Nucleic Acids Research Advance Access published October 7, 2014

doi: 10.1093/nar/gku916

Large-scale modeling of condition-specific gene regulatory networks by information integration and inference

Daniel Christian Ellwanger[1,2,*], Jörn Florian Leonhardt[2] and Hans-Werner Mewes[1,2]

[1]Chair of Genome-Oriented Bioinformatics, Technische Universität München, Center of Life and Food Sciences Weihenstephan, 85354 Freising, Germany and [2]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany

Received October 23, 2013; Revised September 10, 2014; Accepted September 22, 2014

ABSTRACT

Understanding how regulatory networks globally coordinate the response of a cell to changing conditions, such as perturbations by shifting environments, is an elementary challenge in systems biology which has yet to be met. Genome-wide gene expression measurements are high dimensional as these are reflecting the condition-specific interplay of thousands of cellular components. The integration of prior biological knowledge into the modeling process of systems-wide gene regulation enables the large-scale interpretation of gene expression signals in the context of known regulatory relations. We developed COGERE (http://mips.helmholtz-muenchen.de/cogere), a method for the inference of condition-specific gene regulatory networks in human and mouse. We integrated existing knowledge of regulatory interactions from multiple sources to a comprehensive model of prior information. COGERE infers condition-specific regulation by evaluating the mutual dependency between regulator (transcription factor or miRNA) and target gene expression using prior information. This dependency is scored by the non-parametric, nonlinear correlation coefficient η^2 (eta squared) that is derived by a two-way analysis of variance. We show that COGERE significantly outperforms alternative methods in predicting condition-specific gene regulatory networks on simulated data sets. Furthermore, by inferring the cancer-specific gene regulatory network from the NCI-60 expression study, we demonstrate the utility of COGERE to promote hypothesis-driven clinical research.

INTRODUCTION

Cellular processes are programed through regulatory control and are conditionally modulated. Gene expression is a highly regulated mechanism that has a profound impact on crucial processes such as cell division, differentiation and apoptosis. Its malfunction can lead to the pathogenesis of fatal diseases (1,2). The regulation of gene expression covers a number of sequential processes controlling the RNA concentration of target genes (TGs) selectively regulating the quantity of gene products in the cell. Transcriptional regulation is controlled through proteins called transcription factors (TFs). Combinatorial interactions of RNA-binding proteins and non-coding RNAs with regulatory elements located on target RNA molecules determine the functional outcome of target RNA processing, such as splicing, polyadenylation, export, stability and translation (3). At this, a family of small RNAs of about 22 nucleotides in length without protein-coding potential called microRNAs (miRNAs) has attracted a lot of attention. Integrated within a multiprotein complex they bind to target sites preferably located in the 3′ untranslated region (4) or the coding sequence (5) of mRNAs to govern stability and translational efficiency. Post-transcriptional regulation by miRNAs is an essential regulation layer for higher eukaryotes. One miRNA is able to regulate a large number of protein-coding genes and vice versa one mRNA can be regulated by several miRNAs. By intertwining with transcriptional gene regulatory networks (GRNs), miRNA regulation induces extensive interacting control structures. Both types of regulator genes (RGs), namely TFs and miRNAs, span a global GRN that controls thousands of mammalian TGs and forms multilayer regulatory circuits (6).

Novel technologies promote the ongoing transformation of biology from a data-poor to an increasingly data-rich science. The attendant increase in the number, size and diversity of data sources features knowledge for both, TF:TG and miRNA:TG interactions. The integration of this information offers unprecedented and as yet, largely unrealized

*To whom correspondence should be addressed. Tel: +49 8161 712131; Fax: +49 8161 712186; Email: ellwanger@wzw.tum.de

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

opportunities for discoveries from the analysis of large-scale GRNs. However, each data source has its unique bias and inherent potential drawbacks. Sequence-based predictions are rather exhaustive but yield a significant fraction of false positives due to the limited comprehension of the molecular basis of the regulator:target pairing process. Databases with experimentally verified data and high-profile studies provide an impressive amount of information but are far from complete. The biomedical literature is rich in known regulatory interactions but these are difficult to extract. All biological data sources naturally exhibit semantic differences that are caused by varying levels of granularity or abstraction at which objects and their relationships are described. These aspects illustrate the potential and importance of sophisticated data-driven integration approaches.

The fact that integrated networks contain regulatory interactions that were described under varying conditions makes these GRNs comprehensive, but also unspecific and static; also the regulatory sign (stimulation/repression) of potential relations is largely unknown. Since transcriptional and miRNA-mediated post-transcriptional regulation is context-dependent, it is evident that static GRNs are not sufficient to represent regulatory interactions taking place under changing conditions. Modeling condition-specific GRNs using prior information from integrated networks aims to overcome these problems and will facilitate a better understanding on how gene expression is modulated.

With rapidly increasing amounts of gene expression profiles, an exhaustive insight into their underlying large-scale condition-specific GRNs becomes feasible and attractive. Therefore, we developed COGERE (modeling of COndition-specific GEne REgulation; from the Latin 'to collect'), an approach to infer condition-specific gene regulation from gene expression data integrating existing knowledge of regulatory interactions. This approach enables the interpretation of multi-dimensional expression profiles reflecting the dynamic interplay of thousands of cellular components in the context of known regulatory relations. We build a data structure of transcriptional and miRNA-mediated gene regulation (prior model) by integrating automatically and manually mined interactions from all available biomedical text with information from relevant databases, recent studies and computational predictions from sequence data. In addition to an increased sensitivity, COGERE is able to suggest references for inferred interactions that were described in the literature. This will facilitate the generation of novel, testable hypotheses. To compute the condition-specific strength of association from gene expression data, COGERE uses a two-way nonlinear non-parametric analysis of variance (ANOVA) considering prior information. This association metric overcomes the disadvantages of common approaches utilizing linear correlation (7,8) and mutual information (8,9). Linear correlation requires miRNA and mRNA expression profiles to be obtained from the same set of individuals (matched data), and inherently detects only linear relations. Mutual information needs careful discretization of the expression data to avoid loss of signal and, in addition, is non-negative, and as such does not provide information about the condition-specific sign of interaction.

In this work, we report the construction of the COGERE framework and show that our approach significantly improves existing methods for the large-scale modeling of miRNA-mediated condition-specific GRNs. Further, we demonstrate the utility of COGERE by inferring a cancer-specific regulatory network from the NCI-60 (10) microarray project.

MATERIALS AND METHODS

COGERE maps regulatory complexity by reconstructing GRNs involving TFs or miRNAs as regulators (Figure 1A). The workflow of COGERE is outlined in Figure 1B. In the following each step of the framework (information integration, inference), the evaluation and the data analysis of the use-case are described in detail.

Construction of the prior model by information integration

The prior network is composed of *in vivo*, *in vitro* and computationally determined regulator:target interactions. Several heterogeneous data sources were combined to a single-graph data model. All genes with their symbols, gene synonyms and identifiers as listed in NCBI Entrez Gene (11) and miRBase version 19 (12) were added as vertices to the regulatory graph. Regulatory associations were stored as directed interactions between two gene nodes. Each interaction was weighted by a prior score that ranks its regulatory potential; this score is computed from the integrated evidences as specified below.

Integration of transcriptional regulatory interactions. To predict transcriptional regulatory associations, we obtained human and murine promoter sequences of protein-coding genes from the ElDorado database version 08-2011 (ElDorado; http://www.genomatix.de). For miRNA genes, we collected promoters from Fujita *et al.* (13) and CoVote (14) and transcriptional starts from CoreBoost_HM (15), Corcoran *et al.* (16), Marson *et al.* (17), Ozsolak *et al.* (18), miRStart (19) and Eponine-TSS (20). Given a median promoter length of 448 nt in the study of Fujita *et al.* and 350 nt predicted by CoVote, we extracted adequate promoter sequences from 500 nt upstream to 100 nt downstream relative to a transcriptional start site. We obtained chromosomal locations of all miRNA hairpins from miRBase and calculated the distances between a miRNA hairpin start position and all promoter start positions. For each miRNA gene, we selected the promoter located closest to its hairpin sequence. If miRNA genes shared the same promoter and had an inter-gene distance of up to 50 kb, they were proposed to form a transcriptional unit (21). We filtered promoters located up to 50 kb upstream of a miRNA gene or a transcriptional unit (22). Additional promoter regions of intragenic miRNAs located on the same strand and within an intron of a protein-coding gene were considered coincident with the one defined for the host gene (23). Gene annotations were obtained from Ensembl (24). All promoter sequences were scanned for vertebrate TF matrix matches using the MatInspector algorithm (matrix family library version 8.4) (25). We utilized ModelInspector (module library version

