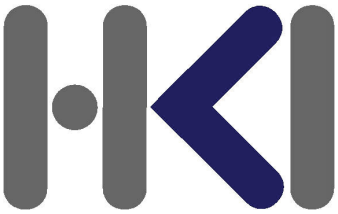


Integrative methods for reconstruction of dynamic networks in chondrogenesis

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der
Biologisch-Pharmazeutischen Fakultät der
Friedrich-Schiller-Universität Jena

von Dipl.-Bioinf. Michael Weber
geb. am 20.06.1984 in Suhl

Gutachter:

1. Prof. Dr. Reinhard Guthke (Hans-Knöll-Institut Jena)
2. Prof. Dr. Raimund W. Kinne (Universitätsklinikum Jena)
3. Prof. Dr. Francesco Falciani (University of Liverpool)

Die Disputation der Doktorarbeit erfolgte am 21.11.2013 in Jena.

Contents

Chapter 1

Introduction

1.1 Microarray technology

More than a decade ago, the first full draft of the human genome sequence has been published [1–3]. Since then, molecular biological research has benefited from the availability of those sequences. Knowledge about the location of protein-coding gene loci and non-protein-coding transcripts has been augmented and updated on the basis of the given sequence data [4]. While the genome can be regarded as relatively constant information which is stored in most of the human cells, transcription of genes varies and is directed by dynamical processes that depend on the type of cell, tissue and environmental conditions. Therefore, the most promising research apart from the exploration of the genome has been the analysis of the transcriptome, which is the total set of all RNA molecules expressed in the cell [5]. One task of transcriptome studies is the investigation about how mRNA expression leads to cellular function. This connection is of enormous interest, because it helps to assign function to genes and therefore can lead to a better understanding of cellular processes and associated disorders [6]. This includes the study of gene regulation, which aims to investigate the dependencies of genes and how their regulation governs the activity of those processes.

Due to technological advances made in automatisation and robotic industry, the production of high-density probe arrays (e.g. glass slides or silicon chips) is possible and is mainly applied in biological and medical research [7]. Hundred thousands of artificially generated oligonucleotides (probes) can be fixed to the array's surface. Probes were designed to specifically match a subsequence of the target transcript and allow for perfect hybridisation between probe and transcript. Because of the tremendous number of probes, a single microarray can measure transcript concentration on a genome-wide scale. The corresponding experimental procedure is standardised and typically involves the following steps: RNA extraction, reverse transcription of mRNA, labelling of the molecules, hybridisation to the microarray, signal detection (e.g. by laser scan) and data quantification [8]. This leads to an array of measurement values for all genes which are captured by the microarray. On the basis of such data, gene expression can be explored in an unbiased manner for all annotated genes of the genome. In the human genome, this includes measurements for more than 20,000 genes. Therefore, cellular researchers have obtained a powerful tool for various applications, which has become immensely popular and initiated a new era of gene expression studies and molecular research [9].

1.1.1 Application of microarrays

Besides gene expression analysis, microarray technology allows for other possible applications in molecular biology including SNP genotyping and transcriptome mapping [10]. Nevertheless,

expression analysis is the most common task performed by microarrays and applications can be found in fundamental research as well as in the biomedical field [6, 11]. A typical experimental strategy is to perturb the cells of interest and to monitor their response on the transcript level. Cellular perturbations include external stimulation, gene knock-out or knock-down and the over-expression of specific genes [12–14]. Furthermore, several human diseases have been investigated by using genome-wide arrays including diverse types of cancer [15, 16], rheumatoid arthritis [17] and osteoporosis [18]. One major aim of those studies is the detection of genes which show significant expression differences between the diseased state and an appropriate control state. Moreover, gene expression patterns of patients have been analysed to distinguish disease subtypes, to predict a clinical outcome (diagnosis) and to personalise therapy [16]. As a result, medical diagnosis could be improved by assigning more precise and detailed disease characteristics to the individual patient. Therapy which is tailored to the patients specific characteristics, is the main objective of personalised medicine [19].

1.1.2 Human genome microarrays from Affymetrix

While this thesis exclusively deals with human data, microarray experiments can be conducted for a wide range of model organisms, such as *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster* and *Arabidopsis thaliana* [7]. The following text will outline some facts about the design and the application of the microarray technology used in this thesis. Gene expression experiments were performed using the Affymetrix GeneChip Human Genome U133A (HG-U133A), which contains more than 20,000 probe-sets for 14,500 genes of the human genome. Another more advanced microarray platform of the GeneChip family named Human Genome U133 Plus 2.0 contains more than 47,000 probe-sets and is one of the most widely used arrays (more than 80,000 entries in GEO database). Expression of each gene is measured by a probe-set, which is a collection of 11-20 probe pairs assigned to the same transcript. Each probe pair consists of a perfect match probe (PM) and a mismatch probe (MM), which are both oligonucleotides of length 25. The PM probe is designed to perfectly match a specific sequence within the target mRNA, while the MM probe carries one mismatch at the 13th base. Originally, this single mismatch strategy should help to quantify the background signal caused by non-specific binding. However, some standard microarray pre-processing methods (e.g RMA) do not incorporate the measurements of MM probes in their gene expression estimate, because they found that the use of those values is not always appropriate [20].

1.1.3 MicroRNA microarrays

Further advances in microarray technology have enabled the measurement of very short RNAs such as microRNAs (miRNAs). In this thesis, data of custom microarrays were analysed, which have been designed to capture the expression of about one thousand miRNAs from the database miRBase [21]. The main issue of measuring miRNAs on microarrays concerns the design of probes to measure miRNAs, which have a very short sequence length (around 22 bp) [22]. Therefore, microarray experiments employed locked nucleic acid (LNA) capture probe-sets from Exiqon spotted on Schott Nexterion E slides [23]. Compared to DNA probes, LNA-RNA hybrids are extremely stable, but get destabilised when a single mismatch occurs. Therefore, they have been found to be very appropriate for measuring miRNA concentration. Moreover, the probe design aimed at a normalised melting temperature in order to ensure an optimal hybridisation process. This led to capture probes of varying length. The performance of this platform was tested and the specificity of the measurements was evaluated on the basis of mismatch probes, reproducibility tests and quantitative PCR. As a result, almost no cross-hybridisation was observed when using the developed capture probe-sets [23].

1.1.4 Microarray data pre-processing

Analysis of microarray data requires adequate processing of the generated values. One reason is that the measured signal intensities do not only reflect biological quantities, i.e. the true RNA concentrations in the cell, but can also be affected by measurement error during the microarray experiment. Such error can arise for different reasons, including variation in sample preparation and non-specific hybridisation to probes [24]. To ensure that gene expression is reliably estimated from the data and comparative analysis of microarray samples is feasible, microarray measurements need to be adjusted for noise effects. Procedures which address this issue are generally referred to as pre-processing, because they are performed before the actual analysis of gene expression. Pre-processing methods typically involve three steps including background correction, normalisation and summarisation [25]. Two of the most commonly applied pre-processing methods for GeneChips are the Microarray Suite 5.0 (MAS5) from Affymetrix and the robust multi-array average (RMA) method [20, 26].

Both take a raw microarray dataset as input data and compute gene expression estimates for all annotated probe-sets. The main difference lies in their estimation of the background signal intensity. While MAS5 relies on the array's MM probes and their ability to capture non-specific binding, RMA employs empirical distributions for background and signal intensities by only considering PM probes. Specifically, each PM probe signal is adjusted by subtracting the modelled background intensity. By comparing the performance of RMA and MAS5, it could be shown that RMA can lead to a better precision of expression values and more consistent estimates of fold-changes [20, 27]. Nevertheless, biological interpretations are reported to remain similar independent of the chosen pre-processing method [7]. After microarray data pre-processing, the logarithmic gene expression values are usually analysed in a relative manner, i.e. relative expression changes are calculated by subtracting expression values from control samples. The resulting values are often referred to as log2 fold-changes. In fact, relative measures are regarded to be more reliable compared to absolute measures, because microarrays intensity values can usually not be interpreted as total RNA concentrations [28].

1.1.5 Candidate gene selection

The selection of genes from a microarray dataset is regarded as an essential task prior to model construction [29]. In principal, one aims to reduce the large amount of available genes down to a manageable number of candidate genes, which are associated with the respective biological question. Since the main assumption of microarray analysis is that most of genes are not differentially expressed, it is reasonable to remove those genes before continuing analysis [30]. Potentially interesting genes display variation in their expression profile across the obtained samples. In the case of time series data, this means that expression dynamics undergo significant changes over time. To assess the significance of those temporal changes, appropriate statistical tests are employed to detect differentially expressed genes. However, most microarray datasets have properties which are not ideal to perform standard statistcal tests, such as the classic t-test [31]. Typical microarray features are the inequality between few replicates and the large number of measured genes per array. Both issues are considered by methods which perform adapted statistcal tests (e.g. LIMMA, SAM) [32, 33]. To estimate variance from very few replicates, adapted tests take account of the global expression variation in the dataset. This approach was found to stabilise the computation of the p-values [31, 32]. Furthermore, the vast number of statistical tests requires a multiple testing correction in order to reliably estimate the p-values for all measured genes. A common approach is the Benjamini Hochberg adjustment, which calculates the estimated proportion of false positives also known as the false discovery rate (FDR) [34]. Additionally, analysis of time series data requires specific tests (e.g. F-test), to identify genes which are significantly differentially expressed at any of the measured time points.

The result of the differential expression analysis is typically a ranked gene list, which is ordered by their adjusted p-values and/or fold-changes. To choose a certain p-value cutoff for gene selection is a user-defined task, which depends on the number of genes to be analysed and the proportion of false positives that is tolerable. Additionally, to eliminate statistically significant genes which display only a minor average difference to control, a two-fold-change criterion is often employed. Since small changes in the expression level might be dominated by measurement noise, the additional fold-change cutoff helps to require a minimum amount of change in concentration. Genes which double their expression level are typically concerned of biological interest [35], in contrast to genes which are merely significant according to their p-value. In general, the combination of p-value and fold-change criteria for gene filtering was reported to enhance the reproducibility of results from microarray experiments [36].

1.2 Inference of gene regulatory networks

1.2.1 Network inference from microarray data

Gene expression studies on a genome-wide level using microarray technology enables researchers to perform analysis in an explorative and data-driven manner. Investigations of molecular processes benefit from such large-scale gene expression data, because it captures the expression patterns of all involved genes. To unravel the underlying gene regulation on the basis of such data is a scientific challenge. The principal aim is to understand how the complex gene regulatory processes are structured in the cell. Particularly, researchers seek to investigate how genes influence each other in the cellular system and how the sum of those regulatory interactions gives rise to the phenotype of the cell. Systems biological approaches apply this idea of an integrated analysis of cellular components. They explain cellular behaviour by the orchestrated regulation of its components, e.g. genes, proteins or metabolites, instead of focusing on individual components. One prevalent objective is to generate a network which describes gene regulation in the cell. This type of network is composed of genes and connections between the genes, which reflect observed or predicted relationships in the cellular system. The reconstruction or inference of such a gene regulatory network (GRN) is a typical system biological approach, which aims to derive the underlying regulatory network from gene expression data. In this thesis, network inference is part of a multi-step analysis process, which aims to result in testable experimental hypotheses (see Figure 1.1).

Various computational approaches have been developed for the purpose of network reconstruction by applying diverse modelling strategies. One common strategy seeks to learn an influence network model [14], which describes indirect regulatory relationships among the genes instead of direct physical interactions. Particularly in eukaryotic cells it is advantageous to model influences, since gene regulation is complex and controlled by the interplay of various signalling pathways, whose activity is not always reflected on the transcript level [14]. Those pathways mediate an external environmental signal (e.g. by a growth factor) to the cell's nucleus, where it typically results in the change of target gene expression. On an abstract level, cellular regulation is regarded as a network of regulatory relationships between genes, which represent direct and indirect dependencies. To unravel the structure of a cellular GRN represents an enormous challenge due to its complex intertwined nature.

Analysis of known GRNs has helped to determine some of their general properties. Their network structure has been found to be sparsely connected and the distribution of node degrees appeared to follow a power law distribution [37]. Most of the genes are regulated by few central

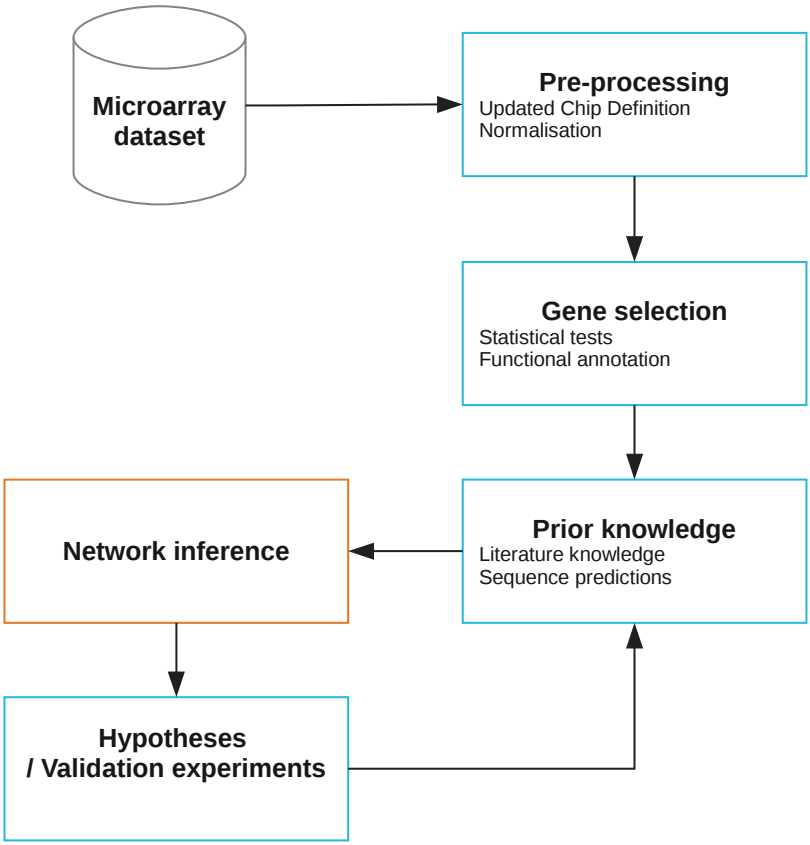


Figure 1.1: Network inference as one part of a multi-step process, which includes microarray data analysis, prior knowledge collection and validation of hypotheses.

regulators, which are also known as hub genes. Various GRN inference methods have translated the observed property of sparseness into an optimisation principle, which favours minimal connected network structures during the model reconstruction [38–41]. Various types of computational methods which aim at the reconstruction of GRNs have been proposed, including methods based on correlation measures, Bayesian probability and Boolean approaches [42–44]. This thesis deals with network models using a system of ordinary linear differential equations (ODEs) to explain observed changes in gene expression. The network structure of the model is determined by the model's parameters, which can be visualised by a directed network in which connections represent regulatory influences between the components. To infer those parameters, time series data are required because they indicate the dynamic behaviour of gene regulation. The resulting dynamic model can be simulated to evaluate the concordance with the measured time series. There are several studies which implemented ODE-based models and successfully applied them to reconstruct the GRN for diverse biological contexts, e.g. *E.coli* [38], murine hepatocytes [41] and *Candida albicance* [45]. The tools which were used in those studies to perform the network inference are termed ExTILAR and NetGenerator. While ExTILAR is based on constrained linear regression, NetGenerator applies a heuristic to optimise the GRN structure.

This thesis primarily focuses on an extended version of the NetGenerator tool and its application for network inference [38,46,47]. The underlying modelling approach is based on the assumption that the investigated system is initially (at the first time point of the measured time series) in a steady state. Then, it undergoes a significant external perturbation (e.g. stimulation by a growth factor), which induces a dynamic response characterised by concentration changes of the system's components (mRNAs, miRNAs). Time series measurements capture the dynamic response of all components at preselected time points. Therefore, most of the variation that occurs in the

time series data is directly or indirectly due to the applied experimental stimulus. In order to reconstruct the underyling network NetGenerator optimises a parametrised ODE-based model by applying a heuristic search strategy to solve the combinatorial network structure optimisation problem, i.e. to identify the optimal set of regulatory connections between the components. The interaction parameters of the model are optimised to minimise the deviation between the simulated data and the measured time series data.

This inference strategy strongly depends on the provided microarray data and therefore one has to deal with several issues that are connected to this type of experimental data. One issue concerns the dimensions of the applied microarray data, i.e. measurements for ten thousands of genes are available for a usually relatively small number of microarray samples. This imbalance between genes (network components) and samples make it difficult to unambiguously infer the network structure, due to the existence of multiple solutions which explain the data equally well. This so-called dimensionality problem can be resolved by various strategies including (1) the selection of an reduced number of genes, (2) the grouping of genes into clusters, (3) the extension of the amount of data and (4) the integration of prior knowledge. To select a reduced number of network genes, one typically identifies differentially expressed genes, which are regarded as the set of genes that are responsible for the investigated cellular phenotype. Another way to deal with the vast amount of genes is to cluster genes according to their gene expression profiles and perform network inference for few cluster representatives [38, 45]. Instead of further decreasing the number of network nodes, one can also collect further expression data from additional experiments or public databases to increase the number of samples. The integration of multiple experimental data is the main focus of this thesis and therefore section 1.2.2 will introduce already existing approaches which apply this idea.

Another important aspect is the application of adequate prior knowledge, because it constrains the potential network structures and therefore may facilitate the inference process [14]. The general idea of prior knowledge and the heterogeneous types of sources used in this thesis will be outlined in chapter 1.2.3. Furthermore, network inference depends very much on the quality of the given data. Therefore, it is necessary to cope with the inherent noise of microarray data to ensure that the resulting model is primarily based on true expression changes rather than on technical variation in the data and biological variation (e.g. between the individuals included in a study). One way to address this issue is the application of a resampling approach. Such methods can help to assess the impact of noise in the input data on the resulting network model and lead to a more robust model with less dependency on noise and therefore better quality [14].

1.2.2 Inference from multiple experimental data

The particular focus of this thesis is the combination of experimental data in order to increase the reliability of network inference results. There are various studies which have proposed diverse approaches to reconstruct a regulatory network on the basis of multiple experimental data. One of the earliest algorithms was introduced by [12], which requires data from specific gene perturbation experiments in order to gradually reconstruct a Boolean network. A similar algorithm was introduced by [48], which also applied the concept of systematic perturbations of the investigated network. However, both studies presented the successful application of their algorithm solely on artificial data. A more recent modelling approach, which is also based on Boolean networks, applied a gene expression dataset from synovial fibroblasts which were treated with two different stimuli. This work included large parts of knowledge curation and manual network improvement steps [49].

Based on a linear model, the genome-wide scale network published by [50], illustrated network inference from a large number of published datasets that included *Candida albicans* expression

data under approximately 200 different conditions. They applied an extended version of the LASSO algorithm on a combined dataset which included all available expression data. In the case of time series data, such merging of data is regarded as inappropriate, because it ignores the temporal relationships among the datasets [51]. Instead, an approach based on linear differential equations has been proposed by [51] to model principal components of the different datasets. A more general framework which allows to integrate distinct types of data including steady-state as well as time series data was presented by [52]. Their particular focus is the combination of data from stimulation and knock-out experiments.

Even though some inference approaches have dealt with multiple time series data, there is a general lack of tools which performs this task in a relatively automatic and time efficient manner. Therefore, this thesis proposes the NetGenerator V2.0 tool which is capable to automatically infer GRNs from multiple time series data. Such data is typically gained from multiple stimulation experiments of the same biological system. The performance of this tool is evaluated on artificial benchmark examples and is applied to infer a multiple stimulus network of human mesenchymal stem cells which undergo chondrogenic differentiation. Furthermore, a second modelling task demonstrates the applicability of NetGenerator V2.0 to a combined dataset of miRNA and mRNA expression time series to elucidate the involvement of miRNAs in chondrogenesis. Due to advanced microarray technology and miRNA annotation, such experimental datasets have become more frequent in the recent years. Those data contributes to the analysis of miRNA target interactions and understanding of post-transcriptional regulation. There are a number of network inference approaches which address such questions by an integrated analysis of miRNA and mRNA data. Tools such as MAGIA [53], MMIA [54] and miRConnX [55] are based on the integrated analysis of miRNA target gene predictions and correlation between miRNA and mRNA expression profiles. Although they have particular strengths in the discovery of regulatory motifs, their correlation-based approach does not consider the temporal information of time series data.

1.2.3 Prior knowledge

Generally, the inference of GRNs benefits from the integration of additional biological data [14]. This includes prior biological knowledge, which summarises all previously collected information about regulatory interactions among the network components. Possible sources for prior knowledge include scientific literature, transcription factor binding site (TFBS) predictions in promoter regions and potential regulations derived from knock-out experiments. In this context, predicted interactions are also referred to as knowledge although their reliability is apparently lower compared to published experimental results. Nonetheless, the additional information contained in accurate predictions is a valuable source for network inference, because this data is independent of the used time series data.

A collection of heterogeneous knowledge can assist the inference process by providing network structure proposals. As the inference of the network structure represents a combinatorial problem, accurate proposals can contribute to the performance of this task. However, the obtained prior knowledge may not be explainable by the given gene expression measurements, which are used for network model construction. Such contradictions can occur if the biological context of both data sources is different, e.g. a different type of tissue or diseased cells. Therefore, network inference approaches which use prior knowledge have adopted ways to deal with such inconsistencies. The extended NetGenerator tool is now able to perform flexible integration, which allows the rejection of contrary prior knowledge to retain a good model adaptation to the data. Another concept is applied by the inference approach ExTILAR, which is based on heterogeneous knowledge including TFBS predictions and literature knowledge. Incorporation of the diverse knowledge is performed on different levels of the GRN inference to result in a model

of high biological plausibility. A general issue of using prior knowledge for network inference is the distribution across diverse resources such as databases, literature or additional datasets. The collection of this information can require a significant amount of manual work, before the network inference is initiated.

1.3 Differentiation of human mesenchymal stem cells

Human mesenchymal stem cells (hMSCs) are progenitor cells which are characterised by self-renewal and multipotency. Self-renewing cells maintain a pool of undifferentiated stem cells, which are capable of differentiating into multiple cell types. Specifically, hMSCs have been found to give rise to cells of the mesodermal lineage including chondrocytes, osteocytes and adipocytes [56]. Therefore, they are responsible for the supply of fresh cells to maintain and replace non-functional (e.g. dead or mutated) cells. For example, human osteoblasts have a half-life of 8-10 days and get replaced by hMSCs [57]. Regeneration of tissue is often described as one of the most essential features of hMSC [58]. In this process, they are able to induce a regenerative microenvironment around the area and thereby support the repair of injured tissue [57].

One of the earliest descriptions of hMSCs were made by [59], who discovered a rare cell type in the bone marrow, which allowed for generating cell colonies from a single cell [60]. Those colonies were found to differentiate into cells which display similarities to bone or cartilage. The discovered stem cells were termed "mesenchymal", after it became clear that they are typical precursor cells of mesenchymal tissues [56]. However, it became apparent that their differentiation capacity allows them also to transdifferentiate into other lineages, such as neurons under appropriate conditions [61]. The traditional source of hMSC is the bone marrow, although they have also been derived from almost all types of connective tissue, including adipo tissue and synovial membrane tissue [56, 58]. However, they usually occur in very small amounts and are heterogeneous in their individual commitment stages [56].

Commitment of hMSCs towards a specific lineage is characterised by subsequent cell fate decisions [62]. This developmental process is organised by differentiation pathways, which are characterised by mutual dependencies. For example, blocking the adipogenic pathway leads to hMSCs differentiating towards osteocytes and vice versa [61]. There is a wide range of existing knowledge about the involved regulatory factors, including growth factors and transcription factors which have been identified and associated with the differentiation processes. On the genetic level, the transcription factors are responsible for the regulation of target genes which are required for activation of downstream processes and the cellular commitment [63, 64]. However, the full extent of the involved regulatory factors as well as their complex interaction network has not been completely elucidated.

More precise knowledge about the regulatory mechanisms would allow for advances in biomedical applications, such as tissue engineering in regenerative medicine. Experiments and culture conditions have been designed for the purpose of cell cultivation and production of certain cell types. Due to their flexible character hMSCs are suitable for application in medical therapies. Particularly, they can be employed with the objective to repair damaged tissue in the case of a degenerative disease. Preclinical trials already successfully implemented stem cells to form bone, cartilage, muscle and other connective tissues. For example, autologous hMSC were applied at long-bone repair sites and polymeric hMSC scaffolds were used for cartilage repair [57]. Bone and joint diseases like osteoporosis, osteoarthritis could be treated with artificially generated cartilage and bone cells. Therefore, advances in molecular hMSC research are likely to have direct positive impact on the medical treatment of patients. Particularly, the understanding of stem

cell regulation and elucidation of regulatory mechanisms can give rise to adapted and improved stem cell therapies. Furthermore, hMSCs have also been found to modulate immune cells, such as T-cells. Surprisingly, they appear to have an anti-proliferative and anti-inflammatory effect on the target cells [56]. This feature is extremely valuable for the therapeutical application of hMSCs, since their immunosuppressive effects might facilitate the use of donor cells.

Chapter 2

Overview of manuscripts

Creation and Comparison of Different Chip Definition Files for Affymetrix Microarrays

C. Hummert[1], F. Mech[1], F. Horn[1], M. Weber[1], S. Drynda[2], U. Gausmann[3], R. Guthke[1]
[1]Research Group: Systems Biology / Bioinformatics, Hans Knöll Institute Jena
[2]Clinic of Rheumatology, Otto von Guericke University Magdeburg, Medical Faculty
[3]Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI)

Status published in Proc. International Conference on Bioinformatics and Computational Biology BIOCOMP'11, 2011, Vol.I, pp. 16-22

Summary Microarray technology is based on a vast number of probes which are designed to hybridise to specific target transcripts. The quality of the microarray's measurements therefore depends on the specificity of the probe sequences. This study compared four different chip definition files, including one novel approach, which all aimed to identify and discard non-specific probes from the probe-sets. Evaluation was performed on the basis of two datasets which include microarray data as well as qRT-PCR data for a selection of genes. Since qRT-PCR is regarded as a reliable method to determine gene expression, it was used to validate the gene expression estimates from the corresponding microarray data. As a result, one chip definition file was found to show the best performance, while the original annotation was found to perform worst.

Authors' contributions CH and FM performed the analysis and developed the annotation files. MW carried out the statistical evaluation of the files. SD and UG provided some of the applied experimental data. RG supervised the study and contributed to the manuscript. CH, FM and FH wrote the manuscript. All authors read and approved the final manuscript.

Chapter 4

Gene Expression Regulation underlying Osteo-, Adipo-, and Chondro-Genic Lineage Commitment of Human Mesenchymal Stem Cells

Ana M. Sotoca
Radboud University, The Netherlands

Michael Weber
Hans Knöll Institute, Germany

Everardus J. J. van Zoelen
Radboud University, The Netherlands

Status: published in Daskalaki A: Medical Advancements in Aging and Regenerative Technologies: Clinical Tools and Applications. IGI Global, Vol. 1, pp. 76-94. Posted by permission of the publisher.

Summary: Human mesenchymal stem cells can be found in various adult tissues and are characterised by their multipotent state, which enables them to differentiate into multiple cell types including cartilage, bone and fat cells. The task of this study was the further examination of the differentiation process with a focus on gene regulation. Diverse experimental conditions were applied and found to induce differentiation with distinct efficacy. A comprehensive microarray dataset was analysed to investigate the underlying gene expression levels. Differentially expressed genes were identified which are potentially involved in the differentiation process. Furthermore, some lineage-specific genes were found to show an increased or accelerated upregulation, which reflects the observed differentiation performance. Taken together, this study provided a comprehensive overview about differentiation conditions as well as the behaviour and function of regulated genes.

Authors' contributions: AS and MW wrote the manuscript. AS and EJvZ contributed to the design of the experiments and the biological interpretation of the results. MW peformed the data pre-processing and the bioinformatic analysis. EJvZ supervised the study. All authors read and approved the final manuscript.

Weber *et al. BMC Systems Biology* 2013, **7**:1
http://www.biomedcentral.com/1752-0509/7/1

BMC Systems Biology

METHODOLOGY ARTICLE Open Access

Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0

Michael Weber[1], Sebastian G Henkel[2*], Sebastian Vlaic[1], Reinhard Guthke[1], Everardus J van Zoelen[3] and Dominik Driesch[2]

Status: published in BMC Systems Biology; 2013; 7:1

Summary: Inference of gene regulatory networks is a common task to investigate the behaviour of regulatory processes. Typically, gene expression data from microarrays is employed to model the effects on the transcript level. This study proposes the extended NetGenerator V2.0 tool which is capable of inferring networks from multiple time series data investigating the effect of multiple stimuli. Data integration into a single network inference is beneficial, because it increases interpretability and reliability of the model. Successful inference of three benchmark examples demonstrated the tool's ability to infer different types of cross-talk structures. Microarray data from a chondrogenesis experiment were used to model the effect of multiple stimuli. The network inference procedure illustrates the initial gene selection and the combination of diverse prior knowledge. The resulting network proposes interesting hypotheses about gene regulation during chondrogenesis. Overall, the described NetGenerator V2.0 is an automatic and efficient approach to infer gene regulatory networks from multiple experimental data.

Authors' contributions: MW and SGH drafted the manuscript. SGH and DD contributed to the development and programming of the NetGenerator algorithm and software as well as to the mathematical and modelling background. MW and RG contributed to data processing, application of NetGenerator to examples, statistical evaluation and the biological interpretation. SV contributed to the generation of the benchmark systems and their artificial data. EJvZ contributed to experimental set-ups, measurements and biological interpretation of the chondrogenic investigation. All authors read and approved the final manuscript.

Dynamic modelling of microRNA regulation during mesenchymal stem cell differentiation

Michael Weber*[1] , Ana M Sotoca[2] , Peter Kupfer[1] , Reinhard Guthke[1] and Everardus J van Zoelen[2]

[1]Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Beutenbergstr. 11a, 07745 Jena, Germany
[2]Department of Cell and Applied Biology, Radboud University, Heijendaalseweg 135, 6525 AJ Nijmegen, The Netherlands

Email: MW - michael.weber@hki-jena.de; AS - a.sotoca@science.ru.nl; PK - peter.kupfer@hki-jena.de; RG - reinhard.guthke@hki-jena.de; EJvZ - vzoelen@science.ru.nl;

*Corresponding author

Status: submitted to BMC Systems Biology; March 2013

Summary: Integrated modelling of microRNA and mRNA interactions has become feasible through the availability of adequate microarray data. This study applied NetGenerator V2.0 on time series expression data, which includes simultaneous measurements of both types of RNAs. The most promising microRNAs were identified in a multi-step procedure which included differential expression analysis, microRNA target predictions and functional annotation of the targets. Afterwards, NetGenerator inferred a dynamical model of high stability. As a result, this network proposes four microRNAs which potentially influence the expression level of regulatory factors in chondrogenesis. Additional experiments confirmed the effect of miR-524-5p on the expression of its downstream targets. Taken together, this study demonstrates the successful inference of microRNA regulation on the basis of data integration.

Authors' contributions: MW and AS drafted the manuscript. MW performed pre-processing and modelling of the network. PK contributed to the network component selection. AS and EJvZ contributed to the experimental set-ups, measurements and biological interpretation of the network. All authors read and approved the final manuscript.

Title	Status	Journal	Authors	Contribution
Manuscript 1	published	Proc. Internat. Conf. on Bioinformatics and Computational Biology BIOCOMP'11, 2011, Vol.I, pp. 16-22	Hummert C Mech F Horn F Weber M Drynda S Gausmann U Guthke R	27 % 27 % 20 % 20 % 2 % 2 % 2 %
Manuscript 2	published	Daskalaki A: Medical Advancements in Aging and Regenerative Technologies: Clinical Tools and Applications. IGI Global, Vol. 1, pp. 76-94	Sotoca AM Weber M Zoelen EJ	50 % 30 % 20 %
Manuscript 3	published	BMC Systems Biology 2013; 7:1	Weber M Henkel SG Vlaic S Guthke R van Zoelen EJ Driesch D	50 % 30 % 5 % 5 % 5 % 5 %
Manuscript 4	submitted	BMC Systems Biology	Weber M Sotoca AM Kupfer P Guthke R van Zoelen EJ	70 % 10 % 10 % 5 % 5 %

Table 2.1: Overview of manuscripts

Chapter 3

Manuscripts

3.1 Creation and Comparison of Different Chip Definition Files for Affymetrix Microarrays

Creation and Comparison of Different Chip Definition Files for Affymetrix Microarrays

C. Hummert[1], F. Mech[1], F. Horn[1], M. Weber[1], S. Drynda[2], U. Gausmann[3], R. Guthke[1]
[1]Research Group: Systems Biology / Bioinformatics, Hans Knöll Institute Jena
[2]Clinic of Rheumatology, Otto von Guericke University Magdeburg, Medical Faculty
[3]Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI)

Creation and Comparison of Different Chip Definition Files for Affymetrix Microarrays

C. Hummert[1], F. Mech[1], F. Horn[1], M. Weber[1], S. Drynda[2], U. Gausmann[3], R. Guthke[1]
[1]Research Group: Systems Biology / Bioinformatics, Hans Knöll Institute Jena
[2]Clinic of Rheumatology, Otto von Guericke University Magdeburg, Medical Faculty
[3]Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI)

Abstract— *Microarrays are broadly used for high-throughput gene expression analyses in molecular biology and medicine. Nevertheless, the quality of the technology is still capable for further improvements. One of the main problems is cross-hybridization of the transcripts to non-corresponding probes on the array by unspecific binding.*

Four different Affymetrix GeneChip arrays are analyzed, namely the Human Genome arrays HG-U133A, HG-U133B, HG-U133 Plus 2.0 and the Mouse Genome 430 2.0 array. It is shown that putative cross-hybridizations are common for the examined arrays (e.g., 45 % of all probes for the U133A). Furthermore, a considerable amount of probes does not match the annotated transcript correctly. A new set of CDFs is created avoiding putative cross-hybridization completely. It is compared with three other CDFs (Affymetrix, Dai et al.*, Ferrari* et al.*) with the help of correlation between microarray and qRT-PCR results for two datasets. The newly created and the Ferrari CDFs perform significantly better than the original Affymetrix CDFs. The new CDFs are available as R-packages at http://www.sysbio.hki-jena.de/software and have been submitted to BioConductor.*

Keywords: microarrays, unspecific binding, cross-hybridization, Chip Definition Files

1. Background

Microarrays are broadly used for high-throughput gene expression analyses in molecular biology and medicine. They are applied to measure changes in expression levels for thousands of genes simultaneously. Until 2011, more than 20,000 measurement series based on microarray technology have been published in public repositories like NCBI's Gene Expression Omnibus.

Nevertheless, the quality of the technology is still capable for further improvements [1], [2]. Several studies tried to compare data derived from different types of arrays and showed a rather poor consistency [3], [4]. Although microarrays are commonly used, this is a daunting problem. In addition, although there has been extended work on this field [5], there is still a lack of standardized experimental protocols among different laboratories [6].

The main problem of microarray analysis is unspecific binding of transcripts by cross-hybridization. This means that RNA fragments hybridize to a probe which is not designed for this gene. It was shown that fragments longer than 8 nucleotides are able to hybridize and that cross-hybridization can emerge from alignments ranging from 10 to 16 nucleotides. Further, the 5'-ends were found to cross-hybridize more likely than the 3'-ends [7].

Unspecific binding may lead to false-positive and false-negative results following in incorrect hypotheses about gene expression [8], [9]. Affymetrix, a technology widely used [10], accounts for the influence of cross-hybridization by introducing internal controls: each probepair comprises a Perfect Match (PM) and a Mismatch (MM) probe which are statistically evaluated [11]. Unfortunately, this procedure cannot solve the problem of cross-hybridization completely [12] and further refinements are suggested [13]. For example, Wu *et al.* [7] stated that the MM probes can also cross-hybridize, even though by another mechanism as the PM probes. Therefore, they recommended ignoring the MM probes.

Generally, expressed transcripts are represented on the array by a series of probepairs called probesets. The signal intensities are summarized to a single value per probeset. A large number of single transcripts are represented by multiple probesets. Multiple probesets representing the same gene are expected to show similar fold changes calculated from the signal intensities of the hybridized samples. However, this is in fact not the case [14], [15], [16]. This problem arises from single probes in the probeset which are capable of cross-hybridization. Ways to deal with this problem is either a probe-based analysis, leaving out the probe-to-probeset summarization step [17], [18], or the composition of the probesets could be improved by setting up alternative Chip Definition Files (CDFs) based on information contained in different sequence databases. For example, the group of Ferrari *et al.* [19] created a set of custom CDFs based on the GeneAnnot database [20]. In these CDFs the probesets that match the same gene were merged into one probeset. Hence, the existence of more than one probeset per gene was eliminated, avoiding discordant expression signals for the same transcript.

Another set of custom CDFs relying on a broad repertoire

of databases like RefSeq or Unigene has been created by the group of Dai *et al.* [21]. Probesets matching the same gene were merged, but remained divided if they were able to discriminate different isoforms of a gene. Probes causing cross-hybridizations were removed from the new probesets, but the filter had been not very strict.

Several groups dealt with the question of the minimum probeset size [19], [21]. For example, the group of Lu *et al.* [22] sets the minimum probeset size to 4 probes because smaller probesets result in high error rates. In this study the minimum probeset size was set to 4 [19], [21]. From these new probesets custom CDFs and the corresponding Bioconductor libraries for Affymetrix GeneChips were created.

In the work presented here, a new set of CDFs is introduced avoiding putative cross-hybridization completely. These CDFs are compared with those from Affymetrix, Ferrari, and Dai by validation of the respective microarray results using qRT-PCR for two different datasets.

2. Results

Four different Affymetrix GeneChip arrays are analyzed, namely the HG-U133A, HG-U133B, HG-U133 Plus 2.0 designed for human samples, and the Mouse Genome 430 2.0 array. For the detection of putative cross-hybridizations, the sequences of all Affymetrix probes (only the PM probes, the MM probes are discarded) are aligned against the RefSeq database using blastn [23] as described in the methods section.

The GeneChip HG-U133A consists of 22,283 probesets, each of 11–20 probepairs and 247,937 probepairs in total. Additional 1,155 probepairs are controls and are furthermore ignored. About 44 % of the PM probes (109,245) match exactly one single gene. 11 % of the probes (26,159) do not match any annotated gene. 45 % of the probes (112,533) match more than one gene and thus have cross-hybridization potential.

Furthermore, the direction of the probes was analyzed. Normally, sense strand RNA fragments are expected, although there are some loci in the human genome [24], as well as in the mouse genome [25], where both sense and antisense strands are transcribed. However, mixing up probes detecting sense or antisense strands in one single probeset could cause wrong expression results. Here, only probes matching the sense strand are considered as correct. For the U133A microarray all probes match the sense strand.

The GeneChip HG-U133B consists of 22,645 probesets, each of 11–20 probepairs and 249,491 probepairs in total. Again, there are additional probesets containing more than 11 probes as controls and are ignored (1,100). About 35 % of the probes (87,067) are found to match exactly one gene. 2 % of the probes (5,453) match more than one gene, so they possibly cross-hybridize, 5 % of the probes (12,805) match at least one gene but in the wrong direction (antisense direction) and no gene in the sense direction, and 58 % of the probes (144,166) do not match any annotated gene.

The GeneChip HG-U133 Plus 2.0 consists of 54,675 probesets and 604,247 probepairs. Like in the other arrays, additional probesets containing more than 11 probes are controls and are discarded. Here, 37 % of the remaining probes (221,821) match exactly one gene, 23 % of the probes (141,146) match more than one gene, 11 % of the probes (65,327) match at least one gene but in the wrong direction (antisense direction) and no gene in the sense direction, and 29 % of the probes (175,953) do not match any annotated gene.

The Mouse Genome 430 2.0 array consists of 45,036 probesets and 496,457 probepairs. About 52 % of the counted probes (257,331) match exactly one gene and 5 % of the probes (27,112) match more than one gene. About 1 % of the probes (4,661) match genes only in the wrong direction and 42 % of the probes (207,353) do not match any annotated gene.

Nearly all Affymetrix probesets contain at least one probe which has cross-hybridization potential. In fact, for the HG-U133 Plus 2.0 Chip about 65 % of all probesets include more cross-hybridizing probes than non-ambiguous ones.

All probes matching exactly one single gene are classified as good and all probes matching more than one gene are classified as problematic. Those probes, that match in the wrong direction or do not match any RefSeq sequence are also classified as problematic. Only the good probes are used to create the new CDFs as described in the methods chapter. Accordingly, for the HG-U133A microarray originally measuring 14,500 genes by 22,283 probesets the newly created CDF contains 12,400 probesets representing 12,400 genes. For the HG-U133 Plus 2.0 the number of probesets is reduced from 54,675 (representing 38,500 genes) to 18,800 (representing 18,800 genes). The HG-U133B comprises 22,645 probesets measuring the expression of 18,400 genes. Here, the number of probesets is reduced to 6,500 matching 6,500 transcripts. The Mouse 430 2.0 microarray consists of 45,036 probesets for 39,000 genes. With the new CDF there are 16,400 probesets matching 16,400 genes. Hence, the number of identifiable genes is reduced in order to achieve a higher specificity of the probesets. The result for the HG-U133 Plus 2.0 is in good agreement to the results of Barnes *et al*. [26], who used BLAT and the Golden Path database and achieved a number of 17,143 genes that can be measured.

Small probesets lead to higher error rates and result in lower statistical significance. In the Affymetrix CDFs the size is 11 for nearly all probesets, but in the newly created probesets the size is not fixed. Some probesets are smaller than those from Affymetrix due to the removal of the problematic probes. However, many probesets increase in size due to useful probes on the array that have not been used for the matching gene before and probesets measuring

the same gene beeing merged. For example, for the HG-U133 Plus 2.0 the mean probeset size increases from 11 to 17.

For the validation of all CDFs two test datasets are chosen: (i) the Etanercept (ETC) and (ii) the MAQC dataset. The first of the two datasets is derived from a study analyzing the effect of the TNF-α blocker Etanercept, a rheumatoid arthritis drug, using data from 17 patients at three time points [27]. It is a typical dataset that arises in medical studies and is rather representative. One Affymetrix HG-U133A array experiment was performed for each time point. The second dataset is the Microarray Quality Control (MAQC) reference dataset [28]. It contains data from more than 1,300 microarrays and qRT-PCR data for more than 1,000 genes. The subset of the 120 Affymetrix U133 Plus 2.0 expression results and all the qRT-PCRs are selected for the analysis presented here.

qRT-PCR results are considered to reflect the real transcript concentrations with higher reliability than those determined by microarrays. Therefore, qRT-PCR experiments are regarded as a 'gold standard' for chip analyses [29], [30]. The Pearson correlation coefficient (PCC) of the microarray and the qRT-PCR data is computed for each gene using the different CDFs.

