Review

Translational disease interpretation with molecular networks

Anaïs Baudot*, Gonzalo Gómez-López* and Alfonso Valencia

*These authors contributed equally to this work

Address: Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro 3, E-28029 Madrid, Spain.

Correspondence: Alfonso Valencia. Email: avalencia@cnio.es

Published: 29 June 2009

Genome Biology 2009, **10:**221 (doi:10.1186/gb-2009-10-6-221)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/6/221

Abstract

Molecular networks are being used to reconcile genotypes and phenotypes by integrating medical information. In this context, networks will be instrumental for the interpretation of disease at the personalized medicine level.

Genes and proteins do not function in isolation in the cell, but are integrated into a global network of interactions between cellular components. Even if current networks mainly describe protein-protein interactions, other biological relations, including gene regulation, control by small RNAs, enzymatic reactions and other interactions, are progressively being integrated. The complete network of interactions, along with addition of the fundamental dimensions of time and space, will ultimately provide a complete picture of cellular functions.

As phenotypic disorders can arise from abnormalities in genes, knowing the functions of the corresponding proteins can provide clues to understanding the molecular basis of disease, especially of complex diseases such as diabetes and cancer. High-throughput genomic analyses have been applied to study these complex multifactorial diseases. They produce a tremendous amount of raw data that are, however, difficult to interpret due, for instance, to problems of reproducibility, functional interpretation and statistical shortcomings, which have often led to controversial findings [1]. To better interpret such high-throughput genomic experiments, ways of integrating network information - for example, on protein-protein interactions - have been developed.

We will first discuss how mapping disease genes or proteins into their corresponding interaction networks can facilitate the study of their cellular functions. We will then consider the use of network analysis and bioinformatics to integrate high-throughput information on networks of interactions to better understand the functional cellular defects underlying complex multifactorial diseases. Finally, we consider how molecular networks could be used to link disease genotypes and phenotypes, and propose the use of networks to integrate scattered information - connecting genomic knowledge, detailed molecular information and precise medical descriptions of diseases, and ultimately taking into account an individual's genetic background to provide effective personalized medicine.

Unraveling disease from a network perspective

A large number of gene variants are known to cause phenotypic disorders in humans. The Online Mendelian Inheritance in Man (OMIM) database [2] stores information on more than 2,000 genes related to such disorders. These disease-causing genes have been historically identified by linkage analysis of affected families and mutational screening. When the relationship between a particular disease and a small set of gene variants (or a single variation) is well characterized, protein functions can then be deciphered to provide direct insight into the molecular basis and progression of the disease, and, ultimately, to identify valid targets for therapy. For instance, the identification of the enzyme deficiency responsible for the metabolic disease

phenylketonuria, which causes mental retardation, led to the adoption of a specialized diet that reduces the impact of the gene defect.

Functionally similar proteins tend to be connected in molecular networks - for instance, by being involved in the same molecular complexes [3]. Therefore, the analysis of the network surrounding disease proteins can provide clues about their functional roles in the cell. This assumption was behind an interaction screen for the poorly understood huntingtin protein, in which a polyglutamine tract expansion induces Huntington's disease. A number of protein partners related to transcriptional regulation and DNA maintenance were identified, predicting the involvement of huntingtin in these processes [4]. Similar studies have constructed molecular networks around other known disease genes, such as ataxia-causing genes [5], and even around virus proteins to pick up their interactions with host proteins and reveal a host-pathogen hybrid protein-interaction network [6]. Overall, deciphering the molecular networks surrounding disease proteins might reveal pathogenic mechanisms, new candidate disease proteins and modifiers of phenotype, and so expand the list of potential therapeutic targets [4,5] and the possibility of multi-targeted therapy [7].

As interacting proteins are functionally close, one can hypothesize that mutations in linked genes might lead to similar clinical manifestations or phenotypes. A bioinformatics study in yeast showed that among many possible functional links (for example, gene interactions, gene coexpression, co-citation in the literature), stable protein interactions, and in particular protein complexes, are the best predictors of phenotypic similarities in growth rates [8]. In humans, the inherited ataxias, a set of neurodegenerative disorders manifested by a loss of movement coordination and sharing some phenotypic traits, have also been studied through a protein-interaction approach. The deciphering of the protein-interaction network around genes already known to be directly involved in more than 20 inherited ataxias shows that most of the corresponding proteins interact with each other, either directly or indirectly [5]. Hence, ataxia-causing genes are functionally related at the cellular level (for example, the corresponding proteins interact or participate in the same complex).

Obviously, this wealth of information about the molecular basis of diseases could not have been reached by studying the functions of isolated proteins. Altogether, these results show that disorders with similar phenotypes may be the consequence of mutation in genes that are related by their cellular function. This conclusion is complemented by the finding that a 'disease network', in which two genes are related if they are known to be responsible for the same disease, overlaps significantly with a protein-interaction network [9]. Molecular networks, and in particular protein-interaction networks, could thus provide a valuable framework for relating genotypes and disease phenotypes.

