Spring 2017 – Epigenetics and Systems Biology
Lecture Outline (Systems Biology)
Michael K. Skinner – Biol 476/576
CUE 418, 10:35-11:50 am, Tuesdays & Thursdays
January 24 & 31, 2017
Weeks 3 and 4

Systems Biology (Components & Technology)

Components (DNA, Expression, Cellular, Organ, Physiology, Organism, Differentiation, Development, Phenotype, Evolution)

Technology (Genomics, Transcriptomes, Proteomics)
(Interaction, Signaling, Metabolism)

Omics (Data Processing and Resources)

Required Reading

ENCODE (2012) ENCODE Explained. Nature 489:52-55.

Street ME, et al. (2013) Artificial Neural Networks, and Evolutionary Algorithms as a systems biology approach to a data-base on fetal growth restriction. Prog Biophys Mol Biol. 113(3):433-8.

Literature

Sun YV, Hu YJ. (2016) Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. Adv Genet. 2016;93:147-90.

Horgusluoglu E, Nudelman K, Nho K, Saykin AJ. (2016) Adult neurogenesis and neurodegenerative diseases: A systems biology perspective. Am J Med Genet B Neuropsychiatr Genet. 2016 Feb 16. doi: 10.1002/ajmg.b.32429. [Epub ahead of print]

Zenil H, Kiani NA, Tegnér J. (2016) Methods of information theory and algorithmic complexity for network biology. Semin Cell Dev Biol. 51:32-43.

Ostaszewski M, Skupin A, Balling R. (2016) Neurological Diseases from a Systems Medicine Point of View. Methods Mol Biol. 2016;1386:221-50.

Nishi A, Milner DA Jr, Giovannucci EL, et al. (2016) Integration of molecular pathology, epidemiology and social science for global precision medicine. Expert Rev Mol Diagn. 2016;16(1):11-23.

Davidsen PK, Turan N, Egginton S, Falciani F. (1985) Multilevel functional genomics data integration as a tool for understanding physiology: a network biology perspective. J Appl Physiol (1985). 2016 Feb 1;120(3):297-309.

Prathipati P, Mizuguchi K. (2016) Systems Biology Approaches to a Rational Drug Discovery Paradigm. Curr Top Med Chem. 2016;16(9):1009-25.

Qin Y, Jiao X1, Simpson JL, Chen ZJ. (2015) Genetics of primary ovarian insufficiency: new developments and opportunities. Hum Reprod Update. 2015 Nov-Dec;21(6):787-808.

Parikshak NN, Gandal MJ, Geschwind DH. (2015) Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. Nat Rev Genet. 16(8):441-58.
Xie L, Draizen EJ, Bourne PE. (2016) Harnessing Big Data for Systems Pharmacology. Annu Rev Pharmacol Toxicol. 2016 Oct 13. [Epub ahead of print]
Macovei A, Pagano A, Leonetti P, Carbonera D, Balestrazzi A, Araújo SS. (2016) Systems biology and genome-wide approaches to unveil the molecular players involved in the pre-germinative metabolism: implications on seed technology traits. Plant Cell Rep. 2016 Oct 11. [Epub ahead of print]
Tosto G, Reitz C. (2016) Use of "omics" technologies to dissect neurologic disease. Handb Clin Neurol. 2016;138:91-106.
Altaf-Ul-Amin M, Afendi FM, Kiboi SK, Kanaya S. (2014) Systems biology in the context of big data and networks. Biomed Res Int. 2014;2014:428570.
Putri SP, Nakayama Y, Matsuda F, et al. (2013) Current metabolomics: practical applications. J Biosci Bioeng. 115(6):579-89.
Manning T, Sleator RD, Walsh P. (2013) Naturally selecting solutions: the use of genetic algorithms in bioinformatics. Bioengineered. 4(5):266-78.
Shekari F, Baharvand H, Salekdeh GH. (2014) Organellar proteomics of embryonic stem cells. Adv Protein Chem Struct Biol. 95:215-30.
Wu X, Hasan MA, Chen JY. (2014) Pathway and network analysis in proteomics. J Theor Biol. 2014 Jun 6. pii: S0022-5193(14)00304-X. [Epub ahead of print]
Liu Z, Wang Y, Xue Y. (2013) Phosphoproteomics-based network medicine. FEBS J. 280(22):5696-704.
Dharuri H, Demirkan A, van Klinken JB, et al. (2014) Genetics of the human metabolome, what is next? Biochim Biophys Acta.1842(10):1923-1931.
Stumpf MP. (2014) Approximate Bayesian inference for complex ecosystems. F1000Prime Rep. 17;6:60.
Purcell O, Lu TK. (2014) Synthetic analog and digital circuits for cellular computation and memory. Curr Opin Biotechnol. 29:146-55.
Mason CE, Porter SG, Smith TM. (2014) Characterizing multi-omic data in systems biology. Adv Exp Med Biol. 799:15-38.
Sarpeshkar R. (2014) Analog synthetic biology. Philos Trans A Math Phys Eng Sci. 24;372(2012):20130110.
Rekhi R, Qutub AA. (2013) Systems approaches for synthetic biology: a pathway toward mammalian design. Front Physiol. 9;4:285.
Renda BA, Hammerling MJ, Barrick JE. (2014) Engineering reduced evolutionary potential for synthetic biology. Mol Biosyst. 10(7):1668-78.
Cronin RM, Field JR, Bradford Y, et al. (2014) Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. Front Genet. 5;5:250.
Svahn AJ, Becker TS, Graeber MB. (2014) Emergent properties of microglia. Brain Pathol. 24(6):665-70.
Caterino M, Aspesi A, Pavesi E, et al. (2014) Analysis of the interactome of ribosomal protein S19 mutants. Proteomics. 14(20):2286-96.
Singh R, Dangol S, Jwa NS. (2014) Yeast two-hybrid system for dissecting the rice MAPK interactome. Methods Mol Biol. 1171:195-216.
Petrey D, Honig B. (2014) Structural bioinformatics of the interactome. Annu Rev Biophys. 43:193-210.

Garcia B, Datta G, Cosgrove GP, Strong M. (2014) Network and matrix analysis of the respiratory disease interactome. BMC Syst Biol. 22;8:34.
Blomme J, Inzé D, Gonzalez N. (2014) The cell-cycle interactome: a source of growth regulators? J Exp Bot. 65(10):2715-30.
Stevens A, De Leonibus C, Hanson D, et al. (2014) Network analysis: a new approach to study endocrine disorders. J Mol Endocrinol. 19;52(1):R79-93.
Salvo SA, Hirsch CN, Buell CR, Kaeppler S, Kaeppler HF. (2014) Whole Transcriptome Profiling of Maize during Early Somatic Embryogenesis Reveals Altered Expression of Stress Factors and Embryogenesis-Related Genes. PLoS One. 30;9(10):e111407.
Xuan J, Yu Y, Qing T, Guo L, Shi L. (2013) Next-generation sequencing in the clinic: promises and challenges. Cancer Lett. 1;340(2):284-95.
Robinson SW, Fernandes M, Husi H. (2014) Current advances in systems and integrative biology. Comput Struct Biotechnol J. 11(18):35-46.
Sharma A, Rai A, Lal S. (2013) Workflow management systems for gene sequence analysis and evolutionary studies - A Review. Bioinformation. 17;9(13):663-72.
Street ME, Buscema M, Smerieri A, Montanini L, Grossi E. (2013) Artificial Neural Networks, and Evolutionary Algorithms as a systems biology approach to a data-base on fetal growth restriction. Prog Biophys Mol Biol. 113(3):433-8.
Caccia D, Dugo M, Callari M, Bongarzone I. (2013) Bioinformatics tools for secretome analysis. Biochim Biophys Acta. 1834(11):2442-53.
Ecker JR, et al. (2012) Genomics: ENCODE explained. Nature. 6;489(7414):52-5.
Gerstein MB, et al. (2012) Architecture of the human regulatory network derived from ENCODE data. Nature. 6;489(7414):91-100.
Thurman RE, et al. (2012) The accessible chromatin landscape of the human genome. Nature. 6;489(7414):75-82.
Sanyal A, Lajoie BR, Jain G, Dekker J. (2012) The long-range interaction landscape of gene promoters. Nature. 6;489(7414):109-13.
Afacan NJ, Fjell CD, Hancock RE. (2012) A systems biology approach to nutritional immunology - focus on innate immunity. Mol Aspects Med. 33(1):14-25.
Wilson RA. (2012) The cell biology of schistosomes: a window on the evolution of the early metazoa. Protoplasma. 249(3):503-18.
Murphy BF, Thompson MB. (2012) A review of the evolution of viviparity in squamate reptiles: the past, present and future role of molecular biology and genomics. J Comp Physiol B. 181(5):575-94.
Fritzsch FS, Dusny C, Frick O, Schmid A. (2012) Single-cell analysis in biotechnology, systems biology, and biocatalysis. Annu Rev Chem Biomol Eng. 3:129-55.
Tian Q, Price ND, Hood L. (2012) Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. J Intern Med. 271(2):111-21.
Weckwerth W. (2011) Green systems biology - From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. J Proteomics. 10;75(1):284-305.
St-Denis N, Gingras AC. (2012) Mass spectrometric tools for systematic analysis of protein phosphorylation. Prog Mol Biol Transl Sci. 106:3-32.
Amit I, Regev A, Hacohen N. (2011) Strategies to discover regulatory circuits of the mammalian immune system. Nat Rev Immunol. 18;11(12):873-80.
Zhao S, Iyengar R. (2012) Systems pharmacology: network analysis to identify multiscale mechanisms of drug action. Annu Rev Pharmacol Toxicol. 10;52:505-21.

Habibi E, Masoudi-Nejad A, Abdolmaleky HM, Haggarty SJ. (2011) Emerging roles of epigenetic mechanisms in Parkinson's disease. Funct Integr Genomics. 11(4):523-37.
Liu ZP, Chen L. (2012) Proteome-wide prediction of protein-protein interactions from high-throughput data. Protein Cell. 3(7):508-20.
Markowitz VM, et al. (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 40(Database issue):D115-22.
Diercks A, Aderem A. (2012) Systems Approaches to Dissecting Immunity. Curr Top Microbiol Immunol. 2012 Aug 11. [Epub ahead of print]
Scholz B, Marschalek R. (2012) Epigenetics and blood disorders. Br J Haematol. 158(3):307-22.
Borenstein E. (2012) Computational systems biology and in silico modeling of the human microbiome. Brief Bioinform. 13(6):769-80.
Gupta RK, Rosen ED, Spiegelman BM. (2011) Identifying novel transcriptional components controlling energy metabolism. Cell Metab. 7;14(6):739-45.
Galliot B, Quiquand M. (2011) A two-step process in the emergence of neurogenesis. Eur J Neurosci. 34(6):847-62.
Rodin AS, Gogoshin G, Boerwinkle E. (2011) Systems biology data analysis methodology in pharmacogenomics. Pharmacogenomics. 12(9):1349-60.
Sobie EA, Lee YS, Jenkins SL, Iyengar R. (2011) Systems biology--biomedical modeling. Sci Signal. 6;4(190):tr2
Habibi E, Masoudi-Nejad A, Abdolmaleky HM, Haggarty SJ. (2011) Emerging roles of epigenetic mechanisms in Parkinson's disease. Funct Integr Genomics. 11(4):523-37.
Day JJ, Sweatt JD. (2012) Epigenetic treatments for cognitive impairments. Neuropsychopharmacology. 37(1):247-60.
Prezioso C, Orlando V. (2011) Polycomb proteins in mammalian cell differentiation and plasticity. FEBS Lett. 7;585(13):2067-77.
Zhang X, Yap Y, Wei D, Chen G, Chen F. Novel omics technologies in nutrition research. Biotechnol Adv. 2008 Mar-Apr;26(2):169-76.

FORUM: Genomics

ENCODE explained

The Encyclopedia of DNA Elements (ENCODE) project dishes up a hearty banquet of data that illuminate the roles of the functional elements of the human genome. Here, five scientists describe the project and discuss how the data are influencing research directions across many fields. SEE ARTICLES P.57, P.75, P.83, P.91, P.101 & LETTER P.109

Serving up a genome feast

JOSEPH R. ECKER

Starting with a list of simple ingredients and blending them in the precise amounts needed to prepare a gourmet meal is a challenging task. In many respects, this task is analogous to the goal of the ENCODE project[1], the recent progress of which is described in this issue[2–7]. The project aims to fully describe the list of common ingredients (functional elements) that make up the human genome (Fig. 1). When mixed in the right proportions, these ingredients constitute the information needed to build all the types of cells, body organs and, ultimately, an entire person from a single genome.

The ENCODE pilot project[8] focused on just 1% of the genome — a mere appetizer — and its results hinted that the list of human genes was incomplete. Although there was scepticism about the feasibility of scaling up the project to the entire genome and to many hundreds of cell types, recent advances in low-cost, rapid DNA-sequencing technology radically changed that view[9]. Now the ENCODE consortium presents a menu of 1,640 genome-wide data sets prepared from 147 cell types, providing a six-course serving of papers in *Nature*, along with many companion publications in other journals.

One of the more remarkable findings described in the consortium's 'entrée' paper (page 57)[2] is that 80% of the genome contains elements linked to biochemical functions, dispatching the widely held view that the human genome is mostly 'junk DNA'. The authors report that the space between genes is filled with enhancers (regulatory DNA elements), promoters (the sites at which DNA's transcription into RNA is initiated) and numerous previously overlooked regions that encode RNA transcripts that are not translated into proteins but might have regulatory roles. Of note, these results show that many DNA variants previously correlated with certain diseases lie within or very near non-coding functional DNA elements, providing new leads for linking genetic variation and disease.

The five companion articles[3–7] dish up diverse sets of genome-wide data regarding the mapping of transcribed regions, DNA binding of regulatory proteins (transcription factors) and the structure and modifications of chromatin (the association of DNA and proteins that makes up chromosomes), among other delicacies.

Djebali and colleagues[3] (page 101) describe ultra-deep sequencing of RNAs prepared from many different cell lines and from specific compartments within the cells. They conclude that about 75% of the genome is transcribed at some point in some cells, and that genes are highly interlaced with overlapping transcripts that are synthesized from both DNA strands. These findings force a rethink of the definition of a gene and of the minimum unit of heredity.

Moving on to the second and third courses, Thurman *et al.*[4] and Neph *et al.*[5] (pages 75 and 83) have prepared two tasty chromatin-related treats. Both studies are based on the DNase I hypersensitivity assay, which detects genomic regions at which enzyme access to, and subsequent cleavage of, DNA is unobstructed by chromatin proteins. The authors identified cell-specific patterns of DNase I hypersensitive sites that show remarkable concordance with experimentally determined and computationally predicted binding sites of transcription factors. Moreover, they have doubled the number of known recognition sequences for DNA-binding proteins in the human genome, and have revealed a 50-base-pair 'footprint' that is present in thousands of promoters[5].

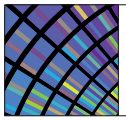
The next course, provided by Gerstein and colleagues[6] (page 91) examines the principles behind the wiring of transcription-factor networks. In addition to assigning relatively simple functions to genome elements (such as 'protein X binds to DNA element Y'), this study attempts to clarify the hierarchies of transcription factors and how the intertwined networks arise.

Beyond the linear organization of genes and transcripts on chromosomes lies a more complex (and still poorly understood) network of chromosome loops and twists through which promoters and more distal elements, such as enhancers, can communicate their regulatory information to each other. In the final course of the ENCODE genome feast, Sanyal and colleagues[7] (page 109) map more than 1,000 of these long-range signals in each cell type. Their findings begin to overturn the long-held (and probably oversimplified) prediction that the regulation of a gene is dominated by its proximity to the closest regulatory elements.

> ***"These findings force a rethink of the definition of a gene and of the minimum unit of heredity."***

One of the major future challenges for ENCODE (and similarly ambitious projects) will be to capture the dynamic aspects of gene regulation. Most assays provide a single snapshot of cellular regulatory events, whereas a time series capturing how such processes change is preferable. Additionally, the examination of large batches of cells — as required for the current assays — may present too simplified a view of the underlying regulatory complexity, because individual cells in a batch (despite being genetically identical) can sometimes behave in different ways. The development of new technologies aimed at the simultaneous capture of multiple data types, along with their regulatory dynamics in single cells, would help to tackle these issues.

A further challenge is identifying how the genomic ingredients are combined to assemble the gene networks and biochemical pathways that carry out complex functions, such as cell-to-cell communication, which enable organs and tissues to develop. An even greater challenge will be to use the rapidly growing body



ENCODE
Encyclopedia of DNA Elements
nature.com/encode

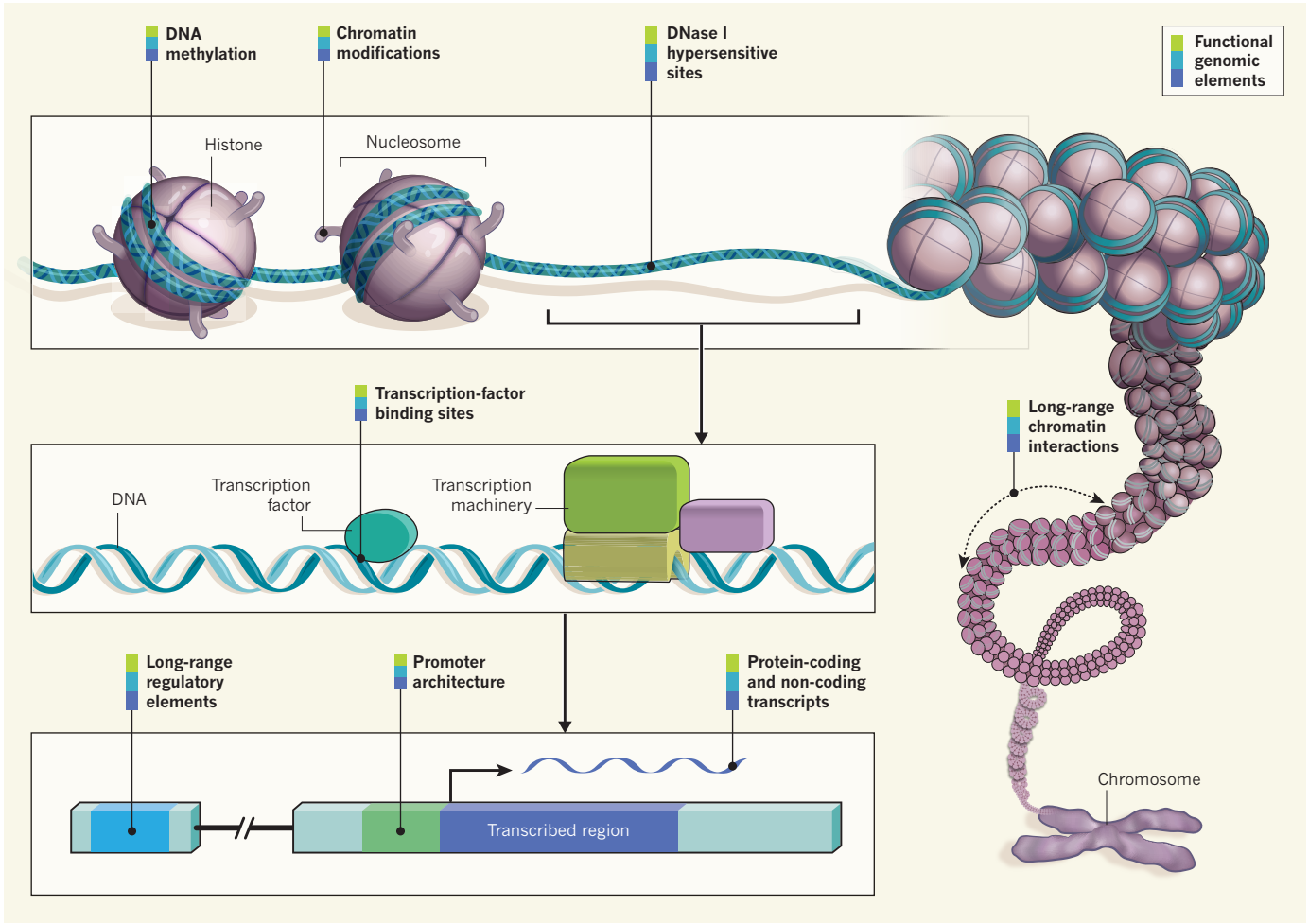


Figure 1 | Beyond the sequence. The ENCODE project[2–7] provides information on the human genome far beyond that contained within the DNA sequence — it describes the functional genomic elements that orchestrate the development and function of a human. The project contains data about the degree of DNA methylation and chemical modifications to histones that can influence the rate of transcription of DNA into RNA molecules (histones are the proteins around which DNA is wound to form chromatin). ENCODE also examines long-range chromatin interactions, such as looping, that alter the relative proximities of different chromosomal regions in three dimensions and also affect transcription. Furthermore, the project describes the binding activity of transcription-factor proteins and the architecture (location and sequence) of gene-regulatory DNA elements, which include the promoter region upstream of the point at which transcription of an RNA molecule begins, and more distant (long-range) regulatory elements. Another section of the project was devoted to testing the accessibility of the genome to the DNA-cleavage protein DNase I. These accessible regions, called DNase I hypersensitive sites, are thought to indicate specific sequences at which the binding of transcription factors and transcription-machinery proteins has caused nucleosome displacement. In addition, ENCODE catalogues the sequences and quantities of RNA transcripts, from both non-coding and protein-coding regions.

of data from genome-sequencing projects to understand the range of human phenotypes (traits), from normal developmental processes, such as ageing, to disorders such as Alzheimer's disease[10].

Achieving these ambitious goals may require a parallel investment of functional studies using simpler organisms — for example, of the type that might be found scampering around the floor, snatching up crumbs in the chefs' kitchen. All in all, however, the ENCODE project has served up an all-you-can-eat feast of genomic data that we will be digesting for some time. Bon appétit!

Joseph R. Ecker *is at the Howard Hughes Medical Institute and the Salk Institute for Biological Studies, La Jolla, California 92037, USA.*
e-mail: ecker@salk.edu

Expression control

WENDY A. BICKMORE

Once the human genome had been sequenced, it became apparent that an encyclopaedic knowledge of chromatin organization would be needed if we were to understand how gene expression is regulated. The ENCODE project goes a long way to achieving this goal and highlights the pivotal role of transcription factors in sculpting the chromatin landscape.

Although some of the analyses largely confirm conclusions from previous smaller-scale studies, this treasure trove of genome-wide data provides fresh insight into regulatory pathways and identifies prodigious numbers of regulatory elements. This is particularly so for Thurman and colleagues' data[4] regarding DNase I hypersensitive sites (DHSs) and for Gerstein and colleagues' results[6] concerning DNA binding of transcription factors. DHSs are genomic regions that are accessible to enzymatic cleavage as a result of the displacement of nucleosomes (the basic units of chromatin) by DNA-binding proteins (Fig. 1). They are the hallmark of cell-type-specific enhancers, which are often located far away from promoters.

The ENCODE papers expose the profusion of DHSs — more than 200,000 per cell type, far outstripping the number of promoters — and their variability between cell types. Through the simultaneous presence in the same cell type of a DHS and a nearby active promoter, the researchers paired half a million enhancers with their probable target genes. But this leaves

11 Years Ago

The draft human genome

OUR GENOME UNVEILED

Unless the human genome contains a lot of genes that are opaque to our computers, it is clear that we do not gain our undoubted complexity over worms and plants by using many more genes. Understanding what does give us our complexity — our enormous behavioural repertoire, ability to produce conscious action, remarkable physical coordination (shared with other vertebrates), precisely tuned alterations in response to external variations of the environment, learning, memory … need I go on? — remains a challenge for the future.

David Baltimore

From *Nature* 15 February 2001

GENOME SPEAK

With the draft in hand, researchers have a new tool for studying the regulatory regions and networks of genes. Comparisons with other genomes should reveal common regulatory elements, and the environments of genes shared with other species may offer insight into function and regulation beyond the level of individual genes. The draft is also a starting point for studies of the three-dimensional packing of the genome into a cell's nucleus. Such packing is likely to influence gene regulation … The human genome lies before us, ready for interpretation.

Peer Bork and Richard Copley

From *Nature* 15 February 2001

more than 2 million putative enhancers without known targets, revealing the enormous expanse of the regulatory genome landscape that is yet to be explored. Chromosome-conformation-capture methods that detect long-range physical associations between distant DNA regions are attempting to bridge this gap. Indeed, Sanyal and colleagues[7] applied these techniques to survey such associations across 1% of the genome.

The ENCODE data start to paint a picture of the logic and architecture of transcriptional networks, in which DNA binding of a few high-affinity transcription factors displaces nucleosomes and creates a DHS, which in turn facilitates the binding of further, lower-affinity factors. The results also support the idea that transcription-factor binding can block DNA methylation (a chemical modification of DNA that affects gene expression), rather than the other way around — which is highly relevant to the interpretation of disease-associated sites of altered DNA methylation[11].

The exquisite cell-type specificity of regulatory elements revealed by the ENCODE studies emphasizes the importance of having appropriate biological material on which to test hypotheses. The researchers have focused their efforts on a set of well-established cell lines, with selected assays extended to some freshly isolated cells. Challenges for the future include following the dynamic changes in the regulatory landscape during specific developmental pathways, and understanding chromatin structure in tissues containing heterogeneous cell populations.

Wendy A. Bickmore *is in the Medical Research Council Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK.*
e-mail: wendy.bickmore@igmm.ed.ac.uk

Non-coding but functional

INÊS BARROSO

The vast majority of the human genome does not code for proteins and, until now, did not seem to contain defined gene-regulatory elements. Why evolution would maintain large amounts of 'useless' DNA had remained a mystery, and seemed wasteful. It turns out, however, that there are good reasons to keep this DNA. Results from the ENCODE project[2–8] show that most of these stretches of DNA harbour regions that bind proteins and RNA molecules, bringing these into positions from which they cooperate with each other to regulate the function and level of expression of protein-coding genes. In addition, it seems that widespread transcription from non-coding DNA potentially acts as a reservoir for the creation of new functional molecules, such as regulatory RNAs.

What are the implications of these results for genetic studies of complex human traits and disease? Genome-wide association studies (GWAS), which link variations in DNA sequence with specific traits and diseases, have in recent years become the workhorse of the field, and have identified thousands of DNA variants associated with hundreds of complex traits (such as height) and diseases (such as diabetes). But association is not causality, and identifying those variants that are causally linked to a given disease or trait, and understanding how they exert such influence, has been difficult. Furthermore, most of these associated variants lie in non-coding regions, so their functional effects have remained undefined.

"The results imply that sequencing studies focusing on protein-coding sequences risk missing crucial parts of the genome."

The ENCODE project provides a detailed map of additional functional non-coding units in the human genome, including some that have cell-type-specific activity. In fact, the catalogue contains many more functional non-coding regions than genes. These data show that results of GWAS are typically enriched for variants that lie within such non-coding functional units, sometimes in a cell-type-specific manner that is consistent with certain traits, suggesting that many of these regions could be causally linked to disease. Thus, the project demonstrates that non-coding regions must be considered when interpreting GWAS results, and it provides a strong motivation for reinterpreting previous GWAS findings. Furthermore, these results imply that sequencing studies focusing on protein-coding sequences (the 'exome') risk missing crucial parts of the genome and the ability to identify true causal variants.

However, although the ENCODE catalogues represent a remarkable tour de force, they contain only an initial exploration of the depths of our genome, because many more cell types must yet be investigated. Some of the remaining challenges for scientists searching for causal disease variants lie in: accessing data derived from cell types and tissues relevant to the disease under study; understanding how these functional units affect genes that may be distantly located[7]; and the ability to generalize such results to the entire organism.

Inês Barroso *is at the Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK, and at the University of Cambridge Metabolic Research Laboratories and NIHR Cambridge Biomedical Research Centre, Cambridge, UK.*
e-mail: ib1@sanger.ac.uk

Evolution and the code

JONATHAN K. PRITCHARD & YOAV GILAD

One of the great challenges in evolutionary biology is to understand how differences in DNA sequence between species determine differences in their phenotypes. Evolutionary change may occur both through changes in protein-coding sequences and through sequence changes that alter gene regulation.

There is growing recognition of the importance of this regulatory evolution, on the basis of numerous specific examples as well as on theoretical grounds. It has been argued that potentially adaptive changes to protein-coding sequences may often be prevented by natural selection because, even if they are beneficial in one cell type or tissue, they may be detrimental elsewhere in the organism. By contrast, because gene-regulatory sequences are frequently associated with temporally and spatially specific gene-expression patterns, changes in these regions may modify the function of only certain cell types at specific times, making it more likely that they will confer an evolutionary advantage[12].

However, until now there has been little information about which genomic regions have regulatory activity. The ENCODE project has provided a first draft of a 'parts list' of these regulatory elements, in a wide range of cell types, and moves us considerably closer to one of the key goals of genomics: understanding the functional roles (if any) of every position in the human genome.

Nonetheless, it will take a great deal of work to identify the critical sequence changes in the newly identified regulatory elements that drive functional differences between humans and other species. There are some precedents for identifying key regulatory differences (see, for example, ref. 13), but ENCODE's improved identification of regulatory elements should greatly accelerate progress in this area. The data may also allow researchers to begin to identify sequence alterations occurring simultaneously in multiple genomic regions, which, when added together, drive phenotypic change — a process called polygenic adaptation[14].

However, despite the progress brought by the ENCODE consortium and other research groups, it remains difficult to discern with confidence which variants in putative regulatory regions will drive functional changes, and what these changes will be. We also still have an incomplete understanding of how regulatory sequences are linked to target genes. Furthermore, the ENCODE project focused mainly on the control of transcription, but many aspects of post-transcriptional regulation, which may also drive evolutionary changes, are yet to be fully explored.

Nonetheless, these are exciting times for studies of the evolution of gene regulation. With such new resources in hand, we can expect to see many more descriptions of adaptive regulatory evolution, and how this has contributed to human evolution.

