

RESEARCH ARTICLE

Global gene network exploration based on explainable artificial intelligence approach

Heewon Park^{1*}, Koji Maruhashi², Rui Yamaguchi^{3,4}, Seiya Imoto⁴, Satoru Miyano^{1,4}

1 M&D Data Science Center, Tokyo Medical and Dental University, Tokyo, Japan, **2** Fujitsu Laboratories Ltd, Kanagawa, Japan, **3** Division of Cancer Systems Biology, Aichi Cancer Center Research Institute, Aichi, Japan, **4** Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan

* hwpark.dsc@tmd.ac.jp

Abstract

In recent years, personalized gene regulatory networks have received significant attention, and interpretation of the multilayer networks has been a critical issue for a comprehensive understanding of gene regulatory systems. Although several statistical and machine learning approaches have been developed and applied to reveal sample-specific regulatory pathways, integrative understanding of the massive multilayer networks remains a challenge. To resolve this problem, we propose a novel artificial intelligence (AI) strategy for comprehensive gene regulatory network analysis. In our strategy, personalized gene networks corresponding specific clinical characteristic are constructed and the constructed network is considered as a second-order tensor. Then, an explainable AI method based on deep learning is applied to decompose the multilayer networks, thus we can reveal all-encompassing gene regulatory systems characterized by clinical features of patients. To evaluate the proposed methodology, we apply our method to the multilayer gene networks under varying conditions of an epithelial–mesenchymal transition (EMT) process. From the comprehensive analysis of multilayer networks, we identified novel markers, and the biological mechanisms of the identified genes and their reciprocal mechanisms are verified through the literature. Although any biological knowledge about the identified genes was not incorporated in our analysis, our data-driven approach based on AI approach provides biologically reliable results. Furthermore, the results provide crucial evidences to reveal biological mechanism related to various diseases, e.g., keratinocyte proliferation. The use of explainable AI method based on the tensor decomposition enables us to reveal global and novel mechanisms of gene regulatory system from the massive multiple networks, which cannot be demonstrated by existing methods. We expect that the proposed method provides a new insight into network biology and it will be a useful tool to integrative gene network analysis related complex architectures of diseases.

OPEN ACCESS

Citation: Park H, Maruhashi K, Yamaguchi R, Imoto S, Miyano S (2020) Global gene network exploration based on explainable artificial intelligence approach. PLoS ONE 15(11): e0241508. <https://doi.org/10.1371/journal.pone.0241508>

Editor: Gary Stein, University of Vermont, UNITED STATES

Received: July 18, 2020

Accepted: September 3, 2020

Published: November 6, 2020

Copyright: © 2020 Park et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data are available from: <https://www.kaggle.com/heewonn/poned2022295>.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Gene regulatory networks are crucial for understanding complex mechanisms of diseases. To reveal heterogeneous genetic networks that underlie complex human diseases, various large-

scale projects (e.g., The Cancer Genome Atlas and Cancer Genome Project) have been conducted and provided considerable amounts of omics data. The scale of gene networks is increasing, and strategies to comprehensively analyze large-scale gene networks have been claimed. In particular, there is currently substantial discussion regarding integrative analysis of sample-specific gene networks for personalized cancer diagnostics and therapeutics.

Shimamura et al. [1] proposed a statistical method for sample-specific network construction, NetworkProfiler, which groups samples according to their specific genomic characteristics (e.g., drug response and survival time) and constructs a network for a target sample based only on samples having characteristics similar to those of the target sample. Thus, we can reveal gene regulatory networks under varying conditions of clinical characteristics of patients. The NetworkProfiler was applied to construct gene networks for 762 cancer cell lines characterized by EMT process, where EMT-related modulators for each cell line were measured based on 50 EMT-related genes labeled in the Molecular Signatures Database. They focused on E-cadherin, which connects epithelial cells at adherens junctions, and identified 24 candidate regulators. Interestingly, the identified genes did not consist of just the 50 genes defining the modulators, i.e., only one of the 24 identified genes was a member of the 50 genes, even though the regulators of E-cadherin were identified from the networks under varying conditions of the EMT modulators computed from the 50 genes. Among the 24 identified genes, approximately half were verified as regulators of E-cadherin from the literature. They selected KLF5 from the remaining genes and performed validation experiments. Through the experiments, the mechanism of KLF5 was demonstrated: knockdown of KLF5 decreased the expression of E-cadherin and led to morphological changes of the characteristics of EMT. Their validation results are also supported by later studies, e.g., Zhang et al. [2]. Although the other half were not verified at that time, a majority of those have been demonstrated as crucial EMT markers in the past decade [3, 4]. Shimamura et al. [1] provided crucial indicators and made major contributions to reveal tumor progress related to EMT. Park et al. [5] suggested that cancer characteristics are not uniformly distributed, and the Gaussian kernel function used to control the effect of samples in the NetworkProfiler leads to extremely small amount of weight for modeling a target sample having rare cancer characteristic, because the Gaussian kernel function is based on a constant bandwidth. To address this problem, Park et al. [5] proposed a robust version of NetworkProfiler based on an adaptive bandwidth via the k -nearest neighbor rule, and constructed a drug sensitivity-specific gene network based on the Sanger dataset from the Cancer Genome Project.

Although existing studies have provided crucial tools for precision medicine, understanding of large-scale multilayer gene networks is limited. In other words, a significant number of multilayer networks cannot be interpreted comprehensively using existing approaches (e.g., Shimamura et al. focused only on E-cadherin, and other regions could not be revealed).

To resolve this problem, we propose a novel strategy based on an explainable artificial intelligence (AI) methodology using tensor decomposition. Although machine learning and AI methods show remarkable performance in modeling accuracy, most of the existing approaches cannot explain how they obtain results (referred to as the black-box problem). This limits AI usage because their results cannot be verified. Maruhashi et al. [6, 7] developed novel explainable AI approaches (i.e., DeepTensor and Tensor Reconstruction-based Interpretable Prediction (TRIP)) for learning multiway relations, which are deep learning approaches using tensor decomposition. Our strategy is based on two stages, i.e., constructing sample-specific gene networks and comprehensive analysis of the constructed multilayer networks by using the explainable AI methodology. That is, we construct a personalized gene regulatory network for each patient and the constructed network is considered as a second-order tensor. We then explore the massive multiple gene networks by using the AI method, TRIP. The use of the

interpretable AI method based on tensor decomposition enables us to overcome limitation of existing gene network analysis, i.e., narrow angle in regulatory networks, and this leads to a greater understanding of biological systems of the regulatory interactions between genes. To the best of our knowledge, this is the first study on revealing gene regulatory networks using an explainable AI method.

To illustrate our strategy, we apply the proposed method to EMT-related gene regulatory networks constructed using NetworkProfiler [1]. We learn a low-dimensional subspace of 762 network tensors using the TRIP and explore multilayer networks on the constructed subspace. From the comprehensive analysis of multilayer networks, we identified novel markers, and the biological mechanisms of the identified genes and their reciprocal mechanisms are verified through the literature. Although any biological knowledge about mechanism of the identified genes was not incorporated in our analysis, the revealed genes by our method have strong evidences. In other words, our data-driven approach provides biologically reliable results. Although we illustrate our method based only on EMT-related networks, it can be expected that the proposed method will be a useful tool to global explore gene regulatory system involved in various clinical characteristics.

The remainder of this paper is organized as follows. In the Method section, we introduce the proposed a novel strategy for global exploration of multilayer personalized gene networks. In the Results section, we describe the evaluation of our strategy based on the gene regulatory networks varying according to the EMT status. Conclusions are provided in the Discussion section.

Method

Suppose X_1, \dots, X_q is q possible regulators that may control transcription of the l^{th} target gene Y_l . Consider the linear regression model for the target gene Y_l ,

$$Y_l = \sum_{j=0}^q \beta_{jl} \cdot X_j + \varepsilon_l, \quad (1)$$

where β_{jl} is the regression coefficient that represents the effect of regulator X_j on target Y_l and ε_l is a random error vector $\varepsilon_l = (\varepsilon_{l1}, \dots, \varepsilon_{lm})^T$ that is assumed to be independently and identically distributed with mean 0 and variance σ^2 . To reveal gene regulatory interactions based on the regression model, various statistical and machine learning methods have been proposed and applied to gene network construction [8, 9]. However, patient-specific gene regulatory systems cannot be revealed via the regression model because the strengths of the relationships between genes are given as β_{jl} for all samples.

We develop a novel strategy for integrative analysis for multilayer gene networks, which are a crucial tool for precision medicine. In our method, gene regulatory networks are constructed under varying conditions of samples and the multilayer networks are analyzed comprehensively using an explainable AI method. The gene network for a target sample is considered as a second-order tensor, and a deep learning method for tensor decomposition is applied to construct low-dimensional subspace of the multiway interaction between genes. Prediction and interpretation are performed on the constructed human-readable low-dimensional subspace, and thus we can effectively understand the constructed large-scale gene networks. Our strategy consists of two stages of constructing sample-specific gene regulatory networks and global investigating large-scale multiple gene networks based on an explainable AI technology.

Stage 1: Constructing personalized gene regulatory network based on sample-specific analysis

We consider the varying-coefficient structural equation model to construct a sample-specific gene regulatory network [10],

$$Y_i = \sum_{j=0}^q \beta_{ji}(m_\alpha) \cdot X_j + \varepsilon_i, \tag{2}$$

$\beta_{ji}(m_\alpha)$ is the regression coefficient of X_j on Y_i for the α^{th} target sample of the modulator $M = m_\alpha$, such as drug sensitivity and survival risk of cell lines.

To estimate the varying coefficient $\beta_{ji}(m_\alpha)$ describing strength of the relationship between the regulator and the target genes for each sample, we considered the the following kernel-based L1-type regularization method [1, 5]

$$L(\beta_{i\alpha}|b_i) = \frac{1}{2} \sum_{i=1}^n \{y_{ii} - \sum_{j=1}^q \beta_{ji\alpha} X_{ij}\}^2 G(m_i - m_\alpha|b_i) + P(\beta_{i\alpha}), \tag{3}$$

where $P(\beta_{i\alpha})$ is the recursive elastic net penalty, and

$$G(m_i - m_\alpha|b_i) = \exp\left\{-\frac{(m_i - m_\alpha)^2}{b_i}\right\} \tag{4}$$

is the Gaussian kernel function used to group samples according to specific cancer characteristics (i.e., modulator m_i for $i = 1, \dots, n$). The Gaussian kernel function enables us to estimate $\beta_{ji}(m_\alpha)$ for the α^{th} sample based only on samples having characteristics of m_i similar to the target sample modulator value m_α . Thus, we can construct a gene network for a specific clinical status of samples, and it leads to evidences for personalized therapy.

In the first stage, we measure cancer characteristics m_i ($i = 1, \dots, n$) of each sample and construct personalized gene regulatory networks under varying conditions of clinical characteristics using NetworkProfiler. This enables us to reveal patient-specific gene regulatory characteristics that are vital information of precision medicine. To comprehensively analyze the multiway interaction between genes, we proposed the use of the AI method in the second stage.

Stage 2: Extracting knowledge from the multilayer networks by explainable AI

The constructed multilayer networks (targets \times regulators \times samples) are considered as the input of the explainable AI method developed in our previous study, called TRIP [7]. The TRIP is a deep learning method for tensor decomposition. In this study, we consider a gene network matrix as a second-order tensor for a data point and then estimate projection matrices for the first and second modes based on the tensor decomposition. By using the projection matrices, we construct low-dimensional subspace of the network tensor. Thus, we can reduce dimensionality of large-scale multiple gene networks and extract crucial components to predict EMT-modulators.

For the K -mode tensor \mathcal{X} for size $I_1 \times \dots \times I_K$, the TRIP estimates a projection matrix $C^{(k)} \in \mathbb{R}^{I_k \times k}$ and then projects the network tensors onto the constructed subspace by using $C^{(k)}$. Prediction or classification is conducted on the constructed human-readable low-dimensional subspace. This leads to more explainable and interpretable results of the multilayer gene network analysis, as compared to the results on the complex high-dimensional data space.

