Methods 85 (2015) 62–74

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

ARMADA: Using motif activity dynamics to infer gene regulatory networks from gene expression data

Peter J. Pemberton-Ross, Mikhail Pachkov, Erik van Nimwegen *

Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland

ARTICLE INFO

Article history:
Received 23 March 2015
Received in revised form 22 June 2015
Accepted 23 June 2015
Available online 8 July 2015

Keywords:
Transcription regulation
Gene regulatory network
Transcription factor
Network inference
Auto-regressive models
Regulatory motif
Motif activity

ABSTRACT

Analysis of gene expression data remains one of the most promising avenues toward reconstructing genome-wide gene regulatory networks. However, the large dimensionality of the problem prohibits the fitting of explicit dynamical models of gene regulatory networks, whereas machine learning methods for dimensionality reduction such as clustering or principal component analysis typically fail to provide mechanistic interpretations of the reduced descriptions. To address this, we recently developed a general methodology called motif activity response analysis (MARA) that, by modeling gene expression patterns in terms of the activities of concrete regulators, accomplishes dramatic dimensionality reduction while retaining mechanistic biological interpretations of its predictions (Balwierz, 2014).

Here we extend MARA by presenting ARMADA, which models the activity dynamics of regulators across a time course, and infers the causal interactions between the regulators that drive the dynamics of their activities across time. We have implemented ARMADA as part of our ISMARA webserver, ismara.unibas.ch, allowing any researcher to automatically apply it to any gene expression time course. To illustrate the method, we apply ARMADA to a time course of human umbilical vein endothelial cells treated with TNF. Remarkably, ARMADA is able to reproduce the complex observed motif activity dynamics using a relatively small set of interactions between the key regulators in this system. In addition, we show that ARMADA successfully infers many of the key regulatory interactions known to drive this inflammatory response and discuss several novel interactions that ARMADA predicts. In combination with ISMARA, ARMADA provides a powerful approach to generating plausible hypotheses for the key interactions between regulators that control gene expression in any system for which time course measurements are available.

1. Introduction

Understanding the structure, dynamics, and functioning of the genome-wide regulatory networks that control gene expression is one of the central challenges in systems biology. Gene regulatory networks allow individual cells to respond and adapt to changes in their environments, and allow multi-cellular eukaryotes to express a single underlying genotype, shared by all their cells, into a large variety of phenotypically and functionally distinct cell types. More than half a century has passed since the discovery of the basic biophysical mechanism underlying gene regulation [2], and during this time much has been learned about the molecular players involved in gene regulation, and the specific mechanisms through which they act. Very roughly speaking, in the cells of multi-cellular eukaryotes there are hundreds of regulatory proteins and RNAs expressed that bind in a sequence-specific manner to short sequence motifs within the DNA and RNA. The binding constellations of these regulatory proteins determine the rates at which genes are being transcribed, the stability of mRNAs, and the rates at which they are being translated.

We not only understand the basic molecular mechanisms, in well-studied model organisms most of the molecular players are known as well, i.e. comprehensive lists of transcription factors TFs [3] and regulatory RNAs such as miRNAs [4] are available, and for many of these there is also information about their targets and their functioning. However, knowing the molecular players and understanding the molecular mechanisms involved does not mean that we understand how gene regulatory networks function *as systems*. For example, how the actions of regulatory genes are coordinated to maintain and stabilize cell identity is not understood. Similarly, although it has recently become clear that, given an appropriate perturbation in the expression of regulatory proteins, cells can be driven from one cell type to another [5], what

* Corresponding author at: Biozentrum, University of Basel, Basel, Switzerland.
E-mail addresses: peter.pemberton-ross@unibas.ch (P.J. Pemberton-Ross), mikhail.pachkov@unibas.ch (M. Pachkov), erik.vannimwegen@unibas.ch (E. van Nimwegen).

http://dx.doi.org/10.1016/j.ymeth.2015.06.024
1046-2023/

perturbations would be needed to transdifferentiate cells from a particular cell type to a given desired target type is not understood. How to recognize the breakdown in control of gene expression, which may be associated with particular disease states, is another example of a systems-level question to which we currently have little insight.

To appreciate the magnitude of the challenge we face in answering such questions, it helps to recognize just how fragmentary our knowledge of genome-wide gene regulatory interactions still is in higher eukaryotes. For example, of the roughly 1500 TFs present in mammalian genomes [6], binding specificities are known for less than half, e.g. [7]. The ability of TFs to bind to their cognate sites depends on the local state of the chromatin which can be modified in a large number of ways, i.e. through chemical modification of the histone tails within nucleosomes. These epigenetic marks are both 'read' and 'written' by chromatin modifying enzymes which in turn may be recruited to specific loci by TFs bound to the DNA. This potentially complex feedback between chromatin state and TF binding is currently poorly understood. TFs may interact through direct protein–protein contacts with each other and with a large number of co-factors, and our knowledge of these interactions is very incomplete. Although regulation of transcription initiation is of crucial importance for the control of gene expression, expression is also regulated at the level of transcript processing (splicing, poly-adenylation), mRNA transport, transcript stability, translation initiation and elongation, and protein degradation. Although some aspects of this post-transcriptional regulation have been investigated in some detail, e.g. the role of micro-RNAs in regulating transcript stability and translation, by-and-large our knowledge of this post-transcriptional regulation is extremely limited. In addition, the 'activity' of regulatory factors is not only determined by their mRNA and protein expression level, but also by post-translational modification (e.g. phosphorylation at specific residues), by their localization within the cell, by their interaction with co-factors, and so on. In other words, although our knowledge of the individual players and interactions in gene regulatory networks has been steadily increasing, the things we *do not know* still outnumber the things we know by many-fold. Given this, it is clear that we are still very far removed from being able to meaningfully simulate detailed models of the functioning of gene regulatory networks. There is little point in taking all the information we happen to know, and pouring them into a mathematical model or computational simulation, without realistically dealing with the fact that there is much more we do not know.

1.1. Using gene expression data to infer regulatory networks

Instead of expecting to establish a detailed model of the functioning of the genome-wide gene regulatory network, much research focuses on more modest goals, such as identifying the key regulators operating in a particular model system. Since there are at least hundreds of potential regulators, it is generally unfeasible to experimentally investigate the role of all potential regulators. However, with the advent of high-throughput technologies such as next-generation sequencing, it has become relatively easy to measure gene expression and chromatin state genome-wide. Over the last decade, many researchers have thus turned to such methodologies with the aim of identifying the key regulatory interactions acting within their specific model systems.

From the point of view of computational methods, the question has thus become of how we can most efficiently use high-throughput data, such as genome-wide gene expression data, to learn about the key regulatory interactions acting in a given system. Indeed, a large number of methods for performing inference of regulatory networks from gene expression data has been proposed over the years, ranging from mostly descriptive methods that aim to summarize the structure of these high-dimensional datasets in terms of a relatively small number of statistical features, to highly specific methods that fit the data in terms of concrete models of the genome-wide gene expression dynamics, e.g. using coupled differential equations, Gaussian models, or Bayesian network models, see e.g. [8–10] for reviews.

On one end of this scale, methods that aim to fit the data using specific models of the underlying gene regulatory network generally suffer from the 'curse of dimensionality'. Put simply, because the number of possible regulatory network architectures is huge, the amount of data that would be necessary to reliably infer the regulatory network is many orders of magnitude larger than even the largest high-throughput datasets can provide. To uniquely predict a regulatory network from the data, these methods employ regularization schemes that aim to minimize either the total number of regulatory interactions, their magnitudes, or a combination of both. However, it is unclear to what extent we should expect such 'minimal' networks to match the true underlying biological network. Moreover, in order for the network inference to be computationally feasible, these methods are often forced to treat all genes as equivalent, thereby ignoring all kinds of relevant prior biological information. For example, many of such methods simply investigate the correlation or mutual information between all pairs of genes, and consider possible regulatory interactions between any pair of genes, even though prior biological knowledge indicates that most genes do not act as regulators.

On the other end of the scale, many methods focus simply on reducing the dimensionality of the data by identifying statistical descriptors that capture the main features of the data. These include well-known methods such a principal component analysis (PCA), which finds linear combinations of the variables (e.g. genes and conditions) which carry most of the variance in the data, as well as various clustering methods that divide the genes and or samples into a relatively small number of subsets that show similar expression profiles. Although such methods are very valuable in clarifying the structure of the data, it is generally difficult to relate the structures that they identify to underlying biological mechanisms. For example, when a particular subset of genes is predicted to form a 'co-regulated module', it is generally unclear what follow-up experiments could be done to further characterize or even validate this prediction.

1.2. Motif activity response analysis

In our view, the challenge facing methods for gene regulatory network reconstruction consist in reducing the dimensionality of the problem, so that models can be meaningfully fitted to the data, on the one hand, while at the same time incorporating relevant prior biological information, and formulating the models in terms of concrete biological mechanisms that are amenable to direct experimental follow-up, on the other hand. A few years ago we proposed an approach to regulatory network inference, called motif activity response analysis (MARA), which combines these desirable features [11]. First, recognizing that much of genome-wide mRNA expression levels are controlled by transcriptional and post-transcriptional regulators, MARA models gene expression levels explicitly in terms of the *activities* of TFs and miRNAs. To do this, MARA makes use of the fact that, both for miRNAs and for many TFs, targets genes can be computationally predicted based on DNA and RNA sequence analysis. That is, MARA first computationally predicts, for each of hundreds of TFs and miRNAs, which transcripts are regulated by each of these regulators. MARA then uses a very simple linear model to relate the observed expression levels of all transcripts in terms of the activities of the regulators. In this way, the very high-dimensional gene expression data, i.e. involving expression levels of tens of

thousands of different transcripts, are reduced to a few hundred *motif activities*. Moreover, since MARA also rigorously quantifies the uncertainty in the inferred motif activities, motifs can be sorted by how much of the observed expression data they explain. Dimensionality can then further be reduced by considering only those motifs that significantly affect gene expression levels, which is typically on the order of one or two dozen motifs. In this way the high-dimensional gene expression data is reduced to the activities of a modest number of key regulators. Importantly, however, since these motif activities represent the action of specific TFs and miRNAs, these predictions are directly amenable to experimental follow-up, e.g. through perturbation of the levels of these regulators, through mapping their genome-wide targets using ChIP-seq, and so on.