Jonathan K. Pritchard *and* **Yoav Gilad** *are in the Department of Human Genetics, University of Chicago, Chicago 60637 Illinois, USA.* **J.K.P.** *is also at the Howard Hughes Medical Institute, University of Chicago.*
e-mails: pritch@uchicago.edu; gilad@uchicago.edu

From catalogue to function

ERAN SEGAL

Projects that produce unprecedented amounts of data, such as the human genome project[15] or the ENCODE project, present new computational and data-analysis challenges and have been a major force driving the development of computational methods in genomics. The human genome project produced one bit of information per DNA base pair, and led to advances in algorithms for sequence matching and alignment. By contrast, in its 1,640 genome-wide data sets, ENCODE provides a profile of the accessibility, methylation, transcriptional status, chromatin structure and bound molecules for every base pair. Processing the project's raw data to obtain this functional information has been an immense effort.

For each of the molecular-profiling methods used, the ENCODE researchers devised novel processing algorithms designed to remove outliers and protocol-specific biases, and to ensure the reliability of the derived functional information. These processing pipelines and quality-control measures have been adapted by the research community as the standard for the analysis of such data. The high quality of the functional information they produce is evident from the exquisite detail and accuracy achieved, such as the ability to observe the crystallographic topography of protein–DNA interfaces in DNase I footprints[5], and the observation of more than one-million-fold variation in dynamic range in the concentrations of different RNA transcripts[3].

"The high quality of the functional information produced is evident from the exquisite detail and accuracy achieved."

But beyond these individual methods for data processing, the profound biological insights of ENCODE undoubtedly come from computational approaches that integrated multiple data types. For example, by combining data on DNA methylation, DNA accessibility and transcription-factor expression. Thurman *et al.*[4] provide fascinating insight into the causal role of DNA methylation in gene silencing. They find that transcription-factor binding sites are, on average, less frequently methylated in cell types that express those transcription factors, suggesting that binding-site methylation often results from a passive mechanism that methylates sites not bound by transcription factors.

Despite the extensive functional information provided by ENCODE, we are still far from the ultimate goal of understanding the function of the genome in every cell of every person, and across time within the same person. Even if the throughput rate of the ENCODE profiling methods increases dramatically, it is clear that brute-force measurement of this vast space is not feasible. Rather, we must move on from descriptive and correlative computational analyses, and work towards deriving quantitative models that integrate the relevant protein, RNA and chromatin components. We must then describe how these components interact with each other, how they bind the genome and how these binding events regulate transcription.

If successful, such models will be able to predict the genome's function at times and in settings that have not been directly measured. By allowing us to determine which assumptions regarding the physical interactions of the system lead to models that better explain measured patterns, the ENCODE data provide an invaluable opportunity to address this next immense computational challenge. ■

Eran Segal *is in the Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel.*
e-mail: eran.segal@weizmann.ac.il

1. The ENCODE Project Consortium *Science* **306,** 636–640 (2004).
2. The ENCODE Project Consortium *Nature* **489,** 57–74 (2012).
3. Djebali, S. *et al. Nature* **489,** 101–108 (2012).
4. Thurman, R. E. *et al. Nature* **489,** 75–82 (2012).
5. Neph, S. *et al. Nature* **489,** 83–90 (2012).
6. Gerstein, M. B. *et al. Nature* **489,** 91–100 (2012).
7. Sanyal, A., Lajoie, B., Jain, G. & Dekker, J. *Nature* **489,** 109–113 (2012).
8. Birney, E. *et al. Nature* **447,** 799–816 (2007).
9. Mardis, E. R. *Nature* **470,** 198–203 (2011).
10. Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. *Annu. Rev. Med.* **63,** 35–61 (2012).
11. Sproul, D. *et al. Proc. Natl Acad. Sci. USA* **108,** 4364–4369 (2011).
12. Carroll, S. B. *Cell* **134,** 25–36 (2008).
13. Prabhakar, S. *et al. Science* **321,** 1346–1350 (2008).
14. Pritchard, J. K., Pickrell, J. K. & Coop, G. *Curr. Biol.* **20,** R208–R215 (2010).
15. Lander, E. S. *et al. Nature* **409,** 860–921 (2001).

Progress in Biophysics and Molecular Biology 113 (2013) 433–438

Contents lists available at SciVerse ScienceDirect

Progress in Biophysics and Molecular Biology

journal homepage: www.elsevier.com/locate/pbiomolbio

Original research

Artificial Neural Networks, and Evolutionary Algorithms as a systems biology approach to a data-base on fetal growth restriction

Maria E. Street [a,*], Massimo Buscema [b,c], Arianna Smerieri [a], Luisa Montanini [a], Enzo Grossi [b]

[a] Department of Pediatrics, University Hospital of Parma, Via Gramsci, 14-43126 Parma, Italy
[b] Semeion Research Centre of Sciences of Communication, Rome, Italy
[c] Dept. of Mathematical and Statistical Sciences, University of Colorado, Denver, CO, USA

ARTICLE INFO

Article history:
Available online 1 July 2013

Keywords:
IUGR
Systems biology
IGF
IL-6
IGFBP-2
Fetal growth

ABSTRACT

One of the specific aims of systems biology is to model and discover properties of cells, tissues and organisms functioning. A systems biology approach was undertaken to investigate possibly the entire system of intra-uterine growth we had available, to assess the variables of interest, discriminate those which were effectively related with appropriate or restricted intrauterine growth, and achieve an understanding of the systems in these two conditions. The Artificial Adaptive Systems, which include Artificial Neural Networks and Evolutionary Algorithms lead us to the first analyses. These analyses identified the importance of the biochemical variables IL-6, IGF-II and IGFBP-2 protein concentrations in placental lysates, and offered a new insight into placental markers of fetal growth within the IGF and cytokine systems, confirmed they had relationships and offered a critical assessment of studies previously performed.

1. Systems biology and the need for an approach that uses Artificial Adaptive Systems

Systems biology is a biology-based inter-disciplinary field of study that focuses on complex interactions within biological systems. Systems biology is a term used to describe an approach applied to biomedical and biological scientific research. Since the year 2000, this term has been used widely in biosciences in a variety of contexts.

One of the specific aims of systems biology is to model and discover properties of cells, tissues and organisms functioning as a system whose theoretical description is only possible using techniques which fall under the orders of systems biology.

At the basis of this new paradigm of systems biology research is a shift from reductionism to complexity. The theory of complexity conceives the behavior of a system as the result (or consequence) of the behavior of its parts (or agents). This intuition applies to all those scientific problems for which it is not possible to come to a complete formulation of the system's objective function. This happens in all those cases, in which it is not possible to derive the motion equations for the system parts or components from a top-down perspective.

The theory of complexity, instead, rests on the concept that a system's behavior can be described using a bottom-up approach, namely by obtaining the system's behavior as a result of the laws or rules followed by the agents in the system. This approach is used increasingly in sociology, political sciences, demography and environmental sciences (Tesfatsion and Judd, 2006; Myrskyla et al., 2009; Farmer and Foley, 2009). These considerations are the same that lie behind the development of Artificial Adaptive Systems algorithms (AAS, for short) which we shall describe below.

Complexity is related often to the high number of variables interacting in biological systems with many connections among these, including feedback loops, and usually several nonlinear relationships which become described by specific equations. However, it is important to note that complexity per se does not need necessarily a high number of variables and that "several" therefore has a relative meaning. For example; a properly interacting system of just three variables can show a highly complex pattern due to deterministic chaos. The remarkable progress in measurement techniques of chemical and biological species and concurrent

Abbreviations: IUGR, intra-uterine growth restriction; IGF, insulin-like growth factor; IGFBP, insulin-like growth factor binding protein; IL-6, interleukin-6; ANNS, Artificial Neural Networks; EA, Evolutionary Algorithms.

* Corresponding author. Tel.: +39 (0)521 702723; fax: +39 (0)521 702209.
E-mail address: mariaelisabeth.street@unipr.it (M.E. Street).

0079-6107/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.pbiomolbio.2013.06.003

substantial advances in mathematical theories behind urged the need of analyzing and understanding complex systems of fundamental importance.

Unfortunately, even the most powerful and well established statistical methods available to date, were developed in the first half of the past century when the scenario was dominated by acute infectious diseases, and the available information was much more simple, or at the most "complicated" rather than "complex" (i.e., Expert Systems and multivariate linear statistics).

It is now the time for fundamental questions as: does the mathematics used currently in medicine give the necessary information on the complexity of the diseases being investigated?

In a complex system each component loses its identity outside of the system of which it is part of. Complexity involves a new kind of mathematics, able to handle chaotic behavior, highly non-linear dynamics, and fractal geometry (Steeb, 2006; Witten and Frank, 2005; Mandelbrot, 1978; Zimmermann, 2006). There are a number of different reasons to apply complex system mathematics on predictive medicine and some of these are listed in Table 1.

The use of computers has opened the floodgates to methods of data collection that were impossible just a decade ago, solving the quantitative problem of information load, but computers are also responsible for allowing computationally intensive medical analyses with newer numerical algorithms addressing the qualitative challenge.

Therefore, computational and mathematical medicine is a new research area where:

a. Complex biological problems, whose theory is unclear, are represented as a large amount of atomic data, measuring actions, structures and functions of the biological landscape under study;
b. No predefined model is hypothesized for those data, but each atomic data uses a highly nonlinear algorithm to interact with the other data;
c. The interaction along time, among the data, changing the data, gives rise to a new dynamic model, top-down from the multiple and distributed local data interactions;
d. The computer becomes the environment within which this process happens and the place where this experimentation is validated in a double blind protocol.

In a few words: the computer becomes a second biological laboratory, and the AAS algorithms work as a meta-model, able to generate different models each time according to the data (data driven algorithms). Therefore, newer statistical approaches, based on new mathematical and logic assumptions broadly belonging to the artificial adaptive system family and complex theory setting, allow to tame intractable data sets. Seen in this perspective computer science (the science of algorithms) is now playing the role which mathematics did from the seventeenth through to the twentieth century: providing an orderly, formal framework and exploratory apparatus for the progress of knowledge.

Table 1
Motivations to apply complex system mathematics to predictive medicine.

Processes are based on complex networks of interacting genes and proteins. (Casey et al., 2012)
Health status is the consequence of dynamic processes that regulate these networks. (Xochitl, 2012)
Non-linear critical thresholds link to pathology. (Grossi, 2001)
Predictions need to be applied to individual patients
Huge amount of data per subject hamper statistical tests (President's council, 2008)

2. Artificial Adaptive Systems

The coupling of computer science with these new theoretical bases allows the creation of "intelligent" agents able to adapt themselves dynamically to problems of high complexity: the Artificial Adaptive Systems (AAS), which include Artificial Neural Networks (ANNs) and Evolutionary Algorithms (EA) (Grossi and Buscema, 2006, 2007).

ANNs and EA are able to reproduce the dynamic interaction of multiple factors simultaneously, allowing the study of complexity.

ANN and EA are adaptive models analyzing data which are inspired by the functioning processes of the human brain and of evolution. These are systems which are able to modify their internal structure in relation to a function objective. These are particularly suited for solving nonlinear type problems, being able to reconstruct the approximate rules that put a certain set of data – which describes the problem being considered – with a set of data which provides the solution (ANN) or to reconstruct the optimal data for a given set of rules or constraints (EA).

2.1. Artificial Neural Networks (ANN)

The basic elements of ANN are the nodes, also called processing elements (PE), and their connections. Each node has its own input, from which it receives communications from other nodes and/or from the environment and its own output, from which it communicates with other nodes or with the environment. Finally, each node has a function through which it transforms its own global input into an output (Fig. 1).

Each connection is characterized by the strength with which pairs of nodes are excited or inhibited. Positive values indicate excitatory connections, the negative ones, inhibitory connections.

The connections between the nodes can modify themselves over time. This dynamic starts as a learning process in the entire ANN. The way through which the nodes modify themselves is called "Law of Learning". The total dynamic of an ANN is tied to time. In fact, for the ANN to modify its own connections, the environment has to necessarily act on the ANN more times. Data are the environment which acts on the ANN.

The learning process is, therefore, one of the key mechanisms that characterize the ANN, which are considered adaptive processing systems. The learning process is one way to adapt the connections of an ANN to the data structure that make-up the environment and, therefore, a way to "understand" the environment and the relations that characterize it.

2.2. Evolutionary Algorithms (EA)

At variance with neural networks which are adaptive systems, able to discover the optimal hidden rules explaining a certain data set, Evolutionary Algorithms (EAs) are Artificial Adaptive Systems able to find optimal data when fixed rules or constraints must be respected. These are in other words optimization tools which become fundamental when the space of possible states in a dynamic system tends to be very huge.

An EA for example can help to distribute the original sample in two or more sub-samples with the aim of obtaining the maximum performance possible from an ANN that is trained on the first sample and tested on the second. In order to limit eventual optimistic polarizations in the evaluation of the performance, it is possible also, to reverse the two samples and to consider the mean between the two approximations obtained as fitness of the algorithm and as an estimate of the model's quality (Buscema et al., 2005).

EA can approach also the problem of selecting the variables most related to a particular outcome, without the use of linear

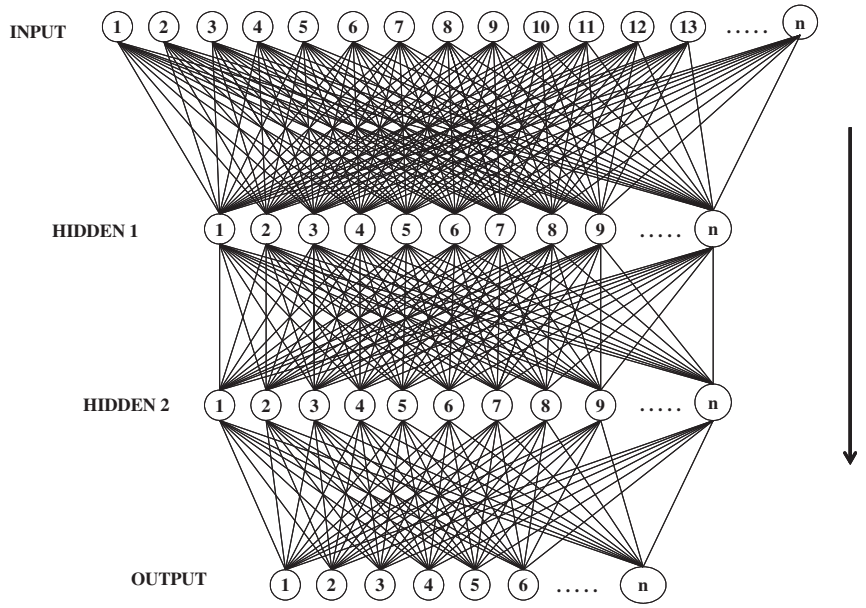


Fig. 1. Typical neural network architecture. The basic elements of ANN are the nodes, also called processing elements (PE), and their connections. Each node has its own input, from which it receives communications from other nodes and/or from the environment and its own output, from which it communicates with other nodes or with the environment. Finally, each node has a function through which it transforms its own global input into an output.

correlation. When linear systems are used, the correlation index, which indicates the degree of relationship existing between the input and output variables of the system, suggests which of the variables available should be used in order to build a model of the problem.

The problem of selecting a subset of variables on which to build "the model for the process under examination" stems from the fact that, when data are gathered to build a Data Base, the relationship between the collected variables and the function of the process being examined is unknown. In this case the natural approach is to include all of the variables that may have a connection with the event being studied. The result of this approach is that often a series of variables, which do not contain any information regarding the process being examined, are present. These variables, introduced into the model, cause an increase of the noise, and, therefore, a greater difficulty for the ANN to learn the data correctly.

The coupling of ANN and EA brings to the concept of artificial organisms, able to optimize classification performance and predictability.

The use of Artificial Adaptive Systems (AAS) in biology and medicine is rapidly spreading, underlining a growing acknowledgment of its relevance as a fundamental tool of research and experimentation. A clear sign of this trend is the proliferation of papers that explicitly address this approach or use it to test a given theory or empirical procedure. Moreover, the publications are "percolating" from highly specialized journals, to more widespread journals.

3. Intra-uterine growth restriction

Most cases of fetal growth restriction (IUGR) are still of unknown origin (Resnik, 2002). The interest in IUGR has grown because the concept of a "Fetal Origin of Adult Disease" has developed with time to describe modifications in utero that might influence adult patho-physiology (Karlberg and Albertsson-Wikland, 1995; Godfrey and Barker, 2000).

Many data are present in the Literature that have analysed and compared separately single or few peptides between IUGR and appropriate for gestational age newborns (AGA), mainly studies performed on serum samples from the mothers, most a few years after birth, few on cord serum and placental samples (Street et al., 2006a).

The system is certainly very complex, and many subsystems are involved, however, some are yet unidentified, others are partially known, and it would be virtually impossible at present to put them all together. Therefore, we focused our attention mainly of the IGF system which is well recognized to be crucial for fetal growth. In particular, this has been proven by experiments in knockout mice (DeChiara et al., 1990; Baker et al., 1993; Liu et al., 1993; Ohlsson et al., 1989). Furthermore, the main peptides of this system, IGF-I and IGF-II are both known to be synthesised in the placenta (Shen et al., 1986; Wang et al., 1998; Zollers et al., 2001).

We previously showed that IGF-I, IGFBP-2 and IGF-II gene expressions were increased in the placentas and that cord serum IGFBP-1 and IGFBP-2 concentrations were also increased in IUGR newborns (Street et al., 2006a).

Cytokines are thought to play also an important role in regulating placental development and growth although they have been poorly studied (Bartha, 2003). In different conditions cytokine and IGF interactions have been previously shown (Street et al., 2006b, 2008b; Senn et al., 2002).

We showed (Street et al., 2006a) that IL-6 gene expression and protein content in the placenta were significantly increased in IUGR newborns and were positively related with IGFBP-1 and IGFBP-2 gene expressions, suggesting cytokine and IGF system relationships in the placenta. However, relationships are difficult to demonstrate in observational studies, in particular, when linear statistical methods are being applied to virtually non-linear relationships, thus, we had no certainty that this really existed and was crucial to "the system of intra-uterine growth".

Furthermore, the selection of cases is extremely important in observational studies, and when referring to IUGR, one must consider at least gestational age, gender, and known causes of IUGR. It is also unknown how gene expression is related to protein content and metabolism, and clinical information is different from maternal serum data which is different to that of the fetus, and different from the information obtained from placenta.

Factors from different compartments, and different kinds of information could explain differently IUGR, and some information might be relevant whereas some other might not, and some information might be hidden to common reasoning. These are the main reasons for a systems biology approach to be undertaken to investigate a system in a wider/comprehensive manner as possible. The ultimate aim was to discriminate variables effectively related with appropriate or restricted intrauterine growth, and achieve an understanding of the systems in these two conditions (Street et al., 2008a).

Artificial Neural Networks have been used to predict the risk of mortality in very low birth weight newborns (Namasivayam and Waldemar, 2001) and to predict extubation outcome in preterm infants proving that they can be particularly useful to evidence-based medicine (Mueller et al., 2004). Other studies have used a systems biology approach to investigate whether a mouse model is actually useful to understand the human placenta. Cox et al. (2009) to address the effects of tobacco smoking on embryonic development (Feltes BC et al., 2013), and evaluate how development and aging are interconnected (Feltes BC, 2011), however, in the Literature to date studies addressing specifically intra-uterine growth are substantially lacking, and thus, it is difficult to compare any of our findings with the work of other authors.

As example of the potential role of AAS in IUGR, we will describe their use in a small data set, which has been object of previous studies (Street, 2006a,b; Smerieri, 2011) in which we followed from pregnancy 20 IUGR and 29 AGA births of comparable gestational age (35.3 ± 0.5 vs 36.6 ± 0.5 weeks, respectively, n.s.). All pregnancies were dated correctly by ultrasound during the first trimester of gestation, and only cases of idiopathic IUGR were included in the studies.

AAS seemed to us particularly appropriate to accomplish the task of creating a modular but comprehensive model to study fetal development for the following reasons:

a) AAS focuses upon the behavior of agents as part of a complex system. These take into account the rules according to which they act and interact;
b) It has been demonstrated by many authors that AAS are capable of handling agent heterogeneity (Lisboa, 2002; Zou et al., 2008);
c) AAS make it possible to integrate different disciplinary perspectives. A particular portion of the system architecture can be described by a separate set of mathematical equations describing the behavior of the "agent" represented;
d) AAS allow to structure the software architecture across different levels of analysis and to explicitly reveal in the model the understanding of the rules with which agents interact.

3.1. Understanding intra-uterine growth using supervised ANNs

In recent years ANNs have been used successfully in medicine as previously mentioned (Penco et al., 2005; Baldassarre et al., 2004; Lahner et al., 2005; Mecocci et al., 2002; Lapuerta et al., 1998; Buchman et al., 1994; DiRusso et al., 2002), and have been applied to solve clinical problems in pediatrics including those related to preterm births (Mueller et al., 2004; Jaing et al., 2001; DeGroff et al., 2001; Namasivayam and Waldemar, 2001; Zernikow et al., 1998).

As previously explained, first we fed to the system all the variables we considered in our data base (Table 2), then we identified, using the optimisation method, the most significant (Table 3), and subsequently worked with these. To do this we used ANNS.

The following step took into consideration that if the relationships among variables were nonlinear, they were not able to preserve, with adequate accuracy, the geometrical structure of the original space. Moreover, mapping is generally based on a specific kind of "distance" among variables and gives origin to a "static" projection of possible associations losing the intrinsic dynamics due to active interactions of variables in living systems of the real world (Buscema and Grossi, 2007). A new paradigm of variable mapping which creates a sort of semantic connectivity has overcome, however, these limitations. This method was described by Buscema and Grossi (2008), and, thus, applied.

Using this latter method and supervised ANNS we showed that IGF-II, IGFBP-2 and IL-6 concentrations in placental lysates were the most important determinants of foetal growth (Tables 3 and 4, Fig. 2), and could be potentially interesting for the development of future therapeutic interventions possibly aiming at reducing IL-6

Table 2
List of variables analysed using Artificial Neural Networks (ANNS) and their linear correlation index (r^2) with each target variable.

Variables		Correlation index (r^2)
1	Chronological age of the mother at delivery (years)	0.11
2	Sex of the newborn	0
3	Gestational age (weeks)	0.22
4	Other siblings (yes or No)	0.001
5	Number of other siblings	0.005
6	Number of previous abortions	0.000
7	Total placental protein content in lysates (mg/ml)	0.007
8	IGF-II concentration in placental lysates (ng/mg)	0.16
9	IGFBP-2 concentration in placental lysates (ng/mg)	0.05
10	IL-6 concentration in placental lysates (pg/mg)	0.09
11	TNF-α concentration in placental lysates (ng/mg)	0.05
12	IGF-I relative gene expression in placenta (A.U./18S, UBQ)	0.01
13	IGF-II relative gene expression in placenta (A.U./18S, UBQ)	0.02
14	IGFBP-1 relative gene expression in placenta (A.U./18S, UBQ)	0.05
15	IGFBP-2 relative gene expression in placenta (A.U./18S, UBQ)	0.08
16	IL-6 relative gene expression in placenta (A.U./18S, UBQ)	0.16

Modified Street et al., 2008a.

Table 3
Variables selected by the optimized (I.S.) system which were subsequently administered to the ANNs.

Variables	IS 1° selection	IS 2° selection
Chronological age of the mother at delivery	X	X
Sex of the newborn		
Gestational age	X	X
Other siblings	X	X
Number of other siblings		
Number of previous abortions		
Total placental protein content in lysates		
IGF-II concentration in placental lysates	X	X
IGFBP-2 concentration in placental lysates	X	X
IL-6 concentration in placental lysates		
TNF-α concentration in placental lysates		
IGF-I relative gene expression in placenta	X	X
IGF-II relative gene expression in placenta		X
IGFBP-1 relative gene expression in placenta		
IGFBP-2 relative gene expression in placenta	X	X
IL-6 relative gene expression in placenta		X

From Street et al., 2008a.

Table 4
Summary of results obtained by different ANN models on the variables of the first (7 variables-Table 2) and second (9 variables-Table 2) random selection from the data base using an optimisation protocol (IS system). The columns, IUGR% and AGA% give the percentage of correct diagnoses based on the variables selected by the optimization protocol, reported in Table 3.

ANN model	Target		Mean accuracy %		Errors
	IUGR (%)	AGA (%)	Arithmetic	Weighted	
BM 7 VAR.	91.67	84.62	88.14	88.00	3
BM 9 VAR.	91.67	76.92	84.29	84.00	4
BP 7 VAR.	91.67	84.62	88.14	88.00	3
BP 9 VAR.	91.67	84.62	88.14	88.00	3
SN 7 VAR.	75.00	92.31	83.65	84.00	4
SN 9 VAR.	75.00	84.62	79.81	80.00	5

IUGR: Intra-uterine growth restriction.
AGA: appropriate for gestational age.
The ANN used were: BM Bimodal neural network; BP: Back propagation neural network; SN: Sine net neural network.
Modified Street et al., 2008a.

and IGFBP-2 concentrations to preserve IGF bioactivity in both placenta and fetus (Street et al., 2008b).

Concluding, the emerging picture was that IL-6, and IGF system peptides in placenta, although with some differences, were ruling factors in intra-uterine growth, both in conditions of appropriate fetal growth and intra-uterine growth restriction. These results overall offered a new insight into placental players of fetal growth within the IGF and cytokine systems.

These finding were in agreement with experimental data in in vitro studies, in animals, and in humans where the IGF-II peptide is a well recognized and important determinant of fetal growth (DeChiara et al., 1990; Baker et al., 1993; Liu et al., 1993; Ohlsson et al., 1989; Shen et al., 1986). Our previous data showed increased IGF-2 placental concentrations and gene expression in IUGR newborns, reflecting a possible protective mechanism to promote growth in unfavorable conditions (Street et al., 2006a).

IGFBP-2 could have yet unknown effects in utero on fetal growth and on placental metabolism. Altogether, to date, IGFBP-2 has been poorly studied, and previously was not considered an important bio-regulator of IGF bio-availability (Han et al., 1996). In cord serum, we previously showed a clear positive effect of IGF-II, and negative effect of IGFBP-2 on both birth length and weight (Smerieri et al., 2011).

IL-6 has been studied only recently in placenta, and with respect to fetal growth, and few significant data are available in humans

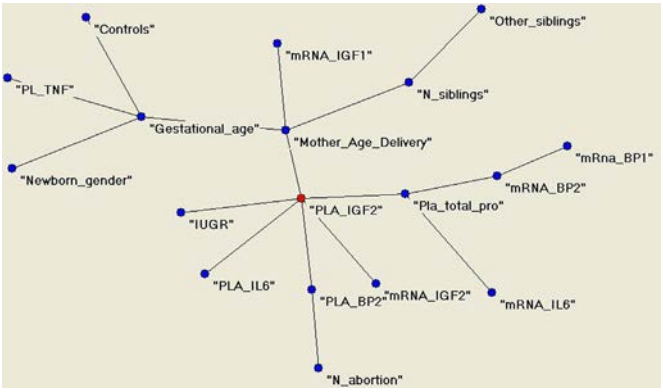


Fig. 2. Connectivity map clarifying the clusters of variables, and single relationships among variables. The placental IGF-II content resulted in a central node (in red), and closely related with IUGR, and placental content in IL-6, IGFBP-2. It was also directly a function of its gene expression and of total placental protein content. Interestingly it was also dependent on the mothers age at delivery.

(Amu et al., 2006; Nahum et al., 2004) showing both unchanged (Bartha et al., 2003) and increased IL-6 (Nahum et al., 2004). Studies have been performed in the maternal serum (Elfayomy et al., 2013), and in cord serum, showing increased IL-6 when acquisition of lipid tissue by the fetus is inadequate (Martos-Moreno et al., 2009), in fetal distress (Hsta et al., 1996), and when birth weight is lower than expected (Amarilyo et al., 2011). IL-6 has been described reduced in the presence of eclampsia (Ødegård et al., 2001).

With the data from our analyses we confirmed a central role of IL-6 content in placenta in IUGR. We showed previously that IL-6 mRNA was significantly increased in the placenta of IUGR newborns (Street et al., 2006a). This pro-inflammatory cytokine is of particular interest as it has shown interactions with the IGF system in many chronic inflammatory diseases (Street et al., 2006b, 2008b; Senn et al., 2002), and interesting molecular mechanisms of insulin-resistance have been shown (Ozes et al., 2001; Klover et al., 2003; Li et al., 2010; Aguirre et al., 2002). Insulin-resistance is considered to be the cause of the metabolic syndrome in later life, and subjects born IUGR have been shown to have a greater prevalence of this condition compared with subjects born AGA.

4. Advantages and disadvantages of Artificial Adaptive Systems (AAS)

Once the problem is reproduced through the model, one gets an extraordinary tool for research and experimentation: a virtual laboratory. Hence, one can explore the many combinations deriving from the interactions of agents by means of virtual experiments. Compared to the real experiments there are a number of advantages, and, of course, some disadvantages. The first advantage consists in the speed at which it is possible to run the experiment, and to analyze it. This can lead to time saving of three or four order of magnitudes. The second advantage is that the results have the strength of evidence driven data, because an agent-based model can be seen as a dynamic system which products are just optimal functions. Of course, the empirical validity of such results depends on the validity of the model and on the calibration and validation of its parameters. The third advantage is that building an AAS forces the researcher to an extreme clarity of his ideas about the problem under modeling. The fourth advantage consists in the relatively easy reproducibility and extendibility of the results, especially if the source code of the computer program were made publicly available in some sort of open access style (Grossi and Buscema, 2007b).

The major disadvantages are that building, calibrating, validating, planning experiments, and interpreting results is all but easy. These tasks require time, expertise, and a strong collaboration between the modelers and the researchers. Basically, proper teamwork is required. Moreover, such difficulties grow with the complexity of the model, which, in turn, replicates the complexity of the problem under investigation. Indeed, as for all forms of bioinformatics, biological and medical laboratories require increasingly, inter-disciplinary skills and heterogenous theoretical and methodological perspectives.

