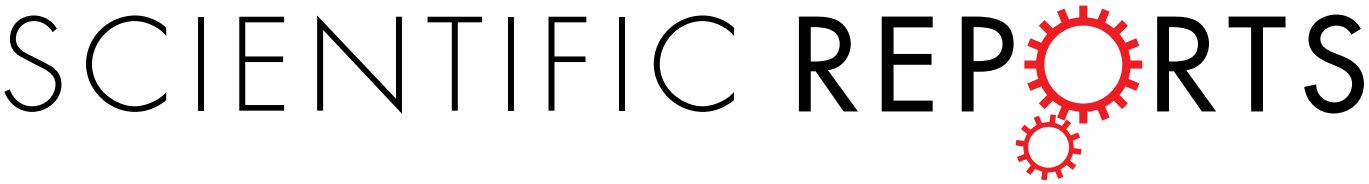

OPEN

Received: 08 September 2016
Accepted: 24 October 2016
Published: 23 November 2016

BGRMI: A method for inferring gene regulatory networks from time-course gene expression data and its application in breast cancer research

Luis F. Iglesias-Martinez[1], Walter Kolch[1,2,3] & Tapesh Santra[1]

Reconstructing gene regulatory networks (GRNs) from gene expression data is a challenging problem. Existing GRN reconstruction algorithms can be broadly divided into model-free and model–based methods. Typically, model-free methods have high accuracy but are computation intensive whereas model-based methods are fast but less accurate. We propose Bayesian Gene Regulation Model Inference (BGRMI), a model-based method for inferring GRNs from time-course gene expression data. BGRMI uses a Bayesian framework to calculate the probability of different models of GRNs and a heuristic search strategy to scan the model space efficiently. Using benchmark datasets, we show that BGRMI has higher/comparable accuracy at a fraction of the computational cost of competing algorithms. Additionally, it can incorporate prior knowledge of potential gene regulation mechanisms and TF hetero-dimerization processes in the GRN reconstruction process. We incorporated existing ChIP-seq data and known protein interactions between TFs in BGRMI as sources of prior knowledge to reconstruct transcription regulatory networks of proliferating and differentiating breast cancer (BC) cells from time-course gene expression data. The reconstructed networks revealed key driver genes of proliferation and differentiation in BC cells. Some of these genes were not previously studied in the context of BC, but may have clinical relevance in BC treatment.

Cellular functions depend on the precise regulation of thousands of genes which are activated or silenced by transcription factors (TFs)[1]. The networks representing the interactions between TFs and their target genes are typically known as GRNs and can be reconstructed from temporal measurements of gene expressions[2–5]. The GRN reconstruction methods can be classified into two main categories; model-based and model-free methods. Model-based methods aim to capture the regulatory interactions by fitting mathematical models of gene regulation to observed gene expression data[2,4,5]. On the other hand, model-free approaches use information-theoretic criteria to infer the structure of the network[3,6]. Though the performances of GRN reconstruction methods depend on several aspects such as data type, network properties of the GRN etc[7]., generally, model-based methods tend to be faster but have lower predictive performance than model-free methods 2015[6]. However, model-free methods are often not scalable enough to reconstruct genome-wide GRNs[5,6] in reasonable time. Typically, model-based methods formulate the expression of a gene as a function of its regulators, evaluate competing models containing different sets of regulators and chose those which closely predict the target gene expression[4,5,8]. Although vast majority of model based methods assume that the expressions of a gene and its regulators are linearly dependent[5,9–12], these methods use different model search algorithms e.g. Least Absolute Shrinkage and Selection Operator (LASSO), Dantzig Selector, elastic net, Markov Chain Monte Carlo and Heuristic search[5,12–19]. Some of these methods such as LASSO and elastic net choose the best model, others such as MCMC or Heuristic search based Bayesian Model Averaging (BMA) methods[5,12,13,15,18] select multiple models that provide close fits to the

[1]Systems Biology Ireland, University College Dublin, Belfield, Dublin 4, Republic of Ireland. [2]Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland. [3]School of Medicine and Medical Science, University College Dublin, Belfield, Dublin 4, Ireland. Correspondence and requests for materials should be addressed to T.S. (email: tapesh.santra@ucd.ie)

data and use these to estimate an average model along with its confidence interval. It is also possible to incorporate different types of existing data in BMA[5,12,18] to increase the accuracy of the reconstructed GRN.

We developed BGRMI, a model-based method that relies on the principles of BMA for inferring GRNs from time course gene expression data. BGRMI uses discretized ordinary differential equation (DODE) based mathematical models to formulate the interactions between each gene and its regulators. It formulates the rate of change in a gene's expression as a function of the expressions of its regulators, takes basal expression and self-regulation into account and therefore provides a more realistic model of gene regulation than many existing methods. These models are then used in a Bayesian framework to evaluate how likely a set of TFs is to regulate a certain gene. We developed a greedy heuristic search algorithm to explore different combinations of TFs and find the most likely TF combinations for each gene. The proposed algorithm is faster and more scalable than many existing methods. The average of some of the most likely models was then used to represent the regulatory model of the gene. We compared the accuracy of BGRMI against other methods using *in-silico* and *in-vivo* benchmark datasets. BGRMI consistently out-performed most of the other competing methods in our benchmarking study. We then showed how additional data sources, e.g. ChIP-seq and the protein-protein interaction (PPI) between TFs can be incorporated as prior knowledge in the core BGRMI formulation. Finally, we applied BGRMI to study the transcriptional mechanisms that lead to proliferation and differentiation in BC cells by combining ChIP-seq, PPI and time course gene expression profiles. Our study uncovered previously unknown transcriptional mechanisms that drive phenotypic changes in BC cells.

Method

A brief overview of the BGRMI algorithm is as follows. We first developed a mathematical model of TF-mediated gene regulations. The model can predict the temporal changes in the expressions of target genes using the temporal expression patterns of TFs as input. This model is then used to evaluate different combinations of TFs to find those that closely predict target gene expressions. This is done by iteratively exploring different TF combinations, calculating the posterior probability of each of these combinations to predict target gene expression, and selecting those with high probabilities. The average of selected gene regulation models for each gene is used as its regulation model. Below we describe each step of our algorithm in detail.

The mathematical model of gene regulation. We used a discretized form of ODEs to formulate the dependence of a target gene on its regulators as shown below.

$$\Delta mRNA_i(t) = \frac{[mRNA_i(t) - mRNA_i(t-\Delta t)]}{\Delta t} = \alpha_i + \boldsymbol{\beta_i}^T \begin{bmatrix} mRNA_i(t-\Delta t) \\ \boldsymbol{TF_i}(t-\Delta t) \end{bmatrix} + \varepsilon_i(t) \quad (1)$$

Here, $mRNA_i(t)$ is the expression of gene i at time t, α_i is the basal gene expression rate, $\boldsymbol{\beta_i}$ is a vector that contains the coefficients of self regulation (by means of degradation, auto-activation/inhibition) and the regulation by a set of TFs ($\boldsymbol{TF_i}$), $\boldsymbol{TF_i}(t-\Delta t)$ are the expressions of the TFs that regulate gene i at time $(t-\Delta t)$. $\varepsilon_i(t)$ is the model fitting error caused by the measurement noise in expression data. Since measurement noise is random $\varepsilon_i(t)$ is a random variable and typically has Gaussian distribution with zero mean and variance σ^2, i.e. $\varepsilon_i(t) \sim N(0, \sigma^2)$[12,18]. The error variance (σ^2) depends on many factors such as biological variability and measurement noise, and is typically unknown.

The posterior probability of a gene regulation model. We used Bayesian statistics to calculate this probability. There are two main components of Bayesian formulations; (a) the prior probability ($p(M_k)$) which represents how well a model (M_k) is supported by prior knowledge, and (b) the likelihood function ($p(\boldsymbol{mRNA_i}|M_k)$) which evaluates how well a model explains experimental data. By Bayes' rule[20], the posterior probability ($p(M_k|\boldsymbol{mRNA_i})$) is proportional to the product of these two entities and represents how well a model (M_k) is supported by prior knowledge and experimental data combined.