After the initial presentation of MARA and its application to inferring the core regulatory network of a differentiating human cell line [11], MARA has since been applied to a large number of different mammalian systems [12–27] and extended to model genome-wide chromatin marks in terms of epigenetic motif activities [28]. Remarkably, in all these systems, MARA's predictions of key regulators and their interactions were subsequently confirmed by experimental validation. It may seem surprising that a simple linear model, which ignores much of the known biological complexity, so robustly identifies key regulators across a wide variety of mammalian systems. Indeed, MARA's linear model generally performs very poorly at predicting the expression profiles of individual genes. However, because typical regulators target hundreds of genes, the inferred motif activities are *statistical averages* of the behaviors of all predicted targets and these averages are highly robust. That is, motif activities can be accurately estimated because the complexities of the regulation at each individual target are effectively averaged out. We recently implemented MARA as a completely automated tool, called the Integrated System for Motif Activity Response Analysis (ISMARA) [1], which is available through a web interface at ismara.unibas.ch. Here users can upload gene expression data (micro-array or RNA-seq) and have MARA performed automatically, with the results provided through an interactive graphical web interface.

1.3. Toward a causal dynamics of motif activities

So far, MARA infers motif activities independently for each experimental condition in which gene expression was measured. Here we propose the first steps toward modeling the *dynamics* of motif activities across a *time course* of gene expression measurements. Time-series data provide not only the opportunity to separate out events which occur on different timescales and transient behavior, but also illustrate the sequence of events. This extra information allows us to begin moving from establishing correlations to establishing causality, an important step in reaching the goal of a mechanistic understanding of a process.

Here we present ARMADA (Auto Regressive Motif Activity Dynamics Analysis) which combines MARA with auto-regressive modeling to infer causal interactions between regulators. In particular, we first apply MARA to a time course of gene expression measurements to infer a time course of motif activities. We then use an auto-regressive model that models the motif activities at time t as a function of the motif activities at time $t-1$. To control the complexity of the model we make a number of simplifying assumptions. First, we assume that the function relating the motif activities at time t to those of time $t-1$ is *time invariant*, i.e. the same at each time point. Second, we assume that the *change* in motif activities from time $t-1$ to t is a simple *linear* function of the current motif activities. Note, however, that even if we only consider the dynamics of the top M most significant motifs, there are still M^2 possible motif–motif interaction parameters in our model, and typical gene expression time courses still do not contain enough data to unambiguously determine all parameters. We will thus additionally assume that motif–motif interactions are sparse, favoring models with few connections.

The organization of the paper is as follows. We will first review the methods employed in motif activity response analysis, including the predictions of binding sites for TFs and miRNAs, the assumptions underlying our linear model, and the target predictions. After this we will introduce our ARMADA method, explain how we optimally fit its parameters to the motif activity dynamics, and describe the results its outputs. To illustrate the method we apply it to a time course of gene expression data of human umbilical vein endothelial cells (HUVECs) that are treated with TNFα. As is well-known, such treatment will trigger an inflammatory response in HUVECs and we show how ARMADA infers interactions between key regulators of this response, including several direct interactions that are known in the literature, and several novel predictions. We have implemented ARMADA as part of our ISMARA web server, allowing any user to automatically perform ARMADA on any dataset that was analyzed by ISMARA.

2. Methods

2.1. Motif activity response analysis

MARA first quantifies genome-wide gene expression patterns in terms of the expression levels of *promoters*. In particular, for a given model organism MARA starts from a curated *promoterome*, i.e. a genome-wide collection of promoters. For human and mouse these promoteromes where constructed primarily from deep-sequencing data of transcription start sites (deepCAGE data [29]). As described previously, promoters were defined as sets of neighboring transcription start sites (TSSs) on the genome that, within measurement noise, are co-regulated across a large panel of conditions, and were identified from deepCAGE data using a Bayesian method described previously [30].

2.1.1. Expression data processing

Most algorithms for network inference do not consider the pre-processing of raw input data to be part of the inference problem, and assume processed data to be provided. This can sometimes lead to a somewhat careless attitude toward the details of raw data-processing, e.g. not carefully distinguishing between transcript-level and gene-level expression values, between relative and absolute expression levels, between log-transformed and non log-transformed expression levels, etcetera. However, in our experience the quality of the network inference crucially depends on the care given to the pre-processing of raw data and in ISMARA all pre-processing steps are an integral part of the method. We here briefly summarize ISMARA's pre-processing of gene expression data and refer the reader to [30] for details. ISMARA uses comprehensive transcript collections, such as the Gencode transcript collection [31], to associate a set of transcript isoforms to each promoter. The micro-array or RNA-seq input data is then used to estimate, for each promoter p and each sample s, the relative log-expression level E_{ps}, which is defined as the logarithm of the number of mRNAs in sample s deriving from promoter p per million mRNAs.

To calculate E_{ps} for micro-array data, the data are first corrected for background and non-specific binding and probes that show no statistical evidence of specific expression in all input samples are removed. Instead of relying on annotation from the array manufacturer, ISMARA maps all probe sequences to the transcript set, taking into account that a given probe may map to multiple transcript

isoforms, which may or may not all be associated with the same promoter. The log-expression E_{ps} of a promoter is a weighted average of the log-intensities all probes that map to transcripts that are associated with the promoter.

Similarly, for RNA-seq data all reads are mapped to the transcripts of the transcript set. When a read maps to n different transcripts, each transcript is assigned a weight $1/n$. A transcript's expression level is estimated as the total weight of reads mapped to it divided by the transcript length. The expression level of the promoter is the sum of the expression levels of the transcripts associated to it. To reduce artefactual fluctuations in the estimated expression of low expressed promoters due to the Poisson noise in read counts, a pseudo-count is added to each promoter which correspond to the fifth-percentile of all promoters' expression levels in the sample. Finally, the expression values E_{ps} are obtained by normalizing all expression levels to the sum of expression levels in the sample, multiplying by one million, and log-transforming.

2.1.2. Regulatory site prediction

The second key ingredient of ISMARA consists of computationally predicted regulatory sites for a large collection of mammalian TFs and miRNAs. We first curated a set of 190 regulatory motifs, i.e. position specific weight matrices, that represent the binding specificities of ≈ 350 mammalian TFs using data from motif databases [32,33] together with data from the literature and our own analysis of ChIP-chip and ChIP-seq data. For each promoter we obtained a multiple alignment of a 1 kilobase region centered on its TSS together with orthologous segments from 6 other mammals (the 7 species being human, mouse, rhesus macaque, dog, cow, horse, and opossum). We then used the MotEvo algorithm [34] to predict functional transcription factor binding sites (TFBSs) for all motifs in each alignment. MotEvo is a Bayesian algorithm that predicts TFBSs by combining a physical binding model with phylogenetic information from the multiple alignment to give a higher posterior probability to binding sites which show evidence of having being conserved. The TFBS predictions are finally summarized in a matrix **N**, where N_{pm} is the sum of the posterior probabilities of all predicted TFBSs for motif m in promoter p. Similarly, to incorporate post-transcriptional regulation by miRNAs we use target site predictions from TargetScan using preferential conservation scoring (P_{CT}) [35]. The target 'site count' N_{pm} for a miRNA seed family m targeting promoter p is calculated as the average of the TargetScan scores across all transcripts associated with promoter p. Our collections of regulatory motifs and target site predictions can all be obtained from our SwissRegulon database [36].

2.1.3. Bayesian inference using a linear model

MARA then fits the observed expression data E_{ps} in terms of the computationally predicted site-counts N_{pm} and (unknown) motif activities A_{ms} using a simple linear model

$$E_{ps} = \tilde{c}_s + c_p + \sum_m N_{pm} A_{ms} + \text{noise}, \tag{1}$$

where $\tilde{c}_s$ is a sample-dependent normalization constant which reflects the total expression in the sample, and c_p is the 'basal' or average expression of promoter p.

MARA uses a Bayesian procedure in which a Gaussian prior is assigned to each motif activity $P(A_{ms}) \propto \exp\left[-\lambda A_{ms}^2\right]$ and the likelihood model $P(E|A)$ assumes that the noise term, the difference between the predicted and observed expression levels resulting from measurement error, biological fluctuation and model error, are Gaussian distributed with unknown variance σ^2, i.e.

$$P(E|A) \propto \exp\left[-\sum_{p,s} \frac{\left(E_{ps} - c_p - \tilde{c}_s - \sum_m N_{pm} A_{ms}\right)^2}{2\sigma^2}\right]. \tag{2}$$

The parameter λ of the Gaussian prior is fitted through a cross-validation procedure, whereby λ is first fit on a randomly-selected set comprising 80% of promoters and then used to predict expression values for all promoters. The average squared-deviation of these predicted values from the observed expression levels is then minimized, using the remaining 20% of all promoters as a test set.

The model (2) can be easily solved and the posterior distribution $P(A|E)$ is given by a multi-variate Gaussian whose maximum and covariance matrix can be expressed in terms of the singular value decomposition of the site-count matrix N. Since the entire posterior distribution $P(A|E)$ can thus be explicitly calculated, we can calculate any particular statistics of interest. Full details of the form and calculation of the solution can be found in Eqs. (4)–(10) of the supplementary material of [30].

2.1.4. Interpretation of the linear model

Whereas in many network inference algorithms the expression levels of regulators are used to directly estimate their 'activity' in each of the samples, MARA does not use the observed mRNA levels of the regulators. Instead, the motif activities A_{ms} are inferred from the observed expression of the predicted *targets* of the motif m and since the precise biological interpretation of these motif activities may be unclear, we here provide a more detailed interpretation of the linear model (2) and its associated motif activities.

Since the E_{ps} correspond to *log* expression values, the linear model calculate the absolute expression $X_{ps} = e^{E_{ps}}$ as

$$X_{ps} = a_s b_p \prod_m e^{N_{pm} A_{ms}}, \tag{3}$$

with the constants $a_s = e^{\tilde{c}_s}$ and $b_s = e^{c_p}$. If we furthermore define $\lambda_{ms} = e^{A_{ms}}$ we have

$$X_{ps} = a_s b_p \prod_m (\lambda_{ms})^{N_{pm}}. \tag{4}$$

Thus, the model effectively assumes that, in sample s, each occurrence of a binding site for motif m multiplies the transcript level of promoter p by a factor λ_{ms}. Whenever $A_{ms} > 0$ the multiplicative factor $\lambda_{ms} > 1$, leading to an increased expression, whereas when $A_{ms} < 0$ the occurrence of a binding site for motif m leads to a repression of promoter p's transcript level. These multiplicative factors can be interpreted as follows. Imagine that the TFs binding to motifs m function as activators. Whenever a site for motif m is added to promoter p, this site will be bound a certain fraction of the time f_{ms} which will generally be a function of the nuclear concentration of TFs that can bind to sites of motif m in sample s. The model now assumes that, whenever this site is bound, this increases the general affinity or binding energy of the RNA polymerase to the promoter, leading to an increase in the transcription rate by a factor c_{ms} compared to the situation when the promoter is not bound. This factor may depend on the condition s because it may depend on other variables, such as the presence of co-factors, which may vary across conditions. Thus, adding the binding site for motif m to promoter p increases the transcription rate of promoter p by a factor $(1 + c_{ms} f_{ms}) = \lambda_{ms}$. Note that for TFs acting as repressors we would have $c_{ms} < 1$, i.e. the binding of the site would lower the transcription rate. In summary, the model assumes that each binding site in promoter m will be bound some condition-dependent fraction f_{ms} of the time, and that whenever a site is bound, this leads the transcription rate to be altered by a factor c_{ms}. The main simplifying assumption that the model makes is that the factor $\lambda_{ms} = (1 + c_{ms} f_{ms})$ is the *same* at each promoter, and independent of the other sites occurring

at the promoter. Although we do not expect this assumption to hold in general, it does not seem unreasonable to expect that it may hold approximately for a substantial fraction of promoters that are targeted by a given motif.

