Network-Guided Key Gene Discovery for a Given Cellular Process

Feng Q. He and Markus Ollert

Abstract Identification of key genes for a given physiological or pathological process is an essential but still very challenging task for the entire biomedical research community. Statistics-based approaches, such as genome-wide association study (GWAS)- or quantitative trait locus (QTL)-related analysis have already made enormous contributions to identifying key genes associated with a given disease or phenotype, the success of which is however very much dependent on a huge number of samples. Recent advances in network biology, especially network inference directly from genome-scale data and the following-up network analysis, opens up new avenues to predict key genes driving a given biological process or cellular function. Here we review and compare the current approaches in predicting key genes, which have no chances to stand out by classic differential expression analysis, from gene-regulatory, protein-protein interaction, or gene expression correlation networks. We elaborate these network-based approaches mainly in the context of immunology and infection, and urge more usage of correlation network-based predictions. Such network-based key gene discovery approaches driven by information-enriched 'omics' data should be very useful for systematic key gene discoveries for any given biochemical process or cellular function, and also valuable for novel drug target discovery and novel diagnostic, prognostic and therapeutic-efficiency marker prediction for a specific disease or disorder.

F.Q. He (✉)
Department of Infection and Immunity, Group of Immune Systems Biology, Luxembourg Institute of Health, 29, rue Henri Koch, 4354 Esch-sur-Alzette, Luxembourg
e-mail: feng.he@lih.lu

M. Ollert
Department of Infection and Immunity, Group of Allergy and Clinical Immunology, Luxembourg Institute of Health, 29, rue Henri Koch, 4354 Esch-sur-Alzette, Luxembourg

Odense Research Center for Anaphylaxis, Department of Dermatology and Allergy Center, Odense University Hospital, University of Southern Denmark, 5000 Odense C, Denmark

Keywords Correlation network, Key gene prediction, Network biology, Systems biology, Systems immunology

Contents

1 Introduction

Network biology has emerged as one of the most promising angles of systems biology and systems medicine [1–4]. As realization of any biological function or process is fully dependent on the interactions of small and macro molecules rather than individual proteins, RNAs or metabolites, it is almost undisputed to require biomedical researchers to work on interactions of cellular constituents and the resulting networks composed of nodes and interactions. The researchers from network biology, as the name suggests, mainly deal with various types of networks in biology. The major conceptual contribution of network biology to the entire biomedical community is to help and guide basic, translational and clinical researchers to investigate systematically the functions of genes/proteins or metabolites based on interlinked molecular networks rather than isolated single pathways or even individual genes. With the ability to gain a holistic understanding of cellular activities, network biologists have a unique power in characterizing and predicting robustness, flexibility, redundancy, and many other evolutionarily-conserved systems features in various kinds of uni- and multicellular organisms. There are numerous types of intra- and intercellular networks. For example, in the classic metabolic networks, where metabolites react with each other to consume and produce metabolites, the nodes represent metabolites and links between nodes designate metabolic reactions between linked metabolites. Another important type of molecular network is the gene regulatory network (GRN), in which one node corresponds to one gene and one arrow from one gene to another indicates a transcription regulatory relationship. In contrast to these directed networks, researchers have also started to reveal undirected networks such as protein–protein interaction (PPI) networks in which one node represents one protein and a link

between two nodes indicates a physical interaction of the given two proteins. In real life, not to mention intercellular networks, all the various network layers are crosslinked with each other, introducing further complexity in understanding the molecular mechanism underlying any biological process. To handle better the obstacles to understanding the complex cellular mechanisms driven by dynamic multiscale interlocked networks [5, 6], network biologists visualize, construct, and analyze various molecular networks from different entry points.

Currently, the major trends in network biology are to infer and reconstruct various types of molecular interaction networks, for example to reconstruct and infer metabolic networks, GRNs, signaling transduction networks, PPI networks, and other types of networks. Particularly, with the development of large-scale 'omics' measurement techniques, many types of genome-scale data from different tissues and organs and cell types with and without different perturbations are increasingly available, including genome, transcriptome, proteome, metabolome, miRNAs, and others which further drives the fast advances in various computational approaches in network inference and reconstruction. For technical details in network reconstruction and inference, also known as data-driven reverse engineering, the reader should refer to other reviews [7–10].

Following the accomplishment of the reconstruction or inference of the given molecular networks, we cannot just stay at the stage of being impressed by the visualized large complex networks. Although it is apparent that the current available large-scale datasets, knowledge-based interaction databases, and computational approaches do not yet allow us to obtain a complete and fully-accurate molecular network for a given organism or cell, it cannot stop us from going forward. The next wave of objectives are to make use of these complex molecular networks that are still incomplete and far from being fully accurate.

Until now, many efforts have already been made mainly in analyzing metabolic networks, which are relatively more accurate and complete than other types of molecular interaction networks. Among many applications, in the metabolic network it has been well established that one could calculate all the potential elementary pathways or extreme pathways between the starting metabolites and final products [11–13]. Such pure stoichiometric and topological properties-based approaches are extremely valuable for investigating the redundancy nature of molecular networks. However, all the physiological and pathological processes take place with certain resource constraints, indicating that the solution space should hopefully narrow down. Based on the reconstructed metabolic networks and given constraints, pioneers have already been able to predict effects of gene deletions on cell viability of *Saccharomyces cerevisiae* with a pretty high accuracy using flux balance analysis [14]. More practically, the metabolic network has recently been used to predict drug targets or a combination of targets against human cancer cells by identifying essential genes for cell proliferation or synthetic lethal genes [15]. Predicting essential genes for cell proliferation or survival based on metabolic networks is in fact the first successful application of network biology. Possibly because of the nature and essential function of metabolic networks, namely maintaining cell survival and multiplicity, one can hardly predict genes

critical to other important functions rather than only viability and proliferation. Many more functions can be executed even by bacterial cells, not to mention mammalian cells. In the meantime, many more layers of molecular networks fortunately exist to complement the onefold functions of metabolic networks. The extra complexity also increases the difficulty for us to predict key genes for a given cellular function or process.