In the absence of prior knowledge, we assumed that sparse regulatory models (i.e. those involving fewer TFs) are a priori more likely than dense models (i.e. those involving a large number of TFs). This assumption was formulated by assigning the following prior distributions over regulatory models (M_k)[5]: $p(M_k) = L^{-2.66}$, where L is the number of regulators in the model. However, there is a wealth of publicly available information about GRNs of several organisms such as yeast, E. coli, humans etc. This information can be used to formulate more informative priors for reconstructing GRNs of these organisms. We shall discuss the formulation of priors for human GRNs using ChIP-seq data[12] in a later section where we describe the implementation of BGRMI on human transcriptomic data.

The likelihood of a gene regulation model (M_k) is the probability that the observed expression pattern ($\boldsymbol{mRNA_i}$) of gene i, can be predicted by the model (M_k), and has the following form[5,12]:

$$P(\boldsymbol{mRNA_i}|\boldsymbol{TF_i}, \alpha_i, \boldsymbol{\beta_i}, \sigma^2) = \prod_{t=\Delta t}^{T} N\left(\boldsymbol{\beta_i}^T \begin{bmatrix} mRNA_i(t-\Delta t) \\ \boldsymbol{TF_i}(t-\Delta t) \end{bmatrix}, \sigma^2\right) \quad (2)$$

The likelihood function in Eq. 2 depends on the model parameters ($\alpha_i, \boldsymbol{\beta_i}\, \sigma^2$), whose values are typically unknown. Therefore, we evaluated the average of the likelihood (Eq. 2) over all possible values of the model parameters. The average likelihood is also called the marginal likelihood ($P(\boldsymbol{mRNA_i}|M_k)$). To analytically calculate the marginal likelihood we assigned conjugate prior distributions to each of the unknown variables. These distributions represent our prior knowledge of how likely a parameter is to have a certain value. Following Fernandez *et al.*[13] we assigned uninformative Jeffrey's prior[13,21] distribution to the basal expression rates α_i, $p(\alpha_i) = 1$, which

implies that α_i is equally likely to have any real value. The regulation coefficients ($\boldsymbol{\beta}_i$) were assigned Zellner's g prior[13,22].

$$P(\boldsymbol{\beta}_i|\sigma^2, g, \mathbf{Z}) = N(0,\ \sigma^2(g\mathbf{Z}'Z)^{-1})$$

$$\mathbf{Z} = [\boldsymbol{mRNA_i}\ \ \boldsymbol{TF_i}] \tag{3}$$

which implies that $\boldsymbol{\beta}_i$ may have a wide range of positive and negative values depending on the Zellner's constant g and error variance σ^2. Note that σ^2 is unknown, and therefore we assigned a non-informative Jeffrey's prior[13,21], $p(\sigma) = \frac{1}{\sigma}$ which suggests that the probability of σ^2 is inversely proportional to itself, i.e. it is more likely to have smaller values than larger ones. The marginal likelihood ($P(\boldsymbol{mRNA_i}|M_k)$) is calculated by integrating the product of the likelihood and the above priors with respect to the unknown parameters and has the following form[5,12].

$$P(\boldsymbol{mRNA_i}|M_k) \propto \left(\frac{g}{g+1}\right)^{\frac{p}{2}} V. \tag{4}$$

$$V = \left(\frac{1}{g+1}SSE + \frac{g}{g+1}K\right)^{-\frac{n-1}{2}}$$

$$K = (\boldsymbol{mRNA_i} - \overline{\boldsymbol{mRNA_i}})'(\boldsymbol{mRNA_i} - \overline{\boldsymbol{mRNA_i}})$$

Here p is the number of TFs in the model M_k, and n is the number of observations, SSE is the squared error between the observed expressions of gene i and those predicted by the model when its parameters are estimated using linear regression. The marginal likelihood in Eq. 4 depends on the Zellner's constant g which is set to the following value recommended by Fernandez *et al.*[13]:

$$g = argmin\left(\frac{1}{n},\ \sqrt{\frac{p}{n}}\right) \tag{5}$$

The posterior probability ($P(M_k|\boldsymbol{mRNA_i})$) of a potential regulatory model (M_k) of gene i, is then calculated using the following formula:

$$P(M_k|\boldsymbol{mRNA_i}) = \frac{P(M_k)P(\boldsymbol{mRNA_i}|M_k)}{\sum_{k=1}^{K}P(M_k)P(\boldsymbol{mRNA_i}|M_k)} \tag{6}$$

Model Search. We developed a heuristic algorithm to search for models with high posterior probabilities. The proposed algorithm (Fig. 1) is inspired by Occam's Up[23] and Branch and Bound algorithms[24]. The procedure starts by evaluating the posterior probability of the null model (M^0) which does not have any regulator except itself. In the next step, the null model is expanded by adding one TF. Each candidate TF is added one by one and the posterior probabilities of the new models with a single TF (M^1) are evaluated. The models that have higher posterior probabilities than the null model are selected and their posterior probabilities are compared. The highest posterior probability is used as a cut-off for the next stage. The selected models are further expanded by adding a new TF. Each of the remaining TFs (the TFs other than the ones already in the model) is added one by one. The models which have higher posterior probability than the cut-off are then kept and compared, and the highest posterior probability is then selected as the new cut-off for the next stage. This process is repeated until adding a new TF does not improve the posterior probability any further. Below we provide a pseudocode for our algorithm.

Pseudocode.

```
P ← 0_{N×N} # Initialize probability matrix
For each gene G_i
{
Active ← 1 # Flag to terminate while loop
Th ← P(M_0) # Non-normalized posterior probability of the null model
MQ ← {G_j, j = 1…N, j ≠ i} # Initialize Model queue which contains all genes but G_i.
NC ← 0 # Initialize normalization constant
AM ← ∅ # Initialize accepted models.
While (Active==1)
  {
    For each M_j in MQ # For each model in the model queue
      {
        PM_j ← P(M_j) # Calculate non normalized posterior of model M_j
        if PM_j > Th

            {  AM ∪ M_j # Add model j to the set of accepted models
        I_{M_j} ← Indexes of genes in M_j
        P(i, I_{M_j}) = P(i, I_{M_j}) + PM_j # Update the posterior of an edge
```

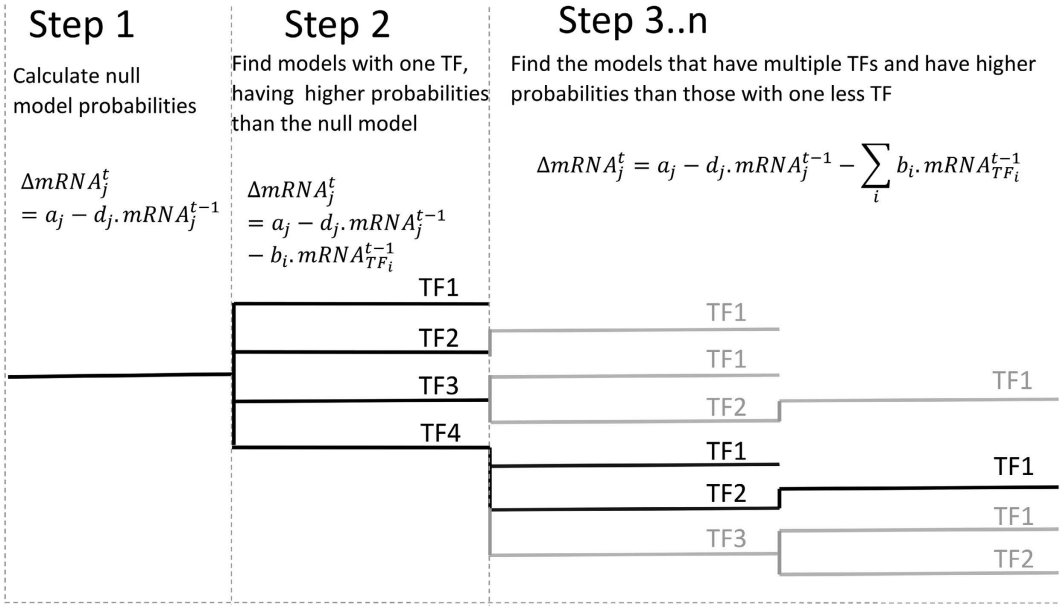

Figure 1. Workflow of the heuristic model search algorithm. On the first step the marginal likelihood of the null model is calculated. Then each TF is evaluated as a variable independently and only TFs whose marginal likelihood is higher than the null model's are further expanded. The highest marginal likelihood of the single TF models is selected as a threshold or bound to evaluate the nested models with two TFs.

