Systems Biology Analysis of Large-Scale Gene Expression Data

Dissertation

zur Erlangung des Grades

„Doktor

der Naturwissenschaften“

am Fachbereich Biologie

der Johannes Gutenberg-Universität

in Mainz

Kolja Becker

geb. in Kettering, Ohio, USA
Mainz, den 08.05.2018

Dekan:
1. Berichterstatter:
2. Berichterstatter:

Tag der mündlichen Prüfung: 19.10.2018

Contents

Abstract

Dynamics of gene expression in the context of gene regulatory networks are key to our understanding of cellular function. Particular with the advent of genome wide measurement of mRNA and protein abundances, large-scale gene expression data to investigate gene expression and gene expression networks are made available. Systems biology analysis provides means of extracting relevant information from this data and to further improve our quantitative understanding of mRNA transcription and consecutive translation into protein.

In this thesis a variety of biological topics related to transcriptional and translational gene regulation are addressed. Topics range from the identification of circadian expressed genes in the context of circadian rhythm, prediction of transcriptional regulator-target interactions from time-course gene expression data, dynamic modelling of the gene regulatory network coordinating the epithelial-to-mesenchymal transition, and the identification of post-transcriptionally regulated genes during *Drosophila* embryogenesis.

In each case study, collected large-scale gene expression data serves as the basis for computational analysis using a combination of different pre-existing as well as newly formulated methods. Whenever feasible, the performance of computational methods is evaluated and an experimental validation of predictions is pursued. As a result, the detailed computational analysis of large-scale gene expression data performed in this study not only provides valuable insight into the biological problem at hand, but further offers the opportunity for the development of systems biology tools and their evaluation under realistic experimental conditions.

Chapter 1

Introduction

1.1 Differential gene expression in the context of gene regulatory networks

Each somatic cell of the body carries the exact identical genomic information imprinted in its DNA sequence. DNA polymerase dependent gene transcription permits the read-out of this genomic information by producing RNA transcripts. RNA transcripts in turn are altered by chemical modification as well as alternative splicing before being exported from the cell nucleus. Outside of the cell nucleus, ribosomes further translate processed mRNA into their respective amino acid sequences, which may fold into functional three-dimensional protein structures, controlling for most of the cells structural and functional properties.

At each step of gene expression, the processing of DNA into RNA and protein can be differentially controlled by a variety of mechanisms. Epigenetic regulators for example control the accessibility of genomic regions by reshaping the three-dimensional chromatin structure of the DNA. Binding of transcription factors to cis-regulatory elements, such as enhancer regions, modulates DNA polymerase dependent transcription of DNA into RNA. At the level of RNA, splicing factors control which parts of mRNA transcripts are retained and further translated into protein. In addition, various post-translational protein modifications are known, impacting the final form of a protein.

The self-referential property of gene expression, in which genomic information is processed into protein, which in turn can affect the read-out of genomic information, allows for the formation of complex gene regulatory networks determinant for a cells gene expression state. Since a cells behaviour is to a large extent controlled by its mRNA and protein content, the understanding of gene expression dynamics and gene regulatory networks is therefore key to our understanding of cellular function itself.

Recently the development of high-throughput methods have allowed for the in-depth investigation of gene expression dynamics, as well as reconstruction of gene regulatory networks from genome wide gene expression data. By RNA-Sequencing the expression of thousands of genes can be assessed simultaneously. Similarly, mass-spectrometry allows for the global characterization of proteins in a given sample. With regard to this large-scale gene expression data, systems biology provides a rigorous and quantitative

framework to analyse and understand dynamics of mRNA and protein expression in the context of gene regulatory networks.

1.2 Systems biology analysis of gene expression and gene regulatory networks

At the core of systems biology lie mathematical models, describing the dynamical interplay of the various elements in a biological system. One of the earliest mathematical models of gene regulation was used to analyse oscillatory behaviour of a closed transcription/translation feedback loop, in which translated protein inhibits the production of its own mRNA [Goodwin, 1965]. Since then, numerous mathematical models of genetic and molecular circuits have been developed, providing critical and non-intuitive insight into the inner workings of a cell (reviewed in [Karlebach and Shamir, 2008, Ay and Arnosti, 2011]).

In mathematical models of gene regulation, the expression state of genes on the level of mRNA or protein is typically encoded by the models state variables. Mathematical expressions referring to these state variables then describe the interdependency of gene expression states. Often system equations further include kinetic parameters detailing production or degradation rates of gene products. Aside from internal state variables and model parameters, external inputs, relating for example to the concentration of signalling molecules, may impact the dynamics of the system. Finally, in order to compare a model to experimental data, the mathematical relationship between the internal state variables of a model and experimental observables must be defined.

Although most mathematical models are formulated according to the general features outlined above, models may vary in their degree of detail. One basic approach to model biological systems is the use of boolean models, in which the state of genes in the network are described using binary values. In boolean models, a state variable of 0 typically corresponds to the non-expressed state while a value of 1 describes expression of a gene. Upon iteration the system proceeds through various gene expression states based on a set of logical rules describing the cross-regulation between genes. Because of their simplicity, boolean models of gene expression have been extensively studied to understand the complex dynamics of gene regulatory networks. One prominent study on boolean models of gene regulation for example, suggested that the structural properties of gene regulatory networks may have evolved to a critical balance point between a order and chaos, in which both the robustness necessary for survival, but also sufficient adaptability to react to changing environmental conditions are provided [Kauffman, 1969]. In addition to such conceptual studies, also boolean models of specific gene regulatory networks have been studied. For example, a boolean model of early sea urchin development was developed based on an extensive review of the scientific literature as well as targeted molecular experiments [Davidson et al., 2002]. In a similar case, a logical model of genetic interactions present in the yeast cell-cycle was constructed. Interestingly, upon artificial perturbation this network showed a return of its gene expression state to states corresponding to important cell-cycle checkpoints [Li et al., 2004].

Particularly for large biological systems containing many genes, boolean models of gene regulation pose an attractive mathematical formalism. They however come with the disadvantage of only qualitatively describing system dynamics. Alternatively, a more detailed representation of gene regulatory networks can be gained by expressing mRNA or protein abundance using real-valued instead of discrete state variables. In this context Ordinary Differential Equation (ODE) based models, which describe the change in activity or expression of a gene in relation to other genes in the system, have widely been adopted. An early ODE model of gene expression dynamics for example was used to investigate the genetic basis of switching from lysogenic to lytic mode of growth in the bacteriophage lambda [Shea and Ackers, 1985]. Also anterior-posterior patterning during *Drosophila* development has been extensively studied using ODE models, providing insight into the read-out of positional information and pattern formation [Jaeger et al., 2004].

Both boolean models and ODE based models are deterministic, implying that the trajectory of the system state is fully predictable based only on the knowledge of the current system state. Here stochastic models of gene expression represent the more realistic model, since they account for experimental as well as intrinsic noise apparent in any biological system. Many intermediate mathematical formalisms exist on the spectrum of possible frameworks to model gene regulatory networks. In general, which specific type of model formalism is chosen to describe a gene regulatory network depends both on the complexity of the system, as well as the amount and quality of available experimental data. Rather than there being a 'one-size-fits-all' solution to describing biological systems, the choice of modelling formalism needs to be considered on a case-to-case basis.

1.3 The systems biology modelling cycle

The first essential step when applying any modelling strategy is the design of a mathematical model based on prior knowledge of the biological system. In the best case scenario, available experimental evidence on the cis-regulatory structure of the system can be used to determine existing regulatory interactions between genes involved in the process studied. Here for example knock-down experiments, computational analysis of enhancer and promoter sequences, as well as Chromatin Immunoprecipitation (ChIP) experiments, provide valuable information on the structure of a gene regulatory network. In case such detailed experimental evidence is missing however, data-based network inference may be used reconstruct the structure of gene regulatory networks. These network inference typically apply a range of computational strategies to infer regulatory relationships between genes from large-scale time-course or perturbation gene expression data [De Smet and Marchal, 2010, Villaverde and Banga, 2013, Maetschke et al., 2014].

After having established the structure of a model, mechanistic properties of genetic interactions need to be modelled in more detail, based on physical laws or heuristics describing the quantitative relationships between genes in the network. These mathematical expression still carry a degree of uncertainty, as in most cases the values of kinetic parameters included in the model are unknown. Unfortunately, the direct measurement of kinetic

parameters *in vivo*, using molecular biology techniques, remains challenging. Therefore, in an alternative strategy, kinetic parameters are often estimated by fitting a dynamic model to experimental data. The general idea behind such a parameter estimation strategy is to simulate model dynamics using different possible parameter combinations and iteratively search for the parameter set best describing the data. In order to solve such complex optimization problems, in which the distance between modelled and measured gene expression values is minimized with respect to model parameters, various computational optimization strategies are available [Banga, 2008, Ashyraliyev et al., 2009].

As an alternative approach, instead of estimating specific parameter values of the model, the complete dynamic range of model behaviours can be explored by simulation of the model under a variety of experimental conditions or sampling from different parameter values. An analysis of model behaviour focuses on the question under which conditions model dynamics may exhibit interesting properties, for example bi-stability in certain state variables or robustness of timing events. In the context of cell mitosis for example, it was shown how robust temporal control of the splitting of chromosomes in anaphase can be achieved by adaptive rather than fixed thresholds of the cell cycle inhibitor securin [Kamenz et al., 2015]. Another example, originating from exhaustive model analysis, includes the observation that heterogeneity in signalling protein expression during TGFβ-signalling can lead to the emergence of different classes of cells, each class showing qualitatively distinct system dynamics [Strasen et al., 2018].

One particular interesting question seeked to answer by systems biology, is whether a formulated model sufficiently explains the given gene expression data. Often statistical tests are applied to the observed differences between modelled and measured gene expression, also known as model residuals, in order to test if a given model needs to be rejected. Rejection of a model can give critical insight into missing parts in our understanding of a biological systems and help formulate more realistic models. In case multiple alternative models describe the same biological system, model selection analysis attempts to identify the mechanistic model best explaining the data while at the same time making the least possible assumptions [Cedersund and Roll, 2009, Kirk et al., 2013].

Finally, in addition to model rejection or model selection analysis applied to existing data, predictions made by a model can help to design new and more optimal experiments in order to test different hypothesis about the system or enhance parameter identifiability [Bandara et al., 2009, Mélykúti et al., 2010].

The circular process of formulating a mathematical model, estimating model parameters, and analysing a model with the aim of designing new and informative experiments, leads to an iterative improvement of our understanding of a biological system. From this point of view, systems biology poses a valuable tool of biological epistemology providing a number of clear benefits [Dougherty and Braga-Neto, 2006, Gunawardena, 2014]: By the use of formal language, ambiguity in the formulation of a model is avoided, critical assumptions are exposed, and communication of the model is enhanced. The holistic viewpoint of model based systems biology further allows for the integration of existing knowledge with experimental data collected from various sources. Subsequent simulation and analysis of quantitative models aids the researcher in understanding the complex

dynamics exhibited by non-linear biological systems, revealing unexpected non-trivial behaviour of a system or providing explanations for emergent properties not visible on the level of single genes. Finally, mathematical formalisms ensure correct deductive reasoning, crucial for the design of new experiments by which a given model can be evaluated.

1.4 Scope of this work

Given the long history and benefits of systems biology analysis applied to the process of gene regulation, in this thesis we use systems biology modelling to investigate a variety of biological topics related to dynamics of gene expression and gene regulatory networks.

In the context of circadian rhythm, oscillations in gene expression form the basis for an organisms ability to adapt their physiology and behaviour to the 24-h day-night cycle to which they are exposed. Despite progress in understanding circadian gene expression, only a few players involved in circadian transcriptional regulation, including transcription factors, epigenetic regulators, and long noncoding RNAs, are known. Aiming to discover such genes, we perform a high-coverage transcriptome analysis of a circadian time-course in murine fibroblast cells. In combination with a newly developed algorithm, we identify many transcription factors, epigenetic regulators, and long intergenic noncoding RNAs that are cyclically expressed (Chapter 2). The mere identification of circadian gene expression genes however, does not provide information on how oscillations in gene expression arise. One possible hypothesis is the regulation of target genes by transcription factors, which themselves show circadian expression. Following this argumentation, we again make use of the transcriptomic time-course data obtained from synchronized NIH 3T3 cells in Chapter 2, and show that mathematical models of transcriptional regulation are able to predict putative regulator-effector interactions between identified circadian genes (Chapter 3).

The intricate interplay of genes in a biological system allows for the formation of complex gene regulatory networks controlling cell behaviour. The process of epithelial-to-mesenchymal transition (EMT), in which individual cells gradually disseminate from an integrated mesenchymal tissue, is thought to be coordinated by such a gene regulatory network. By applying state-of-the-art data-driven network inference methods to an extensive mRNA expression dataset measured in TGFβ-treated NMuMG cells, we predict the structure of the gene regulatory network controlling EMT (Chapter 4). Although informative, a mere structural representation of the EMT network does not capture the kinetic properties of gene expression changes relevant for the progression of EMT. On the basis of results obtained in Chapter 4, we therefore formulate a dynamical model of the EMT network, allowing for a more detailed analysis of gene expression during EMT (Chapter 5).

One important aspect of gene expression, often ignored by models of gene regulatory networks, is processing of mRNA transcripts and translation into protein. Despite the close relationship between mRNA and protein, the direct correlation between mRNA levels and protein abundances is moderate in most studies. In Chapter 6 we therefore determine whether the relation of mRNA and protein can be well explained by

simple mathematical models based on ordinary differential equations (ODEs) incorporating a temporal dimension. To this end, we generate a paired transcriptome/proteome time-course dataset with 14 time points measured in biological quadruplicates during *Drosophila* embryogenesis. On the basis of this data we formulate a systems biology framework for the identification of post-transcriptional gene regulation from large-scale time-resolved mRNA and protein expression patterns.

In each case study, the relevant biological and computational background is introduced and the quality of collected experimental data is validated. Whenever appropriate, the performance of computational methods is evaluated prior to their application to real gene expression data. Results originating from computational analysis of large-scale gene expression data are augmented further by the use of established bioinformatics tools and the discussion of results in light of the current scientific literature. Whenever possible, additionally performed experiments serve as a validation of computational predictions. While this thesis is mostly based in the field of systems biology, its target audience should include also readers yet unfamiliar with the topic. Therefore, throughout the chapters of this thesis, relevant concepts of systems biology are outlined and also reviewed.

Chapter 2

Identification of Circadian Expressed Genes from Large-Scale Gene Expression Data

Preamble

This project was carried out in close collaboration with SS, ST, DF, MHH, SL, and VT. Parts of this chapter have been published in:

Schick, S., Becker, K., Thakurela, S., Fournier, D., Hampel, M. H., Legewie, S., and Tiwari, V. K. (2016). Identifying Novel Transcriptional Regulators with Circadian Expression. *Molecular and Cellular Biology*, 36(4):545–558

2.1 Introduction

The availability of large scale datasets of gene expression is becoming more and more abundant given the use of high-throughput technology. Via RNA-Sequencing (RNA-Seq) the expression of thousands of genes can be assessed simultaneously. Decreased costs have further allowed for snapshot measurements of the full transcriptome under various experimental conditions as well as time-series measurements with high temporal resolution. A major challenge in biology today, is therefore the analysis of large scale data with the aim of identifying genes whose expression pattern meet certain criteria. Differential expression analysis, for example, uses statistical models in order to identify genes significantly changing between one or more experimental conditions. Various clustering approaches have been formulated to group genes based on their observed expression pattern. In this chapter we employ simple mathematical models to identify transcripts showing periodic expression in time-resolved RNA-Seq data.

2.1.1 The circadian clock and approaches to identify periodic gene expression

Organisms adapt to the 24-h day-night cycle, which leads to oscillations in physiology and behaviour. This is coordinated by an intrinsic molecular clock originating from the interplay of transcriptional-translational feedback loops [Lowrey and Takahashi, 2011]. In mammals, the core loop consists of the transcriptional activators Clock and Bmal1 and the repressors Period (Per) and cryptochrome (Cry). In a second loop, retinoid-related orphan receptors (ROR) activate transcription while Rev-Erb factors (Nr1d1 and Nr1d2) repress transcription [Partch et al., 2014]. These core clock components also regulate the expression of additional genes, possibly resulting in an oscillating transcription of these so-called clock-controlled genes and finally in the circadian phenotype.

Recent genome-wide studies of circadian time-courses in various mouse tissues indicated that each tissue expresses its own particular set of cyclical genes, which only partly overlap with each other [Panda et al., 2002, Storch et al., 2002, Menger et al., 2007, Miller et al., 2007, Zhang et al., 2014]. Nearly half of all genes in the mouse genome show circadian oscillation in at least one tissue [Zhang et al., 2014]. The basic mechanisms causing transcriptional rhythms of the core clock components are similar among all tissues, but how tissue-specific circadian output is achieved remains unknown, although several mechanisms have been proposed [Partch et al., 2014]. Among these are the use of tissue-specific transcription factors (TFs) or co-regulators [Miller et al., 2007, Menet et al., 2012] and different temporal control of RNA polymerase II recruitment [Koike et al., 2012, Le Martelot et al., 2012], as well as defined rhythms in histone modifications accompanied by differential gene regulation [Koike et al., 2012, Le Martelot et al., 2012, Etchegaray et al., 2003, Doi et al., 2006, Etchegaray et al., 2006, Nakahata et al., 2008, Jones et al., 2010, DiTacchio et al., 2011, Masri et al., 2012, Vollmers et al., 2012, Valekunja et al., 2013].

Novel factors, particularly TFs and epigenetic regulators, which are under circadian control and might either feed back into the core clock network or are involved in establishing the circadian phenotype via downstream target regulation remain to be identified. Another set of genes with an emerging role in the regulation of transcription are long intergenic noncoding RNAs (lincRNAs). It is not yet fully understood how lincRNAs control gene expression, but case-specific studies suggest that they might act as scaffolds to target chromatin-modifying complexes and transcriptional regulatory proteins to the genome [Vance and Ponting, 2014]. Recently, circadian expression of long noncoding RNAs was reported [Zhang et al., 2014, Coon et al., 2012].

Various strategies to identify periodic expressed genes have been formulated, the most elementary being visual inspection of gene expression time-courses [Cho et al., 1998]. In a more systematic approach, Fourier scores calculate the similarity of gene expression time-courses to periodic curves of predefined phase and period length [Spellman et al., 1998]. Other methods operating in the time domain use least squares regression to fit models of periodicity to gene expression data [Johansson et al., 2003].

A method operating in both time- and frequency-domain is `ARSER`. `ARSER` applies

a combination of autoregressive spectrum analysis and harmonic regression to identify cyclically expressed genes [Yang and Su, 2010]. First, the signal is converted into the frequency domain in order to estimate the main period length of the oscillation. Then sinusoidal models with fixed period length are regressed to the time-course data and statistical significance is assessed.

A different, computationally efficient approach to identify periodic signals is to iteratively group subsequent data-points, corresponding to different possible period lengths and phase shifts, and detect monotonic orderings of data-points within these groups. As a consequence of this strategy however, period lengths can only be estimated as discrete multiples of the sampling interval. One implementation making use of the outlined strategy is `JTK_CYCLE` [Hughes et al., 2010]. The algorithm `RAIN` is a generalization of the `JTK_CYCLE` algorithm, but less stringent regarding the shape of the waveform, hereby allowing for asymmetric oscillations, for example consisting of a steep rise and comparatively slower decay [Thaben and Westermark, 2014].

In this chapter we propose our own computational approach to identify circadian expressed genes based on model fitting. We evaluate its performance and compare it to the two existing methods `JTK_CYCLE` and `ARSER`. Using high-throughput mRNA measurements obtained by RNA-Seq in synchronized NIH 3T3 cells, we formulate a set of high confidence circadian expressed mRNA and lincRNA. The classification of genes is followed by extensive bioinformatics analysis, testing for the enrichment of specific biological functions or TF binding motifs in circadian expressed genes peaking at different times in the day-night cycle.

2.2 Results

2.2.1 Model based method to identify periodic expressed genes from large-scale expression data

The general approach to identify periodic expressed genes from time-course data formulated in this chapter makes use of a sine function describing the periodic expression pattern of genes. This specific formulation includes the assumption that gene expression patterns do not contain any additional expression changes apart from their oscillating behaviour, for example steady up- or down-regulation. As a consequence, in order to fulfil model assumptions, expression data for each gene needs to be detrended by subtracting a fitted line from the data. In addition to removing linear trends from the data, this further normalizes data of each gene to its mean value. Accordingly, the function describing periodic gene expression is dependent on the three parameters amplitude (A), period length (T), and phase shift (ϕ):

$$y^{model}(\theta, t) = A\ sin(\frac{2\pi}{T}t + \phi). \tag{2.1}$$

Model parameters $\theta = \{A, T, \phi\}$ are initially unknown, but can be estimated from the data. The general idea behind parameter estimation is to iteratively simulate the model with different parameters and select the set of parameter values with the best match between modelled and measured gene expression. In order to maximize the match between model and data, a measure for the goodness of fit must be formulated. Here the goodness of fit between model and data is defined by the squared difference between modelled and measured gene expression values. Additionally, the contribution of individual data-points to the goodness of fit can be weighted according to its measurement error (σ):

$$\chi^2(\theta) = \sum_{i=1}^{n} \frac{(y_i^{model}(\theta) - y_i^{data})^2}{\sigma^2}. \tag{2.2}$$

In the above expression, y denotes measured ($data$) or modelled ($model$) gene expression, n the number of measured data-points, and θ the three-dimensional parameter vector - one dimension for amplitude, period length, and phase shift. The mapping between three dimensional parameter space and the goodness of fit given in Equation 2.2 is often visualized as a rigid landscape with 'hills' corresponding to regions of high dissimilarity between model and data, and 'valleys' of better agreement in between. A large number of optimization strategies exist, systematically exploring the parameter landscape and efficiently locating minima in the cost function [Ashyraliyev et al., 2009].

In one particular optimization method, for example, a grid is defined over the defined

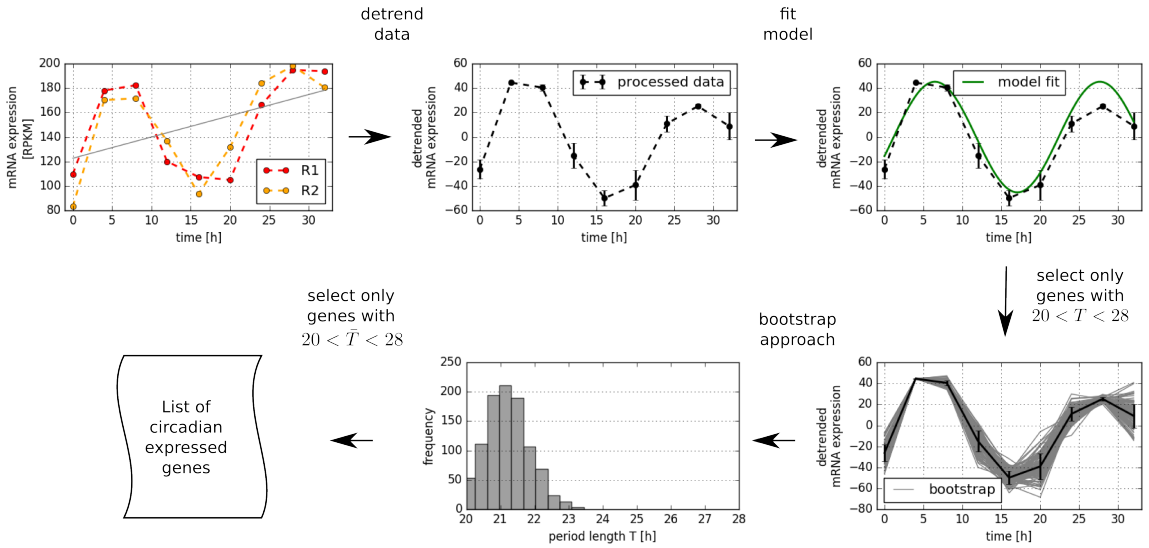


Figure 2.1: Workflow for the identification of circadian expressed genes. In order to identify circadian expressed genes from time-course gene expression data, raw expression values are averaged across replicates and the signal is detrended. A model for periodic gene expression (see Equation 2.1) is fit to the processed data, simultaneously estimating model parameters. In the first phase only genes with an estimated period length between 20h and 28h are selected for further analysis. In the second phase, a non-parametric bootstrap approach is applied to selected gene expression time-courses, generating a distribution of estimated model parameters with respect to potential experimental variation. Finally, only genes with an estimated mean period length across the bootstrap distribution are selected as circadian expressed.

parameter space with a fixed distance between grid-points. On each of the grid points the cost function is evaluated and parameters corresponding to the lowest cost function value are selected as the optimum. Although grid search covers all regions of parameter space equally well, the global optimum of the cost function might not be identified simply because of the fact that it may lie between two or more grid points. As function evaluations are typically costly in terms of computation time, and the number of grid points grows exponentially with the number of parameters as well as the resolution of the grid, brute force searching algorithms such as grid search can be computationally demanding.

By exploiting structural properties of the parameter landscape, gradient-based optimization methods can more reliably and efficiently identify minima in the cost function: An initial point in parameter space is proposed as the trial solution and from the derivatives of the cost function the direction in which the model fit increases most is calculated. Taking a step of appropriate size into this direction can decrease the value of the cost function. The procedure is repeated until certain termination-criteria are met. Since gradient-based optimization methods always select for parameters decreasing the cost function, the possibility of the algorithm becoming trapped in local minima exists. Therefore, in order to increase the chances of optimization to converge to the global optimum, gradient-based methods are typically run from multiple starting parameters.

In our approach to identify circadian expressed genes based on non-linear regression (Classification by Non-Linear Regression - `CBNLR`), we adopt a multi-start local optimiza-

tion procedure carried out using the Levenberg-Marquardt algorithm in order to estimate model parameters. Latin-hypercube-sampling, a pseudo-random sampling approach, ensures that parameter space is equally well sampled along all dimensions. While fitting the model to the data, each data-point of the cost function (Equation 2.2) is weighted equally. As a first step, we select genes with an estimated period length between 20h and 28h as an initial set of circadian expressed genes. Since measurement noise can carry over uncertainty to the estimated parameters, we further asses whether parameter values are robust against changes in the initial measurements. To determine the variance in estimated model parameters with respect to changes in the data, we therefore adopt a non-parametric bootstrap approach. In the bootstrap approach each data-point of the gene expression data is re-sampled, assuming a normal distribution with a standard deviation corresponding to the experimental error. Model parameters are again estimated from each bootstrapped time-course, resulting in a distribution of parameters over the bootstrap samples. Finally, only genes showing a mean period length between 20h and 28h over the bootstrap distribution and a small relative error are selected as the final set of circadian expressed genes.

2.2.2 Model based method to identify periodic expressed genes shows competitive performance on realistic benchmark data

Before applying the proposed method to identify circadian expressed genes to the measured gene expression data, its performance was evaluated on a benchmark dataset, in which the set of circadian genes is known. Structural details of the benchmark data can however critically influence the performance measurement [Futschik and Herzel, 2008]. Similar to benchmark data proposed by Yang and Su (2010) [Yang and Su, 2010], we therefore generated four different benchmark datasets, each with its own level of complexity, . In benchmark datasets A and B, circadian expressed genes were simulated using trigonometric functions with various period lengths between 20 and 28h, while shifts in phase could vary over the entire cycle. In reality however, due to the loss of cell synchronization, often a decay in signal strength is observed. Datasets C and D therefore contain non-stationary periodic signals, which decay in signal amplitude over time. For all periodic signals, a signal-to-noise ratio of 5:1 and 1:1 was considered. While in datasets A and C the set of non-circadian expressed genes was modelled by a white-noise process, in datasets B and D the respective background consisted of auto-regressive processes of order 1. In contrast to white-noise processes, auto-regressive processes better capture the existing degree of auto-correlation in gene expression data, as consecutive time-point measurements in gene expression are not independent of one another.

We applied the proposed method to identify circadian expressed gene to the generated benchmark data and compared its performance to existing approaches `JTK_CYCLE` and `ARSER`. Different metrics from binary classification strategies were used to evaluate the performance of the applied methods to identify circadian expressed genes. The number of true positives for example indicates the number of genes correctly identified as

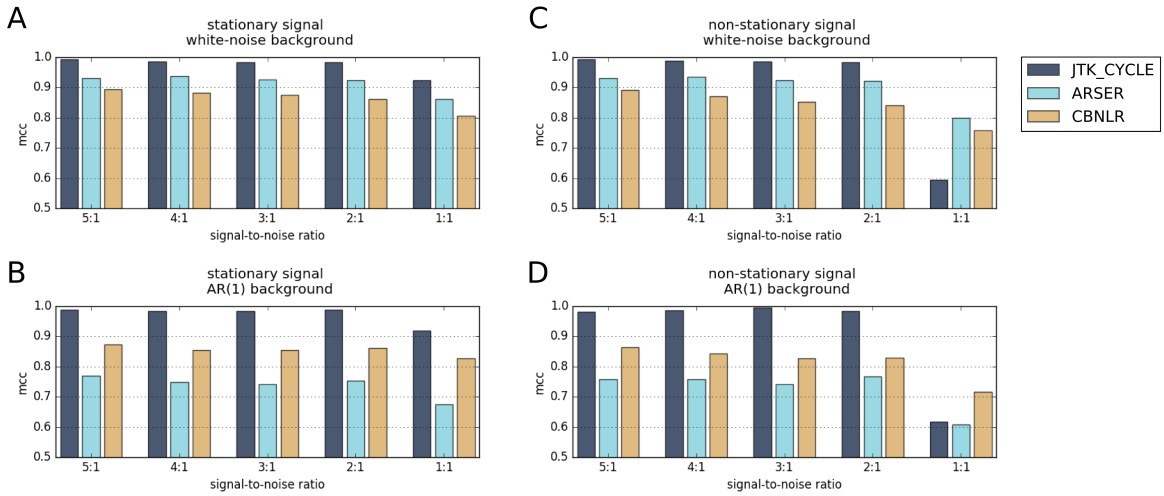


Figure 2.2: Performance evaluation of methods to identify circadian expressed genes. Matthews correlation coefficients for three different methods designed to identify circadian expressed genes applied to various benchmark datasets. The benchmark data was composed of either stationary (A and B) or non-stationary periodic signals (C and D). Background signals were modelled by either white-noise (A and C) or auto-regressive (AR) processes (B and D). In each plot the signal-to-noise ratio decreases from left to right.

circadian expressed, while false positives describes the number of falsely identified circadian expressed genes. By combining true and false positives as well as negatives, the Matthews correlation coefficient (MCC), a measure for the correlation between the true set of circadian expressed and the corresponding prediction was calculated.

As a result of the evaluation our method performed less well compared to `JTK_CYCLE` and `ARSER` on datasets with a white-noise background (Figure 2.2A and B). Our approach did however show increased perform compared to `ARSER` on datasets C and D, which are the respective datasets with an AR(1)-background. `JTK_CYCLE` demonstrated high performance in all four datasets. Although our own approach did not meet the performance of `JTK_CYCLE`, the method performed well even in the context of a low signal-to-noise ratio.

Combining results using the overlap of the three sets of genes identified as circadian by all three methods, the fraction of correctly identified circadian genes in the set of genes predicted to be circadian (precision value) was $> 99\%$ (data not shown). Recall - or the number of correctly identified circadian genes divided by the total number of circadian genes in the data - in this case only reached 80-90%. Choosing signals identified by at least two out of three methods provided a good balance between precision (94-97%) and recall (95-99%).

2.2.3 Integration of methods identifies high-confidence set of circadian expressed genes

After evaluating the different methods for the identification of circadian expressed genes, including the newly formulated approach, we turned to real mRNA expression data to study transcriptional regulation during circadian rhythm in a mammalian system. For

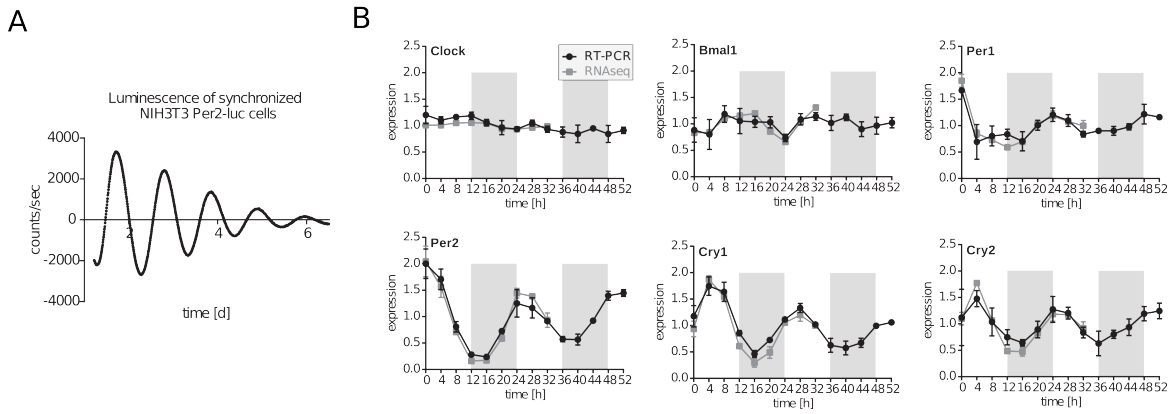


Figure 2.3: Validation of the synchronization protocol in NIH 3T3 cells. (A) Background-subtracted luminescence signal of dexamethasone synchronized NIH 3T3 Per2:luc cells. **(B)** Expression of core clock components as measured by RNA-seq (n = 2, normalized to average expression) or measured by RT-qPCR (n = 4, normalized to Hsp90ab1 and to average expression; error bars represent standard errors of the means [SEM]). Dashed lines mark the 24-h and 48-h time points.

this we synchronized NIH 3T3 cells with dexamethasone, a cell line which has been routinely used as a model system for circadian rhythmicity [Isojimaa et al., 2009, Ma et al., 2013, Nagoshi et al., 2004, Nagoshi et al., 2005, Hughes et al., 2009, Morf et al., 2012, Engelen et al., 2013, Bieler et al., 2014].

NIH 3T3 cells, in which luciferase serves as a reporter for Per2 promoter activity (Per2:luc cells), showed a cyclical luminescence signal over time, hereby validating the synchronization protocol (Figure 2.3A). Furthermore, we confirmed the cyclical expression pattern of known core clock genes in NIH 3T3 cells by using RT-quantitative PCR (RT-qPCR) over two 24h cycles. Consistent with previous studies [Hamilton and Kay, 2008], Clock showed nearly stable expression, whereas Cry and Per mRNAs showed cyclical expression peaking around 4h, 28h, and again at 52h (Figure 2.3B). Bmal1 (Arntl) expression peaked in between the Per and Cry maxima, at approximately 12h and 36h.

We further performed genome-wide RNA expression analysis over a 32h time-course with 4h resolution via high-throughput RNA-Sequencing. Previously, similar datasets were generated via microarray analysis [Menger et al., 2007,Hughes et al., 2009]; however, RNA-Sequencing technology enables a deeper coverage and a more extensive analysis of the transcriptome [Zhao et al., 2014]. Furthermore, it allows analysis of not-yet-annotated coding RNAs as well as noncoding RNAs. The data therefore add substantially to the published microarray datasets. Based on our RNA-Seq data, half of the genes (50.5%, 11,082 genes) were expressed at least at one time point and were included for further analysis. The expression patterns of core clock genes (Clock, Cry, Per, and Bmal1 genes) coincided well with RT-qPCR measurements (Figure 2.3B).

Given the fact that the experimental synchronization protocol was successfully implemented, we next applied our method to identify circadian expressed genes to the full RNA-Seq dataset. In the first phase of the approach, 3691 genes showed an estimated period length between 20h and 28h. By applying the non-parametric bootstrap approach,

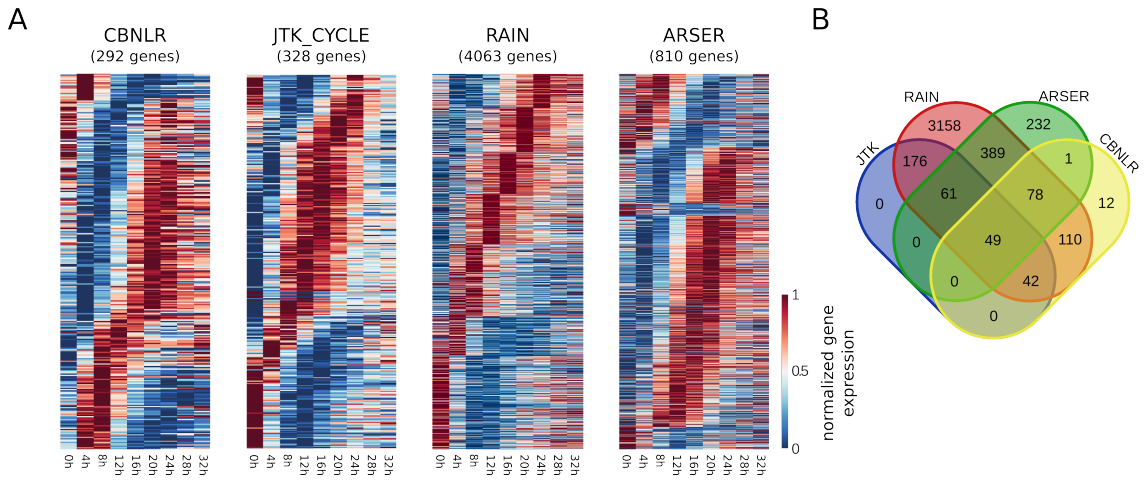


Figure 2.4: Identification of circadian expressed genes in NIH 3T3 cells. (A) Heatmaps of cyclical expressed genes after minimum-maximum normalization (blue - lowest expression, red - highest expression) detected by either `CBNLR`, `JTK_CYCLE`, `RAIN`, or `ARSER`. Genes are sorted according to the phase given by the corresponding method. **(B)** Venn diagram of cyclical expressed genes detected by `JTK_CYCLE`, `RAIN`, `ARSER`, and the `CBNLR` approach.

only 292 of the initially chosen circadian expressed genes remained. This set represents 2-3% of the complete set of analysed genes (Figure 2.4A) and contained most of the known core clock components, such as the Period and cryptochrome genes, as well as Bmal1, Nr1d1, and Nr1d2. In order to compare these results to existing methods, we further identified circadian expressed genes in NIH 3T3 cells using `JTK_CYCLE`, `RAIN` and `ARSER`. `JTK_CYCLE` applied to the NIH 3T3 data returned 328 cyclical transcripts (3% of all expressed genes). Again there were most established core clock components among the genes identified as circadian. `RAIN` identified 4,063 cyclical genes with a period length between 20h and 28h. Since `RAIN` presents a generalization of `JTK_CYCLE`, it is expected that both methods show a large overlap. Indeed, all genes identified as circadian by `JTK_CYCLE` have also been detected by `RAIN`. `ARSER` detected 810 cyclically expressed genes (Figure 2.4B). As noted also by others, and with the exception of `JTK_CYCLE` and `RAIN`, we found little overlap between the oscillating genes detected by `ARSER`, `JTK`, `RAIN`, and our approach [Koike et al., 2012, Doherty and Kay, 2010].

2.2.4 Circadian expressed genes contain known core-clock components and are distributed across various phases

The common set of genes identified by all four methods consisted of 49 genes and included genes for known periodic expressed core clock components, such as Cry1, Cry2, Nr1d2, Per2, and Per3 (Figure 9.1). For further analysis, we selected high-confidence cyclical genes from our dataset, by considering only genes which were classified as cyclical by at least three of the four analysis methods. This resulted in a set of 230 genes. The expression patterns of these genes fall into various oscillation phases, as shown in the exemplary time-courses (Figure 2.5).

When we compared the identified cyclical genes to the results of previous microarray

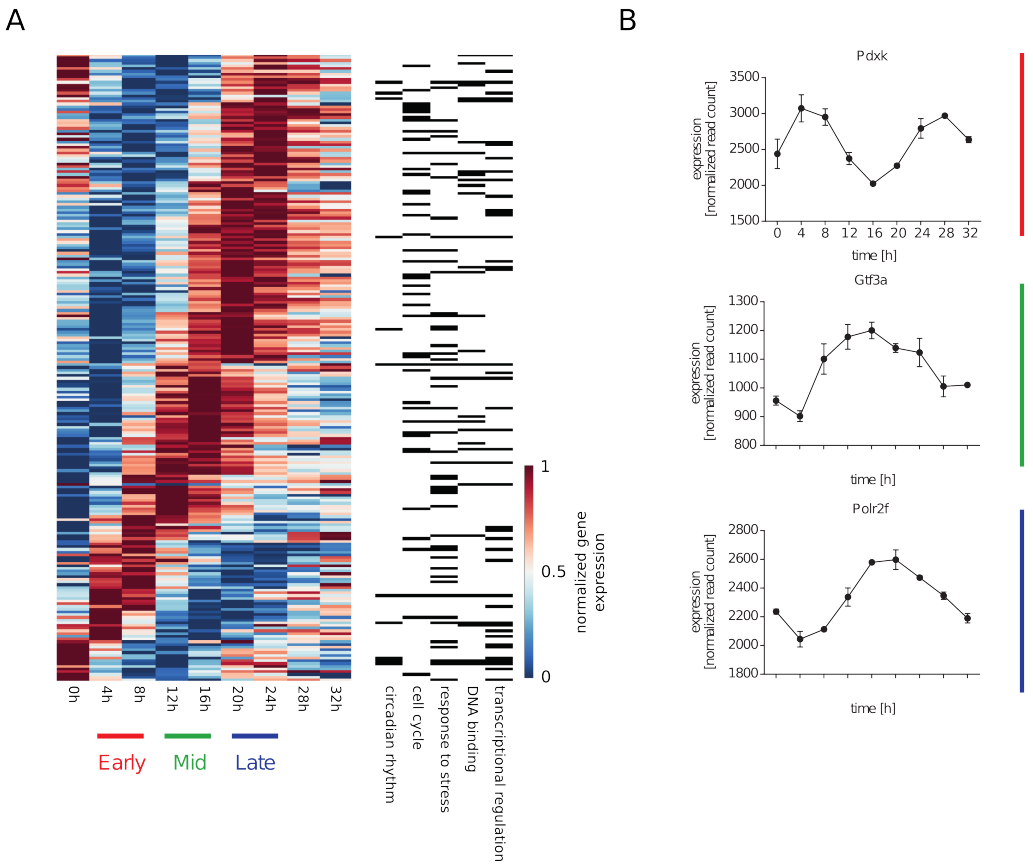


Figure 2.5: Characterization of circadian genes in NIH 3T3 cells. (A) Heat map showing the expression of 230 cyclically expressed genes (detected by at least three methods). The expression values per gene were minimum-maximum normalized. Genes are sorted by the phase determined with `JTK_CYCLE`. GO terms associated with each gene are indicated in black on the right. **(B)** RNA-Seq data for exemplary genes with circadian expression showing different peak times ($n = 2$, normalized read counts, error bars represent SEM). Coloured lines indicate the time point of their highest expression during the time-course as marked in panel A (early, mid, and late).

studies performed in NIH 3T3 cells [Menger et al., 2007, Hughes et al., 2009], we found only minimal overlap between all studies, possibly due to differences in the synchronization protocols or classification method for cyclical genes (Figure 9.2). The total overlap between the two microarray studies was restricted to only four core clock genes (Per2, Per3, Cry1, and Tef). Our data showed greater overlap with each of the microarray experiments, and our results also revealed the four above-mentioned genes as well as additional core clock genes not identified by the microarray studies.

We next investigated, whether there was evidence that our identified cyclically expressed genes are under circadian control in other cell or tissue types. CircaDB, a database that compiles published transcriptome data of circadian time-courses in different tissues and cellular systems, is a valuable resource for this purpose [Pizarro et al., 2013]. We found that 176 out of 230 cyclically expressed genes in our dataset also oscillate in at least one other cell or tissue type (76%) (Figure 9.3A).

Moreover, we examined whether the promoters of these genes are bound by core

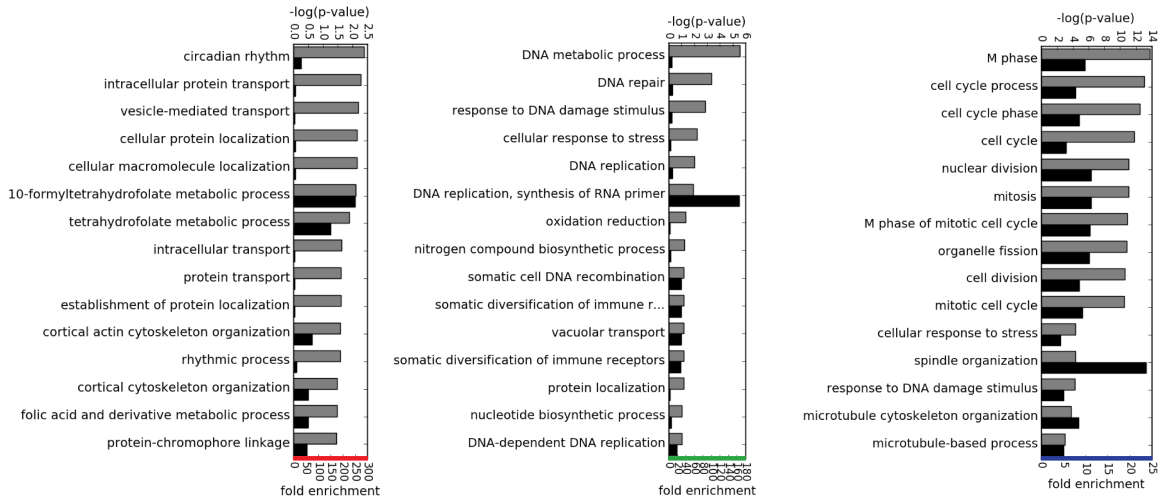


Figure 2.6: Enrichment of biological functions in different classes of circadian expressed genes. GO term enrichment for early (red), middle (green), and late (blue) peaking genes (top 15 GO-terms are shown based on p-values). Gray bars indicate the negative logarithm of the p-value (top axis); black bars show fold enrichment (bottom axis).

clock factors identified in different chromatin immunoprecipitation sequencing (ChIP-Seq) studies performed with liver cells or macrophages [Menet et al., 2012, Koike et al., 2012, Rey et al., 2011, Annayev et al., 2014, Fang et al., 2014, Cho et al., 2012, Bugge et al., 2012, Lam et al., 2013]. Strikingly, 225 of these genes have been determined to be bound by at least one core clock factor (Clock, Per1, Per2, Cry1, Cry2, Npas2, Bmal1, Nr1d1, Nr1d2, or Rora; 97.8%, 1.36-fold increase above background; $p < 0.001$) (Figure 9.3B and Figure 9.4), indicating that most of the 230 cyclically expressed genes might be under the control of the circadian network.

2.2.5 Circadian expressed genes show distinct biological functions

We further investigated potential biological functions enriched in the set of high-confidence circadian genes using Gene Ontology terms (GO terms). GO terms represent a controlled vocabulary in which genes and their products are assigned information regarding their molecular function, involvement in biological processes, or localization in different cellular components. According to this analysis, the set of 230 genes periodically expressed with a period length between 20h and 28h includes circadian regulators, stress response genes, and cell cycle genes (GO category: biological processes), as well as many genes that bind to nucleotides and/or are implied to be transcriptional regulators (GO category: molecular function) (Figure 2.5).

Groups of genes with common functions may peak at certain phases of the circadian cycle. To analyse this, we performed Gene Set Enrichment Analysis (GSEA) for all cyclically expressed genes, sorted by their circadian phase [Mootha et al., 2003, Subramanian et al., 2005]. Interestingly, some GO terms showed enrichment among genes within a certain range of phases, while others displayed nearly uniform distribution. For example,

genes related to metabolic processes were evenly distributed throughout the time-course, while genes involved in the cell cycle tend to peak late in the circadian cycle (Figure 9.5).

To investigate this further, we determined the maximal expression of the cycling genes between 4h and 24h of the time-course and divided them into early (maximal peak at 4h or 8h), middle (maximal peak at 12h or 16h), and late (maximal peak at 20h or 24h) cycling genes (compare Figure 2.5A and B). For these three sets of genes, we performed GO term analysis [Huang et al., 2009a,Huang et al., 2009b] as well as motif prediction [Heinz et al., 2010]. This revealed that genes related to circadian rhythm were mainly enriched in the set of genes showing high expression early in the time-course (Figure 2.6). The genes of this class also exhibited enrichment of the NF1 motif in their promoter (Table 2.1). Genes of the second category, peaking at 12h or 16h, were enriched for the E2F1, NRF, and E2F7 motifs and GO terms related to processes such as DNA metabolic process, DNA repair, cellular response to stress, and DNA replication. As observed by GSEA, the cycling genes peaking late were highly enriched for cell cycle-related genes and stress response genes. Motif analysis detected enriched HIF-1a, NFY, and the E-box-containing motifs c-MYC, USF2, MITF, NPAS2, and bHLHE40 in the promoters of these genes.

2.2.6 Circadian expressed genes contain transcriptional regulators

The GO term analysis indicated that in the set of cyclically expressed genes many genes are associated with transcriptional regulation (GO:0006355). Since the fate of a cell is largely defined by its transcriptome, we investigated the genes grouped under the GO terms 'DNA binding' and 'transcriptional regulation' in more detail. In addition, we screened the list of 230 cyclically expressed genes for putative transcription factors

Table 2.1: Motif enrichment in different classes of circadian genes. Enriched sequence motifs in the promoters (± 1 kb of TSS) of early, middle, and late peaking genes. The percentage of target as well as background sequences with motif are shown.

	motif name	consensus	% (target)	% (background)
Early (45)	NF1		93.33%	76.51%
Mid (52)	E2F1 (E2F)		57.69%	34.90%
	NRF1 (NRF)		44.23%	25.78%
	NRF (NRF)		44.23%	27.16%
	E2F7 (E2F)		34.62%	20.01%
Late (120)	HIF-1a (bHLH)		36.67%	22.76%
	NFY		65.00%	52.32%
	GFY-Staf (Zf)		15.00%	7.59%
	c-Myc (bHLH)		54.17%	41.93%
	Usf2 (bHLH)		30.00%	20.00%
	MITF (bHLH)		60.00%	48.43%
	NPAS2 (bHLH)		65.00%	53.74%
	Sp1 (Zf)		62.50%	51.28%
	bHLHE40(bHLH)		34.17%	24.25%

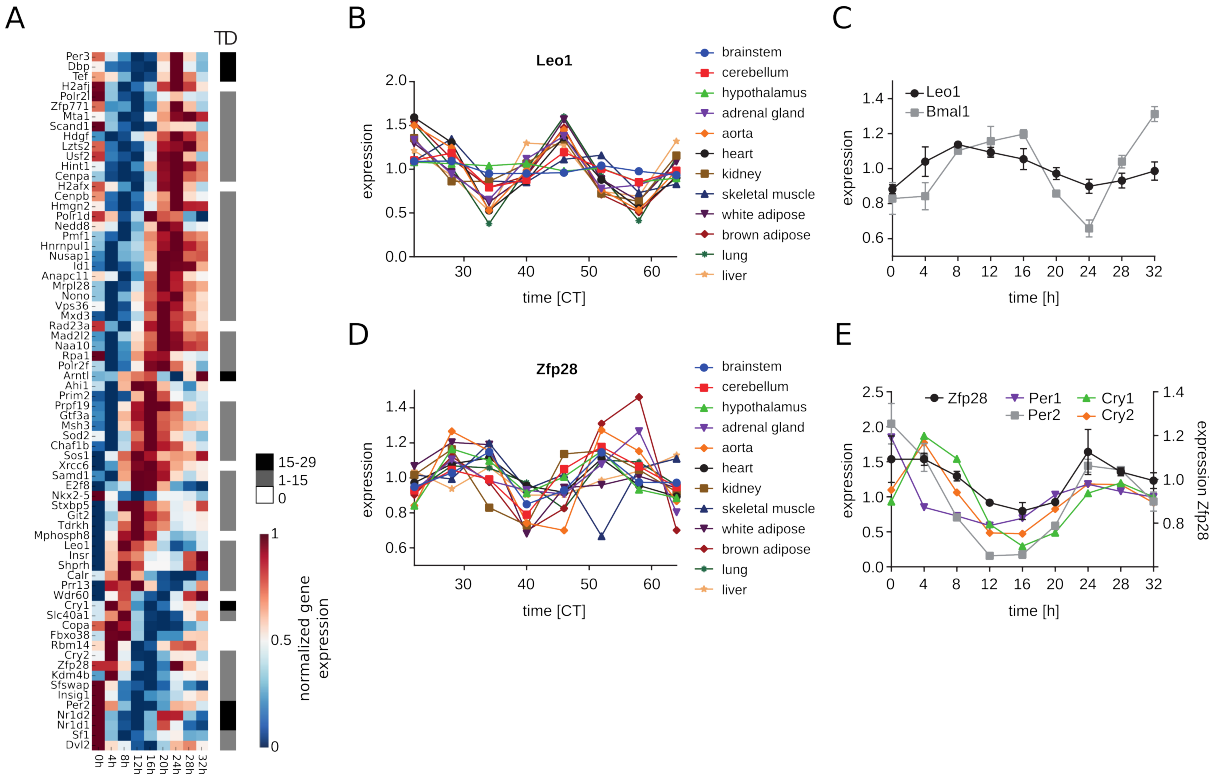


Figure 2.7: Circadian expressed transcriptional regulators, epigenetic regulators, and DNA binding genes. **(A)** Heat map of transcription factors, epigenetic regulators, and additional DNA binding genes with circadian expression in NIH 3T3 cells (minimum-maximum normalized). Genes were sorted by the estimated phase given by `JTK_CYCLE`. The right column (TD) indicates how many times out of 29 published time-course datasets (obtained from CircaDB) a gene was classified by `JTK_CYCLE` to be cyclically expressed. White, in no dataset; gray, in 1 to 15 datasets; black, in $>$ 15 datasets. **(B)** Expression of Leo1 in different mouse tissue time-course experiments (data obtained from GEO, GSE54651). Data were normalized to average expression of Leo1 per tissue. **(C)** Expression of Leo1 compared to Bmal1 in the NIH 3T3 time-course RNA-seq data. Normalized read counts of two replicates were normalized to average expression and are plotted with error bars (standard errors of the means [SEM]). **(D)** Expression of Zfp28 in different mouse tissue time-course experiments (data obtained from GEO, GSE54651). Data were normalized to average expression of Zfp28 per tissue. **(E)** Expression of the Zfp28 compared to Cry and Per genes in NIH 3T3 time-course RNA-seq data. Normalized read counts of two replicates were normalized to average expression and are plotted with error bars (SEM).

and epigenetic regulators, based either on prior knowledge from the literature or on functional domains within the proteins. This filtering approach resulted in a list of 70 putative transcriptional regulators with cyclical expression (Figure 2.7A), also containing genes previously studied in the context of circadian rhythm (e.g., Tef, Dbp, Nono, Mta1, Id1) [Wuarin and Schibler, 1990, Fonjallaz et al., 1996, Humphries et al., 2002, Brown et al., 2005, Li et al., 2013, Kowalska et al., 2013].

It has been shown that the overlap of genes with circadian expression between different tissues is low [Panda et al., 2002, Storch et al., 2002, Menger et al., 2007, Miller et al., 2007, Zhang et al., 2014]. However, if any of identified circadian expressed transcription factors plays a general role in the regulation of circadian rhythm, it should be cyclically

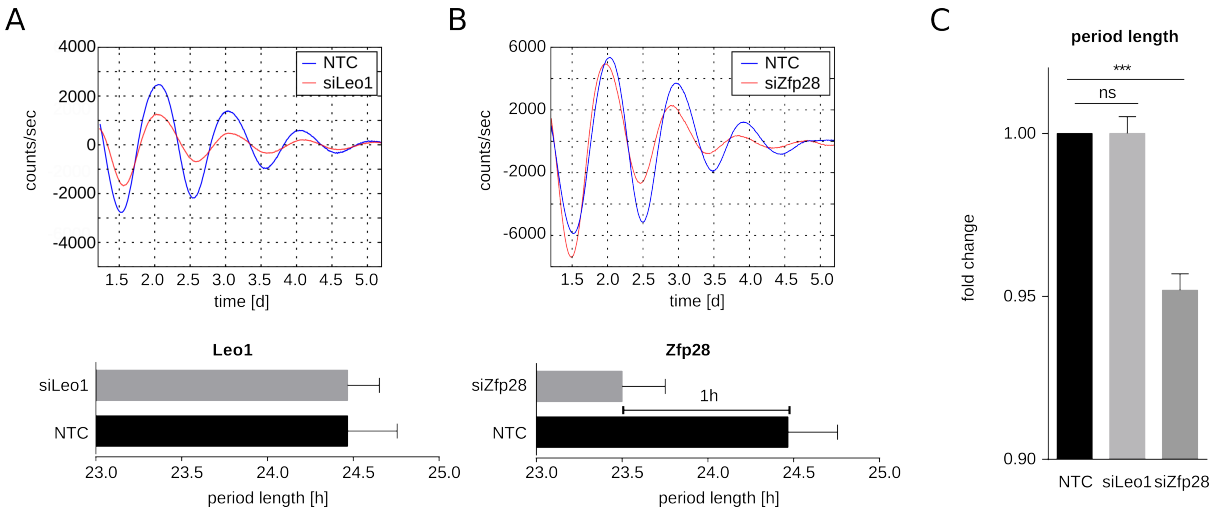


Figure 2.8: Cyclically expressed transcription factors and epigenetic regulators can affect the core clock network. (A) Luminescence measurements of NIH 3T3 Bmal1:luc cells under knock-down of Leo1 (siLeo1) and control (NTC) conditions (top). The bar graph shows the calculated period lengths of the luminescence signal under siLeo1 and NTC conditions of three independent replicates (bottom). **(B)** Luminescence measurements of NIH 3T3 Bmal1:luc cells under knock-down of Zfp28 (siZfp28) and control (NTC) conditions (top). The bar graph shows the calculated period lengths of the luminescence signal under siZfp28 and NTC conditions of three independent replicates (bottom). **(C)** Fold change in period length of knock-down cells (siLeo1, siZfp28) compared to control cells (NTC), measured by luminescence in NIH 3T3 Bmal1:luc cells (n = 3). *, p = 0.0318, Mann-Whitney U test.

expressed in most tissues. We therefore analysed which of the 70 putative transcriptional regulators were also cyclically expressed in other tissues, again using the datasets provided by CircaDB [Pizarro et al., 2013]. Indeed, we found all core clock components and well-established players in circadian rhythm to be cyclically expressed in nearly all investigated tissues (Figure 2.7A, right column). Furthermore, among the 70 potential transcriptional regulators, we found additional genes showing cyclical expression in several tissues. This suggests that some of the identified factors might play a general role in the regulation of circadian rhythm or in circadian-regulated processes. For a more extensive characterization, we selected 2 candidates out of these 70 factors that had not been studied in detail with regard to circadian rhythm, Leo1 and Zfp28.

Leo1 is a component of the PAF complex, which interacts with RNA polymerase II. Inspecting recently published transcriptome data of time-courses in different mouse tissues covering all three germ layers, Leo1 showed circadian expression in 10 out of 12 tested mouse tissues (Figure 2.7B) [Zhang et al., 2014]. Interestingly, Leo1 expression, which we validated by RT-qPCR (Figure 9.6), is cyclically expressed in phase with Bmal1 in all of the 10 tissues, as well as in NIH 3T3 cells (Figure 2.7C). Anafi et al. (2014) provide a ranked list of 1,000 potential core clock components based on multiple metrics (cycling, phenotype, network interaction, ubiquity, and phylogenic conservation) derived from various datasets [Anafi et al., 2014]. Leo1 was ranked number 200 on this list of potential core clock factors.

An additional factor, the zinc finger protein Zfp28, showed cyclical expression in phase

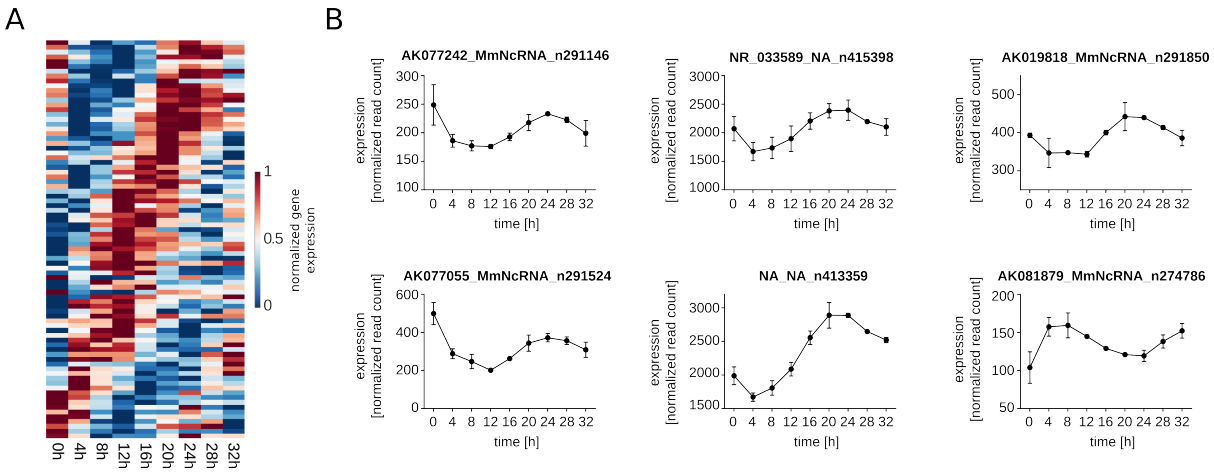


Figure 2.9: Identification of Circadian Expressed lincRNAs in NIH3T3 cells. (A) Heat map showing expression of 83 lincRNAs detected to be cyclical by at least three out of four computational methods (minimum-maximum normalized). lincRNAs were sorted by the phase, determined by `JTK_CYCLE`. **(B)** RNA-seq expression data for six cycling lincRNAs (n = 2; error bars represent standard errors of the means [SEM]).

with downstream targets of Clock and Bmal1 (Figure 2.7D and E). Although cyclical expression of Zfp28 in several tissues was clearly visible, CircaDB failed to identify it as such, given the standard parameters (q-value of < 0.05). Perhaps this resulted from the low fold change in Zfp28 expression.

We performed knock-down experiments of Leo1 and Zfp28 in order to explore their impact on the core clock network. While knock-down of Clock and/or Bmal1 abolished the cyclical expression of luciferase in NIH 3T3 Bmal1:luc cells (Figure 9.7), knock-down of Leo1 and Zfp28 did not suppress circadian cycling (Figure 2.8A and B). Hence, Leo1 and Zfp28 are not essential for circadian rhythm. However, Zfp28 knock-down decreased the period length of the circadian rhythm of Bmal1 by approximately 1h, indicating a possible role of Zfp28 in modulating the core clock network. In contrast, Leo1 did not affect the period length of circadian Bmal1 cycling.

Recently, also lincRNAs have emerged as critical regulators of gene expression [Vance and Ponting, 2014, Kornienko et al., 2013]. Having performed high-coverage RNA sequencing, we were able to investigate the expression patterns of these RNA transcripts. Again, we performed the analysis using `JTK_CYCLE`, `RAIN`, `ARSER`, and our own approach. We identified around 130 to 400 cyclically expressed lincRNAs with each method (`JTK_CYCLE` identified 134; `RAIN`, 392; `ARSER`, 161; `CBNLR`, 211 - Figure 9.8); among them, 31 lincRNAs were detected by all four methods (Figure 2.9A and B), indicating that periodically expressed lincRNA are involved in circadian rhythm, either in maintaining circadian gene expression or its phenotypical output.

2.3 Discussion

In this chapter we proposed a model of periodic gene expression in combination with non-linear regression to identify circadian expressed genes from genome-wide time-course mRNA expression data. In contrast to other approaches for the identification of circadian expressed genes, in the approach presented here all parameters of the underlying model for periodic gene expression - including the period length - are estimated directly from the time-course data and do not need to be predefined. In addition, the method explicitly takes into account the experimental variability between biological replicates by using a non-parametric bootstrap approach. Bootstrapping time-series data is inherently challenging, since auto-correlation properties of the original time-course can be lost while individual data-points are perturbed [Härdle et al., 2003]. As an alternative approach a parametric bootstrap, in which re-sampled gene expression time-courses are generated from the best model fit, should be considered. In any case, bootstrap procedures in combination with non-linear regression are computationally demanding. Reasonable computation times therefore require an efficient parallelization strategy.

As stated previously, the performance evaluation of computational approaches to identify circadian expressed genes relies heavily on realistic benchmark data. Following this argumentation, Lichtenberg et al. (2005) compiled a list of potential cell-cycle genes in yeast based on a combination of criteria, including previously identified cell-cycle genes or association of genes with known transcriptional regulators of the cell-cycle [de Lichtenberg et al., 2005]. In the study, various methods to identify periodic expressed genes were applied to yeast gene expression data and the overlap to putative cell-cycle genes was evaluated. As a result, it was shown that no single-best performing method exists with each method identifying a different set of periodic genes. However, a method using partial least-squares [Johansson et al., 2003], similar to the approach presented here, was among the best performing methods applied in the study. In this chapter, instead of evaluating methods on gene expression data measured in a real biological system, we made use of *in silico* benchmarks, in which the set of period expressed genes is known. This permitted performance evaluation in a setting in which crucial properties of the benchmark data, such as the signal-to-noise ratio, could be controlled. Particularly in the presence of high noise, non-stationary periodic signals and realistic background gene expression, our proposed method showed competitive performance compared to `JTK_CYCLE` and `ARSER`.

As part of our observations, each method identified a different set of circadian expressed genes, only partially overlapping with each other. In the benchmark data we integrated results from various approaches to identify periodic expressed genes, hereby obtaining a significant match between predicted and true circadian expressed genes. Likewise, the high confidence set of circadian expressed genes derived from NIH 3T3 data contained most known core clock components.

Further analysis of circadian expressed genes in NIH 3T3 cells indicated their involvement in various biological processes, including circadian rhythm, stress response, and cell cycle regulation. Some groups of genes with a common functional annotation tend to peak in the same phase of the cycle. For example, cell cycle genes were preferentially

enriched for middle to late phases during the time-course, consistent with the finding of enriched cell-cycle regulator motifs in the promoters of these genes, e.g., the E2F1, c-MYC, HIF-1a, and MITF genes [Amati et al., 1998, Carreira et al., 2005, Goda et al., 2003, Koshiji et al., 2004, Johnson et al., 1993]. The middle and late categories were also enriched for stress response genes, which is consistent with the motif enrichments of factors involved in stress responses, such as E2F7, HIF-1a, and SP1 [Panagiotis Zalmas et al., 2008, Majmundar et al., 2010, Carvajal et al., 2012, Li et al., 2014]. Furthermore, gene promoters of each category were enriched for motifs related to circadian rhythm. NF1, detected in the early class, has been shown to play a role in circadian rhythm regulation in *Drosophila* [Williams et al., 2001]. Previous findings showed that NRF1, enriched in the middle class, regulate numerous circadian regulatory genes in NIH 3T3 cells [Zhu, 2011]. The transcription factor NF-Y, enriched in the late class, has been found to regulate the transcription of the core clock gene Bmal1 [Xiao et al., 2013]. Overall, GO terms and motifs assigned to each group of genes (early, middle, and late) are consistent with respect to their assigned functions. Furthermore, the detected motifs are in agreement with identified motifs enriched in promoters of clock-controlled genes in other cell and tissue types [Bozek et al., 2009, Bozek et al., 2010].

