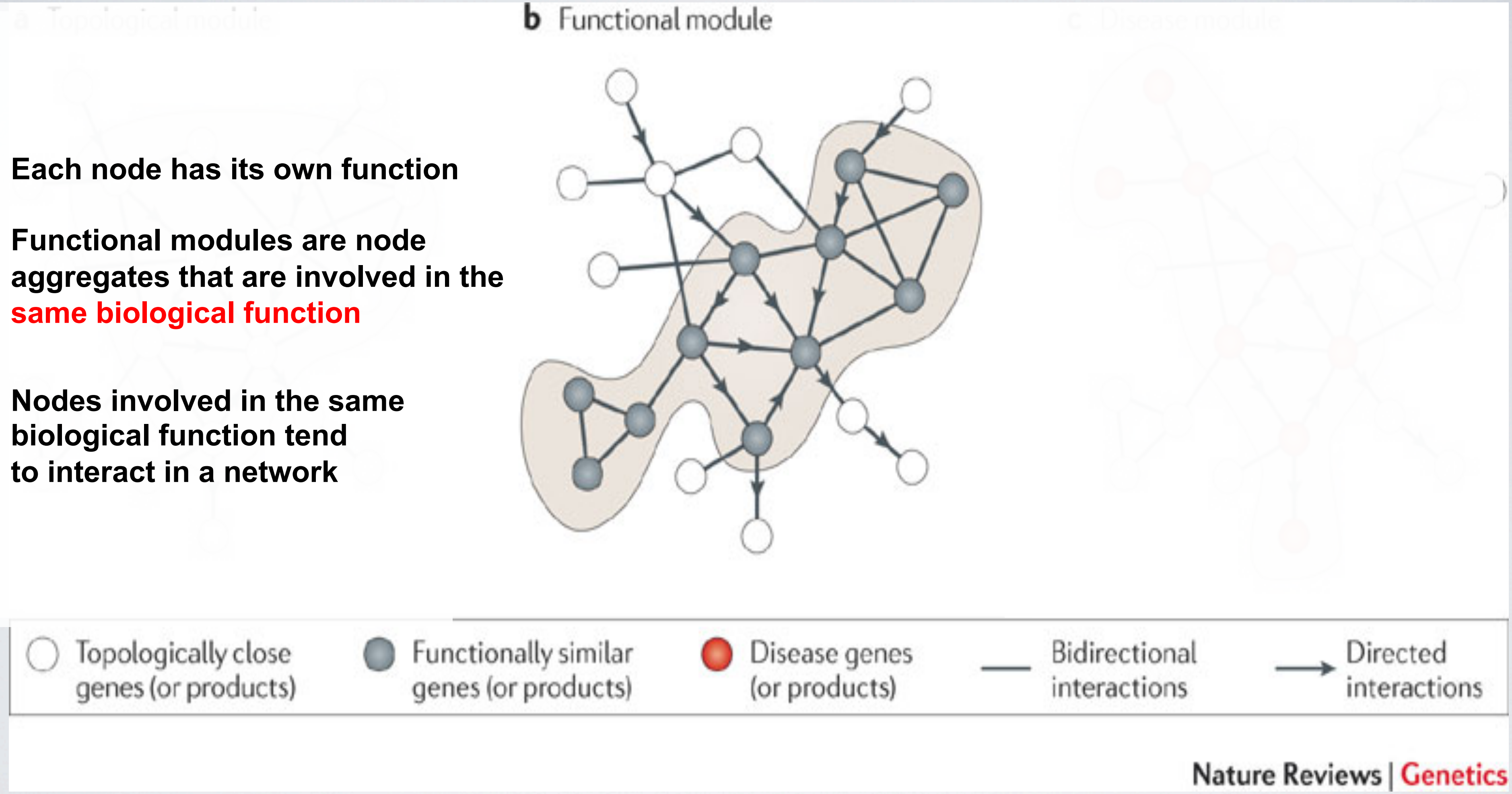
Barabasi et al., 2011

2. Disease modules

Each node has its own function

Functional modules are node aggregates that are involved in the same biological function

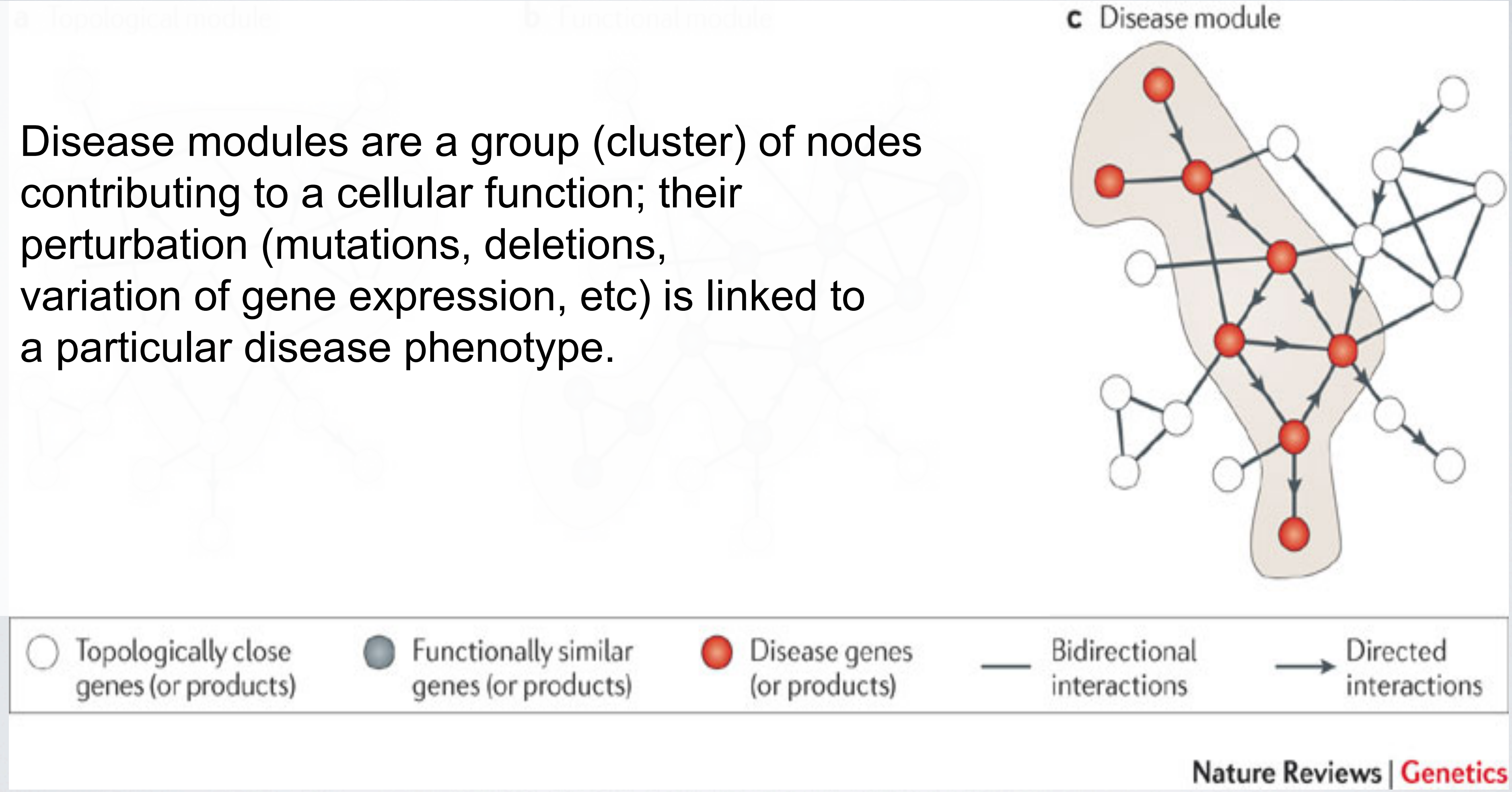
Nodes involved in the same biological function tend to interact in a network


a Topological module
b Functional module
c Disease module
○ Topologically close genes (or products)
● Functionally similar genes (or products)
● Disease genes (or products)
— Bidirectional interactions
→ Directed interactions
Nature Reviews | Genetics

Barabasi et al., 2011

2. Disease modules

Disease modules are a group (cluster) of nodes contributing to a cellular function; their perturbation (mutations, deletions, variation of gene expression, etc) is linked to a particular disease phenotype.


a Topological module
b Functional module
c Disease module
○ Topologically close genes (or products)
● Functionally similar genes (or products)
● Disease genes (or products)
— Bidirectional interactions
→ Directed interactions
Nature Reviews | Genetics

Barabasi et al., 2011

2. Disease modules

Hypotheses in network medicine:

- **Topological modules correspond with functional modules**

- **The disease is viewed as a perturbation of a functional module;**
Thus, functional modules are also disease modules

- **<u>Topological, functional and disease modules often OVERLAPP</u>**

2. Disease modules

Further considerations:

1. Disease modules are not IDENTICAL to, but have high probability to overlap to topological and functional modules

2. A disease module is defined in relation to a Specific disease; each disease has its own module

3. A gene, a protein or a metabolite can be involved in several disease modules, thus, some disease modules may overlap

Steps for the identification of disease modules (a-e)

a. Interactome reconstruction

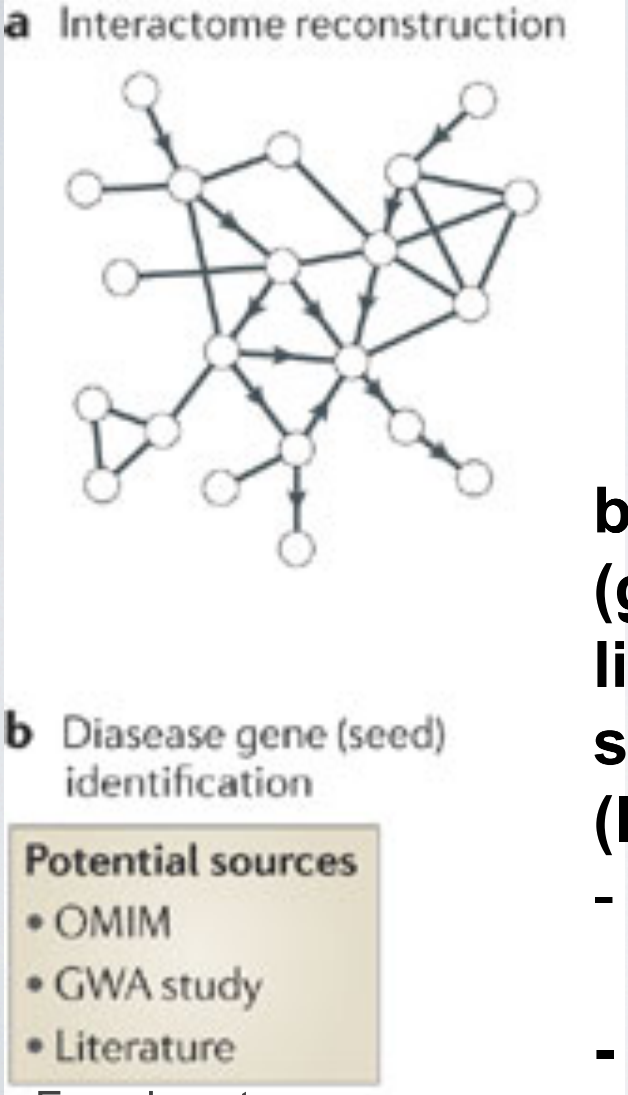
b. Disease gene identification (“disease seed”)

c. Disease module identification

d. Pathway identification (Identificazione di vie molecolari specifiche)

e. Disease module validation

Steps for the identification of disease modules (a-e)


a Interactome reconstruction
b Diasease gene (seed) identification
Potential sources
• OMIM
• GWA study
• Literature