NC ← *NC* + *PM*$_j$; # *Update the normalization constant*
} #*End of if*
} # *End of for*
MQ ← *All models that can be generated by extending each model in* ***AM*** *with one more gene.*

if AM==∅ # *If* ***AM*** *is empty*
{

Active = 0 # *set Active to zeros*
}# *End of if*
AM ← ∅ # *Empty* ***AM***
} # *End of while*
P(*i*,:) = ***P***(*i*,:)/*NC*;
} # *End of for*

Model averaging. The models selected by the above algorithm are used to estimate the probability of each TF-gene interaction and its strength. The probability that a TF (*j*) regulates a gene (*i*) is the sum of the probabilities of the models which include the TF (*j*)[15], i.e.

$$P_{ij} = \sum_{k=1}^{K} (\delta_{ik}) P(M_k | \boldsymbol{mRNA_i}). \tag{7}$$

Here *K* is the number of selected models, $\delta_{ik} = 1$ if TF *j* is part of model *k* and $\delta_{ik} = 0$ otherwise. The interaction strength between a TF (*j*) and its target gene (*i*) is calculated by taking weighted average of its expected value in each selected model (M_k), the weight being the posterior probability $P(M_k|\boldsymbol{mRNA_i})$ of the model (M_k)[15]

$$\beta_{ij} = \sum_{k=1}^{K} \left(E\left(\beta_{i,j}^k\right) \right) P(M_k | \boldsymbol{mRNA_i})$$

$$E\left(\beta_{i,j}^k\right) = \widehat{\beta_{i,j}^k}(1 + g) \tag{8}$$

Here $\widehat{\beta_{i,j}^k}$ is the maximum likelihood estimate of the regulation coefficient of the TF *j* on gene *i* in model M_k. If β_{ij} is positive then we assume that the TF *j* is an activator of gene *i* and if it is negative, the opposite is true.

	Size 10 Networks						
	BGRMI	**Jump3**	**GENIE3**	**CLR**	**Inferelator**	**ScanBMA**	**G1DBN**
Net1	0.635	0.498	0.555	0.465	**0.643**	—	0.564
Net2	**0.524**	0.396	0.351	0.447	0.443	—	0.392
Net3	**0.566**	0.44	0.407	0.414	0.509	—	0.499
Net4	0.751	0.584	0.519	0.555	0.653	—	**0.76**
Net5	0.615	0.646	0.787	**0.885**	0.637	—	0.77
Average	**0.618**	0.513	0.524	0.553	0.577	0.505	0.597
Net 1	0.245	**0.27**	0.228	0.179	0.126	—	0.089
Net2	**0.118**	0.11	0.096	0.109	0.101	—	0.055
Net3	0.185	0.2	0.23	**0.238**	0.198	—	0.155
Net4	**0.213**	0.18	0.157	0.154	0.147	—	0.153
Net5	0.154	**0.174**	0.168	0.163	0.148	—	0.117
Average	0.184	**0.187**	0.176	0.167	0.144	0.101	0.114
	Overall average						
All Nets	**0.401**	0.35	0.35	0.36	0.3605	0.303	0.3555

Table 1. AUPRs for the DREAM4 Networks. The numbers in bold represent the best performer. The authors of the scanBMA algorithm published only the average AUPR for size 10 and size 100 category, therefore we showed only the average AUPRs for scanBMA.

	BGRMI	**Jump3**	**GENIE3**	**CLR**	**Inferel-ator**	**Scan BMA**	**TSN1**	**G1DBN**
Switch -On Dataset	**0.904**	0.685	0.62	0.423	0.718	0.455	0.706	0.6
Switch-Off Dataset	0.574	**0.682**	0.347	0.372	0.649	0.232	0.511	0.313

Table 2. AUPRs of the *In Vivo* IRMA Network. The numbers in bold represent the best performers.

Network	**No. of Genes**	**No. of Observations**	**No. of Regulators**	**Running Time**
IRMA	5	~62	5	0.03 secs
DREAM4 10	10	105	10	0.32 secs
DREAM4 100	100	210	100	~4 mins

Table 3. Execution times of the BGRMI algorithm.

Results

We first evaluated BGRMI's accuracy on several *in silico* and *in vivo* benchmark datasets and compared its performance with other algorithms. Then we applied BGRMI to study human BC transcription regulatory network. Below we discuss the results of our analysis in detail.

The DREAM4 *In Silico* Network Inference Challenge dataset. The DREAM4 *In Silico* Network Challenge contains ten *in silico* GRNs, five of which consist of *10* genes each and the remaining five have *100* genes each. The dynamics of each of these networks in response to a series of perturbations were simulated and the resulting time course gene expression profiles were published by the DREAM consortium[6] for benchmarking network inference methods. We used BGRMI to analyse these data and calculate the probabilities and strengths of all possible interactions in each of these networks. The interactions which have higher probabilities than a pre-determined threshold constitute the reconstructed GRNs. The accuracy of the reconstructed GRNs is estimated by comparing these with the gold-standard networks and is dependent on the choice of the threshold probability. We used Precision Recall (PR) curve[25] to estimate these accuracies in an unbiased manner, independently of particular choices of threshold probabilities. PR curve is calculated by gradually increasing the threshold probability from 0 to 1, and for each threshold, calculating the precision and recall of the GRN reconstructed at that threshold[25]. Precision and recall are the ratios of the numbers of correctly inferred interactions vs all interactions in the reconstructed and the gold standard networks respectively[25]. The Area under the PR curve (AUPR) provides an unbiased scalar estimate of the accuracies of the reconstructed GRNs[5,6]. The AUPR values of the 10 and 100 gene networks reconstructed by the BGRMI algorithm are provided in Table 1. For comparison, we have also provided the AUPR values of the networks reconstructed by several other state-of-the-art algorithms, e.g. Jump3 the lagged time variant of GENIE3[3], CLR[26], Inferelator[4], G1DBN[27], and ScanBMA[5], which also claimed to have performed very well on the same datasets. BGRMI consistently performed well, achieving the highest AUPRs in 4 out of 10 networks (2 each of the 10 and 100 genes networks). It also achieved the highest average AUPR (0.401) across all ten datasets (Table 1), a noticeable improvement over its closest competitor Inferelator (avg. AUPR = 0.3605 across all ten datasets.)

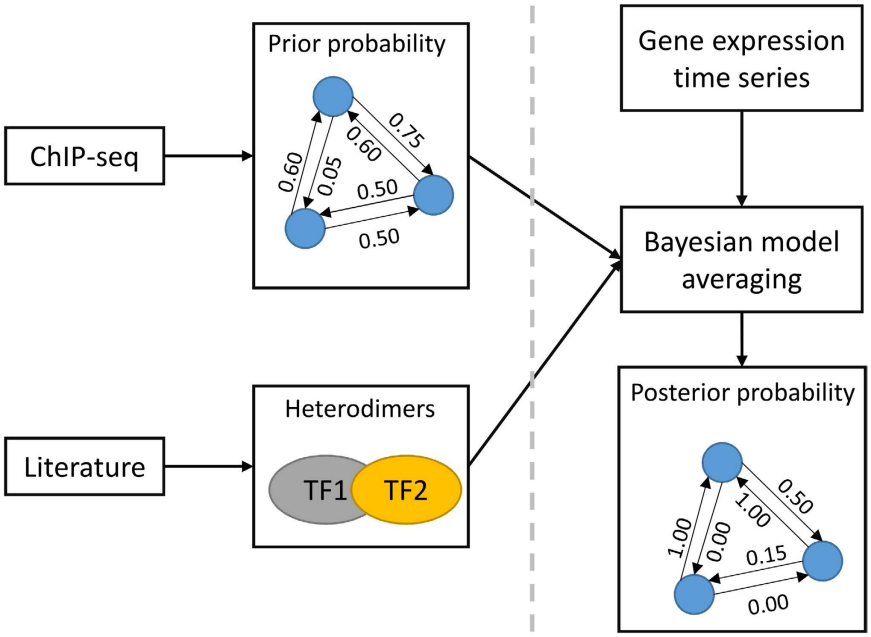

Figure 2. Workflow of BGRMI implementation on time course gene expression profiles on human BC cells.

