

RESEARCH ARTICLE

Open Access



# Estimating genome-wide regulatory activity from multi-omics data sets using mathematical optimization

Saskia Trescher<sup>\*</sup> , Jannes Münchmeyer and Ulf Leser

## Abstract

**Background:** Gene regulation is one of the most important cellular processes, indispensable for the adaptability of organisms and closely interlinked with several classes of pathogenesis and their progression. Elucidation of regulatory mechanisms can be approached by a multitude of experimental methods, yet integration of the resulting heterogeneous, large, and noisy data sets into comprehensive and tissue or disease-specific cellular models requires rigorous computational methods. Recently, several algorithms have been proposed which model genome-wide gene regulation as sets of (linear) equations over the activity and relationships of transcription factors, genes and other factors. Subsequent optimization finds those parameters that minimize the divergence of predicted and measured expression intensities. In various settings, these methods produced promising results in terms of estimating transcription factor activity and identifying key biomarkers for specific phenotypes. However, despite their common root in mathematical optimization, they vastly differ in the types of experimental data being integrated, the background knowledge necessary for their application, the granularity of their regulatory model, the concrete paradigm used for solving the optimization problem and the data sets used for evaluation.

**Results:** Here, we review five recent methods of this class in detail and compare them with respect to several key properties. Furthermore, we quantitatively compare the results of four of the presented methods based on publicly available data sets.

**Conclusions:** The results show that all methods seem to find biologically relevant information. However, we also observe that the mutual result overlaps are very low, which contradicts biological intuition. Our aim is to raise further awareness of the power of these methods, yet also to identify common shortcomings and necessary extensions enabling focused research on the critical points.

**Keywords:** Gene regulation, Regulatory network, Systems biology, Mathematical optimization

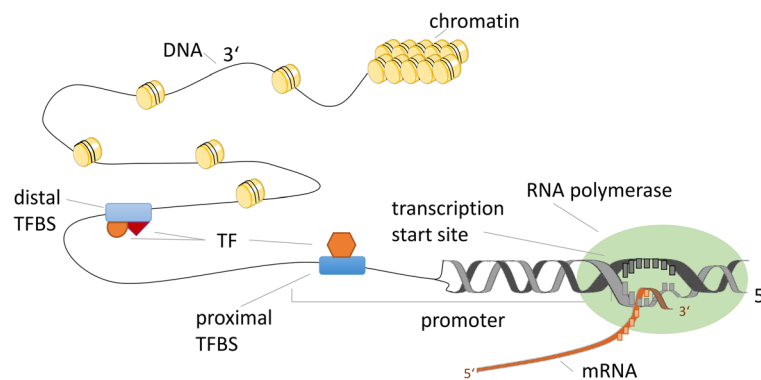
## Background

Gene regulation is one of the most important biological processes in living cells. It is indispensable for adapting to changing environments, stimuli, and developmental stage and plays an essential role in the pathogenesis and course of diseases. Mechanistically, the transcription of DNA into RNA is predominantly controlled by a complex network of transcription factors (TFs) (see Fig. 1). These proteins bind to enhancer or promoter regions adjacent to the genes they regulate [1], which may

enhance or inhibit the recruitment of RNA polymerase and thereby activate or repress gene transcription [2]. Gene products also can be modified post-translationally via microRNAs (miRNAs) degrading the transcript or inhibiting their translation [3]. Besides, a multitude of other mechanisms influence gene regulation, such as chromatin remodelling [4], epigenetic effects [5], and compound-building of transcription factors [2]. Distortion of regulatory processes is inflicted with various diseases [6, 7], especially with cancer [8, 9].

Due to this importance, many efforts have been devoted to the elucidation of human regulatory relationships and networks. Wide-spread experimental techniques are transcriptome measurements to quantify gene

<sup>\*</sup> Correspondence: [saskia.trescher@informatik.hu-berlin.de](mailto:saskia.trescher@informatik.hu-berlin.de)  
Knowledge Management in Bioinformatics, Computer Science Department,  
Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany



**Fig. 1** Transcription of DNA into RNA. Transcription factors (TFs) bind to distal or proximal TF binding sites (TFBS) enhancing the binding of RNA polymerase and activating the transcription of DNA into RNA

and transcription factor co-expression [10], chromatin immunoprecipitation (ChIP) on chips or followed by sequencing for identifying binding patterns of specific TFs [11], and bisulfite sequencing to find epigenetic signals of regulation [12]. Many large-scale datasets of such experiments have been published and are available in public repositories such as the Gene Expression Omnibus (GEO) [13], the Cancer Genome Atlas (TCGA) [14] or the Encyclopedia of DNA Elements (ENCODE) [15]. Computational methods are also used, for instance, to identify transcription factor binding sites (TFBS) [16] or to find known TFBS within the genome (e.g., [17, 18]). Several databases have been created which store relevant information, such as lists of binding motifs (TRANSFAC [19] or JASPAR [20]) or targets of regulatory miRNAs [21].

Such measurements and predictions are used by network reconstruction algorithms to predict regulatory relationships and regulatory networks [22]. A plethora of different methods have been proposed, ranging from purely qualitative methods [23] over simple statistical approaches [24] to more advanced probabilistic frameworks [25]. Early methods were plagued by insufficient data and a general scarcity of background knowledge, which led to rather unstable results [26]. This situation has changed dramatically over the last years, as results of more and more large screens have been made publicly available [27] and also the knowledge on principal regulatory relationships has increased [28, 29]. This, in turn, has increased the interest in methods which predict genome-wide networks using a systematic, unified, mathematical framework.

Here, we review five rather recent methods and conduct a quantitative comparison of their results with the goal to identify their mutual strengths and weaknesses. They all have in common that they assume both the set of regulators (transcription factors or micro RNAs) to be known and the topology of the regulatory network to be given. By combining this background knowledge with

specific omics data sets, especially transcriptome data, they try to infer the activity of regulators in a certain experimental condition or disease using mathematical optimization. All presented methods are global methods in the sense that they compute activities genome-wide (as much as represented by the underlying network), thus removing the shortcomings of local methods which ignore cross-talk between sub-models and global effects within samples. The methods predominantly produce a ranked list of regulators, sorted by their activity in a given group of samples; given that a multitude of biological influences is ignored during inference, especially kinetic and temporal effects, their goal cannot be to produce absolute snapshots of regulatory activity. We describe each method in detail and compare them with respect to the most important properties, such as the data being used, the method applied for deriving optimized activity values, or the evaluation performed to show effectiveness. We further implemented a quantitative comparison including four of the presented methods to objectively analyze their results. As contrast, we also include ARACNE [30] as sixth method; this algorithm uses only local reasoning and requires no background knowledge, but is still rather popular.

## Methods

We describe in detail five methods which infer transcription factor activity from omics data sets using a background network of transcription factors and the genes they regulate. All use some form of mathematical optimization. To emphasize the common ground of these at-first-sight rather different methods, we explain their underlying models using a simple framework for defining the relationships of transcription factors and genes. This framework is presented first; it should be understood as a least common denominator, not as a proper method for network inference by itself. We then describe five recently published methods for genome-

wide TF activity estimation as extensions or constraints to this general framework, namely the approach by Schacht et al. [31] (estimation of TF activity by the effect on their target genes), RACER [32], RABIT [33], ISMARA [34] and biRte [35]. Additionally, we contrast these more comprehensive methods with the local inference algorithm ARACNE [30], a popular tool for the de-novo reconstruction of gene regulatory networks. Key properties of all methods (input, mathematical model, computation, output) are summarized in Table 1.

### Mathematical framework

To combine regulatory networks and quantitative omics data and to thereby deduce regulatory activity, all methods described here use a genome-wide mathematical model. Sample specific gene expression values  $g_{i,s}$ , derived from one biological condition, i.e., grouped into a single class, for in total  $G$  genes and  $S$  samples need to be provided as input. The background regulatory network is represented as a directed graph where the nodes designate regulators and regulated entities (mostly TFs and genes, but also miRNAs, regulatory sites, or TF complexes) and directed edges indicate a regulatory relationship between the two connected nodes, for example the influence of a TF on the expression of a gene (see Fig. 2).

We will use the variable  $t$  for regulators,  $i$  for regulated entities, and  $b_{t,i}$  for the strength of an edge from a TF/miRNA  $t$  to a gene  $i$  representing, for instance, a binding affinity. As abstract framework for explaining the different methods we propose a simple linear model predicting gene expression  $\widehat{g}_{i,s}$  of gene  $i$  in sample  $s$  in terms of the activity of all  $T$  transcription factors  $\beta_{t,s}$ , which regulate  $i$ , and the binding affinities  $b_{t,i}$ . In contrast to Fig. 2, where TFs can influence each other, this model ignores TF – TF relations and feedback loops:

$$\widehat{g}_{i,s} = \sum_{t=1}^T \beta_{t,s} b_{t,i}$$

Given this model and a set of quantitative measurements of gene expression  $g_{i,s}$ , the goal of the mathematical optimization is to find parameters  $\beta$  such that the sum of squared errors of measured vs predicted gene expression over all genes and samples is minimized using a certain norm, for example the  $L_2$  norm:

$$\min \sum_{i=1}^G \sum_{s=1}^S (g_{i,s} - \widehat{g}_{i,s})^2$$

### Estimation of TF activity by the effect on their target genes [31]

The idea of this method is to use the expression levels of TF's target genes to infer their integrated effect (see

Fig. 3). The method uses expression data and database curated TF binding information as input whereby the TF – gene network is restricted to genes regulated by more than 10 TFs and TFs with at least 5 target genes. The model is closely related to the abovementioned general framework, only adding a term for the sample specific effect of a TF. Specifically, the activity of a TF is modelled linearly by its cumulative effect on its target genes normalized by the sum of target genes or the TF's gene expression level:

$$\widehat{g}_{i,s} = c + \sum_t \beta_t b_{t,i} (\theta_{a,t} act_{t,s} + \theta_{g,t} g_{t,s})$$

where  $\widehat{g}_{i,s}$  denotes the predicted gene expression of gene  $i$  in sample  $s$ ,  $c$  is an additive offset,  $\beta_t$  describes the estimated activity of TF  $t$  and  $b_{t,i}$  refers to the underlying strength of the relation between TF  $t$  and gene  $i$  reflecting the binding affinity. The estimated effect of a TF in a certain sample is calculated via the switch-like term in parentheses, where either the ac-

tivity definition  $act_{t,s} = \frac{\sum_i b_{t,i} g_{i,s}}{\sum_i b_{t,i}}$  or the gene expression of the TF itself  $g_{t,s}$  is taken into account using the restrictions  $\theta_{a,t}, \theta_{g,t} \in \{0, 1\}$  and  $\theta_{a,t} + \theta_{g,t} = 1$ . This switch term represents a meta-parameter to find the best model and has no biological interpretation. The model outputs an activity value and the information which switch parameter is chosen for each TF of the reduced network.

