Gene network analysis: from heart development to cardiac therapy

Fulvia Ferrazzi¹; Riccardo Bellazzi¹; Felix B. Engel¹

¹Institute of Human Genetics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany; ²Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy; ³Experimental Renal and Cardiovascular Research, Department of Nephropathology, Institute of Pathology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Summary
Networks offer a flexible framework to represent and analyse the complex interactions between components of cellular systems. In particular, gene networks inferred from expression data can support the identification of novel hypotheses on regulatory processes. In this review, we focus on the use of gene network analysis in the study of heart development. Understanding heart development will promote the elucidation of the aetiology of congenital heart disease and thus possibly improve diagnostics. Moreover, it will help to establish cardiac therapies. For example, understanding cardiac differentiation during development will help to guide stem cell differentiation required for cardiac tissue engineering or to enhance endogenous repair mechanisms. We introduce different methodological frameworks to infer networks from expression data such as Boolean and Bayesian networks. Then we present currently available temporal expression data in heart development and discuss the use of network-based approaches in published studies. Collectively, our literature-based analysis indicates that gene network analysis constitutes a promising opportunity to infer therapy-relevant regulatory processes in heart development. However, the use of network-based approaches has so far been limited by the small amount of samples in available datasets. Thus, we propose to acquire high-resolution temporal expression data to improve the mathematical descriptions of regulatory processes obtained with gene network inference methodologies. Especially probabilistic methods that accommodate the intrinsic variability of biological systems have the potential to contribute to a deeper understanding of heart development.

Keywords
Gene networks, temporal expression data, reverse engineering, heart development

Introduction
Systems biology aims at analysing biological systems through the study of the complex interplay of cellular components, including genes, proteins and metabolites (1–3). It is well agreed that there are at least three elements that are necessary to systems biology: parts list, i.e. the list of elements to be studied, interactions, i.e. the network of interactions between the elements in the parts list, and finally the dynamical model, i.e. the mathematical description of temporal relationships between such variables (4). Thus, the notion of network plays a central role in systems biology research, representing the way to synthesize the complexity and offering a structure amenable to be exploited by different data analysis algorithms. Networks can be derived by combining data and knowledge already collected in biological data repositories. In particular, systems biology can nowadays benefit from data generated by high-throughput technologies, microarrays and deep sequencing, which make it possible to investigate cellular systems at a genome-wide level and with increasing resolution. We will mainly concentrate on gene expression data and, thus, we will limit our discussion to gene networks. In this area, large-scale expression analyses allow observing the transcriptional response of systems to environmental perturbations as well as during development and differentiation processes.

Expression data analysis typically involves identifying differentially expressed genes and employing functional annotation of known genes to associate the differentially expressed genes with the over-represented biological processes in which they are involved. In addition, clustering algorithms can be employed to group genes with similar expression profiles. This helps to identify novel candidate genes according to the “guilt-by-association” principle, i.e. the hypothesis that genes with similar expression profiles are likely to be involved in the same biological processes (5). Gene network analysis moves one step forward and aims at supporting...
the identification of the regulatory interactions between genes. The main assumption underlying the inference of gene networks from expression data is that a regulatory relationship between genes generates an observable statistical dependency between their respective expression levels (6). It is important to consider that statistical dependencies might arise not only as the result of an actual regulatory relationship of a gene on another (i.e. a transcription factor regulating its target gene). Indeed a dependency between the expression levels of two genes A and B might also be the effect of an indirect regulation (gene A regulates gene C, which in turn regulates gene B) or be the effect of a common, unobserved cause (gene D regulates genes A and B). It is unlikely that the application of the methodologies for network inference from expression data will result in the identification of the whole set of actual regulatory processes. However, they can significantly contribute to this goal by identifying sets of genes linked by potential cause-effect relationships. Thus, network analysis can be considered as a useful tool to provide novel biological hypotheses that can be specifically tested by biological experiments.

A large variety of methods have been proposed in the literature to build gene expression networks. Once the network has been inferred it is possible to rely on computational techniques to analyse the system of interest. The analysis of the structure of these networks can reveal key genes in the regulation, such as for example those involved in feedback loops or the so-called "hub" genes (7, 8). Given the very nature of gene expression data and the complexity of the biological processes underlying gene regulation, some authors have proposed to derive "module networks", i.e. networks in which the expression of several co-expressed genes are clustered in one module (9–11). Moreover one can also try to relate such modules with components of the pathways involved in specific cell activities.