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

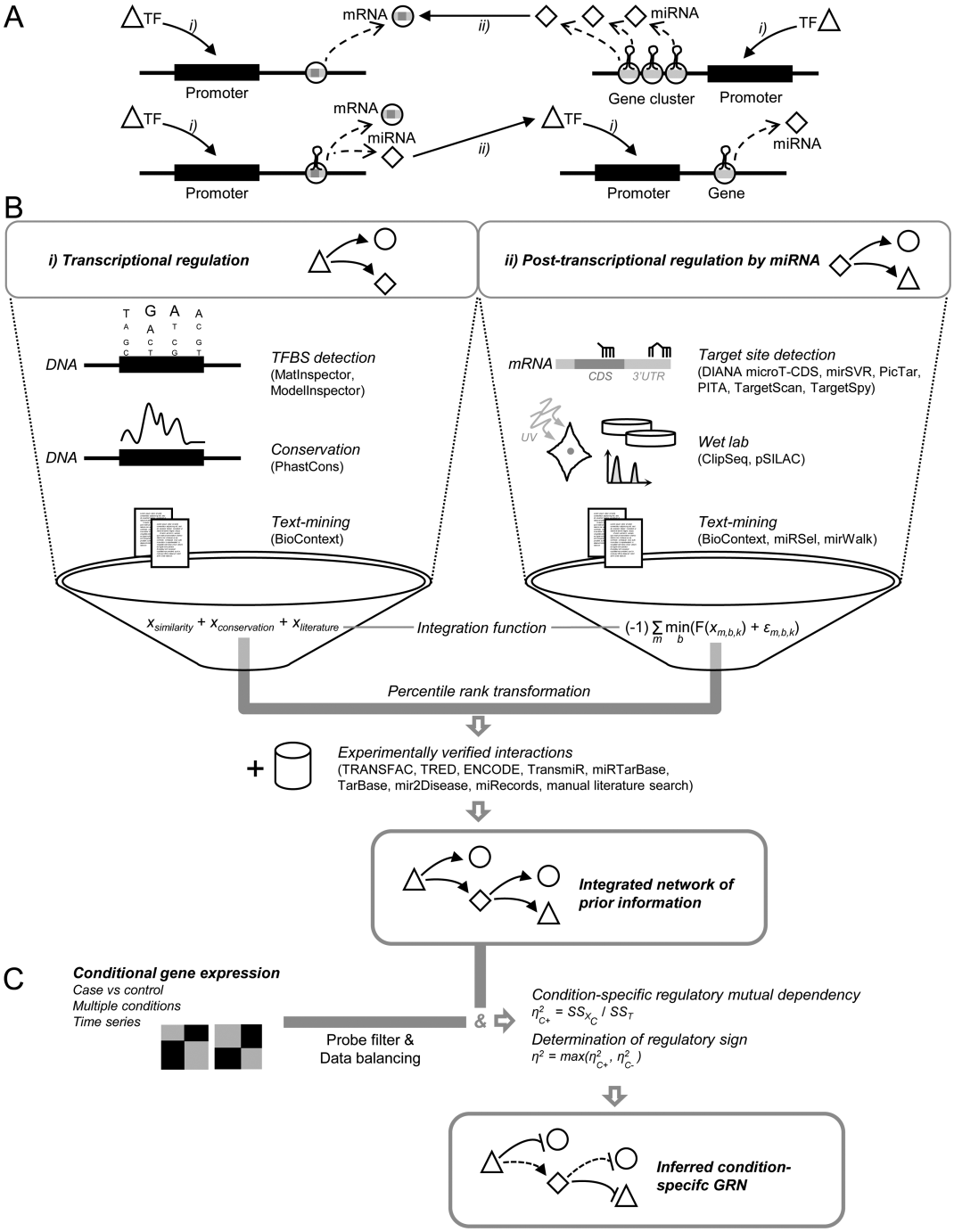


Figure 1. Overview of the COGERE workflow. (**A**) Outline of the biological paradigm of TF and miRNA interplay in gene expression regulation considered by COGERE: (i) transcriptional regulation is conducted by TFs binding to sites in promoter regions on the DNA of genes encoding either proteins or non-coding RNAs such as miRNAs. Here, miRNAs can be co-regulated with its protein-coding host gene, within a transcriptional unit (gene cluster), and/or through its own promoter (ii) miRNA regulation takes place post transcription by binding to sites mainly located on the 3′ untranslated region (3′-UTR) and/or the coding sequence (CDS) of the target mRNA. The transcriptional and post-transcriptional regulatory pathways are interconnected. (**B**) Construction of the prior network by information integration. For transcriptional regulation we combined predicted transcription factor-binding sites (TFBS), their conservation and mined interactions from biomedical text by a linear integration function. For post-transcriptional regulation, we integrated individual scores of six miRNA target prediction algorithms and text-mining results to a unified score weighting the regulatory potential of a miRNA:TG interaction using AGO-bound CLIPseq data and proteomics (pSILAC) data. All scores computed by the relevant integration function were normalized to percentile ranks (= prior score). Experimentally verified interactions were added to the prior network (prior score = 1). (**C**) Determination of condition-specific regulation. For user-specified normalized and $\log_2$-transformed mRNA and/or miRNA expression data of at least two conditions, COGERE computes for each interaction of the prior network, the strength of the conditional dependency and the condition-specific regulatory sign (stimulation/repression) by deriving the coefficient η^2 with its corresponding P by a two-way ANOVA.

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

5.5) (26) to filter experimentally verified vertebrate modules of transcriptional regulatory units, functional composite elements consisting of at least two TF-binding sites in conserved order and distance. PhastCons (27) scores from 46-way (human) and 30-way (mouse) alignments of vertebrates available through UCSC (http://genome.ucsc.edu/) were used to calculate mean conservation levels of potential TF-binding sites. We required each candidate target site to correspond to the most conserved nucleotide at 95% of all positions of the TF matrix or to be conserved with an average score of at least 95%. Moreover, we extracted all regulatory interactions found by the text-mining tool BioContext (28). We required the associations of two biological entities to be organism-specific and the interaction type to be included in the set of terms: regulation, positive regulation and negative regulation. BioContext provides a score for each event mirroring the precision of the identified association based on specific event features. This allows for each TF:target interaction the computation of the prior score based on the TF matrix similarity score $x_{\text{similarity}}$, the conservation score $x_{\text{conservation}}$ and the text-mining score $x_{\text{literature}}$ as follows:

$$\text{prior} = x_{\text{similarity}} + x_{\text{conservation}} + x_{\text{literature}} \quad (1)$$

at which $x_{\text{similarity}}$, $x_{\text{conservation}}$ and $x_{\text{literature}}$ are scaled between 0 and 1 by

$$F(x) = \frac{x_i - \min(x)}{\max(x) - \min(x)}. \quad (2)$$

Integration of miRNA-mediated post-transcriptional regulatory interactions. Due to the diverse feature and model selection of miRNA:target prediction approaches (29,30), we selected a set of six current algorithms to cover a wide range of different miRNA targeting characteristics: DIANA-microT-CDS (5), mirSVR (31), PicTar (32), PITA 3/15 (33), TargetScan 6.1 (34) and TargetSpy (35). Additionally, we integrated predicted interactions from literature mining provided by miRSel (36), miRWalk (37) and BioContext (28). For miRSel and miRWalk, we scored each interaction by the number of retrieved documents containing a co-occurrence between the miRNA and its target. To minimize the false positive rate, we obtained only the most confident predictions of each tool as recommended by the authors, respectively (Supplementary Table S1). We utilized CLIP-Seq data from starBase version 1.0 (38) to identify predicted target sites located in an Argonaute (AGO) CLIP-Seq peak cluster. Here, each cluster holds a biological complexity score b describing a measure of reproducibility between biological replicates or experiments. We prepared six score vectors $x_{m,b}$ with $\{b = 0, b = 1, b = 2, b = 3, b = 4, b \geq 5\}$ for each prediction method $m \in M$. A biological complexity of $b = 0$ denotes target sites not located in any annotated AGO2-binding region. For each miRNA:target interaction the best score was retained. We transformed each prediction score vector $x_{m,b}$ into an efficiency score vector $y_{m,b}$ of protein downregulation based on miRNA transfection data from Selbach *et al.* (39) (Supplementary Figure S1). We computed a regression function $F(x_{m,b,k})$ for the k-th prediction score and the average log fold change $y_{m,b,k}$ of all miRNA:target pairs with $x_{m,b,l} \geq x_{m,b,k}$ and a random error $\varepsilon_{m,b,k}$:

$$y_{m,b,k} = F(x_{m,b,k}) + \varepsilon_{m,b,k}. \quad (3)$$

We used the locally weighted least-squares method to fit the polynomial function of the predictor (40). For each miRNA:target pair we computed the prior score:

$$\text{prior} = (-1) \sum_{m \in M} \min_b(y_{m,b,k}). \quad (4)$$

Transformation of prior scores. The integration of independent sources inherently results in non-identical, heterogeneous prior score distributions. To obtain unified weights for each interaction type $k \in \{\text{TF:TG, miRNA:TG}\}$, we converted the raw prior scores $x_i \in \text{prior}_k$ to percentile rank scores:

$$F(x_i) = \frac{0.5|\{x_j : x_j = x_i\}| + |\{x_j : x_j < x_i\}| + 0.5}{|\text{prior}_k|}. \quad (5)$$

This equation allows an intuitive interpretation of the prior scores, e.g. a transformed prior_k score of 0.90 denotes an interaction with a higher regulatory potential than 90% of all interactions of type k contained in the prior model; in return a prior cutoff of 0.90 retains the 10% most reliable predicted regulatory associations. Note that Equation (5) computes the mean rank for ties (see Supplementary Methods S1).

Integration of verified regulatory interactions. We collected experimentally verified TF:TG interactions from ENCODE (41), TRED (42), TRANSFAC (43), TransMir (44) and from manual literature search. We obtained miRNA:TG interactions from miRecords (45), miRTarBase (46), miR2Disease (47) and Tarbase (48). For each interaction contained in one of these sources the prior score was set to 1.0.

Determination of condition-specific regulation by inference

Preprocessing of expression data. We apply two preprocessing steps to input data in order to improve the discriminatory power of our approach.