The second stage of our strategy is based on the following two problems,

- Constructing human-readable low-dimensional subspace and projecting the network tensor onto the subspace,

$$\bar{\mathcal{X}}_i = \mathcal{X}_i \prod_k \times_k \mathbf{C}^{(k)} \tag{5}$$

- Predicting a response variable based on the projected network tensor $\bar{\mathcal{X}}_i$

$$\hat{y}_i = f(\bar{\mathcal{X}}_i, \boldsymbol{\theta}) \quad \text{for } i = 1, \dots, n, \tag{6}$$

where $\mathbf{r}_i = \mathcal{W}_i^T \text{vec}(\bar{\mathcal{X}}_i)$, \mathcal{W}_i is the weight tensor for prediction and $\boldsymbol{\theta}$ are the remaining parameters other than \mathcal{W}_i .

The optimization of the TRIP for the two aforementioned problems (i.e., projection of the network tensor and prediction of the response variable) is based on the following objective function

$$O_T = \frac{1}{n} \sum_{i=1}^n \{ \mathcal{L}(y_i, \hat{y}_i) + \gamma \| \mathcal{X}_i - \bar{\mathcal{X}}_i \prod_k \times_k \mathbf{C}^{(k)T} \|_2^2 \}, \tag{7}$$

subject to $\mathbf{C}^{(k)T} \mathbf{C}^{(k)} = \mathbf{I}$,

where $\gamma > 0$ is the tuning parameter for the projection error and $\mathcal{L}(y_i, \hat{y}_i)$ is the loss function of the prediction given in (6). The second term is the loss function for projection of the network tensor onto the constructed subspace given in (5). As shown in the objective function (7), the TRIP estimates the projection matrix $\mathbf{C}^{(k)}$ and simultaneously predicts the response variable. This implies that the TRIP enables us to achieve effective prediction results while retaining as much of the original data variance as possible.

The projection matrix $\mathbf{C}^{(k)}$ that satisfies the orthonormal condition (i.e., $\mathbf{C}^{(k)T} \mathbf{C}^{(k)} = \mathbf{I}$) is derived from singular value decomposition (SVD) of a latent variable $\mathbf{Z}^{(k)}$ of the same size as $\mathbf{C}^{(k)}$. That is, we first perform SVD of $\mathbf{Z}^{(k)}$,

$$\mathbf{Z}^{(k)} = \mathbf{P}^{(k)} \mathbf{S}^{(k)} \mathbf{Q}^{(k)T}, \tag{8}$$

and then set

$$\mathbf{C}^{(k)} = \mathbf{P}^{(k)} \mathbf{Q}^{(k)T}. \tag{9}$$

The latent variable $\mathbf{Z}^{(k)}$ is estimated from the derivatives of the objective function O_T by setting them to zero. Maruhashi et al. [7] showed that the derivatives of the objective function O_T with respect to $\mathbf{Z}^{(k)}$ are derived from a function of $\partial O_T / \partial \mathbf{C}^{(k)}$ and SVD of $\mathbf{Z}^{(k)}$, and proposed iterative algorithm for optimization problem of the TRIP given in (7). That is, the optimization problem of the TRIP is based on simultaneously taking the derivatives of the objective function O_T with respect to $\mathbf{C}^{(k)}$, $\mathbf{Z}^{(k)}$, \mathbf{r}_i , and $\boldsymbol{\theta}$ and setting them to zero.

The projection matrices are considered as the regression coefficients of the crucial components. They learnt a multi-linear surrogate model $\hat{y}_i = \langle \mathcal{W}, \bar{\mathcal{X}}_i \rangle + b$ that minimizes the sum of the differences with the results of the prediction model $\sum_i \| \hat{y}_i - \hat{y}'_i \|_2^2$, by setting the regression coefficient \mathcal{W} to be a multi-linear tensor, i.e., $\mathcal{W} = \mathbf{g}^{(1)} \circ \dots \circ \mathbf{g}^{(K)}$, where \circ denotes outer product. The crucial components in prediction of response variable are extracted by principal

component analysis of the vectors

$$\mathbf{u}_i^{(k)} = \mathcal{X}_i \prod_{l \neq k} \mathbf{C}^{(l)} \mathbf{g}^{(l)} \times \mathbf{C}^{(k)} \quad (10)$$

That is, the principal components of $\mathbf{U}^{(k)}$ having the normalized $\mathbf{u}_i^{(k)}$ as the column vector are extracted as crucial components for prediction of the response. For details of the notation and algorithm for the TRIP, please refer to Maruhashi et al. [7].

In the second stage, we perform integrative gene network analysis based on the constructed subspace of the network tensors.

Results

To illustrate our strategy, we apply the proposed method to personalized gene regulatory networks varying depending on the EMT process [1]. Because the EMT modulator is uniformly distributed, we consider EMT-related networks constructed by ordinary NetworkProfiler instead of the robust NetworkProfiler.

Personalized gene regulatory networks under varying conditions of the EMT process

EMT-related gene networks were constructed based on the expression profiles of 762 cell lines from the Sanger Cell Line Project (<http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi>). A 13,508 (genes) \times 762 (cancer cell lines) gene expression matrix was constructed based on the expression profiles of 13,006 mRNAs from Affymetrix GeneChip Human Genome U133 Array set and 502 microRNAs from bead-based oligonucleotide arrays. A total of 1732 regulator genes consisting of 1183 transcription factors, 47 nuclear receptors, and 502 human miRNA were extracted. The modulator describing the EMT process (epithelial-like or mesenchymal-like) of cell lines, the so-called EMT modulator, was extracted using the module discovery method [11] based on 50 genes labeled as the EMT-related genes (i.e., EMT-UP, EMT-DN, JECHLIN-GER-EMT-UP, and JECHLIN-GER-EMT-DN) in Molecular Signatures Database v2.5 (<http://www.broadinstitute.org/gsea/msigdb/index.jsp>). In short, 762 EMT-related gene regulatory networks for 13,508 targets with 1732 regulator genes were constructed under varying conditions of the EMT modulator values. In this study, we consider the network for 13,508 targets \times 1762 regulators as a second-order tensor for a specific EMT process and apply the TRIP for integrative gene network analysis of the 762 network tensors.

Comprehensive interpretation of the EMT-related network tensors

We learn a 50 \times 50 subspace of the 762 EMT-related gene networks using the TRIP, i.e., $J_1 = 50$ and $J_2 = 50$. From the projection matrices of the constructed subspace $\mathbf{C}^{(k)}$ $k = 1, 2$ in (9), each 50 crucial factors describing importance of target and regulator genes to predict EMT-modulator are extracted by PCA of $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$, respectively. In this study, we consider the crucial components of the subspace for regulator genes, i.e., $\mathbf{U}^{(2)}$. Fig 1 shows the variabilities explained by the extracted 50 independent components. As shown in Fig 1, the first component explains more than half of the total variability (i.e., 56%) of the EMT-related gene networks, and the first three components explain approximately 70% of the variability. We focus on the first three components and interpret the EMT-related gene networks based on the three components.

Table 1 shows the distribution of the tissue origin of 100 cell lines corresponding to the 100 highest and lowest values of each of the three components. The brain cell lines are

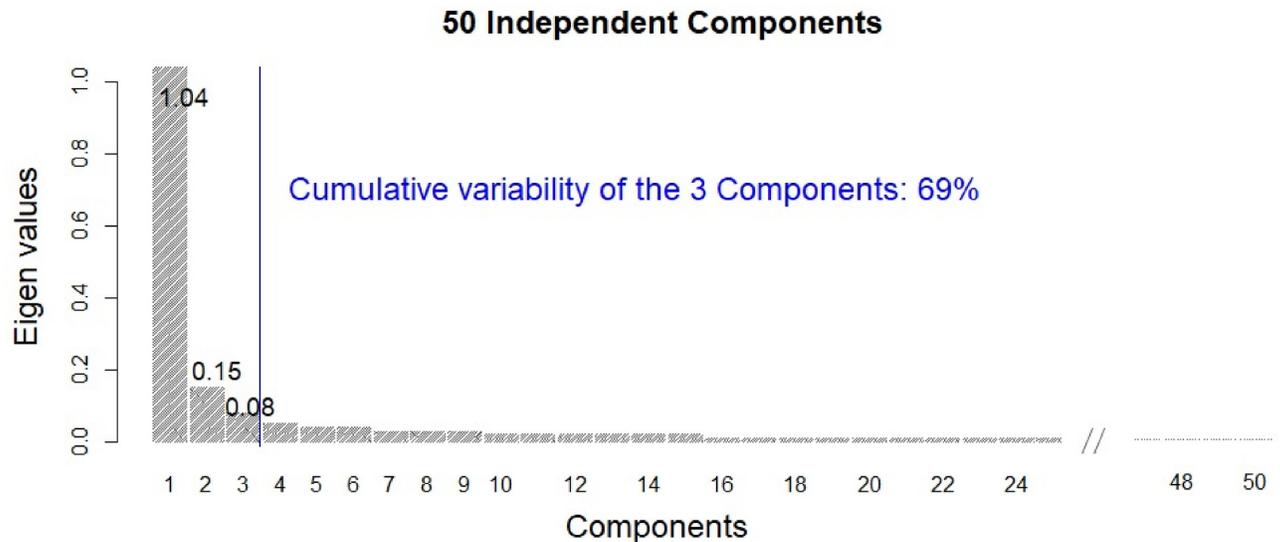


Fig 1. Variabilities of the 50 independent components.

<https://doi.org/10.1371/journal.pone.0241508.g001>

concentrated in the region with high values of the first component (i.e., 83% of brain cell lines are in the region with the 100 highest values of component 1), and leukemia cell lines are concentrated in the region with high values of the second component (i.e., 83% of leukemia cell lines are in the region with the highest values of component 2).

Colorectal cell lines are concentrated in the region with low values of the first two components. In short, the first and second components can be characterized by the ‘Brain with Colorectal’ and ‘Leukemia with Colorectal’, respectively. An independent test (i.e., Chi-squared test) between high/low values of each component (e.g., leukemia cells/nonleukemia cells and high/low values of each component) is conducted, and significant diseases ($p < .01$) for each component are given in the column χ^2 in Table 1, where high and low regions are classified based on the median value of each component. The brain seemed to be most associated with the first component, whereas Leukemia can be considered as the most significant disease for both the second and third components. The Brain, Leukemia and Colorectal seem to be crucial diseases for all three components. Fig 2 shows the scatter plot of the EMT modulator and extracted three components. As shown in Fig 2, the first component is strongly associated with the EMT modulator (i.e., the correlation coefficient of the EMT modulator and the first component is 0.997). This indicates that the extracted first component can be considered as another EMT modulator, and the second and third components may have information other than the EMT-related mechanism.

In order to globally explore the EMT-related mechanisms from the multilayer networks, we combine our results with the well-known EMT markers (i.e., five EMT transcription factors (EMT-TFs): ZEB1, ZEB2, SNAIL1, SNAIL2, and TWIST1, see Table 2). For each component, the target networks of the five EMT-TFs are constructed in the region of high and low values of each component. From the multiple networks in each region, we first extract target genes (TG.EMT-TFs) of the EMT-TFs and then extracted the target genes of the TG.EMT-TFs, where the genes are extracted as targets in at least one network are considered as target genes. The target networks for each region are given as binary adjacency matrices (1: two genes are associated in at least one cell line, 0: otherwise). Next, we compute absolute differences of the two adjacency matrices for genes, then extract 10 genes showing considerably

Table 1. Distribution of diseases for cells corresponding the 100 highest and lowest three components (%).