For the Etanercept dataset we performed qRT-PCR experiments for 16 genes. In total, this dataset now contains results from 51 microarrays and 816 qRT-PCR experiments. In addition, the genes with qRT-PCR data in both records are analyzed in more detail.

The perfomance of these CDFs were compared: the original Affymetrix CDFs (A), the two alternative CDFs of Ferrari *et al.* (F) [19] and Dai *et al.* (D) [21], and the new CDFs (H) presented here. The CDFs from Ferrari, using the GeneAnnot database, contain merged probesests (see background chapter), and cross-hybridization was not considered. The group of Dai offers a broad spectrum of different CDFs based on different databases. The one using RefSeq is chosen for comparison because it corresponds best to the new CDFs, using RefSeq as well. In the Dai CDFs different probesets matching a single gene are combined, although there are exceptions for genes comprising different isoforms. A check for cross-hybridization is also included. However, it applies a different algorithm than the new CDFs and the filter is much less strict.

For the probe to probeset summarization step two algorithms are used as described in the methods section: (i) the Robust Multi-array Analysis Algorithm (RMA) [13], [31] and (ii) the Affymetrix Microarray Suite MAS5 [32]. These were compared repeatedly, but it is difficult or even impossible to decide which of the both algorithms performs better in any case [33], [34], [35].

For the Etanercept dataset, the mean correlation coefficient of all 16 genes for the Affymetrix CDF is 0.61 using the robust multi-array analysis algorithm (RMA) and 0.60 using the Affymetrix Microarray Suite MAS5. These values include 31 probesets in total matching these 16 genes according to the Affymetrix annotation file. If only the best correlating probeset for each gene is considered, the average correlation coefficient increases to 0.73 for RMA and 0.71 for MAS5. However, this value is more of theoretical interest because the knowledge which probeset will perform best is gained not until the qRT-PCR experiments and correlation analysis is finished. On average, the incorporated probesets contain 5.58 putative cross-hybridizations calculated by BLAST (4.47 including only the best performing probesets).

The Dai CDF contains 23 probesets for the 16 genes of the Etanercept dataset. Their mean correlation coefficient increases to 0.67 for both RMA and MAS5 compared to the 0.60 using the Affymetrix CDF. Considering the best correlating Dai probesets only, the values further increase to 0.73 for RMA and 0.69 using MAS5. The mean size of the Dai probesets increases to 20.59 probes containing 8.82 putative cross-hybridizations. This number changes to 4.71 if normalized to a probeset size of 11. Here, normalization means the number of putative cross-hybridizations calculated for a hypothetical Dai probeset size of 11. Considering only the best Dai probesets, the number of putative cross-hybridizations decreases to 7.88 on average.

For the Ferrari CDF, the mean correlation coefficient equals 0.73 for RMA and 0.69 using MAS5 on average. The mean probeset size increases to 19.56, harboring 10.81 possible cross-hybridizations (6.07 if normalized).

Using the new CDF the mean correlation coefficient amounts to 0.72 for RMA and 0.68 for MAS5. The mean probeset size decreases to 10.25 with no cross-hybridizations at all. The detailed results are shown in the table below:

Gene	Probeset	PCC ETC (RMA)	PCC ETC (MAS5)	PCC MAQC (RMA)	Number of ambiguous probes	Probeset-size
TNF	A: 207113_s_at	0.88	0.85	N/A	8	11
	D: NM_000594_at	0.88	0.85	N/A	8	11
	F: GC06P031652_at	0.88	0.85	N/A	8	11
	H: gi_25952110	0.86	0.81	N/A	0	3
IL1B	A: 205067_at	0.95	0.90	0.37	6	11
	A: 39402_at	0.95	0.87	0.82	6	16
	D: NM_000576_at	0.96	0.89	0.74	12	27
	F: GC02M113303_at	0.96	0.89	0.74	12	27
	H: gi_27894305	0.95	0.88	0.86	0	15
IL6	A: 205207_at	0.69	0.71	0.81	3	11
	D: NM_000600_at	0.69	0.71	0.81	3	11
	F: GC07P022732_at	0.69	0.71	0.81	3	11
	H: gi_10834983	0.65	0.72	0.71	0	8
IL8	A: 202859_x_at	0.88	0.81	0.90	6	11
	A: 211506_s_at	0.86	0.73	0.98	6	11
	D: NM_000584_at	0.88	0.73	0.96	12	22
	F: GC04P074845_at	0.88	0.73	0.96	12	22
	H: gi_28610153	0.89	0.73	0.95	0	10
IL1RN	A: 212657_s_at	0.75	0.87	N/A	2	11
	A: 212659_s_at	0.77	0.84	N/A	4	11
	A: 216243_s_at	0.75	0.86	N/A	6	11
	A: 216244_s_at	0.13	0.07	N/A	4	11
	A: 216245_at	0.21	0.11	N/A	10	11
	D: NM_173841_at	0.80	0.88	N/A	12	33
	D: NM_000577_at	0.80	0.88	N/A	12	33
	D: NM_173842_at	0.80	0.88	N/A	12	33
	D: NM_173843_at	0.84	0.86	N/A	15	42
	F: GC02P113591_at	0.83	0.86	N/A	16	44
	H: gi_27894315	0.78	0.88	N/A	0	23
ICAM1	A: 202637_s_at	0.63	0.73	0.97	7	11
	A: 202638_s_at	0.62	0.72	0.98	4	11
	A: 215485_s_at	0.71	0.73	0.94	3	11
	D: NM_000201_at	0.70	0.76	0.99	14	33
	F: GC19P010247_at	0.70	0.77	0.99	14	33
	H: gi_4557877	0.72	0.74	0.97	0	20
SOD2	A: 215078_at	0.25	0.35	N/A	10	11

Continued on next page

Gene	Probeset	PCC ETC (RMA)	PCC ETC (MAS5)	PCC MAQC (RMA)	Number of ambiguous probes	Probeset-size
	A: 215223_s_at	0.15	0.28	N/A	7	11
	A: 216841_s_at	0.18	0.39	N/A	3	11
	A: 221477_s_at	0.32	0.44	N/A	10	11
	D: NM_001024466_at	0.16	0.33	N/A	6	12
	D: NM_000636_at	0.19	0.37	N/A	10	22
	D: NM_001024465_at	0.16	0.33	N/A	6	13
	F: GC06M160020_at	0.20	0.36	N/A	20	33
	H: gi_67782304	0.20	0.39	N/A	0	12
TRAF1	A: 205599_at	0.61	0.50	0.88	6	11
	D: NM_005658_at	0.61	0.50	0.88	6	11
	F: GC09M122704_at	0.61	0.50	0.88	6	11
	H: gi_53759116	0.59	0.47	0.89	0	5
ZFP36	A: 201531_at	0.84	0.86	N/A	5	11
	A: 213890_x_at	-0.01	-0.46	N/A	8	11
	D: NM_003407_at	0.84	0.86	N/A	5	11
	F: GC19P044589_at	0.84	0.86	N/A	5	11
	H: gi_141802261	0.85	0.82	N/A	0	6
PTGS2	A: 204748_at	0.91	0.71	0.97	4	11
	D: NM_000963_at	0.91	0.71	0.97	4	11
	F: GC01M184907_at	0.91	0.71	0.97	4	11
	H: gi_4506264	0.89	0.72	0.95	0	9
TNFAIP3	A: 202643_s_at	0.78	0.82	0.97	4	11
	A: 202644_s_at	0.87	0.85	0.93	6	11
	D: NM_006290_at	0.82	0.83	0.96	10	22
	F: GC06P138230_at	0.82	0.83	0.96	10	22
	H: gi_26051241	0.80	0.82	0.98	0	13
DUSP2	A: 204794_at	0.75	0.66	N/A	5	11
	D: NM_004418_at	0.75	0.66	N/A	5	11
	F: GC02M096230_at	0.75	0.66	N/A	5	11
	H: gi_12707563	0.74	0.60	N/A	0	6
ADM	A: 202912_at	0.80	0.67	0.92	5	11
	D: NM_001124_at	0.80	0.67	0.92	5	11
	F: GC11P010283_at	0.80	0.67	0.92	5	11
	H: gi_4501944	0.82	0.67	0.94	0	6
CROP	A: 203804_s_at	0.44	0.56	N/A	5	11
	A: 208835_s_at	0.43	0.36	N/A	5	11
	A: 220044_x_at	0.43	0.44	N/A	4	11
	D: NM_016424_at	0.49	0.50	N/A	13	32
	D: NM_006107_at	0.49	0.45	N/A	13	30
	F: GC17P046151_at	0.48	0.48	N/A	14	33
	H: gi_52426741	0.46	0.47	N/A	0	17
NFκBIA	A: 201502_s_at	0.81	0.73	N/A	4	11
	D: NM_020529_at	0.81	0.73	N/A	4	11
	F: GC14M034940_at	0.81	0.73	N/A	4	11
	H: gi_10092618	0.82	0.77	N/A	0	7
JUNB	A: 201473_at	0.44	0.44	0.94	7	11
	D: NM_002229_at	0.44	0.44	0.94	7	11
	F: GC19P012763_at	0.44	0.44	0.94	7	11
	H: gi_44921611	0.54	0.44	0.73	0	4
Ø	all Affymetrix	0.61	0.59	0.88	5.58	11.16
	best Affymetrix	0.73	0.71	0.92	4.47	11.00
	Dai	0.67	0.67	0.91	8.82	20.59
	best Dai	0.73	0.69	0.91	7.88	18.69
	Ferrari	0.73	0.69	0.91	10.81	19.56
	Hummert	0.72	0.68	0.89	0.00	10.25

Evaluating the PM and MM probes statistically, the MAS5 software assigns 'present', 'absent' or 'marginal' to each expression value, and Affymetrix recommends to use only the 'present' detection call for further analysis. Following this recommendation and using only those results for the correlation analysis that are marked as 'present' the mean correlation coefficient increases from 0.59 to 0.66 (0.74 including only the best performing probesets). Hence, incorporating the Affymetrix detection call indeed improves the correlation, but using alternative CDFs is still better than using the Affymetrix probesets and the detection call.

Analyzing the MAQC reference dataset using the RMA suite, the results are almost in accordance with those of the Etanercept data described above. The mean correlation coefficient for all 1,000 genes is 0.47 for the Affymetrix CDF (0.71 incorporating only the best probeset for each gene). Using the Dai CDF, the mean correlation increases to 0.63 (0.64 for the best probesets). With the Ferrari and the new CDF the mean correlations are 0.63 and 0.58, respectively. The detailed results for all MAQC genes can be downloaded.

Discussion

Results from microarray experiments contain considerably high error rates [36]. Due to error propagation, it is of particular importance to minimize errors in the beginning of the analysis chain [37]. Therefore, especially the pre-processing of the chip data has to be done as accurate as possible. Many efforts were spent on these problems before [38], such as the notable results of the 'Golden Spike Project' [6]. The question which statistical method should be adequately chosen is even more complicated if experimental data from different laboratories are incorporated in one single analysis [39].

For microarray analyses algorithms are essential which combine the 11-20 probepair intensities for a given gene and define a measure of expression that represents the amount of the corresponding mRNA species. In this study, two of these algorithms are compared, the robust multi-array analysis algorithm (RMA) and the Affymetrix Microarray Suite MAS5. Applying both algorithms to the Etanercept dataset RMA outperforms MAS5 on average. Other studies revealed similar results. However, their performance is assumed to be dependent on the actual dataset [40]. In fact, normalisation steps are applied after the probe to probeset summarization. Some of these steps depend on global parameters (e.g. mean of total gene expression) which depend on the total set of probesets. Therefore, identical probesets within different CDFs vary slightly in the final gene expression values.

Analyzing the probes of the Affymetrix microarrays discloses many inaccuracies. A large number of problematic probes are based on the fact that Affymetrix had to rely on genome annotation available at the time the chips were designed (U133A and U133B: 2001; U133 Plus 2.0 and Mouse 430 2.0: 2003). Because genome annotation improves permanently, the chip design does not properly match the present annotations anymore. Due to compatibility reasons, Affymetrix is not able to keep the design of their microarrays up to date.

The problem of cross-hybridization is well known. The first work on custom CDFs examining this error source was published by the group of Dai in 2005 [21]. They created a large amount of high quality custom CDFs related to different reference databases. Some probes, causing cross-hybridizations, are deleted from the probesets, but the filter is quite loose, so the number of problematic probes decreased but did not vanish. The use of the new CDFs can avoid full length, i.e., 25 mer long, cross-hybridizations completely. Cross-hybridization of shorter fragments are very difficult to handle due to the fact that the number of putative bindings grows exponentially the shorter the considered fragments are. Hence, if all putatively cross-hybridizing probes are excluded the amount of measurable genes will be reduced extremely.

The underlying gene annotation which is used for sequence alignment has a big impact on the number of cross-hybridizations. Manually curated mRNA sequences have a high chance of missing transcripts. Therefore, the inclusion of computational proposed gene annotations decreases the

number of false negative predicted cross-hybridizations. The drawback is that a number of false positive hybridizations increases. A more strict approach should be preferred, because it does not significantly decrease the number of covered transcripts as there is a high amount of availabe probes. In this study, the exclusion of XM-RefSeq-accessions results in smaller differences between the different CDFs in the number of putative cross-hybridzing transcripts. Interestingly, the correlation coeefficents of the newly created probesets do not change significantly.

Evaluating the four different CDFs, we figured out that the usage of the original Affymetrix CDFs leads to poorer results than the usage of the custom CDFs, although the best Affymetrix probesets give equally good or even better results than the other CDFs. However, as already mentioned, this cannot be taken into account, because it is not known which probeset will perform best before the correlation analysis is completed. The Dai probesets perform better, but the problem of several probesets representing a single gene had not been solved. Although multiple probesets representing the same gene are expected to show similar signal intensities, this is in fact not the case [14], [15]. Thus, it is difficult to decide which of the probesets matching the same gene is the most reliable. The Ferrari and the new CDFs comprise only one probeset per gene, which is of great advantage. The Ferrari CDFs perform slightly better on the Etanercept dataset and both CDFs perform equally well on the MAQC data.

The analysis of the genes for which qRT-PCR results are available in the Etanercept dataset as well as in the MAQC dataset clearly shows higher correlation coefficients in the MAQC dataset. This is most likely due to the fact that the U133 Plus 2.0 arrays which were used in the MAQC dataset outperform the older U133A microarrays.

The results show that probesets consisting of more probes, i.e., larger probesets, lead to better correlation results in general, whereas smaller probesets perform poorer. This finding correlates to the results of the study of Cui *et al.* [14] that merges probesets matching the same transcript. Interestingly, probesets containing many putative cross-hybridizations do not considerably perform poorer than probesets containing only a few. This result is very surprising, because it is obvious that cross-hybridization is one of the main error sources in microarray experiments [8], [9]. The normalization step in the two summarizing algorithms RMA and MAS5 may explain for that because they possibly eliminate some cross-hybridization effects. Another explanation is that leaving out the problematic probes does not compensate the influence of cross-hybridization. Unspecific binding leads to two types of error: (i) false-positives because RNA fragments bind to problematic probes of the probeset, and (ii) gene expression events are missed or underestimated, leading to a false-negative error if the RNA fragments are already bound to problematic probes of other probesets (competitive binding). Custom CDFs can only account for the first type of error by leaving out the problematic probes, the second effect could only be overcome by better array design.

The newly created CDFs perform slightly poorer than the Ferrari probesets (0.72 vs. 0.73) on the Etanercept dataset and equally well on the much larger MAQC dataset. On the one hand, the Ferrari CDFs can obviously countervail the negative effect by their much larger probesets in comparison to the new CDFs. On the other hand, using the new CDFs, putative cross-hybridizations are systematically excluded whereas using the Ferrari CDFs, the negative effect vanishes for statistical reasons due to the larger probesets. For exact studies, it is better to avoid a putative error source instead of averaging the cross-hybridization effects out as the Ferrari CDFs do. In addition, it has to be mentioned that the new CDFs provide as good or better results as the other CDFs using only about half the amount of probes (HG-U133A: 44 %, HG-U133B: 35 %, HG-U133 Plus 2.0: 37 %, Mouse Genome 430 2.0 Array: 52 %). Hence, designing new microarrays without the problematic probes, the dimension can be reduced by half without loosing any information and minimize the costs of the technology tremendously. Future microarray design using only the good probes and incorporating probesets of large sizes like in the Ferrari CDFs will certainly provide optimal solutions.

Methods

Probe Analysis

For the detection of putative cross-hybridizations by sequence alignment, the sequences of all Affymetrix probes (only the PM probes, the MM probes are discarded) are aligned against the RefSeq database using blastn [23]. For the U133A and the U133 Plus 2.0 the RefSeq release from 05/14/07 was used (download from ftp://ftp.ncbi.nih.gov/-refseq/H_sapiens/mRNA_Prot/human.rna.fna.gz), for the U133B the realease from 01/10/08, and for the Mouse 430 2.0 microarray the release from 05/09/08 (∼M_musculus/mRNA_Prot/mouse.rna.fna.gz) was used. These parameters were applied: ValW = 7, ValE = 1000, ValHspmax = 1.

In this work all those RefSeq accession numbers beginning with XM or NM are used. The XM-identifiers indicate mRNA-RefSeq-accessions which are produced by computationally annotated genome submissions. The NM-identifier show that the RefSeq records are subsequently curated. Using both accessions in our model leads to more predicted cross-hybridizations which increases the reliability of the specificity of the probes.

The strand direction of the probes is analyzed. For each probe it is counted how many genes match and checked whether the match has the correct direction, i.e., the sense direction.

All BLAST hits for different transcript isoforms are merged, i.e., if the probe hybridizes to alternative splice variants of one gene but not to another gene, it is considered as unambiguous. Different gene isoforms of one gene are identified by screening the gene descriptions of the RefSeq database.

All probes matching only one single gene are classified as good and all probes matching more than one gene are classified as problematic. Those probes that match in the wrong direction or do not match any RefSeq sequence are also classified as problematic. For the creation of the new CDFs only the good probes are used. The probe sequences are annotated with GeneIDs derived from RefSeq. The GeneID is a database cross-reference qualifier, which supports access to the Entrez Gene database and provides a distinct tracking identifier for a gene or locus. Probes sharing the same GeneID are grouped together into a new probeset. The intersection between two different probesets is therefore always empty for all probesets. The size of the newly created probesets is variable and not fixed to 11 like in the Affymetrix CDFs.

Datasets

Two datasets were chosen for the validation of the different CDFs. The first of the two datasets chosen is derived from a study published by Koczan *et al.* [27] analyzing the effect of the TNF-α blocker Etanercept, a rheumatoid arthritis drug, using data from 17 patients at three time points. One Affymetrix HG-U133A array was performed for each time point. The data are available at the Array Express archive [41] with the accession number E-MTAB-11.

Expression levels of 16 genes were measured by quantitative real-time RT-PCR (qRT-PCR) performed with TaqMan assay reagents according to the manufacturer's instructions on a 7900 High Throughput Sequence Detection System (Applied Biosystems, Foster City, CA, USA) using predesigned primers and probes (GAPDH Hs99999905_m1, ICAM1 Hs00164932_m1, TNFAIP3 Hs00234713_m1, IL1B Hs00174097_m1, NFκBIA Hs00153283_m1, IL8 Hs00174103_m1, ADM Hs00181605_m1, TNF Hs00174128_m1, IL6 Hs00174131_m1, IL1RN Hs00277299_m1, SOD2 Hs00167309_m1, TRAF1 Hs00194638_m1, ZFP36 Hs00185658_m1, PTGS2 Hs00153133_m1, DUSP2 Hs00358879_m1, CROP Hs00538879_s1, JUNB HS00357891_s1).

The threshold cycle values (C_T) for specific mRNA expression in each sample were normalized to the C_T values of GAPDH mRNA in the same sample. This provides ΔC_T values that were used for the correlation analysis. In total, 816 qRT-PCR experiments were performed and complement the 51 microarray experiments (17 patients, 3 time points) described in [27]. The results of the qRT-PCR experiments can be downloaded.

The second dataset is the Microarray Quality Control (MAQC) reference dataset [28]. It contains data from more than 1,300 microarrays and qRT-PCR data for more than 1,000 genes. All available 120 Affymetrix U133 Plus 2.0 expression results and all the qRT-PCRs are selected for the analysis presented here. The MAQC data discussed in this publication are available in NCBI's Gene Expression Omnibus with accession number GSE5350. In addition, the nine genes for which qRT-PCR results are available in both datasets, are analyzed in more detail.

Comparison of the CDFs

For the comparison of different CDFs, the correlation between the microarray and the qRT-PCR experiments is used [29], [30]. As a performance index the Pearson correlation coefficient of the microarray results and the qRT-PCR experiments is calculated. Calculation of the Spearman correlation coefficient showed very similar results (data available at http://sysbio.hki-jena.de/software).

The raw chip data (CEL Files) are analyzed using the Robust Multi-array Analysis Algorithm (RMA) [13], [31] and the Affymetrix Microarray Suite MAS5 [32] in combination with the different CDFs.

The MAS5 software assigns 'present', 'absent' or 'marginal' to each expression value, and Affymetrix recommends to use only the 'present' detection call for further analysis [32]. For an additional correlation analysis only the 'present' probesets are used to check if the calculated detection call from MAS5 gives a good prediction for the probeset quality.

Availability

The newly created CDFs as R-packages and additional files are available for download at http://www.sysbio.hki-jena.de/software. Using the CDFs does not interfere with all further steps of microarray analysis.

Acknowledgements

This work was supported by the ILRS - International Leibniz Research School for Microbial and Molecular Interactions (CH, FH) and by the ERASysBio+ project Linconet (MW).

References

[1] S. Heber and B. Sick, "Quality assessment of Affymetrix GeneChip data," *OMICS A Journal of Integrative Biology*, vol. 10, no. 3, pp. 358–368, Fall 2006.

[2] O. Modlich and M. Munnes, "Statistical framework for gene expression data analysis," *Methods in Molecular Biology*, vol. 377, pp. 111–130, May 2007.

[3] P. K. Tan, T. J. Downey *et al.*, "Evaluation of gene expression measurements from commercial microarray platforms," *Nucleic Acids Research*, vol. 31, no. 19, pp. 5676–5684, October 2003.

[4] A.-K. Järvinen, S. Hautaniemi *et al.*, "Are data from different gene expression microarray platforms comparable?" *Genomics*, vol. 83, no. 6, pp. 1164–1168, June 2004.

[5] A. Brazma, P. Hingamp *et al.*, "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data," *Nature Genetics*, vol. 29, no. 4, pp. 365–371, December 2001.
[6] S. E. Choe, M. Boutros *et al.*, "Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset," *Genome Biology*, vol. 6, no. 2, p. R16, January 2005.
[7] C. Wu, R. Carta, and L. Zhang, "Sequence dependence of cross-hybridization on short oligo microarrays," *Nucleic Acids Research*, vol. 33, no. 9, p. e84, May 2005.
[8] Z. Chen, M. McGee *et al.*, "A distribution free summarization method for Affymetrix GeneChip® arrays," *Bioinformatics*, vol. 23, no. 3, pp. 321–327, February 2007.
[9] A. C. Cambon, A. Khalyfa *et al.*, "Analysis of probe level patterns in Affymetrix microarray data," *BMC Bioinformatics*, vol. 8, no. 146, May 2007.
[10] H. R. Ueda, S. Hayashi *et al.*, "Universality and flexibility in gene expression from bacteria to human," *The Proceedings of the National Academy of Sciences (US)*, vol. 101, no. 11, pp. 3765–3769, March 2004.
[11] Affymetrix Inc, "GeneChip custom express array design guide. part no. 700506 rev. 4," Tech. Rep., 2003.
[12] L. Zhang, M. F. Miles, and K. D. Aldape, "A model of molecular interactions on short oligonucleotide microarrays," *Nature Biotechnology*, vol. 21, no. 7, pp. 818–821, July 2003.
[13] B. M. Bolstad, R. A. Irizarry *et al.*, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 3, pp. 185–193, January 2003.
[14] X. Cui and A. E. Loraine, "Consistency analysis of redundant probe sets on Affymetrix three-prime expression arrays and applications to differential mRNA processing," *PLoS One*, vol. 4, no. 1, p. 4229, January 2009.
[15] T. R. Hughes, M. Mao *et al.*, "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer," *Nature Biotechnology*, vol. 19, no. 4, pp. 342–347, April 2001.
[16] M. A. Stalteri and A. P. Harrison, "Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips," *BMC Bioinformatics*, vol. 8, no. 13, January 2007.
[17] X. Liu, M. Milo *et al.*, "Probe-level measurement error improves accuracy in detecting differential gene expression," *Bioinformatics*, vol. 22, no. 17, pp. 2107–2113, September 2006.
[18] G. Sanguinetti, M. Milo *et al.*, "Accounting for probe-level noise in principal component analysis of microarray data," *Bioinformatics*, vol. 21, no. 19, pp. 3748–3754, October 2005.
[19] F. Ferrari, S. Bortoluzzi *et al.*, "Novel definition files for human GeneChips based on GeneAnnot," *BMC Bioinformatics*, vol. 8, no. 446, November 2007.
[20] V. Chalifa-Caspi, I. Yanai *et al.*, "GeneAnnot: Comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes," *Bioinformatics*, vol. 20, no. 9, pp. 1457–1458, June 2004.
[21] M. Dai, P. Wang *et al.*, "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data," *Nucleic Acids Research*, vol. 33, no. 20, p. e175, November 2005.
[22] J. Lu, J. C. Lee *et al.*, "Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays," *BMC Bioinformatics*, vol. 8, no. 108, March 2007.
[23] S. McGinnis and T. L. Madden, "BLAST: at the core of a powerful and diverse set of sequence analysis tools," *Nucleic Acids Research*, vol. 32, pp. W20–W25, July 2004.
[24] R. Yelin, D. Dahary *et al.*, "Widespread occurrence of antisense transcription in the human genome," *Nature Biotechnology*, vol. 21, no. 4, pp. 379–386, April 2003.
[25] H. Kiyosawa, N. Mise *et al.*, "Disclosing hidden transcripts: Mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized," *Genome Research*, vol. 15, no. 4, pp. 463–474, April 2005.
[26] M. Barnes, J. Freudenberg *et al.*, "Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms," *Nucleic Acids Research*, vol. 33, no. 18, pp. 5914–5923, October 2005.
[27] D. Koczan, S. Drynda *et al.*, "Molecular discrimination of responders and nonresponders to anti-TNFalpha in rheumatoid arthritis therapy by Etanercept," *Arthritis Research & Therapy*, vol. 10, p. R50, May 2008.
[28] L. Shi, L. H. Reid *et al.*, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, no. 9, pp. 1151–1161, September 2006.
[29] J. S. Moray, J. C. Ryan, and F. M. Van Dolah, "Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR," *Biological Procedures Online*, vol. 8, no. 1, pp. 175–193, December 2006.
[30] R. D. Canales, Y. Luo *et al.*, "Evaluation of DNA microarray results with quantitative gene expression platforms," *Nature Biotechnology*, vol. 24, no. 9, pp. 1115–1122, September 2006.
[31] R. A. Irizarry, B. Hobbs *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, April 2003.
[32] Affymetrix Inc, "Statistical algorithms description document. whitepaper. part no. 701137 rev. 3," Tech. Rep., 2002.
[33] R. A. Irizarry, Z. Wu, and H. A. Jaffee, "Comparison of Affymetrix GeneChip expression measures," *Bioinformatics*, vol. 22, no. 7, pp. 789–794, July 2006.
[34] J. Seo and E. P. Hoffman, "Probe set algorithms: is there a rational best bet?" *BMC Bioinformatics*, vol. 7, no. 395, August 2006.
[35] S. D. Pepper, E. K. Saunders *et al.*, "The utility of MAS5 expression summary and detection call algorithms," *BMC Bioinformatics*, vol. 8, no. 273, July 2007.
[36] M. Eisenstein, "Microarrays: Quality control," *Nature*, vol. 442, pp. 1067–1070, August 2006.
[37] M. Grabe, *Measurement Uncertainties in Science and Technology*. New York: Springer Press, 2005.
[38] P. Boutros, "Systematic evaluation of the microarray analysis pipeline," in *Proceedings of the First 11th MGED Meeting: 1-4 September 2008; Riva del Garda*, G. Sherlock, Ed. MGED, 2008, pp. 16–27.
[39] H.-C. Liu, C.-Y. Chen *et al.*, "Cross-generation and cross-laboratory predictions of Affymetrix microarrays by rank-based methods," *Journal of Biomedical Informatics*, vol. 41, no. 4, pp. 570–579, August 2008.
[40] K. Shedden, W. Chen *et al.*, "Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data," *BMC Bioinformatics*, vol. 6, no. 26, 2005.
[41] H. Parkinson, M. Kapushesky *et al.*, "ArrayExpress update — from an archive of functional genomics experiments to the atlas of gene expression," *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D868–D872, January 2009.

3.2 Gene Expression Regulation underlying Osteo-, Adipo-, and Chondro-Genic Lineage Commitment of Human Mesenchymal Stem Cells

Chapter 4

Gene Expression Regulation underlying Osteo–, Adipo–, and Chondro–Genic Lineage Commitment of Human Mesenchymal Stem Cells

Ana M. Sotoca
Radboud University, The Netherlands

Michael Weber
Hans Knöll Institute, Germany

Everardus J. J. van Zoelen
Radboud University, The Netherlands

Medical Advancements in Aging and Regenerative Technologies:

Clinical Tools and Applications

Andriani Daskalaki
Max-Planck Institute for Molecular Genetics, Germany

Medical Information Science Reference

Managing Director: Lindsay Johnston
Editorial Director: Joel Gamon
Book Production Manager: Jennifer Romanchak
Publishing Systems Analyst: Adrienne Freeland
Development Editor: Hannah Abelbeck
Assistant Acquisitions Editor: Kayla Wolfe
Typesetter: Travis Gundrum
Cover Design: Nick Newcomer

Published in the United States of America by
Medical Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Medical advancements in aging and regenerative technologies: clinical tools and applications / Andriani Daskalaki, editor.
p. cm.
Includes bibliographical references and index.
Summary: “This book translates basic science discoveries into regenerative therapies with the application of clinical tool in aging and tissue regeneration”-- Provided by publisher.
ISBN 978-1-4666-2506-8 (hardcover) -- ISBN 978-1-4666-2507-5 (ebook) -- ISBN 978-1-4666-2508-2 (print & perpetual access) 1. Regenerative medicine. 2. Regeneration (Biology) I. Daskalaki, Andriani, 1966-
QH499.M43 2013
571.8’89--dc23

2012023095

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 4
Gene Expression Regulation underlying Osteo-, Adipo-, and Chondro-Genic Lineage Commitment of Human Mesenchymal Stem Cells

Ana M. Sotoca
Radboud University, The Netherlands

Michael Weber
Hans Knöll Institute, Germany

Everardus J. J. van Zoelen
Radboud University, The Netherlands

ABSTRACT

Human mesenchymal stem cells have a high potential in regenerative medicine. They can be isolated from a variety of adult tissues, including bone marrow, and can be differentiated into multiple cell types of the mesodermal lineage, including adipocytes, osteocytes, and chondrocytes. Stem cell differentiation is controlled by a process of interacting lineage-specific and multipotent genes. In this chapter, the authors use full genome microarrays to explore gene expression profiles in the process of Osteo-, Adipo-, and Chondro-Genic lineage commitment of human mesenchymal stem cells.

INTRODUCTION

Human Mesenchymal Stem Cells (hMSCs) can be obtained in relatively large numbers from a variety of connective tissues sources including adipose tissue, umbilical cord and bone marrow (De Bari, et al., 2003; Pittenger, et al., 1999; Zuk, et al., 2002). The cells are multipotent cells and can differentiate *in vivo* into a variety of mesenchymal tissues, including bone, muscle, cartilage, and fat. Although they lack specific markers, upon *in vitro* culturing they can be identified by the expression

DOI: 10.4018/978-1-4666-2506-8.ch004

of surface molecules such as CD105 and CD73, while they are negative for the hematopoietic markers CD34, CD45, and CD14 (Chamberlain, Fox, Ashton, & Middleton, 2007). They have the ability to expand many-fold *in vitro* while maintaining their growth potential and multipotency (Bouchez, et al., 2011), giving rise to cultures ranging from narrow spindle shaped to large polygonal cells (Javazon, Beggs, & Flake, 2004). Also *in vitro* they have the ability to differentiate into osteoblasts, chondrocytes, and adipocytes (Dezawa, et al., 2005; Pittenger, et al., 1999). The fact that these cells can be differentiated into several different cell types, in combination with their immune-modulatory properties, make MSCs a promising source of stem cells for tissue repair and gene therapy.

With aging of the population, degenerative diseases such as osteoporosis and arthritis will have an increasing impact. The increase in marrow adipogenesis associated with osteoporosis and age-related osteopenia is well known clinically, and classical *in vitro* and *in vivo* studies strongly support an inverse relationship between the commitment of bone marrow-derived mesenchymal stem cells to the adipoctye and osteoblast lineage pathways (Nuttall & Gimble, 2004). Restoration of damaged bone and cartilage or of an unbalanced cell fate by stimulating hMSCs to differentiate into a specific lineage, provides a novel and attractive therapeutic approach. This interest in developing new therapies with cells that can repair non-hematopoietic tissues is currently of high interest and the first successful clinical trials with MSCs claimed to improve osteogenesis in children with osteogenesis imperfecta (Horwitz, et al., 2001). Currently hMSCs are being employed in clinical trials in heart disease, Crohn's disease, cartilage repair, stroke, spinal cord injury, and several other diseases (Giordano, Galderisi, & Marino, 2007; Körbling & Estrov, 2003; Prockop & Olson, 2007) with positive results. In addition, implanted cell–host interaction needs to be addressed carefully (Shi, et al., 2012), which requires detailed knowledge of the pathways involved in hMSC differentiation, as key factor for understanding normal development and disease processes.

This study aims to apply a high-throughput screening of gene expression regulation underlying lineage commitment in hMSCs to understand tissue development and to identify key genes involved in lineage-specific differentiation. hMSCs were induced to differentiate *in vitro* into three distinct lineages, i.e. bone, cartilage and fat, by applying different culture conditions. The percentage of differentiated cells was determined for each differentiation condition. Analysis of the multiple gene expression data sets, obtained upon specific treatments and time points during the course of lineage-specific differentiation, was used to confirm and understand hMSC fate.

MATERIALS AND METHODS

Culture and Differentiation of Human Mesenchymal Stem Cells

Human Mesenchymal Stem Cells (hMSCs), harvested from normal human bone marrow, were purchased from Lonza (Walkersville, MD) at passage 2. Cells were tested by the manufacturer and were found to be positive by flow cytometry for expression of CD105, CD166, CD29, and CD44, and negative for CD14, CD34, and CD45. We confirmed multipotency of all donor batches based on *in vitro* osteo-, chondro- and adipogenic differentiation capacity. The cells were expanded for no more than 5 passages in 'Mesenchymal Stem Cell Growth Medium' (MSCGM; Lonza, Walkersville, MD) at 37ºC in a humidified atmosphere containing 7.5% CO_2. Studies were performed with hMSCs from multiple donors, including 5F0138 and 1F1061.

Osteogenesis Data Set: For osteogenic differentiation, 2.0 x 10^4 cells per cm^2 were seeded

in MSCGM with 10% fetal bovine serum (a selected lot from Lonza Walkersville, Inc.). The next day, sub-confluent cultures were switched to osteogenic differentiation medium, consisting of high glucose-containing Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (Lonza Walkersville, Inc.), 100 U/ml penicillin, 100 g/ml streptomycin, 10^{-7}M dexamethasone (Dex), 0.2 mM ascorbic acid and 10 mM β-glycerophosphate, in the absence or presence of 250 ng/ml BMP2 (Bone Morphogenetic Protein 2; R&D Systems) or 10^{-8} M 1α, 25-dihydroxyvitamin D3 (1,25(OH)2D3 or Vitamin D; Calbiochem). These osteogenic treatments will be further referred to as DX, DB, and DV, respectively. Cells treated with differentiation medium in the absence of dexamethasone (further referred to as MD) were used as negative controls. Medium was refreshed every 3 days as indicated by Piek and coworkers (Piek, et al., 2010).

Adipogenesis Data Set: For adipogenic differentiation, 4.0 x 10^4 cells per cm^2 were seeded DMEM supplemented with 10% fetal bovine serum (a selected lot from Lonza Walkersville, Inc.), 100 U/ml penicillin, and 100 µg/ml streptomycin. The next day, the cells were switched to adipogenic differentiation medium (AD), which consisted of the above proliferation medium now supplemented with 10^{-6} M dexamethasone, 10 µg/ml insulin (R&D Systems), 10^{-7} M rosiglitazone (Campro Scientific, The Netherlands), and 500 µM IBMX (3-isobutyl-1-methylxanthine, Sigma-Aldrich, St. Louis, MO). Medium was refreshed every 3 days. In parallel cultures cells were also treated with osteogenic differentiation medium supplemented with 250 ng/ml BMP2, using the same cell density and dexamethasone concentration as during adipogenic differentiation (OS). Cells treated with proliferation medium (PR) or dexamethasone alone (DX) were used as controls.

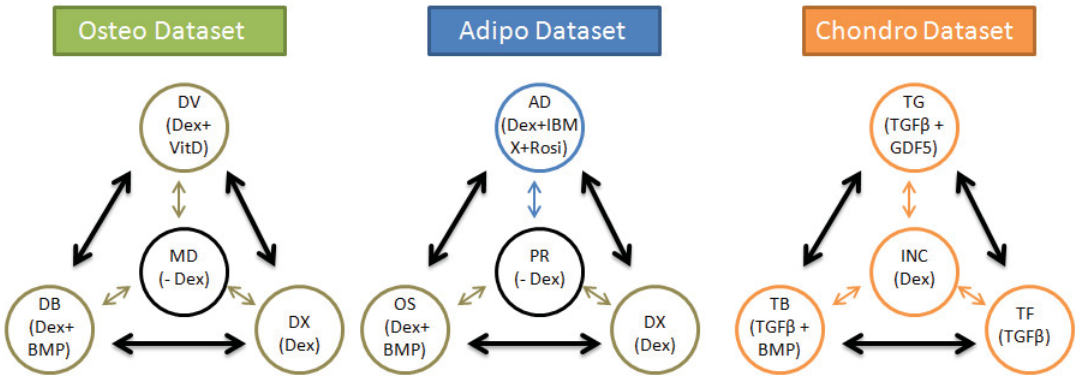
Chondrogenic Data Set: For chondrogenic differentiation, hMSCs were trypsinized and 2.5x 10^5 cells pelleted in a 10 ml round bottom tube (Greiner Bio-One, Monroe, NC) for 10 min at 250xg. Cell pellets were subsequently cultured for 21 days in chondrogenic differentiation medium, consisting of proliferation medium supplemented with 6.25 µg/ml insulin, 6.25 µg/ml transferrin, 6.25 ng/ml sodium selenite, 5.35 µg/ml linoleic acid, 400 µg/ml proline, 1 mg/ml sodium pyruvate, 10^{-7} M dexamethasone, 50 µg/ml sodium L-ascorbate (all obtained from Sigma-Aldrich, St. Louis, MO), in the absence (INC) or presence (TF) of 10 ng/ml recombinant TGFβ1, and supplemented with either 50 ng/ml recombinant human BMP2 (TB) or 50 ng/ml GDF5 (TG). Medium was refreshed every 3 days. Growth factors were obtained from R&D Systems.

Microarray Processing and Identification of Significantly Regulated Genes

In total, 396 RNA samples were obtained from triplicate experiments of all three lineages, four biological conditions and each measured at 11 time points (0, 1, 3, 6, 12, 24, 48, 72, 120, 192, and 288 hours after onset of treatment) as indicated in the experimental design (Figure 1). RNA was extracted using TRIzol® according to the protocol provided by the manufacturer (Invitrogen). For each sample, 5 µg of RNA were reverse transcribed into double-stranded cDNA, and used as a template for the preparation of biotin-labeled cRNA, as previously described (Vaes, et al., 2006). A total of 10 µg of biotin-labeled cRNA was hybridized to the Human Genome U133A Array (Affymetrix), after which hybridization signals were amplified using a streptavidin-biotin amplification procedure. Arrays were hybridized and scanned with a GeneChip G3000 scanner (Affymetrix). Data were quantified using GCOS 1.2 software (Affymetrix). Normalization and statistical analysis of the data were performed

Gene Expression Regulation

Figure 1. Experimental design of the present study based on the three differentiation lineages. For details about the symbols, see materials and methods.



using the error model developed for Data were imported into R (http://www.r-project.org) using Bioconductor (http://www.bioconductor.org) affy Package. The model-based Robust Multiarray Average (RMA) algorithm was used to generate the probe set summary based on the full annotation at the gene level, after which normalization was done according to quantiles method. The Limma algorithm was used to compute a linear model fit and to decide which probe sets should be considered as statistically differentially expressed. Ratios were calculated using Limma in R, applying moderated F-tests. To correct for multiple hypothesis testing, adjusted p-values were obtained using Benjamini and Hochberg correction, indicating the significance of the corresponding ratio. This resulted in a selection of genes significantly regulated during differentiation for each of the (3×10 =) 30 biological conditions tested, i.e. for each combination of osteogenic treatment (MD, DX, DB, DV) and 10 different time points (except time zero), based on triplicate experiments (Piek, et al., 2010). We considered all genes with an adjusted p-value < 10^{-5} and a log-fold-change > 1 as significant. The use of p-value and additional fold-change criterion was chosen to ensure that selected genes showed at least a doubling of their expression level over time.

Alkaline Phosphatase Assay on Osteoblasts

To quantify Alkaline Phosphatase (ALP) enzymatic activity, hMSCs were seeded in 96-well tissue culture plates and cultured for 7 days under the above osteogenic differentiation conditions. ALP activity was measured enzymatically and corrected for differences in cell number, based on neutral red staining (Piek, et al., 2010). For histochemical analysis of ALP activity, cells were fixed with 1% formaldehyde. After washing with Phosphate-Buffered Saline (PBS), cells were incubated for 1 hour at 37°C in a mixture of 0.1 mg/ml naphtol AS-MX phosphate (Sigma-Aldrich), 0.01% $MgSO_4 \cdot 7H_2O$, and 0.6 mg/ml Fast Blue BB salt (Sigma-Aldrich) in 0.1 M Tris–HCl, pH 8.5.

