The use of network analyses for elucidating mechanisms in cardiovascular disease†

Diego Diez,[a] Åsa M. Wheelock,[abc] Susumu Goto,[a] Jesper Z. Haeggström,[d] Gabrielle Paulsson-Berne,[e] Göran K. Hansson,[e] Ulf Hedin,[f] Anders Gabrielsen[e] and Craig E. Wheelock*[ad]

Received 19th June 2009, Accepted 28th August 2009
First published as an Advance Article on the web 16th October 2009
DOI: 10.1039/b912078e

Systems biology offers the potential to provide new insights into our understanding of the pathogenesis of complex diseases such as atherosclerosis. It seeks to comprehend the system properties of the non-linear interactions of the multiple biomolecular components that characterize a living organism. An important component of this research approach is identifying the biological networks that connect the differing elements of a system and in the process describe the characteristics that define a shift in equilibrium from a healthy to a diseased state. The utility of this method becomes clear when applied to multifactorial diseases with complex etiologies such as inflammatory-related diseases, herein exemplified by cardiovascular disease. In this study, the application of network theory to systems biology is described in detail and an example is provided using data from a clinical biobank database of carotid endarterectomies from the Karolinska University Hospital (Biobank of Karolinska Endarterectomies, BiKE). Data from 47 microarrays were examined using a combination of *Bioconductor* modules and the *Cytoscape* resource with several associated plugins to analyze the transcriptomics data and create a combined gene association and correlation network of atherosclerosis. The methodology and workflow are described in detail, with a total of 43 genes found to be differentially expressed on a gender-specific basis, of which 15 were not directly linked to the sex chromosomes. In particular, the APOC1 gene was 2.1-fold down-regulated in plaques in women relative to men and was selected for further analysis based upon a purported role in cardiovascular disease. The resulting network was identified as a scale-free network that contained specific sub-networks related to immune function and lipid biosynthesis. These sub-networks link atherosclerotic-related genes to other genes that may not have previously known roles in disease etiology and only evidence small alterations, which are challenging to find by statistical and comparison-based methods. A number of Gene Ontology (GO), BioCarta and KEGG pathways involved in the atherosclerotic process were identified in the constructed sub-network, with 19 GO pathways related to APOC1 of which 'phospholipid efflux' evidenced the strongest association. The utility and functionality of network analysis and the different *Cytoscape* plugins employed are discussed. Lastly, the applications of these methods to cardiovascular disease are discussed with focus on the current limitations and future visions of this emerging field.

[a] *Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan*
[b] *Lung Research Lab L4:01, Respiratory Medicine Unit, Department of Medicine, Karolinska Institutet, 171 76, Stockholm, Sweden*
[c] *Karolinska Biomics Center Z5:02, Karolinska University Hospital, SE-171 76, Stockholm, Sweden*
[d] *Department of Medical Biochemistry and Biophysics, Division of Physiological Chemistry II, Karolinska Institutet, SE-171 77, Stockholm, Sweden. E-mail: craig.wheelock@ki.se; Fax: +46-8-736-0439; Tel: +46-8-5248-7630*
[e] *Center for Molecular Medicine, L8:03, Karolinska University Hospital Solna and Department of Medicine, Karolinska Institutet, SE-171 76, Stockholm, Sweden*
[f] *Center for Molecular Medicine and Department of Molecular Medicine and Surgery, Karolinska Institutet, SE-171 76, Stockholm, Sweden*

† Electronic supplementary information (ESI) available: extended information on a number of different components of the results, including the OPLS analysis, lists of DE genes and probes sets, manual curation of the literature associations and supplementary figures from the *ClueGO* analyses section. See DOI: 10.1039/b912078e

Introduction

Systems biology approaches to investigating cardiovascular disease

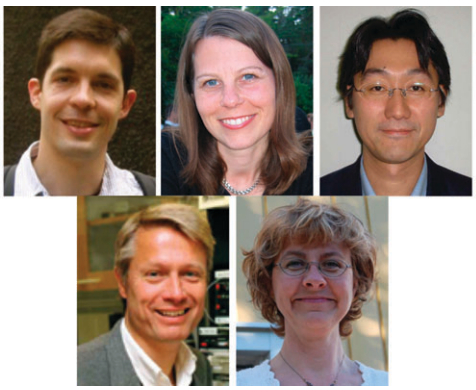
Systems biology seeks to understand how system properties emerge from the non-linear interactions of multiple components.[1–3] The connections and interactions between individual constituents including genes, proteins, and metabolites are examined at the level of the cell, tissue, and organ to ultimately describe the entire organism or system.[1,4–6] The intent is to identify the biological networks that connect the differing system elements, thereby defining the characteristics that describe the overall system.[4] This information can then be used to derive mechanistic information on biological processes as well as identify potential target sites for therapeutic intervention.[7–9] The utility of this approach becomes clear when applied to multifactorial diseases with complex etiologies.

Cardiovascular disease, diabetes, metabolic syndrome, chronic obstructive pulmonary disease (COPD) and asthma all involve complicated etiologies that resist efforts to identify a single gene or pathway responsible for disease onset and progression. Much effort has been made to develop biological networks to describe cardiovascular disease;[10–19] however, any comprehensive model must account for the variety of systemic influences on the disease including lifestyle, diet, body mass index, (epi)genetics, hypertension, dyslipidemia, inflammation and environment. The current research paradigm addresses these individual risk factors in isolation, even though they are known to concomitantly contribute to disease pathogenesis. This problem is further confounded by the fact that discrete biological functions can only rarely be attributed to an individual molecule, and that small defects in many genes rather than large defects in a few genes are most likely responsible for the observed pathology.[20–24] Accordingly, an integrative systems approach involving investigations of the corresponding biological networks is required to address the complex issues of these multifactorial pathologies. However, current network approaches focus on the system properties of individual sub-systems (*e.g.*, the gene regulatory network, or the protein–protein interaction network) and integration is a challenge that requires understanding how the elements in one network affect those in other networks. This review provides an overview of network theory and the computer-assisted generation of biological networks, then presents an example of an atherosclerosis-specific biological network generated from microarray data using the program *Cytoscape* and associated plugins. The applications and utility of these plugins are presented followed by a discussion on the future directions of this research approach.

Network biology

What is a network?

The accumulation of large amounts of biological data from omics projects is providing the foundation for the development of systems biology. Accordingly, the new challenge is to combine information from multiple high-throughput experiments involving multiple platforms and formats and extract the

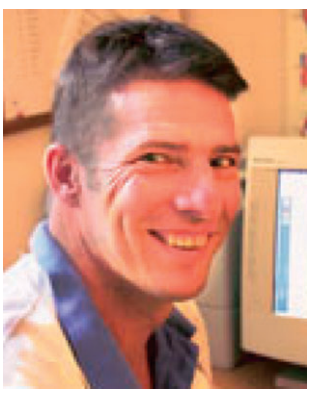


Diego Diez, Åsa M. Wheelock, Susumu Goto, Jesper Z. Haeggström and Gabrielle Paulsson-Berne

Dr Diego Diez is a postdoctoral researcher at the Kyoto University Bioinformatics Center working on applying systems biology approaches to cardiovascular disease.
Assistant Professor Åsa M. Wheelock heads a research group at the Karolinska Institute that investigates pneumotoxicants and inflammatory lung diseases, as well as quantitative intact proteomics.
Associate Professor Susumu Goto is interested in the development of databases for molecular interaction networks and network analysis using the KEGG database suite. His work also involves in silico *metabolic reconstruction.*
Professor Jesper Z. Haeggström heads a research group at the Karolinska Institutet that examines the role of bioactive lipid mediators in inflammatory disease.
Associate Professor Gabrielle Paulsson-Berne is a molecular biologist and member of the steering group for the BiKE project. Her main research areas include: 1. Identify link(s) between local and systemic inflammatory response in patients with unstable plaque. 2. Elucidate the role of lipid-induced immune processes linked to atherosclerosis, specifically activation of CD1d restricted NKT cells.

Göran K. Hansson

Göran K. Hansson is Professor of Cardiovascular Research at the Karolinska Institute and head of the Cardiovascular Research Laboratory in the Center for Molecular Medicine at Karolinska University Hospital in Stockholm, Sweden. His research deals mainly with proatherogenic and atheroprotective immunity and the role of inflammation in cardiovascular disease.



Ulf Hedin

Professor Ulf Hedin holds the academic chair in vascular surgery at the Karolinska University Hospital and directs a translational research group at the Center for Molecular Medicine. The research is focused on identifying cellular and molecular pathways in plaque instability and healing processes in the vessel wall. As a part of this work, the group has together with other investigators at CMM established a biobank for carotid endarterectomies (BiKE) containing carotid plaque gene arrays, histology and clinical data from 500 patients.

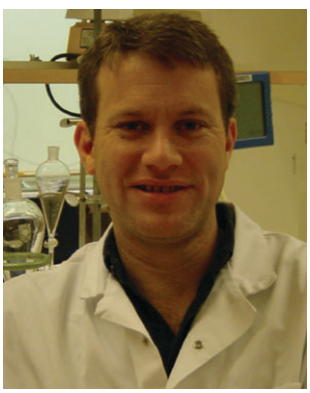
relevant system properties.[25–27] A common approach to the visualization and examination of omics data involves the generation of a network of all the individual components of a given set of experiments.[28–30] These approaches are not novel and methods for analyzing systems and networks have already been developed in other fields, for example social and information networks.[31,32] Network theory is widely used to analyze and visualize systems level relationships without losing detailed relations between components of the system.[33] Network theory, or more generally graph theory, is a branch of mathematics devoted to the study of networks (graphs), which are mathematical structures used to model pairwise relations between objects from a "collection".[34–36] This representation is suitable for different kinds of complex data including technological networks (*e.g.*, the Internet, GPS, WiFi, *etc.*)[37,38] and social networks (*e.g.*, Facebook, MySpace, WoW, *etc.*)[39,40] as well as biological networks.[4,41] In a biological context, a collection could be the proteome of a cell and the relations are defined by their interactions.[42] A network is usually represented as a set of nodes, connected to each other with links or edges (Fig. 1). A node represents an element of the collection and the edge connecting two nodes represents the relation (*e.g.*, the node is a protein and the edge connecting two proteins is their interaction). This relation can be either symmetric or asymmetric, depending upon if the relation in one direction implies a relation in the opposite direction. For example, a protein–protein interaction (PPI) network is a symmetric network where nodes represent proteins and edges represent interactions between them (Fig. 2C).[43] Conversely, a gene regulatory network (GRN) is an asymmetric network, where nodes represent genes and edges the relationships between genes (*e.g.*, "gene A activates gene B" or "gene D represses gene C"; Fig. 1, directed network).[41]

Although network theory was developed for mathematical applications, the use of network representation is widespread in molecular biology and biochemistry to represent cellular signaling and metabolic pathways.[44,45] For example, the MAPK pathway is a signal transduction pathway that couples growth factor binding to plasma membrane receptors to changes in gene expression that control cell proliferation, differentiation and survival.[46] The signal is transduced to the nucleus by tyrosine and serine/threonine kinases (Raf1, MEK and ERK, also known as MAPKKK, MAPKK and MAPK, respectively) that ultimately activate transcription factors, which regulate the expression of target genes (Fig. 2). Ras proteins play a major role in the regulation of the pathway's homeostasis by alternating between active GTP-bound and inactive GDP-bound states. Ras must be in its active form in order to interact with downstream effector proteins that transduce the signal. Guanine nucleotide exchange factors (GEFs; *e.g.*, SOS, RasGRF, RasGRP) and GTPase activator proteins (GAPs; *e.g.*, p120GAP, NF1, Gap1m) regulate the activity of Ras, and hence, modulate the entire pathway.