***In Vivo* benchmark data.** To further test BGRMI we used time course gene expression data from a synthetic GRN, called *In vivo* Reverse-engineering and Modeling Assessment (IRMA) network, which was purposefully built to assess the performances of network reconstruction methods[28]. The IRMA network was synthesized in the yeast *Saccharomyces cerevisiae*. The network has 5 genes and 6 regulatory interactions and can be switched on or off by culturing cells in galactose or glucose, respectively. The expression levels of the genes in the network were measured using quantitative RT-PCR at different time points in two different sets of experiments. In the first set, cells were stimulated with galactose and the network was switched on, whereas in the second set the network was switched off by adding glucose.

Table 2 shows the AUPRs of the GRNs reconstructed by all methods which were used for performance comparison of in-silico data. Additionally, we added the performance of the TSNI algorithm[29] which was originally used to reconstruct the IRMA network[28]. BGRMI had the highest accuracy for the Switch-On dataset by a large margin. However, on the Switch-Off dataset, Jump3 performed the best. These results suggest BGRMI performs well, not only on in-silico datasets but also on *in-vivo* experimental data.

Execution time of BGRMI. We measured the execution time of our method on the DREAM4 and IRMA networks. We used a 32-GB RAM, 1.7 GHz Intel core i7 computer. The results are summarized in Table 3.

Uncovering transcriptional mechanism governing proliferation and differentiation in BC cells. Several types of BCs are formed when breast tissue cells stop differentiating and keep proliferating[30]. Therefore, it is important to determine the molecular mechanisms that govern proliferation and differentiation in these cells. For this purpose, Mina *et al.*[31] measured time course gene expression profiles of MCF-7 BC cells after artificially inducing proliferation and differentiation by stimulating these cells with Heregulin (HRG) and Epidermal Growth Factors (EGF), respectively[31]. We used the resulting data to reconstruct the GRNs that orchestrate differentiation and proliferation in MCF-7 cells (Fig. 2). To increase the accuracy of reconstructed GRN we integrated ChIP-seq and PPI data[12] into the core network reconstruction algorithm (Fig. 2). The ChIP-seq data, which give us quantitative measurements of bindings between TFs and DNA molecules, was used to formulate prior probabilities of different gene regulation models. Additionally, PPIs between TFs were used to incorporate TF-heterodimers into our gene regulation model. Below we describe our implementation of BGRMI on the aforementioned dataset.

Formulating the prior probability of the gene regulation models. The prior probability of a gene regulation model (M_k) is formulated as the probability that a certain set of TFs ($TF_j, j = 1 \ldots K$) regulate a specific gene and the probability of observing a certain number of TFs on a target gene. To calculate this probability, we first estimated the probability (P_{ij}) that an individual TF (j) binds to gene (i). This probability (P_{ij}) is defined as the product of two quantities:

$$P_{ij} = Q_{ij} R_j \tag{9}$$

where Q_{ij} is the probability that the position in which the TF (j) was bound affects the expression of the target gene (i), and R_j is the probability that the TF (j) binds to the same position across different cell-types. Q_{ij} was calculated by combining different datasets from Gerstein *et al.*[32] who built models of consensus human transcription

regulatory networks by analysing the ENCODE data. They generated three models of human transcription regulatory networks, a proximal unfiltered network, a proximal network, and a distal network. The unfiltered proximal network consists of TF-gene interactions where the TF binds close to the promoter of the gene. The proximal filtered network consists of only those TF-gene interactions where the TF binds close to the promoter of the gene and their expressions are significantly correlated. The distal network represents the TF- gene interactions where the TF binds to the enhancer region of the gene. Q_{ij} was assigned a value of 1 for the TF-gene interactions found in the proximal network, 0.5 for those found only in the unfiltered proximal and distal networks, 0.05 for those not found in any of the above networks. R_j was estimated from the ENCODE ChIP-seq data using an in-house MATLAB script (freely available from https://github.com/Luisiglesiasmartinez/Peak-Merging). It should be noted that the ENCODE database does not have sufficient data to estimate (R_j) for each individual TF. Therefore we selected CTCF, a TF which has the most ChIP-seq data (98 datasets) in the ENCODE database, calculated its R_j ($\approx$0.26) and used this value for all TFs.

The probabilities (P_{ij}) of individual TFs are then combined using the following formula[5] to calculate the probability of a gene regulation model involving multiple TFs.

$$P(M_k) = \prod_{j=1}^{k} P_{ij}^{\delta_{ij}} \left(1 - P_{ij}\right)^{1-\delta_{ij}} L^{-2.66} \tag{10}$$

Here $\delta_{ij} = 1$ if the TF j is included in the model M_k and $\delta_{ij} = 0$ otherwise. L is the number of regulators in the model.

Incorporating TF-TF heterodimers in the formulation of gene regulation models. We gathered information about heterodimer formation between TFs from the literature[33–48]. Inspired by the interaction terms in linear regression models (https://en.wikipedia.org/wiki/Interaction_(statistics)), the expression of a heterodimer (TF_{j-l}) composed of any two TFs, j and l is calculated by multiplying the expressions of their individual mRNAs, i.e.

$$TF_{j-l} = mRNA_{TFj} \times mRNA_{TFl} \tag{11}$$

The heterodimers were then treated as separate potential regulators, along with the monomer forms of TFs, in different gene regulation models.

Data pre-processing. Gene expressions in Mina *et al.*'s dataset[31] were measured using cap analysis of gene expression (CAGE). CAGE uses tags from 5′ ends of cDNAs, which can be used to identify the specific expression of transcription start sites (TSSs) of the same gene[49]. For simplicity, we combined the normalized read counts for different isoforms of the same genes. The resulting data was then analysed using BGRMI. Note that the ENCODE database has ChIP-seq data for only 140 TFs, and therefore BGRMI inferred regulatory interactions involving these TFs only.

Post-processing of reconstructed networks. BGRMI estimated the posterior probabilities of each possible TF-DNA interaction for differentiating and proliferating MCF-7 cells. We kept only the interactions with posterior probabilities higher than 0.75. The interactions were assumed to be either inhibitory of activating depending on the sign of the regulation coefficients (β_{ij}).

Large differences between the GRNs that regulate differentiation and proliferation in MCF7 cells. BGRMI found 22692 and 19016 regulatory interactions for the HRG and EGF stimulated cells (Fig. 3A,B). The complete list of interactions is available as Supplementary Data 1. 10804 and 8997 of all interactions in the HRG and EGF induced networks were inhibitory regulations and the remaining were activating regulations. Surprisingly, only 286 of all the inferred interactions were common in both networks. However, the number of common interactions depends on the cut-off probability and for lower cut-offs more interactions were found common between these networks (Supplementary Fig. S1). The large difference between EGF and HRG induced GRNs suggests that the same genes are regulated by different sets of TFs in these two networks.

Transcriptional hubs in HRG and EGF induced GRNs. We sorted the TFs based on the number of their predicted targets (out-degree) in both GRNs and found that these networks have different sets of transcriptional hubs, i.e. TFs with a large number of targets (Fig. 3A,B).