Among the cyclic genes, we found many potential transcription factors and epigenetic regulators. This emphasizes that transcriptional control is an important mechanism to regulate circadian rhythm. Two factors that were previously not implicated in circadian rhythmicity were characterized further: Leo1, a component of the PAF complex, displayed cyclical expression in phase with Bmal1 in many tissues. The cyclical expression of Leo1 in many tissues suggests an important role of Leo1 for the circadian phenotype. However, it is most likely not involved in the core clock network, as its knock-down did not affect cyclical expression of Bmal1. Leo1 has been shown to be important for the recruitment of the Paf1 complex to nucleosomes as well as for active H3K4 tri-methylation in yeast [Dermody and Buratowski, 2010, Chu et al., 2013], hinting at an essential role for Leo1 in transcriptional regulation. Its coordinated expression with Bmal1 further indicates that Leo1 might enhance transcriptional activation of clock-controlled genes by Clock and Bmal1.

The zinc finger protein Zfp28 has so far not been studied in the context of circadian rhythm. In humans, it showed strong expression in various adult tissues, but expression in embryonic tissue is development specific [Zhou et al., 2002]. Apart from being cyclically expressed in NIH 3T3 cells, the cyclical behaviour was also present in various tissues among different lineages. However, its amplitude is very low, which could explain why it is rarely detected as being cyclically expressed by computational approaches. Zfp28 cycled in phase with the classical down-stream targets of Clock and Bmal1, such as the Period and cryptochrome genes. As mentioned before, there is evidence that in liver Zfp28 is bound at the promoter by Clock and Bmal1 [Rey et al., 2011, Annayev et al., 2014]. Knock-down of Zfp28 shortened the period of Bmal1 oscillations by around 1h, indicating a potential role of this factor in the core clock regulatory network. To elucidate the mechanisms behind this, it would be interesting to investigate the genomic occupancy of Zfp28 by ChIP analysis and Zfp28 interaction partners by mass spectrometry.

We further identified cyclically expressed lincRNAs in our time-course data. lincRNAs constitute a different layer of transcriptional regulation, but their specific role in circadian rhythm remains to be investigated. For example, it seems that lincRNAs often accomplish their gene regulatory function in *trans* at distal binding sites [Vance and Ponting, 2014]. In the future the binding sites of cyclically expressed lincRNAs and their interaction partners, as well as their impact on the transcription of coding genes, should be studied in more detail.

In summary, measurement of high-throughput gene expression time-course data in synchronized NIH 3T3 cells and the identification of circadian expressed genes provides a valuable resource to the research community. Further, extensive analysis of circadian expressed genes adds to existing knowledge on circadian rhythm, ubiquitously present in all life forms.

Chapter 3

Predicting Gene Regulatory Interactions from Small-Scale Models of Transcriptional Regulation

Preamble

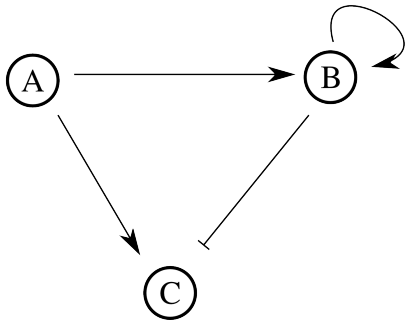
This project was carried out in close collaboration with SS, ST, DF, MHH, SL, and VT. Parts of this chapter have been published in:

Schick, S., Becker, K., Thakurela, S., Fournier, D., Hampel, M. H., Legewie, S., and Tiwari, V. K. (2016). Identifying Novel Transcriptional Regulators with Circadian Expression. *Molecular and Cellular Biology*, 36(4):545–558

3.1 Introduction

In Chapter 2 we identified a set of circadian expressed coding and non-coding genes from high-throughput time-resolved mRNA expression data. Following up on this set of circadian expressed genes, we tested for the enrichment of various biological functions or TF binding motifs in genes peaking at different time-points during the day-night cycle. Further, potential transcriptional regulators were selected from the list of circadian expressed genes and their circadian expression in multiple tissue types was confirmed. Using knock-down experiments, a role of the circadian expressed transcriptional regulator Zfp28 in regulating the period length of the core clock could be established.

In the following chapter we investigate potential cross-regulation among circadian expressed genes using low-parametric models of transcriptional regulation. In this computational approach, which we term `NodeInspector`, non-linear models of regulator-target interactions are fit to time-course gene expression data, allowing for the assessment of the likelihood of an interaction. The performance of the approach is demonstrated using time-course gene expression data produced by gene regulatory networks with known structure and compared to existing approaches with the aim to identify regulatory interactions among genes. We further apply `NodeInspector` to the NIH 3T3 dataset derived in Chapter 2, hereby testing the feasibility of potential regulatory interactions between



Regulator

Target	Gene A	Gene B	Gene C
Gene A	0	0	0
Gene B	3.2	1.5	0
Gene C	2.1	-0.9	0

Figure 3.1: Graph representation of a network. On the left a schematic of a gene regulatory network consisting of three genes is shown. Genes are represented as nodes, while interactions between them are shown as edges. An arrow represents activating regulation, while a crossed line indicates repression. On the right an interaction matrix of the corresponding network is shown.

genes showing circadian expression. The resulting prediction of potential one-to-one regulator-target interactions is evaluated, again making use of knock-down experiments.

3.1.1 Data-driven inference of gene regulatory networks

Throughout our day networks surround us: In the morning we take a network of roads and streets to work, where we meet with a network of colleagues and friends. Information and ideas are shared on a world-wide network of computers. Even while reading these lines, electrical charges are transmitted through a complex network of neurons and brain structures. On a cellular level, genes and their mRNA and protein products can interact to form molecular-interaction networks. The circadian clock for example is made up of a network of molecular species interacting with one another to produce the characteristic day-and-night rhythms exhibited in gene expression.

To understand and analyse gene expression networks, mathematical formalisms are useful. One such formalism is given by the graph representation of network, with the network being pictured as a set of nodes connected by a number of edges (Figure 3.1A). In the case of gene regulatory networks, nodes can be interpreted as genes or gene products while edges depict regulatory interactions between them. A compressed description of the structural properties of the network is provided by the interaction matrix, in which columns contain potential regulators and rows contain potential target genes (Figure 3.1B). Regulatory interactions between genes in the network are defined as entries in the corresponding fields of the interaction matrix. A value of 1 in a field of the matrix can for example imply the existence of an interaction, while a value of 0 encodes its absence. The degree of detail with which interactions are represented in the network however can vary. Accordingly, interactions can also be symbolized using a range of values, describing either the strength or possibly the likelihood of an interaction.

Various strategies to determine the structure of gene regulatory networks exist. Detailed molecular assays for example can provide evidence for the existence of specific

regulatory interactions. By assessing which genes react to a genetic perturbation, for instance using knock-down experiments, the relative order of genes in the network can be inferred. The existence of *direct* binding of the transcriptional regulator to its target can however not be resolved using a knock-down approach. In this case the analysis of motif binding sites or Chromatin Immunoprecipitation (ChIP) experiments can provide further insight, indicating binding of the regulator within a certain range of the target gene.

As an alternative approach to identify genetic interactions using tools from molecular biology, it is possible to determine the structure of gene regulatory networks using data-based network inference methods. The basic principle of this approach is the idea, that genes sharing similar dynamical behaviour under a variety of measurements obtained under changing experimental conditions, can be assumed to interact in some way or another. Despite network inference being a challenging problem in systems biology, quite a diverse number of approaches with the aim to infer the regulatory structure of a network from experimental data have been developed (reviewed for example in [De Smet and Marchal, 2010, Villaverde and Banga, 2013, Maetschke et al., 2014]).

One of the earliest network inference algorithms, termed relevance networks, estimates the association between genes based on the correlation of gene expression measurements [Butte et al., 2000]. By computing the linear correlation between each pair of expressed genes in a given dataset, a square distance matrix between genes is calculated. Correlations are subsequently removed from the network if they fall below a given threshold, which in turn can be determined by random permutation of the gene expression data. Due to its low computational complexity, relevance networks are capable of handling networks with many thousands of species.

By relying on the correlation coefficient as a measure of association between genes however, a method implicitly assumes linearity and continuity in the relationship between expression patterns of different genes. Due to the non-linear and complex nature of transcriptional gene regulation, this assumption is almost never fulfilled. Another set of network inference methods therefore identifies interactions based on the mutual information shared between genes (reviewed in [Villaverde et al., 2013]). Mutual information is defined from the individual and joint probability distribution of two or more genes, and has the advantage that it provides a measure of dependence free from linearity or continuity assumptions.

One particular implementation of a network inference approach making use of mutual information is the method of context likelihood of relatedness (`CLR`) [Faith et al., 2007b]. The distance matrix between genes is calculated based on pairwise mutual information instead of correlation. Further, rather than estimating the threshold for removing interactions from a background distribution of mutual information values obtained after perturbation, only the gene-specific context is considered for threshold calculation. The relevant context in `CLR` is defined by all possible interactions containing either one of the two genes for which mutual information is calculated.

A different network inference algorithm also based on mutual information, termed minimum redundancy networks (`MRNET`), adopts a more stepwise approach [Meyer et al.,

2007]. The algorithm at first selects for each gene the regulator with maximal shared mutual information and adds this interaction to the predicted network. The second potential regulator of a gene however is chosen according to a score, once again maximizing the mutual information between regulator and target, but also minimizing the mutual information of the proposed regulator and previously selected regulators. By adopting this strategy, the approach balances the trade-off between selecting strong interactions in the network and removing associations which are potentially introduced by indirect interactions.

Spurious associations of genes caused by indirect interactions between genes present a substantial challenge faced by network inference methods. One network inference method accounting for the problem of falsely identified indirect interactions is Algorithm for the Reconstruction of Accurate Cellular Networks (`ARACNE`) [Margolin et al., 2006]. The procedure of `ARACNE` is based on the data processing inequality principle, which states that the association between two genes introduced via an indirect interaction must be smaller than the association between genes which are linked directly. Similar to other network inference methods, `ARACNE` first calculates the mutual information matrix and removes interactions beneath a previously estimated threshold. It then proceeds by identifying triplets (circular regulation between three genes) in the inferred network and removing the weakest link, based on the assumption that this association is caused only by indirect regulation.

In all approaches mentioned so far, the calculated interaction matrix, upon which network inference tools base their structural prediction, is symmetric. This implies that from an inferred interaction it is impossible to discriminate which of the genes is the regulator and which the target. In order to resolve the directionality of interactions, time-course data is particularly useful. The main advantage of time-course gene expression data, is that mutual information between two genes will increase if target gene expression is delayed relative to gene expression of its regulator, due to delays caused by mRNA transcription and processing.

One approach making use of time-resolved gene expression data to infer directionality in interactions is `PREMER` [Villaverde et al., 2014, Villaverde et al., 2017]. In the strategy adopted by `PREMER` a measure related to mutual information, the relative reduction of entropy in the target variable with respect to knowledge of the input variable, is estimated. Additionally, discrete time-shifts between regulator and target gene expression are introduced and taken into account while selecting the regulators which maximize the relative entropy reduction for each target gene. Another advantage of `PREMER` is its ability to detect higher order or synergistic interactions based on three-way mutual information: While often interactions are inferred based on the expression patterns of two genes only, the calculation of higher order mutual information specifically allows for the detection of combinatorial regulation between genes.

Although not based on mutual information, the network inference approach `GENIE3` (GEne Network Inference with Ensemble of trees), is also capable of inferring higher order interactions [Huynh-Thu et al., 2010]. The `GENIE3` method assumes that the expression of a gene in the network can be represented as a function of all other genes present in the

dataset. Using a feature selection algorithm based on decision trees, the regulators most predictive for the expression of each gene can be determined.

Model-free methods of network inference, such as those based on mutual information or feature selection, do not make any assumptions about the form of the relationship between the expression pattern of genes. This makes them easily generalizable to other contexts, such as chemical, ecological, or social networks. Model-free network inference methods generally scale well and are applicable to networks containing many thousand of nodes. Model-based network inference methods on the other hand are formulated explicitly for a certain type of network. In the case of gene regulatory networks they make use of known mechanistic and kinetic properties of transcriptional regulation, making up for the loss in generality by providing a more straightforward interpretation of regulatory interactions. A useful mathematical framework for model-based network inference methods are ordinary differential equation (ODE) based dynamical models. In ODE models the change in activity or expression of a gene in the network is modelled as a function of the expression of potential regulators present in the network, with model equations typically combining both a production and degradation term.

The `Inferelator` approach for instance models transcriptional regulation for each gene as a truncated linear function, where the change in expression of the target gene is linearly dependent on the expression of potential upstream regulators [Bonneau et al., 2006]. Gene expression models are fit to time-course or steady state gene expression data, while at the same time the number of regulators controlling each target gene is minimized via a regularization approach. Although is has been shown that the `Inferelator` method shows good performance across various test cases, the methods assumes linearity in the functions describing transcriptional regulation.

In this chapter we therefore investigate the utility of non-linear ODE models of transcriptional regulation to infer the structure of gene regulatory networks. In our approach, low-parametric models of one-to-one regulation between regulator and target gene are formulated for each possible pair of genes and solved numerically to obtain a simulated time-course of the target gene expression. The model is fit to available expression data by minimizing the weighted difference between the model simulation and the data. Finally, the feasibility of interactions is evaluated by ranking interactions according to the ability of the model to explain the data. The method, which we term `NodeInspector`, is tested using a number of *in-silico* benchmark networks with known regulatory structure and its performance is compared to other existing network inference strategies. After its evaluation, `NodeInspector` is applied to NIH 3T3 gene expression data and potential regulatory interactions between circadian genes are tested. Using a knock-down approach, regulator-target interactions predicted by `NodeInspector` are further evaluated experimentally.

3.2 Results

3.2.1 Formulation of low-parametric dynamical models of transcriptional regulation

Per definition, model-based methods to infer gene regulatory networks require the formulation of a mathematical model representing the kinetic properties of transcriptional and post-transcriptional gene regulation. The regulation function can be derived from thermodynamical considerations taking into account the structure of enhancer promoter systems and cooperativity between transcription factor and co-factor binding [Bintu et al., 2005b, Bintu et al., 2005a]. As a result, transcriptional regulation is often modelled by sigmoid or Hill-type functions dependent on the transcriptional input [Santillan, 2008], where the transcriptional output of a gene can be scaled between 0 and its maximum value. Naturally, processing of mRNA and its translation into a functional protein, acting as the transcriptional regulator, will impact the dynamics of gene regulation. Often however, because of experimental and practical limitations, maturation of mRNA is neglected in models of transcriptional regulation, and mRNA expression is used as a proxy for the expression of its functional protein.

While formulating the `NodeInspector` approach three different regulation functions are considered and compared. The first function, formulated by Reinitz and Sharpe (1995) [Reinitz and Sharp, 1995], describes transcriptional dynamics of gene a regulated by gene b as a sigmoid function dependent on four parameters:

$$\frac{dy_a(\theta,\ y_b)}{dt} \ = \ R_{tc}\ \frac{1}{2}\left(\frac{wy_b + h}{\sqrt{(wy_b + h)^2 + 1}} + 1\right) \ - \ \lambda_{mRNA}\ y_a \tag{3.1}$$

In the above model, y_a and y_b describes the expression of gene a and gene b respectively. The maximum transcription rate of gene a is given by the parameter R_{tc}, while w represents the regulatory weight of gene b on gene a. Note that a negative value of w leads to inhibition of gene a. The parameter h of the production term describes the sensitivity of the promoter of gene a to changes in expression of b. In addition to production of gene a in the model, its linear degradation is assumed proportional to the degradation rate λ.

As a second case we consider a Hill-type regulation function. The Hill-function was first formulated based on mass-action kinetics in order to describe the binding of oxygen binding to haemoglobin [Hill, 1910], but has since been widely used to model cooperative binding phenomena, including TF binding to the promoter. In the context of ligand binding, the Hill-function contains two parameters relating to the affinity of ligand and receptor (k) and cooperativity in binding (n). In order to adapt the Hill-function to transcriptional regulation, parameters for maximum transcription rate (R_{tc}) and mRNA degradation (λ_{mRNA}) are included:

$$\frac{dy_a(\theta,\ y_b)}{dt} \ = \ R_{tc}\ \frac{y_b^n}{y_b^n + k^n} \ - \ \lambda_{mRNA}\ y_a \tag{3.2}$$

As a third option to model transcriptional regulation between genes we consider an ODE system explicitly taking into account translation of mRNA into protein. As before, transcriptional regulation is described using a Hill-function (Equation 3.2). Then, in order to model protein translation from mRNA, a second term describing linear protein production and degradation is added to the system of equations:

$$\frac{dp_a(\theta, y_a)}{dt} \ = \ R_{tl}\ y_a - \lambda_{prot}\ p_a \tag{3.3}$$

Here p denotes protein expression, while R_{tl} and λ_{prot} describe protein production and degradation rates respectively.

In order to determine the transcriptional input into genes, in all models gene expression of the regulator is interpolated between measured data-points. Since we further assume gene expression to be unchanged in the absence of any external perturbation, production and degradation rates need to be balanced at the beginning of an experimental time-course. Therefore, by expressing production rates R_{tc} and R_{tl} as a function of the other parameters, the total number of model parameters can be reduced. Both the regulation function adopted from Reinitz and Sharpe and the Hill-function without protein production therefore contain three unknown parameters, while adding a protein translation term to the Hill-function results in a model with four parameters.

Typically network inference methods generate a structural prediction of a network by returning a ranked list of all possible interactions in the network. In the `NodeInspector` work-flow, models of transcriptional regulation are fit to time-course gene expression data by minimizing the weighted least-squares distance (χ^2-value) between modelled and measured gene expression (see Chapter 2). Since the concept of model and genetic interaction are interchangeable in this context, the likelihood of interactions can be determined by ranking them based on the quality of fit between modelled and measured gene expression data.

3.2.2 Utility of NodeInspector in predicting the structure of benchmark networks from time-course data

In annual competitions, such as the Dialogue for Reverse Engineering Assessment and Methods (DREAM) project [Marbach et al., 2010a], the performance of network inference algorithms is evaluated using computationally generated networks with known structure. One tool to generate gene expression data from *in-silico* gene regulatory networks is

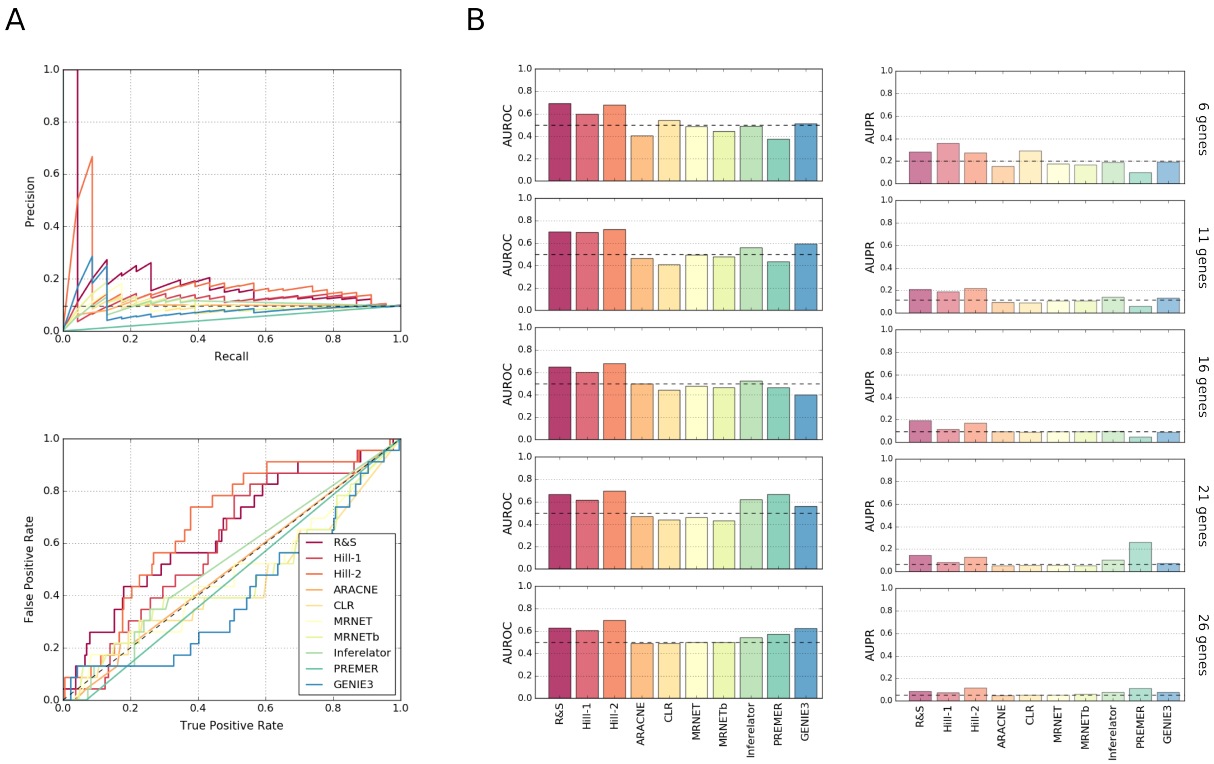


Figure 3.2: Performance evaluation of network inference methods on time-course gene expression data. (A) Exemplary AUROC and AUPR curves for different network inference methods applied to the benchmark network with 16 genes. The `NodeInspector` approach using the regulation function given in Reinitz and Sharpe (1995) (Equation 3.1) is denoted by R&S. Hill-1 and Hill-2 represent the `NodeInspector` approach implemented using Equation 3.2 and Equation 3.3 respectively. **(B)** Barplots of AUROC and AUPR performance values of network inference methods applied to gene regulatory networks of different size (6, 11, 16, 21, 26 genes). Black dashed line indicates the average performance of a random classifier.

`GeneNetWeaver` (`GNW`), which is designed to extract biologically plausible networks of different sizes from *E. coli* or Yeast transcriptional networks [Schaffter et al., 2011a]. After converting extracted sub-networks to dynamical models, a variety of experimental data types can be simulated, ranging from steady-state to time-course data under different perturbation conditions. Using the `GNW` program, we extracted 5 networks of different size (6, 11, 16, 21, or 26 genes) from the provided *E. coli* transcriptional network and generated quantitative mRNA time-course expression data for each of the networks, assuming a perturbation in one of the genes at the beginning of the time-course. We then applied a variety of existing network inference methods as well as the proposed `NodeInspector` approach to this simulated time-course gene expression data. As a result, for each of the extracted sub-networks, ranked predictions of the likelihood of interactions were generated, with each network inference method producing a different ranking.

Rankings generated by each network inference approach were evaluated using standard performance measures of binary classification. Similarly to binary classifiers, which group items into two separate classes, network inference may group the set of potential interactions into two categories: existing (positive) or non-existing (negative) interactions. By

comparing predicted and true interactions, the performance of network inference methods can be evaluated. Precision for example describes the fraction of true interactions in all predicted interactions, while recall is related to the number of existing interactions correctly recovered by network inference.

As typically network inference methods produce ranked lists of interactions instead of definite groups, different sets of predicted interactions can be created based on an altered threshold applied to the ranking. Each applied threshold will therefore result in different values of precision and recall. Choosing a high threshold, very few high-confidence interactions will be predicted resulting in high precision and low recall. While lowering the threshold more interactions will be predicted. Accordingly, in this case precision will decrease while recall increases. This trade off between precision and recall can be visualized in the precision and recall curve (Figure 3.2A). In order to obtain a single measure of performance of a network inference method, the Area Under the Precision and Recall curve (AUPR) can be calculated. A perfect prediction of the structure of the network will result in an AUPR value of one, while the average performance of a random prediction is determined by the ratio of existing interactions to all conceivable interactions in the network.

A different but related measure of performance calculates the Area under Receiver Operating Characteristic (AUROC), which represents the area beneath the false positive rate plotted against the true positive rate. Although it has been argued that AUPR curves supply the superior measure to evaluate binary classification methods, especially in imbalanced datasets with a low number of true interactions [Saito and Rehmsmeier, 2015], we calculated both AUPR and AUROC values for the network inference methods applied to gene regulatory networks of different size.

As a result of our performance evaluation, in most cases all three versions of `NodeInspector` performed better than a random prediction of the gene regulatory network, with increased performance in networks with fewer number of genes (Figure 3.2B). With regard to AUROC values `NodeInspector` further showed superior performance compared to all other tested network inference methods. In most cases this superior performance was also visible based on AUPR values.

`NodeInspector`, in contrast to the other network inference methods considered here, also returned additional information about the sign of an interaction, whether it is activating or inhibiting. We therefore further tested for all the true interactions in the benchmark networks, how often the sign of an interaction was correctly predicted by `NodeInspector`. For the smallest benchmark network, the signs of all six existing interactions matched between the benchmark network and the `NodeInspector` prediction Figure 9.9. In the larger networks the fraction of correctly predicted signs of interactions ranged between 85% and 92%, adding even more weight to the superior performance of the `NodeInspector` approach in comparison to tested methods.

3.2.3 NodeInspector predicts potential regulatory interactions from NIH3T3 expression data

After having established its positive performance, we applied `NodeInspector` to a real biological dataset with the intention to identify potential interactions between circadian genes. Previous theoretical work on the circadian clock established qualitative networks of clock-controlled genes [Bozek et al., 2009, Yan et al., 2008], or formulated mechanistic models of the core clock, mainly based on prior knowledge [Hughes et al., 2009, Ueda et al., 2005, Korenčič et al., 2012, Tovin et al., 2012]. Here we focus on a network reconstruction approach, in which we ask in an unbiased way whether the cyclical genes identified from the RNA-Seq data (See Chapter 2) could, in principle, be regulated by one of the core clock factors or any other cyclically expressed transcription factor.

Since evaluating the complete network of 49 circadian expressed genes would result in testing on the order of 2000 interactions we instead used `NodeInspector` in a query-based manner, where the feasibility of specific individual interactions is tested, rather than generating a ranked prediction for all possible interactions in the network. With this aim in mind, we adopted a parametric bootstrap approach in which the likelihood of each model/interaction was assessed based on the background distribution of χ^2-values generated under the assumption that the regulatory model is in fact true [Johansson et al., 2014]. Following this argumentation, time-courses were re-sampled 1000 times from modelled gene expression simulated by the initial model in combination with estimated parameters. The distribution of expected χ^2-values under the assumption that the initially fitted model is true, was calculated by refitting the model to re-sampled gene expression time-courses. By comparing the quality of the original model fit with this distribution of χ^2-values, the likelihood with which it can be expected to measure a given gene expression time-course under the assumption that the tested model is in fact true was calculated (p-value). In order to evaluate tested interactions, we interpreted a p-value above 0.05 as the possibility of the interaction being true, while a p-value below 0.05 implied its absence.

Since the transcriptional regulators Zfp28 and Leo1, which we studied in more detail in Chapter 2, showed distinct phases of oscillation, we first considered regulation of these genes by core clock genes. Leo1 showed cyclical expression in phase with Bmal1 in many tissues. Similar to the expression of Bmal1, Leo1 could be regulated by ROR factors, since its promoter contains a RORE motif (AACTA GGTCA; 66 bp upstream). Moreover, in livers of Rev-ErbA-/- mice, Leo1 is de-repressed [Fang et al., 2014], suggesting that, additionally, Rev-ErbA (Nr1d1) represses the transcription of Leo1. Nr1d1, Nr1d2, and Rora have also been found to bind to the Leo1 promoter in liver cells and macrophages [Fang et al., 2014, Cho et al., 2012, Bugge et al., 2012, Lam et al., 2013]. Another possibility is that Leo1 is regulated by the cryptochrome genes, which have been shown to bind to the intron of Leo1 in liver cells [Koike et al., 2012]. Accordingly, we designed ODE models of these selected interactions, with transcriptional regulation modelled as a Hill-function combining both mRNA transcription and protein translation (Equation 3.2 and Equation 3.3), and applied the above mentioned modified `NodeInspector` approach. Analysis of

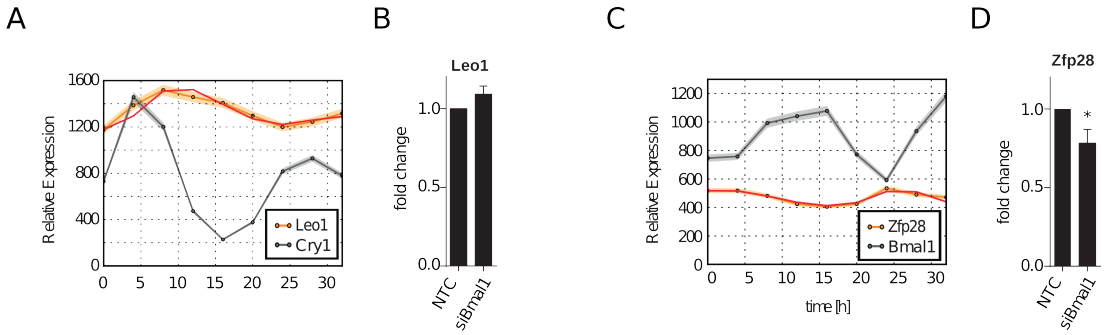


Figure 3.3: Potential regulators of Leo1 and Zfp28 tested by model evaluation. (A) RNA-Seq data and simulation of the model considering Leo1 regulation by Cry1. Gray color represents expression of the input gene, while orange shows expression of the target gene and the corresponding standard deviation (shaded area). In red the model simulation resulting from the best fit is plotted. **(B)** Fold change of Leo1 expression in NIH 3T3 Per2:luc cells treated with siBmal1 for 3 days, compared to control cells (NTC), measured by RT-qPCR (n = 4, normalized to Hsp90ab1 and NTC control [$\Delta\Delta$CT]). Error bars represent standard errors of the means (SEM). No significant change was observed (Mann-Whitney U-test). **(C)** RNA-Seq data and simulation of the model considering Zfp28 regulation by Bmal1 (Arntl). Colour code is equal to the one presented in (A) **(D)** Fold change of Zfp28 expression in NIH 3T3 Per2:luc cells treated with siBmal1 for 3 days compared to control cells (NTC), measured by RT-qPCR (n = 4, normalized to Hsp90ab1 and NTC control [$\Delta\Delta$CT]). Error bars represent the SEM. A significant change was observed (Mann-Whitney U-test, p = 0.011).

the proposed models showed that Leo1 is potentially regulated by Cry1 but not by Cry2 alone (p-value < 0.05, Figure 3.3A; see also Table 9.1). Based on a p-value threshold of < 0.05, regulation of Leo1 by Rev-ErbA (Nr1d1) without contributions from other factors is possible but not likely. Regulation of Leo1 by Rev-ErbB (Nr1d2) or Bmal1 alone was rejected. Indeed, in Bmal1 knock-down experiments, Leo1 expression was not affected (Figure 3.3B).

Zfp28 is cyclically expressed with typical Bmal1 downstream targets, indicating that it might be regulated by Bmal1. Clock and Bmal1 were found to be enriched at the promoter of Zfp28 in liver [Rey et al., 2011, Annayev et al., 2014]; also, Nr1d1, Nr1d2, and Per2 showed binding to the Zfp28 promoter [Koike et al., 2012, Lam et al., 2013]. The models considering regulation of Zfp28 by Bmal1, Cry1, Cry2, Nr1d1, or Nr1d2 could indeed not be rejected, implying that one of these core clock factors might establish transcriptional activation of Zfp28. Of all tested regulators, Bmal1 was considered the most likely (Figure 3.3C and Table 9.1), and Bmal1 knock-down experiments revealed a significant down-regulation of Zfp28 (Figure 3.3D), showing regulation of Zfp28 by the core clock.

Encouraged by these promising predictions, we formulated a larger number of regulatory models, testing if circadian expressed genes identified from the RNA-Seq data could in principle be regulated by selected regulators. Nr1d1 (Rev-ErbA) constitutes a core clock factor as part of the ROR/Bmal1/Rev-Erb feedback loop. Out of the set of cyclically expressed genes, 20 genes contained a Rev-ErbA motif (GTAGGTCACTGGGTCA) in their promoter. Among these 20 genes, 15 have been found to be bound by Nr1d1 in

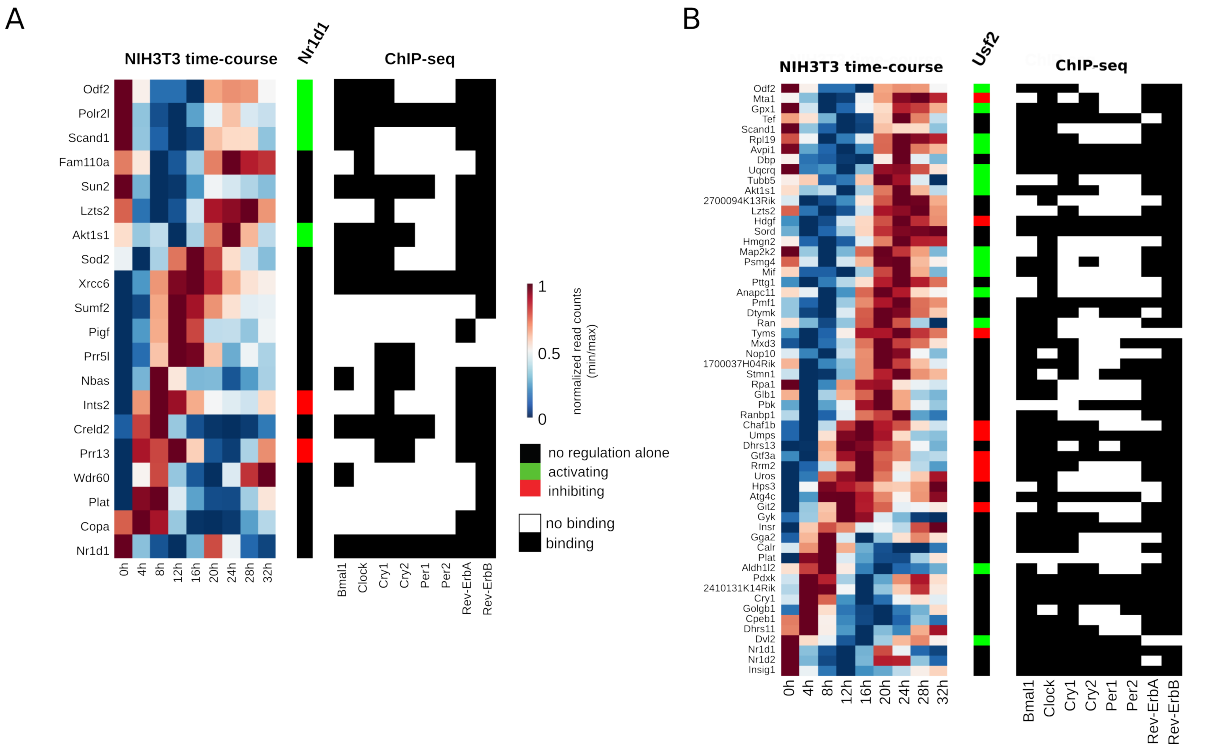


Figure 3.4: Model evaluation of potential Nr1d1 or Usf2 targets. (A) Heatmap of potential Nr1d1 (Rev-ErbA) targets (genes identified by three out of four methods as cyclical and containing an Rev-ErbA motif in their promoter, minimum-maximum normalized, sorted by estimated phase given by `JTK_CYCLE`). The middle panel indicates whether the regulation of this gene by Nr1d1 alone was rejected (black) or not (red/green). Red represents inhibition of the target gene by Nr1d1, whereas green indicates activation. Columns on the right show binding of core clock factors to target genes (black) in at least one of the analysed ChIP-Seq datasets of the respective factor. **(B)** Heatmap of potential Usf2 targets (genes identified by three out of four methods as cyclical and containing an Usf2 motif in their promoter, minimum-maximum normalized, sorted by estimated phase given by `JTK_CYCLE`). Middle and right panel are designed according to (A).

liver tissue or macrophages (Figure 3.4A) [Cho et al., 2012, Lam et al., 2013]. All of the remaining potential Nr1d1 targets showed Nr1d1 (Rev-ErbA) binding in at least one of the considered ChIP-Seq studies. Consequently, we formulated regulatory models of the 20 potential Nr1d1 targets with Nr1d1 as regulatory input. As a result, 14 of 20 potential one-to-one interactions with Nr1d1 as the only regulator needed to be rejected, implying that in theses cases additional regulators need to be considered (Figure 3.4A).

Although so far not discussed as a core clock factor, in Chapter 2 Usf2 was identified as a cyclically expressed gene by three out of four methods. Interestingly, the Usf2 binding motif was also enriched in the set of identified circadian genes. In total, we found 87 occurrences of the Usf2 motif distributed among the promoters of 58 cyclically expressed genes. No preference for a specific phase, fold change, or period length was found among these 58 genes compared to genes that did not show the Usf2 motif (data not shown). To answer the question of whether Usf2 alone could explain the expression data of each of the 58 genes containing a Usf2 motif in their promoter, we designed regulatory models for these genes with Usf2 as the main input. Following this analysis, 35 out of the 58 genes

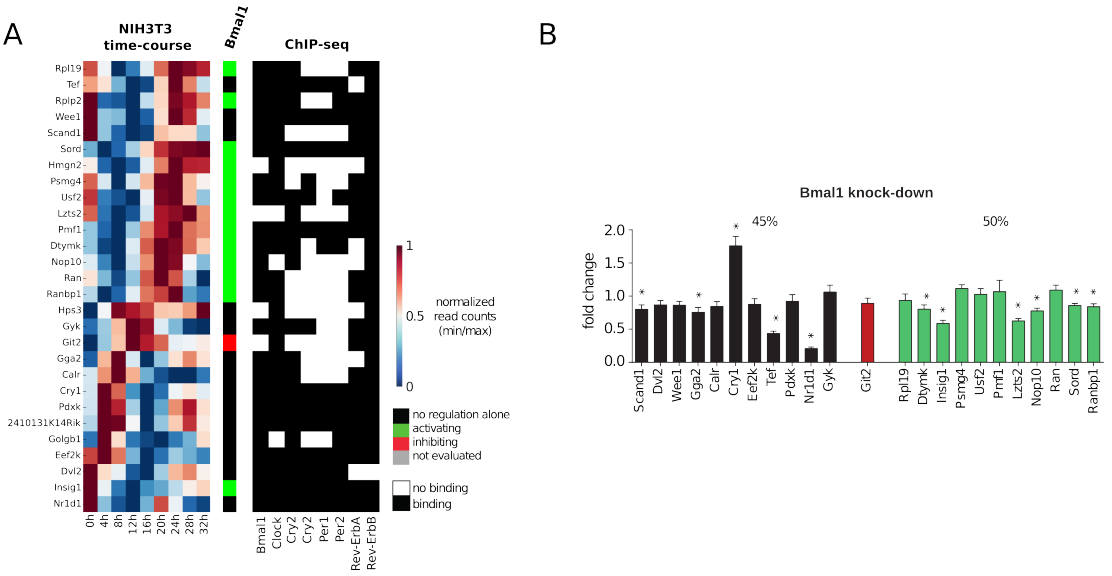


Figure 3.5: Model evaluation of potential Bmal1 targets. (A) Heat map of potential Bmal1 targets (genes identified by three out of four methods as cyclical and containing an E-box motif in their promoter, minimum-maximum normalized, sorted by the estimated phase given by nimJTK_CYCLE). The middle panel indicates whether the regulation of the gene by Bmal1 alone was rejected (black) or not (red/green). Red represents inhibition of the target gene by Bmal1, whereas green indicates activation. Columns on the right show binding of core clock factors to target genes (black) in at least one of the analysed ChIP-Seq datasets of the respective factor. **(B)** Fold change in expression of putative Bmal1 target genes in NIH 3T3 Per2:luc cells treated with siBmal1 for 3 days compared to control cells (NTC), measured by RT-qPCR (n = 4, normalized to Hsp90ab1 and NTC control [$\Delta\Delta$CT]). Error bars represent the SEM. Black bars indicate rejected targets, while green and red bars indicate inhibited or activated potential Bmal1 target genes, respectively. The fraction of deregulated genes in rejected (black) and potential target groups (red and green) are given above the respective bars. Significance was assessed by a Mann-Whitney U-test; *, $p < 0.05$.

with a Usf2 binding motif in their promoter could not be regulated by Usf2 alone, and additional regulators needed to be considered (Figure 3.4B). In principle, the remaining 23 genes are potentially regulated by Usf2 only. These 23 genes are mainly distributed among the genes that peak late during the day-night cycle.

We further tested regulation of cyclical genes by Bmal1, which is an essential factor that participates in both core clock feedback loops. Again only cyclical genes with a corresponding Bmal1 binding site in their promoter (E-box; CCGGTCACGTGA) were considered potential targets. 24 of these 28 genes with the Bmal1 motif also showed Bmal1 binding in their promoter based on ChIP experiments in the liver [Menet et al., 2012, Koike et al., 2012, Rey et al., 2011, Annayev et al., 2014]. Out of 28 potential Bmal1 targets, 14 regulator-target interactions were rejected by our analyses (Figure 3.5).

We evaluated the performance of the network inference approach on real mRNA expression data by performing knock-down of Bmal1 and measuring the mRNA expression of 23 of the 28 genes showing Bmal1 binding. Bmal1 is widely known as a transcriptional activator [Partch et al., 2014] and, consequently, all but one deregulated gene was down-regulated. In accordance with an observation in Bmal1-/- mice [Menet et al., 2012],

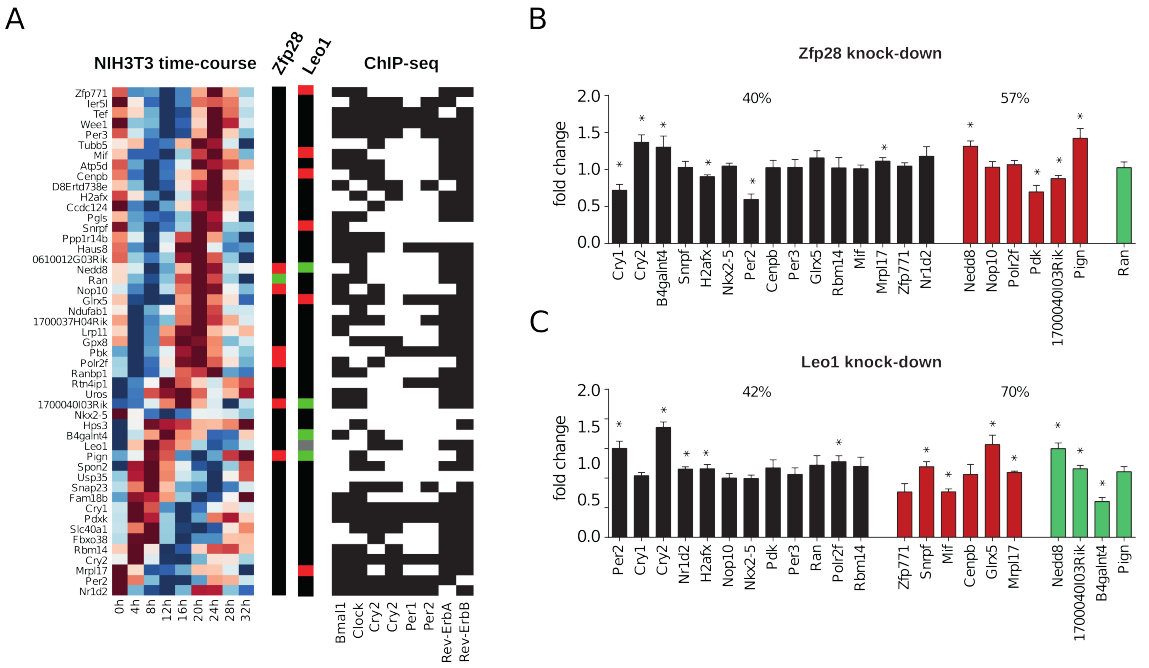


Figure 3.6: Model evaluation of potential Leo1 or Zfp28 targets. (A) Heat map of potential Zfp28 or Leo1 targets (genes identified by all four methods as cyclical, minimum-maximum normalized, sorted by estimated phase given by `JTK_CYCLE`). The middle panels indicate whether the regulation of the target gene by Zfp28 or Leo1 alone was rejected (black) or not (red/green). Red represents inhibition of the target gene by Zfp28, whereas green indicates activation. Gray indicates the model was not evaluated. Columns on the right show binding of core clock factors to target genes (black) in at least one of the analysed ChIP-Seq data sets of the respective factor. **(B)** Fold change in expression of putative Zfp28 (top) or Leo1 (bottom) target genes in NIH 3T3 Per2:luc cells treated with the respective siRNA for 3 days compared to control cells (NTC) measured by RT-qPCR (n = 4, normalized to Hsp90ab1 and NTC control [$\Delta\Delta$CT]). Error bars represent the SEM. Black bars indicate rejected targets, while green and red bars indicate inhibited or activated potential Zfp28 or Leo1 target genes, respectively. The fraction of deregulated genes in rejected (black) and potential target groups (red and green) are given above the respective bars. Significance was assessed by a Mann-Whitney U-test; *, $p < 0.05$.

Cry1 was up-regulated upon Bmal1 knock-down. However, the regulation of Cry1 by Bmal1 was rejected by the modelling analysis, reflecting the fact that additional factors cooperate with Bmal1 in regulation of Cry1 [Ma et al., 2013]. Overall, we found that among the measured target genes 50% (6 out of 12) of predicted Bmal1 targets were indeed differentially regulated upon Bmal1 knock-down (Figure 3.5B, for a measurement of knock-down efficiency see Figure 9.10). However, also 45% (5 out of 11) of targets for which one-to-one regulation needed to be rejected, were also significantly deregulated upon Bmal1 knock-down. Since Bmal1 constitutes a core clock factor and its removal leads to a substantial disruption of the circadian clock, a large impact on the set of tested circadian genes was expected. As a result, potential indirect effects on target gene expression will inevitably mask direct one-to-one interactions and therefore interfere with the assessment of our predictions.

With regard to the transcriptional regulators Zfp28 and Leo1, which we have characterized in more detail, we do not possess information about putative binding sites. Therefore, we chose the 49 genes identified by all four methods as potential target genes of Leo1 or Zfp28 and tested which of these could possibly be explained by one of these factors alone. In the case of Zfp28, almost all putative targets (42 out of 49) could not be explained by Zfp28 alone, and additional regulators needed to be taken into account (Figure 3.6A). For Leo1, this number was only slightly higher; 10 out of the 49 putative targets can be explained by regulation of Leo1 alone. Many potential Zfp28 or Leo1 putative targets showed binding of at least one core clock factor to their promoter, meaning that in principle they could also be directly regulated by the core clock network. However, in knock-down experiments of Zfp28 or Leo1, we found that indeed many predicted targets were deregulated (Figure 3.6). Moreover, in the predicted target group, we found more deregulated genes than in the group of rejected genes, supporting the predictive power of our network inference approach.

3.3 Discussion

In this chapter we presented a newly designed computational method capable of inferring the structure of gene regulatory networks from time-course gene expression data using non-linear dynamical models of transcriptional regulation. Similar approaches have been proposed in the past. In Honkela et al. (2010) simple linear ODEs are used in a Gaussian process frame-work to model transcriptional activation [Honkela et al., 2010]. Similar to our approach, models are fit to measured time-course expression data and ranked according to their quality of the fit. The approach was applied to microarray expression data obtained during *Drosophila* embryogenesis in order to identify potential target genes for the transcriptional regulators Twist and Mef2. Validating predicted interactions by ChIP-experiments, it could be shown that in the presented case ODE based models perform better in identifying genetic interatcion, than simple correlation based methods or methods applied to knock-down data.

Also the ODE-based `Inferelator` method has been tested extensively using data obtained from the archaeon *Halobacterium* NRC-1 and its power to predict global expression level changes from inferred gene expression networks was demonstrated using gene perturbation experiments [Bonneau et al., 2006]. Both mentioned network inference approaches however, make simplifying assumptions about linearity in gene regulation, do not consider translation of regulator mRNA into protein, or restrict themselves to model only activatory regulation.

A recent study by Hillenbrand et al. (2016) addressed the question of non-linearity in gene expression by formulating alternative non-linear ODE-based models of transcriptional regulation. The feasibility of different gene regulation functions was determined by fitting models of transcriptional regulation to dynamic transcriptome analysis data obtained for cell cycle genes in yeast [Hillenbrand et al., 2016]. In this study, also the utility of non-linear ODE models of transcriptional regulation to infer potential interactions between cell-cycle genes was demonstrated. The `NodeInspector` approach formulated here carries on the idea of using non-linear models of transcriptional regulation to infer the structure of gene regulatory networks. Once again, the performance such an approach was illustrated and compared to other state-of-the-art network inference methods using realistic gene expression data derived from benchmark networks of different size.

As an additional test case, we evaluated the potential of non-linear models of transcriptional regulation to infer novel regulatory relationships among cyclically expressed genes identified in Chapter 2. Regarding transcriptional dynamics of circadian rhythm, mathematical models have largely contributed to our understanding of the core clock network and its impact on core clock genes. Several of these studies formulated mechanistic models of (parts of) the core clock and conceptually analysed how different transcriptional regulatory modes among core clock regulators affect the dynamic behaviour of clock-controlled genes [Ueda et al., 2005, Korenčič et al., 2012, Westermark and Herzel, 2013, Korenčič et al., 2014]. Others created a network description of the core clock genes based on large-scale promoter analyses [Bozek et al., 2009] or based on a careful analysis of different TF knock-out or mutant strain datasets as well as binding site predictions [Yan

et al., 2008, Relógio et al., 2011]. Hence, most of these studies draw from prior information provided in the literature, are limited to a conceptual level, or are restricted to only a small number of clock genes.

Our analysis revealed that in most cases, even though a gene contains a specific motif, such as Bmal1, Nr1d1, or Usf2, most likely more than one factor is involved in establishing the transcriptional output, as a large number of the 213 tested interactions had to be rejected. In addition, we have identified cyclical genes which are potentially regulated by Bmal1, Nr1d1, Usf2, Leo1, or Zfp28 alone. We evaluated our predictions by performing knock-down experiments and found that, on average, more potential targets were deregulated compared to targets which needed to be rejected by our computational analysis.

In the test case presented above transcriptional regulation is modelled by Hill-type kinetics, while translation of mRNA into protein is considered linear. However, circadian regulation is also known to occur on post-transcriptional and post-translational levels [Kwak et al., 2006, Morf et al., 2012, Kim et al., 2015, Chen et al., 2014, Kojima and Green, 2015, Beckwith and Yanovsky, 2014, Reddy and Rey, 2014, Ki et al., 2015, Woo et al., 2009, Woo et al., 2010, Lück et al., 2014, Mauvoisin et al., 2015, Mauvoisin et al., 2014, Robles et al., 2014]. As regulatory models are merely heuristics, more realistic functions including mechanisms of post-transcriptional regulation should be tested in the future. This however requires the integration of data from multiple layers of gene regulation, namely mRNA and protein expression [Lück et al., 2014, Mauvoisin et al., 2014, Robles et al., 2014] (see also Chapter 6). As soon as such data is made available, the generality of `NodeInspector` can easily be extended to a scenario in which both protein and mRNA expression data are considered.

In summary, using the `NodeInspector` approach, we demonstrate once again the utility of non-linear ODE models in inferring genetic interactions from time-course gene expression data. Further, predicted regulator-target interactions between circadian expressed genes were validated experimentally, adding to existing knowledge on the structural properties of the gene regulatory network governing circadian rhythm.

Chapter 4

Inferring the Structure of the Gene Regulatory Network Controlling the Epithelial-to-Mesenchymal transition in NMuMG cells

Preamble

This project was carried out in collaboration with SP, SS, AG, and SSu who performed all experiments described in this chapter. Valuable feedback was further provided by VT and SL.

4.1 Introduction

In the previous chapter we demonstrated that low-parametric models of transcriptional regulation, combined with model fitting and model selection, can be used to infer the topology of gene regulatory networks. We applied this approach to evaluate potential one-to-one regulatory interactions between circadian expressed genes and tested predictions experimentally.

In this chapter we extend the formulated network inference approach to a different biological system: The gene regulatory network controlling the early steps of the epithelial-to-mesenchymal transition (EMT). We apply the network inference methods described in Chapter 3 to gene expression data measured under a combination of wild-type and knock-down conditions. Integrating resulting structural predictions by each applied network inference method provides a 'community prediction' of the topology of the network controlling EMT. Finally, predicted interactions compared to known interactions reported in the scientific literature, interactions determined by analysis of transcription factor motif occurrences in the regulatory sequence of each gene, and identification of genetic interactions based on publicly available ChIP data.

4.1.1 TGFβ-induced EMT is important in health and disease

The process of epithelial-to-mesenchymal transition (EMT) describes an important trans-differentiation event, in which cells lose their epithelial characteristics and acquire a more migratory, mesenchymal phenotype. During progression of EMT many properties of the cell undergo a drastic change, including the disassembly of epithelial cell-cell contacts and the loss of cell polarity. The cells cytoskeletal architecture reorganizes and cells acquire motile and invasive capacities.

EMT is essential for numerous developmental processes including mesoderm and neural tube formation, but has also been shown to be indispensable for wound healing. Accordingly, the physiological as well as morphological changes occurring during EMT need to be tightly controlled. The lack of this control of gene expression may critically impact an organisms health, and mis-regulation of EMT can be the source of various diseases, such as organ fibrosis. In the context of cancer metastasis, EMT can be the source of cells disseminating from the primary tumor and travelling to distinct sites in the body. For current reviews on EMT see for example [Lamouille et al., 2014, Nieto et al., 2016, Skrypek et al., 2017].

In many cellular systems, EMT can be induced by stimulation of cells with TGFβ [Valcourt, 2005]. TGFβ is a small extracellular soluble factor that signals through a complex of type I and type II receptors to phosphorylate receptor Smads (Smad2, Smad3), which upon activation bind to co-Smad (Smad4). Trimeric Smad2-Smad3-Smad4 protein complexes then translocate into the nucleus, where they act as transcription factors (TFs) to repress or activate target genes mediating phenotypic changes during EMT. Additional TFs that cooperate with or are induced by Smads make up the gene regulatory network initiating and ultimately driving early EMT.

Judging by the current literature, a large number of genes has been implicated in EMT or EMT related processes. However, most research agrees that the gene regulatory network controlling EMT in various cell systems includes three families of EMT-inducing TFs, Snail1/Snail2, Zeb1/2, and Twist. Depending on the cellular context and research focus, additional factors involved in EMT may be important.

In general the network of genes controlling EMT is thought to be organized in a hierarchical fashion. The TFs of the network are induced by phosphorylated Smad protein either directly or indirectly. Among the TFs there is a substantial amount of cross-regulation. TFs then repress the expression of epithelial markers, while activating the expression of mesenchymal markers. For instance, one hallmark of EMT is the down-regulation of the cell adhesion molecule E-cadherin (Cdh1) and concomitant up-regulation of N-cadherin (Cdh2), also known as the cadherin switch.

The main goal of this chapter is to determine the structure of the regulatory network controlling EMT downstream of TGFβ/Smad-signalling. With this aim in mind, an initial set of known interactions is derived based on experimental evidence reported in the literature. As this set of interactions is still incomplete to permit the full reconstruction of the EMT network, an extensive dataset of mRNA expression of selected EMT related

genes is generated from wild-type and various knock-down cell lines. Based on this data, structural predictions of the gene regulatory network controlling EMT are obtained using data-driven network inference methods. Finally, structural predictions are evaluated using previously identified interactions reported in the literature, as well as interactions derived by testing for the occurrence of TF motifs in the promoter sequence of genes, or analysis of publicly available ChIP data.

4.2 Results

4.2.1 Identification of network interactions based on detailed molecular experiments

We made use of a large body of experimental evidence reported in the literature, in order to identify main factors associated with the genetic program of EMT and interactions between them. Although some factors are frequently discussed in the context of EMT, discrepancies exist between different EMT model systems. For example, it was shown that direct Smad-targets differ largely between different cell types. Based on Smad4 ChIP-Seq experiments, low overlap between Smad4 target genes in three different cell types (A2780, IOSE, and HaCaT), with only one gene bound by Smad4 simultaneously in all three cell types, was observed [Kennedy et al., 2011]. Another example illustrating differences between model systems of EMT is indicated by the fact that knock-down of Snail1 results in a decrease of Cdh1 promoter activity in MDCK cells, while a corresponding effect is missing in NMuMG cells [Shirakihara et al., 2007]. While attempting to reconstruct the EMT network from experimental evidence reported in the literature we therefore mainly focused on NMuMG cells as a standard model for investigating TGFβ-induced EMT. Within 24h of TGFβ-treatment, NMuMG cells show pronounced changes in cell morphology accompanied by expression changes in EMT relevant molecular markers [Miettinen et al., 1994].

Based on an extensive literature review we selected a number of EMT related genes as core genes of the EMT network. This selection of EMT genes included two criteria: First, genes should have been shown to be involved in TGFβ-induced EMT of NMuMG cells and their relevance to EMT needed to be sufficiently established in previous publications. As a second criteria, the expression of genes in NMuMG cells within the first 24h after TGFβ-treatment was required. The expression of EMT genes was assessed using RNAseq data obtained in unstimulated NMuMG cells as well as NMuMG cells 24h after TGFβ-stimulation [Sahu et al., 2015]. The set of genes qualifying for the above outlined criteria included important EMT related epithelial markers Cdh1, Ocln, and Crb3. Also mesenchymal markers Cdh2, Vim, Ncam, and Fn1 were selected. In addition to the critical TFs downstream of TGFβ - Smad2, 3, and 4 - we selected Atf3, Ets1, Hmga2, Id1, Id2, Id3, JunB, Serpine1, Snail1, Sox4, Zeb1, and Zeb2 as core EMT genes (For an overview of selected EMT factors and relevant publications see Table 9.2 and Table 9.3). Although Twist1 and Twist2 are frequently discussed in an EMT context, their expression was not detected in NMuMG cells within the first 24h after TGFβ-treatment.

After having selected a number of EMT related factors, we next determined potential interactions between genes in order to reconstruct the structure of the gene regulatory network controlling EMT. For example, the TF Snail1 is rapidly up-regulated after treatment of NMuMG cells with TGFβ, which may suggest a direct regulation of Snail1 by Smad TFs. Indeed, evidence for a direct regulation of Snail1 by Smads can be found in multiple published experiments [Gervasi et al., 2012, Thuault et al., 2008]. For example it was shown that Smad4 knock-down partially inhibits the TGFβ-dependent increase in

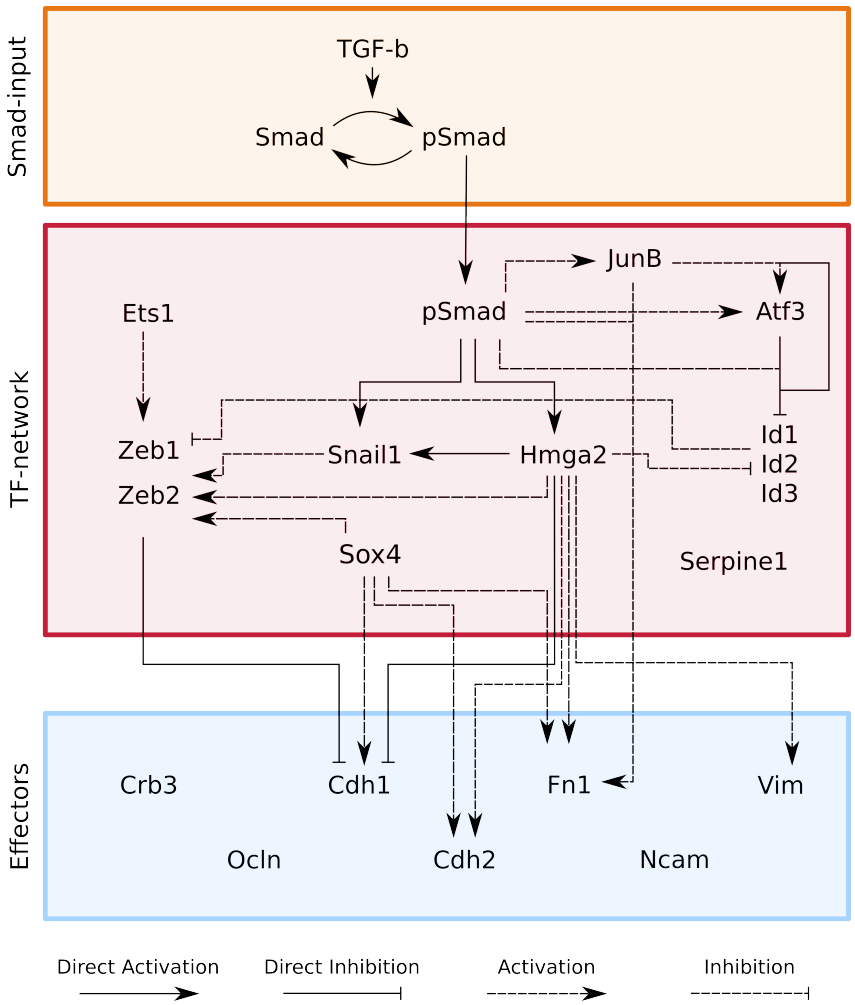


Figure 4.1: Schematic representation of the gene regulatory network controlling EMT. EMT in NMuMG cells can be induced by TGFβ-treatment. The input of the gene regulatory network controlling EMT consists of TGFβ-dependent phosphorylation of Smad proteins, which translocate to the cell nucleus to regulate potential target genes. TFs cross-regulate one another and impact downstream effector genes. Interactions between genes in the network can be assigned using experimental evidence reported in the literature. In some cases this evidence indicates a direct regulation of the target by its regulator (solid lines), in other cases evidence of direct binding is missing (dashed lines).

Snail1 mRNA expression. Further, Smad3 and Smad4 co-expression can induce Snail1 promoter activity as well as protein expression in the absence of TGFβ-stimulation. Interestingly, the increase in Snail1 activity is enhanced even further by co-expression of Smad3 and Smad4 with the TF Hmga2. Using promoter deletion constructs, it was possible to demonstrate that a -170 to -110bp promoter region relative to the Snail1 transcription start site is responsible for the Smad/Hmga2 dependent activation of Snail1. Finally, chromatin immunoprecipitation experiments indicate TGFβ-dependent binding of Smad4 and Hmga2 to the Snail1 promoter region. From these experiments, we concluded that Smads and Hmga2 cooperatively and directly regulate Snail1 expression during TGFβ-induced EMT in NMuMG cells (Figure 4.1).

In addition to Snail1, Hmga2 may also be directly regulated by TGFβ/Smad-signalling. Accordingly, Hmga2 mRNA is up-regulated 2h after TGFb treatment but decreases again

after 36h [Thuault et al., 2006]. Cycloheximide, which is a protein synthesis inhibitor, is not able to block the TGFβ-dependent increase in Hmga2, hinting at a direct impact of TGFβ-signalling on Hmga2 transcription. Indeed, expression of a dominant negative form of Smad2 blocks Hmga2 mRNA induction and promoter activation by TGFβ. Finally, it was shown that Smad4 binds to multiple regions in the Hmga2 promoter, hereby establishing the direct regulation of Hmga2 by Smad proteins.

As EMT is mainly a morphological event, also effector genes affecting cell structure need to be differentially regulated during the transition. For example, Cdh1 expression is directly regulated via Hmga2. Accordingly, in cells ectopically expressing Hmga2, strong association of Hmga2 with the Cdh1 promoter was observed [Tan et al., 2014]. By measuring various activating or repressing histone marks in the Cdh1 promoter region, it was shown that Hmga2 together with Snail1 and Twist1 participates in the epigenetic silencing of Cdh1.

The above presented case studies represent only a fraction of experimental evidence hinting at regulatory interactions present in the EMT network. Additional experimental evidence from which further interactions can be derived are listed in supplemental material 9.3.3. In contrast to evidence pointing at a *direct* regulation of a target gene by the respective regulator, in some cases merely indirect regulation may be established based on the limited experimental data. Shirakihara et al. (2007) [Shirakihara et al., 2007] for example show differential regulation of Zeb1, Zeb2, and Cdh1 upon knock-down of Ets1 in NMuMG cells. A direct association of Ets1 with these respective targets however was not tested. As a consequence, regulation of the target genes by Ets1 could in principle proceed via additional factors.

Clearly, evidence of regulation in such cases is less informative compared to evidence of direct binding. In the following we therefore refer to evidence of *direct* binding as grade A evidence, while all other evidence concerning genomic interactions, which does not allow the inference of a direct regulation, is referred to as grade B. In total, from our literature search we were able to extract evidence for the existence of 8 direct interactions (grade A - Figure 4.1 and Table 9.4). In additional 23 cases, experimental evidence pointed to some form of regulation between two genes, but it was not possible to determine whether this regulation was direct or indirect (grade B).

In summary, the gene regulatory network regulating EMT could partially be inferred from available literature evidence, however critical information on the regulation of important EMT factors is still missing. A complete reconstruction of the structure of the gene regulatory network controlling EMT entirely based on experimental evidence reported in the literature therefore was not possible. As an alternative strategy to determine the regulatory interactions present in the EMT network, we apply state-of-the-art network inference tools as outlined in Chapter 3 to gene expression data derived from TGFβ-stimulated NMuMG cells.