These protein interactions and dependencies can be represented in a number of ways, as illustrated in Fig. 2. The classical representation of cellular signaling pathways uses circles and boxes to symbolize proteins (*e.g.*, PKA, PKC, hemoglobin) and metabolites (*e.g.*, ATP, DAG).[47] Interacting proteins are drawn in proximity to their partners, and proteins affecting the activity of other proteins are indicated with directional arrows (Fig. 2A). Although not a network *sensu stricto*, this depiction resembles the representation used in network theory. In a more advanced representation, the arrows represent interactions detailing the type of effect they have on the partner. For example, signaling and metabolic pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) are constructed in this fashion (Fig. 2B), enabling some automatic manipulation exemplified with the software *KegArray*[48] and others tools (*e.g.*, the *Bioconductor* packages *KEGGSOAP*[49] and *keggorth*[50]). While this representation is useful for providing an overview of all existing interactions, it is quite limited. KEGG pathways appear to be networks, but are actually quasi-static images with restricted flexibility (however, a new *Bioconductor* package called *KEGGgraph* is capable of converting KGML files representing KEGG pathways into a network structure[51]). The pathways are snapshots of the state of the system in a hypothetical controlled situation and do not contain any quantitative information regarding the interactions and activities of the elements involved. Furthermore, the pathways do not indicate the

Anders Gabrielsen

Anders Gabrielsen, MD, PhD, is a cardiologist with an interest in measuring gene expression in cardiovascular disease to elucidate possible mechanisms of disease development, and integrated applications of system biology and pathway analysis.



Craig E. Wheelock

Associate Professor Craig E. Wheelock heads a research group at the Karolinska Institute that examines the role of bioactive lipid mediators in inflammatory diseases, with a focus on cardiovascular and pulmonary diseases. His group is developing lipidomics and metabolomics methods for applications in investigating multiple inflammatory disorders. He is also broadly interested in the development of bioinformatics tools for probing inflammatory diseases at the systems level.

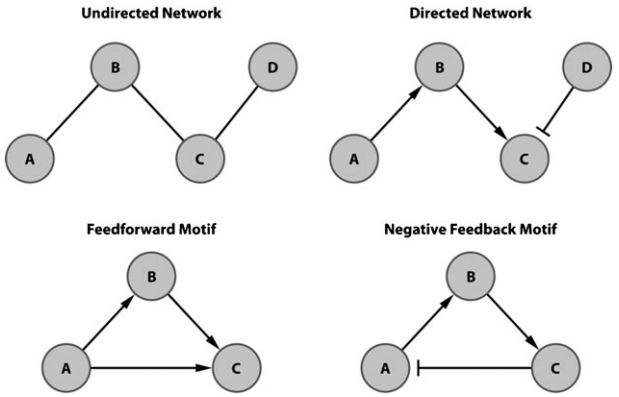


Fig. 1 The upper panel shows four nodes involved in two different networks. The left network is undirected or symmetric; all connected nodes have bidirectional relations. Protein–protein interaction networks are an example of symmetric networks. If protein A interacts with protein B, then protein B also interacts with protein A. The directed network in the right is asymmetric (the arrows indicate activation and the "⊥" indicates inhibition). Regulatory networks are asymmetric; in this example if the network was a gene regulatory network, gene A activates gene B, but gene B does not activate gene A. Gene C is activated by gene B and inactivated by D. The lower panel shows two different network motifs found in biological networks. In the feedforward motif, gene A activates gene B, which activates gene C. Likewise, gene A activates gene C directly. The negative feedback motif shows an example of autoregulation. Gene A activates gene B which activates gene C. Gene C inhibits its own expression through the inhibition of A.

relative importance of the different elements of the network. Network theory places this kind of representation into a framework where the biologically significant key components of the system can be identified. For example, if a network of the interacting proteins found in the KEGG MAPK pathway is constructed, some highly connected proteins are observed (*e.g.*, Ras, Grb2), whereas other proteins are only sparsely connected (Fig. 2C). An analysis of this network's connectivity shows that Ras is a hub node (further discussed below), suggesting that Ras has an important role in the regulation and stability of this network and the subsequent processes that the network regulates. This finding has been thoroughly demonstrated by years of experimental analysis and the fact that Ras is mutated and constitutively activated in ~20% of cancers.[52,53]

Network properties

An important component of network analysis is the generation of meaningful graphical output. Data visualization using network representation is not new to biology and has been used for years to represent elements and relations in metabolic and signaling pathways (*e.g.*, KEGG[54,55]), GRNs, *etc.* However, instead of a simple representation tool, network theory provides a framework for the quantitative representation of the general properties of the system.[56–59] The network characteristics are defined by a series of topological parameters that summarize the behavior and the importance of specific nodes, as well as the entire network.[60] An understanding of this terminology and its application to deciphering a network is vital for interpreting the results of systems biology studies.[61] First of all, it should be stressed that a biological network is a quantifiable structure, enabling direct comparisons of different networks.[4] The primary components of a network are nodes and the edges that define their relationships. The concept of a node comes from graph theory, where a node (also called a vertex) is the fundamental unit from which graphs are formed. The definition of a node is a point on which the graph is defined, which may in turn be connected by graph edges or links.[34,35] The most fundamental characteristic of a node is its degree, which is defined as the number of edges incident to the node (*i.e.*, how many links a given node has to other nodes in the network, denoted as the variable k). For example, according to the undirected network in Fig. 1, node A has degree $k = 1$ and node B has degree $k = 2$. The degree distribution is the probability that a node has a specific number of links (k) over the entire network and is given by the function $P(k)$. The node degree distribution can be used to distinguish between scale-free and random networks (discussed below).[4] The relative importance of any given node within a network is determined by the centrality (*i.e.*, how important is a specific gene within a disease network), which can be described in terms of the degree, betweenness and closeness centrality as well as other centrality terms.[62] The degree centrality is the number of edges that a node possesses or the number of links incident upon a node (sometimes a normalized degree centrality is used, in which the number of edges that a node possesses is divided by the total number of edges minus one). The betweenness centrality is the fraction of shortest paths (discussed below), counted over all pairs of nodes, that pass through that node, and reflects the amount of control that this node exerts over the interactions of other nodes in the network, with nodes that occur on many shortest paths between other nodes having higher betweenness.[63,64] In other words, nodes that participate in denser sub-networks evidence a greater betweenness centrality. Closeness centrality is a measure of how fast information spreads from a given node to other reachable nodes in the network and is defined as the inverse of the mean geodesic distance (*i.e.*, the shortest path) between a node and all other reachable nodes. An example of the importance of centrality was demonstrated by Jeong *et al.*,[65] who showed that in a *Saccharomyces cerevisiae* protein network, highly connected proteins with a central role in the network's architecture were three times more likely to be essential than proteins with only a small number of links to other proteins.

There are a number of distances that can be measured in a network, including the network diameter, radius and path length, all of which give important information on the network. For a given path, the path length provides the number of links between two nodes in the network of which there are usually multiple path alternatives. The shortest path is the one with the smallest number of links connecting two selected nodes. The mean path length is the average over the shortest path of all nodes in the network and is a measure of the general navigability of the network. There are multiple algorithms for calculating distance measures, which can profoundly affect the outcome of the analysis. Accordingly, this choice should be based upon what is appropriate for the

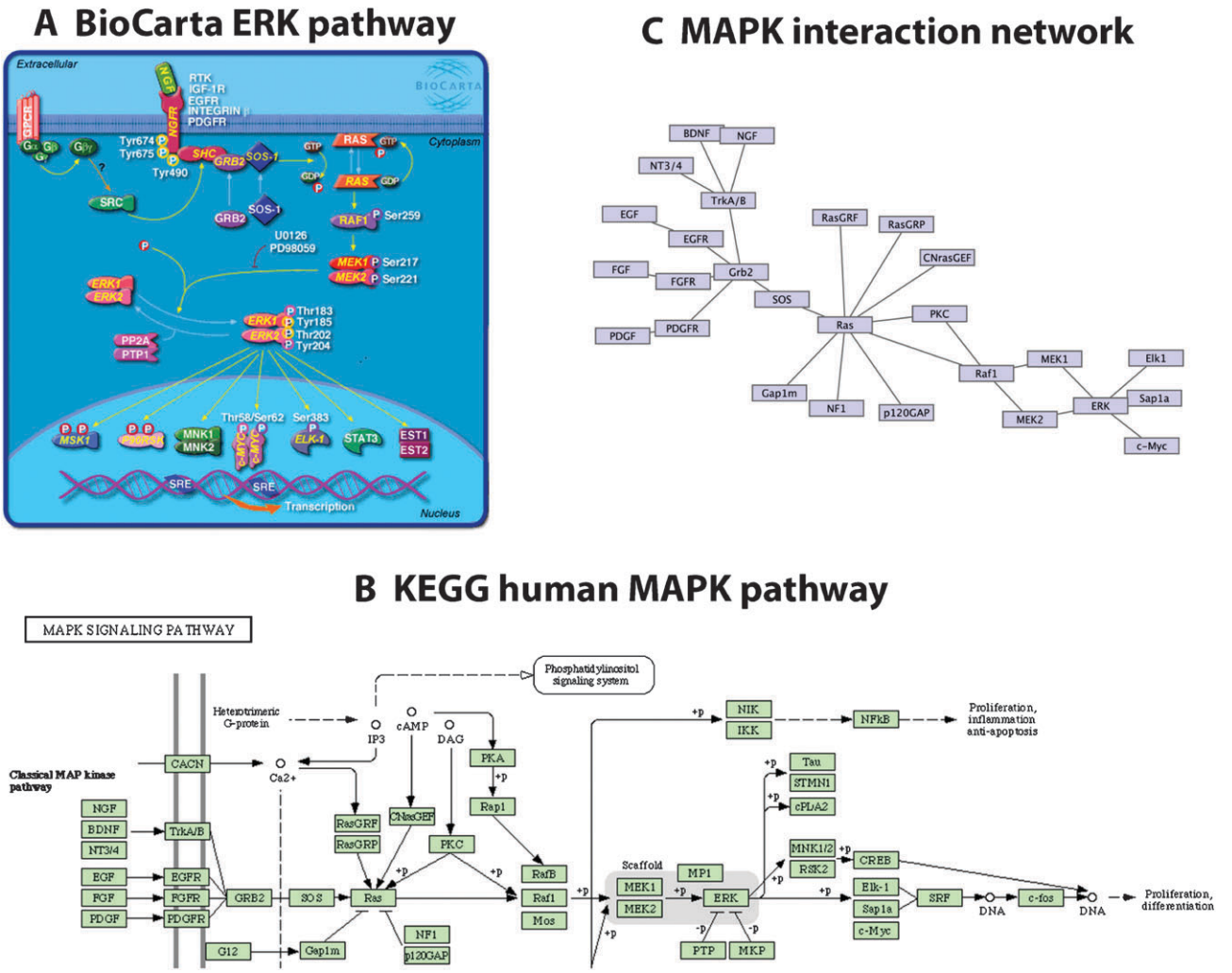


Fig. 2 (A) BioCarta representation of the MAPK pathway. Proteins (*e.g.*, Grb2, SOS, Ras) and metabolites (*e.g.*, GTP, GDP, phosphate) are described with globular shapes. Interacting proteins are drawn close to each other. Chemical activities like phosphorylation are indicated with arrows. (B) KEGG representation of the MAPK pathway in humans. Proteins are described as squares and metabolites (*e.g.*, Ca^{2+}, cAMP, DAG) as circles. Interactions are described as arrows, and special activities like phosphorylation are indicated with "$+p$" over the corresponding arrow. These images can be manipulated. For example, proteins absent (or not yet annotated) in specific organisms are shown as clear squares instead of green. (C) Protein–protein interaction network derived from the KEGG map. Ras is the most highly connected node with 9 edges, suggesting that it is a hub in the interaction network and therefore most likely has an important role in the functionality of the network.

biological question of interest.[36] Finally, the clustering coefficient provides information on how a specific node is included in a densely connected sub-network. If a node has a high clustering coefficient, many nodes that have connections to that node also have connections to each other. This value is then averaged over the whole network and is a measure of compactness and modularity. These different network characteristics and properties enable the quantitative analysis and comparison of multiple complex network structures and provide a means for estimating the importance of any node within a given network. Accordingly, there are multiple parameters that describe node/network behavior and an understanding of their meaning and relative importance will greatly assist the reader in deciphering network analyses.