As already mentioned, the third step of a systems biology approach is to describe the dynamics of the system under study with a suitable computational strategy. When network edges are directional and an explicit mathematical function (deterministic or probabilistic) is employed to describe the represented regulations, it is possible to rely on computational simulations to perform in silico experiments aimed at identifying novel biological hypotheses. It is for example possible to simulate the knock down of selected genes and rank them on the basis of the magnitude of the corresponding effect on the network, thus effectively prioritising them for experimental validation.

This review will discuss the use of network-based approaches to study gene expression in cardiac development. First, methodologies for the analysis of high-throughput expression data will be introduced with a focus on gene networks. Then, currently available expression datasets relevant for heart development will be presented and published studies using network analysis to elucidate heart development will be discussed. Conclusions will highlight perspectives for future research.

From heart development to treatment of cardiac disease

Cardiovascular diseases (CVDs) present a major socio-economic burden that has been constantly rising (12). In addition, congenital heart disease (CHD) is the most common birth defect (13). Thus, there is an urgent need for the development of novel therapies. A better understanding of heart development will help in elucidating the aetiology of congenital heart disease and thus possibly improve diagnostics (14). Moreover, it has the potential to improve current therapies or to aid in the development of novel therapies to treat CVDs and CHD (15–17). For example, the single leading cause of death in CVDs is ischaemic heart disease (12). Therefore, therapies that protect the heart under hypoxic conditions would be of great value. In contrast to the adult heart, the foetal heart develops under hypoxic conditions (18, 19). Another treatment option is accelerating vascularisation/neo-angiogenesis. One of the most publicised approaches to treat cardiac disease is the use of stem cell-based therapies (15, 20); a major limitation of these approaches is the inefficient differentiation of stem cells into mature cardiomyocytes (21). Finally, it has been shown that zebrafish (22), newts (23) and possibly also newborn mice (24) can regenerate the heart based on cardiomyocyte proliferation, a process that also underlies heart growth during development (25).

Mathematical and statistical methods for gene expression data analysis: from clustering to networks

Two categories of technological platforms are currently available for high-throughput expression analysis: microarrays and high-throughput sequencing. Microarrays are a mature technology, while high-throughput sequencing is more recent and is constantly being improved. These technologies have been reviewed in detail elsewhere (for example [26–28]). High-throughput sequencing offers some significant advantages over microarrays, mainly the much greater dynamic range of measurable expression values and the possibility to detect also previously unknown transcripts. The initially high costs of high-throughput sequencing have been steadily decreasing, and therefore sequencing appears to be the elective technology for a deep understanding of expression patterns. Nonetheless, microarrays are still of great interest for research. For example, when the project’s goal is to assay expression at the gene level concentrating on the known transcriptome of the studied organism, microarrays constitute an effective platform. A variety of well-established techniques are indeed available for the pre-processing of raw data into numerical expression values associated with different transcripts, which allows concentrating analysis efforts on higher-level processing.

Expression data can be static, such as snapshots of gene expression in different environmental conditions, or dynamic, such as expression profiles measured at different time points during developmental or differentiation processes. Typically the first step in the post-processing of both expression data types is to identify
differentially expressed genes, i.e. genes that exhibit significant changes of expression in the examined conditions. In the case of temporal data, it is advisable to use differential analysis methods specifically designed for time series since they explicitly take into account the dependencies existing across time points (29, 30). This allows them to have more power in detecting dynamic differential expression, such as for example when genes exhibit monotonic increasing or decreasing profiles. The number of differentially expressed genes often ranges from a few tens to a few hundreds. Clustering constitutes a useful method to identify sets of genes with similar expression profiles and facilitate exploration of the differentially expressed gene list. Profiles in a cluster should be as similar as possible to each other (“homogeneity”) and as different as possible from profiles in other clusters (“separation”). Also in this case, clustering techniques specifically designed for temporal data have been developed (31–34). Plots of temporal expression profiles in each cluster or alternatively heatmap-based representations constitute an intuitive graphical display of clustering results. After clustering, functional enrichment analysis is typically employed in order to identify the over-represented biological functions associated with the cluster genes and to make hypotheses on novel candidate genes on the basis of guilt-by-association.