Oil-Red-O Staining of Adipocytes

To stain lipid droplets in mature adipocytes, cell cultures were differentiated for 12 days, as indicated. Cells were then washed twice with PBS, fixed for 30 min with 1% formaldehyde in PBS, washed once with water and then twice with 60% isopropanol. Cells were subsequently stained for 1 hour with 0.3% w/v Oil Red O (Sigma Aldrich) in 60% isopropanol, after which they were washed once with 60% isopropanol and twice with water.

Histology and Immunohistochemistry of Chondrocytes

Chondrocyte pellets were fixed in phosphate-buffered formalin and cut at 7 μm sections for histological staining with Safranin O. Sections were deparaffinized and hydrated with distilled water, followed by staining for 10 min with hematoxylin. Subsequently the sections were rinsed in distilled water for 10 min and stained with fast green solution for 5 min. Sections were then rinsed briefly in 1% acetic acid solution after which they were stained in Safranin O solution for 5 min. Finally, sections were brought to xylene solution during several dehydration steps and mounted. Sections were incubated overnight at 4°C with a specific primary antibody against Aggrecan (Millipore, Billerica, MA). As a negative control, the primary antibody was replaced with the corresponding control IgG. A biotin-streptavidin detection system was used according to the manufacturer's protocol (Vector Laboratories, Burlingame, CA). Bound complexes were visualized by reacting with 3',3'-diaminobenzidine (Sigma-Aldrich) and H_2O_2 resulting in a brown precipitate.

RESULTS

hMSCs Differentiation towards Bone, Fat, or Cartilage

In this study, we applied gene expression microarray analysis to determine the candidate genes that might mediate hMSCs lineage commitment in response to defined *in vitro* conditions. Following the experimental design outlined in Figure 1, we tested 12 different protocols, resulting in distinct time-dependent data sets for osteogenic, adipogenic, and chondrogenic differentiation.

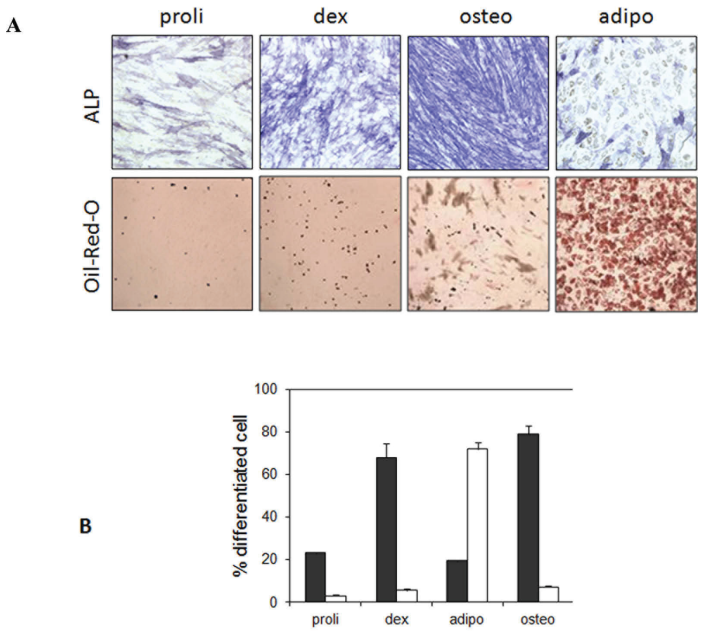
hMSCs could be induced to differentiate into osteoblasts by treating them with just 10^{-7} M dexamethasone. However, under these conditions (DX) not more than 50% of the hMSC population differentiated to osteoblasts, as revealed previously by us on the basis of histochemical staining of Alkaline Phosphatase (ALP)-positive cells (Piek, et al., 2010). Therefore, we used for our analysis two other osteogenic treatments whereby dexamethasone was combined with either BMP2 (DB) or VitD3 (DV). This strongly enhanced the percentage of ALP-positive cells to 71 and 75% respectively (Piek, et al., 2010).

Adipogenic differentiation was induced by treating hMSCs with a combination of dexamethasone, IBMX, and rosiglitazone (AD), as previously indicated (see Materials and Methods). In order to compare adipogenic and osteogenic differentiation under similar experimental conditions, we also included two osteogenic treatments (DX: only Dex; OS: Dex+BMP2) and used proliferation medium as a negative control (PR). Oil Red O-positive adipogenic cells were detected from day 9 on in adipogenic differentiation medium whereas undifferentiated controls and cells treated with osteogenic media did not stain for fat droplets at any time point or treatment (Figure 2a). However, undifferentiated control cells as well as cells treated with adipogenic media did show some ALP background staining, which was strongly enhanced upon osteogenic differentiation (Figure 2a). Using FACS analysis after 9 days of differentiation, we examined the percentage of hMSCs cells that differentiated towards bone or fat cells (Figure 2b). In agreement with the above staining data, the FACS analyses also showed some bone background. Overall our data demonstrate that both under osteogenic and adipogenic conditions 80% of the cells were positive for differentiation.

At last, the ability of hMSCs to undergo chondrogenic differentiation was tested on centrifuged pellets, cultured in chondrogenic differentiation medium. This medium contained insulin and dexamethasone (INC) in the additional presence of either TGFβ1 alone, TGFβ1 in combination with BMP2 or TGFβ1 in combination with GDF5. Over time, pellets aggregated and formed a white-

Gene Expression Regulation

Figure 2. a) Staining of hMSC cultured for 9 days in osteogenic differentiation medium (top) or 14 days in adipogenic differentiation medium (bottom). Osteoblasts were stained for ALP activity and adipocytes with oil-red-O. proli: DMEM with 10% FBS (control medium); dex: control medium with dexamethasone; osteo: osteogenic differentiation medium; adipo: adipogenic differentiation medium. b) FACS analysis of hMSCs at 9 days after treatment with the indicated media. Black bars represent % of ALP positive cells, white bars the % of nile-red positive cells.



transparent round-shaped structure at the bottom of the 10 ml tube. Chondrogenic differentiation was validated by histological staining of pellet sections for proteoglycans within the matrix using Safranin O and antibodies against aggrecan (Figure 3). Pellets cultured in chondrogenic medium containing TGFβ1 (TB) stained positively for proteoglycans, while pellets cultured for 16 days in control medium (INC) were not only smaller in size, but also did not show proteoglycan expression. Enhanced staining was observed when cells were treated with TGFβ1 + BMP2 (TB) compared to cells treated with TGFβ1 + GDF5 (TG) or TGFβ1 alone (TF). In all cases positive staining for chondrogenic differentiation was observed in treated cells.

These experiments show that expanded hMSC cultures can differentiate, in a controlled manner, into cells of multiple lineages. Based on phenotypic characterization and histological analysis it appeared that the entire culture of expanded mesenchymal cells progressed to the desired lineage, which allowed us to focus on gene expression profiling during each of the lineage-specific differentiation processes.

Differentiation of hMSCs as Revealed from Microarray Data Analysis

RNA from each lineage, treatment, and time point was isolated and used for microarray analysis. Triplicate genome-wide expression studies were performed as a function of time (11 time points) and treatment, as schematically shown in Figure 1. Data were normalized independently for each differentiation data set, in order to account

Figure 3. Chondrogenic differentiation. Cell pellets were maintained in chondrogenic differentiation medium containing insulin and dexamethasone (INC) in the additional presence of either 10 ng/ml TGFβ1 alone, TGFβ1 in combination with 50 ng/ml BMP2 or TGFβ1 in combination with 50 ng/ml GDF5. Sections of chondrogenic pellets are shown after histological staining for proteoglycans within the matrix using Safranin O and antibodies against Aggrecan.

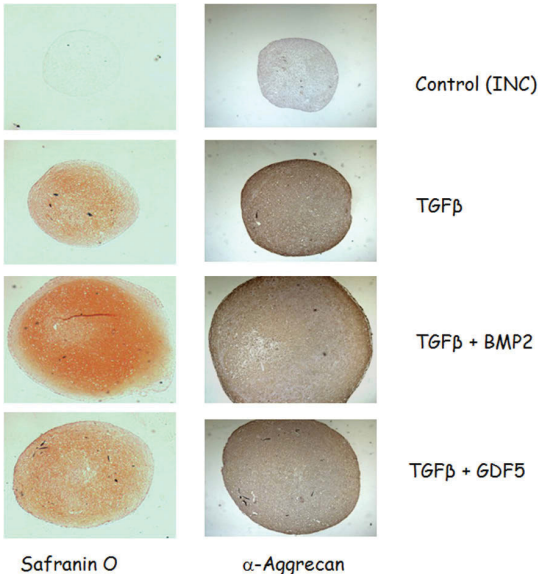
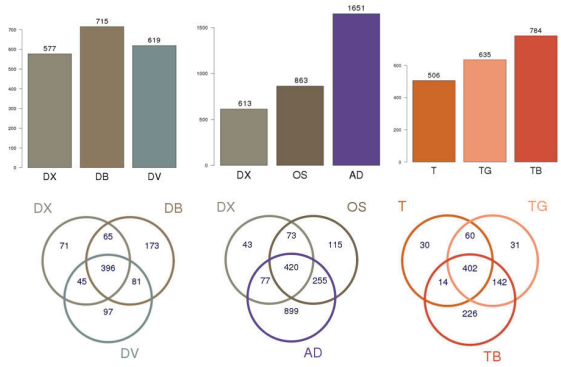


Figure 4. Representation of the number of differentially regulated genes per data set and treatment. Venn diagrams are presented to visualize overlapping genes as identified in this study.



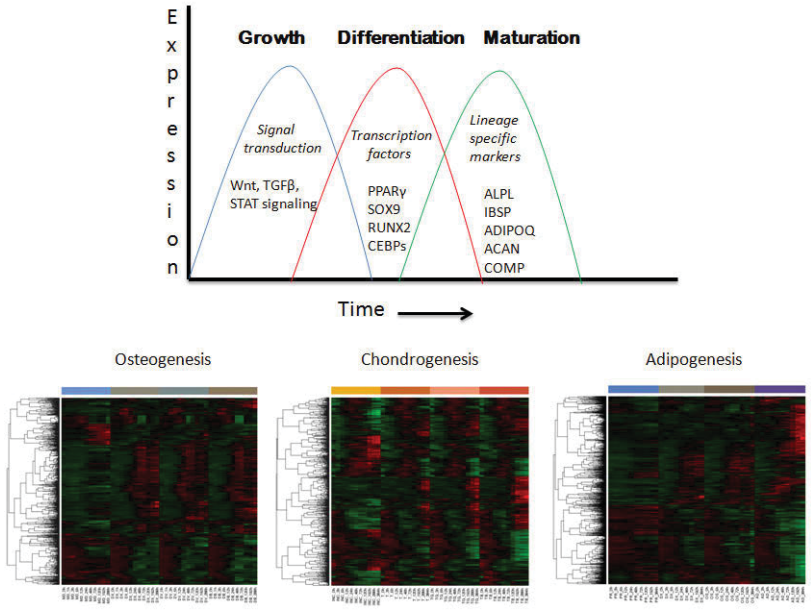
for differences in culturing methods (pellet vs. monolayer), donor, cell density, or dexamethasone concentration. Expression intensities were calculated for each time point in combination with each treatment, and subsequently expressed as base-2 logarithm of the ratios compared to time-matched untreated samples. We considered all genes with an adjusted p-value < 10^{-5} and a log-fold-change > 1 as significant. Complete lists of differentially regulated genes resulting from microarray analysis of Osteo-, Adipo-, and Chondro-Genesis sets are presented in Appendix A-I at the end of the book. Figure 4 shows the total number of differentially expressed genes, as well as Venn diagrams of the overlap between different treatments within the same lineage. The observed high degree of overlap between the treatments within a specific lineage suggests that a core pattern of altered gene expression is common in the three models.

In each lineage model, a similar number of genes was found to be regulated by Dex alone (Figure 4), in the order of 577 and 613 genes, which formed the majority (~ 70%) of all genes regulated within that lineage (396-420). These common genes showed a functional gene annotation enrichment of biological processes such as blood vessel development, vasculature development, regulation of cell proliferation, bone development, and ossification (data not shown) according to the Web-based platform DAVID Bioinformatics Resources that identify Gene Ontology (GO) terms (of the biological process category) in the set of significantly regulated genes relative to all probes represented on the array. Moreover, a similar number of genes was found significantly regulated during chondro- and osteogenesis treatments (approximately 700 genes), while a significantly higher number of regulated genes was observed upon adipocyte differentiation (1651 genes).

The growth and differentiation characteristics of hMSCs can be divided into three distinct stages based on gene expression kinetics, as schemati-

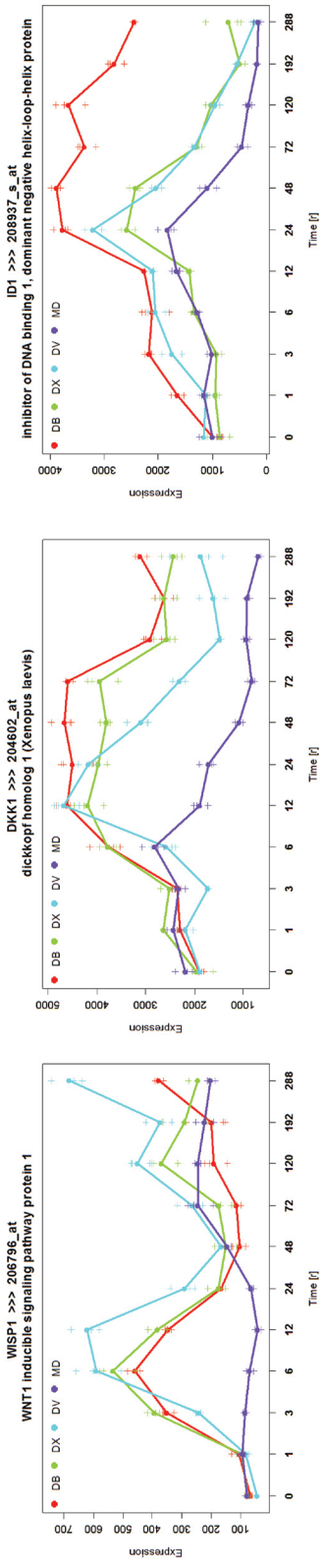
Gene Expression Regulation

Figure 5. Schematic representation of growth, differentiation, and maturation characteristics of hMSCS based on gene expression kinetics. Hierarchical clustering of differentially expressed genes is shown per lineage.

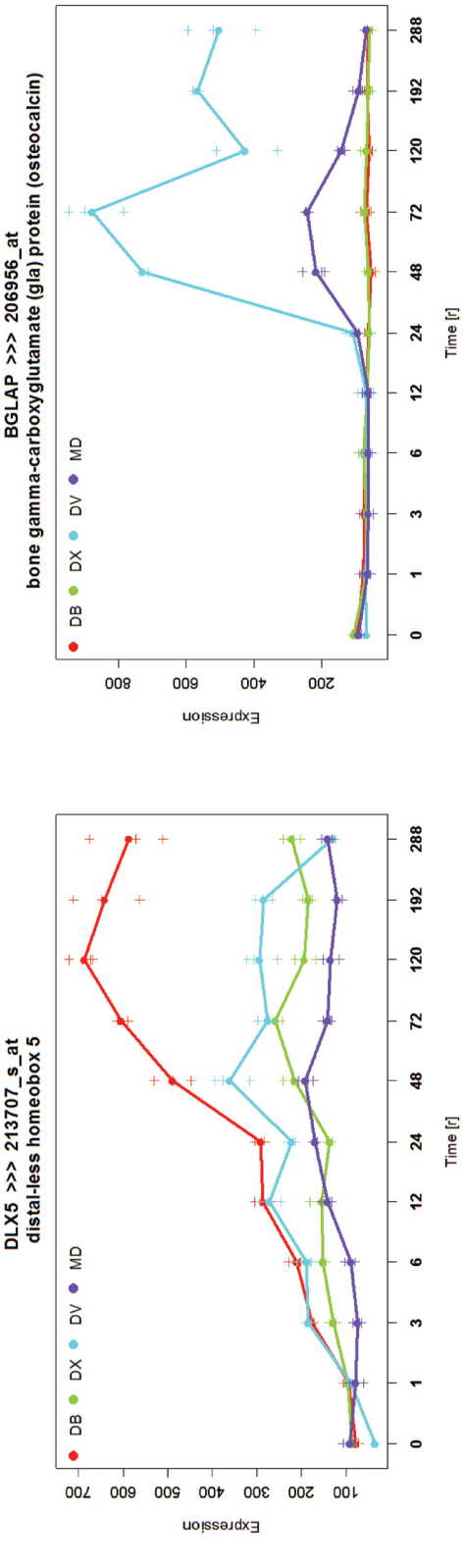


cally illustrated in Figure 5: (1) rapidly regulated genes that directly result from the signal transduction of the applied stimuli; (2) key transcription factor genes that induce lineage-specific differentiation of hMSCs; (3) genes that mark the fully differentiated phenotype, including those that encode extracellular matrix proteins in osteoblasts, fatty acid synthesizing enzymes in adipocytes or cartilage forming proteins in chondrocytes.

We extended lineage and treatment comparison of our three models by performing a hierarchical cluster analysis in an effort to further reduce the list of relevant candidate genes (Figure 5). The end product is a heatmap representation of complex gene expression data that, through statistical organization and graphical display, allowed us to examine the data. Genes were selected based on adjusted p-value $< 10^{-5}$ and a log-fold-change > 1 as significant, and each differentially expressed gene is represented by a single row of coloured boxes; each time point/replicate/treatment is represented by a single column. By using hierarchical cluster analysis we observed that genes represented by more than one array element and genes with similar temporal expression sequence are clustered next to each other. When larger groups of clustered genes were examined, we observed a significant predisposition for these genes to share common roles in biological processes. More importantly, it allowed us to identify genes and pathways, which are able to drive cells to differentiate into a specific lineage not only faster but also using divergent mechanisms. Accelerated genes expression and osteoblast differentiation was induced by Dex in combination with Vitamin D (DV) as compared to Dex alone or in combination with BMP2 (Piek, et al., 2010). In addition, our chondrogenesis data revealed accelerated chondrocyte maturation upon TB treatment compared to TGFβ alone (TF), whereas TG induced upregulation of a number of articular chondrocyte genes such as *ACAN, COL10A1* and *COL2A1* (encoding aggrecan, collagen type II, and collagen type X, respectively) at later stages than with TB.

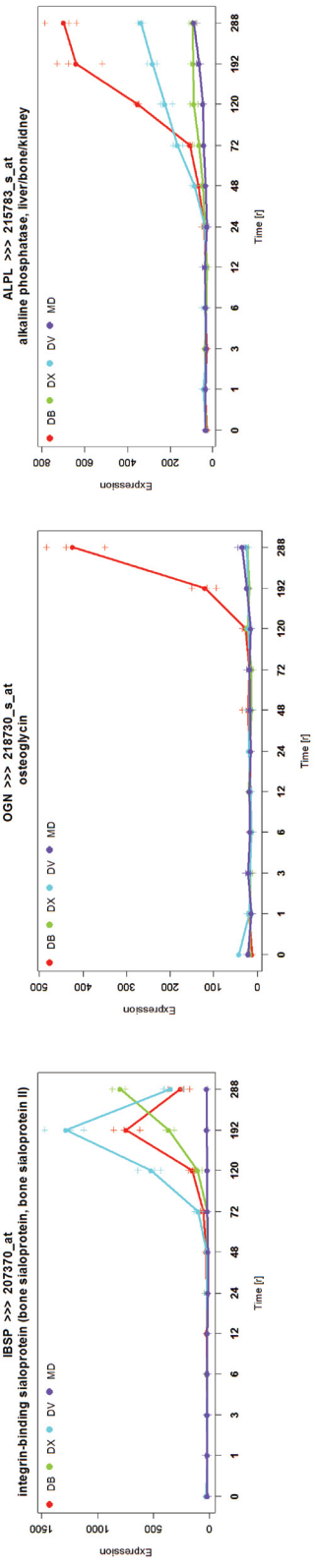


Figures 6. Examples of expression regulation of osteogenic marker genes during time and treatment (part 1)

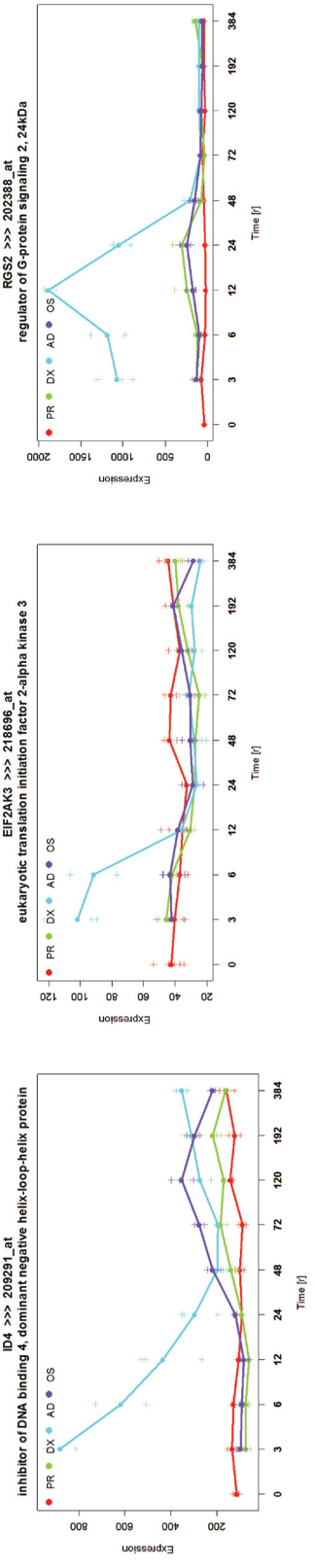


Figures 7. Examples of expression regulation of osteogenic marker genes during time and treatment (part 2)

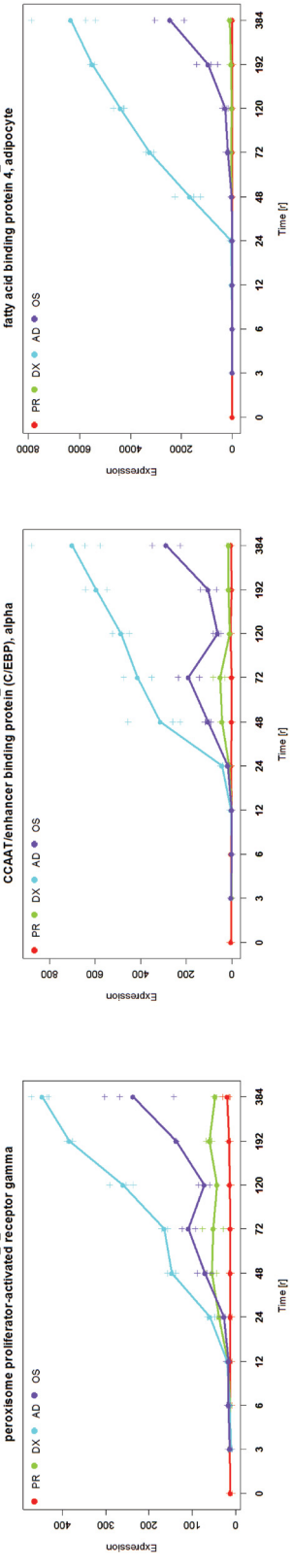
Gene Expression Regulation



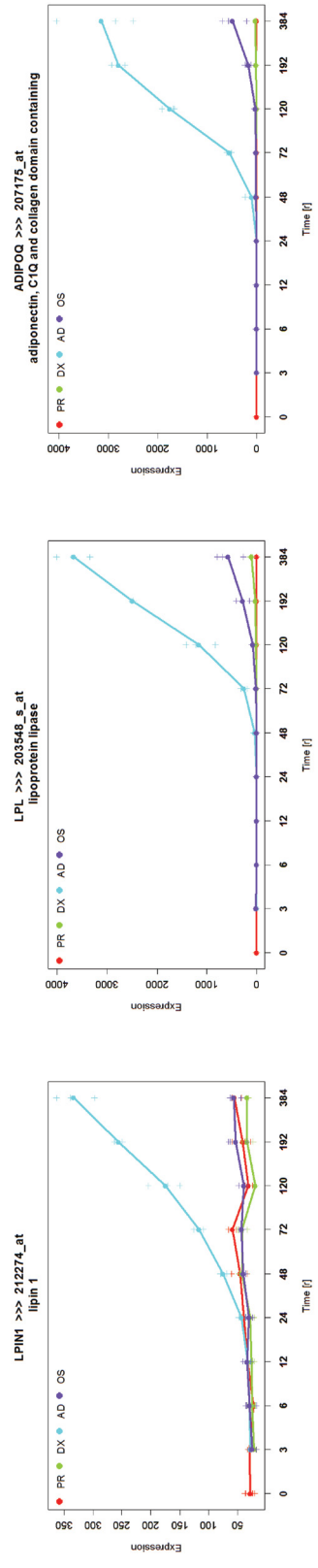
Figures 8. Examples of expression regulation of osteogenic marker genes during time and treatment (part 3)



Figures 9. Examples of expression regulation of adipogenic marker genes during time and treatment (part 1)



Figures 10. Examples of expression regulation of adipogenic marker genes during time and treatment (part 2)



Figures 11. Examples of expression regulation of adipogenic marker genes during time and treatment (part 3)

Osteogenic Differentiation

Bone homeostasis is dependent on the balance between mineral deposition by osteocytes (mature osteoblasts) and bone resorption by osteoclasts. This continuous remodelling of bone tissue is a dynamic process and is crucial for maintaining the proper bone homeostasis. Disruption of this process accompanies disorders that include osteoporosis, arthritis, and many other skeletal diseases. Therefore, we studied the functional gene annotation of the differentially expressed genes obtained from the osteogenic treatments (DX, DB, and DV). Enrichment analysis of upregulated genes revealed a functional gene annotation enrichment of biological processes such as response to hormone stimulus, skeletal system development, regulation of cell proliferation, ossification and bone development, similarly for each of the three treatments although with different p-value distribution (data not shown).

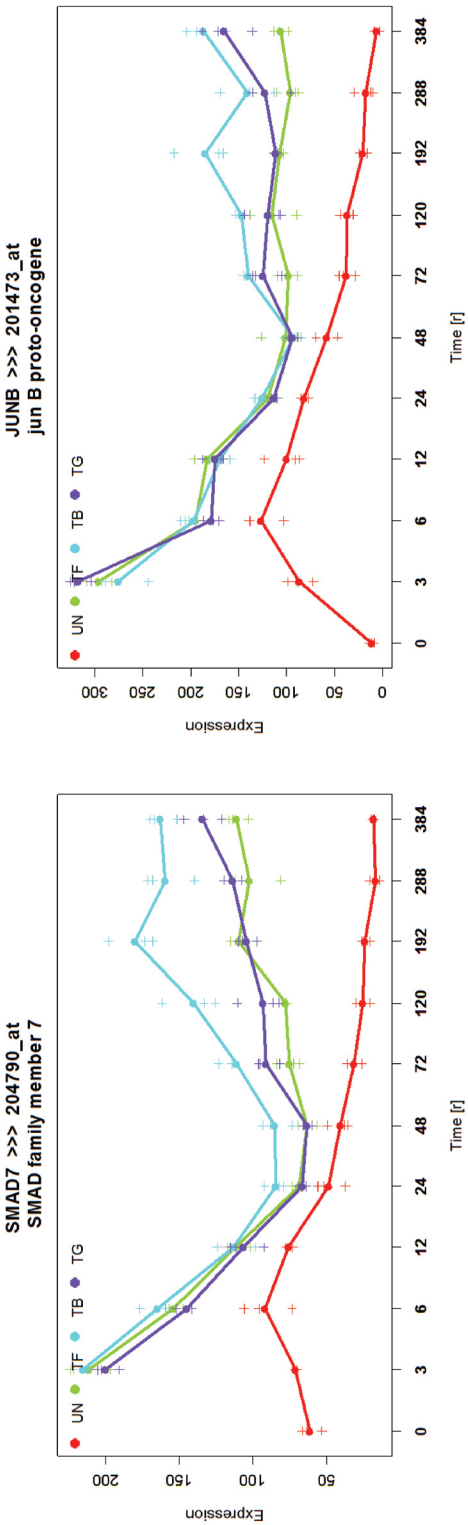
In Figures 6-8, expression profiles are presented of a selected set of marker genes generally known to be involved in bone development. As indicated previously, during the 11 days period of the experiment hMSCs cultures treated with osteogenic stimuli progressively develop into a bone tissue-like organization consisting of multilayered nodules of cells within an ordered, mineralized collagen extracellular matrix. The temporal sequence of gene expression during the osteoblast developmental program showed activation of signal transduction pathways like Wnt signalling. In particular, *WIPS1* (Wnt1 inducible signalling pathway protein 1) and *DKK1* (dickkopf homolog 1) were upregulated during the early time points. *ID1* (Helix-loop-helix protein inhibitor of DNA binding 1), a gene involved in cell proliferation and differentiation, was also regulated during early stages of osteogenesis, while upon DB treatment it showed a continuous high regulation during late time points.

DLX5 (homeodomain protein distal-less homeobox 5) is an activator of *RUNX2* (Runt-related transcription factor 2) expression, and both are thought to be key regulators in bone formation. In our data, *RUNX2* expression was not differently regulated, but its activity is known to be controlled by post-translational modification. As a result of osteoblast maturation, high expression of *ALPL*, *BGLAP*, *IBSP,* and *OGN* (encoding genes for alkaline phosphatase, osteocalcin, bone sialoprotein, and osteoglycin, respectively) was observed from 72 hours on, which is indicative for the onset of the final differentiation process. Surprisingly, the late osteogenic marker *BGLAP* was highly expressed in DV treatment from 24 hours on, but not up to 288 hours upon DB treatment.

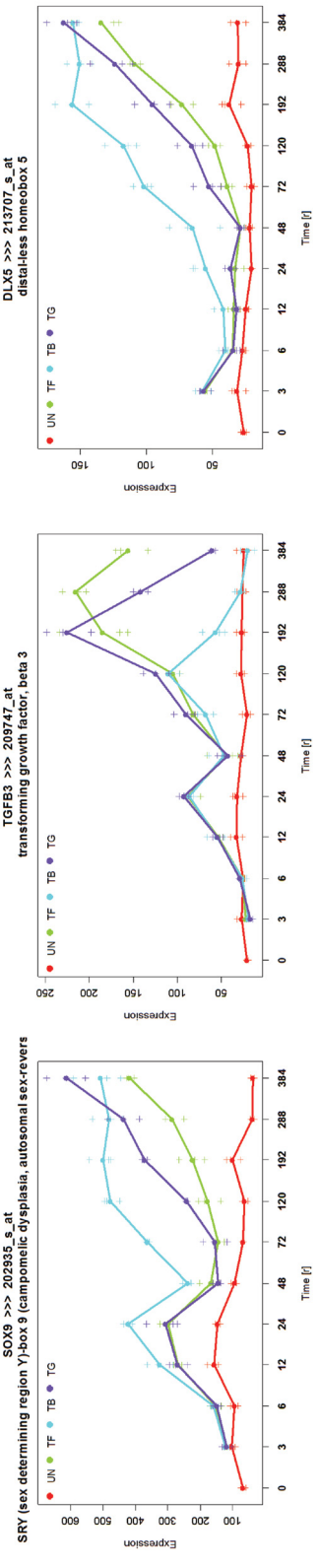
Adipogenic Differentiation

Adipocytes cells are responsible for the synthesis and storage of fat. Morphologically, adipocytes are lipid-bloated cells with displaced nuclei and a thin compartment of cytoplasm. Compared to the control the number of genes regulated during adipogenesis was significantly higher than upon osteogenic treatment. Enrichment of biological processes of upregulated genes revealed GO-terms related to oxidation reduction, generation of precursor metabolites and energy, energy derivation by oxidation of organic compounds, fatty acid metabolic and lipid biosynthetic processes.

As indicated in Figures 9-11, expression of genes such as *ID4* and *RGS2* (encoding helix-loop-helix protein inhibitor of DNA binding 1 and regulator of G-protein signalling, respectively) was highly modulated during early commitment towards adipogenesis. At the same time, early expression of *EIF2AK3* (encoding eukaryotic translation initiation factor-2alpha kinase 3) was decreased specifically during adipogenic conditions, suggesting a role in adipocyte differentiation. Expression of the differentiation controlling transcriptional regulators *PPARG* and *CEBPA* (peroxisome proliferator-activated receptor gamma and CCAAT/enhancer binding protein alpha, respectively) was observed from 24 hours on. Their downstream target genes, including *LPL*,

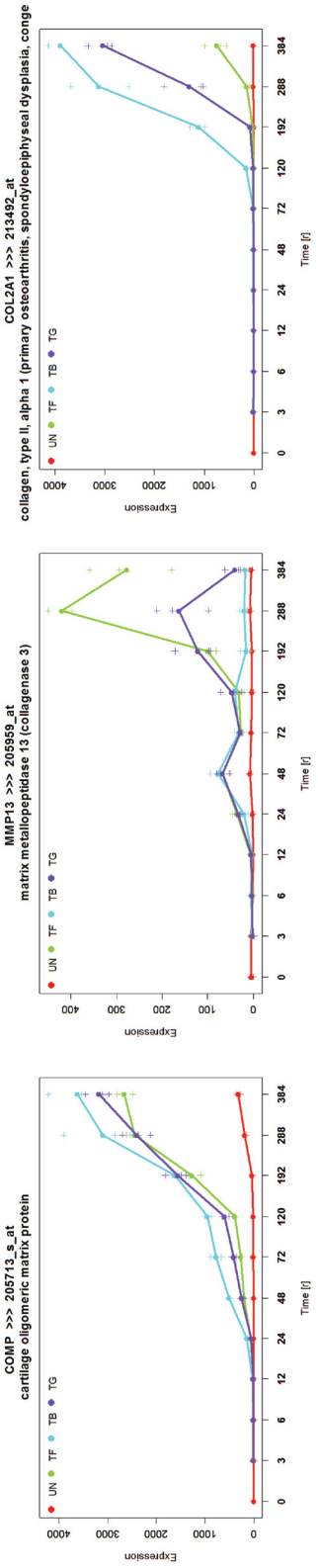


Figures 12. Examples of expression regulation of chondrogenic marker genes during time and treatment (part 1)

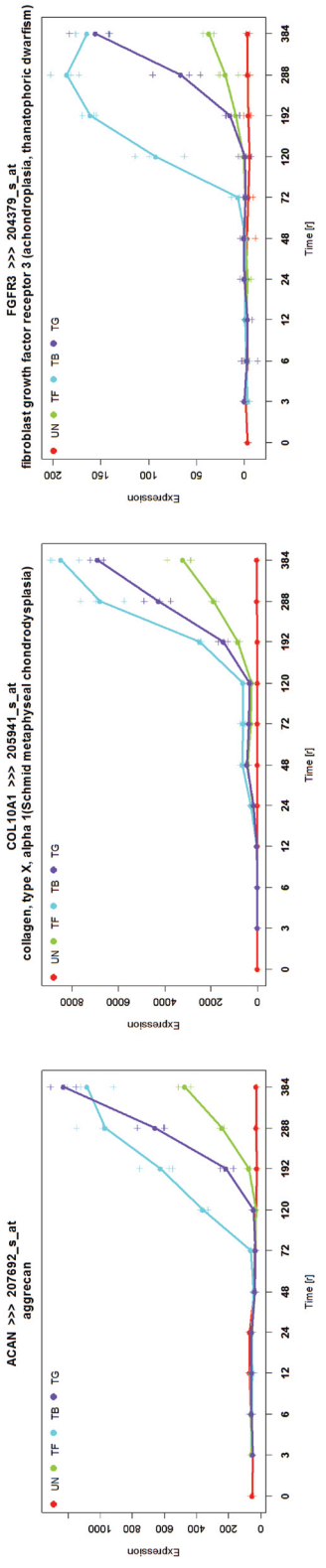


Figures 13. Examples of expression regulation of chondrogenic marker genes during time and treatment (part 2)

Gene Expression Regulation



Figures 14. Examples of expression regulation of chondrogenic marker genes during time and treatment (part 3)



Figures 15. Examples of expression regulation of chondrogenic marker genes during time and treatment (part 4)

ADIPOQ or *FABP4* (encoding lipoprotein lipase, adiponectin, or fatty acid binding protein 4, respectively) were highly expressed after at least 24 hours and specifically under adipogenic conditions.

Chondrogenic Differentiation

Chondrocytes, which differentiate following the condensation of mesenchymal stem cells, are responsible for the secretion of extracellular matrix proteins, such as collagens and proteoglycans that form the cartilage anlagen. This eventually forms the cartilage of the articular joint or undergoes hypertrophy and endochondral ossification to form bone. Enrichment of biological processes of upregulated genes during TF, TB, and TG treatments revealed enrichment of GO-terms related to skeletal system development, cartilage development, collagen fibril organization, extracellular matrix organization, regulation of cell proliferation, response to hypoxia, ossification and bone development. Overall, the three treatments showed positive chondrogenic differentiation based on gene expression. Moreover, as we previously showed in Figure 3, chondrogenic pellets grown in TGFβ alone or with GDF5 exhibited a smaller increase in size and matrix production than pellets that were treated in addition with BMP2.

Figures 12-15 shows that TGFβ pathway is activated at early time points based on the expression of downstream factors such as *SMAD7* (encoding smad family member 7) and *JUNB* (encoding JunB proto-oncogen). The transcription factor *SOX9* (encoding sex determining region-y box 9), which plays a critical role in chondrocyte differentiation and function, is highly regulated from the early time points on. Moreover, *SOX9* expression showed stronger upregulation upon treatment with TB than with TG or TF. Expression of cartilage-specific genes such as *COL2A1*, *COL10A1* and *ACAN* was induced at later time points, from 72 hours on during TB treatment compared to 120 hours with TG.

DISCUSSION

Mesenchymal stem cells in the bone marrow have the capacity to differentiate towards osteoblasts (bone), adipocytes (fat), and chondrocytes (cartilage) (Pittenger, et al., 1999). The developmental switch from self-renewal to differentiation is controlled by transcription factors that downregulate stemness genes and upregulate lineage-specific differentiation genes. Activation of the Glucocorticoid Receptor (GR) is essential for loss of self-renewal capacity (Wessely, Deiner, Beug, & von Lindern, 1997). Therefore, we used dexamethasone as a potent glucocorticoid activator, and once cells were out of the "stem cells compartment" they were subsequently induced to differentiate into the three distinct lineages by using a different but comprehensive selection of treatments.

For that reason, the aim of this study was to understand normal lineage specification of human mesenchymal stem cells by using a system biology approach from time-course microarray data following *in vitro* differentiation. By comparing patterns of gene expression in three independent models of hMSCs differentiation, we have been able to identify candidate genes that might mediate the core lineage commitment features of hMSCs.

We have demonstrated on the basis of phenotypic characterization, histological analysis and gene expression profiling that hMSCs have the ability to differentiate into bone under osteogenic conditions (β-glycerophosphate and ascorbic acid) either in the presence of dexamethasone alone or in combination with either BMP2 or Vitamin D. Our observations also indicated that Vitamin D speeds up the kinetics of gene expression suggesting that the osteogenic program is initiated faster than in Dex or Dex+BMP2-treated cultures. This effect has recently been described by Piek et al. (2010).

Additionally, we studied the hMSCs differentiation towards fat. Adipogenesis, a highly regulated process, was coordinated by a cascade of transcription factors and leads to the forma-

Gene Expression Regulation

tion of mature adipocytes (Farmer, 2006). Early commitment was marked by the expression of *ID4*, a candidate molecular switch in osteoblast and adipocyte differentiation. Downregulation of *ID4* drastically reduced osteoblast differentiation and enhanced differentiation toward adipocytes by increasing *PPARG* and *CEBPA* expression (Tokuzawa, et al., 2010). This cascade begins with transient expression of *CEBPB* and *CEBPD* which in turn activate *CEBPA* and *PPARG*. Their protein products act together to coordinate the expression of adipogenic genes underlying the phenotype of terminally differentiated adipocytes (Darlington, Ross, & MacDougald, 1998; Kang, et al., 2008). This terminal differentiation process is characterized by the induction of genes encoding fatty acid binding proteins (FABPs), lipid phosphatases such as *LPIN* (Lipin 1), and those encoding adipokines such as *LEP* (leptin) and *ADIPOQ* (adiponectin) (Ntambi & Young-Cheul, 2000; Ponce, et al., 2008). In addition, we were able to identify genes that might regulate the commitment of hMSCs into pre-adipocyte or pre-osteoblasts based on their unique common regulation. Additional studies will explore the precise role of these identified genes in regulating the possible switch between adipogenesis and osteoblastogenesis (Gimble, Zvonic, Floyd, Kassem, & Nuttall, 2006), since at present there are no reliable cell markers to identify intermediate cells.

The TGFβ superfamily members are pleiotropic factors, which may have different effects when they are applied alone or in combination with other growth factors. Formation of cartilage essentially requires the presence of TGFβ, alone or in combination with BMP5 or GDF5. The kinetics of gene expression during chondrogenesis suggest that this differentiation program is initiated faster in chondrogenic pellets treated with TGFβ in combination with BMP2 than with TGFβ alone or TGFβ in combination with GDF5. During mesenchymal condensation and limb cartilage development, TGFβ and BMP act distinctly in terms of their chondrogenic potential, such that TGFβ has an early positive role in promoting chondrogenesis, while BMP has an early negative and late positive role during chondrogenesis (Hatakeyama, Tuan, & Shum, 2004). GDF5 is a divergent member of the TGFβ/BMP superfamily that is required for proper skeletal patterning and joint development in the vertebrate limb (Heidaran, et al., 2000). As shown in Figures 12-15, TB caused a higher expression of *SOX9* over time, but TG caused a more sustained elevated expression level at late stage. Expression of cartilage-specific genes was accelerated by TB treatment, such that *COL2A1*, *COL10A1,* or *ACAN* appeared regulated from 72 hours on, whereas during TG treatment this was observed only at later time points. This might be due to the early expression of *FGFR3* (fibroblast growth factor receptor 3) during TB, since FGFR3 is known to inhibit chondrocyte proliferation (L'Hôte & Knowles, 2005; Schibler, et al., 2009). In conclusion, these finding are suggestive of reduced proliferation, premature cell cycle exit, and induction of earlier differentiation.