It is important to notice the essential discrepancy between key gene predictions for a given biological process or function and that for a specific disease. The latter has been well reviewed [16, 17] whereas the former is still open to be summarized. Although any dysfunction for the disease key genes often cause a catastrophic consequence to human health, the expression or functional changes of key genes for the former cases might only alter the activities of a particular biological pathway which might not necessarily lead to a severe event at the whole cellular or organism level. To identify those playing a quantitatively regulatory role, which might be affected or compensated by the existence of redundant pathways of molecular networks, is even more challenging than to identify the death-or-survival essential nodes from the networks. Furthermore, the readout of particular activity changes in many given cellular processes or functions is often not so obvious in contrast to disastrous symptoms or not so easily examined by current techniques.

In this review we mainly summarize the recent advances in the key gene discovery for a given physiological or pathological process based on GRNs, PPI, and expression correlation networks. We also survey several selected examples in the context of immunology and infection. Through comparisons we support more implementation of expression correlation network-based key gene predictions that are still non-mainstream. Furthermore, we present our views on potential challenges and future directions.

2 GRNs-Based Key Gene Discovery

2.1 *Hub-Based Prediction*

The long-term dream of biologists is to understand fully how one or multiple transcription factors dynamically and quantitatively regulate the target genes [18–20], the regulatory cascades composed of chains or trees of linked interactions, and eventually the overall GRN in the given cell type or organism. In the last few decades, enormous progress has been made in understanding the regulatory relationships between individual regulators and some of their target genes. Accumulating results in individual regulatory interactions by finer-scale or large-scale high throughput experiments cannot automatically help us to understand how the GRNs control cellular functions and processes. To understand better these regulatory processes systematically, the first step is to assemble and reconstruct the GRNs from various public available or commercial interaction databases. In parallel,

although significant progress has already been made, computational techniques also need to be further developed to infer GRNs from a single type or multi-type of genome-scale data such as ChIP-chip, ChIP-seq, RNA-seq/microarray, proteome, metabolome, and others with or without time-series and knockdown/knockout measurements [8, 21–23]. Following the reconstruction and inference of GRNs, researchers aim to predict whether there are any novel key regulators for the biological process or the given disease of interest. Biomedical researchers are of course also highly interested in identifying novel drug targets for therapy of the disease of interest by analyzing GRNs. The most easily acceptable and straightforward approach is probably to identify the most highly interlinked regulators, also known as hubs [24]. It is intuitive that with more target genes of a given regulator in the GRN network, the more important the corresponding regulator is in terms of controlling or 'influencing' the given network or subnetwork, indicating the important role of the hubs in the given cellular activities or diseases.

For instance, by analyzing the hubs, Della Gatta et al. [25] identified a novel tumor suppressor in T cell acute lymphoblastic leukemia (T-ALL) from the inferred TLX interaction networks. The authors first predicted the potential interaction network from T-ALL patient samples using the widely-used ARACNe approach based on mutual information [26]. They further constrained such a huge general network down to a subnetwork by using genes that match with both criteria, not only direct targets of TLX1 and TLX3 based on ChIP-chip results, but also differentially expressed in TLX1- and TLX3-expressing T-ALL samples. Extraction of such a subnetwork greatly helped them to pinpoint the most highly linked regulator, RUNX1, the top ranked hub in that subnetwork, as a novel mediator in T-ALL. Furthermore, both human RUNX1 mutation analysis and murine Runx1 heterozygous knockout mice analysis have confirmed that RUNX1 is indeed a tumor suppressor in T-ALL.

It is also highly attractive for researchers to predict novel key regulators controlling the mode of action for a novel drug. We recently attempted to identify novel key mediators underlying the mode of action for a novel inhibitor Carolacton against the biofilm formation of the lead human oral pathogen *Streptococcus mutans* [27]. We first generated a time-series transcriptome following treatment of Carolacton and predicted the time-shifted correlation network using a Trend Correlation (TC) approach [28]. Furthermore, we combined a time-shifted correlation network with computed regulator-binding site maps and consequently obtained the Carolacton-response GRN. Following this, we identified CodY as the top ranked hub, which is a well-known global master regulator and its deletion in *Streptococcus mutans* signficantly reduces biofilm formation [29]. We thus no longer need to validate CodY. Next to CodY, CysR is the most linked regulator and has been selected as the most promising novel candidate. Interestingly, our experimental results show that deletion of CysR has successfully reduced the sensitivity of *Streptococcus mutans* to Carolacton treatment, indicating a successful prediction of a novel key mediator underlying the mode of action for the inhibitor.

2.2 Biological Feature-Based Prediction

Rather than simply checking the topological features, for example the top hubs of the inferred GRNs, some groups have examined a combination of several biological features for candidate regulators. For instance, Yoself et al. aimed to identify novel regulators controlling Th17 cell differentiation [30]. They first generated a time-series transcriptome dataset while they were reconstructing a general network of regulator-target associations from various published genomics profiles. They then assigned a regulatory interaction between a regulator and a given gene for a certain time window if there is a significant overlap between the regulator's putative targets and the cluster in which the given gene is located. Following the inference of the potential GRN controlling Th17 cell differentiation, they ranked all regulators by emphasizing certain biological features over others (e.g., they first order the regulators based on whether they are predicted regulators of key Th17 genes, then, based on the criteria, whether they are differentially expressed over the time-series data). It makes sense to order in that manner as, apparently, activities of many transcriptional factors are mainly modulated at posttranslational and/or post-transcriptional levels (also refer to the discussion below). With a joint effort from different groups, they could successfully demonstrate at least 4 out of the 12 selected novel factors significantly mediate the Th17 program in knockout mice models or knockdown cells.