(i) *Balancing the data*. To avoid a condition-dependent bias, the sets of microarrays measured under the same condition are pruned to equal size n. We compute the $M_{i,j}$ value for each probe j on microarray i by dividing the intensity of j by the median intensity of the same probe across all microarrays. According to (49) $M_{i,j}$ can be decomposed to the probe effect z_j, the differential expression effect $\beta_{i,j}$ and an error term $\varepsilon_{i,j}$. As z_j and $\beta_{i,j}$ are the same across all k samples within one condition, computing the sum of all L_1 distances enables to filter the n microarrays with minimal technical variation:

$$d_i = \sum_k \sum_j |M_{i,j} - M_{k,j}|. \quad (6)$$

We ranked all samples of each condition by their increasing order of d_i and selected the top n microarrays for further processing.

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

(ii) *Filtering of non-present and uninformative transcripts*. In order to assess the context-specific strength of associations, the transcripts of both regulator and target have to meet the following two requirements: the genes need to be expressed in all samples of interest and to show significant variation across the different conditions. Regarding the latter, TGs whose expression does not alter between the different conditions are unlikely to be under context-specific regulation. We filter all probe sets that have sufficient expression intensities ($>\log_2 20$) on more than 5% of the microarrays and exhibit an adequate variation across samples (probe set expression interquartile range > median expression interquartile range) (50).

Inferring condition-specific regulation by ANOVA. To score regulatory associations of the prior model in terms of condition-specific relevance, we utilize the non-parametric, nonlinear correlation coefficient η^2 (eta squared) (51). This variable is derived from a two-way ANOVA and enables the quantification of the mutual dependency between a regulatory pair based on gene expression profiles over different experimental conditions. We model observed expression data with n replicates and k conditions as responses of two factors X_C (condition) and X_G (RG and TG), their potential interaction $X_C \times X_G$ and the proportion of variation which cannot be explained by the model ε (measurement noise). Variance can be expressed in terms of the sum of squared deviations from the mean (sum of squares, SS) (52). Accordingly, a two-way ANOVA splits the total sum of squares (SS_T) into four parts:

$$SS_T = SS_{X_C} + SS_{X_G} + SS_{X_C \times X_G} + SS_\varepsilon. \quad (7)$$

Here, SS_{X_C} reflects the effect of differential gene expression between the conditions, SS_{X_G} is the difference in means of the expression profiles of RG and TG, $SS_{X_C \times X_G}$ denotes the joint effect of both factors and SS_ε quantifies the variation due to inaccuracy of measurement. For each regulatory pair we extract two matrices of size $n \times k$ containing the expression values of the RG and the TG, respectively. From Equation (7), we calculate for each Z-score standardized expression matrix the mutual dependence in gene expression between the different conditions. It is defined as the fraction of total variation explained by the variation in the data between conditions:

$$\eta^2_{C+} = \frac{SS_{X_C}}{SS_T}, \eta^2_{C+} \in [0, 1]. \quad (8)$$

This value can be interpreted in the same way as common correlation coefficients. We take into account that η^2 does not explicitly test for negative regulation: we reversed the sign of the RG data and calculate η^2_{C-}. As a final score we define $\eta^2 = \max(\eta^2_{C+}, \eta^2_{C-})$ (53). Interactions with $\eta^2_{C+} < \eta^2_{C-}$ are signed as repression, otherwise as stimulation.

Regulatory associations showing a strong conditional dependency between RG and TG, i.e. having a high η^2 score, were assumed to be of high relevance. We test this dependency for statistical significance by an F-test. For each η^2 the corresponding F-value is calculated by dividing the effect variance of factor X_C by the total variance:

$$F_{X_c} = \frac{MS_{X_C}}{MS_T}, MS_i = \frac{SS_i}{df_i}, i \in \{X_C, T\}, \quad (9)$$

where the degrees of freedom are chosen $df_{X_C} = k - 1$ and $df_T = 2 \times n \times k - 1$. P-values (P) are obtained from the F-distribution and adjusted by the Benjamini–Hochberg procedure (54) to control the false discovery rate (FDR). The pseudocode of the algorithm can be found in Supplementary Methods S2.

Evaluation

Performance assessment of the integration function. The set of mRNA expression data was taken from the miRNA transfection study performed by Linsley *et al.* (55). We obtained the data from the NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under accession GSE6838. The expression profiles were measured at 24 h post-transfection featuring maximal mRNA silencing but minimal secondary effects by protein depletion. The expression profiles of HeLa, HCT116 Dicer^{ex5} and DLD-1 Dicer^{ex5} miRNA-transfected cells relative to mock-transfected cells were computed. We mapped probe identifiers to NCBI Gene accession numbers and selected the probe with the lowest log ratio P for each gene. To obtain statistically meaningful results, we considered only experiments for which each prediction tool scored at least 150 interactions. Finally, we obtained 18 expression profiles containing 10 miRNAs (miR-106b, miR-16, miR-15a, miR-20a, miR-195, miR-103, let-7c, miR-107, miR-17-5p and miR-103). We recomputed the COGERE prior scores without the information of validated interactions to make the scores comparable among approaches. To mimic the integration functions of mirConnX (7) and MAGIA2 (8) we used the six miRNA target prediction algorithms incorporated in the COGERE prior score. We implemented the scoring scheme of mirConnX by weighting a miRNA:TG association by the proportion of algorithms predicting the interaction. For MAGIA2, we computed all 57 possible intersections between the six miRNA target prediction algorithms. We calculated the rank correlation between the observed gene $\log_2$ fold changes following miRNA transfection and the scores computed for the miRNA:TG interaction. Further, we performed a precision–recall analysis using the top and bottom 20% of candidate TGs based on their expression changes.

Benchmark of prediction accuracy. We generated an *in silico* gold standard of 80 regulatory networks extracted from a human source network composed of 64 029 experimentally verified interactions using GeneNetWeaver (56). Each sub-network contained 500 nodes and a varying number of edges (min = 852, median = 1226 and max = 1421) of which 50% were set to occur in a given set of conditions to obtain a balanced test set. Stochastic dynamical models of gene regulation accounting for molecular and experimental noise were applied to simulate matched expression data of mRNA and miRNA. We simulated the steady-state expression of all genes for 60 conditions [c.f. NCI-60 cancer microarray project (10)], with five replicated measurements

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

each. We applied a precision–recall analysis for the determination of the prediction accuracy of condition-specific regulation as well as a $\text{precision}_{\text{sign}}$–recall analysis for the prediction of the sign of a regulatory interaction. For further details about the construction of the benchmark suite, the evaluation metrics and the application of the prediction methods refer to Supplementary Methods S3 and Supplementary Figure S2A.

Case study data. We extracted the raw total gene signals of the NCI-60 Agilent microarray measurements from the Liu *et al.* study (57). Six cell samples (MCF7, HCT116, HT29, K562, SK-MEL-2 and CAK1-1) were labeled in quadruplicated, and the remaining samples were labeled in duplicate. In accordance with the manufacturer, we set probe intensities <5.0 to 5.0 and removed spots if the gene was not detected on the microarray. The data were quantile normalized (58) and $\log_2$ transformed. All probes were assigned a miRBase ID or Entrez Gene ID, respectively. We obtained 789 miRNA probes measuring 533 genes and 26 091 mRNA probes measuring 16 651 genes. Processed data from the NCI-60 DTP human tumor cell line screen measuring the activity of 19 941 chemical compounds (drugs) in NCI-60 cell lines were obtained from CellMiner (http://discover.nci.nih.gov/cellminer/). As proposed by Liu *et al.* (57) relationships between drug activity and gene expression were computed by Pearson correlation. At this, we averaged the expression of a gene transcript for replicates of cell lines. The FDA status of each compound was taken from the CellMiner database (annotation of 07/30/2013, version 1.4).

RESULTS

Comprehensive information integration

As the prior model of COGERE defines the hypothesis space for the inference of condition-specific regulation, the information integration step has to be extensive. We combined several sources containing regulatory interaction information to a unique, directed graph constituting a static model of feasible gene regulation. Each interaction was weighted by a prior score computed by a domain-specific integration function. COGERE contains a regulatory network with 5 481 057 interactions for human and 3 472 682 interactions for mouse. Comparing the amount of high-confident interactions (prior score > 0.9) to recent data pools, our model contains an extensive set of qualitative information: 294 394 TF:TG, 11 258 TF:miRNA and 316 875 miRNA:TG interactions in human. In comparison, ENCODE (41) features about 27 386 TF:TG, TransmiR (44) 353 TF:miRNA and the recent release of miRTarBase (46) about 45 540 miRNA:TG human regulatory associations. Since there is less data available for mouse, the information gain is even higher: 199 308 TF:TG, 4105 TF:miRNA and 156 779 miRNA:TG high-confident interactions. In comparison: TRANSFAC (43) has 1118 TF:TG, TransmiR 16 TF:miRNA and miRTarBase 13 405 murine interactions. The TransmiR database provides regulatory associations for only 9% of human miRNA genes and 2% of murine genes. This shortcoming was substantially improved by extensively collecting data from existing studies and carefully predicting promoter sequences. We considered that miRNA genes can be embedded within a protein-coding host gene, and/or being part of an independent transcriptional unit, and/or can have their own promoter. COGERE predicts transcriptional regulation for 51% of all human and 50% of all murine miRNA genes (as annotated in miRBase 19). This is an increase compared to existing integrative approaches that model transcriptional regulation of about 29% [MAGIA2 (8)] to 31% [mirConnX (7)] of human miRNA genes and between 46% and 47% of murine miRNA genes, respectively.

For the first time, we automatically integrated mined interactions from all available biomedical text with information from databases and their corresponding references. Our current prior model contains 141 713 references for 97 816 interactions in human and 44 950 references for 25 142 interactions in mouse.

