doi: 10.1093/nar/gky210

TGMI: an efficient algorithm for identifying pathway regulators through evaluation of triple-gene mutual interaction

Chathura Gunasekara[1,2], Kui Zhang[3], Wenping Deng[1], Laura Brown[4] and Hairong Wei[1,2,4,5,6,*]

[1]School of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI 49931, USA, [2]Program of Computational Science and Engineering, Michigan Technological University, MI 49931, USA, [3]Department of Mathematical Sciences Michigan Technological University, Houghton, MI 49931, USA, [4]Department of Computer Science, Michigan Technological University, MI 49931, USA, [5]Life Science and Technology Institute, Michigan Technological University, Houghton, MI 49931, USA and [6]Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing Forestry University, Beijing, China

Received December 07, 2017; Revised March 07, 2018; Editorial Decision March 10, 2018; Accepted March 12, 2018

ABSTRACT

Despite their important roles, the regulators for most metabolic pathways and biological processes remain elusive. Presently, the methods for identifying metabolic pathway and biological process regulators are intensively sought after. We developed a novel algorithm called triple-gene mutual interaction (TGMI) for identifying these regulators using high-throughput gene expression data. It first calculated the regulatory interactions among triple gene blocks (two pathway genes and one transcription factor (TF)), using conditional mutual information, and then identifies significantly interacted triple genes using a newly identified novel mutual interaction measure (MIM), which was substantiated to reflect strengths of regulatory interactions within each triple gene block. The TGMI calculated the MIM for each triple gene block and then examined its statistical significance using bootstrap. Finally, the frequencies of all TFs present in all significantly interacted triple gene blocks were calculated and ranked. We showed that the TFs with higher frequencies were usually genuine pathway regulators upon evaluating multiple pathways in plants, animals and yeast. Comparison of TGMI with several other algorithms demonstrated its higher accuracy. Therefore, TGMI will be a valuable tool that can help biologists to identify regulators of metabolic pathways and biological processes from the exploded high-throughput gene expression data in public repositories.

INTRODUCTION

It is known that most organisms have at least several hundred metabolic pathways and a multitude of biological processes, but our understanding of how these biological pathways or processes are regulated is limited. For example, *Arabidopsis thaliana* has 549 annotated metabolic pathways and a few thousand biological processes as defined with gene ontology terms, but the regulators for most of these pathways have not yet been revealed (1–3). With the advent of the whole-genome approach and the explosion of biological data in public repositories, demands have heightened for computational algorithms that can be used to predict pathway regulators using high-throughput gene expression datasets. Although some methods for building gene regulatory networks have been developed during the last decade, accurate algorithms tailored specifically for identifying pathway regulators have not been developed. Currently, methods for identifying regulatory relationships from time-series gene expression data include dynamic Bayesian networks (4–6), differential equations (7), control logic (8), Boolean networks (9), stochastic networks (10) and finite state linear models (11). These methods and algorithms are primarily suitable for time-course data generated from bacteria, yeast and some cell lines of eukaryotic organisms.

Gene expression datasets in public repositories have increased exponentially. Most are non-time series static gene expression datasets, which include both treatment versus control datasets and those that have very large time intervals of a few hours to even several days (12). During each time interval, too many biological events elapsed to abolish the attempt to perform dynamic simulation using temporal variables. A few highly efficient methods have

[*]To whom correspondence should be addressed. Tel: +1 906 487 1473; Fax: +1 906 487 2915; Email: hairong@mtu.edu

been developed to infer regulatory relationships from these types of static data, such as the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (13), the Backward Elimination Random Forest (BWERF) algorithm (14), and the Bottom-up Graphical Gaussian Model (Bottom-up GGM) algorithm (15,16). The ARACNE algorithm uses 'mutual information' to identify dependent relationships between pairwise genes and then applies a data processing inequality to remove indirect links. It can therefore be used to identify pathway regulators through analysis of transcription factor (TF)-pathway gene dependence. BWERF and Bottom-up GGM were developed and tailored for building multilayered hierarchical gene regulatory networks (ML-hGRNs) that operate above a given pathway. BWERF is based on a random forest algorithm with a recursive evaluation process to reduce the number of TFs that have greater importance values to pathway genes; this process is repeated with the newly acquired layer to be set as the new bottom layer and the rest of TFs until a multi-layered ML-hGRN is obtained. The Bottom-up GGM method also constructs a ML-hGRN using a set of pathway genes as the bottom layer and TFs as inputs for building multiple upper layers in a layer-by-layer fashion. When Bottom-up GGM method is used to evaluate the significant interference within a triple gene block, i.e. the interference of a candidate TF in the higher hierarchical layer on two pathway genes. The interference can be determined by examining whether the difference between the correlation coefficient of two pathway genes and the partial correlation coefficient of the two bottom-layer genes after removing the effect of the upper-layer TF exceeds the significance level. This difference represents the interference strength of the TF on the two bottom-layer genes.

In this study, we present a novel algorithm called triple-gene mutual interaction (TGMI) for identifying regulators of metabolic pathway and biological process from high-throughput gene expression data by evaluating triple gene blocks. The reasons that we used triple gene blocks instead of other combinations of genes in TGMI are based on two biological and one statistical facts. First, genes involved in the same pathway or biological process are tended to be coexpressed (17) or coordinated (18,19) in expression patterns. Second, genes with same or similar expression patterns are often under the regulation of the same molecular mechanism (20–22). Third, several studies also showed that the evaluation of three variables together is more powerful than the evaluation of two variables for recognizing causal relationships (43,86,87). TGMI algorithm therefore used each paired pathway genes as a 'bait' to fish the TFs that had highest regulatory interactions with two pathway genes. When applied to a gene expression dataset, it first dissected all interactions among a triple gene block into seven interactive components, based on which we developed an efficient measure to screen for the significantly interacted triple gene blocks in which paired pathway genes have dependency while TFs and paired pathway genes showed significant conditional interactions. Then, the frequencies of all TFs' presence in significantly interacted triple gene blocks were calculated and ranked to show the importance of each TF to a pathway or a biological process. The algorithm was applied to several metabolic or non-canonical pathways using microarray gene expression data from *A. thaliana* and mouse embryonic stem cells. The ranked TFs were compared with those obtained from the other three algorithms, BWERF, Bottom-up GGM and ARACNE, using the same input data. The results indicated that the TGMI algorithm was more efficient and accurate in identifying true pathway regulators than the other three algorithms.

MATERIALS AND METHODS

Arabidopsis thaliana microarray gene expression datasets

A wood formation compendium dataset (128 microarray samples) was constructed using the microarray data from *A. thaliana* hypocotyledonous stem tissue treated under short-day condition that is known to induce wood formation (23). These data are publicly available in the NCBI GEO repository (http://www.ncbi.nlm.nih.gov/geo/) with the following accession numbers: GSE607, GSE6153, GSE18985, GSE2000, GSE24781 and GSE5633. Affymetrix 25k ATH1 microarrays were used to generate all these datasets. The CEL files were downloaded and processed with the Robust Multi-Array Analysis algorithm available at https://www.bioconductor.org (24). A previously published method was used to perform quality control of the datasets (25).

Mouse microarray gene expression datasets

Pluripotency maintenance pathway was downloaded from the embryonic stem cells atlas of pluripotency evidence (ESCAPE) repository. This time-course dataset was generated using the Affymetrix MOE439A arrays from embryonic stem cells under undirected differentiation, with three replications at each of the following time points: at 0, 6, 12, 18, 24, 36 and 48 h, and also 4, 9 and 14 days. Validated regulatory relationships were obtained from ChIP-X studies available in the ESCAPE repository (http://www.maayanlab.net/ESCAPE/). The pluripotency maintenance pathway with 24 genes and 35 known regulatory TFs was used to construct three datasets with 100, 200 or 300 random noise genes for the evaluation of the TGMI algorithm.

The TGMI algorithm

The interactions between a regulatory TF and a pair of pathway genes were first evaluated by conditional mutual information plus a novel mutual interaction measure (MIM) we discovered. Given a triple gene block with a TF being represented by a variable X and a pair of pathway genes being represented by variables Y_1 and Y_2 (Figure 1A), the gene expression data for each gene were discretized by using the equal frequency discretization algorithm (26). The entropy (H) of each variable is then calculated using the following formula:

$$H(X) = -\sum_{x} p(x) \log p(x) = -E[\log(p(x))],$$

where x represents each discretized value in X and $p(x)$ is the probability mass function. H(Y_1) and H(Y_2) were calculated in the similar way.