Biological networks were initially studied in model organisms such as bacteria, yeast and nematodes,[56,66–71] from which important general properties have been derived. These analyses are now gradually being applied to the understanding of complex human diseases, including inflammation and cardiovascular disease.[10,72–75] For example, one important aspect derived from the analysis of biological networks is that the degree distribution follows a power-law, which is any polynomial relationship that exhibits the property of scale invariance where $P(k) \approx k^{-\gamma}$.[76] One of the main features of a power-law is the scale invariance such that scaling by a constant simply multiplies the original power-law relation by the constant.[77] Accordingly, a log–log plot of the connectivity distribution will give a linear slope of $-\gamma$ (where γ is the degree exponent). The value of γ gives information on the type of network model, with $2 < \gamma < 3$ commonly observed for most biological as well as non-biological networks (these values of γ indicate that the network is ultra-small[78,79]).[4] As a consequence,

these networks maintain the same structure over different scales and are accordingly called scale-free networks.[4,80–82] The scale-free property is common in other networks derived from complex systems, such as social networks or the Internet.[83–85] In a scale-free network the majority of nodes have few edges, whereas some nodes (termed hubs or nexus nodes) are highly connected and contain many edges. In biological terms, a hub node is a protein, gene or metabolite that is connected to many other proteins, genes or metabolites. In a scale-free network, the ratio of hub nodes to nodes in the rest of the network remains constant as the network changes in size. Hub nodes are of particular interest as they are potentially involved in critical regulatory processes that maintain the structure of the network, and therefore, the homeostasis of the system. For example, in Fig. 2C the small protein Ras appears as a hub node, in concordance with the fact that this protein has a key role in the regulation of the signaling processes that lead to cell differentiation and proliferation. An alteration in this function, through for example a point mutation, modifies the regulatory network, leading to cancer.[52] Another characteristic of scale-free networks is the high level of redundancy, where different paths can result in the same outcome. This quality combined with the fact that most nodes have only a few edges renders these networks very robust to random elimination of nodes (*e.g.*, these networks evidence increased stability).[81,86,87]

Although network analysis can help explain the behavior of the system as a whole, the importance of individual elements is not lost in this global view. The study of biological networks shows that complex networks are constructed of recurrent simple motifs.[67,68,88] For example, a feedforward motif is composed of three genes A, B and C (Fig. 1). In one of its simplest forms, gene A interacts with gene B, which interacts with gene C. Gene A also interacts with gene C, thus augmenting the signal on gene C. For example, a transcription factor X regulates the expression of another transcription factor Y, and both control the expression of a third gene Z. An example of this is the L-arabinose system in *Escherichia coli*, where the transcription factor *crp* regulates the expression of transcription factor *araC* and these two transcription factors control the expression of the operon *araBAD*.[89] Another example is that of negative and positive feedback motifs. In a negative feedback motif, gene A activates gene B, which activates gene C, while gene C on the other hand inhibits gene A, thus autoregulating its own expression (Fig. 1). For example, in the hypothalamic–pituitary–thyroid axis, thyroid hormone autoregulates its own synthesis using a negative feedback mechanism, by negatively regulating the secretion of TRH (thyrotropin releasing hormone) and TSH (thyroid stimulating hormone), two hormones that regulate the secretion of thyroid hormone.[90] Initially described in simple bacteria, these motifs are also found in the regulatory networks of higher eukaryotes and are fundamental to understanding the behavior of complex networks, including biological networks.

Modularity is a key concept in biology that assumes that cellular functionality can be divided into independent self-autonomous modules.[91] These modules can perform functions without being affected by external components. Although widely recognized in biology, modularity directly conflicts with scale-free networks. This is because in a scale-free topology, a few nodes contain many links implying that they participate in a lot of interactions in the network, thus helping to integrate the information across the network. This situation, however, explicitly prohibits the existence of separated modules. This dichotomy is solved by the definition of a higher network topology structure, termed a hierarchical network.[56,92] The existence of hierarchical networks was derived by the realization that the metabolic networks of several organisms, including all three domains of life, have a cluster coefficient that is two times larger than that expected for a scale-free network of the same size.[66] The proper understanding of how modularity is integrated into the network topology is an important step in understanding how biological networks are organized.

Our understanding of specific biological networks is increased rapidly as more data accumulate. For example, the advancements in identification of *cis*- and *trans*-regulatory elements are pushing the development of models to disentangle the transcriptional regulatory networks of genes.[93] Using such models, when sufficient resolution is achieved, we could predict the expression level of a gene under a given condition. Moreover, disagreement with expected expression levels might be an indicator of unknown regulatory processes, helping to better understand the structure of the real network.[94] Finally, the mathematical models used to generate the network itself can be used to predict the behavior of the network when specific elements are altered. For example, what are the effects if a specific node of a GRN is removed by a knockout mutation? How does this change affect the global stability and robustness of the network, and eventually, the phenotype of the studied system? How would a drug targeted to a specific protein product of the network affect disease phenotype as well as other associated up- or downstream processes?[95] Systems biology seeks to answer these and other questions by modeling the relationship between the individual components. Accordingly, network analysis can aid in elucidating if therapeutic intervention shifts the individual directly from a disease to healthy state or whether the individual goes through a novel pharmacological state before returning to equilibrium (*e.g.*, healthy). This type of information would increase our understanding of the concept of a "healthy" individual and provide significant insight into disease and resolution processes.

Network analysis for the study of atherosclerosis

The analysis of biological networks has enabled the discovery of the general properties of biological systems, including the scale-free structure of biological networks and the hierarchical structure of modular networks.[4] However, network analysis can also be used to extract new biological information such as identifying unknown modules (based upon the clustering coefficient or other network parameters) or finding previously unknown associations between elements in large-scale analysis. A number of research groups have successfully used this approach to examine networks in cardiovascular disease,[10–18] which has been extensively reviewed elsewhere.[96–98] To serve as an example of the utility of network analysis in examining

these large datasets, we performed analyses on a microarray study combining literature mining and correlation analysis to provide information on biological associations in cardiovascular disease.

This approach was employed to analyze atherosclerotic plaques in order to identify gender-specific relationships that correlated with gene transcripts. Clinical samples from a biobank database of endarterectomies at the Karolinska Institute (Biobank of Karolinska Endarterectomies, BiKE) were utilized in the analysis. BiKE contains samples from patients who have undergone stroke-preventive carotid endarterectomy for asymptomatic or symptomatic carotid stenosis. Microarray analyses on global gene expression patterns were performed previously with the individual patient endarterectomies using Affymetrix GeneChip® Technology with Affymetrix platform HGU133-plus2. This platform consists of 54 675 probe sets corresponding to approximately 20 326 known human genes. A subset of the BiKE microarray data was utilized for the network analyses (n = 47; females, n = 8; males, n = 39). Probe intensities were background corrected, normalized and summarized using the robust multichip average (RMA) method.[99] Statistical analyses were performed probe set-wise using the *limma* package (linear models for microarray data) for *Bioconductor*.[100] The *limma* package fits linear models, and can be used to extract information about differential expression between different contrasts. Here, a linear model was fitted for each probe set based on a specified experimental design using the factor "gender" (male/female levels). Subsequently, the gene-wise expression levels were estimated by calculating the average of all probe sets corresponding to the same gene, based on the Entrez Gene annotation, and exported in a format suitable for use with *Cytoscape* (discussed below).

Statistical analysis revealed 43 differentially expressed (DE) genes, with 23 down-regulated and 20 up-regulated in females relative to males (the complete list of DE genes and probes with the corresponding p- and q-values is provided in Tables S1 and S2, respectively).† Of these 43 DE genes, 28 are located on X or Y chromosomes. Although we cannot exclude the possibility that genes linked to sexual chromosomes are related to atherosclerosis, for the purpose of this study we focused on non-sexual chromosome genes. Of the remaining genes, APOC1 was selected for further analysis based upon an observed 2.1-fold lower expression level in females and a reported potential role in atherosclerosis and coronary artery disease.[101,102] APOC1 is the major plasma inhibitor of cholesteryl ester transfer protein (CETP), inhibits lipoprotein binding to the LDL and VLDL receptors, and appears to interfere directly with fatty acid uptake,[102,103] suggesting that further investigation into its biological activity is warranted. Because atherosclerosis is more prevalent in males than females of the same age, sample availability is highly unbalanced reflecting the clinical reality of working with this disease. To estimate the effect of the unbalance on these results, a bootstrapping approach was performed in which all 8 female samples were compared against 8 male samples randomly selected from the total pool of males (n = 39), using the same statistical model described above. The process was repeated 1000 times, and each male sample was selected on average 228 ± 12 times in any experiment. The bootstrap analysis indicated that APOC1 was significantly DE on average 937.5 ± 17 out of the 1000 iterations (925 times for probe set 204 416_x_at and 950 times for probe set 213 553_x_at), indicating that the results are robust in spite of the discrepancy between the number of female and male subjects included in the study. The robustness against differences in group size was further confirmed with both supervised (orthogonal partial least square of latent structure [OPLS]) and unsupervised (principal component analysis [PCA]) multivariate statistics using SIMCA-P$^+$ (Umetrics AB, Umeå, Sweden). The OPLS analysis did, however, reveal four weak outliers among the male subjects (Fig. S1, ESI†). These four subjects were over-represented among the bootstrap iterations where APOC1 failed to produce a significantly DE between men and women. Furthermore, exclusion of these four subjects resulted in a drastic improvement in the separation between genders in the OPLS analysis (Fig. S1, ESI†). Overall, the OPLS analysis confirmed the importance of the two APOC1 probe sets in driving the separation between genders, an effect that was further pronounced following exclusion of the four outliers. This observation is important for application in studies where the sample sets are significantly unbalanced.

In omics studies, the high false positive rate resulting from multiple univariate testing is often adjusted using p-value correction (*e.g.*, the false discovery rate [FDR] method of Benjamini and Hochberg[104]). However, the risk of detecting false positives and the ability to detect true positives (*i.e.*, statistical power) are inversely related. Even so, the drastic decrease in statistical power that is inevitable when a p-value correction is utilized on datasets with a large number of variables, *e.g.* microarray data, is seldom discussed. To address the issue in this study, we compared the statistical

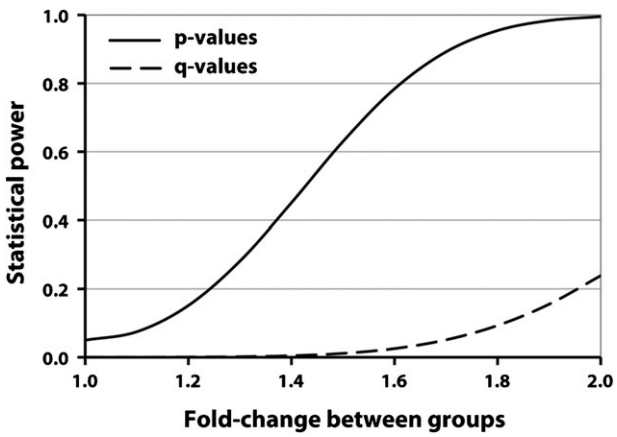


Fig. 3 The statistical power to detect true positives (*i.e.*, differential expression between men and women in an atherosclerotic plaque) at the respective fold-change level is displayed. The graphs are based on a 0.05 significance level (solid line) or false discovery rate (FDR, dashed line), and a variance level corresponding to the coefficient of variance (CV) of the 84th percentile (mean + 2S.D.). Using the conventional significance level of 0.05 (p-value) resulted in a 95% statistical power to detect a 1.8-fold increase in expression at the probe level. Controlling the FDR by means of Benjamini and Hochberg[104] method resulted in a drop of the statistical power to 9.3%.