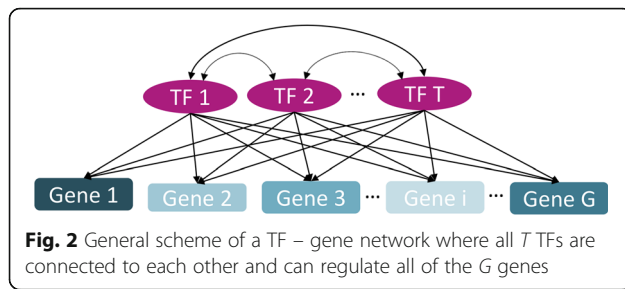
During the optimization, the sum of error terms (absolute value of the difference between predicted and measured gene expression) is minimized which is achieved via mixed-integer linear programming using the Gurobi 5.5 optimizer.<sup>1</sup> The authors of this method state that the activity definition (see above) was used in 95% of their test cases, but the switch-like combination of both terms yielded still better optimization results. In the paper, the optimization task is greatly simplified as the model is computed for each gene separately and allows only a maximum number of 6 regulating TFs. The TF – gene network indicating the strength of a relation between a TF and a gene is created for 1120 TFs using knowledge from the commercial MetaCore™ database,<sup>2</sup> ChEA [36] and ENCODE [15]. Due to the restriction of the network mentioned above, the actual model is then based on 521 TFs and 636 target genes only.

Evaluation of the results was performed using expression data from 59 cell lines of the NCI-60 panel [37, 38] and from melanoma cell lines (“Mannheim cohort”) [39]. A sample based leave-one-out and 10-fold cross validation of predicted and measured gene expression yielded Pearson correlation scores of about 0.6 for both

**Table 1** Overview of methods for estimating regulatory activity from transcriptome data comparing input data, modelling, computational aspects and outcome variables

Method	Input	Model	Computation	Output
Approach by Schacht et al.	<ul style="list-style-type: none"> <li>- mRNA expression data</li> <li>- TF binding information</li> </ul>	<p>Linear model</p> $\widehat{g}_{i,s} = c + \sum_t \beta_{i,t} b_{i,t} (\theta_{a,t} act_{t,s} + \theta_{g,t} g_{t,s})$ <p>with</p> $act_{t,s} = \sum_{b_{i,t}} \beta_{i,t} b_{i,t}, \theta_{a,t} + \theta_{g,t} = 1, \theta_{a,t}, \theta_{g,t} \in \{0, 1\}$	<ul style="list-style-type: none"> <li>- Optimization criterion: minimize sum of absolute errors</li> <li>- Mixed-integer linear programming</li> <li>- Optimization via Gurobi 5.5</li> </ul>	<ul style="list-style-type: none"> <li>- parameter for each TF: <math>\beta_i</math></li> <li>- decision for each TF if <math>\theta_{a,t}</math> or <math>\theta_{g,t}</math> was chosen</li> </ul>
RACER	<ul style="list-style-type: none"> <li>- mRNA expression data</li> <li>- copy number variation</li> <li>- DNA methylation</li> <li>- miRNA expression signals</li> <li>- TF binding information</li> <li>- miRNA target site info (c)</li> </ul>	<p>Linear models:</p> <p>1) <math>\widehat{g}_{i,s} = c + \theta_{CNV,s} CNV_{i,s} + \theta_{DM,s} DM_{i,s} + \sum_t \beta_{i,t,s} b_{i,t}</math>  <math>+ \sum_m \beta_{mi,s} c_{i,mi} miRNA_{mi,s}</math></p> <p>2) <math>\widehat{g}_{i,s} = \tilde{c} + \tilde{\theta}_{i,CNV} CNV_{i,s} + \tilde{\theta}_{i,DM} DM_{i,s} + \sum_t Y_{i,t} \beta_{i,t,s}</math>  <math>+ \sum_m Y_{i,mi} \beta_{mi,s}</math></p>	<ul style="list-style-type: none"> <li>- Optimization criterion: minimize sum of squared errors with <math>L_1</math> norm penalty on linear coefficients</li> <li>- Elastic-net regularized generalized linear models and LASSO</li> </ul>	<ol style="list-style-type: none"> <li>1) sample-specific TF and miRNA activities <math>\beta_{i,s}</math> and <math>\beta_{mi,s}</math></li> <li>2) TF-gene <math>Y_{i,t}</math> and miRNA-gene <math>Y_{i,mi}</math> interactions across all samples</li> </ol>
RABIT	<ul style="list-style-type: none"> <li>- differential mRNA expression data</li> <li>- somatic mutations</li> <li>- DNA methylation</li> <li>- copy number variation</li> <li>- TF binding info</li> <li>- recognition motifs for RNA-binding protein (RBP)</li> </ul>	<p>Linear model:</p> $\widehat{g}_i = \sum_t \theta_i b_{i,t} + \sum_t \beta_{i,t}$ <p>With <math>B_i</math>: background factors (gene CNA, promoter DNA methylation, promoter degree promoter CpG content)</p>	<ul style="list-style-type: none"> <li>- Frisch-Waugh-Lovell method, select subset of significant TFs via model selection procedure and remove TFs with insignificant correlation across tumors</li> </ul>	<ul style="list-style-type: none"> <li>- regulatory activity score for each TF (t value of linear regression coefficient of t-test)</li> </ul>
ISMARA	<ul style="list-style-type: none"> <li>- gene expression or chromatin state measurements</li> <li>- annotation of promoters (number of predicted sites for motifs)</li> <li>- transcripts and associated promoters</li> <li>- miRNA target site predictions</li> </ul>	<p>Linear model</p> $\widehat{g}_{p,s} = c_p + c_s + \sum_m N_{p,m} \beta_{m,s}$	<ul style="list-style-type: none"> <li>- Optimization criterion: minimize sum of errors</li> <li>- Bayesian procedure, ridge regression</li> <li>- Gaussian prior for <math>\beta_{m,s}</math> to avoid overfitting</li> </ul>	<ul style="list-style-type: none"> <li>- inferred motif activity profiles <math>\beta_{m,s}</math> with set of TFs and miRNAs binding to sites of these motifs (= key regulators)</li> <li>- predicted target promoters, associated transcripts and genes</li> <li>- Network of known interactions between predicted targets and predicted regulatory interactions</li> <li>- enriched ontology categories</li> </ul>
biRte	<ul style="list-style-type: none"> <li>- mRNA differential expression</li> <li>- miRNA, TF measurements, CNV (optionally)</li> <li>- regulator (R) – target network</li> </ul>	<p>Likelihood model:</p> $L_{D,R}(R) = p(D R, \theta) = \prod_b p(\hat{D}_b   R, \theta) = \prod_b \prod_c \prod_i p(\hat{D}_{b,c}   R_c, \theta)$	<ul style="list-style-type: none"> <li>- data specific marginal likelihoods using estimation of hidden state variables with via MCMC</li> <li>- Nested effects model structure Learning to reconstruct transcriptional network</li> </ul>	<ul style="list-style-type: none"> <li>- Estimation of active regulators</li> <li>- Estimation of associated transcriptional network</li> </ul>
ARACNE	<ul style="list-style-type: none"> <li>- microarray expression profiles</li> </ul>	none	<ul style="list-style-type: none"> <li>- local estimation of pairwise gene expression profile mutual information</li> </ul>	<ul style="list-style-type: none"> <li>- Reconstruction of gene regulatory network</li> </ul>

Gene expression data is named "g" with index i, estimated parameters with "β", TF binding information with "b", TFs with "a", samples with "s", miRNAs with "mi" and model constants with "c". Other variables are explained in the text



**Fig. 2** General scheme of a TF – gene network where all  $T$  TFs are connected to each other and can regulate all of the  $G$  genes

data sets. A gene set enrichment analysis of the target genes for TFs modelled by the activity definition yielded 64 significantly enriched concepts including cell cycle, immune response and cell growth for the data from the NCI-60 panel. Additionally, a  $t$ -test was computed between melanoma and other cell lines of the NCI-60 panel to find differentially expressed genes of melanogenesis. For the resulting genes, regulation models were built and used to predict gene expression in the melanoma cell line data set yielding good prediction performances.

#### RACER [32]

RACER (Regression Analysis of Combined Expression Regulation) aims to integrate generic cell-line data with sample-specific measurements using a two-stage regression (see Fig. 4). Firstly, sample-specific regulatory activities for TFs and miRNAs are calculated. Subsequently, general TF/miRNA – gene interactions are derived.

Compared to our general framework, RACER includes additionally miRNA binding information. It assumes a linear combination, which is not further justified, of the regulatory effects of TFs and miRNAs on mRNA level. RACER can incorporate a variety of sample specific data including mRNA and miRNA expression values, CNV and DNA methylation. Optimization is applied twice to reduce model complexity, where the method first infers sample-specific TF and miRNA activities and uses these, in a second step, to compute general TF/miRNA – gene interactions.

In the first regression step, mRNA, miRNA, CNV and DNA methylation data are used to calculate the sample specific activities:

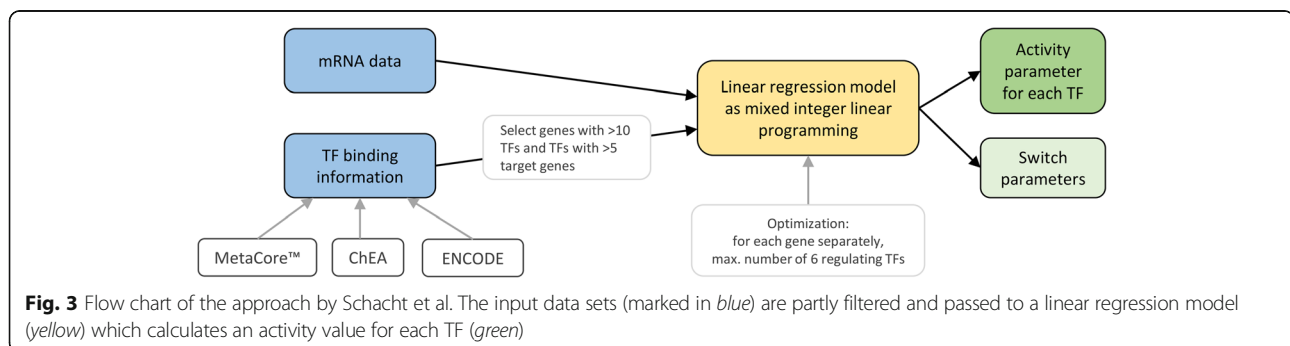
$$\widehat{g}_{i,s} = c + \theta_{CNV,s} CNV_{i,s} + \theta_{DM,s} DM_{i,s} + \sum_t \beta_{t,s} b_{t,i} + \sum_{mi} \beta_{mi,s} c_{i,mi} miRNA_{mi,s}$$

where  $\widehat{g}_{i,s}$  denotes the predicted gene expression of gene  $i$  in sample  $s$ ,  $c$  is an intercept,  $\beta_{t,s}$  describes the estimated activity of TF  $t$  in sample  $s$  and  $b_{t,i}$  is the TF – gene binding score for TF  $t$  and gene  $i$ . The parameter  $\beta_{mi,s}$  stands for the estimated activity of miRNA  $mi$  in sample  $s$  and is multiplied by  $c_{i,mi}$ , the number of conserved target sites on 3'UTR of the target gene  $i$  for miRNA  $mi$ , and by the expression level of miRNA  $mi$  in sample  $s$ .  $\theta_{CNV,s}$  (respectively  $\theta_{DM,s}$ ) are the regression parameters for CNV signals  $CNV_{i,s}$  (respectively DNA methylation data  $DM_{i,s}$ ). Using  $\beta_{t,s}$  and  $\beta_{mi,s}$  from the first regression step, TF – gene and miRNA – gene interactions across all samples are calculated in a second model:

$$\widehat{g}_{i,s} = \tilde{c} + \tilde{\theta}_{i,CNV} CNV_{i,s} + \tilde{\theta}_{i,DM} DM_{i,s} + \sum_t \gamma_{i,t} \beta_{t,s} + \sum_{mi} \gamma_{i,mi} \beta_{mi,s}$$