Improved weighting of miRNA:TG interactions *a priori*

Since COGERE integrates six miRNA target prediction algorithms into a unique scoring framework under consideration of individual target scores, we were interested to know whether our integration function improves previous approaches such as the ordinary intersection of several tools.

First, we compared the COGERE prior scores with the prior scores computed by the integration function used by mirConnX (7). The latter weights each miRNA:TG interaction by the fraction of target prediction tools confirming a potential regulation. The outcome is a prior network with a discrete score distribution composed of {0, 1/6, 1/3, 1/2, 2/3, 5/6, 1}. To obtain the intrinsic value of how well the weights describe the regulatory potential of an interaction, we evaluated the overall ranking performance of both scoring schemes. For this purpose, we computed the Spearman's rank correlation between the observed $\log_2$ expression change following miRNA transfection and the prior weights of the miRNA:TG interactions, respectively. Figure 2A shows that both attempts for combining multiple target prediction tools exhibit a better performance compared to the average performance of all individual tools. At this, the COGERE prior score strongly outperforms the basic weighting used by mirConnX. In contrast to the prior score, the simple combination of target prediction tools is not optimized to describe potential miRNA-induced expression changes. In 16 of the 18 experiments, the performance of the prior score outperforms the basic scoring framework that constitutes a significant improvement (paired signed rank test $P = 1.7 \times 10^{-4}$). In all except one case, the COGERE integration function is superior to a blindfolded random selection of a single algorithm.

Second, to analyze the performance of the prior score and the intersection of tools as used in MAGIA2 (8) we generated all 57 possible intersections composed of at least two of the six algorithms. We defined precision as the fraction of predictions that are true positives and recall as the proportion of actual positives that are correctly identified as such. Figure 2B shows that the prior score strongly improves the precision of the prior network over almost all values of recall. On average, the ranking by prior scores yields a signif-

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

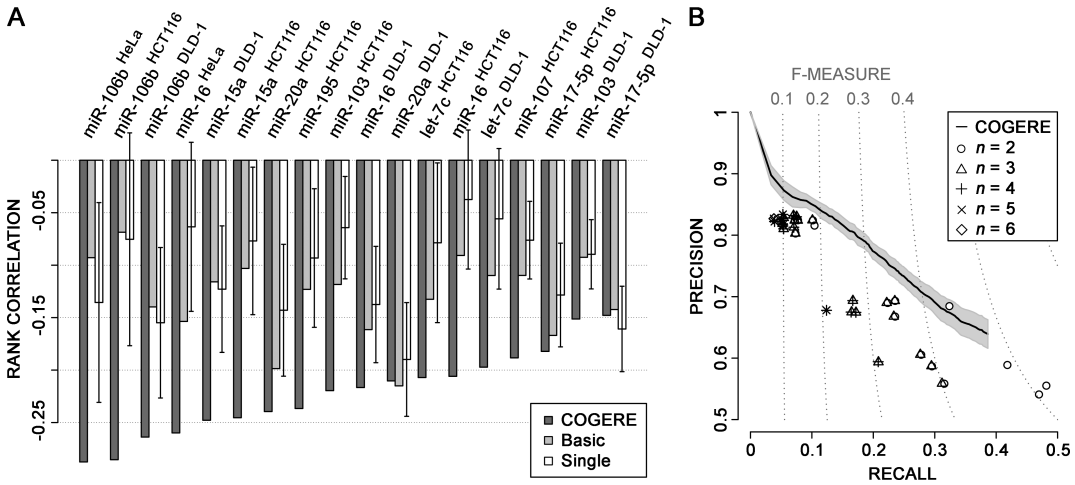


Figure 2. Evaluation of the prior score of miRNA:TG interactions. (**A**) Rank correlations (vertical bars) between predicted interaction weights and observed mRNA $\log_2$ expression changes measured post-transfection of 11 miRNAs in three cell lines (55). The lower the correlation coefficient, the better represents the scoring framework the potential of a miRNA-mediated regulation. Weighting miRNA:TG associations using the COGERE prior score outperforms the basic scoring framework applied by mirConnX (7) in 94% of cases. Both scoring frameworks improve the average performance of all single target prediction algorithms. The error bars denote the 95% confidence interval for the mean. (**B**) Mean precision–recall curve of the COGERE prior score ranking the top 20% most downregulated targets (positives) and 20% least downregulated targets (negatives) of each transfection data set. Shown are also the mean precision–recall values for all intersections of *n* miRNA target prediction algorithms. For a given recall the ranking by the prior score yields an average advantage of 7.5% points in precision compared to the simple tool intersection applied by MAGIA2 (8). The F-measure denotes the harmonic mean between precision and recall. The shaded area indicates the 95% confidence interval for the mean.

icant advantage of 7.5% points in precision (paired signed rank test $P = 8.8 \times 10^{-11}$) compared to any tool intersection. Interestingly, the intersection method is not straightforward and thus does not assure a gain of precision for a higher number of intersected tools on the expense of recall.

Advanced inference of condition-specific interactions

We generated an *in silico* benchmark set (80 networks of size 500 nodes with corresponding steady-state expression data) according to the framework used in the Dialogue for Reverse Engineering Assessments and Methods (DREAM) competition (56). This allows us to test COGERE against a known ground truth and to compare it to the common approaches mirConnX (7) and MAGIA2 (8). To measure prediction accuracy, we used the area under the precision–recall curve ($AUPR$) and the area under the $\text{precision}_{\text{sign}}$–recall curve ($AUP_{\text{sign}}R$). At this, recall describes the fraction of predicted condition-specific interactions defined by the gold standard, precision denotes the proportion of true condition-specific predictions in the result set and $\text{precision}_{\text{sign}}$ the fraction of correctly predicted regulatory signs. We computed the performance advancement of each algorithm over the null model (random guessing), denoted as $\Delta AUPR$ and $\Delta AUP_{\text{sign}}R$, respectively.

First, we evaluated how well the algorithms infer condition-specific edges from the expression data. Figure 3A shows that all tested algorithms perform better than random guessing predicting the whole condition-specific model ($\Delta AUPR > 0$ for RG:TG). Here, COGERE (median $\Delta AUPR = 0.294$) exhibits a significantly higher accuracy (Mann–Whitney U-test $P < 4 \times 10^{-15}$) than mirConnX (median $\Delta AUPR = 0.079$) and MAGIA2 (median $\Delta AUPR = 0.054$). COGERE achieves major overall improvements for the prediction of TF:TG as well as miRNA:TG interactions compared to existing tools. The major drawback of mirConnX and MAGIA2 is the low accuracy in predicting transcriptional regulation; both tools have their strength in detecting post-transcriptional regulation by miRNAs ($\Delta AUPR$ TF:TG < $\Delta AUPR$ miRNA:TG).

Second, we investigated the accuracy of predicted signs of the regulatory interactions (Figure 3B). Again, the $AUP_{\text{sign}}R$ values obtained by the tools were higher than the values obtained by the null model, whereas COGERE ($\Delta AUP_{\text{sign}}R = 0.456$) substantially outperforms mirConnX ($\Delta AUP_{\text{sign}}R = 0.127$) and MAGIA2 ($\Delta AUP_{\text{sign}}R = 0.073$). Apparently, COGERE precisely determines the signs for both kinds of regulatory interaction for all values of recall. mirConnX and MAGIA2 exhibit similar lower accuracy profiles for TF:TG interactions compared to miRNA regulatory associations, whereas the $\Delta AUP_{\text{sign}}R$ miRNA:TG values obtained by mirConnX are significantly higher than the values of MAGIA2 (Mann–Whitney U-test $P = 2.2 \times 10^{-12}$).

Case study: human cancer GRN

mRNA and miRNA profiles from tumor samples are frequently published. Having only been used to extract tumor-classifying molecular signatures (57) or confirming predicted miRNA:TG interactions (59), these expression data sets contain more information to be exploited. We computed the condition-specific relevance of regulatory interactions for the NCI-60 data panel which involves 60 cell lines originating from prostate cancer, lung cancer, breast cancer, melanoma, ovarian cancer, hematologic cancer, kidney cancer, colorectal cancer and malignant glioma. We considered the top 10% predictions by COGERE (Table 1) as highly relevant tumor specific interactions and will refer this network in the following as the cancer GRN. The resultant

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

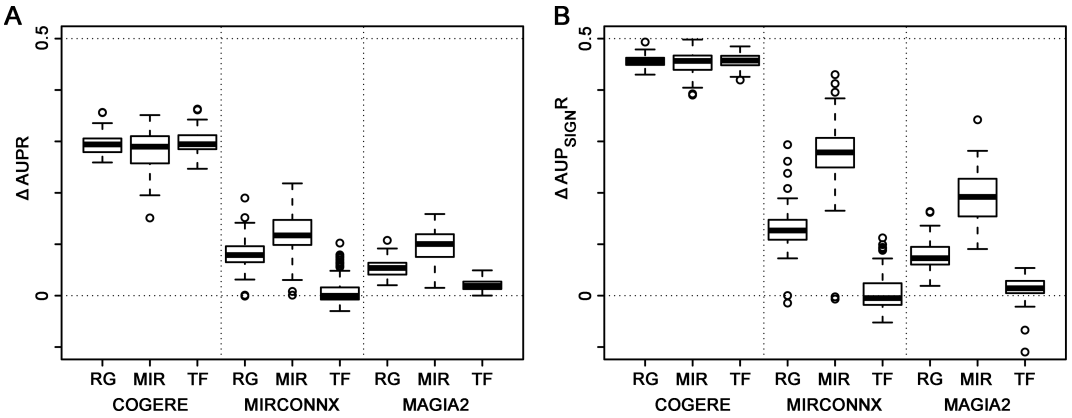


Figure 3. Accuracy of predicted condition-specific regulation. (**A**) *AUPR* values for each inference method for predicting condition-specific interactions. Shown is the deviation Δ from the null model (random guessing). COGERE outperforms mirConnX (7) and MAGIA2 (8) on the prediction of condition-specific gene regulation tested against TF- and miRNA-mediated regulation (RG), only miRNA-mediated regulation (MIR) and only transcriptional regulation (TF). (**B**) $AUP_{sign}R$ values for each inference method for predicting the condition-specific sign of an interaction. Shown is the deviation Δ from the null model (random guessing). We computed $precision_{sign}$–recall curves to determine the fraction of correctly predicted condition-specific regulatory signs for each value of recall. COGERE holds an excellent accuracy tested against TF- and miRNA-mediated regulation (RG), only miRNA-mediated regulation (MIR) and only transcriptional regulation (TF). mirConnX and MAGIA2 exhibit low accuracy for transcriptional regulation.

miRNA-mediated GRN enables the systematic analysis of gene regulation in human cancers and demonstrates the potential of COGERE to reveal conditional regulatory landscapes.