2.3 Translated mRNA-Based Prediction

Researchers have recently moved beyond the classic transcriptome-based GRN analysis. As it is now well known that a big fraction of mRNAs might not necessarily be translated into proteins, Brichta et al. have aimed at identifying neurodegenerative factors using the predicted regulatory networks from the so-called translatome, that is all the translated mRNAs [31]. The authors have first reverse engineered a general murine brain regulatory network from a mouse whole brain transcriptome database. They then profiled the translatome by measuring the translated mRNAs specifically expressed in dopaminergic neurons (DNs) of a model of Parkinson's disease in which DNs are under progressive loss. Projecting the differentially expressed translatome to the general murine brain GRNs, the authors predicted 19 candidate regulators which drive the molecular signature mediating the degeneration of DNs at an early stage. Remarkably, none of the 19 are significantly differentially expressed in mRNAs based on classic statistical analyses [32]. As substantia nigra pars compacta DNs are more vulnerable to degeneration than ventral tegmental area DNs, the authors have further tested two predicated regulators which were expressed higher in substantia nigra pars compacta DNs. Interestingly, virus-mediated knockdown of the two selected

genes has successfully caused loss of substantia nigra pars compacta DNs, recapitulating the results obtained in a Parkinson's disease murine model.

2.4 Multilayer Network-Based Prediction

Analysis based on a single layer of network, for example GRN, might be limited because the activities of many transcription factors are mainly tuned at the post-transcriptional and/or posttranslational levels [33, 34]. Therefore, we have recently extended the analysis to integrate multilayer molecular networks. We aim to identify novel factors contributing to survival of CD4+ cells following chronic infection of SIV [35]. To this end, we first integrated a general human molecular network by including GRN, PPI, and signaling transduction networks from various databases. Then we projected the transcriptome measured from human CD4+ cells with chronic infection vs acute infection of SIV to the general integrated molecular network. We predict a regulator as a key regulator only if the number of differentially-expressed putative targets are significantly enriched and meanwhile there exist short length of significantly affected interactive molecular chains in the upstream of the given non-differentially-expressed regulator. Consistent with the discoveries aforementioned by Brichta et al. [31], we predicted 13 'hidden' non-differentially-expressed key regulators by using our network approach, named as Inter-Chain-Finder (ICF). Six out of the 13 predicted key regulators are known to interact with HIV. These predicted key 'hidden' regulators are all involved in the regulation of cell growth, again underscoring the predictive power of the ICF approach. These non-differentially expressed 'hidden' key regulators cannot be detected or prioritized by other classic approaches, indicating the advantages of the ICF approach.

3 PPI Network-Based Key Gene Discovery

3.1 Centrality-Based Approaches

With the development and wide application of high throughput techniques to detect protein–protein interactions, for example yeast two-hybrid (Y2H) and co-immunoprecipitation-mass spectrometry approaches, we are moving toward generating comprehensive PPI networks for various organs, tissues, and cell types. The availability of increasingly comprehensive PPI networks provides us with another great opportunity to dig out novel key genes for a given biological function or process as most, if not all, biological functions can only be realized through physical interactions between various proteins and complexes. The question is which characteristics of the PPI network are unique or helpful to predict key

genes. To address this question one first needs to find out which general characteristics the PPI networks possess for most organisms. In the early contribution of Barabasi and others, it was found that the PPI networks for all the studied organisms share a common characteristic, the scale-free property, namely most nodes have only a few links whereas a few other nodes, the so-called hubs, have a very high degree of connection [1, 36]. Similar to what researchers have done with the GRNs, the first proteins predicted to be disease-relevant are also the hubs or nodes as defined by other centrality measurements [37] as disease genes are more inclined to encode hubs in the PPI networks than non-disease genes [38]. However, this observation could be biased by the fact that researchers are more willing to investigate disease-related genes and consequently mainly identify PPIs for those disease-related genes, which is possibly further supported by all the leading funding bodies. Furthermore, analysis shows that it is mainly lethal-causative or essential disease genes that often cause spontaneous early severe diseases which are more likely to be the hubs than genes mediating other types of diseases [39]. Parallel ideas to distinguish the so-called date and party hubs, where the former appears more dynamic than the latter, might be helpful in identifying non-lethal disease genes [16, 40]. Last but not least, researchers are often interested in which genes/proteins are important for the given biological process or disease of particular interest rather than the general diseased conditions. Therefore, focusing on hubs of the PPI networks alone might not be the best choice to predict key genes for a given biological process. We need to develop better approaches or integrate more data information to predict or prioritize key genes for a given biological process or a given disease from the PPI networks. For instance, one could adapt the newly developed epidemiology analysis approach, the so-called expected force to quantify the spreading power of a protein, to evaluate the effects of the given proteins because the expected force outperforms other centrality measurement, especially for the non-hub nodes [41].

As discussed above, it is apparent that hub- or centrality-based analysis has weaknesses as do all the other approaches. Nevertheless, Wu et al. have recently successfully demonstrated the value of centrality analysis from the extracted network by integrating the general PPI network, knockout transcriptome, and knowledge on the specific signaling transduction pathway [42]. From the general PPI network, the authors have constructed a PPI network model connecting known proteins of the key Th17-stablizing pathway, that is the IL-23R signaling pathway to the transcription factors that have been dysregulated in murine Il23r$^{-/-}$ cells. They then ranked the network nodes by centrality analysis and identified SGK1 as the top 1 ranked node. Using the sgk$^{-/-}$ mice, they have already successfully validated that SGK1 plays a critical role in the induction of pathogenic Th17 cells. Apparently, the essential reason behind their success in applying centrality-based analysis is not attributable to the general PPI network but to the extracted specific PPI network.

3.2 Closeness-Based Approaches

In the last decade, enormous progress has been made in predicting disease genes from PPI networks using approaches based on the widely-accepted 'guilt-by-association' hypothesis [17, 43]. The assumption 'guilt-by-association' can easily be accepted because proteins having similar functions or sharing phenotypes have a high chance of interacting with each other [44] or become 'close' in terms of distance by certain measurements in the given PPI network. There are different ways to measure proximity of proteins in the given PPI network. These methods can be largely clustered into local and global distance measurements [17]. A well-established local distance measurement is the so-called direct neighbor counting (also known as first-degree neighbor counting) approach, which has at least two variants, the absolute count on the number of given disease genes linked to the candidate and the percentage of the number of given disease genes among all the genes directly linked to the candidate in the PPI network [4]. Initially, researchers employed such an approach to predict protein functions [45]. One of the first successful showcases to predict key genes was performed by Oti et al., in which they predicted disease genes by counting the number of proteins known to be linked to the given disease among the first-degree nodes of the candidate gene [46]. As not all the proteins in the same pathways directly interact with each other, the shortest pathway measurement has been introduced to calculate the closeness between the known disease proteins and the candidate proteins. For instance, as pioneers, Krauthammer et al. once used the shortest pathway approach from the constructed PPI network to predict candidate genes in Alzheimer's disease [47]. Later on, Guney and Oliva improved the shortest pathway approach by integrating not only the pathway length but also the disease-associated nodes included in the pathway in the PPI network [48]. They showed a high predictive power in prioritizing top genes related to Alzheimer's disease, diabetes and AIDS.