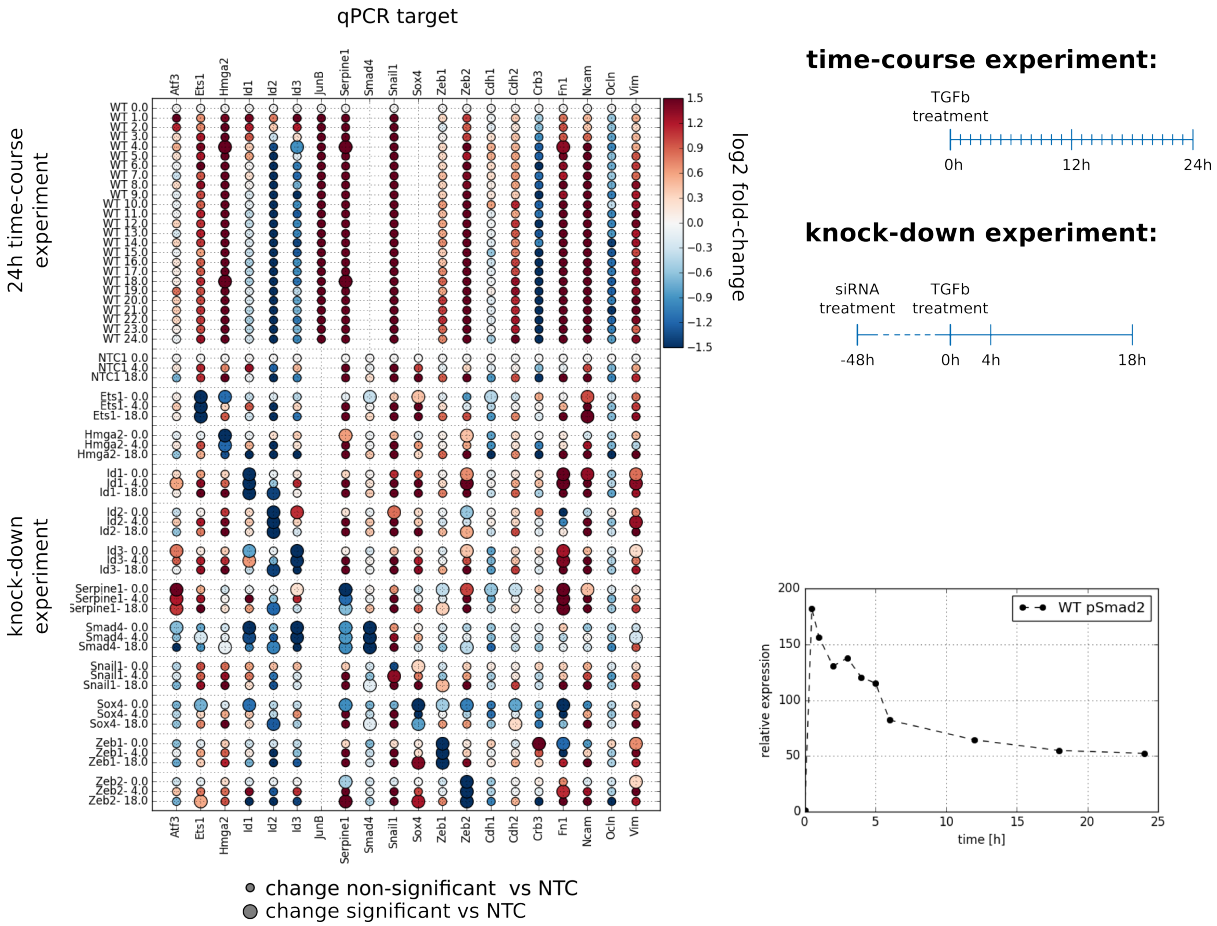


Figure 4.2: Expression of EMT related genes in TGFβ-stimulated NMuMG cells. In the 24h time-course experiment, expression of EMT related genes was measured by RT-qPCR each hour for 24h after TGFβ-stimulation. In the knock-down experiment, cells were treated with TGFβ 48h after siRNA mediated knock-down of EMT relevant TFs, and selected EMT genes measured at 0h, 4h, and 18h after TGFβ-treatment. Large dots correspond to a statistical significant difference between target gene expression in knock-down cells relative to non-targeting control (NTC) cells at the same time-point (independent t-test, $p < 0.025$) The lower right corner shows expression of pSmad2 in TGFβ-treated wild-type cells measured by western blot relative to expression at the 0h time-point.

4.2.2 Gene expression time-course and perturbation data of TGFβ-stimulated NMuMG cells shows good reproducibility

In order to infer the EMT network using data-driven network inference methods, gene expression for the 20 selected EMT genes were determined by RT-qPCR. In a first experiment, we measured gene expression in wild-type cells at each hour after TGFβ-stimulation over a period of 24h (Figure 4.2). An additional experiment provided RT-qPCR measurement of EMT related genes at 0h, 4h, and 18h relative to TGFβ-treatment in NMuMG cells, in which one of 11 EMT relevant TFs was knocked down using siRNA. Out of the 20 EMT genes, the expression of 17 genes was determined both in the time-course and perturbation experiment. However, JunB was exclusively measured in the time-course experiment, while Smad4 and Sox4 were measured only in knock-down experiments.

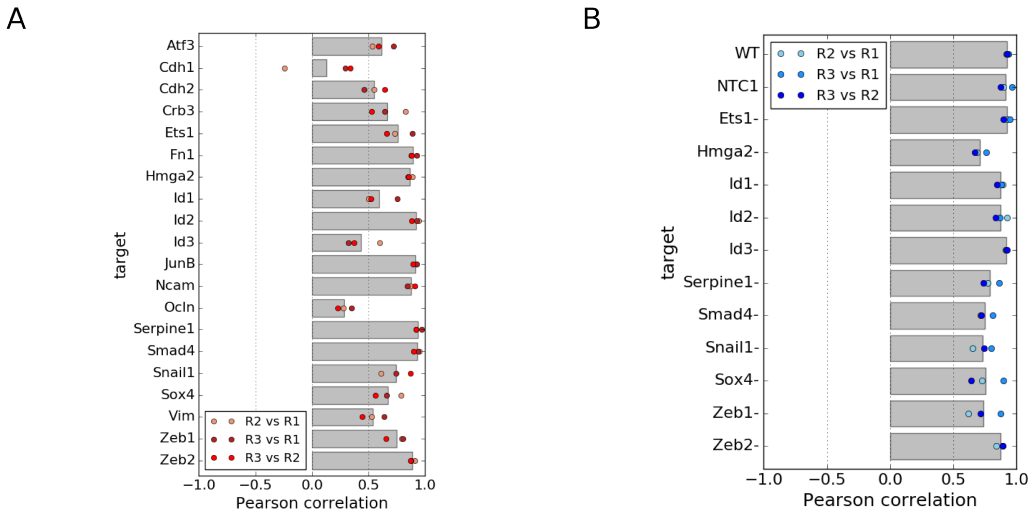


Figure 4.3: Reproducibility of qPCR measurements. Shown are Pearson correlation coefficients calculated between the three different biological replicates of RT-qPCR measurements, either for each target gene across the different cell-lines **(A)** or each cell-line across the different target genes **(B)**.

According to genome-wide RNA-Seq data, the expression of Smad2, Smad3, and Smad4 does not change within the first 24h after TGFβ-stimulation (Table 9.2). Therefore, in order to more accurately determine the transcriptional input into the gene regulatory network, a Western blot of pSmad2 expression at various time-points after TGFβ-treatment was quantified in wild-type cells. In total, the combination of all three experiments - RT-qPCR time-course and perturbation data, as well as measurement of pSmad2 by Western blot - amounts to a number of 1070 unique data points.

RT-qPCR experiments depend on the normalization of target genes to a set of reference genes, whose expression in an ideal case should not change across the measured conditions. As reference genes we selected Rpl19 and Tbp, which merely exhibited significant differential regulation compared to the control sample in two experimental conditions (Snail1- at 0h and Id3- at 0h - independent t-test 2-sided, $p < 0.025$). Although, as demonstrated these gene expression changes are significant, they are associated with fold-changes below 25%.

In addition to the quality of reference genes, we checked the reproducibility of experiments between the three performed biological replicates. Indeed, qPCR experiments showed good agreement with a mean Pearson correlation coefficient of 0.92 between log-normalized fold-changes of the time-course experiment or 0.82 between log-normalized fold-changes of the knock-down experiment. Over all data points measured in both experiments, the mean Pearson correlation coefficient between individual replicates was 0.87 (Figure 9.11).

Correlation between replicates however may vary greatly between the individual genes or cell lines. In general however, we observed good correlation between biological replicates if only single target genes were considered. The largest overall correlation was reached by Serpine1 (0.93) while the lowest correlation was exhibited for Cdh1 (0.13 -

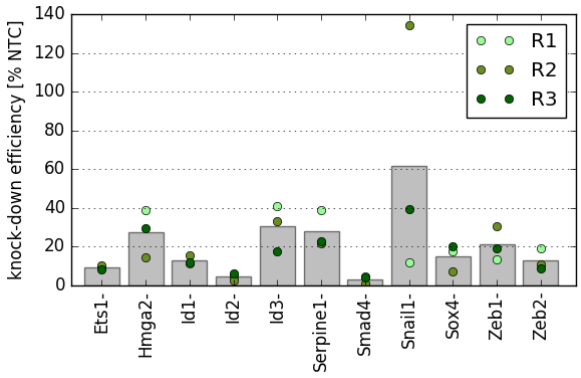


Figure 4.4: Knock-down efficiency in siRNA treated NMuMG cells. Barplots show the level of target gene expression in siRNA-treated cells relative to cells treated with a non-targeting control siRNA 48h after treatment. Individual replicates are plotted as circles.

Figure 4.3A). In fact, correlation between replicate 2 and replicate 1 for Cdh1 turned out to be negative. Variance in the correlation could for example reflect technical problems related to primer efficiencies. However, in addition to technical issues, varying speeds in the dynamics of gene expression may impact the assessment of correlation: Meaningful evaluation of correlation between experimental replicates for genes displaying fast dynamics depends critically on temporally precise measurements, while genes exhibiting slower gene expression changes may even indicate good correlation in the presence of variability in the time of measurement. Genes showing no significant trend but only random variation around a mean value, as is the case for the majority of Cdh1 measurements, will also exhibit low correlation.

Similar to good agreement between biological replicates on the level of single genes, the we also observed good correlation when comparing all genes measured in a given cell (Figure 4.3 right). Overall our results indicate that experiments show good reproducibility across the three performed biological replicates.

In knock-down experiments it is important to ensure the effectiveness of the siRNA utilized to knock-down individual genes. Knock-down efficiency over all knock-down experiments was in the range between 99% (Smad4-, biological replicate 2) and 60% (Id3-, biological replicate 1) of mRNA depleted compared to control cells (Figure 4.4). Merely knock-down of Snail1 in biological replicate 2 showed insufficient knock-down effect measured by RT-qPCR. However, this lack of knock-down may potentially be caused by large variability in the raw measurement values of Snail1 due to its low concentration. Since effects of Snail1 knock-down in replicate 2 were reproducible compared to other biological replicates, we nevertheless decided to include the data from this biological replicate.

Brightfield microscopy images obtained at 0h and 24h relative to TGFβ-treatment indicated abnormal progression of EMT in cells treated with siRNA against Id1, Id2, Serpine1, Sox4, and Zeb1 while knock-down of Atf3 resulted in reduced cell survival. In knock-down cell lines of the remaining selected EMT factors there appeared no visible effects on EMT within the first 24h of TGFβ-treatment (data not shown).

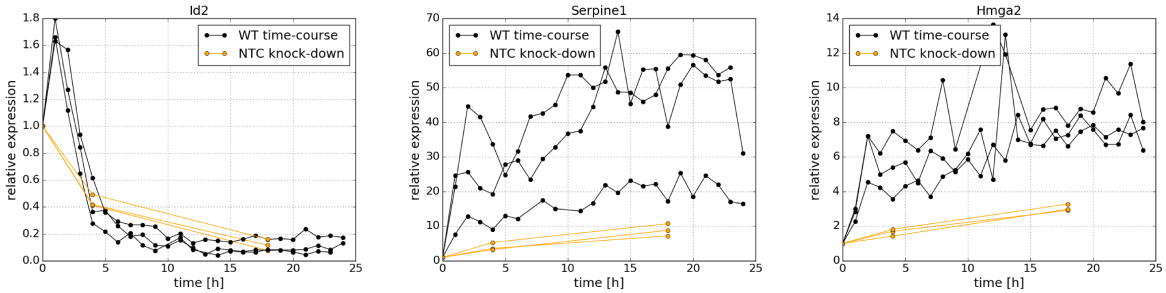


Figure 4.5: Comparison of gene expression between wild-type and NTC-treated NMuMG cells. A comparison of the wild-type time-course (0-24h) and NTC-treated cells (0h, 4h, 18h) is shown for three example genes (Serpine1, Id2, and Hmga2)

In order to assure that wild-type gene expression in time-course and knock-down experiments agree, we checked for correlation of RT-qPCR measurements from the time-course experiment with wild-type and control cells of the knock-down experiment. As a result, wild-type cells of the knock-down experiment matched well with the original time-course experiment at 4h and 18h with a Pearson correlation coefficient of 0.73 (WT1) and 0.79 (WT2). The Pearson correlation between cells with non-targeting siRNA (non-targeting control - NTC) and the wild-type cells from the original time-course experiment was 0.74 for both control cell-lines used. The largest differences between time-course and knock-down experiment was observed for Hmga2 and Serpine1, where the time-course was qualitatively reproduced, but fold-changes in the knock-down were decreased compared to the time-course experiment (Figure 4.5).

When comparing the obtained RT-qPCR time-course data with expression changes reported in the literature, most of the hallmark changes of EMT could be recapitulated. For example, mesenchymal markers Fn1, Ncam, and Vim as well as others showed up-regulation upon TGFβ-treatment, while epithelial markers such as Ocln and Crb3 were down-regulated. Qualitatively changes in EMT genes agree with those reported in the literature, with the exception of Atf3, which observed a sharp but transient up-regulation in our RT-qPCR data, while a 5-fold up-regulation was reported in RNA-Seq data of NMuMG cells even after 24h after TGFβ-stimulation [Sahu et al., 2015]. Strikingly, in our data Cdh1 down-regulation was not observed within the first 24 hours after TGFβ-treatment. This is in accordance with multiple studies, which also demonstrated missing down-regulation of Cdh1 protein [Bakin et al., 2000, Bhowmick et al., 2001, Maeda et al., 2005] as well as mRNA [Gervasi et al., 2012] at 24h in NMuMG cells. Also Chang et al. (2016) showed only minor down-regulation of Cdh1 protein in A549 cells on day 1 after TGFβ-treatment [Chang et al., 2016]. Other studies however reported Cdh1 down-regulation within the first 24h after EMT induction [Miettinen et al., 1994, Piek et al., 1999]. At this point we can only speculate, that these differences arise from differences in TGFβ-treatment, cell-line, or cell culture conditions.

Knock-down experiments which are directly comparable to the knock-down conditions applied in this study are rarely found in the literature. Gervasi et al. 2012 for instance report a number of EMT genes measured 24h after TGFb treatment via qPCR

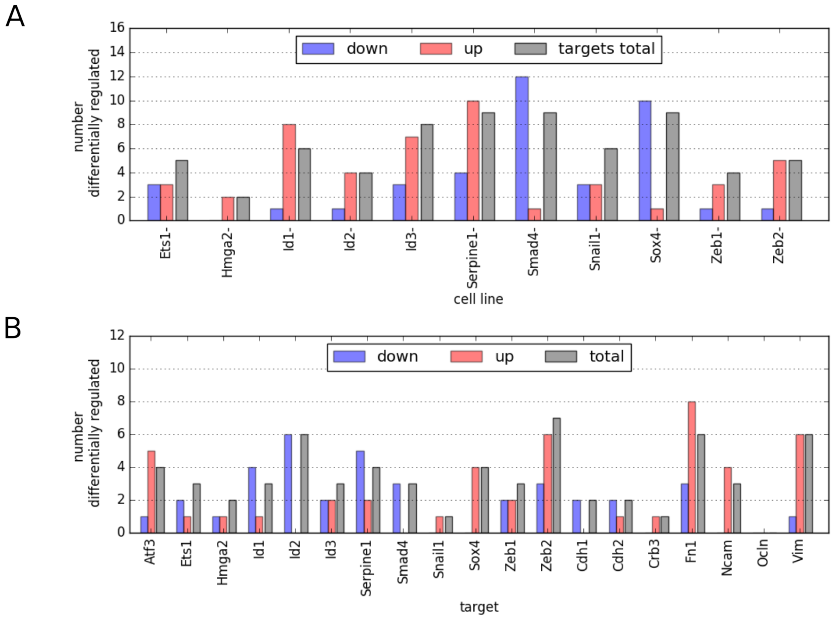


Figure 4.6: Analysis of significant changes in EMT expression data. (A) For each cell-line of the knock-down experiment, the number of significantly different measurements (independent t-test, two-sided, $p < 0.025$) is shown. While the number of samples showing up-regulation compared to control cells is shown in red, the number of down-regulated samples is shown in blue. The total number of differentially expressed genes, regardless of measurement time-point is indicated in gray. **(B)** For each measured gene, the number of significant changes in the knock-down experiment compared to control cells is shown. Colour code corresponds to panel A, with gray indicating the total number of cell-lines in which a target was differentially expressed at any of the three measured time-points.

in Smad4 deficient NMuMG cells [Gervasi et al., 2012]. Here no change in Cdh1 expression upon Smad4 knock-down was observed, while differential expression was reported for Id2, Hmga2, Snail1, and Fn1. All of these effects could be qualitatively confirmed in the Smad4 knock-down measured at 18h performed in this study. Dave et al. (2011) further report a block of TGFβ-dependent increase in Zeb1 mRNA as well as protein in Snail1 deficient cells at 8h and 24h after TGFβ-treatment, but no expression changes early after EMT induction [Dave et al., 2011]. Indeed, Zeb1 expression in our dataset was decreased in the Snail1 knock-down cell line relative to the non-targeting control at 18h but not at 0h.

In summary, analysis of the experimental RT-qPCR data indicated good reproducibility and agreement with experimental evidence reported in the scientific literature.

4.2.3 Analysis of gene expression changes provides general insight into the topology of the gene regulatory network controlling EMT

Before applying data-driven network inference methods to gene expression data measured in NMuMG cells, an initial analysis of the dataset may help to general conclusions about the overall structure of the network controlling EMT. For example, the number

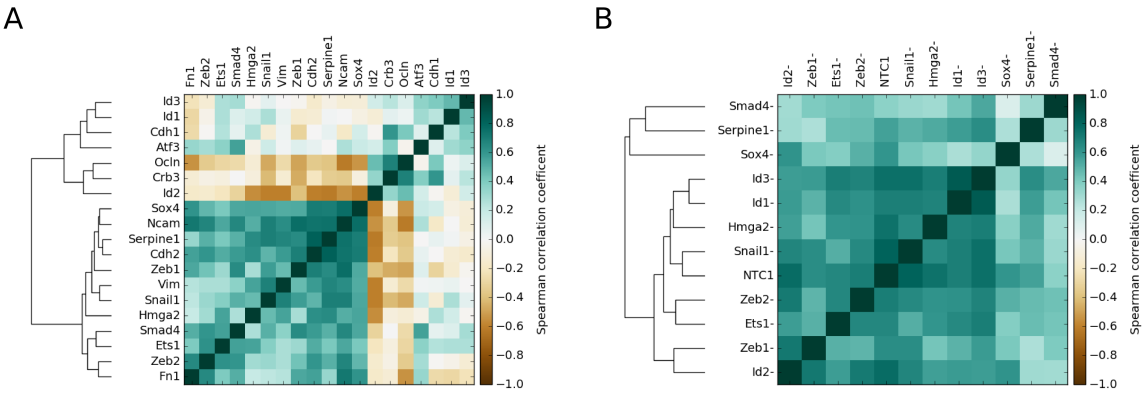


Figure 4.7: Clustering of EMT genes and knock-down cell lines based on similarity of gene expression changes. (A) Heatmap shows Pearson correlation coefficients between gene expression measurements of each gene across all measured samples of the knock-down experiment. The dendogramm indicates the trajectory of hierarchical clustering. **(B)** Heatmap shows Pearson correlation coefficients between gene expression measurements of each cell-line across all measured target genes.

of differentially expressed genes in a given knock-down experiment will indicate how far upstream a factor is located within the EMT network. As expected, Smad4 knock-down cells showed a large number of 9 significantly differentially regulated targets at any of the measured time-points, which is in accordance with its central role in TGFβ-signalling (Figure 4.6A). In more detail, in most cases Smad4 knock-down led to a decrease in target gene expression relative to control cells, implying that this factor is largely involved in activation of target gene expression.

An equal number of 9 genes were differentially expressed upon knock-down of either Serpine1 or Sox4, from which we conclude that these genes serve as additional hubs in the EMT network. The majority of differential expressed genes showed down-regulation upon Sox4, indicating that Sox4 serves mostly as an activator of gene expression. Serpine1 knock-down on the other hand resulted mostly in increased target gene expression, pointing to a mostly inhibitory role of this EMT factor. Although discussed as a central regulator of EMT, Hmga2 knock-down led to the lowest number of differentially expressed targets with only 2 targets being significantly increased relative to control cells.

As we consider Smad4 to be upstream of the gene regulatory network controlling EMT, we expect no significant changes of Smad4 under the measured knock-down conditions. Smad4 expression however did show significant change in Ets1 knock-down cells at 0h as well as significant down-regulation in cells treated with Snail1 or Sox4 siRNA. This impact of EMT genes on Smad4 expression may hint at a potential feedback of gene expression changes back on TGFβ/Smad-signalling (Figure 4.2 and Figure 4.6B). However, observed significant fold-changes of Smad4 in the mentioned cell-lines were minor compared to the overall profile of fold-changes, possibly indicating false positives.

Zeb2 showed the highest count of significant changes in the perturbation experiment,

with overall 7 significant changes in any of the knock-down conditions. Fn1, Id2, and Vim closely followed Zeb2 as highly affected targets with 6 significant changes. The target with the lowest number of significant changes was Ocln, which did not significantly change in any of the measured perturbation conditions compared to NTC cells. Ocln expression is therefore either independent of the perturbed TFs measured in the knock-down experiment, or the effect of TF knock-down is delayed until after 18h of TGFβ-treatment.

We further investigated the existence of groups of genes or cell-lines exhibiting similar gene expression dynamics. For this we chose to correlate gene expression measurements observed in the knock-down data. Clustering genes based on the correlation of gene expression produced two distinct groups of genes (Figure 4.7A). The first cluster included epithelial marker genes Crb3, Cdh1, and Ocln but also also EMT genes Atf3, Id1, Id2, and Id3, which exhibited transient behaviour upon TGFβ-treatment. The second cluster of genes consisted of mesenchymal markers and genes up-regulated during EMT (Sox4, Ncam, Serpine1, Cdh2, Zeb1, Vim, Snail1, Hmga2, Smad4, Ets1, Zeb2, Fn1). Interestingly, although belonging to the same family of TFs, Zeb2 gene expression does not correlate well with expression of Zeb1.

In a corresponding approach we compared the effects of TF knock-down on gene expression with one another. In general, measurements grouped by the similarity of cell lines showed higher agreement compared to measurements grouped by genes. When correlating gene expression in individual cell lines, high impact knock-downs Smad4, Sox4 and Serpine1 cluster together, while other cell-lines preferentially group with control cells.

In summary, the presented gene expression dataset generated from TGFβ-stimulated NMuMG cells under wild-type and knock-down conditions is highly complex and its interpretation far from trivial. For example, whether changes of a gene upon knock-down are due to direct regulation of the target by the TF, or regulation proceeds via additional TFs remains unclear. We therefore turn to data-driven network inference methods in an attempt to reconstruct topological features of the gene regulatory network producing the observed gene expression changes.

4.2.4 Network inference strategy shows good performance on benchmark data

In Chapter 3 various data-driven methods to reconstruct the structure of gene regulatory networks based on gene expression data have been discussed. The basic concepts with which the different applied network inference methods attempt to identify interactions of a network range from methods based on information theory (`PREMER`, `CLR`, `ARACNE`, `MRNET`, `MRNETb`), to machine learning (`GENIE3`), to fitting dynamical models of gene expression (`Inferelator`, `NodeInspector`). After having established the quality of gene expression measurements obtained from the EMT model system and performing qualitative analysis of the data, we intend to apply network inference methods described in Chapter 3 to the measured time-course and knock-down gene expression data. Before doing so however,

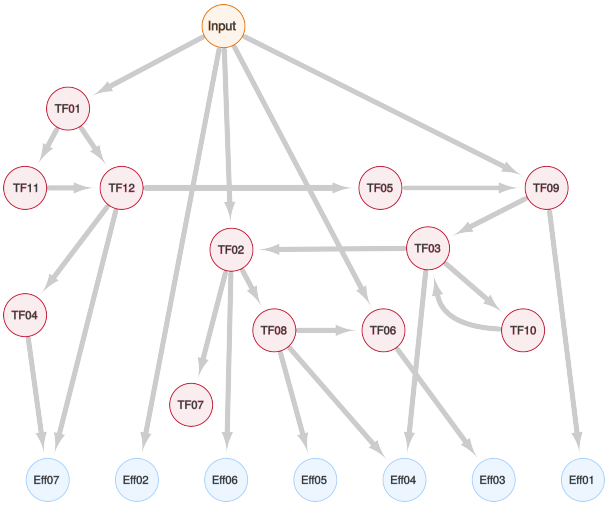


Figure 4.8: Structure of the benchmark network. The benchmark network for evaluating network inference approaches contains one upstream input, 12 TFs and 7 effector genes. A total of 26 interactions were assigned between the genes.

we evaluated the performance of different network inference methods using artificial gene expression data generated from a gene regulatory network of known structure.

Similar to the experimental design of gene expression measurements in NMuMG cells, the generated benchmark data was chosen to include time-course and knock-down expression data with measurements closely corresponding to the collected EMT data. The benchmark network, from which gene expression data was simulated, consisted of one input node, 12 TF nodes and 7 effector nodes (Figure 4.8). In total 26 interactions between genes were assigned before simulating wild-type time-course data as well as knock-down data for 11 different knock down conditions (Figure 4.9). As previously mentioned, JunB in the EMT data was not measured under knock-down, while Sox4 was not measured under wild-type conditions. In order to realistically capture this scenario, expression data for each one TF of the benchmark network was removed for either the time-course (TF10) or the knock-down (TF07) experiment.

Providing gene expression data as input, the different network inference methods rank interactions of the network by their likelihood. In other words, they generate a *prediction* of the structure of the inferred network. Since most network inference methods, with the exception of `NodeInspector`, are not able to treat missing gene expression values, it was necessary to split the full dataset into time-course and knock-down experiments and carry out network inference on both datasets independently. As a result, the prediction of each network inference method was based on either the time-course or knock-down data separately.

Further, as some network inference methods are not designed to infer interactions based on time-course data (see Chapter 3 - `ARACNE`, `CLR`, `MRNET`, `MRNETb`), we chose to evaluate these methods on knock-down data only. In a total, 11 different structural predictions of the benchmark network were generated, based on combinations of the different network inference methods and knock-down or time-course data.

In principle a network of 20 genes produces a total number of 400 potential interac-

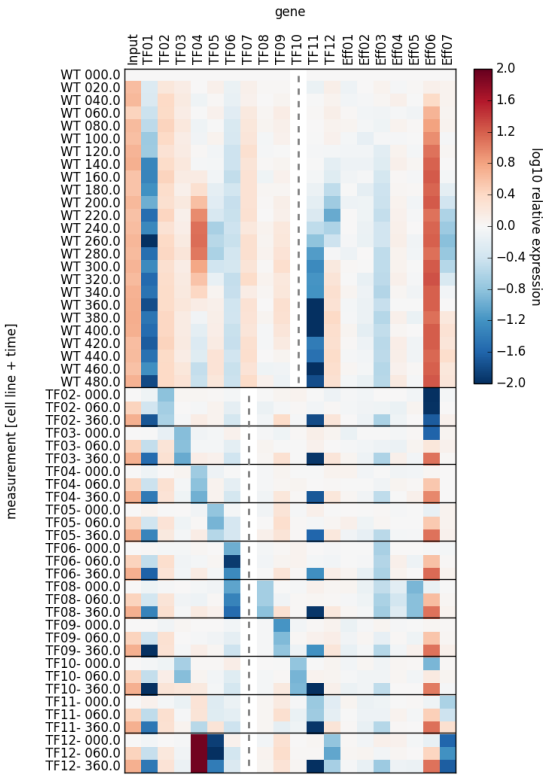


Figure 4.9: Data simulated for the benchmark network. Heatmap showing fold-changes of gene expression values in simulated time-course (top) or knock-down (bottom) experiments. Dashed lines indicate measurements removed for TF10 in the time-course experiment or TF07 in the knock-down experiments.

tions. However, not all interactions may be ranked by network inference. For example, due to conceptual limitations, network inference methods are limited in their ability to infer auto-regulation of genes. In addition to removing auto-regulatory interactions, in the EMT system we do not assume feedback from downstream genes back to TGFβ/Smad-signalling within the duration of the experiment. Accordingly, interactions which involve the regulation of Smad by any EMT gene, were excluded from structural predictions. Additional interactions removed from the network included interactions in which effector genes served as potential regulators. Lastly, it was not possible for network inference methods to determine the existence of interactions between TF07 and TF10, due to the fact, that these two genes have not simultaneously been measured at any of the considered data-points.

The final set of ranked interactions therefore was reduced to only 233 interactions out of the original 400 potential interactions. In Figure 4.10 predicted ranks of the 26 true interactions present in the benchmark network are displayed. As shown in this figure, true interactions group within the top ranked interactions for all network inference methods, already qualitatively indicating good performance of the applied network inference methods.

In order to more quantitatively evaluate and compare the performance of the in-

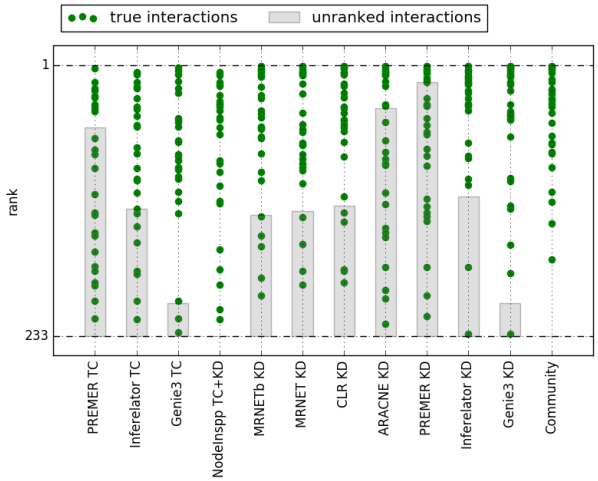


Figure 4.10: Position of true interactions in ranked structural predictions obtained by different network inference methods. The position of true interactions in each prediction is indicated in green. Gray shaded area corresponds to unranked interactions in each prediction.

dividual network inference predictions, we calculated Area under Receiver Operating Characteristic (AUROC) and Area under Precision Recall (AUPR) values (see Chapter 3.2.2). Note once again, that an AUROC value of 0.5 indicates performance comparable to a random prediction, while a perfect prediction will result in an AUROC value of 1. In the case of AUPR, the baseline for performance of a random prediction is given by the fraction of true interactions to all ranked interactions in the benchmark network ($26/233 = 0.11$).

As a result of the analysis of AUROC values, all tested methods performed better than a theoretical random classifier (See Table 9.5 and Figure 4.11). Judging from AUPR values however, methods applied to time-course data only showed performance marginally better than random. Making use of knock-down compared to time-course data therefore improved performance of network inference methods in terms of both AUPR and AUROC values. Indeed, the best performance across all predictions was given by the `Inferelator` method applied to the knock-down data. The modularized network inference work-flow `NodeInspector` (see Chapter 3), for which the complete knock-down and time-course dataset served as input, showed intermediate performance with an AUPR value of 0.29 and AUROC value of 0.69. In this case study, the lowest performance was observed for `PREMER` applied to time-course data.

As suggested by [Marbach et al., 2010b], individual predictions of the network structure obtained by different network inference methods show some degree of dissimilarity, reflecting the fact that each network inference algorithm has its particular set of strengths and weaknesses. We therefore integrated individual structural predictions in order to make use of the complementary information provided by each approach, resulting in a community prediction of the network. As a strategy to integrate predictions, we applied Borda's method of rank integration, which integrates results by calculating the mean rank of each interaction among the individual predictions [Lin, 2010]. In this way, also

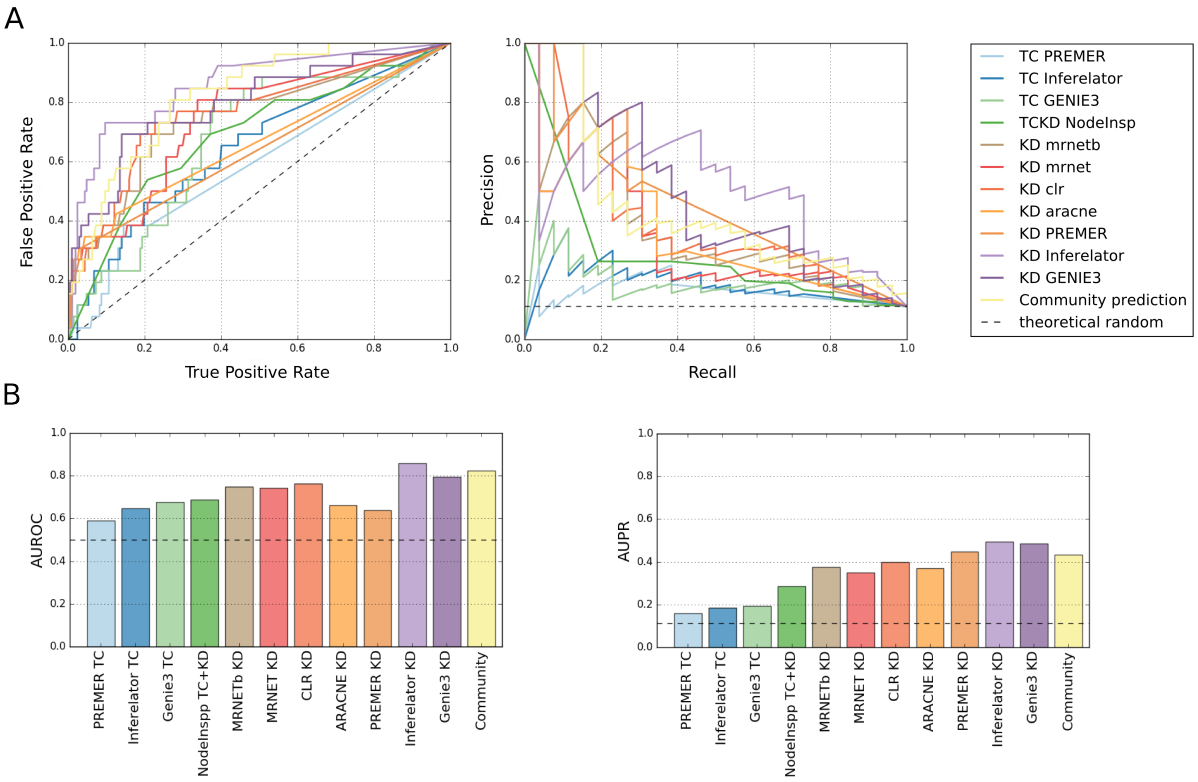


Figure 4.11: Performance evaluation of network inference methods. (A) AUROC (left) and AUPR (right) curves for the different network inference methods applied to the benchmark network. Theoretical random classifier in both plots are indicated by the dashed line. **(B)** AUROC (left) and AUPR (right) values for the different network inference methods applied to the benchmark network.

predictions resulting from time-course or knock-down experiments could be joined.

Choosing only the top 25 best ranking interactions from the community prediction, 10 interactions (Figure 4.12) were correctly predicted. This corresponds to a precision of 0.4 (number of true interactions in the set of predicted interactions) and recall of 0.38 (number of recovered interactions compared to all existing interactions). In contrast, a random pick of 25 interactions would have on average correctly predicted 3 interactions, which translates to a precision of 0.12 and recall of 0.115. Also in terms of AUPR and AUROC values, the community prediction showed performance comparable to the top performing methods: Integrating all network predictions into a community prediction resulted in an AUPR value of 0.43 and an AUROC value of 0.82.

In conclusion, integrating predictions from network inference methods applied to the generated benchmark data showed good performance. In the next section, we apply data-driven network inference methods to EMT time-course and knock-down gene expression data measured in TGFβ-treated NMuMG cells, with the aim to further investigate the structure of the gene regulatory network controlling EMT.

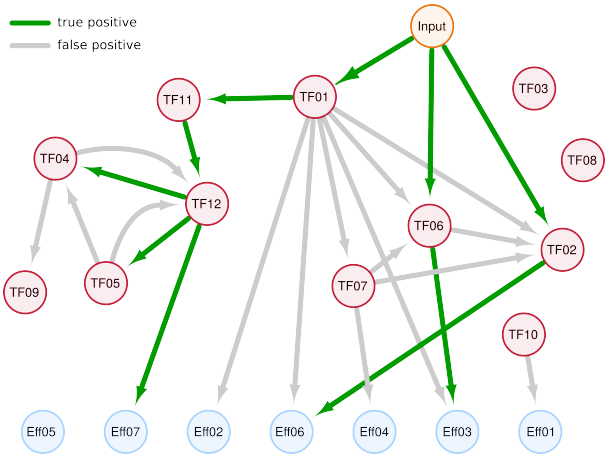


Figure 4.12: Top 25 ranked interactions of the benchmark network in the community prediction. Graphical representation of the top 25 ranked interactions of the benchmark network as resulting from integrating multiple network inference approaches. True interactions are shown in green.

4.2.5 Network inference predicts structure of EMT network

Analogous to the network inference work-flow applied to the benchmark data, in total 11 predictions of the structure of the EMT network were generated based on combinations of network inference methods and NMuMG time-course or knock-down gene expression measurements. By comparing individual predictions with information from the literature we attempted to evaluate the performance of the different network inference methods Figure 4.13. Since each of the network inference methods ranked at least 25 interactions, we were able to compare these top 25 interactions with interactions for which evidence for a direct or indirect regulation exists (grade A and grade B evidence - see Section 4.2.1). In addition to its good performance on the benchmark network, `Inferelator` showed the highest number of known interactions within the top 25 predicted interactions (Figure 4.13). In total, 4 direct interactions (grade A) and an additional 3 indirect - potentially direct - interactions (grade B) were found. By picking 25 interactions at random, on average a number of 0.85 grade A and 2.46 grade B interactions would be expected. The chance of detecting a total number of 7 interactions supported by experimental evidence (grade A and B) within 25 randomly picked interactions is lower than 1% (hypergeometric-test).

Using Borda's method of rank aggregation, individual predictions were integrated, resulting in a community prediction (Figure 4.14A). A summary of this community prediction and the network structure corresponding to the top ranked 25 interactions are shown in Figure 4.14B. In the community prediction, the interaction of Sox4 on Ncam was ranked highest. Indeed, a regulatory role for Sox4 on Ncam (also known as CD56) has been proposed for human myeloma cells [Iqbal et al., 2010]. In total, within the top 25 interactions of the community prediction we found 5 interactions for which experimental evidence was reported in the literature. For example, regulation of Atf3 by Smad was ranked 7th in the community prediction. Ranked 10th in the community prediction was

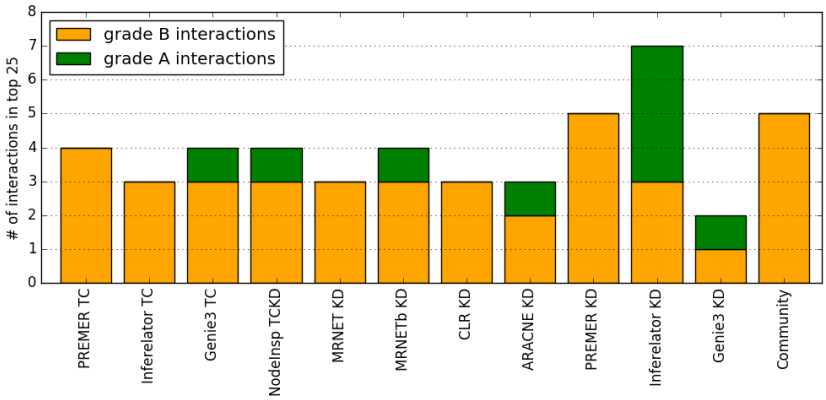


Figure 4.13: Comparison of predictions with known interactions of the EMT network. For each prediction the number of interactions with either grade A (green) or grade B (orange) evidence within the top 25 ranked interactions is shown.

regulation of JunB by Smad. The highest ranked interaction for which there is evidence of *direct* regulation was regulation of Snail1 by Smad, which was ranked 48th out of 233, closely followed by the regulation of Id2 by JunB ranked 54th. Although evidence for a direct regulation of Cdh1 by Zeb2 exists, this interaction was merely ranked 197th.

In general, known direct and indirect interactions from the literature are spread out over the full range of possible ranks (Figure 4.14C), with a mean rank of 106.6 for interactions for evidence of a direct regulation (grade A) and a mean rank of 98.0 for interactions with evidence for indirect regulation (grade B). By calculating the mean of rank of interactions involving a TF we further tried to estimate its relevance for the EMT network. Here the trio Serpine1, Smad and Sox4 again emerged as central regulators of the EMT network with mean ranks of 72.2, 81.9, and 83.7 respectively. TFs with lowest mean ranks were Id1 (161.2), Id2 (173.4), and Atf3 (179.9).

4.2.6 Evaluation of network predictions by motif analysis and publicly available ChIP-data

A complementary approach to data-driven network inference is to infer regulatory interactions from the existence of TF binding motifs in the promoter region of target genes. For 8 out of the 13 TFs in the EMT network (Atf3, Ets1, Id2, JunB, Sox4, Zeb1, Smad2/3/4, Snail1), known motifs could be found in existing databases [Mathelier et al., 2016, Heinz et al., 2010] (Table 8.5). The position frequency matrices of each motif, indicating the probability of each nucleotide in the motif sequence, were compared to the putative promoter sequence around the transcription start site and hereby potential TF binding sites identified. Based on this analysis, in total 53 interactions between the 8 TFs and the 19 targets genes were predicted (Figure 4.15).

Out of the 8 known interactions qualifying as direct (grade A), 3 interactions also appeared in the motif analysis (Atf3 → Id2, Zeb1 → Cdh1, Smad → Hmga2), 3 could not be compared due to missing position frequency matrices (Hmga2 → Snail1, Hmga2 → Cdh1, Zeb2 → Cdh1), and 2 of the known direct interactions could not be confirmed

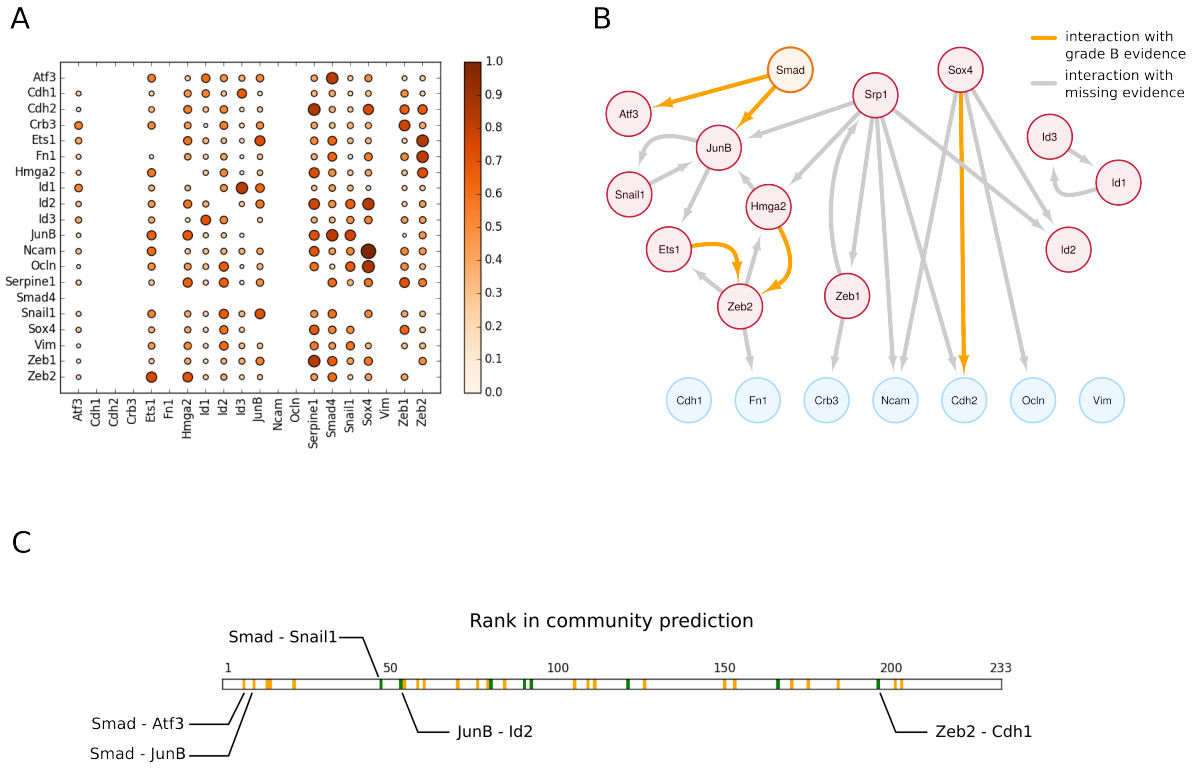


Figure 4.14: Community prediction of the EMT network. (A) All 233 potential TF-target interactions of the EMT network are ranked according to the community prediction. Size and color of the circle indicate the mean rank across all individual predictions. **(B)** Schematic network of the top ranked 25 interactions of the community prediction is shown. Orange arrow indicates an interaction with grade B evidence, while a gray arrow indicates an interaction with no supporting evidence from the literature. **(C)** Schematic representation of the ranked community prediction. Green represents interactions with grade A evidence, while orange indicates interactions with grade B evidence

by motif analysis (JunB → Id2, Smad → Snail1). From the combined set of *direct* and *indirect* interactions (grade A and B) reported in the literature, 9 interactions could be confirmed by motif analysis, 10 interactions could not be evaluated due to missing motif information, and 12 interactions were able to be evaluated using motif analysis but not confirmed by literature evidence. This number of matching interactions between interactions reported in the literature and interactions identified by motif analysis is only slightly above the expected average overlap 7.32 interactions.

We further checked whether TF-target interactions predicted based on motif analysis were to be found preliminary in the set of high ranked interactions predicted by network inference. Within the top 25 ranked interactions from network inference also evaluated by motif analysis, we found 10 interactions confirmed by motif analysis. The average expected overlap between the two sets of interactions is 8.57. This limited enrichment of interactions determined by motif analysis within the top ranked interactions obtained by network inference was observed to be independent of the cut-off applied for top ranking interactions (Figure 4.15B).

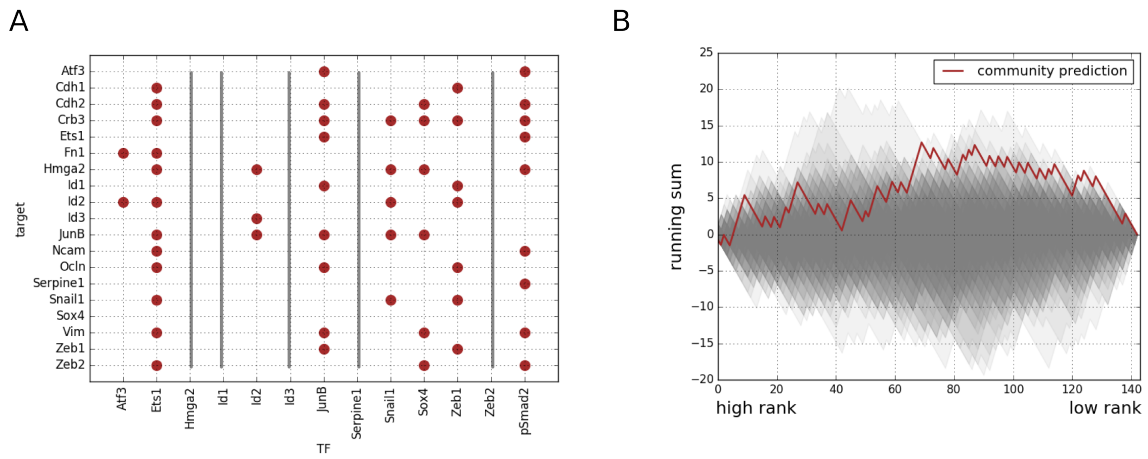


Figure 4.15: Structural prediction of the EMT network based on motif analysis. (A) Regulatory interaction matrix derived by motif analysis. Each circle indicates the occurrence of a particular TF motif in the promoter sequence of a target. For TFs with gray bars no positional frequency matrix could be obtained in the considered databases. **(B)** Running sum plot of interactions found by motif analysis along the ranks of the community prediction is displayed (red line). Along the ranks of the community prediction each occurrence of an interaction supported by motif analysis results in an up-step, while missing support produces a down-step in the running sum. Gray shaded area corresponds to the running sum of 100 randomly re-sampled community predictions. For the running sum analysis only interactions evaluated by both network inference and motif analysis are considered.

ChIP experiments produce a more reliable assessment of TF binding compared to motif analysis, since here the physical association of a TF to the DNA is measured. In order to determine existing regulator-target interactions based on ChIP data, we compiled a set of 44 different publicly available ChIP experiments from the cistrome database, spanning different cell types and experimental conditions (cistrome.org [Wang et al., 2014, Mei et al., 2017]). The 44 selected ChIP experiments contained binding information for 6 out of the 13 TFs considered in the EMT network (Atf3, Ets1, Hmga2, JunB, Smad2/3/4, Sox4). For each of the experiments in the ChIP data we extracted the top 600 ranked TF-target associations as high-confidence interactions. Target genes of the same TF but from different datasets were combined and a list of putative TF-target interactions was compiled. This list of high-confidence TF-target interactions derived from ChIP data contained a total of 20 interactions between selected EMT genes.

Within the list of interactions determined from the ChIP data, only for one interaction (JunB $\rightarrow$ Id2) evidence for a direct binding could be found in the literature. Considering both interactions supported by grade A as well as grade B evidence, in total 5 out of the 20 interactions identified by ChIP analysis were supported by experimental evidence. The corresponding average expected overlap between both sets is 4.56 interactions.

We further observed limited agreement between the the top 25 ranked interactions from network inference and the interactions derived from the ChIP data, with 5 interactions appearing in both sets. This lack of agreement between predictions from network inference and analysis of ChIP data again was independent of the cut-off chosen for the top interactions (Figure 4.16B).

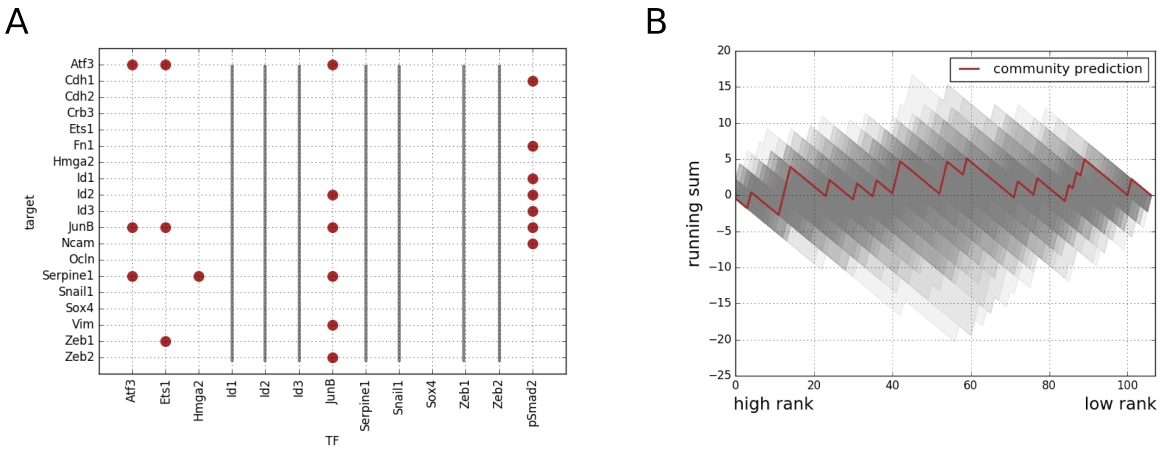


Figure 4.16: EMT network based on analysis of ChIP-data. (A) Interactions matrix indicating interactions derived by analysis of ChIP-data. Each circle represents the existence of an association of a TF to the respective target gene. TFs for which no ChIP data was available are overlayed by gray bars. **(B)** Running sum plot showing the position of each TF-target interaction derived from analysis of ChIP data in the ranked community prediction. For the running sum only interactions evaluated by both network inference and motif analysis are considered.

4.3 Discussion

In Chapter 3 a novel network inference method based on non-linear models of transcriptional regulation was introduced and compared to existing state-of-the art network inference methods. In contrast to testing individual regulator-target interactions as in Chapter 3, in this chapter we applied a combination of different network inference methods to a complex dataset of gene expression measured under time-course and knock-down conditions in order to unravel the structure of the gene regulatory network responsible for the progression of EMT.

As EMT is a highly complex but also important process in both health and disease, in the past attempts have been made to decipher the gene regulatory network controlling EMT. In A549 cells, for example, time-course gene expression data obtained by RNA-Seq was combined with genome wide binding-data of 52 TFs in order to build a dynamic gene regulatory network controlling EMT [Chang et al., 2016]. This model identified JunB, Ets2, and HNF4a as key regulators involved in early EMT. Using immunoprecipitation analysis, indeed physical association of the three mentioned TFs could be shown.

However, data-based network inference remains a challenging problem in systems biology, facing many conceptual as well as practical limitations. Often network inference methods need to deal with sparse and noisy data, and datasets containing missing values. Despite these challenging conditions, good performance of our network inference work-flow, in which multiple structural predictions were integrated, could be demonstrated using an artificial network from which realistic experimental data was simulated (Chapter 4). Particularly, the availability of knock-down gene expression data, resulted in increased performance of most network inference methods compared to a setting, in which only time-course gene expression under wild-type conditions was provided (Chap-

ter 3). However, good performance of network inference strategies under benchmark conditions does not necessarily directly transfer to similar performance when applied to real biological data. Particularly, the EMT system poses some additional challenges not enforced by the benchmark network.

For instance, in the benchmark network we assume missing feedback of the TF network on TGFβ/Smad-signalling. In the EMT network however, there is evidence of Sox4 over-expression affecting expression of TGFβ-receptors as well as Smad2 phosphorylation in MCF10A cells [Zhang et al., 2012]. Interestingly, also in our data knock-down of Sox4 leads to changes in Smad4 expression, possibly further affecting the stoichiometry of phosphorylated Smad-trimers. In summary, pour data as well as multiple other studies argue for a tight co-regulation of Smad signalling and Sox4 gene expression. As a major caveat however, dynamics of the input into the EMT network have only been measured on the level of pSmad2 under wild-type conditions, while the activity of pSmad2 has not been determined under knock-down conditions, posing a critical limitation for the network inference approach.

In the benchmark network all components contributing to the gene regulatory network are known. An additional difficulty imposed by the real biological data however, is that in NMuMG cells only a subset of EMT relevant factors have been measured. Missing gene expression data of critical EMT factors however, may lead to the inability of network inference methods to identify important interactions relevant for the genetic program of EMT. One example further illustrating the failure to evaluate potential interactions, is missing measurement of Sox4 in the time-course and JunB in the knock-down data. Indeed, motif analysis shows occurrence of a Sox4 motif in the promoter sequence of JunB. There is no doubt that in the generated gene expression dataset, essential EMT factors are missing entirely. For example it is known that miRNAs of the miR-200 family and Zeb, as well as miR-34 and Snail1, form two distinct double-negative feedback loops important for the switch like behaviour of EMT [Korpal et al., 2008,Siemens et al., 2011].

Adding even further to these complications, there is evidence that gene expression dynamics associated with EMT further depend on post-transcriptional gene regulation. Accordingly, while transient up-regulation of Atf3 was observed in RT-qPCR data, a more sustained response of Atf3 could be shown on protein level [Bakin et al., 2005]. Differences in mRNA and protein expression however, may further impede the ability of network inference to correctly predict existing regulatory interactions. Therefore, in the future more comprehensive measurement of gene expression, including measurement of protein abundances, are required in order to more reliably reconstruct the gene regulatory network of EMT.

Despite the conceptual as well as practical challenges faced by data-based network inference methods, our detailed analysis of the generated knock-down data as well as structural predictions supported the important role of Smads in TGFβ-induced EMT. From this analysis in addition to Smad, Sox4 and Serpine1 emerge as critical regulators of EMT, while traditional factors such as Snail and Zeb seem less relevant to regulation of early EMT.

In order to evaluate interactions predicted by different network inference methods,

additional information on the structure of the gene regulatory network controlling EMT was derived from analysis of TF binding motifs and analysis of publicly available ChIP data. However, as a result we observed only minor overlap between interactions identified by ChIP-analysis, motif-analysis and known interactions reported in the literature.

Potential causes for this missing agreement between the different sources of information are manifold. The identification of motif occurrences in motif analysis, for example, relies on comparing position frequency matrices of each motif with putative promoter/enhancer regions of genes. Position frequency distributions are themselves derived from experimental data in combination with down-stream computational analysis, potentially introducing uncertainty on the level of predicted motifs. On the other hand, the exact position of enhancer or promoter sequences is not well defined, with TFs being able to bind to even distant genetic regions, while impacting target gene expression. Accordingly, Kennedy et al. (2011) found 75% of Smad4 binding loci within 100kb around a known target gene, while only 13% of Smad4 binding signals were found within 8kb of the promoter region [Kennedy et al., 2011]. In case of regulator-target interactions of the EMT network identified by the analysis of ChIP data, the largest caveat was certainly the fact that currently none of the publicly available ChIP-Seq data was specific for NMuMG cells or was carried out in an EMT/TGFβ-context.

In comparison to analysis of TF binding motifs and ChIP data, evidence from detailed molecular experiments performed in TGFβ-treated NMuMG cells represents the most reliable source of information to identify potential TF-target interactions of the EMT model system. Indeed, we observed a significant overlap of interactions predicted by the network inference approach `Inferelator`, and interactions supported by the literature. Also while integrating multiple network inference approaches into the community prediction top ranked interactions were supported by experimental evidence. The magnitude of this overlap however was decreased compared to the best performing method. In our case, limited performance of the community prediction may result from the similarity of structural predictions originating from network inference methods applying a similar strategy to identify regulatory interactions. These similar prediction could potentially bias the final community prediction (See supplemental information 9.3.7). As it is impossible to judge the performance of network inference methods on real gene expression data, we nevertheless decided to include all considered network inference approaches in the community prediction, regardless of their performance on the benchmark data. As a solution, a larger variety of network inference methods should be applied to NMuMG expression data and integrated in order to avoid that specific predictions disproportionately impact the community prediction.

Finally, lack of support of the top ranked interactions predicted by the community approach does not necessarily imply the absence of these interactions, as critical experimental evidence might still be missing. In the future therefore, highly ranked interactions predicted for the EMT network should be tested using specifically designed ChIP-Seq experiments carried out in TGFβ-stimulated NMuMG cells to establish a more reliable assessment of the structural predictions made by the network inference approach.

Chapter 5

Formulating a Dynamical Model of the Gene Regulatory Network Controlling the Epithelial-to-Mesenchymal transition in NMuMG cells

Preamble

This project was carried out in collaboration with SP, SS, AG, and SSu who performed all experiments described in this chapter. Valuable feedback was further provided by VT and SL.

5.1 Introduction

In the last chapter we predicted interactions present in the gene regulatory network controlling EMT using a combination of network inference algorithms applied to an extensive time-course and perturbation dataset. As a result, potential regulatory interactions in the EMT network were ranked according to their probability. This structural prediction was evaluated and compared to interactions derived from motif-, ChIP-, and literature analysis.

Although arguably useful, these structural predictions do not capture the dynamical gene expression changes related to EMT. Therefore, in order to gain a deeper understanding of the gene regulatory network controlling EMT, we need to move away from static networks and towards a dynamical model of gene expression. In this chapter we present the framework for the formulation of such a dynamical model of the gene regulatory network controlling EMT. While fitting the model to gene expression data obtained from TGFβ-treated NMuMG cells, we further include the structural prediction originating from data-driven network inference methods applied in Chapter 4.

5.1.1 Simultaneous network inference and parameter estimation in models of gene regulatory networks

A large selection of possible mathematical formulations to model dynamical systems exist: In the simplest case the variables describing the expression states of genes in the system can take on either the value 0 - representing the non-expressed state - or the value 1 - according to the situation where the gene is expressed. The state of the complete network is then described by a vector of binary values, with each position of the vector describing the expression state of a gene. Upon iteration, the network proceeds to a different state based on a set of logical rules summed up in the so called regulation function. In boolean models, regulation functions are typically encoded as logical gates, such as the AND or OR functions, which describe the expression state of a gene at time t+1 in dependence of the expression states of a subset of genes in the system at time t. The advantage of a boolean formulation of a gene expression network is its scalability: Many variables can be included in the system while computation times remain comparatively small. On the other hand however, dynamics are merely described on a qualitatively.

A more quantitative description of gene regulatory networks can be achieved by formulating a dynamical model using on ordinary differential equations (ODEs). In contrast to boolean model of gene regulation, in ODE models expression levels of genes are no longer limited to binary values. Also integration of the system equations does not proceed on discrete time-intervals. Instead, the change in expression of a gene is described as a differential equation composed of a production and degradation term and system equations are solve by numerical integration methods. While in most ODE models of gene regulation degradation of gene products is assumed to be linear and independent of the expression of other genes, the production term typically takes on more complicated forms and involves the expression levels of other genes in the network.

In order to formulate a dynamical model of gene expression the upstream regulators for each gene must be identified, i.e. the regulatory structure of the network must be determined. In the best case scenario a large body of experimental evidence exists, from which conclusions on the networks structure can be drawn. Alternatively, as discussed in the previous two chapters, the network topology can be inferred from gene expression data using network inference approaches.

Once the structure of a gene regulatory network is determined, ODE models further include a number of relevant parameters which need to be defined before its dynamics can be simulated. These parameters include gene specific production and degradation rates, but also parameters describing the strength of regulation between different genes. Assigning parameter values experimentally using molecular biology techniques remains difficult, particular *in vivo*. Therefore the standard procedure to estimate model parameters is to fit the model to existing gene expression data and select the set of parameters with which the highest agreement between model and data can be achieved. In some cases however, neither the model topology or its parameters are sufficiently constrained. By incorporating structural features of the model into the model parameters however, it is possible to infer both features of the model simultaneously while fitting the model to the data.

Jaeger et al. (2004), for example, designed a model of the gap genes involved in anterior-posterior patterning of the early *Drosophila* embryo [Jaeger et al., 2004]. Although cross-regulation between gap genes has been extensively studied [Nüsslein-volhard and Wieschaus, 1980, Nusslein-Volhard et al., 1987], the goal was to reverse-engineer the gap gene network based on gene expression data only. For this, the approach made use of spatially and temporally resolved protein expression data obtained by fluorescent antibody staining. The data was combined with a generic model of gene regulatory networks based on coupled ordinary differential equations. Without including any prior information on the cross-regulation of the gap genes, it was possible to correctly determine the structure of the gap gene network by fitting the generic model to the data. The approach further revealed that negative regulation of overlapping gap genes leads to an anterior shift in gap domains [Jaeger et al., 2004] and allowed for a detailed analysis of the dynamics exhibited by the gap gene network [Manu et al., 2009].

In another study, simultaneous inferring both structure and parameters of a gene regulatory network, a core dynamic model of the KdpD/KdpE system of *E. coli* was formulated and subsequently different competing model extensions were evaluated [Rodriguez-Fernandez et al., 2013]. In this approach the main topology of the model was fixed while other interactions remained optional. The formulation of the ODE model contained the conventional real valued parameter values encoding for production or degradation rates of the different species, as well as interaction strengths between them. In addition however, model parameters included binary parameter values describing the existence or non-existence of potential interactions in the model. Estimating model parameters composed of a combination of real valued and integer values is particularly challenging, since differences between model and data are no longer described by a smooth, real-valued cost function. This prohibits the use of common gradient based optimization methods. Therefore, in order to estimate integer valued parameters, specific optimization schemes capable of handling mixed integer non-linear programming (MINLP) problems had to be adopted. Using a suitable optimization algorithm together with *in silico* generated data, the true structure of the underlying dynamic model of the KdpD/KdpE system could be inferred and parameters estimated with high precision.

In this chapter we formulate a dynamical model of the gene regulatory network controlling EMT, based on the structural information obtained by network inference in the previous chapter. First, we design a dynamic model of a generic gene regulatory network where no structural prior is imposed, meaning that almost all interactions between the different genes in the network are in principle possible. We then proceed to fit this model to expression data obtained under wild-type and various knock-down conditions. While fitting the model, we penalize interactions according to their likelihood, hereby increasing the match between model structure and interactions inferred by the previously obtained community prediction. The resulting model fits are evaluated with respect to their ability to explain the data, the agreement between network structure and known interactions, as well as the consistency of model predictions with experimental evidence reported in the literature.

5.2 Results

5.2.1 Model formulation

The framework for the formulation of a dynamical model of the EMT network presented here is related to the general mathematical formulation of gene regulatory networks found in Jaeger et al. (2004) [Jaeger et al., 2004]. In this formulation the rate of change of an individual gene in the network is represented by an ordinary differential equation combining both a production and degradation term. While degradation is assumed linear, the production term can scale between zero and its maximum production value depending on the additive input received from other genes in the network (Figure 5.1).

$$\frac{dy_k^a(t,\theta,u(t))}{dt} = R^a g\left(\sum_{b\in G}\beta^{ab}w^{ab}y^b + \beta^{au}w^{au}u(t) + h^a\right) - \lambda^a e_k^a y^a, \qquad (5.1)$$

The regulatory input of a gene y^a is determined by adding up the weighted contribution of each upstream factor y^b and the input received from TGFβ/Smad-signalling $u(t)$. In our model formulation, the information on which specific factors contribute to the expression of a gene is encoded in the binary regulation matrix β. The columns of this matrix represent regulating genes (TFs) while rows depict the different possible target genes. A value of 1 in the corresponding field of the binary regulation matrix indicates the existence of an interaction between a TF and its target while 0 indicates its absence, hereby encoding the general topology of the system. A secondary matrix, the regulatory weight matrix w, contains information on the regulatory strength a TF asserts on its target gene. A positive weight indicates activation, while a negative weight specifies inhibition of the target gene by the TF. By element-wise multiplication of the binary regulation matrix and the regulatory weight matrix and summing over the columns, the regulatory input for each target gene can efficiently be calculated. The total regulatory input received by each gene is then passed through the sigmoid regulation function g, which can scale the production R^a of a gene between 0 and its maximum value. The impact a change in the regulatory input may have on the transcriptional output of a gene is determined by the promoter sensitivity parameter h^a, describing the steepness of the regulation function. As already stated, we assume linear degradation of mRNA with degradation rate λ. While EMT genes represent dependent variables in the model, TGFβ/Smad-signalling is considered as an external input. Accordingly, in order to numerically solve ODEs, pSmad2 expression is linearly interpolated between existing data-points.