The link between protein interactions and phenotypic similarities can also be exploited to predict new candidate disease proteins; mutations of proteins in the network neighborhood of a disease-causing protein are more likely to cause a similar disorder. An integrated network of gene coexpression combined with other high-throughput datasets (for example, direct protein-protein interactions, membership of protein complexes, genetic interactions) has been constructed around four known breast cancer proteins in order to obtain insights into cancer mechanisms and to identify new cancer-associated proteins. The hyaluronan-mediated motility receptor (HMMR), a protein that may be involved in centrosome function, was found to be closely linked in this integrated network to one of these cancer genes, *BRCA1*, and thus is predicted to have a role in breast cancer [10].

Similar prediction methods can also be applied to lists of candidate genes - for instance, the genes in a disease locus identified by linkage analysis of cancer-prone families. If one of the genes mapped to the locus interacts with a protein known to cause the disease, then it is predicted as the best disease candidate [11]. This principle can be refined by comparing the disease phenotypes induced by the different proteins of the complex containing the disease candidate [12], or by computing a correlation between phenotype similarities and closeness - a measurement of topological proximity in the molecular-interaction network [13].

All the methods described above rely on previously known disease-causing genes, either to study their cellular functions in the cell or to predict other genes that will lead to similar phenotypes when mutated. However, complex disorders cannot be adequately described as lists of implicated genes and require different conceptual and technical approaches.

From high-throughput data to networks for complex diseases

The importance of analyzing information in terms of networks is most obvious for the study of complex diseases, such as cancer or diabetes, in which illness is caused by the combined actions of multiple genes, the individual's genetic background and environmental factors. The frequency and penetrance of complex diseases vary greatly among individuals. For instance, mutations in slightly different sets of genes can converge onto similar phenotypes, whereas the same set of mutated genes can lead to significant phenotypic differences in different individuals. Furthermore, many mutated genes show very little effect independently, but behave cooperatively to predispose to disease, a phenomenon called epistasis. Deciphering the impact of epistatis on complex disease phenotypes represents a current challenge in human genetics [14].

Table 1

Information for complex diseases provided by high-throughput projects and gene variation databases

(a) Disease	Project	Reference
Cancer	Cancer Genome Project	[70]
	The Cancer Genome Atlas	[16]
	Cancer Genome Anatomy Project	[17]
	The International Cancer Genome Consortium	[71]
	Cancer Genetic Markers Susceptibility	[72]
Diabetes	Diabetes Genome Anatomy Project	[18]
Alzheimer's disease	Alzheimer's Genome Project	[73]
Autism	The Autism Genome Project	[19]
Schizophrenia	The Schizophrenia Genome Project	[74]
(b) Variation	Database	
Polymorphisms	HapMap	[75]
Polymorphisms	HGVMap	[76]
Cancer mutations	Cosmic	[77]
Genome-wide association studies	Genome-wide association studies catalog	[78]

Despite their huge impact on public health and massive investment in research, the causes, progression, and mechanisms of complex disorders and the impact of treatments on them still remain largely unknown [15]. Multidisciplinary projects based on high-throughput genomic analyses (including massive sequencing, genotyping, transcriptomic and proteomic experiments) have been launched to study common complex diseases (Table 1a). They include cancer (for example, the Cancer Genome Atlas [16] and the Cancer Genome Anatomy Project [17]); diabetes (the Diabetes Genome Anatomy Project [18]); and autism (the Autism Genome Project Consortium [19]). Such high-throughput studies aim first at elucidating the causal genetic mechanisms of diseases by examining different genetic characteristics in a large number of sick and healthy individuals (for example, gene mutations, chromosomal abnormalities, or copy-number variation).

Disease loci can be identified in the first instance by high-throughput linkage analysis of disease-prone families, an approach that has been applied, for example, to autism [20] and schizophrenia [21]. For autism, linkage analysis in more than 1,400 families highlighted the chromosomal region 11p12-p13 and neurexin, a protein involved in synaptogenesis, as candidate loci [20]. Disease-associated loci can also be identified by whole-genome association studies, which systematically assay for genetic variation such as single nucleotide polymorphisms (SNPs) across the genome [22]. This type of association study can be applied to both affected and healthy cohorts, or in relation to particular phenotypes, such as disease susceptibility (for example, diabetes [23]), or to study individual responses to drugs. Finally, genetic variations can be identified through comprehensive resequencing studies. This approach has been applied to identifying cancer-related mutations in colon and breast tumors, leading to the identification of around 80 DNA alterations in a typical cancer [24]. A number of databases provide information on genetic variations associated with disease (Table 1b).

Complementary high-throughput studies, commonly called functional genomic experiments, aim to go beyond the identification of variants and regions associated with disease phenotypes; they intend to decipher the molecular processes underlying illness. They can, for example, assess gene expression through transcriptomic approaches [25] or use proteomics to assay for the presence of the corresponding proteins in cellular fractions, and so gain information about protein activity and localization [26].