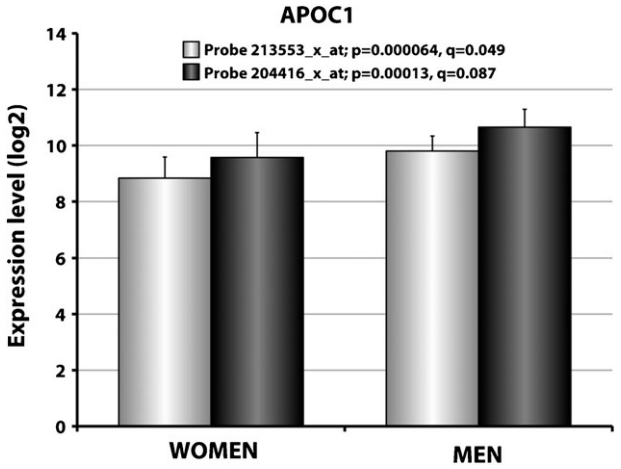


Fig. 4 Bar graph showing the expression levels (mean + S.D.) of the two probe sets representing the APOC1 gene (213 553_x_at; white bars, 204 416_x_at; black bars). When using a *p*-value, both probe sets are defined as statistically significantly altered between genders ($p < 0.001$). In contrast when FDR is applied, only one of the probes is defined as significantly altered ($q = 0.049$ and 0.087, respectively). This discrepancy is likely to be a consequence of the low statistical power resulting from *p*-value correction of datasets consisting of a very large number of variables.

power when using both *p*-values and *q*-values for the current data (Fig. 3). Conventional statistical methods ($p < 0.05$) resulted in a statistical power of 0.95 to detect a 1.8-fold increase in expression levels. Application of the *p*-value correction method of Benjamini and Hochberg[104] with a FDR of 5% ($q < 0.05$) resulted in a decrease in the statistical power to 0.093. Furthermore, use of the FDR approach decreased the number of significant changes from 4304 to 72. While the 72 probe sets found to be significantly altered using the FDR approach have a 95% probability of being true positives, an additional 700 true positives evade the discovery process due to the lower power. For example, the use of FDR resulted in one of the two probe sets representing the APOC1 gene falling outside of the significance cutoff (Fig. 4), even though its relevance for gender differences was confirmed by the bootstrap and OPLS analyses (Fig. S1, ESI†). In contrast, the conventional *p*-value approach by definition results in 3734 false positives (5% of 54 675 probe sets). Accordingly, 87% of the 4304 probe sets found to be significantly altered when using the *p*-value approach can be expected to be false positives. Both approaches indicate that the number of true positives, here defined as probe sets differentially expressed between men and women in atherosclerotic plaques, is in the range of 570–770. The appropriate method largely depends on whether the goal of the study is to achieve a high stringency in terms of avoiding false positives, or a high statistical power to detect true positives. In contrast, you do not have to "mind your *p*'s and *q*'s" in network analysis or multivariate approaches, since the encompassing nature of these unbiased approaches does not require any pre-selection based on significance level. The fact that all variables are considered in extracting global trends from the data represents one of the main strengths of network and multivariate statistical analyses.

Software tools employed in network analysis

A plethora of tools designed for the construction and analysis of networks have been developed in the last few years and it is not the intention of the current review to describe all of them (see Ng *et al.*[105] and Bauer-Mehren *et al.*[106] for a comprehensive list). Instead, we will focus on one popular tool and the accompanying plugins to demonstrate how network analysis can be applied to the study of cardiovascular disease. *Cytoscape* is free software developed for the visualization, manipulation and analysis of biological networks. It is available for most common computer platforms, is easy to use, and comes with extensive documentation.[107–110] The main strength of this software is the ability to extend its functionality through the addition of plugins, which can be used for diverse tasks ranging from network inference to network analysis and visualization (Table 1).

The study was initiated by loading the expression data into *Cytoscape* using a tabular format, and the *ExpressionCorrelation* plugin[111] was then used to compute the correlation network. This plugin enables the construction of a network from microarray data, by computing the Spearman correlation coefficient for all pairwise comparisons. The plugin visualizes a histogram of all calculated correlations, and a lower and upper cutoff can be selected (*e.g.*, a correlation cutoff of −0.9/0.9 was used in this analysis), after which a network is generated from all the nodes that meet the specified criteria. This functionality enables the identification of correlated groups of genes or modules. Next, an association network was constructed from a list of DE genes (see Table S1 (ESI†) for a list of the genes used) using the *AgilentLiteratureSearch* plugin,[112] which is a user-friendly literature mining tool. This software accepts as input a list of gene symbols and associated aliases, and then performs a search for each symbol in several databases, including Pubmed (by default), OMIM (Online Mendelian Inheritance in Man, NCBI) and USPTO (The US patents and trademark office). The number of hits per symbol can be defined (default 10), but there is a 1000 record limitation on the number of hits per analysis to avoid overloading search engines. Some context keywords as well as organism limitations can be added to restrict the analysis. The use of relaxed relationships can be specified in order to increase the number of results obtained, although the strength of the subsequently detected relations may be weak. The software searches for association keywords (*e.g.*, "gene A activates gene B" or "gene C is repressed by gene A"), then constructs a network based on these associations. The network can be further extended by querying individual genes for more associations. In this analysis 50 queries were allowed per gene and aliases were used with no context limitations, giving a total of 948 papers to be analyzed. Based upon interest, these papers can be individually queried to examine for direct biological evidence of identified associations. We performed a manual curation for the 24 nodes linked directly to APOC1 (Fig. S2 and Table S3, ESI†). Of these 24 nodes, we found 5 direct associations (20.8%), 14 primary associations (58.3%), two secondary associations (8.3%), and 3 incorrect associations (12.5%). A few clear mismatches were identified, due to incorrect gene name and/or association detection; however, these mismatches

Table 1 *Cytoscape* plugins for applications in network analyses of transcriptomics data[a]

Software	Description
Cytoscape http://www.cytoscape.org	Manipulation of networks
ExpressionCorrelation http://www.baderlab.org/Software/ExpressionCorrelation	Network inference from microarray data
AgilentLiteratureSearch http://www.agilent.com/labs/research/litsearch.html	Text mining
NetworkAnalyzer http://med.bioinf.mpi-inf.mpg.de/netanalyzer	Compute topological parameters
jActiveModules http://www.cytoscape.org/plugins/index.php	Detect modules based on topological parameters
ClueGO http://www.ici.upmc.fr/cluego	Functional enrichment
MONET http://delsol.kaist.ac.kr/~monet/home/index.html	Network inference from microarray data
MCODE http://baderlab.org/Software/MCODE	Network motif identification
Cerebral http://www.pathogenomics.ca/cerebral	Visualization of omics data using cellular compartments
OmicsViz http://metnet.vrac.iastate.edu/MetNet_fcmodeler.htm	Visualization of omics data across species
BiNGO http://www.psb.ugent.be/cbd/papers/BiNGO	Gene ontology enrichment
GOlorize http://www.pasteur.fr/recherche/unites/Biolsys/GOlorize	Coloring of nodes

[a] This list is non-exhaustive and is solely provided to give an example of some of the available resources. See http://www.cytoscape.org/plugins/ for a complete list of available *Cytoscape* plugins.

evidenced very few associations in the overall network. Accordingly, the impact of these errors in the network is very limited. Overall, these results highlight the utility of this approach to discover associations between genes, but demonstrate the need to verify the most important associations to avoid the impact of false positives.

The expression correlation and the association network were then merged into a single undirected network using the set operations available in *Cytoscape* (Fig. 5). Next, the network parameters were computed using *NetworkAnalyzer*,[113] another plugin that computes and displays a comprehensive set of topological parameters, including the degree of nodes and degree distribution, the network diameter, the centrality and clustering coefficient, *etc*. A linear or power-law model can be fitted to examine whether the parameters follow a scale-free topology. The degree distribution for this network was shown to follow a power-law with $R^2 = 0.905$ and $\gamma = 1.6$ indicating that the network behaves like a scale-free network (Fig. 6). All the computed parameters can be mapped to the network for visualization and can be used to detect putative modules. For example, the *jActiveModules* plugin[114] enables the discovery of modules or sub-networks based on specific topological parameters. A sub-network can be created from the detected modules, enabling in-depth analyses of the components. For this analysis, active modules were identified based on the closeness centrality (with default parameters) and a sub-network was generated based on one of the detected modules (Fig. 5, right panel). Lastly, the *BiNGO* plugin[115] was used to determine whether specific sub-networks contained over-represented Gene Ontology (GO) terms. This application enables the identification of functionally distinct modules. The plugin allows the specification of several parameters, including the type of statistical test to perform (default hypergeometric), the multiple testing correction method (default Benjamini and Hochberg[104]), whether to test for over-represented categories (default), under-represented or both, the ontology to use and the organism. In this analysis, the sub-network was analyzed to detect functional categories of GO over-represented as shown in Fig. 7 (selecting human as the organism and otherwise default parameters). *ClueGO*[116] is another tool designed to find (under)over-represented ontology terms and pathways from GO and pathway databases such as KEGG or BioCarta. In addition to the options available in *BiNGO*, further parameters can be selected, such as restricting the analysis to specific evidence codes. *ClueGO* also performs a clustering of the detected ontology terms, and hence, the clustering parameters can be tuned. Both *BiNGO* and *ClueGO* can make use of *GOlorize*,[117] a plugin designed to visualize overlapping GO terms present in the same nodes. These additional plugins were subsequently used to examine the sub-network from Fig. 7 to generate the GO and pathway information in Fig. 8 as well as Fig. S3 and S4 (ESI†).

The sub-network in Fig. 7 is composed of three different clusters (clusters A–C). Cluster A is mainly composed of genes from the associative network, although a few genes correlated to them (either positively or negatively) are included (CAPG, DDR2, C16ORF14 and ITGAX). The composition of cluster A is enriched in GO categories for 'lipid homeostasis' and visual inspection of the node composition shows that many of the genes are relevant to atherosclerosis (LDLR, VLDLR and SOAT1). In particular, a number of apoliprotein genes are present, most noticeable APOC1, which is expected as this gene was used to create the network due to its 2.1-fold down-regulation in females relative to males. Cluster B is composed exclusively of genes found *via* literature mining (associative network), with the majority belonging to the superfamily of small GTPases Ras (GO category 'small GTPase signaling'). Cluster C is composed only of genes