In conclusion, the exponent of the motif activity $e^{A_{ms}}$ can be interpreted as the fold-change in the expression level of a promoter that is predicted to be observed in sample s if a binding site for motif m were to be added to the promoter. Since MARA also predicts the locations of the binding sites determining the site-count N_{pm} for each motif and promoter, these predictions are also directly amenable to experimental verification, i.e. using experiments in which promoter sites are mutated.

2.2. ARMADA: an autoregressive model of motif activity dynamics

Here we propose to go beyond the inference of independent motif activities A_{ms} for each sample s, and present a model for inferring causal interactions between motif activities from time course gene expression measurements. We assume that, at least formally, the motif activity profile $\vec{A}(t)$ across the time course obeys a differential equation of the form

$$\frac{d\vec{A}(t)}{dt} = F(\vec{A}, t), \tag{5}$$

where the function F is a function of all current motif activities and possible other 'outside' influences, and which may also be time-dependent. In this general form the problem is hugely underdetermined and to make progress we make a number of assumptions. First of all, we will assume that the function F does not explicitly depend on time, i.e. $F(\vec{A}, t) = F(\vec{A})$. This means, in particular, that any outside influences are assumed approximately constant over the length of the time course.

Second, since ISMARA infers motif activity *changes*, the average activity of each motif across the time course is zero, i.e. $\sum_t A_{mt} = 0$, for all m. In other words, the activity A_{mt} at time point t denotes the *deviation* at time point t of motif m from its average activity over the time course. We will make the assumption that these deviations are small enough such that the function $F(\vec{A})$ can be approximated to first order in these 'deviations', i.e. we write

$$\frac{d\vec{A}(t)}{dt} = \vec{F}(0) + \vec{A}(t) \cdot W. \tag{6}$$

In this equation, the components W_{mn} of matrix W, denote the strength of the 'causal' regulatory interaction from motif m to motif n. The vector $\vec{F}(0)$ can be thought of as a constant vector of outside influence that consistently drives motif activities (either upward or downward). For notational simplicity, we will write

$$F_m(0) = A_0 W_{0m}, \tag{7}$$

where A_0 is the (constant) activity of an outside influence and W_{0m} is the interaction of the outside influence with motif m, which may be positive or negative.

The final assumption that we will make is that the time points are spaced densely enough such that, over each time interval, the solution of the differential equation can be approximated assuming the right-hand side constant. We then have, for each time point t and motif m:

$$A_{t,m} = A_{t-1,m} + \delta_{t-1} \sum_n A_{t-1,n} W_{nm}, \tag{8}$$

where δ_{t-1} is the time interval from time point $t-1$ to time point t. The form of the model (8) is known in the literature as an auto-regressive model of order 1 (AR(1) model). Autoregressive models have been one of the most important tools for modeling and forecasting of multidimensional time series data, with established application in fields ranging from finance to ecology. There is a large corpus of knowledge regarding various extensions and postprocessing of the basic autoregressive model which have been well summarized in [37,38], see also [39,40].

The basic premise of an order-1 autoregressive model (AR(1) model) is that the state of the system at time t, i.e. in our case an M-dimensional motif activity vector $\vec{A}_t$, with M the number of motifs, can be modeled as a weighted sum of the system state at the previous time point, $t-1$, up to some additive Gaussian noise:

$$A_{t,m} = A_{t-1,m} + \sum_n A_{t-1,n} \delta_{t-1} W_{nm} + e_{t,m} \quad \forall t \in 1 \ldots T \tag{9}$$

where e is the zero-mean Gaussian noise process with $\langle e_{t,m} \rangle = 0$ and $\langle e_{t,m} e_{t,n} \rangle = \Lambda^{-1}_{mn}$ is the covariance in the noise. That is, the matrix Λ is the inverse of the covariance matrix of the noise.

Expressed in this form, the problem can be thought of as a simultaneous multiple linear regression: let us construct a matrix of "predictors", X, by stacking the row vectors of our state vector from $t = 1 \ldots T-1$:

$$X = [A_{1,m}, A_{2,m}, \ldots, A_{T-1,m}]. \tag{10}$$

Furthermore, let's define the discrete motif activity derivatives

$$a_{t,m} = \frac{A_{t,m} - A_{t-1,m}}{\delta_{t-1}}, \tag{11}$$

and a matrix of "responses", Y, by stacking the row vectors of these discrete time derivatives from $t = 2 \ldots T$:

$$Y = [a_{2,m}, a_{3,m}, \ldots, a_{T,m}] \tag{12}$$

We can now rewrite our $T-1$ simultaneous copies of Eq. (9) in the simpler form:

$$Y = X \cdot W + E \tag{13}$$

where $E = e_{t,m}$ is a $(T-1) \times M$ noise matrix.

Under this Gaussian noise model and the autoregressive model for time evolution, we find that the probability of obtaining the time course of activities A given a particular W and Λ is:

$$p(A|W,\Lambda) = (2\pi)^{\frac{-M(T-1)}{2}} (\det \Lambda)^{(T-1)/2} \times \exp\left\{ -\frac{1}{2} Tr\left[\Lambda \cdot (Y - X \cdot W)^T \cdot (Y - X \cdot W) \right] \right\} \tag{14}$$

Since the log-likelihood (14) is a quadratic function of the components W_{mn}, the maximum likelihood (ML) solutions W_{ML} and Λ_{ML} can be easily obtained and are given by

$$W_{ML} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y \tag{15}$$

and

$$\Lambda_{ML} = \frac{1}{N(N-1)} (Y - X \cdot W_{ML})^T \cdot (Y - W_{ML} \cdot X) \tag{16}$$

However, this simple ML solution is unsatisfactory for two reasons. First of all, since there are M^2 possible interactions W_{mn} and only $M \cdot T$ independent activities $A_{t,m}$, over-fitting is a concern unless $T \gg M$. To control over-fitting, we thus want to introduce a prior $P(W)$ on the interactions and restrict the number of motifs M that are used in the fitting. Second, in the above equations we have assumed that the motif activities $A_{t,m}$ are known with infinite precision, whereas in reality the $A_{t,m}$ are themselves estimated by MARA from gene expression measurements.

Instead of inferring the interactions W directly from motif activities, we ultimately are inferring them from a time course of expression measurement E and use the motif activities A as an intermediate in this inference. Thus, formally we want to calculate the likelihood in terms of the expression data E only, i.e. we want

to calculate $P(E|W,\Lambda)$. This likelihood can obtained by marginalizing over the motif activities

$$P(E|W,\Lambda) = \int P(E|A)P(A|W,\Lambda)dA, \tag{17}$$

where the integral is over the activities $A_{t,m}$ for all motifs m and time points t. In the model used by MARA, the likelihood of motif activity vector $\vec{A}_t$ at time point t, given a gene expression state E_t, is given by a multi-variate Gaussian [1]

$$P(E_t|\vec{A}_t) = P(E_t|\vec{A}_t^*)\exp\left[-\frac{1}{2}(\vec{A}_t - \vec{A}_t^*)\cdot\Delta_t^{-1}\cdot(\vec{A}_t - \vec{A}_t^*)\right], \tag{18}$$

where $\vec{A}_t^*$ is the maximum likelihood value of the motif activities, and Δ_t is the covariance matrix of motif activities at time t. Consequently, the probability $P(E|A)$ is a multi-variate Gaussian in terms of all motif activities $A_{t,m}$, so that in principle the integral in (17) can be performed analytically. Unfortunately, the resulting expression is a complex function of powers W^t of the interaction matrix W, which cannot be easily optimized.

Thus, instead of performing the integral (17) analytically, we approximate this integral by sampling a large number, e.g. $R = 100$, of 'replicate' motif activity time courses $\vec{A}_t$ from the distribution (18), and let ARMADA infer the interactions W from the entire set of R replicate time courses. That is, we write

$$\begin{aligned} P(E|W,\Lambda) &= \int dA P(E|A)P(A|W,\Lambda) \approx \frac{1}{R}\sum_{r=1}^{R} P(A^r|W,\Lambda) \\ &\approx \exp\left[\frac{1}{R}\sum_r \log(P(A^r|W,\Lambda))\right], \end{aligned} \tag{19}$$

where the replicate motif activity time courses A^r are sampled from the distribution (18) and in the last approximation we have replaced the mean with the geometric mean. It can be shown that, as long as all replicates have similar likelihood near the global maximum likelihood value of W, this leads to the same maximal likelihood value of W and also to a similar curvature of the likelihood function near its maximum. This approximation allows us to simply pass the R replicate time courses to ARMADA and perform the ML inference as defined above, taking care to scale all log-likelihoods by a factor $1/R$ in the end.

To control over-fitting, we assume that the interaction weights W_{mn} are drawn from a Gaussian distribution with zero-mean and precision α:

$$p(W|\alpha) = \left(\frac{\alpha}{\sqrt{2\pi}}\right)^{M^2}\exp\left\{-\frac{\alpha}{2}\sum_{m,n}W_{mn}^2\right\}, \tag{20}$$

where the parameter α controls how strongly interactions are penalized. Our tests on a number of datasets indicate that values in the range $\alpha \in [1,5]$ work well in practice to minimize the deviation between the observed and predicted activities; for the results reported here we used $\alpha = 2$. Apart from allowing the user to set the value of α, we have also implemented a cross-validation scheme in which one transition $t \to t+1$ is left out, and α is set to maximize the prediction on this transition. Finally, although not used in the results presented here, we have also implemented an *informative prior* where for each edge (m,n) there is a separate parameter $\alpha_{m,n}$ which can be specified by the user. The values of $\alpha_{m,n}$ can, for example, be set either depending on the occurrence of predicted TFBSs for motif m in promoters of TFs associated with motif n, i.e. a high value of $\alpha_{m,n}$ whenever there is no known site for motif m, and a lower value when there are such predicted TFBSs. Alternatively, ISMARA predicts regulatory interactions from each motif m to the promoters of TFs associated with motif n, and the chi-squared scores of these predicted regulatory interactions can also be used to set the $\alpha_{m,n}$.