In most cases, high-throughput approaches to complex diseases do not provide lists of directly altered genes or proteins but genomic and proteomic information for groups of genes that are likely to be related to the pathology under study. Cancer gene-expression profiling illustrates this well, as numerous microarray-based studies have proposed gene markers, or signatures, related to clinical phenotypes (for example, metastatic capability or survival rates): for instance, a six-gene signature involving proteins mainly functioning in cell adhesion and/or signal transduction has

recently been implicated in the prediction of breast cancer metastasis into the lung [25]. However, such experiments are barely reproducible, leading to inconsistencies in signatures between different experiments and, more importantly, they do not reveal the underlying molecular mechanisms accounting for the signatures.

In such high-throughput experiments, the molecular mechanisms are typically analyzed through functional bioinformatics analysis, mainly based on Gene Ontology (GO) annotations of proteins (for example, FatiGO [27]), which can highlight molecular processes shared by the genes in a disease signature. However, this approach has several shortcomings: nonspecific terms tend to be overrepresented (for example, 'extracellular matrix', 'cell communication' and 'cell growth' in the invasive front of colorectal metastasis [28]), interesting proteins can be superficially annotated, and GO can lack direct associations with pathways and disease. In view of these limitations, some authors have proposed strategies focused on *a priori* defined gene sets (for example, gene-set enrichment analysis [29]), such as genes belonging to a particular signaling pathway, that search for global trends in their expression levels - for example, all the genes are upregulated in a given disease. A recent high-throughput resequencing study for human pancreatic cancer revealed a shift from a gene-centric view, with the identification of many genetic alterations, to a pathway-centric view, with the description of core pathways enriched in mutations [30]. The pathway-centric view fits with a current consideration of complex diseases as pathway diseases more than gene diseases [31]. This shift in the analysis provides more biologically consistent results and can be extended to related problems, such as disease classification [32], assessment of progression [33] or evaluation of chemotherapy resistance [34] in cancers.

Unfortunately, the majority of human genes are not assigned to well-characterized pathways [35]. This limitation can be overcome by analyzing molecular interactions between proteins. Indeed, public databases, such as the BioGRID database [36], store a lot of interaction data, even for proteins that are poorly described at the molecular and biochemical levels. These interactions can not only complement pathway-based approaches, but also provide information on other biological processes and regulations in which proteins are involved. In the context of high-throughput studies of complex diseases, networks can provide valuable indications. For instance, subnetworks important for breast cancer metastasis can be identified by mapping changes in gene expression onto a protein-interaction network. These subnetworks are used to provide metastasis markers, with the advantage that subnetwork markers are potentially more robust than single gene signatures [37]. In the same way, global pathway consistencies and activities distinguish between different breast cancer subtypes such as estrogen-receptor positive/negative status [38].

Variation in coexpression between proteins and their interaction partners has also been assessed to predict the outcome of disease. In breast cancers, expression of the DNA-damage repair protein BRCA1 is strongly correlated with the expression of its interaction partners in tumors from patients with a good outcome, whereas it is uncorrelated with their expression in tumors from patients with a poor prognosis [39]. The value of molecular network integration is not restricted to microarray analyses. For example, integration of microRNA profiling and proteomic analyses has been used to reveal three subnetworks involved in different aspects of osteoarthritis, a multifactorial disease characterized by destruction of the articular cartilage [40]. Finally, with regard to genotyping studies, in which thousands of variations appear for each particular individual, networks offer a way of interpreting the significance of these variations at the molecular level. For example, the connectivity provided by a molecular network can shortcut the huge combinatorial space of possible gene-gene epistasis, a problem currently addressed by expensive computational approaches [14].

Integrating clinical and genomic information into networks

The high-throughput studies of disease discussed above mainly emerge from a culture of molecular biology and are still rather disconnected from the medical field. It is clear that to gain insights into complex diseases, new approaches will have to go beyond simple phenotypic descriptions and use more precise clinical information. We would like to argue here that networks can play an instrumental role in the integration of medical information required for the translation of high-throughput genomics into a greater understanding of disease and, ultimately, into personalized medicine.

Molecular networks have been used to link disease genotypes. An initial set of published studies has pioneered the inclusion of disease descriptions with high-throughput genomic data. For example, Butte and Kohane [41] applied text-mining strategies to organize microarray experiments into similar disease classes, according to the Unified Medical Language System Metathesaurus terms (UMLS; a compendium of ontologies) associated with their experimental annotations. Box 1 lists the main standards for disease description and databases of disease phenotypic information. Specific associations between individual genes and diseases, principally extracted from OMIM [2], have been exploited to study relationships between phenotype and underlying molecular mechanism. Using this approach, Van Driel *et al.* [42] showed that disease-related proteins are correlated with various attributes, including their organization in protein interactions. They established phenotypic and disease similarities between protein pairs by comparing their corresponding Medical Subject Heading (MeSH) biomedical terms extracted from the OMIM descriptions of

Box 1. Sources of standard disease phenotype terminology

International standards for describing disease phenotypes

The World Health Organization's **International Classification of Diseases (ICD)** is a widely used standard terminology for classification of diseases and health disorders [46]. The current version is available in more than 30 languages, covers more than 14,000 medical terms and includes adaptations focused on specific health areas such as oncology, mental disorder or primary care.