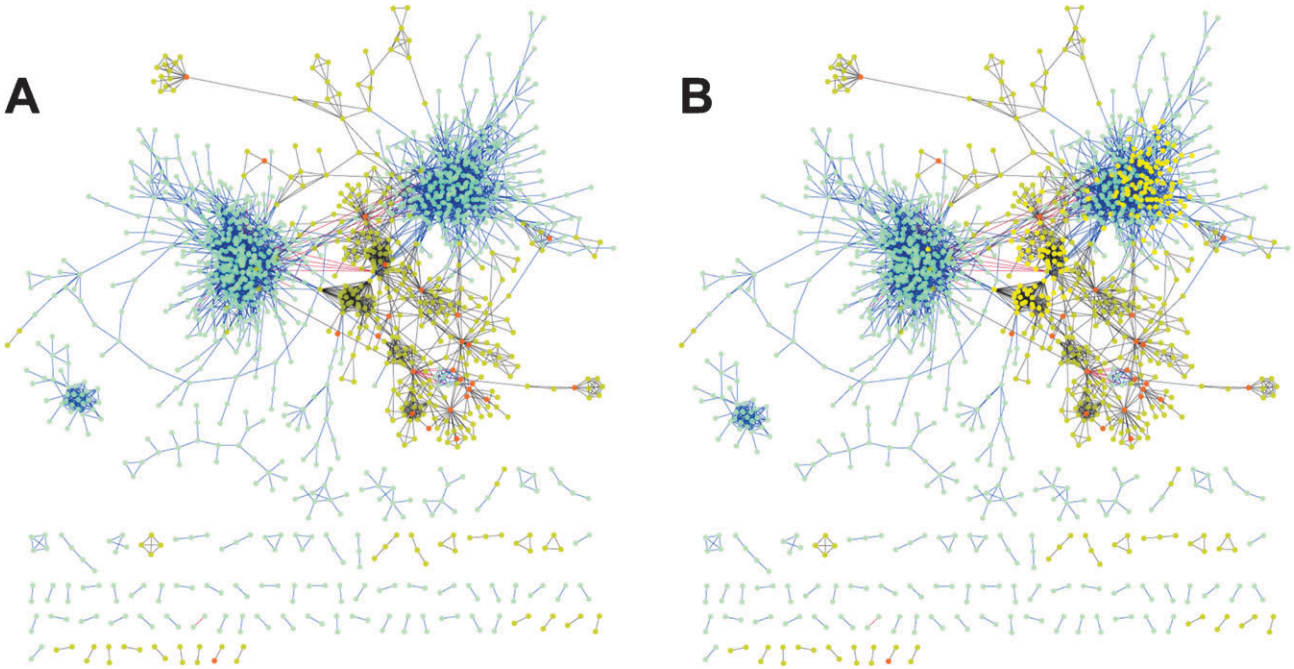


Fig. 5 Gene correlation and association network for gender-dependency in atherosclerosis. In network A, nodes from the expression correlation are colored in cyan. Orange nodes exhibited gender-dependent differential expression (DE). Dark yellow-green nodes were identified *via* the literature mining step in association with the DE nodes. Edges colored in red indicate negative correlations, whereas blue edges show positive correlations. Black edges correspond to relations derived from the literature association alone. Network B shows in bright yellow the nodes detected in one active module by the analysis with *jActiveModules* and is displayed in greater detail in Fig. 7.

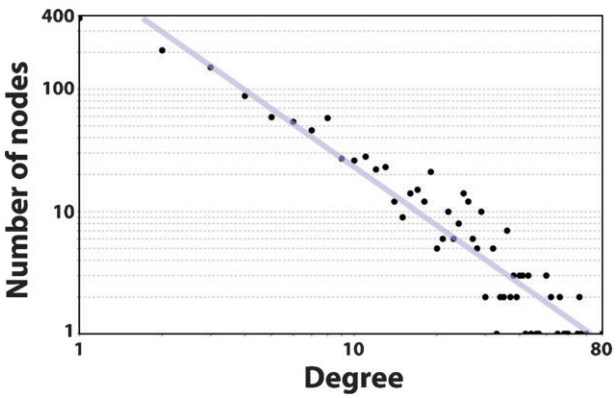


Fig. 6 Degree distribution for the expression correlation and the association network shown in Fig. 5. The linear slope illustrates that this network follows a power-law $P(k) \approx k^{-\gamma}$, where the scale parameter $\gamma = 1.6$, demonstrating that the network has scale-free behavior.

extracted from the correlation network, with an over-representation of GO categories in 'immune response'. This information is summarized in Fig. 7 (right panel). Accordingly, this sub-network whose topological features were recognized *via* the *jActiveModules* plugin links together the functions 'small GTPase signaling', 'immune response' and 'lipid homeostasis' *via* a few hub nodes, specifically RAB10 and SOAT1. RAB10 belongs to the RAS superfamily of small GTPases, to the exocytic and endocytic compartments, and is involved in regulating intracellular vesicle trafficking. There are suggestions that it is involved in glucose transport *via* intracellular retention of the glucose transporter 4 (GLUT4).[118] SOAT1 is acyl-coenzyme A:cholesterol acyl-transferase (ACAT; EC 2.3.1.26), which is an intracellular protein located in the endoplasmic reticulum that esterifies free cholesterol[119] and is a potential target for the control of atherosclerosis.[120] The accumulation of cholesterol esters as lipid droplets within macrophages and smooth muscle cells is a characteristic feature of the early stages of atherosclerotic plaques.[121] Interestingly, SOAT1 and RAB10 are linked in the network by correlation, meaning that both genes evidence similar expression profiles and potentially may be involved in related processes. Accordingly, this sub-network links together functions in cholesterol biosynthesis with glucose transport and immune function, providing interesting information on potential interactions between multiple biological categories in disease etiology.

ClueGO was used to further investigate whether specific ontologies or pathways were over-represented in this sub-network. An examination using KEGG pathways revealed several pathways to be over-represented, including 'PPAR signaling pathway' (APOA1, APOC3, CD36 and LPL) and 'B cell receptor signaling pathway' (BTK, GSK3B, LYN, RAC2 and SYK). An analysis with BioCarta revealed 'PPAR signaling pathway' (APOA1 and LPL), 'B cell receptor signaling pathway' (BTK, LYN and SYK) and 'low-density lipoprotein (LDL) pathway during atherogenesis' (LSLR, LPL and SOAT1). An examination with Immunome showed that 'macrophage cells' (APOA, BCAT1, GM2A, MS4A4A, MSR1 and SCARB2) and 'Th1' (APBB2, APOD and LRP8) were over-represented. This information supports what was observed above for the individual hubs in that there is an overall interaction between lipid transport/biosynthesis and

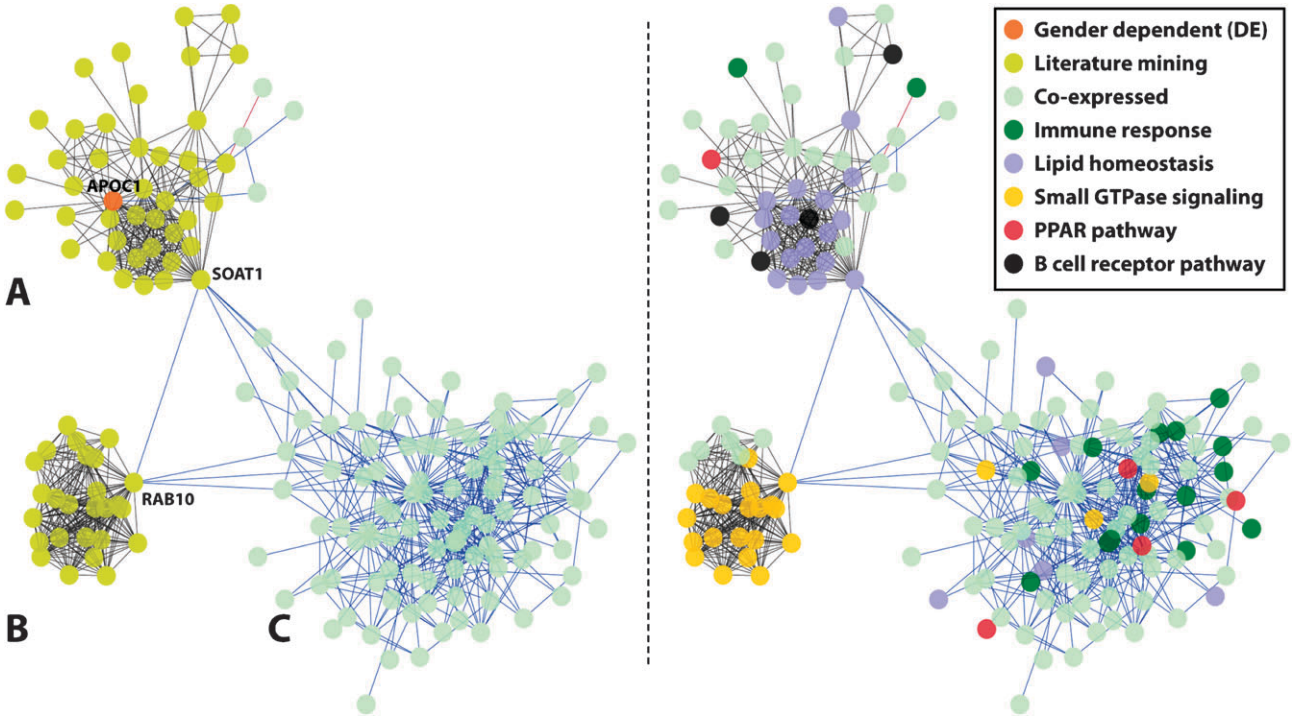


Fig. 7 Sub-network extracted from the analysis in Fig. 5 with *jActiveModules*. The left panel shows the network using the same coloring scheme as described in Fig. 5. This sub-network contains three distinct clusters (named A, B and C). The right panel shows the same sub-network colored according to the analysis of enriched gene ontology (GO) terms, performed with *BiNGO*.[115] Green nodes are over-represented in the 'immune response' category that includes the terms 'immune system process', 'cytokine production' and 'regulation of cytokine biosynthesis'. Violet nodes are enriched in 'lipid homeostasis' including the terms 'lipid transport', 'cholesterol metabolism' and 'lipoprotein metabolism'. Light orange nodes are genes involved in 'small GTPase signaling' pathways. An analysis performed with *ClueGO*[116] to find over-represented KEGG pathways revealed 'PPAR signaling pathway' (shown with red nodes) and 'B cell receptor signaling pathway' (shown with black nodes) genes in the network. Overall, this example shows how networks can be used to discover non-evident relations between genes.

immune function in the plaques, further suggesting that the nodes involved in these interactions should be examined in detail.

A GO analysis (category: biological processes) of the sub-network from Fig. 7 using all information inferred from electronic annotations terms (IEA) resulted in 102 different groups of clustered GO terms (Fig. S3, ESI†). Each cluster was mapped to the nodes of the sub-network (Fig. S4, ESI†) from which a detail of cluster A is shown in Fig. 8. It can be challenging to extract data from this type of analysis due to the sheer volume of information as demonstrated by the complexity of Fig. S3 (ESI†). However, it is possible to sort the network into sub-networks and individual components of interest. For example, some of the main groups relevant to atherosclerosis identified in Fig. 8 include: 'cholesterol transport', which contains many terms related to lipid homeostasis, transport and metabolism; 'regulator of immune system process', which contains terms related to the immune response; 'regulation of mast cell activation', with terms related to leukocyte migration, activation and cytokine production; 'lipoprotein particle clearance', with terms involved in lipoprotein metabolism, *etc*. The node of interest identified *via* statistical analysis, APOC1, is present in 19 different GO categories as shown in the inset in Fig. 8 including: 'cholesterol homeostasis', 'cholesterol metabolic process', 'cholesterol transport', 'innate immune response', 'lipoprotein particle clearance', 'melanocyte differentiation', 'membrane protein ectodomain proteolysis', 'negative regulation of blood vessel endothelial cell migration', 'Notch receptor processing', 'phospholipid efflux', 'regulation of exocytosis', 'regulation of immune response', 'regulation of immune system process', 'regulation of lipid transport', 'regulation of mast cell activation', 'reverse cholesterol transport', 'secretion by cell', and 'triacylglycerol metabolic process'. This level of information can still be challenging to analyze; however, when IEA GO terms are excluded from the *ClueGO* analysis, only six groups are retained from the original 102 groups, of which only one group is present in the APOC1 node ('phospholipid efflux'). Accordingly, by changing the stringency of the analysis, the level of output can be controlled. Taken together, this analysis shows that a node which was identified to be differentially expressed in plaques contains a number of GO categories that are of interest in disease mechanism and etiology. In addition, the main biological features identified within the different clusters are all represented in this node, further suggesting that it plays an important role in plaque biology and pathogenesis. Accordingly, the genes in this sub-network could be treated as putative candidates for further investigating their relation to gender-specific differences in disease.