Gene networks aim at representing and analysing potential regulatory interactions between genes. Gene networks are graphically represented as a set of nodes, associated with gene expression levels, connected by edges representing the existence of an interaction between the connected nodes. Different paradigms have been proposed in the literature to build gene networks from expression data (10, 35, 36). In the case of “association networks” or “co-expression networks”, network construction starts with the calculation of pairwise similarity measures between gene expression data, such as Pearson’s correlation. Then the significant interactions are extracted, either by establishing an arbitrary threshold for the measured correlations or by calculating an empirical significance for example through a permutation-based approach. The inferred network of interactions allows the identification of co-expression modules. Analysis of co-expression networks can provide useful insights into the complex regulatory mechanisms underlying the analysed biological process (10). Yet these networks lack a dynamical model of the represented interactions and thus do not allow performing in silico experiments.

In order to infer both the network of interactions between genes and a dynamical model of the interactions, model based network inference methodologies rely on the assumption that a real gene regulatory network exists, and that such network generated the observed expression data. The aim is therefore to “reverse engineer” this network starting from the data itself. Reverse engineering thus requires testing different hypotheses on gene regulatory relationships and selecting the model that can best represent the observed data (Figure 1). A large variety of reverse engineering approaches have been proposed in the literature, both deterministic and probabilistic. The proposed methodologies employ a wide range of mathematical frameworks to model gene interactions, ranging for example from Boolean functions to detailed differential equation models (35). In this review we will summarize the key features of two paradigms, namely Boolean networks and Bayesian networks, which have been increasingly used in the context of gene network modelling (36–38).

In Boolean networks gene expression values are discretized and associated with binary (on/off) variables while dependencies between expression values are represented by Boolean functions. The goal of methodologies to infer Boolean networks is to identify a rule able to represent the expression level of a given gene as a function of the levels of other genes using combinations of the logical operations AND-OR-NOT (Figure 2). Both deterministic and probabilistic Boolean models have been proposed for gene network modelling, since early work by Somogyi and Sniegoski (39) and Shmulevich et al. (40). A prototype approach to reverse

| gene | e1 | e2 | e3 | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>off</td>
<td>off</td>
<td>on</td>
<td>...</td>
</tr>
<tr>
<td>Y</td>
<td>off</td>
<td>on</td>
<td>on</td>
<td>...</td>
</tr>
<tr>
<td>Z</td>
<td>off</td>
<td>on</td>
<td>on</td>
<td>...</td>
</tr>
</tbody>
</table>

Figure 1: Reverse engineering networks from expression data involves testing different hypotheses on regulation. Assuming a toy example in which expression values for three genes (X, Y, and Z) have been measured (A), four different regulation scenarios for Z are possible when self-regulation is not considered (B): Z is regulated by neither X nor Y (a), regulated by X alone (b), by Y alone (c), or by X and Y jointly (d).
engineer Boolean models from gene expression data is the algorithm REVEAL (REVerse Engineering Algorithm) (41). REVEAL relies on the information theory concepts of entropy and mutual information to identify a minimum set of input genes able to predict the behaviour of a given output gene. The more recent algorithm ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) relies on entropy extended to continuous domain, in order to avoid discretisation of gene expression data (42, 43).

Bayesian networks (BNs) describe gene expression values by means of random variables and model the dependencies between them by means of conditional probabilities. The probabilistic framework of BNs is indeed very suited to model gene networks as it allows taking into account the intrinsic variability of cellular systems and the presence of noise in the data (44). A BN is completely specified by a directed acyclic graph (DAG), constituting the network structure, and the set of conditional probabilities of each node in the graph given its parent nodes. BNs can be inferred from both discrete and continuous expression values, and are not limited to binary values in the former case. Since the use of BNs to model expression data was proposed by Friedman et al. (45), they have received increasing interest. Inference of BNs from data has the goal of finding the network model with the highest probability with respect to the measured expression data. To this aim, different approaches have been proposed in the literature (37). An interesting strategy to learn BNs from expression data is given by the module network approach by Segal et al. (11, 46, 47), which is based on the assumption that the dynamic behaviours of co-expressed genes are not distinguishable. These genes are thus grouped into a unique variable before network inference. In the case of temporal expression data, dynamic Bayesian networks (DBNs), an extension of BNs to take into account the dynamics of a system, are particular suitable (48–50).

As mentioned above, the availability of a mathematical model to describe gene relationships is a key feature that distinguishes model based network inference approaches from association networks. Mathematical models allow using the inferred network for simulations aimed at predicting different scenarios, thus effectively enabling in silico experiments. In the case of BNs, this is achieved through probabilistic inference. For example, when the values of some variables in the system are known (this is called evidence), it is possible to make inferences about the probability distribution of other variables in the system by propagating this evidence.