- Esperiments

a. Interactome reconstruction
(of a cell or tissue)

b. Disease gene (seed) identification
(genes associated to disdease obtained for
linkage analysis, genome-wide association
studies (GWA) and other resources
(literature, microarrays)

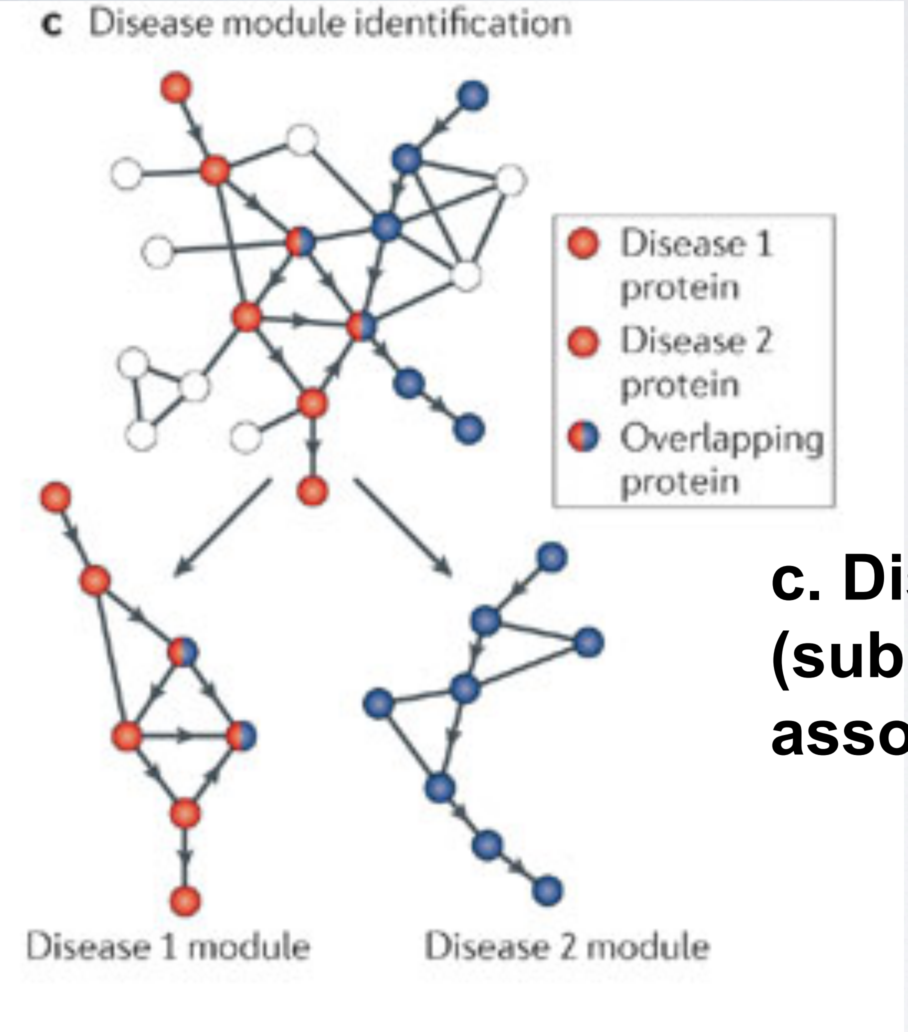
- Proteins associated to disease (esperiments, proteomics)
- Metabolites (esperiments, metabolomics)

Linkage analysis = studio che stabilisce il linkage tra geni.
Genetic linkage = la tendenza dei certi loci o alleli di essere ereditati assieme a causa della loro localizzazione uno vicino ad altro sullo stesso cromosoma.
Loci genetici che sono vicini fisicamente sullo stesso cromosoma tendono a rimanere assieme durante la meiosi e quindi sono geneticamente "linked".

Genome-wide association study (GWA study, o GWAS), noto anche come **whole genome association study** (WGA study, or WGAS) = esaminare il genoma di diversi individui (es malati versus sani) per capire come i geni variano da individuo a individuo.
- Le variazioni nei geni sono associate a diversi tratti (malattia).
- Si fanno tests per single DNA mutations - SNPs

Dicembre 2010: >1200 human GWAS; sono state esaminate e trovate > 200 malattie e tratti e trovati > 4.000 SNPs.

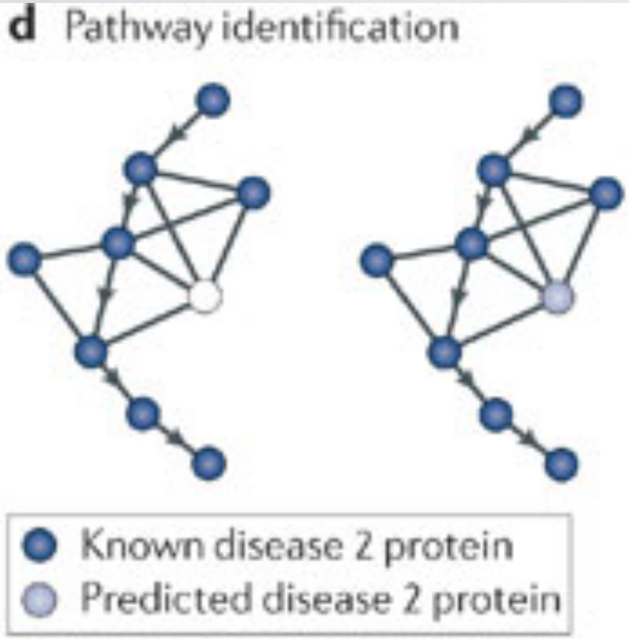
Steps for the identification of disease modules



c. Disease module identification (subnetwork containing nodes associated to the specific disease

clustering tools

Steps for the identification of disease modules



d. Pathway identification (Identificazione di vie molecolari specifiche - quando il modulo è molto grande);
- la loro disfunzione è responsabile della malattia

Si assume che tra i nodi della pathway esiste la shortest path - network "parsimony principle"

Steps for the identification of disease modules

e. Disease module validation

Functional homogeneity

- **Gene ontology (**1) **biological processes**,
  2) **cellular components** and 3) **molecular functions)**

Tissue specificity

Dynamic homogeneity

- **Co-expression**

Other tests:

- **GWA (SNPs identified for the nodes/genes)**

Steps for the identification of disease modules

e. Disease module validation:

Analysis of disease module:

Predictions on:
-Disease genes
-Disease pathways
-Drug targets

Test the predictions (make experiments!):

1. Gene expression data, proteomics, metabolomics, etc
2. Validate molecules in in vitro and in vivo assays
3. Drug targets

Hypotheses and principles of network medicine

1. Hubs

2. Disease module hypothesis

3. Local hypothesis

4. Network parsimony principle

5. Shared components hypothesis

Ipotesi e principi organizzativi della network medicine:

3. Local hypothesis

Se alcuni componenti di un disease network vengono identificati, altri componenti che sono implicati in quella malattia possono essere identificati in base alla "network-vicinity"

Hypotheses and principles of network medicine:

3. Local hypothesis

- Proteins involved in the same disease have the tendency to interact between them

- If a gene (or protein) is involved in a biological process or disease, its direct interactors may have the same role in the process/disease

- Mutations in interacting proteins lead to similar disease phenotypes

- Genes correlated to diseases with similar phenotypes have high probability to interact in a network

Hypotheses and principles of network medicine:

1. Hubs

2. Disease module hypothesis

3. Local hypothesis

4. Network parsimony principle

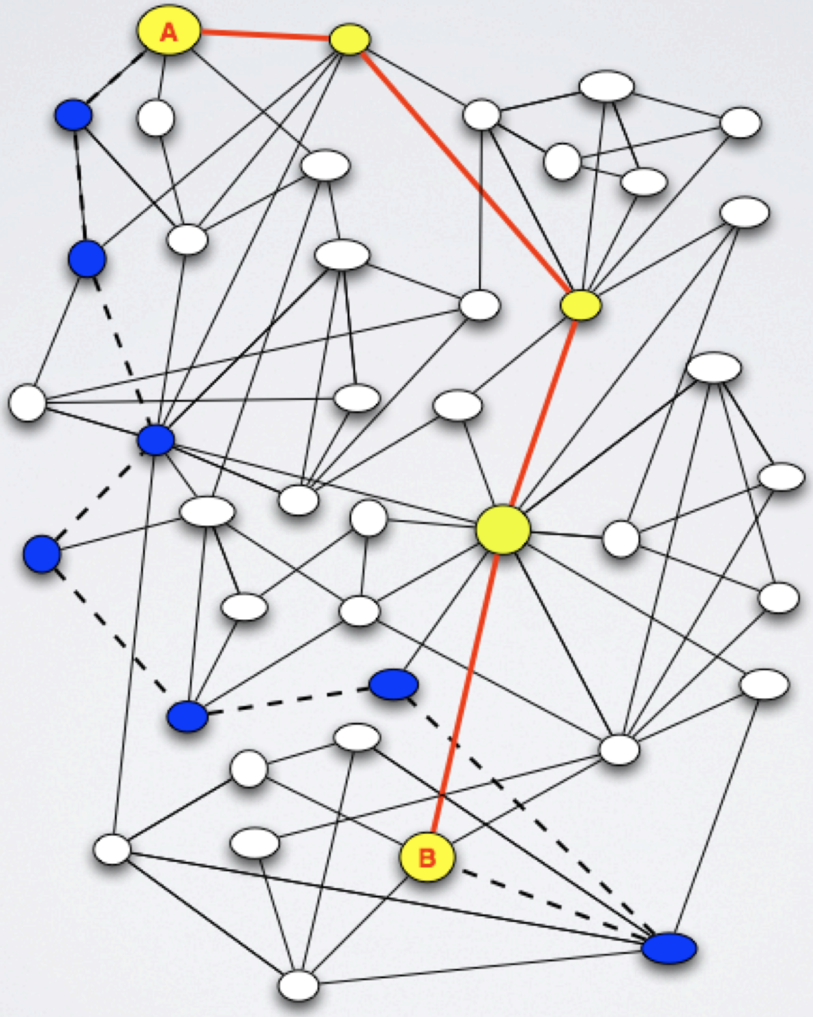
5. Shared components hypothesis

Ipotesi e principi organizzativi della network medicine:

4. Network parsimony principle = The molecular pathways involved in a disease often coincide with the “shortest path” between network components associated to the disease

"Shortest path" = il tragitto minimo fra due nodi

Shortest path A-B = 4

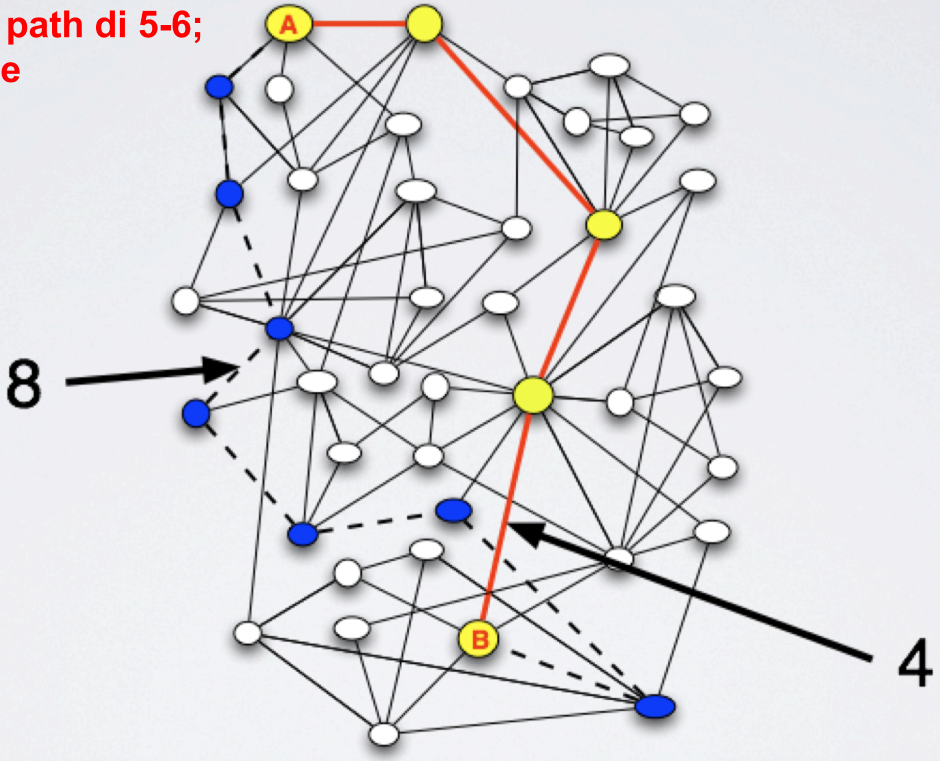


Minimo numero di archi (edges) fra due nodi

"Average shortest path" =
il tragitto minimo "medio" fra tutti i nodi

Misura la "navigabilità" globale della rete,
cioè la facilità con cui si si sposta da un
nodo all'altro

Le reti biologiche
Hanno una average shortest path di 5-6;
sono reti compatte



"Average path length" = distanza media fra i nodi
(non minima);
"Diameter" = shortest path (distanza) massima fra i nodi;
Indici di connettività

Ipotesi e principi organizzativi della network medicine:

5. Shared components hypothesis = Diseases that share cellular components (genes, proteins, metabolites, miRNAs) show phenotipic similarities and comorbidity

Per epigenetica si intende una qualunque attività di regolazione dei geni tramite processi chimici che non comportino cambiamenti nel codice del DNA, ma possono modificare il fenotipo dell’individuo e/o della progenie.

L’epigenetica studia la trasmissione di caratteri ereditari non attribuibili direttamente alla sequenza di DNA.

Malattia

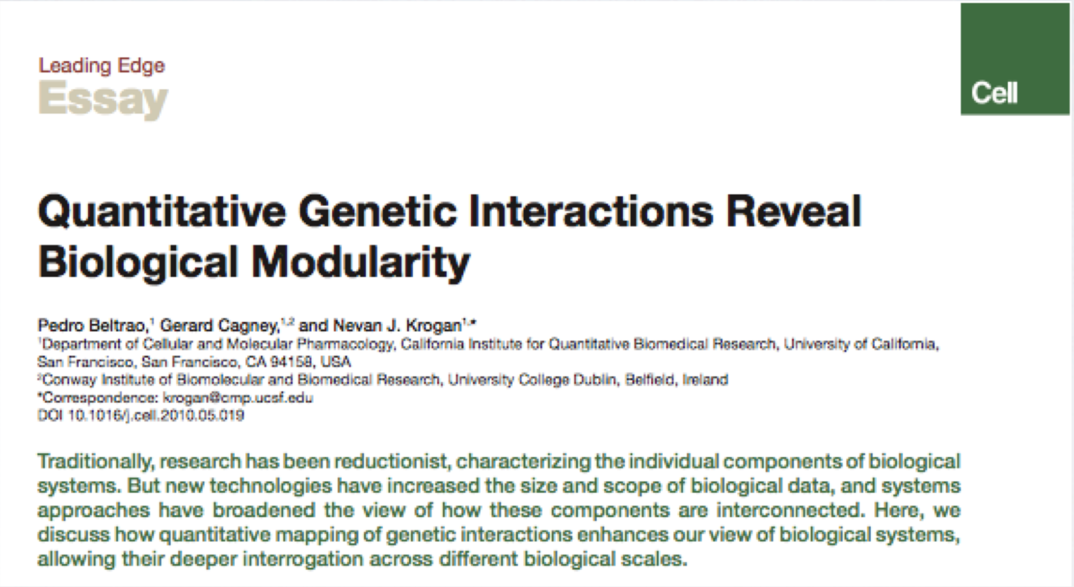
Malattia **localizzata** - riguarda una parte del corpo

Malattia **disseminata** - è estesa ad altre parti del corpo

Malattia **sistemica** - concerne tutto il corpo

Altri tipi di rete

2. Genetic interaction networks - due geni sono linked se il fenotipo del doppio mutante è diverso rispetto al fenotipo dei singoli mutanti.



Leading Edge

Essay

Cell

Quantitative Genetic Interactions Reveal Biological Modularity

Pedro Beltrao,[1] Gerard Cagney,[1,2] and Nevan J. Krogan[1,*]
[1]Department of Cellular and Molecular Pharmacology, California Institute for Quantitative Biomedical Research, University of California, San Francisco, San Francisco, CA 94158, USA
[2]Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Ireland
[*]Correspondence: krogan@cmp.ucsf.edu

DOI 10.1016/j.cell.2010.05.019

Traditionally, research has been reductionist, characterizing the individual components of biological systems. But new technologies have increased the size and scope of biological data, and systems approaches have broadened the view of how these components are interconnected. Here, we discuss how quantitative mapping of genetic interactions enhances our view of biological systems, allowing their deeper interrogation across different biological scales.

Le interazioni genetiche possono essere negative o positive

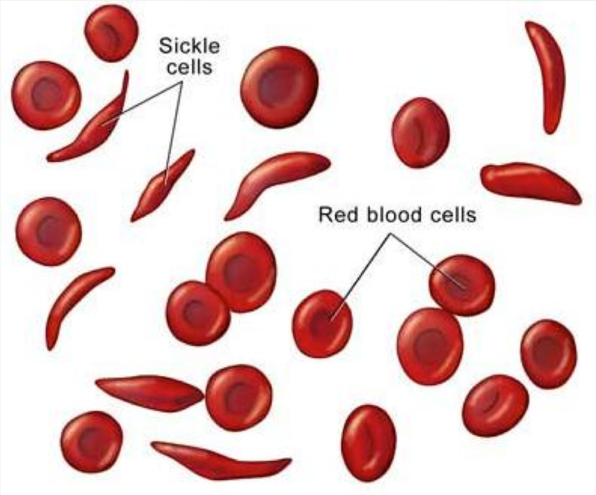
Malattia = risultato di meccanismi combinatoriali ---- diversi difetti e perturbazioni nel modulo di malattia e nella rete ---- diversi fenotipi patologici di malattia

1. Malattie poligeniche
2. Malattie monogeniche

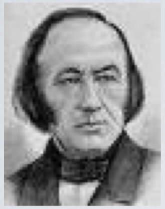
Esempio: l'anemia a cellule falciformi

Mutazione E6V

- anemia severa o moderata
- crisi dolorose
- osteonecrosi
- sindrome polmonaria acuta
- ictus
- ischemia renale
- ittero
- infezioni



OMEOSTASI IN BIOLOGIA



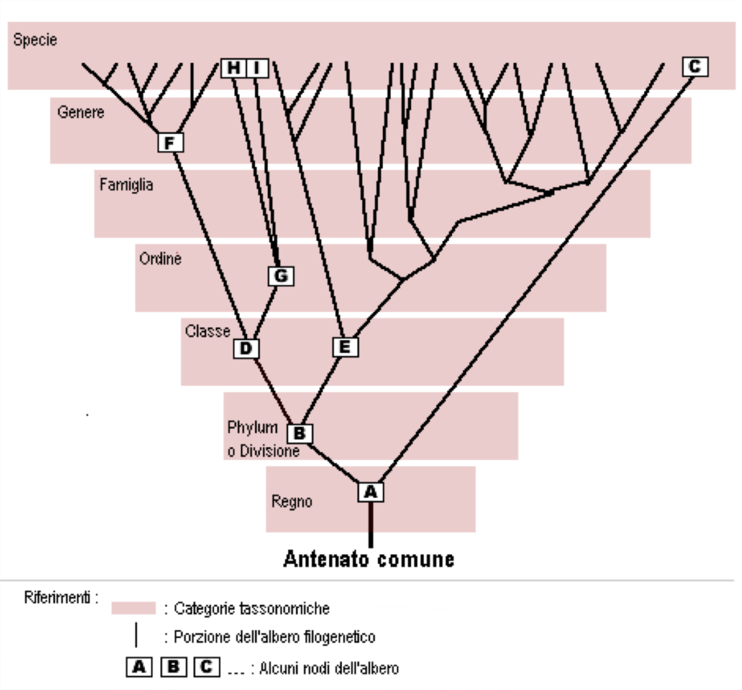
Claude Bernard (1813-1878) - *Un'introduzione allo studio della medicina sperimentale (1865)*
- concetto che la vita si svolge in quanto l'organismo ha la possibilità di **adattare la funzione dei suoi organi e sistemi agli stimoli esogeni in modo da riuscire a <u>mantenere costante il proprio ambiente interno</u>**, "le milieu intérieur"

Walter Bradford Cannon (1871-1945) - *The wisdom of the body (1932)*
- concetto di **omeostasi** nel 1926 (***homoios*** = simile, e ***stasis*** = posizione) - in riferimento alla capacità del corpo di regolare la composizione e il volume del sangue e, di conseguenza, di tutti i fluidi extracellulari in cui sono immerse le cellule

Tassonomia

La **tassonomia** (dal greco, *taxis*, "ordinamento", e *nomos*, "norma" o "regola") è la disciplina della classificazione. La **tassonomia biologica** = i criteri con cui si ordinano gli organismi in un sistema di classificazione composto da una gerarchia di taxa.