The model formulation so far is restricted to the simulation of the gene regulatory network under wild type conditions. To facilitate simulation of knock-down experiments, an additional parameter - the knock-down efficiency e_k^a - is included in the model, by which the degradation rate of the affected mRNA is multiplied when carrying out model simulations. Analogous to siRNA treatment in knock-down experiments, knock-downs in our model formulation are simulated for 48h before TGFβ-treatment is applied. Since

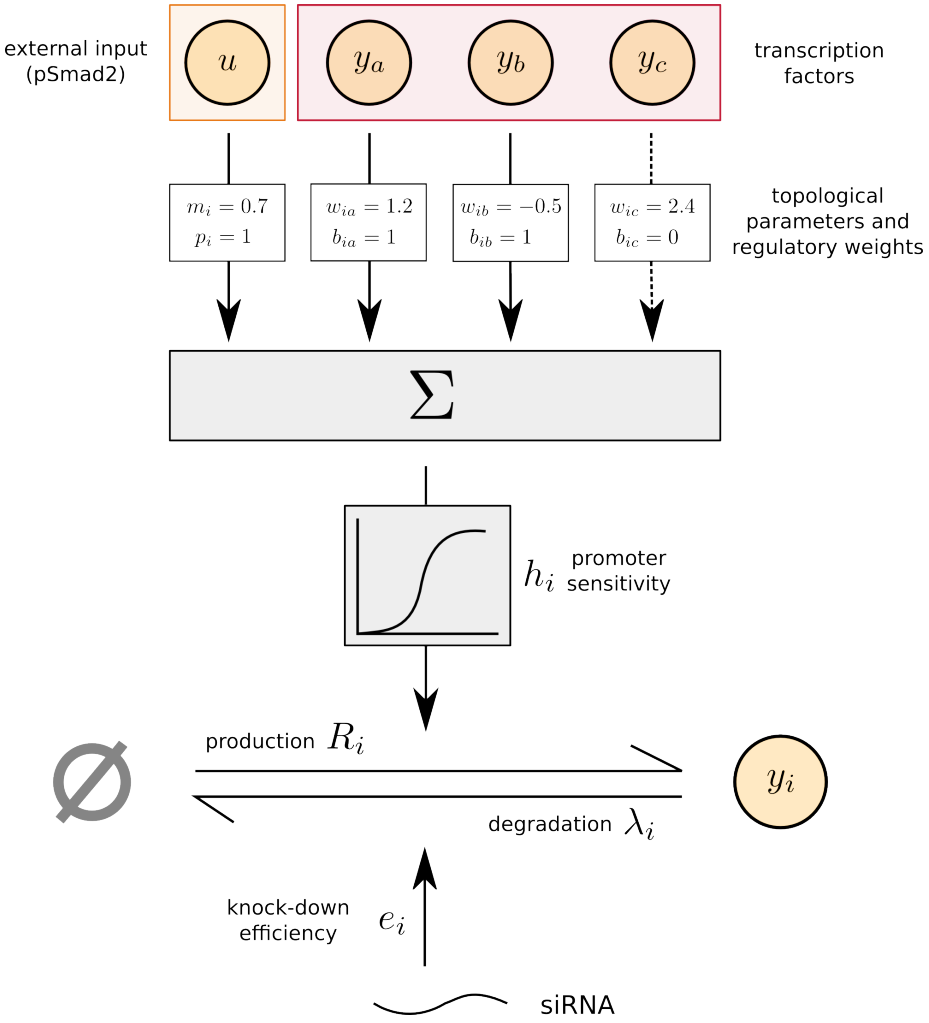


Figure 5.1: Schematic representation of the model formulation. The rate of change of gene y_i in the model is described as a combination of production and degradation term. Each gene of the network y_j as well as the input u may impact the production of gene y_i by contributing to its regulatory input. The information on which specific genes regulate a given factor, are encoded in binary model parameters p_i and b_{ij}. Regulatory weights m_i and w_{ij} define the regulatory strength of gene y_j (or the input u) on the production of gene y_i. Negative weights represent inhibition of gene y_i by gene y_j. The regulatory input into each gene is summed up and passed to the sigmoid regulation function, describing the relation of TF input and transcriptional output of gene y_i. In addition to production, linear degradation of gene y_i is assumed with degradation rate λ_i. When simulating knock-down conditions of gene y_i the degradation rate of gene y_i is multiplied by the knock-down efficiency factor e_i.

pSmad2 expression was not measured under knock-down conditions, it is assumed that neither the expression nor phosphorylation of Smad2 changes under any of the considered knock-down conditions.

Model parameters such as production and degradation rates, but also values of the binary regulation matrix, the regulatory weight matrix, knock-down efficiencies, and promoter sensitivities are not known a priori but need to be estimated from the data. Not only due to the non-linearity of model equations, but also because of the large number of model parameters, this presents itself as a demanding task. In order to reduce the

number of parameters, and therefore the complexity of model fitting, further simplifying assumptions can be made: Assuming gene expression does not change in absence of any form of perturbation, production rates are chosen such that production and degradation terms are balanced in unperturbed conditions. Further, specific interactions are removed from the model, hereby further decreasing the number of parameters. This includes regulation of target genes by genes which do not function as TFs (effector genes: Cdh1, Cdh2, Crb3, etc), auto-regulation of genes as well as regulation between Sox4 and JunB (see Chapter 4.2.4). Including two scaling factors, which account for the observed difference in Hmga2 and Snail1 expression in the knock-down compared to the WT data (see Section 4.2.2), the final model contains a total of 516 model parameters.

5.2.2 Model fitting

Since no information on specific values of model parameters was available, model parameters were estimated by fitting the model to RT-qPCR expression data of relevant EMT genes measured under wild-type and knock-down conditions in NMuMG cells. Instead of using the complete gene-expression dataset described in Chapter 4, we decided to drop gene expression measurements obtained under Smad4 knock-down conditions. This decision was based on the fact that we did not possess any information on pSmad2 activity in Smad4 deficient cells. As a result, the final dataset utilized to estimate model parameters included a total of 972 data-points.

The key principle of estimating model parameters from measured data is to simulate the model with many different parameter values and to select parameter values for which there is the best agreement between the simulated and measured gene expression values (see Section 2.2.1). The agreement between model simulation and data is commonly measured by the weighted squared difference between measured and simulated gene expression, with weights corresponding to the estimated error in each data-point (Equation 5.2). In this way, measurements associated with large errors are less relevant for the model fit, while more weight is allocated to measurements with small uncertainty:

$$\chi^2(\theta) = \sum_{i=1}^{n} \frac{(y_i^{model}(\theta) - y_i^{data})^2}{\sigma^2}. \tag{5.2}$$

A large variety of heuristics to minimize the distance between model and data exist. However, since our model formulation included both continuous parameters (degradation rates, regulatory weights, etc) and non-continuous parameters (topological parameters), strategies which can deal with MINLP problems needed to be employed. Therefore, in order to estimate model parameters we chose the scatter search optimization method implemented by Egea et al. (2009) [Egea et al., 2009, Egea et al., 2014]. Scatter search is a well established global optimization strategy capable of handling MINLP problems, such as the one described here. In scatter search a set of initial and diverse set of trial solutions is generated. Using parameter re-combination techniques, these trial solutions

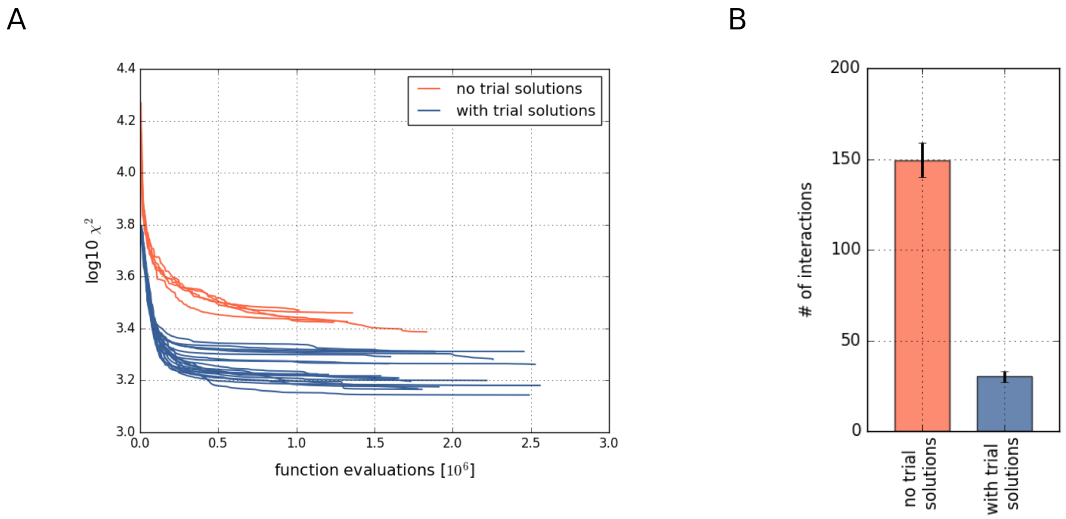


Figure 5.2: Convergence of model fitting. **(A)** Minimum value of the cost function (Equation 8.5) with respect to the number of function evaluations. Red curves show convergence for model fits without supplying initial trial solutions obtained from `NodeInspector`, while blue represents model fits including trial solutions. **(B)** Barplot showing the mean number of interactions selected from model fits with (blue) or without (red) supplying trial solutions.

are improved and eventually gain access to the reference set. The reference set contains a small number solutions which show a good quality of model fit but also a certain degree of diversity. Upon each iteration a subset of solutions is selected from the reference set and by re-combination of these solutions new trial solutions are generated. Following this procedure, the quality of the model fit is iteratively improved, while at the same time efficiently exploring the available parameter space. Scatter search terminates its search for better parameter values either after certain time or number of model simulations is reached, the cost function decreases below a predefined value, or there is no further significant improvement in the model fit.

In an enhanced version of scatter search at each iteration trial solutions are improved using a local optimization algorithm. However, because of the large number of solutions needed to estimate the structure of the parameter space around the current solution, local optimization techniques are not generally suitable for problems with many parameters. Accordingly, only very few local optimization techniques exist capable of solving MINLP problems. In order to be able to include a local search strategy in our model fitting approach, we therefore fixed the topology of the model after 20 hours of optimization, hereby reducing the complexity of the search space. Subsequently local optimization with respect to the remaining real-valued parameter values was carried out via a dynamic hill climbing algorithm [Yuret and de la Maza, 1993].

Good initial parameters values aid the scatter search method in the establishment of a high quality reference set. Therefore, to further increase the performance of the parameter estimation method, we supplied the scatter search algorithm with initial parameter values originating from one-to-one regulator-target interactions estimated by `NodeInspector` (See Chapter 3). This was made possible by the explicit compatibility

of the gene regulation function utilized in `NodeInspector` with the model formulation of the gene regulatory network: For each parameter in the regulation function, there exists a corresponding parameter in the formulation of the dynamical model of the gene regulatory network. In order to impose the least restrictions possible in the trial solutions, for each target in the network a single regulator was selected at random. When running scatter search, 25% (1290) of the initial trial solutions were composed of one-to-one networks obtained from `NodeInspector`. The remaining trial solutions of the reference set were randomly sampled within the defined parameter ranges.

As in the case of data-driven network inference methods, we first applied the proposed model fitting approach to the benchmark data described in Section 4.2.4. Here, we observed that initializing scatter search using results from `NodeInspector` led to an overall increase in model fit (Figure 5.2A). At the same time, the total number of function evaluations within the chosen optimization time was greater compared to fitting without supplying initial parameter values (1.85×10^6 compared to 1.35×10^6 mean function evaluations). As an additional effect of adding high quality trial solutions obtained from `NodeInspector`, we observed a decreased number of interactions in the model structure (Figure 5.2B).

In total we carried out 20 model fits on the benchmark data (Table 5.1). From these 20 fits, the lowest χ^2-value obtained was 1393.01 with a mean χ^2 of 1.43 per data-point. When testing the correlation between model and data, we could further show that gene expression values of the model and data aligned well, with a Pearson correlation coefficient of 0.89 produced by the best fit (Figure 9.14) and a mean Pearson correlation coefficient of 0.87 over all fits.

While there exist only 26 interactions in the benchmark network, the best model fit selected a total of 31 interactions to describe the data, which is close to the average number of interactions across the 20 performed fits (30.05). Out of all 233 possible interactions, six interactions were consistently chosen among all 20 model fits, while 91 interactions never appeared (Figure 5.3). Hence, although some structural features between model fits were shared, most of the chosen interactions differed between the model fits.

In our model fitting approach no constraint on the interactions present in the network was imposed. In order to compare our fitting results with that of fitting a model with the correct network structure, we performed three fits including only the true interactions realized in the benchmark network. The best fit resulted in a χ^2-value of 1060.3 and a Pearson correlation between modelled and measured gene expression of 0.89.

5.2.3 Model fitting strategy shows performance comparable to network inference

After having established the ability of the mathematical model to describe benchmark gene expression, we assessed how well interactions predicted by model fits matched with those of the benchmark network. One important aspect to notice however, is that a single model fit produces a definite set of interactions, unlike network inference, where each interaction is ranked according to its likelihood. Calculating rank-dependent per-

formance measures, such as AUROC and AUPR (see Section 3.2.2), in order to assess the performance of individual model fits therefore was not possible. Instead we chose to calculate standard performance measures of precision and recall (see Section 2.2.2).

As mentioned previously, the best model fit to the benchmark data selected a total of 31 interactions. Out of these 31 interactions predicted by the best fit, 13 interactions were also present in the benchmark network, resulting in a precision value of 0.42 (Table 5.1). Accordingly, 13 out of the existing 26 interactions where correctly recovered, translating to a recall value of 0.5. In comparison, selecting 31 interactions at random would have correctly predicted an average number of 3.46 interactions, corresponding to a precision value of 0.11 and a recall of 0.13. The mean precision value over all model fits was 0.37 while mean recall was 0.43. There was no apparent correlation between quality of model fit and precision and recall values (data not shown).

The results above indicate that each model fit showed substantial better performance in terms of precision and recall compared to a random classifier. But how does this performance relate to the performance of the community prediction in Chapter 4? To answer this question, we chose the same number of top ranking interactions from the community prediction as were selected in each model fit and compared values of precision

Table 5.1: Summary of 20 model fits carried out on benchmark data. Shown are quality of model fit (χ^2-value), number of interactions predicted (n), precision and recall values obtained from model fitting (left), and corresponding precision and recall values obtained from the community prediction (see Section 4.2.4) using the top n ranked interactions.

				Model Fitting		Network Inference	
fit	χ^2	n	ρ	precision	recall	precision	recall
1	1393.01	31	0.89	0.42	0.50	0.39	0.46
2	1445.89	35	0.88	0.31	0.42	0.37	0.50
3	1463.87	29	0.90	0.41	0.46	0.38	0.42
4	1498.50	26	0.81	0.35	0.35	0.38	0.38
5	1516.48	29	0.88	0.31	0.35	0.38	0.42
6	1536.12	30	0.88	0.43	0.50	0.40	0.46
7	1585.22	30	0.87	0.33	0.38	0.40	0.46
8	1616.61	33	0.89	0.33	0.42	0.39	0.50
9	1650.40	33	0.89	0.39	0.50	0.39	0.50
10	1668.26	28	0.88	0.46	0.50	0.36	0.38
11	1669.90	34	0.86	0.35	0.46	0.38	0.50
12	1678.92	34	0.87	0.44	0.58	0.38	0.50
13	1783.30	22	0.88	0.45	0.38	0.41	0.35
14	1837.96	32	0.86	0.28	0.35	0.38	0.46
15	1904.76	28	0.88	0.39	0.42	0.36	0.38
16	1957.79	28	0.85	0.39	0.42	0.36	0.38
17	2044.20	29	0.86	0.28	0.31	0.38	0.42
18	2049.19	32	0.87	0.31	0.38	0.38	0.46
19	2054.82	28	0.88	0.39	0.42	0.36	0.38
20	2060.22	30	0.88	0.40	0.46	0.40	0.46
Mean	1720.77	30.05	0.87	0.37	0.43	0.38	0.44

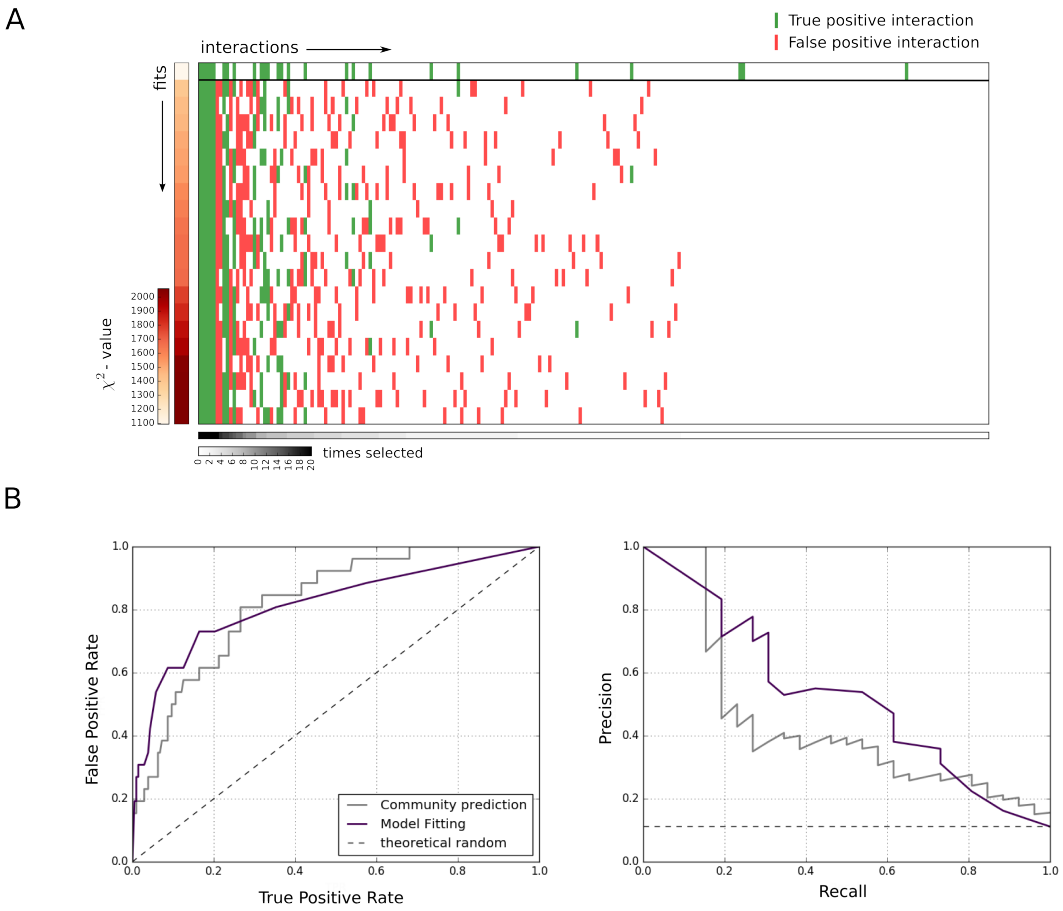


Figure 5.3: Evaluation of individual model fits to the benchmark data. (A) Along the x-axis true positive (green) and false positive (red) interactions of each model fit (y-axis) are indicated. Fits are sorted according to their final χ^2-value (indicated left), while interactions are sorted according to the frequency their appear across all 20 model fits (indicated below). The top row of the heatmap shows the result of fitting a model with only true interactions. **(B)** AUROC and AUPR curves for the community prediction (see Chapter 4) and a ranked prediction obtained from model fitting by summing up the occurrence of each interaction over all 20 fits.

and recall values. In 9 out of 20 cases, precision and recall values calculated from model fits were greater compared to the top ranking interactions of the community prediction, while in 2 cases equal performance was achieved (Table 5.1). Also in terms of average precision and recall values, the performance of both approaches in inferring the underlying gene regulatory network were comparable.

As already stated, calculating AUROC and AUPR values from individual fits was not feasible, since interactions selected by each fit were not ranked. However, to also be able to compare our model fitting strategy with the AUROC and AUPR values produced by the community prediction, we investigated a strategy in which multiple model fits were integrated. In this strategy, we summed up the number of times an interaction was selected by any of the 20 model fits and ranked interactions according to this sum. Interestingly, out of the top 6 interactions consistently chosen among all 20 model fits, 5 were indeed true (Figure 5.3A). When evaluating the ranked prediction obtained by

integrating individual model fits, we obtained an AUROC value of 0.82 and AUPR value of 0.52 (Figure 5.3B). For comparison, the performance of the community prediction resulted in an AUROC value of 0.82 and and AUPR value of 0.43.

In conclusion, the ability of the model fitting strategy to infer the structure of a gene regulatory network is comparable or even superior to the network inference approach applied in Chapter 4. In the next section we demonstrate that a combination of both methods, in which we include structural predictions from data-driven network inference while fitting the dynamical model to gene expression data, can yield even further improvement in reconstructing the network topology.

5.2.4 Integrating network inference and model fitting increases predictive power

Having shown that both, the proposed model fitting strategy and network inference methods applied in Chapter 4, perform well in predicting the structure of the benchmark network from simulated gene expression data, we next investigated whether a combination of both approaches could yield even further increase in performance. We implemented this strategy by adding a regularization term to the cost function, which is minimized while fitting the model to the data. This regularization term penalizes lowly ranked, i.e. less likely interactions, according to the community prediction obtained by integrating data-based network inference methods. The final cost function then consists of the weighted least squares distance between simulated and measured gene expression values, plus the regularization term describing interaction specific penalties derived from the community prediction:

$$\chi^2(\theta) \quad = \quad \sum_{i=1}^{n} \frac{(y_i^{model}(\theta) - y_i^{data})^2}{\sigma^2} \quad + \quad \alpha \left(\frac{1}{\nu} \sum_{a \in G, b \in G} \beta^{ab} \pi^{ab} \right). \tag{5.3}$$

In this extended cost function, β^{ab} represents regulation of gene a by gene b, with a value of 1 implying its existence and a value of 0 its absence. The penalty value assigned to the corresponding interaction of a being regulated by b is given by π^{ab}. Finally, the sum of all interaction penalties is normalized by the total number of interactions in the network ν.

In the formulated two-component approach, a good balance between the model fit, and agreement between model structure and the interactions of the community prediction must be established. Accordingly, the degree with which penalization occurs can be tuned via a regularization parameter α. By increasing the regularization parameter α the contribution of the model fit to the overall cost function decreases, while more weight is assigned to the match between model structure and the community prediction. Accordingly, for different values of the regularization parameter α each 20 model fits to

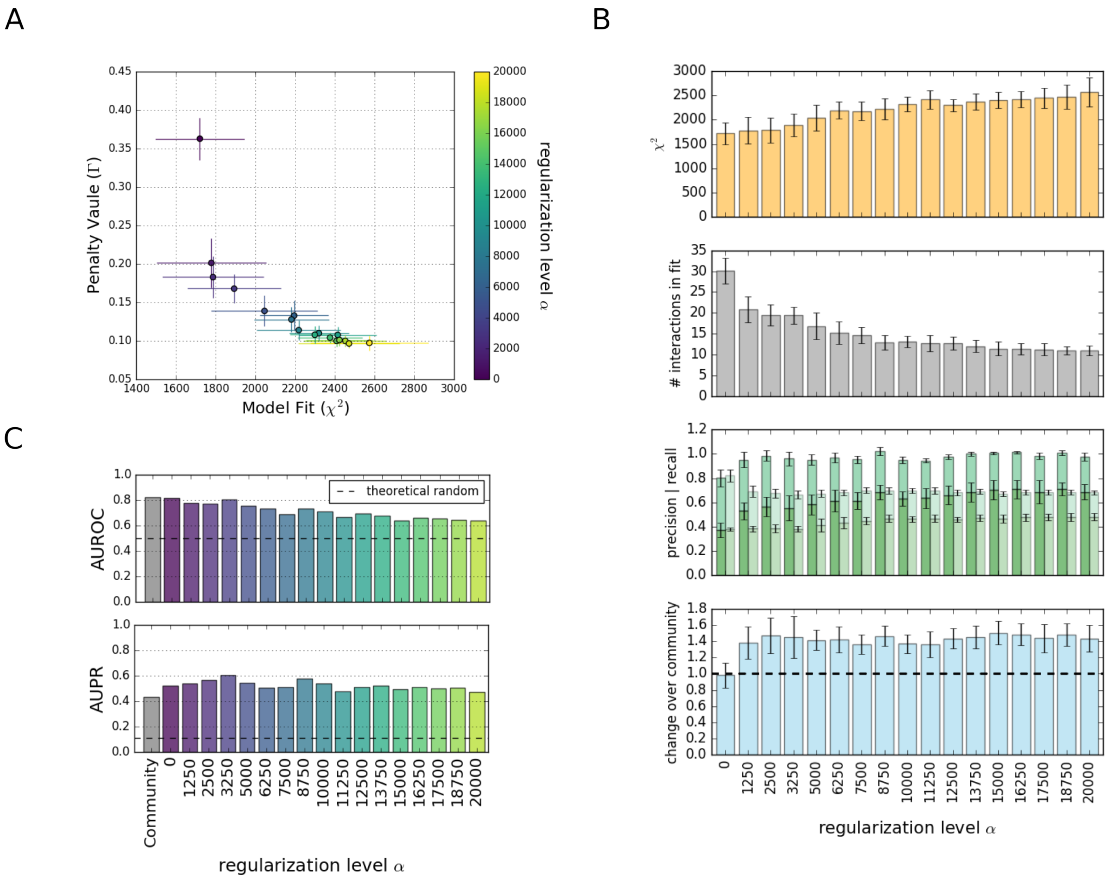


Figure 5.4: Evaluation of regularized model fitting approach. (A) Scatter plot of mean penalty value (Γ) and weighted least squares (χ^2) for different values of α. **(B)** Mean final χ^2, number of interactions, precision and recall, and fold-change of performance compared to the community prediction for model fits with each level of α. **(C)** AUROC and AUPR values of integrated model fits for each level of the regularization parameter parameter α. Corresponding performance values of the community prediction (see Section 4.2.4) are shown in gray.

the benchmark data were carried out. The anticipated trade off between model fit and agreement with interactions predicted by network inference is shown in Figure 5.4A. In theory, this graphical representation also known as the L-curve, can aid in selecting the proper regularization parameter [Gabor and Banga, 2015], with the optimal choice of α located approximately at the inflection point of the curve.

When increasing the regularization parameter α, an increase of the distance between modelled and measured data was observed, while simultaneously the value of the penalty term decreased. In addition, also the number of inferred interactions by each model fit decreased with higher levels of α (Figure 5.4B - second panel). As the combined cost function considers the mean penalty over all chosen interactions, and therefore the absolute number of interactions was not penalized, this effect may only have resulted from the distribution of penalty values over all interactions (Figure 9.15), with only very few interactions being assigned a low penalty. As a consequence, as the weight of penalization term is increased, fewer interactions corresponding to a low penalty are chosen.

Analogous to the evaluation of the unregularized model fitting approach ($\alpha = 0$), we compared performance of the two-component approach using precision and recall

values. Mean precision values of the regularized model fitting approach for each level of α fell between 0.53 ($\alpha = 1250$) and 0.71 ($\alpha = 18750$), while recall values between 0.29 ($\alpha = 20000$) and 0.42 ($\alpha = 1250$) were observed (Figure 5.4B - third panel and Table 9.7). When comparing the individual model fits with the community prediction, all 320 fits carried out with a regularization value $\alpha > 0$, displayed superior performance in terms of precision and recall. Accordingly, mean precision and recall values from the regularized model fitting approach showed a performance increase between 36% ($\alpha = 11250$) and 50% ($\alpha = 15000$) compared to the community prediction alone (Figure 5.4B - third and bottom panel).

For each level of the regularization parameter α we further generated a ranked prediction by summing up the number of times an interaction was selected. Following this strategy, we obtained AUROC values of up to 0.82 ($\alpha = 0$) and AUPR values of 0.60 ($\alpha = 3250$) (Figure 5.4C). Although AUROC values of the community prediction and the regularized model fitting approach were comparable over a large range of α, the performance of the regularized model fitting approach in terms of AUPR was significantly increased compared to the community prediction (AUPR: 0.43). In general, we observed a decrease of performance of the model fitting approach with larger levels of α. The optimal performance of the regularized model fitting approach was reached with a regularization parameter $\alpha = 3250$, which also presented a good balance between model fit (χ^2-value) and agreement with the community prediction (Γ).

As demonstrated, combining model fitting and network inference strategies resulted in increased performance in inferring the structure of the benchmark network, compared to each strategy on its own. This improvement however depends to a large extent on the choice of the regularization parameter α, describing the balance between the model fitting the data and agreement of model structure with the community prediction.

5.2.5 Applying combined network inference and model fitting to NMuMG gene expression data

Having thoroughly evaluated the combined model fitting approach on the benchmark data, we applied it to the EMT expression data comprising both time-course and knockdown experiments measured in NMuMG cells. Again we chose different levels of the regularization parameter α to penalize less probable interactions according to the community prediction obtained via network inference. For a value of $\alpha = 0$, implying no regularization, 35 fits were carried out, while for each level $\alpha > 0$ an additional 15 fits were performed. As in the case of the benchmark data, with increasing levels of the regularization parameter α, we observed a trade-off between the quality of model fit and agreement with the community prediction (Figure 5.5A).

As expected, the overall best model fit was obtained with a regularization parameter $\alpha = 0$. This fit predicted 41 interactions and resulted in a χ^2-value of 1151.21, which corresponds to an average χ^2 of 1.18 per data-point (Figure 9.16). The Pearson correlation between measured and simulated gene expression was 0.89 (Figure 9.17). Interestingly, in this fit we observed an average penalty per interaction over the 41 selected interactions of

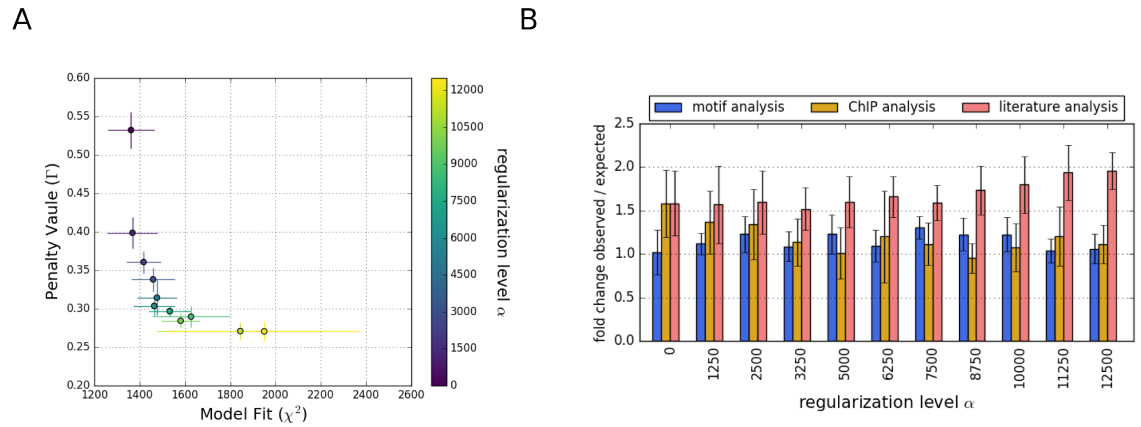


Figure 5.5: Regularized model fitting applied to NMuMG gene expression data. **(A)** Scatter plot of mean final χ^2-values and mean final penalty values (Γ) for each level of α, describing the trade-off between model fit and penalty term **(B)** Mean fold-change of expected vs observed matching interactions between model fits and interactions derived from motif analysis (blue), ChIP analysis (orange), or literature analysis (red)

0.50 compared to an average penalty of 0.57 over all ranked interactions. This indicates, that although the community prediction was not taken into account while minimizing the cost function, interactions partially matched with those inferred by the community prediction.

In the benchmark data a more detailed analysis of the performance of the regularized model fitting approach was possibly by comparing selected interactions of each model fit with the structure of the benchmark network. The true structure of the EMT network however is not known. We therefore compared model fits to the available information on putative interactions derived by literature-, motif-, or ChIP-analysis (see Chapter 4).

Out of all 125 fits carried out, only two fits showed significant overlap between interactions derived by motif analysis and interactions selected by model fitting (hypergeometric test, $p < 0.05$). In one of these model fits, performed with a regularization parameter of $\alpha = 2500$, 13 interactions were supported by motif analysis, while the other fit ($\alpha = 0$) produced an overlap of 14 interactions. For comparison, the expected overlap produced by selecting the same number of interactions at random is 8.6 and 8.2 respectively. In general, there is no visible trend between the regularization parameter α and the overlap of interactions derived from motif analysis and interactions selected by each model fit (Figure 5.5B).

Interactions derived from analysis of the publicly available ChIP data were significantly enriched in interactions selected by 5 model fits (hypergeometric test, $p < 0.05$). The majority of these model fits (4 fits), were carried out with an α value of 0. Over all fits, regardless of significant enrichment, the mean ratio of observed to expected overlap between interactions was highest for model fits applying no regularization and steadily declined with increasing level of alpha.

While hardly any significant overlap between interactions selected by the model fits and interactions predicted by motif-, or ChIP-analysis could be observed, there exist 36 fits with a significant enrichment of interactions for which experimental evidence was reported in the literature (hypergeometric test, $p < 0.05$). The largest overlap was observed

in a fit performed with regularization parameter $\alpha = 1250$, with 9 out of 27 selected interactions supported by the literature. Selecting the same number of interactions at random would have by chance produced an average overlap of 3.6 interactions. For comparison, the overlap between the top 25 ranked interactions in the best performing method of Chapter 4 and interactions reported in the literature was 7. In case of the model fits, over all levels of α the mean observed versus expected overlap between interactions selected by model fits and known interactions was above 50%, increasing with higher levels of α.

In conclusion, interactions selected by the combined model fitting strategy applied to the NMuMG gene expression data were hardly supported by the analysis of TF motif occurrences or publicly available ChIP data. However, we did observe a significant overlap between interactions resulting from combined model fitting and interactions reported in the scientific literature.

5.2.6 Evaluation of model fitting strategy using known interactions

As noted, we detected a significant overlap between interactions selected by each model fit originating from the combined model fitting approach and interactions reported in the literature. In more detail, this overlap increased with higher levels of the regularization parameter α. At the same time however, the quality of the model fit declined. Therefore, we again turned to inspection of the L-curve, in order to define an optimal value of the regularization parameter α for the case of combined model fitting applied to NMuMG gene expression data (Figure 5.5). In case of the benchmark data, optimal performance in reconstructing the structure of the underlying network was reached with regularization parameter of $\alpha = 3250$. At the same time, fits for this particular regularization parameter were located in the vicinity of the inflection point of the L-curve, balancing both model fit and agreement of model structure with the community prediction. In case of the NMuMG data, the L-curve suggested an optimal trade off between the two features of the model fitting approach using a regularization parameter of $\alpha \approx 5000$.

Among the 150 fits performed with $\alpha = 5000$, the best overall fit resulted in a χ^2-value of 1202.2, which corresponds to an average χ^2 of 1.23 per data-point (Figure 9.18). On average, the 150 fits produced a χ^2-value of 1462.3 relating to a mean χ^2-value of 1.5 per data-point. With regard to the penalty term, the best match between model topology and the community prediction was observed for a fit, which showed an interaction penalty (Γ) of 0.28 over 24 selected interactions. The mean penalty over all model fits was 0.31 combined with an average of 29.5 interactions selected by each fit.

Out of the 150 model fits, 39 observed a significant enrichment of interactions supported by the literature. The fit achieving the greatest enrichment of known interactions contained 8 interactions confirmed by evidence from the literature, out of 27 selected interactions. Additional model fits also showed an equal number of overlapping interactions with those reported in the literature. At the same time, these fits however produced a higher number of total interactions, hereby resulting in comparatively lower enrichment.

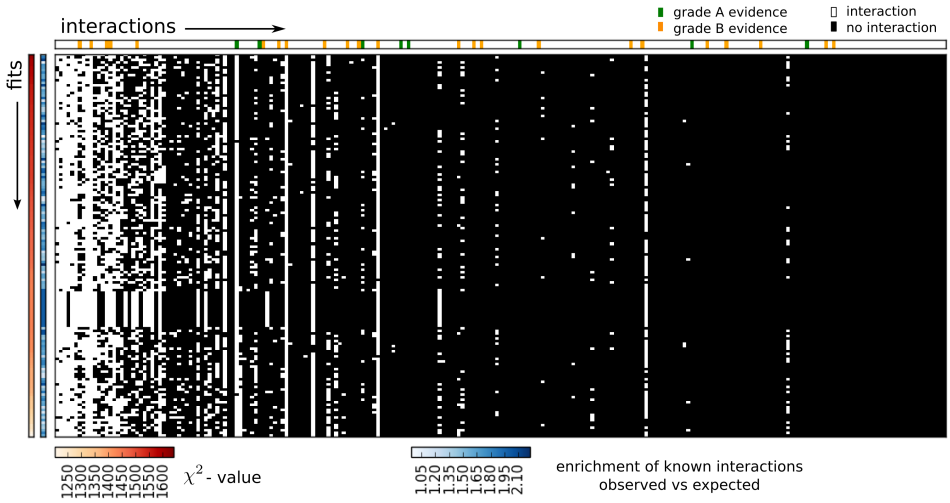


Figure 5.6: Summary of model fits applied to NMuMG gene expression data. Shown are interactions chosen by each model fit with regularization parameter $\alpha = 5000$. Interactions are sorted according to their rank in the community prediction (x-axis), while fits are sorted by their final χ^2-value (y-axis - indicated in orange). Blue column shows enrichment of known interactions (grade A and B) in interactions chosen by each fit. In the upper plot interactions supported by the literature are indicated (grade A evidence - green, grade B evidence - orange).

Interestingly, 15 out of the 150 performed fits resulted in exactly the same model structure, with all fits showing 8 interactions supported by the literature.

As in the benchmark data, some interactions appeared consistently across all fits with $\alpha = 5000$ (Table 9.8). Remarkably, consistently appearing interactions are distributed across all ranks in the community prediction, meaning that although some interactions have a high penalty, they seem to produce a better model fit than other interactions associated with a lower penalty (Figure 5.6).

Interestingly, 4 out of the top 5 most persistently chosen interactions were supported by evidence reported in the literature. For example, one of the interactions selected across all model fits was regulation of JunB by Smad. Indeed, there exists evidence of Smad4 knock-down inhibiting TGFβ-dependent JunB up-regulation in NMuMG cells [Gervasi et al., 2012]. Another persistently chosen interactions was regulation of Fn1 by Sox4. Accordingly, Fn1 up-regulation was observed in NMuMG cells ectopically expressing Sox4, while TGFβ-dependent upregulation of Fn1 was blocked by Sox4 knock-down [Tiwari et al., 2013]. Regulation of Snail1 by Smad appeard in 148 out of the 150 performed model fits. As mentioned in Section 4.2.1, there is evidence of direct regulation of Snail1 by Smad proteins.

Among the 150 fits carried out, 151 out of 233 possible interactions never appeared in any of the fits. Ideally these should represent interactions which are not existent in the EMT network. However, 6 out of the 8 interactions for which evidence for a direct regulation exists, were included within the set of interactions (Atf3 $\rightarrow$ Id2, Hmga2 $\rightarrow$ Cdh1, Hmga2 $\rightarrow$ Snail1, JunB $\rightarrow$ Id2, Zeb1 $\rightarrow$ Cdh1, Zeb2 $\rightarrow$ Cdh2). There are multiple potential causes for this observation: Possibly the regularization parameter α was set to

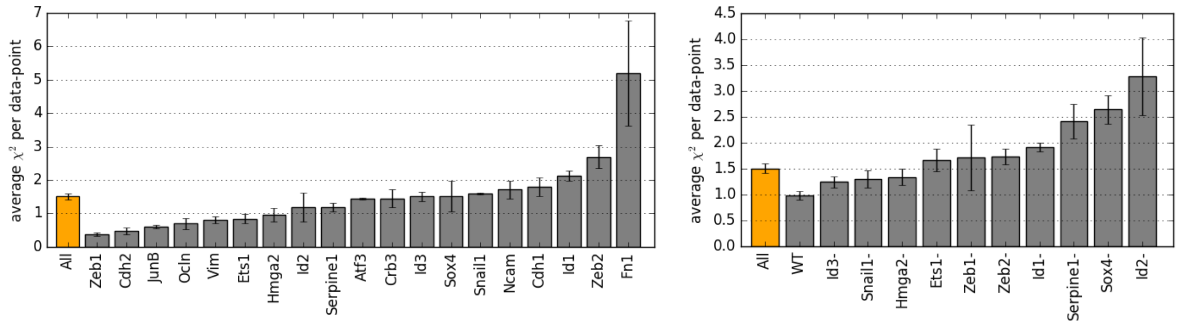


Figure 5.7: Mean residuals grouped by target or cell-line. Shown is the mean χ^2-value per gene (left) or cell-line (right) over all fits carried out with regularization level $\alpha = 5000$. The overall mean χ^2-value per data-point is indicated in orange.

high, leading to interactions associated with a high penalty but also evidence of direct regulation to be selected less likely by the fitting approach. Also, in some cases although evidence of a direct interaction may exist, it is possible that the measured data was fit sufficiently well without it, hinting at non-identifiability issues of the data in combination with model complexity.

In summary, while performing fits with a regularization parameter $\alpha = 5000$, we observed a significant overlap of interactions selected by model fits with interactions reported in the literature. Further, interactions frequently appearing across all model fits match well with those supported by experimental evidence reported in the literature.

5.2.7 Evaluation of model fits based on model residuals

Across al fits not all genes were equally well described. We therefore calculated average χ^2-values per data-point for each gene separately. The mean normalized χ^2-value per data-point over all fits with $\alpha = 5000$ was 1.5 (Figure 5.7). Among all fits, Zeb1 expression was described best with an average normalized χ^2 of 0.37. Fn1 on the other hand was fit less well with an average normalized χ^2 of 5.2. Similarly, we calculated the average normalized χ^2-value for each cell-line. Here the wild-type cell line was best described across all model fits, with an average normalized χ^2 of 0.98, while Id2 deficient cells exhibited the largest contribution to the overall χ^2 with an average normalized χ^2 of 3.28.

Clustering the individual model fits based on the deviations between modelled and measured gene expression can provide information on different classes of model fits, with groups of models describing different target genes with different quality. For example Zeb1 was described well over all model fits (Figure 5.8), while other genes exhibited larger differences in their contribution to the model fit. Id2, for example, showed above average contribution to the overall difference between modelled and measured gene expression in a subset of fits, while in other fits this contribution to the overall χ^2 was below average. Further, it is possible that an antagonistic relationship between different features of the

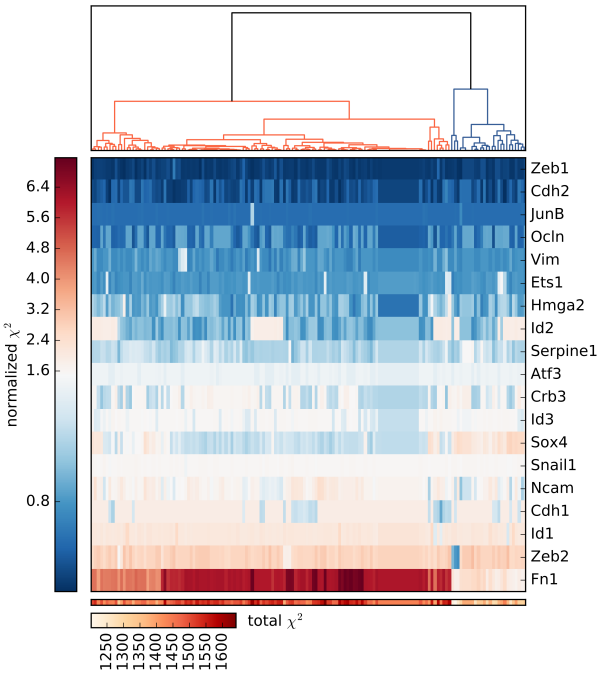


Figure 5.8: Model fits clustered by their average χ^2-values per gene. Heatmap shows model fits clustered by the average χ^2-values per gene. Red denotes average χ^2-values above the total average, while blue denotes mean χ^2-values below. The dendrogram in the top panel shows the clustering trajectory. Distinct clusters are marked by in red or blue. Total χ^2 values of each model fit are given in the column underneath the heatmap.

model fits exists, in which certain genes are described better, at the cost of other genes being less well explained by the model. Indeed there seems to be a trade-off in the representation of Fn1 and Sox4 gene expression, in the sense that whenever a model fits the gene expression measurements of one gene well, expression for the other gene is less well reproduced.

5.2.8 Evaluation of model fits based on model predictions

In addition to analysing models with regard to their structural features or differences between modelled and measured gene expression, model simulations can be carried out, hereby predicting the dynamic behaviour of gene expression under various assumed experimental conditions. One hallmark of EMT is the down-regulation of Cdh1 during later stages of EMT. When simulating expression of Cdh1 past 24h, we could observe that only 29 out of all 150 model fits showed Cdh1 down-regulation of at least 10% within the first 14 days of TGFβ-treatment Figure 5.9). Out of these 29 fits in which Cdh1 was down-regulated, 9 predicted elevated Cdh1 down-regulation of at least 50%.

An additional observation reported in the literature is the complete phenotypic reversal of EMT after removal of the TGFβ ligand or inhibition of TGFβ/Smad-signalling.In the experimental literature this reversibility was observed as late as 30-60 days after TGFβ treatment and was shown to occur also on a molecular level, with the expression of Zeb2, Cdh1, Cdh2 and Snail1 returning to almost basal levels within 7 days after

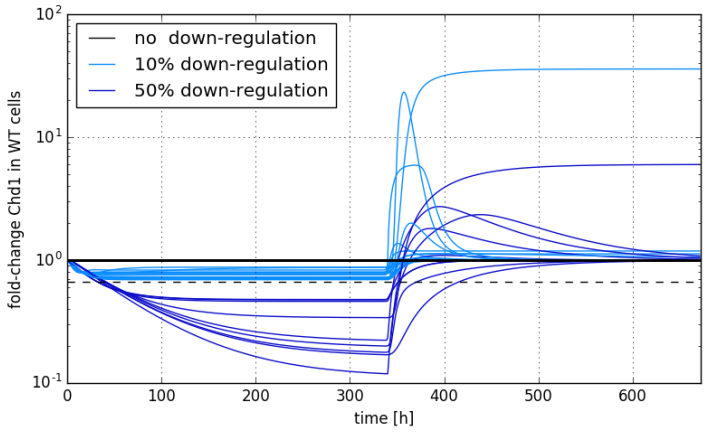


Figure 5.9: Prediction of Cdh1 gene expression beyond 24h of TGFβ-treatment. Plots show predicted dynamics of Cdh1 expression until 28 days (672h) after TGFβ-treatment. After 14 days a reduction of pSmad2 expression to basal levels is assumed. Shown in black are predictions of models fits in which no down-regulation of Cdh1 occurred, lightblue indicate predictions of model fits with 10% down-regulation of Cdh1, while dark blue shows predictions of model fits exhibiting at least 50% down-regulation.

TGFβ removal [Maeda et al., 2005]. In order to check the ability of the different models to predict this return of gene expression to basal levels, we simulated gene expression time-courses after depletion of pSmad2 14 days after TGFβ-treatment and again examined expression levels of Cdh1. Out of the 29 models showing Cdh1 down-regulation in the wild-type time-course, 13 predicted an overshoot of Cdh1 expression after pSmad2 depletion. A return of Cdh1 to its initial gene expression sate was observed for 27 models, while 2 models predicted sustained up-regulation of Cdh1 even after removal of pSmad2.

As an additional analysis, instead of focusing on Cdh1 gene expression only, we further predicted the reaction of all genes in the EMT network to the removal of the pSmad2 input. Model predictions were then clustered based on simulated gene expression dynamics after pSmad2-depletion. As a result, model predictions roughly split into 4 groups (Figure 5.10): While cluster 1 displayed a gradual return of gene expression to unstimulated levels, cluster 2 observed sustained gene activation particularly in Cdh2, Fn1, Ncam, and Sox4 even days after pSmad2-removal. Cluster 3 again predicted a return of gene expression to basal levels, albeit more rapidly than seen in cluster 1. Cluster 4 consisted of a small group of models predicting an overshoot of Crb3, Id1, Id2, and Id3 gene expression after pSmad2-removal, together with sustained activation of Fn1, Ncam, as well as multiple other genes.

Over all fits, both Atf3 and Snail1 returned to within 10% of their basal expression level within 14 days after inhibition of Smad-signalling. Other genes in the network exhibited a more variable pattern of sustained or transient gene expression. In 41 out of the 150 fits, Ncam was expressed above basal expression levels even 14 days after pSmad2-depletion. In total, 49 model fits show some form of memory effect in gene expression even days after inhibition of EMT, while 101 model fits predicted a return of all EMT genes to basal levels within 2 weeks after TGFβ-removal.

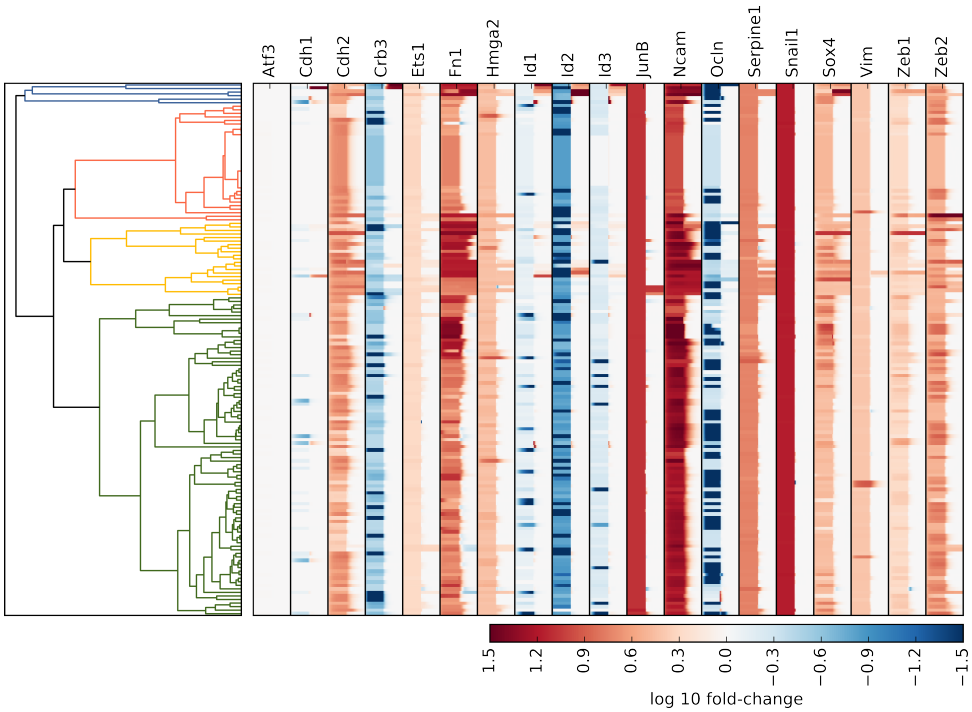


Figure 5.10: Model fits clustered by prediction of gene expression dynamics in EMT reversal experiment. The heatmap shows the predicted behaviour of EMT genes in the 150 model fits carried out with $\alpha = 5000$, assuming a decrease of pSmad2 to basal expression after 14 days of TGFβ-treatment. Fits are clustered according to simulated model dynamics after TGFβ-removal.

5.3 Discussion

The current view on EMT is that of a gradual, multi-state process with many intermediate meta-stable states, in which important epithelial and mesenchymal markers are expressed at intermediate levels [Nieto et al., 2016, Chang et al., 2016]. A dynamic model of the gene regulatory network controlling EMT more realistically captures such complex gene expression changes, compared to a static representation of the structure of the EMT network.

In most systems biology studies, the formulation of a dynamical model depends on the integration of experiments providing evidence of regulatory interactions present in the system. In the case of the gene regulatory network controlling EMT however, such experimental evidence on specific regulatory interactions between EMT relevant genes remains incomplete, hereby preventing a full reconstruction of the structure of the network based on literature information. As an alternative strategy, we therefore attempted to infer the structure of the gene regulatory network controlling EMT by applying various data-based network inference approaches to an extensive time-course gene expression dataset measured under wild-type as well as knock-down conditions. Although the integration of multiple network inference methods applied to gene expression data in Chapter 4 was able to constrain the structural topology of the EMT network to some extent, we further investigated the possibility of reconstructing the structure of gene regulatory network using a dynamic model of gene expression. The strategy followed in this approach

was based on the formulation of a generic model of gene expression dynamics, in which only few constraints on the topology of the underlying gene regulatory network were imposed. The mathematical formulation of this network model contained both real-valued parameters, describing kinetic rates, but also integer-valued parameters, encoding the structural features of the network. Both, kinetic and structural features of the model, were estimated by fitting the dynamical model to the available gene expression data.

On its own, estimating real-valued parameters poses a difficult optimization problem, due to the high-dimensionality of parameter space, non-convexity of the problem solution, and the existence of multiple local optima. Introducing integer parameters to the model further increases its difficulty, since dependence of the cost function on model parameters is no longer described by a continuous function. Despite this challenge, attempts to simultaneously infer model structure and model parameters using mixed integer optimization have been made in the past. In a study on the *E. coli* KdpD/KdpE system, a MINLP problem containing a total of 17 real valued, 5 integer valued and 3 binary parameter values was solved, hereby discriminating between different possible model structures while at the same time inferring model parameters [Rodriguez-Fernandez et al., 2013]. In a similar case study, focusing on signalling in liver hepatocytes, the formulated MINLP problem consisted of 135 continuous and 109 integer valued parameters [Henriques et al., 2015].

The model formulation of the gene regulatory network controlling EMT presented here, includes a total of 283 real-valued and 233 integer-valued parameters, once again increasing the complexity of the model fitting approach. In order to leverage our chances of solving this challenging optimization problem, we modified the model fitting approach using two separate strategies: First we supplied the chosen optimization strategy with high quality trial solutions, obtained from fitting potential one-to-one regulator-target interactions [Guillén-Gosálbez et al., 2013]. As a result, convergence of the optimization algorithm was increased, while at the same time the number of selected interactions by each model fit remained within a reasonable range. Secondary, the cost function was extended in such a way, that interactions selected by each model fit matched those given by a structural prior of the EMT network. In our case, this structural prior originated from integrating multiple data-driven network inference approaches Chapter 4.

Using a realistic benchmark network, we could show that the modified fitting procedure converges to minima with a cost function comparable to that of fitting a model with known interactions. We further demonstrated that the combined model fitting strategy performed better in predicting the networks topology than the unregularized model fitting approach or any of the network inference methods tested in Chapter 4. As an additional benefit, model fits provided the ability to examine gene regulatory networks not only from a structural point of view, but also according to their ability to reproduce measured gene expression data and predict dynamics of gene expression under various perturbation conditions.

Instead of using known interactions to evaluate model fits, in the future the model fitting strategy could further benefit from including these interactions into the structure of the model. Hereby, modelled gene regulatory networks would not only more faithfully capture our understanding of the EMT network, but also the complexity of the optimiza-

tion problem would be reduced. In turn, decreased dimensionality of the optimization problem would open up the possibility of withholding a subset of gene expression measurements from model fitting. In a cross-validation strategy, in which gene expression measurements of the test data is predicted from fitting a model to initial training data, these preliminarily retained gene expression measurements could then be used as a means to more reliably assess the optimal regularization parameter balancing model fit, predictive power of the model, and agreement of model interactions with the structural prior.

As demonstrated by an abundance of studies on EMT, it is clear that this transdifferentiation event is a highly complex process, regulated on many different layers of cell signalling, gene expression, and cell or tissue morphology. In its current form however, the formulated dynamic model of EMT is focused exclusively on transcriptional cross-regulation between selected EMT genes, accompanied by a number of further simplifying assumptions. For example, degradation rates in the model are independent of the expression of other factors in the network. In contrast to this assumption, an increase of Zeb1 protein half-life upon TGFβ-treatment could be shown [Dave et al., 2011]. Moreover, discrepancies in gene expression on the level of mRNA and protein have been observed for at least a subset of EMT genes (Chapter 4), further highlighting the relevance of transcriptional and post-transcriptional gene regulation to EMT. Therefore, in order to more realistically capture the complexity of gene regulation governing EMT, more detail on the mechanistic features of transcriptional and post-transcriptional control should be added to the model. Extension of the model however, would first require the further collection and integration of experimental data, describing gene expression on the level of mRNA as well as protein.

At this point, it "seems to be a daunting task to construct a systems landscape that encompasses the regulatory networks involving miRNA-TF loops, transcript processing, and the epigenetic control of protein stability to determine what controls the continuous transitions through the complete EMT spectrum." [Nieto et al., 2016]. Despite the challenging setting and the imperfection of formulated models, dynamic models generated via the combined network inference and model fitting strategy can already be analysed with respect to their structural and functional properties, and further provide a valuable primer for a deeper understanding of gene expression dynamics governing EMT.

Chapter 6

Quantifying Post-Transcriptional Regulation in the Development of *D. melanogaster*

Preamble

This project was carried out in collaboration with NCV, AB, ND, SSP, MD, JYR, FB, and SL.

6.1 Introduction

In the previous chapters, analysis of gene expression patterns and inference of gene regulatory networks was based exclusively on mRNA expression data. However, both in the gene regulatory network controlling circadian rhythm, as well as the transcriptional network of EMT, post-transcriptional gene regulation is known to occur. In addition, generally a limited correlation between mRNA and protein expression has been observed. Accordingly, the need to integrate both levels of gene expression has been expressed [Gomez-Cabrero et al., 2014].

In this chapter we therefore integrate paired transcriptome and proteome time-course data with 14 time points measured in biological quadruplicates during *Drosophila* embryogenesis. Mathematical models are used to classify genes based on their kinetic mode of translation, including the identification of a set of genes potentially post-transcriptionally regulated during *Drosophila* embryogenesis. The analysis is followed up by extensive bioinformatics analysis in order to determine biological functions enriched within each defined group and further propose potential mechanisms of post-transcriptional regulation.

6.1.1 Correlation between mRNA and Protein

According to the central dogma of molecular biology, protein is translated from mRNA, suggesting that mRNA levels can be predictive of protein concentrations. However, the relationship between mRNA and the concentration of its protein does not follow the simple monotonic correlation that higher mRNA levels always relate to concordantly more protein. This non-trivial mRNA-protein relation is a widespread observation ranging from

yeast to human [Beyer et al., 2004, Bonaldi et al., 2008, Brockmann et al., 2007, Eichelbaum and Krijgsveld, 2014, Fournier et al., 2010, Gouw et al., 2009, Greenbaum et al., 2003, Griffin et al., 2002, Gygi et al., 1999, Le Roch et al., 2004, Li et al., 2014, Maier et al., 2011, Nagaraj et al., 2011, Schrimpf et al., 2009, Schwanhäusser et al., 2011, Taniguchi et al., 2010, Wu et al., 2008] (reviewed for example in [de Sousa Abreu et al., 2009, Liu et al., 2016, Maier et al., 2009]), that could arise from extensive post-transcriptional gene regulation in eukaryotic organisms. One important mechanism of post-transcriptional gene regulation is controlled protein translation. In line with translational regulation of protein expression, global protein pulse labelling studies in mammalian cells revealed that cellular mRNAs show distinct translation rates [Schwanhäusser et al., 2011]. Accordingly, it was demonstrated that the ribosome occupancy of a transcript is a better predictor of protein expression when compared to its mRNA concentration [Ingolia et al., 2009]. Furthermore, protein levels are subject to active degradation via the Ubiquitin-proteasome system, controlling protein degradation rates independently of its transcript abundance [Goldberg, 1995, McShane et al., 2016].

Even in the absence of regulated protein translation or turnover, a mismatch between mRNA and protein levels can occur, because of a temporal delay that occurs when protein is translated from mRNA. Accordingly, it has been noted that while mRNA-protein correlations are generally low, the discrepancy is even more pronounced during dynamical cellular transitions [Liu et al., 2016]. Theoretical analyses revealed that the delayed dynamics of a protein relative to its corresponding mRNA during a dynamical transition lead to a non-linear relationship between the two species [Becker et al., 2013, Gedeon and Bokes, 2012, Lee et al., 2011]. This is influenced by many factors including protein turnover rates [Taniguchi et al., 2010], as long protein half-lives lead to slower dynamics of protein accumulation. Furthermore, at these cellular transitions active regulation of protein translation and turnover is enforced. For instance, RNA binding proteins, micro-RNAs, or RNA modifications can affect processing, translation and/or turnover of a transcript [Bushati et al., 2008, Filipowicz et al., 2008, Glisovic et al., 2008, Roignant and Soller, 2017, Roundtree et al., 2017, Valencia-Sanchez et al., 2006].

These complex dynamics of post-transcriptional regulation make it difficult to intuitively understand the relations between mRNA and protein expression changes. Mechanistic models of protein translation provide a possible mathematical framework for an improved understanding of gene expression regulation. For instance, we investigated the spatio-temporal mRNA-protein relationship for three *Drosophila* gap genes during early embryonic development. Using simple models of protein translation, we concluded that in these cases protein merely represents a time-delayed version of its corresponding mRNA, i.e., that no post-transcriptional regulation needs to be assumed [Becker et al., 2013]. On a genome-wide level, Peshkin et al. (2015) concluded for *Xenopus laevis* development that the majority of protein expression changes can be explained by assuming proportional protein production from mRNA and first-order protein degradation [Peshkin et al., 2015]). Likewise, Jovanovic et al. (2015) conclude that simple models of protein translation can account for most of the variance in protein expression in dendritic cells responding to external lipopolysaccharide treatment [Jovanovic et al., 2015]. Going even further, they identify mRNA abundance as having the largest contribution in explaining

the variance in relative protein expression changes.

Drosophila embryogenesis poses a particular interesting case to study post-transcriptional regulation. Although the early embryo is in principle capable of transcription, no transcription takes place during the first two to three hours of development after oocyte fertilization. Instead, the embryo completely relies on maternally deposited mRNA and protein until the transcription of embryonic genes is switched on in a process called maternal-to-zygotic transition (MZT). Thus, the developmental dynamics occurring within the first hours after fertilization need to be controlled at the post-transcriptional level [Gouw et al., 2009, Tadros and Lipshitz, 2009]. For example, it is well established that the maternally deposited positional information genes hunchback, caudal, nanos, and oskar are regulated post-transcriptionally [Murata and Wharton, 1995, Rivera-Pomar et al., 1996, Sonenberg and Hinnebusch, 2009]. On a global scale, Qin et al. (2007) measured ribosome occupancy of 10,000 transcripts at various time points after deposition of fertilized eggs, concluding that post-transcriptional regulation is a widespread phenomenon in the developing embryo [Qin et al., 2007]. Given that RNA-binding proteins (RBPs) are major regulators of post-transcriptional gene regulation, it is not surprising that the mRNA-bound proteome is highly dynamic during MZT [Sysoev et al., 2016].

In this study, we investigate the relation between mRNA and protein levels during *Drosophila* embryogenesis (including MZT) using highly time-resolved paired transcriptome/proteome measurements. We fit gene expression models of different complexity to thousands of mRNA-protein profiles to infer the underlying molecular mechanisms of (post-) transcriptional regulation.

6.2 Results

6.2.1 Paired mRNA/protein measurements reduce experimental variation

We previously followed proteome changes during *Drosophila* development with high temporal resolution [Casas-Vila et al., 2017]. This dataset included measurement time points every 1h within the first 6h after fertilization, and every 2h hereafter until 20h (Figure 6.1A). When comparing our proteome dataset to the published developmental RNA-Seq time-course [Graveley et al., 2011], we observed limited correlation between RNA and protein levels. To further investigate post-translational gene regulation in more detail and to exclude that the discrepancies between RNA and protein levels arose from experimental variation between laboratories, *Drosophila melanogaster* strains, and chosen measurement time points, we generated a new developmental RNA-Seq dataset from exactly the same fly embryo collections previously used for our proteome measurement (Figure 9.19).

Our transcriptome matches the published RNA-Seq dataset when calculating pairwise correlations of all common reads between samples (Figure 9.19 and Figure 9.20A). As expected, the strongest correlation between both datasets was observed when comparing similar developmental time points (marked with + signs in Figure 9.20A). Of note, the correlation of corresponding developmental time points tended to be higher early in development, and declined at later stages, hinting at a systematic deviation between both datasets. Indeed, late in development, the maximum correlation of our time points occurs with slightly earlier developmental time points of the previously published RNA-Seq data. This suggests that the embryonic development in the published dataset was accelerated when compared to our conditions.

We also observed a strong correlation of both RNA-Seq datasets when relating mRNA time-courses for individual genes: The average Spearman correlation coefficient over all genes was $\rho = 0.7$, and only very few genes showed insignificant or negative correlation (Figure 9.20B). To test for our hypothesis of altered developmental speed, we introduced a stretching factor by which we scaled the time axis in the previous dataset to various degrees, while quantifying the match between datasets using ordinary least squares. In line with faster developmental progress in Graveley et al. (2011), we could increase the overlap between both datasets by assuming a 8 acceleration (ca. 1.5h at the last measured time point) in our developmental progression (Figure 9.21).

This observation highlights that laboratory conditions can indeed affect the kinetics of gene expression even for such a robust biological process as embryo development. *Drosophila* development is subject to manipulations under different temperature conditions, which offers a potential explanation for the discrepancy observed between datasets. Although in our case, both RNA-Seq datasets show similarity, pairing of RNA and proteome measurements avoids the risk of systematic mis-estimation.

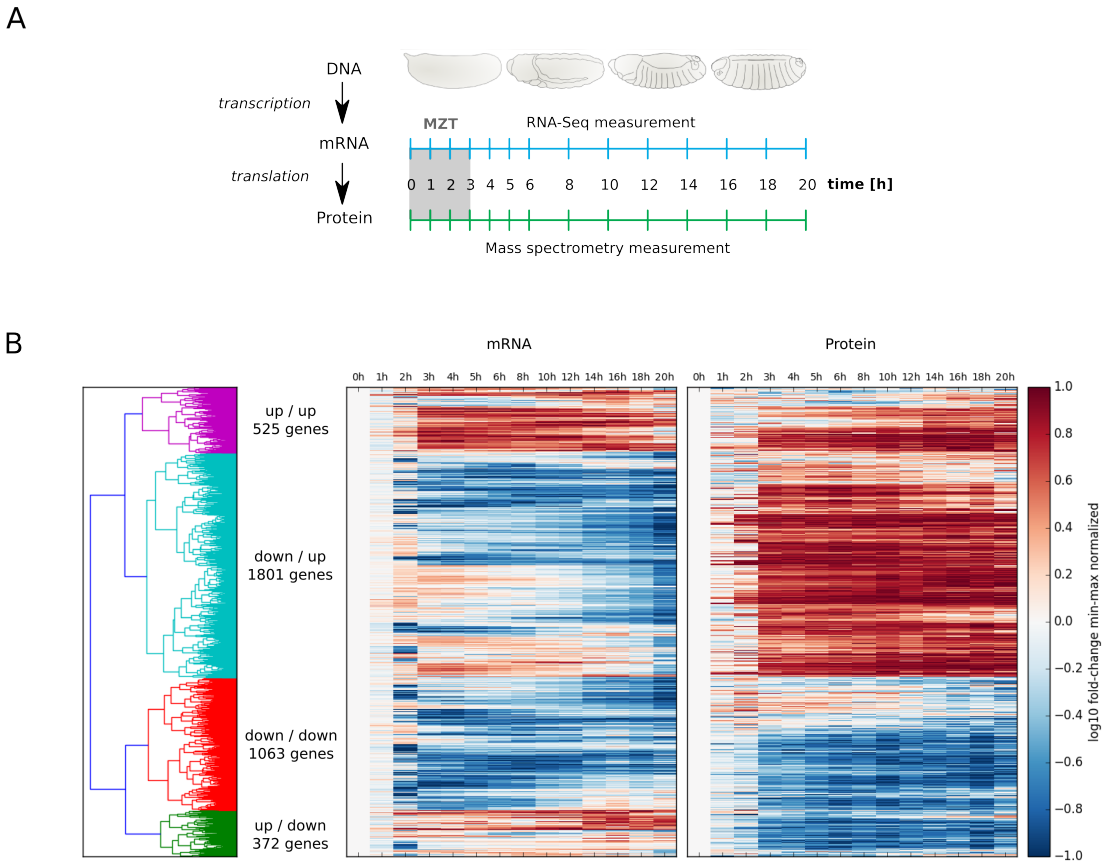


Figure 6.1: Paired transcriptome and proteome measurements during *Drosophila* development. (A) Time points of paired mRNA and protein measurements during *Drosophila* embryonic development using RNAseq and mass spectrometry, respectively. The initial time point (0h) represents egg deposition; the maternal-to-zygotic transition (MZT) occurs in the first 3h of development. **(B)** Heatmaps of mRNA and protein time-courses. The 3761 mRNA/protein pairs (y-axis), for which protein could be detected in at least 10 out of 14 time points, are shown at various times of development (x-axis). The color code indicates the mRNA and protein fold changes relative to t = 0h after min-max normalization between -1 and 1. Time-courses are sorted according to hierarchical clustering using the Euclidean distance as distance metric (see also dendrogram on the left). Time-courses within each of the four clusters (green, red, blue, purple) roughly follow similar dynamics, reflecting concomitant or opposing mRNA and protein dynamics (increase (up) or decrease (down)).