All reverse engineering methods suffer from the low amount of samples usually contained in available datasets, which often leads to a high number of spurious inferred dependencies. This effect is strengthened by the presence of experimental and technical noise in expression measurements that can have a profound effect of the quality of the data. Relying on multiple expression datasets relative to the same biological process has the potential to improve the accuracy and robustness of inferred models, as shown in the context of BN learning by Steele and Tucker (51) and by Werhli and Husmeier (52).

Another powerful strategy to improve the accuracy of inferred network models is to integrate prior knowledge and different types of data in network learning (30, 35, 53). In the Bayesian network framework, the incorporation of prior knowledge and different data can be done in a principled way, by leveraging the possibility to specify suitable priors for the network models to be explored (35). For example Bernard and Hartemink relied on genome-wide transcription factor binding location data to assign prior probabilities to network edges in a DBN model (54). Praveen and Fröhlich proposed two computational strategies to integrate multiple heterogeneous knowledge sources into a probabilistic structure prior, one assuming a high degree of correlation between the sources and the other capturing the strongest support for an interaction (55).

In large-scale expression datasets the number of genes is typically in the order of tens of thousands while the number of samples is in the order of a few tens or a few hundreds. This poses important challenges to the application of BNs, since both probabilistic inference and learning from data are NP-hard problems (56, 57). As concerns BN learning, the number of possible DAGs is super-exponential in the number of variables and this makes searching across all possible network structures unfeasible even for a very small number of genes. To tackle this issue several heuristic strategies have been proposed, which aim at reducing the search space by setting constraints on the network structures to be explored (57). Notable examples are the K2 algorithm by Cooper and Herskovits (58), the sparse candidate algorithm by Friedman et al. (59), and the structure learning algorithm with L1-regularisation by Schmidt et al. (60).

When the amount of samples is small compared to the number of variables, there are likely many networks able to explain the data reasonably well (61), thus the use of sampling-based approaches has been proposed (45, 61–63). It is then possible to extract from the inferred models the higher-confidence structural features.

An interesting area of research is the development of methodologies that are able to effectively exploit the increasing amount of available genomics datasets and prior knowledge to learn large-scale gene networks. An example is the ScanBMA algorithm by Young et al. (64), which relies on a Bayesian inference scheme that integrates temporal expression data with prior information on regulatory relationships and on an efficient strategy to search the model space.

**Available temporal expression datasets in heart development**

In order to identify publicly available high-throughput expression datasets related to heart development, we relied on the ArrayExpress database, a repository for functional genomics data maintained at the European Bioinformatics Institute (65). ArrayExpress contains both datasets directly submitted to it and datasets which are imported, on a weekly basis, from Gene Expression Omnibus (GEO) database, the functional genomics repository maintained at NCBI (66). We searched ArrayExpress database in May 2014 using the terms “heart development”, “cardiac development”, “cardiovascular.
development”, “myocardial development”, and “myocardium development”. This yielded in total 80 datasets, which we inspected to identify those belonging to one of the following categories most relevant for heart development: temporal expression studies performed in vivo at different time points along embryonic (E) and postnatal (P) development (Table 1) or temporal expression studies related to stem cell differentiation towards cardiomyocytes (Table 2).

Out of the identified temporal expression datasets the majority have been measured in murine or rat models. Microarray expression analyses during heart development have been performed in the early 2000s in the context of the CardioGenomics Project, profiling mouse embryos from E10.5 to E14.5 at daily intervals as well as E16.5 and E18.5 (seven time points in total). From E12.5 onwards, ventricles and atrial chambers have been separately analysed. Up to now these two datasets are still among the reference dataset for the other CardioGenomics projects investigating disease conditions. Gene expression changes occurring during embryonic heart maturation were investigated also in another dataset (E-GEOD-1479) produced within the CardioGenomics Project, profiling mouse embryos from E10.5 to E14.5 at daily intervals as well as E16.5 and E18.5 (seven time points in total). From E12.5 onwards, ventricles and atrial chambers have been separately analysed. Up to now these two datasets are still among

Table 1: Temporal expression datasets for heart development identified through a search in the ArrayExpress database.