The mutual information between each pair of variables, i.e. $(Y1,\ Y2)$, $(Y1,\ X)$ and $(Y2,\ X)$), was calculated using the

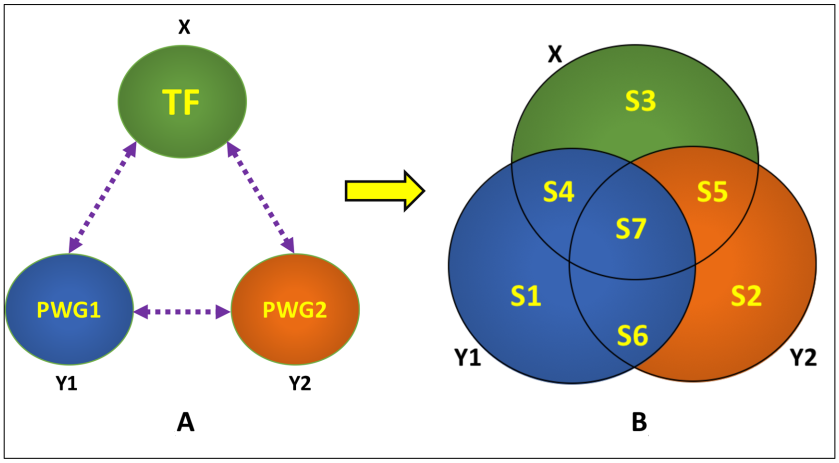


Figure 1. Dissection of interactive components for a given triple gene block. A. Pathway Gene 1 (PWG1) and 2 (PWG2), and the TF are represented by Y1, Y2 and X, respectively. B. Dissected components of interaction among triple genes.

following formulae. We used $(Y1,\ Y2)$ gene pair as an example to show how to calculate the mutual information, and for the other pairs of variables, $(Y1,\ X)$ and $(Y2,\ X)$, could be calculated in the similar way.

The conditional entropy, $H(Y1|Y2)$, was calculated as:

$$H(Y1|Y2) = -\sum_{y_1,y_2} p(y_1, y_2) \log p(y_1|y_2)$$

and the joint entropy, $H(Y1,\ Y2)$, as:

$$H(Y1, Y2) = -\sum_{y_1,y_2} p(y_1, y_2) \log p(y_1, y_2)$$

From this, mutual information, $I(Y1;\ Y2)$, could be obtained:

$$I(Y1; Y2) = H(Y1) + H(Y2) - H(Y1, Y2)$$

The S1, S2 and S3 segments as shown in Figure 1B can be represented by the conditional entropies of each variable given the other two variables, i.e. by $H(Y1|X, Y2)$, $H(Y2|X, Y1)$ and $H(X|Y1, Y2)$. S4 is the conditional mutual information of Y1, X given Y2; S5 is the conditional mutual information of Y2, X given Y1; S6 is conditional mutual information of Y1, Y2 given X; S7 is the difference between mutual information between Y1, Y2 and conditional mutual information Y1, Y2 given X. These can be calculated using the definition of multivariate conditional entropy, as follows:

The joint entropy, $H(Y1, Y2, X)$, is calculated in the same way as described in (27).

$$H(Y1, Y2, X) = -\sum_{y_1,y_2,x} p(y_1, y_2, x) \log p(y_1, y_2, x)$$

From this, the conditional entropy for S1, $H(Y1|X, Y2)$, can then be calculated as:

$$H(Y1|X, Y2) = H(Y1, Y2, X) - H(Y2, X) - H(X)$$

S2 and S3 can be calculated in a similar manner. Conditional mutual information for S6 in Figure 1B, i.e. $(Y1; Y2|X)$, can be calculated as:

$$I(Y1; Y2|X) = H(Y1|X) - H(Y1|X, Y2)$$

Then, multivariate mutual information for S7, $I(Y1; Y2; X)$ can then be calculated from:

$$I(Y1; Y2; X) = I(Y1; Y2) - I(Y1; Y2|X)$$

Note that a positive S7 indicates there exist significant regulatory interactions in the given triple gene block. If S7 is negative, there is no regulatory interactions and the given triple gene block should be discarded.

The MIM for a triple gene block is calculated as follows:

$$MIM = \frac{S7}{S1 + S2 + S3} = \frac{I(Y1; Y2; X)}{H(Y1|X, Y2) + H(Y2|X, Y1) + H(X|Y1, Y2)}$$

This MIM reflects the regulatory strength exerted by the TF on two pathway genes in the triple gene block. The larger the MIM, the more significant the TF controls two pathway genes.

To test if an MIM calculated from a triple gene block is significant, a P-value for each triple gene block was calculated using the randomized permutation method (28). First, 1000 permuted datasets were created by random permutation of the data vector of the TF in the triple gene block. This randomization of all TF data vectors broke original triple gene blocks and generated many new recombined ones. The MIM for each randomly combined block was then calculated and compared to the MIM of original triple gene block to estimate the permutation P-value. After obtaining P-values for all triple gene combinations, the Benjamini–Hochberg method (Benjamini and Hochberg, 1995) was employed to perform multiple testing correction. Triple gene blocks with a false discovery rate (FDR) less than the significance level of 0.05 were considered to have significant interactions.

The importance of a TF to a pathway in overall is represented by the frequency of this TF's presence in significant triple gene blocks with an FDR > 0.05. The selection of top k TFs with highest frequency and then linking them with the pathway genes in significant triple gene blocks led to a two-layered network. Based on this network, we extracted putative combinatorial regulation relationships, namely TF_i–PWG_k-TF_j, representing the kth pathway gene PWG_k is regulated by $TF_i$and TF_j. Combinatorial regulation refers to multiple TFs control the same target gene directly. Identification of them enables genetic engineering of pathways and biological processes more precisely.

Validation of TGMI algorithm on multiple pathways using high-throughput gene expression data

The TGMI algorithm was tested for its accuracy in identifying pathway regulators and building regulatory layers above several biological pathways using data from *A. thaliana* hypocotyledonous stem tissues and mouse embryonic stem cells. The pathway genes were downloaded from *Arabidopsis* Information Resource (https://www.arabidopsis.org/). Alternatively, the genes involved in a biological process defined by gene ontology term can be viewed as non-canonical pathway genes for identification of regulatory layers above pathway genes using TGMI algorithm.

Lignin biosynthesis pathway in A. thaliana. Lignin is the second most abundant plant biopolymer found in secondary cell walls and fibers of wood (29,30). Understanding how lignin is synthesized has long been a research focus of plant biologists and the wood industry because of the importance of lignin in plant structural integrity and stem stiffness (31,32). To identify pathway regulators that govern lignin biosynthesis, we used a compendium dataset that comprised 128 microarray samples (33). The data in this compendium were generated from *A. thaliana* stem tissues under short-day conditions, which induced secondary wood formation. The expression data of the lignin pathway genes and all the TFs were extracted from this compendium data and were subjected to the analysis by the TGMI algorithm. The performance of TGMI was compared with those of the BWERF, Bottom-up GGM and ARACNE algorithms.

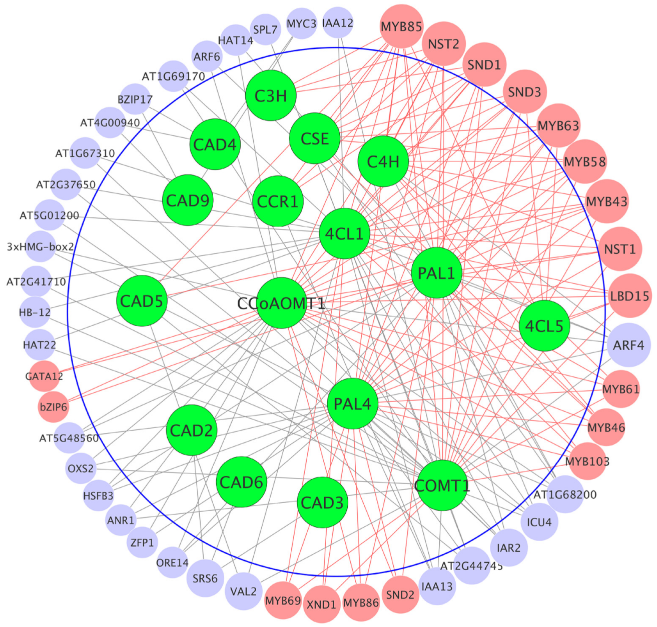
Pigment biosynthesis pathways in A. thaliana. The performance of the TGMI algorithm was also tested with a unified pigment biosynthesis pathway in *A. thaliana*. In plants, pigments provide a broad range of colors from red and orange to blue and violet, serving as important compounds that attract insects for pollination and protect against ultraviolet-B radiation (34). Previous studies have suggested coordinated activity among four linked plant pigment biosynthesis pathways: the anthocyanin, flavonol, flavonoid, and leucopelar–gonidin and leucocyanidin biosynthesis pathways. Leucopelargonidin and leucocyanidin are colorless intermediates synthesized during the course of colored anthocyanin pigmentation (35). Chemical reactions involving leucopelargonidin and leucocyanidin compounds result in red or pink anthocyanin pigmentation in a variety of plant species including *A. thaliana* (36). Flavonoids are modified by a number of reactions that contribute to pigmentation in seeds and flowers (37). Visible patterns of anthocyanins in plants are intermediated by chemical compounds synthesized by flavonol biosynthesis (38). In this study, we treated anthocyanin, flavonol, flavonoid, and leucopelar–gonidin and leucocyanidin biosynthesis as a single large unified pigment biosynthesis pathway for identifying pathway regulators. First, we performed a co-expression analysis to identify co-expressed pathway gene pairs across the four pigment pathways; the significantly co-expressed pathway gene pairs were used in triple gene blocks. The co-expression among these four pigmentation-related pathways was analyzed using four different pairwise association methods: Spearman rank correlation coefficients, Pearson product moment correlation coefficients, Kendall rank correlation coefficients (39) and the maximum information coefficient (MIC) (40). Of these methods, the Spearman and Kendall coefficients can capture monotonic relationships, whereas the MIC can capture varying degrees of both linear and non-linear relationships between genes. After removing duplicated pairs, all the significant co-expressed pathway gene pairs identified by these four gene association methods were used in triple gene blocks to identify pigment pathway gene regulators.