	Component 1			Component 2			Component 3			Total 762 cells (#)
	High	Low	χ^2	High	Low	χ^2	High	Low	χ^2	
AdrenalGland	0	0		100	0		0	0		1
AutoGanglia	5	0	****	3	0		0	54	***	37
BiliaryTrack	17	17		0	33		0	0		6
Bone	29	0	****	3	6		10	13		31
Brain	83	0	****	0	17	****	41	5	****	59
Breast	7	26	***	10	17	**	12	10		42
Cervix	0	43		21	36		7	14		14
Colorectal	0	63	****	5	63	***	8	5	****	38
Endometrium	9	9		0	9		0	18		11
Eye	0	0		100	0		0	0		2
Headneck	4	13	***	0	8		33	0		24
Kidney	9	0	****	0	0		5	32		22
Leukemia	0	1	****	50	1	****	16	0	****	109
Liver	18	0		18	0		9	9		11
Lung	2	15	**	8	12		9	15	**	128
Lymphoma	0	0		32	0	****	16	0		19
Muscle	0	0		0	13		0	63		8
Oesophagus	0	38	***	0	33	***	4	8		24
OtherSarcoma	70	0	**	0		0	20	10		10
Ovary	4	0		33	0		17	21		24
Pancreas	0	41		6	29		12	0		17
Placenta	0	0		0	0		0	0		2
Pleura	33	0		0	0		0	17		6
Prostate	0	17		0	17		17	0		6
Skin	26	2	***	0	7		7	33		46
SmoothMuscle	0	0		0	0		0	100		1
StomachGI	0	28		12	24		12	8		25
Testis	0	0		0	0		50	0		4
Thyroid	42	8		0	8		8	0		12
Unknown	0	0		100	0		0	0		1
UrinaryTrack	5	40		0	30		10	15		20
Vulva	0	0		0	0		50	50		2

Significant of χ^2 test (p-value):** $p < .01$;*** $p < .001$;**** $p < .0001$ <https://doi.org/10.1371/journal.pone.0241508.t001>

different edges between the two adjacency matrices for high and low regions of each component (i.e., for each genes, sum of the absolute differences are computed for all genes in the EMT target network, and then 10 genes having the largest sum of the absolute differences are extracted). The identified 10 genes can be considered as markers having specific regulatory characteristics depending on the each component.

Table 3 shows the identified genes for the three components and their evidence sources. We focus only on the newly identified genes other than the five EMT-TFs, even though the five EMT-TFs are selected for all three components.

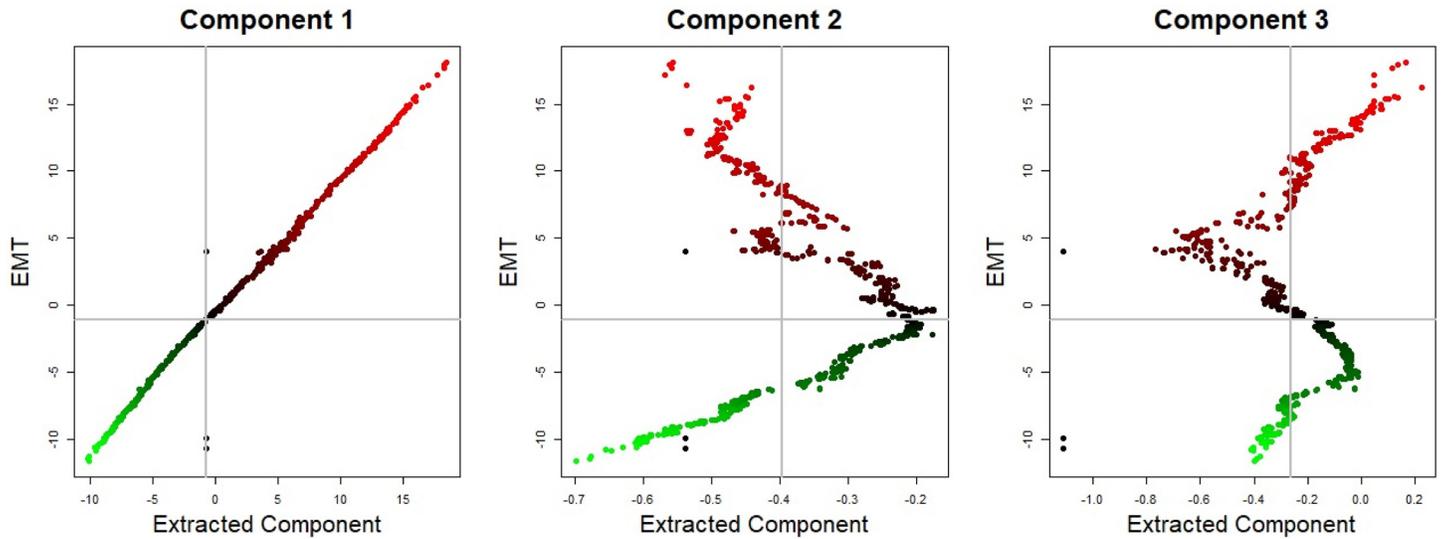


Fig 2. EMT modulator and top three components (Green: low EMT modulator, red: high EMT modulator). The first three components seem to be independent of each other, i.e., each component may have different information to describe EMT-related gene networks.

<https://doi.org/10.1371/journal.pone.0241508.g002>

Among the newly identified 17 genes, only IRF6 is used for defining the EMT modulator. For the identified genes, we compute the regulatory effect change (REC) for 13,508 target genes [1, 5]. In this study, we consider direct connection between genes, i.e., in the 762 EMT-related gene regulatory networks, the effect of the j^{th} regulator on the l^{th} target gene at the α^{th} sample can be measured by the following regulatory effect (RE),

$$RE_{jl\alpha} = \hat{\beta}_{jl}(m_\alpha) \cdot X_{\alpha j}, \tag{11}$$

and the RECs according to the EMT modulator values are computed as follows

$$REC_{jl} = \max\{RE_{jl\alpha}; \alpha = 1, \dots, n\} - \min\{RE_{jl\alpha}; \alpha = 1, \dots, n\}. \tag{12}$$

which measures the effect of the EMT modulator on strength of the relationship between the regulator and the target genes.

Table 2. EMT Transcription Factors (TFs) and their mechanisms.

EMT-TFs	EMT-related mechanism	Evidences
ZEB family: ZEB1, ZEB2	: Snail1 and Twist1 can up-regulate and cooperate with ZEB1 to induce EMT. : MYC-or serum-induced EMT were characterized by increased expression of ZEB1, ZEB2, and SNAIL1.	[3, 12]
SNAIL1	: mediator in different signaling pathways that induce EMT, such as the NBS1-SNAIL1 axis and the TGF- β /SMADS/HMGA2/SNAIL1 axis. : main role in EMT, the process by which epithelial cells acquire a migratory, mesenchymal phenotype as a result of its repression of E-cadherin.	[13, 14]
SNAIL2	: up-regulated by Notch to induce EMT with an increase of cell migration and loss of cell-cell junctions. : activates ZEB1 and cooperates with it to promote EMT. : direct induction of SNAIL1 is essential for TWIST1 to induce EMT.	[4, 15, 16]
TWIST1	: HIF-1 α directly binds to the promoter of TWIST1 to induce EMT in hypoxic microenvironments. : needs to induce SNAIL1 to suppress the epithelial branch of the EMT program. : acts together with SNAIL1 to promote EMT.	[4, 17]

<https://doi.org/10.1371/journal.pone.0241508.t002>

Table 3. Identified novel candidate markers involved in EMT related mechanism.

Genes	Components	Related diseases	Reference
AFF1	3	-	
ANKRD5	1	Brain, Pharynx, and Swim bladder cancers	[18]
FOXF1	2	Colorectal and Lung cancers	[19, 20]
FOXF2	1	Basal-like breast cancer	[21, 22]
GLI3	2	Oral squamous cell carcinoma and Colorectal cancer	[23, 24]
GRHL2	1	Gastric, Ovarian, and Breast cancers	[13, 25–28]
IFI16	2, 3	Prostate Cancer	[29–32]
IRF6	3	Embryonic palate, Breast, Gastric, and Prostate cancers	[33, 34]
KANK2	3	-	
LSR	2	Endometrial and Breast cancers as well as Head and Neck Squamous Cell Carcinomas	[35–38]
MAFB	3	-	
OVOL2	2	Colorectal tumor, Osteosarcoma, and Breast and Prostate cancers	[39–42]
PCBD1	2	-	
SOX13	2	Colorectal cancer	[43]
TGFB11	2	Lung adenocarcinoma, Kidney disease, and Renal fibrosis	[44–47]
TP63	2, 3	Breast cancer	[48–50]
ZNF91	1	-	

<https://doi.org/10.1371/journal.pone.0241508.t003>

As mentioned above, each 762 cell lines have EMT modulator values and corresponding networks. Fig 3 shows the average of RE of the identified genes other on the 50 EMT-related genes for 10 networks corresponding to the 10 highest (top) and lowest (middle) EMT modulator values, where edges having zero mean for 10 cell lines are deleted. As shown in Fig 3, FOXF1 regulates the EMT-related genes, COL6A1, COL6A2, HTRA1, IL11, MMP2, PCOLCE, PDGFRA, PDGFRB, and PMP22. The regulation system of FOXF1 with the EMT-related genes is observed in both epithelial-like and mesenchymal-like cell lines. E-cadherin (CDH1), which is one of the important genes for cell–cell adhesion in epithelial cells, is positively regulated by LSR and GRLH2 and negatively regulated by the ZEB family (i.e., ZEB1 and ZEB2). We then focus on genes not discovered as EMT markers in the literature, but selected by our method, AFF1, KANK2, PCBD1, ZNF91, and MAFB, i.e., their EMT-related mechanisms have not yet been revealed. Although a majority of the unrevealed genes regulate the EMT-related genes, their REs are extremely small. From these results, it can be considered that the genes may have mechanisms that are not directly involved in the EMT process.

For the selected 10 genes from the analysis of the three components, we compute REC matrices consisting of 10 columns (extracted 10 genes) and 13,493 rows (13,493 target genes) and then perform PCA of the REC matrix. The identified genes for each component are grouped in the first two PC spaces, and the reciprocal mechanisms between the grouped genes can be found in the published literature. Overall flowchart of our strategy for global exploration of EMT-related networks is given in Fig 4.