<table>
<thead>
<tr>
<th>ArrayExpress dataset identifier</th>
<th>Dataset title</th>
<th>Year</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-GEOD-75</td>
<td>Transcription profiling of heart ventricles from mice at different developmental stages to monitor changes in cardiac gene expression over time</td>
<td>2002</td>
<td>(77)</td>
</tr>
<tr>
<td>E-GEOD-1479</td>
<td>Transcription profiling of mouse embryo at E16.5 and E18.5, to monitor changes in gene expression related to maturation of the heart</td>
<td>2004</td>
<td>(77)</td>
</tr>
<tr>
<td>E-GEOD-17020</td>
<td>Genome-wide analysis of murine cardiomyocytes at different developmental stages</td>
<td>2009</td>
<td>(67)</td>
</tr>
<tr>
<td>E-MEXP-3590</td>
<td>Rat heart developmental hypertrophy</td>
<td>2012</td>
<td>(68)</td>
</tr>
<tr>
<td>E-GEOD-49906</td>
<td>Transcriptome modulation of ventricles, cardiomyocytes and cardiac fibroblasts during postnatal mouse development</td>
<td>2014</td>
<td>(69)</td>
</tr>
<tr>
<td>E-GEOD-51483</td>
<td>Transcriptional Atlas of Cardiogenesis Maps Congenital Heart Disease Interactome</td>
<td>2014</td>
<td>(70)</td>
</tr>
<tr>
<td>E-GEOD-32078</td>
<td>Global gene expression analysis of developing heart in embryos of diabetic mice</td>
<td>2011</td>
<td>(71)</td>
</tr>
<tr>
<td>E-GEOD-56048</td>
<td>Gene profile in fetal human heart and brain</td>
<td>2014</td>
<td>(72)</td>
</tr>
<tr>
<td>E-GEOD-13614</td>
<td>Expression profiling of the forming atroventricular node using a novel Tbx3-based node-specific transgenic reporter</td>
<td>2008</td>
<td>(73)</td>
</tr>
<tr>
<td>E-GEOD-45821</td>
<td>Gene expression analysis of human heart valves</td>
<td>2013</td>
<td>(74)</td>
</tr>
<tr>
<td>E-GEOD-13923</td>
<td>Divergent gene-expression profiles between proepicardium and epicardium differentiation</td>
<td>2010</td>
<td>(75)</td>
</tr>
<tr>
<td>E-GEOD-37070</td>
<td>Microarray analysis of normal and abnormal chick myocardial development</td>
<td>2012</td>
<td>(76)</td>
</tr>
<tr>
<td>E-GEOD-7689</td>
<td>Genome-wide expression profiling of Drosophila adult heart organogenesis</td>
<td>2007</td>
<td>(77)</td>
</tr>
</tbody>
</table>

Table 2: Temporal expression datasets for human stem cell differentiation into cardiomyocytes identified through a search in the ArrayExpress database.

<table>
<thead>
<tr>
<th>ArrayExpress dataset identifier</th>
<th>Dataset title</th>
<th>Year</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-GEOD-5671</td>
<td>Transcription profiling of mouse embryonic stem cell line mESC D3 reveals cardiac differentiation of embryonic stem cells recapitulates embryonic cardiac development</td>
<td>2006</td>
<td>(78)</td>
</tr>
<tr>
<td>E-GEOD-43197</td>
<td>Natural cardiogenesis-based template predicts cardiogenic potential of induced pluripotent stem cell lines</td>
<td>2013</td>
<td>(79)</td>
</tr>
<tr>
<td>E-GEOD-50704</td>
<td>Transcriptome-guided functional analyses reveal novel biological properties and regulatory hierarchy of human embryonic stem cell derived ventricular cardiomyocytes</td>
<td>2013</td>
<td>(80)</td>
</tr>
<tr>
<td>E-GEOD-35671</td>
<td>Comparison of mRNA expression profiling of differentiating human-induced pluripotent stem cell (hiPSC)–derived cardiomyocytes, biopsies from fetal, adult and hypertensive human hearts and primary cardiomyocytes</td>
<td>2012</td>
<td>(81)</td>
</tr>
</tbody>
</table>
the expression datasets for mammalian heart development with the highest number of analysed time points.