References

Aguirre, V., Werner, E.D., Giraud, J., Lee, Y.H., Shoelson, S.E., White, M.F., 2002. Phosphorylation of Ser307 in insulin receptor substrate-1 blocks interactions with the insulin receptor and inhibits insulin action. J. Biol. Chem. 277, 1531–1537.

Amalilyo, G., Oren, A., Mimouni, F.B., Ochshorn, Y., Deutsch, V., Mandel, D., 2011. Increased cord serum inflammatory markers in small-for-gestational-age neonates. J. Perinatol. 31, 30–32.

Amu, S., Hahn-Zoric, M., Malik, A., Ashraf, R., Zaman, S., Kjellmer, I., Hagberg, H., Padyukov, L., Hanson, L.A., 2006. Cytokines in the placenta of Pakistani newborns with and without intraueterine growth retardation. Pediatr. Res. 59, 254–258.

Baldassarre, D., Grossi, E., Buscema, M., Intraligi, M., Amato, M., Tremoli, E., Pustina, L., Castelnuovo, S., Sanvito, S., Gerosa, L., Sirtori, C.R., 2004. Recognition of patients with cardiovascular disease by artificial neural networks. Ann. Med. 36, 630–640.

Baker, J., Liu, J.P., Robertson, E.J., Efstratiadis, A., 1993. Role of insulin-like growth factors in embryonic and postnatal growth. Cell 75, 73–82.

Bartha, J.L., Romero-Carmona, R., Comino-Delgrado, R., 2003. Inflammatory cytokines in intrauterine growth retardation. Acta Obstet. Gynecol. Scand. 82, 1099–1102.

Buchman, T.G., Kubos, K.L., Seidler, A.J., Siegforth, M.J., 1994. A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care unit. Crit. Care Med. 22, 750–762.

Buscema, M., Grossi, E., Intraligi, M., Garbagna, N., Andriulli, A., Breda, M., 2005. An optimized experimental protocol based on neuro-evolutionary algorithms application to the classification of dyspeptic patients and to the prediction of the effectiveness of their treatment. Artif. Intell. Med. 34, 279–305.

Buscema, M., Grossi, E., 2007. A novel adapting method for emergent properties discovery in data bases: experience in medical field. In: Proceedings IEEE International Conference on Systems, Man, and Cybernetics (SMC 2007), pp. 3457–3463.

Buscema, M., Grossi, E., 2008. The semantic connectivity map: an adapting self-organizing knowledge discovery method in data bases. Experience in gastro-oesophageal reflux disease. Int. J. Data Min. Bioinform. 2, 362–404.

Casey, S., Greene, O.G., Troyanskaya, 2012. Data-driven view of disease biology. In: Translational bioinformatics. (Chapter 2), PLOS Computational Biol. http://dx.doi.org/10.1371/journal.pcbi.1002816. Published 27 Dec 2012.

Cox, B., Kotlyar, M., Evangelou, A.I., Ignatchenko, V.L., Whiteley, K., Jurisica, I., Adamson, S.L., Rossant, J., Kislinger, T., 2009. Comparative systems biology of human and mouse as a tool to guide the modeling of human placental pathology. Mol. Syst. Biol. 5, 279–293.

DeChiara, T.M., Robertson, E.J., Efstratiadis, A., 1990. A growth-deficient phenotype in heterozygous mice carrying an insulin-like growth factor-II gene disrupted by targeting. Nature 344, 78–80.

DeGroff, C.G., Bhatikar, S., Hertzberg, J., Shandas, R., Valdes-Cruz, L., Mahajan, R.L., 2001. Artificial neural network-based method of screening heart murmurs in children. Circulation 103, 2711–2716.

DiRusso, S.M., Chahine, A.A., Sullivan, T., Risucci, D., Nealon, P., Cuff, S., Savino, J., Slim, M., 2002. Development of a model for prediction of survival in pediatric trauma patients: comparison of artificial neural networks and logistic regression. J. Pediatr. Surg. 37, 1098–1104.

Elfayomy, A.K., Habib, F.A., Almasry, S.M., Safwat, M.D., Eldomiaty, M.A., 2013. Serum levels of adrenomedullin and inflammatory cytokines in women with term idiopathic intrauterine growth restriction. J. Obstet. Gynaecol. 33, 135–139.

Farmer, J.D., Foley, D., 2009. The economy needs agent-based modeling. Nature 460, 685–686.

Feltes, B.C., de Faria Poloni, J., Bonatto, D., 2011. The developmental aging and origins of health and disease hypotheses explained by different protein networks. Biogerontology 12, 293–308.

Feltes, B.C., de Faria Poloni, J., Notari, D.L., Bonatto, D., 2013. Toxicological effects of the different substances in tobacco smoke on human embryonic development by a systems chemo-biology approach. PLoS One 8, e61743.

Godfrey, K.M., Barker, D.J., 2000. Fetal nutrition and adult disease. Am. J. Clin. Nutr. 71, 1344S–1352S.

Grossi, E., 2001. Non linearity in medicine: a problem or an opportunity? Br. Med. J.. Online: <http://www.bmj.com/rapid-response/2011/10/28/non-linearity-medicine-problem-or-opportunity>.

Grossi, E., Buscema, M., 2006. Artificial intelligence and outcome research. Drug Dev. Res. 67, 227–244.

Grossi, E., Buscema, M., 2007. Introduction to artificial neural networks. Eur. J. Gastroenterol. Hepatol. 19, 1046–1054.

Han, V.K., Bassett, N., Walton, J., Challis, J.R.G., 1996. The expression of insulin-like growth factor (IGF) and IGF-Binding Protein (IGFBP) genes in the human placenta and membranes: evidence for IGF-IGFBP interactions at the feto-maternal interface. J. Clin. Endocrinol. Metab. 81, 2680–2693.

Hata, T., Kawamura, T., Inada, K., Fujiwaki, R., Ariyuki, Y., Hata, K., Kitao, M., 1996. Interleukin-6, interleukin-8, and granulocyte elastase in newborns with fetal distress. Gynecol. Obstet. Invest. 42, 174–177.

Jaing, J.T., Sepulveda, J.A., Casillas, A.M., 2001. Novel computer-based assessment of asthma strategies in inner-city children. Ann. Allergy Asthma Immunol. 87, 230–237.

Karlberg, J., Albertsson-Wikland, K., 1995. Growth in full-term small-for-gestational-age infants: from birth to final height. Pediatr. Res. 995 (38), 733–739.

Klover, P.J., Zimmers, T.A., Koniaris, L.G., Mooney, R.A., 2003. Chronic exposure to interleukin-6 causes hepatic insulin resistance in mice. Diabetes 52, 2784–2789.

Lahner, E., Grossi, E., Intraligi, M., Buscema, M., Delle Fave, G., Annibale, B., 2005. Possible contribution of advanced statistical methods (artificial neural networks and linear discriminant analysis) in the recognition of patients with suspected atrophic body gastritis. World J. Gastroenterol. 11, 5867–5873.

Lapuerta, P., L'Italien, G.J., Paul, S., Hendel, R.C., Leppo, J.A., Fleischer, L.A., Cohen, M.C., Eaglek, A., Giugliano, R.P., 1998. Neural network assessment of perioperative cardiac risk in vascular surgery patients. Med. Decis. Making 18, 70–75.

Li, Y., Samuel, D.J., Sundaraj, K.P., Lopes-Virella, M.F., Huang, Y., 2010. IL-6 and high glucose synergistically upregulate MMP-1 expression by U937 mononuclear phagocytes via ERK 1/2 and JNK pathways and c-Jun. J. Cell. Biochem. 110, 248–259.

Lisboa, P.J., 2002. A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Netw. Jan. 15, 11–39.

Liu, J.P., Baker, J., Perkins, A.S., Robertson, E.J., Efstratiadis, A., 1993. Mice carrying null mutations of the genes encoding insulin-like growth factor-I (IGF-I) and Type-1 IGF Receptor (Igf1r). Cell 75, 59–72.

Maldelbrot, B.B., 1978. The Factal Geometry of Nature. Freeman, New York.

Martos-Moreno, G.A., Barrios, V., Saenz de Pipaon, M., Pozo, J., Dorronsoro, I., Martinez-Biarge, M., Quero, J., Argente, J., 2009. Influence of prematurity and growth restriction on the adipokine profile, IGF1, and gherlino levels in cord blood: relationship with glucose metabolism. Eur. J. Endocrinol. 161, 381–389.

Mecocci, P., Grossi, E., Buscema, M., Intraligi, M., Savare, R., Rinaldi, P.C., Cherubini, A., Sanin, U., 2002. Use of artificial networks in clinical trials: a pilot study to predict responsiveness to Donezepil in Alzheimer's disease. J. Am. Geriatr. Soc. 50, 1857–1860.

Mueller, M., Wagner, C.L., Annibale, D.J., Hulsey, T.C., Knapp, R.G., Almeida, J.S., 2004. Predicting extubation outcome in preterm newborns: a comparison of neural networks with clinical expertise and statistical modelling. Pediatr. Res. 56, 11–18.

Myrskylä, M., Kohler, H.P., Billari, F.C., 2009. Advances in development reverse fertility declines. Nature 460, 741–743.

Namasivayam, A., Waldemar, A.C., 2001. Comparison of the prediction of extremely low birth weight neonatal mortality by regression analysis and by neural networks. Early Hum. Dev. 65, 123–137.

Nahum, R., Brenner, O., Zahalka, M.A., Traub, L., Quintana, F., Moroz, C., 2004. Blocking of the placental immune-modulatory ferritin activates Th1 type cytokines and affects placenta development, fetal growth and the pregnancy out come. Hum. Reprod. 19, 715–722.

Ødegård, R.A., Vatten, L.J., Nilsen, S.T., Salvesen, K.A., Vefring, H., Austgulen, R., 2001. Umbilical cord plasma interleukin-6 and fetal growth restriction in pre-eclampsia: a prospective study in Norway. Obs. Gynecol. 98, 289–294.

Ohlsson, R., Holmgren, L., Glaser, A., Szpecht, A., Pfeifer-Ohlsson, S., 1989. Insulin-like growth factor 2 and short-range stimulatory loops in control of human placental growth. EMBO J. 8, 1993–1999.

Ozes, O.N., Akca, H., Mayo, L.D., Gustin, J.A., Maehama, T., Dixon, J.E.., Donner, D.B., 2001. A phosphatidylinositol 3-kinase/Akt/mTOR pathway mediates and PTEN antagonizes tumor necrosis factor inhibition of insulin signalling through insulin receptor substrate-1. Proc. Natl. Acad. Sci. U S A 98, 4640–4645.

Penco, S., Grossi, E., Cheng, S., Intraligi, M., Maurelli, G., Patrosso, M.C., Marocchi, A., Buscema, M., 2005. Assessment of the role of genetic polymorphism in venous thrombosis through artificial neural networks. Ann. Hum. Gen. 69, 693–706.

President's Council Of Advisors On Science And Technology, September 2008. Priorities for Personalized Medicine. Washington, USA.

Resnik, R., 2002. Intrauterine growth restriction. Obs. Gynecol. 99, 490–496.

Senn, J.J., Klover, P.J., Nowak, I.A., Mooney, R.A., 2002. Interleukin-6 induces cellular insulin resistance in hepatocytes. Diabetes 51, 3391–3399.

Shen, S.J., Wang, C.Y., Nelson, K.K., Jansen, M., Ilan, J., 1986. Expression of insulin-like growth factor II in human placentas from normal and diabetic pregnancies. Proc. Natl. Acad. Sci. U S A 83, 9179–9182.

Smerieri, A., Petraroli, M., Ziveri, M.A., Volta, C., Bernasconi, S., Street, M.E., 2011. Effects of cord serum insulin, IGF-II, IGFBP-2, IL-6 and cortisol concentrations on human birth weight and length: pilot study. PLoS One 6, e29562.

Steeb, W.H., 2006. The Non Linear Workbook. World Scientific, NJ.

Street, M.E., Seghini, P., Fieni, S., Ziveri, M.A., Volta, C., Martorana, D., Gramellini, D., Bernasconi, S., 2006a. Interleukin (IL)-6 and IGF-IGFBP relationships in placenta and cord blood: possibile determinants of fetal growth restrction (IUGR). Eur. J. Endocrinol. 155, 567–574.

Street, M.E., Ziveri, M.A., Spaggiari, C., Viani, I., Volta, C., Grzincich, G.L., Virdis, R., Bernasconi, S., 2006b. Inflammation is a modulator of the IGF-IGFBP system inducing reduced bioactivity of IGFs in cystic fibrosis. Eur. J. Endocrinol. 154, 1–7.

Street, M.E., Grossi, E., Volta, C., Faleschini, E., Bernasconi, S., 2008a. Placental determinants of fetal growth: identification of key factors in the IGF and cytokine systems using artificial neural networks. BMC Pediatr. 8, 24.

Street, M.E., Volta, C., Ziveri, M.A., Zanacca, C., Banchini, G., Viani, I., Rossi, M., Virdis, R., Bernasconi, S., 2008b. Changes and relationships of IGFs and IGFBPs and cytokines in coeliac disease at diagnosis and on gluten-free diet. Clin. Endocrinol. 68, 22–28.

Tesfatsion, L., Judd, K., 2006. Handbook of Computational Economics. Elsevier, Amsterdam.

Wang, C.Y., Daimon, M., Shen, S.J., Engelman, G.L., Ilan, J., 1998. Insulin-like growth factor-I messenger ribonucleic acid in the developing human placenta and in term placenta of diabetics. Mol. Endocrinol. 2, 217–229.

Witten, I.H., Frank, E., 2005. Data Mining. Morgan Kaufmann Publishers, Elsevier, San Francsico.

Xochitl, C., Morgan, L., Huttenhower, C., 2012. Human microbiome analysis. In: Translational bioinformatics. (Chapter 12), PLOS Computational Biol. http://dx.doi.org/10.1371/journal.pcbi.1002816. Published 27 Dec 2012.

Zernikow, B., Holtmannspoetter, K., Michel, E., Theilhaber, M., Pielemeier, W., Hennecke, K.H., 1998. Artificial neural network for predicting intracranial haemorrhage in preterm neonates. Acta Paediatr. 87, 969–975.

Zimmermann, H.J., 2006. Fuzzy Theory Set and Its Application. Kluwer Academic Publisher, Boston.

Zollers, W.G., Babischkin, J.S., Pepe, G.J., Albrecht, E.D., 2001. Developmental regulation of placental insulin-like growth factor (IGF)-II and IGF-binding protein-1 and −2 messenger RNA expression during primate pregnancy. Biol. Reprod. 165, 1208–1214.

Zou, J., Han, Y., So, S.S., 2008. Overview of artificial neural networks. Methods Mol. Biol. 458, 15–23.

Spring 2017 – Epigenetics and Systems Biology
Lecture Outline (Systems Biology)
Michael K. Skinner – Biol 476/576
CUE 418, 10:35-11:50 am, Tuesdays & Thursdays
January 24 & 31, 2017
Weeks 3 and 4

Systems Biology (Components & Technology)

Components (DNA, Expression, Cellular, Organ, Physiology, Organism, Differentiation, Development, Phenotype, Evolution)

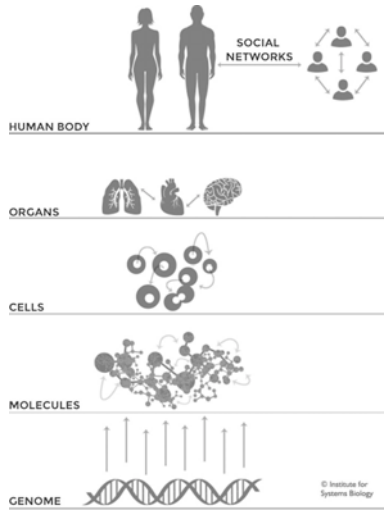
Technology (Genomics, Transcriptomes, Proteomics) (Interaction, Signaling, Metabolism)

Omics (Data Processing and Resources)

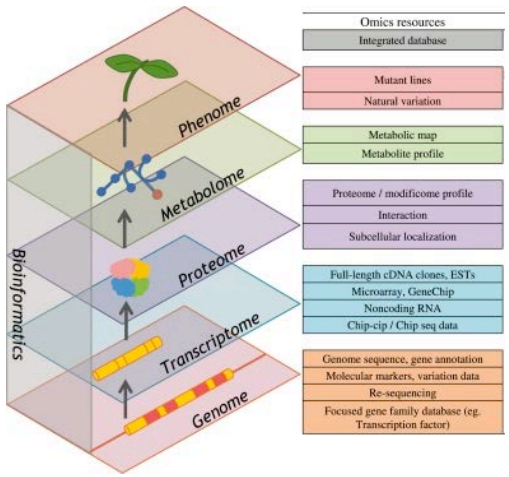
Required Reading

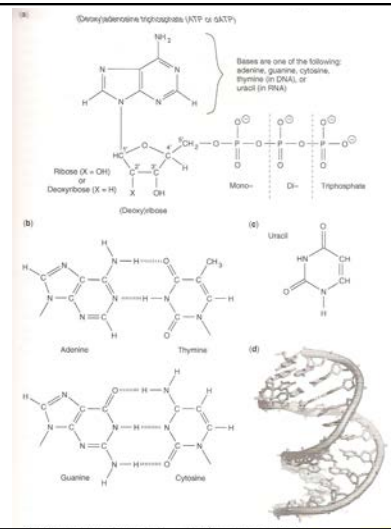
ENCODE (2012) ENCODE Explained. Nature 489:52-55.

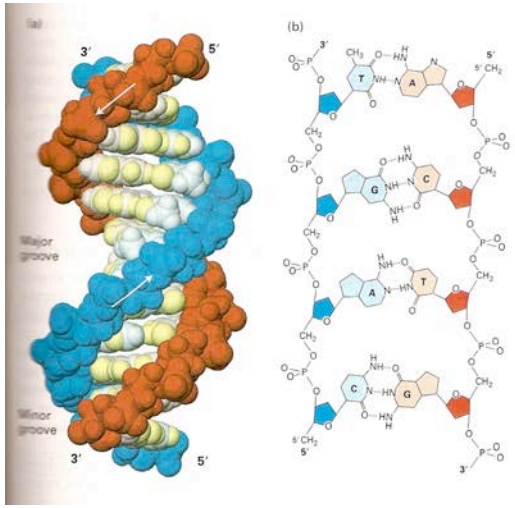
Street ME, et al. (2013) Artificial Neural Networks, and Evolutionary Algorithms as a systems biology approach to a data-base on fetal growth restriction. Prog Biophys Mol Biol. 113(3):433-8.

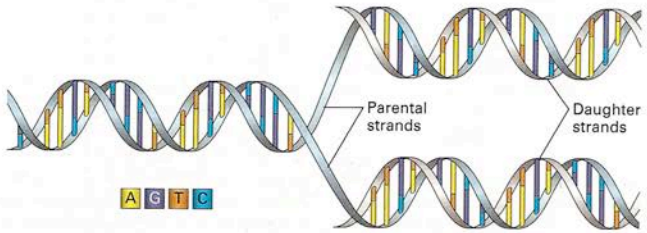


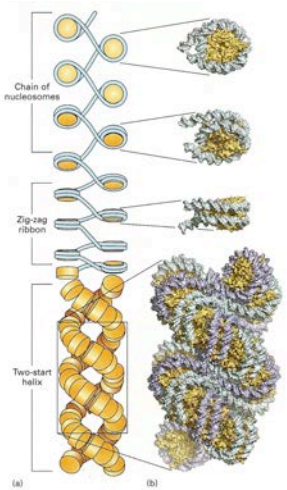
The human being as a dynamic network of networks. In systems medicine the human organism is envisioned as a system of systems or network of networks. At every scale of biological organization (molecular, cellular, organ, individual and social/ environmental) systems are portrayed as giving rise to and embedding each other. At all levels the network of networks is seen as a dynamic or four-dimensional process (as opposed to a static thing) (Copyright: The Institute for Systems Biology, used with permission)











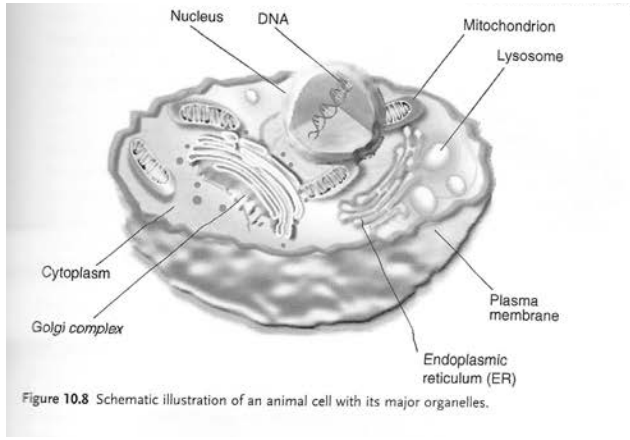
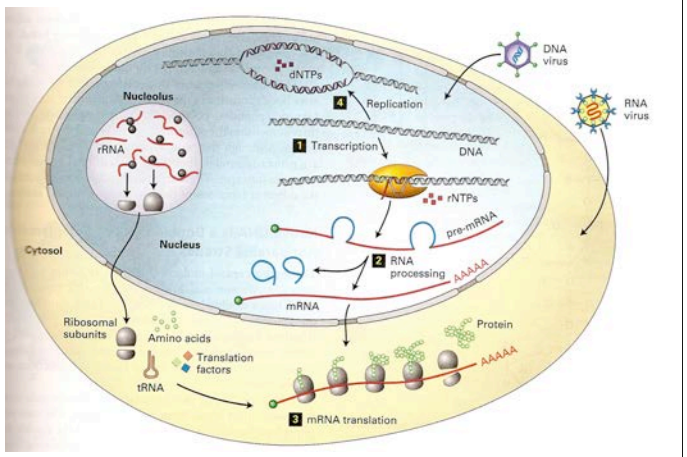
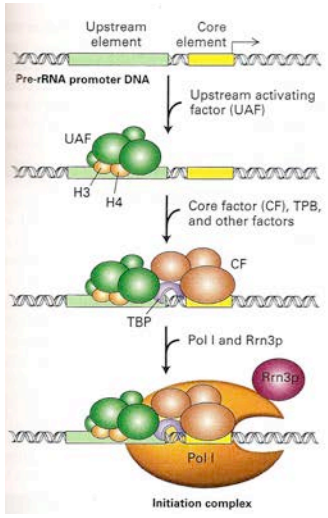
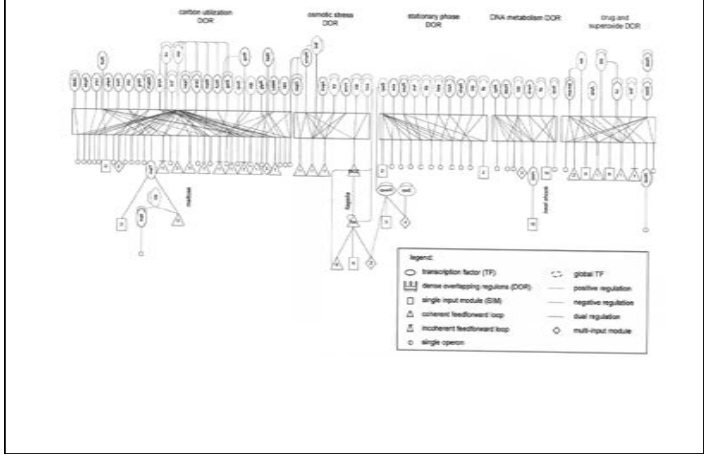
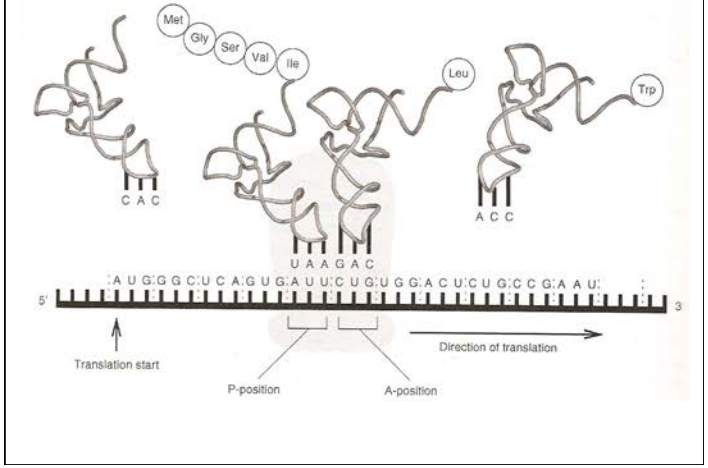


Figure 10.8 Schematic illustration of an animal cell with its major organelles.


DNA virus
RNA virus
dNTPs
4 Replication
Nucleolus
rRNA
1 Transcription
DNA
rNTPs
pre-mRNA
Cytosol
Nucleus
2 RNA processing
AAAAA
mRNA
Ribosomal subunits
Amino acids
Translation factors
tRNA
Protein
AAAAA
3 mRNA translation


Upstream element
Core element
Pre-rRNA promoter DNA
Upstream activating factor (UAF)
UAF
H3
H4
Core factor (CF), TPB, and other factors
CF
TBP
Pol I and Rrn3p
Rrn3p
Pol I
Initiation complex




Met
Gly
Ser
Val
Ile
Leu
Trp
C A C
A C C
U A A G A C
5'
3'
Translation start
Direction of translation
P-position
A-position

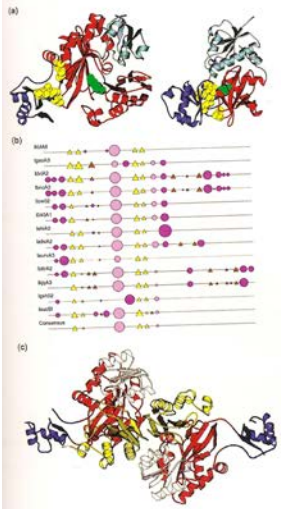
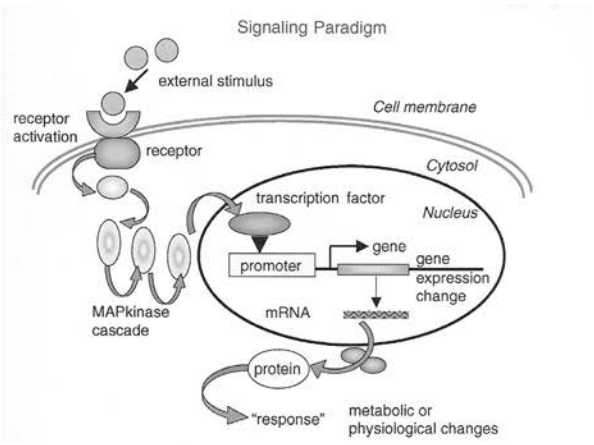
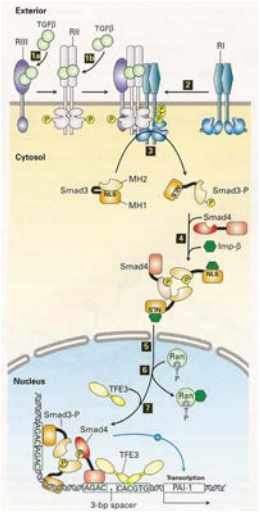
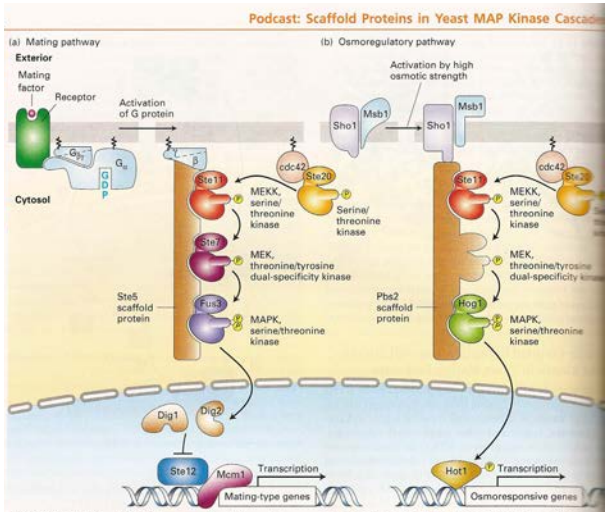


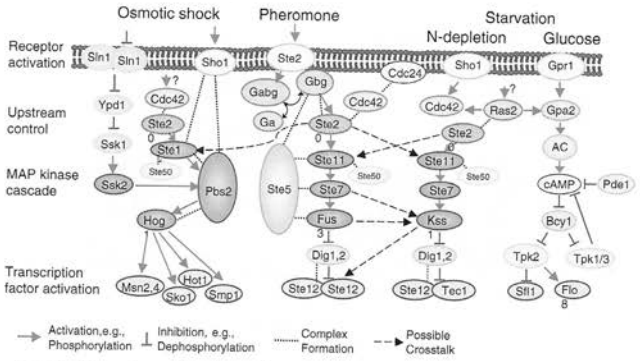
TABLE 2.2 Timescales for the Reactions in the Transcription Network of the Bacterium *E. coli* (Order of Magnitude)

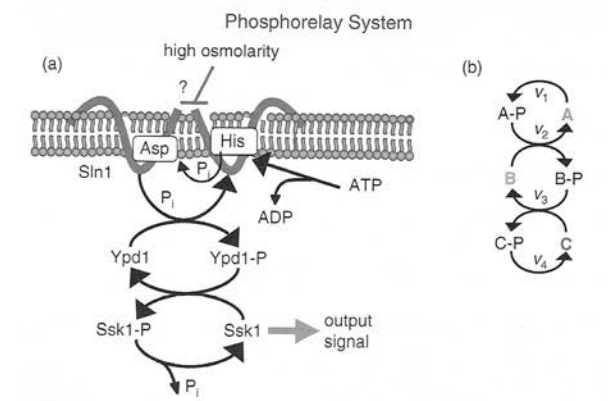
Binding of a small molecule (a signal) to a transcription factor, causing a change in transcription factor activity	~1 msec
Binding of active transcription factor to its DNA site	~1 sec
Transcription + translation of the gene	~5 min
Timescale for 50% change in concentration of the translated protein (stable proteins)	~1 h (one cell generation)











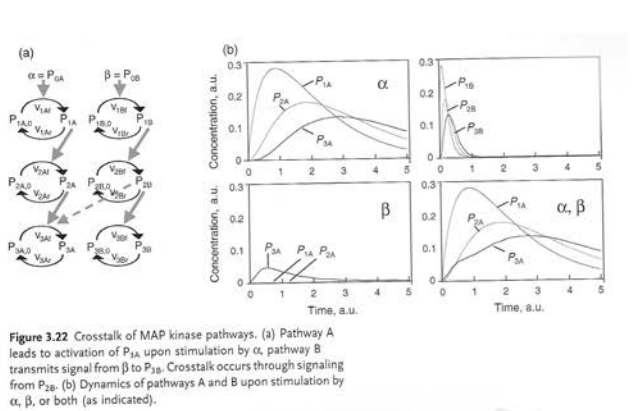


Figure 3.22 Crosstalk of MAP kinase pathways. (a) Pathway A leads to activation of P_{1A} upon stimulation by α, pathway B transmits signal from β to P_{3B}. Crosstalk occurs through signaling from P_{2B}. (b) Dynamics of pathways A and B upon stimulation by α, β, or both (as indicated).