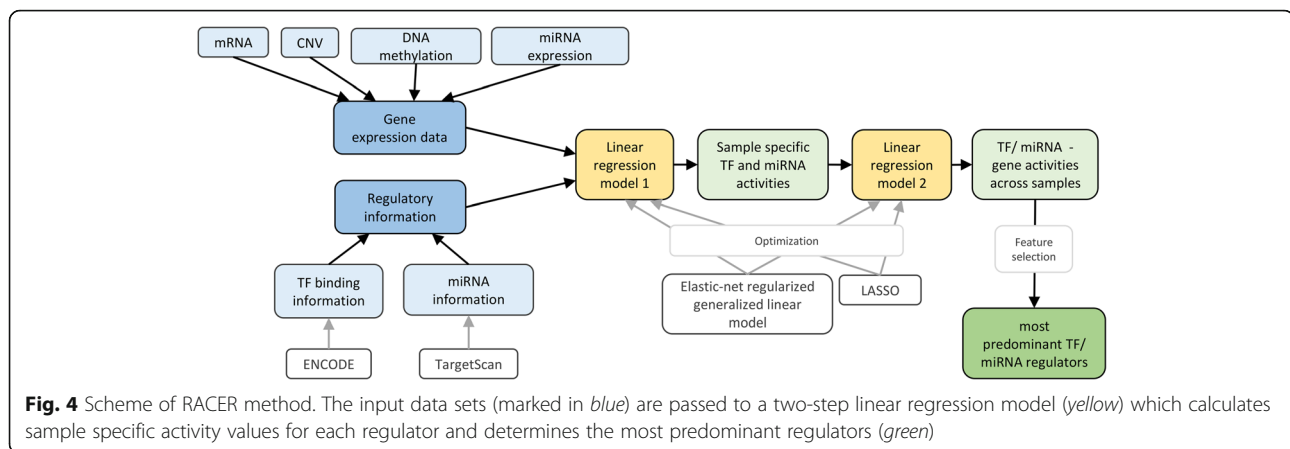
where the sums apply only to a number of selected TFs and miRNAs with nonzero binding signals  $b_{t,i} > 0$  and conserved target sites  $c_{i,mi} > 0$ . The resulting parameters  $\gamma_{i,t}$  and  $\gamma_{i,mi}$  indicate the strength of a TF/miRNA – gene relationship across all samples. To obtain robust estimates,  $\gamma_{i,mi}$  is additionally weighted by the averaged activities of the miRNA.

In each of the two regression steps, the optimization criterion is to minimize the sum of squared errors with  $L_1$  penalty on the linear coefficients to induce a sparse solution and to set irrelevant parameters to zero after the fitting. This sparse LASSO solution is obtained through elastic-net regularized generalized linear models. A supplementary feature selection procedure comparing the full model to a restricted model leaving one TF or miRNA out provides the most predominant TF/miRNA regulators. TF binding scores are collected from the generic cell line of erythroleukemia cells K562 from ENCODE



**Fig. 3** Flow chart of the approach by Schacht et al. The input data sets (marked in blue) are partly filtered and passed to a linear regression model (yellow) which calculates an activity value for each TF (green)





**Fig. 4** Scheme of RACER method. The input data sets (marked in blue) are passed to a two-step linear regression model (yellow) which calculates sample specific activity values for each regulator and determines the most predominant regulators (green)

for 97 TFs and 16653 genes. Further, the number of conserved target sites on 3'UTR is taken from sequence-based information from TargetScan for 470 miRNAs and 16653 genes. The RACER method is implemented in R and publicly available under <http://www.cs.utoronto.ca/~yueli/racer.html>.

The method was evaluated using expression data from an acute myeloid leukemia (AML) data set from TCGA with 173 samples [40] via a sample based 10-fold cross validation on the prediction of gene expression. To assess the quality of predictions, the Spearman rank correlation was calculated resulting in a reassuring value of approximately 0.6. Further, the full model was compared to models excluding one type of the input variables. The full model performed best and a substantial reduction of Spearman correlation was observed by omitting TF regulation (20%) and DNA methylation (5%). RACER also performed with competitive accuracy in predicting known miRNA – mRNA and TF – gene relationships compared to other methods like GenMiR++ [41] or ENCODE TF binding scores [15] using e.g., validated interactions from the MirTarBase [42] and knockdown studies. The feature selection procedure revealed 18 predominant transcriptional regulators in the AML dataset. Using their associated targets, a functional enrichment analysis showed that DNA repair and the tumor necrosis factor pathway were enriched. When applying this panel to cluster patients at different cytogenetic risks, the clustering pattern of the regulatory activities was largely consistent with the risk groups. Further, a literature survey on AML showed that many TF regulators among the top predictions had a role in leukemogenesis.

#### RABIT [33]

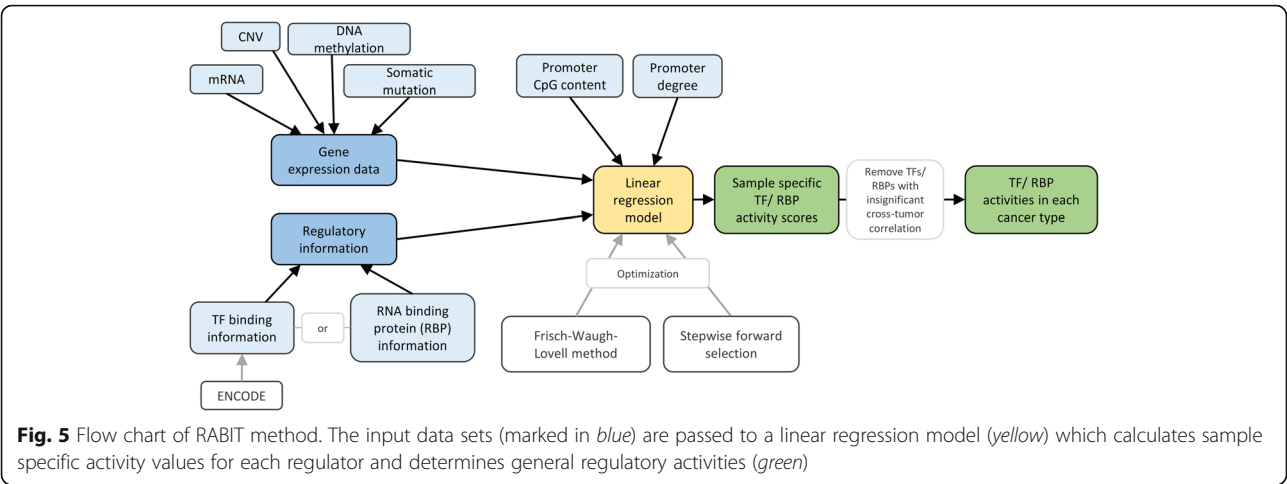
Regression Analysis with Background Integration (RABIT) is a method for finding expression regulators in cancer by a large scale analysis across diverse cancer types. It integrates TF binding information with tumor profiling data to search for TFs driving tumor-specific

gene expression patterns (see Fig. 5). It can be applied to predict cancer-associated RNA-binding protein (RBP) recognition motifs which are key components in the determination of miRNA function [43].

In contrast to our general framework, RABIT can, like RACER, make use of CNV and DNA methylation data additionally integrating promoter CpG content and promoter degree information (total number of ChIP-seq peaks near the gene transcription start site) and takes RBP or TF binding information as regulatory input. The computational model consists of three steps (see Fig. 5). First, RABIT tests in each tumor whether the target genes, identified by the BETA method [44], show differential expression compared to the normal controls including a control for background effects from CNVs, promoter DNA methylation, promoter CpG content and promoter degree:

$$\hat{g}_i = \sum_f \theta_f B_{f,i} + \sum_t \beta_t b_{t,i}$$

where  $\hat{g}_i$  represents the predicted differential gene expression between tumor and normal samples in gene  $i$ ,  $B$  includes values of the  $f$  different background factors for gene  $i$ ,  $b$  contains RBP or TF binding information and  $\theta$  and  $\beta$  are the respective regression parameter vectors. The regression coefficients  $\beta$  are estimated by minimizing the squared difference between measured and predicted gene expression. The regulatory activity score for each TF/RBP is defined by a  $t$ -value (regression coefficient divided by standard error) and its significance by the corresponding  $t$ -test. If multiple profiles exist for the same TF from different conditions or cell lines, the profile with the highest absolute value of TF regulatory activity score is selected. In a second step, a stepwise forward selection is applied to find a subset of TFs among those screened in step one optimizing the model error. Lastly, TFs with insignificant cross-tumor correlation are removed from the results.



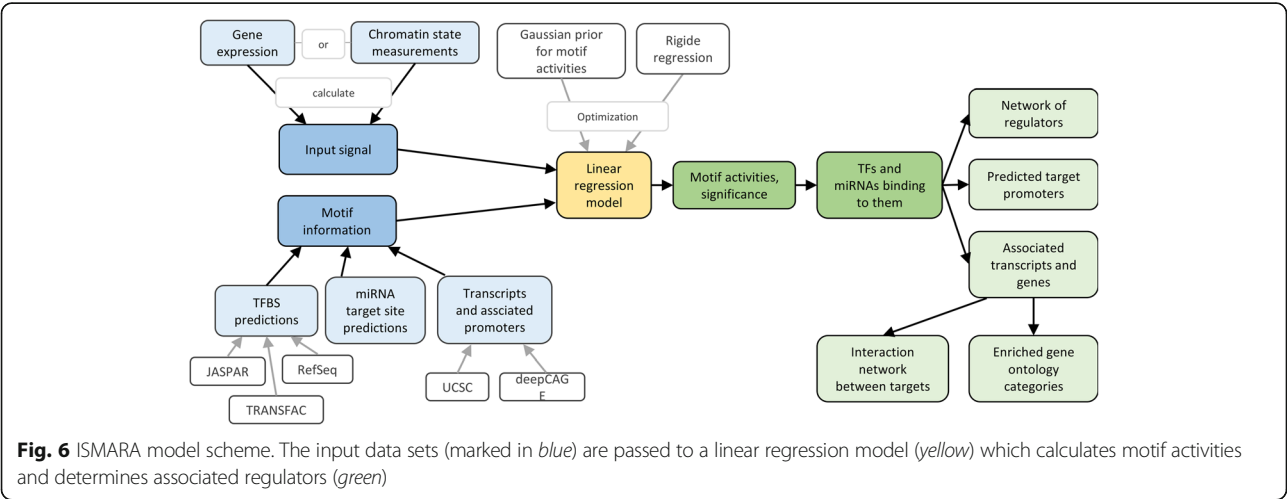
Computationally, the regression coefficients are calculated via the efficient Frisch-Waugh-Lovell method. TF binding information is taken from 686 TF ChIP-seq profiles from ENCODE representing 150 TFs and 90 cell types. Additionally, recognition motifs for 133 RBPs and their putative targets are collected by searching recognition motifs over the 3'UTR regions [45]. An implementation of the RABIT method can be downloaded from <http://rabbit.dfci.harvard.edu/download>.