The inferred GRN discovers causal RGs in cancer. To investigate whether the genes contained in the predicted GRN were substantially related to the condition of cancer, we extracted 2760 known gene–cancer associations from HuGE-Navigator (60) for all cancer cell lines contained in the NCI-60 data. Altogether, 2477 cancer-related genes were measured by the NCI-60 microarrays, of which 1192 were contained in the inferred GRN. This denotes a significant enrichment of cancer-related genes (odds ratio = 1.2, Fisher test $P = 2.1 \times 10^{-7}$), consistent with the expectation that the inferred GRN should hold a higher fraction of cancer-related genes than expected by chance. Further, cancer-related TFs with at least one TG were significantly over-represented (odds ratio = 2.2, Fisher test $P = 4.1 \times 10^{-10}$). Five of the 10 most highly connected TFs (ELF3, EHF, ETS2, ETV5 and KLF6) were known to play a role in carcinogenesis. We examined the enrichment by using all 518 genes listed in the cancer Gene Census database (61). Again, the GRN shows a significant high content of cancer-related genes (odds ratio = 1.4, Fisher test $P = 7.7 \times 10^{-6}$) and regulatory TFs (odds ratio = 2.1, Fisher test $P = 2.2 \times 10^{-6}$) even without filtering the database for NCI-60 tumors. This result suggests that the inferred GRN can be a valuable resource to extract information regarding cancer-specific gene regulation in general.

Next, we were interested to know whether the human cancer GRN was able to recapitulate miRNAs that are both, namely dysregulated in malignant cells and at the same time causally linked to specific oncogenic processes. We compared the miRNAs contained in the GRN to entries in PhenomiR (62), a manually curated database of miRNAs that are dysregulated in diseases including the nine cancers of the NCI-60 panel. We used the Disease Ontology resource (63) to manually map the NCI-60 cell lines to PhenomiR diseases (Supplementary Table S2). Remarkably, a highly significant enrichment of known dysregulated miRNAs was observed: 164 miRNAs in the inferred GRN were previously shown to be dysregulated in tumors of the NCI-60 data set (odds ratio = 6.0, Fisher test $P = 4.5 \times 10^{-12}$; Supplementary Table S3). To investigate whether the dysregulated miRNAs contained in the human cancer GRN are also known to hold a causal influence on cancer phenotypes, we manually mapped the causal relationships annotated in mirR2Disease to PhenomiR. It was striking that 48% of the miRNAs in the GRN that were known to be dysregulated were also annotated to causally affect cancer phenotypes (odds ratio = 1.7, Fisher test $P = 4.3 \times 10^{-3}$). Among the top 10 of the most highly connected miRNAs, all were known to be dysregulated and seven were assigned a known causal relationship (mir-27a, mir-23a, mir-17, mir-21, mir-29a, mir-20a and let-7b); among the top 25, all were dysregulated and 80% causal (Supplementary Table S4). In general, the higher the number of predicted condition-specific targets by COGERE, the higher the probability that a miRNA exhibits a causal relationship to cancer (Figure 4A); e.g. of the 5% of miRNAs with the highest number of targets 67% were causal, whereas for the 5% of miRNAs with the lowest number of regulatory interactions no causal relationship was known.

RGs associated with the hallmarks of cancer. The fact that a miRNA or a TF is contained in the inferred cancer GRN does not implicate that this RG plays a role in key oncogenic processes. Hanahan and Weinberg (64,65) proposed 10 traits of cancer that govern the transformation of normal cells to tumor cells. We used the set of Gene ontology (66) biological process terms representing the 10 hallmarks of cancer prepared by Plaisier *et al.* (67) to analyze TGs for functional enrichment. We found 1393 genes involved in key oncologic processes in our cancer GRN which denotes a highly significant over-representation (odds ratio = 1.3, Fisher test $P = 6.6 \times 10^{-11}$). Next, we were interested to know which RGs in detail interact with these

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

Table 1. Network characteristics of the human cancer GRN

NCI-60 GRN	RG = TF	RG = miRNA	TG	Interactions (prior = 1, reference)	Maximum *P*
Full	473	251	8853	634 863 (4%, 5%)	0.67
Study	387	180	5869	63 486 (5%, 7%)	$<10^{-5}$

Listed are the network statistics for the full inferred GRN and the subnetwork used in this study: the number of RGs and TGs, the number of their interactions and the highest predicted *P* of a condition-specific interaction; prior = 1 denotes the fraction of interactions with a prior score of 1 and reference denotes the proportion of interactions with a reference.

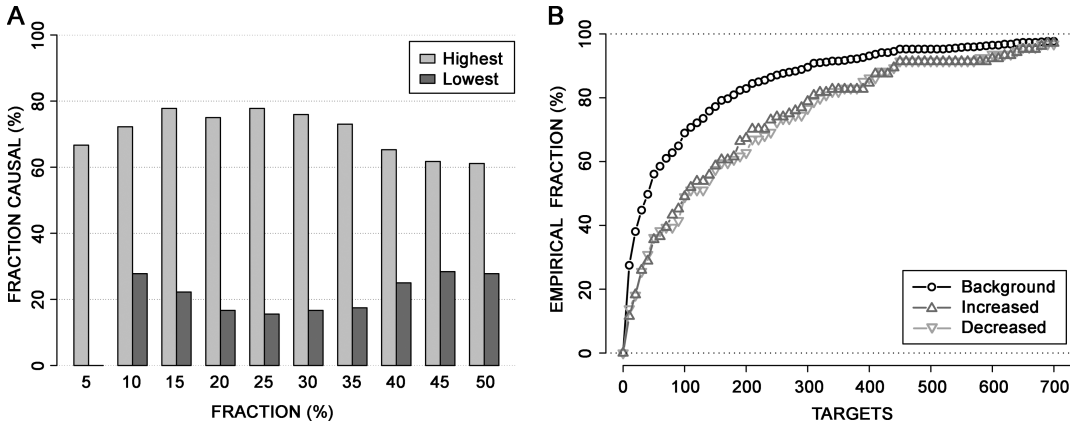


Figure 4. Degree distributions of the cancer GRN. (**A**) Fraction of miRNAs that have been reported to be causally linked to specific oncogenic processes (y-axis) for each fraction of miRNAs with the highest (light gray) or lowest (dark gray) number of targets in the cancer GRN (x-axis). miRNAs with a high number of predicted cancer-specific TGs were more often reported to be causal than miRNAs with a low number of predicted cancer-specific TGs, e.g. 78% of the top 15% of miRNAs with a high out-degree have a causal role in cancer compared to only 22% of the bottom 15% of miRNAs with a low out-degree. (**B**) Empirical out-degree distributions of all RGs. In average, RGs with a predicted association with an altered chemosensitivity of cancer cells exhibited 75 (increased drug response) or 81 (decreased drug response) more targets than any RG contained in the cancer GRN (background).

genes and are subsequently associated with the hallmarks of cancer. The functional enrichment analysis of the target sets of each RG recovered 31 miRNAs and 85 TFs that were predicted to regulate at least one process in oncogenesis (FDR adjusted Fisher test $P < 0.05$; Supplementary Table S5). Notably, 10 TFs and nine miRNAs were associated with at least five hallmarks of cancer (E2F2, ELF1, FLI1, JUN, KLF2, KLF4, KLF6, KLF10, NFKB2, TFDP2, mir-7-1, mir-18b, mir-21, mir-23a, mir-23b, mir-24-1, mir-24-2, mir-29a and mir-181b-1) suggesting that these genes are promising candidates for follow-up studies. Further, we observed the evasion of inhibition mechanisms blocking proliferation and the metastatic potential of a cell to be under strong control. Together 100 RGs (71 TFs and 29 miRNAs) were predicted to regulate 'insensitivity to antigrowth signals' followed by 97 RGs (71 TFs and 26 miRNAs) associated with 'tissue invasion and metastasis'. The latter hallmark is one of the defining features of malignant tumors making putative regulators excellent biomarker candidates. COGERE proposes a mechanistic explanation of how TFs and miRNAs act together to directly regulate genes involved in metastatic processes (Figure 5 and Supplementary Table S6). In this example, our cancer model suggests that mir-20a, a member of the miR-17∼92 miRNA cluster, regulates cell adhesion and cell migration in tumor metastasis through direct suppression of thrombospondin 1 (THBS1). An upregulation of miR-17∼92 was described to promote angiogenesis and tumor growth (68), whereas increased THBS1 expression suppresses growth or metastasis of some tumors *in vivo* and inhibits angiogenesis (69). The THBS1 downregulation was observed primarily at the level of mRNA turnover (70) which is probably induced by miRNA-mediated mRNA degradation. These findings return a predicted cancer-specific interaction as an interesting subject for further investigations.