Component 1:

High.E&High.C1 versus Low.E&Low.C1

We consider high EMT&high component 1 (High.E&High.C1) and Low EMT&Low component 1 (Low.E&Low.C1) regions consisting of 380 and 379 cell lines, respectively. For the first component, ZEB2, SNAI2, ZEB1, SNAI1, TWIST1, MAFB, GRHL2, ANKRD5, FOXF2, and ZNF91 are selected as markers having specific characteristics depending on the first component (i.e., these genes show significantly different regulatory systems between the High.E&High.C1 and Low.E&Low.C1 regions). In addition to the EMT-TFs, the novel five genes

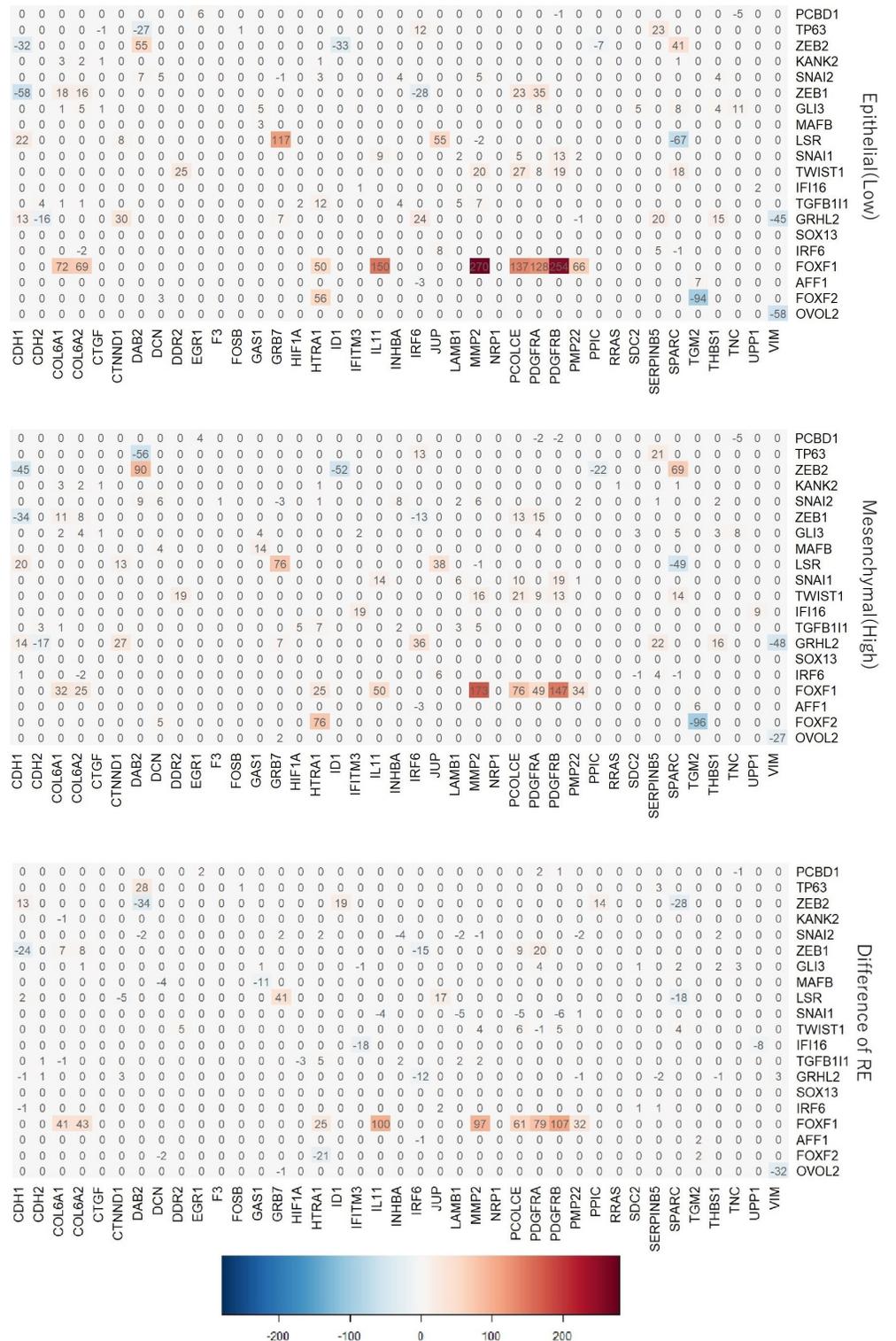


Fig 3. HeatMap of the regulatory effect of the selected genes on the 50 EMT-related genes.

<https://doi.org/10.1371/journal.pone.0241508.g003>

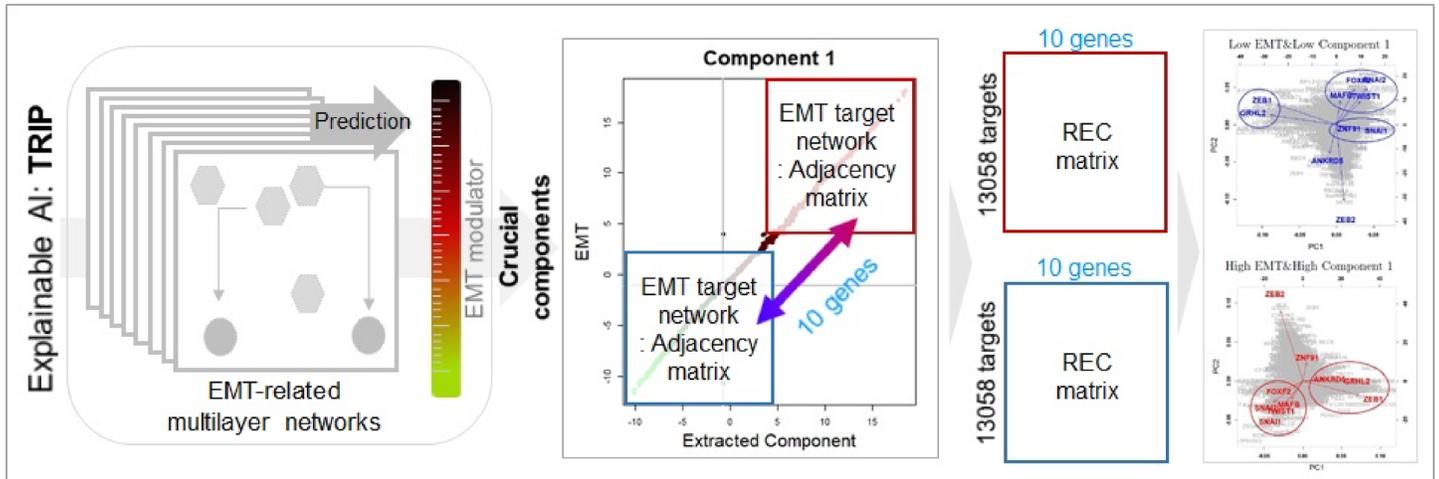


Fig 4. Overall flowchart of our strategy.

<https://doi.org/10.1371/journal.pone.0241508.g004>

are identified, i.e., MAFB, GRHL2, ANKRD5, FOXF2, and ZNF91. Fig 5 shows the target networks of the five genes. A relatively sparse network can be seen in the Low.E&Low.C1 region, as compared to the High.E&High.C1 region. That is, only five EMT-TFs have a strong relationship between each other in Low.E&Low.C1, whereas there are many relationships between not only the identified 10 genes, but also their target and regulator genes in the High.E&High.C1 region. As shown in Table 3, more than half of the identified genes are confirmed as EMT markers (i.e., their EMT-related mechanisms have been reported in the literature). For

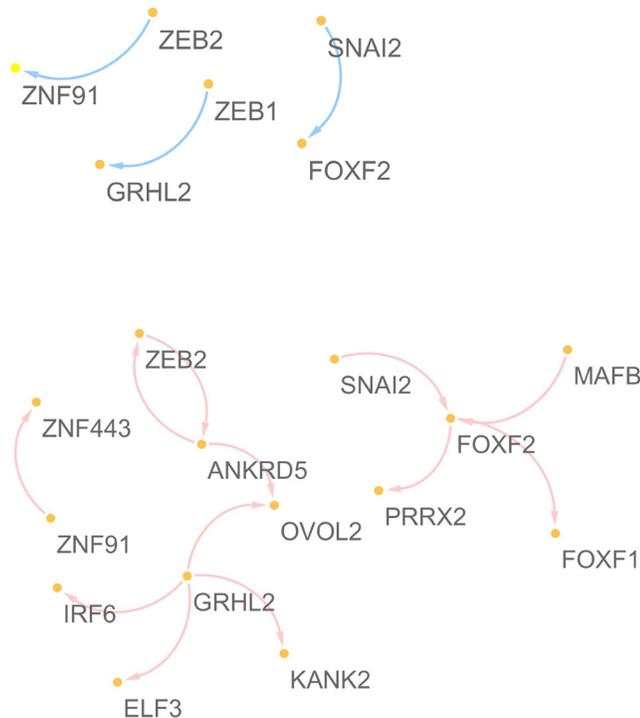


Fig 5. Target networks of the five regulators discovered from the analysis of the first component.

<https://doi.org/10.1371/journal.pone.0241508.g005>

instance, GRHL2 reduces the invasion and migration through the inhibition of TGF- β -induced EMT in gastric cancer [26]; GRHL2 leads to mesenchymal–epithelial transition (MET) through the inhibition of ZEB1 [13]. ANKRD5 plays roles in protocadherin-mediated cell protrusion and adhesion, and participates in cell adhesion [18]. FOXF2 was identified as a novel TF related to EMT-suppressing in basal-like breast cancer (BLBC) FOXF2 negatively targets TWIST1 in the EMT programming and metastasis progress of BLBC [22].

Fig 6 shows the projected 10 genes and 13,493 target genes on the first two PC spaces of the REC matrix. From the biplot, the identified genes can be grouped as follows,

High.E&High.C1 region

- Group 1: ZEB1, ANKRD5, GRHL2
- Group 2: FOXF2, MAFB, SNAL1, SNAL2

Low.E&Low.C1 region

- Group 1: ZEB1, GRHL2
- Group 2: FOXF2, MAFB, TWIST1
- Group 3: ZNF91, SNAL1

We focus only on the newly discovered five genes because the reciprocal mechanisms between the five EMT-TFs were well known in previous studies. For both high and low regions of the first component, GRHL2 is grouped with well-known EMT marker ZEB1. Their reciprocal mechanisms in the EMT-related process are demonstrated as follows [25, 27, 28]: GRHL2 suppresses EMT and restores sensitivity to anoikis by repressing ZEB1 expression; Combination of TGF- β and Wnt activation represses GRHL2 expression by direct interaction of ZEB1 with the GRHL2 promoter, inducing EMT; Reciprocal feedback loop between GRHL2 and ZEB1 controls epithelial versus mesenchymal phenotypes and EMT-driven tumor progression; GRHL2 is the main gatekeeper of EMT in EOC via miR-200-ZEB1, and their axis forms the core of EMT signaling. The EMT-related interaction between FOXC2 and TWIST1 is also demonstrated as follows [21, 22]: FOXC2 transcriptionally represses the expression of two EMT-TFs TWIST1 and FOXC2; FOXC2 negatively targets TWIST1 in the EMT programming and metastasis progress of BLBC. Genes MAFB and ZNF91 are members of Gene-Ontology-Terms Class: GO:0006355—regulation of transcription, DNA-templated. However, the interaction between MAFB and ZNF91 in diseases has not been yet demonstrated.

Component 2:

High.E&High.C2 vs High.E&Low.C2

Low.E&High.C2 vs Low.E&Low.C2

For the second component, we consider regions of high and low values of the second component for high and low EMT modulators, respectively (i.e., “High.E&High.C2 (195 cell lines) vs High.E&Low.C2 (186 cell lines)” and “Low.E&High.C2 (186 cell lines) vs Low.E&Low.C2 (195 cell lines)”). In addition to the well known five EMT-TFs, GLI3, ANKRD5, PCBD1, FOXF1, OVOL2 and LSR, TP63, SOX13, IFI16, and TGFB11 are identified for low and high EMT regions, respectively.