The global 'closeness' measurements such as random walk with restart and diffusion kernel have also been successfully applied to predict disease genes. For instance, Navlakha and Kingsford once demonstrated that random-walk approaches outperformed clustering and neighbor counting methods when predicting disease-related proteins from the constructed PPI network [49]. Interestingly, the comparison of different 'closeness' calculation approaches demonstrates that not only do different methods predict some unique novel key genes but also the prediction performance varies a lot among different diseases [49]. Therefore, as already proposed, a consensus approach combining various prediction methods into the ensemble method, the random forest classifier, needs to be used for a higher predictive performance. As it has been observed that the distance between known disease genes/proteins is inversely correlated with the predictive performance [49], the successful prediction of key genes is only favorable for some diseases in which the known disease proteins are already observed to interact closely with others. Such a variance among diseases might be biased by the incompleteness of the current human PPI networks and be backed by current research focusing only on

certain diseases. Therefore, given the fact that the current human PPI network is still very incomplete, we need to incorporate further different types of data with the PPI networks, for example genome, transcriptome, proteome, post-protein modifications, and other layers of molecular interactions, and in parallel develop alternative better approaches to predict genes important for a given disease.

A similar challenge to the prediction of disease genes is to infer key genes for a given biological process or function from the PPI networks. As a starting point, one can also use the 'closeness' measurement approaches to predict novel key targets from the PPI networks. Compared with disease gene prediction, there are advantages and disadvantages in predicting genes important for a given biological process. On one hand, there is much more information available on the known genes involved in most of the biological processes, indicating a higher predictive power to infer key genes for most biological processes. On the other hand, genes important for many biological processes except for those essential cellular processes, for example RNA processing and protein synthesis, might have even less chances to occupy the hub nodes than those non-lethal disease genes. It is fairly understandable that imbalance caused by dysfunction of some key genes in certain biological processes might be compensated by other pathways or functions. As a result, identifying non-hub but key genes for a given biological process might be even more challenging.

4 Expression Correlation Network-Guided Key Gene Discovery

We would already be able to predict or prioritize successfully some key genes for a given biological function or process, or even a specific disease, if the GRN and PPI networks were virtually complete and accurate for mammals, especially for human beings. However, the available GRN, PPI, and signaling transduction networks are far from being complete, which severely affects the predictive power no matter which advanced computational approaches we utilize. Fortunately, the proposal to utilize expression correlation networks provides us with alternatives [28, 50, 51]. Expression correlation network (shortened as correlation network) is often purely data-driven, being fully based on large-scale transcriptome data by using various 'distance' measurements, for example the most-widely used Pearson correlation or the variants partial Pearson correlation [52]. If the correlation coefficient between two genes through various conditions or time points is higher than the given thresholds, a link is assigned between the two genes. A systematic caculation for each pair of genes generates a correlation network [8]. It is not always possible to obtain an accurate and complete GRN, PPI, and signaling transduction network for many types of cells and organisms. However, fortunately, with the decreasing cost of microarray measurements and RNA-seq, obtaining a correlation network is almost always possible for any type of cell or organism as long as genome-scale

transcriptional profiling has been performed for a certain number of conditions or time points. Athough undirected, the expression correlation network has its potential, as indicated by the early observation [44] that a high correlation between two human genes has equal or even slightly higher power to predict the possible shared phenotypes than the interacting protein pairs from the highly curated human protein reference databse (HPRD) [53].

As the proteins stably interacting are intended to be coexpressed rather than random pairs [54–56], one could use similar approaches as those applied to the PPI networks, for example 'centralities' and 'closeness'-based methods. For instance, we have recently inferred a correlation network that only exist in CD4+ regulatory T cells (Tregs) but not in CD4+ effector T cells (Teffs) to identify novel key genes for the suppressive function of Tregs [57]. The correlation network is predicted from high-time-resolution time-series transcriptome data of either Tregs or Teffs followed by T cell receptor stimulation by using a combination of two different association calculation approaches allowing time shifts using dynamic programming [28, 58]. Before carrying out any deeper analysis, we first examined the connection degrees of each node in the Treg-specific correlation and ranked all the candidates based on a simple centrality measure, namely connection degrees. Excitingly, the top-1 ranked gene, STUB1 (E3 ubiquitin ligase) was later independently shown to mediate the Treg suppressive function using stub1 knockout mice models [59]. Although this gene has never been emphasized in our original publication, the original table is available at our website (http://wwwen.uni.lu/lcsb/publications/databases_networks_tools). Other groups have also investigated the correlation networks in the cancer research fields. For instance, Yang et al. constructed four correlation networks for four cancer types, respectively [60]. Surprisingly, they found that almost all the prognostic genes are depleted among the hub genes, indicating that it might be not so appropriate to predict prognostic genes using hubs from correlation networks. This difference might be not only because of the different research subjects but also because we used time-series information rather than the static measurement they used. Incorporation of time-shift information into the prediction of correlation edges, which might not exist if calculation is solely based on synchronized correlation coefficients, in reality allows the potential candidates to interact with known key genes from different pathways. This in principle could increase the chances of identifying some novel key genes from different pathways as most genes in the same pathway might be coexpressed in a synchronizing way.