A review of the literature on APOC1 suggests that this gene has a potential role in atherosclerosis and cardiovascular disease. For example, APOC1 has been shown to increase

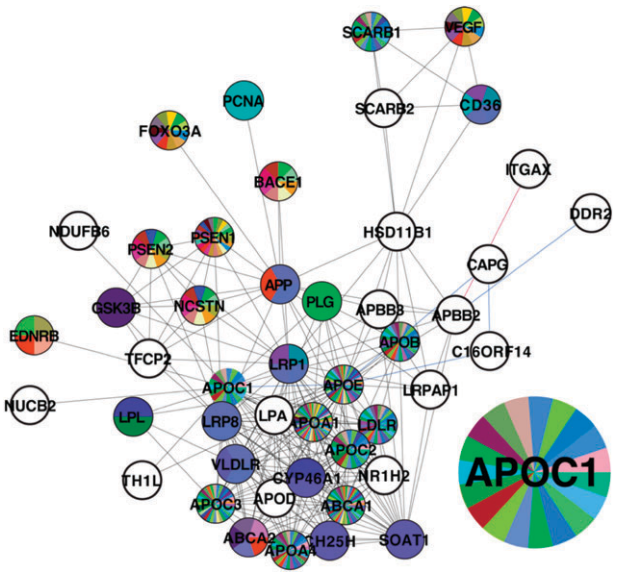


Fig. 8 Detail of cluster A from Fig. 7 in the selected sub-network following GO analysis with *ClueGO*,[116] using all evidence codes. White nodes indicate genes without significant ontology terms. *ClueGO* enables (through the use of the plugin *GOlorize*[117]) the representation of genes with overlapping GO categories. An example is shown with the inset node APOC1, which contains 19 different GO categories including: 'cholesterol homeostasis', 'cholesterol metabolic process', 'cholesterol transport', 'innate immune response', 'lipoprotein particle clearance', 'melanocyte differentiation', 'membrane protein ectodomain proteolysis', 'negative regulation of blood vessel endothelial cell migration', 'Notch receptor processing', 'phospholipid efflux', 'regulation of exocytosis', 'regulation of immune response', 'regulation of immune system process', 'regulation of lipid transport', 'regulation of mast cell activation', 'reverse cholesterol transport', 'secretion by cell', and 'triacylglycerol metabolic process'.

hyperlipidemia in *APOE*−/− mice by stimulating VLDL production and inhibiting lipoprotein lipase (LPL).[122] However, the localization of *APOC*1 appears to be important, with atherosclerosis development increasing with systemic *APOC*1, but remaining unchanged with local macrophage production in the arterial wall.[123] *APOC*1 was also shown to be involved in lipopolysaccharide (LPS)-induced atherosclerosis in *APOE*−/− mice, suggesting that plasma levels of APOC1 can contribute to accelerated atherosclerosis development in individuals with chronic infection.[124] These trends have been further observed in human studies with, for example, the number of APOC1 molecules on small chylomicron remnants strongly associating with the degree of atherosclerosis in normolipidemic men.[125] There is also a potential role of APOC1 in plaque stability, with APOC1 and APOC1-enriched HDL being shown to activate the *N*-SMase–ceramide signaling pathway, leading to apoptosis in human aortic smooth muscle cells—an effect that may promote plaque rupture *in vivo*.[126] The APOC1 content of VLDL particles has been shown to be associated with plaque size in patients with carotid atherosclerosis,[127] and to evidence a postprandial increase in APOC1-containing VLDL in normolipidemic patients with coronary artery disease.[128] Subsequently, the APOC1 content of postprandial triglyceride-rich lipoproteins has been proposed as an independent risk factor for early atherosclerosis and coronary artery disease risk.[125,129] An interesting gender component was observed with male transgenic mice with high *APOC*1 expression in the liver showing elevated levels of serum cholesterol and triglycerides in the VLDL fraction compared with control mice, while females showed less pronounced elevated serum levels.[130] Accordingly, the identification of the APOC1-centric network as potentially important in atherosclerosis is supported by the published literature and suggests that experimental validation should be pursued.

The tools presented in this review enabled us to combine information from three sources: expression correlation, differential expression and literature association. This information was integrated in a network, enabling the extraction of modules or sub-networks evidencing specific topological properties. Using this methodology a sub-network enriched in genes relevant to atherosclerosis was identified. In addition to containing one gene explicitly found to be down-regulated in females, the linked genes provided an important source of additional genes that will enable hypothesis generation. Taken together, this network analysis shows that many of the genes in the selected sub-network are involved in processes related to lipid homeostasis, immune response and atherosclerosis-related pathways. A number of specific processes potentially related to atherosclerosis were identified and provided suggestions for further investigations into the disease mechanism.

Discussion of the tools

We have shown how using freely available tools, a plethora of new analysis techniques can be used to analyze complex data, extracting information that may otherwise be difficult to uncover using classical analysis techniques. Although this methodology is useful, there are several caveats that need to be understood before extracting relevant biological conclusions. For example, although the *ExpressionCorrelation* plugin allows flexible selection of cutoffs, it lacks the ability to determine statistically whether the correlation cutoff selected is significantly different from the observed distribution. In other words, it would be desirable to have a way to detect outliers that will be used in generating the corresponding network. This should increase the probability that the observed correlations correspond to real relations. The *AgilentLiteratureSearch* plugin enables the straightforward extraction of information from the literature that may correspond to experimentally verified relationships. However, it lacks the ability to check the reliability of these associations, or to extract directional information from the analysis. Although manual checking of the literature sources can be performed, this task is time-consuming (Table S3 and Fig. S2, ESI†). The plugin enables the use of gene aliases, but it is unclear from where this information is obtained and how it is updated. Since gene definition and annotation is a dynamic process, with information constantly changing, it would be desirable if the user could have more control over this step. Accordingly, while useful, this approach is solely based upon perceived associations from a parsing of the literature that need to be confirmed with additional evidence. Therefore, the results derived from this analysis should be validated before deriving

conclusions. Interested readers are directed to the literature to learn more about this approach.[131–134] The example of literature mining provided here was conducted with gene data, but it is also possible to perform this type of analysis with proteins and metabolites.[43,135–137] The analysis of GO term enrichment is an established methodology to find functional associations in lists of genes.[138] The *BiNGO* and *ClueGO* plugins provide a convenient interface to perform this kind of analysis in *Cytoscape*. However, one problem with enrichment analyses is that the results depend heavily on the quality of the annotations. Whereas some well-studied genes are richly annotated, others contain annotations only derived from electronic sources or even lack any annotation at all, reducing the significance of the test.

To perform this analysis, several *Cytoscape* plugins were used. However, there are a plethora of alternative plugins that can be used to carry out similar or different tasks. For example, another tool that enables the inference of GRNs from gene expression data is the *MONET* plugin.[139] This method generates a Bayesian network by using prior knowledge about the modeled genes. The *MCODE* plugin,[140] on the other hand, is a very flexible tool that can be used to obtain motifs of highly interconnected nodes. Other plugins like *Cerebral*[141] enable the visualization of omics data from multiple experiments using cellular compartment information. *OmicsViz*[142] enables the visualization of omics data across different species. A list of all available plugins with a brief description of their functionality can be found at the *Cytoscape* web site (http://www.cytoscape.org).

To summarize, as in any other experimental study, these tools are useful to the extent that they provide information to organize and probe the tremendous amount of data derived from omics experiments. Accordingly, these tools can be useful for obtaining results when classical analysis tools fail to find significant associations, something that is especially important in the study of complex diseases like atherosclerosis. It is the responsibility of the user to check the different steps that drive the experiment from hypothesis generation to the derived conclusions in order to minimize misinterpretations of the results. However, as these tools further evolve, it is expected that they will become increasingly integrated and automatic, which should increase their utility to a wider audience. Nonetheless, it will remain important to manually curate and validate important findings as demonstrated by the incorrect associated genes in Fig. S2 (ESI†).

Conclusions and future directions

The analysis of biological networks is still in its infancy. Although general properties such as scale-free topology or hierarchical networks are capable of explaining some of the properties of living systems, there are many unsolved questions. It is necessary to understand why some metabolites dramatically affect the stability of the system, whereas drastic alterations of other system components such as knocking down a gene can sometimes have little effect. More in-depth knowledge regarding the general properties inherent to biological networks will come from the integration of all data already available, and should facilitate the development of specific drugs targeting a desired effect in the network.

The ultimate goal from a clinical perspective is an increase in our understanding of disease etiology and pathogenesis leading to concomitant increase in the development of new therapies. A translational systems biology approach may be a feasible option to solve crucial clinical issues. With respect to atherosclerosis, one of the most important current clinical problems is determining if and when a patient will develop a symptomatic disease, as well as identification or imaging of vulnerable lesions. A systems biology approach as described for carotid disease may be capable of identifying molecular pathways and targets that operate in plaque instability and help to develop molecular tools that can be applied to imaging modalities such as MRI or PET CT to identify vulnerable lesions, improve patient selection or monitoring of stroke-preventive intervention. Accordingly, a systems approach could provide concrete clinical applications to address the needs of the medical community. In particular, a systems biology analytical approach may provide the opportunity to identify medications with a desirable "network" effect rather than the traditional approach of seeking treatments targeting single-gene effects, and therefore enable more holistic targeted treatments—the promise of personalized medicine.

Systems biology is rapidly changing the way we examine living organisms. Biological network construction is a useful tool, but to realize the ultimate goal of systems biology, *i.e.* the understanding of the organisms as a whole, the next major challenge is to combine and analyze data from multiple sources.[143] Currently, individual networks are examined as independent entities, which is an oversimplification. Each network is integrated into the entire system, which works together and the system cannot be understood without considering all individual components to eventually generate a "network of networks". For example, GRNs, PPI networks, protein–DNA interaction networks and metabolic networks are all integrated into a single compartment that comprises the cell. Cells in turn constitute a network in which different cell types evidence specific interactions, which are then networked into the organ and finally the whole organism level. These effects will be further complicated by interactions with microbiota and the environment, which can have profound effects upon disease.[144–147] Therefore it is not only a problem of integrating different networks at a given scale (*e.g.*, the cell) but to integrate the information at different scales (*e.g.*, how a mutation in a gene affects the overall state of the whole organism). Some initial steps in that direction have already been made,[148] but increased integration of heterogeneous data and networks is non-trivial. The virtual physiological human (VPH) is an exciting step in this direction and aims to "enable" collaborative investigation of the human body as a single complex system.[149] This approach to quantitatively studying human physiology can be combined with phenotypic information such as that provided in the Phenotypic Disease Network (PDN[150]) to link together dynamic networks, physiology and phenotype towards the goal of complete system understanding. The potential of combining the knowledge from different networks and the accumulation of high-throughput data will move us one step further towards an unprecedented

understanding of a living organism. The rapid advances in computer sciences coupled with advances in high-throughput technologies are moving the field into new areas. However, even more important are the paradigm shifts in the way that clinicians, computer scientists, engineers and laboratory-based scientists approach research. A realization has come that in order to develop accurate models of living systems; a truly interdisciplinary approach is required.