A study in 2009 (E-GEOD-17020) aimed at identifying genes associated with the switch from hyperplastic to hypertrophic heart growth through the analysis of murine cardiomyocytes at different stages of development. To obtain a pure cardiomyocyte population the authors utilised transgenic mice in which GFP is expressed exclusively by cardiomyocytes and isolated the cardiomyocytes by FACS. Expression was profiled at E11–12, P3–4, and adults using Illumina arrays. Analysis revealed a subset of 32 cell cycle-associated genes with expression specific to the neonatal stage with respect to embryonic and adult stage (67). Another study (E-MEXP-3590) investigated expression during rat heart developmental hypertrophy by profiling expression at E18, neonatal, 10-day old (P10), and young adult (six months) Sprague-Dawley rat hearts as well as young adult (six months) and old (>18 months) Wistar rat hearts. The study revealed major changes in gene expression during the postnatal period, in large part due to genes related to mitochondrial biogenesis and mitochondrial DNA maintenance (68). A study in 2014 (E-GEOD-49906) employed RNA-seqencing to assess the transcriptional changes during postnatal mouse heart development using RNA from ventricles (E17, P1, P10, P28, and P90), freshly isolated cardiomyocytes (P1–2, P30, P67) and cardiac fibroblasts (P1–3, P28, P60). The study revealed significant changes in expression and alternative splicing between P1 and P28 (69). In the same year a temporal developmental dataset in mouse was made available (E-GEOD-51483), in which authors assayed gene expression at eight stages (E7.5, E8.5, E9.5, E12.5, E14.5, E18.5, newborn, and adult) with Affymetrix arrays. Starting from E9.5 left and right ventricles have been separately analysed (70).

Our search also retrieved a study in which expression in developing wild type hearts was assessed at only two time points, E13.5 and E15.5, in both diabetic and control mice (E-GEOD-32078, 71) and another study in which heart and brain gene expression was profiled in human foetuses at three time points during late first and early second semester (E-GEOD-56048, 72)). Other studies focused on a key heart developmental sub-process, namely valve formation: one study in mouse (E-GEOD-13614) analysed expression at E10.5 and E17.5 of atrioventricular canal as well as working myocardium (73); a study in human (E-GEOD-45821) assayed expression in foetal cardiac valves from first and second trimester, adolescent heart valves, and adult heart valves (74). These datasets can be interesting to compare with expression data of whole hearts to show whether genes involved in sub-processes such as valve development act independently of genes involved in other developmental sub-processes.

Our search yielded also temporal expression analyses performed in non-mammalian organisms. A study in chicken (E-GEOD-13923) investigated eight time points during proepicardium differentiation and four points during epicardium differentiation (75); another study in chicken (E-GEOD-37070) profiled heart and ventricle tissues at E4, E5, E6, E8, and E10 and compared their profiles with expression profiles at E5 and E8 of chicken with altered haemodynamic loading (76). A study in Drosophila (E-GEOD-7689) investigated adult heart organogenesis by profiling eight time points (77).

Expression studies of stem cells differentiating into cardiomyocytes are also of interest for heart development as they allow identifying the genes and pathways activated during the differentiation process and comparing those with the pathways activated during normal cardiac development (▶Table 2). In an earlier study performed in 2006 (E-GEOD-5671) microarray expression analyses of a selected population of mouse cardiac progenitor cells were performed at differentiation days 5, 6, 7, and 8. The study showed that the transcriptional profile of cells during differentiation was consistent with known active pathways during embryonic heart development (78). In order to investigate the transcriptional differences between stem cells and mature cardiomyocytes, the same group who published the heart expression profiling at eight mouse developmental stages (E-GEOD-51483) analysed expression at three time points (day 0, day 5, and day 11) in three different stem cell lines (E-GEOD-43197, 79)). In addition an analysis in 2013 (E-GEOD-50704) profiled gene expression in undifferentiated human embryonic stem cells, stem-cell-derived ventricular cardiomyocytes, foetal ventricular cardiomyocytes, and adult ventricular cardiomyocytes (80). Another study (E-GEOD-55671) performed a comparison of the mRNA transcriptomes across human-induced pluripotent stem cell-derived cardiomyocytes (12 time points over a 120 day period) with biopsies from foetal, adult, and hypertensive human hearts (81). The authors included in the study also miRNA expression profiling (E-GEOD-35672).

Temporal expression analyses in heart development have been often limited to very few time points (two or three), with the most extensive ones consisting of less than 10 time points. A reason for the limited number of time points is probably the cost for performing high-throughput data analyses. This is still an important factor; however, its importance is likely to diminish as the development of novel more affordable high-throughput technologies is rapidly advancing. Another important limiting factor is tissue collection. For example the recommended starting quantity of RNA for hybridisation to a single Affymetrix GeneChip® 3’ Expression array with the one-cycle target labelling protocol (1–15 µg RNA) required the isolation of around 100 hearts of early mouse and rat embryos. One way to circumvent this issue was to amplify the sample. However, this could have significant influence on the accuracy of the analysis. The development of new technologies and protocols requiring a reduced initial sample amount (e.g. 50–300 ng RNA with the 3’ IVT Express Kit by Affymetrix, which replaced the one-cycle protocol in 2009) will make high resolution temporal high-throughput expression analyses more affordable in the future.