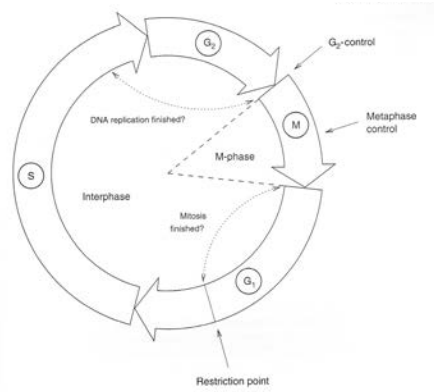


Figure 3.23 The cell cycle is divided into the interphase, which is the period between two subsequent cell divisions, and the M phase, during which one cell separates into two. Major control points of the cell cycle are indicated by arrows. More details are given in the text.

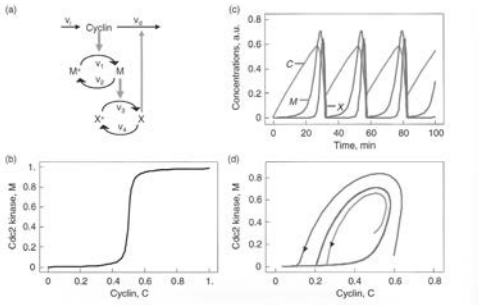
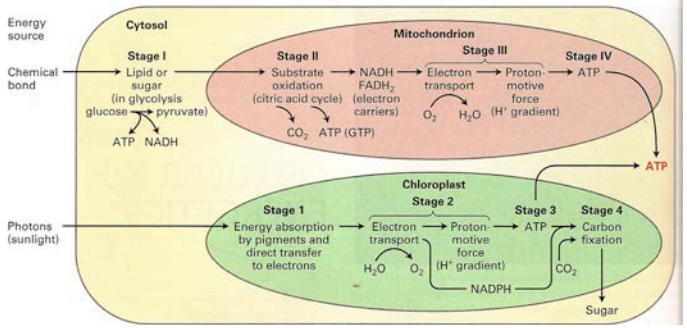
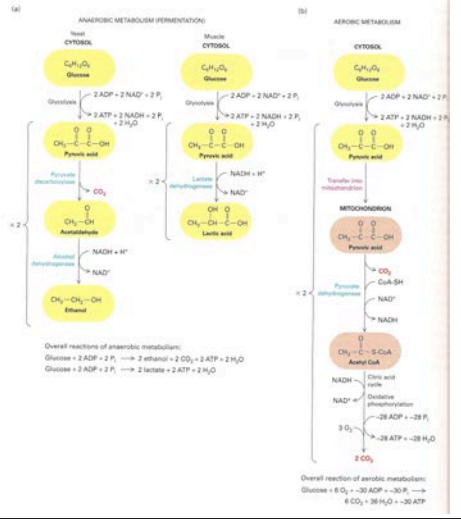
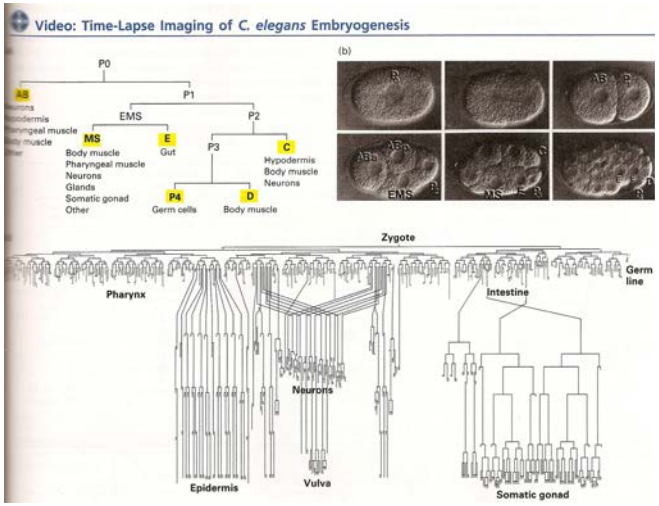
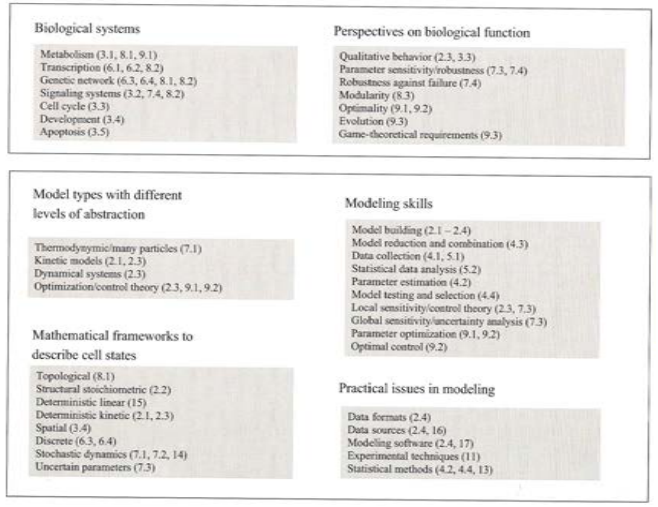


Figure 3.24 Goldbeter's minimal model of the mitotic oscillator. (a) Illustration of the model comprising cyclin production and degradation, phosphorylation and dephosphorylation of Cdc2 kinase, and phosphorylation and dephosphorylation of the cyclin protease (see text). (b) Threshold-type dependence of the fractional concentration of active Cdc2 kinase on the cyclin concentration. (c) Time courses of cyclin (*C*), active Cdc2 kinase (*M*), and active cyclin protease (*X*) exhibiting oscillations according to Equation system in Eq. (3.1). (d) Limit cycle behavior, represented for the variables *C* and *M*. Parameter values: $K_{mi} = 0.05$ ($i = 1, \ldots, 4$), $K_{m1} = 0.5$, $k_d = 0.01$, $v_i = 0.025$, $v_d = 0.25$, $V_{m1} = 3$, $V_{m3} = 1.5$, $V_{m2} = 1$, $V_{m4} = 0.5$. Initial conditions in (b) are $C(0) = M(0) = X(0) = 0.01$, and in (c) are $X(0) = 0.01$. Units: μM and min^{-1}.

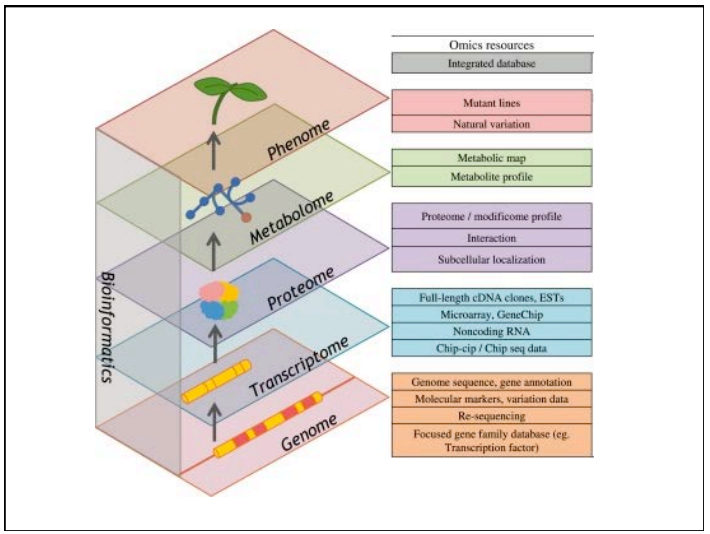


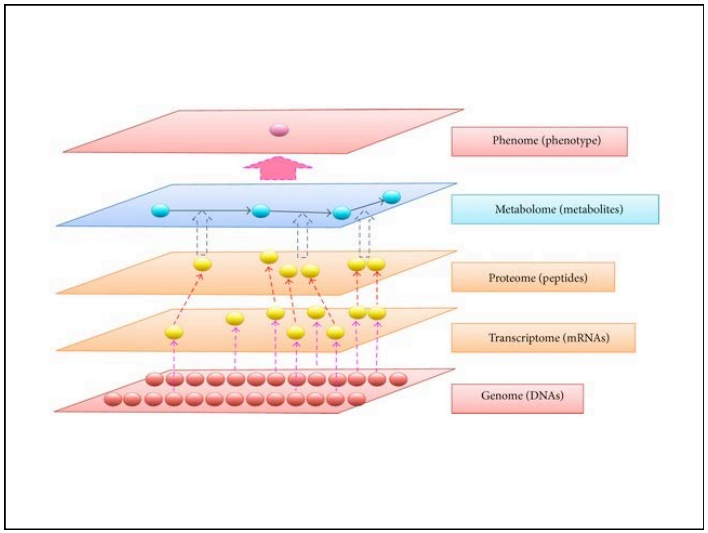


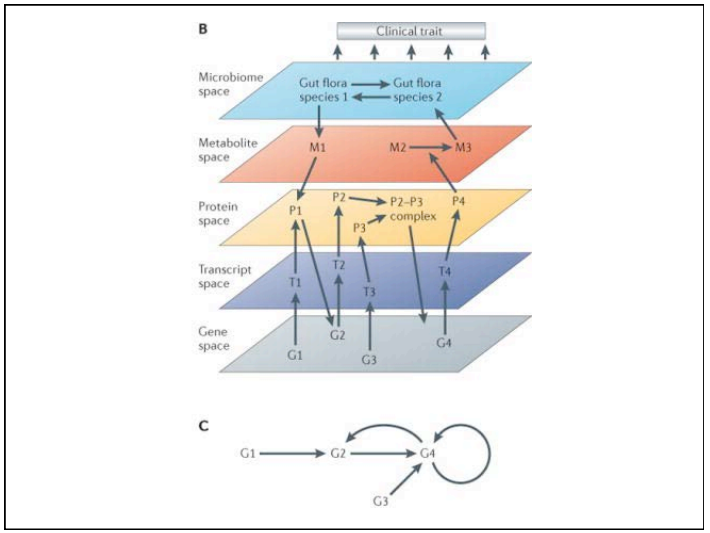


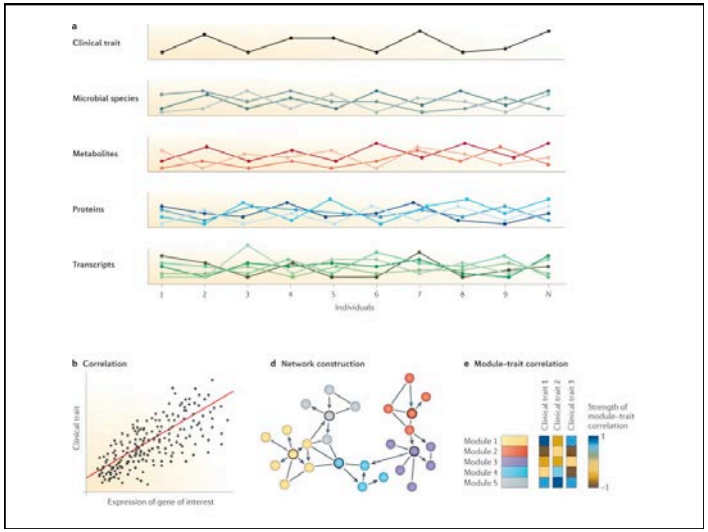


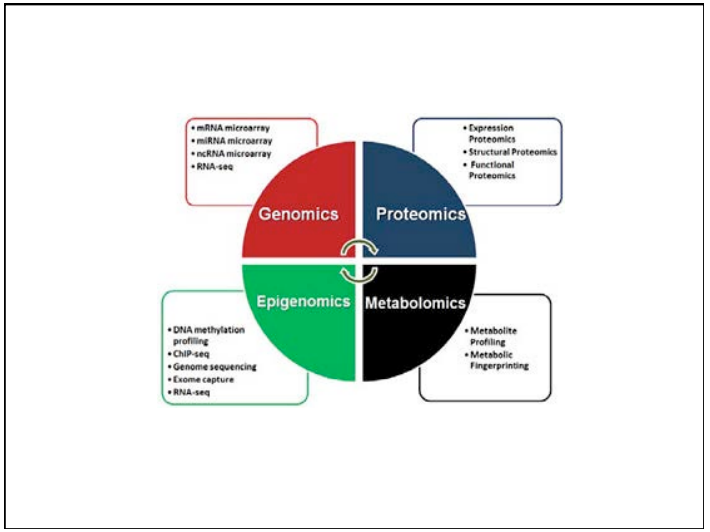
Omics Technology



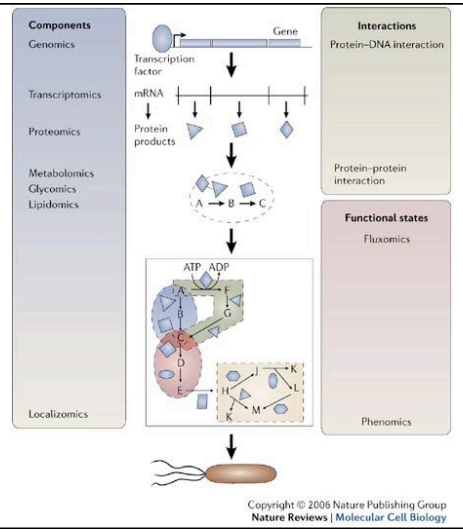

Phenome (phenotype)
Metabolome (metabolites)
Proteome (peptides)
Transcriptome (mRNAs)
Genome (DNAs)

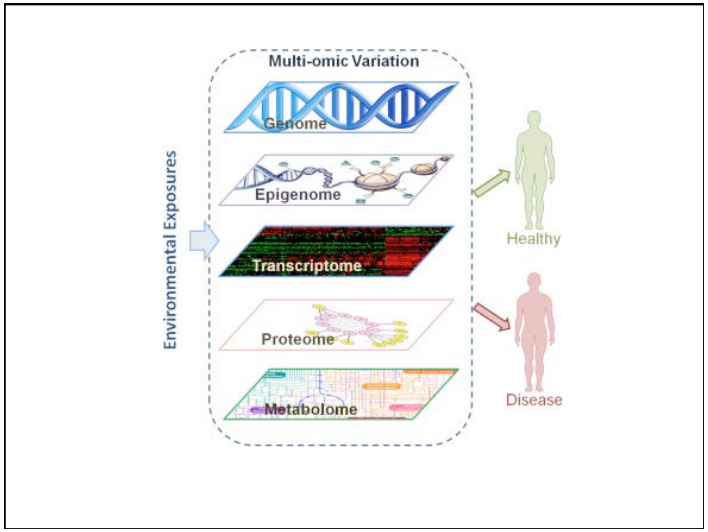

B
Clinical trait
Microbiome space
Gut flora species 1
Gut flora species 2
Metabolite space
M1
M2
M3
Protein space
P1
P2
P2–P3 complex
P3
P4
Transcript space
T1
T2
T3
T4
Gene space
G1
G2
G3
G4
C
G1
G2
G4
G3

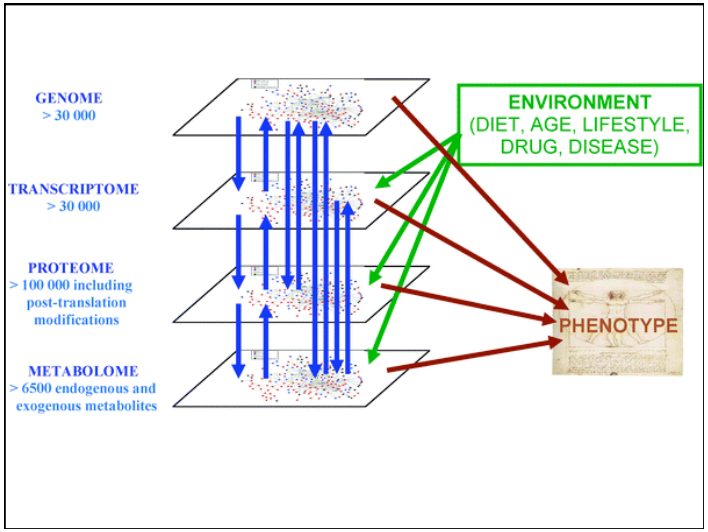

a
Clinical trait
Microbial species
Metabolites
Proteins
Transcripts
1
2
3
4
5
6
7
8
9
N
Individuals
b Correlation
Clinical trait
Expression of gene of interest
d Network construction
e Module–trait correlation
Module 1
Module 2
Module 3
Module 4
Module 5
Clinical trait 1
Clinical trait 2
Clinical trait 3
Strength of module–trait correlation
1
–1


• mRNA microarray
• miRNA microarray
• ncRNA microarray
• RNA-seq
• Expression Proteomics
• Structural Proteomics
• Functional Proteomics
Genomics
Proteomics
Epigenomics
Metabolomics
• DNA methylation profiling
• ChIP-seq
• Genome sequencing
• Exome capture
• RNA-seq
• Metabolite Profiling
• Metabolic Fingerprinting

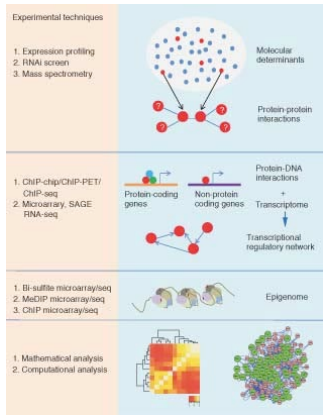
Genomics	Transcriptomics	Proteomics	Metabolomics	Protein–DNA interactions	Protein–protein interactions	Fluxomics	Phenomics
Genomics (sequence annotation)	• ORF validation • Regulatory element identification[74]	• SNP effect on protein activity or abundance	• Enzyme annotation	• Binding-site identification[75]	• Functional annotation[76]	• Functional annotation	• Functional annotation[77,100] • Biomarkers[101]
	Transcriptomics (microarray, SAGE)	• Protein: transcript correlation[78]	• Enzyme annotation[102]	• Gene-regulatory networks[74]	• Functional annotation[80] • Protein complex identification[81]		• Functional annotation[103]
		Proteomics (abundance, post-translational modification)	• Enzyme annotation[82]	• Regulatory complex identification	• Differential complex formation	• Enzyme capacity	• Functional annotation
			Metabolomics (metabolite abundance)	• Metabolic–transcriptional response		• Metabolic pathway bottlenecks	• Metabolic flexibility • Metabolic engineering[104]
				Protein–DNA interactions (ChIP–chip)	• Signalling cascades[84–86]		• Dynamic network responses[44]
					Protein–protein interactions (yeast 2H, coAP–MS)		• Pathway identification activity[87]
						Fluxomics (isotopic tracing)	• Metabolic engineering
							Phenomics (phenotype arrays, RNAi screens, synthetic lethals)







Genomics

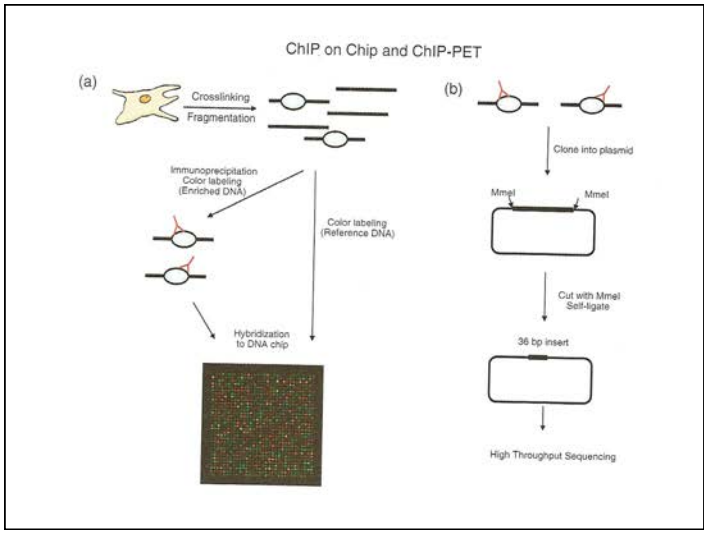


Tiling array

A high-density microarray that contains evenly spaced, or 'tiled', sets of probes that span the genome or chromosome, and can be used in many experimental applications such as transcriptome characterization, gene discovery, alternative-splicing analysis, ChIP-chip, DNA-methylation analysis, DNA-polymorphism analysis, comparative genome analysis and genome resequencing.

ChIP-chip

A high-throughput experimental technique that combines chromatin immunoprecipitation (ChIP) and microarray technology (chip) that directly identifies protein-DNA interactions.



Complete strategy for TFBSs focused ChIP-Seq and RNA-Seq data analysis.

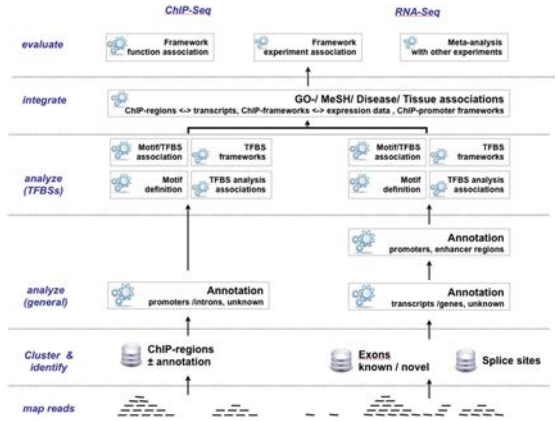


Table 2.1 Types of high-throughput sequencing technologies

Optical sequencing					
Platform	Instrument	Template preparation	Chemistry	Average length	Longest read
Illumina	HiSeq2500	BridgePCR/cluster	Rev. term., SBS	100	150
Illumina	HiSeq2000	BridgePCR/cluster	Rev. term., SBS	100	150
Illumina	MiSeq	BridgePCR/cluster	Rev. term., SBS	250	300
GnuBio	GnuBio	emPCR	Hyb-assist sequencing	1,000[a]	64,000[a]
Life Technologies	SOLiD 5500	emPCR	Seq. by Lig.	75	100
LaserGen	LaserGen	emPCR	Rev. term., SBS	25[a]	100[a]
Pacific biosciences	RS	Polymerase binding	Real-time	1,800	15,000
454	Titanium	emPCR	PyroSequencing	650	1,100
454	Junior	emPCR	PyroSequencing	400	650
Helicos	Heliscope	Adaptor ligation	Rev. term., SBS	35	57
Intelligent BioSystems	MAX-Seq	Rolony amplification	Two-step SBS (label/unlabell)	2×100	300
Intelligent BioSystems	MINI-20	Rolony amplification	Two-step SBS (label/unlabell)	2×100	300
ZS Genetics	N/A	Atomic labeing	Electron microscope	N/A	N/A
Halcyon Molecular	N/A	N/A	Direct observation of DNA	N/A	N/A
Electical sequencing					
Platform	Instrument	Template preparation	Chemistry	Average length	Longest read
IBM DNA transistor	N/A	None	Microchip nanopore	N/A	N/A
NABsys	N/A	None	Nanochannel	N/A	N/A
Bionanogenomics	N/A	Anneal 7mers	Nanochannel	N/A	N/A
Life Technologies	PGM	emPCR	Semi-conductor	150	300
Life Technologies	Proton	emPCR	Semi-conductor	120	240
Life Technologies	Proton 2	emPCR	Semi-conductor	400[a]	800[a]
Genia	N/A	None	Protein nanopore (a-hemalysin)	N/A	N/A
Oxford nanopore	MinION	None	Protein nanopore	10,000	10,000[a]
Oxford nanopore	GridION 2K	None	Protein nanopore	10,000	500,000[a]
Oxford nanopore	GridION 8K	None	Protein nanopore	10,000	500,000[a]

[a]Values are estimates from companies that have not yet released actual data

Table 1. A comparison of representative third-generation DNA sequencing companies

	Helicos	Pacific Biosciences	Oxford Nanopore	Complete Genomics	Ion Torrent
Key technology	Amplification-free sequencing	Zero-mode waveguide nanostructure arrays	Protein nanopores	Self-assembling DNA nanoarrays	Chemical sensitive field-effect transistor arrays
Single-molecule detection	Yes	Yes	Yes	No	Undisclosed
Commercialization	Launched in 2008	Launched in March 2010	Undisclosed estimated launch	Sequencing services	Launched in March 2010
Funding to date (millions USD)[a]	$115	$266	$64	$45	$23
Refs	[43]	[44]	[46]	[47]	[66]

[a]Company website data.

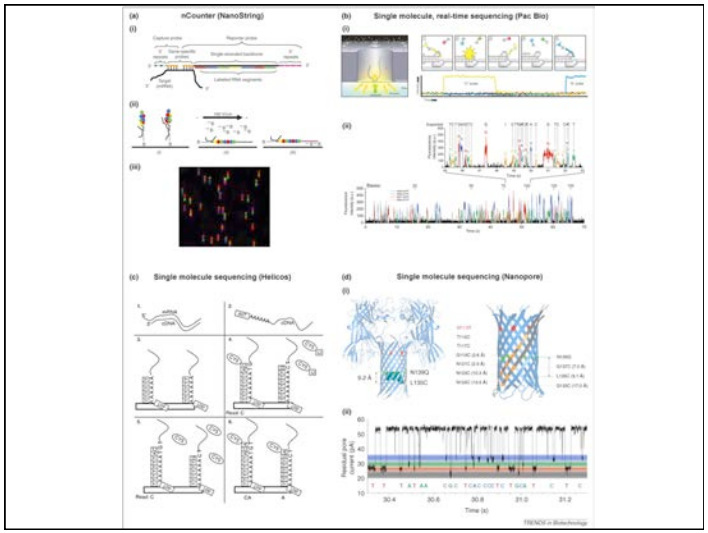
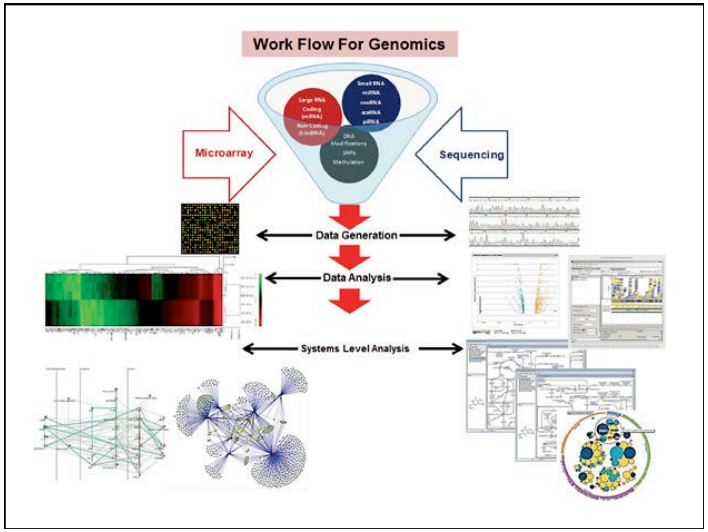


Table 2
Cancer driver mutations discovered by large-scale next generation sequencing.