Another parameter controlled by the user is the activity value A_0 of the external source. Since the prior (20) determines the a priori likely values of W, the size of A_0 relative to the typical size $A_{t,m}$ for the motifs in the model, determines how likely it is a priori to couple to an outside source rather than to another motif. In ARMADA A_0 is set to match the highest activity $A_{t,m}$ observed across all motifs and time points. For the precision matrix we use a simple uninformative prior

$$p(\Lambda) = (\det\Lambda)^{\frac{M+1}{2}} \tag{21}$$

over a range in which the likelihood is non-zero. In this case, we find

$$W_{ML} = (\alpha I + X^T\cdot X)^{-1}\cdot X^T\cdot Y, \tag{22}$$

and Λ_{ML} is still given by (16).

Finally, as already noted above, as the number of parameters in the model grows quadratically with the number of motifs M, it becomes hard to control over-fitting as the number of motifs becomes large, and we would thus like to restrict ARMADA to use only those motifs that are likely key regulators in the process under study. Indeed, MARA calculates for each motif m an overall significance z_m. In particular, for each time point t, the z-statistic $z_{mt} = A_{t,m}^*/\sqrt{\Delta_{mm}}$ calculates how many standard-deviations the motif is away from zero, i.e. no activity, at time point t, and the overall z-score z_m is defined as

$$z_m = \sqrt{\frac{1}{T}\sum_t z_{mt}^2}. \tag{23}$$

The higher z_m, the more important the motif is in explaining the observed gene expression variation across the time course. To reduce dimensionality, ARMADA uses only the M motifs that have a *z-score* over a cut-off, which can be set by the user but is typically of order $z_m \geqslant 2.0$.

2.3. Algorithm output

2.3.1. Evidence score for interactions between regulators

The primary output from the algorithm is a set of Gaussian posterior distributions for each element $W_{m,n}$ of the matrix of interactions W. Each of these posterior distributions is characterized by an estimated interaction term $\hat{W}_{m,n}$ and its variance $C_{m,n}$. To allow a more simple and concrete interpretation, this distribution is used to assign a *link-confidence z-score* to each interaction defined as $Z_{m,n} = \hat{W}_{m,n}/\sqrt{C_{m,n}}$, which quantifies how much evidence exists that the coupling $W_{m,n}$ is non-zero, i. e. the evidence that one regulator (be it a TF or a miRNA) influences another. ARMADA provides a list of all predicted interaction strengths $W_{m,n}$, sorted by their significance $Z_{m,n}$.

2.3.2. Graphical representation of the regulatory network

Those significant couplings with a high link-confidence z-score $Z_{m,n}$ can be represented as a network, i.e. a graph, where nodes representing motifs are joined by a directed edge if the absolute value of the link-confidence z-score is above a certain threshold. ARMADA automatically generates such graphs for a desired threshold Z_c. In addition, the shape of the arrow and the color of the edge can be used to indicate whether the interaction was activating $\hat{W}_{m,n} > 0$ or repressing $\hat{W}_{m,n} < 0$. See Fig. 2 below for an example of a network inferred by ARMADA.

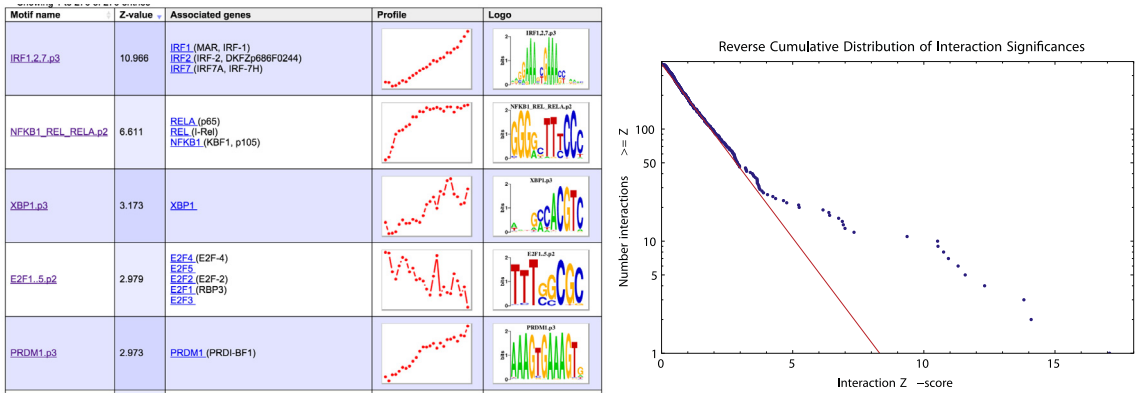

Fig. 1. *Left panel*: the top 5 regulatory motifs, as identified by ISMARA, for the time course of HUVEC cells treated with TNF. The table shows, for each motif, the motif name, its z-score z_m, the associated TFs, a thumbnail of its motif activity profile, and its sequence logo. *Right panel*: reverse cumulative distribution of Z scores $Z_{m,n}$, as calculated by ARMADA, of all possible interactions among the 19 regulatory motifs with $z_m > 2$. The vertical axis is shown on a log scale and the red line shows an exponential fit to the initial part of the distribution.

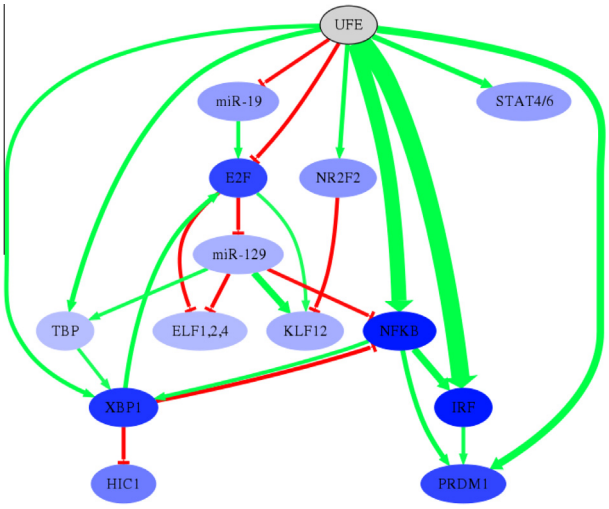

Fig. 2. The core regulatory network of the gene expression dynamics observed in HUVEC cells upon treatment with TNFα as predicted by application of ARMADA. Each node corresponds to a regulatory motif and each edge to a predicted causal regulatory interaction. The intensity of the node's color corresponds to the overall significance z_m of the corresponding motif, and the thickness of the edges corresponds to the significance $Z_{m,n}$ of the regulatory interaction. Activating interactions are shown in green and repressing interactions are shown in red. The UFE (unknown functional element) node corresponds to an unknown outside influence that drives motif activities consistently up or consistently down across the time course.

It is important to distinguish the regulatory network predicted by ARMADA from the regulatory network that ISMARA provides in its results. In ISMARA, the motif m is predicted to target a gene g when the promoter of g has one or more binding sites for motif m, and the motif activity of m contributes significantly to explaining the mRNA expression profile of g. ISMARA then identifies, among the target genes of each motif m, those genes that are themselves encoding TFs corresponding to motifs m' within ISMARA's collection of motifs. In this way ISMARA also provides a regulatory network with edges from motifs m to m', where the edge $m \to m'$ indicates that the activity of motif m contributes to explaining the mRNA level of a TF associated with motif m'. In contrast, in ARMADA's network an edge from motif m to m' indicates that the motif activity of m at time t, contributes to explaining the motif activity of m' at time point $t+1$.

2.3.3. Ability to recapture ISMARA activity dynamics

To further assess the performance of the model, ARMADA first compares the observed discrete-time derivatives $a_{t,m}$ with those predicted by the model, i.e.

$$a^{\text{theo}}_{t,m} = \sum_n \hat{W}_{m,n} A^*_{t-1,n}. \tag{24}$$

ARMADA provides, for each motif m, a scatter plot of the observations $(a^{\text{theo}}_{t,m}, a_{t,m})$ (examples are shown in left panels of Fig. 3 below) and calculates its Pearson correlation coefficient. In the results, ARMADA provides these correlation coefficients and scatter plots for each motif.

Secondly, using the predicted motif activity changes $a^{\text{theo}}_{t,m}$, ARMADA constructs a 'one step' predicted motif activity profile by estimating

$$A^{\text{os}}_{t,m} = a^{\text{theo}}_{t,m}\delta_{t-1} + A^*_{t-1,m}. \tag{25}$$

As explained above, to estimate the interactions W, ARMADA uses R replicate motif activity time courses $A_{t,m}$ and the ARMADA output reports both the mean and standard-deviation of the one-step-forward projected time course $A^{\text{os}}_{t,m}$. As a heuristic of ARMADA's performance, plots are provided that show this predicted motif activity time course side-by-side with the observed motif activities (middle panels in Fig. 3 below).

2.3.4. Long-term behavior of the predicted dynamics

ARMADA also calculates a 'forward projected' motif activity time course $A^{\text{fp}}_{t,m}$ by starting from the motif activities at the initial time point and iterating the matrix $\hat{W}$, i.e.

$$A^{\text{fp}}_{t,m} = A^{\text{fp}}_{t-1,m} + \delta_{t-1}\sum_n A^{\text{fp}}_{t-1,n}\hat{W}_{n,m}. \tag{26}$$

The ARMADA output also reports both the mean and standard-deviation of this long-term forward-projected time course $A^{\text{fp}}_{t,m}$, again side-by-side with the observed motif activity dynamics $A^*_{t,m}$ (right panels of Fig. 3 below). Regarding the long-term forward-projected time course, it should be noted that a poor correspondence between the predicted and observed motif activities using this method does not necessarily mean that the predicted interactions are unreliable. This method of propagating forward in time from an initial observation is of course very sensitive to the initial activities $A_{1,m}$ used, and fluctuations in $A_{1,m}$ may be amplified as t increases, so that $A^{\text{fp}}_{t,m}$ may not be close to $A^*_{t,m}$ for large t.