The family of TGFβ can be classified into two groups: the first group includes the Bone Morphogenetic Proteins (BMP) and the Growth Differentiation Factors (GDF) and the second one comprises TGFβ, activin, and nodal (Pucéat, 2007). Members of both groups have been found to play a crucial role during hMSCs self-renewal and differentiation (Jian, et al., 2006; Zhou, 2011). In agreement with our findings, TGFβ signaling has been reported to be important in hMSC differentiation into the osteogenic and chondrogenic lineages, whereby *in vitro* studies suggest a high degree of plasticity between the adipocytic and osteoblastic pathways. In addition, we have observed that TGFβ promotes osteogenesis while blocking adipogenesis in human mesenchymal stem cells (data not shown), in agreement with previous data (Choy & Derynck, 2003; Ponce, et al., 2008; Zamani & Brown, 2011).

In conclusion, using microarray analysis we have identified a panel of genes whose expression across multiple datasets and treatments suggests that they may be associated with lineage-specific differentiation of hMSCs. Some of these genes play

previously well-documented roles in bone, cartilage, and fat biology. Moreover, this study identifies genes for further research on their functional role during hMSCs differentiation and diseases, providing a comprehensive and important resource for future studies on regenerative medicine of degenerative diseases in mesenchymal tissues.

ACKNOWLEDGMENT

Thanks to Linconet (ERASysBio+ initiative) and N.V. Organon for financial support. Thanks to Ester Piek, Laura Sleumer, Bas van der Woning, Sander Caerteling, Jose Roelofs-Hendriks, Ingrid de Grijs, Eugene van Someren, Peter van der Kraan, and Koen Dechering for their contribution in obtaining the described data.

REFERENCES

Bouchez, L. C., Boitano, A. E., de Lichtervelde, L., Romeo, R., Cooke, M. P., & Schultz, P. G. (2011). Small-molecule regulators of human stem cell self-renewal. *ChemBioChem*, *12*(6), 854–857. doi:10.1002/cbic.201000734

Chamberlain, G., Fox, J., Ashton, B., & Middleton, J. (2007). Concise review: Mesenchymal stem cells: Their phenotype, differentiation capacity, immunological features, and potential for homing. *Stem Cells (Dayton, Ohio)*, *25*(11), 2739–2749. doi:10.1634/stemcells.2007-0197

Choy, L., & Derynck, R. (2003). Transforming growth factor-Î² inhibits adipocyte differentiation by Smad3 interacting with CCAAT/enhancer-binding protein (C/EBP) and repressing C/EBP transactivation function. *The Journal of Biological Chemistry*, *278*(11), 9609–9619. doi:10.1074/jbc.M212259200

Darlington, G. J., Ross, S. E., & MacDougald, O. A. (1998). The role of C/EBP genes in adipocyte differentiation. *The Journal of Biological Chemistry*, *273*, 30057–30060. doi:10.1074/jbc.273.46.30057

De Bari, C., Dell'Accio, F., Vandenabeele, F., Vermeesch, J. R., Raymackers, J.-M., & Luyten, F. P. (2003). Skeletal muscle repair by adult human mesenchymal stem cells from synovial membrane. *The Journal of Cell Biology*, *160*(6), 909–918. doi:10.1083/jcb.200212064

De Bari, C., Dell'Accio, F., Vanlauwe, J., Eyckmans, J., Khan, I. M., & Archer, C. W. (2006). Mesenchymal multipotency of adult human periosteal cells demonstrated by single-cell lineage analysis. *Arthritis and Rheumatism*, *54*, 1209–1221. doi:10.1002/art.21753

Dezawa, M., Ishikawa, H., Itokazu, Y., Yoshihara, T., Hoshino, M., & Takeda, S.-I. (2005). Bone marrow stromal cells generate muscle cells and repair muscle degeneration. *Science*, *309*(5732), 314–317. doi:10.1126/science.1110364

Farmer, S. R. (2006). Transcriptional control of adipocyte formation. *Cell Metabolism*, *4*, 263–273. doi:10.1016/j.cmet.2006.07.001

Gimble, J. M., Zvonic, S., Floyd, Z. E., Kassem, M., & Nuttall, M. E. (2006). Playing with bone and fat. *Journal of Cellular Biochemistry*, *98*, 251–266. doi:10.1002/jcb.20777

Giordano, A., Galderisi, U., & Marino, I. R. (2007). From the laboratory bench to the patient's bedside: An update on clinical trials with mesenchymal stem cells. *Journal of Cellular Physiology*, *211*(1), 27–35. doi:10.1002/jcp.20959

Hatakeyama, Y., Tuan, R. S., & Shum, L. (2004). Distinct functions of BMP4 and GDF5 in the regulation of chondrogenesis. *The Journal of Biological Chemistry*, *91*, 1204–1217.

Gene Expression Regulation

Heidaran, M. A., Daverman, R., Thompson, A., Ng, C. K., Pohl, J., & Poser, J. W. (2000). Extracellular matrix modulation of rhGDF-5-induced cellular differentiation. *Journal of Regenerative Medicine*, *1*(9), 121–135. doi:10.1089/152489000420294

Horwitz, E. M., Prockop, D. J., Gordon, P. L., Koo, W. W. K., Fitzpatrick, L. A., & Neel, M. D. (2001). Clinical responses to bone marrow transplantation in children with severe osteogenesis imperfecta. *Blood*, *97*(5), 1227–1231. doi:10.1182/blood.V97.5.1227

Javazon, E. H., Beggs, K. J., & Flake, A. W. (2004). Mesenchymal stem cells: Paradoxes of passaging. *Experimental Hematology*, *32*(5), 414–425. doi:10.1016/j.exphem.2004.02.004

Jian, H., Shen, X., Liu, I., Semenov, M., He, X., & Wang, X.-F. (2006). Smad3-dependent nuclear translocation of Î²-catenin is required for TGF-Î²1-induced proliferation of bone marrow-derived adult human mesenchymal stem cells. *Genes & Development*, *20*(6), 666–674. doi:10.1101/gad.1388806

Kang, Q., Song, W.-X., Luo, Q., Tang, N., Luo, J., & Luo, X. (2008). A comprehensive analysis of the dual roles of BMPs in regulating adipogenic and osteogenic differentiation of mesenchymal progenitor cells. *Stem Cells and Development*, *18*(4), 545–558. doi:10.1089/scd.2008.0130

Körbling, M., & Estrov, Z. (2003). Adult stem cells for tissue repair: A new therapeutic concept? *The New England Journal of Medicine*, *349*(6), 570–582. doi:10.1056/NEJMra022361

L'Hôte, C. G. M., & Knowles, M. A. (2005). Cell responses to FGFR3 signalling: Growth, differentiation and apoptosis. *Experimental Cell Research*, *304*(2), 417–431. doi:10.1016/j.yexcr.2004.11.012

Ntambi, J. M., & Young-Cheul, K. (2000). Adipocyte differentiation and gene expression. *The Journal of Nutrition*, *130*(12), 3122–3126.

Nuttall, M. E., & Gimble, J. M. (2004). Controlling the balance between osteoblastogenesis and adipogenesis and the consequent therapeutic implications. *Current Opinion in Pharmacology*, *4*(3), 290–294. doi:10.1016/j.coph.2004.03.002

Piek, E., Sleumer, L. S., van Someren, E. P., Heuver, L., de Haan, J. R., & de Grijs, I. (2010). Osteo-transcriptomics of human mesenchymal stem cells: Accelerated gene expression and osteoblast differentiation induced by vitamin D reveals c-MYC as an enhancer of BMP2-induced osteogenesis. *Bone*, *46*(3), 613–627. doi:10.1016/j.bone.2009.10.024

Pittenger, M. F., Mackay, A. M., Beck, S. C., Jaiswal, R. K., Douglas, R., & Mosca, J. D. (1999). Multilineage potential of adult human mesenchymal stem cells. *Science*, *284*(5411), 143–147. doi:10.1126/science.284.5411.143

Ponce, M. L., Koelling, S., Kluever, A., Heinemann, D. E. H., Miosge, N., & Wulf, G. (2008). Coexpression of osteogenic and adipogenic differentiation markers in selected subpopulations of primary human mesenchymal progenitor cells. *The Journal of Biological Chemistry*, *104*(4), 1342–1355.

Prockop, D. J., & Olson, S. D. (2007). Clinical trials with adult stem/progenitor cells for tissue repair: Let's not overlook some essential precautions. *Blood*, *109*(8), 3147–3151. doi:10.1182/blood-2006-03-013433

Pucéat, M. (2007). TGFb in the differentiation of embryonic stem cells. *Cardiovascular Research*, *74*(2), 256–261. doi:10.1016/j.cardiores.2006.12.012

Schibler, L., Gibbs, L., Benoist-Lasselin, C., Decraene, C., Martinovic, J., & Loget, P. (2009). New insight on FGFR3-related chondrodysplasias molecular physiopathology revealed by human chondrocyte gene expression profiling. *PLoS ONE*, *4*(10), e7633. doi:10.1371/journal.pone.0007633

Shi, Y., Su, J., Roberts, A. I., Shou, P., Rabson, A. B., & Ren, G. (2012). How mesenchymal stem cells interact with tissue immune responses. *Trends in Immunology*, *33*(3), 136–143. doi:10.1016/j.it.2011.11.004

Tokuzawa, Y., Yagi, K., Yamashita, Y., Nakachi, Y., Nikaido, I., & Bono, H. (2010). Id4, a new candidate gene for senile osteoporosis, acts as a molecular switch promoting osteoblast differentiation. *PLOS Genetics*, *6*(7), e1001019. doi:10.1371/journal.pgen.1001019

Vaes, B. L. T., Ducy, P., Sijbers, A. M., Hendriks, J. M. A., van Someren, E. P., & de Jong, N. G. (2006). Microarray analysis on Runx2-deficient mouse embryos reveals novel Runx2 functions and target genes during intramembranous and endochondral bone formation. *Bone*, *39*(4), 724–738. doi:10.1016/j.bone.2006.04.024

Wessely, O., Deiner, E.-M., Beug, H., & von Lindern, M. (1997). The glucocorticoid receptor is a key regulator of the decision between self-renewal and differentiation in erythroid progenitors. *The EMBO Journal*, *16*(2), 267–280. doi:10.1093/emboj/16.2.267

Zamani, N., & Brown, C. W. (2011). Emerging roles for the transforming growth factor-Î² superfamily in regulating adiposity and energy expenditure. *Endocrine Reviews*, *32*(3), 387–403. doi:10.1210/er.2010-0018

Zhou, S. (2011). TGF-β regulates β-catenin signaling and osteoblast differentiation in human mesenchymal stem cells. *The Journal of Biological Chemistry*, *112*(6), 1651–1660.

Zuk, P. A., Zhu, M., Ashjian, P., De Ugarte, D. A., Huang, J. I., & Mizuno, H. (2002). Human adipose tissue is a source of multipotent stem cells. *Molecular Biology of the Cell*, *13*(12), 4279–4295. doi:10.1091/mbc.E02-02-0105

KEY TERMS AND DEFINITIONS

Adipogenesis: Is the process of cell differentiation by which preadipocytes become adipocytes. Adipocytes, also known as lipocytes and fat cells, are the cells that primarily compose adipose tissue, specialized in storing energy as fat.

Chondrogenesis: Is the process by which cartilage is developed. Cartilage is composed of specialized cells called chondrocytes that produce a large amount of extracellular matrix.

Gene Ontology (GO): Gene Ontology is a controlled method for describing terms related to genes in any organism. As more gene data is obtained from organisms, it is annotated using Gene Ontology. Gene Ontology is made of three smaller ontologies or aspects: Molecular Function, Biological Process, and Cellular Component. Each of these ontologies contains terms that are organized in a directed acyclic graph with these three terms as the roots. The roots are the broadest terms relating to genes.

Human Mesenchymal Stem Cells (hMSC): Human mesenchymal stem cells are thought to be multipotent cells, which are present in adult marrow, that can replicate as undifferentiated cells and that have the potential to differentiate to lineages of mesenchymal tissues, including bone, cartilage, fat, tendon, muscle, and marrow stroma.

Osteogenesis: Or ossification is the process of laying down new bone material by cells called osteoblasts. It is synonymous with bone tissue formation.

Transcriptomics: Also referred to as expression profiling, often using high-throughput techniques based on DNA microarray technology available for humans and other species.

3.3 Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0

Weber *et al. BMC Systems Biology* 2013, **7**:1
http://www.biomedcentral.com/1752-0509/7/1

BMC Systems Biology

METHODOLOGY ARTICLE Open Access

Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0

Michael Weber[1], Sebastian G Henkel[2*], Sebastian Vlaic[1], Reinhard Guthke[1], Everardus J van Zoelen[3] and Dominik Driesch[2]

Weber *et al. BMC Systems Biology* 2013, **7**:1
http://www.biomedcentral.com/1752-0509/7/1

BMC Systems Biology

METHODOLOGY ARTICLE Open Access

Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0

Michael Weber[1], Sebastian G Henkel[2*], Sebastian Vlaic[1], Reinhard Guthke[1], Everardus J van Zoelen[3] and Dominik Driesch[2]

Abstract

Background: Inference of gene-regulatory networks (GRNs) is important for understanding behaviour and potential treatment of biological systems. Knowledge about GRNs gained from transcriptome analysis can be increased by multiple experiments and/or multiple stimuli. Since GRNs are complex and dynamical, appropriate methods and algorithms are needed for constructing models describing these dynamics. Algorithms based on heuristic approaches reduce the effort in parameter identification and computation time.

Results: The NetGenerator V2.0 algorithm, a heuristic for network inference, is proposed and described. It automatically generates a system of differential equations modelling structure and dynamics of the network based on time-resolved gene expression data. In contrast to a previous version, the inference considers multi-stimuli multi-experiment data and contains different methods for integrating prior knowledge. The resulting significant changes in the algorithmic procedures are explained in detail. NetGenerator is applied to relevant benchmark examples evaluating the inference for data from experiments with different stimuli. Also, the underlying GRN of chondrogenic differentiation, a real-world multi-stimulus problem, is inferred and analysed.

Conclusions: NetGenerator is able to determine the structure and parameters of GRNs and their dynamics. The new features of the algorithm extend the range of possible experimental set-ups, results and biological interpretations. Based upon benchmarks, the algorithm provides good results in terms of specificity, sensitivity, efficiency and model fit.

Keywords: Gene-regulatory networks, Network inference, Heuristic algorithm, ODE, NetGenerator

Background

For the adaptation of biological systems towards external and environmental stimuli usually a complex interaction network of intracellular biochemical components is triggered. That includes changes in the gene expression at both the mRNA and protein level. Considering a certain stimulus as an input signal to the system and mRNA or protein levels as outputs, the connecting network may include interactions between signal transduction intermediates: transcription factors and target genes. Generally, the term "gene-regulatory network" (GRN) summarises genetic dependencies, which describe the influence of gene expression by transcriptional regulation, [1].

The inference (elucidation) of GRNs is important for understanding intracellular processes and for potential manipulation of the system either by specific gene mutations, knock-downs or by treatment of the cells with drugs, e.g. for medical purposes. Towards a full understanding in terms of a complete network, partial models of the network give intermediate results which help to refine the knowledge and to design new experiments. Still, many gene-regulated cellular functions, e.g. stem cell differentiation, depend on more than one stimulus and the cross-talk within the GRN, e.g. [2]. On the other

*Correspondence: sebastian.henkel@biocontrol-jena.com
[2]BioControl Jena GmbH, Wildenbruchstr. 15, 07745 Jena, Germany, www.biocontrol-jena.com
Full list of author information is available at the end of the article

hand, the stimuli might influence distinct components of a GRN. Such biologically relevant dependencies can be investigated by applying two or more stimuli and measuring the influenced genes. This approach can be called multi-stimuli experiment. If this is carried out in two or more separate experiments, one derives multi-stimuli multi-experiment data. Only algorithms with the ability to consider those data can infer such dependencies.

As shown in review articles, e.g. [1,3,4], there are different inference methods using various sources of information thus leading to different results. Amongst the typically mathematical models the application of differential equations describing time-resolved gene expression data ("time series") has been proven successful. Unfortunately the potential complexity of the networks leads to a high number of structural connections and parameters in contrast to the comparably small number of available measurement data. Apart from the problem of identifiability, the number of possible parameter combinations is very large, thus resulting in high computational costs. Therefore, appropriate heuristic approaches can reduce this amount while providing comparably good inference results. NetGenerator is a heuristic algorithm, which considers time series data to automatically infer GRNs influenced by an external stimulus, [5] and [6]. The approach combines a structure (network topology) and parameter optimisation. The final result in form of a differential equations model can be simulated and displayed graphically. An earlier version with less functionality was applied successfully to biological problems, e.g. the regulatory network of iron acquisition in *Candida albicans* and the analysis of the *Aspergillus fumigatus* infection process, [7] and [8].

In the present article, we propose NetGenerator V2.0, an extended version of the algorithm which enables the use of multi-stimuli multi-experiment data, thus increasing the number of addressable biological questions. This causes significant changes in the algorithmic procedures, especially the processing of this kind of data as well as the structure and parameter optimisation. Also, some other updated features will be outlined, for example the different modes of prior knowledge integration, further knowledge-based procedures, options of graphical outputs, changed non-linear modelling and re-implementation in the programming language / statistical computing environment R, [9]. Further, in comparison to the previous version, some of the algorithmic procedures will be explained in more detail, because they are important for understanding the overall method.

The successful application of the novel NetGenerator will be shown by inference of relevant multi-stimuli multi-experiment benchmark examples, namely systems with a different degree of cross-talk. Two aspects will be assessed: (i) reproduction of the benchmark systems (data and structure) and (ii) refinement / extension of a network structure by combination of different experimental data. Furthermore, the applicability of NetGenerator to a real-world problem is presented: after describing necessary data pre-processing steps, the underlying GRN of chondrogenic differentiation of human mesenchymal stem cells, a process influenced by the two stimuli TGF-beta1 and BMP2, is inferred.

Methods

In the following subsections the necessary background knowledge and methodology of the NetGenerator algorithm is described. In comparison to previous publications this includes new, updated and more detailed algorithmic procedures. First, the motivation and the goals are defined by considering the biological data. Necessary steps of data pre-processing are also explained within this subsection. Subsequently, ordinary differential equations and some of their properties are presented as a means for modelling the dynamics of gene regulatory networks. Then the heuristic approach of the algorithm is explained including the structure and parameter identification (here: optimisation-based determination). The next important topic will be the consideration of prior knowledge, followed by a subsection about the numerical simulation as well as the representation of modelling and graphical results. Finally, some important options and their influence to the algorithm are presented.

Time series data and pre-processing

Gene expression time series data as required by NetGenerator are typically derived from microarray measurements. Before starting the network inference, raw microarray data have to be processed comprising a series of steps. The three main steps are displayed in Figure 1: (i) microarray pre-processing, (ii) gene selection and (iii) time series scaling.

Microarray pre-processing applies multiple procedures to remove non-biological noise from the data and to estimate gene expression levels. Custom probe-sets, as assembled by [10], reduce the number of cross-hybridising probes. This initial reduction accomplishes a one-to-one correspondence between probe-set and gene. Background correction, normalisation and summarisation are provided by the RMA package, [11], resulting in logarithmised gene expression estimates, which can be used for the next processing step.

Gene selection ("filtering") is the important second step of processing, since reliable network inference is only feasible for a sufficient number of measurements per gene [1]. This number is often limited and therefore a selection of genes for modelling is inevitable. Candidate genes should show pronounced temporal dynamics and significant differences compared to the control

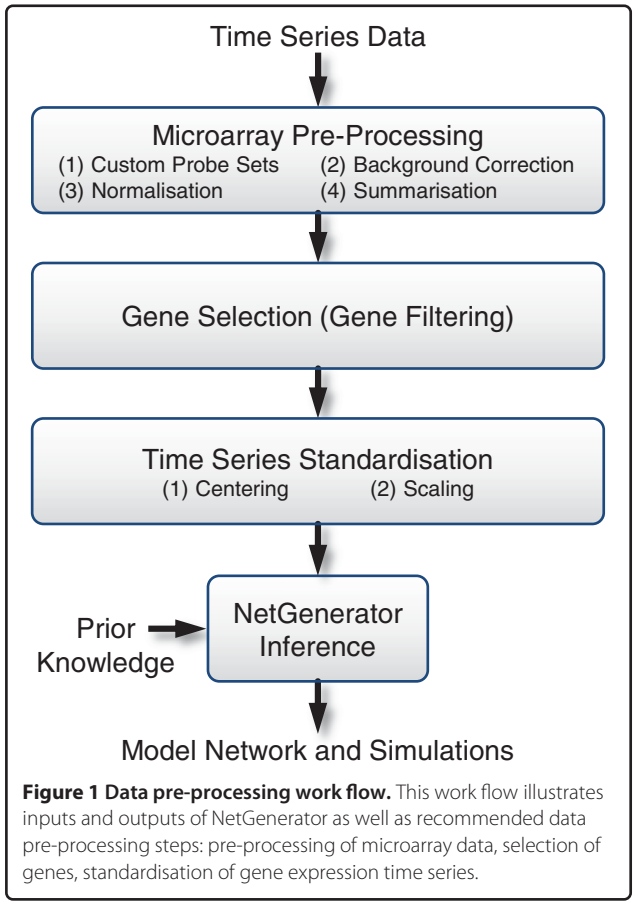


Figure 1 Data pre-processing work flow. This work flow illustrates inputs and outputs of NetGenerator as well as recommended data pre-processing steps: pre-processing of microarray data, selection of genes, standardisation of gene expression time series.

group. Statistical methods for identification of differentially expressed genes are widely used for gene selection. We use the LIMMA tool, which can operate on time series data determining significance of gene expression changes over time [12]. The statistical test (moderated t-statistics) operates on contrast terms, defined by subtracting the control group at each time point. LIMMA returns a ranked table for all genes containing columns for gene name, fold-change and adjusted p-values. Differentially expressed genes are selected by a combination of adjusted p-value cut-off and fold-change criterion.

Time series standardisation is the last processing step including centering and scaling of each time series. The centering procedure subtracts the original initial value at the starting time point from all values such that the transformed time series starts from zero. In the subsequent scaling procedure each time series is divided by its maximum (absolute value) across all provided experimental data. This leads to gene-wise scaled data and gene expression time series varying within -1 and 1. The resulting data provided to the NetGenerator algorithm are stored in $\underline{\underline{X}}_e$ and $\underline{\underline{U}}_e$, i.e. matrices for the time series (output) and stimuli (input) data, respectively, for all experiments $e = 1, \ldots, E$. Therefore, the dimensions are $\underline{\underline{X}}_e : T_e \times N$ and $\underline{\underline{U}}_e : T_e \times M$ with T_e being the number of experimental time points, N being the number of time series and M being the number of inputs. Furthermore, NetGenerator provides the option of introducing additional artificial data points by cubic spline interpolation.

GRNs considered as linear time-invariant systems

The NetGenerator algorithm infers dynamical models of GRNs. Their general non-linear dynamics can be described by a set of first-order time-invariant ordinary differential equations (ODEs), initial conditions, and time range (validity period)

$$\begin{aligned} \dot{\underline{x}}(t) &= \underline{f}\left(\underline{x}(t), \underline{u}(t), \underline{\theta}\right) \\ \underline{x}_0 &= \underline{x}(t_0) \\ t &\geq t_0 \end{aligned} \tag{1}$$

with the vector of state variables $\underline{x}$ and their changes $\dot{\underline{x}}$ as a function $\underline{f}$ of state variables, input vector $\underline{u}$ and parameter vector $\underline{\theta}$. The state variables and inputs depend on time t, the independent variable, that is dropped in further equations. The description is valid for a certain time range starting at t_0 from the initial conditions for the state variables $\underline{x}_0$. If not stated otherwise each of the state variables corresponds to one specific output variable, i.e. one time series. The dimensions of the variables are $\underline{x} : N \times 1$, $\underline{u} : M \times 1$, and $\underline{\theta} : P \times 1$, with N being the number of state variables, M the number of inputs and P the number of parameters.

Even though NetGenerator has a non-linear modelling option, the core mechanisms are based on linear modelling. Under the assumption that most networks can be considered linear and time-invariant, the differential equation system in (1) can be modified resulting in the linear state-space equation system

$$\dot{\underline{x}} = \underline{\underline{A}}\,\underline{x} + \underline{\underline{B}}\,\underline{u} \tag{2}$$

with the system or interaction matrix $\underline{\underline{A}} : N \times N$ and the input matrix $\underline{\underline{B}} : N \times M$. Most important for the understanding of the biological systems properties and the heuristic approach of the NetGenerator algorithm is the system matrix $\underline{\underline{A}}$ and its elements $a_{i,j}, i, j \in N$, because they describe the dynamics and the coupling of state variables.

Under the assumption, that the behaviour of a GRN is described sufficiently by indirect transcriptional events and not by a conversion of material, activation ($a_{i,j} > 0$) or inhibition ($a_{i,j} < 0$) of state variable x_i is not changing the value of the originating state variable x_j.

Without any further assumptions all elements of $\underline{\underline{A}}$ and $\underline{\underline{B}}$, adding up to $N^2 + M \cdot N$, had to be determined. Typically,

in GRNs there are far less connections than theoretically possible leading to a sparse matrix $\underline{\underline{A}}$. Regarding this property and avoiding problems occurring by the number of usually available measurement data (parameter identifiability, local or unique solutions, computational effort) the NetGenerator algorithm applies a heuristic approach as described in the next subsection.

Heuristic multi-stimuli multi-experiment approach

The novel NetGenerator algorithm is a heuristic multi-stimuli and multi-experiment approach. The heuristic is based on the observation that in GRNs the number of connections is much lower than all possible connections. Further, since the applied stimulus is the cause of the observed dynamical changes, the network can be considered as a hierarchical structure originating from the input. The NetGenerator algorithm implements both observations by an iterative development of the state-space system (2) by including *coupled sub-models* for each time series based on a structure optimisation iteratively increasing the number of connections. Structural changes are taking place only if they result in a better adaptation of simulated to measured behaviour. The terms multi-stimuli and multi-experiment mean that the extended algorithm can handle more than one changed input and data of several experiments, respectively.

In Figure 2 (A) the main work flow of the algorithm is displayed. One outer loop, starting with empty $\underline{\underline{A}}$ and $\underline{\underline{B}}$, iterates over all sub-models (state variables) to which the measured time series should be linked. At the ith iteration step of the outer loop already $i-1$ time series have been included in the model as sub-models. There are $N-i+1$ remaining time series to be included. The ith state equation (sub-model) would be described by

$$\dot{x}_i = \sum_{n \in N_i} a_{i,n} x_n + \sum_{m \in M_i} b_{i,m} u_m \tag{3}$$

containing connections from state variables, N_i being the indices of state-state connections including the self-regulatory term $a_{i,i}x_i$, and connections from inputs with M_i being the indices of input-state connections for the considered state variable x_i. That means that only the parameters of sub-models have to be identified.

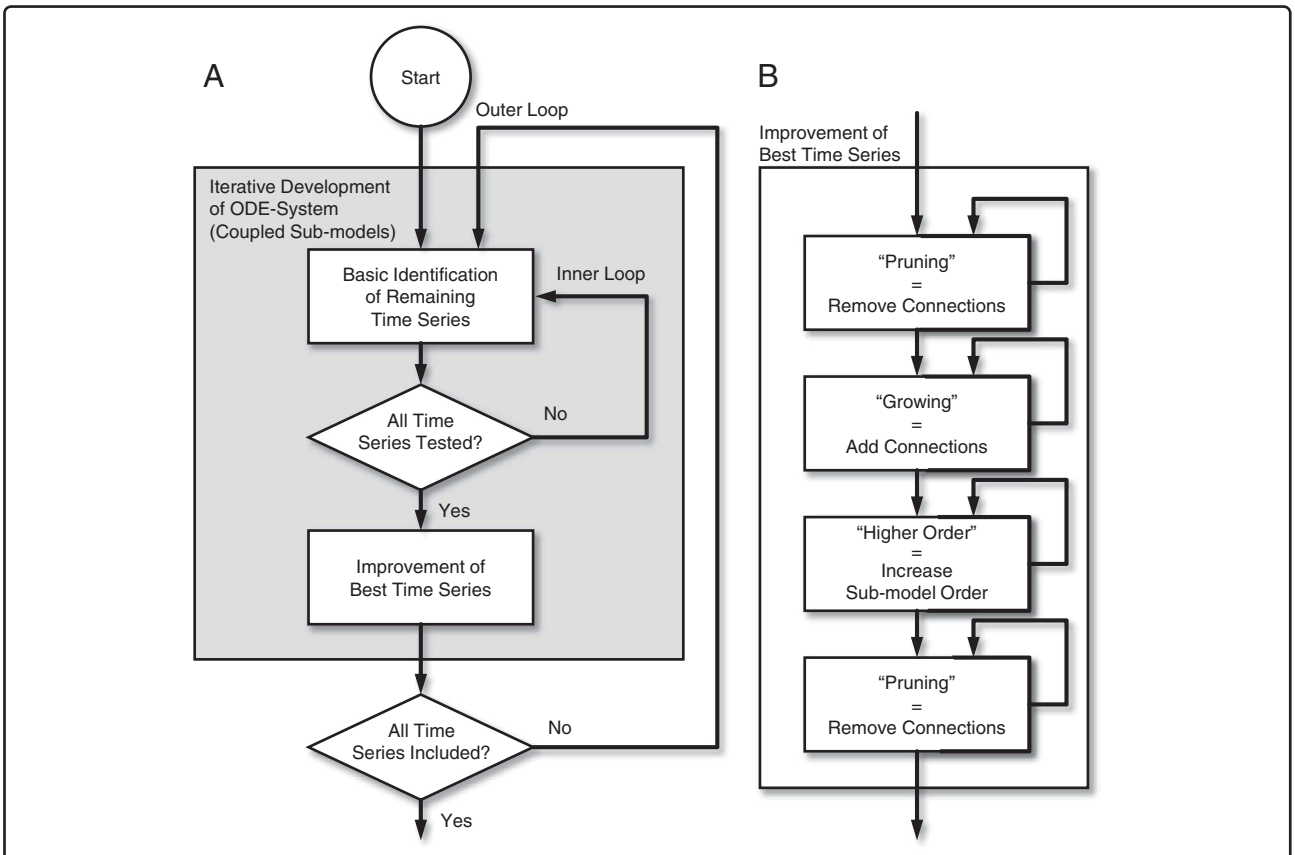


Figure 2 NetGenerator work flow. NetGenerator work flow displaying the main steps of the algorithm **(A)** and the improvement of best time series **(B)**. In the main work flow, the outer loop iterates over all state variables (sub-models), while in the inner loop the remaining time series are tested, i.e. a basic structure and parameters are identified. The best time series is improved further ("Growing", "Higher Order", "Pruning") and included into the model as a state variable.

Since the algorithm aims at a low number of parameters, i.e. small $|N_i| \leq N$ and $|M_i| \leq M$, the inner loop starts with basic models for the remaining time series containing only self-regulation, one input term as well as connections from "fix" prior knowledge if available, see respective subsection. Those basic structures can be extended by further incoming connections ("growing") from already included sub-models and further inputs. Every structural change requires a parameter identification of the active connections with respect to the considered time series, as will be explained later in the corresponding subsection. For every different set of parameters the resulting model needs to be simulated, that is the numerical solution of an initial value problem has to be found, as will be described later in another subsection.

The basic sub-model which reproduces one of the remaining time series best, is chosen for further improvement, for details see Figure 2 (B), and included into the model as a state variable. The most important structural improvements are

- "Growing": further connections added
- "Higher Order": increase sub-model order
- "Pruning": connections removed

In the improvement step "growing" is not restricted to connections from time series that are already included in the model. For describing the influence of time series that have not yet been included as sub-models, the corresponding measured and interpolated data are used as inputs. Those connections form global feedbacks in the final model.

The increase of the dynamical order within the description of a time series is realised by $r-1$ *additional* equations or intermediate state variables leading to the following form:

$$\begin{aligned} \dot{x}_i &= a_{i,i}x_i + \sum_{n\in N_i\setminus\{i\}} a_{i,n}x_n + \sum_{m\in M_i\setminus\{i\}} b_{i,m}u_m \\ \dot{x}_{i+1} &= a_{i,i}x_{i+1} + x_i \\ &\vdots \\ \dot{x}_{i+r-1} &= a_{i,i}x_{i+r-1} + x_{i+r-2} \end{aligned} \tag{4}$$

In this way the dynamics of a certain sub-model are described by an rth order integrator chain allowing for reproduction of processes with more complex time courses. It should be emphasised that by applying this approach the number of parameters is not increased but on the other hand the number of state variables becomes larger than the number of time series data. In that case only the state variable with the highest order in such a sub-model is to be compared to time series data. Still, for the sake of simplicity all following algorithmic procedures are described for first-order sub-models.

In terms of the iterative process of including sub-models the different elements of the *final* system matrix

$$\underline{\underline{A}} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N_s} \\ a_{2,1} & a_{2,2} & & \\ \vdots & & \ddots & \\ a_{N_s,1} & & & a_{N_s,N_s} \end{bmatrix} \tag{5}$$

describe forward, local feedback and global feedback connections. Elements below the main diagonal become forward connections, whereas the main diagonal elements $a_{1,1},\ldots,a_{N_s,N_s}$ describe local feedbacks or self-regulations, while the elements above the main diagonal represent global feedbacks. From a biological point of view the local feedbacks describe different mechanisms including not only feedback regulation, but also the important process of mRNA-degradation.

All the previously described structural procedures and the corresponding parameter identification are controlled by *a-priori* defined settings and options of the algorithm. Some of them are balancing network complexity and error between measurement and simulation. For example, additional connections are rejected if they are not improving the objective function value to a significant extent while on the other hand connections are removed only if they are not worsening the result significantly. Further important options of the algorithm are explained in the respective subsection.

Parameter identification

The parameter values of an active sub-model are identified by a non-linear optimisation, minimising the error between simulated and measured time series data of multiple experiments. The initial parameters required for this optimisation are obtained by a linear regression. For one specific first order state variable x_i equation (3) can be rewritten as

$$\dot{x}_i = \left[u_1,\ldots,u_{M_i},\, x_1,\ldots,x_{N_i}\right]\cdot \underline{\theta}_{i,\text{init}} \tag{6}$$

with

$$\underline{\theta}_{i,\text{init}} = \left[b_1,\ldots,b_{M_i},\, a_1,\ldots,a_{N_i}\right]^{\mathrm{T}} \tag{7}$$

being the parameter vector of some of the elements of $\underline{\underline{B}}$ and $\underline{\underline{A}}$, respectively, as determined by structural optimisation using only a subset of inputs and state variables influencing the considered ith state variable. Satisfying the measured data in an optimal way the unknown parameters can be determined by the following equation of linear regression, see e.g. [13],

$$\begin{aligned} \underline{\theta}_{i,\text{init}} = &\left(\left[\underline{\underline{U}} \quad \underline{\underline{X}}\right]^{\mathrm{T}} \underline{\underline{W}} \left[\underline{\underline{U}} \quad \underline{\underline{X}}\right]\right)^{-1} \\ &\times \left(\left[\underline{\underline{U}} \quad \underline{\underline{X}}\right]^{\mathrm{T}} \underline{\underline{W}}\,\underline{\dot{x}}_{i,\text{num}}\right) \end{aligned} \tag{8}$$

with the weight matrix $\underline{\underline{W}}$ and the state variable derivatives $\underline{\dot{x}}_{i,\text{num}}$. The latter are calculated by numeric differentiation of the respective time series (output) data. This means the vector $\underline{\dot{x}}_{i,\text{num}}$ is not the vector of state variables but the vector of time points of the considered time series derivatives. Its length ($T = K + K_{\text{interp}}$) equals the sum of the number of measurement time points and interpolation time points, as outlined in the subsection on data preprocessing. The reason for the use of interpolated data is the avoidance of over-fitting. The different influence of measured and interpolated values is considered in the elements of the weight matrix $\underline{\underline{W}}$ possessing the dimensions $T \times T$. Since the model must be valid for all E experiments, the respective input and time series data are *concatenated*, indeed resulting in $T = \sum T_e$, $e = 1, \ldots, E$. This becomes possible because the regression approach implicitly assumes a "dynamic independence" of data points. The dimensions of the other variables are $\underline{\underline{U}} : T \times |M_i|$ and $\underline{\underline{X}} : T \times |N_i|$, with the number of rows of each matrix also equalling to the total number of time points. Both dimensions of $\underline{\underline{U}}$ reflect necessary algorithmic changes due to the consideration of multi-input multi-experiment data in this NetGenerator version, because $|M_i| > 1$ represents multiple inputs while the concatenated data of length T considers multiple experiments. For the sake of completeness it should be mentioned that higher-order sub-models are initialised first by their first-order equation and then adapted such that total time constant and static gain remain the same.

The non-linear optimisation of the parameters for the ith sub-model, initialised by the solution of the linear regression (8), is based on the minimisation of the objective function (model error)

$$J_{i,\text{output}} = \sum_{e=1}^{E} \sum_{k=1}^{T_{e,i}} \left[w(t_k) \cdot \left(x_{e,i}(t_k) - \hat{x}_{e,i}(t_k, \underline{\theta}_i) \right)^2 \right] \quad (9)$$

describing the deviation between measured $x_{e,i}$ and simulated $\hat{x}_{e,i}$ time series at different time points t_k depending on the parameter vector $\underline{\theta}_i$. The minimisation following (9) is an optimisation problem of the least squares type featuring a double sum of experiments $e = 1, \ldots, E$ and time points $k = 1, \ldots, T_{e,i}$. In contrast to the objective function applied in former NetGenerator versions, now E multiple experiments are considered. The simulated time series are compared to measured and also interpolated data weighed by different $w(t_k)$ avoiding over-fitting. A further weighing based on properties of the data, like for example the maximal range, is not necessary since the described pre-processing normalises and scales the data. For the optimisation problem, the new NetGenerator implementation applies the "L-BFGS-B" algorithm, [14], of the `optim` R-function, which has the ability to solve bounded non-linear optimisation problems.

Consideration of prior knowledge

For improving the results, prior knowledge about the network connections can be integrated into the network inference. This version of NetGenerator provides two modes for integration of prior knowledge about connections of stimuli on time series as well as between the time series: (i) "fix" and (ii) "flexible". For both modes the knowledge can be provided in form of connection matrices $\underline{\underline{A}}_{\text{fix|flexible}}$ and $\underline{\underline{B}}_{\text{fix|flexible}}$ resembling the system matrix and input matrix, respectively, as well as additional matrices containing reliability scores of the connections. The connection matrices can contain single-valued information about connection (1), no connection (0), activation (10) and inhibition (−10). Fix integration represents rigid model requirements that cannot be ignored by the heuristic. Therefore fix connections are always included in the model structure.

Flexible integration allows the inference heuristic to ignore prior knowledge when the model fit is substantially worsened. This is represented by an additional term in the objective function (model error) now resulting in

$$J_i = J_{i,\text{output}} + \lambda \left[\sum_{j \in N_i} s^A_{i,j} d^A_{i,j} + \sum_{k \in M_i} s^B_{i,k} d^B_{i,k} \right]. \quad (10)$$

The term $J_{i,\text{output}}$ corresponds to the previously in (9) defined evaluation of output deviation, while λ weighs the overall consideration of prior knowledge, s represent the score values of the respective prior knowledge and d describe the distances between the prior knowledge and the modelled structure (incoming connections) evaluated by comparison of signs. That means the resulting elements of $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are converted into the described notation of 0, $\overline{1}$, 10, and −10, thus permitting a comparison with elements of flexible prior knowledge connection matrices. Here we consider two types of prior knowledge *origin*: (i) gene interactions automatically extracted from published literature and (ii) predicted transcription factor binding sites (TFBS) in the proximal promoter region of target genes.

For the extraction from published literature the software Pathway Studio V9 provides a gene relation database termed ResNet Mammalian, which has been compiled by automatic extraction of interactions from PubMed, as evaluated by [15]. As shown in the latter publication, gene relations derived from Pathway Studio V9 can be considered of high quality, since in general scientific literature is a reliable resource and the false positive rate is reported to be about 10 %.

Weber *et al. BMC Systems Biology* 2013, **7**:1
http://www.biomedcentral.com/1752-0509/7/1

Further, the tool matrix-scan from the RSAT toolbox determines putative TFBS in the promoter regions of target genes, which might be involved in transcriptional regulation [16]. This approach requires known sequence motifs of the investigated transcription factors as well as promoter sequences. Sequence motifs are stored in form of position weight matrices (PWM), which describe relative nucleotide frequencies for each motif position, as can be obtained from the Transfac database (Version 2010) [17]. Gene promoter sequences are available from Ensembl using biomaRt, [18].

Additional prior knowledge about the regulatory potential of the individual genes can be obtained by examining the known molecular functions. For example, the interaction between genes coding for non-regulatory proteins, such as structural proteins, and target genes can be assigned "no connection".

Simulation and graphical output

For every comparison of measurement and simulation as well as the generation of results the model equations (2) must be integrated. This corresponds to an initial value problem that is solved numerically. Since the recent implementation of the NetGenerator algorithm is in R, repeated operations of certain types take a long time. Therefore, the model itself is implemented in C, created iteratively and simulated applying the implicit method "impAdams" of the R-package `deSolve`, [19]. The necessary initial conditions $\underline{x}_0 = \underline{x}(t_0)$ are either measurement data or extrapolated measurement data typically at $t_0 = 0$ of the respective time scale.

The final result of the NetGenerator algorithm is a parametrised model of the considered GRN. Moreover, the new implementation of the algorithm contains important graphical output facilities which have been extended to meet the needs of displaying multi-input multi-experiment data as well as different results concerning prior knowledge. First, there is a graphical comparison of measurements and simulations, showing the single measured data points and the corresponding simulated trajectory. This can be done either by comparison of each component (gene) over all experiments or by displaying the data for each experiment independently. Second the resulting network structure can be displayed as a directed graph applying the language DOT and the software collection Graphviz, [20]. Nodes denote the biochemical components, e.g. genes, and edges display connections of either activation or inhibition. In case of applying prior knowledge (see respective subsection), a comparison between the inferred network and this knowledge is displayed with a colour code. Black edges denote inferred connections without prior knowledge, green edges present an agreement, red edges could either have a wrong sign (e.g. activation instead of inhibition) or be connections that do not comply with prior knowledge, while grey dashed edges stand for prior knowledge not reproduced in the inferred network.