Gene	Aberration type	Tumor type	Biological function	Tumor effect	Sequencing method	Number of samples	Sample type	Reference
EBF1-PDGFRB, BCR-JAK2, NUP214-ABL1	Fusion	ALL	Kinase signaling	Activating	Whole-genome	15	Acute lymphoblastic leukemia	22
IL7R, SH2B3	Mutation	ALL	Cytokine signaling	Activating	Whole-genome	15	Acute lymphoblastic leukemia	22
TP53	Mutation	Cell Carcinoma	Cell cycle regulation	Inactivating	Whole-genome	457	Peripheral blood	23
VTI1A-TCF7L2	Fusion	Colon	Transcription factor	Activating	Whole-genome	9	Colorectal adenocarcinomas	24
ARID1A, ARID1B, ARID2, MLL, MLL3	Mutation	Liver	Chromatin regulation	Inactivating	Whole-genome	27	Hepatocellular carcinoma	25
PREX2	Mutation	Melanoma	Rac exchange factor	Inactivating	Whole-genome	25	Melanomas	26
ATRX	Mutation	Neuroblastoma	Telomere maintenance	Inactivating	Whole-genome	40	Neuroblastomas	27
BRIP1	Mutation	Ovary	DNA repair	Inactivating	Whole-genome	457	Peripheral blood	28
DNMT3A	Mutation	AML	DNA methylation	Inactivating	Exome	112	Acute monocytic leukemias	54
CBFB	Mutation	Breast	Transcription factor	Inactivating	Exome	103	Breast cancers	55
MAGI3-AKT3	Fusion	Breast	Cell signaling	Activating	Exome	103	Breast cancers	55
NOTCH1	Mutation	Cell carcinoma	Cell signaling	Inactivating	Exome	32	Head and neck squamous cell carcinomas	56
SF3B1	Mutation	CML	mRNA splicing	Inactivating	Exome	105	Chronic lymphocytic leukemias	57
MXRA5	Mutation	Lung	Matrix remodeling	Activating	Exome	14	Non-small cell lung carcinomas	58
CSMD3	Mutation	Lung	Unknown	Inactivating	Exome	31	Non-small cell lung carcinomas	59
RAC1	Mutation	Melanoma	Cell signaling	Activating	Exome	147	Melanomas	60
GRIN2A	Mutation	Melanoma	Glutamate receptor	Unknown	Exome	14	Melanomas	61
SPOP, FOXA1, MED 12	Mutation	Prostate	Transcription regulation	Unknown	Exome	112	Prostate tumors	62
FAT4	Mutation	Stomach	Cell adhesion	Inactivating	Exome	15	Gastric adenocarcinomas	63
ARID1A	Mutation	Stomach	Chromatin remodeling	Inactivating	Exome	15	Gastric adenocarcinomas	63



Transcriptome

Table 1

Comparison of Approaches to Identify Transcriptional Components

Identifying TFs Through Expression Profiling		
Method	**Pros**	**Cons**
Gene expression arrays	Technology and analysis tools very widespread; nonTF target genes can be assessed in concert with TF genes	Small dynamic range for quantitative analysis of gene expression; low sensitivity; limited to known transcripts
qPCR (Quanttrx)	Real-Time PCR technology available to most molecular labs; data easy to analyze; very sensitive; quantitative over wide range of expression; highly reproducible; direct analysis of TFs	Limited to pre-defined list of TFs and isoforms
Sequencing (RNA-seq)	Very Sensitive; Quantitative over wide dynamic range; Identifies all known and unknown transcripts; nonTF target genes can be assessed in concert with TF genes	Technology and analysis methods not yet widespread; large datasets require bioinformatics expertise
Nanostring	Very sensitive; quantitative over wide range of expression; high-throughput; direct analysis of TFs	Technology not currently widespread; requires up front investment in specialized equipment; limited to pre-defined list of TFs and isoforms
DNase-Seq	Increased sequencing depth can reveal 'footprinted' motifs; requires less starting material than ChIP-Seq	Does not distinguish between enhancers, promoters, or other regulatory elements; optimization can be troublesome; cost of deep sequencing is high (but declining); does not immediately reveal relevant TF involved; requirement for significant downstream analysis at genome-wide level
FAIRE-Seq	Technically simple; requires less starting material than ChIP-Seq	Does not distinguish between enhancers, promoters, or other regulatory elements; cost of deep sequencing is high (but declining); does not immediately reveal relevant TF involved; requirement for significant downstream analysis at genome-wide level
ChIP-seq of modified histone marks	Can identify enhancers specifically, and can distinguish between poised and active enhancers.	Can require significant amounts of starting material (>10^6 cells per epitope); need for high quality antibody; cost of deep sequencing is high (but declining); does not immediately reveal relevant TF involved; requirement for significant downstream analysis at genome-wide level

5.1
High-Throughput Experiments

Summary

The analysis of transcriptome data has become increasingly popular over the last decades due to the advent of new high-throughput technologies in genome research. Often, these data build the basis for defining the essential molecular read-outs for a particular disease, developmental state, or drug response being subject to computational modeling. In particular, DNA arrays have become the most prominent experimental technique to analyze gene expression data. A DNA array consists of a solid support that carries DNA sequences representing genes – the probes. In hybridization experiments with the target sample of labeled mRNAs and through subsequent data capture a numerical value, the signal intensity, is assigned to each probe. It is assumed that this signal intensity is proportional to the amount of molecules of the respective gene in the target sample. Changes in signal intensities are interpreted as concentration changes. Several experimental platforms are available that enable the genome-wide analysis of gene expression. Another recently emerging high-throughput technology is next generation sequencing. These new sequencing techniques provide in many cases flexible alternatives to DNA array techniques in identifying the abundance of specific sequences, providing information on transcript abundance or RNA processing events.

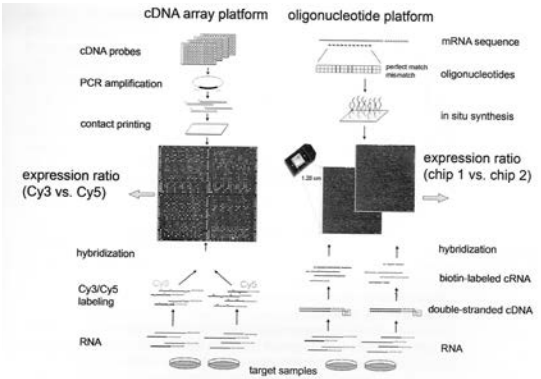
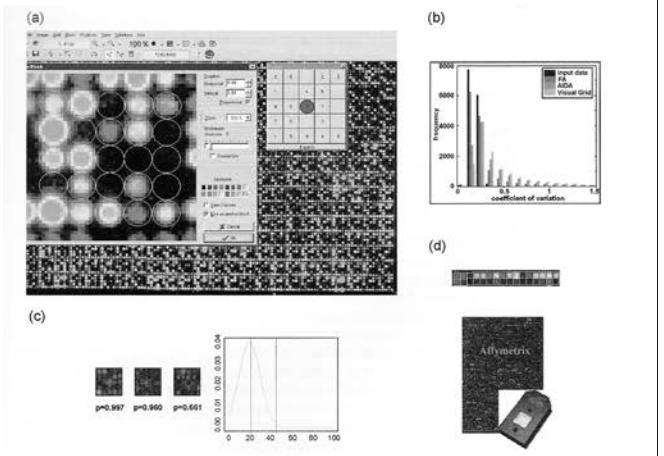


Figure 5.1 Scheme of DNA array analysis using cDNAs (left) and oligonucleotides (right, here Affymetrix technology) as probes. Probe design and construction of the array is shown in the upper part of the figure and the different preparation steps of the target samples are shown in the lower part. In the center of the figure, the strategy for case-control studies is shown using a two-color approach (left panel) and a single-color approach (right panel). Array images were taken from [69] and from http://www.affymetrix.com.



5.2
Analysis of Gene Expression Data

Summary

The analysis of genome-wide gene expression data involves basic concepts from multivariate statistics. Most applications belong to two groups: the first group consists of case-control studies comparing a certain transcriptome state of the biological system (e.g., disease state, perturbed state) to the control situation; the second group of applications consist of multiple case studies involving different states (e.g., drug response time series, groups of patients, etc.). The analysis of case-control studies involves testing of statistical hypotheses. Here, expression changes are observed that deviate from a predefined hypothesis and this deviation is judged for significance. The basic methods for multicase studies are clustering and classification. Here, groups of coexpressed genes serve to identify functionally related groups of genes or experiments. These types of analysis result in the identification of marker genes and their related interactions, which are the basis for further network studies.

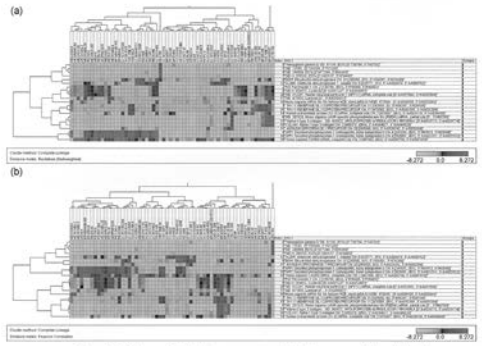


Figure 5.5 Influence of similarity measure on clustering. Two dendrograms of a subgroup of genes using the microarray expression data of Ross *et al.* [70] were generated using hierarchical clustering with Euclidean distance (a) and Pearson correlation (b) as pairwise similarity measure. Although all other parameters are kept constant, results show differences in both gene and cancer cell line groupings. Clustering was performed with the *J-Express Pro* software package (Molmine, Bergen Norway).

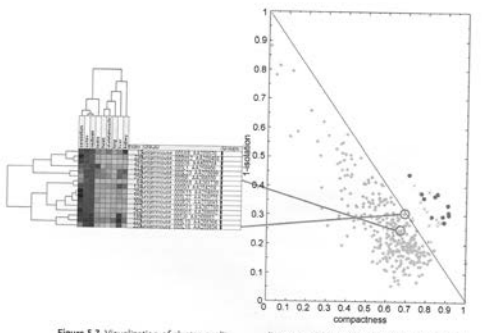
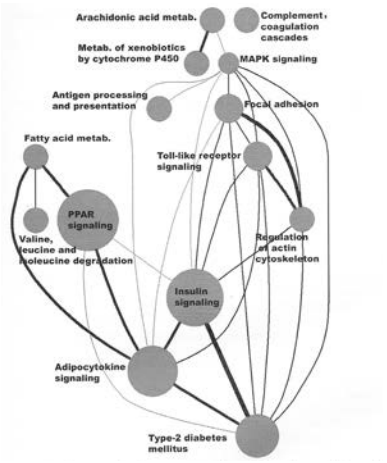
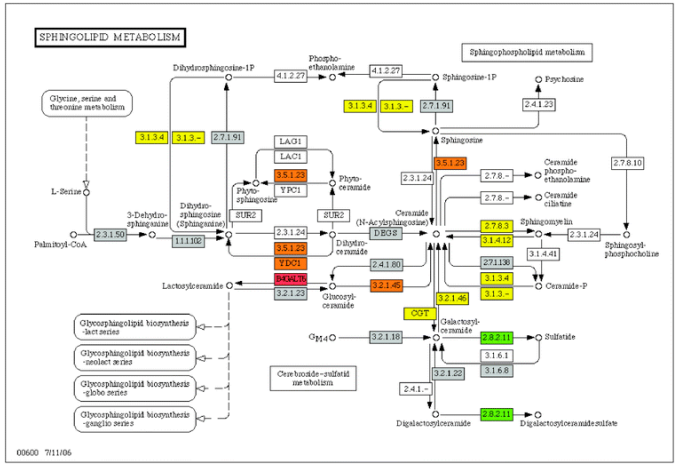


Figure 5.7 Visualization of cluster quality. Left: cluster of cDNA sequences that have a tissue-specific expression in brain tissue. In this study nine different tissues have been compared in the mouse using a whole-genome approach. Genes in that specific cluster show a high expression in three brain regions (cortex, cerebellum, and midbrain). Right: Compactness (X-axis) and isolation (Y-axis) can be used to visualize the cluster quality in a two-dimensional plot. Crosses represent the brain-specific cDNAs and circles represent another cluster of liver-specific sequences. Green diamonds represent random assignments of compactness and isolation. The visualization allows the identification of false positives in each cluster which can enhance following analyses, for example with respect of promoter searches.





6.3
Dynamic Models of Gene Regulation

Summary

In order to comprehend the functioning of organisms at the molecular level, we wish to know which genes are expressed, to what level, where, and when. A network of interactions between DNA, mRNA, proteins, and other molecules realizes the regulation of gene expression. This network comprises many components. According to the central dogma of molecular biology formulated by Francis Crick [71], there is a forward flow of information from gene to mRNA to protein. Moreover, positive and negative feedback loops and information exchange with signaling pathways and energy metabolism ensure the appropriate regulation of expression according to the current state of the cell and the environment.

Modeling of gene expression is used here as an example to apply different modeling techniques. The dynamics and the regulatory patterns of gene expression will be mathematically described with various graphs, Boolean networks, Bayesian networks, ordinary and partial differential equation systems, stochastic processes, and with rule-based formalisms.

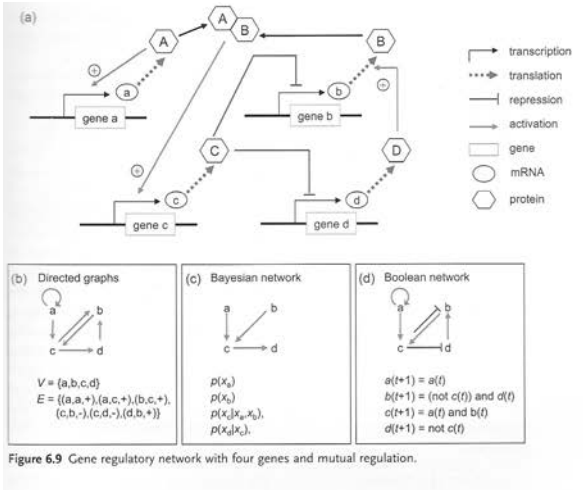
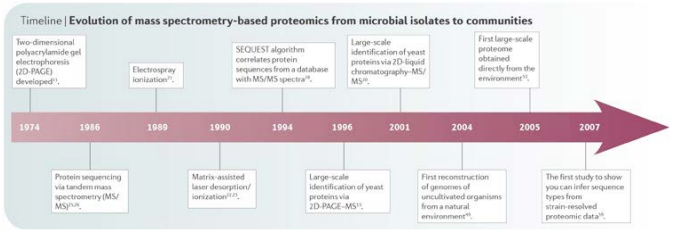


Figure 6.9 Gene regulatory network with four genes and mutual regulation.

Proteome

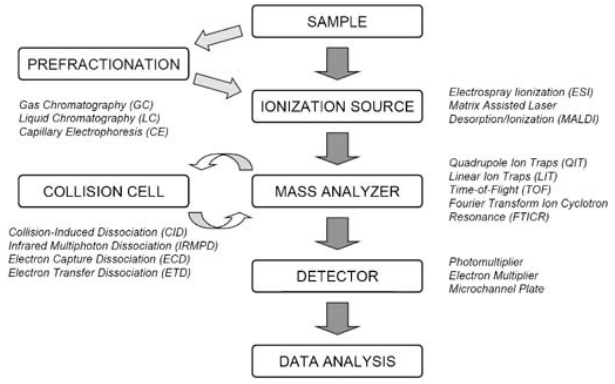
Mass spectrometry

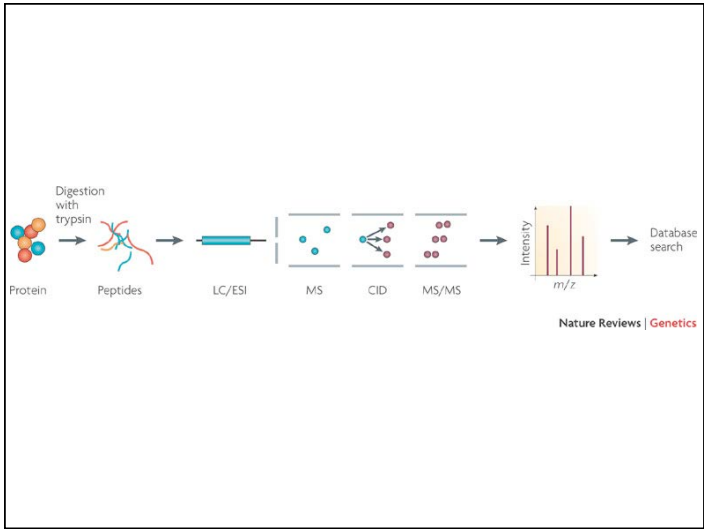
An analysis technique that identifies biochemical molecules (such as proteins, metabolites or fatty acids) on the basis of their mass and charge.



Tandem mass spectrometry

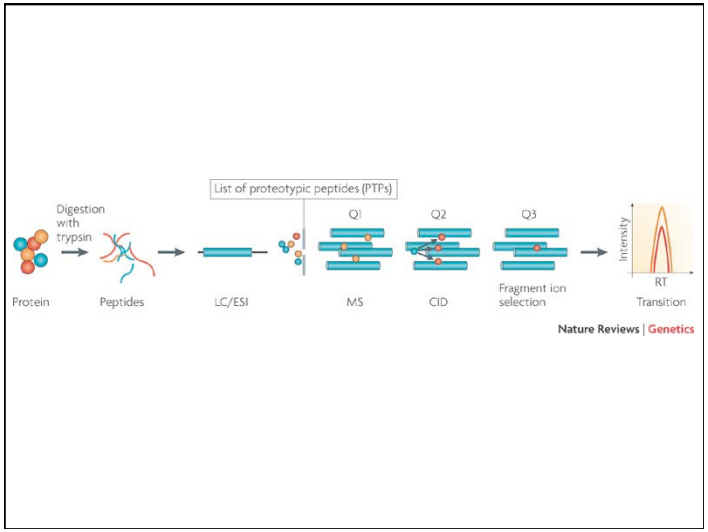
This combines two mass spectrometers: one (MS1) for the detection and selection of precursor ions, which is followed by a second (MS2) for the analysis of fragment ion spectra generated from selected precursor ions after collision-induced fragmentation. The information from the fragment ion spectra is used for peptide identification.

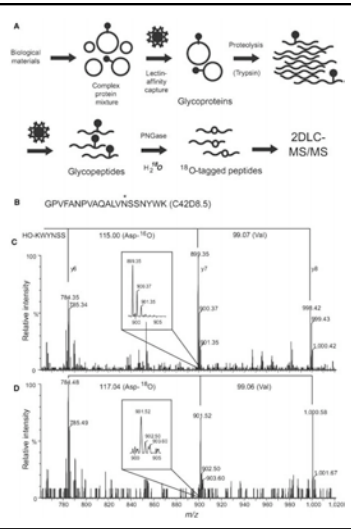


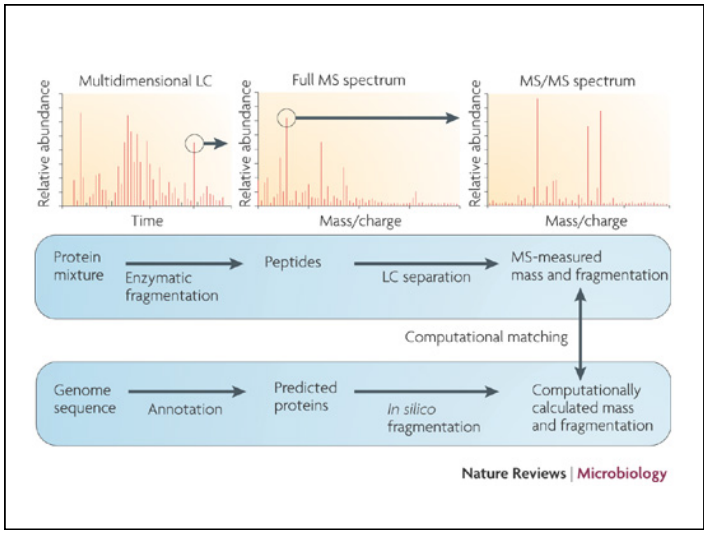


Liquid chromatography–tandem mass spectrometry

Liquid chromatography is used in MS-based proteomics to separate peptides in complex mixtures primarily on the basis of their charge or hydrophobicity using strong cation exchange or reversed-phase chromatography columns.







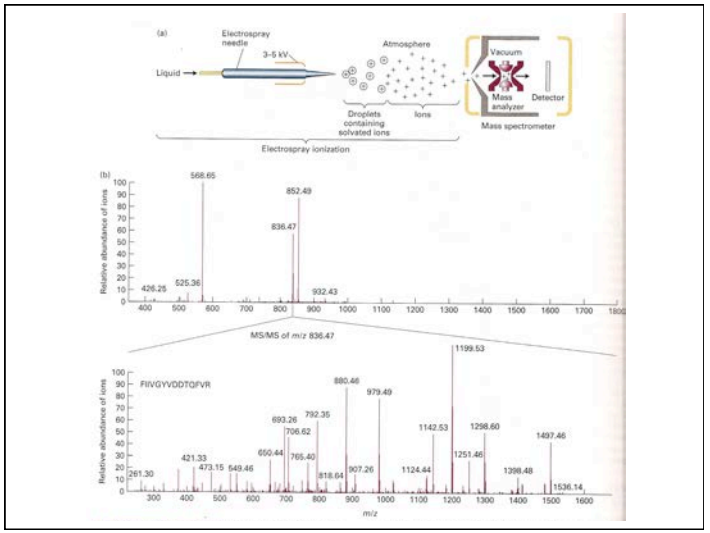
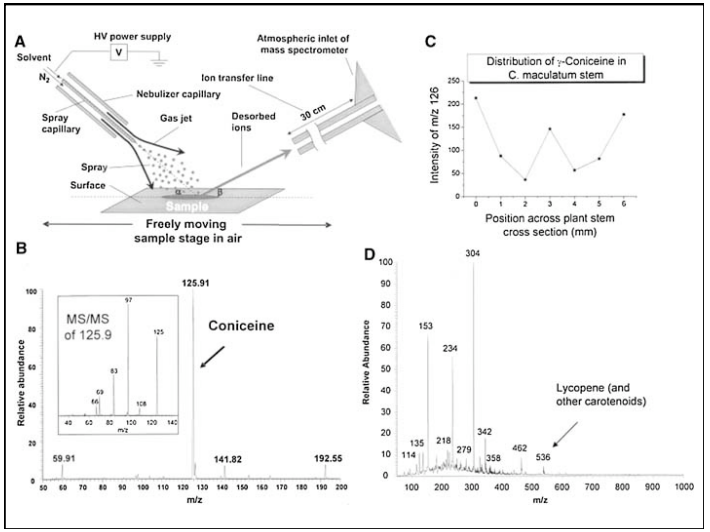
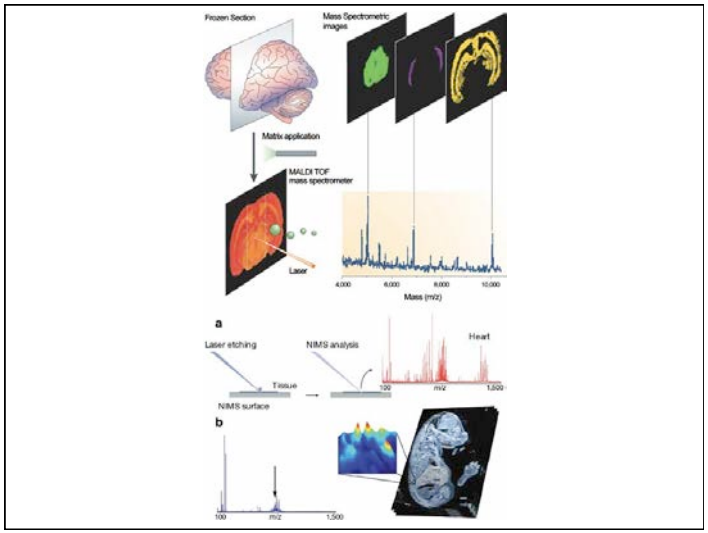
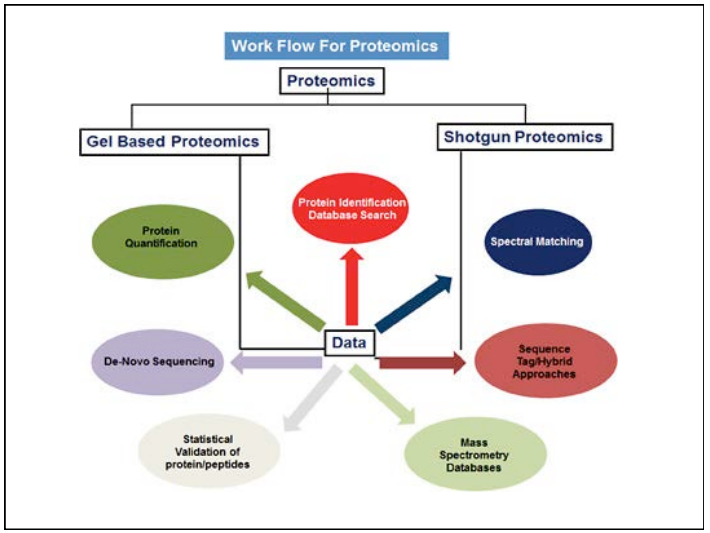


TABLE 1. Potential Cancer Biomarkers Identified by Mass Spectrometry-Based 'Omics' Technologies

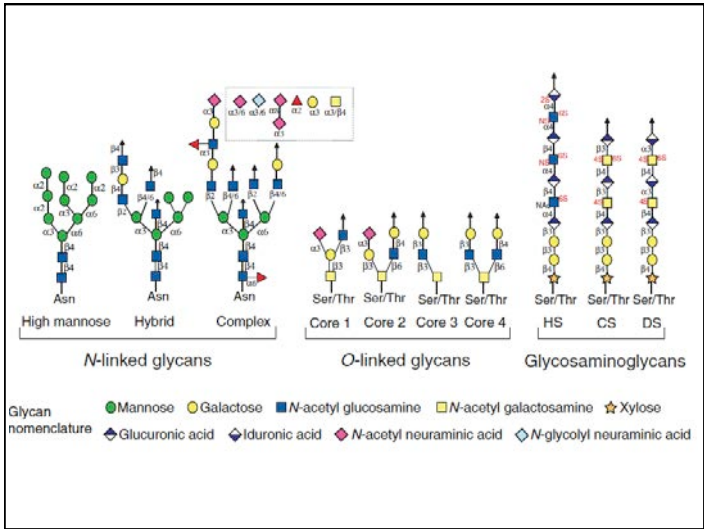
Biomarkers	"omics" platforms	MS methods	Sample source	Cancer type	References
Apolipoprotein A1, Inter-α-trypsin inhibitor Haptoglobin-a-subunit Transthyretin	Proteomics	SELDI-TOF	Serum	Ovarian	Ye et al., 2003; Zhang et al., 2004
Vitamin D-binding protein	Proteomics	SELDI-TOF	Serum	Prostate	Hlavaty et al., 2003
Stathmin (Op18), GRP 78 14-3-3 isoforms, Transthyretin Protein disulfide Isomerase	Proteomics	ESI-MS	Tissue	Lung	Chen et al., 2003
Peroxiredoxin, Enolase Protein disulfide Isomerase HSP 70, α -1-antitrypsin	Proteomics	MALDI-TOF, LC-MS	Tissue	Breast	Somiari et al., 2003
HSP 27	Proteomics	MALDI-TOF	Serum	Liver	Feng et al., 2005
Annexin I, Cofilin, GST Superoxide dismutase Peroxiredoxin, Enolase Protein disulfide Isomerase	Proteomics	MALDI-TOF, ESI-MS, Q-TOF	Tissue	Colon	Seike et al., 2003; Stierum et al., 2003
Neutrophil peptides 1-3	Proteomics	SELDI-TOF	Nipple aspirate fluid	Breast	Li et al., 2005b
PCa-24	Proteomics	MALDI-TOF	Tissue	Prostate	Zheng et al., 2003
Alkanes, Benzenes	Metabonomics	GC-MS	Breath	Lung	Phillips et al., 1999
Decanes, Heptanes	Metabonomics	GC-MS	Breath	Breast	Phillips et al., 2003
Hexanal, Heptanal	Metabonomics	LC-MS	Serum	Lung	Deng et al., 2004
Pseu, m1A, m11	Metabonomics	HPLC, LC-MS	Urine	Liver	Yang et al., 2005b




Frozen Section
Mass Spectrometric images
Matrix application
MALDI TOF mass spectrometer
Laser
Mass (m/z)
a
Laser etching
NIMS analysis
Heart
Tissue
NIMS surface
b


Work Flow For Proteomics
Proteomics
Gel Based Proteomics
Shotgun Proteomics
Protein Identification Database Search
Protein Quantification
Spectral Matching
Data
De-Novo Sequencing
Sequence Tag/Hybrid Approaches
Statistical Validation of protein/peptides
Mass Spectrometry Databases

Glycomics


Asn
High mannose
Hybrid
Complex
N-linked glycans
Ser/Thr
Core 1
Core 2
Core 3
Core 4
O-linked glycans
HS
CS
DS
Glycosaminoglycans
Glycan nomenclature
Mannose
Galactose
N-acetyl glucosamine
N-acetyl galactosamine
Xylose
Glucuronic acid
Iduronic acid
N-acetyl neuraminic acid
N-glycolyl neuraminic acid

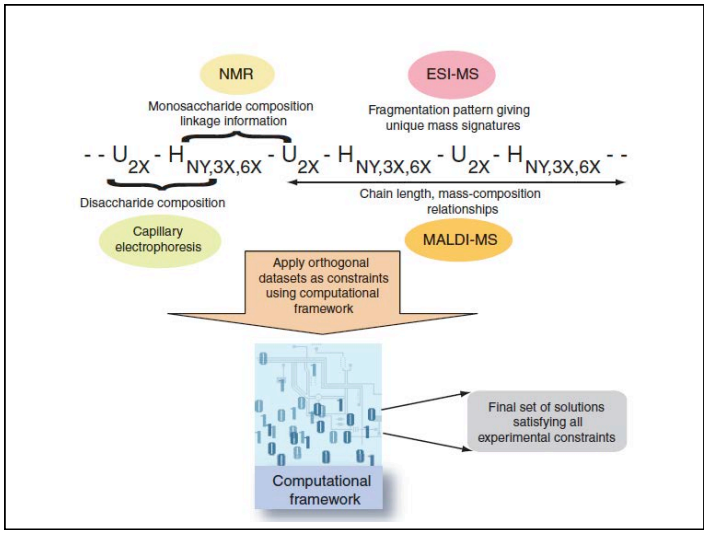
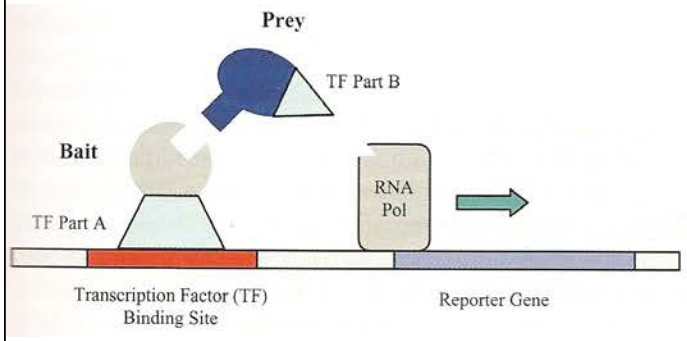
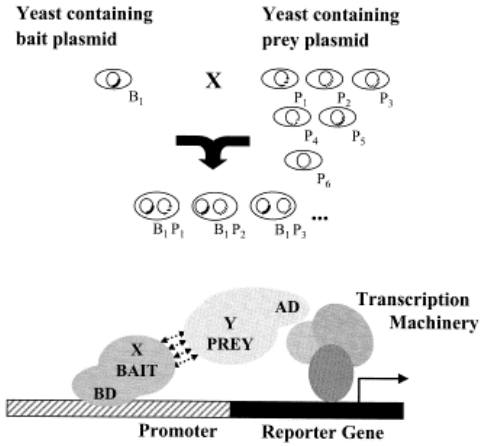