To utilize our Treg-specific correlation network better as a hypothesis generator, we have used the naive Bayesian integration approach to incorporate various information such as closeness measure and expression difference to obtain a uniformed score to rank all the candidate genes [57]. To prioritize the candidate genes (Fig. 1) we used the 'closeness' measure for the given gene in the correlation network, called the queen-bee-surrounding (QBS) principle, defined as the chance (calculated by cumulative binomial distribution tests) to obtain the number of known potential T cell functioning genes among the total number of first-degree neighbors. We have also calculated the expression difference of the candidate gene

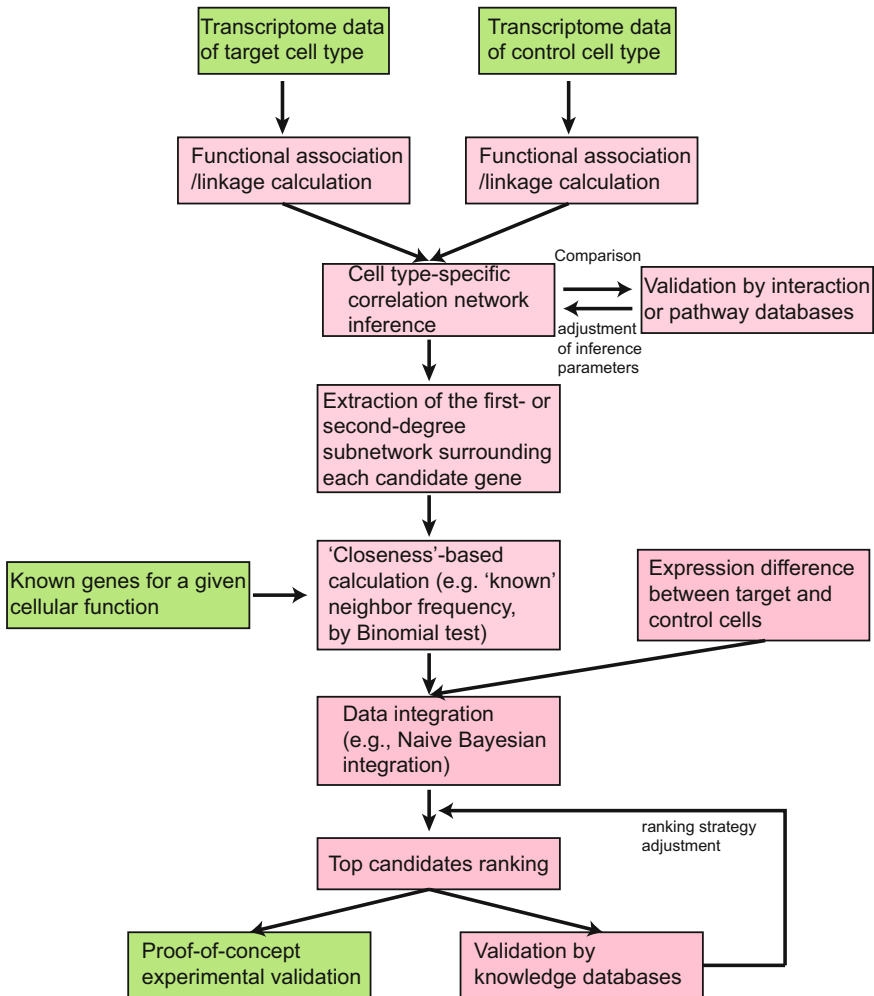**The pipeline of correlation network-based key gene discovery for a given cellular process**

Fig. 1 Integrated pipeline abstracted from the original work [57] to predict key genes for a given cellular process in a specific cell type based on correlation networks inferred from 'omics' data. *Red*, representing computation-orientated steps; *green*, designating experiment-related steps

between Tregs and Teffs over the measured time period (Fig. 1). Remarkably, six out of the top ten ranked key genes predicted from the Treg-specific correlation network are known to be important for the Treg suppressive function or to play a critical role in autoimmune diseases or disorders, encouraging us to validate further the top-ranked novel key genes. Interestingly, we were able to demonstrate successfully that the novel key gene PLAU plays an important role in Treg suppressor function by using specific antibody against PLAU in human Tregs and PLAU$^{-/-}$ murine Tregs [57]. One of the essential advantages of our QBS approach may lay in

Table 1 Comparisons between correlation, GRN and PPI network-based key gene discovery approaches

	Correlation network-based[a]	GRN-based[a]	PPI-network based[a]
Difficulty in network inference or reconstruction	Relatively straightforward	Very difficult in both computational and experimental techniques	Requiring a lot of experimental resources
Centrality-based predictions	Difference is shown when time-series data is utilized compared with static measurements; application has been shown using cell-type-specific networks.	Performance is very good combining ChIP-seq or transcription binding sites, or other layers of molecular networks.	Performance is very good when integrating conditional specific information (e.g., knockout transcriptome); performance is better for predicting severe lethal diseases.
'Closeness'-based predictions	Performance is very good when integrating expression difference information and cell-type-specific network.	Not done yet	Performance is dependent on diseases and various 'closeness' measurements.
Consensus approach-based predictions	Not done yet	Not done yet	Performance is better than that of a single approach.
Biological-feature-based predictions	Not done yet	Performance is very good	Not done yet
Ability to identify key transcription factors	Not done yet (regulatory network layers to be incorporated)	Yes	Possible
Ability to identify upstream key genes of transcription factors	Possible	Impossible	Possible

[a]The comparisons are based on the whole review and therefore we did not mention a particular reference here

its ability to prioritize key novel genes, for example the gene PLAU that does not show a very strong expression difference between Tregs and Teffs, which thus cannot easily stand out by classic differential-expression analysis. As demonstrated here, another advantage of correlation network-based approaches is the ability to predict key genes in the upstream of transcription factors, for example signaling proteins (here PLAU), which is almost impossible by the GRN-based approaches (Table 1). As both centrality- and 'closeness'-based approaches were able to predict some interesting novel key genes as exemplified in our Treg correlation network, it might be worth checking whether it increases the predictive power by combing both approaches into a consensus. So far, through the comparisons between correlation, GRN, and PPI network-based key gene prediction approaches, correlation network-

based methods do not display obvious weaknesses, even if not showing apparent advantages over the others (Table 1). Therefore, for some cases it is beneficial to explore the use of correlation network-based key gene discoveries, as the inference of correlation network is much easier than that of others.