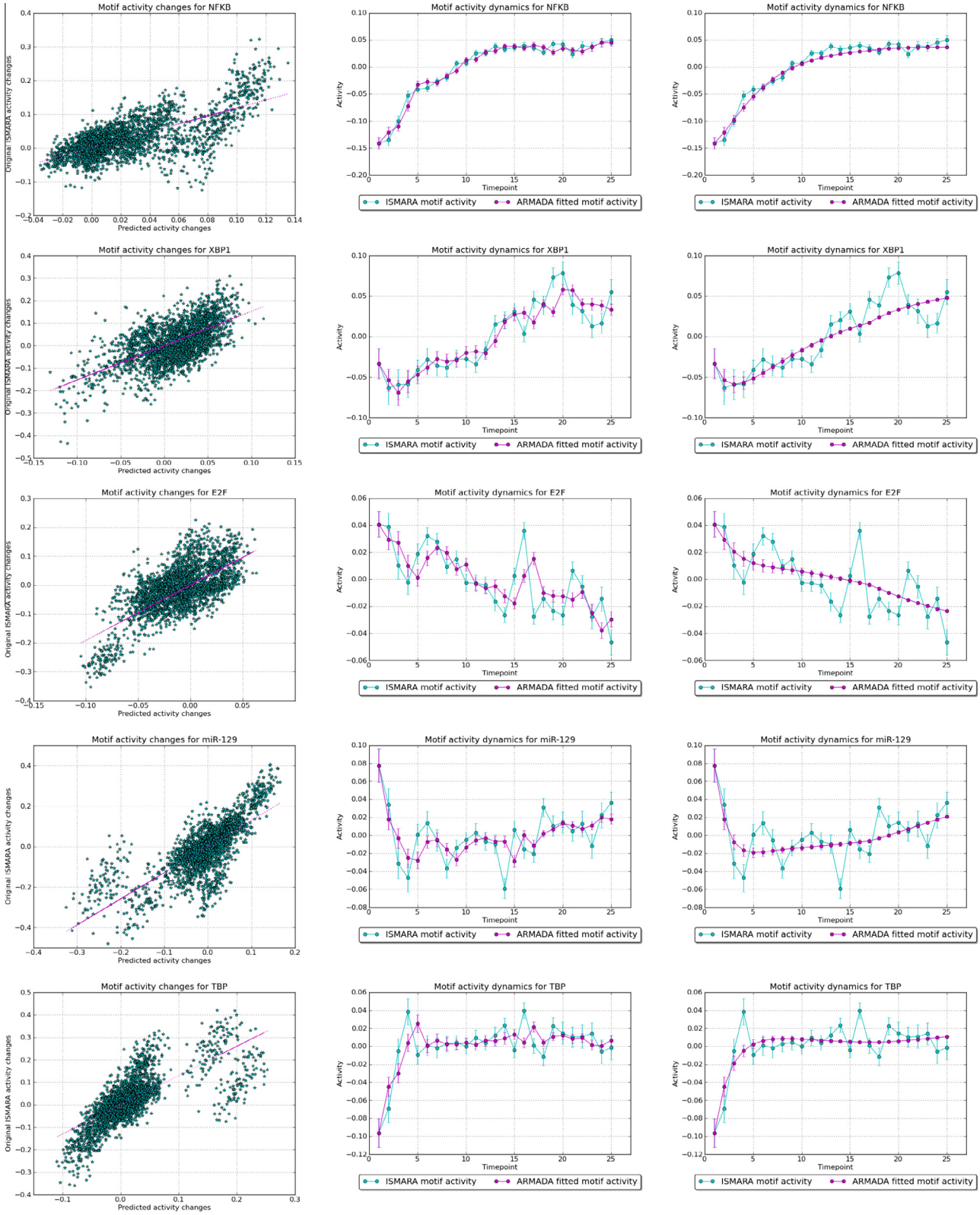

Fig. 3. Motif activities for some of the top regulatory factors, as inferred by ISMARA and modeled by ARMADA. The rows correspond, from top to bottom, to the motifs NFκB, XBP1, E2F, miR-129, and TBP. The left panels show a scatter plot of the ARMADA predicted motif activity changes at consecutive time points (horizontal axis) against the observed motif activity changes (vertical axis). The middle panels show the ARMADA one step predicted activity profiles and the right panels the ARMADA forward projected activity profiles (red curves). The observed motif activity profiles are shown as the blue curves. The error bars correspond to standard deviations across 100 samples of the posterior probability of motif activities, as inferred by ISMARA.

However, it may still be that the $W_{m,n}$ provides a good time-domain update between most of the $a_{t,m}$ for $t > 1$.

3. Results

To illustrate ARMADA's performance we applied it to a time series of gene expression measurements of human umbilical vein endothelial cells (HUVECs) that were treated with tumor necrosis factor (TNF, also known as TNF-alpha). We previously analyzed this dataset with ISMARA [1]. Messenger RNA expression was measured every 15 min for the first 4 h after treatment, and every 30 min for the next 4 h after that [41]. As is well-known, TNF treatment causes an inflammatory response in HUVEC cells, and since innate immunity and inflammation have been the subject of intense study, much is known about the regulation of this process, allowing us to compare ARMADA's predictions with knowledge from the literature.

As discussed previously [1], ISMARA successfully identified the known key regulators of this inflammatory response and Fig. 1 shows the top 5 motifs identified by ISMARA, together with their significance z_m, a thumbnail of their motif activity profile across the time course, and their sequence logo (left panel). The complete ISMARA results on this dataset are available at ismara.unibas.ch/supp/dataset3/ismara_report. In total ISMARA identified 19 motifs with $z_m > 2$ and we selected these motifs for further analysis with ARMADA. The full results of the ARMADA analysis on this dataset are available at ismara.unibas.ch/supp/dataset3/ismara_report/armada. Since ARMADA allows each motif to be regulated by an unknown functional element (UFE), corresponding to A_0 in Eq. (7), which provides a constant driving force across the time course, there are $20 * 19 = 380$ possible regulatory interactions that ARMADA considers for this dataset.

The right panel of Fig. 1 shows the reverse cumulative distribution of interaction z-scores $Z_{m,n}$ that ARMADA calculated for all 380 possible regulatory interactions. As shown by the fitted red line, at low significances the z-scores are roughly exponentially distributed, while for z-scores above approximately 3 the curve breaks away from this exponential and shows an enrichment of interactions with much higher z-scores. Fig. 2 shows the core regulatory network for the HUVEC time course, as inferred by ARMADA, retaining all regulatory links with $Z_{m,n} \geqslant 3$. In Fig. 3 below, the motif activities as inferred by ISMARA, and as modeled by ARMADA, are shown for some of the most important motifs.

Before discussing the biological significance of the inferred network, we first mention a technical point. We have noticed that ARMADA consistently infers negative regulatory interactions of motifs on themselves. This is likely due to the fact that ISMARA's motif activities are already normalized to the average activity across the time course, such that an activity $A_{t,m} = 0$ corresponds not to zero activity of the motif, but rather an activity equal to its time average. Positive motif activities are thus upward deviations from this average, and negative motif activities downward variations. The negative self-interactions likely reflect the fact that, due to this normalization, deviations generally tend to shrink as time increases, and they generally ensure that the motif activities stay bounded with time. For visual clarity we omitted these self-interactions from the network picture in Fig. 2.

The predicted core network has a hierarchical structure, so the nodes in Fig. 2 are placed such that the causal flow roughly runs from top to bottom in the figure. We see that, at the top of the figure is the UFE, which represents an outside influence that drives motif activities either consistently upward or downward across the time course. This UFE likely reflects the direct effects of TNF and the signaling pathways that it activates. The most significantly activated regulatory motifs are, in order of significance, NFκB, IRF, STAT4/6, PRDM1, NR2F2, TBP, and XBP1, whereas miR-19 and E2F are downregulated by the UFE. We will first discuss the activated motifs.

By far the most significant and most strongly regulated motifs are NFκB (bound by the complex NFKB1/REL/RELA) and the IRF motif which is bound by multiple interferon response regulators. Indeed it is well-known that these regulators are key drivers of the inflammatory response [42]. In addition, ARMADA predicts that NFκB directly activates IRF and this interaction is confirmed by the literature [43]. Interestingly, ARMADA also predicts that IRF and NFκB both activate the downstream regulator PRDM1 (also known as BLIMP-1), which is an important developmental regulator in the B-cell and T-cell lineages. A literature search reveals that, indeed, there is evidence from other systems that PRDM1 is targeted by NFκB as well as several IRF factors [44,45]. That there is a general cross-talk between IRF and PRDM1 is further supported by the fact that these TFs can compete for binding to target sites on the DNA at a subset of their target genes [46].

Another upregulated motif is the motif bound by the STAT2, STAT4, and STAT6 TFs. From the mRNA expression data it is clear that, in this system, it is mostly STAT4 and STAT6 that are upregulated and these factors are indeed known to play a crucial role in the inflammatory response and the associated cytokine signaling [47].

In contrast to the previous regulators, relatively little is known about the role of NR2F2 and TATA-binding protein (TBP) in the inflammatory response. NR2F2 binds to the regulatory motif that is also bound by NR2F1 and HNF4A, but the mRNA expression in this system makes clear that NR2F2 is the crucial factor here and that it likely acts as a repressor (see ismara.unibas.ch/supp/dataset3/ismara_report/pages/HNF4A_NR2F1,2.p2). NR2F2 (also known as COUP-TFII) is known to play a crucial role in lymph vessel differentiation [48] and has recently been reported to, in a different context, regulate genes involved in inflammation [49]. These observations together are consistent with ARMADA's prediction that NR2F2 plays an important role in the inflammatory response in HUVEC cells. ARMADA also predicts that NR2F2 directly regulates KLF12, which is a novel prediction as far as we are aware.

The role of TATA-binding protein in the inflammatory response in HUVEC cells is currently unclear. Although its activity is quickly upregulated in the first hour, the mRNA levels of TBP appear to be decreasing. This would imply TBP acts as a repressor in this system, which is at odds with current knowledge about the role of TBP. It is conceivable that, rather than TBP itself, the activity of the TBP regulatory motif is mediated by cofactors or, more generally, by the chromatin state of the promoters that contain TATA boxes, which are known to represent a special class of mammalian promoters. Indeed, it was recently found that Immediate/Early genes in the inflammatory response are characterized by having TATA-boxes [50]. Our literature search suggests that ARMADA's prediction that TBP upregulates XBP1 is also novel.

XBP1 is the final motif predicted to be externally upregulated and it is indeed known to be a key regulator in the inflammatory response. XBP1 is the main regulator of ER stress and the unfolded protein response (UPR) [51] and UPR is known to be a general characteristic of the inflammatory response resulting from TNF activation [52,53]. XBP1 activity thus likely reflects the general activation of the UPR.

An interesting prediction made by ARMADA is that there is direct cross-talk between NFκB and XBP1, with the most significant prediction being that XBP1 downregulates NFκB. Notably, XBP1 is upregulated rather late in the time course, and its upregulation coincides with the leveling off of NFκB motif activity (see Fig. 3). Indeed, there is recent evidence in the literature for direct cross-talk between NFκB and XBP1 in breast cancer [54]. Moreover, several studies have shown that the UPR can attenuate

the inflammatory response mediated by NFκB [55–57], supporting the feedback of XPB1 on NFκB.