RABIT was applied to 7484 tumor profiles of 18 cancer types from TCGA using gene expression, somatic mutation, CNV and DNA methylation data. To systematically assess the results, the cancer relevance level of a TF was calculated as percentage of tumors with the TF target genes differentially regulated (averaged across all TCGA cancer types). A comparison to cancer gene databases, i.e., the NCI cancer gene index project [46], the Bushman Laboratory cancer driver gene list [47, 48], the COSMIC somatic mutation catalog [49] and the CCGD mouse cancer driver genes [50], showed a consistent

picture. Further, RABIT's performance was compared to other regression models like LAR or LASSO where RABIT had the best classification results when classifying all TFs into three categories by NCI cancer index and achieved better cross-validation error and shorter running time. The regulatory activity of RBPs showed that some alternative splicing factors could affect tumor-specific gene expression by binding to target gene 3'UTR regions.

# ISMARA [34]

In contrast to the previous three methods and to our general framework which directly scores TFs or other regulators, ISMARA (Integrated System for Motif Activity Response Analysis) infers the activity of regulatory motifs (short nucleotide sequences) and thereby indirectly deduces the effects of TFs and miRNAs (see Fig. 6). ISMARA is a web service where no parameter settings or specific processing of the input data, gene expression or ChIP-seq data are necessary. It can also be used to



calculate regulatory activity differences between samples and consider replicates or data from time series.

ISMARA takes sample specific measurements and information about regulatory motifs for TFs and miRNAs into account. Based on the input of gene expression data or chromatin state measurements, the input signal is calculated for each promoter in each sample. The input signal levels are modelled linearly in terms of the binding site predictions and unknown motif activities:

$$\widehat{g}_{p,s} = c_p + c_s + \sum_m N_{p,m} \beta_{m,s}$$

where  $\widehat{g}_{p,s}$  refers to the input signal for a promoter  $p$  in sample  $s$ ,  $c_p$  and  $c_s$  are intercepts for each promoter and sample,  $N_{p,m}$  summarizes the TF/miRNA binding site predictions (sum of the posterior probabilities of all predicted TF/miRNA binding sites for motif  $m$  in promoter  $p$ ) and  $\beta_{m,s}$  stands for the estimated motif activities. Like in the other presented methods, the optimization criterion is to minimize the sum of squared error terms between predicted and measured gene expression. Primarily, ISMARA provides the inferred motif activity profiles ( $\beta_{m,s}$ ) sorted by significance and a set of TFs and miRNAs that bind to these motifs representing the key regulators. Further, a list containing their predicted target promoters, associated transcripts and genes, a network of known interaction between these targets and a list of enriched gene ontology categories is displayed. The web service ISMARA is available under <http://ismara.unibas.ch>.

ISMARA employs a Bayesian procedure with a Gaussian likelihood model and a Gaussian prior distribution for  $\beta_{m,s}$  to avoid overfitting. Information about regulatory motifs is provided via the annotation of promoters based on deep sequencing data of transcription start sites. To obtain a set of promoters and their associated transcripts, the 5' ends of mRNA mappings from UCSC genome database are clustered with the promoters. TF binding site predictions in the proximal promoter region are collected using 190 position weight matrices representing 350 TFs from JASPAR, TRANSFAC, motifs from the literature and their own analyses of ChIP-seq and ChIP-chip data. Additionally, miRNA target sites for about 100 seed families are annotated in the 3'UTRs of transcripts associated with each promoter.

For evaluation, ISMARA was applied to data from well-studied systems and results were compared to the literature. Inferred motif activities were highly reproducible and even more robust than the expression profiles from which motif activities were derived. When comparing samples from 16 human cell types (GEO accession number GSE30611) from younger and older donors,

ISMARA was able to identify a key regulator of aging-related changes in expression of lysosomal genes. A joint analysis of the human GNF atlas of 79 tissues and cell lines [51] and the NCI-60 reference cancer cell lines [52] revealed that many of the top dysregulated motifs were well-known in cancer biology like HIF1A and has-miR-205 miRNA. They also suggested novel predictions for regulating TFs in innate immunity, mucociliary differentiation and cancer.

### biRte [35]

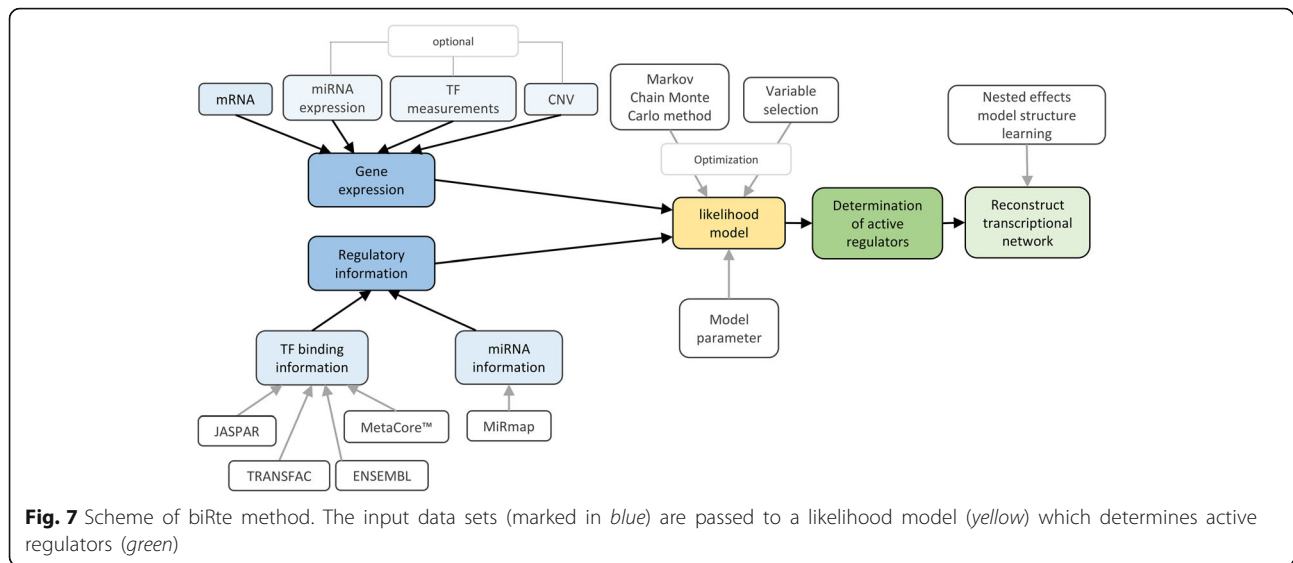
BiRte (Bayesian inference of context-specific regulator activities and transcriptional networks) takes a mathematically different approach compared to the abovementioned methods integrating TF/miRNA target gene predictions with sample specific expression data into a joint probabilistic framework (see Fig. 7). Compared to our general scheme of a TF – gene network (Fig. 2), biRte takes the TF/miRNA – gene network without the interactions between regulators to estimate regulatory activities and infers the network between regulators in a second step.

BiRte takes as input differential gene expression data (mRNA), an underlying regulatory network including TF/miRNA – target gene binding information and optionally CNV data, miRNA and TF expression measurements. As opposed to our general framework, biRte defines a likelihood model for the set of active TFs/miRNAs (called regulators  $R$  which can be seen as hidden variables) based on the entire gene expression data  $D$  and certain model parameters  $\theta$ :

$$\begin{aligned} L_{D,\theta}(R) &= p(D|R, \theta) = \prod_{\hat{D}} p(\hat{D}|R, \theta) \\ &= \prod_{\hat{D}} \prod_c \prod_i p(\hat{D}_{ic}|R_c, \theta) \end{aligned}$$

Here  $D$  represents the set of all available experimental data including mRNA, CNV, miRNA and TF expression data and  $D_{ic}$  refers to its  $i$ th feature measured under experimental condition  $c$ . The condition specific hidden state variables  $R_c$  are estimated with help of the Markov Chain Monte Carlo (MCMC) method where a regulator can switch from an active to an inactive state (switch) or an inactive and an active regulator exchange their activity states (swap). Thereby, the posterior probability for each regulator and condition to influence the expression of its target genes is estimated. Simultaneously, a variable selection procedure is applied to achieve sparsity of the model. The optimization goal is not, as one would expect, to return the configuration with highest posterior probability among all sampled ones but to take marginal selection frequencies during sampling into account and filter those above a defined cutoff. After the





determination of active regulators, the associated transcriptional network containing TFs and miRNAs is inferred from the observable differential expression of target genes and target gene predictions for individual regulators.

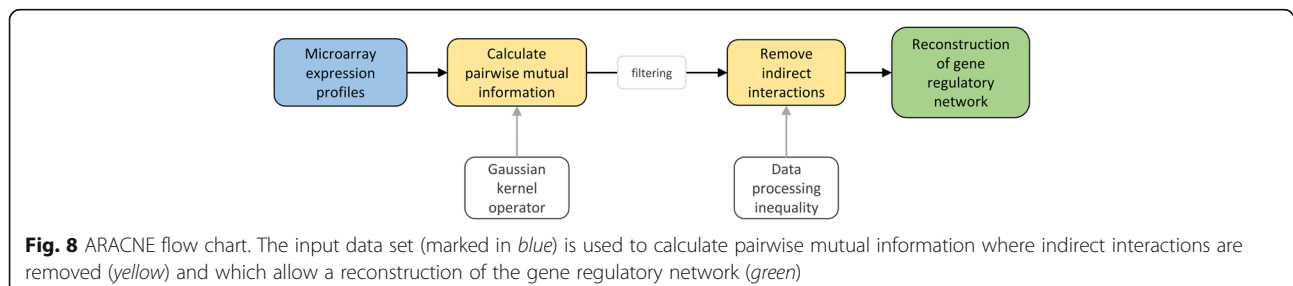
In practice, the stochastic sampling scheme based on MCMC allows swap operations only when regulators show a significant overlap of regulated targets. The variable selection procedure is implemented via a spike and slab prior [53] which can integrate prior knowledge about the activity of regulators. To infer the associated transcriptional network, Nested Effects Model (NEM) [54] structure learning is applied. An input miRNA – gene network is constructed based on MiRmap [55] for 356 miRNAs. The TF – target gene network with 344 TFs is compiled by computing TF binding affinities to promoter sequences according to the TRAP model [56] using data from ENSEMBL, TRANSFAC, JASPAR and MetaCore™. An implementation of biRte is available for R on Bioconductor under <https://bioconductor.org/packages/release/bioc/html/birte.html>.