The cancer GRN predicts potential targets for cancer pharmacology. Given a condition-specific GRN, a key next step for the extraction of novel testable hypotheses is the integration of orthogonal information. Drug insensitivity or drug resistance are major obstacles in the successful treatment of cancer. Several studies suggested that robustly positive or negative correlations between drug activity and gene expression reflect a role in chemosensitivity of cancer cells. A negative correlation may indicate that cancer cells with an increased expression level of mRNA or miRNA are less sensitive to the drug compound than other cells. On the contrary, if the correlation is positive, co-treatment with mRNA or miRNA might be used to enhance drug potency or reduce toxicity (57,71). We calculated the correlation of miRNA and mRNA expression profiles versus drug activities over all NCI-60 cancer cell lines. First, we validated the informative value of the correlation coefficients by comparing our results to GI_{50} values measuring the growth inhibitory power of the test agent provided by Blower *et al.* (71). They experimentally tested the activity pattern of 10 drugs following either inhibitor or precursor transfection of three miRNAs (let-7, mir-16 and mir-21) in A549 cell lines. The correlation coefficients were in good agreement with the average $\log_{10}$ fold-changes of GI_{50} values between

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

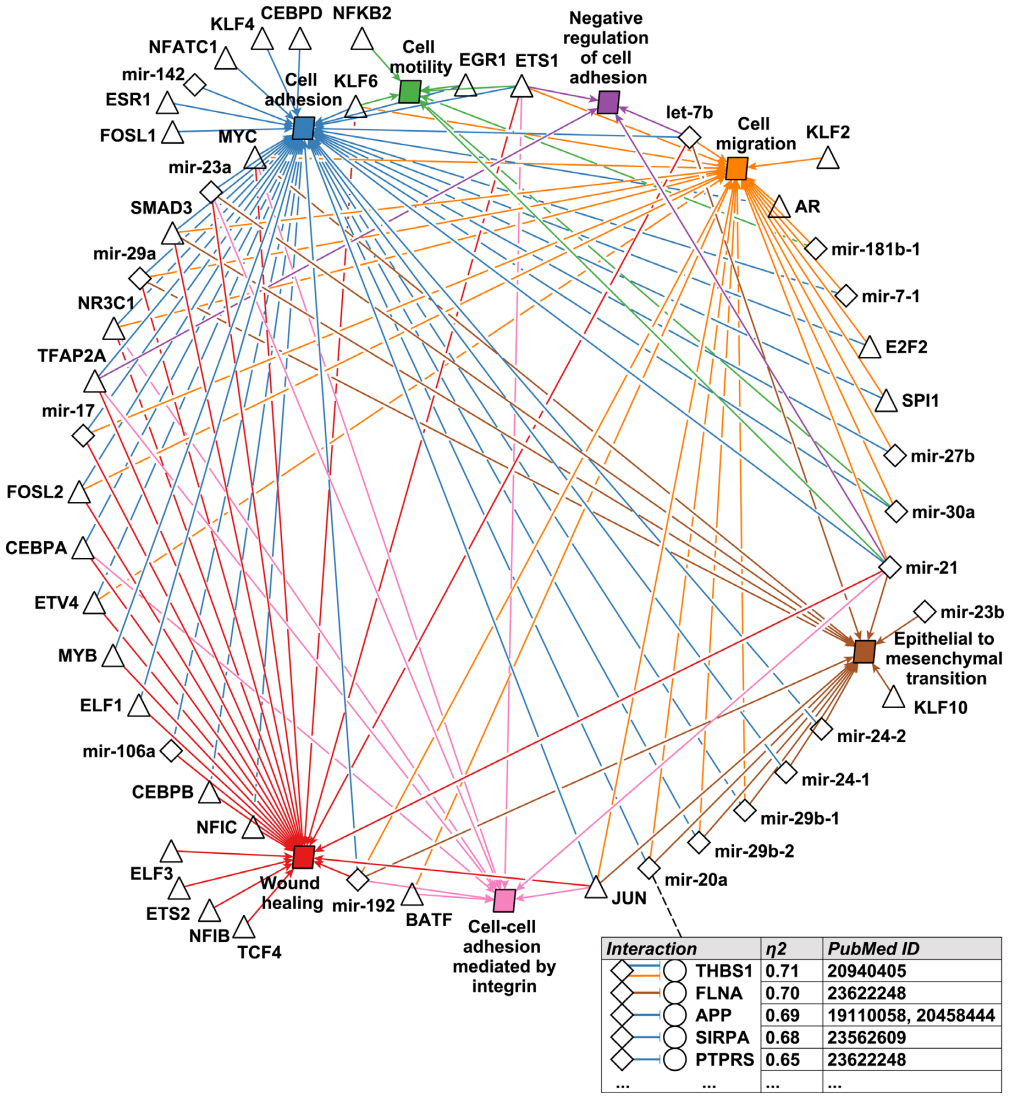


Figure 5. Metastatic interplay of TFs and miRNAs. Nodes are biological processes (colored parallelogram), TFs (triangle) and miRNAs (diamond). Arcs denote an enrichment of RG targets in a metastatic process and are colored, respectively. The top five predicted negative regulations of mir-20a are listed exemplary in the table shown at the lower right corner; e.g. the THBS1 repression by mir-20a which was described in (68) and holds a condition-specific regulation score of 0.71. This interaction affects cell adhesion and cell migration (blue and orange arcs). Note that the shown network was filtered by regulatory interactions having at least one literature reference (PubMed ID).

lowered and raised miRNA levels ($R^2 = 0.38$, $P = 2.7 \times 10^{-4}$; Supplementary Figure S3).

To gain a first broad perspective on the potential roles of the predicted RGs in cancer therapy, we analyzed the associations of 163 anti-cancer compounds that are in clinical trial or were approved by the FDA (U.S. Food and Drug Administration) and all genes contained in our cancer GRN. We observed 45 miRNAs and 125 TFs accounting for 105 drug–miRNA and 309 drug–TF correlations reaching the α-level of $P = 10^{-4}$ suggested by Blower *et al.* (71). This denotes a significant amount of potential drug targets (miRNA odds ratio = 2.0, Fisher test $P = 9.6 \times 10^{-3}$; TF odds ratio = 2.9, Fisher's $P = 1.9 \times 10^{-16}$). Among these, 23 miRNAs and 71 TFs were predicted to decrease the cancer cells' chemosensitivity. This set of chemoresistance factors exhibited on average 1.7 times more targets (factor 2.9 for miRNAs and factor 1.5 for TFs) than any RG contained in the whole GRN (Mann–Whitney U-test $P = 6.7 \times 10^{-6}$; Figure 4B). For example, mir-22 was predicted with the highest amount of negative effects to compound potencies; it had the third most regulatory interactions in the cancer GRN. The aberrant expression of this oncogene has been reported to correlate with poor survival (72) and our results indicate that tumor cells expressing mir-22 are less sensitive to drug treatment. Based on its high number of targets, mir-22 might be an interesting subject for further assessments of its role in resistance to anticancer agents. It remains to be evaluated if mir-22 is suitable as a prognostic biomarker. However, if mir-22 plays a causal role in drug resistance its inhibition may enhance the response of malignant cells to cancer drug treatment.

Further, we observed 25 miRNAs and 79 TFs that exhibited a positive correlation coefficient and thus were assumed

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

to increase the susceptibility of NCI-60 cells to the action of at least one cancer drug. Interestingly, we found the proto-oncogene MYC as the RG which was predicted to positively affect the potency of the highest number of compounds. This TF is constitutively expressed in many cancers causing augmentation of cell proliferation (73). To investigate whether this TF plays a substantial role in chemosensitivity, we extracted all positive correlated drug–gene associations composed of the 591 predicted MYC targets and the eight MYC-affected compounds (Supplementary Figure S4). The expression of the MYC targets POLG2, CAMKV, VASH2 and OGFOD2 in cancer cells was predicted to increase the potency of oxaliplatin. Active derivatives of this compound form both inter- and intra-strand DNA crosslinks resulting in inhibition of DNA replication and transcription and cell-cycle nonspecific cytotoxicity. POLG2 polymerase promotes DNA synthesis. Oxaliplatin has been described to induce lesions in the human MYC gene (74). Cancer treatment with oxaliplatin might reduce the positive cancer-specific regulation of POLG2 by MYC, which in turn might cause an induced inhibitory effect on DNA synthesis resulting in an enhanced cytotoxic effect of this compound. In addition VASH2 is involved in positive regulation of angiogenesis, a typical process taking place in cancer cells. Loss of induced regulation of this gene might induce a secondary anti-cancer effect. Further, we found two compounds that lower estrogen levels: calusterone and dromostanolone propionate. It has been proposed that the human MYC gen-regulatory region embeds an estrogen-responsive *cis*-acting element (75) inducing rapid MYC expression in the presence of estrogen. Further, estrogen repletion is accompanied by significant reduction in leukocyte adhesion (76). The MYC target ICAM3 was predicted to increase the susceptibility of cancer cells to the action of both anti-estrogen compounds. This gene is a member of the intercellular adhesion molecule family and has been reported to induce cancer cell proliferation, cellular radio-resistance, cancer cell migration and invasion (77). Based on the COGERE predictions we can hypothesize that the reduction in estrogen might reduce MYC expression resulting in reduced ICAM3 function resulting in increased drug potency. Another interesting compound for further investigations might be imexon, a 2-cyanoaziridine derivate with antitumor activity, which was predicted to be positively affected by the highest number of MYC targets. These 27 TGs contained among others BCL2, a well-known oncogene encoding an anti-apoptotic protein.