5 Concluding Remarks and Future Directions

The exponentially-increasing amount of genomic and functional genomic data, such as coding and noncoding transcriptome (e.g., from microarray, RNA-seq, and microRNA assays), proteome, metabolome, ChIP-seq, and others, empower us with the possibility to unveil better than ever the cellular components, their interactions, and the resulted molecular networks. The next essential task is to interpret properly and make use of those large-scale datasets. To this end, many excellent computational approaches, some of which we have already briefly discussed here, have been developed to analyze those large-scale datasets. Some of them allow us to infer or reconstruct molecular networks from large-scale 'omics' datasets. Following the network reconstruction or inference, some of the reports have already successfully demonstrated the application value of various types of molecular networks. One of the essential but challenging applications for systems biologists is to predict key genes for a given biological process or, eventually, for a specific disease. At first sight, predicting key genes seems to be quite paradoxical because the way we predict key genes almost reverses the route by which we infer networks. In fact, network prediction starts from a huge number of potential genes and one does not know whether and how these genes interact or regulate. However, key gene prediction is to seek and position precisely a few genes or nodes among complex networks with a huge number of nodes and interactions. As we have already discussed above, several network analysis approaches allow us to predict novel key genes which are not easily made to stand out or even detected by classic differential-expression-based statistical analyses. However, we are still a long way off from being able to apply network analysis to predict successfully novel drug targets or novel prognostic or diagnostic markers for a given disease. Possibly because of the extremely long period required for drug development, including all the phases of clinical trials (at least 10 years), to date there has been no successful report on network-based novel drug target identification and subsequent drug development.

To predict more successfully novel key genes for a given cellular process from the inferred or reconstructed molecular networks, we believe the following directions should be further developed:

1. First of all, context- and cell-type-specific network inference needs to be developed further by integrating various condition- and type-specific and various layers of 'omics' data into the general molecular networks. The context- and cell-type-specific networks should be highly relevant to human disease gene

prediction as particular human diseases are mainly driven by specific cell types, at least down to specific tissues [61]. For instance, it is also only relevant to utilize CD4+ T cell-specific networks rather than fibroblast cellular networks to infer key genes important for CD4+ T cell differentiation and function.

2. We need to infer or reconstruct personalized or at least patient-subgroup-related molecular networks as many patients with the same disease can be further stratified into several subgroups based on their molecular patterns [62, 63]. For instance, responders to the promising tumor immunotherapy checkpoint inhibitors apparently have very different molecular patterns than those non-responders [64, 65]. It only makes sense to predict novel key drug targets from molecular networks inferred from samples of non-responders rather than a general tumor molecular network. Current efforts have already started to predict disease differential networks [66, 67], which might further increase the power to predict some disease key genes. A further step to finely divide the disease-specific networks or disease maps [68] into disease-subgroup networks is needed, as discussed above.

3. A few examples of pioneering work [35, 69] have already been carried out to integrate multiple layer molecular networks to infer novel key nodes for a given biological process or specific disease. Nevertheless, there are still many improvements needed and a few challenging questions to be addressed. For instance, what is the minimal amount of information required for successful key gene prediction? The more data we use, the better the predictive power we can achieve. Possibly, the answer is not that simple, in that many different molecular layers of data measured might generate redundant information, unfortunately introducing more noise to the hide true signals. Another quite new and challenging problem is how to integrate various layers of molecular networks and multilayer 'omics' data to predict novel key targets, as in our body all these layers of networks intertwine together per se.

We are very confident that, with the further development of computational and experimental techniques in revealing cell-specific and patient-specific cellular networks, we can soon better predict and efficiently disclose novel key genes or drug targets mediating any given biological process or function or a particular disease by means of a combination of data-driven and knowledge-driven network approaches, rather than classic trial-and-error methods. The network-based strategy to discover key genes for a given cellular process should be the next driving force to revolutionize biomedicine research, complementing the well-established GWAS or QTL-based approaches or the newly-developed resource-intensive systemic phenotyping of murine mutants which, although with clear caveats [70], have already substantially leveraged the systemic identification of key genes or loci associated with a given disease or phenotype or symptom [71–76].

Completing Interests Statement The authors declare that they have no competing financial interests.

References

1. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113
2. Kitano H (2002) Systems biology: a brief overview. Science 295(5560):1662–1664
3. Ma'ayan A (2011) Introduction to network analysis in systems biology. Sci Signal 4(190):tr5
4. Barabasi AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12(1):56–68
5. Meier-Schellersheim M, Fraser IDC, Klauschen F (2009) Multi-scale modeling in cell biology. Wiley Interdiscip Rev Syst Biol Med 1(1):4–14
6. Mizeranschi A, Groen D, Borgdorff J, Hoekstra AG, Chopard B, Dubitzky W (2016) Anatomy and physiology of multiscale modeling and simulation in systems medicine. Methods Mol Biol 1386:375–404
7. Karlebach G, Shamir R (2008) Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol 9(10):770–780
8. He F, Balling R, Zeng AP (2009) Reverse engineering and verification of gene networks: principles, assumptions, and limitations of present methods and future perspectives. J Biotechnol 144(3):190–203
9. Emmert-Streib F, Glazko GV, Altay G, de Matos Simoes R (2012) Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Front Genet 3:8
10. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D (2007) How to infer gene networks from expression profiles. Mol Syst Biol 3:78
11. Trinh CT, Wlaschin A, Srienc F (2009) Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. Appl Microbiol Biotechnol 81(5):813–826
12. Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat Biotechnol 18(3):326–332
13. Papin JA, Price ND, Edwards JS, Palsson BB (2002) The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy. J Theor Biol 215(1):67–82
14. Forster J, Famili I, Palsson BO, Nielsen J (2003) Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. OMICS 7(2):193–202
15. Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, Shlomi T (2011) Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol 7:501
16. Vidal M, Cusick ME, Barabasi AL (2011) Interactome networks and human disease. Cell 144(6):986–998
17. Wang X, Gulbahce N, Yu H (2011) Network-based methods for human disease gene prediction. Brief Funct Genomics 10(5):280–293
18. Banerjee N, Zhang MQ (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. Nucleic Acids Res 31(23):7024–7031
19. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74

20. He F, Buer J, Zeng A-P, Balling R (2007) Dynamic cumulative activity of transcription factors as a mechanism of quantitative gene regulation. Genome Biol 8(9):1–18

21. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Consortium D, Kellis M, Collins JJ, Stolovitzky G (2012) Wisdom of crowds for robust gene network inference. Nat Methods 9(8):796–804

22. Maetschke SR, Madhamshettiwar PB, Davis MJ, Ragan MA (2014) Supervised, semi-supervised and unsupervised inference of gene regulatory networks. Brief Bioinform 15(2): 195–211

23. Gitter A, Carmi M, Barkai N, Bar-Joseph Z (2013) Linking the signaling cascades and dynamic regulatory networks controlling stress responses. Genome Res 23(2):365–376

24. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431(7006): 308–312

25. Della GG, Palomero T, Perez-Garcia A, Ambesi-Impiombato A, Bansal M, Carpenter ZW, De Keersmaecker K, Sole X, Xu L, Paietta E, Racevskis J, Wiernik PH, Rowe JM, Meijerink JP, Califano A, Ferrando AA (2012) Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL. Nat Med 18(3):436–440

26. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. Nat Genet 37(4):382–390

27. Sudhakar P, Reck M, Wang W, He FQ, Wagner-Dobler I, Zeng AP (2014) Construction and verification of the transcriptional regulatory response network of Streptococcus mutans upon treatment with the biofilm inhibitor carolacton. BMC Genomics 15:362

28. He F, Zeng AP (2006) In search of functional association from time-series microarray data based on the change trend and level of gene expression. BMC Bioinformatics 7:69

29. Lemos JA, Nascimento MM, Lin VK, Abranches J, Burne RA (2008) Global regulation by (p) ppGpp and CodY in Streptococcus mutans. J Bacteriol 190(15):5291–5299

30. Yosef N, Shalek AK, Gaublomme JT, Jin H, Lee Y, Awasthi A, Wu C, Karwacz K, Xiao S, Jorgolli M, Gennert D, Satija R, Shakya A, Lu DY, Trombetta JJ, Pillai MR, Ratcliffe PJ, Coleman ML, Bix M, Tantin D, Park H, Kuchroo VK, Regev A (2013) Dynamic regulatory network controlling TH17 cell differentiation. Nature 496(7446):461–468

31. Brichta L, Shin W, Jackson-Lewis V, Blesa J, Yap EL, Walker Z, Zhang J, Roussarie JP, Alvarez MJ, Califano A, Przedborski S, Greengard P (2015) Identification of neurodegener-ative factors using translatome-regulatory network analysis. Nat Neurosci 18(9):1325–1333

32. Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 7(1):55–65

33. Tootle TL, Rebay I (2005) Post-translational modifications influence transcription factor activity: a view from the ETS superfamily. Bioessays 27(3):285–298

34. Bar-Joseph Z, Gitter A, Simon I (2012) Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet 13(8):552–564

35. He FQ, Sauermann U, Beer C, Winkelmann S, Yu Z, Sopper S, Zeng AP, Wirth M (2014) Identification of molecular sub-networks associated with cell survival in a chronically SIVmac-infected human CD4+ T cell line. Virol J 11:152

36. Seebacher J, Gavin AC (2011) SnapShot: protein-protein interaction networks. Cell 144(6): 1000, 1000 e1

37. Koschutzki D, Schreiber F (2008) Centrality analysis methods for biological networks and their application to gene regulatory networks. Gene Regul Syst Bio 2:193–201

38. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) The human disease network. Proc Natl Acad Sci U S A 104(21):8685–8690

39. Feldman I, Rzhetsky A, Vitkup D (2008) Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A 105(11):4323–4328

40. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430(6995):88–93

41. Lawyer G (2015) Understanding the influence of all nodes in a network. Sci Rep 5:8665
42. Wu C, Yosef N, Thalhamer T, Zhu C, Xiao S, Kishi Y, Regev A, Kuchroo VK (2013) Induction of pathogenic TH17 cells by inducible salt-sensing kinase SGK1. Nature 496 (7446):513–517
43. Wolfe CJ, Kohane IS, Butte AJ (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics 6:227
44. Fraser HB, Plotkin JB (2007) Using protein complexes to predict phenotypic effects of gene mutation. Genome Biol 8(11):R252
45. Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol 3:88
46. Oti M, Snel B, Huynen MA, Brunner HG (2006) Predicting disease genes using protein-protein interactions. J Med Genet 43(8):691–698
47. Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A (2004) Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. Proc Natl Acad Sci U S A 101(42):15148–15153
48. Guney E, Oliva B (2012) Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. PLoS One 7(9), e43557
49. Navlakha S, Kingsford C (2010) The power of protein interaction networks for associating genes with diseases. Bioinformatics 26(8):1057–1063
50. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559
51. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4:Article17
52. de la Fuente A, Bing N, Hoeschele I, Mendes P (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics 20(18):3565–3574
53. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A (2009) Human Protein Reference Database—2009 update. Nucleic Acids Res 37(Database issue):D767–D772
54. Jansen R, Greenbaum D, Gerstein M (2002) Relating whole-genome expression data with protein-protein interactions. Genome Res 12(1):37–46
55. Ge H, Liu Z, Church GM, Vidal M (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet 29(4):482–486
56. Bhardwaj N, Lu H (2005) Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics 21(11):2730–2738
57. He F, Chen H, Probst-Kepper M, Geffers R, Eifes S, Del Sol A, Schughart K, Zeng AP, Balling R (2012) PLAU inferred from a correlation network is critical for suppressor function of regulatory T cells. Mol Syst Biol 8:624
58. Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. J Mol Biol 314(5):1053–1066
59. Chen Z, Barbi J, Bu S, Yang HY, Li Z, Gao Y, Jinasena D, Fu J, Lin F, Chen C, Zhang J, Yu N, Li X, Shan Z, Nie J, Gao Z, Tian H, Li Y, Yao Z, Zheng Y, Park BV, Pan Z, Zhang J, Dang E, Li Z, Wang H, Luo W, Li L, Semenza GL, Zheng SG, Loser K, Tsun A, Greene MI, Pardoll DM, Pan F, Li B (2013) The ubiquitin ligase Stub1 negatively modulates regulatory T cell suppressive activity by promoting degradation of the transcription factor Foxp3. Immunity 39(2): 272–285
60. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H (2014) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nat Commun 5:3231