Funding

This research was supported by the Åke Wibergs Stiftelse, the Fredrik and Ingrid Thurings Stiftelse, the Jeanssons Stiftelse, the Royal Swedish Academy of Sciences, the Swedish Research Council (10350; 20854; 6816; the Linnaeus grant CERIC 70870301), CIDaT, EU FP6 and FP7 (005033; 201668), VINNOVA and the Swedish Heart-Lung Foundation. The BiKE biobank is supported by grants from the AFA Fund, Swedish Heart-Lung Foundation, and the Swedish Research Council (14121; K2009-65X-2233-01-3). C.E.W. was supported by a Center for Allergy Research Fellowship (Cfa), D.D. was supported by the Japanese Society for the Promotion of Science (JSPS) and S.G. was supported by KAKENHI from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1 H. Kitano, *Science*, 2002, **295**, 1662–1664.
2 H. Kitano, *Nature*, 2002, **420**, 206–210.
3 A. Aderem, *Cell (Cambridge, Mass.)*, 2005, **121**, 511–513.
4 A. L. Barabasi and Z. N. Oltvai, *Nat. Rev. Genet.*, 2004, **5**, 101–113.
5 A. C. Ahn, M. Tewari, C. S. Poon and R. S. Phillips, *PLoS Med.*, 2006, **3**, e208.
6 A. C. Ahn, M. Tewari, C. S. Poon and R. S. Phillips, *PLoS Med.*, 2006, **3**, e209.
7 L. Hood, J. R. Heath, M. E. Phelps and B. Lin, *Science*, 2004, **306**, 640–643.
8 L. Hood and R. M. Perlmutter, *Nat. Biotechnol.*, 2004, **22**, 1215–1217.
9 A. D. Weston and L. Hood, *J. Proteome Res.*, 2004, **3**, 179–196.
10 M. Stoll, A. W. Cowley, Jr., P. J. Tonellato, A. S. Greene, M. L. Kaldunski, R. J. Roman, P. Dumas, N. J. Schork, Z. Wang and H. J. Jacob, *Science*, 2001, **294**, 1723–1726.
11 R. Tabibiazar, R. A. Wagner, E. A. Ashley, J. Y. King, R. Ferrara, J. M. Spin, D. A. Sanan, B. Narasimhan, R. Tibshirani, P. S. Tsao, B. Efron and T. Quertermous, *Physiol. Genomics*, 2005, **22**, 213–226.
12 J. Y. King, R. Ferrara, R. Tabibiazar, J. M. Spin, M. M. Chen, A. Kuchinsky, A. Vailaya, R. Kincaid, A. Tsalenko, D. X. Deng, A. Connolly, P. Zhang, E. Yang, C. Watt, Z. Yakhini, A. Ben-Dor, A. Adler, L. Bruhn, P. Tsao, T. Quertermous and E. A. Ashley, *Physiol. Genomics*, 2005, **23**, 103–118.
13 E. A. Ashley, R. Ferrara, J. Y. King, A. Vailaya, A. Kuchinsky, X. He, B. Byers, U. Gerckens, S. Oblin, A. Tsalenko, A. Soito, J. M. Spin, R. Tabibiazar, A. J. Connolly, J. B. Simpson, E. Grube and T. Quertermous, *Circulation*, 2006, **114**, 2644–2654.
14 P. S. Gargalovic, M. Imura, B. Zhang, N. M. Gharavi, M. J. Clark, J. Pagnon, W. P. Yang, A. He, A. Truong, S. Patel, S. F. Nelson, S. Horvath, J. A. Berliner, T. G. Kirchgessner and A. J. Lusis, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 12741–12746.
15 A. Ghazalpour, S. Doss, B. Zhang, S. Wang, C. Plaisier, R. Castellanos, A. Brozell, E. E. Schadt, T. A. Drake, A. J. Lusis and S. Horvath, *PLoS Genet.*, 2006, **2**, e130.
16 J. Skogsberg, J. Lundström, A. Kovacs, R. Nilsson, P. Noori, S. Maleki, M. Köhler, A. Hamsten, J. Tegner and J. Björkegren, *PLoS Genet.*, 2008, **4**, e1000036.
17 M. Toenjes, M. Schueler, S. Hammer, U. J. Pape, J. J. Fischer, F. Berger, M. Vingron and S. Sperling, *Mol. BioSyst.*, 2008, **4**, 589–598.
18 S. Cagnin, M. Biscuola, C. Patuzzo, E. Trabetti, A. Pasquali, P. Laveder, G. Faggian, M. Iafrancesco, A. Mazzucco, P. F. Pignatti and G. Lanfranchi, *BMC Genomics*, 2009, **10**, 13.
19 W. Zhu, L. Yang and Z. Du, *PLoS One*, 2009, **4**, e6288.
20 N. J. Samani, J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino, B. Mayer, R. J. Dixon, T. Meitinger, P. Braund, H. E. Wichmann, J. H. Barrett, I. R. Konig, S. E. Stevens, S. Szymczak, D. A. Tregouet, M. M. Iles, F. Pahlke, H. Pollard, W. Lieb, F. Cambien, M. Fischer, W. Ouwehand, S. Blankenberg, A. J. Balmforth, A. Baessler, S. G. Ball, T. M. Strom, I. Braenne, C. Gieger, P. Deloukas, M. D. Tobin, A. Ziegler, J. R. Thompson and H. Schunkert, *N. Engl. J. Med.*, 2007, **357**, 443–453.
21 T. W. T. C. C. Consortium, *Nature*, 2007, **447**, 661–678.
22 S. Kathiresan, C. J. Willer, G. M. Peloso, S. Demissie, K. Musunuru, E. E. Schadt, L. Kaplan, D. Bennett, Y. Li, T. Tanaka, B. F. Voight, L. L. Bonnycastle, A. U. Jackson, G. Crawford, A. Surti, C. Guiducci, N. P. Burtt, S. Parish, R. Clarke, D. Zelenika, K. A. Kubalanza, M. A. Morken, L. J. Scott, H. M. Stringham, P. Galan, A. J. Swift, J. Kuusisto, R. N. Bergman, J. Sundvall, M. Laakso, L. Ferrucci, P. Scheet, S. Sanna, M. Uda, Q. Yang, K. L. Lunetta, J. Dupuis, P. I. de Bakker, C. J. O'Donnell, J. C. Chambers, J. S. Kooner, S. Hercberg, P. Meneton, E. G. Lakatta, A. Scuteri, D. Schlessinger, J. Tuomilehto, F. S. Collins, L. Groop, D. Altshuler, R. Collins, G. M. Lathrop, O. Melander, V. Salomaa, L. Peltonen, M. Orho-Melander, J. M. Ordovas, M. Boehnke, G. R. Abecasis, K. L. Mohlke and L. A. Cupples, *Nat. Genet.*, 2009, **41**, 56–65.
23 E. E. Schadt, S. H. Friend and D. A. Shaywitz, *Nat. Rev.*, 2009, **8**, 286–295.
24 X. Yang, J. L. Deignan, H. Qi, J. Zhu, S. Qian, J. Zhong, G. Torosyan, S. Majid, B. Falkard, R. R. Kleinhanz, J. Karlsson, L. W. Castellani, S. Mumick, K. Wang, T. Xie, M. Coon, C. Zhang, D. Estrada-Smith, C. R. Farber, S. S. Wang, A. van Nas, A. Ghazalpour, B. Zhang, D. J. Macneil, J. R. Lamb, K. M. Dipple, M. L. Reitman, M. Mehrabian, P. Y. Lum, E. E. Schadt, A. J. Lusis and T. A. Drake, *Nat. Genet.*, 2009, **41**, 415–423.
25 S. Bornholdt, *Science*, 2005, **310**, 449–451.
26 A. R. Joyce and B. O. Palsson, *Nat. Rev. Mol. Cell Biol.*, 2006, **7**, 198–210.
27 A. Ma'ayan, *IET Syst. Biol.*, 2008, **2**, 206–221.
28 G. W. Bell and F. Lewitter, *Methods Enzymol.*, 2006, **411**, 408–421.
29 J. D. Han, *Cell Res.*, 2008, **18**, 224–237.
30 R. E. Bumgarner and K. Y. Yeung, *Methods Mol. Biol. (Totowa, N. J.)*, 2009, **541**, 225–245.
31 R. Albert and A. Barabási, *Rev. Mod. Phys.*, 2002, **74**, 47–97.
32 M. E. J. Newman, *Soc. Ind. Appl. Math.*, 2003, 167–256.
33 A. Clauset, C. Moore and M. E. J. Newman, *Nature*, 2008, **453**, 98–101.
34 S. Bornholdt and H. G. Schuster, *Handbook of Graphs and Networks: From the Genome to the Internet*, Wiley-VCH, Berlin, 2003.
35 S. N. Dorogovstev and J. F. F. Mendes, *Evolution of Networks from Biologicals Nets to the Internet and WWW*, Oxford University Press, Oxford, 2003.
36 W. Huber, V. J. Carey, L. Long, S. Falcon and R. Gentleman, *BMC Bioinformatics*, 2007, **8**(suppl 6), S8.
37 H. Hu, S. Myers, V. Colizza and A. Vespignani, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 1318–1323.
38 A. Vespignani, *Science*, 2009, **325**, 425–428.
39 N. A. Christakis and J. H. Fowler, *N. Engl. J. Med.*, 2007, **357**, 370–379.
40 J. Raacke and J. Bonds-Raacke, *Cyberpsychol. Behav.*, 2008, **11**, 169–174.