This brings us to the part of the network that is *negatively* regulated by external influences, starting from the external negative regulation of miR-19 and E2F, and including the downstream effects on miR-129, ELF1, and KLF12. Of these motifs, E2F is predicted to be by far the most significant. E2F is a complex regulatory motif that is bound by the family of E2F TFs, which includes both activators and suppressors that are involved in regulating various checkpoints in the cell cycle. In previous analysis with ISMARA we have found that E2F activity tends to reflect the proliferate state of the cells, with increased E2F activity reflecting increased proliferation [1]. Since E2F is down-regulated across the time course (Fig. 3), this may thus reflect progressive cell cycle arrest.

However, of all E2F family members, it is the expression of E2F8 that most closely resembles the E2F motif activity, suggesting that E2F motif activity may reflect the activity of E2F8, which is known to respond to DNA damage [58]. Moreover, TNF is known to induce DNA damage during an inflammatory response [59]. Together these observations suggest that at least part of the E2F motif activity may reflect a response to DNA damage, and indeed among the predicted targets of E2F in ISMARA genes involved in DNA repair are over-represented (ismara.unibas.ch/supp/dataset3/ismara_report/pages/E2F1..5.p2).

According to ARMADA's predictions, E2F activity is involved in cross-talk with XBP1 and two miRNAs in the 'repressive' branch of the regulatory network. The most upstream of these miRNAs is miR-19 whose activity is down-regulated across the time course meaning that the *targets* of miR-19 are progressively down-regulated (Fig. 3). This implies that miR-19 is itself upregulated over the time course. Indeed, miR-19 is upregulated in inflammatory response and it has been found that miR-19 expression enhances the inflammatory response by down-regulating a number of repressors of NFκB [60]. Although it is not known that miR-19 directly targets E2F factors, miR-19 is part of the miR-17-92 cluster which is known to be involved in a regulatory feedback loop with E2F factors and proliferation [61,62].

The other miRNA predicted to interact with E2F is miR-129. miR-129's activity drops quickly in the first 45 min of the time course, and subsequently slowly recovers. This again implies that miR-129 is sharply upregulated in the initial phase. Indeed, miR-129 has been previously reported to be upregulated during an innate immune response [63]. ARMADA predicts E2F to regulate miR-129 activity and, at a lower z-score of about 2, ARMADA also predicts miR-129 to negatively effect E2F activity. That miR-129 negatively affects proliferation and directly target E2F TFs is confirmed in the literature [64,65]. ARMADA additionally predicts that miR-129 targets ELF1,2,4 and KLF12. Indeed, both ELF1 and KLF12 are predicted targets of miR-129 [35]. It may seem counter-intuitive that miR-129 is predicted to activate KLF12 since micro-RNAs generally act as repressors. However, the prediction is that miR-129 positively affects the activity of KLF12. Since KLF12 is known to be a repressor, this implies that miR-129 activity negatively impacts on the expression of KLF12.

In summary, ARMADA's core regulatory network recapitulates much that is known about the inflammatory response in endothelial cells, and a remarkable number of predicted links are supported by literature evidence. In addition, ARMADA predicts a substantial number of novel regulatory interactions that would be of interest for targeted experimental follow-up.

A further heuristic evaluation of ARMADA's performance can be obtained by comparing the observed motif activity dynamics, i.e. the motif activities inferred by ISMARA, to those predicted by ARMADA's dynamical model. The left panels of Fig. 3 show scatter plots of the predicted motif activity changes from one time step to the next against the observed motif activity changes for a few selected motifs. We see that the predicted motif activity changes generally correlate well with the observed changes. Consequently, the one-step projected motif activity profiles tend to closely follow the observed motif activity profiles (middle panels of Fig. 3).

Strikingly, the forward projected motif activity profiles, i.e. the motif activity dynamics obtained by solving ARMADA's inferred model forward in time, using only the initial motif activities at the first time point as a starting condition, show a remarkable match to the observed activity profiles. This close match not only applies to motifs showing a simple up- or down-regulation, but also to motifs showing more complex time dynamics. The fact that ARMADA is able to recreate these complex time dynamics using only a limited number of interactions further suggests the validity of the inferred interactions. ARMADA's full predictions, including networks at different cut-offs on the z-score, as well as all predicted motif activity profiles for all motifs, can be viewed at ismara.unibas.ch/supp/dataset3/ismara_report/armada.

Finally, to make ARMADA easily available to any researcher in possession of time course gene expression data, we have implemented ARMADA as an extension of our ISMARA webserver. As we have described previously [1], users can perform automated ISMARA analysis of gene expression data by simply uploading raw gene expression data to our webserver at ismara.unibas.ch. Once the ISMARA analysis has been performed, the results page includes a button for automatically performing ARMADA analysis on the results. The only thing the user has to specify is the order of the time points, and the assignment of replicate numbers when there are replicate time courses. All results are displayed within our graphical web interface and including networks and motif activity plots exactly as those presented in the figures in this paper.

4. Discussion

Analysis of genome-wide gene expression data is one of the most promising approaches for inferring gene regulatory networks and we have here discussed how motif activity response analysis (MARA), by leveraging TF binding site predictions to model gene expression in terms of regulatory motif activities, can simultaneously reduce the dimensionality of the inference problem, while retaining the ability to give specific mechanistic interpretations that can be directly experimentally validated. A substantial number of recent applications of MARA confirm that it successfully identifies key regulators and their regulatory roles across a wide range of mammalian systems. In this paper we have further extended this methodology by introducing ARMADA, which infers causal regulatory interactions *between* the regulators from time course measurements.

Of course, both MARA and ARMADA make a large number of hugely simplifying assumptions about the genome-wide gene regulatory networks, and ignore many factors that are known to be crucial to a full understanding of gene expression regulation, such as regulation of protein stability, post-translational modifications, protein localization within the cell, the local chromatin state of the DNA and DNA accessibility, interaction of regulators with co-factors, and so on. In fact, it is clear that gene regulatory networks are extremely complex and, in all likelihood, we are currently aware of only a tiny sliver of the genome-wide molecular interactions that impact on gene regulation. As such, we do not pretend that either MARA or ARMADA are able to explain even a moderate fraction of the expression changes which occur, even if perfectly noiseless data at high sampling rates were available. The main aims of MARA and ARMADA are to provide experimental researchers with ways to generate plausible hypotheses that can

be investigated by targeted experimental work. Recent applications of MARA strongly suggest that it successfully identifies key regulators *ab initio* and the results presented here suggest that ARMADA is a valuable extension that makes plausible hypotheses about the ways in which regulators causally affect each other's activities.

We note that, as ARMADA provides an explicit formula for predicting motif activities forward in time, it is straight forward to make *in silico* predictions about the effects of perturbing the expression of a particular regulator. For example, ARMADA could easily predict the expected effects of the knock-out of a particular TF or the interruption of one TF–TF interaction. This will be especially valuable to prioritize which validation experiments are most likely to be informative about regulatory network structure and function.

The specific assumptions that ARMADA makes suggest that it will likely be most effective on data-sets in which the time points are sampled densely enough such that motif activity changes are relatively small from one time point to the next. In addition, the assumption that external influences are approximately constant over the time course is likely most appropriate for systems in which the external conditions are held relatively constant.

There are several ways in which ARMADA can be further extended. Probably the most important of these is the incorporation of specific prior information regarding regulatory interactions that are more or less plausible and a major advantage of our Bayesian frame work is that such 'informative priors' can be easily incorporated. For example, instead of a constant precision α in the prior of each putative interaction $W_{m,n}$, we can easily give independent precisions $\alpha_{m,n}$ depending on the prior information that regulator m may target regulator n, e.g. through the existence of binding sites for regulator m in the promoter of the gene of regulator n, or through ISMARA's prediction that motif m targets a promoter of a TF associated with motif n. Indeed, we have already implemented such a scheme in ARMADA, allowing users to specify a matrix of prior parameters $\alpha_{m,n}$. However, at the time of completion of this manuscript we were still experimenting with different ways for setting the $\alpha_{m,n}$ in terms of TFBS and ISMARA target predictions, and we intend to report on a specific approach to setting informative priors in a future publication. Further in the future we intend to also provide meta-analysis of the interactions $W_{m,n}$ across multiple data-sets. That is, since the gene regulatory networks are ultimately hard-coded into the genome through the constellations of regulatory sites, we expect the same interactions $W_{m,n}$ to reappear across different systems and conditions. Another possible extension is to relax the assumption that outside influences are constant in time by, for example, allowing them to fluctuate over time, or to only exist for a fraction of the time course. In some situations we may have specific prior information about the time dependence of the driving forces and these could be directly incorporated in the model.

One of the ultimate goals of modeling gene regulatory networks is to understand how the genome-wide regulatory interactions implement the gene expression 'attractors' corresponding to different cell types in cellular eukaryotes. That is, we imagine that the interactions between regulators that shape gene expression dynamics as cells move from one state to another are ultimately also responsible for stabilizing the gene expression states of discrete cell types against perturbations. From this point of view it would be extremely interesting to compare the regulatory interactions that ARMADA infers from time course data, with regulatory interactions that can be inferred from the fluctuations in gene expression across single cells of a given cell type.

Author's contributions

EvN and PPR designed the investigation and the algorithm, MP and PPR implemented the webserver, PPR and EvN wrote the paper.

Acknowledgements

The authors express thanks to Piotr Balwierz for help with processing the sequencing data and providing access to ISMARA, and to Florian Geier and Olin Silander for useful suggestions. This work was supported by the CellPlasticity project grant of SystemsX.ch, the Swiss Initiative in Systems Biology.