In the EGF induced GRN, SIX5, CHD2, GATA2, ZEB1, NR4A1, ESRRA, and FOXA1 were found to be the largest hubs. GATA2, ZEB1, NR4A1, ESRRS and FOX1 are known to play crucial roles in the proliferation of BC cells[50–54]. In a recent study, SIX5 was shown to correlate with clinic-pathological parameters, e.g. tumour stage, size etc., of BC patients[55]. However, its specific role in BC cell proliferation is largely unknown. To the best of our knowledge, CHD2 was not previously studied in the context of BC. We analysed survival and gene expression data of BC patients from several sources, e.g. the TCGA database (http://cancergenome.nih.gov/) and all sources used by the kmplot webtool (http://kmplot.com/analysis/)[56] to further investigate the role of CHD2 and SIX5 in breast cancer progression. Firstly, in the TCGA dataset, the expression of the SIX5 and CHD2 were found to be significantly different (p-values 0.00000071, 0.0016 respectively, based on the Kruskal-Wallis test (Supplementary Figs S2 and S3) among patients of different BC subtypes including normal like, Luminal A, Luminal B, Her2 positive, and triple negative BC (TNBC), which vary in their aggressiveness. Also, in TNBC, the most aggressive and highly proliferative form of BC, patients survived significantly longer when they had low

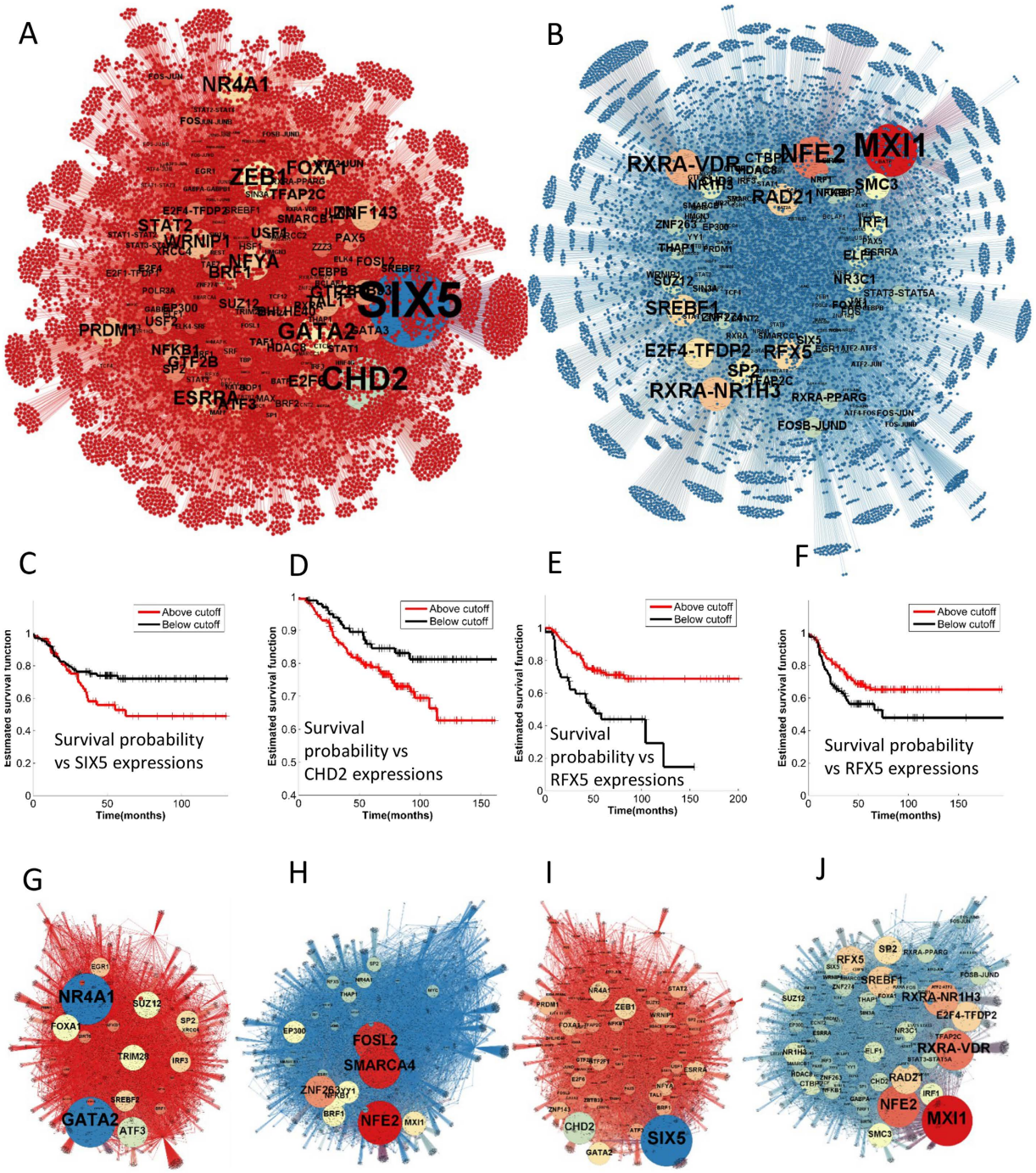

Figure 3. The EGF and HRG induced GRN in BC cells and the clinical relevance of some of its transcriptional hubs. (**A**,**B**) EGF and HRG induced GRNs in MCF7 cells with node size proportional to outdegree (number of targets). (**C**) Kaplan Meier plot for BC patient survival probability for different levels of SIX5 expression. (**D**) Kaplan Meier plot for survival probabilities of BC patients who underwent endocrine therapy for different levels of CHD2 expression. (**E**) Kaplan Meier plot for HER2 positive BC patient survival for different levels of RFX5 expression. (**F**) Kaplan Meier plot for TNBC patients survival probabilities for different levels of RFX5 expression. In (**C**–**F**) the red and black curves show survival probabilities for higher and lower expression of the corresponding markers respectively. (**G**,**H**) EGF and HRG induced GRNs in MCF7 cells with node size proportional to their betweenness centralities. (**I**,**J**) EGF and HRG induced GRNs in MCF7 cells with node size proportional to their page-rank.

SIX5 expression than when they featured high levels of SIX5 (Fig. 3C). In Liu *et al.*'s study[57], SIX5 expression had a statistically significant association with the response of cancer cells to the HER2 inhibitor Lapatinib (p-value *0.014*) and the MEK inhibitor PD-0325901 (p-value *0.0184*), both of which inhibit proliferation in cancer cells[58,59].

The expression of CHD2, a chromatin remodeller, did not correlate with BC patient survival. However, we found that patients who have undergone endocrine therapy, a chemopreventive measure targeting the estrogen receptor which promotes proliferation in BC cells, are significantly more likely to survive if they have relatively low level of CHD2 expression than those who have a high level of CHD2 (Fig. 3D). Furthermore, CHD2 expression has statistically significant association (p-value *0.029*) with the response of cancer cells to CDK inhibitor PD-0332991[57] which inhibits proliferation[60]. The above results not only supports our finding that SIX5 and CHD2 may play a crucial role in the proliferation of BC cells but also indicates that they may have potential clinical relevance in designing new BC treatments.

In the HRG induced GRN, MXI1, NFE2, RXRA-VDR complex, RXRA-NR1H3 complex, RAD21, RFX5 and SREBF1 are some of the largest transcriptional hubs. HRG induced differentiation of mammary cells is characterized by the synthesis of lipid droplets. Interestingly, two of the aforementioned transcriptional hubs, RXRA-NR1H3 complex and SREBF1, have been previously described as master regulators of lipid synthesis in mammary epithelial cells[61,62], corroborating our results. Among the remaining hubs, MXI1, NFE2, RXRA-VDR and RAD21 have known role in cell differentiation[63–66]. To the best of our knowledge, RFX5 does not have any previously known association with BC cell differentiation. Our analysis of gene expression and patient survival data reveals that RFX5 expression varies significantly (p-value $1.99415e^{-17}$, see Supplementary Fig. S4) among normal, Luminal A, Luminal B, Her2 positive and TNBC patients. Furthermore, patients of poorly differentiated BC subtypes, e.g. basal or HER2 positive BC[67,68] with higher RFX5 expression are significantly more likely to survive longer than those with lower levels of RFX5 (Fig. 3E,F). These data reveal a potential clinical relevance of RFX5 in designing new BC treatment.