61. Guan Y, Gorenshteyn D, Burmeister M, Wong AK, Schimenti JC, Handel MA, Bult CJ, Hibbs MA, Troyanskaya OG (2012) Tissue-specific functional networks for prioritizing phenotype and disease genes. PLoS Comput Biol 8(9):e1002694

62. Trusheim MR, Berndt ER, Douglas FL (2007) Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. Nat Rev Drug Discov 6(4):287–293

63. Hofree M, Shen JP, Carter H, Gross A, Ideker T (2013) Network-based stratification of tumor mutations. Nat Methods 10(11):1108–1115

64. Glinsky GV, Berezovska O, Glinskii AB (2005) Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. J Clin Invest 115(6):1503–1521

65. Chang YH, Chen CM, Chen HY, Yang PC (2015) Pathway-based gene signatures predicting clinical outcome of lung adenocarcinoma. Sci Rep 5:10979

66. de la Fuente A (2010) From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends Genet 26(7):326–333

67. Zickenrott S, Angarica VE, Upadhyaya BB, Del Sol A (2016) Prediction of disease-gene-drug relationships following a differential network analysis. Cell Death Dis 7, e2040

68. Fujita KA, Ostaszewski M, Matsuoka Y, Ghosh S, Glaab E, Trefois C, Crespo I, Perumal TM, Jurkowski W, Antony PM, Diederich N, Buttini M, Kodama A, Satagopam VP, Eifes S, Del Sol A, Schneider R, Kitano H, Balling R (2014) Integrating pathways of Parkinson's disease in a molecular interaction map. Mol Neurobiol 49(1):88–102

69. Chen Y, Wu X, Jiang R (2013) Integrating human omics data to prioritize candidate genes. BMC Med Genomics 6:57

70. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, Barrett JC, Biesecker LG, Conrad DF, Cooper GM, Cox NJ, Daly MJ, Gerstein MB, Goldstein DB, Hirschhorn JN, Leal SM, Pennacchio LA, Stamatoyannopoulos JA, Sunyaev SR, Valle D, Voight BF, Winckler W, Gunter C (2014) Guidelines for investigating causality of sequence variants in human disease. Nature 508 (7497):469–476

71. Weidinger S, Gieger C, Rodriguez E, Baurecht H, Mempel M, Klopp N, Gohlke H, Wagenpfeil S, Ollert M, Ring J, Behrendt H, Heinrich J, Novak N, Bieber T, Kramer U, Berdel D, von Berg A, Bauer CP, Herbarth O, Koletzko S, Prokisch H, Mehta D, Meitinger T, Depner M, von Mutius E, Liang L, Moffatt M, Cookson W, Kabesch M, Wichmann HE, Illig T (2008) Genome-wide scan on total serum IgE levels identifies FCER1A as novel susceptibility locus. PLoS Genet 4(8):e1000166

72. Bush WS, Moore JH (2012) Chapter 11: genome-wide association studies. PLoS Comput Biol 8(12), e1002822

73. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6(2):95–108

74. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet 37(3):233–242

75. Civelek M, Lusis AJ (2014) Systems genetics approaches to understand complex traits. Nat Rev Genet 15(1):34–48

76. Hrabe de Angelis M, Nicholson G, Selloum M, White JK, Morgan H, Ramirez-Solis R, Sorg T, Wells S, Fuchs H, Fray M, Adams DJ, Adams NC, Adler T, Aguilar-Pimentel A, Ali-Hadji D, Amann G, Andre P, Atkins S, Auburtin A, Ayadi A, Becker J, Becker L, Bedu E, Bekeredjian R, Birling MC, Blake A, Bottomley J, Bowl MR, Brault V, Busch DH, Bussell JN, Calzada-Wack J, Cater H, Champy MF, Charles P, Chevalier C, Chiani F, Codner GF, Combe R, Cox R, Dalloneau E, Dierich A, Di Fenza A, Doe B, Duchon A, Eickelberg O, Esapa CT, Fertak LE, Feigel T, Emelyanova I, Estabel J, Favor J, Flenniken A, Gambadoro A, Garrett L, Gates H, Gerdin AK, Gkoutos G, Greenaway S, Glasl L, Goetz P, Da Cruz IG, Gotz A, Graw J, Guimond A, Hans W, Hicks G, Holter SM, Hofler H, Hancock JM, Hoehndorf R, Hough T,

Houghton R, Hurt A, Ivandic B, Jacobs H, Jacquot S, Jones N, Karp NA, Katus HA, Kitchen S, Klein-Rodewald T, Klingenspor M, Klopstock T, Lalanne V, Leblanc S, Lengger C, le Marchand E, Ludwig T, Lux A, McKerlie C, Maier H, Mandel JL, Marschall S, Mark M, Melvin DG, Meziane H, Micklich K, Mittelhauser C, Monassier L, Moulaert D, Muller S, Naton B, Neff F, Nolan PM, Nutter LM, Ollert M, Pavlovic G, Pellegata NS, Peter E, Petit-Demouliere B, Pickard A, Podrini C, Potter P, Pouilly L, Puk O, Richardson D, Rousseau S, Quintanilla-Fend L, Quwailid MM, Racz I, Rathkolb B, Riet F, Rossant J, Roux M, Rozman J, Ryder E, Salisbury J, Santos L, Schable KH, Schiller E, Schrewe A, Schulz H, Steinkamp R, Simon M, Stewart M, Stoger C, Stoger T, Sun M, Sunter D, Teboul L, Tilly I, Tocchini-Valentini GP, Tost M, Treise I, Vasseur L, Velot E, Vogt-Weisenhorn D, Wagner C, Walling A, Wattenhofer-Donze M, Weber B, Wendling O, Westerberg H, Willershauser M, Wolf E, Wolter A, Wood J, Wurst W, Yildirim AO, Zeh R, Zimmer A, Zimprich A, Consortium E, Holmes C, Steel KP, Herault Y, Gailus-Durner V, Mallon AM, Brown SD (2015) Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics. Nat Genet 47(9):969–978