DISCUSSION

The experimentalist is confronted with large data sets of high dimensionality reflecting the interplay of thousands of cellular components. Therefore, it is an imperative computational challenge to develop predictive and actionable models to investigate functionality as well as spatial and temporal behavior of these components. As the availability of experimental evidence in databases and the biomedical literature sharply increase, the systemic integration of existing knowledge to support the analysis of genome-wide molecular expression signatures of complex diseases becomes a bare requirement.

Firstly, we presented a method for the graph-oriented integration of several millions of annotated, literature-mined as well as pure sequence-based miRNA:TG and TF:TG interactions to a uniform scoring framework (prior score) of prior knowledge for human and mouse. We have illustrated that our integrated model comprehensively covers current knowledge provided by common experimental databases, the biomedical literature and computational predictions. The presented comparison to existing attempts reveals that the COGERE prior score constitutes a major improvement in the task of weighting miRNA regulation by their feasible regulatory effect on a TG. A basic combination of multiple prediction tools as conducted by mirConnX (7) performs better than a blindfolded random selection of any individual algorithm. Compared to a sighted systematic selection this scoring scheme performs effectively worse than several individual tools (Supplementary Figure S5). In contrast, the COGERE prior score improves the accuracy in 78% of all transfection experiments (median rank 1) directly compared to any of the six integrated target prediction algorithms. Further, priors based on the COGERE scoring framework exhibit effectively more accurate information than a simple intersection of tools as used by MAGIA2 (8). Our evaluation shows that a basic intersection of tools also implies a strong limitation in usability: it remains unclear to the user which tool combination fits best his requirements regarding recall and precision. Despite the current success of the COGERE prior score, ongoing progress in data collection by high-throughput '-omics' techniques will further improve the prior knowledge.

Secondly, to detect condition-specific regulation from mRNA and miRNA expression data, COGERE scores the relevance of prior interactions by measuring the mutual dependency between a RG and its TG. By applying an ANOVA we derived the non-parametric and nonlinear correlation coefficient η^2 and its corresponding FDR adjusted P. Here, neither a discretization of the expression data nor a setup with matching samples is required, increasing the robustness of COGERE. We showed that COGERE strongly outperforms existing approaches in predicting condition-specific GRNs from synthetic expression data and holds an excellent performance for predicting the regulatory sign of an interaction. The presented analysis denotes, in addition, a comparative evaluation of MAGIA2 and mirConnX performance for the first time.

COGERE is capable to infer GRNs from unmatched data implying two advantages: (i) expression data can be obtained from different studies/measurements with identical experimental setups, (ii) detection of signals in at least a subset of experiments increases the robustness of the method against noise. COGERE balances the gene expression data by a condition-specific and individual-independent filtering of microarrays. The discriminatory power of the inference is sharpened as the variation within the conditions (technical variation) is reduced, whereas the differences between the conditions (biological variation) become more pronounced. We observed increased robustness of accuracy to detect context-specific effects due to differential TF- or miRNA-mediated regulation in a benchmark with noisy expression data (Supplementary Figures S2B and S6).

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

We remind that the performance assessment is based on simulated data. The *in silico* benchmark set is based on subnetworks from a human GRN with known interactions and thus holds similar types of structural properties and regulatory dynamics as realized in biological gene networks. The evaluation represents a simplified model of gene regulation. An *in silico* benchmark does not replace the careful evaluation *in vivo*, but enables a systematically and efficiently performance validation and comparison of prediction methods over multiple networks. Unfortunately, to date an elaborate *in vivo* data set composed of mRNA and miRNA expression for several conditions as well as the corresponding experimentally verified condition-specific GRN is not available for human or mouse. It is likely that methods that do not perform well in an *in silico* benchmark will perform even worse with real biological data (78). In contrast to artificial data, linear correlation between a RG and a TG is a weak indicator of true condition-specific regulatory relationships in real expression measurements. This assumption is supported by a recent comprehensive and comparative evaluation of inference methods rating a two-way ANOVA-based approach best on the prediction of real GRNs from *Escherichia coli* and *Saccharomyces cerevisiae* expression data (53).

We used the NCI-60 cancer expression study to show that COGERE is a valuable resource to promote hypothesis-driven clinical research. We were able to demonstrate that the GRN inferred by COGERE captured disease-relevant regulation of cancer. A significant reliable proportion of known cancer-related genes and miRNAs were found in the predicted network. At this, causal miRNAs exhibited a higher number of condition-specific targets mirroring their central role in cancerogenesis. We identified a relatively small subset of RGs that play a role in multiple oncogenic processes in cancer. By using the inferred GRN, we provided a mechanistic insight into the TF and miRNA interplay during the regulation of metastatic processes. Since many somatic passenger mutations may also alter expression profiles, we do not expect that all condition-specific correlations are necessarily related to cancer driving processes.

Our results suggest that the GRN contains novel, testable and interesting hypotheses regarding cancer-specific regulation beyond what is documented in existing databases. Moreover, the network predicted TFs and miRNAs that play a role in the chemosensitivity to approved cancer drugs and made novel predictions regarding the role of 116 RGs mediating the expression of genes associated with oncogenic processes. A predicted strong drug–gene relation may indicate a causal role in drug response (57,71). If such a relationship proves to be causal, it could be exploited to improve cancer therapy. We showed that condition-specific GRN information inferred by COGERE enables the analysis of potential drug targets in the context of gene regulation. Based on our observations, we suggest that the predicted GRN contains several hypotheses promoting cancer pharmacogenomics.

In summary, we introduced COGERE, a novel, generalizable approach that boosts signal to noise for the modeling of large-scale condition-specific regulatory landscapes in any cellular contexts. COGERE implements a robust inference method together with a concept of high-level data integration. It features the capacity of rational interpretation of expression signals in very large data sets in the context of known regulatory relations driving the discovery of new biology.

AVAILABILITY

The application to infer GRNs from expression data is freely available for academic use under http://mips.helmholtz-muenchen.de/cogere. Furthermore, to facilitate reader access and usability we provide all data contained in the NCI-60 cancer GRN: regulatory interactions and gene associations with approved compounds. We hope to provide experimentalists with a tool to infer GRNs for their condition of interest and cancer researchers with a valuable resource to explore the cancer-specific GRN.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We thank Gregor Sturm (Technische Universität München) for his technical support. The contribution of all anonymous reviewers to the improvement of the manuscript is gratefully acknowledged.

FUNDING

German Center for Diabetes Research [DZD e.V. to J.F.L.]. Funding for open access charge: Helmholtz Zentrum München (German Research Center for Environmental Health).
Conflict of interest statement. None declared.

REFERENCES

1. Benayoun,B.A., Caburet,S. and Veitia,R.A. (2011) Forkhead transcription factors: key players in health and disease. *Trends Genet.*, **27**, 224–232.
2. Mendell,J.T. and Olson,E.N. (2012) MicroRNAs in stress signaling and human disease. *Cell*, **148**, 1172–1187.
3. Keene,J.D. (2007) RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.*, **8**, 533–543.
4. Pillai,R.S. (2005) MicroRNA function: multiple mechanisms for a tiny RNA? *RNA*, **11**, 1753–1761.
5. Reczko,M., Maragkakis,M., Alexiou,P., Grosse,I. and Hatzigeorgiou,A.G. (2012) Functional microRNA targets in protein coding sequences. *Bioinformatics*, **28**, 771–776.
6. Shalgi,R., Lieber,D., Oren,M. and Pilpel,Y. (2007) Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS Comput. Biol.*, **3**, e131.
7. Huang,G.T., Athanassiou,C. and Benos,P.V. (2011) mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic Acids Res.*, **39**, W416–W423.
8. Bisognin,A., Sales,G., Coppe,A., Bortoluzzi,S. and Romualdi,C. (2012) MAGIA^2: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). *Nucleic Acids Res.*, **40**, W13–W21.
9. Margolin,A.A., Nemenman,I., Basso,K., Wiggins,C., Stolovitzky,G., Dalla Favera,R. and Califano,A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**(Suppl. 1), S7, 16723010.