The **Unified Medical Language System Metathesaurus** (**UMLS**) is also a well-known source of ontology standards, integrating more than 2 million medical terms, and 12 million relationships between them [43]. UMLS-associated projects include the **Medical Subject Headings** (**MeSH**) thesaurus, a controlled vocabulary used for cataloging biomedical and health-related documents that provides one of the most popular searching facilities as the MeSH terms are used to label Medline abstracts. It also contains the **Logical Observation Identifiers Names and Codes** (**LOINC**) [47], a catalogue of universal identifiers designed for the electronic exchange of laboratory and clinical test results [48].

Another source of standard terminology is the **Systematized Nomenclature of Medicine-Clinical Terms** (**SNOMED-CT**) [49], supported by the International Health Terminology Standards Development Organization [50]. This computer-readable collection of medical terms covers diverse clinical areas such as diseases, medical procedures and drugs. SNOMED-CT currently contains more than 310,000 concepts with unique meanings and formal logic-based definitions organized into hierarchies. SNOMED-CT has already been extended to Spanish, and translations to other languages such as Danish, French and Swedish are currently taking place, addressing one of the pressing needs in the multilingual environment of medical records.

Complementary disease-related ontologies are the **Human Phenotype Ontology** (**HPO**) [51], with more than 8,000 terms representing individual phenotypic abnormalities [52] and the **Disease Ontology** (**DOID**) [53], which is part of the Open Biological Ontologies Foundry (OBO) [54].

Information on disease phenotypes related to particular genes and proteins

The **Online Mendelian Inheritance in Man (OMIM) database** stores information such as gene descriptions, inheritance patterns, localization maps and polymorphisms for more than 12,500 gene loci and phenotypic descriptions [55].

SwissProt, the key source of information about protein function, even though not specifically dedicated to disease-related annotations, also includes information linking proteins and associated mutations with pathologies. It provides a very useful link between MeSH disease terminology and specific proteins [56].

Disease description standardization is also fundamental for the exchange of electronic medical records and for their interoperability. Major efforts such as Health Level Seven (HL7) [57] and Digital Imaging and Communication in Medicine (DICOM) [58] protocols provide standards for sharing and retrieving electronic health information and medical images. A more detailed description of standards for electronic medical charts is provided in specialized reviews [59].

the corresponding genes. Lage and collaborators [12] predicted 113 new disease-candidate genes by comparing their protein-interaction neighborhood with the associated phenotypes. In this case, the phenotypes were defined by identifying UMLS terms [43] in the OMIM descriptions. Each disease was then described as a vector of medical terms that can be directly compared. These are perhaps the best current examples of how protein-interaction network data can be used to interpret phenotypic proximities between diseases. However, only basic descriptions of the diseases are used, far from the complete - and individual - information contained in medical records.

For a greater insight into complex diseases, it will be necessary to access detailed information such as symptoms, diagnosis, treatment and disease progression. The main source of detailed information are patients' medical records, authored by physicians. Medical records store private

Box 2. Biobanks

The efficient mining of large collections of clinical and epidemiological data requires the availability of electronic and standardized records coupled to organized collections of samples in biological banks (biobanks). The concept of a biobank covers efforts with different goals and organization, from efforts to obtain samples from the general population, to collections dedicated to specific diseases, in particular cancer types. Biobanks also vary greatly in the type of sample-associated information they contain. In some cases this comprises very detailed clinical and epidemiological records, and in others only basic descriptions of population characteristics. At a very general level, three main types of biobanks can be distinguished [60].

Population biobanks gather germline DNA from healthy donors representing a particular regional population. Their major goal is to obtain biomarkers of susceptibility and population characteristics.

Disease-oriented biobanks focus on the identification of disease biomarkers for patient selection. They store collections of pathological and healthy samples commonly associated with clinical data or trials. Well-known examples are tumor biobanks.

Epidemiology-oriented biobanks focus on exposure biomarkers. Samples are recruited from healthy exposed individuals or from case-control studies.

Current efforts in biobank development include the European Biobanking and Biomolecular Resources Infrastructure (BBMRI), which intends to coordinate biobanks from 19 European countries, including the organization of compatible infrastructures and annotations [61]. The European Life-sciences Infrastructure for Biological Information project (ELIXIR [61]), another project of BBMRI, represents an effort to link biological and biomedical databases and computational resources [61]. In the same way, the NCI Biomedical Informatics Grid project (CaBIG) supports the integration of medical oncology and cancer research genome projects [62]. The Public Population Project in Genomics consortium (P3G) gathers together more than 20 international institutions to promote effective collaborations between biobanks involved in population studies [63].