Fig 7 shows target networks of the newly identified five markers other than 5 EMT-TFs, where the edges appeared in each network of all cell-lines are only extracted. Similar to the component 1, a relatively sparse network can be seen in the low regions of component 2, as compared to the high regions. Especially in the low EMT region, it can be seen that the target network consists of only well known EMT markers, ZEB1, SNAI1, SNAI2, OVOL2, FOX1 and

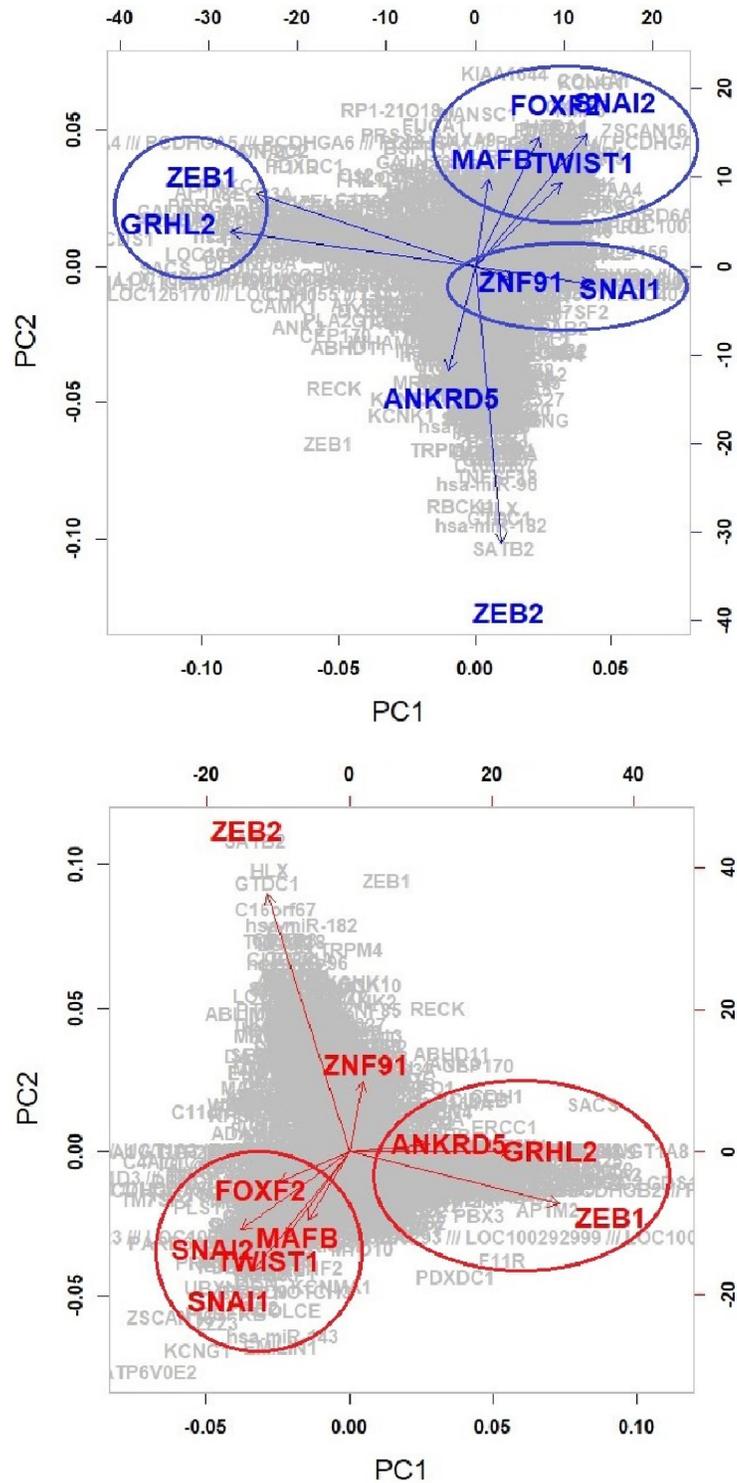


Fig 6. Projection of the discovered genes onto the first two principal components of REC.

<https://doi.org/10.1371/journal.pone.0241508.g006>

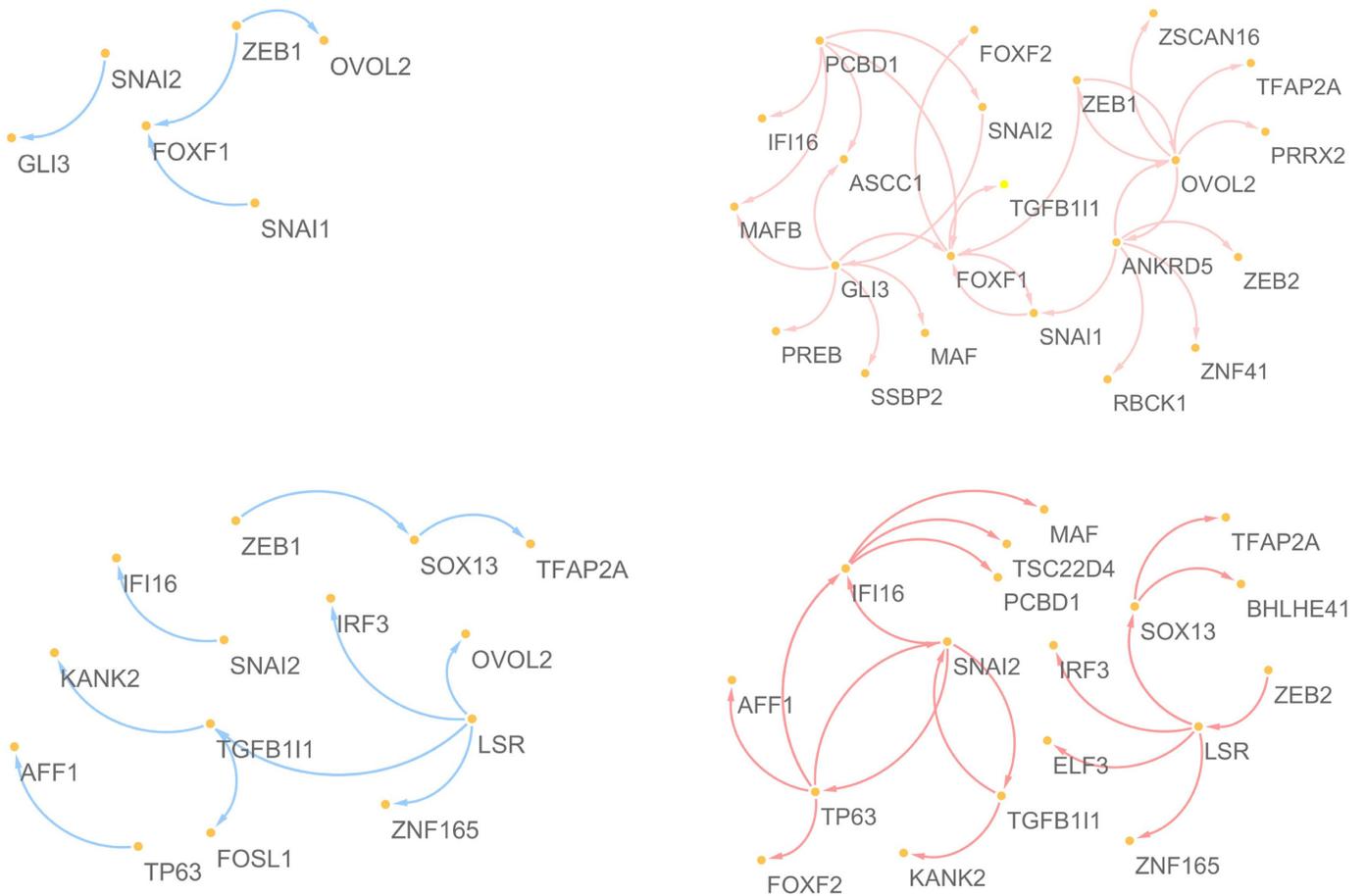


Fig 7. Target networks of the five regulators discovered from the analysis of the second component for high and low EMT regions.

<https://doi.org/10.1371/journal.pone.0241508.g007>

GLI3 in region for the low component 2, whereas the target network for the high component 2 involves many genes. Fig 8 shows the results of PCA for the REC matrices. We focus on the results of the low EMT region (i.e., Low.E&High.C2 versus Low.E&Low.C2) and group the genes as follows,

ZEB1 and OVOL2: OVOL2 is one of the well-known EMT markers, and the interaction in the EMT process of OVOL2 and ZEB1 has been demonstrated in many studies: OVOL-TFs control MET through a regulatory feedback loop with EMT-inducing TF, ZEB1 [41]; OVOL2 restricts EMT by directly inhibiting EMT-inducing factors including the ZEB1 system; A regulatory network containing OVOL2–ZEB1 mutual repression results in a four-state EMT, i.e., epithelial, intermediate, intermediate, and mesenchymal states [42]; OVOL2 suppresses ZEB1 expression by binding to the ZEB1 promoter [39].

FOXF1 and SNAI1: FOXF1 is also confirmed as an EMT-related marker, and reciprocal mechanisms of genes in this group have been demonstrated: the expression of FOXF1 inhibits cancer cell invasion and migration, whereas the inactivation of FOXF1 stimulates cell invasion and migration (Wei et al., 2014); higher level of FOXF1 is positively associated with enrichment of EMT gene signatures [19]; Overexpression of FOXF1 induces EMT by transcriptionally activating SNAI1 in colorectal cancer metastasis [19].

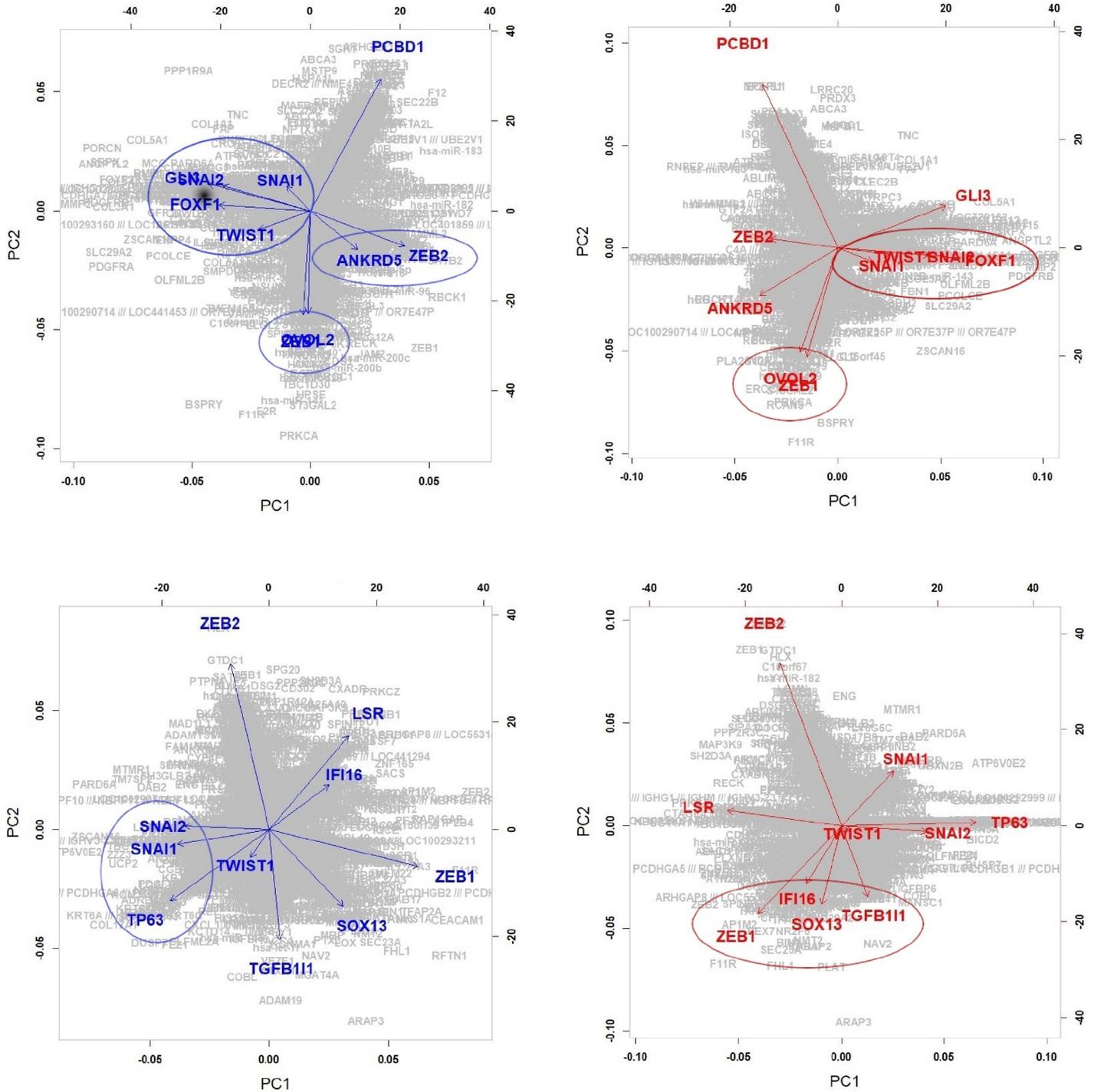


Fig 8. Projection of the discovered genes onto the first two principal components of REC: Component 2.