6.2.2 Proteome and transcriptome changes show limited correlation

We related our paired transcriptome and proteome measurements to explore whether both gene expression layers exhibit a high degree of coordination during *Drosophila* embryogenesis. We based our analysis on 3,761 RNA-protein pairs that have been reproducibly (in at least 10 time points) measured during our time-course. For both, mRNA or protein, we calculated the median value across all four replicates for each time point.

We classified the temporal behaviour of these 3,761 mRNA-protein pairs into four groups with qualitatively distinct dynamics using a hierarchical clustering approach (Fig-

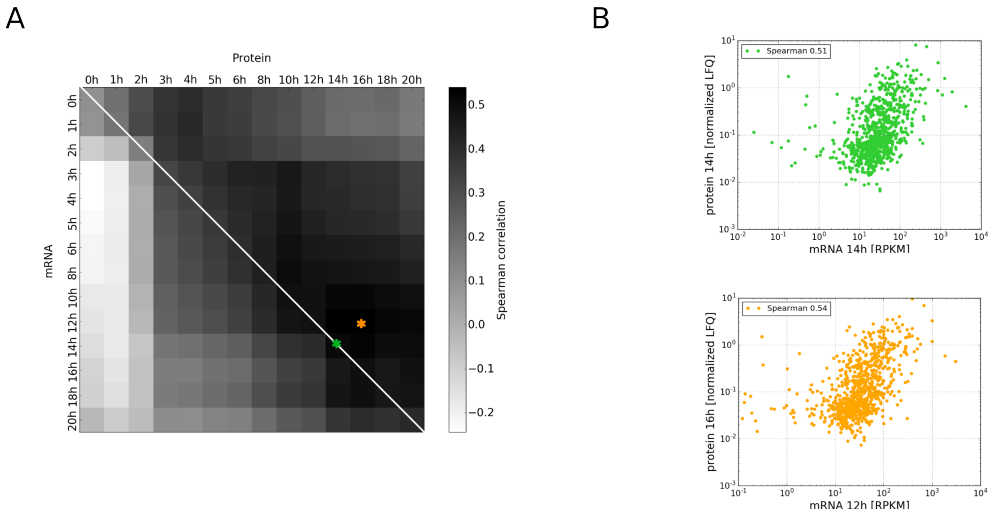


Figure 6.2: Limited correlation between mRNA and protein samples. (A) Global RNA-protein correlation over all genes at fixed time points. Heatmap of Spearman correlation coefficients between protein (x) and mRNA (y) levels at the same (diagonal) or distinct (off-diagonal) time points of embryonic development. Green and orange stars indicate maximum correlation between same and distinct time points, respectively. **(B)** Maximum global RNA-protein correlation. Scatter plots showing correlation of mRNA and protein levels at 14h (top), or between mRNA at 12h and protein at 16h (bottom). Chosen time points correspond to the orange and green stars in A, respectively.

ure 6.1B). These four groups are: (1) mRNA and protein levels increase concordantly (525 genes), (2) mRNA and protein levels decrease concordantly (1063 genes), (3) mRNA increases, while protein levels decrease (372 genes) and (4) mRNA decreases, while protein levels increase (1801 genes). In total, 58% of genes showed inverse abundance changes between mRNA and protein, indicating that the temporal expression dynamics of most mRNA transcripts and the corresponding proteins are not monotonous.

To further investigate the quantitative relationship between mRNA and protein abundance, we calculated the Spearman correlation coefficient globally, relating all mRNA and protein levels separately for each time point. In contrast to the Pearson correlation coefficient, Spearman correlation makes no assumptions about normality of the data or linearity in the dependence of the variables [de Sousa Abreu et al., 2009, Maier et al., 2009]. We included only significantly changing proteins (n = 866) to avoid that random variations around the mean might mask existing trends between mRNA and protein. In line with previous studies, the average correlation between mRNA and protein levels of the same time point is limited ($\rho = 0.36$) (Figure 6.2A, diagonal) with a maximum at 14h ($\rho = 0.51$) (Figure 6.2B - green panel). Of note, this limited correlation is not due to experimental variation, as the mean correlation across four biological replicates at 14h is $\rho = 0.96$ for the transcriptome and $\rho = 0.96$ for the proteome.

To account for the time delay associated with mRNA processing and translation, we also assessed non-synchronous correlations by globally relating mRNA and protein levels at distinct time points. In this scenario, the highest mRNA-protein correlation reaches $\rho = 0.54$ between mRNA levels at 12h and protein abundances at 16h (Figure

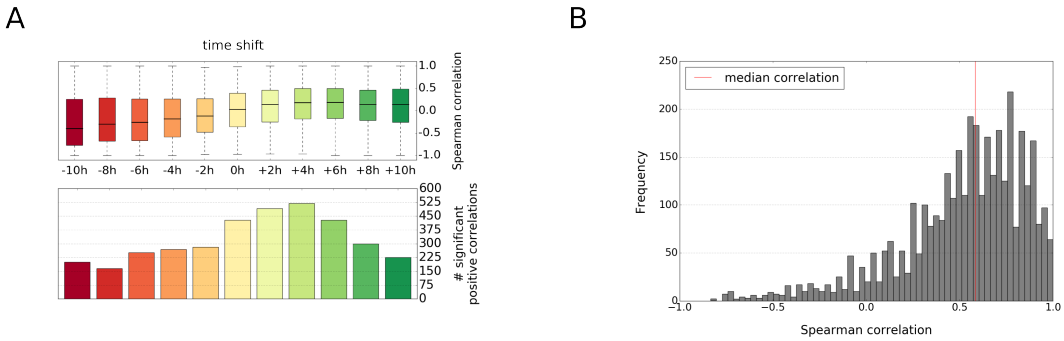


Figure 6.3: Limited correlation between mRNA and protein time-courses. (A) Correlations of individual mRNA-protein pairs over the complete embryonic time-course. **Top panel:** Boxplots indicating the distribution of Spearman correlation coefficients for all 3761 mRNA-protein pairs (white line: median; boxes: quartiles; whiskers: 95-percentile). Time shifts between mRNA and protein dynamics were introduced by adding a constant to the time axis of the protein (0h: no shift). Positive and negative shifts reflect that protein lags behind or is advanced relative to mRNA, respectively. **Bottom panel:** Number of mRNA-protein pairs with a significant ($p < 0.05$), positive correlation for each time shift. **(B)** Distribution of maximum Spearman correlation coefficients across all time shifts. For each individual mRNA-protein pair, the maximum correlation at any time shift between 0h and +10h was considered.

6.2B - orange panel). Despite this modest value, we observed a general trend of better transcriptome/proteome correlation when relating mRNA samples to later protein time points, most likely due to delays in protein synthesis relative to mRNA accumulation. Furthermore, higher correlations are generally observed at later developmental time points, possibly because less pronounced dynamical changes of mRNA and protein occur after MZT.

To further test for the correspondence of mRNA and protein dynamics, we calculated the correlation between the mRNA and protein time-courses for each gene individually. Only a subset of mRNA-protein pairs (429 - 11.4%) exhibits a significant positive Spearman correlation coefficient (Figure 6.3A - bottom panel, $p < 0.05$), and the median correlation coefficient over all genes is close to zero (Figure 6.3A - top panel). To account for delays in protein synthesis, we additionally calculated the same correlation coefficients after introducing a time shift between mRNA and protein measurements. These shifts only marginally improved the median correlation coefficient over all genes, with a maximum median correlation being observed if the protein was assumed to lag behind the mRNA for 4-6h. At a time shift of 4h, we also observe the highest number (520 - 13.8%) of mRNA-protein pairs with significant positive correlation. A similar time delay of 2h to 6h between peaking of circadian mRNA and protein was also reported by Robles et al. (2014) [Robles et al., 2014].

Taken together, these data indicate a poor correlation of mRNA and protein dynamics. Overall, only 33.7% (1268) of genes show a significant ($p < 0.05$) positive monotonic relationship between mRNA and protein at positive time shifts with protein lagging behind mRNA. As each mRNA-protein pair may be characterized by a distinct delay, we further selected the maximal correlation estimate of each mRNA-protein pair over all time

shifts, regardless of significance. This procedure improves the median correlation to $\rho = 0.58$ (Figure 6.3B). Thus, the relation between mRNA and protein abundances requires a more elaborate mathematical framework, incorporating gene-specific parameters. We therefore turned to mathematical modelling to better describe the mechanisms underlying protein production from mRNA.

6.2.3 Kinetic models quantitatively relate mRNA and protein dynamics

We used a set of dynamical models describing protein expression based on ordinary differential equations (ODEs) to further investigate whether the dynamics of each protein can be explained as a function of the corresponding mRNA time-course. By fitting these models to the experimentally measured protein expression time-courses, we assigned each mRNA-protein pair to a model representing one of four regulatory scenarios described below (schematically depicted in Figure 6.4A), or excluded it from any of these.

The default model ('production') follows the simple assumption that protein is synthesized from mRNA and additionally subject to degradation. The mRNA concentration serves as an input and proportionally affects the translation rate, implying the absence of complex post-transcriptional regulation. We also considered a 'delayed production' model, which assumes a temporal delay in translation. In this model, the protein is initially degraded until translation sets in, from which point the 'delay' model corresponds to the 'production' model. The implementation of the 'delay' model was motivated by very low protein accumulation prior to the MZT (Figure 6.1B). Additionally, we assumed two mathematically less complex models: In the degradation model *de novo* translation can be neglected throughout embryonic development and therefore protein synthesis was omitted. In the 'stationary' model, the protein is assumed to be stable during embryogenesis, thereby eliminating protein production or degradation altogether.

The unknown model parameters such as protein synthesis and degradation rates, initial protein concentrations, and delay times had to be determined by fitting our models to the experimental data (see Chapter 2.2.1). We fitted each of the 3,761 mRNA-protein pairs individually using the four model variants, and classified the genes into one of the regulatory scenarios ('stationary', 'degradation', 'production' and 'delayed-production') using a two-step strategy: In the first classification step we assessed whether a given model is capable of describing the measured mRNA-protein dynamics based on the difference between model fit and data using a χ^2-test ($\alpha = 0.05$). We also tested for systematic deviations between model and data using a Durbin-Watson test ($\alpha = 0.05$), by which we exclude that the differences are correlated in time. Only if both tests were passed, a model variant was considered to be feasible. The second classification step selects for the most appropriate model in case multiple models for a mRNA/protein pair could not be rejected in the first step. To this end, we performed a stepwise likelihood ratio test, balancing the goodness-of-fit against the risk of over-fitting to select for the simplest model explaining the measured data. If the model selection assigned a protein to the 'degradation' or 'stationary' classes, but the mathematically more complex 'production' model could not be rejected with a non-zero protein translation rate (see Supplementary

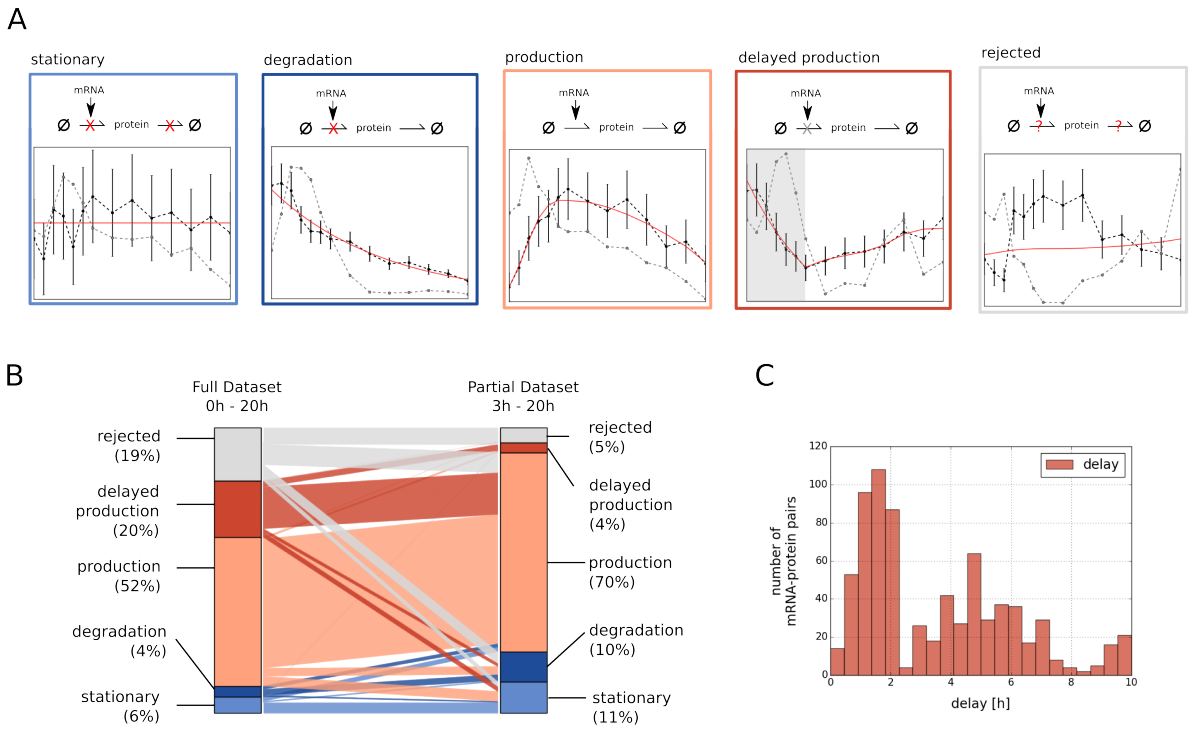


Figure 6.4: Kinetic models quantitatively relate mRNA and protein dynamics. (A) Schematic representation of model variants incorporating protein synthesis and degradation (thin arrows). Red and gray x indicate absence or delayed onset of a reaction, respectively. Measured mRNA time-courses were used as an input to the model, and simulated protein output was fitted to the corresponding experimental data by tuning the kinetic parameters. An exemplary model fit of a protein belonging to each class is shown (red line), alongside with corresponding mRNA (gray) and protein (black) data. If all four models shown on top are rejected, the protein is classified as potentially post-transcriptionally regulated, with unknown function of synthesis and/or degradation. **(B)** Model-based classification results for 3761 mRNA-protein pairs. Barplot showing fractions of proteins in each class for the full dataset (left; 0-20h) or post-MZT data only (right; 3-20h). Connecting lines indicate the migration from one model class to another between both scenarios. **(C)** Distribution of estimated delay times of 743 proteins with delayed translation after egg deposition.

Text 9.5.4), we re-assigned the protein to the 'production' class. Thus, we consider the 'production' model as the simplest, unregulated protein expression scenario in biological terms, whereas the competing (mathematically simpler) models require the existence of an additional biological factor blocking translation or degradation. Exemplary classifications to corresponding models are shown in Figure 6.4A.

Slightly more than half (52%) of mRNA-protein expression patterns fit the 'production' model, which assumes continuous protein synthesis from mRNA and are thus likely determined by pure transcriptional control (Figure 6.4B). Furthermore, 6% of proteins are stable during embryogenesis and can be explained by the 'stationary' model. Another 4% of proteins show decaying expression levels and are fitted best using the 'degradation' model. Finally, the extended assumption of a delay time for translation ('delayed-production' model) explained another 20% of genes. Altogether, 81% of genes

were explained by any of these four simple regulatory models, suggesting that proteome and transcriptome dynamics can be quantitatively described using simple ODE models for gene expression.

For the remaining 19% of protein-mRNA pairs all four proposed models needed to be rejected. We believe this fraction of proteins to be under complex post-transcriptional control, and accordingly find a statistically significant enrichment (1.68 fold enrichment, $p < 0.01$, Figure 9.23A) when overlaying this list with a published set of translationally regulated genes identified by ribosome profiling during *Drosophila* development [Qin et al., 2007]. These potentially post-transcriptionally regulated proteins are found in all four dynamical groups obtained by hierarchical clustering (Figure 9.23B). This demonstrates the complementarity of our mathematical modelling analysis and suggests that post-transcriptional regulation occurs via a variety of mechanisms rather than being under control by a global factor. Interestingly, two proteomic studies on mammalian circadian rhythmicity [Robles et al., 2014] and on yeast amino acid starvation [Ingolia et al., 2009] report very similar fractions of 20-30% of potentially post-transcriptionally regulated proteins, suggesting that the frequency of this type of control may be conserved across organisms.

During the first 2-3h until MZT, no transcription occurs in the fly embryo [Langley et al., 2014, Tadros and Lipshitz, 2009]. Thus, any protein must be maternally deposited or translated from maternally deposited RNA. MZT is also visible in our data as striking changes in mRNA as well as protein expression patterns at 3h (Figure 6.1B). Since MZT marks the advent of transcriptional activity, we expected a large fraction of post-transcriptional regulation specifically within the first 3h of development. To test this hypothesis, we repeated the protein classification including only data from time points beyond MZT (3h 20h). In line with a strong decline of post-transcriptional control after MZT, we found that 95% of proteins can be explained by one of the four above-mentioned ODE model variants if only MZT data is considered (Figure 6.4B, right).

Among these 95%, only 4% of proteins (initially 20%) are still assigned to the 'delayed-production' model, i.e. only few proteins show delayed translation past MZT. This observation again supports that MZT is a major point of post-transcriptional control at which mRNA translation may be activated 'on demand' [Beyer et al., 2004]. In line with this hypothesis, estimated delay times in the delayed-production model fitted to the full dataset (0-20h) show a bimodal distribution with a large group of proteins exhibiting a delay time of 0.5 - 2.5h (344 proteins) and another group of proteins with translation delayed even longer for about 3 - 10h (353 proteins) (Figure 6.4C). This suggests two waves of translation: a first translation burst coinciding with MZT and a secondary delayed translation phase. The general ability of our classification method to distinguish between different translational models is shown in Figure 9.24.

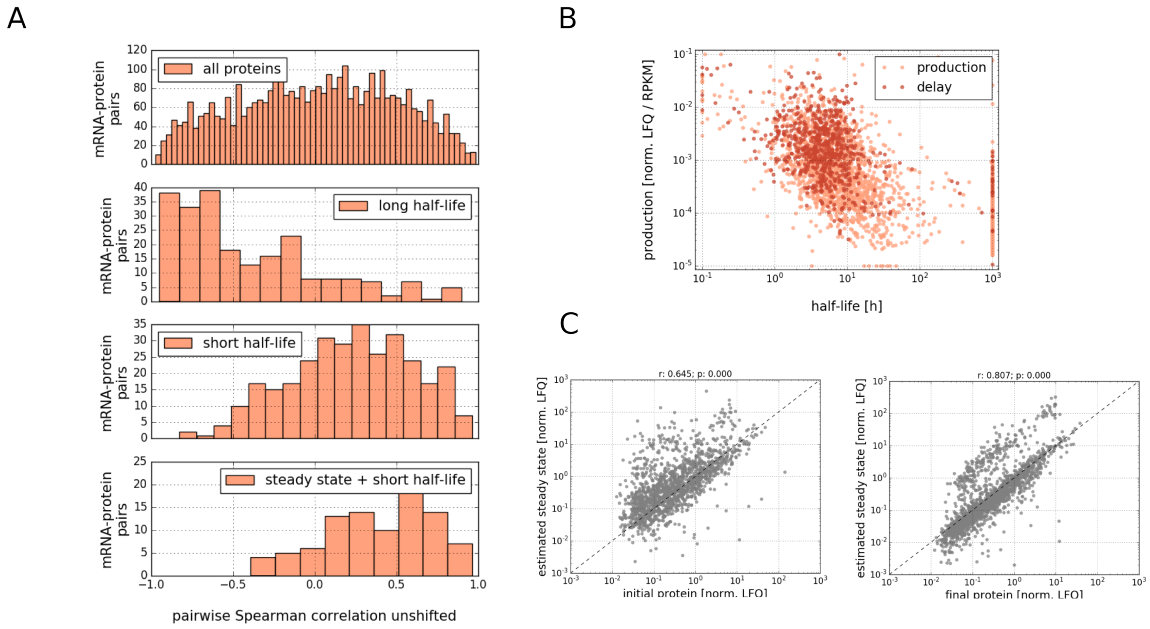


Figure 6.5: Lack of mRNA-protein correlation partially explained by long protein half-lives and unbalanced production and degradation. (A) mRNA-protein correlation depends on dynamics protein dynamics and initial level. Distribution of Spearman mRNA-protein correlations for all 3761 proteins (top panel), proteins with long half-life (second panel), proteins with short half-life (third panel) and proteins with short half-life, which are also close to their estimated steady state at the onset of embryogenesis (bottom panel). B Inverse relation of protein production and turnover. Protein half-lifes and production rates are related across all mRNA-protein pairs for the production and delay models (1960 and 743 proteins, respectively). Protein half-lives were calculated as ln(2) / λ, with λ denoting the degradation rate, and are limited between 10e-03 and 10e03 due to finite parameter ranges during model fitting. **(C)** Protein levels early in development tend to deviate from the model-predicted protein steady-state. Measured protein levels at 0h (left) and 20h (right) of proteins falling into the production class are plotted against protein steady states, estimated as $\frac{\alpha}{\lambda}\ mRNA_t$, from the fitted model parameters (α: production rate; λ degradation rate) and from measured mRNA with t={0h, 20h}.

6.2.4 ODE models account for lack of mRNA-protein correlation

Given that 81% of mRNA-protein pairs are described without assuming complex post-transcriptional regulation, the apparent lack of direct mRNA-protein correlation (Figure 6.2 and Figure 6.3) might be surprising. The proposed mathematical models provide insight into the observed missing correlation between mRNA and protein. For the 'stationary', 'degradation' and 'delay' classes, the model predicts simple biological control mechanisms on protein translation and/or degradation, which may lead to a lack of mRNA-protein correlation.

However, even for the 'production' model, where no post-transcriptional control mechanisms are assumed, a strong correlation of the mRNA and protein time-course of each gene is no necessary consequence, and the corresponding correlation coefficients of proteins in this group show a broad distribution (Figure 6.5A, top panel). One explanation

for the lack of correlation may be that only proteins with short half-life (fast degradation) closely follow the dynamics of the corresponding mRNA time-course [Hargrove and Schmidt, 1989], hereby increasing the mRNA-protein correlation [Wu et al., 2008]. Based on our fitting results, we obtained estimates for the kinetic parameters controlling translation and degradation of each protein, and can use these to test this hypothesis. As expected, the lack of positive mRNA-protein correlation ($\rho = 0.03$) is particularly evident when considering only stable proteins whose half-life confidence interval exceeds the median (8.4h) of all protein half-lives (237 proteins, Figure 6.5A, second panel). In contrast, positive correlations are much more common in the inversely defined group of proteins with shorter half-lives (313 proteins, Figure 6.5A, third panel), with an average correlation coefficient of $\rho = 0.24$, supporting that protein degradation rates determine how closely protein dynamics follow the mRNA time-course. Interestingly, estimated protein translation rates and half-lifes are negatively correlated (Figure 6.5B; ρ = -0.58, $p < 0.01$). An according observation has already been reported in early *Xenopus* development [Peshkin et al., 2015] suggesting that, on a global level, proteins are tuned to exhibit similar expression levels (i.e., similar synthesis-degradation ratios) irrespective of their dynamical behaviour (protein half-life). This hints to a global "make-fast break-fast" mode of gene regulation, in which highly dynamic proteins with a strong mRNA-protein correlation are also synthesized at a higher level to achieve considerable expression levels.

In our study, the dynamical behaviour of a protein is not fully predictive for the mRNA-protein correlation, as even proteins with short half-lives can show a weak mRNA-protein correlation (Figure 6.5A, third panel, 'fast dynamics'). Our model explains many of these mRNA-protein discrepancies with an out-of-steady-state protein concentration in the fertilized egg (t = 0h): It is known for example that certain mRNA deposited in the egg remain untranslated in the inactive egg and are only translated upon fertilization [Lasko, 2012, Smibert et al., 1996]. In this case, initial protein production at t=0 exceeds protein degradation, leading to a net increase in protein level even if mRNA levels remain constant. In an extreme case, this scenario can result in an inverse protein-mRNA relationship during the developmental time-course, where the protein level increases, while mRNA abundance decreases over time (Figure 9.25A). Likewise, the opposite behaviour, mRNA increase with a concomitant protein decay, can also occur if net protein degradation exceeds synthesis (Figure 9.25B).

To test our prediction that a large fraction of proteins is out of steady-state at fertilization, we compared the actual protein expression values with the theoretical steady-state predicted by the fitted synthesis and degradation rates (Figure 6.5D). We found that correlation of measured protein levels with the theoretical steady state is limited at fertilization, whereas a better agreement is observed later in development. Furthermore, we can show that such a perturbation from steady state weakens the mRNA-protein correlation during the developmental time-course: If we consider only proteins with fast dynamics (i.e., short half-life), whose abundance is also close to their estimated steady state at 0h (95 proteins), the median correlation for individual mRNA-protein pairs significantly increases to $\rho = 0.45$ (Figure 6.5A, bottom panel) compared to $\rho = 0.24$ for all

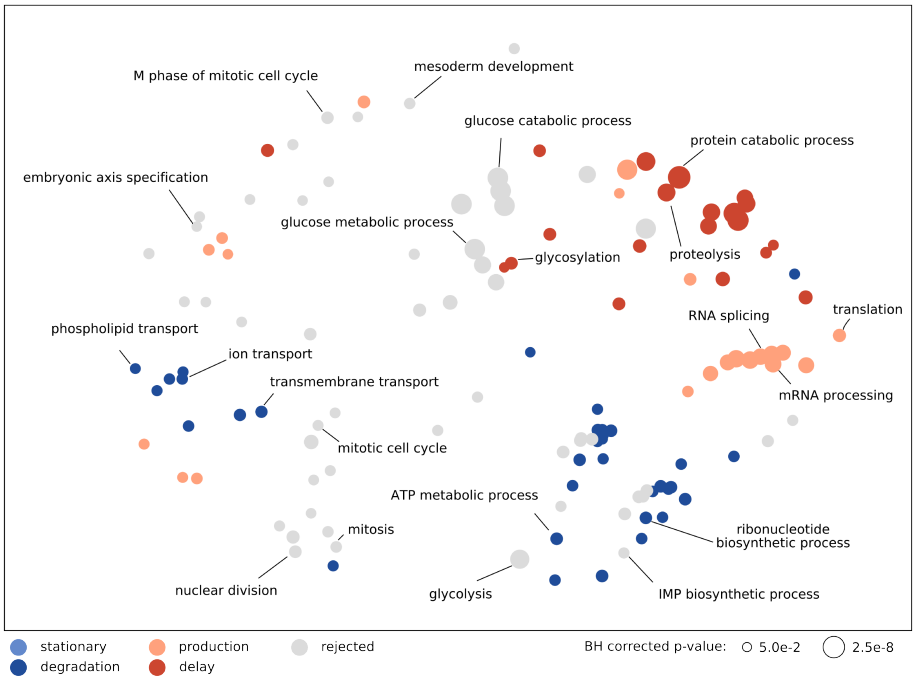


Figure 6.6: Classes of protein expression regulation reflect biological function. Distinct GO terms are over-represented in each model category. The five protein classes obtained from combined model fitting and model selection on the full dataset (0-20h) are indicated by colors (see legend). GO terms significantly enriched in one of the five protein groups (BH corrected p-value < 0.05) are arranged according to their semantic similarity and a 2D projection was generated via multidimensional scaling, each circle representing a GO term. The protein classes according to our model fits to the full dataset (0-20h) are indicated by colors (see legend). The size of each circle is proportional to the p-value for enrichment.

fast changing proteins, (two-sided Kolmogorov-Smirnov test: $p = 0.0054$). In a dynamical context, a prerequisite for a high correlation between mRNA and protein is therefore not only fast protein turnover, but also an appropriate initial mRNA-protein ratio with balanced net production and degradation rates.

In summary, the central dogma of mRNA being translated into proteins does not necessarily imply that expression levels of the two species correlate. Even in simple ODE models of protein expression, long protein half-lives as well as differences between actual protein abundance and the steady-state level result in a lack of mRNA-protein correlation.

6.2.5 Classes of protein expression regulation reflect biological function

Given that we are able to group proteins into four classes reflecting their respective modes of expression regulation, the question arises, whether proteins in a given group serve common biological functions, not shared with the other classes. To answer this question we performed Gene Ontology (GO) analyses and found enrichment of several GO terms for each model class (Figure 6.6).

For instance, proteins described by the production model showed enrichment for pro-

cesses related to mRNA processing, splicing and RNA metabolism, suggesting that this class contains critical regulators of post-transcriptional gene expression. Moreover, proteins in the delay group are enriched for GO terms related to protein catabolic processes, possibly indicating widespread expression of factors mediating the degradation of maternal protein at the onset of MZT [De Renzis et al., 2007]. We further noted an enrichment of cell cycle related genes among the proteins for which all of our four models needed to be rejected, i.e. proteins which are putatively under post-transcriptional control. In fact, the early (synchronous) nuclear divisions in *Drosophila* embryos cannot be controlled transcriptionally, since transcription is virtually absent before MZT. Accordingly, widespread post-transcriptional regulation has been reported for maternal mRNAs involved in cell cycle regulation [Groisman et al., 2002, Mendez and Richter, 2001]. In addition, GO-terms related to 'sugar metabolism' are also enriched in the group of potentially post-transcriptionally regulated proteins. This agrees with previous evidence showing post-transcriptional control of genes functioning in glucose metabolism [Robles et al., 2014].

The mode and time scale of post-transcriptional regulation, however, appears to be distinct for proteins involved in cell cycle or sugar metabolism: When we repeated our GO enrichment analysis for proteins with post-transcriptional regulation occurring only after MZT (rejected class in 3-20h dataset), most cell cycle proteins remain under post-transcriptional control, whereas this is not true for the group of sugar metabolic proteins. In line with early post-transcriptional regulation before or at MZT, protein levels of genes related to sugar metabolism predominantly up-regulate 3h after fertilization and afterwards remain nearly stable, while their corresponding mRNAs sharply decrease already within 2h post fertilization (Figure 9.26A). In contrast, a subset of proteins related to the cell cycle shows an abrupt down-regulation of both mRNA as well as protein at 14h, potentially indicating (post-) transcriptional regulation long after the completion of MZT (Figure 9.26B).

In summary, differential enrichment analysis reveals how proteins associated with different biological functions show distinct modes of translation.

6.2.6 Post-transcriptionally regulated transcripts are enriched for RBP binding motifs

To uncover potential regulators of post-transcriptional gene regulation in the early fly embryo, we searched for enriched sequence motifs in the mRNA sequence of proteins in our protein classes. To this end, we considered 67 *Drosophila* specific RNA Binding Protein (RBP) motifs corresponding to different 51 RBPs identified by Ray et al. (2013) [Ray et al., 2013] and tested for enrichment using the Analysis of Motif Enrichment tool provided by the Meme-Suite [McLeay and Bailey, 2010a]. Significant motif enrichment was strongest for the full length mRNA and weaker when analysing 3'- or 5'-UTRs or the coding sequence, indicating that motifs are predominantly located within the intronic sequences of transcripts, in line with a previous findings for Hrb98DE [Blanchette et al., 2009]. Moreover, when considering full-length RNAs, we found no or little enrichment

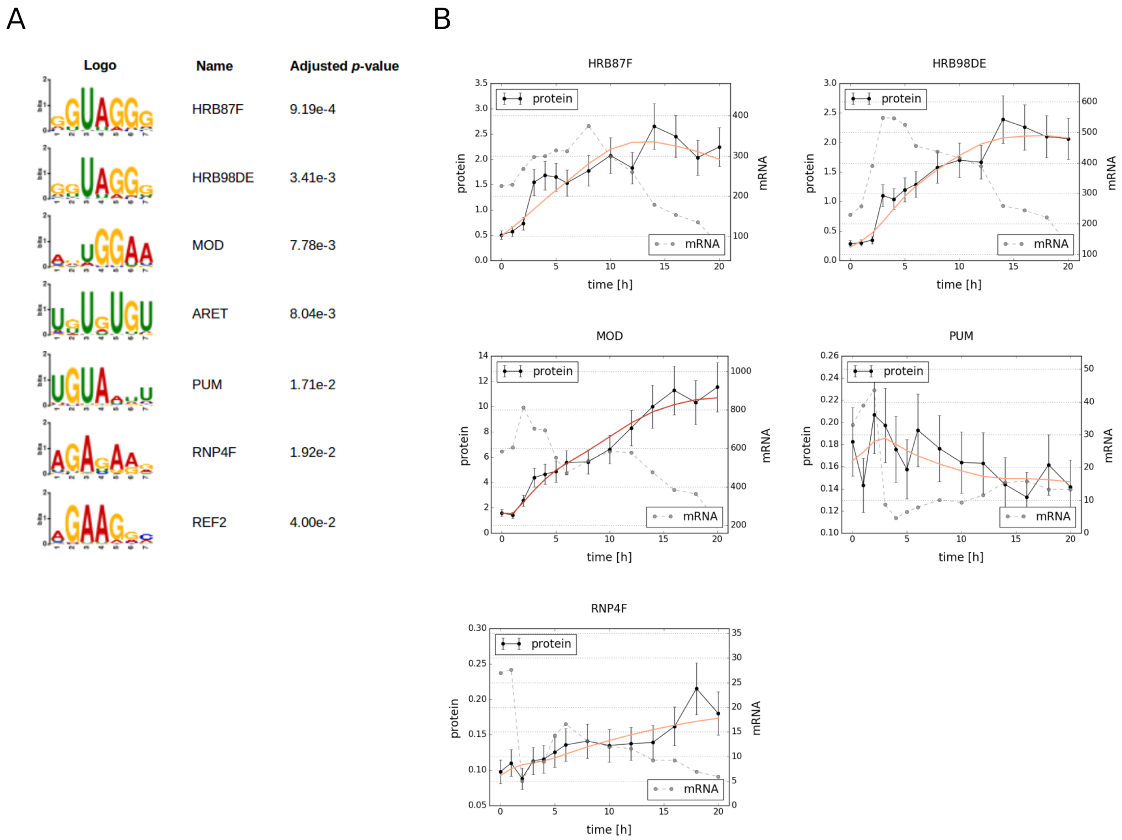


Figure 6.7: Post-transcriptionally regulated proteins are enriched for RBP binding motifs. (A) Seven RBP sequence motifs are enriched in the group of potentially post-transcriptionally regulated proteins. Only motifs with an adjusted enrichment p-value < 0.05 in the longest full length transcript reported. If two motifs for the same RBP were enriched, only the one with stronger enrichment is shown. In total we scanned for enrichment of 67 *Drosophila* specific motifs corresponding to 51 distinct RBPs identified in Ray et al. (2013). **(B)** Putative post-transcriptional regulators are regulated at the protein expression level during *Drosophila* development. Protein time-courses of RBPs identified (black), alongside with best model fit (coloured) and corresponding mRNA dynamics (gray dashed). The expression of remaining RBPs ARET and REF2 is below the LFQ detection threshold at all developmental time points.

of RBP binding motifs in 'stationary', 'degradation' and 'delay' classes, whereas 7 RBPs showed significant motif enrichment in the group of potentially post-transcriptionally regulated proteins (Figure 6.7A). This further supports that complex post-transcriptional regulation may occur in those proteins whose mRNA-protein relationship cannot be explained by any of the simple gene expression models.

Among the enriched motifs in this class were the RBPs Pumilio and Bruno (also known as Aret), both of which are known to be post-transcriptional regulators of positional information genes in early *Drosophila* development [Murata and Wharton, 1995, Sonenberg and Hinnebusch, 2009]. The strongest motif enrichment was observed for Hrb87F and Hrb98DE (also known as Hrp36 and Hrp38), two proteins recognizing highly similar RNA sequence motifs. During our time-course, Hrb87F and Hrb98DE both show a 5-fold up-regulation in protein expression (Figure 6.7B), suggesting that their activity may be developmentally regulated in the embryo. Accordingly, a recent RNA

interactome study showed that the mRNA bound fraction of Hrb87F and Hrb98DE changes during MZT in *Drosophila* [Sysoev et al., 2016]. Interestingly, the known functions of Hrb87F and Hrb98DE match with results of our GO term analysis, as these proteins have been identified in large screens for regulators of the cell cycle and sugar metabolism, respectively [Ducat et al., 2008, Ugrankar et al., 2015]. Also an enrichment of Hrb98DE and Hrb87F motifs in proteins related to sugar metabolism compared to all post-transcriptionally regulated proteins can be observed (Hrb98DE: 51.7% vs. 37.0%, hypergeometric test: $p = 0.033$; Hrb87F: 44.8% vs. 19.6%, hypergeometric test: $p = 0.006$).

6.2.7 Hrb98DE may post-transcriptionally regulate glucose metabolism

For further analysis, we focused on Hrb98DE, because this protein (and its vertebrate homologue hnRNPA1) has been implicated in the regulation of protein translation as well as splicing [Blanchette et al., 2009, Brooks et al., 2015, Despic et al., 2017, Ji and Tulin, 2016]. To assess the role of Hrb98DE in post-transcriptional regulation, we made use of *Drosophila* S2R+ cell culture system and performed paired RNA-Seq and quantitative proteomics experiments under Hrb98DE knock-down conditions. We expected the Hrb98DE knock-down to differentially perturb mRNA and protein pools, especially for genes involved in sugar metabolism, with additional effects on pre-mRNA splicing.

The efficiency of Hrb98DE knock-down was confirmed at the mRNA and protein level 5 days after treatment of SR2+ cells with siRNA targeting Hrb98DE. Based on RNA-Seq and mass spectrometry data, we observe a strong depletion of Hrb98DE mRNA and protein of about 84% and 78%, respectively, when compared to S2R+ cells treated with a dsRNA targeting LacZ (Figure 9.27). On a genome-wide level, we detected an overlapping set of 4156 genes at both mRNA and protein levels in control cells. Out of these, 233 genes are differentially expressed at the mRNA level in response to Hrb98DE knock-down (Benjamini-Hochberg corrected p-value < 0.05, absolute fold-change $> 30\%$), whereas 41 change significantly at the protein level (p-value < 0.01, absolute fold-change $> 50\%$) (Figure 6.8A and B).

Interestingly, a large fraction of genes responding at the protein level (26 out of 41, i.e. 63%) do not show a concomitant change at the mRNA level (Figure 6.8C). Accordingly, it was shown in a recent study that Hrb98DE binding to mRNA leads to a change in the transcript levels of only a small fraction of genes bound by Hrb98DE [Ji and Tulin, 2016]. Thus, we hypothesize that Hrb98DE may bind to these 26 RNAs to affect protein translation, but not mRNA turnover. In line with Hrb98DE controlling sugar metabolism at the post-transcriptional level, three out of the 26 candidate genes (14-3-3ζ, dorsal, domino) were previously shown to have an effect on glucose metabolism [Ugrankar et al., 2015].

Investigating Hrb98DE splicing effects, we determined differentially spliced genes in the Hrb98DE knock-down transcriptome data. To this end, we focused on 3549 genes consistently detected at the mRNA level in the S2R+ cell lines and our embryonic time-

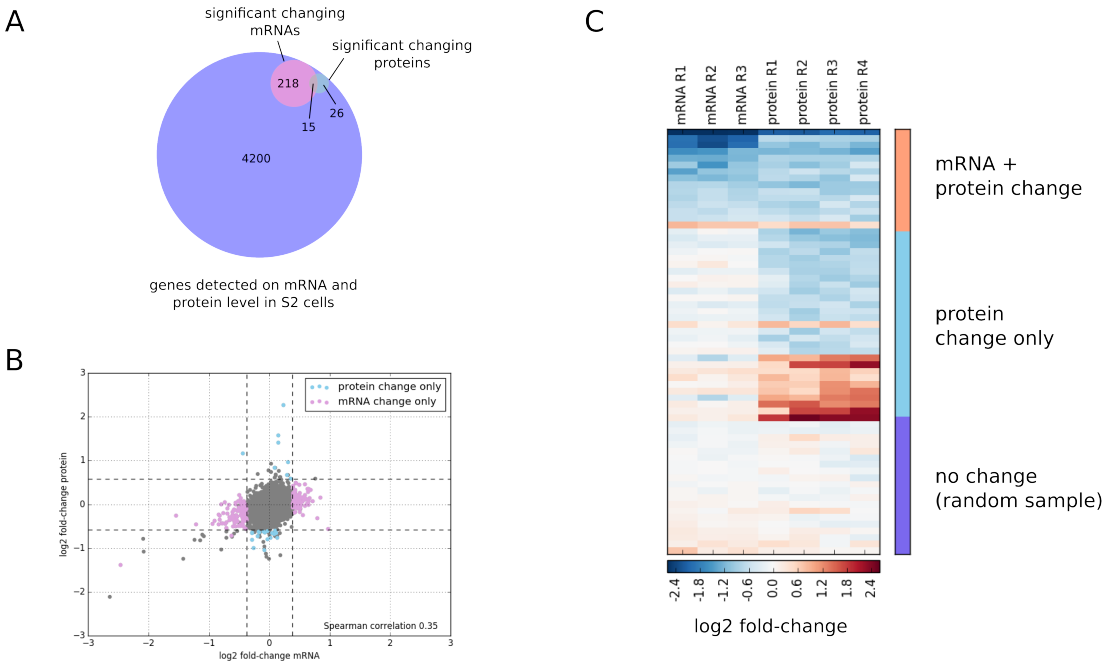


Figure 6.8: Hrb98DE post-transcriptionally regulates glucose metabolism. Venn diagram **(A)** and scatter plot **(B)** showing significantly changing mRNAs (pink) and proteins (light blue) 5 days after Hrb98DE knock-down in S2R+-cells. Dashed lines in B indicate thresholds of fold-changes used as a first criterion to identify significantly changing mRNA (log2 fold-change ¿ 30%) or protein (log2 fold-change $>$ 50%). Significantly changing mRNAs additionally required a BH corrected p-value of < 0.05 and proteins a non-corrected p-value of < 0.01. **(C)** Fold-changes for individual mRNA (3) and protein (4) replicates are shown for the groups of genes showing differential Hrb98DE knock-down responses (mRNA and protein significantly changing, significant protein change only, no change of mRNA or protein). For genes with no significant change in either mRNA or protein only a subset of 20 randomly chosen genes out of 4200 genes is shown.

course. Out of these we found 288 differentially spliced genes upon Hrb98DE knock-down (Figure 6.9A), which substantially overlap with genes identified as direct Hrb98DE targets either by motif analysis or RIPseq [Blanchette et al., 2009,Ji and Tulin, 2016]. Specifically, 57% of the differentially spliced genes were identified as a potential direct Hrb98DE target (1.96-fold increase observed vs expected, p-value 7.4e-25) (Figure 6.9B). In line with our hypothesis, we found that the differentially spliced genes are significantly enriched for the GO-term 'regulation of glucose metabolism' (18 genes, 3.36-fold enrichment observed vs expected, p-value 6.4e-06), out of which 11 genes have also previously been identified as Hrb98DE targets. This suggests that, in our knock-down setup, the majority of Hrb98DE-dependent effects on glucose metabolism are visible at the level of pre-mRNA splicing. For example, we show that alternative splicing of domino, one of the genes involved in glucose metabolism, leads to an isoform switch at the protein level (Figure 9.27B).

In summary, the identification of post-transcriptionally regulated genes together with the subsequent bioinformatic analysis suggested post-transcriptional regulation of glucose metabolism by the RBP Hrb98DE, which we were able to confirm experimentally.

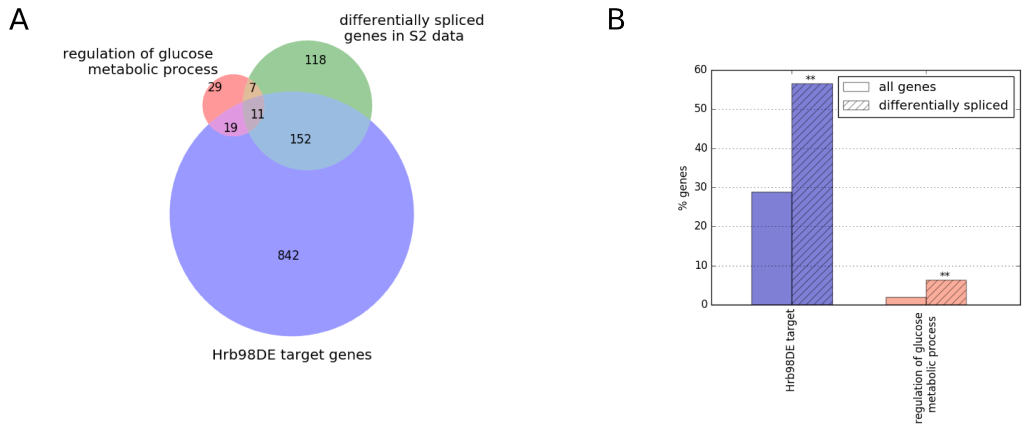


Figure 6.9: Hrb98DE causes splicing changes in genes regulating glucose metabolism. (A) Venn diagram showing differentially spliced genes upon Hrb98DE knock-down and their overlap with genes previously identified as Hrb98DE targets or genes annotated as regulators of glucose metabolic processes (GO:0010906). Hrb98DE targets were defined by the union of Hrb98DE targets defined by Tulin et al (2016), Blanchette et al (2012) or having the Hrb98DE binding motif. **(B)** Percentage of Hrb98DE targets (blue) or genes annotated as regulators of glucose metabolic processes (red) in the set of background genes (clear) vs. the set of differentially spliced genes upon Hrb98DE knock-down (striped).

6.2.8 Discussion

Early *Drosophila* development is initiated in the absence of *de novo* transcription, suggesting extensive gene expression regulation at the post-transcriptional level. In this study, we set out to quantify how mRNA and protein levels are dynamically coordinated in this system. To this end, we combined our *Drosophila* embryogenesis proteome of 15 time points within 20h with its paired transcriptome. The data shows high biological and technical reproducibility among the four quadruplicates for proteome and transcriptome. This creates an extensive systems-level dataset to study translational and post-translational gene regulation in a highly resolved temporal fashion.

There is a recent interest in explaining the moderate mRNA-protein correlation ($\rho \sim$ 0.4 - 0.6) observed in most systems-level studies. It is appealing to attribute the lack of correlation between mRNA and protein expression to a widespread post-transcriptional regulatory network [Mittal et al., 2009]. However, correlation measures, although practical, do not reflect the complex temporal connection between mRNA and protein dynamics. This is especially true if the correlation analysis is limited to the global mRNA-protein relationship at a single time point, as in most currently available studies. Using mathematical models, we could show that a low correlation of mRNA and protein time-courses does not necessarily imply post-transcriptional regulation, but could be explained based on protein turnover rates or deviations from steady state at the onset of development. In our opinion, the analysis of time-resolved mRNA and protein expression data using mechanistic mathematical models of translation is superior when compared to simple correlation analysis.

Only few studies have quantitatively modelled time-course data to investigate the

transcriptome/proteome relationship. Investigating the response of dendritic cells to LPS stimulation, Jovanovic et al. (2015) can explain 79% of the protein variance based on a simple model of protein translation [Jovanovic et al., 2015]. They further concluded that post-transcriptional regulation is relevant predominantly for the regulation of absolute protein levels rather than relative fold-changes upon stimulation. The other study investigated early *Xenopus* embryonic development and - similar to our study - classified proteins into four simple modes of translation using a model fitting and model selection approach [Peshkin et al., 2015]. The authors found that only few proteins are not classified into one of the four groups, and concluded that post-transcriptional regulation is not relevant for early embryonic development.

In this study, we combined model fitting and model selection with a quantitative model rejection approach to explicitly name mRNA-protein pairs insufficiently described by the four considered model variants (production, degradation, stationary and delay). We thus derive and analyse lists of potentially post-transcriptionally regulated proteins, which represent roughly 20% of genes during *Drosophila* embryonic development. Most of this post-transcriptional control occurs within the first few hours after fertilization, i.e. before or during MZT. These 20% of post-transcriptionally regulated proteins are unlikely under control of one global regulatory factor, but rather several regulatory mechanisms (e.g. miRNA and RNA-binding proteins) exert control on multiple levels. More sophisticated models may explicitly take into account time-courses of relevant RNA-binding proteins to model dynamical changes in protein translation and impact of these RBPs on the RNA-protein relationship. In the future, with further advances in technology, more variables can be included in dynamical models, such as protein modification and spatial heterogeneity.

According to our models, we evidenced wide-spread delay of protein production during MZT. Translation-on-demand mechanisms have been postulated in yeast [Beyer et al., 2004,Brockmann et al., 2007], and neurons [Liu-Yesucevitz et al., 2011,Wang et al., 2010] but to our knowledge not yet for metazoan embryonic development. The idea of proteins being translationally silent during the first hours of development, although mRNA is present, implies that the appearance of the proteins needs to be timely regulated. From our analysis, proteins delayed in translation are enriched for GO terms related to protein catabolic processes, suggesting widespread degradation of maternal proteins at the onset of MZT [De Renzis et al., 2007]. Based on sequence motif analyses, we hypothesized that the RNA-binding protein Hrb98DE is a post-transcriptional regulator of genes involved in sugar metabolism during early embryogenesis. We were able to validate this experimentally by performing a Hrb98DE knock-down in S2R+ cells, advancing from computational prediction to *in vivo* model verification. In line with our findings, sugar metabolism has been identified as a target for post-transcriptional regulation, with possible implications in the pathogenesis of diabetes [Kim and Lee, 2012]. During development, post-transcriptional control of sugar metabolism could be attributed to the fact, that the energy demand of the embryo changes dramatically just before global transcription is activated during the maternal-zygotic transition. The identification of Hrb98DE targets using our knock-down approach was potentially limited by the lack of knock-down ef-

fect propagation to target protein levels due to stability of target proteins. Accordingly, we find moderate effects of Hrb98DE knock-down on the proteome, but identify more extensive changes at the mRNA level, in particular differential splicing.

Overall, we generated a large-scale paired RNA-protein dataset that we used for a complete systems biology analysis cycle including validation of our proposed hypothesis by *in vivo* experiments. Hopefully our data will serve as a valuable resource for future studies to come.

Chapter 7

Discussion

7.1 Summary of applied systems biology approaches

In this thesis we applied various methods borrowed from the field of systems biology to investigate biological problems of increasing complexity, placed within the context of differential gene expression. In Chapter 2 we classified individual genes as circadian based on their mRNA time-course expression pattern in synchronized NIH 3T3 cells. Exploiting the same dataset, but focusing on expression measurements of two genes instead of one, we further predicted potential regulator-target interactions between circadian genes (Chapter 3). As part of our investigation of gene expression dynamics governing TGFβ-mediated EMT, we attempted to reconstruct the complete structure of a regulatory network of many genes, based on time-course and knock-down gene expression data (Chapter 4). As an additional level of complexity, in Chapter 5 we formulated a kinetic model of the genetic program relevant for EMT, allowing for more detailed model analysis and prediction of model dynamics. Finally, investigating dynamics of protein translation instead of merely focusing on mRNA expression, in Chapter 6 we integrated genome-wide data on the level of mRNA and protein and utilized simple models of protein translation to identify post-transcriptionally regulated genes.

7.2 Employing model rejection analysis for the classification of gene expression time-courses

Despite the variety of biological topics addressed throughout this thesis, a reoccurring theme is the use of model rejection and model selection analysis to classify genes based on their measured mRNA and/or protein time-courses. Model rejection analysis seeks to answer the question whether a given model is sufficiently able to explain a given set of measured data. In this context, a model can also be viewed as a quantitative hypothesis capturing the mechanistic properties of a system, and model rejection analysis provides the tools to rigorously test this hypothesis. Typically model rejection in systems biology focuses on one particular biological system, testing if the complete set of experimental measurements can be sufficiently explained by one or multiple models. In this study however, we repeatedly applied model rejection analysis to gene expression data measured

for many thousands of genes, and successively tested for which of the measured gene expression time-courses the formulated model may hold. As a result, we were able to classify gene expression measurements into various groups, based on the ability of a proposed model to sufficiently describe the underlying data.

In Chapter 2 for example, we identified circadian expressed genes by testing whether measured expression data was well explained by a simple model of periodic gene expression. The approach was further enhanced by a non-parametric bootstrap strategy, in which the identifiability of the estimated model parameters, in particular period length, was assessed with respect to experimental variability. As demonstrated, combined model fitting and identifiability analysis displayed satisfiable performance, both on benchmark as well as real biological data. However, a number of further improvements may be applied in the future. In particular for time-course data, bootstrapping approaches are inherently difficult, due to of the potential loss of auto-correlation properties of the time-course while re-sampling data points. As an alternative strategy, the use of parametric bootstrap strategies to identify periodic expressed genes should be tested.

The general concept behind a parametric bootstrap strategy, is the assumption that the true data generating process is well represented by the initial model in combination with the estimated model parameters. Following this argumentation, the only deviation between modelled and measured gene expression is contributed by experimental noise. Accordingly, by re-sampling gene expression values from the parametrized model, theoretical experimental outcomes can be generated. The mechanistic model is subsequently fit to these theoretical experimental outcomes, and the empirical distribution of various test-statistics, such as the size and correlation of the residuals, can be calculated. From its empirical distribution, the likelihood of observing a test-statistic as extreme as the one originating from the initial measurement can be determined (p-value). In the context of identifying circadian expressed genes, the probability of the gene expression time-course being well explained by a model of periodic gene expression could be assessed based on this p-value, rather than on the value and identifiability of model parameters. Then, only as a secondary criteria of the validity of a model, the estimated parameter value and its identifiability should be considered. In Chapter 3 such a proposed parametric bootstrap was implemented, in order to evaluate putative regulator-target interactions. The increased statistical power of the bootstrap however comes at the cost of an increased computational demand, requiring high-performance optimization procedures as well as efficient parallelization strategies.

In model rejection analysis, the p-value relates to the possibility of falsely rejecting or accepting a model out of pure chance. To avoid misinterpretation of results, the careful interpretation of the null hypothesis and its rejection is required. In Chapter 6 for example, the existence of post-transcriptional regulation of genes is inferred by rejection of the null hypothesis of protein being translated from mRNA. Setting a significance threshold of 5%, this implies that in 1 out of 20 cases post-transcriptional gene regulation is detected, when in fact protein expression can be explained via simple mRNA translation (type I error). Especially when testing many gene expression patterns, the number of

falsely rejected models can grow quite large. In order to control the type I error rate therefore additional techniques correcting for multiple testing need to be applied.

Another difficulty arising from model rejection analysis is the fact, that the acceptance of the null hypothesis does not necessarily imply its validity. In the context provided above, this implies that the failure to reject the standard model of protein production from mRNA cannot be interpreted as the definite absence of post-transcriptional gene regulation. The type II error of falsely accepting the null hypothesis is of particular disadvantage when predicting potential regulator-target interactions from gene expression time-courses, with only few data-points obtained under one experimental condition (Chapter 3). In this context, the inability to reject the null hypothesis again does not imply the existence of the regulator-target interaction, but rather states that in light of the given data the inspected interaction remains feasible. Ultimately, only an increase in measured variables under various experimental conditions will increase the ability to more reliably assess the existence of putative regulator-target interactions and increase the predictive power of the approach.

As detailed above, model rejection analysis inspects whether a given model can sufficiently reproduce a given set of measurements. If multiple alternative models exist, model selection is used to identify the one model which best reproduces the measured data while making the least number of assumptions. Problematic for model selection however is the circumstance, that a model with more degrees of freedom (parameters) will be able to better fit the measured data. In order to deal with this problem of over-fitting, one needs to assess if the addition of new model parameters is justified given the increase in the quality of model fit. If models are nested, i.e. one model can be expressed as a special case of another model, the significance of the increase in model fit can be determined using a likelihood ratio test. Accordingly, in Chapter 6 we applied such a model selection strategy with the aim to assign measured mRNA/protein time-courses to a variety of classes, corresponding to different kinetic models of translational regulation.

From a mathematical perspective, the formulated model of delayed protein translation contains the largest degree of freedom, while production, degradation, and stationary models show decreasing complexity. However, this reduction of model complexity in models excluding protein translation (stationary, degradation), seems unjustified from a biological point of view, as the existence of additional factors inhibiting the read-out of mRNA must be assumed. In this case an *ad hoc* formulated strategy to re-classify genes based on a number of criteria was designed to successfully address the problem of miss-classification. As a general rule, the situation described above once again highlights the necessity to consciously inspect model assumptions and not to confuse mathematical formalism with biological reality.

7.3 Benchmarking facilitates performance and interpretability of systems biology analysis

While careful consideration should be placed on the statistical power of methods to classify gene expression patterns, the performance of system biology approaches should always be evaluated using independent data. Often *in silico* benchmarking methods represent the method of choice. Here artificial gene expression data is produced from a system, in which the set of genes or interactions to be identified is known. Comparing the known solution with the prediction obtained by a method, method performance can be evaluated. *In silico* benchmarks carry the advantage that critical properties of the data, such as the timing and amount of measurements, as well as the extent of measurement noise, can be controlled. However, when generating benchmark data, realistic experimental conditions should be assured.

Marbach et al. (2009) for example implement realistic structural properties of gene regulatory networks by extracting sets of interacting genes from widely accepted *E. coli* or yeast transcriptional networks [Marbach et al., 2009, Schaffter et al., 2011a]. Kinetic properties of gene regulation are then captured using a thermodynamic model of gene expression and random Gaussian noise is added to simulated gene expression data. In the DREAM3 challenge, simulated gene expression data was provided from a range of networks and participants of the challenge were asked to rank potential interactions of the network based on their inferred likelihood. Simulated data included steady-state gene expression measurements under knock-down as well as knock-out conditions. Also multiple time-courses simulated under various perturbation conditions were supplied. Although some participants scored exceptionally well in reconstructing network structures, most of the participating teams performed only marginally better than a random guess.

This is in agreement with our observation in Chapter 3, in which most network inference methods showed poor performance while inferring the structure of specific benchmark networks. We speculate however, that this low performance was a result of the sparsity of gene expression measurements with only one time-course serving as the basis for network inference, instead of the low predictive power of the methods. Owing to this sparsity of the measurement data, it is possible that many network interactions are simply not identifiable from the provided time-course gene expression data. Consequently, the performance of most of the tested network inference approaches improved significantly when additional gene expression data, measured under knock-down conditions, was provided (Chapter 4).

While *in silico* benchmarking is certainly useful, realistic benchmarking conditions can only be guaranteed if a true biological system is used to acquire measurements of gene expression. In one such example, Lichtenberg et al. (2005) compiled lists of potential yeast cell-cycle genes based on a combination of criteria, including previous identification as a cell-cycle gene or association of genes with known cell-cycle regulators [de Lichtenberg et al., 2005]. Various methods to identify periodic expressed genes could then be applied to time-course yeast gene expression data and evaluated against this known set of cell-cycle genes. Similarly, gene expression datasets have been compiled from real biolog-

ical systems with the aim to evaluate the performance of data-driven network inference methods. For example, a large collection of *E. coli* gene expression data under various experimental conditions was integrated by the Many Microbe Microarrays Database (M3D) [Faith et al., 2007a]. In combination with the RegulonDB database, which collects experimental evidence of genetic interactions in *E. coli*, the quality of structural predictions made by network inference methods can be tested [Gama-Castro et al., 2015].

A major caveat of drawing on true biological systems to evaluate systems biology tools however, is that experimental evidence from which the sets of periodic expressed genes or existing genetic interactions are identified, is still incomplete. Therefore, reproducing old results is often favoured over new predictions. In the future, tools from synthetic biology may provide more reliable means of generating real biological benchmark datasets, in which the true set of periodic expressed genes or genetic interactions is specified by design [Cantone et al., 2009].

In the long run, thorough benchmarking of systems biology approaches will provide further insight into their respective strengths and weaknesses, give critical guidance when interpreting results, and help identify the type, amount, and quality of data by which the performance of a given computational method may be improved.

7.4 Integration of information enhances predictive power of systems biology analysis

As a general strategy in this thesis we further follow the principle of integrating multiple sources of information in order to enhance the predictive power of systems biology analysis. For example, it has been previously shown that the performance of a particular network inference tool may vary greatly across different datasets [Marbach et al., 2012]. Regarding this uncertainty, it is an open question, which network inference method performs best on a given dataset. In order to alleviate this problem, the performance of network inference tools could be evaluated by matching predicted interactions with known interactions present in the network. As such information is typically sparse, in this thesis we followed an alternative strategy, in which predictions made by different available systems biology tools were integrated in order to increase overall predictive power. In Chapter 2 for example, high confidence circadian expressed genes were identified from the overlap of predictions originating from multiple applied computational approaches to identify periodic expressed genes. The integration of results produced an exceptionally good agreement between predicted and known circadian genes.

Following this strategy, in Chapter 4 a community prediction of the gene regulatory network controlling EMT was obtained by combining the predicted rank of interactions across multiple network inference methods. As a result, this community prediction showed performance comparable to the best performing network inference approach. In general, this positive performance of the community prediction is believed to be robust also across different datasets. Finally, by further combination of the community prediction with a dynamical modelling approach in Chapter 5, we not only observed an additional increase

in the ability to correctly reconstruct a given network structure, but also acquire the potential to predict gene expression dynamics under a variety of experimental conditions.

Apart from integrating various systems biology tools methods applied to the same dataset, throughout this thesis we further augment systems biology analysis by various experimental as well as bioinformatics approaches operating on independent data. In Chapter 3 for example, potential interactions between circadian genes were determined based on the putative binding of a regulator to its target gene. Also the analysis of biological functions distributed across defined sets of genes provided further context for the interpretation of results (Chapter 2 and Chapter 6). Such integrative analysis, using tools from systems biology as well as bioinformatics, will in the future not only add to our understanding of the biological process at hand, but may further enhance the predictive power of applied systems biology approaches.

Lastly, the holistic mindset of systems biology systems biology not only allows, but rather *requires* the integration of multiple independent datasets obtained from different levels of gene expression. Accordingly, in Chapter 6 we integrated genome-wide mRNA and protein data, allowing for the identification of post-transcriptional regulated genes during *Drosophila* embryogenesis. In the future hopefully more such integrated datasets will be made available, possibly further revealing information on relevant histone modifications, chromatin structure, active translation and many more.

7.5 Experimental evaluation of systems biology analysis

An isolated model without any relation to measurable observables can not be tested and consequently not falsified. Therefore, the ultimate test for the predictive power of any hypothesis must be its experimental validation. In Chapter 6 for example, we experimentally validated the hypothesis of Hrb98DE being involved in post-transcriptional regulation of glucose metabolism. Further, in Chapter 4 and Chapter 5, structural predictions of the gene regulatory network controlling EMT were compared to interactions identified by careful review of the current literature, motif analysis, and publicly available ChIP data. Particularly in the case of publicly available ChIP data however, neither the provided cell-types nor applied experimental treatments matched those of our study. In the near future therefore, predictions on the structure of the gene regulatory network of EMT as well as anticipated dynamics of gene expression under various perturbation conditions should be carefully evaluated in a more appropriate context.

In this as well as any other case, systems biology analysis requires the close collaboration between experimentalists and theoreticians. Only by this close collaboration can we ensure that formulated models realistically capture our current knowledge of the system, facilitate the interpretation of results, and design new and meaningful experiments to test our understanding of the biological process.

Chapter 8

Material and Methods

8.1 Materials and Methods for Chapter 2

8.1.1 Cell culture, synchronization, small interfering RNA knockdown, and luminescence measurement.

NIH 3T3 Per2:luc cells (kindly provided by Hiroki R. Ueda [Isojimaa et al., 2009]) were cultured at 37°C, 5 or 7% CO_2 , and in high humidity. Culture medium was Dulbeccos modified Eagles medium (DMEM; catalogue number 21969035; Life Technologies) supplemented with 10% fetal bovine serum (catalogue number 10270106; Life Technologies), 2 mM L-glutamine (catalogue number 25030024; Life Technologies), and 1 non-essential amino acids (catalogue number 11140-035; Life Technologies). Two days after seeding the cells at a density of 3,000 cells/cm^2 , the culture medium was removed and changed to medium containing 100 nM dexamethasone (catalogue number D4902; Sigma-Aldrich) and, 2h later, medium was changed back to normal medium (time point 0h) [Ma et al., 2013, Kwak et al., 2006, Lee et al., 2012, Takarada et al., 2012, Lee et al., 2014, Onishi and Kawano, 2012, Izumo et al., 2006, Balsalobre et al., 2000, Goriki et al., 2014].

For luminescence measurements, NIH 3T3 Bmal1:luc cells were generated by transfecting NIH 3T3 cells (CRL-1658; ATCC) with a plasmid expressing Bmal1:luc (Bmal1:luc-pT2A) and a Tol2 transposase-expressing plasmid (pCAGGS-TP) (plasmids were kindly provided by K. Yagita [Yagita et al., 2010]) followed by selection with 250 μg/ml hygromycin. Two days before the start of the measurement, cells were seeded at a density of 3,000 cells/cm^2 and at the same time transfected with 30 pmol ON-TARGETplus SMARTpool siRNA (i.e., a mixture of four siRNAs targeting the same gene) against the gene of interest or a non-targeting control (NTC; Dharmacon) per 3.5-cm plate by using Lipofectamine RNAiMAX (catalog number 13778150; Invitrogen). After 2 days, cells were synchronized with dexamethasone as described above, and after 2h medium was changed to DMEM without phenol red (catalog number 31053-028; Life Technologies) containing 10% fetal bovine serum, 2 mM L-glutamine, 1× non-essential amino acids, 1× sodium pyruvate (catalog number 11360-039; Life Technologies), 25 mM HEPES (catalog number H0887; Sigma), and 500 μM beetle luciferin (catalog number E1601; Promega), and cells were transfected again with siRNA. Luminescence measurements were performed with a LumiCycle 32 (ActiMetrics) and analysed with the LumiCycle analysis program,

version 2.3. A running average of 24h was used for the baseline fit, and the period length was calculated by fitting a damped sine wave to the data.

8.1.2 Quantitative reverse transcription-PCR

Total RNA was prepared using TRIzol (catalog number 15596; Invitrogen) or a Purelink RNA Minikit (catalog number 12183025; Life Technologies) and reverse transcribed with a First-Strand cDNA synthesis kit (catalog number K1612; Fermentas). Transcripts were quantified by PCR using SYBR green PCR master mix (catalog number 4334973; Life Technologies) on a ViiA7 PCR machine (Life Technologies).