Another issue is to decide on the best experimental strategy to elucidate regulatory mechanisms controlling heart development. The heart consists of several different cell types including cardiomyocytes (e.g. left and right ventricular as well as myocardial conduction cells), cardiac fibroblasts, vascular smooth muscle cells, and endothelial cells (82–84). Profiling the whole heart complicates data analysis and interpretation but it might be important, as it is known that heart development depends on the interaction of
the different cell types (e.g. [85]). Several researchers also chose to utilise only parts of the heart such as left ventricle, right ventricle, atria or valves (Table 1). This can make sense as left and right ventricle for example arise from different cell populations (86). However, the complexity of CHDs with overlapping phenotypes indicates that despite the modular building plan of the heart there must be interactions between the different modules (13). An advantage of datasets of individual heart parts or individual cell types is the opportunity to identify genes that are exclusively involved in sub-processes. A comparison to whole heart datasets will help to identify dependencies of different sub-processes. However, the analysis of sub-structures or individual cell types increases the issue related to the necessary starting material. Therefore, in vitro differentiation studies of the individual cell-types are of interest as this eliminates the issue of extracting a sufficient quantity of RNA. However, current protocols for differentiation of stem cells do not recapitulate the entire differentiation resulting in fully matured adult cardiomyocytes (21). Finally, another reason for the lack of highly resolved temporal datasets is likely to be that collaborations between experimentally focused and bioinformatics researchers are still nowadays often initiated only after high-throughput data have been acquired. Thus, experimental design is not optimized for higher level data processing.

Expression network studies in heart development

In this section we will discuss some examples of use of expression network-based approaches to elucidate regulatory mechanisms during heart development.

The mRNA and microRNA (miRNA) expression datasets measured during differentiation of human-induced pluripotent stem cell derived cardiomyocytes (E-GEOD-35671 and E-GEOD-35672) have been integrated by Babiarz et al. to develop a miRNA network (81). First, differentially expressed mRNA and microRNAs over the time course have been identified; then potential targets of each microRNA have been identified relying on a curated database of miRNA targets and on the (anti)-correlation of each pair of miRNA/mRNA. Finally the calculation of a target overlap score between miRNAs was calculated and the most significant scores visualised as a network relying on the Cytoscape software, an open source platform for the visualisation of complex networks (87). Thus, interactions in the network are indicative of a potential functional association between miRNAs.

Relying on the temporal dataset at eight mouse developmental stages (E-GEOD-51483), Li et al. identified pathways active during heart development and relied on known genes for CHD to build a disease-centric interaction network (70). The authors first performed Principal Component Analysis (PCA) on the whole dataset. PCA showed that replicates at the same time point cluster together and are distinct from the other developmental stages. Furthermore, PCA revealed the similarity between left and right ventricle samples. Then differentially expressed genes across time were identified and clustered to identify the main expression patterns and the associated over-represented Gene Ontology categories. Also temporal patterns of genes involved in some predefined canonical pathways required for heart development are visualised and discussed. Finally authors started from a curated set of known CHD genes to extract their interaction partners from the STRING protein interaction database (88). A custom R package was developed by the authors to visualise the extracted interactions. The extracted view also integrates information on the expression profiles of the interacting genes in the measured time course as well as information on the enriched biological pathways.

As already pointed out, the available temporal datasets contain a limited amount of time points, thus impairing the inference of expression networks. In order to both overcome this limitation and exploit the wealth of information contained in publicly available datasets, some studies relied on the integration of many datasets measuring expression in heart tissues, of which temporal studies constitute the minority. Chen and VanBuren have recently proposed a novel computational strategy to infer gene regulatory networks in mouse heart development (89). The authors retrieved from GEO a selection of 239 microarray expression datasets measured in mouse heart tissues or mouse embryonic stem cells and devised a comprehensive clustering algorithm to identify an optimal set of 765 overlapping gene co-expression modules. Differential expression of the modules in the heart development expression dataset by the CardioGenomics project (E-GEOD-1479) was assessed and employed to assign modules to developmental phase categories. Furthermore, potential regulatory transcription factors (TFs) were inferred by relying on the curated transcription factor binding sites (TFBSs) collection in JASPAR database (90) and graphically represented in a transcriptional regulatory atlas. DNA binding data for selected TFs available within the ENCODE Project (91) were employed to evaluate the built transcriptional network.