<https://doi.org/10.1371/journal.pone.0241508.g008>

GLI3 and SNAI2: GLI3 is one of the zinc finger protein and well-known marker of Sonic Hedgehog, and interaction between GLI3 and SNAI2 has been demonstrated: shRNA-GLI3-transfected cells were associated with the decreased expression of stem cell- and EMT-related genes (CD44, BMI1, POU5F1, and SNAI2) [24].

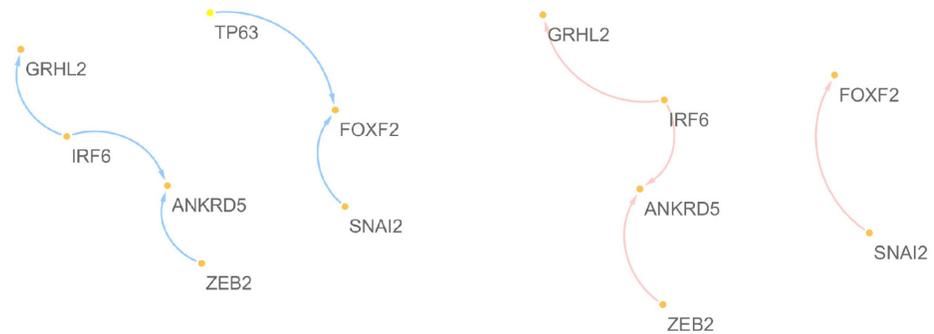


Fig 9. Target networks of the five regulators discovered from the analysis of the third component.

<https://doi.org/10.1371/journal.pone.0241508.g009>

Component 3:

High.E&High.C3 vs High.E&Low.C3

For the third component, we consider High.E&High.C3 and High.E&Low.C3 regions consisting of 154 and 227 cells, respectively. TP63, IFI16, AFF1, IRF6, and KANK2, other than EMT-TFs, show significantly different edges between the High.E&High.C3 and High.E&Low.C3 regions, where TP63 and IFI16 are also identified as crucial genes in the analysis of the second component. IRF6 is one of the 50 EMT-related genes defining the EMT modulator. The EMT-related mechanism of IRF6 has been demonstrated in previous studies; especially, there are many studies on the association between well-known EMT markers with IRF6: ectopic expression of IRF6 increases the expression of SNAI2 and diminishes the expression of various epithelial markers (e.g., E-cadherin) in EMT; TGF β 3 increases IRF6 expression, and IRF6 appears to regulate EMT during palatal fusion via SNAI2 [33]; IRF6 is downregulated during the EMT process of breast cancer and prostate cancer [34]. Although EMT-related mechanisms of KANK2 have not yet been revealed, it has been demonstrated that KANK2 concentrates around most mature focal adhesions and binds talin in migrating cells [51]. AFF1 is known as the mixed lineage leukemia fusion-associated gene and plays a role in osteogenic differentiation of human mesenchymal stem cells [52–54]. However, EMT-related mechanisms of AFF1 have not yet been demonstrated.

The target networks of the newly identified five genes are presented in Fig 9. There are no significant differences between the two regions, except for the association between TP63 and FOXF2. It can be considered that the target networks related to the third component may be dominated by the mechanism of the high EMT region, and only regulating FOXF2 by TP63 can be considered as a specific characteristic related to component 3.

Fig 10 shows the projected 10 regulators and target genes on the first two PC spaces of REC. We group the genes as follows,

Group 1: IRF6 and TP63

IRF6 regulated by TP63 plays a tumor suppressor role in squamous cell carcinomas through a Notch-dependent mechanism, which plays critical roles in EMT pathway [34].

Group 2: AFF1 and SNAI1, SNAI2, TWIST1

Group 3: KANK2 and ZEB1, ZEB2

The reciprocal mechanisms of genes in groups 2 and 3 have not yet been demonstrated.

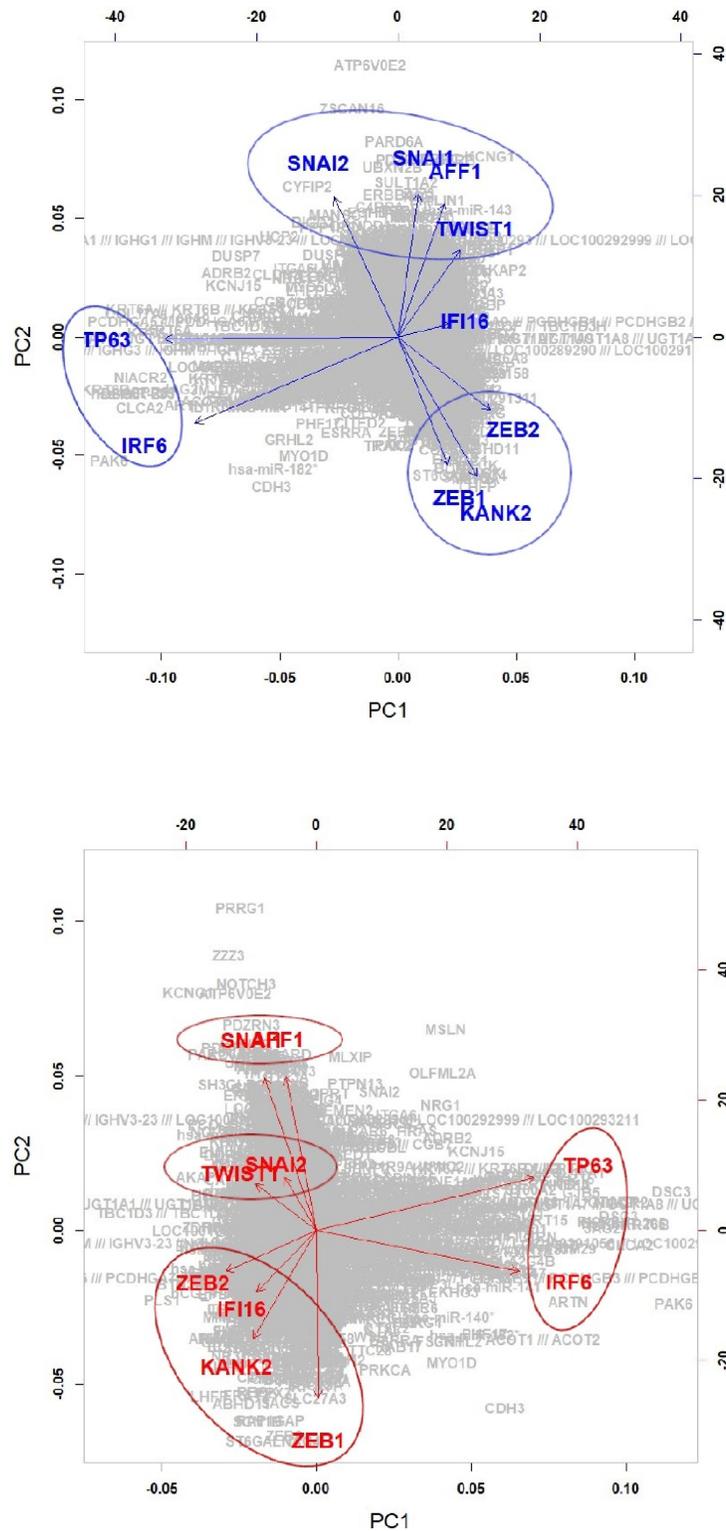


Fig 10. Projection of the discovered genes onto the first two principal components of REC: Component 3.

<https://doi.org/10.1371/journal.pone.0241508.g010>

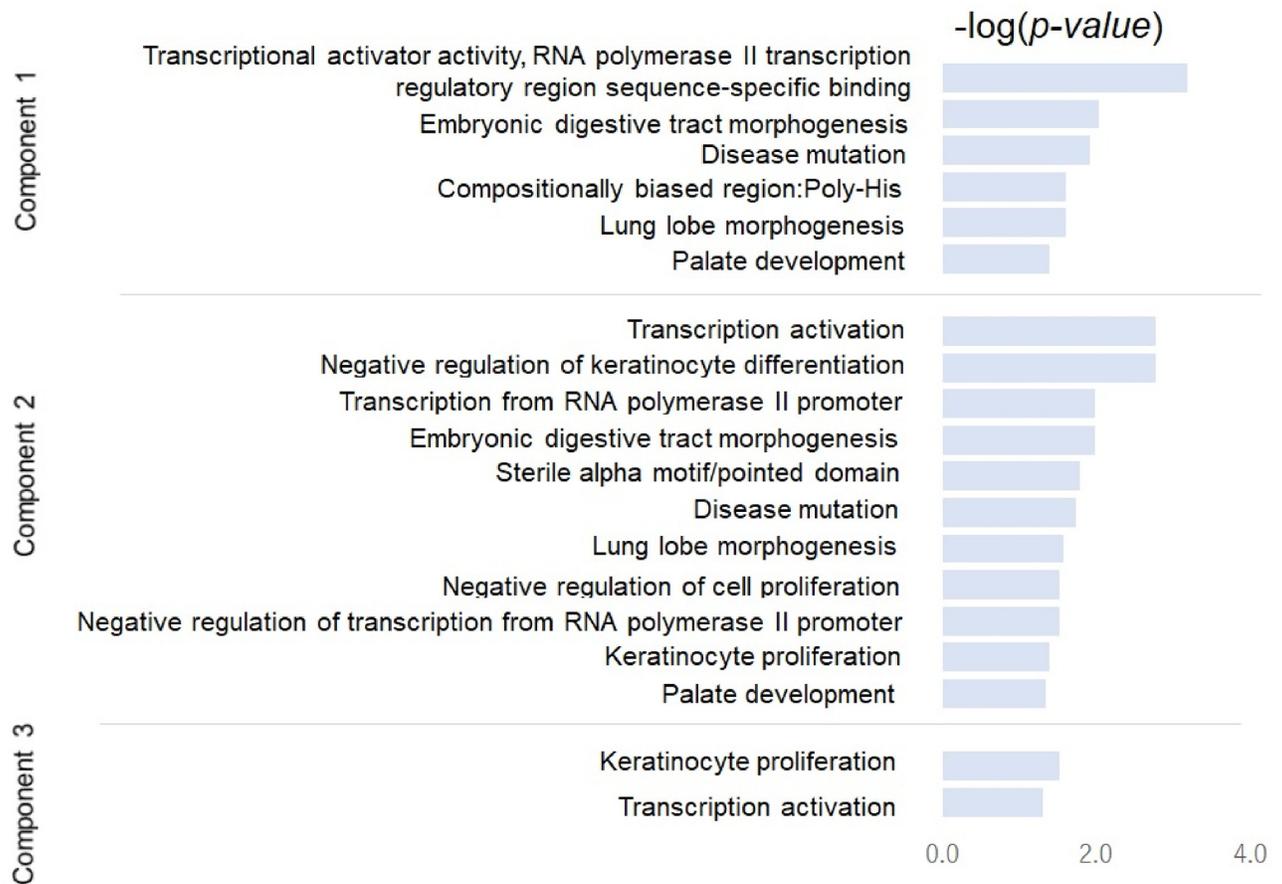


Fig 11. Gene enrichment analysis of the identified crucial genes from the REC matrix.