Several simulations were conducted to study model behavior. On the basis of a human regulatory sub-network and accordingly simulated expression data of 900 target genes biRte was compared to BIRTA [57],

GEMULA [58] and a hypergeometric test and further to other network reconstruction algorithms like ARACNE [30], GENIE3 [59] and GeneNet [60]. BiRte performed best in regulator activity predictions including a favorable computation time and was robust against false positive and false negative target gene predictions. Additionally, biRte was applied to an E.coli growth control and to a prostate cancer data set including 44 normal and 47 cancer samples from GEO (GSE29079) with corresponding array data from 464 human miRNAs (GSE54516) and the results showed a principal agreement with the biological literature.

#### ARACNE [30]

We compare ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [30] as an established, yet local, tool for the reconstruction of gene regulatory networks to the previous five recent genome-wide approaches. The algorithm is background knowledge-free and identifies transcriptional interactions based on mutual information including non-linear and non-monotonic relationships and distinguishes between direct and indirect relationships (see Fig. 8). ARACNE is a free tool available under <http://califano.c2b2.columbia.edu/aracne>.



ARACNE uses as input only microarray expression profiles and estimates candidate interactions by calculating the pairwise gene expression profile mutual information  $I$  defined as

$$I(g_i, g_j) = I_{i,j} = S(g_i) + S(g_j) - S(g_i, g_j)$$

where  $S$  denotes the entropy.  $I_{i,j}$  measures the relatedness of genes  $g_i$  and  $g_j$  and equals zero if both are independent. In a second step, the mutual information values are filtered using an appropriate threshold depending on the distribution of all mutual information values between random permutations of the original data set and indirect interactions are removed.

Computationally, a Gaussian kernel operator is used to calculate mutual information scores. In a subsequent step, the data processing inequality (DPI) [61] is applied to remove probably indirect candidate interactions. The DPI states that if the genes  $g_i$  and  $g_k$  interact only through a third gene  $g_j$ , then

$$I(g_i, g_k) \leq \min(I(g_i, g_j), I(g_j, g_k))$$

Thus, the least of the three mutual information scores can come from indirect interactions only [30].

ARACNE's performance was evaluated on the reconstruction of realistic synthetic datasets [62] and on an expression profile dataset consisting of about 340 B lymphocytes derived from normal, tumor-related and experimentally manipulated populations [63] against Relevance Networks and Bayesian networks. Regarding the synthetic networks, ARACNE had consistently better precision and recall values compared to the two other algorithms and reached very good precision at significant recall levels. It recovers far more true connections and fewer false connections than the other methods with better performance on tree-like topologies compared to scale-free topologies. A reconstructed B-cell specific regulatory network was found to be highly enriched in known c-MYC targets where about 50% of the predicted genes to be first neighbors were reported in the literature.

## Results

We described five recent methods for the genome-wide inference of regulatory activity, namely the approach by Schacht et al., RACER, RABIT, ISMARA, and biRte. They all assume the topology of the regulatory network to be known, cast activity estimation as an optimization problem regarding the difference between predicted and measured values, take different types of sample specific omics data into account, and eventually produce a list of regulators like transcription factors or miRNAs, ranked by their estimated activities in the samples under study.

We also included ARACNE which is background knowledge-free and uses only local dependency measures to reconstruct a regulatory network and indirectly infer activities. All of the presented methods essentially follow the same goal, i.e., accurate ranking of regulatory activity, but differ in the types of measurements being integrated, the background knowledge necessary for their application, the complexity and refinement of the underlying model of gene regulation, and the concrete paradigm used for solving the optimization problem. Most of the methods, except for the approach by Schacht et al., are available online via a downloadable implementation, a web service, or an R package providing an operable solution for the interested user. Whereas an overview of the main features of each method can be found in Table 1, we now first compare the algorithms regarding their general properties in a descriptive way.

The data sets used for evaluation vary between all methods. Therefore, we further implemented an evaluation framework to compare the method by Schacht et al., RACER, RABIT and biRte in an objective and quantitative way. We used experimental data of three publicly available data sets from TCGA [64] and a regulatory network as background knowledge. We first used only mRNA expression data as input to the four methods to ensure the result's comparability, whereas in a second evaluation step, also other omics data sets were included where possible. We further analyzed the relevance of regulators found by different methods using a literature search.

## General properties

### Experimental data types included

The methods differ in the types of measurements being integrated, which corresponds to the level of detail of their model of gene regulations. All six methods use mRNA as input. RACER, RABIT and biRte can also integrate CNV, DNA methylation, TF/miRNA expression data, or somatic mutations. ISMARA calculates an input signal from microarray, RNA-seq, or ChIP-seq data.

Additionally, all presented methods use prior knowledge about the underlying regulatory network. These networks are extracted from different data sources and pre-processed in different manners. All methods require at least knowledge about TF – gene relationships, yet RACER, biRte and ISMARA also incorporate information about miRNAs. When using RABIT, the user can choose whether to provide TF or RNA binding protein information. The approach of Schacht et al. and biRte extract regulatory information partly from the commercial MetaCore™ database, whereas the other methods use only publicly available databases, like ENCODE, JASPAR or TRANSFAC. The networks which are used for the evaluations published in the respective papers are

publicly available for the case of RACER (network for 16653 genes, 97 TFs and 470 miRNAs), RABIT (predicted binding scores of 63 RBP motifs and 17463 genes) and biRte (network for E.coli including 160 TFs). Neither Schacht et al. nor ISMARA make this data available.

#### Mathematical models of regulatory activity

The methods use different mathematical models to infer regulatory activity. The approach by Schacht et al., RACER, RABIT and ISMARA use linear regression whereas biRte applies a probabilistic framework. ARACNE, as a local method, is based on mutual information. RACER and RABIT can be seen as extensions of the approach by Schacht et al. since they essentially use the same model structure but incorporate more input data types and more classes of regulatory information. Further, RACER applies a two-stage regression to infer regulatory activity.

#### Optimization frameworks

For assessing regulator activities, Schacht et al., RACER, RABIT and ISMARA minimize the sum of error terms between measured and predicted gene expression. However, the methods use rather different algorithms for solving the resulting optimization problem, and also apply different constraints to achieve model sparsity, robustness of inference, and feature selection. In the approach by Schacht et al., the regression model is computed for each gene separately and allows only a maximum number of six regulating TFs. RACER uses a LASSO approach, while ISMARA follows a Bayesian model that infers regulator activities as posterior distributions. LASSO can be interpreted as a Bayesian model using Laplacian priors instead of Gaussian priors in the regression framework obtaining point estimates of the regulatory activities and enforcing sparseness of the solution [32]. In contrast, biRte uses a likelihood model with a spike and slab prior to induce model sparsity. This approach implements a selective shrinkage of model coefficients such that estimates are less biased compared to a LASSO prior [65]. With the help of the spike and slab prior, sparsity can be controlled in a variable dependent manner allowing the inclusion of prior belief in the activity of each regulator [35].

#### Computed outputs

Schacht et al. and biRte determine activity of regulators over all samples at once, whereas RACER and biRte first infer sample-specific activities which are combined to cross-tumor activities only in a second optimization step. In contrast, ISMARA in first place infers motifs activity; these activities are used to deduce the effects of TFs and miRNAs by their motif binding profiles.

ISMARA primarily provides sample specific TF and miRNA activity but also offers an option to group samples and compare average regulatory activity between different conditions. Like biRte and ARACNE, it also infers the network of the regulators themselves.

#### Methods and data sets used for evaluation

The type and extent of evaluation performed for the different methods vary greatly. They range from direct application to biological problems over the comparison of results to the biological literature to simulation studies. All methods published evaluations results on publicly available datasets, e.g., from the National Cancer Institute, TCGA or GEO, but unfortunately address different tissues and cancer types. Sample-based cross-validation is applied in the work by Schacht et al., RACER, RABIT and ISMARA. The first two of these methods use correlation coefficients between measured and predicted gene expression for assessing prediction quality. RACER, RABIT and biRte compare their results to the outcome of other algorithms and to those of restricted models, for example excluding one type of the input variables. All methods search the literature to compare their predictions to previously published studies on the respective biological question. Overall, ISMARA provide the most extensive biological evaluation using a battery of relevant use cases, whereas biRte excels in systematic simulation studies. Sadly, there are very few works which compare any of the methods presented on the same problem; the only result we are aware of compared ARACNE and biRte regarding their performance in network reconstruction on simulated data, in which biRte attained higher robustness against false positive and false negative target gene predictions [35].

#### Quantitative comparison

Although certain evaluation steps were carried out for all methods, results in the original papers are not comparable as they used different input datasets, different background regulatory networks, and different evaluation metrics. Therefore, in addition to the comparison of general properties of the methods, we implemented an evaluation framework using three independent and publicly available test data sets to compare the method by Schacht et al., RACER, RABIT and biRte in an objective and quantitative way. All evaluated methods were given the same regulatory network as input.

#### Data sets

For the evaluation we used experimental data from TCGA [64] for three cancer types: Colon adenocarcinoma (COAD), liver hepatocellular carcinoma (LIHC) and pancreatic adenocarcinoma (PAAD). For all three cancer types, mRNA expression, CNV, DNA methylation

and miRNA expression data is available for primary tumor and normal tissue samples. These data sets are openly accessible via the NCI Genomic Data Commons Data Portal<sup>3</sup> or the NCI Genomic Data Commons Legacy Archive<sup>4</sup> (DNA methylation data).

For mRNA gene expression we used processed RNA-seq data in the form of FPKM (fragment per kilobase of exon per million mapped reads) values. The files included Ensembl Gene IDs which were converted to HGNC symbols using the Ensembl [66] BioMart tool<sup>5</sup> to match the IDs of the TF – gene network. In two cases, when multiple Ensembl Gene IDs mapped to one HGNC symbol, we chose the gene with highest log2 fold change between case and control group. miRNA expression was given as RPM (reads per million miRNA mapped) measurements. Both mRNA and miRNA data were centered using a weighted mean such that the mean of the case group equaled the negative mean of the control group, and normalized via a weighted standard deviation. CNV data was retrieved as masked copy number segment where the Y chromosome and probe sets with frequent germline copy-number variation had already been removed. Chromosomal regions were mapped to genes using the R package biomaRt [67]. If multiple records mapped to one gene, the median of the segment mean values was calculated. For DNA Methylation data we used the beta-values of Illumina Human Methylation 450 arrays as methylation scores. Multiple scores for the same gene were averaged within a sample.