Further settings and updated methods

The NetGenerator algorithm itself can be controlled by parameters (settings) and also contains further methods that will be summarised in the following. An important setting is the "allowedError" that controls the structure optimisation. If the objective function value of a certain sub-model structure is worse than this value the model structure must be extended as described. Therefore smaller values of "allowedError" are indirectly leading to more complex structures. Further important settings are the maximal number of connections and sub-model order.

Additional updated or new methods, not described extensively here, include non-linear modelling and knowledge-based methods. The optional non-linear modelling approach contains an additional sigmoid transformation of the linear combination described in this publication. This transformation has its biological background in the saturating behaviour of gene expression. The additional non-linear parameters of each sub-model are determined by the described non-linear parameter identification, too. Amongst further knowledge-based methods, the most important presents the possibility of retrieving network information from databases and combining this information with the inferred model in a directed graph. In that way, the biological interpretation can be extended by introducing unmeasured components into the network structure.

Availability

The algorithm has been implemented as a package in the programming language / statistical computing environment R, [9]. It is available in form of a testing bundle containing both the algorithm and the examples at www.biocontrol-jena.com/NetGenerator/NetGeneratorBundle.zip.

Results

Example networks

We applied the NetGenerator algorithm, which has been described extensively in the Methods section, to 3 benchmark examples and 1 real-world example to examine the performance of our approach. At first, we consider the three benchmark systems, their corresponding artificial data and inferred networks in order to test the reliability and performance of our algorithm. Particularly, we investigated whether network inference from multiple data sets, originating from different stimulation experiments, is beneficial. Finally, we applied NetGenerator to microarray time series data gained from human mesenchymal stem cells. We focussed on the modelling of gene regulation

that occurs during *in vitro* stimulation of chondrogenic differentiation of these cells, with emphasis on the different effects triggered by multiple stimuli in the inferred model.

Benchmark examples

We constructed three fully parametrised benchmark systems based on linear time-invariant descriptions, i.e. they are composed of differential equations representing the time series of genes and two external stimuli (u_1 and u_2). The systems are characterised by a different degree of cross-talk between the components with respect to the external stimuli, that is "full cross-talk" (FCT): all components are influenced by all stimuli, "limited or low cross-talk" (LCT): some of the components are influenced by more than one stimulus, and "no cross-talk" (NCT): the stimuli influence distinct components resulting in separate networks. They also differ in the number of genes (FCT: 5, LCT: 4, NCT: 7) and the parameters. The artificial data were generated exhibiting characteristics of real microarray time series data, i.e. low number of time points (six), exponentially increasing time intervals, and additional normally distributed noise $\mathcal{N}(0, 0.05^2)$. In summary, this procedure led to sample data sets containing matrices with number of rows equalling number of genes and six columns (time points).

Evaluation measures. The network inference of benchmark systems can be evaluated by determining the final objective function value (model error) J according to equation (10), the computation time t_C, and statistical measures that quantify the performance of the network inference by comparing the known structure with the inferred structure. The indicated computation times resulted from running the examples on a x86-PC with a 2.33 GHz CPU. The measures comprise sensitivity (SE), specificity (SP), precision (PR) and F-measure (FM). The definitions of the measures take into account the correctly integrated edges (true positives, TP), the falsely integrated edges (false positives, FP), the truly missing edges (true negatives, TN) and true edges that are not contained in the model result (FN). False positives (FP) were further grouped into FP_s, connections integrated with wrong sign and FP_n, modelled interactions which are not present in the real network. This leads to the following definitions:

$$\begin{aligned} SE &= TP/(TP + FN + FP_s) \\ SP &= TN/(TN + FP_n) \\ PR &= TP/(TP + FP_n + FP_s) \\ FM &= 2 \cdot PR \cdot SE/(PR + SE) \end{aligned}$$

For all three benchmark examples, we evaluated the inference by those statistical measures showing the reproduction of the system structure and time series by the model.

FCT scenarios and network inference evaluation. For FCT, artificial data generation and subsequent network inference was performed within three scenarios: (i) "S_1": single experiment applying only u_1, (ii) "S_2": single experiment applying only u_2 and (iii) "M": multiple experiment integrating experiments "S_1" and "S_2". For the special case of FCT, the scenarios allowed us to directly compare the inference of multiple stimuli data sets with the inferences of single stimulus data sets.

We applied the network inference to each of the three scenarios ("M", "S_1", "S_2") for a series of 10 different settings varying the previously described "allowedError"= 0.001, 0.002, ..., 0.01 resulting in 10 models, see Figure 3. Results for all statistical measures are depicted as connected points in individual boxes. The three scenarios are plotted in distinct colours ("M": blue, "S_1": red, "S_2": green) in each box. With regard to sensitivity, M models performs best, showing gradually decreasing values. Specificity obtains highest values for the first and second model. F-measure results, which benefit from high sensitivity values, display good performance for all M models. The resulting model error increases gradually as expected, due to the increased "allowedError", which is defined per time series. Analysing these results, we found "M 1" ($TP = 18$, $TN = 15$, $FP_n = 2$, $FP_s = 0$, $FN = 0$) to be optimal with respect to the evaluation measures ($SE = 1$, $SP = 0.88$, $FM = 0.94$, $J = 0.004$). For this model, the computation time was $t_C = 92$ s.

Dynamics of this model are displayed in Figure 4 showing a good reproduction of all time series for each of the two experiments. In Figure 5, the corresponding regulatory network is presented in form of a directed graph. Here, the colour code is not denoting a reproduction of prior knowledge but a graphical means displaying TP (green), FP (red) and FN (grey/dashed) connections.

LCT and NCT network inference evaluation. In order to test whether NetGenerator is capable of inferring different cross-talk structures, we generated benchmark systems LCT and NCT. Both contain biologically motivated types of cross-talk, such as cross-talk of downstream components or separate sub-networks (no cross-talk). Inference of both networks was successful, shown by high statistical measures ($SE_{LCT} = 1$, $SP_{LCT} = 1$, $FM_{LCT} = 1$, $J_{LCT} = 0.0007$, $SE_{NCT} = 0.9$, $SP_{NCT} = 0.98$, $FM_{NCT} = 0.92$, $J_{NCT} = 0.003$), the inferred network structures in Figure 6 and Figure 7, and the good reproduction of the time courses (Additional file 1 and Additional file 2). The computation time for inference of LCT and NCT was $t_C = 28$ s and $t_C = 33$ s, respectively.

Chondrogenesis model

Background of chondrogenic data. Human mesenchymal stem cells (hMSC) are multi-potent adult stem cells that have the capacity to differentiate into a variety of cell

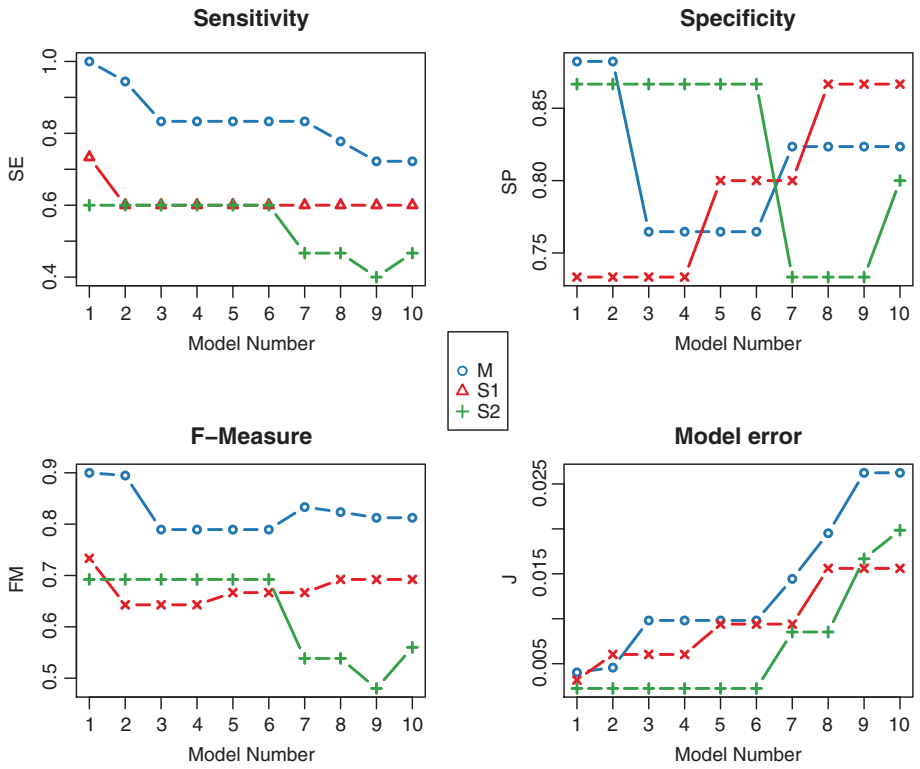


Figure 3 "Full cross-talk" example: inference statistics. Comparison of different NetGenerator inference results (varying the "allowedError" parameter) of three "full cross-talk" (FCT) scenarios ("M", "S_1", "S_2") using three statistical evaluation measures (sensitivity, specificity, and F-measure) and the model error.

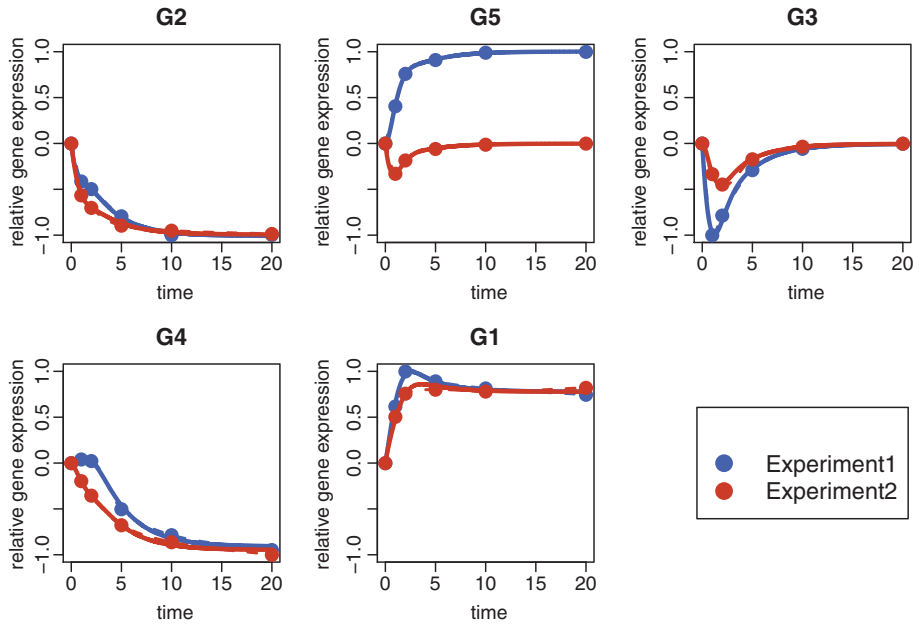


Figure 4 "Full cross-talk" example: time courses. Comparison of the "full cross-talk" (FCT) time courses. Each panel displays the results of one gene: the simulated time course (solid line), interpolated measurements (dashed line) and the measured time series (dots) for both data sets (Experiment1 and Experiment2).

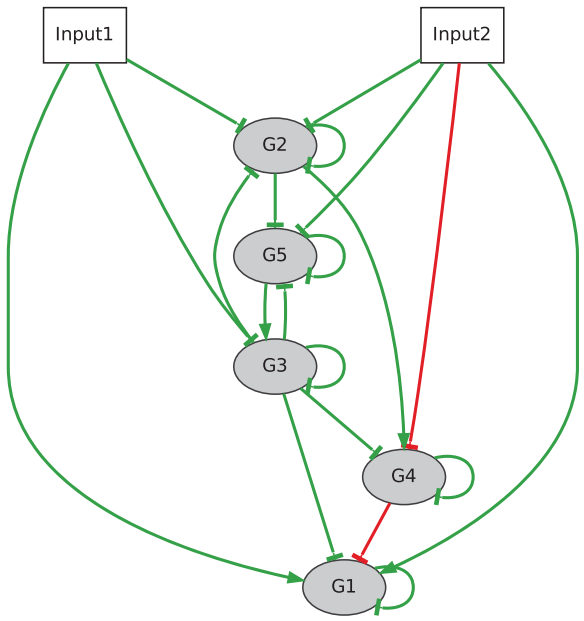


Figure 5 "Full cross-talk" example: inferred network. Network structure of the "full cross-talk" (FCT) model containing two simulated inputs (Input1, Input2) and five gene nodes. Inferred connections are highlighted in green (TP = 18), red (FP_n = 2) and dashed grey (FN = 0).

types depending on the external stimulus, [2]. Regulation of lineage-specific genes is crucial in this temporal process, [21]. Transforming growth factor (TGF)-beta1 is essential for induction of chondrocyte differentiation of hMSC, a process which is strongly enhanced by the additional presence of bone morphogenetic protein (BMP)2, [22] and [23]. In this section, we describe the complementary effects of TGF-beta1 and BMP2 by multi-stimuli multi-experiment inference applying the NetGenerator algorithm.

Microarray time series data. hMSC from bone marrow were commercially obtained (Lonza) and cultured as described in [2]. To induce chondrogenic differentiation trypsinised hMSC were pelleted and subsequently incubated in culture medium supplemented with 100 nM dexamethasone, 10 ng/mL TGF-beta1 and, if applicable, 50 ng/mL BMP2. Time-dependent gene expression was studied under three experimental conditions: (i) following treatment with TGF-beta1 ("T"), (ii) following treatment with TGF-beta1 + BMP2 ("TB") and (iii) untreated hMSC as a control. At 10 different time points (0, 3, 6, 12, 24, 48, 72, 128, 256, 384) h after addition of the stimuli, RNA was isolated from three technical replicates per time point and measured on Affymetrix HG-U133a microarrays.

Pre-processing and filtering. Raw microarray data was pre-processed as described in the corresponding subsection. This included the use of custom chip definition files provided by [10] and application of the RMA method [11]. This procedure resulted in logarithmised gene expression estimates for 12 095 genes.

Modelling a small-scale GRN using microarray data requires adequate filtering of genes. We tested all genes for differential expression, used functional annotation and expert knowledge. Differentially expressed genes were identified for both experiments ("T", "TB") by computing adjusted p-values using LIMMA. All genes with an adjusted p-value less than 10^{-10} and an absolute fold-change greater than 2 for any time point were considered significant. Using those criteria, 192 genes were found to be differentially expressed compared to control as well as between "T" and "TB". Subsequently, we selected from this group 10 annotated transcription factors (GO:0003700, sequence-specific DNA binding transcription factor activity) and associated 5 of them (SOX9, MEF2C, MSX1, TRPS1, SATB2) with our investigated process (GO:0051216, cartilage development). Those genes may be involved in promoter-dependent regulation, which is important for binding site predictions. Furthermore, we added COL2A1, ACAN, COL10A1, all three essential marker genes of chondrocyte differentiation, which encode essential structural proteins of the extracellular matrix.

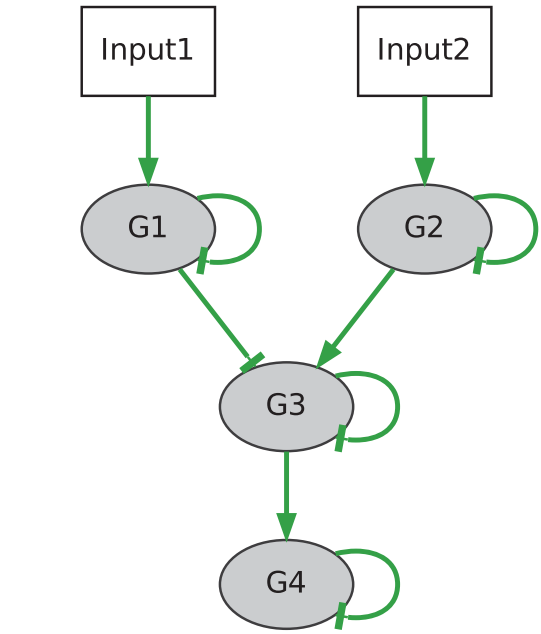


Figure 6 "Limited cross-talk" example: inferred network. Network structure of the "limited cross-talk" (LCT) model containing two simulated inputs (Input1, Input2) and four gene nodes. Inferred connections are highlighted in green (TP = 9), red (FP_n=0) and dashed grey (FN = 0).

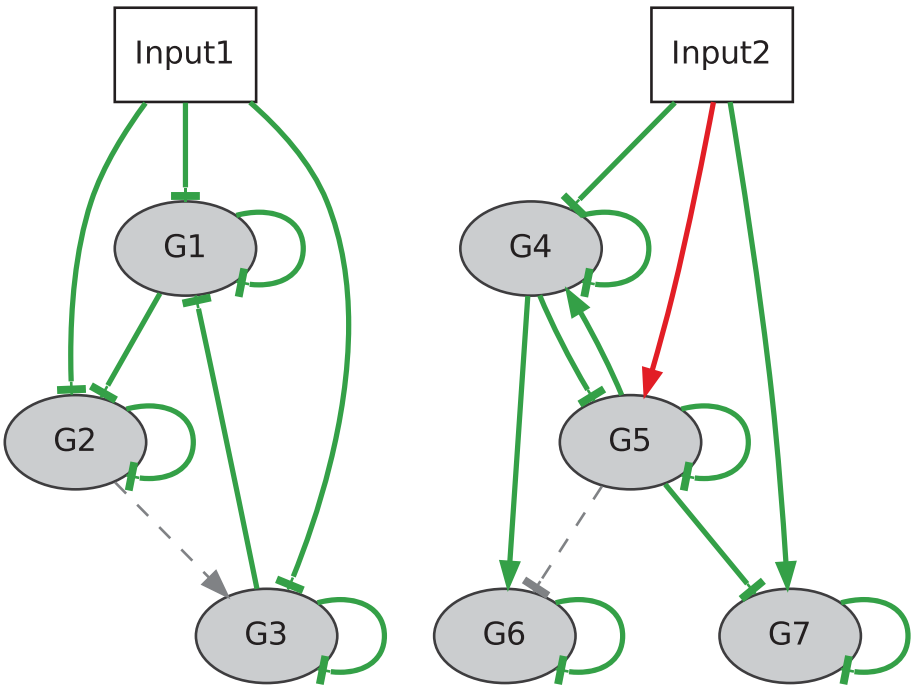


Figure 7 "No cross-talk" example: inferred network. Network structure of the "no cross-talk" (NCT) model containing two simulated inputs (Input1, Input2) and seven gene nodes. Inferred connections are highlighted in green (TP = 18), red (FP_n=1) and dashed grey (FN = 2).

Prior knowledge. Prior knowledge was taken into account as described in the corresponding sub-section. Gene interactions were retrieved from the Pathway Studio ResNet Mammalian database. We obtained 6 gene-gene and 5 input-gene regulatory interactions. Gene-gene interactions were passed as flexible prior knowledge to NetGenerator. Input-gene interactions were not integrated. Additionally, potential gene interactions were determined by binding site predictions. For this purpose, we obtained PWMs for SOX9, MEF2C and MSX1 from the Transfac database and promoter sequences 1000 bp upstream from the transcription start site. Both PWMs and sequences were loaded into matrix-scan from RSAT, which is performed with default options (weight-score $>$ 1, p-value $< 10^{-4}$) and organism-specific estimation of background nucleotide frequencies. The resulting significant binding sites have been listed in the table of Additional file 3. The observed high significance of all matches minimises the risk obtaining similar results from random sequences.

Network inference of multi-stimuli (TGF-beta1 and BMP2) multi-experiment data. After pre-processing, the input and time series data of the microarray experiments were passed to NetGenerator for automatic network inference. According to the experimental set-up, the available data sets describe two experiments: only TGF-beta1 stimulation ("T") and TGF-beta1 + BMP2 stimulation ("TB"). This is mirrored by the two distinct input data matrices both describing the respective stimuli by step functions

$$\underline{\underline{U}}_{\mathrm{T}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix}, \qquad \underline{\underline{U}}_{\mathrm{TB}} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

Model evaluation and validation. The inference results of the chondrogenic system, the GRN and the graphical comparison of time series, are displayed in Figure 8 and Additional file 4, respectively. The resulting network contains 20 connections: 14 gene-gene connections and 6 input-gene connections. Compared to the prior knowledge, there are 10 green connections (consistent), 1 red connection (wrong sign) and 3 blue connections (additional colour code for predicted binding site).

For validation, we performed resampling which is based on random perturbation of time series data. A Gaussian noise component $\mathcal{N}(0,\ 0.05^2)$ was added to the time series data which is used for subsequent model inference. Repeated performance (100×) led to a series of inference results as well as relative frequencies for each of the connections of the nominal model, i.e. the proportion of models containing that specific connection. Those frequencies imply a reliability ranking of all nominal connections. Most of the connections were inferred with high frequency, (76±24) %, see Figure 8. Particularly,

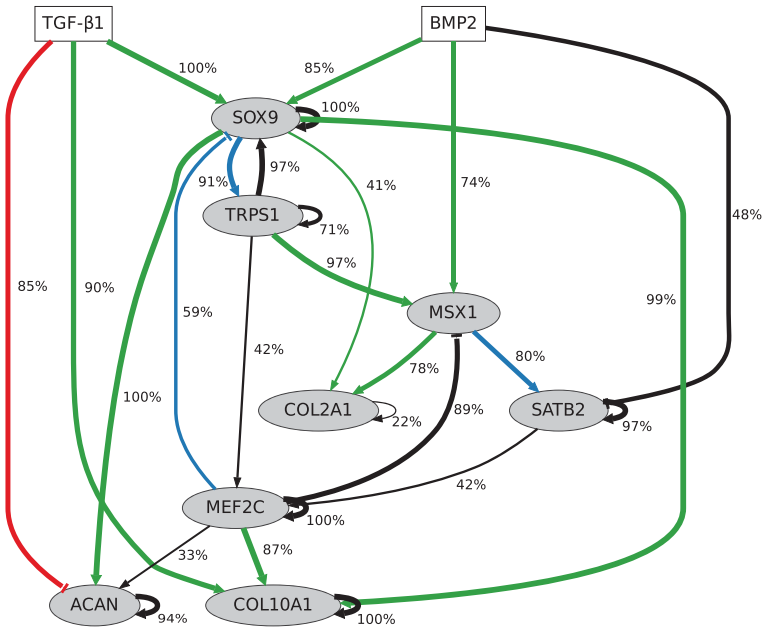


Figure 8 Chondrogenesis system: inferred network. Network of the chondrogenesis system, which contains two inputs (TGF-beta1 and BMP2). Nodes represent either transcription factor genes (SOX9, MEF2C, MSX1, TRPS1, SATB2) or genes coding for structural proteins (COL2A1, COL10A1, ACAN). Connections are coloured in green (consistent with prior knowledge), red (contrary to prior knowledge), blue (predicted connection with existing binding site) and black (predicted interaction). Connection widths and percentage labels illustrate the frequency of occurrence in the validation procedure.

this applies to connections which reflect prior knowledge. Also, inferred connections which are associated with a predicted binding site (blue colour) were present in more than 50 % of the models.

Network interpretation. SOX9 exhibits a central role in this chondrogenic network and is activated by both TGF-beta1 and BMP2. This indicates a complementary effect of both stimuli on the expression of SOX9. Activated SOX9 drives the expression of its target genes COL2A1, ACAN and COL10A1 [24-26]. This regulation marks the essential formation of cartilage-specific structural components of the extra-cellular matrix and the differentiation of hMSC towards chondrocytes. Beside this process, SOX9 also activates the repressor gene TRPS1 and vice versa. Regulatory interactions between both factors has not yet been addressed in the literature. Additionally, the SOX9 binding motif is present in the proximal promoter of the TRPS1 gene according to the prior knowledge. There is also a modelled effect from TRPS1 on MEF2C, which in turn activates COL10A1 and ACAN, but represses SOX9. This represents a negative global feedback from MEF2C on SOX9 in our model. MEF2C also represses the expression of MSX1, which is solely activated by BMP2 stimulus and activates COL2A1 according to the prior knowledge [27]. MSX1 also activates the SATB2 gene, which in turn activates MEF2C expression. Negative regulation of ACAN by TGF-beta1 is contrary to prior knowledge, as indicated by the red connection of the network graph. However, TGF-beta1 can also activate ACAN indirectly through SOX9, [24]. In summary, the central player SOX9 is influenced by both TGF-beta1 and BMP2. Essential structural proteins are not solely regulated by SOX9, but also by other transcription factors (MEFC, MSX1). Moreover, SOX9 and MSX1 are repressed by MEF2C through negative feedback that involves TRPS1 and SATB2.

Discussion

The NetGenerator algorithm for automatic network inference from multi-input multi-experiment time series data and prior knowledge, described in the methods section, will be classified and distinguished from other methods in the next sub-section. Therefore, its properties will be reviewed and justified showing advantages and disadvantages to other approaches. Our discussion contains a wide spectrum of other methods, but will only go into detail for the ones closely related to NetGenerator. Also, further specifications of NetGenerator will be summarised without a detailed comparison to other methods.

Classification of the algorithm

Good review articles on methods for automatic inference of GRNs can be found in [1,3,4]. The different methods can be classified by the data type (static or dynamic), the

mathematical approach (e.g. probabilistic vs. deterministic) and the result (e.g. undirected vs. directed graphs, algebraic correlation vs. dynamic models) whereby various combinations are possible. Mutual information methods (for a review of ARACNE, CLR and MRNET see [28]) are based on evaluating the statistical dependencies of large data sets resulting in undirected graphs. In comparison to NetGenerator they possess far different preconditions and purposes, for example they do not consider a concerted influence of the variables or the dynamics of the state-space concept, and therefore a more detailed comparison is set aside. Even though dynamical Boolean networks for gene-regulation, first proposed by [29], possess some similarities to discrete-valued state-space models, their rule-based approaches typically lead to rather qualitative results (for an overview of recent methods see the aforementioned review articles).

Very often, like in case of the core elements of NetGenerator, GRNs are based on linear modelling, i.e. the behaviour of one variable depends on a linear combination of other variables. Still the method can be a combination of either probabilistic or deterministic approach as well as algebraic correlation modelling (equations system) or dynamic modelling (differential equations system). In the case of the probabilistic modelling which is especially covered by static and dynamic Bayesian networks (see aforementioned review articles) the inference is based on the application of probability distributions to describe the uncertainties or noise inherent in GRNs. Beside the differences in the mathematical approach, probabilistic modelling includes the determination of statistical parameters and therefore generally more data replicates are required in comparison to deterministic modelling approaches such as NetGenerator.

Deterministic linear modelling applied to automatic network inference, [30], can be distinguished into at least two types depending on the results: (i) algebraic equations systems, e.g. [31], and (ii) differential (difference) equations systems, e.g. [32]. Although they have different prepositions on the dynamics of time series data, both types can be solved by linear regression. Still, there is a disproportion between the number of free parameters and available measurement data on the one hand and the property of sparsity of GRNs on the other hand. For the former interpolated data points can be applied under the assumption that the influence of the chosen interpolation on the results can be neglected. For the reproduction of sparse networks the regression can be combined with model reduction, for example using the large group of LASSO-based algorithms, see e.g. [33-36], on the basis of PCA (SVD), [37], or a combination of both, [38]. For further approaches, see the aforementioned review articles.

In contrast to all these methods, we propose the NetGenerator algorithm dealing with the problem of data number and sparsity in a different way. The algorithm is not inferring the network structure and parameters in one go. Instead we applied an heuristic approach of explicit structure optimisation, which iteratively generates a system of sparsely coupled sub-models. In that way, the GRN property of possessing more or less hierarchical input to output structures is reproduced. Thus, only the parameters of sub-models describing one time series have to be determined. A major drawback of regression-based solutions of linear differential (difference) equations systems is the necessity of applying numerical derivatives of small sample size and noisy data, which have a strong influence on the resulting network and modelled dynamics. NetGenerator uses a different solution, whereby the regression just provides initial parameters for a non-linear optimisation of an objective function of the least squares type. Overall, the final dynamic network can be obtained by a lower computational effort, because in comparison to the total number of parameters ($N^2 + M \cdot N$) in the model description (2) only a small number of parameters has to be determined.

Inference from multi-stimuli multi-experiment time series data

The concept of inferring from multiple data sets is also applied by [38], however on the basis of principal components of those data sets. The work of [39] provides a multiple methods framework to integrate distinct types of data like steady-state and time series data, focussing mainly on the combination of knock-out and stimulation data.

The proposed NetGenerator V2.0 algorithm allows for integrating data sets of multiple experiments with multiple stimuli. In the inferred models, weighed input terms represent external stimuli and resulting GRNs represent the merged effects of the diverse experiments. Therefore, from a biological point of view, the algorithm is able to handle experiments which investigate the degree of cross-talk.

We applied and tested this feature for 3 benchmark examples and 1 real-world example, the gene regulation during chondrogenic differentiation. The evaluation of the benchmark examples' results showed the power of the algorithm to infer the network structure and to reproduce the time series. Further, for a special system of "full cross-talk", i.e. all components are influenced by all stimuli, we could show that the simultaneous utilisation of different data sets leads to higher model quality compared to modelling data sets individually. The reason for this effect is due to the different stimulation by another external input which alters the time series data qualitatively and quantitatively, something that could not be achieved by biological replicates of a single input experiment. This underlines the benefit of using our integrated approach. Further, the

presented examples LCT and NCT are possible outcomes of GRN investigations. In the first case, there are two different types of genes: some are induced by one stimulus only and some are induced by multiple stimuli. The model inferred by NetGenerator contains both the separate and common structural elements. The special case of NCT occurs, if network parts are stimulated that are not connected at all. In summary, the extended NetGenerator takes advantage of multi-stimuli multi-experiment data by network refinement and extension.

We further inferred a two-stimulus network for hMSC differentiating towards chondrocytes. This network model contains gene regulatory events following the stimulation with two distinct chondrogenic factors, therefore providing a view on how genes involved in differentiation might be controlled by external molecules. Applying a subsequent resampling gives further information about the connections of this inferred GRN: (i) the majority of connections, especially the ones of prior knowledge and predicted binding sites, occur with a high frequency which can be considered a measure of reliability and (ii) the ranking of the frequencies can be used in interpreting the results with regard to biological hypotheses. Overall, this shows the importance for an extension of NetGenerator to deal with multiple data sets.

Consideration of prior knowledge

The means to integrate prior knowledge (fix and flexible) into the network inference is a distinctive feature of the extended NetGenerator algorithm achieved by modifying the objective function. This feature can reduce the complexity of the structure optimization, although it strongly depends on the origin and quality of the given knowledge, see e.g. [7]. Using prior knowledge for network inference can also be found in several other algorithms, see [1,3,4].

For our example of chondrogenic differentiation, we exemplarily showed network inference using flexible prior knowledge about regulatory interactions extracted from a database (Pathway Studio). The graphical evaluation of the inferred network showed very good reproduction of the proposed prior knowledge. Further predicted connections could be associated with potential regulatory binding sites generated from sequence data (Transfac, Ensembl).

Further aspects

Apart from the linear modelling presented in detail, the ability of NetGenerator to infer a non-linear model has been mentioned as a further option. The additional sigmoid function describing saturation in gene-expression has been proven successful before, e.g. [40-42]. Since the sigmoid transformation has also been used for neural network models, those inference methods are sometimes classified as such.

Besides the many advantages and possible application areas, there are minor restrictions of NetGenerator: it should be applied to pre-processed data without high correlations, it infers networks from measured time series data and due to the heuristic approach it cannot be proven that the global solution was found. The latter can be improved by decreasing the influence of noisy data using a bootstrap (resampling) approach, see chondrogenesis example and [1]. One feature which might be introduced in subsequent versions is the application of "interventional" multi-experiment data, i.e. data originating from perturbations within the system. This can be dealt with by applying either experiment-wise prior knowledge or an additional module in the structure optimisation explicitly dealing with that kind of data.

Conclusions

We presented the novel NetGenerator algorithm for automatic inference of GRNs, which applies multi-stimuli multi-experiment time series data and biological prior knowledge resulting in dynamical models of differential equations systems. This heuristic approach combines network structure and parameter optimisation of coupled sub-models and takes into account the biological properties of those networks: indirect transcriptional events for information propagation, limited number of connections and mostly hierarchical structures. The analysis of benchmark examples showed a good reproduction of the networks and emphasised the biological relevance of inferred networks with a different degree of cross-talk. The ability to infer a real-world example based on multi-stimuli multi-experiment data was shown by application of NetGenerator to a system of growth factor-induced chondrogenesis.

Additional files

Additional file 1: Figure: "Limited cross-talk" example, time courses. Comparison of the "limited cross-talk" (LCT) network time courses. Each panel displays the results of one gene: the simulated time course (solid line), interpolated measurements (dashed line) and the measured time series (dots) for both data sets (Experiment1 and Experiment2).

Additional file 2: Figure: "No cross-talk" example, time courses. Comparison of the "no cross-talk" (NCT) network time courses. Each panel displays the results of one gene: the simulated time course (solid line), interpolated measurements (dashed line) and the measured time series (dots) for both data sets (Experiment1 and Experiment2).

Additional file 3: Table: Results of RSAT. Results of RSAT matrix-scan tool using Transfac PWMs and genomic DNA sequences from Ensembl. Each row represents a predicted binding site with Transfac motif ("PWM"), target gene, start and end coordinates, the matched sequence, match score ("Weight") and associated *p*-value.

Additional file 4: Figure: Chondrogenesis system, time courses. Comparison of the chondrogenesis system time courses. Each panel displays the results of one gene: the simulated time course (solid line), interpolated measurements (dashed line) and the measured time series (dots) for both data sets ("T" and "TB").

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MW and SGH drafted the manuscript. SGH and DD contributed to the development and programming of the NetGenerator algorithm and software as well as to the mathematical and modelling background. MW and RG contributed to data processing, application of NetGenerator to examples, statistical evaluation and the biological interpretation. SV contributed to the generation of the benchmark systems and their artificial data. EJvZ contributed to experimental set-ups, measurements and biological interpretation of the chondrogenic investigation. All authors read and approved the final manuscript.

Acknowledgements
We would like to thank all our LINCONET project partners of the ERASysBio+ initiative. Also, we kindly acknowledge the support of this work by the BMBF (German Federal Ministry of Education and Research) funding MW within this initiative (Fkz. 0315719). Further, we acknowledge the Virtual Liver Network initiative of the BMBF for granting support (Fkz. 0315760 and Fkz. 0315736).

Author details
[1]Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Beutenbergstr. 11a, 07745 Jena, Germany. [2]BioControl Jena GmbH, Wildenbruchstr. 15, 07745 Jena, Germany, www.biocontrol-jena.com. [3]Department of Cell and Applied Biology, Radboud University, Heijendaalseweg 135, 6525 AJ Nijmegen, The Netherlands.

Received: 6 August 2012 Accepted: 15 December 2012
Published: 2 January 2013

References

1. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R: **Gene regulatory network inference: Data integration in dynamic models—A review.** *Biosystems* 2009, **96:**86–103.
2. Piek E, Sleumer LS, van Someren EP, Heuver L, Haan JRd, Grijs Id, Gilissen C, Hendriks JM, van Ravestein-van Os RI, Bauerschmidt S, Dechering KJ, van Zoelen EJ: **Osteo-transcriptomics of human mesenchymal stem cells: accelerated gene expression and osteoblast differentiation induced by vitamin D reveals c-MYC as an enhancer of BMP2-induced osteogenesis.** *Bone* 2010, **46**(3):613–627.
3. de Jong H: **Modeling and simulation of genetic regulatory systems: a literature review.** *J Comput Biol* 2002, **9:**67–103.
4. Bansal M, Belcastro V, Ambesi-Impiombato A, Di Bernardo D: **How to infer gene networks from expression profiles.** *Mol Syst Biol* 2007, **3:**78.
5. Guthke R, Möller U, Hoffmann M, Thies F, Töpfer S: **Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection.** *Bioinformatics* 2005, **21**(8):1626–1634.
6. Toepfer S, Guthke R, Driesch D, Woetzel D, Pfaff M: **The NetGenerator algorithm: reconstruction of gene regulatory networks.** In *Lecture Notes in Computer Science.* Edited by Tuyls K, Westra R, Saeys Y. Nowé A , Berlin and Heidelberg: Springer Berlin Heidelberg; 2007:119–130.
7. Linde J, Wilson D, Hube B, Guthke R: **Regulatory network modelling of iron acquisition by a fungal pathogen in contact with epithelial cells.** *BMC Syst Biol* 2010, **4:**148.
8. Albrecht D, Kniemeyer O, Mech F, Gunzer M, Brakhage A, Guthke R: **On the way toward systems biology of Aspergillus fumigatus infection.** *Int J Med Microbiol* 2011, **301**(5):453–459.
9. R Development Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria; 2008. [http://www.R-project.org].
10. Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, Shmoish M, Ferrari S, Lancet D, Danieli G, Bicciato S: **Novel definition files for human GeneChips based on GeneAnnot.** *BMC Bioinformatics* 2007, **8:**446.
11. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**(2):249–264.
12. Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**Article3.
13. Draper NR, Smith H: *Applied regression analysis*. 3. edition. A Wiley-Interscience publication, New, York: Wiley; 1998.
14. Byrd RH, Lu P, Nocedal J, Zhu C: **A limited memory algorithm for bound constrained optimization.** *SIAM J Sci Comput* 1995, **16**(5):1190.
15. Yuryev A, Mulyukov Z, Kotelnikova E, Maslov S, Egorov S, Nikitin A, Daraselia N, Mazo I: **Automatic pathway building in biological association networks.** *BMC Bioinf* 2006, **7:**171.
16. Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J: **RSAT 2011: regulatory sequence analysis tools.** *Nucleic Acids Res* 2011, **39**(Web Server issue):W86–W91.
17. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34**(Database issue):D108–D110.
18. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart – biological queries made easy.** *BMC Genomics* 2009, **10:**22.
19. Soetaert K, Petzoldt T, Setzer RW: **Solving differential equations in R: package deSolve.** *J Stat Software* 2010, **33**(9):1–25.
20. Gansner ER, North SC: **An open graph visualization system and its applications to software engineering.** *Software–Practice and Experience* 1999, **00:**1–5.
21. Hartmann C: **Transcriptional networks controlling skeletal development.** *Curr Opin Genet Dev* 2009, **19**(5):437–443.
22. Jin EJ, Lee SY, Choi YA, Jung JC, Bang OS, Kang SS: **BMP-2-enhanced chondrogenesis involves p38 MAPK-mediated down-regulation of Wnt-7a pathway.** *Mol Cells* 2006, **22**(3):353–359.
23. van der Kraan PM, Blaney Davidson EN, Blom A, van den Berg WB: **TGF-beta signaling in chondrocyte terminal differentiation and osteoarthritis: modulation and integration of signaling pathways through receptor-Smads.** *Osteoarthritis and Cartilage / OARS , Osteoarthritis Res Soc* 2009, **17**(12):1539–1545.
24. Sekiya I, Tsuji K, Koopman P, Watanabe H, Yamada Y, Shinomiya K, Nifuji A, Noda M: **SOX9 enhances aggrecan gene promoter/enhancer activity and is up-regulated by retinoic acid in a cartilage-derived cell line, TC6.** *J Biol Chem* 2000, **275**(15):10738–10744.
25. Yamashita S, Andoh M, Ueno-Kudoh H, Sato T, Miyaki S, Asahara H: **Sox9 directly promotes Bapx1 gene expression to repress Runx2 in chondrocytes.** *Exp Cell Res* 2009, **315**(13):2231–2240.
26. Oh Cd, Maity SN, Lu JF, Zhang J, Liang S, Coustry F, Crombrugghe Bd, Yasuda H: **Identification of SOX9 interaction sites in the genome of chondrocytes.** *PLoS ONE* 2010, **5**(4):e10113.
27. Craft AM, Krisky DM, Wechuck JB, Lobenhofer EK, Jiang Y, Wolfe DP, Glorioso JC: **Herpes simplex virus-mediated expression of Pax3 and MyoD in embryoid bodies results in lineage-related alterations in gene expression profiles.** *Stem Cells* 2008, **26**(12):3119–3129.
28. Meyer PE, Lafitte F, Bontempi G: ***minet*: A R/Bioconductor package for inferring large transcriptional networks using mutual information.** *BMC Bioinf* 2008, **9:**461.
29. Kauffman S: **Metabolic stability and epigenesis in randomly constructed genetic nets.** *J Theor Biol* 1969, **22**(3):437–467.
30. D'haeseleer P, Wen X, Fuhrman S, Somogyi R: **Linear modeling of mRNA expression levels during CNS development and injury.** *Pac Symp Biocomput* 1999, **4:**41–52.
31. Altwasser R, Linde J, Buyko E, Hahn U, Guthke R: **Genome-wide scale-free network inference for *Candida albicans*.** *Frontiers Microbiol* 2012, **3:**51.
32. Gustafsson M, Hornquist M, Lombardi A: **Constructing and analyzing a large-scale gene-to-gene regulatory network–Lasso-constrained inference and biological validation.** *IEEE/ACM Transac Comput Biol Bioinf* 2005, **2**(3):254–261.
33. Tibshirani R: **Regression shrinkage and selection via the Lasso.** *J R Stat Soc: Ser B (Stat Methodology)* 1996, **58:**267–288.
34. van Someren E, Wessels L, Reinders M, Backer E: **Regularization and noise injection for improving genetic network models.** In

Computational and Statistical Approaches to Genomics. 1. edition. Edited by Zhang W, Shmulevich I. NJ, USA: World Scientific Publishing Co.; 2002:211–226.

35. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V: **The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*.** *Genome Biol* 2006, **7**(5):R36.
36. Hecker M, Goertsches R, Engelmann R, Thiesen HJ, Guthke R: **Integrative modeling of transcriptional regulation in response to antirheumatic therapy.** *BMC Bioinformatics* 2009, **10:**262.
37. Bansal M, Gatta GD, Di Bernardo D: **Inference of gene regulatory networks and compound mode of action from time course gene expression profiles.** *Bioinformatics* 2006, **22**(7):815–822.
38. Wang Y, Joshi T, Zhang XS, Xu D, Chen L: **Inferring gene regulatory networks from multiple microarray datasets.** *Bioinformatics* 2006, **22**(19):2413–2420.
39. Gupta R, Stincone A, Antczak P, Durant S, Bicknell R, Bikfalvi A, Falciani F: **A computational framework for gene regulatory network inference that combines multiple methods and datasets.** *BMC Syst Biol* 2011, **5:**52.
40. Weaver DC, Workman CT, Stormo GD: **Modeling regulatory networks with weight matrices.** *Pac Symp Biocomput* 1999, **4:**112–123.
41. Wahde M, Hertz J: **Coarse-grained reverse engineering of genetic regulatory networks.** *Biosystems* 2000, **55**(1-3):129–136.
42. Mjolsness E, Mann T, Castaño R, Wold B: **From coexpression to coregulation: An approach to inferring transcriptional regulation among gene classes from large-scale expression data.** In *Advances in Neural Information Processing Systems, Volume 12*. Edited by Solla SA, Leen TK. Müller KR: MIT Press; 2000:928–934.

doi:10.1186/1752-0509-7-1
Cite this article as: Weber *et al.*: Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0. *BMC Systems Biology* 2013 **7**:1.

3.4 Dynamic modelling of microRNA regulation during mesenchymal stem cell differentiation

Dynamic modelling of microRNA regulation during mesenchymal stem cell differentiation

Michael Weber*[1] , Ana M Sotoca[2] , Peter Kupfer[1] , Reinhard Guthke[1] and Everardus J van Zoelen[2]

[1]Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Beutenbergstr. 11a, 07745 Jena, Germany
[2]Department of Cell and Applied Biology, Radboud University, Heijendaalseweg 135, 6525 AJ Nijmegen, The Netherlands

Email: MW - michael.weber@hki-jena.de; AS - a.sotoca@science.ru.nl; PK - peter.kupfer@hki-jena.de; RG - reinhard.guthke@hki-jena.de; EJvZ - vzoelen@science.ru.nl;

*Corresponding author

Dynamic modelling of microRNA regulation during mesenchymal stem cell differentiation

Michael Weber*[1] , Ana M Sotoca[2] , Peter Kupfer[1] , Reinhard Guthke[1] and Everardus J van Zoelen[2]

[1]Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Beutenbergstr. 11a, 07745 Jena, Germany

[2]Department of Cell and Applied Biology, Radboud University, Heijendaalseweg 135, 6525 AJ Nijmegen, The Netherlands

Email: MW - michael.weber@hki-jena.de; AS - a.sotoca@science.ru.nl; PK - peter.kupfer@hki-jena.de; RG - reinhard.guthke@hki-jena.de; EJvZ - vzoelen@science.ru.nl;

*Corresponding author

Abstract

Background Network inference from gene expression data is a typical approach to reconstruct gene regulatory networks. During chondrogenic differentiation of human mesenchymal stem cells (hMSCs), a complex transcriptional network is active and regulates the temporal differentiation progress. As modulators of transcriptional regulation, microRNAs (miRNAs) play a critical role in stem cell differentiation. Integrated network inference approaches determine interrelations between miRNAs and mRNAs on the basis of expression data as well as miRNA target predictions. We applied the tool NetGenerator to infer an integrated gene regulatory network.

Results Time series experiments measured mRNA and miRNA abundances of TGF-beta1+BMP2 stimulated hMSCs. Network nodes were identified using information about differential expression, predicted interactions, time series correlation and associated literature knowledge. Network inference was performed using NetGenerator to reconstruct a dynamical regulatory model for the given expression and knowledge data. The resulting model is robust and optimal with regard to data-fit and prior knowledge. It predicts the influence of miRNAs on the expression of chondrogenic marker genes and therefore proposes novel relations in differentiation control. Analysing the inferred network, we identified a previously unknown regulatory effect of miR-524-5p on the expression of the transcription factor SOX9 and the chondrogenic marker genes COL2A1, ACAN and COL10A1.

Conclusions Integrated network inference is beneficial to elucidate the impact of miRNAs on gene expression. Especially in biological processes which are driven by transcription factors, it reveals a new layer of transcriptional control. The tool NetGenerator is able to perform integrated network inference and leads to valid network results.

Background

Modelling of gene regulatory networks (GRNs) has become a widely used computational approach in systems biology [19]. This development has been greatly promoted by the availability of high-throughput data of adequate amount and quality, such as genome-wide expression data. Inference of regulatory dependencies between genes on the basis of such data is the major task. The inferred gene interactions constitute a network, which contains predictions about cellular regulation. Those can motivate the design of new experiments, which might validate and potentially elucidate unknown regulatory interactions. Successful applications of GRNs can be found in studies about specific human diseases (cancer [12], rheumatoid arthritis [18]), murine hepatocytes [41], the infection by *E. coli* [16, 25] and the pathogenic fungus *Candida albicans* [28]. In this study, we focus on the involvement of microRNAs (miRNAs) in the gene regulation of human mesenchymal stem cells (hMSCs) which differentiate towards chondrocytes. Therefore, we provide a biological background about hMSCs, characteristics and function of miRNAs and modelling approaches which integrate miRNA regulation.

hMSCs are multi-potent adult stem cells, which have the capacity to differentiate into multiple cell types, such as chondrocytes, osteoblasts and adipocytes [31,36]. Lineage commitment towards a certain type of cell depends on specific environmental factors. Those factors can activate intracellular signalling pathways which control developmental genes and other signalling pathways. Here, we focus on chondrogenic differentiation, which is characterised by a sequence of intermediate developmental stages, including cell condensation, proliferation, differentiation and hyperthrophy [7]. Each of the individual processes is associated with the activity and regulation of lineage-specific genes [17]. These genes encode e.g. transcription factors (e.g. SOX9, MEF2C) and ligands of distinct signalling pathways (e.g. TGF-beta1, BMP2, IHH, WNT) [24]. Stimulation of hMSCs by TGF-beta1 initiates the process of chondrogenic differentiation [39]. Although key genes were determined, the entire process of regulation in chondrogenesis is still not fully understood. In the recent years, it has become apparent that miRNAs are active regulators in the development of stem cells [23] [15].

MiRNAs are short ($\backsim$ 22 nucleotides), noncoding RNA molecules, which bind to complementary sequences in target mRNAs and repress translation or induce degradation [14]. Silencing of gene expression by these

post-transcriptional processes has revealed a new level of gene regulation which is capable of modulating expression levels after DNA-dependent regulation by transcription factors. For the human genome, more than two thousand mature miRNAs have been identified [13]. Much effort has been invested into unravelling the complex functional network of miRNA and target gene regulation. According to sequence predictions, one miRNA can target hundreds of potential target genes, while in turn one gene can be regulated by multiple miRNAs [10]. Considering the function of the target genes, miRNAs were found to target biological processes such as various signalling pathways and the cell cycle. Interestingly, transcription factor genes have been found to be overrepresented targets of miRNAs [5, 6].

Network inference approaches have considered the emerging knowledge about miRNA-dependent regulation by taking account of interactions between miRNAs and mRNAs. Those approaches utilise data about miRNA target predictions as well as miRNA and mRNA expression data. Consideration and integration of diverse data led to the extension of GRNs by the inclusion of post-transcriptional gene regulation. This new feature has promoted the analysis of dependencies between miRNAs and target genes. For example, tools like MAGIA [4], MMIA [30] and mirConnX [20] perform integrated network analysis on the basis of miRNA target predictions and correlation between miRNA and mRNA expression profiles.

In this study, we applied the tool NetGenerator for an integrated network inference based on mRNA and miRNA time series data as well as prior knowledge [16, 37, 43]. The resulting network predicts the activity of selected miRNAs in the chondrogenic regulatory network. In comparison to correlation-based approaches (e.g. MAGIA), the NetGenerator applies a dynamical model, which is based on ordinary differential equations.

Results and discussion

Chondrogenesis data and node selection

We analysed a dataset which contained mRNA and miRNA microarray measurements of cultured hMSCs in pellet cultures after stimulation with growth factors TGF-beta1 and BMP2. Both factors are known to induce the process of chondrogenesis [24, 39]. Microarray samples were available for 9 time points (0, 3, 6, 12, 24, 48, 72, 120 and 192) h with 3 (mRNA microarray) and 2 (miRNA microarray) replicates per time point. For mRNA microarray data pre-processing, custom chip definition files [9] were used in order to improve the accuracy of the expression estimates. Quantile normalisation was applied to mRNA and miRNA microarray data, respectively (see Methods). This resulted in time series expression data for 12,175 protein-coding genes (mRNAs) and 1,023 miRNAs. Integrated network inference requires the filtering of relevant

mRNAs and miRNAs, which constitute the network nodes in the model. A multi-step selection strategy was applied, which included statistical filtering, knowledge-based filtering and time series correlation, to identify miRNAs and genes that are associated with the investigated differentiation process. A workflow which illustrates the sequence of selection procedures, starting from microarray data and resulting in network components, is displayed in Figure 1. The statistical method LIMMA identified differentially expressed genes (DEGs) and miRNAs (DEMIRs) from the time series expression data, as described in the Methods section. It resulted in the selection of 192 DEGs and 485 DEMIRs. Subsequently, both sets were used to perform knowledge-driven selection, which is based on miRNA target predictions and literature about gene regulation during hMSC differentiation. It is known that the transition from stem cells to terminally differentiated cells is mainly controlled by transcription factors [7]. Moreover, transcription factor genes are reported to be potential targets of miRNAs, because they are significantly overrepresented among the miRNA target genes [6]. Based on this information, annotated transcription factor genes were selected using the Gene Ontology term GO:0003700 (sequence-specific DNA binding transcription factor activity) resulting in 10 differentially expressed transcription factor genes (DETFs). In the next step, predicted interactions between miRNAs and target genes were considered, as they provide a useful link between miRNA and mRNA data. Specifically, interaction data represents a subset of miRNAs and mRNAs and therefore promotes the selection process. There are numerous resources to obtain experimentally validated or computationally predicted interactions between miRNAs and target mRNAs, such as TarBase [40], miranda [3], miRBase [13], MirTarget2 [42] and TargetScan [10]. Access to those databases is provided by the R package RmiR.Hs.miRNA, which downloads the data in form of interaction tables. Overall, there are more than 1 million predicted interactions stored in the combination of the provided tables. However, most prediction approaches are reported to have relatively low specificity [1]. This issue can be addressed by combining the sequence-based predictions with the correlation of the corresponding time series expression data. To obtain the most reliable interaction predictions, two criteria were applied: (1) at least two of the five above mentioned databases store the interaction and (2) the associated miRNA-mRNA expression time series are negatively correlated. The first criterion ensures that the considered interactions were found by different approaches. For the second criterion, Pearson correlation between miRNA and mRNA time series was calculated for each interaction pair. Under the assumption that the predicted miRNA target gene interaction is functional, we would intuitively expect a negative correlation coefficient, due to the negative regulatory effect of the miRNA on the expression of its mRNA target. This assumption could be confirmed by [29], who successfully identified miRNA targets by correlation. However, the authors also emphasise that a strong miRNA effect on target gene expression

might be better recognisable on target protein level or downstream gene expression levels. Moreover, we noted that positive as well as negative correlations were observed for functional miRNA target relations in the literature [38]. In this study, we focused on strongly negatively correlated predictions only, which can be modelled by a repressing interaction between the miRNA and its target. As a result, four interactions with a correlation smaller than -0.8 were extracted (see Table 1). Therefore, the focus on negatively correlated interactions resulted in the selection of 4 DEMIRs and 4 DETFs.

As reported in the literature, there are prominent chondrogenesis marker genes such as COL2A1, ACAN (aggrecan) and COL10A1, whose expression level indicates the progress of differentiation [2,8]. They encode for structural proteins of the extra-cellular matrix (ECM) and are differentially expressed in our time series data. Therefore, we added them to the selection of network nodes, because marker genes help to monitor the effects of regulation by miRNAs and transcription factors on chondrogenic differentiation.

In summary, the applied multi-step selection procedure resulted in a set of 11 network components, including 4 miRNAs (miR-524-5p, miR-494, miR-298 and miR-500), 4 transcription factor genes (SOX9, TRPS1, MEF2C and SATB2) and 3 chondrogenic marker genes coding for components of the extra-cellular matrix (COL2A1, ACAN and COL10A1).

Network inference

The tool NetGenerator was applied to infer regulatory interactions among the network components and the influence of the external stimulus (TGF-beta1+BMP2). Input data of the algorithm comprised time series data and prior knowledge about potential regulatory interactions between the components. Time series data were extracted from the available miRNA and mRNA microarray datasets, averaged across replicates at each time point, centered and scaled by their maximum absolute value (see Methods). The resulting time series matrix has 9 rows (time points) and 11 columns (nodes). Prior knowledge about regulatory interactions was collected from diverse sources, which will be described below.

Extraction of prior knowledge

We considered knowledge about the general regulatory potential of each component as well as knowledge about regulatory interactions among the components for GRN inference. On the basis of the three component classes ((1) miRNA, (2) transcription factor gene, (3) marker/target gene), each of which was linked to its typical biological function. Those functions were translated into prior knowledge as follows: (1) miRNAs primarily function by degradation of their target mRNAs [14]. Therefore, they are expected to negatively regulate the expression of their respective target genes. (2) Transcription factors positively or negatively

regulate the expression of their target genes, which can be protein-coding genes as well as miRNA precursor genes. (3) Genes encoding for structural components of the extracellular matrix (ECM) are not known to have an effect on the expression of neither protein-coding genes nor miRNA genes. Therefore, they were considered to be pure target genes, whose expression is regulated by transcription factors and miRNAs. In addition to this functional annotation-based knowledge, a set of potential regulatory interactions was obtained from miRNA target predictions, as described in the gene selection section, and scientific literature. This included four predicted interactions from miRNAs on target genes, which have not been reported in the literature (see Table 1). To extract regulatory interactions from published literature, Pathway Studio V9 was applied, which provides a database of automatically derived interactions from PubMed [44]. In total, four interactions from transcription factors on target genes were retrieved from the database. SOX9 was found to regulate the expression of COL2A1, ACAN and COL10A1 by specifically binding to regulatory elements of those genes [26]. The chondrocyte hypertrophic marker COL10A1 is activated by MEF2C, which binds to conserved sequences in the promoter region [2]. Finally, the collected prior knowledge data were stored in form of interaction matrices (see Methods), which can be processed by NetGenerator.

Model inference and interpretation

Inference of the network model, which is based on ordinary differential equations, aims to find a solution which is optimal with respect to the given input data and the presumption of a sparse interaction matrix. Consequently, the algorithm's objective is to minimise the model error J, which quantifies the deviation between measured and simulated data, and to consider prior knowledge. The balance between network complexity and an adequate model error is controlled by the parameter "allowedError", which is the requested minimum error for each time series. A series of inference results varying this parameter was analysed with respect to model error, model complexity (total number of connections) and number of integrated prior knowledge connections (see Methods and Figure 2). This resulted in the selection of one model, which shows a good fit to the measured time series data (J=0.0833) and contains 8 prior knowledge interactions. Simulated model time courses and measured time series are displayed in Figure 3. The simulated time courses (blue lines) show a good reproduction of the measured time series data (black points). Although this inferred model seems to be excellent in terms of model fit and prior knowledge, model validation is necessary. The main reason is to prevent over-fitting of the measured time series data by the model. Therefore, the model's robustness against noise in the data was evaluated by using an approach which is based on repeated resampling of the time series data (see Methods). This procedure resulted in a table of occurrence frequencies for each interaction of the initial model (Additional File 1). While most of the

connections attained a high frequency $(86 \pm 22)\%$, two connections with a frequency less than 40% were discarded from the network. All remaining interactions were considered robust against minor fluctuations in the expression data. The network (Figure 4) consists of 11 nodes, i.e. 4 miRNAs, 4 transcription factors (TF), 3 target genes (chondrogenesis marker), and the external stimulus (TGF-beta1+BMP2). There are 19 stable connections in the network, which indicate transcriptional or post-transcriptional regulation, depending on the type of the connected components. Considering the proportion of nodes and connections (11 nodes / 19 connections) the network appears to be sparsely connected. There are 6 input-to-node and 13 node-to-node (miRNA/ TF/ target) interactions. Latter type of interactions can be further grouped into 1 (miRNA,miRNA), 6(miRNA,TF), 2(TF,miRNA) and 4(TF,target gene). Four connections are coloured in green, which reflects their concordance with literature knowledge. Four connections are coloured in blue, because they are underpinned by predicted miRNA target sites. Connections coloured in black represent regulatory hypotheses without further evidence.

Biological interpretation

In the following, the network model will be described and interpreted. The interpretation will be based on transcription factor nodes (SOX9, MEF2C, TRPS1, SATB2), to identify regulator and target nodes for each of them. This promotes the understanding of the model, particularly about how miRNAs might interfere with transcriptional regulation and control the differentiation process. Additional knowledge about the regulation of chondrogenic differentiation will be provided for interpretation of selected model characteristics. The input stimulus (TGF-beta1+BMP2) inhibits the expression of 3 miRNAs (miR-494, miR-524-5p, miR-298) and activates miR-500, which is in turn suppressed by TRPS1. Consequently, the negative effect of downregulated miRNAs on their target genes is attenuated, which leads to the activation of the transcription factor genes SOX9, MEF2C, TRPS1 and SATB2. SOX9, the main regulatory factor in chondrogenesis [27], is inhibited by miR-524-5p, which is supported by a predicted miRNA target site (Table 1). Since miR-524-5p expression is suppressed by the TGF-beta1+BMP2 stimulus, SOX9 expression increases and leads to activation of differentiation markers COL2A1, ACAN and COL10A1. This transactivation is achieved through the detection of a consensus binding motif ((A/T)(A/T)CAA(A/T)G), which is shared by the SOX family members [26]. In COL2A1, this motif could be identified multiple times in an enhancer located in intron 1. Activation of ACAN could be associated with the binding of SOX9 in its first intron [34] and the COL10A1 promoter contains a distal enhancer element 4.3 kb upstream from transcription start site [27]. Therefore, primary chondrogenesis might be under control of miR-524-5p by modulating the expression of SOX9 and its target genes. The MADS box transcription factor MEF2C, which controls chondrogenic

hypertrophy, positively regulates expression of COL10A1 through binding to conserved sequences in the promoter region [2]. Negative regulation of MEF2C by miR-298 might be a mechanism to prevent early activation of hyperthrophic genes. The transcriptional repressor TRPS1 is known to be activated by a specific type of BMP-signalling and promotes chondrogenic differentiation by transcriptional repression of only few known target genes [22]. In our model, its expression is regulated by the stimulus as well as by miR-494 and miR-524-5p. The interaction from miR-494 is underpinned by prior knowledge (blue connection in Figure 4), but surprisingly there also exists a predicted binding site of miR-524-5p within the TRPS1 mRNA. However, the positive sign of the connection suggests that the assumed inhibitory effect may be not reflected by the given data. In recent literature, extensive control of TRPS1 by at least 7 miRNAs has been described for the process of skeletal development [45]. In the network, TRPS1 inhibits the expression of miR-500 and miR-298, which controls the chondrogenic transcription factor MEF2C. While knowledge about target genes of TRPS1 is rare, it is known that TRPS1 can act positively on the chondrogenic marker gene COL10A1 and thereby promote chondrogenic differentiation [22]. SATB2, a transcription factor mainly associated with osteogenesis [17], is repressed by miR-500 and miR-494, as predicted by the model. A potential regulation of SATB2 by miR-500 is supported by the associated binding sequence (Table 1). However, since there is no influence of SATB2 on the expression of chondrogenic marker genes in our model, it is less relevant for chondrogenesis according to our model.

Overall, the involvement of transcription factor genes is a central part of the model. The model integrates transcriptional regulation (by transcription factor genes) and post-transcriptional regulation (by miRNAs) and thereby displays the interrelationship between miRNAs and transcription factors. Since all four investigated miRNAs are ultimately downregulated, the model proposes the suppression of miRNA activity, which gives rise to the activation of the transcriptional regulators of chondrogenic differentiation such as SOX9. The model comprises miRNAs acting on different stages of the differentiation process including early proliferation and late hyperthrophic stages. The downregulation of miR-524-5p constitutes an interesting prediction about how chondrogenic differentiation might be modulated on the level of post-transcriptional mRNA interference. Furthermore, we found expression of miR-524-5p to be oppositely (positively) regulated during osteogenic and adipogenic hMSC differentiation (data not shown). This indicates that the repression of miR-524-5p activity may be relevant for lineage specificity during hMSC differentiation.

Experimental validation

The inferred chondrogenic network implies that mir-524-5p is able to target SOX9 mRNA and thereby repressing expression of SOX9 and its target genes (COL2A1, ACAN, COL10A1). Therefore, we performed overexpression experiments of mir-524-5p in hMSCs to validate if chondrogenesis was impaired. Changes in chondrogenic differentiation were measured in basis of the expression of specific marker genes. For this, hMSCs were transfected with lentivirus harbouring the mir-524-5p coding sequence and a non-related murine Jnk RNAi lentivirus used as a negative control. Then, hMSCs were allowed to differentiate for 14 days into chondrocytes after which the expression of chondrogenic marker genes was measured by using qPCR. Relative expression of marker genes of transfected cells was compared to the negative control (incomplete: differentiation in culture medium without any growth factors added) and a positive control (TGFB1+BMP2: medium containing TGF-beta1 and BMP2) in which differentiation occurs in culture medium but without lentiviral transduction (see Figure 5). The positive control sets the baseline for comparison of the different expression levels, because differentiation is optimal. The results showed that mir-524-5p overexpression decreases the relative expression of all measured marker genes. The decrease in relative expression is significantly stronger than the decrease observed when using the non-related murine Jnk RNAi lentivirus. This negative control had no effect on chondrogenesis observed for all marker genes tested. In addition, hMSCs were transfected with mir-524-5p lentivirus and the empty pMIRNA backbone (PM_40) vector lentivirus (as negative control) and subsequently cells were allowed to differentiate for 14 days prior RNA qPCR analysis. The relative expression of SOX9 decreased when mir-524-5p was overexpressed and the control virus (PM_40) remained comparable to the relative expression of the positive control (TGFB1+BMP2). In conclusion, experimental validation showed lentiviral based overexpression experiments of mir-524-5p in differentiating hMSCs resulted in a significant inhibition of several chondrogenic marker genes compared to either non-transfected hMSCs or transfected with a control lentivirus.

Conclusions

In this study, we have demonstrated how miRNA regulation can be modelled by a dynamical GRN inference approach. This required the integration of mRNA and miRNA time series data of stimulated hMSCs, which underwent chondrogenic differentiation. We presented a filtering approach, in which specific biological knowledge (literature knowledge, transcription factor annotation, miRNA target gene predictions) was utilised in conjunction with statistical criteria. This filtering helped in dealing with the vast number of miRNA target gene predictions and in selecting highly relevant network components. Hereby, we detected 4 miRNAs

(miR-524-5p, miR-494, miR-298 and miR-500) and predicted their involvement in the gene regulation of chondrogenic differentiation. Applying the NetGenerator algorithm to the given data, we inferred a network model of moderate complexity, good data fit and robustness. We analysed this model by interpreting the interactions between miRNAs and transcription factors, while also considering the potential effect on chondrogenic marker genes. This analysis resulted in hypotheses and additional experiments which verified model predictions by showing that miR-524-5p can affect the expression of the central transcription factor gene SOX9 and differentiation marker genes. Therefore, this work showed how dynamic modelling of miRNA regulation can enhance the understanding of a specific biological process, such as hMSCs differentiation, and lead to the discovery of new regulatory interactions.

Methods

Culture and differentiation of human mesenchymal stem cells

Human mesenchymal stem cells (hMSCs), harvested from normal human bone marrow, were purchased from Lonza (Walkersville, MD) at passage 2. Cells were tested by the manufacturer and were found to be positive by flow cytometry for expression of CD105, CD166, CD29 and CD44 and negative for CD14, CD34 and CD45. We confirmed multipotency of all donor batches based on in vitro osteo-, chondro- and adipogenic differentiation capacity [36]. The cells were expanded for no more than 5 passages in 'mesenchymal stem cell growth medium' (MSCGM; Lonza, Walkersville, MD) at 37^{o}C in a humidified atmosphere containing 7.5% CO_2. Studies were performed with hMSCs from multiple donors, including 5F0138, 5F0138 and 1F1061. For chondrogenic differentiation, hMSCs were trypsinised and 2.5x 10^5 cells pelleted in a 10 ml round bottom tube (Greiner Bio-One, Monroe, NC) for 10 min at 250xg. Cell pellets were subsequently cultured for 21 days in chondrogenic differentiation medium, consisting of proliferation medium supplemented with 6.25 µg/ml insulin, 6.25 µg/ml transferrin, 6.25 ng/ml sodium selenite, 5.35 µg/ml linoleic acid, 400 µg/ml proline, 1 mg/ml sodium pyruvate, 10^{-7} M dexamethasone, 50 µg/ml sodium L-ascorbate (all obtained from Sigma-Aldrich, St. Louis, MO) , in the absence (incomplete or control) or presence of 10 ng/ml recombinant TGF-beta1 in combination with 50 ng/ml recombinant human BMP2 (TGF-beta1+BMP2). Growth factors were obtained from R&D Systems.

mRNA and microRNA profiling

Affymetrix Human Genome U133A (HG-U133A) microarrays were employed in triplicate experiments at 9 time points (0, 3, 6, 12, 24, 48, 72, 120 and 192 hours after onset of treatment with TGF-beta1+BMP2).

Further experimental details can be found in [36]. For miRNA profiling, 18 RNA samples were obtained from duplicate experiments, one biological condition and measured at 9 time points (0, 3, 6, 12, 24, 48, 72, 120 and 192 hours after onset of treatment with TGF-beta1+BMP2). RNA was extracted using TRIzol® according to the protocol provided by the manufacturer (Invitrogen). For each sample, 5 µg of RNA was used for miRNA profiling. Hybridisation and profiling were performed using Exiqon (Vedbaek, Denmark) capture probe sets spotted on Schott Nexterion Hi-Sense E glass slides [32].

Determination of relative expression levels of chondrogenic marker genes using quantitative PCR (qPCR)

Total RNA was isolated from chondrogenic pellets using the Mirvana (Ambion) kit according to manufacturer's instructions. The isolated total RNA (≈100 ng) was then used as a template in a 20 µl reverse transcriptase reaction using superscript reverse transcriptase from Invitrogen according to manufacturer's instructions using random hexamers to prime the reaction. The following cycling conditions were used: 10 min at 20°C, 45 min at 42°C and 10 min at 94°C. The resulting cDNA solution was diluted 5x by adding 80µl water. qPCR of chondrogenic markers was performd using the following human primers: COL2A1 (forward: 5'-CTGCCAGTGGGCAACCA-3'; reverse: 5'-TTTGGGTCCTACAATATCCTTGATG-3'), COL10A1 (forward: 5'-AAAGCTGCCAAGGCACCAT-3' and reverse: 5'-AGGATACTAGCAGCAAAAAGGGTATT-3'), ACAN (forward: 5'-GACAGAGGGACACGTCATATGC-3' and reverse: 5'-CGGGAAGTGGCGGTAACA-3') and SOX9 (forward: 5'-GCAAGCTCTGGAGACTTCTGAAC-3' and reverse: 5'-ACTTGTAATCCGGGTGGTCCTT-3'), expression values were normalised and corrected using RPS27a housekeeping gene (Forward: 5'-GTTAAGCTGGCTGTCCTGAAA-3' and reverse: 5'-CATCAGAAGGGCACTCTCG-3'). Relative expression was calculated using the following formula: Relative expression: $2^{-Ct} \cdot 10^6$ marker gene / $2^{-Ct} \cdot 10^6$ RPS27a. Data are presented as a fraction of RPS27a expression and all qPCRs were performed in duplicates.

Microarray data analysis

Microarray data pre-processing and network inference was entirely performed in the statistical programming environment R [33] using Bioconductor software tools [11]. Pre-processing aims to remove non-biological noise from the data and to estimate gene expression levels.

Pre-processing of mRNA microarray data

Data from mRNA microarray experiments were pre-processed using the customised chip definition pack-

age "gahgu133a" and the robust multi-array average (RMA) procedures [21]. The chip definition package provides custom probe-sets for the Affymetrix HG-U133A chip, which reduces the number of cross-hybridising probes [9]. The remaining probes allow for a one-to-one correspondence between probe-set and gene. RMA procedures were applied for background correction, quantile normalisation and summarisation. The resulting signal matrix contains the logarithmised gene expression estimates for 12,175 genes.

Pre-processing of miRNA microarray data

First, mean signal values were extracted for each of the measured miRNAs, respectively. Second, quantile normalisation was applied, which is provided by the RMA package. This led to logarithmised miRNAs expression estimates for 1,023 miRNAs. In contrast to mRNA microarray data, there can be multiple probe-sets representing the same miRNA.

Statistical filtering

We applied the statistical tool LIMMA [35], which is available as an R package, on the miRNA and the mRNA dataset, respectively. It provides routines for identification of differentially expressed genes using an empirical bayes approach. Time series data can be analysed by contrast terms, which were defined by subtracting the control group from the stimulus group at each time point. Statistical significance was determined by applying a moderated F-statistics. Finally, LIMMA returned a ranked table, which contains columns for gene name, fold-change and adjusted p-values. By applying thresholds for adjusted p-value and fold-change, a list of significantly regulated mRNAs and miRNAs was determined. Due to the fact that the replicate number in the miRNA dataset is low (2 replicates per time point), differentially expressed miRNAs were selected by using a 2-fold-change criterion, while for mRNA selection the fold-change criterion was combined with a p-value threshold (Benjamini-Hochberg adjusted p-value $\leq 10^{-10}$).

Time series standardisation

Time series standardisation is a required processing step before starting network inference applying NetGenerator [43]. It includes centering and scaling of each time series. Centering implies subtraction of the initial value at the starting time point from all values such that the transformed time series starts from zero. Subsequent scaling divides each time series by its maximum absolute value, which leads to gene-wise scaled data and time series varying within -1 and 1.

Network inference

Network inference was performed using the tool NetGenerator, which models gene regulation by a system of ordinary differential equations (Equation 1).

$$\dot{x}_i(t) = \sum_{j}^{N} a_{i,j} x_j(t) + b_i u(t) \tag{1}$$

Dynamic change of expression x_i of gene i is described by the sum of weighted gene expressions of N genes and the weighted input $u(t)$, which is a stepwise constant function representing the external stimulus (e.g. TGF-beta1+BMP2). Regulatory interactions are modelled by the interaction parameters $a_{i,j}$ and the input parameters b_i. A positive parameter value denotes an activating connection, a negative value denotes an inhibitory connection and the value zero denotes no connection. Consequently, the GRN structure is determined by the model's interaction parameters, which have to be identified by the NetGenerator algorithm. The algorithm's central part is an optimisation heuristic, which performs network structure and parameter optimisation. Structure optimisation applies the principle of sparseness. Iterative development of sparse sub-models explicitly restricts the number of found connections. In each development step, parameter optimisation is applied to obtain interaction and input parameter values. The resulting model contains a minimal number of parameters that are necessary to obtain a good fit between simulated model and measured time series. A more detailed description of the algorithm can be found in [16, 37, 43].

NetGenerator also allows for integration of additional information about regulation among the components, referred to as prior knowledge. As this knowledge is usually independent of the time series data, it represents valuable additional data for the network inference. NetGenerator is capable of using prior knowledge as proposals during the structure optimisation process, while also dealing with contradictions between prior knowledge and time series data. Knowledge data is provided in form of an interaction matrix, which contains values for information about a connection (coded by 1), no connection (0), activation (10), inhibition (-10) or not available (NA). NetGenerator provides a flexible integration mode which ignores prior knowledge interactions in case the model fit is worsened.

Since NetGenerator contains a heuristic core, it depends on the setting of configuration parameters. The central parameter "allowedError" controls the allowed deviation between simulated and measured data of each time series. To determine an optimal result, we performed a series of network inference runs varying the value of this parameter (0.001, 0.01 (0.005) 0.05) resulting in ten models (see Figure 2). The resulting models were assessed on the basis of the actual model error J and the number of successfully integrated

prior knowledge connections. An optimal model reproduces the data with a low error (high accuracy), while attaining a relatively low model complexity (number of interactions). Considering the ten models, we found the second model (0.01) to be optimal with respect to model error (J=0.0833), model complexity (21 interactions) and integrated prior knowledge connections (8). The network model is shown in Figure 4 and simulated time courses are shown in Figure 3.

For model validation, robustness of the inferred network against small changes in the time series data was tested. Such changes may occur in the data due to technical or biological variance in the data. A robust inference result is expected to maintain a similar network structure when the input data is perturbed. Therefore, we applied random perturbation of the time series data by sampling from a Gaussian noise distribution ($\mathcal{N}(0, 0.05^2)$). This noise was added to the time series data, which was used for subsequent network inference. This procedure was repeated 100x leading to a series of models, from which relative frequencies for each of the connections of the initial model were derived. Connections which were inferred with a frequency of at least 50% were considered stable and therefore reliable.

Maintenance, lentiviral transfection and induced chondrogenesis of hMSCs

hMSCs were maintained in DMEM medium supplemented with 10% FBS, 1% pyruvate, 1% L-glutamine, 100 U/ml penicillin and 100 µg/ml streptomycin (referred to as proliferation medium, PM) and incubated at 37°C and an humidified atmosphere containing 7,5% CO_2. The day before lentiviral transduction, about $5 \cdot 10^5$ cells were transferred to 25cm^2 flasks in PM and incubated for 18 hours, as before. Then, cells were transfected using lentivirus containing either the empty pMIRNA backbone vector (control) or pMIRNA vector with mir-524-5p premature DNA sequences (purchased from System Biosciences). Lentiviruses were added in various concentrations (20 ng, 40 ng and 80 ng virus/30.000 cells) in addition to 1 mg/L polybrene (Milipore). The transfected cells were incubated for 2-3 days to allow for lentiviral integration and expression of the introduced transgenes. Transfected hMSCs were grown as pellets (by centrifugation) in high-glucose DMEM supplemented with 100 U/ml penicillin, 100 µg/ml streptomycin, 1% L-glutamate 6,25 µg/ml insulin, 6,25 ng/ml sodium selenite, 6,25 µg/ml transferrin, 5,35 µg/ml linoleic acid, 400 µg/ml proline, 1% pyruvate, 100 nM dexamethasone, 50 µg/ml sodium ascorbate and 1,25 mg/ml bovine albumin (listed compound from Sigma). This medium will be further referred to as incomplete medium. Differentiation experiments were performed using incomplete medium in the presence or absence of 10 ng/ml TGF-beta1 and 50 ng/ml BMP2 (both purchased from R&D Systems). Differentiation of hMSCs chondrogenic pellets was allowed for 14 days.

Authors' contributions

MW and AS drafted the manuscript. MW performed pre-processing and modelling of the network. PK contributed to the network component selection. AS and EJvZ contributed to the experimental set-ups, measurements and biological interpretation of the network. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank all our LINCONET project partners of the ERASysBio+ initiative. Also, we kindly acknowledge the support of this work by the BMBF (German Federal Ministry of Education and Research) funding MW and PK within this initiative (Fkz. 0315719). We are grateful to Dr. Joris Pothof (ErasmusMC, Rotterdam) for his contribution to the microarray analysis of the miRNAs.

References

1. P. Alexiou, M. Maragkakis, G. L. Papadopoulos, M. Reczko, and A. G. Hatzigeorgiou. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25(23):3049–3055, Dec 2009.
2. M. A. Arnold, Y. Kim, M. P. Czubryt, D. Phan, J. McAnally, X. Qi, J. M. Shelton, J. A. Richardson, R. Bassel-Duby, and E. N. Olson. MEF2C transcription factor controls chondrocyte hypertrophy and bone development. *Dev. Cell*, 12(3):377–389, Mar 2007.
3. D. Betel, M. Wilson, A. Gabow, D. S. Marks, and C. Sander. The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, 36(Database issue):D149–153, Jan 2008.
4. A. Bisognin, G. Sales, A. Coppe, S. Bortoluzzi, and C. Romualdi. MAGIA^2: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). *Nucleic Acids Res.*, 40(Web Server issue):13–21, Jul 2012.
5. L. Croft, D. Szklarczyk, L. J. Jensen, and J. Gorodkin. Multiple independent analyses reveal only transcription factors as an enriched functional class associated with microRNAs. *BMC Syst Biol*, 6:90, 2012.
6. Q. Cui, Z. Yu, E. O. Purisima, and E. Wang. Principles of microRNA regulation of a human cellular signaling network. *Mol. Syst. Biol.*, 2:46, 2006.
7. B. de Crombrugghe, V. Lefebvre, R. R. Behringer, W. Bi, S. Murakami, and W. Huang. Transcriptional mechanisms of chondrocyte differentiation. *Matrix Biol.*, 19(5):389–394, Sep 2000.
8. B. de Crombrugghe, V. Lefebvre, and K. Nakashima. Regulatory mechanisms in the pathways of cartilage and bone formation. *Curr. Opin. Cell Biol.*, 13(6):721–727, Dec 2001.
9. F. Ferrari, S. Bortoluzzi, A. Coppe, A. Sirota, M. Safran, M. Shmoish, S. Ferrari, D. Lancet, G. A. Danieli, and S. Bicciato. Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics*, 8:446, 2007.
10. R. C. Friedman, K. K. Farh, C. B. Burge, and D. P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, 19(1):92–105, Jan 2009.
11. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5(10):R80, 2004.
12. J. Goutsias and N. H. Lee. Computational and experimental approaches for modeling gene regulatory networks. *Curr. Pharm. Des.*, 13(14):1415–1436, 2007.
13. S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, 36(Database issue):D154–158, Jan 2008.
14. H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840, Aug 2010.
15. L. Guo, R. C. Zhao, and Y. Wu. The role of microRNAs in self-renewal and differentiation of mesenchymal stem cells. *Exp. Hematol.*, 39(6):608–616, Jun 2011.
16. R. Guthke, U. Möller, M. Hoffmann, F. Thies, and S. Töpfer. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, 21:1626–1634, Apr 2005.
17. C. Hartmann. Transcriptional networks controlling skeletal development. *Curr. Opin. Genet. Dev.*, 19(5):437–443, Oct 2009.
18. M. Hecker, R. H. Goertsches, R. Engelmann, H. J. Thiesen, and R. Guthke. Integrative modeling of transcriptional regulation in response to antirheumatic therapy. *BMC Bioinformatics*, 10:262, 2009.
19. M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke. Gene regulatory network inference: data integration in dynamic models-a review. *BioSystems*, 96(1):86–103, Apr 2009.
20. G. T. Huang, C. Athanassiou, and P. V. Benos. mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic Acids Res.*, 39(Web Server issue):W416–423, Jul 2011.
21. R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, Apr 2003.

22. S. Itoh, S. Kanno, Z. Gai, H. Suemoto, M. Kawakatsu, H. Tanishima, Y. Morimoto, K. Nishioka, I. Hatamura, M. Yoshida, and Y. Muragaki. Trps1 plays a pivotal role downstream of Gdf5 signaling in promoting chondrogenesis and apoptosis of ATDC5 cells. *Genes Cells*, 13(4):355–363, Apr 2008.

23. K. N. Ivey and D. Srivastava. MicroRNAs as regulators of differentiation and cell fate decisions. *Cell Stem Cell*, 7(1):36–41, Jul 2010.

24. E. J. Jin, S. Y. Lee, Y. A. Choi, J. C. Jung, O. S. Bang, and S. S. Kang. BMP-2-enhanced chondrogenesis involves p38 MAPK-mediated down-regulation of Wnt-7a pathway. *Mol. Cells*, 22(3):353–359, Dec 2006.

25. C. Kaleta, A. Gohler, S. Schuster, K. Jahreis, R. Guthke, and S. Nikolajewa. Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis. *BMC Syst Biol*, 4:116, 2010.

26. V. Lefebvre, W. Huang, V. R. Harley, P. N. Goodfellow, and B. de Crombrugghe. SOX9 is a potent activator of the chondrocyte-specific enhancer of the pro alpha1(II) collagen gene. *Mol. Cell. Biol.*, 17(4):2336–2346, Apr 1997.

27. V. Y. Leung, B. Gao, K. K. Leung, I. G. Melhado, S. L. Wynn, T. Y. Au, N. W. Dung, J. Y. Lau, A. C. Mak, D. Chan, and K. S. Cheah. SOX9 governs differentiation stage-specific gene expression in growth plate chondrocytes via direct concomitant transactivation and repression. *PLoS Genet.*, 7(11):e1002356, Nov 2011.

28. J. Linde, D. Wilson, B. Hube, and R. Guthke. Regulatory network modelling of iron acquisition by a fungal pathogen in contact with epithelial cells. *BMC Syst Biol*, 4:148, 2010.

29. T. Liu, T. Papagiannakopoulos, K. Puskar, S. Qi, F. Santiago, W. Clay, K. Lao, Y. Lee, S. F. Nelson, H. I. Kornblum, F. Doyle, L. Petzold, B. Shraiman, and K. S. Kosik. Detection of a microRNA signal in an in vivo expression set of mRNAs. *PLoS ONE*, 2(8):e804, 2007.

30. S. Nam, M. Li, K. Choi, C. Balch, S. Kim, and K. P. Nephew. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Res.*, 37(Web Server issue):W356–362, Jul 2009.

31. E. Piek, L. S. Sleumer, E. P. van Someren, L. Heuver, J. R. de Haan, I. de Grijs, C. Gilissen, J. M. Hendriks, R. I. van Ravestein-van Os, S. Bauerschmidt, K. J. Dechering, and E. J. van Zoelen. Osteo-transcriptomics of human mesenchymal stem cells: accelerated gene expression and osteoblast differentiation induced by vitamin D reveals c-MYC as an enhancer of BMP2-induced osteogenesis. *Bone*, 46(3):613–627, Mar 2010.

32. J. Pothof, N. S. Verkaik, W. van IJcken, E. A. Wiemer, V. T. Ta, G. T. van der Horst, N. G. Jaspers, D. C. van Gent, J. H. Hoeijmakers, and S. P. Persengiev. MicroRNA-mediated gene silencing modulates the UV-induced DNA-damage response. *EMBO J.*, 28(14):2090–2099, Jul 2009.

33. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

34. I. Sekiya, K. Tsuji, P. Koopman, H. Watanabe, Y. Yamada, K. Shinomiya, A. Nifuji, and M. Noda. SOX9 enhances aggrecan gene promoter/enhancer activity and is up-regulated by retinoic acid in a cartilage-derived cell line, TC6. *J. Biol. Chem.*, 275(15):10738–10744, Apr 2000.