Pluripotency maintenance pathway in mouse embryonic stem cells. TGMI was also tested with three mouse microarray datasets we prepared in our previous study from a data with 34 samples, which were originally downloaded from the ESCAPE repository. From this original data, the expression data of 35 TFs that were known to control pluripotency were extracted. Following that, the three datasets, namely, Datasets 1, 2 and 3 all contain 24 genes involved in pluripotency maintenance plus 100, 200 and 300 randomly selected noise genes, respectively. The expression data of these noise genes were simulated. For more detail, please consult our previous publication (14).

Cell cycle pathway in Saccharomyces cerevisiae. The TGMI algorithm was also tested for identifying pathway regulators with a microarray dataset from *Saccharomyces cerevisiae* (41). 29 cell cycle pathway genes were selected from a previous publication (42), as shown in Supplemental File 1, and all TF in yeast were used. BWERF, Bottom-up GGM and ARACNE algorithms were also used as comparisons.

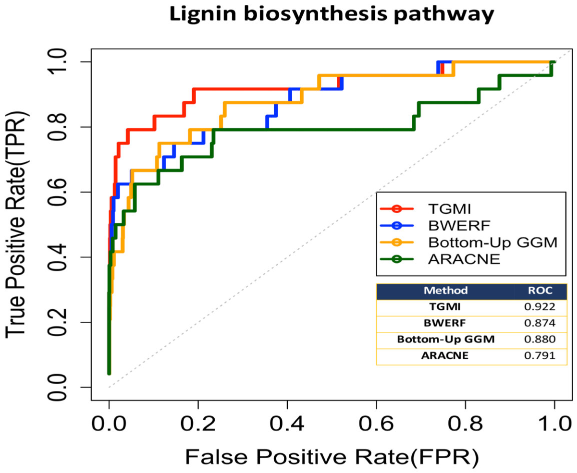
Comparing the accuracy of TGMI with that of other algorithms

Receiver operating characteristic (ROC) curves were used to compare the accuracy of TGMI algorithm with those of the BWERF, Bottom-up GGM and ARACNE algorithms. A pair of true positive rate (TPR) and false positive rate (FPR) values were calculated for each cut-off point in a TF list presorted in descending order based on interference frequencies on pathway genes. In this study, the cut-off point progressed at one gene each time and the same cut-off points were used to threshold the ranked TF lists obtained from all methods to ensure a fair comparability. The acquired TPR and FPR values for all cut-off points were used to plot the ROC curves. An ROC curve that closely follows the TPR axis and then the upper FPR axis indicates a high accuracy of the algorithm, whereas a curve that is closer to the 45-degree diagonal line indicates a low accuracy. To quantify



Rank	TF	Freq.	Reference
1	MYB85	11	(Zhou, 2009)
2	NST2	9	(Zhong & Ye, 2015)
3	SND1	9	(Zhong & Ye, 2015)
4	MYB63	8	(Zhou et al., 2009)
5	SND3	8	(Zhong, Lee, 2008)
6	MYB43	7	(Yoon, Choi, 2015)
7	MYB58	7	(Zhou et al., 2009)
8	ARF4	6	-
9	LBD15	6	(Shuai, 2002)
10	NST1	6	(Mitsuda et al., 2007)
11	AT1G68200	5	-
12	IAR2	5	-
13	ICU4	5	-
14	MYB103	5	(Ohman et al., 2013)
15	MYB46	5	(Kim, 2014)
16	MYB61	5	(Zhong & Ye, 2009)
17	AT2G44745	4	-
18	IAA13	4	-
19	ANR1	3	-
20	AT5G48560	3	-
21	HSFB3	3	-
22	MYB69	3	(Zhong et al., 2008)
23	MYB86	3	(Zhong et al., 2008)
24	ORE14	3	-
25	OXS2	3	-
26	SND2	3	(Hussey et al., 2011)
27	SRS6	3	-
28	VAL2	3	-
29	XND1	3	(Zhao, 2008)
30	3xHMG-box2	2	-
31	ARF6	2	-
32	AT1G67310	2	-
33	AT1G69170	2	-
34	AT2G37650	2	-
35	AT2G41710	2	-
36	AT4G00940	2	-
37	AT5G01200	2	-
38	BZIP17	2	-
39	bZIP6	2	(Zhong & Ye, 2012)
40	GATA12	2	(Nishitani, 2015)

Figure 2. Regulatory network generated by TGMI algorithm for the *Arabidopsis thaliana* lignin biosynthesis pathway using the microarray data generated from hypocotyledonous stem tissues under the short-day condition. Green nodes represent pathway genes. All other nodes are TFs regardless of what colors they are. Light coral nodes represent positive known pathway regulators in current knowledgebase with the literature evidence being provided.



Method	ROC
TGMI	0.922
BWERF	0.874
Bottom-Up GGM	0.880
ARACNE	0.791

Figure 3. The performance comparison of TGMI with the other algorithms in recognition of lignin pathway regulators using ROC curves. The ROC curves that closely follow the TPR axis, and then the top FPR axis, represent the higher accuracies in identifying positive regulatory TFs. Area under each curve (AUROC) was also calculated and provided for comparing the accuracies of different algorithms.

the differences between the accuracy levels, the areas under the ROC curves (AUROCs) were calculated. AUROCs can vary from 0.5 (no positive TFs were found) to 1.0 (all positive TFs were identified).

RESULTS

Lignin biosynthesis pathway in *A. thaliana*

The triple gene blocks identified by TGMI algorithm with a cut-off significance level of 0.05 were shown in the right panel of Figure 2. The interference frequencies of TFs on pathway genes were displayed in the descending order and the TFs highlighted in red are known positive TFs in existing literature. SND1 is a high hierarchical regulator that controls SND2, SND3, MYB103, MYB85, MYB52, MYB54, MYB69, MYB42, MYB43, MYB86, MYB61, MYB46, MYB20 and KNAT7 (43–45). TGMI algorithm identified 9 of these 15 TFs (SND1, SND2, SND3, MYB103, MYB85, MYB43, MYB46, MYB86, MYB61). NST1, NST2, VND6 and VND7, the functional NAC family homologs of SND1, regulate the same downstream targets in different cell types (46). TGMI recognized NST1 and NST2. In addition, MYB58 and MYB63, which are transcriptional activators of lignin biosynthesis in the SND1-mediated transcriptional regulatory network (47), were identified by TGMI algorithm. TGMI also identified LBD15 (48), XND1 (49), bZIP6 (50) and GATA12 (51), which are involved in regulating various aspects of secondary cell wall synthesis.

The triple gene blocks with significant regulatory interactions were combined to generate a circular network, as shown in the left panel of Figure 2, with the TFs arranged in a clockwise direction, from the most frequent to the least frequent. Each directed edge from a TF to a pathway gene represents a regulatory relationship. It is perceivable that the known positive lignin pathway regulators, highlighted in a light coral color, are those that were most frequently connected to lignin biosynthesis pathway genes.

Figure 3 showed the ROC curves representing the accuracies of the TGMI, BWERF, Bottom-up GGM and ARACNE algorithms. As what can be observed from the shapes of the ROC curves, the TGMI algorithm showed much higher accuracy than the other three algorithms in identifying lignin pathway regulators. This was supported

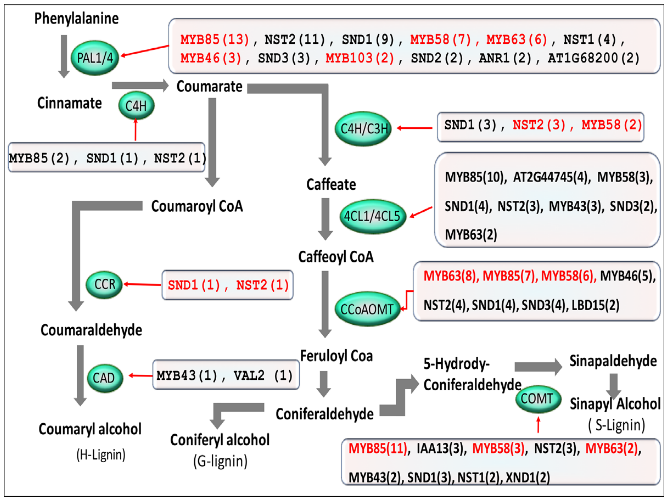

Phenylalanine
PAL1/4
MYB85 (13) , NST2 (11) , SND1 (9) , MYB58 (7) , MYB63 (6) , NST1 (4) , MYB46 (3) , SND3 (3) , MYB103 (2) , SND2 (2) , ANR1 (2) , AT1G68200 (2)
Cinnamate
C4H
Coumarate
MYB85 (2) , SND1 (1) , NST2 (1)
C4H/C3H
SND1 (3) , NST2 (3) , MYB58 (2)
Caffeate
4CL1/4CL5
MYB85(10), AT2G44745(4), MYB58(3), SND1(4), NST2(3), MYB43(3), SND3(2), MYB63(2)
Coumaroyl CoA
Caffeoyl CoA
CCR
SND1 (1) , NST2 (1)
CCoAOMT
MYB63(8), MYB85(7), MYB58(6), MYB46(5), NST2(4), SND1(4), SND3(4), LBD15(2)
Coumaraldehyde
CAD
MYB43 (1) , VAL2 (1)
Feruloyl Coa
5-Hydrody-
Coniferaldehyde
Sinapaldehyde
COMT
Sinapyl Alcohol
(S-Lignin)
Coniferaldehyde
Coumaryl alcohol
(H-Lignin)
Coniferyl alcohol
(G-lignin)
MYB85(11), IAA13(3), MYB58(3), NST2(3), MYB63(2), MYB43(2), SND1(3), NST1(2), XND1(2)