Transcriptional junctions in HRG and EGF induced GRNs. In a typical GRN, information flows through intricate networks of successive activation and/or deactivation of TFs. Some TFs play crucial roles in the genetic information flow by residing at the junction of several transcriptional pathways. A network theoretic measure, 'betweenness centrality'[69], quantifies how busy a transcriptional junction is. The betweenness centrality (b_i) of gene i is calculated as follows.

$$b_i = \sum_{j \neq k} \frac{n_{jk}(i)}{n_{jk}} \tag{12}$$

Here b_i is the betweenness centrality of gene i, n_{jk} is the number of shortest paths from gene j to gene k, and $n_{jk}(i)$ the number of shortest paths from gene j to gene k that pass through gene i. We calculated the betweenness centralities for each transcription factor in the EGF and HRG induced GRNs (Fig. 3G,H). Our results suggest that NR4A1, GATA2, ATF3, SUZ12 and FOXA1 are some of the busiest junctions (have highest betweenness centralities) in the EGF induced GRN. GATA2, FOXA1 and NR4A1 were also found as transcriptional hubs in the same network, whereas, ATF3 and SUZ12 were recently shown to play crucial roles in the proliferation of breast cancer cells[70,71]. In the HRG induced GRN, NFE2, SMARCA4, FOSL2, ZNF263 and MXI1 were found to be the largest junctions. Among these, NFE2 and MXI1 were also found to large hubs, whereas SMARA4, FOSL2 and ZNF263 were previously shown to play important role in mammary cell differentiation[72–74].

Transcriptional master regulators in EGF and HRG induced GRNs. Another important class of TFs is the master regulators which regulate large transcriptional hubs. These can be identified by calculating the 'page rank' (see Brin *et al.*[75] for details) of each TF, and then find the TFs with the highest page ranks. SIX5, CHD2, GATA2, ZEB1, NR4A1 were found to have the highest page ranks in the EGF induced network, whereas, MXI1, NFE2, RXRA-VDR, RXRA-NR1H3, SREBF1 had the highest page-rank in the HGR induced networks, further highlighting the importance of these molecules in proliferation and differentiation of breast cancer cells (Fig. 3I,J).

Discussion

Deciphering GRNs is fundamental to understanding cellular decision making. Experimental reconstruction of GRNs is not feasible since current experimental methods produce snapshots of the genomic activities, but such data do not reveal the underlying regulatory mechanisms. Several computational methods had been proposed to reconstruct GRNs from experimental data. Many of these methods fail to strike a balance between scalability and accuracy. In this paper, we presented BGRMI, a Bayesian algorithm that can reconstruct quantitative models of GRNs from time course gene expression data. The main advantages of BGRMI are its speed/scalability while having comparable or higher accuracy than the current state of the art methods. Additionally, BGRMI can incorporate prior information from other data sources such as ChIP-seq and PPI databases to increase the accuracies of the reconstructed GRNs. Many recent GRN reconstruction methods, e.g. RNEA[76], PANDA[77], PTHGRN[78], APG[79], CMGRN[80], BVS[12] also have this feature. However, these algorithms have their advantages and disadvantages. For instance, PANDA[77] and RNEA[76] use gene expression data to find co-expressed and differentially expressed genes respectively, which are then combined with ChIP-seq and PPI data to reconstruct GRN topologies. Therefore, these approaches are not suitable for reconstructing GRNs if there are no prior ChIP-Seq/PPI data available. While most algorithms use PPI data to determine transcriptional co-regulators, BGRMI uses this data to infer regulatory programs of TF-complexes. Arguably, this yields clearer and more realistic pictures of GRNs than those containing interactions between individual TFs and their target genes. To demonstrate the practical applicability of BGRMI, we used it to reconstruct the GRNs of proliferating and differentiating BC cells, revealing strikingly different regulatory programs governing these phenotypes. Topological comparison of reconstructed GRNs revealed a number of key transcriptional regulators which play essential roles in BC cell proliferation and differentiation. Three of these TFs, SIX5, CHD2 and RFX5, were not previously studied in these contexts and

therefore may shed new light in understanding how BC cells decide to proliferate or differentiate. Expressions of these TFs were found to be predictive of BC patient survival or their responsiveness to Endocrine therapy. Therefore, these molecules may have clinical relevance in treating BC patients. Furthermore, the reconstructed GRNs can potentially be used to predict new therapeutic targets for BC. For instance, recent studies[81–84] demonstrated that it is possible to predict therapeutic targets for different types of cancer by integrating the respective GRNs with mutation data, miRNA data and functional RNAi/phenotypic screens.

Nevertheless, BGRMI has some limitations. Firstly, it uses mRNA levels of TFs as proxy for their activities. This can lead to spurious results since the activity of a TF can depend on posttranslational modifications of its protein form and may not always be directly related to its expression[85]. Secondly, changes in gene expressions may be induced by mechanisms other than transcription regulation, e.g. epigenetic regulation. However, BGRMI cannot differentiate between the mechanisms of gene regulation and assumes that any observed change in the gene expression is caused by transcriptional regulation. Finally, BGRMI uses prior knowledge on DNA binding preferences of TFs and PPIs among TFs, which is available for a limited number of TFs. However, other data such as gene ontology (GO) annotations, protein abundance, protein phosphorylation datasets may provide important clue in the transcriptional activities of relatively less studied TFs, but are not currently used by BGRMI.

References

1. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* **13,** 613–626, doi: 10.1038/nrg3207 (2012).
2. Bonneau, R. *et al.* The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology* **7,** doi: 10.1186/gb-2006-7-5-r36 (2006).
3. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *Plos One* **5,** doi: 10.1371/journal.pone.0012776 (2010).
4. Madar, A., Greenfield, A., Ostrer, H., Vanden-Eijnden, E. & Bonneau, R. The Inferelator 2.0: a scalable framework for reconstruction of dynamic regulatory network models. *Paper presented at Annual International Conference of the IEEE Engineering in Medicine and Biology Society* Washington DC, USA.New York, USA, IEEE, doi: 10.1109/iembs.2009.5334018 (2009, Nov 1–4).
5. Young, W. C., Raftery, A. E. & Yeung, K. Y. Fast Bayesian inference for gene regulatory networks using ScanBMA. *Bmc Systems Biology* **8,** doi: 10.1186/1752-0509-8-47 (2014).
6. Huynh-Thu, V. A. & Sanguinetti, G. Combining tree-based and dynamical systems for the inference of gene regulatory networks. *Bioinformatics* **31,** 1614–1622, doi: 10.1093/bioinformatics/btu863 (2015).
7. Madhamshettiwar, P. B., Maetschke, S. R., Davis, M. J., Reverter, A. & Ragan, M. A. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Medicine* **4,** 41, doi: 10.1186/gm340 (2012).
8. Michailidis, G. & d'Alche-Buc, F. Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical Biosciences* **246,** 326–334, doi: 10.1016/j.mbs.2013.10.003 (2013).
9. Huang, X. & Zi, Z. Inferring cellular regulatory networks with Bayesian model averaging for linear regression (BMALR). *Molecular BioSystems* **10,** 2023–2030 (2014).
10. Kim, H. & Gelenbe, E. Reconstruction of Large-Scale Gene Regulatory Networks Using Bayesian Model Averaging. *IEEE Transactions on NanoBioscience* **11,** 259–265, doi: 10.1109/TNB.2012.2214233 (2012).
11. Li, Z., Li, P., Krishnan, A. & Liu, J. Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. *Bioinformatics* **27,** 2686–2691, doi: 10.1093/bioinformatics/btr454 (2011).
12. Santra, T. A Bayesian Framework that integrates heterogeneous data for inferring gene regulatory networks. *Frontiers in Bioengineering and Biotechnology* **2,** doi: 10.3389/fbioe.2014.00013 (2014).
13. Fernandez, C., Ley, E. & Steel, M. F. J. Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100,** 381–427, doi: 10.1016/s0304-4076(00)00076-2 (2001).
14. Ghanbari, M., Lasserre, J. & Vingron, M. Reconstruction of gene networks using prior knowledge. *BMC Systems Biology* **9,** 1–11, doi: 10.1186/s12918-015-0233-4 (2015).
15. Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. Bayesian model averaging: A tutorial. *Statistical Science* **14,** 382–401 (1999).
16. Omranian, N., Eloundou-Mbebi, J. M. O., Mueller-Roeber, B. & Nikoloski, Z. Gene regulatory network inference using fused LASSO on multiple data sets. *Scientific Reports* **6,** 20533, doi: 10.1038/srep20533 (2016).
17. Ruyssinck, J. *et al.* NIMEFI: Gene Regulatory Network Inference using Multiple Ensemble Feature Importance Algorithms. *PLoS ONE* **9,** e92709, doi: 10.1371/journal.pone.0092709 (2014).
18. Santra, T., Kolch, W. & Kholodenko, B. N. Integrating Bayesian variable selection with Modular Response Analysis to infer biochemical network topology. *BMC Systems Biology* **7,** 1–19, doi: 10.1186/1752-0509-7-57 (2013).
19. Vignes, M. *et al.* Gene Regulatory Network Reconstruction Using Bayesian Networks, the Dantzig Selector, the Lasso and Their Meta-Analysis. *PLoS ONE* **6,** e29165, doi: 10.1371/journal.pone.0029165 (2011).
20. Bayes, M. & Price, M. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions* **53,** 370–418, doi: 10.1098/rstl.1763.0053 (1763).
21. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **186,** 453–461 (1946).
22. Tiao, G. C. & Zellner, A. Bayes's Theorem and the Use of Prior Knowledge in Regression Analysis. *Biometrika* **51,** 219–230, doi: 10.2307/2334208 (1964).
23. Madigan, D. & Raftery, A. E. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89,** 1535–1546 (1994).
24. Narendra, P. M. & Fukunaga, K. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers* **100,** 917–922 (1977).
25. Davis, J. & Goadrich, M. *The relationship between Precision-Recall and ROC curves. Paper presented at Proceedings of the 23rd International Conference on Machine Learning.* Pittsburgh, USA.New York, USA, ACM (2006, June 29).
26. Faith, J. J. *et al.* Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5,** e8, doi: 10.1371/journal.pbio.0050008 (2007).
27. Lebre, S. Inferring Dynamic Genetic Networks with Low Order Independencies. *Statistical Applications in Genetics and Molecular Biology* **8,** doi: 10.2202/1544-6115.1294 (2009).
28. Cantone, I. *et al.* A Yeast Synthetic Network for *In Vivo* Assessment of Reverse-Engineering and Modeling Approaches. *Cell* **137,** 172–181, doi: 10.1016/j.cell.2009.01.055 (2009).
29. Gardner, T. S., Di Bernardo, D., Lorenz, D. & Collins, J. J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301,** 102–105 (2003).