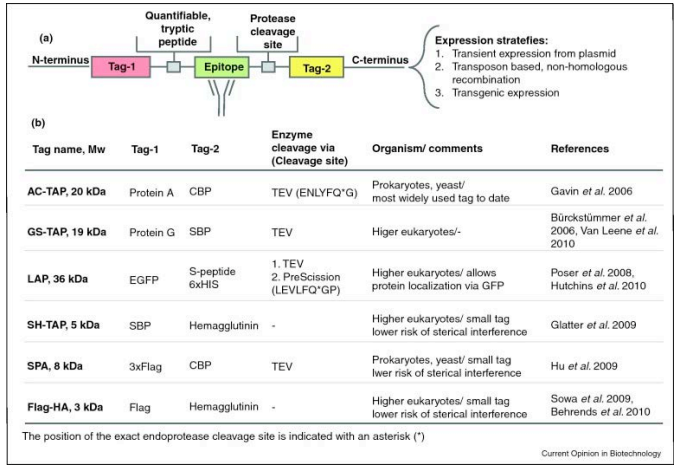
Table 2 | Web-based resources for glycomics

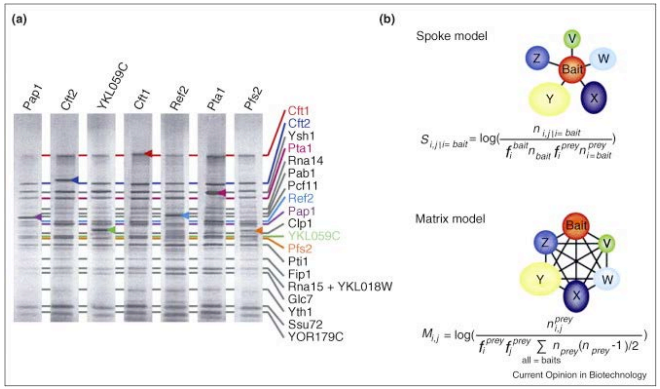
Web resource (URL)	Key datasets or information
Consortium for Functional Glycomics (CFG; USA)	
GBP Molecule Page (http://www.functionalglycomics.org/glycomics/molecule/jsp/gbpMolecule-home.jsp)	Information portal with access to CFG and public databases
Glycan Database (http://www.functionalglycomics.org/glycomics/molecule/jsp/carbohydrate/carbMoleculeHome.jsp)	Database of glycan structures with search interfaces and links to CFG glycan array and MALDI-MS data
Glycan Profiling Data (http://www.functionalglycomics.org/glycomics/publicdata/glycoprofiling.jsp)	Raw and annotated MALDI-MS profiles of glycans from mouse and human cells and tissues
Glycan Array Screen Data (http://www.functionalglycomics.org/glycomics/publicdata/primaryscreen.jsp)	Raw data, bar graph of mean binding signal of GBP to each glycan in the array with links to their structures in glycan database
Gene Microarray Data (http://www.functionalglycomics.org/glycomics/publicdata/microarray.jsp)	Gene expression profiles of glycan biosynthesis enzymes and GBPs in various cells and tissues supplied by investigators
Transgenic Mice Phenotyping Data (http://www.functionalglycomics.org/lycomics/publicdata/phenotyping.jsp)	Experimental protocols, data files corresponding to various phenotyping analysis of transgenic mice
Kyoto Encyclopedia of Genes and Genomes (KEGG; Japan)	
KEGG Glycan Database (http://glycan.genome.jp)	Database of glycan structures obtained from CarbBank and updated with structures from other labs
KEGG Pathways dDatabase (http://www.genome.jp/kegg/pathway.html)	Collection of 15 glycan biosynthesis pathways with links to around 100 glycan biosynthesis enzymes
Glycomics Initiative of the German Cancer Research Institute (Glycosciences.de; Germany)	
Glycan Database (http://www.glycosciences.de/sweetdb/structure/)	Database of glycan structures
Glycans in Protein Data Bank (PDB) (http://www.glycosciences.de/sweetdb/start.php?action=form_pdb_data)	Glycan structures extracted from PDB entries using computational tools
Glycan NMR Profiles (http://www.glycosciences.de/sweetdb/nmr/)	Characteristic chemical shifts monosaccharides in different glycans
Computational Tools for Glycans (http://www.glycosciences.de/tools/index.php)	Collection of tools to analyze and query glycan structures and predict glycosylation sites on glycoproteins
Three-dimensional Modeling of Glycans (http://www.glycosciences.de/modeling/index.php)	Collection of tools to investigate conformational aspects of glycans and model their 3D structures
Other glycomics resources	
Glycosuite Database (http://www.glycosuitedb.org)	Commercial database and tools for glycans
Sugabase (http://www.boc.chem.uu.nl/sugabase/sugabase.html)	Glycan NMR database: chemical shifts of glycan structures
Lectin Database (http://www.imperial.ac.uk/research/animallectins/)	Collection of information on animal lectins
Three-dimensional Lectin Database (http://www.cermav.cnrs.fr/lectines/)	Three-dimensional structures of lectins in the PDB
Bacterial Glycan Database (http://www.glyco.ac.ru/bcsdb/)	Database of bacterial glycan structures
CAZy (http://afmb.cnrs-mrs.fr/CAZY/)	Carbohydrate active enzymes database

Protein Interactome









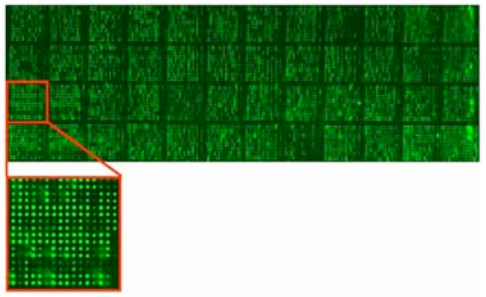
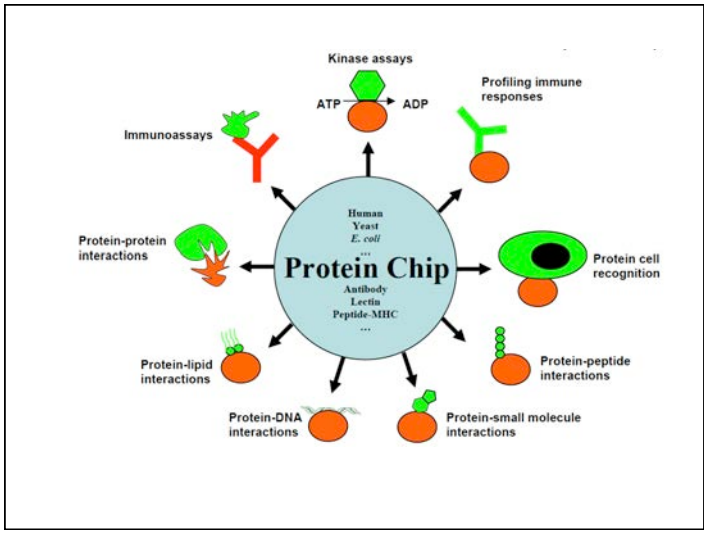
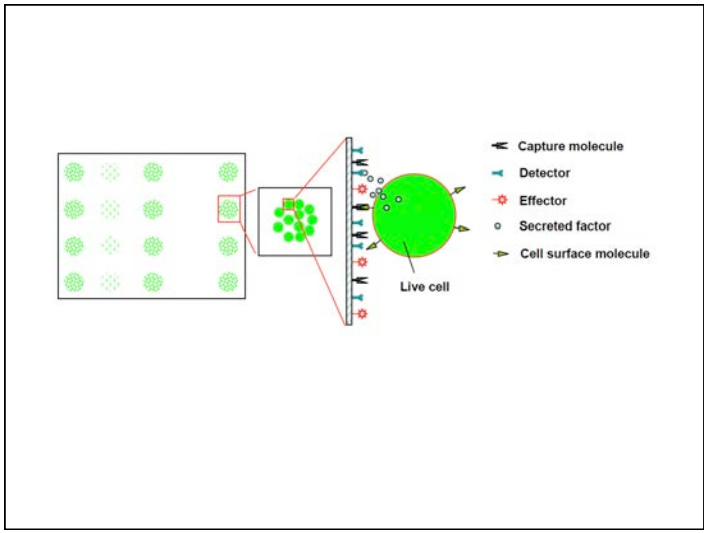


Fig. (1). Image of a typical protein microarray. A yeast protein microarray has been probed with an anti-GST antibody, followed by a Cy3-conjugated secondary antibody. An enlarged image of one of the 48 blocks is shown below the protein chip.





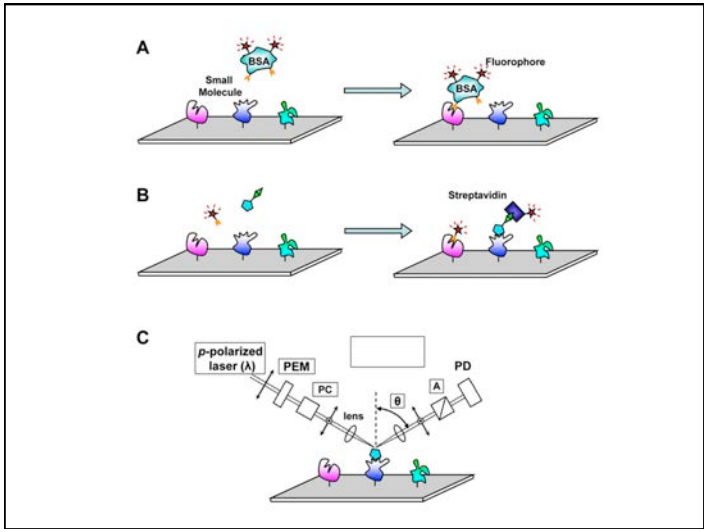
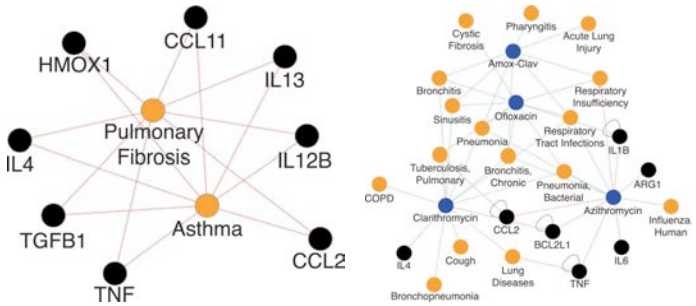


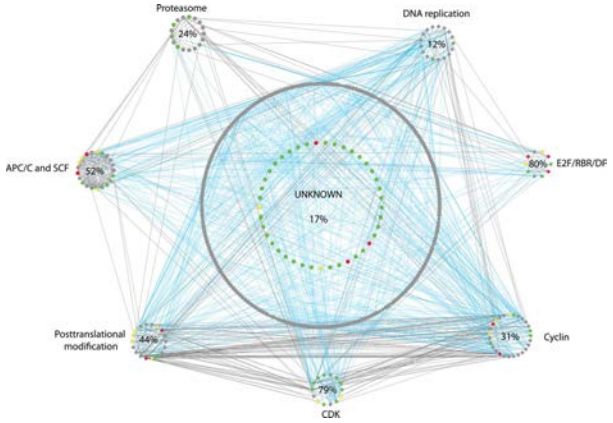
Table 1
Selected pathway/network analysis resource that can benefit Proteomics data analysis.

Name	Description	Link	Reference	Functional info using	Topological info using
GoMiner	Gene ontology (GO) analysis for omic data	http://discover.nci.nih.gov/gominer/	Zeeberg et al. (2003)	Single molecule	Non
KEGG	Kyoto encyclopedia of genes and genomes	http://www.genome.jp/kegg/	Kanehisa and Goto (2000)	Molecular pathway	Non
DAVID	The database for annotation, visualization and integrated discovery	http://david.abcc.ncifcrf.gov/	Dennis et al. (2003)	Molecular pathway	Small-scale
PID	Pathway interaction database	http://pid.nci.nih.gov/	Schaefer et al. (2009)	Cellular pathway	Non
HPD	Human pathway database	http://bio.informatics.iupui.edu/HPD	Chowbina et al. (2009)	Cellular pathway	Small-scale
GESA	Gene set enrichment analysis	http://www.broadinstitute.org/gsea/	Subramanian et al. (2005)	Cellular pathway	Small-scale
IPA	Ingenuity pathway analysis	http://www.ingenuity.com/	N/A	Molecular pathway	Small-scale
MetaCore	Thomson Reuters pathway analysis and knowledge mining	http://thomsonreuters.com/metacore/	N/A	Cellular pathway	Small-scale
Pathway-Express	A systems biology approach for pathway level impact analysis	http://vortex.cs.wayne.edu/projects.htm	Draghici et al. (2007)	Molecular pathway	Mid-Scale
SPIA	Signaling pathway impact analysis	http://www.bioconductor.org/packages/2.12/bioc/html/SPIA.html	Tarca et al. (2009)	Molecular pathway	Mid-Scale
PAGED	An integrated pathway and gene enrichment database	http://bio.informatics.iupui.edu/PAGED	Huang et al. (2012)	System pathway	Mid-Scale
HAPPI	Human annotated and predicted protein interaction database	http://bio.informatics.iupui.edu/HAPPI	Chen et al. (2009)	Single molecule	Large-scale
STRING	Search tool for the retrieval of interacting genes/proteins	http://string.embl.de/	Franceschini et al. (2013)	Single molecule	Large-scale
CytoScape	An open source platform for complex network analysis and visualization	http://www.cytoscape.org/	Shannon et al. (2003)	Molecular pathway	Large-scale
ACOR	Ant colony optimization reordering	N/A	Wu et al. (2009), (2009b), (2009c), (2012)	Molecular pathway	Large-scale
Gene-Terrain	Terrain-based visual analysis for complex networks	N/A	Kim et al. (2001), You et al. (2010)	Network module	Large-scale

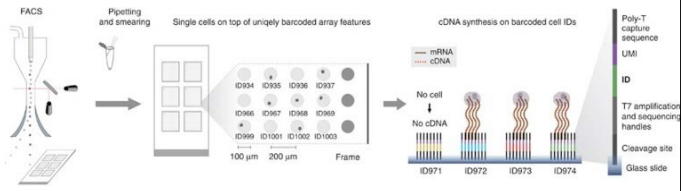
Network and matrix analysis of the respiratory disease interactome.
Garcia B, Datta G, Cosgrove GP, Strong M.
BMC Syst Biol. 2014 Mar 22;8:34.



The cell-cycle interactome: a source of growth regulators?
Blomme J, Inzé D, Gonzalez N.
J Exp Bot. 2014 Jun;65(10):2715-30.



Massive and parallel expression profiling using microarrayed single-cell sequencing.
Vickovic S, Ståhl PL, Salmén F, et al.
Nat Commun. 2016 Oct 14;7:13182.



MASC-Seq overview.
A FACS machine sorts single cells onto a barcoded microarray, printed with six replicates on an activated glass slide. The throughput of the method and microarray design as a 33 × 35 ID matrix is illustrated. An alternative is to pipette and smear cells which then distribute randomly onto the array. Positions of the cells and IDs are noted in a high-resolution image and cDNA is only transcribed when an individual cell lands on top of the barcoded oligo-dTVN primer (ID).

Spring 2017 – Epigenetics and Systems Biology
Lecture Outline (Systems Biology)
Michael K. Skinner – Biol 476/576
CUE 418, 10:35-11:50 am, Tuesdays & Thursdays
January 24 & 31, 2017
Weeks 3 and 4

Systems Biology (Components & Technology)

Components (DNA, Expression, Cellular, Organ, Physiology, Organism, Differentiation, Development, Phenotype, Evolution)

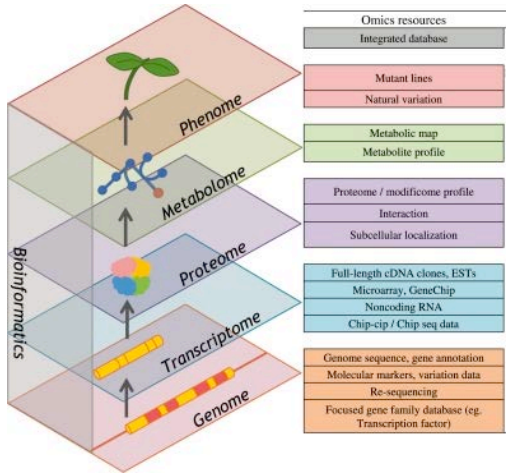
Technology (Genomics, Transcriptomes, Proteomics) (Interaction, Signaling, Metabolism)

Omics (Data Processing and Resources)

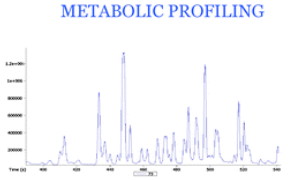
Required Reading

ENCODE (2012) ENCODE Explained. Nature 489:52-55.

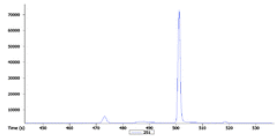
Street ME, et al. (2013) Artificial Neural Networks, and Evolutionary Algorithms as a systems biology approach to a data-base on fetal growth restriction. Prog Biophys Mol Biol. 113(3):433-8.



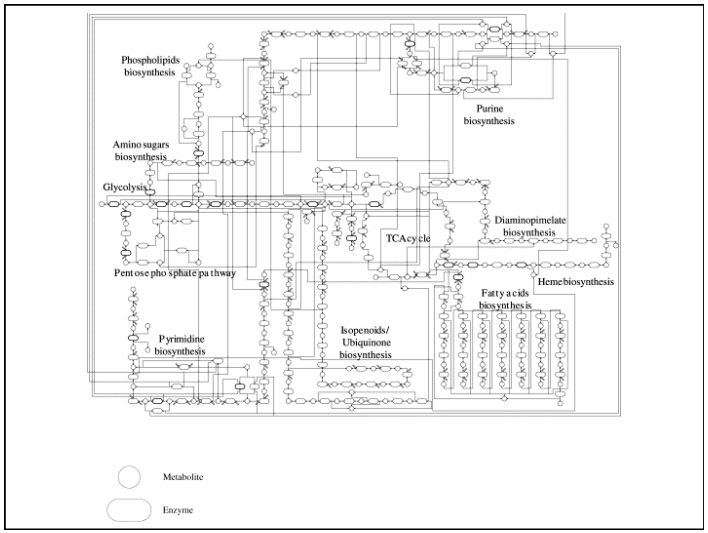
Metabolome

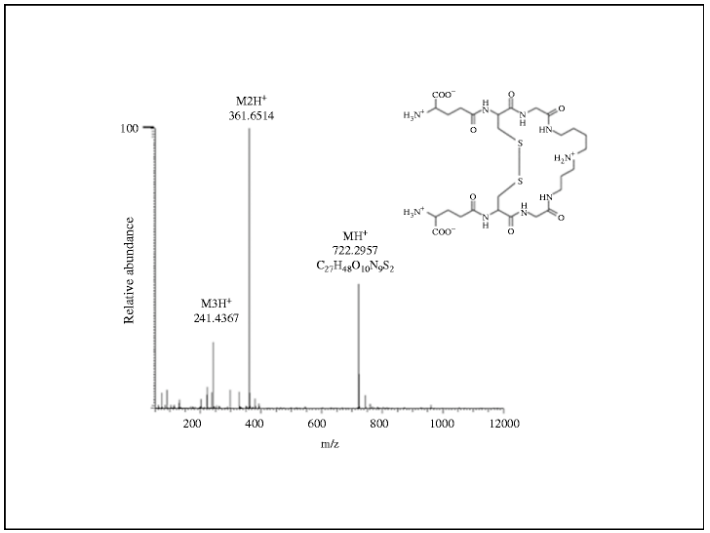


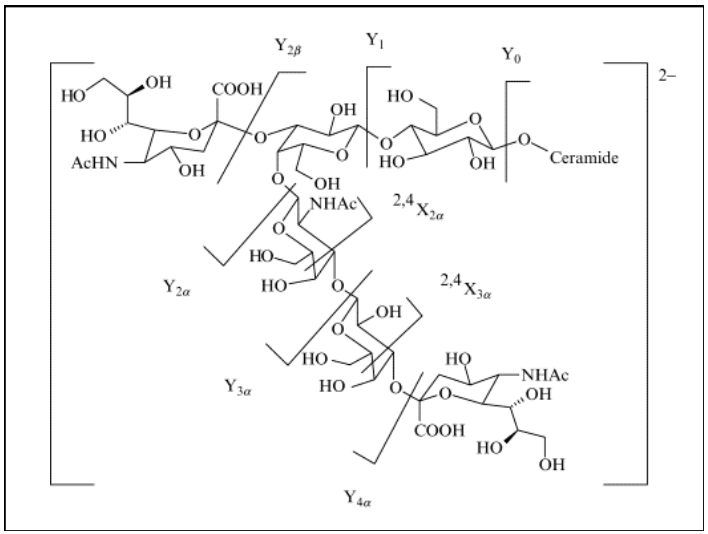
- (semi)-quantitative (global) detection of a wide range of metabolites
- data acquisition without *a priori* knowledge of biologically interesting metabolites
- search for the proverbial 'needles in the metabolic haystack'
- **inductive/hypothesis generating** experiment
- appropriate experimental design and data analysis is essential
- **MIDDLE-IN strategy**

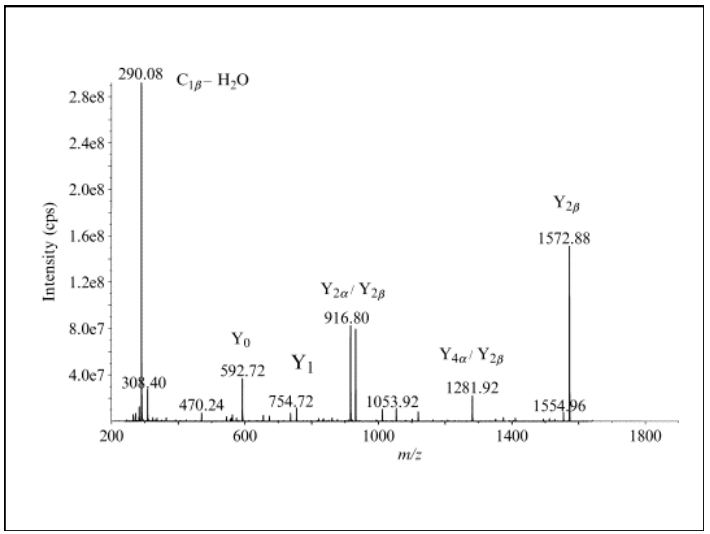


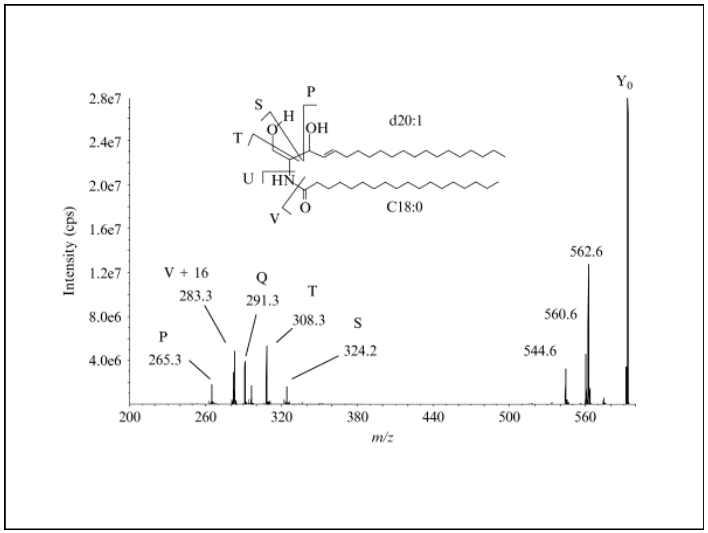
- quantification of a small number of related metabolites for **hypothesis testing** or systems biology modelling
- specificity provided by extensive sample preparation and MS/MS
- absolute quantification using isotopic internal standards or standards addition
- **BOTTOM-UP strategy**

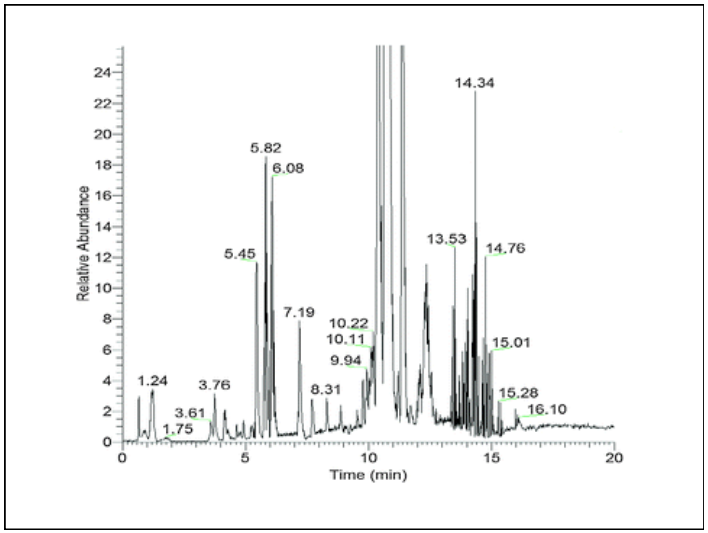

Phospholipids biosynthesis
Purine biosynthesis
Amino sugars biosynthesis
Glycolysis
Diaminopimelate biosynthesis
TCAcycle
Pentose phosphate pathway
Heme biosynthesis
Fatty acids biosynthesis
Pyrimidine biosynthesis
Isopenoids/ Ubiquinone biosynthesis
Metabolite
Enzyme

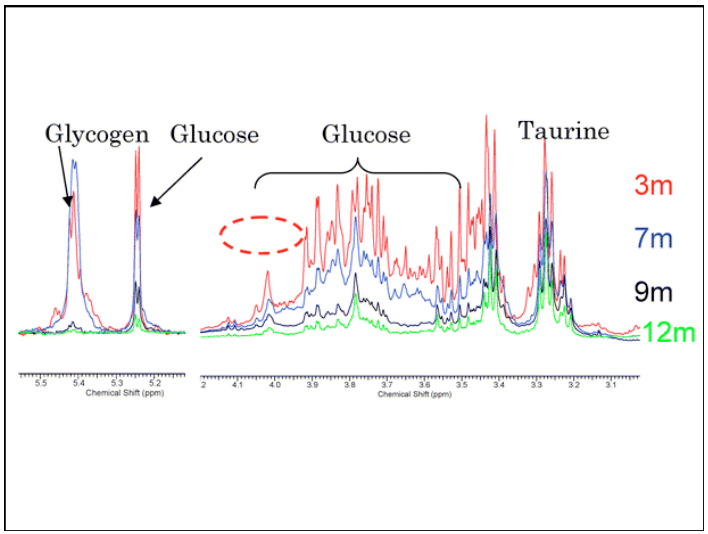

M2H+
361.6514
MH+
722.2957
C27H48O10N6S2
M3H+
241.4367
Relative abundance
100
200
400
600
800
1000
12000
m/z


Y2β
Y1
Y0
2–
HO
OH
COOH
HO
AcHN
Ceramide
NHAc
2,4X2α
Y2α
2,4X3α
Y3α
NHAc
Y4α


290.08
C1β– H2O
Y2β
1572.88
Y2α/ Y2β
916.80
Y0
592.72
Y1
Y4α/ Y2β
1281.92
308.40
470.24
754.72
1053.92
1554.96
Intensity (cps)
2.8e8
2.4e8
2.0e8
1.6e8
1.2e8
8.0e7
4.0e7
200
600
1000
1400
1800
m/z







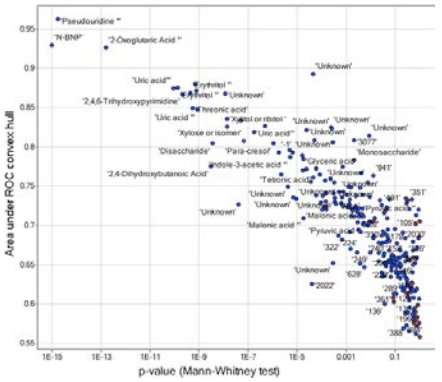


Fig. 13 An example of plots describing the relationship between area under ROC curve and p-values for various metabolites. These plots are applicable when comparing univariate biomarkers or multiple model predictions. The more effective biomarkers approach the top left hand corner of the plot (i.e., low p-value and high AuROC). Kindly reprinted from a study related to heart failure203 with permission from Springer.

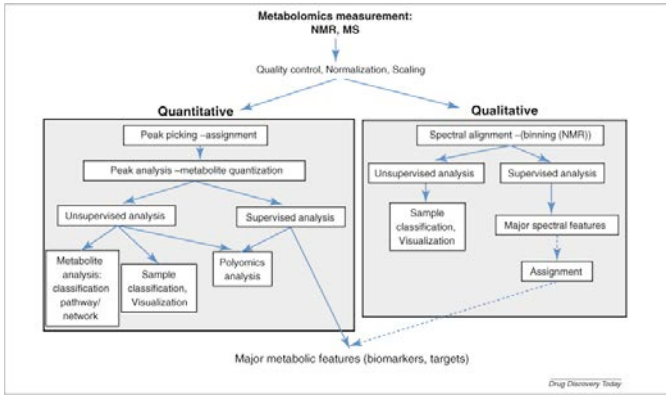


TABLE 1

Comparison of characteristics of major experimental methods for metabolomic analysis.