35. G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

36. Ana M. Sotoca, Michael Weber, and E.J. van Zoelen. Gene expression regulation underlying osteo-, adipo-, and chondro-genic lineage commitment of human mesenchymal stem cells. *Medical Advancements in Aging and Regenerative Technologies: Clinical Tools and Applications*, 7 Web:226–94, 2013.

37. S. Toepfer, R. Guthke, D. Driesch, D. Woetzel, and M. Pfaff. The NetGenerator algorithm: reconstruction of gene regulatory networks. 4366:119–130, 2007.

38. J. Tsang, J. Zhu, and A. van Oudenaarden. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol. Cell*, 26(5):753–767, Jun 2007.

39. P. M. van der Kraan, E. N. Blaney Davidson, A. Blom, and W. B. van den Berg. TGF-beta signaling in chondrocyte terminal differentiation and osteoarthritis: modulation and integration of signaling pathways through receptor-Smads. *Osteoarthr. Cartil.*, 17(12):1539–1545, Dec 2009.

40. T. Vergoulis, I. S. Vlachos, P. Alexiou, G. Georgakilas, M. Maragkakis, M. Reczko, S. Gerangelos, N. Koziris, T. Dalamagas, and A. G. Hatzigeorgiou. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res.*, 40(Database issue):D222–229, Jan 2012.

41. S. Vlaic, W. Schmidt-Heck, M. Matz-Soja, E. Marbach, J. Linde, A. Meyer-Baese, S. Zellmer, R. Guthke, and R. Gebhardt. The Extended TILAR Approach: A novel tool for dynamic modeling of the transcription factor network regulating the adaption to in vitro cultivation of murine hepatocytes. *BMC Syst Biol*, 6(1):147, Nov 2012.

42. X. Wang and I. M. El Naqa. Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, 24(3):325–332, Feb 2008.

43. M. Weber, S. G. Henkel, S. Vlaic, R. Guthke, E. J. van Zoelen, and D. Driesch. Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0. *BMC Syst Biol*, 7(1):1, Jan 2013.

44. A. Yuryev, Z. Mulyukov, E. Kotelnikova, S. Maslov, S. Egorov, A. Nikitin, N. Daraselia, and I. Mazo. Automatic pathway building in biological association networks. *BMC Bioinformatics*, 7:171, 2006.

45. Y. Zhang, R. L. Xie, J. Gordon, K. LeBlanc, J. L. Stein, J. B. Lian, A. J. van Wijnen, and G. S. Stein. Control of mesenchymal lineage progression by microRNAs targeting skeletal gene regulators Trps1 and Runx2. *J. Biol. Chem.*, 287(26):21926–21935, Jun 2012.

Figures

Figure 1 Gene selection workflow

This workflow illustrates the steps from pre-processed miRNA/mRNA microarray data to the selection of 11 network components, which includes statistical filtering (2), transcription factor annotation (3), negative correlation of predicted interactions (4) and identification of chondrogenesis marker genes (5).

Figure 2 Network model selection

A series of ten network inference results, varying the NetGenerator parameter "allowedError", is shown. For each inference result the model error (left ordinate) and the number of model connections / prior knowledge connections (right ordinate, orange/green) are displayed. For further analysis one model (highlighted in red) was selected (allowedError = 0.01).

Figure 3 Chondrogenesis model: time courses

Comparison of the measured and simulated time courses. Each panel displays the results of one model component: the simulated time course (blue solid line), interpolated measurements (black dashed line) and the measured time series (black dots).

Figure 4 Chondrogenesis model: inferred network

Network structure of the chondrogenesis model, which contains the input TGF-beta1+BMP2 and 11 nodes. Nodes represent either a miRNA (miR-524-5p, miR-494, miR-298, miR-500), a transcription factor gene (SOX9, MEF2C, TRPS1, SATB2) or a chondrogenic marker gene (COL2A1, COL10A1, ACAN). Connections are coloured in green (consistent with prior knowledge), blue (predicted miRNA target site) and black (predicted interaction).

Figure 5 Experimental validation

Barplots depict the relative expression of COL2A1, COL10A1, ACAN and SOX9, respectively, under a series of distinct conditions. Those include untreated cells (Incomplete), TGF-beta1+BMP2-treated cells (TGFB1+BMP2), lentiviral based miR-524-5p overexpression with three different concentrations (Mir-524_20, Mir-524_40, Mir-524_80) and negative control experiments (Jnk RNAi_20, Jnk RNAi_40, Jnk RNAi_80, PM_40).

Tables

Table 1 Predicted miRNA targets and associated time series correlation

miRNA	gene	method	Pearson correlation
hsa-miR-494	TRPS1	miranda, mirtarget2, targetscan	-0.89
hsa-miR-298	MEF2C	miranda, mirtarget2	-0.86
hsa-miR-500	SATB2	miranda, mirtarget2	-0.86
hsa-miR-524-5p	SOX9	miranda, mirtarget2	-0.88

Predicted miRNA target genes, corresponding prediction methods and the attained time series correlation.

Additional Files

Additional file 1 — Table of connection frequencies from model validation

A table of the network model connections and their relative frequencies in the model validation.

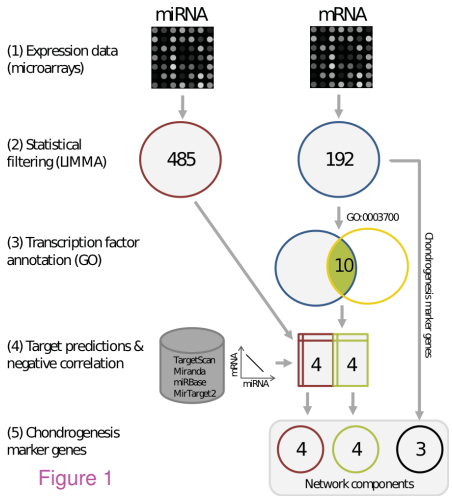

miRNA
mRNA
(1) Expression data (microarrays)
(2) Statistical filtering (LIMMA)
485
192
GO:0003700
(3) Transcription factor annotation (GO)
10
Chondrogenesis marker genes
(4) Target predictions & negative correlation
TargetScan
Miranda
miRBase
MirTarget2
mRNA
miRNA
4
4
(5) Chondrogenesis marker genes
4
4
3
Network components

Figure 1

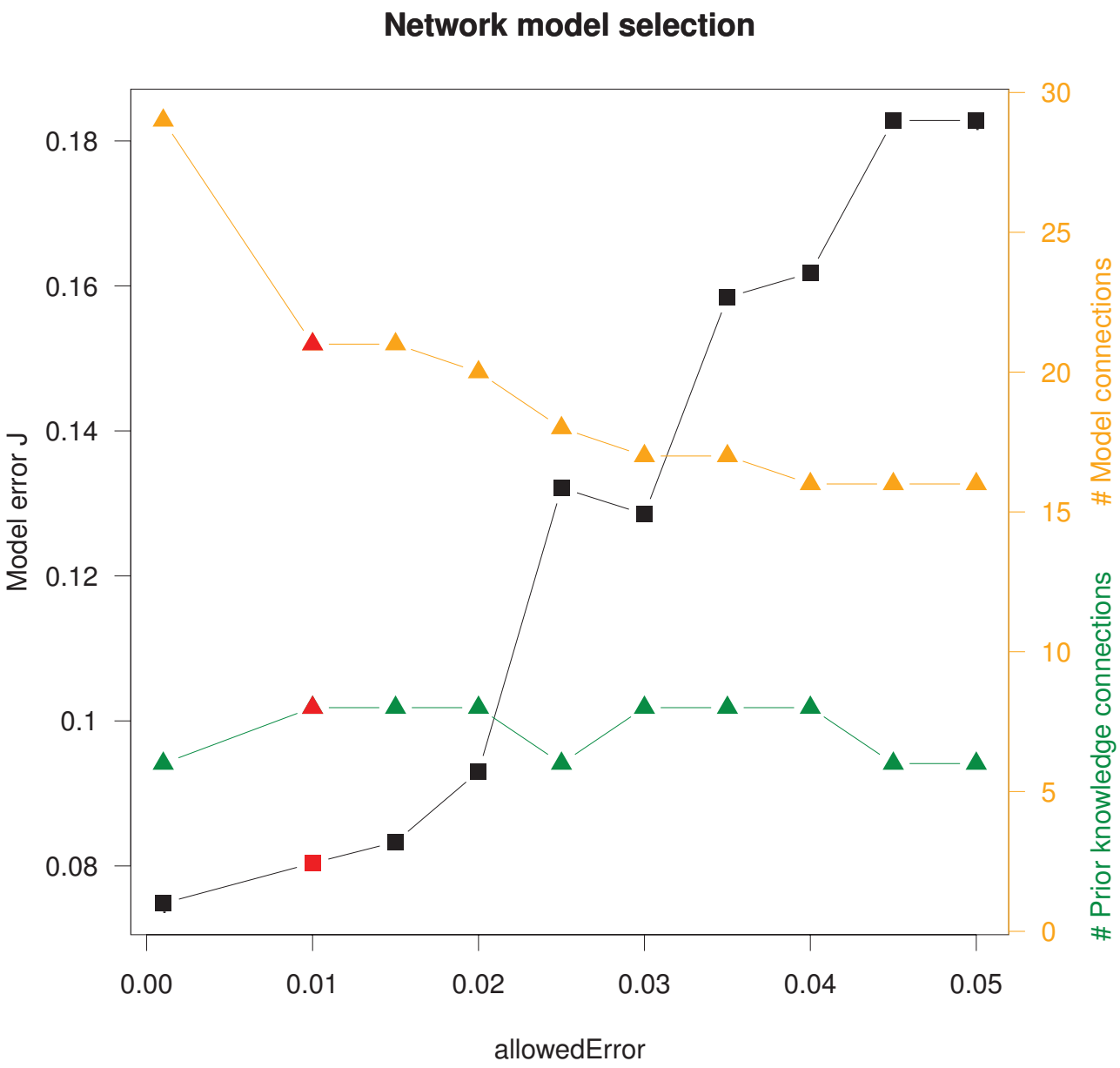

Network model selection
Model error J
0.18
0.16
0.14
0.12
0.1
0.08
0.00
0.01
0.02
0.03
0.04
0.05
allowedError
30
25
20
15
10
5
0
Model connections
Prior knowledge connections

Figure 2

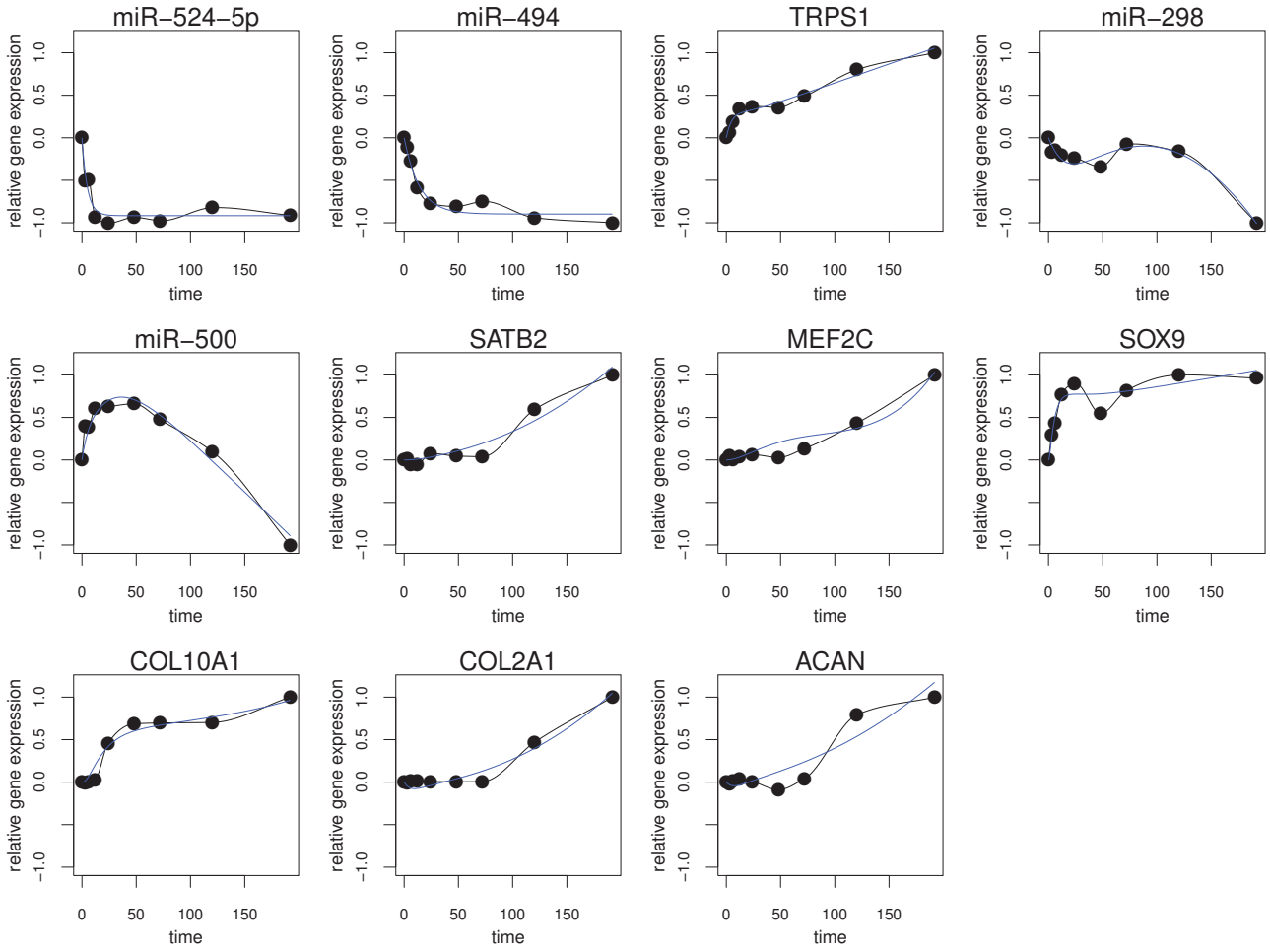

miR−524−5p
miR−494
TRPS1
miR−298
miR−500
SATB2
MEF2C
SOX9
COL10A1
COL2A1
ACAN
relative gene expression
time
−1.0
0.0
0.5
1.0
0
50
100
150

Figure 3

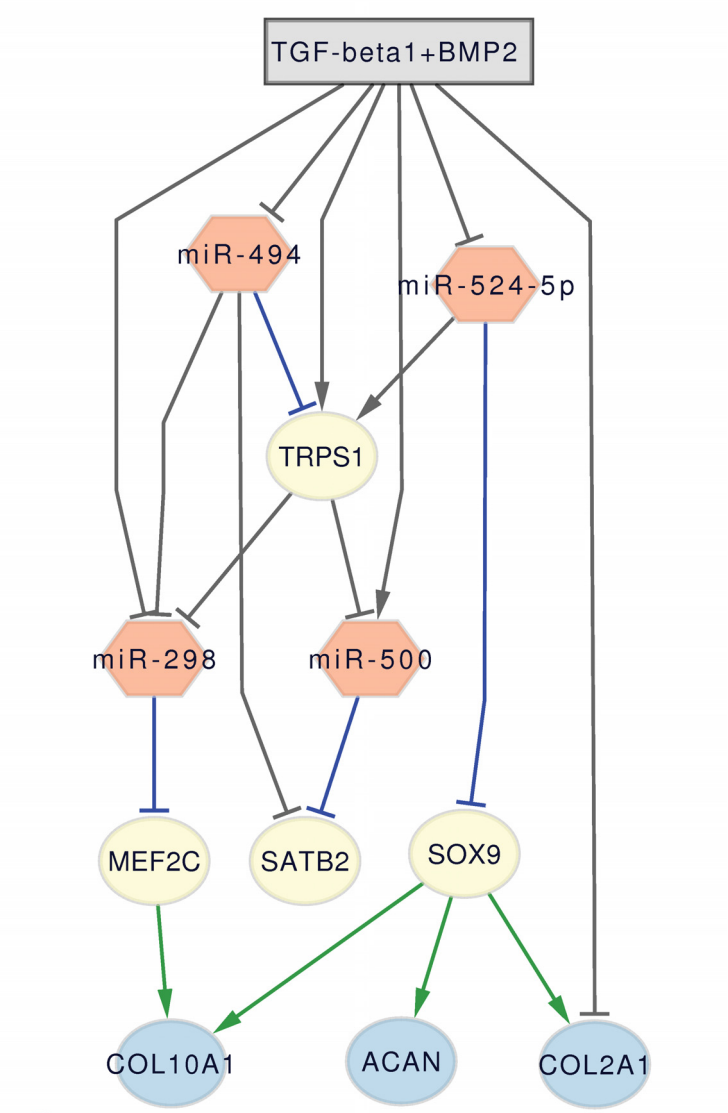
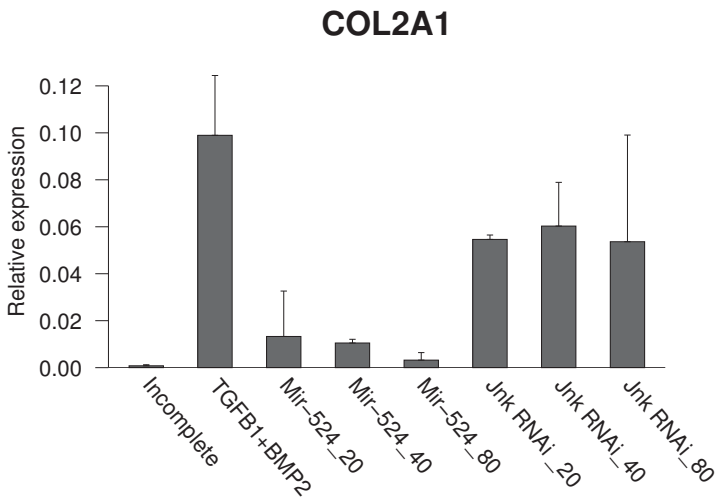
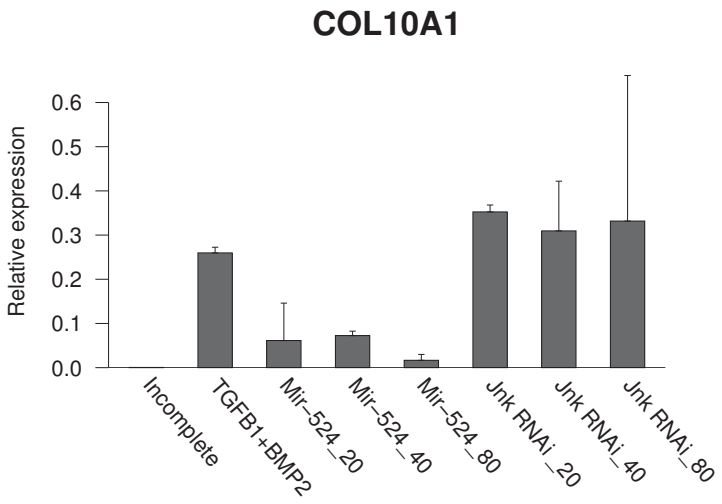
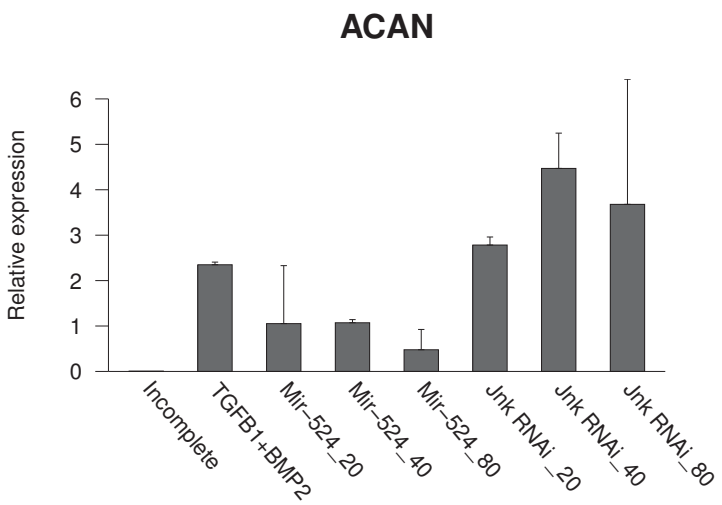

TGF-beta1+BMP2
miR-494
miR-524-5p
TRPS1
miR-298
miR-500
MEF2C
SATB2
SOX9
COL10A1
ACAN
COL2A1

Figure 4


COL2A1
Relative expression
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Incomplete
TGFB1+BMP2
Mir−524_20
Mir−524_40
Mir−524_80
Jnk RNAi_20
Jnk RNAi_40
Jnk RNAi_80


COL10A1
Relative expression
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Incomplete
TGFB1+BMP2
Mir−524_20
Mir−524_40
Mir−524_80
Jnk RNAi_20
Jnk RNAi_40
Jnk RNAi_80


ACAN
Relative expression
6
5
4
3
2
1
0
Incomplete
TGFB1+BMP2
Mir−524_20
Mir−524_40
Mir−524_80
Jnk RNAi_20
Jnk RNAi_40
Jnk RNAi_80

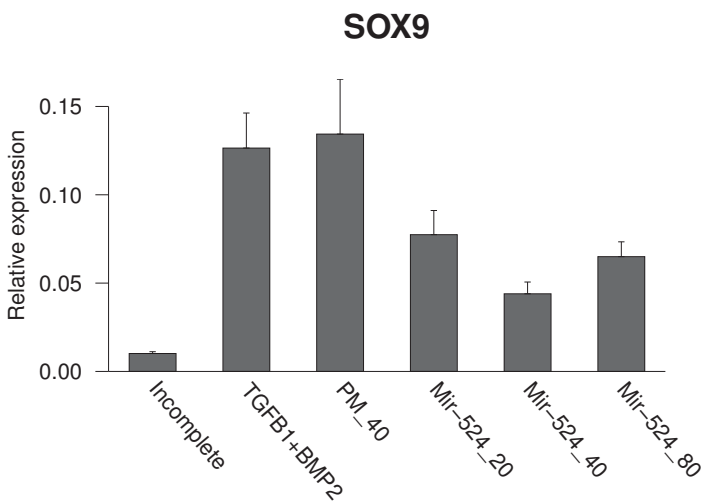

SOX9
Relative expression
0.15
0.10
0.05
0.00
Incomplete
TGFB1+BMP2
PM_40
Mir−524_20
Mir−524_40
Mir−524_80

Figure 5

Additional files provided with this submission:

Additional file 1: AdditionalFile1_ConnectionFreq.csv, 0K
http://www.biomedcentral.com/imedia/1374529957926817/supp1.csv

Chapter 4

Discussion

This thesis includes four papers which cover different parts of the process of network inference in the context of chondrogenic differentiation. The main biological objective has been the investigation of regulatory processes which are involved in the development of chondrocytes. To identify interesting regulatory hypotheses, this work applied microarray data analysis and an integrative network inference approach, which combines diverse types of biological data. The following chapter will discuss the presented results and highlight the main issues in network modelling on the basis of multiple microarray datasets.

4.1 Updated chip definition for microarrays

The application of an updated chip definition file (CDF) is the recommended initial step in microarray data pre-processing [65]. Replacement of the standard probe-set annotation intends to improve the reliability of the gene expression estimates, which are based on the corresponding probe intensities. Since all subsequent gene expression analyses operate on those estimates, it is crucial to ensure an accurate and specific mapping from probes to gene transcripts. To achieve this goal, probes have been classified according to their potential to cross-hybridise, e.g. bind RNA fragments of multiple mRNA species. Non-specific binding is known to be one reason for measurement errors in microarray data and therefore needs to be minimised by exclusion of potentially cross-hybridising probes.

Newly developed and publicly available CDFs were tested and evaluated on the basis of the resulting gene expression estimates. Each of the CDFs is characterised by its individual probe selection, which is based on different sequence annotation resources (e.g. GeneAnnot, RefSeq) and classification criteria. The subsequent reassembly of updated probe-sets therefore incorporates only the detected specific probes. The crucial objective of this study was to evaluate the performance of the different CDFs and to identify the most suitable for subsequent microarray analysis. Evaluation of the diverse gene expression estimates required the use of reference expression data in form of qRT-PCR data. Since the qRT-PCR technology is reported to show high assay specificity and detection sensitivity, the resulting data is suitable to be used as a validation reference [66]. Nonetheless, data obtained by this technology might also be affected by measurement errors, for example due to an amplification bias [67]. Furthermore, given that qRT-PCR experiments are generally more time-consuming, they are typically performed for only a small subset of genes which are measured by the microarray.

In section 3.1, two datasets were chosen for evaluation which included microarray as well as quantitative expression data from qRT-PCR experiments: (1) an illustrative collection of 16 genes obtained from a study of rheumatoid arthritis patients [68] and (2) a collection of 1000

genes from the MAQC dataset [36]. To avoid the effects of systematic bias, normalised microarray data and standardised qRT-PCR data were employed. The most important finding of this study is that the three custom CDFs performed, on average, considerably better than the standard annotation from Affymetrix on both datasets. This underscores the positive effect of the updated CDFs on the generated expression data and generally recommends the use of a custom CDF. Overall, the CDF provided by Ferrari et al. accomplished the best average correlation values between the microarray data and the qRT-PCR validation data. Moreover, this CDF provides a one-to-one correspondence between genes and probe-sets and therefore resolves the problem of multiple probe-sets per gene. The effectiveness of this approach is best reflected by the fact that 29% of the original probes are no longer considered by the updated CDF for the HG-U133A GeneChip, which is the central microarray platform used in this thesis [69].
A further question addressed the influence of microarray normalisation on the results. The standard MAS5 algorithm from Affymetrix [26] was compared to the alternative RMA normalisation method proposed by [20]. According to the correlation of the analysed datasets, RMA normalisation led to slightly better expression estimates. Nevertheless, the differences in the normalised measurements generated by using either RMA or MAS5 are generally considered to have a minor impact on the subsequent biological interpretation [7].

4.2 Genome-wide microarray analysis of hMSC differentiation

The investigation of multipotent hMSCs, which undergo differentiation towards diverse cell types, was the central aim of the underlying biological study presented in section 3.2. To analyse the biological processes taking place during cellular development, cells were exposed to various experimental conditions and were allowed to differentiate for a comprehensive time course. Genome-wide expression measurements were performed using Affymetrix GeneChips at predefined time points. The resulting microarray data contained time series datasets for chondrogenic, osteogenic and adipogenic differentiation that include diverse experimental condition groups. Time series experiments were designed to cover several developmental stages including growth, differentiation and maturation of the cells. The main objective was to monitor the expression levels of all genes which are involved in the cellular development. Since each of the developmental stages is associated with the activity of specific genes, it is important to understand how actively transcribed genes can influence the expression of target genes during the ongoing development. Another important aspect concerned the distribution of the time point samples along the differentiation process. The ten experimental time points were distributed on a logarithmic scale, which allowed for a high resolution of the early events i.e. the direct response to the stimulation, while fewer time points cover the late stages of cellular maturation.

Furthermore, to verify the successful differentiation of the investigated cells, staining experiments were conducted to detect specific characteristics of the mature cells. Such characteristics included staining intensity or proportion of positively stained cells as well as the size of cultured pellets. As a result, the observed patterns were found to show variation among the different experimental condition groups. On this basis, the individual performance of each condition was further evaluated. For example, the TGF-beta1+BMP2 (TB) stimulation showed enhanced staining effects, which indicated a better differentiation performance compared to the two other groups (TGF-beta1 and TGF-beta1+GDF5). In the case of osteogenic treatment, the differentiation performance of cells was enhanced up to 75% by adding VitaminD3.

The investigation of the underlying gene expression, which might give rise to those differences, was based on the computational analysis of microarray time series data. One bioinformatic challenge in microarray data analysis is the choice and correct application of tools which pre-process

the data, identify differentially expressed genes (DEGs) and model the gene expression profiles [35]. Even though those tasks are regarded as standard procedures for microarray analysis, it is crucial to be aware of the specific requirements and assumptions of the applied methods as well as the underlying experimental conditions of the given data. Otherwise the results might be affected by undesired properties and of low significance. Some methods are preferably used, because they have been adapted to the specific features of microarray data. This makes them superior to other methods. An updated chip definition file, which improves the mapping between probes and target genes, was applied in the first step of pre-processing as already described in section 4.1. In comparison to the standard CDF, the resulting gene expression estimates are more reliable, since they are less affected by inaccurate measurements. Subsequently, data normalisation was applied using RMA in order to remove systematic bias from the data and to generate gene expression estimates from the signal intensities. One assumption of the involved quantile normalisation is that the underlying distribution of gene expression intensities should be the same over all arrays [30]. Consequently, to achieve an optimal normalisation result, the three datasets (chondro, osteo, adipo) were normalised separately in order to account for their underlying experimental differences with regard to culture condition, cell density and cell donor.

The identification of DEGs was performed using the tool LIMMA [32]. This method can deal with time series data and applies statistical tests which take account of the low number of replicates per time point. Particularly, the variance estimation is not only based on the three available replicates, but also on the global expression of all genes, which results in more stable estimates [32]. The major objective of the gene selection process is the identification of biologically relevant genes. Therefore, genes were selected which display pronounced expression differences compared to the control group by applying a stringent adjusted p-value cutoff in combination with two-fold-change criterion. Particularly, this combination helped to considerably reduce the number of identified DEGs. In comparison, application of a conventional p-value cutoff without fold-change criterion led to the selection of gene time series which showed only marginal difference to control and therefore have only minor relevance. Sets of DEGs were identified for each stimulation group and therefore enabled a comparative analysis among the different groups. The functional analysis revealed genes that are associated with developmental processes, activity of signalling pathways and factors involved in cellular regulation. This provided the basis for identification of genes whose expression kinetics vary among the differentiation groups and therefore might contribute to the differentiation process. Overall, a large number of candidate genes has been identified, which reflect the variety of underlying processes involved in the cellular response. Based on the behaviour of the corresponding time series three distinct differentiation stages (growth, differentiation, maturation) can be distinguished. Each of those stages is associated with specific genes which are activated in the respective biological processes.

Since this thesis focuses particularly on chondrogenesis, one aim was to investigate genes which are associated with the three chondrogenic stimulations that might explain the distinct differentiation efficacy. Generally, the formation of chondrocytes requires the presence of TGF-beta1, which is considered as the essential growth factor to induce chondrogenic lineage commitment [68]. This fact is supported by the observation that the majority of DEGs is regulated by the TGF-beta1 stimulus. Interestingly, stimulation with TGF-beta1+BMP2 displays a large number of genes, which can be associated to the additional BMP2 stimulus. Moreover, the gene expression time series indicate a rapid progress of chondrocyte differentiation after TGF-beta1+BMP2 stimulation. SOX9, the main player in chondrocyte differentiation, showed a stronger upregulation, while chondrocyte marker genes COL2A1, ACAN and COL10A1 were found to show upregulation at earlier time points. Those results initiated further analysis which identified regulatory factors, that might be responsible for this behaviour. Similar behaviour has been observed in osteocyte differentiation and has been described as an acceleration phenomena driven by specific transcription factors [64].

4.3 Integrative inference of gene regulatory networks

4.3.1 Modelling multi-stimuli multi-experiment data

This thesis described and evaluated a novel tool for network inference from multiple experimental time series data. The availability of time series datasets, which investigate the influence of different experimental conditions, and the demand for rapid network inference results, have led to the development of NetGenerator V2.0. This extended implementation of the previously published NetGenerator algorithm represents an easy applicable and automatic network inference tool that includes several features which increase reliability and interpretability of the resulting model. The integration of multiple experimental data leads to one inferred network model instead of several individual network models for each dataset. Consequently, network model evaluation including analysis, interpretation and comparison is necessary for only one network and therefore more efficient. Furthermore, the applied optimisation procedures generally benefit from more data. The result is a refined and more reliable network structure, which accounts for all integrated datasets. The network visualisation of NetGenerator V2.0 includes input nodes, which represent the applied stimulus of each experiment. This directly supports the interpretation of the inferred model, since it allows to discriminate the effects of the multiple inputs on the network genes.

Besides those advantages, there is a basic assumption of the tool on the underlying biology of the data. In each of the incorporated experimental datasets, the structure of the underlying GRN is assumed to be identical. This implies that all variation in gene expression is only due to the impact of the applied experimental conditions, but not a result of structural changes in the GRN. In fact, the strength of the regulatory interactions among the genes is assumed to be independent of the applied stimulation. This modelling assumption is motivated by the knowledge about the nature of cellular regulation. Gene expression is controlled by signalling pathways, or more precisely by the activity of the involved transcription factors. Those are capable of binding to specific DNA motifs and thereby control gene expression of their associated target genes. Different experimental conditions may alter the activity of corresponding signalling pathways which in turn leads to a change in activity of the involved transcription factors. However, while transcription factor activity is modulated, their corresponding binding sites in the DNA sequence and the associated target genes remain unchanged. Nevertheless, there are biological mechanisms which have the potential to change binding site sequences, e.g. DNA mutation. Therefore, NetGenerator V2.0 is principally not applicable to integrate data from experiments based on a different genetic context. All network models which were inferred on the basis of multiple expression data and presented in this thesis satisfy the discussed biological assumption. In the case of the chondrogenesis model in section 3.3, both expression datasets were obtained from equally cultivated hMSC, whose differences in expression are assumed to be exclusively due to the applied stimulus.

Another point is the pre-processing of time series data before applying the NetGenerator tool. Centering and scaling are introduced as required processing steps in order to transform the data into an adequate format for the model optimisation process. Subtraction of the value at the initial time point, which represents the expression before stimulation, is referred to as centering. The centred time series start from zero and represent relative changes from the steady-state, rather than absolute measurement values. This supports the NetGenerator modelling concept which assumes the initiation of the time series from the steady-state. In the subsequent scaling procedure, each time series is divided by its maximum absolute value across all provided datasets. Specifically, the scaled measurements are all relative to the largest absolute expression value. This emphasizes the qualitative properties of the time series dynamics and prevents the influence of gene-specific expression magnitudes. Taken together, this thesis proposes GRN inference on the basis of centered and scaled time series, which start from a cellular steady-state

and range in the same interval for all genes.

4.3.2 Network inference validation using benchmark examples

Three benchmark networks were generated to test the performance of the NetGenerator V2.0 tool. Each of the computational examples consists of two weighted inputs, which represent the applied external stimuli. Since NetGenerator V2.0 allows for integration of multiple stimuli, cross-talk between the included inputs can be analysed. In the study of signalling pathways, there is a great interest to quantify the degree of cross-talk e.g. the extent of shared pathway components and interactions. Even though this study focuses on GRNs which generally summarise the effects of signalling pathways on the expression of downstream target genes, it is still interesting to investigate the cross-talk on this level. Therefore, examples were designed to contain different types of cross-talk, which may occur in the cellular response to two distinct external stimuli. The full cross-talk (FCT) network exclusively consists of genes which are affected by both inputs, while in the no cross-talk (NCT) network each gene is exclusively affected by one input. The limited cross-talk model (LCT) contains genes which are affected by either both or one of the inputs. Notably, all genes in all models are regulated by at least one input, because NetGenerator only deals with DEGs e.g. genes that are directly or indirectly affected by any of the considered stimuli.

To reconstruct the example networks, input data in form of time series samples needed to be generated. By individual simulation of each input, two time series datasets per network were generated. This data generation process considered some characteristics of real-world microarray data in order to ensure a relatively realistic benchmark evaluation. Those characteristics included the logarithmic scale of the sampled time series, the few number of samples per time point and the inherent noise of microarray data. Three replicates were generated by adding random Gaussian noise to the simulated time series. Applying the NetGenerator V2.0 to the three benchmark examples resulted in models which reproduce the time series with high accuracy. More importantly, the underlying network structures were inferred with a relatively high quality. In the case of the FCT model, it was shown that the combination of the two datasets improved the inference quality compared to individual application of each dataset. Overall, this benchmark study demonstrated the tool's general capability to infer networks of diverse cross-talk structure by combining multiple experimental data. Therefore, application of NetGenerator V2.0 provides a novel way to systematically and automatically investigate the common regulatory effects of various signalling pathways. This approach can enhance biological understanding about the complexity of cellular regulation in an efficient and powerful manner.

Another study which also deals with the generation of benchmark examples for the validation of network inference approaches is the Dream Challenge [70]. The provided *in silico* challenges are available for small-scale as well as large-scale networks and provide a range of experimental data including time series, knock-out and knock-down data. The difference to our benchmark data is that they generally provide more equidistant time points (21 in DREAM3 Challenge4). Additionally, different perturbations of the same network are provided. Future work on NetGenerator should evaluate its performance on the various *in silico* challenges provided by the Dream project.

4.3.3 Integration of prior knowledge

One distinctive feature of NetGenerator is its ability to integrate prior biological knowledge into the inference process. Inclusion of such information is regarded as beneficial to infer networks of high biological plausibility [14]. As a novel feature, NetGenerator V2.0 supports the flexible

integration of prior knowledge which is implemented by an extended objective function of the optimisation algorithm. Soft integration of the known connections represents a valuable feature, because it enables the rejection of inappropriate connections which cannot be reproduced by the model with regard to the given time series data. Therefore, prior interactions constitute proposals for the inference rather than fixed structural constraints. This way of knowledge utilisation has already been proven beneficial in other network inference studies [41, 45].

In this thesis, different types of prior knowledge were collected using various methods and resources. Published gene regulatory interactions were retrieved from the Pathway Studio database [71]. All interactions in that database have been identified with the help of automatic text mining, which represents a rapid alternative approach to comprehensive manual literature research. Interactions derived from Pathway Studio can be considered of high quality, since scientific literature is in general a reliable resource and the false positive rate is reported be about 10% [71]. Because of the few number of ultimately used regulatory interactions, the automatically found text parts were manually checked and additional information such as binding site locations were extracted from the publications. Furthermore, prediction of transcription factor binding sites (TFBS) was performed in order to obtain additional data about potential TF interactions. TFBS were predicted using the toolbox RSAT and position weight matrices (PWMs), which represent experimentally determined binding motifs of transcription factors, from the Transfac database V10 [72, 73]. For three transcription factors (SOX9, MEF2C, MSX1), which are differentially expressed during chondrogenic differentiation, corresponding PWMs were available in the database. Promoter sequences of the analysed network genes were extracted from the human genome sequence GRCh37 stored in Ensembl [74]. Since the length of the promoter or regulatory upstream region of a gene is not defined, proximal binding sites in the region of 1000 base pairs upstream from transcription start were considered. The main problem of TFBS prediction is due to the short length and heterogeneous conservation of the binding motifs, which make it difficult to distinguish between a random and a functional sequence. The toolbox RSAT addresses this issue by employing a background nucleotide distribution to assess the significance of the detected TFBS and thereby reduces the number of false positive matches. For binding site detection, a stringent p-value cutoff was applied in order to ensure a high quality of the found sites. In summary, the identification of DNA sequence sites which can be detected by transcription factors is a wide field of bioinformatic research. This thesis applied a suitable tool on genomic sequence data to generate prior knowledge for the task of network inference.

Furthermore, predictions about miRNA target genes are supplied by various resources including TargetScan, Miranda, miRBase and MirTarget2 [21, 75–77]. The associated target prediction algorithms employ diverse criteria, such as sequence complementarity and evolutionary conservation to improve their prediction quality. The enormous number of miRNA target predictions represents a tremendous data resource to initiate miRNA investigations. However, it is difficult to judge how to proceed with the heterogeneity and low specificity of the supplied data [78]. In a preliminary step, predicted interactions which were detected by just one resource were eliminated, i.e. each considered miRNA target prediction occurred in at least 2 of the 5 available databases. The resulting interaction set comprised about a fifth of the total set. MiRNAs are known for their regulatory effects by translational repression and transcript degradation [22]. The latter effect was investigated with the help of the available microarray data for both miRNAs and mRNAs. If a relation is functional, one would expect a negative correlation of the predicted pair. This assumption was confirmed by [79] who successfully identified miRNA targets by correlation. Therefore, Pearson correlation was applied in order to identify highly correlated miRNA target predictions. The selection resulted in four miRNA target gene predictions, which are most reliable according to the analysed expression data. On the one hand this series of selection criteria has dramatically limited the analysis of potential miRNAs and their associated target genes. On the other hand, this result includes only hypotheses of high reliability, which

represents meaningful prior knowledge for the network inference.

Overall, the combination of diverse prior knowledge of different reliability was the focus of this thesis to provide useful additional data for the task of network inference. Three different types of prior knowledge were applied: literature knowledge, predicted binding sites of transcription factors and miRNA target interactions. Clearly, the first source of information is most reliable as it represents experimental findings from scientific literature. Predicted interactions reflect significant findings of the prediction method which is based on previous biological findings. Even though the performance of the applied methods has been evaluated in previous publications, they are associated with a certain false positive rate. In case of predicted miRNA target genes, the number of predictions was markedly decreased by only considering consistently found interactions. Eventually, network inference provides an approach to further confirm or reject those interactions on the basis of further experimental data. Therefore, combination of heterogeneous prior knowledge represents a reasonable way to augment the process of network inference.

4.3.4 Dynamic modelling of chondrogenesis

In this thesis, network inference was applied to infer regulatory networks for hMSC which undergo chondrogenic differentiation. Two network models (MULTI-CHONDRO, MIRNA-CHONDRO) have been presented, which describe different aspects in the regulation of chondrogenesis and also have a different purpose. Both models were inferred using the NetGenerator V2.0 and are therefore based on a system of ODEs. First of all, the common network inference strategy and the related issues will be discussed. Generally, those issues also apply to any network inference task which is based on microarray data. Subsequently, this subsection focuses on the individual behaviour and purpose of each model.

In the process of network model generation, several efforts were made in order to obtain a network model of high quality including (1) the reliable gene expression estimation using an updated CDF and the normalisation of microarray data, (2) the balance between network size and available data, (3) the use of additional biological knowledge and (4) the robustness validation on the basis of data resampling. One inherent problem of network structure optimisation is the immense number of possible combinations. Particularly, the number of possible network structures grows exponentially with the number of included network nodes. The existence of many alternative network structures, which explain the underlying time series data, can hamper the NetGenerator heuristic to infer a stable network result. To avoid or at least minimise such effects, statistical and biological selection criteria were combined to identify network components (genes and miRNAs) which are most suitable for network inference. The main selection strategy involved a differential expression analysis, as discussed in section 4.2.