Examples of specific biobank developments are the Estonian Gene Bank Project [64], the private initiative of deCODE project in Iceland [65], the Spanish National Tumor Bank network [66], the DNA scanning project in children, in the Children's Hospital of Philadelphia (CHOP) [67], the Personalized Medicine Research Project DNA Biobank [68] in the United States, and the BioBank Japan Project [69].

patient data as well as clinical information on their illnesses. Unfortunately, the mining of electronic medical records is exposed to well-known legal difficulties such as intellectual property and patient confidentiality. Furthermore, the lack of standardization between hospitals and institutions and the recruitment of poorly annotated samples make gathering clinical data a major burden, as demonstrated in large-scale projects such as the Cancer Genome Atlas [44].

The availability of biological samples, combined with adequate clinical and epidemiological information, is of paramount importance in correlating disease phenotypes with their molecular underpinnings. This is where 'biobanks' come in (Box 2). The information recorded in biobank entries is more accessible to research projects as, in general they contain less direct personal information. Brief standardized health summaries describing the minimal, but relevant, clinical information, without damaging confidentiality and intellectual property rights, can provide an intermediate solution between the extremes of complete medical records and minimal pathological information associated with biological samples [45]. In this evolving situation, biobanks will facilitate the integration of high-throughput genomic information with disease descriptions using information standards and medical ontologies.

In conclusion, the effective translation of high-throughput genomic data on complex diseases into molecular mechanisms and potential therapies requires taking precise medical information into account. Molecular networks are currently being used to interpret high-throughput data generated in functional genomics or genotyping studies, but they can also be used as an instrument to interpret phenotypic data in molecular terms. Molecular networks are flexible enough to integrate high-throughput genomic information with phenotypic descriptions of complex diseases (Figure 1). To achieve this goal, however, the networks will have to be reliable, complete, and combine the various types of molecular interactions present in living cells. Furthermore, to fully understand disease mechanisms, molecular