We restricted our analyses to the samples for which all four input data types were available. When multiple measurements for one sample and data type were available, we used only the first one in alphabetical order of the file name. After this selection procedure, 165 samples remained for COAD, 404 for LIHC and 180 for PAAD. A list including sample and file information is available in Additional file 1.

Together with the experimental data, all evaluated methods were given the same regulatory network as input. We used a publicly available human TF – gene network [28] based on a text-mining approach and complemented it with TF – gene interactions from the public TRANSFAC<sup>6</sup> database [19]. This network included 2894 interactions between 429 TFs and 1218 genes. The network is provided in Additional file 2.

#### Evaluated methods

We conducted the quantitative comparison for the method proposed by Schacht et al., RACER, RABIT and biRte. ISMARA was not included since it is (a) only available as a web service, (b) can only be used with its own, proprietary underlying regulatory network model, and (c) requires the upload of raw data which is prohibited by TCGA's terms of use. Also ARACNE [30] was

not included in the quantitative evaluation since it does not use background knowledge and we therefore consider its results as incomparable to the other methods.

- For the approach by Schacht et al. we re-implemented their method as closely as possible to the original design using Python and the Cuneiform workflow language [68, 69]. Due to the high number of integer parameters in the original method, the complexity of optimizing the whole network at once would have by far exceeded computational measures. Therefore, like in the original paper, we computed the model for each gene separately and restricted the number of regulating TFs per gene to six. We added a second step where we used these TF – gene interactions building a sub-network to optimize TF activity globally to describe the interplay of the TFs' effects on their target genes. As in the implementation of Schacht et al., we used the Gurobi Optimizer.<sup>7</sup>
- For RACER we used the available R scripts<sup>8</sup> and extracted the resulting sample-specific regulatory activities.
- RABIT published a C++ implementation which they provide on their website<sup>9</sup> and which we used with the FDR option set to 1. As RABIT takes differential expression into account, we used the difference of expression values between case and control group as input and ordered the TFs by t-value as proposed in the RABIT paper.
- BiRte is available as a bioconductor R package. We used R version 3.3.2 with biRte version 1.10.0 and applied the method "birteLimma" to estimate regulatory activities with the options niter and nburnin set to 10000. As biRte has a randomized component, the resulting TF activities are not exactly the same for different runs. We averaged the final activity scores over 1000 iterations of birteLimma.

For our re-implemented method by Schacht et al. and RACER we computed separate models for case and control group and ranked the TFs by their activity difference between the two groups.

To ensure the result's comparability, we first used only mRNA expression data as input to the four methods. In a second evaluation, we included also other omics data sets where possible. BiRte was evaluated on mRNA and CNV data, RABIT on mRNA, CNV and DNA methylation data, and RACER additionally used miRNA expression as input. We obtained lists with the regulators ranked according to the absolute value of their computed activity for each cancer type and method, with and without the use of additional inputs. For each cancer type we calculated the size of the overlaps in the four different results using the



top 10 and top 100 regulators. The results for the top 10 regulators using either only mRNA or multiple omics data sets as input are shown in Table 2.

#### Only mRNA as input

When only mRNA is used as input, one TF is commonly found by the three methods RACER, RABIT and biRte in each data set, respectively: PHOX2B for COAD, EPAS1 for LIHC and ELF1 for PAAD. A literature search of these TFs and their targets revealed clear associations to the respective cancer type. The TF obtained commonly for COAD, PHOX2B, is related to TLX2, a gene which has been shown to play a role in the tumorigenesis of

gastrointestinal stromal tumors [70]. EPAS1, which was found in the LIHC top 10 TFs of three methods, is linked to CXCL12, which plays an important role in metastasis formation of hepatocellular carcinoma by promoting the migration of tumor cells [71, 72]. For PAAD, three methods ranked TF ELF1 high, which is related to 14 genes in our network, inter alia to BRCA2 and LYN. Mutations in the BRCA2 gene have been implicated in pancreatic cancer susceptibility [73, 74], whereas the knockdown of LYN reduced human pancreatic cancer cell proliferation, migration, and invasion [75]. These results underline that the methods are able to find biologically relevant information about regulation processes in cancer.

**Table 2** HGNC Symbols of the top 10 regulators found by each method for COAD (using 165 samples), LIHC (404 samples) and PAAD (180 samples) and the use of only mRNA data as input (left panel) and multiple input data sets (RACER: mRNA, miRNA, CNV and DNA methylation; RABIT: mRNA, CNV and DNA methylation; biRte: mRNA and CNV; right panel). TFs with equal activity values are marked with\*. TFs found by several method's top 10 are marked in bold (when found by RACER, RABIT and biRte), blue (RACER and RABIT), red (RABIT and biRte) or yellow (RACER and biRte)

Data set	Only mRNA				Multi-omics		
	Schacht et al.	RACER	RABIT	biRte	RACER (mRNA, miRNA, CNV, DNA methyl.)	RABIT (mRNA, CNV, DNA methyl.)	biRte (mRNA, CNV)
COAD	INSM1	HOXA5	MYC	AHR*	MIR130A	MYC	GUCA2A
	NR0B1	SP4	KLF5	NR1I3*	MIR598	NRF1	SLC25A34
	SNAI1	MECOM	CDX2	KLF5	MIR640	KLF5	PLCD1
	FOXC1	MLXIPL	NRF1	PRDM1	MIR554	RARG	AHR
	PHOX2A	CDX2	PRDM1	CDX1	MIR921	GF1B	FAM163B
	FOXA1	NRF1	NFYA	PHOX2B	MIR631	E2F1	NR1I3
	SREBF2	MYC.MAX.ZBTB17	NFKB1	ESRRA	MIR1202	CDX2	KLF4
	NR4A1	PHOX2B	PHOX2B	HOXA5	MIR548G	NFYA	TRPM6
	SNAI2	HOXA10	RARG	TCF7L2	MIR602	HOXA5	ADAMDEC1
	ARNT.HIF1A	MYC	PITX2	SOX2	MIR623	PITX2	TMIGD1
LIHC	NFIL3	GBX2	HNF4A	PHOX2A	MIR187	HNF4A	PHOX2A
	NR0B1	STAT5B	MYC	EPAS1	MIR892A	EGR1	EPAS1
	ELF2	POU3F1	NRF1	HNF4A	MIR638	SP1	HNF4A
	NR4A2	EPAS1	HNF1A	FLI1	MIR517A	NRF1	ADRA1A
	ZNF384	POU5F1	SP1	MTF1	MIR493	DNA methyl.	MTF1
	INSM1	ELK3	RARB	IKZF1	MIR572	MYC	IKZF1
	ATOH1	PHOX2A	MTF1	NFATC1	CNV	SOX10	EGR1
	SP4	FOXF2	SOX10	POU3F1	MIR192	MTF1	FLI1
	KLF11	MMP3	NR1I3	POU3F2	MIR1281	RARB	CEBPB
	POU4F1	GCM1	EPAS1	NFKB1	MIR1244	NR1I3	FOS
PAAD	RARB	ELF1	SPI1	SPI1	MIR653	DNA methyl.	RNU6-830P
	RBPJ	SATB1	GATA2	PRDM1	MIR552	SPI1	RN7SKP94
	USF1.USF2	IRF1	PES1	PES1	MIR381	PES1	RNA5SP60
	BARX2	STAT1.STAT2.IRF9	FOXO3	BACH1	MIR668	NFKB1.REL	SPI1
	USF2	IKZF1	ELF1	ELF1	MIR587	PURA	PHBP14
	STAT3.STAT1	NFATC2	RELA.REL	PURA	CNV	NFE2	TOMM22P6
	ETV4	MYF5	NFE2	TFAP2B	MIR596	ATF1	IL22
	HOXA1	GATA2	CTCF	SATB1	MIR1180	FOXO3	EEF1A1P24
	STAT4	NFYC	ATF1	NR2C2	MIR190B	NFATC2	LINC01375
	ESR1	PHOX2A	PURA	STAT1	MIR216A	IRF1	EIF4EP4



Several TFs in the top 10 are found by two of the four methods. For instance, RACER and RABIT have four common top 10 TFs (CDX2, NRF1 and MYC next to PHOX2B) in the COAD data set. However, the top 10 TFs found by the method by Schacht et al. do not overlap with any top 10 TFs of the other methods in any data set. The agreement of RACER, RABIT and biRte in the top 10 TFs hints to the biological importance of the found TFs since this overlap is statistical significant as the probability of finding common TFs in three sets of ten randomly chosen ones out of 429 TFs ( $p$ -value) is below 0.006. Additionally, the methods do identify different TFs for different data sets, indicating the importance of the actual cancer specific mRNA expression values and that results are not dictated by the background network.

The results for the number of overlapping regulators in the top 100 between the four methods and the three different data sets are shown in Fig. 9. For RABIT, only 76 TFs for COAD (resp. 67 for LIHC and 57 for PAAD) could be ranked since all other TFs had an activity value equal to zero.

When looking at the overlap of three of the four methods, the number of overlapping TFs is still the highest for the triplet RACER, RABIT and biRte. For the LIHC dataset two TFs are found in the top 100 of all four methods (E2F4 and SOX10). E2F4 is a downstream target of ZBTB7, which was associated to the expression of cell cycle-associated genes in liver cancer cells [76]. Two target genes of E2F4, CDK1 and TP73 were also involved in liver cancer development [77] and proposed as prognostic marker of poor patient survival prognosis in hepatocellular carcinoma [78]. Further, epigenetic alterations of the EDNRB gene, a target of SOX10, might play an important role in the pathogenesis of hepatocellular carcinoma [79]. Even if the result of four methods finding two common TFs is not statistically significant ( $p$ -value = 0.36), their association to liver hepatocellular carcinoma shows that the methods reach their goal of identifying relevant TFs.

However, when comparing different data sets, the methods tend to rank the same TFs under the top 100 to a greater or lesser extent. For example, the overlap of all top 100 TFs of the three cancer types is only one TF for RABIT and nine TFs for biRte, but 16 TFs for the method by Schacht et al. and even 32 TFs for RACER. Therefore, the results from RABIT and biRte seem to be more cancer type specific and less dependent on the regulatory network than the results from RACER. However, we did not specifically investigate the influence of the underlying network and its topology on the results which would be an interesting point for further research.

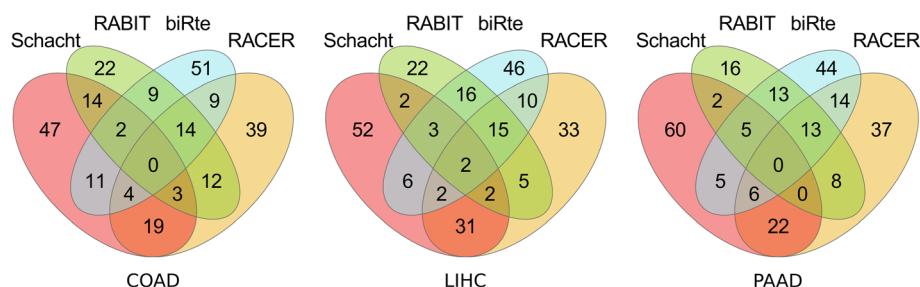
#### Multi-omics data as input

When not only taking mRNA into account but also miRNA, CNV and DNA methylation, the results are more difficult to compare between the methods, since they all use a different way of combining different types of data due to their models and implementations.