<https://doi.org/10.1371/journal.pone.0241508.g011>

Functional enrichment analysis based on the bioinformatics tool DAVID

To identify biological processes involved in the extracted components, we performed gene enrichment analysis using the bioinformatics tool Database for Annotation, Visualization and Integrated Discovery (DAVID) [55]. We used all regulatory genes as a background and performed functional enrichment analysis for 17 genes (1% of the regulatory genes) other than five EMT-TFs showing different regulatory systems in the high and low regions of each component. Fig 11 shows the functional annotation chart ($p < .05$) and the corresponding p-value (i.e., $-\log(p\text{-value})$). For the first component, six clusters are found, where the most significant cluster corresponding to the lowest p-value (i.e., the highest score for enrichment) is “*transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding*” grouping genes OVOL2, MAFB, FOXF1, FOXF2, GRHL2, and ALX1. The “*RNA polymerase II transcription regulatory region sequence-specific binding*” is a GO annotation of ZEB1. “*Embryonic digestive tract morphogenesis*”, “*disease mutation*”, “*lung lobe morphogenesis*”, and “*palate development*” are common factors enriched in the first and second components. Except for “*disease mutation*”, the three clusters are the GO functional terms (<http://www.informatics.jax.org/>),

- Embryonic digestive tract morphogenesis (GO:0048566): the anatomical structures of the digestive tract are generated and organized during embryonic development. The digestive tract is the anatomical structure through which food passes and is processed.

- Lung lobe morphogenesis (GO:0060463): The process in which the anatomical structures of a lung lobe are generated and organized. A lung lobe is a projection that extends from the lung.
- Palate development (GO:0060021: Roof of mouth development): The biological process whose specific outcome is the progression of the roof of the mouth from an initial condition to its mature state. This process begins with the formation of the structure and ends with the mature structure. The roof of the mouth is the partition that separates the nasal and oral cavities.

We focus on *sterile alpha motif/pointed domain* (grouping FLI1, ELF3, and TP63), which is involved in interactions with proteins, DNA and RNA. In a previous study, it was demonstrated that sterile alpha motif-pointed domain containing ETS TF (SPDEF) negatively regulates CCL2 and the EMT markers in prostate cancer cells, and the interaction between SPDEF and CDH1 (E-cadherin) related to the EMT process was also demonstrated: decreased SPDEF levels significantly induce CCL2 and CDH2 (N-cadherin), and decrease CDH1 (E-cadherin) mRNA and protein expression, confirming the association between SPDEF inhibition and EMT in cells. [56]. Although only ELF3 was demonstrated as a crucial regulator of E-cadherin in Shimamura et al., [1], interactions FLI1 and TP63 with E-cadherin are also verified as follows: Cav1-Snail-E-cadherin pathway plays a central role in the expression of the oncogenic transformation functions of fusion gene EWS/FLI1 [57]; reactivation of $\Delta Np63a$ is linked to the maintenance of epithelial markers and suggests that E-cadherin has a dual role in lung squamous cell carcinoma [58].

For the second and third components, “transcription activation” (grouping TP63, TGFBI1, and GRHL2) and “keratinocyte proliferation” (grouping IRF6 and TP63) are the most enriched terms, respectively. *keratinocyte* is frequently reported to be related to the second and third components in Fig 11. *keratinocyte* plays a role in cell–cell adhesion [59, 60]. We found that gene TP63, which is known as a keratinocyte TF, plays a key role in the EMT process through interaction with the well-known EMT markers, i.e., TGF- β , GRHL2, and miR-200n family: Ectopic $\Delta Np63a$ expression in normal human epidermal keratinocytes yields the EMT phenotype in a TGF- β -dependent manner; Knockdown of all isoforms of p63 leads to the EMT phenotype through loss of GRHL2 and miR-200 family genes [61].

Furthermore, it has been found that IRF6 (target of TP63) is induced by the NOTCH signaling pathway, which plays vital roles in the development and progression of cancers through regulating ZEB1 expression and EMT pathway, in breast cancer and keratinocytes [34]. Interestingly, the cluster consisting of IRF6 and TP63 related to keratinocyte proliferation was also identified by PCA of REC for the third component (see Fig 8). In addition, it was shown that the interaction of TP63 and FOXF2 is uniquely different in the target networks for the high and low regions of the third component (see Fig 7). In short, the third component and a part of the second component of the EMT-related networks can be described as a keratinocyte-related factor and the interaction of TP63 and IRF6, which may play a vital role in EMT in keratinocytes.

Discussion

We introduced a novel methodology for a comprehensive analysis of large-scale personalized network tensors. In this study, we considered a gene regulatory network as a tensor for a data point, and decompose the multilayer networks represented as tensors by using an AI approach, TRIP. Unlike existing studies for sample-specific gene network construction, our strategy analyzes whole multilayer networks based on tensor decomposition, thus we can perform wide

exploration of the large-scale gene regulatory networks for all patients. To illustrate our method, we applied it to personalized networks constructed for 762 cell lines having varying conditions of the EMT process. We identified novel candidate markers and verified biological mechanisms of a majority of the identified markers based on the literature. Although most of the identified markers were found in previous studies, some of the revealed genes could not be verified. Further work is required towards experimental validation of the newly revealed markers. In this study, our strategy was illustrated based only on EMT-related gene regulatory networks. As one of our future works, we consider the comprehensive analysis of dynamic systems of personalized gene networks in accordance with various clinical characteristics (e.g., drug sensitivity of cell lines).

Global exploration of multilayer networks along with clinical characteristics of a patient provides crucial information for evidence-based personalized medicine. We expect that the proposed strategy based on an explainable AI approach will provide a novel insight into network biology.

Acknowledgments

The authors thank the associate editor and anonymous reviewers for the constructive and valuable comments that improved the quality of the paper considerably. This research used computational resources of the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo.

Author Contributions

Data curation: Heewon Park.

Formal analysis: Heewon Park, Koji Maruhashi, Rui Yamaguchi.

Methodology: Heewon Park, Koji Maruhashi, Rui Yamaguchi.

Project administration: Satoru Miyano.

Supervision: Seiya Imoto, Satoru Miyano.

Writing – original draft: Heewon Park.

Writing – review & editing: Koji Maruhashi, Rui Yamaguchi, Seiya Imoto.

References

1. Shimamura T, Imoto S, Shimada Y, Hosono Y, Niida A, et al. (2011) A novel network profiling analysis reveals system changes in epithelial-mesenchymal transition. *PLoS ONE* 6(6): e20804.
2. Zhang B, Zhang Z, Xia S, Xing C, Ci X, Li X, et al. (2013) KLF5 Activates microRNA 200 Transcription to Maintain Epithelial Characteristics and Prevent Induced Epithelial-Mesenchymal Transition in Epithelial Cells. *Mol Cell Biol* 33(24):4919–35. <https://doi.org/10.1128/MCB.00787-13> PMID: 24126055
3. Dave N, Guaita-Esteruelas S, Gutarra S, Frias A, Beltran M, Peiro S, et al. (2011) Functional cooperation between Snail1 and twist in the regulation of ZEB1 expression during epithelial to mesenchymal transition. *J Biol Chem* 286:12024–12032. <https://doi.org/10.1074/jbc.M110.168625> PMID: 21317430
4. Casas E, Kim J, Bendesky A, Ohno-Machado L, Wolfe CJ et al. (2011) Snail2 is an essential mediator of Twist1-induced epithelial mesenchymal transition and metastasis. *Cancer Res* 71(1):245–254. <https://doi.org/10.1158/0008-5472.CAN-10-2330> PMID: 21199805
5. Park H, Shimamura T, Imoto S, Miyano S (2018) Adaptive NetworkProfiler for identifying cancer characteristic-specific gene regulatory networks. *J Comput Biol* 25(2). <https://doi.org/10.1089/cmb.2017.0120> PMID: 29053381
6. Maruhashi K, Todoriki M, Ohwa T, Goto K, Hasegawa Y, Inakoshi H, et al. (2018) Learning multi-way relations via tensor decomposition with neural networks. In: *AAAI* 3770-3777.

7. Maruhashi K, Park H, Yamaguchi R, Miyano S (2020) Linear Tensor Projection Revealing Nonlinearity. arXiv: 12007.03912, <http://arxiv.org/abs/2007.03912>.
8. Kc K, Li R, Cui F, Yu A, Haake AR (2019) GNE: a deep learning framework for gene network inference by aggregating biological information BMC Syst Biol 13(Suppl 2):38. <https://doi.org/10.1186/s12918-019-0694-y> PMID: 30953525
9. Mochida K, Koda S, Inoue K, Nishii R (2018) Statistical and Machine Learning Approaches to Predict Gene Regulatory Networks From Transcriptome Datasets Frontiers in Plant Science, 871. <https://doi.org/10.3389/fpls.2018.01770> PMID: 30555503
10. Hastie T, and Tibshirani R (1993) Varying-Coefficient Models. J Royal Stat Soc Ser B 55(4): 757–796.
11. Niida A, Smith AD, Imoto S, Aburatani H, Zhang MQ, et al. (2009). Gene setbased module discovery in the breast cancer transcriptome. BMC Bioinformatics 10:71. <https://doi.org/10.1186/1471-2105-10-71> PMID: 19243633
12. Larsen JE, Nathan V, Osborne JK, Farrow RK, Deb D et al. (2016) ZEB1 drives epithelial-to-mesenchymal transition in lung cancer. J Clin Invest 126:3219–3235. <https://doi.org/10.1172/JCI76725> PMID: 27500490
13. Chen W, Yi JK, Shimane T, Mehrazarin S, Lin YL, et al. (2016) Grainyhead-like 2 regulates epithelial plasticity and stemness in oral cancer cells. Carcinogenesis 37(5):500–510. <https://doi.org/10.1093/carcin/bgw027> PMID: 26933170
14. Kaufhold S, Bonavida B (2014). Central role of Snail1 in the regulation of EMT and resistance in cancer: a target for therapeutic intervention. J Experim & Clinic Cancer Res 33:62. <https://doi.org/10.1186/s13046-014-0062-0> PMID: 25084828
15. Leong KG, Niessen K, Kulic I, Raouf A, Eaves C, Pollet I, et al. (2007) Jagged1-mediated notch activation induces epithelial-to-mesenchymal transition through slug-induced repression of E-cadherin. J Exp Med 204:2935–2948. <https://doi.org/10.1084/jem.20071082> PMID: 17984306
16. Wels C, Joshi S, Koefinger P, Bergler H, Schaidler H (2011). Transcriptional activation of ZEB1 by slug leads to cooperative regulation of the epithelial-mesenchymal transition-like phenotype in melanoma. J Invest Dermatol 131:1877–1885. <https://doi.org/10.1038/jid.2011.142> PMID: 21593765
17. Yang MH and Wu KJ (2008) TWIST activation by hypoxia inducible factor-1 (HIF-1): implications in metastasis and development. Cell cycle 7:2090–2096. <https://doi.org/10.4161/cc.7.14.6324> PMID: 18635960
18. Daniel JG, Panizzi JR (2019) Spatiotemporal expression profile of embryonic and adult ankyrin repeat and EF-hand domain containing protein 1-encoding genes ankef1a and ankef1b in zebrafish. Gene Expr Patterns 34:119069. <https://doi.org/10.1016/j.gep.2019.119069> PMID: 31520739
19. Wang S, Yan S, Zhu S, Zhao Y, Yan J et al. (2018) FOXF1 Induces Epithelial-Mesenchymal Transition in Colorectal Cancer Metastasis by Transcriptionally Activating SNAI1. Neoplasia 20(10):996–1007. <https://doi.org/10.1016/j.neo.2018.08.004> PMID: 30189360
20. Wei HJ, Nickoloff JA, Chen WH, Liu HY, Lo WC et al. (2014) FOXF1 mediates mesenchymal stem cell fusion-induced reprogramming of lung cancer cells. Oncotarget 5(19):9514–9529. <https://doi.org/10.18632/oncotarget.2413> PMID: 25237908
21. Lo PK (2017) The controversial role of forkhead box F2 (FOXF2) transcription factor in breast cancer. PRAS Open 1:009. PMID: 29046905
22. Cai J, Tian AX, Wang QS, Kong PZ, Du X et al. (2015) FOXF2 suppresses the FOXC2-mediated epithelial-mesenchymal transition and multidrug resistance of basal-like breast cancer. Cancer Lett 367:129–137. <https://doi.org/10.1016/j.canlet.2015.07.001> PMID: 26210254
23. Iwasaki H, Nakano K, Shinkai K, Kunisawa Y, Hirahashi M et al. (2013) Hedgehog Gli3 activator signal augments tumorigenicity of colorectal cancer via upregulation of adherence-related genes. Cancer Sci 104(3):328–36. <https://doi.org/10.1111/cas.12073> PMID: 23176625
24. Rodrigues MFSD, Miguita L, De Andrade NP, Heguedusch D, Rodini CO (2018) GLI3 knockdown decreases stemness, cell proliferation and invasion in oral squamous cell carcinoma. Int J Oncol 53(6):2458–2472. <https://doi.org/10.3892/ijo.2018.4572> PMID: 30272273
25. Chung VY, Tan TZ, Tan M, Wong MK, Kuay KT et al. (2016) GRHL2-miR-200-ZEB1 maintains the epithelial status of ovarian cancer through transcriptional regulation and histone modification. Sci Rep 6:19943. <https://doi.org/10.1038/srep19943> PMID: 26887977
26. Xiang J, Fu X, Ran W, Wang Z (2017) Grhl2 reduces invasion and migration through inhibition of TGFβ-induced EMT in gastric cancer. Oncogenesis 9; 6(1):e284. <https://doi.org/10.1038/oncsis.2016.83> PMID: 28067907
27. Cieply B, Farris J, Denvir J, Ford HL, Frisch SM (2013). Epithelial-mesenchymal transition and tumor suppression are controlled by a reciprocal feedback loop between ZEB1 and Grainyhead-like-2. Cancer Res 15; 73(20):6299–309. <https://doi.org/10.1158/0008-5472.CAN-12-4082> PMID: 23943797