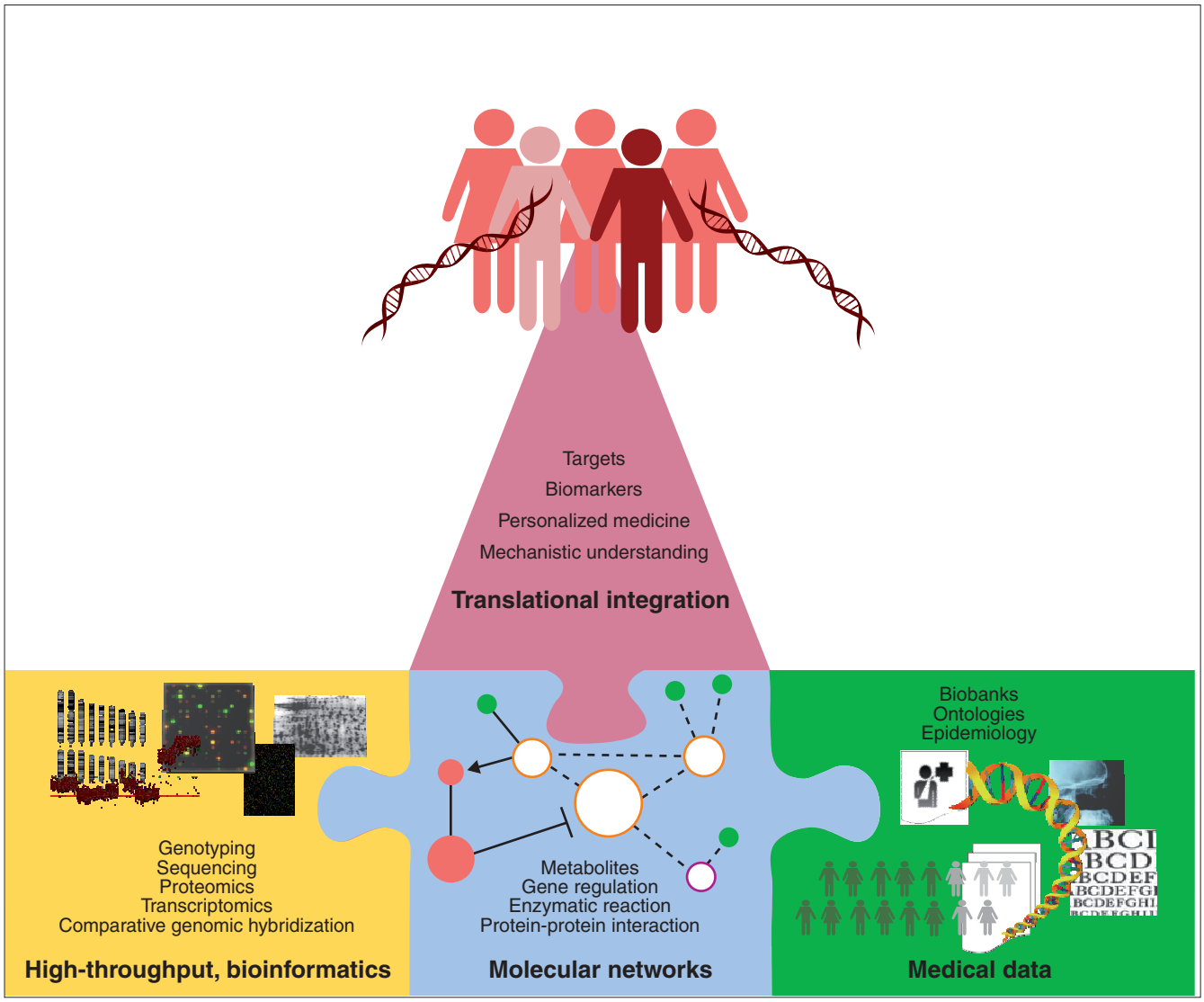

Figure 1
Bioinformatics high-throughput experiments and medical resources can be integrated through molecular networks.

networks will have to switch from their current static description of interactions to dynamic information, describing the evolution of the network in time and space. Such integrated networks will be particularly relevant for complex diseases, where targeted therapy against single proteins is not sufficient, and critical therapeutic decisions can be better taken in the knowledge of integrated molecular profiles. Finally, molecular networks could ultimately take into account the mutations and polymorphisms specific to individual cases. In this vision, the future integration of information using molecular networks as frameworks is the basis for the development of personalized medicine.

Acknowledgements

We thank Gert-jan van Ommen, Manuel Morente, Trey Ideker, Gary Bader and Søren Brunak for valuable comments and suggestions. This work is supported by ISCIII grant COMBIOMED (RD07/0067/0014), MICINN grant BIO2007-66855, Spanish Ministry of Education and Science, EU grants LSHGCT-2003-503265 (BioSapiens), LSHG-CT-2004-503567 (ENFIN) and Eurocancercoms EU Seventh Framework Programme. AB is supported by the "Juan de la Cierva" fellowship; GG-L is partially supported by the Spanish National Institute for Bioinformatics (INB), a platform of Genoma España.

References

1. Chng WJ: **Limits to the Human Cancer Genome Project?** *Science* 2007, **315:**762; author reply 764-765.
2. **OMIM** [http://www.ncbi.nlm.nih.gov/omim]
3. Danchin A: **The Delphic boat or what the genomic texts tell us.** *Bioinformatics* 1998, **14:**383.
4. Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, Droege A, Lindenberg KS, Knoblich M, Haenig C, Herbst M, Suopanki J, Scherzinger E, Abraham C, Bauer B, Hasenbank R, Fritzsche A, Ludewig AH, Büssow K, Coleman SH, Gutekunst C, Landwehrmeyer BG, Lehrach H, Wanker EE: **A protein interaction**