8.1.3 Identification of circadian expressed genes

In the `CBNLR` (classification by non-linear regression) approach, a sine function dependent on three parameters (amplitude (A), period length (T), and phase shift (ϕ) - see Equation 2.1) is fit to the detrended time-course data of each gene. Prior to fitting, detrending of time-courses is carried out using the `python scipy.signal.detrend` package. Parameter estimation of the model is carried out using a multi-start local optimization scheme. In total 5 initial parameter vectors are calculated using latin-hypercube sampling provided by the `pyDOE` toolbox. The Levenberg-Marquardt algorithm used for parameter estimation is implemented in the `LMfit` package. The step length for the forward-difference approximation of the Jacobian was set to ϵ = 1.0e-03, while the initial step bound was set to α = 1.0e-01. A maximum number of 400 function evaluations was carried out. The tolerance in both the relative error desired in the sum of squares and the relative error desired in the approximate solution was set to 1.0e-05. For each gene with a period length between 20h and 28h we created 1,000 non-parametric bootstrap samples. For each data point a normal distribution with a standard deviation corresponding to the estimated experimental error between the two biological replicates was assumed. By fitting the model to bootstrapped data, a distribution of estimated parameters over the bootstrap samples is obtained. For the classification of cyclical genes, we only selected genes with a mean period (T) between 20h and 28h and a small relative error.

Table 8.1: Parameter ranges selected for `CBNLR`. Shown are chosen lower and upper boundary of parameters values while fitting Equation 2.1 to time-course gene expression data using the `CBNLR` approach.

parameter	description	lower boundary	upper boundary
A	Amplitude	1.0e01	1.0e06
T	Period Length	10.0	32.0
ϕ	Phase Shift	0.0	2π

In addition to our approach, we further identified cyclically expressed genes using `JTK_CYCLE` [Hughes et al., 2010], `ARSER` [Yang and Su, 2010], (benchmark and NIH 3T3 data) and `RAIN` [Thaben and Westermark, 2014] (NIH 3T3 data only). The published methods were applied as described in the respective manuals. For all four methods,

genes with a period length between 20h and 28h and a false-discovery rate below 0.05 were selected to be circadian expressed.

8.1.4 Benchmarking methods to identify circadian expressed genes

In order to evaluate the proposed method of classifying circadian genes a benchmark dataset according to Yang & Su (2011) [Yang and Su, 2010] was constructed: Periodic genes were simulated using trigonometric functions with various period lengths between 20h and 28h as well as phase shifts over the entire cycle. 1000 time-courses for both stationary signals (non-decaying) and non-stationary signals (decaying in signal amplitude and signal strength) were constructed with a signal-to noise ratio ranging from 1 to 5. Two datasets (datasets A and B) were constructed by combining either 1000 stationary or non-stationary periodic signals with a white-noise background (non-periodic signals). In test datasets C and D the stationary or non-stationary periodic signals were combined with a background composed of autoregressive processes of order AR(1). The AR coefficients α for each AR signal were estimated by selecting random genes from the NIH 3T3 data. Also timing and amount of measurements were chosen according to the time points of the RNAseq experiment presented in this study, where measurements took place every 4h until 32h after cell synchronization. Two replicates of the experiment including measurement error were generated and the mean and standard deviation calculated in order to simulate experimental duplicates.

8.1.5 RNA-seq data analysis

RNA-Seq data for multiple time points of the circadian cycle were generated in biological duplicates by using Illumina sequencing [50-bp reads; poly(A) RNA sequencing, non-strand specific; Illumina HiSeq 2000]. Reads were aligned to the mouse genome (mm9) by using TopHat (version 2.0.9) with default options [Trapnell et al., 2009]. Normalized read counts for each samples were calculated using the DESeq package after applying library size normalization [Anders and Huber, 2010]. Library size-normalized .wig files were created for all the samples by using the QuasR package [Gaidatzis et al., 2015]. The RNA-seq data were deposited with NCBIs Gene Expression Omnibus database [Edgar, 2002]. Only genes with a normalized read count above 64 at any of the time points were considered for further analyses. The lincRNA annotations were collected from NonCode [Liu et al., 2005], ENSEMBL [Flicek et al., 2014], UCSC [Kent et al., 2002], and other published resources [Ramos et al., 2013]. Normalized read counts for each sample were generated using the DESeq package after applying library size normalization [Anders and Huber, 2010]. Only lincRNAs expressed above 32 normalized read counts at any of the time points and which were found to be cyclically expressed by three out of four methods were considered for further analysis.

8.1.6 GO term analysis

Functional annotation tables were generated using the Gene Ontology (GO) browser supplied by the Mouse Genome Database (MGD; Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME) [Blake et al., 2014]. GO terms included were GO:0007623 circadian rhythm, GO:0006950 response to stress, GO:0007049 cell cycle, GO: 0003677 DNA binding, and GO:000635 regulation of transcription, DNA-templated, all with corresponding child terms. GO term analyses for all cyclical genes and subgroups were performed using the DAVID functional annotation tool [Huang et al., 2009a, Huang et al., 2009b]. Gene set enrichment analysis (GSEA) was performed by sorting the cyclical genes according to their phase, which was determined with `JTK_CYCLE` [Subramanian et al., 2005] (http://www.broadinstitute.org/gsea/index.jsp).

8.1.7 Motif analysis

Motif analysis was carried out using the `HOMER` software [Heinz et al., 2010]. A promoter region was defined as bp -1000 to +1000 around the transcription start site. The cutoff values for the different promoters were set to the default values given by the `HOMER` program.

8.1.8 Tissue data and CircaDB

Tissue data were obtained from [Zhang et al., 2014]. Cyclical genes of these and other published tissue datasets, as well as cell line datasets, were obtained from CircaDB database (http://bioinf.itmat.upenn.edu/circa/) [Pizarro et al., 2013].

8.1.9 Analysis of ChIP-Seq data

ChIP-Seq data were downloaded from the Gene Expression Omnibus (GEO) database (NCBI). Fastq files were processed using `Bowtie` (version 0.12.7) with default parameters for alignment to mouse genome mm9 (available from UCSC) [Langmead, 2010]. Positions of promoters (-800 to +200 with respect to the transcription start site [TSS]) were determined using the `GenomicFeatures R bioconductor` package. Promoter enrichments over inputs were determined by using the same method as described by Schick et al. [Schick et al., 2015]. Only peaks showing an enrichment 1.5-fold above input were considered for further analysis. The datasets used were the following: for Bmal1 ZT08, Clock ZT08, and Cry1 ZT20 in liver, GSE53828 (corresponding inputs at ZT08 and ZT20 were kindly obtained from the authors for that GEO entry); for the Bmal1 time-course from ZT02 to ZT22 in liver and associated input, GSE26602; for Bmal1 ZT08, Clock ZT08, and input in liver, GSE36916; for Rora in liver at ZT22, GSE59486 (associated input at ZT22, GSE26345); for Rev-ErbB in liver at ZT10 and ZT22, GSE36375 (associated inputs at ZT10 and ZT22, GSE26345); for Rev-ErbA in liver, Rev-ErbB in liver, and associated input, GSE34020; for Rev-ErbA_macrophage, Rev-ErbB_macrophage, and associated input, GSE45914.

8.2 Materials and Methods for Chapter 3

8.2.1 Inference of regulator-target interactions by NodeInspector

In `NodeInspector` models of transcriptional regulation are fit to time-course gene expression data by minimizing the weighted least squares difference between modelled and measured gene expression values (χ^2, see Equation 2.2). To estimate variance in gene expression measurement, we adopted a linear error model using linear regression to find the best fit between expression values and corresponding standard deviations. Chosen parameter ranges are given in Table 8.2. In Equation 3.1 the parameter R_{tc} was formulated in terms of other parameters as

$$R_{tc} = \frac{2\lambda_{mRNA}\ y_a(0)\ \sqrt{(h + y_b(0)\ w)^2 + 1}}{\sqrt{(h + y_b(0)\ w)^2 + 1} + h + y_b(0)\ w} \tag{8.1}$$

where y_i represents gene expression measurements of gene a or gene b respectively. In case of transcriptional regulation modelled by a Hill-function R_{tc} is given by

$$R_{tc} = \lambda_{mRNA}\ y_a(0)\ \frac{y_b(0)^n + k^n}{y_b(0)^n}. \tag{8.2}$$

In the Hill-function containing additional translation of mRNA into protein, R_{tl} was set equal to λ_p. Gene expression of the regulating gene in all cases is given by the linear interpolation between data points.

The cost function is minimized using the Augmented Lagrangian Particle Swarm Optimizer (`ALPSO`) provided by the `pyOpt` optimization package [Perez et al., 2012]. The swarm size of the particle swarm optimization scheme was set to 20, while cognitive and social parameters were set to $c_1 = 2.8$ and $c_2 = 1.3$ respectively. As neighbourhood model `dlring` was chosen. As soon as optimization by `ALPSO` terminated because internal convergence criteria of the optimization method were reached (`stopCriteria = 1`), additional local optimization was carried out using the Sequential Penalty Derivative-free method for Non-linear constrained optimization (`SDPEN`) method of the `pyOpt` package. For local optimization via `SDPEN` the standard parameters set by `pyOpt` were selected. In case of the benchmark gene expression data generated by `GNW`, a prediction of the network structure was obtained by ranking all possible interactions by the estimated χ^2-value, where the lowest χ^2-value corresponds to the best rank. Activating regulation was assumed if model parameters n (Equation 3.2 and Equation 3.3) or w (Equation 3.1) were positive, otherwise the regulation was considered inhibitory.

In the modified `NodeInspector` approach, bootstrap sample size of the parametric

Table 8.2: Parameter ranges selected for `NodeInspector`. Shown are parameter ranges selected while estimating model parameters in the `NodeInspector` approach for each of the three chosen regulation functions (Reinitz and Sharpe: Equation 3.1, Hill - mRNA only: Equation 3.2, Hill - mRNA and protein: Equation 3.3).

regulation function	parameter	description	lower boundary	upper boundary
Reinitz and Sharpe	h	promoter sensitivity	-5.0	5.0
Reinitz and Sharpe	w	regulatory weight	-20.0	20.0
Reinitz and Sharpe	λ_{mRNA}	degradation mRNA	0.014 (48.0h)	1.39 (0.5h)
Hill mRNA only	k	TF affinity	1.0e-03	1.0e-06
Hill mRNA only	n	TF cooperativity	-5.0	5.0
Hill mRNA only	λ_{mRNA}	degradation mRNA	0.014 (48.0h)	1.39 (0.5h)
Hill mRNA and protein	k	TF affinity	1.0e-03	1.0e+03
Hill mRNA and protein	n	TF cooperativity	-4.0	4.0
Hill mRNA and protein	$\lambda_{protein}$	degradation protein	0.014 (48.0h)	1.39 (0.5h)
Hill mRNA and protein	λ_{mRNA}	degradation mRNA	0.014 (48.0h)	1.39 (0.5h)

bootstrap chosen to evaluate regulator-target interactions from NIH 3T3 data was set to 1000. Bootstrap samples were generated by perturbing simulated gene expression of the fitted model according to normally-distributed uncorrelated noise with standard deviation corresponding to the respective gene expression measurements. Models were fit to perturbed gene expression data using the optimization scheme presented above. P-values were calculated based on the empirical cumulative density of the χ^2-distribution obtained via the `python statsmodels` package. A right-sided test with a cut-off p-value of < 0.05 served as rejection criteria for the proposed models.

8.2.2 Evaluation of network inference methods using benchmark networks

Time-course gene expression data for benchmark networks of different size were generated using GeneNetWeaver (`GNW`) version 3.1 beta [Schaffter et al., 2011a]. From the *E. coli* gene regulatory network provided, sub-networks of different sizes were extracted (6, 11, 16, 21, 26 genes) and auto-regulation was removed. From these networks kinetic models were formulated and gene expression simulated in which at $t = 0$ of the experiment the input gene was perturbed (mutifactorial gene perturbation, factor -1.0). Data was simulated three times independently for 600 time-units containing 25 equally spaced gene measurements before it was normalized to the 0h time-point. A linear error model was fit to the standard deviation and mean expression values across three replicates to estimate the experimental error. Network inference methods `ARACNE`, `CLR`, `MRNET`, and `MRNETb` are implemented in the `MINET R` package [Meyer et al., 2008]. The mutual information between gene expression time-courses in these methods was estimated based on the Miller-Madow asymptotic bias corrected empirical estimator (`option: mi.mm`) together with the `equalfreq` discretization method. Network inference by `PREMER` was carried using the implementation provided by `pyPREMER` [Villaverde et al., 2017]. In total three entropy reduction rounds were carried out using default parameters set by the method with $\tau^{min} = 5$. The network inference method `GENIE3` was run using default settings recom-

mended by the method [Huynh-Thu et al., 2010]. For network inference via `Inferelator` the `2015.08.05` release was applied. Grouping of genes based on the similarity of gene expression was disabled. Metadata-files describing the temporal relationships of data-points were provided. Otherwise parameters were set to their default values. All network predictions obtained from tested network inference methods were evaluated using the packages `roc_curve`, `auc`, `precision_recall_curve`, `average_precision_score` provided by the `python` `sklearn.metrics` toolbox.

8.2.3 Evaluation of predicted interactions between circadian expressed genes identified from NIH 3T3 data

In order to evaluate predicted interactions by the `NodeInspector` approach, NIH 3T3 cells were seeded at a density of 3,000 cells/cm^2 and at the same time transfected with 30 pmol ON-TARGETplus SMARTpool siRNA (i.e., a mixture of four siRNAs targeting the same gene) (Dharmacon) per well of a 6-well plate by using Lipofectamine RNAiMAX (catalogue number 13778150; Invitrogen). After 3 days, cells were harvested for further analysis by RT-qPCR 8.1.2. The statistical significance of deregulation observed in the experiments was tested using a Mann-Whitney U test (with significance set at a p-value of < 0.05) using the `python` function `scipy.stats.mannwhitneyu`.

8.3 Materials and Methods for Chapter 4

8.3.1 qPCR experiments

NMuMG cells were cultured in DMEM supplemented with 10% FBS, 2 mM L-Glutamine, 1x non-essential amino acids) at 37°C with 7% CO^2 in a humid incubator. For the wild-type time-course experiment NMuMG cells were seeded at a density of 3×10^5 cells per 6 cm plate in a total volume of 3 ml NMuMG medium (440 ml DMEM, 50 ml FBS, 5 ml Glutamine, 5 ml non-essential amino acids) and incubated 24h before media was removed and cells treated with 3 ml TGFβ diluted at 2 ng/ml in NMuMG medium. At each hour after TGFβ-treatment, cells were harvested using trypsin. Harvested cells were collected in PBS and spinned down at 180g for 5 min. Before further processing, pellets were stored at −80°C. RNA isolation was carried out according to the protocol provided by Chomczinski 1987 [Chomczyński and Sacchi, 1987] while cDNA was synthesized using a First Strand cDNA Synthesis Kit (Thermo Scientific). Primers used for qPCR are listed in Table 8.3. As reference genes Tbp and Rpl19 were chosen. All measurements were performed in technical duplicates, apart from measurement of housekeeping genes and untreated samples, which were measured in triplicates. The experiment was repeated using three biological replicates.

Table 8.3: Table of primer sequences for selected EMT genes. Listed are primer sequences for selected EMT relevant genes measured by qPCR in NMuMG cells.

	forward	reverse
Atf3	gctggagtcagttaccgtcaacaa	cgcctccttttcctctcatcttc
Cdh1	gagacaggctggctgaaagtgac	tgacacggcatgagaatagagga
Cdh2	ggtggaggagaagaagaccagga	tggcatcaggctccacagtatct
Crb3	gtgctcttggcgtttggtttg	atttgtgaaagggtccggtgct
Ets1	cgtgctgacctcaacaaggacaa	cagaagaaactgccacagctgga
Fn1	cggagagagtgcccctacta	cgatattggtgaatcgcaga
Hmga2	ggtgccacagaagcgaggac	gtttttgctgcctttgggtcttc
Id1	ggggcgactaataggaaaaagctc	ggagaggcgttcccatttctcta
Id2	accagagacctggacagaaccag	agctcagaagggaattcagatgc
Id3	gaggagctttgccactgacc	gaagctcatccatgccctcag
JunB	caccacggagggagagaaaatc	agttggcagctgtgcgtaaagg
Ncam	aagcacacagagcccaacgag	aaggcagcatgtcctcgacagt
Ocln	catgtccgtgaggccttttga	tggtgcataatgattgggtttga
Serpine1	gaggtaaacgagagcggcacagt	atgcgggctgagatgacaaag
Snail1	tctctaggccctggctgcttc	cagcaaaagcacggttgcagt
Sox4	ctcgctctcctcgtcctct	cgtcttcgaactcgtcgtc
Vim	tgccaacctttttcttccctgaac	tttgagtgggtgtcaaccagagg
Zeb1	gtgatccagccaaacggaaacc	tctttttgggtggcgtggagt
Zeb2	ccagaggaaacaaggatttcagg	cggagtctgtcatgtcatctaggc

In the knock-down experiment 5 nmol siRNA against each target gene were diluted in 250 µl 1x siRNA buffer. The siRNA premix contained 200 µl Opti-MEM, 1.5 µl siRNA (30 pmol), 5 µl Lipofectamine RNAimax. The premix was incubated 15 min before 200

µl was added to the cells, which were seeded at 120000 cells/6 well in 1.8 ml medium buffer. Cells were cultured for 48h before being seeded at 240000 cells per 6 cm plate in 3.591 ml medium, three dishes per condition. After seeding, cells were again treated with 419 µl siRNA premix. After 24h, one dish was harvested in 1 ml TRIzol as an untreated control. Remaining plates were treated with 2 ng/ml 0.5 µl TGFβ. At each time point (4h, 18h) cells were harvested in 1 ml TRIzol. In total the experiment was performed in biological triplicates. RNA isolation, cDNA synthesis and RT-qPCR was performed as described above.

8.3.2 Data processing

Raw threshold cycle values (ct) for time-course (WT; 0-24h) and knock-down experiments (NTC, siEts1, siHmga2, siId1, siId2, siId3, siSerpine1, siSmad4, siSnail1, siSox4, siZeb1, siZeb2; 0h, 4h, 18h) were obtained by qPCR measurement. ct values were averaged over technical replicates. In case only one technical replicate existed, the data-point was discarded. If the calculated standard deviation between the two technical replicates of a data-point exceeded a threshold defined by the linear equation $0.025ct + 0.125$, the data-point was discarded. For Hmga2 the 11h time-point in replicate 2 of the wild-type gene expression data was identified as an extreme outlier upon visual inspection and removed manually. Remaining ct values were normalized to housekeeper genes Rpl19 and Tbp independently, obtaining Δct values normalized to each of the housekeeping genes. Δct^{Rpl19} and Δct^{TBP} values were normalized to the 0h time-point of the wild-type data (time-course experiment) or NTC data (knock-down experiment) resulting in $\Delta\Delta ct$ values. $\Delta\Delta ct$ values were then averaged over both housekeeping genes and the fold-change relative to either the wild-type 0h sample (time-course experiment) of NTC 0h sample (knock-down experiment) was calculated ($2^{-\Delta\Delta ct}$). Biological replicates were averaged to obtain the final fold-change data. Based on the correlation between of standard deviation and mean fold-changes a linear error model with a relative error of 25% and an absolute error of 2.5% was chosen.

8.3.3 Data analysis

Differential expression of genes in knock-down cell lines was determined using an independent t-test (two-sided, p-value < 0.025) comparing target gene expression against expression in cells treated with non-targeting control siRNA (NTC). Similarity of target gene expression over the different knock-down cell lines (including NTC cells) was determined by calculating the Spearman correlation between log10 transformed fold-changes of each gene. The clustering of the corresponding dendogram was carried out using a complete linkage criterion. Accordingly the similarity of the different cell lines with respect to target gene expression was assessed.

8.3.4 Benchmark network and data

The benchmark data for the RE approach could not be generated via gene net weaver, since it is not possibly to apply two separate perturbations at different times. For the RE

approach however it is necessary that the knock-down is applied before TGFb stimulation. Benchmark data was generated similar to the approach used in [Schaffter et al., 2011b]. Using the original software provided by the developers however it was not possible to simulate two perturbations (knock-down and TGFβ-stimulation) at different time points. Generating a reliable Benchmark we had to fall back to our own implementation of gene regulatory networks based on the benchmark equations given in [Marbach et al., 2010b]. In the benchmark network both mRNA and protein expression are accounted simulated. We chose the network to include one independent input gene, 12 TFs and 7 effector genes. Between these genes an total of 26 interactions were assigned using the *E. coli* network of the original GNW software as a template. While in theory a gene is potentially regulated by multiple upstream regulates, we restricted the number of upstream regulators to a maximum of two genes. Parameters of the benchmark network were randomly assigned in between the ranges given in 8.4.

Table 8.4: Parameter ranges for the simulation of benchmark data. Parameter ranges chosen to generate simulations of the benchmark network. Parameters correspond to those detailed in [Marbach et al., 2010b].

parameter	lower limit	upper limit	type
m_i	0.015	0.045	real
r_i	0.015	0.045	real
λ^{RNA}	0.015	0.045	real
λ^{Prot}	0.015	0.045	real
a_0	0	1	real
a_1	0	1	real
a_2	0	1	real
a_3	0	1	real
n_j	1	6	real
k_j	0	1	real
ρ	0	1	integer

Final parameter values of the benchmark network are given in `/home/bleep/ofc/projects/EMT_new/0.1_Data_Benchmark/GNW2/candidates/GNW2_Benchmark_26.csv`. Using these parameter values, simulation was carried out for 2000 time-units until steady-state was reached, before simulating wild-type and knock-down conditions. In order to simulate knock-down conditions the degradation rate of the affected gene was multiplied by its corresponding knock-down efficiency parameter. Knock-down condition were simulated for 1000 time-units before expression of the input gene changes. Hereafter all simulations (wild-type and knock-down) are carried out for 500 time-units. The wild-type time-course was sampled at an interval of 20 time-units, while for the knock-down experiment gene expression at time 0, 80 and 360 were extracted. Three independent replicates of the time-course and knock-down experiment were generated adding a normally distributed error with an absolute error of 5% and a relative error of 20%. All measurements were normalized to the expression at 0h relative to the expression change in the input gene and the mean over all three replicates was calculated. All experiments are assigned the same realization of the input time-course (wild-type). Expression of

TF07 was removed from the knock-down data, while expression of TF10 was removed from wild-type data.

8.3.5 Network inference on NMuMG expression data

Network inference on data obtained from TGFβ-treated NMuMG cells was carried out using multiple state-of-the-art network inference tools (`PREMER`, `Inferelator`, `GENIE3`, `MRNET`, `MRNETb`, `ARACNE`, `CLR`, `NodeInspector`). Due to missing data (Sox4 in time-course, JunB in knock-down data), network inference for all network inference approaches - with the exception of `NodeInspector` - was based on time-course and knock-down data separately. In the time-course data pSmad2 expression measured by western blot was used as a measure of the activity of the input (Smad), while in the knock-down data Smad4 expression measured by qPCR was selected. All network inference methods were carried as described in Section 8.2.2 out using default settings, with the exception `PREMER`, which was carried out with $\tau^{min} = 0$. In the `NodeInspector` approach the function given by Reinitz and Sharpe (1995) was chosen to represent transcriptional regulation. Respective ranges are given in Table 8.2. From all possible interactions auto-regulation, regulation of the input (Smad), regulation by effectors, and interactions Sox4 $\rightarrow$ JunB and JunB $\rightarrow$ Sox4 were excluded. If possible this was done *a priori* (before network inference: `PREMER`, `GENIE3`, `Inferelator`, `NodeInspector`) or *a posteriori* (after network inference: `MRNET`, `MRNETb`, `ARACNE`, `CLR`).

Individual predictions were integrated by a modified Rank-Borda method. Ranks of items not evaluated in a certain dataset (Sox4 in time-course data and JunB in knock-down data) were set to None, whereas items which are evaluated but resulted in a score of 0 are ranked as $k - 1$, where k is equal to number of ranked elements by the respective network inference approach. While integrating individual predictions, interactions containing Smad species (pSmad2/Smad4) therefore were merged.

8.3.6 Motif analysis

PWM files for 8 TFs were obtained from either HOMER software or the JASPAR Database. Snail1 is known to bind to the E-box motif, which is why this motif has been selected to identify Snail1 binding. The following motifs were selected for further motif analysis.

Promoter sequences were defined as -1000 to 1000bp around the TSS and retrieved using the HOMER software [Heinz et al., 2010]. 1000 background genes were selected from expressed genes in previously published RNA-sequencing data obtained from NMuMG cells [Sahu et al., 2015] and background base transition probabilities calculated from a 3rd-order Markov model. Motif occurrences were identified using FIMO software [Grant et al., 2011]. Only motif occurrences with $p < 0.0001$ were selected for further analysis. The existence of at least one motif occurrence in the promoter sequence of a gene qualified as an interaction. Motif occurrences for all Smad motifs (Smad2, Smad3, Smad4) were combined. In total 53 interactions between the 8 TFs and the 19 targets genes were predicted using this approach.

Table 8.5: Overview of motifs selected for motif analysis. Listed are details of positional frequency matrices selected for motif analysis: The species from which matrices are derived, the motif length, the database from which the motif extracted and - if possible - the source publication in which the motif was defined.

TF	species	length	database	source
Atf3	mouse	8	Jasper	Weirauch PBM (2014)
Ets1	mouse	15	Jasper	- unspecified -
Id2	mouse	8	Jasper	- unspecified -
JunB	human	11	Jasper	- unspecified -
Smad2	- unspecified -	8	Homer	ES-SMAD2-ChIP-Seq(GSE29422)
Smad3	- unspecified -	8	Homer	NPC-Smad3-ChIP-Seq(GSE36673)
Smad4	- unspecified -	10	Homer	ESC-SMAD4-ChIP-Seq(GSE29422)
Snail1 (E-Box)	- unspecified -	12	Homer	- unspecified -
Sox4	mouse	10	Homer	proB-Sox4-ChIP-Seq(GSE50066)
Zeb1	human	9	Jasper	- unspecified -

The expected overlap of interactions found by motif analysis and interactions reported in the literature is calculated based on the set of 152 background interactions assessed by both approaches. In this common background set 21 interactions were identified by literature research, while 53 interactions were derived using motif analysis. The common set between interactions found by both approaches contained 9 interactions. Similarly the overlap between interactions derived from motif analysis and top interactions ranked by network inference was calculated. As background the set of 143 interactions evaluated by both approaches was chosen, leaving 49 confirmed interactions by motif analysis (excluded interactions are: JunB $\rightarrow$ JunB, Snail1 $\rightarrow$ Snail1 ,Zeb1 $\rightarrow$ Zeb1, and Sox4 $\rightarrow$ JunB).

8.3.7 ChIP analysis

Lists of putative targets derived from 44 ChIP experiments (Table 8.6) were downloaded from the cistrome database [Wang et al., 2014, Mei et al., 2017]. These 44 experiments cover 6 factors selected as EMT relevant in this study (Atf3, Ets1, Hmga2, JunB, Sox4, Smad2/3/4) and are mouse specific. In each target list binding occurrences or each TF are ranked. An interaction between TF and target was assigned, if the EMT relevant factor appeared within the top 600 ranked target genes of at least one of the considered experiments ChIP experiments for the TF. From this analysis 20 interactions were identified.

To evaluate the overlap between interactions identified by analysis of existing ChIP data and interactions derived from literature only the common background of 114 genes was considered. Within this background 26 interactions derived from the literature are found (grade A and B). The background set of interactions evaluated both by network inference and ChIP data contains 107 interactions.

Table 8.6: Overview of selected ChIP data. Listed are selected publicly available mouse ChIP datasets used to reconstruct the gene regulatory network controlling TGFβ-induced EMT in NMuMG cells.

accession_id	cell	gene	reference
GSM1334038	Macrophage; Bone Marrow	ATF3	Krebs W- et al. Nucleic Acids Res. 2014
GSM1334037	Macrophage; Bone Marrow	ATF3	Krebs W- et al. Nucleic Acids Res. 2014
GSM1334041	Macrophage; Bone Marrow	ATF3	Krebs W- et al. Nucleic Acids Res. 2014
GSM539546	38B9; Pre-B Lymphocyte; Blood	SMAD3	Mullen AC- et al. Cell 2011
GSM1334036	Macrophage; Bone Marrow	ATF3	Krebs W- et al. Nucleic Acids Res. 2014
GSM1334039	Macrophage; Bone Marrow	ATF3	Krebs W- et al. Nucleic Acids Res. 2014
GSM1334040	Macrophage; Bone Marrow	ATF3	Krebs W- et al. Nucleic Acids Res. 2014
GSM654875	Th2; Blood	ETS1	Wei G- et al. Immunity 2011
GSM1558587	Embryonic Fibroblast	HMGA2	Singh I- et al. Cell Res. 2015
GSM1558589	Embryonic Fibroblast	HMGA2	Singh I- et al. Cell Res. 2015
GSM1558591	Embryonic Fibroblast	HMGA2	Singh I- et al. Cell Res. 2015
GSM1558593	Embryonic Fibroblast	HMGA2	Singh I- et al. Cell Res. 2015
GSM1022318	Macrophage; Bone Marrow	JUNB	Ostuni R- et al. Cell 2013
GSM539549	V6.5; Embryonic Stem Cell; Embryo	SMAD2/3	Mullen AC- et al. Cell 2011
GSM539543	C2C12; Myoblast; Muscle	SMAD3	Mullen AC- et al. Cell 2011
GSM539544	C2C12; Myoblast; Muscle	SMAD3	Mullen AC- et al. Cell 2011
GSM898371	Cortex	SMAD3	Estar?s C- et al. Development 2012
GSM1360718	Thymocyte	ETS1	Zacaras-Cabeza J- et al. J. Immunol. 2015
GSM1360719	Thymocyte	ETS1	Zacaras-Cabeza J- et al. J. Immunol. 2015
GSM1908613	P5424	ETS1	Cauchy P- et al. Nucleic Acids Res. 2015
GSM726992	T Lymphocyte; Blood	ETS1	Koch F- et al. Nat. Struct. Mol. Biol. 2011
GSM999186	T Lymphocyte; Blood	ETS1	Samstein RM- et al. Cell 2012
GSM1022319	Macrophage; Bone Marrow	JUNB	Ostuni R- et al. Cell 2013
GSM1288393	Dendritic Cell	JUNB	Vander Lugt B- et al. Nat. Immunol. 2014
GSM1309514	T Lymphocyte	JUNB	Kurachi M- et al. Nat. Immunol. 2014
GSM1467432	Cortical Neuron; Neuric	JUNB	Malik AN- et al. Nat. Neurosci. 2014
GSM978769	Th17; Spleen	JUNB	Li P- et al. Nature 2012
GSM578474	E14; Embryonic Stem Cell; Embryo	SMAD2	Unknown
GSM578475	E14; Embryonic Stem Cell; Embryo	SMAD2	Unknown
GSM578476	E14; Embryonic Stem Cell; Embryo	SMAD2	Unknown
GSM1335483	Embryonic Stem Cell; Embryo	SMAD3	Lee BK- et al. Cell Rep 2015
GSM539542	V6.5; Embryonic Stem Cell; Embryo	SMAD3	Mullen AC- et al. Cell 2011
GSM539545	38B9; Pre-B Lymphocyte; Blood	SMAD3	Mullen AC- et al. Cell 2011
GSM590100	Embryonic Stem Cell; Embryo	SMAD3	Mullen AC- et al. Cell 2011
GSM590102	Embryonic Stem Cell; Embryo	SMAD3	Mullen AC- et al. Cell 2011
GSM590103	Embryonic Stem Cell; Embryo	SMAD3	Mullen AC- et al. Cell 2011
GSM1213461	Progenitor B cell; Fetal Liver	SOX4	Mallampati S- et al. Blood 2014
GSM1468717	Pancreas	ATF3	Unknown
GSM1003774	CH12; Lymphoblastoid; Blood	ETS1	Beer MA- et al. Nature 2014
GSM1003777	MEL; Erythroid Progenitor Cell; Blood	ETS1	Beer MA- et al. Nature 2014
GSM777093	G1ME; Megakaryocyte; Bone Marrow	ETS1	Chlon TM- et al. Mol. Cell 2012
GSM999187	T Lymphocyte; Blood	ETS1	Samstein RM- et al. Cell 2012
GSM1370451	3T3-L1; Preadipocyte; Adipose	JUNB	SiersbR- et al. Cell Rep 2014
GSM1467437	Cortical Neuron; Neuric	JUNB	Malik AN- et al. Nat. Neurosci. 2014
GSM1288310	Embryonic Stem Cell; Embryo	SMAD2/3	Yoon SJ- et al. Nat Commun 2015
GSM1288315	Endoderm	SMAD2/3	Yoon SJ- et al. Nat Commun 2015
GSM590101	Embryonic Stem Cell; Embryo	SMAD3	Mullen AC- et al. Cell 2011
GSM1376735	Lung	SMAD4	Liu J- et al. Cell Rep 2015

8.4 Materials and Methods for Chapter 5

8.4.1 Model equations

The ordinary differential equations describing a gene regulatory network containing 19 genes as well as one input factor are given as

$$\frac{dy_k^a(t,\theta,u(t))}{dt} = R^a g\left(\sum_{b\in G}\beta^{ab}w^{ab}y^b + \beta^{au}w^{au}u(t) + h^a\right) - \lambda^a e_k^a y^a, \tag{8.3}$$

where g denotes the sigmoid regulation function (see Section 3.2.1). In the model equations, y^a denotes gene expression of gene a in the set of all genes G. The time-dependent input into the gene regulatory network is given by $u(t)$. θ denotes the parameter vector with production rate for gene a (R^a), the binary interaction parameter of gene $b \in G$ on gene a (β^{ab}), the regulatory weight of gene $b \in G$ on gene a (w^{ab}), the binary interaction parameter of gene a regulated by the external input (β^{au}), the regulatory weight for regulation of gene a by the external input (w^{au}), the promoter activity threshold (h^a), the degradation rate of gene a (λ^a), and the knock-down efficiency of gene a in cell-line k (e_k^a). Assuming steady state conditions at the time of knock-down R^a can be expressed in relation to other parameters of the system as

$$R^a = \frac{\lambda^a}{g(input)}, \tag{8.4}$$

with *input* denoting the total total regulatory input to gene a received by all other genes in the ODE system. When simulating knock-down conditions the degradation rate of the affected gene is multiplied by the corresponding knock-down efficiency e_k^a and simulations are carried out for 48h before the change in gene expression of the input function occurs. Binary interaction parameters as well as regulatory weights for excluded interactions are set to 0 (see Section 8.3.5). When simulating NMuMG data, two additional scaling factors are introduced for Snail1 and Hmga2 (s_{Snail1} and s_{Hmga2} respectively) due to differences in fold-changes between the wild-type time-course and the knock-down experiments. The expression time-course of the input factor $u(t)$ is defined by linear interpolation between measurements and assumed to be identical in all simulated cell-lines. Model equations are integrated using the Matlab implemented version of `CVode` [Serban and Hindmarsh, 2005] (Adams-Moulton Solver; Order 1-12) with a maximal relative error of 10^{-6} and an maximal absolute error of 10^{-8}.

8.4.2 Model fitting

Model parameters $\theta = (\beta, w, h, \lambda, e, s)$ are estimated by minimizing the regularized weighted least squares difference between modelled and measured gene expression values

$$V(\theta) = \sum_{i=1}^{n} \frac{(y^{model}(\theta) - y^{data})^2}{\sigma^2} + \alpha \frac{1}{\nu} \left(\sum_{a \in G, b \in G} \beta^{ab} \pi^{ab} + \sum_{a \in G} \beta^{au} \pi^{au} \right). \tag{8.5}$$

In the above equation, y^{model} is given by the integration of model equations, while y^{data} consists of measured mRNA expression at different time-points under different experimental conditions (see Section 8.3.2; 972 data-points excluding the Smad4 knock-down cell-line). Weights for each data-point (σ) are based on a linear error-model (see Section 8.3.2). The right part of Equation 8.5 described the penalization term in which less likely interactions are penalized during model fitting. The penalties (π) for each interaction were calculated from their respective ranks given in the community prediction integrating different network inference methods (see Section 8.3.5), by normalizing ranks between 0 and 1:

$$\pi^{ab} = \frac{(r^{ab} - 1)}{max(\mathbf{r}) - 1}. \tag{8.6}$$

The penalty over interactions is scaled by the total number of interactions present (ν). By increasing the regularization parameter α, the weight of the the penalty term can be increased relative to the model fit. Minimization of the cost function (Equation 8.5) was carried out using the enhanced Scatter Search (`eSS`) optimization method implemented in the `MEIGO` package [Egea et al., 2014]. The number of initial trial solutions was set to 5160, out of which 1290 initial solutions were supplied using parameters of direct TF-target regulations estimated from `NodeInspector` (see Section 8.3.5) with a randomly selected regulator for each of the 19 genes present in the network. The size of the `eSS` reference set was set to 74, while the time until end of optimization was set to 20h at which point structural model parameters were fixed and additional local optimization was carried out using using the direct hill climbing (`dhc`) optimization method provided by the `MEIGO` package. Ranges chosen for parameter estimation are given in Table 8.7.

Table 8.7: Ranges for parameter estimation. m corresponds to regulatory weights from the input factor to a given target gene, h describes the promoter activity threshold, k is the mRNA half-life related to the degradation rate λ as log(2) / λ, w corresponds to regulatory weights between regulator and target gene, e is the knock-down efficiency, s describes the scaling factor for Hmga2 or Snail1 while fitting to NMuMG data. Binary parameters β can only take on values of the set {0, 1}.

parameter	description	lower boundary	upper boundary
h	promoter threshold	-5	5
log(2) / k	degradation rate	0.1	48
w	regulatory weight	-20	20
e	knock-down efficiency	1	200
s	scaling factor	1	20

8.4.3 Model evaluation & model predictions

Significant enrichment of interactions derived from the literature-, motif-, and ChIP-analysis (see Section 8.3.7 and Section 8.3.6) was carried out using a hypergeometric test ($p < 0.05$). As in Section 8.3.7 and Section 8.3.6, only the appropriate background of interactions, evaluated both by the respective analysis and the model fitting approach was chosen.

Mean residuals per gene and data-point for each gene were clustered using a hierarchical clustering approach based on a cosine distance metric and complete linkage-criterion, considering only fits with $\alpha = 5000$. All clustering described here was carried out via the `python` clustering package `scipy.cluster.hierarchy`).

In order to simulating model fits after 24h, pSmad2 expression was assumed to be identical to its value at 24h. After 336h (14 days) of TGFβ-stimulation, pSmad2 expression was assumed to return to its basal level with a half-life of 30 min. Clustering of model predictions was based on time-course simulations after TGFβ-removal only. A hierarchical clustering approach was adopted using a cosine distance metric and complete-linkage criterion.

8.5 Materials and Methods for Chapter 6

8.5.1 Collection of embryos for proteome measurement and RNA-Seq

Population cages of wild-type Oregon R flies containing only fertilized females were maintained at 25°C. Embryos were collected on standard agar apple juice plates in 30 min laying time windows and processed immediately (0h time point) or aged at 25°C for the required period (1h, 2h, 3h, 4h, 5h, 6h, 8h, 10h, 12h, 14h, 16h, 18h, 20h). After collection, embryos were dechorionated using 7.5% hypochlorite for 2 min and rinsed with water. At this point, approximately 30% of the embryos (20 µl embryo pellets) were transferred to PBS buffer for lysis and mass spectrometry measurement. To check for correct and homogeneity of stages, approximately 10% of each sample was fixed and staged. The remaining samples were snap-frozen in liquid nitrogen and stored at -80°C. For proteome measurements, snap-frozen embryos in PBS were homogenized with a microtube pestle, cells were pelleted at 1000 g for 5 min at 4°C and resuspended in 1x LDS buffer complemented with 0.1M DTT. Samples were boiled for 10 min at 80°C and proteins were separated on a 4-12% NuPAGE Bis/Tris gel for 10 min at 180V in MOPS buffer. In-gel digestion and MS analysis was done as described [Bluhm et al., 2016]. Total RNA was extracted from approximately 20 µl embryo pellets with the RNeasy Mini Kit (Qiagen) and RNA integrity checked by Bioanalyzer.

8.5.2 Cell culture

Drosophila S2R+ cells were cultured at 25°C in Schneider's Drosophila Medium (GIBCO, Cat-No 21720) supplemented with 10% FBS and 2% Penicillin/Streptomycin. For knockdown experiments, dsRNA was synthesized overnight at 37°C using the Hi-Scribe T7 kit (NEB, Cat-No-E2040). dsRNA was transfected in S2R+ cells by serum starvation for 6h. The treatment was repeated twice and cells were harvested 5 days after the first treatment.

Table 8.8: Primer sequences to amplify the dsRNA template.

Hrb98DE dsRNA fwd	TAATACGACTCACTATAGGGACTACCGTACCACCGACGAG
Hrb98DE dsRNA rev	TAATACGACTCACTATAGCACCTCCCTGTTGGTCATTCTG

8.5.3 qRT-PCR

3 µg total RNA was transcribed into cDNA using MMLV reverse transcriptase (Promega). qRT-PCR analysis was performed using a ViiA7 real-time PCR system (Applied Biosystems). Measurements were done in triplicates and relative RNA levels were normalized to Rpl15 levels.

Table 8.9: qRT-PCR primer sequences.

Hrb98DE qPCR fwd	CAAGGAGGTGGTGGATTCAAAG
Hrb98De qPCR rev	CAATATCTGCGGTTGTTGCCAC
Rpl15 qPCR fwd	GCGCAATCCAATACGAGTTC
Rpl15 qPCR rev	AGGATGCACTTATGGCAAGC

8.5.4 Mass spectrometry measurement and label-free analysis

Peptides were separated by nanoflow liquid chromatography on an EASY-nLC 1000 system (Thermo) coupled to a Q Exactive Plus mass spectrometer (Thermo). Separation was achieved by a 25 cm capillary (New Objective) packed in-house with ReproSil-Pur C18-AQ 1.9 µm resin (Dr. Maisch). The column was mounted on an Easy Flex Nano Source and temperature controlled by a column oven (Sonation) at 40°C using SprayQC. Peptides were separated chromatographically by a 240 min gradient from 2% to 40% acetonitrile in 0.5% formic acid with a flow rate of 200 nl/min. Spray voltage was set between 2.4-2.6 kV. The instrument was operated in data-dependent mode performing a top10 MS/MS per MS full scan. Isotope patterns with unassigned and charge state 1 were excluded. MS scans were conducted with 70,000 and MS/MS scans with 17,500 resolution. The raw measurement files were analysed with MaxQuant 1.5.2.8 standard settings except LFQ quantitation [Cox et al., 2014] and match between runs option was activated as well as quantitation was performed on unique peptides only. The raw data was searched against the translated ENSEMBL transcript databases (release 79) of *D. melanogaster* (30,362 translated entries) and the *S. cerevisiae* protein database (6,692 entries). Known contaminants, protein groups only identified by site and reverse hits of the MaxQuant results were removed. A distribution calculated via the `logspline` `R` package of each replicate per time point as density function was used to impute the missing values. The mean of measured replicates or the average of two surrounding time points were used as a central value for the imputation distribution calculated using the `zoo` `R` package [Zeileis and Grothendieck, 2005]. In case the gap was bigger than a single time point, as well as single measurements with no surrounding values, they were replaced by a fixed small value of 22.5 in log2 scale.

8.5.5 Sequencing library preparation

NGS library prep was performed with Illuminas TruSeq stranded mRNA LT Sample Prep Kit following Illuminas standard protocol (Part #15031047 Rev. E). Libraries were prepared with a starting amount of 500 ng and amplified in 11 PCR cycles. Libraries were profiled in a High Sensitivity DNA on a 2100 Bioanalyzer (Agilent technologies) and quantified using the Qubit dsDNA HS Assay Kit, in a Qubit 2.0 Fluorometer (Life technologies). All 68 embryo samples were pooled in equimolar ratio and sequenced on 8 HiSeq 2500 lanes, SR for 1x 51 cycles plus 7 cycles for the index read. S2R+ samples were sequenced with a Mid Output kit using PE2x 79bp.

8.5.6 Analysis of RNA-Seq data

The RNA-Seq measurement of the embryo time-course yielded an average 18M reads per sample. Reads were mapped to the BDGP6 fly reference from Ensembl version 79 [Yates et al., 2016] using `STAR` [Dobin et al., 2013] version 2.4.0h, allowing up to 2 mismatches, a minimum intron length of 21, discarding reads mapping to more than 10 loci, and eventually keeping only the primary alignment. We assessed the quality of the sequenced reads with `FastQC` [Andrews, 2010], `dupRadar` [Sayols et al., 2016] and other in-house developed tools. We then counted reads per gene using `htseq-count` [Anders et al., 2015] from the `HTSeq` package with the default "union" mode and using the gene model provided by Ensembl for the same assembly version (BDGP6 version 79). We estimated the isoform abundance with the `Miso` package [Katz et al., 2010], as described in their pipeline.

In the S2R+ samples, we obtained an average of 31M paired reads per sample, assessing the quality using the same strategy described above. For these samples, we used `STAR` version 2.5.1b to align the reads to the BDGP6 fly reference from Ensembl version 90, using the same parameters as before. `FeatureCounts` [Liao et al., 2014] from the `Subread` package version 1.4.6-p2 was used in order to count reads per gene, also with the default "union" parameters and using the gene model provided by Ensembl for the same assembly version used for mapping (BDGP6 version 90).

8.5.7 Comparison of our RNA-Seq data with RNA-Seq data from Graveley et al. (2011)

Calculating the Spearman correlation between individual mRNA in our data and data by Graveley et al. (2011), we chose the set of 3125 genes present in both datasets. Samples were paired according to the lower limit of the measurement interval. To identify the extent of differences in developmental progress between data from Graveley et al. (2011) and our own data time-course expression of both mRNA time-courses was normalized by its mean and standard deviation (z-score), in order to focus only on time-course behaviour and avoid issues resulting from differences in absolute gene expression. Data from Graveley et al. (2011) was defined at $t_{Graveley} = \{0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22\}$, while our data was defined at $t_{Becker} = \{0, 1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 18, 20\}$. Expression data from Graveley et al. (2011) was compressed along its time axis by division of t with a constant factor a chosen between 0.5 and 1.5. By linear interpolation of the compressed time-course at time points corresponding to t_{Becker}, the ordinary least squares difference between our RNAseq data and compressed data from Graveley et al. (2011) was calculated over all genes.

8.5.8 Time-course clustering

From the mass spectrometry data, we selected 3761 proteins for further analysis based on their number of missing values (maximum 4 LFQ values below the detection limit out of 14 time points). Both mRNA and protein data was normalized by the respective value at

0h, log10 transformed, and scaled to minimum and maximum values between -1 and 1. A hierarchical clustering approach was applied to the processed data with imputed values: For this, pairwise distances between all mRNA-protein time-courses were calculated using the Euclidean distance. Clusters were merged using a complete-linkage criterion.

8.5.9 Correlation analysis

Global Spearman correlation between mRNA and protein samples was calculated using only significantly changing proteins, as identified in [Casas-Vila et al., 2017].

We calculated the Spearman correlation between individual mRNA and protein for all 3761 time-courses based on even time points only ($t = \{0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$). Shifts were introduced by matching mRNA at time t_n with protein at time t_{n+i} with $i \in \mathbb{Z} : -5 \leq i \leq 5$. Unmatched time points were left out, leading to a minimum number of 6 paired values for the assessment of correlation. mRNA/protein correlations with a p-value < 0.05 (two-sided) were chosen as significant.

8.5.10 Model fitting and evaluation

Before fitting, protein data was rescaled by its global mean value over all proteins to avoid numerical issues due to overly large LFQ values. By relating mean and standard deviation of the four biological replicates, a linear error-model was chosen for the protein data (slope: 0.169, intercept: 0.0). On each time interval between two subsequent measured time points the model variants (stationary, degradation, production, delayed-production) can be solved analytically if the input mRNA time-course on each interval is described using a linear function $u(t) = mt + b$. The explicit solution for the delay model

$$\frac{dy(t, \theta, u)}{} = \alpha h(t - \tau)(mt + b) - \lambda y \tag{8.7}$$

with initial value $y(t = 0) = y0$ then becomes

$$y(t) = \alpha h(t - \tau)(b\lambda^{-1} - m\lambda^{-2} + mt\lambda^{-1} + ce^{\lambda t}). \tag{8.8}$$

In the above equation y denotes measured protein, u measured mRNA, α and λ the production and degradation rate respectively. The heavy-side function h is used to suppress the production term in the delay model with τ being the time delay.

Model parameters were estimated by minimizing the weighted least-squares distance between modelled values (Equation 8.9) and non-imputed protein data using a trust region reflective optimization scheme (`scipy.optimize.least_squares` in `python`). Missing protein values (or their imputed counterparts) were not considered in the cost function.

$$\chi^2(\theta) = \sum_{i=1}^{n} \frac{(y_i^{model}(\theta) - y_i^{data})^2}{\sigma^2}. \tag{8.9}$$

Here n denotes the number of data-points, y modelled or measured protein expression, $\theta = \{y_0, \alpha, \lambda, \tau\}$ the parameter vector, and σ the weight for each data-point i. Ranges for the parameters are given in Table 8.10. In all cases, multi-start local optimization was carried out via latin-hypercube sampling and parameters were sampled on a logarithmic scale [Raue et al., 2013]. Depending on the model complexity, a different number of initial parameter samples was chosen (degradation model: 5, production model: 5, delayed-production model: 25).

Table 8.10: Ranges chosen for parameter estimation to classify paired mRNA and protein time-courses. Shown are parameter ranges selected while estimating model parameters for models of post-transcriptional regulation.

parameter	description	lower boundary	upper boundary
y_0	initial protein value	1.0e-01	5.0e03
λ	protein degradation	ln(2) / 1.0e03	ln(2) / 1.0e-01
α	translation rate	1.0e-05	1.0e-01
τ	translational delay	1.0e-05	1.0e01

Model rejection was carried out by applying a χ^2-test to the weighted squared residuals between model and data [Cedersund and Roll, 2009] with $n - m$ degrees of freedom (n: number of data-points, m: number of model parameters) and $\alpha = 0.95$ (one-tailed). To further exclude strong systematic deviations between model and data, the residuals were subjected to a Durbin-Watson-test with the number of data-points n and m degrees of freedom ($\alpha = 0.95$). The resulting Durbin-Watson value of the fit was evaluated against the lower limit (dL) of the test statistic. For the stationary model, a value of $m = 2$ was chosen, since Durbin-Watson significance tables do not cover the case of 1 regressor only. Only if a model fit passed both χ^2- and Durbin-Watson-tests, a model variant was considered possible for the mRNA-protein pair under consideration. If multiple models remain possible for a given mRNA-protein pair after testing, model selection was carried out using a stepwise likelihood-ratio test with $\alpha = 0.95$ (one-tailed). For both the stationary and the degradation model, correction criteria were applied according to the procedure described in Section 9.5.4. For all estimated parameters 95% confidence intervals are calculated using a profile likelihood approach [Raue et al., 2009].

8.5.11 Analysis of mRNA and protein correlation dependent on protein half-life and protein steady-state

Proteins with an estimated upper confidence interval limit of the half-life below the median value of all protein half-lives were selected as proteins with short half-lives (313 proteins). Inversely, the lower confidence interval limit needed to be above the median estimated half-life for a protein to be classified as having a long half-life (237 proteins).

Theoretical protein steady-states were calculated based on estimated parameter values as

$$y_t^{ss} = \alpha \frac{u(t)}{\lambda} \tag{8.10}$$

where α denotes the estimated protein production rate, λ the estimated protein degradation rate and $u(t)$ the measured mRNA abundance either at $t = 0$ or $t = 20$. A protein was considered in steady-state at $t = 0$ if its measured concentration at this time point was within 20% of the estimated steady state as $t = 0$. This resulted in 481 proteins to be classified as in steady state, while a combination of this set with proteins with short half-lives consisted of 95 proteins. Correlation was calculated as in Section 8.5.9 with no time-shift considered between mRNA and protein.

8.5.12 Enrichment analysis

GO term, KEGG and miRNA enrichment for protein groups lists was carried out using `GeneTrail1` [Backes et al., 2007]. As background the list of 3761 reproducibly measured proteins was chosen. Results were filtered using a corrected p-value of < 0.05 (Benjamini & Hochberg correction). Plotting was done using Multi-Dimensional Scaling implemented in the `sklearn` package provided for `python` based on a distance metric obtained using semantic similarity between GO terms. Semantic similarities were retrieved from the `bioconductor` package `GOSemSim` [Yu et al., 2010] using the Wang distance metric. 29 Proteins with the following GO terms have been selected for Figure 9.24A: 'glucose catabolic process (GO:0006007)', 'hexose catabolic process (GO:0019320)', 'alcohol catabolic process (GO:0046164)', 'monosaccharide catabolic process (GO:0046365)', 'glucose metabolic process (GO:0006006)', 'cellular carbohydrate catabolic process (GO:0044275)', 'glycolysis (GO:0006096)', 'carbohydrate catabolic process (GO:0016052)', 'hexose metabolic process (GO:0019318)', 'monosaccharide metabolic process (GO:0005996)', 'alcohol metabolic process (GO:0006066)'. Proteins (65) assigned the following GO terms are plotted in Figure 9.24B: 'nuclear division (GO:0000280)', 'M phase of mitotic cell cycle (GO:0000087)', 'mitosis (GO:0007067)', 'mitotic cell cycle (GO:0000278)'.

8.5.13 Motif analysis

Motif analysis was carried out using `AME` [McLeay and Bailey, 2010b]. Input sequences were defined as one of the following: the common transcript sequence between all mapped transcripts for each protein, the longest transcript sequence, longest 3'UTR sequence, longest 5'UTR sequence, or longest coding sequence. The motif database was taken from Ray et al., which included 67 motifs mapping to 51 RBPs [Ray et al., 2013]. Using `AME`, sequence motif scoring method was based on total hits. Motif match threshold was 0.0002 (p-value). Background was estimated from the sequences of the 3761 reproducibly measured proteins. As control all sequences not in the group were chosen. Statistical test

for enrichment/association was Fisher's exact. The threshold values for reported results was set to an adjusted p-value of 0.05 (Benjamini & Hochberg correction). Individual motif occurrences were identified using `FIMO` [Grant et al., 2011] with a p-value threshold of 0.0002 for reported motifs. Multiple motif copies in the motif database were grouped by calculating the union over all transcripts.

8.5.14 Statistics of Hrb98DE knock-down

Protein with NaN values in any of the tested conditions were filtered out before differential enrichment was assessed using an independent t-test between Hrb98DE knock-down and LacZ control. As thresholds for differentially expressed protein a p-value < 0.01 (two-tailed, n=4) and fold-change cut-off $> 50\%$ was chosen. In the RNAseq data, non-expressed mRNA were removed based on a mean RPKM over all samples > 1. We obtained mRNA statistics using `DESeq2` [Love et al., 2014] version 1.14.1, using the default Wald test to calculate the significance and applying automatic independent filtering to avoid testing genes which were poor candidates of being differentially expressed (maximizes the number of adjusted p-values less than $\alpha = 0.1$). Differentially expressed mRNA were identified using a threshold of < 0.05 for BH corrected p-values (n=3) and a fold-change cut-off of $> 30\%$ change. To identify differentially spliced genes we used `DEXSeq` [Anders et al., 2012] in the following way: we recounted reads on exons using the default `htseq union` mode; not discarding reads spanning multiple exons; taking into account data is paired-end but without discarding singletons. Finally, we considered that genes could show a differential splicing pattern if they included exon abundance changes of at least 50% between conditions at FDR 1%. `Dupradar`, `DESeq2` and `DEXSeq` are part of the `Bioconductor` [Huber et al., 2015] project. Genes are considered targeted by Hrb98DE if identified as such by the study of Ji and Tulin (2009) [Ji and Tulin, 2016], the study of Blanchette et al. (2012) [Blanchette et al., 2009] or containing a Hbr98DE binding motif (Section 8.5.13). Target genes identified in Ji and Tulin (2009) were kindly provided by Alexei Tulin. Target genes according to Blanchette et al. (2012) were determined re-analysing the raw data according to the methods described in [Blanchette et al., 2009], and using a fold-change cut-off of log2 > 10 between immunoprecipitation and control samples.

Chapter 9

Supplemental Information

9.1 Supplemental Information for Chapter 2

9.1.1 Circadian genes detected by all four methods

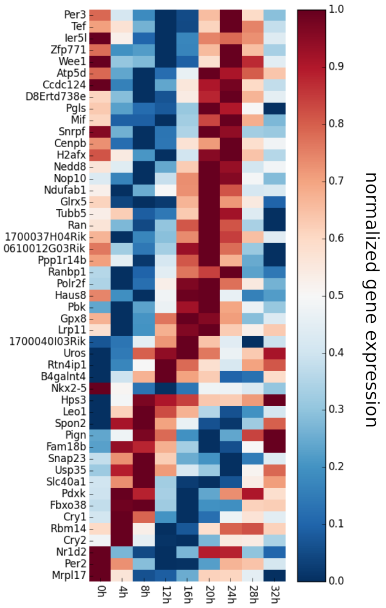


Figure 9.1: Circadian genes detected by all four methods. Heatmap of cyclical expressed genes detected by all four methods after minimum-maximum normalization (blue - lowest expression, red - highest expression). Genes are sorted by the phase given by `JTK_CYCLE`.

9.1.2 Comparison of circadian genes found in this study with genes identified in Menger et al. (2007) and Hughes et al. (2009)

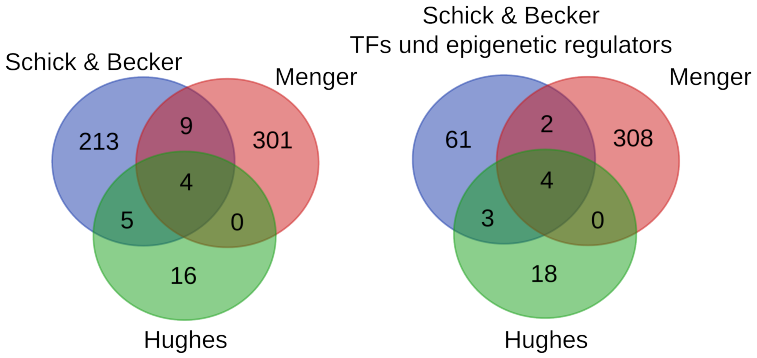


Figure 9.2: Comparison of circadian genes with genes identified in Menger et al. (2007) and Hughes et al. (2009). Venn diagrams comparing cyclically expressed genes found in NIH3T3 cells in Menger et al. (2007), Hughes et al. (2009) and this study (left: genes detected by three out of four applied methods, right: transcription factors and epigenetic regulators (TF and EpiR)).

9.1.3 Number of circadian genes cyclical expressed in other tissue types or bound by core clock genes

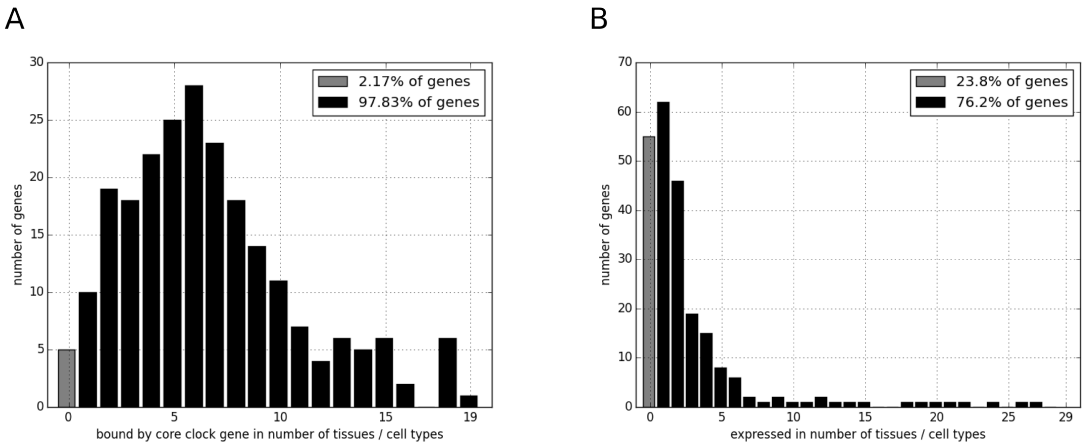


Figure 9.3: Number of circadian genes cyclical expressed in other tissue types or bound by core clock genes. (A) Barplot showing the number of genes identified by three out of four methods found to be bound by at least one core clock factor (black) compared to genes not bound (grey). **(B)** Barplot showing the number of genes identified by three out of four methods found to be cyclically expressed in different cell or tissue types. Black bars indicate genes cyclically expressed in at least one cell or tissue type.

9.1.4 Binding of core clock factors to circadian expressed genes

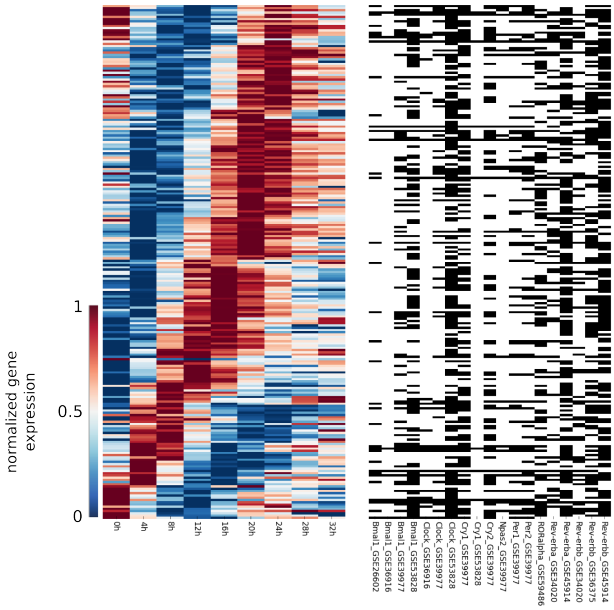


Figure 9.4: Binding of core clock factors to circadian expressed genes. Heatmap of cyclically expressed genes detected by three out of four methods after minimum-maximum normalization. Columns on the right indicate binding of a core clock factor in the promoter (-800, +200 around TSS) of the respective gene.

9.1.5 Unequal distribution of selected GO terms among circadian phase

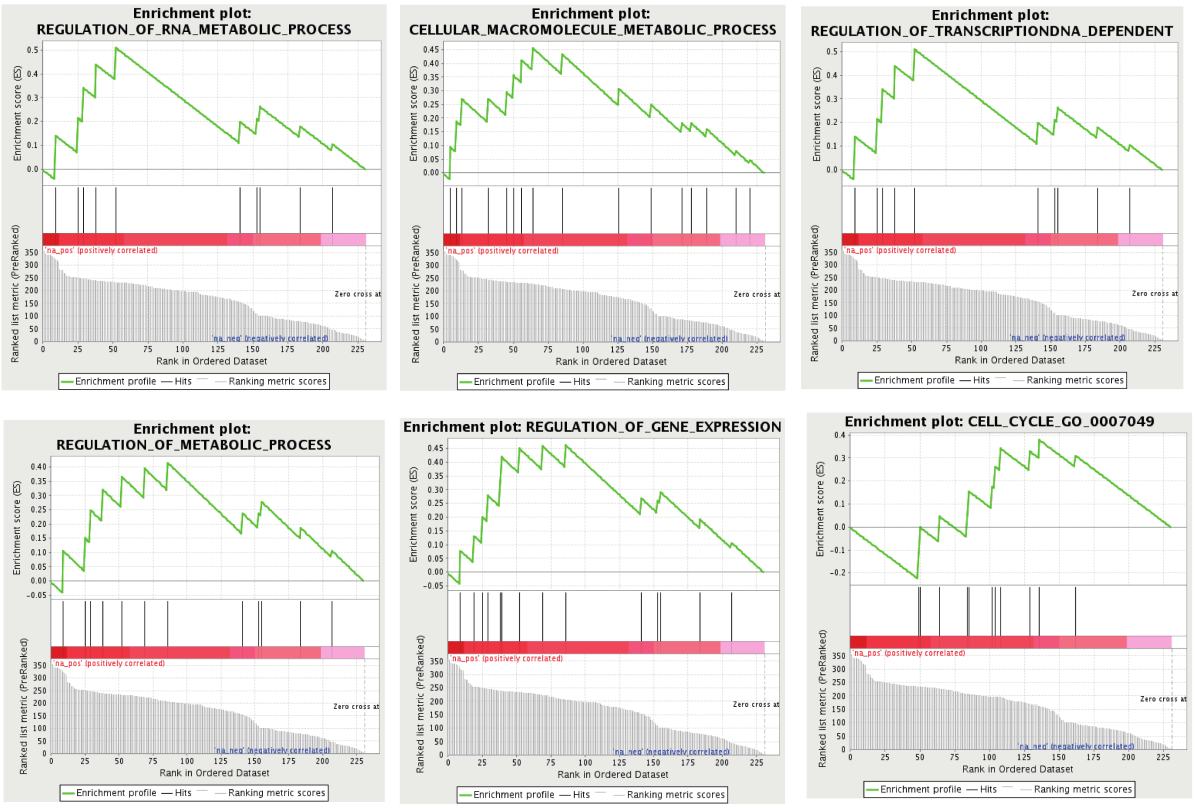


Figure 9.5: Distribution of selected GO terms among circadian phase. Examples of Gene Set Enrichment Analysis results. Genes are sorted by the phase determined by JTK_CYCLE, with largest phase value on the left.

9.1.6 Validation of cyclical expression of Leo1 and Zfp28 by qPCR

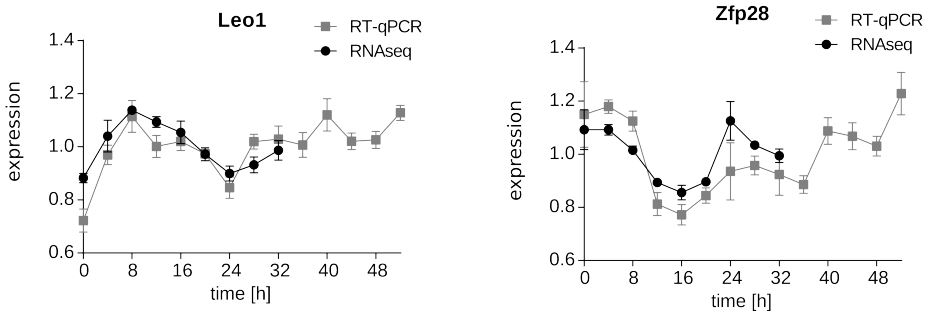


Figure 9.6: Validation of cyclical expression of Leo1 and Zfp28 by qPCR. Expression of Leo1 (left) and Zfp28 (right) determined by RNA-seq (n=2, normalized read counts normalized to average expression, error bars represent s.e.m.) and RT-qPCR (n=4, relative expression normalized to Hsp90ab1 and to the average expression: $\triangle\triangle$CT, error bars represent s.e.m.).

9.1.7 Expression of luciferase in NIH 3T3 Bmal1:luc cells after knock-down of core clock factors

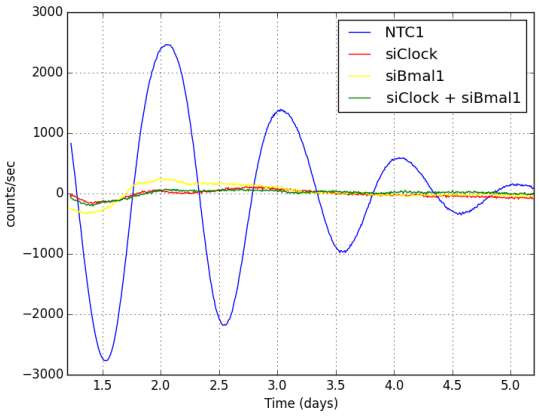


Figure 9.7: Expression of luciferase in NIH 3T3 Bmal1:luc cells after knock-down of core clock factors. Luminescence measurement of NIH3T3 Bmal1:luc cells with knock-down of Clock (siClock), Bmal1 (siBmal1), or Clock + Bmal1 (siClock + siBmal1), respectively. Data is background subtracted.