Figure 4. Identified combinatorial TFs are depicted in the pathway diagram of lignin biosynthesis. The green oval shapes show pathway genes involved in the lignin biosynthesis. The combinatorial TFs are shown in square shapes. The frequencies of interactions for each TF in significant triple gene blocks (TF_i-PWG_k-TF_j) are given in parentheses next to each TF. The combinatorial TFs, which have supporting literature evidence for regulating pathway genes, are highlighted in the squares.

by TGMI's significantly larger AUROCC (0.92), indicating it performed better than the other three algorithms. It should be noted that the performances of BWERF and Bottom-up GGM were based on only one layer of TFs inferred, but these algorithms were originally designed to build ML-hGRNs.

The network resulting from TGMI algorithm can be used to identify the possible combinatorial regulatory genes for each pathway gene. We generated a combinatorial map for the lignin pathway (Figure 4). The frequency of each TF–pathway edge (TF_i–PWG_k–TF_j) present in the triple gene blocks with significant interactions is given in parentheses next to the TFs. The results of combinatorial regulation appear to align well with the existing knowledge base. For example, TGMI identified MYB85, MYB58 and MYB63 as combinatorial TFs for the lignin pathway genes PAL1/4, COMT and CCoAMT1, which is consistent with previous reports (52). In addition, MYB103 and MYB46 are combinatorial TFs for PAL1/4 pathway genes, which again is in agreement with the literature (52). SND1 is a regulator that appears in multiple combinatorial TF groups, as shown in Figure 4. A previous study has suggested that SND1 regulates many lignin pathway genes, including PAL1, CCoAOMT and 4CL1 (53). The results also showed that NST2 and MYB58 are combinatorial TFs that regulate both C3H and C4H genes. This is consistent with the conclusion of an earlier study that C3H and C4H are regulated by common TFs that include NST2 and MYB58 (54). The algorithm identified 4CL1/5 pathway genes that are directly regulated by several MYB domain TFs, which include combinatorial TFs, MYB85, MYB58 and MYB63 (Figure 4). A previous study showed that these three TFs regulated multiple 4CL family gene members, including 4CL1/5 (55). In addition, it has been suggested that the CCR family is directly regulated by secondary wall thickening TFs, including SND1 and NST2 (56). Current knowledge suggests that the CCR and CAD genes, which are located in the later steps of lignin biosynthesis, have a milder influence on lignin deposition (57). Consistent with this, our analysis found a relatively low frequency of combinatorial TFs (SND1, NST2) that regulate the CAD family of genes (Figure 4).

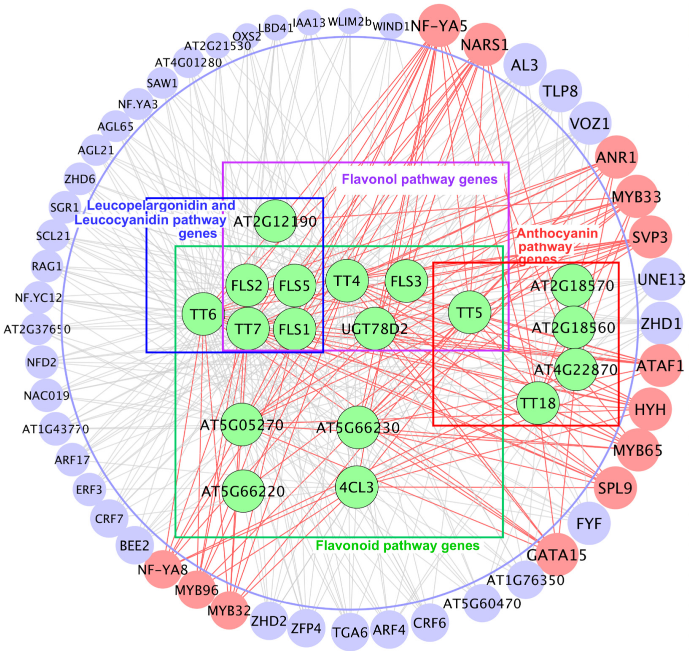
Pigment biosynthesis pathways in *A. thaliana*

The co-expression analysis was performed between all pairwise genes involved in the unified pigmentation pathway including anthocyanin, flavonol, flavonoid, and leucopelargonidin and leucocyanidin biosynthesis using Spearman, Pearson, Kendall and MIC coefficients. The identified co-expressed pathway gene pairs were used to bait the regulators. The TFs captured with higher frequency are listed at the right-side in Figure 5 and many of them are functionally associated with the pigment biosynthesis. NFYA5 is a stress-responsive regulator related to anthocyanin synthesis, which controls purple pigmentation under drought conditions (58). NARS1 is involved in anthocyanin pigmentation in the epidermal cells of the *A. thaliana* (59). ANR1 and ANR2 have been shown to induce the over-accumulation of flavonoid intermediates by suppressing the anthocyanin pathway genes (60). MYB33 and MYB65 are involved in anthocyanin accumulation and seed color pigmentation (61). Overexpression of the SVP3 gene in kiwi fruit has been shown to interfere with anthocyanin biosynthesis in petals (62). ATAF1 is involved in anthocyanin synthesis in *A. thaliana* in adverse growth conditions (63). HYH is an *Arabidopsis* bZIP TF that is directly involved in anthocyanin and chlorophyll estimation (64). SPL9 negatively regulates anthocyanin by directly suppressing anthocyanin biosynthesis genes (65). GATA15 is involved in various activities to modify chlorophyll pigment content in response to different environmental conditions (66). MYB32 indirectly regulates anthocyanin biosynthesis through MYB112, which is a known regulator of the anthocyanin pathway (67). In addition, MYB96 is a drought stress response regulator that affects anthocyanin synthesis in *A. thaliana* (68). NFYA8 is a major regulator of tomato ripening; this TF is also present in *A. thaliana* (68).

ROC curves that compared the accuracy of TGMI algorithm for identification of pigment biosynthesis regulators with those of the other three existing algorithms were shown in Figure 6. The results demonstrated that TGMI algorithm had greater accuracy and recognized more positive regulatory TFs than the other three algorithms; this is supported by the higher AUROC for the TGMI algorithm (0.8). It should be noted that the performance BWERF and Bottom-up GGM is based on only one layer of TFs, but these algorithms were designed to build ML-hGRNs.

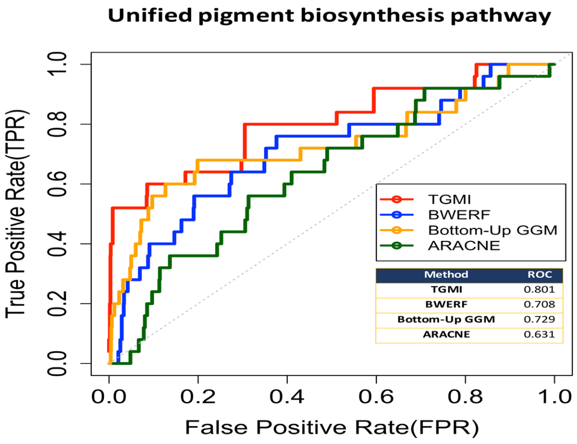
Pluripotency maintenance pathway in mouse embryonic stem cells

The regulatory network operating above the mouse pluripotency maintenance pathway was constructed with TGMI algorithm using mouse Dataset 3, which contains 24 pathway genes, 35 TFs and 300 randomly selected noise genes. The



Rank	TF	Freq.	Reference
1	NFYA5	14	(W. X. Li et al., 2008)
2	NARS1	12	(Kunieda et al., 2013)
3	AL3	11	-
4	TLP8	11	-
5	VOZ1	11	-
6	ANR1	10	(Kovinich et al., 2012)
7	MYB33	10	(Ambawat, Sharma,2013)
8	SVP3	10	(Wu et al., 2014)
9	UNE13	10	-
10	ZHD1	10	-
11	ATAF1	9	(Mahmood, Xu, 2016)
12	HYH	9	(Zhang, Zheng, 2011)
13	MYB65	9	(Ambawat et al., 2013)
14	SPL9	9	(Chao et al., 2017)
15	FYF	9	-
16	GATA15	8	(Manfield, Devlin, 2007)
17	AT1G76350	8	-
18	AT5G60470	8	-
19	CRF6	8	-
20	ARF4	8	-
21	TGA6	8	-
22	ZFP4	8	-
23	ZHD2	8	-
24	MYB32	7	(Lotkowska et al., 2015)
25	MYB96	7	(Seo & Park, 2010)
26	NFYA8	7	(S. Li et al., 2016)
27	BEE2	7	-
28	CRF7	7	-
29	ERF3	7	-
30	ARF17	7	-
31	AT1G43770	7	-
32	NAC019	7	-
33	NFD2	7	-
34	AT2G37650	7	-
35	NF.YC12	7	-
36	RAG1	7	-
37	SCL21	7	-
38	SGR1	7	-
39	ZHD6	7	-
40	AGL21	6	-

Figure 5. Regulatory network for the unified pigment biosynthesis pathway that consists of anthocyanin, flavonol, flavonoid, and leucopelar-gonidin and leucocyanidin biosynthesis. Green nodes represent pathway genes. All other nodes are TFs regardless of their colors. Light coral nodes represent positive known TFs of pigment biosynthesis. References are provided for those positive known TFs that are evidenced to regulate pigment synthesis pathway genes.