28. Mooney SM, Talebian V, Jolly MK, Jia D, Gromala M et al. (2017) The GRHL2/ZEB Feedback Loop-A Key Axis in the Regulation of EMT in Breast Cancer. *J Cell Biochem* 118(9):2559–2570. <https://doi.org/10.1002/jcb.25974> PMID: 28266048
29. Alimirah F, Chen J, Davis FJ, Choubey D (2007) IFI16 in human prostate cancer. *Mol Cancer Res* 5(3):251–9. <https://doi.org/10.1158/1541-7786.MCR-06-0269> PMID: 17339605
30. Lin W, Zhao Z, Ni Z, Zhao Y, Du W, Chen S (2017) IFI16 restoration in hepatocellular carcinoma induces tumour inhibition via activation of p53 signals and inflammasome. *Cell Prolif* 50(6). <https://doi.org/10.1111/cpr.12392> PMID: 28990231
31. Unterholzner L, Keating SE, Baran M, Horan KA, Jensen SB et al. (2010) IFI16 is an innate immune sensor for intracellular DNA. *Nature Immunology* 11:997–1004. <https://doi.org/10.1038/ni.1932> PMID: 20890285
32. Roy A, Ghosh A, Kumar B, Chandran B (2019) IFI16, a nuclear innate immune DNA sensor, mediates epigenetic silencing of herpesvirus genomes by its association with H3K9 methyltransferases SUV39H1 and GLP. *Elife* 8: e49500. <https://doi.org/10.7554/eLife.49500> PMID: 31682228
33. Ke CY, Xiao WL, Chen CM, Lo LJ, Wong FH (2015) IRF6 is the mediator of TGF β 3 during regulation of the epithelial mesenchymal transition and palatal fusion. *Sci Rep* 5:12791. <https://doi.org/10.1038/srep12791>
34. Li D, Cheng P, Wang J, Qiu X, Zhang X (2019) IRF6 Is Directly Regulated by ZEB1 and ELF3, and Predicts a Favorable Prognosis in Gastric Cancer. *Frontiers in Oncol* 9:200. <https://doi.org/10.3389/fonc.2019.00220> PMID: 31019894
35. Shimada H, Abe S, Kohno T, Satohisa S, Konno Y (2017) Loss of tricellular tight junction protein LSR promotes cell invasion and migration via upregulation of TEAD1/AREG in human endometrial cancer *Sci Rep* 7:37049. <https://doi.org/10.1038/srep37049> PMID: 28071680
36. Parsana P, Amend SR, Hernandez J, Pienta KJ, Battle A (2017) Identifying global expression patterns and key regulators in epithelial to mesenchymal transition through multi-study integration. *BMC Cancer* 17(1):447. <https://doi.org/10.1186/s12885-017-3413-3> PMID: 28651527
37. Reaves DK, Fagan-Solis KD, Dunphy K, Oliver SD, Scott DW et al. (2014) The role of lipolysis stimulated lipoprotein receptor in breast cancer and directing breast cancer cell behavior. *PLoS One* 9(3): e91747. <https://doi.org/10.1371/journal.pone.0091747> PMID: 24637461
38. Takano K, Kakuki T, Obata K, Nomura K, Miyata R et al. (2016) The Behavior and Role of Lipolysis-stimulated Lipoprotein Receptor, a Component of Tricellular Tight Junctions, in Head and Neck Squamous Cell Carcinomas. *Anticancer Res* 36(11):5895–5904. <https://doi.org/10.21873/anticancer.11176> PMID: 27793914
39. Liu J, Wu Q, Wang Y, Wei Y, Wu H et al. (2018) Ovovl2 induces mesenchymal-epithelial transition via targeting ZEB1 in osteosarcoma. *Oncotargets Ther* 11:2963–2973. <https://doi.org/10.2147/OTT.S157119> PMID: 29872308
40. Nilsson G and Kannius-Janson M (2016) Forkhead Box F1 promotes breast cancer cell migration by upregulating lysyl oxidase and suppressing Smad2/3 signaling. *BMC Cancer* 16:142. <https://doi.org/10.1186/s12885-016-2196-2> PMID: 26908052
41. Roca H, Hernandez J, Weidner S, McEachin RC, Fuller D (2013) Transcription factors OVOL1 and OVOL2 induce the mesenchymal to epithelial transition in human cancer. *PLoS One*. 8(10):e76773. <https://doi.org/10.1371/journal.pone.0076773> PMID: 24124593
42. Hong T, Watanabe K, Ta CH, Villarreal-Ponce A, Nie Q, Dai X. (2015) An Ovovl2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-step Transition between Epithelial and Mesenchymal States. *PLoS Comput Biol* 11(11):e1004569. <https://doi.org/10.1371/journal.pcbi.1004569> PMID: 26554584
43. Zhang Y, Liao Y, Chen C, Sun W, Sun X et al., (2019) p38 regulated FOXC1 stability is required for colorectal cancer metastasis, *Pathology* 250(2). PMID: 31650548
44. Chandhoke AS, Karve K, Dadakhujaev S, Netherton S, Deng L et al. (2016) The ubiquitin ligase Smurf2 suppresses TGF β -induced epithelial-mesenchymal transition in a sumoylation-regulated manner. *Cell Death Differ* 23(5):876–88. <https://doi.org/10.1038/cdd.2015.152> PMID: 26679521
45. Huang Y, Tong J, He F, Yu X, Fan L, et al. (2015) miR-141 regulates TGF- β 1-induced epithelial-mesenchymal transition through repression of HIPK2 expression in renal tubular epithelial cells. *Int J Mol Med* 35(2):311–8. <https://doi.org/10.3892/ijmm.2014.2008> PMID: 25421593
46. Moustakas A, Heldin CH (2016) Mechanisms of TGF β Induced Epithelial-Mesenchymal Transition. *J Clin Med* 5(7). pii: E63.
47. Saito RA, Watabe T, Horiguchi K, Kohyama T, Saitoh M et al. (2009) Thyroid transcription factor-1 inhibits transforming growth factor-beta-mediated epithelial-to-mesenchymal transition in lung adenocarcinoma cells. *Cancer Res* 69(7):2783–91. <https://doi.org/10.1158/0008-5472.CAN-08-3490>

48. Zhang Y, Yan W, Chen X (2014) P63 regulates tubular formation via epithelial-to-mesenchymal transition. *Oncogene* 33:1548–1557. <https://doi.org/10.1038/onc.2013.101> PMID: 23542170
49. Olsen JR, Oyan AM, Rostad K, Hellem MR, Liu J et al. (2013) p63 attenuates epithelial to mesenchymal potential in an experimental prostate cell model. *PLoS One* 8(5):e62547. <https://doi.org/10.1371/journal.pone.0062547> PMID: 23658742
50. Lindsay J, McDade SS, Pickard A, McCloskey KD, McCance DJ (2011) Role of DeltaNp63gamma in epithelial to mesenchymal transition. *J Biol Chem* 286(5):3915–24. <https://doi.org/10.1074/jbc.M110.162511> PMID: 21127042
51. Pascalis CD, Etienne-Manneville S (2017) Single and collective cell migration: the mechanics of adhesions. *Molecular Biology of the Cell* Vol 28(14). <https://doi.org/10.1091/mbc.E17-03-0134> PMID: 28684609
52. Mak AB, Nixon AM, Moffat J (2012) The mixed lineage leukemia (MLL) fusion-associated gene AF4 promotes CD133 transcription. *Cancer Res* 72(8):1929–34. <https://doi.org/10.1158/0008-5472.CAN-11-3589> PMID: 22337994
53. Thomas AM (2017) Mouse Models of MLL Leukemia: Recapitulating the Human Disease *Blood* 129(16): 2217–2223.
54. Zhou C, Xiong Q, Zhu X, Du W, Deng P, Li X, et al. (2017) AFF1 and AFF4 differentially regulate the osteogenic differentiation of human MSCs. *Bone Res* 5:17044. <https://doi.org/10.1038/boneres.2017.44> PMID: 28955517
55. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4:P3. <https://doi.org/10.1186/gb-2003-4-5-p3> PMID: 12734009
56. Tsai YC, Chen WY, Abou-Kheir W, Zeng T, Yin JJ, Bahmad H, et al. (2018) Androgen deprivation therapy-induced epithelial-mesenchymal transition of prostate cancer through downregulating SPDEF and activating CCL2. *Biochim Biophys Acta Mol Basis Dis* 1864(5 Pt A):1717–1727. <https://doi.org/10.1016/j.bbadis.2018.02.016> PMID: 29477409
57. Notario V, Lozano S, Villar J, Tirado OM (2010) Targeting EWS/FL1 Driven Signaling Pathways as Therapy for Tumors of the Ewing's Sarcoma Family. *Handbook of Cell Sign (Second Ed)*: 2859–2871.
58. Pallier K, Cazes Aurelie, El Khattabi Laila, Lecchi Cristina, Desroches Marine, Danel Claire, et al. Helene Blons DeltaN TP63 Reactivation, Epithelial Phenotype Maintenance, and Survival in Lung Squamous Cell Carcinoma. *Tumour Biol.* 2012 Feb; 33(1):41–51. <https://doi.org/10.1007/s13277-011-0239-5> PMID: 21986963
59. Geiger T, Sabanay H, Kravchenko-Balasha N, Geiger B, Levitzki A (2008) Anomalous Features of EMT during Keratinocyte Transformation. *PLoS One* 6; 3(2):e1574. <https://doi.org/10.1371/journal.pone.0001574> PMID: 18253510
60. Leopolda Philip L., Vincentb Jan, Wanga Hongjun (2012) A comparison of epithelial-to-mesenchymal transition and reepithelialization *Semin Cancer Biol* 22(5-6):471–83. <https://doi.org/10.1016/j.semcancer.2012.07.003>
61. Mehrzarin S, Chen W, Oh JE, Liu ZX, Kang KL, et al. (2015) p63 Gene is Regulated by Grainyhead-Like 2 (GRHL2) Through Reciprocal Feedback and Determines Epithelial Phenotype in Human Keratinocytes *J Biol Chem* 129(16), 2217–2223. <https://doi.org/10.1074/jbc.M115.659144> PMID: 26085095