We are aware of the lower level of comparability of this approach regarding the multi-omics results in contrast to a scenario, where all methods are evaluated on the same set of input data. However, we intended to use maximum set of input data for each method to cover the effect of the use of multiple omics data sets compared to only mRNA as input.

BiRte was evaluated on mRNA and CNV data, RABIT on mRNA, CNV and DNA methylation data, and RACER additionally used miRNA expression as input. Whereas RACER and RABIT considered CNV or DNA methylation data as one background factor and compute only one activity value, biRte evaluated the influence of each CNV separately.

The results (see Table 2, right panel) show that RACER exclusively ranks miRNAs high; not a single TF is found among the top 10 regulators. Also, the influence of CNVs was high in LIHC and PAAD. However, the TFs that RACER found in the top 10 when using only mRNA data as input are still ranked high in the multi-omics scenario, e. g. the COAD top three TFs of the mRNA results are ranked 13th, 16th and 14th in the



**Fig. 9** Number of overlapping TFs in the top 100 of ranked TFs per method (for RABIT the overlap with the top 76/67/57 TFs (having activity > 0) in COAD/LIHC/PAAD is shown)

results of the multi-omics input. The difference of the results coming from the two input types is less for RABIT: seven TFs are still in the top 10 for COAD (8 for LIHC and 6 for PAAD) when using CNV and DNA methylation additionally to mRNA data. Therefore, the contribution of additional input data seems not to be crucial for the performance of RABIT. BiRte considers each CNV as a potential regulator which increases the total number of regulators enormously. Still, two commonly present TFs in the top 10 of the COAD data set (even six for LIHC and one for PAAD) are found by either the sole mRNA input and the multi-omics approach.

The overlap of the top 10 of RABIT and biRte in the multi omics case is considerable with three TFs in LIHC (HNF4A, EGR1 and MTF1;  $p$ -value = 0.001), and one TF in PAAD (SPI1;  $p$ -value = 0.21). Three of them (HNF4A, MTF1 and SPI1) were already found when using only mRNA data as input.

The results for the use of different input data sets show that the top ranked regulators are drastically changed when using additionally miRNA data in RACER, but change less when only CNV or DNA methylation data is provided in RABIT and biRte. However, the results from multi omics analyses are difficult to compare since the combination of input data sets is not consistent across the three different methods.

## Discussion

### Background networks

A crucial input to the models is the underlying regulatory network which is needed to reduce the search space for actual regulatory activity. However, the construction of comprehensive TF/miRNA – gene regulatory networks is difficult for various reasons. Firstly, a comprehensive characterization of the human regulatory repertoire is lacking since only about half of the estimated 1,500–2,000 TFs in the mammalian genome is known [80]. ChIP experiments, prone to a high false positive rate [81], were used to identify TF binding patterns but each assay is limited to the detection of one TF in one condition and therefore TF binding has not been characterized for many TFs in most cell types. Further, the local proximity of a binding site to the transcriptional start site of a gene does not automatically implicate transcriptional regulation. With regard to post-transcriptional regulators, the functions for only a few of the around 1,200 different miRNAs have been experimentally determined and current data on miRNA targets is mostly based on computational predictions [82]. Generally, the knowledge about TF and miRNA binding is scattered over the biological literature and different, partly commercial, databases, impeding the construction of comprehensive networks [28]. Therefore, any comparative

evaluation of the methods presented here would have to make sure that the same background network is used for each computation. Besides, studies on the impact of network incompleteness or different error rates in networks would be important to assess the ability of the methods to cope with such common problems. Simulation studies will be vital in this regard.

### The graph view on regulation

The modelling of regulatory networks as graphs, as used in all presented methods, is perhaps not the optimal representation for the underlying biological regulatory processes. A graph cannot easily account for important effects such as TF complex formation and temporal and spatial synchronization of activities. Furthermore, TF binding is affected by chromatin state and the impact of posttranslational modifications on transcriptional activity which are difficult to include in a graph view on regulation. The model's dependence on the topological structure and the robustness to changes in the underlying network have not been evaluated or discussed in any of the presented methods even if these issues are known to have a severe influence in network analysis [83].

### Underlying mathematical model

Linear models, widely spread in different fields of science, provide a simple and easily understandable design but over-simplify the underlying biological processes. Nonlinear behavior, e.g., saturation effects, cannot be represented. Considering that the number of available samples is typically relatively small, the incorporation of many different data types and according parameters into the model could result in excessively complex designs prone to overfitting, but this issue lacks general awareness. Only two of the presented methods incorporate parameter priors (ISMARA and biRte), and two apply cross validation techniques to estimate prediction performance (method by Schacht et al. and RACER). Further, the effect of temporal buffering between TF binding and the actual effect on gene expression is not included in any of the methods.

### Comparability

All methods produce a ranked list of regulators. Comparing these results across different methods, even when applied on the same data set and using the same background network, is difficult since no generally accepted benchmarks are available. Therefore, there currently is no objective measure to designate a best method. The closest comparable evaluation effort we are aware of is implemented in the “DREAM5 – Network Inference” challenge [84], which targets gene regulatory network reconstruction. The invited participants reverse-engineered a network from gene expression data, including a simulated

network, and evaluated the results on a subset of known interactions or the known network for the in-silico case. The approach of GENIE3 [59] which trains a random forest to predict target gene expression performed best and the integration of predictions from multiple inference methods showed robust and high performance across diverse data sets. However, an extensive competitive evaluation to determine active regulators based on a given regulatory network has, to the best of our knowledge not been carried out yet.

We therefore compared the results of four methods in a quantitative way. The experimental data and the regulatory network we used as input are publicly available to ensure transparency of our results. The results suggest that the methods are able to find biologically relevant information about regulation processes in cancer. However, the result overlaps are rather low (though sometimes statistically significant). This seems surprising as all methods essentially follow the same goal, i.e., identification of the most differentially active TFs or genes. We think further research is necessary to exactly characterize the specific strengths of each method. Furthermore, we did not investigate the influence of the underlying network on the results, which is another topic for further research.

## Conclusion

Despite their often rather involved procedures and models, none of the presented methods adequately reflects the biological reality of regulatory activity in cells. A specific disease phenotype is rarely caused by a single gene but rather a product of the interplay of genetic variability, epigenetic modifications and post-transcriptional regulation mechanisms [85]. The presented methods ignore a multitude of such factors like the effects of chromatin state and alternative splicing, nonlinear relationships between regulatory activity and gene expression, or kinetic and temporal effects. Furthermore, TFs themselves regulate the expression of other TFs forming feedback loops which are not considered in any of the presented methods. Nevertheless, the methods apparently are able to detect strong signals and produced promising results in terms of ranking transcription factors by their activity and are thus valuable tools for identifying biomarkers for specific phenotypes.

## Endnotes

<sup>1</sup><http://www.gurobi.com/products/gurobi-optimizer>

<sup>2</sup><http://lsresearch.thomsonreuters.com/pages/solutions/1/metacore>

<sup>3</sup><https://gdc-portal.nci.nih.gov>

<sup>4</sup><https://gdc-portal.nci.nih.gov/legacy-archive>

<sup>5</sup><http://www.ensembl.org/biomart/martview>, release 87

<sup>6</sup><http://www.gene-regulation.com/pub/databases.html>, release 7.0

<sup>7</sup>version 6.04, available under a free academic license

<sup>8</sup><http://www.cs.utoronto.ca/~yueli/racer.html> (accessed 17 October 2016)

<sup>9</sup><http://rabit.dfc.harvard.edu> (accessed 05 February 2016)

## Additional files

**Additional file 1:** Lists information about the samples and files from TCGA included in our quantitative evaluation for all three cancer types (COAD, LIHC and PAAD). (XLS 697 kb)

**Additional file 2:** Includes an adjacency list of the connected nodes of the TF – gene network. The list includes three columns ("TF", "gene", "edge") where each row indicates an association with the value of "edge" between a TF and a gene. Complexes of TFs are indicated with a separating "." between their components. (TXT 39 kb)

## Acknowledgements

We thank Dr. Holger Fröhlich, the author of biRte, for his help in the usage of biRte with multiple omics data sets as input and Christopher Schiefer for his contribution to the re-implementation of the method proposed by Schacht et al. We acknowledge the advice of Prof. Dr. Erik van Nimwegen concerning ISMARA. The results in this work are in part based upon data generated by the TCGA Research Network.

## Funding

We would like to acknowledge the funding provided to S.T. and J.M. from the Berlin School of Integrative Oncology (BSIO, Graduate School 1091) which is supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) in the framework of the Excellence Initiative of the German federal and state governments.

## Availability of data and materials

The experimental datasets analyzed during the current study are available in the TCGA repository, <https://gdc-portal.nci.nih.gov> under the project names TCGA-COAD, TCGA-LIHC and TCGA-PAAD. TF – gene interactions were obtained from the TRANSFAC database (release 7.0, <http://www.gene-regulation.com/pub/databases.html>) and complemented with interactions from text mining based on the paper by Thomas et al. [28] (available via the FastForward DNA database under <http://fastforward.sys-bio.net>). The TF – gene network is provided in Additional file 1.

## Authors' contributions

ST performed literature research, quantitative comparisons and drafted the manuscript with the help of JM and UL. All authors read and approved the final manuscript.

## Authors' information

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Consent for publication

Not applicable.