41 G. Karlebach and R. Shamir, *Nat. Rev. Mol. Cell Biol.*, 2008, **9**, 770–780.
42 N. Blow, *Nature*, 2009, **460**, 415–418.
43 T. Ideker and R. Sharan, *Genome Res.*, 2008, **18**, 644–652.
44 S. H. Strogatz, *Nature*, 2001, **410**, 268–276.
45 D. Bray, *Science*, 2003, **301**, 1864–1865.
46 L. Chang and M. Karin, *Nature*, 2001, **410**, 37–40.
47 H. Kitano, A. Funahashi, Y. Matsuoka and K. Oda, *Nat. Biotechnol.*, 2005, **23**, 961–966.
48 C. E. Wheelock, A. M. Wheelock, S. Kawashima, D. Diez, M. Kanehisa, M. van Erk, R. Kleemann, J. Z. Haeggstrom and S. Goto, *Mol. BioSyst.*, 2009, **5**, 588–602.
49 R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang and J. Zhang, *GenomeBiology*, 2004, **5**, R80.
50 M. Reimers and V. J. Carey, *Methods Enzymol.*, 2006, **411**, 119–134.
51 J. D. Zhang and S. Wiemann, *Bioinformatics*, 2009, **25**, 1470–1471.
52 J. L. Bos, *Cancer Res.*, 1989, **49**, 4682–4689.
53 J. Downward, *Nat. Rev. Cancer*, 2003, **3**, 11–22.
54 M. Kanehisa and S. Goto, *Nucleic Acids Res.*, 2000, **28**, 27–30.
55 M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu and Y. Yamanishi, *Nucleic Acids Res.*, 2008, **36**, D480–484.
56 E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabasi, *Science*, 2002, **297**, 1551–1555.
57 E. Almaas, *J. Exp. Biol.*, 2007, **210**, 1548–1558.
58 A. L. Barabasi, *Science*, 2009, **325**, 412–413.
59 C. T. Butts, *Science*, 2009, **325**, 414–416.
60 A. Barrat, M. Barthelemy, R. Pastor-Satorras and A. Vespignani, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 3747–3752.
61 P. Grindrod and M. Kibble, *Expert Rev. Proteomics*, 2004, **1**, 229–238.
62 O. Mason and M. Verwoerd, *IET Syst. Biol.*, 2007, **1**, 89–119.
63 K. I. Goh, E. Oh, B. Kahng and D. Kim, *Phys. Rev.*, 2003, **67**, 017101.
64 J. Yoon, A. Blumer and K. Lee, *Bioinformatics*, 2006, **22**, 3106–3108.
65 H. Jeong, S. P. Mason, A. L. Barabasi and Z. N. Oltvai, *Nature*, 2001, **411**, 41–42.
66 H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A. L. Barabasi, *Nature*, 2000, **407**, 651–654.
67 S. S. Shen-Orr, R. Milo, S. Mangan and U. Alon, *Nat. Genet.*, 2002, **31**, 64–68.
68 R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, *Science*, 2002, **298**, 824–827.
69 A. P. Burgard, E. V. Nikolaev, C. H. Schilling and C. D. Maranas, *Genome Res.*, 2004, **14**, 301–312.
70 E. V. Nikolaev, A. P. Burgard and C. D. Maranas, *Biophys. J.*, 2005, **88**, 37–49.
71 V. Vermeirssen, M. I. Barrasa, C. A. Hidalgo, J. A. Babon, R. Sequerra, L. Doucette-Stamm, A. L. Barabasi and A. J. Walhout, *Genome Res.*, 2007, **17**, 1061–1071.
72 N. K. Hollenberg, *Curr. Hypertens. Rep.*, 2002, **4**, 412–413.
73 S. E. Calvano, W. Xiao, D. R. Richards, R. M. Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, S. K. Tschoeke, C. Miller-Graziano, L. L. Moldawer, M. N. Mindrinos, R. W. Davis, R. G. Tompkins and S. F. Lowry, *Nature*, 2005, **437**, 1032–1037.
74 K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A. L. Barabasi, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 8685–8690.
75 X. Wu, R. Jiang, M. Q. Zhang and S. Li, *Mol. Syst. Biol.*, 2008, **4**, 189.
76 J. C. Nacher and T. Akutsu, *Cell Biochem. Biophys.*, 2007, **49**, 37–47.
77 M. Arita, *J. Biochem.*, 2005, **138**, 1–4.
78 D. J. Watts and S. H. Strogatz, *Nature*, 1998, **393**, 440–442.
79 R. Cohen and S. Havlin, *Phys. Rev. Lett.*, 2003, **90**, 058701.
80 A. L. Barabasi and R. Albert, *Science*, 1999, **286**, 509–512.
81 R. Albert, H. Jeong and A. L. Barabasi, *Nature*, 2000, **406**, 378–382.
82 C. Song, S. Havlin and H. A. Makse, *Nature*, 2005, **433**, 392–395.
83 A. Barabási and H. Jeong, *Physica A (Amsterdam)*, 2000, **281**, 69–77.
84 M. Girvan and M. E. Newman, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 7821–7826.
85 N. Lin and H. Zhao, *BMC Bioinformatics*, 2005, **6**, 119.
86 R. Cohen, K. Erez, D. ben-Avraham and S. Havlin, *Phys. Rev. Lett.*, 2000, **85**, 4626–4628.
87 D. Deutscher, I. Meilijson, M. Kupiec and E. Ruppin, *Nat. Genet.*, 2006, **38**, 993–998.
88 U. Alon, *Nat. Rev. Genet.*, 2007, **8**, 450–461.
89 R. Schleif, *Trends Genet.*, 2000, **16**, 559–565.
90 P. M. Yen, *Physiol. Rev.*, 2001, **81**, 1097–1142.
91 R. P. Alexander, P. M. Kim, T. Emonet and M. B. Gerstein, *Sci. Signaling*, 2009, **2**, pe44.
92 E. Ravasz, *Methods Mol. Biol.*, 2009, **541**, 145–160.
93 S. A. Ramsey, S. L. Klemm, D. E. Zak, K. A. Kennedy, V. Thorsson, B. Li, M. Gilchrist, E. S. Gold, C. D. Johnson, V. Litvak, G. Navarro, J. C. Roach, C. M. Rosenberger, A. G. Rust, N. Yudkovsky, A. Aderem and I. Shmulevich, *PLoS Comput. Biol.*, 2008, **4**, e1000021.
94 T. Kuhlman, Z. Zhang, M. H. Saier, Jr. and T. Hwa, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 6043–6048.
95 A. Ma′ayan, S. L. Jenkins, J. Goldfarb and R. Iyengar, *Mt. Sinai J. Med.*, 2007, **74**, 27–32.
96 A. Ghazalpour, S. Doss, X. Yang, J. Aten, E. M. Toomey, A. Van Nas, S. Wang, T. A. Drake and A. J. Lusis, *J. Lipid Res.*, 2004, **45**, 1793–1805.
97 E. E. Schadt and P. Y. Lum, *J. LipidRes.*, 2006, **47**, 2601–2613.
98 J. Tegner, J. Skogsberg and J. Bjorkegren, *J. Lipid Res.*, 2007, **48**, 267–277.
99 L. Gautier, L. Cope, B. M. Bolstad and R. A. Irizarry, *Bioinformatics*, 2004, **20**, 307–315.
100 G. K. Smyth, in *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, ed. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry and W. Huber, Springer, New York, 2005, pp. 397–420.
101 L. Dumont, T. Gautier, J. P. de Barros, H. Laplanche, D. Blache, P. Ducoroy, J. Fruchart, J. C. Fruchart, P. Gambert, D. Masson and L. Lagrost, *J. Biol. Chem.*, 2005, **280**, 38108–38116.
102 N. S. Shachter, *Curr. Opin. Lipidol.*, 2001, **12**, 297–304.
103 J. F. Berbee, C. C. van der Hoogt, D. Sundararaman, L. M. Havekes and P. C. Rensen, *J. Lipid Res.*, 2005, **46**, 297–306.
104 Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc. Ser. B (Methodological)*, 1995, **57**, 289–300.
105 A. Ng, B. Bursteinas, Q. Gao, E. Mollison and M. Zvelebil, *Briefings Bioinf.*, 2006, **7**, 318–330.
106 A. Bauer-Mehren, L. I. Furlong and F. Sanz, *Mol. Syst. Biol.*, 2009, **5**, 290.
107 P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res.*, 2003, **13**, 2498–2504.
108 M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P. L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker and G. D. Bader, *Nat. Protoc.*, 2007, **2**, 2366–2382.
109 N. Yeung, M. S. Cline, A. Kuchinsky, M. E. Smoot and G. D. Bader, *Current protocols in bioinformatics*, ed. A. D. Baxevanis, L. D. Stein, G. D. Stormo and J. R. Yates III, Wiley, New York, 2008, ch. 8, Unit 8,p. 13.
110 S. Killcoyne, G. W. Carter, J. Smith and J. Boyle, *Methods Mol. Biol.*, 2009, **563**, 219–239.
111 ExpressionCorrelation (Cytoscape Plugin), http://www.baderlab.org/Software/ExpressionCorrelation.
112 A. Vailaya, P. Bluvas, R. Kincaid, A. Kuchinsky, M. Creech and A. Adler, *Bioinformatics*, 2005, **21**, 430–438.
113 Y. Assenov, F. Ramirez, S. E. Schelhorn, T. Lengauer and M. Albrecht, *Bioinformatics*, 2008, **24**, 282–284.

114 T. Ideker, O. Ozier, B. Schwikowski and A. F. Siegel, *Bioinformatics*, 2002, **18**(suppl 1), S233–240.
115 S. Maere, K. Heymans and M. Kuiper, *Bioinformatics*, 2005, **21**, 3448–3449.
116 G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W. H. Fridman, F. Pages, Z. Trajanoski and J. Galon, *Bioinformatics*, 2009, **25**, 1091–1093.
117 O. Garcia, C. Saveanu, M. Cline, M. Fromont-Racine, A. Jacquier, B. Schwikowski and T. Aittokallio, *Bioinformatics*, 2007, **23**, 394–396.
118 C. B. Dugani and A. Klip, *EMBO Rep.*, 2005, **6**, 1137–1142.
119 T. Y. Chang, B. L. Li, C. C. Chang and Y. Urano, *Am. J. Physiol.: Cell Physiol.*, 2009, **297**, E1–9.
120 T. A. Bell, 3rd, J. M. Brown, M. J. Graham, K. M. Lemonidis, R. M. Crooke and L. L. Rudel, *Arterioscler., Thromb., Vasc. Biol.*, 2006, **26**, 1814–1820.
121 G. K. Hansson, *N. Engl. J. Med.*, 2005, **352**, 1685–1695.
122 M. Westerterp, W. de Haan, J. F. Berbee, L. M. Havekes and P. C. Rensen, *J. Lipid Res.*, 2006, **47**, 1203–1211.
123 M. Westerterp, M. Van Eck, W. de Haan, E. H. Offerman, T. J. Van Berkel, L. M. Havekes and P. C. Rensen, *Atherosclerosis*, 2007, **195**, e9–16.
124 M. Westerterp, J. F. Berbee, N. M. Pires, G. J. van Mierlo, R. Kleemann, J. A. Romijn, L. M. Havekes and P. C. Rensen, *Circulation*, 2007, **116**, 2173–2181.
125 J. Bjorkegren, A. Silveira, S. Boquist, R. Tang, F. Karpe, M. G. Bond, U. de Faire and A. Hamsten, *Arterioscler., Thromb., Vasc. Biol.*, 2002, **22**, 1470–1474.
126 A. Kolmakova, P. Kwiterovich, D. Virgil, P. Alaupovic, C. Knight-Gibson, S. F. Martin and S. Chatterjee, *Arterioscler., Thromb., Vasc. Biol.*, 2004, **24**, 264–269.
127 A. T. Noto, E. B. Mathiesen, J. Brox, J. Bjorkegren and J. B. Hansen, *Lipids*, 2008, **43**, 673–679.
128 J. Bjorkegren, S. Boquist, A. Samnegard, P. Lundman, P. Tornvall, C. G. Ericsson and A. Hamsten, *Circulation*, 2000, **101**, 227–230.
129 A. Hamsten, A. Silveira, S. Boquist, R. Tang, M. G. Bond, U. de Faire and J. Bjorkegren, *J. Am. Coll. Cardiol.*, 2005, **45**, 1013–1017.
130 M. C. Jong, V. E. Dahlmans, P. J. van Gorp, K. W. van Dijk, M. L. Breuer, M. H. Hofker and L. M. Havekes, *J. Clin. Invest.*, 1996, **98**, 2259–2267.
131 S. Dickman, *PLoS Biol.*, 2003, **1**, E48.
132 P. Sharma, R. D. Senthilkumar, V. Brahmachari, E. Sundaramoorthy, A. Mahajan, A. Sharma and S. Sengupta, *Lipids Health Dis.*, 2006, **5**, 1.
133 R. B. Altman, C. M. Bergman, J. Blake, C. Blaschke, A. Cohen, F. Gannon, L. Grivell, U. Hahn, W. Hersh, L. Hirschman, L. J. Jensen, M. Krallinger, B. Mons, S. I. O'Donoghue, M. C. Peitsch, D. Rebholz-Schuhmann, H. Shatkay and A. Valencia, *GenomeBiology*, 2008, **9**(suppl 2), S7.
134 M. Krallinger, A. Valencia and L. Hirschman, *GenomeBiology*, 2008, **9**(suppl 2), S8.
135 C. J. Baker, R. Kanagasabai, W. T. Ang, A. Veeramani, H. S. Low and M. R. Wenk, *BMC Bioinformatics*, 2008 **9**(suppl 1), S5.
136 A. Krishnan, J. P. Zbilut, M. Tomita and A. Giuliani, *Curr. Protein Pept. Sci.*, 2008, **9**, 28–38.
137 R. Chowdhary, J. Zhang and J. S. Liu, *Bioinformatics*, 2009, **25**, 1535–1542.
138 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nat. Genet.*, 2000, **25**, 25–29.
139 P. H. Lee and D. Lee, *Bioinformatics*, 2005, **21**, 2739–2747.
140 G. D. Bader and C. W. Hogue, *BMC Bioinformatics*, 2003, **4**, 2.
141 A. Barsky, J. L. Gardy, R. E. Hancock and T. Munzner, *Bioinformatics*, 2007, **23**, 1040–1042.
142 T. Xia and J. A. Dickerson, *Bioinformatics*, 2008, **24**, 2557–2558.
143 U. Sauer, M. Heinemann and N. Zamboni, *Science*, 2007, **316**, 550–551.
144 J. K. Nicholson, E. Holmes, J. C. Lindon and I. D. Wilson, *Nat. Biotechnol.*, 2004, **22**, 1268–1274.
145 J. M. Ordovas and J. Shen, *J. Periodontol.*, 2008, **79**, 1508–1513.
146 J. M. Ordovas and E. S. Tai, *Curr. Opin. Lipidol.*, 2008, **19**, 158–167.
147 P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight and J. I. Gordon, *Nature*, 2009, **457**, 480–484.
148 N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P. Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori and M. Tomita, *Science*, 2007, **316**, 593–597.
149 P. Kohl and D. Noble, *Mol. Syst. Biol.*, 2009, **5**, 292.
150 C. A. Hidalgo, N. Blumm, A. L. Barabasi and N. A. Christakis, *PLoS Comput. Biol.*, 2009, **5**, e1000353.