References

[1] P.J. Balwierz, M. Pachkov, P. Arnold, A.J. Gruber, M. Zavolan, E. van Nimwegen, ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs, Genome Res. 24 (5) (2014) 869–884.
[2] F. Jacob, J. Monod, Genetic regulatory mechanisms in the synthesis of proteins, J. Mol. Biol. 4 (1961) 318–356.
[3] S.K. Kummerfeld, S.A. Teichmann, DBD: a transcription factor prediction database, Nucl. Acids Res. 34 (2006) D74–D81.
[4] A. Kozomara, S. Griffiths-Jones, miRBase: annotating high confidence microRNAs using deep sequencing data, Nucleic Acids Res. 42 (2014) 68–73. Database issue.
[5] K. Takahashi, S. Yamanaka, Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors, Cell 126 (4) (2006) 663–676.
[6] J.M. Vaquerizas, S.K. Kummerfeld, S.A. Teichmann, N.M. Luscombe, A census of human transcription factors: function, expression and evolution, Nat. Rev. Genet. 10 (4) (2009) 252–263.
[7] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K.R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J.M. Vaquerizas, R. Vincentelli, N.M. Luscombe, T.R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, J. Taipale, Dna-binding specificities of human transcription factors, Cell 152 (1–2) (2013) 327–339. <http://www.ncbi.nlm.nih.gov/pubmed/23332764>.
[8] H.J. Bussemaker, B.C. Foat, L.D. Ward, Predictive modeling of genome-wide mRNA expression: from modules to molecules, Annu. Rev. Biophys. Biomol. Struct. 36 (2007) 329–347.
[9] H.D. Kim, T. Shay, E.K. O'Shea, A. Regev, Transcriptional regulatory circuits: predicting numbers from alphabets, Science 325 (5939) (2009) 429–432.
[10] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, R. Guthke, Gene regulatory network inference: data integration in dynamic models-a review, BioSystems 96 (1) (2009) 86–103.
[11] FANTOM Consortium, H. Suzuki, A.R. Forrest, E. van Nimwegen, C.O. Daub, P.J. Balwierz, K.M. Irvine, T. Lassmann, T. Ravasi, Y. Hasegawa, M.J. de Hoon, S. Katayama, K. Schroder, P. Carninci, Y. Tomaru, M. Kanamori-Katayama, A. Kubosaki, A. Akalin, Y. Ando, E. Arner, M. Asada, H. Asahara, T. Bailey, V.B. Bajic, D. Bauer, A.G. Beckhouse, N. Bertin, J. Björkegren, F. Brombacher, E. Bulger, A.M. Chalk, J. Chiba, N. Cloonan, A. Dawe, J. Dostie, P.G. Engström, M. Essack, G.J. Faulkner, J.L. Fink, D. Fredman, K. Fujimori, M. Furuno, T. Gojobori, J. Gough, S.M. Grimmond, M. Gustafsson, M. Hashimoto, T. Hashimoto, M. Hatakeyama, S. Heinzel, W. Hide, O. Hofmann, M. Hörnquist, L. Huminiecki, K. Ikeo, N. Imamoto, S. Inoue, Y. Inoue, R. Ishihara, T. Iwayanagi, A. Jacobsen, M. Kaur, H. Kawaji, M.C. Kerr, R. Kimura, S. Kimura, Y. Kimura, H. Kitano, H. Koga, T. Kojima, S. Kondo, T. Konno, A. Krogh, A. Kruger, A. Kumar, B. Lenhard, A. Lennartsson, M. Lindow, M. Lizio, C. Macpherson, N Maeda, C.A. Maher, M. Maqungo, J. Mar, N.A. Matigian, H. Matsuda, J.S. Mattick, S. Meier, S. Miyamoto, E. Miyamoto-Sato, K. Nakabayashi, Y. Nakachi, M. Nakano, S. Nygaard, T. Okayama, Y. Okazaki, H. Okuda-Yabukami, V. Orlando, J. Otomo, M. Pachkov, N. Petrovsky, C. Plessy, J. Quackenbush, A. Radovanovic, M. Rehli, R. Saito, A. Sandelin, S. Schmeier, C. Schönbach, A.S. Schwartz, C.A. Semple, M. Sera, J. Severin, K Shirahige, C. Simons, G. St Laurent, M. Suzuki, T. Suzuki, M.J. Sweet, R.J. Taft, S. Takeda, Y. Takenaka, K. Tan, M.S. Taylor, R.D. Teasdale, J. Tegnér, S. Teichmann, E. Valen, C. Wahlestedt, K. Waki, A. Waterhouse, C.A. Wells, O. Winther, L. Wu, K. Yamaguchi, H. Yanagawa, J. Yasuda, M. Zavolan, D.A. Hume, Riken Omics Science Center, T. Arakawa, S. Fukuda, K. Imamura, C. Kai, A. Kaiho, T. Kawashima, C. Kawazu, Y. Kitazume, M. Kojima, H. Miura, K Murakami, M. Murata, N. Ninomiya, H. Nishiyori, S. Noma, C. Ogawa, T. Sano, C Simon, M. Tagami, Y. Takahashi, J. Kawai, Y. Hayashizaki, The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line, Nat. Genet. 41 (5) (2009) 553–562.
[12] K.M. Summers, S. Raza, E. van Nimwegen, T.C. Freeman, D.A. Hume, Co-expression of FBN1 with mesenchyme-specific genes in mouse cell lines: implications for phenotypic variability in Marfan syndrome, Eur. J. Hum. Genet. 18 (11) (2010) 1209–1215.
[13] N. Aceto, N. Sausgruber, H. Brinkhaus, D. Gaidatzis, G. Martiny-Baron, G. Mazzarol, S. Confalonieri, M. Quarto, G. Hu, P.J. Balwierz, M. Pachkov, S.J. Elledge, E. van Nimwegen, M.B. Stadler, M. Bentires-Alj, Tyrosine phosphatase