network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol Cell* 2004, **15:**853-865.
5. Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual J, Fisk CJ, Li N, Smolyar A, Hill DE, Barabási A, Vidal M, Zoghbi HY: **A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration.** *Cell* 2006, **125:**801-814.
6. de Chassey B, Navratil V, Tafforeau L, Hiet MS, Aublin-Gex A, Agaugué S, Meiffren G, Pradezynski F, Faria BF, Chantier T, Le Breton M, Pellet J, Davoust N, Mangeot PE, Chaboud A, Penin F, Jacob Y, Vidalain PO, Vidal M, André P, Rabourdin-Combe C, Lotteau V: **Hepatitis C virus infection protein network.** *Mol Syst Biol* 2008, **4:**230.
7. Zimmermann GR, Lehár J, Keith CT: **Multi-target therapeutics: when the whole is greater than the sum of the parts.** *Drug Discov Today* 2007, **12:**34-42.
8. Fraser HB, Plotkin JB: **Using protein complexes to predict phenotypic effects of gene mutation.** *Genome Biol* 2007, **8:**R252.
9. Goh K, Cusick ME, Valle D, Childs B, Vidal M, Barabási A: **The human disease network.** *Proc Natl Acad Sci USA* 2007, **104:**8685-8690.
10. Pujana MA, Han JJ, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B, Assmann V, Elshamy WM, Rual J, Levine D, Rozek LS, Gelman RS, Gunsalus KC, Greenberg RA, Sobhian B, Bertin N, Venkatesan K, Ayivi-Guedehoussou N, Solé X, Hernández P, Lázaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, *et al.*: **Network modeling links breast cancer susceptibility and centrosome dysfunction.** *Nat Genet* 2007, **39:**1338-1349.
11. Oti M, Snel B, Huynen MA, Brunner HG: **Predicting disease genes using protein-protein interactions.** *J Med Genet* 2006, **43:**691-698.
12. Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tümer Z, Pociot F, Tommerup N, Moreau Y, Brunak S: **A human phenome-interactome network of protein complexes implicated in genetic disorders.** *Nat Biotechnol* 2007, **25:**309-316.
13. Wu X, Jiang R, Zhang MQ, Li S: **Network-based global inference of human disease genes.** *Mol Syst Biol* 2008, **4:**189.
14. Pattin KA, Moore JH: **Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases.** *Hum Genet* 2008, **124:**19-29.
15. Buchanan AV, Weiss KM, Fullerton SM: **Dissecting complex disease: the quest for the Philosopher's Stone?** *Int J Epidemiol* 2006, **35:**562-571.
16. **The Cancer Genome Atlas** [http://cancergenome.nih.gov]
17. **Cancer Genome Anatomy Project** [http://www.ncbi.nlm.nih.gov/ncicgap]
18. **Diabetes Genome Anatomy Project** [http://www.diabetesgenome.org]
19. Hu-Lince D, Craig DW, Huentelman MJ, Stephan DA: **The Autism Genome Project: goals and strategies.** *Am J Pharmacogenomics* 2005, **5:**233-246.
20. Autism Genome Project Consortium: Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu X, Vincent JB, Skaug JL, Thompson AP, Senman L, Feuk L, Qian C, Bryson SE, Jones MB, Marshall CR, Scherer SW, Vieland VJ, Bartlett C, Mangin LV, Goedken R, Segre A, Pericak-Vance MA, Cuccaro ML, Gilbert JR, Wright HH, Abramson RK, Betancur C, Bourgeron T, Gillberg C, *et al.*: **Mapping autism risk loci using genetic linkage and chromosomal rearrangements.** *Nat Genet* 2007, **39:**319-328.
21. Stefansson H, Rujescu D, Cichon S, Pietiläinen OPH, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller H, Hartmann A, *et al.*: **Large recurrent microdeletions associated with schizophrenia.** *Nature* 2008, **455:**232-236.
22. The International HapMap Consortium: **The International HapMap Project.** *Nature* 2003, **426:**789-796.
23. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Boström K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Råstam L, Speliotes EK, *et al.*: **Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.** *Science* 2007, **316:**1331-1336.
24. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JKV, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PVK, *et al.*: **The genomic landscapes of human breast and colorectal cancers.** *Science* 2007, **318:**1108-1113.
25. Landemaine T, Jackson A, Bellahcène A, Rucci N, Sin S, Abad BM, Sierra A, Boudinet A, Guinebretière J, Ricevuto E, Noguès C, Briffod M, Bièche I, Cherel P, Garcia T, Castronovo V, Teti A, Lidereau R, Driouch K: **A six-gene signature predicting breast cancer lung metastasis.** *Cancer Res* 2008, **68:**6092-6099.
26. Corbett BA, Kantor AB, Schulman H, Walker WL, Lit L, Ashwood P, Rocke DM, Sharp FR: **A proteomic study of serum from children with autism showing differential expression of apolipoproteins and complement proteins.** *Mol Psychiatry* 2007, **12:**292-306.
27. Al-Shahrour F, Minguez P, Tárraga J, Medina I, Alloza E, Montaner D, Dopazo J: **FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments.** *Nucleic Acids Res* 2007, **35:**W91-W96.
28. Bandapalli OR, Geheeb M, Kobelt D, Kuehnle K, Elezkurtaj S, Herrmann J, Gressner AM, Weiskirchen R, Beule D, Blüthgen N, Herzel H, Franke C, Brand K: **Global analysis of host tissue gene expression in the invasive front of colorectal liver metastases.** *Int J Cancer* 2006, **118:**74-89.
29. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102:**15545-15550.
30. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong S, Fu B, Lin M, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, *et al.*: **Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.** *Science* 2008, **321:**1801-1806.
31. Jones D: **Pathways to cancer therapy.** *Nat Rev Drug Discov* 2008, **7:**875-876.
32. Lee E, Chuang H, Kim J, Ideker T, Lee D: **Inferring pathway activity toward precise disease classification.** *PLoS Comput Biol* 2008, **4:**e1000217.
33. Ruminy P, Jardin F, Picquenot J, Parmentier F, Contentin N, Buchonnet G, Tison S, Rainville V, Tilly H, Bastard C: **S(mu) mutation patterns suggest different progression pathways in follicular lymphoma: early direct or late from FL progenitor cells.** *Blood* 2008, **112:**1951-1959.
34. Riedel RF, Porrello A, Pontzer E, Chenette EJ, Hsu DS, Balakumaran B, Potti A, Nevins J, Febbo PG: **A genomic approach to identify molecular pathways associated with chemotherapy resistance.** *Mol Cancer Ther* 2008, **7:**3141-3149.
35. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P: **Reactome knowledgebase of human biological pathways and processes.** *Nucleic Acids Res* 2009, **37:**D619-D622.
36. Stark C, Breitkreutz B, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Res* 2006, **34:**D535-D539.
37. Chuang H, Lee E, Liu Y, Lee D, Ideker T: **Network-based classification of breast cancer metastasis.** *Mol Syst Biol* 2007, **3:**140.
38. Efroni S, Schaefer CF, Buetow KH: **Identification of key processes underlying cancer phenotypes using biologic pathway analysis.** *PLoS ONE* 2007, **2:**e425.
39. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL: **Dynamic modularity in protein interaction networks predicts breast cancer outcome.** *Nat Biotechnol* 2009, **27:**199-204.
40. Iliopoulos D, Malizos KN, Oikonomou P, Tsezou A: **Integrative microRNA and proteomic approaches identify novel osteoarthritis genes and their collaborative metabolic and inflammatory networks.** *PLoS ONE* 2008, **3:**e3740.
41. Butte AJ, Kohane IS: **Creation and implications of a phenome-genome network.** *Nat Biotechnol* 2006, **24:**55-62.
42. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM: **A text-mining analysis of the human phenome.** *Eur J Hum Genet* 2006, **14:**535-542.

43. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology.** *Nucleic Acids Res* 2004, **32:**D267-D270.
44. Compton C: In *Scientific Workshop Report 2008* [http://www.icgc.org/documents]
45. Tierney WM, Beck EJ, Gardner RM, Musick B, Shields M, Shiyonga NM, Spohr MH: **Viewpoint: a pragmatic approach to constructing a minimum data set for care of patients with HIV in developing countries.** *J Am Med Inform Assoc* 2006, **13:**253-260.
46. **ICD** [http://www.who.int/classifications/icd]
47. **LOINC** [http://loinc.org]
48. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, Forrey A, Mercer K, DeMoor G, Hook J, Williams W, Case J, Maloney P: **LOINC, a universal standard for identifying laboratory observations: a 5-year update.** *Clin Chem* 2003, **49:**624-633.
49. **SNOMED** [http://www.ihtsdo.org/snomed-ct]
50. **IHTSDO** [http://www.ihtsdo.org]
51. **HPO** [http://www.human-phenotype-ontology.org]
52. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S: **The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease.** *Am J Hum Genet* 2008, **83:**610-615.
53. **DOID** [http://diseaseontology.sourceforge.net]
54. **OBO** [http://obofoundry.org]
55. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2009, **37:**D5-D15.
56. Mottaz A, Yip YL, Ruch P, Veuthey A: **Mapping proteins to disease terminologies: from UniProt to MeSH.** *BMC Bioinformatics* 2008, **9(Suppl 5):**S3.
57. **Health Level 7** [http://www.hl7.org]
58. **DICOM** [http://medical.nema.org]
59. Kalra D: **Electronic health record standards.** *Yearb Med Inform* 2006:136-144.
60. Riegman PHJ, Morente MM, Betsou F, de Blasio P, Geary P: **Biobanking for better healthcare.** *Mol Oncol* 2008, **2:**213-222.
61. Yuille M, van Ommen G, Bréchot C, Cambon-Thomsen A, Dagher G, Landegren U, Litton J, Pasterk M, Peltonen L, Taussig M, Wichmann H, Zatloukal K: **Biobanking for Europe.** *Brief Bioinformatics* 2008, **9:**14-24.
62. **caBIG** [https://cabig.nci.nih.gov]
63. **P3G** [http://www.p3g.org]
64. Metspalu A: **Estonian Genome Project - before the take-off and take-off.** *Bioinformatics* 2002, **189 Suppl 2:**S152.
65. **deCODE** [http://www.decode.com]
66. Morente MM, de Alava E, Fernandez PL: **Tumour banking: the Spanish design.** *Pathobiology* 2007, **74:**245-250.
67. Kaiser J: **Genetics. U.S. hospital launches large biobank of children's DNA.** *Science* 2006, **312:**1584-1585.
68. McCarty CA, Mukesh BN, Kitchner TE, Hubbard WC, Wilke RA, Burmester JK, Patchett RB: **Intraocular pressure response to medication in a clinical setting: the Marshfield Clinic Personalized Medicine Research Project.** *J Glaucoma* 2008, **17:**372-377.
69. Nakamura Y: **The BioBank Japan Project.** *Clin Adv Hematol Oncol* 2007, **5:**696-697.
70. **Cancer Genome Project** [http://www.sanger.ac.uk/genetics/CGP]
71. **The International Cancer Genome Consortium** [http://www.icgc.org]
72. **Cancer Genetic Markers Susceptibility** [http://cgems.cancer.gov]
73. **Alzheimer's Genome Project** [http://www.curealzfund.org/content/view/105/79]
74. **The Schizophrenia Genome Project** [http://schizophrenia.ncgr.org/index.jsp]
75. **HapMap** [http://www.hapmap.org]
76. **HGVMap** [http://www.hgvbaseg2p.org]
77. **Cosmic** [http://www.sanger.ac.uk/genetics/CGP/cosmic]
78. **Genome-wide association studies catalog** [http://www.genome.gov/page.cfm?pageID=26525384#searchForm]