In biologia, un **taxon** (plurale **taxa**) o **unità tassonomica**, è un raggruppamento di organismi reali, distinguibili morfologicamente e geneticamente da altri e riconoscibili come unità sistematica, posizionata all'interno della struttura gerarchica della classificazione scientifica.

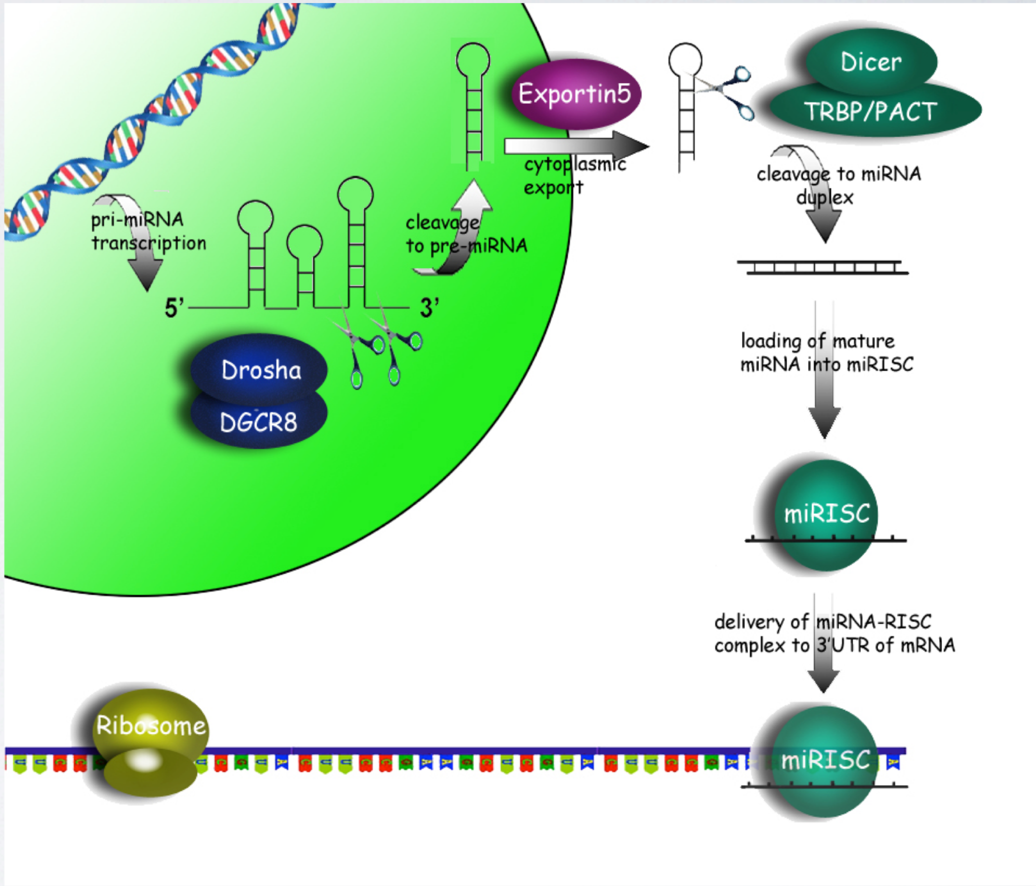


http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy

Cosa sono i microRNAs?

= post-transcriptional regulators

MicroRNA
I **miRNA** sono piccole molecole di RNA, che svolgono diverse funzioni, la più nota attualmente è una regolazione post-trascrizionale. Si legano a sequenze complementari di mRNA e bloccano la traslazione (gene silencing).



Le reti nelle malattie

Considerazioni e limiti

1. I nodi e le interazioni devono essere considerate nel contesto specifico tessutale

2. L'interactoma umano è incompleto e "noisy"

3. Le conoscenze attuali delle reti nelle malattie si riferiscono soprattutto a reti INTRACELLULARI; esiste una mancanza di dati riguardo le reti molecolari che connettono cellule, tessuti e organi

2. Disease modules

Further considerations:

1. Disease modules are not IDENTICAL but have High probability to overlap to topological and functional modules

2. A disease module is defined in relation to a Specific disease; each disease has its own module

3. A gene, a protein or a metabolite can be involved in several disease modules, thus, some disease modules may overlap

Tappe per identificare e validare i moduli di malattia (a-e)

a. Interactome reconstruction

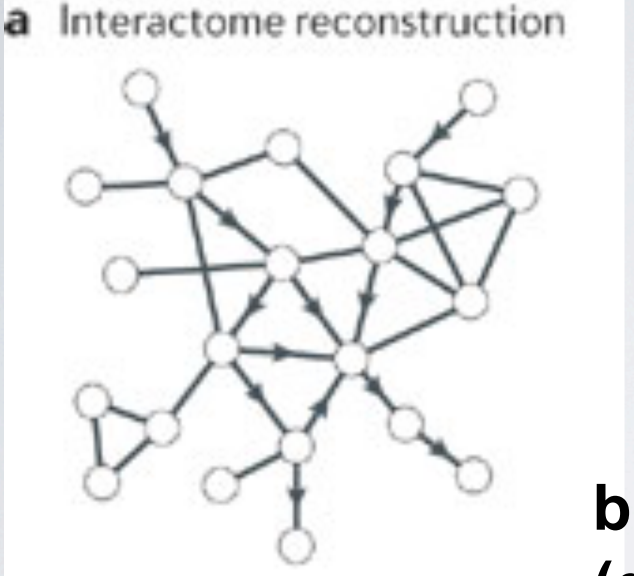
b. Disease gene identification (“disease seed”)

c. Disease module identification

d. Pathway identification (Identificazione di vie molecolari specifiche)

e. Disease module validation

Tappe per identificare e validare i moduli di malattia (a-e)



a. Interactome reconstruction
(della cellula o tessuto d'interesse)

b Diasease gene (seed) identification

Potential sources
- OMIM
- GWA study
- Literature

- Esperimenti

b. Disease gene (seed) identification (geni associati alla malattia ottenuti da linkage analysis, studi genome-wide association (GWA) e altre risorse (letteratura, microarrays)

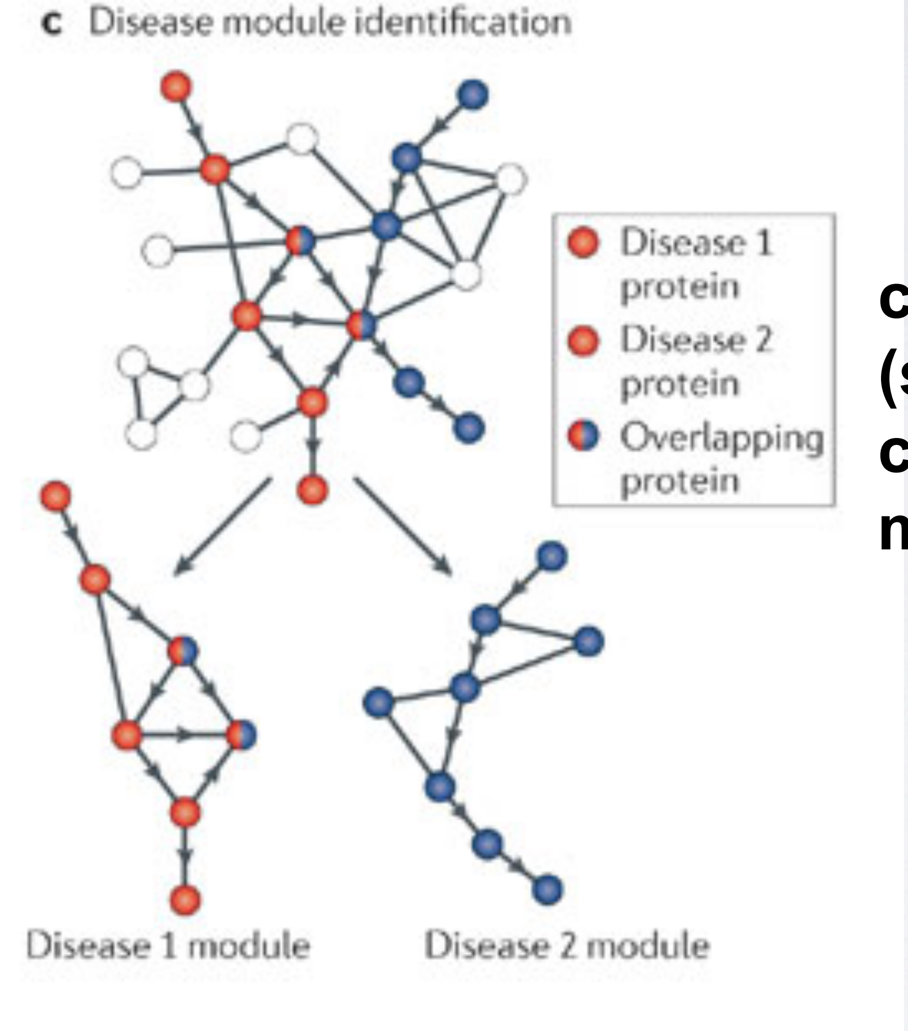
- Proteine associate alla malattia (esperimenti di proteomica)
- Metaboliti (esperimenti di metabolomica)

Linkage analysis = studio che stabilisce il linkage tra geni.
Genetic linkage = la tendenza dei certi loci o alleli di essere ereditati assieme a causa della loro localizzazione uno vicino ad altro sullo stesso cromosoma.
Loci genetici che sono vicini fisicamente sullo stesso cromosoma tendono a rimanere assieme durante la meiosi e quindi sono geneticamente “linked”.

Genome-wide association study (GWA study, o GWAS), noto anche come **whole genome association study** (WGA study, or WGAS) = esaminare il genoma di diversi individui (es malati versus sani) per capire come i geni variano da individuo a individuo.
- Le variazioni nei geni sono associate a diversi tratti (malattia).
- Si fanno tests per single DNA mutations - SNPs

Dicembre 2010: >1200 human GWAS; sono state esaminate e trovate > 200 malattie e tratti e trovati > 4.000 SNPs.