Method	ROC
TGMI	0.801
BWERF	0.708
Bottom-Up GGM	0.729
ARACNE	0.631

Figure 6. The performance of TGMI in recognition of authentic pigment pathway regulators using ROC curves. The ROC curves that closely follow the TPR axis, and then the top FPR axis, represent the higher accuracies in identifying positive regulatory TFs. In the contrary, the ROC curves that more closely follow the 45-degree diagonal line reflect lesser accuracies. AUROC was also calculated and provided for comparing the accuracies of different algorithms.

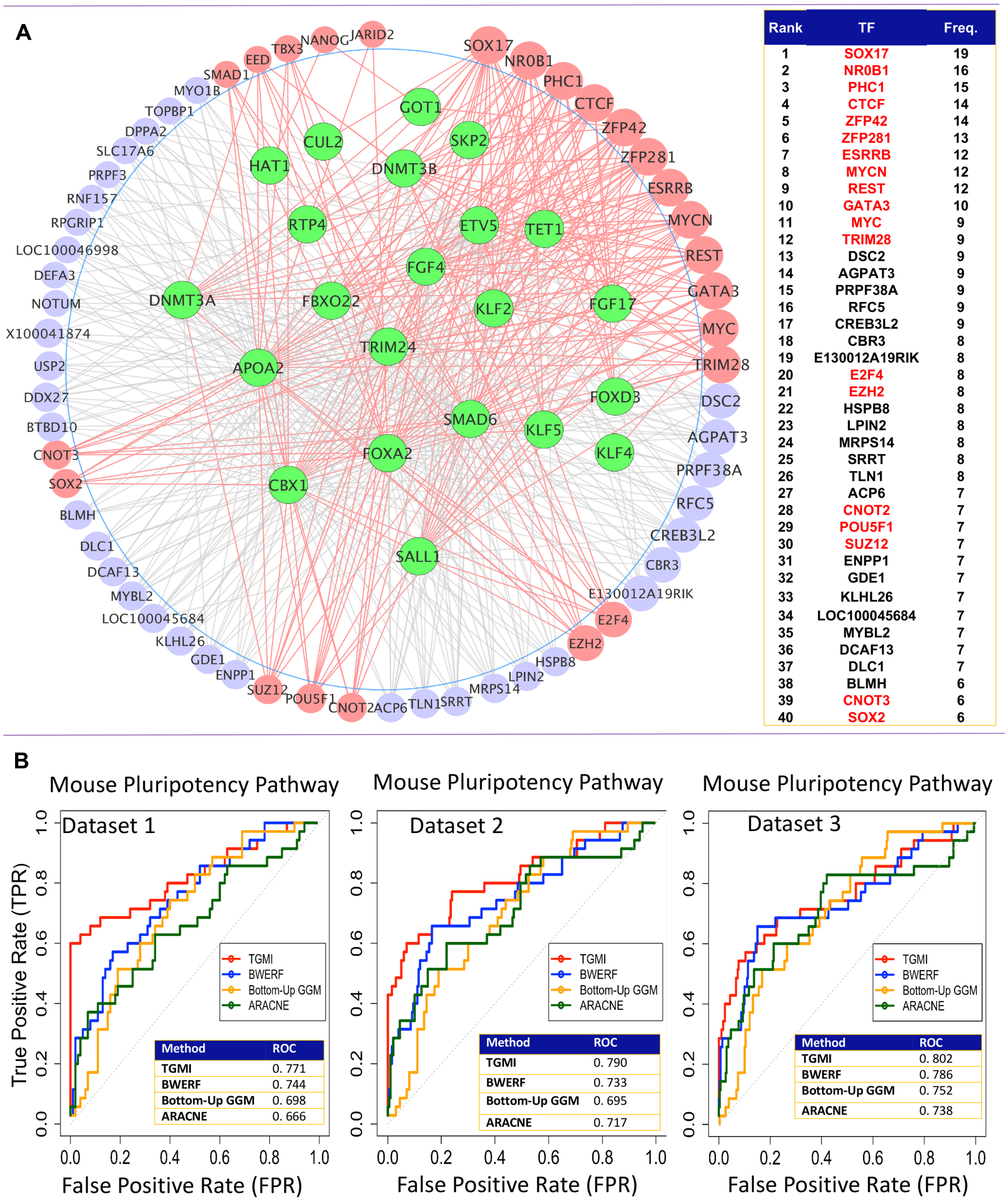
results were shown in Figure 7A. The TFs recognized by TGMI included 12 positive TFs out of the 35 known TFs involved in pluripotency maintenance in mouse embryonic stem cells: SOX17 (69), NROB1 (70), PHC1 (71), CTCF (72), ZFP42 (73), ZFP281 (74), ESRRB (75), MYCN (76), REST (77), GATA3 (78), MYC (79) and TRIM28 (80). POU5F1, SOX2 and NANOG are well-known master regulatory genes that govern stem cell renewal in mice (81–86). Although they were identified by the TGMI algorithm, their frequencies of interactions with pathway genes were relatively low (Figure 7A). A possible explanation is that these three master regulators are located at higher hierarchical levels, and are relatively distal from the lower hierarchical pathway genes used to identify them. ROC curves were used to compare the accuracy of TGMI with those of three other existing algorithms using all the three mouse datasets, namely, Dataset 1, 2 and 3. The results indicated that TGMI algorithm had a higher accuracy than the other three algorithms when tested with all three datasets (Figure 7B). The AUROCs of the TGMI algorithm for identifying the pluripotency pathway regulatory TFs using Datasets 1, 2 and 3 were 0.77, 0.79 and 0.80, respectively. Note that the performance of the BWERF and Bottom-up GGM algorithms is based on only one layer of TFs, but these algorithms were designed and tailored to build ML-hGRNs.

Cell cycle pathway in *S. cerevisiae*

We predicted the pathway regulators that govern cell cycle pathway in yeast. The resulting rankings of TFs are provided in the Supplementary Table S1. ROC curves and AUROCs of TGMI and three other comparative algorithms, BWERF, Bottom-up GGM and ARACNE, were enclosed in the Supplementary Figure S1. Obviously, TGMI had the highest sensitivity/specificity among all methods and AUROCs for TGMI, Bottom-up GGM, BWERF and ARACNE were 0.91, 0.72, 0.71 and 0.5, respectively.

DISCUSSION

The TGMI algorithm was developed for identifying novel regulators that control pathways and biological processes. The underlying principle includes: (i) dissection of the interactions among triple gene block into multiple compo-



Rank	TF	Freq.
1	SOX17	19
2	NR0B1	16
3	PHC1	15
4	CTCF	14
5	ZFP42	14
6	ZFP281	13
7	ESRRB	12
8	MYCN	12
9	REST	12
10	GATA3	10
11	MYC	9
12	TRIM28	9
13	DSC2	9
14	AGPAT3	9
15	PRPF38A	9
16	RFC5	9
17	CREB3L2	9
18	CBR3	8
19	E130012A19RIK	8
20	E2F4	8
21	EZH2	8
22	HSPB8	8
23	LPIN2	8
24	MRPS14	8
25	SRRT	8
26	TLN1	8
27	ACP6	7
28	CNOT2	7
29	POU5F1	7
30	SUZ12	7
31	ENPP1	7
32	GDE1	7
33	KLHL26	7
34	LOC100045684	7
35	MYBL2	7
36	DCAF13	7
37	DLC1	7
38	BLMH	6
39	CNOT3	6
40	SOX2	6