30. Mueller, E. *et al.* Terminal differentiation of human breast cancer through PPARγ. *Molecular cell* **1,** 465–470 (1998).
31. Mina, M. *et al.* Promoter-level expression clustering identifies time development of transcriptional regulatory cascades initiated by ErbB receptors in breast cancer cells. *Scientific Reports* **5,** doi: 10.1038/srep11999 (2015).
32. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489,** 91–100, doi: 10.1038/nature11245 (2012).
33. Butler, A. J. & Parker, M. G. COUP-TF-II Homodimers are formed in preference to heterodimers with RXR-alpha or TR-beta in intact-cells. *Nucleic Acids Research* **23,** 4143–4150, doi: 10.1093/nar/23.20.4143 (1995).
34. Chen, F. E., Huang, D. B., Chen, Y. Q. & Ghosh, G. Crystal structure of p50/p65 heterodimer of transcription factor NF-kappa B bound to DNA. *Nature* **391,** 410–413 (1998).
35. Delgoffe, G. M. & Vignali, D. A. A. STAT heterodimers in immunity: A mixed message or a unique signal? *Jak-Stat* **2,** e23060–e23060, doi: 10.4161/jkst.23060 (2013).
36. Garvie, C. W., Hagman, J. & Wolberger, C. Structural studies of Ets-1/Pax5 complex formation on DNA. *Molecular Cell* **8,** 1267–1276, doi: 10.1016/s1097-2765(01)00410-5 (2001).
37. Glover, J. N. M. & Harrison, S. C. Crystal-structure of the heterodimeric bZIP transcription factor c-FOS-c-JUN bound to DNA. *Nature* **373,** 257–261, doi: 10.1038/373257a0 (1995).
38. Hai, T. W., Liu, F., Coukos, W. J. & Green, M. R. Transcription factor ATF CDNA clones - an extensive family of leucine zipper proteins able to selectively form DNA-binding heterodimers. *Genes & Development* **3,** 2083–2090, doi: 10.1101/gad.3.12b.2083 (1989).
39. Helin, K. *et al.* Heterodimerization of the transcription factors E2F-1 and DP-1 leads to cooperative transactivation. *Genes & Development* **7,** 1850–1861, doi: 10.1101/gad.7.10.1850 (1993).
40. Malnou, C. E. *et al.* Heterodimerization with Different Jun Proteins Controls c-Fos Intranuclear Dynamics and Distribution. *Journal of Biological Chemistry* **285,** 6552–6562, doi: 10.1074/jbc.M109.032680 (2010).
41. Mangelsdorf, D. J. & Evans, R. M. The RXR heterodimers and orphan receptors. *Cell* **83,** 841–850, doi: 10.1016/0092-8674(95)90200-7 (1995).
42. Menet, J. S., Pescatore, S. & Rosbash, M. CLOCK: BMAL1 is a pioneer-like transcription factor. *Genes & Development* **28,** 8–13, doi: 10.1101/gad.228536.113 (2014).
43. Orlov, I., Rochel, N., Moras, D. & Klaholz, B. P. Structure of the full human RXR/VDR nuclear receptor heterodimer complex with its DR3 target DNA. *Embo Journal* **31,** 291–300, doi: 10.1038/emboj.2011.445 (2012).
44. Pufall, M. A. & Graves, B. J. Autoinhibitory domains: Modular effectors of cellular regulation. *Annual Review of Cell and Developmental Biology* **18,** 421–462, doi: 10.1146/annurev.cellbio.18.031502.133614 (2002).
45. Shrivastava, T., Mino, K., Babayeva, N. D., Baranovskaya, O. I. & Tahirov, T. H. Structural basis of Ets1 activation by Runx1. *Leukemia* **28,** 2040–2048, doi: 10.1038/leu.2014.111 (2014).
46. Westin, S. *et al.* Interactions controlling the assembly of nuclear-receptor heterodimers and co-activators. *Nature* **395,** 199–202 (1998).
47. Wu, Y. & Zhou, B. P. Snail: more than EMT. *Cell Adh Migr* **4,** doi: 10.4161/cam.4.2.10943 (2010).
48. Zheng, N., Fraenkel, E., Pabo, C. O. & Pavletich, N. P. Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. *Genes & Development* **13,** 666–674, doi: 10.1101/gad.13.6.666 (1999).
49. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nature Methods* 3, 211–222, doi: 10.1038/nmeth0306-211 (2006).
50. Hedrick, E., Lee, S.-O., Doddapaneni, R., Singh, M. & Safe, S. Nuclear receptor 4A1 as a drug target for breast cancer chemotherapy. *Endocrine-related cancer* **22,** 831–840 (2015).
51. Hugo, H. J. *et al.* Direct repression of MYB by ZEB1 suppresses proliferation and epithelial gene expression during epithelial-to-mesenchymal transition of breast cancer cells. *Breast Cancer Research* **15,** 1–19, doi: 10.1186/bcr3580 (2013).
52. Li, Y.-W. *et al.* Decreased Expression of GATA2 Promoted Proliferation, Migration and Invasion of HepG2 *In Vitro* and Correlated with Poor Prognosis of Hepatocellular Carcinoma. *PLoS ONE* **9,** e87505, doi: 10.1371/journal.pone.0087505 (2014).
53. Meyer, K. B. & Carroll, J. S. FOXA1 and breast cancer risk. *Nat Genet* **44,** 1176–1177 (2012).
54. Tiwari, A., Swamy, S., Gopinath, K. S. & Kumar, A. Genomic amplification upregulates estrogen-related receptor alpha and its depletion inhibits oral squamous cell carcinoma tumors *in vivo*. *Scientific Reports* **5,** 17621, doi: 10.1038/srep17621 (2015).
55. Xu, H.-X. *et al.* Expression profile of SIX family members correlates with clinic-pathological features and prognosis of breast cancer: A systematic review and meta-analysis. *Medicine* **95,** e4085, doi: 10.1097/md.0000000000004085 (2016).
56. Szasz, A. M. *et al.* Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1065 patients. *Oncotarget* **7,** 49322–49333, doi: 10.18632/oncotarget.10337 (2016).
57. Liu, X. *et al.* A systematic study on drug-response associated genes using baseline gene expressions of the Cancer Cell Line Encyclopedia. *Scientific reports* **6,** doi: 10.1038/srep22811 (2016).
58. Leary, A. *et al.* Anti-proliferative effect of lapatinib in HER2-positive and HER2-negative/HER3-high breast cancer: results of the pre-surgical randomized MAPLE trial (CRUK E/06/039). *American Association for Cancer Research* **21,** 2932–2940, doi: 10.1158/1078-0432.ccr-14-1428 (2014).
59. Zhou, Y. *et al.* MEK inhibitor effective against proliferation in breast cancer cell. *Tumor Biology* **35,** 9269–9279, doi: 10.1007/s13277-014-1901-5 (2014).
60. Finn, R. S. *et al.* PD 0332991, a selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines *in vitro*. *Breast Cancer Research*: *BCR* **11,** R77–R77, doi: 10.1186/bcr2419 (2009).
61. Rudolph, M. C. *et al.* Sterol regulatory element binding protein and dietary lipid regulation of fatty acid synthesis in the mammary epithelium. *American Journal of Physiology-Endocrinology and Metabolism* **299,** E918–E927, doi: 10.1152/ajpendo.00376.2010 (2010).
62. McFadden, J. W. & Corl, B. A. Activation of liver X receptor (LXR) enhances de novo fatty acid synthesis in bovine mammary epithelial cells. *Journal of Dairy Science* **93,** 4651–4658, doi: 10.3168/jds.2010-3202 (2010).
63. Meinhardt, G. & Hass, R. Differential expression of c-myc, max and mxi1 in human myeloid leukemia cells during retrodifferentiation and cell death. *Leukemia Research* **19,** 699–705 (1995).
64. Chung, J. H. *et al.* Deferoxamine promotes osteoblastic differentiation in human periodontal ligament cells via the nuclear factor erythroid 2-related factor-mediated antioxidant signaling pathway. *Journal of Periodontal Research* **49,** 563–573, doi: 10.1111/jre.12136 (2014).
65. de la Fuente, A. G. *et al.* Vitamin D receptor–retinoid X receptor heterodimer signaling regulates oligodendrocyte progenitor cell differentiation. *The Journal of Cell Biology* **211,** 975–985, doi: 10.1083/jcb.201505119 (2015).
66. Nitzsche, A. *et al.* RAD21 Cooperates with Pluripotency Transcription Factors in the Maintenance of Embryonic Stem Cell Identity. *PLoS ONE* **6,** e19470, doi: 10.1371/journal.pone.0019470 (2011).
67. Liu, X. *et al.* Expression of SATB1 and HER2 in breast cancer and the correlations with clinicopathologic characteristics. *Diagnostic Pathology* **10,** 50, doi: 10.1186/s13000-015-0282-4 (2015).
68. Brouckaert, O., Wildiers, H., Floris, G. & Neven, P. Update on triple-negative breast cancer: prognosis and management strategies. *International Journal of Women's Health* **4,** 511–520, doi: 10.2147/IJWH.S18541 (2012).
69. Brandes, U. A faster algorithm for betweenness centrality. *Journal of mathematical sociology* **25,** 163–177 (2001).
70. Kwok, S. *et al.* Transforming growth factor-β1 regulation of ATF-3 and identification of ATF-3 target genes in breast cancer cells. *Journal of cellular biochemistry* **108,** 408–414 (2009).