Additionally, genes which encode for TFs were selected in the construction of both chondrogenesis models. A functional category from Gene Ontology was used to identify those DEGs which encode for transcription factors and are characterised by sequence-specific DNA binding activity. There are at least two reasons which make transcription factor genes ideal candidates for network inference. One relates to the interpretation of inferred GRNs, which is generally difficult because network connections generally indicate rather indirect relationships between the connected nodes. In comparison, TF gene to target gene interactions can be interpreted as direct relationships in which the active TF is involved in the direct regulation of the target gene by promoter binding. Apparently, this interpretation is based on the assumption that the changed TF gene expression results in the change of TF activity. In the analysis of both models, most of the included TFs were found to be associated with hMSC differentiation in the literature, which indicates the activity of those regulators. On the other hand, active TFs might not be identified by differential expression analysis if they are primarily regulated by post-translational modifica-

tion. One example is the SOX9 antagonist RUNX2, which is not differentially expressed in any of the analysed time series, but is well-known to be active and controlled on the post-translational level [80]. Those factors should be taken into account for a more complete modelling of the underlying regulation process. There are specialised modelling strategies which address the issue of transcriptionally silent TF, for example TILAR and ExTILAR [40, 41]. Their approach implements a template network, which can include unmeasured TF nodes, and is based on specific prior knowledge about TF binding interactions. Therefore, knowledge data and gene expression data is combined in a very elegant way to extend the scope of the resulting network.

Overall, the selection of genes using statistical criteria and TF annotation resulted in the small-scale character of the presented network models, which consist of less than about twenty nodes per network. Since additional biological knowledge was available from diverse sources, this data was used in the network inference by means of soft integration and for the evaluation of the resulting network model. The last common concept applied in network inference with NetGenerator considers the computational validation of the inferred model. Generally, NetGenerator inference primarily results in one model without any information about the variance of the inferred model connections. This can be disadvantageous if the inferred connections are due to noise i.e. do not reflect a biological change in expression. Since microarray data have been employed, the appearance of noise in the data is likely. Therefore, it is important to deal with such effects to avoid a negative impact on the result. Both inferred network models were computationally validated by a resampling approach. This approach was based on the random perturbation of the input time series data using Gaussian noise. A similar approach was used by [38, 81] in their validation. The repeated resampling and model inference led to occurrence frequencies for each connection of the initial model. Those frequency values were interpreted as scores which imply a ranking of the network connections. Connections which attained higher scores can be considered more stable and therefore more reliable. With respect to experimental validation, those scores can help to select reliable hypotheses from the network, which represents the ultimate goal of network inference. From both models, hypotheses with acceptable scores and high biological plausibility were selected.

After the common steps in the inference of both models, individual features and the interpretation of both models will be discussed. The inference of the MULTI-CHONDRO model illustrated the ability of the novel NetGenerator tool to reconstruct a biologically relevant multi-stimuli model from multi-experiment data. In comparison to the computationally generated benchmark models, this dataset included microarray time series data from an induced biological process. Differences in the gene expression datasets are explained by two inputs, which correspond to the applied growth factors TGF-beta1 and BMP2. While the first dataset represents the expression levels after TGF-beta1 stimulation, the second dataset contains the effects of combined stimulation with TGF-beta1 and BMP2. Simultaneous network inference from both datasets demonstrated NetGenerator's ability to integrate multiple experimental data into one resulting model. While the TGF-beta1 input is responsible for inducing the dynamics of the first dataset as well as partially for the second dataset, the BMP2 stimulus accounts for the significant higher upregulation in the second dataset.

The focus of the model is the regulation of specific transcription factors (SOX9, MEF2C, MSX1, TRPS1, SATB2) and their activation by the two stimuli. All five were previously found to be active during hMSC differentiation [63, 82–85]. According to the time series data, they can be discriminated into early activated factors (SOX9, MSX1) and late factors (TRPS1, SATB2, MEF2C). The central purpose of the network model is to propose gene regulation hypotheses which are relevant in the process of chondrogenic differentiation. Particularly, this includes regulatory events which have an influence on the expression of the three essential chondrogenic marker genes (COL2A1, ACAN, COL10A1). With regard to the applied prior knowledge, the

network model contains connections which represent literature knowledge as well as connections which represent regulatory hypotheses that are associated with a potential transcription factor binding site. The latter type of knowledge provides valuable links in the interpretation of the model. For example, given the knowledge that SOX9 is the central regulatory factor in chondrogenesis and TRPS1 is known to promote chondrogenic differentiation via MSX1, the TFBS-supported hypothesis that SOX9 regulates TRPS1 appears to be relevant and interesting. Similarly, downregulation of SOX9 by the late transcription factor MEF2C can be interpreted as a negative feedback of the differentiation process to initiate cellular hypertrophy. MEF2C is known to control chondrogenic hypertrophy, while SOX9 is rather associated with the early stages of chondrocyte differentiation [82]. Taken together, the MULTI-CHONDRO model showed that network inference from multiple microarray datasets is feasible and that the interpretation of the resulting network leads to interesting regulatory hypotheses about the underlying biological process. Further experiments are required to verify those findings.

In contrast, the MIRNA-CHONDRO network model includes only one input stimulus, which represents the combined action of TGF-beta1 and BMP2. This is due to the provided miRNA time series data, which is available for TGF-beta1+BMP2 stimulation only. Considering the network components, the miRNA network model focused on the regulation of transcription factors by specific miRNAs. Transcription factors were previously found to be significantly over-represented targets of miRNAs [86]. Interestingly, apart from the miRNAs, all considered network genes were also part of the MULTI-CHONDRO model. However, in this model, the main objective was to integrate miRNA and mRNA expression data into a single network inference. Concerning the dynamic behaviour of the miRNAs, all four are ultimately down-regulated either by the stimulus or a transcription factor. The transcription factors SOX9, TRPS1 and MEF2C have an influence on the expression of the three chondrogenic markers (COL2A1, ACAN, COL10A1). A miRNA-dependent modulation of transcription factor gene expression which in turn changes the expression of a chondrogenesis marker gene reflects an interesting regulatory hypothesis. Previous studies have already described the impact of other miRNAs on the expression of SOX9 and TRPS1 in the context of skeletal development [87, 88].

According to the model, the stimulus-driven downregulation of miR-524-5p leads to the activation of SOX9 expression and therefore enables chondrogenic differentiation, which includes the activation of cartilage-forming genes. This model hypothesis was found to be most interesting, especially because miR-524-5p shows opposite regulation during osteogenic and adipogenic hMSC differentiation. Such behaviour indicates that miR-524-5p may be involved in the process of lineage commitment. To verify the model predictions, additional experiments were performed which overexpressed the concentration of miR-524-5p and monitored the response of the potentially affected downstream genes (SOX9, COL2A1, ACAN, COL10A1). Overall, the effect could be demonstrated by significant downregulation of the analysed genes for varying miR-524-5p concentration, whereas control experiments did not show this effect. Consequently, the integrated network inference using NetGenerator V2.0 resulted in the discovery of a so-far unknown regulation, which might play a role in the process of chondrogenesis.

4.4 Future perspectives

Consideration of data from multiple experiments which analyse multiple stimuli was found to extend interpretability and reliability of the inferred network. Apart from experiments which are based on external stimulations, there are also studies which investigate the effects of genetic perturbations, such as knock-down or knock-out of specific genes. The automatic integration of heterogeneous data including data from diverse types of perturbation experiments with NetGenerator is a promising and challenging future task. Another more apparent issue to be addressed is

the relatively time-consuming step of prior knowledge collection. Particularly, predicted binding sites of transcription factors can be stored in a database attached to NetGenerator to circumvent the repeated application of prediction tools. This would minimise the time needed for obtaining this data. Additionally, this thesis applied one tool (RSAT) for prediction of binding sites, but there are various other powerful tools available which could be combined to enhance the reliability of the detected binding sites.

With regard to the modelling of chondrogenesis, the proposed network models focus on the effect of differentially expressed transcription factors. Future modelling needs to include essential regulators whose dynamical behaviour is not reflected on the transcriptional level. To solve this issue, one could take advantage of an advanced modelling concept, which allows for integration of unmeasured network nodes. More preferable would be the application of data that measures transcription factor activity, such as protein phosphorylation measurements.

The inference of regulatory networks which include miRNA nodes to investigate miRNA target gene interactions was found to be a very promising. In the presented modelling, the selection procedure involved several criteria that resulted in very few miRNA nodes. Application of less stringent criteria can lead to more comprehensive models which provide information about general characteristics of miRNA regulation and a better understanding about how chondrogenesis is controlled by miRNAs.

Chapter 5

Summary

Application of human mesenchymal stem cells, which can be extracted from a variety of adult tissues such as the bone marrow, represents a promising approach in the field of regenerative medicine. Adult stem cells have preserved their multi-potent differentiation capability, i.e. can still differentiate into multiple cell types. Specific stimulation activates the commitment of the cells to a certain cellular lineage. Depending on the applied experimental conditions, this can give rise to chondrocytes, osteocytes or adipocytes. Investigation of the underlying biological processes which induce the observed cellular differentiation is essential to efficiently generate specific tissues for therapeutic purposes. A particular issue is the analysis of the differentiation performance upon treatment with different combination of molecules. A large-scale gene expression study was conducted to address those questions. Upon treatment with diverse stimuli, gene expression levels of cultivated human mesenchymal stem cells were monitored using time series microarray experiments for three lineages (chondro, osteo, adipo). The chosen time points covered the cellular response from the early undifferentiated stage until late terminal differentiation of the cells. On the basis of specific marker genes, differentiation progress of the analysed cells could be observed.

Analysis of microarray data involved a series of procedures including data pre-processing, gene selection and time series modelling. The initial task in pre-processing of microarray data from Affymetrix GeneChips applied an updated chip definition file. Those alternative microarray annotations claim to improve the estimation of gene expression by redefinition of the probe-sets. To evaluate the performance of four probe-set annotations, pre-processed microarray data were compared to data from qRT-PCR experiments. As a result, one annotation was found to perform best and was therefore used in this work. In the subsequent analysis, differentially expressed genes were identified, which are regulated during the differentiation process and helped to distinguish differentiation stages. Furthermore, expression of differentiation regulators and marker genes was found to be accelerated in case of the most efficient experimental stimulation. Further analysis aimed at a more precise understanding of the underlying gene regulation.

Application of gene network inference is a common approach to identify the regulatory dependencies among a set of investigated genes. The general aim is to understand the complex behaviour of gene regulation and how this gives rise to the observed expression changes. There are various network inference approaches which apply different mathematical concepts to learn the network structure from experimental data. Frequently, such reconstruction of the underlying network structure is based on microarray data. This thesis applies the NetGenerator V2.0 tool, which is capable to deal with multiple time series data, which investigates the effect of multiple external stimuli. In a system biological view, those data are based on the same biological system under different environmental conditions. The applied model is based on a system of linear ordinary differential equations. Model parameters are optimised to reproduce the given time

series datasets and to describe the underlying gene regulatory network. Since this model optimisation represents a combinatorial challenge, NetGenerator applies a heuristic search strategy to identify the optimal model parameters. Several procedures in the inference process, including pre-processing, optimisation and visualisation, were adapted in this new version in order to allow for the integration of multiple datasets. The inference result is a single network which contains an individual input node for each of the associated experimental stimuli and therefore allows for cross-talk analysis among the inputs.

Three different cross-talk benchmark examples were generated *in silico* in order to evaluate the performance of the NetGenerator V2.0 tool. Inference of the examples was accomplished with high accuracy, even though the applied time series were sparsely sampled and perturbed with artificial noise. In one case, network inference quality improved due to the application of multiple datasets. Afterwards, network inference was applied on the given multi-experiment microarray data of mesenchymal stem cells, which are stimulated to differentiate towards chondrocytes. This task involved the selection of an adequate number of network nodes and the collection of additional biological knowledge about known and predicted regulatory interactions among the components. Since the underlying regulatory network is unknown, the resulting chondrogenesis model was evaluated on the basis of several features including the model adaptation to the data, total number of connections, proportion of connections associated with prior knowledge and the model stability in a resampling procedure. Altogether, the proposed chondrogenesis model was found to have a high quality and includes interesting hypotheses about specific gene regulation in the differentiation process. Generally, NetGenerator V2.0 has provided an automatic and efficient way to integrate experimental datasets and to enhance the interpretability and reliability of the resulting network.

In a second chondrogenesis model, the influence of miRNAs on the differentiation process was investigated. On the basis of miRNA microarray data, this work demonstrated the integration of mRNA and miRNA time series expression data for the purpose of network inference. An initial challenge was to identify the most interesting miRNAs from the microarray dataset. A series of procedures was applied including differential expression analysis, functional annotation, miRNA target gene prediction and miRNA-mRNA time series correlation. The resulting network nodes included four miRNAs, their predicted target genes and chondrogenic differentiation marker genes. Similarly to the multi-stimuli chondrogenesis model, evaluation was based on the quality of the data reproduction, the number of total / prior knowledge connections and stability. In a subsequent validation experiment, one hypothesis of the model was verified by overexpression experiments, which demonstrated the negative effect of miR-524-5p on downstream genes including the transcription factor SOX9 and the marker genes COL2A1, COL10A1 and ACAN.

Chapter 6

Zusammenfassung

Die Verwendung von humanen mesenchymalen Stammzellen, welche aus einer Vielzahl von adulten Geweben, zum Beispiel dem Knochenmark, gewonnen werden können, ist ein vielversprechender Ansatz im Gebiet der regenerativen Medizin. Adulte Stammzellen haben ihre Fähigkeit zur multipotenten Differenzierung erhalten, d.h. sie können in verschiedene Zelltypen differenzieren. Die Stimulation mit spezifischen Molekülen aktiviert den Differenzierungsprozess der Zellen in Richtung einer bestimmten Zelllinie. Abhängig von den angewandten experimentellen Bedingungen, können Chondrozyten, Osteozyten oder Adipozyten entstehen. Die Untersuchung der ablaufenden biologischen Prozesse, welche die beobachtete Zelldifferenzierung bewirken, ist entscheidend um spezielle Gewebetypen für therapeutische Zwecke effizient herzustellen. Ein spezielles Problem ist die Analyse der Differenzierungsleistung nach Behandlung mit verschiedenen Kombinationen von Molekülen. Eine umfangreiche Untersuchung der Genexpression wurde durchgeführt, um diese Fragestellungen zu untersuchen. Genexpression wurde nach unterschiedlicher Behandlung von kultivierten humanen mesenchymalen Stammzellen gemessen. Es wurden Zeitreihenexperimente mit Mikroarrays für drei Zelllinien (Chondro, Osteo und Adipo) durchgeführt. Die gewählten Zeitpunkte umfassten die zelluläre Antwort vom frühen undifferenzierten Stadium bis zur späten terminalen Differenzierung der Zellen. Mit Hilfe von spezifischen Markergenen konnte der Differenzierungsverlauf der untersuchten Zellen beobachtet werden.

Die Untersuchung der erzeugten Mikroarraydaten umfasste eine Reihe von Analyseprozeduren, inklusive Datenvorverarbeitung, Genselektion und Zeitreihenmodellierung. Der erste Schritt in der Vorverarbeitung von Mikroarraydaten von Affymetrix GeneChips beinhaltete die Verwendung einer aktualisierten “Chip Definition File”. Diese alternativen Annotationsdateien für Mikroarrays behaupten die Genexpressionsschätzung durch eine Neudefinition der Probesets zu verbessern. Für die Evaluation von vier verschiedenen Probesetannotationen wurden vorverarbeitete Mikroarraydaten mit Daten von qRT-PCR Experimenten verglichen. Eine der untersuchten Annotationen erreichte die besten Ergebnisse und wurde deshalb in dieser Arbeit verwendet. In der folgenden Untersuchung wurden differentiell exprimierte Gene, welche während des Differenzierungsprozesses reguliert wurden, identifiziert und unterstützten die Unterscheidung von verschiedenen Differenzierungsstadien. Weiterhin wurde entdeckt, dass die Expression von Differenzierungsregulatoren und Markergenen im Falle der wirkungsvollsten experimentellen Stimulierung beschleunigt wurde. Weitere Untersuchungen zielten auf ein genaueres Verständnis der zugrunde liegenden Genregulation ab.

Die Inferenz von genregulatorischen Netzwerken ist ein verbreiteter Ansatz um regulatorische Abhängigkeiten zwischen einer Menge von untersuchten Genen zu finden. Das generelle Ziel ist das komplexe Verhalten der Genregulation zu verstehen und herauszufinden wie dieses zu den beobachteten Expressionsänderungen führt. Es gibt zahlreiche Ansätze zur Netzwerkinferenz, welche verschiedene mathematische Konzepte einsetzen um die Netzwerkstruktur aus experimentellen

Daten zu lernen. Häufig basiert diese Rekonstruktion des Netzwerkes auf Mikroarraydaten. Diese Doktorarbeit nutzt erstmalig das Programm NetGenerator V2.0, welches in der Lage ist mit multiplen experimentellen Daten umzugehen, welche den Effekt von multiplen externen Stimulierungen untersuchen. Aus systembiologischer Sicht representieren solche Daten dasselbe biologische System, unter verschiedenen Umgebungsbedingungen. Das verwendete Modell basiert auf einem System von linearen gewöhnlichen Differentialgleichungen. Modellparameter werden optimiert um die vorliegenden Zeitreihendaten nachzubilden und um das zugrunde liegende genregulatorische Netzwerk zu beschreiben. Da diese Modelloptimierung ein kombinatorisches Problem darstellt, verwendet NetGenerator eine heuristische Suchstrategie um die optimalen Modellparameter zu identifizieren. Mehrere Prozeduren des Inferenzprozesses einschließlich Vorverarbeitung, Optimierung und Visualisierung wurden in dieser neuen Version angepasst um die Integration von mehreren Datensätzen zu ermöglichen. Das Ergebnis der Inferenz ist ein Netzwerk, welches für jeden assoziierten experimentellen Stimulus einen Eingangsknoten enthält und daher eine Crosstalk-Analyse zwischen den Eingangsknoten ermöglicht.

Drei verschiedene Crosstalk-Beispiele wurden *in silico* generiert um die Leistung von NetGenerator V2.0 zu evaluieren. Diese Beispielnetzwerke wurden mit hoher Genauigkeit rekonstruiert, obwohl die verwendeten Zeitreihen wenig Datenpunkte enthielten und mit künstlichem Rauschen gestört waren. In einem Fall konnte gezeigt werden, dass sich die Qualität der Netzwerkinferenz durch die Verwendung von mehreren Datensätzen verbesserte. Danach wurde Netzwerkinferenz auf die vorliegenden Mikroarraydaten von mesenchymalen Stammzellen, welche zur chondrogenen Differenzierung stimuliert wurden, angewandt. Diese Aufgabe umfasste die Auswahl einer angemessenen Zahl von Netzwerkknoten und die Zusammentragung von biologischem Wissen über regulatorische Interaktionen zwischen den Komponenten. Da das zugrunde liegende regulatorische Netzwerk unbekannt ist, wurde das resultierende Chondrogenese-Modell auf der Basis von verschiedenen Merkmalen bewertet: Modellanpassung an die Daten, Gesamtanzahl der Verbindungen, Anteil der Verbindungen welche Vorwissen widerspiegeln und die Modellstabilität mittels Resampling. Zusammenfassend hat das vorgestellte Chondrogenese-Modell eine hohe Qualität und beinhaltet interessante Hypothesen über die spezifische Genregulation im Differenzierungsprozess.

In einem zweiten Chondrogenese-Modell wurde der Einfluss von miRNAs auf den Differenzierungsprozess untersucht. Auf der Grundlage von miRNA Mikroarraydaten konnte diese Arbeit die Anwendung von Netzwerkinferenz auf kombinierte mRNA und miRNA Zeitreihendaten zeigen. Eine initiale Herausforderung war die Identifizierung der interessantesten miRNAs aus dem Mikroarraydatensatz. Eine Reihe von Prozeduren wurde angewandt, einschließlich einer Analyse der differentiellen Expression, funktioneller Annotation, Vorhersage von miRNA Zielgenen und Korrelation von miRNA-mRNA Zeitreihen. Die resultierenden Netzwerkknoten enthielten vier miRNAs, ihre vorhergesagten Zielgene und Markergene der chondrogenen Differenzierung. Wie beim Chondrogenese-Modell mit multiplen Eingangsknoten basierte die Bewertung des Modells auf der Qualität der Datenwiedergabe, der Kantengesamtanzahl, der Anzahl von Vorwissenskanten und der Stabilität. In einem anschließenden Validierungsexperiment konnte eine Modellhypothese mittels Überexpression verifiziert werden. Die Experimente zeigten den negativen Effekt einer miRNA (miR-524-5p) auf Zielgene, einschließlich den Transkriptionsfaktor SOX9 und die Markergene COL2A1, COL10A1 und ACAN.

Bibliography

[1] International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860–921.

[2] Venter JC, et al.: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304–1351.

[3] International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931–945.

[4] The ENCODE Project Consortium: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799–816, [http://dx.doi.org/10.1038/nature05874].

[5] Velculescu VE, Madden SL, Zhang L, Lash AE, Yu J, Rago C, Lal A, Wang CJ, Beaudry GA, Ciriello KM, Cook BP, Dufault MR, Ferguson AT, Gao Y, He TC, Hermeking H, Hiraldo SK, Hwang PM, Lopez MA, Luderer HF, Mathews B, Petroziello JM, Polyak K, Zawel L, Kinzler KW: **Analysis of human transcriptomes**. *Nat. Genet.* 1999, **23**(4):387–388.

[6] Butte A: **The use and analysis of microarray data.** *Nat Rev Drug Discov* 2002, **1**(12):951–960, [http://dx.doi.org/10.1038/nrd961].

[7] Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG: **The affymetrix GeneChip platform: an overview.** *Methods Enzymol* 2006, **410**:3–28, [http://dx.doi.org/10.1016/S0076-6879(06)10001-4].

[8] Stoughton RB: **Applications of DNA microarrays in biology.** *Annu Rev Biochem* 2005, **74**:53–82, [http://dx.doi.org/10.1146/annurev.biochem.74.082803.133212].

[9] Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays**. *Nat. Genet.* 1999, **21**(1 Suppl):33–37.

[10] Hoheisel JD: **Microarray technology: beyond transcript profiling and genotype analysis**. *Nat. Rev. Genet.* 2006, **7**(3):200–210.

[11] Dupuy A, Simon RM: **Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting.** *J Natl Cancer Inst* 2007, **99**(2):147–157, [http://dx.doi.org/10.1093/jnci/djk018].

[12] Ideker TE, Thorsson V, Karp RM: **Discovery of regulatory interactions through perturbation: inference and experimental design**. *Pac Symp Biocomput* 2000, :305–316.

[13] Markowetz F, Spang R: **Inferring cellular networks–a review.** *BMC Bioinformatics* 2007, **8 Suppl 6**:S5, [http://dx.doi.org/10.1186/1471-2105-8-S6-S5].

[14] Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R: **Gene regulatory network inference: data integration in dynamic models-a review**. *BioSystems* 2009, **96**:86–103.

[15] Mischel PS, Cloughesy TF, Nelson SF: **DNA-microarray analysis of brain cancer: molecular classification for therapy**. *Nat. Rev. Neurosci.* 2004, **5**(10):782–792.

[16] Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR: **Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling**. *Cancer Cell* 2002, **1**(2):133–143.

[17] Huber R, Hummert C, Gausmann U, Pohlers D, Koczan D, Guthke R, Kinne RW: **Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane**. *Arthritis Res. Ther.* 2008, **10**(4):R98.

[18] Trost Z, Trebse R, Prezelj J, Komadina R, Logar DB, Marc J: **A microarray based identification of osteoporosis-related genes in primary culture of human osteoblasts**. *Bone* 2010, **46**:72–80.

[19] Dietel M, Sers C: **Personalized medicine and development of targeted therapies: The upcoming challenge for diagnostic molecular pathology. A review**. *Virchows Arch.* 2006, **448**(6):744–755.

[20] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data**. *Nucleic Acids Res.* 2003, **31**(4):e15.

[21] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics**. *Nucleic Acids Res.* 2008, **36**(Database issue):D154–158.

[22] Guo H, Ingolia NT, Weissman JS, Bartel DP: **Mammalian microRNAs predominantly act to decrease target mRNA levels**. *Nature* 2010, **466**(7308):835–840.

[23] Pothof J, Verkaik NS, van IJcken W, Wiemer EA, Ta VT, van der Horst GT, Jaspers NG, van Gent DC, Hoeijmakers JH, Persengiev SP: **MicroRNA-mediated gene silencing modulates the UV-induced DNA-damage response**. *EMBO J.* 2009, **28**(14):2090–2099.

[24] Tu Y, Stolovitzky G, Klein U: **Quantitative noise analysis for gene expression microarray experiments.** *Proc Natl Acad Sci U S A* 2002, **99**(22):14031–14036, [http://dx.doi.org/10.1073/pnas.222164199].

[25] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol.* 2004, **5**(10):R80.

[26] Pepper SD, Saunders EK, Edwards LE, Wilson CL, Miller CJ: **The utility of MAS5 expression summary and detection call algorithms**. *BMC Bioinformatics* 2007, **8**:273.

[27] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**:249–264.

[28] Gardner TS, Faith JJ: **Reverse-engineering transcription control networks.** *Phys Life Rev* 2005, **2**:65–88, [http://dx.doi.org/10.1016/j.plrev.2005.01.001].

[29] Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23**(19):2507–2517, [http://dx.doi.org/10.1093/bioinformatics/btm344].

[30] Stekel D: *Microarray Bioinformatics*. Cambridge University Press 2003.

[31] Cui X, Churchill GA: **Statistical tests for differential expression in cDNA microarray experiments.** *Genome Biol* 2003, **4**(4):210.

[32] Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**:Article3.

[33] Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2001, **98**(9):5116–5121, [http://dx.doi.org/10.1073/pnas.091062498].

[34] Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**:289–300, [http://dx.doi.org/10.2307/2346101].

[35] Gillespie CS, Lei G, Boys RJ, Greenall A, Wilkinson DJ: **Analysing time course microarray data using Bioconductor: a case study using yeast2 Affymetrix arrays**. *BMC Res Notes* 2010, **3**:81.

[36] MAQC Consortium: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements**. *Nat. Biotechnol.* 2006, **24**(9):1151–1161.

[37] Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM: **Protein interaction networks from yeast to human.** *Curr Opin Struct Biol* 2004, **14**(3):292–299, [http://dx.doi.org/10.1016/j.sbi.2004.05.003].

[38] Guthke R, Möller U, Hoffmann M, Thies F, Töpfer S: **Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection**. *Bioinformatics* 2005, **21**:1626–1634.

[39] Tibshirani R: **Regression shrinkage and selection via the lasso**. *J. Roy. Statist. Soc. Ser. B* 1996, **58**:267–288, [http://www.ams.org/mathscinet-getitem?mr=1379242].

[40] Hecker M, Goertsches RH, Engelmann R, Thiesen HJ, Guthke R: **Integrative modeling of transcriptional regulation in response to antirheumatic therapy**. *BMC Bioinformatics* 2009, **10**:262.

[41] Vlaic S, Schmidt-Heck W, Matz-Soja M, Marbach E, Linde J, Meyer-Baese A, Zellmer S, Guthke R, Gebhardt R: **The Extended TILAR Approach: A novel tool for dynamic modeling of the transcription factor network regulating the adaption to in vitro cultivation of murine hepatocytes**. *BMC Syst Biol* 2012, **6**:147.

[42] Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7**(3-4):601–620, [http://dx.doi.org/10.1089/106652700750050961].

[43] Liang S, Fuhrman S, Somogyi R: **Reveal, a general reverse engineering algorithm for inference of genetic network architectures.** *Pac Symp Biocomput* 1998, :18–29.

[44] Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302**(5643):249–255, [http://dx.doi.org/10.1126/science.1087447].

[45] Linde J, Wilson D, Hube B, Guthke R: **Regulatory network modelling of iron acquisition by a fungal pathogen in contact with epithelial cells**. *BMC Syst Biol* 2010, **4**:148.

[46] Toepfer S, Guthke R, Driesch D, Woetzel D, Pfaff M: **The NetGenerator algorithm: reconstruction of gene regulatory networks.** 2007, **4366**:119–130.

[47] Weber M, Henkel SG, Vlaic S, Guthke R, van Zoelen EJ, Driesch D: **Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0**. *BMC Syst Biol* 2013, **7**:1.

[48] Tegner J, Yeung MKS, Hasty J, Collins JJ: **Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling.** *Proc Natl Acad Sci U S A* 2003, **100**(10):5944–5949, [http://dx.doi.org/10.1073/pnas.0933416100].

[49] Wollbold J, Huber R, Pohlers D, Koczan D, Guthke R, Kinne RW, Gausmann U: **Adapted Boolean network models for extracellular matrix formation.** *BMC Syst Biol* 2009, **3**:77, [http://dx.doi.org/10.1186/1752-0509-3-77].

[50] Altwasser R, Linde J, Buyko E, Hahn U, Guthke R: **Genome-Wide Scale-Free Network Inference for Candida albicans.** *Front Microbiol* 2012, **3**:51, [http://dx.doi.org/10.3389/fmicb.2012.00051].

[51] Wang Y, Joshi T, Zhang XS, Xu D, Chen L: **Inferring gene regulatory networks from multiple microarray datasets.** *Bioinformatics* 2006, **22**(19):2413–2420, [http://dx.doi.org/10.1093/bioinformatics/btl396].

[52] Gupta R, Stincone A, Antczak P, Durant S, Bicknell R, Bikfalvi A, Falciani F: **A computational framework for gene regulatory network inference that combines multiple methods and datasets.** *BMC Syst Biol* 2011, **5**:52, [http://dx.doi.org/10.1186/1752-0509-5-52].

[53] Bisognin A, Sales G, Coppe A, Bortoluzzi S, Romualdi C: **MAGIA^2: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update)**. *Nucleic Acids Res.* 2012, **40**(Web Server issue):13–21.

[54] Nam S, Li M, Choi K, Balch C, Kim S, Nephew KP: **MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression**. *Nucleic Acids Res.* 2009, **37**(Web Server issue):W356–362.

[55] Huang GT, Athanassiou C, Benos PV: **mirConnX: condition-specific mRNA-microRNA network integrator**. *Nucleic Acids Res.* 2011, **39**(Web Server issue):W416–423.

[56] Uccelli A, Moretta L, Pistoia V: **Mesenchymal stem cells in health and disease**. *Nat. Rev. Immunol.* 2008, **8**(9):726–736.

[57] Caplan AI: **Adult mesenchymal stem cells for tissue engineering versus regenerative medicine**. *J. Cell. Physiol.* 2007, **213**(2):341–347.

[58] Pittenger M, Vanguri P, Simonetti D, Young R: **Adult mesenchymal stem cells: potential for muscle and tendon regeneration and use in gene therapy**. *J Musculoskelet Neuronal Interact* 2002, **2**(4):309–320.

[59] Friedenstein AJ, Chailakhyan RK, Latsinik NV, Panasyuk AF, Keiliss-Borok IV: **Stromal cells responsible for transferring the microenvironment of the hemopoietic tissues. Cloning in vitro and retransplantation in vivo.** *Transplantation* 1974, **17**(4):331–340.

[60] Augello A, De Bari C: **The regulation of differentiation in mesenchymal stem cells.** *Hum Gene Ther* 2010, **21**(10):1226–1238, [http://dx.doi.org/10.1089/hum.2010.173].

[61] Zhang W, Yang N, Shi XM: **Regulation of mesenchymal stem cell osteogenic differentiation by glucocorticoid-induced leucine zipper (GILZ).** *J Biol Chem* 2008, **283**(8):4723–4729, [http://dx.doi.org/10.1074/jbc.M704147200].

[62] Schittler D, Hasenauer J, Allgower F, Waldherr S: **Cell differentiation modeled via a coupled two-switch regulatory network**. *Chaos* 2010, **20**(4):045121.

[63] Hartmann C: **Transcriptional networks controlling skeletal development**. *Curr. Opin. Genet. Dev.* 2009, **19**(5):437–443.

[64] Piek E, Sleumer LS, van Someren EP, Heuver L, de Haan JR, de Grijs I, Gilissen C, Hendriks JM, van Ravestein-van Os RI, Bauerschmidt S, Dechering KJ, van Zoelen EJ: **Osteo-transcriptomics of human mesenchymal stem cells: accelerated gene expression and osteoblast differentiation induced by vitamin D reveals c-MYC as an enhancer of BMP2-induced osteogenesis**. *Bone* 2010, **46**(3):613–627.

[65] Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, Shmoish M, Ferrari S, Lancet D, Danieli GA, Bicciato S: **Novel definition files for human GeneChips based on GeneAnnot**. *BMC Bioinformatics* 2007, **8**:446.

[66] Canales RD, Luo Y, Willey JC, Austermiller B, Barbacioru CC, Boysen C, Hunkapiller K, Jensen RV, Knight CR, Lee KY, Ma Y, Maqsodi B, Papallo A, Peters EH, Poulter K, Ruppel PL, Samaha RR, Shi L, Yang W, Zhang L, Goodsaid FM: **Evaluation of DNA microarray results with quantitative gene expression platforms.** *Nat Biotechnol* 2006, **24**(9):1115–1122, [http://dx.doi.org/10.1038/nbt1236].

[67] Morey JS, Ryan JC, Van Dolah FM: **Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR.** *Biol Proced Online* 2006, **8**:175–193, [http://dx.doi.org/10.1251/bpo126].

[68] van der Kraan PM, Blaney Davidson EN, Blom A, van den Berg WB: **TGF-beta signaling in chondrocyte terminal differentiation and osteoarthritis: modulation and integration of signaling pathways through receptor-Smads**. *Osteoarthr. Cartil.* 2009, **17**(12):1539–1545.

[69] Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, Shmoish M, Ferrari S, Lancet D, Danieli GA, Bicciato S: *gahgu133a.db: Genecards derived annotations for gahgu133a custom probeset definitions.* [R package version 2.2.0].

[70] Stolovitzky G, Monroe D, Califano A: **Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference.** *Ann N Y Acad Sci* 2007, **1115**:1–22, [http://dx.doi.org/10.1196/annals.1407.021].

[71] Yuryev A, Mulyukov Z, Kotelnikova E, Maslov S, Egorov S, Nikitin A, Daraselia N, Mazo I: **Automatic pathway building in biological association networks**. *BMC Bioinformatics* 2006, **7**:171.

[72] Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J: **RSAT 2011: regulatory sequence analysis tools.** *Nucleic Acids Res* 2011, **39**(Web Server issue):W86–W91, [http://dx.doi.org/10.1093/nar/gkr377].

[73] Wingender E: **The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation**. *Brief. Bioinformatics* 2008, **9**(4):326–332.

[74] Flicek P, et al.: **Ensembl 2013**. *Nucleic Acids Res.* 2013, **41**(Database issue):48–55.

[75] Friedman RC, Farh KK, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome Res.* 2009, **19**:92–105.

[76] Betel D, Wilson M, Gabow A, Marks DS, Sander C: **The microRNA.org resource: targets and expression**. *Nucleic Acids Res.* 2008, **36**(Database issue):D149–153.

[77] Wang X, El Naqa IM: **Prediction of both conserved and nonconserved microRNA targets in animals**. *Bioinformatics* 2008, **24**(3):325–332.

[78] Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG: **Lost in translation: an assessment and perspective for computational microRNA target identification**. *Bioinformatics* 2009, **25**(23):3049–3055.

[79] Liu T, Papagiannakopoulos T, Puskar K, Qi S, Santiago F, Clay W, Lao K, Lee Y, Nelson SF, Kornblum HI, Doyle F, Petzold L, Shraiman B, Kosik KS: **Detection of a microRNA signal in an in vivo expression set of mRNAs**. *PLoS ONE* 2007, **2**(8):e804.

[80] Sotoca AM, Weber M, van Zoelen E: **Gene Expression Regulation underlying Osteo-, Adipo-, and Chondro-Genic Lineage Commitment of Human Mesenchymal Stem Cells**. *Medical Advancements in Aging and Regenerative Technologies: Clinical Tools and Applications* 2013, **7 Web**:226–94.

[81] D'haeseleer P, Liang S, Somogyi R: **Genetic network inference: from co-expression clustering to reverse engineering.** *Bioinformatics* 2000, **16**(8):707–726.

[82] Leung VY, Gao B, Leung KK, Melhado IG, Wynn SL, Au TY, Dung NW, Lau JY, Mak AC, Chan D, Cheah KS: **SOX9 governs differentiation stage-specific gene expression in growth plate chondrocytes via direct concomitant transactivation and repression**. *PLoS Genet.* 2011, **7**(11):e1002356.

[83] Arnold MA, Kim Y, Czubryt MP, Phan D, McAnally J, Qi X, Shelton JM, Richardson JA, Bassel-Duby R, Olson EN: **MEF2C transcription factor controls chondrocyte hypertrophy and bone development**. *Dev. Cell* 2007, **12**(3):377–389.

[84] Huang Z, Xu H, Sandell L: **Negative regulation of chondrocyte differentiation by transcription factor AP-2alpha.** *J Bone Miner Res* 2004, **19**(2):245–255, [`http://dx.doi.org/10.1359/jbmr.2004.19.2.245`].

[85] Wuelling M, Kaiser FJ, Buelens LA, Braunholz D, Shivdasani RA, Depping R, Vortkamp A: **Trps1, a regulator of chondrocyte proliferation and differentiation, interacts with the activator form of Gli3**. *Dev. Biol.* 2009, **328**:40–53.

[86] Cui Q, Yu Z, Purisima EO, Wang E: **Principles of microRNA regulation of a human cellular signaling network**. *Mol. Syst. Biol.* 2006, **2**:46.

[87] Yang B, Guo H, Zhang Y, Chen L, Ying D, Dong S: **MicroRNA-145 regulates chondrogenic differentiation of mesenchymal stem cells by targeting Sox9**. *PLoS ONE* 2011, **6**(7):e21679.

[88] Zhang Y, Xie RL, Gordon J, LeBlanc K, Stein JL, Lian JB, van Wijnen AJ, Stein GS: **Control of mesenchymal lineage progression by microRNAs targeting skeletal gene regulators Trps1 and Runx2**. *J. Biol. Chem.* 2012, **287**(26):21926–21935.

Abbreviations

CDF	Chip Definition File
DEG	Differentially Expressed Gene
FCT	Full Cross-Talk
FDR	False Discovery Rate
GEO	Gene Expression Omnibus
GRN	Gene Regulatory Network
hMSC	Human Mesenchymal Stem Cells
LCT	Limited Cross-Talk
LIMMA	Linear Models for MicroArray data
LNA	Locked Nucleic Acid
MAS5	MicroArray Suite 5.0
miRNA	microRNA
NCT	No Cross-Talk
ODE	Ordinary Differential Equation
PWM	Position Weight Matrix
qRT-PCR	quantitative Real-Time Polymerase Chain Reaction
RMA	Robust Multichip Average
RNA	RiboNucleic Acid
RSAT	Regulatory Sequence Analysis Tools
SAM	Significance Analysis of Microarrays
TF	Transcription Factor
TFBS	Transcription Factor Start Site

Danksagung

Ich möchte mich bei allen Personen bedanken, die mich in den letzten drei Jahren unterstützt haben. Meinem Betreuer Prof. Dr. Reinhard Guthke danke ich für die Möglichkeit in einem europaweiten Systembiologieprojekt zu arbeiten und für die kontinuierliche Beratung und die Hinweise während meiner Arbeit. Meinen beiden Bürogenossen Sebastian Vlaic und Peter Kupfer danke ich für viele inspirierende Gespräche und ihr musikalisches Feingespür. Einige zentrale Ideen meiner Arbeit sind durch unsere gemeinsame Diskussion entfaltet wurden und in unserem Büro voll ausgereift.
Mein Dank gilt auch der gesamten SBI Arbeitsgruppe für die gemeinsame Zeit und unbezahlbare Motivation durch frischen Kaffee. Ein besonderer Dank geht an Sebastian Henkel von der BioControl Jena GmbH, der meine Arbeit entscheidend vorangebracht hat und durch seine Ausdauer und Teamarbeit einen großen Beitrag zu meiner Doktorarbeit geleistet hat.

Im Rahmen des ErasysBio-Projektes hatte ich die Möglichkeit mit Personen aus vier Städten Europas zusammenzuarbeiten, welche zur Hälfte einen bioinformatischen und biologischen Hintergrund hatten. Daran beteiligt waren Prof. Dr. Raimund W. Kinne und Dr. Franziska Dees. Beiden möchte ich für die zahlreichen Gruppengespräche in unserem Büro danken und ihr Bemühen unsere Ergebnisse im biologischem Sinne zu interpretieren. I am grateful to Nil Turan and Ana Sotoca for our common work and the patience they had in our transnational cooperation, which has been mainly based on Skype calls. Thanks to Prof. Dr. Francesco Falciani and Prof. Dr. Joop van Zoelen for their advice, the helpful discussions and the shared time during our meetings.

I am also grateful to Prof. Dr. Thiesen and the opportunity to work one month in the Shanghai SIBS institute. This was an extraordinary experience of Chinese science and culture. I had wonderful colleagues and I enjoyed the meals, meetings and the friendly atmosphere in the lab. Particulary, I want to thank Dong, Xiao for the nice time in Shanghai, Munich and Jena.

Ich möchte meiner Familie dafür danken, dass sie während der gesamten Zeit für mich da war und mir dadurch viel Kraft und Ausdauer geschenkt hat. Meinen Freunden danke ich für die gemeinsame Zeit, den Zusammenhalt und den Austausch über teils weite Distanzen. Diese Zeit hat mir einiges an Motivation und Realität zurückgegeben um meine Doktorarbeit letztendlich zu vollenden.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass mir die geltende Promotionsordnung der biologisch-pharmazeutischen Fakultät bekannt ist und ich mich mit bestem Wissen an diese Ordnung gehalten habe. Die vorliegende Dissertation habe ich selbständig und nur unter Verwendung der angegebenen Hilfsmittel, Daten und Quellen angefertigt. Unterstützung während meiner wissenschaftlichen Arbeit und zur Erstellung des vorliegenden Dissertationstextes habe ich nur von den genannten Co-Autoren und in der Danksagung genannten Personen erhalten. Ich habe keine Hilfe von externen Vermittlungs- oder Beratungsdiensten in Anspruch genommen. Niemand hat mittelbare oder unmittelbare geldwerte Leistungen erhalten, für Arbeiten die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die vorgelegte Dissertation wurde bisher nicht als Prüfungsarbeit für eine andere wissenschaftliche Prüfung eingereicht. Im Speziellen habe ich sie an keiner anderen Hochschule eingereicht, um einen akademischen Grad zu erhalten.

_____