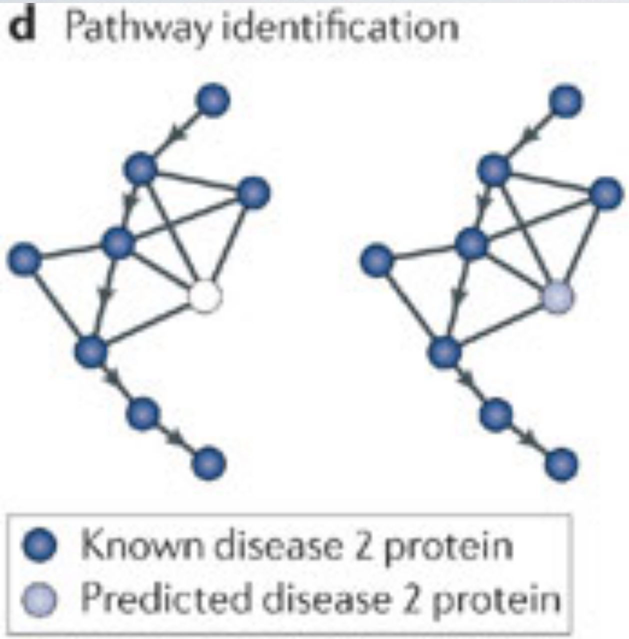
Tappe per l'identificazione e validazione dei moduli di malattia



c. Disease module identification (sottorete che contiene i nodi associati alla malattia).

clustering tools

Tappe per l'identificazione e validazione dei moduli di malattia



d. Pathway identification (Identificazione di vie molecolari specifiche - quando il modulo è molto grande);
- la loro disfunzione è responsabile della malattia

Si assume che tra i nodi della pathway esiste la shortest path - network "parsimony principle"

Tappe per l'identificazione e validazione dei moduli di malattia

e. Disease module validation

Omogeneità funzionale

- **Gene ontology (**1) **biological processes**, 2) **cellular components** and 3) **molecular functions)**
- **Specificità tessutale**

Omogeneità dinamica

- **Co-espressione**

Altri tests:

- **GWA (SNPs nei componenti cellulari predetti)**

Tappe per l’identificazione e validazione dei moduli di malattia

e. Disease module validation:

L’analisi del modulo di malattia:

Predizioni su:
-Disease genes
-Disease pathways
-Drug targets

Testare le predizioni (fare esperimenti!):

1. Gene expression data, proteomics, metabolomics, etc
2. Validare le molecole in saggi in vitro e in vivo
3. Drug targets

Ipotesi e principi organizzativi della network medicine:

1. Hubs

2. Disease module hypothesis

3. Local hypothesis

4. Network parsimony principle

5. Shared components hypothesis

Ipotesi e principi organizzativi della network medicine:

3. Local hypothesis

Se alcuni componenti di un disease network vengono identificati, altri componenti che sono implicati in quella malattia possono essere identificati in base alla "network-vicinity"

Ipotesi e principi organizzativi della network medicine:

3. Local hypothesis

- Le proteine coinvolte nella stessa malattia hanno la tendenza di interagire tra di loro.

- Se un gene (o una proteina, etc) è coinvolta in un processo biologico o malattia, i suoi interratori diretti possono avere lo stesso ruolo nello stesso processo/malattia

- Mutazioni in proteine interagenti spesso portano a fenotipi simili di malattia

- Geni che sono correlati con malattie con fenotipi simili hanno alta probabilità di interagire tra di loro

Ipotesi e principi organizzativi della network medicine:

1. Hubs

2. Disease module hypothesis

3. Local hypothesis

4. Network parsimony principle

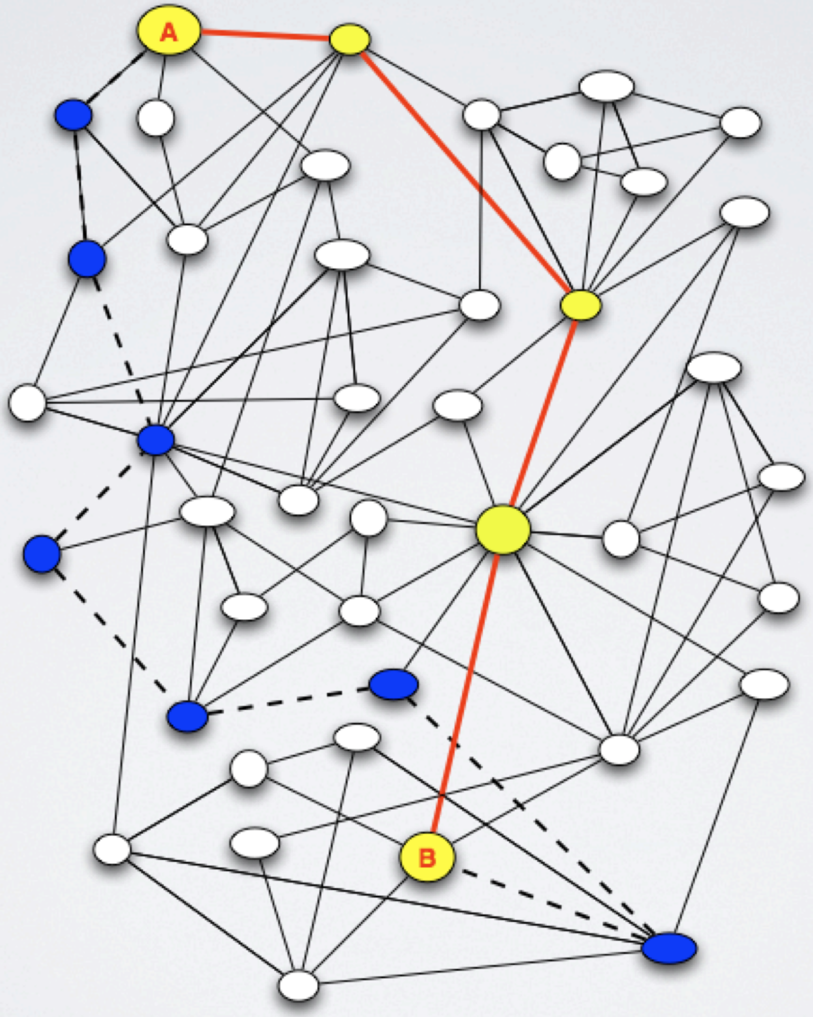
5. Shared components hypothesis

Ipotesi e principi organizzativi della network medicine:

4. Network parsimony principle = Le vie molecolari implicate nella malattia spesso coincidono con le “shortest path” tra i componenti della rete associati alla malattia.

"Shortest path" = il tragitto minimo fra due nodi

Shortest path A-B = 4

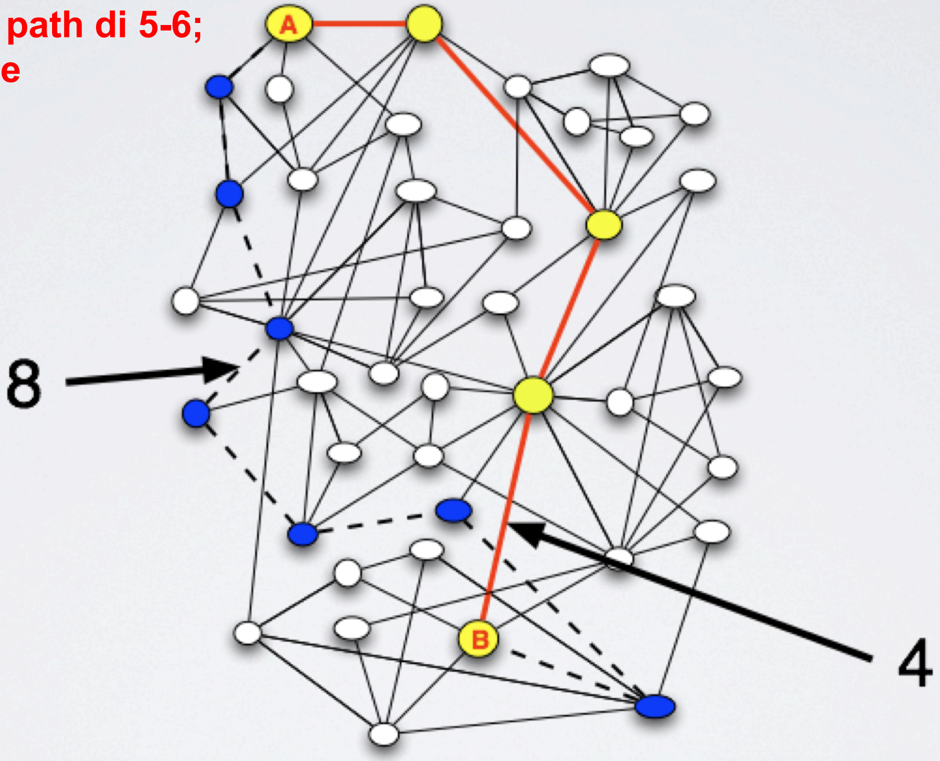


Minimo numero di archi (edges) fra due nodi

"Average shortest path" = il tragitto minimo "medio" fra tutti i nodi

Misura la "navigabilità" globale della rete, cioè la facilità con cui si si sposta da un nodo all'altro

Le reti biologiche
Hanno una average shortest path di 5-6;
sono reti compatte



"Average path length" = distanza media fra i nodi *(non minima);*
"Diameter" = shortest path (distanza) massima fra i nodi;
Indici di connettività

Ipotesi e principi organizzativi della network medicine:

5. **Shared components hypothesis = Le malattie che condividono componenti cellulari (geni, proteine, metaboliti, miRNAs) mostrano similarità fenotipiche e comorbidity**

Methods for disease gene identification

A. Classical Linkage studies
B. GWA
C. Network based tools

C. Network based tools:

1. Linkage methods
2. Disease module-based methods
3. Diffusion-based methods

Linkage analysis = studio che stabilisce il linkage tra geni.
Genetic linkage = Linked genes sit close together on a chromosome, making them likely to be inherited together (la tendenza dei certi loci o alleli di essere ereditati assieme a causa della loro localizzazione uno vicino ad altro sullo stesso cromosoma). Loci genetici che sono vicini fisicamente sullo stesso cromosoma tendono a rimanere assieme durante la meiosi e quindi sono geneticamente "linked".