9.1.8 Identification of circadian expressed lincRNA by four different methods

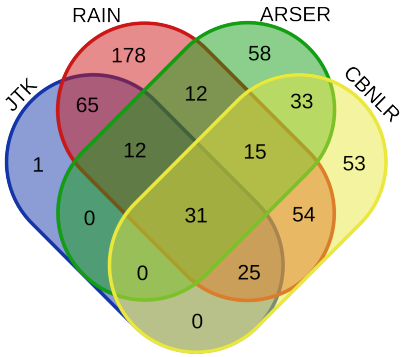


Figure 9.8: Identification of circadian expressed lincRNA by four different methods. Venn diagram showing cyclical expressed long intergenic non-coding RNAs (lincRNAs) detected by `JTK_CYCLE`, `RAIN`, `ARSER`, and the `CBNLR` approach.

9.2 Supplemental Information for Chapter 3

9.2.1 Testing for correctly predicted signs of interactions

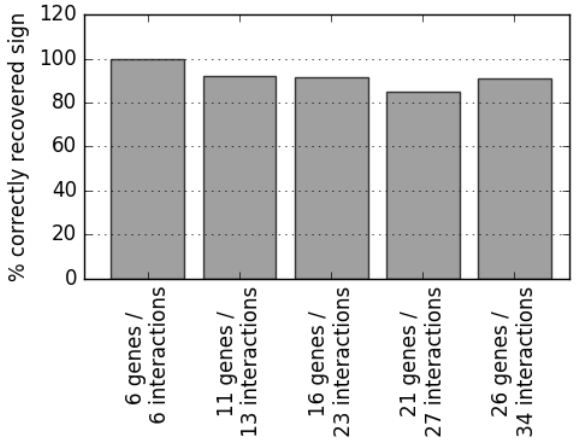


Figure 9.9: Performance of `NodeInspector` in recovering the sign of interactions. Shown is the fraction of correctly recovered signs (activation or inhibition) of true interactions in a benchmark as predicted by `NodeInspector`. The size of the benchmark networks increase from left to right.

9.2.2 Evaluation of formulated regulator-target interactions between circadian expressed genes in NIH 3T3 data

Table 9.1: Evaluation of regulatory models describing gene expression of Leo1 and Zfp28. Table shows χ^2- and p-values of regulatory models considering Leo1 or Zfp28 as the affected target genes. P-values are given under the null hypothesis that the respective model is true with interactions being rejected if the p-value lies above 0.95.

target	regulator	χ^2	p-value
Leo1	Cry1	7.849	0.404
Leo1	Arntl	55.778	1.000
Leo1	Cry2	130.807	1.000
Leo1	Nr1d1	33.665	0.999
Leo1	Nr1d2	17.981	0.950
Zfp28	Nr1d1	14.579	0.943
Zfp28	Cry1	13.349	0.864
Zfp28	Nr1d2	7.803	0.788
Zfp28	Cry2	28.139	0.738
Zfp28	Arntl	7.469	0.628

9.2.3 Measured knock-down efficiencies in cells treated with siRNA

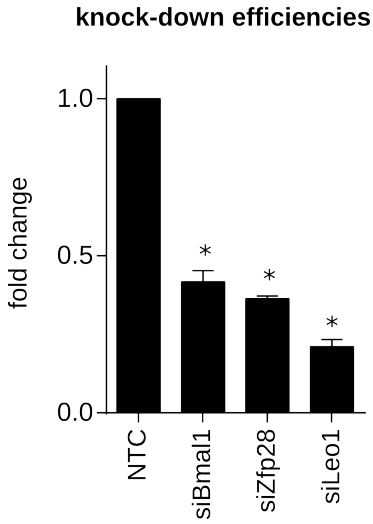


Figure 9.10: Efficiency of mRNA knock-down in NIH 3T3 cells by siRNA. Bar graphs showing the expression of the respective target gene after knock-down with siRNA in NIH3T3 Per2:luc cells measured by RT-qPCR. The expression is normalized to Hsp90ab1 as well as to the expression in the control cells (NTC) ($\Delta\Delta$CT, n=4, error bars represent s.e.m.). Significance was assessed using a Mann-Whitney U-test (*: $p < 0.05$).

9.3 Supplemental Information for Chapter 4

9.3.1 Expression of selected EMT factors in NMuMG RNA-Seq data

Table 9.2: Summary of selected EMT factors Listed are EMT factors selected for further analysis and their expression in RNA-Seq data obtained in NMuMG cells at 0h and 24h relative to TGFβ-treatment.

GeneID	mean RPKM 0h	mean RPKM 24h	fold-change
Atf3	239.41	1099.83	4.59
Ets1	1370.25	3069.24	2.24
Hmga2	221.81	986.16	4.45
Id1	1143.57	539.77	0.47
Id2	2278.65	227.70	0.10
Id3	251.59	142.14	0.56
Junb	985.96	4537.79	4.60
Serpine1	3383.52	45509.22	13.45
Smad2	873.25	975.12	1.12
Smad3	1499.77	1530.69	1.02
Smad4	1163.61	1078.22	0.93
Snai1	9.83	65.63	6.68
Sox4	2163.39	5763.46	2.66
Zeb1	419.07	870.24	2.08
Zeb2	143.59	791.64	5.51

9.3.2 Selected publications reporting experimental evidence of regulator-target interactions between EMT genes in NMuMG cells

Table 9.3: Selected publications reporting experimental evidence of regulator-target interactions between EMT genes in NMuMG cells. Shown is a list of publications consulted to identify interactions between EMT related genes.

Publication	Reference number
[Bakin et al., 2005]	3
[Peinado et al., 2003]	8
[Gervasi et al., 2012]	12
[Kondo et al., 2004]	14
[Korpal et al., 2008]	15
[Kowanetz et al., 2004]	16
[Shirakihara et al., 2007]	17
[Thuault et al., 2006]	18
[Tan et al., 2014]	19
[Richards et al., 2015]	20
[Thuault et al., 2008]	24
[Tan et al., 2012]	25
[Dave et al., 2011]	26

9.3.3 Interactions between selected EMT genes identified based on published experimental data

Table 9.4: Interactions between selected EMT genes identified based on published experimental data. Table shows interactions identified in the EMT system based on experiments reported in the literature. Columns show regulators, while target genes are given as rows. Numbers in each cell indicate the reference in which experimental evidence supporting the interaction is published (see Table 9.3). Orange coloured background corresponds to interactions supported by grade B evidence, while interactions supported by grade A evidence are marked in green. Rows denoted 'Motif' and 'ChIP' indicate whether the respective regulator given in the column was included in motif or ChIP analysis.

	Atf3	Ets1	Hmga2	Id1	Id2	Id3	JunB	Srp1	Snail1	Sox4	Zeb1	Zeb2	pSmad2
Atf3							12						3
Ets1													
Hmga2													12, 18
Id1													
Id2	12		18				12						12
Id3													
JunB													12
Srp1													
Snai1			18,24,25										12,24
Sox4													
Zeb1		17	24		17				24,26				
Zeb2		17	24						24	27			
Cdh1		17	18,19	14	14,16					27	17	17	
Cdh2			18							27			
Crb3													
Fn1			18				12			27			
Ncam													
Ocln													
Vim			18										
Motif	TRUE	TRUE			TRUE		TRUE		TRUE	TRUE	TRUE		TRUE
ChIP	TRUE	TRUE	TRUE				TRUE			TRUE			TRUE

Listed below is a short summary of experiments reported in the literature from which the specific interactions given in Table 9.4 have been identified. Reference numbers correspond to those given in Table 9.3.

Atf3: In NMuMG cells Atf3 expression does not appear to change at the level of mRNA within the first 24h after TGFβ-treatment. However, there is evidence that Atf3 is post-trancriptionally regulated and therefore protein abundance increases shortly after TGFb treatment [3]. As CHX is able to block Atf3 protein up-regulation by TGFβ it appears that up-regulation of protein can not be a direct effect of Smad phosphorylation and there need to be additional factors regulating the translation of this gene. One of the factors which might be involved in this function is JunB, as JunB knock-down blocks induction of Atf3 protein by TGFβ [12].

Cdh1: Although Cdh1 mRNA expression hardly changes within the first 24h after TGFb treatment, there have been several regulators proposed: Cdh1 is directly regulated by Zeb1 and Zeb2 [17]. Also indirect evidence of regulation comes from Id1 and Id2 [14,16], Hmga2 (by promoter methylation) [18], Ets1 [17], E47 and E12 [14,17], the microRNA-family miR200 [15], and the lincRNA HIT [20]. Cdh1 expression in NMuMG cells was not influenced by knock-down of JunB [12] or Snail1 [17].

Cdh2: So far no clear evidence has been reported for the regulation of Cdh2. In NMuMG cells Cdh2 protein expression is dependent on Hmga2 [18] but not on Zeb1, Zeb2 or miR-200 [17]. Possibly Cdh2 is also a direct or indirect target of YB-1 and/or Snail1, as ectopic expression of YB-1 and knock-down of YB-1 or Snail1 affected protein expression levels of Cdh2.

Fn1: As in the case of Cdh2, no direct activator or repressor of Fn1 has been found either in NMuMG or other cell lines. However, there exists evidence of Fn1 being either directly or indirectly regulated by JunB [12], Smad3 [12] and Hmga2 [18]. Knock-down of either Zeb1 and Zeb2 or Snail1 did not have an affect on Fn1 expression in NMuMG cells [17].

Hmga2: In NMuMG cells there is evidence for Smad4 binding to the Hmga2 promoter [18]. Also knock-down of Smad4, but not Smad2 or Smad3 resulted in mis-expression of Hmga2 [18,12]. Smad2 phosphorylation is critical for Hmga2 up-regulation [18]. In NMuMG cells Hmga2 expression was not influenced by JunB [12].

Id2: Most likely Id2 is directly regulated by JunB together with Aft3, possibly in cooperation with Smad4, which is dependent on the AP1 binding site within the (-1570/-1703) region of the Id2 promoter [12]. Hmga2 also had an indirect effect on Id2 expression [18], possibly via one of the above TFs.

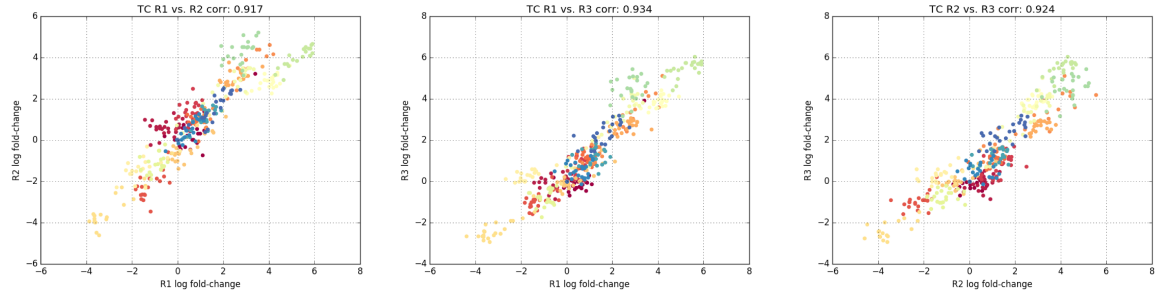
Snail1: In NMuMG cells Snail1 is regulated directly by the Smad pathway [12,24] possibly in cooperation with Hmga2 [18,24,25]. Twist might also pose an additional regulatory input into Snail1 expression [25,26]. Snail1 is however independent of lincRNA HIT [20] or JunB [12]. In MDCK cells there is evidence that Snail1 up-regulation by TGFβ is Smad4 independent [8].

Zeb1: As Zeb1 up-regulation is dependent on *de-novo* protein synthesis, its is likely that Zeb1 regulation is an indirect target of Smads or possibly co-regulators are involved [17]. Zeb1 mRNA and protein expression is regulated at multiple levels: mRNA stability of Zeb1 is regulated by miR-200 [15]. Also a yet unknown factor might contribute to Zeb1 protein stability [26]. Although there is no direct evidence for regulation of Zeb1 by Ets1, it is possible that Ets1 and Id2 are direct regulators of Zeb1 [17]. There exists also evidence of indirect regulation of Zeb1 by Snail1 [24, 26], Hmga2 [24], and Twist [25,26]. No effect on Zeb1 was seen in over-expression experiments with lincRNA HIT [20].

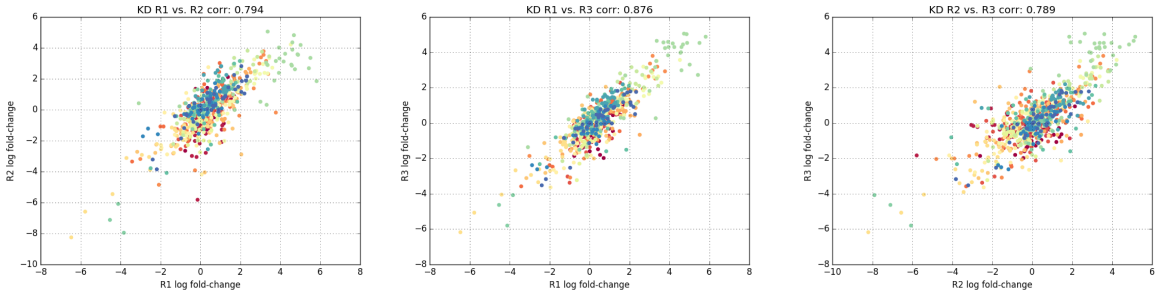
Zeb2: As Zeb2 up-regulation is dependent on *de-novo* protein synthesis, its is likely that Zeb2 regulation is an indirect target of Smads or possibly co-regulators are involved [17]. Zeb2 mRNA stability can be regulated by miR-200 [15]. Although there is no direct evidence for regulation of Zeb2 by Ets1, it is possible that Ets1 and Id2 are direct regulators of Zeb2 [17]. There exists also evidence of indirect regulation of Zeb2 by Snail1 [24, 26], Hmga2 [24], and Twist [25,26].

9.3.4 Correlation between gene expression measurements in biological replicates

Time-course data:



Knock-down data:



Combined data:

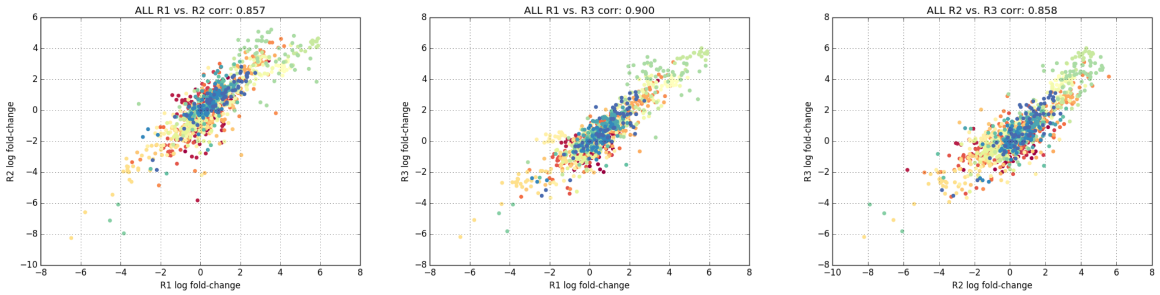


Figure 9.11: Correlation between replicates of NMuMG data. Scatter plots showing correlation of the different biological replicates either for time-course data (top), knock-down data (middle), or both data combined (bottom). Pearson's correlation coefficients are given above each subplot.

9.3.5 AUROC and AUPR values of network inference predictions obtained from benchmark data

Table 9.5: AUROC and AUPR values of network inference predictions obtained from benchmark data. Shown are performance values (AUROC and AUPR) of the different network inference methods applied to either knock-down (KD) or time-course (TC) gene expression data simulated by the benchmark network.

Method	Data	AUROC	AUPR
Random	–	0.50	0.11
Genie3	KD	0.80	0.49
Inferelator	KD	0.86	0.49
PREMER	KD	0.64	0.45
ARACNE	KD	0.66	0.37
CLR	KD	0.76	0.40
MRNET	KD	0.74	0.35
MRNETb	KD	0.75	0.38
NodeInspector	TC+KD	0.69	0.29
Genie3	TC	0.68	0.19
Inferelator	TC	0.65	0.18
PREMER	TC	0.59	0.16
Community	TC+KD	0.82	0.43

9.3.6 Example of limited temporal resolution of the knock-down experiment

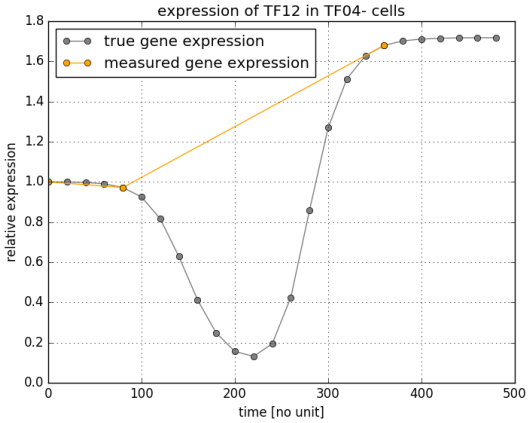


Figure 9.12: Example of inadequate time-course sampling. Shown is the time-course expression of TF12 simulated under knock-down of TF04. Gray indicated are dynamics of gene regulation as measured at an interval of 20 time-units, while orange shows gene expression at 0, 80, and 360 time-units.

Apart from conceptual problems facing all network inference methods alike, each network inference methods displays its own strengths and weaknesses. NodeInspector, for example, although applied to knock-down and time-course data simultaneously, shows only mediocre performance on the benchmark data of 4. This is possibly due to low

temporal resolution of the knock-down data, where gene expression is measured only at 0h, 4h, and 18h after TGFb stimulation. As can be seen in Figure 9.12, where gene expression for TF12 in TF04- cells of the benchmark data is shown, a linear interpolation between only few data-points does not satisfy the complex non-linear dynamics of gene expression in the network. As a consequence, the dynamics of transiently expressed genes will be affected by insufficient temporal resolution. Such missing information may not only affect the performance of NodeInspector, but also that of other network inference methods.

9.3.7 Similarity of network inference predictions

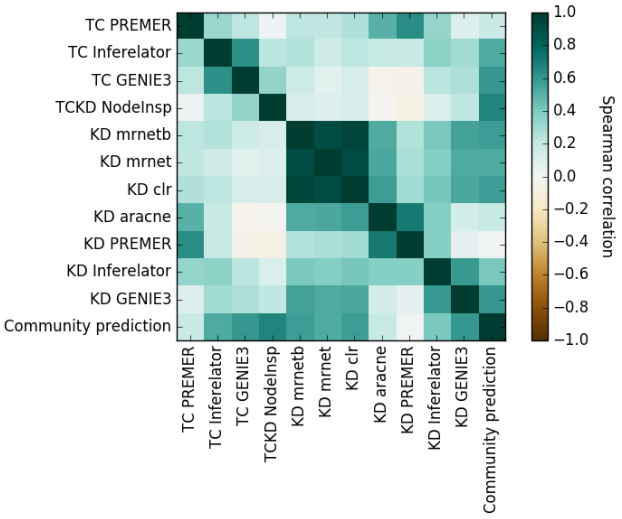


Figure 9.13: Similarity of network inference predictions. Shown are Spearman rank correlations between the different network predictions based on each applied combination of network inference method and either time-course or knock-down data.

The general procedure implemented by each of the network inference might differ greatly. Some methods however show a greater similarity, since their rankings are based largely on estimates of mutual information. If prediction between some methods agree heavily this might introduce a bias towards these methods into the community prediction. We therefore checked the similarity of the different prediction by calculating the Spearman correlation coefficient between predictions, including the community prediction. The highest similarity between predictions is observed between MRNET, MRNETb and CLR applied to the knock-down data Figure 9.13. Strikingly, all of these methods rely on mutual information to assess the probability of interactions. When comparing the similarity of individual predictions with the community prediction, a higher degree of similarity of the named predictions with the community prediction can be observed. This potentially introduces a bias towards highly similar predictions in the community while other predictions are under-represented. In principle an appropriate weighting scheme could alleviate this problem of bias and guarantee a more even contribution of all sources to the community prediction.

9.3.8 Top 25 interactions predicted for the EMT network by network inference

Table 9.6: Top 25 interactions predicted for the EMT network by network inference Table lists the top 25 ranked interactions predicted for the gene regulatory network controlling EMT. The prediction is obtained by calculating the mean rank of each interaction across the various network inference approaches (Community prediction). Additional columns indicate whether evidence for an interaction was found in the literature (grade A, grade B), by motif analysis (TRUE, FALSE), or by analysis of the ChIP data (TRUE, FALSE). Empty cells in the table correspond to interactions which have not been evaluated by one of the three mentioned analysis.

Rank	TF	Target	Community prediction	Literature evidence	Motif evidence	ChIP evidence
1	Sox4	Ncam	14.1		FALSE	FALSE
2	Sox4	Ocln	29.4		FALSE	FALSE
3	Serpine1	Cdh2	31.8			
4	Zeb2	Ets1	32.5			
5	Serpine1	Zeb1	33.4			
6	Zeb2	Fn1	33.8			
7	pSmad2	Atf3	33.9	grade B	TRUE	FALSE
8	Sox4	Id2	34.0		FALSE	FALSE
9	Id3	Id1	34.4			
10	pSmad2	JunB	34.5	grade B	FALSE	TRUE
11	Zeb1	Crb3	39.7		TRUE	
12	Serpine1	Id2	40.7			
13	Snail1	JunB	41.3		TRUE	
14	Ets1	Zeb2	43.3	grade B	TRUE	FALSE
15	Sox4	Cdh2	44.8	grade B	TRUE	FALSE
16	Serpine1	Hmga2	45.4			
17	Zeb2	Hmga2	45.6			
18	JunB	Ets1	46.0		TRUE	FALSE
19	Serpine1	JunB	47.0			
20	JunB	Snail1	47.3		FALSE	FALSE
21	Id1	Id3	47.8			
22	Hmga2	Zeb2	47.9	grade B		FALSE
23	Zeb1	Serpine1	48.3		FALSE	
24	Serpine1	Ncam	48.5			
25	Hmga2	JunB	50.3			FALSE

9.4 Supplemental Information for Chapter 5

9.4.1 Correlation between benchmark data and modelled gene expression values

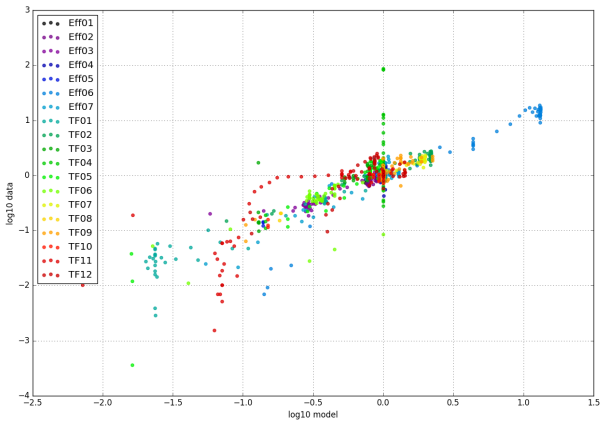


Figure 9.14: Scatter plot of the correlation between simulated gene expression and benchmark data. Scatter plot comparing gene expression fold-changes originating from the benchmark network and modelled gene expression fold-changes obtained by the best model fit.

9.4.2 Distribution of interaction penalty values for interactions the benchmark network

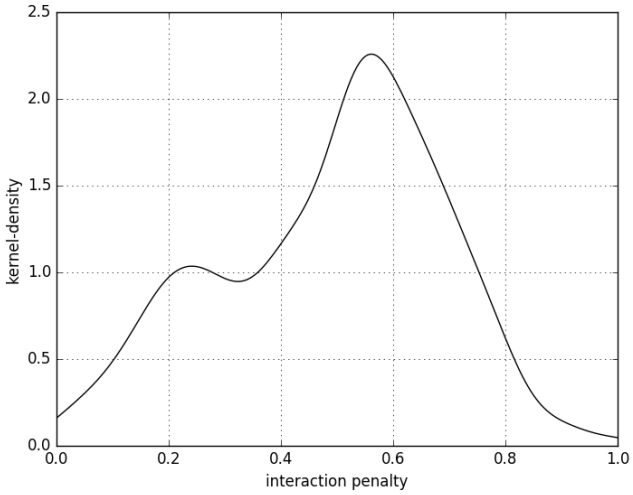


Figure 9.15: Distribution of individual penalty values for interactions in the benchmark network. Shown is the estimated density of interaction penalty values for interactions in the benchmark network as derived from the community prediction. The density function is estimated using a Gaussian kernel density estimator with covariance factor 0.25.

9.4.3 Performance evaluation of regularized network inference approach

Table 9.7: Summary of results obtained by regularized model fitting on the benchmark data. Shown are results of the regularized model fitting approach applied to the benchmark data with different values of the regularization parameter α. Each column presents the mean value over 20 fits per α: Weighted least squares (WLS), number of interactions (n), Pearson correlation between modelled and measured gene expression, interaction penalty, precision of the regularized model fitting approach, recall of the regularized model fitting approach, precision of the community prediction, recall of the community prediction, change in performance regularized model fitting vs community prediction.

					Model Fitting		Network Inference		
α	WLS (χ^2)	n	ρ	Penalty (Γ)	precision	recall	precision	recall	increase
0	1720.77	30.05	0.87	0.36	0.37	0.43	0.38	0.44	-02%
1250	2030.41	20.80	0.87	0.20	0.53	0.42	0.39	0.31	38%
2500	2244.28	19.45	0.87	0.18	0.57	0.42	0.39	0.29	47%
3250	2439.10	19.40	0.87	0.17	0.55	0.41	0.38	0.28	45%
5000	2740.85	16.80	0.86	0.14	0.58	0.37	0.41	0.26	41%
6250	3026.22	15.25	0.85	0.13	0.62	0.35	0.43	0.25	42%
7500	3146.46	14.80	0.86	0.13	0.62	0.35	0.45	0.25	37%
8750	3208.69	13.80	0.86	0.12	0.67	0.35	0.45	0.24	49%
10000	3388.09	13.40	0.85	0.11	0.63	0.32	0.46	0.23	38%
11250	3576.65	13.20	0.85	0.11	0.62	0.31	0.47	0.23	34%
12500	3661.07	12.70	0.85	0.11	0.66	0.32	0.46	0.22	44%
13750	3811.04	11.90	0.85	0.11	0.70	0.32	0.48	0.22	45%
15000	3885.65	11.10	0.85	0.10	0.72	0.30	0.47	0.20	53%
16250	4084.30	11.50	0.85	0.10	0.69	0.30	0.48	0.21	45%
17500	4228.13	11.00	0.85	0.10	0.72	0.30	0.48	0.20	48%
18750	4198.73	11.30	0.85	0.10	0.69	0.30	0.48	0.21	46%
20000	4479.37	10.70	0.86	0.10	0.70	0.29	0.48	0.20	47%
Mean	3286.46	15.13	0.86	0.14	0.63	0.35	0.44	0.25	41%

9.4.4 Comparison of NMuMG gene expression data and modelled gene expression values

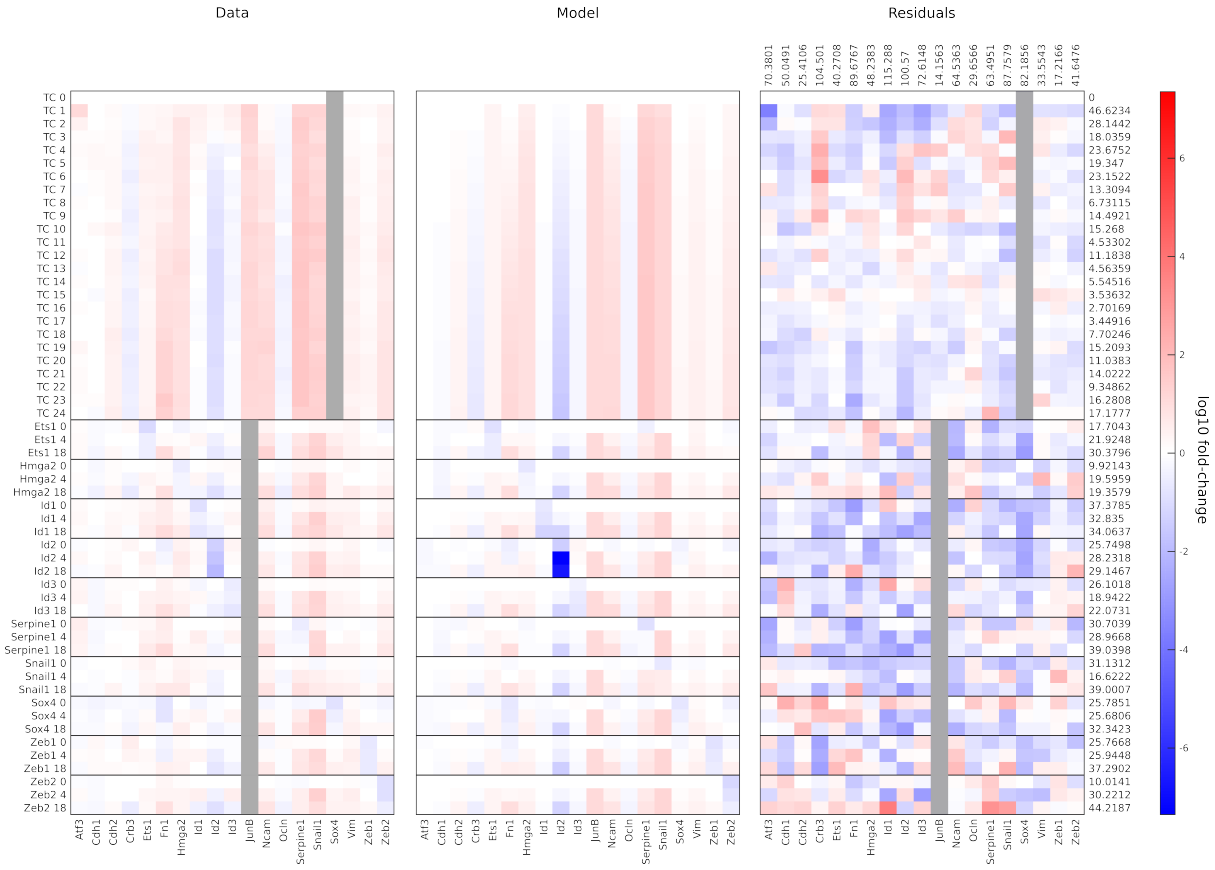


Figure 9.16: Comparison of NMuMG gene expression data and modelled gene expression values. Heatmaps show qPCR gene expression data determined in NMuMG cells after TGFβ-stimulation (left) and simulated gene expression for the best fit obtained by unregularized model fitting (middle, $\chi^2 = 1151.21$). Gray bars correspond to data-points for which no measurement exists. Heatmap on the right shows weighted least squares distance (residuals) between modelled and measured gene expression values. Numbers indicate combined χ^2-values for each target (rows) or each column (condition: cell-line + time-point).

9.4.5 Correlation between NMuMG gene expression data and modelled gene expression values

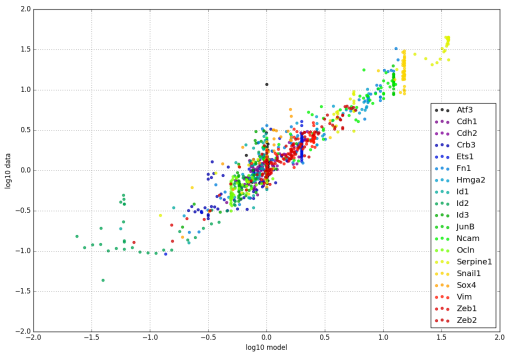


Figure 9.17: Scatter plot of the correlation between simulated gene expression and NMuMG gene expression data. Shown is the correlation between log10-transformed qPCR gene expression data collected in NMuMG cells after TGFβ-stimulation and simulated gene expression for the best fit obtained by un-regularized model fitting ($\chi^2 = 1151.21$).

9.4.6 Distribution of χ^2- and penalty-values for model fits with regularization level $\alpha = 5000$

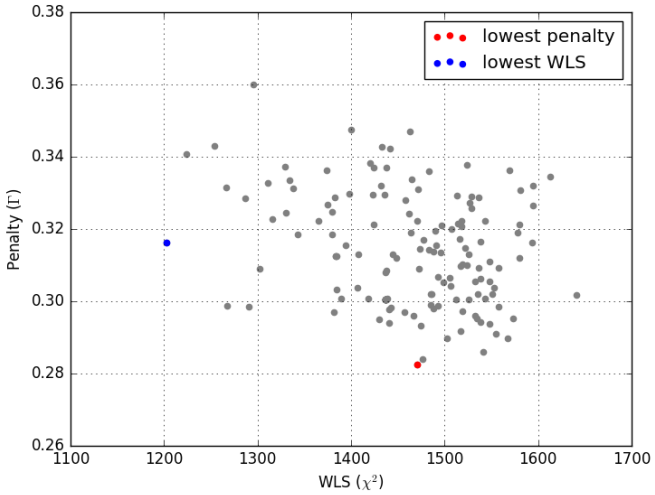


Figure 9.18: Distribution of χ^2- and penalty-values for model fits with $\alpha = 5000$. Scatter plot indicates the χ^2- and penalty-values for each model fit obtained with a regularization parameter of $\alpha = 5000$. The model fit resulting in the lowest χ^2-value is indicated in blue, while the fit with lowest penalty is indicated in red.

9.4.7 Top 25 interactions frequently selected by model fits with regularization parameter $\alpha = 5000$

Table 9.8: Persistently chosen interactions by model fits with $\alpha = 5000$. Top 25 most frequently selected interactions from model fitting with regularization parameter $\alpha = 5000$. Additional columns indicate interactions for which evidence for a direct regulatory relationship was found in the literature (grade A or grade B), interactions which could be supported by motif analysis (TRUE, FALSE), or by interactions determined by analysis of publicly available ChIP data (TRUE, FALSE). Empty cells correspond to interactions which are not evaluated by the respective analysis.

Rank	TF	Target	Times selected	Literature evidence	Motif evidence	ChIP evidence
1	Sox4	Fn1	150	grade B	FALSE	FALSE
2	pSmad2	JunB	150	grade B	FALSE	TRUE
3	Serpine1	Cdh2	149			
4	pSmad2	Snail1	148	grade A	FALSE	FALSE
5	pSmad2	Id2	148	grade B	FALSE	TRUE
6	Id3	Id1	143			
7	Serpine1	Zeb1	138			
8	Sox4	Ncam	136		FALSE	FALSE
9	Serpine1	Hmga2	134			
10	Id2	Id3	133		TRUE	
11	Sox4	Ocln	132		FALSE	FALSE
12	Snail1	Ocln	130		FALSE	
13	Zeb2	Fn1	129			
14	Zeb2	Ets1	124			
15	pSmad2	Serpine1	118		TRUE	FALSE
16	pSmad2	Crb3	113		TRUE	FALSE
17	Ets1	Zeb2	110	grade B	TRUE	FALSE
18	Sox4	Id2	109		FALSE	FALSE
19	Serpine1	Sox4	107			
20	Sox4	Cdh2	89	grade B	TRUE	FALSE
21	Zeb1	Crb3	87		TRUE	
22	JunB	Ets1	81		TRUE	FALSE
23	Zeb1	Sox4	80		FALSE	
24	pSmad2	Zeb2	75		TRUE	FALSE
25	Serpine1	Id2	74			

9.5 Supplemental Information for Chapter 6

9.5.1 Quality measures of paired transcriptome and proteome data

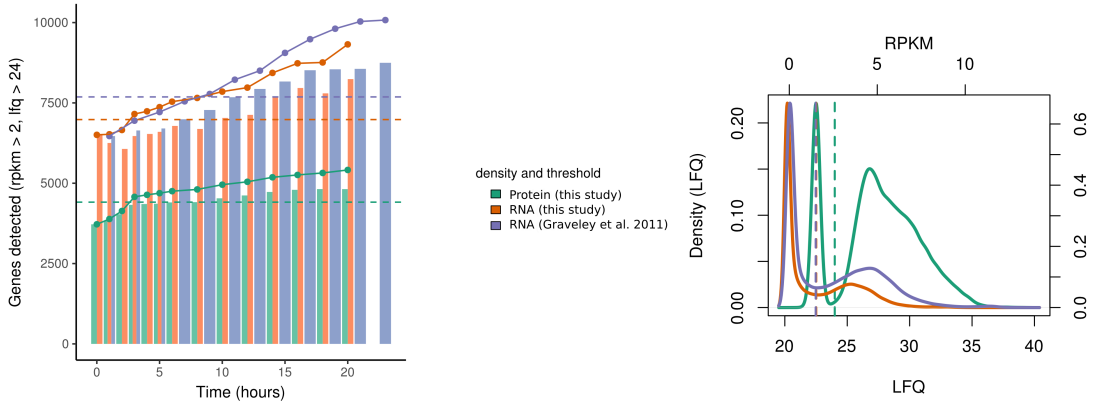


Figure 9.19: Quality measures of paired transcriptome and proteome data. Coverage of transcriptomic and proteomic data. Left panel: Number of genes detected in each sample for mass-spectrometry (green), RNA-sequencing in this study (orange) or RNA-Seq from Graveley et al. (2011) (purple). Right panel: distribution of either Reads Per Kilobase Million (RPKM - own RNA-Seq data and Graveley et al. (2011)) or Label-Free Quantification (LFQ) values obtained from mass-spectrometry (this study) over all samples. Same color code as in left panel.

9.5.2 Comparison of own and published RNAseq datasets

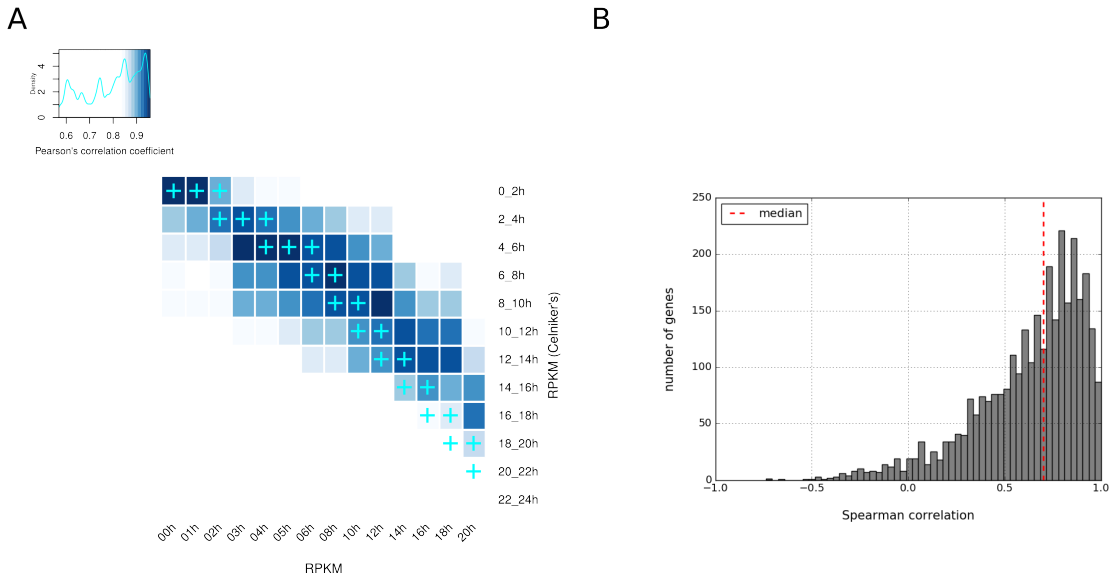


Figure 9.20: Comparison of own and published RNAseq datasets. (A) Strong global correlation between samples of RNAseq data from Graveley et al. (2011) and RNA-Seq in this study. Pearson correlation relating the RPKM values over all genes commonly identified in both datasets (12478 genes). Matching time points are indicated by a (+) sign. **(B)** Individual mRNAs show similar dynamics in Graveley et al (2011) and our dataset. For each of the 12478 common genes, the Spearman correlation was calculated by relating complete mRNA time-courses (time points in our data being paired to the earlier corresponding time point in the published data). The histogram shows the distribution of correlation coefficients over all 12478 genes.

9.5.3 Comparison of developmental progress in our experiments with Graveley et al. (2011)

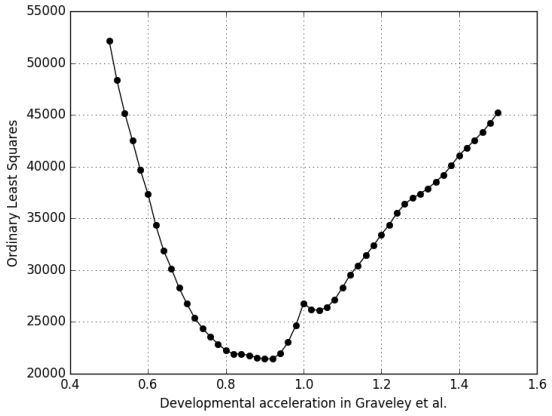


Figure 9.21: Comparison of developmental progress in our experiments with Graveley et al. (2011). Developmental progress is accelerated in the Graveley mRNA time-courses compared to our RNA-Seq data. The time axis in the Graveley et al (2011) dataset was compressed by an acceleration factor (x-axis; 1: no acceleration; ¡1: acceleration; ¿1: deceleration of mRNA dynamics in Graveley et al (2011)). Differences in the mRNA time-courses of both datasets were quantified by calculating the ordinary least squares difference summed up over the RPKM values of all time points and genes (y-axis). On data in Graveley et al. (2011) (multiplied by the acceleration factor) linear interpolation to obtain time points matching those of our RNA-Seq dataset was performed.

9.5.4 Correction of model selection due to conflicting biological and mathematical assumptions

In the model selection process we select between four different proposed models with varying numbers of parameters. The stationary and degradation model contain one and two parameters respectively, while the production and delay model contain three and four parameters. For a more complex model (i.e. a model with more parameters) to be preferred over a simpler one, it is necessary that the increase in the quality of the model fit of the more complex model compared to the simpler one exceeds a certain threshold. This ensures that both model fit and number of parameters are balanced and over-fitting is prevented. When comparing for example the degradation model with the production model from a biological point of view (Figure 9.22A), it is at least questionable whether the degradation model is indeed the simpler one: In reality one would need to assume an additional factor blocking translation, meaning that the 'implicit' number of parameters should be larger in the degradation model compared to the production model. Simply adding an extra parameter to the degradation model does not solve this problem. We therefore introduce a set of correction criteria to account for wrongly classified proteins:

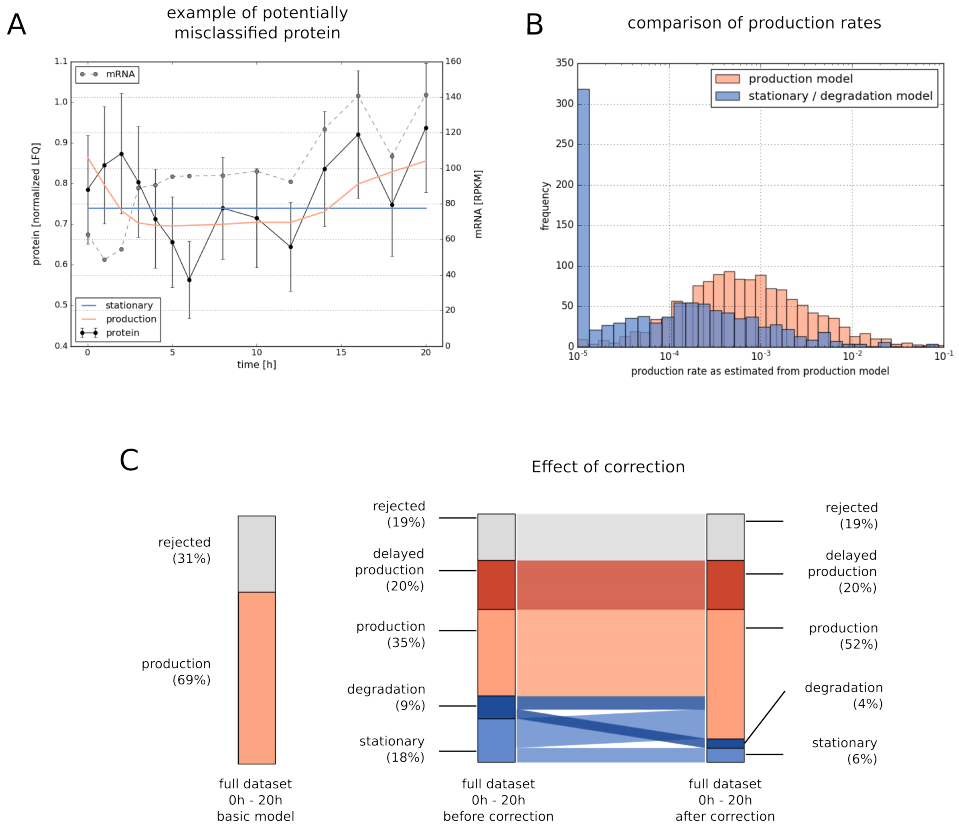


Figure 9.22: Correction of model selection due to conflicting biological and mathematical assumptions. **(A)** Exemplary time-course of a potentially misclassified protein. **(B)** Histogram of estimated production rates in the 'production'-model for proteins initially classified into the 'production'-group (red) or the 'stationary'- and 'degradation'-group (blue). **(C)** Schematic representation of the impact of correction step on protein classification. The bar on the left shows the percentage of proteins for which the 'production' model is assumed if no other models are considered. The bar in middle bar shows results of protein classification considering all four proposed models without correction. The bar on the right shows results of protein classification considering all four proposed models after correcting for misclassified proteins (see text below).

- The protein is assigned to the stationary or degradation class.
- The production model for this protein was not rejected based on the χ^2-test or Durbin-Watson test.
- The estimated production rate in the production model is bigger than 2e-5 (Figure 9.22B).

Based on this correction step, 646 out of 1002 (64%) proteins assigned to the stationary or degradation class are re-classified into the production group of proteins (Figure 9.22C). In the case where only data post-MZT is considered 1677 out of 2485 (67%) proteins are potentially wrongfully assigned to either the stationary or degradation class.

An additional justification of the performed correction step, is the consideration that a protein which can be described via the production model if the whole data is considered,

should not be classified as stationary or degradation if only post-MZT data is used for model selection. Such a case would imply that while the corresponding mRNA fully accounts for the protein time-course in the first case, additional regulators must be assumed in the second case. However, this again is a result of the assumption that both the stationary and degradation models are simpler compared to the production model. Proteins shifting from a simpler model (production) to a - biologically speaking - more complex model (stationary or degradation) after considering only the partial dataset should therefore be avoided. Indeed, the use of the proposed correction step reduces this number of 'undesired' shifts between protein classes from 707 to 234.

The situation described here reinforces the need for being aware that mathematical formulations not necessarily reflect biological reality.

9.5.5 Post-transcriptionally regulated genes and mRNA-protein dynamics in different protein groups

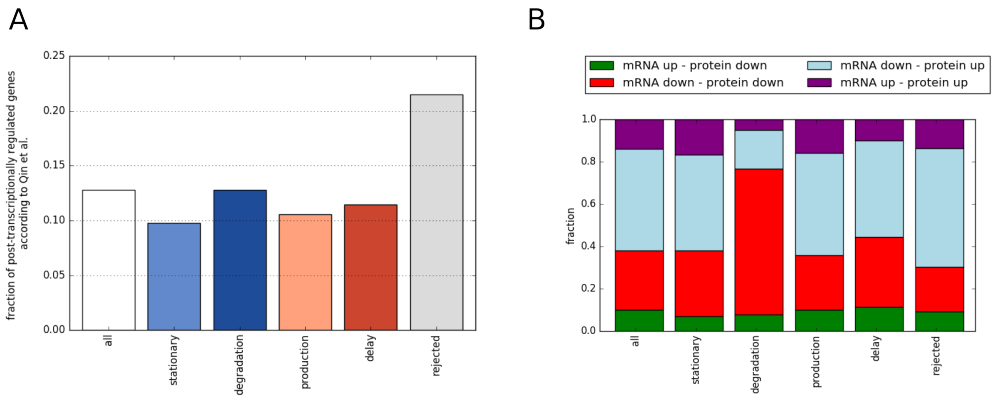


Figure 9.23: Post-transcriptionally regulated genes and mRNA-protein dynamics in different protein groups. (A) Genes previously identified as post-transcriptionally regulated during embryonic development by ribosome profiling [Qin et al., 2007], are enriched in the class of rejected proteins not explainable by the simple protein production and degradation model in our dataset. Based on the results by Qin et al. (2007) we derived a list of 1522 post-transcriptionally regulated genes by considering the union of individual lists provided in the supplement. **(B)** Association of protein groups with mRNA-protein dynamics inferred by clustering. Shown is the fraction of genes displaying one of the four possible mRNA/protein dynamics determined by hierarchical clustering.

9.5.6 mRNA and protein time-courses grouped by selected protein class

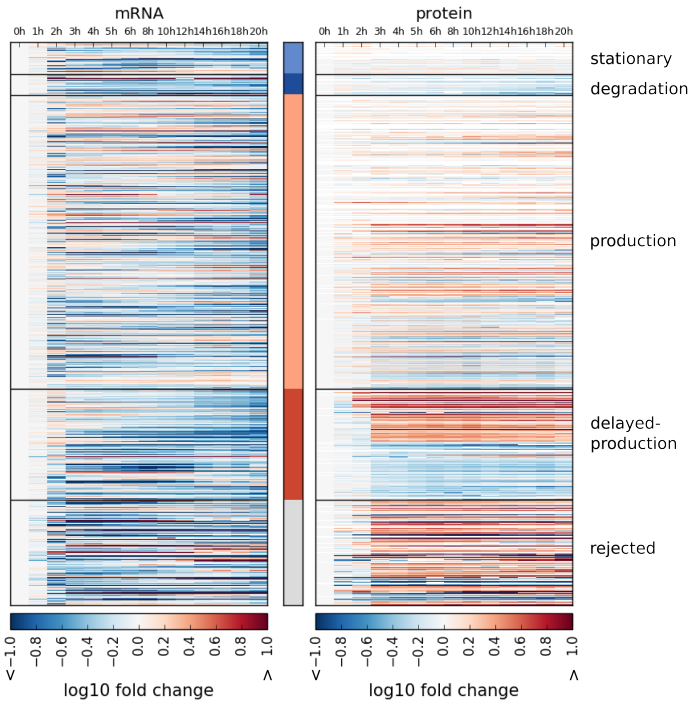


Figure 9.24: mRNA and Protein Time-Courses Grouped by Selected Protein Class. Time course heatmap of mRNA and protein dynamics sorted according to the model-based classification result (full dataset). Within each class, time-courses are sorted according to the estimated initial protein level (stationary), the degradation rate (degradation), the production rate (production) or the delay (delay). Fold-changes are shown within the range -1 and 1.

9.5.7 Inverse changes of mRNA and protein

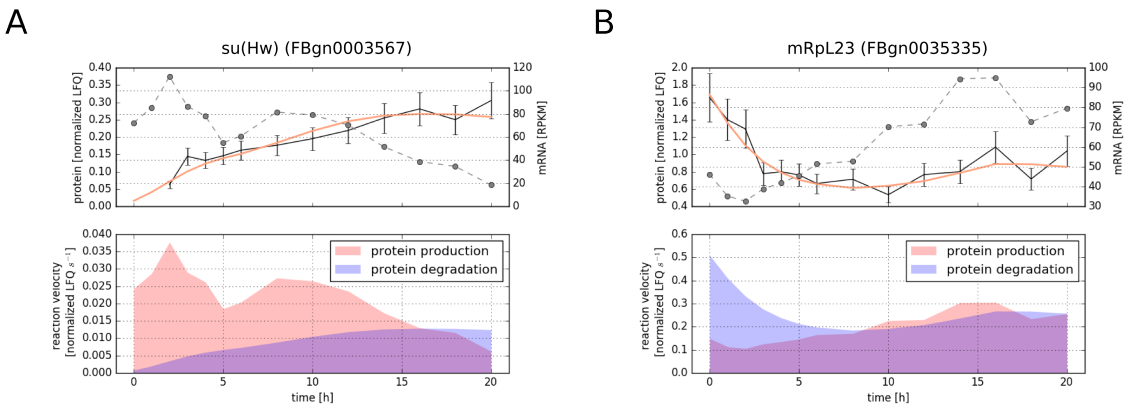


Figure 9.25: Inverse Changes of mRNA and Protein. Examples of genes with inverse mRNA and protein dynamics. **(A)** For mRpL23 (FBgn0035335), mRNA (dashed line) increases while protein decreases (black line - data; coloured line - model). The lower panel shows the velocities of protein production and degradation, which are initially unbalanced. **(B)** Conversely, su(Hw) (FBgn0003567) shows a protein increase while mRNA decreases, since the protein production velocity initially exceeds that of degradation.

9.5.8 Temporal dynamics of sugar metabolic and cell cycle regulatory proteins

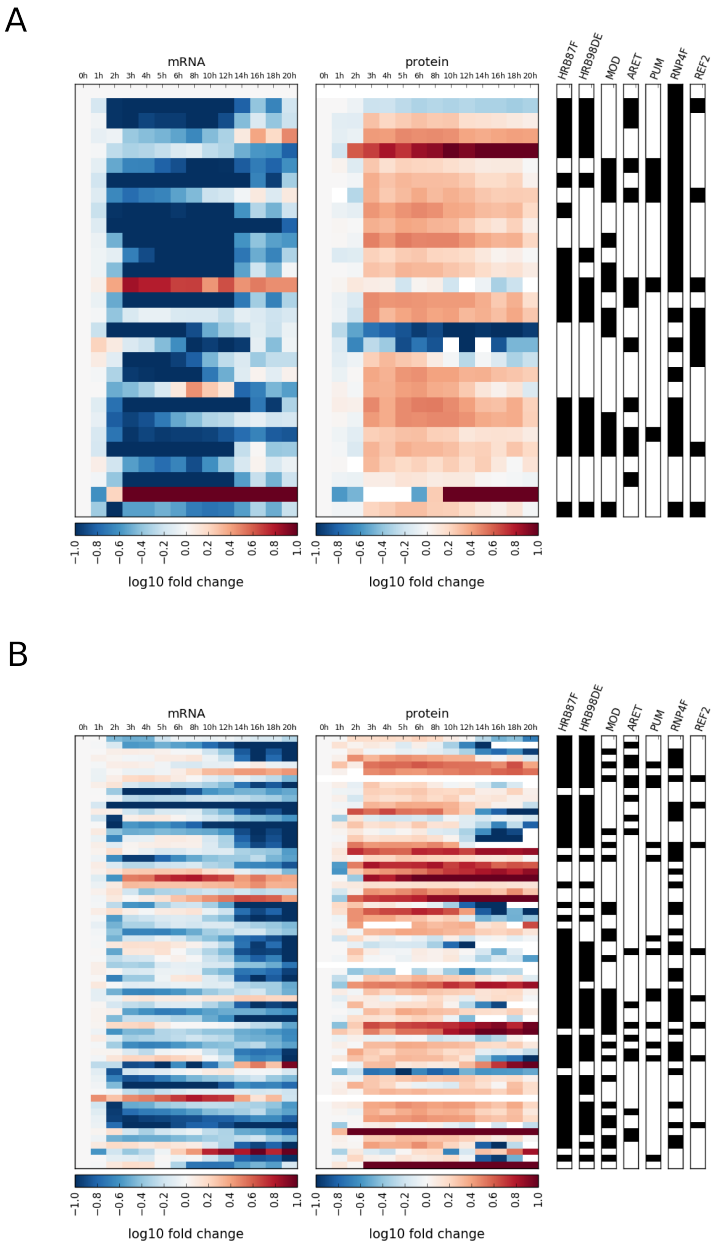


Figure 9.26: Temporal dynamics of sugar metabolic and cell cycle regulatory proteins. mRNA and protein expression time-courses of annotated with GO-terms associated with glucose metabolism **(A)** and cell cycle **(B)** in the rejected class of potentially post-transcriptionally regulated proteins. Shown are log10 fold-changes (relative to t=0h) of 29 genes annotated by at least one GO-term related to sugar metabolism or 65 genes annotated by at least one GO-term related to cell-cycle and mitosis. The center column indicates the presence of a Hrb98DE motif within the longest coding sequence of the gene.

9.5.9 Hrb98DE knocked down efficiency in S2R+ cells

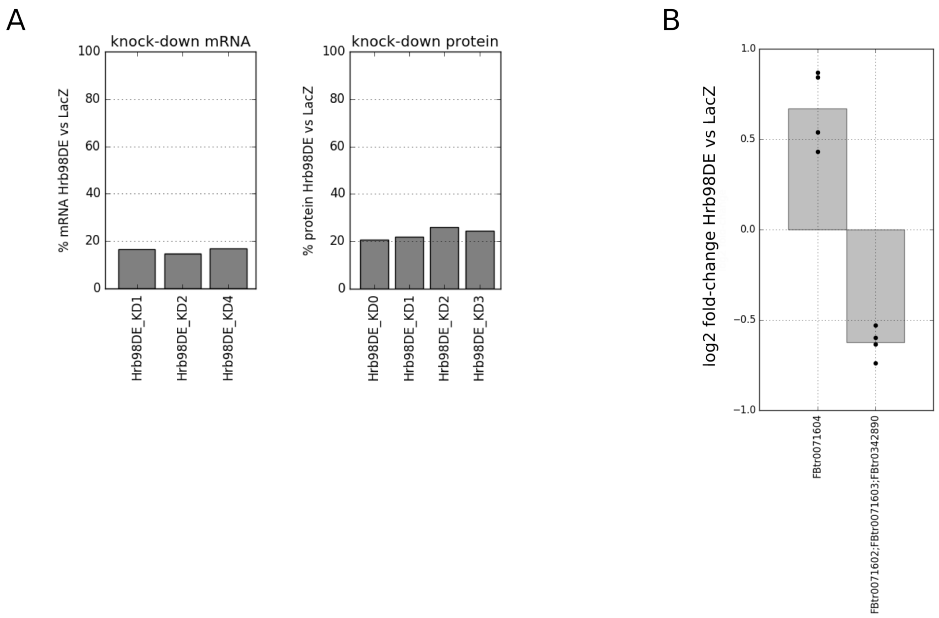


Figure 9.27: Hrb98DE knocked down with high efficiency in S2R+ cells. (A) Knock-down efficiency using Hrb98DE specific dsRNA on mRNA and protein levels. Shown are residual fractions of Hrb98DE mRNA (left, RNA-Seq) or protein (right, mass spectrometry) compared to LacZ knock-down control treated S2R+-cells 5 days after the first dsRNA treatment. **(B)** Isoform switch of domino (dom) on protein level. Barplots represent log2 fold-changes of two alternative domino proteingroups (FBtr0071604 and FBtr0071602;FBtr0071603;FBtr0342890). Individual replicates are shown in black.

List of Figures

List of Tables

Bibliography

[Amati et al., 1998] Amati, B., Alevizopoulos, K., and Vlach, J. (1998). Myc and the cell cycle. *Frontiers in bioscience : a journal and virtual library*, 3:d250–68.

[Anafi et al., 2014] Anafi, R. C., Lee, Y., Sato, T. K., Venkataraman, A., Ramanathan, C., Kavakli, I. H., Hughes, M. E., Baggs, J. E., Growe, J., Liu, A. C., Kim, J., and Hogenesch, J. B. (2014). Machine Learning Helps Identify CHRONO as a Circadian Clock Component. *PLoS Biology*, 12(4).

[Anders and Huber, 2010] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106.

[Anders et al., 2015] Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.

[Anders et al., 2012] Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017.

[Andrews, 2010] Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. *Babraham Bioinformatics*, page http://www.bioinformatics.babraham.ac.uk/projects/.

[Annayev et al., 2014] Annayev, Y., Adar, S., Chiou, Y.-Y., Lieb, J. D., Sancar, A., and Ye, R. (2014). Gene Model 129 (*Gm129*) Encodes a Novel Transcriptional Repressor That Modulates Circadian Gene Expression. *Journal of Biological Chemistry*, 289(8):5013–5024.

[Ashyraliyev et al., 2009] Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J. A., and Blom, J. G. (2009). Systems biology: Parameter estimation for biochemical models.

[Ay and Arnosti, 2011] Ay, A. and Arnosti, D. N. (2011). Mathematical modeling of gene expression: a guide for the perplexed biologist. *Critical reviews in biochemistry and molecular biology*, 46(2):137–151.

[Backes et al., 2007] Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y. A., Müller, R., Meese, E., and Lenhof, H.-P. (2007). Genetrailadvanced gene set enrichment analysis. *Nucleic acids research*, 35(suppl_2):W186–W192.

[Bakin et al., 2005] Bakin, A. V., Stourman, N. V., Sekhar, K. R., Rinehart, C., Yan, X., Meredith, M. J., Arteaga, C. L., and Freeman, M. L. (2005). Smad3–atf3 signaling mediates tgf-β suppression of genes encoding phase ii detoxifying proteins. *Free Radical Biology and Medicine*, 38(3):375–387.

[Bakin et al., 2000] Bakin, a. V., Tomlinson, a. K., Bhowmick, N. a., Moses, H. L., and Arteaga, C. L. (2000). Phosphatidylinositol 3-kinase function is required for transforming growth factor beta-mediated epithelial to mesenchymal transition and cell migration. *The Journal of biological chemistry*, 275(47):36803–10.

[Balsalobre et al., 2000] Balsalobre, A., Brown, S. A., Marcacci, L., Tronche, F., Kellendonk, C., Reichardt, H. M., Schutz, G., and Schibler, U. (2000). Resetting of circadian time in peripheral tissues by glucocorticoid signaling. *Science*, 289(5488):2344–2347.

[Bandara et al., 2009] Bandara, S., Schlöder, J. P., Eils, R., Bock, H. G., and Meyer, T. (2009). Optimal experimental design for parameter estimation of a cell signaling model. *PLoS computational biology*, 5(11):e1000558.

[Banga, 2008] Banga, J. R. (2008). Optimization in computational systems biology. *BMC systems biology*, 2(1):47.

[Becker et al., 2013] Becker, K., Balsa-Canto, E., Cicin-Sain, D., Hoermann, A., Janssens, H., Banga, J. R., and Jaeger, J. (2013). Reverse-Engineering Post-Transcriptional Regulation of Gap Genes in Drosophila melanogaster. *PLoS Computational Biology*, 9(10).

[Beckwith and Yanovsky, 2014] Beckwith, E. J. and Yanovsky, M. J. (2014). Circadian regulation of gene expression: at the crossroads of transcriptional and post-transcriptional regulatory networks. *Current Opinion in Genetics & Development*, 27:35–42.

[Beyer et al., 2004] Beyer, A., Hollunder, J., Nasheuer, H. P., and Wilhelm, T. (2004). Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. *Mol Cell Proteomics*, 3(11):1083–1092.

[Bhowmick et al., 2001] Bhowmick, N. A., Ghiassi, M., Bakin, A., Aakre, M., Lundquist, C. A., Engel, M. E., Arteaga, C. L., and Moses, H. L. (2001). Transforming Growth Factor-1 Mediates Epithelial to Mesenchymal Transdifferentiation through a RhoA-dependent Mechanism. *Molecular Biology of the Cell*, 12(1):27–36.

[Bieler et al., 2014] Bieler, J., Cannavo, R., Gustafson, K., Gobet, C., Gatfield, D., and Naef, F. (2014). Robust synchronization of coupled circadian and cell cycle oscillators in single mammalian cells. *Molecular Systems Biology*, 10(7):739–739.

[Bintu et al., 2005a] Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., and Phillips, R. (2005a). Transcriptional regulation by the numbers: Applications.

[Bintu et al., 2005b] Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005b). Transcriptional regulation by the numbers: Models.

[Blake et al., 2014] Blake, J. A., Bult, C. J., Eppig, J. T., Kadin, J. A., and Richardson, J. E. (2014). The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res*, 42(Database issue):D810–7.

[Blanchette et al., 2009] Blanchette, M., Green, R. E., MacArthur, S., Brooks, A. N., Brenner, S. E., Eisen, M. B., and Rio, D. C. (2009). Genome-wide Analysis of Alternative Pre-mRNA Splicing and RNA-Binding Specificities of the Drosophila hnRNP A/B Family Members. *Molecular Cell*, 33(4):438–449.

[Bluhm et al., 2016] Bluhm, A., Casas-Vila, N., Scheibe, M., and Butter, F. (2016). Reader interactome of epigenetic histone marks in birds. *Proteomics*, 16(3):427–436.

[Bonaldi et al., 2008] Bonaldi, T., Straub, T., Cox, J., Kumar, C., Becker, P. B., and Mann, M. (2008). Combined Use of RNAi and Quantitative Proteomics to Study Gene Function in Drosophila. *Molecular Cell*, 31(5):762–772.

[Bonneau et al., 2006] Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L., Baliga, N. S., and Thorsson, V. (2006). The inferelator: An algorithn for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology*, 7(5).

[Bozek et al., 2009] Bozek, K., Relógio, A., Kielbasa, S. M., Heine, M., Dame, C., Kramer, A., and Herzel, H. (2009). Regulation of clock-controlled genes in mammals. *PLoS ONE*, 4(3).

[Bozek et al., 2010] Bozek, K., Rosahl, A. L., Gaub, S., Lorenzen, S., and Herzel, H. (2010). Circadian transcription in liver. *BioSystems*, 102(1):61–69.

[Brockmann et al., 2007] Brockmann, R., Beyer, A., Heinisch, J. J., and Wilhelm, T. (2007). Posttranscriptional expression regulation: What determines translation rates? *PLoS Computational Biology*, 3(3):0531–0539.

[Brooks et al., 2015] Brooks, A. N., Duff, M. O., May, G., Yang, L., Bolisetty, M., Landolin, J., Wan, K., Sandler, J., Booth, B. W., Celniker, S. E., Graveley, B. R., and Brenner, S. E. (2015). Regulation of alternative splicing in Drosophila by 56 RNA binding proteins. *Genome Research*, 25(11):1771–1780.

[Brown et al., 2005] Brown, S. A., Ripperger, J., Kadener, S., Fleury-Olela, F., Vilbois, F., Rosbash, M., and Schibler, U. (2005). PERIOD1-associated proteins modulate the negative limb of the mammalian circadian oscillator. *Science*, 308(5722):693–696.

[Bugge et al., 2012] Bugge, A., Feng, D., Everett, L. J., Briggs, E. R., Mullican, S. E., Wang, F., Jager, J., and Lazar, M. A. (2012). Rev-erbα and Rev-erbβ coordinately protect the circadian clock and normal metabolic function. *Genes and Development*, 26(7):657–667.

[Bushati et al., 2008] Bushati, N., Stark, A., Brennecke, J., and Cohen, S. M. (2008). Temporal Reciprocity of miRNAs and Their Targets during the Maternal-to-Zygotic Transition in Drosophila. *Current Biology*, 18(7):501–506.

[Butte et al., 2000] Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., and Kohane, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182–12186.