Dataset 1

Method	ROC
TGMI	0. 771
BWERF	0. 744
Bottom-Up GGM	0. 698
ARACNE	0. 666

Dataset 2

Method	ROC
TGMI	0. 790
BWERF	0. 733
Bottom-Up GGM	0. 695
ARACNE	0. 717

Dataset 3

Method	ROC
TGMI	0. 802
BWERF	0. 786
Bottom-Up GGM	0. 752
ARACNE	0. 738

Figure 7. Regulatory network for mouse pluripotency maintenance non-canonical pathway built with TGMI and comparison of TGMI with other algorithms. (**A**) Regulatory network generated using Dataset 3 (335 TF). The green nodes are pathway genes, and all others are TFs. The light coral colored nodes represent known authentic TFs. (**B**) The performance of TGMI in comparison with those of three other algorithms in identifying mouse pluripotency pathway authentic regulatory TFs. Left: ROC curves generated using 135 TF dataset (Dataset 1). Middle: ROC curves generated using 235 TF dataset (Dataset 2). Right: ROC curves generated using 335 TF dataset (Dataset 3). AUROC was also calculated and provided for comparing the accuracies of different algorithms.

nents using conditional mutual information; (ii) use of a novel MIM to identify significantly interacted triple genes; (iii) ranking TFs present in the significantly interacted triple genes based on their interference frequency with pathway genes. When testing with four pathways from three species, the TGMI algorithm consistently performed well in identifying true pathway regulators through constructing two-layered regulatory networks with TFs being operating immediately above pathway genes. In addition, TGMI was able to yield combinatorial TFs that co-regulated several of specific lignin pathway genes. TGMI algorithm is based on use of mutual information and conditional mutual information, which have several advantages over linear pairwise association methods (39) because they can be generalized to identify both linear and non-linear relationships (87). As conditional mutual information was introduced and applied to triple gene blocks, more causal relationships could be captured. This is because the evaluation of trivariate variables is better than evaluation of pairwise variables for recognizing causal relationships (45,88,89). The implementa-

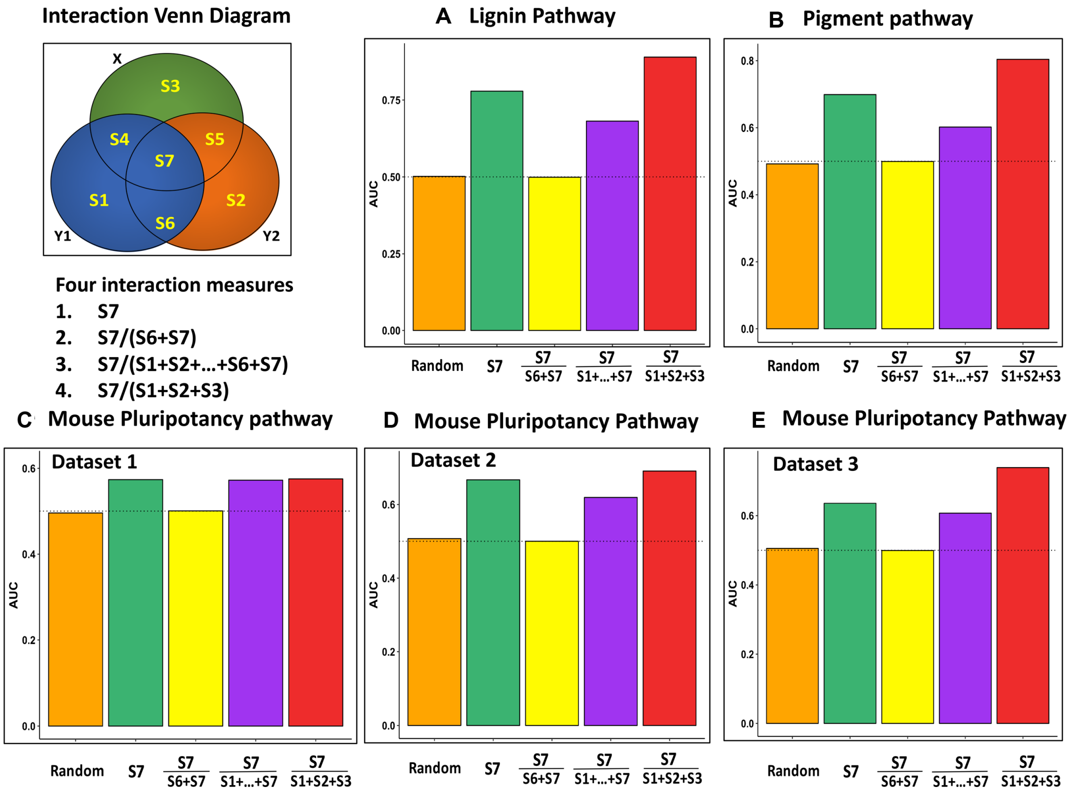


Figure 8. Comparison of the efficiencies of MIM and several other candidate measures in identification of authentic pathway regulators using AUROC curves. Y1, Y2 represent pathway gene 1 and 2, respectively, and X represents a TF. (**A**) AUROC of authentic lignin pathway regulators; (**B**) AUROC of pigment biosynthesis pathway regulators; (**C**), (**D**) and (**E**) are AUROCs of mouse pluripotency pathway regulators from Dataset 1 (135 TFs), Dataset 2 (235 TFs) and Dataset 3 (335 TFs), respectively.

tion of TGMI algorithm to four pathways and five datasets obtained from different species demonstrated that TGMI performed better than Bottom-up GGM algorithm, which uses the difference between correlation and partial correlation to identify causal relationships among triple genes. TGMI algorithm enables the emergence of more true regulators by using conditional mutual information on the same triple gene model, and especially the MIM, which was evidenced to serve as a powerful filter to eliminate non-authentic regulatory relationships. To test whether the MIM was most efficient in capturing the authentic regulatory interactions by evaluating triple gene blocks, we also evaluated MIM together with several derivative candidate measures that could also reflect the magnitudes of the interactions within triple gene blocks. The results are shown in Figure 8.

The results, as shown in Figure 8A–D, consistently supported that MIM is the best filter for capturing authentic pathway regulators. We created a training dataset and a test dataset from each of the two *Arabidopsis* datasets and three mouse datasets. Each training dataset and test dataset contained 75 and 25% randomly drawn rows of all triple gene blocks in each of the five datasets. For each triple gene block, 1 was assigned if the TF was a positive TF, 0 otherwise. Each training dataset was used to build a logistic regression classification model, which was then used to predict the positive TFs in the respective testing data. The prediction performance was measured by determining the ROC AUROCs for the four different triple gene interaction measures plus a random predictor. The results were shown in Figure 8, which indicated that MIM consistently exhibited the best prediction performance among the four interaction measures, and thus could be used to capture positive pathway regulators by evaluating triple gene blocks. The distinct differences between the AUROCs of the four different measures suggest that MIM indeed contributed greatly to identification of positive pathway regulators.

CONCLUSION

We developed a novel and efficient algorithm called TGMI which could generate a regulatory layer operating above a metabolic pathway or biological process for facilitating identification of important regulators that govern various pathways or biological processes. In addition, TGMI algorithm could yield some combinatorial TFs that collectively regulate each individual pathway gene, enabling more precise genetic engineering of regulators that regulate rate-limited steps in a pathway or biological process. The algorithm accomplishes these objectives through dissecting all interaction components among triple gene blocks, each containing two pathway genes and one TF, using conditional mutual information, and then through extracting and summarizing regulatory edges from significantly interacted triple gene blocks identified with MIM. The algorithm was validated with four pathways from plants, animals and yeast, and the results consistently supported that TGMI was highly effective in capturing true pathway regulators. Performance tests using ROC curves and AUROC values also showed that MIM consistently outperformed several others candidate measures, manifesting MIM's critical role in bringing the true pathway regulators to light. Since TGMI algorithm does not require time-course data, it is widely applicable to the high-throughput data yielded from treatment versus control or large time-interval time course microarray or RNA-seq experiments, which are most frequently conducted in modern genomics. We thus think TGMI can meet the great needs of the biological research community in identifying pathway regulators for more efficient genetic modification of various metabolic pathways and biological processes.

DATA AVAILABILITY

The R package for TGMI is freely available at: http://sys.bio.mtu.edu/sample_output/TGMI/

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank National Science Foundation (NSF) Advances in Biological Informatics for supporting C.G., and W.D.

Authors' contributions: C.G. and H.W. designed the project; C.G. developed the algorithm with the input from H.W., K.Z., W.D. and L.E.B. C.G. and H.W. contributed on data collection and evaluation as well as data interpretation. H.W. and C.G. wrote the manuscript, and K.Z., W.D. and L.E.B. r.

FUNDING

NSF Advances in Biological Informatics [dbi-1458130 to H.W.], USDA McIntire-Stennis Fund to H.W. and NSF Division of Integrative Organismal Systems [134001 to H.W.]. Funding for open access charge: National Science Foundation Advances in Biological Informatics [dbi-1458130 to H.W.].

Conflict of interest statement. None declared.