Genome-wide association study (GWA study, o GWAS), noto anche come **whole genome association study** (WGA study, or WGAS) = esaminare il genoma di diversi individui (es malati versus sani) per capire come i geni variano da individuo a individuo.
- Le variazioni nei geni sono associate a diversi tratti (malattia).
- Si fanno tests per single DNA mutations - SNPs

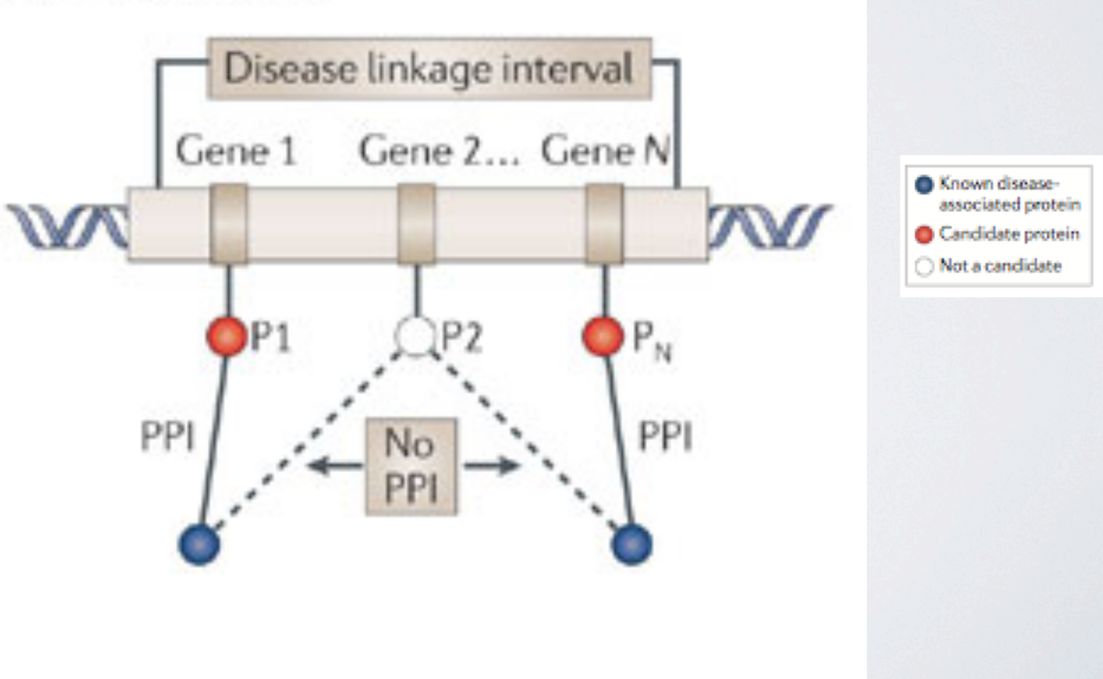
Dicembre 2010: >1200 human GWAS; sono state esaminate e trovate > 200 malattie e tratti e trovati > 4.000 SNPs.

Metodi di identificazione di geni implicati nella malattia

Network based tools:

1. **Linkage methods**

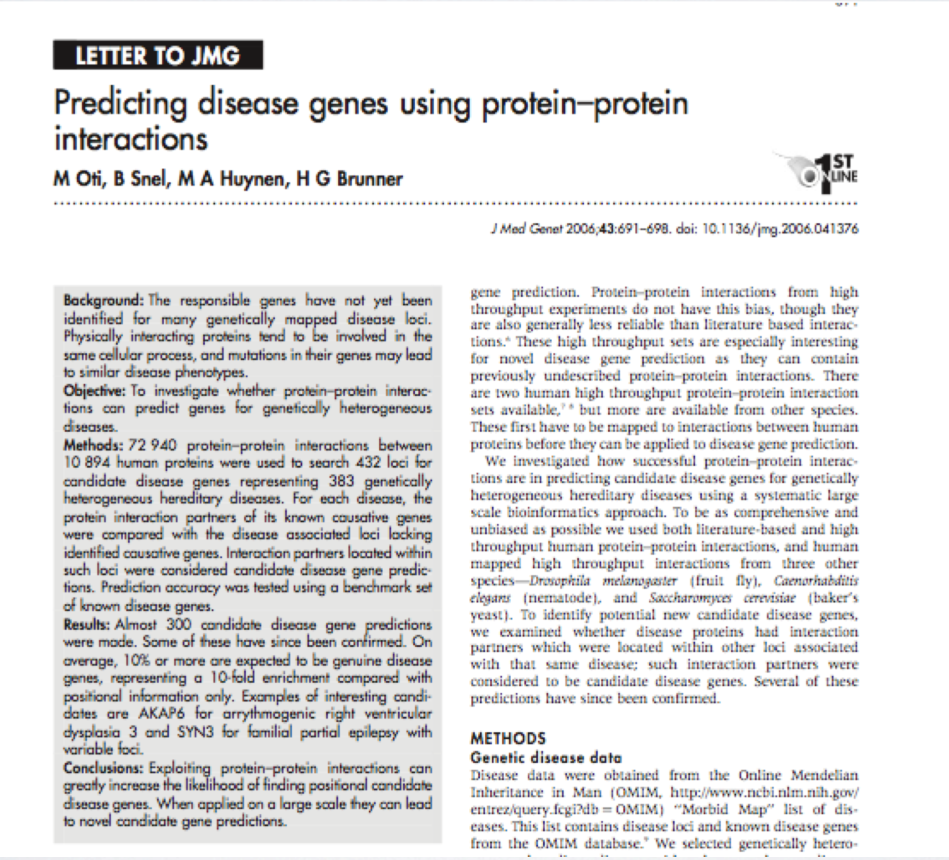
Genes located in the linkage interval of a disease whose protein products (labelled P1, P2, and so on) interact with a known disease-associated protein are considered likely candidate disease genes (assume that the direct interaction partners of a disease protein are likely to be associated with t



1. Linkage methods

Linkage-based methods are based on the assumption that proteins that interact directly tend to be involved in the same cellular process, thus their mutations may lead to similar disease phenotypes.

Le proteine che interagiscono direttamente con una proteina coinvolta in una malattia sono probabilmente associate alla stessa malattia



LETTER TO JMG

Predicting disease genes using protein–protein interactions

M Oti, B Snel, M A Huynen, H G Brunner

J Med Genet 2006;**43**:691–698. doi: 10.1136/jmg.2006.041376

Background: The responsible genes have not yet been identified for many genetically mapped disease loci. Physically interacting proteins tend to be involved in the same cellular process, and mutations in their genes may lead to similar disease phenotypes.

Objective: To investigate whether protein–protein interactions can predict genes for genetically heterogeneous diseases.

Methods: 72 940 protein–protein interactions between 10 894 human proteins were used to search 432 loci for candidate disease genes representing 383 genetically heterogeneous hereditary diseases. For each disease, the protein interaction partners of its known causative genes were compared with the disease associated loci lacking identified causative genes. Interaction partners located within such loci were considered candidate disease gene predictions. Prediction accuracy was tested using a benchmark set of known disease genes.

Results: Almost 300 candidate disease gene predictions were made. Some of these have since been confirmed. On average, 10% or more are expected to be genuine disease genes, representing a 10-fold enrichment compared with positional information only. Examples of interesting candidates are AKAP6 for arrythmogenic right ventricular dysplasia 3 and SYN3 for familial partial epilepsy with variable foci.

Conclusions: Exploiting protein–protein interactions can greatly increase the likelihood of finding positional candidate disease genes. When applied on a large scale they can lead to novel candidate gene predictions.

gene prediction. Protein–protein interactions from high throughput experiments do not have this bias, though they are also generally less reliable than literature based interactions.[6] These high throughput sets are especially interesting for novel disease gene prediction as they can contain previously undescribed protein–protein interactions. There are two human high throughput protein–protein interaction sets available,[7 8] but more are available from other species. These first have to be mapped to interactions between human proteins before they can be applied to disease gene prediction.

We investigated how successful protein–protein interactions are in predicting candidate disease genes for genetically heterogeneous hereditary diseases using a systematic large scale bioinformatics approach. To be as comprehensive and unbiased as possible we used both literature-based and high throughput human protein–protein interactions, and human mapped high throughput interactions from three other species—*Drosophila melanogaster* (fruit fly), *Caenorhabditis elegans* (nematode), and *Saccharomyces cerevisiae* (baker's yeast). To identify potential new candidate disease genes, we examined whether disease proteins had interaction partners which were located within other loci associated with that same disease; such interaction partners were considered to be candidate disease genes. Several of these predictions have since been confirmed.

METHODS

Genetic disease data

Disease data were obtained from the Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = OMIM) "Morbid Map" list of diseases. This list contains disease loci and known disease genes from the OMIM database.[9] We selected genetically hetero-

Metodi di identificazione dei geni implicati nella malattia

2. Disease module-based methods

a. Interactome reconstruction

b. Disease gene identification (“disease seed”)

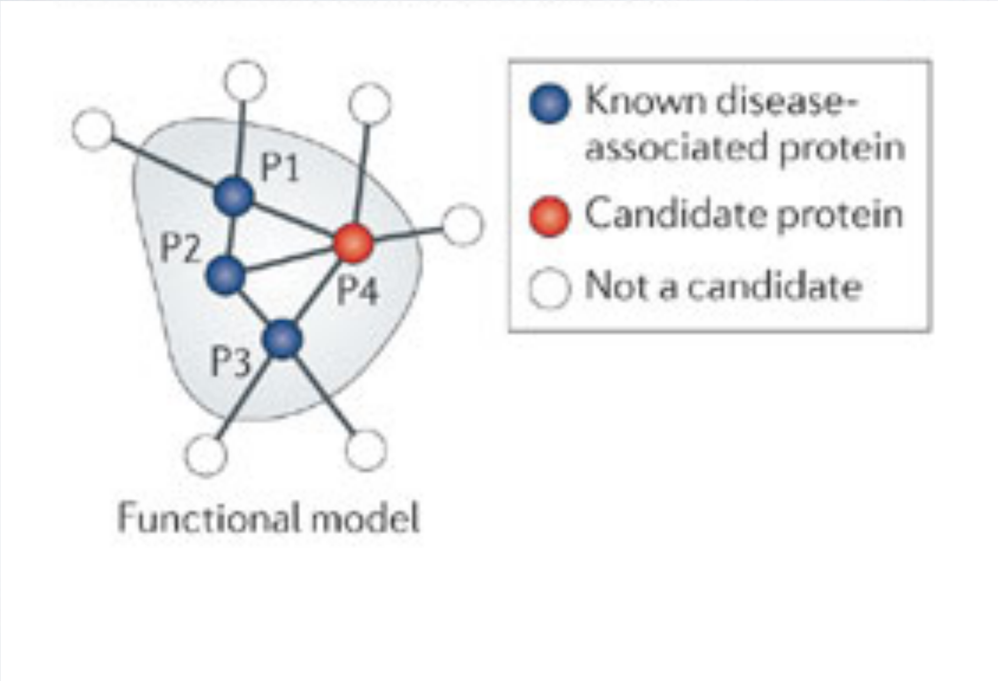
c. Disease module identification

d. Pathway identification (Identificazione di vie molecolari specifiche)

e. Disease module validation

Metodi di identificazione dei geni implicati nella malattia

2. **Disease module-based methods**
   Tutti i componenti cellulari che appartengono allo stesso modulo funzionale o di malattia hanno probabilità di essere coinvolti nella stessa malattia



Tuttavia:

1. Le interazioni cellulari delle molecole coinvolte nelle malattie sono poco conosciute

2. Sono necessari altri esperimenti

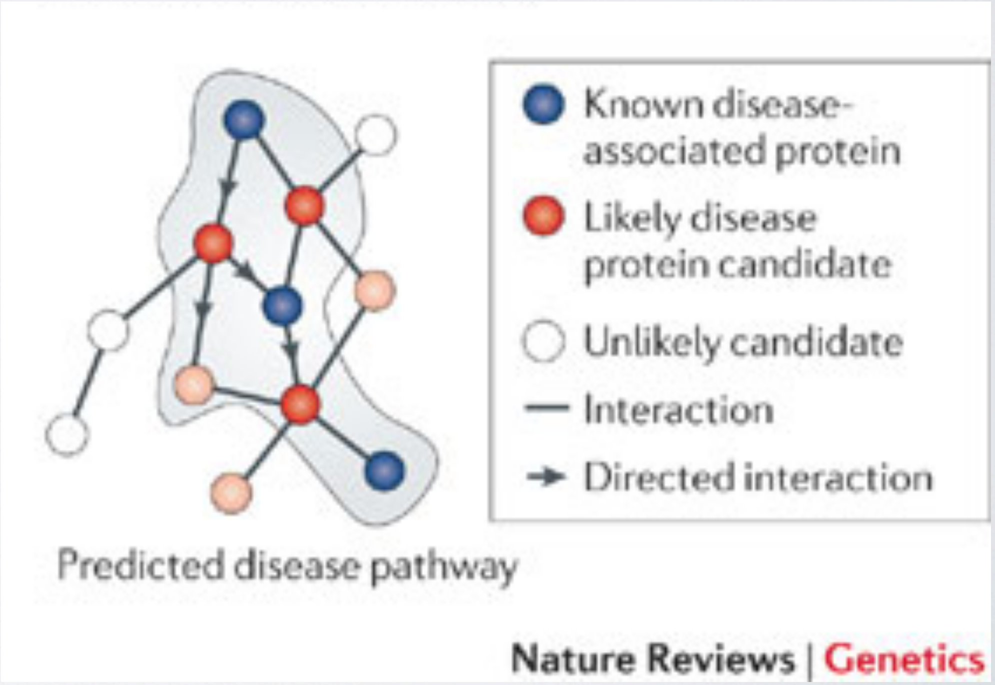
Metodi di identificazione dei geni implicati nella malattia

3. Diffusion-based methods

Starting from proteins that are known to be associated with a disease, a random walker visits each node in the interactome with a certain probability.
The outcome of this algorithm is a disease-association score that is assigned to each protein, that is, the likelihood that a particular protein is associated with the disease.

Prioritization algoritms:

1) **Random walk**
2) **CIPHER**
3) **PRINCE**


● Known disease-associated protein
● Likely disease protein candidate
○ Unlikely candidate
— Interaction
→ Directed interaction

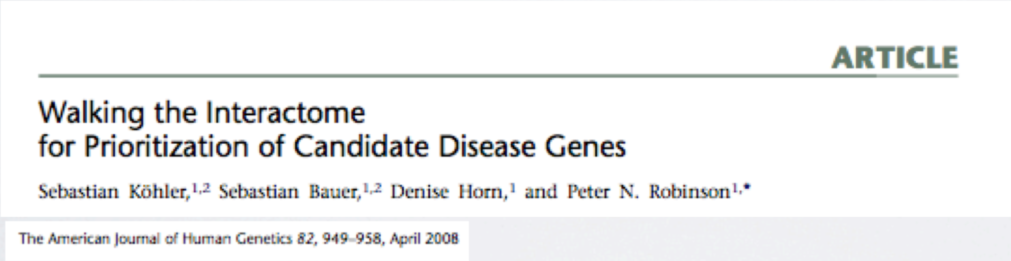
Predicted disease pathway

Nature Reviews | Genetics

Metodi di identificazione dei geni implicati nella malattia

3. Diffusion-based methods

Algoritmi "Random walk" e "PRINCE"



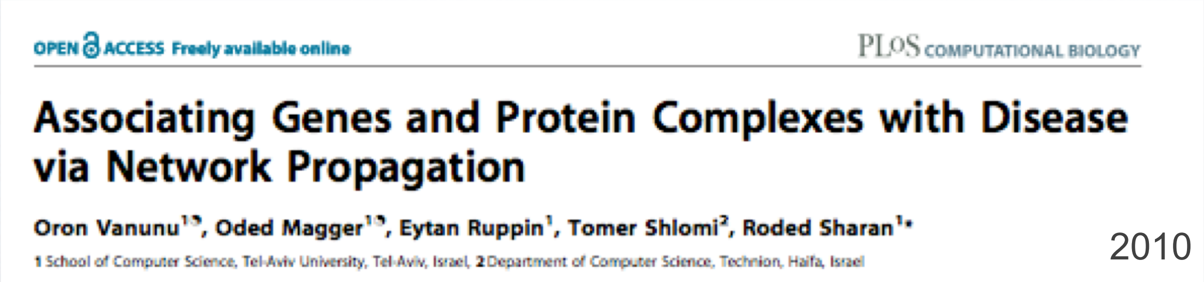
ARTICLE

Walking the Interactome
for Prioritization of Candidate Disease Genes

Sebastian Köhler,[1,2] Sebastian Bauer,[1,2] Denise Horn,[1] and Peter N. Robinson[1,*]

The American Journal of Human Genetics 82, 949–958, April 2008

<u>Prioritize proteins and interactions</u> on the basis of their potential involvement in the particular disease



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Associating Genes and Protein Complexes with Disease via Network Propagation

Oron Vanunu[1,9], Oded Magger[1,9], Eytan Ruppin[1], Tomer Shlomi[2], Roded Sharan[1*]

1 School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel, 2 Department of Computer Science, Technion, Haifa, Israel

2010

Metodi di identificazione dei geni implicati nella malattia

3. Diffusion-based methods

Algoritmo Random walk
In matematica, la passeggiata aleatoria è la formalizzazione dell'idea di prendere passi successivi in direzioni casuali (**succession of random steps in random/casual direction**)

(Kohler et al. : il metodo è meglio degli algoritmi basati sulle interazioni dirette oppure shortest path)

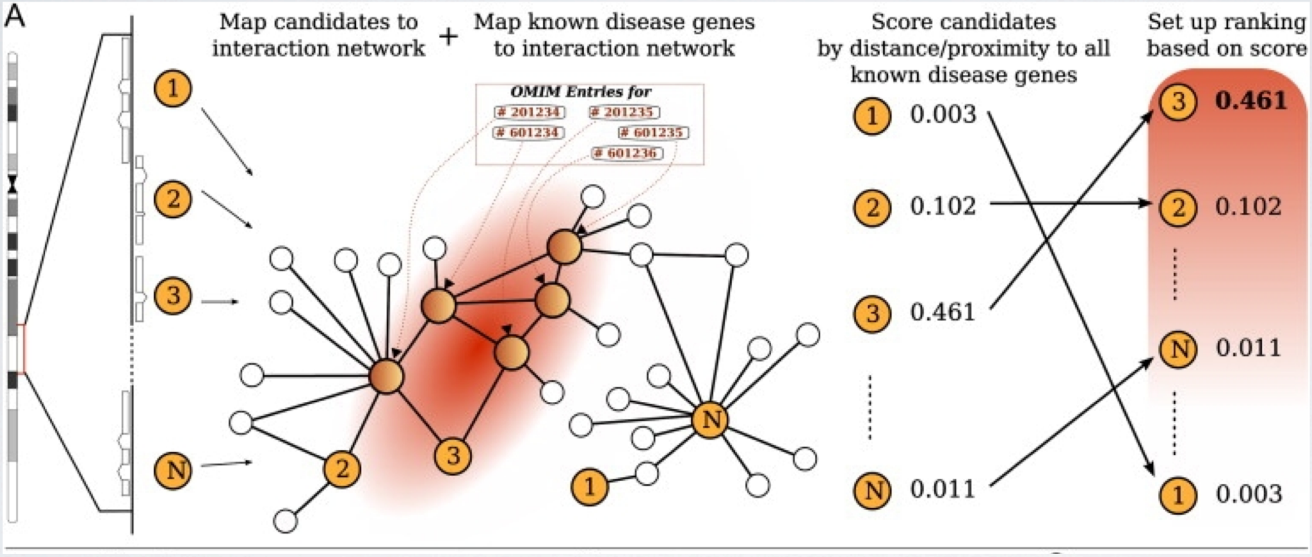
Studio effettuato da *Kohler et al., 2008*:

1.Input: 110 disease genes da OMIM più altri geni correlati (literature, databases) per un totale di **783** geni
2.Per ogni gene sono stati considerati circa 100 geni candidati più vicini nel loro linkage interval
3. PPI network (PPI verificate sperimentalmente o predette): grafo non-direzionato ottenuto con: HPRD, BIND, BioGrid, IntACT, DIP, STRING.
4. Random walk algoritm per identificare nuovi disease genes

Metodi di identificazione dei geni implicati nella malattia

3. Diffusion-based methods

Random walk algoritm


A
Map candidates to interaction network
+
Map known disease genes to interaction network
OMIM Entries for
# 201234
# 201235
# 601234
# 601235
# 601236
1
2
3
N
Score candidates by distance/proximity to all known disease genes
1 0.003
2 0.102
3 0.461
N 0.011
Set up ranking based on score
3 0.461
2 0.102
N 0.011
1 0.003

Prioritizzazione dei geni (Disease-Gene Prioritization)

1. Tutti i geni candidati contenuti nel linkage interval sono stati mappati nella PPI

2. Score a ogni candidato in base alla localizzazione della proteina in relazione ai prodotti genici coinvolti nelle malattie usando misure di distanza sul global network (random walk).

3. Rank per i geni nel linkage interval in base allo score per definire una lista di priorità (prioritizzare i geni)

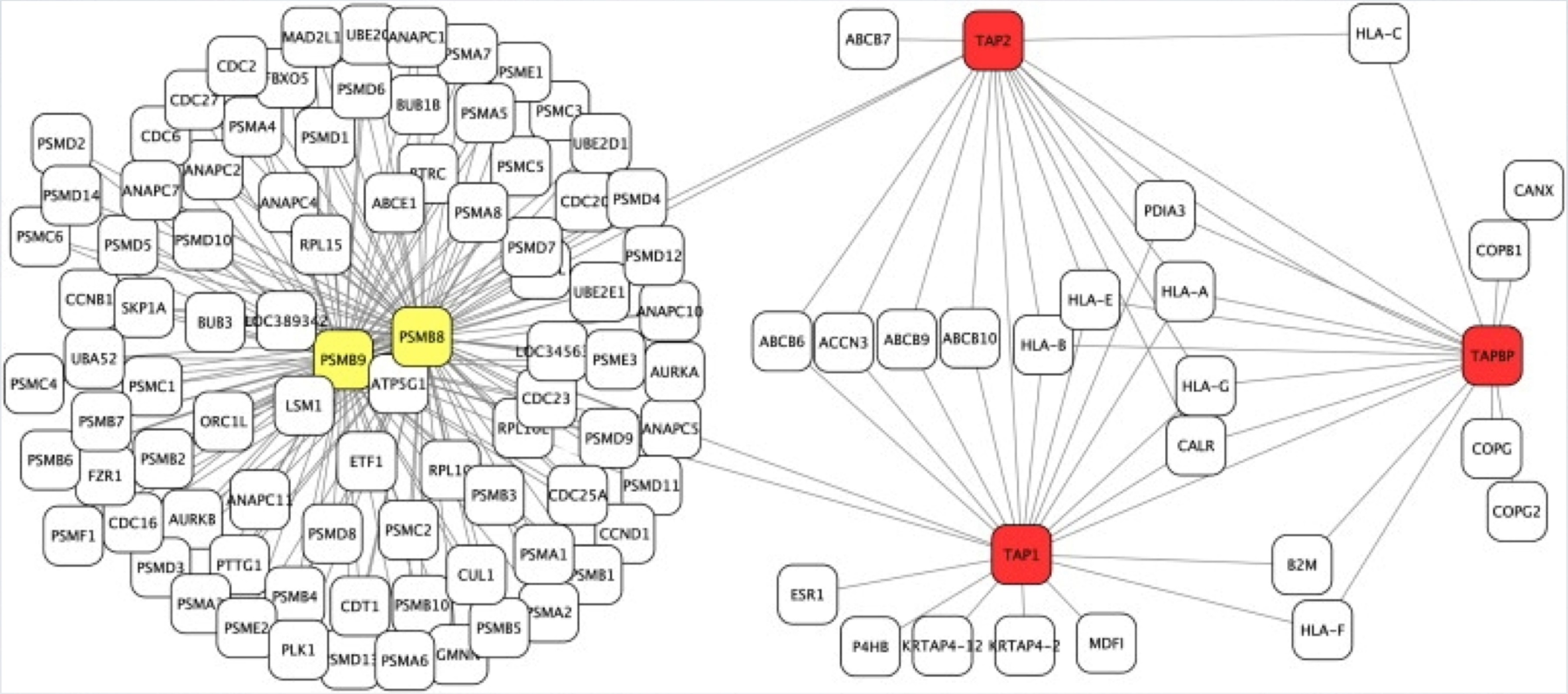
da validare in futuri esperimenti.