Analysis	NMR	MS
High throughput – metabolites	No	Medium
High throughput – samples; automation	Yes	No
Quantitative	Yes	Yes
Availability in clinic	No	No
Equipment cost	High	High
Maintenance cost	Medium	High
Per sample cost	Low	High
Required technical skills	Yes	Yes
Sensitivity	Medium	High
Reproducibility	High	Low
Data analysis automation	Yes	Yes
Identification of new metabolites	Difficult	Possible
Chemical exchange analysis	Yes	No
Stereoisomers analysis	Yes	Difficult
Sample preservation	Yes	No
***In vivo* measurement**	Possible	Impossible

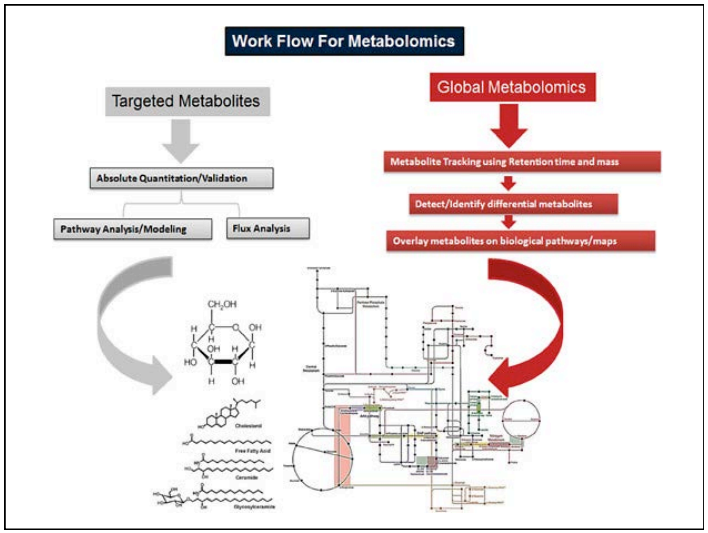
TABLE 2

Some major non-commercial databases of metabolomic standard data for quantification and assignment.

Name and availability	Instrument	Additional information
Human Metabolome Project [81] (http://www.hmdb.ca)	NMR, MS	Biological data; chemical and clinical data specific to humans
BMRD (http://www.bmrb.wisc.edu)	NMR	Database search for NMR peaks assignment
Prime (Akiyama [82]) (http://prime.psc.riken.jp)	MS, NMR	
Golm metabolome database (http://csbdb.mpimp-golm.mpg.de)	MS	Specific to plants
METLIN metabolite database (http://metlin.scripps.edu)	MS	Drug and drug metabolites; specific to humans
NIST Chemistry WebBook (http://Webbook.nist.gov/chemistry)	NMR, MS, IR	
Madison metabolomics database (http://mmcd.nmrfam.wisc.edu)	MS, NMR	
NMR Lab of biomolecules (http://spinportal.magnet.fsu.edu)	NMR	Database search for NMR peaks assignment

TABLE 3. Summary of medical metabolomics applications cited in this review.

Disease	Species	Material	Method	Approach	Specific biomarker species	Reference
Myocardial ischemia	Human	Plasma	LC/MS/MS	Targeted	Gamma-amino- butyric acid, uric acid, citrate	90,91
Type 2 diabetes	Mouse	Urine	NMR	Targeted	Mannose, 1,5–anhydroglucitol, phenylacetylglutamine	92
Type 2 diabetes	Human	Plasma	GC/MS, LC/MS/MS, NMR	Targeted	3-Indoxyl sulfate, glycerophospholipids, bile acids	93
Obesity	Human	Serum	LC/MS/MS	Non-targeted	Lysophosphatidylcholine	94
Obesity	Human	Serum	MS/MS	Targeted	Phosphatidylcholine	95
Cardiovascular disease	Human	Plasma	LC/MS/MS	Non-targeted	Trimethylamine *N*-oxide, choline, betaine	96
Ovarian carcinoma	Human	Tumor tissue	GC/MS	Non-targeted	Alpha-glycerolphosphate, uracil, glycine	97
Lung cancer	Human	Tissue, plasma	GC/MS, NMR	Stable isotope resolved analysis	(^{13}C-enrichment in lactate, alanine, succinate)	98
Pancreatic cancer	Human	Serum	GC/MS	Targeted	Thiodiglycolic acid, lactic acid, 7-hydroxyoctanoic acid	99
Hepatocellular carcinoma	Human	Urine	GC/MS	Non-targeted	Xylitol, urea, hydroxy proline dipeptide	100
Colorectal cancer	Human/rat	Urine/tissue	GC/MS	Targeted	Succinate, *N*-acetyl-aspartate, 2-hydroxyhippurate	101
Oral cancer	Human	Saliva	CE/MS	Non-targeted	Pyrroline, leucine + isoleucine, taurine	102
Breast cancer	Human	Saliva	CE/MS	Non-targeted	Taurine, putrescine, leucine + isoleucine	102
Pancreatic cancer	Human	Saliva	CE/MS	Non-targeted	Leucine + isoleucine, phenylalanine, alpha-amino butyric acid	102
Schizophrenia	Human	Cerebrospinal fluid	NMR	Non-targeted	Lactate, citrate, glucose	103
Parkinson's disease	Human	Plasma	NMR	Targeted	Threonate, myoinositol, suberate	104
Huntington's disease	Human/mouse	Serum	GC/MS	Non-targeted	Glycerol, urea, valine	105
Schizophrenia	Human	Plasma	LC/MS/MS, GC/MS	Targeted	Free fatty acids, triglycerides, phosphatidylethanolamine	106
Depression	Rat	Plasma	GC/MS	Targeted	Glucose, glutamine, butanedioic acid	107



Databases

16
Databases

Summary

With the rapid increase of biological data, it has become even more important to organize and structure the data in a way so that information can easily be retrieved. As a result, the number of databases has also increased rapidly over the past few years. Most of these databases have a web interface and can be accessed from everywhere in the world, which is an enormously important service for the scientific community. In the following, various databases are presented that might be relevant for systems biology.

Moreover, the journal *Nucleic Acids Research* offers a database issue each year in January dedicated to factual biological databases and in addition to this a web server issue each year in July presenting web-based services.

Databases Sources

- National Center for Bioinformatics NCBI
- European Bioinformatics Institute
- EMBL
- Ensembl
- Interpro
- Protein databank
- Bionumbers
- Gene Ontology
- Pathway- KEGG
- Consensus Path DB

Omics data set
A generic term that describes the genome-scale data sets that are emerging from high-throughput technologies. Examples include whole-genome sequencing data (genomics) and microarray-based genome-wide expression profiles (transcriptomes).

Data mining
An analytical discipline that is focused on finding unsuspected relationships and summarizing often large observational data sets in new ways that are both understandable and useful to the data owner.

***In silico* prediction**
A general term that refers to a computational prediction that usually results from the analysis of a mathematical or computational model.

Unsupervised analysis: Unsupervised analysis includes methods used for grouping of features (sample, metabolites and spectral features) according to the molecular data measured. These methods are used for the analysis of features when no prior information is available about the system. Depending on the method, the analysis might or might not require the user to define the number of clusters. In terms of cell culture metabolomics, this method is ideal for discovery of novel classes.

Supervised analysis: Supervised analysis defines methods for sample grouping or classification and for selection of major sample defining features. In supervised analysis, a set of features is pre-assigned to a class and it is used as a training set for the method of choice to define a classifier that will be used for classification of an unknown sample. Supervised analysis creates a model from the training set and, thus, can only be accurately used for classification of a different dataset (i.e. supervised analysis requires application of cross-validation for the determination of accuracy of the classifier).

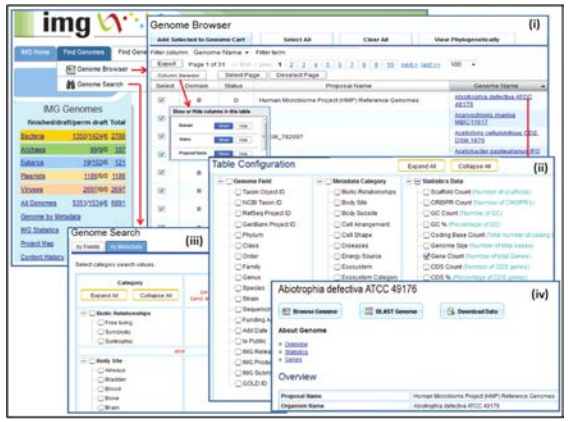
Table 1 | **'Omics' data repositories***

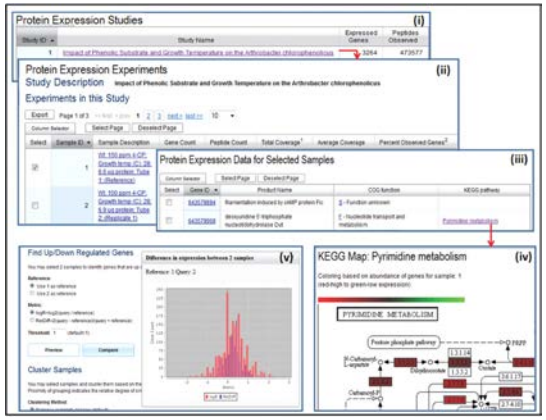
Data types	Online resource	Description	URL
Components			
Genomics	Genomes OnLine Database (GOLD)	Repository of completed and ongoing genome projects	http://www.genomesonline.org
Transcriptomics	Gene Expression Omnibus (GEO)	Microarray and SAGE-based genome-wide expression profiles	http://www.ncbi.nlm.nih.gov/geo
	Stanford Microarray Database (SMD)	Microarray-based genome-wide expression data	http://genome-www.stanford.edu/microarray
Proteomics	World-2DPAGE	Links to 2D-PAGE data	http://us.expasy.org/ch2d/2d-index.html
	Open Proteomics Database (OPD)	Mass-spectrometry-based proteomics data	http://bioinformatics.icmb.utexas.edu/OPD
Lipidomics	Lipid Metabolites and Pathways Strategy (LIPID MAPS)	Genome-scale lipids database	http://www.lipidmaps.org
Localizomics	Yeast GFP Fusion Localization Database	Yeast genome-scale protein-localization data	http://yeastgfp.ucsf.edu
Interactions			
Protein–DNA	Biomolecular Network Database (BIND)	Published protein–DNA interactions	http://www.bind.ca/Action/
	Encyclopedia of DNA Elements (ENCODE)	Database of functional elements in human DNA	http://genome.ucsc.edu/ENCODE/index.html
Protein–protein	Munich Information Center for Protein Sequences (MIPS)	Links to protein–protein-interaction data and resources	http://mips.gsf.de/proj/ppi
	Database of Interacting Proteins (DIP)	Published protein–protein interactions	http://dip.doe-mbi.ucla.edu
Functional states			
Phenomics	RNAi database	*C. elegans* RNAi screen data	http://rnai.org
	General Repository for Interaction Datasets (GRID)	Synthetic-lethal interactions in yeast	http://biodata.mshri.on.ca/grid
	A Systematic Annotation Package For Community Analysis of Genomes (ASAP)	Single-gene-deletion microarray data for *E. coli* phenotypes	http://www.genome.wisc.edu/tools/asap.htm

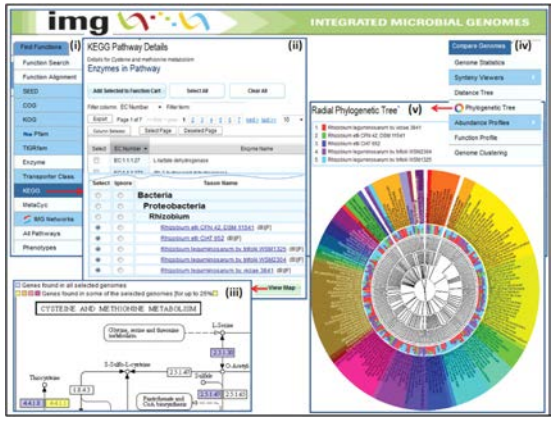
*This table details some of the databases that store and distribute genome-scale omics data sets through publicly accessible Web sites. Some omics technologies do not yet have associated data-dissemination resources — notably metabolomics, glycomics and fluxomics — and are therefore not included in this table. It should also be noted that this table does not represent all publicly available omics data resources, but, rather, provides a reasonably broad sample of the data that are readily accessible to researchers today. *C. elegans*, *Caenorhabditis elegans*; 2D-PAGE, two-dimensional polyacrylamide-gel electrophoresis; *E. coli*, *Escherichia coli*; GFP, green fluorescent protein; RNAi, RNA interference; SAGE, serial analysis of gene expression.

Table I: Useful online resources for systems biology and modeling of the human microbiome

Resources	References
Microbial genomic data and analysis	
IMG	[80]
DACC	[81]
GOLD	[31]
Microbes online	[82]
RAST	[83]
Metagenomic data and analysis	
IMG/M	[84]
MG-RAST	[85]
METAREP	[86]
Metabolic databases	
KEGG	[23]
MetaCyc	[24]
Brenda	[87]
Metabolic model reconstruction, visualization and analysis	
The Model Seed	[34]
Systems Biology Research Group	[88]
iPath	[89]
Pathway Tools	[90]
Cytoscape	[91]
Cobra	[92]
Reverse ecology software	
NetSeed	[44]







Networks

Network scaffold

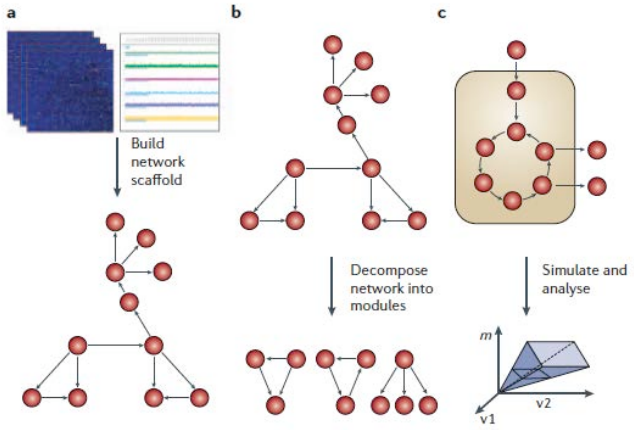
Refers to the structure of a network that specifies the components of the network and the interactions between them, and represents the end product of the network-reconstruction process.

Network module

A portion of a biological network that is composed of multiple molecular entities (such as genes, proteins or metabolites) that work together as a distinct unit within the cell, for example, in response to certain stimuli or as part of a developmental or differentiation programme.

Table 1
Comparison of various types of mathematical models

Type of model	Characteristics	Advantages	Disadvantages	Example
Unstructured	Excludes intracellular reactions.	Easy to build. Well established model.	Cannot reproduce the complex behavior of cells.	Monod (1949)
Structured	Includes intracellular reactions.	Capacity to express complex systems.	Accurate, detailed information required.	Tomita et al. (1999)
Lumped constant system	Regards cells as homogeneous.	Easy to build. Well established model.	Expresses only average state.	(Most cases)
Distributed constant system	Considers parameter distribution within cells.	Good approximation for real cell mass.	Subjectivity in division rules and parameter measurement.	Dairaku et al. (1982)
Dynamic	Variables change with time.	Transition states can be calculated.	Kinetic parameters and initial values required.	Chassagnole et al. (2002)
Static	Variables remain constant.	Can incorporate a number of reactions.	Cannot express transition states.	Schilling et al. (1999)



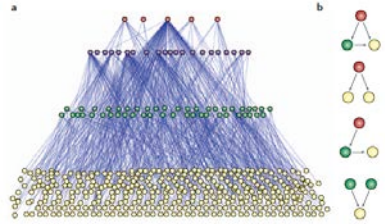
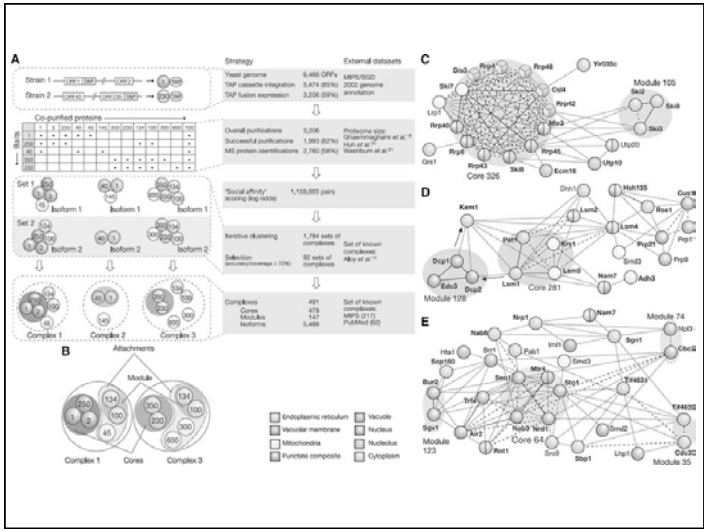


Figure 3 | Network-motif enrichment: an example of network decomposition. One approach to decomposing complex networks involves the identification of enriched motifs, or interaction modules, that exist within networks. **a** | A reconstructed *Escherichia coli* regulatory network. This diagram represents a hierarchical representation of the regulatory network that was previously used in an integrated regulatory and metabolic model of *E. coli*[96]. This relatively complex network structure that contains 104 regulators and 479 target-enzyme genes is organized so that global regulators are depicted in red, major regulators in purple, minor regulators in green, and target genes — in this case model enzymes — in yellow. The blue lines represent direct regulatory reactions, either between regulators or between regulator and target enzymes. Genes were arbitrarily assigned to the regulator class (global versus major versus minor) on the basis of the total number of target genes that are either directly under their control, or that are indirectly controlled by the direct interaction with other regulators. **b** | Network-motif-enrichment analysis. To illustrate one method that researchers use to decompose biological networks, network-motif-enrichment analysis[46] was conducted to identify the relative enrichment of the 13 possible types of three-node motifs that could be present within the depicted *E. coli* regulatory network. The four three-node motifs that were detected within the network are shown. Only the top motif was identified as occurring more frequently than would be expected by chance, as determined by comparisons with randomized networks. This motif, which is commonly known as the feed-forward loop, has been frequently observed in biological-network analyses and is thought to serve as a filter against noise within networks that are responding to various signals.

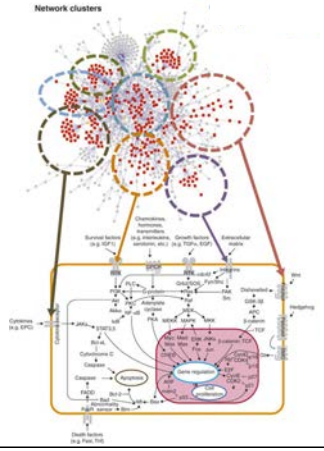


Network reconstruction
The process of integrating different data sources to create a representation of the chemical events that underlie a biochemical reaction network.

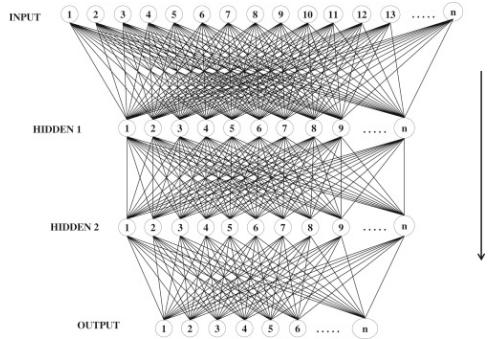
Governing constraints
Biochemical networks and cellular systems are constrained by natural law. These governing constraints include physico-chemical constraints (such as enzyme turnover), topobiological constraints (such as cellular crowding), environmental constraints (such as nutrient availability) and regulatory constraints (such as gene repression in response to external signals).

Constraint-based reconstruction and analysis
(COBRA). A genome-scale modeling approach that involves: first, the reconstruction of biochemical reaction of networks; then, applying constraints to the network; and finally, analyzing the characteristics and capabilities of the network using various computational techniques.

Network analysis: a new approach to study endocrine disorders.
Stevens A, et al.
J Mol Endocrinol. 2013 Dec 19;52(1):R79-93.



Artificial Neural Networks, and Evolutionary Algorithms as a systems biology approach to a data-base on fetal growth restriction.
Street M, et al.
Prog Biophys Mol Biol. 2013 Dec;113(3):433-8.

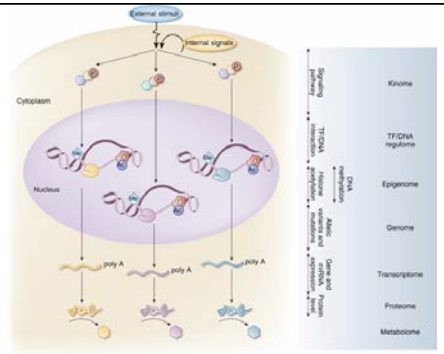


Typical neural network architecture. The basic elements of ANN are the nodes, also called processing elements (PE), and their connections. Each node has its own input, from which it receives communications from other nodes and/or from the environment and its own output, from which it communicates with other nodes or with the environment. Finally, each node has a function through which it transforms its own global input into an output.

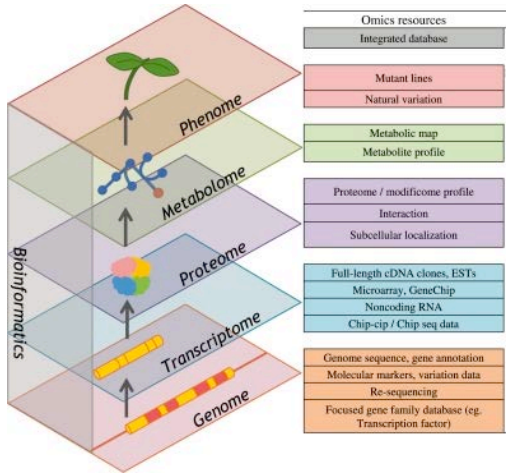
Omics Data Integration

Omics data integration

The simultaneous analysis of high-throughput genome-scale data that is aimed at developing models of biological systems to assess their properties and behavior.



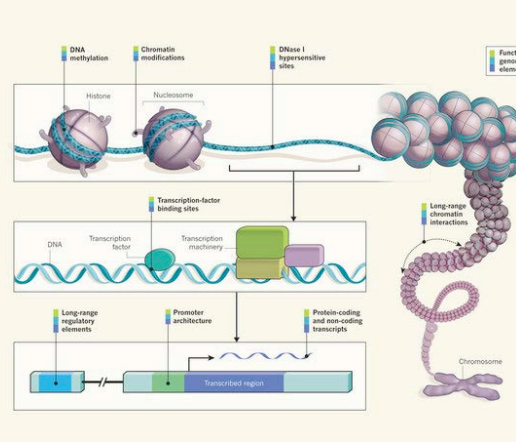
'Omic' approaches to profile cellular processes. Both external stimuli and internal signals can initiate the activation of downstream signaling pathways through post-transcriptional modification of signaling molecules, such as protein kinases, by phosphorylation. Activation of the signaling cascade leads to either transcriptional activation by recruitment and assembly of the transcriptional complex on the DNA, transcription factor/DNA (TF/DNA) interaction, or inhibition of transcriptional activation by exclusion of key TFs from the complex. Covalent modifications of the DNA by methylation and of histones by acetylation have a major role in the regulation of gene expression. Variations in the coding region of the genome result in mutations of the gene transcripts, and changes in non-coding region could affect the regulation of gene expression. Gene transcripts as messenger RNAs in the cytoplasm are translated into proteins. The protein products synthesized in response to the initial signal serve to maintain cellular integrity and react to perturbations in the systems by initiating additional signals or catalyze reactions that generate specific metabolites as a by-product. Characterization of these processes at the systems level is known as the 'omics' approaches: kinome for phosphorylation of proteins, TF/DNA regulome for regulation of transcription by the interaction of TF with DNA, epigenome for modification of histones and DNA, genome for the sequence of DNA, transcriptome for the mRNA transcripts of the expressed genes, proteome for the protein composition, and metabolome for the metabolites that are generated in a specific tissue or cell (Red P=phosphorylation, Me=methylation, Ac=acetylation, line with poly A tail=mRNA). miRNA, microRNA.

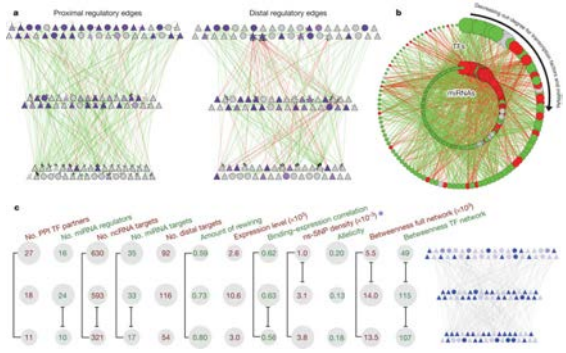


FORUM: Genomics

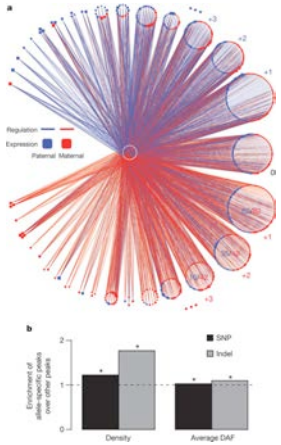
ENCODE explained

The Encyclopedia of DNA Elements (ENCODE) project dishes up a hearty banquet of data that illuminate the roles of the functional elements of the human genome. Here, five scientists describe the project and discuss how the data are influencing research directions across many fields. See Articles p.57, p.75, p.83, p.91, p.101 & Letter p.109

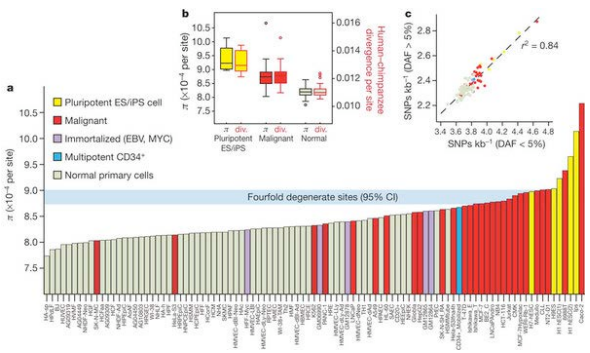




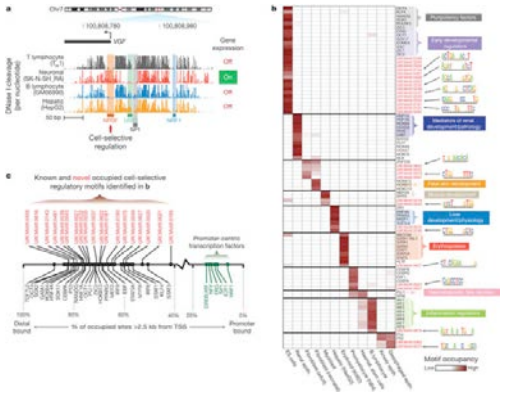
Overall network. a, Close-up representation of the transcription factor hierarchy. Nodes depict transcription factors. TFSSs are triangles, and non-TFSSs are circles. Left: proximal-edge hierarchy with downward pointing edges coloured in green and upward pointing ones coloured in red. The nodes are shaded according to their out-degree in the full network (as described in Table 1). Right: factors placed in the same proximal hierarchy but now with edges corresponding to distal regulation coloured green and red, and nodes re-coloured according to out-degree in the distal network. The distal edges do not follow the proximal-edge hierarchy. **b**, Close-up view of transcription-factor–miRNA regulation. The outer circle contains the 119 transcription factor, whereas the inner circle contains miRNAs. Red edges correspond to miRNAs regulating transcription factors; green edges show transcription factors regulating miRNAs. Transcription factors and miRNAs each are arranged by their out-degree, beginning at the top (12:00) and decreasing in order clockwise. Node sizes are proportional to out-degree. For transcription factors, the out-degree is as described in Table 1; for miRNAs, it is according to the out-degree in this network. Red nodes are enriched for miRNA–transcription factor edges and green nodes are enriched for transcription factor–miRNA edges. Grey nodes have a balanced number of edges (within ±1). **c**, Average values of various properties (topological, dynamic, expression-related and selection-related—ordered consistently with Table 1) for each level are shown for the proximal-edge hierarchy. The top, middle and bottom rows correspond to the top, middle and bottom of the hierarchy, respectively. The sizing of the grey circles indicates the relative ordering of the values for the three levels. Significantly different values ($P < 0.05$) using the Wilcoxon rank-sum test are indicated by black brackets. The proximal-edge hierarchy depicted on the right shows non-synonymous SNP (ns-SNP) density, where the shading corresponds to the density for the associated factor. (See Supplementary Fig. 4 for more details.)



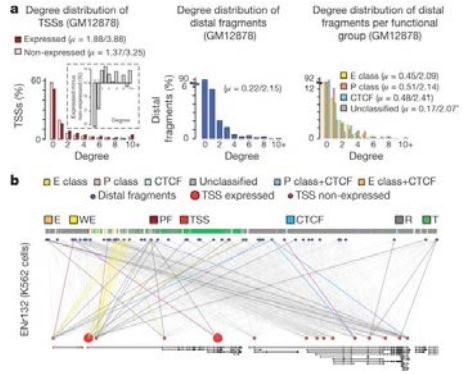
Allelic effects. a, An 'allelic effects network' depicting the increasing coordination between allele-specific binding and allele-specific expression as the number of factors regulating a target increases. Central white nodes denote transcription factors, and peripheral nodes denote targets, which are blue (red) if they are expressed from the paternal (maternal) allele. Blue (red) edges denote allele-specific binding to the paternal (maternal) allele. This network represents the strongest differences between the paternal- and maternal-specific regulatory networks. As one goes around the larger circle anticlockwise (clockwise), each of the small circular clusters represents targets with progressively more paternal (maternal) regulation, indicated by the small blue (red) numbers to the side of the clusters. Moreover, within each of the clusters the fraction of predominantly paternally (maternally) expressed targets increases as one goes around the larger circle. As an illustration, this fraction is explicitly indicated by the ratios within three of the larger clusters at the bottom right. **b**, Relationship between transcription factor allelicity and selection. The bar height is the ratio of the degree of selection (as measured by SNP density or average DAF) in those binding peaks showing allelic behaviour to the degree of selection in all other binding peaks. Asterisks represent significant differences ($P < 0.05$, Wilcoxon rank-sum test). (See Supplementary Information section I.2 and Supplementary Fig. 10b, c for details.)