71. Peng, F. *et al.* Direct targeting of SUZ12/ROCK2 by miR-200b/c inhibits cholangiocarcinoma tumourigenesis and metastasis. *Br J Cancer* **109,** 3092–3104, doi: 10.1038/bjc.2013.655 (2013).
72. Ambele, M. A., Dessels, C., Durandt, C. & Pepper, M. S. Genome-wide analysis of gene expression during adipogenesis in human adipose-derived stromal cells reveals novel patterns of gene expression during adipocyte differentiation. *Stem Cell Research* **16,** 725–734 (2016).
73. Coradini, D., Boracchi, P., Oriana, S., Biganzoli, E. & Ambrogi, F. Differential expression of genes involved in the epigenetic regulation of cell identity in normal human mammary cell commitment and differentiation. *Chinese Journal of Cancer* **33,** 501–510, doi: 10.5732/cjc.014.10066 (2014).
74. Langer, S. *et al.* Jun and Fos family protein expression in human breast cancer: correlation of protein expression and clinicopathological parameters. *European journal of gynaecological oncology* **27,** 345–352 (2005).
75. Brin, S. & Page, L. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks* **56,** 3825–3833 (2012).
76. Chouvardas, P., Kollias, G. & Nikolaou, C. Inferring active regulatory networks from gene expression data using a combination of prior knowledge and enrichment analysis. *BMC Bioinformatics* **17,** 319–332, doi: 10.1186/s12859-016-1040-7 (2016).
77. Glass, K., Huttenhower, C., Quackenbush, J. & Yuan, G.-C. Passing Messages between Biological Networks to Refine Predicted Interactions. *PLoS One* **8,** e64832, doi: 10.1371/journal.pone.0064832 (2013).
78. Guan, D. *et al.* PTHGRN: unraveling post-translational hierarchical gene regulatory networks using PPI, ChIP-seq and gene expression data. *Nucleic Acids Research* **42,** W130–W136, doi: 10.1093/nar/gku471 (2014).
79. Wang, J. *et al.* APG: an Active Protein-Gene Network Model to Quantify Regulatory Signals in Complex Biological Systems. *Scientific Reports* **3,** 1097, doi: 10.1038/srep01097 (2013).
80. Guan, D. *et al.* CMGRN: a web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data. *Bioinformatics* **30,** 1190–1192, doi: 10.1093/bioinformatics/btt761 (2014).
81. Wang, E. *et al.* Predictive genomics: A cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Seminars in Cancer Biology* **30,** 4–12, doi: http://dx.doi.org/10.1016/j.semcancer.2014.04.002 (2015).
82. Zaman, N. *et al.* Signaling Network Assessment of Mutations and Copy Number Variations Predict Breast Cancer Subtype-Specific Drug Targets. *Cell Reports* **5,** 216–223, doi: http://dx.doi.org/10.1016/j.celrep.2013.08.028 (2013).
83. Hamed, M., Spaniol, C., Zapp, A. & Helms, V. Integrative network-based approach identifies key genetic elements in breast invasive carcinoma. *BMC Genomics* **16,** S2, doi: 10.1186/1471-2164-16-s5-s2 (2015).
84. Noh, H. & Gunawan, R. Inferring gene targets of drugs and chemical compounds from gene expression profiles. *Bioinformatics* **32,** 2120–2127, doi: 10.1093/bioinformatics/btw148 (2016).
85. Whitmarsh, A. J. & Davis, R. J. Regulation of transcription factor function by phosphorylation. *Cellular and Molecular Life Sciences* **57,** 1172–1183, doi: 10.1007/pl00000757 (2000).

Acknowledgements

This project was funded by the Irish Cancer Society CCRC BREAST-PREDICT [grant number CCRC13GAL].

Author Contributions

L.F.I.M. performed the analysis and wrote the manuscript. W.K. designed the study and wrote the manuscript. T.S. designed the study, performed the analysis and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Iglesias-Martinez, L. F. *et al.* BGRMI: A method for inferring gene regulatory networks from time-course gene expression data and its application in breast cancer research. *Sci. Rep.* **6**, 37140; doi: 10.1038/srep37140 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2016