Kohler S et al., The American Journal of Human Genetics 82, 949-958, April 2008

Metodi di identificazione dei geni implicati nella malattia

3. Diffusion-based methods

Random walk algoritm



Kohler S et al., The American Journal of Human Genetics 82, 949-958, April 2008

Sindrome linfocitaria di Bare – i geni noti e i geni del linkage interval sono stati mappati nel PPI Network
Il PPI associato alla sindrome di Bare tipo 1, comprende i geni TAP1, TAP2, and TAPBP (in rosso). Il metodo ha identificato i geni PSMB8 e PSMB9 (giallo) come potenziali disease genes nello stesso linkage interval.

http://www.ncbi.nlm.nih.gov/omim

Metodi di identificazione dei geni implicati nella malattia

3. Diffusion-based methods

PRINCE (PRIoritizatioN and Complex Elucidation)- metodo basato sul network globale che prioritizza disease genes e “inferisce” associazioni proteiche complesse.
Input: phenotypic similarities between diseases and **PPI networks (HPRD)**
- Usa un algoritmo iterativo “propagation-based” **(iterative process)**

Starting principle: GENES CAUSING THE SAME OR SIMILAR DISEASES TEND TO LIE CLOSE TO ONE ANOTHER IN A NETWORK OF PROTEIN-PROTEIN OR FUNCTIONAL INTERACTIONS.

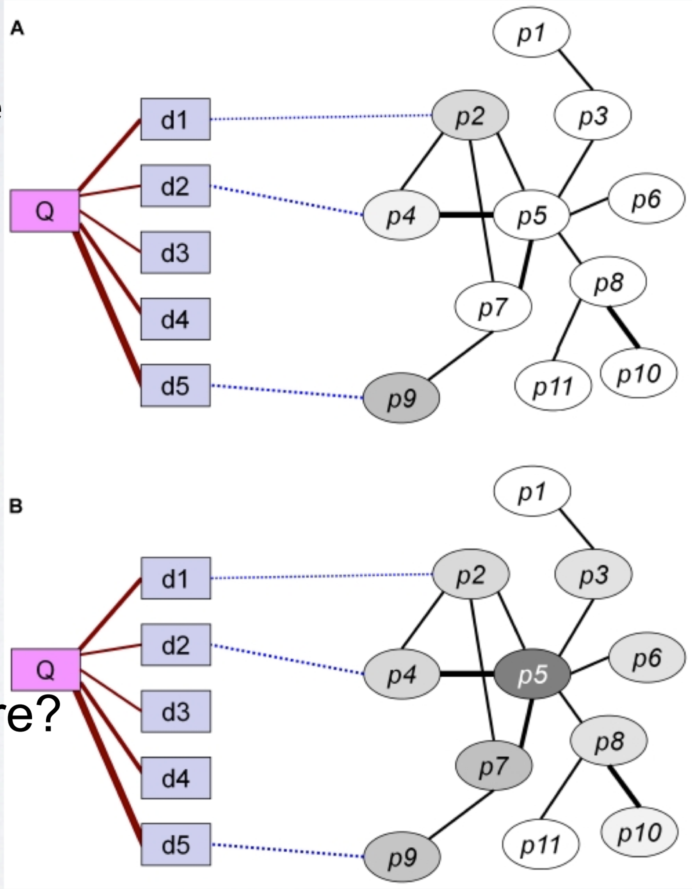
Metodi di identificazione dei geni implicati nella malattia

3. Diffusion-based methods

PRINCE (PRIoritizatioN and Complex Elucidation)

Query disease: Q
d1-d5 malattie che presentano similarità fenotipiche
p2, p4, p9 = proteine codificate dai
disease genes
p1-11 set di proteine del PPI network dello stesso
intervallo di linkage

A= il flusso dopo prima iterazione (info iniziale)
B= lo score dopo varie iterazioni; lo score di
ogni proteina (intensità colore) in base al flusso
che converge su ogni proteina
- quale p ha lo score più alto ed è il candidato migliore?

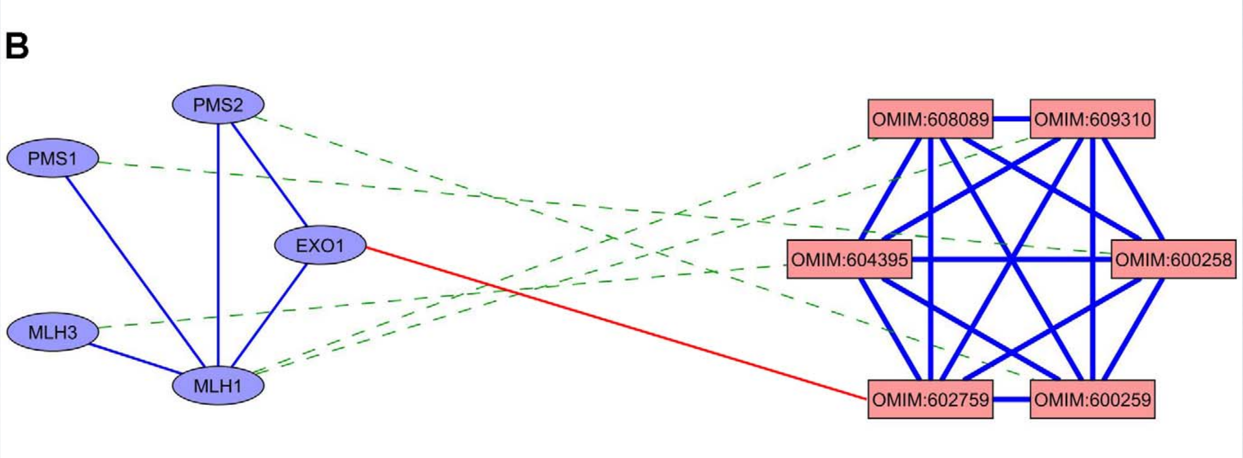


Vanunu O. et al., PLOS Comput. Biol. 2010

Complessi inferiti con PRINCE

Query

Hereditary Prostate Cancer type 8 (HPC8)



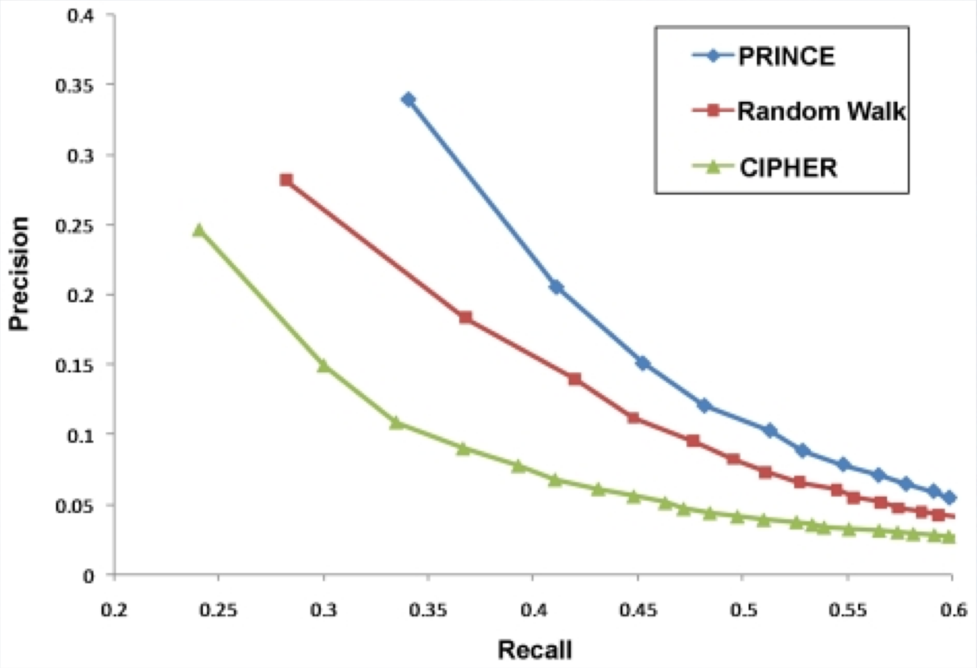
OMIM 602759 (1q42.2-q43); EXO(exonuclease)1 (1q42-q43)

Circular nodes represent proteins and their connecting edges represent protein-protein interactions. Diseases are denoted by square nodes, connected by phenotypic similarity edges. Green dashed edges represent known gene-disease associations; red edges connect a disease to a gene that lies within its associated genomic interval.

- The gene coding for EXO1 is located at genetic locus 1q43, which lies within the region associated with HPC8 (1q42.2-q43).
- EXO1 was ranked first by PRINCE in this interval. In this case, the inferred protein complex provides support also to the prediction that <u>EXO1 is a causal gene for prostate cancer.</u>

Vanunu O. et al., PLOS Comput. Biol. 2010

Performance degli algoritmi di prioritizzazione



Vanunu O. et al., PLOS Comput. Biol. 2010

Performance comparison for PRINCE, Random Walk and CIPHER in a ***leave-one-out cross-validation test*** over **1,369 diseases with a known causal gene**. The figure shows recall versus precision when considering the top proteins for various values of k.

Metodi di identificazione dei geni implicati nella malattia

1. Linkage methods (*interazioni dirette; ipotesi locale*)
2. Disease module-based methods(*neighborhood di disease genes; ipotesi del modulo di malattia*)
3. Diffusion-based methods (*intera topologia della rete; parsimony principle*)