SHP2 promotes breast cancer progression and maintains tumor-initiating cells via activation of key transcription factors and a positive feedback signaling loop, Nat. Med. 18 (4) (2012) 529–537.
[14] E. Arner, N. Mejhert, A. Kulyte, P.J. Balwierz, M. Pachkov, M. Cormont, S. Lorente-Cebrian, A. Ehrlund, J. Laurencikiene, P. Heden, K. Dahlman-Wright, J.F. Tanti, Y. Hayashizaki, M. Ryden, I. Dahlman, E. van Nimwegen, C.O. Daub, P. Arner, Adipose tissue microRNAs as regulators of CCL2 production in human obesity, Diabetes 61 (8) (2012) 1986–1993.
[15] J. Perez-Schindler, S. Summermatter, S. Salatino, F. Zorzato, M. Beer, P.J. Balwierz, E. van Nimwegen, J.N. Feige, J. Auwerx, C. Handschin, The corepressor NCoR1 antagonizes PGC-alpha and estrogen-related receptor alpha in the regulation of skeletal muscle function and oxidative metabolism, Mol. Cell. Biol. 32 (24) (2012) 4913–4924.
[16] N. Tiwari, N. Meyer-Schaller, P. Arnold, H. Antoniadis, M. Pachkov, E. van Nimwegen, G. Christofori, Klf4 is a transcriptional regulator of genes critical for EMT, including Jnk1 (Mapk8), PLoS ONE 8 (2) (2013) e57329.
[17] S.J. Vervoort, A.R. Lourenco, R. van Boxtel, P.J. Coffer, SOX4 mediates TGF-beta-induced expression of mesenchymal markers during mammary cell epithelial to mesenchymal transition, PLoS ONE 8 (1) (2013) e53238.
[18] P.S. Eisele, S. Salatino, J. Sobek, M.O. Hottiger, C. Handschin, The peroxisome proliferator-activated receptor γ coactivator 1α/β (PGC-1) coactivators repress the transcriptional activity of NF-κB in skeletal muscle cells, J. Biol. Chem. 288 (4) (2013) 2246–2260.
[19] T. Suzuki, M. Nakano-Ikegaya, H. Yabukami-Okuda, M. de Hoon, J. Severin, S. Saga-Hatano, J.W. Shin, A. Kubosaki, C. Simon, Y. Hasegawa, Y. Hayashizaki, H. Suzuki, Reconstruction of monocyte transcriptional regulatory network accompanies monocytic functions in human fibroblasts, PLoS ONE 7 (3) (2012) e33474.
[20] R. Hasegawa, Y. Tomaru, M. de Hoon, H. Suzuki, Y. Hayashizaki, J.W. Shin, Identification of ZNF395 as a novel modulator of adipogenesis, Exp. Cell Res. 319 (3) (2013) 68–76.
[21] N. Tiwari, V.K. Tiwari, L. Waldmeier, P.J. Balwierz, P. Arnold, M. Pachkov, N. Meyer-Schaller, D. Schubeler, E. van Nimwegen, G. Christofori, Sox4 is a master regulator of epithelial-mesenchymal transition by controlling Ezh2 expression and epigenetic reprogramming, Cancer Cell 23 (6) (2013) 768–783.
[22] F. Meier-Abt, E. Milani, T. Roloff, H. Brinkhaus, S. Duss, D.S. Meyer, I. Klebba, P.J. Balwierz, E. van Nimwegen, M. Bentires-Alj, Parity induces differentiation and reduces Wnt/Notch signaling ratio and proliferation potential of basal stem/progenitor cells isolated from mouse mammary epithelium, Breast Cancer Res. 15 (2) (2013) R36.
[23] A.J. Gruber, W.A. Grandy, P.J. Balwierz, Y.A. Dimitrova, M. Pachkov, C. Ciaudo, E.v. Nimwegen, M. Zavolan, Embryonic stem cell-specific microRNAs contribute to pluripotency by inhibiting regulators of multiple differentiation pathways, Nucleic Acids Res. 42 (14) (2014) 9313–9326.
[24] M. Baresic, S. Salatino, B. Kupr, E. van Nimwegen, C. Handschin, Transcriptional network analysis in muscle reveals AP-1 as a partner of PGC-α in the regulation of the hypoxic gene program, Mol. Cell. Biol. 34 (16) (2014) 2996–3012.
[25] M. Diepenbruck, L. Waldmeier, R. Ivanek, P. Berninger, P. Arnold, E. van Nimwegen, G. Christofori, Tead2 expression levels control the subcellular distribution of Yap and Taz, zyxin expression and epithelial-mesenchymal transition, J. Cell. Sci. 127 (Pt 7) (2014) 1523–1536.
[26] R. Luisier, E.B. Unterberger, J.I. Goodman, M. Schwarz, J. Moggs, R. Terranova, E. van Nimwegen, Computational modeling identifies key gene regulatory interactions underlying phenobarbital-mediated tumor promotion, Nucleic Acids Res. 42 (7) (2014) 4180–4195.
[27] M.T. Dill, Z. Makowska, G. Trincucci, A.J. Gruber, J.E. Vogt, M. Filipowicz, D. Calabrese, I. Krol, D.T. Lau, L. Terracciano, E. van Nimwegen, V. Roth, M.H. Heim, Pegylated IFN- regulates hepatic gene expression through transient Jak/STAT activation, J. Clin. Invest. 124 (4) (2014) 1568–1581.
[28] P. Arnold, A. Scholer, M. Pachkov, P.J. Balwierz, H. Jørgensen, M.B. Stadler, E. van Nimwegen, D. Schubeler, Modeling of epigenome dynamics identifies transcription factors that mediate polycomb targeting, Genome Res. 23 (1) (2013) 60–73.
[29] M. de Hoon, Y. Hayashizaki, Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference, Biotechniques 44 (5) (2008) 627–628, 630, 632, http://dx.doi.org/10.2144/000112802.
[30] P.J. Balwierz, P. Carninci, C.O. Daub, J. Kawai, Y. Hayashizaki, W.V. Belle, C. Beisel, E. van Nimwegen, Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepcage data, Genome Biol. 10 (7) (2009) R79.
[31] J. Harrow, A. Frankish, J.M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B.L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J.M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, T.J. Hubbard, GENCODE: the reference human genome annotation for the ENCODE project, Genome Res. 22 (9) (2012) 1760–1774.
[32] V. Matys, E. Fricke, R. Geffers, E. Gssling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Mnch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, E. Wingender, TRANSFAC: transcriptional regulation, from patterns to profiles, Nucleic Acids Res. 31 (1) (2003) 374–378.
[33] A. Mathelier, X. Zhao, A.W. Zhang, F. Parcy, R. Worsley-Hunt, D.J. Arenillas, S. Buchman, C.Y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, W.W. Wasserman, JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles, Nucleic Acids Res. 42 (2014) D142–D147 (Database issue).
[34] P. Arnold, I. Erb, M. Pachkov, N. Molina, E. van Nimwegen, Motevo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of dna sequences, Bioinformatics 28 (4) (2012) 487–494. <http://bioinformatics.oxfordjournals.org/content/28/4/487.abstract>.
[35] R.C. Friedman, K.K.-H. Farh, C.B. Burge, D.P. Bartel, Most mammalian mRNAs are conserved targets of microRNAs, Genome Res. 19 (1) (2009) 92–105, http://dx.doi.org/10.1101/gr.082701.108.
[36] M. Pachkov, P.J. Balwierz, P. Arnold, E. Ozonov, E. van Nimwegen, SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates, Nucleic Acids Res. 41 (2013) D214–D220. <http://www.ncbi.nlm.nih.gov/pubmed/23180783> (Database issue).
[37] H. Ltkepohl, New Introduction to Multiple Time Series Analysis, Springer Publishing Company, Incorporated, 2007.
[38] D. Barber, Bayesian Reasoning and Machine Learning, Cambridge University Press, 2012.
[39] W. Penny, S. Roberts, Bayesian multivariate autoregressive models with structured priors, IEEE Proc. Vision Image Signal Process. 149 (1) (2002) 33–41, http://dx.doi.org/10.1049/ip-vis:20020149.
[40] S. Roberts, W. Penny, Variational bayes for generalized autoregressive models, IEEE Trans. Signal Process. 50 (9) (2002) 2245–2257, http://dx.doi.org/10.1109/TSP.2002.801921.
[41] Y. Wada, Y. Ohta, M. Xu, S. Tsutsumi, T. Minami, K. Inoue, D. Komura, J. Kitakami, N. Oshida, A. Papantonis, A. Izumi, M. Kobayashi, H. Meguro, Y. Kanki, I. Mimura, K. Yamamoto, C. Mataki, T. Hamakubo, K. Shirahige, H. Aburatani, H. Kimura, T. Kodama, P.R. Cook, S. Ihara, A wave of nascent transcription on activated human genes, Proc. Natl. Acad. Sci. U.S.A. 106 (43) (2009) 18357–18361, http://dx.doi.org/10.1073/pnas.0902573106.
[42] K. Inoue, M. Kobayashi, K. Yano, M. Miura, A. Izumi, C. Mataki, T. Doi, T. Hamakubo, P.C. Reid, D.A. Hume, M. Yoshida, W.C. Aird, T. Kodama, T. Minami, Histone deacetylase inhibitor reduces monocyte adhesion to endothelium through the suppression of vascular cell adhesion molecule-1 expression, Arterioscler. Thromb. Vasc. Biol. 26 (12) (2006) 2652–2659, http://dx.doi.org/10.1161/01.ATV.0000247247.89787.e7.
[43] H. Harada, E. Takahashi, S. Itoh, K. Harada, T.A. Hori, T. Taniguchi, Structure and regulation of the human interferon regulatory factor 1 (IRF-1) and IRF-2 genes: implications for a gene network in the interferon system, Mol. Cell Biol. 14 (2) (1994) 1500–1509.
[44] K. Calame, Activation-dependent induction of Blimp-1, Curr. Opin. Immunol. 20 (3) (2008) 259–264.
[45] C. Lien, C.M. Fang, D. Huso, F. Livak, R. Lu, P.M. Pitha, Critical role of IRF-5 in regulation of B-cell differentiation, Proc. Natl. Acad. Sci. U.S.A. 107 (10) (2010) 4664–4668.
[46] G.M. Doody, M.A. Care, N.J. Burgoyne, J.R. Bradford, M. Bota, C. Bonifer, D.R. Westhead, R.M. Tooze, An extended set of PRDM1/BLIMP1 target genes links binding motif type to dynamic repression, Nucleic Acids Res. 38 (16) (2010) 5336–5350.
[47] A. Matsukawa, M.H. Kaplan, C.M. Hogaboam, N.W. Lukacs, S.L. Kunkel, Pivotal role of signal transducer and activator of transcription (Stat)4 and Stat6 in the innate immune response during sepsis, J. Exp. Med. 193 (6) (2001) 679–688.
[48] X.L. Aranguren, M. Beerens, G. Coppiello, C. Wiese, I. Vandersmissen, A. Lo Nigro, C.M. Verfaillie, M. Gessler, A. Luttun, COUP-TFII orchestrates venous and lymphatic endothelial identity by homo- or hetero-dimerisation with PROX1, J. Cell. Sci. 126 (Pt 5) (2013) 1164–1175.
[49] X. Li, M.J. Large, C.J. Creighton, R.B. Lanz, J.W. Jeong, S.L. Young, B.A. Lessey, W.A. Palomino, S.Y. Tsai, F.J. Demayo, COUP-TFII regulates human endometrial stromal genes involved in inflammation, Mol. Endocrinol. 27 (12) (2013) 2041–2054.
[50] L. Escoubet-Lozach, C. Benner, M.U. Kaikkonen, J. Lozach, S. Heinz, N.J. Spann, A. Crotti, J. Stender, S. Ghisletti, D. Reichart, C.S. Cheng, R. Luna, C. Ludka, R. Sasik, I. Garcia-Bassets, A. Hoffmann, S. Subramaniam, G. Hardiman, M.G. Rosenfeld, C.K. Glass, Mechanisms establishing TLR4-responsive activation states of inflammatory response genes, PLoS Genet. 7 (12) (2011) e1002401.
[51] L.H. Glimcher, XBP1: the last two decades, Ann. Rheum. Dis. 69 (Suppl 1) (2010) 67–71.
[52] P.S. Gargalovic, N.M. Gharavi, M.J. Clark, J. Pagnon, W.-P. Yang, A. He, A. Truong, T. Baruch-Oren, J.A. Berliner, T.G. Kirchgessner, A.J. Lusis, The unfolded protein response is an important regulator of inflammatory genes in endothelial cells, Arterioscler. Thromb. Vasc. Biol. 26 (11) (2006) 2490–2496, http://dx.doi.org/10.1161/01.ATV.0000242903.41158.a1.
[53] M. Civelek, E. Manduchi, R.J. Riley, C.J. Stoeckert Jr, P.F. Davies, Chronic endoplasmic reticulum stress activates unfolded protein response in arterial endothelium in regions of susceptibility to atherosclerosis, Circ. Res. 105 (5) (2009) 453–461, http://dx.doi.org/10.1161/CIRCRESAHA.109.203711.
[54] R. Hu, A. Warri, L. Jin, A. Zwart, R.B. Riggins, H.B. Fang, R. Clarke, NF-κB signaling is required for XBP1 (unspliced and spliced)-mediated effects on antiestrogen responsiveness and cell fate decisions in breast cancer, Mol. Cell. Biol. 35 (2) (2015) 379–390.
[55] A. Kaser, A.-H. Lee, A. Franke, J.N. Glickman, S. Zeissig, H. Tilg, E.E.S. Nieuwenhuis, D.E. Higgins, S. Schreiber, L.H. Glimcher, R.S. Blumberg, XBP1 links ER stress to intestinal inflammation and confers genetic risk for human

inflammatory bowel disease, Cell 134 (5) (2008) 743–756, http://dx.doi.org/10.1016/j.cell.2008.07.021.
[56] M. Kitamura, Control of NF-kappaB and inflammation by the unfolded protein response, Int. Rev. Immunol. 30 (2011) 4–15.
[57] J. Li, J.J. Wang, S.X. Zhang, Preconditioning with endoplasmic reticulum stress mitigates retinal endothelial inflammation via activation of X-box binding protein 1, J. Biol. Chem. 286 (6) (2011) 4912–4921, http://dx.doi.org/10.1074/jbc.M110.199729.
[58] L.P. Zalmas, X. Zhao, A.L. Graham, R. Fisher, C. Reilly, A.S. Coutts, N.B. La Thangue, DNA-damage response control of E2F7 and E2F8, EMBO Rep. 9 (3) (2008) 252–259.
[59] N.M. Wheelhouse, Y.S. Chan, S.E. Gillies, H. Caldwell, J.A. Ross, D.J. Harrison, S. Prost, TNF-alpha induced DNA damage in primary murine hepatocytes, Int. J. Mol. Med. 12 (6) (2003) 889–894.
[60] M.P. Gantier, H.J. Stunden, C.E. McCoy, M.A. Behlke, D. Wang, M. Kaparakis-Liaskos, S.T. Sarvestani, Y.H. Yang, D. Xu, S.C. Corr, E.F. Morand, B.R. Williams, A miR-19 regulon that controls NF-B signaling, Nucleic Acids Res. 40 (16) (2012) 8048–8058.
[61] J.T. Mendell, miRiad roles for the miR-17-92 cluster in development and disease, Cell 133 (2) (2008) 217–222.
[62] K. Woods, J.M. Thomson, S.M. Hammond, Direct regulation of an oncogenic micro-RNA cluster by E2F transcription factors, J. Biol. Chem. 282 (4) (2007) 2130–2134.
[63] J. Chen, Z. Liu, Y. Yang, In vitro screening of LPS-induced miRNAs in leukocytes derived from cord blood and their possible roles in regulating TLR signals, Pediatr. Res. 75 (5) (2014) 595–602.
[64] J. Wu, J. Qian, C. Li, L. Kwok, F. Cheng, P. Liu, C. Perdomo, D. Kotton, C. Vaziri, C. Anderlind, A. Spira, W.V. Cardoso, J. Lu, miR-129 regulates cell proliferation by downregulating Cdk6 expression, Cell Cycle 9 (9) (2010) 1809–1818.
[65] M. Karaayvaz, H. Zhai, J. Ju, miR-129 promotes apoptosis and enhances chemosensitivity to 5-fluorouracil in colorectal cancer, Cell Death Dis. 4 (2013) e659.