Dewey et al. also relied on the large amount of publicly available microarrays in order to identify whether there are foetal-specific gene expression programs and to what extent these programs are reproduced in cardiac hypertrophy and failure (92). The authors compiled a compendium of 478 expression array datasets relative to murine heart contained in the GEO. Weighted gene co-expression network analysis was employed to identify foetal gene programs, i.e. modules of co-expressed genes that are specific to foetal myocardial tissue. The authors found that 50 out of the 72 identified modules are not reproduced in adult heart and are thus foetal-specific and that a low proportion of foetal modules are recapitulated in hypertrophied and failing myocardium. Furthermore, by representing the modules by their first “eigengene”, authors could explore higher-order network topology and find that foetal intermodular connections are only partially preserved in hypertrophy and heart failure.

These two studies offer an example of the insights given by network analysis and not obtainable from traditional expression analysis. However, the use of association-network approaches did not allow performing in silico experiments on the networks. Network-based simulations have been exploited in a study by Herrmann et al. based on Boolean networks (93). The authors relied on...
an extensive literature search to construct a regulatory network of early cardiac development in the mouse. The network focuses on the biological processes that lead to specification of the first and second heart fields (FHF and SHF), which later develop into different parts of the heart. The network was then implemented as a Boolean model with 15 nodes, translating the published data on the regulations between genes into Boolean functions. Simulations of the model showed that the network was able to reveal stable states corresponding to the expression profiles of the FHF and SHF. Further simulations revealed that the model was able to reproduce knock-out and overexpression phenotypes. This study shows how also relatively simple network models as those represented with Boolean networks can summarise and reproduce complex developmental processes.

Integration of expression data with other data types, such as transcription factor binding measurements offered by ChIP-chip or ChIP-seq experiments, is a promising strategy to improve accuracy of the inferred regulatory networks. An example of integrative approach to reconstruct cardiac transcription networks was proposed by Toenjes et al. (94). The authors collected heart biopsies from healthy individuals as well as patients with a large variety of cardiac malformations and used real-time PCR to measure expression levels of 42 genes associated with heart development. A phenotype ontology was defined and employed to define patient groups and linear models were used to detect differentially expressed genes in the groups. In addition, independently of phenotype, groups of genes with significantly correlated expression levels were identified. Next, in order to identify TFs potentially co-regulating genes in a group, TFBS predictions were employed to identify TFs having binding sites in the promoters of all genes in a group. Finally a cardiac regulatory network based on the correlated gene groups and the predicted TFBSs was built and verified with literature and data from chromatin immunoprecipitation.

Conclusions

Gene networks constitute a promising tool to support the identification of novel therapy-relevant regulatory processes in heart development. The majority of reviewed papers highlight learned co-expression networks, which are useful instruments to summarise the expression patterns and to highlight relationships and interactions. However, the small number of samples in currently available temporal expression datasets has limited the use of reverse engineering approaches able to infer a network of gene interactions as well as a mathematical model of these interactions to be employed for in silico simulations. A first solution is to resort to Boolean models, since they require a relatively low amount of data for learning. However, Boolean models only provide an abstract view of the gene expression behaviour, and can therefore be considered only the first step of a deep mathematical description of the underlying gene regulatory processes. An advantage of Bayesian approaches is their ability to accommodate the intrinsic variability of biological systems, providing a quantitative representation of the temporal patterns. Furthermore Bayesian approaches offer a mathematically well-founded framework to integrate prior knowledge and different types of data in network learning and thus improve the accuracy of the inferred models. However, the application of Bayesian network modelling requires a larger amount of data than Boolean. Recent publications have demonstrated that the application of Bayesian approaches supports the identification of important molecular processes underlying development and disease (95–98). Thus, to understand dynamic processes in heart development it is crucial to improve our modelling capabilities by acquiring temporal expression data spanning from embryonic to postnatal developmental stages at high resolution. Once these data are made available, it will be possible to move further with a systems biology description. Taken together, we propose the use of probabilistic gene network analysis methods as a promising opportunity for future studies in heart development.

Conflicts of interest

None declared

References