[Cantone et al., 2009] Cantone, I., Marucci, L., Iorio, F., Ricci, M. A., Belcastro, V., Bansal, M., Santini, S., Di Bernardo, M., Di Bernardo, D., and Cosma, M. P. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, 137(1):172–181.

[Carreira et al., 2005] Carreira, S., Goodall, J., Aksan, I., La Rocca, S. A., Galibert, M.-D., Denat, L., Larue, L., and Goding, C. R. (2005). Mitf cooperates with Rb1 and activates p21Cip1 expression to regulate cell cycle progression. *Nature*, 433(7027):764–769.

[Carvajal et al., 2012] Carvajal, L. A., Hamard, P. J., Tonnessen, C., and Manfredi, J. J. (2012). E2F7, a novel target, is up-regulated by p53 and mediates DNA damage-dependent transcriptional repression. *Genes and Development*, 26(14):1533–1545.

[Casas-Vila et al., 2017] Casas-Vila, N., Bluhm, A., Sayols, S., Dinges, N., Dejung, M., Altenhein, T., Kappei, D., Altenhein, B., Roignant, J. Y., and Butter, F. (2017). The developmental proteome of Drosophila melanogaster. *Genome Research*, 27(7):1273–1285.

[Cedersund and Roll, 2009] Cedersund, G. and Roll, J. (2009). Systems biology: Model based evaluation and comparison of potential explanations for given biological data.

[Chang et al., 2016] Chang, H., Liu, Y., Xue, M., Liu, H., Du, S., Zhang, L., and Wang, P. (2016). Synergistic action of master transcription factors controls epithelial-to-mesenchymal transition. *Nucleic Acids Research*, 44(6):2514–2527.

[Chen et al., 2014] Chen, W., Liu, Z., Li, T., Zhang, R., Xue, Y., Zhong, Y., Bai, W., Zhou, D., and Zhao, Z. (2014). Regulation of Drosophila circadian rhythms by miRNA let-7 is mediated by a regulatory cycle. *Nature Communications*, 5.

[Cho et al., 2012] Cho, H., Zhao, X., Hatori, M., Yu, R. T., Barish, G. D., Lam, M. T., Chong, L. W., Ditacchio, L., Atkins, A. R., Glass, C. K., Liddle, C., Auwerx, J., Downes, M., Panda, S., and Evans, R. M. (2012). Regulation of circadian behaviour and metabolism by REV-ERB-α and REV-ERB-β. *Nature*, 485(7396):123–127.

[Cho et al., 1998] Cho, R. J., Campbell, M. J., Winzeler, E. a., Steinmetz, L., Conway, a., Wodicka, L., Wolfsberg, T. G., Gabrielian, a. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular cell*, 2(1):65–73.

[Chomczyński and Sacchi, 1987] Chomczyński, P. and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinum thiocyanate-phenol-chlotoform extraction. *Analytical Biochemistry*, 162(1):156–159.

[Chu et al., 2013] Chu, X., Qin, X., Xu, H., Li, L., Wang, Z., Li, F., Xie, X., Zhou, H., Shen, Y., and Long, J. (2013). Structural insights into Paf1 complex assembly and histone binding. *Nucleic Acids Research*, 41(22):10619–10629.

[Coon et al., 2012] Coon, S. L., Munson, P. J., Cherukuri, P. F., Sugden, D., Rath, M. F., Moller, M., Clokie, S. J. H., Fu, C., Olanich, M. E., Rangel, Z., Werner, T., Mullikin, J. C., Klein, D. C., Benjamin, B., Blakesley, R., Bouffard, G., Brooks, S., Chu, G., Coleman, H., Dekhtyar, M., Gregory, M., Guan, X., Gupta, J., Han, J., Hargrove, A., Ho, S.-l., Johnson, T., Legaspi, R., Lovett, S., Maduro, Q., Masiello, C., Maskeri, B., McDowell, J., Montemayor, C., Novotny, B., Park, M., Riebow, N., Schandler, K., Schmidt, B., Sison, C., Stantripop, M., Thomas, J., Vemulapalli, M., and Young, A. (2012). Circadian changes in long noncoding RNAs in the pineal gland. *Proceedings of the National Academy of Sciences*, 109(33):13319–13324.

[Cox et al., 2014] Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics*, 13(9):2513–2526.

[Dave et al., 2011] Dave, N., Guaita-Esteruelas, S., Gutarra, S., Frias, À., Beltran, M., Peiró, S., and García De Herreros, A. (2011). Functional cooperation between snail1 and twist in the regulation of ZEB1 expression during epithelial to mesenchymal transition. *Journal of Biological Chemistry*, 286(14):12024–12032.

[Davidson et al., 2002] Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., et al. (2002). A genomic regulatory network for development. *science*, 295(5560):1669–1678.

[de Lichtenberg et al., 2005] de Lichtenberg, U., Jensen, L. J., Fausbøll, A., Jensen, T. S., Bork, P., and Brunak, S. (2005). Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics*, 21(7):1164–1171.

[De Renzis et al., 2007] De Renzis, S., Elemento, O., Tavazoie, S., and Wieschaus, E. F. (2007). Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. *PLoS Biology*, 5(5):1036–1051.

[De Smet and Marchal, 2010] De Smet, R. and Marchal, K. (2010). Advantages and limitations of current network inference methods.

[de Sousa Abreu et al., 2009] de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., and Vogel, C. (2009). Global signatures of protein and mRNA expression levels. *Molecular bioSystems*, 5(12):1512–26.

[Dermody and Buratowski, 2010] Dermody, J. L. and Buratowski, S. (2010). Leo1 subunit of the yeast Paf1 complex binds RNA and contributes to complex recruitment. *Journal of Biological Chemistry*, 285(44):33671–33679.

[Despic et al., 2017] Despic, V., Dejung, M., Gu, M., Krishnan, J., Zhang, J., Herzel, L., Straube, K., Gerstein, M. B., Butter, F., and Neugebauer, K. M. (2017). Dynamic RNA-protein interactions underlie the zebrafish maternal-to-zygotic transition. *Genome Research*, 27(7):1184–1194.

[DiTacchio et al., 2011] DiTacchio, L., Le, H. D., Vollmers, C., Hatori, M., Witcher, M., Secombe, J., and Panda, S. (2011). Histone lysine demethylase JARID1a activates CLOCK-BMAL1 and influences the circadian clock. *Science*, 333(6051):1881–1885.

[Dobin et al., 2013] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

[Doherty and Kay, 2010] Doherty, C. J. and Kay, S. A. (2010). Circadian Control of Global Gene Expression Patterns. *Annual Review of Genetics*, 44(1):419–444.

[Doi et al., 2006] Doi, M., Hirayama, J., and Sassone-Corsi, P. (2006). Circadian Regulator CLOCK Is a Histone Acetyltransferase. *Cell*, 125(3):497–508.

[Dougherty and Braga-Neto, 2006] Dougherty, E. R. and Braga-Neto, U. (2006). Epistemology of computational biology: mathematical models and experimental prediction as the basis of their validity. *Journal of Biological Systems*, 14(01):65–90.

[Ducat et al., 2008] Ducat, D., Kawaguchi, S.-i., Liu, H., Yates, J. R., and Zheng, Y. (2008). Regulation of microtubule assembly and organization in mitosis by the aaa+ atpase pontin. *Molecular biology of the cell*, 19(7):3097–3110.

[Edgar, 2002] Edgar, R. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

[Egea et al., 2014] Egea, J. a., Henriques, D., Cokelaer, T., Villaverde, A. F., MacNamara, A., Danciu, D.-p., Banga, J. R., and Saez-Rodriguez, J. (2014). MEIGO: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. *BMC bioinformatics*, 15:136.

[Egea et al., 2009] Egea, J. a., Vazquez, E., Banga, J. R., and Martí, R. (2009). Improved scatter search for the global optimization of computationally expensive dynamic models. *Journal of Global Optimization*, 43(2-3):175–190.

[Eichelbaum and Krijgsveld, 2014] Eichelbaum, K. and Krijgsveld, J. (2014). Rapid Temporal Dynamics of Transcription, Protein Synthesis, and Secretion during Macrophage Activation. *Molecular & Cellular Proteomics*, 13(3):792–810.

[Engelen et al., 2013] Engelen, E., Janssens, R. C., Yagita, K., Smits, V. A. J., van der Horst, G. T. J., and Tamanini, F. (2013). Mammalian TIMELESS Is Involved in Period Determination and DNA Damage-Dependent Phase Advancing of the Circadian Clock. *PLoS ONE*, 8(2).

[Etchegaray et al., 2003] Etchegaray, J. P., Lee, C., Wade, P. A., and Reppert, S. M. (2003). Rhythmic histone acetylation underlies transcription in the mammalian circadian clock. *Nature*, 421(6919):177–182.

[Etchegaray et al., 2006] Etchegaray, J. P., Yang, X., Debruyne, J. P., Peters, A. H., Weaver, D. R., Jenuwein, T., and Reppert, S. M. (2006). The polycomb group protein EZH2 is required for mammalian circadian clock function. *Journal of Biological Chemistry*, 281(30):21209–21215.

[Faith et al., 2007a] Faith, J. J., Driscoll, M. E., Fusaro, V. A., Cosgrove, E. J., Hayete, B., Juhn, F. S., Schneider, S. J., and Gardner, T. S. (2007a). Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic acids research*, 36(suppl_1):D866–D870.

[Faith et al., 2007b] Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007b). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1):0054–0066.

[Fang et al., 2014] Fang, B., Everett, L. J., Jager, J., Briggs, E., Armour, S. M., Feng, D., Roy, A., Gerhart-Hines, Z., Sun, Z., and Lazar, M. A. (2014). Circadian enhancers coordinate multiple phases of rhythmic gene transcription in vivo. *Cell*, 159(5):1140–1152.

[Filipowicz et al., 2008] Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: Are the answers in sight?

[Flicek et al., 2014] Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Kulesha, E., Martin, F. J., Maurel, T., McLaren, W. M., Murphy, D. N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S. J., Vullo, A., Wilder, S. P., Wilson, M., Zadissa, A., Aken, B. L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T. J., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D. R., and Searle, S. M. (2014). Ensembl 2014. *Nucleic Acids Research*, 42(D1).

[Fonjallaz et al., 1996] Fonjallaz, P., Ossipow, V., Wanner, G., and Schibler, U. (1996). The two PAR leucine zipper proteins, TEF and DBP, display similar circadian and tissue-specific expression, but have different target promoter preferences. *The EMBO journal*, 15(2):351–62.

[Fournier et al., 2010] Fournier, M. L., Paulson, A., Pavelka, N., Mosley, A. L., Gaudenz, K., Bradford, W. D., Glynn, E., Li, H., Sardiu, M. E., Fleharty, B., Seidel, C., Florens, L., and Washburn, M. P. (2010). Delayed Correlation of mRNA and Protein Expression in Rapamycin-treated Cells and a Role for Ggc1 in Cellular Sensitivity to Rapamycin. *Molecular & Cellular Proteomics*, 9(2):271–284.

[Futschik and Herzel, 2008] Futschik, M. E. and Herzel, H. (2008). Are we overestimating the number of cell-cycling genes? The impact of background models on time-series analysis. *Bioinformatics*, 24(8):1063–1069.

[Gabor and Banga, 2015] Gabor, A. and Banga, J. R. (2015). Robust and efficient parameter estimation in dynamic models of biological systems. *BMC Systems Biology*, 9(1).

[Gaidatzis et al., 2015] Gaidatzis, D., Lerch, A., Hahne, F., and Stadler, M. B. (2015). QuasR: Quantification and annotation of short reads in R. *Bioinformatics*, 31(7):1130–1132.

[Gama-Castro et al., 2015] Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., Castro-Mondragón, J. A., et al. (2015). Regulondb version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic acids research*, 44(D1):D133–D143.

[Gedeon and Bokes, 2012] Gedeon, T. and Bokes, P. (2012). Delayed protein synthesis reduces the correlation between mRNA and protein fluctuations. *Biophysical Journal*, 103(3):377–385.

[Gervasi et al., 2012] Gervasi, M., Bianchi-Smiraglia, A., Cummings, M., Zheng, Q., Wang, D., Liu, S., and Bakin, A. V. (2012). JunB contributes to Id2 repression and the epithelial-mesenchymal transition in response to transforming growth factor-β. *Journal of Cell Biology*, 196(5):589–603.

[Glisovic et al., 2008] Glisovic, T., Bachorik, J. L., Yong, J., and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation.

[Goda et al., 2003] Goda, N., Ryan, H. E., Khadivi, B., McNulty, W., Rickert, R. C., and Johnson, R. S. (2003). Hypoxia-inducible factor 1alpha is essential for cell cycle arrest during hypoxia. *Molecular and cellular biology*, 23(1):359–69.

[Goldberg, 1995] Goldberg, A. L. (1995). Functions of the proteasome: the lysis at the end of the tunnel. *Science*, 268(5210):522.

[Gomez-Cabrero et al., 2014] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). Data integration in the era of omics: current and future challenges.

[Goodwin, 1965] Goodwin, B. C. (1965). Oscillatory behavior in enzymatic control processes. *Advances in enzyme regulation*, 3:425–437.

[Goriki et al., 2014] Goriki, A., Hatanaka, F., Myung, J., Kim, J. K., Yoritaka, T., Tanoue, S., Abe, T., Kiyonari, H., Fujimoto, K., Kato, Y., Todo, T., Matsubara, A., Forger, D., and Takumi, T. (2014). A Novel Protein, CHRONO, Functions as a Core Component of the Mammalian Circadian Clock. *PLoS Biology*, 12(4).

[Gouw et al., 2009] Gouw, J. W., Pinkse, M. W. H., Vos, H. R., Moshkin, Y., Verrijzer, C. P., Heck, A. J. R., and Krijgsveld, J. (2009). In Vivo Stable Isotope Labeling of Fruit Flies Reveals Post-transcriptional Regulation in the Maternal-to-zygotic Transition. *Molecular & Cellular Proteomics*, 8(7):1566–1578.

[Grant et al., 2011] Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

[Graveley et al., 2011] Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., Van Baren, M. J., Boley, N., Booth, B. W., Brown, J. B., Cherbas, L., Davis, C. A., Dobin, A., Li, R., Lin, W., Malone, J. H., Mattiuzzo, N. R., Miller, D., Sturgill, D., Tuch, B. B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R. E., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J. E., Wan, K. H., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P. J., Brenner, S. E., Brent, M. R., Cherbas, P., Gingeras, T. R., Hoskins, R. A., Kaufman, T. C., Oliver, B., and Celniker, S. E. (2011). The developmental transcriptome of Drosophila melanogaster. *Nature*, 471(7339):473–479.

[Greenbaum et al., 2003] Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale.

[Griffin et al., 2002] Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L., and Aebersold, R. (2002). Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in *Saccharomyces cerevisiae*. *Molecular & Cellular Proteomics*, 1(4):323–333.

[Groisman et al., 2002] Groisman, I., Jung, M. Y., Sarkissian, M., Cao, Q., and Richter, J. D. (2002). Translational control of the embryonic cell cycle. *Cell*, 109(4):473–483.

[Guillén-Gosálbez et al., 2013] Guillén-Gosálbez, G., Miró, A., Alves, R., Sorribas, A., and Jiménez, L. (2013). Identification of regulatory structure and kinetic parameters of biochemical networks via mixed-integer dynamic optimization. *BMC Systems Biology*, 7.

[Gunawardena, 2014] Gunawardena, J. (2014). Models in biology:accurate descriptions of our pathetic thinking. *BMC biology*, 12(1):29.

[Gygi et al., 1999] Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between Protein and mRNA Abundance in Yeast. *Molecular and Cellular Biology*, 19(3):1720–1730.

[Hamilton and Kay, 2008] Hamilton, E. E. and Kay, S. A. (2008). SnapShot: Circadian Clock Proteins.

[Härdle et al., 2003] Härdle, W., Horowitz, J., and Kreiss, J.-P. (2003). Bootstrap methods for time series. *International Statistical Review*, 71(2):435–459.

[Hargrove and Schmidt, 1989] Hargrove, J. L. and Schmidt, F. H. (1989). The role of mRNA and protein stability in gene expression. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 3(12):2360–70.

[Heinz et al., 2010] Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589.

[Henriques et al., 2015] Henriques, D., Rocha, M., Saez-Rodriguez, J., and Banga, J. R. (2015). Reverse engineering of logic-based differential equation models using a mixed-integer dynamic optimization approach. *Bioinformatics*, 31(18):2999–3007.

[Hill, 1910] Hill, A. V. (1910). The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *Journal of Physiology*, 40(0):iv–vii.

[Hillenbrand et al., 2016] Hillenbrand, P., Maier, K. C., Cramer, P., and Gerland, U. (2016). Inference of gene regulation functions from dynamic transcriptome data. *eLife*, 5(September2016).

[Honkela et al., 2010] Honkela, A., Girardot, C., Gustafson, E. H., Liu, Y.-H., Furlong, E. E., Lawrence, N. D., and Rattray, M. (2010). Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences*, 107(17):7793–7798.

[Huang et al., 2009a] Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13.

[Huang et al., 2009b] Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57.

[Huber et al., 2015] Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MaCdonald, J., Obenchain, V., Oles, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., and Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121.

[Hughes et al., 2009] Hughes, M. E., DiTacchio, L., Hayes, K. R., Vollmers, C., Pulivarthy, S., Baggs, J. E., Panda, S., and Hogenesch, J. B. (2009). Harmonics of circadian gene transcription in mammals. *PLoS Genetics*, 5(4).

[Hughes et al., 2010] Hughes, M. E., Hogenesch, J. B., and Kornacker, K. (2010). JTK-CYCLE: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of Biological Rhythms*, 25(5):372–380.

[Humphries et al., 2002] Humphries, A., Klein, D., Baler, R., and Carter, D. A. (2002). cDNA array analysis of pineal gene expression reveals circadian rhythmicity of the dominant negative helix-loop-helix protein-encoding gene, ld-1. *Journal of Neuroendocrinology*, 14(2):101–108.

[Huynh-Thu et al., 2010] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9).

[Ingolia et al., 2009] Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223.

[Iqbal et al., 2010] Iqbal, M. S., Otsuyama, K. I., Shamsasenjan, K., Asaoku, H., and Kawano, M. M. (2010). CD56 expression in human myeloma cells derived from the neurogenic gene expression: Possible role of the SRY-HMG box gene, SOX4. *International Journal of Hematology*, 91(2):267–275.

[Isojimaa et al., 2009] Isojimaa, Y., Nakajimac, M., Ukaic, H., Fujishimac, H., Yamadac, R. G., Masumotoc, K.-h., Kiuchia, R., Ishidac, M., Ukai-Tadenumac, M., Minamic, Y., et al. (2009). Cki/-dependent phosphorylation is a temperature-insensitive, period-determining process in the mammalian circadian clock. *PNAS*, 106(37).

[Izumo et al., 2006] Izumo, M., Sato, T. R., Straume, M., and Johnson, C. H. (2006). Quantitative Analyses of Circadian Gene Expression in Mammalian Cell Cultures. *PLOS Comput Biol*, 2(10):e136.

[Jaeger et al., 2004] Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., Vanario-Alonso, C. E., Samsonova, M., Sharp, D. H., and Reinitz, J. (2004). Dynamic control of positional information in the early Drosophila embryo. *Nature*, 430(6997):368–371.

[Ji and Tulin, 2016] Ji, Y. and Tulin, A. V. (2016). Poly(ADP-Ribosyl)ation of hnRNP A1 Protein Controls Translational Repression in Drosophila. *Molecular and Cellular Biology*, 36(19):2476–2486.

[Johansson et al., 2003] Johansson, D., Lindgren, P., and Berglund, a. (2003). A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics*, 19(4):467–473.

[Johansson et al., 2014] Johansson, R., Strålfors, P., and Cedersund, G. (2014). Combining test statistics and models in bootstrapped model rejection: It is a balancing act. *BMC Systems Biology*, 8(1).

[Johnson et al., 1993] Johnson, D. G., Schwarz, J. K., Cress, W. D., and Nevins, J. R. (1993). Expression of transcription factor E2F1 induces quiescent cells to enter S phase. *Nature*, 365(6444):349–52.

[Jones et al., 2010] Jones, M. A., Covington, M. F., DiTacchio, L., Vollmers, C., Panda, S., and Harmer, S. L. (2010). Jumonji domain protein JMJD5 functions in both the plant and human circadian systems. *Proceedings of the National Academy of Sciences*, 107(50):21623–21628.

[Jovanovic et al., 2015] Jovanovic, M., Rooney, M. S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E. H., Fields, A. P., Schwartz, S., Raychowdhury, R., Mumbach, M. R., Eisenhaure, T., Rabani, M., Gennert, D., Lu, D., Delorey, T., Weissman, J. S., Carr, S. A., Hacohen, N., and Regev, A. (2015). Dynamic profiling of the protein life cycle in response to pathogens. *Science*, 347(6226).

[Kamenz et al., 2015] Kamenz, J., Mihaljev, T., Kubis, A., Legewie, S., and Hauf, S. (2015). Robust ordering of anaphase events by adaptive thresholds and competing degradation pathways. *Molecular cell*, 60(3):446–459.

[Karlebach and Shamir, 2008] Karlebach, G. and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770.

[Katz et al., 2010] Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015.

[Kauffman, 1969] Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22(3):437–467.

[Kennedy et al., 2011] Kennedy, B. a., Deatherage, D. E., Gu, F., Tang, B., Chan, M. W. Y., Nephew, K. P., Huang, T. H. M., and Jin, V. X. (2011). Chip-Seq defined Genome-Wide map of TGF??/SMAD4 targets: Implications with clinical outcome of ovarian cancer. *PLoS ONE*, 6(7).

[Kent et al., 2002] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006.

[Ki et al., 2015] Ki, Y., Ri, H., Lee, H., Yoo, E., Choe, J., and Lim, C. (2015). Warming Up Your Tick-Tock: Temperature-Dependent Regulation of Circadian Clocks.

[Kim et al., 2015] Kim, S. H., Lee, K. H., Kim, D. Y., Kwak, E., Kim, S., and Kim, K. T. (2015). Rhythmic control of mRNA stability modulates circadian amplitude of mouse Period3 mRNA. *Journal of Neurochemistry*, 132(6):642–656.

[Kim and Lee, 2012] Kim, W. and Lee, E. K. (2012). Post-transcriptional regulation in metabolic diseases.

[Kirk et al., 2013] Kirk, P., Thorne, T., and Stumpf, M. P. (2013). Model selection in systems and synthetic biology. *Current opinion in biotechnology*, 24(4):767–774.

[Koike et al., 2012] Koike, N., Yoo, S. H., Huang, H. C., Kumar, V., Lee, C., Kim, T. K., and Takahashi, J. S. (2012). Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science*, 338(6105):349–354.

[Kojima and Green, 2015] Kojima, S. and Green, C. B. (2015). Circadian genomics reveal a role for post-transcriptional regulation in mammals. *Biochemistry*, 54(2):124–133.

[Kondo et al., 2004] Kondo, M., Cubillo, E., Tobiume, K., Shirakihara, T., Fukuda, N., Suzuki, H., Shimizu, K., Takehara, K., Cano, A., Saitoh, M., et al. (2004). A role for id in the regulation of tgf-β-induced epithelial–mesenchymal transdifferentiation. *Cell death and differentiation*, 11(10):1092.

[Korenčič et al., 2012] Korenčič, A., Bordyugov, G., Košir, R., Rozman, D., Goličnik, M., and Herzel, H. (2012). The Interplay of cis-Regulatory Elements Rules Circadian Rhythms in Mouse Liver. *PLoS ONE*, 7(11).

[Korenčič et al., 2014] Korenčič, A., Košir, R., Bordyugov, G., Lehmann, R., Rozman, D., and Herzel, H. (2014). Timing of circadian genes in mammalian tissues. *Scientific Reports*, 4.

[Kornienko et al., 2013] Kornienko, A. E., Guenzl, P. M., Barlow, D. P., and Pauler, F. M. (2013). Gene regulation by the act of long non-coding RNA transcription.

[Korpal et al., 2008] Korpal, M., Lee, E. S., Hu, G., and Kang, Y. (2008). The miR-200 family inhibits epithelial-mesenchymal transition and cancer cell migration by direct targeting of E-cadherin transcriptional repressors ZEB1 and ZEB2. *Journal of Biological Chemistry*, 283(22):14910–14914.

[Koshiji et al., 2004] Koshiji, M., Kageyama, Y., Pete, E. A., Horikawa, I., Barrett, J. C., and Huang, L. E. (2004). HIF-1α induces cell cycle arrest by functionally counteracting Myc. *EMBO Journal*, 23(9):1949–1956.

[Kowalska et al., 2013] Kowalska, E., Ripperger, J. A., Hoegger, D. C., Bruegger, P., Buch, T., Birchler, T., Mueller, A., Albrecht, U., Contaldo, C., and Brown, S. A. (2013). NONO couples the circadian clock to the cell cycle. *Proceedings of the National Academy of Sciences*, 110(5):1592–1599.

[Kowanetz et al., 2004] Kowanetz, M., Valcourt, U., Bergström, R., Heldin, C.-H., and Moustakas, A. (2004). Id2 and id3 define the potency of cell proliferation and differentiation responses to transforming growth factor β and bone morphogenetic protein. *Molecular and cellular biology*, 24(10):4241–4254.

[Kwak et al., 2006] Kwak, E., Kim, T. D., and Kim, K. T. (2006). Essential role of 3-untranslated region-mediated mRNA decay in circadian oscillations of mouse Period3 mRNA. *Journal of Biological Chemistry*, 281(28):19100–19106.

[Lam et al., 2013] Lam, M. T., Cho, H., Lesch, H. P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M. U., Kim, A. S., Kosaka, M., Lee, C. Y., Watt, A., Grossman, T. R., Rosenfeld, M. G., Evans, R. M., and Glass, C. K. (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature*, 498(7455):511–515.

[Lamouille et al., 2014] Lamouille, S., Xu, J., and Derynck, R. (2014). Molecular mechanisms of epithelial-mesenchymal transition. *Nature reviews. Molecular cell biology*, 15(3):178–96.

[Langley et al., 2014] Langley, A. R., Smith, J. C., Stemple, D. L., and Harvey, S. A. (2014). New insights into the maternal to zygotic transition. *Development*, 141(20):3834–3841.

[Langmead, 2010] Langmead, B. (2010). Aligning short sequencing reads with Bowtie. *Current Protocols in Bioinformatics*, (SUPP.32).

[Lasko, 2012] Lasko, P. (2012). mRNA localization and translational control in Drosophila oogenesis.

[Le Martelot et al., 2012] Le Martelot, G., Canella, D., Symul, L., Migliavacca, E., Gilardi, F., Liechti, R., Martin, O., Harshman, K., Delorenzi, M., Desvergne, B., Herr, W., Deplancke, B., Schibler, U., Rougemont, J., Guex, N., Hernandez, N., and Naef, F. (2012). Genome-Wide RNA Polymerase II Profiles and RNA Accumulation Reveal Kinetics of Transcription and Associated Epigenetic Changes During Diurnal Cycles. *PLoS Biology*, 10(11).

[Le Roch et al., 2004] Le Roch, K. G., Johnson, J. R., Florens, L., Zhou, Y., Santrosyan, A., Grainger, M., Yan, S. F., Williamson, K. C., Holder, A. a., Carucci, D. J., Yates, J. R., and Winzeler, E. a. (2004). Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome research*, 14:2308–2318.

[Lee et al., 2014] Lee, K. H., Kim, S. H., Kim, H. J., Kim, W., Lee, H. R., Jung, Y., Choi, J. H., Hong, K. Y., Jang, S. K., and Kim, K. T. (2014). AUF1 contributes to Cryptochrome1 mRNA degradation and rhythmic translation. *Nucleic Acids Research*, 42(6):3590–3606.

[Lee et al., 2012] Lee, K.-H., Woo, K.-C., Kim, D.-Y., Kim, T.-D., Shin, J., Park, S. M., Jang, S. K., and Kim, K.-T. (2012). Rhythmic Interaction between Period1 mRNA and hnRNP Q Leads to Circadian Time-Dependent Translation. *Molecular and Cellular Biology*, 32(3):717–728.

[Lee et al., 2011] Lee, M. V., Topper, S. E., Hubler, S. L., Hose, J., Wenger, C. D., Coon, J. J., and Gasch, A. P. (2011). A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Molecular Systems Biology*, 7.

[Li et al., 2013] Li, D. Q., Pakala, S. B., Reddy, S. D. N., Peng, S., Balasenthil, S., Deng, C. X., Lee, C. C., Rea, M. A., and Kumar, R. (2013). Metastasis-associated protein 1 is

an integral component of the circadian molecular machinery. *Nature Communications*, 4.

[Li et al., 2004] Li, F., Long, T., Lu, Y., Ouyang, Q., and Tang, C. (2004). The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14):4781–4786.

[Li et al., 2014] Li, H., Zhang, Y., Ströse, A., Tedesco, D., Gurova, K., and Selivanova, G. (2014). Integrated high-throughput analysis identifies Sp1 as a crucial determinant of p53-mediated apoptosis. *Cell Death and Differentiation*, 21(9):1493–1502.

[Liao et al., 2014] Liao, Y., Smyth, G. K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

[Lin, 2010] Lin, S. (2010). Rank aggregation methods.

[Liu et al., 2005] Liu, C., Bai, B., Skogerbø, G., Cai, L., Deng, W., Zhang, Y., Bu, D., Zhao, Y., and Chen, R. (2005). NONCODE: An integrated knowledge database of non-coding RNAs. *Nucleic Acids Research*, 33(DATABASE ISS.).

[Liu et al., 2016] Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance.

[Liu-Yesucevitz et al., 2011] Liu-Yesucevitz, L., Bassell, G. J., Gitler, A. D., Hart, A. C., Klann, E., Richter, J. D., Warren, S. T., and Wolozin, B. (2011). Local RNA Translation at the Synapse and in Disease. *Journal of Neuroscience*, 31(45):16086–16093.

[Love et al., 2014] Love, M. I., Huber, W., Anders, S., Lönnstedt, I., Speed, T., Robinson, M., Smyth, G., McCarthy, D., Chen, Y., Smyth, G., Anders, S., Huber, W., Zhou, Y.-H., Xia, K., Wright, F., Wu, H., Wang, C., Wu, Z., Hardcastle, T., Kelly, K., Wiel, M. V. D., Leday, G., Pardo, L., Rue, H., Vaart, A. V. D., Wieringen, W. V., Boer, J., Huber, W., Sültmann, H., Wilmer, F., von Heydebreck, A., Haas, S., Korn, B., Gunawan, B., Vente, A., Füzesi, L., Vingron, M., Poustka, A., Gentleman, R., Carey, V., Bates, D., 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, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., Zhang, J., McCullagh, P., Nelder, J., Hansen, K., Irizarry, R., Wu, Z., Risso, D., Schwartz, K., Sherlock, G., Dudoit, S., Smyth, G., Bottomly, D., Walter, N., Hunter, J., Darakjian, P., Kawane, S., Buck, K., Searles, R., Mooney, M., McWeeney, S., Hitzemann, R., Pickrell, J., Marioni, J., Pai, A., Degner, J., Engelhardt, B., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., Pritchard, J., Hastie, T., Tibshirani, R., Friedman, J., Bi, Y., Davuluri, R., Feng, J., Meyer, C., Wang, Q., Liu, J., Liu, X., Zhang, Y., Benjamini, Y., Hochberg, Y., Bourgon, R., Gentleman, R., Huber, W., McCarthy, D., Smyth, G., Li, J., Tibshirani, R., Cook, R., Hammer, P., Banck, M., Amberg, R., Wang, C., Petznick, G., Luo, S., Khrebtukova, I., Schroth, G., Beyerlein, P., Beutler, A., Frazee, A., Langmead, B., Leek, J., Trapnell, C., Hendrickson, D.,

Sauvageau, M., Goff, L., Rinn, J., Pachter, L., Glaus, P., Honkela, A., Rattray, M., Anders, S., Reyes, A., Huber, W., Sammeth, M., Robinson, M., McCarthy, D., Smyth, G., Zhou, X., Lindsay, H., Robinson, M., Leng, N., Dawson, J., Thomson, J., Ruotti, V., Rissman, A., Smits, B., Haag, J., Gould, M., Stewart, R., Kendziorski, C., Law, C., Chen, Y., Shi, W., Smyth, G., Hubert, L., Arabie, P., Witten, D., Irizarry, R., Wu, Z., Jaffee, H., Asangani, I., Dommeti, V., Wang, X., Malik, R., Cieslik, M., Yang, R., Escara-Wilke, J., Wilder-Romans, K., Dhanireddy, S., Engelke, C., Iyer, M., Jing, X., Wu, Y.-M., Cao, X., Qin, Z., Wang, S., Feng, F., Chinnaiyan, A., Ross-Innes, C., Stark, R., Teschendorff, A., Holmes, K., Ali, H., Dunning, M., Brown, G., Gojis, O., Ellis, I., Green, A., Ali, S., Chin, S.-F., Palmieri, C., Caldas, C., Carroll, J., Robinson, D., Chen, W., Storey, J., Gresham, D., McMurdie, P., Holmes, S., Vasquez, J., Hon, C., Vanselow, J., Schlosser, A., Siegel, T., Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y., Wei, W., Cox, D., Reid, N., Robinson, M., Smyth, G., Pawitan, Y., Armijo, L., Di, Y., Schafer, D., Cumbie, J., Chang, J., Abramowitz, M., Stegun, I., Newton, M., Kendziorski, C., Richmond, C., Blattner, F., Tsui, K., Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., Vingron, M., Durbin, B., Hardin, J., Hawkins, D., Rocke, D., Friedman, J., Hastie, T., Tibshirani, R., Cule, E., Vineis, P., Iorio, M. D., Cook, R., Weisberg, S., Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., Carey, V., Anders, S., Pyl, P., Huber, W., Delhomme, N., Padioleau, I., Furlong, E., Steinmetz, L., Liao, Y., Smyth, G., Shi, W., Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

[Lowrey and Takahashi, 2011] Lowrey, P. L. and Takahashi, J. S. (2011). Genetics of circadian rhythms in mammalian model organisms. *Advances in Genetics*, 74:175–230.

[Lück et al., 2014] Lück, S., Thurley, K., Thaben, P. F., and Westermark, P. O. (2014). Rhythmic degradation explains and unifies circadian transcriptome and proteome data. *Cell Reports*, 9(2):741–751.

[Ma et al., 2013] Ma, Y. T., Luo, H., Guan, W. J., Zhang, H., Chen, C., Wang, Z., and Li, J. D. (2013). O-GlcNAcylation of BMAL1 regulates circadian rhythms in NIH3T3 fibroblasts. *Biochemical and Biophysical Research Communications*, 431(3):382–387.

[Maeda et al., 2005] Maeda, M., Johnson, K. R., and Wheelock, M. J. (2005). Cadherin switching: essential for behavioral but not morphological changes during an epithelium-to-mesenchyme transition. *Journal of cell science*, 118(Pt 5):873–887.

[Maetschke et al., 2014] Maetschke, S. R., Madhamshettiwar, P. B., Davis, M. J., and Ragan, M. A. (2014). Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in Bioinformatics*, 15(2):195–211.

[Maier et al., 2009] Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Letters*, 583(24):3966–3973.

[Maier et al., 2011] Maier, T., Schmidt, A., Güell, M., Kühner, S., Gavin, A. C., Aebersold, R., and Serrano, L. (2011). Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Molecular Systems Biology*, 7.

[Majmundar et al., 2010] Majmundar, A. J., Wong, W. J., and Simon, M. C. (2010). Hypoxia-Inducible Factors and the Response to Hypoxic Stress.

[Manu et al., 2009] Manu, Surkova, S., Spirov, A. V., Gursky, I. V., Janssens, H., Kim, A. R., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M., and Reinitz, J. (2009). Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation. *PLoS Biology*, 7(3):0591–0603.

[Marbach et al., 2012] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Aderhold, A., Bonneau, R., Chen, Y., et al. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796.

[Marbach et al., 2010a] Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010a). Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA*, 107(14):6286–6291.

[Marbach et al., 2010b] Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010b). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14):6286–6291.

[Marbach et al., 2009] Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of computational biology*, 16(2):229–239.

[Margolin et al., 2006] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(SUPPL.1).

[Masri et al., 2012] Masri, S., Zocchi, L., Katada, S., Mora, E., and Sassone-Corsi, P. (2012). The circadian clock transcriptional complex: Metabolic feedback intersects with epigenetic control. *Annals of the New York Academy of Sciences*, 1264(1):103–109.

[Mathelier et al., 2016] Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2016). JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–D115.

[Mauvoisin et al., 2015] Mauvoisin, D., Dayon, L., Gachon, F., and Kussmann, M. (2015). Proteomics and circadian rhythms: It's all about signaling!

[Mauvoisin et al., 2014] Mauvoisin, D., Wang, J., Jouffe, C., Martin, E., Atger, F., Waridel, P., Quadroni, M., Gachon, F., and Naef, F. (2014). Circadian clock-dependent and -independent rhythmic proteomes implement distinct diurnal functions in mouse liver. *Proceedings of the National Academy of Sciences*, 111(1):167–172.

[McLeay and Bailey, 2010a] McLeay, R. C. and Bailey, T. L. (2010a). Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, 11.

[McLeay and Bailey, 2010b] McLeay, R. C. and Bailey, T. L. (2010b). Motif enrichment analysis: a unified framework and an evaluation on chip data. *BMC bioinformatics*, 11(1):165.

[McShane et al., 2016] McShane, E., Sin, C., Zauber, H., Wells, J. N., Donnelly, N., Wang, X., Hou, J., Chen, W., Storchova, Z., Marsh, J. A., Valleriani, A., and Selbach, M. (2016). Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell*, 167(3):803–815.e21.

[Mei et al., 2017] Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., Zhu, M., Wu, J., Shi, X., Taing, L., Liu, T., Brown, M., Meyer, C. A., and Liu, X. S. (2017). Cistrome Data Browser: A data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Research*, 45(D1):D658–D662.

[Mélykúti et al., 2010] Mélykúti, B., August, E., Papachristodoulou, A., and El-Samad, H. (2010). Discriminating between rival biochemical network models: three approaches to optimal experiment design. *BMC systems biology*, 4(1):38.

[Mendez and Richter, 2001] Mendez, R. and Richter, J. D. (2001). Translational control by CPEB: A means to the end.

[Menet et al., 2012] Menet, J. S., Rodriguez, J., Abruzzi, K. C., and Rosbash, M. (2012). Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. *eLife*, 2012(1).

[Menger et al., 2007] Menger, G. J., Allen, G. C., Neuendorff, N., Nahm, S.-S., Thomas, T. L., Cassone, V. M., and Earnest, D. J. (2007). Circadian profiling of the transcriptome in NIH/3T3 fibroblasts: comparison with rhythmic gene expression in SCN2.2 cells and the rat SCN. *Physiological Genomics*, 29(3):280–289.

[Meyer et al., 2007] Meyer, P. E., Kontos, K., Lafitte, F., and Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *Eurasip Journal on Bioinformatics and Systems Biology*, 2007.

[Meyer et al., 2008] Meyer, P. E., Lafitte, F., and Bontempi, G. (2008). Minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9.

[Miettinen et al., 1994] Miettinen, P. J., Ebner, R., Lopez, A. R., and Derynck, R. (1994). TGF-beta induced transdifferentiation of mammary epithelial cells to mesenchymal cells: involvement of type I receptors. *The Journal of cell biology*, 127(6 Pt 2):2021–36.

[Miller et al., 2007] Miller, B. H., McDearmon, E. L., Panda, S., Hayes, K. R., Zhang, J., Andrews, J. L., Antoch, M. P., Walker, J. R., Esser, K. A., Hogenesch, J. B., and Takahashi, J. S. (2007). Circadian and CLOCK-controlled regulation of the mouse transcriptome and cell proliferation. *Proceedings of the National Academy of Sciences*, 104(9):3342–3347.

[Mittal et al., 2009] Mittal, N., Roy, N., Babu, M. M., and Janga, S. C. (2009). Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(48):20300–5.

[Mootha et al., 2003] Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273.

[Morf et al., 2012] Morf, J., Rey, G., Schneider, K., Stratmann, M., Fujita, J., Naef, F., and Schibler, U. (2012). Cold-Inducible RNA-Binding Protein Modulates Circadian Gene Expression Posttranscriptionally. *Science*, 338(6105):379–383.

[Murata and Wharton, 1995] Murata, Y. and Wharton, R. P. (1995). Binding of pumilio to maternal hunchback mRNA is required for posterior patterning in drosophila embryos. *Cell*, 80(5):747–756.

[Nagaraj et al., 2011] Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular Systems Biology*, 7(548):548.

[Nagoshi et al., 2005] Nagoshi, E., Brown, S. A., Dibner, C., Kornmann, B., and Schibler, U. (2005). Circadian gene expression in cultured cells.

[Nagoshi et al., 2004] Nagoshi, E., Saini, C., Bauer, C., Laroche, T., Naef, F., and Schibler, U. (2004). Circadian gene expression in individual fibroblasts: Cell-autonomous and self-sustained oscillators pass time to daughter cells. *Cell*, 119(5):693–705.

[Nakahata et al., 2008] Nakahata, Y., Kaluzova, M., Grimaldi, B., Sahar, S., Hirayama, J., Chen, D., Guarente, L. P., and Sassone-Corsi, P. (2008). The NAD+-Dependent Deacetylase SIRT1 Modulates CLOCK-Mediated Chromatin Remodeling and Circadian Control. *Cell*, 134(2):329–340.

[Nieto et al., 2016] Nieto, M. ., Huang, R. Y., Jackson, R. A., and Thiery, J. P. (2016). EMT: 2016.

[Niu et al., 2016] Niu, M., Dai, Z., Lawrence, N., and Becker, K. (2016). Spatio-temporal gaussian processes modeling of dynamical systems in systems biology. *arXiv preprint arXiv:1610.05163*.

[Nusslein-Volhard et al., 1987] Nusslein-Volhard, C., Frohnhofer, H., and Lehmann, R. (1987). Determination of anteroposterior polarity in Drosophila. *Science*, 238(4834):1675–1681.

[Nüsslein-volhard and Wieschaus, 1980] Nüsslein-volhard, C. and Wieschaus, E. (1980). Mutations affecting segment number and polarity in drosophila. *Nature*, 287(5785):795–801.

[Onishi and Kawano, 2012] Onishi, Y. and Kawano, Y. (2012). Rhythmic binding of Topoisomerase i impacts on the transcription of Bmal1 and circadian period. *Nucleic Acids Research*, 40(19):9482–9492.

[Panagiotis Zalmas et al., 2008] Panagiotis Zalmas, L., Zhao, X., Graham, A. L., Fisher, R., Reilly, C., Coutts, A. S., and La Thangue, N. B. (2008). DNA-damage response control of E2F7 and E2F8. *EMBO Reports*, 9(3):252–259.

[Panda et al., 2002] Panda, S., Antoch, M. P., Miller, B. H., Su, A. I., Schook, A. B., Straume, M., Schultz, P. G., Kay, S. A., Takahashi, J. S., and Hogenesch, J. B. (2002). Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, 109(3):307–320.

[Partch et al., 2014] Partch, C. L., Green, C. B., and Takahashi, J. S. (2014). Molecular architecture of the mammalian circadian clock.

[Peinado et al., 2003] Peinado, H., Quintanilla, M., and Cano, A. (2003). Transforming growth factor β-1 induces snail transcription factor in epithelial cell lines mechanisms for epithelial mesenchymal transitions. *Journal of Biological Chemistry*, 278(23):21113–21123.

[Perez et al., 2012] Perez, R. E., Jansen, P. W., and Martins, J. R. R. A. (2012). pyOpt: A Python-based object-oriented framework for nonlinear constrained optimization. *Structures and Multidisciplinary Optimization*, 45(1):101–118.

[Peshkin et al., 2015] Peshkin, L., Wühr, M., Pearl, E., Haas, W., Freeman, R. M., Gerhart, J. C., Klein, A. M., Horb, M., Gygi, S. P., and Kirschner, M. W. (2015). On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development. *Developmental Cell*, 35(3):383–394.

[Piek et al., 1999] Piek, E., Moustakas, a., Kurisaki, a., Heldin, C. H., and ten Dijke, P. (1999). TGF-(beta) type I receptor/ALK-5 and Smad proteins mediate epithelial to mesenchymal transdifferentiation in NMuMG breast epithelial cells. *Journal of cell science*, 112 (Pt 2:4557–4568.

[Pizarro et al., 2013] Pizarro, A., Hayer, K., Lahens, N. F., and Hogenesch, J. B. (2013). CircaDB: a database of mammalian circadian gene expression profiles. *Nucleic Acids Research*, 41(Database issue):D1009–13.

[Qin et al., 2007] Qin, X. L., Ahn, S., Speed, T. P., and Rubin, G. M. (2007). Global analyses of mRNA translational control during early Drosophila embryogenesis. *Genome Biology*, 8(4):R63.

[Ramos et al., 2013] Ramos, A. D., Diaz, A., Nellore, A., Delgado, R. N., Park, K. Y., Gonzales-Roybal, G., Oldham, M. C., Song, J. S., and Lim, D. A. (2013). Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. *Cell Stem Cell*, 12(5):616–628.

[Raue et al., 2009] Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929.

[Raue et al., 2013] Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B. D., Theis, F. J., et al. (2013). Lessons learned from quantitative dynamical modeling in systems biology. *PloS one*, 8(9):e74335.

[Ray et al., 2013] Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L. H., Dale, R. K., Smith, S. A., Yarosh, C. A., Kelly, S. M., Nabet, B., Mecenas, D., Li, W., Laishram, R. S., Qiao, M., Lipshitz, H. D., Piano, F., Corbett, A. H., Carstens, R. P., Frey, B. J., Anderson, R. A., Lynch, K. W., Penalva, L. O., Lei, E. P., Fraser, A. G., Blencowe, B. J., Morris, Q. D., and Hughes, T. R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177.

[Reddy and Rey, 2014] Reddy, A. B. and Rey, G. (2014). Metabolic and Nontranscriptional Circadian Clocks: Eukaryotes. *Annual Review of Biochemistry*, 83(1):165–189.

[Reinitz and Sharp, 1995] Reinitz, J. and Sharp, D. H. (1995). Mechanism of eve stripe formation. *Mechanisms of Development*, 49(1-2):133–158.

[Relógio et al., 2011] Relógio, A., Westermark, P. O., Wallach, T., Schellenberg, K., Kramer, A., and Herzel, H. (2011). Tuning the mammalian circadian clock: Robust synergy of two loops. *PLoS Computational Biology*, 7(12).

[Rey et al., 2011] Rey, G., Cesbron, F., Rougemont, J., Reinke, H., Brunner, M., and Naef, F. (2011). Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. *PLoS Biology*, 9(2).

[Richards et al., 2015] Richards, E. J., Zhang, G., Li, Z.-P., Permuth-Wey, J., Challa, S., Li, Y., Kong, W., Dan, S., Bui, M., Coppola, D., et al. (2015). Long non-coding

rnas regulated by tgfβ: lncrna-hit mediated tgfβ-induced epithelial to mesenchymal transition in mammary epithelia. *Journal of Biological Chemistry*, pages jbc–M114.

[Rivera-Pomar et al., 1996] Rivera-Pomar, R., Niessing, D., Schmidt-Ott, U., Gehring, W. J., and Jackle, H. (1996). RNA binding and translational suppression by bicoid. *Nature*, 379(6567):746–749.

[Robles et al., 2014] Robles, M. S., Cox, J., and Mann, M. (2014). In-Vivo Quantitative Proteomics Reveals a Key Contribution of Post-Transcriptional Mechanisms to the Circadian Regulation of Liver Metabolism. *PLoS Genet*, 10(1):e1004047.

[Rodriguez-Fernandez et al., 2013] Rodriguez-Fernandez, M., Rehberg, M., Kremling, A., and Banga, J. R. (2013). Simultaneous model discrimination and parameter estimation in dynamic models of cellular systems. *BMC systems biology*, 7:76.

[Roignant and Soller, 2017] Roignant, J. Y. and Soller, M. (2017). m6A in mRNA: An Ancient Mechanism for Fine-Tuning Gene Expression.

[Roundtree et al., 2017] Roundtree, I. A., Evans, M. E., Pan, T., and He, C. (2017). Dynamic RNA Modifications in Gene Expression Regulation.

[Sahu et al., 2015] Sahu, S. K., Garding, A., Tiwari, N., Thakurela, S., Toedling, J., Gebhard, S., Ortega, F., Schmarowski, N., Berninger, B., Nitsch, R., Schmidt, M., and Tiwari, V. K. (2015). JNK-dependent gene regulatory circuitry governs mesenchymal fate. *The EMBO Journal*, 34(16):2162–2181.

[Saito and Rehmsmeier, 2015] Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3).

[Santillan, 2008] Santillan, M. (2008). On the Use of the Hill Functions in Mathematical Models of Gene Regulatory Networks. *Mathematical Modelling of Natural Phenomena*, 3(2):85–97.

[Sayols et al., 2016] Sayols, S., Scherzinger, D., and Klein, H. (2016). dupRadar: A Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. *BMC Bioinformatics*, 17(1).

[Schaffter et al., 2011a] Schaffter, T., Marbach, D., and Floreano, D. (2011a). {GeneNetWeaver}: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270.

[Schaffter et al., 2011b] Schaffter, T., Marbach, D., and Floreano, D. (2011b). GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270.

[Schick et al., 2016] Schick, S., Becker, K., Thakurela, S., Fournier, D., Hampel, M. H., Legewie, S., and Tiwari, V. K. (2016). Identifying Novel Transcriptional Regulators with Circadian Expression. *Molecular and Cellular Biology*, 36(4):545–558.

[Schick et al., 2015] Schick, S., Fournier, D., Thakurela, S., Sahu, S. K., Garding, A., and Tiwari, V. K. (2015). Dynamics of chromatin accessibility and epigenetic state in response to UV damage. *Journal of Cell Science*, 128(23):4380–4394.

[Schrimpf et al., 2009] Schrimpf, S. P., Weiss, M., Reiter, L., Ahrens, C. H., Jovanovic, M., Malmstrom, J., Brunner, E., Mohanty, S., Lercher, M. J., Hunziker, P. E., Aebersold, R., von Mering, C., and Hengartner, M. O. (2009). Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes. *PLoS Biology*, 7(3):616–627.

[Schwanhäusser et al., 2011] Schwanhäusser, B., Busse, D., and Li, N. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342.

[Serban and Hindmarsh, 2005] Serban, R. and Hindmarsh, A. C. (2005). CVODES: The Sensitivity-Enabled ODE Solver in SUNDIALS. In *Volume 6: 5th International Conference on Multibody Systems, Nonlinear Dynamics, and Control, Parts A, B, and C*, volume 2005, pages 257–269.

[Shea and Ackers, 1985] Shea, M. A. and Ackers, G. K. (1985). The or control system of bacteriophage lambda: A physical-chemical model for gene regulation. *Journal of molecular biology*, 181(2):211–230.

[Shirakihara et al., 2007] Shirakihara, T., Saitoh, M., and Miyazono, K. (2007). Differential regulation of epithelial and mesenchymal markers by deltaEF1 proteins in epithelial mesenchymal transition induced by TGF-beta. *Molecular biology of the cell*, 18(9):3533–3544.

[Siemens et al., 2011] Siemens, H., Jackstadt, R., Hünten, S., Kaller, M., Menssen, A., Götz, U., and Hermeking, H. (2011). miR-34 and SNAIL form a double-negative feedback loop to regulate epithelial-mesenchymal transitions. *Cell Cycle*, 10(24):4256–4271.

[Skrypek et al., 2017] Skrypek, N., Goossens, S., De Smedt, E., Vandamme, N., and Berx, G. (2017). Epithelial-to-Mesenchymal Transition: Epigenetic Reprogramming Driving Cellular Plasticity.

[Smibert et al., 1996] Smibert, C. A., Wilson, J. E., Kerr, K., and Macdonald, P. M. (1996). smaug protein represses translation of unlocalized nanos mRNA in the Drosophila embryo. *Genes and Development*, 10(20):2600–2609.

[Sonenberg and Hinnebusch, 2009] Sonenberg, N. and Hinnebusch, A. G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets.

[Spellman et al., 1998] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive Identification of Cell Cycleregulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, 9(12):3273 –3297.

[Storch et al., 2002] Storch, K.-F., Lipan, O., Leykin, I., Viswanathan, N., Davis, F. C., Wong, W. H., and Weitz, C. J. (2002). Extensive and divergent circadian gene expression in liver and heart. *Nature*, 417(6884):78–83.

[Strasen et al., 2018] Strasen, J., Sarma, U., Jentsch, M., Bohn, S., Sheng, C., Horbelt, D., Knaus, P., Legewie, S., and Loewer, A. (2018). Cell-specific responses to the cytokine tgfβ are determined by variability in protein levels. *Molecular systems biology*, 14(1):e7733.

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

[Sysoev et al., 2016] Sysoev, V. O., Fischer, B., Frese, C. K., Gupta, I., Krijgsveld, J., Hentze, M. W., Castello, A., and Ephrussi, A. (2016). Global changes of the RNA-bound proteome during the maternal-to-zygotic transition in Drosophila. *Nature Communications*, 7.

[Tadros and Lipshitz, 2009] Tadros, W. and Lipshitz, H. D. (2009). The maternal-to-zygotic transition: a play in two acts. *Development*, 136(18):3033–3042.

[Takarada et al., 2012] Takarada, T., Kodama, a., Hotta, S., Mieda, M., Shimba, S., Hinoi, E., and Yoneda, Y. (2012). Clock Genes Influence Gene Expression in Growth Plate and Endochondral Ossification in Mice. *Journal of Biological Chemistry*, 287(43):36081–36095.

[Tan et al., 2014] Tan, E.-J., Kahata, K., Idas, O., Thuault, S., Heldin, C.-H., and Moustakas, a. (2014). The high mobility group A2 protein epigenetically silences the Cdh1 gene during epithelial-to-mesenchymal transition. *Nucleic Acids Research*, 43(1):162–178.

[Tan et al., 2012] Tan, E.-J., Thuault, S., Caja, L., Carletti, T., Heldin, C.-H., and Moustakas, A. (2012). Regulation of transcription factor twist expression by the dna architectural protein high mobility group a2 during epithelial-to-mesenchymal transition. *Journal of Biological Chemistry*, 287(10):7134–7145.

[Taniguchi et al., 2010] Taniguchi, Y., Choi, P. J., Li, G. W., Chen, H., Babu, M., Hearn, J., Emili, A., and Sunney Xie, X. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538.

[Thaben and Westermark, 2014] Thaben, P. F. and Westermark, P. O. (2014). Detecting rhythms in time series with rain. *Journal of Biological Rhythms*, 29(6):391–400.

[Thuault et al., 2008] Thuault, S., Tan, E. J., Peinado, H., Cano, A., Heldin, C. H., and Moustakas, A. (2008). HMGA2 and Smads co-regulate SNAIL1 expression during

induction of epithelial-to-mesenchymal transition. *Journal of Biological Chemistry*, 283(48):33437–33446.

[Thuault et al., 2006] Thuault, S., Valcourt, U., Petersen, M., Manfioletti, G., Heldin, C. H., and Moustakas, A. (2006). Transforming growth factor-β employs HMGA2 to elicit epithelial-mesenchymal transition. *Journal of Cell Biology*, 174(2):175–183.

[Tiwari et al., 2013] Tiwari, N., Tiwari, V. K., Waldmeier, L., Balwierz, P. J., Arnold, P., Pachkov, M., Meyer-Schaller, N., Schübeler, D., VanNimwegen, E., and Christofori, G. (2013). Sox4 Is a Master Regulator of Epithelial-Mesenchymal Transition by Controlling Ezh2 Expression and Epigenetic Reprogramming. *Cancer Cell*, 23(6):768–783.

[Tovin et al., 2012] Tovin, A., Alon, S., Ben-Moshe, Z., Mracek, P., Vatine, G., Foulkes, N. S., Jacob-Hirsch, J., Rechavi, G., Toyama, R., Coon, S. L., Klein, D. C., Eisenberg, E., and Gothilf, Y. (2012). Systematic Identification of Rhythmic Genes Reveals camk1gb as a New Element in the Circadian Clockwork. *PLoS Genetics*, 8(12).

[Trapnell et al., 2009] Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111.

[Ueda et al., 2005] Ueda, H. R., Hayashi, S., Chen, W., Sano, M., Machida, M., Shigeyoshi, Y., Iino, M., and Hashimoto, S. (2005). System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nature Genetics*, 37(2):187–192.

[Ugrankar et al., 2015] Ugrankar, R., Berglund, E., Akdemir, F., Tran, C., Kim, M. S., Noh, J., Schneider, R., Ebert, B., and Graff, J. M. (2015). Drosophila glucome screening identifies Ck1alpha as a regulator of mammalian glucose metabolism. *Nature Communications*, 6.

[Valcourt, 2005] Valcourt, U. (2005). TGF- and the Smad Signaling Pathway Support Transcriptomic Reprogramming during Epithelial-Mesenchymal Cell Transition. *Molecular Biology of the Cell*, 16(4):1987–2002.

[Valekunja et al., 2013] Valekunja, U. K., Edgar, R. S., Oklejewicz, M., van der Horst, G. T. J., O'Neill, J. S., Tamanini, F., Turner, D. J., and Reddy, A. B. (2013). Histone methyltransferase MLL3 contributes to genome-scale circadian transcription. *Proceedings of the National Academy of Sciences*, 110(4):1554–1559.

[Valencia-Sanchez et al., 2006] Valencia-Sanchez, M. A., Liu, J., Hannon, G. J., and Parker, R. (2006). Control of translation and mrna degradation by mirnas and sirnas. *Genes & development*, 20(5):515–524.

[Vance and Ponting, 2014] Vance, K. W. and Ponting, C. P. (2014). Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends in genetics : TIG*, 30(8):348–355.

[Villaverde et al., 2013] Villaverde, A., Ross, J., and Banga, J. (2013). Reverse Engineering Cellular Networks with Information Theoretic Methods. *Cells*, 2(2):306–329.

[Villaverde and Banga, 2013] Villaverde, A. F. and Banga, J. R. (2013). Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *Journal of The Royal Society Interface*, 11(91):20130505–20130505.

[Villaverde et al., 2016] Villaverde, A. F., Becker, K., and Banga, J. R. (2016). Premer: parallel reverse engineering of biological networks with information theory. In *International Conference on Computational Methods in Systems Biology*, pages 323–329. Springer.

[Villaverde et al., 2017] Villaverde, A. F., Becker, K., and Banga, J. R. (2017). PREMER: a Tool to Infer Biological Networks.

[Villaverde et al., 2014] Villaverde, A. F., Ross, J., Morán, F., and Banga, J. R. (2014). MIDER: Network inference with mutual information distance and entropy reduction. *PLoS ONE*, 9(5).

[Vollmers et al., 2012] Vollmers, C., Schmitz, R. J., Nathanson, J., Yeo, G., Ecker, J. R., and Panda, S. (2012). Circadian oscillations of protein-coding and regulatory RNAs in a highly dynamic mammalian liver epigenome. *Cell Metabolism*, 16(6):833–845.

[Wang et al., 2010] Wang, D. O., Martin, K. C., and Zukin, R. S. (2010). Spatially restricting gene expression by local translation at synapses.

[Wang et al., 2014] Wang, Q., Huang, J., Sun, H., Liu, J., Wang, J., Wang, Q., Qin, Q., Mei, S., Zhao, C., Yang, X., Liu, X. S., and Zhang, Y. (2014). CR Cistrome: A ChIP-Seq database for chromatin regulators and histone modification linkages in human and mouse. *Nucleic Acids Research*, 42(D1).

[Westermark and Herzel, 2013] Westermark, P. and Herzel, H. (2013). Mechanism for 12 Hr Rhythm Generation by the Circadian Clock. *Cell Reports*, 3(4):1228–1238.

[Williams et al., 2001] Williams, J. A., Su, H. S., Bernards, A., Field, J., and Sehgal, A. (2001). A circadian output in Drosophila mediated by neurofibromatosis-1 and Ras/MAPK. *Science*, 293(5538):2251–2256.

[Woo et al., 2010] Woo, K.-C., Ha, D.-C., Lee, K.-H., Kim, D.-Y., Kim, T.-D., and Kim, K.-T. (2010). Circadian Amplitude of Cryptochrome 1 Is Modulated by mRNA Stability Regulation via Cytoplasmic hnRNP D Oscillation. *Molecular and Cellular Biology*, 30(1):197–205.

[Woo et al., 2009] Woo, K. C., Kim, T. D., Lee, K. H., Kim, D. Y., Kim, W., Lee, K. Y., and Kim, K. T. (2009). Mouse period 2 mRNA circadian oscillation is modulated by PTB-mediated rhythmic mRNA degradation. *Nucleic Acids Research*, 37(1):26–37.

[Wu et al., 2008] Wu, G., Nie, L., and Zhang, W. (2008). Integrative analyses of post-transcriptional regulation in the yeast Saccharomyces cerevisiae using transcriptomic and proteomic data. *Current Microbiology*, 57(1):18–22.

[Wuarin and Schibler, 1990] Wuarin, J. and Schibler, U. (1990). Expression of the liver-enriched transcriptional activator protein DBP follows a stringent circadian rhythm. *Cell*, 63(6):1257–1266.

[Xiao et al., 2013] Xiao, J., Zhou, Y., Lai, H., Lei, S., Chi, L. H., and Mo, X. (2013). Transcription factor NF-Y is a functional regulator of the transcription of core clock gene Bmal1. *Journal of Biological Chemistry*, 288(44):31930–31936.

[Yagita et al., 2010] Yagita, K., Horie, K., Koinuma, S., Nakamura, W., Yamanaka, I., Urasaki, A., Shigeyoshi, Y., Kawakami, K., Shimada, S., Takeda, J., and Uchiyama, Y. (2010). Development of the circadian oscillator during differentiation of mouse embryonic stem cells in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 107(8):3846–3851.

[Yan et al., 2008] Yan, J., Wang, H., Liu, Y., and Shao, C. (2008). Analysis of Gene Regulatory Networks in the Mammalian Circadian Rhythm. *PLoS Comput Biol*, 4(10):e1000193.

[Yang and Su, 2010] Yang, R. and Su, Z. (2010). Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics*, 26(12).

[Yates et al., 2016] Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716.

[Yu et al., 2010] Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, 26(7):976–978.

[Yuret and de la Maza, 1993] Yuret, D. and de la Maza, M. (1993). Dynamic Hill Climbing: Overcoming the Limitations of Optimization Techniques. *Tainn'93*, pages 208–212.

[Zeileis and Grothendieck, 2005] Zeileis, A. and Grothendieck, G. (2005). zoo : An S3 Class and Methods for Indexed Totally Ordered Observations. *Wirtschaftsuniversität Wien*, (April).

[Zhang et al., 2012] Zhang, J., Liang, Q., Lei, Y., Yao, M., Li, L., Gao, X., Feng, J., Zhang, Y., Gao, H., Liu, D. X., Lu, J., and Huang, B. (2012). SOX4 induces epithelial-mesenchymal transition and contributes to breast cancer progression. *Cancer Research*, 72(17):4597–4608.

[Zhang et al., 2014] Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E., and Hogenesch, J. B. (2014). A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of the National Academy of Sciences*, 111(45):16219–16224.

[Zhao et al., 2014] Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*, 9(1).

[Zhou et al., 2002] Zhou, L., Zhu, C., Luo, K., Li, Y., Pi, H., Yuan, W., Wang, Y., Huang, C., Liu, M., and Wu, X. (2002). Identification and characterization of two novel zinc finger genes, ZNF359 and ZFP28, in human development. *Biochemical and Biophysical Research Communications*, 295(4):862–868.

[Zhu, 2011] Zhu, W. (2011). *Mechanisms and functional roles of nuclear respiratory factor 1 (NRF1) binding sites in the human genome*. PhD thesis, University of Pittsburgh.

Disclaimer

Parts of this PhD-thesis were carried out in collaboration with multiple cooperation partners. In the following section the contribution of each of the cooperation partners is summarized.

Chapter 2: NIH3T3 luminescence data and expression data was generated by SS in the group of VT at the IMB. GO term and motif analysis was carried out by ST in the group of VT at the IMB. CHiP data was processed by DF in the group of VT at the IMB. Zfp28 and Leo1 knock-down experiments were generated by SS.

Chapter 3: All knock-down experiments were performed by SS.

Chapter 4: qPCR expression data of TGFb-stimulated NMuMG wild-type or knock-down cells were generated by PS, SS, AG, and SSu in the group of Vijay Tiwari at the IMB Mainz. Western blot measurement of pSmad2 was carried out by AG.

Chapter 6: Experimental data for mRNA and protein expression during *Drosophila* development was obtained by NC, AB in the group of FB at the IMB Mainz. Data processing was carried out by MD, and SSP. Hrb98DE knock-down flies were generated by ND in the lab of JYR at the IMB in Mainz. Analysis of RNA-sequencing data was provided by SSP.

Publications

Becker, K., Balsa-Canto, E., Cicin-Sain, D., Hoermann, A., Janssens, H., Banga, J. R., and Jaeger, J. (2013). Reverse-Engineering Post-Transcriptional Regulation of Gap Genes in Drosophila melanogaster. *PLoS Computational Biology*, 9(10)

Schick, S., Becker, K., Thakurela, S., Fournier, D., Hampel, M. H., Legewie, S., and Tiwari, V. K. (2016). Identifying Novel Transcriptional Regulators with Circadian Expression. *Molecular and Cellular Biology*, 36(4):545–558

Villaverde, A. F., Becker, K., and Banga, J. R. (2016). Premer: parallel reverse engineering of biological networks with information theory. In *International Conference on Computational Methods in Systems Biology*, pages 323–329. Springer

Villaverde, A. F., Becker, K., and Banga, J. R. (2017). PREMER: a Tool to Infer Biological Networks

Niu, M., Dai, Z., Lawrence, N., and Becker, K. (2016). Spatio-temporal gaussian processes modeling of dynamical systems in systems biology. *arXiv preprint arXiv:1610.05163*

Acknowledgements

To all those who supported me during my time as a PhD-student, I want to express my deepest gratitude. In particular, I want to thank my supervisor, SL, for valuable feedback regarding all of my projects, as well as moral and financial support.

This thesis would have not been possible without the extensive and enjoyable collaboration with experimental researchers. Here I would particularly like to thank PS, SS, AG, and SSu for their contribution to the EMT project. Once again I need to thank SS for initiating and ambitiously pursuing our joint Clock project, which benefited greatly from the help by DF and ST. The success of the *Drosophila* project largely depended on the tremendous experimental groundwork established by AB and NCV, assisted by SSP and MD. I further want to thank ND who performed the experimental validation of our formulated hypothesis. As senior advisers to the *Drosophila* project, I would further like to express my gratitude to FB, JYR, and SL who guided our efforts in analysing this extremely rich dataset and always contributing professional and valuable feedback to the project.

In addition to our experimental collaborators, I would further like to thank former and current group members UA, SB, ME, CF, LH, MK, SL, TM, LR, and US for interesting discussions and providing an enjoyable working atmosphere. I further profited very much from joint projects carried out in collaboration with AV and JB. Thanks also to IK and once again SS for carefully proofreading my thesis.

Naturally, I would not be where I am today without the continuous support by close friends. In particular, props go out to IK, without whom I would have never found my way to the IMB, as well as TM and MW for being awesome friends. Finally I want to thank my Mother and Father for encouraging me to always pursue my goals, whatever they may be.