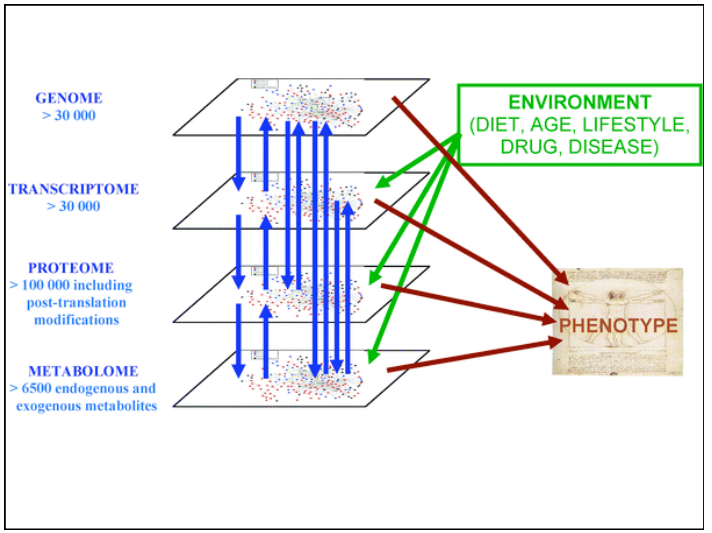
Genetic variation in regulatory DNA linked to mutation rate. a, Mean nucleotide diversity (π, y axis) in DHSs of 97 diverse cell types (x axis) estimated using whole-genome sequencing data from 53 unrelated individuals. Cell types are ordered left-to-right by increasing mean π. Horizontal blue bar shows 95% confidence intervals on mean π in a background model of fourfold degenerate coding sites. Note the enrichment of immortal cells at right. ES, embryonic stem; iPS, induced pluripotent stem. **b**, Mean π (left y axis) for pluripotent (yellow) versus malignancy-derived (red) versus normal cells (light green), plotted side-by-side with human–chimpanzee divergence (right y axis) computed on the same groups. Boxes indicate 25–75 percentiles, with medians highlighted. **c**, Both low- and high-frequency derived alleles show the same effect. Density of SNPs in DHSs with derived allele frequency (DAF) <5% (x axis) is tightly correlated (r2 = 0.84) with the same measure computed for higher-frequency derived alleles (y axis). Colour-coding is the same as in panel a.



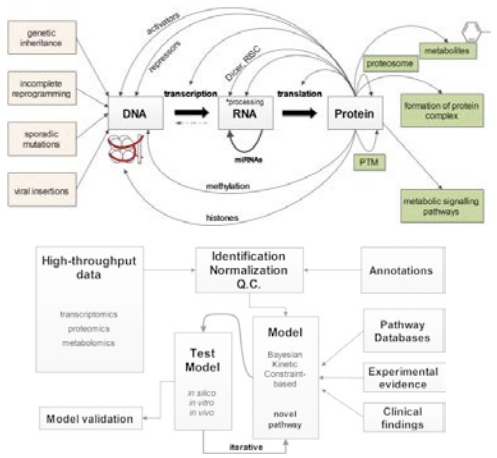
Multi-lineage DNase I footprinting reveals cell-selective gene regulators. a, Comparative footprinting of the nerve growth factor gene (*VGF*) promoter in multiple cell types reveals both conserved (NRF1, USF1 and SP1) and cell-selective (NRSF) DNase I footprints. **b,** Shown is a heat map of footprint occupancy computed across 12 cell types (columns) for 89 motifs (rows), including well-characterized cell/tissue-selective regulators, and novel *de novo*-derived motifs (red text). The motif models for some of these novel *de novo*-derived motifs are indicated next to the heat map. **c,** The proportion of motif instances in DNase I footprints within distal regulatory regions for known (black) and novel (red) cell-type-specific regulators in b is indicated. Also noted are these values for a small set of known promoter-proximal regulators (green). ES, embryonic stem.



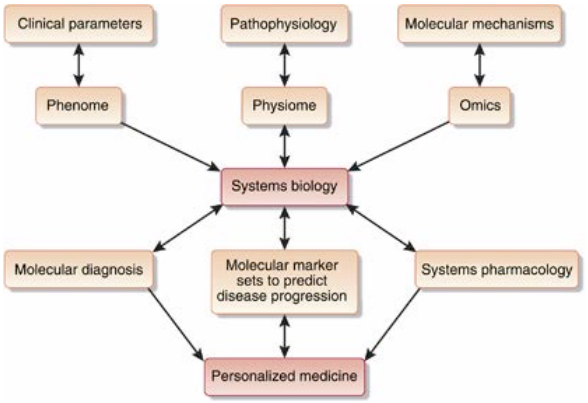
Networks of looping interactions. a, Histogram showing the number of TSSs (left, red) or distal fragments (middle, blue) in percentages that are involved in 0, 1, 2,….10 (and above) looping interactions (degree, x axis) in GM12878 cells. All of the values for degrees that are >9 are grouped under degree 10+. The dark red bars represent the percentages of looping TSSs that are expressed whereas light red bars represent the percentages of looping TSSs that are not expressed. Inset: the difference in percentage between looping TSSs that are expressed and not expressed for each degree is shown. The right panel shows the degree distribution for each functional group of distal fragments. The average degrees (mean, μ) for TSSs and distal fragments are indicated. The first value is the mean degree considering all the TSS/distal fragments (looping plus non-looping), whereas the second value is the mean degree of looping TSS/distal fragments (excluding degree = 0). **b**, Web plot showing the long-range looping interactions in the ENr132 region in K562 cells. The interrogated distal fragments (blue circles) and the TSSs (red circles) are positioned according to genomic coordinates and the GENCODE v7 gene annotation is indicated. The size of the red circles indicates whether that TSS is expressed (large circles) or not expressed (small circles). The thin grey lines show all the interactions that were interrogated. The coloured lines show significant looping interactions between TSSs and distal fragments of a particular group.



Current advances in systems and integrative biology.
Robinson SW, Fernandes M, Husi H.
Comput Struct Biotechnol J. 2014 Aug 27;11(18):35-46.

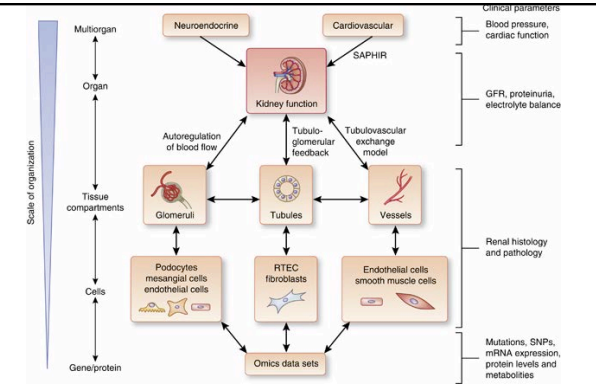


Systems Biology and Medicine

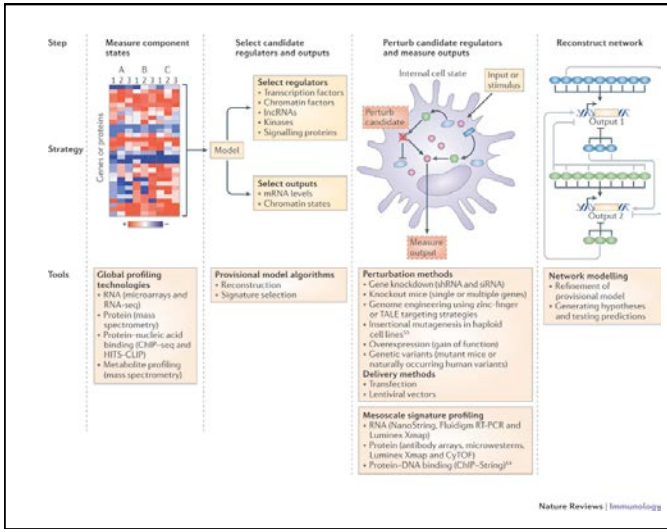


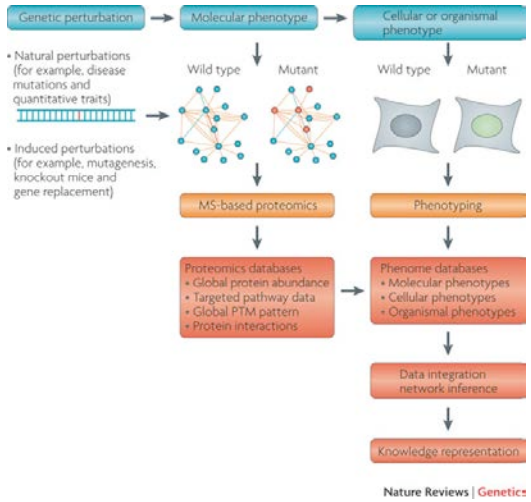
Personalized genomic medicine

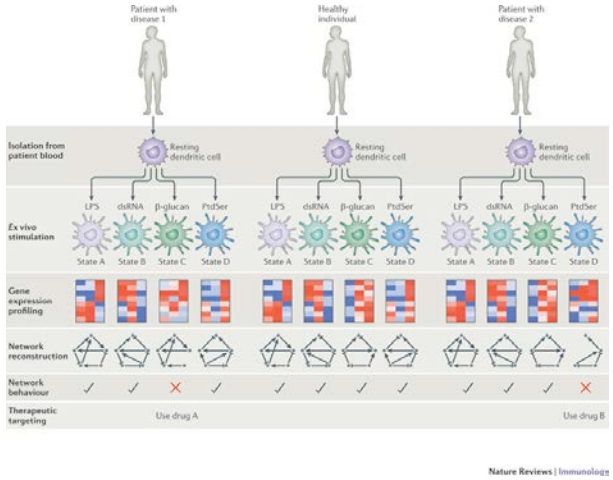
The idea that genome-scale technologies will allow clinicians to apply treatment regimens that are tailored specifically to an individual patient on the basis of their genetic makeup and associated predispositions.

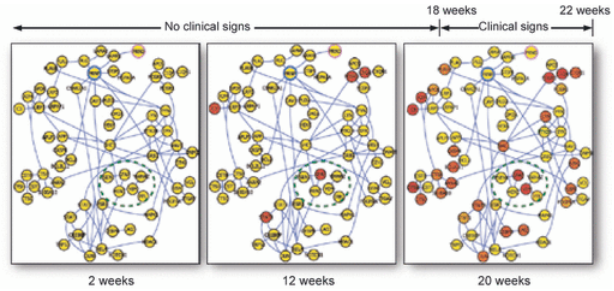


Multiscale analysis of kidney function. Maintenance of renal function requires the coordinated regulation from other organ systems (neuroendocrine and cardiovascular) and various tissue compartments and cells within the kidney. To recapitulate normal renal physiology in biological models, this multiscale organization from organ systems down to cell/gene level will need to be determined. The interactions at several levels have been described: Systems Approach for Physiological Integration of Renal, Cardiac, and Respiratory (SAPHIR) models, as well as models of autoregulation of glomerular blood flow, tubuloglomerular feedback, and tubulovascular exchange. Clinical parameters that we can use to assess and infer the function of organ systems and organs include blood pressure and cardiac function for cardiovascular and neuroendocrine input into the kidney, glomerular filtration rate (GFR), and proteinuria as determinants of the filtration function of the glomeruli, balance of electrolytes as an indicator of tubular function, podocyte number, foot process effacement, mesangial deposition, glomerulosclerosis, and tubulointerstitial fibrosis on renal histology as indicators of disease severity, and genetic variations (mutations, single-nucleotide polymorphisms (SNPs)) and mRNA and protein expression levels as indices of cellular response to internal and external stimuli. RTEC, renal tubular epithelial cell.

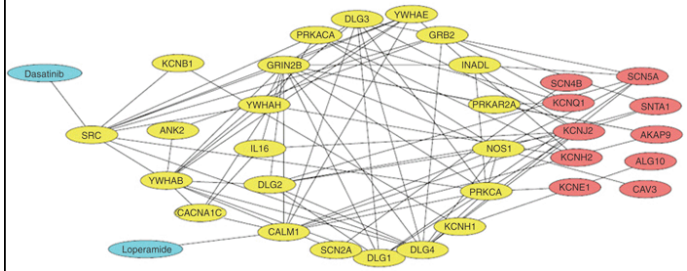




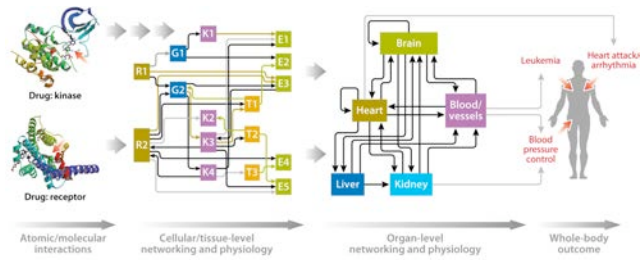




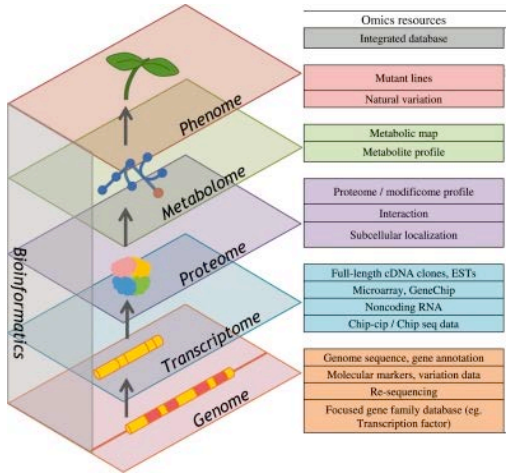
A schematic of the network perturbations of one neural degenerative network over the 20 weeks of the progression of this disease in a mouse model. The red nodes indicate mRNAs that have become disease perturbed as compared with the brain transcripts of normal mice. The spreading of the disease-perturbed networks at the three different times points is striking – indicating the progressive disease perturbation of this neurodegenerative network.



Cellular regulatory networks can explain rare adverse events. A disease module focused on long-QT syndrome connecting targets for two drugs (blue), the Src inhibitor dasatinib (a cancer drug) and loperamide (an antidiarrheal) that interacts with calmodulin (CALM1), through intermediate nodes to genes known to be associated with congenital forms of the long-QT syndrome. Such a subnetwork provides a plausible explanation of how these drugs used to treat different pathophysiologies produce the same adverse event. Figure from Berger et al.[25]



Zhao S, Iyengar R. 2012.
Annu. Rev. Pharmacol. Toxicol. 52:505–21



Evolutionary Systems Biology		Actual Systems Biology		Revolutionary Systems Biology	
	Ecosystem		Ecosystem		Ecosystem
	↑		↑↓		↓
	Populations		Populations		Populations
	↑		↑↓		↓
	Organisms		Organisms		Organisms
Linear Stasis	↑		↑↓		↓
	Physiology		Physiology		Physiology
	↑		↑↓		↓
	Organ Systems		Organ Systems		Organ Systems
Homeostasis	↑		↑↓		↓
	Organs		Organs		Organs
	↑		↑↓		↓
	Tissues		Tissues		Tissues
Mechanism	↑		↑↓		↓
	Cells		Cells		Cells
	↑		↑↓		↓
	Organelles		Organelles		Organelles
Genetic Determinism	↑		↑↓		↓
	Macromolecules		Macromolecules		Macromolecules
	↑		↑↓		↓
	DNA		DNA		DNA
Reductionism					

Nonlinear	
Robustness	
Synergism	
Emergence	
Epigenetics	
Holism	

Application system biology

Systems Biology Approaches to a Rational Drug Discovery Paradigm.
Prathipati P, Mizuguchi K.
Curr Top Med Chem. 2016;16(9):1009-25.

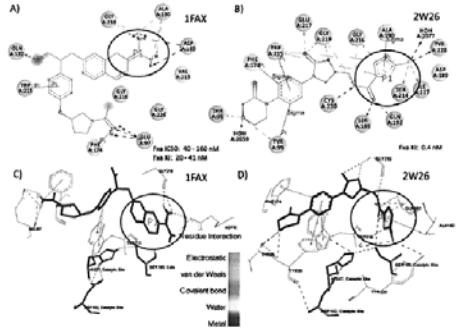


Fig. (1). The advantages of structure-based approaches augmenting ligand-based approaches are exemplified using FXa inhibitors Dabigatran (A and B) and Rivaroxaban (C and D). Using a ligand-based approach, a basic fragment at P1 site (A and B) was shown to be essential for FXa inhibition. Hence, the optimization strategy involved exploring other basic fragments at this site, which led to unfavorable ADMET profiles. A structure-based approach was used for replacing the basic amidine fragment of Dabigatran at the P1 pocket with a neutral chlorothiophene fragment by exploiting FXa's binding site information and the full range of intermolecular interactions such pi-pi, cation-pi and anion-pi, in addition to the usual hydrogen bonding, hydrophobic and salt bridge interactions. This strategy has led to the design of the new inhibitor Rivaroxaban (C and D) with favorable ADMET profiles. For more details, the readers can refer to a review by Nar *et al.* [4].

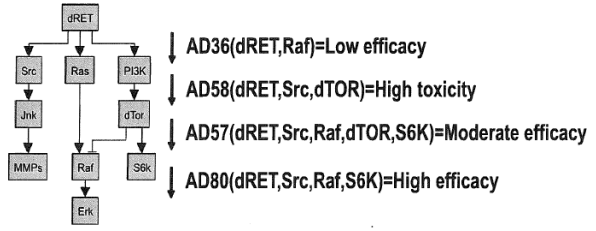


Fig. (2). The cancer model built using the fruit fly Drosophila pathway[†] data was used to rationally arrive at a compound (AD80) that was shown to inhibit four of the 10 selected cancer targets, dRET, Src, Raf and S6K. This chemoproteomic profile (inhibiting these four targets out of the 10) was associated with high efficacy and low toxicity in whole animal screening. Further SAR analysis also led to the identification of dTOR as an anti-target responsible for toxicity. AD80 proved far more effective and less toxic than standard cancer drugs, which generally focus on a single target. This study by Dar *et al.*[6] was the first time that whole-animal screening has been used in a rational, step-wise approach to identifying favorable chemoproteomic profiles and laid the case for a rational systems biology approach to drug discovery. For a more general discussion of chemoproteomics, see section 3.5.

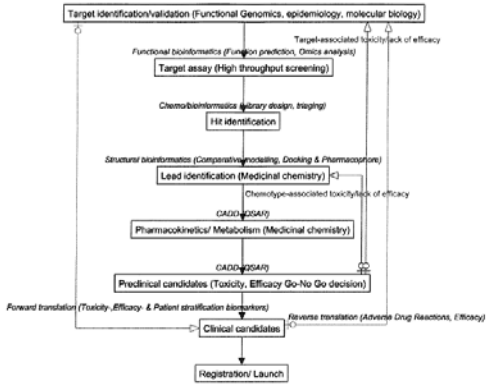


Fig. (3). A schematic flow chart summarizing the process of drug discovery, including major *in silico* contributions (italicized) to chemistry, biology[†] and ADMET.

Neurological Diseases from a Systems Medicine Point of View

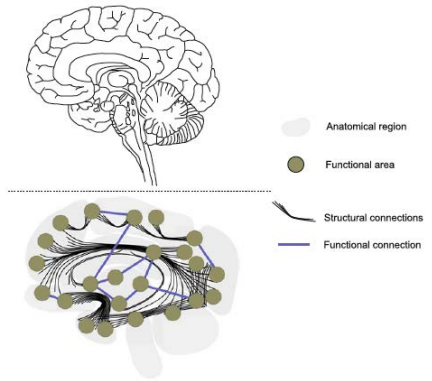


Fig. 3 Brain-level networks represent connections between anatomical and functional brain areas, representing structural connections of directly interacting groups of neuronal cells, or functional associations between areas co-activated during a given type of brain activity

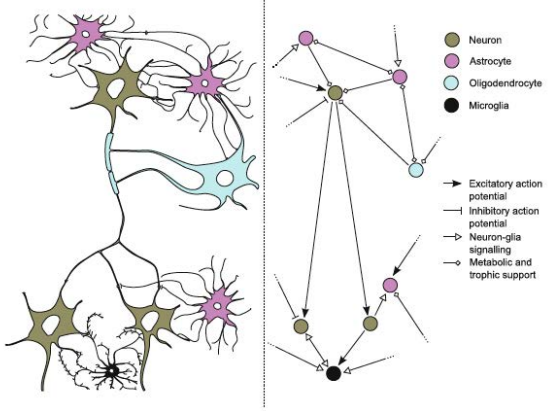
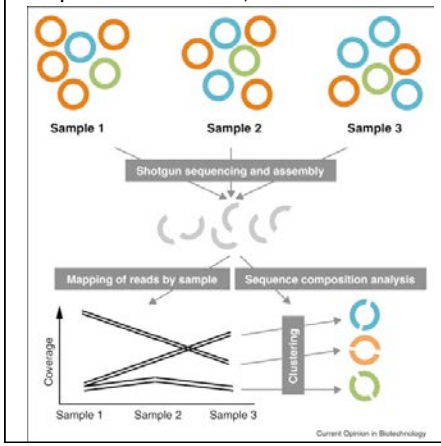


Fig. 2 The diverse cell types within the brain generate a complex interaction network with different classes of interaction edges that also exhibit distinguished dynamic properties. Excitable connections (excitatory or inhibitory) denote action potentials of the neurons. Metabolic and trophic support interactions represent exchange of substrates required for cellular network homeostasis. Topology-altering interactions denote cellular mechanisms leading to changes in the local network structure

High definition for systems biology of microbial communities: metagenomics gets genome-centric and strain-resolved.
Turaev D, Rattei T.
Curr Opin Biotechnol. 2016 Jun;39:174-81.



Multiple coverage binning. Multiple samples of the same microbiome are taken, for example, at different time points. The extracted DNA is sequenced and assembled into contigs. For each contig, the coverage in each sample is estimated by mapping of the reads from each sample. The coverage profiles and sequence composition statistics data are clustered. Contigs having similar coverage profiles and similar sequence composition are thereby grouped into the same bin.

Plant synthetic biology for molecular engineering of signalling and development.
Nemhauser JL, Torii KU.
Nat Plants. 2016 Mar 2;2:16010.

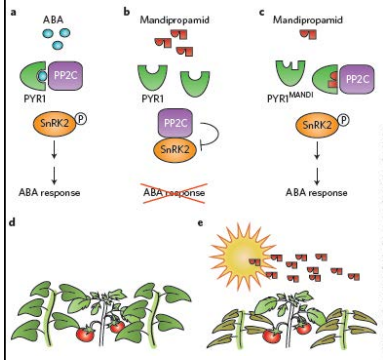
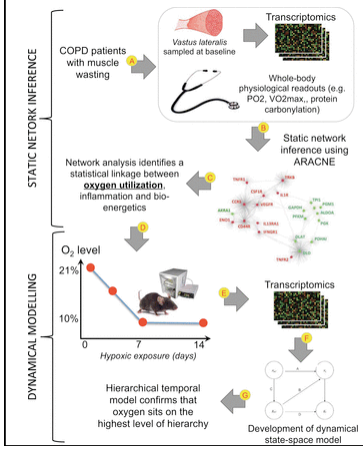
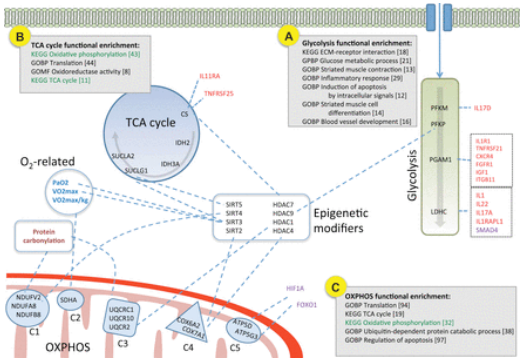


Figure 2 | An engineered ABA receptor can perceive a fungicide and trigger an ABA response. a, Endogenous ABA signalling. Binding of ABA to the PYR1 receptor promotes PYR1–PP2C association, which in turn activates the downstream protein kinase SnRK2. Active SnRK2 triggers ABA responses, including stomatal closure and drought tolerance. **b**, Mandipropamid treatment does not elicit an ABA response in wild-type plants. **c**, In a plant expressing the engineered $PYR1^{MANDI}$ receptor, mandipropamid treatment triggers ABA response and, consequently, drought tolerance. **d**,**e**, Potential and idealized translational application of a synthetic ligand–receptor system in a crop field. Here, a transgenic tomato crop plant expressing $PYR1^{MANDI}$ is outcompeted by nearby weeds (**d**). During drought season, mandipropamid application triggers ABA response to the transgenic tomato plant, thereby boosting its drought tolerance. Surrounding non-transgenic weeds do not respond to the chemical spray. Mandipropamid has already been approved by the Environmental Protection Agency as a fungicide (reg. no. 100-1281) for field application.

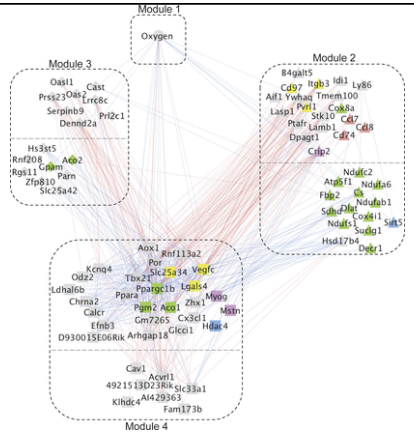
Multilevel functional genomics data integration as a tool for understanding physiology: a network biology perspective.
Davidsen PK, Turan N, Egginton S, Falciani F.
J Appl Physiol (1985). 2016 Feb 1;120(3):297-309.



Schematic representation of the analysis strategy used in the case study, highlighting how the inferred static multiscale network from the clinical chronic obstructive pulmonary disease (COPD) cohort (A–C) can be bridged to the inference of a dynamic network representing the temporal progression of events following an experimental challenge (hypoxic exposure) in a murine animal model (D–G). Having identified a clinical condition with known outcome (exercise intolerance in patients with respiratory disease), we could target unknown mechanisms by focusing on one likely source of functional limitation (skeletal muscle dysfunction ± central limitation on O2 supply) and generate data characterizing the phenotype. Both genomic and physiological readouts were used to construct a network of inferred interactions, which was then interrogated to identify statistically robust linkages among broad biological functions. While very useful in providing a list of useful biomarkers, there remains a potential limitation with single-point associations. The dynamic nature of relationships is captured by repeated measures across a suitable time scale (which will vary for different molecular, physiological, and structural responses) using an animal model of respiratory distress, where the transcriptome-based model demonstrated the central importance of oxygen in the response. Vo_{2max}, maximum O_2 uptake; ARACNE, Algorithm for the Reconstruction of Accurate Cellular Networks.



Graphical representation highlighting putative regulatory associations (significant correlation between two factors is shown as a dashed line) that likely represent robust interactions, based on high mutual information values. The focus is on central metabolism pathways [i.e., glycolysis, TCA, and oxidative phosphorylation (OXPHOS), respectively] and their immediate neighbors. The gray boxes define functional enrichment of the different bioenergetic compartments based on direct neighbors. Individual genes of relevance are grouped into modules with others of related function, as are physiological readouts that may be treated in a similar manner for statistical analysis. C1–C5: the different complexes in the electron transport chain. The value of such an approach is in providing a detailed overview of a complex interaction network, reducing the huge number of potential factors into groups of defined function, and offering a limited number of candidates whose utility as biomarkers or therapeutic targets may be experimentally verified. KEGG, Kyoto Encyclopedia of Genes and Genomes; GOBP, Gene Ontology Biological Process; GPBP, Goodpasture antigen-binding protein; GOMF, Gene Ontology Molecular Function; Pa_{O2}, arterial oxygen tension; SIRT, sirtuin; HDAC, histone deacetylase.



A higher resolution representation of Fig. 5, highlighting the most significant gene interactions between components in the four inferred modules, is shown. Lines represent factor interactions based on mutual information (blue represents temporal repression, red represents temporal induction). Genes are color coded for broad functional categories (red, cytokines; blue, epigenetic modifiers; green, aerobic metabolism; purple, muscle differentiation; yellow, cell interaction).

A "systems medicine" approach to the study of non-alcoholic fatty liver disease.

Petta S, Valenti L, Bugianesi E, Targher G, Bellentani S, Bonino F; Special Interest Group on Personalised Hepatology of the Italian Association for the Study of the Liver (AISF).; Special Interest Group on Personalised Hepatology of Italian Association for Study of Liver AISF..

Dig Liver Dis. 2016 Mar;48(3):333-42.

Abstract

The prevalence of fatty liver (steatosis) in the general population is rapidly increasing worldwide. The progress of knowledge in the physiopathology of fatty liver is based on the systems biology approach to studying the complex interactions among different physiological systems. Similarly, translational and clinical research should address the complex interplay between these systems impacting on fatty liver. The clinical needs drive the applications of systems medicine to re-define clinical phenotypes, assessing the multiple nature of disease susceptibility and progression (e.g. the definition of risk, prognosis, diagnosis criteria, and new endpoints of clinical trials). Based on this premise and in light of recent findings, the complex mechanisms involved in the pathology of fatty liver and their impact on the short- and long-term clinical outcomes of cardiovascular, metabolic liver diseases associated with steatosis are presented in this review using a new "systems medicine" approach. A new data set is proposed for studying the impairments of different physiological systems that have an impact on fatty liver in different subsets of subjects and patients.