## Ethics approval and consent to participate

Not applicable.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 16 July 2016 Accepted: 8 March 2017

Published online: 27 March 2017

## References

- Lemon B, Tjian R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* 2000;14(20):2551–69.
- Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet.* 2012;13(9):613–26.
- Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature.* 2010;466(7308):835–40.
- Clapier CR, Cairns BR. The biology of chromatin remodeling complexes. *Annu Rev Biochem.* 2009;78:273–304.
- Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet.* 2003;33(Suppl):245–54.
- Gong X, Jia P, Zhao Z. Investigating microRNA-transcription factor mediated regulatory network in glioblastoma. 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops; 2010. p. 258–63.
- Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* 2009;37:98–104.
- Mayo MW, Baldwin AS. The transcription factor NF-kappaB: control of oncogenesis and cancer therapy resistance. *Biochim Biophys Acta.* 2000;1470(2):M55–62.
- Esquela-Kerscher A, Slack FJ. Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer.* 2006;6(4):259–69.
- Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics.* 2004;25:5–18.
- Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science.* 2007;316(5830):1497–502.
- Lou S, Lee H-M, Qin H, Li J-W, Gao Z, Liu X, Chan LL, Lam V, So W-Y, Wang Y, Lok S, Wang J, Ma RC, Tsui SK, Chan J, Chan T-F, Yip KY. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biol.* 2014;15(7):408.
- Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;30(1):207–10.
- The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.* 2008;455(7216):1061–8.
- Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Adleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Fietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M. Architecture of the human regulatory network derived from ENCODE data. *Nature.* 2012;489(7414):91–100.
- Tomba M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavese G, Pesole G, Régnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol.* 2005;23(1):137–44.
- Elemento O, Tavazoie S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.* 2005;6:R18.
- Ernst J, Plasterer HL, Simon I, Bar-Joseph Z. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.* 2010;20(4):526–36.
- Wingender E, Dietze P, Karas H, Knüppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* 1996;24(1):238–41.
- Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 2004;32:D91–4.
- Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* 2006;34:D140–4.
- Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models-a review. *Biosystems.* 2009;96(1):86–103.
- Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symp Biocomput.* 1998;18–29.
- Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. *Mol Syst Biol.* 2007;3:78.
- Li P, Zhang C, Perkins EJ, Gong P, Deng Y. Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. *BMC Bioinformatics.* 2007;8 Suppl 7:S13.
- Markowitz F, Spang R. Inferring cellular networks-a review. *BMC Bioinformatics.* 2007;8 Suppl 6:S5.
- Rung J, Brazma A. Reuse of public genome-wide gene expression data. *Nat Rev Genet.* 2013;14:89–99.
- Thomas P, Durek P, Solt I, Klinger B, Witzel F, Schulthess P, Mayer Y, Tikk D, Blüthgen N, Leser U. Computer-assisted curation of a human regulatory core network from the biological literature. *Bioinformatics.* 2015;31(8):1258–66.
- Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics.* 2014;30(4):523–30.
- Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics.* 2006;7 Suppl 1:S7.
- Schacht T, Oswald M, Eils R, Eichmüller SB, König R. Estimating the activity of transcription factors by the effect on their target genes. *Bioinformatics.* 2014;30(17):i401–7.
- Li Y, Liang M, Zhang Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput Biol.* 2014;10(10):e1003908.
- Jiang P, Freedman ML, Liu JS, Liu XS. Inference of transcriptional regulation in cancers. *Proc Natl Acad Sci.* 2015;112(25):7731–6.
- Balwierz PJ, Pachkov M, Arnold P, Gruber AJ, Mihaela Z, van Nimwegen E. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.* 2014;24(5):869–84.
- Fröhlich H. biRte: Bayesian inference of context-specific regulator activities and transcriptional networks. *Bioinformatics.* 2015;31(20):3290–8.
- Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics.* 2010;26(19):2438–44.
- Liu H, D'Andrade P, Fulmer-Smentek S, Lorenzi P, Kohn KW, Weinstein JN, Pommier Y, Reinhold WC. MRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Mol Cancer Ther.* 2010;9(5):1080–91.
- Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer.* 2006;6(10):813–23.
- Hoek KS, Schlegel NC, Brafford P, Sucker A, Ugurel S, Kumar R, Weber BL, Nathanson KL, Phillips DJ, Herlyn M, Schadendorf D, Dummer R. Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. *Pigment Cell Res.* 2006;19(4):290–302.
- The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med.* 2013;368(22):2059–74.
- Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, Hughes TR, Blencowe BJ, Frey BJ, Morris QD. Using expression profiling data to identify human microRNA targets. *Nat Methods.* 2007;4(12):1045–9.
- Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, Tsai WT, Chen GZ, Lee CJ, Chiu CM, Chien CH, Wu MC, Huang CY, Tsou AP, Huang HD. MiRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 2011;39:D163–9.
- van Kouwenhove M, Kedde M, Agami R. MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nat Rev Cancer.* 2011;11(9):644–56.
- Wang S, Sun H, Ma J, Zang C, Wang C, Wang J, Tang Q, Meyer CA, Zhang Y, Liu XS. Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat Protoc.* 2013;8(12):2502–15.
- Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh C, Kelly SM, Nabet B, Mecnas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penava LO, Lei EP, Fraser AG, Blencowe BJ, Morris QD, Hughes TR. A compendium of RNA-binding motifs for decoding gene regulation. *Nature.* 2013;499(7457):172–7.



46. National Cancer Institute Wiki. Cancer gene index End user documentation. 2014. Available: <https://wiki.nci.nih.gov/x/hC5yAQ>. [Accessed 14 Jul 2016].
47. Sadelain M, Papapetrou EP, Bushman FD. Safe harbours for the integration of new DNA in the human genome. *Nat Rev Cancer*. 2012;12(1):51–8.
48. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr LA, Kinzler KW. Cancer genome landscapes. *Science*. 2013;339(6127):1546–58.
49. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. *Nat Rev Cancer*. 2004;4(3):177–83.
50. Abbott KL, Nyrre ET, Abrahante J, Ho YY, Vogel RI, Starr TK. The candidate cancer gene database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res*. 2015;43:D844–8.
51. Su A, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the mouse and human protein encoding transcriptomes. *Proc Natl Acad Sci*. 2004;101(16):6062–7.
52. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*. 2000;24(3):227–35.
53. George EI, McCulloch RE. Approaches for bayesian variable selection. *Stat Sin*. 1997;7:339–73.
54. Markowitz F, Kostka D, Troyanskaya OG, Spang R. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*. 2007;23(13):305–12.
55. Vejnar CE, Zdobnov EM. MiRmap: comprehensive prediction of microRNA target repression strength. *Nucleic Acids Res*. 2012;40(22):11673–83.
56. Roeder HG, Kanhere A, Manke T, Vingron M. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*. 2007;23(2):134–41.
57. Zacher B, Abnaof K, Gade S, Younesi E, Trescher A, Fröhlich H. Joint bayesian inference of condition-specific miRNA and transcription factor activities from combined gene and microRNA expression data. *Bioinformatics*. 2012;28(13):1714–20.
58. Geeven G, van Kesteren RE, Smit AB, de Gunst MC. Identification of context-specific gene regulatory networks with GEMULA-gene expression modeling using Lasso. *Bioinformatics*. 2012;28(2):214–21.
59. Huynh-Thu VA, Irtthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*. 2010;5(9):e12776.
60. Opgen-Rhein R, Strimmer K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst Biol*. 2007;1:37.
61. Cover T, Thomas J. Elements of Information Theory. New York: Wiley; 1991.
62. Mendes P, Sha W, Ye K. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*. 2003;19 suppl 2:i122–9.
63. Klein U, Tu Y, Stolovitzky GA, Mattioli M, Cattoretti G, Husson H, Freedman A, Inghirami G, Cro L, Baldini L, Neri A, Califano A, Dalla-Favera R. Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *J Exp Med*. 2001;194(11):1625–38.
64. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas Pan-cancer analysis project. *Nat Genet*. 2013;45(10):1113–20.
65. Hernández-Lobato D, Hernández-Lobato JM, Suárez A. Expectation propagation for microarray data classification. *Pattern Recognit Lett*. 2010;31(12):1618–26.
66. Yates A, Akanni W, Amodé MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Nuhn M, Parker A, Patricio M, Pignatelli M, Rahtz M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Birney E, Harrow J, Muffato M, Perry E, Ruffier M, Spudich G, Trevanion SJ, Cunningham F, Aken BL, Zerbino DR, Flicek P. Ensembl 2016. *Nucleic Acids Res*. 2016;44(D1):D710–6.
67. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*. 2009;4(8):1184–91.
68. Brandt J, Bux M, Leser U. Cuneiform: a functional language for large scale scientific data analysis. *Proc Work EDBT/CDT*. 2015;1330:17–26.
69. Bux M, Brandt J, Lipka C, Hakimzadeh K, Dowling J, Leser U. SAASFE: scalable scientific workflow execution engine. *Very Large Data Bases*. 2015;8(12):1892–5.
70. Naumov VA, Genozov EV, Zaharjevskaya NB, Matushkina DS, Larin AK, Chernyshov SV, Alekseev MV, Shelygin YA, Govorun VM. Genome-scale analysis of DNA methylation in colorectal cancer using Infinium human methylation 450 bead chips. *Epigenetics*. 2013;8(9):921–34.
71. Liu H, Pan Z, Li A, Fu S, Lei Y, Sun H, Wu M, Zhou W. Roles of chemokine receptor 4 (CXCR4) and chemokine ligand 12 (CXCL12) in metastasis of hepatocellular carcinoma cells. *Cell Mol Immunol*. 2008;5(5):373–8.
72. Rubie C, Frick VO, Wagner M, Weber C, Kruse B, Kempf K, König J, Rau B, Schilling M. Chemokine expression in hepatocellular carcinoma versus colorectal liver metastases. *World J Gastroenterol*. 2006;12(41):6627–33.
73. Couch FJ, Johnson MR, Rabe KG, Brune K, de Andrade M, Goggins M, Rothenmund H, Gallinger S, Klein A, Petersen GM, Hruban RH. The prevalence of BRCA2 mutations in familial pancreatic cancer. *Cancer Epidemiol Biomarkers Prev*. 2007;16(2):342–6.
74. Greer JB, Whitcomb DC. Role of BRCA1 and BRCA2 mutations in pancreatic cancer. *Gut*. 2007;56(5):601–5.
75. Je DW, O YM, Ji YG, Cho Y, Lee DH. The inhibition of SRC family kinase suppresses pancreatic cancer cell proliferation, migration, and invasion. *Pancreas*. 2014;43(5):768–76.
76. Yang X, Zu X, Tang J, Xiong W, Zhang Y, Liu F, Jiang Y. Zbtb7 suppresses the expression of CDK2 and E2F4 in liver cancer cells: implications for the role of Zbtb7 in cell cycle regulation. *Mol Med Rep*. 2012;5(6):1475–80.
77. Bisteau X, Caldez MJ, Kaldis P. The complex relationship between liver cancer and the cell cycle: a story of multiple regulations. *Cancers*. 2014;6(1):79–111.
78. Stiewe T, Tuve S, Peter M, Tannapfel A, Elmaagacli AH, Pützer BM. Quantitative TP73 transcript analysis in hepatocellular carcinomas. *Clin Cancer Res*. 2004;10(2):626–33.
79. Hsu LS, Lee HC, Chau GY, Yin PH, Chi CW, Lui WY. Aberrant methylation of EDNRB and p16 genes in hepatocellular carcinoma (HCC) in Taiwan. *Oncol Rep*. 2006;15(2):507–11.
80. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*. 2009;10(4):252–63.
81. Pickrell JK, Gaffney DJ, Gilad Y, Pritchard JK. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics*. 2011;27(15):2144–6.
82. Rajewsky N. microRNA target predictions in animals. *Nat Genet*. 2006;38(Suppl):S8–13.
83. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*. 2004;431:308–12.
84. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Consortium TD, Kellis M, Collins JJ, Stolovitzky G. Wisdom of crowds for robust gene network inference. *Nat Methods*. 2012;9:796–804.
85. Daviden PK, Turan N, Egginton S, Falciani F. Multi-level functional genomics data integration as a tool for understanding physiology: a network perspective. *J Appl Physiol*. 2016;120(3):297–309.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