REFERENCES

1. Huala,E., Dickerman,A.W., Garcia-Hernandez,M., Weems,D., Reiser,L., LaFond,F., Hanley,D., Kiphart,D., Zhuang,M., Huang,W. *et al.* (2001) The Arabidopsis information resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.*, **29**, 102–105.
2. Lv,Q., Cheng,R. and Shi,T.L. (2014) Regulatory network rewiring for secondary metabolism in Arabidopsis thaliana under various conditions. *BMC Plant Biol.*, **14**, 1–12.
3. Sweetlove,L.J., Last,R.L. and Fernie,A.R. (2003) Predictive metabolic engineering: A goal for systems biology. *Plant Physiol.*, **132**, 420–425.
4. Murphy,S.M. (1999) *Modelling gene expression data using dynamic Bayesian networks*. Technical report,Computer Science Division, University of California.
5. Zou,M. and Conzen,S.D. (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, **21**, 71–79.
6. Chen,X.H., Chen,M. and Ning,K.D. (2006) BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. *Bioinformatics*, **22**, 2952–2954.
7. Chen,T., He,H.L. and Church,G.M. (1999) Modeling gene expression with differential equations. *Pac. Symp. Biocomput*, 29–40.
8. Becskei,A., Seraphin,B. and Serrano,L. (2001) Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J.*, **20**, 2528–2535.
9. Kauffman,S. (1969) Homeostasis and differentiation in random genetic control networks. *Nature*, **224**, 177–178.
10. Chen,B.S., Chang,C.H., Wang,Y.C., Wu,C.H. and Lee,H.C. (2011) Robust model matching design methodology for a stochastic synthetic gene network. *Math. Biosci.*, **230**, 23–36.
11. Ruklisa,D., Brazma,A. and Viksna,J. (2005) Reconstruction of gene regulatory networks under the finite state linear model. *Genome Inform.*, **16**, 225–236.
12. Yang,C. and Wei,H. (2015) Designing microarray and RNA-seq experiments for greater systems biology discovery in modern plant genomics. *Mol. Plant*, **8**, 196–206.
13. 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.
14. Deng,W., Zhang,K., Busov,V. and Wei,H. (2017) Recursive random forest algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways. *PLoS One*, **12**, e0171532.
15. Kumari,S., Deng,W., Gunasekara,C., Chiang,V., Chen,H.S., Ma,H., Davis,X. and Wei,H. (2016) Bottom-up GGM algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways or processes. *BMC Bioinformatics*, **17**, 132.
16. Wei,H. (2017) Construction of a hierarchical gene regulatory network centered around a transcription factor. *Brief. Bioinform.*, doi:10.1093/bib/bbx152.
17. Wei,H., Persson,S., Mehta,T., Srinivasasainagendra,V., Chen,L., Page,G.P., Somerville,C. and Loraine,A. (2006) Transcriptional coordination of the metabolic network in Arabidopsis. *Plant Physiol.*, **142**, 762–774.
18. Birnbaum,K., Shasha,D.E., Wang,J.Y., Jung,J.W., Lambert,G.M., Galbraith,D.W. and Benfey,P.N. (2003) A gene expression map of the Arabidopsis root. *Science*, **302**, 1956–1960.
19. Williams,E.J. and Bowles,D.J. (2004) Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome Res.*, **14**, 1060–1067.
20. Clements,M., van Someren,E.P., Knijnenburg,T.A. and Reinders,M.J. (2007) Integration of known transcription factor binding site information and gene expression data to advance from co-expression to co-regulation. *Genomics Proteomics Bioinformatics*, **5**, 86–101.
21. Yeung,K.Y., Medvedovic,M. and Bumgarner,R.E. (2004) From co-expression to co-regulation: how many microarray experiments do we need? *Genome Biol*, **5**, R48.
22. Allocco,D.J., Kohane,I.S. and Butte,A.J. (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, **5**, 18.
23. Chaffey,N., Cholewa,E., Regan,S. and Sundberg,B. (2002) Secondary xylem development in Arabidopsis: a model for wood formation. *Physiol. Plant*, **114**, 594–600.
24. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
25. Persson,S., Wei,H., Milne,J., Page,G.P. and Somerville,C.R. (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 8633–8638.
26. Boulle,M. (2005) Optimal bin number for equal frequency discretizations in supervized learning. *Intell. Data Anal.*, **9**, 175–188.
27. Cover,T.M. and Thomas,J.A. (2006) *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.
28. Sham,P.C. and Purcell,S.M. (2014) Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.*, **15**, 335–346.

29. Vanholme,R., Demedts,B., Morreel,K., Ralph,J. and Boerjan,W. (2010) Lignin biosynthesis and structure. *Plant Physiol.*, **153**, 895–905.
30. Dixon,R.A. and Paiva,N.L. (1995) Stress-induced phenylpropanoid metabolism. *Plant Cell*, **7**, 1085–1097.
31. Chabannes,M., Ruel,K., Yoshinaga,A., Chabbert,B., Jauneau,A., Joseleau,J.P. and Boudet,A.M. (2001) In situ analysis of lignins in transgenic tobacco reveals a differential impact of individual transformations on the spatial patterns of lignin deposition at the cellular and subcellular levels. *Plant J.*, **28**, 271–282.
32. Donaldson,L.A. (2001) Lignification and lignin topochemistry—an ultrastructural view. *Phytochemistry*, **57**, 859–873.
33. Kumari,S., Deng,W., Gunasekara,C., Chiang,V., Chen,H.S., Ma,H., Davis,X. and Wei,H. (2016) Bottom-up GGM algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways or processes. *BMC Bioinformatics*, **17**, 132.
34. Tanaka,Y., Sasaki,N. and Ohmiya,A. (2008) Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. *Plant J.*, **54**, 733–749.
35. Springob,K., Nakajima,J., Yamazaki,M. and Saito,K. (2003) Recent advances in the biosynthesis and accumulation of anthocyanins. *Nat. Prod. Rep.*, **20**, 288–303.
36. Burbulis,I.E. and Winkel-Shirley,B. (1999) Interactions among enzymes of the Arabidopsis flavonoid biosynthetic pathway. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 12929–12934.
37. Forkmann,G. and Martens,S. (2001) Metabolic engineering and applications of flavonoids. *Curr. Opin. Biotechnol.*, **12**, 155–160.
38. Martens,S., Teeri,T. and Forkmann,G. (2002) Heterologous expression of dihydroflavonol 4-reductases from various plants. *FEBS Lett.*, **531**, 453–458.
39. Kumari,S., Nie,J., Chen,H.S., Ma,H., Stewart,R., Li,X., Lu,M.Z., Taylor,W.M. and Wei,H. (2012) Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS One*, **7**, e50411.
40. Reshef,D.N., Reshef,Y.A., Finucane,H.K., Grossman,S.R., McVean,G., Turnbaugh,P.J., Lander,E.S., Mitzenmacher,M. and Sabeti,P.C. (2011) Detecting novel associations in large data sets. *Science*, **334**, 1518–1524.
41. Caba,E., Dickinson,D.A., Warnes,G.R. and Aubrecht,J. (2005) Differentiating mechanisms of toxicity using global gene expression analysis in Saccharomyces cerevisiae. *Mutat. Res.*, **575**, 34–46.
42. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I., Zeitlinger,J. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.
43. Zhong,R. and Ye,Z.H. (2015) The Arabidopsis NAC transcription factor NST2 functions together with SND1 and NST1 to regulate secondary wall biosynthesis in fibers of inflorescence stems. *Plant Signal. Behav.*, **10**, e989746.
44. Zhong,R., Richardson,E.A. and Ye,Z.H. (2007) The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in Arabidopsis. *Plant Cell*, **19**, 2776–2792.
45. Lin,Y.C., Li,W., Sun,Y.H., Kumari,S., Wei,H., Li,Q., Tunlaya-Anukit,S., Sederoff,R.R. and Chiang,V.L. (2013) SND1 transcription factor-directed quantitative functional hierarchical genetic regulatory network in wood formation in Populus trichocarpa. *Plant Cell*, **25**, 4324–4341.
46. Zhong,R., Lee,C., Zhou,J., McCarthy,R.L. and Ye,Z.H. (2008) A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. *Plant Cell*, **20**, 2763–2782.
47. Zhou,J., Lee,C., Zhong,R. and Ye,Z.H. (2009) MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. *Plant Cell*, **21**, 248–266.
48. Shuai,B., Reynaga-Pena,C.G. and Springer,P.S. (2002) The lateral organ boundaries gene defines a novel, plant-specific gene family. *Plant Physiol.*, **129**, 747–761.
49. Zhao,C., Avci,U., Grant,E.H., Haigler,C.H. and Beers,E.P. (2008) XND1, a member of the NAC domain family in Arabidopsis thaliana, negatively regulates lignocellulose synthesis and programmed cell death in xylem. *Plant J.*, **53**, 425–436.
50. Zhong,R. and Ye,Z.H. (2012) MYB46 and MYB83 bind to the SMRE sites and directly activate a suite of transcription factors and secondary wall biosynthetic genes. *Plant Cell Physiol.*, **53**, 368–380.
51. Nishitani,K. and Demura,T. (2015) An emerging view of plant cell walls as an apoplastic intelligent system. *Plant Cell Physiol.*, **56**, 177–179.
52. Hussey,S.G., Mizrachi,E., Creux,N.M. and Myburg,A.A. (2013) Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Front. Plant Sci.*, **4**, 325.
53. Ohashi-Ito,K., Oda,Y. and Fukuda,H. (2010) Arabidopsis VASCULAR-RELATED NAC-DOMAIN6 directly regulates the genes that govern programmed cell death and secondary wall formation during xylem differentiation. *Plant Cell*, **22**, 3461–3473.
54. Poovaiah,C.R., Nageswara-Rao,M., Soneji,J.R., Baxter,H.L. and Stewart,C.N. (2014) Altered lignin biosynthesis using biotechnology to improve lignocellulosic biofuel feedstocks. *Plant Biotechnol. J.*, **12**, 1163–1173.
55. Liu,Y., Wei,M., Hou,C., Lu,T., Liu,L., Wei,H., Cheng,Y. and Wei,Z. (2017) Functional characterization of populus PsnSHN2 in coordinated regulation of secondary wall components in Tobacco. *Sci. Rep.*, **7**, 42.
56. Mitsuda,N. and Ohme-Takagi,M. (2008) NAC transcription factors NST1 and NST3 regulate pod shattering in a partially redundant manner by promoting secondary wall formation after the establishment of tissue identity. *Plant J.*, **56**, 768–778.
57. Yoon,J., Choi,H. and An,G. (2015) Roles of lignin biosynthesis and regulatory genes in plant development. *J. Integr. Plant Biol.*, **57**, 902–912.
58. Li,W.X., Oono,Y., Zhu,J.H., He,X.J., Wu,J.M., Iida,K., Lu,X.Y., Cui,X.P., Jin,H.L. and Zhu,J.K. (2008) The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. *Plant Cell*, **20**, 2238–2251.
59. Kunieda,T., Shimada,T., Kondo,M., Nishimura,M., Nishitani,K. and Hara-Nishimura,I. (2013) Spatiotemporal secretion of PEROXIDASE36 is required for seed coat mucilage extrusion in arabidopsis. *Plant Cell*, **25**, 1355–1367.
60. Kovinich,N., Saleem,A., Rintoul,T.L., Brown,D.C., Arnason,J.T. and Miki,B. (2012) Coloring genetically modified soybean grains with anthocyanins by suppression of the proanthocyanidin genes ANR1 and ANR2. *Transgenic Res.*, **21**, 757–771.
61. Ambawat,S., Sharma,P., Yadav,N.R. and Yadav,R.C. (2013) MYB transcription factor genes as regulators for plant responses: an overview. *Physiol. Mol. Biol. Plants*, **19**, 307–321.
62. Wu,R., Wang,T., McGie,T., Voogd,C., Allan,A.C., Hellens,R.P. and Varkonyi-Gasic,E. (2014) Overexpression of the kiwifruit SVP3 gene affects reproductive development and suppresses anthocyanin biosynthesis in petals, but has no effect on vegetative growth, dormancy, or flowering time. *J. Exp. Bot.*, **65**, 4985–4995.
63. Mahmood,K., Xu,Z., El-Kereamy,A., Casaretto,J.A. and Rothstein,S.J. (2016) The arabidopsis transcription factor ANAC032 represses anthocyanin biosynthesis in response to high sucrose and oxidative and abiotic stresses. *Front. Plant Sci.*, **7**, 1548.
64. Zhang,Y.Q., Zheng,S., Liu,Z.J., Wang,L.G. and Bi,Y.R. (2011) Both HY5 and HYH are necessary regulators for low temperature-induced anthocyanin accumulation in Arabidopsis seedlings. *J. Plant Physiol.*, **168**, 367–374.
65. Gou,J.Y., Felippes,F.F., Liu,C.J., Weigel,D. and Wang,J.W. (2011) Negative regulation of anthocyanin biosynthesis in arabidopsis by a miR156-targeted SPL transcription factor. *Plant Cell*, **23**, 1512–1522.
66. Xu,Z. (2016) The role of anthocyanins and the GATA transcription factors GNC and CGA1 in the plant response to stress. Ph.D. Thesis. University of Guelph.
67. Lotkowska,M.E., Tohge,T., Fernie,A.R., Xue,G.P., Balazadeh,S. and Mueller-Roeber,B. (2015) The arabidopsis transcription factor MYB112 promotes anthocyanin formation during salinity and under high light stress. *Plant Physiol.*, **169**, 1862–1880.
68. Seo,P.J. and Park,C.M. (2010) MYB96-mediated abscisic acid signals induce pathogen resistance response by promoting salicylic acid biosynthesis in Arabidopsis. *New Phytol.*, **186**, 471–483.
69. Niakan,K.K., Ji,H.K., Maehr,R., Vokes,S.A., Rodolfa,K.T., Sherwood,R.I., Yamaki,M., Dimos,J.T., Chen,A.E., Melton,D.A. *et al.* (2010) Sox17 promotes differentiation in mouse embryonic stem