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

10. Weinstein,J.N. (2006) Spotlight on molecular profiling: "Integromic" analysis of the NCI-60 cancer cell lines. *Mol. Cancer Ther.*, **5**, 2601–2605.
11. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
12. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
13. Fujita,S. and Iba,H. (2008) Putative promoter regions of miRNA genes involved in evolutionarily conserved regulatory systems among vertebrates. *Bioinformatics*, **24**, 303–308.
14. Zhou,X., Ruan,J., Wang,G. and Zhang,W. (2007) Characterization and identification of microRNA core promoters in four model species. *PLoS Comput. Biol.*, **3**, e37.
15. Wang,X., Xuan,Z., Zhao,X., Li,Y. and Zhang,M.Q. (2009) High-resolution human core-promoter prediction with CoreBoost_HM. *Genome Res.*, **19**, 266–275.
16. Corcoran,D.L., Pandit,K.V., Gordon,B., Bhattacharjee,A., Kaminski,N. and Benos,P.V. (2009) Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. *PLoS One*, **4**, e5279.
17. Marson,A., Levine,S.S., Cole,M.F., Frampton,G.M., Brambrink,T., Johnstone,S., Guenther,M.G., Johnston,W.K., Wernig,M., Newman,J. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.
18. Ozsolak,F., Poling,L.L., Wang,Z., Liu,H., Liu,X.S., Roeder,R.G., Zhang,X., Song,J.S. and Fisher,D.E. (2008) Chromatin structure analyses identify miRNA promoters. *Genes Dev.*, **22**, 3172–3183.
19. Chien,C.-H., Sun,Y.-M., Chang,W.-C., Chiang-Hsieh,P.-Y., Lee,T.-Y., Tsai,W.-C., Horng,J.-T., Tsou,A.-P. and Huang,H.-D. (2011) Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. *Nucleic Acids Res.*, **39**, 9345–9356.
20. Down,T.A. and Hubbard,T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, **12**, 458–461.
21. Rodriguez,A., Griffiths-Jones,S., Ashurst,J.L. and Bradley,A. (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res.*, **14**, 1902–1910.
22. Baskerville,S. and Bartel,D.P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, **11**, 241–247.
23. Monteys,A.M., Spengler,R.M., Wan,J., Tecedor,L., Lennox,K.A., Xing,Y. and Davidson,B.L. (2010) Structure and activity of putative intronic miRNA promoters. *RNA*, **16**, 495–505.
24. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
25. Cartharius,K., Frech,K., Grote,K., Klocke,B., Haltmeier,M., Klingenhoff,A., Frisch,M., Bayerlein,M. and Werner,T. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933–2942.
26. Klingenhoff,A., Frech,K., Quandt,K. and Werner,T. (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics*, **15**, 180–186.
27. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
28. Gerner,M., Sarafraz,F., Bergman,C.M. and Nenadic,G. (2012) BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. *Bioinformatics*, **28**, 2154–2161.
29. Alexiou,P., Maragkakis,M., Papadopoulos,G.L., Reczko,M. and Hatzigeorgiou,A.G. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, **25**, 3049–3055.
30. Ellwanger,D.C., Büttner,F.A., Mewes,H.-W. and Stümpflen,V. (2011) The sufficient minimal set of miRNA seed types. *Bioinformatics*, **27**, 1346–1350.
31. Betel,D., Koppal,A., Agius,P., Sander,C. and Leslie,C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90, 20799968.
32. Krek,A., Grün,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
33. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
34. Grimson,A., Farh,K.K.-H., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
35. Sturm,M., Hackenberg,M., Langenberger,D. and Frishman,D. (2010) TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics*, **11**, 292, 20509939.
36. Naeem,H., Küffner,R., Csaba,G. and Zimmer,R. (2010) miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics*, **11**, 135, 20233441.
37. Dweep,H., Sticht,C., Pandey,P. and Gretz,N. (2011) miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J. Biomed. Inform.*, **44**, 839–847.
38. Yang,J.-H., Li,J.-H., Shao,P., Zhou,H., Chen,Y.-Q. and Qu,L.-H. (2011) starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res.*, **39**, D202–D209.
39. Selbach,M., Schwanhäusser,B., Thierfelder,N., Fang,Z., Khanin,R. and Rajewsky,N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.
40. Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.
41. Gerstein,M.B., Kundaje,A., Hariharan,M., Landt,S.G., Yan,K.K., Cheng,C., Mu,X.J., Khurana,E., Rozowsky,J., Alexander,R. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.
42. Jiang,C., Xuan,Z., Zhao,F. and Zhang,M.Q. (2007) TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res.*, **35**, D137–D140.
43. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
44. Wang,J., Lu,M., Qiu,C. and Cui,Q. (2010) TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res.*, **38**, D119–D122.
45. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
46. Hsu,S.D., Tseng,Y.T., Shrestha,S., Lin,Y.L., Khaleel,A., Chou,C.H., Chu,C.F., Huang,H.Y., Lin,C.M., Ho,S.Y. *et al.* (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.*, **42**, D78–D85.
47. Jiang,Q., Wang,Y., Hao,Y., Juan,L., Teng,M., Zhang,X., Li,M., Wang,G. and Liu,Y. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.
48. Vergoulis,T., Vlachos,I.S., Alexiou,P., Georgakilas,G., Maragkakis,M., Reczko,M., Gerangelos,S., Koziris,N., Dalamagas,T. and Hatzigeorgiou,A.G. (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res.*, **40**, D222–D229.
49. Kauffmann,A., Gentleman,R. and Huber,W. (2009) arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**, 415–416.
50. Hackstadt,A.J. and Hess,A.M. (2009) Filtering for increased power for microarray data analysis. *BMC Bioinformatics*, **10**, 11, 19133141.
51. Cohen,J. (1973) Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educ. Psychol. Meas.*, **33**, 107–112.
52. Miller,R.G. (1997) *Beyond ANOVA: Basics of Applied Statistics*. Chapman & Hall/CRC press, Boca Raton, Florida.
53. Küffner,R., Petri,T., Tavakkolkhah,P., Windhager,L. and Zimmer,R. (2012) Inferring gene regulatory networks by ANOVA. *Bioinformatics*, **28**, 1376–1382.

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014

54. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
55. Linsley,P.S., Schelter,J., Burchard,J., Kibukawa,M., Martin,M.M., Bartz,S.R., Johnson,J.M., Cummins,J.M., Raymond,C.K., Dai,H. *et al.* (2007) Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol. Cell. Biol.*, **27**, 2240–2252.
56. Schaffter,T., Marbach,D. and Floreano,D. (2011) GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, **27**, 2263–2270.
57. Liu,H., D'Andrade,P., Fulmer-Smentek,S., Lorenzi,P., Kohn,K.W., Weinstein,J.N., Pommier,Y. and Reinhold,W.C. (2010) mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Mol. Cancer Ther.*, **9**, 1080–1091.
58. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
59. Wang,Y.P. and Li,K.B. (2009) Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC Genomics*, **10**, 218, 19435500.
60. Yu,W., Gwinn,M., Clyne,M., Yesupriya,A. and Khoury,M.J. (2008) A navigator for human genome epidemiology. *Nat. Genet.*, **40**, 124–125.
61. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
62. Ruepp,A., Kowarsch,A. and Theis,F. (2012) PhenomiR: microRNAs in human diseases and biological processes. *Methods Mol. Biol.*, **822**, 249–260.
63. Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.W., Mazaitis,M., Felix,V., Feng,G. and Kibbe,W.A. (2012) Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.
64. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
65. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
66. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
67. Plaisier,C.L., Pan,M. and Baliga,N.S. (2012) A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. *Genome Res.*, **22**, 2302–2314.
68. Dews,M., Fox,J.L., Hultine,S., Sundaram,P., Wang,W., Liu,Y.Y., Furth,E., Enders,G.H., El-Deiry,W., Schelter,J.M. *et al.* (2010) The myc-miR-17∼92 axis blunts TGF{beta} signaling and production of multiple TGF{beta}-dependent antiangiogenic factors. *Cancer Res.*, **70**, 8233–8246.
69. Roberts,D.D. (1996) Regulation of tumor growth and metastasis by thrombospondin-1. *FASEB J.*, **10**, 1183–1191.
70. Janz,A., Sevignani,C., Kenyon,K., Ngo,C.V. and Thomas-Tikhonenko,A. (2000) Activation of the myc oncoprotein leads to increased turnover of thrombospondin-1 mRNA. *Nucleic Acids Res.*, **28**, 2268–2275.
71. Blower,P.E., Chung,J.H., Verducci,J.S., Lin,S., Park,J.K., Dai,Z., Liu,C.G., Schmittgen,T.D., Reinhold,W.C., Croce,C.M. *et al.* (2008) MicroRNAs modulate the chemosensitivity of tumor cells. *Mol. Cancer Ther.*, **7**, 1–9.
72. Song,S.J., Ito,K., Ala,U., Kats,L., Webster,K., Sun,S.M., Jongen-Lavrencic,M., Manova-Todorova,K., Teruya-Feldstein,J., Avigan,D.E. *et al.* (2013) The oncogenic microRNA miR-22 targets the TET2 tumor suppressor to promote hematopoietic stem cell self-renewal and transformation. *Cell Stem Cell*, **13**, 87–101.
73. Dang,C.V. (2013) MYC, metabolism, cell growth, and tumorigenesis. *Cold Spring Harb. Perspect. Med.*, **3**, 23906881.
74. Woynarowski,J.M., Chapman,W.G., Napier,C., Herzig,M.C. and Juniewicz,P. (1998) Sequence- and region-specificity of oxaliplatin adducts in naked and cellular DNA. *Mol. Pharmacol.*, **54**, 770–777.
75. Dubik,D. and Shiu,R.P. (1992) Mechanism of estrogen activation of c-myc oncogene expression. *Oncogene*, **7**, 1587–1594.
76. Santizo,R. and Pelligrino,D.A. (1999) Estrogen reduces leukocyte adhesion in the cerebral circulation of female rats. *J. Cereb. Blood Flow Metab.*, **19**, 1061–1065.
77. Park,J.K., Park,S.H., So,K., Bae,I.H., Yoo,Y.D. and Um,H.D. (2010) ICAM-3 enhances the migratory and invasive potential of human non-small cell lung cancer cells by inducing MMP-2 and MMP-9 via Akt and CREB. *Int. J. Oncol.*, **36**, 181–192.
78. Marbach,D., Prill,R.J., Schaffter,T., Mattiussi,C., Floreano,D. and Stolovitzky,G. (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl Acad. Sci. U.S.A.*, **107**, 6286–6291.

Downloaded from http://nar.oxfordjournals.org/ at GSF Forschungszentrum on October 13, 2014