cells by directly regulating extraembryonic gene expression and indirectly antagonizing self-renewal. *Gene Dev.*, **24**, 312–326.

70. Fujii,S., Nishikawa-Torikai,S., Futatsugi,Y., Toyooka,Y., Yamane,M., Ohtsuka,S. and Niwa,H. (2015) Nr0b1 is a negative regulator of Zscan4c in mouse embryonic stem cells. *Sci. Rep.*, **5**, 9146.
71. Morey,L., Santanach,A. and Di Croce,L. (2015) Pluripotency and epigenetic factors in mouse embryonic stem cell fate regulation. *Mol. Cell. Biol.*, **35**, 2716–2728.
72. Donohoe,M.E., Silva,S.S., Pinter,S.F., Xu,N. and Lee,J.T. (2009) The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting. *Nature*, **460**, U128–U147.
73. Masui,S., Ohtsuka,S., Yagi,R., Takahashi,K., Ko,M.S.H. and Niwa,H. (2008) Rex1/Zfp42 is dispensable for pluripotency in mouse ES cells. *BMC Dev. Biol.*, **8**, 45.
74. Fidalgo,M., Shekar,P.C., Ang,Y.S., Fujiwara,Y., Orkin,S.H. and Wang,J.L. (2011) Zfp281 functions as a transcriptional repressor for pluripotency of mouse embryonic stem cells. *Stem Cells*, **29**, 1705–1716.
75. Papp,B. and Plath,K. (2012) Pluripotency re-centered around Esrrb. *EMBO J.*, **31**, 4255–4257.
76. Ruiz-Perez,M.V., Henley,A.B. and Arsenian-Henriksson,M. (2017) The MYCN protein in health and disease. *Genes (Basel)*, **8**, 113.
77. Singh,S.K., Kagalwala,M.N., Parker-Thornburg,J., Adams,H. and Majumder,S. (2008) REST maintains self-renewal and pluripotency of embryonic stem cells. *Nature*, **453**, 223–227.
78. Shu,J., Wu,C., Wu,Y.T., Li,Z.Y., Shao,S.D., Zhao,W.H., Tang,X., Yang,H., Shen,L.J., Zuo,X.H. *et al.* (2015) Induction of pluripotency in mouse somatic cells with lineage specifiers (vol 153, pg 963, 2013). *Cell*, **161**, 1229–1229.
79. Chappell,J. and Dalton,S. (2013) Roles for MYC in the establishment and maintenance of pluripotency. *Cold Spring Harb. Perspect. Med.*, **3**, a014381.
80. Miles,D.C., de Vries,N.A., Gisler,S., Lieftink,C., Akhtar,W., Gogola,E., Pawlitzky,I., Hulsman,D., Tanger,E., Koppens,M. *et al.* (2017) TRIM28 is an epigenetic barrier to induced pluripotent stem cell reprogramming. *Stem Cells*, **35**, 147–157.
81. Hall,V.J. and Hyttel,P. (2014) Breaking down pluripotency in the porcine embryo reveals both a premature and reticent stem cell state in the inner cell mass and unique expression profiles of the naive and primed stem cell states. *Stem Cells Dev.*, **23**, 2030–2045.
82. Sharov,A.A., Masui,S., Sharova,L.V., Piao,Y., Aiba,K., Matoba,R., Xin,L., Niwa,H. and Ko,M.S. (2008) Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. *BMC Genomics*, **9**, 269.
83. Rodda,D.J., Chew,J.L., Lim,L.H., Loh,Y.H., Wang,B., Ng,H.H. and Robson,P. (2005) Transcriptional regulation of nanog by OCT4 and SOX2. *J. Biol. Chem.*, **280**, 24731–24737.
84. Loh,Y.H., Wu,Q., Chew,J.L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.
85. Kellner,S. and Kikyo,N. (2010) Transcriptional regulation of the Oct4 gene, a master gene for pluripotency. *Histol. Histopathol.*, **25**, 405–412.
86. Zhou,Q., Chipperfield,H., Melton,D.A. and Wong,W.H. (2007) A gene regulatory network in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 16438–16443.
87. Schneidman,E., Still,S., Berry,M.J. and Bialek,W. (2003) Network information and connected correlations. *Phys. Rev. Lett.*, **91**, 238701.
88. Schäfer,J. and Strimmer,K. (2005) Learning Large-Scale Graphical Gaussian Models from Genomic Data. In: Mendes,J (ed). *Science of Complex Networks: From Biology to the Internet and WWW (CNET 2004)*. The American Institute of Physics, Aveiro.
89. Lu,S., Li,Q., Wei,H., Chang,M.J., Tunlaya-Anukit,S., Kim,H., Liu,J., Song,J., Sun,Y.H., Yuan,L. *et al.* (2013) Ptr-miR397a is a negative regulator of laccase genes affecting lignin content in Populus trichocarpa. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 10848–10853.